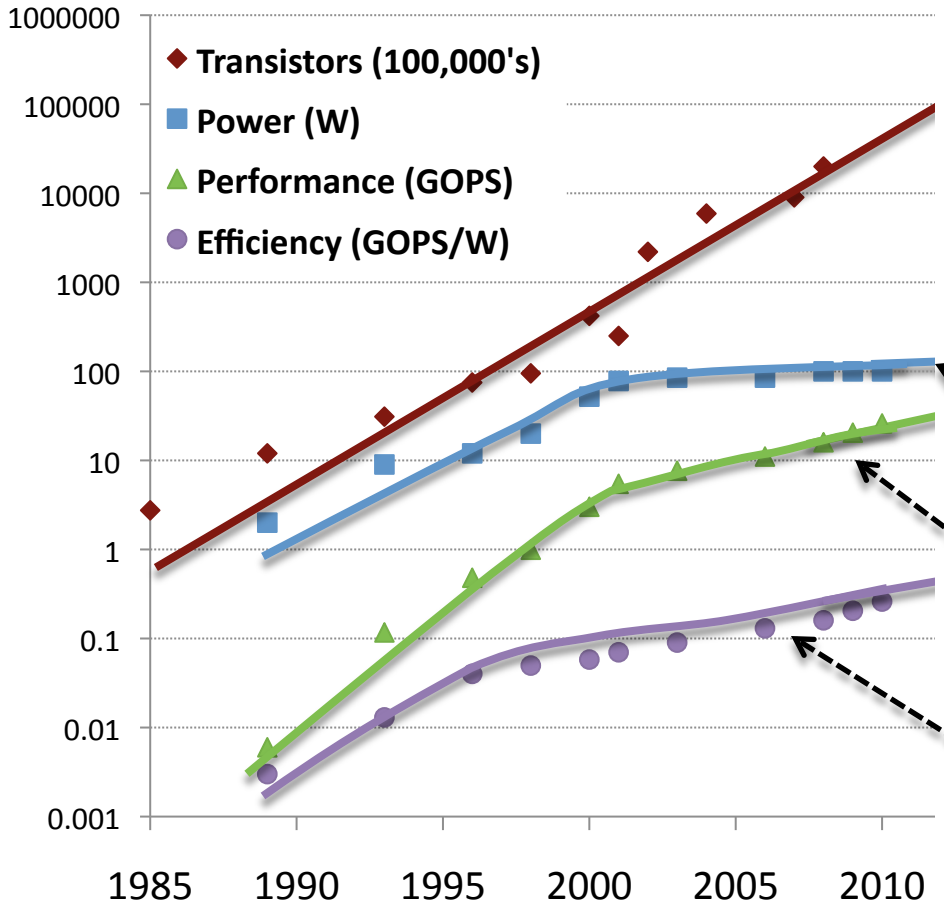
Near-Threshold Computing: Reclaiming Moore's Law

Dr. Ronald G. Dreslinski

Research Fellow

University of Michigan – Ann Arbor

Motivation

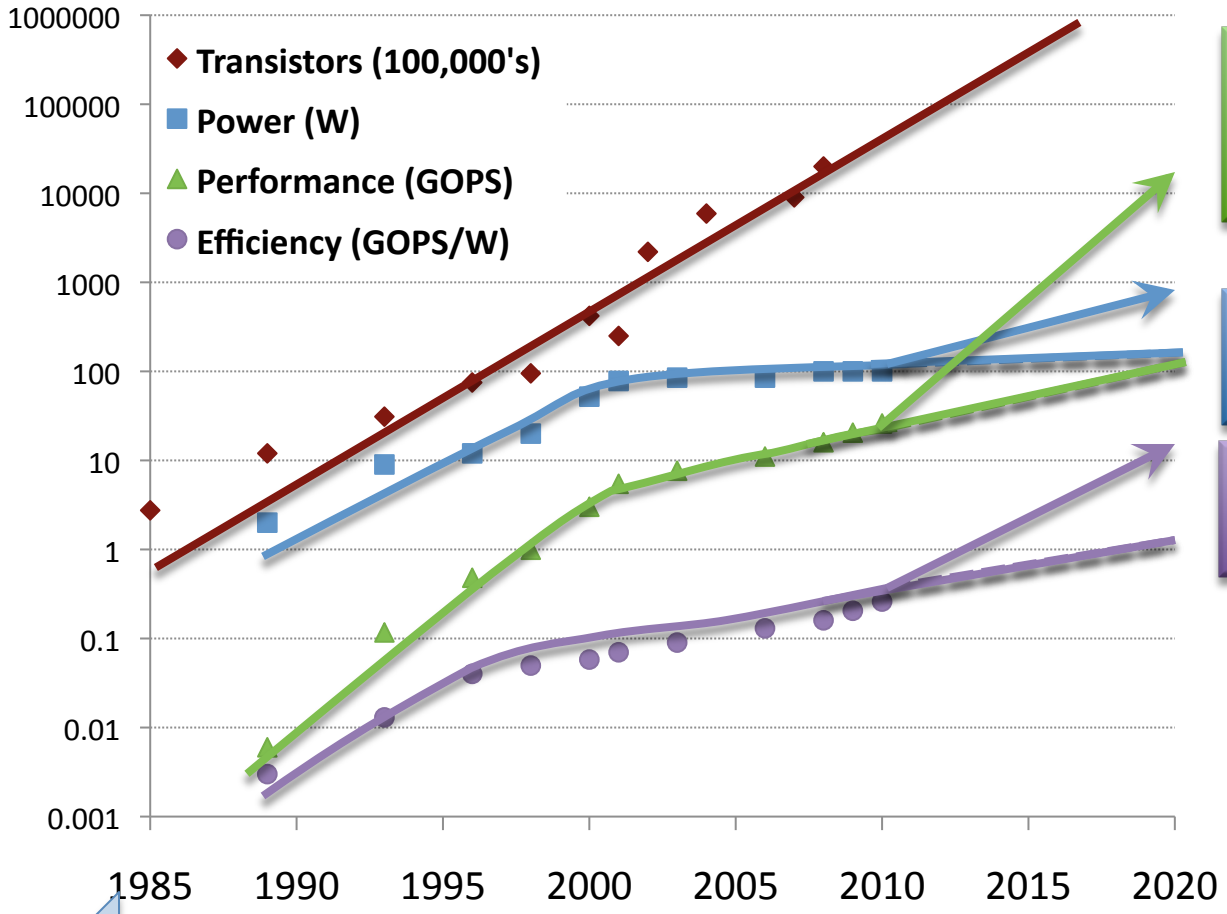


Limits on heat extraction

Stagnates performance growth

Limits on energy-efficiency of operations

Motivation



Result: Continue scaling trends that fueled the computing revolution

With the help of some better thermal management…

Goal: To increase energy-efficiency of operations

Era of High Performance Computing

c. 2000

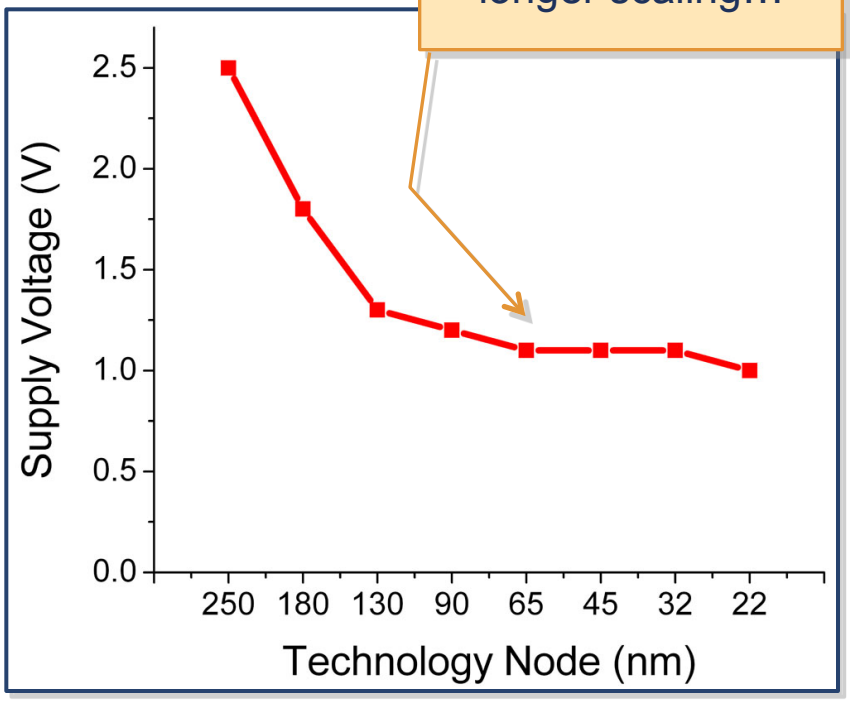
Era of Energy-Efficient Computing

Outline

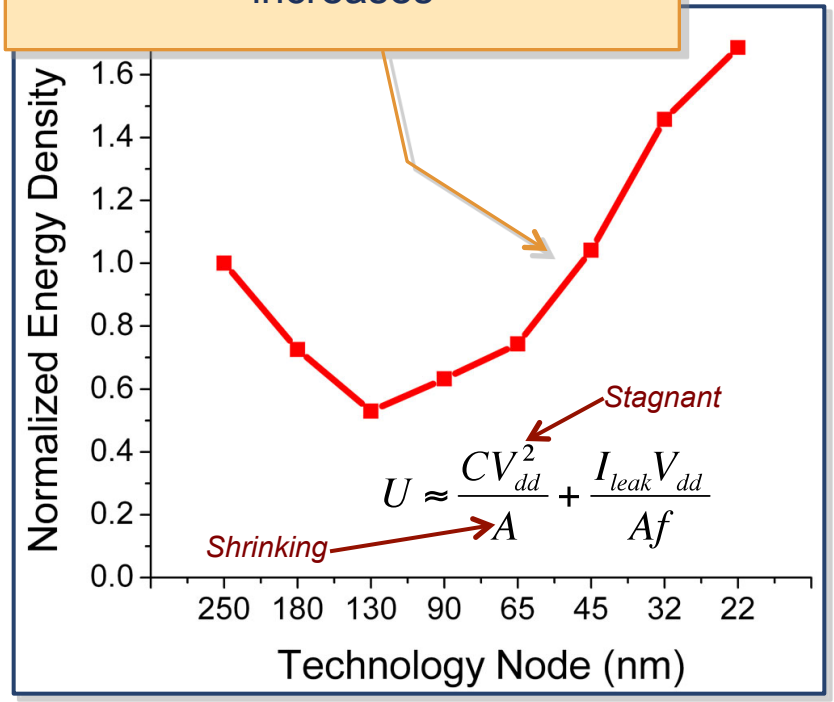
- Define a new region of operation, **Near-Threshold Computing**
- Explore **new architectures** enabled by key insights of computing in the NTC region
- Present an initial design of a 3D stacked NTC system, **Centip3De**

