Flash Memory Summit

Main Memory: Its Future Direction in the Hybrid Cloud & AI

Ju Jin An, Adam McPadden, Jung Yoon

IBM Systems

OMEM-201-2: DRAM, Part 2

8:45AM, 8/3/2022

35 YEARS OF NAND FLASH MEMORY INNOVATION

IT Infrastructure and Hybrid Cloud

: Combined components needed for the operation and management of enterprise IT services and IT environments

Flash Memory Summit

Traditional infrastructure

- Made up of hardware and software components: facilities, data centers, servers, networking hardware desktop computers and enterprise application software solutions
- A traditional infrastructure is typically installed on-premises for company-only, or private, use

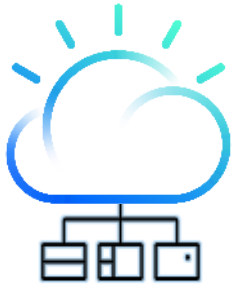
Cloud infrastructure

- End users can access the infrastructure via the internet, with the ability to use computing resources without installing on-premises through virtualization
- Virtualization connects physical servers maintained by a service provider at any or many geographical locations

Hybrid Cloud

- A mix of on-premises infrastructure, private cloud services, and public cloud services

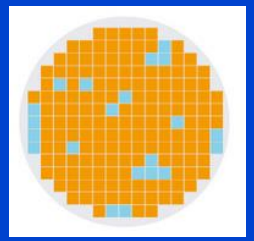
Hybrid Cloud / AI through the lens of DRAM



- System Infrastructure is evolving from compute centric to data centric
  - More parallel processing and a massive increase in the size of neural networks of AI require processors to access more data from memory with low latency
  - Memory with high bandwidth/capacity and low latency is important to run AI applications in hybrid cloud environment
- DRAM technology is critical in Hybrid Cloud & AI technology
  - DRAM has been served as main memory subsystem thanks to superior performance/cost in Hybrid Cloud infrastructure
  - System performance & cost leadership was enabled by successful DRAM technology scaling

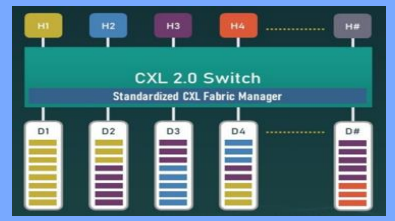
- DRAM technology scaling encountered challenges (e.g., Cs, Sensing Margin, WL-WL interference, etc.)

- DRAM quality is important to minimize “unplanned downtime” of hybrid cloud infrastructure
- Focus on ‘Shift Left’ understanding of reliability mechanisms of advanced DRAM node

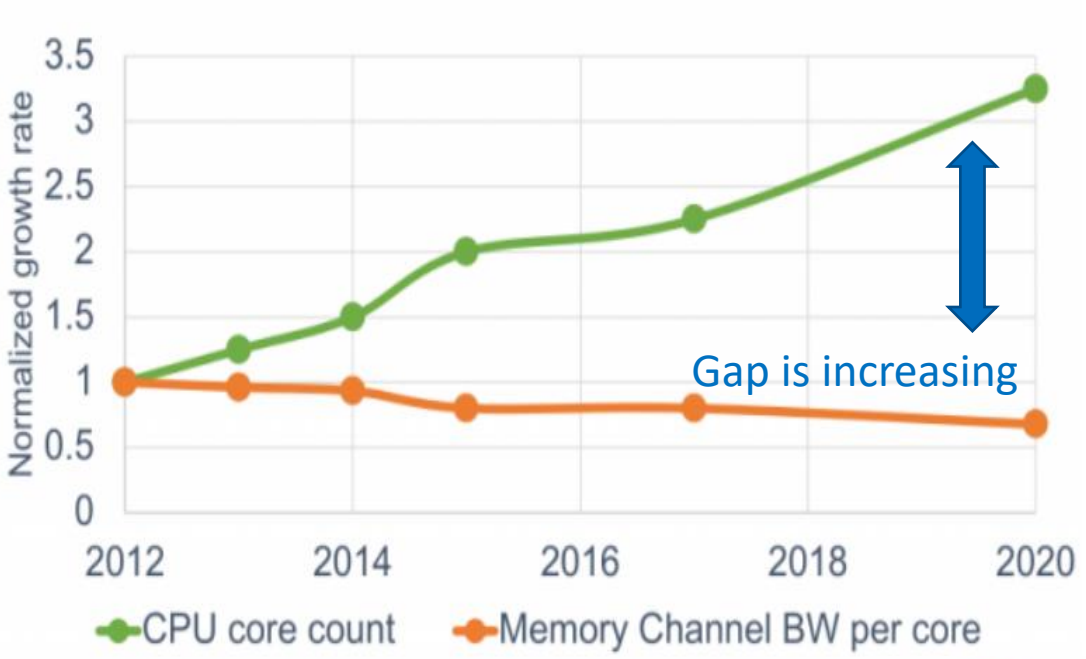


- Differential DIMM (DDIMM) in IBM power and z enterprise server improves main memory’s bandwidth, latency and capacity via OMI (Open Memory Interface)

