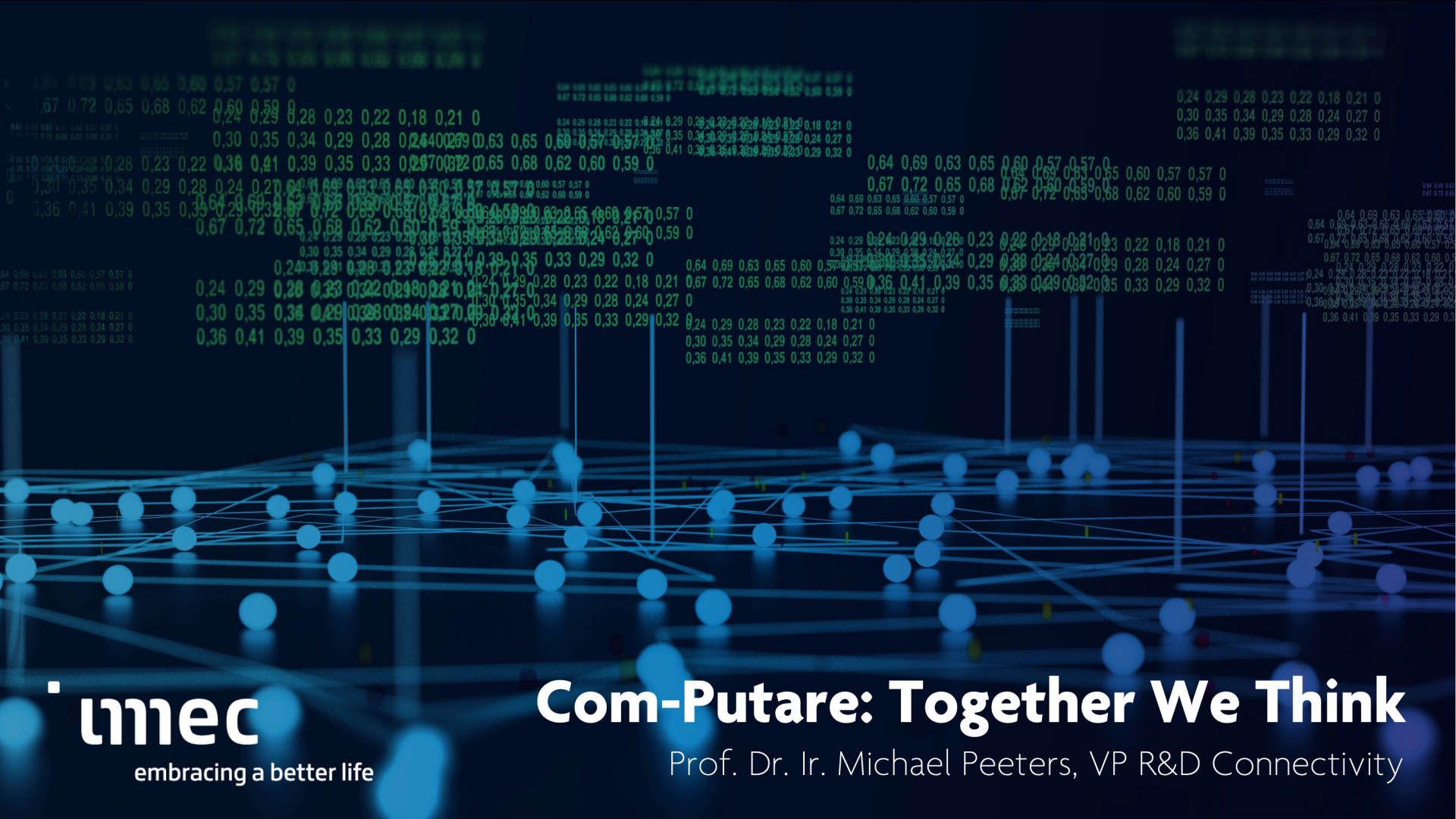
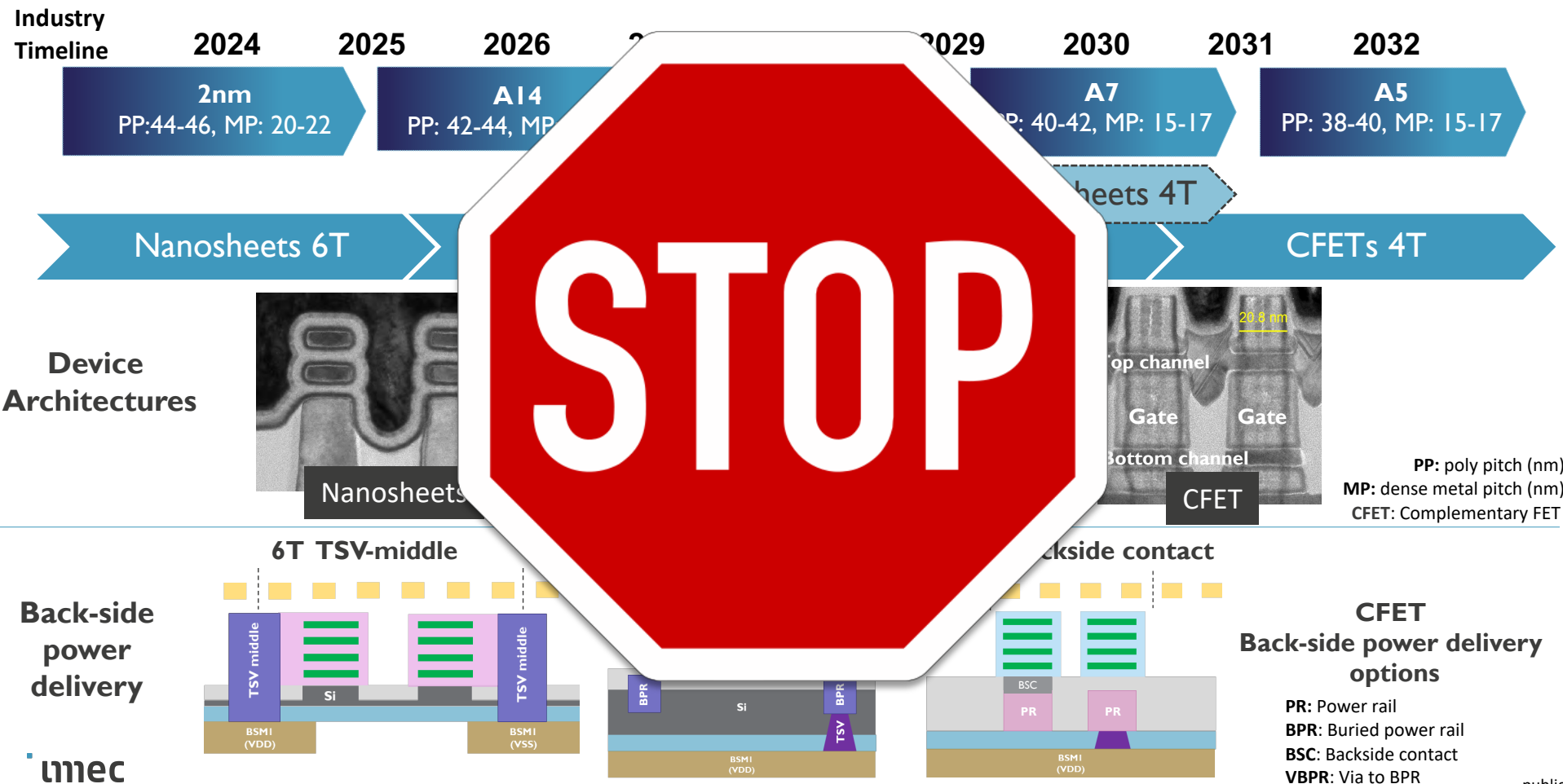




mec
embracing a better life

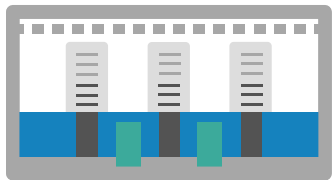
Com-Putare: Together We Think
Prof. Dr. Ir. Michael Peeters, VP R&D Connectivity

