VISION OF PROCESSOR–MEMORY SYSTEMS

MICRO 48

J. Thomas Pawlowski, Chief Technologist, Fellow
jpawlowski@micron.com

- Memory diversity
- Some scaling observations
- Unabated performance pressure
- Memory Technologies
- Emerging Memory
- 3D XPoint™ Memory
- Changing Relationships
- Micron Automata Processor
- Abstraction and Future Systems
- 11 Key Points

Memory Diversity

Diversified Memory Markets

Automotive

Storage

Graphics / Consumer

Networks

Server

Mobile

Personal Computing

IMM

Smartphones and Tablets

Micron Portfolio

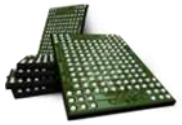
4-32Gb LPDDR4

- Multiple package options
- 4x provides 4GB density
- LPDDR4 2x bandwidth LPDDR3
- LPDDR4 35% power savings over LPDDR3

eMMC 5.x

- Small package options
- 3x random Read/Write performance
- 2x sequential Read/Write performance
- Wide temperature range tolerance and high reliability

eMCP + LPDDR3



- Attractive package/density options
- 2x bandwidth over eMCP+LP2
- Quick time-to-market for system qualification efforts

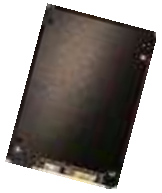
Enterprise and Cloud

Micron Portfolio

Server Modules

- Up to 128GB module densities
- DDR4 solutions driving higher performance
- Load Reduced options for higher system capacity

Enterprise SSD



- Enabled with high performance MLC NAND
- Expanding portfolio of SAS and PCIe
- Partnerships enable efficient go to market

NVDIMM

- Integrated DRAM/NAND Module
- Outstanding latency and bandwidth
- Excellent endurance capability

Hybrid Memory Cube

- Best-in-class bandwidth solution
- 70% less energy per bit
- Industry leading reliability

3D XPoint

- First new memory technology in over 25 years
- 1,000x faster than NAND
- 10x higher density than conventional memory

Networking

Micron Portfolio

DRAM Modules

- Error correction for data integrity
- VLP and custom form factors
- Leading-edge and legacy options

RLDRAM

- Low latency
- Fast random access
- Specialized x9, x18, and x36 configurations

Hybrid Memory Cube

- Best-in-class bandwidth solution
- Provides the bandwidth of as many as 144 DDR3 components

eUSB

- Firmware and OS management
- User tracking data

Client

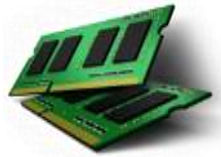
Micron Portfolio

On-Board Memory

- Up to 16Gb per package
- Various package options
- Soldered-on-board for thin profile
- DDR4, DDR3L and LPDDR3/4

Client Modules

- Up to 16GB density
- Thin form factor
- 1.35V DDR3L at 10% lower power than standard DDR3
- DDR4 memory



Serial NOR

- Dependable BIOS
- Secure Boot

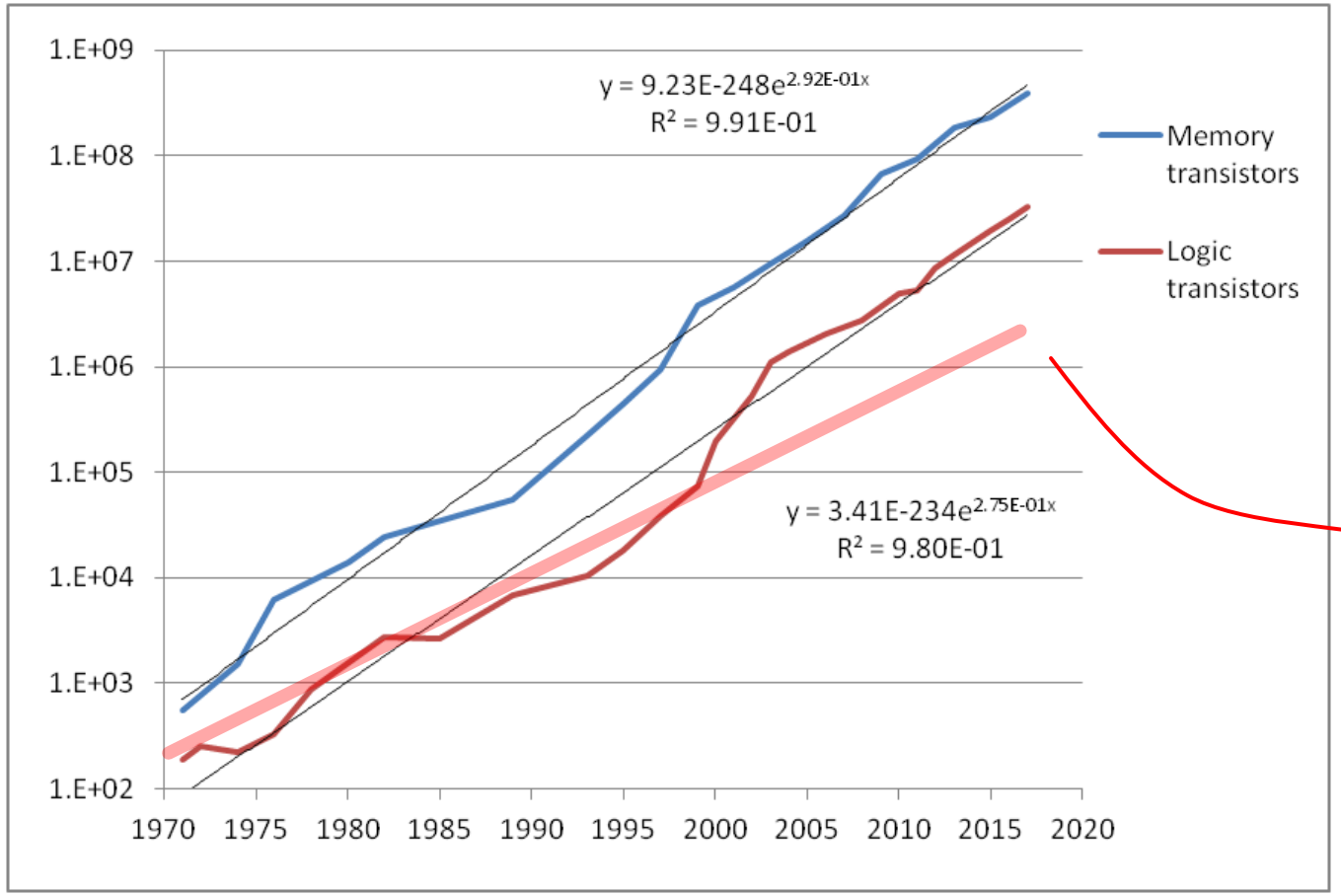
Client SSD

- Vs. HDD: Superior $/IOPS and IOPS/watt
- Faster boot and bandwidth
- Lower dB, ns, W, oz, cu.in.
- Encryption, security
- 100x performance

Some Scaling Observations

DRAM Array and Logic Transistors per square mm

- Suggests using memory wafers more extensively than logic, widening gap
- Logic transistor curve based only on Intel production start dates to maintain harmony of methodologies Dates rounded down to year – Note that this does not show COST



1.339x / 1 year
1.793x / 2 year
2x / 2.4 years

1.316x / 1 year
1.733x / 2 year
2x / 2.5 year

Logic transistor density growing at slower rate (excluding on-die SRAM)

Source: JTPawlowski, Micron

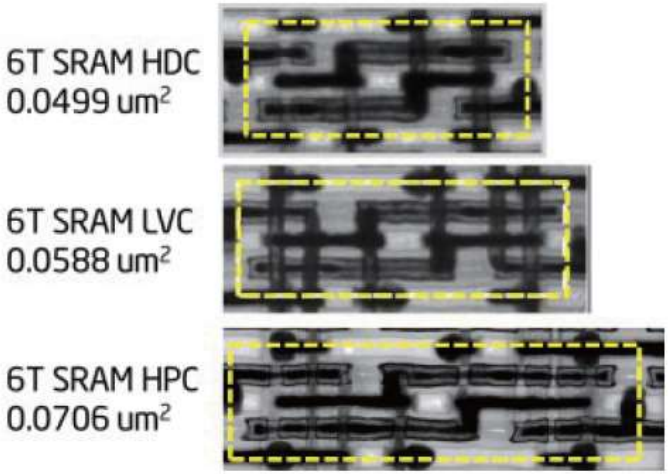
Most Dense Use of Logic Technology is Memory

	22nm	14nm
6T HDC SRAM	Intel 190.1 F^2	Intel 14.5Mb/mm^2 254.6 F^2
6T LVC SRAM	IBM dens-perf 297.5 F^2	Intel 300 F^2
6T HPC SRAM		Intel 360.2 F^2
8T SRAM		Intel 5.1Mb/ mm^2
eDRAM	IBM 53.7 F^2 Intel Macro 17.5Mb/mm^2 59.9 F^2	IBM 89.1 F^2 IBM Macro 18.7Mb/mm^2

Source: various ISSCC and VLSI Symposium Proceedings

- Compare to DRAM now constant 6-7F^2
- SRAM scaling poorly
- eDRAM appears worse – watch future scaling
  Some density and energy benefit for now
- DRAM is challenging but scaling far better than SRAM and eDRAM