Power Density Limitations

Circuit supply voltages are no longer scaling…



Power does not decrease at the same rate that transistor count increases



Dynamic dominates

$$U \approx \frac{CV_{dd}^2}{A} + \frac{I_{leak}V_{dd}}{Af}$$

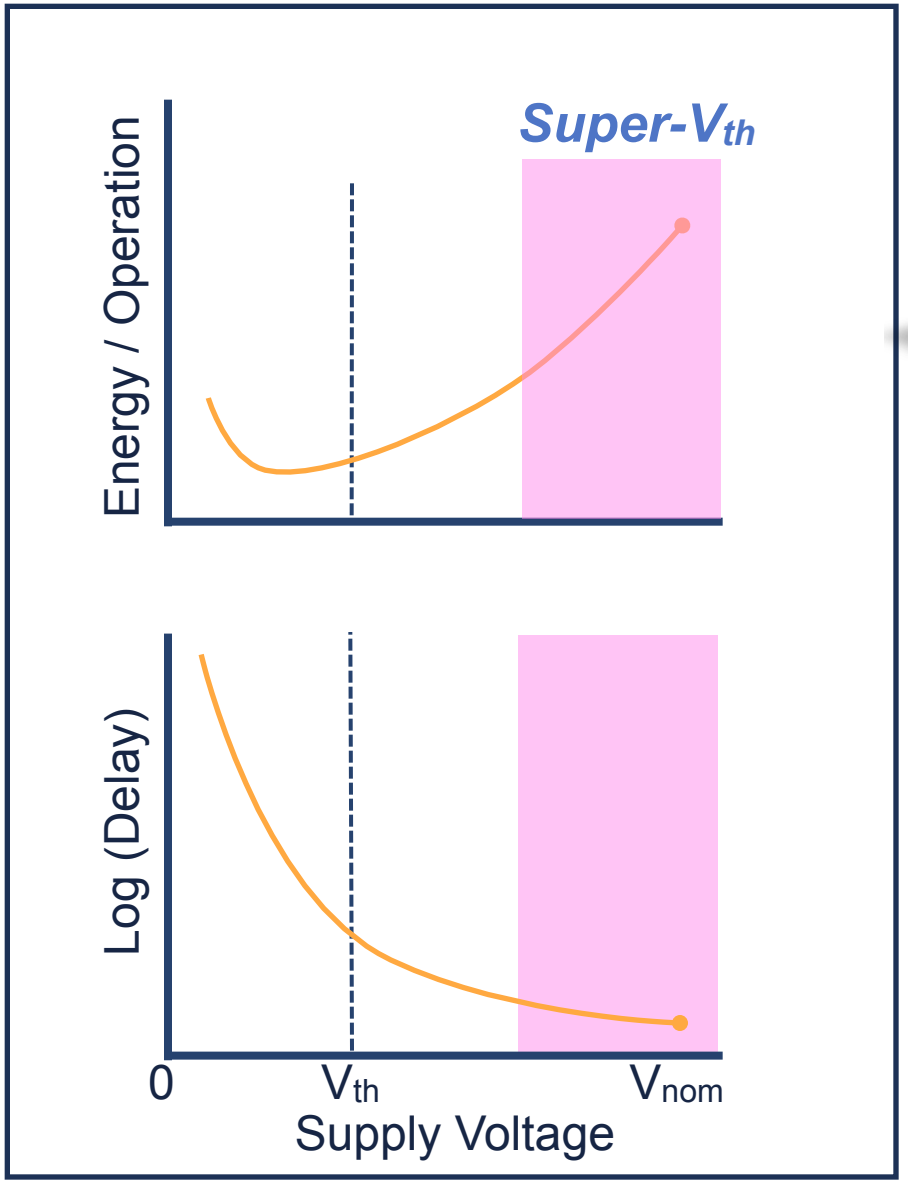
A = gate area → scaling $1/s^2$

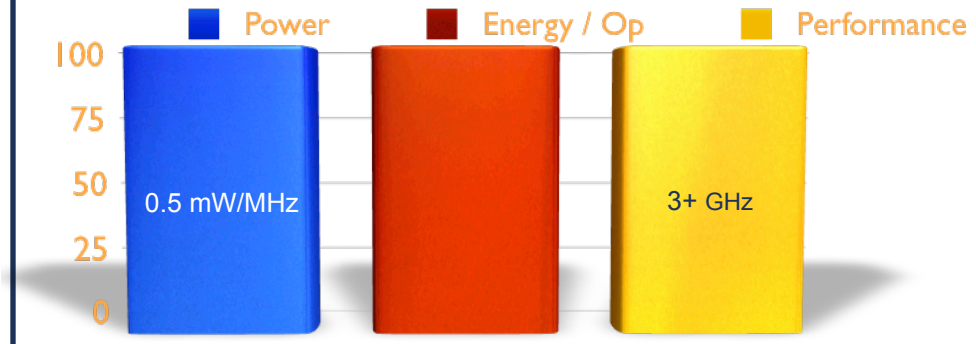
C = capacitance → scaling < 1/s

Dark Silicon—The emerging dilemma:

More and more gates can fit on a die,
but not all can be turned on at the same time

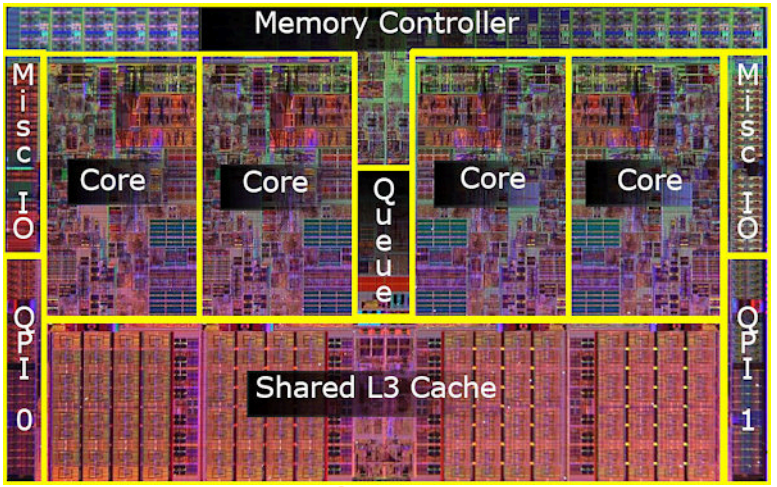
Today: Super-V_{th}, High Performance, Power Constrained





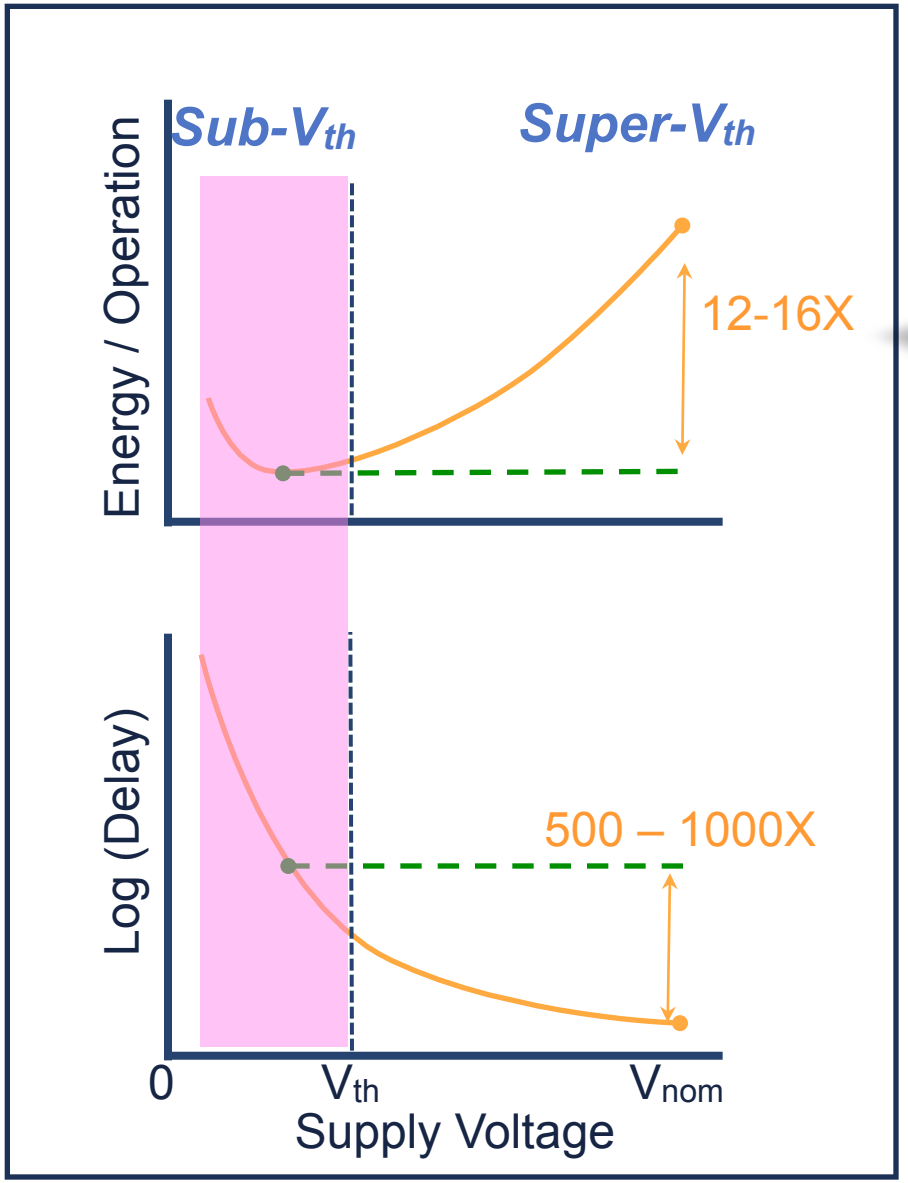
Normalized Power, Energy, & Performance

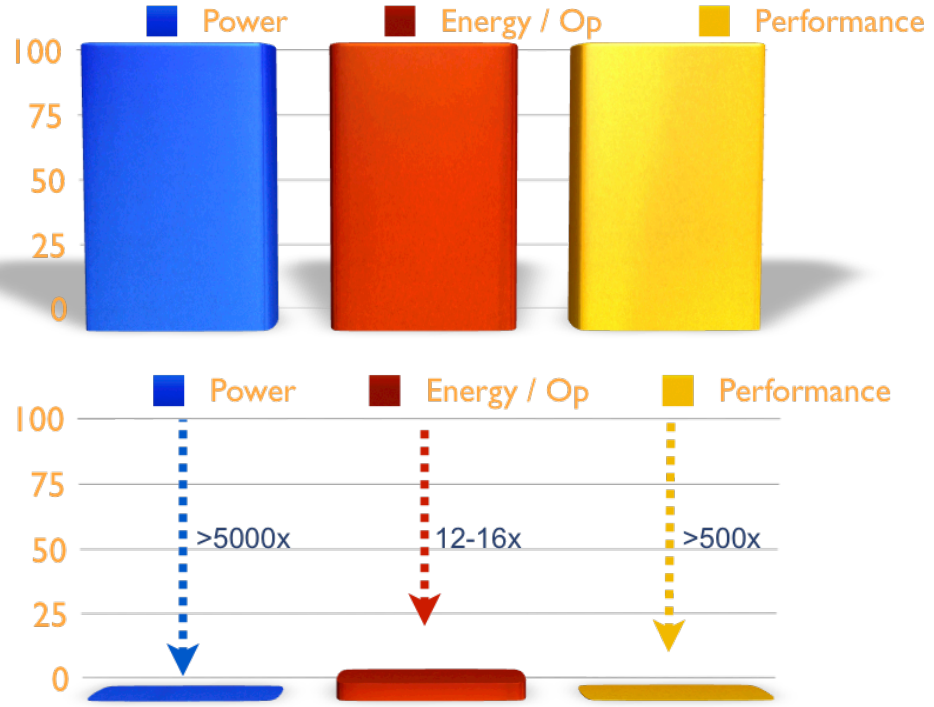
Energy per operation is the key metric for efficiency. Goal: same performance, low energy per operation



Core i7

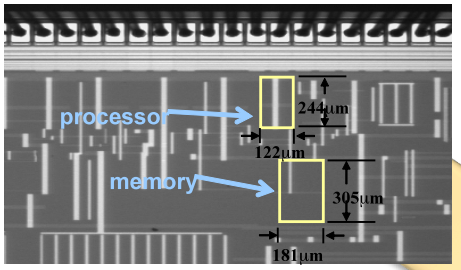
Subthreshold Design





Operating in the sub-threshold gives us huge power gains at the expense of performance → OK for sensors!