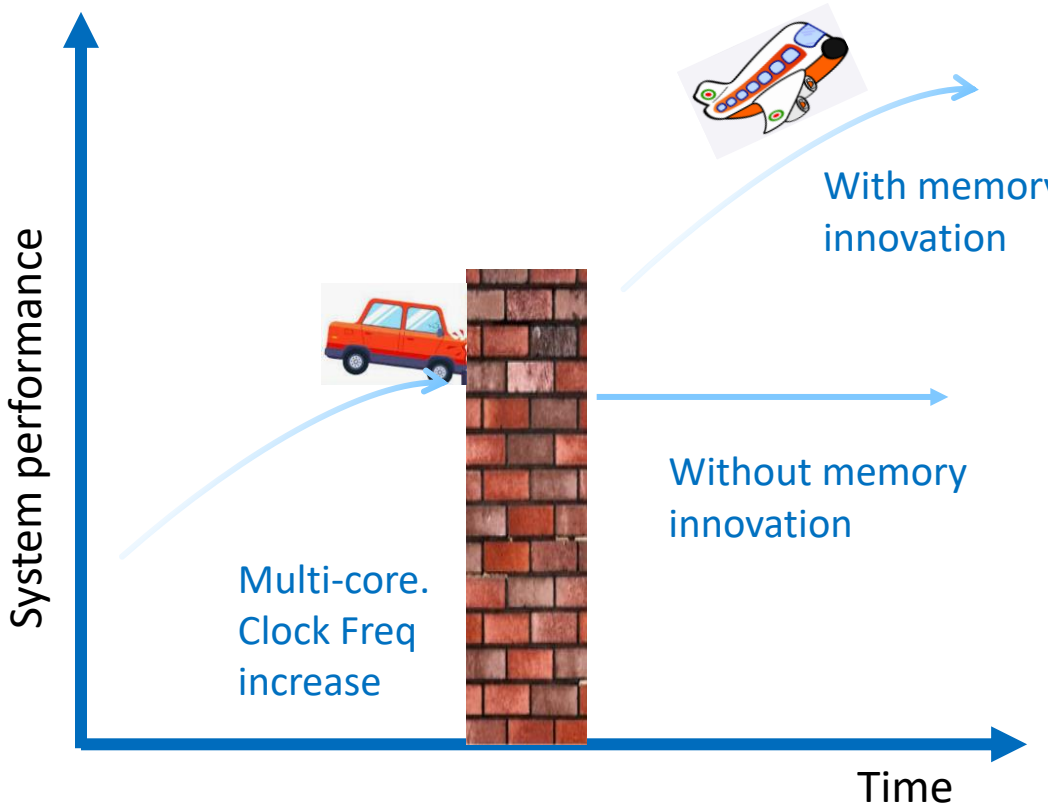
- Compute Express Link (CXL) will open a new era of server architecture
- Discrete units of compute, memory and storage resources are dynamically allocated to meet the requirement of workload



Memory Wall



- Source : Meta, OCP Global Summit, Nov 2021

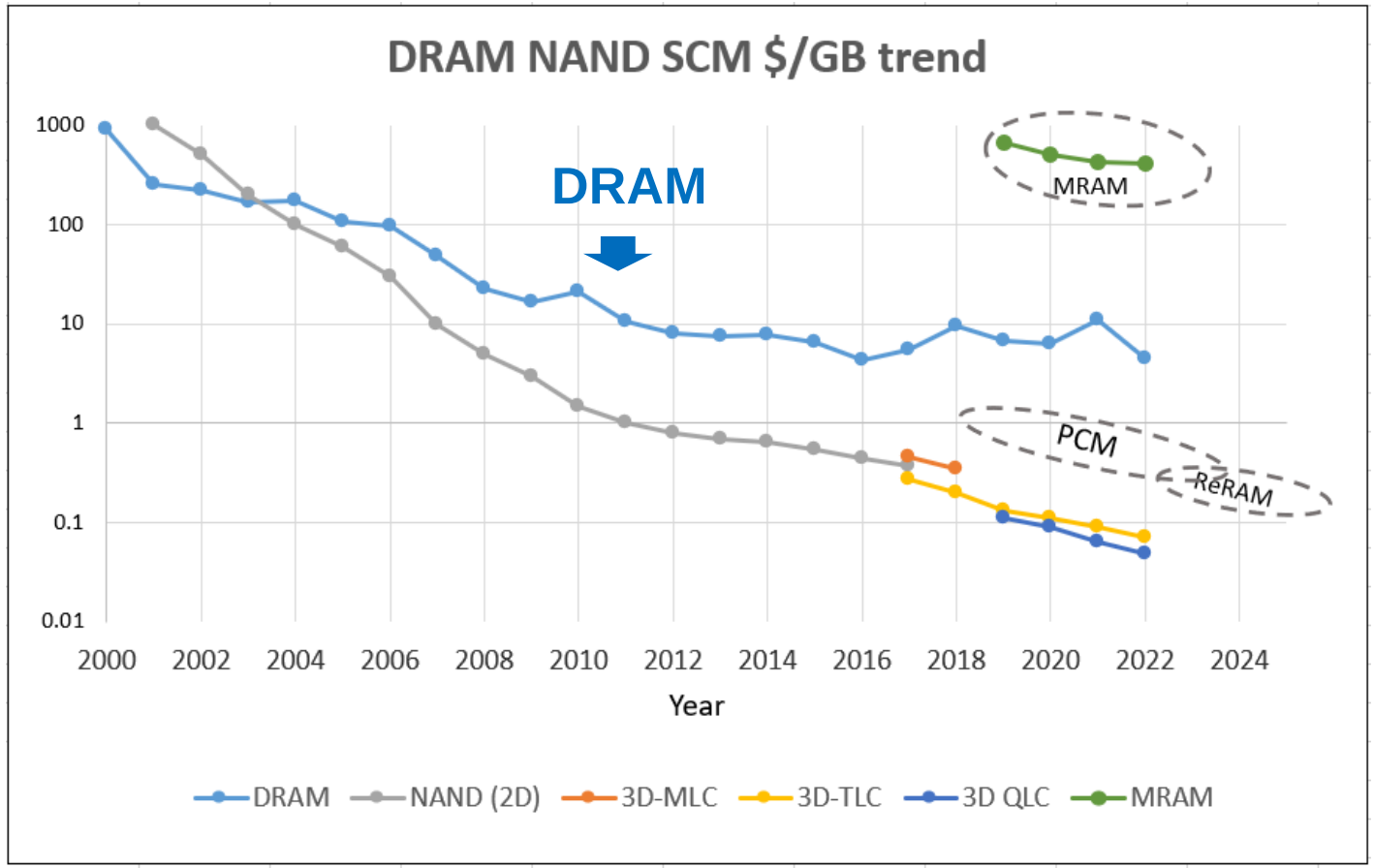


- Increasing CPU core counts is driving memory demand with increased bandwidth and capacity
- More parallel processing and a massive increase in the size of neural networks of AI require processors to access more data from memory in less time

Economics of DRAM scaling

Flash Memory Summit

- Cost
- Performance
- Power
- Quality

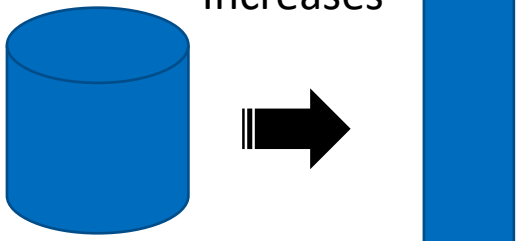


Source: **R. Nair and J. Yoon**, "7 - The evolving role of storage-class memory in servers and large systems," in Semiconductor Memories and Systems, A. Redaelli and F. Pellizzer, Eds. Woodhead Publishing, 2022, pp. 217–251.