Logic Scaling Roadmap



Computing Research Beyond Conventional CMOS

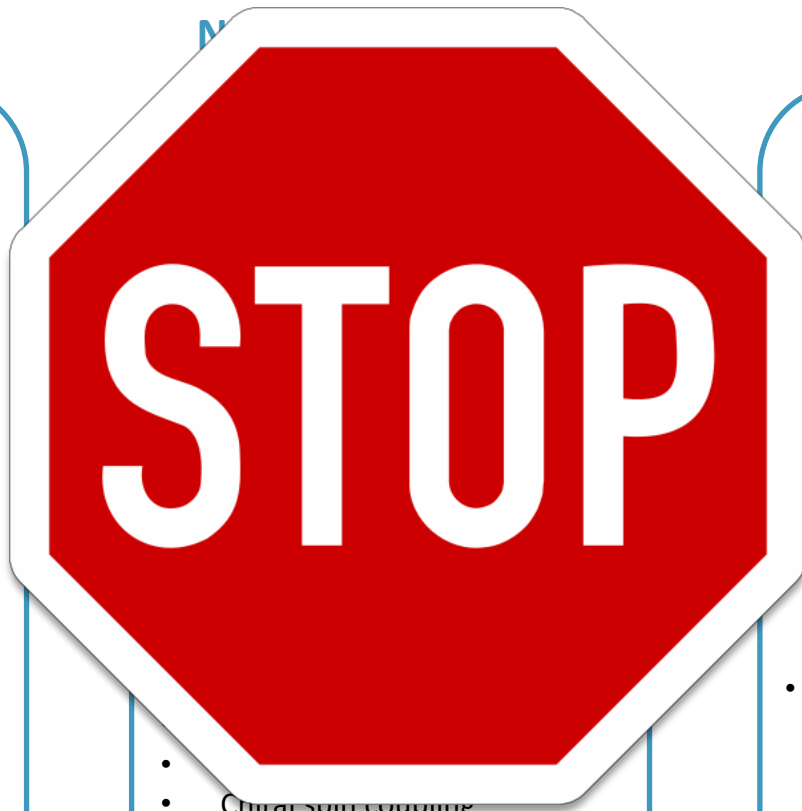
The next CMOS switch



2D atomic channels



2D Atomic Channels

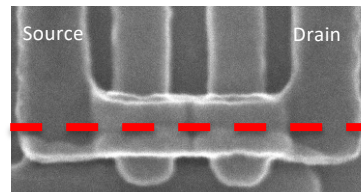


-
- Chiral spin coupling

Magnetics-based Majority Gates

Paradigm change

Superconducting and Si spin qubit platform



- Leveraging 300mm Si learning

Quantum Computing

The background is a dark blue, almost black, field filled with intricate, glowing patterns. Two prominent, thick, glowing blue lines curve across the frame, one from the top left towards the bottom center, and another from the top right towards the bottom center, creating a sense of dynamic movement. These lines are composed of many fine, parallel segments. In the background, there are faint, concentric circular patterns and a grid of thin, glowing lines that recede into the distance, creating a three-dimensional effect. The overall aesthetic is futuristic and technological.

The Future of Computing...
...is Connectivity

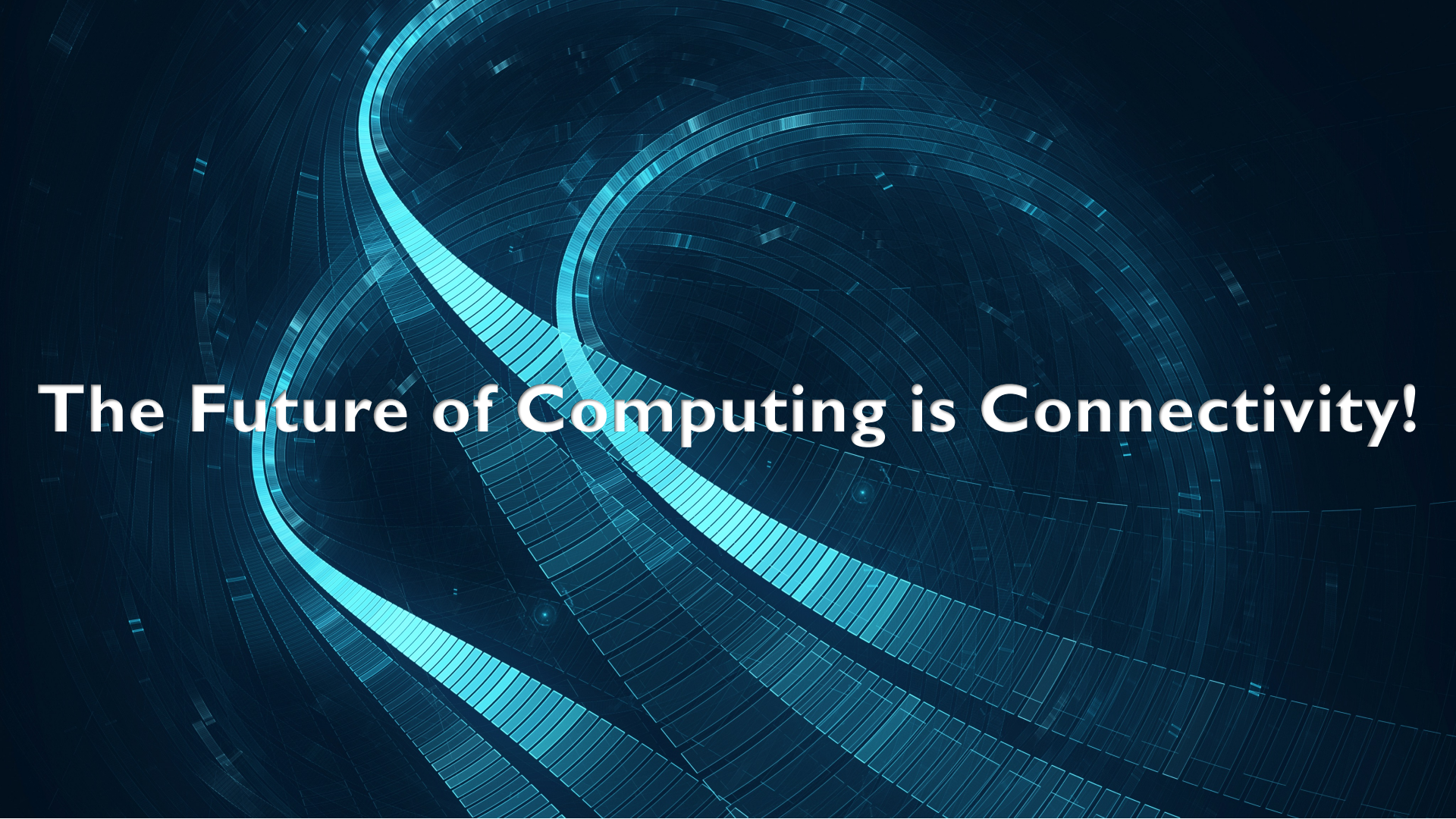


Semiconductors and (sub-)THz

6G and 5G-Advanced

Prof. Dr. Ir. M. ... connectivity, imec

On behalf of the Advanced RF team



The Future of Computing is Connectivity!

clean
purify
prune
simplify
account
consider
think

Parable

The blind men and the elephant

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

— John Godfrey Saxe



From Augusta Stevenson, *Children's Classics in Dramatic Form*, Houghton Mifflin (Boston),

Organisms communicate to improve their chances of achieving their goals



THE ROYAL
SOCIETY

doi 10.1098/rspb.2000.1140

On the advantages of information sharing

Michael Lachmann^{1*}, Guy Sella^{1,2} and Eva Jablonka³

¹*Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA*

²*School of Mathematics, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel*

³*Cohn Institute for the History and Philosophy of Science and Ideas, Tel-Aviv University, Tel-Aviv 69978, Israel*

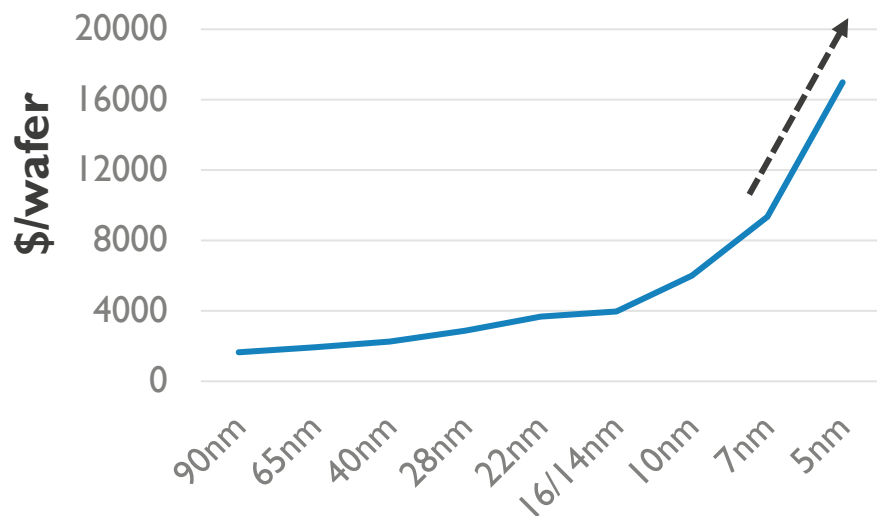
During the evolution of life, there have been several transitions in which individuals began to cooperate, forming higher levels of organization, and sometimes losing their independent reproductive identity. For example, multicellularity and insect societies evolved independently multiple times. Several factors that confer evolutionary advantages on higher levels of organization have been proposed. In this paper we highlight one additional factor: the sharing of information between individuals. Information sharing is not subject to the intrinsic conservation laws that characterize the sharing of physical resources. A simple model will illustrate how information sharing can result in aggregates in which the individuals both receive more information about their environment and pay less for it. This may have played a role in the evolution of higher levels of organization.

