Flash Memory SUMMIT

IBM

Flash & DRAM Si Scaling Challenges, Emerging Non-Volatile Memory Technology Enablement - Implications to Enterprise Storage and Server Compute systems

Jung H. Yoon, Hillery C. Hunter, Gary A. Tressler - IBM Corporation

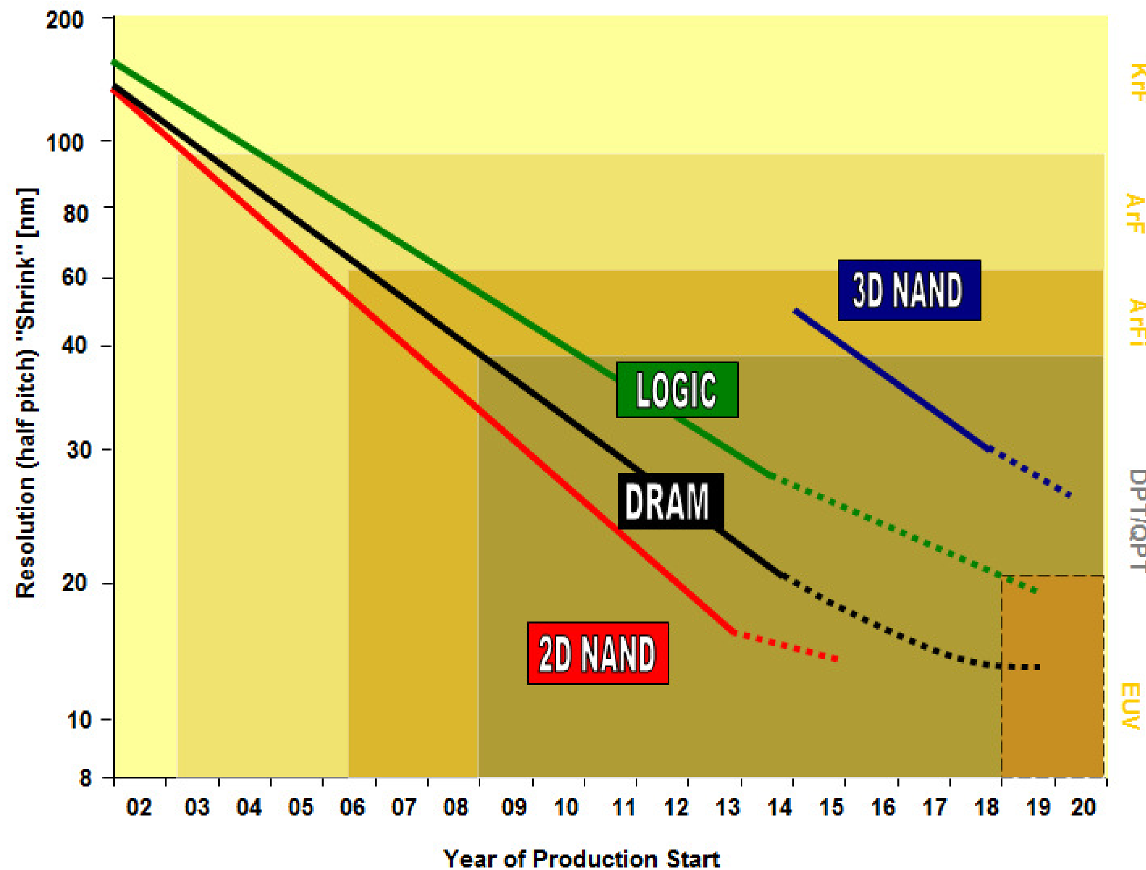
© 2013 IBM Corporation

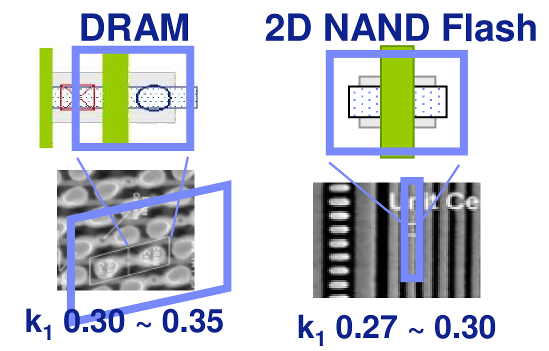
Outline

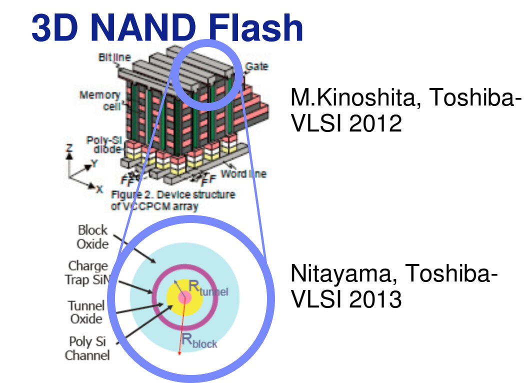
- Flash & DRAM scaling directions & challenges
- Server Storage & Compute Memory Requirements
- Emerging Memory opportunities & challenges
- Summary

