

Respin: Rethinking Near-Threshold Multiprocessor Design with Non-Volatile Memory

Xiang Pan, Anys Bacha, and Radu Teodorescu
Computer Architecture Research Lab

<http://arch.cse.ohio-state.edu>



THE OHIO STATE UNIVERSITY



Universal Demand for Low Power



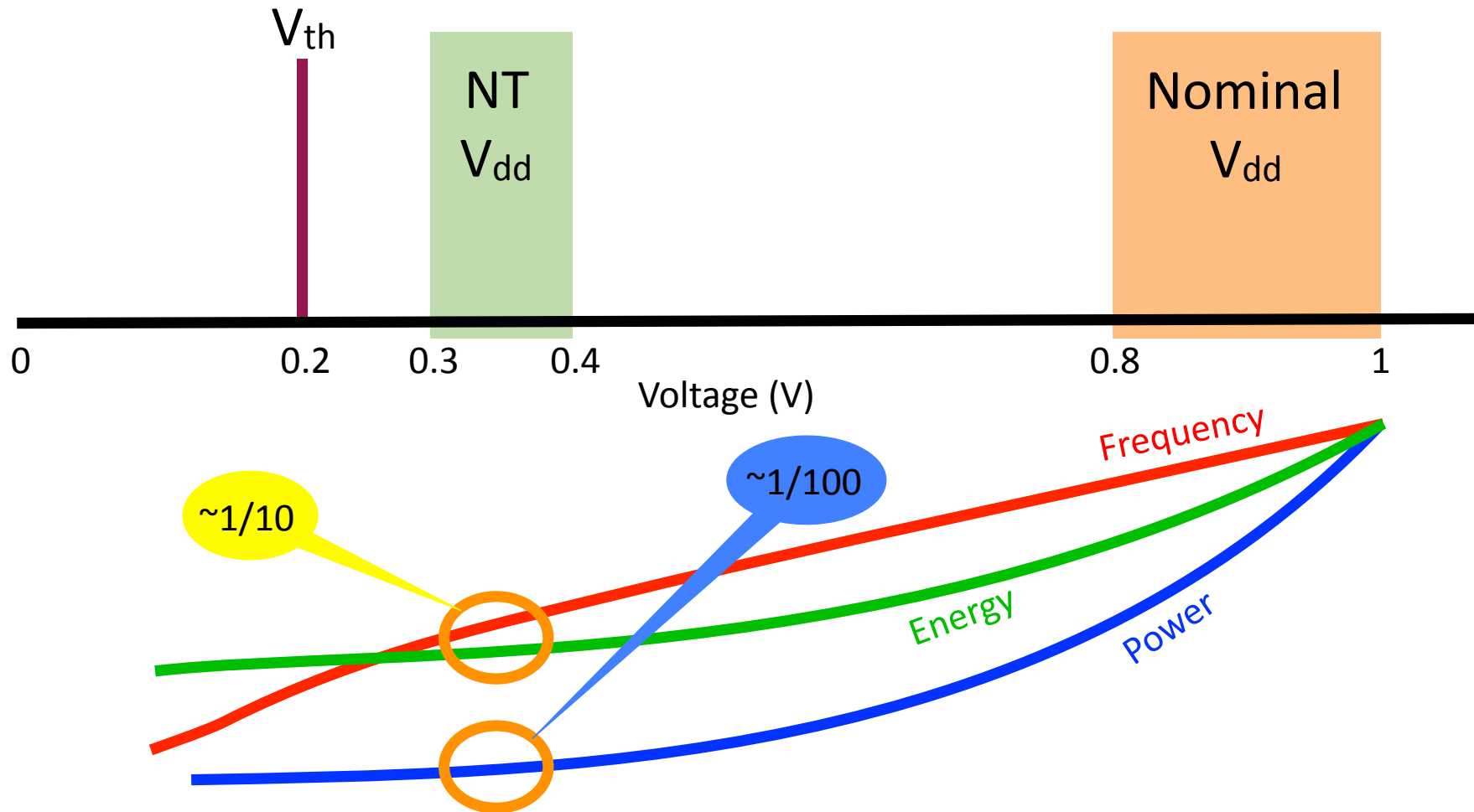
- Mobility
- Battery life

- Performance
- Power constraints

- Energy cost
- Environment



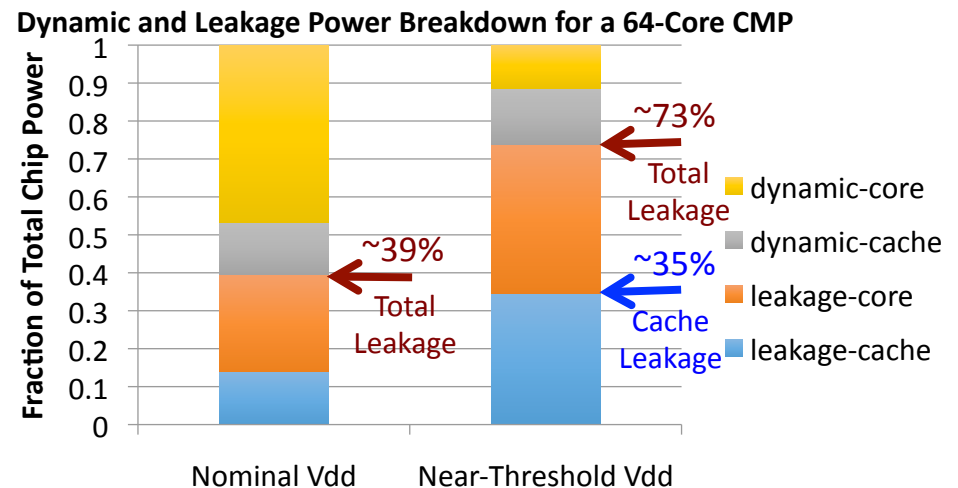
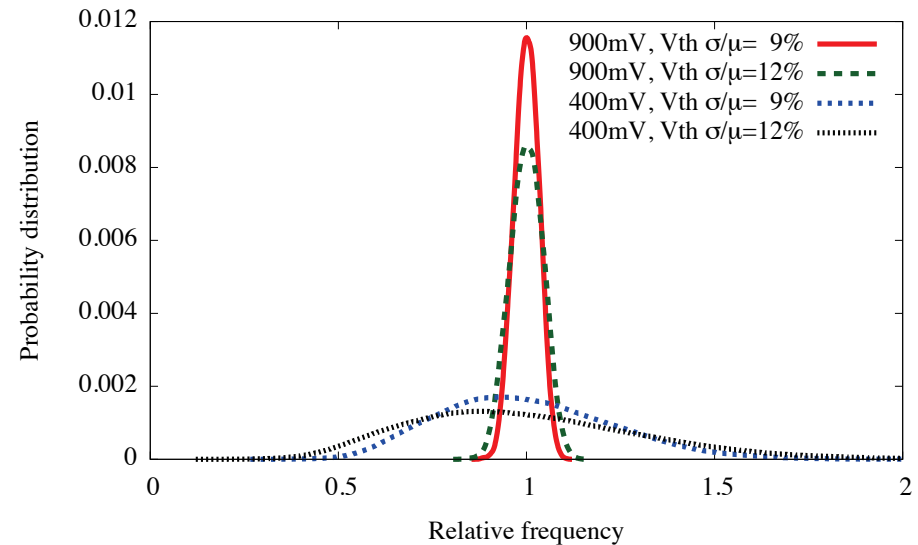