Evolution of Subthreshold Designs

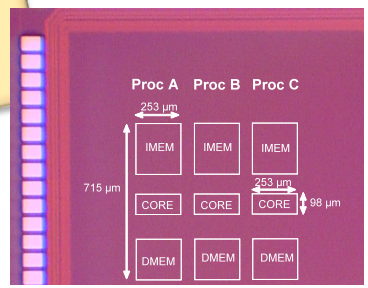


Subliminal 1 Design (2006)

-0.13 µm CMOS

-Used to investigate existence of Vmin

-2.60 µW/MHz



Subliminal 2 Design (2007)

-0.13 µm CMOS

-Used to investigate process variation

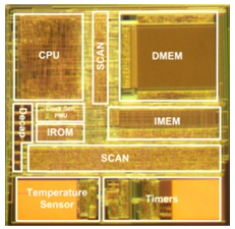
-3.5 µW/MHz

Phoneix 1 Design (2008)

- 0.18 µm CMOS

-Used to investigate sleep current

-2.8 µW/MHz / 30pW sleep power



Phoenix 2 Design (2010)

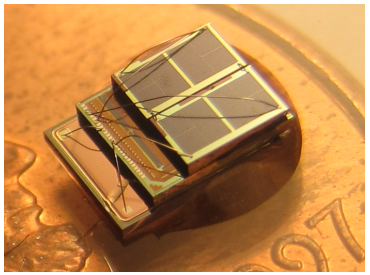
- 0.18 µm CMOS

-Commercial ARM M3 Core

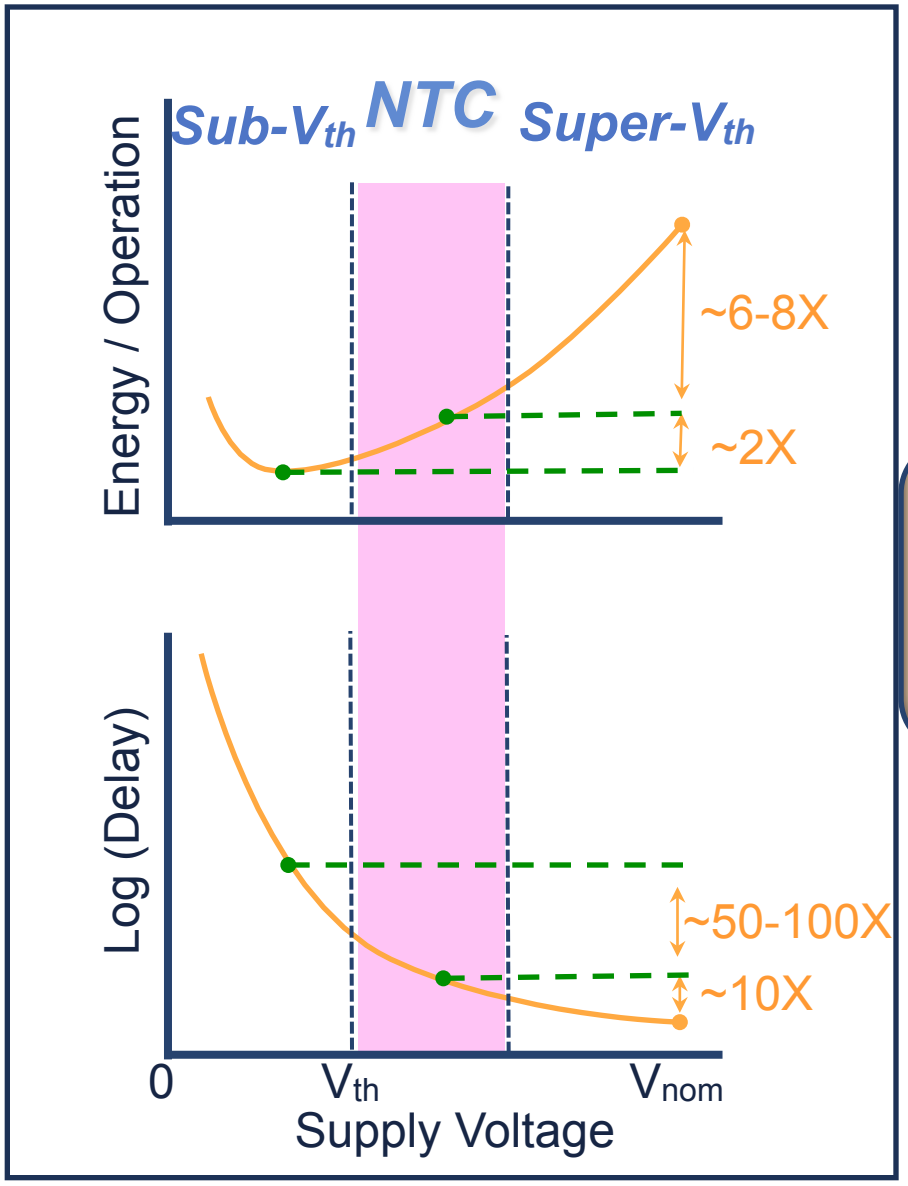
-Used to investigate:

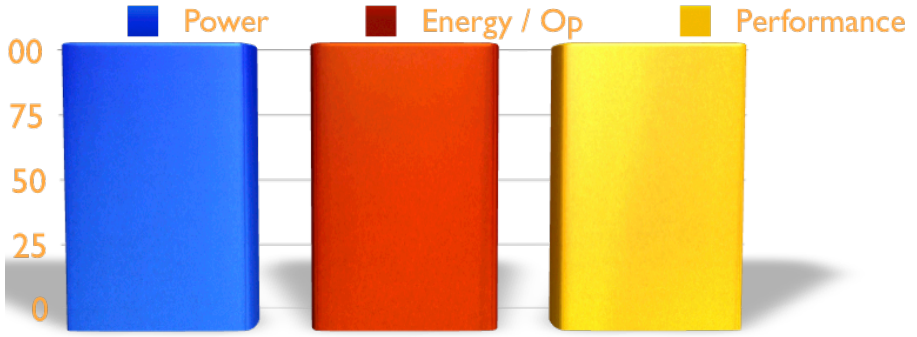
- •Energy harvesting
- •Power management

-37.4 µW/MHz



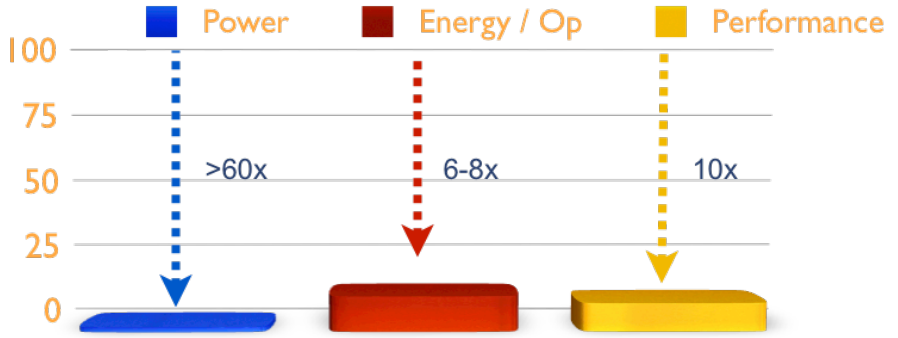
Near-Threshold Computing (NTC)



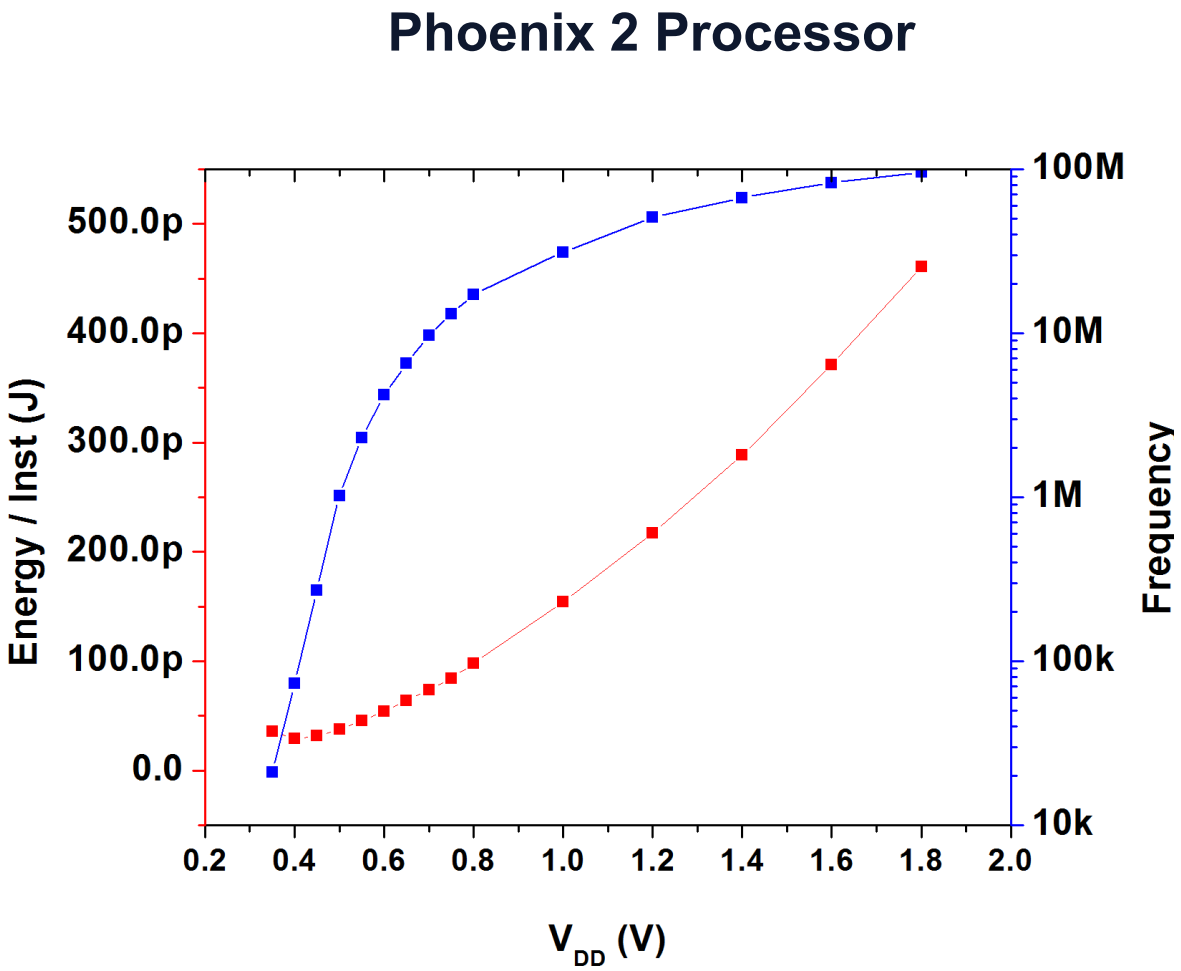


Near-Threshold Computing (NTC):

- *>60X power reduction*
- *6-8X energy reduction*
- *Invest portion of extra transistors from scaling to overcome barriers*



Silicon Verification of Trends



Phoenix 2 Design [Seok'11]

180nm Design

1.8V -> 700mV
~10x NTC Performance Loss
~7x NTC Energy Reduction

Seok ISSCC 2011

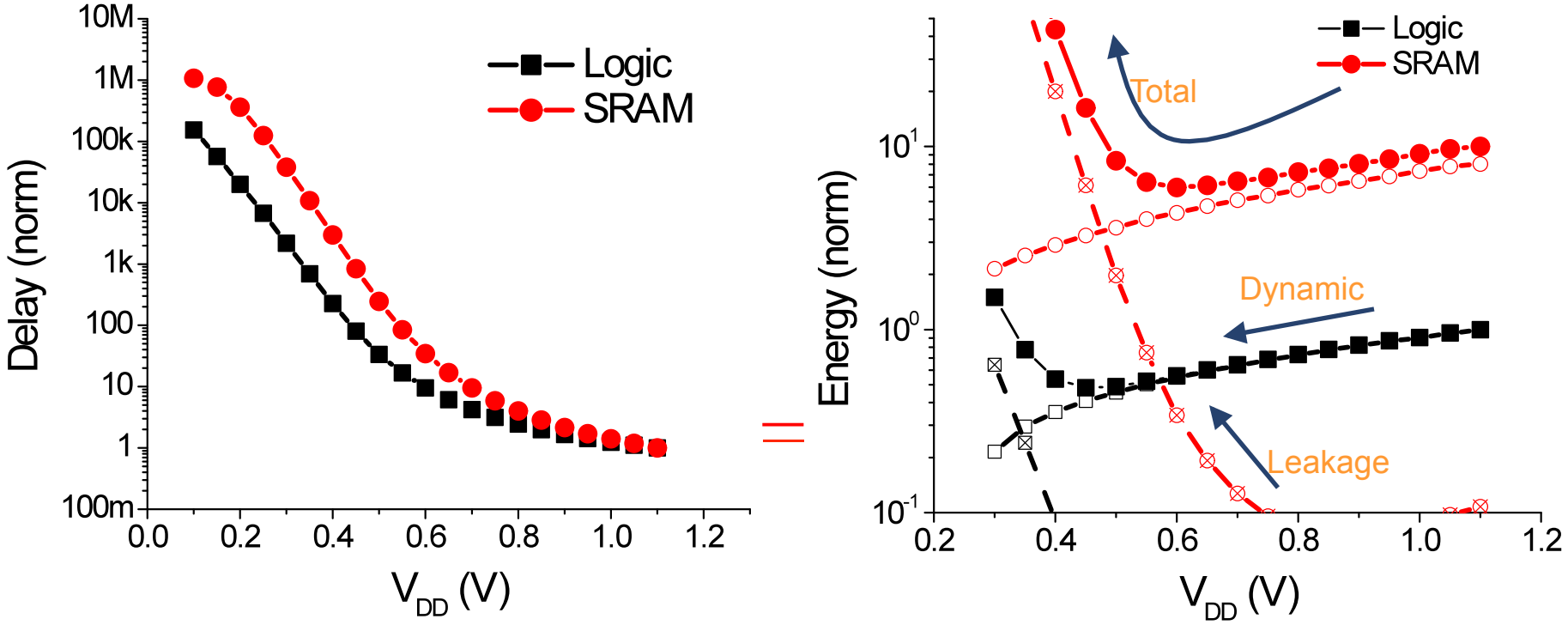
NTC – Opportunities and Challenges

- Opportunities:
  - New architectures
  - Optimized Processes
  - 3D Integration – less thermal restrictions
- Challenges:
  - Low Voltage Memory
    - New SRAM designs
    - Robustness analysis at near-threshold
  - Variation
    - Razor [Ernst'03] and other in-situ delay monitoring
    - Adaptive body biasing
  - Performance Loss
    - Many-core designs to improve parallelism
    - Core boosting to improve single thread performance