© 2013 IBM Corporation

Si Technology Scaling Trends & Outlook





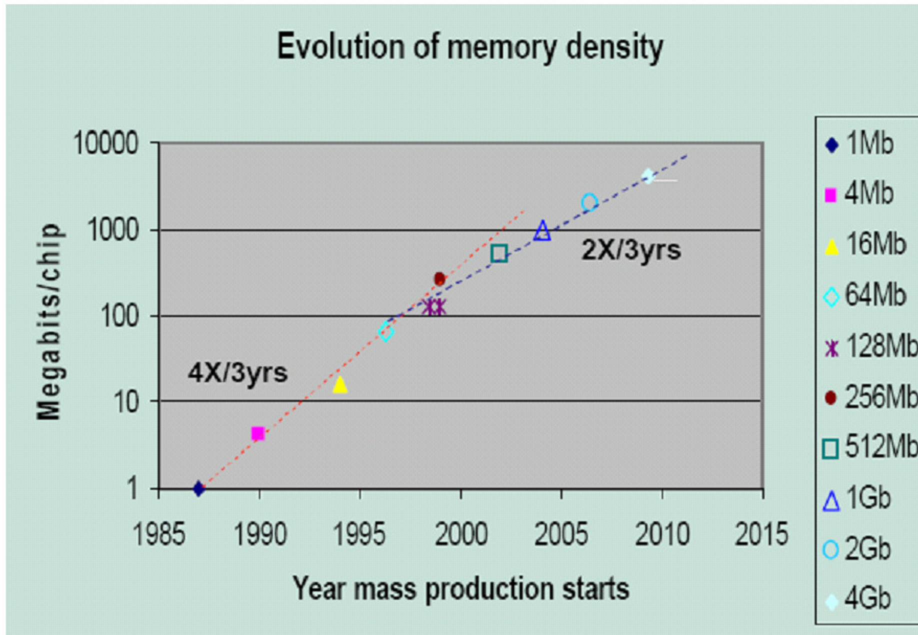


- NAND continue to drive Si minimum feature scaling – 2H13 NAND @ 1xnm and DRAM @ 2xnm node in mass production
- NAND Floating gate scaling extends to 1y nm > transition to 3D NAND cell in 2014-2015 timeframe enabling path for ‘Effective Sub10nm’ scaling
- Reduced rate & level of Flash technology shrinks @ sub 20nm - maintain cost reduction cadence, deal with scaling challenges incrementally