Near-Threshold Operation





Challenges in Near-Threshold

- Performance degradation
- Amplified process variation
- Leakage power dominates
- **The initial idea of Respin** –
Build caches in NT-CMP
with “leakage-free” non-
volatile memories to
reduce power consumption





Outline

- **Basics of NVM**
- Respin Architecture
- Methodology and Evaluation
- Conclusion



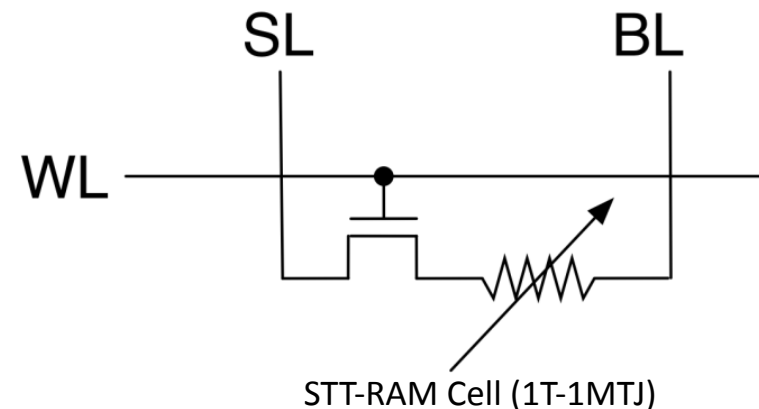
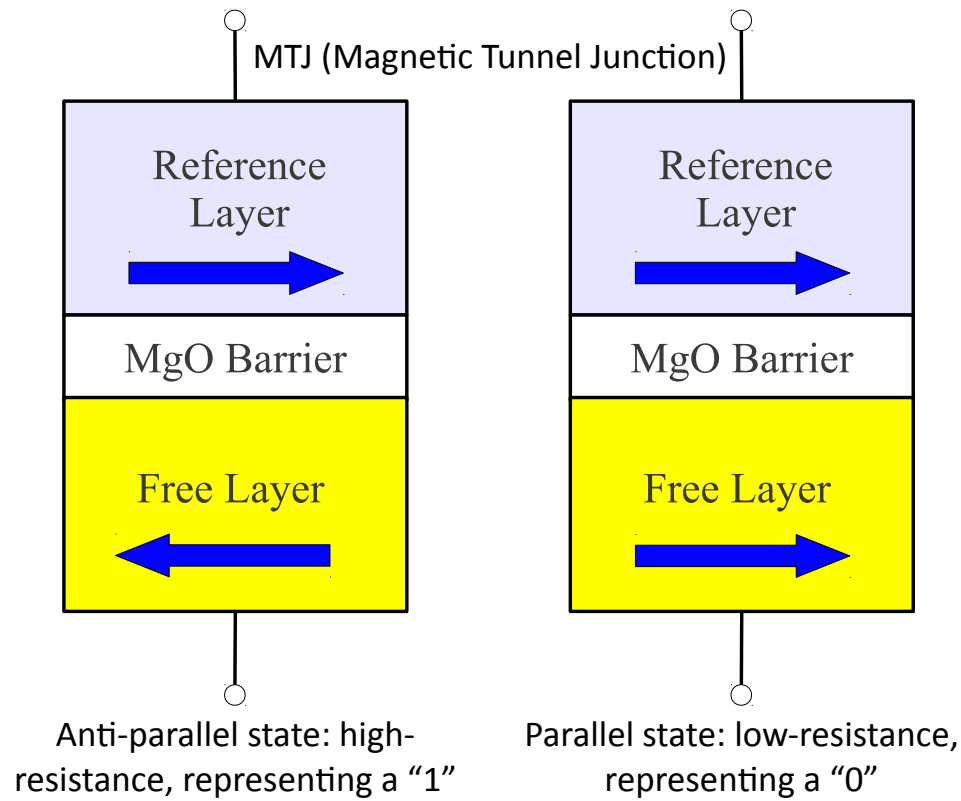
Non-Volatile Memory Basics