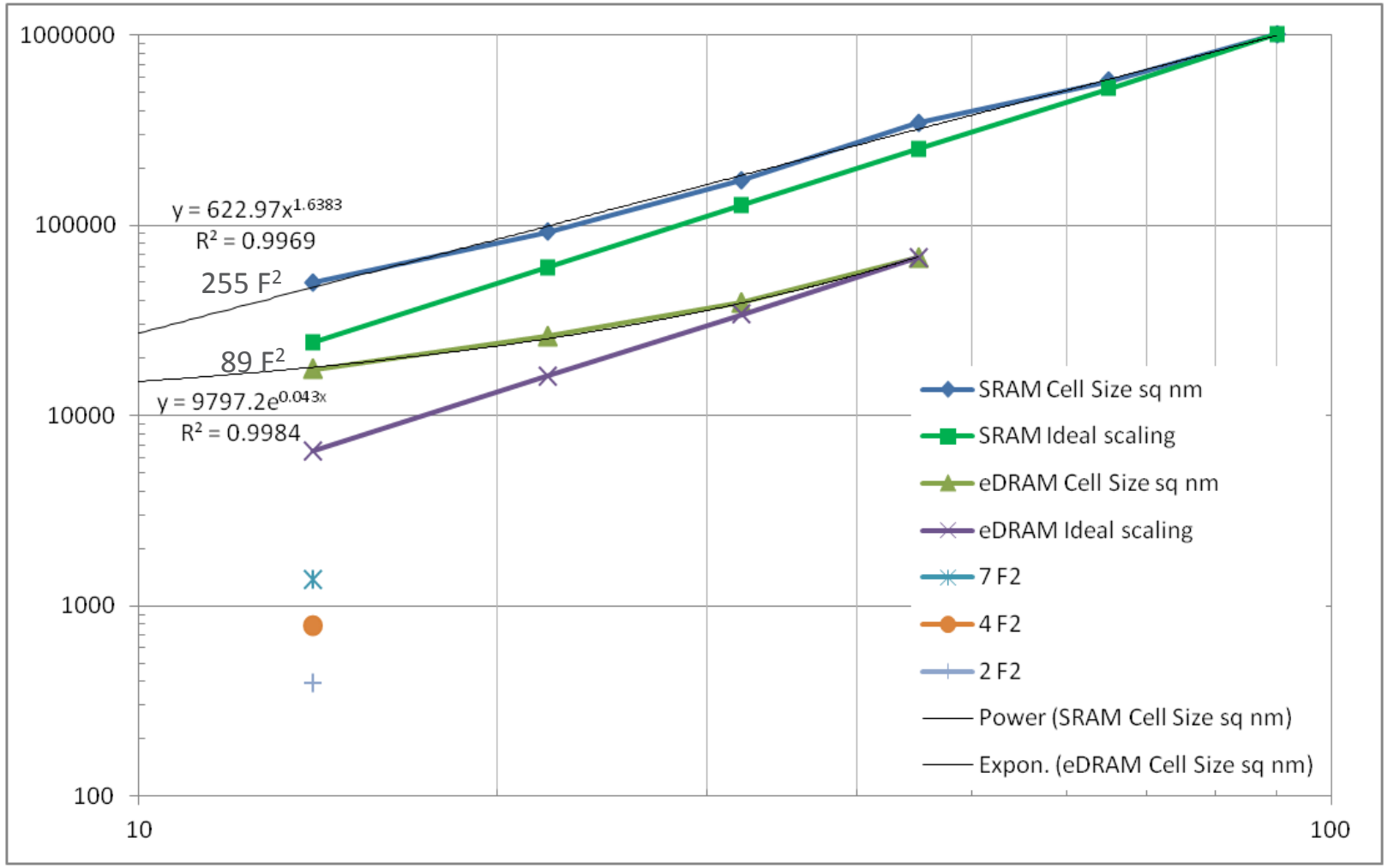
Intel 14nm SRAMs



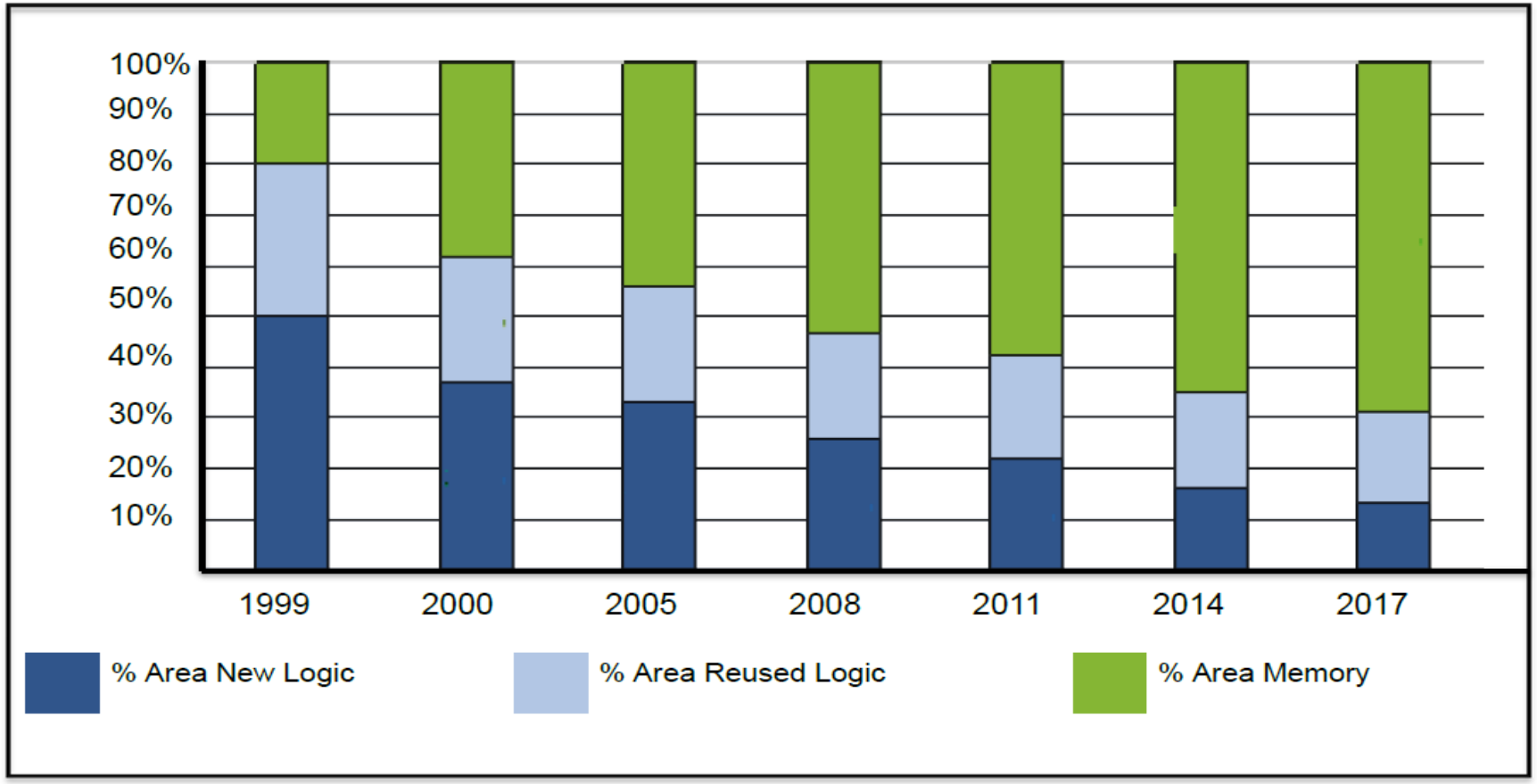
Intel HC SRAM and IBM eDRAM Cell Scaling

Area [nm²] vs. Node [nm]

- SRAM 2x relative size since 90nm, eDRAM 2.7x since 45nm, macros scaling worse
- Contrast: smallest nFET available in 14nm:
  Intel 25F²
  Micron 4F²



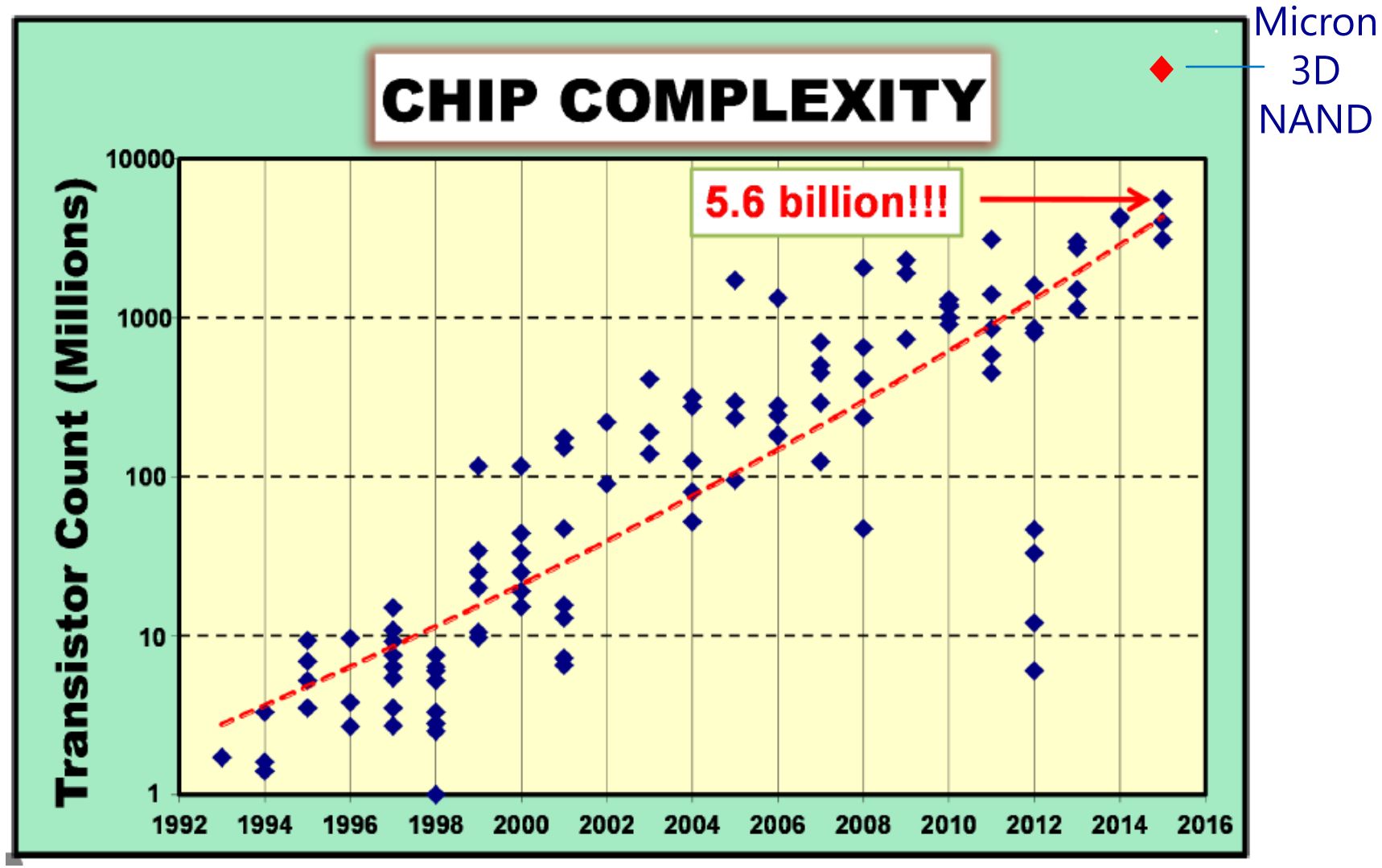
Memory Growing As Percentage Area in Logic Chips



