Microprocessor Trends and Implications for the Future

John Mellor-Crummey

Department of Computer Science
Rice University

johnmc@rice.edu

Context

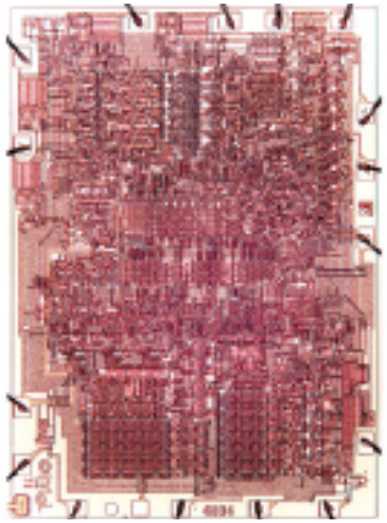
- Last two classes: from transistors to multithreaded designs
  - —multicore chips
  - —multiple threads per core
    - simultaneous multithreading
    - fine-grain multithreading
- Today: hardware trends and implications for the future

The Future of Microprocessors

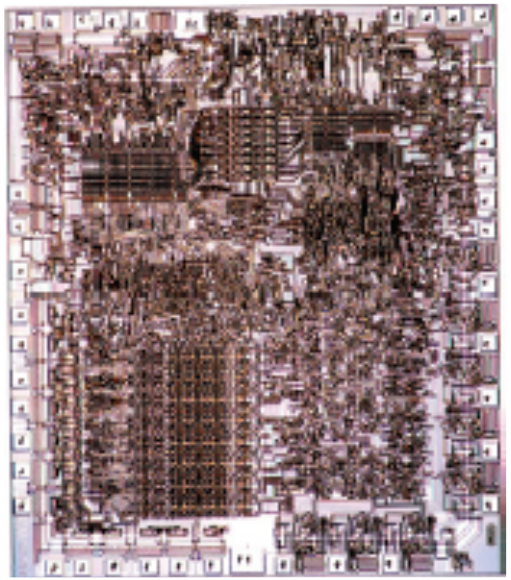
Review: Moore's Law

- Empirical observation
  - —transistor count doubles approximately every 24 months
    - features shrink, semiconductor dies grow
- Impact: performance has increased 1000x over 20 years
  - —microarchitecture advances from additional transistors
  - —faster transistor switching time supports higher clock rates

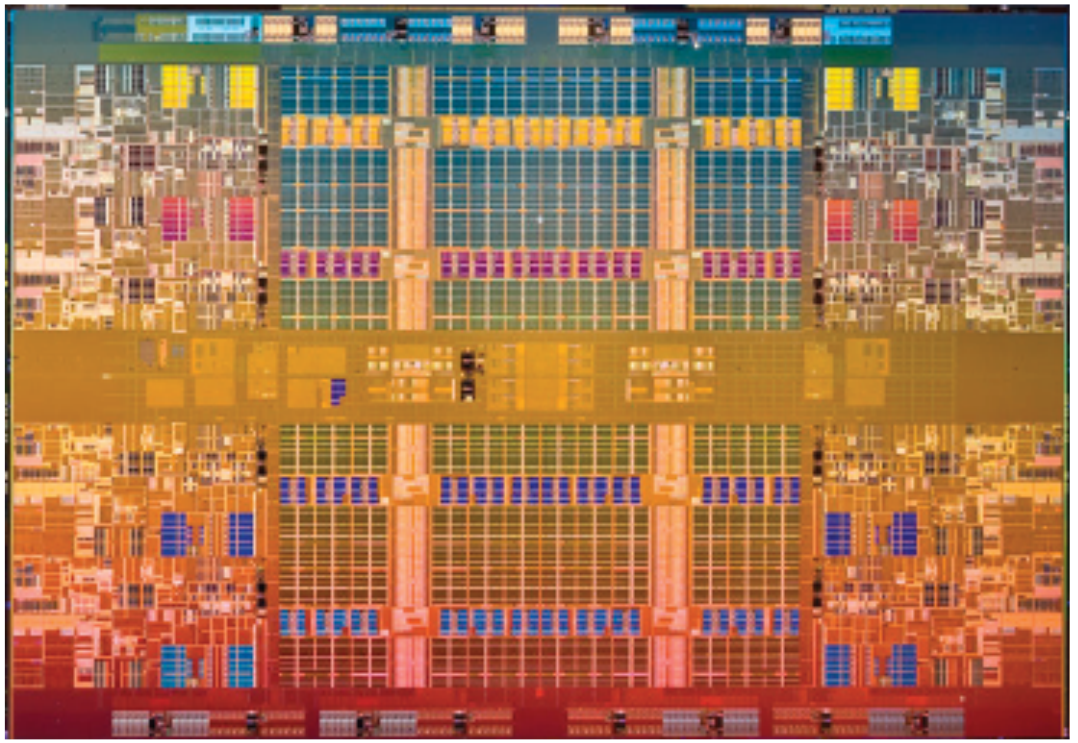
Evolution of Microprocessors 1971-2015



Intel 4004, 1971
1 core, no cache
23K transistors

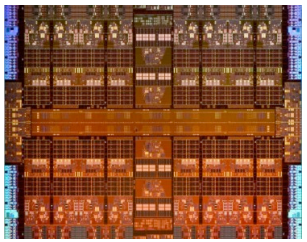


Intel 8008, 1978
1 core, no cache
29K transistors



Intel Nehalem-EX, 2009
8 cores, 24MB cache
2.3B transistors

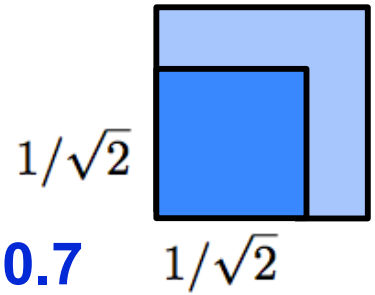
Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.



Oracle SPARC M7 (2015)
32 cores; > 10B transistors

Dennard Scaling: Recipe for a "Free Lunch"

Scaling properties of CMOS circuits



- **Linear scaling of all transistor parameters**
  - **—reduce feature size by a factor of $1/\kappa$, $\kappa \approx \sqrt{2}$; $1/\kappa \approx 0.7$**

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\varepsilon A/t$	$1/\kappa$
Delay time per circuit VC/I	$1/\kappa$
Power dissipation per circuit VI	$1/\kappa^2$
Power density VI/A	1

Delay time ↓ ~ .7x
Frequency ↑~1.4x

power density is constant

- **Simultaneous improvements in transistor density, switching speed, and power dissipation**
- **Recipe for systematic & predictable transistor improvements**

R. Dennard, et al. Design of ion-implanted MOSFETs with very small physical dimensions. IEEE Journal of Solid State Circuits, vol. SC-9, no. 5, pp. 256-268, Oct. 1974.

Impact: 1000x Performance over 20 Years

- **Dennard scaling**
  - —faster transistor switching supports higher clock rates
- **Microarchitecture advances**
  - —enabled by additional transistors
  - —examples: pipelining, out of order execution, branch prediction