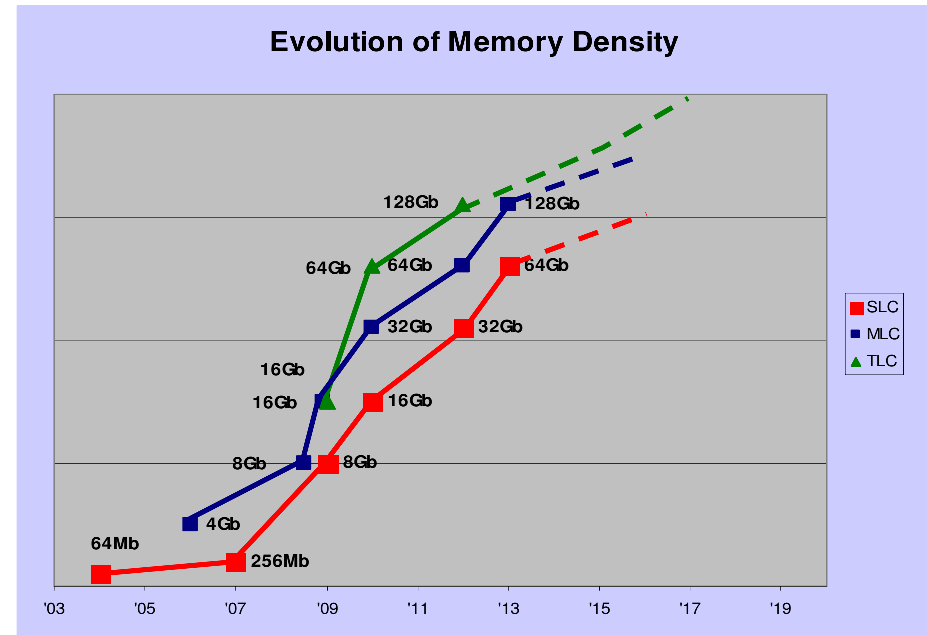
© 2013 IBM Corporation

DRAM & Flash : Density trends

DRAM



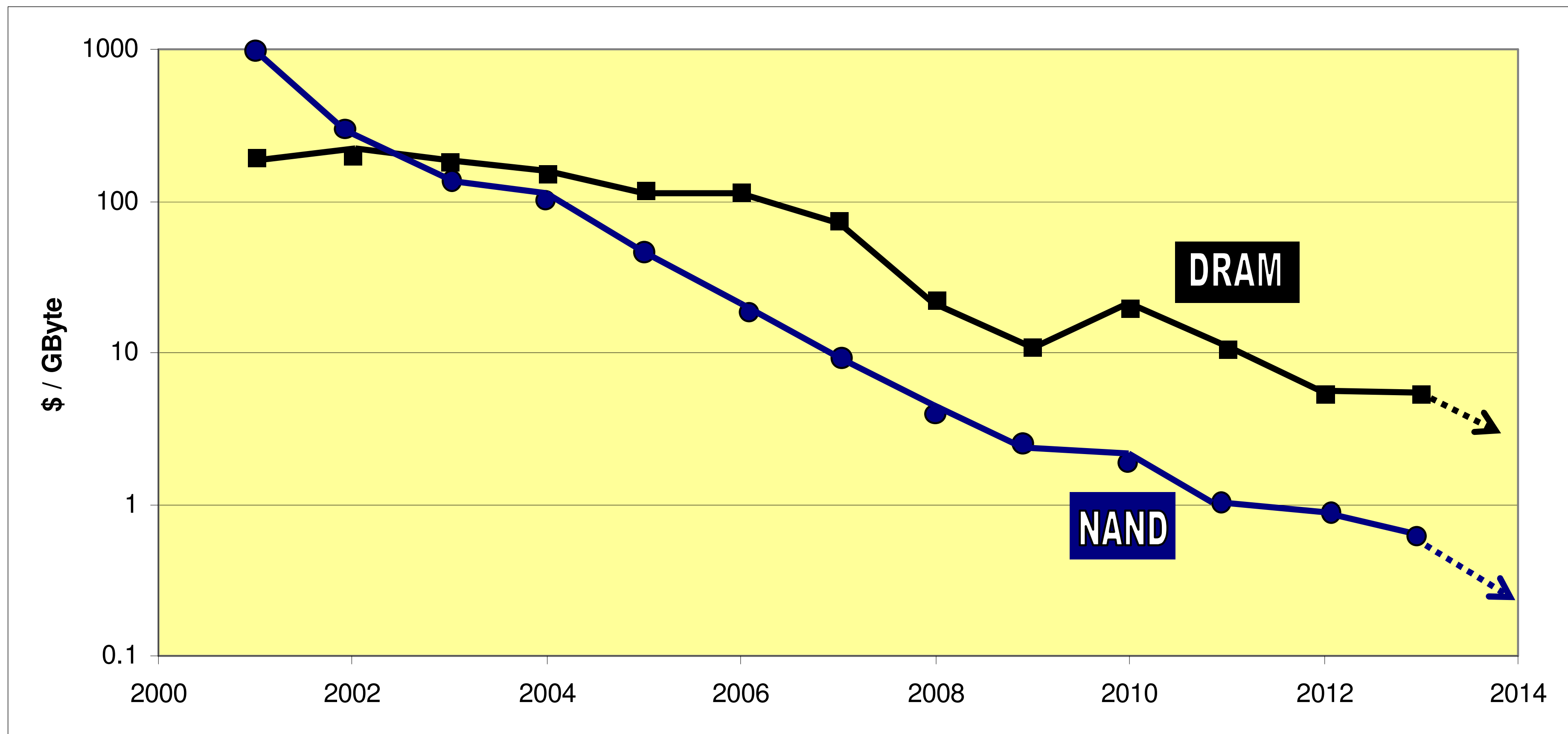
Flash



- DRAM scaling at sub 20nm node for 8/16Gbit – bit cost reduction vs. increased technology complexity & fab investment requirements
- Flash scaling at sub 1x nm node for 256Gb & higher density Flash – 3D NAND technology enablement, yield and quality maturity is key

© 2013 IBM Corporation

NAND & DRAM Cost per GB Trend*



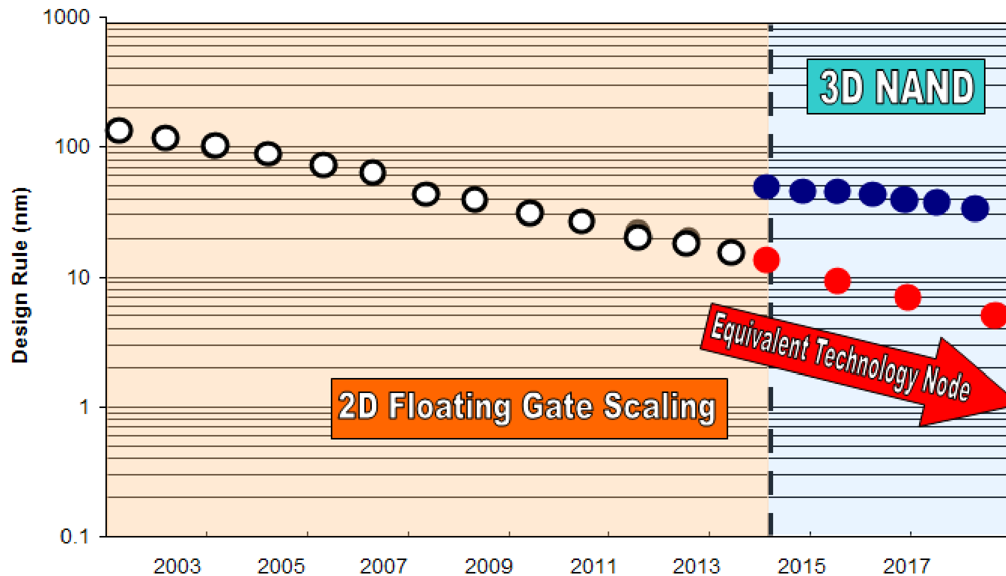
* Commercial grade NAND/DRAM pricing

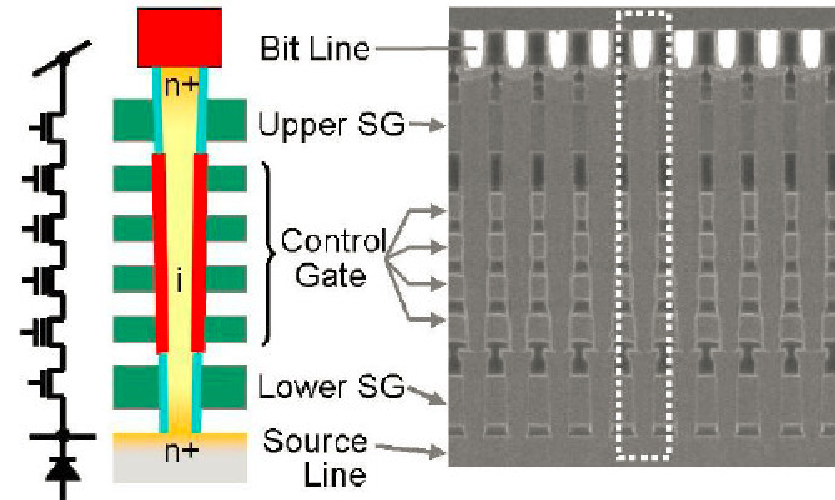
Source = DRAM Exchange/IDC/ASML

- NAND & DRAM market driven by cost per bit – scaling fueled by die cost reduction opportunities
- NAND flash cost slope fueled by Mobile & Consumer applications in 2008-2014 timeframe
  - SSD cost reduction driven by NAND scaling

© 2013 IBM Corporation

NAND Scaling Outlook





Fukuzumi, Toshiba - IEDM Tech. Dig, 2007

- 2D Floating Gate Flash expected to face significant scaling challenges below 15nm Technology Node
- 2D scaling challenges driven by significantly degrading reliability and cell-to-cell interference effects causing degradation in operating window
- 3D NAND allows ‘Effective Flash scaling’ into sub 10nm
  - 3D-NAND offers cost effective path for sub 10nm Flash scaling – non EUV based solution, # of litho layers
  - Manufacturing, Yield & Quality critical learning must occur
  - 3D-NAND long term scalability may be challenging for highly stack cells
- Resistive RAM (ReRAM) in development as post 3D-NAND memory technology

