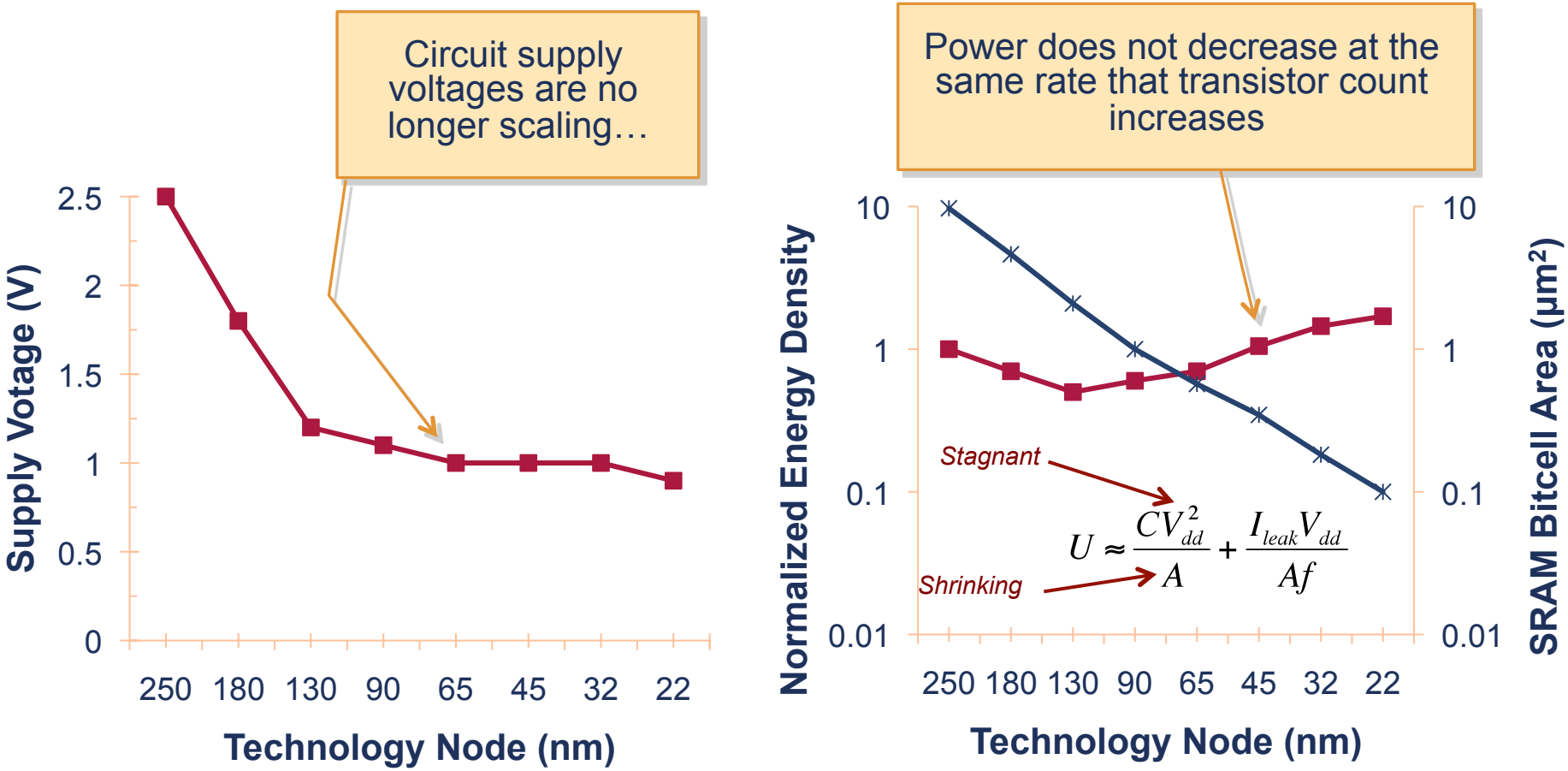
Limits of Voltage-Scaled Parallel Architectures to Combat Dark Silicon

Nathaniel Pinckney, **Ronald G. Dreslinski**, Korey Sewell,
David Fick, David Blaauw, Dennis Sylvester, and Trevor Mudge

EECS Department University of Michigan, Ann Arbor, MI.

The Problem of Power

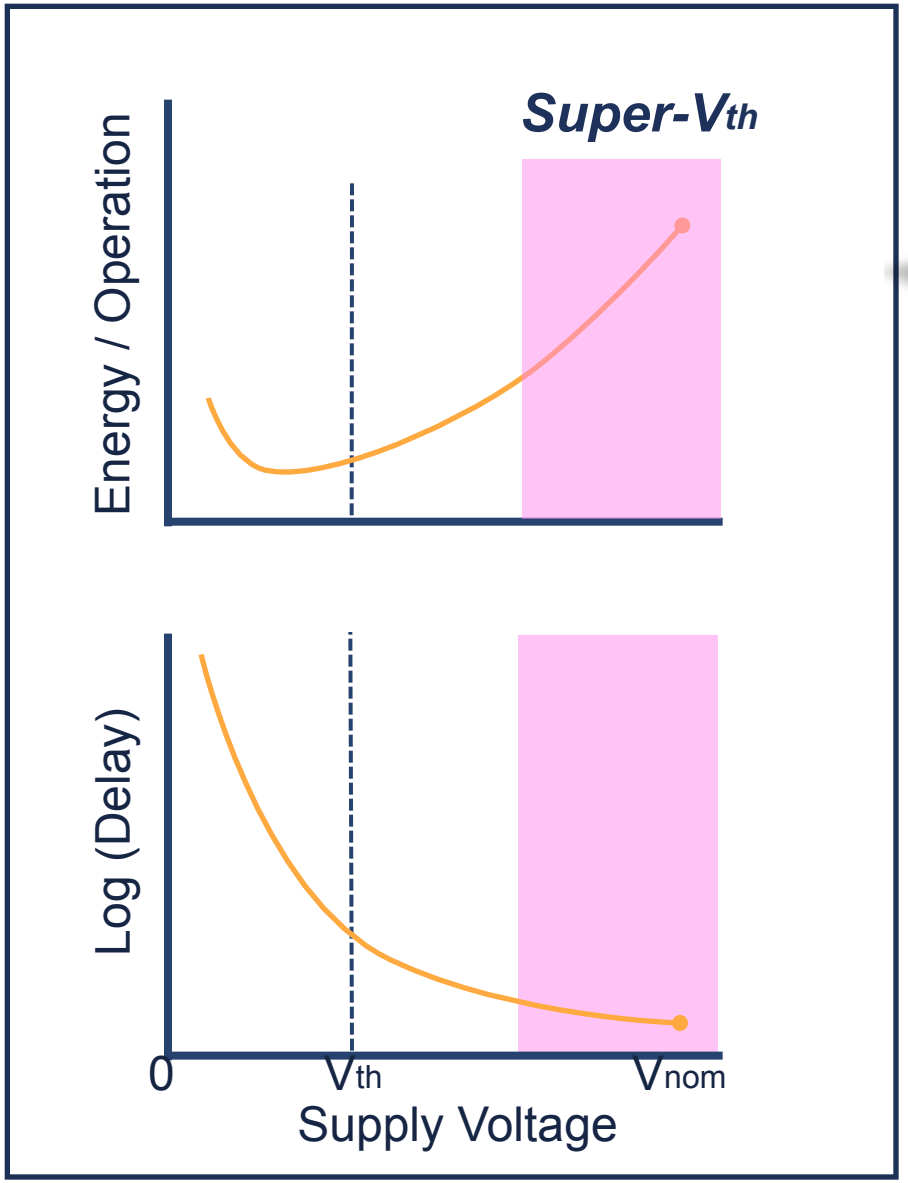


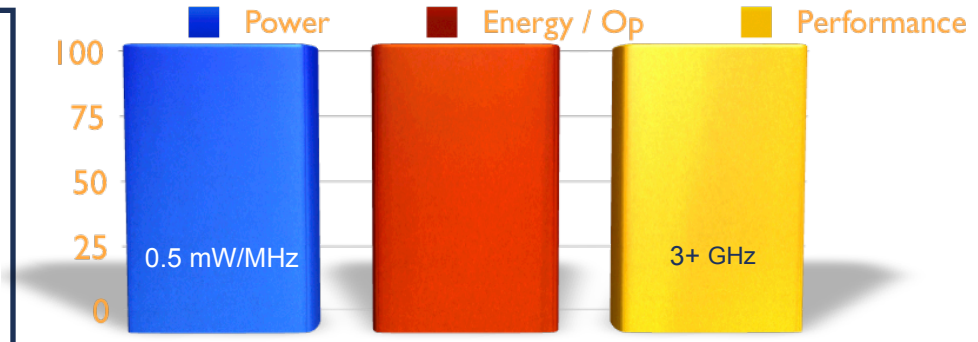
The emerging dilemma:

More and more cores can fit on a die, but we can not turn them all on!

"Dark Silicon."

Today: Super-Vth

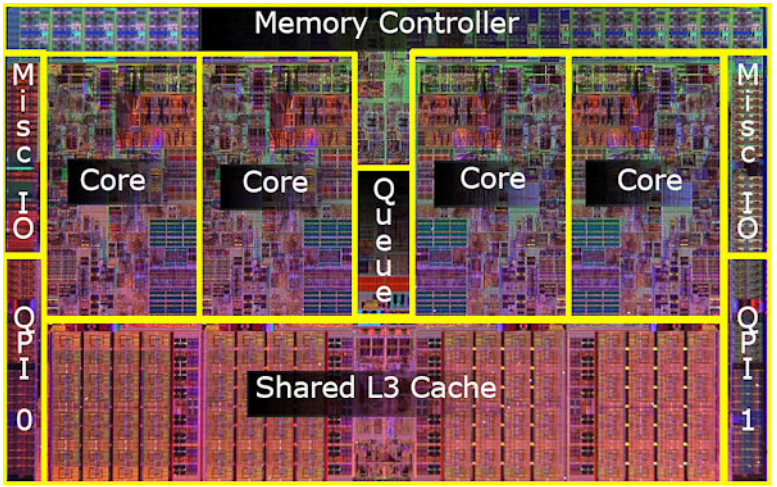




Normalized Power, Energy, & Performance

Energy per operation is the key metric for efficiency.

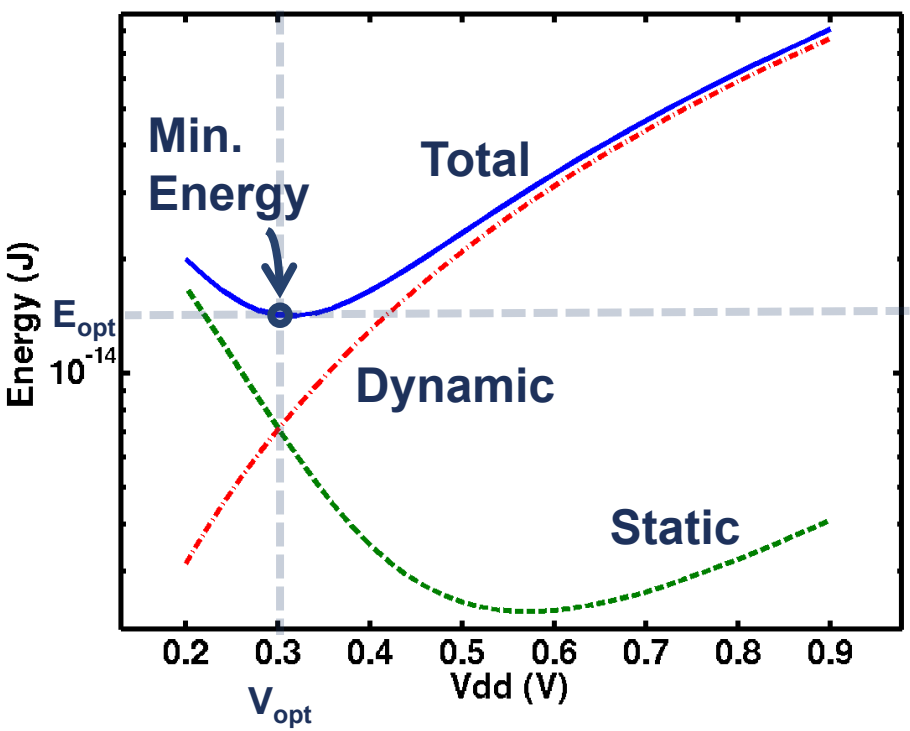
Goal: same performance, lower energy per operation

