

Memory Challenges and Solutions in the age of AI

Dr. Martin Brox
Micron Semiconductor (Germany) GmbH
Munich

The Past: Systems of Limited Size

Main consumers of memory

- PC
- Small/Mid scale server
- Cell phones
- Game consoles



The Future: The Age of AI

Germany's Merz inaugurates Nvidia supercomputer for research

By Reuters

September 5, 2025 5:46 PM GMT+2 · Updated September 5, 2025



[1/4] Jupiter, Nvidia's new high-performance computer, during its ceremonial launch, marking the start of operations for the first European supercomputer of the exascale class, at Jülich research center, in Jülich, Germany, September 5, 2025. REUTERS/Jana Rutenfranz Purchase Licensing Rights

Reuters

Jülich Supercomputing Center



Nvidia

Telekom data center with 10,000 Nvidia GPUs being built in Munich

In partnership with Nvidia, Telekom is building an AI data center in Munich and wants to "test the waters" with it. The services will run under the German flag.

heise.de



- Requires hyperscale data-centers
 - Massive amounts of memory
 - Memory is key to performance
 - Stability
 - Serviceability
 - Speed
 - An increased challenge
 - NAND for storage
 - DRAM for compute
- This presentation

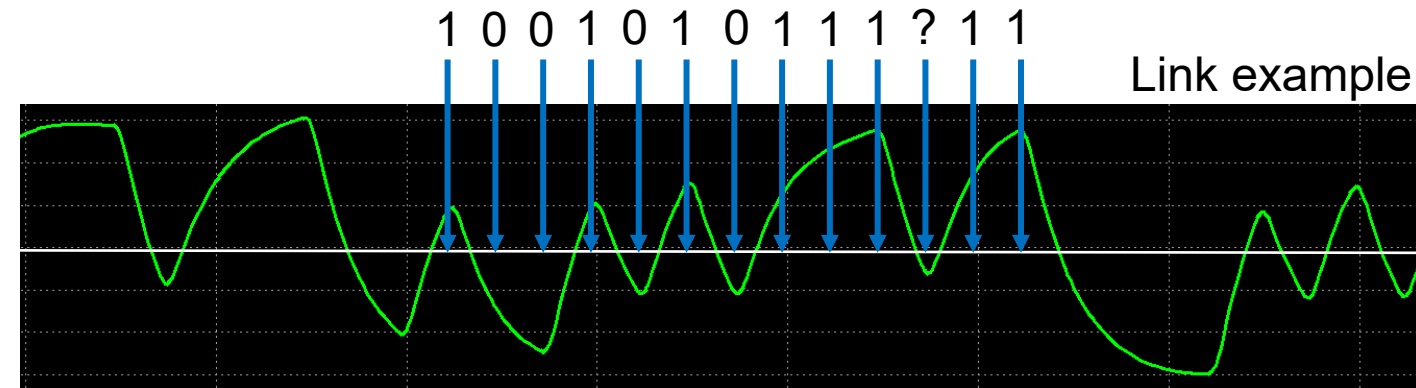
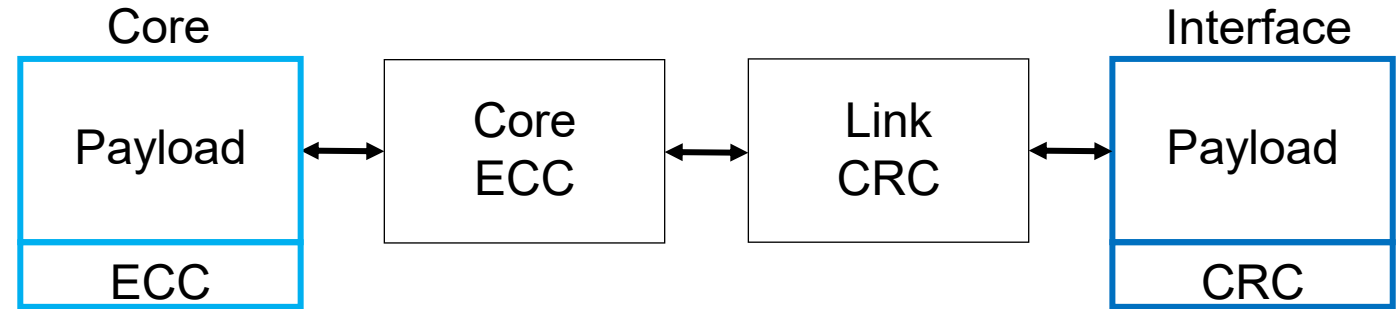
Solutions: Stability

