PCIe FLASH and Stacked DRAM for Scalable Systems

Joe Jeddeloh
Director
Controller Development
Micron Technology, Inc.

Presentation Overview

- Introduction
- HMC Architecture
- PCIe SSD
- Future Trends

The Memory / Storage Performance Wall

- We started talking about the memory wall in the ‘90s; it’s always been “a few years out...”
  - Why is this time different?
  - → Bandwidth AND Power must be addressed
- System Power limits have constrained performance; performance can grow if power is addressed
- Storage = Dollar, Watts/IOP
- Memory bandwidth = PJ/bit transferred

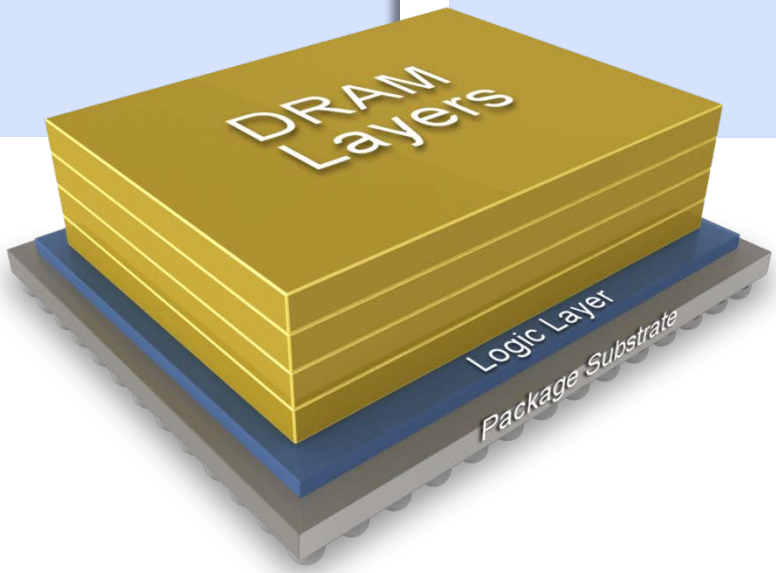
Optimize the Media

- How can we create the most value from the silicon?
  - New Architectures
  - New Technology
    - Process shrinks
    - TSV
- Media management
  - Error correction
  - Repair

Hybrid Memory Cube

Enabling Technologies

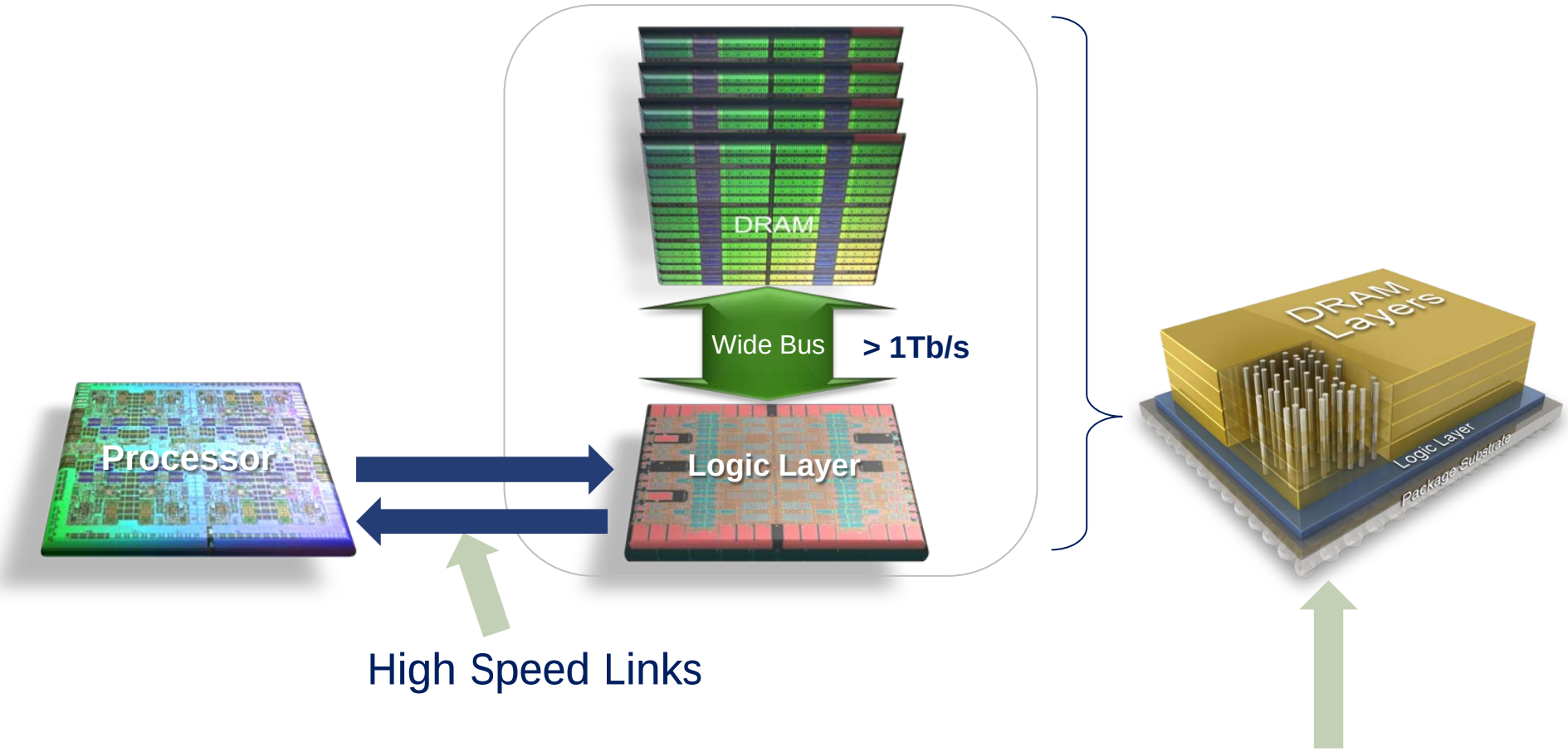
Abstracted Memory Management	Through-Silicon Via (TSV) Assembly
Memory Vaults Versus DRAM Arrays	Process Flow
Logic Base Controller	Package Assembly



HMC Goals

- A standard DRAM building block which can be combined with various versions of application specific logic
  - Primary Targets: Networking (Hubs, switches, routers) & HPC (High Performance Computing)
    - Increased bandwidth
    - High concurrency for multi-core/many-core processors
    - Lowest energy per unit of work done
- Ultimately a new paradigm of *memory system* implementations

Hybrid Memory Cube (HMC)



HMC Architecture

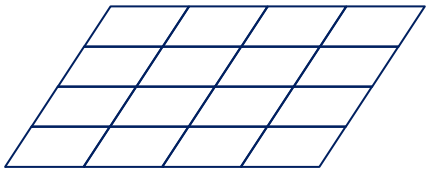
Start with a clean slate

DRAM

HMC Architecture

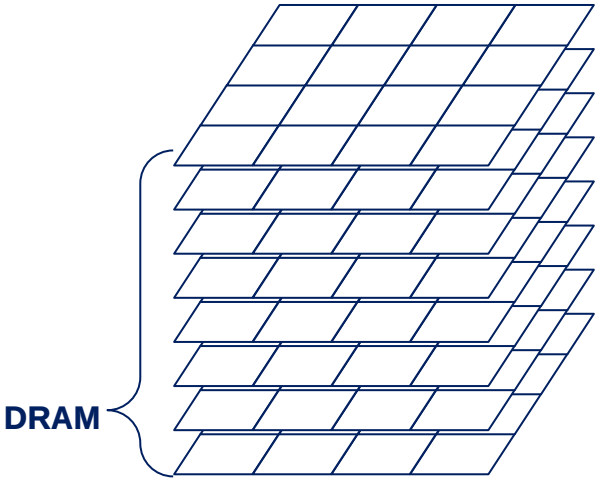
Re-partition the DRAM and strip away the common logic