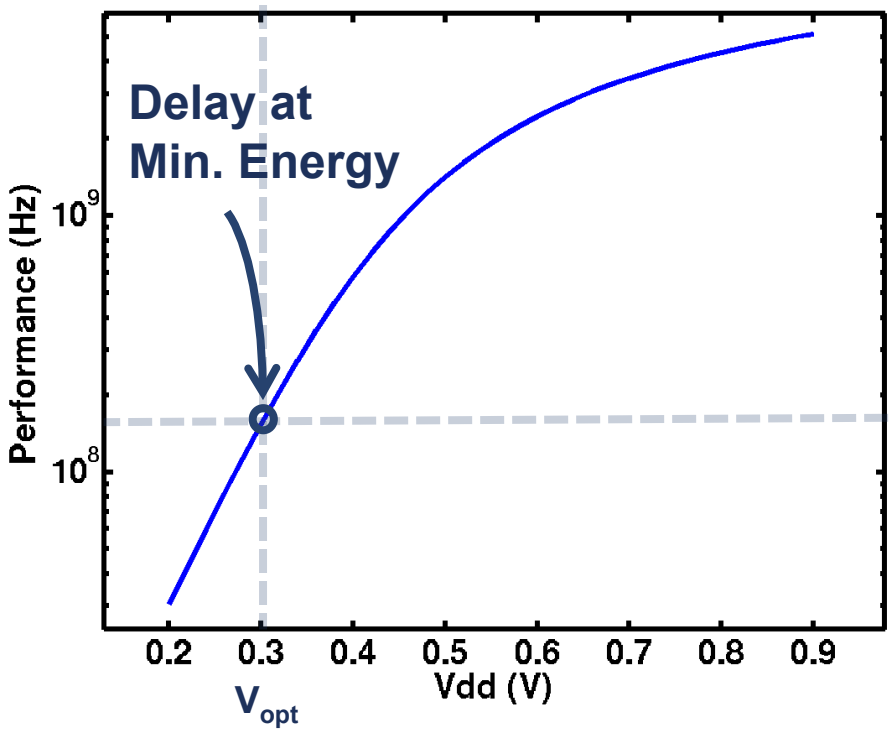


Core i7

Source: Kurd et. al., JSSC 2009.

Energy/Delay with Voltage Scaling

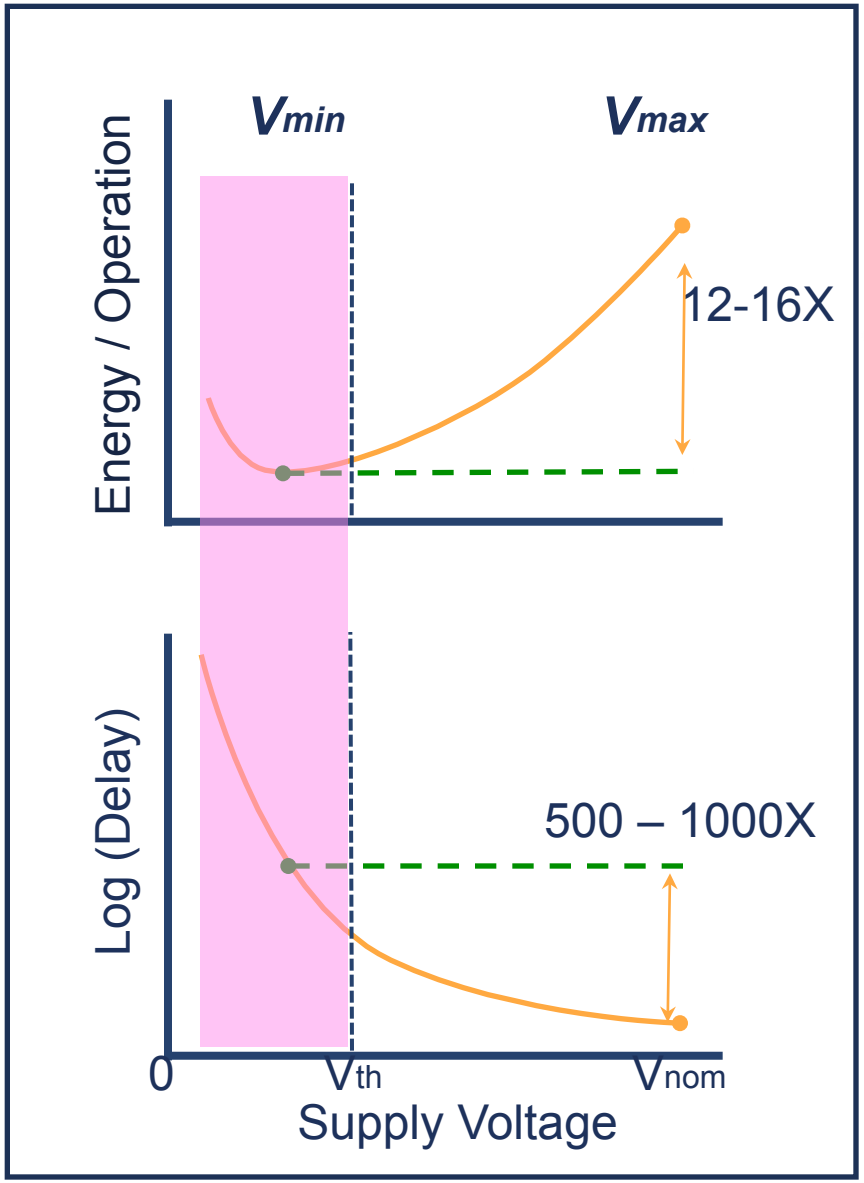


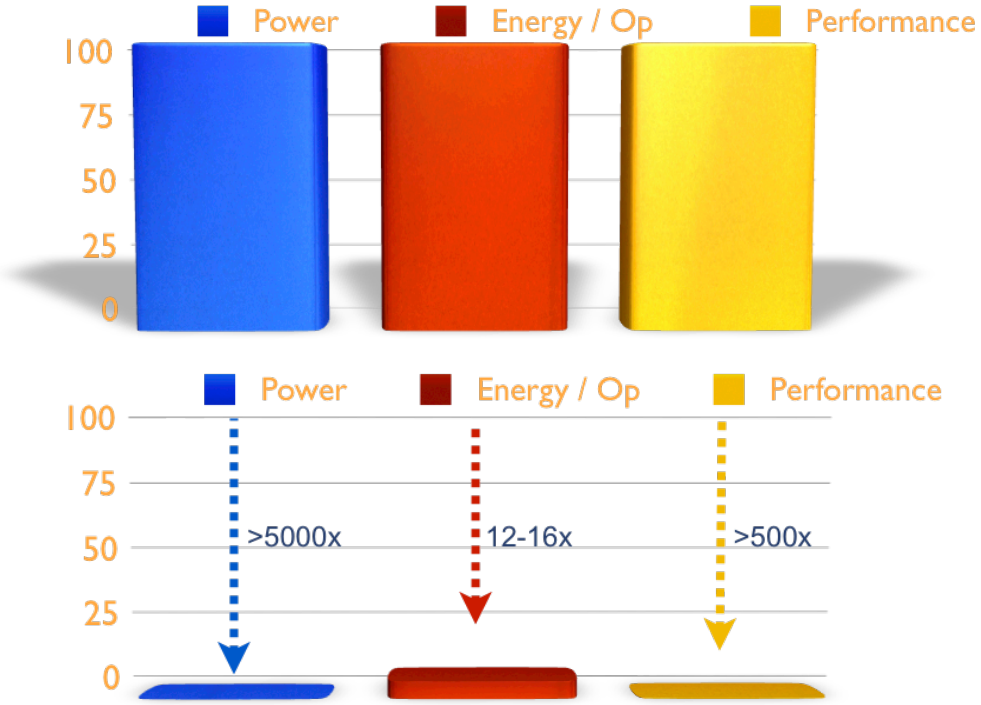


- Minimum energy reached when further voltage scaling increases leakage more than dynamic energy is reduced
  - In older technologies, $V_{opt} < V_{th}$ giving rise to *subthreshold design* [Zhai, DAC 2004]

32nm CMOS

Lowest Energy Operation





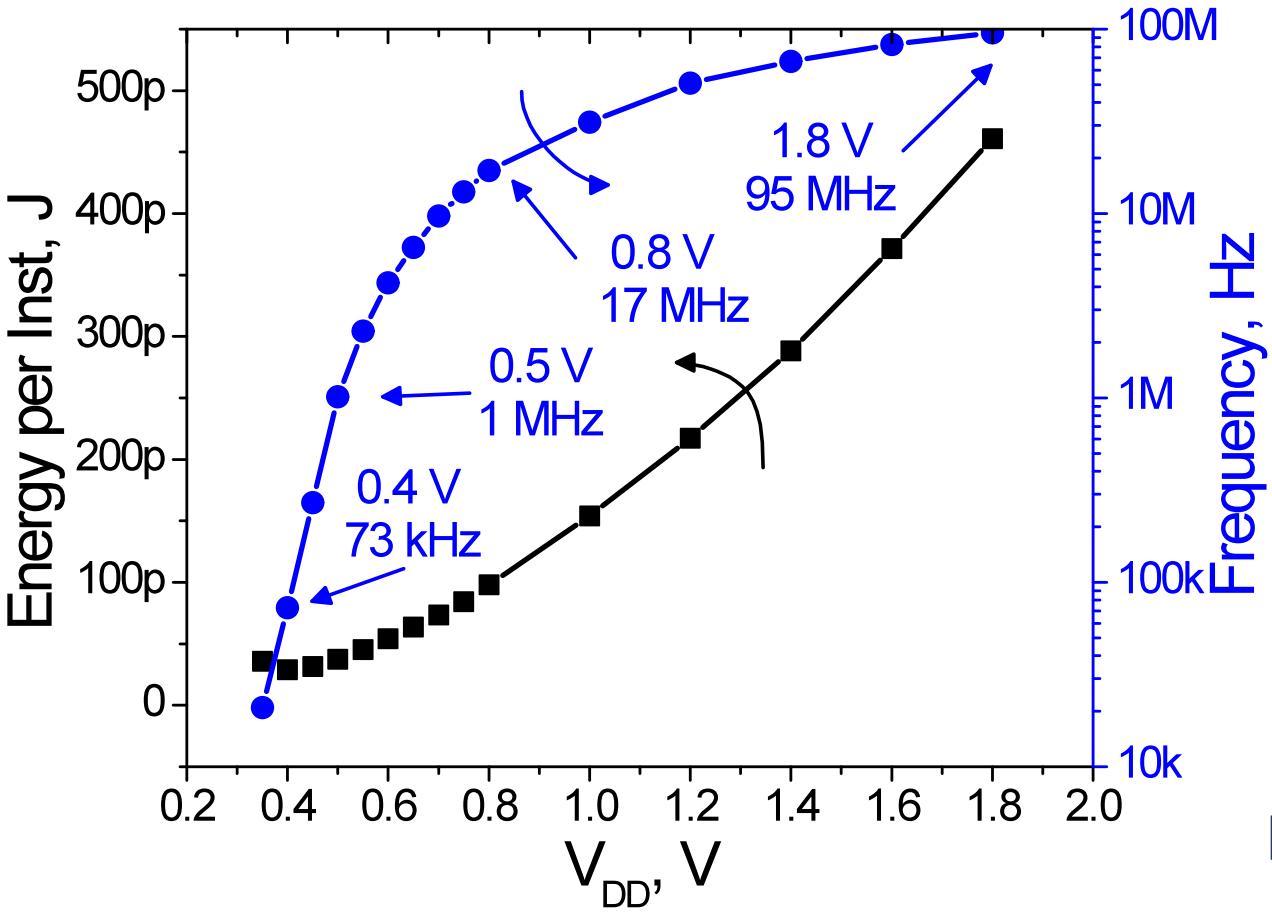
Operating in the sub-threshold gives us huge power gains at the expense of performance

Latency is very high →

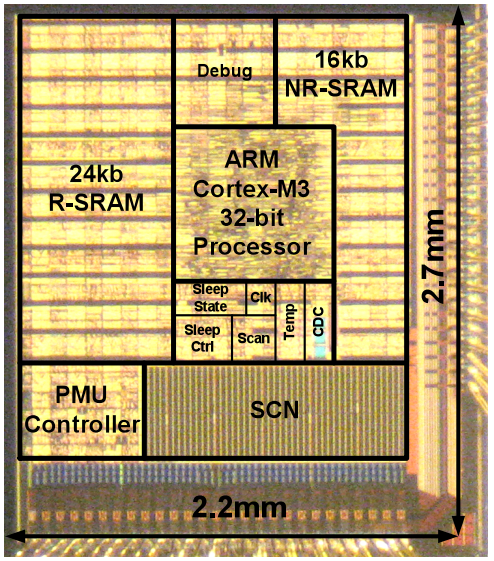
OK for sensors but not for general purpose computing!

