Lecture 1

Design and Technology Trends

R. Saleh
Dept. of ECE
University of British Columbia
res@ece.ubc.ca
Recently Designed Chips

- Itanium chip (Intel), 2B tx, 700mm$^2$, 8 layer 65nm CMOS (4 processors)
- TILE64 Processor, 64-Core SoC with Mesh NoC Interconnect, 90nm CMOS
- 153Mb-SRAM (Intel), 45nm, high-k metal-gate CMOS
- FPGAs recently fabricated in 45nm

- What are the major technology and design issues that are driving the IC industry?
  
  Let’s start from the simple rules of MOS scaling…
MOS Transistor Scaling
(1974 to present)

Scaling factor $s=0.7$ per node (0.5x per 2 nodes)

- **Metal pitch**
- **Technology Node set by 1/2 pitch**
  (interconnect)
- **Poly width**
- **Gate length**
  (transistor)
### Ideal Technology Scaling (constant field)

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Before Scaling</th>
<th>After Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel Length</td>
<td>L</td>
<td>L’ = L * s</td>
</tr>
<tr>
<td>Channel Width</td>
<td>W</td>
<td>W’ = W * s</td>
</tr>
<tr>
<td>Gate Oxide thickness</td>
<td>t_{ox}</td>
<td>t’<em>{ox} = t</em>{ox} * s</td>
</tr>
<tr>
<td>Junction depth</td>
<td>x_{j}</td>
<td>x’<em>{j} = x</em>{j} * s</td>
</tr>
<tr>
<td>Power Supply</td>
<td>V_{dd}</td>
<td>V’_{dd} = Vdd * s</td>
</tr>
<tr>
<td>Threshold Voltage</td>
<td>V_{th}</td>
<td>V’<em>{th} = V</em>{th} * s</td>
</tr>
<tr>
<td>Doping Density, p n+</td>
<td>N_A</td>
<td>N_A’ = N_A / s</td>
</tr>
<tr>
<td></td>
<td>N_D</td>
<td>N_D’ = N_D / s</td>
</tr>
</tbody>
</table>
Technology Nodes 1999-2019

Two year cycle between nodes until 2001, then 3 year cycle begins.
## Forecast Technology Parameters

<table>
<thead>
<tr>
<th>Year</th>
<th>Technology Node (nm)</th>
<th>Physical Gate (nm)</th>
<th>tox (nm)</th>
<th>Dielectric K</th>
<th>Vdd (V)</th>
<th>Vth (V)</th>
<th>Na (/cm$^3$)</th>
<th>Nd (/cm$^3$)</th>
<th>xj (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>130</td>
<td>90</td>
<td>3.0</td>
<td>3.7</td>
<td>1.2</td>
<td>0.34</td>
<td>1.0e16</td>
<td>1.0e19</td>
<td>67.5</td>
</tr>
<tr>
<td>2004</td>
<td>90</td>
<td>53</td>
<td>2.4</td>
<td>3.0</td>
<td>1.1</td>
<td>0.32</td>
<td>1.4e16</td>
<td>1.4e19</td>
<td>46.7</td>
</tr>
<tr>
<td>2007</td>
<td>65</td>
<td>32</td>
<td>1.7</td>
<td>2.5</td>
<td>0.9</td>
<td>0.29</td>
<td>2.0e16</td>
<td>2.0e19</td>
<td>33.8</td>
</tr>
<tr>
<td>2010</td>
<td>45</td>
<td>22</td>
<td>1.5</td>
<td>2.0</td>
<td>0.8</td>
<td>0.29</td>
<td>2.9e16</td>
<td>2.9e19</td>
<td>23.4</td>
</tr>
<tr>
<td>2013</td>
<td>32</td>
<td>16</td>
<td>1.4</td>
<td>1.9</td>
<td>0.7</td>
<td>0.25</td>
<td>4.0e16</td>
<td>4.0e19</td>
<td>16.6</td>
</tr>
<tr>
<td>2016</td>
<td>22</td>
<td>11</td>
<td>1.3</td>
<td>1.7</td>
<td>0.6</td>
<td>0.22</td>
<td>5.9e16</td>
<td>5.9e19</td>
<td>11.4</td>
</tr>
</tbody>
</table>
Where are we now?