DRAM

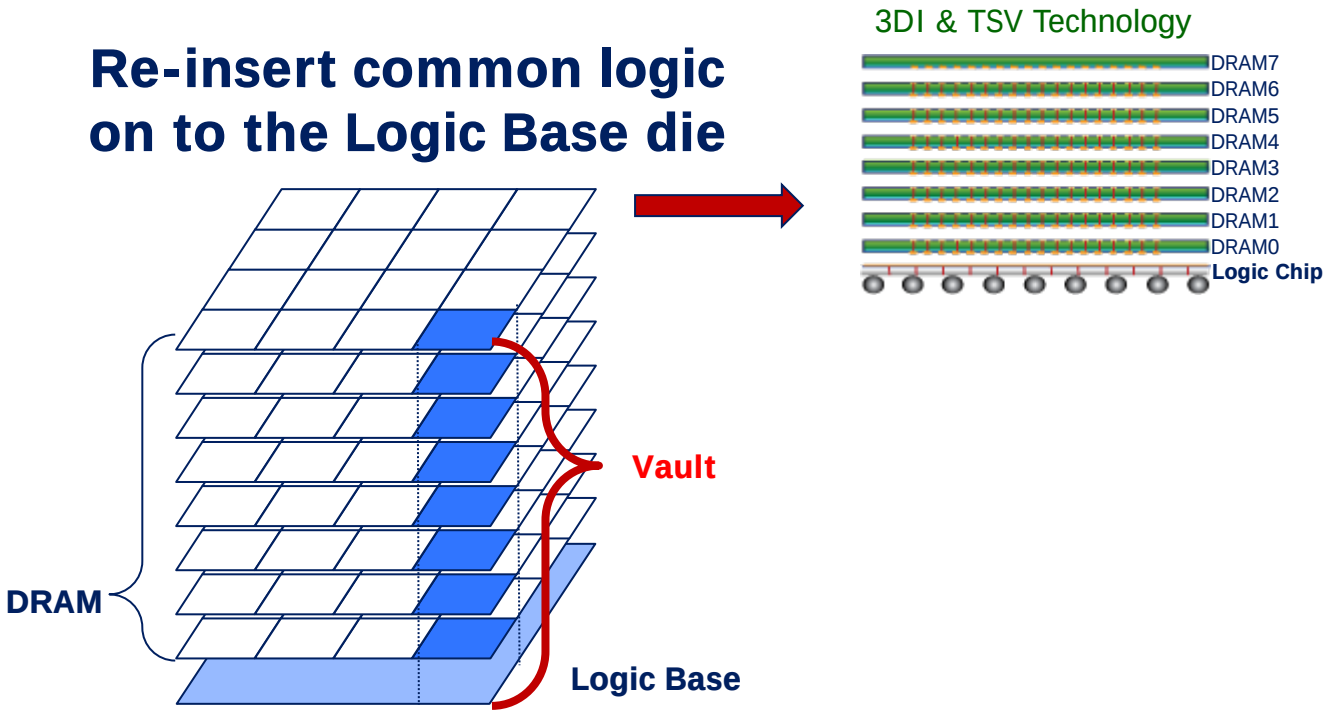


HMC Architecture

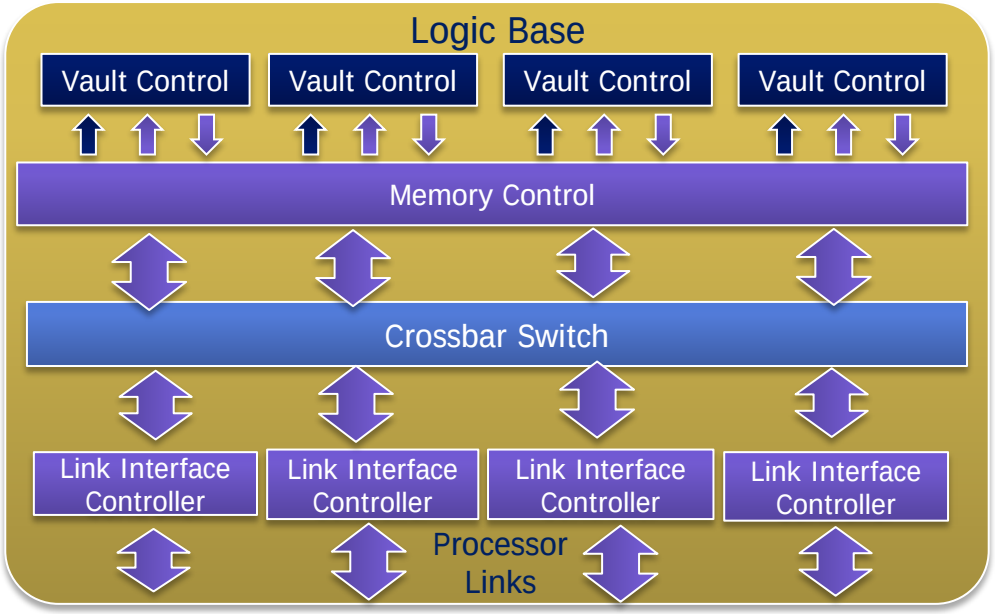
Stack multiple DRAMs



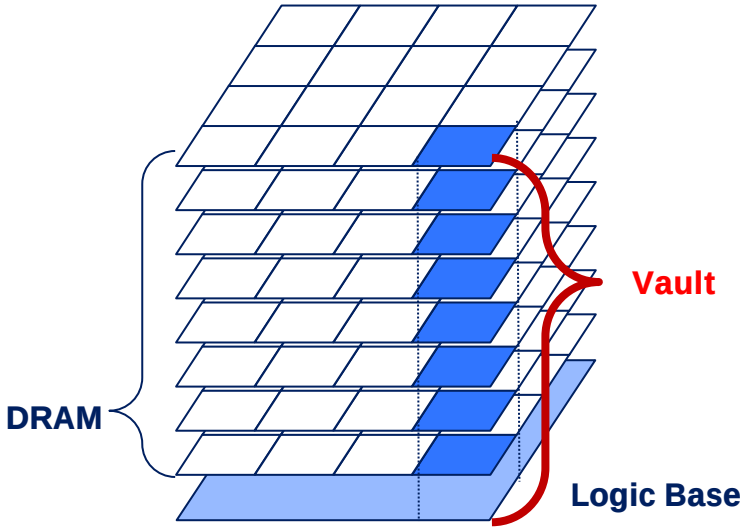
HMC Architecture



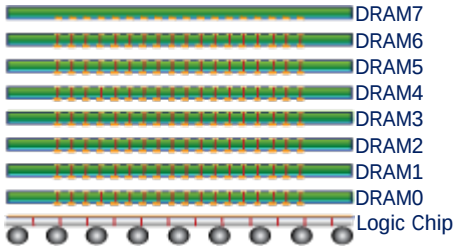
HMC Architecture



Add cross bar switching, optimized memory control and simple interface to host processor(s)…