Outline

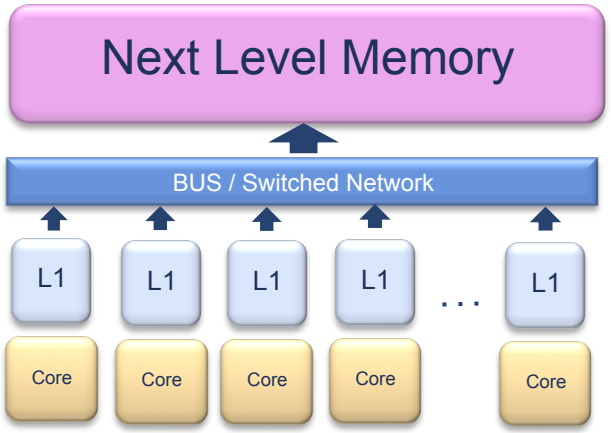
- Define a new region of operation, **Near-Threshold Computing**
- Explore **new architectures** enabled by key insights of computing in the NTC region
- Present an initial design of a 3D stacked NTC system, **Centip3De**

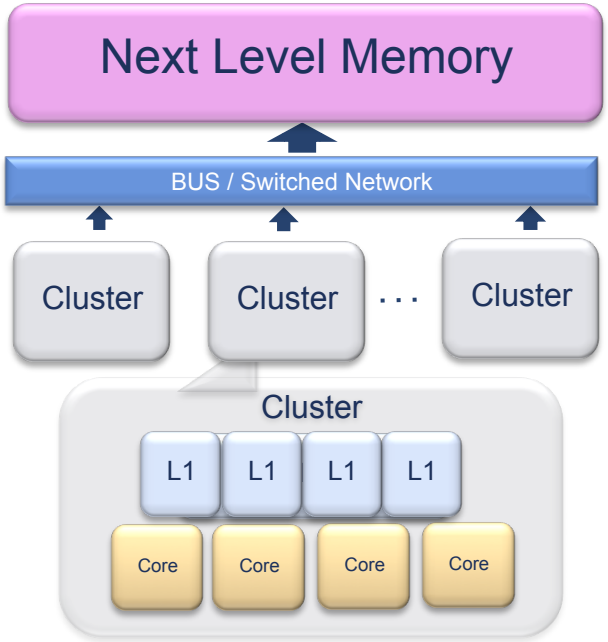
Minimum Energy SRAM



- SRAM has a lower activity rate than logic
- VDD for minimum energy operation (V_{MIN}) is higher
- Running logic at V_{MIN} for SRAM has a small energy penalty with increased performance

New NTC Architectures





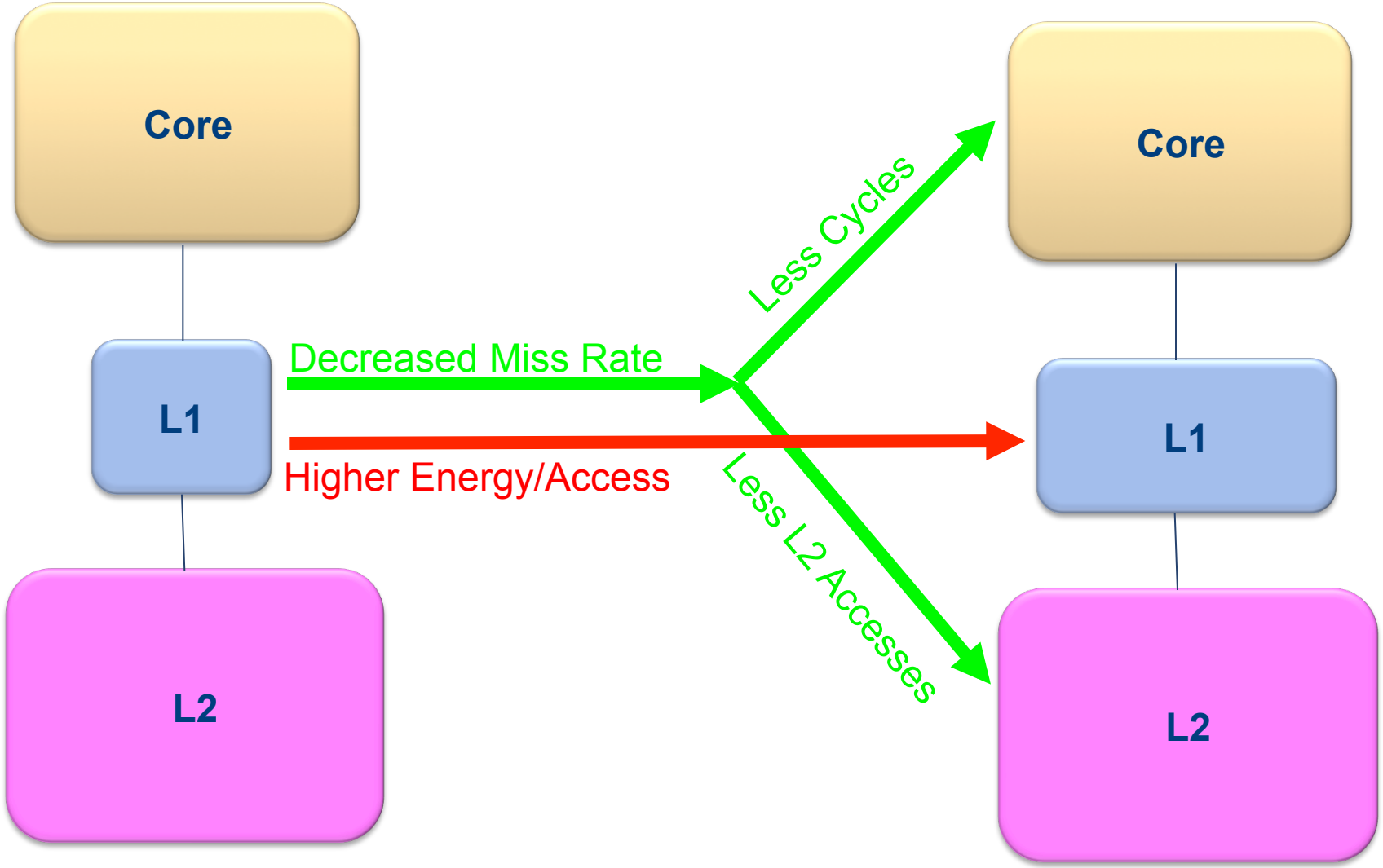
Key Insight:

- SRAM is run at a higher V_{DD} than cores with little energy penalty, allowing caches to operate faster than the core

Design Levers:

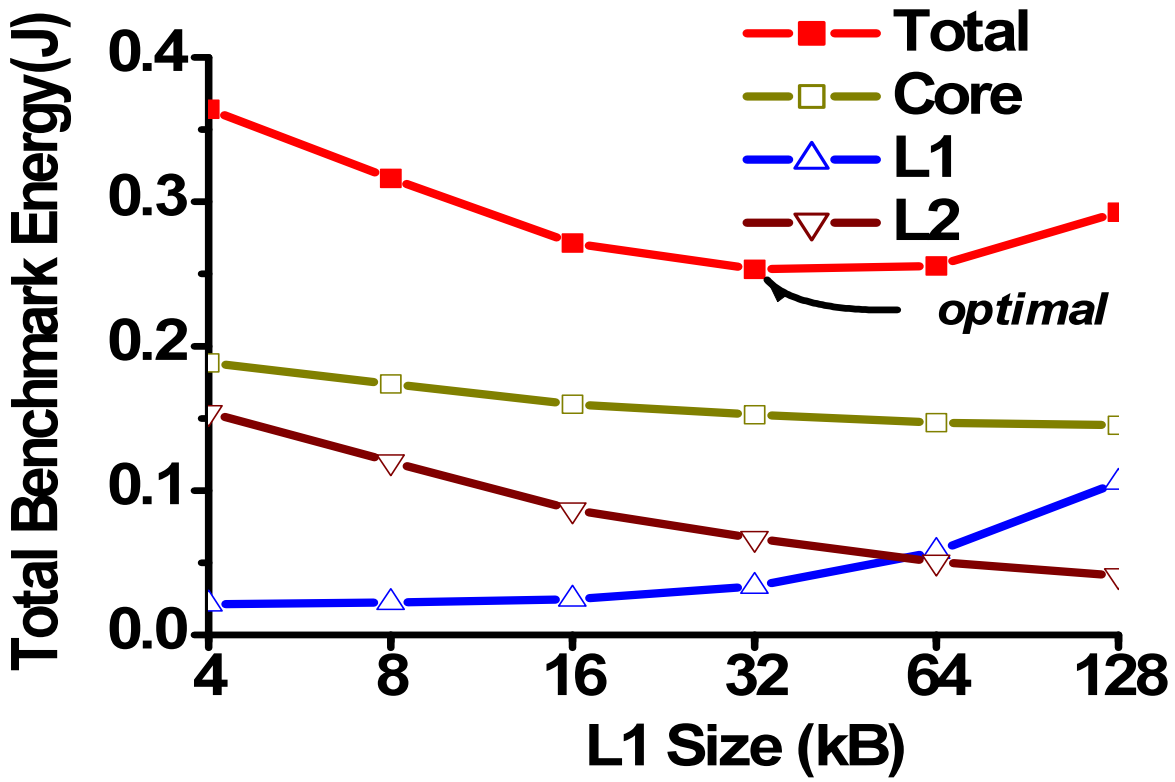
- Operating Voltage
- L1 Size
- Number of Cores per Cluster
- Number of Clusters

L1 Cache Size Tradeoff

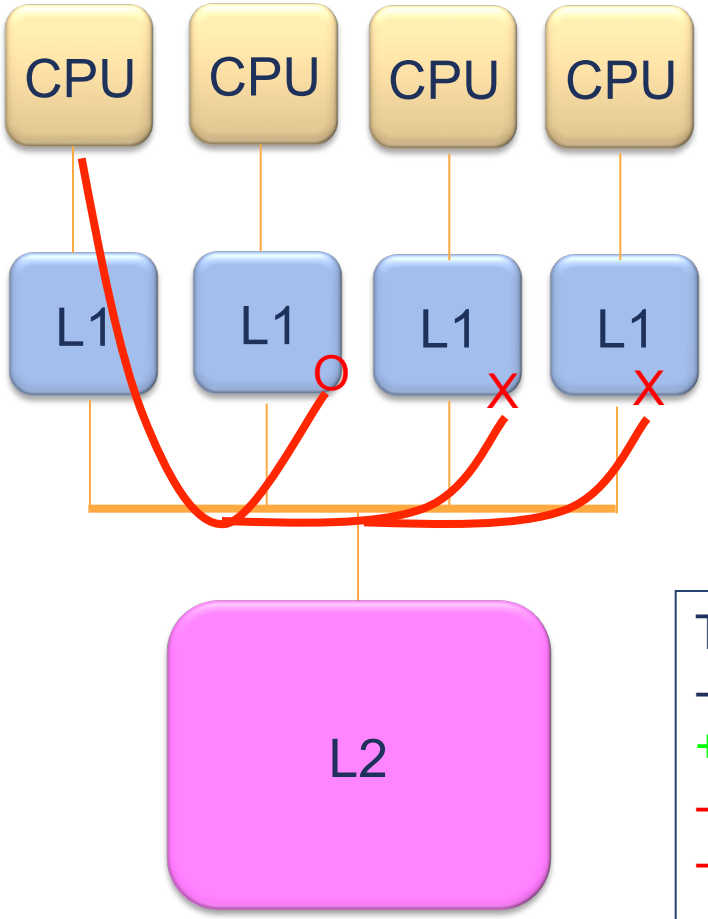


Results – Energy Optimal L1 Size (Single Core)