- Non-Volatility – Resistance as data representation (e.g. PCRAM, STT-RAM, ReRAM, etc.)
- Near-Zero Leakage Power – Good fit for future power-constrained computing
- High Density – Great design candidate in the big data era
- Good Performance – Feasible for on-chip storage replacement



STT-RAM

- Unique features of STT-RAM: fast read speed, low read energy, unlimited write endurance, and good compatibility with CMOS technology
 - Shortcomings: long write latency and high write energy











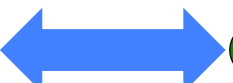





STT-RAM is Good Fit for NT-CMP

NT-CMP

STT-RAM

High leakage power				Near-zero leakage power
Low operating frequency				Long latency writes
Large numbers of cores requiring high cache capacity				High density (~4x denser than SRAM)
Unreliable functional units				Robust from soft-errors

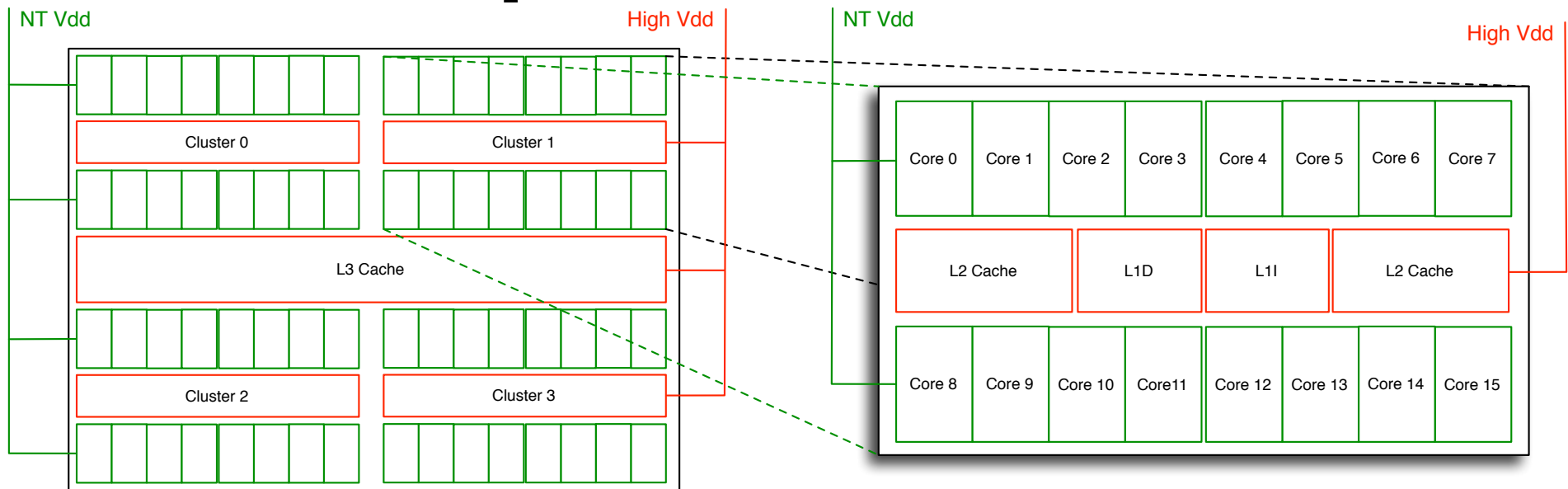


Outline

- Basics of NVM
- **Respin Architecture**
- Methodology and Evaluation
- Conclusion