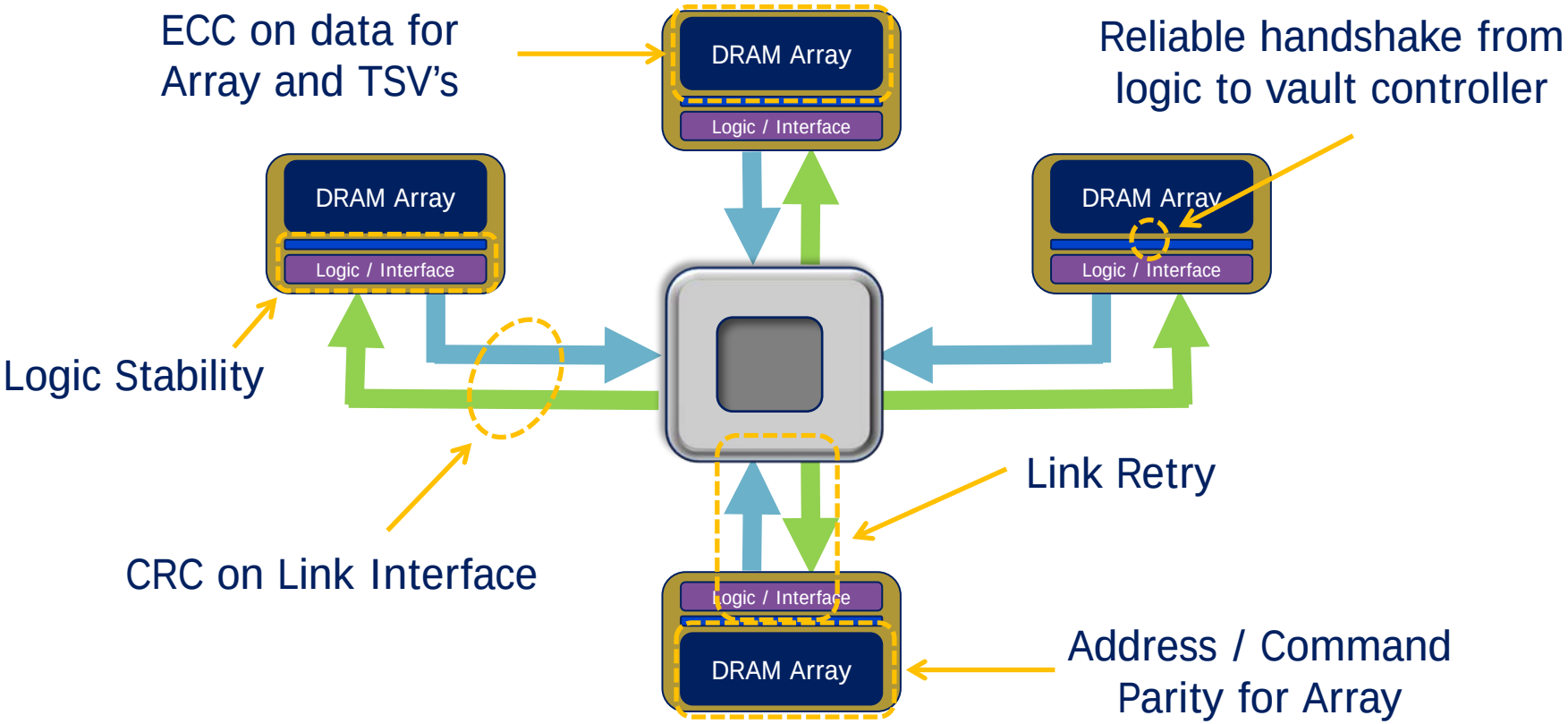


3DI & TSV Technology



HMC Reliability

Built-In RAS features



DRAM Design Goals

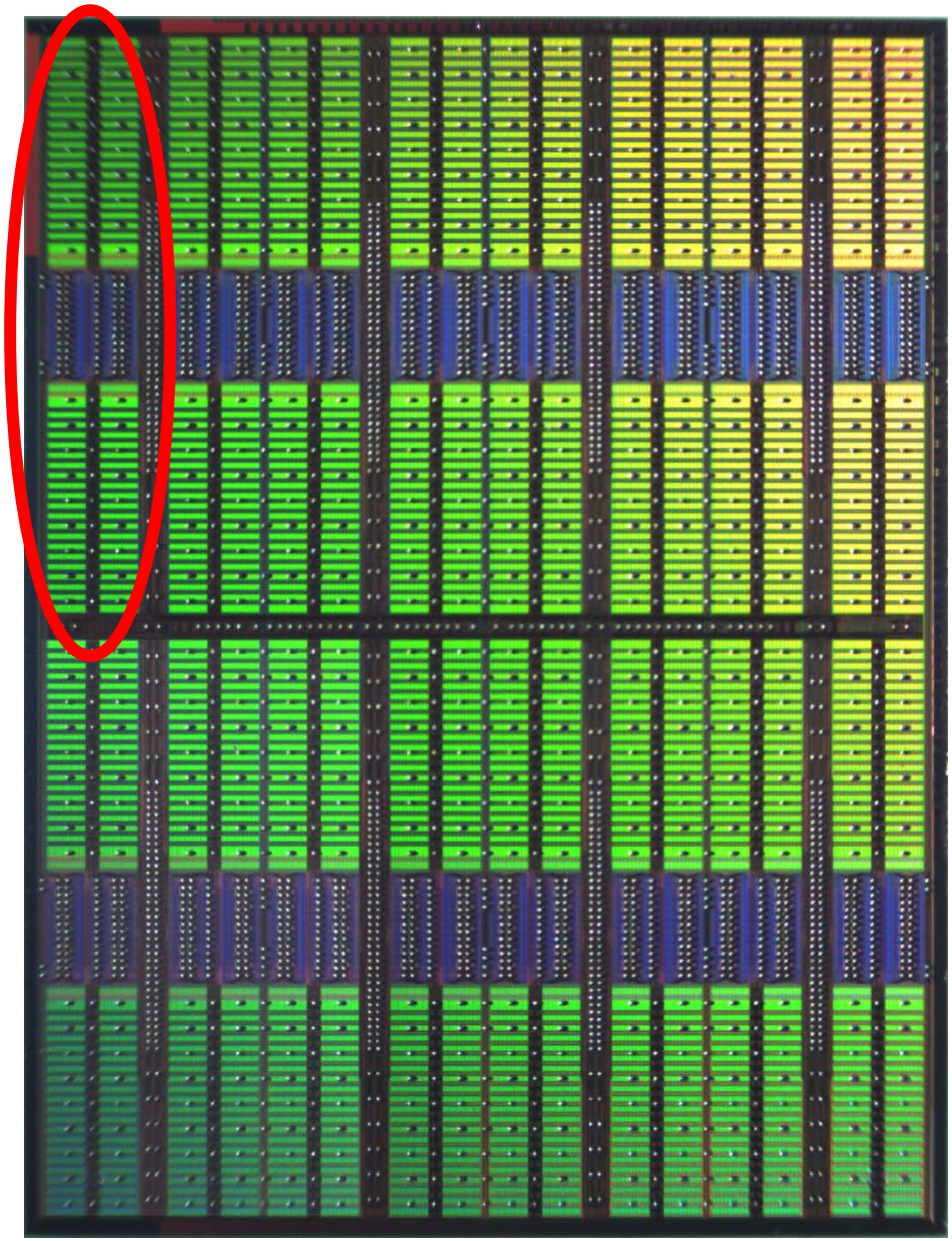
- We started with the following goals to support multi-core, multi-thread/core CPUs:
  - Aggregate bandwidth goal of 128GBps (1024Gbps)
  - Low energy/bit (best in class)
  - High levels of concurrency (≥16 simultaneous operations)
  - TSV signaling rates not exceeding 2Gbps
  - Optimization for full cache line transfers
  - ECC data to support reliability, availability, and serviceability (RAS) features
  - Offload array repair function to logic layer

DRAM Partitions & Vaults

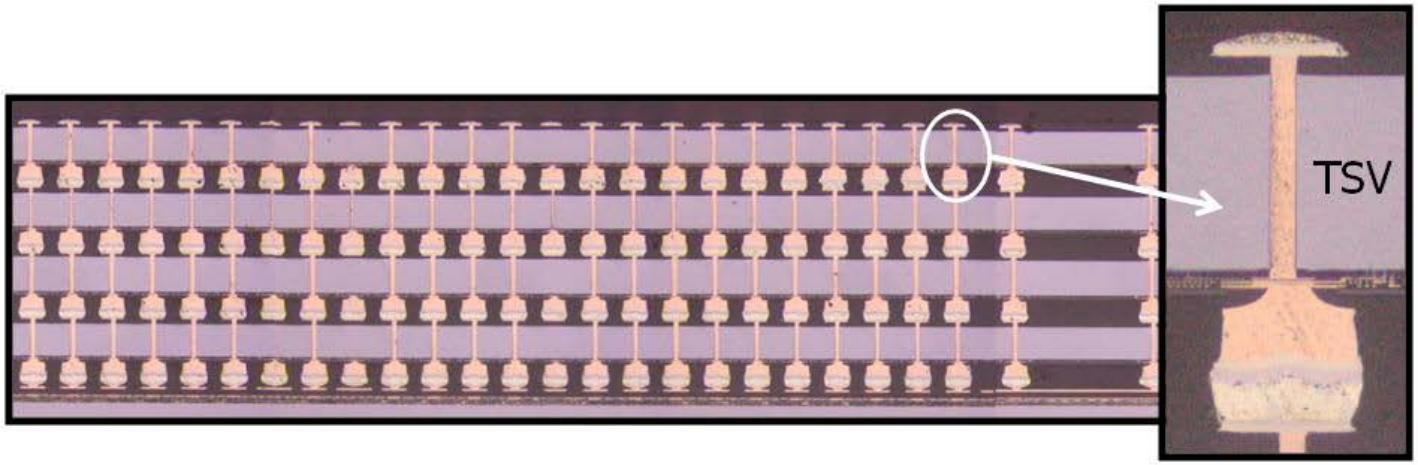
- The DRAM die is segmented into multiple autonomous partitions
  - Gen 1 target: 16 partitions
- Each partition includes multiple independent memory banks (2 to 8)
- Each partition supports full cache line transfers for each access (32 to 256bytes)
- Memory vaults are essentially vertical stacks of DRAM partitions (3D)