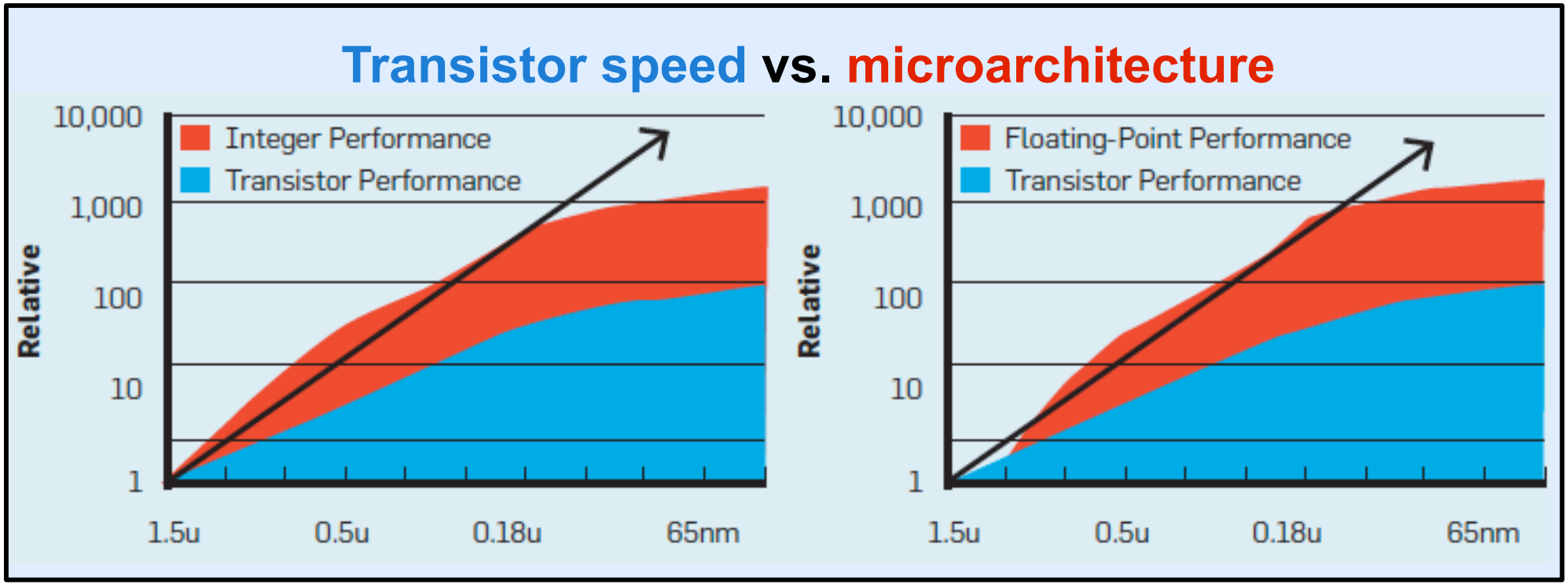
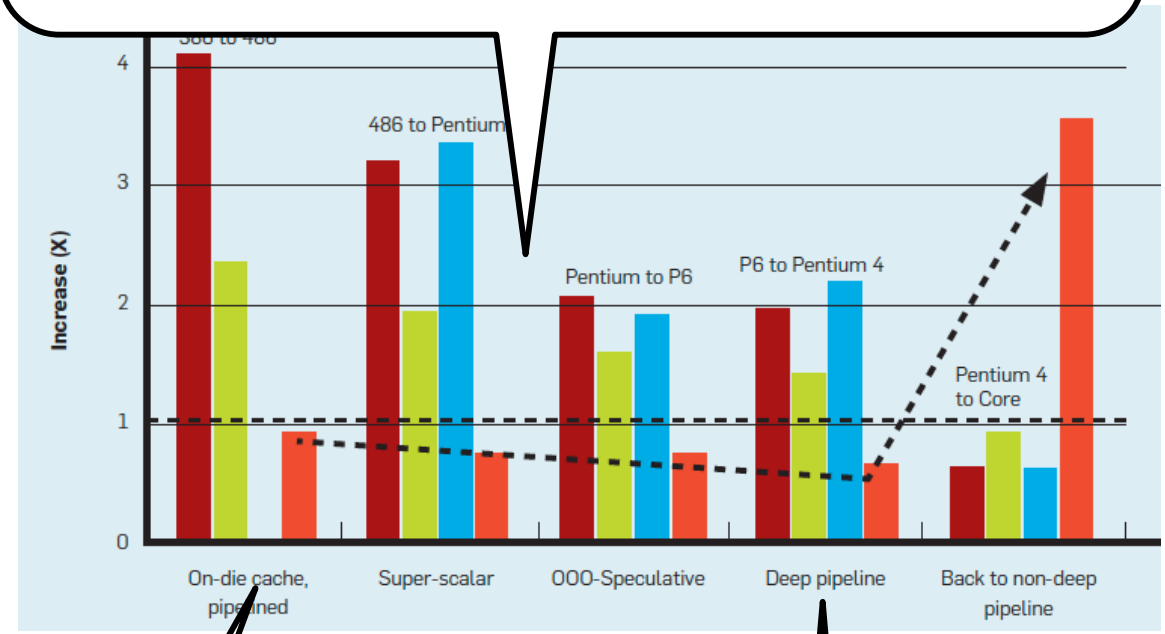


Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Core Microarchitecture Improvements

- **Improvements**
  - —pipelining
  - —branch prediction
  - —out of order execution
  - —speculation
- **Results**
  - —higher performance
  - —higher energy efficiency

superscalar and OOO provided performance benefits at a cost in energy efficiency



Measure performance with SPEC INT 92, 95, 2000

Die Area

Integer Perform ce (X)

FP Perfo mance (X)

Int Perf mance/Watt (X)

on-die cache and pipelined architectures beneficial: significant performance gain without compromising energy

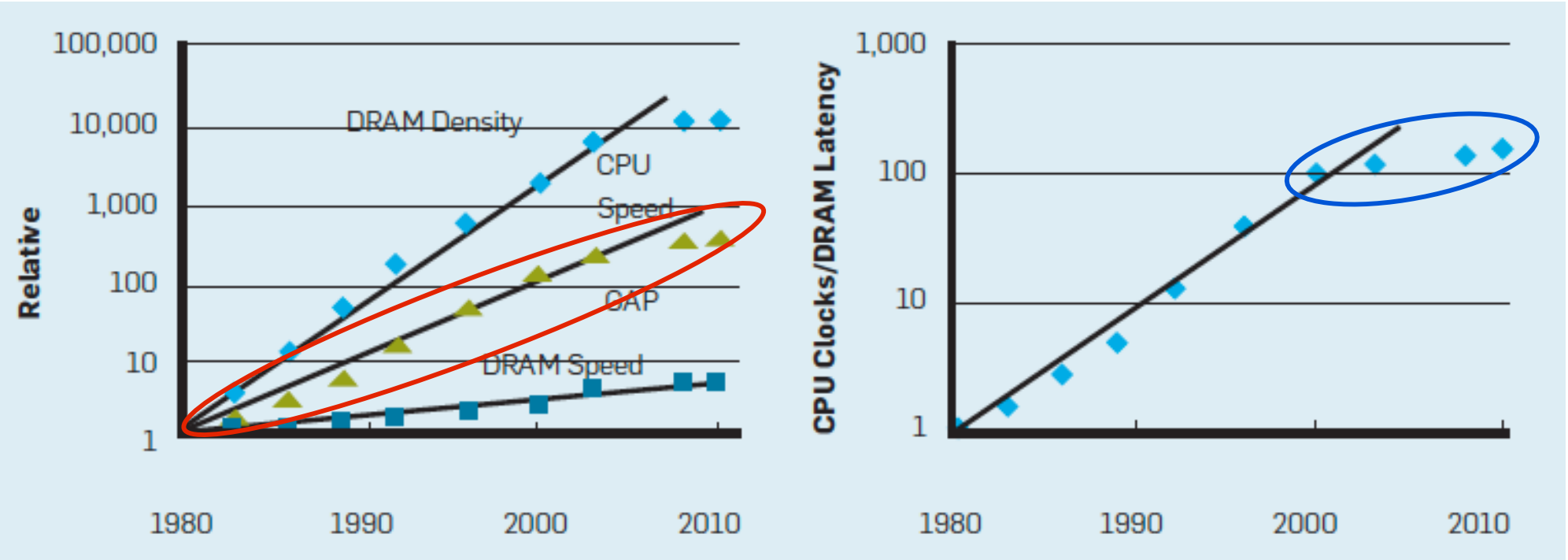
deep pipeline delivered lowest performance increase for same area and power increase as OOO speculative

The End of Dennard Scaling

- Decreased scaling benefits despite shrinking transistors
  - —complications
    - transistors are not perfect switches: leakage current
      substantial fraction of power consumption now due to leakage
    - keep leakage under control: can't lower threshold voltage
      reduces transistor performance
  - —result
    - little performance improvement
    - little reduction in switching energy
- New constraint: energy consumption
  - —finite, fixed energy budget
  - —key metric for designs: energy efficiency
  - —HW & SW goal: energy proportional computing
    - with a fixed power budget: ↑ energy efficiency = ↑ performance

Problem: Memory Performance Lags CPU

- **Growing disparity between processor speed and DRAM speed**
  - **—DRAM speed improves slower b/c optimized for density and cost**



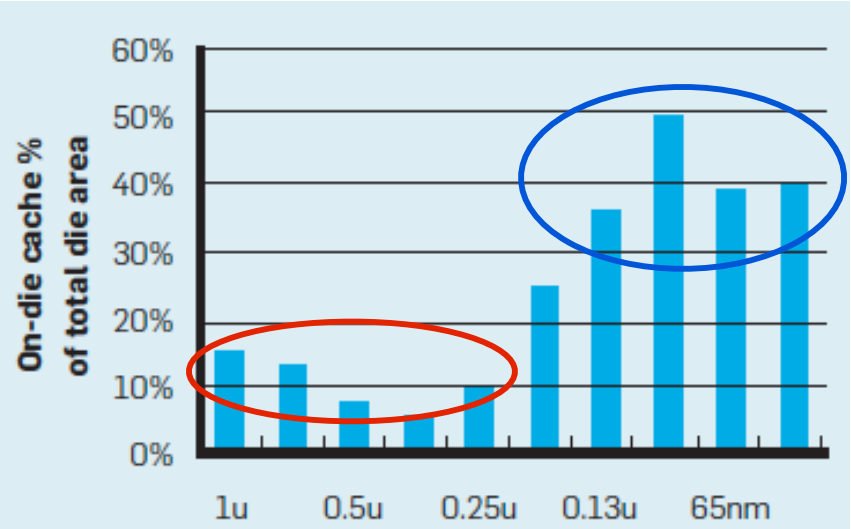
DRAM Density and Performance, 1980-2010

- **Speed disparity growing from 10s to 100s of processor cycles per memory access**
- **Speed flattens out due to flattening of clock frequency**

Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Cache-based Memory Hierarchies

- DRAM design: emphasize density and cost over speed
- 2 or 3 levels of cache: span growing speed gap with memory
- Caches
  - —L1: high bandwidth; low latency → small
  - —L2+: optimized for size and speed



- Initially, most transistors devoted to microarchitecture
- Later, larger caches became important to reduce energy