Performance vs. Energy

- Measured Performance and Energy of ARM Cortex M3 in 180nm CMOS
  - V_{opt} = 400mV with E_{opt} = 30 pJ/instr and 73 kHz frequency
  - At 500 mV, 40 pJ/instr and 1 MHz frequency



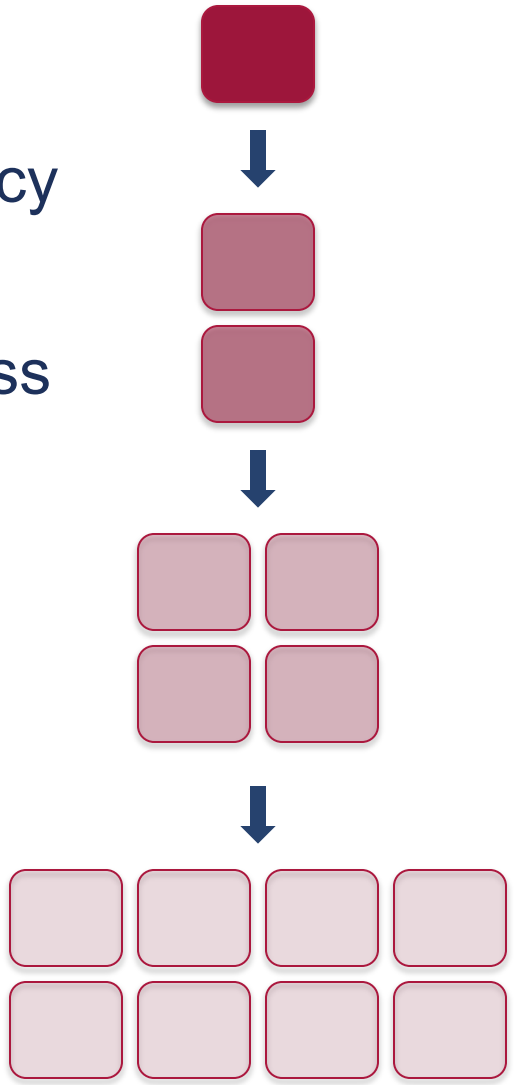
Photomicrograph of 180nm M3 core



[Seok, VLSI Symp 2008]

Parallelization – The Dim Horseman

- Start with runtime at nominal voltage
- Lower the voltage to increase energy-efficiency
- Add more cores to overcome performance loss (meet the original runtime)
- Repeat the process until reaching Vopt



How do parallelization overheads impact Vopt?

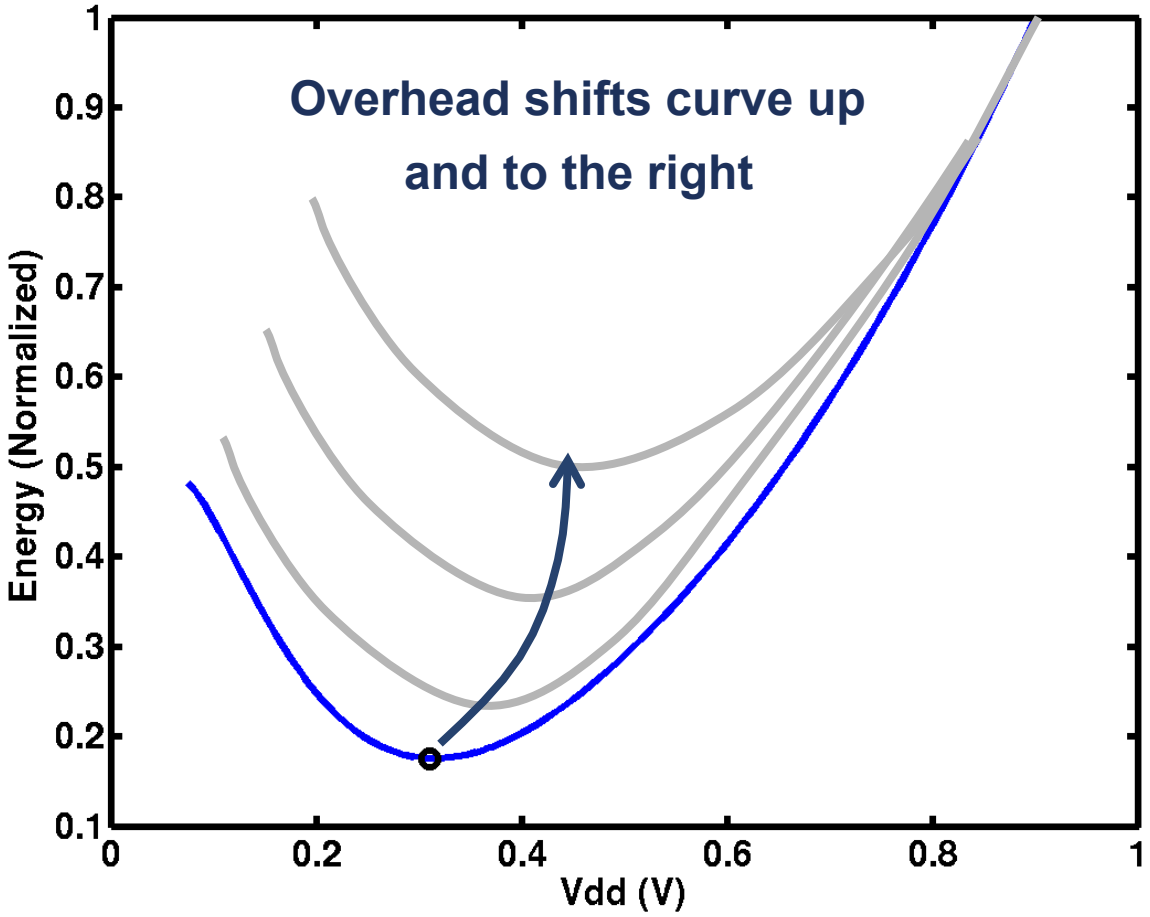
What is the number of cores @ Vopt?

How do these numbers change with technology scaling?

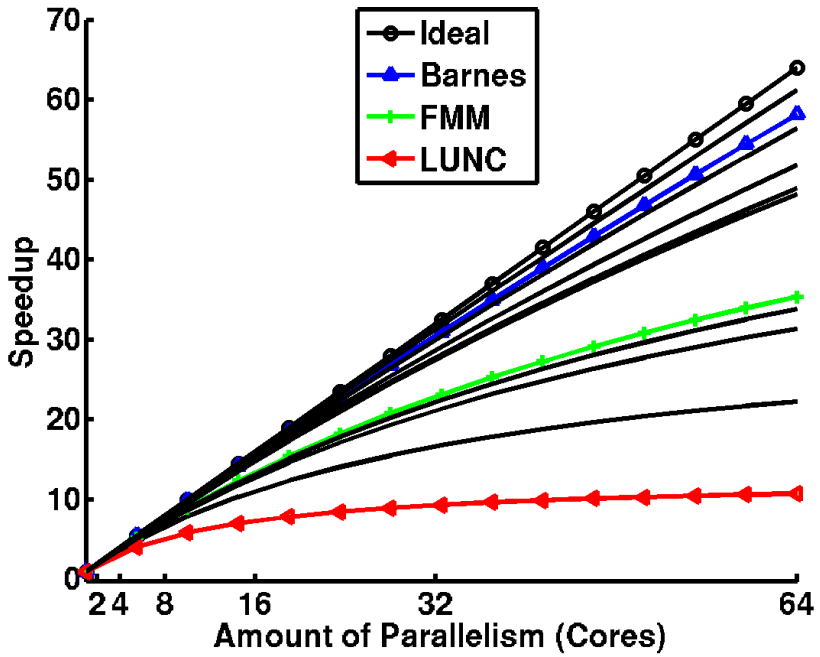
Voltage Scaling w/ Performance Recovery

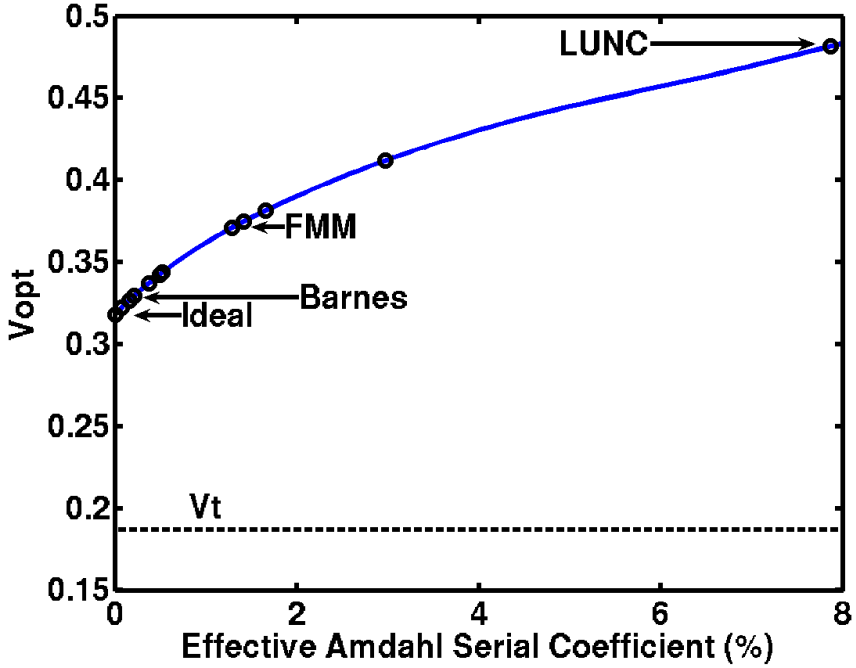
- Non-ideal parallelism → energy overhead → increase in V_{opt} and E_{opt}

- Overhead Sources
  - Non-Ideal Parallelism (Synchronization, Amdahl's law)
  - Architecture (Coherency, cache overhead)
  - Physical (Routing, clock distribution)