© 2013 IBM Corporation

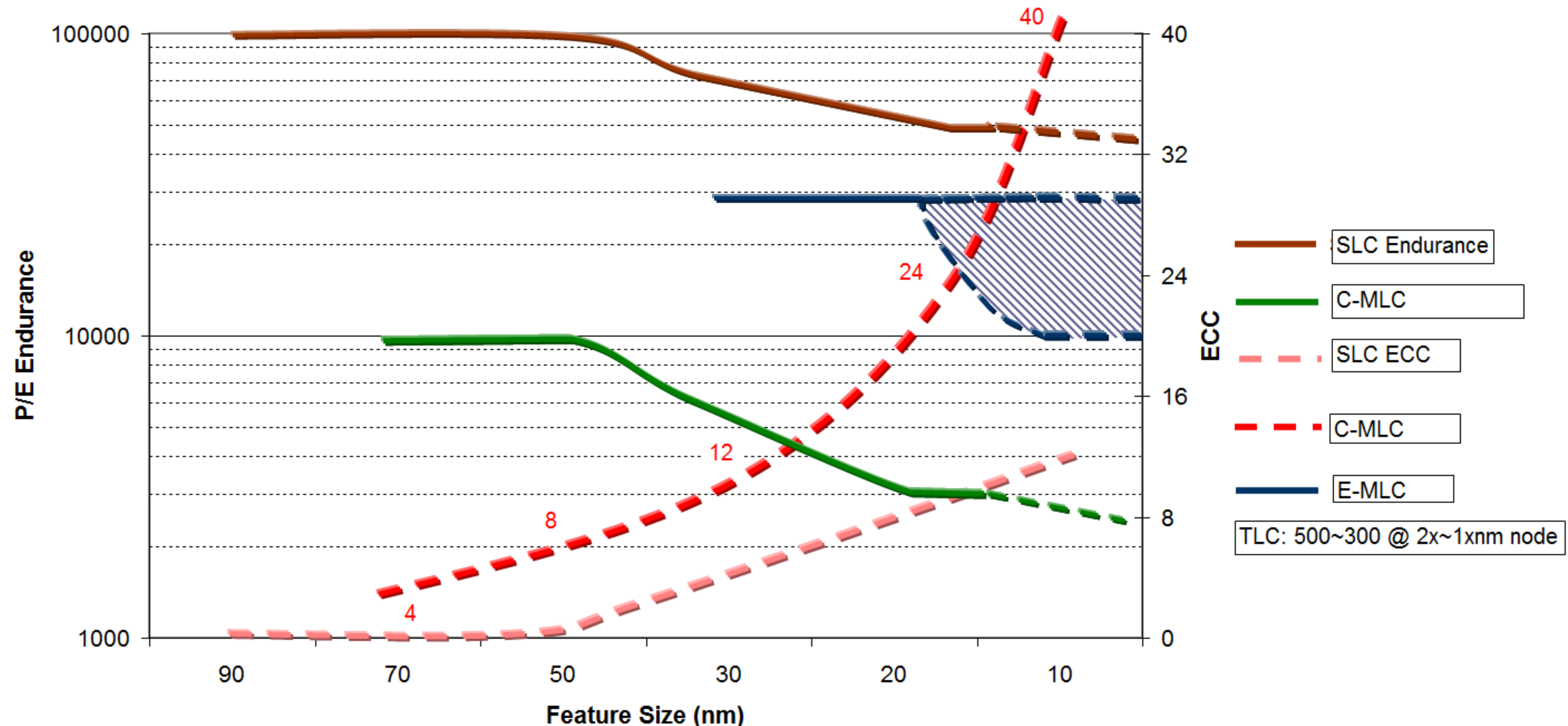
3D NAND Cell Structures

Overview of 3D NAND Designs				
3D Cell (Supplier)	p-BiCS (A)	TCAT (B)	Smart (C)	(D)
Structure	Tanaka. H. VLSI 2007	J. Jang, VLSI 2009	Choi, IEDM 2012	G. Hawk, FMS 2011
Charge Storage	SONOS	SONOS	SONOS	-
Key Issues	High Etching A/R	Very High Etching A/R	High Etching A/R	-
	Memory Hole RIE	Memory Hole RIE WL Separation	Memory Hole RIE WL Separation	-

- 3D NAND Technical Challenges
  – Yield – Multi-stack patterning, Metrology, Defect monitoring deep within 3D structure
  – Data Retention – Cell optimization critical

© 2013 IBM Corporation

NAND (cMLC, eMLC, SLC) Endurance & ECC Trends



- Endurance & Data Retention cycles have been decreasing with process shrinks – 5x~2xnm node
- ECC requirements increasing exponentially with process shrinks
- Advanced ECC algorithms, signal processing techniques critical for Enterprise SSD enablement in 2014-2016 timeframe (using sub 1x nm MLC flash)
- Increased Industry focus on Enterprise MLC – higher endurance with extended lifespan support

© 2013 IBM Corporation

DRAM Capacitor Scaling Challenges

- DRAM scaling major challenges due to:
  - Reduction in cell capacitance, retention time, sensing margin
  - Lithography complexities – Low K1 ArF immersion extension, EUV lithography
  - Increase in cell parasitic resistance/capacitance and interconnect RC delay
- DRAM scaling rate vs. Equipment cost increases, process complexities
- Parametric degradation & increased bit error rate at sub 20nm node?

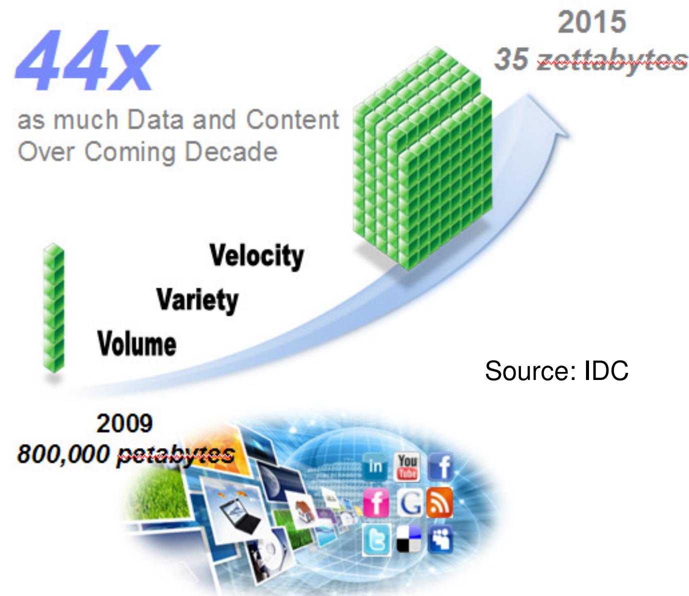
SW Park SK Hynix, Flash Memory Summit 2012