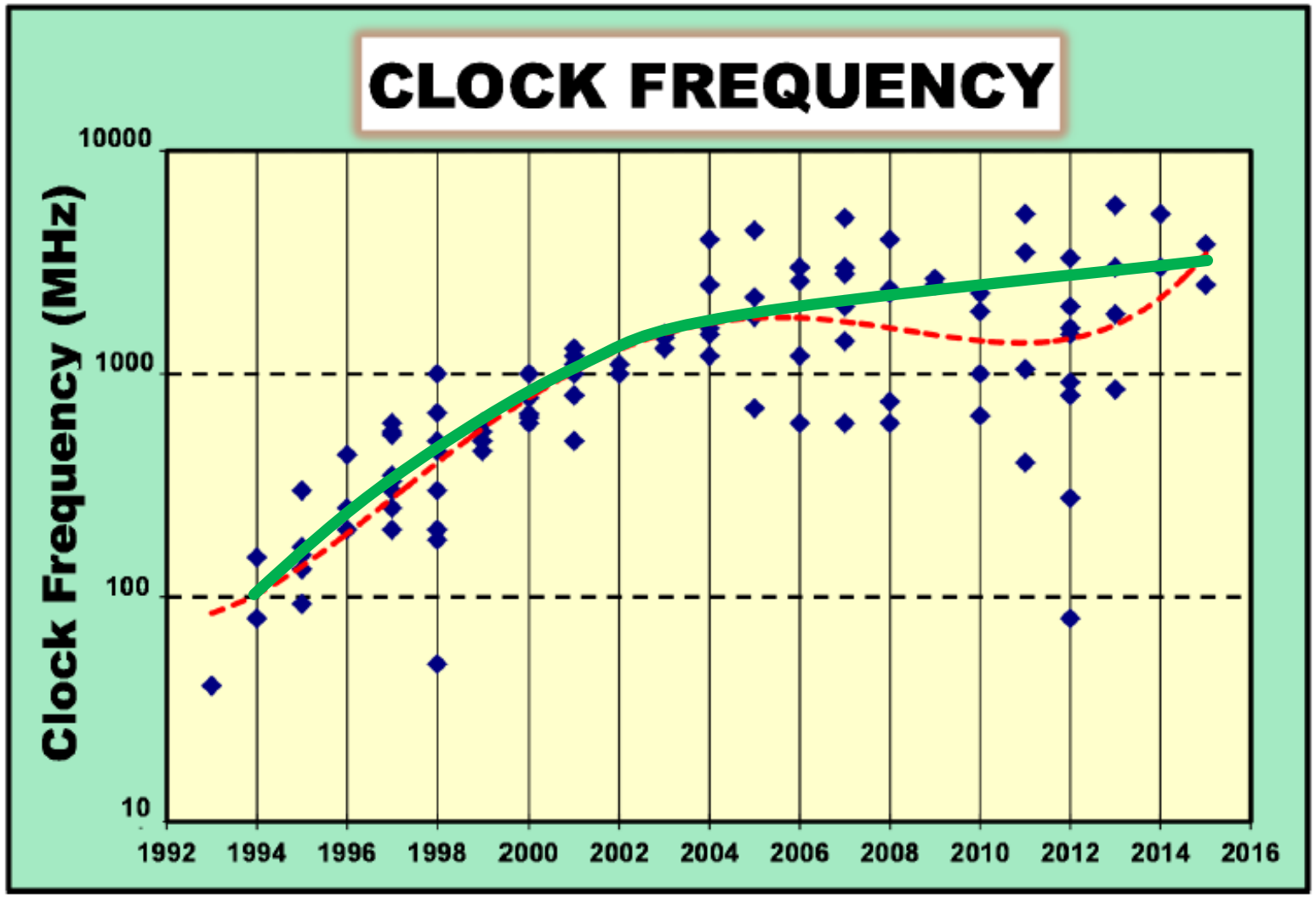
Source: Semico

Much of this area can be displaced by cheaper process technology

Unabated Performance Pressure



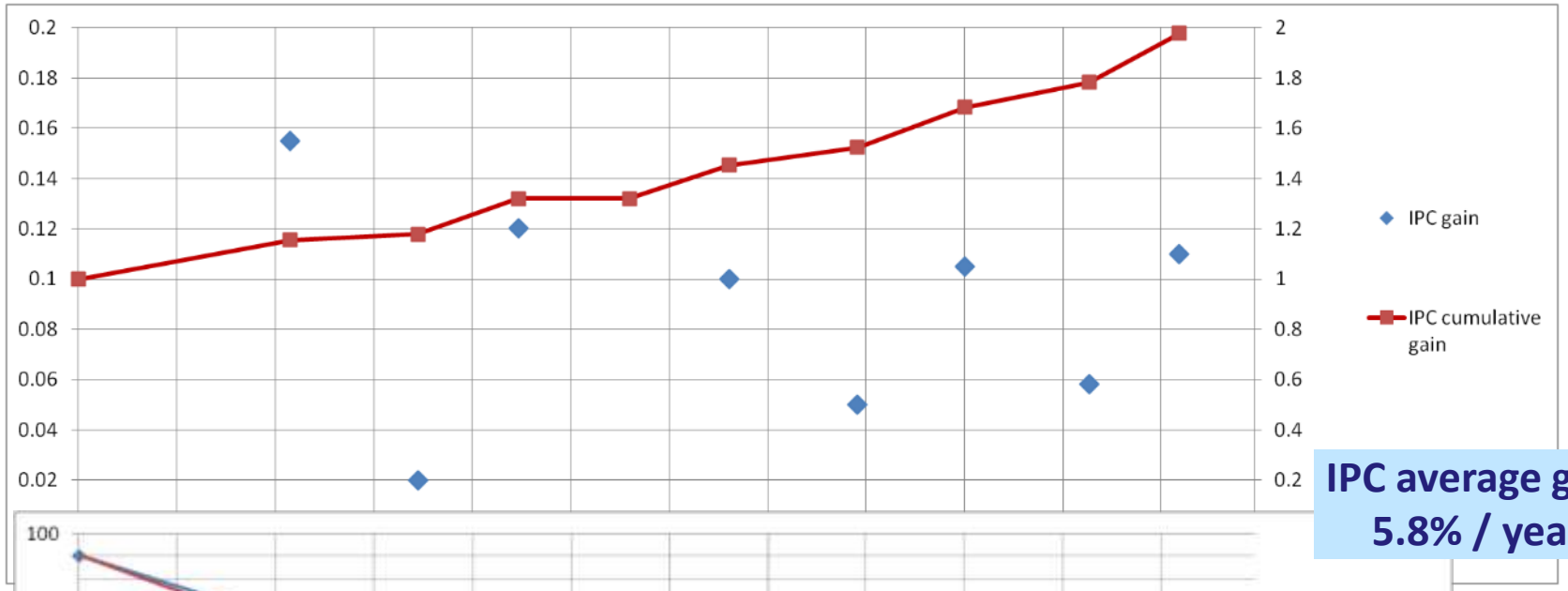
- Source: ISSCC 2015 Trends – Logic chips only
- Memory transistor count dwarfs logic (8Gb DRAM ~10B, latest NAND ~400B)



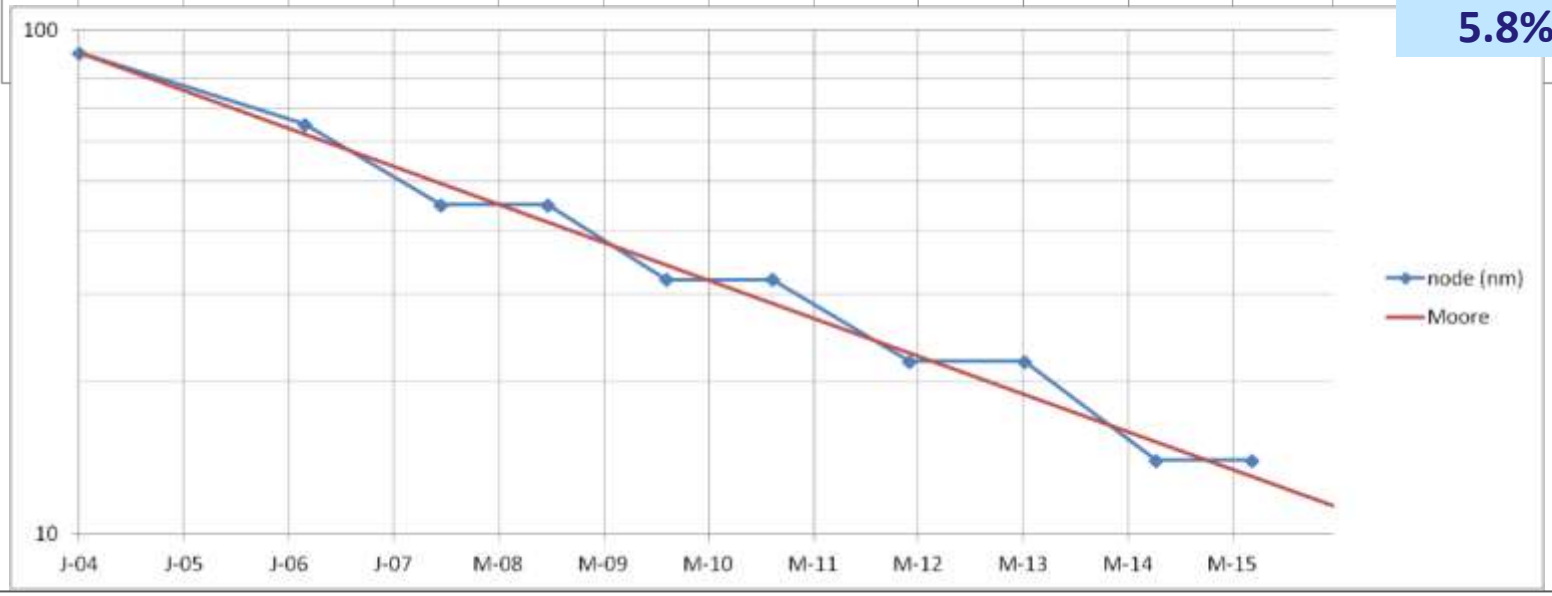
Lately
~6% / year

- Source: ISSCC 2015 Trends

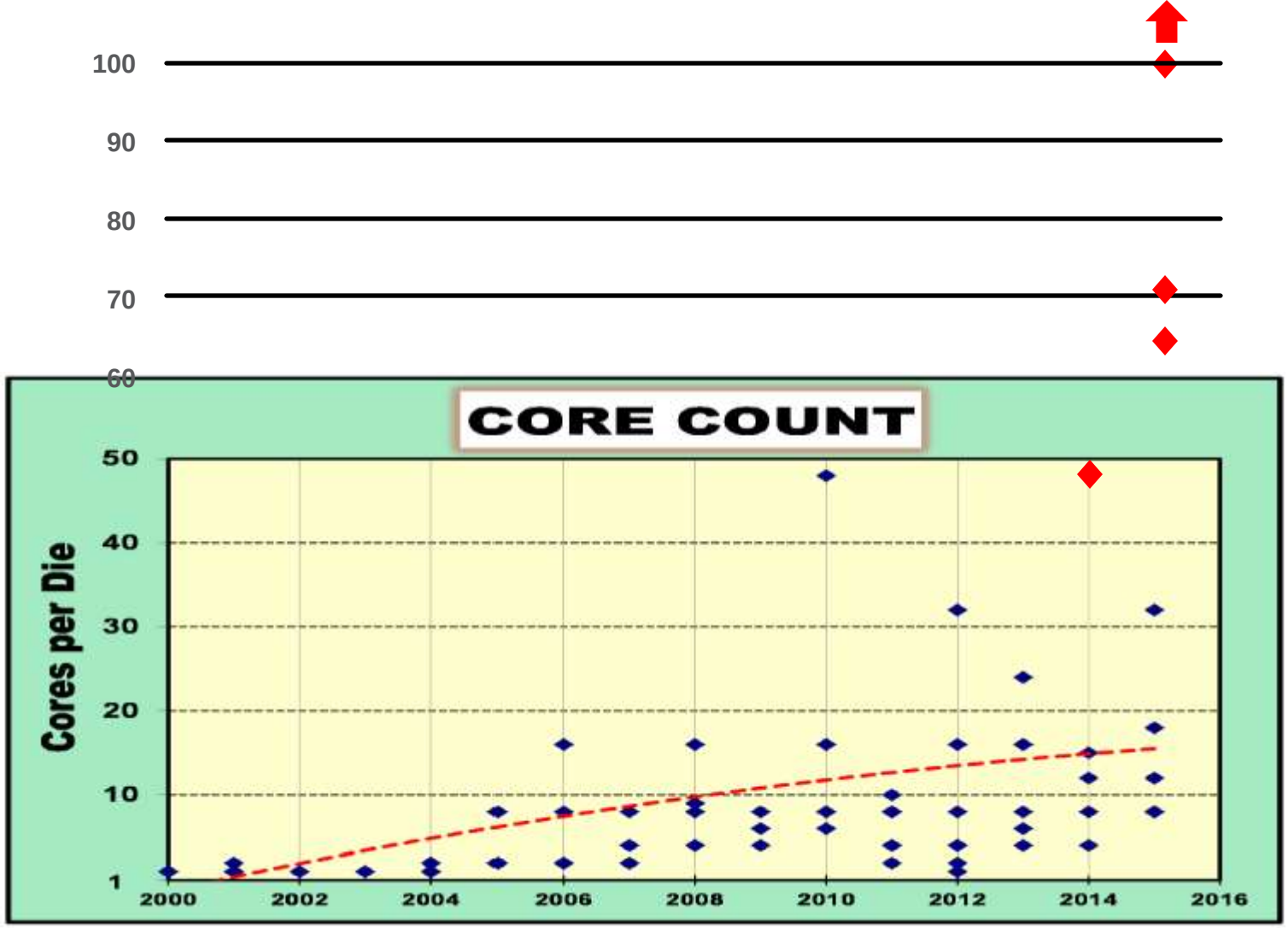
Single Core IPC Gain vs. Introduction Year for Intel Xeon Family



IPC average gain 5.8% / year



- Compiled from Wikipedia and other internet articles Sep 2015 by JTPawlowski



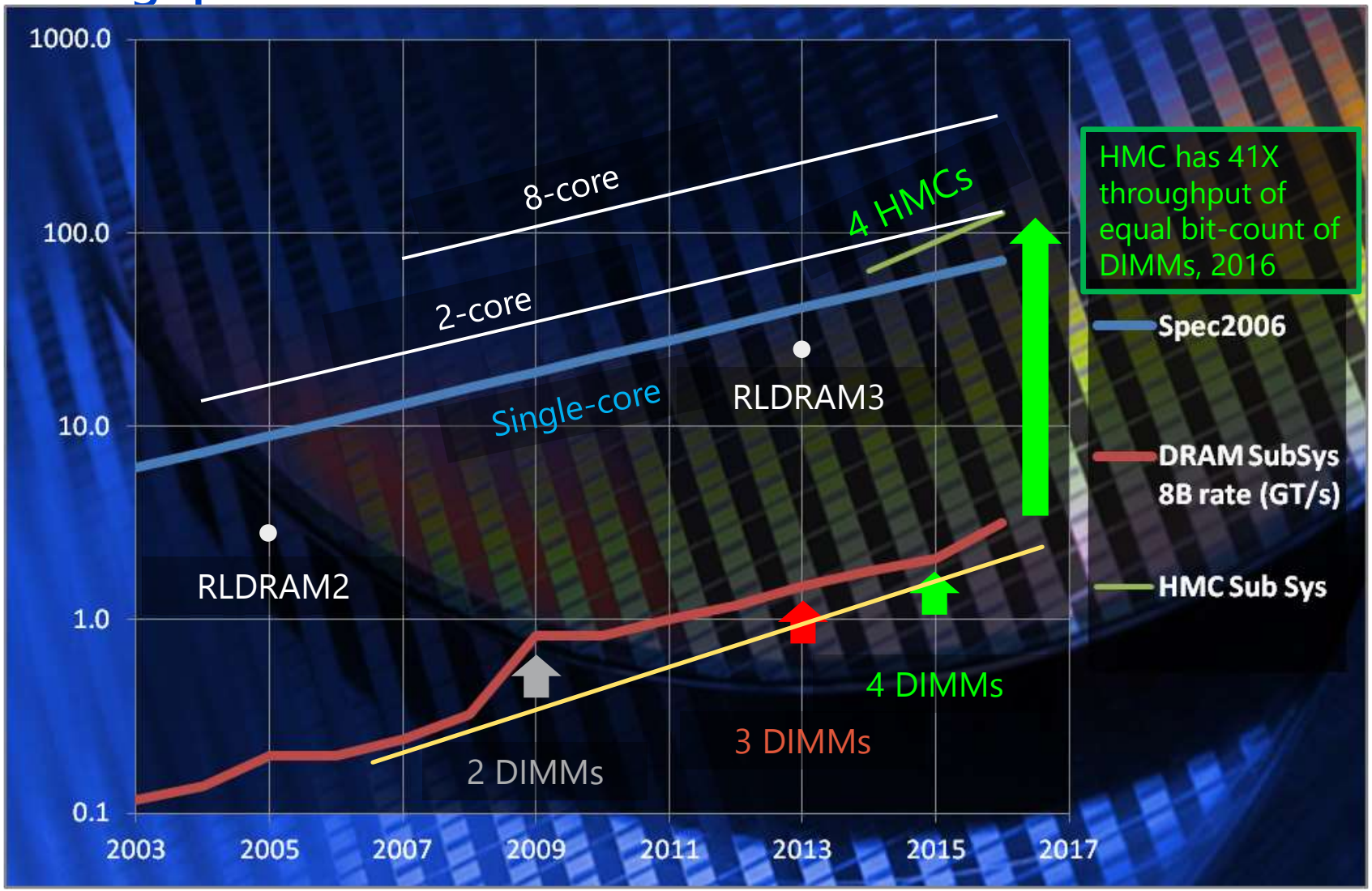
GPCPU
Core Count grows at 19% / year with high variability