Overhead: Amdahl's Law

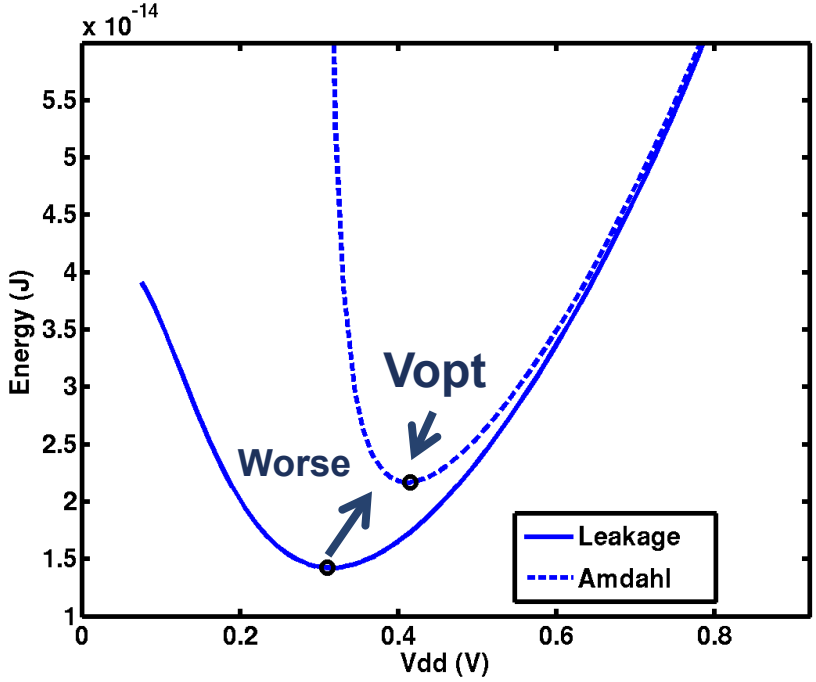


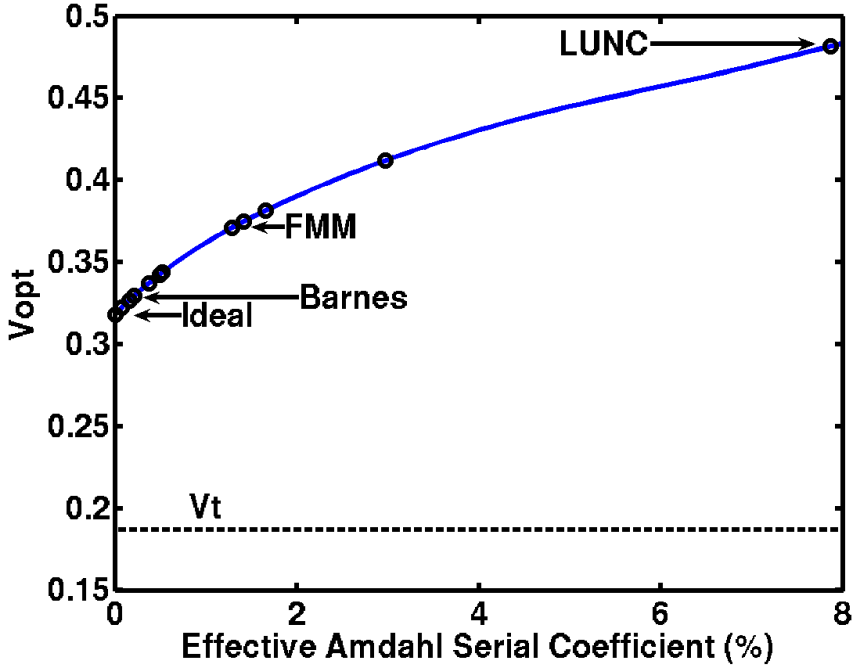


- Application-dependent scalability impacts energy optimality
- Model non-ideal speedup with Amdahl's equation
  - Obtain Amdahl serial coefficient $S(N) = 1/[(1 – P) + P/N]$
- Simulated results of SPLASH-2 benchmarks in M5
  - Ideal memory/caches

Source: Amdahl et. al., AFIPS 1967.

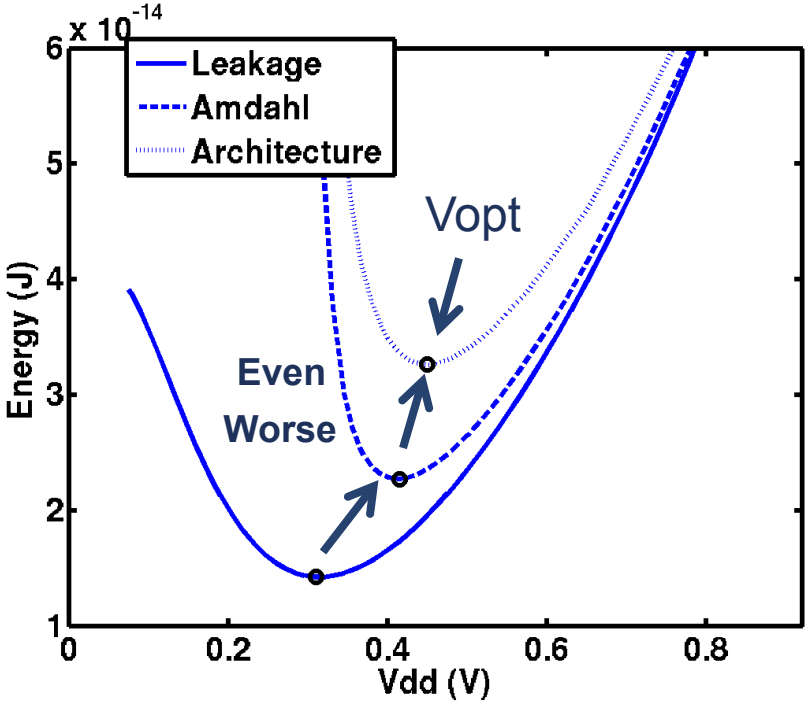
Overhead: Amdahl's Law, cont.





- Vopt increases (e.g. from 300 mV to 400 mV)
  - Parallelization overheads increase penalty for low-voltage operation
  - More energy savings if operated at a higher voltage
  - Minimum energy point increases
  - Dependent on Amdahl serial coefficient

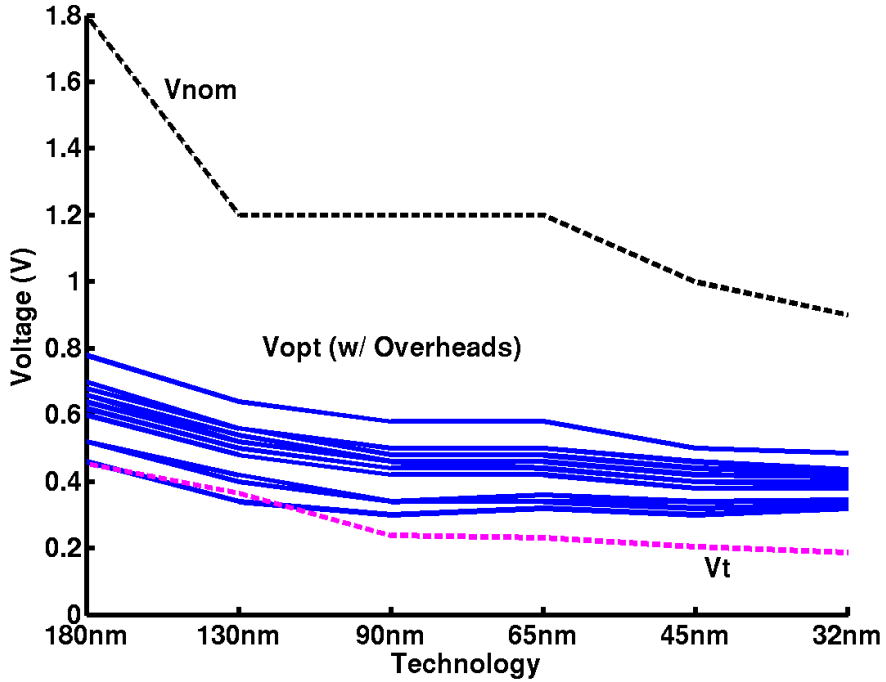
Overhead: Architectural & Physical



Feature	Description
Cores	1 to 64 one-IPC Alpha cores @ 1GHz
L1 Caches	32 kB, 1 cycle latency, 4-way associative, 64-byte line size
L2 Caches	Shared 1MB divided evenly between cores, 10 cycle latency, 8-way associative, 64-byte line size
Interconnect	2-GHz Routers, 128-bit, 2-stage routers, 50 cycle-access to main memory

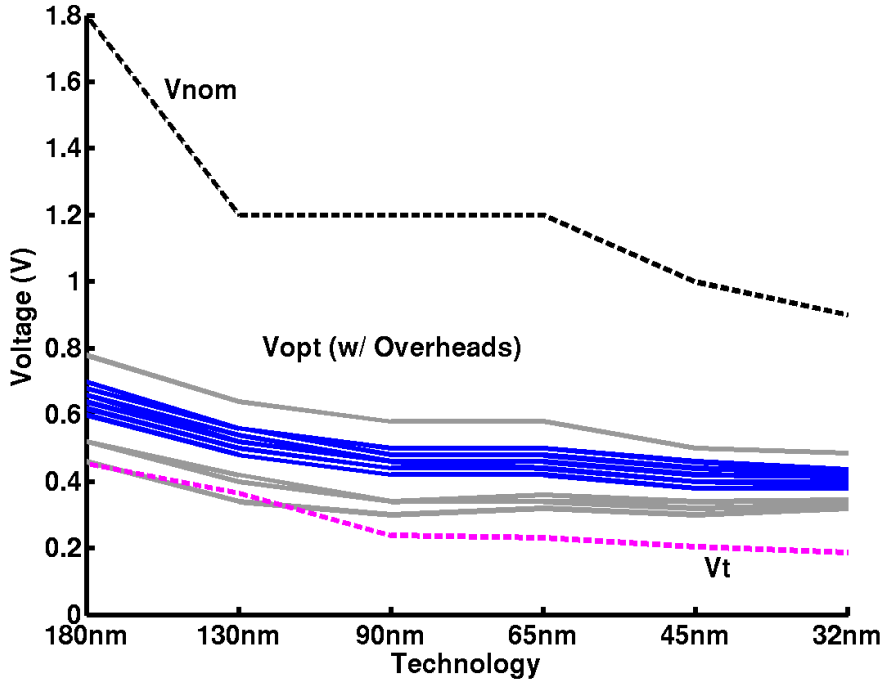
- Architectural & physical overheads include
  - Coherence protocols, cache and memory hierarchy
  - Physical routing, interconnect networks
- SPLASH-2 benchmarks simulated on gem5 architecture
- Further reduces achievable energy-efficiency, increases Vopt

Optimal Operating Voltage Across Technology



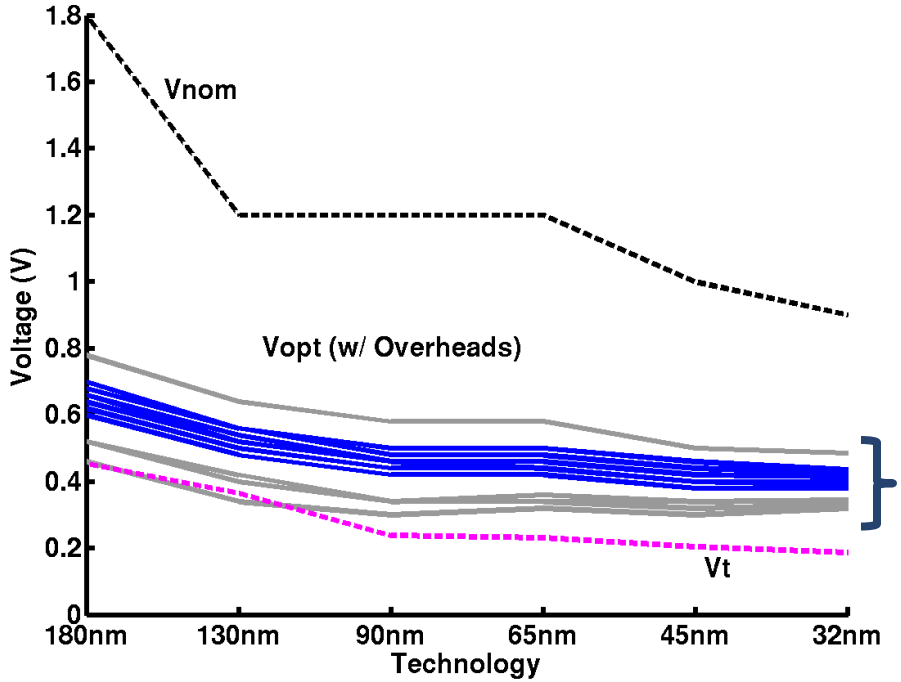
- Tested on 6 commercial technologies from 180 – 32nm
- NTC region is 200 mV to 400 mV above Vt in recent techs
  - Approximately tracks with Vt
  - Optimal energy for single, parallelized task – best way to use extra cores on CMP

Optimal Operating Voltage Across Technology



- Tested on 6 commercial technologies from 180 – 32nm
- NTC region is 200 mV to 400 mV above Vt in recent techs
  - Approximately tracks with Vt
  - Optimal energy for single, parallelized task – best way to use extra cores on CMP