- DRAM scaling challenges and associated bit density increase within fixed power/thermal budget ?
- DRAM Quality & Reliability challenges ahead – driven by capacitor single bit refresh, disturb effects, contact integrity/controls & particle controls

© 2013 IBM Corporation

Data and Server Performance Growth Drive DRAM & NAND

Information is rapidly expanding and changing



Source: IDC

Exascale Systems Targets

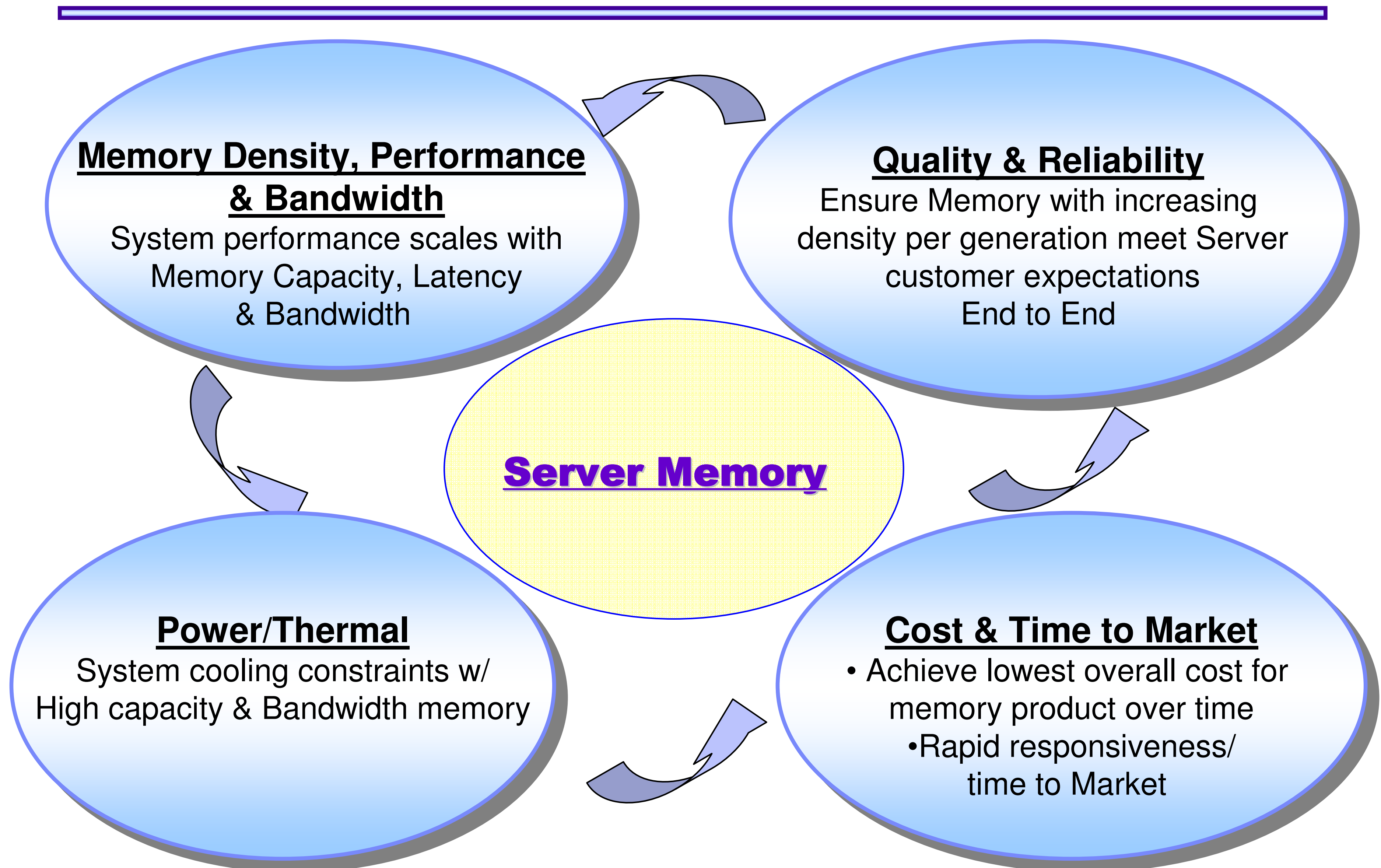
Systems	2009	2018	Difference Today & 2018
System peak	2 Pflop/s	1 Eflop/s	O(1000)
Power	6 MW	~20 MW (goal)	
System memory	0.3 PB	32 - 64 PB	O(100)
Node performance	125 GF	1.2 or 15TF	O(10) – O(100)
Node memory BW	25 GB/s	2 - 4TB/s	O(100)
Node concurrency	12	O(1k) or O(10k)	O(100) – O(1000)
Total Node Interconnect BW	3.5 GB/s	200-400GB/s (1:4 or 1:8 from memory BW)	O(100)
System size (nodes)	18,700	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	225,000	O(billion) + [O(10) to O(100) for latency hiding]	O(10,000)
Storage Capacity	15 PB	500-1000 PB (>10x system memory is min)	O(10) – O(100)
IO Rates	0.2 TB	60 TB/s	O(100)
MTTI	days	O(1 day)	- O(10)

http://www.exascale.org/mediawiki/images/d/db/PlanningForExascaleApps-Steven.pdf

- Server performance gains drive DRAM and NAND capacity requirements
- Storage performance (IOPS) motivates NAND integration
- Innovations in server architecture possible with Emerging Memory Technologies

