Chip Making

&

Microprocessor Technology

Donald P. Greenberg

Lecture #2

August 31, 2015
• **Required Reading:**
  
  
Turing Machine

Alan Turing
Cloud Computing - 2010
Microprocessor Transistor Counts 1971-2011 & Moore’s Law

- IBM 360 Model 75 (1965)
- VAX-11/780 (1980)
- IBM 360 Model 75
- Cray T3E (1995)

Curve shows ‘Moore’s Law’: transistor count doubling every two years.
Transistor
Shockley & the Traitorous Eight

• William Shockley - Receives the Nobel Prize in Physics with Bardeen and Brattain (1956) leaves Bell Laboratory and forms Fairchild Semiconductor

• Julius Blank - founded Xicor

• Jean Hoerni - invented the planar process founded Amelco → Teledyne

• Jay Last - founded Amelco → Teledyne

• Sheldon Roberts - founded Amelco → Teledyne
Shockley & the Traitorous Eight

- Gordon Moore - founded Intel in 1968
- Robert Noyce - founded Intel in 1968
- Eugene Kleiner - founded Kleiner-Perkins
- Victor Grinich - only a poor professor at UC Berkeley & Stanford
From Sand to Silicon – Manufacturing an Integrated Circuit

Scientific American: The Solid-State Century, Special Issue 1998
Chip Design
Silicon Crystal
Layering
Masking & Etching
Interconnections & Dicing
Probing Electrical Connections
Dicing
Chip Selection
## International Technology Roadmap for Semiconductors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>130nm</td>
<td>90nm</td>
<td>65nm</td>
<td>45nm</td>
<td>32nm</td>
<td>22nm</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>193</td>
<td>386</td>
<td>1546</td>
<td>3092</td>
<td>6184</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clock Speed (Ghz)</th>
<th>2001</th>
<th>2004</th>
<th>2007</th>
<th>2010</th>
<th>2013</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5Ghz</td>
<td>4.1Ghz</td>
<td>9.3Ghz</td>
<td>15Ghz</td>
<td>23Ghz</td>
<td>40Ghz</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>200mm</td>
<td>300mm</td>
<td>300mm</td>
<td>300mm</td>
<td>450mm</td>
<td>450mm</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>140 mm²</td>
<td>140 mm²</td>
<td>140 mm²</td>
<td>140 mm²</td>
<td>140 mm²</td>
<td>140 mm²</td>
<td>140 mm²</td>
</tr>
</tbody>
</table>

Roughly 0.5 shrink every 3 years. Intel released 22 nm chips in 2013.
## Moore’s Law – CPU Transistor Counts