GPGPU
Core Count grew at 42% / year
But now slowing to 32% / year
(JTP estimates)

- ISSCC 2015 Trends, plus: 2014 Cavium 48 core, 2015 EZChip/Tilera 100 core, Knights Landing ~72 Atom cores, Phytium 64 ARM cores.
- Off the chart: Kalray and REX 256 cores, GPUs e.g. Nvidia TitanX 3,360 cores

Impressive Gains in Memory Throughput

System Throughput a.u. vs. Time



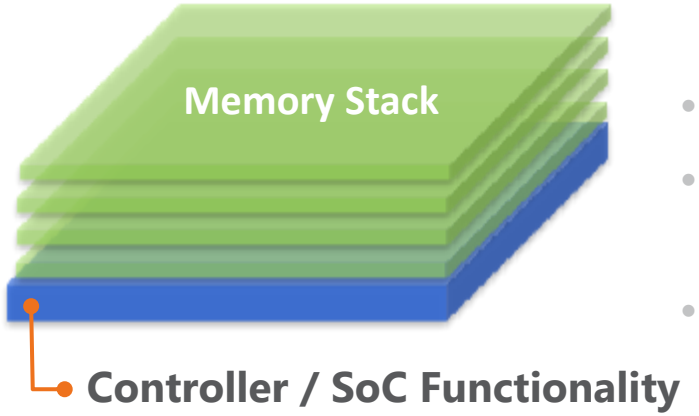
1.06 f
x
1.06 IPC
x
1.19 core count

Source: JTPawlowski, Micron

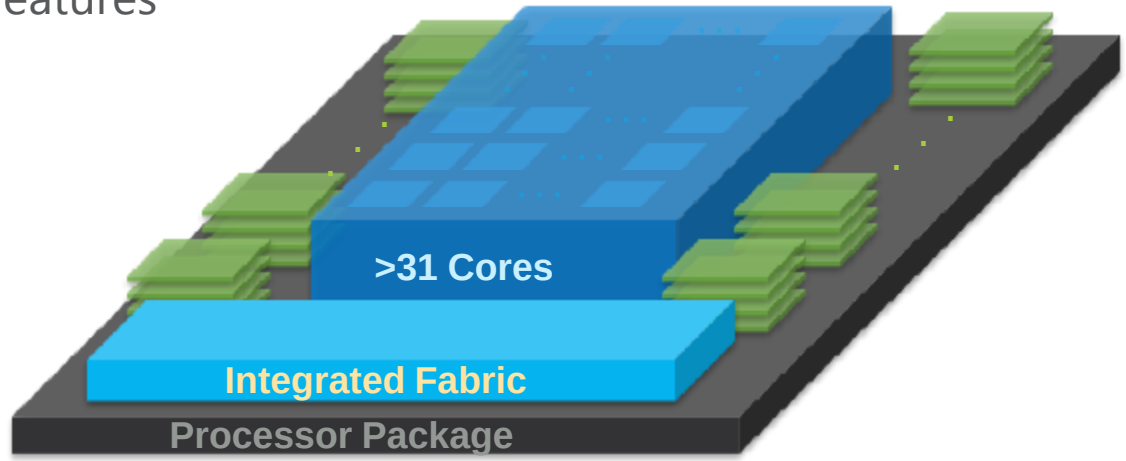
Memory Technologies

In-Package Memory

Bandwidth | **Efficiency** | **Form Factor**



- Unprecedented memory bandwidth keeps pace with multiple CPU cores
- Increased savings in energy/bit
- TSV technology with logic layer allows roadmap to higher performance and lower energy
- Compelling RAS features



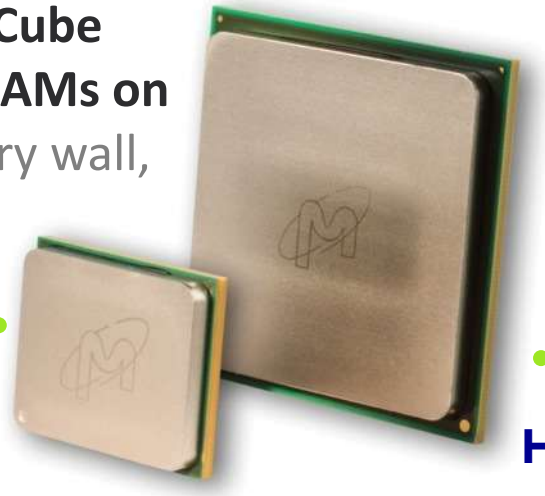
The Innovative Hybrid Memory Cube

Revolutionary Approach

- Evolutionary **DRAM roadmaps hit limitations** of bandwidth and power efficiency
- Micron introduces a new class of memory: **Hybrid Memory Cube**
- Unique **combination of DRAMs on Logic** addresses the memory wall, scalability and **RAS**

Unparalleled Performance

- Provides **15X the bandwidth** of a DDR3 module
- Uses **70% less energy per bit** than existing memory technologies
- **Reduces the memory footprint by nearly 90%** compared to today's RDIMMs



Key Applications

- **Data packet processing**, data packet buffering, and storage applications
- Enterprise and **high performance computing applications**

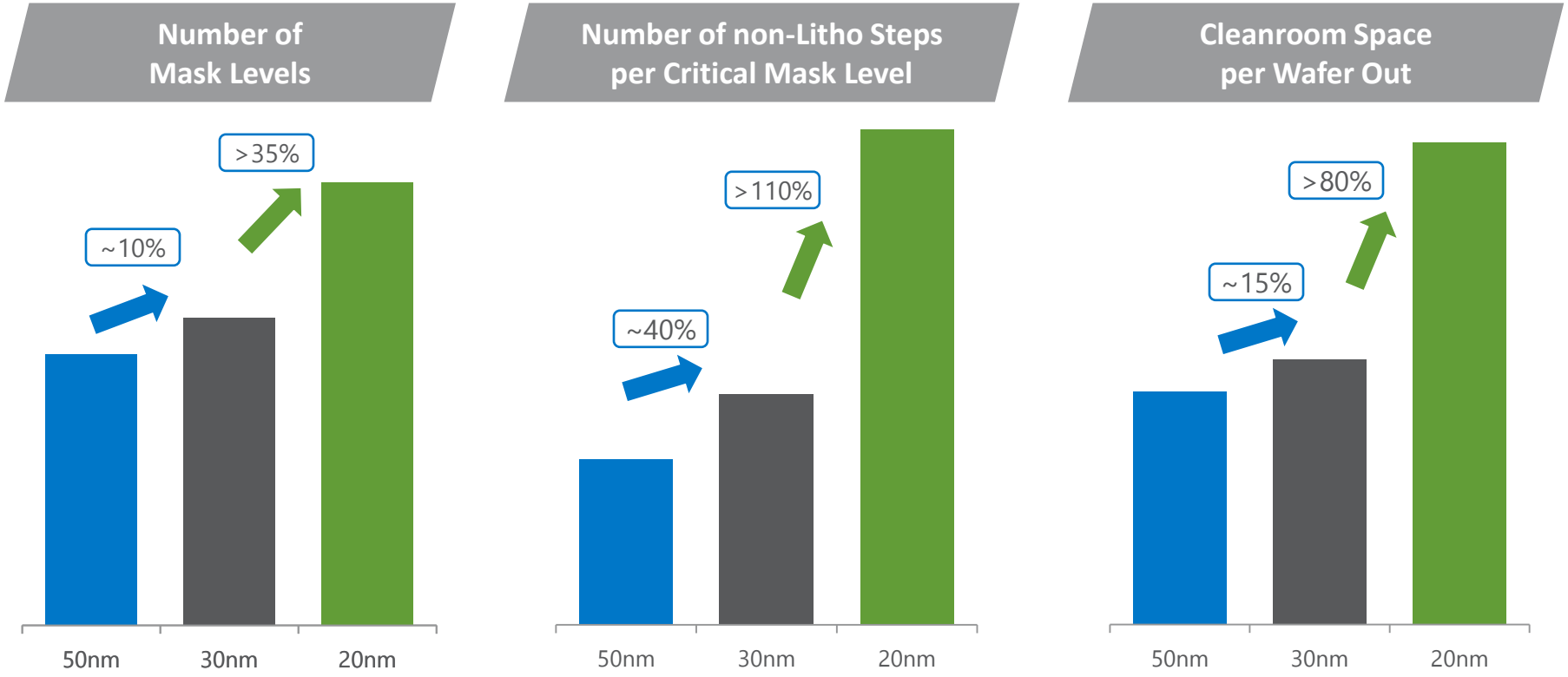
How did we do it?

- Micron-designed **logic controller**
- **High speed link** to CPU
- Massively parallel "**Through Silicon Via**" connection to DRAM

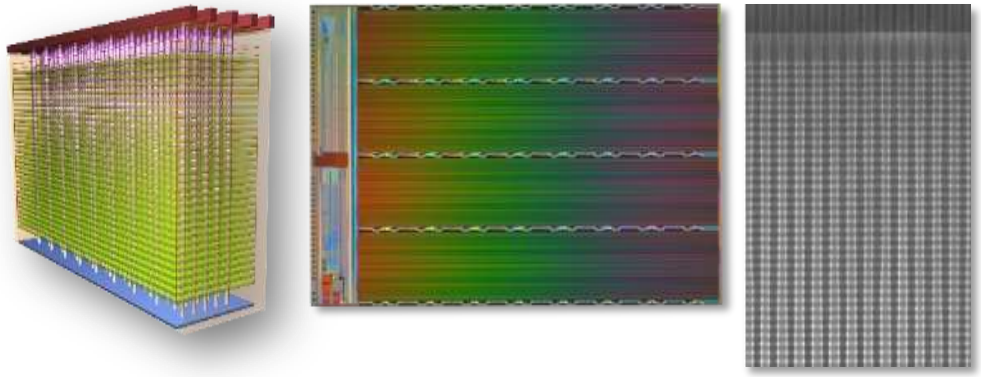
Impacts of DRAM Process Complexity

- Large increase in number of process steps to enable shrink
- Conversion Capital Expenditure scales with number of steps
- Significant reduction in wafer output per existing cleanroom area

Complexity comparison for enablement of ~100% bits/wafer increase

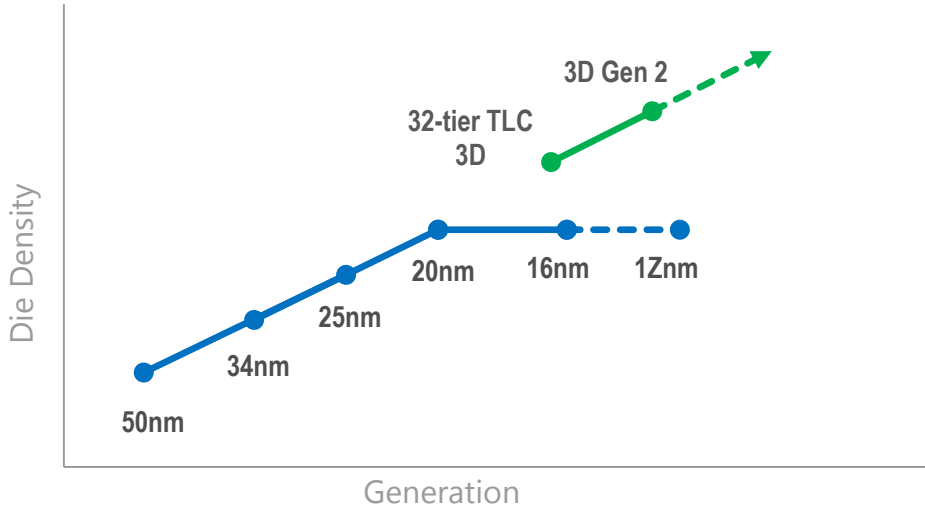


3D NAND vs. Planar NAND Scaling

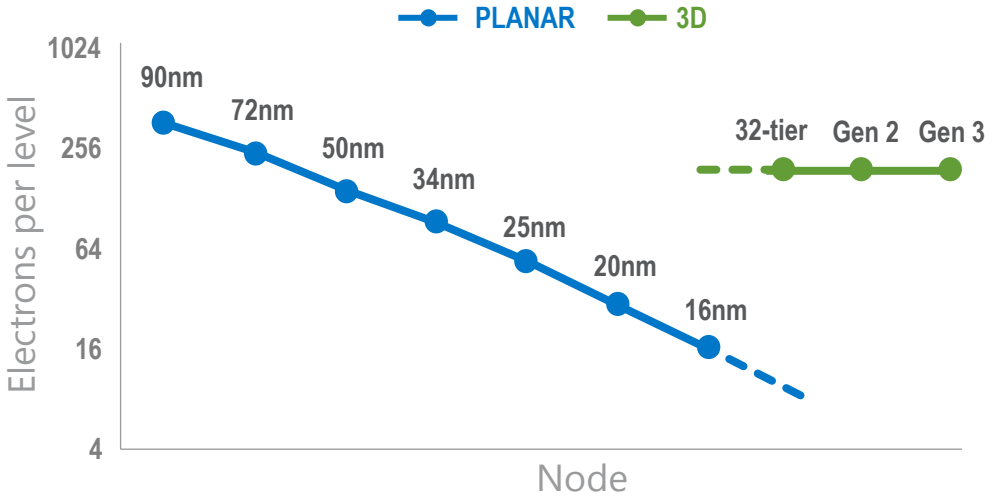


- Planar NAND scaling
  - Planar can be scaled below 16nm, but performance and cost are not competitive with 3D NAND
  - Micron focused 100% on 3D NAND after 16nm
- 3D NAND scaling
  - 3D NAND cost/capacity improvement over planar expands with subsequent nodes
  - 3D NAND cell architecture enables significant performance improvement relative to planar technology