- Energy dependency on L1 size
  - Trade-off between L1 and L2 access

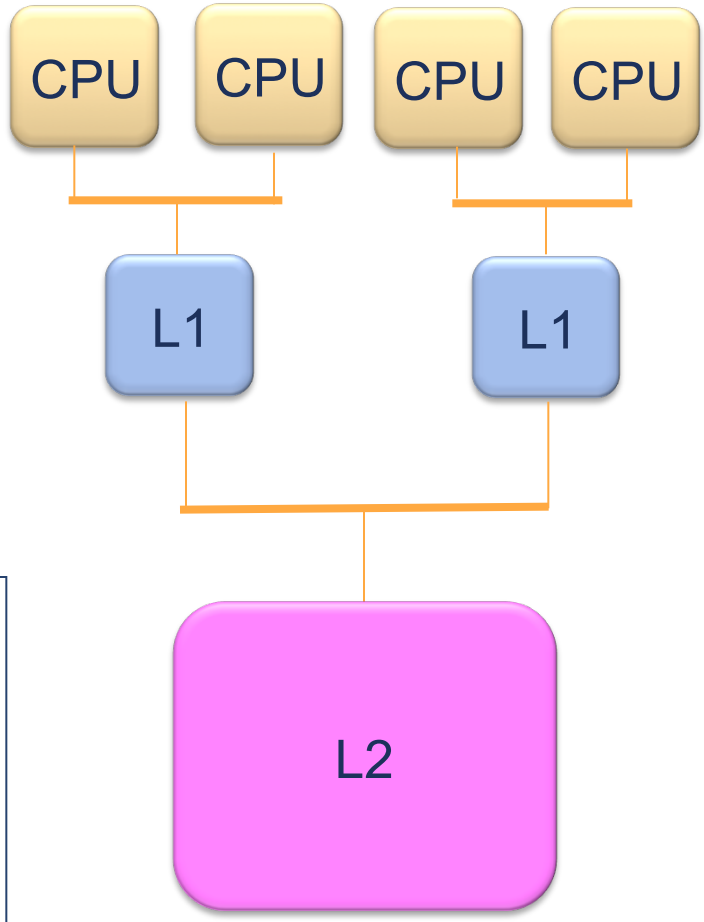


Clustering Tradeoffs

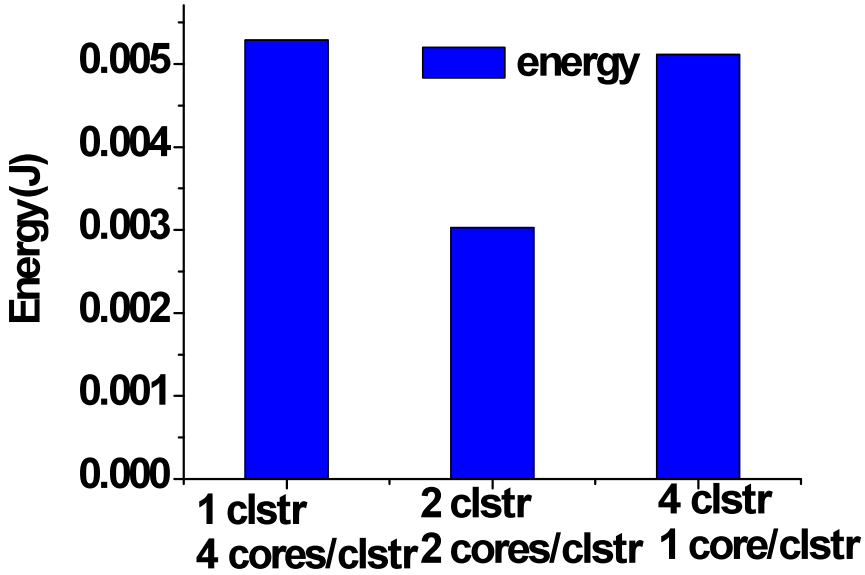


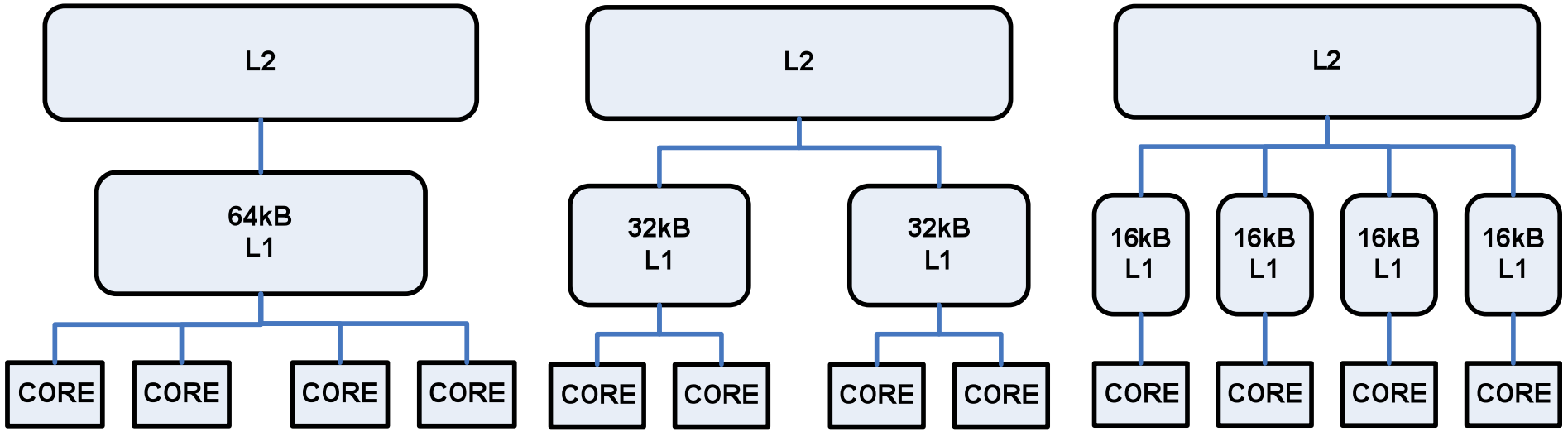
Tradeoffs

+ Clustered Sharing
- Cluster Conflict
- New Bus
- L1 Speed

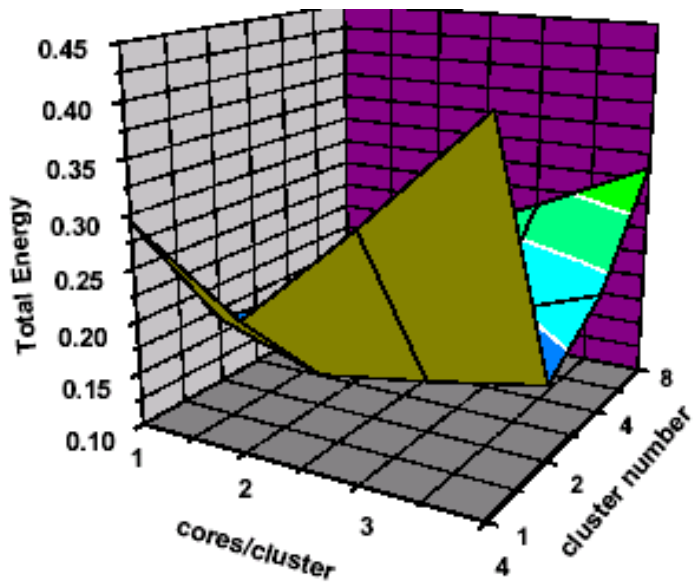


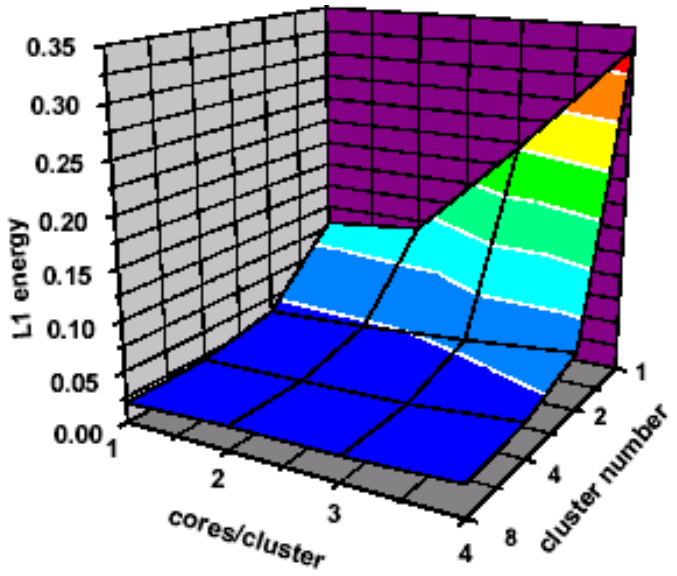
Energy Optimal Cluster-based CMP (Fixed Die Size)