Optimal Operating Voltage Across Technology

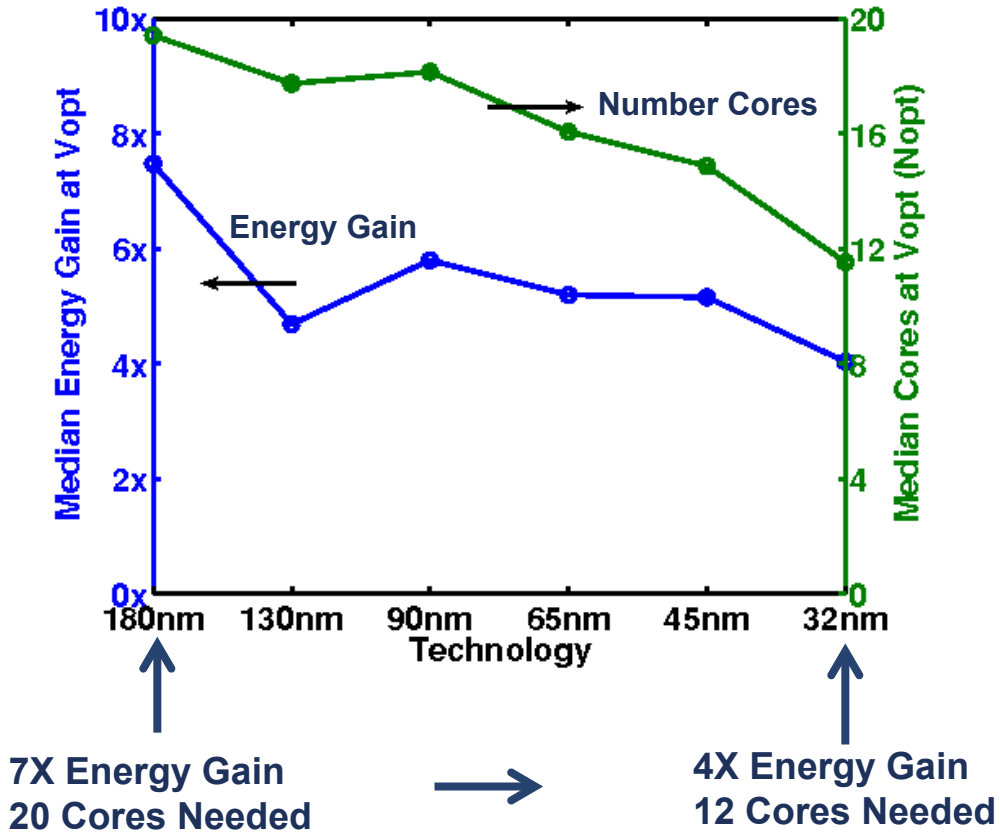


NTC Region 200 mV to 400 mV above Vt and tracks.

Zhai, ISLPED 2007

- Tested on 6 commercial technologies from 180 – 32nm
- NTC region is 200 mV to 400 mV above Vt in recent techs
  - Approximately tracks with Vt
  - Optimal energy for single, parallelized task – best way to use extra cores on CMP

Gain from V_{nom} to V_{opt}

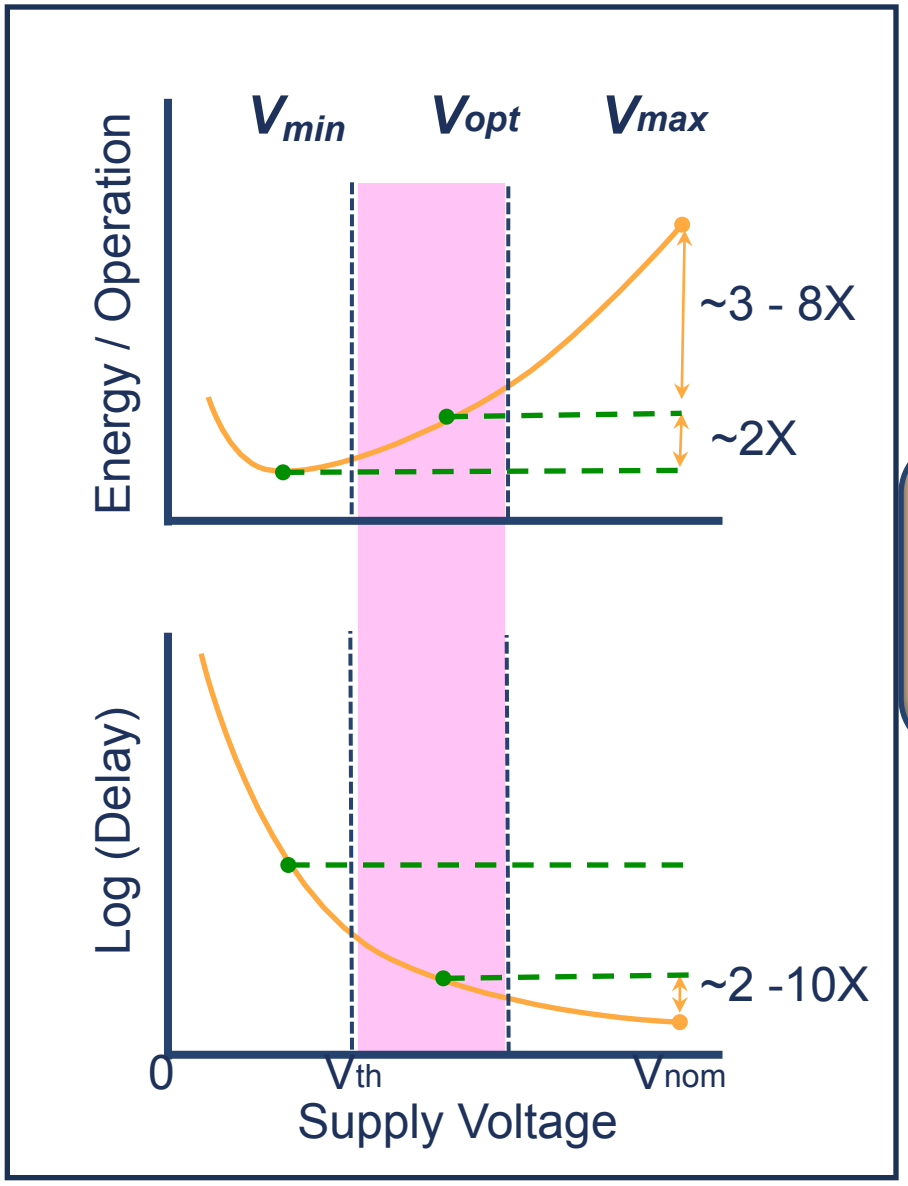


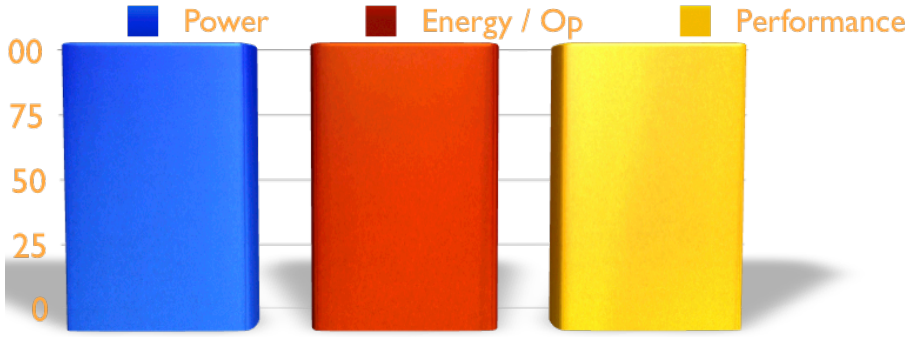
Maximum of 12 cores required for one task in 32nm.

Energy gain of 4x from operating at NTC.

- Energy gain over nominal supply voltage
  - Cores used to parallelize decreased from 20 to 12 cores
  - Energy gains decrease from 7.5x to 4x in 32nm

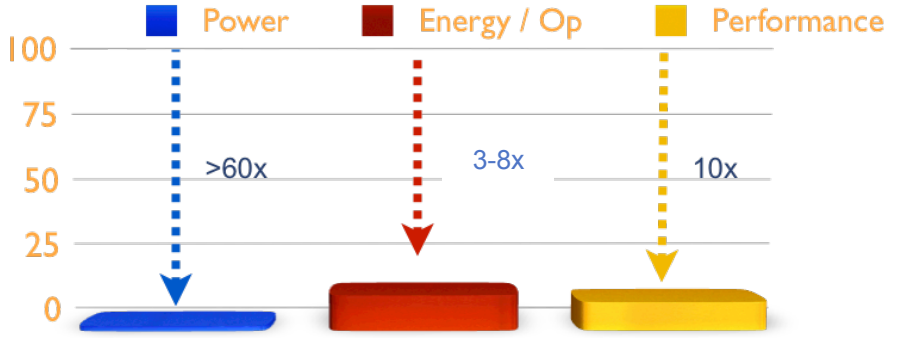
Near-Threshold Computing



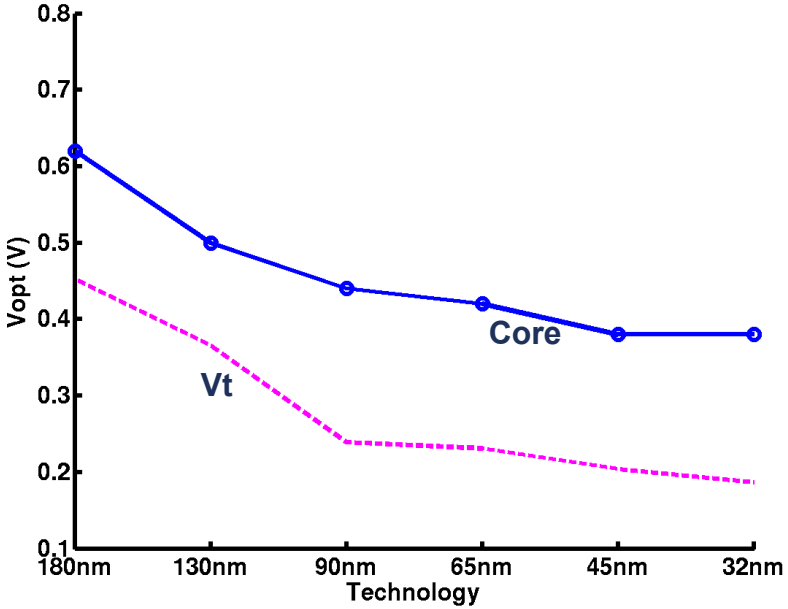


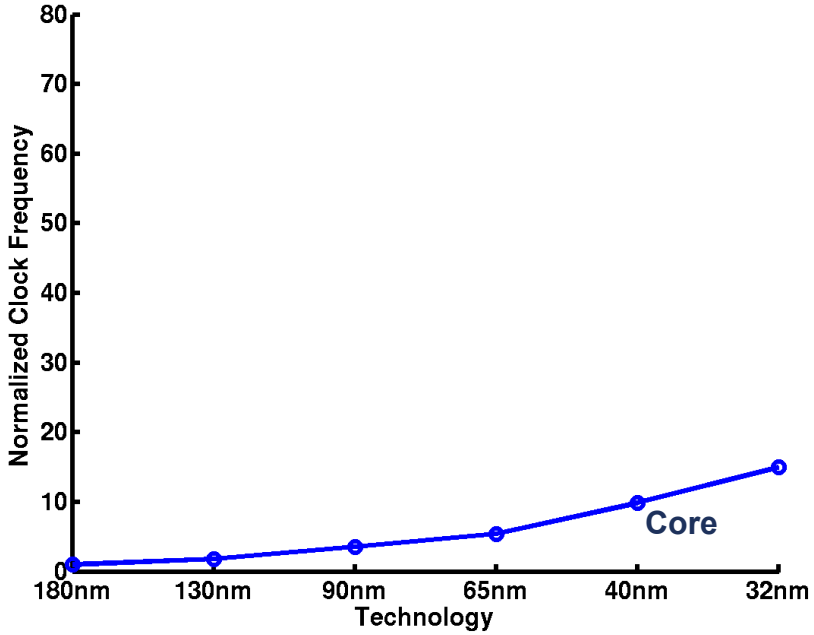
Near-Threshold Computing (NTC):

- *10 - 50X power reduction*
- *3 - 8X energy reduction*
- *Minimize energy while considering latency*