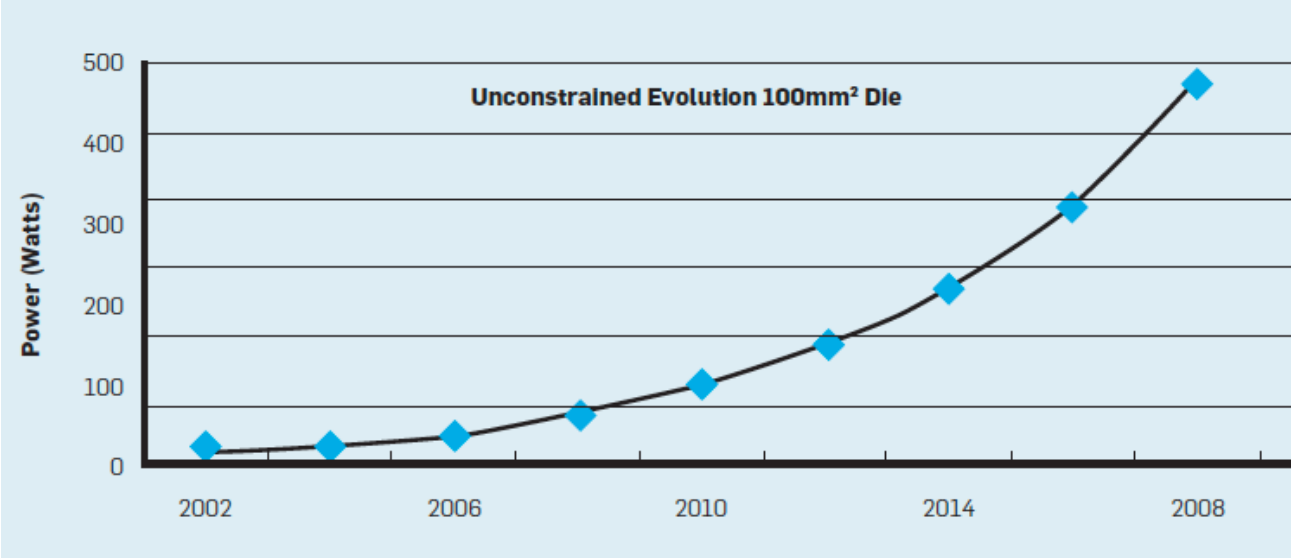
Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

The Next 20 Years (2011 and Beyond)

- **Last 20 years: 1000x performance improvement**
- **Continuing this trajectory: another 30x by 2020**

Unconstrained Evolution vs. Power

- If
  - —add more cores as transistors and integration capacity increases
  - —operate at highest frequency transistors and designs can achieve
- Then, power consumption would be prohibitive

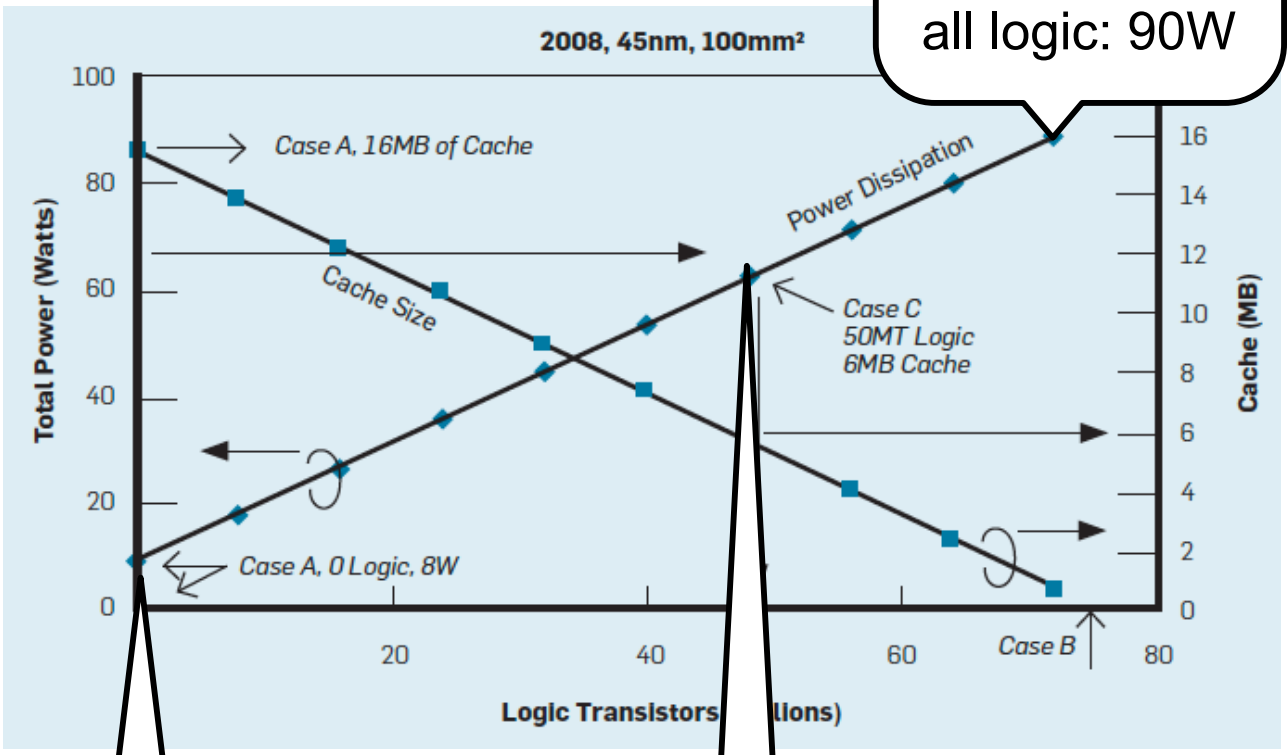


- Implications
  - —chip architects must limit number of cores and frequency to keep power reasonable
    - – severely limits performance improvements achievable!

Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Transistor Integration @ Fixed Power

- **Desktop applications**
  - —power envelope: 65W; die size 100 mm²
- **Transistor integration capacity at fixed power envelope**
  - — analysis for 45nm process technology
    - – ↑ # logic T
    - – size of cache ↓
  - —as # logic T ↑, power dissipation increases
- **Analysis assumes avg activity seen in ~2011**



~ Core 2 Duo

Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

What about the Future (Past 2011)?

Projections from Intel

- **Modest frequency increase per generation 15%**
- **5% reduction in supply voltage**
- **25% reduction of capacitance**
- **Expect to follow Moore's law for transistor increases, but increase logic 3x and cache > 10x**

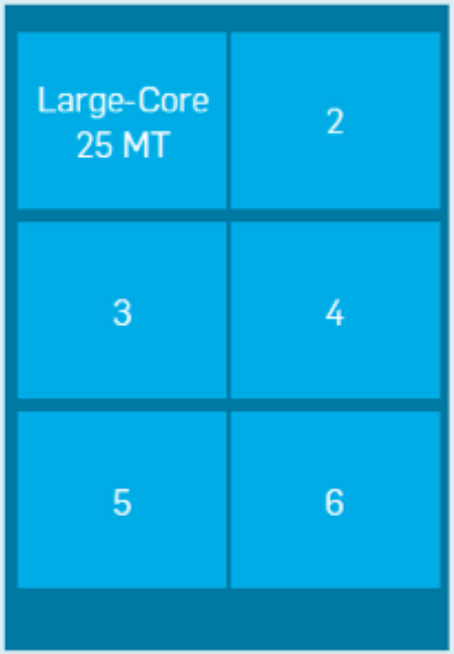
Year	Logic Transistors (Millions)	Cache MB
2008	50	6
2014	100	25
2018	150	80

Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Key Challenges Ahead

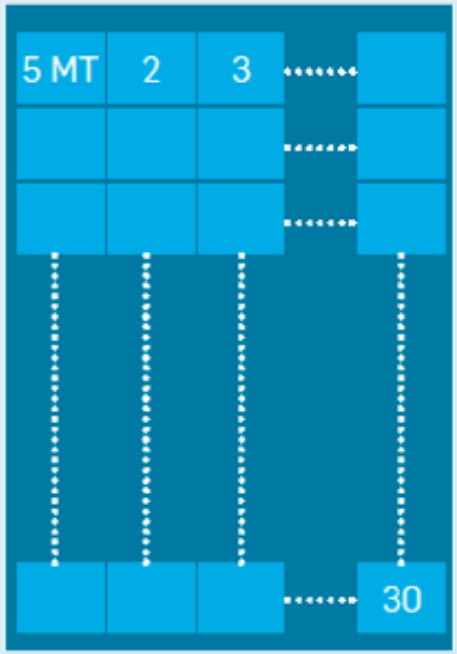
- **Organizing the logic: multiple cores and customization**
  - —single thread performance has leveled off
  - —throughput can increase proportional to number of cores
  - —customization can reduce execution latency
  - —multiple cores + customization can improve energy efficiency
- **Choices for multiple cores**

Three Scenarios for a 150M Transistor Chip



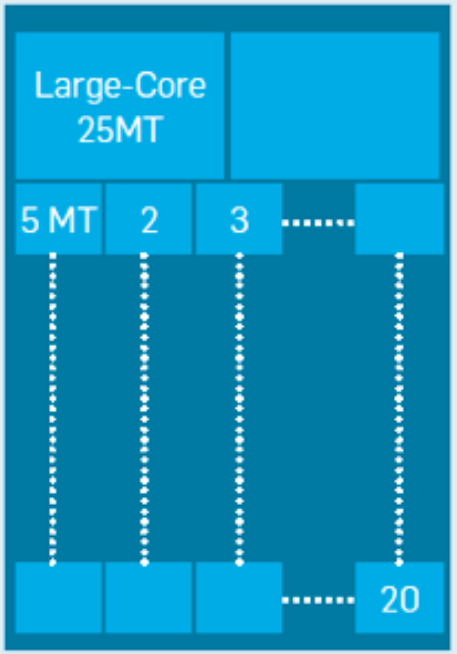
Large-Core Homogeneous	
Large-core throughput	1
Small-core throughput	
Total throughput	6