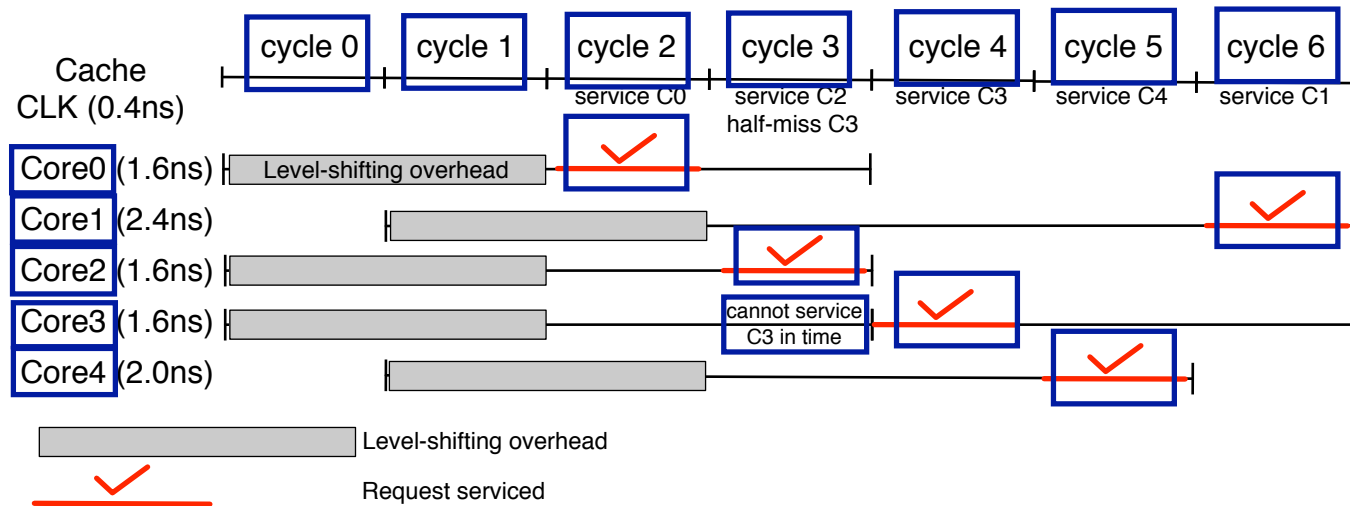
Respin Architecture



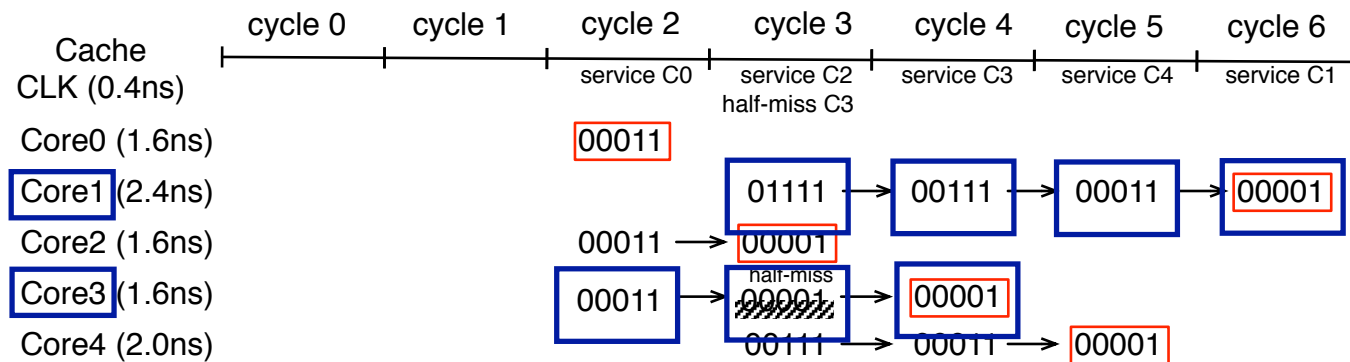
- Cores operate at NT-Vdd rail with low frequencies
- Caches are built with STT-RAM and operate at high-Vdd rail making read speed extremely fast
- Clustered-CMP with fast STT-RAM read enables within-cluster shared L1 cache design, removing coherence costs



Shared Cache Controller



(a)

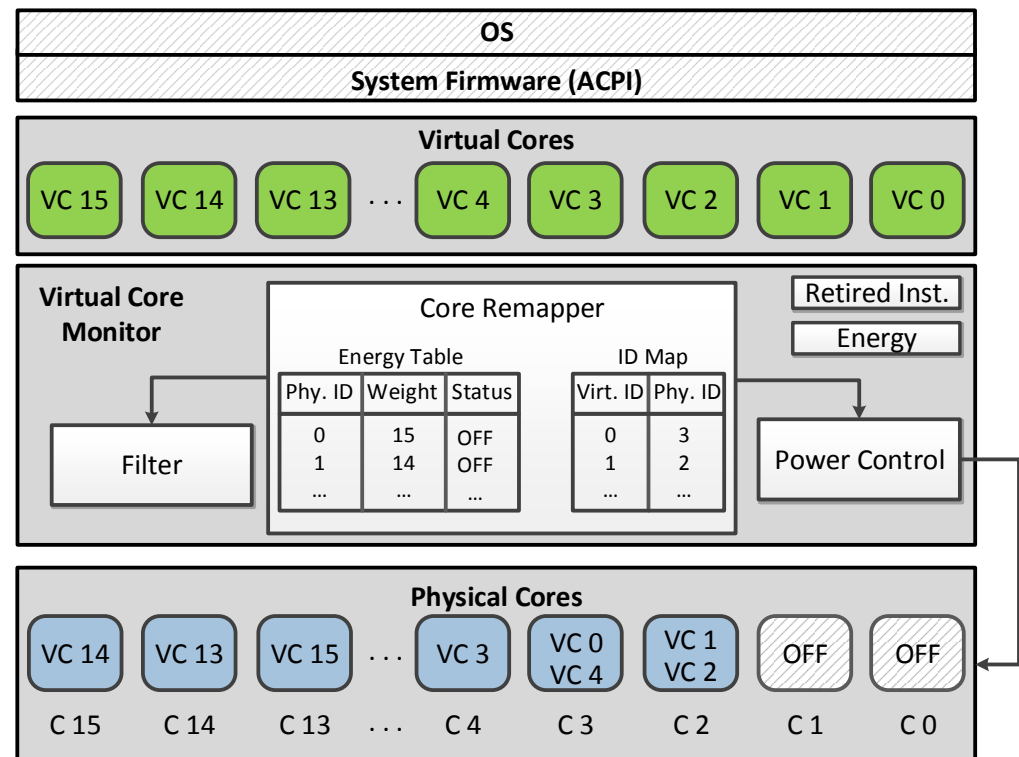


(b)



