MANUFACTURING ROADMAP FOR HETEROGENEOUS INTEGRATION AND ELECTRONICS PACKAGING (MRHIEP)

Final Report

authorship

This roadmap addresses advanced packaging and was lead by UCLA CHIPS and SEMI (USA) and was sponsored by NIST under award 70NANB22H038, and is submitted as final report.

Principal Investigators: Subramanian Iyer (till September 2023) and Gity Samadi.

Chapters

Acknowledgements

Authors and Editors

Glossary of terms

Chapter 1: MRHIEP Executive Summary

Chapter 2: Advanced Packaging & Heterogeneous Integration

Chapter 3: Medical/Hybrid Electronics

Chapter 4: Reliability and Thermal Challenges

Chapter 5: Modeling and Simulation

Chapter 6: Modular Chiplet Packaging for an Open Chiplet Economy

Chapter 7: Security in Heterogeneous Integration and Advanced Packaging

Chapter 8: Heterogeneous Integration Test Technology

Chapter 9: Advanced Packaging Supply Chain for High Performance Computing

Chapter 10: Smart Manufacturing Technology for Heterogeneous Integration & Advanced Packaging

Please select a chapter to continue reading.

Acknowledgement

The MRHIEP organizing committee extends its heartfelt gratitude to the dedicated volunteers and their respective companies for their unwavering support and significant contributions to the success of the MRHIEP Package Roadmap effort. Your commitment and expertise have played a pivotal role in development of practical and sustainable roadmaps for key advanced packaging areas including materials, manufacturing process flows, cross-cutting technologies, chiplet architectures, simulations, electrical & mechanical standards, Supply Chain, Security, Test and Smart Manufacturing.

We also thank our executive committee: Akshay Singh (Micron), Tom Rucker (Intel), Melissa Grupen-Shemansky (SEMI), Gity Samadi (SEMI), Subu Iyer (UCLA CHIPS till September 2023), Om Nalamasu (AMAT), Eric Forsythe (US ARL) for providing constant and valuable guidance.

We also extend our gratitude to the members of previously established Technical Working Groups (TWG) in the Heterogeneous Integration Roadmap (HIR) for their pioneering efforts and the wealth of insights they provided throughout the road-mapping process. Your groundwork has paved the way for our current initiative, and we appreciate the valuable knowledge and experience you have shared.

We thank Hetal Jain (UCLA CHIPS) for administrative support.

The chapters developed through the MRHIEP efforts have aimed to identify the gaps in the existing advanced packaging infrastructure in the US with viable approaches to overcome these gaps. In addition, they provide targeted metrics to be achieved over the next 15 years for a sustainable advanced packaging ecosystem. We expect MRHIEP to shape the future of heterogeneous integration, chiplet ecosystem and advanced packaging. We look forward to continuing this collaborative journey as we shape the future of manufacturing, heterogeneous integration and electronics packaging together.

Sincerely,

Krutikesh Sahoo
(on behalf of Subramanian Iyer, UCLA CHIPS and Gity Samadi, SEMI)
UCLA CHIPS
Los Angeles, CA
January 5, 2024

Authors:

Abhijit Dasgupta (Univ of Maryland)
Anu Ramamurthy (microchip)
Arvind Kumar (IBM)
Bapi Vinnakota (ODSA)
Benson Chan (Binghamton)
Benjamin Fasano (Consultant)
Boris Vaisband (McGill)
Daniel Berger (Consultant)
Dave Armstrong (Advantest)
Dharmesh Jani (META)
Eric Forsythe (US Army Research Lab)
Gamal Refai Ahmed (AMD)
Gerald Pasdast (Intel)
Gity Samadi (SEMI)
Habib Hichri (Ajinomoto Fine Techno US)
Hanwen Chen (Applied Materials)
Harry Chen (MediaTek)
Ira Leventhal (Advantest)
Jeorge Hurtarte (Teradyne)
Jerry McBride (Micron)
Jobert van Eisden (Atotech/MKS)
John Shalf (LBL)
Josh Dillon (Marvell)
Joy Watanabe (EMD Electronics)
Kanda Tapily (Tokyo Electron US)
Kemal Aygun (Intel)
Ken Butler (Advantest)
Ken Lanier (Teradyne)
Krutikesh Sahoo (UCLA CHIPS)
Lou Dadok (Fujifilm US)
Marc Hutner (Siemens)
Mark da Silva (SEMI)
Markus Leitgeb (AT&S)
Mary Ann Maher (SoftMems)
Michel Koopmans (Micron)
Morten Jensen (Intel)
Nader Sehatbakhsh (UCLA)
Naveed Hussain (AMAT)
Ram Kambhampati (Resonac US)
Reza Mahmoodian (Ulvac)
Rich Dumene (Renesas)
Robert Rodriquez (InnovaFlex Foundry)
Snehamay Sinha (Texas Instruments)
Steven Verhaverbeke (Applied Materials)
Subramanian S. Iyer (UCLA CHIPS) (till September 2023)
Tom Rucker (Intel)
Venky Sundaram (Consultant)
Vineeth Harish (UCLA CHIPS)
Wendy Chen (KYEC)

(Names in Bold indicate section leads)

Editors:

Gity Samadi
Krutikesh Sahoo
Subramanian S. Iyer (until September 2023)
Venky Sundaram
Vineeth Harish

Glossary of terms

3D IC Stacking	Techniques involving the integration of different technologies and the stacking of multiple integrated circuits (ICs) in three-dimensional configurations
3D Printing	Additive manufacturing technique creating three-dimensional objects layer by layer
ADAS	Advanced driver-assistance systems
Additive Manufacturing	Manufacturing techniques that build objects layer by layer, adding material
Advanced Driver Assist Systems (ADAS)	Vehicle safety systems that aid the driver in the driving process
Advanced Packaging	New generation of packaging technologies such as 2
Advanced Substrates	New generation of high-density substrates and interposers, including silicon, glass and organic substrates, essential for advanced packaging
AI (Artificial Intelligence)	The simulation of human intelligence processes by machines, especially computer systems
AI-Driven Inspection	Inspection processes utilizing artificial intelligence algorithms and techniques for automated and precise quality control
AOA	Angle of arrival
AR/VR (Augmented Reality/Virtual Reality)	Technologies that create immersive, computer-generated environments or enhance real-world experiences through digital overlays
ASIC (Application-Specific Integrated Circuit)	A customized integrated circuit designed for a specific purpose or application, often used in high-performance computing and specialized devices
ASP	Average sales price
ATE	Automatic test equipment
ATPG	Automatic test pattern generation
Autonomous Driving	The ability of a vehicle to operate without human intervention, using sensors and software to navigate and control the vehicle
AXI/CHI	Common bus protocols used in System-on-Chip architectures
Backend Issues	Challenges related to chiplet integration that typically occur after the initial design phase, including packaging, inventory management, and testing
BERT	Bit-error-rate tester
BGA	Ball grid array
Biocompatibility	The ability of materials and substances to be compatible with living tissues and biological systems without causing harm or rejection

BISC	Built-in self-correlation/compensation
BISD	Built-in self-diagnostics
BIST	Built-in self-test
BOST	Built-out self-test
BOW	Bunch of wires
Bumping and Assembly	Manufacturing processes involving the attachment of copper and/or solder bumps and assembly of electronic components onto substrates
BW	Bandwidth
CAGR	Compound annual growth rate
Capital Investments	Financial resources allocated for the development, manufacturing, and improvement of chiplet-based products, including investments in packaging technologies
CCC	Current carrying capacity
CHB	Copper hybrid bonding
Chiplet	Small, individual semiconductor components that need to be integrated with other chiplets to create a functional product, as opposed to monolithic devices which are standalone integrated circuits
Co-Packaged Optics (CPO)	Integration of optical components within electronic packaging for high-bandwidth and low-power data transmission
COT	Cost of test
CPS	Cyber-physical systems
CPU	Central processing unit
CPU (Central Processing Unit)	The primary component of a computer that performs most of the logic processing inside the computer
Cu-Cu Hybrid Bonding	Bonding technique involving the use of direct copper-to-copper bonding without solder, in combination with oxide-to-oxide bonding to form ultra-fine pitch interconnections, typically between two silicon chips or substrates
D2D Interconnect and PHY	Protocol and analog logic used to connect two chiplets in a package
DARPA	Defense Advanced Research Projects Agency
DDR	Dual data rate (memory)
DFT	Design for testability
DIB	Device interface board
Die-to-Die Interfaces	Connections established between individual chiplets within a chip package, allowing them to communicate and work together
Dielet	Hard instantiation of a chiplet design
DIP	Dual inline package
DPW	Die per wafer
DRAM	Dynamic random-access memory

DSP	Digital signal processing
DUT	Device under test
DVFS	Dynamic voltage and frequency scaling
ECID	Electronic chip identifier
EDA	Electronic design automation
EIC	Electronic integrated circuit
Electrochemical Sensing	Sensing technology based on chemical reactions involving electricity
Electronics	Refers to components and systems involving electrical circuits and devices
Emerging Technologies	Novel and developing technologies that have the potential to significantly impact various industries and everyday life
EPDA	Electronic/photonic design automation
EVM	Error vector magnitude
Extreme Environmental Conditions	Harsh or challenging environments that require specialized electronic components
Fan-Out Package	Packaging technology where redistribution layers are used to expand the on-chip IO area and enable direct board-level assembly of chips
FFT	Fast Fourier transform
Flexible Fanout Wafer Level Packaging	Flexible packaging technique at the wafer level, enabling miniaturized hybrid electronic systems
Flexible Substrates	Materials that can be bent and shaped without breaking, used in flexible electronics
Flextrate™	Die-first integration on flexible substrate followed by molding and RDL buildup using wafer level processes.
Foundry Capacity	The ability of semiconductor foundries to produce chips in large quantities
Front-end Device Manufacturing	The process of fabricating semiconductor devices on the front-end of the production line
Fugaku Supercomputer	A high-performance supercomputer developed by RIKEN and Fujitsu, currently one of the fastest supercomputers in the world, used as a reference for deriving the modular architecture in the report
GAN	Gallium nitride
Gbps	Gigabits per second
Geo-politics	The study of the effects of geography on international politics and international relations, specific to semiconductor supply chains in this report
GPIO	General purpose input/output
GPU	Graphics Processing Unit: A specialized electronic circuit designed to accelerate the processing of images and videos in a computer

Guardrails	Set boundaries or limitations within a modular architecture, defining constraints such as die size, bandwidth, thermals, and other attributes critical to the final product's design and manufacture
GUI	Graphical user interface
HB	Hybrid Bonding
HBM	High bandwidth memory
HCI	Hot carrier injection
HDD	Hard disk drive
Heterogeneous Integration	Combining different types of chiplets, each optimized for specific tasks, within a single package to enhance overall performance and functionality
High Performance Computing (HPC)	Computing systems that deliver high performance for solving complex computational problems
HIR	Abbreviation for the Heterogenous Integration Roadmap, a comprehensive technology roadmap for the future of semiconductor devices, packages and electronics systems
HSIO	High speed input/output
HVM	High-Volume Manufacturing, indicating large-scale production of electronic components
Hybrid Electronics	Combining diverse technologies into a unified and flexible substrate, enhancing the functionality of electronics used for medical and wearable applications
IJTAG	Internal Joint Test Action Group, refers to the IEEE 1687 family of standards
Interposer	A component used to connect semiconductor components within a package, typically at a level between the chips and the package substrate
IoT	Internet of Things, a network of interconnected devices and objects exchanging data
IP	Intellectual property
JTAG	Joint Test Action Group, refers to IEEE 1149 family of standards
KGD	Known good die
LBIST	Logic built-in self-test
LGA	Land grid array
LiDAR	Light Detection and Ranging, a remote sensing method using laser light for measuring distances
Lithography	The process of creating intricate patterns on surfaces using light or radiation, a crucial step in manufacturing electronic components and packages
Low-cost Regions	Geographical areas with lower labor and production costs, often targeted for outsourcing purposes
M2M	Machine-to-machine

Manufacturing Blueprint	A detailed plan outlining generic process flows, material and tool sets, and major suppliers for various packaging platforms
Manufacturing Equipment	Machinery and tools used in the manufacturing process of semiconductor devices
Materials & Chemicals	Raw materials and chemicals used in the production of semiconductors
MBIST	Memory built-in self-test
Memory	Electronic components used to store data and instructions temporarily or permanently in a computer system
Micro-fluidic Components	Miniaturized devices used for manipulating small amounts of fluids
MISR	Multiple input signature register
Modular Architecture	A design approach where a system is divided into smaller, manageable modules, allowing for flexibility, scalability, and ease of integration
Moore's Law	The observation that the number of transistors on a microchip doubles approximately every two years, leading to increased computing power
MQTT	Message queueing telemetry transport
MRHIEP	Manufacturing Roadmap for Heterogenous Integration and Electronics Packaging
MSE	Multi-site efficiency
NBTI	Negative bias temperature instability
Noninvasive	Procedures or devices that do not penetrate the body
NRE	Non-recurring engineering
OEE	Overall equipment efficiency
Off-shoring	The practice of relocating a business operation or process to another country
Onshore Supply Chain	Manufacturing processes and resources located within the domestic boundaries of a country, ensuring self-sufficiency and reduced dependency on external sources
Onshoring	The practice of bringing manufacturing operations and jobs back to the domestic country from overseas locations
OOK	On-off keying
Optical Sensing	Sensing technology using light properties to measure various parameters
OSAT	Outsourced semiconductor assembly and test
OTA	Over the air
Outsourcing	The practice of contracting out certain business functions or processes to external third-party vendors

Overlay Accuracy	Precision in aligning different layers or components during the manufacturing process, ensuring high density multi-layer RDL structures, or chip assemblies
PAM4	Pulse amplitude modulation 4-level
Panel-Level Packaging (PLP)	Fan-out packaging performed at the panel level, enabling cost reduction and larger package sizes than wafer-level fanout packages
PCB	Printed circuit board
PDK	Process design kit
Photonic Integrated Circuit (PIC)	Integrated circuit technology for manipulating light in optical systems
Photonic IO	Optical interconnections between components related to data transmission and communication using light
Photonics	Technology related to the generation, transmission, and manipulation of light
PIC	Photonic integrated circuit
PKG	Package or packaging
Plastronics	Technology combining plastic and electronic components
Polarization Maintaining Fiber (PMF)	Optical fiber that maintains the polarization state of light, used in advanced optical systems
PoP	Package on Package, a stacking technique where one chip package is placed on top of another
PRBS	Pseudo-random binary sequence
PSS	Portable stimulus standard
PTE	Parallel test efficiency
PV	Photovoltaic
QAM	Quadrature amplitude modulation
QED	Quick error detect
QFN	Quad flat no-lead package
QFP	Quad flat pack
QPSK	Quadrature phase shift keying
RDL (Redistribution Layer)	A layer of metal traces used to redistribute electrical connections on semiconductor devices and packages
RF	Radio frequency
RMA	Return material authorization
Scale-down	The reduction in electrical and/or photonic IO pitch, enabling increased channels per package and higher bandwidth
Scale-out	Increasing system-level computing capacity by adding more discrete units based on a massively parallel architecture
Scaling	The ability to increase the performance, capacity, or capabilities of a chiplet-based product, often

	accomplished by optimizing existing technology or adopting new packaging methods
SDC	Silent data corruption
SECS	Semiconductor equipment communications standard
SFDR	Spurious-free dynamic range
SIC	Silicon carbide
Si-IF	Silicon Interconnect Fabric
Single Mode Fiber (SMF)	Optical fiber designed to carry a single light mode, used in high-speed data transmission
SiP	System in Package, a packaging technology where multiple chips and passive components are integrated within a single package
SIP	System in package
SLT	System level test
SNR	Signal to noise ratio
SOC	System on chip
Soft Robotics	Field of robotics dealing with soft and flexible robots
SOP	Small-outline package
SSD	Solid state drive
Standards	Established guidelines and specifications that define various aspects of chiplet integration, including packaging, mechanical properties, thermal management, and power delivery, ensuring compatibility and interoperability among different vendors' products
STDF	Standard test data format
SuperCHIPS	Simple Universal Parallel intERface for CHIPS – high bandwidth, low power, low latency dielet-to-dielet communication protocol.
Supply Chain Networks	Interconnected systems of organizations, people, activities, information, and resources involved in the production and distribution of goods and services
Supply Chain Resiliency	The ability of a supply chain to recover and adapt swiftly in the face of challenges, ensuring consistent production and delivery
Tailwinds	Favorable external factors or trends that support a particular industry or business
TAM	Test access mechanism
TCB (Cu - Cu)	Cu to Cu Thermal Compression Bonding without solder
TDDB	Time-dependent dielectric breakdown
TDE	Touchdown efficiency
THD	Total harmonic distortion
Thermal Dissipation	The process of dissipating heat generated by electronic components to prevent overheating and ensure optimal performance

Thermal Management	Techniques and technologies for controlling and optimizing the temperature of electronic devices and systems
Throughput	The rate at which a process or system can complete tasks or transactions within a specific time frame, crucial for efficient manufacturing operations
TOF	Time of flight
TSOP	Thin small-outline package
TSV	Through silicon via
TSVs (Through-Silicon Vias)	Vertical conduits passing through a silicon wafer, enabling connections between stacked ICs
TTM	Time to market
TWG	Abbreviation for Technical Working Group, focused on key identified areas
TWR	Two way ranging
UCIe, Bunch of Wires, XSR	Standards for D2D interconnect
UPH	Units per hour
UWB	Ultra wideband
VCSEL	Vertical-cavity surface-emitting laser
Wireless Power Transfer	Technology enabling the transfer of power without physical connections
WLCM	Wafer-level camera module
WLCSP	Wafer-level chip-scale packaging
WLO	Wafer-level optics
WLP	Wafer-level packaging

Chapter 1: Report Summary

Contents:

1.1 MRHIEP Goals and Organization 1

1.2 MRHIEP Roadmap Challenges 2

1.3 TWG1: Advanced Packaging Platforms 9

1.4 TWG2: Cross-cutting Technologies 12

1.5 TWG3: Chiplet Architectures and Standards 12

1.6 TWG4: Supply Chain, Security, Test and Smart Manufacturing 13

1.7 Example of Manufacturing Gaps and Challenges: A Global Supply Chain Perspective on Advanced Substrates 14

1.1 MRHIEP Goals and Organization

The goal of MRHIEP is to develop an operational road map for jump starting advanced packaging in the US, with the creation of a quick-start guide for on-shore rapid development, piloting prototyping and manufacturing. This manufacturing roadmap is inspired by the Heterogeneous Integration Roadmap (HIR). MRHIEP focuses on leveraging on-shore skills, capabilities, and infrastructure, towards building on-shore resiliency with a diverse, robust, and secure global supply chain. MRHIEP is focused on defining a manufacturing-centric packaging roadmap for two major segments, (1) High performance computing (HPC) and (2) Medical electronics & hybrid device packaging. It is believed that these two sectors can provide a foundational developmental roadmap for other applications sectors such as rf/mm wave, automobile, and power electronics as well.

MRHIEP was organized into four technical working groups (TWGs) with major themes as shown below:

TWG1: Advanced Packaging Platforms
TWG2: Cross-cutting Technologies (Thermal, Reliability, Modeling and Simulation)
TWG3: Chiplet Architectures and Standards
TWG4: Supply Chain, Security, Test and Smart Manufacturing

The TWG members provided detailed input to the TWG leaders and this forms the basis of this report.

Additionally, the roadmap was validated by an industrial board of advisors. A three person steering committee from UCLA CHIPS and SEMI provided day-day operational guidance.

1.2 MRHIEP Roadmap Challenges

The roadmap challenges can be summarized as shown in Figure 1.1 and elements of each will be covered in the technical working group (TWG) summaries in this report.

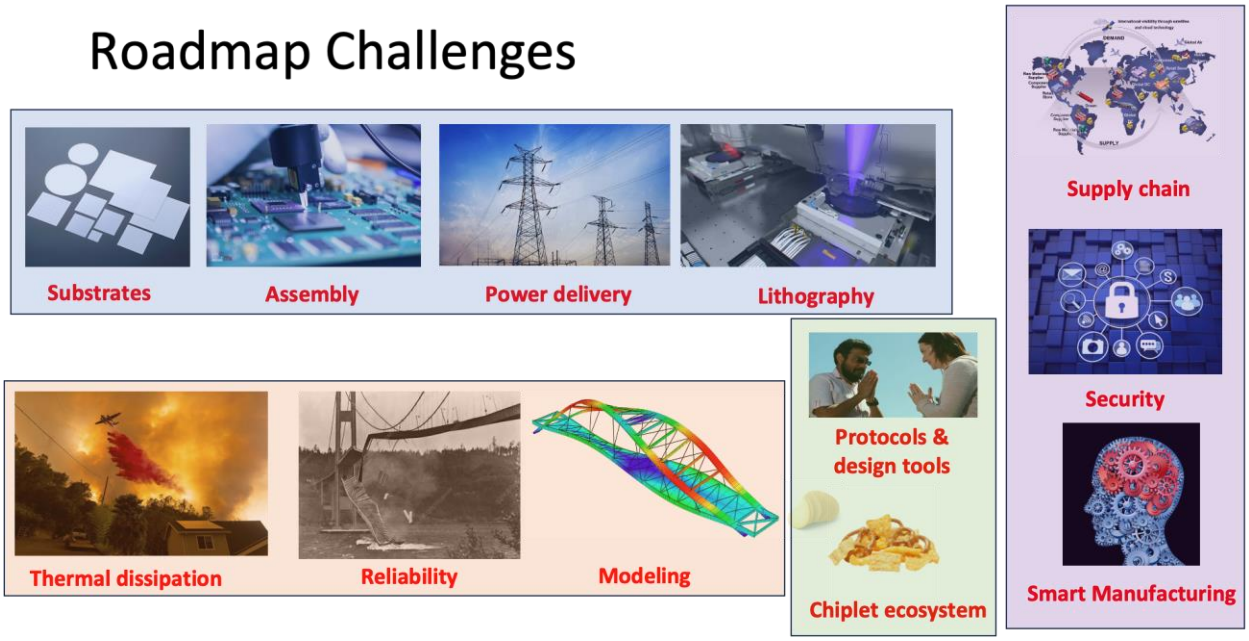


Figure 1.1. MRHIEP Manufacturing Roadmap Challenges

Several detailed roadmap charts were compiled by the MRHIEP team to represent the system-level roadmap requirements for high performance computing. These were translated to key advanced packaging metrics as shown below in Figure 1.2(a-f). These charts were extrapolated from the HIR roadmap to provide a more manufacturing-based visual of future trends until 2035. The same information is also provided in tabular form in Table 1.1 (a-f).

Table 1.1 (a) MRHIEP roadmap requirements for wafer-to-wafer bond pitch.

year of manufacturing	-3σ (µm)	Nominal wafer-to-wafer bond pitch (µm)	+3σ (µm)
1995	8	10	12
2005	4	5	6
2015	2.4	3	3.6
2025	1.12	1.4	1.68
2035	0.56	0.7	0.84

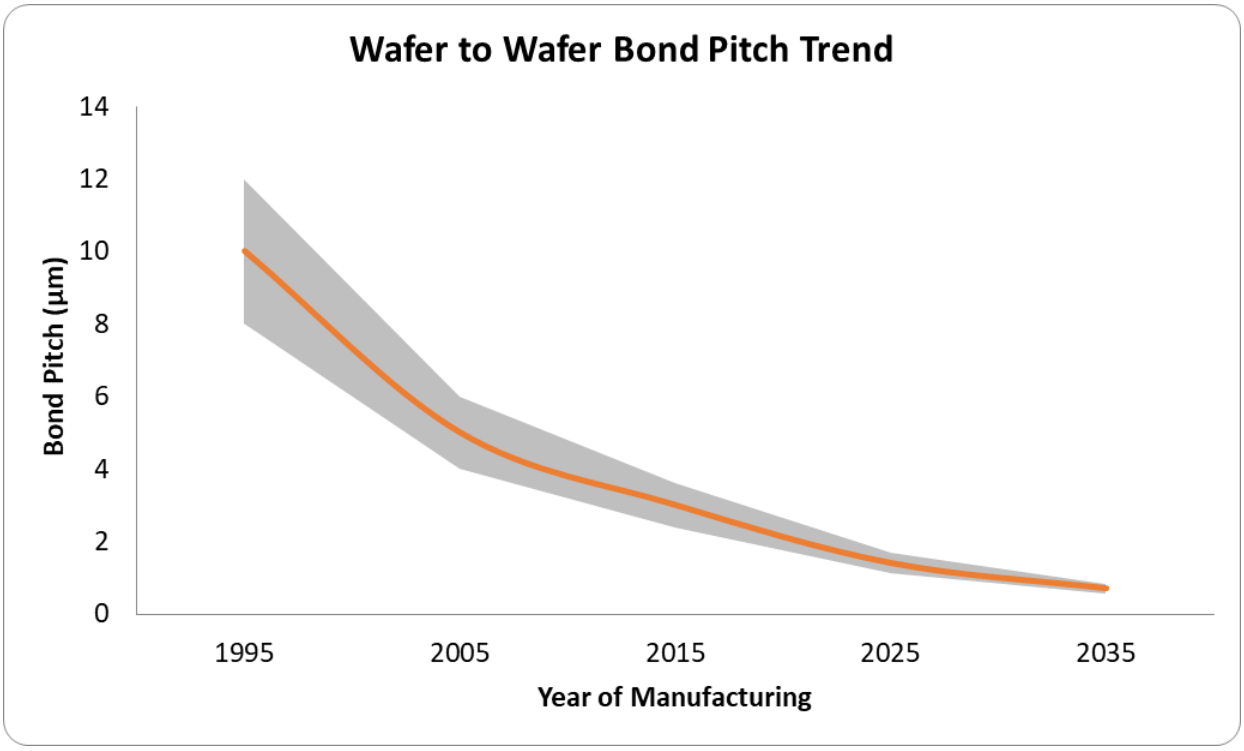


Figure 1.2 (a) MRHIEP roadmap requirements for wafer-to-wafer bond pitch with ± 3σ.

Table 1.1 (b) MRHIEP roadmap requirements for die-to-die and die-to-wafer bond pitch.

Year of manufacturing	-3σ (µm)	Nominal die-to-die & die-to-wafer bond pitch (µm)	+3σ (µm)
2023	7	10	13
2026	3.5	5	6.5
2029	1.75	2.5	3.25
2032	0.875	1.25	1.625
2035	0.49	0.7	0.91

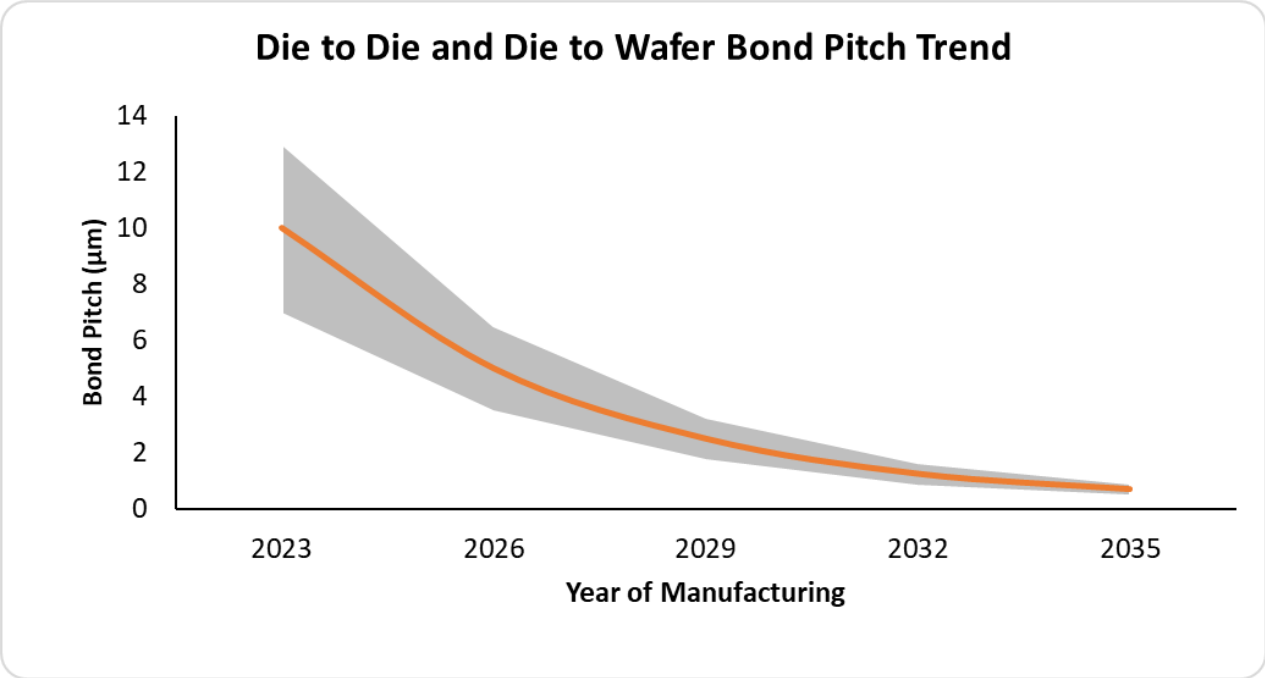


Figure 1.2 (b) MRHIEP roadmap requirements for die-to-die and die-to-wafer bond pitch with ± 3σ.

Table 1.1 (c) MRHIEP roadmap requirements for fanout wafer level packaging (FOWLP) pitch.

Year of manufacturing	Contact pitch, C (µm)	Trace pitch, T (µm)	C +3σ (µm)	C -3σ (µm)	T +3σ (µm)	T -3σ (µm)
2023	40	20	44	36	22	18
2025	28	14	30.8	25.2	15.4	12.6
2027	19.6	9.8	21.56	17.64	10.78	8.82
2029	13.72	6.86	15.09	12.34	7.54	6.17
2031	9.60	4.80	10.56	8.64	5.28	4.32
2033	6.72	3.36	7.39	6.05	3.69	3.02
2035	4.70	2.35	5.17	4.23	2.58	2.11

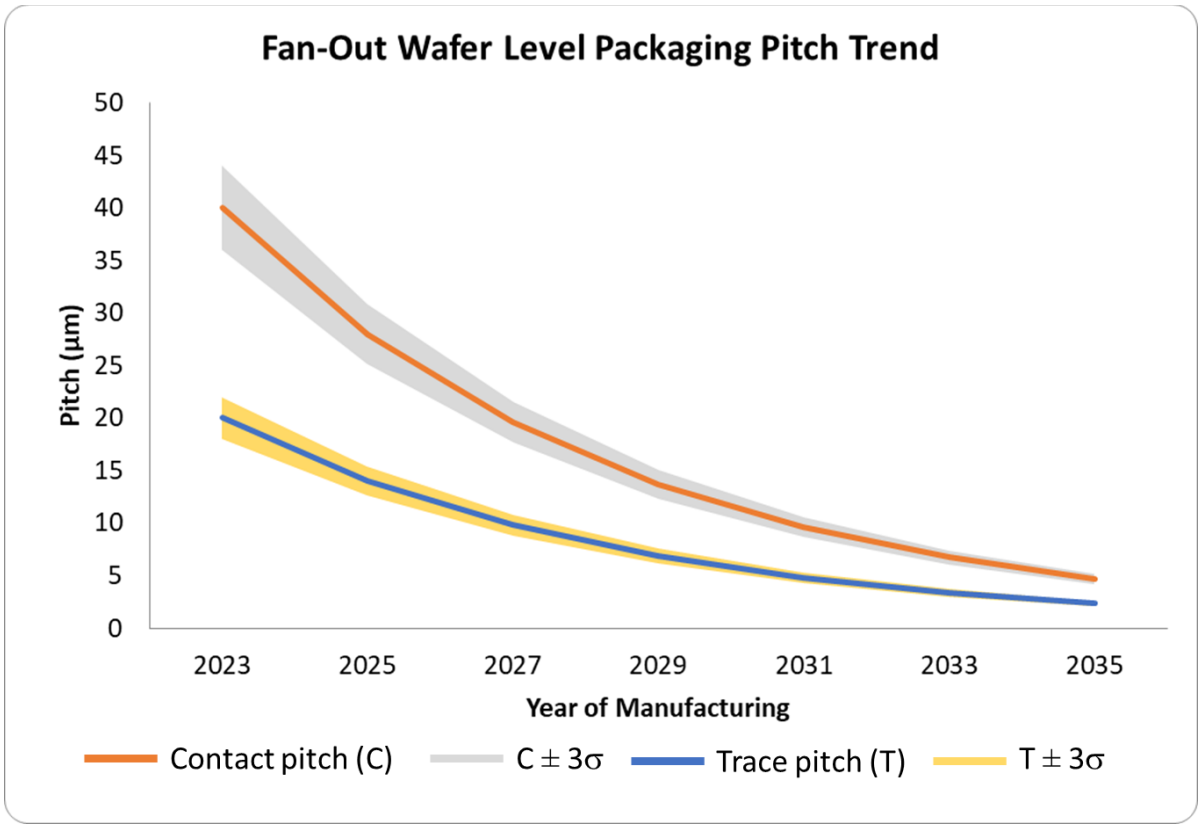


Figure 1.2 (c) MRHIEP roadmap requirements for fanout wafer level packaging (FOWLP) contact and trace pitch with ± 3σ.

Table 1.1 (d) MRHIEP roadmap requirements important silicon substrate parameters.

Year of Manufacturing	Number of Wiring Layers	Wiring Pitch (µm)	Landed Via size (µm)	Overlay (µm)
2023	4.00	2.00	1.00	0.70
2026	6.00	1.40	0.70	0.49
2029	9.00	0.98	0.49	0.34
2032	14.00	0.69	0.34	0.24
2035	20.00	0.48	0.24	0.17

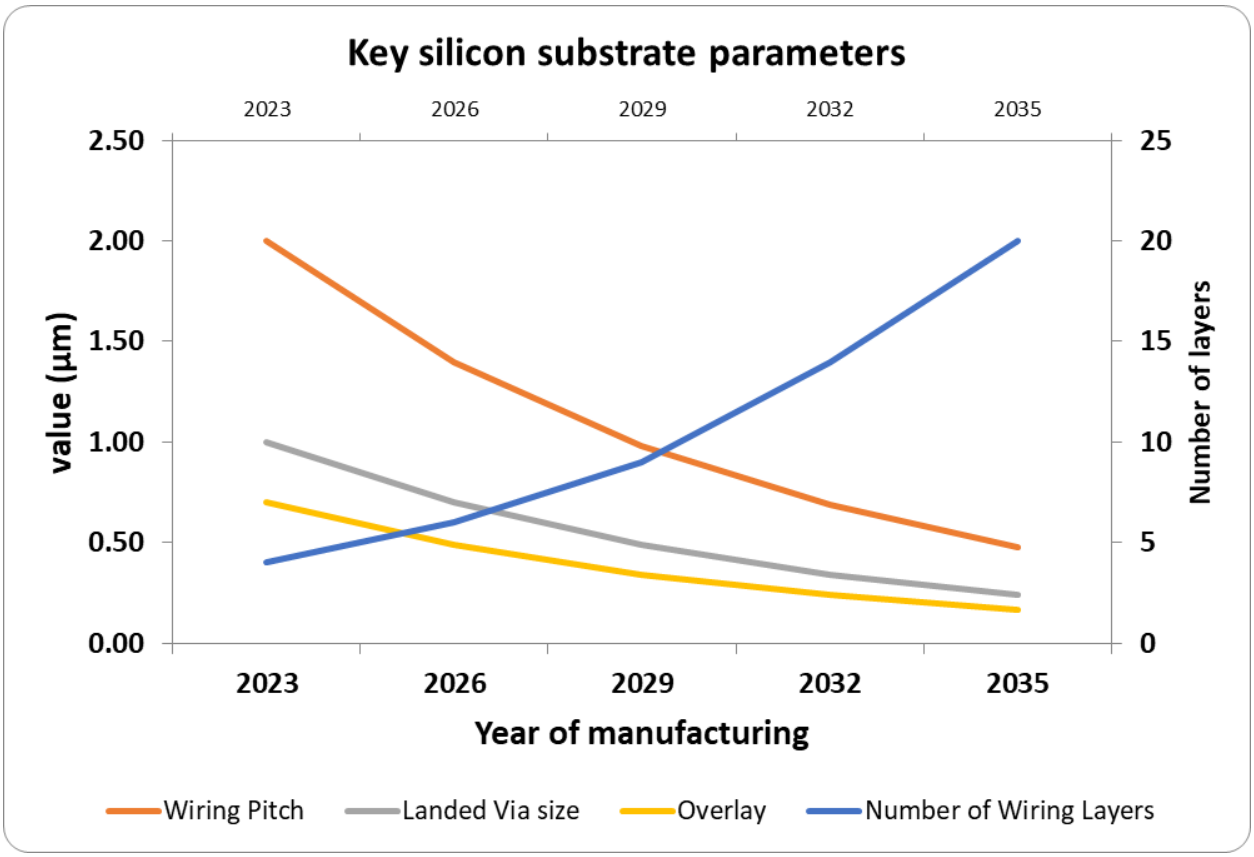


Figure 1.2 (d) MRHIEP roadmap requirements for important silicon substrate parameters such as substrate wiring pitch, via size and number of wiring layers.

Table 1.1 (e) MRHIEP roadmap requirements for thermal density for high bandwidth memory.

Year of Manufacturing	Thermal Density (W/mm^2)	# of Stacked Dies
2023	0.15	12
2026	0.3	24
2029	0.6	36
2032	1.2	48
2035	2.04	55

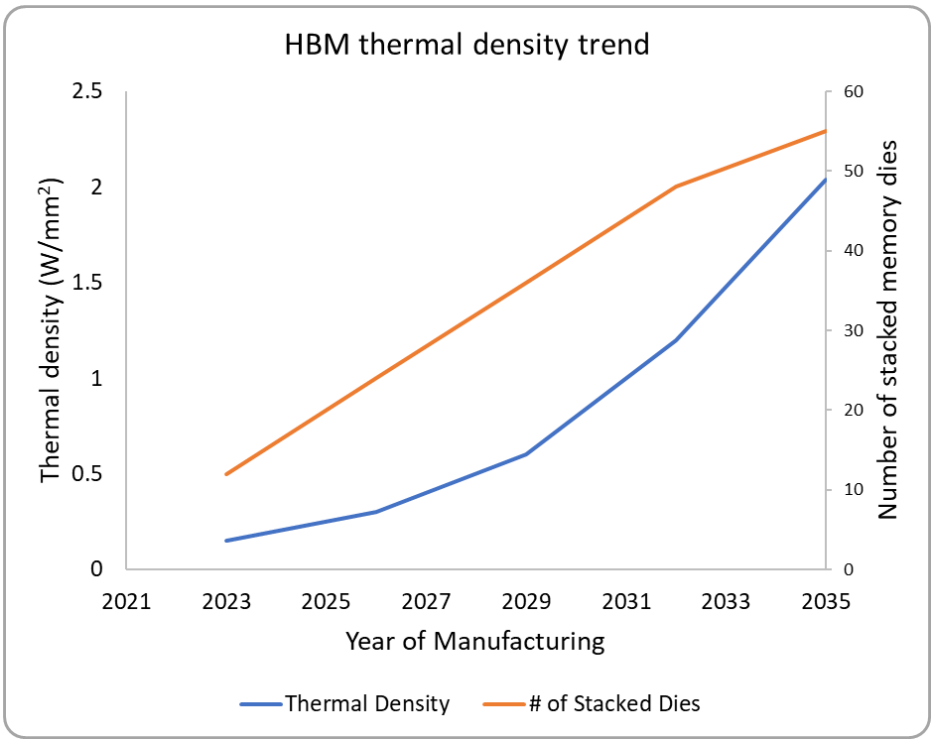


Figure 1.2 (e) MRHIEP roadmap requirements for thermal density for high bandwidth memory.

Table 1.1 (f) MRHIEP roadmap requirements for thermal density for logic strata.

Year of Manufacturing	Thermal Density (W/mm^2)	# of Stacked logic strata
2023	1	2
2026	2	3
2029	3	4
2032	4	5
2035	5	5

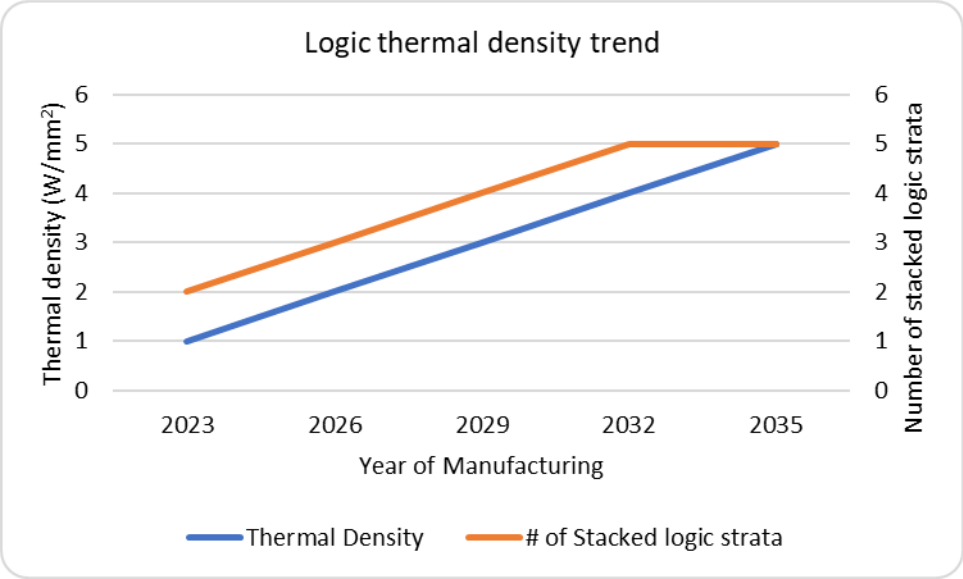


Figure 1.2 (f) MRHIEP roadmap requirements for thermal density for logic strata.

High Performance Computing (HPC): The roadmap specifies the scale-down and scale-out of packaging solutions that will integrate an ever increasing number of heterogeneous dielets to provide more functionality than can be provided by monolithic solutions alone, and, at lower cost, higher performance and lower power. Scale-down refers to the increase of channels per package through a steady reduction of all packaging dimensions. For example, bump pitches of today's advanced packaging will need to scale-down from ~30-50µm to approach the via pitches of on-chip via numbers of 1um or even <1um pitches. Photonic I/Os will require decreases in fiber pitch to 80um and less with increasing fiber count from 4-8 fibers today to numbers approaching 100. The main goal of scaling-down pitches is to reduce area per IO, reduce energy /bit for communication across the system, and reduce latency. Scale-out refers to more intimately connected semiconductors (Si, III-Vs etc), and other functional elements (passives, sensors, energy storage) through an expanding use of chiplets/dielets (rather than large chips) that are architected to work together in a system of computation, uniform shared memory with "uniform and everything everywhere" connectivity. Heterogeneous Integration (HI) will require standards to allow for their reuse in an ever-increasing number of applications of these chiplet building blocks. These chiplets will need to be designed, modeled, and integrated within the application performance, reliability, thermal budget, cost, and system level link budget requirements with a special emphasis on dielet reuse. A chiplet warehousing strategy will need to be developed based on a chiplet discovery methodology. Chiplet designs will need to be widely available as bare dielets that can be integrated into user defined customizable assemblies with minimal design resources and a versatile automated design system. Assembly tool improvements to address combining "round" package elements and "square/rectangle" package elements will be needed for high volume manufacturing. These assembly tools and associated test methodologies with be a significant add to the existing packaging tooling available. Managing thermal dissipation will be in increasing focus as well, especially with the expansion of 3D stacking of dielets, closer dielet to dielet spacing (also to approach ~1um) and the use of a broad array chiplet types to include CPUs, GPUs, other domain specific compute engines, and diverse memory types and connectors (including photonics) that connect the packages to other packages and to the outside world. We

foresee an eventual radical departure from today's multi hierarchical packaging architecture to a simpler hierarchy, and the gradual limited use of interposers being replaced by the direct assembly of heterogeneous dielets on advanced substrates that exceed the connection densities of interposers and which also provide significant more functionality than interposers. These advanced substrates are essentially interconnecting fabrics based on silicon, (including heterogeneous semiconductors substrates such as Si, GaN on Si and high conductivity SiC). These silicon interconnect fabrics (Si-IFs) will have built-in passives, power delivery features and capability of both organic and inorganic buildup layers to extend their functionality. Furthermore, substrates based on glass cores with organic and inorganic build up layers and embedded active dies and passives as well as laterally composite substrates (also called compliant substrates) to allow for thermal expansion slack between rigid segments (similar to the rubberized fill between concrete slabs) are also possible for low power applications. For high performance applications, we see an eventual phasing out of organic cores and their replacement with semiconductor and glass cores each with multiple stress-balanced fine pitch wiring layers on both sides. We expect both sides of the substrate to be populated. We expect multiple substrates to be electrically or optically connected to further expand the electrical footprint of complex HPC systems.

Another major driver of the HPC roadmap going forward is the expanding use of high speed connectors that may also include both rf/mm wave as well as co-packaged optics (CPO,) using new advanced packaging techniques. This is particularly important as AI data center and inter data center applications expand. Development needed to increase both wire and fiber density in these connectors with significant improvements in integration. Miniaturization and integration of rf/mm wave and photonic elements is a key to widespread adoption. The trade-off between bandwidth, bit error rate and power will be a major activity in the coming years with the emphasis being on reach, overall miniaturization (light source modulators, de-modulators and other electronics) cost and net power. We believe that wherever possible a wired solution will outperform a rf and photonics solution, though rf/mm wave solutions do present security vulnerabilities. Improvements in EDA systems to incorporate wired, rf/mm wave and photonic elements including their thermal environments will need to be made.

1.3 TWG1: Advanced Packaging Platforms

This technical working group consists of two sub-groups, focused on (a) High performance computing electronics, and (b) Medical and wearable hybrid electronics.

The Goal of TWG1 is to create a generic manufacturing roadmap and blueprint for manufacturing execution in key identified areas, building from HIR Roadmap and other relevant industry roadmaps. Advanced Packaging platforms have become critical to scaling electronic systems, yet a number of critical gaps exist in the onshore supply chain. The first edition of the manufacturing blueprint is focused on manufacturing gaps and potential solutions driven by the relevant technology roadmap chapters extracted from multiple editions of the HIR roadmap. The blueprint lays out detailed generic process flows with material and tool sets, as well as major suppliers (non-exhaustive) for each of the major packaging platforms. The working group has created 3-, 5- and 7-year targets and expanded on the onshoring gaps and solutions to create an actionable manufacturing roadmap. The major technology platforms included in the manufacturing blueprint with a focus on onshoring needs, are (a) Advanced Substrates, (b) Bumping and Assembly, (c) Hybrid Bonding and 3D IC stacking, and (d) Fan out packaging – wafer and panel level.

Key manufacturing roadmap gaps & challenges highlighted in this report are listed below.

- No advanced substrate manufacturing capability in the US.
- Equipment and material enablement gaps exist in all technology platforms to meet end user needs in a 3-10 year time frame.
- Onshore gaps exist in assembly and test at fine pitches.
- Power and thermal challenges in the roadmap need new solutions to support scaling.
- Hybrid electronics needs significant investments and additional focus inside HIR and other roadmaps.

The gaps and opportunities have been further sub-divided into two categories, namely,

A. Leading-edge Gaps that Create Opportunities
   - There is currently no high-volume, silicon-based package manufacturing infrastructure in the US.
   - Die-to-Die interconnect pitch scaling roadmaps create new opportunities to address lithographic tools and process gaps for large area patterning.
   - Bond pitch scaling with hybrid bonding and alternate assembly methods require innovations in plasma or other dicing, cleans and metrology steps to achieve high yields and cost-effective volume manufacturing.

B. Supply Chain Resiliency Gaps
   - The biggest gap in the onshore packaging supply chain for high performance computing is the lack of any advanced organic substrate manufacturing infrastructure in the US.
   - Addressing the lack of non-captive, high volume bumping and assembly infrastructure in the US is another key to ensuring supply chain resiliency.

Substrates/Interposers: Wide-area lithography that can scale to sub-micron dimensions is a major gap in the global as well as onshore supply chains. The basis for this gap is the combination of reticle stitching & layer count escalation, resulting in worst case scenarios with >100 unique masks to build RDL for one interposer design. Although large area projection printing can scale to 2um today, there is concern whether this platform can be extended to sub-micron pitches, while maintaining large image field areas. Direct write lithography has emerged in recent years as a viable alternative. However, concerns remain about the ability to achieve high throughput, high overlay accuracy, and high resolution, as package sizes increase beyond 100mm x 100mm and substrate warpage increases. Other major needs in substrates and interposers include metrology and electrical test methods and tools for yield management. The emergence of automated AI driven inspection has been identified as a promising direction for future substrates and interposers. Passive integration for power delivery efficiency will continue to be adopted in wafer and panel formats and further material as well as process innovations are necessary. Advances in GaN on Si and SiC substrates offer the ability to revolutionize power delivery making highly segmented, multi-domain efficient power delivery a reality within the next few years.

Bond Pitch Scaling & Assembly: Manufacturing challenges and gaps exist in increasing throughput for Cu-Cu hybrid bonding(HB) and direct thermo-compression bonding)TCB) (die to substrate) as bond pitches scale to below 10 microns. Key challenges for HB lie in process

tolerance, wafer reconstitution and shear strength especially at high connection densities. New methods such as plasma dicing and cleans to eliminate particle contamination will need to be introduced as hybrid bonding scales to high volume manufacturing. While TCB is more forgiving from a process tolerance perspective, TCB equipment needs significant improvement from an automation and alignment perspective. Continued focus needs to be placed on improved handling methods for thinner die with through semiconductor/substrate vias (TSVs). Another challenge will be on reducing the die to die spacing for use in assembly of die to wafer, or collective die to wafer integration techniques. Lastly, Wafer-scale RDL lithography and cross die topography management will be another manufacturing gap as bond pitch scales.

Fan-out WLP/PLP: Managing die shift and warpage as well as overlay accuracy are the major concerns in scaling the IO pitch for fanout packages. Better materials are needed for improved thermal dissipation as power density increases. The large area lithography challenge identified for substrates and interposers is even more critical to scale bond pitches for panel-level fanout packages (PLP). Recent innovations in carrier bonding and debonding will need to continue to progress to achieve the target process yields. Fanout approaches may also be used to build laterally heterogeneous substrates (including organic and glass core) to allow for accommodation of thermal expansion. This combined with AI mediated direct write lithography will allow for finer overall features over large areas enabling scale-down and scale-out.

Medical/Hybrid Electronics: For medical/hybrid device packaging, increasing utilization and extension of flexible substrates through materials development and tooling to allow for broader technology application space that will include, asset monitoring such as electronics integrated onto large 3D-surfaces, communications arrays and associated electronics, soft robotics, electronics for extreme environmental conditions, to name a few. Medical applications are extensive and ensuring extreme flexibility, wireless power transfer, ultra-thin components below 100um in thickness and the incorporation of micro-fluidic components are all development extensions needed. The roadmap addresses increasing miniaturization, lower power consumption and energy production/harvesting, increasing accuracy, increasing connectivity with improvement in shape, flexibility, and conformance improvements for wearability. Packages will need to drive toward being noninvasive skin wearable with shifts from electrochemical toward improved optical sensing. Significant development and manufacturing investments will need to be focused towards 3D printing and other additive manufacturing methods.

The major challenges and opportunities in this area are summarized below.

- Hybrid Integration combining various technologies into flexible substrates is of high interest to medical and wearable applications.
- Panel-level packaging is a major focus area for hybrid electronics, leveraging flexible display manufacturing infrastructure (e.g. DPiX).
- Requirements for medical/wearable electronics are significantly different from consumer and computing electronics, and this needs more emphasis in HIR and other advanced packaging roadmaps.
- Biocompatibility for materials, substrates, chip assembly, hermetic and bio-compatible encapsulation, and terminal metals needs to be addressed to enable new applications.
- Flexible fanout wafer level packaging is an important area with several emerging approaches that need to be scaled to volume manufacturing.

- The integration of functional batteries, other energy storage elements , and wireless charging are key enabling technologies.

1.4 TWG2: Cross-cutting Technologies

For TWG2 covering Reliability, Thermal and Modeling, there are gaps within the US electronics ecosystem.

However, there needs to be a major shift in cooling technologies to keep up with the scale-down and scale-out theme of advanced packaging. As we package more higher power density dielets closer to one another, conventional heat spreading is no longer an option. Instead, heat needs to be extracted vertically and liquid cooling, immersion cooling and flash cooling for hot spot elimination and transient heat loads need to be developed. Additionally, thermal dissipation in 3D stacks is the limiting constraint. Heat needs to be extracted vertically in high conductivity strata and transported laterally to vertical heat pipes. See Fig. 1.4.1. Another more fundamental issue that limits heat transport is the interfacial thermal resistance. Materials engineering to improve phonon transport across interfaces will need focus. Additionally, current thermal interface materials are sourced from outside the US and this presents a supply chain concern. Fig. 1.1 (e,f) shows the heat loads of concern.

The US is in reasonably good shape relative to EDA, mechanical and electrical modeling software. However a holistic design methodology that includes electrical, thermal, thermomechanical and optical parameters still eludes us. To achieve faster time to market in designing and fabricating advanced electronics packaging, we need to stress the need to develop strategies to implement co-design methods for packages which includes not only electrical, mechanical, and thermal, but also power delivery, design for manufacturability, design for test and design for reliability. Adopting co-design strategies will reduce the cost of advanced packaging and will ensure packages that can be manufactured at a lower cost.

Advanced packaging presents unique issues with respect to yield and reliability. Unlike conventional packaging, Advanced packaging is not amenable to rework. Advanced packaging assemblies are very high value. While very high yield processes are needed, novel redundancy approaches will be needed so that every assembly is a good assembly. From a reliability perspective a different approach will be needed. These complex systems will need continuous repair via an in situ lifetime built-in self-test and repair system. Another concept that needs to be explored is the idea of graceful rather than catastrophic failure similar to complex biological systems.

1.5 TWG3: Chiplets, chiplet architectures and standards

Chiplets present a game changing paradigm that can enable a revolutionary method of building complex systems. While a lot of lip service has been paid to chiplets, a chiplet or dielet marketplace does not yet exist as yet. To be useful, dielets/chiplets need to be small (a few mm on a side) and highly reusable in a variety of applications. Small dielets make more sense only when we have an extremely fine pitch dielet to substrate connections. We expect that as the bump and trace pitch on

substrates approach sub-10μm dimensions, the chiplet/dielet infrastructure will develop. To facilitate this, a methodology for chiplet discovery must be developed, that is based on a statistical analysis of existing SoCs and ASICS. What kind of IPs should be combined to build chiplets that can be easily handled, reused, and connected to other complementary chiplets. One needs to worry about chiplet/dielet warehousing. Chiplet mechanical and electrical standards are also essential.

Recently, die-to-die interfaces for chiplets have received attention in standardization efforts from multiple organizations. Chiplet-based products require a new integration of the supply chain, not just a new interconnect. Unlike monolithic devices, chiplets must be integrated with other chiplets to form a usable product. Therefore, chiplet-based designs must be cognizant of several factors that are usually considered "back end" issues in monolithic ASIC design such as packaging, inventory, and test. These factors have limited chiplet-based designs to large companies that largely control their supply chain.

In this report, we identify several gaps in standards needed to address these "backend" issues in product development that hinder the integration of chiplets from multiple vendors. We propose the development of domain-specific modular architectures to close these gaps. A modular architecture can develop guardrails for die size, die-to-die bandwidth, thermals, mechanicals, packaging technology, heat dissipation and other attributes relevant to final product design and manufacture.

We develop an example reference modular architecture for high-performance computing (HPC). We derive the reference architecture from the ASIC used to develop the recent Fugaku supercomputer. We show that this modular architecture with bounds on die size, bandwidth, mechanicals, and thermals can meet current HPC requirements for performance, heterogeneous integration, and scale into the future. We also show that scaling can be accomplished in one of two ways - to preserve capital investments in packaging manufacture or to leverage packaging technology. Future development for the modular HPC proposal will require the development of a complete set of standards for packaging, mechanical, thermal, power delivery and other attributes.

While the goal of establishing a signaling standard is ideal, we expect a few standards to co-exist because of application specifics. Within a scaled down assembled system, fine pitch interconnects make inter dielet communication simple using energy efficient protocols such as SuperCHIPS. However, to connect to dielets not using this protocol, translator dielets may be needed to ensure communication between dielets with incompatible protocols.

1.6 TWG4: Supply Chain, Security, Test and Smart Manufacturing

TWG4 group focused on 4 different topics for Heterogeneous Integration – Security, Test, Supply Chain and Smart Manufacturing.

In Chapter 7, we discuss the cybersecurity landscape in heterogeneous integration and electronics packaging (MRHIEP) which is impacted by the rise of hardware-based vulnerabilities which have been created by malicious actors across the supply chain and the advent of fresh integration and packaging technologies, such as chiplets, which have opened the door to an unprecedented chance to reconsider security in hardware design and production. The next generation of ONSHORE manufacturing methods must account for these factors. Designers and manufacturers must recognize that security is a critical concern, which can lead to significant business consequences.

Therefore, they must make appropriate tradeoffs to ensure security is on par with other critical metrics like performance, power, and cost.

In chapter 8, we focus on testing of HI systems. Semiconductor Test was for multiple decades dominated by structured test methods such as full scan and built-in self-test (BIST). As chip manufacturing transitions from monolithic ICs towards heterogeneous integration (HI), and complexity increases dramatically at the same time as access to circuit internals decreases. In this chapter, we elaborate on various test methodologies for HI 7 chiplet systems under the following domains - RF test, Photonics Test, Logic, Specialty Test, Memory, Analog/Mixed Signal, System level, Data Analytics, 2.5/3D test and test cost.

There are many challenges that the test industry must address in order to keep up with this rapidly evolving industry and solving these problems requires specialized skills which are increasingly scarce in the US for a variety of underlying reasons including fewer university programs, test equipment cost, lagging funding for graduate level test research, etc. The chapter proposes key approaches to address the challenges through a concerted effort on the education front.

Chapter 9 addresses supply chain resiliency and concerns for onshoring – The recent pandemic brought into sharp focus the need for more resilient supply chains in the semiconductor industry, which has perhaps one of the most complex and globalized supply chain networks of any industry. Fortunately for the semiconductor supply chain, the USA has significant if not dominant positions across most of the value layers including EDA & Design, front-end device manufacturing, manufacturing equipment, and materials & chemicals. However, one link of the value layer – chip packaging (assembly and test) - has traditionally been outsourced to low-cost regions and as a result the supply chain related to this value step has faced pressure to localize outside of the USA.

The chapter discussion focuses on inflections in packaging sub-assembly technology that could offer a serendipitous opportunity to secure the packaging value layer related supply chain for the USA, especially for high performance computing (HPC), AI and other technology intensive medical devices. Not exploiting these inflection opportunities to onshore and secure packaging supply chains for the USA, would not only endanger its leading position in technology and defense capability, it may also lead to a permanent off-shoring of R&D for emerging technologies such as advanced packaging.

Lastly Chapter 10 focuses on the deployment of Industry 4.0 or Smart Manufacturing tools, technologies, and methods for HI & Chiplet systems and provides roadmap guidance of Smart Manufacturing methods in development and in current production, where the use of digital twins, AI/ML techniques will facilitate through closed-loop smart control of manufacturing processes to improve quality, yield, and reliability at a reduced overall manufacturing cost for HI systems. These technologies are essential to create a technology led-path to re-shoring package manufacturing into the US by reducing the dependence on low cost labor. Adoption of Smart Manufacturing techniques and methodologies will reduce the cost of assemblies by reducing the manpower required to run the assembly processes to produce assemblies but will also increase the quality of the components that are made.

1.7 Example of Manufacturing Gaps and Challenges: A Global Supply Chain Perspective on Advanced Substrates

The major challenges and potential solution pathways to onshoring manufacturing of advanced panel-based substrates (organic, glass, silicon) are summarized in this section, compiled from discussions with key global suppliers of materials, process chemistries, and manufacturing process equipment for package substrates. The same challenges and opportunities cut across the other advanced packaging platforms in this report.

Substrate Manufacturing Onshoring:

- Factory investments for onshoring: Investing in existing production onshoring will not command premium pricing required to meet the return-on-investment targets for high volume manufacturing infrastructure. Additional investment challenges come from the need to re-capitalize the factories with upgrades that could represent up to 10% of the initial capital on an annual basis to remain on the leading edge. Advanced substrates that enable multiple levels of fine pitch connections, active and passive components will provide significant value add to substrates making the return on investment favorable.
- Automation and Smart Manufacturing – Extremely high levels of automation will be needed in any onshore manufacturing locations to be competitive with the lower cost structures present in leading edge Asian manufacturing locations. The overseas cost is lower due to several factors, including sustained government incentives over decades, built-up manufacturing yield know-how, innovations in processes, tools and materials.
- US infrastructure is PCB based, transitioning to advanced substrates is quite challenging – starting with a blueprint for package substrates would be a better path than converting existing PCB factories to manufacture advanced substrates.
- An additional avenue to expedite onshore package substrate capacity is to incentivize leading global substrate manufacturers to initiate or expand their onshore footprint. Even more important could be to provide support to leading edge package substrate manufacturers that already have other types of manufacturing footprints in the US.
- Skills gap is quite significant in the US, training programs need to focus on process development and integration know-how, and end to end materials and process tool knowledge development.

Substrate Materials/Chemistry/Equipment Supply Chain Onshoring:

- **What would motivate a leader in the global materials supply chain to invest in onshore manufacturing?** The lack of onshore high-volume demand from immediate customers (i.e. Substrate manufacturers) is a major barrier for such investments. Avenues to incentivize onshoring of global material and chemistry suppliers include, (i) investing in cost-competitive, but leading-edge raw material supply chain to enable the final material and chemistry suppliers, (ii) expanding scientific centers of excellence in US universities and research institutes to support future roadmaps, (iii) integral involvement of end users who drive future material and chemistry specifications , and (iv) value-add advanced substrates outline in section 1.2.

Innovation and Manufacturing Hubs: Innovation hubs serve as a center for global supply chain companies to collaborate with their customers and develop their future products. Such innovations hubs are usually followed by investments in manufacturing at those same locations.
There are a number of challenges in setting up innovation and manufacturing hubs for advanced packaging in the US.

- Innovation and tech centers for global leaders are currently located in their overseas HQ and in markets such as Asia where the high-volume customers are located.
- There are specific additional challenges for high volume equipment manufacturing – a key challenge is that sufficient capacity already exists in their multiple sites, and significant overall market growth for high capex equipment is limited and constrains the creation of new development and manufacturing centers.

Chapter 2:
Advanced Packaging & Heterogeneous Integration

Contents

2.1 CHARTER 3

2.2 APPROACH & FOCUS AREAS 3

2.3 TEAM 3

2.4 HIR CHAPTER REVIEWS 4

2.4.1. Multi-Chip Packages (Chapter 8) 4

2.4.2 Photonics (Chapter 9) 5

2.4.3 Interconnects for 2D and 3D (Chapter 22) 5

2.4.4 RF/mm-wave, Power, Analog, MEMS 6

2.5 MANUFACTURING BLUEPRINT 6

2.5.1 High Performance Computing Manufacturing Roadmap Targets 7

2.5.2 Process Flow, Material and Tool Sets 7

2.5.2.1 Advanced Interposers and Substrates 8

2.5.2.2 Bond Pitch Scaling and Assembly 17

2.5.2.3 Fanout Wafer and Panel Level Packaging 24

2.5.2.4 Silicon Photonics Packaging 29

2.6 MANUFACTURING GAP ANALYSIS (ROADMAP & ONSHORE NEEDS) 32

List of Figures

Figure 2. 1 Major sections of the HPC Manufacturing Blueprint..... 7
Figure 2. 2 Typical TSV and BEOL Silicon Interposer Process Flow Sequence..... 9
Figure 2. 3 Typical Organic FCBGA Substrate Fabrication Process Flow [1] 12
Figure 2. 4 An Example of Through Glass Via Structuring and Metallization Process Flow [2] 12
Figure 2. 5 Solder-based Historic Interconnect Roadmap & Fine-Pitch Cu-SnAg Microbumps. 17
Figure 2. 6 Two step high throughput thermal compression bonding process [8] 18
Figure 2. 7 Bonding cross-section in a sample Cu-Cu thermal compression bonding process [8] 18
Figure 2. 8 Hybrid Bonding Approaches and Use Cases [11]..... 20
Figure 2. 9 Wafer-to-Wafer Hybrid Bonding Process Flow..... 21
Figure 2. 10 Die-to-Wafer Hybrid Bonding Process Flow 21
Figure 2. 11 Three Major Fanout WLP/PLP Technology Categories based on Process Flows. .. 25
Figure 2. 12 Generic Process Flows for Chip-First Fanout Package Fabrication (a) Face Down FO-WLP, (b) Face Up FO-WLP..... 25
Figure 2. 13 Generic Process Flow for Chip-Last or RDL-First Fanout Package Fabrication..... 25

Figure 2. 14 Examples of fiber arrays A: Schematic, B: Photo. Free-space micro-optical couplers that are printed on a fiber array (PHIX), C: SEM image & photo (Nanoscribe, PHIX), D: Photonic-Plug® (Teramount), E: Microcantilever-based fiber coupling, (MicroAlign); [13] 30
Figure 2. 15 Ayar Labs showcased a 4 Tbps optically-enabled Intel FPGA design at SC23, which offers 5x current industry bandwidth at 5x lower power and 20x lower latency, all packaged in a common PCIe form factor. (credit: Ayar Labs) 31
Figure 2. 16 Co-packaged photonic assembly with six detachable optical interfaces [14] 31
Figure 2. 17 III-V laser integration on the monolithic SiPh platform. (a)-(d) 3D perspective views of various PICs with different SSCs formed on Si or SiN layer. (e)-(f) Optical image and SEM of laser cavities with and without flip-chip-bonded laser. (g)-(j) Light- current curve, RIN, spectral characterization and SMSR performance. (k)-(l) Wafer-level accelerated reliability results [15]31
Figure 2. 18 Summary of Gaps and Challenges that create opportunities for New Innovation and Investments in Future HPC Packages. 32

List of Tables

Table 2.1 Summary of Roadmap Targets for HPC Extracted from HIR Roadmap (2019, 2021) 7

Table 2.2 Advanced Interposers and Substrates Roadmap Highlighting Key Manufacturing Challenges (Source: 2023 HIR Roadmap Update) 8

Table 2.3 Silicon Interposer Materials and Process Tool Lists with Key Identified Gaps 10

Table 2.4 Manufacturing flow along with unit process tools and associated materials for organic substrate, with key identified gaps. 15

Table 2.5 Manufacturing flow along with unit process tools and associated materials for Silicon-core substrate. 16

Table 2.6 Manufacturing flow along with unit process tools and associated materials for direct metal-metal thermal compression bonding, with key identified gaps. 19

Table 2.7 Wafer-to-Wafer Hybrid Bonding Process Flow, Materials, Equipment, and Gaps 22

Table 2.8 Die-to-Wafer Hybrid Bonding Process Flow, Materials, Equipment, and Gaps 23

Table 2.9 Chip-First, Face-Down, FOWLP/PLP Process Flow, Materials, Equipment, & Gaps 26

Table 2.10 Chip-First, Face-Up, FOWLP/PLP Process Flow, Materials, Equipment, & Gaps 27

Table 2.11 Chip-Last FOWLP/PLP Process Flow, Materials, Equipment, and Gaps 28

Table 2.12 Summary of Global and Onshore Capabilities in HPC Package Platforms Highlighting On-Shore Gaps in Most Platforms 33

2.1 CHARTER

The charter of this technical working group was to create a manufacturing roadmap and generic blueprint for manufacturing execution in key identified areas within advanced packaging and heterogenous integration for high performance computing (HPC) applications, building from the HIR Roadmap and other relevant industry roadmaps.

2.2 APPROACH & FOCUS AREAS

The approach starts from a detailed review of the HIR Roadmap (2019, 2021 editions with selected information from the 2023 update) and builds a manufacturing blueprint in three key technology platforms for high performance computing applications, as listed below:

1. Advanced Substrates for Chiplet and Multi-Chip Integration
2. Bond Pitch Scaling and Assembly Processes
3. Fan-out Wafer-Level and Panel-Level Packages

The roadmap targets, gaps/challenges and potential solutions are built for each platform, leveraging the HIR roadmap, and the collective experiences of the team of industry experts, to create a manufacturing blueprint. Once the key platforms are selected, a comprehensive benchmark is undertaken to show the state-of-the-art technologies in manufacturing in the US and Globally against the 3-, 5-, and 7-year HIR roadmap targets. Onshoring gaps are then identified for the selected process flows to ensure complete end to end supply chain coverage. All these activities will culminate in the creation of a manufacturing implementation strategy and generic blueprint for advanced packaging and heterogeneous integration, with a focus on ONSHORE end-to-end supply chain.

2.3 TEAM

The large and diverse scope of this working group was supported by participants from several leading semiconductor, materials, process tools, and packaging supply chain companies.

Venky Sundaram (3D System Scaling LLC)
Tom Rucker (Intel)
Joy Watanabe (EMD Electronics)
Ram Kambhampati (Resonac US)
Steven Verhaverbeke (Applied Materials)
Hanwen Chen (Applied Materials)
Kanda Tapily (Tokyo Electron US)
Reza Mahmoodian (Ulvac)
Habib Hichri (Ajinomoto Fine Techno US)
Lou Dadok (Fujifilm US)
Kruthikesh Sahoo (UCLA CHIPS)
Vineeth Harish (UCLA CHIPS)
Markus Leitgeb (AT&S)
Jobert van Eisden (Atotech/MKS)

2.4 HIR CHAPTER REVIEWS

The first task undertaken was to review the key HIR roadmap chapters and provide brief chapter summaries that include key points highlighted, potential solutions, gaps and future challenges. The team reviewed several chapters in the IEEE HIR roadmap to initiate the advanced packaging manufacturing blueprint development process. Although the initial scope included RF/mm-wave content, due to the significant overlap with the iNEMI MAESTRO project on 5G/6G/mm-wave materials and testing, the group leadership connected with iNEMI leadership and agreed to cross-reference the work scopes for mutual benefit and avoid duplication. This chapter focuses on the HIR roadmap chapters relevant to High Performance Computing.

2.4.1. Multi-Chip Packages (Chapter 8)

The key points from the chapter review are summarized below:

- Advanced Substrates is a major gap in the Onshore Supply Chain
  - HIR calls for 1/1 um lines/spaces for Chiplet integration by 2025, 0.5/0.5 um by 2030
  - Both wafer and panel solutions will be needed considering application diversity and large range of package body sizes
  - Majority of recent investments in fine pitch RDL manufacturing have been focused in Asia – exceptions such as EMIB investments by Intel in the US.
  - Materials and several tools need to be upgraded to enable at-scale alternatives to silicon interposers
- Power Integration at package level (substrates/interposers or fanout) is a critical requirement to continue bandwidth scaling (recent trends indicate 2x increase every 3 years, timeline accelerating)
  - Bulk of the manufacturing investments continue to be in traditional discrete components

Further details of the chapter review including gaps and future challenges highlighted are listed below in the context of multi-chip packages.

- System level performance metrics roadmap is not broken out into single and multi-chip, needs extraction and consultation with chapter authors.
  - 4-6 Gbps per lane data rates required for HBM3-logic and logic die-to-die interconnects
  - Number of HBMs will increase 1.4x for each silicon node transition
  - HBM3 will require 2048 I/Os per link
- Substrate solutions for 0-5 years ahead have been called out
  - Improving organic substrates/panel substrates to 2/2um and 1/1um in the longer term (2-5um range has been identified as optimal based on line resistance)
  - Extending existing EMIB and silicon interposer solutions
  - High density ceramic carriers called out as an emerging option
- Power delivery identified as major challenge
  - Both inductor and switched capacitor based in-package voltage regulators called out
  - 200-400W TDP will require package integration of power delivery components

- Thermal management issues escalating with chiplet and 2.5D/3D integration driven increase in power density at package level

2.4.2 Photonics (Chapter 9)

This chapter talks about how integrated photonics will be a key enabler of delivering increased bandwidth density, low latency, low power and low cost to meet the demand associated with the data deluge. It also covers challenges that need to be addressed. Other salient points are listed below.

- Requires co-packaging of electronics, photonics and plasmonic.
- Same challenges as IC packages exist with integrated photonics with the added complexity to integrate both passive and active photonics elements.
  - May lean on other chapters.
- Many photonics elements have unique thermal, electrical, mechanical characteristics that will require specialized materials and system integration, processes, and equipment such as microfluidics and temperature control.
- Examples of growing technology is Lidar (fueled by automotive market)
- Integration Platform for photonics use electronics technology whenever possible.
  - Chip level integration: photonics + electronics into single product w/ sequential chip connection
    - This process is slow and costly.
  - Wafer level integration: fabrication and assembly for photonics at wafer level and cost is reduced.
  - System level integration offers lowest latency, cost and power.

2.4.3 Interconnects for 2D and 3D (Chapter 22)

This chapter presented a comprehensive guide to 2D and 3D architecture related nomenclature, and also identified a number of challenges for future interconnects.

- **Converged Nomenclature Framework for 2D & 3D Architectures**
  - **2D architecture** An architecture where two or more active silicon devices are placed side by side on a package and are interconnected on the package. A 2D architecture with "enhanced" interconnect, i.e., higher interconnect density than mainstream organic packages, is further sub-categorized as below.
    - **2DO (2D Organic) architecture** A 2D architecture with "enhanced" interconnect accomplished using an organic medium.
    - **2DS architecture** A 2D architecture with "enhanced" interconnect accomplished using an inorganic medium (e.g. a Silicon/glass/ceramic interposer or bridge).
  - **3D architecture** An architecture where two or more active Silicon devices are stacked and interconnected **without** the agency of the package.
- **Interconnect Nomenclature**
  - **Die-Die Interconnects** Interconnects between stacked dies for vertical connectivity between multiple dies in a 3D stack.
  - **On-package Die-to-Die Interconnects** 2D and Enhanced-2D interconnects.

- **Die-to-Package Interconnects** Interconnects between the die and the package, typically known as the first level interconnect (FLI).
- **Within-package Interconnects** Interconnects within the package that enable lateral connections between two or more dies.
- **Package-to-Board Interconnects** Interconnects between the package and the next level, which is typically the motherboard, are referred to as the second level interconnect (SLI).
- **POP (Package-on-Package) Interconnects** The PoP construction allows packages to be placed on top of other packages using peripheral package interconnects, also referred to as VI (**V**ertical **I**nterconnects).

The following Challenges and Requirements for the 2D/3D Interconnect Roadmap were extracted from Chapter 22.

- When line pitch scaling is combined with increasing signal speeds, signal integrity is a concern due to increased crosstalk caused by the reduced line spacing. Solutions that minimize impact to signal integrity and provide physical links with improved power efficiency are required.
- Key challenges for stacked-die architectures will continue to be in fine pitch sort/test, thermal management, power delivery network development, design process co-design, in-line process control and equipment readiness for high volume.
- Greater need to enable novel assembly technologies for ultra-fine pitch enhanced-2D and 3D architectures using both solder and non-solder-based approaches.
- Ability to integrate the right thermal features will define the physical envelope (i.e. form factor and number of die/die stacks that can be integrated on the package) and the warpage characteristics that will ensure manufacturability.

2.4.4 RF/mm-wave, Power, Analog, MEMS

A summary is included here for the sake of completeness, however, as stated in the introduction, the activities in this sub-group were limited to leverage the work done in the iNEMI 5G/6G MAESTRO project.

- Recommend Leveraging IEEE International Network Generations Roadmap (INGR), which provides system level guidance and design requirements.
- Multiple HIR Roadmap Chapters relevant to this sub-group
- Key Areas of Focus in RF/mm-wave
  - *Advanced low loss dielectrics in 5G mm-wave and 6G bands needs attention.*
    - Several emerging materials in the supply chain but needs high frequency characterization data as well as design library development.
  - RF/mm-wave substrates (especially onshore manufacturing) need significant development.

2.5 MANUFACTURING BLUEPRINT

This section describes the key sections of the manufacturing blueprint (5 & 10 year targets; process flows, tool lists, material lists, onshoring gaps and options, future challenges and potential solutions).

2.5.1 High Performance Computing Manufacturing Roadmap Targets

The HIR roadmap Chapter 8 presents a complex set of system parametric targets for high performance computing applications and associated advanced packaging technology targets. Multiple tables and figures and sections from Chapter 8 were used to derive a much more simplified set of targets for the manufacturing roadmap, shown in **Table 2.1**. The targets for 2029 are estimated since targets for that year are not yet available in the HIR roadmap and likely to be published in the 2023 update.

2.5.2 Process Flow, Material and Tool Sets

Typical industry process flows for various platforms were compiled within the defined scope, with detailed materials and tool sets for each flow, ending with major manufacturing challenges and gaps identified for each platform. **Figure 2.1** illustrates the organization of the selected platforms in the HPC manufacturing blueprint. Please note that the process flows and material/tool lists have been listed as "For Use in US and Canada Only".

Table 2.1 Summary of Roadmap Targets for HPC Extracted from HIR Roadmap (2019, 2021)

Parameter	Unit	2025	2027	2029
Silicon Node	nm	3nm	2nm	1nm
I/O Bandwidth (Logic-HBM)	Gbps	1024 x 2.4	2048 x 3.6	4096 x 6.4
I/O per mm per layer (shoreline)	#	250	500	1000
I/O lines and spaces (and vias)	micron	2/2/2	1/1/1	0.5/0.5/0.5
Package to Board I/O BW	Gbps	64 per I/O	112 per I/O	256 per I/O
Package to Board Pin Count	#	9600	11200	12800
Power Density	W/mm^2	1	1.05	1.1
Package Dimension (Minimum)	mm	95	103	120

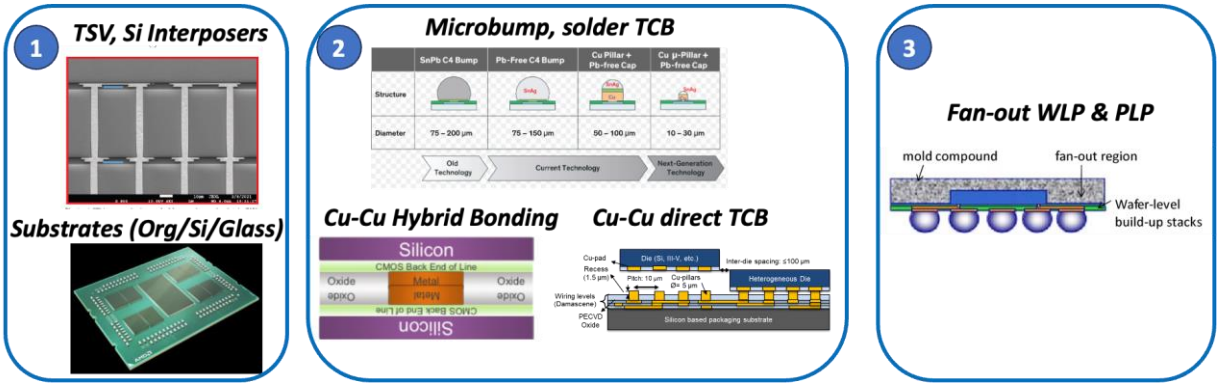


Figure 2. 1 Major sections of the HPC Manufacturing Blueprint

2.5.2.1 Advanced Interposers and Substrates

This section of the blueprint covers silicon interposers with through silicon vias (TSV) and back-end of line (BEOL) RDL, and advanced substrates (Si, organic or glass core with through vias and polymer-Cu RDL or Cu-inorganic dielectric RDL). A major focus area to enable future HPC roadmaps is on chiplet integration and die-to-die interconnect on advanced interposers and substrates. The 2023 update to the HIR roadmap outlines die-to-die interconnect parameters for various platforms including silicon interposers, organic FCBGA substrates and RDL/organic interposers. **Table 2.2** shows a more detailed parametric roadmap with the TWG1 team assessment of key roadmap manufacturing challenges.

Table 2.2 Advanced Interposers and Substrates Roadmap Highlighting Key Manufacturing Challenges (Source: 2023 HIR Roadmap Update)

	Year of Production		2018	2019	2020	2021	2022	2025	2028	2031	2034
Technology	Parameter	Units									
FCBGA substrate	Maximum layer Count	N/A	20	20	20	20	22	24	24	26	26
	Maximum body size	mm2	3000	3500	4000	4500	5000	6500	8000	9000	>10000
	Minimum Bump Pitch	µm	110	110	100	100	100	90	90	80	80
	Roughness	nm	500	300	300	300	150	150	100	<100	<100
	Dielectric loss tangent (Df)		0.007	0.007	0.007	0.007	0.004	0.004	0.002	0.002	0.002
Chiplet (Fan-out, Organic interposer)	Min. Bump Pitch (um)	µm	50	50	50	45	45	40	40	30	30
	Min Line width (um)	µm	2.0	2.0	2.0	1.5	1.5	1.0	1.0	0.5	0.5/0.5
	Min. uVia diameter (um)	µm	30	30	30	20	20	10	10	5	5
Chiplet (Si Interposer) solder based	Min. Bump Pitch (um)	µm	40.0	40.0	40.0	35.0	30.0	22.0	16.0	13.0	10.0
	Min Line width (um)	µm	0.6	0.6	0.6	0.6	0.6	0.5	0.4	0.3	0.2
	Min. uVia diameter (um)	µm	0.6	0.6	0.6	0.6	0.6	0.5	0.4	0.3	0.2
Chiplet (Si Interposer) hybrid bonding	Min. Bump Pitch (um)	µm				9.0	9.0	6.0	6.0	3.0	3.0

Organic Core Running out of Steam for Warpage Control and >100mm package sizes

Polymer RDL scaling to sub-micron needs significant development

Reticle stitching driving low throughput and high cost as interposer sizes escalate to support chiplet integration

Major challenges that need to be addressed include (a) lithographic scaling to sub-micron copper wiring, especially for large interposer/substrate sizes greater than 60mm x 60mm, (b) polymer RDL scaling to reduce RC delay and enable longer wire lengths consistent with UCIe, BoW and other industry standards, and (c) new inorganic core materials such as silicon and glass, as well as improved organic laminates to address the warpage and reliability concerns of current organic core materials for large body size packages.

a) Silicon Interposers (BEOL, TSV)

Silicon interposers were introduced in 2011 with the Xilinx FPGA products based on die splitting of one large die into multiple tiles and re-connecting them using BEOL wiring on a thin silicon interposer with TSVs. This technology has subsequently been adopted by AMD for GPU to HBM high bandwidth connectivity, and by many other companies in chiplet-based and non-chiplet based products, all involving heterogenous integration of logic and memory. This is a mature technology practiced in high volume manufacturing by TSMC (CoWoS-S), Intel (Foveros active interposer) and other foundries. A typical process flow for a silicon interposer is shown in **Figure 2.2** (Source: X. Zhang, IEEE ECTC 2009).

Major gaps identified for scaling silicon interposers for the future roadmap targets include bond/debond yield as wafers become ultra-thin (e.g. less than 50 microns), and metrology tool throughput as wiring density and TSV density escalates to support bandwidth scaling.

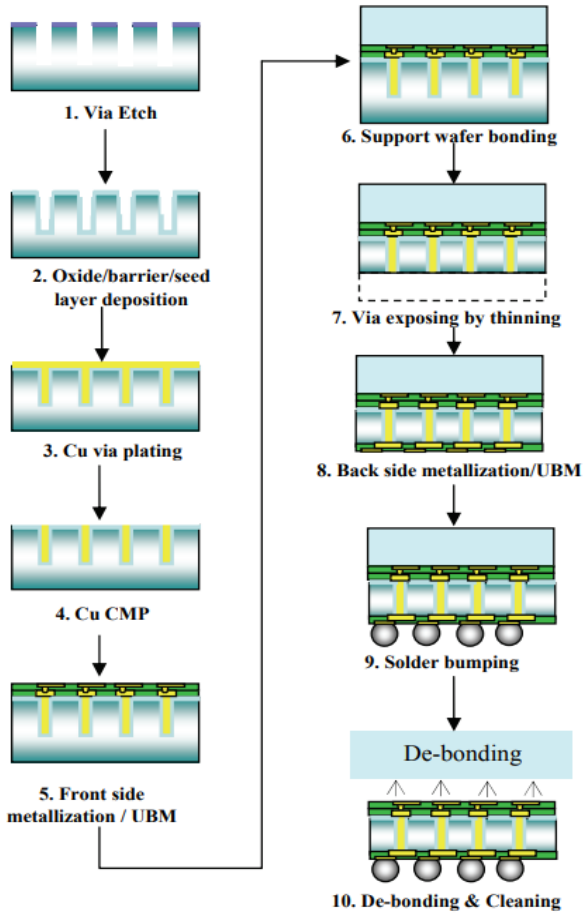


Figure 2. 2 Typical TSV and BEOL Silicon Interposer Process Flow Sequence

A set of materials with major suppliers (list not exhaustive), and process tools used and major suppliers (list not exhaustive) is shown in **Table 2.3**.

Table 2.3 Silicon Interposer Materials and Process Tool Lists with Key Identified Gaps

				Equipment			Material			
	Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks
TSV	1	Oxidation	mask	furnace, PECVD	TEL, HKE, AMAT, LAM	N	SiH4, TEOS, H2, O2	JSR, TOK, DuPont	N	
	2	Litho/patterning	Mask patterning	track, developer, exposure	TEL, ASML, Canon, Carl Zeiss	N	resist	JSR, TOK, DuPont	N	
	3	Via etch	Via etch	BOSCH Etch (Si)	Lam, AMAT	N	SF6, O2, C4F8, Ar	Air Products, Air Liquide, SK Chem,	N	
	4	Oxidation/Oxide deposition	Oxide deposition	PECVD/Furnace	TEL, HKE	N	SiH4, TEOS, H2, O2	Air Products, Air Liquide, SK Chem,	N	
	5	barrier seed layer deposition	Cu barrier and seed	PVD, ALD	AMAT, Ulvac, Shibaura, ASMPT, Evatec	N	Ti, Ta, TaN, Cu	JX, Honeywell, Linde, ToSoh, KFMI	N	
	6	Cu plating	Cu plug	Electroplating	AMAT, ASMPT	N	CuSO4, H2SO4, Cl-, additives	Entegris, BASF, MLI, DuPont	N	
	7	Anneal	Cu anneal	furnace/oven	TEL, HKE	N			N	
	8	Inspection for void	metrology	metrology	ASM Pacific	N			N	
	9	CMP and CMP clean	planarization	CMP	AMAT, Ebara	N	CMP Pads, Slurries, Cleans	Entegris/CMC, Fujimi, DuPont, etc.	N	
	10	Metrology for planarization	metrology	interferometry		N			N	
RDL	11	RDL build, CVD dep & pattern	Interconnect dielectric	PECVD	AMAT, LAM, TEL	N	TEOS, SiH4, O2	EMD Electronics, Air Liquide, SK Chem,	N	
	12	RDL build (optional CVD nitride)	Interconnect dielectric	PECVD	AMAT, LAM, TEL	N	SiH4, NH3	Air Products, Air Liquide, SK Chem,	N	
	13	RDL Litho / Etch	Patterning	Track, Develop, Etch	AMAT, LAM, TEL, ASML, Canon, Zeiss	N	Resist / Etch Gases (CF4, etc.	JSR, TOK, DuPont	N	
	14	RDL build, PVD barrier & Cu seed	Cu barrier & seed	PVD -> ALD	AMAT, Ulvac, Shibaura, ASMPT, Evatec	N	Ti, Ta, TiN, TaN Cu targets	JX, Honeywell, Linde, ToSoh, KFMI	N	
	15	RDL build, Electroplated Cu	Interconnect metal	Electroplating (EP)	AMAT, ASMPT	N	Cu plating	Entegris, BASF, MLI, DuPont	N	
	16	CMP & post CMP clean	Planarization	CMP	AMAT, Ebara	N	CMP Pads, Slurries, Cleans	Entegris/CMC, Fujimi, DuPont, etc.	N	
	17	AFM	Cu dishing, Erosion Meas.	Metrology - AFM	Bruker, Park systems	N				
Bonding	18	Carrier wafer preparation	cleaning	Wets		N			N	
	19	Bond layer deposition and activation	bond layer	ALD, CVD, spin-on	TEL, LAM, AMAT,…	N	SiO2, SiOCN, SiCN…		N	
	20	Bonding	Carrier wafer bonding	Bonder	EVG, TEL	Y	water, activation (plasma)		Y	Ultra-Thin Wafer Handling
	21	Annealing	annealing	furnace	TEL, small vendors	N			N	
Grind/thinning	22	Wafer flip for backside removal				N			N	
	23	Wafer thinning	Via exposing	Grinder, CMP, etch	Disco, AMAT	N			N	
	24	Backside metallization/RDL	backside interconnects	see RDL section	see RDL section	N			N	
UBM	25	Seed layer	Seed layer/UBM	PVD, ALD	AMAT, Ulvac, Shibaura, ASMPT, Evatec	N	Ti, Ta, TaN, Cu , Al		N	
Solder Bump	26	Litho	patterning	track, developer, exposure	TEL, ASML, Canon, Carl Zeiss	N			N	
	27	Plating	Metal fill	Cu/solder or Sn/Ag..	AMAT, ASMPT, Pactech	N	Cu, Sn, Ag, Ni		N	
	28	Resist strip	resist removal			N	resist remover	Dupont	N	
	29	Seed layer patterning	seed layer etch	etch		N			N	
	30	Bump reflow	reflow	furnace	Air products/Sikama international	N			N	
De-bonding	31	De-bonding	Carrier wafer removal		TEL, EVG	Y	Temporary Bonding Film	Resonac, Brewer Science, DuPont, 3M	Y	Ultra-Thin Wafer Yield
	32	Cleaning	cleans			N			N	
	33	Metrology	Inspection for Defects	Optical, Laser, IR	KLA, Onto Innovation, AMAT	Y				Higher Throughput Needed

b) Organic, Silicon and Glass Substrates (Polymer Build-up and RDL)

The package substrate traditionally served the functions of connecting the ICs to the PCB motherboard, providing a stable base to assemble one or more active and passive components, protection, dissipating heat through thermal vias and copper planes, and routing power from the motherboard to the ICs. Substrates play a critical role in product reliability and electrical testing. With the introduction of 2.5D architectures and chiplets, package substrates in some cases have been used for die-to-die interconnections and embedding of components into the substrate core or build-up layers. In recent years, heterogeneous integration of 2.5D/3D architectures with chiplets and/or multiple electronic components into systems in package (SiP) has become the driver for pitch scaling and integration at the package substrate level.