© 2013 IBM Corporation

Server Memory Requirements



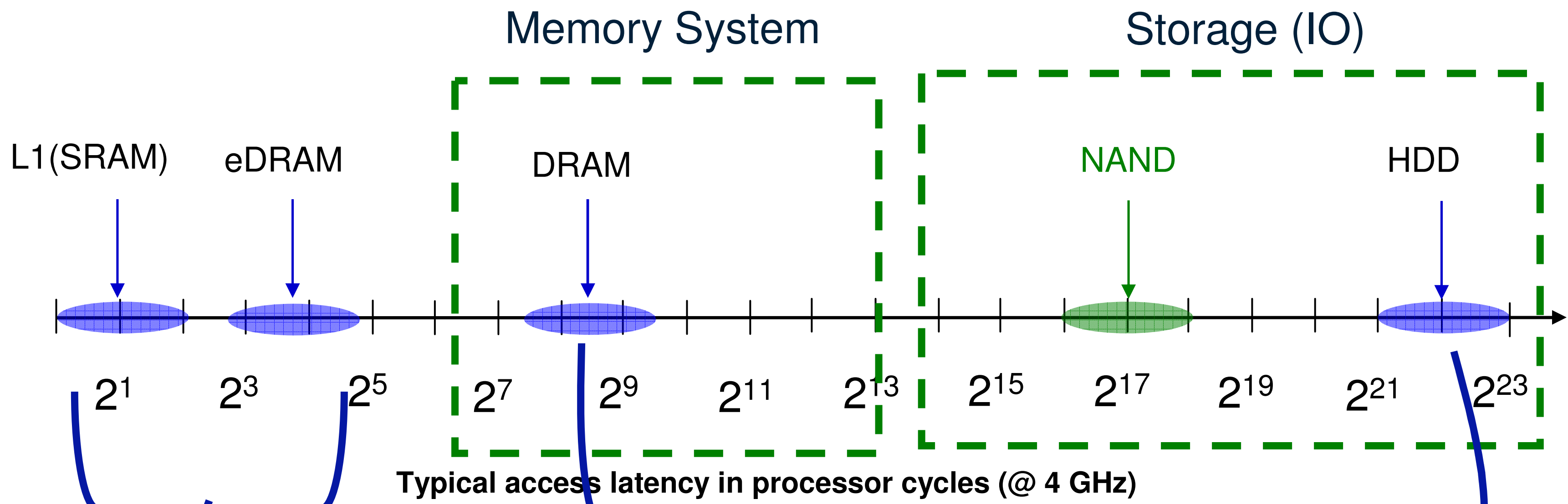
© 2013 IBM Corporation

Server Memory Key Considerations

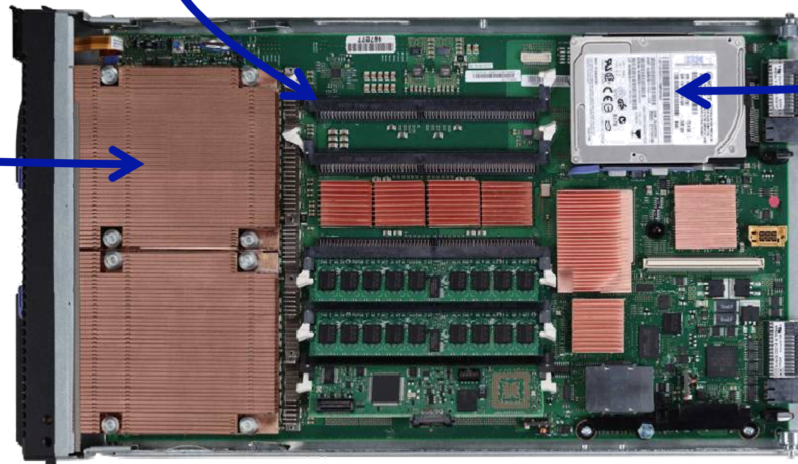
- **Server & mainframe memory**
  - Performance and reliability are important
  - Highly threaded multi-cores, virtualization, and datacenter consolidation all drive memory capacity & bandwidth
  - Datacenter and server-level power delivery and cooling constraints impose both nominal and peak power consumption requirements
  - Other domains, e.g. graphics, netbook, have somewhat different challenges & may have motivators for non-DRAM technologies at different times than servers
- **Storage systems**
  - Storage performance (IOPS) driving pervasive HDD → Flash storage transition
  - Capacity, latency and reliability are important for the enterprise.
  - Main focus – High density MLC-based flash storage solutions
- **Innovations in server architecture will enable benefits from Emerging Memory Technologies as they become mature**