1Gb HMC DRAM

Partition

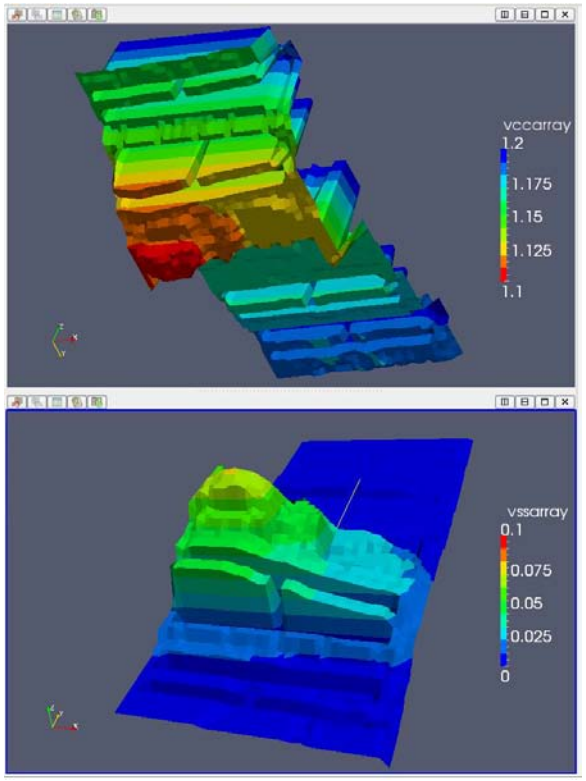


TSV Cross Section

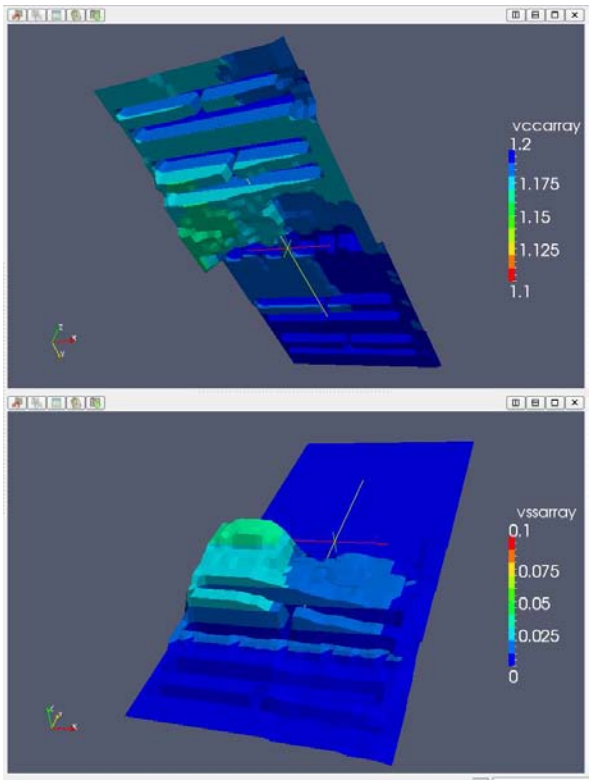


Cross Sectional Photo of HMC Die Stack Including TSV Detail (Inset)

PDN Analysis



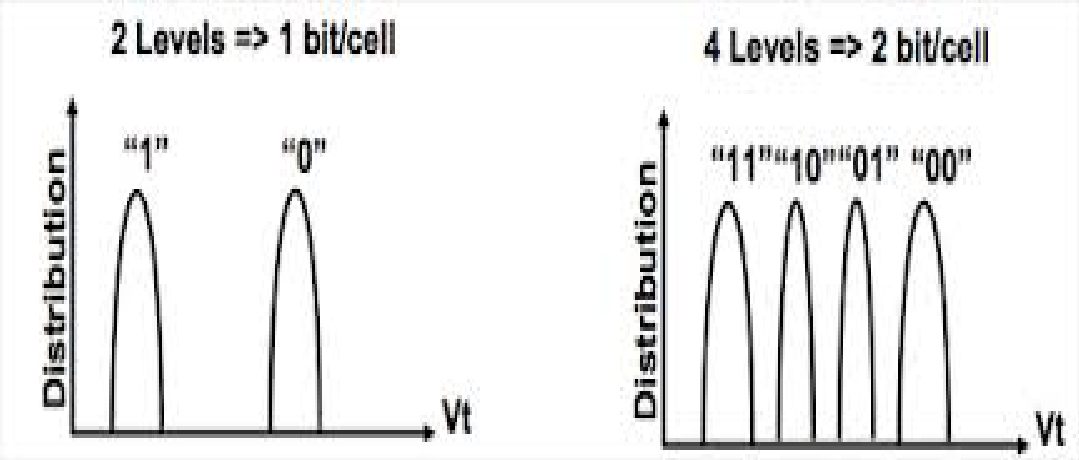
HMC DRAM Array Supply Noise Before Design Modifications, >100mV



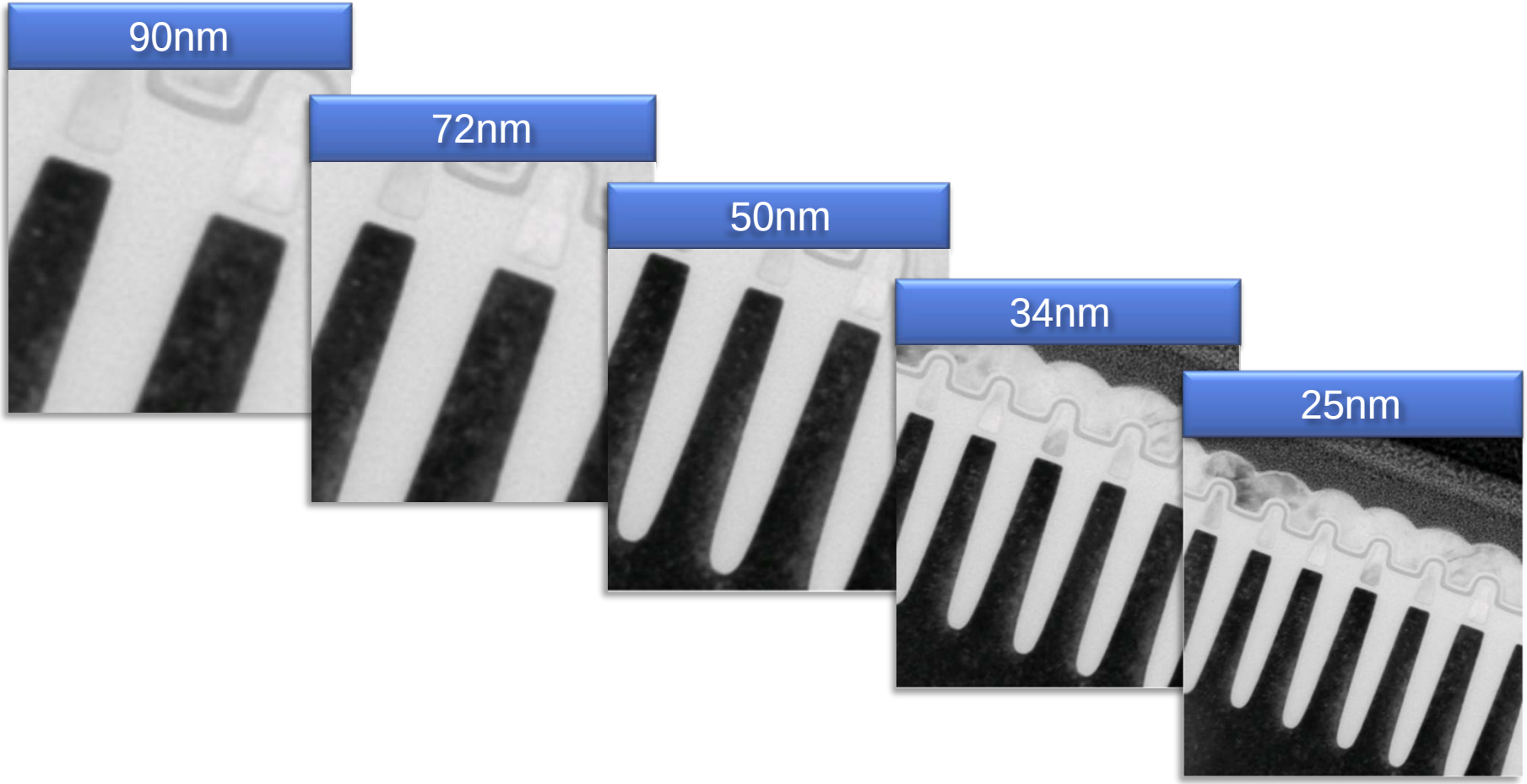
HMC DRAM Array Supply Noise After Design Modifications, < 30 mV