Keywords: information sharing; multicellularity; ensociality; public goods; evolution

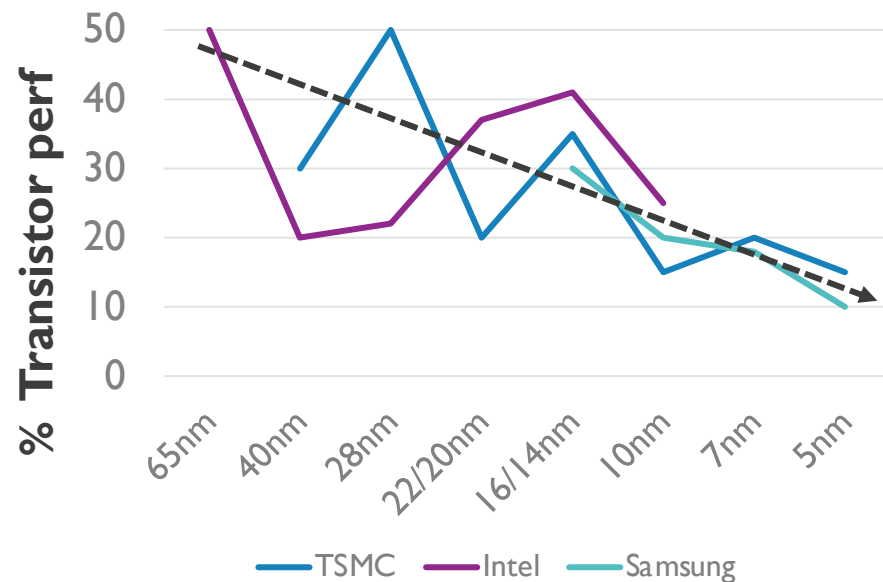


Economic Realities of Logic Scaling

Wafer prices increasing significantly
node to node



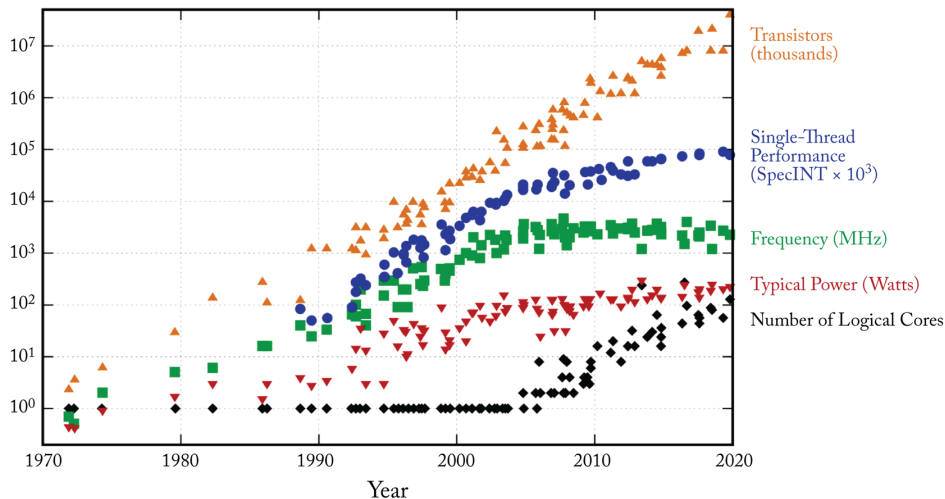
Node to node perf improvement
is decreasing



The compute trend

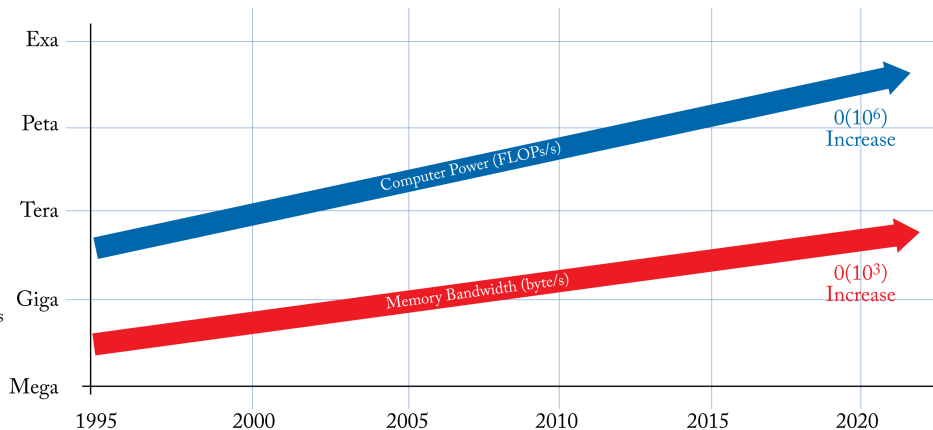
Where are the bottlenecks?

48 Years of Microprocessor Trend Data



*Source: K. Rupp, 2020

- Growth in the number of cores may slow down
- More transistors utilized for **domain-specific acceleration**.
- Single-thread performance continues to increase due to better instruction pipeline, improved branch prediction, out-of-order execution, larger instruction vectors, and specialized execution units, **resulting in more IPC**.



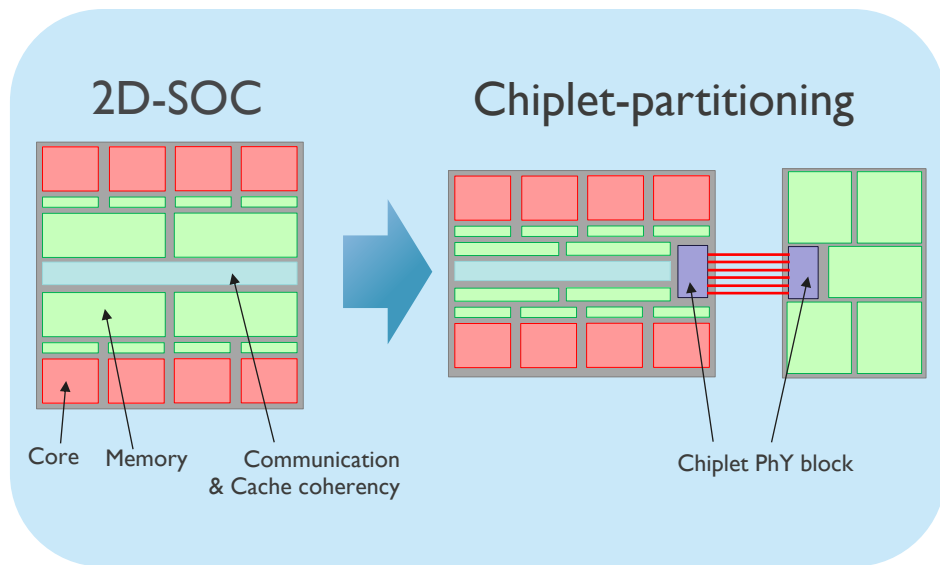
*Source: P. Lindstrom, IPAM, Oct. 2018

Bottlenecks executing DNN workloads are:

- main memory bandwidth;
- local (SRAM) memory; and
- power (primarily from data movement).

SOC Functional partitioning : Memory-Logic

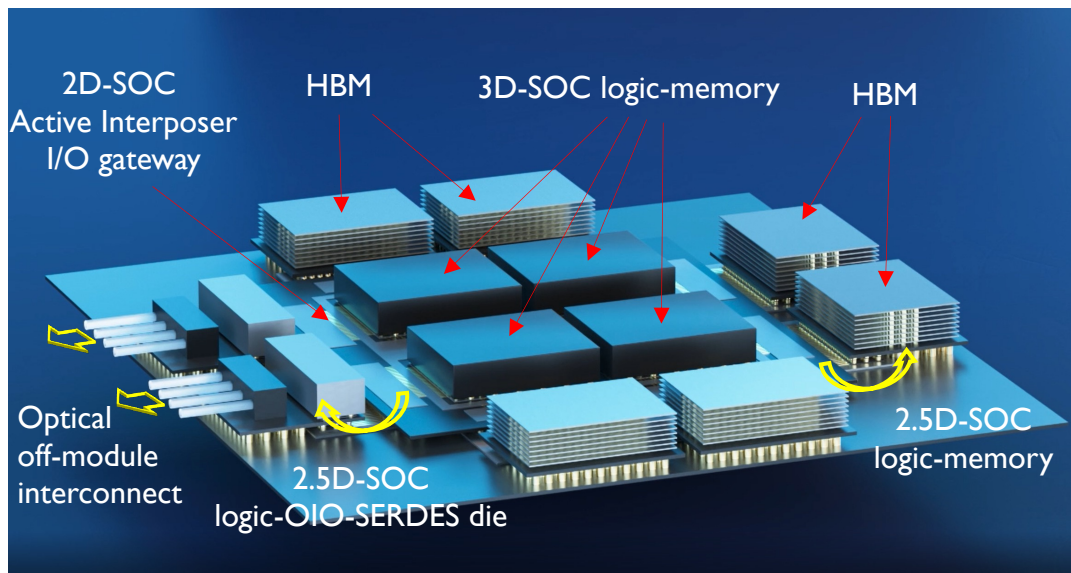
The “Chiplet” approach



- Separate design and processing of chiplet die
- Die-to-die communication through (standardized) interface PhY IP-block and wide bus interface
 - ⇒ introduces additional latency
 - ⇒ only for last level cache
 - ⇒ Today mainly course bump pitch: 2.5D centric
 - ⇒ With highly scaled I/O pitch: also enabling 3D