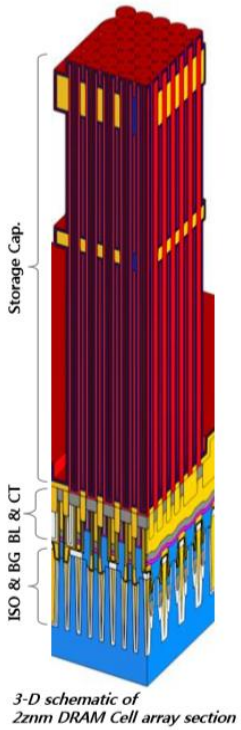
- DRAM technology scaling is driving the dominant use of DRAM as a main memory in computing systems
  - DRAM scaling enabled the reduction in the device power consumption and cost while improving the device performance
- Historical bit cost reduction at ~30% per technology node no longer holds at sub 1x nm due to challenges with technology scaling

35 YEARS OF NAND FLASH MEMORY INNOVATION

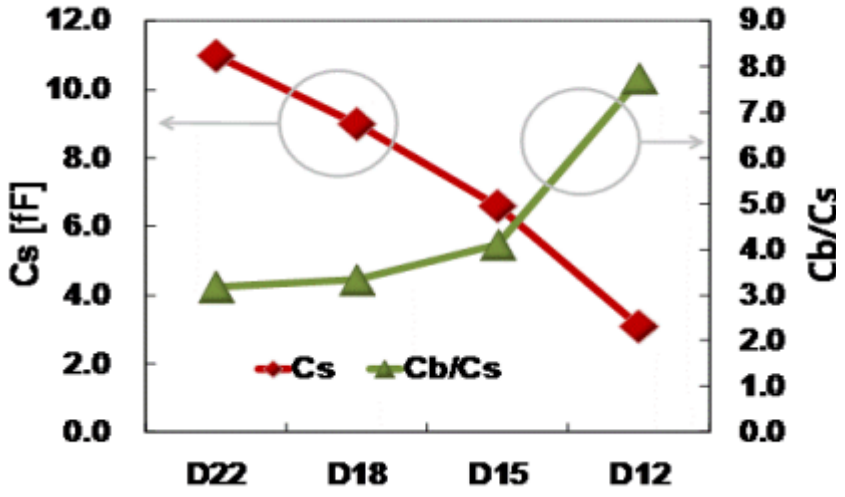
Scaling challenges of DRAM technology



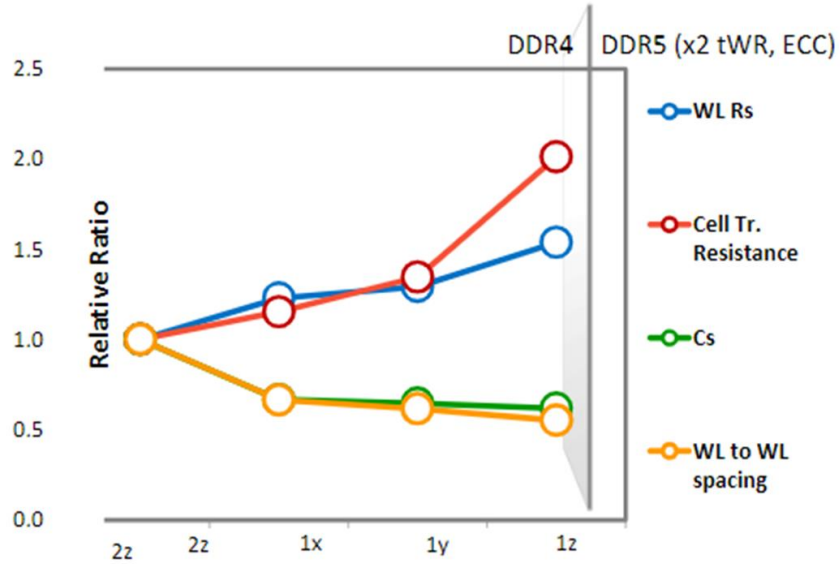
Capacitor AR Increases



3-D schematic of 2znm DRAM Cell array section



Hwang, Y., Park, J., Jin, G.Y., & Chung, C. (2012). An Overview and Future Challenges of High Density DRAM for 20 nm and Beyond.



SK Hynix

- Capacitor scaling challenges (Cs. Cb)
  - Cell capacitor (Cs) is decreasing because of smaller dimension of the capacitor
  - AR (Aspect Ratio) of physical cell capacitor is increased to boost capacitance in addition to high k dielectric material adoption

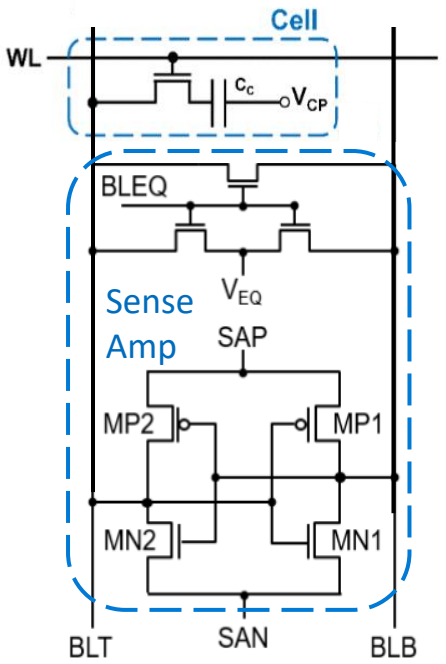
- Retention time continues to decrease with cell capacitance (Cs) reduction and leakage current increase
  - Main source of leakage current is GIDL and junction leakage

