Memory Architecture for Exa-scale Computing: Challenges and Solution Directions

SPARSH MITTAL

IIT HYDERABAD, INDIA

PARCOMPTECH, Bengaluru, 2017

Presentation Overview

- Scaling HPC to Exascale systems
- Challenges in achieving Exascale performance
- Solution directions
  - Non-volatile and heterogeneous memory systems
  - Die stacking
  - Approximate Computing
  - Near-threshold computing,
- Summary

Scaling HPC to Next-generation Systems

Parameter	2009	Exascale
Peak performance	2 PF	1 Exaflop
Power budget	6 MW	20MW
Memory	0.3 PB	>100 PB
Storage	15 PB	>500 PB
I/O Bandwidth	0.2 TB/s	30-60 TB/s
Reliability level	Nearly same	

Challenges:

- 50X performance at 3X power
- Much higher memory density and capacity
- Same reliability

Challenges of Exascale

- 20MW power is already enough for supporting city of lakhs of people
- 100 PB DDR3 DRAM memory => 52MW power
- Using hard disk to support checkpoint/restart can degrade performance of supercomputers by >50% => hardly any useful progress can be made

Challenges in designing memory-systems

- **“90% of supercomputer design is the memory system… and so is the other 10%”: Shekhar Borkar, Intel Fellow**

- Challenges are many
  - Perf., power, reliability, cost, area….

- We will focus on few opportunities and obstacles

Novel memory designs

Given the limitations of conventional memories, e.g., SRAM, (e)DRAM, researchers are exploring novel memories to address memory requirements of future systems.

A Comparison of Memory Technologies

	SRAM	DRAM	eDRAM	2D NAND Flash	3D NAND Flash	PCM	STT-RAM	2D ReRAM	3D ReRAM
Non-volatile	N	N	N	Y	Y	Y	Y	Y	Y
Cell size (F^2)	50-200	4-6	19-26	2-5	<1	4-10	8-40	4	<1
Minimum F shown (nm)	14	25	22	16	64	20	28	27	24
Read time (ns)	<1	30	5	10^4	10^4	10-50	3-10	10-50	10-50
Write time (ns)	<1	50	5	10^5	10^5	100-300	3-10	10-50	10-50
Read power	Low	Low	Low	High	High	Low	Medium	Medium	Medium
Write Power	Low	Low	Low	High	High	High	Medium	Medium	Medium
Write endurance	10^{16}	10^{16}	10^{16}	10^4-10^5	10^4-10^5	10^8-10^9	10^{15}	10^8-10^{12}	10^8-10^{12}
Power (other than R/W)	Leakage	Refresh	Refresh	None	None	None	None	Sneak	Sneak
Maturity									

Non-volatile memories

Optimum Very good Tolerable Challenge

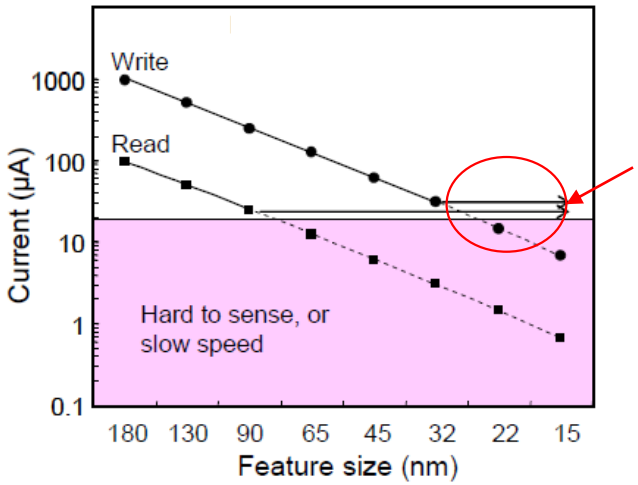
Benefits of non-volatile memories (NVM)

- Data retention
- Near-zero standby power
- Higher density than SRAM/DRAM
- MLC (multi-level cell) allows very high density
- Allow unifying main memory and storage

- However, they present several reliability challenges, as we will show now.

