Update on Physical Scalability
Sabotaging Performance Gains!

Dallas IEEE Computer Society Meeting
Friday Jan 21, 2011

Douglas J. Matzke, Ph.D.
IEEE Senior Member
matzke@IEEE.org
Abstract

In September 1997, Dr. Matzke wrote the lead off paper entitled "Will Physical Scalability Sabotage Performance Gains?" for the special issue of Computer Magazine on "Billion Transistor Computer". This paper is now required reading for most computer architecture courses in the world and cited by 257 other papers. The prediction in that paper was architectures would become more fine grain due to wire scaling and most likely the billion transistor computer would be a multiple CPU machine. This paper will give an update on this prediction and talk about other trends in the architecture and device arena, including multi core cpus, hybrid core machines, Memristors and quantum computing trends.
Introduction and Outline

Topics in Presentation

- Review of Wire Scaling Prediction
- Billion Transistor computers
- Current Multi-core processors – Core Wars
- Process Trends and Intel roadmap
- Limits of semiconductor/computer scaling
- Design Trends
- Memristor Fundamentals
- Scaling predictions
- Summary
Wire Scaling Prediction 1997

\[ t_{\text{gate}} = t_{\text{wire}} \sim R \times C \]

Signal Drive
Distance/region

\[ t_{\text{clock}} \sim 12 \text{ gates} \]

Assumption:
25 simple gate delays per clock or 12 drive distance
## Non-scaling Drive Distance

<table>
<thead>
<tr>
<th>Node</th>
<th>0.6 um</th>
<th>0.06 um</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die length</td>
<td>16 mm</td>
<td>32 mm</td>
<td>2 X</td>
</tr>
<tr>
<td>Gates on Die</td>
<td>1 million</td>
<td>400 million</td>
<td>400 X</td>
</tr>
<tr>
<td>Clock Frequency</td>
<td>166 MHz</td>
<td>2.5 GHz</td>
<td>15 X</td>
</tr>
<tr>
<td>( t_{\text{gate}} )</td>
<td>250 ps</td>
<td>15.6 ps</td>
<td>16 X</td>
</tr>
<tr>
<td>( d_{\text{wire locality (raw)}} )</td>
<td>5 mm</td>
<td>0.03125 mm</td>
<td>1/160 X</td>
</tr>
<tr>
<td>( d_{\text{wire locality (improv)}} )</td>
<td>5 mm</td>
<td>0.125 mm</td>
<td>1/40 X</td>
</tr>
<tr>
<td>( d_{\text{clock locality}} )</td>
<td>Die if &gt;.18 ( \mu \text{m} )</td>
<td>1.5 mm</td>
<td>1/40 X</td>
</tr>
<tr>
<td>Gates in Region</td>
<td>100,000</td>
<td>6,000</td>
<td>1/16 X</td>
</tr>
</tbody>
</table>

**Assumptions over 8 process steps:**
- 150% increase in gate speed per process step
- 20% wire improvement per process step
- 10% die size increase per process step
Die reachable per clock

Process nodes: .6, .35, .25, .18, .13, .1, .08, .06 µm
Trends Since 1997 paper

- Clock speeds have maxed out ~3 GHz
- Moore’s law w/high dielectric materials
- Process nodes are now at 32 nm (next 22)
- 10 chips since 2003 w/ > 1 B transistors
- Multiple CPU chips are the norm
- Large fine grain GPU and FPGA chips
- Power major design constraint (>200 W)
Transistor Counts 1971-2008

Curve shows ‘Moore’s Law’: transistor count doubling every two years

from wikipedia
**Billion Transistor Chips by 2010**

<table>
<thead>
<tr>
<th>Product</th>
<th>Date</th>
<th>Trans</th>
<th>Proc</th>
<th>Cores</th>
<th>Codename</th>
<th>Developer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itanium</td>
<td>Feb 2011</td>
<td>3.1 B</td>
<td>32 nm</td>
<td>8*4</td>
<td>Poulson</td>
<td>Intel</td>
</tr>
<tr>
<td>Nvidia GTX 570</td>
<td>Dec 2010</td>
<td>3 B</td>
<td>40 nm</td>
<td>480</td>
<td>GF110</td>
<td>NVIDIA/GPU</td>
</tr>
<tr>
<td>UltraSPARC T3</td>
<td>Sep 2010</td>
<td>1 B</td>
<td>45 nm</td>
<td>16*8</td>
<td>Niagara 3</td>
<td>Sun/Oracle</td>
</tr>
<tr>
<td>Core i7-980X</td>
<td>Mar 2010</td>
<td>1.17 B</td>
<td>32 nm</td>
<td>6*2</td>
<td>Gulftown</td>
<td>Intel</td>
</tr>
<tr>
<td>Intel Xeon</td>
<td>May 2009</td>
<td>2.3 B</td>
<td>45 nm</td>
<td>8*2</td>
<td>Beckton</td>
<td>Intel</td>
</tr>
<tr>
<td>Intel Itanium</td>
<td>Feb 2008</td>
<td>2 B</td>
<td>65 nm</td>
<td>4</td>
<td>Tukwila</td>
<td>Intel</td>
</tr>
<tr>
<td>Power7 (8-core)</td>
<td>Aug 2009</td>
<td>1.2 B</td>
<td>45 nm</td>
<td>8*4</td>
<td>Power 7</td>
<td>IBM</td>
</tr>
<tr>
<td>GeForce GTX 280</td>
<td>Dec 2008</td>
<td>1.4 B</td>
<td>65 nm</td>
<td>240</td>
<td>GPX 200</td>
<td>NVIDIA/GPU</td>
</tr>
<tr>
<td>Itanium-2</td>
<td>Oct 2005</td>
<td>1.72 B</td>
<td>90 nm</td>
<td>2*2</td>
<td>Montecito</td>
<td>Intel</td>
</tr>
<tr>
<td>Stratix IV FPGA</td>
<td>May 2008</td>
<td>2.5 B</td>
<td>40 nm</td>
<td>680K</td>
<td>FPGA Gates</td>
<td>Altera</td>
</tr>
<tr>
<td>Virtex FGPA</td>
<td>Sep 2003</td>
<td>1 B</td>
<td>70 nm</td>
<td>4 PPC</td>
<td>FPGA w/PPC</td>
<td>Xilinx</td>
</tr>
</tbody>
</table>
Core Wars