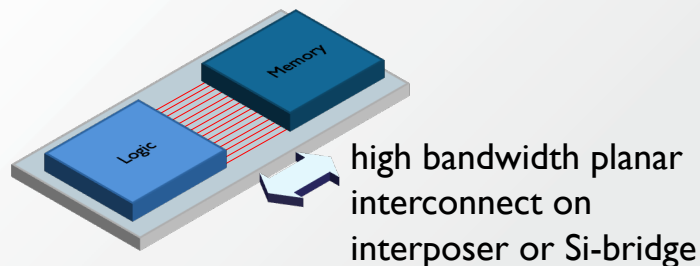
High Bandwidth 3D interconnect

System driver for 3D integration technology

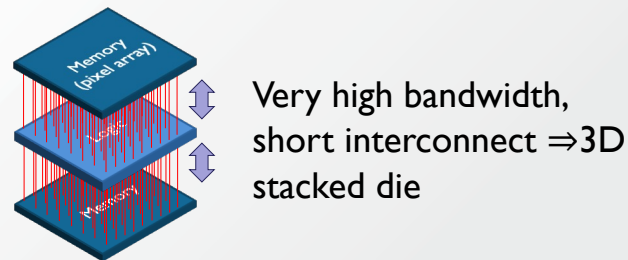


- Need to match and optimize bandwidth density, data latency and Energy per bit
- Need for different technologies with different 3D integration densities

Lateral interconnect



Area-array interconnect

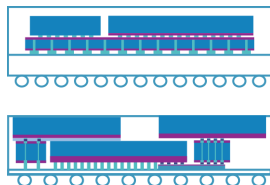


The 3D Interconnect Technology Landscape

3D-SIP

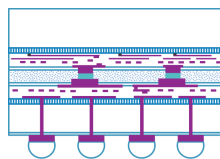
Package stacking

- Multi-die Packaging
- Interposer “2.5D”
- Embedded Die

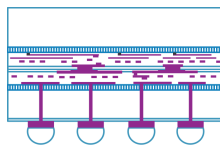


3D-SIC

Die Stacking
 μ bump

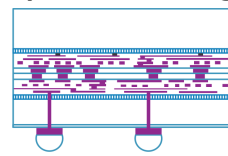


Die Stacking
Hybrid bonding

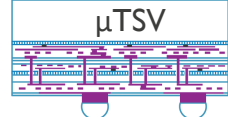


3D-SOC

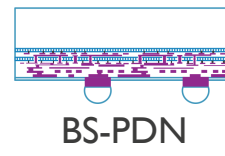
Wafer-to-Wafer
Hybrid bonding



Dielectric bonding



Wafer-to-Wafer
Sequential



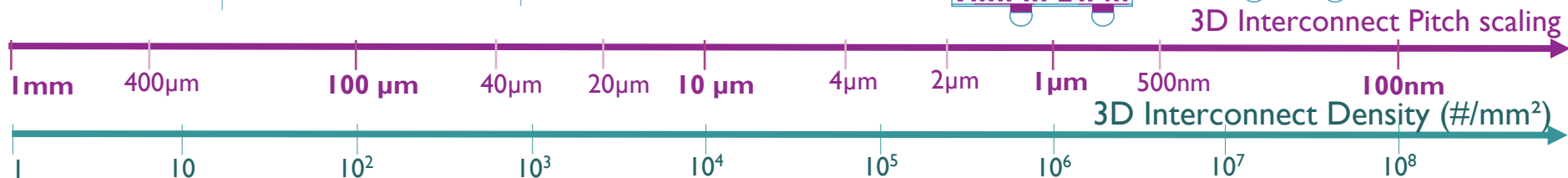
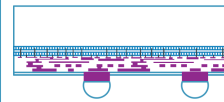
BS-PDN

nTSV



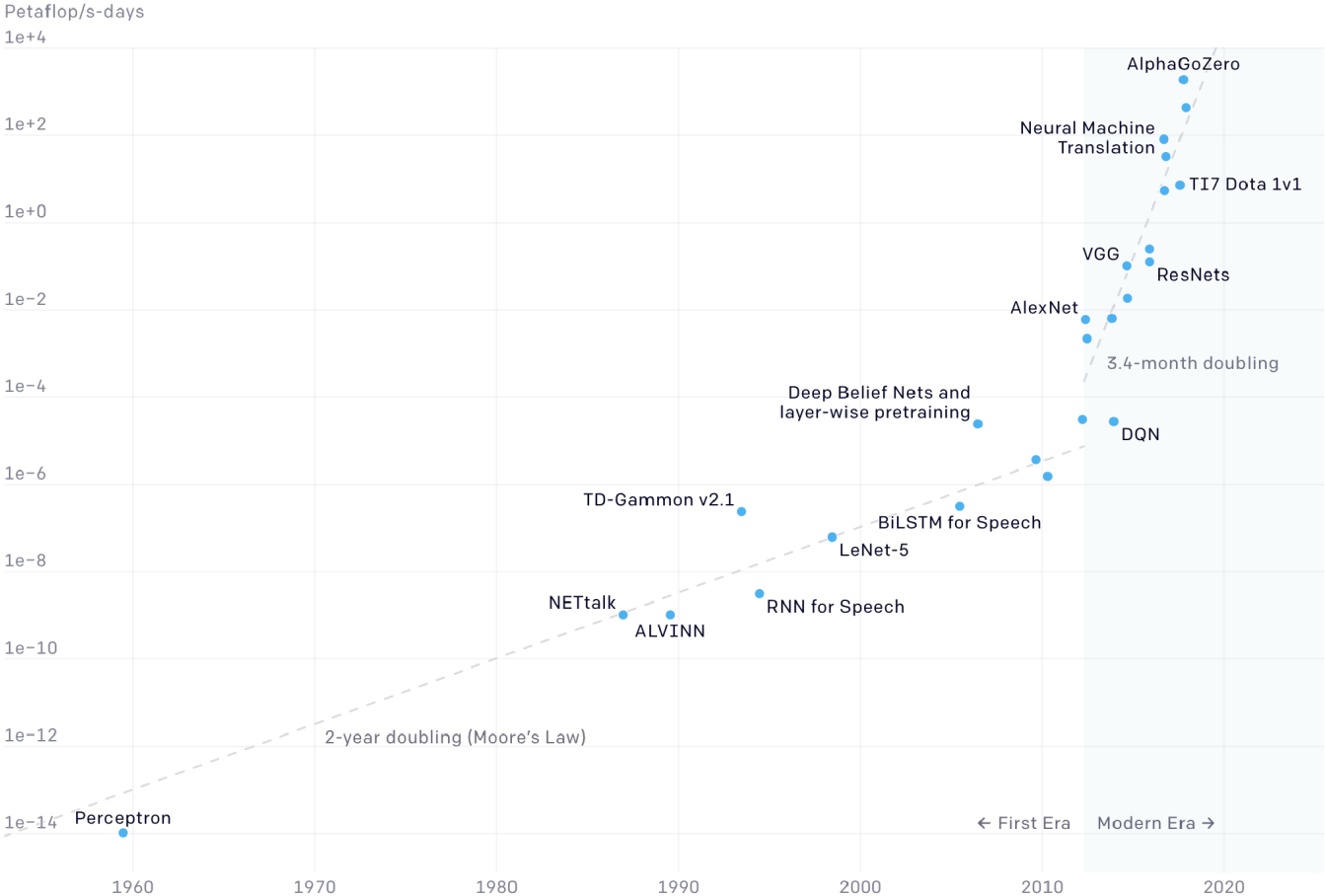
3D-IC

Transistor
Stacking
CFET



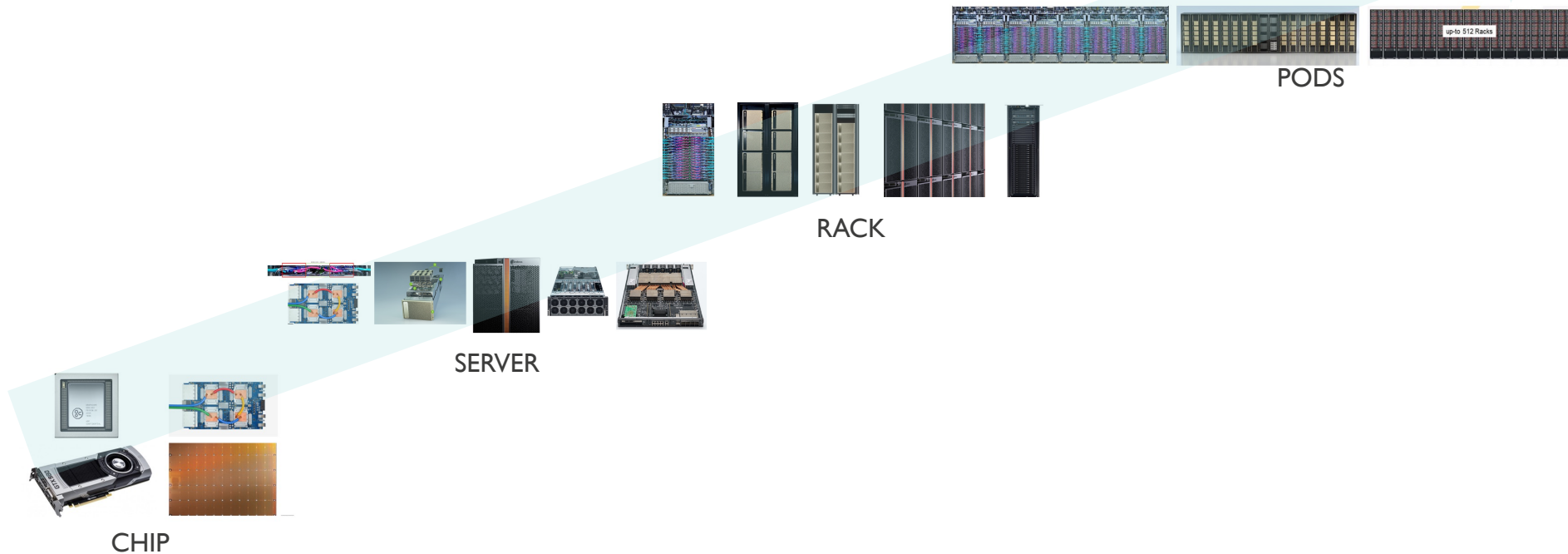
The big wait...