1. Read disturbance in STT-RAM

- With ongoing process scaling, read and write currents are becoming so close that reading a cell may inadvertently modify it

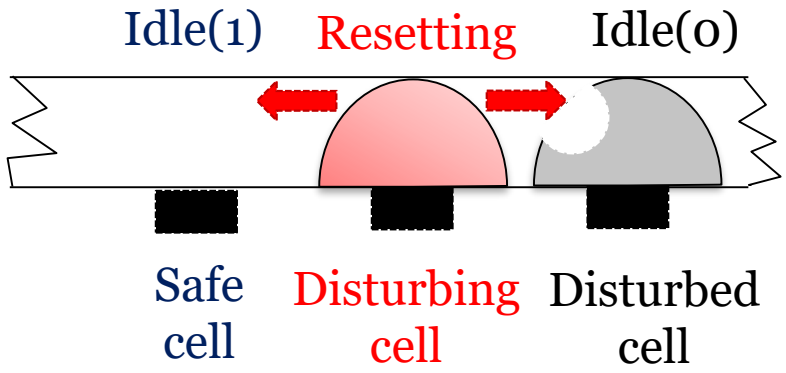


R/W currents becoming close

Takemura et al., IMW 2010.

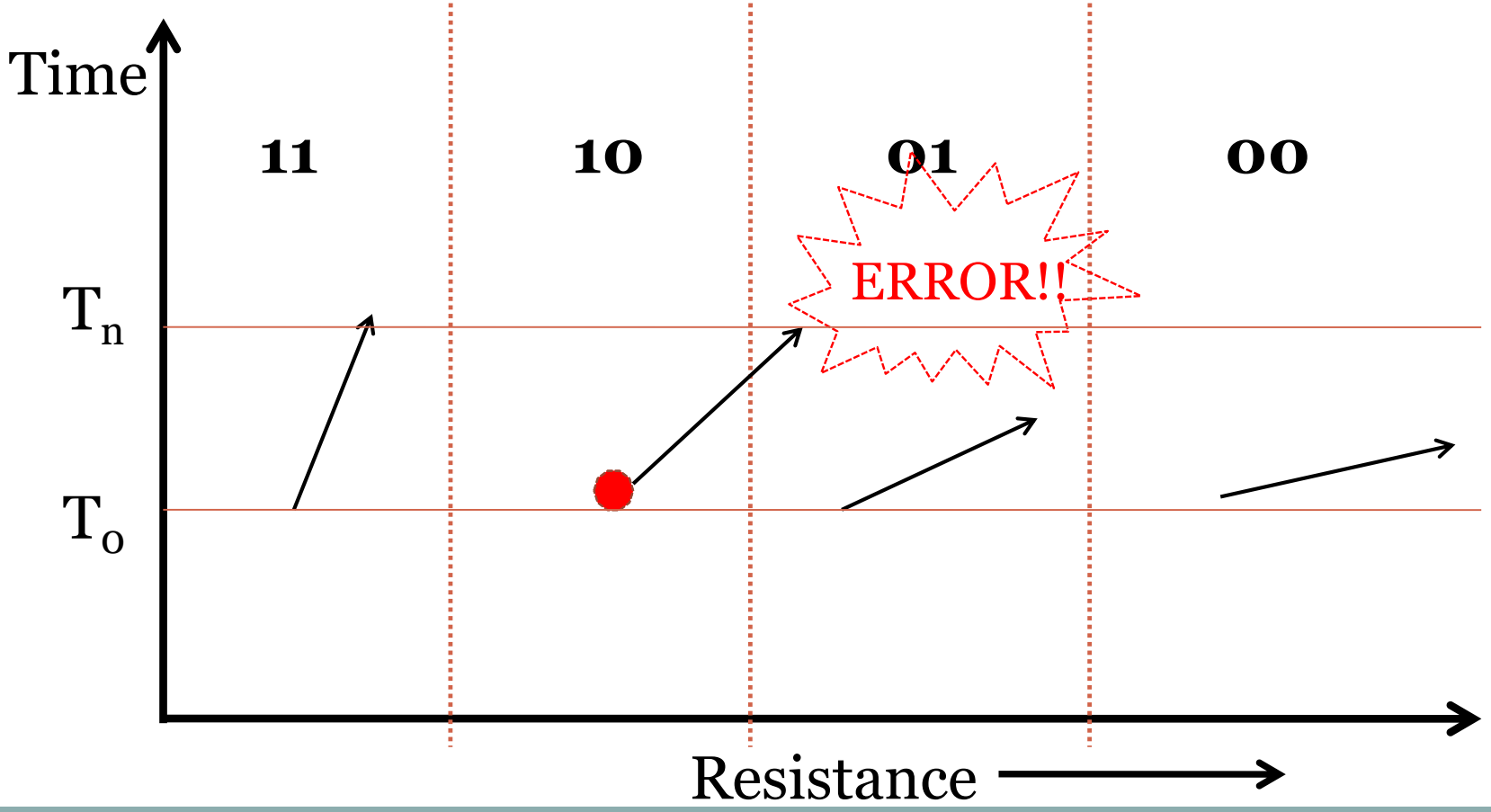
2. Write disturbance in PCM

- Heat generated in resetting a cell can corrupt neighboring cells storing ‘zero’ value.
  - Becoming severe with process scaling due to decreasing inter-cell distance