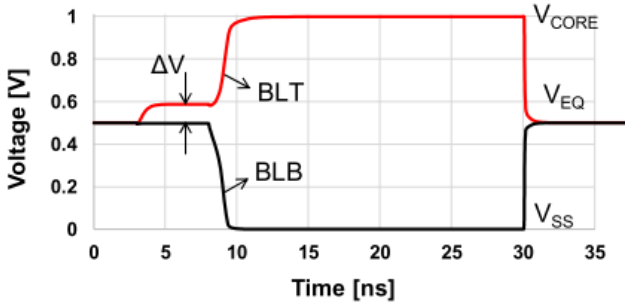
$$t_{ref} \propto \frac{Cell\ Capacitance}{I_{leakage}}$$

- WL resistance, Cell Transistor resistance increase slows down the operation of DRAM due to RC delay
- Limited WL to WL spacing
  - RowHammer
- V_{TH} variation driven by SCE (Short Channel Effect)

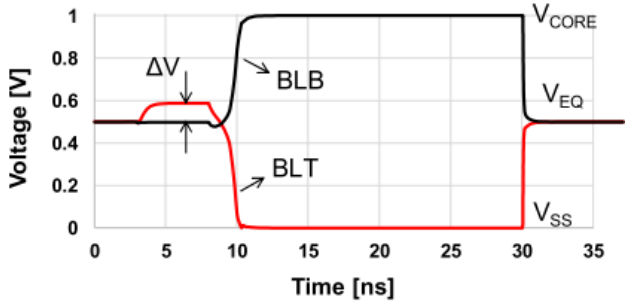
Scaling challenges of DRAM technology

$$\Delta V = \frac{V_{cell} - V_{EQ}}{1 + Cb/Cs}$$





Successful Data sensing



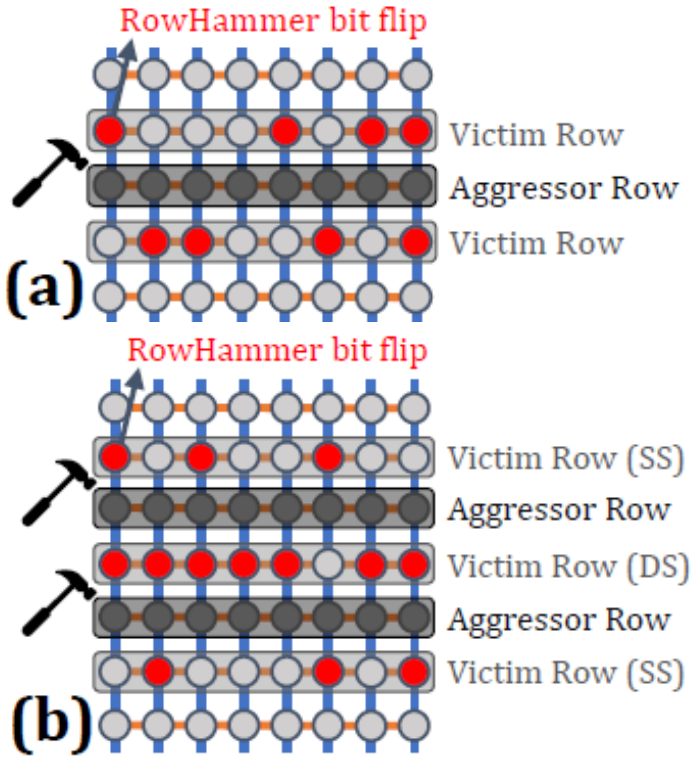
Failing Data sensing

- Sensing voltage difference (ΔV) ↓↓
  - Voltage difference between a charge shared BL and a reference BL

- Offset Voltage (V_{offset}) ↑↑
  - Fab process variation and random V_{TH} variation increase significantly with technology scaling
  - V_{TH} mismatches between paired transistors in sense amplifier increase the offset voltage (V_{offset})

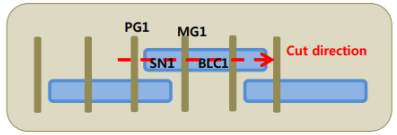
- Sensing margin (ΔV - V_{offset})
  - Sensing margin is required to be large enough to maintain successful data sensing in DRAM operation
- DRAM technology scaling is moving in the negative direction from sensing margin perspective