PCIe SSD

SLC vs MLC


2 Levels => 1 bit/cell
Distribution
"1"
"0"
Vt
4 Levels => 2 bit/cell
Distribution
"11" "10" "01" "00"
Vt

NAND Process Technology

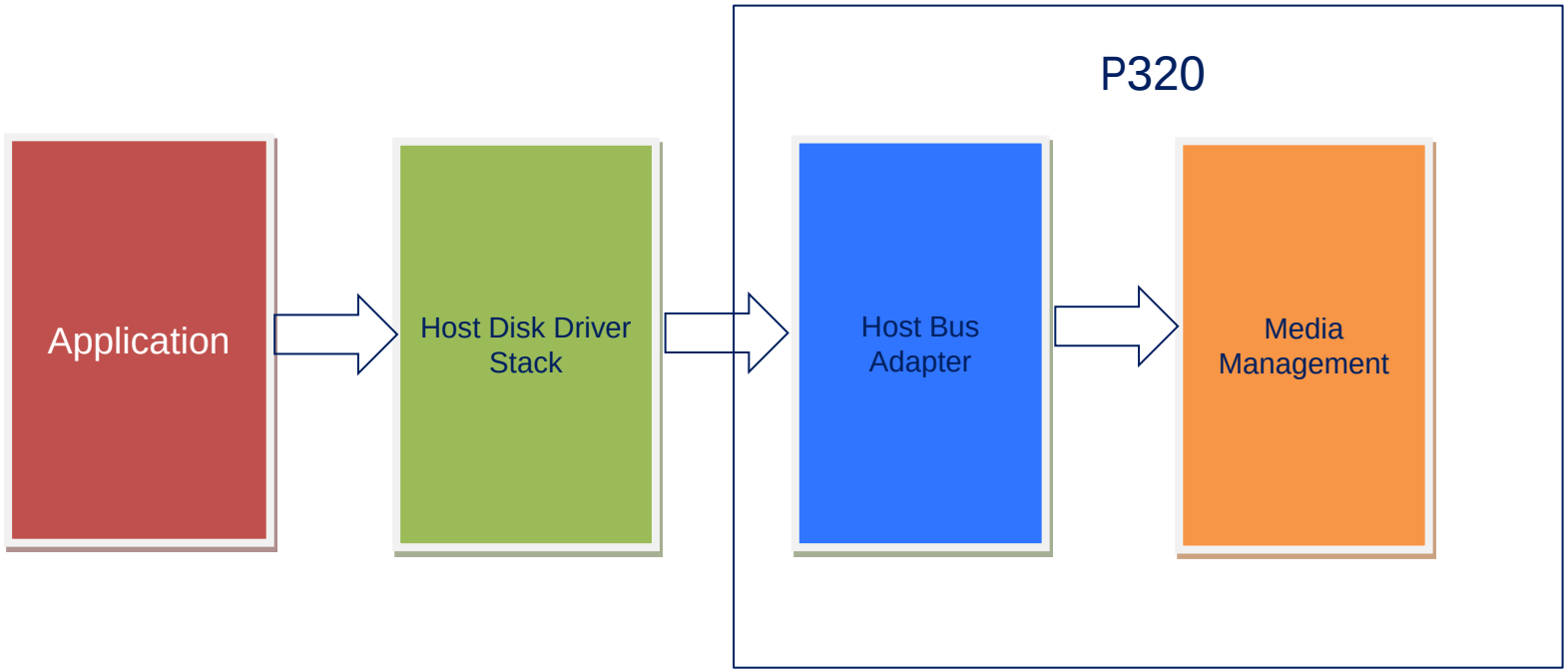


NAND Flash Cross-Sections

Design Priorities

- Maximize performance for given flash footprint
  - How to get 512 flash die working concurrently?
- Optimize random 4KByte writes
  - Traditional SSD weak point
- Maximum concurrency
  - Good steady state performance
- Easy attach
  - Use existing infrastructure
- Optimize performance/power/cost/form factor/ease of use

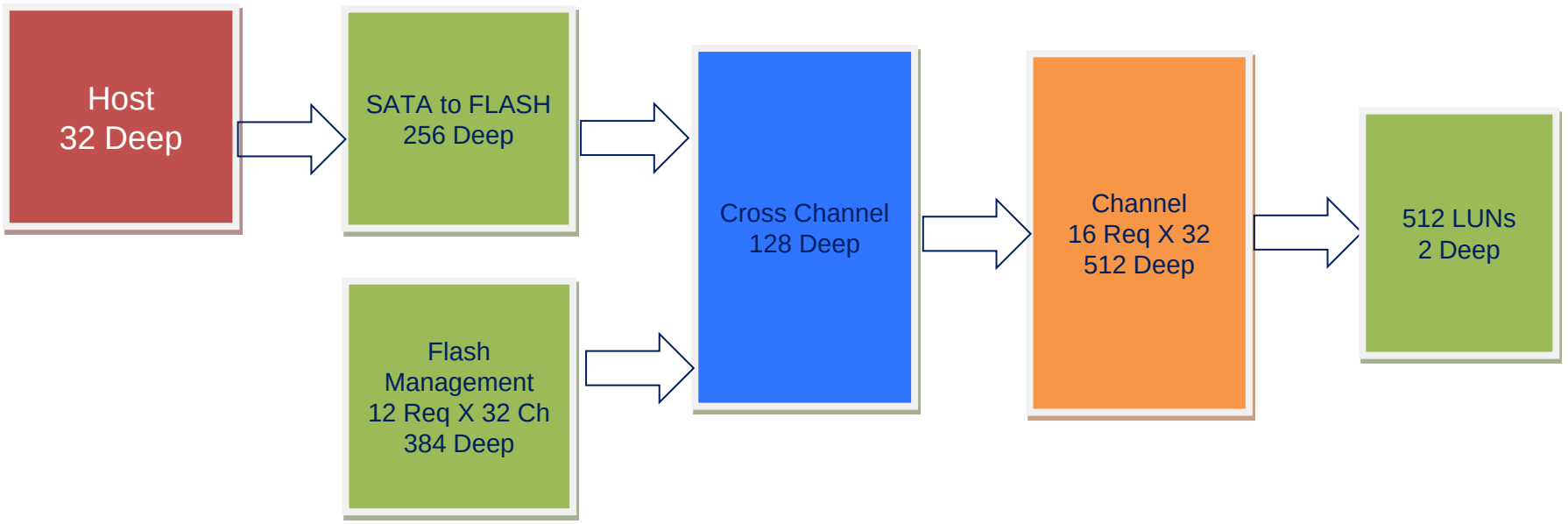
Major Pieces of the Puzzle



Hardware Architecture Overview

- PCIe SSD is Optimized for Performance
  - Media path in hardware
  - Non-media path in firmware
- Three Major Subsystems
  - Host Bus Interface: NVMe support, hw/fw split
  - Array Controller: Manages flash array for IOs
  - Flash Sequencer: NAND channel control and ECC
- Enterprise Data Integrity
  - Overlapping integrity checks
  - RAID / RAIN

Command Queues P320



Host Bus Interface NVMe

- Two NVMe Controller instances
  - For Dual Port operation
  - 256 queues per port
- Breaks commands into native drive chunk size
  - 4k sized and aligned operations
- Command lookup for hardware/firmware routing
- Data integrity checks
  - NVMe Protection Information generate/check
  - Internal integrity fields : Flash LBA and flash CRC

Temp / Power Throttling

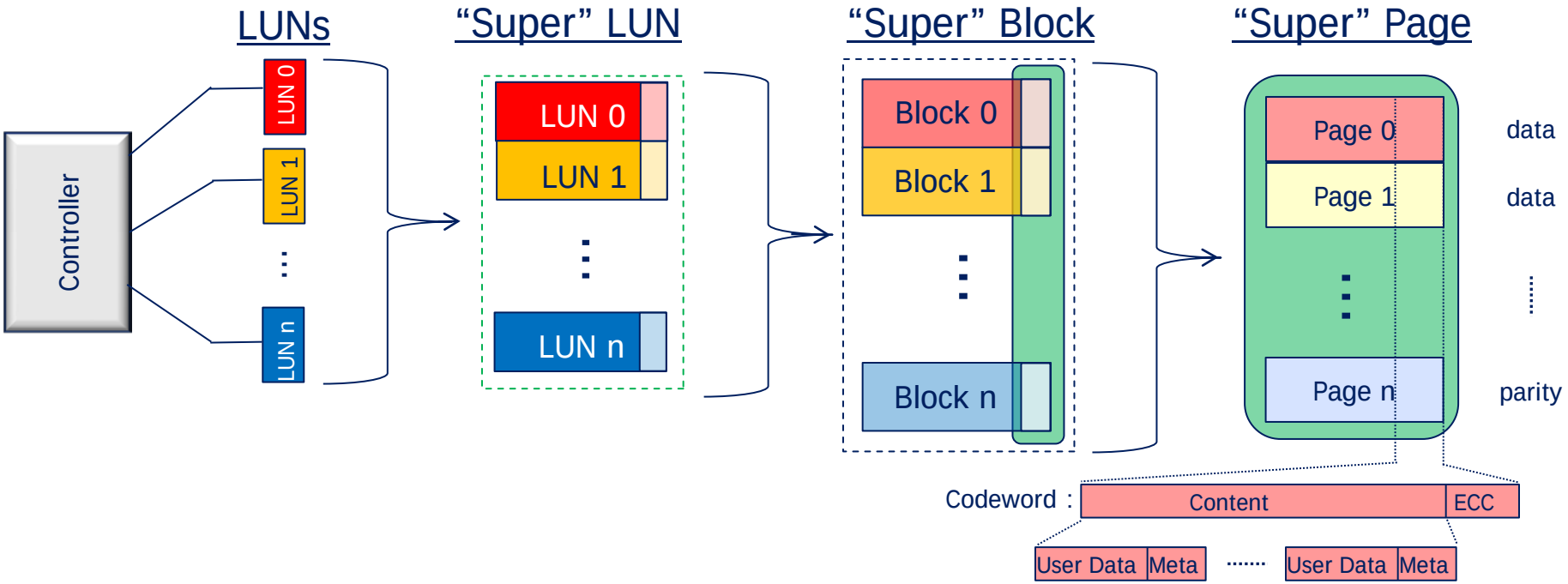
- Temperature Control
  - PCIe SSD is specified to operate at 0C to 50C with 1.5m/s airflow
  - At 50C ambient (87C SMART temp[1]) temperature, drive will throttle writes
    - Drive will exit this mode when temperature returns to normal
- Many systems have a 25W slot power limit
  - Request traffic is monitored to stay within 25W envelope