3. Resistance Drift in MLC PCM

- Drift can change the stored value over time

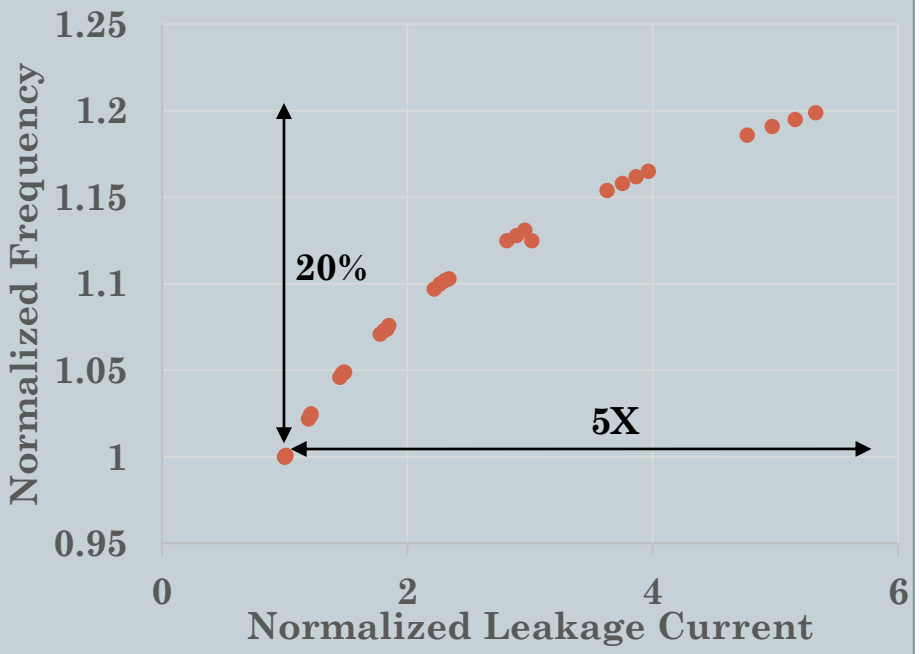


Awasthi et al., HPCA 2012.

4. Process variation: variation in parameters

M_0 f = 1.186 I_{leak} = 4.771	P_1 f = 1.191 I_{leak} = 4.974	P_0 f = 1.195 I_{leak} = 5.161	M_1 f = 1.199 I_{leak} = 5.329
P_{14} f = 1.154 I_{leak} = 3.628	M_{15} f = 1.158 I_{leak} = 3.753	M_{14} f = 1.162 I_{leak} = 3.866	P_{15} f = 1.165 I_{leak} = 3.964
P_2 f = 1.125 I_{leak} = 2.812	M_3 f = 1.128 I_{leak} = 2.890	M_2 f = 1.131 I_{leak} = 2.958	P_3 f = 1.125 I_{leak} = 3.015
M_{12} f = 1.097 I_{leak} = 2.219	P_{13} f = 1.102 I_{leak} = 2.308	P_{12} f = 1.100 I_{leak} = 2.268	M_{13} f = 1.103 I_{leak} = 2.341
M_4 f = 1.071 I_{leak} = 1.780	P_5 f = 1.073 I_{leak} = 1.810	P_4 f = 1.074 I_{leak} = 1.834	M_5 f = 1.076 I_{leak} = 1.851
P_{10} f = 1.046 I_{leak} = 1.450	M_{11} f = 1.048 I_{leak} = 1.468	M_{10} f = 1.049 I_{leak} = 1.481	P_{11} f = 1.049 I_{leak} = 1.490
P_6 f = 1.022 I_{leak} = 1.196	M_7 f = 1.024 I_{leak} = 1.207	M_6 f = 1.024 I_{leak} = 1.214	P_7 f = 1.025 I_{leak} = 1.217
M_8 f = 1.000 I_{leak} = 1.000	P_9 f = 1.000 I_{leak} = 1.006	P_8 f = 1.001 I_{leak} = 1.009	M_9 f = 1.001 I_{leak} = 1.009