Organic core substrates with polymer-Cu build-up layers were first introduced in the early 1990s and kick started the flip-chip BGA (FCBGA) package revolution that continues to be the backbone for high performance computing chipset packaging. FCBGA package sizes remained stable at around 55mm x 55mm for more than 20 years. To improve warpage and electrical performance, the organic core materials, typically constructed using glass fabric reinforced epoxy or other resins, have advanced significantly in electrical and mechanical properties. However, the advent of chiplets in recent years has resulted in a sudden escalation in FCBGA package body sizes up to 80-100 mm on a side. It is predicted that body sizes as large as 140mm x 140mm could be needed to support HPC heterogeneous integration in the next 5-10 years. As a result of this body size increase, the organic core material thickness has increased from 0.6mm to 1.2mm, an upward trend never seen before in the history of FCBGA packaging. This has led some end users to explore inorganic core materials such as silicon and glass, with significant R&D investments having gone into these advanced substrates. The silicon core substrate represents initial R&D conducted at Georgia Tech, with inputs from other universities conducting silicon core substrate R&D (UCLA) and industry members (Applied Materials and others) involved in exploring the scale up of this approach to manufacturing. Glass core substrates follow a similar process flow as the organic core substrates shown in this section. Manufacturing investments are being considered by several suppliers in Asia and some in the US for brownfield or greenfield substrate factories that can support handling and fabrication of glass and new panel-based substrates.

Panel-based Organic and Glass Substrates

The process flows and manufacturing tools/materials discussed in this section are based on typical organic substrates currently in high volume manufacturing, mostly in Asia. However, similar material and tool sets can be utilized to build glass core package substrates, with significant differences coming from the new processes used to fabricate the glass cores with metallized through vias. A typical process flow for an organic FCBGA substrate is shown in **Figure 2.3**.

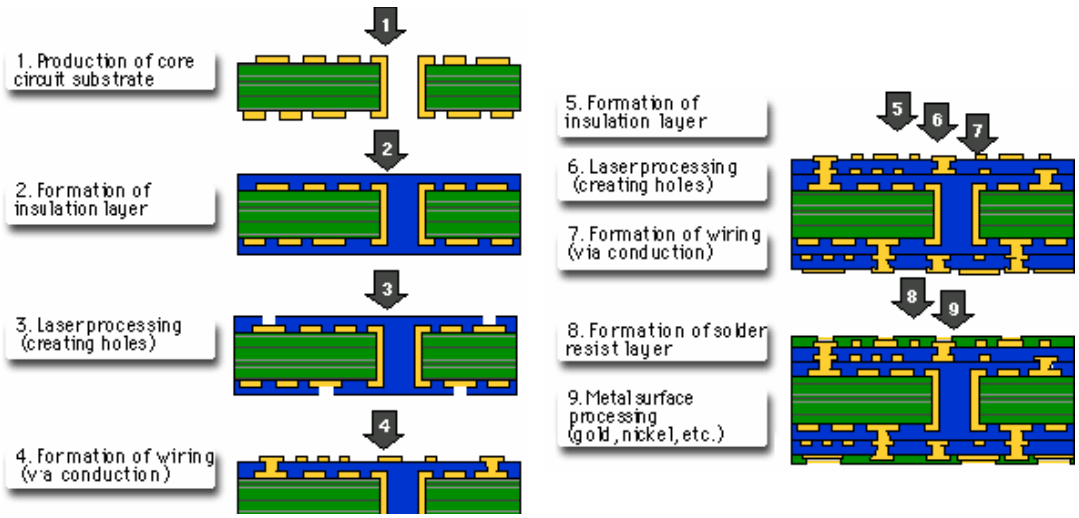


Figure 2. 3 Typical Organic FCBGA Substrate Fabrication Process Flow [1]

The list of materials and tools used to construct FCBGA organic substrates is shown in **Table 2.4**.

Panel-Based Glass Core Substrates

Glass panels for substrates and interposers were explored by Georgia Tech and several other groups starting in 2008. Glass promises to combine the best dimensional stability and ultra-smooth surface properties of silicon with the large panel scalability and low-cost manufacturing of current organic cores. One of the foundations of the glass core substrate technology is the ability to leverage the mature and high-volume LCD panel infrastructure for the glass core material. Several leading glass manufacturers including Corning in the US, AGC in Japan and Schott Glass in Germany have been actively investing in through glass vias and other building blocks required to enable glass core substrates and glass interposers. The first pilot line and low volume manufacturing investments for glass substrate development and production have been made in the past few years, with Intel and others making public announcements on glass substrate capabilities and plans. Several chipmakers have expressed interest in introducing glass substrates into their product roadmaps within the next ten years, starting with high performance computing packages that are pushing the package size and pitch scaling limits of organic substrates. The biggest difference between glass and organic substrates is the glass core structuring and metallization processes. A typical process flow for through glass via (TGV) creation and metallization is shown in **Figure 2.4**.

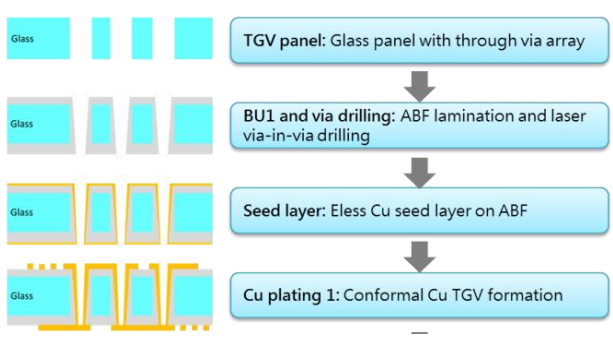


Figure 2. 4 An Example of Through Glass Via Structuring and Metallization Process Flow [2]

Major technical requirements for organic and glass panel-based substrates are associated with key process modules such as precise through-core via drilling, RDL via opening and interlayer alignment, fine line/feature imaging, Cu electroplating, barrier/Cu seed deposition and etch, copper surface treatment for adhesion enhancement, and descum/cleaning for high yields. The next generation of substrate manufacturing will also require investments into new process modules such as planarization for multi-layer fine pitch RDL, higher resolution metrology and inspection for yield management, and ISO5 (Class 100) and even ISO4 (Class 10) cleanrooms as line pitches scale towards 1-2um. The recent trend towards larger body size packages, especially in high performance computing and AI use cases will drive demand for panel-based substrate manufacturing with improved pitch scaling.

Double-Sided Silicon-cored Substrate (an example of wafer-level processed substrate)

This emerging technology has been developed organically within the US (Applied Materials, Georgia Tech, UCLA and others). Both copper-polymer RDL and Cu-SiO2 RDL have been implemented on this platform, enabling a wafer toolset for pitch scaling. The major gap in this platform is the investment in a pilot manufacturing line with provision for expansion to HVM. In addition, certain materials and equipment gaps (**Table 2.5**) will need to be addressed to establish onshore manufacturing capabilities and supply chains that leapfrog other countries.

c) Photonic Integration and Co-Packaged Optics

Package-level integration of photonic ICs with electronic ICs is now mandatory for many high-speed networking, data centers and servers, and other high performance computing and communication systems. Co-packaged optics must interface seamlessly with single- and multi-mode optical fiber with less than 2dB and in leading-edge packages, less than 1dB of channel loss from fiber to photonic IC. Two primary platforms have emerged in recent years for electronic-photonic integration, (a) wafer BEOL silicon interposers with TSVs, which integrate thin-film silicon nitride optical waveguides, and (b) panel substrates (organic or glass), which integrate polymer or glass waveguides. Beam steering structures such as diffraction gratings, microlenses or mirrors, and optical coupling structures such as V- or U-grooves for precision fiber assembly need to be integrated into the substrate or interposer fabrication process flows. Forward looking challenges include precision alignment and dimensionally stable substrates to enable passive alignment, fiber array integration into substrates and interposers, temperature/humidity/light aging stability of embedded waveguides, thermo-mechanical stress management during process integration and operation, and high throughput assembly at sub-micron precision for photonic chip-to-substrate interconnections. Co-packaged optics and electronic-photonic packages require ultra-high speed signal channels in the substrate or interposer, which in turn necessitates low loss dielectrics and precise copper trace formation processes. The power delivery and thermal management challenges highlighted elsewhere in this roadmap, are amplified for photonic integrated packages due to the increased power diversity, power density and heat dissipation brought on by silicon photonic ICs. Integration of high-power lasers and other light sources represent the outer portions of the roadmap and bring in enormous complexity in signal, power and thermal management. The evolution from single fibers to 2D fiber arrays will continue into 3D arrays of fibers, necessitating vertical fiber integration in addition to the current horizontal fiber coupling modules and structures. Co-packaged optics and photonic package integration represent a critical area for global leadership and onshoring investments in R&D and manufacturing.

Advanced Substrates Manufacturing Roadmap Gaps and Challenges: The most critical gaps identified include (a) Warpage and thickness limits of organic core materials limited by low modulus and CTE mismatch to silicon chips, (b) Dimensional instability of organic core materials causing via pitch scaling limits, resulting in layer count escalation to > 24-26 build-up layers for HPC packages, and (c) bond pitch scaling limits induced by insufficient resolution of typical solder resist passivation materials and processes used for organic FCBGA substrates.

Onshoring Approaches: Recommendations for achieving on-shore capabilities of high-volume panel-level and wafer-level substrate manufacturing include multiple approaches. The first approach is to incentivize existing market leading substrate suppliers (both wafer and panel) to invest in capacity expansion for their US customers with manufacturing facilities in the US. An example of this type of investment is for companies like TSMC to setup advanced packaging wafer fab capacity onshore. The second approach is to incentivize existing onshore PCB manufacturers to invest in new capabilities for package substrate manufacturing. This approach will require bridging a major technology gap that exists between PCB and high-end package substrate processes, through setting up of advanced technology pilot lines that can feed a pipeline of technologies to the US package substrate/PCB manufacturers. Both these approaches have three common pre-requisites, (a) support from customers to procure advanced substrates from the new onshore locations, at potentially higher initial costs, (b) onshoring the materials, chemistry and equipment supply chains for advanced substrates, and (c) targeted workforce development skilled in advanced substrate technologies and manufacturing processes.

Table 2.4 Manufacturing flow along with unit process tools and associated materials for organic substrate, with key identified gaps.

A		**Carrier Preparation**						
			Equipment		**Material**			
Step #	**Process**	**Purpose**	**Equipment Type**	**Gap: Y/N**	**Material Type**	**Gap: Y/N**	**Comments/Remarks**	**Next Step: Recommended**
A-1	Dry-film negative photoresist lamination	Cladded Copper layer	Vacuum laminator	N	Dry-film photoresist	Y	Dry-film photoresist lamination equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
A-2	Photoresist exposure	Cladded Copper layer	Mask aligner/Stepper/Direct Imaging	N	Dry-film photoresist	Y	Direct-write maskless lithography exposure equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
A-3	Photoresist development	Cladded Copper layer	Spray Development Tool	N	Developing solution	Y	Dry-film photoresist development chemicals dominated by foreign suppliers	Need on-shore chemical capability for HVM supply chain
A-4	Copper etch for trace/pad patterning	Cladded Copper layer	Wet etch tool	N	Copper etch chemical	N		
A-5	Photoresist strip	Cladded Copper layer	Photoresist wet stripper or Ash	N	Wet Strip solution	N		
A-6	Pre-impregnated bonding sheet application	Multi-layer stack-up and structuring	Dry-film placement tool	N	Pre-impregnated bonding sheet	N		
A-7	Copper foil application	Multi-layer stack-up and structuring	Dry-film placement tool	N	Copper foil	N		
A-8	Stack pressing under high temperature and high pressure	Multi-layer stack-up and structuring	Pressing tool	N		N		
A-9	Through-core hole drilling	Multi-layer stack-up and structuring	Mechanical drill tool	N		N		
A-10	E-less Copper metallization of boreholes	Cladded Copper layer patterning	E-less Copper deposition tool	N	E-less Copper plating chemicals	N		
A-11	Dry-film positive photoresist lamination	Cladded Copper layer patterning	Vacuum laminator	N	Dry-film photoresist	Y	Dry-film photoresist lamination equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
A-12	Photoresist exposure	Cladded Copper layer patterning	Mask aligner/Stepper/Direct Imaging	N	Dry-film photoresist	Y	Direct-write maskless lithography exposure equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
A-13	Photoresist development	Cladded Copper layer patterning	Spray Development Tool	N	Developing solution	Y	Dry-film photoresist development chemicals dominated by foreign suppliers	Need on-shore chemical capability for HVM supply chain
A-14	Copper plating to reinforce circuit structure to desirable thickness	Cladded Copper layer patterning	Electrolytic Copper plating tool	N	Electrolytic Copper plating chemicals	N		
A-15	Etching mask deposition	Cladded Copper layer patterning	Immersion bath	N	Immersion Tin solution	N		
A-16	Photoresist strip	Cladded Copper layer patterning	Photoresist stripper tool wet or Ash	N	Dry-film photoresist	N		
A-17	Copper etch with Thin mask for trace/pad patterning	Cladded Copper layer patterning	Wet etch tool	N	Wet etch chemicals	N		
A-18	Thin mask removal	Cladded Copper layer patterning	Wet etch tool	N	Wet etch chemicals	N		
B		**Build-up Layer Fabrication**		**Typ. L/S**				
				> 7um				
			Equipment		**Material**			
Step #	**Process**	**Purpose**	**Equipment Type**	**Gap: Y/N**	**Material Type**	**Gap: Y/N**	**Comments/Remarks**	**Next Step: Recommended**
B-1	ABF placement	Dielectric layer application and patterning	Dry-film placement tool	N	Polyimide (PI) Benzocyclobutene (BCB) Polybenzobisoxazole (PBO) Ajinomoto Build-up Film (ABF)	Y	Dry-film dielectrics dominated by foreign suppliers	Need on-shore material capability for HVM supply chain
B-2	Vacuum lamination	Dielectric layer application and patterning	Vacuum laminator	N	Polyimide (PI) Benzocyclobutene (BCB) Polybenzobisoxazole (PBO) Ajinomoto Build-up Film (ABF)	Y	Dry-film photoresist lamination equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
B-3	Pre-cure	Dielectric layer application and patterning	Oven	N		N		
B-4	RDL via formation	Dielectric layer application and patterning	Laser system	N	Polyimide (PI) Benzocyclobutene (BCB) Polybenzobisoxazole (PBO) Ajinomoto Build-up Film (ABF)	Y	RDL via laser drilling equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
B-5	Barrier/ Copper seed deposition	Copper trace formation	E-less Copper deposition tool PVD barrier/ Copper seed deposition tool	N	E-less: E-less Copper plating chemicals PVD: Ti/ Cu	N		
B-6	Dry-film Potoresist lamination	Copper trace formation	Vacuum laminator	N	Dry-film photoresist	Y	Dry-film photoresist lamination equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
B-7	Copper plating	Copper trace formation	Electrolytic Copper deposition tool	N	Electrolytic Copper plating chemicals	N		
B-8	Photoresist strip	Copper trace formation	Photoresist stripper wet tool or asher	N	Dry-film photoresist	N		
B-9	Barrier/ Copper seed etch	Copper trace formation	Wet etch tool	N	Wet etch chemicals	N		
C		**Passivation, Metal Finishing, and Solder Bumping**						
			Equipment		**Material**			
Step #	**Process**	**Purpose**	**Equipment Type**	**Gap: Y/N**	**Material Type**	**Gap: Y/N**	**Comments/Remarks**	**Next Step: Recommended**
C-1	Vacuum lamination	Solder mask application and patterning	Vacuum laminator	N	Dry-film photoresist	Y	Dry-film photoresist lamination equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
C-2	Photoresist exposure	Solder mask application and patterning	Mask aligner/Stepper/Direct Imaging	N	Dry-film photoresist	N		
C-3	Photoresist development	Solder mask application and patterning	Developer Spray tool	N	Dry-film photoresist	Y	Dry-film photoresist development chemicals dominated by foreign suppliers	Need on-shore chemical capability for HVM supply chain
C-4	Cure	Solder mask application and patterning	Oven	N		N		
C-5	E-less protective metal deposition	Metal finishing and solder bumping	E-less deposition tool	N	ENEPIG or ENIG plating chemicals	N		
C-6	Solder ball drop	Metal finishing and solder bumping	Ball Drop tool	N	SAC305	N		
C-7	Reflow	Metal finishing and solder bumping	Reflow oven	N	SAC305	N		

Table 2.5 Manufacturing flow, process tools and materials for Silicon-core substrate.

A		Silicon Core Preparation		Typ. L/S				
				20um				
			Equipment		Material			
Step #	Process	Purpose	Equipment Type	Gap: Y/N	Material Type	Gap: Y/N	Comments/Remarks	Next Step: Recommended
A-1	Si through Hole formation	Si through Hole formation	Laser	N		Y	R&D laser drilling equipment demonstrated TSV pitch ≤ 100µm	Need on-shore equipment capability for HVM
A-2	Dielectric layer application	Dielectric layer application	Vacuum laminator	N	Polyimide (PI) Benzocyclobutene (BCB) Polybenzobisoxazole (PBO) Ajinomoto Build-up Film (ABF)	Y	Dry-film dielectrics dominated by foreign suppliers	Need on-shore material capability for HVM supply chain
A-3	Via-in-Via Hole formation	Via-in-Via Hole formation	laser	N		Y	R&D laser drilling equipment demonstrated via-in-via size ≤ 50µm	Need on-shore equipment capability for HVM
A-4	Copper seed deposition	Core copper trace formation	E-less Copper deposition tool	N	E-less: E-less Copper plating chemicals	N		
A-5	Dry-film Potoresist lamination	Core copper trace formation	Vacuum laminator	N	Dry-film photoresist	Y	Dry-film photoresist lamination equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
A-6	Copper plating	Core copper trace formation	Electrolytic Copper deposition tool	N	Electrolytic Copper plating chemicals	N		
A-7	Photoresist strip	Core copper trace formation	Photoresist wet tool or Asher	N	Wet Stripping solution	N		
A-8	Barrier/ Copper seed etch	Core copper trace formation	Wet etch tool	N		N		
B		Build-up Layer Fabrication		Typ. L/S				
				2um				
			Equipment		Material			
Step #	Process	Purpose	Equipment Type	Gap: Y/N	Material Type	Gap: Y/N	Comments/Remarks	Next Step: Recommended
B-1	ABF placement	Dielectric layer application and patterning	Dry-film placement tool	N	Polyimide (PI) Benzocyclobutene (BCB) Polybenzobisoxazole (PBO) Ajinomoto Build-up Film (ABF)	Y	Dry-film dielectrics dominated by foreign suppliers	Need on-shore material capability for HVM supply chain
B-2	Vacuum lamination	Dielectric layer application and patterning	Vacuum laminator	N	Polyimide (PI) Benzocyclobutene (BCB) Polybenzobisoxazole (PBO) Ajinomoto Build-up Film (ABF)	Y	Dry-film photoresist lamination equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
B-3	Pre-cure	Dielectric layer application and patterning	Oven	N		N		
B-4	RDL via formation	Dielectric layer application and patterning	Laser system	N	Polyimide (PI) Benzocyclobutene (BCB) Polybenzobisoxazole (PBO) Ajinomoto Build-up Film (ABF)	Y	R&D laser drilling equipment for RDL via size ≤ 15µm	Need on-shore equipment capability for HVM
B-5	Barrier/ Copper seed deposition	Copper trace formation	E-less Copper deposition tool PVD barrier/ Copper seed deposition tool	N	E-less: E-less Copper plating chemicals PVD: Ti/ Cu	N		
B-6	Dry-film Potoresist lamination	Copper trace formation	Vacuum laminator	N	Dry-film photoresist	Y	Dry-film photoresist lamination equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
B-7	Copper plating	Copper trace formation	Electrolytic Copper deposition tool	N	Electrolytic Copper plating chemicals	N		
B-8	Photoresist strip	Copper trace formation	Photoresist stripper	N	Dry-film photoresist	N		
B-9	Barrier/ Copper seed etch	Copper trace formation	Wet etch bench	N	Wet etch chemicals	N		
C		Passivation, Metal Finishing, and Solder Bumping						
			Equipment		Material			
Step #	Process	Purpose	Equipment Type	Gap: Y/N	Material Type	Gap: Y/N	Comments/Remarks	Next Step: Recommended
C-1	Vacuum lamination	Solder mask application and patterning	Vacuum laminator	N	Dry-film photoresist	Y	Dry-film photoresist lamination equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
C-2	Photoresist exposure	Solder mask application and patterning	Mask aligner	N	Dry-film photoresist	Y	Direct-write maskless lithography exposure equipment dominated by foreign suppliers	Need on-shore equipment capability for HVM
C-3	Photoresist development	Solder mask application and patterning	Developer bench	N	Dry-film photoresist	Y	Dry-film photoresist development chemicals dominated by foreign suppliers	Need on-shore chemical capability for HVM supply chain
C-4	Cure	Solder mask application and patterning	Oven	N		N		
C-5	E-less protective metal deposition	Metal finishing and solder bumping	E-less deposition tool	N	ENEPIG or ENIG plating chemicals	N		
C-6	Solder ball drop	Metal finishing and solder bumping	Solder ball drop tool	N	SAC305	N		
C-7	Reflow	Metal finishing and solder bumping	Reflow oven	N	SAC305	N		

2.5.2.2 Bond Pitch Scaling and Assembly:

This section of the blueprint covers bond pitch scaling for die-to-die, die-to-interposer and die-to-substrate interconnections.

a) Solder-based TCB (microbump)

Die-to-package interconnections migrated from lead-free solder bumps to copper pillars with lead-free solder caps, to copper microbumps with thin solder caps as the bond pitch scaled from 250 micron pitch die-to-substrate flip-chip interconnections to 35-45 micron pitch die-to-interposer interconnections. Mass reflow processes transitioned to thermo-compression bonding as the solder volume per bump reduced and interconnection areas increased. This historic roadmap trend is illustrated in **Figure 2.5**.

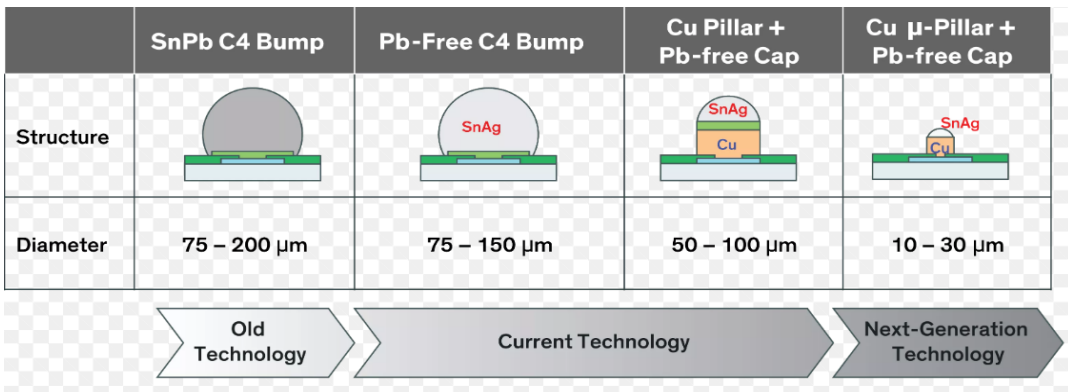


Figure 2. 5 Solder-based Historic Interconnect Roadmap & Fine-Pitch Cu-SnAg Microbumps

b) Solderless direct TCB (metal-metal)

Direct metal to metal thermal compression bonding (TCB) is a solderless bonding process. Intimate contact between metal pads on either side of the bonding interface can result in intermetallic diffusion and grain growth under appropriate conditions of temperature and pressure. This forms the basis of Direct metal-to-metal TCB. Being a solderless bonding process, bonding pitches of < 10 µm can be obtained by metal-to-metal TCB. Unlike hybrid bonding, dielectric is recessed to expose metal pads both on substrate side and dielet side for bonding. There is only metal-to-metal contact and no dielectric-to-dielectric contact. Since there is no dielectric bonding, dielectric roughness requirements are not critical. Surface asperities on bonding pads are flattened by temperature and pressure during thermal compression bonding. D2W-TCB is independent of the type of dicing used, so blade dicing is applicable. Furthermore, the level of particle control obtained through standard wet cleaning processes is adequate for successful assembly. Many choices for metals exist for metal-to-metal TCB. Gold-Gold TCB [3, 4], Gold-Copper TCB [5], Copper-Copper TCB [6, 7, 8], passivation metal-based Cu-Cu TCB [9, 10] have been demonstrated in literature.

To increase the throughput of direct Cu-Cu TCB, a two-step bonding approach discussed in [8] can be taken. The two-step approach constitutes die tacking to wafer scale or interposer substrate, followed by annealing of the wafer-to-wafer or die-to-wafer assembly. During the die tacking

stage, all the dies are aligned at a relatively low temperature of 120 °C and placed within a total time of ≤ 10 seconds/die. This step does not ensure final bonding, but it does guarantee a firm-enough attach with a shear strength > 10 N. Once populated, the assembly is batch annealed (batch size depends on furnace capacity) in vacuum for 1 hour. This step ensures Cu grain growth across the mating surfaces needed for successful bonding. **Figure 2.6** shows the thermal compression bonding process flow and **Figure 2.7** shows the cross-section SEM images of the bonded interconnects. A detailed process flow with manufacturing tools, materials, suppliers and roadmap challenges and gaps is listed in **Table 2.6**.

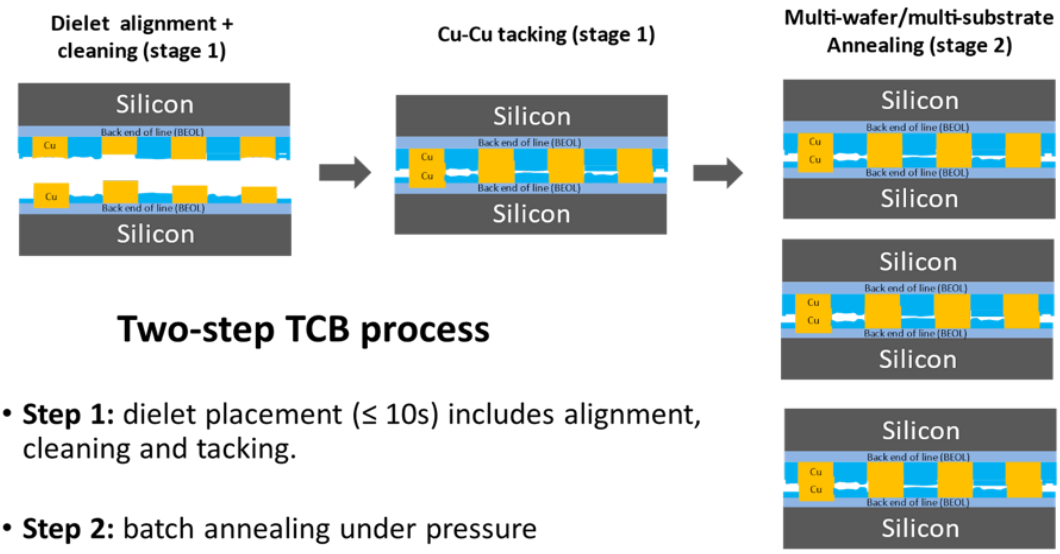


Figure 2. 6 ***Two step high throughput thermal compression bonding process [8]***

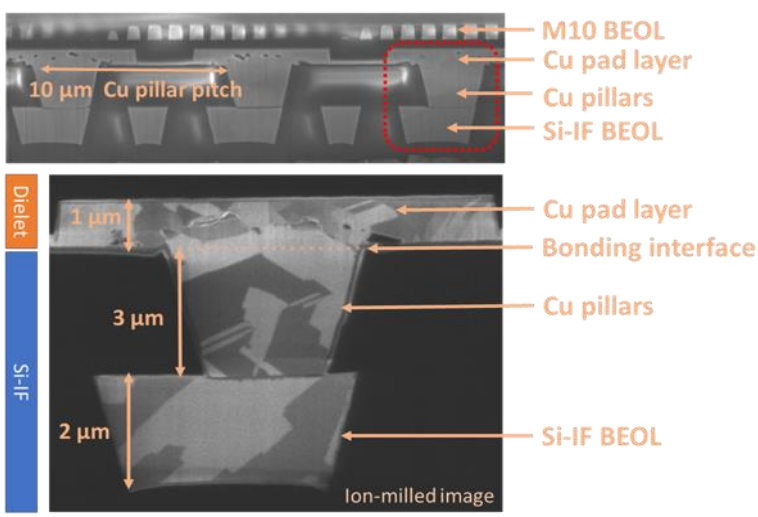


Figure 2. 7 ***Bonding cross-section in a sample Cu-Cu thermal compression bonding process [8]***

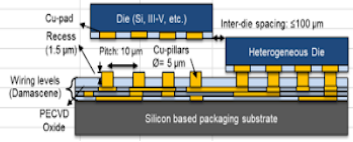
Table 2.6 Manufacturing flow along with unit process tools and associated materials for direct metal-metal thermal compression bonding, with key identified gaps.

Wafer-to-wafer (W2W) TCB process

			Equipment			Material				
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks	Next Step: Recommended
1a	RDL Build, CVD Dielectric dep.	Interconnect Dielect	PECVD	Plasmatherm, AMAT, LAM research	N	SiO2	EMD Electronics, Air Liquide, SK Chem, Hansol	N		
1b	RDL Build, CVD Polish Stop dep.	Interconnect Dielect	PECVD	Plasmatherm, AMAT, LAM research	N	Si3N4	Air Products, Air Liquide, SK Chem, Hansol	N		
2a	RDL Build, Litho and Pattern Pillar	Lithography	Track, Develop	ASML, Karl Suss, Heidelberg, TEL, EVG	N	Resist/CF4	DuPont, Microchemicals (AZ-series)	N		
2b	RDL Build Pillar etch	Etching	RIE etcher	STS, AMAT, LAM research, Oxford, Plasmathern	N	Resist/CF4	DuPont, Microchemicals (AZ-series)	N		
3	RDL Build, PVD Ti Adhesion/Cu Seed	Cu Barrier and Seed	Ebeam Evaporation/Spu	AMAT, CHA, ULVAC,	N	Ti/Cu	JX, Honeywell, Linde, ToSoh, KFMI	N		
4	RDL Build, Electroplate Cu	Interconnect Metal	Electroplating	Technic, LAM Research, AMAT	N	Cu Plating Solr	Technics, Entegris, DuPont	N		
5	CMP and CMP Clean	Planarization	CMP	AMAT, Ebara	N	Cu Slurry	Entegris/CMC, Fujimi, DuPont, etc.	N		
6	AFM	Examination	AFM	Parksystems, Bruker						
7	Pillar Exposure/Insulation Recess	Interconnect Metal	RIE	STS, AMAT, LAM research, Oxford, Plasmathern	N	CF4	Air Products, Air Liquide, SK Chem, Hansol	N		
8	Surface Activation	Prep for Bonding	Bonder/RIE	Oxford, AMAT, Lam research	N	HCOOH, Ar	Stellar, Thermo Scientific Chemicals etc.	N		
9	Align and Bond	Bond	W2W Bonder	SUSS MicroTec Inc., EVG, TEL	Y				Improvement in alignment accuracy for sub 0.5 um pad-to-pad alignment	
10	Annealing (optional)	Bond	W2W Bonder	SUSS MicroTec Inc., EVG	Y				more equipment companies in US need to support wafer-to-wafer bonding	
11	Dicing	separation of bonded assemblies		Disco (blade dicing), EVG, Plasmatherm (plas	Y				Integration is being developed.	

Die-to-wafer (D2W) TCB process

			Equipment			Material				
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks	Next Step: Recommended
1	RDL Build, CVD Dielectric Dep	Interconnect Dielect	PECVD	Plasmatherm, AMAT, LAM research	N	SiO2	EMD Electronics, Air Liquide, SK Chem, Hansol	N		
1b	RDL Build, CVD Polish Stop	Interconnect Dielect	PECVD	Plasmatherm, AMAT, LAM research	N	Si3N4	Air Products, Air Liquide, SK Chem, Hansol	N		
2a	RDL Build, Litho and Pattern Pillar	Lithography	Track, Develop	ASML, Karl Suss, Heidelberg, TEL, EVG	N	Resist/CF4	DuPont, Microchemicals (AZ-series)	N		
2b	RDL Build Pillar etch	Etching	RIE etcher	STS, AMAT, LAM research, Oxford, Plasmathern	N	Resist/CF4	DuPont, Microchemicals (AZ-series)	N		
3	RDL Build, PVD Ti Adhesion/Cu Seed	Cu Barrier and Seed	Ebeam Evaporation/Spu	AMAT, CHA, ULVAC,	N	Ti/Cu	JX, Honeywell, Linde, ToSoh, KFMI	N		
4	RDL Build, Electroplate Cu	Interconnect Metal	Electroplating	Technic, LAM Research, AMAT	N	Cu Plating Solr	Technics, Entegris, DuPont	N		
5	CMP and CMP Clean	Planarization	CMP	AMAT, Ebara	N	Cu Slurry	Entegris/CMC, Fujimi, DuPont, etc.	N		
6	AFM	Examination	AFM	Parksystems, Bruker						
7	Pillar Exposure/Insulation Recess	Interconnect Metal	RIE	STS, AMAT, LAM research, Oxford, Plasmathern	N	CF4	Air Products, Air Liquide, SK Chem, Hansol	N		
8a	Surface Activation	Prep for Bonding	Bonder/RIE	Oxford, AMAT, Lam research	N	HCOOH, Ar	Stellar, Thermo Scientific Chemicals etc.			
8b	inert environment	Prep for Bonding	in-situ within bonder	K&S, BESI, SET	Y	N2 or Ar		Y	More tool companies need to provide inert environment. Inert environment will help to improve D2W bonding thorughput.	
9	Align and tack	Bond	D2W Bonder	K&S, BESI, SET	Y				Improvement in alignment accuracy for sub 1 um pad-to-pad alignment at high temperatures required by TCB	
10	Annealing the fully populated substr	Bond	W2W annealing	SUSS MicroTec Inc., EVG	Y				more equipment companies in US need to support annealing under pressure	
11	Dicing	separation of bonded assemblies		Disco (blade dicing), EVG, Plasmatherm (plas	Y				Integration is being developed.	



c) Hybrid Bonding (Die-to-Wafer and Wafer-to-Wafer)

Hybrid bonding, where dielectric materials are bonded together followed by an anneal which generates the Cu-to-Cu bonding, already has been in high volume manufacturing (HVM) since 2016 when Sony was the first to produce image sensors with hybrid bonding technology. Then in 2021, YMTC leveraged hybrid bonding for their 128L 3D NAND, and in 2022 AMD utilized TSMC's SOIC technology for their Ryzen 7 processor. Currently there are three main approaches for hybrid bonding shown in **Figure 2.8**: (1) wafer to wafer (W2W) approach utilized by CIS and 3D NAND, (2) collective D2W where die are reconstructed on a carrier prior to bonding to a wafer or another set of die on carrier, and (3) single die to wafer or chip to wafer (D2W or C2W) using flip chip bonding.

The main advantage of hybrid bonding over micro bumps is the increase in interconnect density with efforts to reduce W2W pitches to sub-1 or even sub-0.5 um and to reduce D2W pitches to below 4 um. These aggressive pitches create process challenges which include maintaining clean surfaces, having controlled and uniform Cu dishing as well as surface topography, and retaining alignment accuracy during bonding. Surface cleanliness, for example, is driving development of laser and plasma dicing to minimize debris generated. Organic and inorganic temporary bonding and protective layers are also being developed to minimize surface defects. Planarization challenges drives efforts to improve CMP processes and requires efficient in-line post-CMP metrology. In-line, non-destructive characterization is also needed for defect and void detection which is even more critical for multi-die or multi-wafer stacks. Other challenges include mechanical and thermal considerations. Warpage and mechanical issues are concerns as wafers and die are thinned. High Bandwidth Memory (HBM), which could have 8-20 stacked die, requires lower bonding temperatures. A variety of dielectric materials are being developed to reduce bonding temperature while maintaining bond strength, and Cu grain structures are being investigated to reduce the thermal budget required for Cu-to-Cu bond formation. As chiplets and die-to-die (D2D) bonding become more established, multiple bonding approaches will be developed to address additional integration challenges.

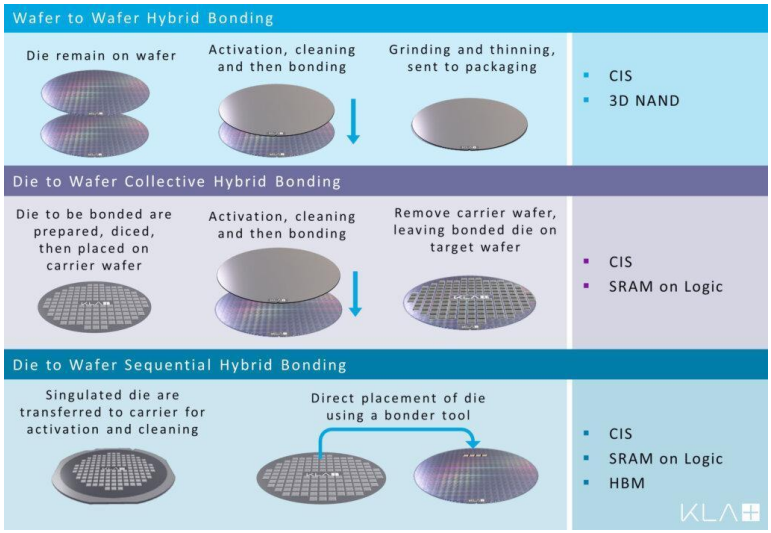


Figure 2. 8 Hybrid Bonding Approaches and Use Cases [11]

Figure 2.9 shows a wafer-to-wafer hybrid bonding process flow and **Figure 2.10** shows a die-to-wafer hybrid bonding process flow. **Table 2.7** shows the detailed manufacturing flow with materials, equipment, selected suppliers, and manufacturing challenges and gaps in the roadmap for W2W hybrid bonding. A similar analysis is summarized in **Table 2.8** for D2W hybrid bonding.

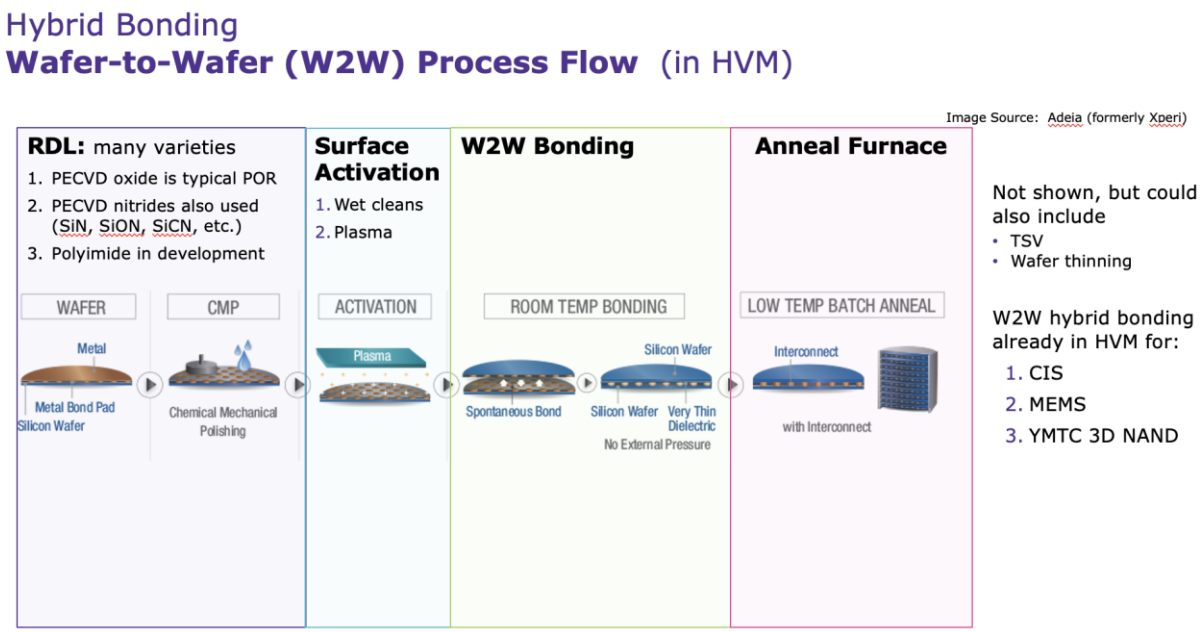


Figure 2. 9 ***Wafer-to-Wafer Hybrid Bonding Process Flow***

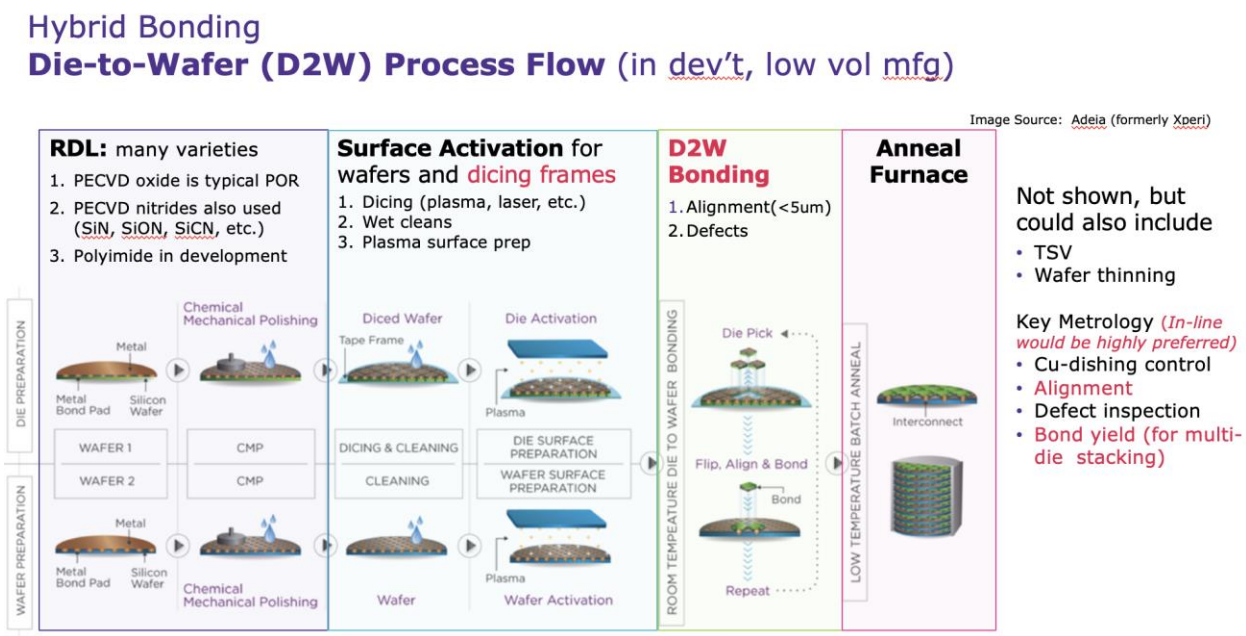


Figure 2. 10 ***Die-to-Wafer Hybrid Bonding Process Flow***

Table 2.7 Wafer-to-Wafer Hybrid Bonding Process Flow, Materials, Equipment, and Gaps

A		Wafer to Wafer Hybrid Bonding			Typ. L/S (um)				
					<1um				
			Equipment*			Material*			
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks
1	RDL build, CVD dep & pattern	Interconnect dielectric	PECVD	AMAT, LAM, TEL	N	TEOS, SiH4, O2	EMD Electronics, Air Liquide, SK Chem,	N	In production for CIS, 3DNAND, MEMs
1b	RDL build (optional CVD nitride)	Interconnect dielectric	PECVD	AMAT, LAM, TEL	N	SiH4, NH3	Air Products, Air Liquide, SK Chem,	N	
2	RDL Litho / Etch	Patterning	Track, Develop, Etch	AMAT, LAM, TEL	N	Resist / Etch Gases (CF4, etc.)	JSR, TOK, DuPont	N	
3	RDL build, PVD barrier & Cu seed	Cu barrier & seed	PVD -> ALD	AMAT, Ulvac, Shibaura, ASMPT, Evatec	N	Ti, Ta, TiN, TaN Cu targets	JX, Honeywell, Linde, ToSoh, KFMI	N	
4	RDL build, Electroplated Cu	Interconnect metal	Electroplating (EP)	AMAT,	N	Cu plating	Entegris, BASF, MLI, DuPont	N	
5	CMP & post CMP clean	Planarization	CMP	AMAT, Ebara	N	CMP Pads, Slurries, Cleans	Entegris/CMC, Fujimi, DuPont, etc.	N	
6	Metrology	Cu dishing, Erosion	Full Wafer AFM	Bruker, Park	N				Higher throughput / Full wafer technique would be nice to have
7	Clean	Surface treatment	Wet Clean (Single Wafer)	Lam,	N	Wet Chemistry		N	
8	Surface Activiation	Surface treatment	Plasma (e.g. etch)	Lam, AMAT, TEL, PlasmaTherm, Oxford	N	Ar, N2 gases	Air Liquide, Air Products, Linde	N	
9	W2W Bonding (with alignment)	Bond	Wafer Bonder	EVG, TEL, Suss	N	N/A	N/A	N	
10	Anneal	Thermal	Furnace **/ Oven**		N	Ar, N2 gases	Air Liquide, Air Products, Linde	N	
11	Dicing	Dicing	Saw Dicing Tool	Disco	N				
12	Metrology	Bond strength, Voids, Alignment	Voids - IR, SAM; BondStrength - Maszara	EVG, KLA					
Optional Steps									
	Backgrind								
	Edge Trim								
	Clean								
	Temp Bond/Debond								

Table 2.8 ***Die-to-Wafer Hybrid Bonding Process Flow, Materials, Equipment, and Gaps***

A - Wafer		Die to Wafer HB			Typ. L/S (um)					
					< 10um -> < 1um					
			Equipment*			Material*				
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks	Next Step: Recommended
0	**Bottom device wafer**	**Base Wafer**								
1	RDL build, CVD dep & pattern	Interconnect dielectric	PECVD	AMAT, LAM, TEL	N	TEOS, SiH4, O2	EMD Electronics, Air Liquide, SK	N	**PI alternative proposed in 2022, didn't see adoption in 2023 - could check with Resonac**	
1b	RDL build (optional CVD nitride)	Interconnect dielectric	PECVD	AMAT, LAM, TEL, SPTS	N	SiH4, NH3, N2, H2	Air Products, Air Liquide, SK	N		
2	RDL Litho / Etch	Patterning	Track, Develop, Etch	AMAT, LAM, TEL	N	Resist / Etch Gases (CF4, etc.)	JSR, TOK, DuPont	N		
3	RDL build, PVD barrier & Cu seed	Cu barrier & seed	PVD -> ALD	AMAT, Ulvac, Shibaura,	N	Ti, Ta, TiN, TaN Cu targets	JX, Honeywell, Linde, ToSoh, KFM	N		
4	RDL build, Electroplated Cu	Interconnect metal	Electroplating (EP)	AMAT,	N	Cu plating	Entegris, BASF, MLI, DuPont	N		
5	CMP & post CMP clean	Planarization	CMP	AMAT, Ebara	N	CMP Pads, Slurries, Cleans	Entegris/CMC, Fujimi, DuPont, e	N		
6	AFM	Cu dishing, Erosion Meas.	Metrology - AFM	Bruker, Park	N				Higher throughput / Full wafer technique would be nice to have	

A -Die - Single Sided										
			Equipment*			Material*				
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks	Next Step: Recommended
0	**Die device wafer**	**Die**								
1	RDL build, CVD dep & pattern	Interconnect dielectric	PECVD	AMAT, LAM, TEL,	N	TEOS, SiH4, O2	EMD Electronics, Air Liquide, SK	N	PI Alternative	
1b	RDL build (optional CVD nitride)	Interconnect dielectric	PECVD	AMAT, LAM, TEL, SPTS	N	SiH4, NH3, N2, H2	Air Products, Air Liquide, SK	N		
2	RDL Litho / Etch	Patterning	Track, Develop, Etch	AMAT, LAM, TEL	N	Resist / Etch Gases (CF4, etc.)	JSR, TOK, DuPont	N		
3	RDL build, PVD barrier & Cu seed	Cu barrier & seed	PVD -> ALD	AMAT, Ulvac, Shibaura,	N	Ti, Ta, TiN, TaN Cu targets	JX, Honeywell, Linde, ToSoh, KFM	N		
4	RDL build, Electroplated Cu	Interconnect metal	Electroplating (EP)	AMAT,	N	Cu plating	Entegris, BASF, MLI, DuPont	N		
5	CMP & post CMP clean	Planarization	CMP	AMAT, Ebara	N	CMP Pads, Slurries, Cleans	Entegris/CMC, Fujimi, DuPont, e	N		
6	Metrology	Cu dishing, Erosion	Full Wafer AFM	Bruker, Park	N					
7a	**Organic Temp Bond**	Connect to carrier	Spin-on, Temp Bond	TEL, EVG, Suss,	N	Organic(s)	Brewer, TOK, 3M,	Integration dev't ongoing	**Temp processing and residue issues with organic temp bond materials**	
7b	**Inorganic Temp Bond**	Connect to carrier	CVD, Temp Bond	AMAT, LAM, TEL, (TBDB-tbd)	N	Inorganic (e.g. SiOx)		Integration dev't ongoing	**Proposal for inorganic temp bond from Disco & Micron in 2023**	
8	Backside grind	Wafer Thinning	Backside Grind	Disco, Axus	N	CMP Pads, Slurries, Cleans	Brewer, 3M, Shin-Etsu	N		
9	Debond	Release wafer	Thermal, mech, laser...	TEL/IBM, EVG, Suss	Integration dev't ongoing			Integration dev't ongoing		
10	**Clean (Front Side)**	Remove temp bond material	Wet or Dry Etch	TEL, Veeco, AMAT, Shibaura	Integration dev't ongoing	Clean	EMD, DuPont,	Integration dev't ongoing	Improve clean for temp bond residue	
Optional	**Sacrificial protective layer**	Defectivity Reduction	tbd		tbd	tbd		Integration dev't ongoing		
11	**Dicing**	Wafer -> Die	Mech, Laser, Plasma	Disco, EVG. SPTS	Integration dev't ongoing	Dicing tape/Frame	3M, Lintech, Sumitomo Bakelite	Integration dev't ongoing	Reduce defects, clean edge desired	

A -Die - Double Sided, Via Middle										
			Equipment*			Material*				
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks	Next Step: Recommended
1	TSV Fabrication	interconnect thru b/s	Etch, Litho, CVD, PVD, ALD, E	TEL, AMAT, LAM, SPTS...						
2	RDL on Frontside	for hybrid bonding	(see above)							
3	Metrology	Cu dishing, Erosion	Full Wafer AFM	Bruker, Park	N				Higher throughput / Full wafer technique would be nice to have	
4	Temp Bond	Attach to Carrier	(see above)							
5	Backgrind	Wafer Thinning	Backside Grind	Disco, Axus	N	CMP Pads, Slurries, Cleans	Brewer, 3M, Shin-Etsu	N		
6	CMP	Polish Si	CMP	AMAT, Ebara	N	CMP Pads, Slurries, Cleans	DuPont, Ferro	N		
7	Dry etch	TSV Reveal	Plamsa etch	AMAT, Lam, TEL	N	Etch gases		N		
8	Passivation	TSV Planarization	CVD (diel)	AMAT, Lam, TEL	N	TEOS, SiH4, O2	EMD Electronics, Air Liquide, SK	N		
9	CMP	TSV Planarization	CMP	AMAT, Ebara	N	CMP Pads, Slurries, Cleans	Entegris, EMD, DuPont,	N		
10	(Optional RDL/Cu Pads)									
11	Metrology	Cu dishing, Erosion	Full Wafer AFM	Bruker, Park	N				Higher throughput / Full wafer technique would be nice to have	
12	Debond	Release wafer	Thermal, mech, laser...	TEL/IBM, EVG, Suss	Integration dev't ongoing			Integration dev't ongoing		
13	Clean	Remove temp bond material	Wet or Dry Etch	TEL, Veeco, AMAT, Shibaura	Integration dev't ongoing	Clean	EMD, DuPont,	Integration dev't ongoing	Improve clean for temp bond residue	
Optional	RDL on backside (TSV-side)	Increase pad CD for bonding	(see RDL flow above)							
14	Dicing	Wafer -> Die	Mech, Laser, Plasma	Disco, EVG. SPTS	Integration dev't ongoing	Dicing tape/Frame	3M, Lintech, Sumitomo Bakelite	Integration dev't ongoing	Reduce defects, clean edge desired	

A - Die to Wafer										
			Equipment*			Material*				
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks	Next Step: Recommended
	Wafer & Die on Tape Frame									
1	Clean (optional)	particle & residue removal	Wet	Besi/AMAT - flip chip EVG/ASMPT - collective Suss/SET - collective	Integration dev't ongoing	Chemical	EMD, DuPont, Entegris	Integration dev't ongoing		
2	Plasma Activation (Wfr & Die)	create dangling bonds	Plasma / Etch			gases	Air Products, EMD, Air Liquide			
3	Wet Clean / Rinse	final surface prep	Wet			Chemical, H2O	EMD, DuPont, Entegris			
4	Alignment (on tool?)	align die to wafer	IR			n/a	n/a			
5	Bond	diel bond	Flip-chip / Collective			n/a	n/a			
6	Anneal	Cu-to-Cu bonding	Furnace/Oven			gases	Air Products, EMD, Air Liquide			
7	Inspection	Void detection, Alignment	IR, Scanning Acoust Tomography/Microsc	Hitachi, EVG	In-line? Cleanliness?	n/a	n/a			
For die stacking										
8	Clean wafer w/ Die1	particle & residue removal	Wet	Besi/AMAT - flip chip EVG/ASMPT - collective Suss/SET - collective	Integration dev't ongoing	Chemical	EMD, DuPont, Entegris	Integration dev't ongoing		
9	Plasma Activation - wafer w/ Die1	create dangling bonds	Plasma / Etch			gases	Air Products, EMD, Air Liquide			
10	Wet Clean / Rinse	final surface prep	Wet			Chemical, H2O	EMD, DuPont, Entegris			
11	Alignment (on tool?)	align die to wafer	IR			n/a	n/a			
12	Bond	diel bond	Flip-chip / Collective			n/a	n/a			
13	Anneal	Cu-to-Cu bonding	Furnace/Oven			gases	Air Products, EMD, Air Liquide			
14	Inspection	Void detection, Alignment	IR, Scanning Acoust Tomography/Microsc	Hitachi, EVG	In-line? Cleanliness?	n/a	n/a			
	Repeat for stacking									

Manufacturing Roadmap Gaps and Challenges: The most critical need in the next 5-10 years for hybrid bonding processes is to increase the manufacturing process throughput, reduce the equipment and cleanroom cost, and bring down the overall process cost closer to parity with current thermo-compression bonding (TCB) manufacturing processes. This will ensure that hybrid bonding expands to high volume applications beyond high-end AI and HPC chipsets, while enabling the pitch scaling beyond the limitations of TCB methods. Other significant challenges are stress management for 3D heterogenous die stacks to meet the long-term reliability requirements, much improved thermal management methods to limit localized heat induced failures, and metrology tools with integrated machine learning to address the electrical test costs associated with millions of fine pitch die-to-die interconnects. Polymer-based hybrid bonding methods are being explored and developed by a number of companies and research groups around the world, and this is an important area for potential future investments to address the throughput, cost and reliability concerns of oxide-based hybrid bonding, and ultimately expand the market for hybrid bonding.

Onshoring Opportunities: Hybrid bonding represents one of the closest processes to front end of line (FEOL) transistor manufacturing, which is one of the few areas that has a significant onshore footprint (>10% share of global manufacturing). The ongoing Chips Act driven investments in onshoring front end transistor fabs in the US can have a positive effect on hybrid bonding and 3D IC onshoring as well, and investments in fabs should be complemented by investments in hybrid bonding and other 3D packaging architectures. Lower cost emerging alternatives to hybrid bonding, such as direct Cu-Cu thermo-compression bonding and polymer hybrid bonding are excellent channels to enable onshoring of leading-edge OSATs, both existing and new players.

2.5.2.3 Fanout Wafer and Panel Level Packaging

A fanout wafer level package (FO-WLP) is a substrateless package that uses a rigid carrier and molding to reconstitute one or more ICs into a wafer form, typically 300mm diameter, and form re-distribution layers (RDL) directly on the reconstituted wafers to create direct copper interconnections to the I/O pads on the ICs [12]. One or more RDL layers are used to “fanout” the I/O on the ICs to a larger pitch for direct BGA assembly to the motherboard. Thus, fanout packages eliminate both the substrate as well as the solder-based chip-to-substrate assembly used in FCBGA and FCCSP packages. Infineon developed and commercialized the first large-scale FO-WLP packages with its e-WLB (embedded wafer-level ball grid array) packages. The introduction of FO-WLPs by TSMC with its InFO (Integrated Fan Out) packaging technology for iPhone application processors put fanout packages on the map of highest volume packaging platforms in use today. More recent trends in fanout packages include the move to 600mm x 600mm panels (FO-PLP), chip-last fanout methods (also called RDL interposers) such as TSMC CoWoS-R, and ASE FoCoS for 2.5D integration, and multi-die fanout packages with embedded silicon bridges for high density interposers (e.g. embedded fanout bridge (EFB) implemented by AMD in high end products). There are many variants in the current fanout wafer and panel-level package manufacturing landscape. These variants have been organized into three major technology categories based on process flows as shown in **Figure 2.11**. Generic process flows for each of these three groups are illustrated in **Figure 2.12 (a)** and **(b)**, and **Figure 2.13**.

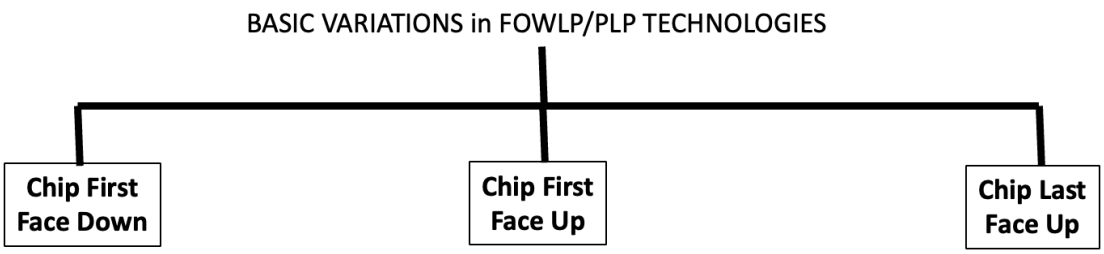
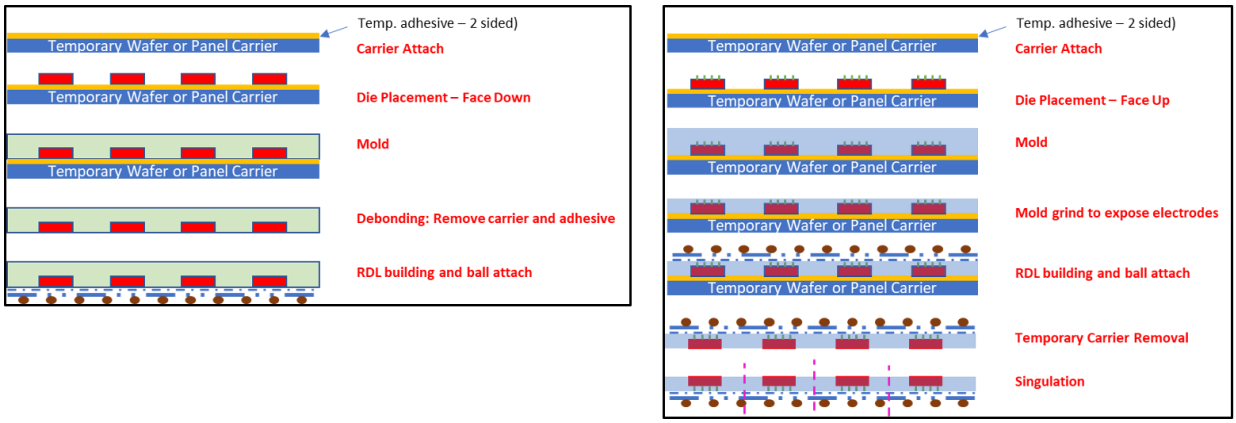


Figure 2. 11 ***Three Major Fanout WLP/PLP Technology Categories based on Process Flows.***



(a) (b)

Figure 2. 12 ***Generic Process Flows for Chip-First Fanout Package Fabrication (a) Face Down FO-WLP, (b) Face Up FO-WLP.***

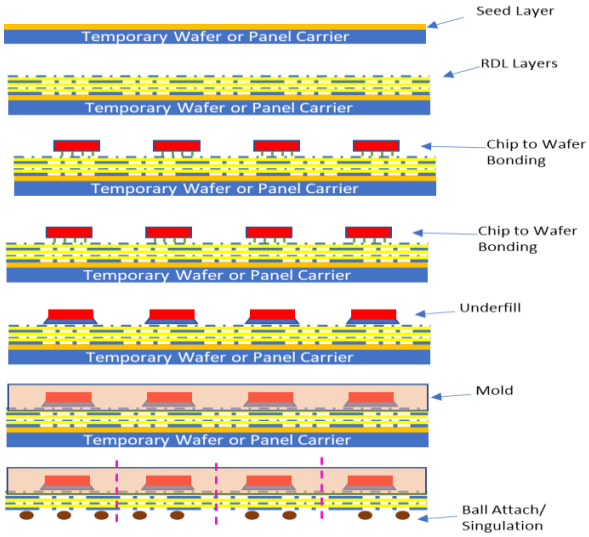


Figure 2. 13 ***Generic Process Flow for Chip-Last or RDL-First Fanout Package Fabrication.***

Table 2.9, **Table 2.10**, and **Table 2.11** summarize a generic view of the key process flows, materials and equipment ecosystem, and gaps.

Table 2.9 Chip-First, Face-Down, FOWLP/PLP Process Flow, Materials, Equipment, & Gaps

A		Chip First Face Down FOWLP/FOPLP:			Typ. L/S (um)				No bumps on the die; potential yield loss (RDL defects). Original eWLB and Fraunhofer Concept	
					10/10					
			Equipment			Material				
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks	Next Step: Recommended
1	Carrier preparation : Attach thermal relase tape	Film lamination	Film lamination	Dynachem, Italy: Dow Chemical	N	Temporary bonding tape. Thermal, laser, mechanical release films and Si / Glass carrier wafers.	3M, Shin-Etsu, AI Technology.			
2	Die placedment	Die placed upside down on carrier.	ASM Nucleus: 5um +/- 3 Sigma	ASMPT, Singapore	N					
3	Mold the panel	Encapsulation: Granular MC or Film	Compression Mold	APIC Yamada, Japan	N	Epoxy Mold Compound	Resonac, Sumitomo, Resonac (Film)		Granular MC is preferred for Panel process	
4	Debond the carrier	Debond Carrier Wafer	Thermal debonder	ERS Panel debonder, Germany	N					
5	Inspection for die position	To check die position and adjust RDL pattern	Optical Inspection	Mahr, Onto	N					
6	Dielectric application	For RDL creation	Film vacuum lamination	Dynachem, Italy	N	Polymer Dry Film	Ajinomoto, DuPont, Resonac, 3M			
7	Drilling: Microvias: Laser process	Microvia drilling	Laser Drill	Schmoll, Germany	N					
8	DRIE	Cleaning and dry etch	Dry etch	Evated, Switzerland	N	CF4	Air Products, Air Liquide, SK Chem, Hansol			
9	Lithography	Maskless, Laser Direct Imaging	RDL Equipment	Schmoll, Germany; Orbotech, Israel, Visitech, Norway	N	Resist/CF4	DuPont, Microchemicals (AZ-series)			
10	Sputtering metal deposit		PVD Seedlayer creation	Creavac, Germany, Evatec, Germany	N	Resist/CF4	DuPont, Microchemicals (AZ-series)			
11	Development, electroplating	Electroplating and resistor etch	RDL Processes	Manz Germany, Atotech USA, Technics	N	Cu Plating Solution	Technics, Entegris, DuPont			
12	Inspection	Optical inspection of RDL patttern	AOI; meed high speed process	Onto Innovation, USA	N					
13	E-Test	Circuit test of RDL sites.	4-point test	SPEA, Italy	N					
14	Solderball attach	Attach solderballs for flip chip bonding	Solder ball mounting tool	PacTech	N					
15	Marking				N					
16	Dicing tape mount	Mount Wafer for Singulation	Wafer Dicing	Disco	N	Polyester Tape (Sticky or UV release)	3M, Nitto Denko			
17	Singulation	Dice Wafer into coupons	Wafer Dicing	Disco	N					

Table 2.10 Chip-First, Face-Up, FOWLP/PLP Process Flow, Materials, Equipment, & Gaps

B		Chip First Face up W/ WO Adaptive Control for RDL			Typ. L/S (um)				Bumped die, L/S 2/2 um and below: under development. W/O adaptive control, potential yield loss. Deca Tech. concept	
					8/8 -> 2/2					
			Equipment			Material				
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks	Next Step: Recommended
1	Die bumping		Solder ball mounting tool	Pac Tech	N					
2	Die thinning		Grinding	Disco	N	Colloidal Silica Slurry	Sun-Tec			
3	Glass carrier with Cu seed layer	Temp. film attach to the carrier plate	Lamination	Dynachem, Nikko Materials	N	Thermal Release Tape or Adhesive	Nitto Materials, 3M		Glass carrier for Panel FOWLP? 600mm x 600mm	
4	Cu Stud plating	To enable second side RDL: Optional	Electroplater	Technic, LAM Research, AMAT	N	Barrier/Cu	Technics, Entegris, DuPont			
5	Chip attach	Chip face-up.	Chip Bonder	Finetech, Germany, UIC, US, Shibaura, Japan, ASM Netherlands, KnS	N					
6	Molding	Encapsulation	Compression Molding	Apic Yamada, Towa	N	Epoxy Mold Compound	Resonac, Sumitomo, Resonac (Film)			
7	Die Front side planarization	Expose electrodes	Grind over mold surface	Disco, Japan	N					
8	Die position data	Optional: Inspect and gather all die position data for the entire panel (implemented by Deca)	Optical Measurement System	Mahr, Onto	N				Adaptive patterning usedby Deca to improve yield and enable finer L/S	
9	Polymer coat & Cure	For RDL Process:	Vacuum Laminator	Dynachem		Polymer Dry Film	Ajinomoto, DuPont, Resonac, 3M			
10	RDL Pattern Creation	Laser Direct Imaging or photolithography	LDI Lithography	Schmoll, Orbotech	N	Resist	AZ Microchemicals, Fujifilm, Dupont, TOK			
11	Polymer coat, pattern & Cure	Second layer RDL	Vacuum Laminator	Dynachem, Nikko Materials	N	Resist/CF4				
12	UBM Pattern & Plate	To create pads for ball attach	Electroplater	Technic, LAM Research, AMAT	N	Cu Plating Solution	Technics, Entegris, DuPont			
13	Ball attach	Ball placement and reflow	Solder ball mounting tool	Pac Tech, BEST	N					
14	Carrier debond	Debond Carrier Wafer	Thermal or laser debonder	EVG		Thermal Release Tape or Adhesive	Nitto Denko, 3M			
15	Seed Layer Etch	Etch Cu Seed Layer				Ammonium Persulfate				
16	Laser Mark	Mark wafers for identification								
17	Remove Temp. film over solder balls		Solder ball mounting tool	Pac Tech						
18	Dicing tape mount	Mount Wafer or Panel for Singulation	Wafer Dicing	Disco		Polyester Tape (Sticky or UV Release)	3M, Nitto Denko			
19	Singulation	Dice Wafer/Panel into coupons	Wafer Dicing	Disco						

Table 2.11 ***Chip-Last FOWLP/PLP Process Flow, Materials, Equipment, and Gaps***

C		RDL First (Chip Last, Face Up): Similar to a FCBGA, except die is attached to multi-layer RDL built on a carrier.			Typ. L/S (um)				Bumped die. KGD on a known good RDL site. TSMC inFo has the same concept. This variation can have plated TMV to enable 3D option.	
					5/10 -->2/2					
			Equipment			Material				
Step #	Process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N	Comments/Remarks	Next Step: Recommended
1	Coat temp. adhesive on carrier wafer	Shoud withstand RDL temperatures and be releaseable.			N	Temporary bonding tape. Thermal, laser, mechanical release films and Si / Glass carrier wafers.	3M, Shin-Etsu, AI Technology.			
2	Coat dielectric material	To create RDL	Lithography, Coaters	Veeco, TEL, AMAT	N	Polymer Dry Film/Resist	AZ Microchemicals, Fujifilm, Dupont, TOK			
3	Build RDL,	To create RDL	Developers, Metal Deposition	TEL, AMAT, LAM	N	Cu Plating Solution	Technics, Entegris, DuPont			
4	Chip with microbumps	Microbumping process	Solder ball mounting tool	Pac Tech	N					
5	Chip attach	Die placement	Chip Bonder	Finetech, Germany, UIC, US, Shibaura, Japan, ASM Netherlands, KnS	N					
6	Reflow	Solder bumps Reflow	Reflow Oven	Heller	N					
7	Molding	Encapsulation	Compression Molding	Apic Yamada, Towa	N	Epoxy Mold Compound	Resonac, Sumitomo, Resonac (Film)			
8	Marking	Mark wafer for identification	Laser		N					
9	Carrier removal	Remove carrier wafer	Temporary Debonder	EVG	N	Thermal Release Tape/Adhesive	Nitto Denko, 3M			
10	Ball attach	Ball placement and reflow	Solder Bump Mounting	Pac Tech	N					
11	Dicing tape mount	Mount Wafer or Panel for Singulation	Wafer Dicing	Disco	N	Polyester Tape (Sticky or UV Release)	3M, Nitto Denko			
12	Singulation	Dice Wafer or Panel	Disco	Disco	N					

Manufacturing Roadmap Gaps and Challenges: Fanout wafer-level packaging is well established in high volume manufacturing today. Fanout panel-level packaging (PLP) is attracting interest from display manufacturers in Asia at 600mm x 600mm panel sizes, however, the process flows for packaging are significantly different than the capabilities of the legacy display fabs, and the lack of knowhow among the display manufacturers is an additional barrier. One of the key challenges for both wafer and panel-level fanout packages is the die shift during molding, which limits bump pitch scaling. Adaptive patterning and software-based correction techniques have been applied to partially address the die shift issue, but new innovations in materials and process flows will be required to meet future bump pitch scaling needs.

Onshoring Opportunities: Although fanout wafer and panel-level packaging is one of the highest volume packaging platforms for mobile and other devices, there are no high volume or even low volume fanout packaging lines in the US. This is a major onshoring gap identified in this roadmap. Migrating current fanout packaging production from Asia to the US is a possibility, however, it will be difficult to compete with the existing high volume production lines in Asia that have been optimized for several years and are running at high yields. Investing in new fanout approaches that address the future roadmap needs for single and multi-die fanout packages needs to be a focus of onshoring investments.

2.5.2.4 Silicon Photonics Packaging

Silicon Photonics (SiPh) packaging has emerged as an important interconnect platform for a large variety of applications including HPC, data center, and AI. The predominant interconnects between optical compute devices are optical fibers, often as legacy single mode fibers installed in and between existing facilities. On chip photonic IO will require decreases in fiber pitch from 250um today (for 125um diameter cladding fibers) to 140 or 125 um pitch enabled by 80um diameter cladding in the next generation. Finer IO pitches over the next decade are anticipated as multicore single mode and polarization maintaining fibers are developed and fiber ribbons are commercialized.

Fibers are attached to Photonic Integrated Circuit (PIC) die or chiplets by methods that include edge (butt) coupling using active or passive alignment self-alignment processes to an edge facet, edge V-groove self-alignment, or top surface grating coupler structures. More advanced coupling using adiabatic coupling or plug/mirror sub-assembles are also being researched and developed to facilitate package or board level integration. Figure 2.14 below shows examples of fiber arrays attached through different methods. Increasing fiber count from 2-8 fibers today to numbers approaching 100s per PIC die will be required in the next 5-10 years.

Future designs will incorporate single mode fiber (SMF) and polarization maintaining fiber (PMF) into co-packaged optics (CPO) using new advanced optical packaging techniques to complement heterogenous integration of electrical chiplets. CPO offers the highest bandwidth density and lowest power requirements for moving data while simultaneously providing thermal and reliability advantages using an external high-power laser. This is particularly important as AI data center and inter data center applications expand. CPO solutions currently in low volume production will require new packaging to efficiently extend data center power and bandwidth limits. Tooling advancements, extensive Design for Test (DFT) implementation and high-speed test and assembly platforms are necessary to enable high volume manufacturing.

Figure 2.15 and figure 2.16 below shows examples of CPO modules and fiber attach connectors.

Development is needed to increase fiber IO density with reduced fiber pitch, increase fiber count per PIC, drive improvements in link budget loss in fiber and laser attach processes as well as improvements in wafer and assembly photonic testing techniques to assure high yields for the most stringent system link budgets. Reliability of the package and system will need to include the interaction of fibers with the traditional chip-package-interaction (CPI) elements to drive chip-fiber-package-interaction (CFPI) requirements to secure acceptable reliability and yield. Many pluggable optical IO PIC connections are bulky compared to direct fiber attach methods. Optical interconnect reliability demonstration expanding beyond TELCORDIA to include JEDEC, MIL and AEC test menus should be a major focus. Passing these tests is very dependent on the choice of package design, materials and assembly operations in collaboration with suppliers. This roadmap will be consistent with the DARPA Photonics in the Package for Extreme Scalability (PIPES) targets of 100 Tbps per package at energies less than 1 picojoule per bit. Photonics will also play important roles in next generation light detecting and ranging (LiDAR), advanced driver assist systems (ADAS), wearable medical device IOT and other consumer applications. Some of these applications also require the assembly of III-V laser diodes onto the PIC die. While today, there may be 1-2 laser(s) per PIC, 4-16 laser diodes per PIC may be needed in the future. Figure 2.17 shows an example of a laser diode integration development data to PIC chiplet. This integration adds additional complexity to substrate assembly, thermal management, module yield, and reliability management of photonic integrated systems since lasers are often a single point of failure (SPOF) concern.

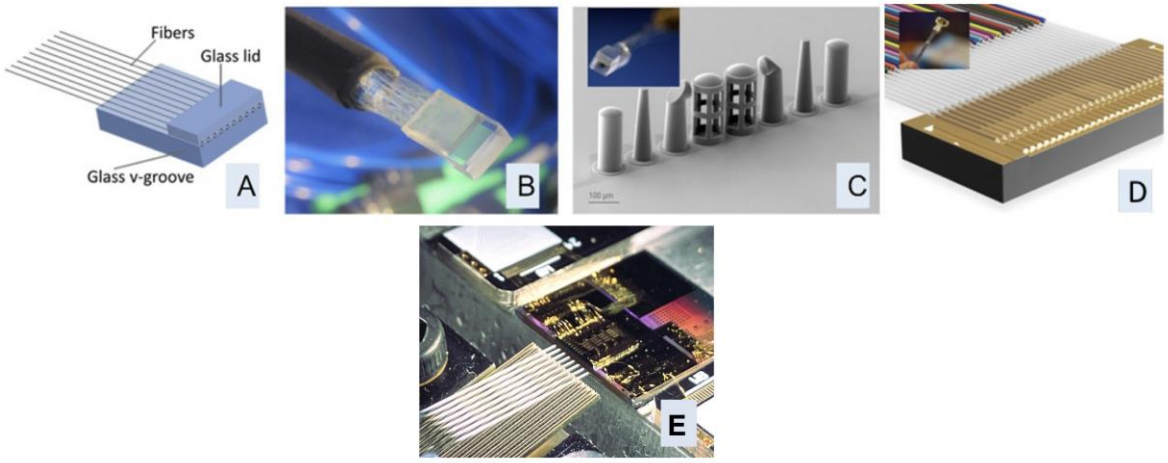


Figure 2. 14 **Examples of fiber arrays A: Schematic, B: Photo. Free-space micro-optical couplers that are printed on a fiber array (PHIX), C: SEM image & photo (Nanoscribe, PHIX), D: Photonic-Plug® (Teramount), E: Microcantilever-based fiber coupling, (MicroAlign); [13]**

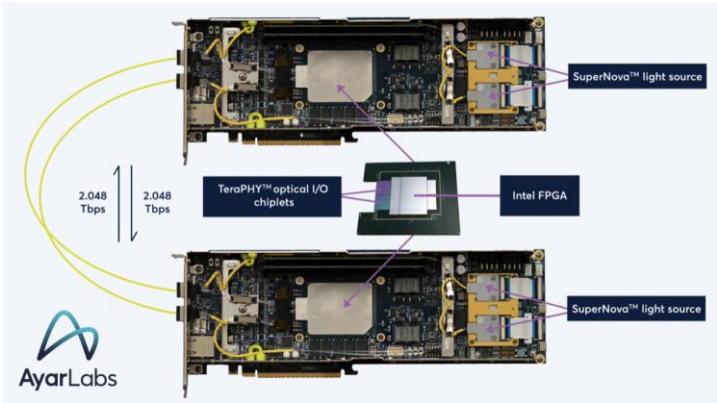


Figure 2. 15 **Ayar Labs showcased a 4 Tbps optically-enabled Intel FPGA design at SC23, which offers 5x current industry bandwidth at 5x lower power and 20x lower latency, all packaged in a common PCIe form factor. (credit: Ayar Labs)**

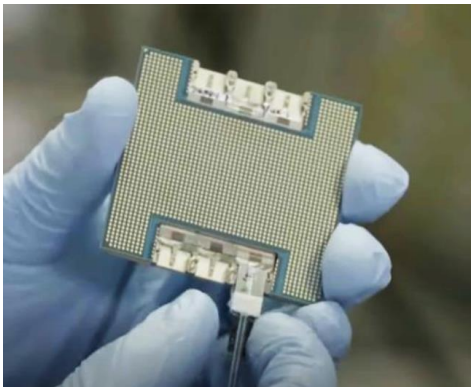


Figure 2. 16 **Co-packaged photonic assembly with six detachable optical interfaces [14]**

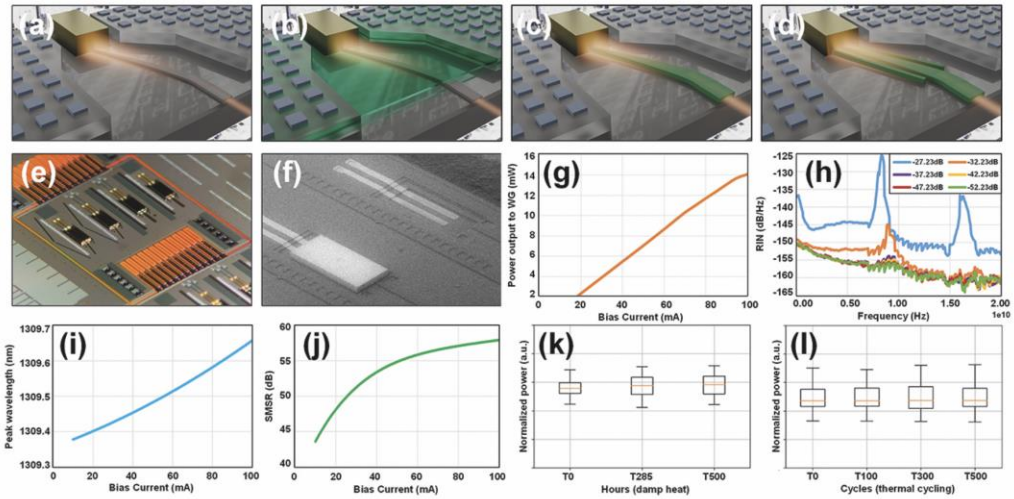


Figure 2. 17 III-V laser integration on the monolithic SiPh platform. (a)-(d) 3D perspective views of various PICs with different SSCs formed on Si or SiN layer. (e)-(f) Optical image and SEM of laser cavities with and without flip-chip-bonded laser. (g)-(j) Light- current curve,

RIN, spectral characterization and SMSR performance. (k)-(l) Wafer-level accelerated reliability results [15]

2.6 MANUFACTURING GAP ANALYSIS (ROADMAP & ONSHORE NEEDS)

This section highlights the most significant gaps identified by the comprehensive gap analysis for future manufacturing of HPC packages and is organized in two categories as listed below.

A. Leading-edge Gaps that Create Opportunities
   - There is currently no high-volume, silicon-based package manufacturing infrastructure in the US.
   - Die-to-Die interconnect pitch scaling roadmaps create new opportunities to address lithographic tool and process gaps for large area patterning.
   - Bond pitch scaling with hybrid bonding and alternate assembly methods require innovations in plasma or other dicing, cleans and metrology steps to achieve high yields and cost-effective volume manufacturing.

B. Supply Chain Resiliency Gaps
   - The biggest gap in the onshore packaging supply chain for high performance computing is the lack of any advanced organic substrate manufacturing infrastructure in the US.
   - Addressing the lack of non-captive, high volume bumping and assembly infrastructure in the US is another key to ensuring supply chain resiliency.

A more detailed view of the key gaps in the leading edge HPC roadmap is shown in **Figure 2.18**.

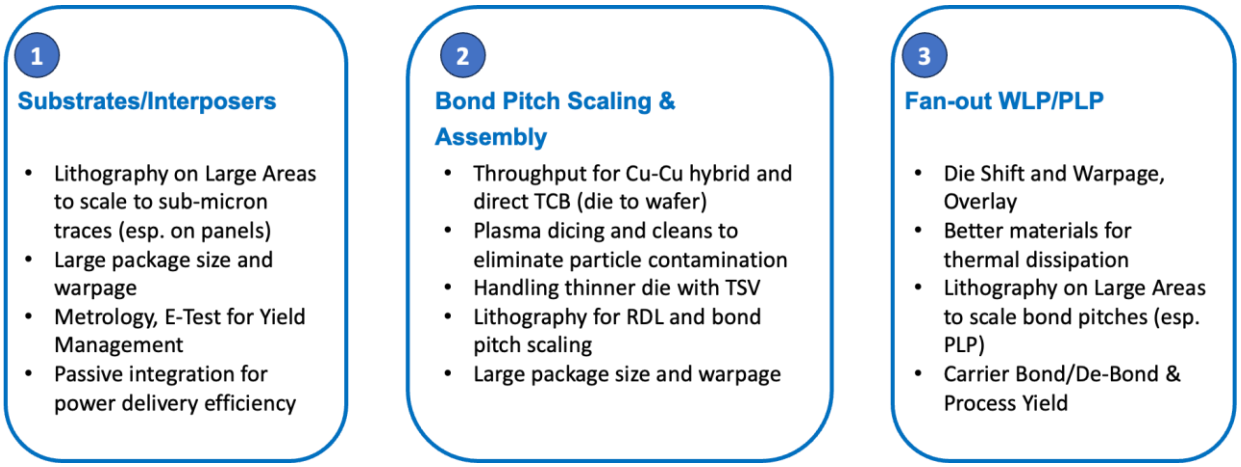


Figure 2. 18 Summary of Gaps and Challenges that create opportunities for New Innovation and Investments in Future HPC Packages.

This technical working group team (TWG1) has conducted a survey of global capabilities in each of the platforms discussed in this section, and a summary of the capabilities with selected examples of companies involved is shown in **Table 2.12**.

Table 2.12 Summary of Global and Onshore Capabilities in HPC Package Platforms Highlighting On-Shore Gaps in Most Platforms

Platform	Technology	On Shore Capability		Off Shore Sources	Global Status (against HIR Targets)
		Captive (IDM/Foundry)	Non-Captive		
Onshore Opportunity 1 Substrates/ Interposers	Organic Substrate	Development Only		Taiwan, S Korea, Japan	**Asia:** 5/5 um – 9/12 um in HVM
	Silicon Interposer	GF, Intel (Bridge), TSMC AZ (?)	In Development – Skywater	TSMC/Others (sub-micron)	**Asia:** Sub-micron on Wafer **US:** Captive Foundries
	Glass Substrate (?)	Not known	In Development - Samtec, 3DGS, Absolics	In Development	
	HD Ceramic	N/A	Kyocera (?)	Japan (RDL in Development)	**Asia:** Mulitple Suppliers **US:** Pilot/LVM (?)
Fan-out Package	Wafer Level	No announced plans	In Development - Skywater	Taiwan, S Korea, China (Wafer FO, FO Emb. Bridge)	**Asia:** TSMC, ASE 2/2 um HVM, others in Dev. **US:** 2/2 um in Dev.
Onshore Opportunity 2	Panel Level	Intel EMIB Embedded Bridge		Taiwan, China	**Asia:** Gap in L/S vs. Wafer FO
Bumping	Copper & Cu/Solder	Development Only	Development/LVM Only	Various, Asia	**Asia:** HVM in many countries **US:** No Volume capability (?)
Assembly	Flip Chip	Intel has site in Costa Rica	Development Only	Taiwan, S Korea, Japan, Asia	**Asia:** Mature HVM **US:** Development/LVM only
Onshore Opportunity 3	Hybrid Bonding & 3D Stacking	Intel New Mexico, TSMC AZ (?), Samsung US (?), GF (?)	In Development - Skywater	Taiwan, S Korea, Japan	**Asia:** TSMC starting 2022 **US:** Intel Foveros Direct in 2023
			KEY MESSAGE: Major On-Shore Gaps in Non-Captive Substrate Mfg/OSAT Assembly		

The key message from this analysis is the fact that there exist major on-shore supply chain gaps in advanced substrates, bumping and in assembly and test infrastructure, and this is the right opportunity for government supported private investments in on-shore manufacturing capability to address supply chain resiliency.

References

[1] He, Lei & Elassaad, Shauki & Shi, Yiyu & Hu, Yu & Yao, Wei. (2011). “System-in-Package: Electrical and Layout Perspectives”, *Foundations and Trends in Electronic Design Automation*. 4. 223-306.

[2] Y. H. Chen et al., “Low Cost Glass Interposer Development”, *IMAPS 2014 Proceedings*, Oct. 13-16, 2014, San Diego, CA USA (ISBN: 978-0-9909028-0-5).

[3] A. A. Bajwa et al., "Heterogeneous Integration at Fine Pitch (≤ 10 μm) Using Thermal Compression Bonding," 2017 IEEE 67th Electronic Components and Technology Conference (ECTC), Orlando, FL, USA, 2017, pp. 1276-1284, doi: 10.1109/ECTC.2017.240.

[4] D. Frye, R. Guino, S. Gupta, M. Sano, K. Sato and K. Iida, "Gold-Gold Interconnects to Copper Pillar using fast Thermal Compression Bonding using Non-conductive paste," 2010 Proceedings 60th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, USA, 2010, pp. 427-430, doi: 10.1109/ECTC.2010.5490938.