(a)



Small-Core Homogeneous	
Large-core throughput	
Small-core throughput	Pollack's Rule $(5/25)^{0.5}=0.45$
Total throughput	13

(b)



Hybrid approach	
Large-core throughput	1
Small-core throughput	Pollack's Rule $(5/25)^{0.5}=0.45$
Total throughput	11

(c)

Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Death of 90/10 Optimization

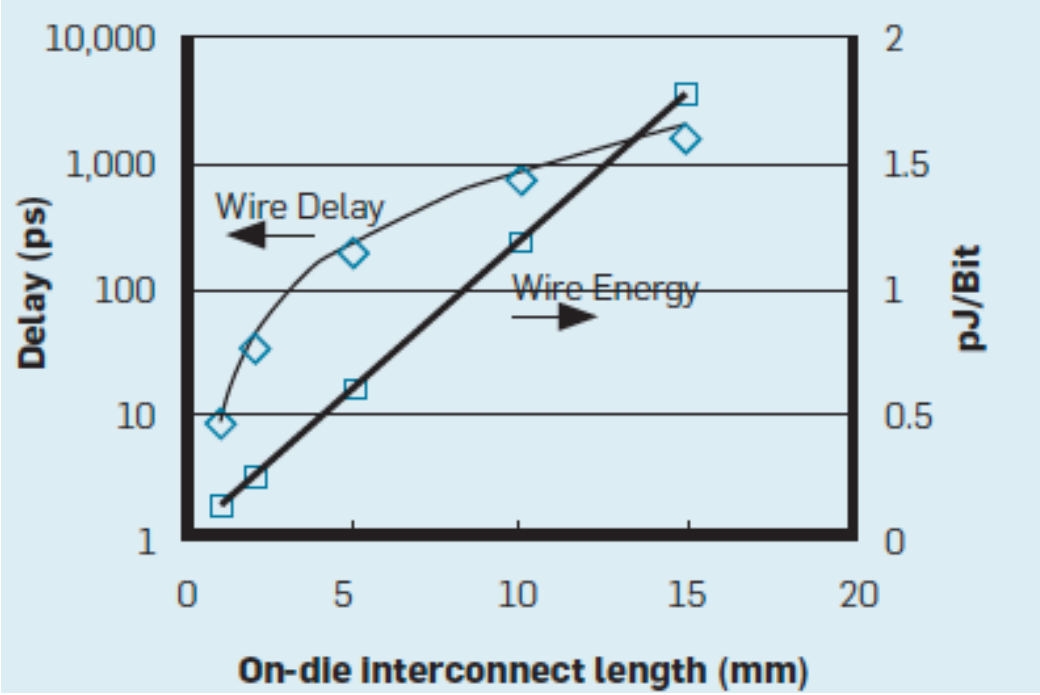
- Traditional wisdom: invest maximum transistors in 90% case
  - —use precious transistors to increase single thread performance that can be applied broadly
- However
  - —new scaling regime (slow transistor performance, energy efficiency) → no sense to add transistors to a single core as energy efficiency suffers
- Result: 90/10 rule no longer applies
- Rise of 10x10 optimization
  - —attack performance as a set of 10% optimization opportunities
    - optimize with an accelerator for a 10% case, another for a different 10% case, and then another 10% case, and so on ...
  - —operate chip with 10% of transistors active, 90% inactive
    - different 10% active at each point in time
  - —can produce chip with better overall energy efficiency and performance

Some Design Choices

- Accelerators for specialized tasks
  - —graphics
  - —media
  - —image
  - —cryptographic
  - —radio
  - —digital signal processing
  - —FPGA
- Increase energy efficiency by restricting memory access structure and control flexibility
  - —SIMD
  - —SIMT - GPUs require expressing programs as structured sets of threads

On-die Interconnect Delay and Energy (45nm)

- As energy cost of computation reduced by voltage scaling, data movement costs start to dominate
- Energy moving data will have critical effect on performance
  - —every pJ spent moving data reduces budget for computation



Improving Energy Efficiency Through Voltage Scaling

- **As supply voltage is reduced, frequency also reduces, but energy efficiency increases**
  - **—while maximally energy efficient, reducing to threshold voltage would dramatically reduce single-thread performance: not recommended**

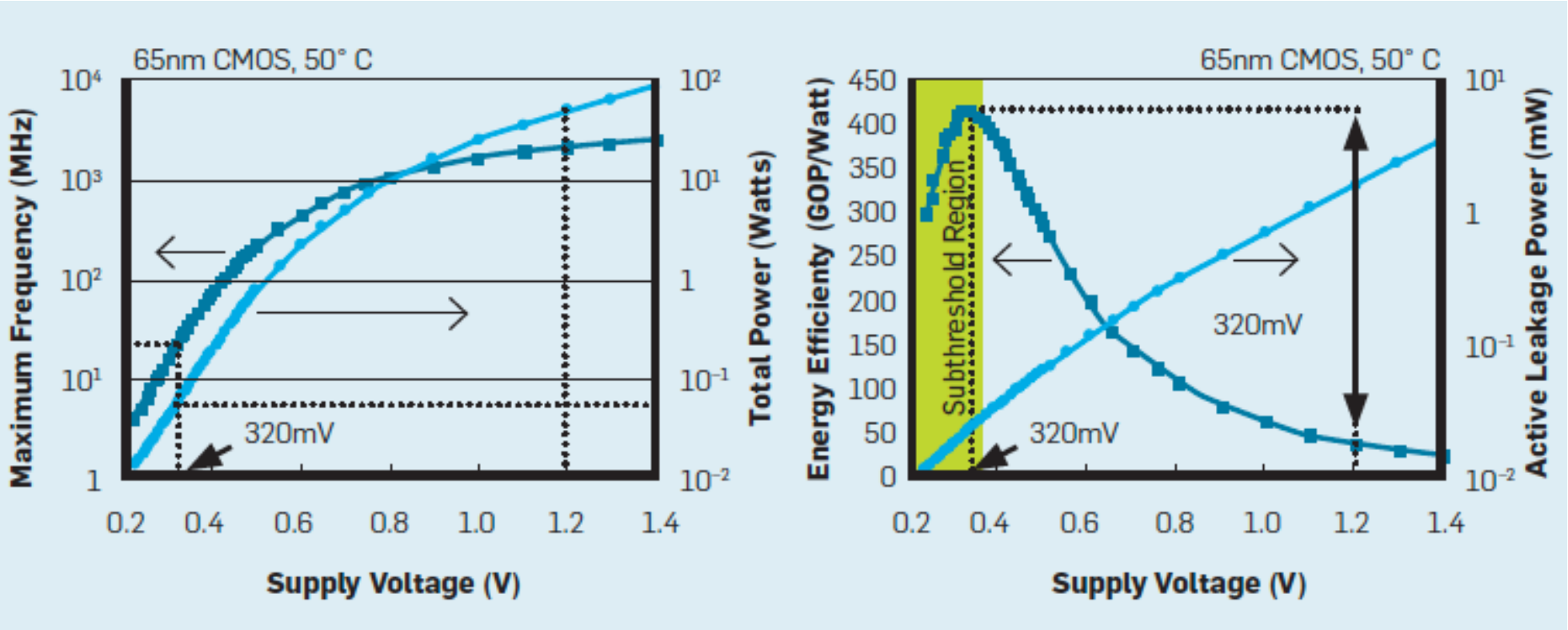


Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Heterogeneous Many-core with Variation

Small cores could operate at different design points to trade performance for energy efficiency