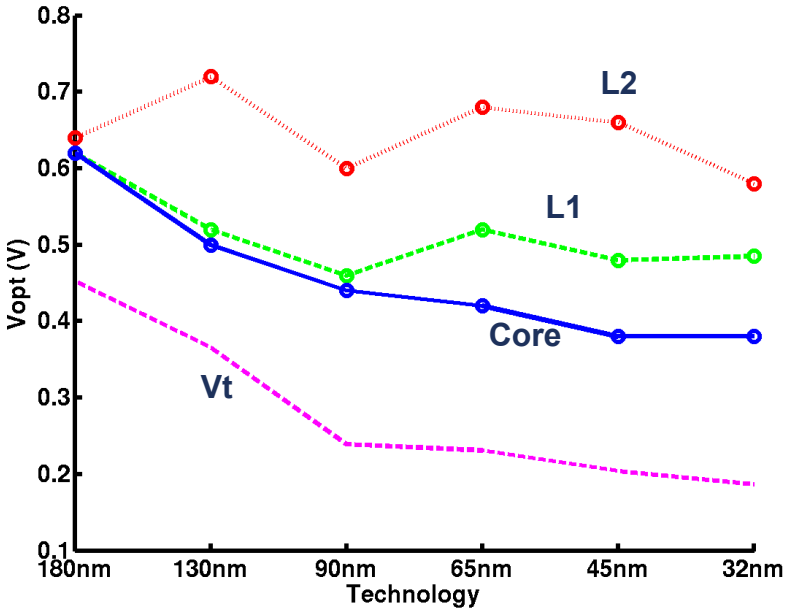
Architectural Impact of NTC

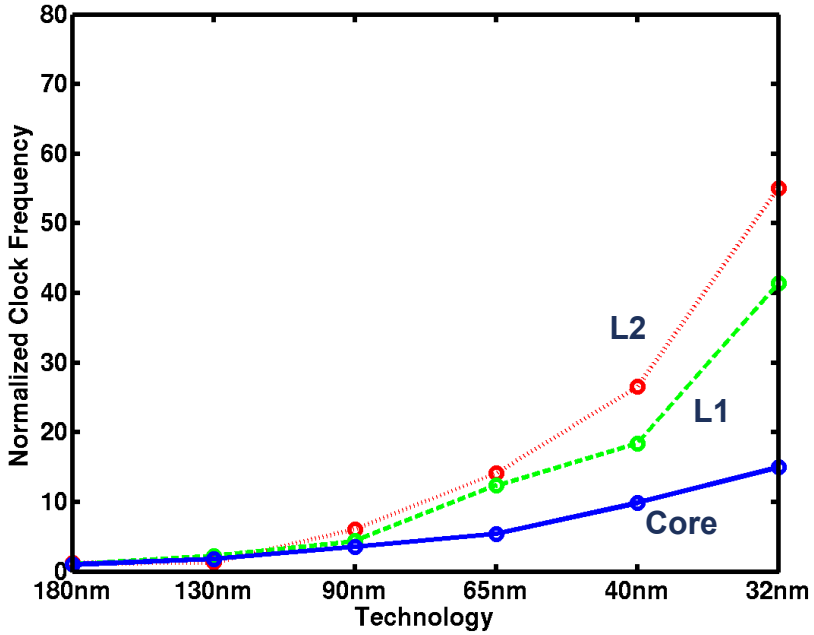




- Caches have higher Vopt and operating frequency
- Smaller activity rate when compared to core logic
- Leakage larger proportion of total power in caches
- Higher Vt can offset this somewhat

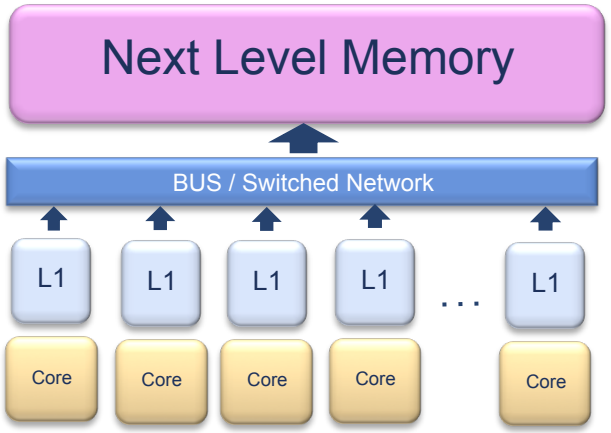
Architectural Impact of NTC

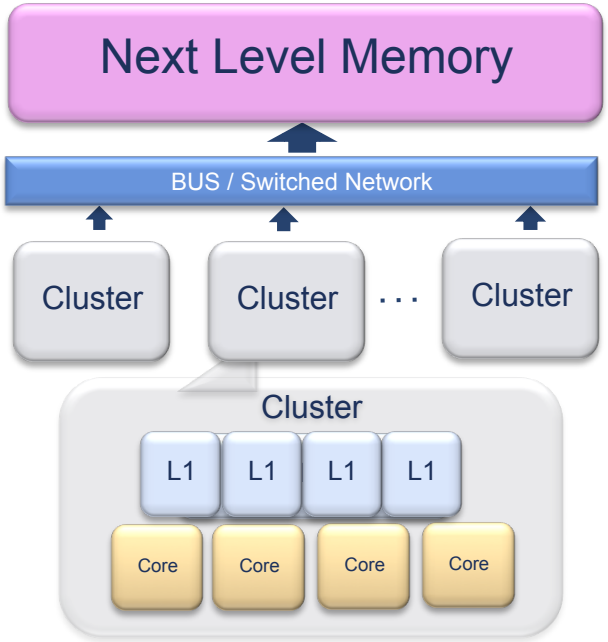




- Caches have higher Vopt and operating frequency
- Smaller activity rate when compared to core logic
- Leakage larger proportion of total power in caches
- Higher Vt can offset this somewhat

New NTC Architectures

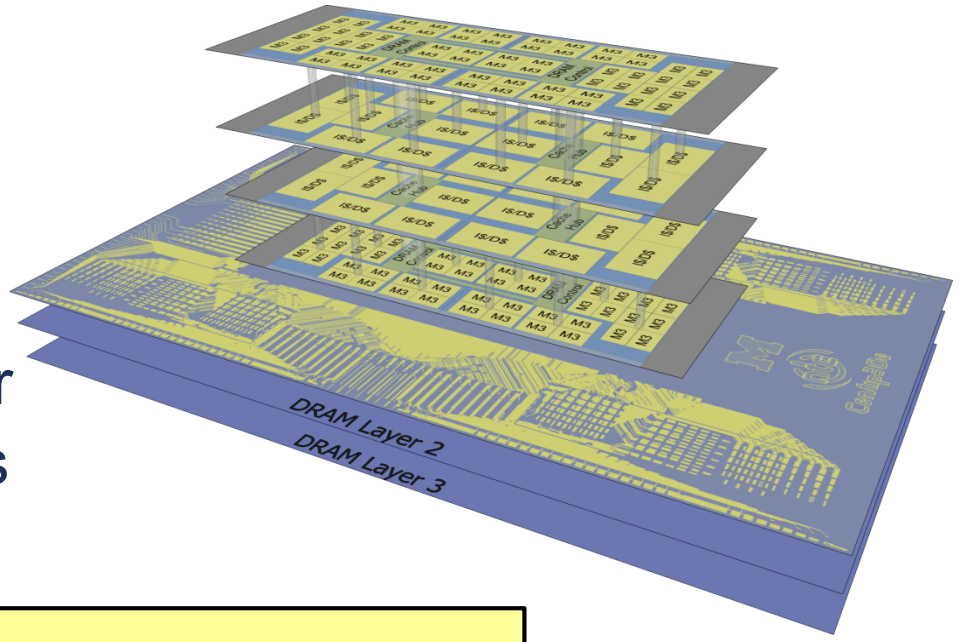


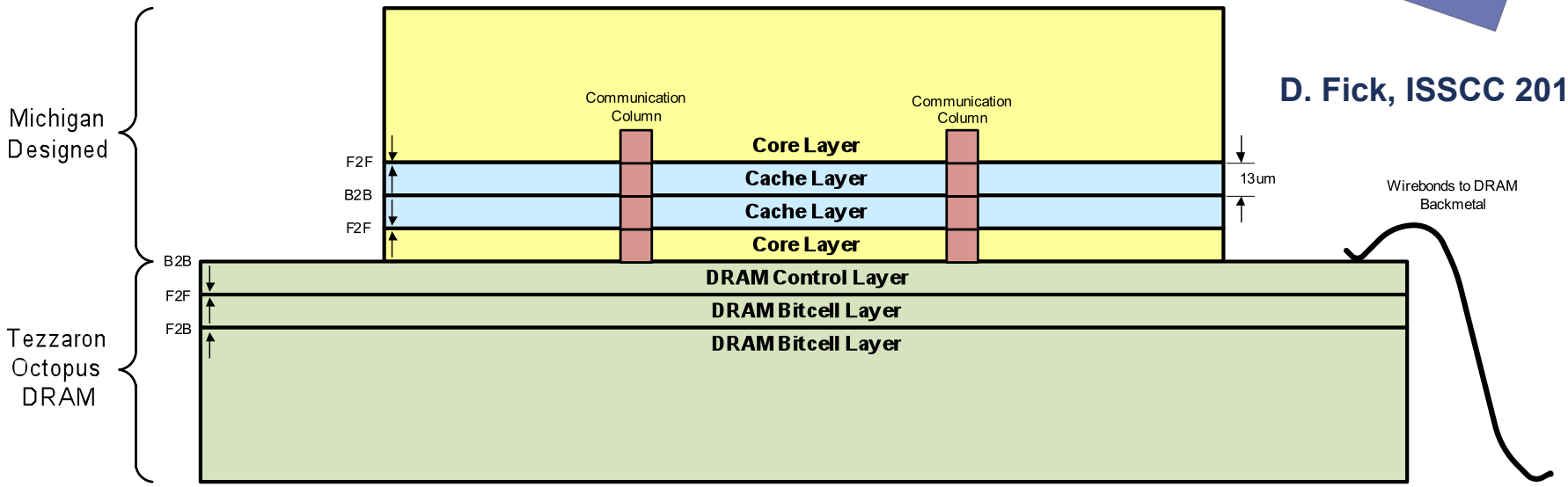


- SRAM runs at a higher V_{DD} than cores with energy efficiency
- Caches / Core Inversion
  - Cache runs faster than core
- Clustered architectures
  - Multiple Cores share L1
  - L1 satisfy all core requests in 1-cycle
  - Cores see view of private single cycle L1
- **Advantages:**
  - Clustered sharing
  - Less coherence/snoop traffic
  - Virtualizes the space (looks larger)
- **Drawbacks:**
  - Core conflicts evicting L1 data
  - Additional Bus/Interconnect from cores to L1
  - Single larger cache increases energy/access

Centip3De Design

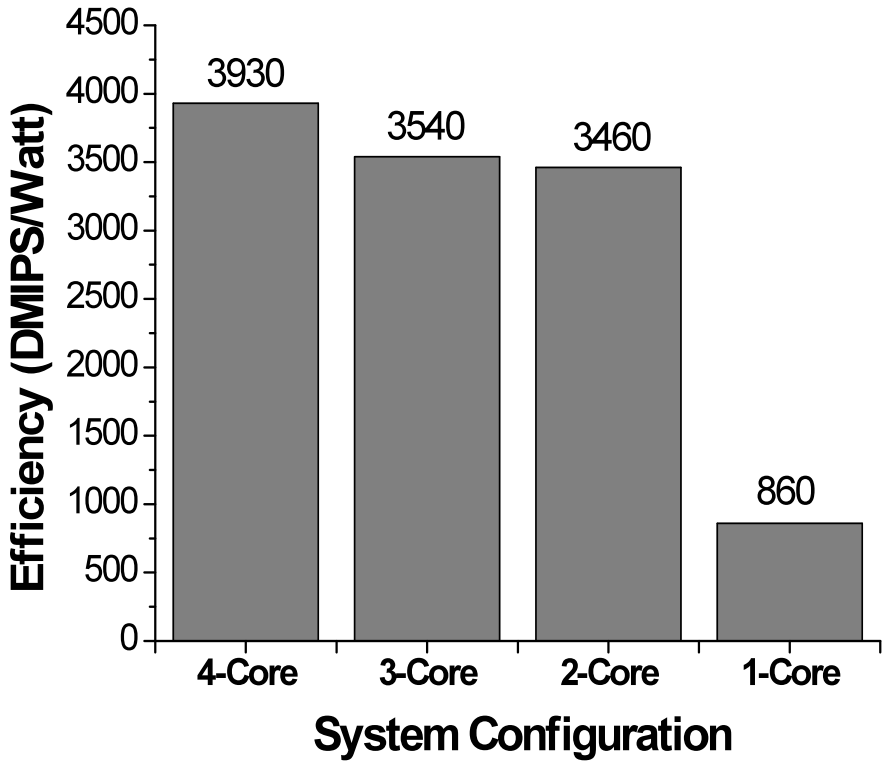
- 128 ARM Cortex-M3 cores
- 7-layer stack architecture
- TSV interconnects
- Physical Specs
  - 130nm technology
  - 12.66 mm x 5 mm per layer
  - 92M device across 4 layers