© 2013 IBM Corporation

Within an Integrated Server: Where "Memory" is in a System

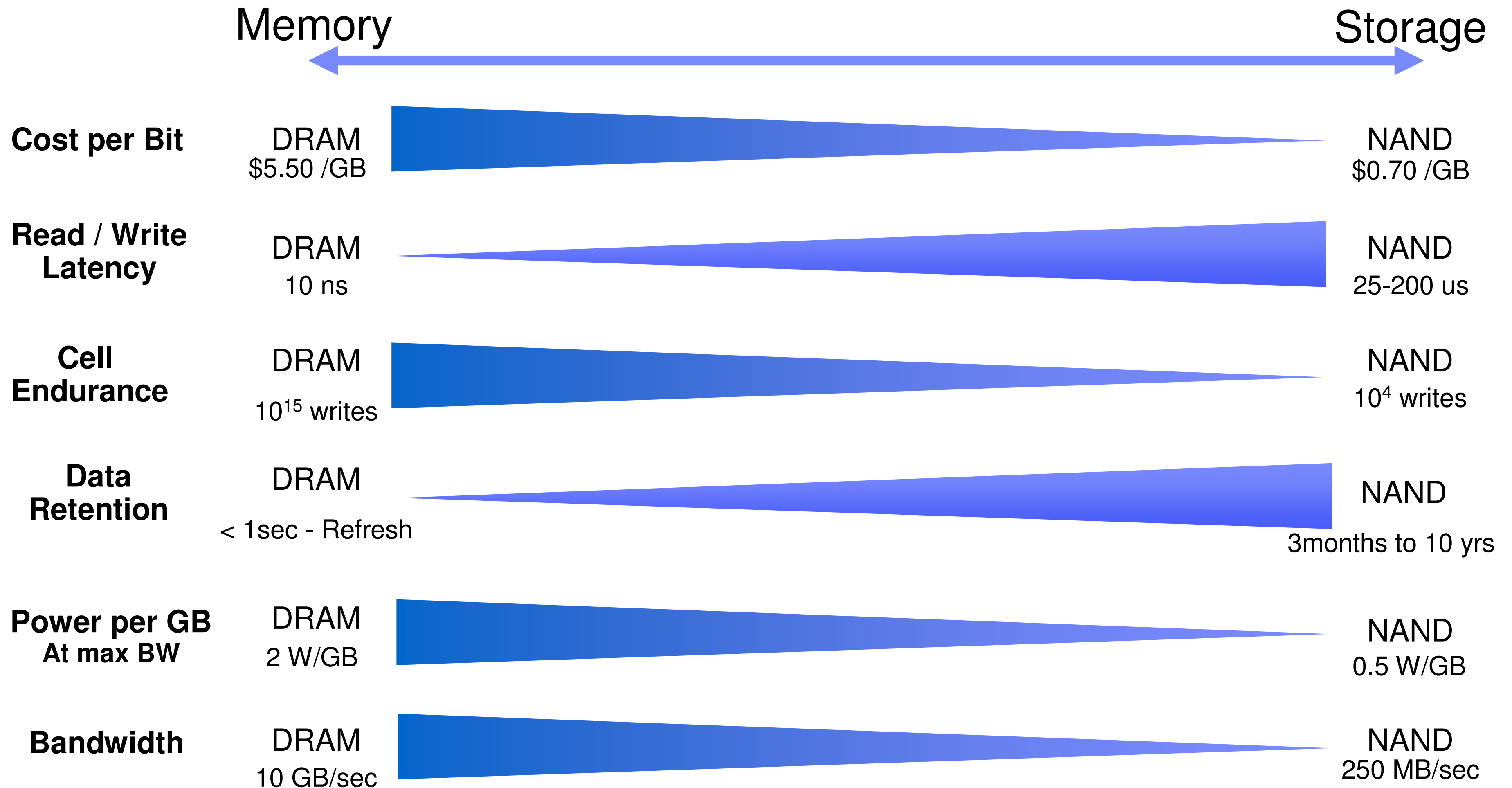


Typical access latency in processor cycles (@ 4 GHz)



© 2013 IBM Corporation

Memory Application Landscape

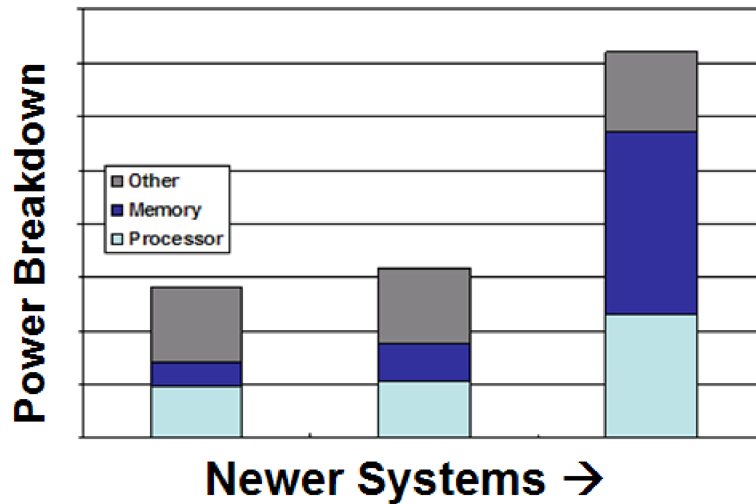


© 2013 IBM Corporation

Industry Power Trends

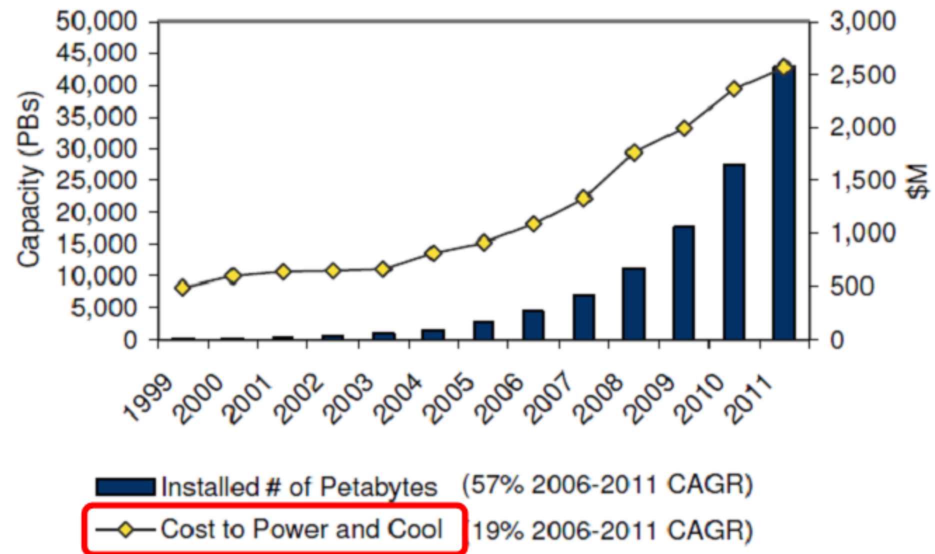
IBM

Server Power Trend



Newer Systems →

Storage System Power & Cooling Cost Trend



SNIA IDC June 2008 – 'The Real Costs to Power and Cool the world's external storage'

- More cores/system drives higher memory capacity & bandwidth
- More memory & bandwidth → more active & standby power
- Massive world-wide data growth → larger-capacity storage systems → more standby power
- Acceleration in rate of stored data analysis → growth in storage active power

 © 2013 IBM Corporation

Server Memory Hierarchy

- Revisiting the server memory hierarchy: need higher capacity, lower power → high density, cheaper, slower memory technologies fit between DRAM & HDD (NAND, PRAM, RRAM)