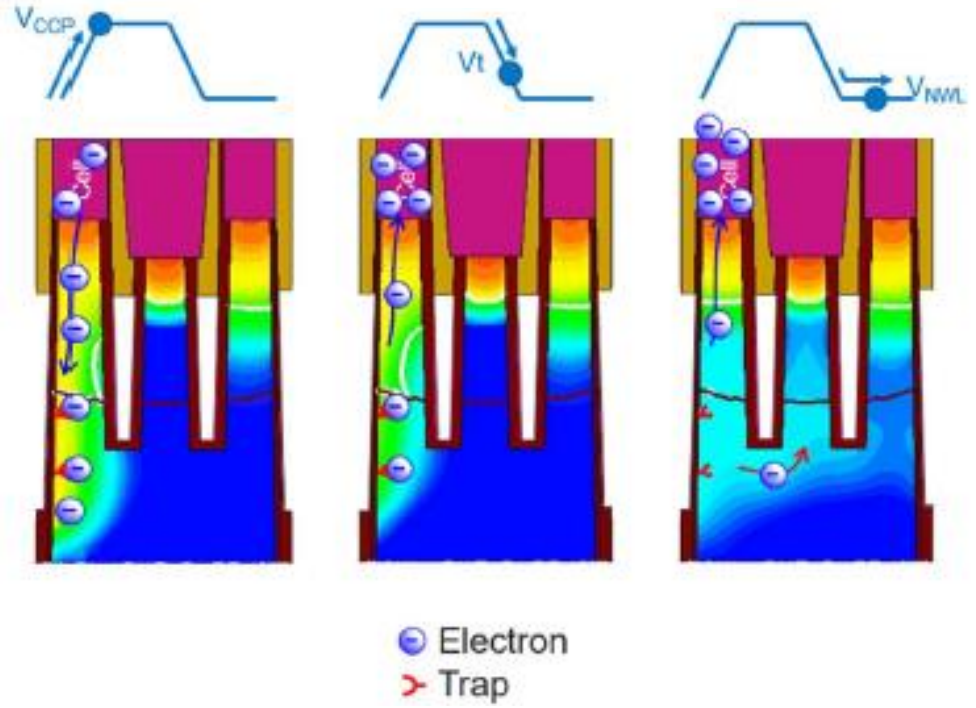
RowHammer - potential security exploits



H. Hassan, Y. C. Tugrul, J. S. Kim, V. van der Veen, K. Razavi, and O. Mutlu, "Uncovering In-DRAM RowHammer Protection Mechanisms:A New Methodology, Custom RowHammer Patterns, and Implications," in MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture, New York, NY, USA, 2021, pp. 1198–1213. doi: 10.1145/3466752.3480110.

DRAM cell (top-down)



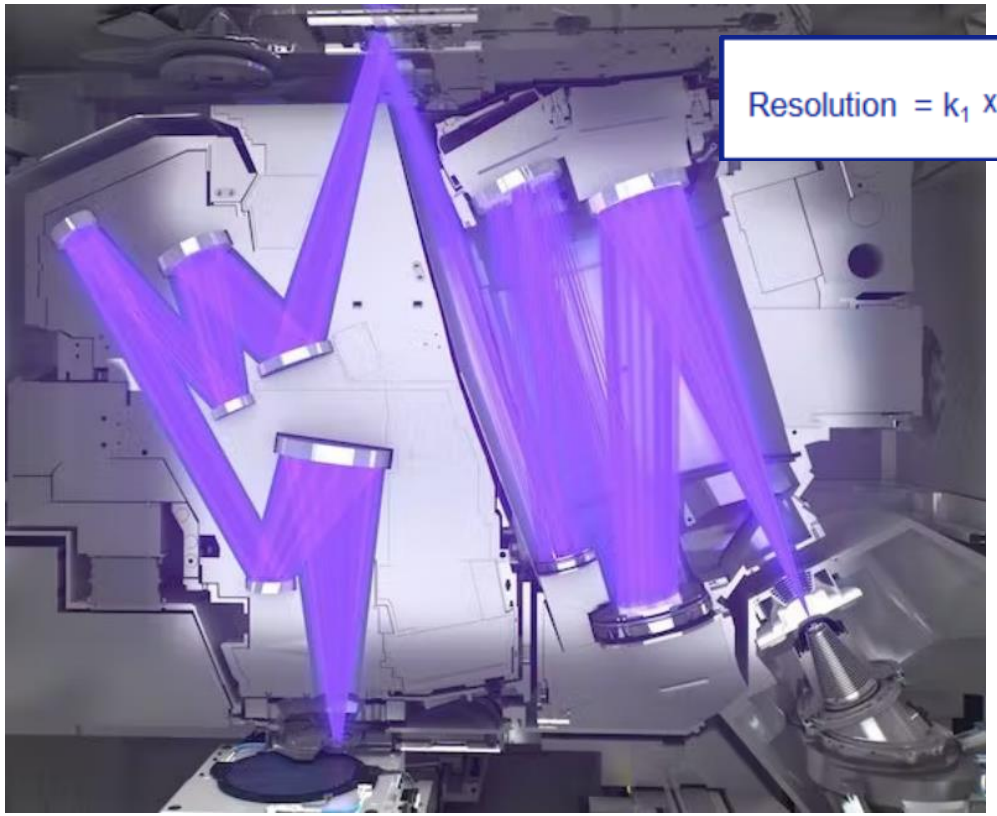


S. Shiratake, "Scaling and Performance Challenges of Future DRAM," 2020 IEEE International Memory Workshop (IMW), 2020, pp. 1-3, doi: 10.1109/IMW48823.2020.9108122.