Values normalized to this tile



PV can cause large variation in performance and power profile

Herbert et al., TVLSI, 2009

Examples of process variation

- 15% difference in energy between nodes of the Eurora supercomputer
- Maximum clock frequency of different cores in an 80-core Intel processor vary between 5.7 to 7.3 GHz
- 9X variation in sleep power in different instances of ARM Cortex M3 processors.
- 50X variation in write endurance of cells of phase change memory
- Timing parameters in a DDR3 DRAM can be 66% lower than the datasheet specifications

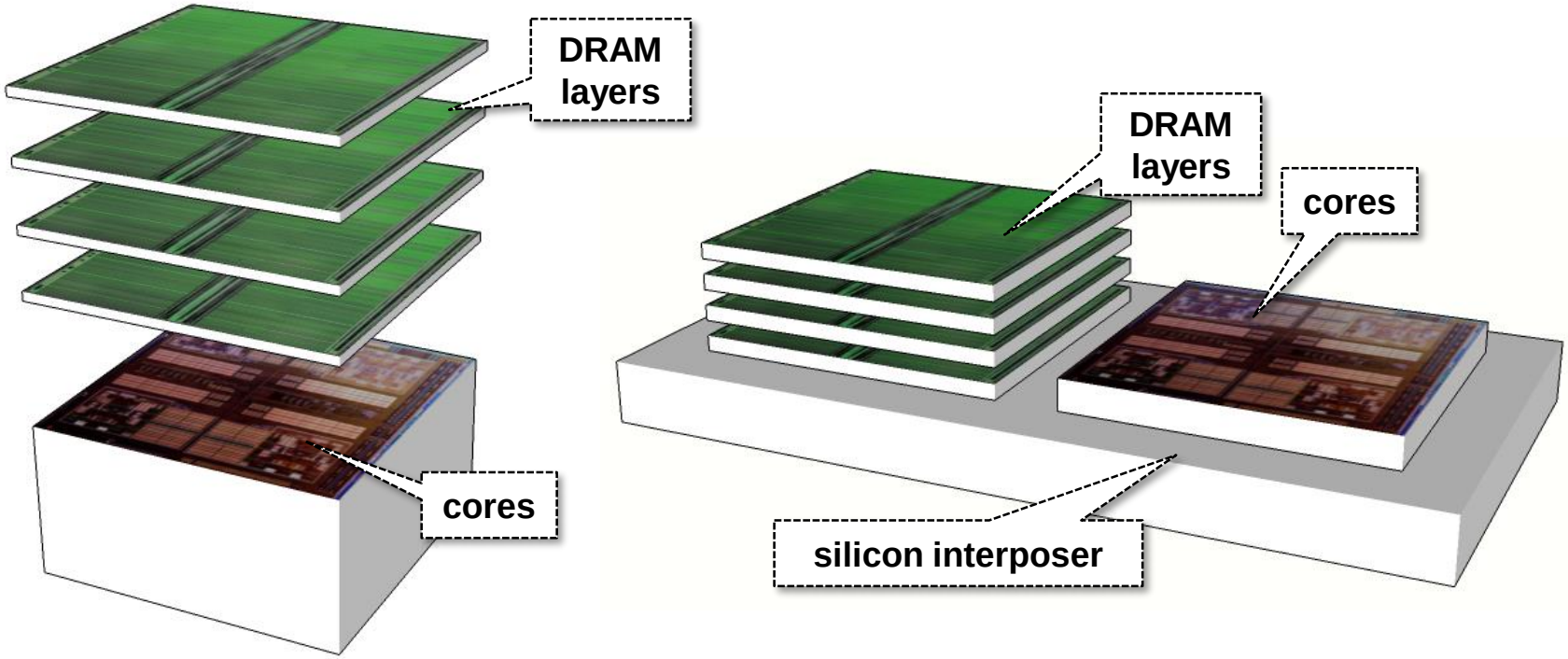
Heterogeneous memory systems

- Allows bringing best of multiple memories together
  - DRAM-PCM main memory
  - Flash-disk storage
  - DRAM-Flash main memory, etc.
- However, heterogeneous memory also has challenges
  - Fabrication and mass-production is challenging
  - Increased complexity
  - Cost challenges

Die-stacking

We will take example of stacked-DRAM. These ideas may apply to other memories also.