Core ECC

- ECC corrects bit-errors in the DRAM core
- Error correction inside the memory
- Single bit or symbol correction capability
- Multi-bit detection capability

Link errors

- CRC against errors on the interface
- Controller based error detection



Solutions: Serviceability

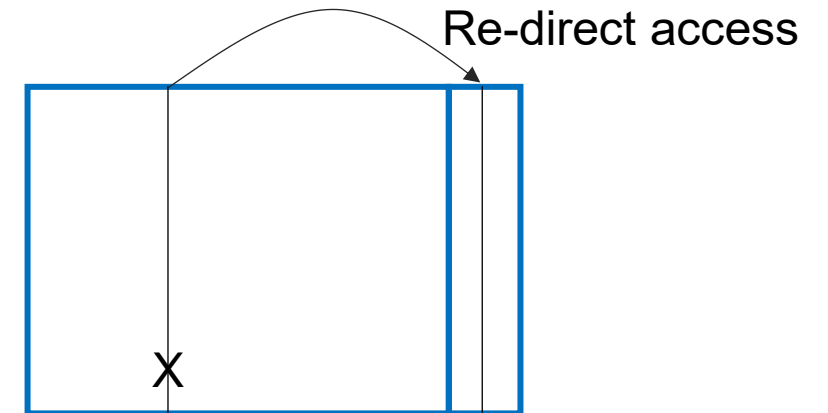
Memory scrubbing

- Periodically interrupt operation by self-repair sequence
- Internally read data and write back after error correction
- Protect against accumulation of bit errors