Two Distinct Eras of Compute Usage in Training AI Systems



Scalability is a primary requirement for HPC/AI HW

From “Chips” to “Pods”



Communication efficiency determines performance

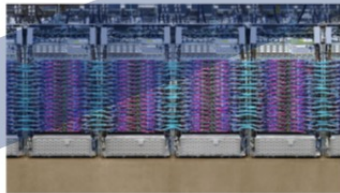
- large-scale system needs to communicate efficiently.
- Google 2D toroidal mesh topology,
- Nvidia Mellanox 100–200 Gb/sec Infiniband with fat-tree topology
- Cerebras and Graphcore IPU1 (standard 100 Gb/sec ethernet links and switches.

11.5 petaflops
4 TB HBM
2-D toroidal mesh network
Training and inference
Alpha



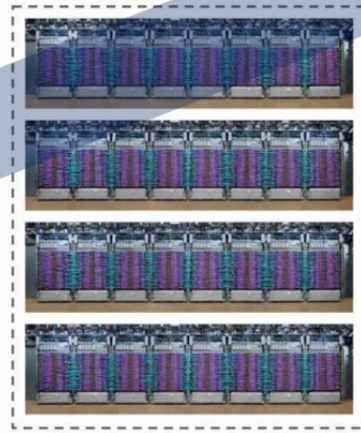
Cloud TPU Pod (v2, 2017)

> 100 petaflops!
32 TB HBM
Liquid cooled
New chip architecture + larger-scale system



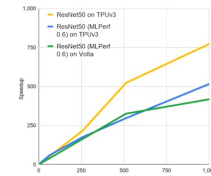
TPU v3 Pod (2018)

> 430 petaflops!
128 TB HBM
> 128x32 2-D mesh topology



TPU SuperPod (2020)

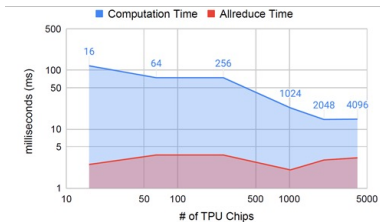
TPUv3 & Volta Scaling for Training: ResNet50



- Only ResNet-50 from MLPerf 0.6 can scale beyond 1000 TPUs and GPUs
- **Gold line** TPUv3 is published result.*
- 77% of perfect scaling at 1024 chips
- **Blue line** TPUv3 is MLPerf 0.6 result: 52% of perfect scaling at 1024 chips
- Why Differ? MLPerf ground rules:
 - MLPerf includes evaluation time; most researchers don't include it
 - Evaluation runs timeout data set after a training finishes to determine accuracy
 - To deter cheating, do 1 evaluation/epoch! No researcher would do it every epoch

*Ying, C. et al. 2018. Image Classification at Supercomputer Scale. arXiv preprint arXiv:1811.06992.

Data communication takes 22% of the total device step time.



*Source: Google cloud



The face of automotive is changing

Mobility-as-a-service



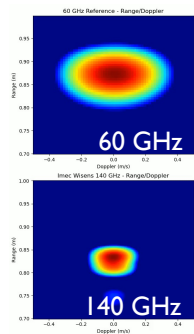
LEVEL 4 AUTONOMY = **MASSIVE** SENSING AND COMPUTE

SENSE & COMPUTE IS A BOTTLENECK TOWARDS ENERGY EFFICIENT MOBILITY-AS-A-SERVICE

IN THE STATUS QUO: UP TO 33% OF THE ENERGY CONSUMPTION, 1 GB/s OF DATA

I 40 GHz CMOS radar

Compact and high-performance radar for next generation mobility



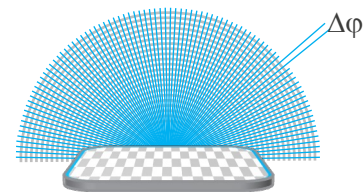
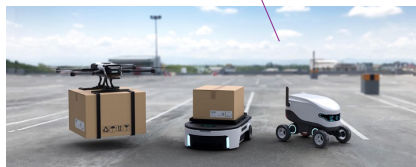
IN-CABIN SENSING

3X IMPROVEMENT IN RANGE RESOLUTION VS. 60 GHZ
5X SMALLER ANTENNA APERTURE

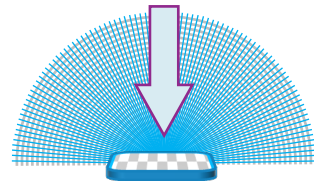


EXTERIOR SENSING

WAVEGUIDE ANTENNAS FOR UP TO 100M RANGE
7 GHZ BANDWIDTH FOR CM RANGE RESOLUTION

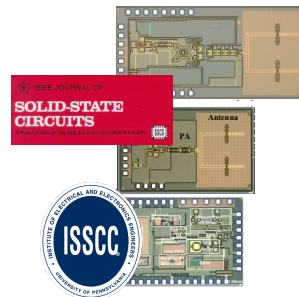


77 GHz

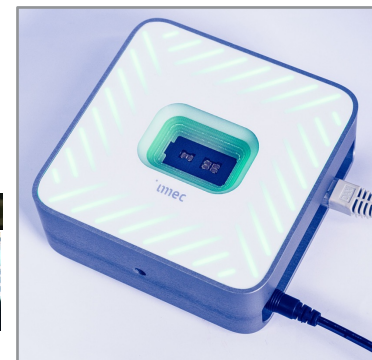


140 GHz (equivalent $\Delta\phi$)

TX/RX published in JSSCC'22
PLL accepted for ISSCC '23



Die embedding
Waveguide-in-PCB



imec 140 GHz CMOS radar development kits

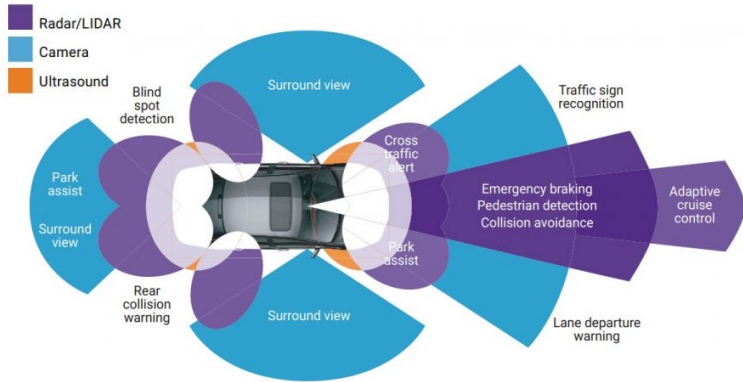
Different versions (on-chip antennas for **short range**, waveguide antennas for **longer range**)

Disruptive innovation

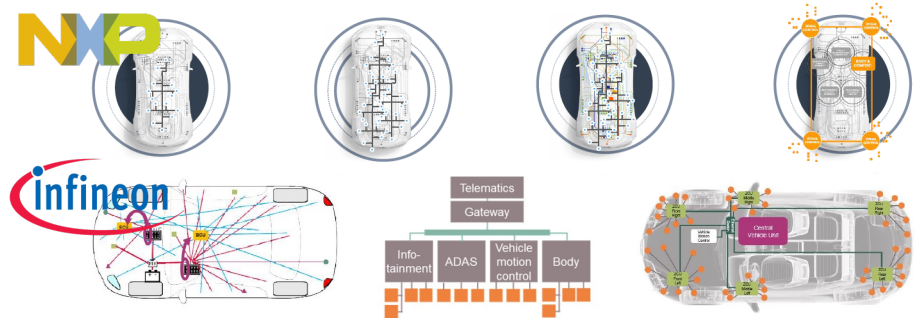
Deeply co-design sensors and AI for emerging architectures