Capacity Projection

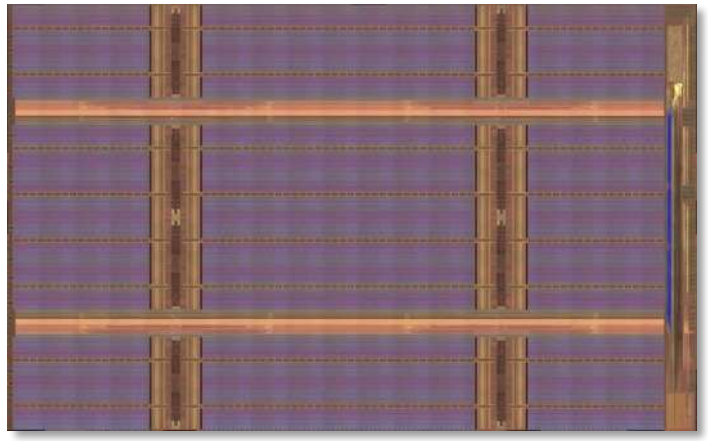


Available Electrons



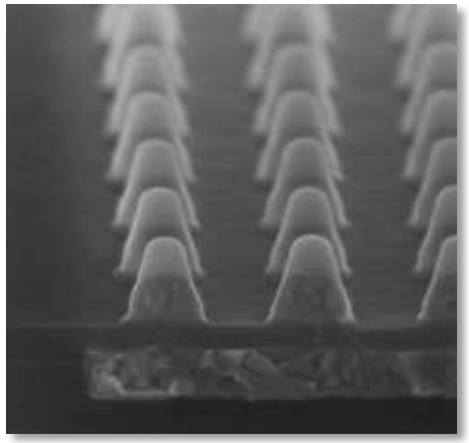
Future Memory Technologies

- Strategic investment in future memory roadmap enablement
- DRAM or DRAM replacement scaling
- Persistent Memory enablement
- Multiple generations of 3D NAND

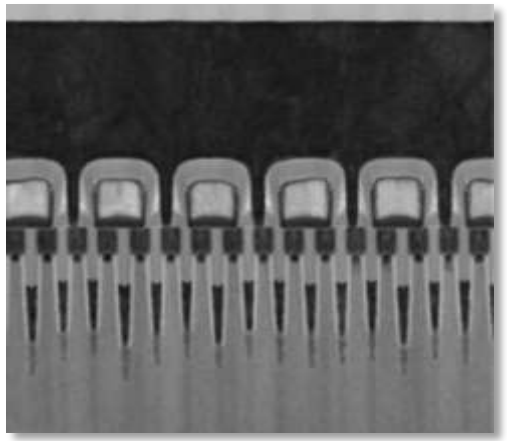


16Gb High Speed Persistent Memory

- Strategic investment in future memory core technologies
- Resistive Random Access Memories (RRAM)
- Spin Torque Transfer Random Access Memories (STTRAM)
- And others…



STTRAM Array



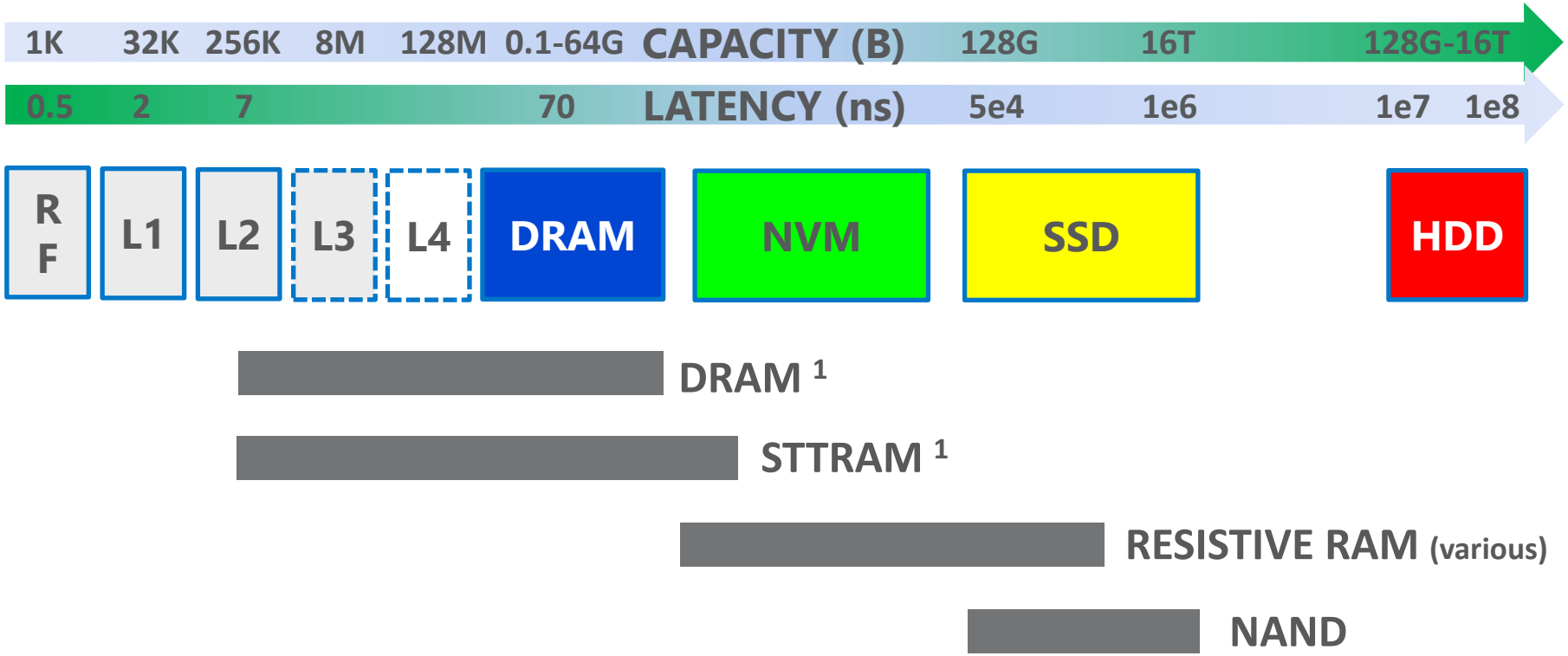
Advanced 27nm RRAM Cell

Emerging Memory

Memory / Storage Replacement Candidates

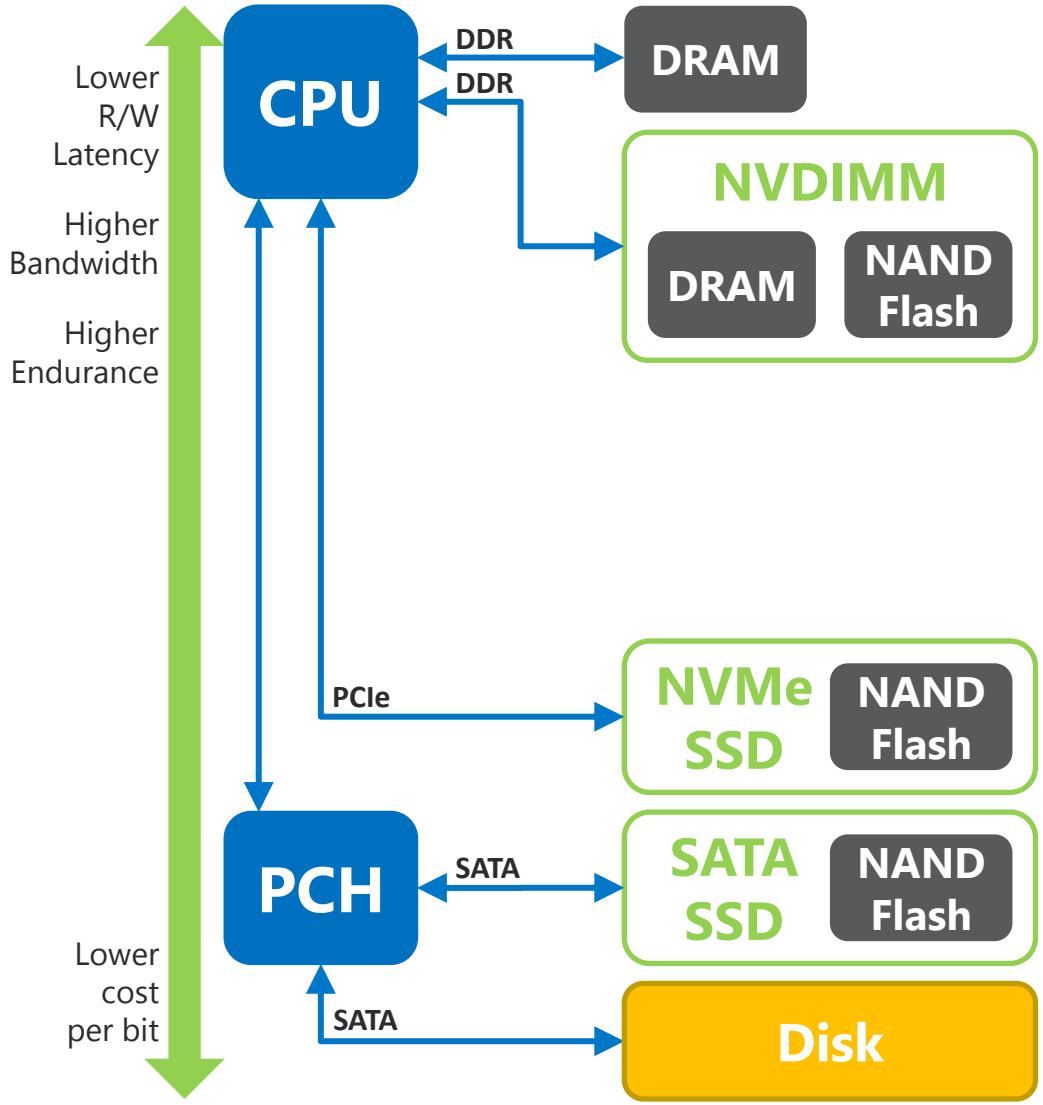
- Spin Torque Transfer RAM is public leading candidate for DRAM
- All candidates - problematic characteristics when formulated for memory
  - Greater latency tRC energy/bit
  - Lesser bandwidth endurance
  - Much greater Raw Bit Error Rate
- Unwieldy fit into existing sockets
- NAND has no realistic direct replacement candidate
  - Everything is projected to have higher cost until NAND has no scaling path
- Numerous emerging memory candidates
  - Many fall between NAND and DRAM

Technology Potential in Memory System Hierarchy



- SSD pushing HDD into cold data / archival role
- Latency gap between DRAM and SSD
- Fill this gap with a non-volatile component
- **Persistent Memory – a new class of NVM (non-volatile memory)**
- Note 1: not exclusively the style from a memory company, includes eDRAM, eSTTRAM

Nonvolatile Memories in Server Architectures



PCM: The Ideal Future Memory?