[1] *SMART temp reads 6-9 degrees hotter than ASIC temperature*

Managed Wear

- Optional write endurance guarantee
  - Allows short write bursts but constrains writes to long term endurance objective
- Monitors long term and short term wear
  - Allows PE cycles to be “banked”
- Allows “burst” traffic
  - Allows max write performance for short durations if it fits overall near term wear profile
  - Continual monitoring and adjustment

RAIN Organization



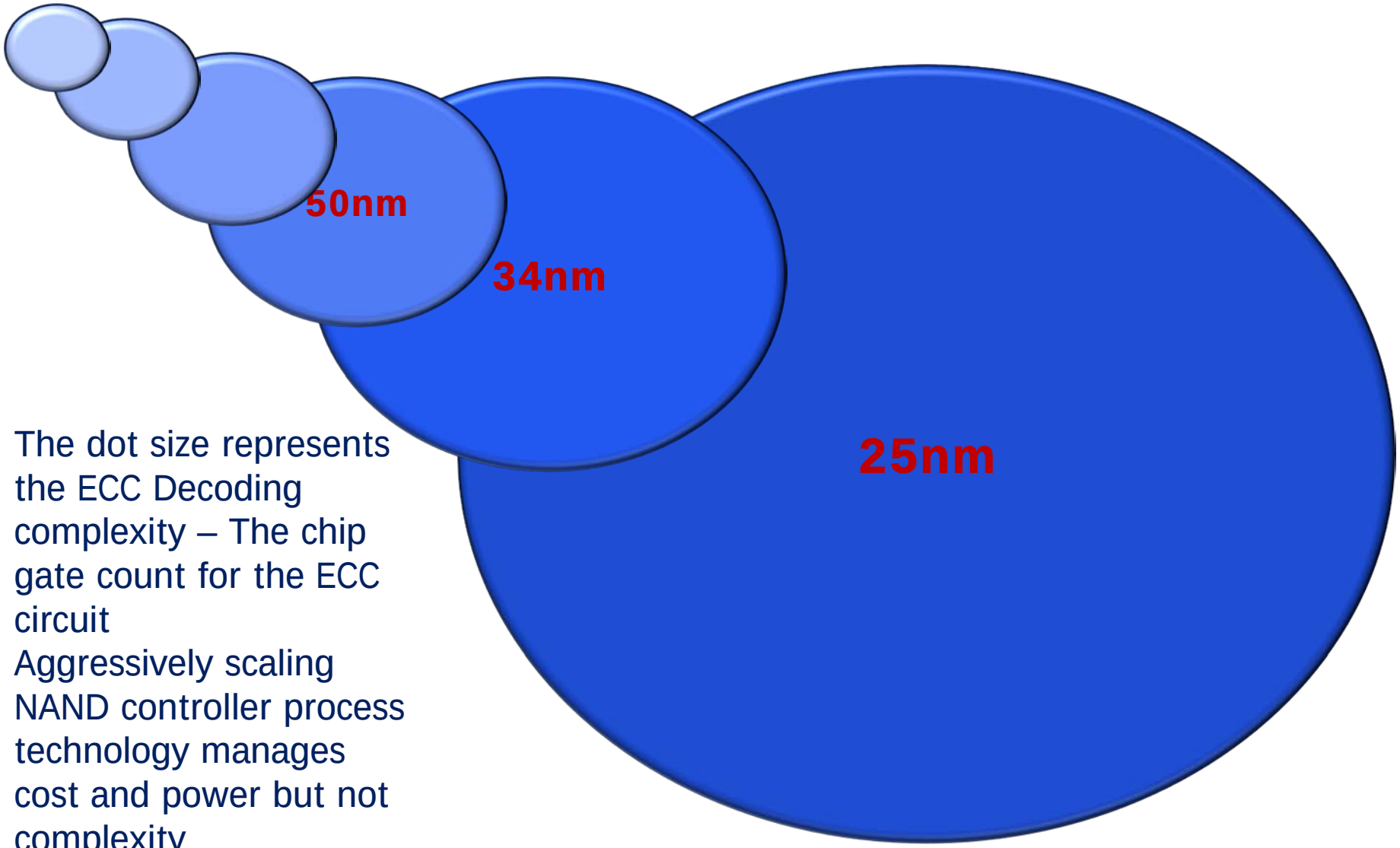
- Up to 16 different LUNs (on separate ONFI buses) associated as a **Super-LUN**
- Blocks from each LUN in Super-LUN associated as a **Super-Block**
- Pages within Super-Block associated as a **Super-Page**
- Pages are comprised of ECC protected Codewords
- Codewords contain user data, meta data and DIF

Error Correction

- Two step process:
  - *Encoding* is preformed as the data is written to the flash
    - User data is supplemented with *parity*. The combined user data and parity is a c*odeword*

    - There is a direct relationship between the amount of parity and the number of errors that the code can correct for a given codeword size
  - *Decoding* is preformed as the data is read and is a multiple step process:
    1. Calculate error signature called the *syndrome*
    2. Determine error pattern from the syndrome
    3. Correct errors

ECC Complexity Scaling Trends



- The dot size represents the ECC Decoding complexity – The chip gate count for the ECC circuit
- Aggressively scaling NAND controller process technology manages cost and power but not complexity

Performance

- Seq Read = 6 GB/s
- Seq Write > 2GB/s
- Random Read 4K IOPs > 1.5 million
- Random Write 4K IOPs > 500K

Capacity

- 16 GB NAND die available today
- 32 GB NAND die soon
- 16 TB PCIe drives coming soon
- Cost curve on rapid decline