Sun Niagara 3: 16 Cores

IBM_Power7: 8 cores

Intel Xeon “Beckton”: 8 Cores
## ITRS: International Technology Roadmap for Semiconductors

### Near-term Years

<table>
<thead>
<tr>
<th>YEAR OF PRODUCTION</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology Node</td>
<td>hp90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRAM 1/2 Pitch (nm)</td>
<td>100</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td>65</td>
<td>57</td>
<td>50</td>
</tr>
<tr>
<td>MPU/ASIC M1 1/2 Pitch (nm)</td>
<td>120</td>
<td>107</td>
<td>95</td>
<td>85</td>
<td>75</td>
<td>67</td>
<td>60</td>
</tr>
<tr>
<td>MPU/ASIC Poly Si 1/2 Pitch (nm)</td>
<td>107</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td>65</td>
<td>57</td>
<td>50</td>
</tr>
<tr>
<td>MPU Printed Gate Length (nm)</td>
<td>65</td>
<td>53</td>
<td>45</td>
<td>40</td>
<td>35</td>
<td>32</td>
<td>28</td>
</tr>
<tr>
<td>MPU Physical Gate Length (nm)</td>
<td>45</td>
<td>37</td>
<td>32</td>
<td>28</td>
<td>25</td>
<td>22</td>
<td>20</td>
</tr>
</tbody>
</table>

### Long-term Years

<table>
<thead>
<tr>
<th>YEAR OF PRODUCTION</th>
<th>2010</th>
<th>2012</th>
<th>2013</th>
<th>2015</th>
<th>2016</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology Node</td>
<td>hp45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRAM 1/2 Pitch (nm)</td>
<td>45</td>
<td>35</td>
<td>32</td>
<td>25</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>MPU/ASIC M1 1/2 Pitch (nm)</td>
<td>54</td>
<td>42</td>
<td>38</td>
<td>30</td>
<td>27</td>
<td>21</td>
</tr>
<tr>
<td>MPU/ASIC Poly Si 1/2 Pitch (nm)</td>
<td>45</td>
<td>35</td>
<td>32</td>
<td>25</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>MPU Printed Gate Length (nm)</td>
<td>25</td>
<td>20</td>
<td>18</td>
<td>14</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>MPU Physical Gate Length (nm)</td>
<td>18</td>
<td>14</td>
<td>13</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

These sizes are close to physical limits and technological limits.


Jan 21, 2011 DJ M
Updated ITRS Forecast

Intel’s Process Roadmap

Source:

See HANS STORK IEEE NanoTech 2010 paper
Computer Scaling Limits

- **Physical Limits**
  - Power density/Dissipation: max is 100 W/cm²
  - Thermal/noise: \( E/f = 100 \)h
  - Molecular/atomic/charge discreteness limits
  - Quantum: tunneling & Heisenberg uncertainty

- **Technology Limits**
  - Gate Length: min > 8 nm (with new materials)
  - Lithography Limits: wavelength of visible light
  - Power dissipation (<100 watts) and Temperature
  - Wire Scaling: multicpu chips at ~ billion transistors
  - Materials for dielectrics etc
2010 Design Trends

- Multicpu Chips will continue
  - Manage power & no more clock increases
  - Requires innovation in parallel computing
  - Designs may top at < 100 cpus
  - CPUs and GPUS integrated (Sandy Bridge at Intel)

- Higher density/lower power solutions
  - DSP/CPUs heterogeneous systems for portable systems
  - CPU/FPGA systems (Convey Computers, IBM)
  - New memory/logic devices (spintronics)
  - Memristor based systems (HP, Numenta)

- Quantum Computing is nitch market
Memristor Fundamentals

Three original 2-terminal circuit elements (based on current, voltage, charge, and magnetic flux relationships)

In 1971, Leon Chua, an electrical engineer professor at UC Berkley, arranged the linear relationships between each of the four basic variables describing the above circuit relationships.

four final 2-terminal circuit elements
Scaling Predictions

- Semiconductors will stop scaling in <10 yrs
  - Nanocomputers won’t stop this; only delay it
  - Breakthroughs required or industry stagnates
  - College students consider non-semiconductor careers

- High dimensional Research in other areas:
  - Deep meaning and automatic learning
  - Programming probabilistic parallel computers
  - Noise as valued resource instead of unwanted
  - Higher dimensional computing
  - Investigate non-local computing
  - Biological inspired computing – Quantum Brain?
Summary

- Predictions in ‘97 came true as expected
- Scaling wall is now visible to industry
- Heat limits my stop multiprocessor count
- Materials innovation allows more of Moore
- New devices may help scaling (more than Moore)
- Fab Costs may slow before physical limits
- Must think outside 3d box (quantum?)
- Watch for unexpected aspects of quonoise
- Tablet/phone computing changes markets
- Clouding computing virtualization trends