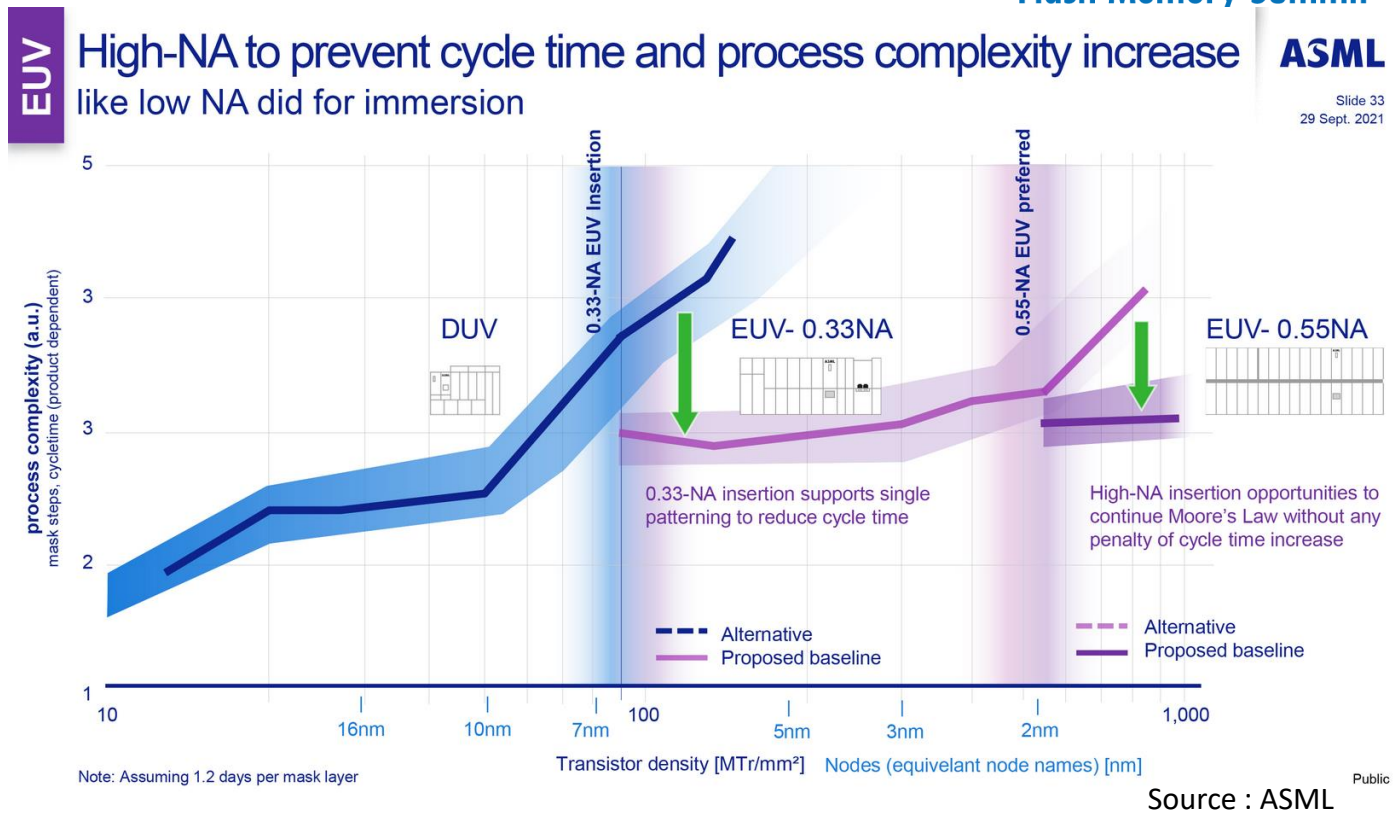
- When aggressor rows are frequently attacked, neighboring cells (victim rows) leak charge at a faster rate than expected, losing their data before the next refresh
- Reliability/Security issue : allowing unprivileged attackers to modify sensitive data without accessing target DRAM cell

- **Mechanism** : Strayed electrons are released from trap site due to frequent cell activation, potentially inducing cell charge gain and increase error rates
- **Root cause** : WL-WL interference due to limited space between DRAM cells due to DRAM technology scaling

EUV for future DRAM technology scaling



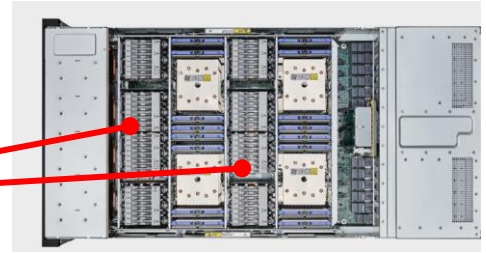
$$\text{Resolution} = k_1 \times \frac{\lambda}{NA}$$



Source : ASML

- DRAM suppliers have started implementing EUV process in their critical process layers of the latest DRAM products to reduce the cost and complexity derived from dual/quadruple patterning
- Higher NA (0.33 → 0.55) enables even smaller feature size patterning (13nm → 8nm)

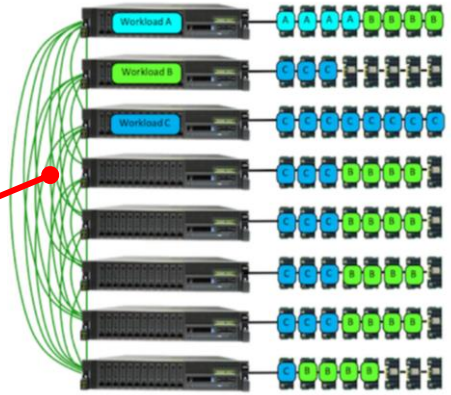
New Power10 Memory Technologies



Differential DIMM Technology (DDIMM) in IBM Power and Z enterprise server
Improving main memory's bandwidth, latency and capacity

OMI (Open Memory Interface) - Providing solutions for near memory computing
Processor communicates with memory through memory controller using differential serial channels

Memory Clustering – Distributed Memory Disaggregation and Sharing
On-demand memory allocation based on workload



Better memory RAS than IS DIMMs (redundancy)†

2x higher memory bandwidth than scalable x86 processors

DDR4 running at up to 3200 Mbps data rate provides 409 GB/s peak memory bandwidth per socket

Transparent memory encryption with no performance impact

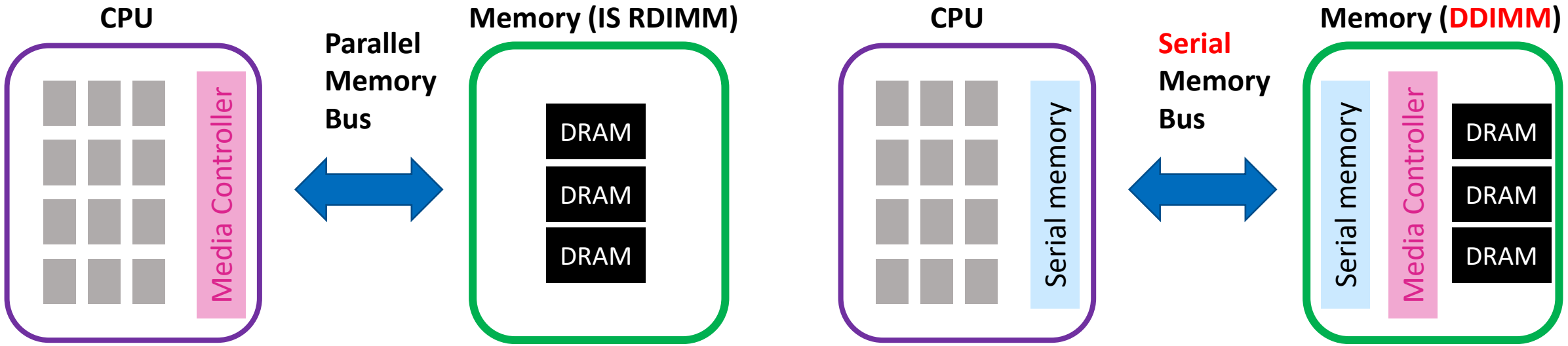
Chipkill technology with advanced ECC protects from memory chip failure

Active Memory Mirroring (AMM) feature supported – Mirrors hypervisor memory to provide resiliency from uncorrectable memory errors

† Applied to mid-range/high-end Power Systems

DDIMM (Differential DIMM) in IBM power/z server

- Reduced number of pins (288 vs 84) enabled high memory bandwidth with minimum latency
- OMI connects standard DDR DRAMs to a host CPU using high-speed serial signaling and “provides near-HBM bandwidth at larger capacities than are supported by DDR.”

