A Quick HPC Perspective on Accelerators

Vincent Betro
Computational Scientist
University of Tennessee National Institute for Computational Sciences
Things started looking harder around 2004...

Historical SpecInt2000 Performance

- **Intel**
- **2x/18 months**
- **2x/24 months**

"Right Hand Turn"

Source: published SPECInt data
Moore’s Law is not at all dead...

Intel process technology capabilities

<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>Feature Size</td>
<td>90nm</td>
<td>65nm</td>
<td>45nm</td>
<td>32nm</td>
<td>22nm</td>
<td>16nm</td>
<td>11nm</td>
<td>8nm</td>
</tr>
<tr>
<td>Integration Capacity</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
</tr>
</tbody>
</table>

Transistor for 90nm Process
Source: Intel

Influenza Virus
Source: CDC
The problem in 2004...didn’t get better.
Not a new problem, just a new scale...

Cray-2 with cooling tower in foreground, circa 1985
How to get same number of transistors to give us more performance without cranking up power?

Many core is more power efficient

Power ~ area

Single thread performance ~ area**.5
Example: Dual core with voltage scaling

RULE OF THUMB

<table>
<thead>
<tr>
<th>Frequency Reduction</th>
<th>Power Reduction</th>
<th>Performance Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>15%</td>
<td>45%</td>
<td>10%</td>
</tr>
</tbody>
</table>

A 15% Reduction In Voltage Yields

SINGLE CORE

Area = 1
Voltage = 1
Freq = 1
Power = 1
Perf = 1

DUAL CORE

Area = 2
Voltage = 0.85
Freq = 0.85
Power = 1
Perf = ~1.8

Frequency Reduction
Power Reduction
Performance Reduction

15%
45%
10%

A 15% Reduction In Voltage Yields

SINGLE CORE

Area = 1
Voltage = 1
Freq = 1
Power = 1
Perf = 1

DUAL CORE

Area = 2
Voltage = 0.85
Freq = 0.85
Power = 1
Perf = ~1.8
GPU Architecture – Kepler: Streaming Multiprocessor (SMX)

- 192 SP CUDA Cores per SMX
  - 192 fp32 ops/clock
  - 192 int32 ops/clock
- 64 DP CUDA Cores per SMX
  - 64 fp64 ops/clock
- 4 warp schedulers
  - Up to 2048 threads concurrently
- 32 special-function units
- 64KB shared mem + L1 cache
- 48KB Read-Only Data cache
- 64K 32-bit registers
Since the days of RISC vs. CISC, Intel has mastered the art of figuring out what is important about a new processing technology, and saying “why can’t we do this in x86?”

The Intel Many Integrated Core (MIC) architecture is about large die, simpler circuit, much more parallelism, in the x86 line.
MIC Architecture

Many cores on the die

L1 and L2 cache

Bidirectional ring network for L2 Memory and PCIe connection
What makes an Accelerator?

CPU Memory

CPU

GPU Memory

GPU

Intel Xeon Phi
## Top 10 Systems in November 2012
(Just got updated last month)

<table>
<thead>
<tr>
<th>Site</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>Country</th>
<th>Cores</th>
<th>Rmax [Pflops]</th>
<th>Power [MW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oak Ridge National Laboratory</td>
<td>Cray</td>
<td><strong>Titan</strong></td>
<td>USA</td>
<td>560,640</td>
<td>17.6</td>
<td>8.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cray XK7, Opteron 16C 2.2GHz,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gemini, NVIDIA K20x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lawrence Livermore National Laboratory</td>
<td>IBM</td>
<td><strong>Sequoia</strong></td>
<td>USA</td>
<td>1,572,864</td>
<td>16.3</td>
<td>7.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BlueGene/Q, Power BQC 16C 1.6GHz, Custom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIKEN Advanced Institute for Computational Science</td>
<td>Fujitsu</td>
<td><strong>K Computer</strong></td>
<td>Japan</td>
<td>795,024</td>
<td>12.6</td>
<td></td>
</tr>
<tr>
<td>Argonne National Laboratory</td>
<td>IBM</td>
<td><strong>Mira</strong></td>
<td>USA</td>
<td>786,432</td>
<td>3.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BlueGene/Q, Power BQC 16C 1.6GHz, Custom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forschungszentrum Juelich (FZJ)</td>
<td>IBM</td>
<td><strong>JuQUEEN</strong></td>
<td>Germany</td>
<td>393,216</td>
<td>4.14</td>
<td>1.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BlueGene/Q, Power BQC 16C 1.6GHz, Custom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leibniz Rechenzentrum</td>
<td>IBM</td>
<td><strong>SuperMUC</strong></td>
<td>Germany</td>
<td>147,456</td>
<td>3.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>iDataPlex DX360M4, Xeon E5 8C 2.7GHz, Infiniband FDR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas Advanced Computing Center/UT</td>
<td>Dell</td>
<td><strong>Stampede</strong></td>
<td>USA</td>
<td>204,900</td>
<td>2.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PowerEdge C8220, Xeon E5 8C 2.7GHz, Intel Xeon Phi</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>National SuperComputer Center in Tianjin</td>
<td>NUDT</td>
<td><strong>Tianhe-1A</strong></td>
<td>China</td>
<td>186,368</td>
<td>4.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Xeon 6C, NVidia, FT-1000 8C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INECA</td>
<td>IBM</td>
<td><strong>Fermi</strong></td>
<td>Italy</td>
<td>163,840</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BlueGene/Q, Power BQC 16C 1.6GHz, Custom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>IBM</td>
<td><strong>DARPA Trial Subset</strong></td>
<td>USA</td>
<td>63,360</td>
<td>1.52</td>
<td>3.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Power 775, Power7 8C 3.84GHz, Custom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Titan System at Oak Ridge National Laboratory
(Just displaced by Chinese system using Xeon Phi)

Upgrade of Jaguar from Cray XT5 to XK6
Cray Linux Environment operating system
Gemini interconnect
- 3-D Torus
- Globally addressable memory
- Advanced synchronization features
AMD Opteron 6274 processors (Interlagos)
New accelerated node design using NVIDIA multi-core accelerators
- 2011: 960 NVIDIA x2090 “Fermi” GPUs
- 2012: 14,592 NVIDIA “Kepler” GPUs
20+ PFlops peak system performance
600 TB DDR3 mem. + 88 TB GDDR5 mem

<table>
<thead>
<tr>
<th>Titan Specs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Nodes</td>
<td>18,688</td>
</tr>
<tr>
<td>Login &amp; I/O Nodes</td>
<td>512</td>
</tr>
<tr>
<td>Memory per node</td>
<td>32 GB + 6 GB</td>
</tr>
<tr>
<td># of Fermi chips (2012)</td>
<td>960</td>
</tr>
<tr>
<td># of NVIDIA “Kepler” (2013)</td>
<td>14,592</td>
</tr>
<tr>
<td>Total System Memory</td>
<td>688 TB</td>
</tr>
<tr>
<td>Total System Peak Performance</td>
<td>20+ Petaflops</td>
</tr>
<tr>
<td>Liquid cooling at the cabinet level</td>
<td>Cray EcoPHLe</td>
</tr>
</tbody>
</table>
Power Efficiency over Time

Linpack/Power [Gflops/kW]

- Accelerator and BG
- multicore
- TOP10
- TOP50
- TOP500

Data from: TOP500 November 2012
Courtesy Horst Simon, LBNL
Trends with ends.
Heterogeneous computing is an important piece of the computing future. GPU accelerators are here to stay.

Now, let’s see how you can use them...