- 130nm and 90nm CMOS volume production
- Early production of 65nm, Leading-edge use of 45nm
- Scaling of gate is leading scaling of wire
- Scaling is driven by DIGITAL design needs
Making Photolithograph Work

- Extensive use of OPC and PSM in 90nm and below:

Optical Proximity Correction (OPC)
- Add scattering features to sharpen corners
- Used extensively for poly gate definition

Phase Shift Masking (PSM)
- Modulate optical path through mask
- Used extensively for contacts & vias
- Complicated for irregular patterns

Source: Socha, ASML (2004)

# Deep Submicron Technology Generations

## Table 1: Time overlap of semiconductor generations

<table>
<thead>
<tr>
<th></th>
<th>95</th>
<th>96</th>
<th>97</th>
<th>98</th>
<th>99</th>
<th>00</th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>08</th>
<th>09</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>350 nm</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>-1</td>
<td>250 nm</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>180 nm</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>130 nm</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-9</td>
<td>-8</td>
<td>-7</td>
<td>-6</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>90 nm</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-11</td>
<td>-10</td>
<td>-9</td>
<td>-8</td>
<td>-7</td>
<td>-6</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>65 nm</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-11</td>
<td>-10</td>
<td>-9</td>
<td>-8</td>
<td>-7</td>
<td>-6</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>45 nm</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each generation spans ~17 years...we are unlikely to be totally surprised
MPU Trends - Moore’s Law

Transistors Double
Every Two Years

2X Growth
in 2 Years!

Source: Intel
More MPU Trends

~40mm Die in 2010?

~7% growth per year
~2X growth in 10 years

Source: Intel
Delay Metric - FO4 Concept

where $\gamma$ is ratio of Parasitic output Capacitance to gate capacitance

Use FO4 delay as optimal delay
For scaling purposes, the alpha-power model is very useful:

\[ I_{\text{dsat}} = K W L_{\text{eff}}^{-0.5} T_{\text{ox}}^{-0.8} (V_{\text{gs}} - V_{\text{th}})^{1.25} \]

If \( L, T_{\text{ox}}, V \) all scale (note \( V \) scaling will be limited by \( V_{\text{th}} \) scaling),

Current should remain constant per micron of width (approx. 600 to 800uA/\( \mu \)m)

\[ \Delta t' = CV/i = s\Delta t \text{ since } C, V, i \text{ all scale down by } s \]

**Fanout = 4** inverter delay at TT, 90% Vdd, 125 °C
MPU Clock Frequency Trend

Intel: Borkar/Parkhurst

- 80386
- 80486
- Pentium
- Pentium II

MHz

Dec-83 Dec-86 Dec-89 Dec-92 Dec-95 Dec-98
MPU Clock Frequency Trend

Forward projection may be too optimistic

Intel: Borkar/Parkhurst
MPU Clock Cycle Trend (FO4 Delays)

Intel: Borkar/Parkhurst

- 80386
- 80486
- Pentium
- Pentium II
MPU Clock Cycle Trend (FO4 Delays)

Forward projection does not make sense

Curve actually flattens at 14-16 FO4

Intel: Borkar/Parkhurst
Power Trend - Ever Increasing

![Chart showing power trend with markers for MPU and DSP, showing an increase of 4 times every 3 years and 1.4 times every 3 years, with data published in ISSCC.]
Dynamic vs. Leakage Power

Krishnamurthy, et al., CICC 2002
Leakage Current Contributions

- Relative contributions of OFF-state leakage (but magnitude of total leakage getting exponentially worse for deeper submicron nodes)

<table>
<thead>
<tr>
<th>130nm</th>
<th>90nm</th>
<th>65nm</th>
</tr>
</thead>
</table>

- $I_{SUB}$: Subthreshold leakage from source
- $I_{GIDL}$: Gate-induced drain leakage (GIDL)
- $I_J$: Junction reverse-bias leakage
- $I_G$: Gate leakage (direct tunneling)

MPU Diminishing Returns

- **Power knob running out**
  - Speed == Power
  - 10W/cm² limit for convection cooling, 50W/cm² limit for forced-air cooling
  - Large currents, large power surges on wakeup
  - Cf. 125A supply current, 150W total power at 1.2V Vdd for EV8 (Compaq)
  - Die size will not continue to increase unless more memory is used to occupy the additional area
  - Additional power dissipation coming from subthreshold leakage

- **Speed knob running out**
  - Historically, 2x clock frequency every process generation
    - 1.4x from device scaling
    - 1.4x from pipelining, hence fewer logic stages (from 40-100 down to around 16 FO4 INV delays)
  - Clocks cannot be generated with period < 6-8 FO4 INV delays
  - Around 14-16 FO4 INV delays is limit for clock period
  - Unrealistic to continue 2x frequency trend!
Low-Power Design Techniques

- Supply Voltage Scaling
- Frequency Scaling
- Multiple Supply Voltages (Voltage Islands)
- Clock Gating
- Power Gating
- Multiple Threshold Voltages: LVT, SVT, HVT
- Substrate Biasing
- Power Shut Off
- HW/SW Power Management
Low-Power Application: PDA