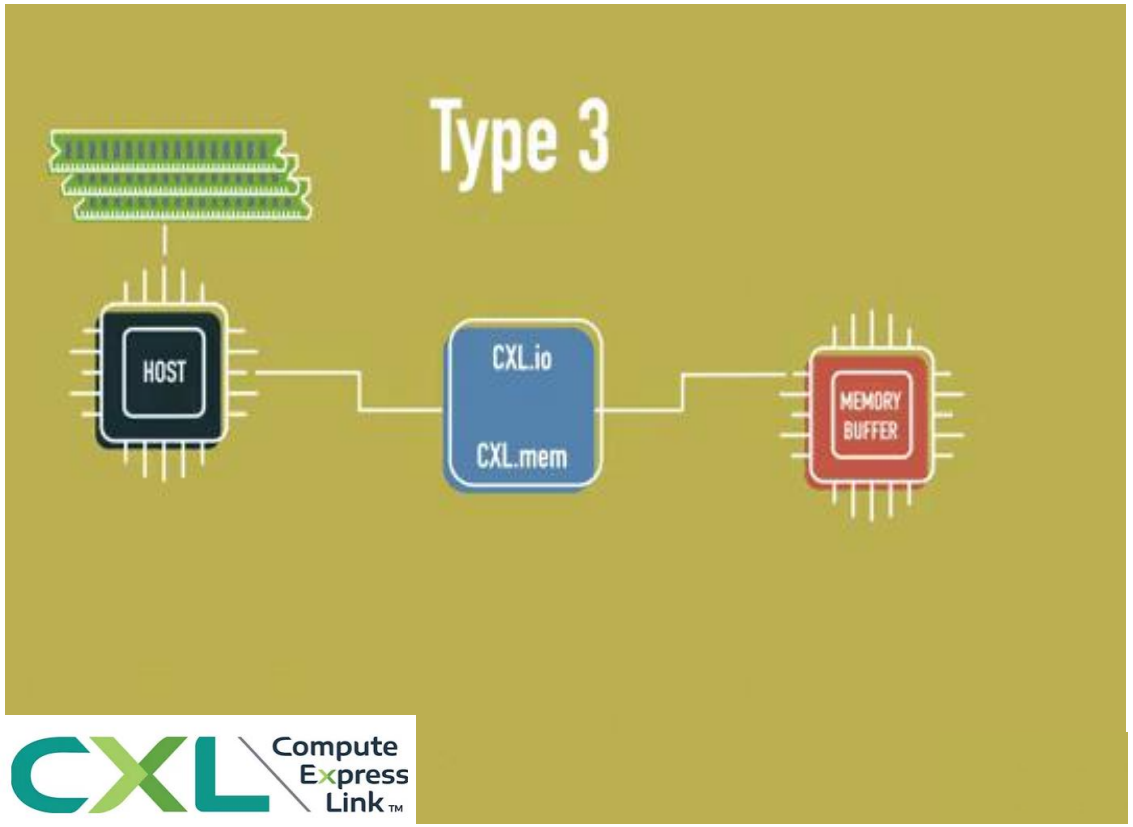


VS.

- 288 Pins per memory channel (limited number of memory channel)
- Media Controller is in SOC
- SOC update is required whenever media controller is updated

- 84 Pins per memory channel enables more memory channels attached to CPU
- Media Controller moves from CPU to DDIMM
- In addition to DRAM, persistent memory can be attached using the same differential interface

Compute Express Link™ (CXL™)