Dynamic Core Consolidation

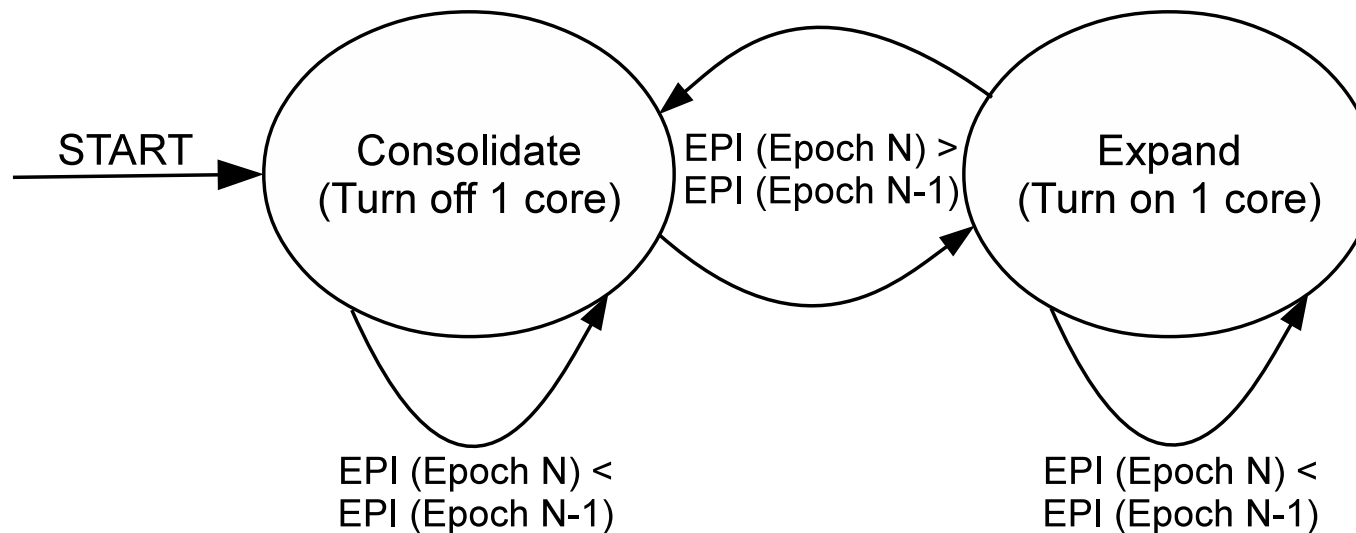
- High process variation and leakage in NT-CMP lead to fast cores more energy-efficient than slow ones
- Dynamically consolidate threads onto more efficient cores with greedy search at runtime can further save energy
- Mechanism implemented in system firmware, energy-per-instruction used as greedy selection metric, and instruction count used as evaluation interval





Greedy Selection

- A greedy approach is used to search for optimal system configurations at application runtime
 - Energy-Per-Instruction (EPI) as evaluation metric
 - Instruction count as evaluation interval





Outline

- Basics of NVM
- Respin Architecture
- **Methodology and Evaluation**
- Conclusion



Methodology

Level	Size (mm ²)	Block Size	Associativity	Read/Write Ports
L1I (Private/Shared within Cluster)	16KB (Private)/256KB (Shared within Cluster)	32B	2-way	1/1
L1D (Private/Shared within Cluster)			4-way	
L2 (Shared within Cluster)	8MB (Small)/16MB (Medium)/32MB (Large)	64B	8-way	
L3 (Shared within Chip)	24MB (Small)/48MB (Medium)/96MB (Large)	128B	16-way	

Table 1. Summary of Cache Parameters.

	Vdd Rail	Area (mm ²)	Read/Write Latency (ns)	Read/Write Energy (pJ)	Leakage Power (mW)
SRAM (16KB × 16)	Low (0.65V)	0.9176	1.337	2.578	573
SRAM (256KB)			0.5336	42.41	881
STT-RAM (256KB)	High (1.0V)	0.2451	0.3774/5.208	29.32/209.3	114

Table 2. Comparison of SRAM vs. STT-RAM Technology Parameters.



Methodology

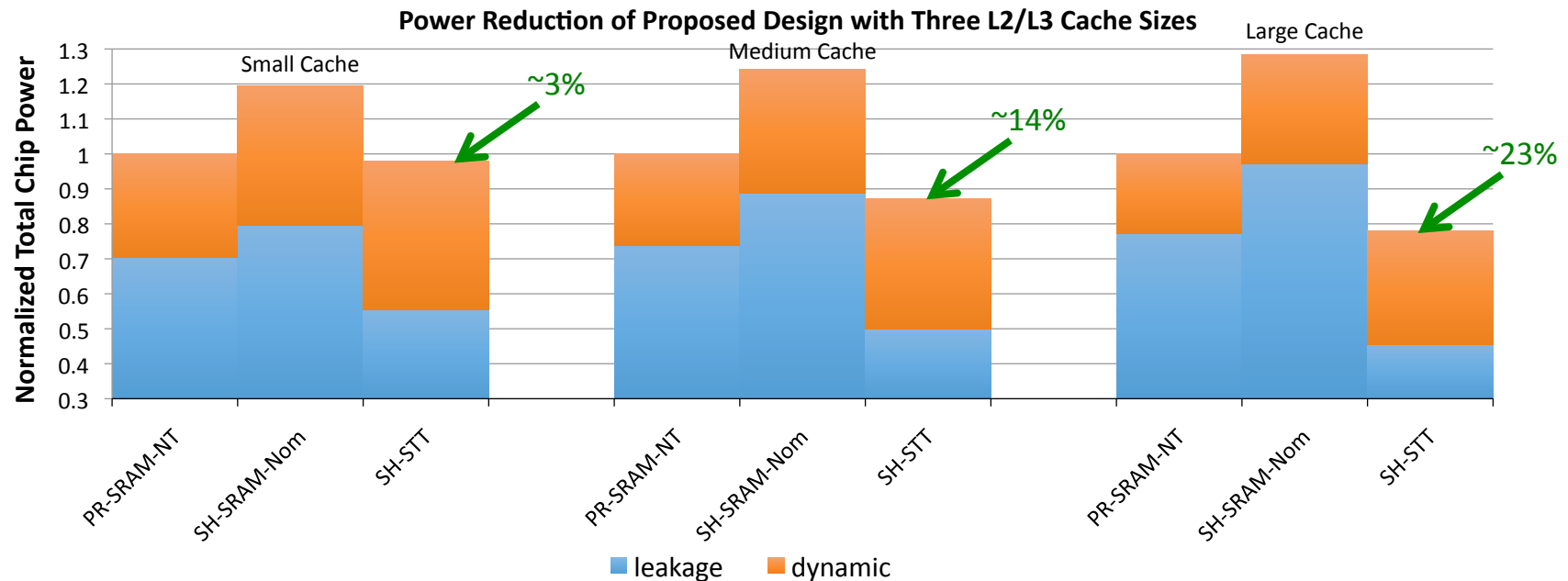
- Simulation Framework:
 - SESC for architectural simulation
 - CACTI, McPAT, and NVSim for latency, power, energy, and area simulation
- Benchmarks:
 - SPLASH2 and PARSEC
- Main Evaluated Configuration:
 - 64-core CMP with four 16-core clusters
 - Medium size L2 and L3 caches
 - 0.4ns shared L1 cache read latency

