Life with “dark” silicon.
Power and thermal problems in future platforms

Michael Kishinevsky
Strategic CAD Labs, Intel
Radboud University, Nijmegen, Netherlands
March 27, 2013
Outline

- Technology trends
- Power scaling and “dark” silicon
- Managing dark silicon
- Research challenges
Acknowledgements

- Raid Ayoub
- Umit Ogras
- Hyungjun Kim
- Steve Gunther
- Efi Rotem
Moore’s Law

Gordon Moore “Cramming more components onto integrated circuits”, Electronics Magazine 19 April 1965

“The complexity for minimum component costs has increased at a rate of roughly a factor of two per year...

Certainly over the short term this rate can be expected to continue, if not to increase.

Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years.

That means by 1975, the number of components per integrated circuit for minimum cost will be 65,000. I believe that such a large circuit can be built on a single wafer.”
## Intel Technology Roadmap

<table>
<thead>
<tr>
<th>Process Name</th>
<th>P1266</th>
<th>P1268</th>
<th>P1270</th>
<th>P1272</th>
<th>P1274</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Production</td>
<td>2007</td>
<td>2009</td>
<td>2011</td>
<td>2013</td>
<td>2015</td>
</tr>
<tr>
<td>Lithography</td>
<td>45 nm</td>
<td>32 nm</td>
<td>22 nm</td>
<td>14 nm</td>
<td>10 nm</td>
</tr>
</tbody>
</table>

Intel continues cadence of introducing a new technology generation every two years.
## Technology Outlook

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>High Volume Manufacturing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technology Node (nm)</td>
<td>45</td>
<td>32</td>
<td>22</td>
<td>15</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Integration Capacity (BT)</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
</tr>
<tr>
<td>Delay Scaling</td>
<td>&gt;0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy Scaling</td>
<td></td>
<td>~0.5</td>
<td></td>
<td></td>
<td></td>
<td>&gt;0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transistors</td>
<td>Planar</td>
<td>3G, FinFET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variability</td>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC Delay</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Metal Layers</td>
<td>8-9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Shekhar Borkar, Intel, 2010
Dennard 1978: Computing capabilities scales as $S^3$ every generation.
For $S \sim 1.4$ : $S^3 \sim 2.8$
Ideal Process Scaling. Constant Power

Dennard 1978: Computing capabilities scales as $S^3$ every generation
For $S \sim 1.4 : S^3 \sim 2.8$
Leakage Limited Process Scaling

- $S^3$
- $S^2$
- $S$

- $S$ faster transistors (x1.4)
- $S^2$ transistors (x2)
- S smaller capacitance (x 1/1.4)
- S smaller Vdd (x 1/$S^2$ = x 1/2)

1
Need in Low Power Computing

Power doubles every 4 years
5-year projection: 200W total, 125 W/cm²!

Fred Pollack, Intel, 1999, keynote at Micro
Utilization Wall

Every process generation the percentage of a chip that can actively switch drops exponentially due to power constraints. Under assumption: no radical process changes.

Experiment with a small datapath.

The remaining silicon that must be left unpowered is referred to as Dark Silicon.

[Mike Muller, ARM, 2009]
[Michael Taylor et al.]
Dark Silicon Apocalypse

Is Dark Silicon Useful?

Harnessing the Four Horsemen of the Coming Dark Silicon Apocalypse

Prof. Michael B. Taylor  UC San Diego

«Four Horsemen of the Apocalypse» (1887)

Viktor Vasnetsov

DAC 2012
Life with Dark Silicon

- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power management
- System-level optimization
22 nm Tri-Gate Transistor

Traditional 2-D planar transistor forms a conducting channel in the silicon region under the gate electrode when in the “on” state.

Used in Intel’s 32 nm technology

3-D Tri-Gate transistor forms conducting channels on three sides of a vertical fin structure, providing “fully depleted” operation.

Only 2-3% cost adder.

Used in Intel’s 22 nm technology
22 nm Tri-Gate Transistor

3-D Tri-Gate transistor can have multiple fins connected together to increase total drive strength for higher performance.

The fully depleted characteristics of Tri-Gate transistors provide a steeper sub-threshold slope that significantly reduces leakage current.