[5] K. Sahoo, S. Pal, N. Shakoorzadeh, Y. -T. Yang and S. S. Iyer, "Copper to gold thermal compression bonding in heterogenous wafer-scale systems," 2021 IEEE 71st Electronic Components and Technology Conference (ECTC), San Diego, CA, USA, 2021, pp. 487-493, doi: 10.1109/ECTC32696.2021.00088.

[6] A. Bajwa, T. Palumbo, T. Colosimo, B. Chylak and S. Goh, "Fluxless Bonding Via In-Situ Oxide Reduction," 2022 IEEE 24th Electronics Packaging Technology Conference (EPTC), Singapore, Singapore, 2022, pp. 498-502, doi: 10.1109/EPTC56328.2022.10013208.

[7] Siva Chandra Jangam, A. Bajwa, U. Mogera, P. Ambhore, T. Colosimo, T. Palumbo, D. DeAngelis, B. Chylak and S. S. Iyer, "Fine-Pitch (≤10 μm) Direct Cu-Cu Interconnects using In-situ Formic Acid Vapor Treatment", IEEE 69th Electronic Components and Technology Conference (ECTC), May 28-31, 2019, Las Vegas, NV.

[8] K. Sahoo, H. Ren and S. S. Iyer, "A High Throughput Two-Stage Die-to-Wafer Thermal Compression Bonding Scheme for Heterogeneous Integration," 2023 IEEE 73rd Electronic Components and Technology Conference (ECTC), Orlando, FL, USA, 2023, pp. 362-366, doi: 10.1109/ECTC51909.2023.00067.

[9] Zhong-Jie Hong, Demin Liu, Han-Wen Hu, Chih-I Cho, Ming-Wei Weng, Jui-Han Liu, Kuan-Neng Chen, Investigation of bonding mechanism for low-temperature CuCu bonding with passivation layer, Applied Surface Science, Volume 592, 2022, 153243, ISSN 0169-4332, https://doi.org/10.1016/j.apsusc.2022.153243.

[10] A. K. Panigrahi, S. Bonam, T. Ghosh, S. R. K. Vanjari and S. G. Singh, "Low temperature, low pressure CMOS compatible Cu -Cu thermo-compression bonding with Ti passivation for 3D IC integration," 2015 IEEE 65th Electronic Components and Technology Conference (ECTC), San Diego, CA, USA, 2015, pp. 2205-2210, doi: 10.1109/ECTC.2015.7159909.

[11] https://www.3dincites.com/2023/05/hybrid-bonding-takes-heterogeneous-integration-to-the-next-level/

[12] Advances in Embedded and Fan-Out Wafer Level Packaging Technologies, Edited by Beth Keser and Steffen Kroehnert, *Wiley*, 2019.

[13] *IEEE SA FIBER ATTACH TECHNOLOGY WHITE PAPER 10/2023.*

[14] N. Psaila, S. Nekkanty, D. Shia and P. Tadayon, "Detachable Optical Chiplet Connector for Co-Packaged Photonics," in Journal of Lightwave Technology, vol. 41, no. 19, pp. 6315-6323, 1 Oct.1, 2023, doi: 10.1109/JLT.2023.3285149.

[15] High-performance monolithically integrated edge couplers, *PIC Magazine.Net I Issue I*, 2023.

Chapter 3: Medical/Hybrid Electronics

TWG1: Advanced Packaging & Heterogeneous Integration

Contents

3.1 Overview 2

3.2 Executive Summary: Flexible Hybrid Electronics 3

3.3 Background 5

3.4 Background Summary for the Heterogeneous Integration Roadmap (HIR) 7

3.5 Technical Summary for the Flexible and Hybrid Electronics Technical Workshop Group 8

3.6 Details by Manufacturing Topics Areas 10

3.6.1 Flexible Hybrid Electronics 10

3.6.2 Panel-Level Packaging and Leveraging of Flexible Display Manufacturing Infrastructure (e.g., InnovaFlex Foundry (formerly known as dpiX, LLC)) 14

3.7 Technical Road Summary and Corresponding Manufacturing Gap Analysis 17

3.8 Appendix: Full Process Flows 19

Figures

Figure 3.1: Macro-trends in Flexible and Hybrid Electronics overlaying minimum bump pitch for HIR context 9

Figure 3.2: Overlay of Flexible and Hybrid Electronics trends and electronics segments with technology application verticals 9

Figure 3.3: Flexible Hybrid Electronics process flow based on the NextFlex Program 11

Figure 3.4: Multiple examples of hybrid electronic manufacturing for single and multiple metal layer processes 12

Figure 3.5: List of the NextFlex Roadmaps 13

Figure 3.6: State of the Art in Flexible Hybrid electronics and associated Taxonomy 14

Figure 3.7: Create a prototype flexible RDL/Interposer consisting of up to four (4) metal and four (4) ILD (ex. AD-BCB) layers on top of polyimide. 15

Figure 3.8: Process flow steps 15

Figure 3.9: X-Section 16

Tables

Table 3.1: Improvement Development areas for AE to Provide SiP solutions 8

Table 3.2: Summary of Technical gaps across 1-3 manufacturing approaches 10

3.1 Overview

The initial step after organizing the TWG1 team into sub-groups involved a comprehensive review of key chapters in the IEEE HIR roadmap. The objective was to initiate the development of an advanced packaging manufacturing blueprint. This process involved providing brief chapter summaries, including key points, potential solutions, gaps, and future challenges.
The chapter focuses on the progress in sub-group 2/Medical/Hybrid Electronics.

For hybrid electronics and medical applications, the HIR Roadmap serves as a foundational resource with multiple chapters addressing relevant topics. During the chapter review, several critical gaps were identified:

- Include materials specifics and corresponding tools. In general HIR has limited specifics on materials. The Materials chapter is almost exclusively device-level materials for future generation (2D CNT etc)
    - Observation: The SiP and WLP are dense in manufacturing technology but a notable gap in coverage to the supply chain.
- CHANGE "Flexible Hybrid Electronics" to Hybrid Electronics (HE) – NextFlex has adopted this change
    - Identify material supply chain gaps. Hybrid has been captured in the Ch 21 and 23 in various sections
- Improve timeline for technology evolution and necessary material and/or tool supply chain to meet the gap. Although the WLP chapter includes a timeline, a similar horizon for SiP manufacturing technology was not observed.
- Specific considerations for Hybrid Electronics:
    - Reduce the public HE roadmaps into a common time horizon based on HIR.
    - Identification of the current position of HE within the HIR chapters.
    - Overlay HE to the HIR manufacturing roadmaps
- Consider refining the term "technology;" for example, "manufacturing technology roadmap" to encompass materials, tools, and packaging manufacturing technology (SiP, WLP, HE). And "System Technology" for example to map medical devices, power devices
- Encourage a section on panel-level processing. This is mentioned in passing in WLP and SiP. In this section, I would capture the domestic panel-level manufacturing not just for packaging. Important to acknowledge domestic electronics manufacturing.

This sub-group (TWG1-Subgroup 2) analyzed and summarized the chapters concerning Emerging Materials (15), SIP and Module (21), and Wafer Level Packaging (23) as below:

***Emerging Materials (Chapter 15)*:** The gaps in materials were organized into two-time frames; 2019-2029 and 2029-2043.

- New conductor & joining processes – known, but have not been integrated into HVM
- Warpage for ever thinner layers – solutions known & demonstrated, but not integrated into HVM
- New Materials:
    - Cobalt & Cobalt/Copper (in use to reduce contact & line resistance)
    - 2D materials (e.g., MoTe2) - ongoing research
- Thermal management – reaching limits
    - Diamond being researched as a solution
- Examples of Material Requirements for the next 25 years
- Examples of future materials:

  - 2D - many types of materials (semi-metal, semiconductor, metal, superconductor, insulator)
  - Nano-infused ceramics (graphene in ceramics)
  - AI designed materials

***SiP and Module (Chapter 21)*:** The summary of this chapter review is organized into three sections as discussed below.

- Toolbox Perspective
  - Technology toolbox description:
    - Interconnects (wirebond, flip chip, hybrid bonding, RDL)
    - Encapsulation
    - Architectures (PoP, Embedded, FOWLP/PLP, chiplets, modules, precision assy)
  - Challenges for the toolbox: chip size, I/O magnitude, chip pitch, chip count, max # of domains served
- Application Perspective & Market Needs
  - Power Functionalities (SiC, GaN, fast switching, low losses, thermal mgmt., etc)
  - MEMS Functionalities (wirebond□WLP□ 3D WLP with TSV)
  - Complex IoT devices, Edge Computing (wiring density, thermal, custom-off shelf, multi-domain testing, etc)
  - AI Integration into SiP (□ mobile AI to be mainstream
  - Modules
- Main Challenges from the Application Perspective Towards SiP Adoption
  - Direct app-related challenges: more functionality, non-electronics need co-design (optics, fluids), assy process will change, reliability requirements adapted,
  - Materials: improvements needed,
  - Physics: thermal, empirics/statistics needed, form factors, signal integrity, power increases, verification, EDA co-design, environmental factors etc.

Wafer-Level Packaging (Chapter 23): This chapter presents a good overview of System-in-Package. Additional points are listed below.

- Interested in hybrid integration where differing technologies can be combined with flexible substrate.
- Are currently existing metals used in flat panel manufacturing sufficient for packaging?
- Interested in panel-level packaging and how current flat panel manufacturing can assist.
- Chiplet technology can also be utilized in flat panel manufacturing.
- Need infrastructure additions.

3.2 Executive Summary: Flexible Hybrid Electronics

The Technical Working Group (TWG) for Flexible & Hybrid Electronics for System in Package, Wafer Level Packaging is driven by the advancements in wearable and health monitoring technologies. The flexible electronics manufacturing domain has displayed a diverse range of applications, encompassing asset monitoring on 3D surfaces, communications arrays, soft robotics, and electronics for extreme environmental conditions. In alignment with the NIST- funded Advanced Packaging Roadmap program, our focus is specifically directed towards the
development of wearable and health monitoring technologies.

Scope and Contribution:

The TWG1 sub-group 2 roadmap draws insights from the IEEE Heterogeneous Integration Roadmap (HIR), knowledge pooled from TWG members, and public summaries from influential entities such as The Office of the Secretary of Defense (OSD) ManTech Office funded NextFlex Manufacturing Innovation Institute, Army Research Laboratory, Air Force Research Laboratory, SEMI, FlexTech Alliance Group, Nano-Bio Manufacturing Consortium (NBMC), UCLA Center for Heterogeneous Integration Performance Scaling (CHIPS), and InnovaFlex Foundry.

Wearable and Health Monitoring Technologies: Wearable and health monitoring technologies have witnessed expansive growth in consumer electronics, empowered by wireless, sensor, and battery technologies. These innovations aim to monitor physiological, cognitive, biological, and situational aspects, paving the way for enhanced medical diagnosis, safety, injury prevention, and performance augmentation capabilities. Wearable devices, equipped with sensors, cover vital signs, cognitive signatures, and access to blood or fluid testing, thus providing a comprehensive health monitoring suite. The technology extends to clinical monitoring systems, including digital x-ray imagers, MRI, Computed Tomography, and emerging devices for clinical analysis.

Roadmap Development:

The roadmap is a culmination of efforts from over 200 industry, academic, and government partners, representing programs and companies such as NextFlex Manufacturing Innovation Institute, InnovaFlex Foundry, UCLA CHIPS, and SEMI FlexTech Alliance. These entities collectively strive to foster a robust U.S. industry network in flexible and hybrid electronics, contributing to a manufacturing ecosystem that offers strategic advantages to the Department of Defense (DOD) and U.S. industry in multibillion-dollar markets.

Economic Impact and Job Creation:

To date, the consensus within the industry strongly suggests that U.S. flexible hybrid electronics technology and manufacturing efforts have the potential to create a substantial number of jobs across a spectrum of businesses, from small enterprises to Fortune-500 companies. This job creation is anticipated to span the entire product supply chain, from the production of raw materials to the retail sales of innovative devices. The flexible electronics sector presents a unique opportunity for the next wave of high-tech manufacturing job creation. Unlike early silicon CMOS manufacturing, which saw the migration of jobs to foreign countries due to low-profit margins on mature Si CMOS technology, flexible electronics offers the potential for novel technologies with higher profit margins. By combining traditional U.S. strengths in plate-to-plate semiconductor manufacturing with roll-to-roll printing, innovative and cost-effective fabrication techniques can be realized, enabling the entry of mid- size companies into manufacturing, and thereby expanding job opportunities within the U.S. The public-private partnerships that are enabling innovation in flexible hybrid electronics encompass the numerous industry, academic, and government participants are advancing the manufacturing goals to realize flexible hybrid electronics products. These partnerships are crucial in harnessing industry expertise and steering basic research towards establishing a new U.S.-based manufacturing paradigm.

Key Contributors:
NextFlex Manufacturing Innovation Institute: Funded by the Office of the Secretary for Defense,

NextFlex aims to grow U.S. competitiveness in Flexible and Additive Hybrid Electronics Manufacturing, supporting both defense and commercial applications.
www.nextflex.us
InnovaFlex Foundry: Formerly known as dpiX, InnovaFlex is a nontraditional semiconductor design and manufacturer with capabilities in creating electronics on glass and flexible substrates, contributing to innovative solutions in military, medical, industrial, and security imaging.
InnovaFlex Foundry
UCLA CHIPS: The Lead Center for Heterogeneous Integration Performance Scaling interprets and implements Moore's Law for heterogeneous systems, developing architectures, methodologies, designs, components, materials, and manufacturable integration schemes.
UCLA CHIPS
SEMI FlexTech Alliance: As a strategic Association Partner, SEMI FlexTech Alliance fosters collaboration between industry, academia, government, and research organizations to advance displays and flexible, printed electronics from R&D to commercialization, contributing to a world-class manufacturing capability.

In conclusion, the TWG roadmap serves as a comprehensive guide, leveraging the collective expertise of industry leaders and researchers to propel the development and application of flexible and hybrid electronics, particularly in the realm of wearable and health monitoring technologies.

CONTRIBUTIONS FROM:

Eric Forsythe (US Army Research Laboratory)
Robert Rodriquez (InnovaFlex Foundry formerly named dpiX)
Gity Samadi (SEMI)
Subramanian Iyer (CHIPS UCLA)
Art Wall (NextFlex)
Executive Agent Printed Circuit Boards and Interconnects (Navy Crane)

3.3 Background

The Technical Working Group for Flexible & Hybrid Electronics for System in Package, Wafer Level Packaging is motivated by wearable and health monitoring technologies. Flexible & Hybrid Electronics manufacturing has demonstrated a broader technology application space that includes, asset monitoring such as electronics integrated onto large 3D-surfaces, communications arrays and associated electronics, soft robotics, electronics for extreme environmental conditions, to name a few. For the purposes of the NIST funded Advanced packaging roadmap program, the Flexible & Hybrid Electronics contribution will focus on wearable and health monitoring technologies. The TWG Flexible & Hybrid Electronics for System in Package, Wafer Level Packaging roadmap will summarize IEEE Heterogeneous Integration Roadmap (HIR), the knowledge from the Technical Working Group members, public summaries from The Office of the Secretary of Defense (OSD) ManTech Office funded NextFlex Manufacturing Innovation Institute, one the Manufacturing USA programs, and Army Research Laboratory and Air Force Research Laboratory SEMI, FlexTech Alliance Group and Nano-Bio Manufacturing Consortium (NBMC), UCLA Center for Heterogeneous Integration Performance Scaling (CHIPS), and InnovaFlex Foundry, Colorado Springs, CO (formerly known as dpiX).

Wearable and health monitoring technologies have realized prolific expansion in consumer electronics markets. Technologies to monitor the physiological cognitive, biological, and situational states of human status are enabled by wireless, sensors and battery technologies. These electronic technologies combine to enable wearable technologies with the objective of providing new capabilities, such as medical diagnosis and therapy, increased safety, injury prevention and performance augmentation capabilities. Wearable electronic devices have integrated sensors to monitor physiological signatures including vital signs such as temperature, heart rate, respiration rate, blood oxygenation, blood pressure and brain activity. Wearable devices are demonstrating value in monitoring cognitive signatures that include, electrophysiological (EEG/EOG/EMG), ultrasound, pupillometry, and other measures of brain activation. Wearable devices that combine access to blood or fluid testing can determine biological signatures. The technology suite for wearable electronic devices encompasses situational and environmental parameters such as external temperature, humidity, noise levels, presence of electrical and electromagnetic hazards, auditory hazards, collision and crush hazards, toxic gases, chemical and biological hazards. In addition, health monitoring devices extend to clinical monitoring systems such as digital x-ray imagers, MRI, Computed Tomography, and many other emerging devices for clinical analysis.

The roadmap was generated from the following programs and companies that represent more than two hundred industry, academic, and government partners.

NextFlex Manufacturing Innovation Institute is funded by the Office the Secretary for Defense (Research and Engineering) Manufacturing Program with the vision to Grow a strong U.S. industry network rallying around electronics integration, leading to a U.S. manufacturing ecosystem that delivers FHE products that give strategic advantage to manufacturing ecosystem that delivers FHE products that give strategic advantage to DOD and U.S. industry in multibillion dollar markets. Network includes chipmakers, DOD, and U.S. industry in multibillion dollar markets. Network includes chipmakers, aerospace and healthcare companies, material and equipment makers, electronics assembly aerospace and healthcare companies, material and equipment makers, electronics assembly and printing companies, and advanced research universities and printing companies, and advanced research universities. The NextFlex program has the mission to Grow U.S. Competitiveness in Flexible and Additive Hybrid Electronics Manufacturing and Design, Prototype, and Manufacture Technologies for the Warfighter and commercial applications.

www.nextflex.us

InnovaFlex Foundry (formerly known as dpiX) is a nontraditional semiconductor, design and manufacturer that has capabilities to create a variety of electronics on both glass and flexible substrates. InnovaFlex provides the foundation for some of today's most innovative solutions in the military, medical, industrial, and security imaging businesses.

https://innovaflexusa.com/

The UCLA Lead Center for Heterogeneous Integration Performance Scaling (CHIPS) has the mission to interpret and implement Moore’s Law to include all aspects of heterogeneous systems and develop architectures, methodologies, designs, components, materials, and manufacturable integration schemes, which will shrink system footprint and improve power and performance.

https://www.chips.ucla.edu/

SEMI FlexTech Alliance, A strategic Association Partner has evolved from prior non-profit consortium programs starting 1997 that contributed to the technical outcomes from the programs described above. FlexTech, a SEMI Technology Community, is devoted to fostering the growth, profitability and success of the electronic display and the flexible, printed electronics and its supply chain. FlexTech offers expanded collaboration between and among industry, academia, government, and research organizations for advancing displays and flexible, printed electronics from R&D to commercialization. To this end, SEMI-FlexTech, based in San Jose, Calif., will help foster development of the supply chain required to support a world-class, manufacturing capability for displays and flexible, printed electronics.

https://www.semi.org/en/communities/flextech

3.4 Background Summary for the Heterogeneous Integration Roadmap (HIR)

Flexible & Hybrid Electronics for System in Package, Wafer Level Packaging Technical working group chapter expands upon portions of Chapter 16 and Section 10 Chapter 8. The following summaries from the chapters are a starting point for the details that follow.: Emerging Research Devices: the IEEE HIR chapter 16 starts from a baseline in 2018. In 2018, commercial practices were essentially three major areas of printed electronics in commercial practice in 2018. First is the oldest usage of polymer thick film conductors principally used for interconnection and things like membrane touch switches. This is complemented by thin film processes that make use of lithography patterning to create everything from thin-film transistors to interconnects. Both approaches converge at touch display and active surface applications that make use of a variety of techniques focused on patterning conductive materials such as Indium Tin Oxide (ITO) or silver, carbon, or copper nanowires. The third major area of flexible hybrid electronics is where elements of each of these techniques are blended together with additive processing and packaging methods to integrate silicon ICs into systems (see section 10 Additive Manufacturing of 2021 HIR Chapter 8). Further HIR Chapter 8 Section 9 Board Assembly process summarize the process steps for printed circuit board and assembly. The TWG will provide a step-by-step summary that provides more detail to the HIR Chapter 8. As a note, significant manufacturing advances are underway in Asia and US to decrease printed circuit board pitches such as substrate-like PCB manufacturing and ultrahigh density interconnects (UHDI) manufacturing technologies to enable fine pitch interconnects for packages to boards. These next generation manufacturing advances for PCB and PCB-A will not be discussed. See for example IPC D-33AP standards working group.

Table 1 below is copied from HIR Chapter 8 section 10.

Table 3.1: Improvement Development areas for AE to Provide SiP solutions

Table 1: Important developmental areas for AE to provide SiP solutions

Development Area	Current Best State	AE Approach for Current Best	Desired State (Depends on use-case)	Developmental Challenges and Suggested Research Areas (Depends on use-case)
Printing Attributes				
Line Width	>40 µm	PRJ, IJ	≤40 µm	Making robust to all print conditions and geometries. Larger line width approaches (DW) brought to inkjet resolution. Higher resolution on inkjet.
Space Width	>100 µm	PRJ, IJ	≤150 µm	Making robust to all print conditions and geometries. DW not at inkjet levels.
Trace Conductivity	12E-8 Ωm	DW, Aerosol Jet	≤10E-8 Ωm	Making robust to all print conditions and geometries and at above width and pitch. PRJ and IJ + PBF need improvements.
Build Speed, Parts per Build	>15 mm/hr, multiple parts	IJ + PBF	Maximize for optimal utility	Improvements to build speed and number of parts/build generally difficult for PRJ and FDM + DW, but likely necessary
Substrate Attributes				
Dielectric Strength	~10 kV/mm	FDM	>15 kV/mm	High dielectric strength materials available, incorporate into AE approaches
HDT	189 °C	FDM (PPSF)	>220 °C	High temperature polymer available, needs development for AE.
Tensile Strength	70 MPa	FDM (ULTEM)	~70 MPa	Highly rigid polymers available for AE
Additional Process Integration				
Component Attachment	Amenable to P&P	FDM, SLA, PRJ	Optimized with P&P	Processes incorporating P&P not optimized: speed, interconnects, in-situ testing, resumption of printing processes, etc.
Print Pausing/Resume	Amenable to PP/R	SLA, FDM	Optimized with PP/R	Processes incorporating P/R not optimized: system integration, workflow optimization, interface mechanical integrity, etc.

https://eps.ieee.org/technology/heterogeneous-integration-roadmap.html

https://eps.ieee.org/technology/heterogeneous-integration-roadmap/2021-edition.html

3.5 Technical Summary for the Flexible and Hybrid Electronics Technical Workshop Group

The technical working group analysis encompassed several complementary manufacturing approaches to enable the next generation of wearable and health monitoring technology applications. The flexible and hybrid electronics manufacturing enables manufacturing processes that can incorporate novel materials and flexible substrates, integrate commercial devices and passives, and combine traditional electronic device manufacturing, such as lithography and pick-and-place to achieve unique technology attributes for the next-generation wearable and health monitoring technologies. The flexible and hybrid electronics manufacturing along with substrate-like PCB and Ultrahigh density interconnects are contributing to the industry convergence of print circuit board manufacturing technologies and advanced packages to meet increasing technology requirements. Fig 3.1 and Fig 3.2 below illustrate the industry manufacturing convergence. The following details will be discussed in context with this convergence.

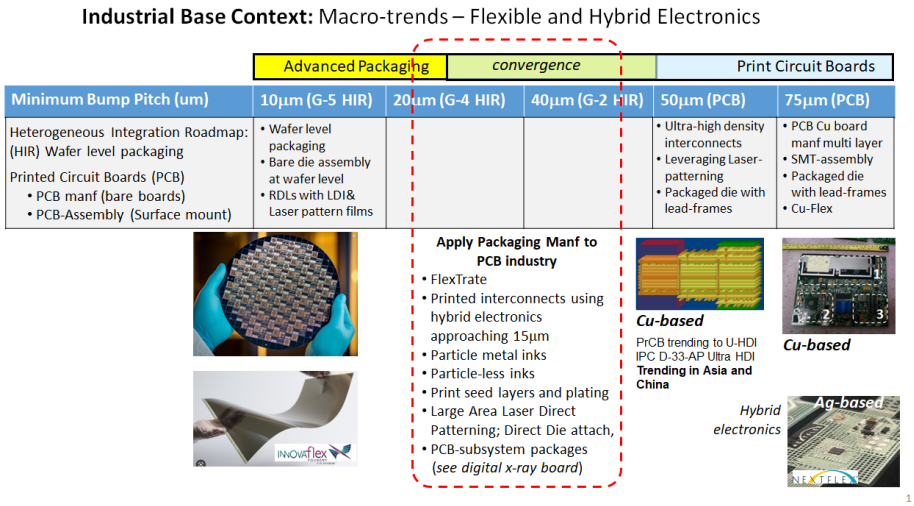


Figure 3.1: Macro-trends in Flexible and Hybrid Electronics overlaying minimum bump pitch for HIR context

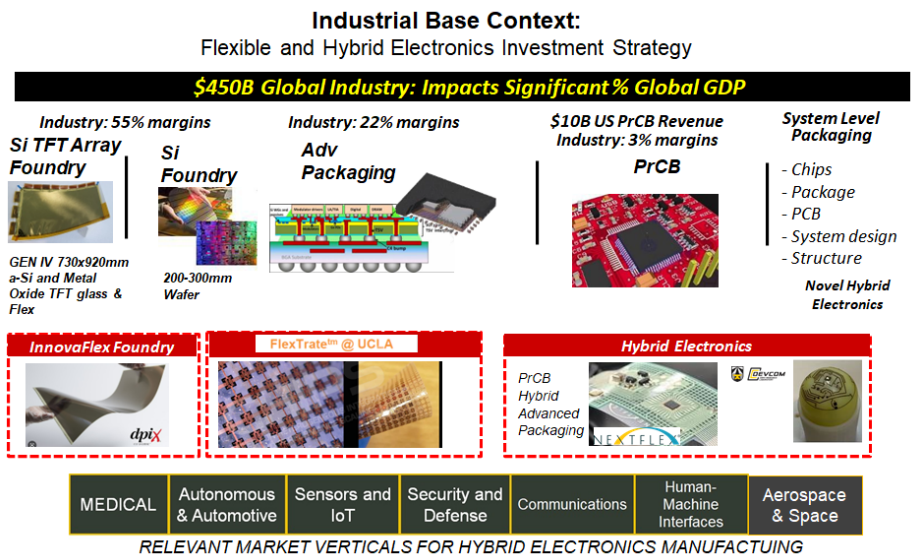


Figure 3.2: Overlay of Flexible and Hybrid Electronics trends and electronics segments with technology application verticals

The following technical section will expand upon the general manufacturing topic areas:

1. Flexible Hybrid electronics (NextFlex MII and SEMI FlexTech Alliance Consortium)
2. Panel-level packaging and leveraging of flexible display manufacturing infrastructure (e.g., innovaFlex Foundry (formerly dpiX))
3. Flexible fanout wafer level packaging – emerging approaches (Flextrate®, UCLA CHIPS)
4. The detailed excel tables include Printed Circuit Board process flows.

Technical gaps across 1-3 manufacturing approaches are summarized in Table 2 below where the (3) manufacturing technologies are delineated by three colors. The flexible and hybrid electronics industry has a common gap of availability of Known Good Die (KGD) in the last ROW.

Table 3.2: Summary of Technical gaps across 1-3 manufacturing approaches

Gap (3) flexible and hybrid electronics	Roadmap Solution needed
1. Lack of robust onshore supply chain for hybrid electronic critical materials, metal inks, die attach.	Hybrid electronics is a maturing manufacturing segment. As such the supply chain remains fragmented while the technology demand is developed. Many of the critical materials are sourced both domestically with increased competitive offshore
1. Reliability requirements for medical and wearable electronics are different than consumer electronics. However, these requirements are far less than national security and defense requirements.	Significant development in understanding reliability and associated manufacturing gaps have been demonstrated from the flexible hybrid electronics ecosystem. SEMI-FlexTech is leading an industry standards working group in FHE.
1. Flexible hybrid electronics throughput must be increased for medium-volume products	Reliability. Materials and tools supply chains must be developed for parallel processing and increased automation for full product throughput
2. Large area panel level processing domestically leverages commercial x-ray imager 2-metal layer thin film transistor platforms at GEN 4.5. Expand domestic capability for multilayer (8 layers minimum) and associated GEN 4.5 die and passive handling.	The panel level processing has demonstrated fine-pitch achievable through lithography-based processing. Traditional multi-layer processing is based on laminating layers and thru-vias. Technical development on multilayer must be developed to identify scalable multilayer processing with end gap compatible with die and passive assembly.
2. Domestic panel manufacturing is based on limited material sets available through vacuum deposition. Expand the materials sets to include copper for traditional die assembly solder approaches	Identify GEN 4.5 processing for copper plating leveraging the lithography-based platform
3. Increasing throughput and yield for the Flextrate process	The wafer level fan processes are adopting traditional wafer level processing that has significant automation capability to scale
3. Reliability requirements for medical and wearable electronics are different than consumer electronics. However, these requirements are far less than national security and defense requirements.	Larger volume testing is required to understand underlying reliability properties. Through these large-scale studies, modifications in process flows will be identified to further enhance reliability. Engaging in the SEMI-FlexTech FHE standards can provide common standards for the community.
Lack of availability of bare Known Good Die (KGD)	Expand and further develop the bare-die marketplace, expand the existing bare-die handling tools and automation, including long-term storage capability

3.6 Details by Manufacturing Topics Areas

3.6.1 Flexible Hybrid Electronics

Hybrid electronics manufacturing and a sub-set of manufacturing capabilities to integrate electronics onto flexible substrates can realize the application challenges by providing highly integrated, unobtrusive,

lightweight, conformable, and low-cost system solutions without sacrificing device functionality and performance. For example,ninety-nine clinical monitoring devices may require high volume at low cost to achieve disposable requirements. Smart sensors and wireless electronics backbones are being demonstrated with hybrid electronics manufacturing. A critical differentiator for hybrid electronics as compared to traditional printed circuit board manufacturing is the integration of bare die and directly fabricated passives to achieve unique low-profile flexible form factors. Hybrid electronics manufacturing is realizing the convergence of circuit board manufacturing and advanced packaging that integrates bare die. As such, the NSIT-funded roadmap chapter for technical working group sub-group 3; hybrid electronics manufacturing will focus on the roadmap for hybrid electronics and emphasize the convergence with advanced packaging heterogeneous integration. Fig 3.3 is a system-level diagram for Hybrid electronics, courtesy of The NextFlex Manufacturing Innovation Institute 2022.

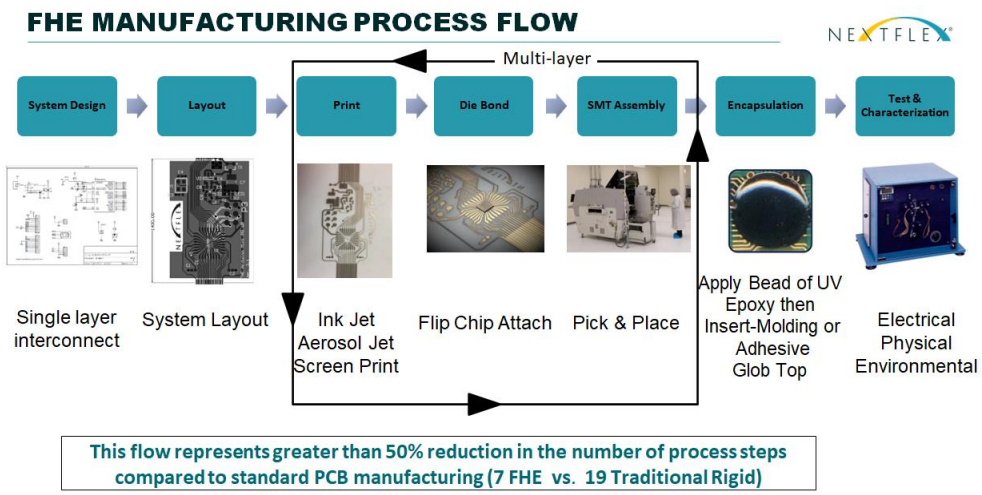


Figure 3.3: Flexible Hybrid Electronics process flow based on the NextFlex Program

Fig. 3.4 below highlights several examples of hybrid electronics manufacturing for the representative "Print" step. The figure below shows examples of a single metal layer or one metal print step. Two-metal layers where a dielectric material is printed between the metal cross overs. Two-metal layer process 2-side figure where a signal metal layer is printed on front and back side then through hole via is drilled and filled to connect the front and back side circuits. Finally, multi-layer processes have been demonstrated by the Boeing Corp that involved printing multiple substrates, including through hole vias, and then laminating multilayers. This approach closely mirrors traditional printed circuit board manufacturing.

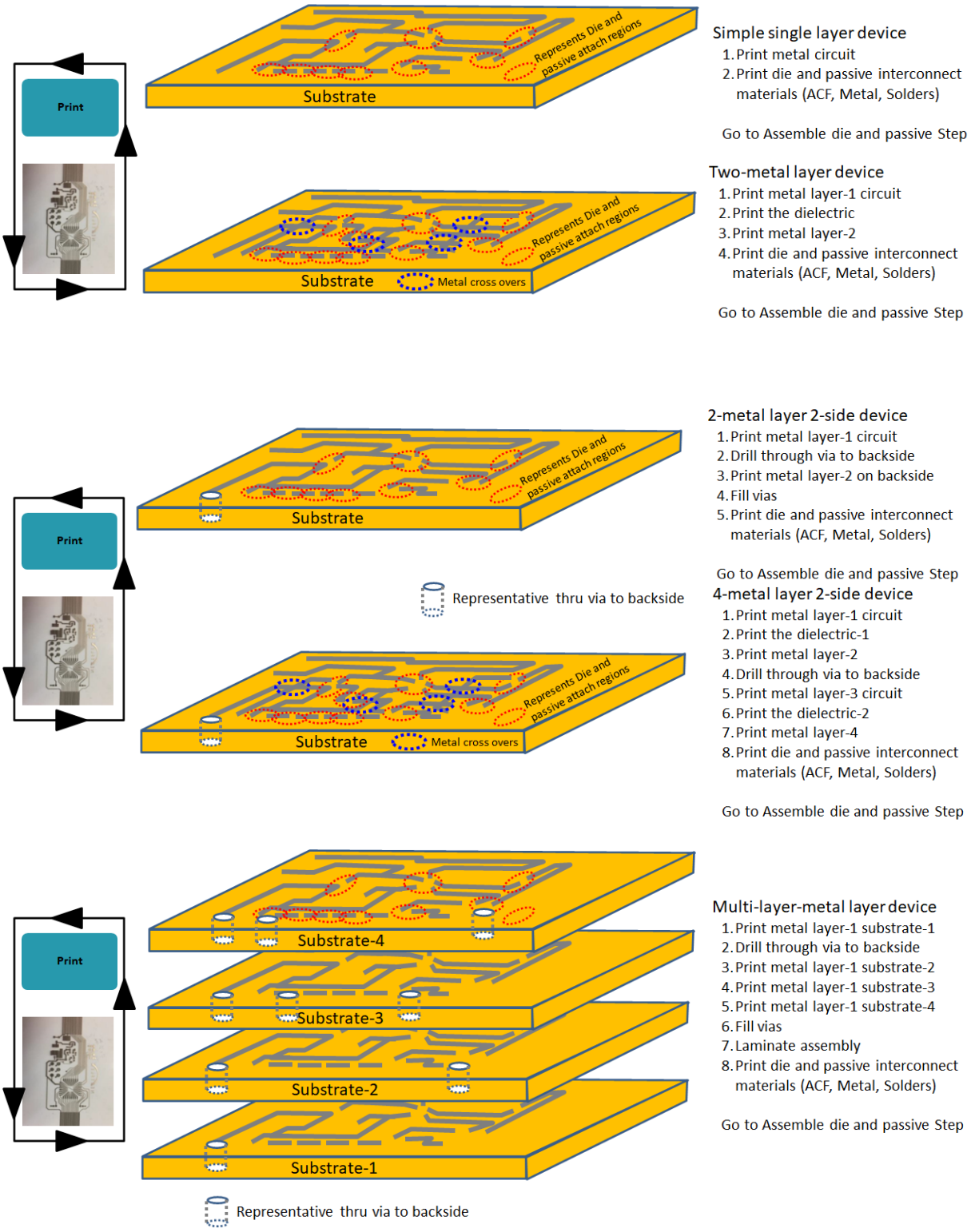


Figure 3.4: Multiple examples of hybrid electronic manufacturing for single and multiple metal layer processes

Fig 3.4 Multiple examples of hybrid electronic manufacturing for single and multiple metal layer processes

The NextFlex program with 153 industry and academic members and subject matter experts from more than seventeen government agencies have developed and evolved the following hybrid electronics roadmaps.
The following is a sample of the public-facing roadmaps from the Nextflex program. (www.nextflex.us)

Fig. 3.5 is the NextFlex roadmaps where human monitoring technology platform demonstrations and device integration and packaging are the most relevant to the TWG 1 report.

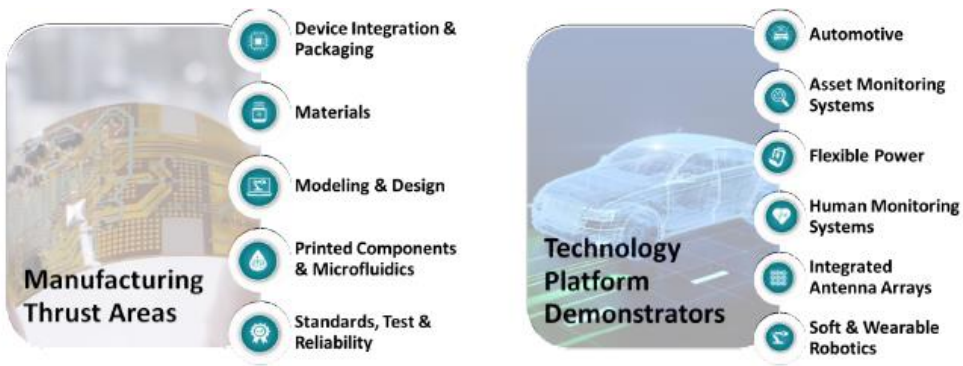
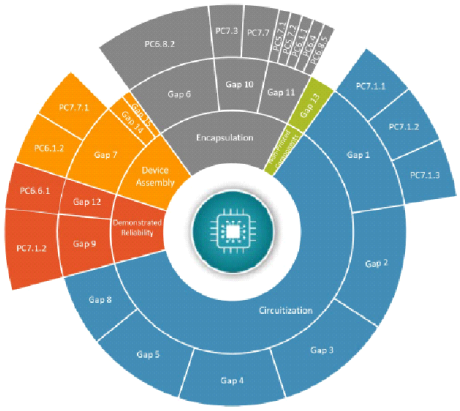


Figure 3.5: List of the NextFlex Roadmaps

The NextFlex manufacturing program then makes investments to address the manufacturing gaps through project calls, leveraging the prototype line in San Jose, CA and supporting the industrial base internal development. The taxonomy for the manufacturing gaps follows the taxonomy below. Fig. 3.6 are the state-of-the-art specifications demonstrated to date and roadmap taxonomies. The flexible hybrid electronics community is advancing the State of the art through project calls to meet the technology gaps identified in the "technology platform demonstration" roadmaps.

STATE OF THE ART

Component / Element	SOTA Specs
Circuit Layers	8
Via Diameter	100-250 µm
Dielectric Thickness	≥25 µm
Bend Radius	>6x thickness
Sheet-to-Sheet Lines & Spaces	50-200 µm
Roll-to-Roll Lines & Spaces	250 µm
Printed conductors	3-20x bulk resistivity
Components	SMTs with solder attach
Printed Resistors	±20% tolerance
Flip-Chip Attach to Flex	100 µm pitch
Die Size	<5 mm^2
Die Thickness	<250 µm
Die I/Os	<100
Pad Area	>75µm sq
Pitch	>150 µm



Roadmap taxonomies (inner ring), identified technical gaps (middle ring), and recent NextFlex-funded projects that align with solving a technical gap (outer ring).

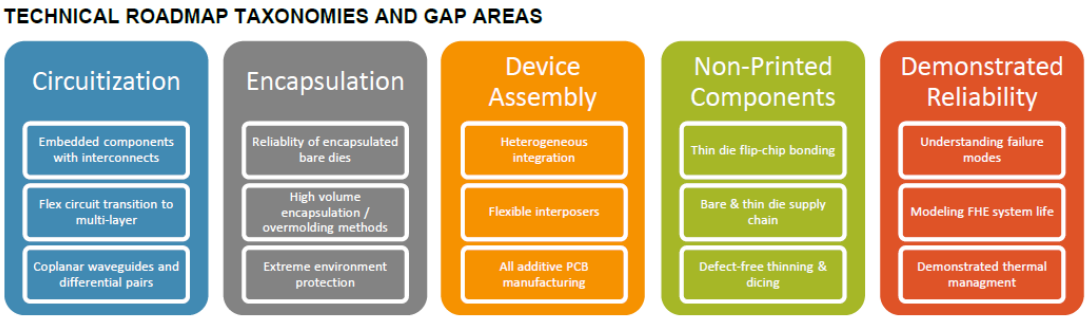


Figure 3.6: State of the Art in Flexible Hybrid electronics and associated Taxonomy

3.6.2 Panel-Level Packaging and Leveraging of Flexible Display Manufacturing Infrastructure (e.g., InnovaFlex Foundry (formerly known as dpiX, LLC))

The second approach to realizing flexible advanced packaging manufacturing is the adoption of flexible thin-film transistor array manufacturing pioneered by the US Army's Flexible Display Center at Arizona State University and the Asian display manufacturing industries. The essential aspect of manufacturing is a bonded flexible substrate and mechanical release process. Metal, semiconductor, and dielectric layers are manufactured using traditional deposition, lithography, and etch processes. The ASU FDC and Innovaflex Foundry (formerly known as dpiX, LLC) adopted the flexible display manufacturing process to demonstrate World's first flexible digital x-ray imagers in partnership with the Xerox Palo Alto Research Center (PARC) funded through the OSD ManTech office, ARL, and Defense Threat Reduction Agency (DTRA).

Innovaflex Foundry's (formerly known as dpiX, LLC) approach for a very high density interconnect (VHDI) is to develop and implement an innovative prototype manufacturing process for substrate, interposer, and redistribution layer (RDL) electronic device packaging products, employing polymer-based materials and VHDI packages with feature sizes below 25 µm (as small as 5 µm). Through-hole via capability would need to be acquired and process development would still need to be done in order to be incorporated into Innovaflex's polymer-based RDL.

Innovaflex Foundry's (formerly known as dpiX, LLC) proposed flexible RDL that could also be applied to its glass substrate for applications where a more rigid structure is desired.

Figure 3.7 & 3.8 highlight Innovaflex Foundry's (formerly known as dpiX, LLC) adoption of the flexible x-ray imager manufacturing processes to flexible, large-area interposers. The figures highlight the manufacturing process flow under development.

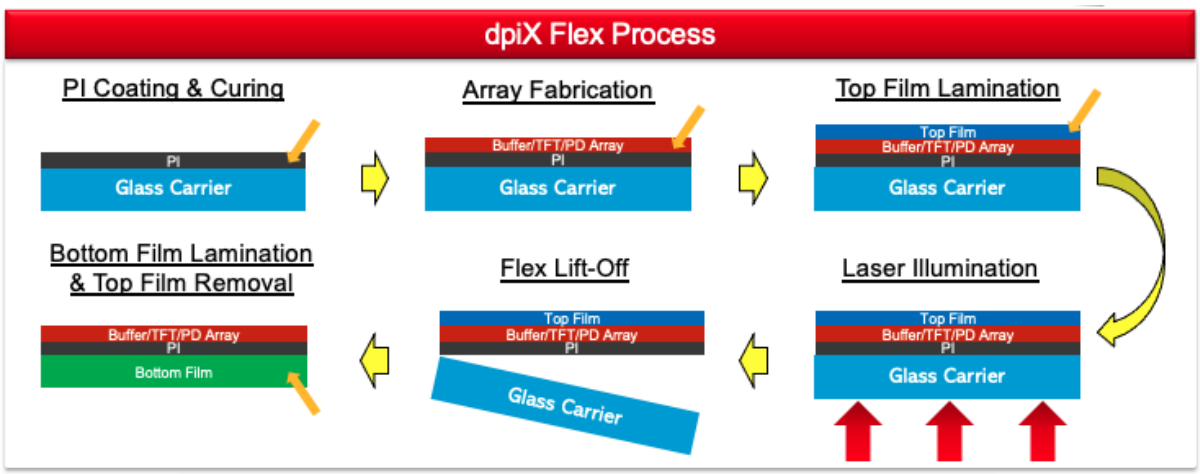


Figure 3.7: Create a prototype flexible RDL/Interposer consisting of up to four (4) metal and four (4) ILD (ex. AD-BCB) layers on top of polyimide.

Project Scope

- Glass Carrier: GEN4.5, 730mm x 920mm x 0.7mm
- Substrate: PI, ~14um final cure
- Bottom Laminate: 100um PET plus ~7Gf/in Glue, Removeable
- Optional Top Laminate: 50um PET plus ~4Gf/in Glue, Removeable
- Samples Requested: TBD, Estimated >500 samples per plate
- Sample Size: ~20mm x ~30mm
- Architecture: 8 Layers (tbc)
- Masking Layers: 8

Process flow steps are highlighted in figure 3.8

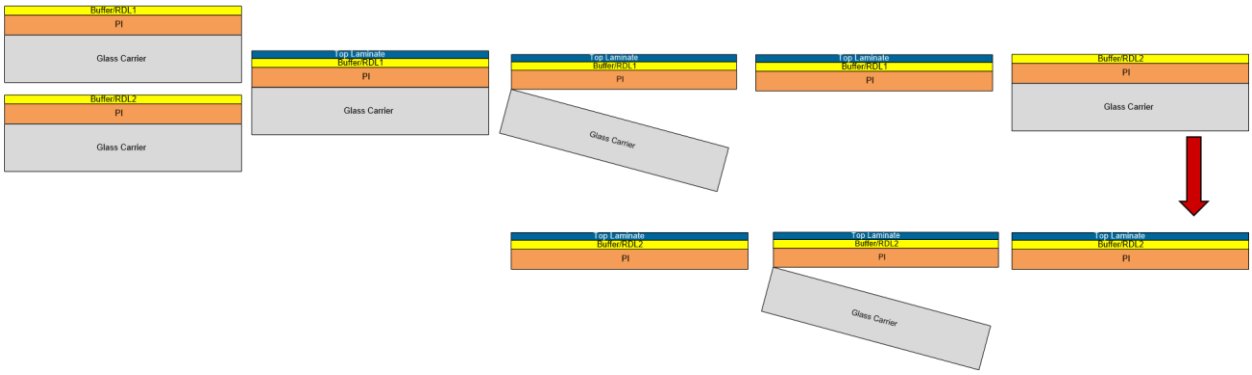


Figure 3.8: Process flow steps

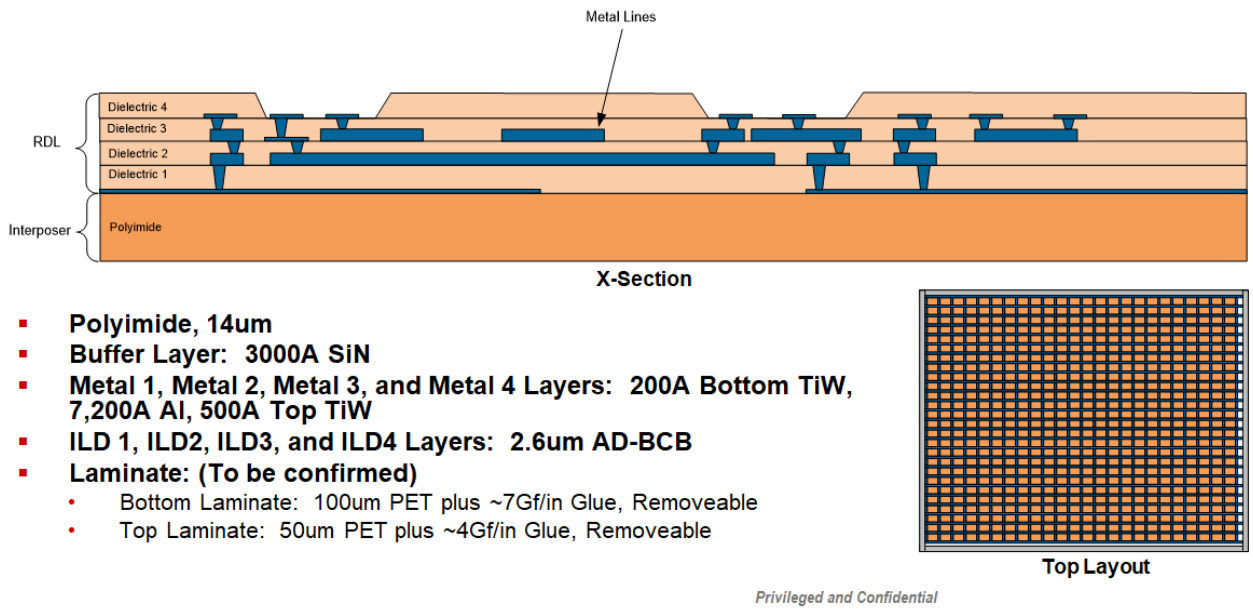


Figure 3.9: X-Section

3.7 Technical Road Summary and Corresponding Manufacturing Gap Analysis

FHE manufacturing technology has roadmaps that realize wearable devices with commercial standards. Military requirements such as durability and environmental operating conditions place larger demands on manufacturing solutions.

The domestic manufacturing industrial base has identified specific system solutions that motivate gap analysis through 2028. Examples include:

- Electrophysiological sensing for cognitive monitoring under high mobility/
- Ultra-small footprint non-invasive sensors for performance monitoring
- Disposable vital sign monitoring with environmentally sustainable materials
- In-vivo biodegradable sensors
- Flexible medical imaging, low-cost impedance sensor arrays
- Biomarker sensors
- E-textiles, soft robotics, and integrated improved power

To realize these wearable product demands, the three technical manufacturing areas need to improve on the following areas through 2028:

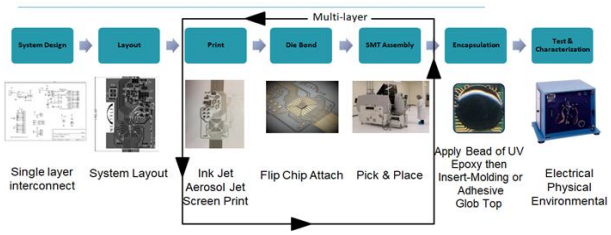
- End-use device level performance for electrophysiological sensing and integrated vital sign monitoring
- Sustainable material sets realize environmentally responsible disposable solutions such as human, biocompatible substrates, metal circuit traces, and interconnects
- Improved signal performance for on-body communications into wireless environments
    - This will require interconnect pitches less than 10μm with reliable die attach metals
- Increased substrate flexibility with ultrathin electronics and sensor devices.
    - This will require flexible or high reliability die interconnects, circuit traces with compliance
- Improve portable power; longer charge, high density and flexible form factors
- Meet commercial reliability standards, thermal cycling, vibration, bend, stretch.
    - SEMI-FlexTech has initiated three standards working groups to define the requirements, manufacturing standards, and tech methodologies to meet applications such as wearables. The NextFlex community has developed large datasets interconnecting reliability, manufacturing processes, and test methodologies. The technical gap is to aggregate this information and develop standards agreed upon by the community.
- Signal Noise for wearable sensors
    - Gap is high conductivity circuit traces to reduce resistance noise such as 1/f noise
    - Improve amplifier devices at the sensor node leverage wafer level fan-out such as FlexTrate (UCLA CHIPs).
    - Integration of Flextrate WLFO technology to integrate system level packaging
    - Optimized sensor signal algorithms
- Large area impedance areas for ultrasound measures
    - FlexTrate integrated array sensors
    - Large area thin film transistor arrays on flexible substrates. Commercially available through InnovaFlex Foundry. However, increased flexibility through thinner substrates will require improved array panel manufacturing
    - Integration of sensors with read-out TFT arrays
- Low-cost
    - Flexible and hybrid electronics manufacturing approaches, flexible hybrid electronics, FlexTrate, and InnovaFlex Foundry all offer the potential for lower-cost solutions.
    - Lower cost is directly linked to commercial volume

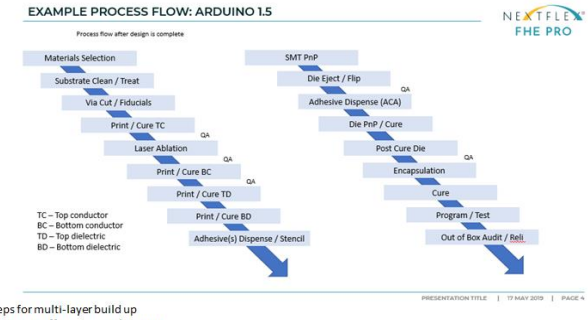
- These manufacturing approaches have fewer manufacturing steps as compared to current manufacturing approaches such as printed circuit boards that inherently could lead to lower costs
- Lower materials that meet commercial requirements but can be processed in digital manufacturing
- Digital manufacturing inherent in FHE improvements in the digital design and digital twin tools to fully realize lower non-requiring engineering (NRE) costs to manufacturing multiple technologies, i.e., no mask sets or tooling.

3.8 Appendix: Full Process Flows

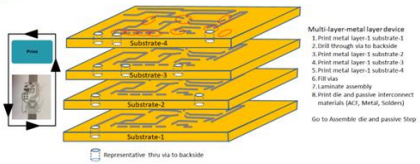
Flexible Hybrid Electronics (NextFlex)

A				flexible Hybrid electronics				Typ. L/S (um)
								< 1um
					Equipment			
Step #	Process	Test	Alternate steps	Purpose	Equipment Type	Equipment maker (US/Overseas)	Alter Equipment maker (US/Overseas)	Gap: Y/N
1	clean substrate			substrate cleaning				N
2	substrate bond to carrier?			handling				N
3	plasma treat			treats and cleans carbon				N
4	surface treatment for adhesion			improves metal and dielectric adhesion				N
	cut vias and fiducials							
5	print ground plane		or Cu plate	creates gournd plane	print inks	Komori screen printer	optomec or nscrypt	N
6	cure			formulates inks				N
7	print first full circuit metal layers	Test conductivelty, digital inspect	print Palladium seed layer	metal circuit traces	print inks	Komori screen printer	optomec or nscrypt	Y
7a			Cu electroless plating or electroplating					
8	cure			formulates inks				N
9	print interlayer dielectrics (cross overs)			allows metal cross-overs	print inks	Komori screen printer	optomec or nscrypt	Y
10	cure			formulates inks				N
11	print second metal layer	Test conductivelty, digital inspect	print Palladium seed layer	metal circuit traces				Y
11a			Cu electroless plating or electroplating					
12	cure			formulates inks				N
13	REPEAT			allows metal cross-overs				Y
14	flip substrates			enables printing double-layer		manual		N
15	REPEAT 3-8 back-side					optomec		N
	print via fill							
16	Laser trim fine-pitch for die assembly			achives sub-10um pitches				N
17	REPEAT 1-17 for multilayer laminate							
18	laminate layers							
19	print fill vias							
	ASSEMBLY							
20	Adhesive dispense		stencile	direct print				
21	SMT Pick and place			commercial-grade part assembly				
22	die eject (flip)			bare wafer die assembly	universal multi-wafer die place			
23	Conductive adhesive dispense (ACA)			ACA, ACF or magnetic materials	BESI Die bonder			
24	die Pick and place onto ACA			bare wafer or SMT	universal multi-wafer die place			
25	Post cure ACA/adhsives							
26	encapsulate			protects metal traces				
27	cure							
28		program tets full circuits						Y
29	overmold				final package			N
30		out of box testing						
								N
								N
Optional Steps								





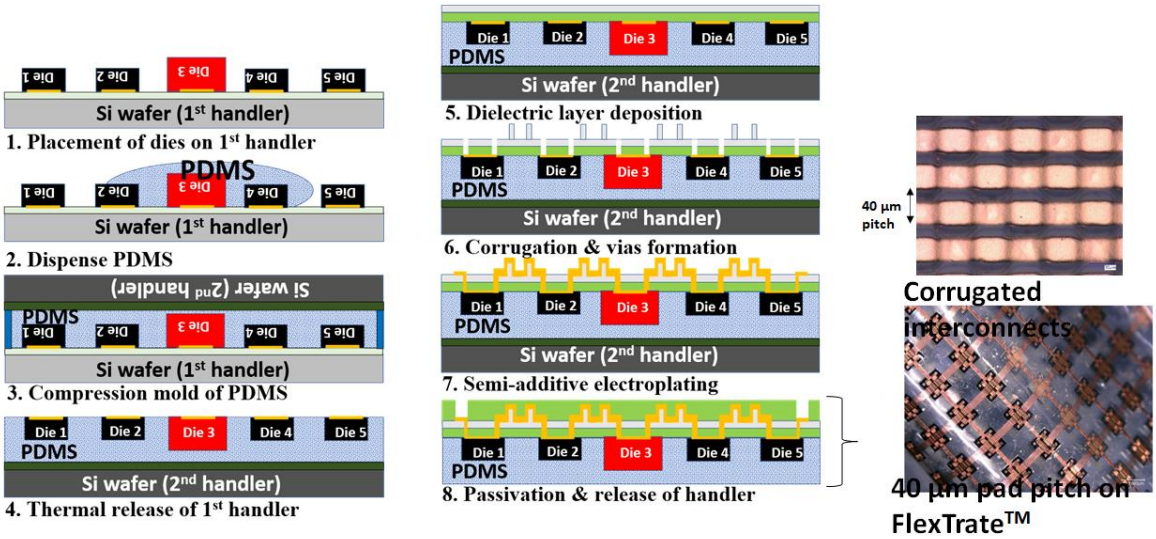
1) repeat steps for multi-layer build up
2) Hybrid electronic offers potential to print interlayer dielectrics eliminates one multilayer laminate



Flextrate (UCLA)

A		flexible Hybrid electronics			Typ. L/S (um) 20/20			
			Equipment			Material		
Step #	Process	Purpose	Equipment Type	ipment maker(US/Over	Gap: Y/N	Material Type	Material Makers	Gap: Y/N
1	clean substrate	prepare 1st handler				glass wafer	University wafer	
2	Adhesive A on Glass Handler						Nitto	
3	Parylene-C deposition	make alignment mark	SCS labcoater PDS 2010	US				
4	spincoat Fluoropolymer		spin coater	-		1700 Novec Electronic Grade Coating	3M	
5	Au Deposition		CHA Solution	us				
6	spincoat Su8 2000.5		spin coater	-		photoresist	MicroChem	
7	Pattern Su8 2000.5 for Die Placement Alignment Marks		Karl Suss MA6-Contact Aligner & Lithography	overseas(Germany)				
8	Etch Au		Ulvac 550 Chlorine Etcher	us				
9	Etch Fluoropolymer and Parylene		Oxford 80 Plus	overseas(UK)				
10	Place Dies		K&S-APRMA	us				
11	Place Teflon Ring	molding flexible substrate						
12	Dispense PDMS		THINKY Mixer	overseas(Japan)		PDMS MDX4210	Dow corning	
13	Adhesive B on Si Handler (2nd handler)		-	-				
14	Place 2nd handler over PDMS		-	-		-		
15	Compression Mold & Curing of PDMS		Karl Suss SB6	overseas(Germany)				
16	Release 1st handler	transfer to 2nd handler	Hot plate	-				
17	Surface Treatment for Adhesion	buffer layer	Oxford 80 Plus	overseas(UK)				
18	spincoat Adhesion Promoter		spin coater	-		adhesion promoter	Dow Chemical	
19	Parylene-C deposition		SCS PDS 2010	US				
20	Surface Treatment for Adhesion		Oxford 80 Plus	overseas(UK)				
21	spincoat Su8 2001			-		photoresist	MicroChem	
22	Hard Bake Su8 2001		Karl Suss MA6-Contact Aligner & Lithography	overseas(Germany)				
23	spincoat Su8 2005	corrugation	-	-		photoresist	MicroChem	
24	Corrugation on Su8 2005		Karl Suss MA6-Contact Aligner & Lithography	overseas(Germany)				
25	spincoat Photoresist AZ P4620	1st metal interconnect	-	-		photoresist	AZ Electronic Materials	
26	Pattern Resist for Vias		Karl Suss MA6-Contact Aligner & Lithography	overseas(Germany)				
27	Etch Vias to Dies		Oxford 80 Plus	overseas(UK)				
28	Remove Resist							
29	Surface Treatment for Adhesion		Oxford 80 Plus	overseas(UK)				
30	Sputter Ti/Cu		Ulvac JSP8000	us				
31	Surface Treatment for Adhesion		Oxford 80 Plus	overseas(UK)				
32	spincoat Photoresist AZ P4620		-	-		photoresist	AZ Electronic Materials	
33	Pattern Resist		Karl Suss MA6-Contact Aligner & Lithography	overseas(Germany)				
34	Thicken Interconnects		electroplating bath, custom made	-		electrolyte	Cu6320 as electrolyte	
35	Remove Resist							
36	Etch Cu Seed Layer with APS-100 Cu Etchant					etchant	Transene	
37	Etch Ti Seed Layer with BOE 6:1					etchant	Fisher Scientific	
38	Spin Su8 2010	2nd metal interconnect	spin coater			photoresist	MicroChem	
39	Pattern Contact Via to M1		Karl Suss MA6-Contact Aligner & Lithography	overseas(Germany)				
40	Su8 2005					photoresist	MicroChem	
41	Corrugation on Su8 2005		Karl Suss MA6-Contact Aligner & Lithography	overseas(Germany)				
42	Photoresist AZ P4620					photoresist	AZ Electronic Materials	
43	Pattern Resist		Karl Suss MA6-Contact Aligner & Lithography	overseas(Germany)				
44	Thicken Interconnects					electrolyte		
45	Remove Resist							
46	Etch Cu Seed Layer with APS-100 Cu Etchant					etchant	Transene	
47	Etch Ti Seed Layer with BOE 6:1					etchant	Fisher Scientific	
48	Parylene-C deposition	passivation and open contacts	SCS PDS 2010	us				
49	Photoresist AZ P4620					photoresist	AZ Electronic Materials	
50	Pattern Resist for Via		Karl Suss MA6-Contact Aligner & Lithography	overseas(Germany)				
51	Etch Vias to Dies		Oxford 80 Plus	overseas(UK)				
52	Remove Resist							
53	Release 2nd handler		Hot plate					

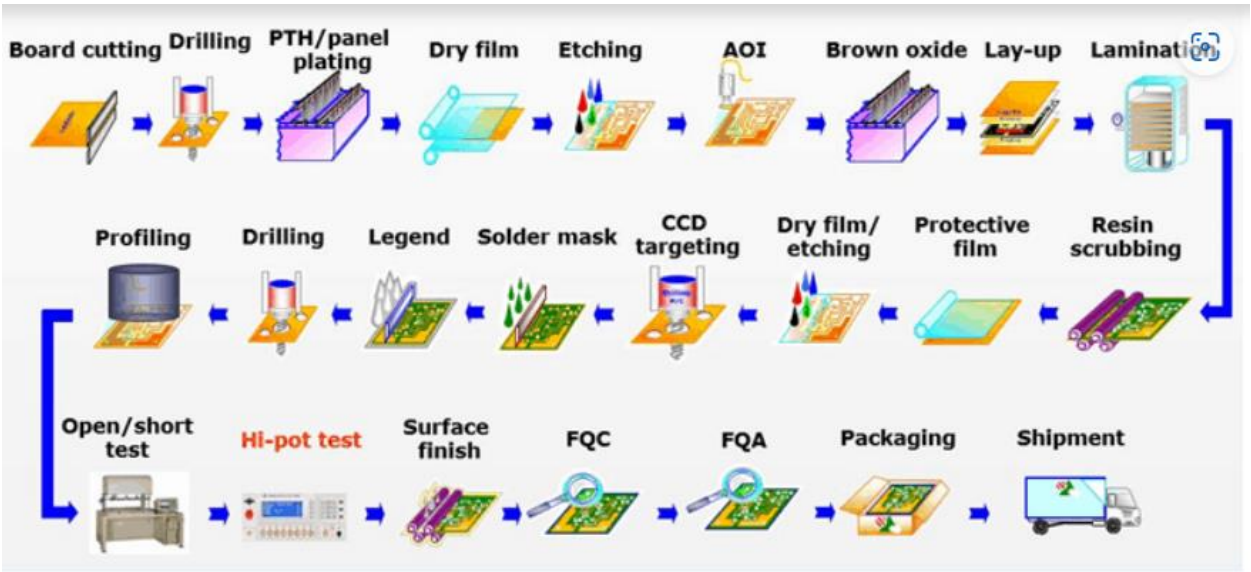
Flextrate (UCLA)



ISSUES: Molding compounds are typically poor thermal conductors
Die - shift can limit connection pitch

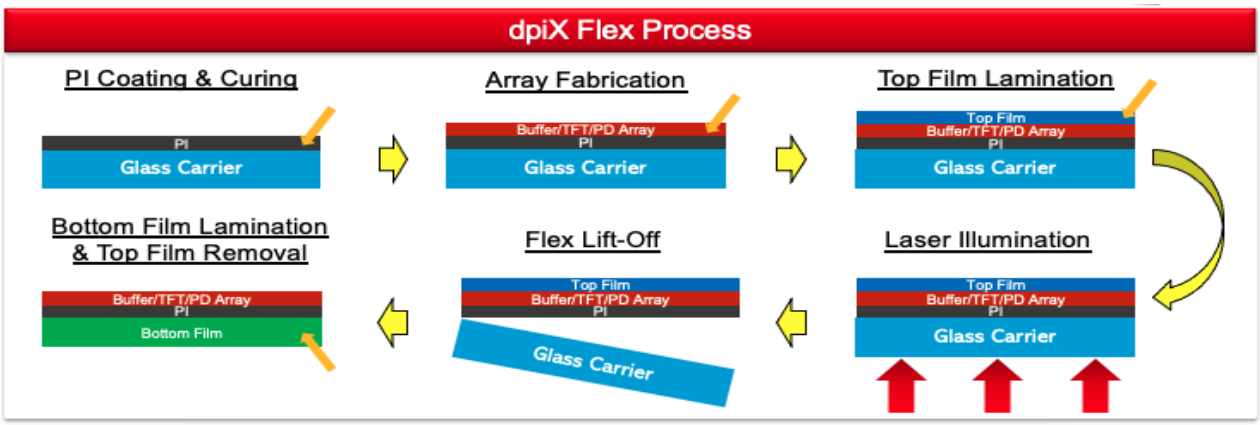
Conventional Printed Circuit Board (Navy Crane)

A	printed circuit boards						Typ. L/S (um)
							< 1um
					Equipment		
Step #	Process	test	alternates process	Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N
1	Step 1 – The Design			EDA software			N
2	Step 2 – Printing the Design						N
3	Step 3 – Creating the Substrate			Cu-plated FR4 glass filled polyimide			N
4	Step 4 – Printing the Inner Layers			The design is printed to a laminate, the body of the structure. A photo-sensitive film made from photo-reactive chemicals			N
5	Step 5 – Ultraviolet Light			Photo-patterned lithography			N
6	Step 6 – Removing Unwanted Copper			chemical etch bath			N
7	Step 7 – Inspection	electrical and manual		test conductivity and trace continuity			
8	Step 8 – Laminating the Layers			A prepreg (epoxy resin) layer goes on the alignment basin.			N
9	Step 9 – Pressing the Layers			Align with pegs manual lamination with clamps and presses. Pins are punch through the layers to keep them properly aligned and secured, these pins can be removed depending on the technology			N
10	Step 10 – Drilling			Holes are drilled into the layers by a computer-guided drill to expose the substrate and inner panels. Any remaining copper after this step is removed.			
11	Step 11 – Plating			The board is now ready to be plated. A chemical solution fuses all of the layers together. The board is then thoroughly cleaned by another series of chemicals. These chemicals also coat the panel with a thin copper layer, which will seep into the drilled holes.			N
12	Step 12 – Outer Layer Imaging			a layer of photoresist, similar to Step 3, is applied to the outside layer before being sent for imaging. Ultraviolet light hardens the photoresist. Any undesired photoresist is removed.			N
13	Step 13 – Plating			Repeat step 11			N
14	Step 14 – Etching			repeat step 6			
	REPEAT layers			multi-layer laminate process			
	Step 15 - cleaning						
Optional S	Step 16 – Solder Mask Application						
	Step 17 – Silkscreening			Silkscreening is a vital step since this process is what prints critical information onto the board. Once applied, the PCB passes through one last coating and curing process.			
	Step 18 – Surface Finish			The PCB is plated with either a solderable finish, depending on the requirements, which will increase the quality/bond of the solder.			
	Step 19 – Testing	electrical and manual		Before the PCB is considered complete, a technician will perform an electrical test on the board. This will confirm the PCB functions and follows the original blueprint designs.			
				https://www.candorind.com/pcb-manufacturing-process/#step8-laminating-the-layers			
Roadmap							
	PCB Assembly						
	Step 1 - Apply solder paste to the circuit board			Place the thin, stainless steel stencil over the board using a mechanical fixture. Solder paste should be applied evenly to the circuit board in the exact locations needed.			
	Step 2 - Pick and place the machine			SMDs, or surface mount components, should be placed on a prepared PCB by a robotic device. Then, the components need to be soldered onto the circuit board surface.			
	Step 3: Let the solder paste solidify			In order to adhere the components to the PCB, the solder paste needs to reflow and remain in place for an extended period of time.			
	Step 4: Inspect the PCB assembly			The assembled board should be tested and inspected for functionality. Ways to check the PCBA for quality control include:			
		Manual checks:		A visual inspection done in person by a designer to ensure the quality of a PCB.			
		Automatic optical inspection		An automatic optical inspection machine, or AOI machine, uses high-powered cameras, set at different angles to view the solder connections.			
		X-ray inspection		An inspection used for more complex PCBs by examining the layers of the PCB and identifying potential problems			
	Step 5: Insert the plated through-hole component			A plated through-hole, or PTH, component is a hole in the PCB that is plated through the board. Rather than soldering paste, more specialized soldering method is required for PTHs.			
	Manual soldering		Wave soldering	A manual, through-hole insertion			
				The automated version of manual soldering where a wave of molten solders all the holes in the bottom of the board at once.			
	Step 6: Complete a final inspection			Once the soldering process of the PCB board assembly is complete, it is time to do a final inspection and functional test. Run power and simulated signals to test the PCBs electrical characteristics. A sign that the PCB has failed is when it shows the fluctuation of electrical signals during the test. If the PCB fails the final inspection, it should be scrapped. And the process begins all over until a successful PCB is produced.			



Flexible Panel Level Processing (Innovaflex-dpiX)

Flexible panel level porcesses			Typ. L/S (um)			
			> 5um			
	Equipment			Material		
Purpose	Equipment Type	Equipment maker (US/Overseas)	Gap: Y/N	Material Type	Material Makers	Gap: Y/N
Substrate	Coater, HVCD, Furnace	Tazmo, TOK, Screen, Chugai Ro, JTEKT, Viatron	N	Polyimide	UBE, Toray, Kaneka	N
Moisture Barrier	PECVD, ALD	AMAT	N	SiH4, NH3	Air Products, Air Liquide, SK Chem, Hansol	N
Pad Metal	PVD	AMAT, ULVAC	N	TiW, Al, ITO, Cr	JSR, TOK, DuPont	N
Insulating Material	PVD -> ALD	AMAT	N	SiH4, NH3	JX, Honeywell, Linde, ToSoh, KFMI	N
Insulating Material	Coater, VCD, Furnace	Tazmo, TOK, Screen, Chugai Ro, JTEKT, Viatron	N		Dupont, TOK, Honeywell, Toray	N
PI separation from Glass	Laser	3D Micromac	N			N
Cu dishing, Erosion	Full Wafer AFM		Y			
Surface treatment	Wet Clean (Single Wafer)	Lam,	Y/N	Wet Chemistry		N
Surface treatment	Plasma (e.g. etch)	Lam, AMAT, TEL, PlasmaTherm, Oxford	Y			N
Post metalization	PVD		Y	TiW, Al		
Post metalization 2	PVD, Plating		Y	Cu plating	Entegris, BASF, MLI, DuPont	N
Chip Placement	Pick-and-Place	ASMPT	Y			
Thermal	Furnace		Y			N
Dicing	Laser		Y			



Chapter 4: Reliability and Thermal Challenges

Contents

4.1. Reliability Introduction 1
4.2. Stakeholders for Reliable HI Systems: Application Domains 7
4.3. Reliability Considerations 11
4.4. Reliability Summary 22
4.5. Thermal 22

Authors: Benson Chan, Mary Ann Maher, Abhijit Dasgupta, Gamal Refai Ahmed

TWG 2 is a cross cutting TWG that looked at technologies, processes and structures that will affect more than one industry/market sector. Examples of these sectors would be HPC, Consumer, IoT, Wearables, Medical etc. An example of a pervasive technology would be solder, most if not all electronics assemblies are made using solders as a means to join two or more electrical component to a circuit board. SAC 305 is used in all consumer devices, while HPC applications may have other solders if their intended usage is medical or military where Eutectic SnPb solders are still allowed due to waivers granted at the start of RoHS legislations. Understanding solders; their failure mechanisms, usage limitations, application in advance packaging will be covered in the Reliability portion of this report.

The Thermal section covers the drive towards higher power and what options are available today to manage the higher thermal loads that these devices will produce. We investigated advanced TIM materials as well as alternate cooling technologies such as single and two phase liquids and immersion cooling to support future systems.

Reliability

4.1. Introduction

Increasing system complexity, functionality, diversity and density, as a result of the twin drives for Heterogeneous Integration (HI) and miniaturization, involving multiple chiplets, will pose new challenges for meeting and verifying customers' reliability targets. HI systems of the future will be multiscale and multi-physics systems and will combine highly resilient designs with self-monitoring, self-cognizance and varying degrees of adaptive reconfiguration and self-healing capabilities to provide high reliability and availability, in spite of distributions of intrinsic material defects, manufacturing flaws and stochastic variabilities. Heterogeneous Integration requires a unified reliability approach across the entire product stack-up from device level to Chip-Package interactions (CPI), package, boards/ modules and systems, to be accomplished by an integrated reliability team across all these levels of integration, to meet the customer's reliability targets. The HI reliability team will also need to meet holistic constraints such as reducing the time required for new product introduction (NPI) and minimizing cost of ownership over the life-cycle of

successive generations of products. Such an integrated approach towards reliability will require a rigorous, disciplined and proactive approach that strategically combines reliability physics with powerful artificial intelligence algorithms, to leverage the unprecedented levels of real-time field performance data, service condition data, product stress data and system/component reliability data that is becoming available via IoT infrastructure. This section lays out the scope, challenges, disruptive opportunities and potential approaches for achieving such an integrated approach, in HI technologies that are likely to emerge over the next 0-5, 5-10 and 10-15 years.

Reliability describes the ability of products to meet intended performance targets throughout their useful life. The metric often used to quantify reliability is the time-dependent probability of meeting the intended performance goals. Related metrics are the histories of the instantaneous failure rates and hazard rates. Reliability risks come from a combination of wearout aging mechanisms and unexpected catastrophic degradation/failures due to overstress events during the lifecycle. The optimum reliability can be achieved by understanding the reliability expectations, product micro/macro environment and impact of the environment on wearout behavior based on product technology characteristics. As illustrated in Figure 1, at the simplest level, reliability risk is often visualized as a stress-strength interference, where unreliability comes from the probability that the applied 'stress' will exceed the inherent 'strength' of the product. The tasks of managing reliability include effective ways to quantify these distributions (and their evolution throughout the life-cycle) and balancing their interactions, as a function of product design, manufacturing variabilities and service expectations, to ensure that the resulting reliability margins will meet the customer's expectations.

The process of quantifying and managing the time-dependent 'stress' and 'strength' interference requires science-based multi-physics, multiscale co-design approaches that leverage the rich disciplines of multi-physics simulations, reliability physics (RP) and artificial intelligence (AI). The 'stress' distributions will have to be identified based on a combination of multi-physics simulation and data-driven AI approaches. AI approaches will have to be based on sophisticated machine learning methods that exploit data analytics and deep learning technologies to correlate reliability outcomes (based on field failures and test failures) with key design and manufacturing attributes. The outcome of such 'stress analysis' will help to identify the intensity of the electrical, thermal, mechanical and chemical fields expected at potential failure sites throughout the expected life cycle of the product. Simultaneously, identifying the corresponding multi-physics 'strength' distributions will require a similar combination of fundamental RP models and advanced AI methods.

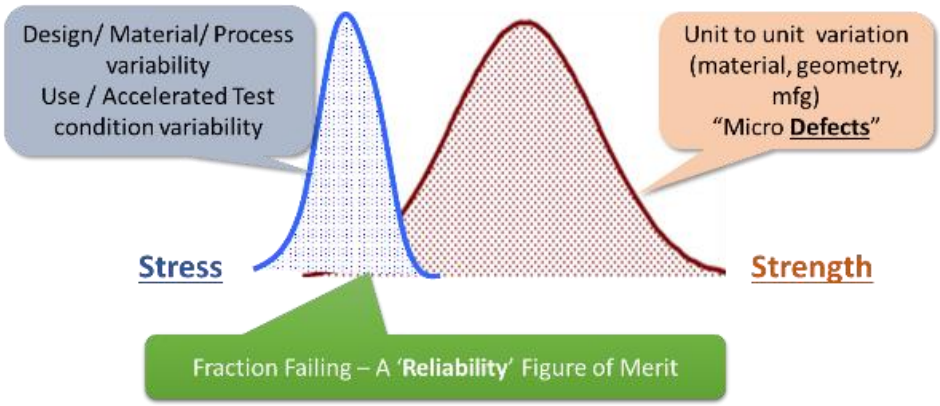
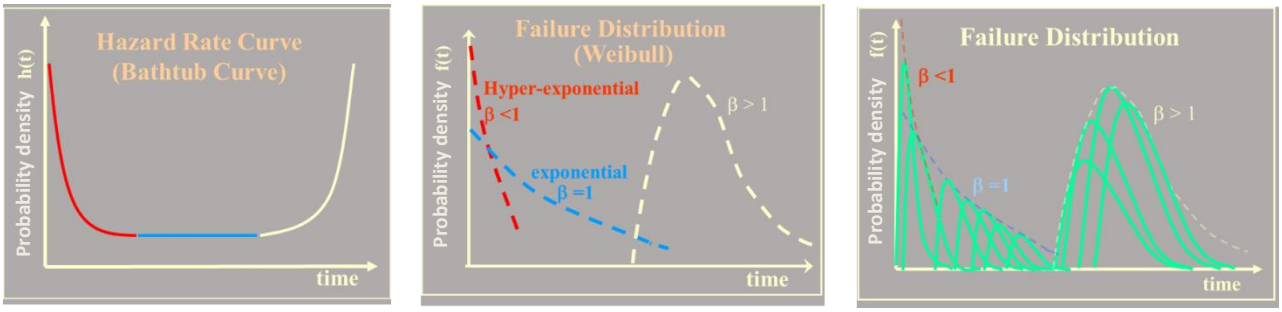


Figure 4.1: 'Stress' vs. 'Strength' interference

RP will use a 'bottom-up' approach to enable robust design margins based on assessment of dominant degradation/failure mechanisms at critical sites, while AI will provide a complementary 'top-down' perspective of system-level risk, based on the unprecedented level of real-time field reliability data that will become available via IoT infrastructure.

The concept of RP perspective of system-level risk is schematically illustrated in Figure 4.2 where the traditional system-level reliability 'bathtub' curve is shown in terms of hazard rates in Figure 4.2a and probability density functions (pdfs) in Figure 4.2b. Such bathtub curves have traditionally been obtained from top-down statistical analysis of failures encountered during accelerated testing and during the life-cycle of fielded products. Such an approach is reactive and will no longer be sufficient for proactive development of reliable HI systems. The challenge facing the reliability community is to evolve AI/ML approaches to extrapolate such reliability data and knowledge of past/current systems to proactively predict such distributions for future HI systems under development, by using appropriate design and manufacturing information. Figure 4.2c emphasizes the corresponding 'bottom-up' RP view that this system-level failure information is actually the result of many competing degradation/failure mechanisms that are active at multiple critical failure sites. End-of-life failures (under the white section of the bathtub curve) in Figure 4.2 are those usually seen in well-manufactured products and depend on the intrinsic robustness of the design. Pre-mature failures (under the red and blue portions of the bathtub curve) depend on the distribution of weak sub-populations due to manufacturing and material variabilities/defects



(a) (b) (c)

Figure 4.2a: Bathtub Curve showing system-level hazard rates for 3 phases ('infant mortality' stage in red, 'usable life' in blue and 'end-of-life' stage in white;

Figure 4.2b: Bathtub curve replotted as scaled probability density functions;

Figure 4.2c: Schematic illustration of underlying competing failure distributions (due to competing failure mechanisms) that constitute the bathtub curve.

In complex, multi-physics, multi-scale, HI systems, developers will have to leverage both RP (bottom-up) and AI (top-down) co-design approaches and digital-twin approaches, to estimate these failure rates. In turn, this will lead to unique opportunities to ensure system robustness and resilience, reduce time to market and minimize cost of ownership.

Figure 4.3a below provides a sample listing of the dominant multi-physics degradation mechanisms in electronic systems. 'Overstress' mechanisms are triggered under the action of sudden catastrophic stress events while 'wearout' mechanisms cause gradual damage accumulation throughout the life cycle because of routine operational and environmental stress exposures. Each of the listed mechanisms represent a rich body of expert knowledge, including quantitative models for assessing design margins and acceleration factors, model constants for different existing material systems, and methods for quantifying the model constants for new materials. These models need to be integrated seamlessly into digital twins that are based on multi-physics RP models along with real-time data-based AI methods, so that management of reliability can truly become a cradle-to-grave function, in a fully integrated environment for:

(i) Co-designing for reliability (DfR)
(ii) Manufacturing for reliability (MfR): assessing role of manufacturing variability on design margins
(iii) Qualifying for reliability (QfR): Verifying product robustness with accelerated stress testing guided by science-based acceleration factors
(iv) Sustaining for reliability (SfR): assessing prognostic metrics such as remaining useful life (RUL)

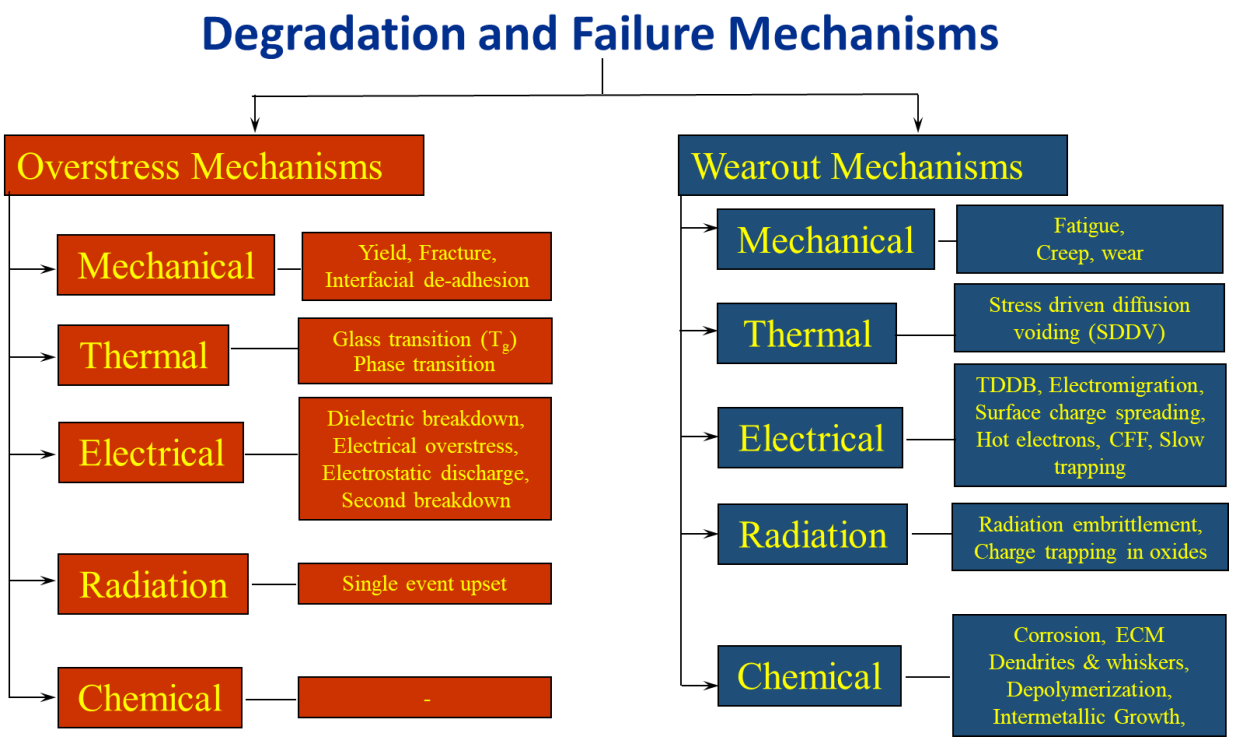


Figure 4.3a. Examples of dominant multi-physics degradation/failure mechanisms in electronic systems, under overstress and wearout stress exposures

(TDDB = Time-dependent dielectric breakdown; CFF = conductive filament formation;

ECM = electrochemical migration)

The top-down concept of data-driven approaches that will use artificial intelligence (AI) algorithmic approach for managing the reliability of complex systems is schematically illustrated in Figure 4.3b. This figure presents a flowchart of the process: (i) collection of system data (performance data and environmental stress data), (ii) smoothing and de-noising of the data using filtering methods; (iii) anomaly detection, using supervised and unsupervised machine learning algorithms; (iv) pattern identification with physics-assisted diagnostic algorithms to identify the root-cause source of the anomaly; (v) pattern extrapolation with physics-assisted prognostic algorithms to assess the remaining useful life (RUL); (vi) actionable responses to RUL estimates (e.g. design support decisions, improvement of manufacturing process flow and process control, self-healing actions, and feedback for improving the data acquisition-analysis cycle).

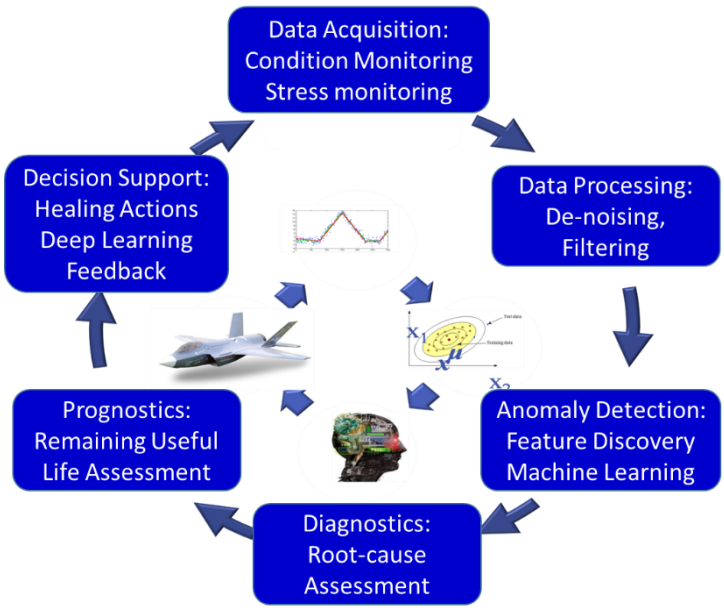


Figure 4.3b. Flowchart of data-driven methods for reliability assurance, using artificial intelligence (AI) algorithms.

As discussed above, the success of future hybrid reliability assurance methods will rely on judicious fusion of the RP and AI methods of Figure 3a and 3b. This fusion prognostics approach is schematically shown in Figure 4.3c.

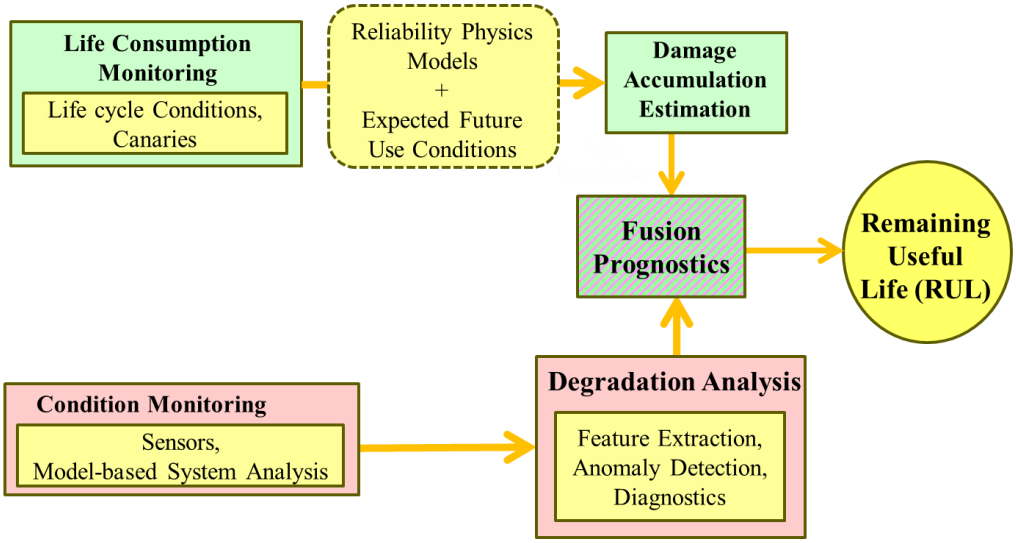


Figure 4.3c. Flowchart showing conceptual schematic of fusion prognostics using combinations of data-driven artificial intelligence (AI) algorithms and reliability physics (RP) models.

