Cross-Layer Power Optimization
Micro-architecture and Systems

BRUCEK KHAILANY
NVIDIA RESEARCH
Which of These are Power Constrained?

~1W  ~3W  ~100W  ~30W  1KW  100KW-20MW
Think about uses cases, not just pJ/op

- 18 kJ per day battery life (phone)
- Bursty perf at electrical limits
- Sustained perf at thermal limits

- Maximize **Perf@Watt** (TDP) at reasonable cost
- Energy-proportional computing
Technology Scaling

Classic Dennard Scaling
- 1/k Voltage and 1/k Capacitance
- \(2.8x\) capability in same power from 2x transistors and 1.4x frequency

Post Dennard Scaling
- \(~\text{Constant (slightly decreasing)}\) Voltage and 1/k Capacitance
- \(~\text{Constant frequency (maybe slightly increasing)}\)
- \(1.4x\) capability in same power
- 32x gap in 10 years compared to classic scaling with increasing freq
- 5.4x gap in 10 years compared to constant-freq scaling (2x capability/generation)

TABLE I
Scaling Results for Circuit Performance

<table>
<thead>
<tr>
<th>Device or Circuit Parameter</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device dimension (L_0, L, W)</td>
<td>(1/\kappa)</td>
</tr>
<tr>
<td>Doping concentration (N_a)</td>
<td>(\kappa)</td>
</tr>
<tr>
<td>Voltage (V)</td>
<td>(1/\kappa)</td>
</tr>
<tr>
<td>Current (I)</td>
<td>(1/\kappa)</td>
</tr>
<tr>
<td>Capacitance (eA/\ell)</td>
<td>(1/\kappa)</td>
</tr>
<tr>
<td>Delay time/circuit (V C/I)</td>
<td>(1/\kappa)</td>
</tr>
<tr>
<td>Power dissipation/circuit (VI)</td>
<td>(1/\kappa^2)</td>
</tr>
<tr>
<td>Power density (VI/A)</td>
<td>1</td>
</tr>
</tbody>
</table>

[Dennard et al, JSSC 1974]
Can micro-architecture & systems help us close the gap?
Opportunity #1: Physical Locality of Data

On-chip

- 64-bit DP FMA: 50pJ
- 256-bit access 8 kB SRAM: 56 pJ, 1.3 nJ
- 40nm technology

Signaling and Packaging

- 256b DRAM Rd/Wr: 10 nJ
- 256b Efficient off-chip link: 256b
- Micron Hybrid Memory Cube
- Silicon Interposer Packaging (Xilinx)

© NVIDIA 2012
Opportunity #2: Heterogeneous Architectures

**Today**
- Latency-optimized CPU cores: 1.7 nJ/flop
- Throughput-optimized cores: 225 pJ/flop (7.6x)

**Future**
- Latency cores: Energy efficiency doesn’t matter!
- Throughput cores: Optimized for parallel workloads
  - Will run the important parallel parts of the workloads
- Fixed-function HW: Wherever SW programmability is not required

Fixed-function HW:
Wherever SW programmability is not required
Opportunity #3: Run Slower @ Lower Voltage

- **Answer #1:** Hard – scale $V_{dd}$ down toward $2*V_t$

- **Answer #2:** Harder – scale down to near $V_t$

**Principal challenges**
- Slow transistors
- Reliability
- Process variation
Don’t ignore cost

Many optimizations trade off throughput/$ for energy efficiency

Throughput/mm$^2$
(Performance/$$)

Perf/W

Perf/$$ Degradation

Perf/Watt improvement
Summary

What not to do
- Stop researching latency-optimized cores optimized for single threads
- Don’t assume scaling Voltage below 2*Vt will work in most systems
- Don’t ignore cost

What to do
- Work on heterogeneity, throughput-optimized cores, data parallelism
- Focus on physical data locality
- Think about use cases, not just pJ/flop