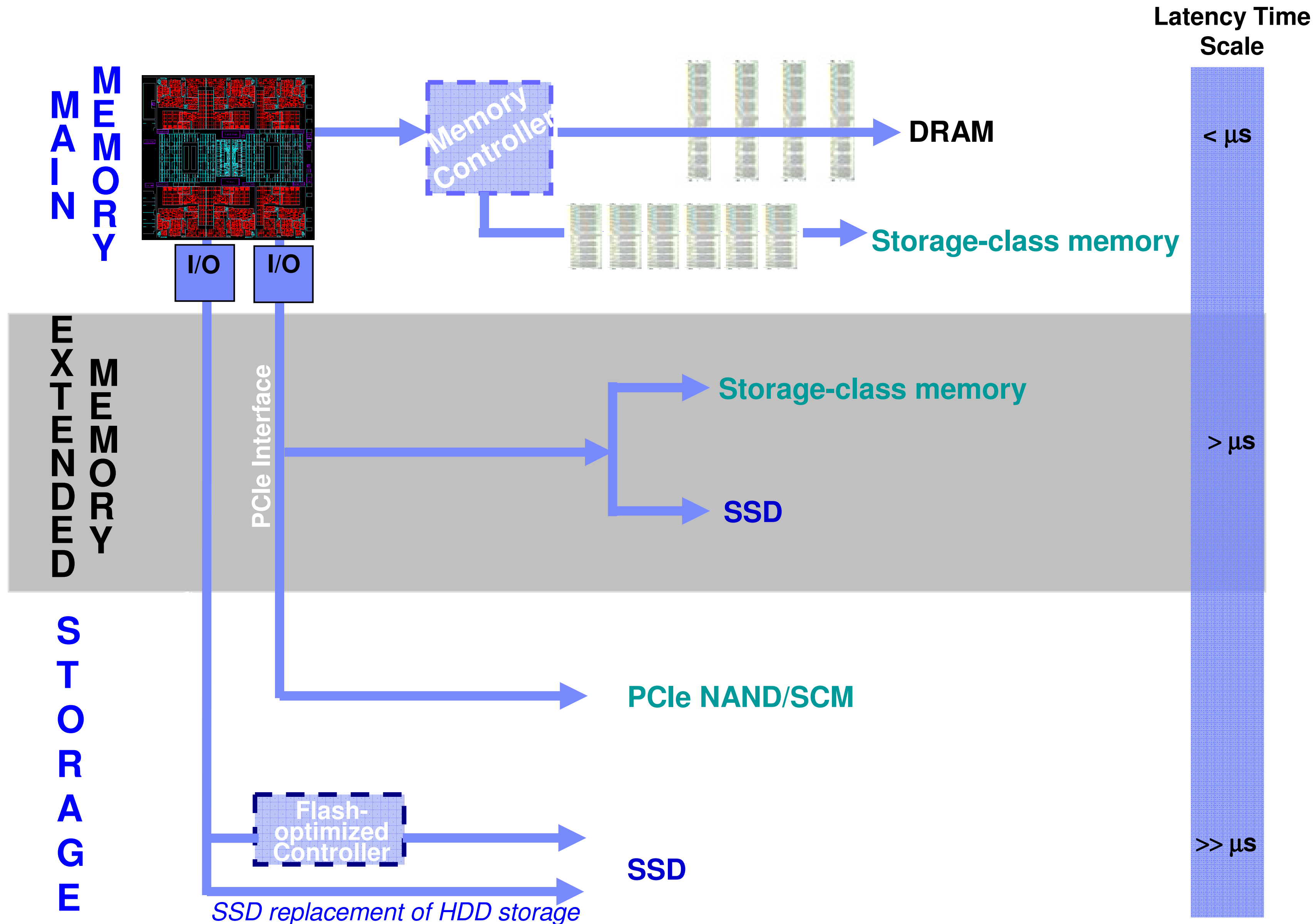
- Potential different roles in the memory/storage hierarchy

- STT-RAM: internal processor cache / main processor memory
- PCRAM: main processor memory / storage class memory
- ReRAM: storage class memory / storage

- Emerging memory technologies must satisfy Performance, reliability, power requirements within the memory/storage hierarchy

© 2013 IBM Corporation

Future architecture possibilities



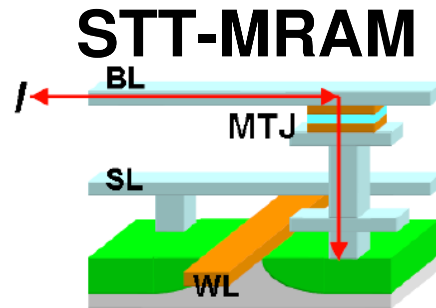
S. Chiras SOCC 2011

© 2013 IBM Corporation

[1] All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

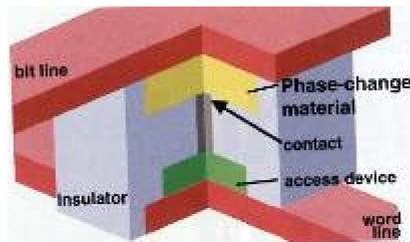
Emerging Memory Opportunities vs Challenges

STT-MRAM



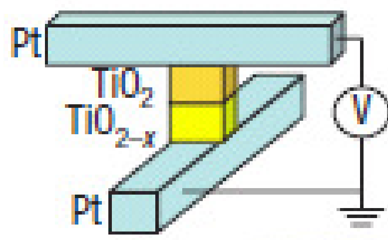
- High speed operation and non-volatility
- Main contender for DRAM replacement
- Eliminating DRAM refresh is a latency, bandwidth & power opportunity for STT-MRAM
- Complicated MTJ stacking structure, Yield challenge
- High temperature process & Low resistance ratio
- Margin Challenges, Soft errors
- 1x nm scaling and cost competitiveness??

PCM



- Most mature amongst emerging memory candidates – low density PCM in production for NOR replacement
- Drift challenges with high density PCM, Stuck Faults – reliability challenge
- Active Power, write current & latency – power/thermal challenges, too slow to work as main memory
- Scaling vs Thermal disturbance ??

ReRAM



- Very simple materials and structure
- Fast access, moderate endurance and low power
- Various and unclear switching mechanisms
- Large cell-to-cell variability
- EUV needed vs 3D NAND
- Stacking required for high density – manufacturing & yield challenges??

© 2013 IBM Corporation

Summary

- Flash & DRAM technology scaling becoming increasingly challenging
  - 2D Floating gate technology facing significant challenges below 15nm
  - 3D NAND will allow 'Effective Flash Scaling' into sub 10nm nodes
- Three emerging non-volatile "Resistive" RAM technologies
  - STT-RAM, PCRAM, and ReRAM
- They could play different roles in the memory/storage hierarchy
  - STT-RAM: internal processor cache / main processor memory
  - PCRAM: main processor memory / storage class memory
  - ReRAM: storage class memory / storage
- For successful emerging memory integration into servers, development must remain focused on cost, reliability, latency and power consumption – combined, these values enable a new era of enterprise computing

© 2013 IBM Corporation