4.2. **Stakeholders for Reliable HI Systems: Application Domains**

Reliability activities for future HI systems will have to evolve in ways that enable the transformative HI Roadmaps proposed by relevant stakeholders in different application domains. In this section the focus is on: (i) High-Power Computing (HPC) and Data Centers; and (ii) Wearables. There are significant differences in the reliability requirements of these two communities. For example, HPC applications require extremely high-performance leading-edge dies along with highly complex HI architectures to integrate processor ad memory, with ultra-high I/O requirements and extremely high steady-state thermal dissipation. Integration of processors and memory systems, energy-efficiency, and thermal management were listed as challenge areas in a past DOE report prepared for the state of High-Performance Computing (HPC) in the US [Luc 14]. These challenges remain at present, not just for HPC but also for data centers, which consume about 71 billion KWh of electricity annually in the US [She 16]. SiPs that integrate processing and memory chiplets address these challenges.

In contrast, wearables may rely on less powerful processors, but often involve integration of multi-functional sensors with processors and wireless communication devices. In addition, wearables sometimes may have to be incorporated onto flexible substrates that might undergo large flexural and stretching deformation, thus exposing the electronics to large mechanical deformation and stresses. Finally, wearables often have on-board power sources like batteries, whose reliability also needs to be considered in system reliability.

Reliability activities need to be customized for the differing needs for each of these two communities. In this chapter, we will briefly address the needs of the HPC community. Similar special needs of the wearables community are deferred to future version of this document.

HPC and Data Systems:

Examples of the key market drivers and milestones presented in the HIR HPC Chapter are summarized next. The dominant degradation/failure modes and mechanisms for these technology milestones are discussed later in this section.

In recent years, several important market drivers have emerged, primarily driven by new applications. These include: • **Internet-of-Things (IoTs)** with processing needs and pre-processing at edge nodes with high IO connectivity to the "things" (sensors/actuators) and final sensor fusion and processing/storage at nodes that would typically be within the cloud.

• **Data analytics** is a growing need in the mass e-commerce, financial industries, smart healthcare, and social networking, with heavy reliance on analytics requiring customized FPGAs, GPUs and customized hardware.

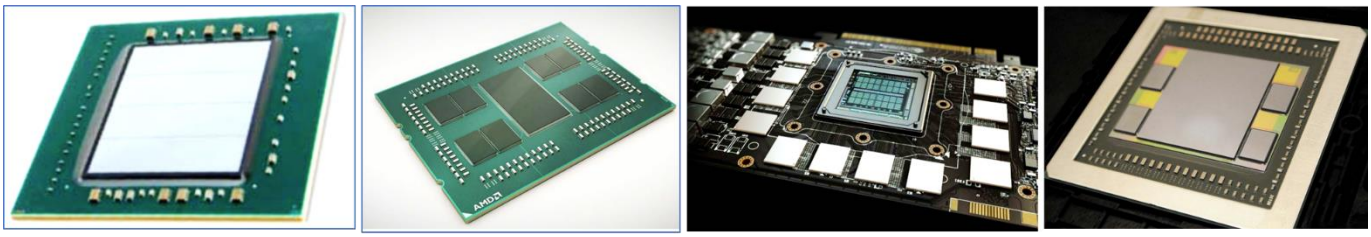
• **Intelligence** needs for recognition and prediction using machine learning techniques, such as convolutional neural networks (CNNs) and deep neural networks (DNNs), have seen the deployment of GPUs, FPGAs and special purpose accelerator chips, especially in cloud-based server platforms.

• **Blockchain processing** is a highly- parallelizable application requiring heterogeneous integration innovations in the data center market, relying on GPUs and integration with memory and I/O chiplets inside SiP packages.

• **Special functions and accelerators**, starting with GPUs and FPGAs, have been targeting applications that go beyond graphics and scientific computing, e.g. to accelerate neural networks, quantum computing, cognitive neuro-morphic computing and graph processing, relying on bit-serial/data-parallel processors, and AI accelerators incorporating analog processing components.

• **Memory-centric computing** involves large data sets and requires a high processing rate or low processing time. Computing logic performing processors are connected to HBMs in a 2.5D configuration on an interposer or implemented in a "logic" layer underneath stacked memory dies in a 3D configuration.

Most of these emerging applications will benefit from special-purpose accelerators, including custom ASICs, FPGAs and GPUs, which provide energy-efficiency, fast implementation strategy and access to significant amounts of data from memory. Heterogeneous integration provides a key solution pathway to meet some of these memory needs by integrating accelerators with Stacked RAM or HBM within a package. Some current examples of relevant packaged systems are shown in Figure 4.4.



Tesselated FPGA | Partitioned Server | GDDR Memory surrounding the CPU on organic substrate | GPU+HBM 2.5D on Si interposer

Figure 4.4. Current examples of HI systems for HPC Applications

The full MRHIEP packaging architecture roadmap is schematically shown in Figure 5 (taken from Refai-Ahmed et al [1-2]).

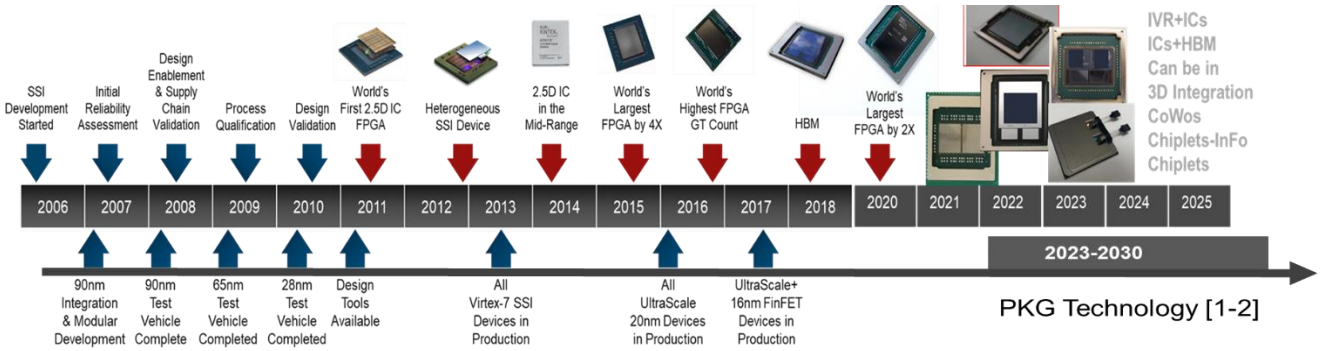


Figure 4.5. Packaging Architecture Roadmap [Source: Refai-Ahmed et al[1-2]]

Table 4.1 shows the corresponding technology roadmap, proposed by the Advanced Packaging TWG, :

Parameter	Unit	2025	2027	2029
Silicon Node	Nm	3nm	2nm	1nm ?
I/O Bandwidth (Logic-HBM)	Gbps	1024 x 2.4	2048 x 3.6	4096 x 6.4 ?
I/O per mm per layer (shoreline)	#	250	500	1000 ?
I/O lines and spaces (and vias)	micron	2/2/2	1/1/1	0.5/0.5/0.5 ?
Package to Board I/O BW	Gbps	64 per I/O	112 per I/O	256 per I/O ?
Package to Board Pin Count	#	9600	11200	12800 ?
Power Density	W/mm^2	1	1.05	1.1 ?
Package Dimension (Minimum)	Mm	95	103	120 ?

Table 4.1: Technology Roadmap [Source: Advanced Packaging TWG, MRHIEP]

Furthermore, as discussed in the MRHIEP Thermal TWG, the enormous thermal management challenge to meet the power density targets listed in Table 4.1, will require future chiplet/package architectures that use bare/exposed die because by doing this, we are reducing the package temperature drop between the die and package. However, merging to exposed/bare die along with power increase has its own challenges, such as TIM selection and package surface warpage, especially as the package sizes progressively increase (as shown in Table 4.1 and in Figure 4.6).

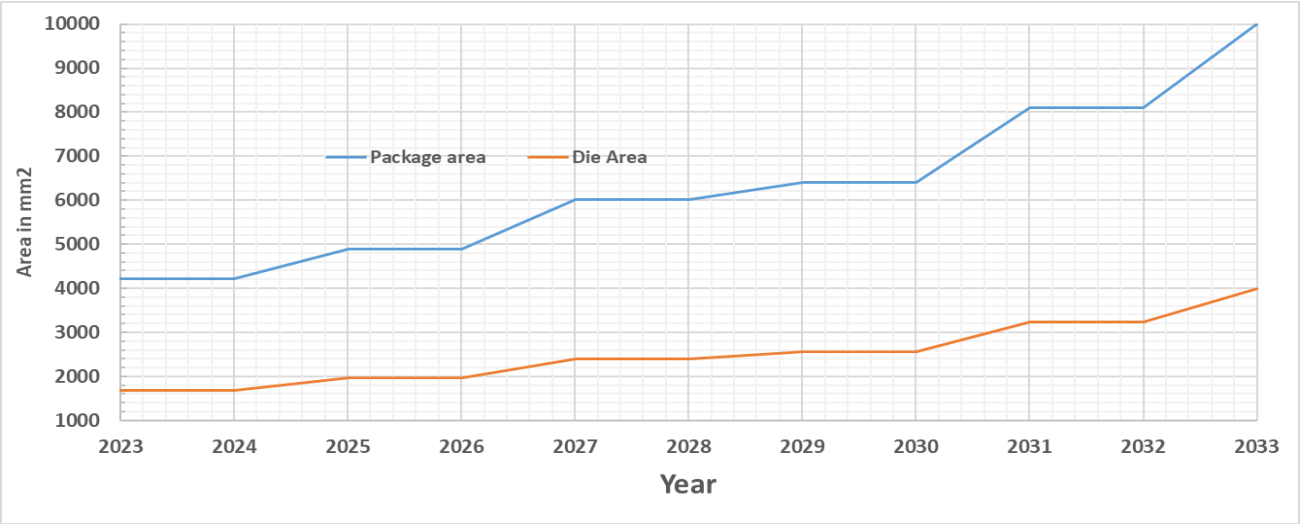


Figure 4.6. Projection of progressive increase in die and package size in HPC applications

[Source: Refai-Ahmed et al [3]]

Typical assessment of corresponding increase in package warpage with increase of package size is shown in Figure 4.7. Needless to say, these warpage estimates are dependent on package architecture and will change with design modifications. The resulting quality and reliability considerations are discussed below in Section 4.3.1.

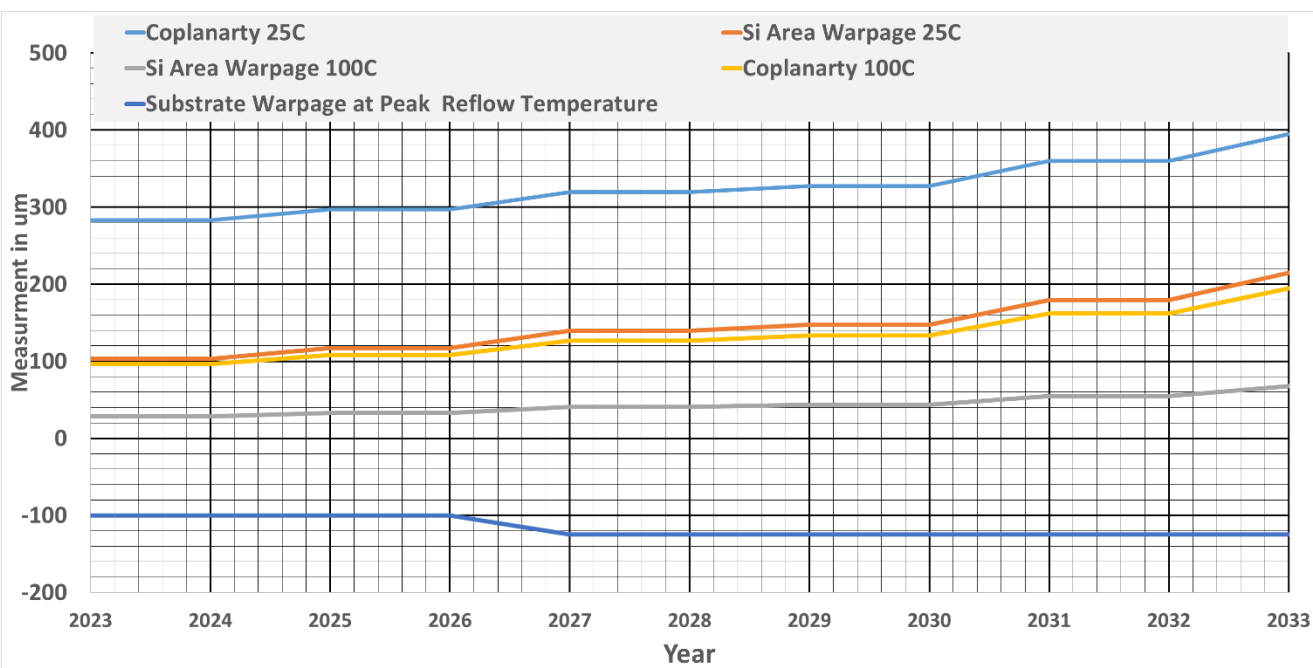


Figure 4.7. Typical warpage estimates as package sizes progressively increase

[Source: Refai-Ahmed et al [3]]

In summary, the requirements dictated by these SiPs on the heterogeneous integration methodologies and processes, are [Source: HIR Roadmap): 1) On-package interconnections; 2) Off-package interconnections; 3) Signal integrity and distribution needs; 4) Power distribution and regulation; 5) SiP-level global power management and overview of thermal management; 6) Security and reliability issues; 7) Design tools; 8) Impact on the supply chain. The reliability implications are discussed in Section 3.

The HPC community understands that developing and supporting reliable HI systems using advanced technologies will require a phased approach. Systems will use technologies of varying maturity. The mature nodes will be relatively easier to qualify, while the advanced semiconductor nodes (with corresponding advance SIP technology nodes) will require more work and time.

Managing the life cycle reliability of such complex systems for such demanding HPC architectures and environments will clearly require digital twins powered by intelligent fusion of co-design for reliability with real-time health prognostics. These functions will have to rely on fusion of reliability physics with data-driven machine learning approaches, discussed earlier in Figure 3c.

4.3. Reliability Considerations

Addressing reliability in complex HI systems requires discussion of hardware reliability, software reliability, human operator reliability and their interactions for fielding, operating and supporting reliable firmware. This section focuses on the hardware reliability issues, while the software reliability aspects and the impact of operator-machine interactions on system reliability are deferred to the next release of this roadmap.

Typical reliability tasks/disciplines related to quantifying and managing the stress and strength distributions are grouped for convenience under 7 headers, shown in Figure 4.8 and discussed below [Source: HIR Roadmap, Reliability TWG]:

(i) Identification of customers' reliability targets for different market segments and different technology segments

(ii) Identification of life-cycle user models that include expected life-cycle environmental & operational stress profiles and understanding of system configurations

(iii) Design for reliability (DfR) tasks using reliability physics (RP), artificial intelligence (AI) methods (based on data analytics and machine learning), materials-centric approaches, co-design simulation methods and resilient, fault-tolerant design approaches

(iv) Manufacturing for reliability (MfR) using knowledge of the effect of processing conditions on material behavior; understanding of process quality, defects and yields; use of appropriate process metrology; AI-based process control; and stress screening approaches, as needed

(v) Qualification for reliability (QfR) which includes knowledge-based accelerated stress testing approaches for engineering verification testing (EVT), design verification testing (DVT) and process verification testing (PVT)

(vi) Supporting for reliability (SfR) which includes, personalized *in-situ* prognostics and health management (PHM) for high availability and system resilience, using fusion of RP models and data driven AI models, based on: real-time detection of early anomalies and failure precursors; system diagnostics and prognostics; and dynamic adaptive healing/reconfiguration

(vii) Integration and managing of reliability best practices across the supply chain.

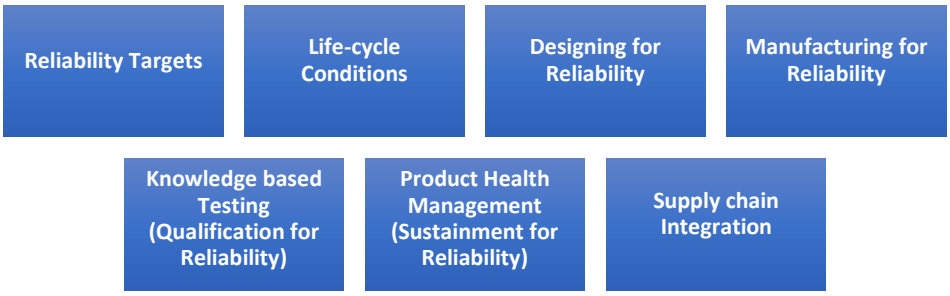


Figure 4.8. Thrust areas for managing reliability risks [Source: HIR Roadmap, Reliability TWG]

Figure 4.9 shows a sample flowchart of tasks for developing and fielding reliable IC technologies, covering the entire spectrum of tasks for Single-chip and Multi-chip IC systems from product concept to volume production to field support. Similar charts can be developed for Substrate/Board Reliability and for Interconnect/Assembly Reliability.

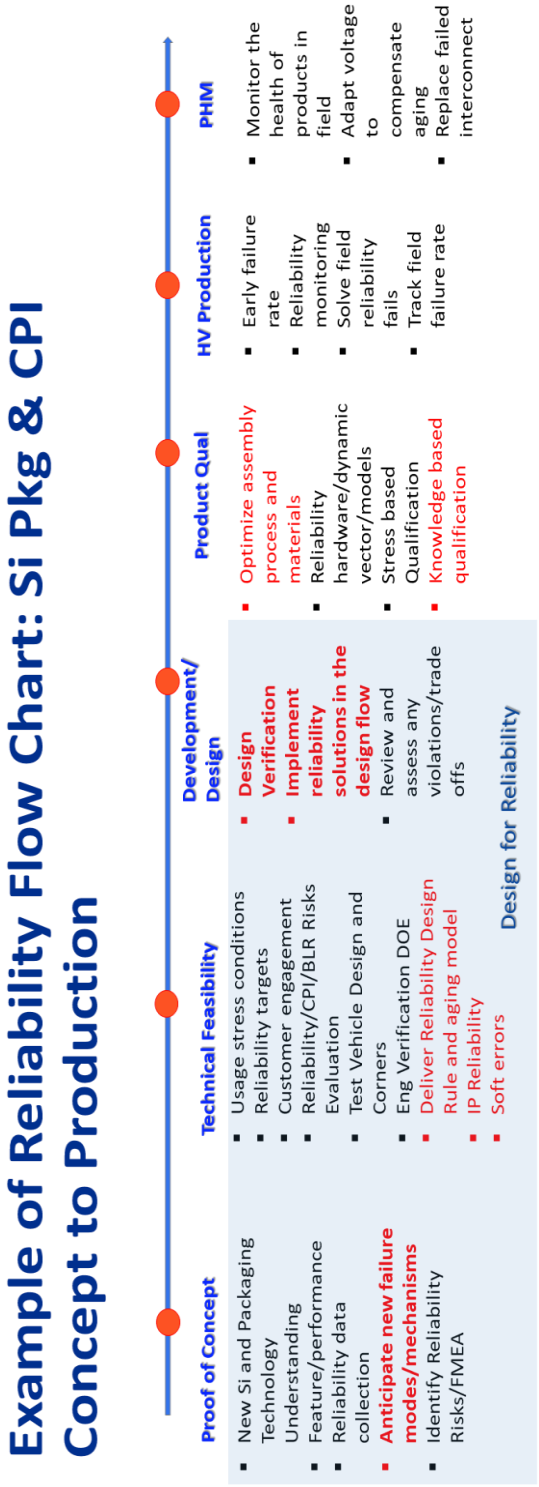


Figure 4.9. Sample flowchart for hardware reliability tasks for IC manufacturers

[Source: HIR Roadmap, Reliability TWG]

4.3.1 Chip/Package/Board Interactions (CPBI):

The failure occurs at the weakest link when chips and chiplets are assembled into packages and populated onto substrates or printed circuit boards (PCBs). Four major CPBI failure mode categories are: chip failures, package failures, device performance shift, and package-to-board interconnect failures. Details of these degradation modes are discussed in the Reliability Chapter of the HIR Roadmap and are briefly summarized here.

Chiplets are vulnerable to both global stresses arising from overall thermal expansion mismatches between the chiplet and the surrounding package, as well as to local stresses arising from gradients of temperature due to self-heating effects (SHE) in complex 3D transistor architectures, such as FinFET and Gate All Around (GAA) devices. As shown in Fig 10, failures driven by global stress can occur in BEOL features such as extremely low-k (ELK) dielectric, Chip corner/edge cracking; Under bump or wire bond pad cratering; UBM cracking; Die backside cracking.

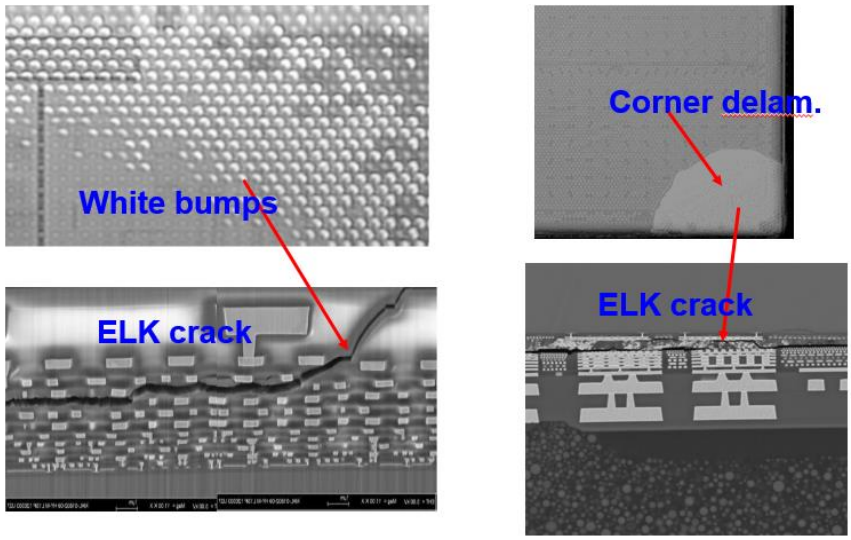


Figure 4.10. Typical CPI induced chip failure modes [Source: HIR Roadmap, Reliability TWG]

In contrast, SHE is known to generate a localized temperature concentration and localized thermal cycling stress profile on top of the global thermal profile, leading to BEOL interconnect failures, accelerated aging of transistors, particularly Hot Carrier Injection (HCI), BEOL stress and electromigration (EM), and bottom layer Cu/ELK cracking. When the SHE is severe enough, it can also burn out the channel. Further scaling requires new materials for metal lines, barrier layers and ELK.

When the chip is stronger than the package, the package experiences CPI failure modes, such as Underfill cracking/delamination; Solder mask cracking/delamination; Substrate failures; Bump cracking near substrate

Due to its piezo-electrical properties, CPBI stress in Si changes the carrier mobility for both NMOS and PMOS transistors. Sources of such stresses include: Local stress caused by Through-Silicon-Vias (TSVs), as shown in Figure 4.11; global stress in thin-die WLCSP; local stress

transmitted by bump and μ-bump. Furthermore, with decreasing TSV pitch, active circuitry may experience higher stress, causing transistor performance drifts for FEOL, MOL and BEOL, due to HCI/BTI, TDDB and EM.

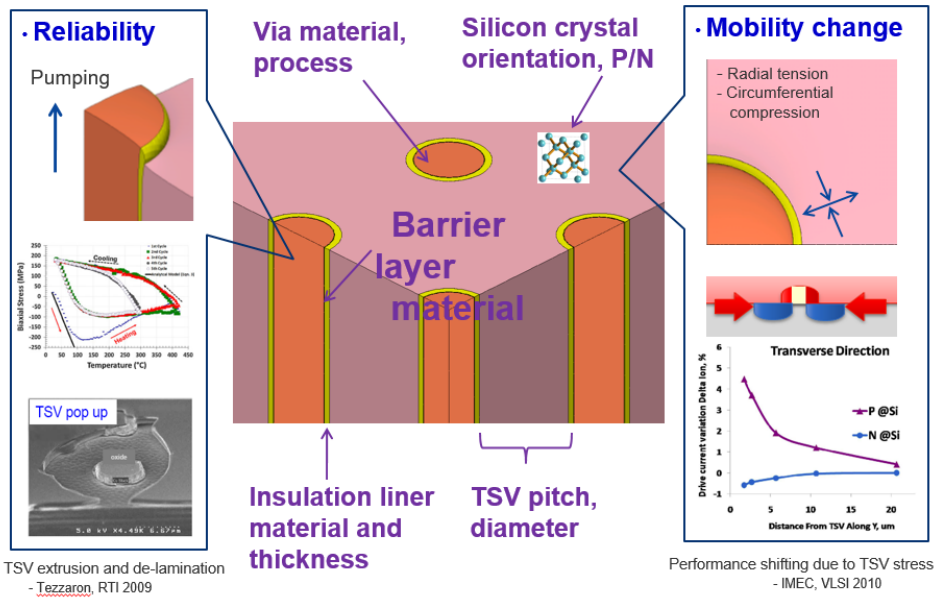


Figure 4.11. TSV-induced stress effect on adjacent transistors [Source: HIR Roadmap, Reliability TWG]

CPBI failures and the board level stress can cause additional failures: WLCSP Die edge cracking after surface mounting; Fan Out Wafer Level Package (FOWLP) RDL layer cracking; Large Flip Chip Ball Grid Array (FCBGA) failure: underfill cracking/delamination or even chip failures due to excessive warpage in Large size FCBGA packages mounted on rigid PCBs. Figure 4.12 shows some typical CBPI failure modes in a FOWLP, e.g. ELK cracking; circular cracks near die-edge in passivation (PSV) layers, redistribution layer (RDL) and BEOL interconnects; initial crack in inner interconnect.

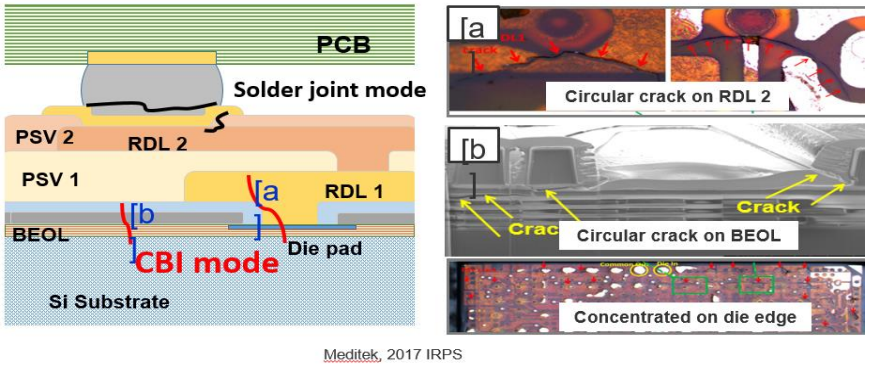


Figure 4.12a. Typical FOWLP CPBI failure modes [Source: HIR Roadmap, Reliability TWG]

For molded packages with heterogeneous integration architecture, interaction with the mold is a key factor and must be characterized. Furthermore, thermal-mechanical concerns for ultra-thin

dies/die-stacks under manufacturing environment need to considered as well; *e.g*. cold plate assembly, system testing, etc.

Unique CPBI interactions exist also in Si-photonics chips and are expected to present special reliability considerations. A schematic of a Si-photonic assembly is shown in Figure 12b (adapted from HIR Photonics chapter 9, https://eps.ieee.org/images/files/HIR_2023 /ch09_photonics.pdf). New reliability challenges and degradation modes are expected due to thermo-mechanical deformations and due to humidity and optical exposure in: optical TSVs, Si waveguides, die-to-die bonding interconnections, laser sources (WBG semiconductors), Ge detectors, misalignment and environmental degradation in optical interconnects and optical couplers, as well as stress corrosion cracking in glass substrates. Figure 4.12c provides a summary of the expected reliability issues.

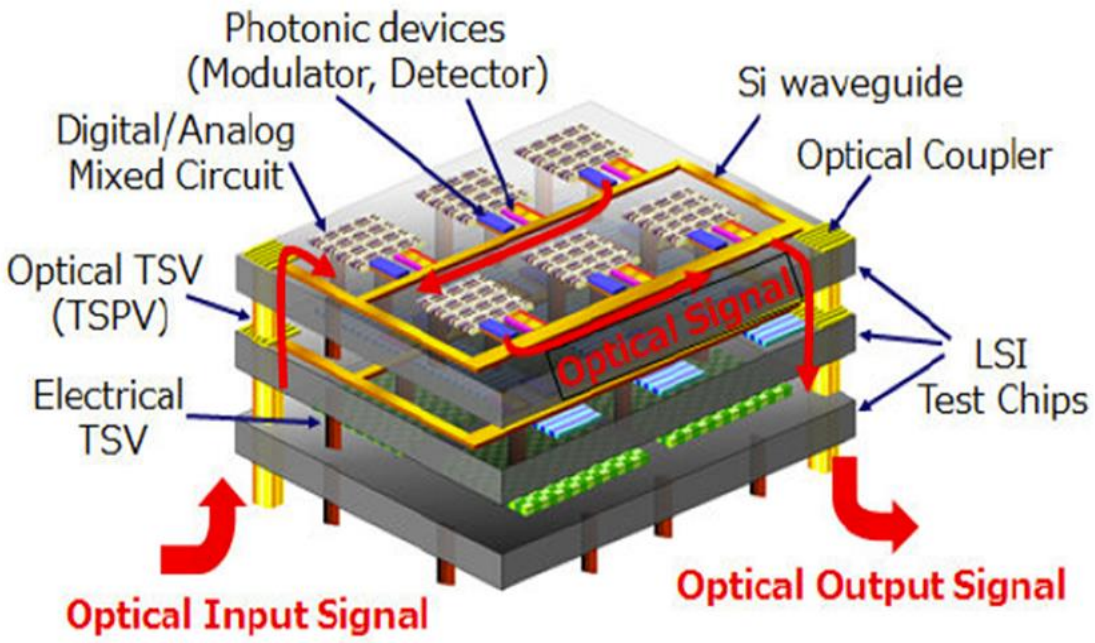
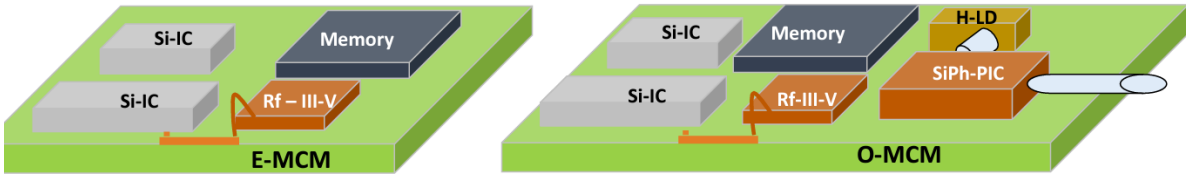


Figure 4.12b. Schematic of future photonic/Electronic 3D-SiP with electro-optical package substrate (Figure 18 of HIR Photonics Chapter)

High Level Comparison E-MCM and O-MCM Reliability Challenges



Failure mechanism	Drivers	E-MCM	O-MCM
Corrosion	%RH, δV/δx, T	☑	☑
Material migration	T, δT/δx, σ, δσ/δx, δI/δv,δI/δx	☑	☑
Thermomechanical	ΔCTE, ΔT, size, Stiffness,IMCs	☑	☑
Mechanical overstress	shock, vibration,corrosion, damage to semiconductor	☑	☑
Polymer moisture absorption	%RH, T, δT/δt, δσ/δx	☑	☑
Stress corrosion cracking	%RH, T, δT/δx, σ, δσ/δx, δV/δx	☑	☑
Light induced degradation	%RH, T, λ, photon density, δσ/δx		☑
Laser degradation	T, δT/δx, σ, δσ/δx, δI/δv, optical power, hermeticity		☑
Loss of hermeticity	T, δT/δx, σ, δσ/δx,		☑

Figure 4.12c. High-level comparison of electronic vs photonic HI systems (Presentation by John Osenbach, Technical Fellow, Infinera, IEEE REPP Conference Nov 2023)

The projected increase in package size and warpage in HPC HI systems was discussed in Section 4.2.1 (Figures 4.6 and 4.7). The resulting concerns in interconnect quality and the corresponding CPBI reliability considerations are discussed here. The expected increase in package complexity and size (presented earlier in Figure 6), will pose significant CPBI reliability challenges. For instance, % change of thermo-mechanical strain in solder I/O interconnects during temperature excursions scale with the package size and die size (diagonal length). The corresponding risk of fatigue failures in interconnects during temperature cycling will scale approximately as the square of the change in package/die size, *i.e.* as the %change in the package or die area. The expected drop in interconnect PTC durability for the next decade is shown in Figure 4.13 (normalized wrt 2023 durability). Clearly, fundamental changes are needed in solder interconnect technology in order to achieve adequate durability as package sizes continue to evolve. Similarly, risk of die-attach fatigue delamination scales with increase in die area. The reliability problem will be further exacerbated by the ever-increasing I/O density and quantity per package, shown in Table 4.1, in view of the increased complexity of multi-layer redistribution layers, multi-layer interposers and substrates, increased number of vias, increased probability of fabrication defects (quality challenges), and shrinking feature sizes.

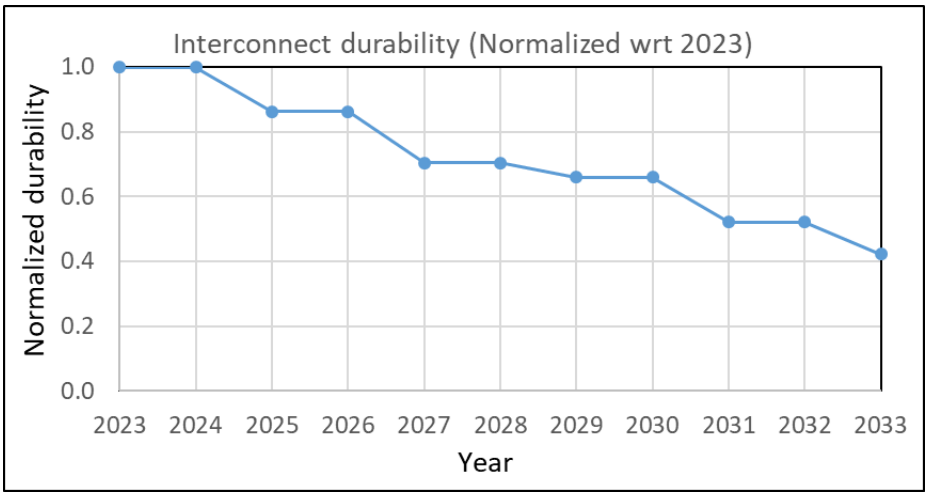


Figure 4.13: Drop in Interconnect PTC durability due to increase in package size

(normalized wrt 2023 durability)

The increase expected in package warpage (presented earlier in Figure 4.7) not only creates flexural residual stresses within the package, but also creates quality challenges for achieving uniform solder interconnects during solder reflow, as shown in Figure 4.14. The defects range from completely open joints to dimensionally distorted joints where the height of the joint is no longer the nominally expected dimension.

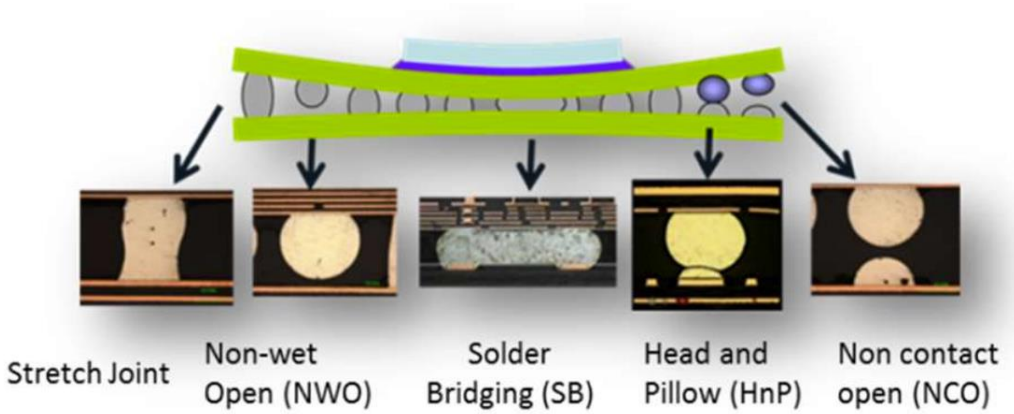


Figure 4.14. Solder quality issues caused by excessive package warpage [Source: Loh, et. al., ICEP 2016]

Solder interconnect reliability under power-temperature cycling (PTC) conditions is known to scale approximately as the square of the joint height. The drop in PTC durability due to the progressively increasing warpage of Figure 4.7, is shown on a normalized scale in Figure 4.15 for the next decade (normalized wrt the PTC durability for expected warpage levels of 2023).

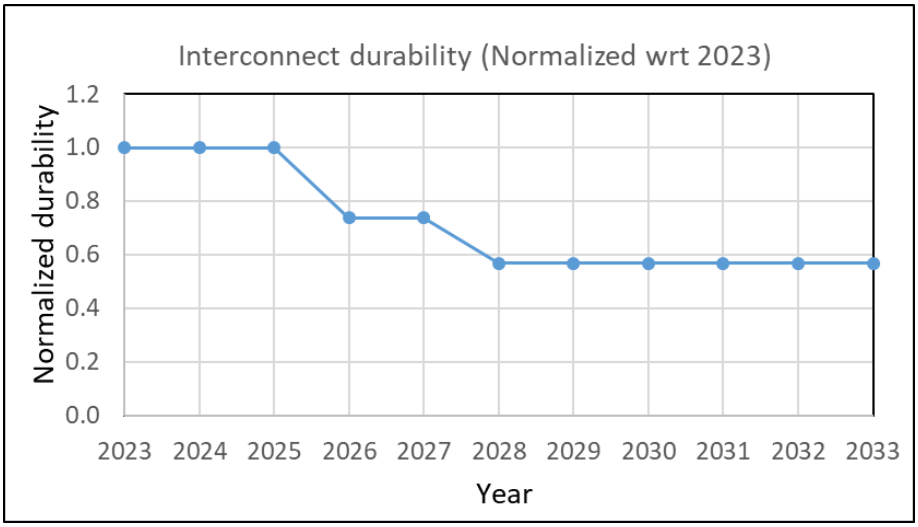


Figure 4.15. Drop in Interconnect PTC durability due to increase in package warpage

(normalized wrt 2023 durability)

4.3.2 Challenges in CPBI Simulation and Risk Management:

CPBI stresses are accumulated from different fab and assembly steps and the hard failures like white bumps are typically on the die BEOL and soft failures like transistor aging on FEOL. In order to completely understand these failures, we will need multi-process, multi-scale simulation flow. Existing commercial tools for CPBI are not adequate to simulate the transistor/circuit level. Figure 4.16 shows the desired simulation capabilities from package, bump to the circuit GDSII levels.

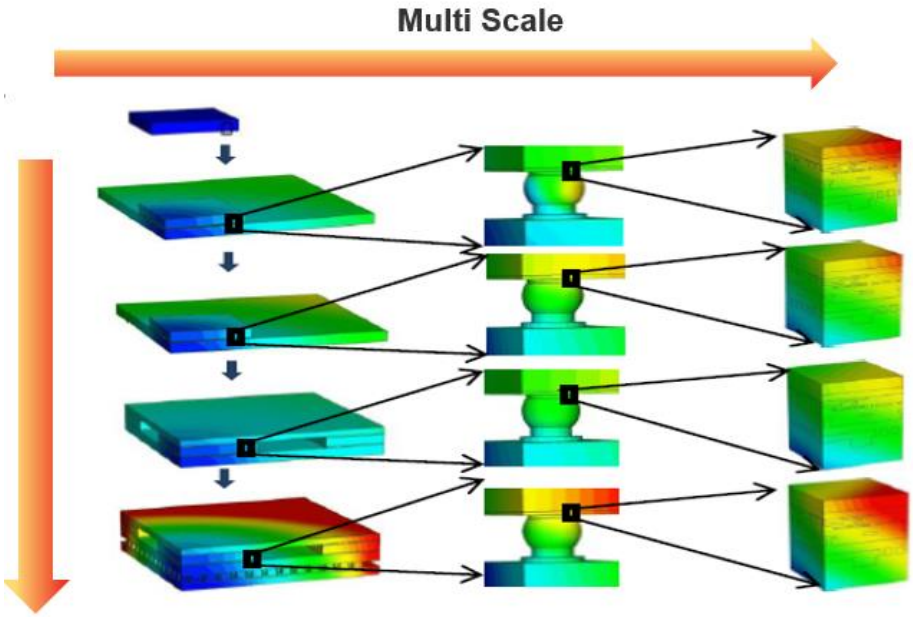


Figure 4.16. Multi-process and Multi-scale CPI simulation flow

[Source: HIR Roadmap, Reliability TWG]

With increasing system complexity and density and miniaturization in HI systems, integration of design, manufacturing and qualification processes to manage CPBI risk will become increasingly difficult. These CPBI reliability challenges include Characterization, design rules, and risk quantification (including CPBI-induced FIT rate prediction).

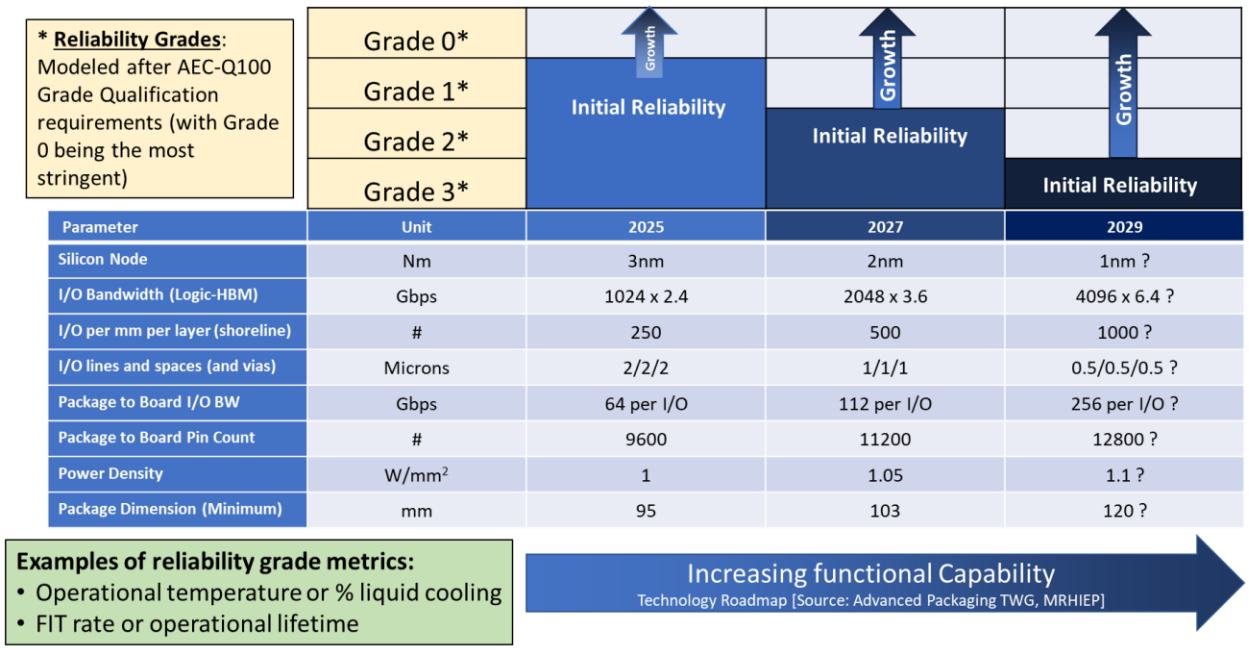
The overall difficult challenges expected in making new leading-edge HI hardware technologies reliable, dependable and affordable, can be broadly grouped under challenges during design, during manufacturing and during qualification/sustainment. For brevity, these are not discussed in detail here but can be found in the HIR Reliability Chapter.

4.3.3 Reliability Roadmap Summary:

A 3-phase timeline is provided in Figure 4.16a, to summarize the milestones that have to be achieved in the Reliability roadmap, in order to assure reliable HPC systems. Similar charts can also be constructed for Medical and wearable electronics roadmaps. Emulating the scheme used in the automotive electronics community (AEC-Q100), reliability targets are classified into 4 categories (Grades 0-3), with Grade 0 representing the most stringent reliability targets. The metrics that can be used in to differentiate between different grade levels could be based on different performance or application metrics. For example, in the HPC roadmap, reliability grade metrics can be based on:

(i) max operating temperature limits (HTOL)
(ii) ability to meet the % liquid-cooling targets presented in the thermal management roadmap.
(iii) FIT rate targets or hazard-rate targets (e.g. Grade 0 could represent FIT rate of 200, with other grades representing progressively higher FIT rate targets)
(iv) Operating life time targets, e.g. Failure free operating period (FFOP) or maintenance-free operating period (MFOP)

Figure 4.16a, represents the fact that the more advanced nodes may present greater challenges for meeting reliability targets. They may start out with a poorer initial reliability grade and may need more development efforts to eventually achieve Grade 0 classification.



Parameter	Unit	2025	2027	2029
Silicon Node	Nm	3nm	2nm	1nm ?
I/O Bandwidth (Logic-HBM)	Gbps	1024 x 2.4	2048 x 3.6	4096 x 6.4 ?
I/O per mm per layer (shoreline)	#	250	500	1000 ?
I/O lines and spaces (and vias)	Microns	2/2/2	1/1/1	0.5/0.5/0.5 ?
Package to Board I/O BW	Gbps	64 per I/O	112 per I/O	256 per I/O ?
Package to Board Pin Count	#	9600	11200	12800 ?
Power Density	W/mm^2	1	1.05	1.1 ?
Package Dimension (Minimum)	mm	95	103	120 ?

Figure 4.16a. 3-phase timeline of roadmap for achieving reliable HI systems for HPC technologies

HPC systems and data centers also present an aggressive market for photonic systems due to explosive growth in user-generated content, IoT, 5G and data-centric applications including AI. The projection is that functionality of faceplate-pluggable (FPP) photonic modules will rapidly increase from 100 Gb/s to 1.6Tb/s. The ever-increasing demand for photonic integrated circuits for data center applications is triggering scalable Si-photonics manufacturing techniques, similar to those already established for microelectronics. This co-integration can reduce power by eliminating the internal I/O functions, while co-packaging can also improve reliability and enable more cost-effective manufacturing. The projected increase in demand will justify investments in manufacturing and reliability of these advanced technologies.

4.3.4 **Gaps for Assuring Reliable HI Systems**:

The main gap is sufficient understanding and tools for combination of physics-based and data-based modeling for:

(i) co-design for reliability and accelerated qualification (acceleration models)
(ii) diagnostic capability (via built-in-testing), ability to prognosticate remaining useful life, and proactive health management based on continuous in-situ monitoring of current condition
(iii) Metrology for assessing process quality and for root-cause assessment of degradation (reliability)
(iv) Holistic methods to address the diversity of technologies and use conditions expected in medical devices, wearable electronics and in HPC applications

Simultaneous advances are needed in:

(i) analytical tools, and improved co-design methods for:
- testability

- reliability and qualification
- process modeling for assessment of:
    - Design for manufacturability,
    - Predictive assessment of manufacturing quality and manufacturing variability,
    - Effect of manufacturing quality on life-cycle reliability.

(ii) Fusion of physics-based & data-based modeling for reliability predictions and acceleration models for accelerated qualification testing

(iii) Incorporating AI and Bayesian for real-time health monitoring throughout the life-cycle

(iv) new experimental methods for materials metrology and reliability metrology

(v) environmental stress test facilities for combined-stresses testing and modeling.

Current generation tools (in US) e.g., Isograph, Relyence, Reliasoft, ANSYS-SHERLOCK, CALCE-SARA, will need significant updates and improvements to be able to address the reliability of future HI systems. Finally, better built-in testing capability is needed for accelerated testing diagnostics when qualifying complex heterogeneous-integration systems.

4.4. Reliability Summary

This section lays out the potential difficult challenges, potential solution approaches, and necessary infrastructure that we envision as important steps for establishing best practices in making future HI hardware technologies highly reliable, dependable and affordable. Particular attention has been paid to Chip-Package-Board Interactions (CPBI). This section has laid out the importance of an integrated approach towards reliable HI systems, based on strategic integration of reliability physics with powerful artificial intelligence algorithms that can leverage the unprecedented level of real-time field reliability data that is becoming available via IoT infrastructure. The business case for such an integrated approach rests in the tremendous opportunities for reducing NPI time and 'cradle-to-cradle' cost of ownership.

The reliability issues discussed in the present version of this document are relevant to the HPC technology roadmap. In future quarters, we will extend this discussion to include wearable HI systems. Furthermore, in future versions, this document will be expanded in scope to include other aspects of system reliability (including software reliability/security and dependability of human/machine interactions) for the entire HI ecosystem.

Significant portions of this section are taken from the reliability chapter of the HI Roadmap and we thank their TWG members.

4.5. Thermal

With the growth of the power density of current/future chips, their thermal management has become more challenging on both the package and system level. Although researchers/developers can be very innovative in their thermal solutions creating chips with small thermal resistance, the thermal resistance within the package itself can limit the flexibility of choosing the right cooling solution.

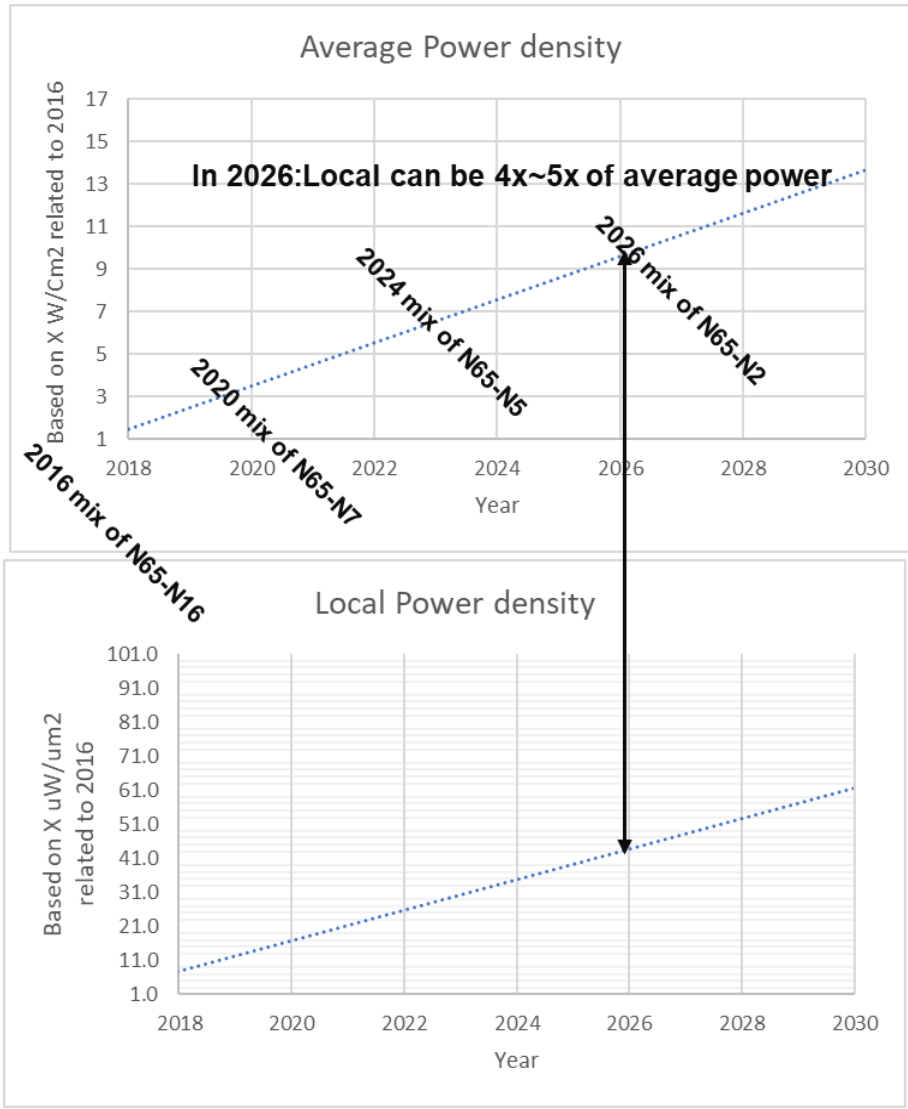
To mitigate this problem, future chips/packages should be bare/exposed die because by doing this, the package temperature drop between the die and package is reduced.

However, moving to exposed/bare die has its own challenges as TIM selection and package surface warpage become issues.

4.5.1 Performance/Cost Trend:

A Roadmap goal is to put a guideline for a future system/ chip/package prediction of performance increase as function of cost. Refai-Ahmed et al [2] stated that the performance increase can be 30-40% from one generation to another generation. However, performance improvement will be associated with a cost increase. Therefore, the incremental cost of the chip shouldn't exceed the global inflation rate. To do so, an optimum balance of technology nodes, Si architecture, package integration, lifetime, OpEx and CapEx should be done. As a conclusion, the future roadmap direction needs to have flexibility to continue the performance increase while controlling the cost.

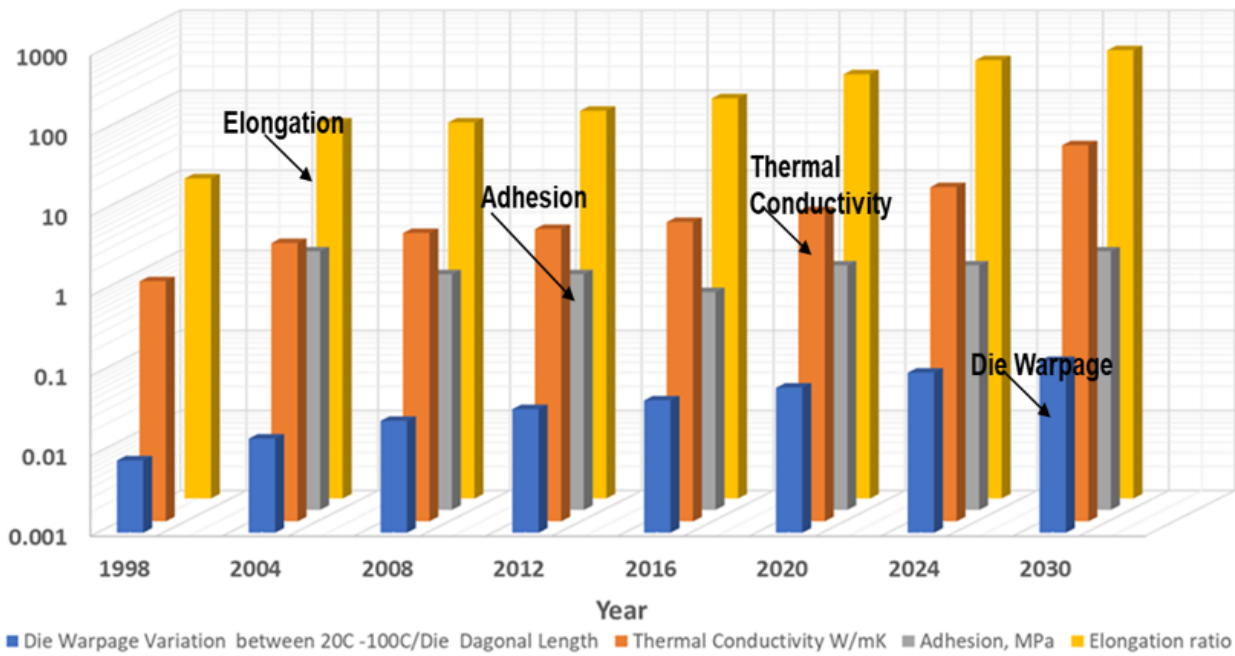
4.5.2 Package Technology Roadmap:



The industry is moving at a fast rate to increase performance through more integration of chiplets in the package. This move will increase the chip power density and thus impose some thermal challenges. Figure 4.17 shows how the power density is increasing over the years.

Figure 4.17. Power Roadmap [source Refai-Ahmed et al [2])

Figure 4.18 reveals the roadmap of the thermal interface materials. The use of thermal interface materials will continue be a crucial factor in the forward-looking thermal solution. Refai-Ahmed et al [1-2] presents challenges for thermal interface materials in Fig. 4.18.



Refai-Ahmed et al [2-3]] put these issues in perspective with respect to the thermal resistance in Fig. 19.

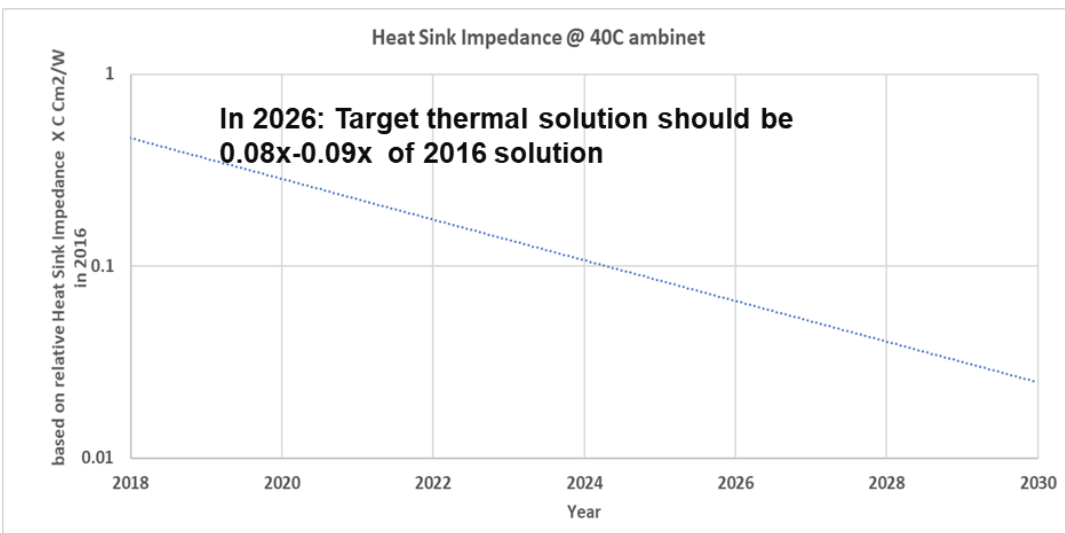


Figure 4.19-a. Power Roadmap [source Refai-Ahmed et al [2])

Figure 4.19-a shows the trend of the thermal solution referred to its value in 2016. Fig. 4.19-b reveals the thermal resistance trend target in the next 10 years graphically, as well as in a table.

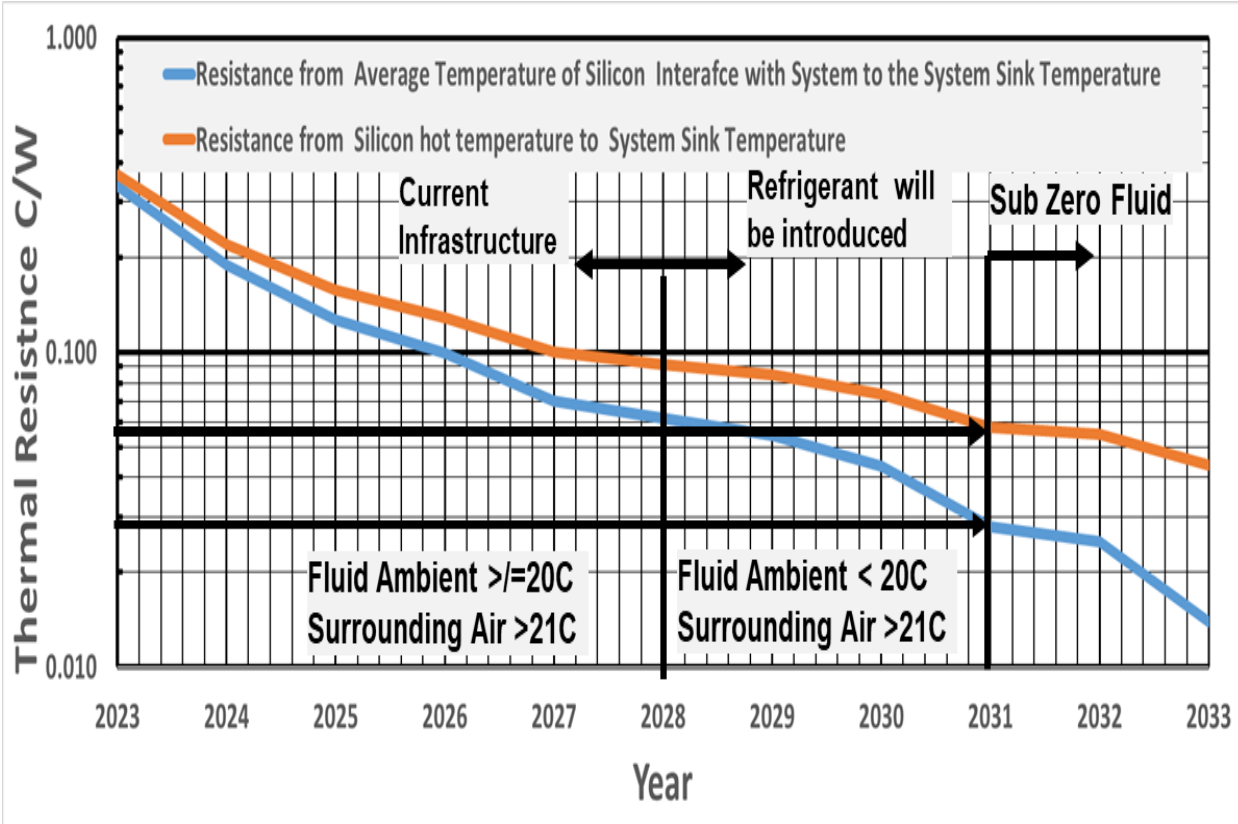


Figure 4.19 -b[source Refai-Ahmed et al [3])

Year	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033
System cooling resistance (ie heat sink)	0.3265	0.1799	0.1167	0.0890	0.0591	0.0507	0.0435	0.0322	0.0162	0.0133	0.0017
RTIM	0.0101	0.0101	0.0105	0.0105	0.0109	0.0109	0.0111	0.0111	0.0117	0.0117	0.0123
Target System resistance (Hot spot assumed to be 110C)	0.3367	0.1900	0.1271	0.0994	0.0700	0.0617	0.0546	0.0433	0.0279	0.0250	0.0140

New innovative thermo-mechanical solutions are needed to address the challenge to provide the low thermal resistance required to cool the chip.

4.5.3 Technology development for the next 15 year:

For system integrators and for overall solution management, we are imposing the following strategy that consists of 7 key principles to establish the present and future thermo-mechanical architecture for high-performance Si-package devices. The seven principles from Refai-Ahmed et al [2] are:

- Enabling and working on/around the current Silicon /Packaging manufacturing infrastructure.
- Accommodating the future heterogeneous integration of new technology such as Silicon Photonics devices.
- Utilizing the current/future infrastructure of board-level manufacturing and assembly.
- Addressing the solution from the System level with a full understanding of the component thermo-mechanical performance and behavior.
- Extending air cooling as one of the primary thermal management strategies.
- Enabling liquid cooling as the next step after the air cooling approach.
- Considering immersion liquid cooling when there is no alternative to thermal management and/or for special applications.

4.5.4 Gaps and Roadmap Solution needed

- **Gap:** Must leverage smaller intimately interconnected dies to minimize warpage and improve planarity.
- **Roadmap Solution needed:** Provide design tools, fabrication process and interconnects to support
- **Gap:** Need to support the TIM roadmap with respect to elongation, adhesion, warpage and thermal conductivity
- **Roadmap Solution needed:** High performance TIM materials development, dependent on materials companies, academic institutions for research and development; Minimize thermal resistance of TIM to support high powered devices
- **Gap:** Need to support the newest generation of data center cooling requirements
- **Roadmap Solution needed:** Develop solutions for all thermal solutions for HPC, including air, refrigerant assisted air, water, two phase and immersion cooling designs, standardize solutions
- **Gap:** Need to support the design of rigid chips on flexible substrates for medical and reduce thermal induced stress effects on sensors due to CTE mismatch
- **Roadmap Solution needed:** Enhance EDA tools to include dynamic analysis of flexible substrates, identify high stress locations and possible failure sites. New encapsulants to provide stress management of high CTE locations (interconnects locations)
- **Gap:** Temperature control for medical devices
- **Roadmap Solution needed:** Define proper limits for medical applications to include allowable temperature ranges for wearable, implantable and other medical devices. Thermal solutions may need to be bio-compatible (fluids)
- **Gap:** New solutions needed to meet goals on power density especially for 3D stacking

- **Roadmap Solution needed:** Develop new cooling solutions to meet the challenges of 3D stacking, dependent on basic research from academic institutions along with industrial process and materials companies to provide cost-effective and reliable solutions

4.6. References

[1]G. Refai-Ahmed et al., "EPTC 2021 Invited Technology Talk: Roadmap Based on Holistic Understanding of Thermo-Mechanical Challenges from Package to System to Maximize Silicon Performance," 2021 IEEE 23rd Electronics Packaging Technology Conference (EPTC), 2021, pp. 530-537, doi: 10.1109/EPTC53413.2021.9663914.

[2]G. Refai-Ahmed et al., "EPTC 2022 Invited Talk: Pathfinding to Maximization of AI/HPC/GPC System Performance," 2022 IEEE 24th Electronics Packaging Technology Conference (EPTC), 2022

[3]G. Refai-Ahmed et al., "EPTC 2023 Forward Looking View to Enable Heterogonies Integration in the Next 10 years," 2023 IEEE 25th Electronics Packaging Technology Conference (EPTC), 2023

[4]Reliability of HI Systems, Chapter 24, HI Roadmap, https://eps.ieee.org/technology/heterogeneous-integration-roadmap/2021-edition.html

[5] Ahn, J-G., "Budget-based reliability management to handle impact of thermal issues in 16nm technology," IRPS 2016.

[6] Dasgupta, A., Contributor, Integrated Circuit, Hybrid, and Multichip Module Package Design Guidelines, Editor M. Pecht, Wiley Interscience, 1994.

[7] Dasgupta, A., "Hardware Reliability," Chapter 5, pp. 95-133, in Product Reliability, Maintainability, and Supportability Handbook, Editor M. Pecht, CRC Press, 1995.

[8] Dong Xiang, Xiaolong Wang, Chuancheng Jia, Takhee Lee, and Xuefeng Guo, "Molecular-Scale Electronics: From Concept to Function," Chem. Rev., 116 (7), 4318–4440, 2016, DOI: 10.1021/acs.chemrev.5b00680.

[9] Karmarkar, A., et. al., " Modeling Copper Plastic Deformation and Liner Viscoelastic Flow Effects on Performance and Reliability in Through Silicon Via (TSV) Fabrication Processes" IEEE Transactions on Device and Materials Reliability, 19(4), Dec. 2019.

[10] Lall, P., Vaidya, R., More, V., Goebel, K., and Suhling, J., "PHM-Based Residual Life Computation of Electronics Subjected to a Combination of Multiple Cyclic-Thermal Environments," IEEE, ITherm Conference, 2010, DOI:10.1109/ITHERM.2010.5501275.

[11] Rao, R., et. al., "Effects of Various Assembly and Reliability Stresses on Chip to Package Interaction", IRPS, Hawaii, June 2014

[12] Rao, R., et. al., "Design for Reliability with a New Modeling Methodology for Chip to Package Interaction (CPI)", ECTC, San Diego, May 2015.

[13] Sun. B., Zeng, S., Kang, R. and Pecht, M., "Benefits Analysis of Prognostics in Systems," IEEE Prognostics and Health Management Conference, Macau, 2010, DOI: 10.1109/PHM.2010.5413503.
[14] Vilan A., Aswal, D. and Cahen, D., "Large-Area, Ensemble Molecular Electronics: Motivation and Challenges," Chem. Rev., 2017, 117 (5), pp 4248–4286, DOI: 10.1021/acs.chemrev.6b00595.
[15] Yu, C. K., et. al. "A unique failure mechanism induced by chip to board interaction on fan-out wafer level package," IRPS, 2017.
[16] Zhang, X., "Chip package interaction (CPI) and its impact on the reliability of flip-chip packages" Ph.D. Dissertation, UT Austin

Chapter 5: Modeling and Simulation

Authors: Benson Chan, Mary Ann Maher, Abhijit Dasgupta, Gamal Refai Ahmed

Contents

5.1 Status: Technology for modeling and simulation software systems…….……. 1

5.2 Key Issues 1

5.3 Challenges for modeling and simulation of HI systems: 2

5.4 Gaps and Roadmap Solution needed 2

The Modeling and Simulation chapter discusses all phases of package design, from circuit designs, layouts, mechanical and thermal modeling, electrical performance and the concept of Digital Twins. The complexities of advance packaging will drive more companies to adopt co-design and co-package techniques. Gone are the days of IC designers designing their devices in a vacuum and throwing this over the wall to the packaging engineers. Such techniques are not feasible with packaging techniques using 2.5 or 3d packaging. Bridge designs such as the Apple Studio Ultra utilizes a common processor (M1 or M2) that can function alone or in a bridged solutions requires that engineering understanding of all aspects of the path from one device to the other device, this includes signal transmission, power delivery, thermal concerns (hotspots) as well as mechanical understanding of materials and warpage control. EDA tools needs to be more integrated where the entire signal path of the HI package be visualized and analyzed to ensure proper function and reliability.

5.1 Status: Technology for modeling and simulation software systems

US based manufacturers and their customers have a broad range of simulation and modeling tools to assist them. CAD/EDA tool vendors from the US and Europe are working with US foundries to improve the manufacturing ecosystem. Active research in US universities in modeling and simulation is also available to US fabs and many have active collaborations with university researchers. However, infrastructure for modeling and simulation for heterogeneous integrated systems is not as organized as for IC development where design flows are well established. Modeling tools are available (FEA, flow models, thermal, electrical, power, signal integrity and others) but putting them together into coherent flows and addressing the specific needs of heterogeneous systems is a major challenge.

5.2 Key Issues

- Packaging determines performance in wearables need to co-design package + device
- Custom processes prevail in wearables leading to device + process co-design
- Heterogeneous integration for both HPC and wearables involves many diverse and new manufacturing processes and new materials.
- Some wearable applications require the HI of sensors made at multiple manufacturing facilities

- Diverse customer needs will make it hard for fabs to offer standard products how can CAD, modeling and simulation help with this problem.
- New materials will play an important role in the manufacturing of HI devices and new simulation, modeling and design tools and strategies will need to be developed for incorporating the new materials
- Fabricators of wearables must deal with flexible substrates and interconnects, also flexible and printed sensors combined with rigid ones
- Demands of HPC packaging requires simulation, modeling of designs of package structures development of materials to handle thermal, mechanical and signal integrity issues
- HPC packaging also incorporates photonic interconnect mechanisms requiring simulation in new energy domains and new links between CAD tools

5.3 Challenges for modeling and simulation of HI systems:

- HPC and wearables are very different applications from a simulation and modeling point of view
- Various applications will require different modeling and simulation flows
- Co-design tools will need to be extended to cover more than packaging/device- i.e. materials/device/electronics/package/subsystem etc
- Depending on the TRL of the manufacturing technology either top-down or bottom up flows or a combination will be required
- Materials characterization will be paramount to simulation success need to scope especially for wearables- need to characterize behavior under bending stretching

5.4 Gaps and Roadmap Solution needed

- **Gap:** Information exchange between fab and designers difficult for new processes and materials
- **Roadmap Solution needed:** PDK improvement PDKs will be important for manufacturing success- need to define/create roadmap for PDK for manufacturing processes in multiple physical domains across multiple heterogenous processes and materials. Some PDKs exist but new extensions are needed especially to support wearables
- **Gap:** Material properties and their dependence on temp, stress, aging not available in all energy domains for modeling of packaging materials and sensors- especially rigid chips on flexible substrates for wearables
- **Roadmap Solution needed:** Materials modeling - mechanical characterization for flexing, bending and characterization of new materials in general. Standards for materials testing and resulting material property data
- **Gap:** Evaluation of incompatible fabrication processes and materials esp. thermal budgets and medical constraints- i.e., implantable - biocompatibility etc.
- **Roadmap Solution needed:** Advances in fabrication modeling and characterization for new packaging processes, sequences, and materials
- **Gap:** Tools for non-linear behavior in MEMS- based systems and efficient simulations of sensors with analog electronics when not in operating range

- **Roadmap Solution needed:** Research and CAD commercialization of advances in reduced order modeling technology and commercialization of the technology so designers can use it
- **Gap:** Difference in tools/databases used by mechanical/thermal/photonics designers and electrical designers
- **Roadmap Solution needed:** Standards- CAD for cross-mechanical, electrical exchange, model exchange, manufacturing formats
- **Gap:** Design/simulation tools for optical routing and fluidic cooling tied to other tools in design process for HPC- Tools exist individually but hard to exchange results
- **Roadmap Solution needed:** Simulation/design tools coupled for cooling, thermo-mechanical-optical-fluidic effects for HPC
- **Gap:** Simulations for biosensors and packaging thermal – mechanical – fluidic-optical-tools exist in individual domains or some coupling exists but stronger bonds are needed for example mechanical-optical simulations needed. Some coupled simulations exist but more coupling still needed
- **Roadmap Solution needed:** Simulations in coupled physics domains for biosensors that allow evaluation and design in the optical/mechanical/thermal/bio signals and mechanical/thermal effects on the bio or chemical signals
- **Gap:** Co-design tools for HPC and sensors for materials/fabrication process/device/package/multiple sensor fusion/electronics/assembly/test/security co-design
- **Roadmap Solution needed:** Stakeholders need to get together from the various design tool companies to come up with co-design flows to span design space and create the couplings and data exchanges needed to support the co-design need. Electronics tools (EDA) companies need to come together with specialty tool suppliers such as those for Photonics, Sensors, materials modeling to come up with an interoperable co-design solution

Chapter 6: Modular Chiplet Packaging for an Open Chiplet Economy

Contents

6.1. Executive Summary 2
6.2. Introduction 3
6.3. Challenges with Chiplet-Based Products 3
6.4. Standards Challenges in Packaging and Assembly 4
6.5. Domain-Specific Modular Reference Architectures 11
6.6. Reference HPC SiP Functional Modules 13
6.7. Mapping Chiplet Modularity in Current HPC 16
6.8. Chiplet Physical Modularity 21
6.9. Reference Architecture Scalability 23
6.10. Discussion 27
6.11. Modular Package Designs 28
6.12. Recommendations 33
6.13. Conclusion 34

List of figures

Figure 6. 1 MRHIEP Manufacturing Roadmap Challenges. 4
Figure 6. 2 Target bandwidth density vs. available package options 10
Figure 6.3 Fugaku ARM based supercomputer schematic (left) and layout (right). 16
Figure 6.4 Key building blocks Fugaku interconnected using a Ring Bus. 17
Figure 6.5 Decomposition of Fugaku HPC node into hypothetical functional chiplets 18
Figure 6.6 NVIDIA Grace Hopper Superchip (source: NVIDIA) 20
Figure 6.7 NVIDIA Grace hopper HPC architecture (source: NVIDIA) 20
Figure 6.8 Comparison between different scaling options limited by thermal dissipation capability. 28
Figure 6.9 Different packaging form factors commonly used in HPC products. 29
Figure 6.10 Reference architecture (with chiplet sizes) implementation in 30mmx55mm interposer package. 30
Figure 6.11 implementation with large 20mmx20mm chiplets on a 55mmx70mm interposer package. 30
Figure 6.12 55mmx70mm interposer package with heterogeneous dielets (these can be compute, memory, IO etc.). 31
Figure 6. 13 Dielet golden regime for identifying optimal bond pitch and dielet size []. 33

List of tables

Table 6. 1 Selected specification of UCIe protocol. 6
Table 6. 2 Selected specifications of ODSA Bunch of wires protocol. 6
Table 6. 3 Selected specifications of SuperCHIPS protocol. 7
Table 6. 4 implementation vs packaging options for required bandwidth densities. 9
Table 6. 5 Current standards in packaging and manufacturing 10
Table 6. 6 Detailed specifications and performance of Fugaku HPC node 17
Table 6. 7 Bandwdith of components in Fugaku HPC node. 18
Table 6. 8 Per Chiplet IO bandwidth without a hub chiplet, NOC element in compute node..... 19
Table 6. 9 With a hub chiplet, NOC element in hub node. 19
Table 6. 10 NVIDIA Grace Hopper Superchip key features (source NVIDIA)..... 21
Table 6. 11 Mapping functional modules to discrete physical chiplet modules. 23
Table 6. 12 Discrete chiplet sizes to be used in decomposition. 23
Table 6. 13 Attributes of Fugaku compute chiplet 24
Table 6. 14 (a) and (b) parameters and multiplication factors considered for scaling estimations. 25
Table 6. 15 Technology centric scaling for reference HPC system. 26
Table 6. 16 Capital centric scaling for reference HPC system. 27
Table 6. 17 Comparing scaling options across process node generations. 27
Table 6. 18 Possible chiplet configurations in different packaging form factors. 30
Table 6. 19 Recommended dielet sizes over the next ten years. 34

6.1. Executive Summary

Recently, die-to-die (D2D) interface protocols for chiplets (UCIe (Universal Chiplets Interconnect Express), BoW (Bunch of Wires), Superchips) have received attention in standardization efforts from multiple organizations. For successful product development, chiplet-based products require a new integration of the supply chain, not just protocols for D2D interconnect. Unlike monolithic devices, chiplets have to be integrated with other chiplets to form a usable product. Therefore, chiplet-based designs have to be cognizant of several factors that are usually considered “back end” issues in monolithic ASIC design such as packaging, inventory and test. These factors have limited chiplet-based designs to large companies that largely control their supply chain.

In this report, we identify several gaps in standards needed to address these “backend” issues in product development that hinder the integration of chiplets from multiple vendors. We propose the development of modular architectures to close these gaps. A modular architecture can develop guardrails or budgets for die size, die-to-die bandwidth, thermals, mechanicals, packaging technology, heat dissipation and other attributes relevant to final product design and manufacture. Modular architectures will need to be domain-specific since the performance, area, power and cost requirements vary by two orders of magnitude across the various applications for chiplets.

We develop an example reference modular architecture for high-performance computing (HPC). We derive the reference architecture from the AF64 ASIC used to develop the recent Fugaku supercomputer. We show that this modular architecture with bounds on die size, bandwidth,

mechanicals and thermals can meet current HPC requirements for performance, heterogeneous integration and scale into the future. We also show that scaling can be accomplished in one of two ways - so as to preserve capital investments in packaging manufacture or to leverage advances in packaging technology. Future development for the modular HPC proposal will require the development of a complete set of standards for packaging, mechanical, thermal, power delivery and other attributes. The approach used to develop the modular HPC architecture can be extended to other domains such as automotive, medical, aerospace, IoT and other applications

6.2. Introduction

TWG3's (Technical Working Group 3) charter is to identify the open standards relevant to a quick start guide for building an advanced packaging factory for chiplet-based products in the United States. Our group includes participants from several leading semiconductors, tools, design automation, systems and hyperscale companies and institutions.

The interim report presented at the end of 2022 focused on a survey of open standards relevant to chiplets and chiplet-based products. TWG3 used the most recent IEEE HIR as a starting point for its activities. The group reviewed multiple chapters relevant to advanced packaging and vertical applications.

This report focuses on the potential for and benefits of modularity within a package. In concert with the other groups, TWG3 has decided to focus on products for which manufacturing costs are a higher percentage than ASP. That is, more expensive, higher performance, power and area products. Of necessity, this focuses on assembly with high-end packaging technologies. Our initial focus will be 2/2.5D packaging technologies. This report does not identify potential sources for these chiplets and expects that to require further effort.

6.3. Challenges with Chiplet-Based Products

Chiplet-based designs offer several advantages over monolithic designs. Because of the advantages they offer, several companies have produced chiplet-based products, indeed multiple generations of these products, in high volume.

However, they also introduce some challenges. Relative to monolithic designs, chiplet-based designs and products require "downstream" attributes in the value chain to be considered in product design. For example, designers of monolithic products rarely consider the packaging technology used in the final product.

Examples of attributes that need to be brought forward include:

- Interconnects: Chiplets need to be interconnected in order to communicate with each other. Chiplet-based designs require careful consideration of the die to die interfaces between the chiplets. These need to meet the power and performance requirements of the design. Beyond these requirements, the designer needs to choose an interconnect supported by all the other chiplets this new design may need to interoperate with.

- Packaging: The choice of interconnect also largely chooses the packaging technology used with the product. Chiplets designed for one packaging technology cannot be used with other packaging technologies. For example, chiplets designed to be used with laminate packaging cannot be used in interposer-based packages. Even within advanced packaging, chiplets designed for bridging technology cannot be used in interposer-based designs.
- Testing: A packaged product works only if all the component chiplets work correctly after insertion into the package. Chiplet-based designs can be more difficult to test than monolithic designs, the chiplets must be tested individually and then tested as a system. Designers may have to allocate more resources to support test than in monolithic products.
- Manufacture: Inventory management with chiplet-based products is more complex. To manufacture a chiplet-based product, there needs to be an adequate supply of inventory of all the component chiplets.

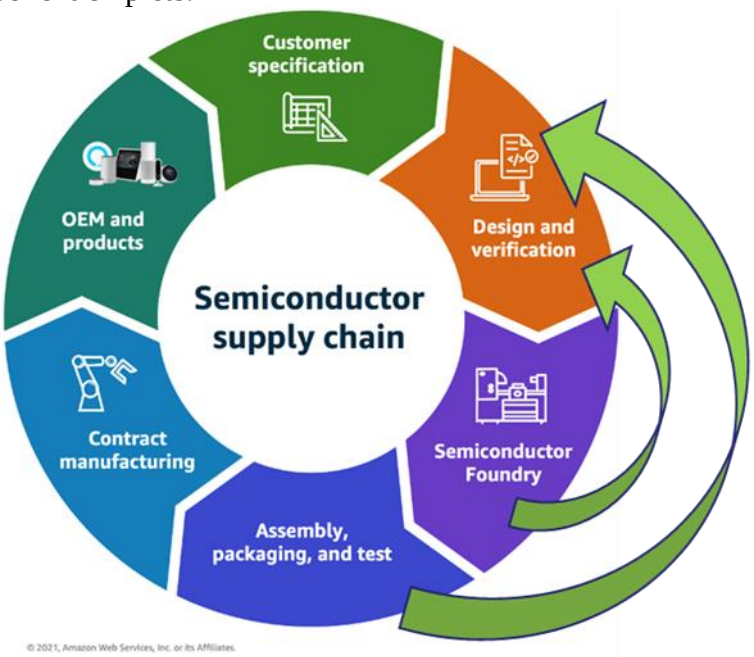


Figure 6. 1 MRHIEP Manufacturing Roadmap Challenges.

These challenges have limited the development of chiplet-based products to vertically integrated designs in large companies. The development of standards for chiplet-based designs could help to make them more widely accessible.

6.4. Standards Challenges in Packaging and Assembly

A substantial interest in logic and design standards for chiplet-based products was triggered by several government and industry activities.

6.4.1. Overview of Standards for Chiplets

Most of the recent work on standards focus on the logic integration in chiplet-based products.

D2D Protocol Standards

When a design is partitioned across multiple die, information that would normally be carried on wires inside a silicon die is instead carried between silicon die with a die-to-die protocol. D2D protocols consist of at least three parts:

1. A PHY protocol for the actual electrical transport of information
2. A link protocol for the transport of data bits between connected die
3. A map to transport common system transaction protocols such as PCIe, CXL, AXI (multiple variants), CHI and even proprietary protocols.
4. The packaging technologies supported by the D2D protocols

There has been a burst of recent activity in protocol standards for the data plane interoperation between die. Early development for D2D protocols extended serial protocols that embedded the clock in the data being transported. XSR is the most well-known serial D2D protocol and has been used in several products. Serial protocols require very few wires for data transport between connected die, but suffer from high transport latency and power consumption.

Recent industry attention has focused on clock-forwarded parallel protocols. That is, protocols in which a set of data signals are transported in parallel with a common clock signal, with 16, 32, 40 or more bits per clock signal. Data is delivered on both rising and falling clock edges (referred to as Double Data Rate (DDR)). The greater the number of data wires per clock, the more complex the protocol is to design and implement. The fewer the number of data wires per clock, the more the overhead of data transport.

D2D protocols are typically evaluated on:

1. Beachfront bandwidth density – the bandwidth density per unit element (usually 1 mm) of die edge. This is usually a function of the highest line rate per wire and bump density of the packaging technology. Bit rates per wire can range from 2 Gbps through to 16 Gbps. While variations exist, on average parallel protocols can offer a edge density of up to 1 Tera bit per second/mm in laminate substrates with regular-sized bumps and several times that with advanced packaging and microbumps.
2. The power per bit for data transport. The higher the data rate, the higher the power used to transport data bits between chiplets. Reported implemented power efficiency for parallel PHYs ranges from 0.3-0.5 pJ/bit and additional power is needed for the link and transaction layers when data is transported at 16 Gbps/data lane. High bump densities can slow data transport across a large number of bumps can make very power efficient data transport possible, asymptotically approaching the power of on-die buses.
3. The transaction protocols supported. Transaction protocols require a low-level link protocol and mappings of common transaction protocols to the link layer.
4. Interoperation constraints. Many D2D protocols fix various aspects of the PHY to ensure interoperation between various implementations.

We briefly review the significant parallel D2D protocols:

UCIe

UCIe is a recent clock-forwarded parallel D2D standard, developed by the UCIe consortium, that has received significant internal support. UCIe specifies bump maps and operating modes and is defined for both laminate and advanced packaging technologies. More information can be found

at https://www.uciexpress.org/. A UCIe PHY supports two modes (a) a CXL mode which fully specifies the transport of CXL transactions between chiplets and (b) a more popular streaming mode that does not specify the link layer or the mapping of transaction layers. AMD has announced an intent to support 3rd party logic in server processors through an outward-facing UCIe port.

PHY technology	Clock-forwarded parallel DDR PHY
Packaging technology	Laminate (100 – 130 µm bumps) Advanced packaging (25 – 55 µm bumps)
Line rates and bump maps	2Gbps – 16Gbps/data link
Transaction protocols	CXL in CXL mode Proprietary protocols in streaming mode
Other	Fixed bump maps for predictable interoperability.

Table 6. 1 Selected specification of UCIe protocol.

ODSA Bunch of Wires

The ODSA Bunch of Wires was the first parallel protocol defined to be scalable across laminate and advanced packaging, and across advanced and mature process nodes. The protocol consists of a BoW PHY Layer, a Transaction and Link Layer and mappings for several common AXI protocols. The PHY is designed to be extensible to adapt to various domains. BoW has been used as a lightweight protocol in multiple products for AI and other domains in process nodes ranging from 65nm to 5nm.

PHY technology	Clock-forwarded parallel DDR PHY 8, 16 data lanes/clock
Packaging technology	Laminate (100 – 130 µm bumps) Advanced packaging (25 – 55 µm bumps)
Line rates and bump maps	2-16 Gbps/data lane
Transaction protocols	ODSA Link Layer DiPort for AXI, other transaction protocols
Other	No fixed bump maps. Interoperability is achieved by specifying wire order at edge exit.

Table 6. 2 Selected specifications of ODSA Bunch of wires protocol.