CMP Architecture	
Cores	64 out-of-order
Fetch/Issue/Commit Width	2/2/2
Register File Size	76 int, 56 fp
Instruction Window Size	56 int, 24 fp
Reorder Buffer Size	80 entries
Load/Store Queue Size	38 entries
NoC Interconnect	2D Torus
Coherence Protocol	MESI
Consistency Model	Release Consistency
Technology	22nm
NT-Vdd	0.4V (Core), 0.65V (Cache)
Nominal-Vdd	1.0V
Core Frequency Range	375MHz – 725MHz
Median Core Frequency	500MHz
Variation Parameters	
Vth std. dev./mean (σ/μ)	12% (Chip), 10% (Cluster)

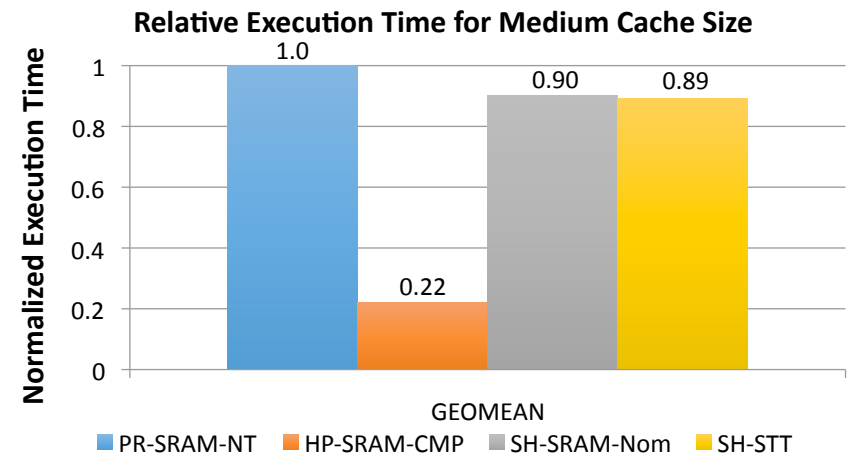
Table 3. Summary of Experimental Parameters.



Power and Performance



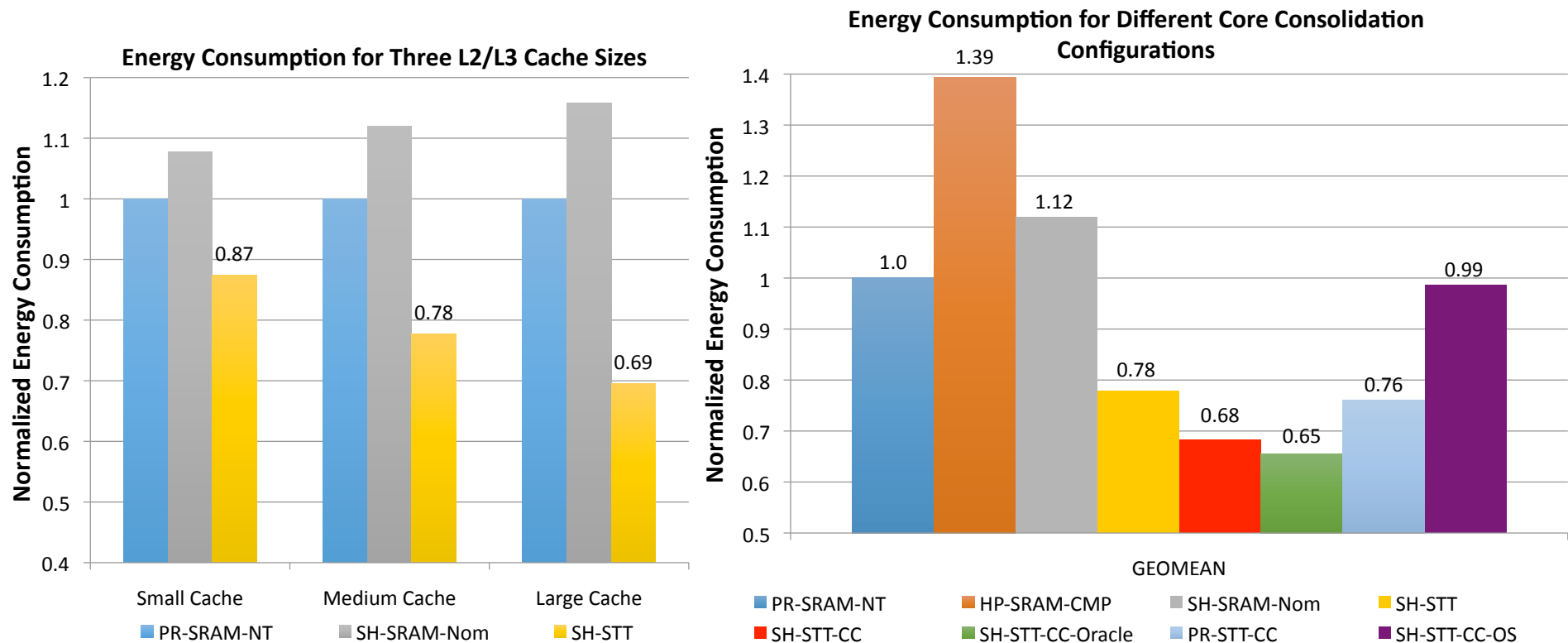
- Respin achieved **14%** power reduction and **11%** performance improvement with medium sized cache