<table>
<thead>
<tr>
<th>Processor</th>
<th>Transistor count</th>
<th>Date of introduction</th>
<th>Manufacturer</th>
<th>Process</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel 4004</td>
<td>2,300</td>
<td>1971</td>
<td>Intel</td>
<td>10 µm</td>
<td></td>
</tr>
<tr>
<td>Intel 8008</td>
<td>3,500</td>
<td>1972</td>
<td>Intel</td>
<td>10 µm</td>
<td></td>
</tr>
<tr>
<td>Intel 8080</td>
<td>4,500</td>
<td>1974</td>
<td>Intel</td>
<td>6 µm</td>
<td></td>
</tr>
<tr>
<td>Intel 8088</td>
<td>29,000</td>
<td>1979</td>
<td>Intel</td>
<td>3 µm</td>
<td></td>
</tr>
<tr>
<td>Intel 80286</td>
<td>134,000</td>
<td>1982</td>
<td>Intel</td>
<td>1.5 µm</td>
<td></td>
</tr>
<tr>
<td>Intel 80386</td>
<td>275,000</td>
<td>1985</td>
<td>Intel</td>
<td>1.5 µm</td>
<td></td>
</tr>
<tr>
<td>Intel 80486</td>
<td>1,180,000</td>
<td>1989</td>
<td>Intel</td>
<td>1 µm</td>
<td></td>
</tr>
<tr>
<td>Pentium</td>
<td>3,100,000</td>
<td>1993</td>
<td>Intel</td>
<td>0.8 µm</td>
<td></td>
</tr>
<tr>
<td>AMD K5</td>
<td>4,300,000</td>
<td>1996</td>
<td>AMD</td>
<td>0.5 µm</td>
<td></td>
</tr>
<tr>
<td>Pentium II</td>
<td>7,500,000</td>
<td>1997</td>
<td>Intel</td>
<td>0.35 µm</td>
<td></td>
</tr>
<tr>
<td>AMD K6</td>
<td>8,800,000</td>
<td>1997</td>
<td>AMD</td>
<td>0.35 µm</td>
<td></td>
</tr>
<tr>
<td>Pentium III</td>
<td>9,500,000</td>
<td>1999</td>
<td>Intel</td>
<td>0.25 µm</td>
<td></td>
</tr>
<tr>
<td>AMD K6-III</td>
<td>21,300,000</td>
<td>1999</td>
<td>AMD</td>
<td>0.25 µm</td>
<td></td>
</tr>
<tr>
<td>AMD K7</td>
<td>22,000,000</td>
<td>1999</td>
<td>AMD</td>
<td>0.25 µm</td>
<td></td>
</tr>
<tr>
<td>Pentium 4</td>
<td>42,000,000</td>
<td>2000</td>
<td>Intel</td>
<td>180 nm</td>
<td></td>
</tr>
<tr>
<td>Atom</td>
<td>47,000,000</td>
<td>2008</td>
<td>Intel</td>
<td>45 nm</td>
<td></td>
</tr>
<tr>
<td>Barton</td>
<td>54,300,000</td>
<td>2003</td>
<td>AMD</td>
<td>130 nm</td>
<td></td>
</tr>
<tr>
<td>AMD K8</td>
<td>105,900,000</td>
<td>2003</td>
<td>AMD</td>
<td>130 nm</td>
<td></td>
</tr>
<tr>
<td>Itanium 2</td>
<td>220,000,000</td>
<td>2003</td>
<td>Intel</td>
<td>130 nm</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Processor</th>
<th>Transistor count</th>
<th>Date of introduction</th>
<th>Manufacturer</th>
<th>Process</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core 2 Duo</td>
<td>291,000,000</td>
<td>2006</td>
<td>Intel</td>
<td>65 nm</td>
<td></td>
</tr>
<tr>
<td>AMD K10</td>
<td>463,000,000</td>
<td>2007</td>
<td>AMD</td>
<td>65 nm</td>
<td></td>
</tr>
<tr>
<td>AMD K10</td>
<td>758,000,000</td>
<td>2008</td>
<td>AMD</td>
<td>45 nm</td>
<td></td>
</tr>
<tr>
<td>Itanium 2 with 9MB cache</td>
<td>592,000,000</td>
<td>2004</td>
<td>Intel</td>
<td>130 nm</td>
<td></td>
</tr>
<tr>
<td>Core i7 (Quad)</td>
<td>731,000,000</td>
<td>2008</td>
<td>Intel</td>
<td>45 nm</td>
<td>263 mm²</td>
</tr>
<tr>
<td>POWER6</td>
<td>789,000,000</td>
<td>2007</td>
<td>IBM</td>
<td>65 nm</td>
<td>341 mm²</td>
</tr>
<tr>
<td>Six-Core Opteron 2400</td>
<td>904,000,000</td>
<td>2009</td>
<td>AMD</td>
<td>45 nm</td>
<td></td>
</tr>
<tr>
<td>Six-Core Core i7</td>
<td>1,170,000,000</td>
<td>2010</td>
<td>Intel</td>
<td>32 nm</td>
<td></td>
</tr>
<tr>
<td>Dual-Core Itanium 2</td>
<td>1,700,000,000</td>
<td>2006</td>
<td>Intel</td>
<td>90 nm</td>
<td>596 mm²</td>
</tr>
<tr>
<td>Six-Core Xeon 7400</td>
<td>1,900,000,000</td>
<td>2008</td>
<td>Intel</td>
<td>45 nm</td>
<td></td>
</tr>
<tr>
<td>Quad-Core Itanium Tukwila</td>
<td>2,000,000,000</td>
<td>2010</td>
<td>Intel</td>
<td>65 nm</td>
<td></td>
</tr>
<tr>
<td>Six-Core Core i7 (Sandy Bridge-E)</td>
<td>2,270,000,000</td>
<td>2011</td>
<td>Intel</td>
<td>32 nm</td>
<td>434 mm²</td>
</tr>
<tr>
<td>8-Core Xeon Nehalem-EX</td>
<td>2,300,000,000</td>
<td>2010</td>
<td>Intel</td>
<td>45 nm</td>
<td>684 mm²</td>
</tr>
<tr>
<td>10-Core Xeon Westmere-EX</td>
<td>2,600,000,000</td>
<td>2011</td>
<td>Intel</td>
<td>32 nm</td>
<td>512 mm²</td>
</tr>
<tr>
<td>Six-core zEC12</td>
<td>2,750,000,000</td>
<td>2012</td>
<td>IBM</td>
<td>32 nm</td>
<td>597 mm²</td>
</tr>
<tr>
<td>8-Core Itanium Poulson</td>
<td>3,100,000,000</td>
<td>2012</td>
<td>Intel</td>
<td>32 nm</td>
<td>544 mm²</td>
</tr>
<tr>
<td>15-Core Xeon Ivy Bridge-EX</td>
<td>4,310,000,000</td>
<td>2014</td>
<td>Intel</td>
<td>22 nm</td>
<td>541 mm²</td>
</tr>
<tr>
<td>62-Core Xeon Phi</td>
<td>5,000,000,000</td>
<td>2012</td>
<td>Intel</td>
<td>22 nm</td>
<td>350 mm²</td>
</tr>
<tr>
<td>Xbox One Main SoC</td>
<td>5,000,000,000</td>
<td>2013</td>
<td>Microsoft</td>
<td>28 nm</td>
<td>363 mm²</td>
</tr>
<tr>
<td>18-core Xeon Haswell-E5</td>
<td>5,560,000,000</td>
<td>2014</td>
<td>Intel</td>
<td>22 nm</td>
<td>661 mm²</td>
</tr>
<tr>
<td>IBM z13 Storage Controller</td>
<td>7,100,000,000</td>
<td>2015</td>
<td>IBM</td>
<td>22 nm</td>
<td>678 mm²</td>
</tr>
</tbody>
</table>
## Moore’s Law – GPU Transistor Counts