SuperCHIPS

SuperCHIPS i.e. Simple Universal intERface for CHIPS is the first parallel hardware protocol to demonstrate cross-dielet communication at sub-10µm bond (bump) pitch with data bandwidths exceeding 2 Tbps/mm. It can be implemented on advanced wafer-scale packages such as the Silicon Interconnect Fabric (Si-IF) and Interposers which can support a sub-10 µm bond pitch. The roadmap for SuperCHIPS is to achieve 0.7 µm bond pitch and 0.48 µm wiring pitch by 2035 as described in Tables 1.1(b) and 1.1(d) of Chapter 1. SuperCHIPS is especially suited for 3D stacking applications. The specifications are described in the table below:

PHY technology	Clock-forwarded parallel SDR/DDR capable PHY with 32, 64, 128 data links/clock

Packaging technology	Advanced package which is capable of thermal compression bonding and/or hybrid bonding at ≤ 10µm bond pitch
Package layers	2 - 6
Line rates	2 - 4 Gbps/data link
Transaction protocols	Agnostic in streaming mode, can use: AXI, other transaction protocols
bump maps	No fixed bump maps. Interoperability is achieved by specifying wire order at edge exit.
Inter-dielet bandwidth	2 Tbps/mm (2023) 8-10 Tbps/mm (2035)

Table 6. 3 Selected specifications of SuperCHIPS protocol.

High-Bandwidth Memory

The most well-known parallel D2D protocol is also the most widely used. High-bandwidth memory is used to provide high bandwidth access to on-package DRAM. The HBM protocol specifies a PHY, bump maps and a memory access protocol. JEDEC has defined three generations of HBM to date.

Optical / co-packaged optics communication

Much like memory, it has been suggested that co-packaged optics, expected to be necessary for high-bandwidth applications such as AI will require custom D2D protocols. This is a nascent area in which significant change is expected in the immediate future. Please refer to Chapter 2 for a detailed description or roadmap for co-packaged optics communication

Physical Design Description Standards

Chiplet-based design requires models of the physicals of chiplets to be available in an EDA tool flow for package design. These models need to capture the size, thermal information, power distribution, dynamic behavior (power model), power and signal integrity and other attributes relevant to package design. These models need to be specified in standard to enable device manufacturers to specify chiplet models for their products and for those models to be used across tool flows from multiple vendors. Two efforts address this modeling challenge and appear to have broad industry support.

TSMC introduced the 3Dblox open standard aims to modularize and streamline 3D IC design solutions for the semiconductor industry. A 3Dblox language aims to standardize the way physical attributes of a chiplet are described for both 2.5D and 3D integration.

The Open Compute Project and JEDEC announced a joint mechanism to standardize Chiplet part descriptions leveraging OCP Chiplet Data Extensible Markup Language (CDXML) specification to become part of JEDEC JEP30: Part Model Guidelines for use with today's EDA tools.

Both approaches aim to provide a standardized Chiplet part description for automating System in Package (SiP) design and build using Chiplets.

Test Standards

Chiplets in a design need to be tested individually before insertion into the package and again within the final product. IEEE 1149 is a test access standard used to apply test vectors to packaged devices. The IEEE has developed standards to access chiplets within a packaged device.

From the IEEE website – the IEEE Std 1838 defines die-level features that, when compliant dies are brought together in a stack, comprise a stack-level architecture that enables transportation of control and data signals for the test of (1) intra-die circuitry and (2) inter-die interconnects in both (a) pre-stacking and (b) post-stacking situations, the latter for both partial and complete stacks in both pre-packaging, post-packaging, and board-level situations. The primary focus of inter-die interconnect technology addressed by this standard is through-silicon vias (TSVs); however, this does not preclude its use with other interconnect technologies such as wire-bonding.

A new standard, the IEEE P3405 is under development to standardize the test and repair of the the lanes in D2D interfaces. This will be particularly relevant with advanced packaging technologies that use wide slow buses on microbumps for data transport between chiplets.

There are still gaps for the test of chiplets before singulation at wafersort. Current practices rely on the use of sacrificial test pads that require area that is not used during regular operation.

Other Standards

Chiplet designs also require standards on telemetry, device management, reset and initialization. There are currently no standards in flight on these topics.

6.4.2. Gaps in Standards

Logic standards alone cannot address the interoperation, packaging and manufacturing challenges with chiplet-based products. The practical impact of these challenges is that a vendor who develops a chiplet may not be able to actually integrate that chiplet with those from other vendors, even if all the chiplets use the same D2D protocol standard.

We explore two sources of aggregation challenges:

1. The large number of ways in which a target D2D bandwidth may be achieved
2. The physical constraints on integrating die to form a product.
3. Tradeoffs in optimal die size

Multiplicity D2D PHY Design Options

This section is based on a presentation by Elad Alon of Blue Cheetah Automation at the ODSA D2D Interface Technical Workshop in October, 2022.

Every chiplet communicates with others through its D2D interfaces. The bandwidth required on its D2D interfaces is specified by the functionality and performance of the systems the chiplet is targeting. For example, a chiplet targeting data centers might require several terabits per second of D2D bandwidth. Whereas, a chiplet targeting embedded systems might require far less.

A chiplet designer might assume that choosing a specific D2D protocol, for example UCIe, and a target bandwidth, say 1 Tbps, would ensure compatibility with any other chiplet that also supports UCIe. In practice, this is very unlikely. The diversity in packaging and bumping options and in product cost, size and other physical requirements means that the target bandwidth can be achieved

The table below shows the number of options for a D2D link across packaging technologies.

- Ultra Wide: Wide slow short-distance connections, possible with advanced packaging and dense bumping.
- Ultra Dense: Wide fast short-distance connections, possible with advanced packaging, dense bumping and more complex circuitry
- Full Reach: Longer reach connections for devices that need to be physically separated, typically over laminate substrate.
- Cost-Optimized Full Reach: Longer reach connections that simplify substrate routing. These implementations have the potential to be used with low cost products, additive manufacturing and flexible substrates.

In practice, as shown in the figure below, this implies a target bandwidth can be achieved in multiple ways. The impact of this diversity of options on interoperability and system design is discussed in more detail below.

Implementation Type	**Package**	**Per-Line Rate**	**Termination**	**Reach**
Ultra Wide (UW)	Advanced only (< 55 µm pitch)	< 8 Gb/s	No	<4mm
Ultra Dense (UD)	Advanced only (40-55 µm pitch)	8 – 32 Gb/s	No	<2mm
Full Reach (FR)	Standard or fanout (55-130 µm pitch)	< 40 Gb/s	Supported	<25mm
Cost-Optimized Full Reach (CO-FR)	Low-layer standard (130-180 µm pitch)	< 40 Gb/s	Supported	<25mm

Table 6. 4 implementation vs packaging options for required bandwidth densities.

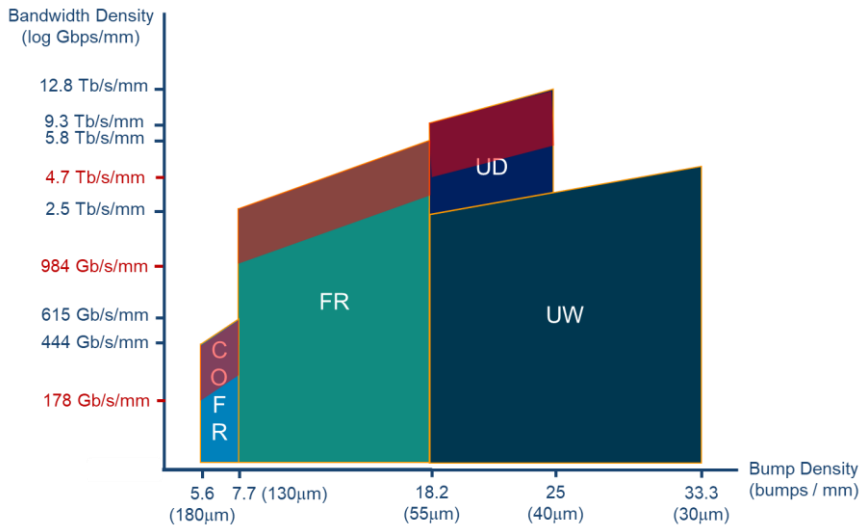


Figure 6. 2 Target bandwidth density vs. available package options

Guardrails for Device Physicals

TWG3 believes that increasing the potential for automation in packaging and assembly is key to reshoring packaging facilities. To further bound the complexity of package manufacturing and assembly, additional aspects of a product will have to be bounded.

.

- Chiplet size, package size, the number of chiplets in a package
- The maximum power per chiplet and the power for the package as a whole.
- The maximum heat to be dissipated by a chiplet and the package as a whole
- The wiring density required between the chiplets in a package.
- Off-package I/O pin count and bandwidth requirements
- The maximum bump pitch for chiplet I/O,
- The maximum mechanical stress expected on the package

Table 6.1 shows that current standards focus on logical protocols and do not address chiplet physicals.

Component	Status
D2D interconnect	UCI, BoW, Superchips, XSR
Chiplet and SiP Test	IEEE 1838, IEEE P3405
Chiplet Description for EDA	JEDEC/OCP CDXML
Chiplet and package size guardrails	Open
Bump and assembly pitch guardrails	Open
Power delivery guardrails	Open
Thermal guardrails	Open
Wiring density guardrails	Open
Mechanical guardrails	Open

Table 6. 5 Current standards in packaging and manufacturing

Placing guardrails on these aspects bounds the physical, mechanical, thermal and electrical requirements of the package and correspondingly the complexity and cost of designing and

manufacturing packaged parts. Automation in packaging and assembly can potentially be increased with guardrails on the complexity of packages and chiplets in packages for chiplet-based products. These guardrails can be used in the tooling for product packaging and assembly. Guardrails that are too tight cannot enable products that meet market requirements. Guardrails that are too loose will increase the complexity of packaging and assembly.

6.4.3 Trade-offs in Chiplet Design

The sections above outline the challenges facing a commercial chiplet designer:

- It is possible for a commercial chiplet designer to choose functionality relevant to a wide range of systems. As an example, develop a design for a large high-performance multicore CPU
- It is also possible for the designer to choose a D2D interconnect protocol popular in the market that is supported by several other chiplets and implement an interconnect that offers adequate bandwidth for the target functionality.

Unlike monolithic ASICs chiplets are not expected to be used in isolation in a package. Every chiplet needs to be aggregated with one or more other chiplets, possibly from other vendors and process nodes to form a product. Even with such considered choices, SiP designers may find it difficult to aggregate this chiplet with others.

- Every chiplet in a product has to be designed for the same packaging technology
    - Within that choice, the D2D interconnect has to be at physically compatible locations
    - The D2D interconnect may also restrict the relative orientation of two connected chiplets
- All the chiplets in a product have to coexist physically
    - The reach of the interconnect has to be enough to not create thermal hotspots
    - The power drawn by one chiplet should not impact the performance of another
- All the chiplets in a product have to share a control framework such that
    - They can either be reset/initialized individually or as a group
    - They can be monitored and operated in the field as a single logical entity
    - The product can be tested at manufacture and in the field as a single logical entity

6.5. Domain-Specific Modular Reference Architectures

Large companies address these challenges **by designing chiplets in families.** That is, a target system is partitioned into functionally-distinct modules, a common packaging technology is chosen and each module is also allocated a specific physical budget. Therefore, though each chiplet is designed individually, aggregation is simplified by the front-end budgeting process.

Today, such a budgeting process is not available for chiplets designed across companies. The time, complexity and cost of solving these challenges directly impacts the commercial viability of any chiplet. In fact, these challenges have impeded the creation of a vibrant multicompany chiplet ecosystem. Chiplet-based designs have largely been restricted to high-volume products from very large companies that largely control their supply chains.

In this report, we propose the development of modular reference architectures that create specific functional and physical modularity budgets. We propose to generate physical, mechanical and thermal guardrails by developing modular packaging and chiplet designs based on modular domain-specific reference architectures. Modularity can enable significantly more automation in packaging and assembly.

It is not possible for one modular architecture to server the entire range of chiplet applications. Therefore, these modular architectures have to be specific to a domain. In this document, as a template, we develop a reference architecture for HPC. This section is an outline of the proposal. Each of the remaining sections, develop the proposal in greater detail. We hope a similar process can be followed for other domains.

Current HPC systems are modular at the system level and consist of a collection of nodes. All the complex functionality in a node is implemented in a small collection of ASICs. Based on this modularity, we develop the following claims:

- Based on an analysis of two recent HPC systems, we claim
    - The functionality within an ASIC package can be partitioned into chiplets that implement industry-standard functionality.
    - The functional chiplets can be mapped to just two types of chiplets, one with a square aspect ratio and the other with a rectangular aspect ratio.
- We claim the scalability of a modular chiplet to advanced process nodes will be limited by heat dissipation density. Within this constraint, functionality can be scaled in one of two ways while preserving architectural stability.
    - In a technology-centric path - leveraging technology to constantly increase bump density for d2d interconnect and lower the power used for data movement
    - In a capital efficient path - preserving bump density across process nodes to enable a manufacturing line to scale across multiple process nodes.

Each of these issues is discussed in greater technical detail in the remainder of the document.

Building on these claims, we develop a proposal for a modular chiplet-based reference architecture and implementation for a HPC node. The reference architecture:

- Specifies two permissible chiplet types and two sizes per type
- Specifies the I/O and off-package bandwidth for each chiplet type and size
- Bounds the mechanical requirements of the bump maps
- Specifies a two package sizes and the mix of chiplets allowed per package type
- Bounds the power and thermal characteristics of chiplets

The modular architecture for chiplet-based designs can meet the performance and functional requirements of HPC systems and offer the following benefits

- Enable easier integration of heterogeneous architectures within a compute node
- Enable more automation in packaging and assembly by reducing design variability
- Enable faster, cheaper packaging and assembly through greater reuse of packages
- Meet requirements for multiple generations of HPC products over several years

A complete implementable specification for a modular architecture for HPC nodes requires substantial additional technical effort and is detailed in the next steps section. The path developed is a promising approach that can be extended to other verticals such as automotive, defense and aerospace, communications and medical devices.

6.6. Reference HPC SiP Functional Modules

The overall vision presented here is to have a reference HPC node with a baseline configurable HPC-oriented System-on-a-Chip (SOC), with the baseline SoC composed of functional modules. The SoC includes cores that run a general-purpose OS, such as Linux, a memory subsystem, standard peripheral interfaces such as PCIe/CXL, and the NIC. The designer can then customize the SOC by incorporating chiplets that are specific to the protocol for the NIC or accelerators that plug into the Network on Chip (NOC) fabric, which stitches all these elements together. Both heterogeneous and homogeneous HPC systems may be built with this reference architecture.

By having a baseline SOC, the designer can reduce development time and cost while ensuring compatibility and interoperability with existing standards-based peripheral interfaces. The specialized chiplets can be added as needed, providing flexibility and customization options for different applications.

This approach allows for more efficient and effective development of complex systems, as the designer can focus on the specialized elements that are most critical for the application. The NOC fabric provides high-bandwidth, low-latency connectivity between the different chiplets, allowing for seamless integration of different components.

Overall, the vision of a baseline configurable SOC with specialized chiplets is a promising approach that could help streamline development and improve performance and efficiency for a range of applications.

The baseline system consists of the following chiplet modules

- A compute subsystem
- A memory subsystem
- A network I/O subsystem
- Other miscellaneous subsystems
- A NOC that integrates all the subsystems

6.6.1. CPU Modules

While light-weight cores are more efficient for HPC workloads, they are often less efficient for running operating systems. To enable a trade-off between these conflicting requirements, the baseline package can be specified to include both types of cores.

One possible approach is to incorporate Fat Linux-capable cores, such as Arm Neoverse or RISC-V, in the baseline. These cores could be clusters of 2 or 8 cores per chiplet module that integrate with the AXI or AMBA NOC. The package may also include an internal NOC that connects to the AXI or AMBA NOC.

In addition to the Fat Linux-capable cores, the baseline could also include a set of light-weight cores optimized for HPC workloads. These cores could be lightweight, with 8-16 or even more per chiplet, and include a shared L2 cache. The baseline may also include an LLC module that connects to the different chiplets to provide a high-speed cache that can be shared among multiple processing elements.

By including both types of cores in the baseline, designers can create a flexible system that can be optimized for different workloads. The Fat Linux-capable cores can be used for running operating systems and performing general-purpose computing tasks, while the light-weight cores can be used for running HPC workloads that require high-performance computing capabilities.

Overall, the baseline design that includes both Fat Linux-capable cores and light-weight HPC-optimized cores can enable a trade-off between efficiency and performance. This approach can help create more flexible and efficient HPC systems that can be tailored to meet the specific requirements of different applications.

6.6.2. GPGPU Modules

GPGPU modules, as the name indicates, implement a programmable general-purpose vector-based processing in architectures derived from GPUs.

6.6.3. Accelerator Modules

Accelerator modules implement application-specific hardware functions, either in programmable hardware, such as an FPGA or in ASIC.

6.6.4. Memory Subsystem

Memory requirements vary considerably across HPC systems. Memory can be external to the package on the node, or internal to the baseline system. For external memory, multiple types need to be supported.

A system architecture that flips between DDR and on-package memory can be challenging to design, but the objective is to create a modular system that can enable different kinds of memory. To achieve this, a module can be created that goes from a standard NOC interface, such as AXI or AMBA, to a DDR memory controller that is complete with the Phy to the DDR DIMMS.

Similarly, modules can be designed that perform the same function as above but target different types of memory. For example, a module that targets HBM (on or more of the variants HBM 2, 2e, 3 or a future revision) or one that targets NVRAM, which may have multiple targets.

In addition to these memory modules, another module can be designed that contains a Last Level Cache (LLC) that connects seamlessly to one or more of those memory modules, potentially as peers on the NOC. This module can help improve performance and reduce latency by providing a high-speed cache that can be shared among multiple processing elements.
Overall, a modular approach that enables different kinds of memory can help create more flexible and efficient systems. By designing modules that can connect to different types of memory and seamlessly integrate with other components, designers can create systems that are optimized for their specific application requirements.

6.6.5. I/O Modules

The HPC community has recently converged on an Ethernet-based solution for HPC systems, with Ethernet serving as the layer 1 link interface. However, the hardware protocol for managing congestion can differ significantly from the Ethernet PHY, which makes it important to have a modular NIC design. This would allow for a common PHY across implementations, while enabling different chiplets to be plugged in to implement the specific protocol.

One possible approach to achieving this is to have an Ethernet chiplet that implements the basic packet engine and physical interface. This chiplet could be designed to include a standard TCP/IP or RDMA/TCP protocol interface for standard Ethernet operations. Additionally, third-party chiplets could be designed to plug in as alternative protocol interfaces, such as the HPE/Cray SlingShot protocol interface or the Broadcom “HPC-Ethernet” protocol interface.

By designing the NIC as a modular package, with a common PHY and pluggable protocol interfaces, HPC designers can create a flexible system that can be optimized for different applications and network requirements. This approach can help ensure that the HPC system is future-proof and can accommodate changes in network requirements as they evolve over time.

Overall, the modular NIC design can enable more efficient and cost-effective HPC systems that are tailored to meet the specific requirements of different applications. By leveraging common hardware components and pluggable protocol interfaces, designers can achieve greater flexibility and agility in their system designs, while reducing the overall complexity and cost of the system.

6.6.6. Miscellaneous Module: NOC Module

The multiple modules in the baseline system need to be connected through a network, usually based on AXI, CHI or a similar protocol. Two approaches are possible.

- The components of the network may be physically distributed across all the modules. Each module comes with a specific degree of fanout. For example, a compute module may have two interfaces to connect to other modules. The connections between modules are made in the package.
- The package supports a NOC chip that is a central connectivity node for all the chiplets. Some of the services listed above may also be provided by the NOC chip. In this approach, every chiplet is connected only to one other chiplet, a NOC chip. Large designs may require multiple NOC chips.

6.7. Mapping Chiplet Modularity in Current HPC

HPC systems have largely evolved into a node-based architecture. A modern HPC system is a collection of several thousands of nodes, integrated across a high-speed network. These nodes may be homogeneous (i.e. identical) or heterogeneous (with different types of logic). In systems with heterogeneous nodes, nodes of one kind may be aggregated into common physical racks. In this section, we demonstrate how HPC nodes can be decomposed into chiplets that implement distinct commercially-identifiable functions.

A node is a board in a specific form factor that contains some combination of general-purpose compute, accelerators, memory and networking I/O. In heterogeneous systems, the proportion of these components varies across the nodes in the system. For example, in a recent Heterogeneous European Computer:

- Every node type has general purpose CPUs, the accelerator nodes just have a pair
- Each node has a power cap of 800 W
- Network (Infiniband) I/O bandwidth of 100 or 200 Gbps
- Higher bandwidth intra-node network only for accelerator cards - a budget of 2 Tbps

We explore two recent examples of supercomputer nodes. In one, we map the monolithic ASIC used in each node onto a potential chiplet-based implementation. The second is already chiplet.

6.7.1. Fugaku supercomputer

Fugaku is an ARM-based supercomputer known for its exceptional performance. Fugaku's architecture uses a custom ASIC in each node. As shown in Figure 6.2, the ASIC consists of three main components: a compute complex, a NOC (Network on a Chip), and I/O (Input/Output) capabilities. The compute complex in each ASIC consists of four powerful clusters, each coupled with an in-package High Bandwidth Memory (HBM). The NOC, shown in Figure 6.2, is a high-speed bidirectional ring, enabling efficient communication between the nodes. In terms of off-package and off-node I/O, Fugaku is equipped with high-speed interfaces that facilitate connectivity with other nodes on a custom protocol Tofu. The other I/O includes PCIe for off-system I/O and interrupt I/O. Table 6.2 and Table 6.3 represent the specifications of Fugaku

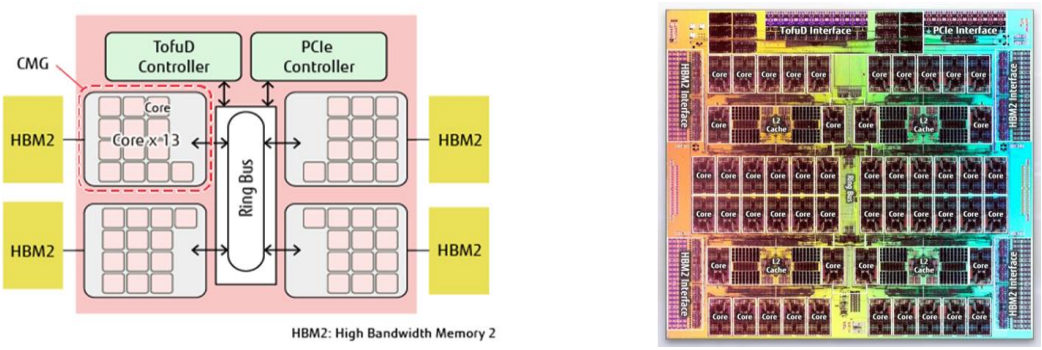


Figure 6.3 Fugaku ARM based supercomputer schematic (left) and layout (right).

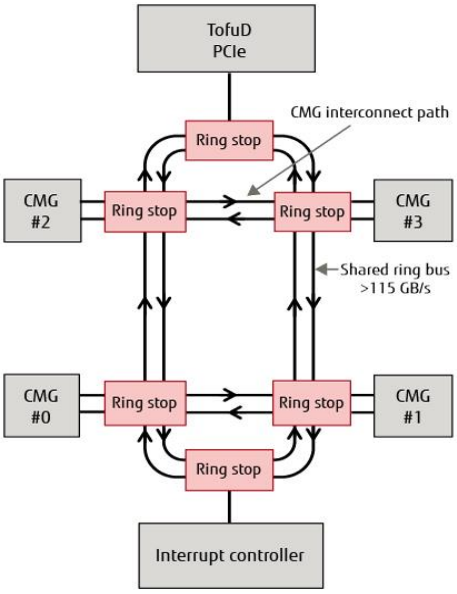


Figure 6.4 Key building blocks Fugaku interconnected using a Ring Bus.

Attribute	Detail
Die size/Power	400 mm^2/122 W/0.3 W per mm^2
Aspect ratio	20x20
Cores/ASIC	52
Core Area/Power	8.3 mm^2/2.2W
Fab/Process	TSMC/7 nm Finfet
ASICs/Node	2
Nodes/System	158,976
System power	40 MW
Tofu Area/Power	25 mm^2/9W

Table 6. 6 Detailed specifications and performance of Fugaku HPC node

I/O Type	Desc	BW per Comp Node	Desc	BW per HPC Node
HBM2	1024 bits	256 GiB		1024 GiB
PCIe			16 lanes/8 gbps per lane	16 GiB

Tofu			20 lanes/28 gbps per lane	68 GiB
RingStop CMG I/O	2 x 57 GBps Bidi	114 GiB		
Ring stop	4x115GBps Bidi	460 GiB	6 x NOC Elements	2760 GiB

Table 6. 7 Bandwdith of components in Fugaku HPC node

We develop hypothetical implementations of each Fugaku node as chiplets based on information from Table 6.2 and Table 6.3. We consider two possible decompositions.
In the first decomposition, there are no explicit NOC chiplets. Instead, each compute node incorporates a portion of the NOC functionality. Figure 6.4 shows the potential decomposition into chiplets. This approach allows for distributed communication within the node, and the required functionality and I/O for each chiplet are captured in a table.

The second decomposition involves an explicit NOC/hub chiplet, to which all the other functional chiplets connect. This central NOC/hub chiplet facilitates communication between the different chiplets. Again, the functionality and I/O requirements for each chiplet are captured in a table 6.4 and table 6.5.

Overall, these two decompositions present different strategies for organizing chiplets within a Fugaku node, impacting system scalability and the interconnect bandwidth required between chiplets. The specific types of chiplets needed, along with their functionalities and interconnect bandwidth, would depend on the chosen decomposition approach.

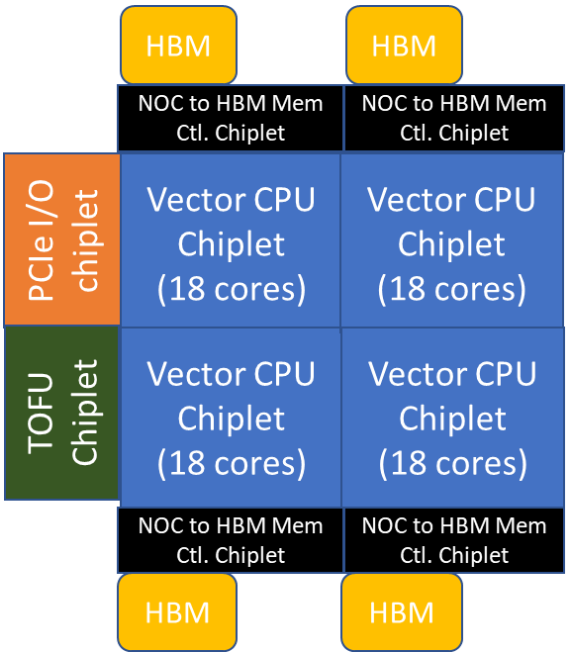


Figure 6.5 Decomposition of Fugaku HPC node into hypothetical functional chiplets

Per-Chiplet I/O Bandwidth

Chiplet	Functionality	D2D I/O bandwidth/ chiplet	Pkg I/O bandwidth
Compute	1 x CMG + 1 x Ring Stop	3x115GBps for RS 256GBps for HBM	None
Package I/O1	HBM2 interface	256GBps for HBM	1x HBM
Package I/O2	PCIe	1x57 GBps for RS	16 GBps for PCIe
Package I/O3	TOFU high-speed network	1x57 GBps for RS	68 GBps for Tofu
Package I/O4	Interrupt controller	1x115 GBps for RS	

Table 6. 8 Per Chiplet IO bandwidth without a hub chiplet, NOC element in compute node

Chiplet	Functionality	D2D I/O bandwidth	Pkg I/O bandwidth
Compute	1 x CMG	2x57GBps for RS 256GBps for HBM	None
NOC/Hub	2K x RS (even)	2K x2x57GBps for RS 6*2*57GBps for RS	None
Package I/O1	HBM2 interface	256GBps for HBM	1x HBM
Package I/O2	PCIe	1x57GBps for NOC	16 GBps for PCIe
Package I/O3	TOFU high-speed network	1x57GBps for NOC	68 GBps for Tofu
Package I/O4	Interrupt Controller		

Table 6. 9 With a hub chiplet, NOC element in hub node.

6.7.2. Nvidia Grace Hopper

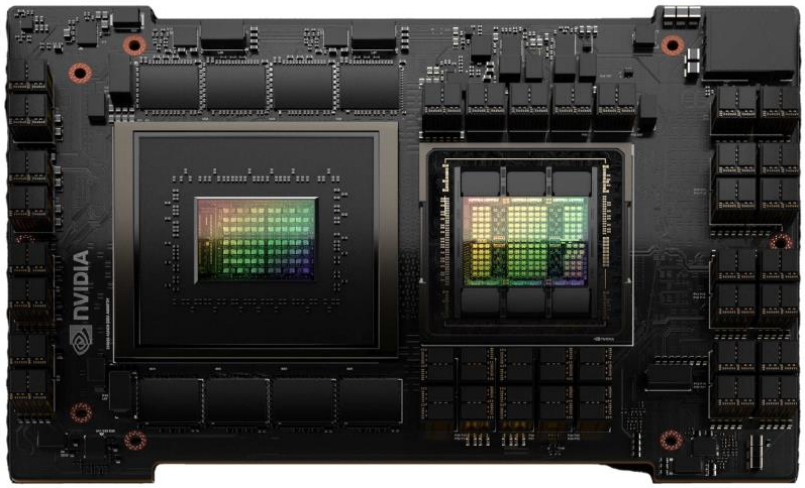
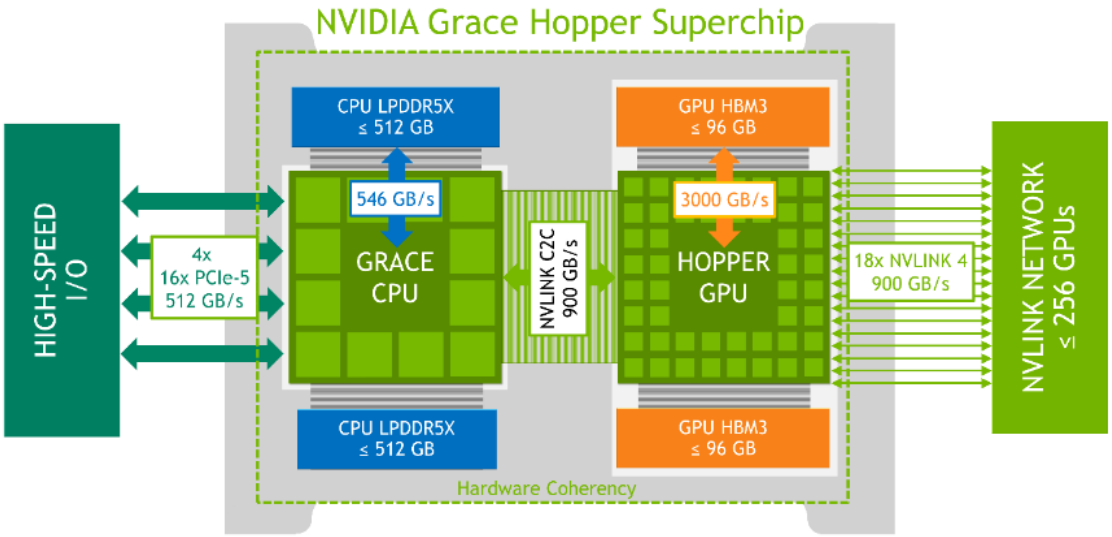


Figure 6.6 NVIDIA Grace Hopper Superchip (source: NVIDIA)



NVIDIA Grace Hopper Superchip Logical Overview

Figure 6.7 NVIDIA Grace hopper HPC architecture (source: NVIDIA)

Grace CPU die area is estimated to be around 600 mm^2. Hopper GPU die area is estimated to be around 800 mm^2.

Feature	Description
Grace CPU cores (number)	Up to 72 cores
CPU LPDDR5X bandwidth (GB/s)	Up to 546 GB/s
GPU HBM3 bandwidth (GB/s)	3 TB/s
NVLink C2C bandwidth (GB/s)	900 GB/s total, 450 GB/s per direction
CPU LPDDR5X capacity (GB)	Up to 512 GB
GPU HBM3 capacity (GB)	Up to 96 GB
PCIe Gen 5 Lanes	64x

Table 6. 10 NVIDIA Grace Hopper Superchip key features (source NVIDIA)

6.8. Chiplet Physical Modularity

The set of diverse HPC functional modules can be mapped to a small set of physical modules. This does not imply that the chiplets themselves will be identical in functionality. Only that the physical variation across chiplets implementing these functional modules can be limited. That is, the dispersion of size, thermals, power requirements, interfaces and other attributes important to package design. To derive physical constraints on chiplets in the baseline system, we first have to bound the physical characteristics of a reference HPC node.

6.8.1. Node Physical Constraints

We assume the following characteristics for each node in an HPC system.

1. The physical form factor of all the nodes is expected to be identical. It may be derived from industry standards, such as standards for modular computing from the Open Compute Project.
2. A power cap of 800-1000 W per node. Power delivery is typically not a limiting factor. The max power at a node is limited by the ability to remove heat.
    a. Usually 100 W/cm^2 can be cooled by forced air
    b. 400 W/cm^2 and above requires liquid cooling.
    c. Expensive two-phase cooling methods are required at ranges of 1000 W/cm^2 and above.
    d. Liquid Cooled 400 W/cm^2. Serviceability of immersion cooling systems is a challenge. Need to drain a machine to upgrade a board.
    e. This power cap cannot grow significantly, so future systems will have to demonstrate gains in power efficiency.
3. Network I/O bandwidth
    a. Every node will have 400/800 Gbps of I/O bandwidth. The network will be based on Infiniband or Converged Ethernet. The bandwidth out of a node is expected to double every 4 years.
    b. Accelerator nodes may have higher additional bandwidth to other accelerator cards of the same type. We budget up to 2 Tbps for this

4. Each node will consist of 1-2 packaged parts, each of which implements the baseline functionality, or some variant). Every node type is expected to have a few general purpose CPUs. In heterogeneous systems, the distribution varies by node type.
5. We set the max power budget per package at 500 W. The Number of packages per node is going to be limited by the hottest package. We are not going to limit the number of packages per node.

6.8.2. Modular Chiplets Proposals

Diverse functional nodes can be mapped onto a limited number of physical chiplet types. Inspecting the range of functional modules, one can observe, there are two classes of functional nodes:

- nodes with no off-package I/O: All their I/O is to other chiplets in the package and potential examples include compute, accelerators, and some peripheral functions.
- nodes with off-package I/O. All their I/O to both and potential examples include network, memory, NOC and PCIe modules.

With this classification, all functional modules can be mapped to two types of physical modules:

- Square: The square chiplet has a low beachfront to internal area ratio, which makes it ideal for nodes with no off-package I/O. All communication is limited to neighbors within the package.
- Tall/thin. The tall/thin chiplet also has a low beachfront to internal area ratio, but it is ideal for nodes with off-package I/O as it allows for communication with neighbors and off-package I/O.

Modules can also be potentially restricted in size, but too few sizes can lead to wasted area and/or limit target functionality. We propose that the baseline design should start with support for two sizes. The specifics of the sizes will need to be derived from the target functionality of each node and the physical constraints outlined above.

Table 6.7 shows functional modules of various types can be mapped to the two types of physical modular chiplets. We originally envisioned that the baseline system would need to support four types of chiplets. Further analysis showed that two types were adequate.

Chiplet Type		Functions
Dense Logic	Large	Accelerators GPU Heavy Cores
	Small	Security Manageability Light cores
Sparse Logic	Large	NIC Host controller

		NOC+I/O Memory I/F
	Small	Memory Controller NVRAM Optical I/O Bridge

Table 6. 11 Mapping functional modules to discrete physical chiplet modules.

	Square	Tall-Thin
Small	Dimensions: 11 x 11	Dimensions: 4 x 11
Large	Dimensions: 20 x 20	

Table 6. 12 Discrete chiplet sizes to be used in decomposition.

Within a fixed chiplet size and aspect ratio represented in Table 6.8, we propose that several other significant attributes will vary as the reference architecture varies across generations. These include, the total power per chiplet, the die-to-die IO bandwidth and correspondingly the area left for logic. In the next section, we assess how these attributes scale with process node technology.

6.9. Reference Architecture Scalability

In this section, we will assess the ability of the reference architecture to scale across process nodes. Scalability is an important factor that determines whether the industry can utilize this architecture for multiple generations of products. To evaluate this, we will analyze the migration of fixed-size compute chiplets across various technology nodes from TSMC. This analysis will serve as an indicator of the reference architecture's ability to scale effectively.

We use the compute chiplet in the hypothetical Fugaku demonstration developed in 7nm as the reference starting point for the analysis. As we progress to more advanced nodes, we expect an increase in the number of compute cores per chiplet. However, this growth is constrained by heat dissipation limits. As the number of cores grows, it becomes crucial to also enhance the chiplet's D2D (Die-to-Die) bandwidth, ensuring that the bandwidth per core remains roughly constant and the architecture remains balanced as it is scaled across process nodes.

In our analysis, we will explore two different options for achieving scalability.

- The first option is a technology-centric approach, which involves constantly advancing the packaging technology used in conjunction with the reference architecture to maximize the performance achievable.
- The second option is a capital-centric approach, which aims to maintain the viability of a packaging and assembly plant across multiple generations of the reference architecture. This approach recognizes the importance of long-term sustainability.

The rest of this section is organized as follows. We detail the assumptions we have made during our evaluation. The results obtained from both the technology-centric and capital-centric approaches and conclusions that can be derived from this analysis.

6.9.1. Reference Baseline Chiplet

The analysis estimates the performance of a reference chiplet as it is scaled across multiple process nodes. The reference chiplet is a single Fugaku compute chiplet in the hypothetical decomposition discussed in Section with the following attributes as shown in Table 6.9.

Attribute	Value
Size	11x11 mm^2
IO Bandwidth	8000 Gbps
Process node	7 nm
Power density	0.3 W/mm^2
Core area	8.3 mm^2
Core power	2.2 W
D2D line rate	16 Gbps
D2D area, power	From standards
Number of compute cores	13
Bandwidth per core	615 Gbps

Table 6. 13 Attributes of Fugaku compute chiplet

For a package with four compute chiplets (with the attributes listed above) and I/O chiplets, the performance and power are approximately consistent with the per package characteristics of Fugaku.

6.9.2. Scaling Constraints

For our analysis, we constrain how the compute chiplet is scaled as follows.

1. We assumed a fixed size chiplet, 121 mm^2 based off the reference chiplet.
2. Across multiple generations of process nodes from TSMC,
    a. Feature size 5 nm, 3 nm, 2nm and 1 nm.
    b. For each successive generation feature sizes shrink and power dissipation decreases.
    c. We assume that in each generation, the power density can improve by 10%
    d. We also assume that bump pitch can decrease by about 30%
3. Compute cores are assumed to scale across process nodes as follows.
    a. To account for scaling inefficiencies, the area of a core scales linearly with feature size, not the square
    b. The power consumption per core scales is linear with feature size. We assume increases in core size are offset by improvements in power management.
4. We assume D2D power and area are largely process node independent and largely defined by bump pitch.
    a. That is, we assume the circuit area is less than the bump area. We use reference bump maps for the Bunch of Wires to estimate the area for a D2D link.
    b. We also assume that power/bit increases with the line rate as shown in Table 6.10.

Process and core parameters						
Units	Metric	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4
	Process Node	7	5	3	2	1
	Relative Power	1	0.8	0.7	0.75	0.75
W/mm^2	Power density	0.3	0.36	0.432	0.5184	0.62208
mm^2	Die area	121	121	121	121	121
W	Die power	36.3	43.56	52.272	62.7264	75.27168
mm^2	Unit Core Area	8.30	5.93	3.56	2.37	1.19
W	Unit Core Power	2.2	1.76	1.232	0.924	0.693

(a)

D2D Parameters	
Line rate	Power (j/bit)
2	2.50E-13
4	3.00E-13
8	3.50E-13
16	4.00E-13
24	8.00E-13
32	1.40E-12

(b)

Table 6. 14 (a) and (b) parameters and multiplication factors considered for scaling estimations.

6.9.3. Scaling Performance Estimation

We estimate the impact of scaling as follows. For each process node, it is expected that the number of cores will grow from the previous node. However, the growth in both core count and D2D IO bandwidth is limited by power consumption or available physical area (listed in the scaling constraints above).

As the total number of cores increases, the direct-to-direct (D2D) input/output (IO) bandwidth must also grow to accommodate the higher workload. The goal is to achieve architectural balance, preserving the bandwidth per core of the reference architecture. To estimate the area and power required for D2D IO, a given total D2D IO bandwidth is taken into account and subtracted from the die size. The remaining space on the die is then allocated for logic components. The scaling process involves iterating on the IO bandwidth and adjusting the growth in D2D bandwidth to maintain architectural stability, as measured by the IO bandwidth per core.

The optimal point is reached when there is a balance between the number of cores, D2D bandwidth, and the cooling capability of the system. The number of cores can be constrained either by the available area or by power limitations. In cases where power is the limiting factor, not all of the die area can be utilized for the core logic or D2D interconnect.

6.9.4. Technology-Centric Scaling

Currently, the connectivity between D2D (direct-to-direct) components is less dense compared to on-die wiring. It has led to the development of D2D protocols in which the line rate between chiplets is higher than the on-chip data transfer rate. The higher the data rate, the higher the energy expended in data transfer. It is expected that technological advancements will lead to increased bump density in the future. With higher bump and wire density, slower D2D links can be utilized, resulting in power savings during data transmission. This, in turn, allows for more power allocation to the cores and logic components.

Building off the reference chiplets, a projected performance table is provided for a fixed 11x11 die, building upon the reference 7 nm chiplet. Table 6.11 lists the expected performance metrics or specifications of the chiplet, serving as a reference point for further analysis or evaluation.

Technology-Centric Roadmap						
Units	Metric	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4
Gbps	Die IO Bandwidth	8,000.00	11,979.64	18,971.75	27,773.67	46,386.91
mm	D2D Bump Pitch	0.055	0.0385	0.02695	0.018865	0.0132055
Gbps	D2D Speed/Lane	16	16	8	4	2
mm^2	D2D Area	6.72	4.83	7.51	10.72	17.55
W	D2D Power	6.40	9.58	13.28	16.66	23.19
mm^2	Area for Cores	114.28	116.17	113.49	110.28	103.45
W	Power for Cores	29.90	33.98	38.99	46.06	52.08
	# Cores - Area Limited	13	19	31	46	87
	# Cores - Power Limited	13	19	31	49	75
	# Cores	**13**	**19**	**31**	**46**	**75**
	%age Area used	94.73	97.09	97.34	100.00	88.00
Gbps	Bandwidth/Core	615.38	630.51	611.99	603.78	618.49

Table 6. 15 Technology centric scaling for reference HPC system.

6.9.5. Capital-Efficient Scaling

Even with a functionally stable reference architecture, the process of changing bump sizes necessitates an ongoing investment in packaging and assembly lines. An alternative approach is to maintain a constant bump size and line rate across multiple generations. Relative to using advanced bumping, this approach consumes more power for D2D connections. However, it offers the potential advantage of being capital efficient. With this method, a single packaging and assembly line could potentially serve multiple generations of products. This approach can also help chiplet designers. Designers can adopt specific form factors, bump maps, and sizes that remain stable for an extended period. For example, the stability of the physicals of High Bandwidth Memory (HBM) has allowed several designers across several generations of product to plan product physicals around HBM, as can be seen in the Grace Hopper chiplets as an example.

By maintaining this stability, it becomes feasible to leverage existing infrastructure and resources for a longer duration. In this context, a Table 6.12 is provided to illustrate the projected performance of a fixed 11x11 die, which builds upon the reference 7 nm chiplet.

Capital Efficient Roadmap						
Units	Metric	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4

Gbps	Die IO Bandwidth	8,000.00	11,200.00	17,737.07	25,966.16	39,357.33
mm	D2D Bump Pitch	0.055	0.055	0.055	0.055	0.055
Gbps	D2D Speed/Lane	16	16	16	16	16
mm^2	D2D Area	6.72	9.24	14.70	21.41	32.33
W	D2D Power	6.40	8.96	14.19	20.77	31.49
mm^2	Area for Cores	114.28	111.76	106.30	99.59	88.67
W	Power for Cores	29.90	34.60	38.08	41.95	43.79
	# Cores - Area Limited	13	18	29	41	74
	# Cores - Power Limited	13	19	30	45	63
	# Cores	**13**	**18**	**29**	**41**	**63**
	%age Area used	94.73	100.00	100.00	100.00	88.45
Gbps	Bandwidth/Core	615.38	622.22	611.62	633.32	624.72

Table 6. 16 Capital centric scaling for reference HPC system.

6.10. Discussion

Table 6.13 compares technology-centric and capital-efficient scaling. The number of cores increases in both scenarios and both approaches use die area efficiently. Within our simple model, in advanced nodes, in both approaches power becomes a limiting factor, in terms of dissipation density for the logic components. As mentioned before, with advanced bumping, faster D2D interfaces may not provide significant advantages, as their primary benefit lies in saving area at the cost of increased power consumption for data delivery. The real roadmap choice revolves around choosing between wide, slow interfaces enabled by advanced technologies and preserving design and manufacturing stability. The optimal solution is expected to lie somewhere between these two extremes, striking a balance that considers both capital efficiency and performance requirements.

Attribute	Technology-Centric Scaling	Capital-Efficient Scaling
Die area	121 mm^2	
Bump density	Increases	Constant
D2D line rate and power	Decreases	Constant
D2D power and area	Higher area	Higher power
Core count	Increases, Higher by 20%	Increases
Area efficiency	High till terminal node	
Terminal limiting factor	Thermal power dissipation	

Table 6. 17 Comparing scaling options across process node generations.

From the Tables the reader may notice that both scaling options are terminally limited by the ability to dissipate power. The analysis assumed that thermal dissipation density can improve by 20% per generation. If the 20% improvement goal cannot be achieved, core count growth will be further constrained. The figure 6.7 below compares three options:

1. Technology centric scaling, assuming 20% thermal improvement per generation (Blue)
2. Capital centric scaling, assuming 20% thermal improvement per generation (Red)
3. Technology centric scaling, assuming 10% thermal improvement per generation (Yellow)

All the benefits of advanced bumping can potentially be lost if thermal efficiency is not improved.

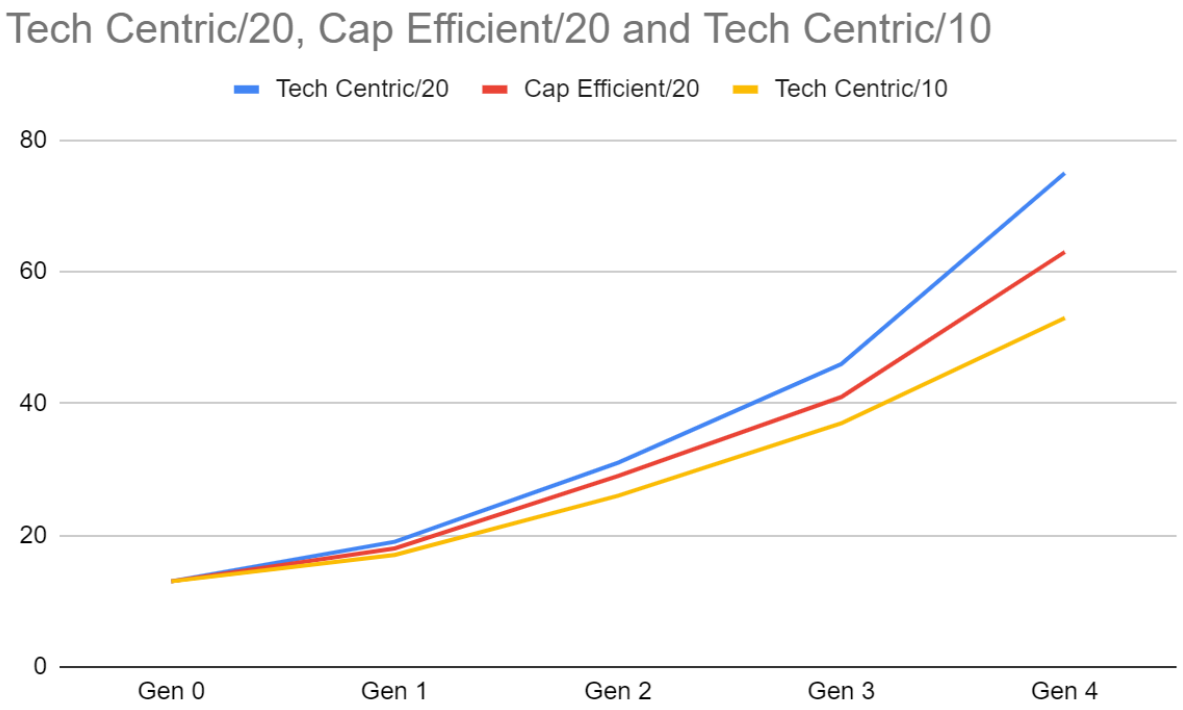


Figure 6.8 Comparison between different scaling options limited by thermal dissipation capability.

Beyond thermal efficiency, further analysis is required to assess the following issues:

1. This analysis used the Fugaku implementation to develop the reference 11x11 chiplet. It is possible that other chipet sizes offer a better solution.
2. The impact of architectures such as accelerators that require far more bandwidth per core. This may impact both scaling options.

6.11. Modular Package Designs

For packaging we have considered two discrete packaging form factor as our starting points:

a. Small 30mm x 55mm interposer package.
b. Large 55mm x 70mm interposer package.

Interposer is chosen as the default packaging option to allow for integration of HBM dielets in any modular HPC system and also allow for generation to generation technology-centric or capital-centric scaling. A silicon bridge based packaging option is also feasible but out of the scope of this study. Packaging and assembly lines for the above package sizes are available and used in many of the current HPC products as can be seen in already manufactured HPC system examples figure 6.8 below.

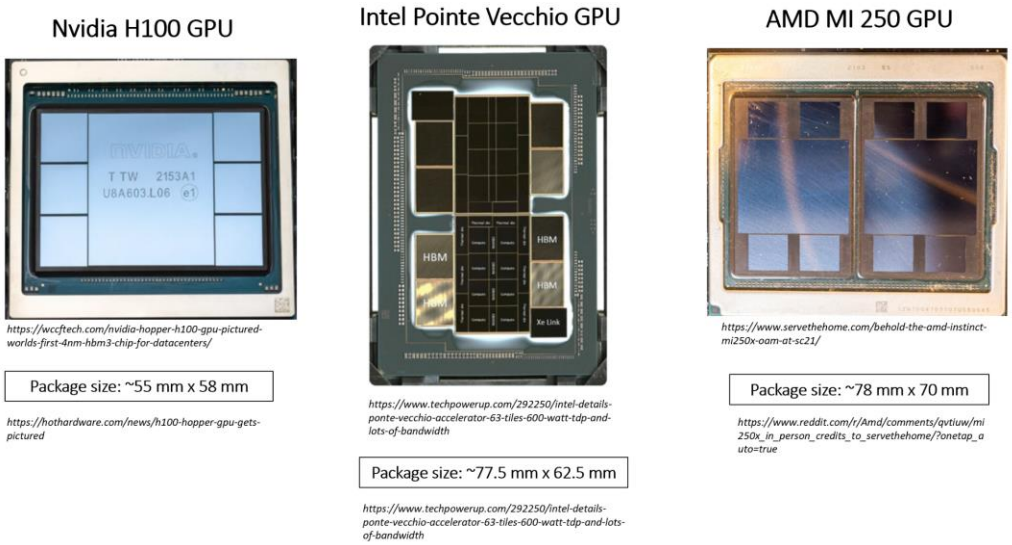


Figure 6.9 Different packaging form factors commonly used in HPC products.

Following chiplet sizes are taken from the reference HPC architecture which can be integrated into the modular package.

a. Reference Fugaku chiplet (11x11): 121 mm^2
b. Reference HBM chiplet (12x9 base die size): 108 mm^2
c. PCIe IP (4x11): 44 mm^2
d. TOFU high-speed network IP (4x11): 44 mm^2
e. NOC + HBM controller/interface die (4x12): 48 mm^2

Assumptions in integration:

a. Our primary goal is to map to the current Fugaku architecture to maintain architectural balance.
b. Routing constraints are relaxed. However up to 20 % routing overhead is included in package form factor design.
c. HBM2e PHY - 1.5 mm x 6 mm in 7 nm which is fitted to the 4 x 12 mm die form factor for ease of die handling, assembly and to have aspect ratio close to recommended tall thin die form factor. Will also help us to change memory type if we consider a fixed NOC size. Table 6.14 below shows different types of packaging form factors used to implement reference HPC system

Packaging form factor	30mmx55mm	55mmx70mm
	No. of chiplets	No. of chiplets
Fugaku ref. chiplets	4	4
HBM chiplets	4	8
NOC+HBM controller chiplet	4	8
PCIe	1	2

TOFU	1	2

Table 6. 18 Possible chiplet configurations in different packaging form factors.

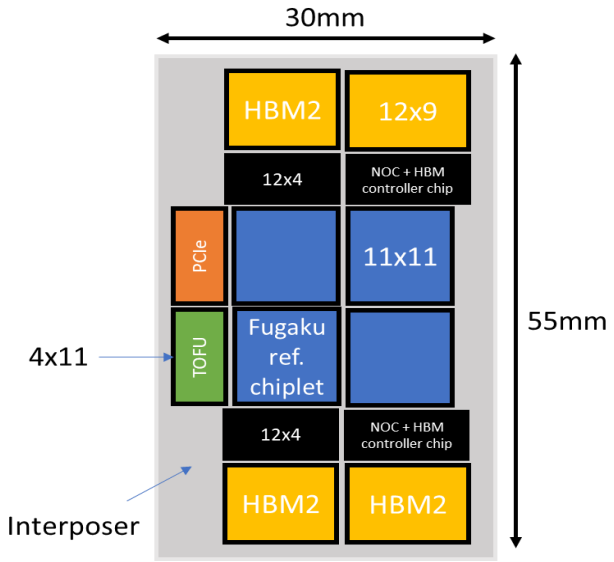


Figure 6.10 Reference architecture (with chiplet sizes) implementation in 30mmx55mm interposer package.

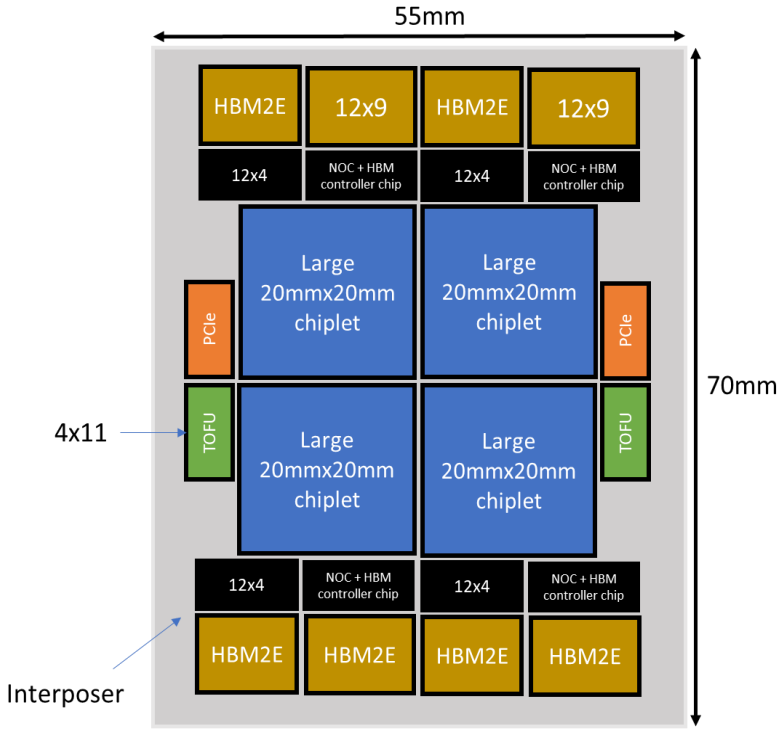


Figure 6.11 implementation with large 20mmx20mm chiplets on a 55mmx70mm interposer package.

A combination of small and large chiplet implementation can also be achieved for taking advantage of heterogeneity in multi-chiplet architectures. These chiplets could be accelerator chiplets or other HPC chiplets.

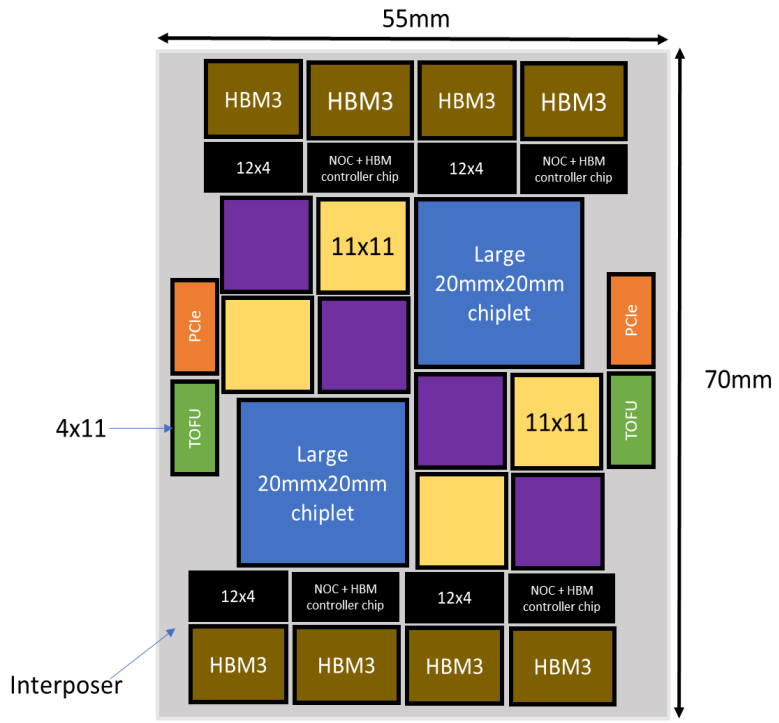


Figure 6.12 55mmx70mm interposer package with heterogeneous dielets (these can be compute, memory, IO etc.).

We also summarize the key considerations for package selection for chiplets to ensure optimal performance and cost-effectiveness:

- Chiplet Size: The physical size of the chiplet plays a crucial role in determining the package size. Larger chiplets may require larger packages to accommodate the necessary bonding pads, interconnects, and power delivery network. Conversely, smaller chiplets can be housed in smaller packages, reducing overall size and cost. The reference HPC node has been implemented in a 121 mm^2 die area for ease of packaging assembly, yield and automation.
- Thermal Considerations: We expect in the next few generations the thermal requirements will increase to 1.5 W/mm^2 which will require immersion and two phase cooling solutions. The Fugaku chiplet system is expected to be immersion cooled for earlier generations but development and implementation of two-phase cooling is critical to address scalability of the system. Our study in fact shows we are thermally limited in future scaling whether we take a technology centric or cost centric approach.
- Package Warpage: Excessive package warpage can strain the chiplets mounted within the package. If the warpage is severe, mechanical reliability issues occur and compromise their electrical performance. Details on acceptable package warpage with scaling over the next few years have been tabulated in TWG1 reports.
- Interconnect Complexity: Higher pin count or more complex interconnect schemes may require larger packages to accommodate the necessary routing and bonding pads.

- Signal Integrity and power integrity: The package design should consider minimizing signal losses, crosstalk, and impedance mismatches. This may impact the package size to allow for proper signal routing and signal integrity optimization. Furthermore as the power requirements grow, package designs need to consider space for adding a greater number of decoupling capacitors, voltage regulator modules etc.
- Mechanical standards: Mechanical standards form a critical part in deciding modular package sizes and chiplet sizes on these packages. Mechanical standards may include JEDEC part model guidelines for electronic-devices packages (JEP30-P101), JEDEC standards for handling, packing, shipping sensitive devices (J-STD-033D), JEDEC reliability standards pertaining to temperature cycling, humidity/moisture bias testing of packages, electrostatic discharge sensitivity tests as well as study of various failure mechanism in these modular packages.
- Power delivery requirements: Power delivery is a critical part of the modular package. Several power standards including IEEE 1801-2018 i.e. the unified power format, IEEE 2416 standard for power modeling to enable system level analysis etc. should be used to analyze chiplets based on equations, measured and simulated data. Unified power format (UPF) is useful for describing power intent especially in a multi-voltage environment such as modular package design. In addition, power delivery analysis is critical to determine the decoupling capacitance to reduce the effect of noise and ground bounce. The DECAP needs to be decided based on DECAP space to die-size and cost, mechanical support required for the chiplet after addition of DECAPs etc.
- Test Strategy During Assembly: Test before, during and after assembly is necessary to ensure highest assembly yield. It is usually expected only known good dies (KGDs) and known good packages are used for assembly. Intermediate testing in case of multi-chiplet assembly and complete system level scan chain tests should be considered to ensure the package functions as expected post assembly. The testing methods should be compliant with standards proposed by JEDEC.

Ultimately, the selection of package sizes for chiplets involves a trade-off between various factors such as chiplet size, thermal considerations, interconnect complexity, signal integrity, power delivery, and cost. We believe that building reference architectures is an effective way to optimize these tradeoffs and provide realizable package designs both for technology and capital centric scaling approaches.

6.12. Further Research and Development

The proposal developed in this document uses current packaging technology and process nodes as a starting point for a reference architecture proposal. As packaging technology advances, two significant advances are possible:

- In sub-nanometer nodes, fairly complex technology can be aggregated in chiplets even as small as a few square millimeters in size

- Packaging technology may be advanced enough for off-chip wire density to asymptotically approach on-die wire density.

In this context, reference architectures may need to evolve to support a large collection of smaller chiplets aggregated on a dense interconnect fabric. Wide slow D2D protocols are best suited to these advanced packaging technologies. Further research will be required to enable this vision.

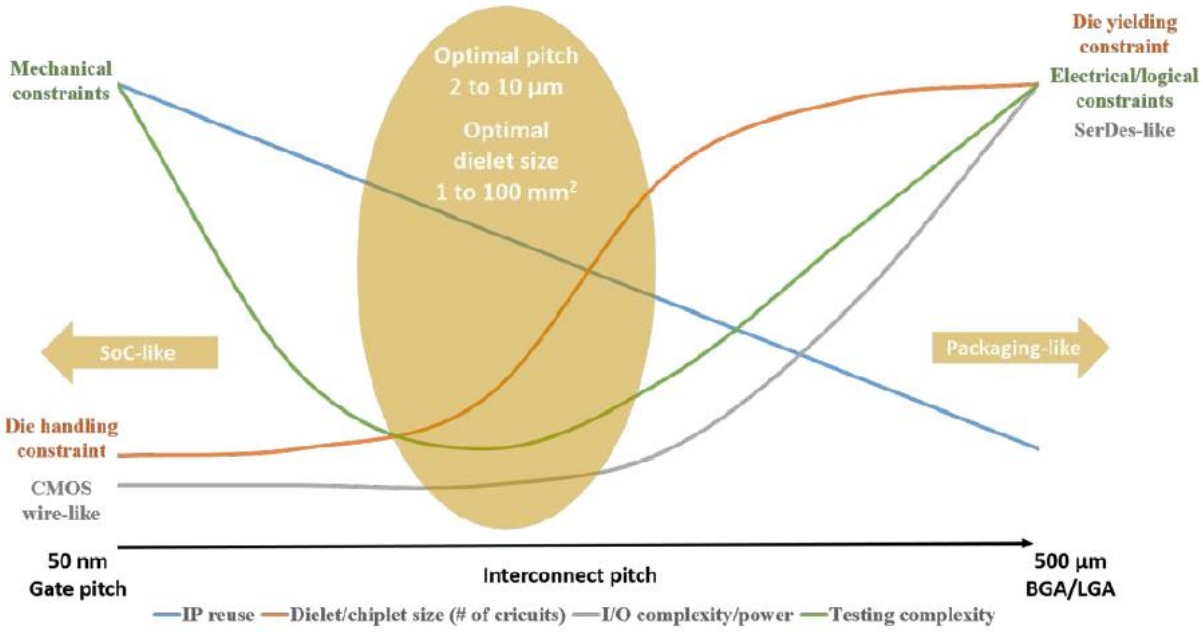


Figure 6. 13 Dielet golden regime for identifying optimal bond pitch and dielet size [8].

6.13. Recommendations

In this work we have attempted to generate a reference modular HPC architecture from a monolithic HPC node (Fugaku), to compare the tradeoffs between monolithic and modular approach in terms of architectural stability (IO bandwidth/core), technology-centric scaling, capital-centric scaling, packaging. We can achieve similar architectural performance (IO bandwidth/core) going from a monolithic to a modular architecture. Maintaining similar architectural balance, we have taken a technology-centric bump pitch reduction approach as well as capital-centric approach to discuss the feasibility of both approaches. In either case we have found a need to improve the thermal dissipation capabilities in future to prevent performance scaling from being thermally limited. Today force air cooled systems are predominantly in use, but in future immersion cooling and two-phase cooling architectures need to be implemented to maintain performance scaling.

Another aspect of this study is that die size is pad limited, referring to the fact that die size and bond pitch are related: large bond-pitches lead to large die sizes. These large dies yield poorly and can be difficult to handle. Furthermore, arbitrary dielet sizes place demands on assembly tooling and manufacturing efficiency and need to be avoided. However, over the course of time, dialets should be made smaller to improve yield, handling and other mechanical constraints while

generating the chiplet performance needed. Below we present a table (Table 6.15) with recommended die size over generation.

Year	**Small (mm)**	**Medium (mm)**	**Large (mm)**	**Super large (mm)**
2023	2X2, 3X3,4X4	5X5 to 10X10	11X11 to 15X15	16X16 to 28X30
2026	1x1 to 4x4	5x5 to 10x10	11X11 to 15X15	16X16 to 20X20
2029	1X1 to 4X4	5X5 to 10X10	11X11 to 15X15	eliminated
2032	1X1 to 4X4	5X5 to 10X10	11X11 to 15X15	eliminated

Table 6. 19 Recommended dielet sizes over the next ten years.

In addition another recommendations can be made: Similar reference architectures need to be created for medical devices, automotive, radio frequency (RF). Developing reference architectures in each field based on already existing products can significantly help with developing chiplet standards in that field. Reference architectures will be used to recommend standards.

6.14. Conclusion

Chiplets make heterogeneous integration possible, the development of products that integrate chiplets from multiple companies and process nodes. The use of standards can make it more feasible for multiple companies to create chiplet-based products. Current efforts in standards have largely focused on open protocols for logical interaction between chiplets such as UCIe, BoW and Superchips. This report is focused on the potential benefits of using these standards, particularly for high mix low volume products that require complex packaging technology. These products are often challenged by high design costs and high per-unit manufacturing costs.

A significant business challenge with chiplet-based products is that product revenue depends on how easily a chiplet can be integrated with chiplets from other vendors to form a product. Challenges in aggregation delay revenue and impede the development of a vibrant chiplet ecosystem. The set of open standards as defined do not adequately address all the challenges in integration. Two chiplets designed to the same protocol may not be usable in one product because they are designed for different packaging technologies, different bump densities, are too hot to be next one another, use too much power etc.

Large companies address these challenges by designing chiplets in families. That is, a target system is partitioned into functionally-distinct modules, a common packaging technology is

chosen and each module is also allocated a specific physical budget. Therefore, though each chiplet is designed individually, aggregation is simplified by the front-end budgeting process.

Today, such a budgeting process is not available for chiplets designed across companies. The time, complexity and cost of solving these challenges directly impacts the commercial viability of any chiplet. In fact, these challenges have impeded the creation of a vibrant multicompany chiplet ecosystem. Chiplet-based designs have largely been restricted to high-volume products from very large companies that largely control their supply chains.

In this report, we propose the development of modular reference architectures that create specific functional and physical modularity budgets. We propose to generate physical, mechanical and thermal guardrails by developing modular packaging and chiplet designs based on modular domain-specific reference architectures.

Modularity can enable significantly more automation in packaging and assembly. By utilizing standards, it is also possible to increase the automation of the packaging and assembly process. These same standards may also expedite the exploration of the search space in system design. To showcase the benefits of these standards, this report focuses on one specific application: high-performance computing. While this application is also covered by other working groups, the report shows that a reference architecture based on acceptable functional partitions can be mapped onto a limited set of physical modules. This limitation in the range of chiplet sizes enables greater design and manufacturing reuse across products. The next step for this work is to generate more detailed proposals for functional and physical modularity to estimate the benefits quantitatively.

Modularity can increase demand for chiplets by easing integration, lead to reduced design costs and lower per-unit costs of packages and accelerate the development of a vibrant open chiplet economy.

References:

[1] https://ieeexplore.ieee.org/document/9731627

[2] https://www.fujitsu.com/global/documents/about/resources/publications/technicalreview/2020-03/article03.pdf

[3] https://hothardware.com/news/h100-hopper-gpu-gets-pictured

[4] https://www.reddit.com/r/Amd/comments/qvtiuw/mi250x_in_person_credits_to_servethehome/?onetap_auto=true

[5] https://www.techpowerup.com/292250/intel-details-ponte-vecchio-accelerator-63-tiles-600-watt-tdp-and-lots-of-bandwidth

[6] S. C. Jangam and S. S. Iyer, "Silicon-Interconnect Fabric for Fine-pitch (≤10 μm) Heterogeneous Integration," in IEEE Transactions on Components, Packaging and Manufacturing Technology, doi: 10.1109/TCPMT.2021.3075219

[7] K. Sahoo, U. Rathore, S. Chandra Jangam, T. Nguyen, D. Markovic, S. S. Iyer, "Functional Demonstration of < 0.4-pJ/bit, 9.8 μm Fine-Pitch Dielet-to-Dielet Links for Advanced Packaging using Silicon Interconnect Fabric," 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC), June 2022.

[8] S. S. Iyer, S. Jangam, and B. Vaisband, "Silicon interconnect fabric: A versatile heterogeneous integration platform for AI systems," in IBM Journal of Research and Development, vol. 63, no. 6, pp. 5:1-5:16, 1 Nov.-Dec. 2019.

Chapter 7: Security in Heterogeneous Integration and Advanced Packaging

Contents

Chapter 7: Security in Heterogeneous Integration and Advanced Packaging 1
7.1 Executive Summary 1
7.2 Challenges 1
7.3 Solutions: Overview and Approach 2
7.4 Opportunities and Solutions by Leveraging Onshoring 2
7.5 General Conclusions 5
References 5

7.1 Executive Summary

The cybersecurity landscape in heterogeneous integration and electronics packaging (MRHIEP) is impacted by two major phenomena. The first is the rise of hardware-based vulnerabilities which have been created by malicious actors across the supply chain. Examples are hardware Trojan which can be injected at various stages of manufacturing, and/or information leakage through side-channels. Our reliance on outsourced designers and fabs has further exacerbated this issue.

The second is the advent of fresh integration and packaging technologies, such as chiplets, which have opened the door to an unprecedented chance to reconsider security in hardware design and production. Numerous existing security concerns could potentially find resolution through these novel technologies, especially with meticulous attention dedicated to the design phase such that a “secure-by-design”approach could be achieved.

The next generation of ONSHORE manufacturing methods must acknowledge these two key factors: the emergence of new hardware vulnerabilities and the opportunity to use innovative technologies to address them. It is essential to develop hardware that is secure, efficient, reliable, and high-performing. Achieving this requires a comprehensive approach that encompasses design, manufacturing, and execution times. Designers and manufacturers must recognize that security is now a paramount concern and cannot be disregarded, as it can have significant financial and other consequences. Therefore, they must make appropriate tradeoffs to ensure security is on par with other critical metrics like performance, power, and cost.

7.2 Challenges

There are three main gaps that requires outmost attention.

- There is a shortage of affordable techniques that can guarantee security without compromising other significant measures such as performance, power, and cost. Successful solutions strike the

ideal balance between these metrics. Nevertheless, designers and stakeholders must acknowledge that security comes at a cost, necessitating certain concessions.

- Insufficient secure-by-design and secure-aware packaging methods, rather than reactive solutions. Considering the widespread and important nature of cyberattacks, particularly targeting the hardware layer, it is crucial that future hardware generations prioritize incorporating security measures during the design and packaging phase, instead of relying solely on post-incident solutions like patching.
- Effectively connecting the overlapping issues in supply-chain and security including detecting altered, counterfeit, and pirated hardware components.

7.3 Solutions: Overview and Approach

The approach for addressing the security concerns in the next generation of MRHIEP[1] technologies should be based on identifying the current issues and developing effective solutions with reasonable overheads. To this end, *three* important focus areas should be considered. It is important to highlight that onshoring, as will be discussed later, has an important impact on all three aspects.

- **Design Time:** Effective solutions are those that start with proper security considerations at the design time. Particularly, "secure-by-design" approaches should be considered and employed. The current known challenges in security, including *side-channel leakage* [1], *fault attacks* [2], *tampering, Trojans* [3], *reverse engineering, and counterfeiting*, should all be properly considered.
- **Manufacturing and Post-Manufacturing Time:** Heterogeneous integration has provided a unique new capability to rethink secure manufacturing. New techniques that can properly address trust concerns should be developed.
- **Execution Time:** Security and trust can be further enhanced by employing execution-time *monitoring* techniques. A multi-layer approach that starts with the design and ends with monitoring can ensure trustworthiness in the next generation of advanced systems.

Given these aspects, a cross-layer approach should be considered to close all security vulnerabilities. Such an approach could also bring down unwanted overheads including cost, area, power, and performance. In the following, we describe how onshoring coupled with recent technological advancements could potentially address the security challenges.

7.4 Opportunities and Solutions by Leveraging Onshoring

The opportunity for new heterogeneous integration technologies as well as onshore manufacturing capabilities could enable us to address many security challenges in state-of-the-art complex systems. Here we propose four promising directions.

New protocols for manufacturing by distributing trust. Offshore production of hardware exposes us to various types of hardware-based attacks. These attacks encompass vulnerabilities in the supply chain, such as counterfeit hardware, as well as more critical threats like hardware Trojans, which can potentially

[1] Manufacturing Readiness for Heterogeneous Integration and Electronics Packaging

compromise the system's integrity through malicious foreign entities. Onshoring, particularly at the packaging stage, introduces an additional manufacturing step that can be utilized to ensure the integrity of the process. However, capitalizing on this opportunity necessitates the development of novel design strategies that distribute trust and maintain system security even if one of the outsourced components is compromised. It is crucial to devise new protocols specifically tailored to this model, while comprehending the threat model and implementing appropriate measures. An example of such measures is the utilization of split manufacturing techniques [4] to distribute trust effectively. An example of such approach is shown in Figure 7.1 below where instead of manufacturing a monolithic chip (2D or 3D), the design is broken down into multiple chiplets, each of which manufactured in a different fabrication facility.

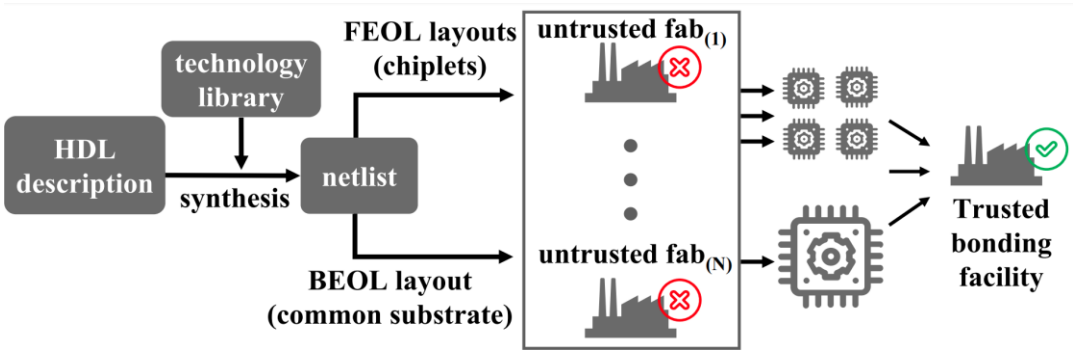


Figure 7.1: Example of split manufacturing technique to distribute trust effectively.

Hardware root of trust approach. Onshoring also provides this new opportunity for building a hardware root of trust in the presence of a foreign adversary. Innovative approaches must be developed to harness this potential. One such approach involves incorporating potentially untrustworthy components into a substrate through a sequence of protocols. Additionally, consideration should be given to implementing specialized detection mechanisms within the hardware root of trust to identify potential malicious components. An example of such approach is shown in Figure 7.2 below [5] where a "secure socket" has been added to the IP (e.g., accelerator) to ensure its trustworthiness. In this example, although the IP has been designed and manufactured in an untrusted facility, the subtract trusts and interacts with this component only if it follows a particular protocol. Formal and hardware methods need to be developed to ensure the correctness and security of this approach.

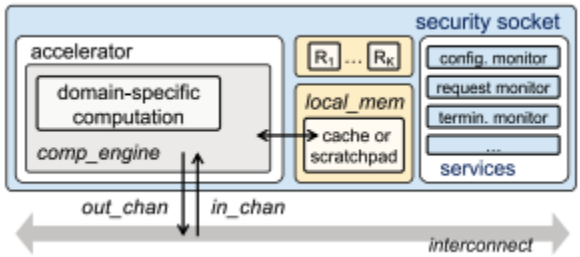


Figure 7.2: Example showing implementation of specialized detection mechanism.

Trust anchors and monitoring. Similar to the idea presented above, instead of having a unified trusted substrate, onshoring could potentially help with manufacturing and/or integrating trusted components into an overall untrustworthy system. These trust "anchor" can then be used as either *monitoring* tools to check the status of the system and/or to enforce the secure and trustworthy operation of the system. New methods [6] should be designed to achieve these design goals. An example of such approach is shown in the Figure 7.3 below. In this design, a secure utility die (UD) is added to the design to provide various *security* features such as monitoring, key management, etc. Similarly, such a unit could be added as an *active* element within the substrate.

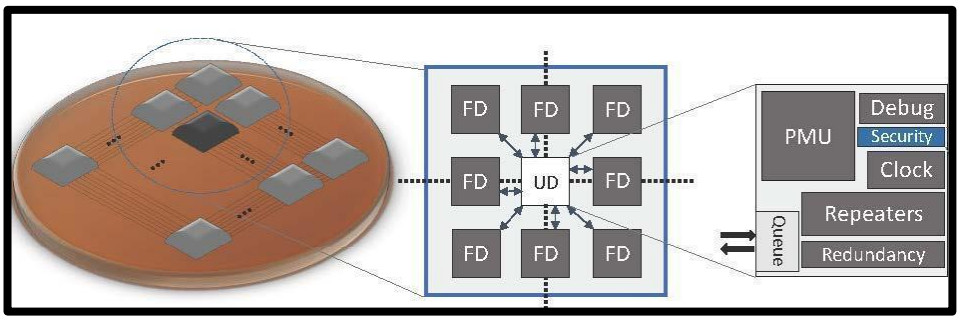


Figure 7.3: Example of new methods to achieve design goals for HI systems

New methods for counterfeit detection and supply chain security. Lastly, instead of integrating untrustworthy components, methods can be developed to detect malicious components during the packaging phase. Novel methods for Trojan and counterfeit detection are needed where the system integrator could use them to detect and eliminate the malicious components. Particularly, focus should be on new methods for counterfeit detection based on leveraging side-channels, machine vision, and advanced test equipment (e.g., laser, SEM, etc.).

Summary of gaps and solutions. The summary of potential gaps and their solutions are summarized in the table below.

Gap	**Roadmap Solution needed**
Lack of low-cost methods that can ensure security without sacrificing other important metrics.	New design methodologies by utilizing the capabilities that onshoring and new packaging technologies could provide. This includes new chiplet and system integration strategies, split manufacturing and packaging techniques that leverage the onshoring capability.
Lack of secure-by-design approaches and secure-aware packaging methods rather than reactive solutions.	New practices for design-by-security strategy. Identifying concerns at different stages and designing new mechanisms that can guarantee security even in the presence of an adversary.

Effectively connecting the overlapping issues in supply-chain and security including detecting altered, counterfeit, and pirated hardware components.	Exploring security-related supply-chain considerations. Developing methods that can authenticate various chiplets during the packaging stage and reject malicious units. Designing methods that ensure security of the system even when a particular component in the system is compromised.

Table 7.1: Summary of Gaps and Proposed Solutions

7.5 General Conclusions

Cybersecurity is an important concern for the next generation of complex systems. As electronic systems become more complex and interconnected, cybersecurity has become a top priority for research, particularly in security-critical applications. By analyzing the very diverse supply chain, more complex system topology, and greater proximity of chips, this chapter has addressed those cybersecurity threats that are most affected by heterogeneous integration. As described in detail above, heterogeneous integration has major security impacts due to changes in interconnect layouts, test protocols, supply chain diversification, and vertically stacked geometries. It is clear that these increased security threats must be addressed by a more system-level approach to security that requires a systematic design-for-security perspective.

References

[1] P. Kocher, J. Horn, A. Fogh, D. Genkin, D. Gruss, W. Haas and M. Hamburg, "Spectre Attacks: Exploiting Speculative Execution," *Communications of the ACM*, vol. 63, no. 7, 2020.

[2] O. Mutlu and J. S. Kim, "Rowhammer: A Perspective," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 8, pp. 1555-1571, 2019.

[3] S. Bhunia, M. S. Hsiao, M. Banga and S. Narasimhan, "Hardware Trojan attacks: Threat Analysis and Countermeasures," *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1229-1247, 2014.

[4] F. Imeson, A. Emtenan, S. Garg and M. Tripunitara, "Securing computer hardware using 3d integrated circuit ({IC}) technology and split manufacturing for obfuscation," in *22nd {USENIX} Security Symposium {USENIX Security 13*, 2013.

[5] L. Piccolboni, D. Giri and L. P. Carloni, "Accelerators & Security: The Socket Approach," *IEEE Computer Architecture Letters*, vol. 21, no. 2, pp. 65-68, 2022.

[6] Y. Safari, P. Aghanoury, S. S. Iyer and N. Sehatbakhsh, "Secure and Scalable Key Management for Waferscale Heterogeneous Integration," in *ECTC Proceedings*, 2023.

Chapter 8: Heterogeneous Integration Test Technology

Contents

Chapter 8: Heterogeneous Integration Test Technology 1
8.1 Executive Summary and Scope 3
8.2 RF Test 5
8.3 Test of Photonic Devices 8
8.3.1 INTRODUCTION 9
8.3.2 Situation Analyses 12
8.3.3 Gaps and Showstoppers 13
8.3.4 HIGHER PIC TECHNOLOGIES 14
8.4 Logic Testing 16
8.4.1 Introduction 16
8.4.2 Addressing the architectural bottlenecks of logic scan infrastructure 18
8.4.3 Traditional scan challenges 18
8.4.4 Emerging Use-Cases 20
8.4.5 Updated scan architectures 20
8.4.6 Evolving logic test beyond scan testing 22
8.4.7 In-System and SLT test requirements driving new logic test requirements 23
8.4.8 ATE equipment challenges with the logic testing 24
8.4.9 Conclusions 26
8.5 Specialty Device Testing 26
8.5.1 Trends Impacting this Technology Area 27
8.5.2 Concerns: Test Challenges 28
8.5.3 Summary 30
8.6 Memory Test 31
8.6.1 Summary 31
8.6.2 NAND 34
8.6.3 DRAM 35
8.7 Analog and Mixed Signal Test 36
8.7.1 Executive Summary 36
8.7.2 DC Accuracy updates for 2020 37
8.7.3 Power updates for 2020 37
8.7.4 Analog Mixed Signal Updates for 2020 39

8.7.5 Rx tests 41
8.7.6 Key Test Trends 41
8.7.7 SUMMARY 44
8.7.8 References 44
8.8 Wafer Probe and Device Handling 46
8.8.1 Device Handling Trends 46
8.8.2 Test Sockets 58
8.9 System Level Test 63
8.9.1 Executive Summary 63
8.9.2 Enablers and Challenges of System-oriented Test 64
8.10 Data Analytics 64
8.10.1 Background 64
8.10.2 Why is Data Analytics Important for Semiconductor Manufacturing and Test? 65
8.10.3 Transforming the Backend to an Industry 4.0 Smart Factory 67
8.10.4 Optimizing Cost of Test 67
8.10.5 Improving Quality Assurance 68
8.10.6 Improving Yield 69
8.10.7 Performance Grading/Binning 69
8.10.8 Traceability Across the Semiconductor Value Chain 70
8.10.9 Data Analytics for Test - Key Enablers Roadmap 70
8.10.10 Impact of COVID-19 on Data Analytics Roadmap (Special Section for 2023) 73
8.10.11 Additional Reading 73
8.10.12 References 73
8.11 2.5D & 3D Device Testing 73
8.11.1 Introduction 73
8.11.2 Challenges for Test 74
8.11.3 Known Good Die (KGD) Test 74
8.11.4 Interposer Testing 75
8.11.5 High Speed Interconnects and Signal Integrity 75
8.11.6 Impact of emerging technologies with respect to test 75
8.11.7 3D TSV/interconnect testing 76
8.11.8 3D probing, 3D die stacks, 3D stack repair 76
8.11.9 Long term prediction 76
8.11.10 Call-for-action 78
8.11.11 References 78

8.12 Key Drivers and Test Costs 78
8.12.1 Key Cost of Test Trends 79
8.12.2 Cost of Test as a Part of Overall Manufacturing Cost 80
8.12.3 Test Cost Models and Cost Improvement Techniques 82
8.12.4 Current Top Cost Drivers 83
8.12.5 Future Cost Drivers 83
8.12.6 Cost Reduction Techniques 83
8.12.7 Summary 86

8.1 Executive Summary and Scope

Semiconductor product sizes and complexities are continuing to increase dramatically, and all the billions of components in a device must be tested to ensure they are functional and meet the product specifications. Semiconductor Test was for multiple decades dominated by structured test methods such as full scan and built-in self-test (BIST). However, the pendulum is swinging back towards the use of functionally based testing such as system level test and other similar methods. Furthermore, as chip manufacturing transitions from monolithic ICs towards heterogeneous integration (HI), and complexity increases dramatically at the same time as access to circuit internals decreases. Finally, defense, automotive, high-performance compute, and other electronics consumers in the US are emphasizing the need for supply chain assurance and security and device traceability across the semiconductor value chain, and test plays a pivotal role since it is the primary communication mechanism for finished devices prior to their integration into the end application. So there are many challenges that the test industry must address in order to keep up with this rapidly evolving industry.

Solving these problems requires specialized skills which are increasingly scarce in the US. There are several reasons for this decrease in the semiconductor test area:

- Relatively few universities in the US have an academic program which includes more than 1-2 courses in semiconductor test and related practices such as design for testability (DFT). Fewer still have labs with the associated test equipment for student use.
- Students are eschewing semiconductors and semiconductor test in favor of software and related disciplines which are more visible and popular and have higher salaries.
- In the past, the Semiconductor Research Corporation (SRC) had significant financial support for test-related research for master's and PhD students. But in recent years, those funds have been redirected to other areas and the associated faculty are moving away from test to be able to support their students. Other countries in Europe and Asia continue to fund test research.
- Semiconductor testers are expensive and are mostly located overseas where labor is cheaper. Test time availability for test engineers in the US often is not available other than overnight hours, which upsets the work life balance and makes semiconductor test a less attractive career choice.

We would recommend several courses of action to address the issues listed above:

- Survey companies to identify their specific needs in semiconductor test.
- Formalize a pipeline of interns throughout the year to support productization of research concepts and push innovations into our products.
- Endow ongoing funding for graduate level semiconductor test research and to support equipment donations or purchases to teach semiconductor test to undergraduates.
- Incentivize semiconductor companies to create lab environments where semiconductor test-related research can be conducted as well as production facilities for test operations.
- Fund the development of outreach materials that can inform middle school, high school, and college students about careers available in semiconductors and semiconductor test and the types of products they enable.

Below we provide a high-level summary of the key test challenges and needs for each of the device types addressed in this chapter on HI & chiplet test. For further background, refer to the IEEE-EPS Heterogeneous Integration Roadmap

RF Test: Need 1) Non-frequency-gapped ATE RF test capability in the 0-100 GHz frequency range, either for characterization, quality assurance, and/or high-volume production testing; 2) Higher ATE RF bandwidth production test capability up to 400 MHz for Wi-Fi 7 (with EVM in the 48+ dB range) and satellite; and up to 2 GHz to support 5G mmWave, UWB, and 6G THz; and 3) High-volume over-the-air (OTA) handler-based testing for mmWave and THz, and possibly automotive radar, will become increasingly relevant as DIB cabling for increased site count becomes cost-prohibitive.

Photonics Test: Need 1) Novel test approaches for testing optics in co-packaged heterogeneous devices in high volume; and 2) Emphasis on test time containment and test time reduction as the number of lanes and wavelengths per fiber increase.

Logic Test: Need 1) New test methods for testing chiplet devices with mixed technologies (for example, need for retargetable test IP for next level of integration into SIP or system); 2) test methodologies using Silent Data Corruption (SDC) logic testing methods; and 3) Standardized test interfaces and methods for chiplets that can be used by both chip foundries and packaging integrators (such as OSATs).

Specialty Test: Need 1) Higher test parallelism to reduce cost of test; and 2) multi-functional and cost-effective test capabilities as specialty devices become part of heterogeneous packages.