Energy Consumption

- For medium sized cache, Respin achieved **22%** energy savings with the basic shared STT-RAM cache design plus additional **10%** with core consolidation enabled

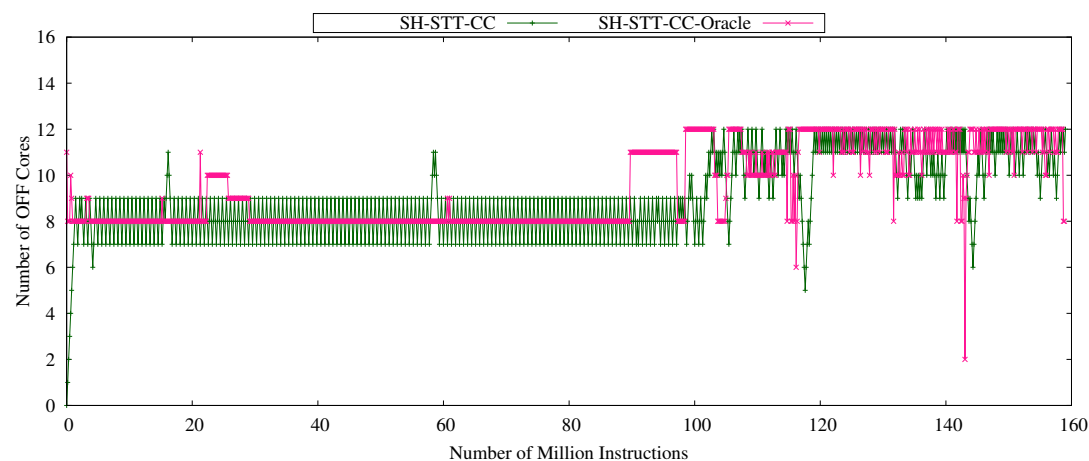




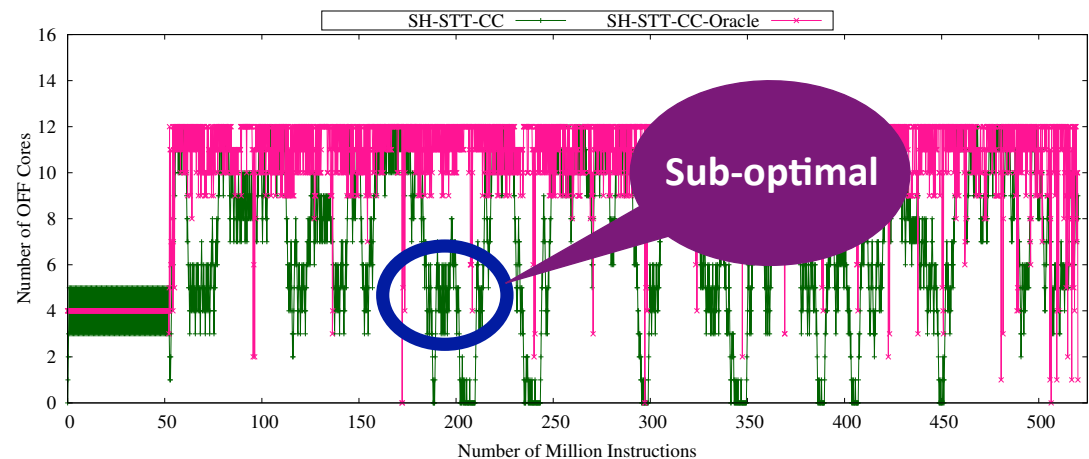
Core Consolidation Analysis

- In most cases our greedy algorithm matches well with the oracle while in very few cases sub-optimal selection becomes the barrier to slow down the pace of our greedy mechanism

radix (SPLASH2):
greedy achieved **48%**
energy savings while
oracle achieved **50%**



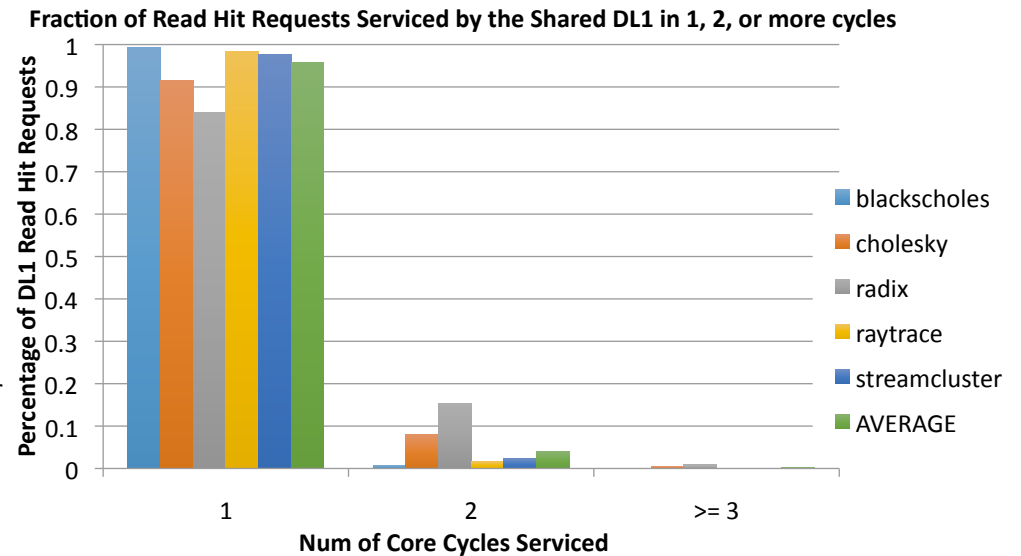
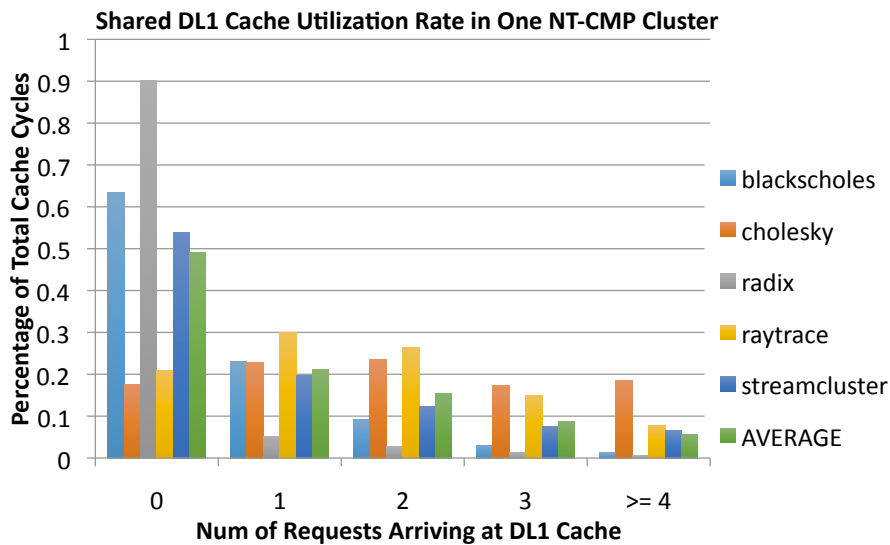
lu (SPLASH2): greedy
achieved **29%** energy
savings while oracle
achieved **38%**