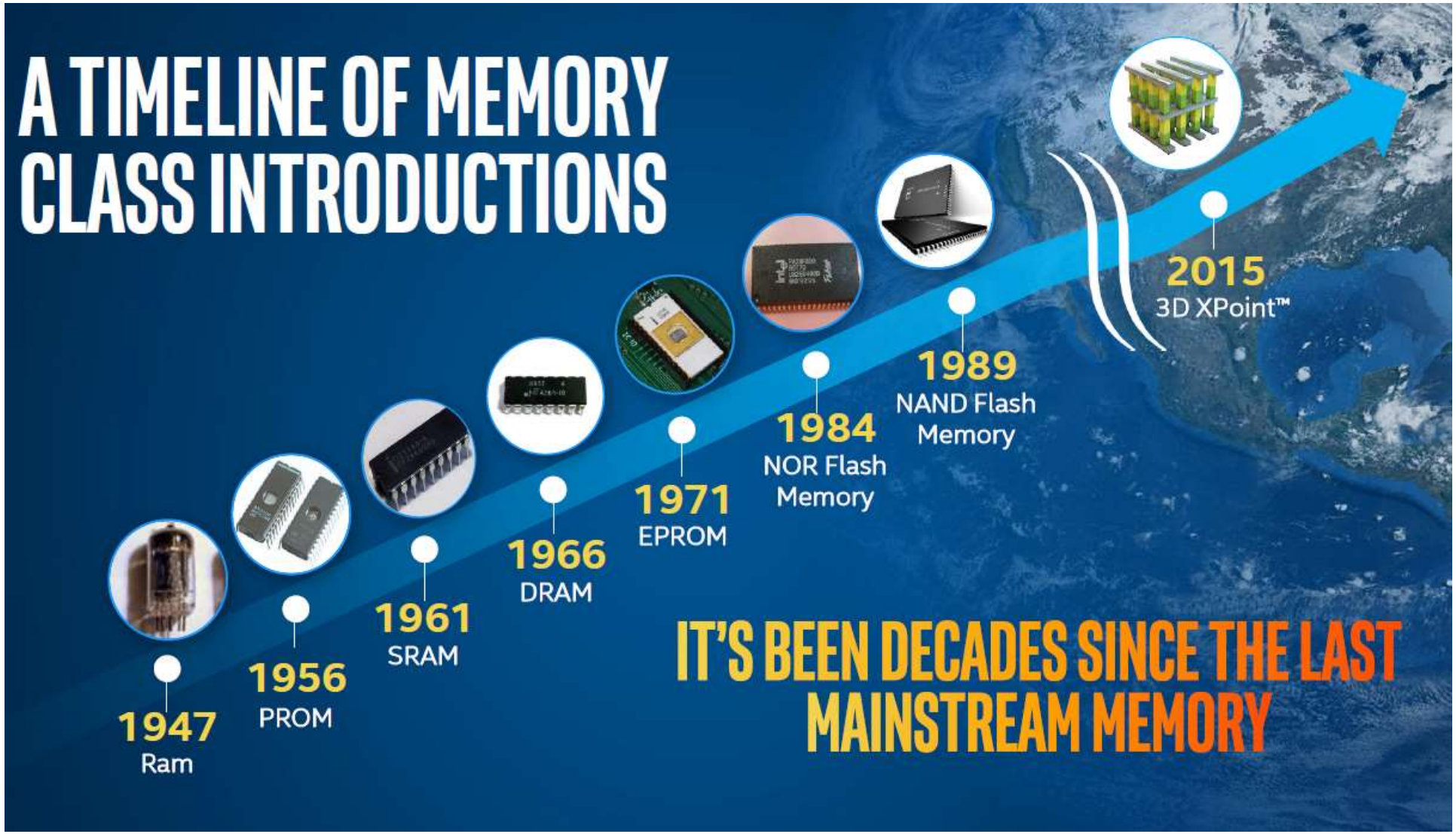
Characteristic	DRAM	PCM	NAND
Nonvolatile	No	Yes	Yes
Latency	Low, Fixed	> DRAM, Fixed	High, Variable
Read/Write Time	1:1	1:5	1:20
Block Management	No	No	Yes
Refresh	Yes	No	No
Addressability	Byte	Byte	Pages
Raw Bit Error Rate	Low	Low	High
Endurance	High	Medium	Low
Energy per Bit	Low	High	High
Cost per Bit	DRAM	~DRAM	Low

- Desire the best attributes of both NAND and DRAM
- PCM's cost per bit was too close to DRAM to justify the higher latency and reduced endurance
- But there is a range of memory possibilities which will provide compelling attributes for the right cost per bit

3D XPoint™ Memory

Micron

2015 is a Historic Year in Memory

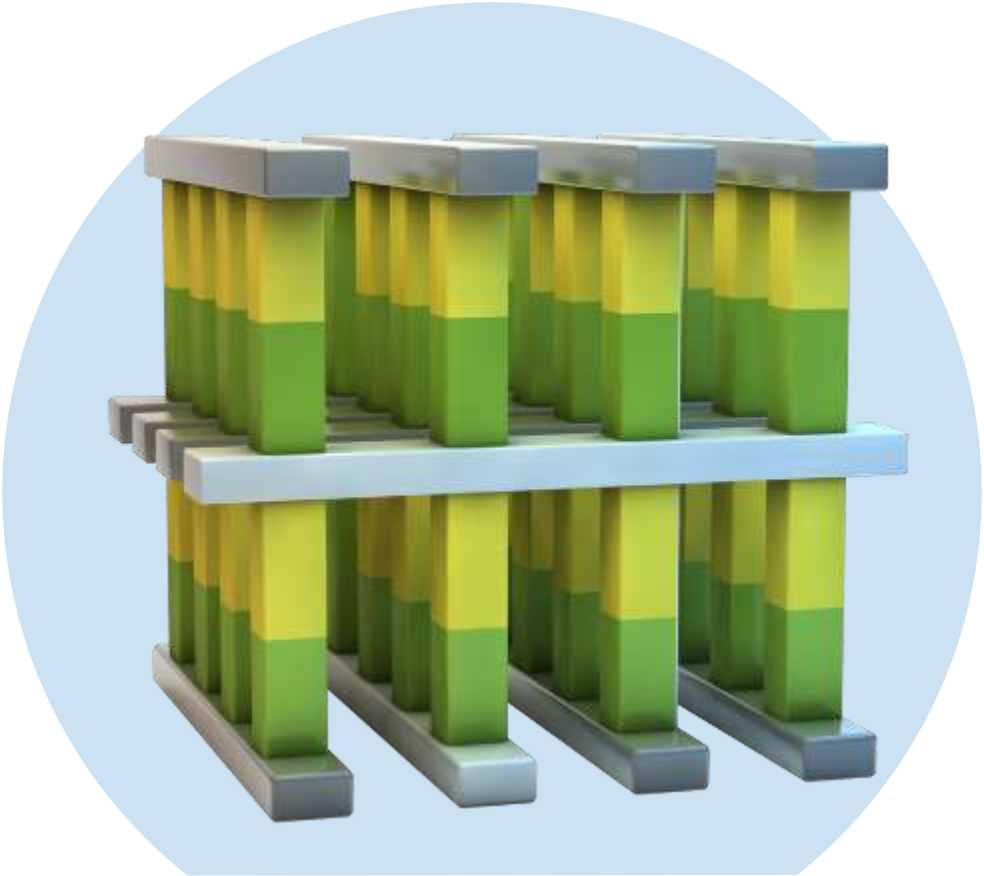


Introducing 3D XPoint™ Memory

1000X
FASTER
THAN NAND

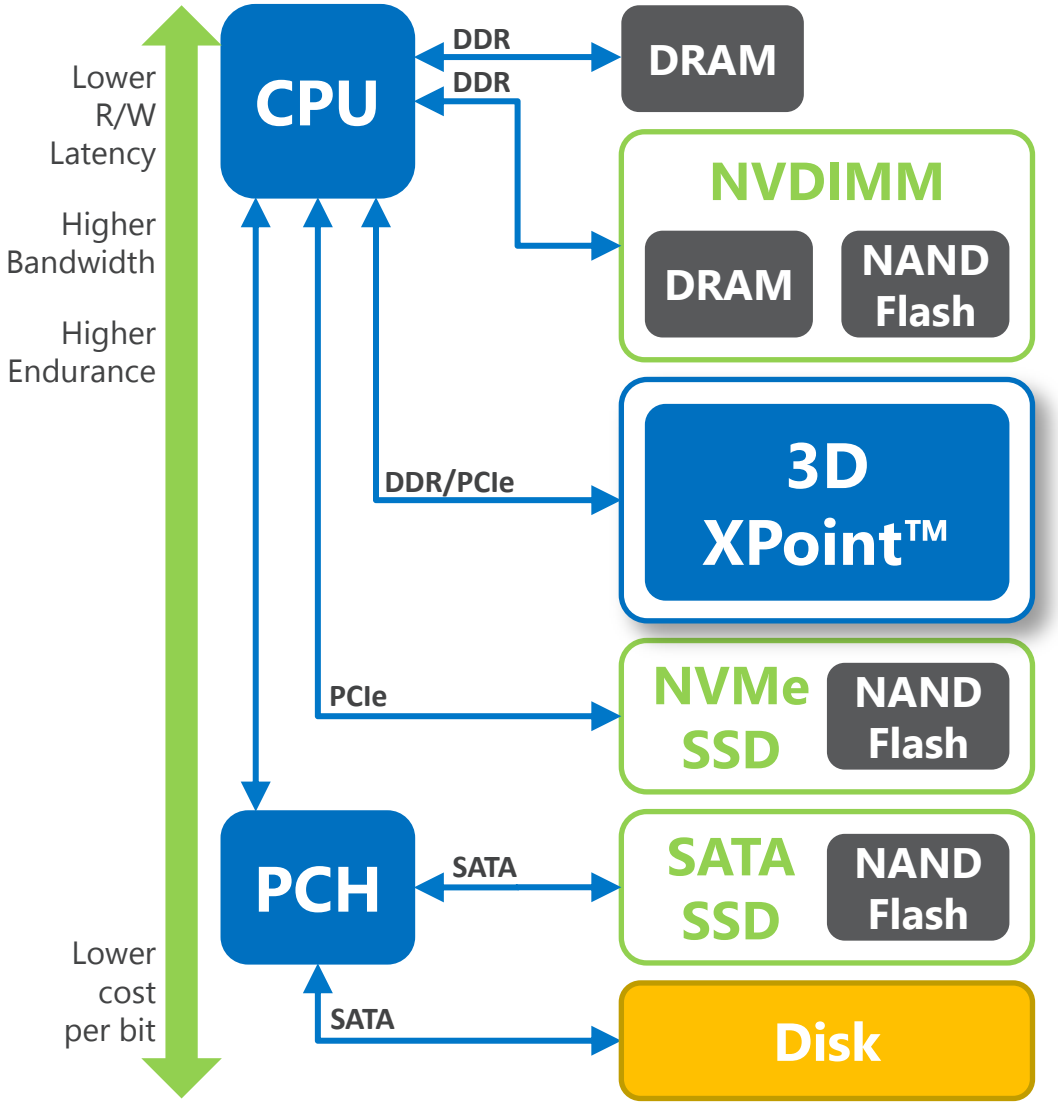
1000X
ENDURANCE
OF NAND

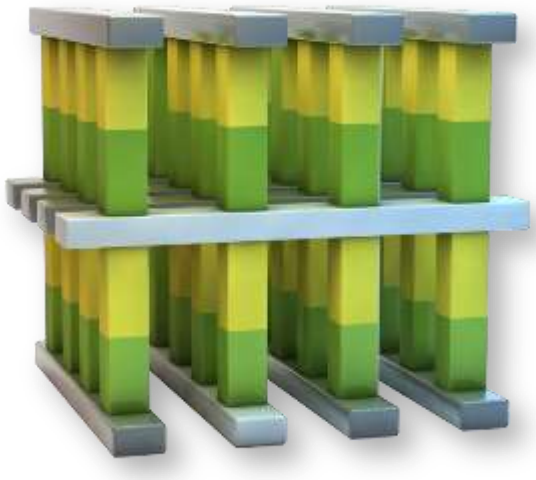
10X
DENSER
THAN CONVENTIONAL MEMORY



3D XPoint

Nonvolatile Memories in Server Architectures





- 3D XPoint™ technology provides the benefit in the middle
- It is considerably faster than NAND Flash
- Performance can be realized on PCIe or DDR buses
- Lower cost per bit than DRAM while being considerably more dense

Changing Relationships

The Great Wall is Energy

Task 256b ops	Energy (Without sequencing overhead)
Two Double Precision Floating Point Operations	15 pJ (10nm logic)
Small L1 Cache SRAM Read	30 pJ (10nm logic)
10mm move on logic die	180 pJ (10nm logic)
Low-Power discrete DRAM off-chip read	800pJ (same timeframe as 10nm logic)

Source: JTPawlowski, Micron

- Numerous walls exist in systems
- Memory throughput wall overcome by HMC and RLDRAM
- The largest system challenge is energy
  - Control overhead and Data movement
- Newest workloads: high volume random data

The 3rd Epoch of Processor-Memory Relationship

Epoch

1 In the early compute days
Processors were cheap compared with memory

2 Memory became much cheaper than processors
Software mostly compiled
Memory usage soared

3 Now using the memory is becoming more
expensive than using processors: energy!

Portends…Increasing memory area to mitigate energy
And the necessity for new fundamental solutions

SHIFT from 60 years of processor-centric compute to memory-centric will be the feature of 3rd Epoch

Core memory board

Early cores were 1/16" in diameter, could be accessed in 5 microseconds, and cost about $1 per bit. Eventually they were 4 times smaller, 10 times faster, and 100 times cheaper.