Die-stacked DRAM



3D stacking

2.5D stacking

Interconnects logic and DRAM dies using TSVs (3D) and silicon interposers (2.5D)

Benefits of Stacked DRAM

- Higher integration density
- 8-15X bandwidth than off-chip DRAM,1/2-1/3 latency
- Reduces cache area => place higher number of cores
- Adopted by major processor and DRAM vendors
  - Samsung, micron, SK Hynix, IBM, Intel, AMD, NVIDIA
- Already integrated in commercial CPUs/GPUs

Challenges of Stacked DRAM

- Memory with very large number of layers => challenges of power delivery and cooling, yield, testability
- 32ms refresh period required instead of 64ms
- Faults in die-stacked memory may not be easily serviceable => may require discarding the entire package including the functioning processor layer
- Large cache/memory also has large metadata

Approximate Computing (AC)

Motivation and Potential

- Exact computation or peak-level service demands => require high amount of resources
- Selective approximation or occasional violation of specification => disproportionate gains in efficiency

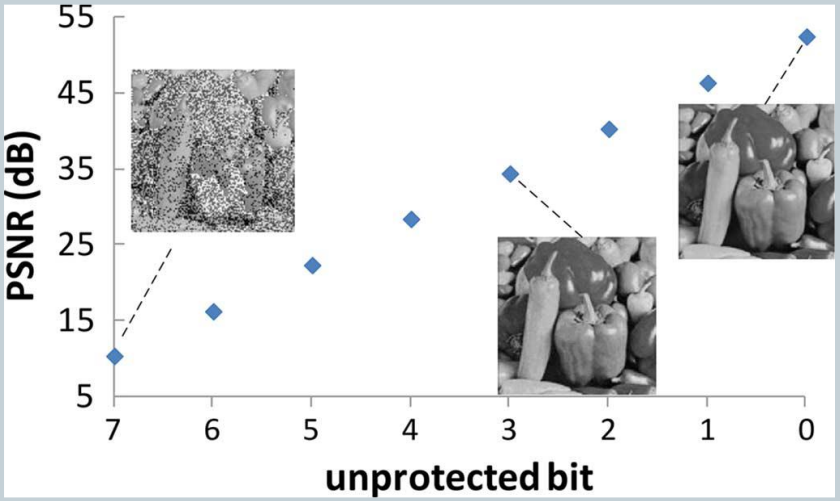
- Exascale systems don't have to be fully accurate!
- Approximate systems are good-enough and acceptable
- Allows using unreliable/inaccurate components

Opportunity: Error-resilient apps and users

- Perceptual limitations of humans
  - PSNR >30db considered acceptable
- Non-critical portions in apps
  - 98% FP operations are approximable in 3D raytracer
  - In clustering, only relative distances matter, not absolute distances
  - In document analysis, processing of common words, such as “a”, “of”, “the” can be avoided

Example results from AC

- 50X energy saving for 5% accuracy loss in k-means clustering
- PSNR loss due to errors in a single bit at various positions



Some approximation strategies

1. Precision scaling
2. Loop-perforation
3. Load-value approximation
4. Memoization
5. Task dropping
6. Memory access skipping
7. Data Sampling
8. Program versions of different accuracy
9. Inexact hardware
10. Voltage scaling
11. (e)DRAM refresh rate reduction
12. Inexact reads/writes
13. Reducing divergence in GPU
14. Lossy compression
15. Neural network