<table>
<thead>
<tr>
<th>Processor</th>
<th>Transistor count</th>
<th>Date of introduction</th>
<th>Manufacturer</th>
<th>Process</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>R520</td>
<td>321,000,000</td>
<td>2005</td>
<td>AMD</td>
<td>90 nm</td>
<td>288 mm²</td>
</tr>
<tr>
<td>R580</td>
<td>384,000,000</td>
<td>2006</td>
<td>AMD</td>
<td>90 nm</td>
<td>352 mm²</td>
</tr>
<tr>
<td>G80</td>
<td>681,000,000</td>
<td>2006</td>
<td>NVIDIA</td>
<td>90 nm</td>
<td>480 mm²</td>
</tr>
<tr>
<td>R600 Pele</td>
<td>700,000,000</td>
<td>2007</td>
<td>AMD</td>
<td>80 nm</td>
<td>420 mm²</td>
</tr>
<tr>
<td>G92</td>
<td>754,000,000</td>
<td>2007</td>
<td>NVIDIA</td>
<td>65 nm</td>
<td>324 mm²</td>
</tr>
<tr>
<td>RV790XT Spartan</td>
<td>959,000,000</td>
<td>2008</td>
<td>AMD</td>
<td>55 nm</td>
<td>282 mm²</td>
</tr>
<tr>
<td>GT200 Tesla</td>
<td>1,400,000,000</td>
<td>2008</td>
<td>NVIDIA</td>
<td>65 nm</td>
<td>576 mm²</td>
</tr>
<tr>
<td>Cypress RV870</td>
<td>2,154,000,000</td>
<td>2009</td>
<td>AMD</td>
<td>40 nm</td>
<td>334 mm²</td>
</tr>
<tr>
<td>Cayman RV970</td>
<td>2,640,000,000</td>
<td>2010</td>
<td>AMD</td>
<td>40 nm</td>
<td>389 mm²</td>
</tr>
<tr>
<td>GF100 Fermi</td>
<td>3,200,000,000</td>
<td>Mar 2010</td>
<td>NVIDIA</td>
<td>40 nm</td>
<td>526 mm²</td>
</tr>
<tr>
<td>GF110 Fermi</td>
<td>3,000,000,000</td>
<td>Nov 2010</td>
<td>NVIDIA</td>
<td>40 nm</td>
<td>520 mm²</td>
</tr>
<tr>
<td>GK104 Kepler</td>
<td>3,540,000,000</td>
<td>2012</td>
<td>NVIDIA</td>
<td>28 nm</td>
<td>294 mm²</td>
</tr>
<tr>
<td>Tahiti RV1070</td>
<td>4,312,711,873</td>
<td>2011</td>
<td>AMD</td>
<td>28 nm</td>
<td>365 mm²</td>
</tr>
<tr>
<td>GK110 Kepler</td>
<td>7,080,000,000</td>
<td>2012</td>
<td>NVIDIA</td>
<td>28 nm</td>
<td>561 mm²</td>
</tr>
<tr>
<td>RV1090 Hawaii</td>
<td>6,300,000,000</td>
<td>2013</td>
<td>AMD</td>
<td>28 nm</td>
<td>438 mm²</td>
</tr>
<tr>
<td>GM204 Maxwell</td>
<td>5,200,000,000</td>
<td>2014</td>
<td>NVIDIA</td>
<td>28 nm</td>
<td>398 mm²</td>
</tr>
<tr>
<td>GM200 Maxwell</td>
<td>8,100,000,000</td>
<td>2015</td>
<td>NVIDIA</td>
<td>28 nm</td>
<td>601 mm²</td>
</tr>
<tr>
<td>Fiji</td>
<td>8,900,000,000</td>
<td>2015</td>
<td>AMD</td>
<td>28 nm</td>
<td>596 mm²</td>
</tr>
</tbody>
</table>
2007