Operation → Scrub → Operation →

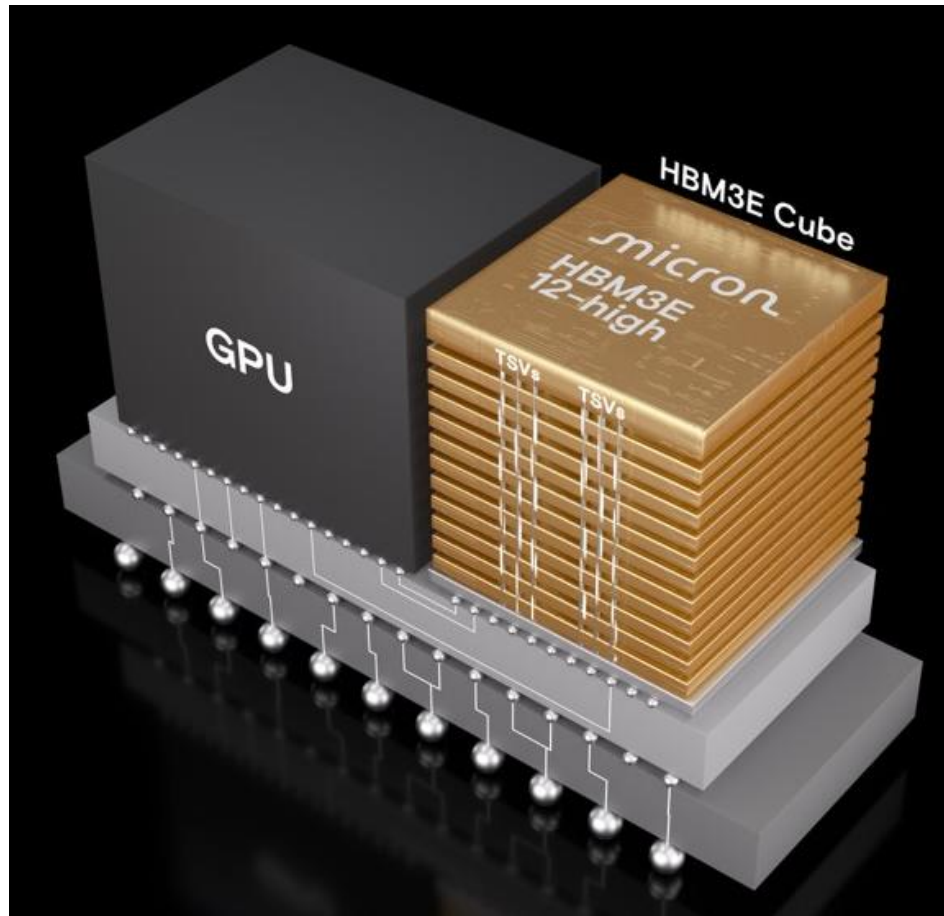
In-system repair

- In rare cases, field errors cannot be corrected by ECC
- Systems are enabled to detect bit errors in the field
 - Memory ECC reports uncorrectable errors
 - System ECC further enhances error detection capabilities
- Internal protocol provided to repair field fails
 - Controller can trigger repair sequence
 - Permanently re-direct access away from field fail



Solutions: Speed

HBM



Pro

- Very high bandwidth per cube
- Very high bit-count per cube
- Perfect fit for AI training
- In demand memory type for AI

Con

- Complex manufacturing
- Complex handling
- Limited supply
- AI inference does not require full HBM performance

→ Industry need for different answers for different tasks

The New Answer: Micron GDDR7

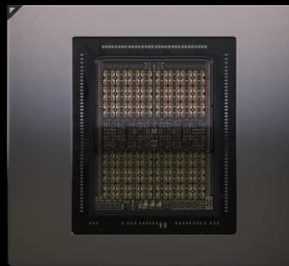
GDDR7 is the new answer
for cost-efficient inferencing

- Conventional discrete component
- Micron added all stability & serviceability features

Nvidia keynote at AI Infra Summit 2025

Announcing Rubin CPX—GPU for Massive Context

Purpose-built CUDA GPU for Context Processing



- **Context Compute**
30 PF NVFP4 CUDA GPU
- **Attention Acceleration**
3x Exponent Operations vs GB300
- **Memory Optimized**
128GB GDDR7
- **Generative Video Ready**
4 NVENC and 4 NVDEC
- **Available End of 2026**

Nvidia keynote at AI Infra Summit 2025

AI Infra Summit—NVIDIA® today announced NVIDIA Rubin CPX, a new class of GPU purpose-built for massive-context processing. This enables AI systems to handle million-token software coding and generative video with groundbreaking speed and efficiency.

Advancements Offered by Rubin CPX

Rubin CPX delivers up to 30 petaflops of compute with NVFP4 precision for the highest performance and accuracy. It features 128GB of cost-efficient GDDR7 memory to accelerate the most demanding context-based workloads. In addition, it delivers 3x faster attention capabilities compared with NVIDIA GB300 NVL72 systems — boosting an AI model's ability to process longer context sequences without a drop in speed.

History of GDDR

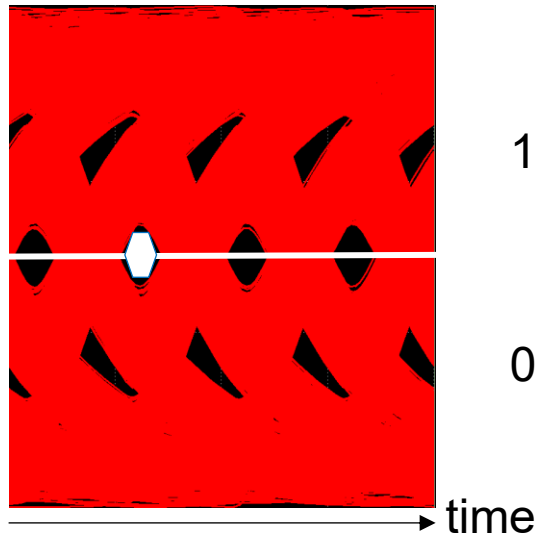
Innovation Inspired by Micron Semiconductor Germany