Steeper sub-threshold slope can be used to target lower threshold voltage, allowing transistor to operate at lower operating voltage.

Improved performance at high voltage and unprecedented performance gain at low voltage.

Can operate at lower voltage with same performance. Active power reduced by > 50%.
Life with Dark Silicon

- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power and Thermal management
- System-level optimization
Parallelism for power efficiency

\[ P_{\text{par}} = C_{\text{par}} V_{\text{par}}^2 f_{\text{par}} \]

\[ = (2.15C_{\text{ref}})(0.58V_{\text{ref}})^2 \left( \frac{f_{\text{ref}}}{2} \right) \approx 0.36 P_{\text{ref}}. \]

Fig. 7. A simple data path with corresponding layout.

Fig. 8. Parallel implementation of the simple data path.

Ideal voltage scaling
Ideal parallelization

Low-Power CMOS Digital Design
Anantha P. Chandrakasan, Samuel Sheng, and Robert W. Brodersen, Fellow, IEEE
Power and Area Constrained Performance

Power constraint is a primary limiter in the future.

Area constraints limiting the number of cores.

Power constrained performance caused transition to multi-cores.

Data partially collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond
Life with Dark Silicon

- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power and Thermal management
- System-level optimization
SoC. Collection of specialized block

- ASIC specialized block is an order more energy-efficient per unit of performance than universal CPU
  - Specialization: only good at one task
  - Specialized logic and shorter wires to deliver computation results

Medfield SOC Platform
Research on quick generation of accelerators (an example)

C-core
Generation

Code to Stylized Verilog and through a CAD flow.

Synopsys
IC Compiler,
P&R, CTS

0.01 mm² in 45 nm TSMC runs at 1.4 GHz
Life with Dark Silicon

- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power and Thermal management
- System-level optimization
Dimming

- Human Brain – very “dark” circuit
- 100 trillion synapses at 20W

- Lincroft SoC IREM
- 50x idle power reduction via power gating
Power management complexity. Intel® Turbo Boost™ Technology (Nehalem)

C6
Zero power for inactive cores

No Turbo

Workload Lightly Threaded or < TDP

Steve Gunther, Intel, 2010
Intel® Turbo Boost™ Technology (Nehalem)

No Turbo

Workload Lightly Threaded or < TDP

Power headroom converted to higher frequency

Steve Gunther, Intel, 2010
Intel® Turbo Boost™ Technology (Nehalem)

No Turbo

Power headroom converted to higher frequency

Workload Lightly Threaded or < TDP

Steve Gunther, Intel, 2010
Sandy Bridge Power Management

- 1-4 CPU cores + Gfx
- Integrated system agent
  - Package control unit
  - Memory controller
- Sliced LLC shared by cores/Gfx
- Ring interconnect + power management link
- PCU - control system
  - Logic and embedded controller running power and thermal management firmware
  - Monitors physical conditions (V, T, P)
  - Controls power states of CPU, Gfx (V,f)
  - Controls voltage regulators, DDR and system
- Excepts external inputs
  - System power management requests and limits
  - Power and temperature reading

Efi Rotem et al., Intel, 2011, HotChips
After idle periods, the system accumulates energy budget and can deliver higher power/performance for up to a minute. Used to enhance user experience.

In steady state power stabilizers on TDP

---

**Efi Rotem et al., Intel, 2011, HotChips**

Turbo Control Dynamics

P-state
Power

Hard Limit
Max Icc

Power limit 2
Power delivery

Power limit 1
Config TDP

Voltage Regulator reported capability
CURRENT_CONFIG_CONTROL MSR

Actual instantaneous power

Allow programmability

TURBO_POWER_LIMIT Control MSR
- Enables and locks
- Package Power limit 2 – Instantaneous
- Package Power limit 1 Time interval
- Package Power limit 1 clamp bit
- Package Power limit 1 - power