0.18um / 400MHz / 470mW (typical)

MM Application
- MP3
- JPEG
- Simple Moving Picture

Available Time
- 6-10Hr

Peripheral Area
- 4 – 48MHz

Processor Area
- 6.5MTrs.
- Max 400MHz

Data Transfer Area
- 100MHz

- I-cache 32KB
- D-cache 32KB
- CPU
- DMA controller
- MEMCnt.
- LCDCnt.
- SDRAM 64MB
- Flash 32MB
- LCD
- PWM
- RTC
- FICP
- SSP
- I2C
- GPIO
- USB
- OST
- MMC
- UART
- AC97
- GPIO
- SSP
- PWM
- RTC
- FICP
- SSP
- I2C
- GPIO
- USB
- OST
- MMC
- UART
- AC97

Sound
USB
MMC
KEY
Trends in Low-Power Design Content

• Today, such designs contain embedded processing engines such as CPU and DSP, and memory blocks such as SRAM and embedded DRAM
• As we scale technology and keep power constant how does the amount of logic vs. memory change?
• Consider the following assumptions to develop trends for on-chip logic/memory percentages
  • Die size is $100\text{mm}^2$
  • Clock frequency starts at 150MHz increases by about 40% per technology node
  • Average power dissipation in limited to 100mW at 100°C
  • Initial condition at Year 2001: area percentage 75% logic, 25% memory
Logic/Memory Content Trend

![Graph showing the trend of Logic Area Contribution (%) and Total Memory Area (%) with years from 2001 to 2016. Die Size = 1cm²]

Die Size = 1cm²
ASIC Logic/Memory Content Trends

- Source: Dataquest (2001)
Design Trend: Productivity Gap

<table>
<thead>
<tr>
<th>Year</th>
<th>Technology</th>
<th>Chip Complexity</th>
<th>ASIC Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>250 nm</td>
<td>50M Tr.</td>
<td>100MHz</td>
</tr>
<tr>
<td>1999</td>
<td>180 nm</td>
<td>150M Tr.</td>
<td>200MHz</td>
</tr>
<tr>
<td>2002</td>
<td>130 nm</td>
<td>250M Tr.</td>
<td>400MHz</td>
</tr>
<tr>
<td>2004</td>
<td>90 nm</td>
<td>500M Tr.</td>
<td>600MHz</td>
</tr>
</tbody>
</table>
Designing a 50M Transistor IC

- Gates Required ~12.5M
- Gates/Day (Verified) 1K (including memory)
- Total Eng. Days 12,500
- Total Eng. Years 35
- Cost/Eng./Year $200K
- Total People Cost $7M
- Other costs (masks, tools, etc.) $8M

Actual Cost is $10-15M to get actual prototypes after fabrication.
Productivity Gap

- Deep submicron (DSM) technology allows hundreds of millions of transistors to be integrated on a single chip
- Number of transistors that a designer can design per day (~1000 gates/day) is not going up significantly
- New design methodologies are needed to address the integration/productivity issues

⇒ “System on a chip” Design with reusable IP
   - new design methodology, IP development
   - new HW/SW design and verification issues
   - new test issues
SoC Design Hierarchy

SOC consists of new logic blocks and existing IP

- New Logic blocks
- Existing IP including memory

Each logic block can be implemented by newly designed portion and a re-use portion based on IPs

- Newly designed portion
- Re-use portion including memory
SoC Platform Design Concept

Pre-Qualified/Verified Foundation-IP*

Hardware IP

SWIP

Programmable IP

Foundation Block + Reference Design

Scaleable bus, test, power, IO, clock, timing architectures

Processor(s), RTOS(es) and SW architecture

Methodology / Flows:
System-level performance evaluation environment
HW/SW Co-synthesis
SoC IC Design Flows
SoC Verification Flow
System-Level Performance Evaluation
Rapid Prototype for End-Customer Evaluation
SoC Derivative Design Methodologies

*IP can be hardware (digital or analog) or software. IP can be hard, soft or ‘firm’ (HW), source or object (SW)
Purpose of this Course

- This course addresses SoC/IP design in DSM technologies
- It is a very broad subject, one that industry is grappling with on a daily basis – one course cannot address all the issue properly
- The goal is to present an overview of the various issues from “Systems to Silicon” to provide a perspective on what is happening in technology and design.
- We will begin with the Systems Level and work our way down to the Silicon Level
- The projects, presentations, and assignments will provide in-depth analysis of the subjects that are of interest to you