TODAY

Sensors designed in isolation
Non-optimized general compute platform

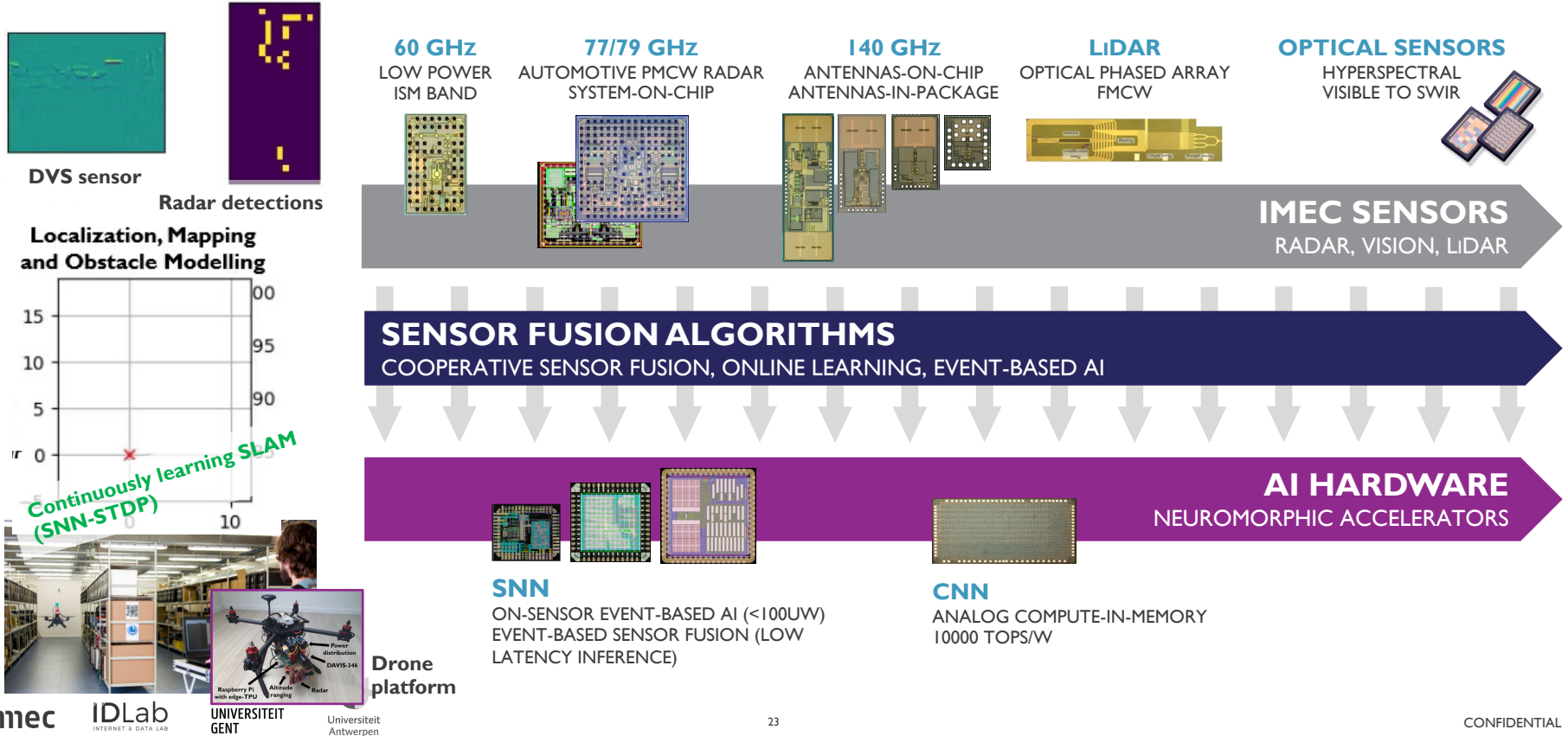


- “**Plumbing**” different sensor data streams (not scalable)
- **Late fusion** of signals (400+ms latency)
- Output of fusion is **huge data representation** (e.g., point cloud)
- Large **stress on higher order AI models** (obstacle avoidance, path planning)



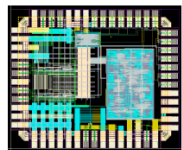
Neuromorphic sensor fusion

From sensors to **senses** – antennas to AI



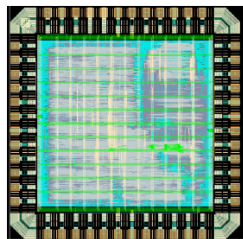
Neuromorphic sensing

From on-sensor feature extraction to event-based sensor fusion



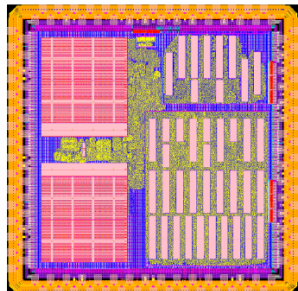
uBrain-X (2020)

- 100 neurons
- On-chip analog-to-spike-convertor
- Insertables (e.g. ExG)
- 40 nm CMOS



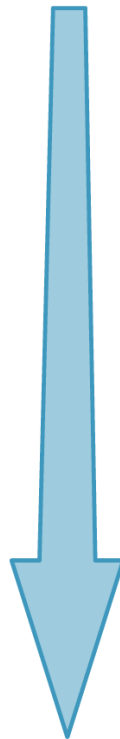
uBrain (2019)

- 336 neurons
- Digital, event-driven
- Small tasks (ECG, audio, SISO radar)
- 70 μ W
- 40 nm CMOS



Argus (2021)

- SoC with 1300 neurons
- High effective # neurons (time-multiplexing)
- Digital, event-driven + RISC-V controller
- Larger tasks (radar/DVS fusion, keyword spotting)
- 1-10 mW (use case dependent)
- 28 nm CMOS



Reference datasets for performance benchmarking (MNIST, Google Soli, Google speech, ECG, DVS128, ...) – **in-house datasets** (e.g. NeuroDrone) for new applications

Expand AI task complexity within **stringent energy/latency/accuracy budget**



Device
Edge

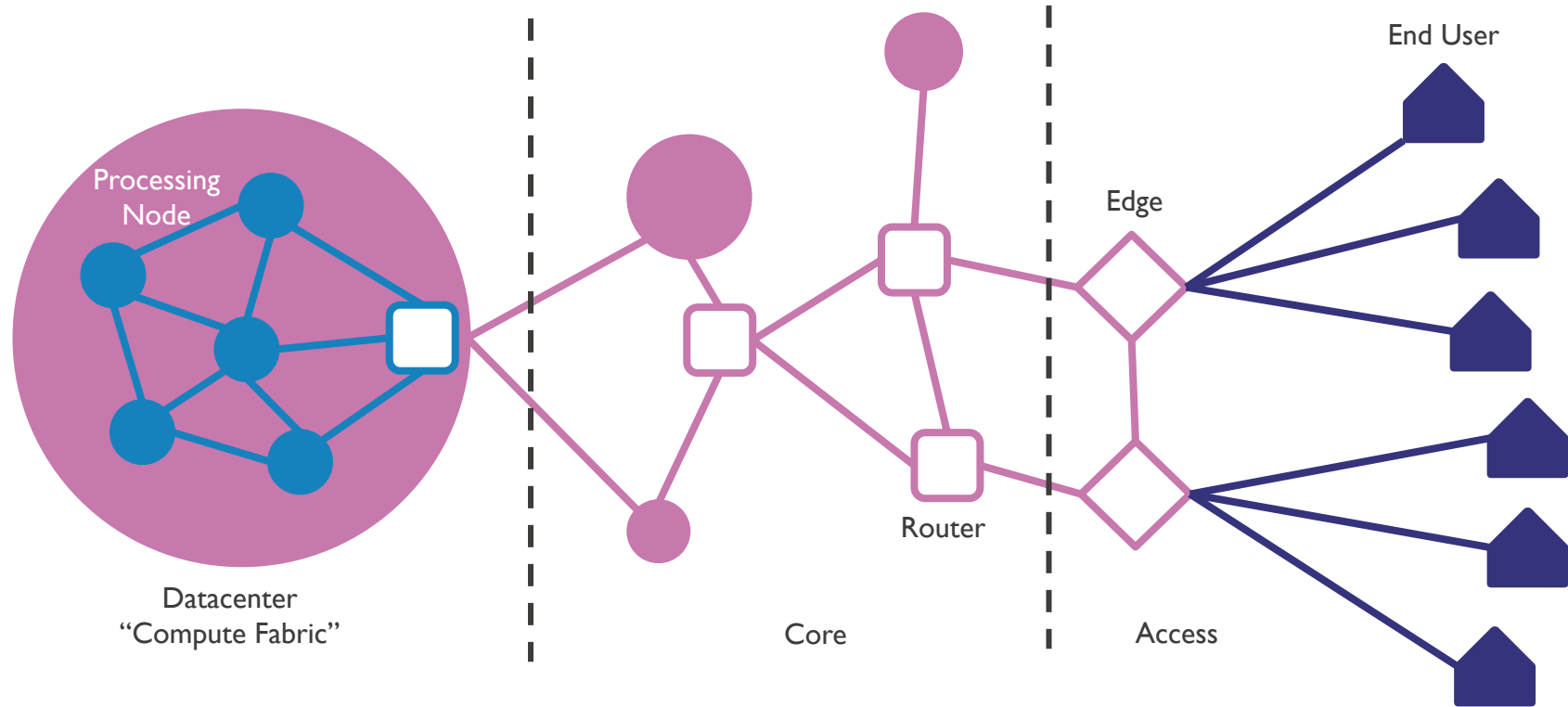
vs.

