Mobile Application Processors are where the action are

- The evolution of mobile application processors mirrors that of microprocessors mirrors that of mainframes ..

- Not much left to borrow!
- Faster iterative cycle than desktops:
  obsolete vs. worn out
- Mobile Application Processors are the new centroid of microprocessor evolution!

![Diagram showing the evolution of microprocessors from 1985 to 2015, comparing Intel and ARM technologies. The diagram highlights key microprocessor generations including 486, 586, 686, StrongARM, Core Duo, Cortex-A8, Cortex-A9, and MPCore.](image-url)
Desktop micros: frequency versus time

- 7 yr / 10x (39%)
- 5 yr / 10x (58%)
- 20 yr / 10x (12%)

Intel x86
Power Density

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

Watts/cm²

1000

100

10

1

Nuclear Reactor

Pentium® 4

Pentium® III

Pentium® II

Hot plate

Pentium® Pro

Pentium®

i386

i486

P=VI: 75W @ 1.5V = 50 A!

From “New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies”
Course Staff

Prof. Michael Taylor  
CSE 3202  
Office Hours: Right After Class  

Dr. Jack Sampson  
CSE 3204  
Office Hours: TBD
Where to find information:

*Wherever you can get it!*

*IEEE Explore* website, free from UCSD network
ISSCC (Chip Tapeout Papers)

*ACM Portal* website, free from UCSD network

*google.com* – I’m serious!
Powerpoint presentations by Qualcomm, Nvidia, Etc.
Commentary articles
Clever search terms
Reverse engineering sites (chipworks)
Microprocessor Report leaked articles

*anandtech.com* – Technical fanboy site

*fandroid.com etc* – I’m serious!

*qualcomm, ti, nvidia, google, arm* – developer sites and blogs
Project Structure

First Two Weeks:
- Literature Review
- Related Work
- Project Proposal

Last 6 Weeks
- Project
- Midpoint Review
- Final Project Report
  - 6 page DAC submission
  - DAC submission deadline shortly after class ends
- Final Project Presentation
  - Last Day of Class
NVidia Tegra 2: Example MAP
Discussion

- RISC vs CISC Battle, Round II
- Vertical Integration (Apple) versus Competitive Bazaar (Wintel/AndroidQualTISamidia)
- AMD vs. Intel
- Nvidia vs. Intel
- Nvidia vs. AMD/ATI

Mobile Monopoly Theory
- Qualcomm
  - (with AMD Adreno)
- Nvidia
  - (with Icera)
- TI
  - (historical, with OMAP)
- Apple
  - Soon enough

Mobile Snafus
- Intel sells Xscale to Marvl
- AMD sold Adreno Mobile GPU ($65m) to Qcom
GreenDroid: An Architecture for the Dark Silicon Era

UCSD Center for Dark Silicon
Department of Computer Science and Engineering,
University of California, San Diego
This Talk

The Dark Silicon Problem

How to use Dark Silicon to improve energy efficiency (Conservation Cores)

The GreenDroid Mobile Application Processor

GreenDroid Highlights
Where does dark silicon come from? And how dark is it going to be?

The Utilization Wall:

With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

[Venkatesh, Chakraborty]
We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

- Scaling theory
  - Transistor and power budgets are no longer balanced
  - Exponentially increasing problem!

- Experimental results
  - Replicated a small datapath
  - More "dark silicon" than active

- Observations in the wild
  - Flat frequency curve
  - "Turbo Mode"
  - Increasing cache/processor ratio
Scaling 101: Moore’s Law

90 65 45 32 22 16 11 8 nm

$S = \frac{22}{16} = \sim 1.4x$
Scaling 101: Transistors scale as $S^2$

- 180 nm
- 16 cores
- Transistors = 4x

- 90 nm
- 64 cores

MIT Raw

Tilera TILE64
Advanced Scaling:

Dennard: “Computing Capabilities Scale by $S^3 = 2.8x$”

If $S=1.4x$ …

Design of Ion-Implanted MOSFETs with Very Small Dimensions
Dennard et al, 1974
Advanced Scaling:

Dennard: “Computing Capabilities Scale by $S^3 = 2.8x$”

If $S = 1.4x$ ...

$S^2 = 2x$

More Transistors
Advanced Scaling:

**Dennard:** “Computing Capabilities Scale by $S^3 = 2.8x$”

If $S = 1.4x$ ...

- $S = 1.4x$
  - Faster Transistors

- $S^2 = 2x$
  - More Transistors
Advanced Scaling:

**Dennard:** “Computing Capabilities Scale by $S^3 = 2.8x$”

If $S = 1.4x$ …

- $S = 1.4x$
  - Faster Transistors

- $S^2 = 2x$
  - More Transistors

But wait: *switching 2.8x times as many transistors per unit time* – what about power??
Dennard: “We can keep power consumption constant”

\[ S = 1.4x \]

- Faster Transistors
- More Transistors

\[ S^2 = 2x \]

- Lower Capacitance
Dennard: “We can keep power consumption constant”

- $S = 1.4x$
  - Faster Transistors

- $S^2 = 2x$
  - More Transistors

- $S^3 = 2x$
  - Lower Capacitance

- Scale Vdd by $S=1.4x$
  - $S^2 = 2x$
Dennard: “We can keep power consumption constant”

- \( S = 1.4x \) Faster Transistors
- \( S^2 = 2x \) More Transistors
- Scale Vdd by \( S=1.4x \) \( S^2 = 2x \)
- \( S^3 \)
- \( S^2 \)
- \( S \)
- 1
Fast forward to 2005:

**Threshold Scaling Problems due to Leakage Prevents Us From Scaling Voltage**

- $S = 1.4x$
  - Faster Transistors

- $S^2 = 2x$
  - More Transistors

- $S^3$
  - $S = 1.4x$
    - Lower Capacitance

- $S^2$
  - Scale Vdd by $S=1.4x$
    - $S^2 = 2x$

- $S$

- 1
Fast forward to 2005:

Threshold Scaling Problems due to Leakage Prevents Us From Scaling Voltage

$S = 1.4x$
Faster Transistors

$S^2 = 2x$
More Transistors

Scale $V_{dd}$ by $S = 1.4x$

$S^2 = 2x$
Lower Capacitance

$S^3$

$S^2$

$S$

1

[Diagram with arrows indicating scaling factors and a red circle with a slash through a cake, indicating limitations.]
Full Chip, Full Frequency Power Dissipation Is increasing exponentially by 2x with every process generation

Factor of $S^2 = 2X$ shortage!!

[ASPLOS 2010, Venkatesh]
We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

- **Scaling theory**
  - Transistor and power budgets are no longer balanced
  - Exponentially increasing problem!

- **Experimental results**
  - Replicated a small datapath
  - More "dark silicon" than active

- **Observations in the wild**
  - Flat frequency curve
  - "Turbo Mode"
  - Increasing cache/processor ratio

### Classical scaling

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device count</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Device frequency</td>
<td>$S$</td>
</tr>
<tr>
<td>Device cap (power)</td>
<td>$1/S$</td>
</tr>
<tr>
<td>Device $V_{dd}$ (power)</td>
<td>$1/S^2$</td>
</tr>
</tbody>
</table>

### Leakage-limited scaling

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device count</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Device frequency</td>
<td>$S$</td>
</tr>
<tr>
<td>Device cap-&gt;power</td>
<td>$1/S$</td>
</tr>
<tr>
<td>Device $V_{dd}$ (power)</td>
<td>$\sim 1$</td>
</tr>
</tbody>
</table>

Utilization ?

1

$S^2$
## We've Hit The Utilization Wall

*Utilization Wall*: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

### Scaling theory
- Transistor and power budgets are no longer balanced
- Exponentially increasing problem!

### Experimental results
- Replicated a small datapath
- More "dark silicon" than active

### Observations in the wild
- Flat frequency curve
- "Turbo Mode"
- Increasing cache/processor ratio

### Classical scaling
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device count</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Device frequency</td>
<td>$S$</td>
</tr>
<tr>
<td>Device cap (power)</td>
<td>$1/S$</td>
</tr>
<tr>
<td>Device $V_{dd}$ (power)</td>
<td>$1/S^2$</td>
</tr>
<tr>
<td>Utilization ?</td>
<td>$1$</td>
</tr>
</tbody>
</table>

### Leakage-limited scaling
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device count</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Device frequency</td>
<td>$S$</td>
</tr>
<tr>
<td>Device cap-&gt;power</td>
<td>$1/S$</td>
</tr>
<tr>
<td>Device $V_{dd}$ (power)</td>
<td>$\sim 1$</td>
</tr>
<tr>
<td>Utilization ?</td>
<td>$1/S^2$</td>
</tr>
</tbody>
</table>
We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

- **Scaling theory**
  - Transistor and power budgets are no longer balanced
  - Exponentially increasing problem!

- **Experimental results**
  - Replicated a small datapath
  - More "dark silicon" than active

- **Observations in the wild**
  - Flat frequency curve
  - "Turbo Mode"
  - Increasing cache/processor ratio

![Graph showing expected utilization for fixed area and power budget with process generations from 90 nm to 32 nm, with a 2x decrease in utilization at each transition.]
We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

- Scaling theory
  - Transistor and power budgets are no longer balanced
  - Exponentially increasing problem!

- Experimental results
  - Replicated a small datapath
  - More "dark silicon" than active

- Observations in the wild
  - Flat frequency curve
  - "Turbo Mode"
  - Increasing cache/processor ratio
We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

- **Scaling theory**
  - Transistor and power budgets are no longer balanced
  - Exponentially increasing problem!

- **Experimental results**
  - Replicated a small datapath
  - More "dark silicon" than active

- **Observations in the wild**
  - Flat frequency curve
  - "Turbo Mode"
  - Increasing cache/processor ratio
We've Hit The Utilization Wall

Utilization Wall: With each successive process generation, the percentage of a chip that can actively switch drops exponentially due to power constraints.

- Scaling theory
  - Transistor and power budgets are no longer balanced
  - Exponentially increasing problem!

- Experimental results
  - Replicated a small datapath
  - More "dark silicon" than active

- Observations in the wild
  - Flat frequency curve
  - "Turbo Mode"
  - Increasing cache/processor ratio

The utilization wall will change the way everyone builds processors.
Utilization Wall: Dark Silicon Leads to Dark Implications for Multicore

Spectrum of tradeoffs between # of cores and frequency

Example:
65 nm \(\rightarrow\) 32 nm (S = 2)

4 cores @ 1.8 GHz

2x4 cores @ 1.8 GHz
(8 cores dark, 8 dim)

(Industry’s Choice, next slide)

4 cores @ 2x1.8 GHz
(12 cores dark)

65 nm

32 nm

[Hotchips 2010]
Utilization Wall in “Real Life”
45 nm → 32 nm (S=1.4x)

Nehalem
3.2 GHz
4 cores
120 W

Gulftown
3.3 GHz
6 cores
120 W

S=1.4x
\Delta \text{cores} = 1.5x
\Delta \text{freq} = 1.03x

As predicted by the utilization wall, compute capability scaled as S and not as S^3

Improvements in performance are gated by improvements in energy efficiency, which primarily result from decreases in capacitance, not improvements in frequency or area.