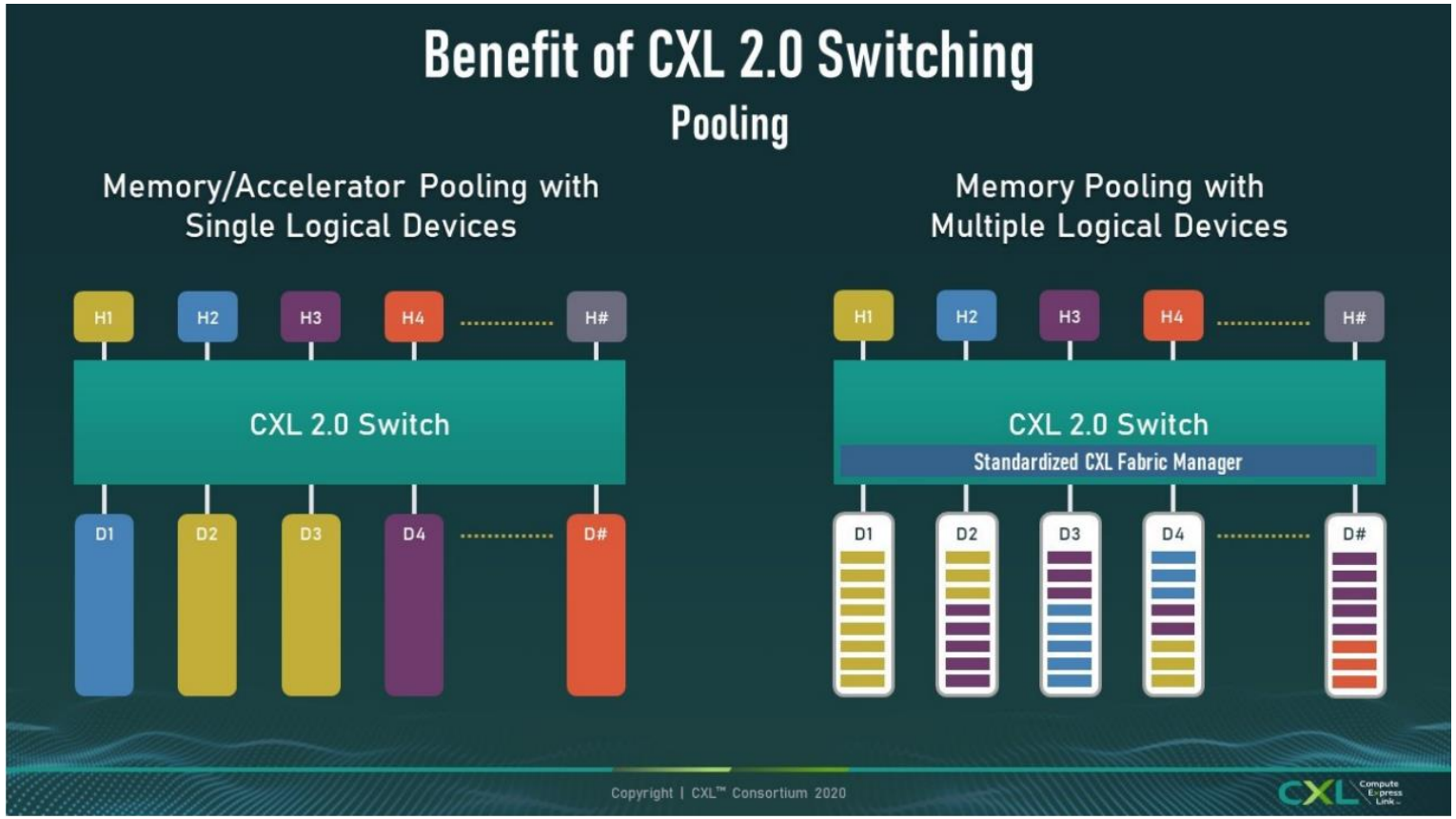


- Compute Express Link™ (CXL™) is an industry-supported Cache-Coherent Interconnect for Processors

- CXL is built on the PCI Express (PCIe) physical/electrical interface
- 3 CXL protocols
  - PCIe-based input/output protocol (CXL.io)
  - Cache-coherent protocols for accessing system memory (CXL.cache)
  - device memory (CXL.mem)

- CXL type 3 device memory increases memory bandwidth and capacity with CXL.mem and CXL.io protocol

CXL 2.0 and beyond for future system architecture



Resource Pooling

Disaggregation & Composability

- CXL 2.0 specification adds support for switching, enabling resource pooling for increased memory utilization efficiency and providing memory capacity on demand

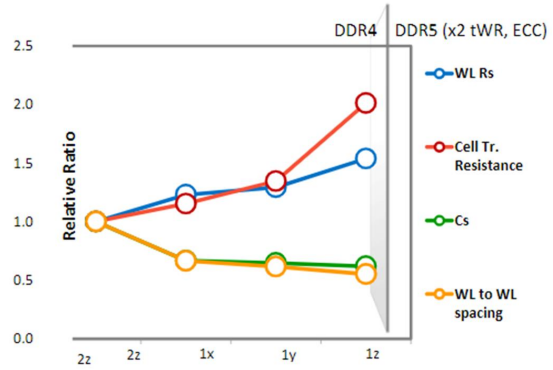
Flexible main memory expansion

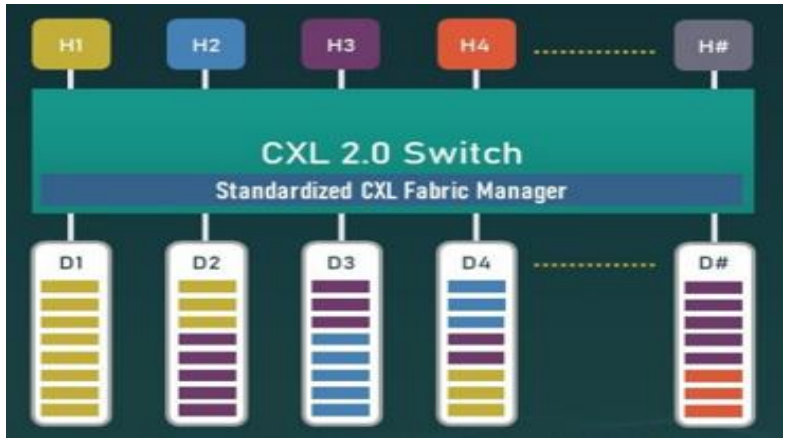
- Memory resource pooling and dynamic allocation/de-allocation capabilities are powerful tools to achieve the performance benefits of main memory expansion with efficiency and total cost of ownership (TCO) benefits

Media Independent

- CXL enables users to attach memory devices with variable latencies/media types (e.g., persistent memory, DDR4, DDR5, etc) while maintaining coherent interface

Summary





CXL Compute Express Link™

1 DRAM technology scaling

Low cost, low latency, high bandwidth memory is an essential element of hybrid cloud infrastructure and DRAM scaling has been meeting such requirements
However, DRAM scaling encountered challenges

2 DDIMM and OMI in IBM Power and Z enterprise server

DDIMM improves main memory's bandwidth, latency and capacity with better RAS capability

3 CXL as a critical step for future architecture development

Memory resource pooling and dynamic allocation/de-allocation capabilities to achieve the performance benefits of main memory and total cost of ownership (TCO) benefits

When you interact with IBM, this serves as your authorization to Flash Memory Summit or its vendor to provide your contact information to IBM in order for IBM to follow up on your interaction.

IBM's use of your contact information is governed by the IBM Privacy Policy.

When you interact with IBM, this serves as your authorization to Flash Memory Summit or its vendor to provide your contact information to IBM in order for IBM to follow up on your interaction.

IBM's use of your contact information is governed by the IBM Privacy Policy.