Paul S. Otellini
Intel Corporation’s fifth CEO
Why are we continuing to strive for smaller and smaller technology?

- More transistors/chip $\rightarrow$ increased functionality and performance

- Higher speeds $\rightarrow$ partially depends on how close together the components are placed

- Cheaper – more chips/wafer, greater yields
Yield Ratio

\[ yield = \frac{n_w}{n_t} \]

\[ n_w = yield \cdot n_t \]

\( n_w = \text{number of working chips/wafer} \)

\( n_t = \text{total number of chips/wafer} \)

Old fab lines, yield \( \rightarrow 90\% \)

New fab lines, yield \( \rightarrow < 40\% \)
Yield per Wafer
Yield Ratio

Number of defects/unit area depends on the process

\[
\frac{1}{\text{Chip Area}}
\]

\( \therefore \text{Yield} \approx \frac{1}{\text{Chip Area}} \)

Total chips \((n_t)\) for a given wafer size is also inversely proportional to the chip area
Why does the shrinking technology make the cost of manufacturing cheaper per component?
Example:

For a 10% shrink in feature size:

\[
\frac{n_{\text{new}}}{n_{\text{old}}} = \left(\frac{1}{0.9}\right)^2 \left(\frac{1}{0.9}\right)^2
\]

\[
\uparrow \quad \uparrow
\]

New yield \quad New n_t

\[
\frac{n_{\text{new}}}{n_{\text{old}}} = 1.52 \frac{n_{\text{old}}}{n_{\text{old}}}
\]
Can this Shrinking Technology Continue?
Human Hair

Integrated Chip

Visible Light

50µm (50,000 nm)

400nm

700nm

32nm
Getting Wafers Wet

By adding a thin layer of water between the projection lens and the wafer, the immersion system can create features 30 percent smaller.
Photolithography

- Defining the smallest components requires short wavelengths of light.
- Currently, most fabrication processors use extreme ultra-violet light at 193nm.
- Can pass the light through water. The water slows the light (less velocity) shrinking its wavelength. It is estimated that this technique will meet demands for 7 more years.
- On February 20, 2006 IBM Almaden & JSR Micro demonstrated a system using an “unidentified” light slowing liquid yielding patterns 29.9nm wide.