Shared Cache Access Load

- More than **95%** of the incoming requests can be serviced in one processor core cycle





Sensitivity Study on Cluster Size

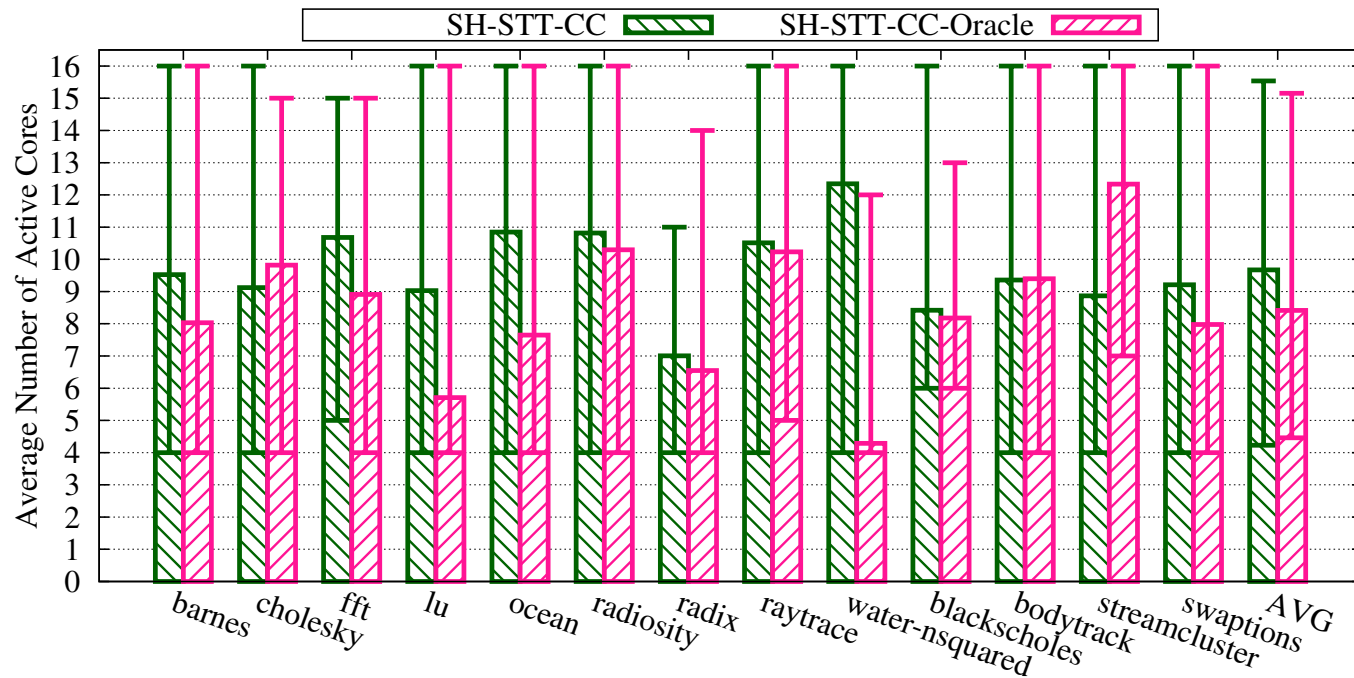
- 16-core cluster provides the best trade-off between data sharing among cores and shared cache access load, giving **~11%** performance gain

Cluster Size (#cores)	Shared L1 (I/D) Size (KB)	Performance Gain (%)
4	64	4.82
8	128	6.29
16	256	10.81
32	512	2.50



Number of Active Cores in Dynamic Core Consolidation

- Strong phased behavior can be observed across all applications, making our dynamic core consolidation mechanism very useful





Outline

- Basics of NVM
- Respin Architecture
- Methodology and Evaluation
- **Conclusion**



Conclusion

- The first work to explore the use of non-volatile caches in near-threshold chip multi-processors
- A novel architecture designed to enhance NT-CMP performance and reduce energy consumption by sharing L1 caches and implementing dynamic core consolidation mechanism
- Achieved energy reduction by **33%** and improved performance by **11%**



Questions?

Thank you!