• Also:
  • Individual power controls available
  • Explicit frequency control

User / OEM / OS preference

C0 P0

Time

Power Limit 2

Power Limit 1

Time 0.001-64 Sec

PL1 time exp. average

Efi Rotem et al., Intel, 2011, HotChips
Sandy Bridge Package Thermal Management

- On-die thermal sensors: 12 on each CPU + Gfx + ring + system agent. Operating range 50-100°C
- Report max T per block and total chip
- Management: critical thermal protection (notification, throttle, shutdown)
- In addition used in power management for
  - Leakage calculation of power meter
  - PM optimization algorithms
  - External controls (e.g. fans)

Efi Rotem et al., Intel, 2011, HotChips
Hetero - extending DVFS. ARM’s Big.LITTLE

System

Big: 3-issue, out-of-order

LITTLE: 2-issue, in-order

Peter Greenhalgh, ARM, 2011
Big.LITTLE Power-performance Trade-off

Power benefit from Cortex-A7 at the same desired performance

Challenge: Migration of the application

Peter Greenhalgh, ARM, 2011
Research on Uncore Power Management

Cores & Local (L1,L2) Cache

Uncore Cache Control, MC, interconnect

Off-chip memory

Cores

C0
C1
Cx

Ring Interconnect

MC

PCIe*

*Not active

Ctrl: Cache controller/Home agent
MC: Integrated memory controller

Xi Chen, Hyungjun Kim et al. DAC 2013
Research on Uncore Power Management

Cores & Local (L1,L2) Cache

Uncore Cache Control, MC, interconnect

Off-chip memory

Cores

Requested Injection Rate

Core request

Uncore response

Throughput

Ring Interconnect

LLC Ctl

LLC Ctl

LLC Ctl

...

MC

DRAM

PCIe*

*Not active

C0

C1

C0

Ctl: Cache controller/Home agent
MC: Integrated memory controller

Xi Chen, Hyungjun Kim et al. DAC 2013
Workload Variation

- Injection rate changes significantly both
  - During the run-time of applications and
  - Across different applications

- Designing the uncore for the worst-case may cause the uncore operate faster than necessary for most of the time
Core Bandwidth-Throughput Characteristics

• **Linear region**: Almost proportional increase in throughput with increasing frequency
Core Bandwidth-Throughput Characteristics

• **Linear region**: Almost proportional increase in throughput with increasing frequency

• **Saturation due to requested injection rate**
Power/Performance Implications

• **Point 1**: Performance requirement is met, but power is wasted. Frequency can be lowered without sacrificing the throughput.
Power/Performance Implications

- **Point 1**: Performance requirement is met, but power is wasted. Frequency can be lowered without sacrificing the throughput.
- **Point 2**: Performance requirement is *not* met. Frequency *should* be increased.

Xi Chen, Hyungjun Kim et al. DAC 2013
Power/Performance Implications

- **Point 1**: Performance requirement is met, but power is wasted. Frequency can be lowered without sacrificing the throughput.
- **Point 2**: Performance requirement is *not* met. Frequency *should* be increased.
- **Point 3**: “Optimum” for this load.

Xi Chen, Hyungjun Kim et al. DAC 2013
• **Point 1:** Performance requirement is met, but power is wasted. Frequency can be lowered without sacrificing the throughput

• **Point 2:** Performance requirement is *not* met. Frequency *should* be increased

• **Point 3:** “Optimum” for this load. Optimum point changes dynamically!
Objective

Power Control Unit (PCU)

Find the minimum frequency, s.t. Throughput ≈ Request rate

Core request

Uncore response

measured request rate

measured throughput

Cores

C0

C1

Cx

Ring Interconnect

LLC Ctl

LLC Ctl

LLC Ctl

...
Improving Power Efficiency

- 31% average uncore power savings at minimal perf. cost of 0.6%
- Up to 73% power savings at 3.5% perf. cost
- Almost no effect on memory bound apps

Xi Chen, Hyungjun Kim et al. DAC 2013
Exascale computing

Goals:
Exascale performance: $10^{18}$ flops
Low energy: 20 pJ/flop (including moving data) -> 20MW per system

John Shalf et al. “Exascale Computing Technology Challenges” VECPAR 2010, LNCS 6449
UHPC experiment

Baseline: 720 nodes on die

Metric: power x latency / throughput = energy x delay/bit for networks (Jxsec/bit)

Balancing hierarchical interconnect

Intra-block network

Total system network with 10x6 2D mesh for inter-block network (12 nodes cluster) Uniform traffic
Evaluation on SAP from SPEC2000

24-node clusters more energy-delay efficient than 12-node
Life with Dark Silicon

- Process Technology Advances
- Parallelism
- Specialization
- Dimming = Power and Thermal management
- System-level optimization
Why system-level?