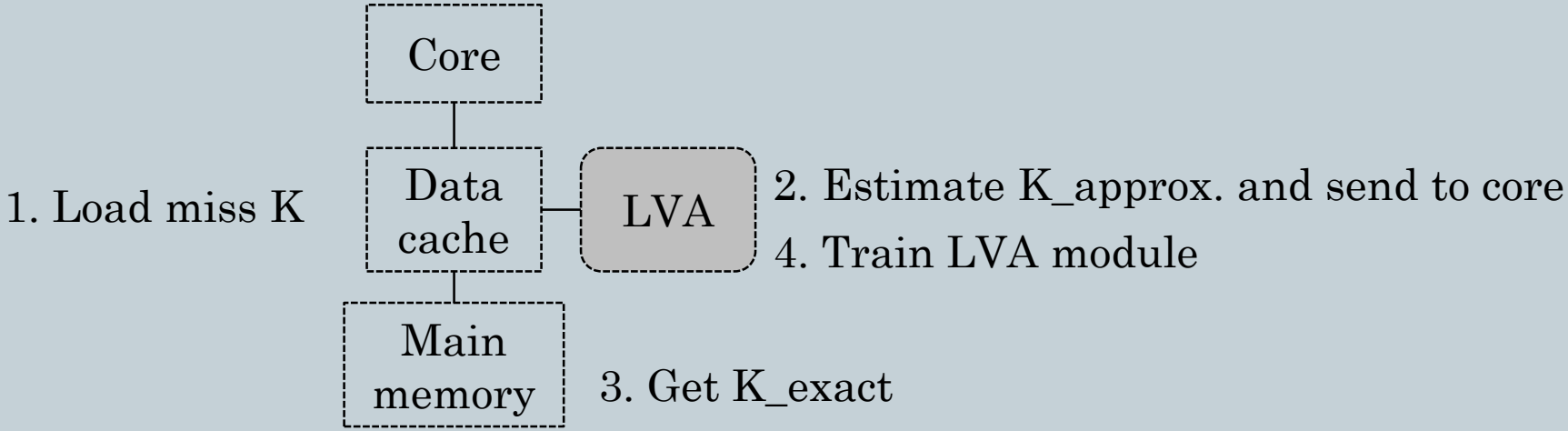
Some memory-related AC strategies

- **Memory access skipping**: skip some memory accesses or fetch only MSBs

- **Memoization:** Reuse previously stored results for *similar* (but not *identical*) functions/inputs
  - Reuse result of an instruction across different parallel lanes of SIMD architecture
  - Used in scatter/gather and map computations

SIMD = single instruction multiple data

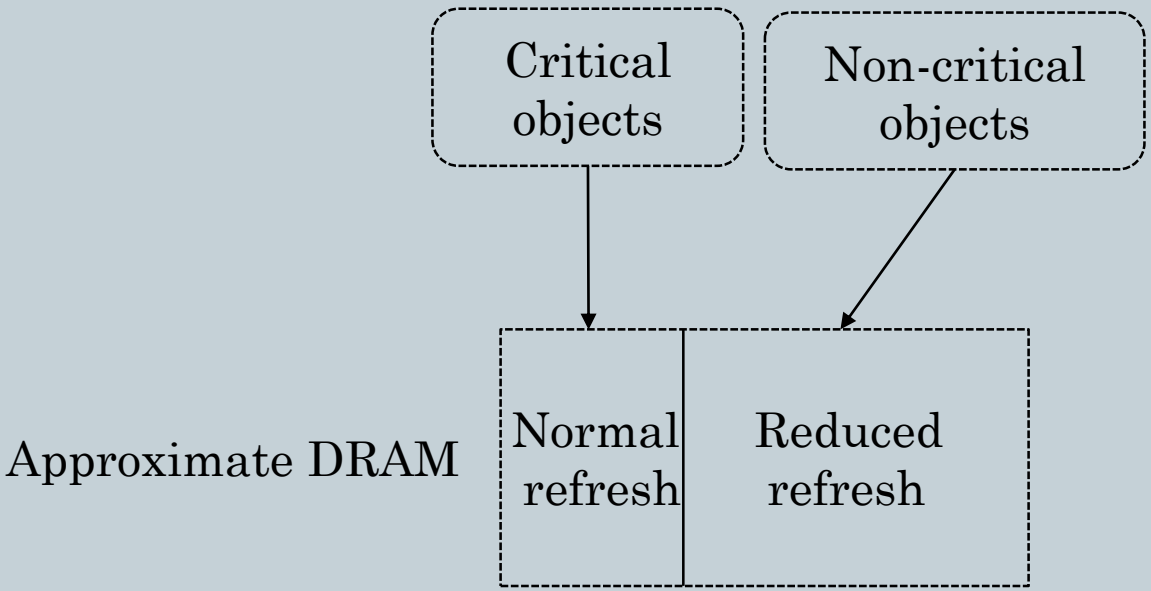
Load-value approximation (LVA)



- Reduces memory accesses and hides latency
- Can offset both latency and BW constraints

DRAM refresh rate reduction

- Lower refresh rate in portion of DRAM to save energy. Store non-critical data in this portion



Voltage scaling (in SRAM)