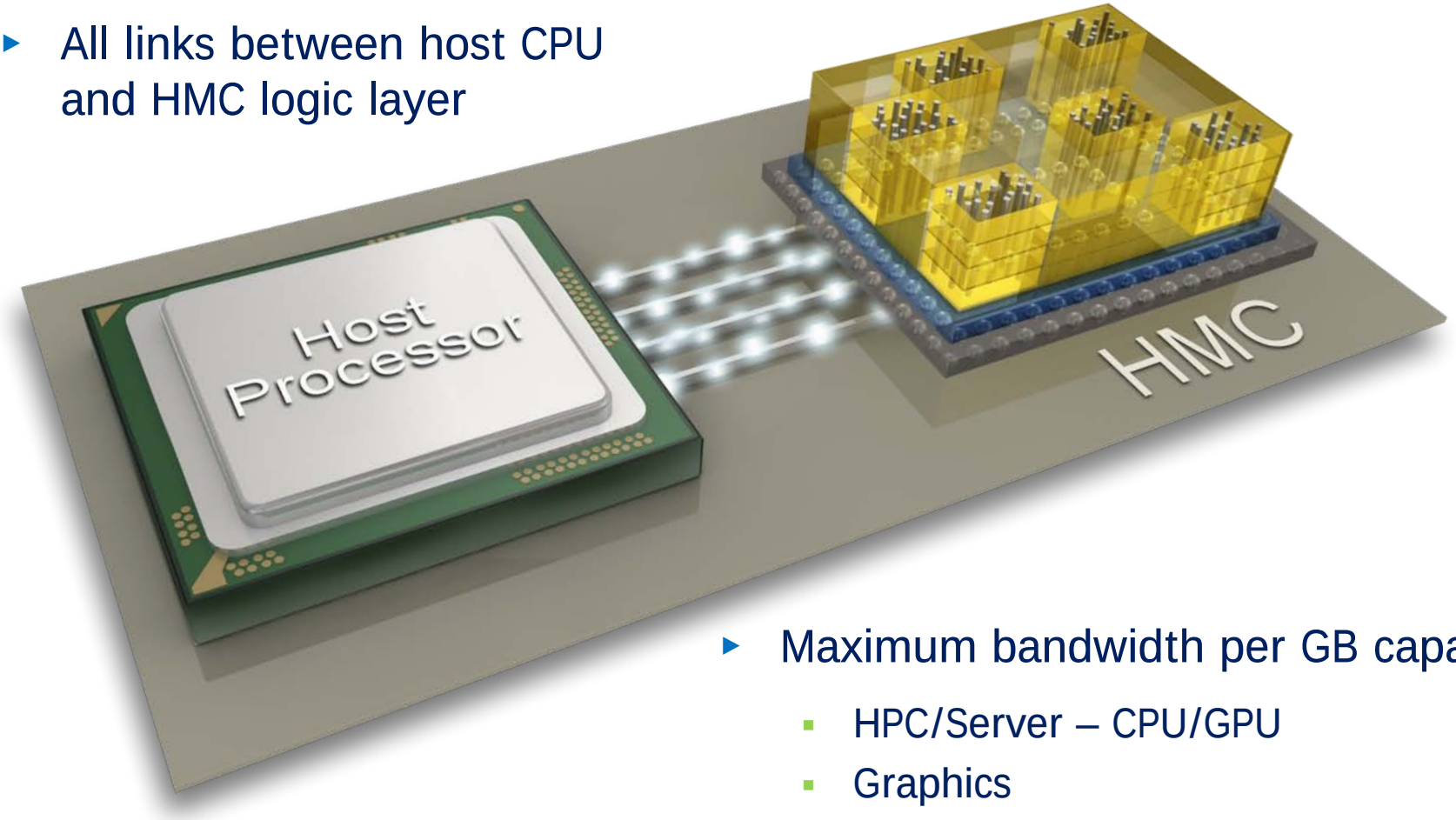
Future Possibilities

Future

- Increase DRAM bandwidth and concurrency
- Additional interfaces
  - Match signaling to interconnect materials
  - Adapt signaling and power to distance and topology
- Add compute and data movement functions to logic die
- Photonics
- Tiered, managed media
  - HMC, DDR4, NAND

HMC Near Memory

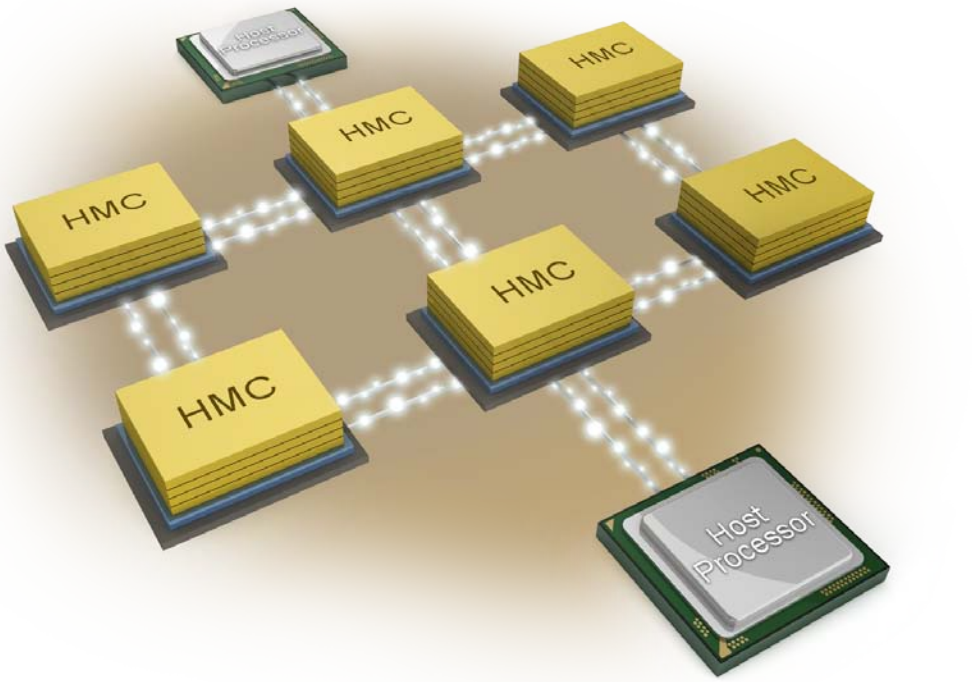
- All links between host CPU and HMC logic layer



- Maximum bandwidth per GB capacity
  - HPC/Server – CPU/GPU
  - Graphics
  - Networking systems
  - Test equipment

HMC Far Memory

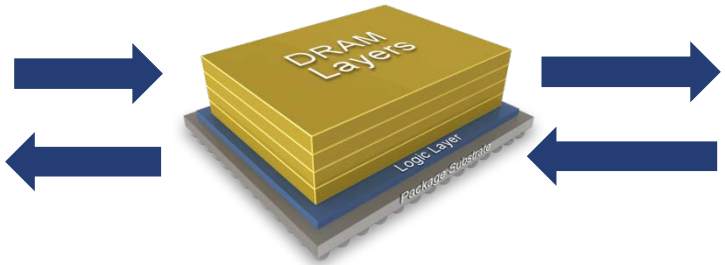
- HMC links connect to host or other cubes
  - Links form networks of cubes
  - Scalable to meet system requirements



- Future interfaces
  - Higher speed electrical
  - Optical
  - Whatever the most appropriate interface is for the job!

