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

Oracle SPARC M7 (2015)
32 cores; > 10B 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.
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 \( \frac{1}{\kappa} \), \( \kappa \approx \sqrt{2} \); \( \frac{1}{\kappa} \approx 0.7 \)

<table>
<thead>
<tr>
<th>Device or Circuit Parameter</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device dimension ( t_{ox}, L, W )</td>
<td>( \frac{1}{\kappa} )</td>
</tr>
<tr>
<td>Doping concentration ( N_a )</td>
<td>( \kappa )</td>
</tr>
<tr>
<td>Voltage ( V )</td>
<td>( \frac{1}{\kappa} )</td>
</tr>
<tr>
<td>Current ( I )</td>
<td>( \frac{1}{\kappa} )</td>
</tr>
<tr>
<td>Capacitance ( \varepsilon A/t )</td>
<td>( \frac{1}{\kappa} )</td>
</tr>
<tr>
<td>Delay time per circuit ( VC/I )</td>
<td>( \frac{1}{\kappa} )</td>
</tr>
<tr>
<td>Power dissipation per circuit ( VI )</td>
<td>( \frac{1}{\kappa^2} )</td>
</tr>
<tr>
<td>Power density ( VI/A )</td>
<td>1</td>
</tr>
</tbody>
</table>

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

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

**Transistor speed vs. microarchitecture**

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

Measure performance with SPEC INT 92, 95, 2000

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

Superscalar and OOO provided performance benefits at a cost in energy efficiency
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 because 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

![Graph showing power versus time]

- 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

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Logic Transistors (Millions)</th>
<th>Cache MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>50</td>
<td>6</td>
</tr>
<tr>
<td>2014</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>2018</td>
<td>150</td>
<td>80</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Large-Core Homogeneous</th>
<th>Small-Core Homogeneous</th>
<th>Hybrid approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large-core throughput</td>
<td>Pollack's Rule</td>
<td>Large-core throughput</td>
</tr>
<tr>
<td>Small-core throughput</td>
<td>(5/25)^0.5 = 0.45</td>
<td>Pollack's Rule</td>
</tr>
<tr>
<td>Total throughput</td>
<td>6</td>
<td>Total throughput</td>
</tr>
</tbody>
</table>

(a) Large-Core Homogeneous
(b) Small-Core Homogeneous
(c) Hybrid approach

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

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

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

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

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

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

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

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

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

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