System Architectures Will Change

- Resource redistribution
- Eventually no memory device will be directly controlled by the host processor without an appropriately-tuned abstraction layer
  - Very light for the lowest capacity, lowest latency memory
  - Highly-abstracted for the highest capacity, highest latency memory
- More processing will be done using memory technology, chiefly as accelerators
  - Also much more near-memory and near-storage processing
- New mentality of “work moving”
  - Migrate control information to the epicenter of data and process there

Enormous Untapped Potential in Memories

- Huge parallelism, tiny little off-chip data pipe
- Costly to provide external access to the full capability
- Great potential for a variety of new accelerators
- And…memory tech is scaling more favorably than logic tech

Standard DDR3 DRAM

30ns to load 1 row into sense amps (tRAS) and stagger fire every 4ns

Internal Bandwidth: 4 Ti bits/sec

External Bandwidth: 25.6 Gb/s on 16b

160x untapped bandwidth (320x on 8b)

1000x

With Numerous Improvements

Can improve internal bandwidth by another 1000x

160,000x – 320,000x untapped bandwidth potential

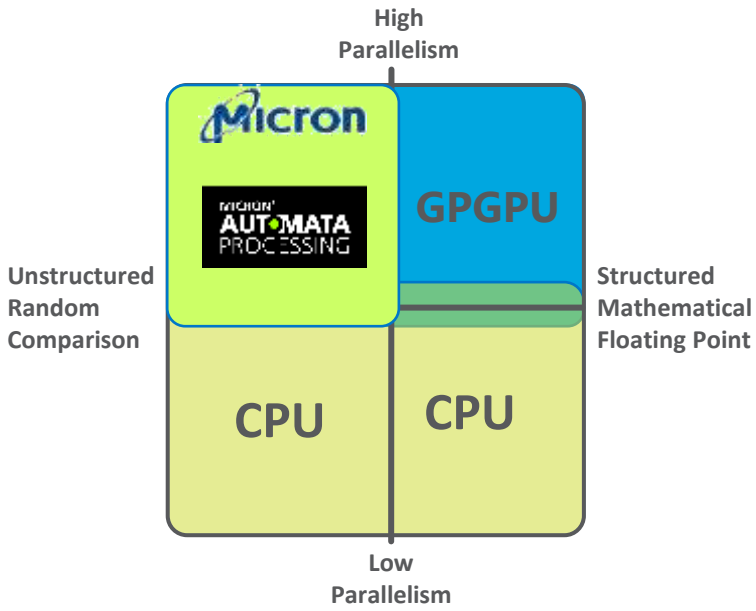
Micron Automata Processor

MICRON®
AUT•MATA
PROCESSING

Micron®

Automata Processor

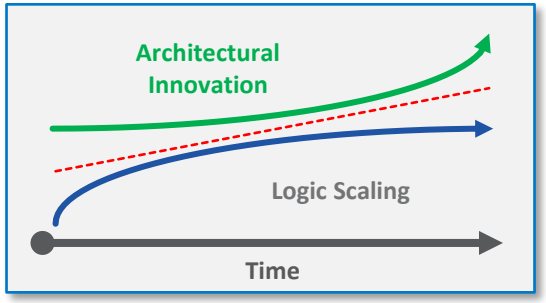
- Micron's ***Automata Processor*** is a revolutionary new class of programmable accelerator
  - An industry-first hardware implementation of highly-parallel Non-deterministic Finite Automata (NFA)
  - Orders of magnitude (>100x) faster pattern matching and graph analysis for:
    - **Financial Services Industry**
    - **Bioinformatics**
    - **Data Analytics**
    - **Network Security…and more!**



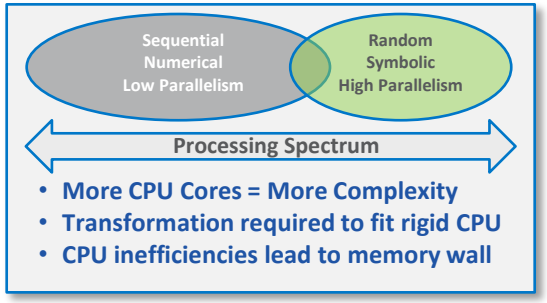
- ➢ Significantly easier to program than GPUs – natural parallelism
- ➢ Dramatically more flexible than FGPAs
- ➢ Unmatched in processing challenges such as graph analysis, fuzzy string matching, and data analytics!

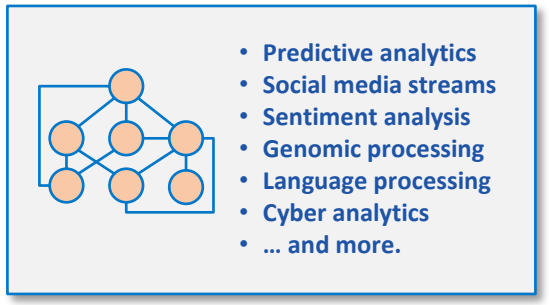
Macro Trends Aligned with the Automata Processor

Value scaling requires new approaches:

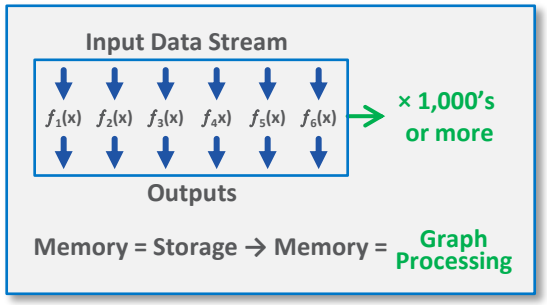


Limitations of conventional processors:





Increasing demand for graph based analytics:



Untapped potential of memory technology:

Problems Aligned with the Automata Processor

Applications requiring **deep analysis** of **data streams** containing **spatial** and **temporal** information are often **memory-bound** and will benefit from the **processing efficiency** and **parallelism** of the Automata Processor.

Network Security:
- Millions of patterns
- Real-time results
- Unstructured data

Bioinformatics:
- Large operands
- Complex patterns
- Unstructured data

Financial Services:
- Highly parallel operation
- Real-time results
- Unstructured data

Data Analytics:
- Highly parallel operation
- Real-time results
- Unstructured data

Automata Value Proposition

Fact: The ability to **generate** and **transport** unstructured data has vastly exceeded our capacity to **analyze** that same information

Market gap: a **configurable** processing engine that can quickly analyze **arbitrarily complex** graph processing problems

Micron response: A massively **parallel** non-von Neumann compute architecture that executes **multiple instructions** for each byte of data

MISD (Multiple Instruction Single Data)

Higher Performance:

- >100x performance increase for complex graphs

Lower Energy:

- As little as 0.9 pJ/calculation
- 5.8W TDP per device
- Natural 8:1 internal transport data compression
- Lower power x
  Shorter execution time x
  One-time data touch (stream)

MICRON AUTOMATA PROCESSING

Lower Cost:

- One PCIe card outperforms a cluster of 48 Xeon processors

Better Quality of Result:

- Directly analyzes complex graphs without approximations

Ease of Parallel Programming:

- No special programming considerations required to perform parallel processing
- No vectorization of data; no timing loops; no race conditions

Breakthrough Performance

Planted Motif Search Problem	Automata Processor Single PCIe Board	UCONN – BECAT Hornet Cluster	
Processors	48 AP + 1 CPU socket	48 CPU (Cluster/OpenMPI)	
Power	245-315W[1]	~2,500W[3]	
Cost	TBA	~$20,000[1]	
(25,10)	12.26 minutes[2]	20.5 minutes	1.67x
(26,11)	13.96 minutes[2]	46.9 hours	202x
(36,16)	36.22 minutes[2]	Unsolved, est. 74 years[1]	1Mx
(39,18)	~10 hours[2]	Unsolved, est. 94,000 years[2]	5Bx

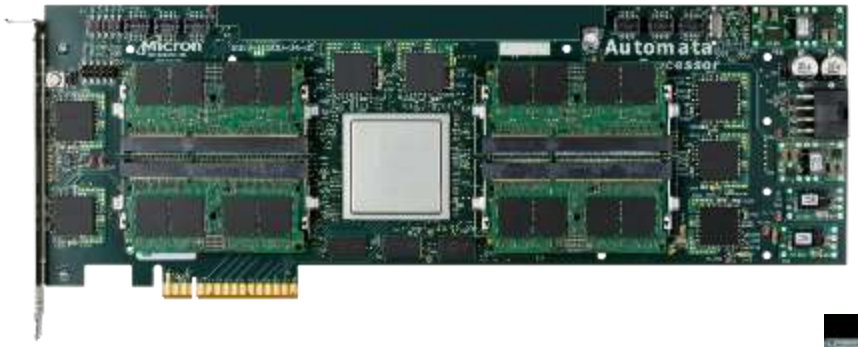
Planted Motif Search - a leading "NP Complete" problem in bioinformatics

Solutions involving high match lengths and substitution counts are often presented to HPC clusters for processing

Independent research predicts the Micron Automata Processor significantly outperforms a multi-core HPC cluster in speed, power and estimated cost

[1] Micron Technology Estimates [2] Research conducted by Georgia Tech (Roy/Aluru) [3] Only power of CPU chips, not including rest of system power such as 4GB DRAM per core

Automata Processor: Support & Tools

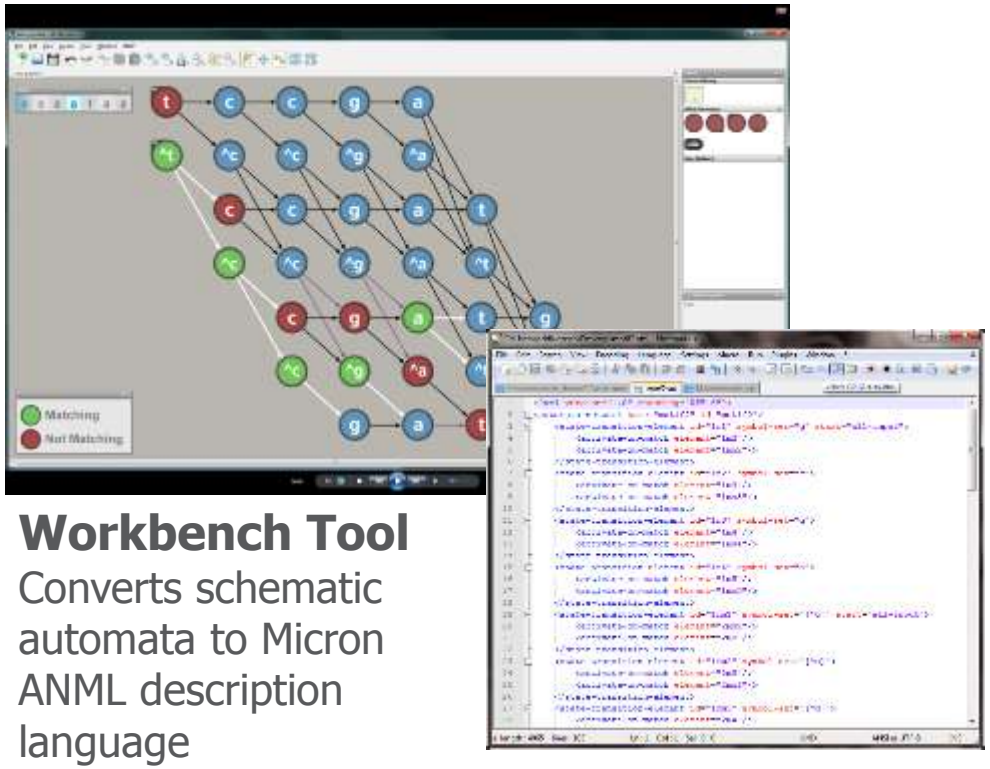


PCIe Accelerator Board

- Industry Standard PCIe (G3) bus interface
- Capacity for up to 32 AP's
- Large FPGA capacity
- DDR3 for local storage

Software Development Kit

AP Optimization, loading & debugging tools & compiler.



Workbench Tool

Converts schematic automata to Micron ANML description language

- How do we enable introduction of new memory technologies?
- How do we enable more economical forms of processing?
- How do we mitigate scaling challenges of current technologies?
- How do we deal with the economics of yield issues, especially 3D?