*Science News, March 2, 2006*
The HIGH-k SOLUTION


In 2007 new 45nm Microprocessors were the result of the first big material redesign in CMOS transistors since the late 1960s
## 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><strong>High Volume</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Manufacturing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Technology Node (nm)</strong></td>
<td>45</td>
<td>32</td>
<td>22</td>
<td>16</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td><strong>Integration Capacity (BT)</strong></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><strong>Delay Scaling</strong></td>
<td>&gt;0.7</td>
<td>~1?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Energy Scaling</strong></td>
<td>~0.5</td>
<td>&gt;0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Transistors</strong></td>
<td>Planar</td>
<td></td>
<td>3G, FinFET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Variability</strong></td>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ILD</strong></td>
<td>~3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RC Delay</strong></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><strong>Metal Layers</strong></td>
<td>8-9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5 to 1 Layer per generation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
10 Nanometer Technology

• Nov. 15, 2012, Samsung unveiled a 64 gigabyte (GB) multimedia card (eMMC) based on 10 nm technology.

• April 11, 2013, Samsung announced it was mass-producing High-Performance 128-gigabit NAND Flash Memory with 10 nm and 20 nm technology.

• April 2015, TSMC announced that 10 nm production would begin at the end of 2016.

• May 23rd 2015, Samsung Electronics showed off a wafer of 10nm FinFET chips.
Factors Contributing to Advancing Microprocessor Performance

- Shrinking Component Size
- Increasing Speed
- Reducing Circuit Resistance
- New Materials
Factors Contributing to Advancing Microprocessor Performance

- RISC vs. CISC
- VLIW
- Multi-level Cache
- Parallelism & Pipelining
Factors Contributing to Advancing Microprocessor Performance

- RISC vs. CISC
- VLIW
- Multi-level Cache
- Parallelism & Pipelining
- Multi-core Technology
Multicore Craze

• For years, the trend was to make chips faster
  Today → 3 Ghz

• But the power required (Watts) and the heat generated is proportional to the frequency squared.

• Therefore, put more computers on the chip but run at slower speeds.
• How long can Moore’s Law continue?
• What are the limits to this integrated circuit technology?

“There are two constraints:

– The finite velocity of light
– The atomic nature of materials”

- Stephen Hawking
IBM researchers have stored and retrieved digital data from an array of just 12 atoms

New York Times, 1/12/12
IBM’s Chip Stacking Technology
Single-phase, miniaturized convective cooling

Distributed return architecture with cross section showing inlet jets with neighboring drainage holes.
Single-phase, miniaturized convective cooling

SEM section of two-level jet plate. Water flow is indicated by blue arrows.
Interior Structure of 3-D chips

3-D Volatile Memory  [Matrix Semiconductor]

2-D Random-Access Memory  [IBM 256-Megabit]

3-D Logic Circuit  [Lab Prototype]

2-D Microprocessor  [Advanced Micro Devices Athlon]

- Monosilicon substrate
- Insulators
- Aluminum wires
- Polysilicon
- Tungsten plugs
- Ion-doped silicon
- Isolation oxides
- Silicide
Intel’s 3D Transistor

The chip maker breaks from conventional approaches to make transistors.

Conventional transistor: Electrons flow between components called a source and a drain, forming a two-dimensional conducting channel. A component called a gate starts and stops the flow, switching a transistor on or off.

Intel’s new transistor: A fin-like structure rises above the surface of the transistor with the gate wrapped around it, forming conducting channels on three sides. The design takes less space on a chip, and improves speed and reduces power consumption.
Intel’s 22nm 3D tri-gate transistor
“Every economic era is based on a key abundance and a key scarcity.”

George Gilder,
Forbes ASAP, 1992

What are the key scarcities?