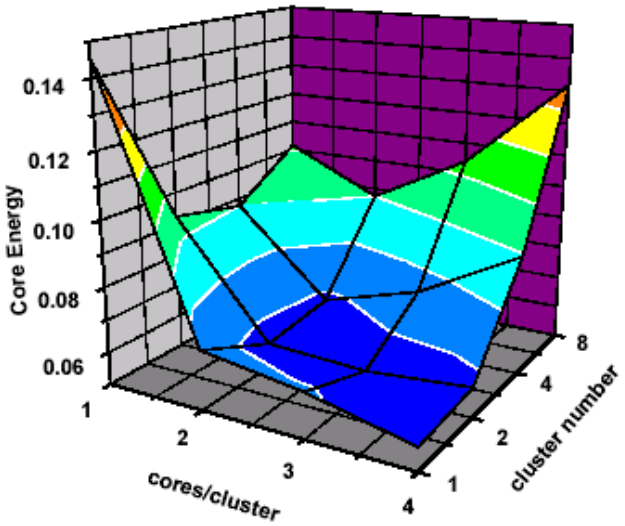


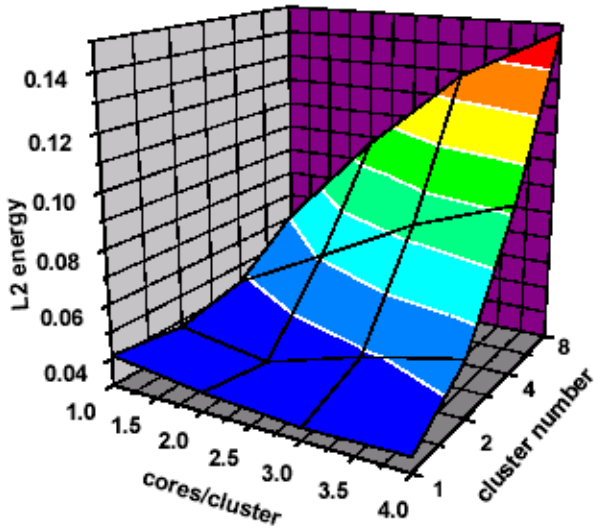


Full Space Analysis

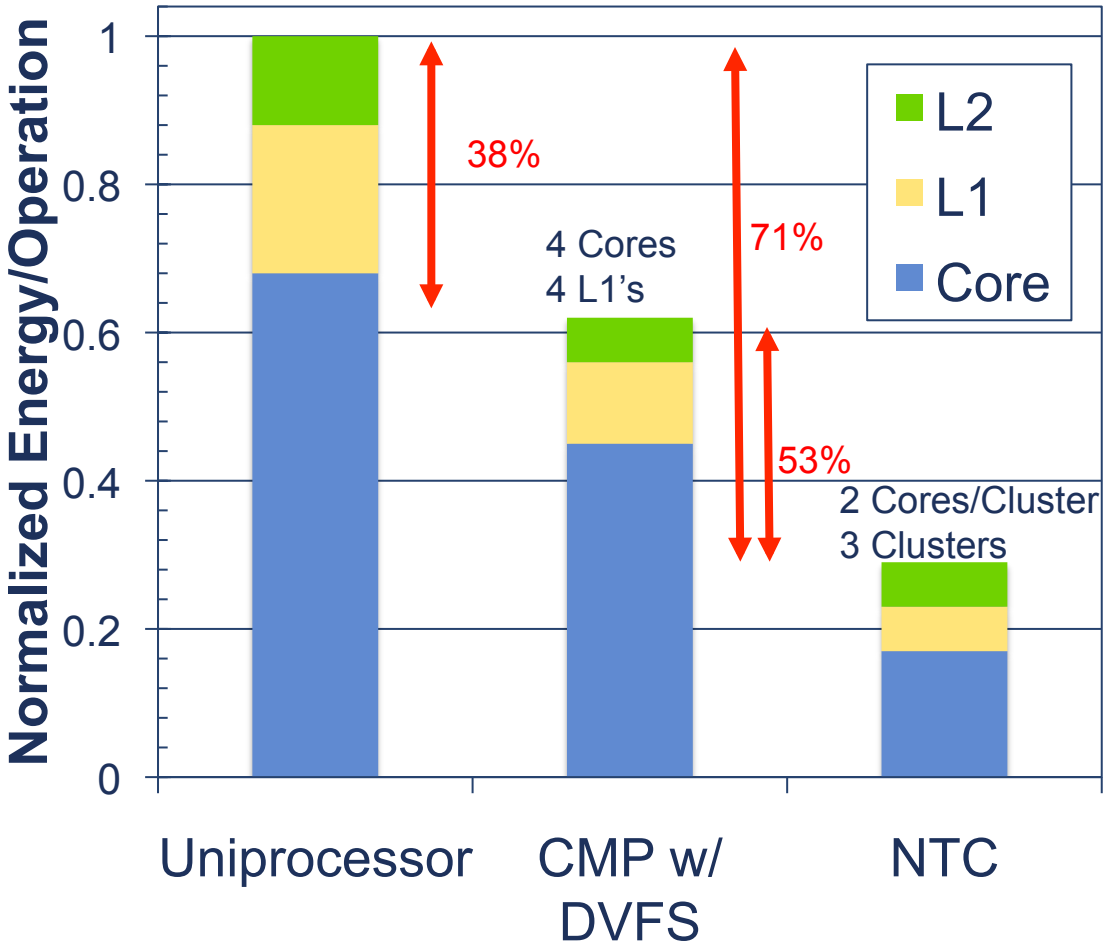








Various Scaling Methods



- Baseline
  - Single CPU @ 233MHz
- Simple CMP
  - One core per L1
  - Vdd scaling
- Proposed cluster-based CMP
  - Multiple cores per L1
  - Vdd scaling

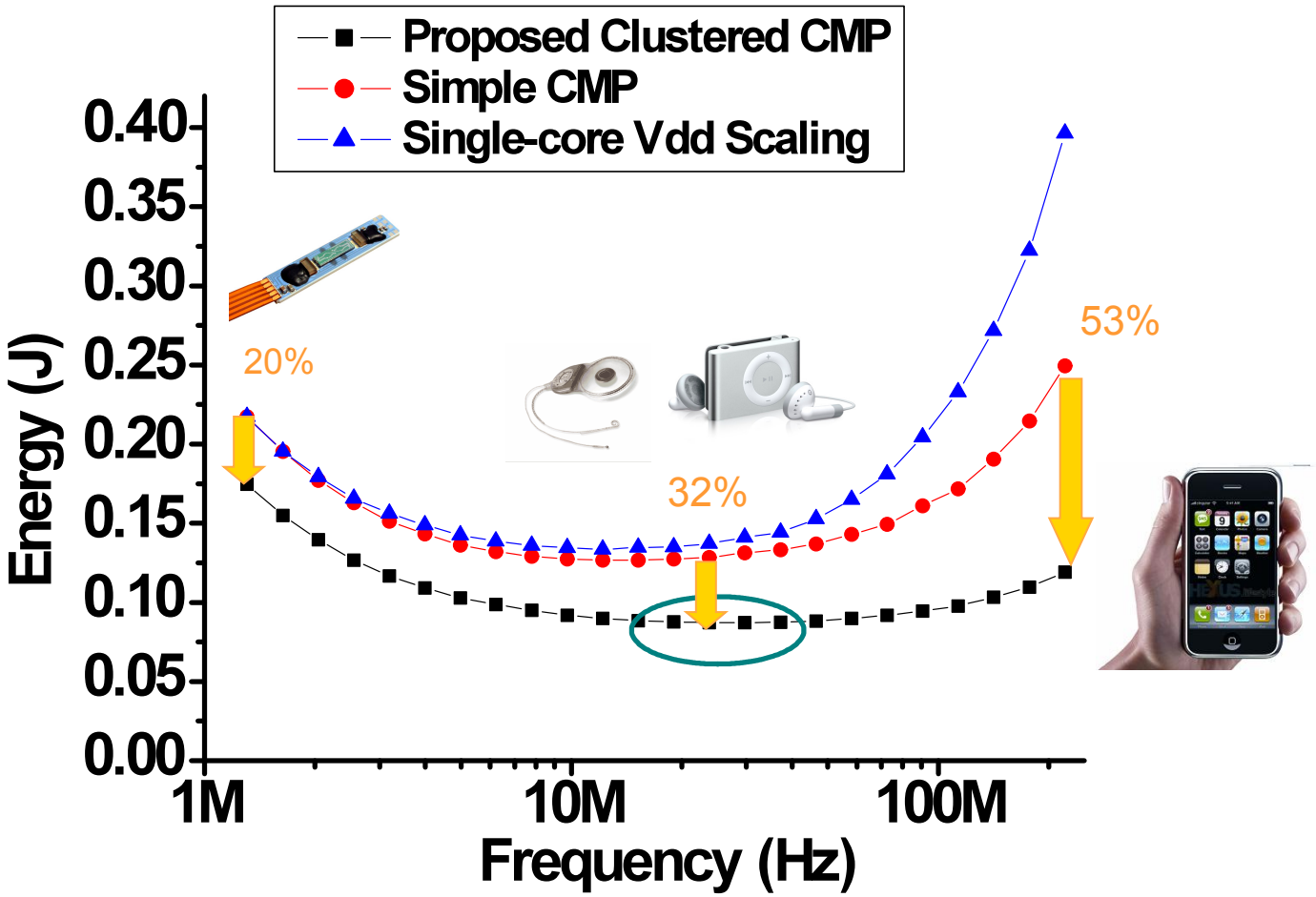
Energy Optima for SPLASH2

- Cluster based architecture with Vdd and Vth scaling
  - Optimal cluster size is 2 for most of the apps
    - Rad choose non-clustered CMP
  - **Average: 74% over baseline, 55% over simple CMP**

	n_c	k	L1 size/kB	energy savings over baseline	energy savings over simple CMP
Cho	3	**2**	64	70.8%	52.8%
Fft	2	**2**	32	72.6%	68.5%
fmm	8	**2**	128	79.7%	41.6%
luc	3	**2**	32	77.8%	64.4%
lun	2	**2**	64	69.2%	58.0%
rad	16	**1**	128	84.2%	35.1%
ray	3	**2**	128	65.1%	54.9%

Energy Optima w/ Performance Requirements

- Cluster based approach provides best savings
  - Traditional approach only saves energy at high end



Outline

- Define a new region of operation, **Near-Threshold Computing**
- Explore **new architectures** enabled by key insights of computing in the NTC region
- Present an initial design of a 3D stacked NTC system, **Centip3De**