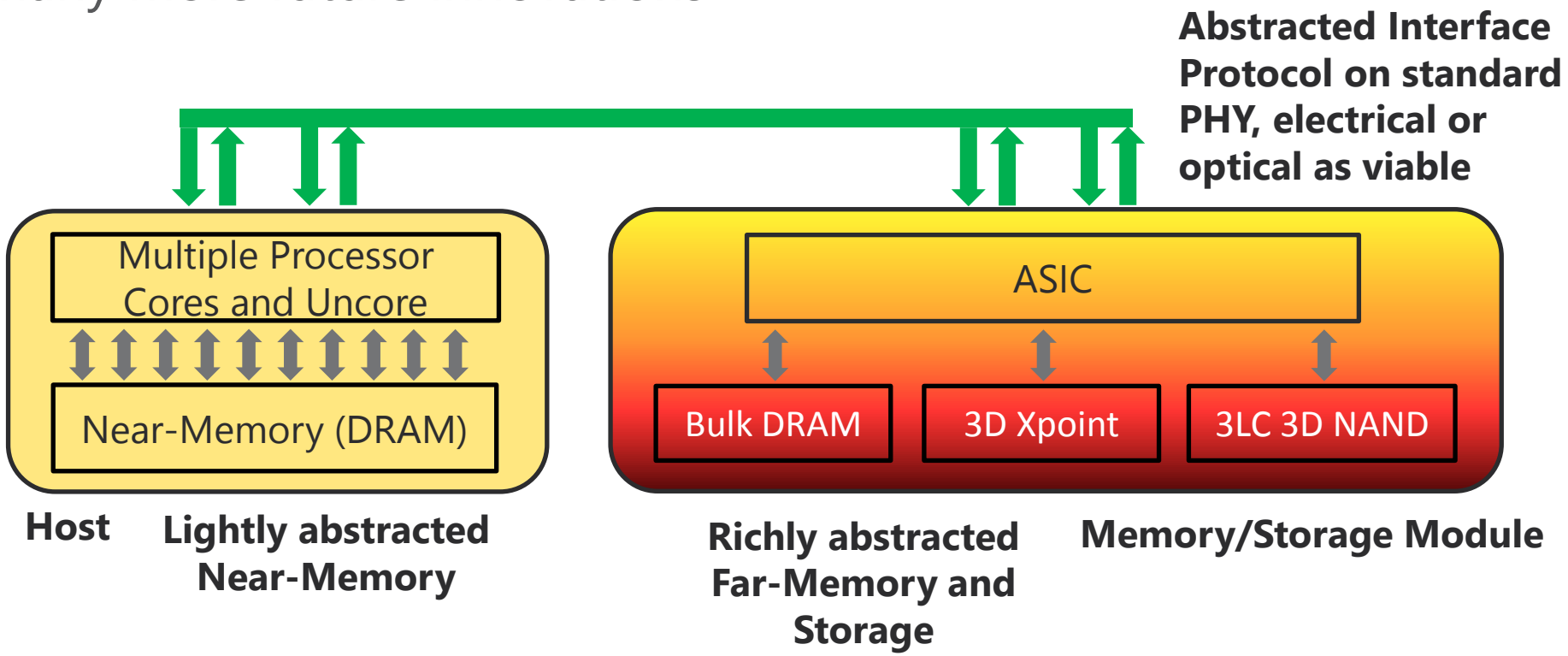

Abstraction

The Need for Abstracted Memory and Storage

- Hide device characteristics and ops not relevant to desired functions
  - Timing, refresh, bit disturb, data errors, wear mechanisms
- Mitigate future device characteristic degradation
  - Capacitive coupling increase as λ decreases, increases intrinsic errors
- Enable unconstrained innovation
- Enable introduction of new technologies, new hierarchical elements
  - New Persistent Memory, DRAM replacement, NAND replacement
  - No need to sync all memory vendors and all users
- Facilitate "bit blending" from numerous tech's
  - Optimal performance/$ = right proportioning
- Facilitate more operations closer to, or in, memory

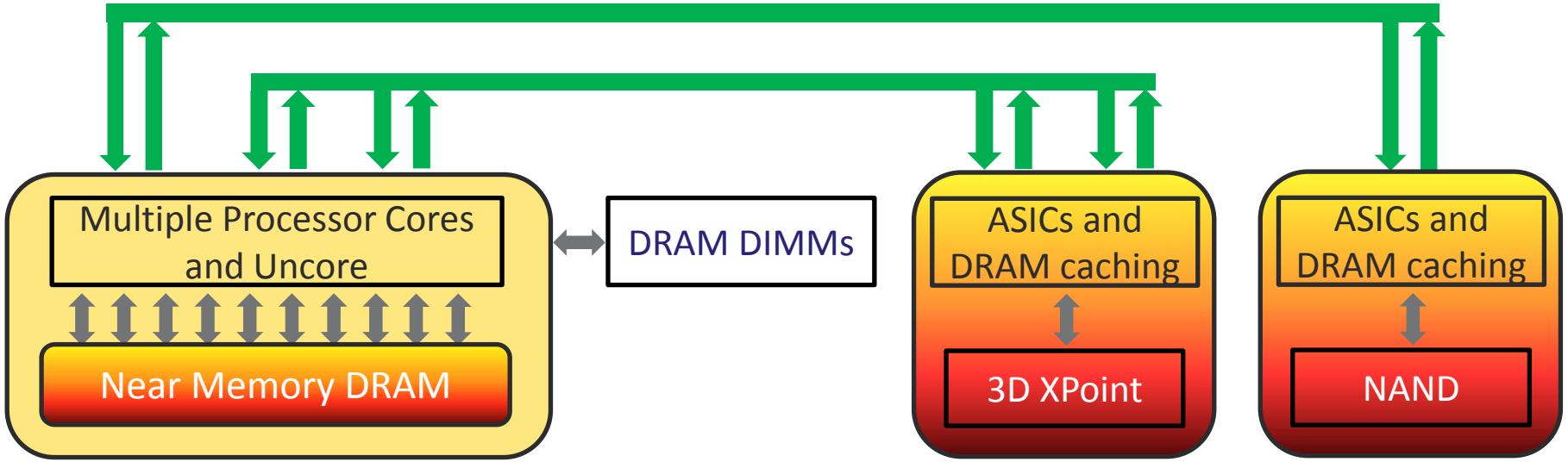
Future System

- Abstraction enables introduction of new memory types
- And many more future innovations



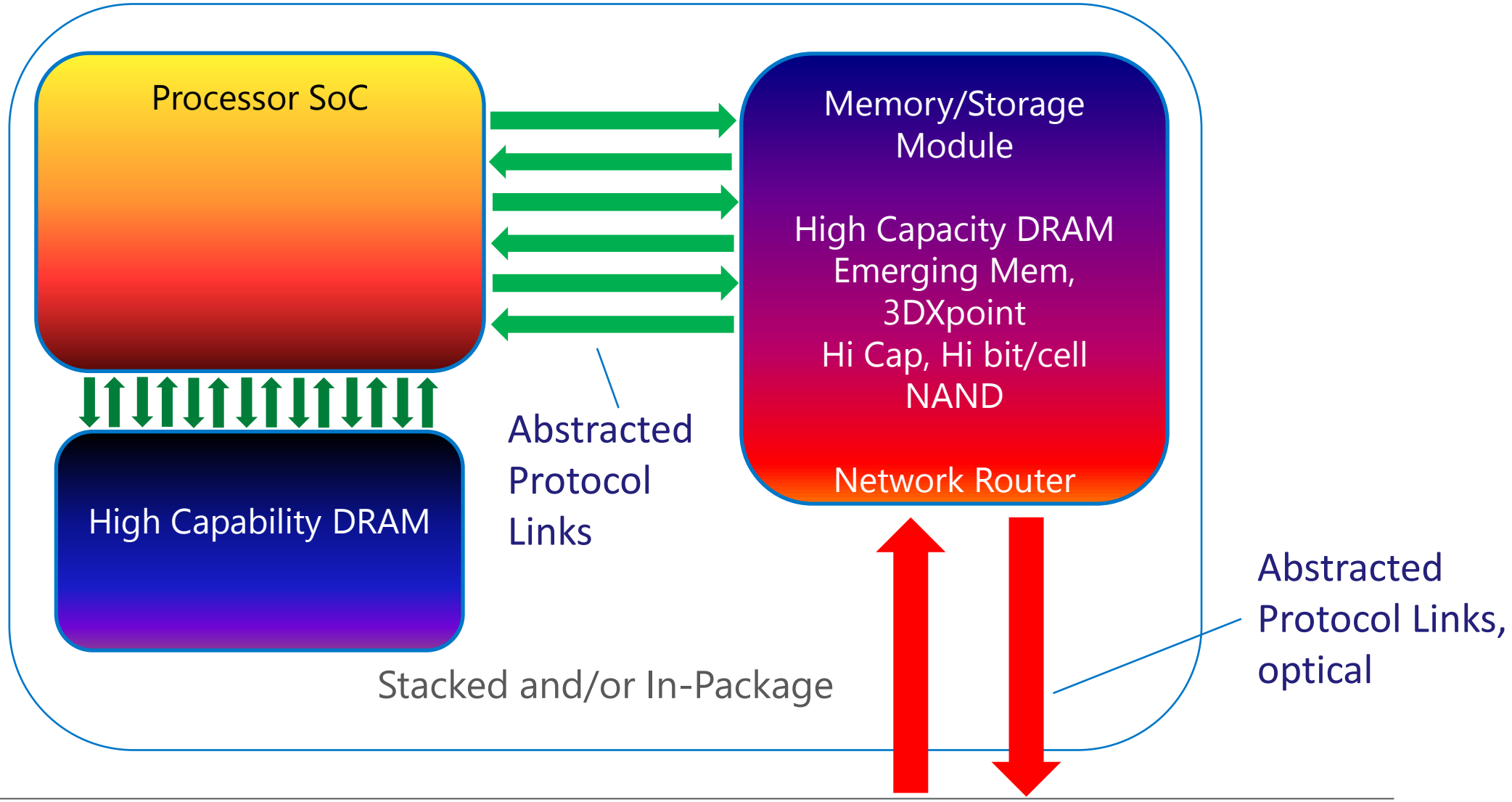
- How much abstraction? The lower the latency, the lighter it must be

Near-term System



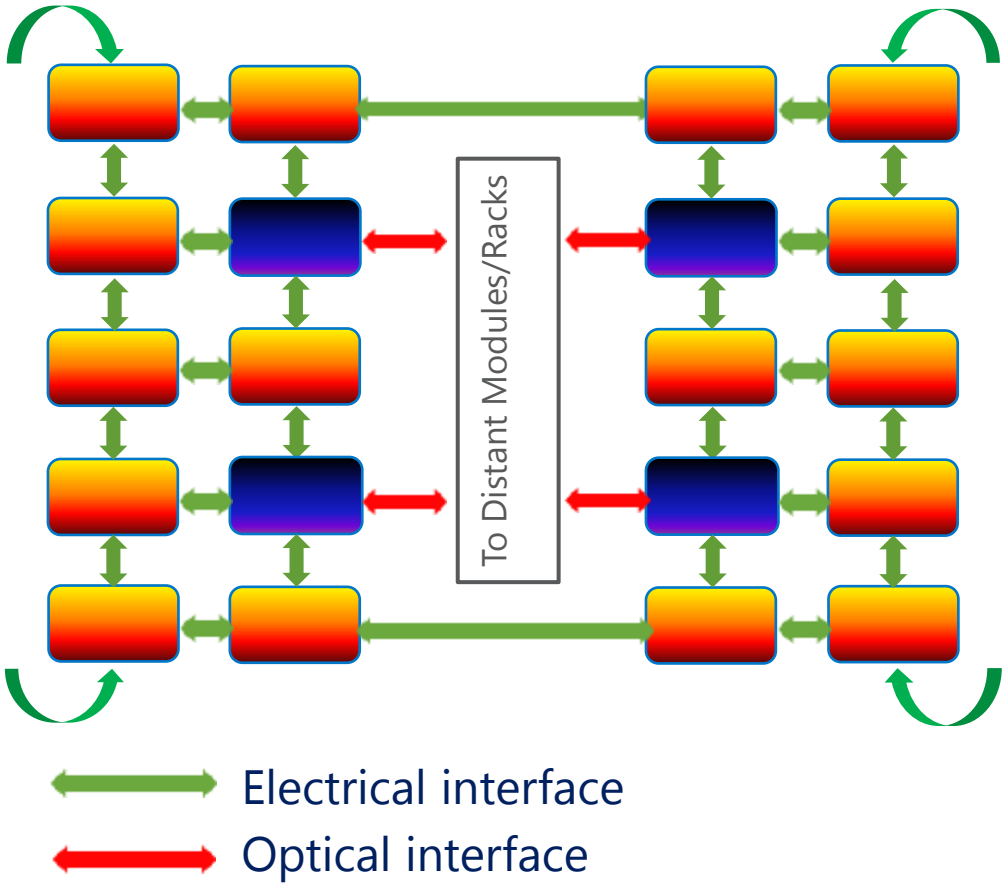
Host with Fast Near Memory	Working Memory	Far Memory	Storage
Fast DRAM Latency to be announced	Bulk DRAM 70ns latency	3D XPoint Latency to be announced	SSD > 50,000 ns latency

Future Integrated Processor/Memory/Storage



Future Hybrid Electrical/Optical Distributed System

- Memory and processor types of any kind in same logical module
  - Processor, accelerator, memory, storage, router, etc.
- All longer reach links – optical, shorter links – most economical
- "Same" protocol on all, i.e. same root protocol



- Logic device scaling continues but with sub-historic benefits
- eDRAM scaling is especially unfavorable: maybe 0 or 1new eDRAM node to come
- Memory tech scaling also with challenges, continues to pull away ahead of logic tech
- Importance of 3D monolithic and 3D-stacked die
- Emerging memory: none for direct replacements coming soon
  but will emerge for different functionality
- New class of memory: 3D Xpoint™ technology
- Energy is the Great Wall
- New Epoch: shift to memory-centric computing, more use of memory technology
- Example: Micron Automata Processor with >100x conventional performance, UNBOUNDED!
- More processing in/near memory, more sensible distribution of components and work
- Memory-Storage-Processing Abstraction mitigates many issues, enables a whole new world

Automata Resources

- SDK Toolchain Download

  http://www.micronautomata.com
- Automata simulator & tutorial : http://www.micronautomata.com/#applications
- Networking - Deep packet inspection for malware signature detection : http://www.micronautomata.com/network_security
- Bioinformatics - DNA sequencing & Motif search : http://www.micronautomata.com/bioinformatics
- More resources :
  - www.micronautomata.com
  - www.cap.virginia.edu

Micron®