Generation	Examples of Usage	Standardization	Importance
GDDR3	Consoles Gfx Cards	JEDEC	1 st high volume GDDR
GDDR5	Consoles Gfx Cards	JEDEC	Clock architecture
GDDR5X	Gfx Cards	Micron/Nvidia	I/O equalization & PLL
GDDR6	Consoles Gfx Cards	JEDEC	"
GDDR6X	Gfx Cards	Micron/Nvidia	Multi-level signaling
GDDR7	Gfx Cards → AI	JEDEC	"

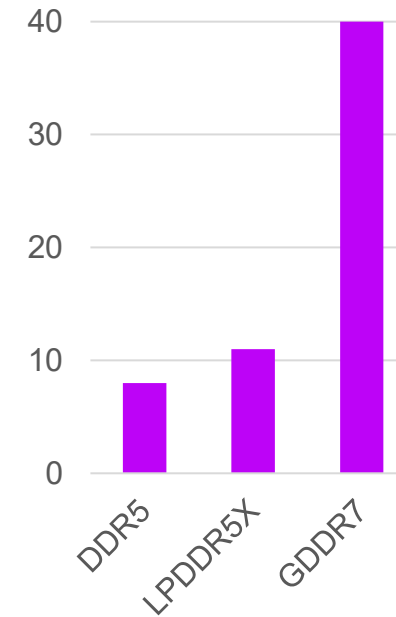
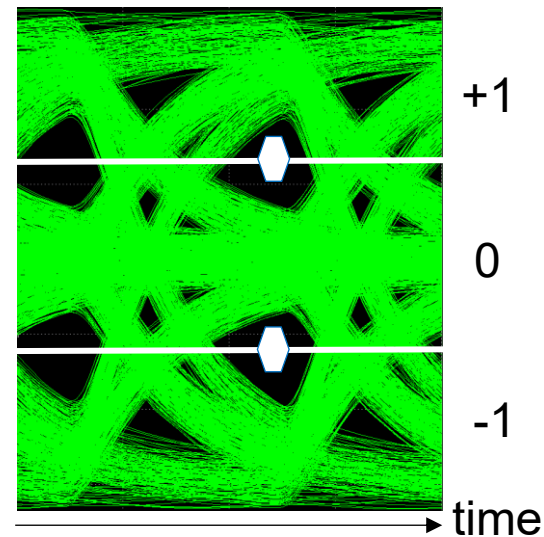
Micron GDDR7: Designed for the speed of AI

- JEDEC introduced **multi-level signaling** to GDDR7
 - Introduced 3-level signaling in combination with 10 DQ per byte of payload data
 - Increases valid data area for link compared to simple doubling of frequency
- Designed to reach 5x bandwidth/pin compared to server-class memory