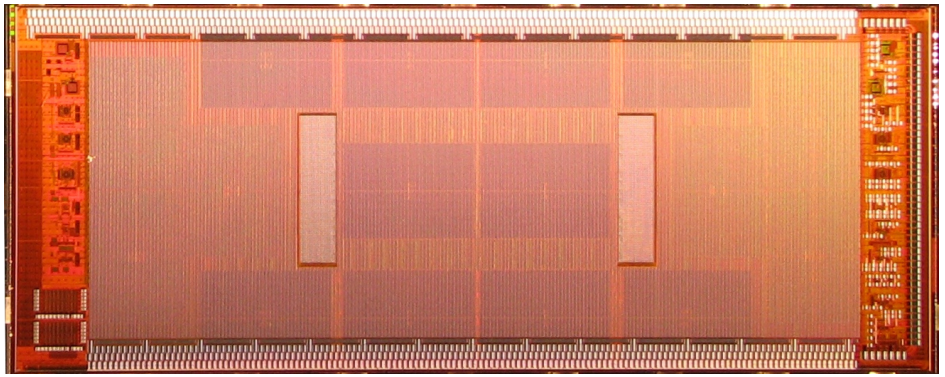




D. Fick, ISSCC 2012

NTC Example: Centip3De Results





- **Measured Results (DMIPS/W) Centip3De – 3,930 (130nm)**
- **Estimated Results: Centip3De – 18,500 (45nm)**

- Many different voltage configurations
  - Cores operate from 0.65 V (10 MHz) to 1.15 V (80 MHz)
  - L1 caches operate from 0.8 V (80 MHz) to 1.65 V (160 MHz)
  - Max of four cores/cache
- 4.5x energy efficiency gain from NT operation

Source: D. Fick, et. al., ISSCC 2012

Conclusions

- **Past**: Area-Constrained Computing
  - Traditional dynamic voltage and frequency scaling
  - Lowers voltage to save energy, but only when idling
  - With latency-sensitive workloads increase voltage to full

- **Today**: Power-Constrained Computing – Dark Silicon
  - Computation is limited by power
  - Glut of Cores: not possible to use all cores simultaneously
  - By improving energy efficiency, more cores can be enabled
  - High energy-efficiency leads to improved computational performance

- **Approach**: Near-Threshold Computing
  - Reduce voltage to increase energy-efficiency
  - Maintain throughput and latency by parallelizing
  - Optimal energy achieves 4x improvement while parallelizing across ~12 cores on average

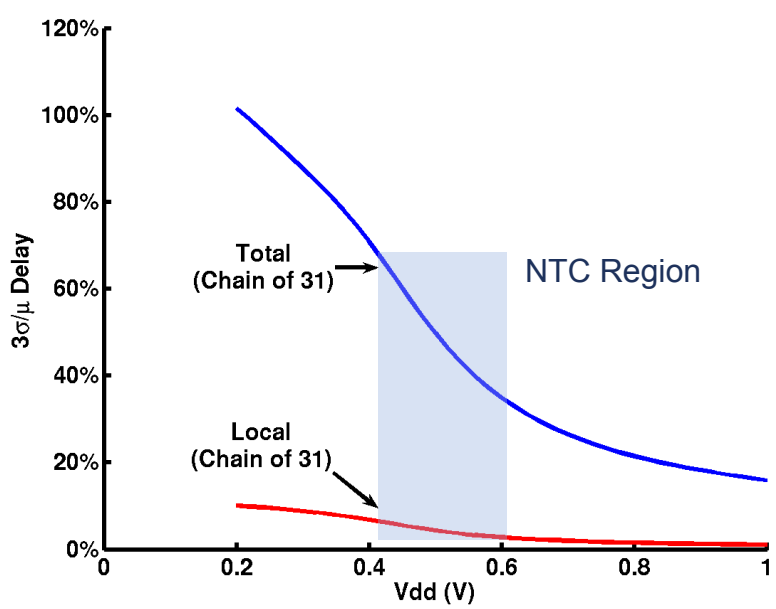
Additional Information Slides

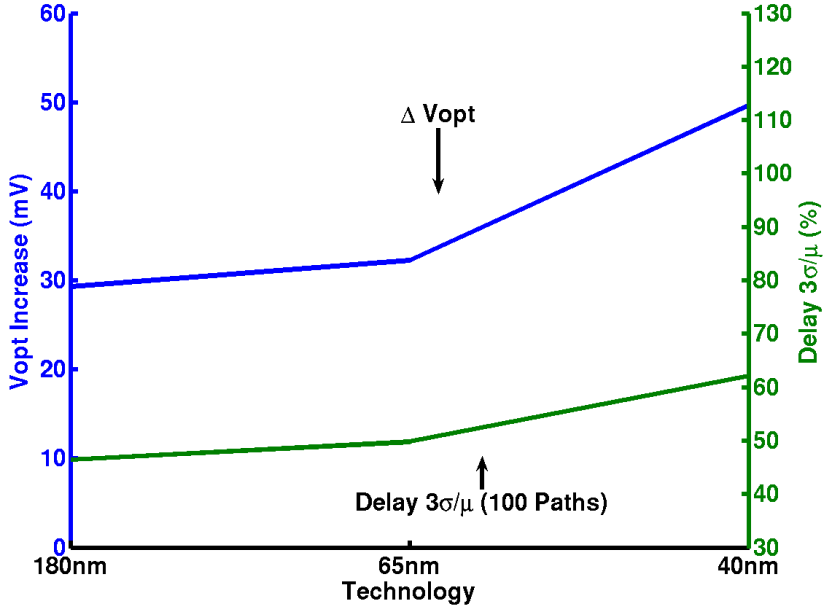
Results Per Benchmark

	180nm	130nm	90nm	65nm	40nm	32nm
Bar	12.7× (71)	7.9× (55)	10.0× (69)	7.8× (43)	5.7× (44)	5.2× (19)
Cho	6.1× (13)	3.9× (10)	4.6× (14)	4.3× (12)	3.2× (11)	3.5× (9)
Fft	13.2× (68)	8.3× (82)	10.4× (67)	8.0× (42)	5.8× (43)	5.2× (23)
Fmm	7.7× (21)	4.8× (20)	6.0× (19)	5.3× (16)	3.9× (16)	4.1× (12)
Luc	8.6× (26)	5.3× (25)	6.7× (24)	5.9× (18)	4.2× (21)	4.4× (13)
Lun	4.1× (8)	2.7× (6)	3.1× (7)	3.0× (6)	2.3× (6)	2.6× (6)
Occ	6.9× (14)	4.3× (16)	5.3× (17)	4.8× (14)	3.5× (13)	3.8× (9)
Ocn	6.8× (15)	4.2× (12)	5.2× (17)	4.7× (14)	3.5× (14)	3.8× (9)
Rad	5.7× (11)	3.6× (12)	4.3× (12)	4.0× (10)	3.0× (9)	3.4× (8)
Ray	7.3× (17)	4.6× (15)	5.6× (16)	5.0× (17)	3.7× (13)	4.0× (11)
Wan	12.6× (71)	7.8× (56)	9.9× (70)	7.8× (32)	5.7× (45)	5.2× (19)
Was	18× (186)	11.3× (250)	13.0× (121)	9.0× (51)	6.8× (79)	5.5× (25)

Table 1: Energy gain and optimal number of cores N_{opt} (in parenthesis) across SPLASH-2 benchmarks and technologies when including the three voltage scaling overheads.

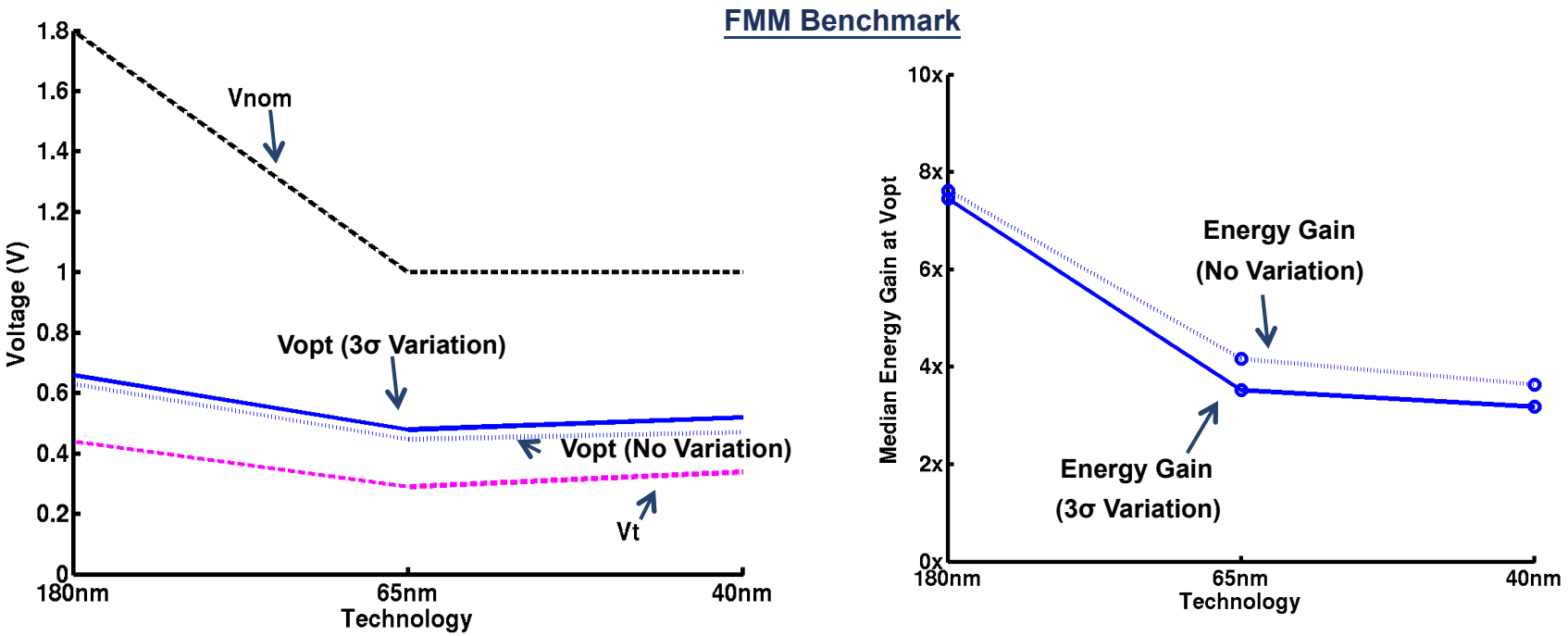
Impact of Variation in 40nm





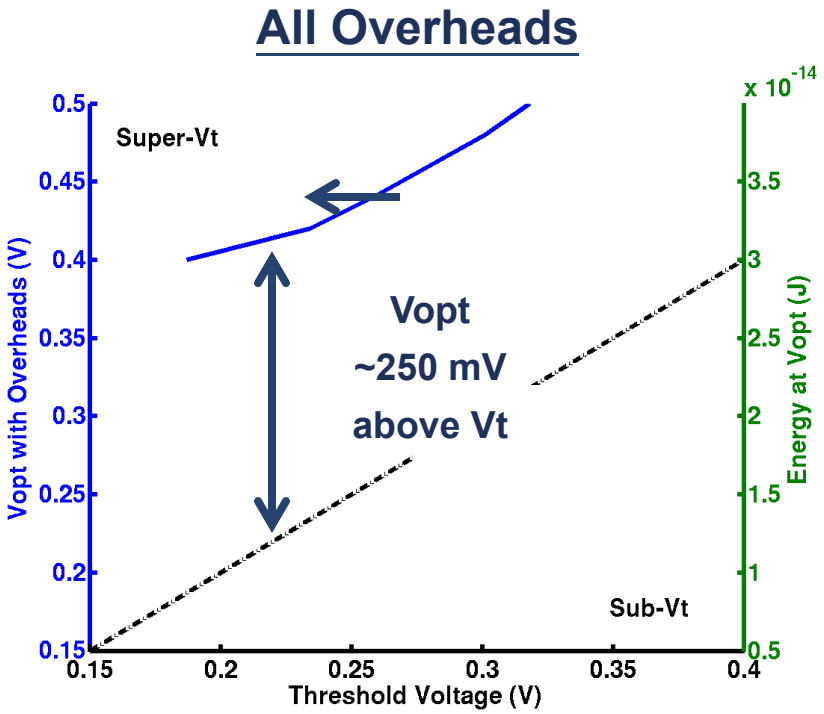
- Long logic chains average out variation
  - Local variation is ~30% of total delay variation for a single gate in NTC
  - Decreases to 10% of total variation for a logic chain of 31 gates
- Multiple paths
  - Higher mean delay, but tighter variation. Increase Vopt 30 to 60 mV.