Memory Test: Need 1) Test capabilities for addressing higher interface speed, power, and thermal management requirements; 2) Test capabilities for overcoming the challenges of electro-mechanical interface capability of wafer and component test as NAND memory density increases due to vertical scaling; and 3) Testing of higher DRAM bandwidth requirements.

Analog/Mixed Signal Test: Need 1) High speed instrumentation that can accept, force, and tolerate higher voltages and currents, driven by wide bandgap materials; 2) DC accuracy below 50 uV over the entire temperature range; 3) Closed-loop temperature forcing test capability at final test; 4) Test capabilities for A/MS devices housed in heterogenous packages; 5) Novel test solutions for overcoming the inherent physics of high voltage test at very high multisite testing; 6) High density floating resources with high accuracy, medium current capability, and large isolation voltages; and 7) Need for fully floating low-speed digital instrumentation for testing chip-to-chip communications devices which are shifted by tens to hundreds of volts above or below system ground.

System Level Test: Need 1) Flexible DFT architectures for both structural and functional test content; 2) Effective SW/HW system failure diagnosis methods; and 3) Deep component parametric data extraction to data analytics.

Data Analytics: Need 1) For advanced and comprehensive data analytics solutions that take full advantage of data from across the entire value chain; 2) Significant improvements in the development and adoption of key enablers such as communications infrastructure, data interchange formats,

traceability, data security, and advanced data analytics algorithms; 3) Efficient methods for accessing, curating, managing, and analyzing data from on-chip sensors IP, equipment sensors, and test results.

2.5/3D Test: Need 1) Known-good-die DFT test methods that enable high quality wafer probe test – thus reducing fallout at final test; 2) Faster die-to-die communication standards that enable thorough testing at final test; 3) Standardized test and repair methodologies that consider new trends in 3D interconnects; 4) Yield prediction and analysis methods that ensure fallout at all levels of testing are understood; and 5) End-to-end data analytics capability that applies to all dies on the package.

Test Cost: Need 1) New probing technology which allows testing of singulated die; 2) New PCB and interposer technology to lower the cost and complexity of consumable materials; 3) Improvements in the test process by increased use of data analysis and machine learning based on measured data; and 4) Cost reduction of system-level testing.

Test Technology Working Group Leadership Team

Co-Chairs: Ken Butler

Jeorge Hurtarte

RF Test: Jeorge Hurtarte **Analog/Mixed Signal Test:** Rich Dumene

Photonics Test: Dave Armstrong **System Level Test:** Harry Chen

Logic Test: Marc Hutner **Data Analytics**: Ira Leventhal

Specialty Test: Wendy Chen **Test Cost:** Ken Lanier

Memory Test: Jerry McBride

2**.5D/3D Test:** Morten Jensen and Boris Vaisband

8.2 RF Test

In the mobile wireless sector, history shows that there is a new "G" every 8-10 years. Thus, while we saw the emergence of 5G in both the sub-8 GHz and mmWave (24-53 GHz) during 2018-2022, we can expect 6G (THz) product prototypes to start emerging in the 2027-2030 timeframe.[1]

In the Wi-Fi connectivity wireless sector, Wi-Fi 7 (802.11be) in the 2.4, 5, and 6 GHz spectrum bands (the latter extending up to 7.125 GHz with up to 320 MHz of bandwidth) will see initial volume production in the 2024-2025 timeframe.[2] We can also expect that micro positioning capability will be added into Wi-Fi 7 at 320 MHz (802.11bk), in addition to 802.15.4z UWB (up to 11 GHz).

[1] https://www.testconx.org/premium/wp-content/uploads/2021/TestConXMesa2021s1p1Hurtarte_9106.pdf

[2] https://www.ieee802.org/11/IEEE%20802-11-Overview-and-Amendments-Under-Development.pptx

In addition to mobile and connectivity, automotive radar applications in the 76-81 GHz frequency range will continue adoption, while satellite connectivity in the Ku-band (12-18 GHz), and Ka-band (26.5-40 GHz) will see higher volume in the 2025-2030 timeframe as an effort to reach remote areas.

These wireless market segment technologies trends translate into the following high level ATE requirements up to 2030, which are further discussed below.

- Non-frequency gapped ATE RF test capability in the 0-100 GHz frequency range, either for characterization, quality assurance, or high-volume production testing. It is also likely that "IF" frequencies for 6G THz will fall within this 0-100 GHz range.
- High-volume over-the-air (OTA) handler-based testing for mmWave and THz, and possibly automotive radar, will become increasingly necessary in the 2025-2030 timeframe as DIB cabling for increased site count becomes cost-prohibitive.
- Higher ATE RF bandwidth production test capability up to 400 MHz for Wi-Fi 7 (with EVM in the 48+ dB range) and satellite; and up to 2 GHz to support 5G mmWave, UWB, and 6G THz.

For mobile devices, we will see the expansion of 5G millimeter wave into the 71 GHz range with the adoption of 3GPP Release 17.[3] In addition, with the advent of 6G, we can expect RF frequencies beyond 100 GHz into the THz range.[4] These two trends will require non-gapped frequency test capabilities from "0-100 GHz" as customers will not want to have multiple instruments to test different frequency ranges. While it is not yet clear that GHz and THz devices will be 100% tested at those frequencies in production, such capabilities need to be present in the tester for characterization and quality assurance purposes (for example, for analyzing field failures).

Millimeter wave and THz will require novel and cost effective over-the-air testing (OTA) methodologies, which started to appear around 2022 from companies such as Teradyne and Advantest, but these will require more maturity to achieve high-volume-handler-ready solutions.[5] [6] OTA test techniques will compete with other more cost-effective methods, yet may not be as reliable for performance testing, such as "leakback" and "radiateback" test techniques. Such alternative solutions will push the limits of the device interface boards (DIB) wiring and cabling for multisite device testing, and thus the need for cost-efficient and high-performance handler-based OTA test techniques.

Table 8.1 shows an increased bandwidth requirement, as a minimum, for characterization testing of various millimeter wave devices, most notably in the 2GHz bandwidth range for higher volume use cases (e.g., 5G FR2-2).

[3] https://www.qualcomm.com/documents/download-our-5g-nr-rel-17-presentation

[4] https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/Qualcomm-Whitepaper-Vision-market-drivers-and-research-directions-on-the-path-to-6G.pdf

[5] https://www.teradyne.com/2022/08/17/the-future-of-wireless-test-is-over-the-air/

[6] https://www.testconx.org/premium/wp-content/uploads/2021/TestConXMesa2021s1p2Semancik_2948.pdf

RF Frequency and Bandwith Requirements for 2020-2030				
Wireless Standard Technology	Min Frequency (GHz)	Max Frequency (GHz)	CC Bandwidth (MHz)[1]	Description
3GPP TS 38.101-1	0.4	7.125	100	5G FR1
802.11ax	0.4	7.125	160	Wi-Fi 6E
802.11be	0.4	7.125	320	Wi-Fi 7
802.15.4z	1	11	1300	UWB
Satellite	12	18	250	Ku VSATs
ETSI TR 101 982	21	27	200	24 GHz SSR Auto Radar SS
Backhaul	18	38	60	BTS Backhaul
Satellite	26	40	250	Ka VSATs
3GPP TS 38.101-2	24.25	52.6	400	5G FR2-1 mmW
3GPP Rel. 17	52.6	71	2000	5G FR2-2 mmW
Backhaul	57	66	4000	BTS Backhaul
802.15.3c	57	66	5500	Motion sense / Hand gesture
3GPP TR 38.806	52.6	71	1000	5G FFS mmW
802.11ay	55	76	4000	WiGig
Backhaul	71	76	4000	BTS Backhaul
ETSI TR 101 983	76	77	1000	77 GHz LRR Auto Radar FMCW
ETSI TR 101 263	77	81	4000	79 GHz SRR Auto Radar FMCW
Backhaul	81	86	4000	BTS Backhaul
4D Imaging Radar	77	86	4000	SRR 4D Imaging Radar
Backhaul	92	95	4000	BTS Backhaul
U-SRR	120	140	> 4000	cm radar
6G (THz)	95	3000	> 4000	6th Generation Mobile Networks
Note 1: CC = Component Carrier				

Table 8.1 RF Frequency and Bandwidth Requirements for 2020-2030

IEEE 802.11 continues to work on new connectivity Wi-Fi standards such as 802.11be (aka Wi-Fi 7) with a maximum channel bandwidth of 320 MHz and 4k QAM modulation.[7] Thus, the key test requirements for Wi-Fi 7 are the capabilities to test waveforms with 320 MHz bandwidth in a single measurement at EVM of greater than 48 dB. The more stringent EVM requirement stems from the 4K QAM (Quadrature Amplitude Modulation) which enables each signal to more densely embed greater amounts of data compared to the 1K QAM with Wi-Fi 6/6E. For high order modulations such as 4096-QAM, which require stringent transmitter accuracy, selecting test equipment with a low EVM floor is critical, otherwise the error uncertainty contributed by the test equipment reduces the confidence in the final measurement.[8]

UWB (Ultra-Wideband) is defined in the IEEE standard 802.15.4 for micro positioning applications. Test requirements will continue to be imposed for testing Time of Flight (ToF), Two Way Ranging (TWR),

[7] https://www.intel.com/content/www/us/en/products/docs/wireless/wi-fi-7.html

[8] https://www.litepoint.com/blog/error-vector-magnitude-why-it-matters-and-how-its-measured/

and Angle of Arrival (AoA), at full spectrum bandwidth (see Table 8.1).[9] In addition to the 802.15.4 UWB standard, a new IEEE 802.11bk standard is emerging for micro positioning at 320 MHz bandwidth and thus such testing capabilities will need to be added when this standard becomes available in late 2024.[10] [11]

Beyond mobile and connectivity device test requirements, Table 8.1 also shows various other RF wireless applications requiring test capabilities in the millimeter wave range, such as Ka/Ku VSATs for satellite rural internet deployments[12], automotive radar in the 77 GHz and 79 GHz frequency bands for SAE levels L4-L5 autonomous driving, and other applications such as base transceiver station (BTS) backhaul, and hand gesture/motion detection applications.[13] These miscellaneous millimeter wave use cases are likely to require similar test capabilities as explained above for 5G FR2-1 and FR2-2 mobile devices.

8.3 Test of Photonic Devices

Executive Summary

In the electronic integrated circuit (EIC) industry, testing has become a mature process supported by practices and equipment that have been heavily optimized to drive down the cost and time spent on IC testing. In contrast, development of similar methods and tools for the photonic integrated circuit (PIC) community is still at an early stage, and the extra complexity that arises from having to measure both in the optical and the electrical domain poses many challenges. In this section. we define a number of key areas where development is needed, and in each of these areas we strive to leverage as much as possible the existing knowledge, practices and infrastructure from the EIC industry.

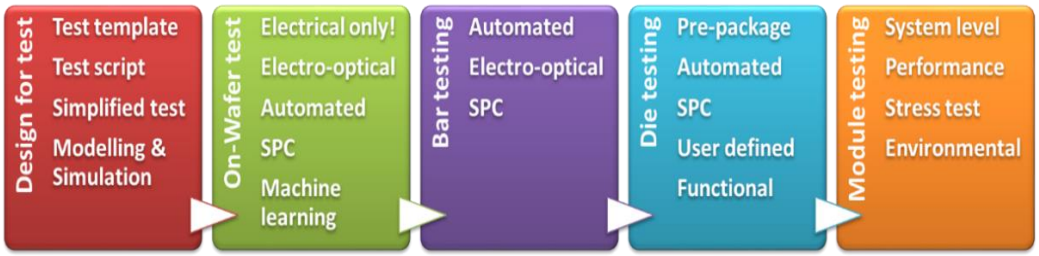


Figure 8.1: Overview of the test processes across the manufacturing chain of photonic integrated circuit based modules. Statistical process controls (SPC) require adequate test methods and data collection plans which should be accounted for already at the design phase

[9] https://www.litepoint.com/uwb/

[10] https://mentor.ieee.org/802.11/dcn/22/11-22-1353-02-00az-11bk-320mhz-ftm-csd.docx

[11] https://www.ieee802.org/11/Reports/802.11_Timelines.htm

[12] https://www.prnewswire.com/news-releases/satellite-internet-roll-out-to-gain-momentum-in-rural-areas-factmr-projects-c-band-to-remain-preferred-frequency-band-301404693.html

[13] https://www.infineon.com/cms/en/product/promopages/60GHz/

A summary of photonic device test methods is available **at this link**. Based on that information, we see three key development areas:

- Standardization of test metrics
- Consolidation of design and test workflows
- Test time reduction

8.3.1 INTRODUCTION

This Test section focuses on unique attributes of testing optical devices, concentrating primarily on testing data communications products.

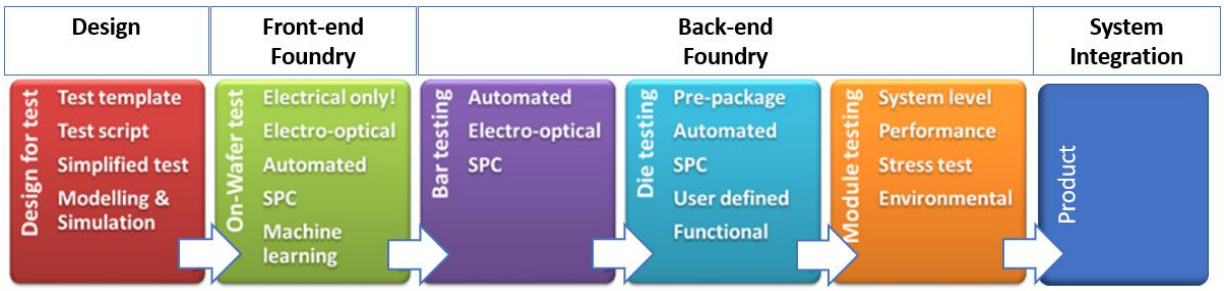


Figure 8.2: An overview of the PIC production chain for test.

In each step of the test chain that is followed by the components that will form an end product, different requirements and methods are used. This chapter will discuss both the separate steps and the connection between those steps, regarding the product and data flow.

Areas of testing needed during a product life cycle are:

- during development to prove functionality and de-bug devices
- qualification testing
- pre-production validation
- in-process production testing to assure product quality, reliability and to improve yield.

This section contains an overview of PICs made on InP, SiN, SiPh, GaAs, Polymers and CMOS platforms. Elements such as fiber couplers, fiber arrays, lenses, optical and electrical interconnects and the standardization of test port positions (optical, DC, RF) will also be discussed. The kinds of testing required vary over the life cycle of a product (Figure 8.2). This figure lists typical optical device test activities and requirements during the life of a device from conception through the in-use and end-of-life phases. A roadmap of quantified key attribute needs is available **at this link**. Considering that data, a projection of the key industry needs is shown in Table 8.2.

Table 8.2: Key challenges with respect to test between 2020 and 2040

	2020	2025	2030	2035	2040
Adopt semiconductor EIC industry test practices					
Test procedures from custom to standardized					
Standardization of test structures					
Test data exchangeability and analysis					
Technology agnostic testing					
Test automation					
Design for test					
Application agnostic testing					

Red: Not current industry practice; **Orange:** Partial industrial coordination; **Yellow:** Significant industrial coordination and compatibility; **Green:** Established Industry standard.

Each category is broken down in more specific subcategories in the following tables (Table 8.3- 8.8), following the same roadmap guidelines. Each table addresses areas such as key challenges, test practices, transition from custom to standardized procedures, transfer of data, adopting semiconductor test practices, and Design-for-Test both at the die level and the software level. The tables show competences going out beyond 5 years and emphasize relative strengths for each area.

Table 8.3: Adopt semiconductor EIC industry test practices

	2020	2025	2030	2035	2040
6 Sigma methodology					
Documenting and reporting					
The same metrics but methods may vary					
Optimized test at wafer-level					
DC testing in electrical – electrical domain					
Revised accept-reject methodology					

Table 8.4: Transition from custom to standardized procedures.

	2020	2025	2030	2035	2040
Standards instead of custom approaches					
Prioritize tests across full PIC value chain					
Testing metrics					
Relevance of a test					
Standardized test structures					

Table 8.5: Transfer of test data across the PIC value chain

	2020	2025	2030	2035	2040
Implementation in PDK **Improved design tools (EPDA)**					
Correlation of the test outcomes **Improved processes** **Identification of redundancies**					
Accessible scope – potential IP issues					

Table 8.6: Technology-agnostic testing

	2020	2025	2030	2035	2040
Across (currently) major technologies InP, SiPH SiN, Electro-Optic (EO) polymers					
Open for emerging platforms polymer, diamond, rare earth ion doped, three-dimensional (3D) PICs, SoC (high temperature)					
Hybrid integration photonic cross platform electronic-photonic chip level (EPICs) electronic-photonic PCB-chip					
Testing PICs with CMOS circuits/testing					

Table 8.7: Automation of test at wafer, bar, die, module and system level testing

	2020	2025	2030	2035	2040
Wafer - level					
Bar and die – level testing					
Standard test interfaces (layout templates)					
Technology agnostic					
Scalability					
On-chip self-diagnostics (Utilizing electrical-to-electrical testing)					

Table 8.8: Design for test

	2020	2025	2030	2035	2040
Test oriented layout templates					
Implementation in PDKs					
Test scripts for generic die testing					
Training of PIC designers					

8.3.2 Situation Analyses

A situation analysis of photonic testing is available **at this link**. It covers topics such as:

- Manufacturing processes
- General Test Equipment
- Critical Infrastructure Issues
- Technology Needs
- Prioritized Research Needs
- Prioritized Development and Implementation Needs
- Workforce Development

8.3.3 Gaps and Showstoppers

8.3.3.1 STANDARIZATION

Standardized testing metrics and procedures are essential for developing PIC markets further. Some specific killer applications (interconnects, automotive, sensors, etc.) are needed to accelerate standardization. Necessary test items depend on a particular application, and a specific application makes them clear. A large market opportunity provides a powerful incentive for PIC companies such as PIC device companies, PIC foundries, and PIC testing equipment companies.

Necessary test items should be standardized across the full PIC value chain. Testing designs and procedures are then standardized. The design tools for testing should be implemented in EPDA and PDK. Testing should be accurate and fast. On-chip self-diagnostics like that for EICs will be needed in the future.

PIC device engineers need to clarify testing equipment specifications (electrical and optical probes, functionalities, accuracy, speed, etc.). They should collaborate closely with PIC testing equipment engineers.

Standardization seems a difficult challenge in this field because it needs many people's efforts and some sufficiently attractive markets. If this challenge is achieved, we will be able to develop various kinds of PIC products with a minimum of effort.

8.3.3.2 PLATFORM-AGNOSTIC TESTING

The basic testing setup is common in a variety of PIC technologies (SiPh, InP, GaAs, SiN, polymer, etc.). Technology-agnostic testing is very important. The standardized testing equipment should be used for a variety of PIC testing with minor modifications. Various PIC companies should cooperate with each other across technical boundaries. The PIC devices are tested at a variety of sample shapes (wafer, bar and die). Sample-shape agnostic testing is also very important.

8.3.3.3 AUTOMATION

Fully automated PIC testing equipment is essential for developing PIC markets further. Mature EIC industry test practices should be emulated, and original PIC industry test practices should be developed. Various types of fully automated transceiver testing (OOK, PAM4, QPSK, 16-64QAM, etc.) will be needed. In addition, as co-packaged PIC and EIC devices ramp, the availability of a comprehensive PIC/EIC ATE based test solution will become critical.

8.3.3.4 HIGH SPEED (RF BANDWIDTH) TESTING

PIC testing equipment must measure both low-speed and high-speed properties. Fully automated high-speed electrical test (>10-100 Gbps) at wafer level is not easy. Adding to this the need to optically connect to the DUT via either a horizontal or vertical coupling approach, and the challenges become both risky and costly.

8.3.3.5 OPTICAL TESTING FOR MANUFACTURE

Contactless and non-destructive inline optical testing equipment with no particle pollution, which is acceptable for a PIC fab, will be needed. Inline optical testing can improve product yield.

8.3.3.6 USER SUPPORT

User-friendly GUIs and a variety of testing scripts are needed. PIC tests are generally difficult because electrical and photonic knowledge are needed. Helpful training manuals and courses are necessary.

8.3.3.7 ANALYSIS OF TESTING RESULTS

We have to research relationships between testing results at each level and product performance. A PIC accept-reject methodology should be established for each product. For example, one faulty sub-system does not necessarily disqualify functionality of the full circuit. In addition, statistics and analysis of testing data should effectively be transferred across the PIC value chain.

8.3.3.8 COST

Fully automated optical and electrical testing equipment will be very expensive. We should share expensive testing tools based on standardization and platform-agnostic testing. Testing time (including setup, calibration, wafer load and unload, etc.) should be short enough because time is money. But testing should be accurate enough.

We have to make the best use of testing results to achieve a good product yield and high product performance. The testing results should also be used to revise a product design and develop new products with much higher performance.

8.3.4 HIGHER PIC TECHNOLOGIES

Some specific applications help to solve the above problems. Higher PIC technologies are necessary to realize such applications. For example, low-loss propagation, low power consumption and high-speed optical modulation, photo detection and amplification, high temperature stability, high r33 materials etc., which translate into high performance, will be expected in SiPh, InP, GaAs, SiN, polymer, etc.

- The 50 GHz barrier resulting from conventional CMOS capability forcing parallel solutions rather than higher baud rates.
- Low speed of suitable assembly, test and other process equipment resulting in high costs.
- Inability to overcome the cost-driving, rate-limiting step/bottleneck of manufacturing/testing such as the number of assembly steps or length of time to perform test, especially BER testing. Lengthy test times increase expense.
- Limits resulting from adapting existing equipment, materials and methods to optical test as more specific equipment is not available. Currently the demand for such specialized equipment is not sufficient to incentivize equipment manufacturers to make it available due to high non-recurring engineering (NRE) costs and low return on investment.
- Designing for Manufacturing and test:
    - Maximizing output to reduce cost
    - Studying designs to trade off accuracy and speed

- Inability to utilize materials or processes due to environment-related constraints (RoHS, REACH, WEEE, etc.)

Recommendations for Potential Alternative Technologies

1. Silicon waveguides to 1D/2D photonic crystal waveguides or plasmonic waveguides. Some devices become much smaller (leading to higher-density photonic integrated circuits).
2. Combinations of active and passive polymers for alternative Silicon (and other) PIC designs and automated test, calibration and verification procedures.
3. Utilize laser processing to make optical waveguides in-situ for effective optical connections and optical structures.
4. Utilization of plasmons to minimize size and maximize functionality.

Contributors

Tom Brown, University of Rochester - chair

Dave Armstrong, Advantest - chair

Sylwester Latkowski, Photonic Integration Technology Center, Eindhoven University of Technology - chair

Robert Pfahl, iNEMI

Dan Evans, Palomar Tech.

Bill Bottoms, 3MTS

Makoto Okano, AIST

Tobias Gnausch, Jenoptik

Jeroen de Coster, IMEC

John MacWilliams, Consultant

Chris Coleman, Keysight

Keren Bergman, Columbia University

Sam Salloum, Tektronix

Scottie Wyatt, Tektronix

Rocio Banos, VLC Photonics

Robert Polster, Columbia University

Ignazio Piancentini, Ficontec

Eugene Atwood, IBM

Graham Reed, University of Southampton

Richard Otte, Promex Industries Inc.

Lionel Kimerling, MIT

Martin Möhrle, Fraunhofer HHI

Michael Lebby, Lightwave Logic Inc.

Chris Roeloffzen, LioniX International

Willem Vos, University Twente

Jan Mink, VTEC

Gordonn Liu, Huawei

Jan Peters Weem, Tektronix

Iñigo Artundo, VLC Photonics

Michael Garner, Stanford University

Philip Schonfield, RIT

Shangjian Zhang, UEST

Zhihua Li, IME

Yi Zhang, Teradyne
Zoe Conroy, Cisco
Fen Guan, Global Foundries
Carl Buck, Aehr Test Systems
Tsuyoshi Horikawa, Photonics Electronics Technology Research Association (PETRA), Japan

We would like to thank all who contributed to the Test TWG, both in workshops, on-line conferences and with written comments.

8.4 Logic Testing

8.4.1 Introduction

The use of heterogeneous integration to combine several chiplets into a multi-die package has more than offset the established slowing of Dennard scaling; the moniker of this being the "More-than-Moore Era" is quite apt. As measured purely by area, the amount of silicon in such a package can now far exceed that possible in a traditional monolithic package. For example, Intel's Ponte Vecchio package contains 47 chiplets with a total active silicon area of 2330 mm^2 [1] compared to the enormous monolithic Nvidia A100 GPU at 862 mm^2 [2]. As measured by logic complexity and the associated test requirements, a package containing this much silicon brings with it the challenge of testing for subtle defects in transistors and wires, but at the scale of what was a motherboard's-worth of functionality only a few years ago. This is in addition to the new test requirements associated with the 2.5D and 3D integration methods themselves. In total, the move to heterogenous integration has created a substantial increase in the number and difficulty of the tasks facing the DFT and test engineering communities. This section considers these tasks by grouping them into categories: access, yield, cost, quality, and time to market.

The first group of these new tasks involve basic access to on-chip test features, both at wafer sort, where the fine pitch of chiplet interconnects makes traditional probing problematic or impossible (see probe section of this Roadmap), and in the package, where only the package pins on the base die are accessible, through which all the other die must be tested. Besides these physical constraints, the bandwidth of the interface through which test data is exchanged with the device is another key consideration: test time and thus cost are directly affected. Furthermore, the emergence of a chiplet ecosystem where third-party providers can contribute silicon for package integrators to utilize is strongly dependent on standardized test interfaces which facilitate interoperability. A standard which should enable test access is IEEE 1838 which provides a method for describing, retargeting and distributing tests as well as physical interfaces for both data and control.

The second group of tasks revolves around yield. In heterogeneous integration, the cost of a test escape (i.e., a defective chiplet which nevertheless passes its (inadequate) wafer sort test) is no longer just the cost of that piece of silicon and the package; it includes the cost of all the other good chiplets as well, since reworking a package is considered to be impossible. This situation will likely drive two different responses. First, integrators may demand known good die from their silicon providers, which in turn will drive the test community to grapple with the cost, quality, yield maximization and die harvesting topics described next. Second, silicon providers and package integrators may collaborate on fault tolerant

schemes such as repair and redundancy for yield recovery, some of which may even be used throughout the life cycle of the product to gracefully deal with degradation over time.

The third group of tasks around cost extends those mentioned in the yield category by also considering the cost of the test features and the production test flow. Internal test features (scan, memory and logic BIST, I/O loopback, on-chip instruments, repair, redundancy, etc.) greatly enhance the testability of a device, but come at the price of silicon area, functional performance, and power. Similarly, adding extra steps in the manufacturing test flow (screening at multiple operating points, performing partial-assembly testing, burn-in, system-level test, etc.) and applying adaptive test techniques (part average testing, good die in bad neighborhoods, outlier detection, etc.) can reduce the number of test escapes, but increase the cost of goods sold. Die-to-Die interfaces between chiplets also present cost challenges as they are expensive to probe with today's methods and coverage is provided at later test steps which can result in higher scrap cost (mitigated with repair and spare lanes). Finding the appropriate features and flows to support the financial models will require many trade-offs.

Quality has a strong bearing on cost and yield as described above, but takes on two other important roles in a heterogeneous integration environment. First, given that a single device may contain silicon from several fabrication facilities and go through a multi-stage assembly process, managing the value delivery chain will be extremely challenging unless each participant in it measures and delivers to very high-quality standards. Second, since the products which utilize these multi-die packages will initially be in high-end markets (e.g., hyper-scale data centers, supercomputers, automotive, etc.) where data integrity is crucial, the absolute level of quality is a key consideration.

Lastly, despite the rising complexity of the devices, levels of integration and the increasing challenges of manufacturing, Time to Market (TTM) is of paramount performance. The time for tests to be developed and qualified for release has not increased with respect to product development time. To ensure that chiplets continue to support TTM efficiency, tests will need to be developed as IP which is retargetable at the various levels of integration and further standardization of test delivery interfaces to ensure interoperability between multiple vendors.

These five groups are clearly intertwined: high quality requires excellent test coverage which often involves expensive test time but can be modulated with high-bandwidth test access and internal test features, but those come at the cost of extra silicon area which can reduce yield and raise both cost and the likelihood of defects (not to mention the negative impact on mission-mode performance and power). Finding the optimal path through these More-than-Moore challenges will require solid engineering. As chiplets are integrated from multiple providers, collaboration on test approaches and coverage methods will be of increasing importance. The following sections address these topics in more detail to help address this engineering work.

Key take-aways in the sections that follow:

- Test content continues to grow with the number of transistors at the die level
- Chiplets will provide additional challenges to traditional logic test with mixing methodologies and approaches
- Quality levels will need to improve to support product economics, and new test methods will be required
- New test methods are emerging for deploying logic test
- Silent Data Corruption (SDC) is driving logic testing methods into deployed products
- Chiplet vendors will need to provide retargetable test IP for the next level of integration into SIP or system

8.4.2 Addressing the architectural bottlenecks of logic scan infrastructure

In previous versions of the HIR logic roadmap section, we have assumed that the fundamental approach for Logic ATPG would remain the same. As such, the roadmap focused on metrics such as scan data volume, data rate of interface, compression factor, and test time. In this version of the roadmap, we are highlighting the impact further integration has had, where the economics of test have driven a repartition of how scan is delivered to a Device Under Test (DUT) and how it is applied. To help delineate which challenges are classical logic scan challenges and which are changes to architecture, our discussion is broken into sections: Traditional scan challenges; emerging use-cases; updated scan architecture; and evolving logic test beyond scan testing.

8.4.3 Traditional scan challenges

With the progression of logic density, we continue to see the proportional growth of test data volume. As was noted in the 2021 roadmap, the effectiveness of compression at the block level is slowing. Other techniques of data compression for multiple instances of identical cores have shown increases at the chip level. If classical scan delivery methods are employed, then the scan frequency is also limited. In the emerging scan challenges, we will highlight new approaches that provide further improvements. The role of continued scan pattern growth is highlighted in Figure 8.3, which illustrates the resultant test time growth that explodes as multiple die are integrated into a single package ("SOC" in the figure). Further modeling will be done in the next roadmap update to capture the impacts of the trends discussed in this document.

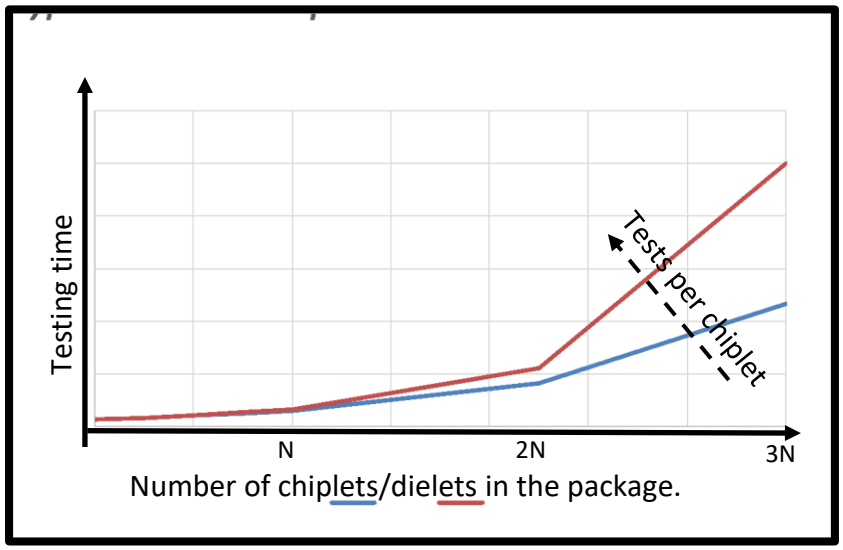


Figure 8.3: Typical Test Time Expectations

Though it is the easiest to model, scan-based testing is not the only driver of test time increases. Other test actions (BIST, functional test, parametric test, analog test, trimming, repair, volume diagnosis data collection, etc.) contribute as well, and have also been growing. The mix of these test types, along with the insertions (wafer sort, package test, system-level test) in which they are applied, factor into the calculation of overall test time. There is no industry consensus on what constitutes the optimal solution

for maximizing quality while minimizing cost, and the specific implementations vary by market segment and by company. Given that reality, the remainder of this analysis will focus on scan-based testing.

Very few SoC sub-cores begin their development process without considering a scan compression architecture into which they will fit. If standard scan architectures are applied, it is typically for very small components or IPs which will later be embedded in larger blocks which will then include a compression architecture. Scan compression has succeeded in reducing test times for manufacturing test and reducing data storage and transport costs. Typical SoCs use a homogenous approach among all cores within a die. But this is not always the case. It is anticipated that heterogeneous integration of disparate die may also include compression schemes from various EDA vendors. Helping to support hierarchical integration and the notion of merging pattern formats from various sources is a common goal of core wrapping. Most SoCs (and therefore heterogeneous package assemblies) are composed of wrapped cores. The patterns for these cores are developed at the core levels and retargeted (or ported) to the top level for eventual ATE application. Die stacking simply adds more hierarchical porting layers to the retargeting solution.

Practical issues users should consider when merging these many core-level pattern sets together into a manufacturing test pattern set include ATE resources, wafer or package-level access resources, test time, and power and thermal requirements and constraints. Test is a power-hungry application, and thermal issues are exacerbated by heterogeneous packaging. Solutions which integrate patterns for all these cores should consider topological proximity, and power and thermal responses when combining patterns for simultaneous application. Compression schemes have incorporated built-in power-reduction techniques for some time to help alleviate the shift switching activity profile for an individual compression codec. In addition, there are hardware resources one can add automatically to further reduce capture power or help ATPG easily reduce capture power. To help automate test scheduling of modules across a stack, more sophisticated power-related data may need to be introduced along with physical topological information to help test schedulers shorten test times while not overrunning power and thermal constraints.

Recently, test data propagation fabrics have emerged from EDA tooling to help address resource allocation issues in multi-core and multi-die packaging applications. Moving large amounts of test data long distances, or simply making use of various data types from circuits sprinkled across a vast surface area, has presented a problem not unlike functional compute and memory applications have always had. Again, heterogeneous packaging applications have only exacerbated the issue. Today, several “scan fabric” solutions are available. These might present a fixed-rate scan bus which adapts its bandwidth to the core endpoints as data moves from tester to core. Moreover, this interface might branch and maintain data speeds as fast as the intervening technology would allow, ramping down clock frequencies and adapting to core-level endpoint resource requirements as necessary.

In addition, there are solutions that seek to reduce ATE data requirements by leveraging the fact that many designs contain multiple identical cores. Of course, broadcasting a single set of stimuli to a group of cores reduces data volumes. But to help further reduce test data volumes, unique solutions exist which collapse response data to a minimum and reduce test times as well. For example, the response data can be scanned in and broadcast along with the stimulus. Each core then can determine its own correctness and store that, or scan out a composite result to help reduce data volumes. Or a MISR can be employed at the compressor outputs to further compress the resulting signature of a passing or failing test pattern or pattern set to a minimal amount of data. One can even initialize the MISR such that the resulting signature for a passing pattern set is zero (all 0 values) and this is easy to compare at the core level to compress the pass/fail result to a single-bit response at the end of the entire test.

8.4.4 Emerging Use-Cases

Several emerging use cases are further driving silicon sensor IP applications and DFT architectural decisions. In particular, re-use of DFT resources past the manufacturing stages and into the field have increased the value of these resources. Performance, safety, reliability, and debuggability applications have emerged as DFT IP and infrastructure has risen to address new functional challenges. Examples of these include solving adaptive voltage and frequency scaling applications, addressing the reliability crisis that silent data corruption (SDC) presents, leveraging system monitor IP to support debug operations in complex system environments, and using functional high-speed IO ports for in-system diagnosis scenarios.

Interestingly, the same IP that is used for in-system process, voltage, and temperature alarms and characterization can be used to support performance enhancements or reactions to measurements which exceed certain thresholds. For example, under a specific operational (software) load, a device could determine that there is headroom left for increasing processor speeds to address the running application. Additionally, one could use embedded monitors to determine that a device will soon fail catastrophically if not replaced due to path margin measurements on internal connections or between devices. Tester failures could be correlated with sensor data to aid the diagnosis process. And all of these resources could be accessed in-system during debugging operations. System debug availability is important. The ambient operational environment afforded by ATE is usually much cleaner and less stressful when compared to system applications. Functional high-speed IO can present a novel entry to solving these problems in-system, where and when they occur.

High-speed IO port use for supporting test and debug operations solves an interesting factory test application problem, as well. By leveraging a high-speed functional port or ports, getting data into and out of the device is no longer slowed by the limited availability of slow-speed pins on a package or die. Once the data is beyond the IO periphery, it can be expanded and slowed to frequencies more in line with the technologies and power constraints presented by each die in the package. When functional ports are leveraged for system-level debug, several considerations should be examined. First, the high-speed port type used has complexities of its own that may need to be tested prior to use. Applying an IEEE Std 1149.10 protocol and architecture to this application may help alleviate some of the manufacturing test complexities associated with high-speed ports in the factory. However, the tester will also need to support the IEEE Std 1149.10 protocol. In addition, for use in the field, 1149.10 may also need to be leveraged by the attached debug environment. On the other hand, the functional architecture and protocol can also be used. Still, one must consider the manufacturing test environment and the field application context before locking in a solution set. Second, data and system security should always be considered. A holistic approach is required to make sure user data and device circuitry is protected from abuse by those wishing to steal that data or leverage those circuits for improper or illegal purposes.

8.4.5 Updated scan architectures

Test compression schemes introduced the first level of separation between external interfaces and scan chains. This helped increase the number of internal scan chains as well as reduce the scan chain length, thereby optimizing both test data volume and test application time. However, with heterogenous integration of multiple dies on a single package, ever-decreasing pin-to-gate ratios, and the dwindling number of available data pins (for example GPIOs), the ability to deliver scan data is a big challenge both for wafer and package-level manufacturing tests. To address the scan bandwidth issue, there are two things that need to be considered:

- Delivery of scan pay-load at a much higher speed via a small number of GPIOs or functional interfaces
- Distribution of scan-data within a die or across dies using a scan network/bus that can be operated at a much higher speed relative to the traditional scan rate

The delivery of scan data at a faster rate addresses the concern related to the volume of scan data that needs to be delivered using a narrow interface. The ability to deliver large amounts of scan data enables concurrent testing of hundreds of cores for large modern designs targeted for a wide-range of applications. The data organization at the interface can now be separated from the structure needed at the IP blocks. As such, the user can think of the scan data as pages of information or packets of information (note this is different than protocol packetization which includes encoding schemes). The packetization of scan data further helps in reducing the dependency on the number of IOs available for every codec within a design. This makes the tasks of test planning and test reuse much simpler, as any number of internal codec pins can be driven when delivering packetized scan data via a scan bus. A benefit of this architecture is that data payloads no longer require padding to balance scan chains, so memory can be used more efficiently by the test equipment.

When test compression was introduced, it relied on having a codec driving a large number of short chains within a core. It exploited the small number of specified bits needed to target faults in a design, and therefore, implemented lossless compression techniques by delivering the required information via a few scan channels. In the 2021 roadmap document, it was indicated that test compression ratios obtained via classical techniques will taper-off with increasing design complexity and improved ability for ATPG tools to pack more faults into a single pattern. Instead, compression will have to rely on a design trend with numerous identical cores, where ATPG tools (in addition to compressing test data) will have to re-use the same pattern set for identical cores within a design, thereby reducing scan data volume. Moving forward, with heterogenous integration of cores, packetized scan data delivery allows usage of data throttling (control the flow of data depending on cores that need the most) to manage integration of tests across multiple cores and pushing compression of test data even further. In other words, test compression improvements in the future will depend on a variety of techniques that are dependent on design characteristics and styles that go beyond just test data sparsity.

One of the characteristics of a modern design is the presence of hundreds of cores. Having a bus-based scan architecture allows delivery of scan data to hundreds (or thousands) of identical cores in parallel, and either observing the test responses or performing a local compare of the responses on-chip assuming the responses along with the masking data is also streamed to individual cores. This results in further improvement of test efficiency by reducing the test data that needs to be stored and improving the performance as data doesn't need to be read back and compared on an ATE.

Power dissipation during test has always been a major concern. With the ability to deliver the scan payload via a high-speed bus to many cores simultaneously, power dissipation becomes a bottleneck related to how many cores can be tested in parallel. This calls for localized generation of scan control signals such that one can perform independent shift and capture for each core in an asynchronous fashion. Asynchronous shift and capture between cores allow one to manage the voltage droop or IR drop that are usually associated with scan test in a much more efficient manner, thereby not only increasing the number of cores that can be tested in parallel but, in many cases, help in increasing the shift frequency.

For designs with hundreds of identical cores, broadcasting the stimuli, responses, and masking data to these cores reduces the volume of test data that need to be stored. However, there is a need for implementing efficient techniques to facilitate volume diagnosis. For example, when implementing on-

chip compare for identical cores, one can determine the failing cores by inspecting a (failure flag) sticky bit at the end of test. Once the failing cores are identified, those cores can be targeted for re-testing and the failing responses can be observed to drive failure capture at ATE and diagnosis. In addition, there is ample opportunity during manufacturing test to optimize a test session based on how different tests can be scheduled and applied by considering test time requirements, as well as various environmental factors such as power, thermal gradients, power supplies, ATE throughput, etc. Additional factors that impact how tests are applied can also be related to failure data collection needs and limits. Based on the conditions in the DUT or the test needs, ATEs can play a significant role to modify and optimize the test sessions. It can drive data collection that would help modify and adapt the tests for subsequent test insertions. The diagnosis and power use-cases highlight that if tests are augmented with additional meta data, the ATE could provide further intelligence for execution which will also result in additional memory savings.

8.4.6 Evolving logic test beyond scan testing

As the cost of test escapes grows, it is important to try to move as much of the test content as far left in the manufacturing process as possible. Many complex devices still rely on some amount of functional testing or system-level tests to close the gap between what is testable though structural DFT techniques and mission mode. As the complexity of the chiplets of the system has grown, more of the design can be put into modes that more closely match mission-mode during test. This in enabled by having enough on-die memory so that tests can be executed internally in the chip. It also requires system hooks to support running without the external devices that would be seen in a full system. Some mission-mode capabilities such as power state and clock control can be quite challenging to shift to a production test environment.

The resurgence of functional testing has driven innovation in how tests are generated and deployed. One issue with functional tests is how effective a generated test is at detecting a fault, given a limited set of interfaces and a finite amount of time. Today, the use of functional tests is largely based on empirical experience of test escapes where symptoms of an undetected fault have impacted a software application running on the hardware. Manual effort identifies and transforms useful code snippets into functional tests; this is analogous to scan testing 30 years ago prior to the extensive automation of structural test. Extending such automation into the functional test domain will require the tools to create tests and measure their fault coverage to enable an efficient test suite for production testing. One promising technology is the Portable Stimulus Standard (PSS) which was proposed by the Accellera System Initiative. PSS takes a requirement definition, design model, and available interface descriptions for tools to generate tests that cover each requirement definition. This technology was developed for chip-level verification to prove that designs meet their operational requirements. The challenge for the test industry is to optimize the mapping of the fault space into the requirement space for coverage while also optimizing the test run time to make each test economical. These functional tests may benefit from another interesting technology called Quick Error Detect (QED, developed at Stanford University) which instruments functional tests using temporal and spatial duplication to speed detection and that can backtrack a detected error to a physical fault condition to guide how to precondition the hardware with minimal test time. These technologies will be required to make SLT testing more effective by limiting the time per test, making each test more effective, and enabling test coverage metrics.

8.4.7 In-System and SLT test requirements driving new logic test requirements

Testing of logic has been extending past the traditional factory test insertions of wafer and package to new areas that span from initial manufacturing throughout the device's operational lifetime. Previous heterogeneous integration roadmaps highlighted the rising use of functional testing with a System Level Test insertion as well as the use of MBIST and LBIST as part of the ISO26262 standards for periodic testing of electronic components in the field. What has gained more attention recently is the vulnerability of circuits to Silent Data Corruption (SDC) impacting complex digital devices in the data center. The risk of SDC is not new, but with the scale of modern data centers the occurrence of such errors has become measurable, and their detection, mitigation, and impact cost to the service provider has become an important topic. In 2022 at the International Test Conference, a major service provider stated that SDC events in a data center could affect as many as 1 in 1000 devices and manifest as applications producing incorrect results. However, there is not yet a consensus on how to measure SDC, nor is there a definitive breakdown of the root causes for these events. The industry sentiment is that we are only seeing the tip of the iceberg of this fault type, and new techniques will be required over the next five to ten years to drive down their rate of occurrence.

Historically, SDCs have been primarily thought of as a symptom of radiation-induced bit flips, and successfully mitigated accordingly. Today, there are multiple additional hypotheses about the possible causes of SDCs: 1) manufacturing defects that were not detected with traditional test flows; 2) latent defects that emerge due to aging effects; and 3) electrical effects (such as di/dt-induced voltage droops, IR drops, thermal gradients, etc.) caused by computational workloads which reduce design margins. The test industry is uniquely positioned to confirm or deny these hypotheses using techniques like extended characterization, root cause diagnosis, and in-situ monitoring. The best-known-method is still being explored and discussed and may well be a combination of approaches.

In the last decade we have seen an improvement in the physical realism of fault models by using the Cell Aware methodology, and this technique is expected to continue to evolve with emerging transistor technology (with related impacts reflected in the vector depth prediction of this section). Using superior fault models addresses the first SDC hypothesis by producing patterns that close the gaps from traditional methods that result in test escapes. It is important to note that, no matter how good the fault models, scan-based structural tests do not mimic the electrical conditions present during mission mode, so functional test will also play a role in catching test escapes. In addition to scan, functional techniques like PSS from Accellera, described in the last section, can be used to augment the test coverage.

Up to this point the discussion has focused on detectable faults at time zero; the degradation of circuits over time is the second hypothesis to consider as a cause of SDCs. Even if devices were all made perfectly, given the tiny size of the transistors as well as the stresses during use (temperature, voltage, current, mechanical, etc.), the way they operate over time will shift. For example, the resistivity of the power grid could increase over time due to thermal variations and current load. When the power grid changes, it will result in lower voltage delivered and the transistors will operate more slowly resulting in less margin. There are also well-known effects at the transistor level that will impact the design margin with respect to operation (NBTI, HCI, TDDB, etc.). One way to prevent SDCs from occurring may be understanding how the performance changes over time with respect to key performance parameters like timing margin, voltage, temperature, and device activity, then compensating for aging by adjusting the supply voltage or clock frequency accordingly. One challenge is how to implement such an in-situ control system to minimize the cost both in circuit area and impact to the end system.

One example from ISO26262 for automotive products involves the application of "key-on/key-off" tests which perform MBIST and LBIST in the field to re-validate the absence of faults before and after every

use. The impact to the end system is defined by a required run time and the periodicity of testing, along with the higher-level architectural features to initiate the test and evaluate the results. It has also been noted that on-chip variation within a device has been increasing, so it is expected that the aging of each individual path will also become more important to measure. The solutions of the future must look at how the critical circuits or paths change over time and be monitored (ideally) while the system is running to measure the reduction of design margin over time.

These requirements are different from our traditional testing techniques that are focused on structural correctness, not operational impact. To better understand the root causes of aging, more sensors at the block level within a device will likely be deployed. These test methods will also need to comprehend how often to collect data, the data flow within the device, and driving measurements to actions within the final product. In some cases, this will be done on-device in the field for mission critical systems or in the cloud for fleet monitoring testing applications. As a result, we are presented with a new opportunity of where and when digital test is applied and how the outcome of testing will impact end-product operation. New features such as extending life with active voltage variation, predictive maintenance, or new repair methods at subsystem levels for compute elements are all within the realm of possibility when test features are made available in the field.

The third hypothesis about SDCs is that they arise when the dynamic effects of stressful workloads push the electrical environment on the chip past design margins. To maximize hardware performance in this era of post-Dennard scaling, aggressive design margining has become common – including the use of Dynamic Voltage and Frequency Scaling (DVFS) to alter operating conditions in real time based on the workload (generally to maximize performance per watt – which corresponds to minimizing excess margin). Furthermore, in modern digital design processes, detailed automation tooling accounts for various parameters like switching factors to predict reasonable device activity which leads to other design features like current estimates and power grid sizing to ensure that no excess margin is left on the table. However, to achieve the highest performance without risk of exceeding design margins, one must be able to understand the impact of the software running on the hardware which enables reducing guard bands and reaching the highest performance. In large multicore architectures, this leads to adjusting the scheduling of cores to ensure balanced activities across the chip with the best performance. To realize this, additional sensors must be deployed to characterize and monitor the impact of software running on the hardware to adjust operating point parameters over time (as DVFS uses to ensure correct operation with optimal energy use). In the future, there will be the need for new data sources (extensions to voltage, temperature and timing margin) to enable further performance improvements.

8.4.8 ATE equipment challenges with the logic testing

Multiple trends are driving the test industry to develop new test methodologies which leverage high-speed IO (HSIO) to communicate data to the DUT in new ways. Most high-end devices (the ones which have the biggest test challenges) also have one or more high-speed protocol-based interfaces such as USB or PCIe available on them. Using this high-speed interface can provide two core values; 1) they provide a high-data bandwidth conduit for test, and 2) they provide a consistent test interface which can be used throughout the lifetime of the device.

Leveraging the existing HSIO interface provides an efficient way to enable many different types of tests, such as:

- Scan Test (including scan test networks)
- Functional Test

- Processor-enabled BIST
- On-chip instrument access (e.g., sensors): e.g., internal I/F to IJTAG
- MBIST and LBIST

Many traditional tests, such as scan and BIST, can be initiated over HSIO. Also, functional tests can be performed because they can be based on data payloads when the ATE interfaces are considered. The HSIO provides a fast way to load test setup information (such as arrays of coefficients) and test data sets (such as training sets) into the device for real-world confirmation of convergence and functionality. Additionally, functional tests can be executed between different cores on the die or between one chip and another in a heterogeneous integration situation, perhaps under the enablement of an on-chip processor.

The consistent HSIO test interface also allows leveraging test content from between test steps such as wafer, final, system level test, and in-situ testing after deployment. As such, it efficiently provides value through test consistency and reuse across many test insertions including end-of-life (RMA).

Enabling this type of testing, however, does require a new type of instrument in the ATE system. Key characteristics of this new ATE resource include:

- High-performance signal integrity
- The ability to enumerate the HSIO successfully, and if unsuccessful to diagnose the problem
- An industrial grade, integrated high-performance compute and software environment which mirrors the targeted real-world
- Very deep data storage array
- The ability to control the device JTAG interface
- The ability to do simple DC continuity testing

It is likely that many devices will retain, if possible, both a HSIO port for scan and functional test as well as traditional GPIO and JTAG interfaces in order to avoid the cost of additional test instrumentation at ATE-based wafer probe and package test insertions. The HSIO interface would be leveraged at system-level and in-situ test insertions where the other interfaces are not accessible.

A critical component of success, if an HSIO is used for scan and functional test, is that the interface adheres to a standardized protocol such as IEEE 1149.10 or standard PCIe. If the interface is based on some proprietary protocol, then it is difficult, if not impossible, to replicate that protocol on commercial test equipment due to implementation or IP protection difficulties.

It is assumed payload information is customized based on the DFT implementation and/or data security concerns. As noted above, this will drive the need for significant computational resources in the test equipment to construct and de-construct payload information in real time using custom software.

Lastly, it is critical that some DFT is available to validate the basic functionality of the high-speed interface prior to any other testing. Ideally, a self-test can be performed using scan and, if possible, an at-speed loopback test that utilizes an internal test path to eliminate the need for high-speed switching on the test fixture that would be needed for an external loopback.

Table 8.9: Updated predictions of Test Metrics

Trend	Short term 0-5 years	Long term >5 years	Challenges
Scan pattern growth of >30%/year	Initial adoption of high-speed interfaces – USB and PCIe. Move towards packetized scan methods with test fabrics	High-speed serial interfaces carrying packetized scan data: more scan bandwidth Extending to D2D interfaces	Rate of adoption of new scan interfaces
Functional test resurgence	Beyond SLT, further adoption focused on portable stimulus	Functional test on ATE and SLT using software test libraries	Establishing coverage metrics
Demand for in-field testing growing due to functional safety	Re-use of DFT-based instruments at power-up e.g., MBIST and LBIST	BIST + software test libraries at power-up and on-line Safety critical requirements driving new functions	Integration of DFT-based BIST with mission mode control and reaction
Increasing IO interface challenges	Sacrificial pads and dedicated DFT interfaces	ATE infrastructure to contact advance interfaces	Electrical, optical, and mechanical interface sensitivities
Logic testing extending into field	Initial methods to describe aging and workload impacts to hardware	Provide coverage methods for Aging and SDCs	Impact and root cause of SDC and aging continues to evolve

8.4.9 Conclusions

It is an exciting time for logic test. In this section we have highlighted the challenges and directions of logic testing (summarized in Table 8.9). We have shown that the classical challenges of increasing logic density are still driving the need for increased testing. Given the volume of data and emerging fault types, we also discuss new methods for test delivery as well as expansion of where logic test will occur. Heterogeneous integration will provide further product economic pressure to accelerate solutions for the challenges outlined above. Many of the emerging use case solutions are starting to be addressed with initial solutions and we will evaluate in the next roadmap the adoption rate and impact to the test economics.

References

[1] https://www.techpowerup.com/292250/intel-details-ponte-vecchio-accelerator-63-tiles-600-watt-tdp-and-lots-of-bandwidth
[2] https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth

8.5 Specialty Device Testing

A classification of *specialty devices* was defined in industry roadmaps beginning in 2006, driven by strong high-volume market demand, but having odd test requirements. Examples are CMOS image

sensors, LCD drivers, MEMS devices (including multimode sensors), actuators, bio-MEMS, and similar non-standard devices.

8.5.1 Trends Impacting this Technology Area

The novel applications of mobile personal devices, IoT, healthcare/artificial organ, automotive/ADAS, smart industry, and emerging energy fields are key drivers of specialty devices where innovative testing technologies are needed to enable future processes such as 3D , chiplets, and heterogeneous integration with high yield during mass production.

The trends for technologies (Near Term < 5 years)

- The trends for multi-mode MEMS sensors are toward fusing multiple sensing functionalities together in one device with artificial intelligence processors.
- The technology trends for image sensors lead to highly integrated multiple wafers using a 3DS (three dimensional stacking) process with Cu-Cu (copper to copper) connection technology for directly connecting pixel chips and logic circuit chips. Cu-Cu connection does not require a specialized area for connecting pixel chips and logic circuit chips, as needed for conventional TSV connections. The first successful implementation of 3DS wafer processing of an image sensor was the BSI (Back Side Illumination) process which bonded a photo-sensor wafer together with a back-side mixed-signal data processing wafer. The next step in the image-sensor wafer-integration process adds a memory-cell wafer between the photo sensor wafer and mix-signal data processing wafer, which could enhance image performance and the speed of data processing in a variety of imaging applications such as 3D imaging, face recognition, and image capture, with frame rates over 1000 frames/second.
- The trends in new WLP (wafer level packaging) for image sensors are WLO (Wafer Level Optics) and WLCM (Wafer Level Camera Module), which stack optical systems on the image sensor wafer using a wafer-level packaging process to reduce the size of optical systems and increase efficiency of mass production.

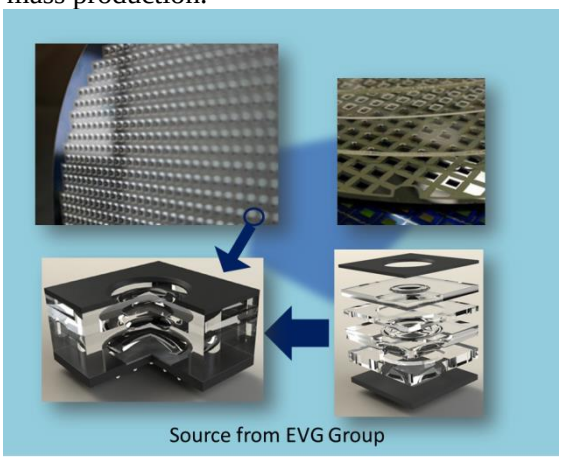


Figure 8.4: Image sensor WLO (Wafer Level Optics) packaging

The trends for technologies (Middle to Long Term < 15 years)

The automotive, robotic, medical and intelligent artificial organ fields are next-wave drivers of specialty devices which impact technologies in the medium and long term:

- Reliability will become critical for specialty devices. Burn-in and tri-temperature testing will become necessary test procedures during mass production.

- Built-in self-diagnostics, self-calibration and compensation, and self-repair technologies will become important design skills to apply to specialty devices for enhancing reliability performance.

8.5.2 Concerns: Test Challenges

LCD display drivers:

LCD display drivers are unique because of die form factor, which can have larger then 10:1 aspect ratio and thousands of very narrow gold bump pads requiring contact for test. In 2022, in-line and stager probing pad width for LCD display drivers already was down to 11μm in production and 8μm in development. Right now, only the cantilever probe card provides a major cost-effective solution for achieving probing of the LCD driver with such narrow and fine pitch pads with gold bumps in mass production.

An upcoming test challenge is that the data transfer speed for I/O will increase to 2.5 Gbps and is predicted to be up to 6.5 Gbps within 10 years. We need to overcome the challenge of probing fine-pitch bumping pads with high-speed signals with economical probing solutions.

Image sensor devices:

Testing of image sensor devices needs to consider special test requirements for optical systems and the resulting massive image data processing. Special requirements for optical test systems will be different and be relative to applications (see Table 8.10).

Table 8.10: Special specifications for optical test systems and applications

Illuminator Specification \ Application		Industry	Automotive	Consumer	Mobile
Wavelength Range	UV (100~400 nm)	V			
	Visible light (400nm~780nm)	V	V	V	V
	NIR (780~1400nm)	V	V		
	SWIR (1400~3000nm)	V			
High Intensity	> 10,000 Lux		V	V	
High Resolution	< 0.1 Lux		V	V	V
Polarized Light	0~360°	V			
Laser	PWM (Pulse Width Modulation)	V	V		
LED	LFM (LED Flicker Mitigation)		V		

Automotive ADAS applications and intelligent machine vision need the functionalities of image sensors with wide spectrum (from UV to FIR). high dynamic range and good S/N (Signal to Noise) ratio, fast data frame rate, and better quality and reliability, which challenges test system design. The burn-in solutions also need to include optical stress for sorting out defects in the coating process on photo sensor surfaces.

MEMS devices (Sensor, Actuator and Biological)

MEMS were successfully applied on various sensors for sensing motion, magnetic field, optic, sound, air pressure and vibration, flow, chemical composition of air, DNA sequencing, and other characteristics, and the market volume is increasing rapidly due to IoT, healthcare and automotive applications. Testing MEMS sensor devices with suitable physical stimulus and cost-effective solutions for the various types of sensors is difficult and tricky (Table 8.11). Testing the expanding kinds of fusion sensors will bring many test challenges.

Table 8.11: Specialty Device Odd test potential solution for a MEMS Fusion Sensor

Table TST20-Speciality Device odd test potential solution

				Year of Production	2022	2023	2024	2025	2026	2031	2036
MEMS Fusion Sensor	Process integration		Test Method	Challenges							
	IMU sensor (Accelerometer + Gyro)	CP / wafer probe	Probing MEMS wafer (DC only)	Probe card technology							
			Full functions (Multi-insertion)	Motion Prober system							
			3DS wafer, full functions (Single-insertion)	DFT design and implement							
		WLP	Test after dicing (Wafer form)	DFT design and implement							
			Test after singular (Package form)	Handling small size package							
		FT	Full functions (Multi- insertion)	Test cost is high							
			Full functions (Single- insertion)	Reduce test coverage rate							
				DFT design and implement							
			Burn In Test	SLT							
				BISX (Build-In-Self Test, Diagnostic, Correlation, Compensation/Repair)							
	Navigation (G-sensor+ Gyro+ Magnetic sensor + Barometer)	FT	Full functions (Multi- insertion)	Test cost is high							
			Full functions (Single- insertion)	Reduce test coverage rate							
				DFT design and implement							
			Burn In Test	SLT							
				BISX (Build-In-Self Test, Diagnostic, Correlation, Compensation/Repair)							
	Environmental Sensor (Pressure + Humidity + Gas) Sensor	FT	Full functions (Multi- insertion)	Test cost is high							
			Full functions (Single- insertion)	Reduce test coverage rate							
				DFT design and implement							
			Burn In Test	SLT							
				BISX (Build-In-Self Test, Diagnostic, Correlation, Compensation/Repair)							

Research Required
Development Underway
Qualification / Pre-Production
Continuous Improvement

DFT for MEMS sensor devices is new technology and needs research and innovative development for different kinds of sensor structure. MEMS sensors DFT needs to develop the stimulus source and sensor together in the MEMS structure as a BIST (Build-In-Self-Test) cell. When testing, the cloned control signal of physical stimulus is generated from the MEMS ASIC to enable the MEMS BIST cell to imitate physical stimulus for testing the sensor cell to achieve the DFT goals. This concept could also implement the technologies of BISD (Build-In-Self-Diagnostic), BSIC (Build-In-Self-Correlation/Compensation) and BSIR (Build-In-Self-Repair) to enhance reliability of MEMS sensors for automotive and medical applications. The key during testing is to make sure this BIST cell works well.

Beyond MEMS sensors, there are also actuator and biological applications such as micro-mirrors, MEMS speakers, RF switches, energy harvesting, microfluidics, micro-dispenser and artificial organs, plus others. The testing challenges for testing MEMS actuators and biological devices are that test methods are hard to standardize and depend on the structure for each different kind of MEMS device. Especially for testing biological devices, the test environment can be severe and there is a need to pass safety certification based on the laws for different grades and countries.

8.5.3 Summary

Specialty devices as defined have odd test requirements and are driven by strong high-volume market demand. Under these two conditions, the trends for specialty devices will be driven toward highly integrated multi-functions in one smaller unit to overcome ASP (Average Sale Price) erosion, and testing procedures will move toward high parallelism to reduce test cost. Test challenges will follow the same trends for heterogeneous integration to address testing for specialty products though cost-effective solutions.

Team Leader: Wendy Chen

References:

1. “EVG’s wafer-level optics (WLO) manufacturing solutions”, EVG press release, September 11 ,2017

8.6 Memory Test

8.6.1 Summary

- Memory is a growing segment within the semiconductor industry (~30% in 2021 up from ~10% in 2000).
- Higher bit density drives increased interface speed, power, and thermal management requirements.
- Smaller physical geometries challenge electro-mechanical interface capability of wafer and component test.
- NAND densities are projected to grow into >8Tb/die by ~2024, driven by continued growth in vertical scaling.
- DRAM bandwidth and densities are growing to meet growing demands of CPU/GPU applications.

From 2020 to 2021 worldwide semiconductor market revenue increased 26.3%, with Memory comprising ~29% of overall production, and growth in the Memory and Storage segments increased from ~10% of overall semiconductor revenue in 2000 to ~30% in 2021 [1]. End applications for the primary segments of the memory space (DRAM and NAND) have shifted slightly over the last few years, with changes in the DRAM segment seeing a flattening in demand for Mobile and PC applications, and an increase in demand for Datacenter applications, and with the NAND segment realizing the largest growth in the SSD segment for both Enterprise and Client applications [2,3]. NOR FLASH remains stable but becoming less relevant as NAND and DRAM growth continues.

As the demand for Memory applications has continued to grow and evolve over the last several years, the associated bit output has also grown due to innovations in architecture and technology that scale the density at a faster pace than package unit output (Figure 8.5) [2,3].

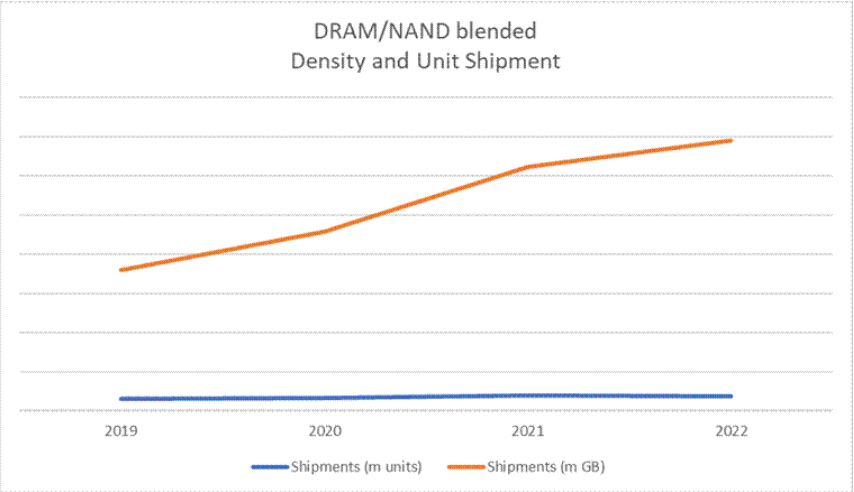


Figure 8.5 DRAM/NAND blended Density and Unit Shipment

Effective use of this increased density relies on higher interface speeds (UFS, PCIe, PAM) to access the data. The scale of these increased speeds for both DRAM [4] and NAND [5] (Figure 8.6) [6] begins driving additional power and thermal management requirements.

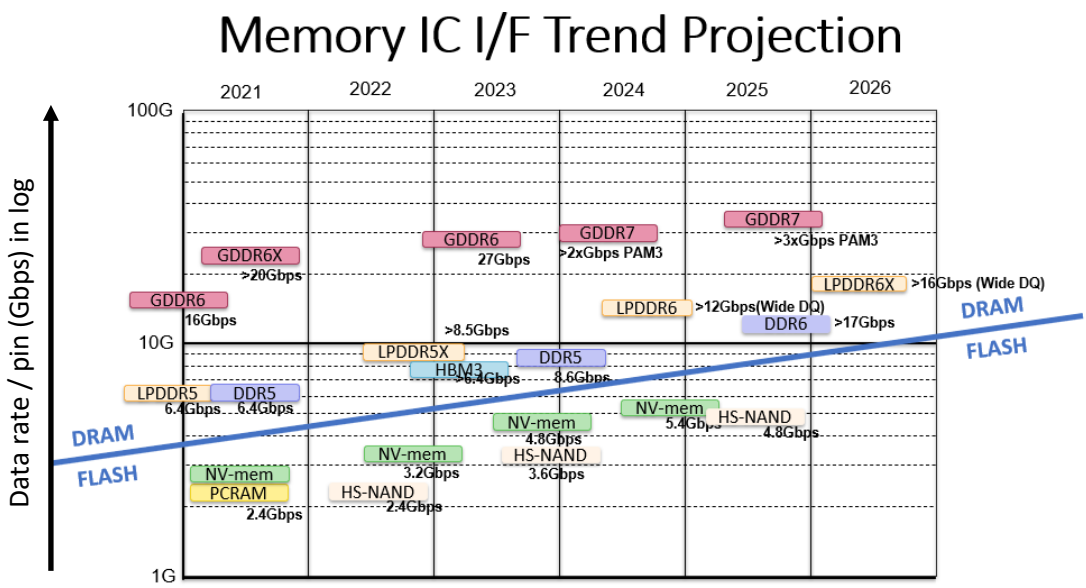


Figure 8.6: Memory IC I/F Trend Projection

From a Memory Test perspective, as the increases in bit output, interface speed, power, and thermal management requirements scale in both NAND and DRAM, challenges arise to meet the intersection of capability. Die sizes continue to shrink either through geometry or integrated scaling, resulting in higher Die Per Wafer at increased device density and speed. These shrinking die sizes create challenges at wafer test in terms of interface constraints – in many cases, the number of die that can be tested must be reduced in order to route signals, and to enable contact to the wafer. The interface pad size and pitch are also

projected to shrink below 50um in size, and the pitch of the pads creates challenges in signal routing, power delivery, and in some cases touching the pads has impact to the bondable pad area used for device assembly. As these key contact interface features scale smaller, and expansions in thermal demands grow to include coverage from -40C to 125C for automotive needs, wafer test interface thermal scaling must be proactively managed to ensure effective test coverage. From a power/thermal management perspective, with more power being delivered to smaller devices through the required range of test temperatures, proactive power dissipation at the device level also becomes a critical concern. For example, the overall growth from 2017 to 2022 shows a 5-year trend of ~9% reduction in voltage, but an increase of ~550% Die Per Wafer (DPW), and ~450% increase of power dissipation requirements at wafer test (Table 8.12)[6].

Table 8.12: Power Dissipation at Device Level

	~2017	2022	Scaling
device voltage	1.2V	1.1V	-9%
dies per wafer (DPW)	~500	2500-3000	500-600%
test equipment device power dissipation (wafer prober)	*100-150W/wafer*	*500-600W/ wafer*	*400-500%*
speculative trend			DPW scaling faster than test equipment power dissipation despite lower device voltages.

As the die size shrinks with higher interface speeds, Signal Integrity (SI) and Power Integrity (PI) become more challenging because the signals become more tightly arranged with smaller interfaces. Similar to challenges faced at wafer test with smaller pads, packaged die are also facing scaling issues, including BGA interfaces that are shrinking below 125um balls and less than 250um pitch. Added challenges include decreasing solderball heights (<100um), thinner packages (<500um), and increasing contact points on the interface which drive issues related to contact, thermal management, power dissipation, and handler drive force to optimally scale interfaces to the desired parallelism.

As bandwidth requirements increase, higher speeds and new interface technologies are emerging (e.g. PAM3, PAM4, wide I/O), and there is often a lack of agreement at standards consortia until very late. This challenges the development of tester technology to meet the evolving device interfaces in terms of technical risk, schedule, and cost.

For all test insertions, as the device density grows, more and more bits are required to stream from each device back to the tester for processing and analysis, potentially driving changes in tester architecture and IT infrastructure to manage growing bandwidth considerations.

8.6.2 NAND

Key NAND applications today include enterprise data and edge compute centers, the ADAS automotive cloud, plus local storage, gaming, and 5G applications. Data creation in these key spaces in 2022 hits a remarkable 100ZB, and is projected to grow to 200ZB by 2026, with a resulting 32% CAGR[7]. NAND device bit density growth from 2015 to 2021 grew from 64Gb/die to 512Gb/die [8], roughly doubling every 2 years, resulting in a 10-20% growth of associated test time every year [9]. Future bit growth is achieved in the transition from today's 2xx layers at ~2Tb/die, into projected >300 layers by 2024 resulting in ~8Tb/die [10]. This density increase is achieved through vertical scaling with thinner layers, lateral scaling with higher density layer interconnect, architecture scaling moving from CNA to Multi-bond, and in logical scaling moving from SLC to PLC. This bit growth in NAND is particularly challenging, as the industry has been testing all die at a wafer level in a single touchdown for the last 10-15 years [9], and with no simple way to scale interface parallelism, there is significant growth in demand for testers to meet Si output. Device speed performance is also increasing to move data from the device to the outside world. Asynchronous random reads enable faster bit access, and SSD interface speed

growth is moving from 2.4Gbps to 3.2Gbps to 4.8Gbps [10] to improve bandwidth and reduce latency. Interface standards compliment the trend as they move (for example) from PCIe G4 to G5 to G6, and the addition of high-speed SERDES interface memory controllers such as UFS 4.0 23 Gbps and PCIe G5 32Gbps provide the necessary support to double interface speeds about every 4 years [11,12].

In next-generation interconnect and speed, CXL3.0 is driving towards the next hyperscale applications. The CXL fabric architecture is intended to solve cost and bandwidth issues that DRAM-only solutions cannot address, all at a projected 64GT/s with no added latency above CXL2.0 [13]. This adds further complexity in signaling and throughput from a Memory Test perspective, as many traditional ATE interfaces are architected for adaptable re-use, and not architected for high bandwidth applications.

8.6.3 DRAM

PC DRAM transitions are beginning to occur from mainstream DDR4 to DDR5, largely in an effort to increase effective bandwidth to CPU cores (Figure 8.7) [14,15]. As this transition occurs, the increase in data volume and speed will result in some key Test challenges both at the wafer and package level. Challenges include: Higher power to service the increase in bandwidth; delivery of power and signal to the device with sufficient fidelity to achieve the higher speeds (while device size shrinks as noted above, which will challenge interface routing and development); and thermal management of the device at wafer and package level to appropriately dissipate and control device heating.

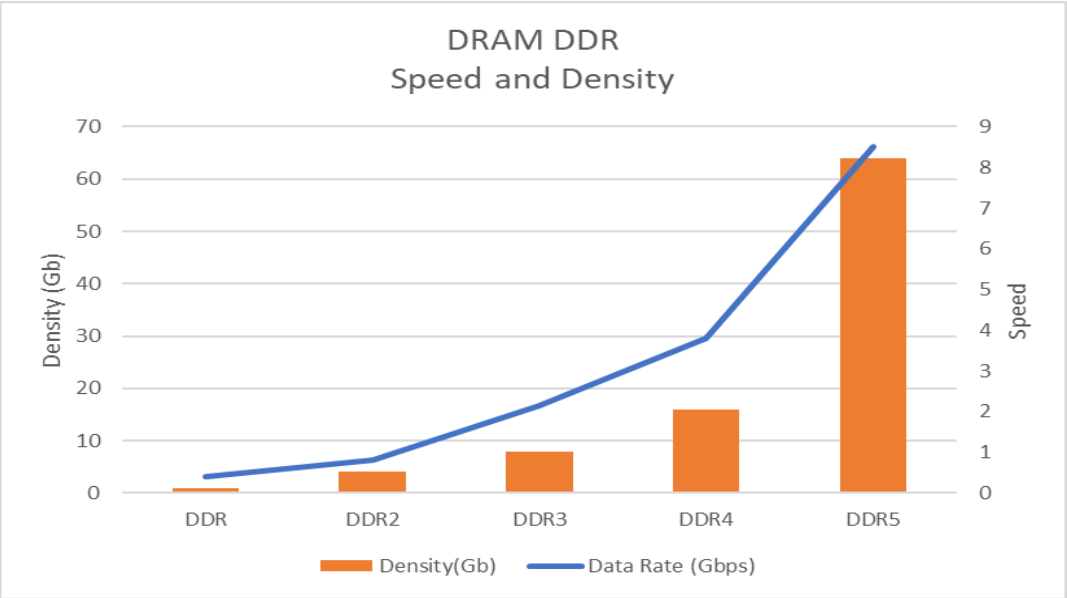


Figure 8.7: DRAM DDR Speed and Density

High Bandwidth Memory (HBM) is also growing in application for near-processor applications to improve graphics and AI applications; and GDDRx speeds are continuing to increase speed to accelerate graphics performance – GDDR6 at 24Gbps [16] is available today, with continued speed increases expected. All these advances improve the speed and ability of users/systems to effectively access data with decreased latency. These advances will further challenge speed, power, and thermal management in similar ways as observed in the DDR transitions noted above.

References:

1) Micron, "Memory Coalition of Excellence"
2) Yole Group "Status of The Memory Industry Report 2022"
3) Forward Insights "NAND Quarterly Insights Q3/22"
4) JEDEC JESD79-5B "DDR5 SDRAM"
5) JEDEC JES220F "UNIVERSAL FLASH STORAGE, Version 4.0"
6) Advantest Summary of Industry trends
7) Jim Elliott, Samsung Keynote, Flash Memory Summit 2022
8) Yole Group "NAND Market Monitor Q4 2022"
9) Teradyne analysis of broad industry trends
10) Scott Nelson, Kioxia Keynote, Flash Memory Summit 2022
11) Wikipedia, "Universal Flash Storage, Version Comparison"
12) Wikipedia, "PCI Express, History and revisions"
13) Compute Express Link "CXL 3.0 Press Release"
14) Synopsis, "DDR5/4/3/2: How Memory Density and Speed Increased with each Generation of DDR"
15) JEDEC, "JEDEC Publishes New DDR5 Standard for Advancing Next-Generation High Performance Computing Systems"
16) Samsung, "Samsung Electronics Launches Industry's First 24Gbps GDDR6 DRAM to Power Next-Generation High-End Graphics Cards"

Contributors:

Contributor	Company Affiliation
Anthony Lum	Advantest America Inc
Vineet Pancholi	Amkor
Paul Okino	Teradyne
Phil Byrd	Micron Technology
Jerry McBride	Micron Technology

8.7 Analog and Mixed Signal Test

8.7.1 Executive Summary

The economic benefit of monolithic integration (SoC) and system in package (SiP) is well established and continues. This integration has combined digital logic with processing, analog, power management, and mixed signal routinely in a single package and often on the same die. This trend has increased the breadth of interface types on a single part and given rise to test equipment that mirrors this range with a corresponding breadth of instruments. Now this trend has again escalated with the emergence of through silicon via (TSV) packaging technology driving the challenge in a 3rd dimension.

An important trend impacting mixed signal and analog testing is the compelling economics of multi-site testing for devices manufactured in extremely high volumes, also called parallel test. To support parallel test, many more instrument channels of each interface type are required to keep test cell throughput and

Parallel Test Efficiency (PTE), also known as Multi-Site Efficiency (MSE), high; this is of increasing importance to avoid severely impacting Units Per Hour (UPH).

A similar concept but in a dimension relating to the single device itself is testing multiple IP cores within the device in parallel (concurrent test). This has many of the requirements and challenges of parallel test, but also includes some unique ones. A key one is having the ability in the design of the IC to test IP cores independently, in parallel. Test Access Mechanisms (TAMs) are the ability of IP cores to be accessed and controlled independently from other IP cores. The most powerful economic advantage results when being able to test multiple IP cores in parallel, while at the same time testing multiple devices in parallel.

The increasing number of interfaces per device and the increasing number of devices tested simultaneously raise the need to process an increasing amount of data in real time. The data from the mixed signal and analog circuitry is typically non-deterministic and must be post processed to determine device quality. This processing must be done in real time or done in parallel with other testing operations to keep test cell throughput high. In fact, as site count increases, overall throughput can decrease if good PTE is not maintained.

Looking forward, the breadth, performance, density, and data processing capability of ATE instrumentation will need to improve significantly to provide the needed economics. The area undergoing the most change is RF/microwave and so it is covered in its own separate section. The digital and high-speed serial requirements for mixed signal devices are equivalent to logic and are covered in that section. The requirements for the TAM are covered in the DFT SOC Device Testing section. The requirements for DC trim accuracy are included in Table 8.13.

8.7.2 DC Accuracy updates for 2020

The 2020 update for DC accuracy includes ever-increasing low-end accuracy requirements driven by lower VDD values and more fuse blowing and servo techniques being used to cost effectively make the DUT more accurate and improve the specifications and yields.

COT is always important and more parallelism in terms of IP blocks within a device (IP block) and multi-site parallelism is key to this.

Quality also needs to be improved with these accuracy improvements. Pre and post inline checking and the comparison of lot runs looking for common tests that always pass or fail will be aided by using Artificial Intelligence (AI) and Machine Learning (ML) to handle and simplify large volumes of data. Other quality improvements include inventorying the tests that have been run and having more quantitative (actual value) versus qualitative (pass/fail) testing. There is always a cost trade-off balance.

8.7.3 Power updates for 2020

The other end of the spectrum for 2020 is high power (current and voltages) being driven primarily by server farm power needs and automotive and battery management systems as shown in Table 8.13 in the Note 8 section.

Because of the higher power, some tests that run a device at full power must be run very quickly and then turned off so as not to damage the parts that require special cooling. In these cases, precision pulses are required on tests like RDSon which pulses a high current at a very short pulse width to test the on-resistance of a switch.

Quality improvements here would include thermal testing and management throughout the test flow. For example, high power tests which would generate a lot of heat could be interleaved with low power test to allow the device to cool down.

Handlers with built-in cooling for the device is another option to be looked at for devices requiring the cooling.

Some process technologies once considered niche are gaining mainstream acceptance, including GaN (gallium nitride) and SiC (silicon carbide) devices.

SiC is projected to hit $1.5B by 2023 for these types of applications[14]:

- Electric Vehicle
- Train
- Charging Infrastructure
- Motor Drivers
- Photo Voltaic (PV)
- Wind Power

GaN is projected to hit $500M by 2023 for these types of applications[15]:

- Data Centers
- Fast Charger
- LiDar
- Wireless Charging
- Electric Vehicle

Power devices using GaN and SiC have higher band gaps compared to their silicon counterparts. The benefits are[16].

- Higher power density
- Smaller size (smaller wafer & die)
- Better high temperature performance because their band gap is higher than silicon
- Higher frequency response
- Lower ON-resistance
- Lower leakage, so there is a need for sourcing higher test voltages, as well as appropriate low current measurement sensitivity.

The test requirements to test GaN and SiC devices are

---

[14] https://www.systemplus.fr/wp-content/uploads/2018/07/YD18027_Power_SiC_2018_Materials_Devices_Applications_July2018_Yole_Sample-1.pdf

[15] https://compoundsemiconductor.net/article/106038/Would_Apple_Change_The_Power_GaN_World%7BfeatureExtra%7D

[16] https://www.powerelectronics.com/technologies/power-electronics-systems/article/21860727/testing-gan-and-sic-devices-faqs

- Breakdown voltages up to 3000 V or even higher
- More than 100 A
- Junction capacitances for dc biases up to 3000 V
- High SiC and GaN voltages and fast switching speeds
- Testing these devices at their specified voltage, current and power rating
- Test fixturing:
    - A proper test fixture solution is extremely important to ensure safety (due to the high voltages and currents used)
- Supporting the wide variety of power device package types.

The breakdown voltage test has special techniques being investigated involving Paschen's Law. To summarize: above a certain pressure, increasing the pressure raises the breakdown voltage or allows a narrower gap without breakdown at a set voltage.

8.7.4 Analog Mixed Signal Updates for 2020

Pulse Amplitude Modulation – 4 levels (PAM4) (Note: Optical PAM4 is not addressed in this update)

The attributes of PAM4 include:

- 4 amplitude levels
- 2 bits of information in every symbol: ~ 2x throughput for the same Baud rate, ie, 28 GBaud PAM4 = 56 Gb/s
- Lower SNR, more susceptible to noise
- More complex Tx/Rx design, higher cost

It is used extensively in the JESD 204B/C standard.

The transmitter (Tx) can be measured with high-speed digitizers, samplers, digital oscilloscopes or even a digital comparator. The receiver (Rx) signal is generated by RF DACs. RF design rules come into play at these high frequencies.

DSP is required to get an optimal eye opening which entails equalization for both PRE and POST processing. PRE processing is used to clean up the stimulus to the Rx, and POST processing is used to clean up the measured data from Tx. Amplitude accuracy is important because of the 4-level algorithm of PAM4. High-accuracy timing and low jitter are important to get a good eye opening.

Challenges in Analyzing PAM4 signals include:

- Sampling Point: Finite rise times and different transition amplitudes create inherent ISI and make clock recovery more difficult (TransImpedance Amplifiers have CDR integrated into them).
- Quantization error plays a role when you take PAM4 measurements versus NRZ. Transition times of the PAM4 data signal can create significant horizontal eye closure due to the higher transition density.
- Noise Tolerance: Instead of having the full amplitude range, there is only 33% of the amplitude because the voltage range is divided into four levels (refer to the Figure 8.8). Lower PAM4 insertion loss compensates for the 9.5-dB loss in SNR because the eye height for PAM4 is 1/3 of the eye height for NRZ, SNR loss = 20* log10 (1/3) = ~9.5 dB. When other non-linearity is included, it is approximately 11 dB.

- Non-Linear Eyes: The system-margin bottleneck lies with the worst eye. Nonlinearity starts right at the Tx output, and is composed of RLM loss + SNDR loss + other losses like SNDR (ISI).
- Clock Recovery is used on the Rx side to minimize low frequency jitter.
- Fixturing – getting the signal to the DUT
    - Integrated resources are difficult to design at these speeds but are sometimes easier to fixture. External Boxes are available but then are more effort and expense to route to the device. Line loss and jitter are a challenge.

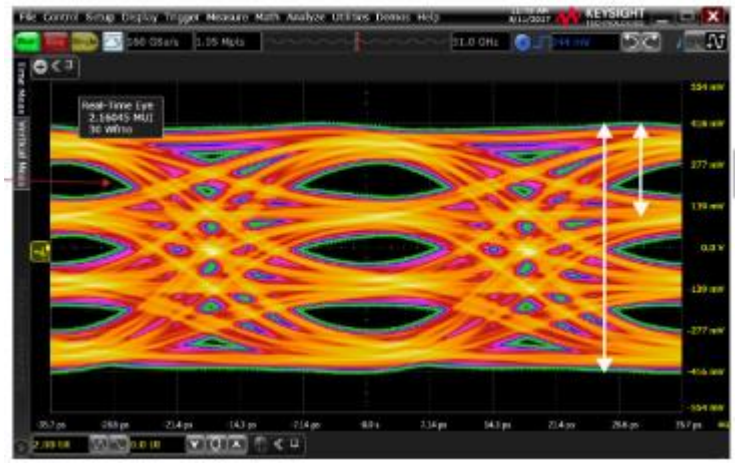


Figure 8.8: Scope Capture of PAM4 Signal

A typical test list for PAM4 looks like this:

Tx using a PRBS13 waveform:

- Output waveform
- Level Separation Mismatch Ratio
- Eye Symmetry
- Eye Height (amplitude) and Width (timing)
- Transition Time
- Signal-to-noise-and-distortion ratio (SNDR)
- Output Jitter
    - Jrms
    - Even-Odd Jitter (EOJ)
- Spacing of the PAM4 levels
- Eye Linearity: ratio of min to max PAM4 eye amplitudes as shown in Figure 8.9
    - Eye linearity = min(AVupp, AVmid, AVlow) / max(AVupp, AVmid, AVlow)

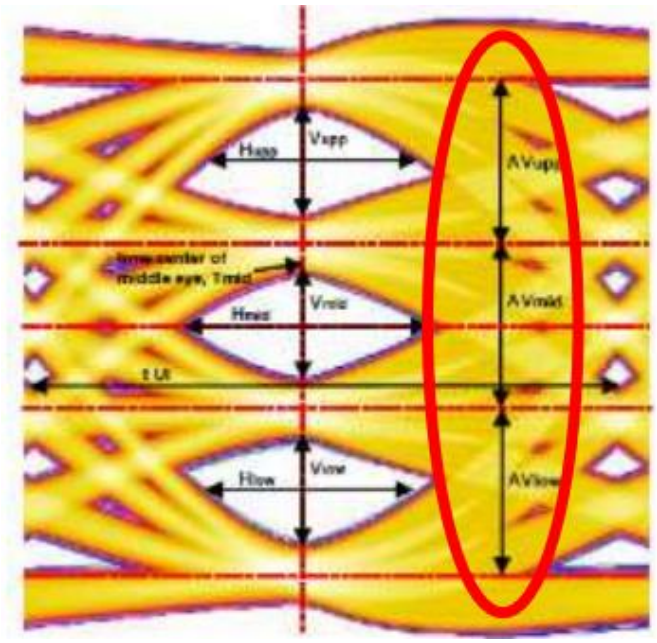


Figure 8.9: Eye Linearity

8.7.5 Rx tests

These tests involve how much distortion and jitter can be placed on the incoming signal to the receiver and still "read" the correct data stream.

- Jitter tolerance defined as how much jitter the receiver can tolerate
- Other potential receiver "stress tests"
  - Eye Skew (Timing)
  - Eye non-linearity (Amplitude between levels)

8.7.6 Key Test Trends

Short-Term Trends (< 5 Years)

There are three important trends. The first is to deliver adequate quality of test. Most analog/mixed-signal testing is done through performance-based testing. This includes functional testing of the device and then analyzing the quality of the output(s). This requires instrumentation capable of accurately generating and analyzing signals in the bandwidths and resolutions of the device's end-market application. Both of these parameters are trending upwards as more information is communicated between devices and/or devices and the physical environment. See the Mixed Signal Test tables (Table 8.13) for updates and future needs.

The second key trend is the need for higher DC accuracy. Many of the converters and precision references are made more accurate by doing a measure and trim step. The trim can be accomplished through several means; one of the more recent and cost-effective ways is through register programming of the device. The trim takes a relatively lower performance device and adds high accuracy to it through a DC test and register programming. In the past, this was done for medium performance devices, but now

the test methodology has matured, and it is being applied to high accuracy/resolution devices. The change is that in this class of devices, much higher DC accuracy is required to make a valid test.

The third key trend is to enable the economics of test through instrumentation density and Parallel Test Efficiency (PTE). The level of parallelism requires an increase in instrumentation density.

These trends of increasing ATE instrument channel count, complexity, and performance are expected to continue, but at the same time the cost of test must be driven lower (see the areas of concern listed below).

Analog/mixed-signal DFT and BIST techniques continue to lag. No proven alternative to performance-based analog testing has been widely adopted and more research in this area is needed. Analog BIST has been suggested as a possible solution and an area for more research. Fundamental research is needed to identify techniques that enable reduction of test instrument complexity, partial BIST, or elimination of the need for external instrumentation altogether.