A Closer Look at Wafer-Level Stacking

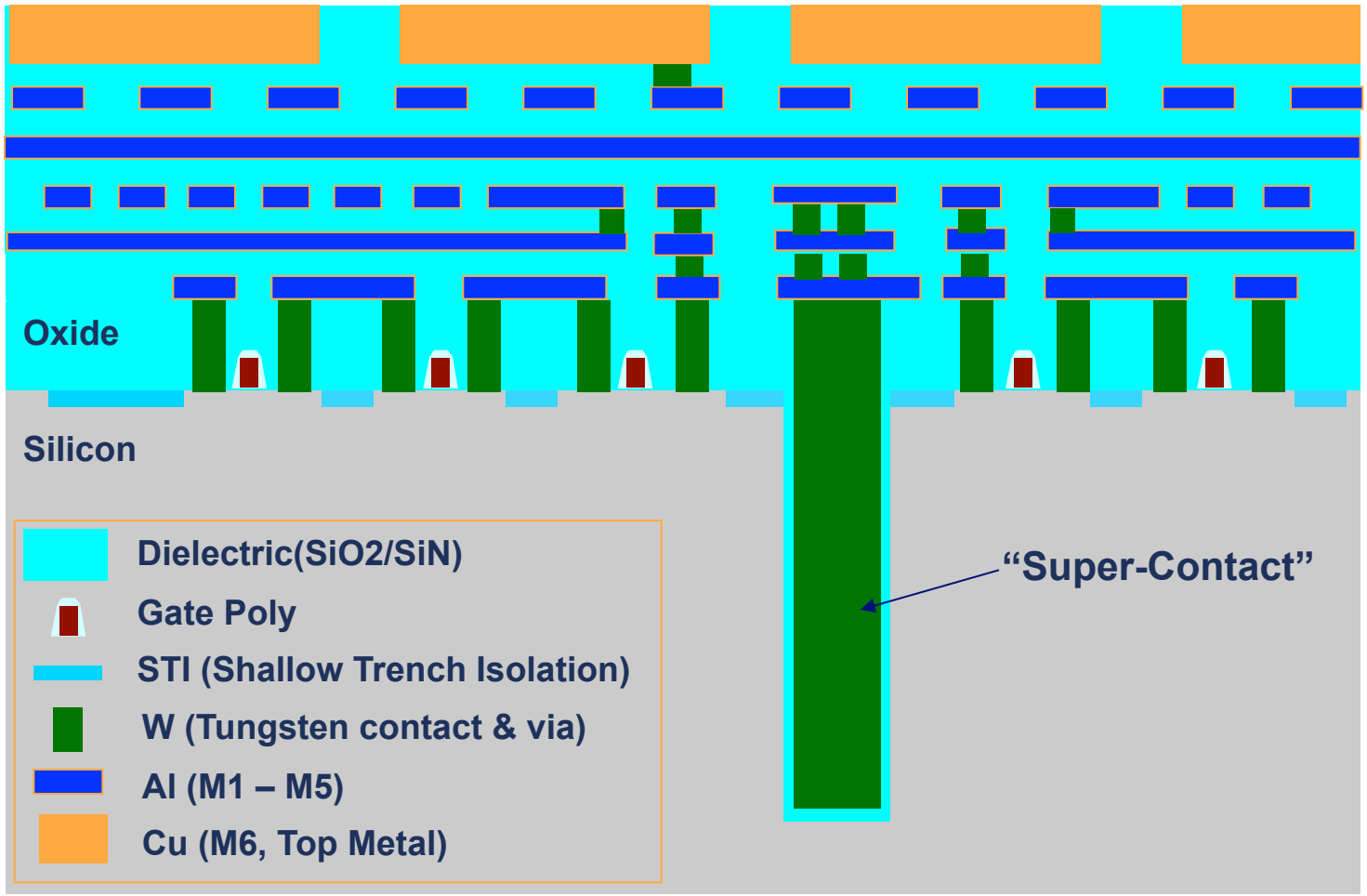
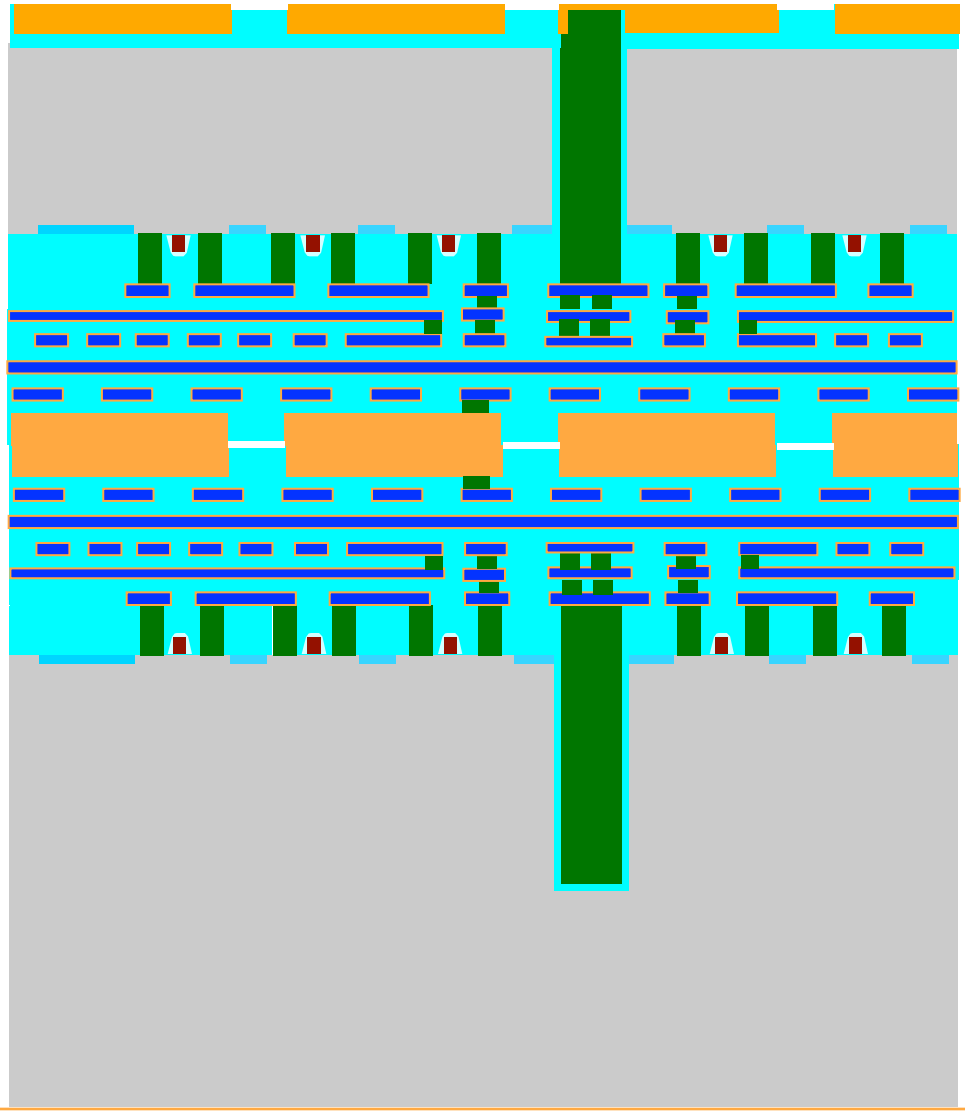
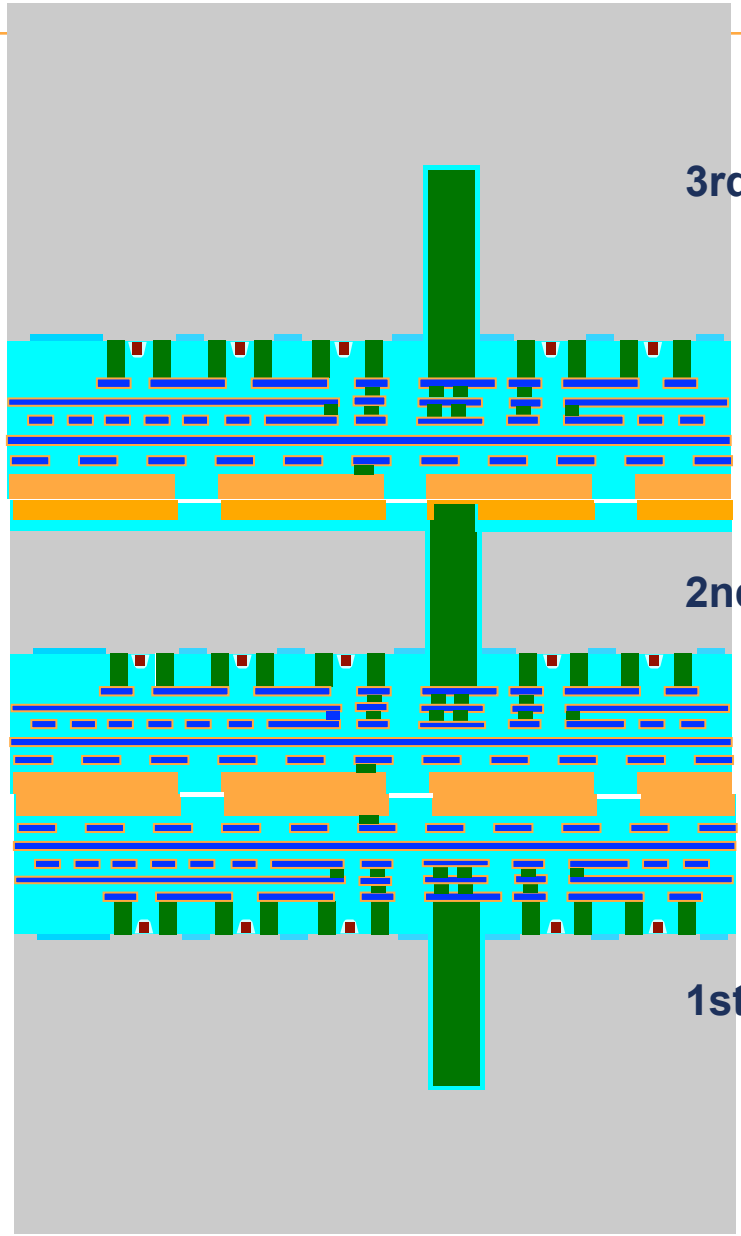


Illustration from Bob Patti, Tezzaron

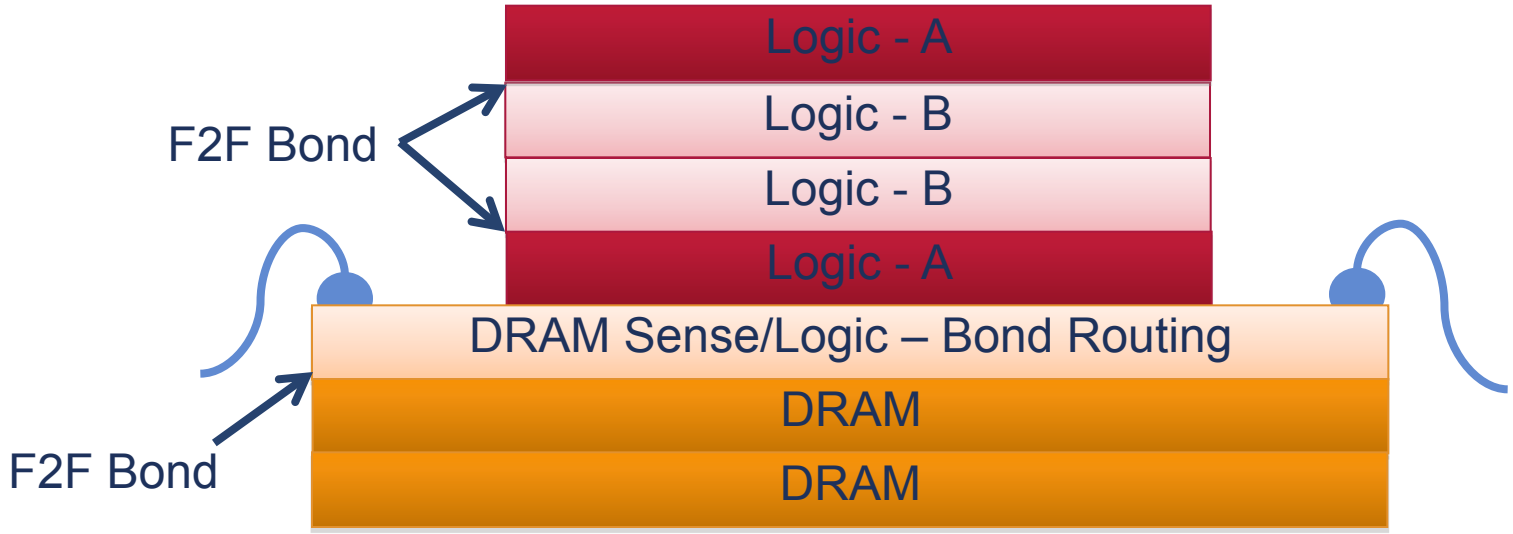
Next, Stack a Second Wafer & Thin:



Then, Stack a Third Wafer:

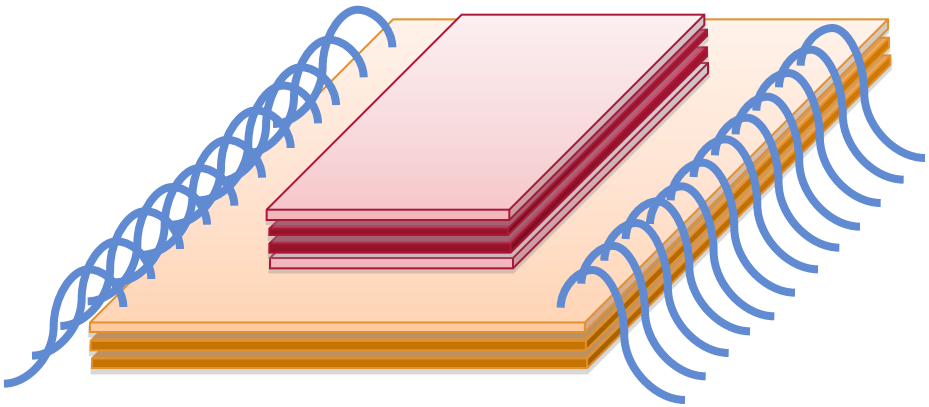


Centip3De – 3D NTC Prototype



Centip3De Design

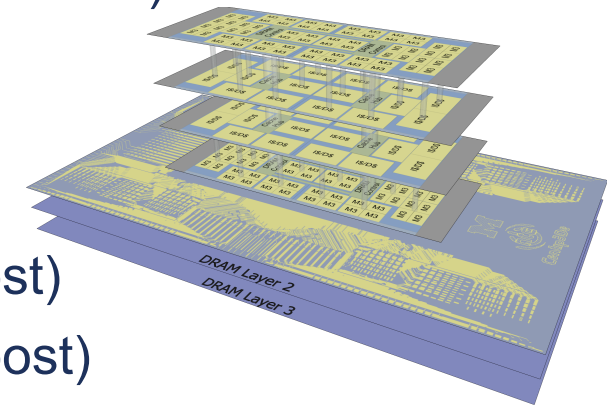
- 130nm, 7-Layer 3D-Stacked Chip
- 128 - ARM M3 Cores
- *150mm²*



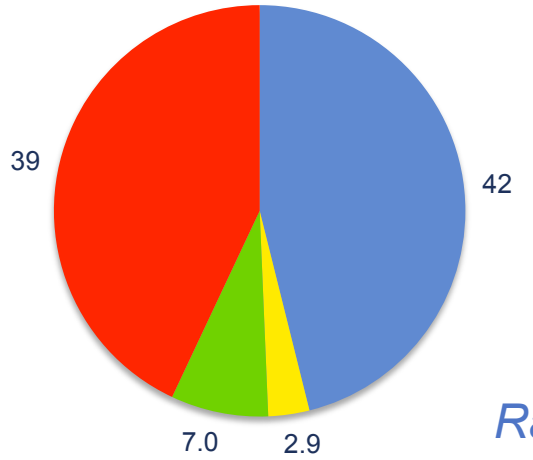
Design Scaling and Power Breakdowns

NTC Centip3De System

- 1.9 GOPS (3.8 GOPS in Boost)
  - Max 1 IPC per core
  - 128 Cores
  - 15 MHz
- 130 mW (691mW in Boost)
- ***14.6 GOPS/W*** (5.5 in Boost)
- **Naïve Scaling to 22nm yields ~200GOPS/W**