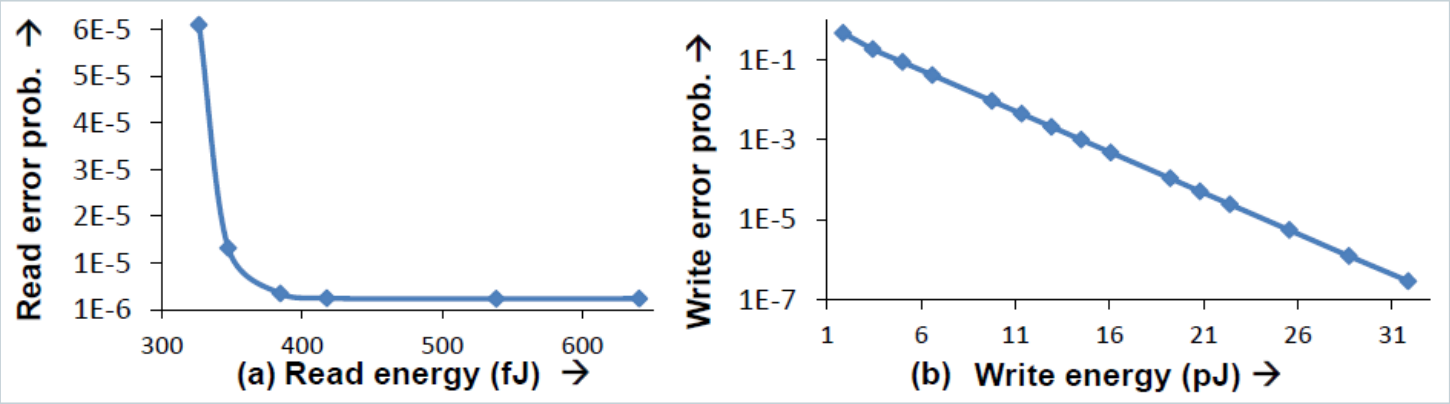
- Reducing SRAM supply voltage
  - saves leakage energy but
  - increases probability of read disturbance and incorrect write
- Store
  - critical data (e.g., MSB) in fault-free blocks
  - non-critical data in faulty blocks.

MSB = most significant bit

Inexact reads/writes in NVMs

1. Low read-current => erroneous reads
2. Lowering either write duration or write current or both => unsuccessful writes

- Use these knobs during reads/writes to control quality



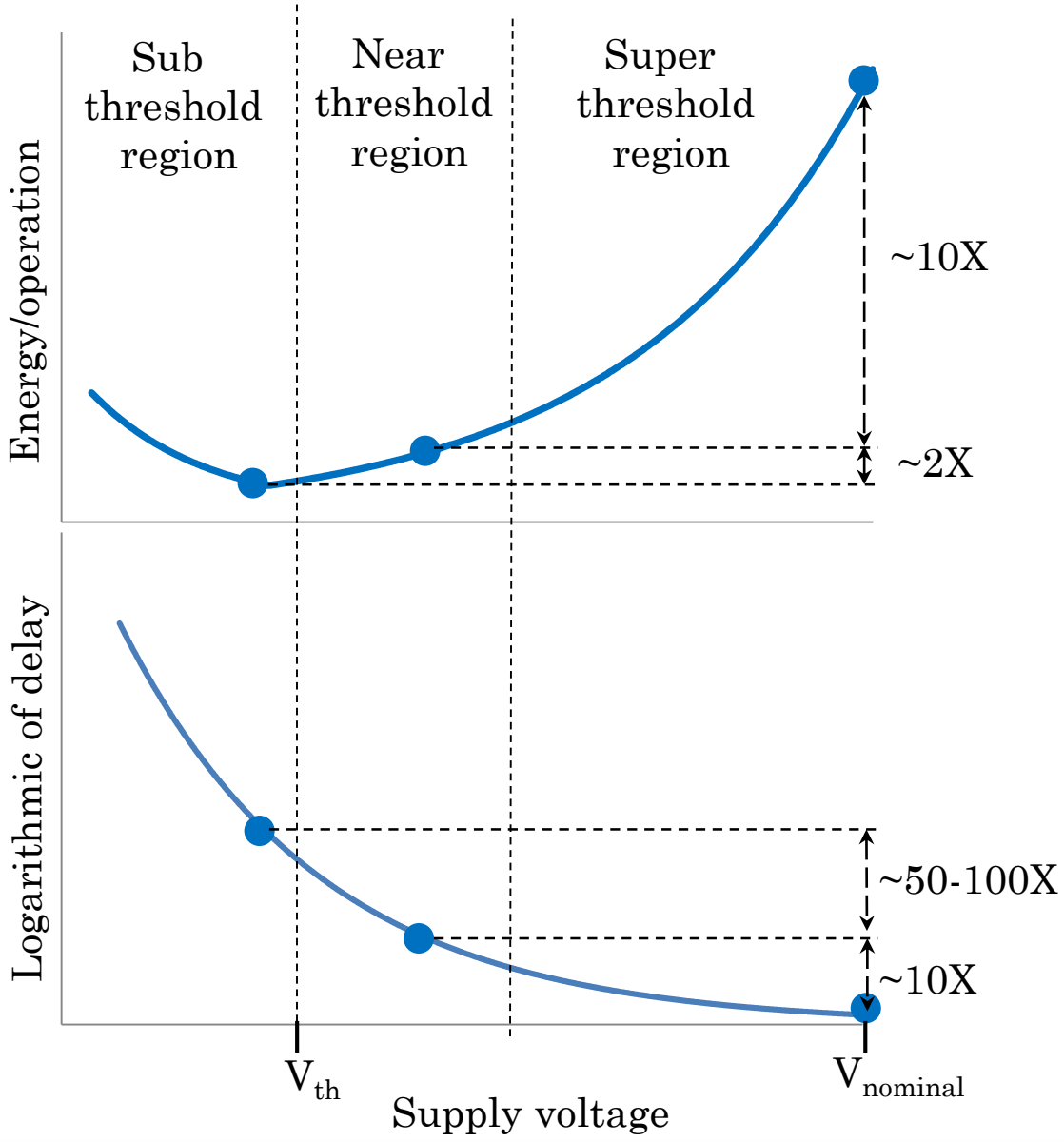
Energy v/s error probability for STT-RAM bit-cell