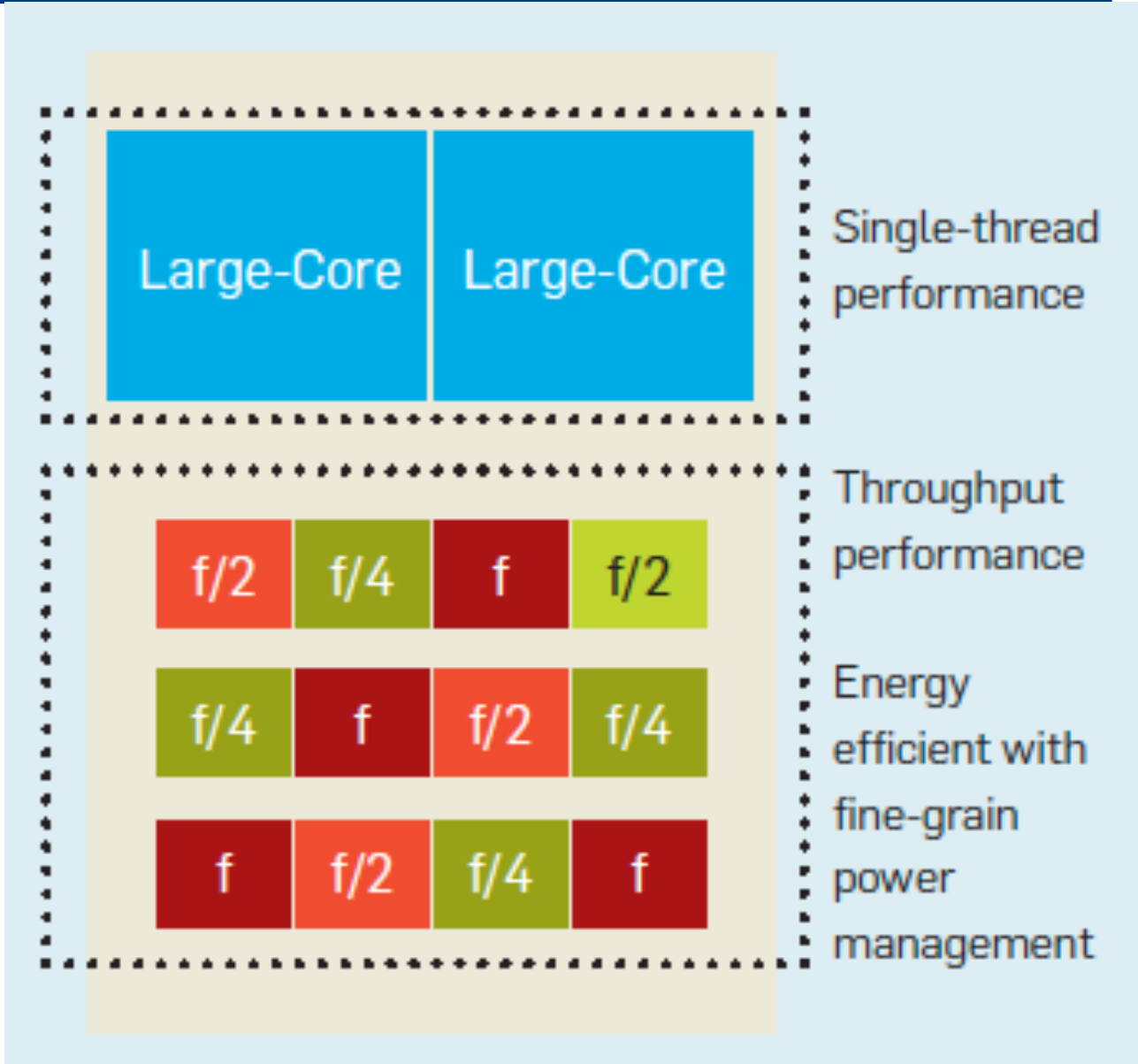


Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Data Movement Challenges, Trends, Directions

Challenge	Near-Term	Long-Term
Parallelism	Increased parallelism	Heterogeneous parallelism and customization, hardware/runtime placement, migration, adaptation for locality and load balance
Data Movement/ Locality	More complex, more exposed hierarchies; new abstractions for control over movement and "snooping"	New memory abstractions and mechanisms for efficient vertical data locality management with low programming effort and energy
Resilience	More aggressive energy reduction; compensated by recovery for resilience	Radical new memory technologies (new physics) and resilience techniques
Energy Proportional Communication	Fine-grain power management in packet fabrics	Exploitation of wide data, slow clock, and circuit-based techniques
Reduced Energy	Low-energy address translation	Efficient multi-level naming and memory-hierarchy management

Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Circuits Challenges, Trends, Directions

Challenge	Near-Term	Long-Term
Power, energy efficiency	Continuous dynamic voltage and frequency scaling, power gating, reactive power management	Discrete dynamic voltage and frequency scaling, near threshold operation, proactive fine-grain power and energy management
Variation	Speed binning of parts, corrections with body bias or supply voltage changes, tighter process control	Dynamic reconfiguration of many cores by speed
Gradual, temporal, intermittent, and permanent faults	Guard-bands, yield loss, core sparing, design for manufacturability	Resilience with hardware/software co-design, dynamic in-field detection, diagnosis, reconfiguration and repair, adaptability, and self-awareness

Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Software Challenges, Trends, Directions

Challenge	Near-Term	Long-Term
1,000-fold software parallelism	Data parallel languages and "mapping" of operators, library and tool-based approaches	New high-level languages, compositional and deterministic frameworks
Energy-efficient data movement and locality	Manual control, profiling, maturing to automated techniques (auto-tuning, optimization)	New algorithms, languages, program analysis, runtime, and hardware techniques
Energy management	Automatic fine-grain hardware management	Self-aware runtime and application-level techniques that exploit architecture features for visibility and control
Resilience	Algorithmic, application-software approaches, adaptive checking and recovery	New hardware-software partnerships that minimize checking and recomputation energy

Figure credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Take Away Points

- **Moore's Law continues, but demands radical changes in architecture and software**
- **Architectures will go beyond homogeneous parallelism, embrace heterogeneity, and exploit the bounty of transistors to incorporate application-customized hardware**
- **Software must increase parallelism and exploit heterogeneous and application-customized hardware to deliver performance growth**

Credit: Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors. Communications of the ACM, Vol. 54 No. 5, Pages 67-77 10.1145/1941487.1941507.

Looking back and looking forward: power, performance, and upheaval

Of Power and Wires

- **Physical power and wire delay limits**
  - —constrain performance of current and future technologies
- **Power is now a first order constraint on designs**
  - —limits clock scaling
  - —prevents using all transistors simultaneously
    - Dark Silicon and the end of multicore scaling. Esmaeilzadeh et al. ISCA 11

Analyzing Power Consumption

- Quantitative performance analysis is the foundation for computer system design and innovation
  - —need detailed information to improve performance
- Goal: apply quantitative analysis to measured power
  - —lack of detailed energy measurements is impairing efforts to reduce energy consumption of modern workloads

Processors Considered

Specifications for 8 processors used in experiments

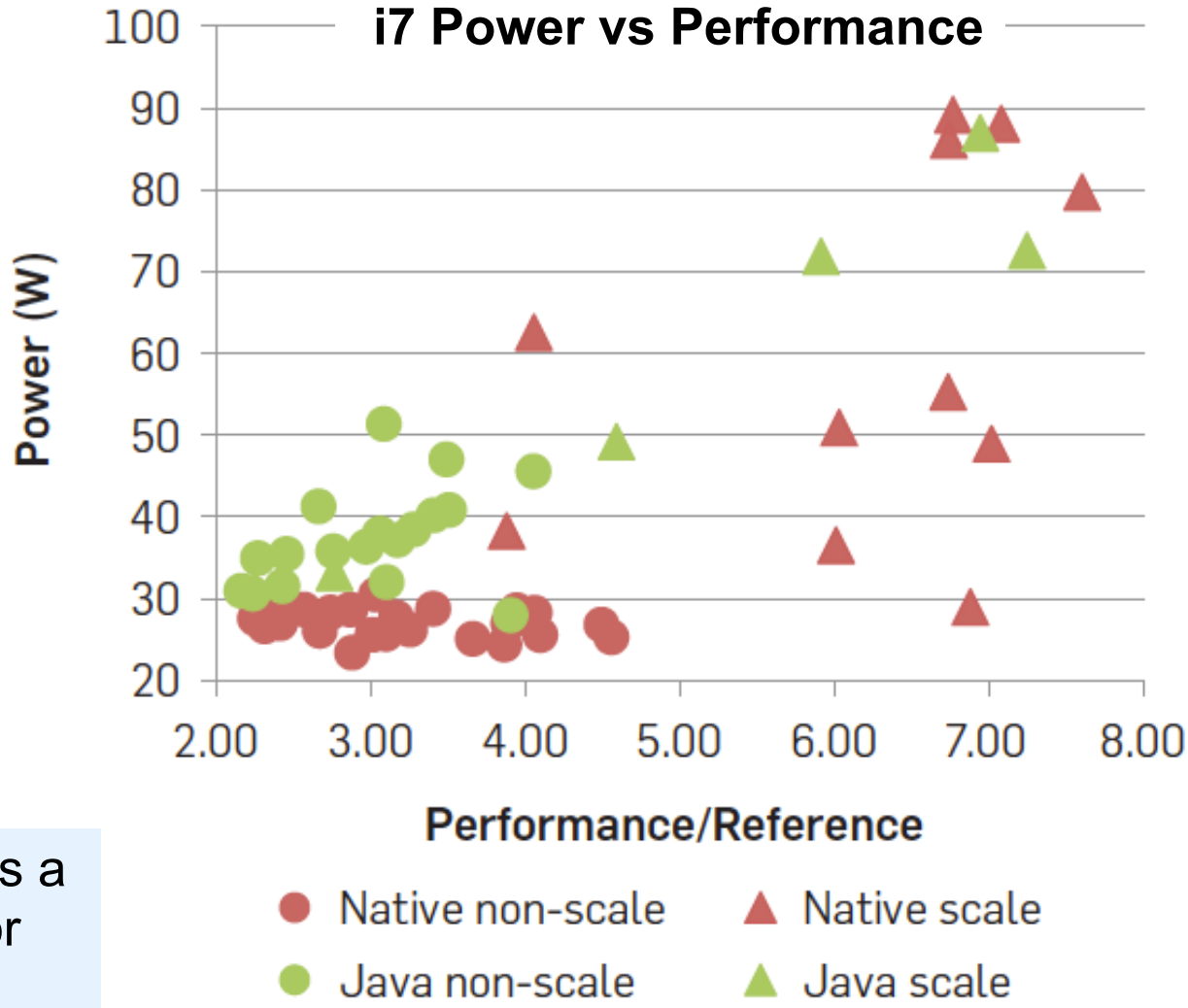
Processor	µArch	Processor	sSpec	Release date	Price (USD)	CMP SMT	LLC (B)	Clock (GHz)	nm	Trans M	Die (mm²)	VID Range (V)	TDP (W)	FSB (MHz)	B/W (GB/s)	DRAM Model
Pentium 4	NetBurst	Northwood	SL6WF	May '03	–	1C2T	512K	2.4	130	55	131	–	66	800	–	DDR-400
Core 2 Duo E6600	Core	Conroe	SL9S8	Jul '06	316	2C1T	4M	2.4	65	291	143	0.85–1.50	65	1066	–	DDR2-800
Core 2 Quad Q6600	Core	Kentsfield	SL9UM	Jan '07	851	4C1T	8M	2.4	65	582	286	0.85–1.50	105	1066	–	DDR2-800
Core i7 920	Nehalem	Bloomfield	SLBCH	Nov '08	284	4C2T	8M	2.7	45	731	263	0.80–1.38	130	–	25.6	DDR3-1066
Atom 230	Bonnell	Diamondville	SLB6Z	Jun '08	29	1C2T	512K	1.7	45	47	26	0.90–1.16	4	533	–	DDR2-800
Core 2 Duo E7600	Core	Wolfdale	SLGTD	May '09	133	2C1T	3M	3.1	45	228	82	0.85–1.36	65	1066	–	DDR2-800
Atom D510	Bonnell	Pineview	SLBLA	Dec '09	63	2C2T	1M	1.7	45	176	87	0.80–1.17	13	665	–	DDR2-800
Core i5 670	Nehalem	Clarkdale	SLBLT	Jan '10	284	2C2T	4M	3.4	32	382	81	0.65–1.40	73	–	21.0	DDR3-1333