Conventional memory



GDDR7

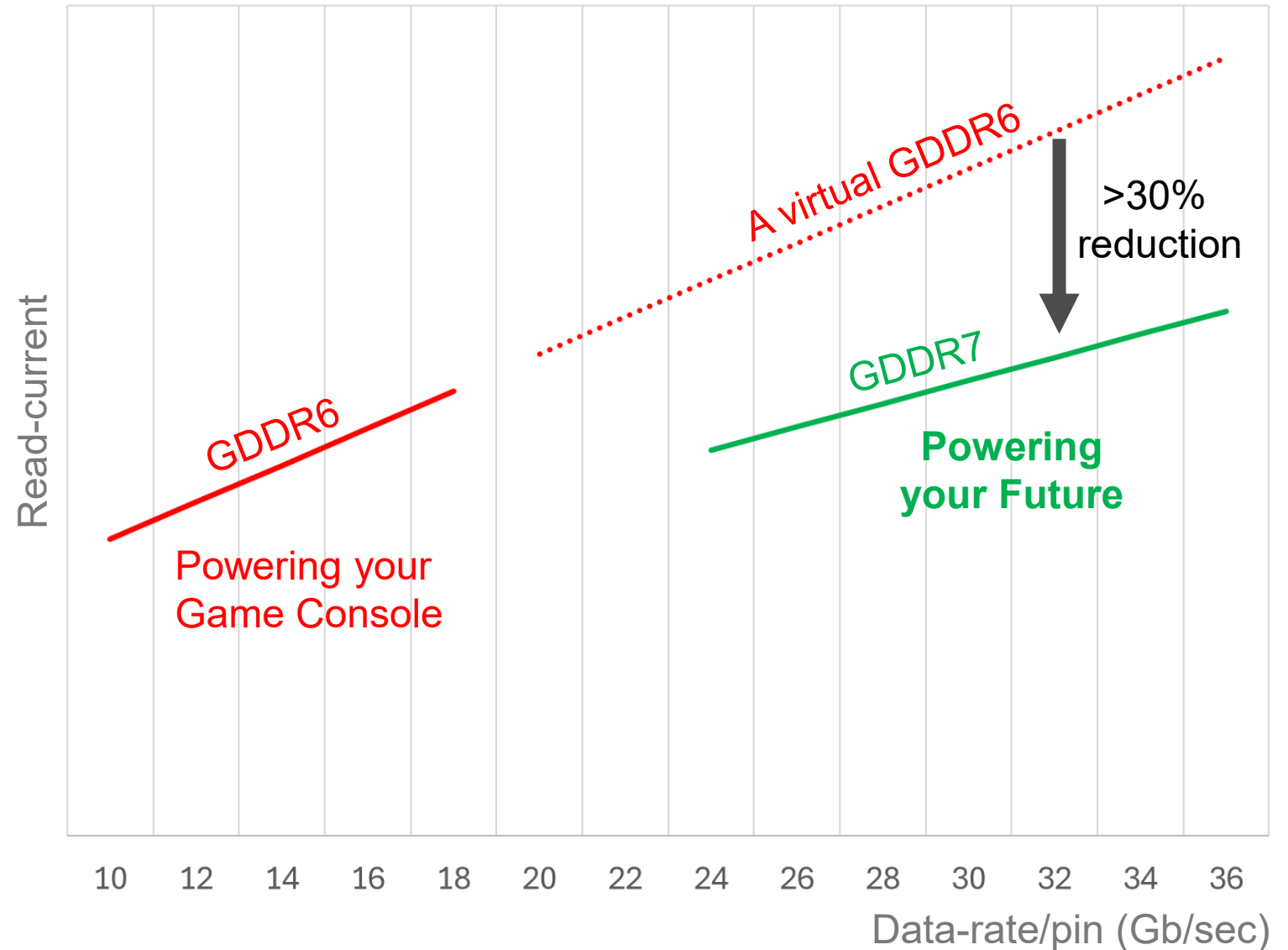


Link examples

 Valid data area

Micron GDDR7: Designed for Low Power

- New interface definition
 - Multi-level signaling lowers clock frequency
- Advanced design
 - Internal low-swing data-transfer
- Advanced manufacturing
 - Leading edge wafer process



Summary

- AI is triggering the build-up of hyperscale data-centers
- AI requires leading edge memory with highest data-rates and lowest power
- AI memory needs to be robust
- Micron has the answers
 - HBM for AI training
 - GDDR7 for AI inference

The Micron logo, featuring the word "micron" in a lowercase, sans-serif font. The letter "m" is stylized with a horizontal bar that extends to the left.

© 2025 Micron Technology, Inc. All rights reserved. Information, products, and/or specifications are subject to change without notice. All information is provided on an "AS IS" basis without warranties of any kind. Statements regarding products, including statements regarding product features, availability, functionality, or compatibility, are provided for informational purposes only and do not modify the warranty, if any, applicable to any product. Drawings may not be to scale. Micron, the Micron logo, and other Micron trademarks are the property of Micron Technology, Inc. All other trademarks are the property of their respective owners.