Tiered Memory

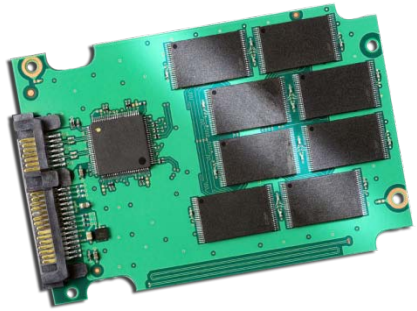
Abstracted Interface



HMC

Low Power Serdes

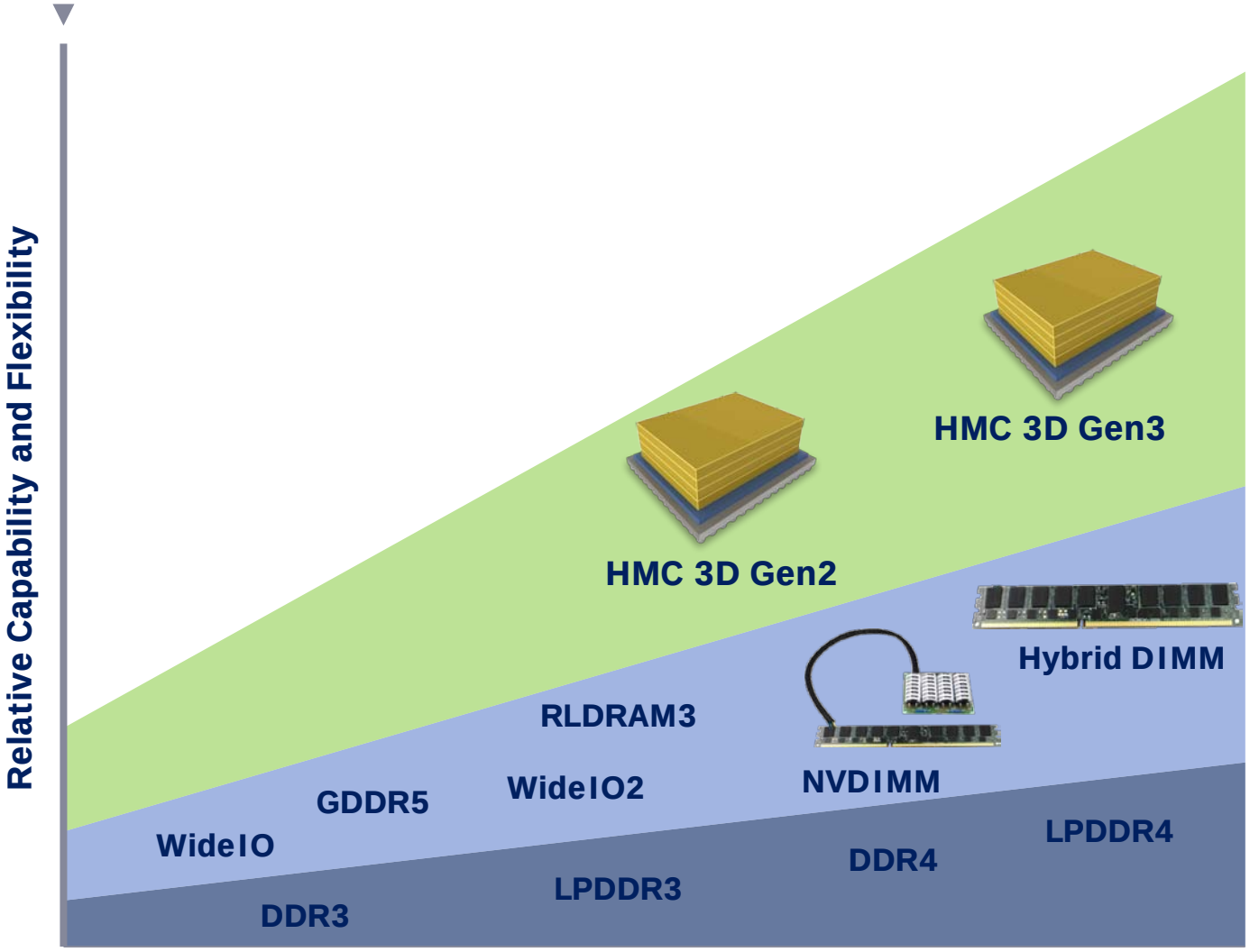
DDR4



NVM

Photonics

Innovative Path to New System Capabilities



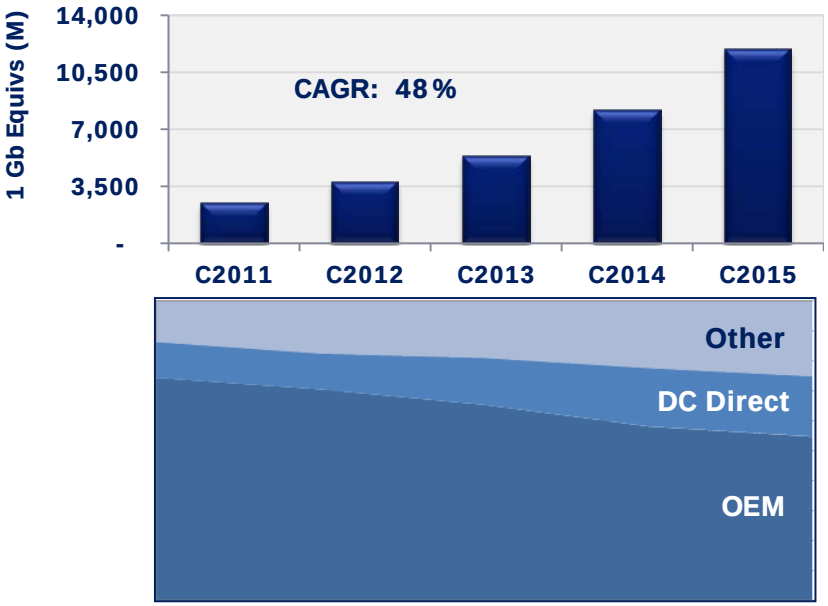
Revolution
- Ultimate Bandwidth
- Abstracted Interface
- Highly Scalable

Evolution
- Increased Bandwidth
- Application Tailored
- Scalability

Variation
- Limited Bandwidth
- Various Applications
- Power efficient

Solutions for Enterprise Storage and Computing

DRAM Growth in Server & Storage



Source: Micron with inputs from Gartner, iSuppli and others

Market Trends

- Enterprise market continues to rapidly develop into distinct Public Cloud and Corporate Segments
- Use Cases, Workloads and economies of scale dictating engagement dynamics and technology roadmaps
- Application development in analytics driving need for in memory database capabilities and storage tiering solutions

Memory Architectures for Enterprise Computing & Storage

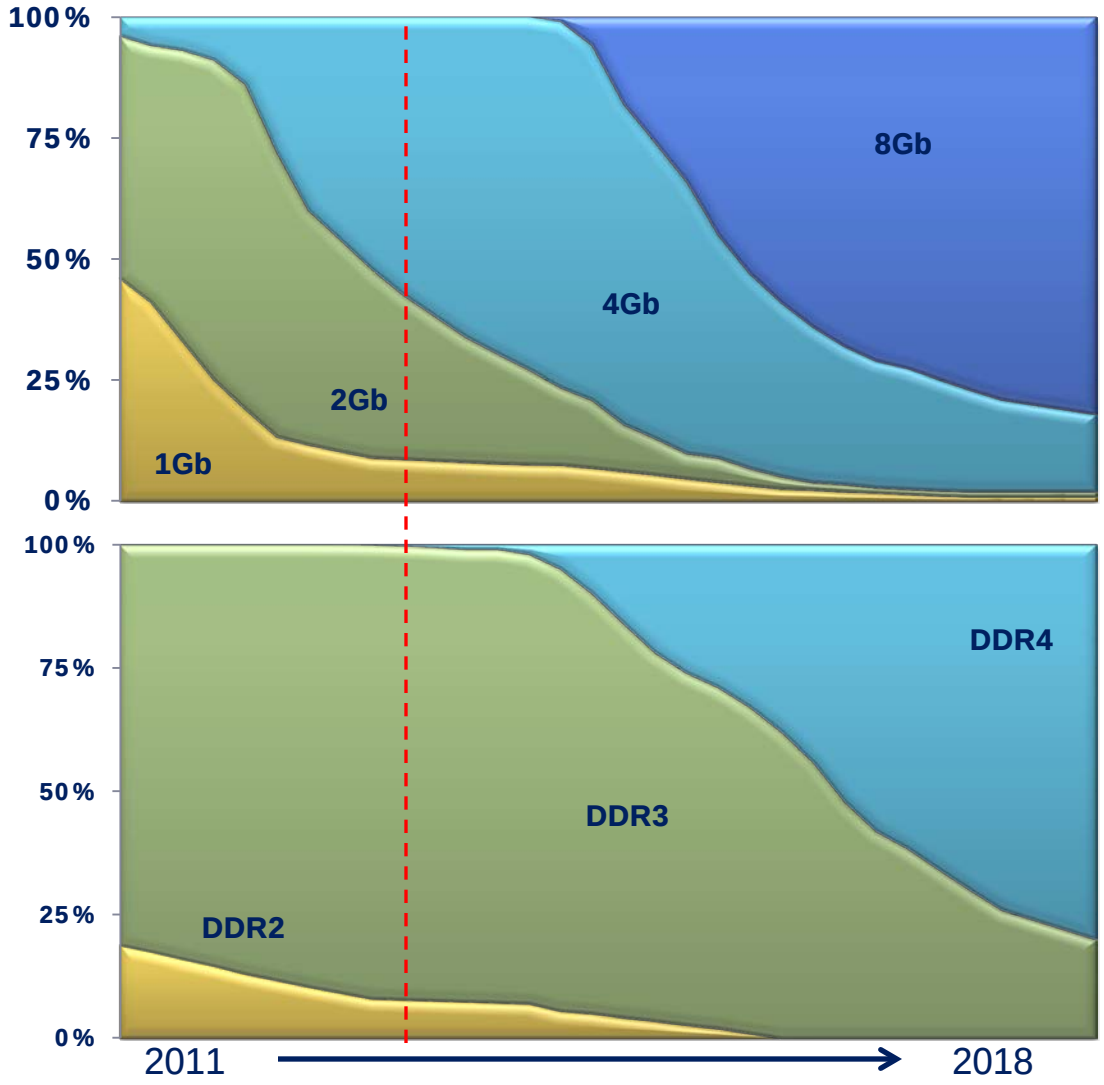
64GB LRDIMM
Maximum Module Density

Micro-Server DIMMs
SORDIMM, ECC SODIMM, Mini-VLP UDIMM

1.5U LRDIMM
Module Density/ Comfortable Price

NVDIMM
Persistent Memory Applications

Server & Storage DRAM Market Transitions



Source: Micron Business Development - 2Q13

Market Trends

- 2Gb to 4Gb DDR3 component transition currently in progress (drives 8GB/16GB module cross over)
- 16GB/32GB module cross over driven by availability of 8Gb components
- 8Gb expected to be mainstream component DRAM density by mid 2016
- One speed grade bump per year anticipated
- DDR4 demand expected mid 2014 with 1866 and 2133 MT/s speed grades
- Prolonged DDR3 to DDR4 transition expected


Micron®