Infrastructure
Edge

“Value”

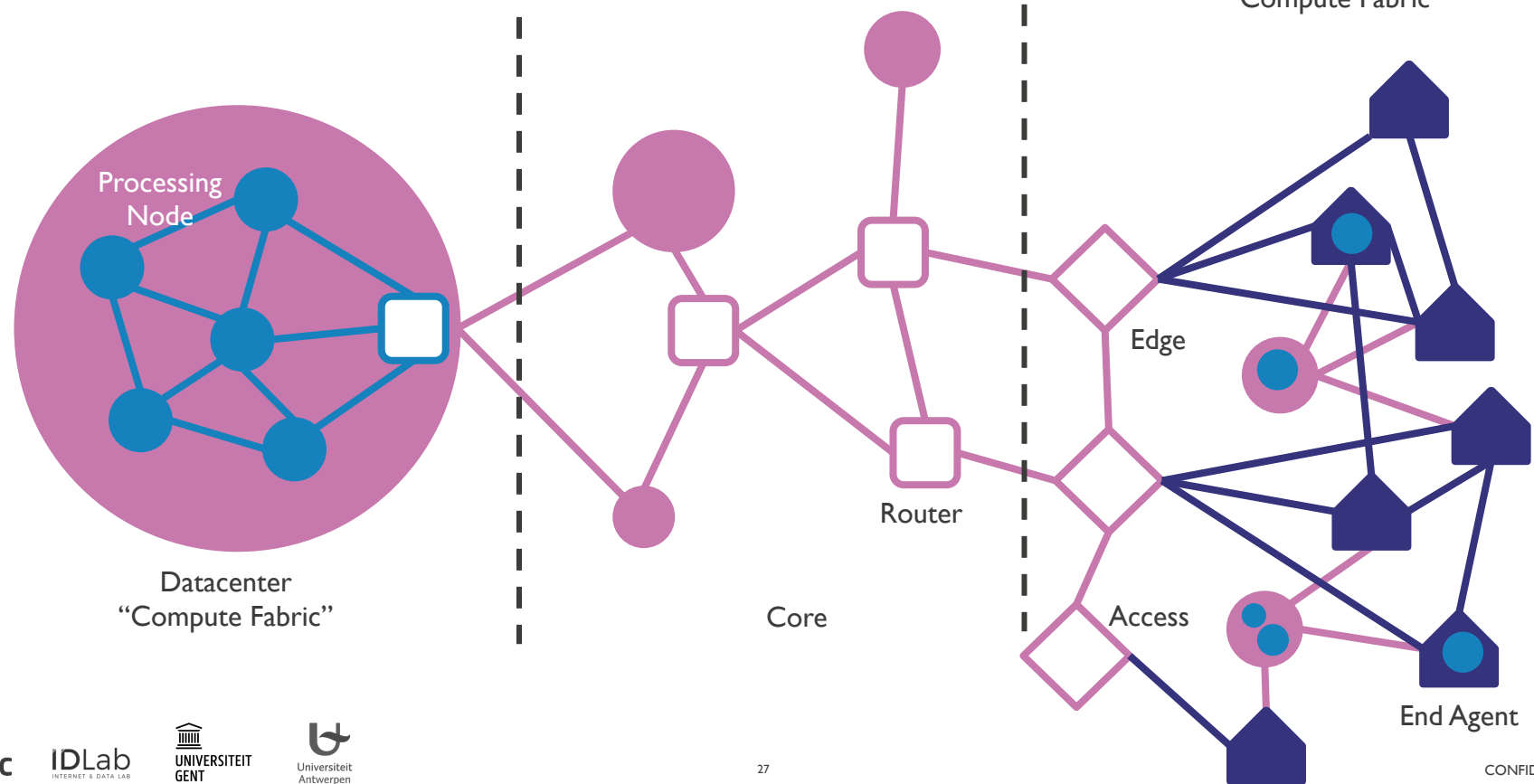
Does a shift make sense?

Current paradigm: data-center-centric.



Does a shift make sense?

Future paradigm: agent-centric





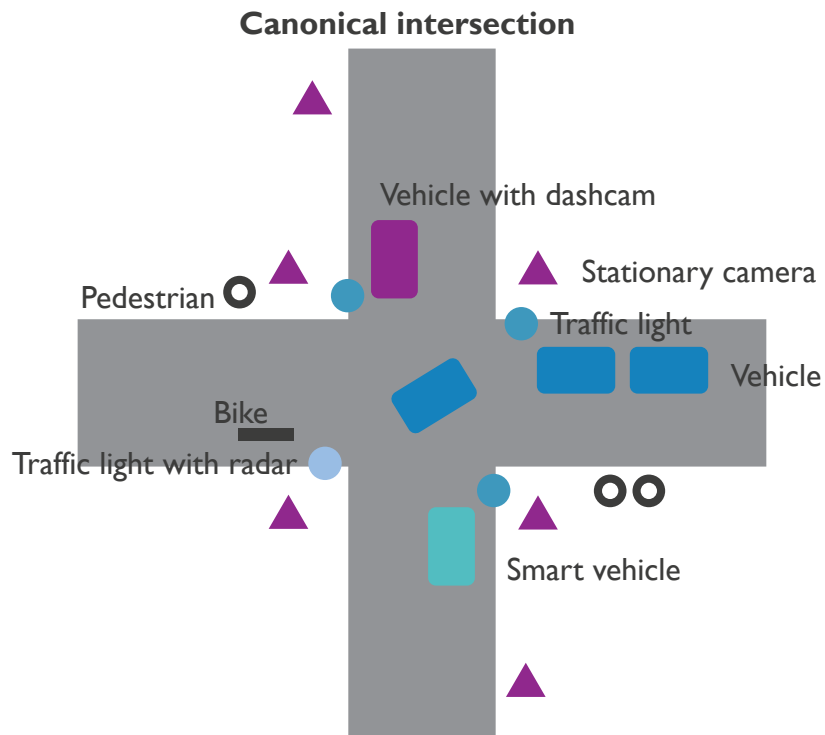
RadioShack

HOUSE
AWESOME
SERVICE
Pearl River, NY
800-735-1500
80 N. MIDDLETOWN RD.

imec

SUPER-VISION

Architecture definition

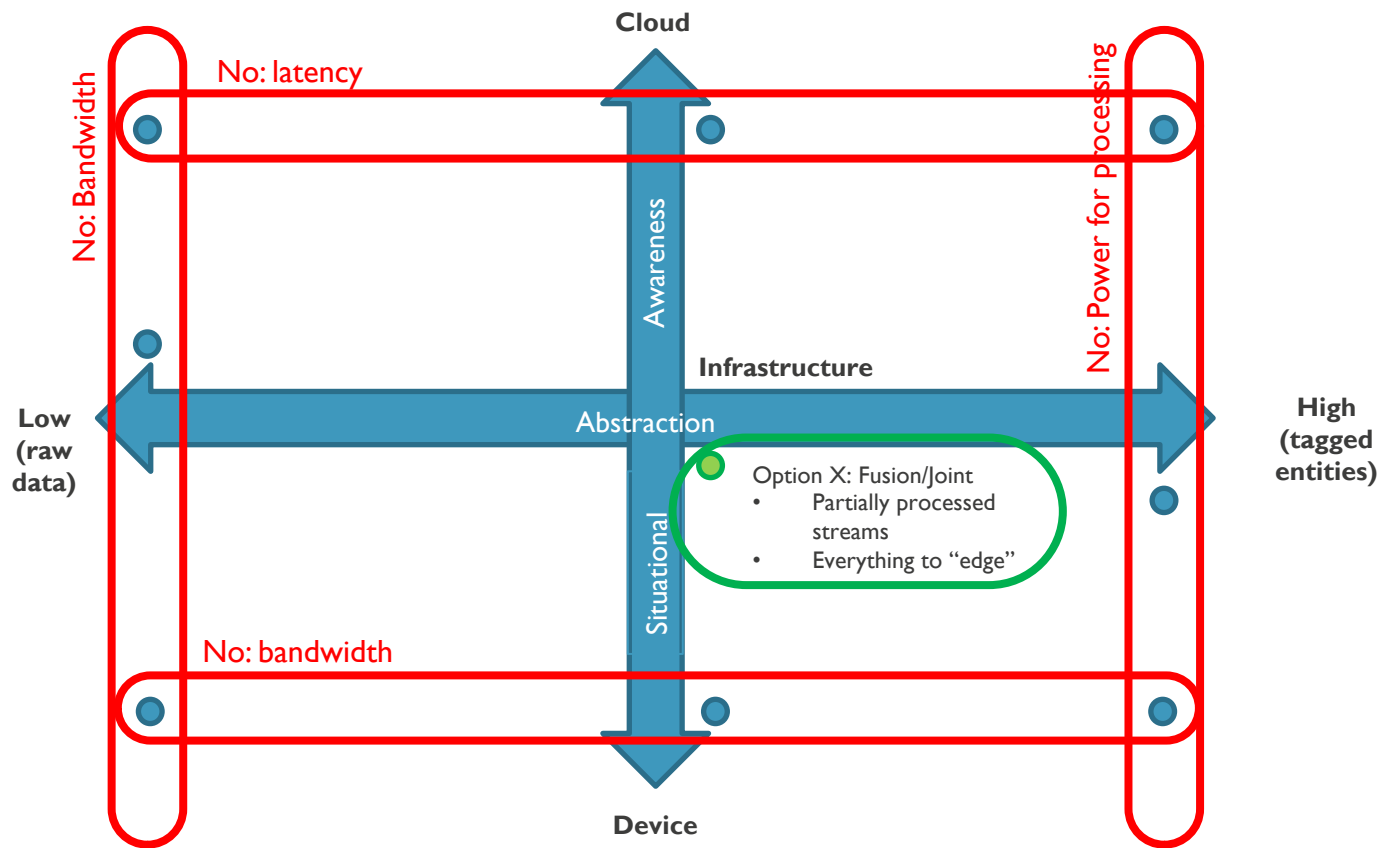


Assumptions:

- Bikes and pedestrians can be smartphone-enabled.
Bikes can be e-bikes
- Vehicles that are smartphone-enabled can have several levels of integration.
- Sensing vehicles (L2+) have multiple cameras and radars
- Local edge connects wirelessly
- Edge-datacenter is wired

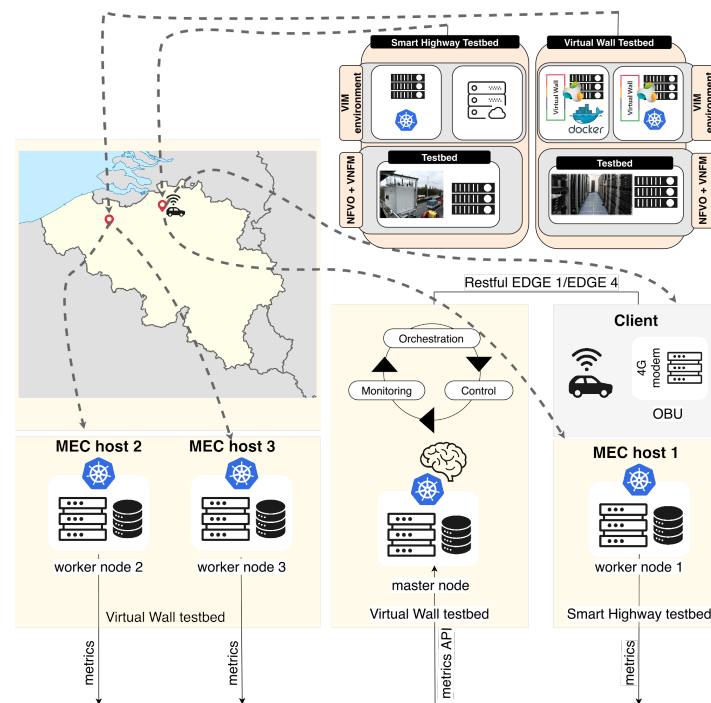
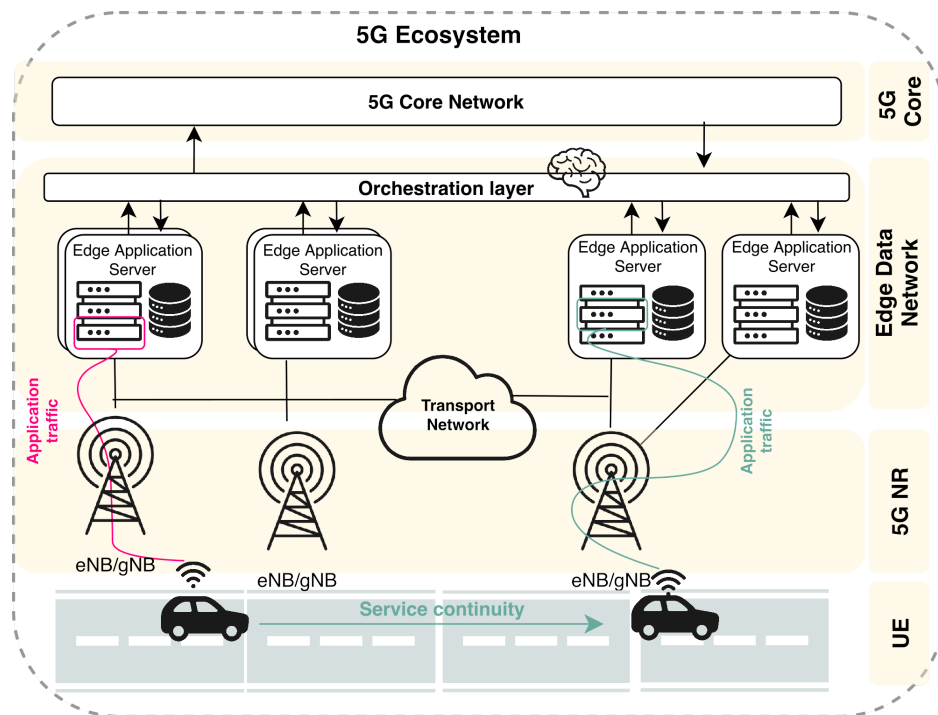
ARCHITECTURE OPTIONS

DEPENDING ON LOCAL PROCESSING AND AWARENESS



Communication is the bottleneck at every layer

Bandwidth AND latency

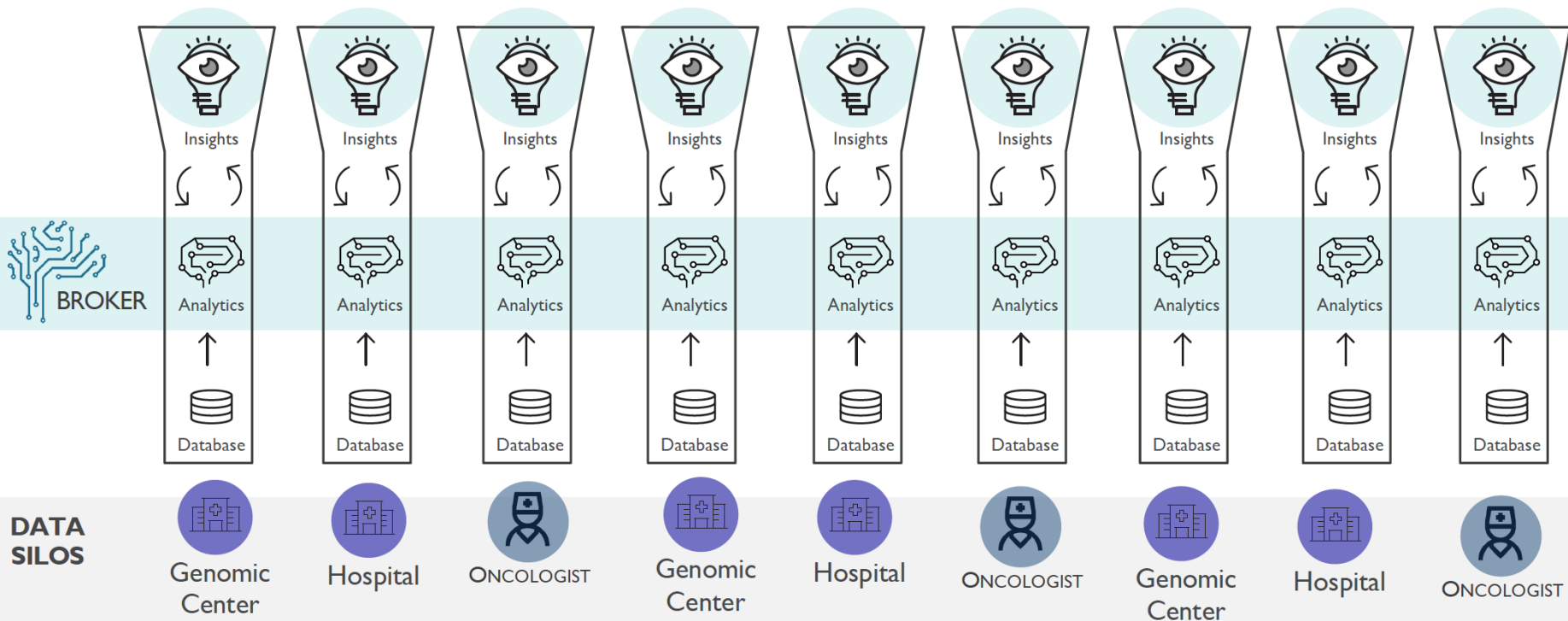


+Distributed
Trust



TRUSTED GENOMICS INSIGHTS PLATFORM

PRIVACY-PRESERVING AMALGAMATED MACHINE LEARNING





SOLID

A personal data vault for sharing data



YOUR POD IS YOUR PERSONAL STORAGE SPACE

Store anything you want in your own Solid POD. PODs are like secure USB sticks for the Web, that you can access from anywhere. When you give others access to parts of your POD, they can react to your photos and share their memories with you. You decide which things apps and people can see.

Think of your Solid POD as your own private website, except that your data interoperates with all your apps, which means you have your own personal API to go along with it. When you post comments or videos online, your friends can view them with whatever app they like, such as an album viewer or a social feed. It's your data, that can be shaped in any way or form.

You can have as many PODs as you like, and they live on Solid enabled Web servers. Install the [Solid Server](#) on your own server at your home or workplace, or [Get a Solid POD](#) from a listed provider.



• The future of computing

- will drive **bandwidth**
 - up to a point...
- is **distributed** at every scale
- needs **trusted** communications



mec

embracing a better life

Thank you to Eric Beyne, Eric Mejdrich, Kathleen Philips, Ilja Ocket, Sri Samavedam and many others...