Near-threshold voltage computing (NTC)

Motivation for NTC

- Peak demands are high => over-provisioning of resources
- Average utilization is low. Large periods of low computational needs
- **Idea:** transition to low-voltage for saving energy with minor loss in performance

Energy/delay variation with voltage



- Super Threshold
  - high performance
  - high energy consumption
- Near Threshold
  - 10x energy reduction
  - 10x performance degradation
- Sub Threshold
  - exponentially decreasing performance
  - increasing leakage

Properties of near-threshold voltage region

- Standard DVFS reduces supply voltage to no lower than 70% of nominal voltage
- Near-threshold operation scales voltage to ~25-35% of nominal level
  - close to V_{th} but nearly 200-400mV above V_{th}
- Near-threshold computing allows trading-off memory capacity for perf./energy/reliability

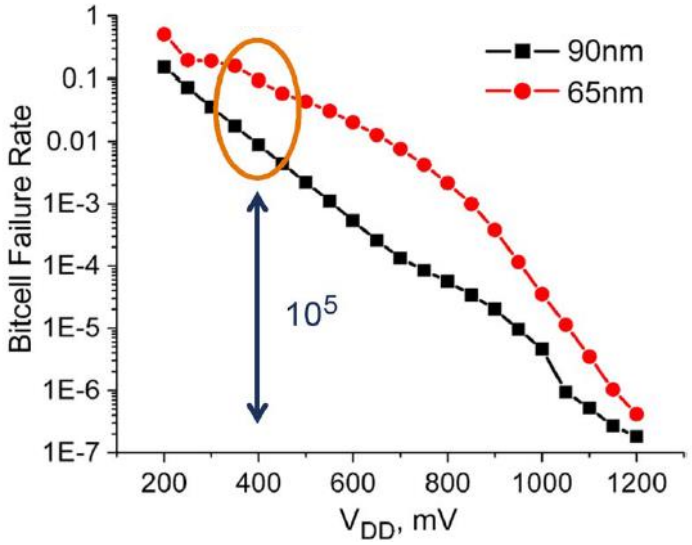
DVFS = dynamic voltage/frequency scaling

NTC challenge: performance

- Challenge: NTC increases delay by 10x
- Solution: Use parallelism. If computation requires N operations, break into N/10 parallel subtasks
- Benefit: Execution time restored, total energy is still 10X less, power 100X less
- Suitable for a wide range of applications

NTC Challenge: Reliability

- Challenge: Failure rate increases rapidly at low voltages



Impact of voltage scaling on SRAM failure rate

- Solution: Disable faulty blocks or use stronger error-correction

Dreslinski et al., Proc. IEEE, 2010.

Conclusion

- Power/perf./reliability targets of next-generation systems mandate magnitude-order improvements
- Memory architecture design: (one of) the most crucial challenge for Exascale
- Many approaches may need to be combined
  - Novel memories, die stacking, approximate computing, near-threshold computing,…