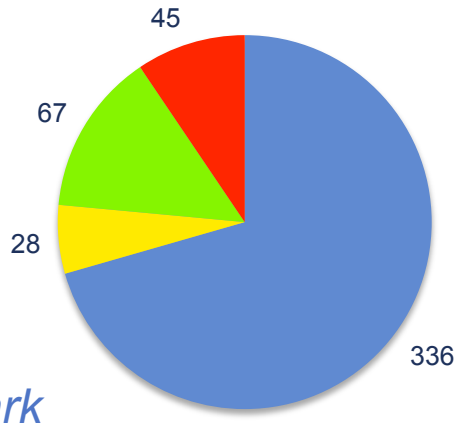


NTC Mode Power (mW)



Cores
I-Caches
D-Caches
DRAM

Boosted Mode Power (mW)



Raytracing Benchmark

Conclusions

- Observed Voltage Scaling and Thermal Limits reducing the gains of Moore's Law
- Defined a new computational operating region: **Near Threshold Computing**
- Leveraged key insights of NTC for **new clustered architectures**
- Initial ideas of a 3D integrated NTC system, **Centip3De**

Related References

• *Ronald G. Dreslinski, Michael Wieckowski, David Blaauw, Dennis Sylvester, Trevor Mudge, "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits," Proceedings of the IEEE, Special Issue on Ultra-Low Power Circuit Technology, Vol. 98, No. 2, February 2010, pg. 253 – 266.*

• *Bo Zhai, Ronald G. Dreslinski, Trevor Mudge, David Blaauw, Dennis Sylvester, "Energy Efficent Near-threshold Chip Multi-processing," ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED), August 2007, Best Paper Nomination.*

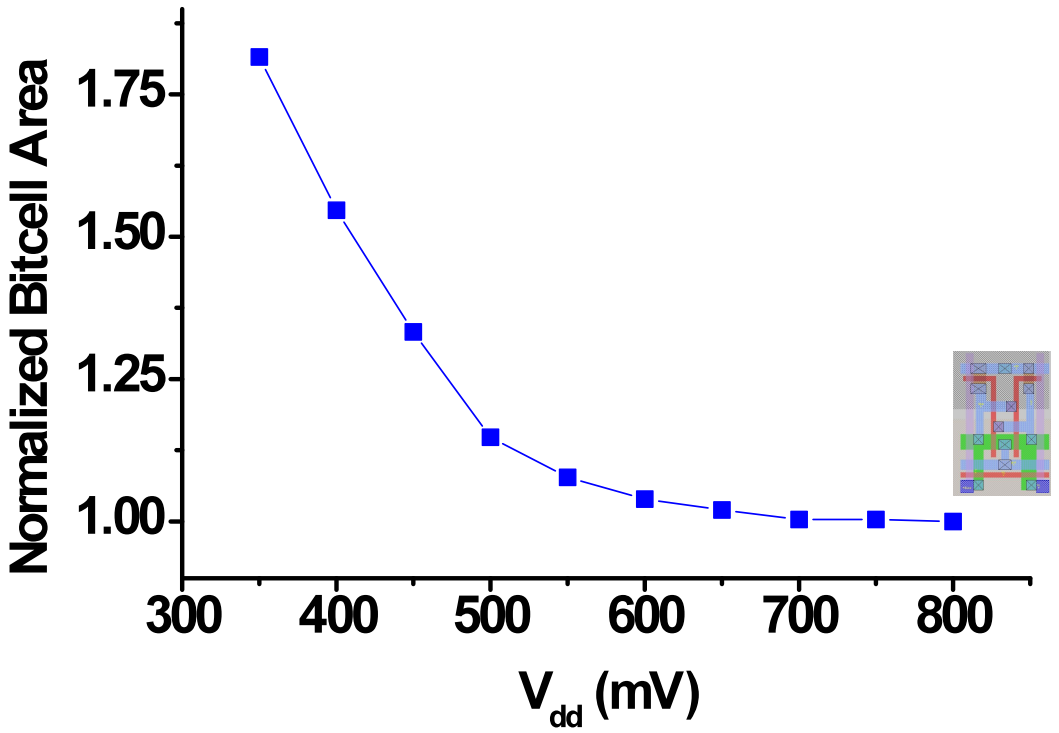
• *Dan Ernst, Shidhartha Das, Seokwoo Lee, David Blaauw, Todd Austin, Trevor Mudge, Nam Sung Kim, Krisztian Flautner, "Razor: Circuit-Level Correction of Timing Errors for Low-Power Operation", IEEE, Vol. 24, No. 6, November-December 2004, pg. 10-20.*

•*Mingoo Seok, Dongsuk Jeon, Chaitali Chakrabarti, David Blaauw, Dennis Sylvester, "A 0.27V, 30MHz, 17.7nJ/transform 1024-pt complex FFT core with super-pipelining," IEEE International Solid-State Circuits Conference (ISSCC), February 2011, to appear*

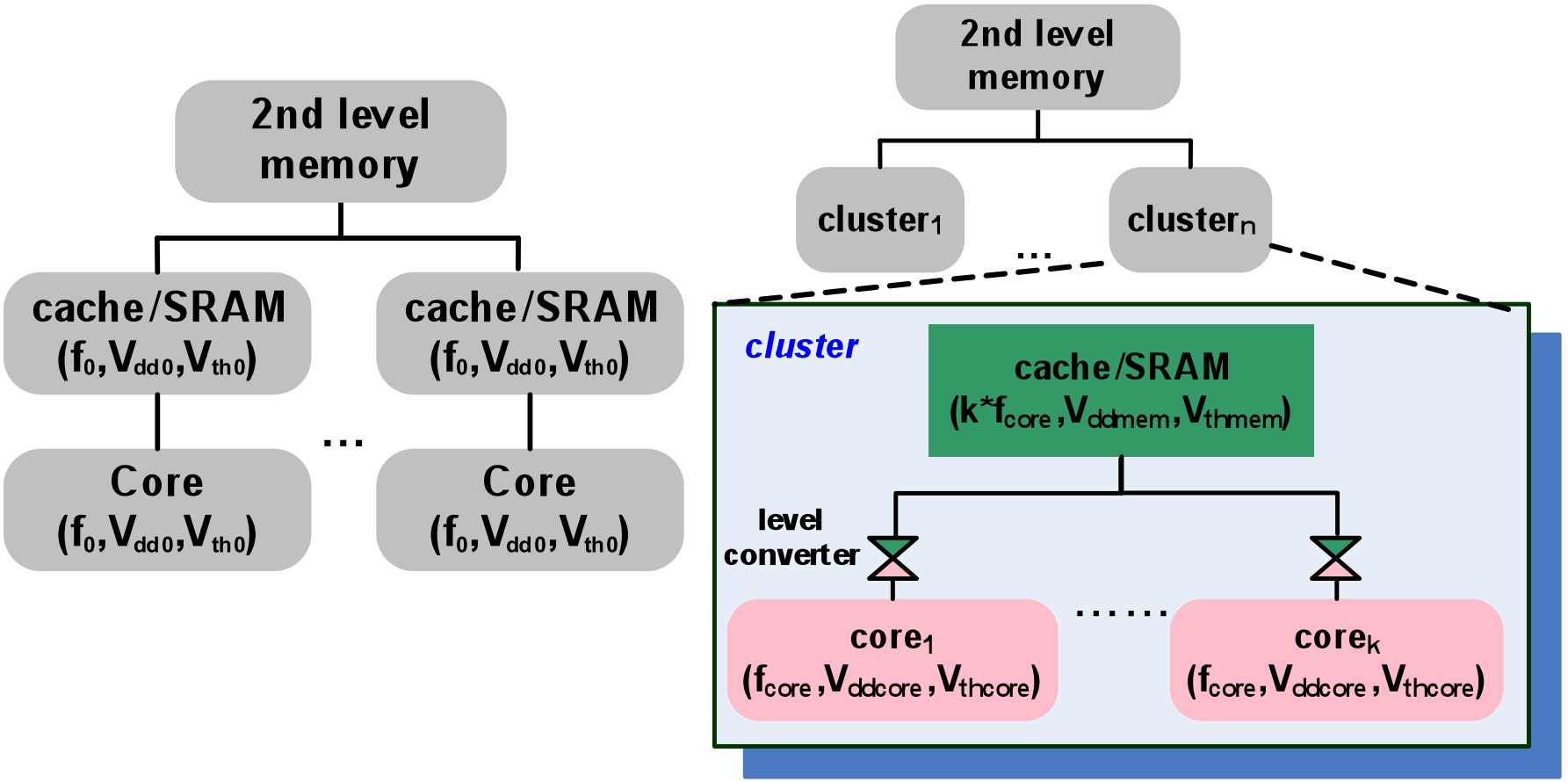
Backup

Logic vs. Memory

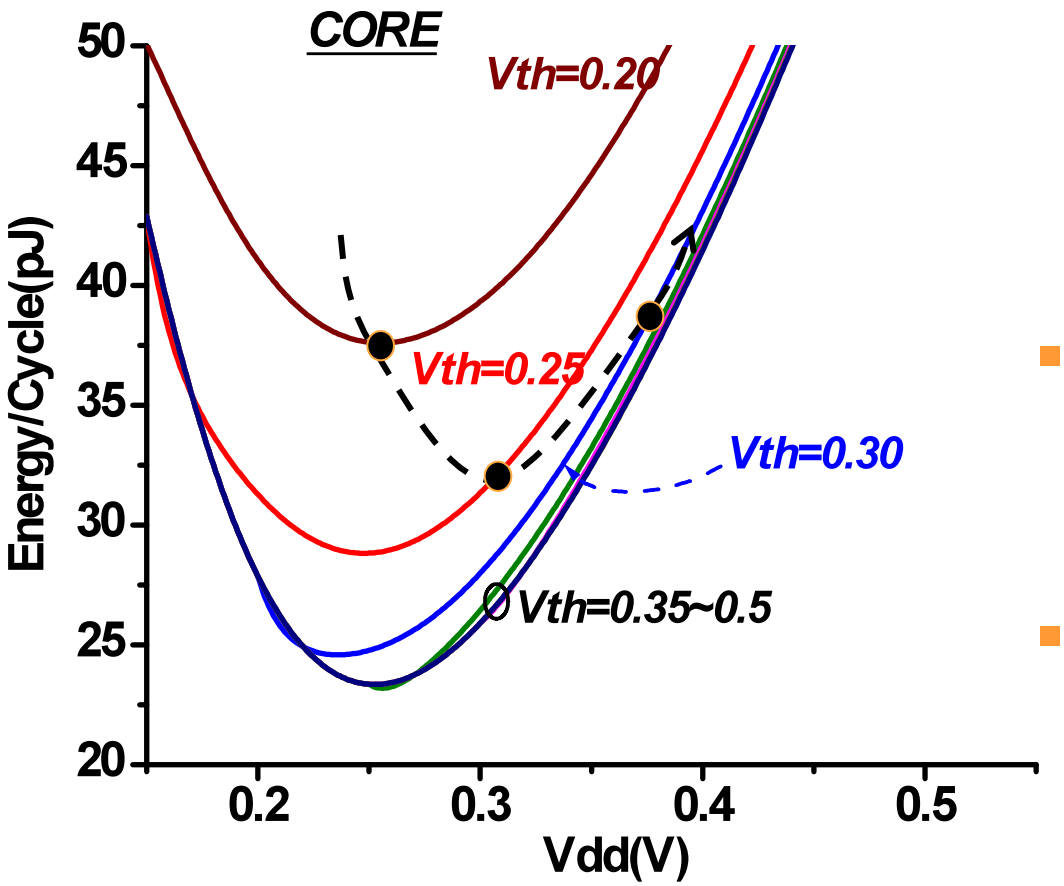
- To maintain same robustness at low voltages SRAM cell sizes needs to be increased to compensate effects of process variation
- Increased size leads to higher energy consumption, and longer interconnects



Proposed Parallel Architecture



Energy Optimal Vth Selection



- Vth is very high
  - Energy optimal Vdd is independent of Vth
  - Free performance gain without consuming more energy
- As Vth reduces
  - Circuit operates faster
  - More leakage, more energy consumption per switching
- Choose Vth
  - Body bias
  - Dopant implant