Vopt Including Process Variations



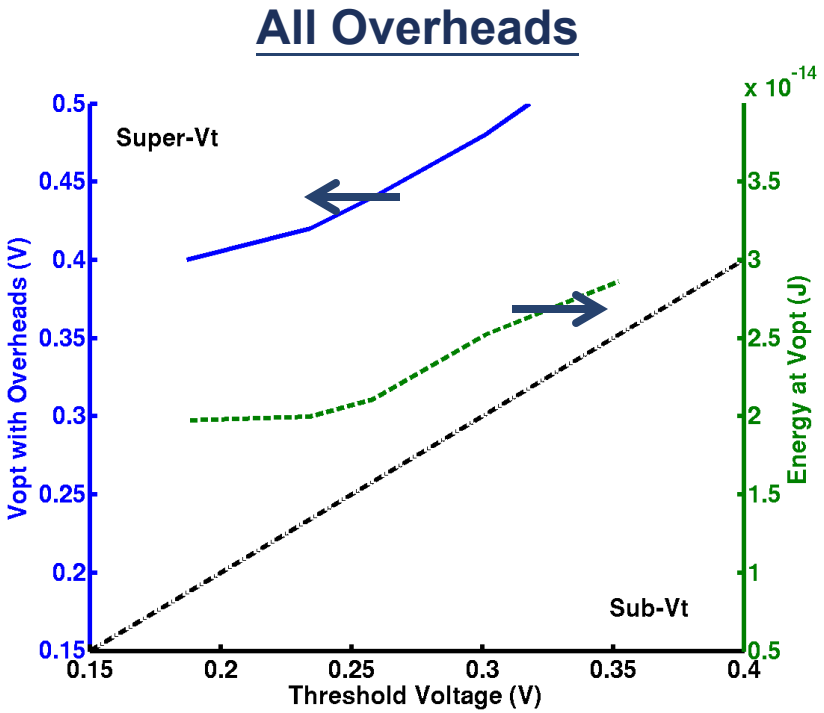
- Vopt increases by 10's of mV when variation included
- Energy gain diminish by roughly 12%
- Overall impact of variation on energy efficiency is managable

Impact of Transistor Vt



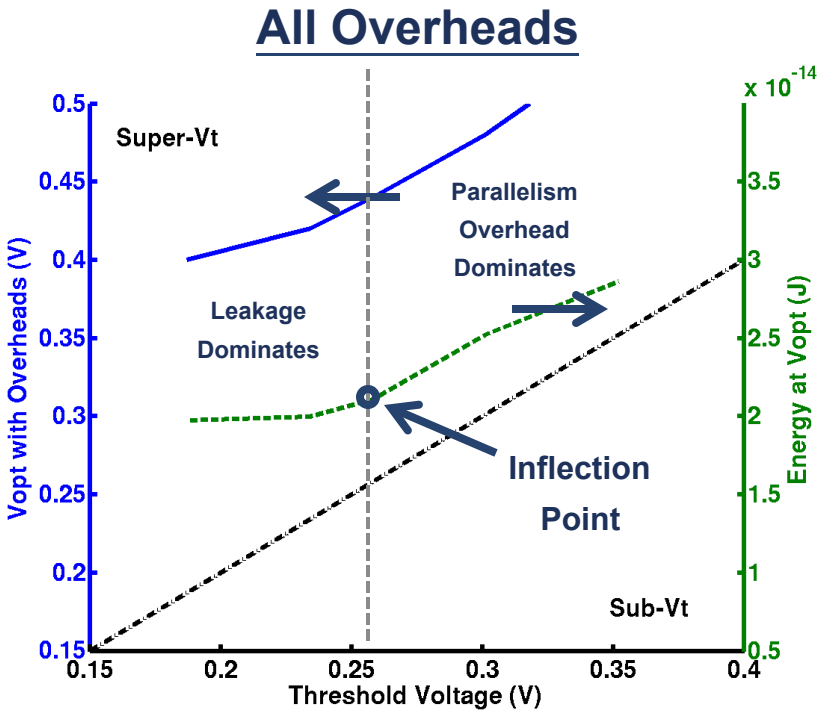
- Leakage, Amdahl, and Architectural overheads
  - At low-Vt leakage dominates achievable energy
  - At high-Vt parallelism overheads dominate (cannot parallelize enough)
  - Strategy: increase Vt to inflection point

Impact of Transistor Vt



- Leakage, Amdahl, and Architectural overheads
  - At low-Vt leakage dominates achievable energy
  - At high-Vt parallelism overheads dominate (cannot parallelize enough)
  - Strategy: increase Vt to inflection point

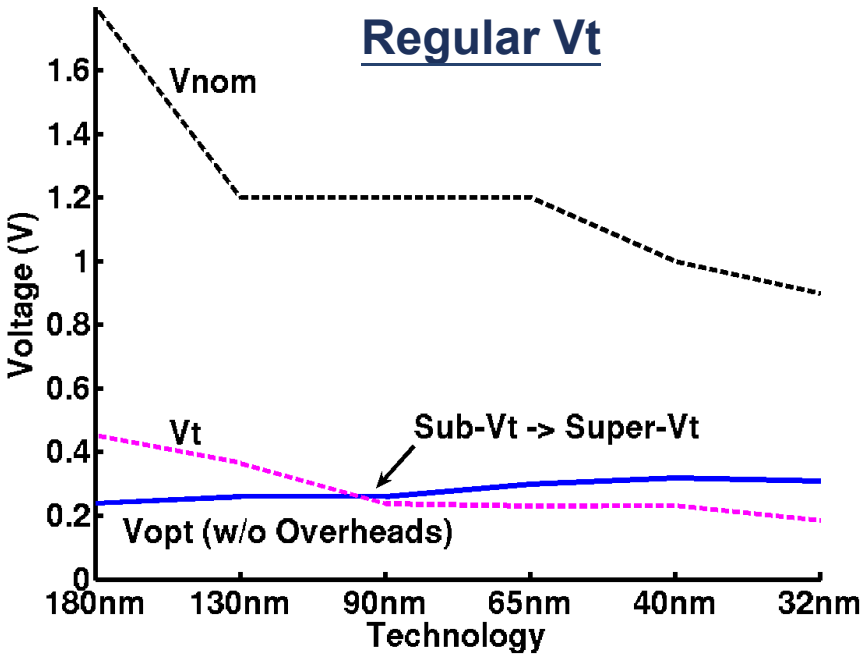
Impact of Transistor Vt



Inflection point indicates where leakage and parallelism overhead are balanced.

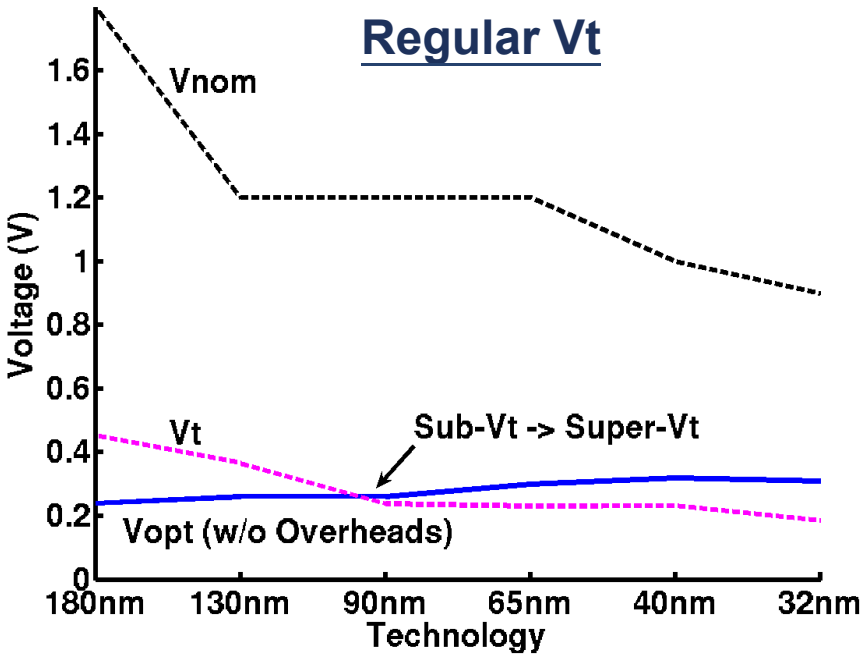
- Leakage, Amdahl, and Architectural overheads
  - At low-Vt leakage dominates achievable energy
  - At high-Vt parallelism overheads dominate (cannot parallelize enough)
  - Strategy: increase Vt to inflection point

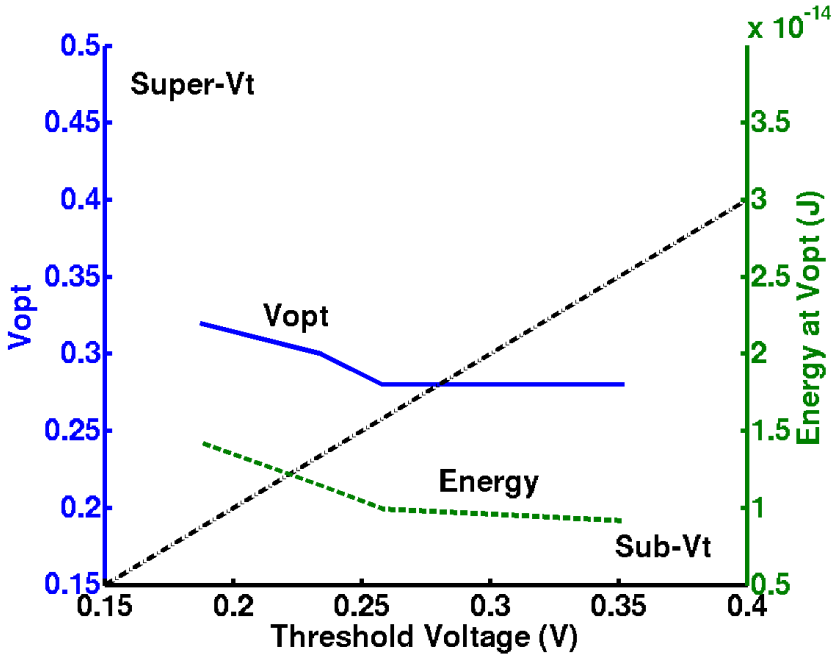
Contrast: No Performance Constraint



- Vopt (no overheads) crossed from Sub-Vt to Super-Vt in 90nm
  - No energy saved by running regular-Vt devices Sub-Vt
- ~100-200 mV above Vt in 32nm
  - Leakage-only case. Not application dependent.

Contrast: No Performance Constraint





- Vopt (no overheads) crossed from Sub-Vt to Super-Vt in 90nm
  - No energy saved by running regular-Vt devices Sub-Vt
- ~100-200 mV above Vt in 32nm
  - Leakage-only case. Not application dependent.