Benchmark Classes

- **Native non-scalable**
  - —single-threaded, compute-intensive C, C++, and Fortran benchmarks from SPEC CPU2006
- **Native scalable**
  - —multithreaded C and C++ benchmarks from PARSEC
- **Java non-scalable**
  - —single and multithreaded benchmarks that do not scale well from SPECjvm, DaCapo 06-10-MR2, DaCapo 9.12, and pjbb2005
- **Java scalable**
  - —multithreaded Java benchmarks from DaCapo 9.12 that scale in performance similarly to native scalable

Power is Application Dependent

Each of 61 points represents a benchmark. Power consumption varies from 23-89W. The wide spectrum of power responses points to power saving opportunities in software.

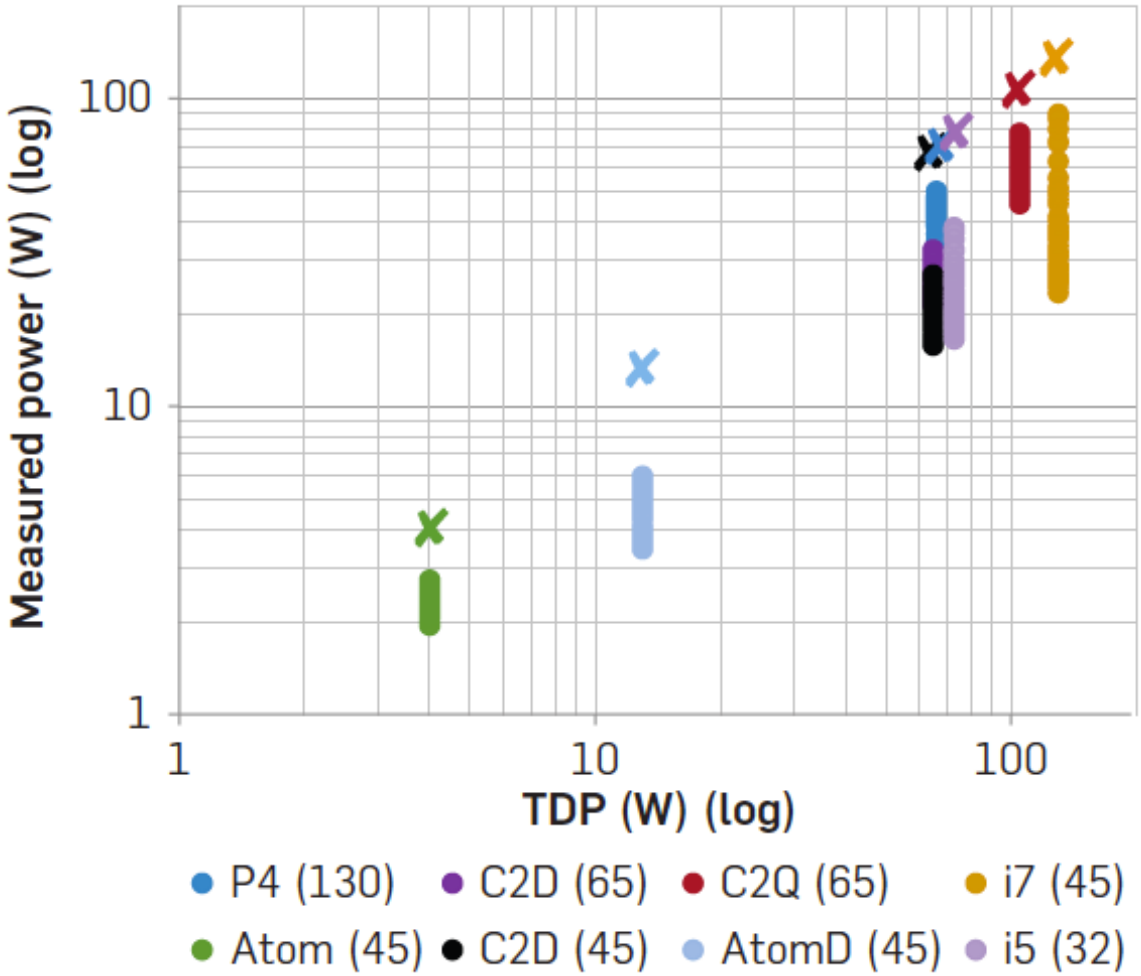


Finding: each workload prefers a different HW configuration for energy efficiency

Figure credit: Hadi Esmaeilzadeh, Ting Cao, Xi Yang, Stephen M. Blackburn, and Kathryn S. McKinley. 2012. Looking back and looking forward: power, performance, and upheaval. *CACM* 55, 7 (July 2012), 105-114.

Power Consumption on Different Processors

Measured power for each processor running 61 benchmarks. Each point represents measured power for one benchmark. The "✗"s are the reported TDP for each processor.



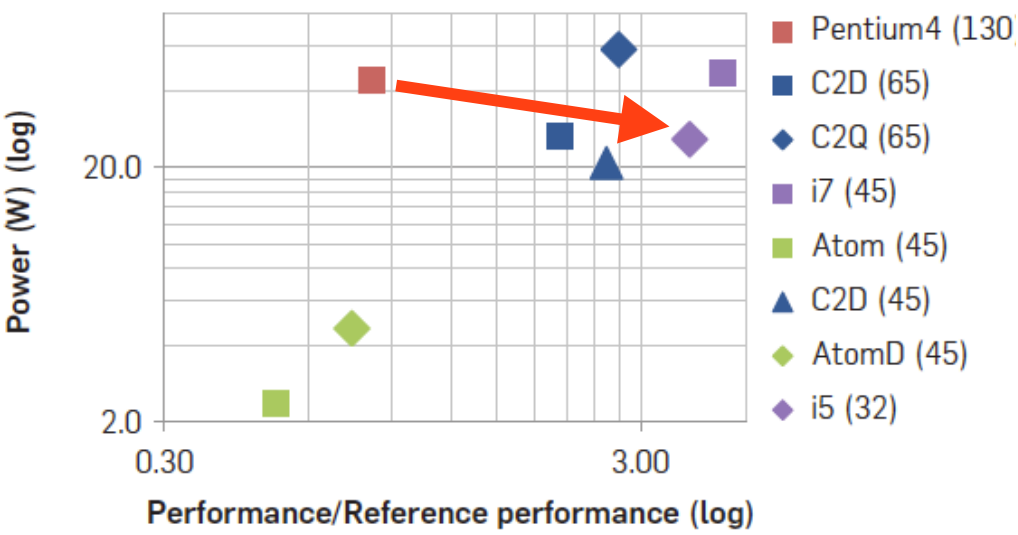
Finding: power is application dependent and does not strongly correlate with TDP

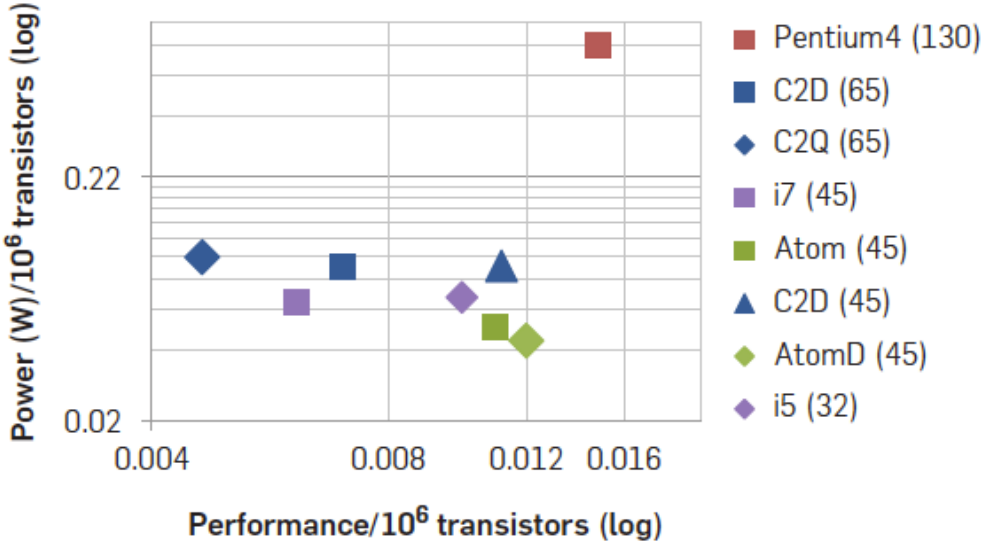
Figure credit: Hadi Esmaeilzadeh, Ting Cao, Xi Yang, Stephen M. Blackburn, and Kathryn S. McKinley. 2012. Looking back and looking forward: power, performance, and upheaval. *CACM* 55, 7 (July 2012), 105-114.