The Ethernet trends are continuing into higher speeds – 28, 40 Gbps per channel and even beyond.[1] There continues to be the need for backwards compatibility to the many existing digital communication standards.

Table 8.13: Mixed-signal and DC Test Requirements

	2020	**2021**	**2026**	**2031**
Low Frequency Waveform [Note 1]				
SFDR	145	145	145	145
SNR	120	120	120	120
THD	140	140	140	140
BW-Minimum (kHz)	50	50	50	50
BW-Maximum (kHz) [Note 2]	500	500	500	500
High Frequency Waveform Source / Measure [Note 3]				
Level V (pk–pk)	<4	<4	<2.5	<2.5
BW (MHz)	250	250	500	500
Sample rate (MS/s) [Note 5]	500	500	1000	1000
Resolution (bits) AWG/Sine	16	16	18	18
Noise floor (dB/RT Hz)	-140	-140	-150	-150
Very High Frequency Waveform Source / Measure [Note 4]				
Level V (pk–pk)	<4	<4	<4	<4
Accuracy (±)	0.50%	0.50%	0.50%	0.50%
Measure BW (GHz) (under sampled)	9.6	9.6	15	15
Capture Depth Mwords	4	4	4	4
Min resolution (bits)	8-10	8-10	8-10	8-10
DC Accuracy (Note 6)				
DC force (uV)	50	50	50	50
DC measure (uV)	50	50	50	50

DC force (nA) (Note 7)	5	5	1	1
DC measure (nA) (Note 7)	5	5	1	1
DC Power (Note 8)				
DC force V Constant	120	120	140	140
DC measure V Constant	120	120	140	140
DC force A Constant	80	80	100	100
DC measure A Constant	80	80	100	100
DC force V Pulse	80	80	100	100
DC measure V Pulse	80	80	100	100
DC force A Pulse	30	30	50	50
DC measure A Pulse	30	30	50	50
Ethernet				
Speeds (Gbps)	40	40	100	400

Manufacturable solutions exist, and are being optimized	
Manufacturable solutions are known	
Interim solutions are known	
Manufacturable solutions are NOT known	

NOTES:

1) Audio / Precision; Source & Measure specifications (22 KHz BW)

2) Major testing condition

3) Target Devices are Wireless Baseband, xDSL, ODD, Digital TV (Track Mobile Baseband)

4) Target Devices are HDD, Radar, WiGig

5) For Measure Sample Rate: Dependent on method, tracking or Front End filter.

6) The purpose of DC accuracy for this table is for high resolution force/measure and trim

7) Devices may also need high current with the less accuracy

8) Markets include Automotive, Battery Management and Power. This does not include high voltage breakdown test.

Difficult Challenges in the Short Term

- As reflected in the tables, manufacturing solutions exist for the immediate future testing needs. However, high DC accuracy for sourcing, measuring and for trim/fuse blowing/register-setting in a manufacturing environment could be at issue depending on how high a resolution/accuracy the DUT is. Also 40 Gbps Ethernet has known manufacturing solutions, but none are optimized.
- Time-to-market and time-to-revenue issues are driving test to be fully ready at first silicon. The analog/mixed-signal test environment can seriously complicate the test fixtures and test methodologies. Noise, crosstalk on signal traces, added circuitry, load board design complexity, and debug currently dominate the test development process and schedule. The test development process must become shorter and more automated to keep up with design. In addition, the ability to re-use analog/mixed-signal test IP is needed.
- Increased use of multi-site parallel and concurrent test of all analog/mixed-signal chips is needed to reduce test time, in order to increase manufacturing cell throughput, and to reduce

test cost. All ATE instrument types, including DC, require multiple channels capable of concurrent/parallel operation and, where appropriate, fast parallel execution of DSP algorithms (FFTs, etc.) to process results. In addition, the cost per channel must continue to drop on these instruments as the density continues to increase in support of parallel test drivers.

- Improvements in analog/mixed-signal DFT and BIST are needed to support the items above.

Medium-term Trends (6 to 10 years out)

- For Wireless Baseband, xDSL, ODD, and Digital TV (Track Mobile Baseband) devices, the source and measure bandwidths, sampling rates and resolutions increase, while the noise floors are decreasing.
- Additionally, DC force and measure accuracies get more challenging.
- Ethernet speeds trending to 100 Gbps [2] have only interim solutions identified.
- Higher speeds and modulation will necessitate PAM to handle the increased data bandwidth – for example, PAM4, 8 or 16 at speeds of 32 GBPS. [3], [4]

Difficult Challenges in the Medium Term

- As the capability requirements increase, there are solutions available, but they do not lend themselves easily to high volume manufacturing.
- Basic physical and electrical properties come more into play. For example, a -150 dB noise floor is possible, but special fixturing is required that is difficult to deploy into a manufacturing environment.
- Ethernet speeds of 100 Gbps [2] have only interim solutions identified.

Long-term Trends (10 years+ out)

- Ethernet speeds trending to 400 Gbps [5], [6]

Difficult Challenges in the Long Term

- Ethernet speeds of 400 Gbps do not have known manufacturing solutions identified.

8.7.7 SUMMARY

Cost continues to be the most critical pressure and concern for analog mixed signal because much of the volume for this is consumer oriented. However, in the medium and long term, performance starts becoming an issue for high-volume manufacturing in terms of bandwidth, sample rate, resolution and noise floor to keep up with the newer devices on the horizon. Ethernet in the medium and long term has manufacturing challenges both in optimization and known solutions.

8.7.8 References

1. https://www.cisco.com/c/dam/en/us/products/collateral/switches/catalyst-6500-series-switches/white-paper-c11-737238.pdf

2. https://www.networkcomputing.com/data-centers/25-50-and-100-gigabit-ethernet-data-center/1422885308

3. http://www.ieee802.org/3/100GNGOPTX/public/mar12/plenary/tremblay_01_0312_NG100GOPTX.pdf

4. http://www.ieee802.org/3/bm/public/sep12/ghiasi_01a_0912_optx.pdf

5. http://searchdatacenter.techtarget.com/feature/Understanding-tomorrows-Ethernet-speeds

6. https://www.keysight.com/main/application.jspx?ckey=2716254&id=2716254&nid=-32176.0.00&lc=eng&cc=US

7. https://en.wikipedia.org/wiki/Effective_number_of_bits

8.8 Wafer Probe and Device Handling

Wafer probe and component test handling equipment face significant technical challenges in each market segment. Common issues on both platforms include higher parallelism and increasing capital equipment and interface cost.

8.8.1 Device Handling Trends

Increased parallelism at wafer probe drives a greater span of probes across the wafer surface and significantly increased probe card complexity. Prober and probe card architecture should evolve to simplify the interface, however just the opposite is happening: ATE tester complexity is decreasing and more technology and complexity is built into the probe card interface. A better thermal solution is a very important parameter along with performance for better yield management. Memory applications are increasing the total power across a 300mm wafer, and wafer probe needs to dissipate this total power to sustain the set-temperature during test. Power density per DUT is increasing and it's very challenging to manage a stable wafer-level test temperature. 3D integration technology requires very precise probing technology in X, Y and Z, as micro-bumps may be easily damaged during the probing process. MEMS applications require a variety of testing environments such as pressure, magnetic, and vacuum environments; also, wafer shape and package style are becoming very unique depending on the application type.

Reducing the cost of wafer-level and package-level test in the face of more challenging technology and performance requirements is a constant goal. The demand for higher throughput must be met by either increased parallelism (even with reduced test times), faster handler speed, or process improvements such as asynchronous test or continuous-lot processing. 3D integration technology requires new contact technology for the intermediate test insertion which will be added between conventional front-end process and back-end process. New contact technology to probe on the singulated and possibly thinned die's micro-bumps or C4 bumps after the die is mounted on an interposer is needed. For the die-level handler, the main tasks are the alignment accuracy to enable fine pitch contact, die level handling without damaging the die, and the tray design that supplies/receives the die.

Packages continue to shrink, substrates are getting thinner, and the package areas available for handling are getting smaller at the same time that the lead/ball/pad count is increasing. In the future, die-level handlers as well as package handlers will need the capability to very accurately pick and place small, fragile parts, yet apply similar or increasing insertion force without inducing damage.

Temperature ranges are expanding to meet more stringent end-use conditions, and there is a need for better control of the junction temperature, immediate heat control technology, and temperature control to enable stable DUT temperature at the start of test. Power dissipation overall appears to be increasing, but multi-core technology is offering relief in some areas.

It is unlikely that there will be one handler that is all things to all users. Integration of all of the technology to meet wide temperature range, high temperature accuracy, high throughput, placement accuracy, parallelism, and special handling needs while still being cost effective in a competitive environment is a significant challenge.

Gravity feed, turret, and strip handlers have been added to the table while retaining the pick and place type handler. The gravity feed handler is used on SOP, QFN, and DIP packages. Turret handlers are

widely used on discrete-type QFN devices. Strip handlers are used on the frame before singulation. Strip test enables high parallelism with fewer interface resources, which enables cheaper test cost. These additional three types of handlers are widely used on relatively low-end or low-cost devices. Evolution of these handlers is quite different but important for various type of LSI.

Table 8.14: Test Handler and Prober Difficult Challenges

Pick and Place Handlers (High Performance)	Temperature control and temperature rise control due to high power densities
	Continuous lot processing (lot cascading), auto-retest, asynchronous device socketing with low-conversion times
	Better ESD controls as products are more sensitive to ESD. On-die protection circuitry increases cost.
	Lower stress socketing, low-cost change kits, higher I/O count for new package technologies
	Package heat lids change thermal characteristics of device and hander
	Multi-site handling capability for short test time devices (1–7 seconds)
	Force balancing control for System in Package and Multi-Chip Module
Pick and Place Handlers (Consumer SoC/ Automotive)	Support for stacked die packaging and thin die packaging
	Wide range tri-temperature soak requirements (-55ºC to 175ºC) increases system complexity for automotive devices
	Device junction temperature control and temperature accuracy +/-1.0°C
	Fine Pitch top and bottom side one shot contact for Package on Package
	Continuous lot processing (lot cascading), auto-retest, low conversion times, asynchronous operation
Pick and Place Handlers (Memory)	Thin die capable kit-less handlers for a wide variety of package sizes, thicknesses, and ball pitches < 0.3mm
	Package ball-to-package edge gap decreases from 0.6 mm to 0 mm require new handling and socketing methods
	Parallelism at greater than x128 drives thermal control +/-1.0°C accuracy and alignment challenges <0.30mm pin pitch
Prober	Consistent and low thermal resistance across the chuck is required to improve temperature control of the device under test. There is a new requirement of active/dynamic thermal control, which can control junction temperature(ΔT) during test
	Both Logic and Memory wafer generates more wattage/heat, demand of Heat dissipation performance improvement is expected. Especially Heat Dissipation at Hot temperature is challenging technology for wafer prober.
	There are wafer handling requirements of non-SEMI standard such as 3DI, MEMS, WLCSP and PsP applications. Those are thin, thick, unique shape so customized wafer handling technique/technology is needed. Wafer cassette is needed to be customized to meet the request as well.
	Probing on micro-bump is technically proven but there are many challenges "parallelism/multi-site", "Thermal conduction" and "bump damages/reliability"
	Advances in probe card technology require a new optical alignment methodology.
	Dicing frame probers can cover a wide temperature range, but a dicing sheet cannot cover the full range.
	Greater parallelism/multi-site, and higher pin counts require higher chuck rigidity and a robust Probe Card changer.
	Power Device application requires very thin wafer which drive need for 'Taiko Wafer' and 'Ring attached wafer' handling and more high voltage chuck technologies.
	Enhanced Probe Z control is needed to prevent damage to pads, there are solution in the market but those must be optimized to integrate onto wafer prober to meet needs of test cost requirement.
Gravity Feed Handlers	Thinner packages and wafer will require a reduction in the impact load to prevent device damage
	Test head size increase due to higher test parallelism may alter handler roadmap
	Reduction of static electricity friction and surface tension moisture friction on very small packages (<1 x 1 mm)
Turret Handlers	Test contactor support for > 100A current forcing on power devices

	Kelvin contact support (2 probes) to very small area (0.2 x 0.2mm) contacts on small signal devices
Strip L/F Handlers	Testing process infrastructure configuration
	Accuracy of the contact position for high temperature testing environment

Table 8.15: (part 1): Wafer Probe Technology Requirements

Year of Production	***2019***		***2020***		***2021***	
MPU, ASIC, SOC and Mixed Signal Products						
Wirebond - inline pad pitch	**40**		**35**		**35**	
Wirebond - stagger pad pitch	**45**		**30**		**30**	
Bump - array bump pitch	**30**		**30**		**30**	
Sacrifical pad pitch in a field of bumps	**100**		**100**		**100**	
I/O Pad Size (µm)	**X**	**Y**	**X**	**Y**	**X**	**Y**
Pad Materials						
Wirebond	**30**	**30**	**30**	**30**	**30**	**30**
Bump	**30**		**30**		**30**	
Sacrifical pad in a field of bumps	**45**	**45**	**42**	**42**	**42**	**42**
Wafer Test Frequency (Hz)	**2.4G**		**2.4G**		**2-10 GHz**	
Wafer Test Frequency (Hz) for HSIO	**25Gbps/12.5GHz**		**56Gbps PAM4 28Gbps NRZ @ 14GHz**		**100Gbps PAM4 @ 28GHz**	
Probe Tip Diameter Wirebond	**7.5**		**6.5**		**6.5**	
Probe Tip Diameter Bump	**25**		**25**		**25**	
Probe Force Bump(gf) - at recommended overdrive	**1.5**		**1.5**		**1.2**	
Size of Probed Area (mm^2)	**20000**		**20000**		**20000**	
Number of Probe Points / Touchdown	**180000**		**200000**		**200000**	
Maximum current per probe >130um pitch	**2A**		**2A**		**2A**	
Maximum current per probe <130um pitch	**1A**		**1A**		**1A**	
Maximum contact resistance	**<0.5**		**<0.5**		**<0.5**	
Probe test temperature range	**-55**	**200**	**-55**	**200**	**-55**	**200**
Automotive Radar						
Wafer Test Frequency (GHz)	**80GHz**		**80GHz**		**80GHz**	
RF I/O Geometry	**Solder Ball**		**Solder Ball**		**Solder Ball**	
I/O Size (um)	**100um Cu Pillar SB**		**100um Cu Pillar SB**		**100um Cu Pillar SB**	
I/O Pitch (um)	**300um**		**300um**		**300um**	
RF Ports per Site	**14**		**14**		**13**	
Sites being probed together	**2**		**4**		**4**	
Total Number of RF Ports	**28**		**56**		**52**	
High Speed Digitial (TIAm CDR, VCSEL, etc.)						
Wafer Test Frequency (GHz)	**67GHz**		**67GHz**		**67GHz**	
RF I/O Geometry	**X**	**Y**	**X**	**Y**	**X**	**Y**
I/O Size (um)	**50**	**50**	**50**	**50**	**50**	**50**
I/O Pitch (um)	**80um**		**80um**		**80um**	
RF Ports per Site	**24**		**24**			
Sites being probed together	**2**		**8**		**8**	
Total Number of RF Ports	**48**		**96**		**96**	
802.11ad						
Wafer Test Frequency (GHz)	**64GHz**		**64GHz**		**64GHz**	
RF I/O Geometry	**Solder Ball**		**Solder Ball**		**Solder Ball**	

I/O Size (um)	**80um**	**70um**	**70um**
I/O Pitch (um)	**150um**	**125um**	**125um**
RF Ports per Site	**32**	**32**	**32**
Sites being probed together	**8**	**8**	**8**
Total Number of RF Ports	**256**	**256**	**256**
5G			
Wafer Test Frequency (GHz)	**45GHz**	**73GHz**	**50-60GHz**
RF I/O Geometry	**Solder Ball**	**Solder Ball**	**Cu Pillar w/o cap**
I/O Size (um)	**100um**	**70um**	
I/O Pitch (um)	**150um**	**130um**	**130um**
RF Ports per Site	**34+**	**38+**	
Sites being probed together	**8**	**8**	
Total Number of RF Ports	**64**	**>100**	

Table 8.16: (part 2): Wafer Probe Technology Requirements. NOTE VCSEL and PIC have different requirements.

Year of Production	***2019***		***2020***		***2021***	
Optical Probe - *NOTE VCSEL and PIC have different requirements*						
Minimum pitch between fibers (um)	**127**		**120**		**120**	
Fiber optical alignment accuracy (Multi-Mode)	**< 5um**		**< 10um**		**<10um**	
Fiber optical alignment accuracy (Single-Mode)	**< 0.1um**		**< 0.1um**			
DRAM						
Wirebond - inline pad pitch	**50**		**50**		**50**	
I/O Pad Size (µm)	**X**	**Y**	**X**	**Y**	**X**	**Y**
Wirebond	**40**	**50**	**35**	**40**	**35**	**40**
Sacrificial Pads	**45**	**50**	**40**	**40**	**40**	**40**
Wafer Test Frequency for Sort(Hz)						
Test Frequency(Hz)	**250M**		**400M**		**400M**	
Shared Signal Line Test Frequency (Hz)	**125M**		**200M**		**250M**	
Minimum pulse width	**2.0nS**		**2.0nS**		**2.0nS**	
At Speed Wafer Test						
Test Frequency(Hz)	**3.2G**		**3.2G**		**3.2G**	
Probe Tip Diameter	**8.5**		**8.5**		**8.5**	
Probe Force(gf) - at recommended overdrive	**2.5**		**2.5**		**2.5**	
Size of Probed Area (mm^2)	**100% of wafer**		**100% of wafer**		**100% of wafer**	
Number of Probe Points / Touchdown - Memory	**130000**		**150000**		**150000**	
Maximum Current (mA)/pin	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**
	250	**<.001**	**250**	**<.001**	**250**	**<.001**
Maximum Resistance (Ohm)	**Contact**	**Series**	**Contac t**	**Series**	**Contac t**	**Series**
	<0.5	**<3**	**<0.5**	**<3**	**<0.5**	**<3**
Probe test temperature range	**-45**	**150**	**-45**	**175**	**-45**	**175**

NAND						
Wirebond - inline pad pitch	**80**		**80**		**65**	
I/O Pad Size (µm)	**X**	**Y**	**X**	**Y**		
Wirebond	**50**	**60**	**50**	**60**	**50**	**60**
Wafer Test Frequency for Sort (Hz)						
Wafer Test Frequency(Hz)	**100M**		**133M**		**133M**	
At Speed Wafer Test						
Test Frequency(Hz)	**600M**		**600M**		**2.4G**	
Probe Tip Diameter	**10**		**10**		**10**	
Probe Force(gf) - at recommended overdrive	**3**		**3**		**3**	
Size of Probed Area (mm^2)	**100% of wafer**		**100% of wafer**		**100% of wafer**	
Number of Probe Points / Touchdown - Memory	**80000**		**80000**		**80000**	
Maximum Current (mA)/pin	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**
	250	**<.001**	**250**	**<.001**	**250**	**<.001**
Maximum Resistance (Ohm)	**Contact**	**Series**	**Contact**	**Series**	**Contact**	**Series**
	<0.5	**<3**	**<0.5**	**<3**	**<0.5**	**<3**

LCD driver Products						
Bump - inline pad pitch	**18**		**16**		**16**	
Bump - stagger pad pitch	**10**		**8**		**8**	
I/O Pad Size (µm)	**X**	**Y**	**X**	**Y**	**X**	**Y**
Inline	**11**	**50**	**11**	**50**	**11**	**50**
Stagger	**15**	**30**	**12**	**40**	**12**	**40**
High speed I/O pin freq (Mobile/TV)	**4.5Gbps / 6.5Gbps**		**4.5Gbps / 6.5Gbps**		**4.5Gbps / 6.5Gbps**	
Probe needle structure	**Cantilever / Vertical**		**Cantilever / Vertical**		**Cantilever / Vertical**	
Probe Tip Diameter (um)	**8**		**8**		**8**	
Probe Force(gf)	**2**		**2**		**2**	
Size of Probed Area (mm^2)	**5600**		**6800**		**6800**	
Number of Probe Points / Touchdown	**12000**		**12000**		**12000**	
Maximum Current (mA)/pin	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**
	300	**<.001**	**300**	**<.001**	**300**	**<.001**
Maximum Resistance (Ohm)	**Contact**	**Series**	**Contact**	**Series**	**Contact**	**Series**
	<0.5	**<3**	**<0.5**	**<3**	**<0.5**	**<3**

Table 8.17: (part 3): Wafer Probe Technology Requirements

Year of Production	***2019***		***2020***		***2021***	
CMOS Image Sensor						
Wirebond - inline pad pitch	**90**		**80**		**70**	
I/O Pad Size (µm)	**X**	**Y**	**X**	**Y**	**X**	**Y**

Wirebond	**60**	**70**	**60**	**65**	**60**	**60**
WLCSP	**46**	**100**				
WLCSP (TSV construction)	**55**	**55**	**40**	**40**	**40**	**40**
	200M		**200M**			
High speed I/O pin freq (Hz)	**2.5G**		**3G**			
Probe needle structure	**Vertical / MEMS**		**Vertical / MEMS**		**Vertical / MEMS**	
Probe Tip Diameter Wirebond (um)	**12**		**10**		**7**	
Probe Force Wirebond(gf)	**2**		**2**		**2**	
Size of Probed Area (mm^2) [3]- Visible light	**300x300**		**300x300**		**300x300**	
Number of Probe Points / Touchdown - IR [4]	**5000**		**10000**		**10000**	
Maximum Current (mA)/pin	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**
Visible light sensor	**250**	**<.001**	**250**	**<.001**	**250**	**<.001**
IR sensor [5]	**1000**	**<.001**	**1200**	**<.001**	**1200**	**<.001**

Visible Light Sensor / Optical Fiberoptic Transmission						
Maximum Current (mA)/pin	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**	**Probe Tip**	**DC Leakage**
Visible light sensor	**250**	**<.001**	**250**	**<.001**	**250**	**<.001**

Parametric (Process monitor)						
Inline pad pitch	**40**		**40**		**40**	
Inter-row pad pitch	**35**		**35**		**35**	
Pad Size (µm)	**X**	**Y**	**X**	**Y**	**X**	**Y**
In line pads	**20**	**20**	**20**	**20**	**20**	**20**
Probe Tip Diameter	**6**		**6**		**6**	
Number of pad rows	**2**		**2**		**2**	
Probe Force(gf) - at recommended overdrive	**2**		**2**		**2**	
Number of Structures /Touchdown	**8**		**8**		**8**	
Maximum Capaciance (pF pin to pin)			**1**		**1**	
Maximum Leakage (pA)/pin (10V / 1 Sec test)	**0.2**		**0.2**		**0.2**	
Maximum Contact resistance (Ohms)/pin	**0.3**		**0.3**		**0.3**	
Maximum Path resistance (Ohms)/pin	**3**		**3**		**3**	
Maximum Probe temperature Range (degrees C)	**-50**	**200**	**-50**	**200**	**-55**	**200**
Maximum test Frequency (GHz)	**3**		**6**		**6**	

Table 8.18: Wafer Prober Requirements

Year of Production	***2019***	***2020***	***2021***
Wafer Handling			
Wafer Size [inch]			
200mm Prober	6, 8	6, 8	6, 8
300mm Prober	8, 12	8, 12	8, 12
Min Bump Size[um]	15	15	15
Min Wafer Thickness[um]	200	100	100
Max Wafer Thickness[um]	3000	3000	3000
Max Wafer Weight[g]	350	350	350
Min Wafer Exchange Time (sec)	30	30	30
Tester Docking			
Test Head Weight[Kg]	1500	1500	1500
Probe Card			
Probe Card diameter[mm]	580	725	725
Probe Card PCB Thickness[mm]	10	18	18
Probecard Total Height [mm]			
Prober			
XY Accuracy (Probe to Pad) [±um]			
200mm Prober	2.0	2.0	2.0
300mm Prober	2.0	1.0	0.8
Z Accuracy (Probe to Pad) [±um]			
200mm Prober	5.0	3.0	2.0
300mm Prober	5.0	2.0	2.0
Chuck Planarity [±um]			
200mm Prober	7.5	7.5	7.5
300mm Prober	7.5	5.0	5.0
Chuck Maximum Force [Kg]			
200mm Prober	60	60	60
300mm Prober	450	450	500
Set temperature range [ºC]			
200mm Prober	-55 to +300	-55 to +300	-55 to +300
300mm Prober	-55 to +250	-55 to +250	-55 to +250
Chuck Temp. Accuracy [±°C]			
200mm Prober	1.0	1.0	1.0
300mm Prober	1.0	1.0	1.0
Chuck Leakage [pA]			
200mm Prober	0.1	0.1	0.1
300mm Prober	0.1	0.1	0.1
Total Power Logic (W/Die)			
300mm Prober	200	200	200
Total Power Memory (Watts Per Die)			
300mm Prober	0.75	0.80	0.80

Max Voltage [V]			
200mm Prober	10000	10000	15000
300mm Prober	10000	15000	15000
Max Electrical current [A]			
200mm Prober	300	300	300
300mm Prober	300	300	300

Table 8.19: (part1) Test Handler Requirements

Year of Production	***2019***	***2020***	***2021***
Pick and Place Handlers (High Performance)			
Temperature set point range (°C)	-20 to 125	-20 to 125	-20 to 125
Temperature accuracy at DUT (°C)	±1.0	±0.5	±0.5
Number of pins/device	2500	4000	5000
Throughput (devices per hour)	2-10K	2-10K	2-10K
Sorting Categories	3-6	3-6	3-8
Maximum Power Dissipation (W/DUT)	400	500	700
Maximum socket load per unit (kg)	80	120	200
Maximum Package Size(mm)	50x50	75x75	90x90
Minimum Package Thickness (mm)			
Pick and Place Handlers (Consumer SoC/Automotive)			
Temperature set point range (°C)	-55 to 190	-60 to 200	-75 to 200
Temperature accuracy at DUT (°C)	±1.0	±1.0	±1.0
Number of pins/device	1000	1200	1200
Throughput (devices per hour)	2-30k	5-30k	5-30k
Sorting Categories	3-6	3-8	3-8
Maximum Power Dissipation (W/DUT)	40	40	40
Maximum socket load per unit (kg)	80	80	80
Minimum Package Size(mm)	2x2	2x2	2x2
Minimum Package Thickness (mm)	0.2-1.8	0.2-1.8	0.2-1.8
Pin/land pitch (mm)	0.3	0.3	0.3
Pick and Place Handlers (Memory)			
Temperature set point range (°C)	-55 to 155	-55 to 155	-55 to 155
Temperature accuracy at DUT (°C)	±1.0	±1.0	±1.0
Number of pins/device	50-1000	50-1000	50-1000
Throughput (devices per hour)	20-75K	20-75K	20-75K
Index time (sec)	2-3	2-3	2-3
Sorting Categories	5-9	5-9	5-9
Minimum Package Size(mm)	4x6	3x5	3x5
Minimum Package Thickness (mm)	0.2-1.8	0.2-1.8	0.2-1.8
Pin/land pitch (mm)	0.2	0.2	0.2
Ball edge to package edge clearance (mm)	>0.1	>0.1	>0.1
Gravity Feed Handlers			
Temperature set point range (°C)	-55 to 175	-55 to 200	-55 to 200
Temperature accuracy at DUT (°C)	±2.0	±1.0	±1.0
Parallel testing:	8 (2x4)	16 (2x8)	16 (2x8)
Throughput (devices per hour)	50k	50k	50k
Index time (sec)	0.6-0.8	0.6-0.8	0.6-0.8
Sorting Categories	3-10	3-10	3-10
Minimum Package Size(mm)			
Minimum Package Thickness (mm)			

Conformity tube type (mm)	280-580	280-580	280-580
Turret Handlers			
Serial testing	2-4	2-4	2-4
Index time (sec)	0.072	0.072	0.072
Throughput (devices per hour)	50k	50k	50k
Minimum Package Size(mm)			
Minimum Package Thickness (mm)			
Sorting Categories	5-9	5-9	5-9
Impact load to PKG (N)	3	3	3

Table 8.20: (part 2): Test Handler Requirements

Year of Production	***2019***	***2020***	***2021***
Strip L/F Handlers			
Temperature set point range (°C)	-55 to 155	-55 to 155	-55 to 155
Temperature accuracy at DUT (°C)	±1.0	±1.0	±1.0
Number of pins/device	6-250	6-250	6-250
Parallel testing:	1-256	1-256	1-256
Throughput (devices per hour) 1-16 parallel	20-120K	20-120K	20-120K
Index time (sec)	0.15	0.15	0.15
Sorting Categories	32	32	32
Min. Pkg. Size(mm)	0.8x0.8	0.8x0.8	0.8x0.8
Max. Strip Size(mm)	300x100	300x100	300x100

8.8.2 Test Sockets

The test socket is an electrical and mechanical interface responsible for good electrical connection and transference of high-integrity signals between the DUT and the PCB/tester through a mechanical contact mechanism in order to determine the electrical characteristics of the DUT. As semiconductor design and manufacturing capabilities have progressed in recent years, the testing process keeps raising the electrical and mechanical requirements of test sockets. Therefore, the socket technologies have been rapidly driven by significantly enhanced electrical and mechanical requirements, both of which are instigated by higher power/voltage/current, reduced package size, tighter pitches, higher pin counts, smaller solder resist opening, and so on. It has been indicated that electrical properties are determined by not only the electrical but also by the mechanical requirements. The multi-physics problems have made socket designs progressively challenging for these higher requirements. Current models show difficulty in making sockets for high ball count devices and achieving I/O bandwidths of > 20GHz.

Socket Trends

Table 3 contains the test socket technology requirements. The requirements have been divided into contacting NAND, DRAM, and SoC devices that are contained in TSOP, BGA, and BGA SoC packages respectively. The TSOP package is assumed to be contacted using a blade; the DRAM BGA is contacted with a spring probe, and the SoC BGA is contacted with a 50-Ohm spring probe. The test socket performance capability is driven by the pitch between balls or leads, so the lead spacing of the assembly and packaging roadmap was used to determine the pitch.

Contact blades are generally used for testing TSOP NAND Flash and contain a spring function in their structure, which is loaded by compressing the DUT into the socket. The structure is very simple and suitable for HVM; however, the contactor blade must be long to maintain the specified contact force and stroke, and to achieve a long mechanical lifetime. A weak point is that the blade contactor is not suitable for fine pitch devices due to the need to have isolation walls between adjacent pins. The thickness of the isolation wall must be thinner for finer pitches, which makes fabrication of the isolation wall more difficult. At the same time, the contactor blade thickness needs to be thinner for finer pitch, which complicates achieving the specified contact force, stroke requirement, and mechanical lifetime.

Spring probes, mainly used for testing BGA-DRAM devices, are formed by use of small-diameter cylindrical parts (probe and socket) and coil springs. Compression of the spring probe creates the contact load. In order to guarantee sufficient mechanical life, the probe diameter should be large enough to

guarantee strength and durability and the length should be long enough to maintain sufficient travel under compression. The spring probe structure is relatively simple and easy to maintain and it is also easy to design a DUT loadboard.

According to the BGA-DRAM roadmap, the spring probe diameter will need to be smaller over time, driven by the finer pitch of the package ball roadmap. In addition, the spring probe will need to be shorter to meet the lower inductance values required to support the high frequencies of the roadmap I/O data rate.

Spring 50-Ohm probes required for BGA-SoC high frequency devices have coaxial structures that can reduce probe length transmission issues through impedance matching. However, advances in the package ball pitch through the roadmap will create restrictions to the coaxial pin arrangement structure (0.5 mm pitch in year 2016). The data rate will increase to 20GT/s in 2016, but the spring 50-Ohm probe will not have good electrical performance due to its multiple parts structure having higher contact resistance than other contactors. To support 50milli-Ohms of contact resistance starting in 2016, advances will be required in materials, plating, and structure.

Table 8.21: Test Socket Technology Requirements

Year of Production	***2019***	***2020***	***2021***
TSOP – Flash (NAND) – Contact blade			
Commodity NAND Memory			
Lead Pitch (mm)	0.3	0.3	0.3
Data rate (MT/s)	133	133	266
Contact blade			
Inductance (nH)	5-10	5-10	5-10
Contact Stroke (mm)	0.2-0.3	0.2-0.3	0.2-0.3
Contact force (N)	0.2-0.3	0.2-0.3	0.2-0.3
Contact resistance (m ohm)	30	30	30
Slit width (mm)	0.17	0.17	0.17

BGA – DRAM – Spring Probe			
Commodity DRAM (Mass production)			
Lead Pitch (mm)	0.25	0.25	0.2
DRAM RM GT/S	5.3	5.4	6.4
Spring Probe			
Inductance (nH)	0.2	0.2	0.15
Contact Stroke (mm)	0.2	0.2	0.2
Contact force (N)	<0.2	<0.2	<0.2
Contact resistance (m ohm)	100	100	100

BGA – SoC – Spring Probe (50 ohm)			
Lead Pitch (mm)	0,3 mm	0,25 mm	0,25 mm
I/O data (GT/s)	56 G/s	56 G/s	112 G/s
Spring Probe (50 ohm)			
Contact force (N)	0,3 (N)	0,2 (N)	0,2 (N)
Contact resistance (m ohm)	28 mOhm	28 mOhm	15 mOhm

BGA – SoC – Conductive Rubber			
Lead Pitch (mm)	0,3 mm	0,25 mm	0,25 mm
I/O data (GT/s)	56 G/s	56 G/s	112 G/s
Conductive Rubber			
Inductance (nH)	0,1 nH	0,1 nH	0,05 nH
Contact Stroke (mm)	0,1 mm	0,1 mm	0,05 mm
Contact force (N)	0.1	0.1	
Contact resistance (m ohm)	20 mOhm	20 mOhm	10 mOhm
Thickness (mm)	0.5	0.5	

QFP/QFN –SoC – Contact blade + Rubber			
QFP/QFN –SoC			
Lead Pitch (mm)	0.3	0.3	0.3
Data rate (GT/s)	20	40	40

Contact blade + Rubber			
Inductance (nH)	0.15	＜0.1	＜0.1
Contact Stroke (mm)	0.2	0.2	0.2
Contact force (N)	0.2-0.3	0.2-0.3	0.2-0.3
Contact resistance (m ohm)	30	30	30

Conductive rubber type contactors are used for BGA high frequency SoC devices. Conductive metal particles are aligned vertically in insulating silicone rubber which enables vertical contact and adjacent conductor isolation. Compared to other contacts, it is superior for uses with high frequency device test due to its low inductance and low contact height, but compression travel is limited. Conductive rubber will meet the fine-pitch requirement in the roadmap, but it is difficult to reduce contact force without decreasing the compression travel.

Contact blade + Rubber, generally used for testing QFP/QFN high frequency SoCs, is a combined structure of a short-length metal contact and compression rubber that makes contact thru force and travel. The required compression force can be varied by changing the rubber material, but the life cycle is normally shorter than for a Contact Blade type contact.

Socket lifetime has not been pursued in this roadmap, but the lifetime problem will become more important in the near future as lead, ball and pad pitch becomes finer and pin counts get higher, which drives lower contact force to avoid lead/ball damage. Pb-free devices require higher contact forces than are required for non Pb-free packages.

Electrical Requirements

Socket electrical requirements include current carrying capacity (CCC) per pin, contact resistance, inductance, impedance, and signal integrity parameters such as insertion loss, return loss, and cross-talk. The higher the power and bandwidth the packages are designed for, the higher the CCC, the lower the resistance, and the better matched the impedance of the pins and/or sockets need to be. Data rate requirements over the roadmap timeframe are expected to exceed 20 GHz, which will greatly challenge impedance matching and potential signal loss. As package size, solder resist opening, and pitches become smaller and pin counts higher, the smaller pins required to fit within tighter mechanical constraints will greatly increase contact resistance and signal integrity issues. One of the critical parameters to stabilize the electrical contact and ensure low contact resistance is the contact force per pin, which generally ranges from 20 ~ 30 grams. As pitches get finer, smaller and more slender pins will be required, which may not be able to sustain a high enough contact force to have reasonable contact resistance. Due to the negative impact of mechanical requirements on electrical properties, it will be necessary to have improved electrical contact technologies or socketing innovations, in which the electrical properties and signal integrity will not be significantly impacted by or will be independent from stringent mechanical requirements. To handle these high-frequency signals, the user has to carefully consider the signal integrity of the overall test system including board design/components/socket.

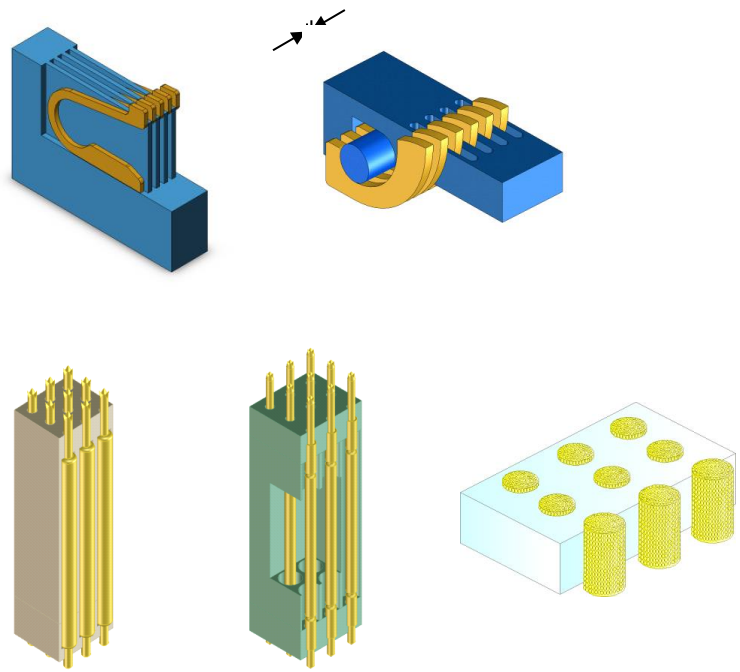


Figure 8.10: Contactor Types

Mechanical Requirements

The mechanical requirements include mechanical alignment, compliance, and pin reliability. Mechanical alignment has been greatly challenged by higher pin counts and smaller solder resist openings, particularly in land grid array (LGA) applications. Currently, the majority of test sockets use passive alignment control in which the contact accuracy between pin and solder resist opening is determined by the tolerance stack-up of mechanical guiding mechanisms. The limit of passive alignment capability is quickly being reached because manufacturing tolerance control is approximately a few microns. The employment of active alignment or an optical handling system is one of the options to enable continuous size reduction of package and solder resist opening, smaller pitches, and higher pin counts.

Compliance is considered as the mechanical contact accuracy in the third dimension (Z-direction), in which the total contact stroke should take into account both the co-planarity of operating pin height and the non-flatness of the DUT pins, in addition to a minimum required pin compression. In general, the total stroke of the contact is between 0.3 mm and 0.5 mm. However, as required pin sizes get smaller, it may not be feasible to maintain the same stroke and thus the compression issue may become the bottleneck of electrical contact performance.

Contactor pin reliability and pin tip wear-out have also experienced challenges because tight geometric constraints prevent adding redundant strength to the pins. The testing environment becomes more difficult with higher temperatures, higher currents, smaller pin tip contacts, etc.

8.9 System Level Test

The section dedicated to system level test (SLT)[17] was introduced for the first time in the 2019 edition of the HIR Test Chapter.[18] As such, it was written much like a whitepaper covering historical background, then-current practices, gaps, challenges, and future needs discernable at the time. In the few years since then, broadened penetration of semiconductor electronics into the multitude of systems that govern our daily lives has significantly affected the role of SLT to meet user expectations in aspects such as quality, reliability, and safety. This update will focus on what's next for SLT from the refreshed perspective of today. For readers less familiar with SLT, a review of the 2019 edition as well as some recent topical papers are recommended.[19] [20] [21]

8.9.1 Executive Summary

While increasing integration and complexity continue to drive the need for SLT, two recent trends are impacting SLT from additional directions:

1. The rise of "bespoke" silicon optimized for specific application domains dictated by system architects.
2. Integration of chiplets in advanced packaging to realize optimized end-system products.

Behind these trends is the accelerating demand in computing and communications far outpacing slowing performance improvements offered by continued semiconductor technology scaling. Both trends impact upstream testing of the components and sub-systems that eventually form the final system. In essence, even if it's not feasible to perform full-fledged SLT, some aspects of the end-system need to be considered in the way individual components are tested. Thus, instead of viewing of SLT as a traditional last-stage test insertion, various forms of system-oriented testing need to occur at every stage from wafer sort, through die stack, packaging, to assembled sub-systems.

Rapid proliferation of AI applications in the cloud and at the edge has made the importance of energy-efficient computing paramount. With the death of Dennard scaling and untenable increase in multi-core complexity, system providers are resorting to novel architectures to meet power and thermal constraints.

[17] https://eps.ieee.org/images/files/HIR_2019/HIR1_ch17_test08.pdf

[18] https://eps.ieee.org/technology/heterogeneous-integration-roadmap/2019-edition/hir-test-chapter.html

[19] Beyond Structural Test, the Rising Need for System-Level Test, https://ieeexplore.ieee.org/document/8373238

[20] Exploring the Mysteries of System-Level Test, https://ieeexplore.ieee.org/document/9301557

[21] System-Level Test: State of the Art and Challenges, https://ieeexplore.ieee.org/document/9486708

Architects resort to bespoke silicon and chiplets to maximize performance for specific use cases and data types via SW-HW co-optimization.[22] [23] [24]

However, during stand-alone testing of individual components prior to integration, more nuanced settings of test conditions and pass/fail criteria are needed when the full system SW-HW context is lacking. System scenarios that may unduly create stress conditions on individual components causing system failure are hard to anticipate. Overall system performance and reliability are degraded by the weakest member in the set of assembled components. The key challenge is thus the mapping of system context to the upstream testing of individual components.

8.9.2 Enablers and Challenges of System-oriented Test

- Flexible DFT architecture allows delivery and execution of both structural and functional test content on multiple tester platforms spanning ATE and in-system.
- Low-cost multi-site high-throughput system functional testers.
- Tight link to system verification for rapid functional test development on both ATE and SLT platforms.
- Transient fault modeling and analysis to better reflect failures at system level.
- Effective SW-HW system failure diagnosis methods for efficient root-causing and yield learning.
- Deep extraction of component internal parametrics that can be correlated with system behavior via advanced data analytics.
- Creation of deep data models can predict how a component will likely behave in the system as well as finding a set of compatible components to meet integrated system performance targets.
- Closer collaboration among supply chain parties to share data and create effective predictive models.
- Standards and practices to meet security requirements despite potentially enlarged threat surface caused by increased data access and sharing.

8.10 Data Analytics

8.10.1 Background

An IEEE Xplore® database search yields publications on Data Analytics for Adaptive Test and Yield Learning dating back 30+ years. While advances have been achieved over the last several decades, it's challenging to apply the techniques holistically across the full semiconductor value chain. Limitations on our ability to efficiently collect, store, and analyze the massive amounts of available data have limited adaptation to well-defined and self-contained applications. During the past 5-10 years, multiple technological advances have combined to change this landscape significantly:

- The Internet of Things has facilitated the efficient collection of massive amounts of data

[22] https://semiengineering.com/ic-architectures-shift-as-oems-narrow-their-focus/
[23] https://semiengineering.com/bespoke-silicon-rattles-chip-design-ecosystem/
[24] https://semiengineering.com/rise-of-the-fabless-idms/

- Cloud Computing and Big Data technologies have turned data silos into Data Lakes and Data Meshes
- Tremendous advances in computational power and parallel processing have facilitated the adoption of advanced Data Analytics and machine learning models
- The combination of all the above has enabled rapid advancements in algorithm design and implementation

The foundation is now in place to strategically improve Adaptive Test and Yield Learning, by implementing Data Analytics, Big Data, and Machine Learning techniques.

8.10.2 Why is Data Analytics Important for Semiconductor Manufacturing and Test?

Today's challenges of increased design complexity including Heterogeneous Integration (HI) packages, functionality, shrinking process nodes, increased quality and reliability requirements, and shortened time to market have combined to drive an exponential level of pressure to improve the semiconductor value chain. A massive amount of data – we conservatively estimate multiple terabytes (TB) of data (device and operational) per day for a fully-loaded high volume back-end operation – is collected across the semiconductor manufacturing supply chain and test flow [1]. That data contains a wealth of information that can help optimize the overall test flow and discover hidden issues and relationships across process steps. For example, if the correlation between process drifts and yield is fully understood, immediate actions can be taken to maximize profit and ensure supply (e.g., predictive analytics). A multitude of key insights can be unlocked by using advanced Data Analytics. Data collection during production test should strategically be designed to take full advantage of new and different analytic techniques.

Data collected at test is critical for driving learning and optimization during the product lifecycle using automated data analytics. This includes:

- Cost of test and back-end operations (including test content optimization across test steps)
- Yield (optimized across all test steps)
    - to drive repair/redundancy, die matching – including chiplets, on-die trim, dynamic voltage scaling and fail data collection for diagnosis
- Product Quality – including shipped DPM and product reliability
- Product Performance – such as speed, power and functionality/repair
    - this includes data gathered from on-die monitors and sensors
- Supply chain traceability – such as all components that are used in a HI package
- Time-to-Market, efficient product introduction, feedback to design

A clear requirement is that all test results (e.g., wafer probe test, final test, SLT) and other data from across the semiconductor value chain will need to be merged and available for these analytics, while maintaining high levels of data security and trust for both data at rest and data in motion across pipelines between entities.

Database and IT infrastructure is critical to enable data analytics. Cloud Technology is a key enabler for end-to-end test data analytics across multiple test steps in the value chain from silicon to system test (full product lifecycle). Analytics will be applied at multiple levels including off-line in the Cloud, local to the tester, and at the Edge (for reduced latency and real-time decision-making).

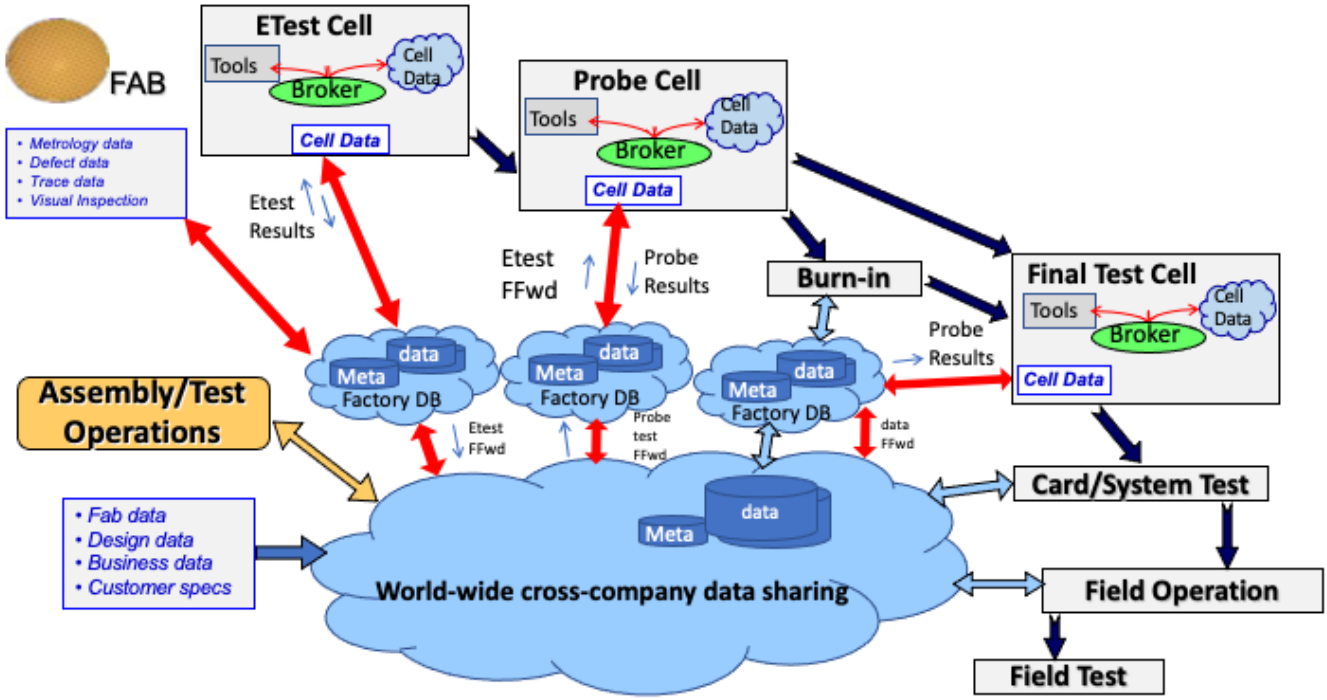


Figure 8.11: The architecture of Adaptive Test organizes each insertion's test data into one or more databases. A waterfall of manufactured parts may insert, join or query databases for test flow decision making.

Data analytics requires real-time analysis to enable capabilities such as Adaptive Testing, as shown in Figure 1. This analysis is done either local to the tester or at the Edge, within the required latency to drive production test and dispositioning.

Heterogeneous integration is increasing the importance of data analytic capabilities due to the complexity of combining many different dies – sometimes from multiple suppliers – onto the same package. Tasks such as yield analysis require the merging and analysis of a wider set of data from these dies and packages.

Failure or delay in applying modern Data Analytics holistically across the semiconductor value chain will lead to increased costs and risks as design, fab, assembly, and test complexities increase, and stop-gap measures are implemented to reach quality targets. Attempts to optimize manufacturing process steps individually, without full consideration of the interactions and dependencies across the entire process flow, will lead to diminishing returns. A test escape anywhere in the process flow reduces the quality level of the overall flow.

Tactical, localized solutions to manufacturing challenges are usually costly. One example is adding a System-Level Test (SLT) insertion as a back-end quality screen. By the time issues are detected, the manufacturing process is typically so far downstream that the effort required to truly root-cause and resolve them is only justified for the most major and systemic issues. With Data Analytics that relate SLT fails to other process and test data, issues can be caught sooner and SLT becomes one of a series of test insertions rather than a backstop.

Multiple benefits of applying advanced Data Analytics are described in the sections below.

8.10.3 Transforming the Backend to an Industry 4.0 Smart Factory

In essence, Smart Manufacturing is the trend towards automation, enhanced data connectivity and advanced analytics to improve efficiency. It involves automating repeatable tasks, using data from process, production, assets maintenance, and production planning to gain actionable insights through analytics. The path to achieving this includes the use of cyber-physical systems (CPS), the Internet of Things (IoT), Industrial Internet of Things (IIoT), cloud computing, cognitive computing, and machine learning.

Today, the big push towards a smart factory is to:

- Reduce costs
- Improve overall equipment effectiveness (increased uptime, accelerated output, decreased faults)
- Improve quality
- Enable secure data exchange across the value chain to improve visibility and productivity

The key requirement is the ability to collect, share and act on the data. For Smart Manufacturing there are three dominant data perspectives:

- Historic state - Review, analyze and model historical performance
- Current state - Monitor current state to enable real-time control
- Future state - Use history and current data to identify and plan for future improvements

It is crucial in Smart Manufacturing to measure and characterize every aspect of the manufacturing process, including logistics, products, machines and processes, without scrambling to consolidate data. A consistent and detailed strategy for collecting, analyzing, and categorizing data is essential.

8.10.4 Optimizing Cost of Test

Savings from cost of test reductions are easily quantifiable, as they drop directly to the bottom line as increased profit. This benefit must be balanced with other factors that can potentially have significantly greater impact on profitability and competitiveness, such as improved yield, quality, and reliability. (See *Section 11: Key Drivers and Test Costs* for a detailed comparison of these impacts.) Advanced Data Analytics are key to achieving this balance, through their ability to identify complex effects and interdependencies, both within a specific test insertion and across the entire test flow. Examples:

- Data Analytics-driven decisions, including those made on-the-fly based on results from the current and/or previous insertions, support a smart, adaptive approach for optimizing test coverage at reasonable cost.
- HI-related technologies such as chiplets drive a “shift-left” of testing to earlier insertions to guarantee known good die (KGD) as well as providing the data necessary for speed binning and die matching. Data Analytics facilitates an efficient and cost-effective shift-left strategy through correlation of results across all test insertions from wafer probe through SLT.
- Correlation of test results across multiple insertions facilitates moving test seconds to lower-cost (or even fully depreciated) equipment while maintaining required test coverage.

- Incorporating additional design-for-test (DFT) circuitry can reduce test system requirements but must be applied carefully as it uses valuable chip real estate. With the complexities that come with advanced packaging, embedded sensor IP provides an effective means for monitoring chip performance and functionality at a deeper level. Data Analytics play a key role in maximizing the information that can be inferred from this sensor data across test insertions as well as in-field.
- Machine Learning models are being successfully used to reduce test cost by replacing time-consuming searches (for example, determining trim values or setting test parameters such as voltage levels) with fast predictions based on previously collected data [2, 3].

This ability to optimize across the entire flow becomes especially important to keep test costs under control for complex devices requiring an added System-Level Test insertion, or devices for automotive applications that have rigorous multi-temperature testing and burn-in requirements. Optimizing test across the entire flow requires tools and standards that support the efficient combination of data from different processes, devices, and equipment.

8.10.5 Improving Quality Assurance

The primary objective of production test is quality assurance. Data analytics provides a powerful means for ensuring that devices are meeting functionality, quality, and reliability requirements by inferring additional information on existing failure modes and potential quality and reliability issues while maintaining an economically viable test strategy. Applying data analytics cohesively on test data from multiple test steps further increases overall effective test coverage. This capability is especially valuable in market segments that require high reliability such as automotive, military, aerospace, and medical devices, which strive for “zero defects” outcomes.

Some defects do not manifest during testing or initial operation. For example, recent advances in the awareness of Silent Data Errors have led to calls for additional screening and outlier detection, particularly on devices that exhibit some degree of abnormal behavior even when passing all tests. For this reason, adding more test coverage and/or insertions, which adds cost, may not meet the stated goal of zero defects. Instead, Advanced Data Analytics can be used to minimize test escapes by optimizing the test content at each insertion, and inferring additional valuable information from the combined results data.

Traditional outlier detection techniques utilizing statistical post-processing are well understood, but may not be adequate for catching potential reliability issues at Final Test or System Level Test. Near-real-time statistical techniques will be particularly valuable for devices that do not have individual device traceability, since re-binning in near real-time allows the prober or handler to re-bin devices before they get lost in the population.

Real-time outlier detection offers a potentially useful addition to the set of tools for achieving high reliability. Near real-time analytics is relatively inexpensive compared to additional test time, and is capable of identifying test process issues such as site-to-site bias, enabling corrective action sooner than would be possible with post processing.

Data feed-forward methods are used to analyze upstream test data, to adaptively determine the appropriate downstream test content and minimize test escape rates. Data feed-backward methods are used to adjust the manufacturing process and shorten the time to achieve entitlement yield and quality.

Correlations across test insertions can identify drift or other issues. When available, historical data should be mined to set baselines, screening limits, and guardbands. As new test methods are deployed, data analysis can measure the impact to ensure no new test escapes are created when displacing other forms of testing.

When applied as described, data analytics can contribute greatly to reduced Time To Quality (TTQ) and therefore to reduced Time To Market (TTM). Achieving this vision will require more standardization of data formats and traceability wherever feasible. Analytics software will need to be demonstrably secure, and capable of running on multiple data systems.

8.10.6 Improving Yield

Heterogeneous integration presents difficult challenges in terms of both yield prediction for the chiplets (the "known good die" or KGD problem) as well as diagnosing yield losses for the packaged product. Full electrical testing of the individual chiplets prior to package assembly is technically challenging and cost prohibitive for the supply chain and, moreover, defects may occur not only at the chiplet level but throughout the entire package manufacturing and assembly process. The chiplets may be manufactured in different process nodes and at multiple foundries, leading to a vast Pareto of possible defect types, and assembly processes such as wafer-to-wafer stacking or die-to-wafer stacking introduce even more defect sources. This creates test coverage challenges in the fully packaged product, leading to extensive and costly electrical testing. Furthermore, late detection of bad chiplets at package test leads to costly loss of the other good chiplets in the package.

Collecting data at all stages of the manufacturing process can provide complete material traceability and overcome gaps in the conventionally recorded genealogy of the packaged part (e.g., ECID). This richer data set enables new data analytics to trade-off cost versus resolution of test and diagnosis throughout the HI process, and enables die matching to improve overall HI product yield. The time lags inherent in chiplet silicon manufacturing and package assembly processes couple with test and diagnosis challenges to create time-to-yield issues resulting in time-to-market issues, making yield improvement throughout the heterogeneous integration process a critical component to product success.

DFT techniques developed originally for SoC products must also be incorporated in the heterogeneously integrated products without driving increased resources or test time. Resilience must also be designed into the chiplet and package architecture to realistically achieve full functionality. This can be accomplished by additional resources such as redundant TSV's or bonds as well as redundant memory and logic circuits. Data analytics across the entire supply chain will play a crucial role in collecting and model building to allow for the optimization of the system resiliency architecture.

8.10.7 Performance Grading/Binning

Performance grading and binning of devices has been a common technique for many years for tiered products, for example memory or processors. More recently, with mobile and energy-conscious applications, there is a need for an improved performance understanding which could lead to either traditional product binning/grading or product applications for improved energy/performance trade-offs. With Heterogeneous Integration, understanding of device performance at the wafer level and concepts like die matching or calibration will be of paramount importance. To help achieve these goals, there have been improvements in on-chip sensing technology for both process variation and operational parameter

monitoring under various operating conditions. The combination of sensor networks and advanced data analytics provides new signatures at the die level for enhanced binning.

Performance binning for the frequency vs voltage trade-off is also changing. Traditionally this has been accomplished by shmooing voltages and frequency to determine the maximum operating point. These test techniques can be expensive from both a test time and test intensity perspective. Trained models are starting to be deployed based on early characterization data to pre-determine operating points. This can be done as a data feed forward to later test/assembly steps or as an in-situ decision for binning. Doing so results in improved product economics - yield, test time and optimal operating conditions.

Previously, data sources have been focused on voltage, temperature and process. Other measurement parameters have also emerged as critical on-die measurements. An example is a margin measurement that is placed on critical timing paths or interfaces. It provides visibility on the amount of timing margin, which will further indicate performance optimizations or more quickly determine the quality of the device grading being performed. This leads to a better understanding of design margin for optimal operation.

It is also expected that innovative approaches will emerge combining financial, sales and device data to tailor deliverables that exactly match customer requirements. This tuning optimizes manufacturing and test processes, which increases margins, improves lead time and increases supply elasticity.

8.10.8 Traceability Across the Semiconductor Value Chain

With increasingly stringent reliability requirements and use of HI, we must have more visibility into the assembly processes where the root cause for hard to pinpoint reliability failures often occur. With the complex supply chain for the HI assembled product, security considerations have become of paramount importance.

Analyzing failures and security events in electronic devices requires traceability at the individual device level to access the manufacturing, test and root of trust data. Virtual identifiers based on SEMI E142 [4] can provide a basis for single device traceability from any point in the supply chain (wafer, package, PCB, field) both downstream and upstream [5]. The data model is applicable from the wafer through traditional and more advanced packaging technologies such as wafer level packaging and the heterogeneous integration of chiplets. This enables precise analysis for pinpointing the root cause of a failure, for example a rare early life failure of a wire-bond in the field, or for pinpointing the source of a security attack, for example rapid detection and mitigation of counterfeits, Trojans, and malware attacks.

An on-chip electronic identifier (ECID) can be used to trace back to wafer test and further back into wafer fab for root cause analysis. However, this only applies to the primary active components with ECID and does not provide any visibility into the assembly processes . We must add traceability to assembly to capture every active and passive component, bump and wire contact, consumables, equipment, Failure Detection Classification (FDC) trace and inspection images.

8.10.9 Data Analytics for Test - Key Enablers Roadmap

In the table below, the key enablers for realizing the full potential of advanced data analytics to optimize the test process across the semiconductor value chain are listed. For each enabler, the current status is described, as well as the 3-5 year projection of how the enabler needs to evolve to support the bold visions described in the sections above. Importantly, progress on the enablers needs to be comprehensive,

as a delay in the development of any of them can hold back overall progress on the successful implementation of advanced data analytics solutions for semiconductor test.

Table 8.22: Key enablers to realize full potential of advanced data analytics to optimize the test process across the semiconductor value chain

Enabler	Current Status	3-5 year Projection
Machine-to-machine IoT communications infrastructure	Mostly Point-to-point ethernet. Early use of MQTT M2M.	Wide use of MQTT M2M-like net.
Standardization of data formats	Several application specific formats; STDF, SECS combined with generics; CSV, text, binary.	Move to more data-centric formats which support real-time analysis, including streaming data formats.
Real-time analytics - more local processing	Dependent on tester capability. Off-line analysis common.	Cells become data centers. Local real-time processing on test. cell or Edge compute server. Distributed analysis and storage
Strategy for collecting, analyzing, and categorizing data	Most data is indexed via file paths and location. Databases are used for access. Data is mutable and hard to find.	All data indexed via metadata. Emphasis on provenance and trust. Data mesh architectures common.
Characterize every aspect of the manufacturing process	Test data is generally available locally. Non-test data is not common.	All collected data available. Continual addition of new data.
Use of device-sourced data, sensors and test structures	Some use of on die sensors for analysis. Mostly post-processing.	Pervasive use of on-die, in-package and in-system test data sourced from the entire life cycle. Efficient real-time access to on-die sensor data.
Big Data technologies - cloud - local	Storage is limited especially globally due to cost.	Distributed analysis to reduce data size impacts.
Advanced Data Analytics and machine learning models	Some well known techniques. Part Average Testing. Outliers, neighborhoods. Limited in scope due to knowledge models. Some use of machine learning models, mainly for COT reduction.	Extension to non test data. Rule based models common. Pervasive use of machine learning for test optimization, yield enhancement, and quality/reliability improvement. Greater use of unsupervised learning algorithms for anomaly detection, correlations, ...
Pipelines between entities	Ad hoc contract-based solutions. Requires experts to share.	Shareable cross domain models. Knowledge shared effectively along with the data.
Data security	Encryption is used. Some data hiding techniques.	Encrypted analytics and models reduce the need to share raw data.

8.10.10 Impact of COVID-19 on Data Analytics Roadmap (Special Section for 2023)

In 2020 we had expected COVID-19 to be an accelerating force in the adoption of advanced Data Analytics, with key drivers being the move to remote (work, data access, support, etc) requirements for efficient meshing of cloud and edge compute resources, increased supply chain stress, and greater reliance on predictive analytics/diagnostics. Together these factors have driven an urgent need to bridge the worlds of test engineering and data science. A key question at the time was how strongly these drivers would persist in the post-pandemic world. This hoped-for post-pandemic world has yet to arrive, and instead the world has adapted to living with COVID-19. Ongoing severe issues such as those with the global supply chain have been exposed as systemic problems requiring new approaches rather than quick and temporary fixes. The bridging of test engineering and data science is in progress but increased focus is required to bring the level of expertise in line with the magnitude of the challenges. Government subsidies such as the US and European CHIPS Acts provide important and timely fuel for furthering the development and application of Data Analytics in the semiconductor industry. The combination of a strong requirement for better capabilities in this space and massive government funding provides a great opportunity to make fast progress, but prudent spending will be key to getting the best return on these investments.

8.10.11 Additional Reading

For further reading on *Adaptive Test* and *Yield Learning* topics, please see ***Data Analytics - Appendix A.***

8.10.12 References

1. A. Meixner, *Too Much Fab and Test Data, Low Utilization*, https://semiengineering.com/too-much-fab-and-test-data-low-utilization/, Semiconductor Engineering, Jan 2021.
2. S. Mier, *ML in Semiconductor Test - A Balanced Approach*, https://events.meptec.org/wp-content/uploads/ChipletsDataTest/DataTest2021Mier.pdf, MEPTEC Road to Chiplets: Data & Test, Nov 2021.
3. M. Eiki, K. Schaub, I. Leventhal, B. Buras, *In Test Flow Neural Network Inference on the V93000 SmarTest Test Cell Controller*, 2019 International Test Conference, Nov 2019.
4. *SEMI E142 Substrate Mapping Standard*, https://www.semi.org/en/standards-watch-2020Sept/revision-to-semi-e142.
5. *The GSA Trusted IoT Ecosystem Security (TIES)*, https://www.gsaglobal.org/iot/ties/.

8.11 2.5D & 3D Device Testing

8.11.1 Introduction

2.5D and 3D technologies (see figure 1) are characteristics of a system and should be tested as such: testing the complete package at an application level and diagnosing failures at the die and interconnect level. This section will address key test challenges, based on the evolution of 2.5D/3D. These test challenges are with respect to known good dies (KGDs), interposers, high speed interconnects and signal

integrity, impact of emerging technologies, 3D TSV/interconnect, as well as 3D probing, die stacks, and stack repair.

Memory die stacks (Wide I/O, High Bandwidth Memory, and Hybrid Memory Cube) were precursors to 2.5D and 3D. Both technologies have provided insights to requirements and challenges associated with 3D and 2.5D test. The best that can be gleaned from these technologies at this time is that reliance on BIST and boundary-scan based technologies, and use of fault tolerance with simple configurations, tend to produce relatively high yields at the stack level. As these adjacent technologies become more mature and as additional 2.5D/3D-TSV applications emerge, more and better data will improve predictions and decision making, with respect to 2.5D/3D-TSV test processes.

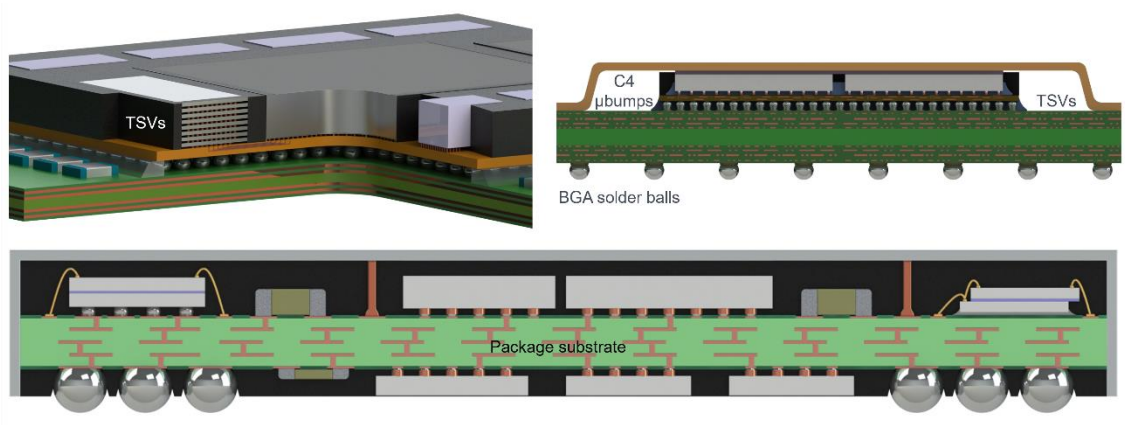


Figure 8.12: 2.5D/3D Technology (Amkor Technology, Inc.) [7]

8.11.2 Challenges for Test

While the current state of 2.5D/3D is maturing, new enabling and supporting technologies will require advances in test access, capabilities, and costs. These emerging technologies will provide significant challenges for testing 2.5D and 3D technologies. The challenges below represent potential impacts to test, including increased costs, longer test times, and reduced yields and reliability.

8.11.3 Known Good Die (KGD) Test

Due to yield concerns at the final package level, incoming bare die should have as good a quality as possible. KGD is the common industry term – but KGD does not mean that 100% of the bare dies will pass all testing at the next level of assembly. Chiplet suppliers should provide an estimate to their customers of the expected fallout at package test or later testing steps.

To achieve high quality, chiplets should see as much testing at wafer probe as possible. But there are challenges since for advanced technology chiplets (such as fine-pitch µ-bump or Copper Hybrid Bonding) not all signal IO will be probed at wafer test. Instead, Design-for-Test (DFT) solutions should enable high test coverage at wafer probe without the requirement to probe all signal IOs. (Using test-only pads is the most common solution). This DFT, combined with the IEEE 1838 standard for chiplet access, should be used to apply tests pre- and post-packaging.

8.11.4 Interposer Testing

Interposer testing can be accomplished primarily by point-to-point continuity probing. Multipoint interposer testing requires significantly more probing (more time and higher costs) and requires embedded logic to coordinate point-to-multipoint connections. Over time, it is expected that probing becomes more challenging, from Point-to-Point to larger-scale Multipoint.

Known good interposers (KGI) are vital to ensure adequate yields for advanced packages. Post-package assembly, IEEE P1838 primary and secondary TAP ports allow for testing the die-to-die test access and interconnect performance integrity.

8.11.5 High Speed Interconnects and Signal Integrity

Testing high speed interfaces (HSIO) requires access and ability to run test patterns specifically on each interface, whether in a loopback mode from transmit ports to receive ports or from one chip transmitting to an adjacent receiving chip in an integration. Design for test is needed to run these tests standalone via an easy-to-use interface such as JTAG 1149.1 or 1149.6, SPI, J2C (JTAG to CPU), or PCIe. These tests should have the ability to be run at wafer sort, package test and system test in characterization and production. Designers need to understand defect mechanisms of the HSIO, and what testing will cover all defect types and guarantee outgoing quality. This can be time-consuming and expensive for silicon area. The IEEE1838 standard is available, but a user must go through details of the IO DFT for each die-to-die connection today to ensure an implementation will work. Some applications have a high count of HSIO (512-1025) lanes. Current ATE generally can only test this number of IO up to low GHz range (<20GHz). High speed add-ons such as bit-error-rate testers (BERTs) and digital sampling oscilloscopes (DSOs) are available but are limited to 32-64 direct connections. Loopback testing needs DFT like BERs and PHY control built in to be most effective. Often a second test step is required to perform the loopback test.

Handling noise and thermal cross talk across multiple chiplets can erode HSIO margin, and considerations as to whether special packaging or shielding may be needed. ATE testing does not lend itself well to testing high-speed optical interfaces of photonic devices. There are no production test solutions for multiple optical port photonic devices.

3D interconnects on a product can exceed 100,000. They are becoming increasingly denser (< 3µm), the interconnect technology is evolving, and each new generation brings complicated failure mechanisms (see ref [1], [2], and [3]). A standardized test and repair methodology that considers these trends in 3D interconnects would be helpful.

8.11.6 Impact of emerging technologies with respect to test

The challenge of wafer probe testing is emerging with many more chip-to-chip connections such as Copper Hybrid Bonding (CHB). Reduced pin count testing will be required for wafer probe. Design-for-Test (DFT) solutions are required to enable complete wafer probe testing.

Die-to-die interconnect testing at final package test will require solutions that enable complete testing and failure diagnostics. Emerging solutions such as UCIe and Bunch-of-Wires (BOW – see ref [5] and [6]) should be explored and ideally an industry standard will emerge to ensure chiplets and SOCs/processor chips can be tested.

Another challenge is the test methods for other types of circuits (not just digital) for 2.5D/3D packaging: for example, test methods for Silicon Photonics, RF and high-speed mixed signal. Some of these circuit types have not typically had the same level of DFT as digital circuits.

8.11.7 3D TSV/interconnect testing

Silicon interposers include interconnect and through-silicon-via (TSV) structures. Mechanical integrity of these structures during the manufacturing process ensures electrical performance. DC and AC transient pre-bond testing of interposers helps in screening micro-void and pinhole defects. Testing of interposers may require custom test fixture development and test insertion, which impacts the overall product cost. Custom implementations may require complex test techniques (see ref [4]).

8.11.8 3D probing, 3D die stacks, 3D stack repair

3D die stacks offer many potential test moments: pre-bond, mid-bond, post-bond, and final test. The more dies that are contained in the stack, the more pressing is the need for a tool that models the cost and yields of wafer processing, stack assembly, testing, packaging, and logistics, to optimize the stack assembly and test flow.

One of the major pre-bond test challenges is getting test access to the non-bottom dies, where the natural functional interfaces consist of large arrays of fine-pitch micro-bumps. State-of-the-art micro-bump pitches are 40 µm; some advanced products already push this down to 30 µm; and the scaling does not stop there. Feasibility of 40 µm probing has been demonstrated but only single-site – future research should push this to multi-site testing and to even smaller pitches (10 µm).

Once the stacking has commenced, we require specific 3D-DFT (i.e., DFT in addition to the conventional 2D-DFT) to transport test stimuli up into the stack and test responses back down. The 3D-DFT in the various dies should collaborate to form a stack-wide test access architecture. For this purpose, in 2020 the IEEE Std 1838-2019 was released; in the meantime, the three major EDA suppliers have started to provide support for IEEE Std 1838 insertion and usage. The standard supports both INTEST (testing or re-testing the internal circuitry of the die) as well as EXTEST (testing inter-die interconnects).

Stack repair makes sense cost-wise only if the spares are already included as redundancy in the stack. Spares could be individual inter-die interconnects or even full dies; for relatively small investments, spares can significantly increase the overall stack yield. Current-generation products that have seen silicon include redundant interconnects and are already implementing repair.

8.11.9 Long term prediction

It is important to note that 2.5D/3D is an evolving technology, and, because of that, it is difficult currently to make any predictions regarding 2.5D/3D test flows. With this said, 2.5D/3D technologies are expected to create increasingly complicated and time-consuming assembly process flows that can add cost as well as yield challenges to the mix. As packaging technologies and the associated Test challenges will continue to evolve, it is expected that DFT features that can enable yield troubleshooting across all process steps will become a focus for the industry.

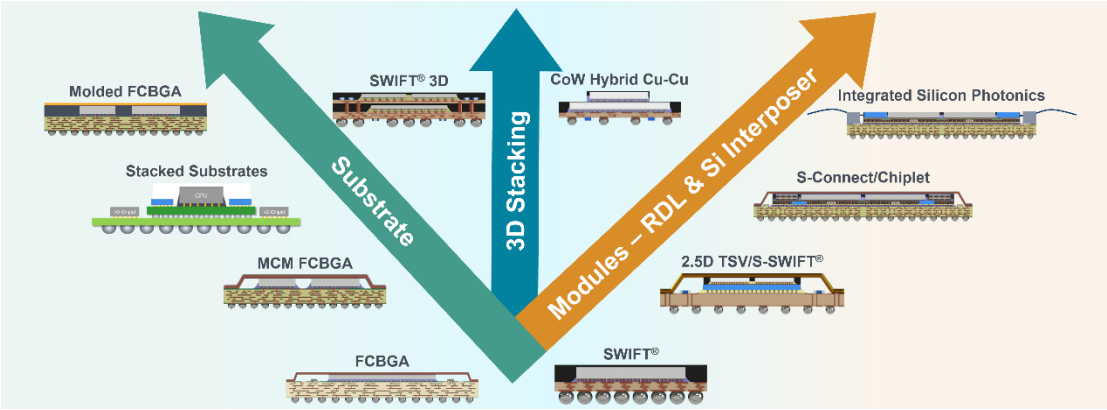


Figure 8.13: Packaging Technology (Amkor Technology, Inc.) [7]

8.11.10Call-for-action

To help 2.5D/3D Device testing capability to mature and improve, there are several areas that need attention:

- Known Good Die DFT methods that enable high quality wafer probe test – thus reducing fallout at final test.
- Die-to-die communication standards that enable thorough testing at final test.
- Repair methods at final test to ensure yield is high.
- A standardized test and repair methodology that considers new trends in 3D interconnects.
- Yield prediction and analysis methods that ensure fallout at all levels of testing is understood.
- End-to-End data analytics capability that applies to all dies on the package (see section 9, Data Analytics)

8.11.11References

[1] A. Elsherbini, et. al., “Hybrid Bonding Interconnect for Advanced Heterogeneously Integrated Processor, IEEE Electronic Components and Technology Conference, 2021

[2] Sreejit Chakravarty, “A Call to Standardize Chiplet Interconnect Testing,” IEEE VLSI Test Symposium, 2022

[3] Sreejit Chakravarty, “3D Interconnect Testing Challenges,” IEEE European Test Symposium, 2022

[4] Sourav Das, Su Fei and Sreejit Chakravarty, “Testing of Pre-bond Through Silicon Vias”, IEEE Design and Test, July/August 2020, pp. 27-34.

[5] R. Farjadrad, M. Kuemerle and B. Vinnakota, "A Bunch-of-Wires (BoW) Interface for Interchiplet Communication," in IEEE Micro, vol. 40, no. 1, pp. 15-24, 1 Jan.-Feb. 2020, doi: 10.1109/MM.2019.2950352.

[6] S. Ardalan, R. Farjadrad, M. Kuemerle, K. Poulton, S. Subramaniam and B. Vinnakota, "An Open Inter-Chiplet Communication Link: Bunch of Wires (BoW)," in IEEE Micro, vol. 41, no. 1, pp. 54-60, 1 Jan.-Feb. 2021, doi: 10.1109/MM.2020.3040410.

[7] © 2022, All content is copyright its respective owners. SWIFT is a registered trademark of Amkor Technology, Inc.

8.12 Key Drivers and Test Costs

Minimizing costs is a key goal of any manufacturing process. Test is no exception, although steady improvements in efficiencies over the last 15 years have lowered the typical cost of test as a percentage of IC revenue to less than 2-3%. The primary drivers of increased efficiency have been reductions in capital costs per resource and lower test times, coupled with increases in parallelism and Built-In Self-Test

(BIST) capability. Most SOC devices are tested 2 to 16 at a time, and memory devices can have more than 1,000 devices tested at once.

The typical perception of test costs is that it is dominated by the test equipment itself, which is extremely costly. In reality, that is not the case. The depreciation cost of the test equipment itself (which has a useful life of 15-20 years) typically constitutes less than half the cost to operate a complete test cell, and that cost is zero after the depreciation period (typically 5 or 6 years) has expired.

For large SOC devices, it is notable that, since 2015, the cost of consumable material – material that is expected to be used and then discarded - has become the leading capital expenditure relative to test. This situation stems from:

- The increased cost of interface material (primarily influenced by probe cards and relative items). This cost is driven by finer-pitch probe pads and sockets, and the need for increased maintenance and repair, especially for high current applications.
- The decreasing depreciation period for materials utilized to produce devices used in the mobile device space, where devices have a shorter life span. In most cases, material is typically discarded not because it has ceased to function, but rather because the devices it is used to test are replaced by newer versions which drive different consumable hardware.
- The increasing use of System-Level Test (SLT), where costs are dominated by device-specific hardware. These costs recur with every new device version and, as noted above, often has a very short useful life.

For lower complexity devices, especially those that are not produced in very high volumes, test costs are dominated by capital equipment and are highly affected by Overall Equipment Efficiency (OEE). OEE measures the amount of time the equipment is doing useful work. For devices that are produced in lower volumes, the test equipment is usually taken out of production to change over to test different devices. If these configuration changes happen frequently, OEE is significantly degraded. For this reason, site counts are intentionally limited to lower idle time and increase OEE, even if cost of test per device is slightly higher.

8.12.1 Key Cost of Test Trends

Looking forward, there are several trends which will counterbalance equipment efficiency and serve to cause cost increases:

- Increases in transistor count that outstrip on-chip test compression technology will increase the amount of external data which must be supplied to the Device Under Test (DUT). Coupled with scan shift rates that are limited by power and thermal concerns, the overall effect will be longer test times. This situation will be addressed primarily with increased parallelism and new scan technology to increase external data rates and reduce the number of clock cycles required for a given scan test.
- Device configuration and one-time programming during test is causing more time to be spent to perform initial device calibrations or to reconfigure devices based on defects or electrical performance. As silicon geometries shrink and defect densities drive circuit redundancy, repair functions will also add to “test” costs, although these are really “repair” costs.
- The drive to multi-die packages will add a requirement for more System Level (“mission mode”) testing owing to lack of access to individual die. Without significant Design For Test (DFT) improvements, this type of testing can take much longer than conventional structural

test. This will also drive more exhaustive test processes at wafer probe to improve the yield of multi-die packages.

- Site count at probe test is limited owing to the attendant increase in the cost of consumable material (discussed above) and the limitations of Touch-Down Efficiency (TDE). TDE is discussed in more detail below.
- The continuing increase of silicon content in automotive and other end-uses such as military and satellite applications that require a high level of reliability, which drives additional test insertions for fault coverage and temperature-related test.

Continuous improvement in equipment efficiency will be offset by new device test requirements, so the overall cost of test will remain relatively flat for the foreseeable future.

8.12.2 Cost of Test as a Part of Overall Manufacturing Cost

While the cost to own and operate test equipment has been reducing, other semiconductor manufacturing costs have been significantly increasing with new silicon technology. Specifically, fab costs for leading-edge processes have increased to about 70-80% of the overall cost of producing a large-scale SOC device. It now costs far more to fab a device than to test it, and that trend will accelerate as new fabrication technologies are deployed.

The figure below represents third-party analysis by VLSI Research of the capital and service costs of equipment used in device fabrication, packaging and test.

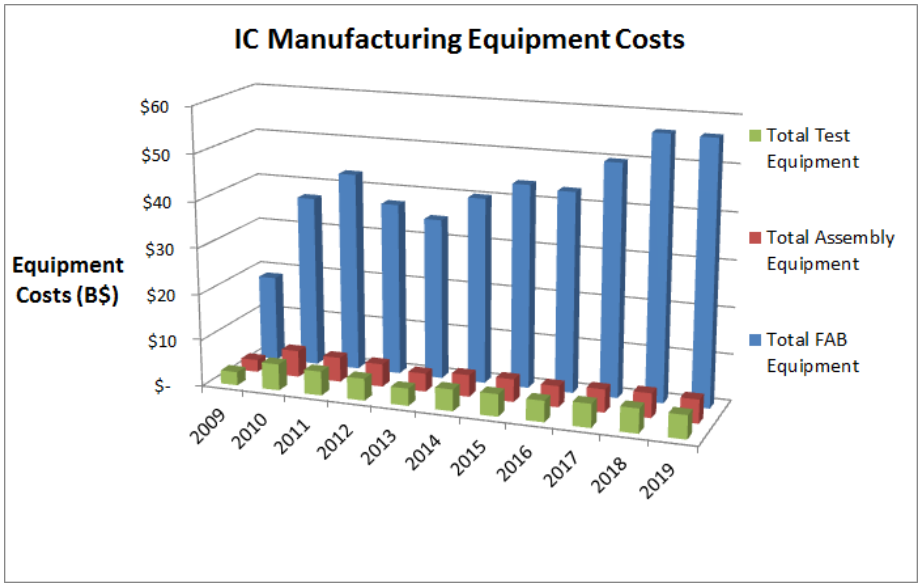


Figure 8.14: Relative cost of Fab, Packaging and Test Equipment

For any device, the worst-case cost scenario is to ship defective devices that cause failures later in the manufacturing process. Presuming this is not the case and defect rates are acceptable, then the next concern is manufacturing costs.

While it is helpful to focus on the cost of test itself, the impact of test costs on overall device costs varies widely by device type. For devices that use smaller die or mature process nodes, tests cost can be a significant part of overall manufacturing costs. For devices that use leading edge processes and have larger die sizes, the contribution to a manufacturer's profitability from lower test costs will be very small, since test is a small part of the device cost overall.

For these more complex devices, the highest avoidable costs in test are devices that are good but are rejected at test for some reason.

Consider the following, simplified example.

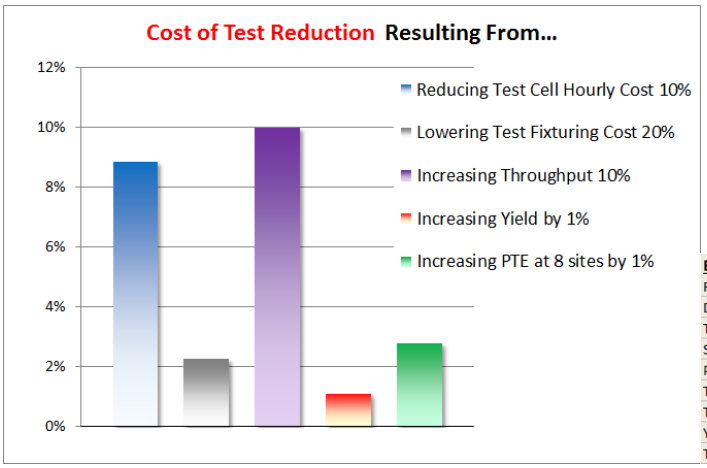
- A device costs $1.00 to manufacture, including fabrication, assembly and packaging, etc.
- Test constitutes 5% of that cost, or $0.05

Reducing the cost of test by 10%, will reduce overall costs by $0.05 X 10% = $0.005 per device.

Improving yield by 1% reduces overall cost by $1.00 * 1% = $0.01 per device.

While the 10% Cost of Test reduction is good, the yield improvement is better.

Figure 2 shows the effect of traditional cost reduction techniques on cost of test.



Example Parameters	
Revenue per device	$10.00
Device production costs before test	$3.00
Test Time (Seconds)	10
Sites	8
Parallel Test Efficiency	98.0%
Test Fixturing Cost	$200,000
Test Fixture Depreciation (months)	18
Yield	90.0%
Test Cell Cost/hour	$100

Figure 8.15: Cost of Test Reduction realized by traditional cost reduction techniques

If one considers the effect on total manufacturing costs, including the cost to scrap devices that are actually good, the cost savings due to improved yield becomes far more significant.

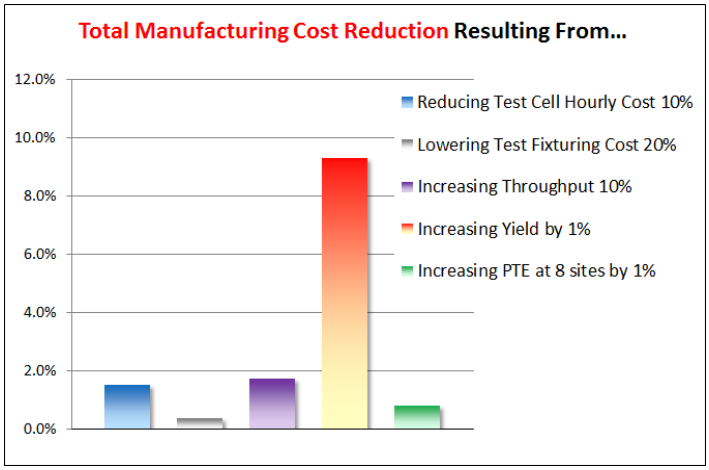


Figure 8.16: Total Cost of Manufacturing Reduction realized by traditional cost reduction techniques.

The risk of yield loss is increasing over time for several reasons:

- Trends such as the reduction of power supply voltages and more complex RF modulation standards will drive higher accuracy requirements for test equipment. Test equipment accuracy is typically added as a "guardband" in testing, reducing the range of acceptable measurements. If measured DC and AC values become smaller and there is no improvement in test accuracy, this guardband will cause more marginal (but good) devices to be scrapped.
- As noted earlier, many devices, especially for mobile applications, require some sort of calibration or trim during the test process to improve DC and AC accuracy. This need dramatically increases both the number of measurements made and the accuracy required of the test equipment. These requirements increase the chance of discarding devices that would otherwise have been good.
- Faster production ramps and short IC product life cycles will reduce the amount of time available to optimize measurements for most devices produced.

Of course, the danger resulting from recovering marginal devices to improve yield is that there may be a greater chance of the device failing in the end application. While test costs for complex devices are lower than silicon and packaging costs, the cost of a failing device in an end product easily swamps out both. Striking the balance between yield and device quality has been the challenge of semiconductor test since the beginning. Optimizing for both can only be achieved through better test accuracy or greater test time, both of which drive up test costs.

The remainder of this section will examine Costs associated with owning and operating test equipment. It must be stressed that reducing these costs must be done in the context of the overall cost of producing devices and to balance reduction in test costs with potential reductions in product yield.

8.12.3 Test Cost Models and Cost Improvement Techniques

The cost of semiconductor test has many drivers, which is further complicated for multi-die packages as shown in Figure 5.