What matters is the final product
Need a system view: **SW stack + Platform + Silicon components**

March 2012
Android Power Management Stack

- **Applications hold “wakelocks”**
- **FM/HW**
- **Linux**
- **Android**

**Application Framework**
- App A
- App B
- App C

**Android Power Manager Services**

**User**

**Kernel**

**Android PM kernel**

**S3 Path** (no wake locks)

**Linux PM Suspend Framework**

**Idle Path – S0***

**Android PMU driver (S0*/S3)**

Applications hold "wakelocks"
System level power delivery

- Power delivery uses significant power

- **Case study:** Increasing CPU frequency-voltage leads to:
  - Higher **SoC** power
  - Higher **MSIC** power

- Ignoring inefficiency in power delivery leads to inaccurate power estimations

Power optimizations must be applied at the **system level**
Example of system level power breakdown

- CPU does not dominate the SoC
- SOC is not the biggest contributor of power
- MSIC dissipates significant portion of platform power
Power savings while using the phone

- Assume current task has a slack
  - **Utilize slack to optimize for execution speed**
    
    Which scenario is better for energy savings?
    - Execute fast and entertain a longer sleep
    - Run slow to reduce dynamic power at the cost of a shorter sleep
Power savings: run CPU faster vs. slower

- **Case 1:** Running a task at two frequencies, enter S3 when no execution (using wait function in Java app)
- Running the CPU faster increases its average power
- Still save at the platform level when CPU runs faster
  - Savings come from putting the platform in a deep system state for a longer time
**Power savings: run CPU faster vs. slower**

- **Case 2:** Running a task at two frequencies, always run in **S0** (long setting on display ON prevents entering S3)
- Running the CPU slower reduces its average power
- Running slower saves power at the platform level
  - Savings in this case come from utilizing DVFS to reduce dynamic power

---

**Finding optimal speed is challenging, requires system level optimizations**
Challenge: Temperature estimation and management

- **Current skin estimation models are off**
  - Challenging problem due to the complexity of thermal distribution in space and time and limitation in number of thermal sensors
  - How to place thermal sensors to maximize thermal observability?
  - Can we create accurate thermal models that are simple enough for run time estimations?

- **Temperature management**
  - Current thermal management techniques are reactive and heuristic
  - Better methods (e.g. MIMO control) are needed to ensure optimality
The model is able to track the temperature of skin and SoC’s die
Skin time constant is much larger than die
Temperature prediction

- Proactive thermal managements provide better thermal decisions than reactive*

- A temperature predictor* can be formulated using our thermal model

\[
T_{\text{skin}}(K+1) = T_{\text{skin}}(k) + (T_{\text{skin}}(k) - T_{\text{skin}}(k-1))e^{-\Delta t/\tau}
\]

\[
\tau = R_{\text{skin}}C_{\text{skin}}
\]

\(\Delta t\) : prediction distance

- Good prediction can be achieved

Challenge: System level power/energy optimizations under given constraints

- How to distribute power between components to maximize performance while meeting power/thermal constraints
  - Runtime optimizations at the system level
    - Low overhead
    - Power, performance and thermal models
    - Better observability of runtime characterization

- Dual problem: optimizing power/thermal under given performance constraints
Summary

- Modern silicon systems are “dark silicon”
  - Cannot operate all parts of the system at full speed simultaneously for a long time
  - This trend increases in the future systems

- Process technology continue delivering breakthrough improvements. However it is not sufficient alone

- Dimming technologies (thermal and power management) getting more elaborate
  - Applying control theory instead of heuristics seem promising
  - Hetero helps if can solve SW thread migration

- Specialization in the form of accelerators is used in all modern SoC
  - Higher energy efficiency
  - Automatic support for quick generation of accelerators and IP blocks is needed

- Need optimization at system-level (software, platform) not just SoC

- Shrinking die size without increase in functionality is an option, but not economically sustainable