Power, Performance, & Transistors

Power/performance trade-off by processor

- Each point is an average of the 4 workloads
  - (native, Java) x (scalable, non-scalable)





- Power/performance trade-offs have changed from Pentium 4 (130) to i5 (32).

Figure credit: Hadi Esmaeilzadeh, Ting Cao, Xi Yang, Stephen M. Blackburn, and Kathryn S. McKinley. 2012. Looking back and looking forward: power, performance, and upheaval. *CACM* 55, 7 (July 2012), 105-114.

- Power and performance per million transistors. Power per million transistors is consistent across different microarchitectures regardless of the technology node. On average, Intel processors burn around 1 W for every 20 million transistors.

Energy/Performance Pareto Frontiers (45nm)

Energy/performance optimal designs are application dependent and significantly deviate from the average case

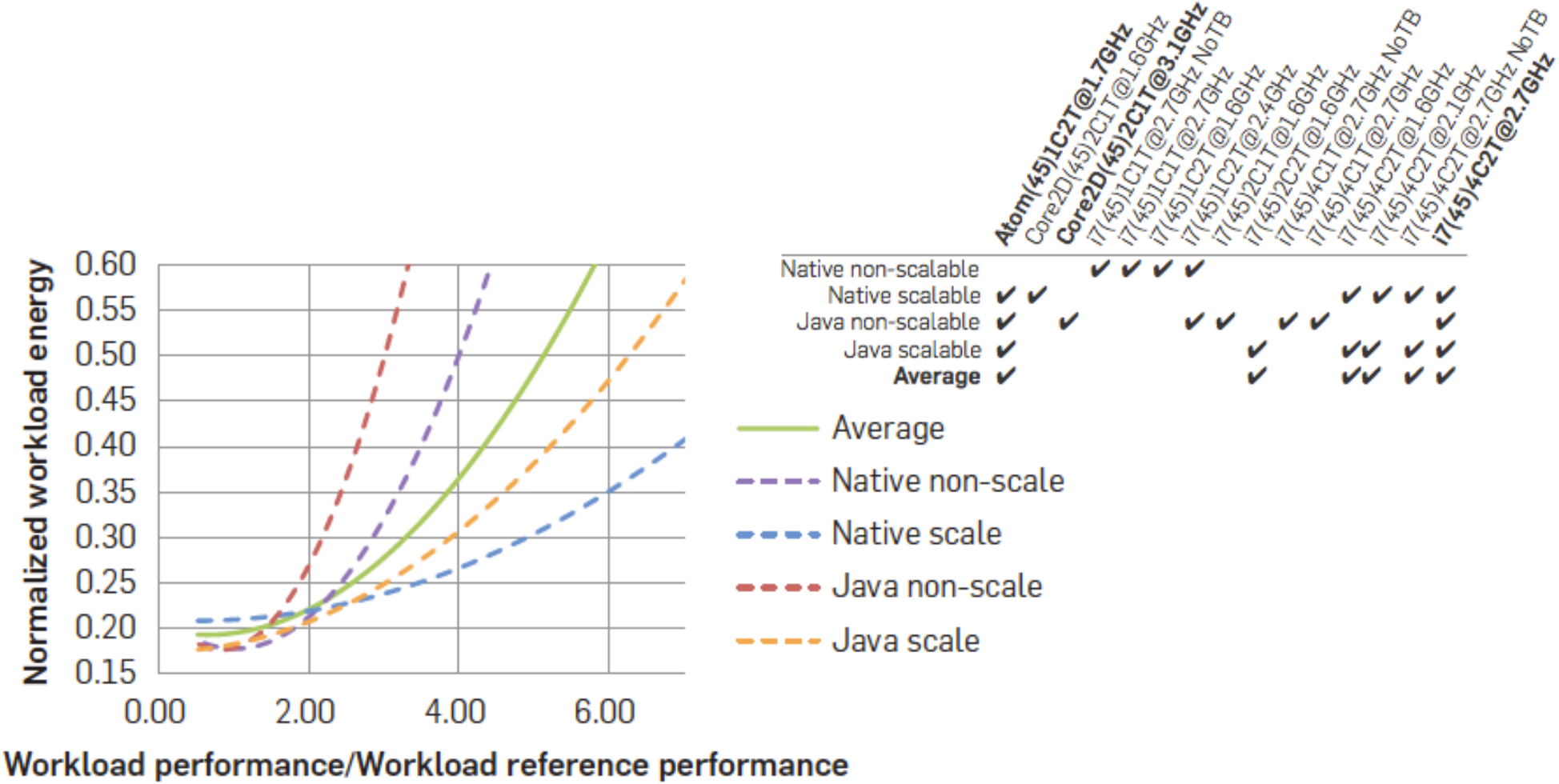
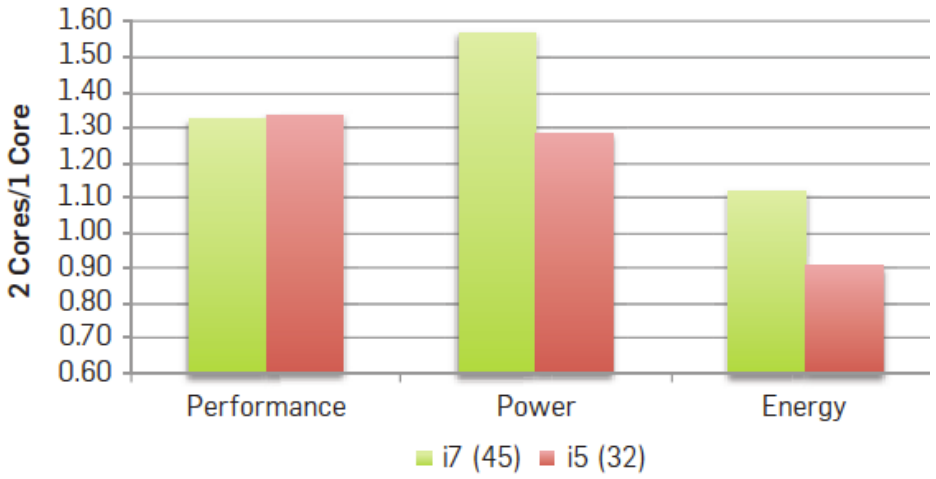
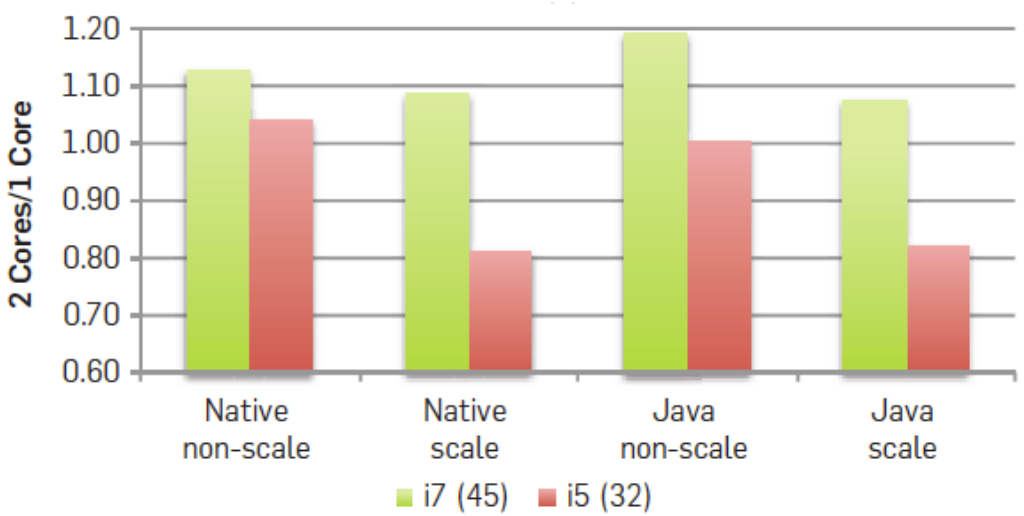


Figure credit: Hadi Esmaeilzadeh, Ting Cao, Xi Yang, Stephen M. Blackburn, and Kathryn S. McKinley. 2012. Looking back and looking forward: power, performance, and upheaval. *CACM* 55, 7 (July 2012), 105-114.