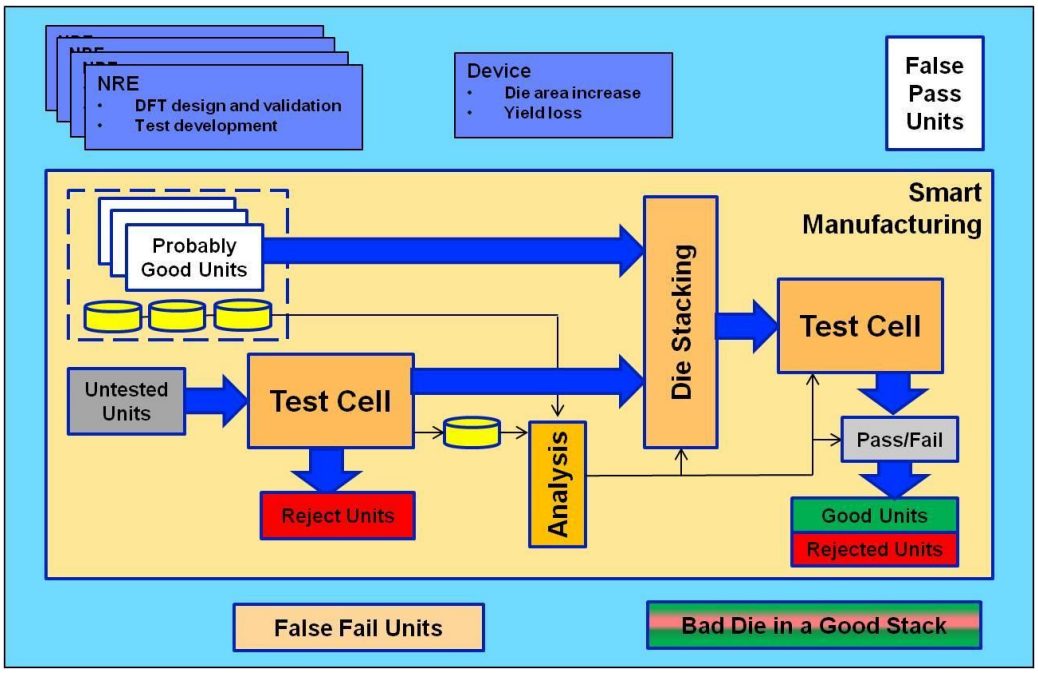


Figure 8.17: Multi-die Flow

8.12.4 Current Top Cost Drivers

The traditional drivers of test costs typically include (in rough order of impact to cost):

- Device yield
- Test time, site count and Parallel Test Efficiency (PTE)
- Overall equipment utilization
- ATE capital and interface expenditures
- Facility/labor costs
- Cost of test program development
- Cost of die space used for test-only functions

8.12.5 Future Cost Drivers

- Increased test time due to additional scan and functional testing
- Increased testing at wafer to produce Known Good Die (KGD)
- Addition of system-level testing to augment traditional ATE test
- Increased cost of handling equipment to support high site count or singulated die
- Increasing use of device calibration/trimming at test or device repair with redundant components

8.12.6 Cost Reduction Techniques

- Multi-site and reduced pin-count
- Structural test and scan
- Compression/BIST/DFT and BOST
- Yield learning and adaptive test
- Concurrent test
- Improvements to test processes based on analysis of collected test data

Multi-site Trend

The simplest way to reduce cost of test is increasing the number of sites. The effectiveness of increasing the number of sites is limited by (1) a high interface cost, (2) a high channel and/or power cost, and (3) a low multi-site efficiency M:

$$M = 1 - \frac{(T_N - T_1)}{(N-1)T_1}$$

where N is the number of devices tested in parallel (N>1), T_1 is the test-time for testing one device, and T_N is the test time for testing N devices in parallel. For example, a device with a test time T_1 of 10 seconds tested using N=32 sites in T_N =16 seconds has a multi-site efficiency of 98.06%. Hence, for each additional device tested in parallel there is an overhead of (1-M) = 1.94%.

There are cases where increased site count is either not possible or not effective:

- Site count is limited by equipment capability. Additional site count required additional test resources and new prober and handler capability that may either not exist or be prohibitively expensive to use as compared to existing equipment that is already depreciated.
- The test time overhead of adding sites will, at some point, begin to reverse the gains achieved by going to higher site count. This situation is discussed below.
- At wafer probe, Touch-down efficiency (TDE) is limited by the size of the die relative to the size of the wafer. Those details are discussed below.
- Additional site count is most effective for high volume devices, which will efficiently occupy test equipment over long periods of time. For lower volume devices, the down time to reconfigure test equipment between different device types will eliminate any gains made as a result of higher site count.

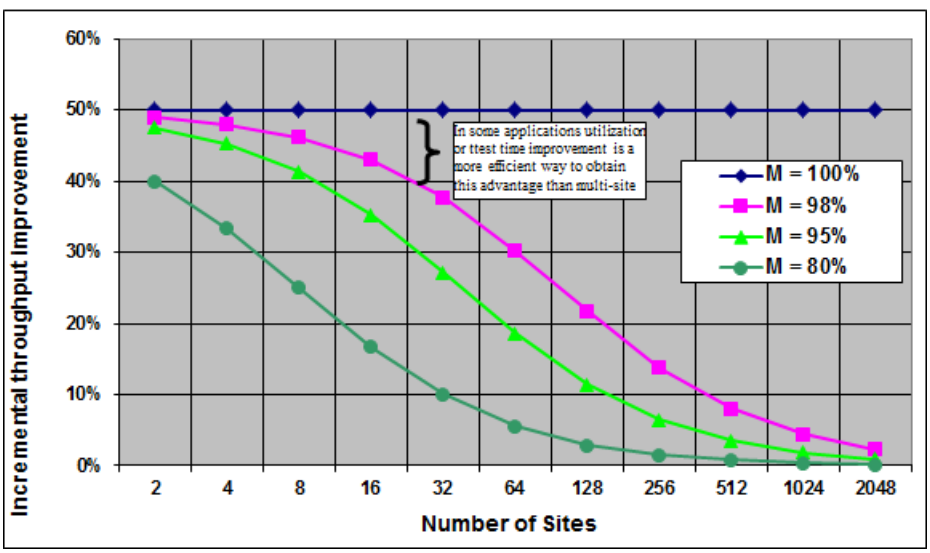


Figure 8.18: Importance of Multi-Site Efficiency in Massive Parallel Test

As one continues to increase the number of sites, a low multi-site efficiency has a larger impact on the cost of test. For example, 98% efficiency is adequate for testing two and four sites. However, much

higher efficiency is needed for testing 32 sites. At 98% efficiency, going from testing a single site to testing four sites will increase a 10 second test time to 10.8 seconds. However, going from testing a single site to testing 32 sites will increase a 10 second test time to 16.4 seconds, significantly reducing the potential advantage of multi-site as shown in Figure 6.

Touch-Down Efficiency (TDE) is defined as the number of wafer touch-downs required to test all devices on a wafer, relative to the theoretical minimum. TDE is influenced for the most part by the die size (and therefore the number of die per wafer) and the pattern used to probe. For example, if a device is tested 10 sites at a time, and there are 1,000 die per wafer, then ideally a probe card would have to touch down 100 times to test the wafer and be 100% efficient. If, due to the mismatch between the round shape of the wafer and the linear or rectangular pattern of the probe card, the probe card must touch down 110 times to test the 1,000 devices, then the TDE is closer to 90%. This result is illustrated in the figures below.

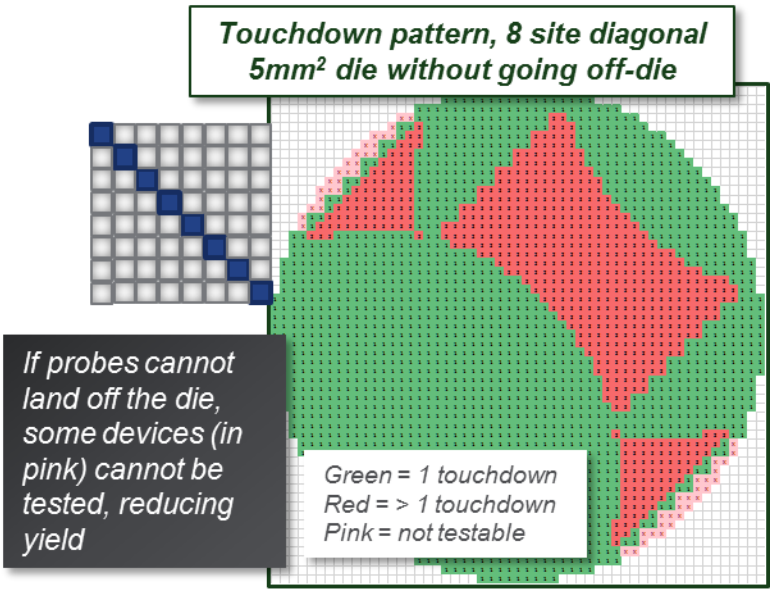


Figure 8.19: Probe Pattern of 5mm2 die using 8-site probe pattern

As die size of a complex device increases, the TDE will continue to degrade as shown in Figure 8. This degradation of efficiency will negate any advantages of increased site count and will eventually increase the cost of test as shown in the example below. In this case, there are gaps in the probe pattern to allow for the inclusion of electrical components on the probe card required for the proper operation of the device under test.

TDE inefficiencies will primarily be addressed by the development of singulated die testing technology. There is significant work underway to allow die to be reassembled in silicon panels that have a rectangular shape as opposed to the round shape of the original silicon wafer. The deployment of this technology will re-start the increase in site count at probe that is currently stalled due to interface costs and TDE limitations.

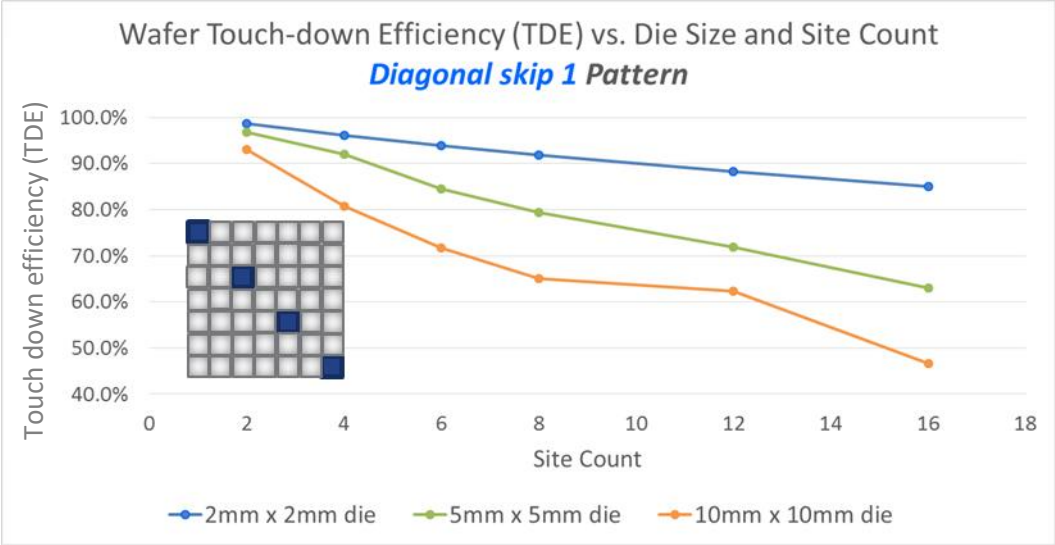


Figure 8.20: Touch-Down Efficiency as function of die size using a 4-site probe pattern

Because of the TDE inefficiencies of probe, more test could be deferred until devices are packaged (where there is no touch-down penalty). The overall trend, however, is to do most testing at probe because:

- Package costs, especially for more complex devices, are significant enough that the cost of discarding a package because of the bad die is greater than the cost of doing more test at probe.
- Multi-die packaging requires known-good die in order to be cost efficient since the cost of discarding good die because of one failing die is never acceptable.
- Devices are used in some form of chip-scale packaging, where traditional package handling equipment cannot be used.

8.12.7 Summary

Major conclusions are:

- Cost of test has been declining for some time, but the rate of reduction has slowed and will remain flat in terms of test costs per device.
- Major reasons for the slower rate of cost reduction are:
  - Packaging trends that drive more test at the wafer probe insertion where site counts are lower.
  - Increased cost of consumable material, which now dominates tester capital cost in terms of test cell costs.
  - Desire for higher yield, which has a much larger impact on overall device production costs than test costs alone.
  - Desire for higher device quality, especially for automotive applications, which necessitates more test.
- Potential solutions to decrease test costs are:
  - New probing technology which allows test of singulated die.
  - New PCB and Interposer technology to lower the cost and complexity of consumable material.

- Improvements to the test process through increased use of data analysis and machine learning based on measured data.
- Factory automation.
- Cost reduction of system-level testing.

Chapter 9: **Advanced Packaging Supply Chain for High Performance Computing**

Contents

Chapter 9: **Advanced Packaging Supply Chain for High Performance Computing** 1
9.1 Executive Summary 1
9.2 Background and Secular Trends (with reference to HIR Packaging Roadmap) 4
9.3 Objectives 5
9.4 Supply Chain Opportunities & Strategies 5
9.4.1 Volatility in demand/cost 5
9.4.2 Proximity of Innovation capability 8
9.5 Supply Chain SWOT 9
9.6 Opportunities 10
9.7 Threats 10
9.8 Conclusion 11
9.9 References 11

9.1 Executive Summary

The rapidly changing geo-politics, the occurrence of a devastating pandemic, and the increased probability of extreme weather conditions due to climate change, have brought into sharp focus the need for more resilient (not just efficient) supply chains in several industries. The semiconductor industry has perhaps one of the most complex and globalized supply chain networks of any industry. Fortunately for the semiconductor supply chain, the USA has significant if not dominant positions across most of the value layers of - circuit design (EDA software), leading edge front-end device manufacturing, manufacturing equipment, and materials & chemicals. However, one link of the value layer – chip packaging (assembly and test) - has traditionally been outsourced to low-cost regions and as a result the supply chain related to this value step has faced pressure to localize outside of the USA.

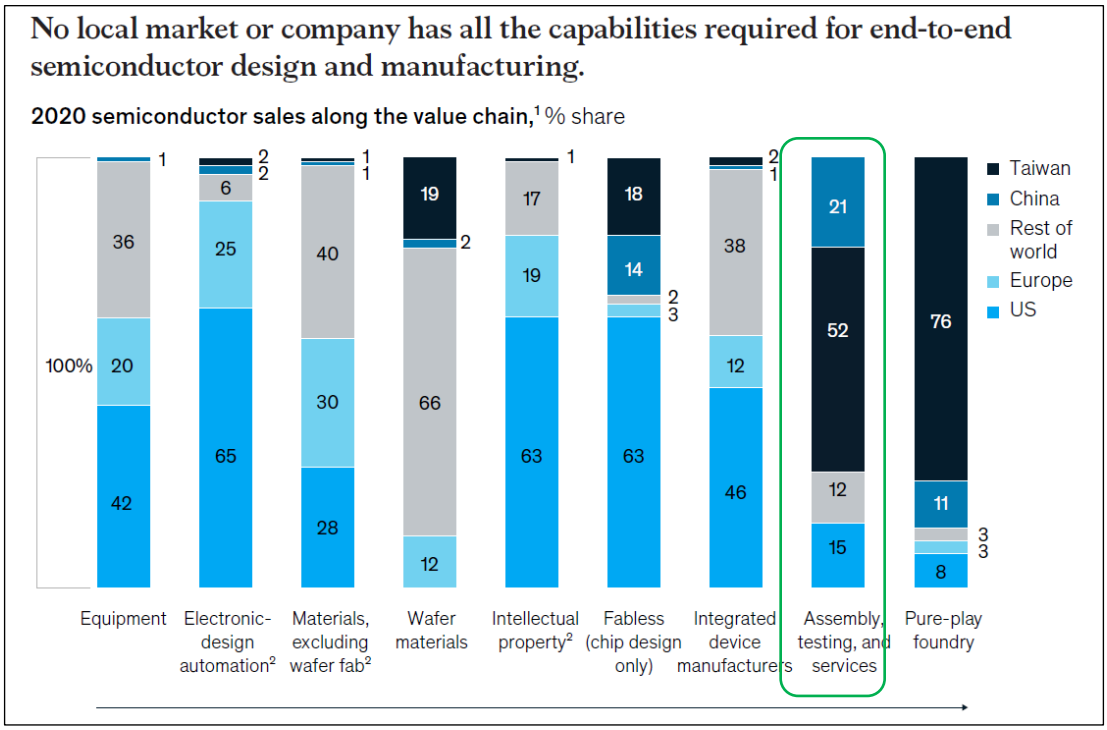
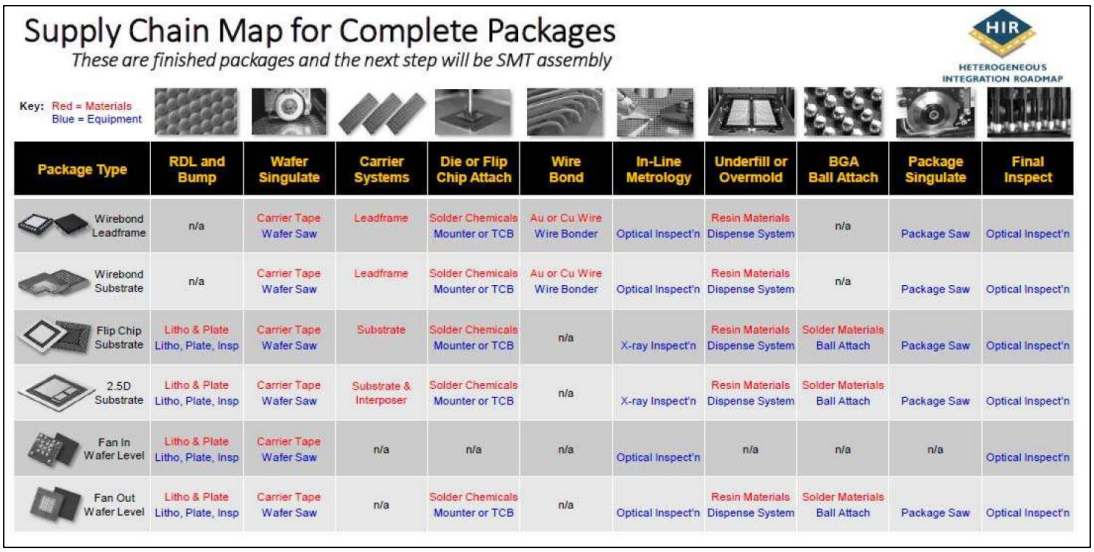


Figure 9.1: Packaging (Included in Assembly and Test) has had weak US presence amongst the various semiconductor value chain links. Several investments have been announced to increase foundry capacity in the US **[7]**. Source: Ondrej Burkacky, Marc de Jong, and Julia Dragon, "Strategies to lead in the semiconductor world," (London: McKinsey & Company, April 15, 2022)

The primary tailwinds to out-source chip legacy packaging technology Figure 9.1 & Figure 9.2 shows value chain share % by region) have been: (1) lower labor costs that are favorable as to the several manual steps that remain, (2) relatively mature, comparatively lower technology, manual, standardized manufacturing processes that benefit from scale/consolidation, (3) proximity to assembly of consumer electronics gadgets providing logistical benefits, and (4) national industrial policies providing various financial incentives and tax advantages.



Supply Chain Map for Complete Packages

These are finished packages and the next step will be SMT assembly

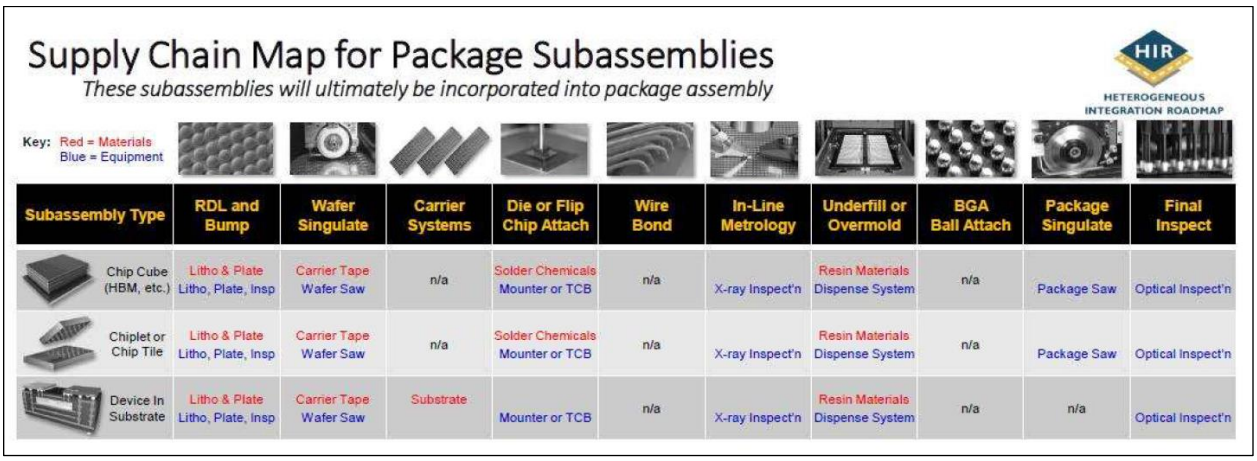
HIR HETEROGENEOUS INTEGRATION ROADMAP

Key: Red = Materials
Blue = Equipment

Package Type	RDL and Bump	Wafer Singulate	Carrier Systems	Die or Flip Chip Attach	Wire Bond	In-Line Metrology	Underfill or Overmold	BGA Ball Attach	Package Singulate	Final Inspect
Wirebond Leadframe	n/a	Carrier Tape Wafer Saw	Leadframe	Solder Chemicals Mounter or TCB	Au or Cu Wire Wire Bonder	Optical Inspect'n	Resin Materials Dispense System	n/a	Package Saw	Optical Inspect'n
Wirebond Substrate	n/a	Carrier Tape Wafer Saw	Leadframe	Solder Chemicals Mounter or TCB	Au or Cu Wire Wire Bonder	Optical Inspect'n	Resin Materials Dispense System	n/a	Package Saw	Optical Inspect'n
Flip Chip Substrate	Litho & Plate Litho, Plate, Insp	Carrier Tape Wafer Saw	Substrate	Solder Chemicals Mounter or TCB	n/a	X-ray Inspect'n	Resin Materials Dispense System	Solder Materials Ball Attach	Package Saw	Optical Inspect'n
2.5D Substrate	Litho & Plate Litho, Plate, Insp	Carrier Tape Wafer Saw	Substrate & Interposer	Solder Chemicals Mounter or TCB	n/a	X-ray Inspect'n	Resin Materials Dispense System	Solder Materials Ball Attach	Package Saw	Optical Inspect'n
Fan In Wafer Level	Litho & Plate Litho, Plate, Insp	Carrier Tape Wafer Saw	n/a	n/a	n/a	Optical Inspect'n	n/a	n/a	n/a	Optical Inspect'n
Fan Out Wafer Level	Litho & Plate Litho, Plate, Insp	Carrier Tape Wafer Saw	n/a	Solder Chemicals Mounter or TCB	n/a	Optical Inspect'n	Resin Materials Dispense System	Solder Materials Ball Attach	Package Saw	Optical Inspect'n

Figure 9.2: Legacy Packaging Technology

These tailwinds coincide with the semiconductor manufacturing pivoting to meet the enormous market demand for mobile computing and communication in the first two decades of the twenty-first century enabled by devices such as laptop/tablet computers and mobile phones which have been accelerated by waves of higher bandwidth wireless connectivity technology. Packaging technology is now navigating multiple inflections as highlighted in Figure 9.3 in the sub-assembly roadmap. These inflections are occurring due to the simultaneous slowing of Moore's Law [2] and the needs for high bandwidth and energy efficient interconnects between various compute functions such as CPU, GPU, Memory, and Specialized ASICs that need strong coupling in High Performance Computing applications such as AI, autonomous driving, and AR/VR [3].



Supply Chain Map for Package Subassemblies

These subassemblies will ultimately be incorporated into package assembly

HIR HETEROGENEOUS INTEGRATION ROADMAP

Key: Red = Materials
Blue = Equipment

Subassembly Type	RDL and Bump	Wafer Singulate	Carrier Systems	Die or Flip Chip Attach	Wire Bond	In-Line Metrology	Underfill or Overmold	BGA Ball Attach	Package Singulate	Final Inspect
Chip Cube (HBM, etc.)	Litho & Plate Litho, Plate, Insp	Carrier Tape Wafer Saw	n/a	Solder Chemicals Mounter or TCB	n/a	X-ray Inspect'n	Resin Materials Dispense System	n/a	Package Saw	Optical Inspect'n
Chiplet or Chip Tile	Litho & Plate Litho, Plate, Insp	Carrier Tape Wafer Saw	n/a	Solder Chemicals Mounter or TCB	n/a	X-ray Inspect'n	Resin Materials Dispense System	n/a	Package Saw	Optical Inspect'n
Device In Substrate	Litho & Plate Litho, Plate, Insp	Carrier Tape Wafer Saw	Substrate	Mounter or TCB	n/a	X-ray Inspect'n	Resin Materials Dispense System	n/a	n/a	Optical Inspect'n

Figure 9.3: Inflections in Advanced Packaging Technology occurring in sub-assemblies

These inflections in packaging sub-assembly technology offer a serendipitous opportunity to secure the packaging value layer related supply chain for the USA, especially for high performance computing (HPC), AI and other technology intensive medical devices. Not exploiting these inflection opportunities

to onshore and secure packaging supply chains for the USA, would not only endanger its leading position in technology and defense capability, it may also lead to a permanent off-shoring of R&D for emerging technologies such as advanced packaging. This would be a devastating loss for the US semiconductor industry as the manufacturing base for such technologies would be permanently established overseas. On the contrary, these new technologies that are presently necessary for high-performance computing will eventually find its way to other device applications, such as auto, battery tech, etc. and every effort must be made that once advanced packaging is secured in the USA, the supply chain for the it is scaled to meet all tiers and applications for the future.

9.2 Background and Secular Trends (with reference to HIR Packaging Roadmap)

Chip packaging is experiencing tremendous innovation and flux in response to the enormous rising costs of advanced transistor nodes [4] due to the enormous costs of lithography (EUV), increasing complexity of devices (planar, to FinFET, to GAA), and sophisticated materials engineering to preserve device performance and yield. This is reversing the decades long historical trend where more functionality was integrated onto the same monolithic chip. Especially for high performance computing, the chip manufacturing techniques are now so dedicated and specialized (DRAM, GPU, CPU, ASICs) that instead of monolithic integration, integration is achieved by integrating “chiplets or tiles” into one chip package that preserves performance (speed and energy efficiency). This integration into one package is where there is great innovation with multiple architectures [5] and scaling both dimensionally (2D, 2.5D, 3D) and pitch of interconnects - reminiscent of scaling of the front-end device in the past half century.

Present Supply Chains are mature and complex: elongated, multiple links and interfaces. This was needed to specialize, standardize and scale to drive down cost, consistent with the previous paradigm of the past half century that advanced nodes were more consistently cost effective for multiple digital electronic products.

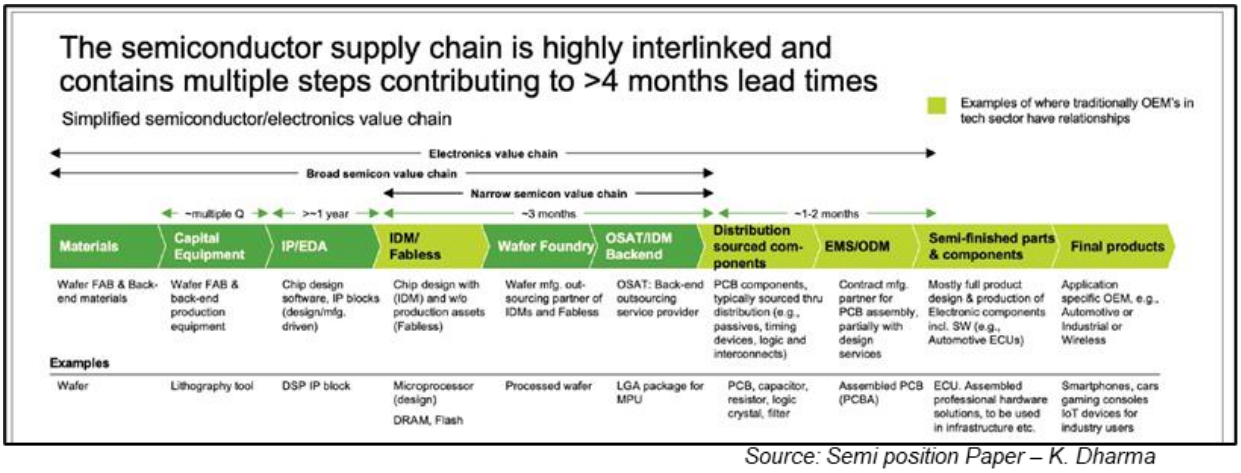


Figure 9.4: Elongated & highly interlinked semiconductor supply chain

9.3 Objectives

1. Identify key inflections and timing that require Supply Chain innovation
    a. Supply Chain Map refresh - The Chips Act funding for the construction of new advanced packaging facilities and new R&D centers will drive development and launch of new Advanced packaging technology and equipment. Market demand and product innovation from fabless companies & IDMs will also present major inflection points. Both product roadmap refreshes should be closely watched for their impact on Supply chain refresh.
2. Identify key Supply Chain Structural changes needed to enable execution of the roadmap
    a. Vertical integration of technologies: Some advanced packaging processes are getting done in the back-end-of-line in the fab on foundry based equipment. Other subsequent processes are being done in back-end packaging facilities. A vertical integration of these processes will make the processes more efficient & economical. [6]
    b. New state of high flux & supply chain realignment - Geo-political tensions and nations' policies are causing huge changes in the traditional semiconductor supply chain. Countries including the US are focusing on building domestic self-contained supply chains, or at best nearshoring or friend-shoring to friendly countries/locales. This may give rise to Toyota-like manufacturing clusters in the US, where IDMs or foundries co-locate their critical suppliers in their physical vicinity.
    c. Rapid cycles of development: New products in both high performance and medical devices are accelerating innovation and development cycles. AI and lately Gen AI is driving the need for high bandwidth memory (HBM) products. Similarly, medical devices with different form factors such as wristbands, monitors, glasses and are driving flexible product development. Both these examples present a large paradigm shift for Supply Chain, as well a huge expansion in product mix (permutations/combinations of devices).

9.4 Supply Chain Opportunities & Strategies

9.4.1 Volatility in demand/cost

9.4.1.1 Domestic/Co-located facilities lead to better planning & control

Domestic OSATs, co-located with foundries or IDMs, or located in the same geographical region as them can better coordinate both the planning and execution process. OEMs can now have a more integrated monthly planning cycle, with better data visibility and fewer blind spots. Any unexpected events, like machine breakdowns, factory shutdowns can be addressed in near real times. Moreover, currently OEMs have limited supply chain visibility into their outsourced remote factories. Many IDMs & foundries have tried to build 'band-aid' solutions to track detailed events like lot starts & lot finishes to get near real-time visibility into the production cycle. However, they are failing and in some cases have deployed a large operational team to manually capture this information via phone & video calls. They lack control over outsourced manufacturing operations, which at times had led to cannibalization of parts on assembly lines to fulfill severely delayed machines. IDMs & foundries are anxious to develop close operational relationships with domestic OSATs. It would be very prudent to exploit this business opportunity.

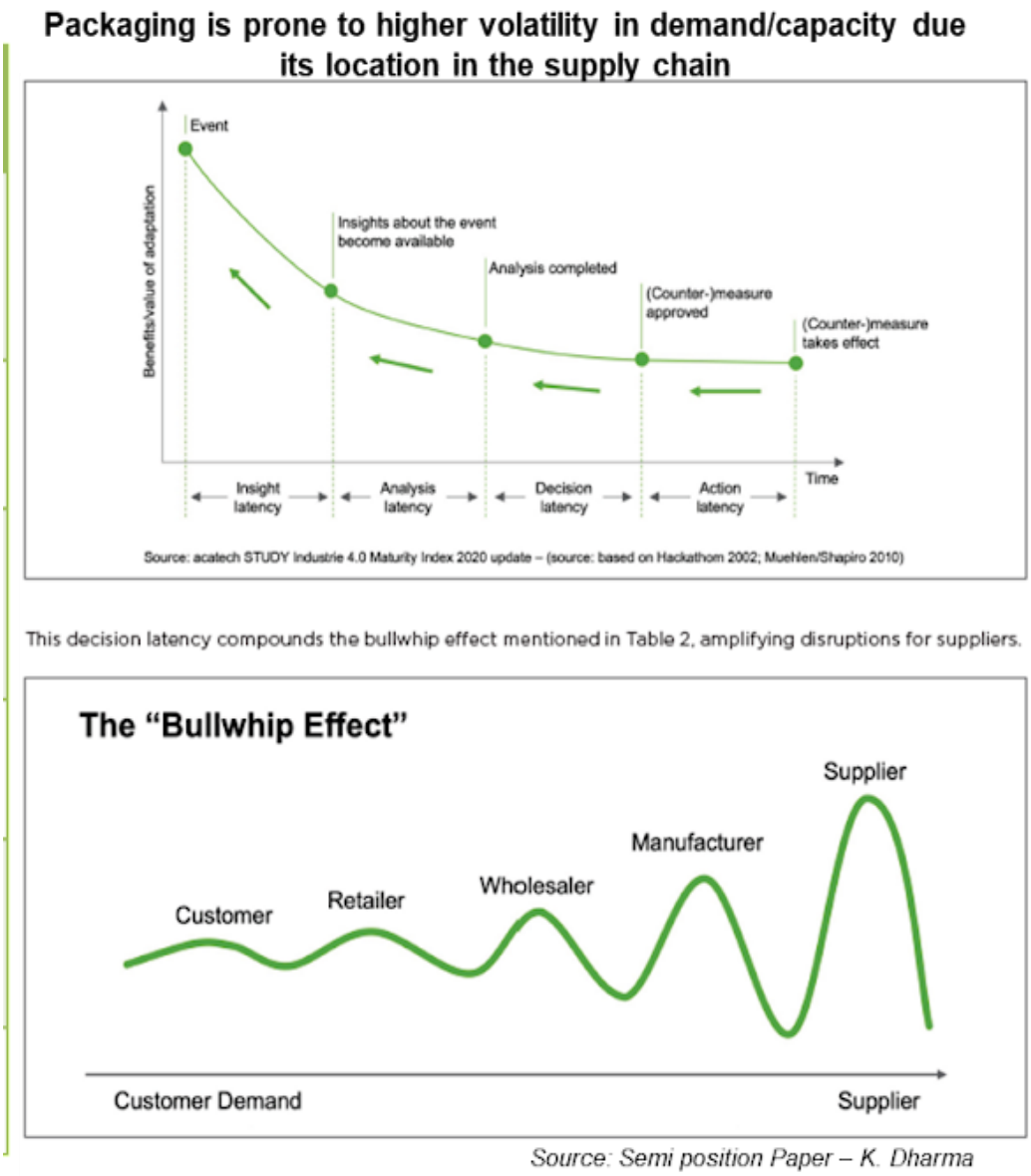


Figure 9.5: How volatility & 'Bullwhip Effect' gets amplified for ATP (Assembly, Test, Packaging)

9.4.1.2 Supply Chain Resiliency

In the post pandemic scenario, with increased geo-political tensions, supply chain resiliency has become a huge issue for IDMs, foundries and critical downstream customers like semi equipment manufacturers. IC manufacturers have little advance knowledge when these global supply chains may get snapped due to political, logistical reasons, natural causes or financial sanctions. Companies are deploying complex solutions to map global inventory flow, multi-tier supplier networks spread in multiple geographies. For many of these Tier 2 & 3 suppliers located in information opaque areas, no information is available about their financial stability or information security, for example. Such suppliers can shut down due to financial or other reasons, such as hacking, with the OEM having little prior information or control. Manufacturers are racing to map their entire supply networks, capture real time logistics data and develop predictive

AI models to predict any future disruptions. A disruption in the commercial sensitive and/or defense sector can lead to major issues. Domestically located OSATs can minimize these risks, commercial US financial monitoring mechanisms, information security tools and technologies can provide a much higher level of security and supply chain resiliency. A fully or largely domestic network of semiconductor backend suppliers will reduce external geo-political risks as well as logistics vulnerability in the extended supply chain for all customers.

a. Long term commitments
   In response to the silicon shortage during the pandemic and as an insurance against future disruptions, Auto OEMs have entered into long term, multi-year contracts with foundries and chip manufacturers. This provides stability to both sides, assuring foundries and IDMs of stable longer term demand, protecting them against bullwhip effect while assuring timely IC delivery to Auto OEMs.

b. Limited US packaging
   With few exceptions, most US IDMs and foundries have no packaging capacity in the US and largely depend on Asian (China & Taiwan based OSATs). Given the current situation, they appear keen to enter into longer term contracts with US based OSATs and help them grow. Such long term commitments between US based IDMs & foundries can lead to capacity expansions, long term partnerships, technical cooperation between both sides. This can be a strong catalyst for re-shoring of semiconductor packaging, especially advanced packaging.

c. Financial Investments
   i. Co-investment: New advanced manufacturing startups as well as traditional OSATs expanding into advanced packaging can also seek co-investment from US IDMs and foundries.. This would be very important given the high cost of establishing highly automated advanced packaging facilities, which are extremely capacity constrained in the short term. Such long term contracts would allow US OSATs to raise capital and invest in building expensive advanced packaging facilities, while providing critical advanced packaging capacity to IDMs & foundries.

      This model can be similar to the co-investment model at ASML where Intel, TSMC and Samsung as co-investors[8] provide critical long term financial commitment to the lithography equipment provider. Another example would be Arm Technologies, where leading customers like Apple and Samsung plan to purchase shares in the Arm IPO, providing it long term commitment [9].

   ii. Subscription/Pay-per-use model: Another investment mechanism would be a pay-per-use model where packaging equipment providers may install high cost, automated equipment at the OSAT sites and would be reimbursed based on equipment usage. This usage fee would capture installation, operations & maintenance expenses borne by the equipment providers.

   iii. PE/VC investment: OSATs can also invite investments from Private Equity or venture capitalist firms, similar to the investment at Intel by Brookfield Asset Management. This would allow private investments into an emerging, high technology, high risk area by professional investors who are very familiar with making such investments.

d. Funding Avenues:

New funding opportunities can be used to establish advanced packaging facilities that can help smoothen demand volatility. The current economic situation has presented several funding sources which are highlighted below.

Chips Act is offering new funding opportunities for traditional and advanced packaging industries to reshore and establish manufacturing facilities. Under the Chips Act, the US Government has allocated $39B for establishing semiconductor manufacturing & packaging facilities. Current companies and startups can apply for funding through the Notice of Funding Opportunities (NOFO), the Chips Program Office. Additionally, the National Advanced Packaging Manufacturing Program (NAPMP) of the Chips Act, $2.5B has been allocated for R&D into Advanced Packaging [10].

The Department of Defense has also been working closely with industry partners on the The State-of-the-Art Heterogeneous Integrated Packaging (SHIP) Program for advanced packaging – SHIP RF led by Qorvo and SHIP Digital headed by Intel. DoD has allocated $560M for custom and dual-use packaging technology, with a strong focus on 3D heterogeneous integration. Out of this, $380M is targeted for developing dual-use technology ecosystems and is expected to grow till 2027. Additional DoD funding is also available through the Accelerate the Procurement and Fielding of Innovative Technologies (APFIT) program [11].

State governments are also offering additional incentives and investments for starting packaging facilities and developing an advanced packaging ecosystem. Primary examples are the State of Kansas which provided more than $300M in incentives to Integra Technologies [12] to locate a packaging plant close to Wichita State University. Similarly, the State of Texas has provided two rounds of funding totalling more than $600M to fund the Texas Institute of Electronics develop an advanced packaging eco system in Austin, TX at the old Sematech fab facility.

There are a number of current and emerging investment options as well as incubators that can also help startup companies fill in business and product gaps in the US. Under the Chips Act, the Federal Government is setting up an investment fund that may have other co-investors, such as venture capitalists, private equity funds, and would help fund startup ideas. There are also silicon focused incubators, such as Silicon Catalyst in the SF bay area, that have experienced advisors to guide a startup and a strong relationship with foundries to quickly fabricate initial designs. In addition, several academic institutions are also setting up incubators to enable rapid prototyping and spin out startups that drive product innovation to fill in supply chain gaps.

9.4.2 Proximity of Innovation capability

e. Advanced Packaging research

The Chips Act and NAPMP are investing $2.5B over the next 5 years in advanced packaging R&D. This would be used to fund AP/HI research in corporate R&D centers and developing academic research centers such as in University of Texas at Austin, Penn State, Georgia Tech and others. The academic R&D centers also have strong support from fabless companies, IDMs & foundries with active plans to develop & support a manufacturing ecosystem in their vicinity. Startups & OSAT companies can take advantage of this option combined with federal and state local incentives to develop & expand a manufacturing base. The ongoing engagement with the R&D centers and the active support of large corporate customers should provide strong technical & business support for these companies.

f. Workforce development
Locating OSAT manufacturing facilities close to academic campuses can also become a key driver for workforce development. Students from the university, local community colleges could also be hired & trained as process engineers, technicians and line workers. This ecosystem would become very powerful, if corporate R&D centers were located close to academic campuses, as this triad of OSAT manufacturing facilities, corporate R&D labs and academic campuses would offer a wide selection of research & operational job opportunities.

g. Technology Cooperation & transfer
US foundries & IDMs have developed advanced packaging technologies, internally but have limited in-house or domestic partner facilities. However, they are keen on domestic partnership options for a number of reasons, which include - national security, possible captive manufacturing facility and the current political and economic move to move away from China. Under these scenarios, smaller domestic OSATs can partner with US IDMs & foundries to develop a close technical and business relationship.

h. US based EDA & Semi equipment companies
Most semi equipment manufacturers design and build their most advanced equipment in the US. With the movement of advanced packaging into the back end of the fab, lithography, etch & chemical mechanical polish (CMP) are becoming key manufacturing operations for Advanced packaging operations like CoWoS, hybrid bonding. Most technology development for advanced packaging is also happening in the US, funded largely by the Chips Act as well as investment by US companies. Hence to capture and retain the business advantage, advanced packaging facilities should be located in the US. In addition, both the major EDA vendors, Synopsys & Cadence, are US based and are already incorporating advanced packaging functionalities in their design tools. Therefore by building and incorporating the advanced packaging piece, the end-to-end supply chain can be securely located in the United States.

9.5 Supply Chain SWOT

2. Strength
   a. Manufacturing & Research parks
   As discussed above private companies have been leading the research in this area and have formed close partnerships with academic institutions across the US. Some of these partnerships are evolving into manufacturing clusters with OSATs located in the vicinity in a manufacturing park. These multiple clusters or developing centers of excellence will spawn a lot of technological, entrepreneurial & workforce development.

   b. Semi equipment & EDA companies
   Nearly all of the semi equipment companies that manufacture front end equipment that is also used in advanced packaging are based in the US. Other smaller backend equipment manufacturers are based in friendly countries like Israel, Japan & Singapore. Both the major EDA companies are also US based. Therefore, it would be easy to build an end-to-end, design-to-test supply chain in the US and friendly shored countries [13].

3. Weaknesses
   a. Packaging Capacity constraints

Some of the leading foundries manufacturing HPC chips have reported major contractions in their advanced packaging capacity and have reported they are completely tapped out in their current installed capacity. Some of the HPC capacity is being cannibalized to meet the huge demand for AI related chips. These foundries are now ordering extra advanced packaging equipment to meet the growing demand from AI & HPC customers.

b. Technology Development
Some of the technology required for advanced packaging, such as automation, water cooling, is still under development in labs. Moreover, other countries in competition may already have been working on them and have advanced solutions. It would take some time for US based research institutions - academic and corporate - to develop these technologies.

c. Advanced Packaging equipment
Equipment suppliers for advanced packaging are also facing a backlog for supplying these equipment. For the fab equipment suppliers, the advanced packaging segment is a small portion of their overall market of tools and therefore receives less attention. The advanced equipment suppliers are relatively smaller and geographically distributed, and therefore are finding it hard to cope up with the increased demand.

9.6 Opportunities

d. Funding
Both the US Government, through the Chips Act, as well as State and local governments are providing unprecedented levels of funding. For example, the State of Texas recently allocated $600M for the Texas Institute of Electronics, with more funding expected from the Federal government. Other states like New York, Arizona, etc are also providing funding in various forms.

e. Nearshoring & Friend-shoring
Advanced packaging offers an excellent opportunity to locate these facilities in nearshore locations, like Mexico, Costa Rica or Puerto Rico, all of which already have electronics or semiconductor manufacturing bases. This would offer trained talent and a lower labor cost basis for US semiconductor companies, given automation for advanced packaging is still under development. Similarly, such operations can also be promoted in friendly countries like Singapore, Malaysia, that would build an end-to-end semiconductor supply chain in a friendly, secure environment.

9.7 Threats

f. Workforce
As highlighted above, shortage of trained workforce, especially in advanced packaging is a big challenge. Universities, community colleges, companies and state employment development departments are all working together to meet this challenge. However, it would still take a few years to train technically advanced personnel like process engineers. This shortfall can be a big challenge to develop and expand advanced packaging in the US.

g. Competition
Companies in other parts of the world, like China, have been doing a lot of research in Advanced Packaging and have also built manufacturing facilities to meet demand in this space. A lot of US & western companies still have strong manufacturing presence and relationships in China and sell

into the Chinese market. They may find it easier to setup and expand their advanced packaging facilities with Chinese companies.

9.8 Conclusion

In conclusion, advanced packaging offers a once-in-a-lifetime opportunity to the US semiconductor industry to re-gain the advantage that was lost to other parts of the world with labor costs. With the corporate and government support as tailwinds, the industry can develop advanced, highly automated technologies that are located in the geographical proximity of US IDMs and foundries. This would also provide a high degree of security for devices being packaged for security, technologically sensitive applications. Moreover, this would act as a large technology moat for any future product development. Other benefits include eliminating the 'bullwhip' effect and bringing a high level of transparency to semiconductor supply chains. Therefore we should move ahead with developing a robust advanced packaging network in the US or friendly nations and enable an end-to-end semiconductor supply chain, largely free of geo-political and logistic issues.

9.9 References

[1] 'Semiconductors and Semiconductor Industry', Congressional Research Service, R47508, April 19, 2023

[2] 'We are not prepared for the end of Moore's law', David Rothman, MIT Technology Review, February 24, 2020

[3] 'How the Slowdown of Moore's Law Has Fueled the Rise of Computational Storage', Hao Zhong, Spiceworks, June 22, 2021

[4] 'What will that chip cost', Brian Bailey, Semiconductor Engineering, Oct 2023.

[5] 'Advanced Packaging: Enabling Moore's Law's Next Frontier through Heterogeneous Integration', Raja Swaminathan, AMD, HotChips 2021.

[6] 'Status of Advanced Packaging Industry, 2021', Santosh Kumar, et al, Yole Group, 2021.

[7] TSMC's front and back-end fabs being located in the same science parks in Taiwan: "TSMC Fabs," TSMC, https://www.tsmc.com/english/aboutTSMC/TSMC_Fabs.

[8] Samsung joins ASML's Customer Co-Investment Program for Innovation, completing the program, August 2012. https://www.asml.com/en/news/press-releases/2012/samsung-joins-asmls-customer-co-investment-program-for-innovation-completing-the-program

[9] Apple, Samsung to invest in Arm as it eyes September IPO, August 2023. Apple, Samsung to invest in Arm as it eyes September IPO - Nikkei Asia

[10] Frequently Asked Questions: CHIPS Act of 2022 Provisions and Implementation, Congressional Research Reports, R47523 pg 9-10. https://crsreports.congress.gov/product/pdf/R/R47523

[11] 'Government CHIPS on the table: How higher DOD microelectronics funding is here to stay', McKinsey article, March 3, 2023.

[12] ‘Semiconductor manufacturer expanding with Kansas taxpayer incentives’. Topeka Capital Journal, February 2, 2023.
https://www.cjonline.com/story/business/economy/2023/02/02/semiconductor-firm-integra-to-be-kansas-second-megaproject-deal/69867169007/

[13] ‘Re-Shoring Advanced Semiconductor Packaging’, John VerWey, Center for Security & Emerging Technology, June 2022, pg 12-15.

Chapter 10: Smart Manufacturing Technology for Heterogeneous Integration & Advanced Packaging

Authors: Benson Chan[1], John Foley[2], & Mark da Silva, Ph.D.[3]

Contents

Chapter 10: Smart Manufacturing Technology for Heterogeneous Integration & Advanced Packaging 1

10.1 Acknowledgements 2

10.2 Introduction 2

10.3 Data Acquisition, Provenance & Governance 5

10.4 Data types/formats – logs, calibration, process, etc. 5

10.5 Data sampling intervals 5

10.6 Traceability 6

10.7 Metrology & Test data 6

10.8 Data Analysis 6

10.8.1 Discovery & Correlations 6

10.9 Yield & Performance Monitoring 7

10.10 Modeling 8

10.10.1 FDC Modeling and Monitoring 8

10.10.2 Digital Twins 8

10.11 Prediction 9

10.11.1 Real-time Process Monitoring 9

10.12 Autonomy 9

10.12.1 "Lights Out" Factory 9

10.12.2 Automatic setup & optimization 10

10.13 Surface Mount Technology 12

10.14 Summary 15

References 16

[1] Associate Director, iMAPS Fellow, IEEE Senior Member; Integrated Electronics Engineering Center (IEEC) Binghamton University

[2] Director Product Management, Ball Bonder Business Unit, Kulicke & Soffa

[3] Sr. Director – Smart Manufacturing Initiative, Advanced Packaging & HI Initiative, SEMI

10.1 Acknowledgements

Special thank you to all working group (TWG) members and especially TWG4 members Benson Chan (Binghamton University), John Foley (Kulick & Soffa) & Mark da Silva (SEMI) who contributed to this chapter. The TWG4 sub-group members would like to acknowledge the many contributions of industry experts for sharing input and perspectives that shaped this chapter. We would also like to acknowledge the contributions of the other TWG members who provided critical feedback. We acknowledge the support from NIST Award (Ref – NIST site) that funded the "Manufacturing Roadmap for Heterogeneous Integration and Electronics Packaging " (MrHIEP)

10.2 Introduction

As feature scaling challenges (due to physics limitations) to develop and commercialize new technology nodes grow, the semiconductor industry is embracing the “More than Moore” paradigm including Heterogeneous Integration with chiplets as a path to increased performance. Multi-die packaged components have evolved over the past decades from multi-chip modules to system-in-package approaches, to interposer-based implementations to today's 2.x/3D ICs. Conventional process flows HI components are shown in Figure 10.1. This chapter focuses on the deployment of Industry 4.0 or Smart Manufacturing tools, technologies, and methods for Heterogeneous Integration. In particular the chapter will focus on HI applications of High-Performance Computing (HPC) and Medical Devices (MD). This chapter will provide roadmap guidance of Smart Manufacturing methods in development and in current production, where the use of artificial intelligence and machine learning is leveraged to improve quality, yield, and reliability at a reduced overall manufacturing cost for HI systems. Smart manufacturing for HI provides a path to re-shoring package manufacturing into the US (and friendly countries) by reducing the dependence on low cost labor currently deployed in off-shore assembly sites. Adoption of Smart Manufacturing techniques and methodologies will reduce the cost of assemblies by reducing the manpower required to run the assembly processes to produce assemblies but will also increase the quality of the components that are made. By ensuring all products are assembled to their required specifications, JEDEC, IPC and others, we can increase first pass yields, reduce rework and scrap as well as improve the reliability of assemblies.

Industry 4.0 as applied to electronics manufacturing, incorporates a wide array of digital technology, and automation, including an enhanced interconnectivity of devices, and related tools deployed with cloud to enable ease of data access and real time analytics. Included in this array of digital technologies are Digital Twins[4], and Machine Learning[5] (ML) operations designed to provide for decision making and self-improvement. When AI/ML are incorporated into specific assembly or process sequences, so called “Smart Manufacturing” operations are created that allow us to work with machines used in manufacturing process operations in new, and highly productive ways. Moreover, because Smart Manufacturing technology typically enables real time manufacturing feedback such as defect detection, and capability for

[4] Digital Twin - A DT is a virtual model designed to accurately reflect a physical object or process.

[5] Machine Learning - Machine learning is a subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed.

real-time corrective action and process optimization, the potential for more efficient, consistent, high quality, high reliability end products are realized at an overall reduced manufacturing cost.

Conventional HI process flows

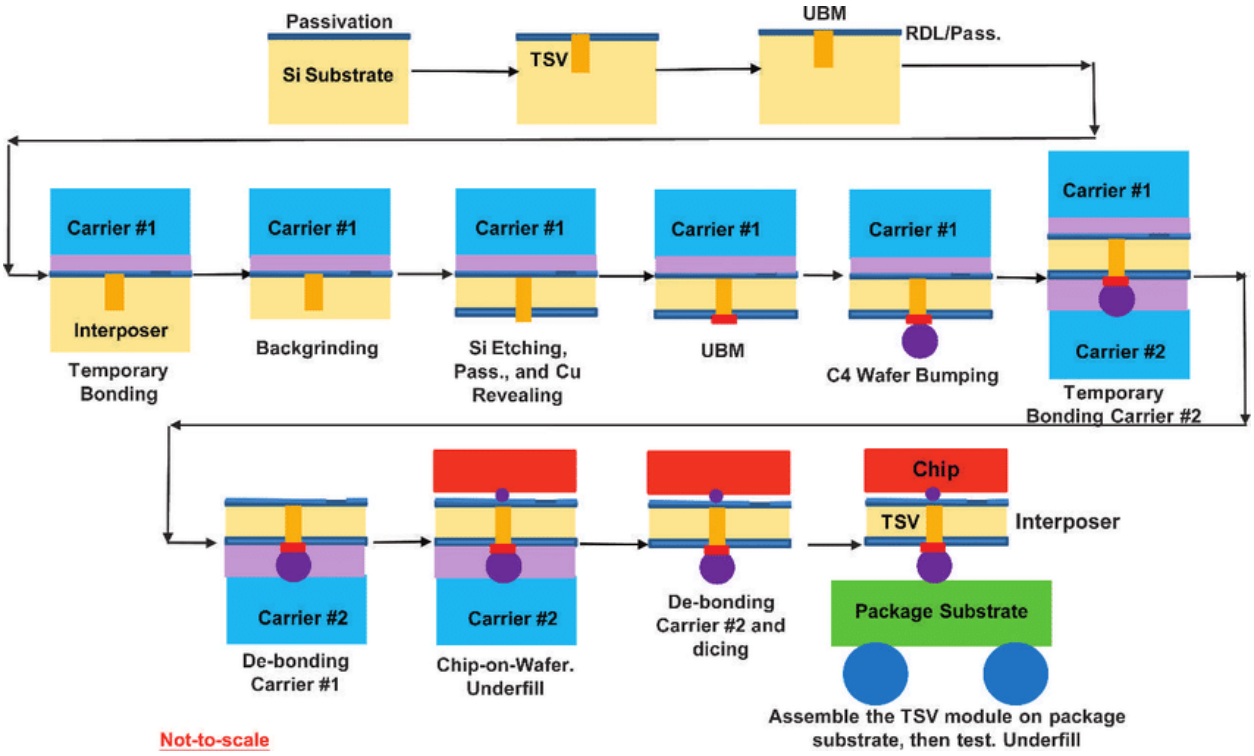


Figure 10.1: Conventional process flow for 2.5D/3D IC integration (chip on interposer wafer on package substrate) - Courtesy Lau et al. (2014) [Ref 2]

As semiconductor companies look to the future of "More than Moore" The examples described in the following sections are systems that are currently being practiced in the various ECATs and captive assembly houses. Real savings can be attributed either through lower yield losses, higher reliability and throughputs and lower cost from direct headcount reduction needed to maintain the process tools. The algorithms and practices of these examples can be applied to any assembly process as long as there is a method to capture real time data of the process variables that would affect the products being produced.

Smart Manufacturing will also play a role in the CHIPS and Science Act and the efforts to re-shore the production of electronics and electronics packaging in the US. Manufacturing products on-shore at a competitive price will require more automation than what is practiced today. This chapter will describe efforts into Industry 4.0 in specific portions of the electronics packaging assembly processes, but to provide the cost advantage of smart manufacturing, the concepts and AI/ML algorithms being developed need to be adopted by all process steps of the assembly process. Smart Manufacturing will not only reduce the cost of the assembly process, but the adaptive process adjustments provided by smart manufacturing will also improve the assembly quality, and reliability by ensuring they are built closer to the center of the specifications that they were designed to be.

SEMI's Smart Manufacturing Initiative's Global Executive Committee (steering body comprised of IDMs, OEM's and solution providers) has released a roadmap vision for back-end assembly that can be adopted, with some modifications, for HI manufacturing as well. Figure 10.2 below shows the Smart Manufacturing Roadmap for Backend assembly broken down into categories that will be discussed in the remainder of this chapter.

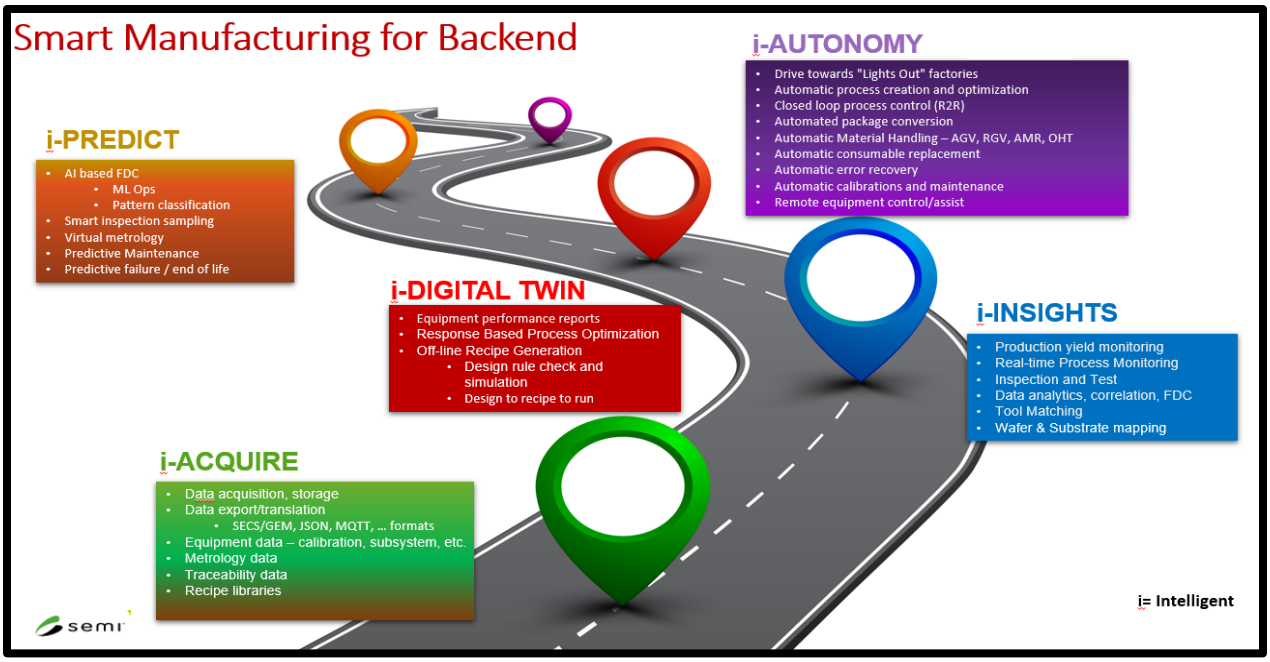


Figure 10.2: SEMI Smart Manufacturing Roadmap Vision for Advanced Packaging Assembly

10.3 Data Acquisition, Provenance & Governance

As the drive towards fully autonomous packaging factories continues, it is critical that assembly equipment suppliers generate the various types of data that are an essential part of developing factory automation capabilities and smart technology. From the beginning of the assembly process, traceability data from assembly tools is used in automating material handling and ensuring the correct product is routed to the correct equipment using the appropriate consumables and process recipes. Data from assembly equipment in real-time during the assembly process to monitor the status and utilization of tools to optimize production efficiency and yield. Data is also used to make intelligent decisions on how to automatically recover from error conditions and even actively compensate to keep the equipment running. For HPC, automotive, medical, and other high reliability packages, traceability data is used extensively throughout the assembly process so each process step can be monitored and analyzed in the event of any unwanted yield excursion.

10.4 Data types/formats – logs, calibration, process, etc.

There are various sources and types of data used throughout the package assembly process. Types of data include but are not limited to application, traceability, process monitoring, inspection, and equipment health data. Factory level management systems such as Manufacturing Execution Systems (MES) and other host level software monitor and control the factory workflow across the fleet of assembly equipment. Data from desired production output capacity levels can be used to route material or balance equipment lines for optimal efficiency based on equipment availability. Within the assembly equipment line, data is often generated in real-time that can be used to monitor process stability. Monitoring calibration, maintenance, or similar data can help ensure equipment health over time.

Data types and formats vary significantly across equipment types and even within the same equipment type from different suppliers. Data file formats often depend on the volume of data being saved, the data sampling rate, or how often it will be accessed. Data formats such as CSV, JSON, or simple text files are commonly used. To realize the value of the data, critical data must be exported from the assembly equipment and imported to a host level database or to the cloud where it can be used by data analytics tools or factory automation software. Various data transfer protocols are used including FTP, SFTP, MQTT, and SECS/GEM. There is an opportunity to standardize both the format of data and transfer protocols in backend assembly.

10.5 Data sampling intervals

Many factors need to be considered when determining data sampling intervals, since the process time varies significantly between processes in backend assembly. For example, die attach equipment place die at a rate of up to 9,000 die per hour, flip chip equipment can bond up to 13,000 units per hour, and wire bonders can bond up to 40,000 units per hour, with individual bonds being formed in less than 10 milliseconds. Processing times, packaging geometry, and number of interconnects all influence processing rates, which therefore define how often data is generated at the package level. The sensor data of interest also has its own data generation rate. Temperature controllers typically generate temperature data used for monitoring at a sampling rate of every 1 second or 0.1 second. Data generated on servo motor controllers generate data in the tens or kilohertz range, while ultras sonic controllers generate data in the hundreds of kilohertz range.

With the faster processing speeds and higher frequency data generated on backend equipment, it is impractical to log time series data in real-time. Doing so would require additional CPU bandwidth on the

equipment which would increase equipment costs. More significant would be the volume of data created. For example, it is estimated that one wire bonder would generate more than 250GB of data per tool per day if raw data were to be logged. Because of these considerations, some equipment suppliers analyze critical data in real time and log key attributes of the data used in monitoring process stability. This data is logged on a per bond, per wire, or per device basis depending on the process type used.

10.6 Traceability

Traceability data is critical in the assembly process and is unfortunately sometimes overlooked. Establishing end to end process traceability is important because it enables tracking a particular package throughout the entire assembly and test process. Without traceability data we can't be sure that we are auditing the correct assembly tool or analyzing the correct assembly or inspection data for a particular package. Traceability data includes identifiers that are on the package materials, equipment, processes, or even operators that are part of the assembly process. Barcodes or 2D codes are now commonly used directly on wafers or substrates for traceability. These codes are read by the assembly equipment so data can then be mapped to that particular package and logged accordingly. Identifiers are also used to keep track of the individual tool used for each assembly process along with the process recipe and any consumables used. Inspection equipment also makes use of traceability data to assure inspection data is properly assigned to the correct package to ensure its quality. Traceability data is an important consideration for all packages including advanced packaging and HI, but especially for those higher reliability packages such as HPC, automotive and medical with targets of zero defects. An additional complexity for HI is the traceability of KGD (Known Good Die) data from different facilities.

10.7 Metrology & Test data

Metrology data is generated throughout the assembly process from various types of manual and automated inspection equipment. During the process optimization phase of a new package, metrology and inspection are performed extensively to ensure an optimal process recipe is created that will be robust enough to extend into high volume production with acceptable performance and yield. In production, inspection is performed during or after each process step as much as is practical to identify and contain any issues as early as possible. It is advantageous to identify the cause of a potential defect at an earlier step in the assembly process rather than waiting for the final electrical test to minimize yield loss. Assembly equipment suppliers have developed inspection capabilities in-situ which provide faster detection of defects and prevent additional yield loss before the issue is detected. For example, vision systems and algorithms directly on tools can detect qualitative defects and also perform quantitative measurements to ensure bonds are the correct size, bonded in the correct location, or wire loops are formed at the correct height. Standalone inspection equipment such as automated optical inspection (AOI) or destructive test equipment such as bond testers are used between assembly processes to inspect for potential defects that could affect yield.

10.8 Data Analysis

10.8.1 Discovery & Correlations

As discussed earlier, there is an abundance of data created during the assembly process. The data includes attributes and identifiers about the package itself, along with data generated by the assembly equipment used and any inspections performed. Understanding key attributes, finding value in the data, and putting it into action is the key. R&D efforts are now focused on identifying which data is most

critical and making sure it is actively monitored and logged. Learning which data from assembly equipment and which specific sensor or subsystem provides valuable data requires extensive analysis to enable the discovery. Once the data is logged on the equipment, it is transferred to a host level database or to the cloud where it can be more efficiently managed and used for analysis. Powerful software analytic tools are used to aid in the discovery process using advanced machine learning algorithms to identify correlations in data that has been integrated from various assembly processes and equipment. Correlations in data are used in various ways including development of fault detection and classification (FDC) algorithms, digital twins and closed loop process control. Discovery and correlations are the foundations of getting value out of assembly data.

10.9 Yield & Performance Monitoring

Continuously monitoring yield and performance data from tools in the assembly process enables real-time tracking of production yield, efficiency, quality, and associated operating costs. Yield Management Systems (YMS) are used to provide management and engineers with real-time and in-depth statistical data from manufacturing that is used to enable data-driven decisions to optimize production efficiency and improve operating expenses. Although some IDM or OSATs are developing their own YMS, there are many robust offerings in the market that can interface with existing MES or ERP systems. , including offerings from providers like PDF Solutions, Applied Materials, ONTO Innovation, and others.

Statistical yield data is logged and classified according to the SEMI-E10 standard, which defines the measurement of equipment reliability, availability, and maintainability. SEMI-E10 specifies six states of equipment operation – productive time, standby time, engineering time, scheduled downtime, unscheduled downtime, and non-scheduled time. States are then combined into categories of manufacturing time, equipment uptime, and equipment downtime as shown in Figure 10.3.

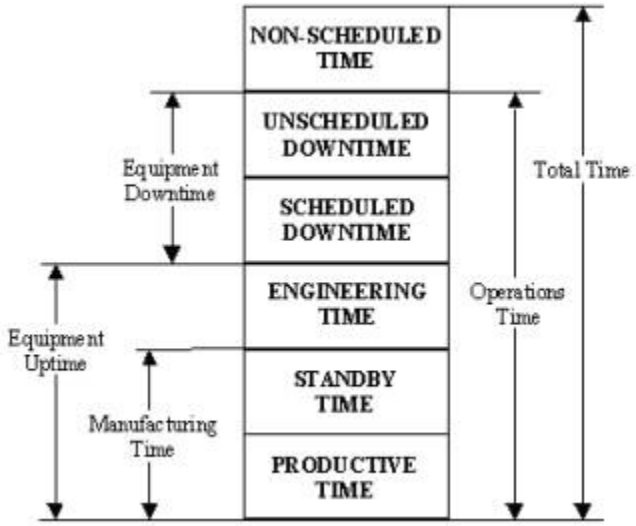


Figure 10.3: SEMI E-10 (add reference) Basic States

10.10 Modeling

10.10.1FDC Modeling and Monitoring

Recent development in smart factory and Industry 4.0 (I4.0) initiatives have helped improve assembly equipment quality assurance, operational efficiency, and time to market. New equipment functionalities have been added to meet the desire for factory automation, real-time monitoring, closed-loop optimization, and traceability. Some of the smart functionalities in today's state of art equipment include:

- Automatic setup and calibration solutions enhance portability and tool matching
- Auto recovery features improve equipment up time and reduce need for operator intervention
- FDC (Fault Detection and Classification)
- Pre and Post Process Inspection
- Digital twins provide offline model for design simulation

10.10.2Digital Twins

An emerging trend in Smart Manufacturing is the use of Digital Twins in the package design phase. A Digital Twin is a digital representation or model of a physical object or process. Digital Twins enable running simulations in the digital realm, prior to any prototype builds. Using these models helps to avoid cost rework or delays in new product introduction. For example, in wire bonding equipment 3D loop models serve as a Digital Twin to the actual wire loops formed on the bonder. Figure 10.4 below shows the wire bond loop Digital Twin that can aid in offline loop design, optimize wire layout and bonding sequence, and perform clearance check for wire to wire spacing and capillary to wire interference. This digital twin can dramatically shorten the time to market and improve production yield.

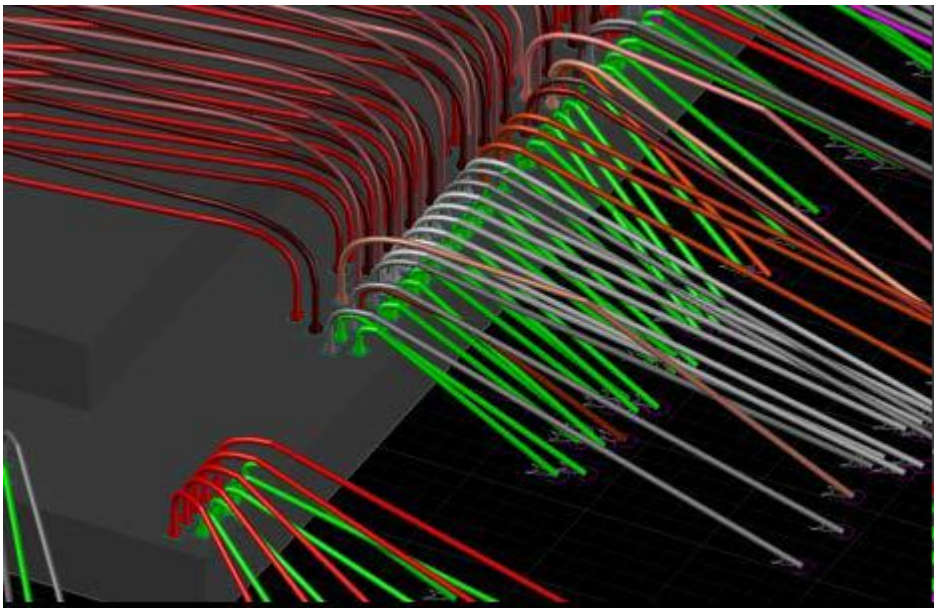


Figure 10.4: Wire looping model digital twin for offline programming, clearance check to improve time to market and fabrication yield. Courtesy of Kulicke and Soffa Industries.

10.11 Prediction

10.11.1Real-time Process Monitoring

The requirement to improve yield for high reliability packages has driven the ability to monitor the health and stability of the processing equipment used in assembly as well as process quality indicators in real-time. Additional sensors are being added to key electromechanical subsystems to enable monitoring of health indicators such as motor temperatures, motion control performance, tracking errors, encoder outputs, pneumatic pressure and flow rates, etc. Actively monitoring these sensor measurements and ensuring the data remain within limits specified by equipment manufacturers prevents equipment performance from drifting undetected. It is also critical to monitor sensor data during the assembly process in an effort to improve and maintain yield.

Depending on the assembly process, processing time, and data sampling rate for the sensor, this real-time process monitoring can be a challenge and generate large amounts of data. For example, on wire bonder equipment bonds are typically formed within about 10 to 20 milliseconds. During this time, the equipment monitors key sensor data such as bond deformation, bonding force, and ultrasonic transducer control data. Sensor data is generated at about 16kHz sampling rate. Rather than export the large volume of data to the cloud or to an external PC for analysis, many equipment manufacturers analyze the data in real-time directly on the equipment and take the appropriate action. This allows for the fastest detection and response time to any potential defects identified. Monitoring this data in real-time will stop the equipment immediately and prevent the chance for continued yield loss versus detecting defects in a process step that occurs later.

10.12 Autonomy

10.12.1“Lights Out” Factory

A “lights out” factory is where there is minimal human intervention in the manufacturing process of the facility and doesn’t necessarily mean that the factory is operating with “lights-out”. In fact, given the complexity of semiconductor and HI assembly processes, it's unlikely that such operations will ever operate in a completely dark environment as there will always be some human presence required. Front-end wafer fab facilities are pioneering these automation methods but the back-end assembly is starting to integrate such methods too.

As an example from [1], IBM Guadalajara Mexico Industry 4.0 manufacturing team has developed and deployed an enclosed customized collaborative robot (COBOT), vision system, and affiliated vision recognition process that identifies socket contact and possible socket related defects to mitigate defects during production. COBOT systems coupled with cameras and integrated software programs to detect defects and guarantee proper assembly. If socket defects are found the system auto-records and highlights specific defect areas, enabling rapid defect identification and resolution. If no defects are found, the vision recognition system and COBOT arm and head assemblies are used to pick, place & secure assemblies with an in-plane alignment precision of approximately 50µm. The system can be used for disassembly or rework as well and similar assemblies are starting to be used in other product assembly lines (e.g. DIMM card technology).

10.12.2 Automatic setup & optimization

Labor shortages during and after the Covid-19 pandemic have dramatically increased the need to automate as much as possible in the back-end assembly process. In addition to reducing labor requirements and associated costs, factory automation solutions can improve yield by eliminating issues caused by product mishandling, which is critical for medical devices. Factory automation equipment is typically integrated with material control scheduling and factory MES systems to ensure the correct materials, consumables, equipment, and process recipes are used in the assembly process. Factory automation systems also improve factory efficiency by reducing equipment stand by time, waiting for operators to load materials. These automation techniques and best practices will be critical in any US re-shoring effort in the near future as they have the benefit of reducing costly specialized labor training and minimizing assembly errors.

There are four major types of factory automation systems currently running in high-volume production in back-end assembly plants to deliver materials and then load and unload the materials to the assembly equipment. Each solution offers a varying degree of automation and compatibility with other assembly processes.

Over-head transfer (OHT) is used extensively in the front-end, and has been adopted by the memory segment for backend assembly. OHT material handling is attractive from a safety perspective since the robot travels along a track near the ceiling, which reduces the chance of contact with any operators or interference with other equipment. However, this type of automation system requires the ceiling height allowance to install the OHT robot track and the infrastructure investment is costly. OHT robots also travel only along a fixed path, so it is less flexible and scalable compared to other alternatives.

Automated guided vehicles (AGV) and *autonomous mobile robots* (AMR) and under evaluation at many back-end assembly plants. These robots can transfer materials between assembly areas, delivering materials to input and output buffer stations or even load and unload materials directly to the equipment. Current trends show AGV systems being replaced by AMR robots. AMR can travel in more densely populated areas with equipment and operators. An AMR's on-board position sensing and safety sensors allow it to self-navigate around obstacles and stop immediately in the presence of any operators. AMR provides the maximum flexibility and not much infrastructure. As shown in Figure 10.5 below, an AMR can service a fleet of the same equipment or could also service various types of equipment within the same factory or even travel on an elevator to different assembly areas.

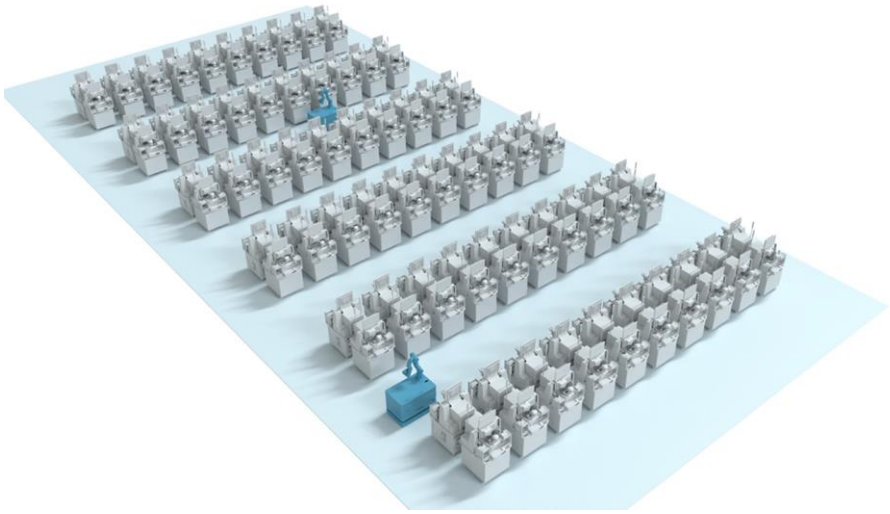


Figure 10.5: Autonomous Mobile Robot Servicing Wire Bonders. Courtesy of Kulicke and Soffa Industries

Rail guided vehicles (RGV) offer a factory automation solution where equipment is arranged with a dedicated robot that travels along a rail located on the floor (Figure 10.6 below). This configuration eliminates many of the concerns with safety as the robot is located behind the equipment, allowing the operator full access from the front. This configuration is becoming popular for high volume assembly equipment, such as wire bonders at the OSATs. Depending on throughput, one RGV can load and unload material for up to about 50 wire bonders.

Figure 10.6: Rail Guided Vehicle Servicing Wire Bonders. Courtesy of Kulicke and Soffa Industries.

Conveyor based factory automation is where equipment is arranged in-line and the quantity of each equipment type is based on individual throughput of each to balance the line without any bottleneck. This configuration is typically used when production lines are dedicated to specific packages. Conveyor based in-line systems are not very flexible or conducive to frequent device changes, so this has not been adopted much by the OSATs. Figure 10.7 below shows an example of an in-line system consisting of a die bonder, snap cure oven, wire bonders, and automated optical inspection equipment. This in-line system has been optimized for assembly of CMOS image sensor packages.

Figure 10.7: In-line System composed of a die-bonder, snap cure oven, and other assembly equipment for CMOS image sensor packages. Courtesy of Kulicke and Soffa Industries.

10.13 Surface Mount Technology

Surface Mount Technology - factory automation research towards Industry 4.0: AI-based Closed-Loop Self-Optimization Platform

With the rapid technology development in Industry 4.0, such as artificial intelligence (AI), electronics manufacturing processes can be more intelligent. Smart manufacturing, which adopts real-time decision-making based on operational and inspectional data, can soon be realized [1]. In Surface Mount Assembly (SMT) lines, data-driven solutions can be applied to diagnose abnormal defects and adjust optimal machine parameters in response to unexpected changes/situations during production with the collected data. Collaborating with various industry partners, the State University of New York at Binghamton research team at Binghamton developed a novel framework based on AI-based closed-loop feedback control and parameter optimization to implement an intelligent manufacturing solution in the

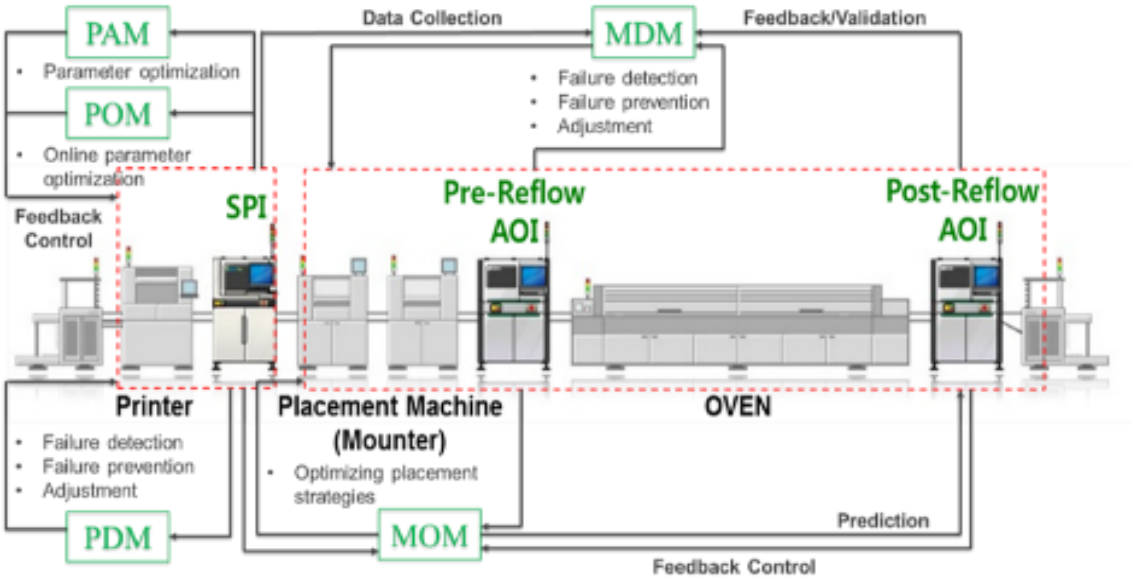


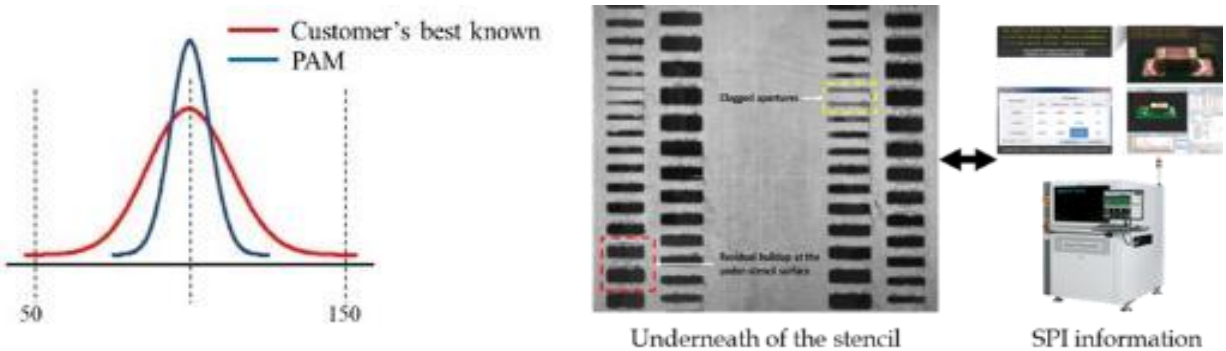
Figure 10.8: Schematic diagram of the AI based, closed loop feedback system

PCB assembly for yield and throughput improvement. This AI-based framework could provide a potential road map for data-driven process control in SMT.

Each SMT process, including solder printing, pick and place (P&P), and soldering reflow (SRP), has a significant impact on the quality and throughput of the final PCB product. As a result, multiple inspection machines, including solder paste inspection (SPI) and automated optical inspection (AOI) machines, are introduced to monitor the manufacturing process. The Binghamton University Smart Electronics Manufacturing Laboratory (SEML) is fully equipped with two solder paste printers, two chip mounters, and a reflow oven, in addition to SPI and AOI machines. At SEML, the research team tested over 10,000 PCBs. The results indicate that numerical methods based solely on physical properties may have practical limitations in explaining the behavioral patterns of small-scale components. However, recent research suggests that methods based on artificial intelligence can improve product quality by up to 35%. It implies that an intelligent SMT process control based on data can advance SMT processes. As a result, the intelligent SMT strives to maintain optimal settings in offline and online environments. **Figure 10.8** depicts the overall schematic of the AI-based closed-loop feedback control framework.

Intelligent SMT Modules

In the solder printing process, four machine intelligence modules are considered: (1) printing advising module (PAM); (2) printing optimization module (POM); (3) printing diagnosis module (PDM); and (4) dynamic stencil cleaning process control (CPC). PAM and POM aim to recommend and adjust the critical printer parameters, such as printing speed, printing pressure, and separation speed, using hybrid machine learning and heuristics optimization techniques offline and online, respectively [3, 4]. The experimental results indicate that by advising and adjusting printing parameters, PAM and POM can improve production quality by more than 60% in the Cpk. PDM identifies potential printing failures to optimize process quality and minimize downtime [5]. The experimental results indicate that the PDM is capable of predicting various types of defects with an accuracy of greater than 87 %. The CPC analyzes the SPI data to determine the amount of residue on the stencil undersurface and evaluate the stencil cleaning profile and cycle control [6]. The CPC improves the robustness and quality of the cleaning process by 34% and 10%, respectively, compared to the best-known cleaning parameters. For instance, Figure 10.9 (a) Illustrates the expected outcome of applying printing modules while showing the AI-based residual prediction.



(a) (b)

Figure 10.9: Application of solder paste printing modules (a) PAM effectiveness (b) smart residue buildup prediction

The mounter optimization module (MOM) and the mounter diagnosis module (MDM) can be used during the P&P procedure to optimize the P&P machine's parameters automatically. The final offsets of the components are predicted in the MOM framework using a hybrid AI model based on data collected by SPI, Pre-AOI, and Post-AOI machines. MOM can determine the optimal placement with the least post-reflow misalignment possible. The experimental results indicate that MOM can reduce final misalignments by 18% compared to a conventional placement method (i.e., placing a component on the

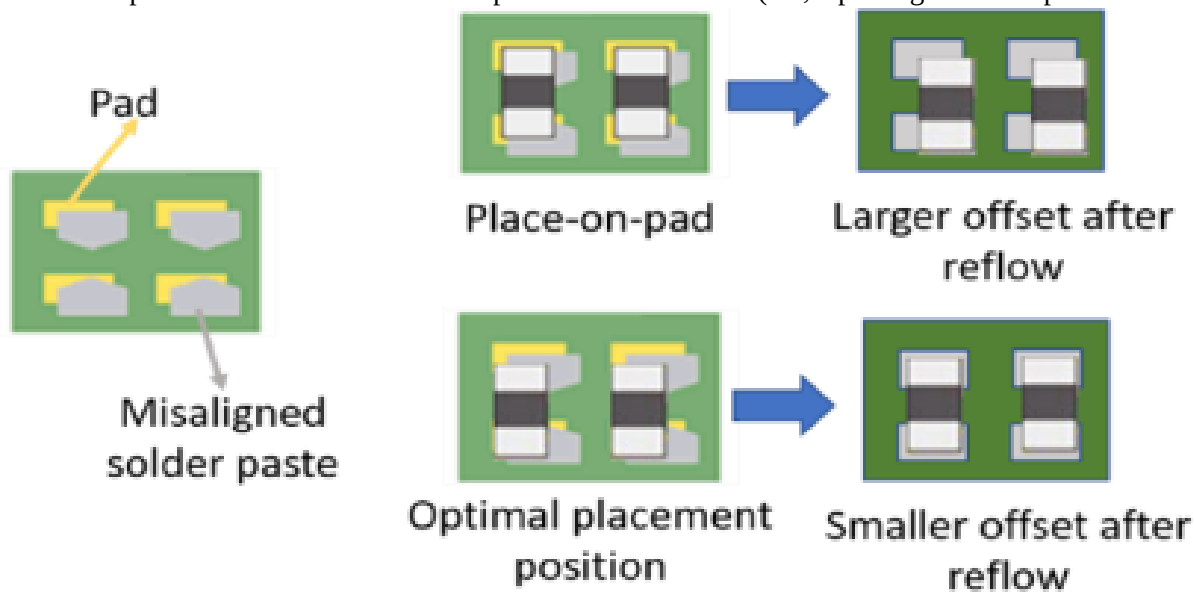


Figure 10.10: Achieving the optimal placement position in the MOM

pad center). Figure 10.10 represents a visual illustration of the mounter placement optimization. MDM is a preventive and prescriptive maintenance method that utilizes operational and AOI inspection data from P&P machines to determine the root causes of P&P defects and to prevent future failure. MDM can identify the known root causes of certain defects, such as improper nozzle size and nozzle contamination, with an accuracy of 84.5%. It demonstrates that when an abnormality is detected using AI-based diagnosis algorithms, various mounting defects can be detected and classified automatically with higher level of accuracy.

The goal of the reflow setting optimization process is to determine the best reflow oven temperature settings that ensure the final quality of the PCB products by fine-tuning the actual thermal profile to the manufacturer's target profile. As a result, the tuning process undertaken by the reflow engineers is time-consuming and costly. We propose an automated recipe optimization model for reflow soldering based on the thermal profile of the printed circuit board and its recipe. First, the initial recipe collects the thermal profile and the recipe and then proposes a simulation model based on the relationship. After that, an AI-based model is used to generate an optimal recipe that minimizes the difference between the simulated and target temperature profiles. The AI-based optimization enables us to achieve 97% fitness in the given target profile within an hour. The AI-based model increases the degree of automation, resulting in time and labor savings. In the future, data from multiple inspection machines will be integrated to improve the reliability of the reflow optimization process.

10.14 Summary

Due to the size of small-scale electronics products, SMT processes have become significantly more complicated to maintain high-quality PCB products. Theoretical interpretations of SMT processes can be complex due to numerous uncertain variables. SMT processes can be intelligent and adaptable to changing environmental conditions with the help of AI and big data. The quality of the final printed circuit board can be improved while maintaining optimal control parameters throughout the SMT processes. Automated and intelligent systems enable the next level of electronics manufacturing, which accelerates the manufacturing of customized products by leveraging data and information from end-users via edge/cloud computing. Smart manufacturing in this sense is intended to produce parts that will more closely adhere to IPC or JEDEC specifications as they are programmed to do. Assemblies will be more consistent with fewer outliers and ultimately better reliability at a lower cost.

References

1. Chan, B., Hoffmeyer, M., Chow, E.M., Qin, I., Won, D. and Park, SB, 2022, Emerging Technology; Smart Manufacturing of Computer Systems and Assemblies:
2. Lau, J., et al. 2014, Redistribution Layers (RDL's) for 2.5D/3D IC Integration, Journal of Microelectronics and Electronic Packaging (2014) 11, 16-24
3. Mfg Roadmap Microelectronics, NIST 2022 - MfgTech Roadmap - Microelectronics | NIST