CMP: Comparing Two Cores to One





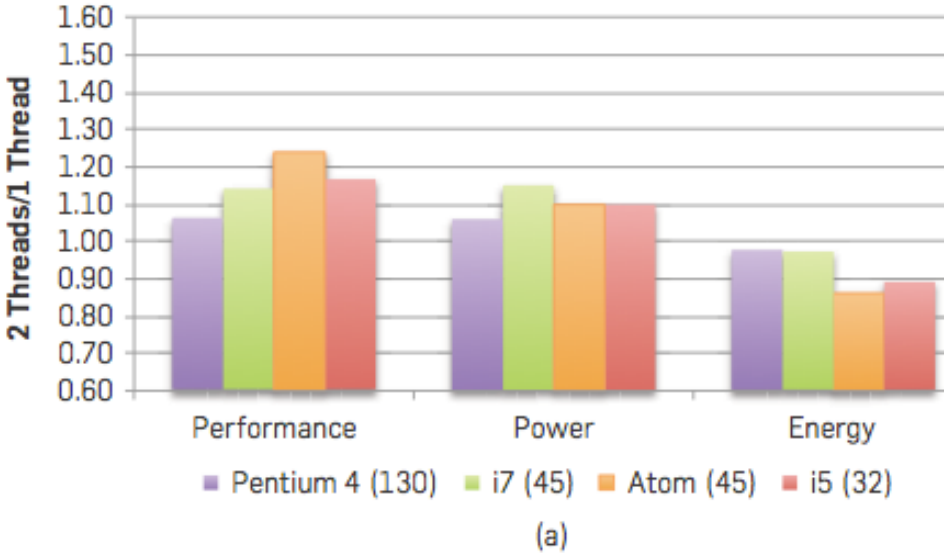
Impact of doubling the number of cores on performance, power, and energy, averaged over all four workloads.

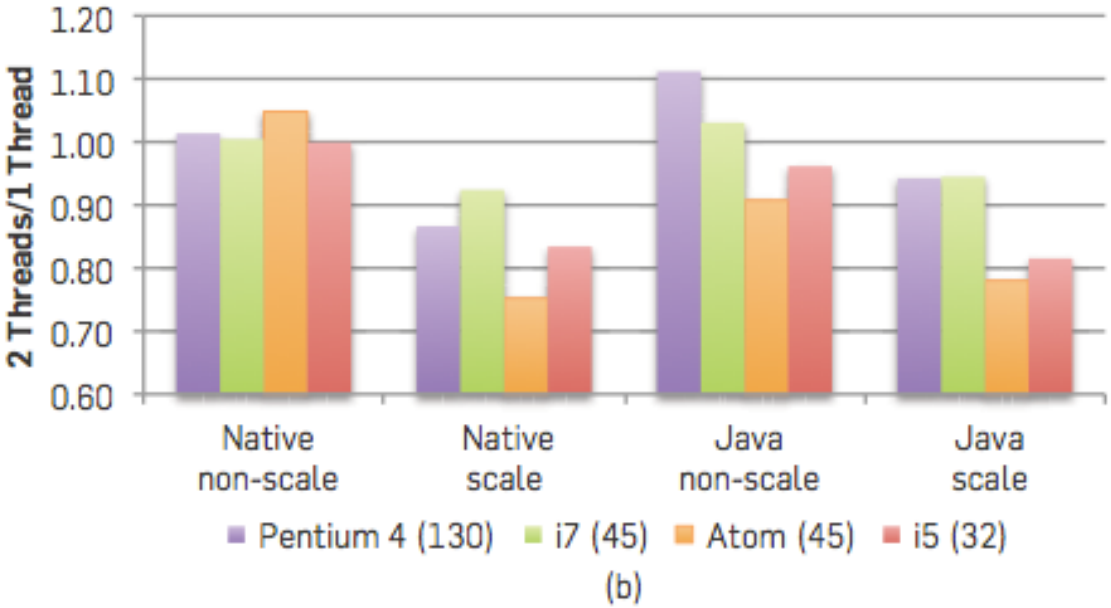
Energy impact of doubling the number of cores for each workload. Doubling the cores is not consistently energy efficient among processors or workloads.

Figure credit: Hadi Esmaeilzadeh, Ting Cao, Xi Yang, Stephen M. Blackburn, and Kathryn S. McKinley. 2012. Looking back and looking forward: power, performance, and upheaval. *CACM* 55, 7 (July 2012), 105-114.

Simultaneous Multithreading

Figure 9. SMT: one core with and without SMT. (a) Impact of enabling two-way SMT on a single-core with respect to performance, power, and energy, averaged over all four workloads. (b) Energy impact of enabling two-way SMT on a single core for each workload. Enabling SMT delivers significant energy savings on the recent i5 (32) and the in-order Atom (45).





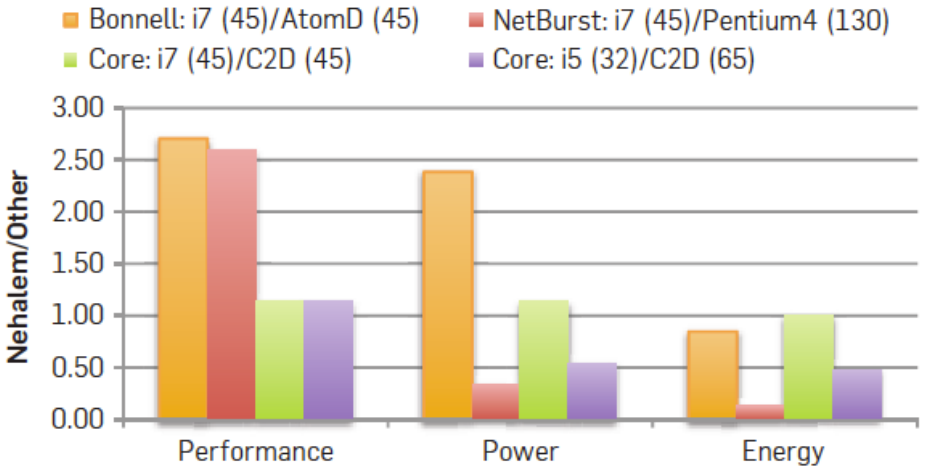
Finding: SMT delivers substantial energy savings for recent hardware and for in-order processors

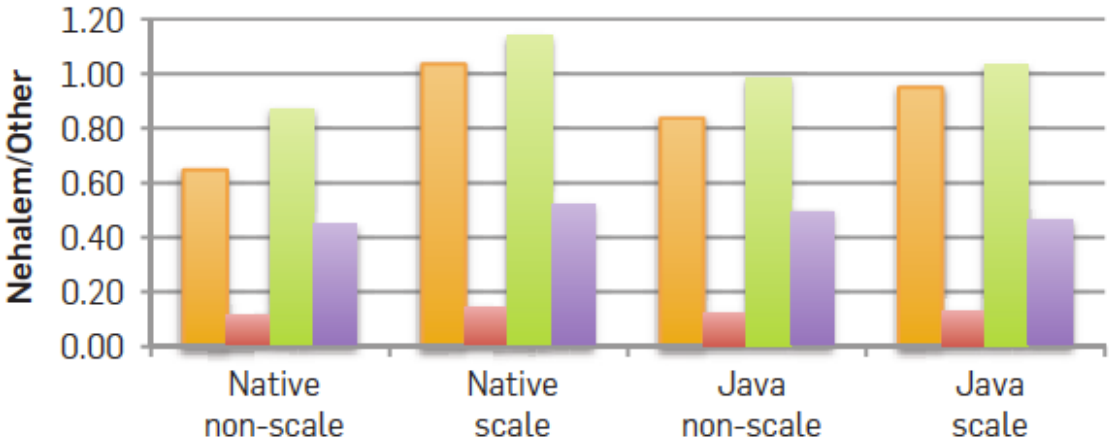
Figure credit: Hadi Esmaeilzadeh, Ting Cao, Xi Yang, Stephen M. Blackburn, and Kathryn S. McKinley. Looking back and looking forward: power, performance, and upheaval. *CACM* 55, 7 (July 2012), 105-114.

Comparing Microarchitectures

Nehalem vs. four other architectures

In each comparison, the Nehalem is configured to match the other processor as closely as possible





Impact of microarchitecture change with respect to performance, power, and energy, averaged over all four workloads.

Energy impact of microarchitecture for each workload. The most recent microarchitecture, Nehalem, is more energy efficient than the others, including the low-power Bonnell (Atom).

Looking Forward: Findings

- Power is application dependent and poorly correlated to TDP
- Power per transistor is relatively consistent within microarchitecture family, independent of process technology
- Energy-efficient architecture design is very sensitive to workload
- Enabling a core is not consistently energy efficient (1 core vs. 2 cores)
- The JVM adds parallelism to single threaded Java benchmarks
- SMT saves significant energy for recent hardware and for in-order processors
- Two recent die shrinks deliver similar and surprising reductions in energy, even when controlling for clock frequency
- Controlling for technology, hardware parallelism, and clock speed, out-of-order architectures have similar energy efficiency as in-order ones
- Diverse application power profiles suggest that applications and system software will need to participate in power optimization and management