GPU COMPUTING AND THE FUTURE OF HPC

Timothy Lanfear, NVIDIA
Power-constrained Computers
EXASCALE COMPUTING WILL ENABLE TRANSFORMATIONAL SCIENCE RESULTS

Comprehensive Earth System Model at 1km scale, enabling modeling of cloud convection and ocean eddies.

First-principles simulation of combustion for new high-efficiency, low-emission engines.

Coupled simulation of entire cells at molecular, genetic, chemical and biological levels.

Predictive calculations for thermonuclear and core-collapse supernovae, allowing confirmation of theoretical models.
EXAFLOP EXPECTATIONS

Titan
8.2 MW

Growing size, cost and power

CM5
~200 KW

First Exaflop Computer

1 EF

1 PF

1 TF

1 GF

Power for CPU-only Exaflop Supercomputer = Power for the Bay Area, CA (San Francisco + San Jose)

HPC’s Biggest Challenge: Power
MOORE’S LAW IS ONLY PART OF THE STORY

- 1993: 3M transistors
- 1997: 7.5M transistors
- 2001: 42M transistors
- 2004: 275M transistors
- 2007: 580M transistors
- 2010: 3B transistors
- 2013: 7B transistors
1968: invented DRAM

1974: postulated all key figures of merit of MOSFETs improve provided geometric dimensions, voltages, and doping concentrations are consistently scaled to maintain the same electric field.
CLASSIC DENNARD SCALING

2.8x chip capability in same power

Chip Power vs. Chip Capability

- 2x more transistors
- 1.4x faster transistors
- 0.7x voltage
- 0.7x capacitance
POST DENNARD SCALING

2x chip capability at 1.4x power
1.4x chip capability at same power

Transistors are no faster
Static leakage limits reduction in $V_{th}$ => $V_{dd}$ stays constant

Chip Power

2x more transistors
0.7x capacitance

0.7x voltage
THE HIGH COST OF DATA MOVEMENT

Fetching operands costs more than computing on them

- Relative cost grows with each generation
- Wire delay (ps/mm) not improving
SO, WHAT TO DO?

1) Stop making it worse...

Multicore CPUs

But still only a tiny fraction of CPU power spent on flops

2) Unwind all that complexity we threw at single thread performance
HPC IS GOING HYBRID

- Do most work by cores optimized for extreme energy efficiency
- Still need a few cores optimized for fast serial work

x86 CPU
Fast single threads
(serial work)

GPU
Extreme power-efficiency
(throughput work)

Sandy Bridge
32nm
690 pJ/flop

Kepler
28nm
132 pJ/flop

Intel MIC

PCIe

Xeon

(AMD Fusion too)
EXPLOSIVE GROWTH OF GPU COMPUTING

- **2008**
  - 1 Supercomputer
  - 60 Universities
  - 4,000 Academic Papers

- **2012**
  - 1.5M CUDA Downloads
  - 52 Supercomputers
  - 560 Universities
  - 22,500 Academic Papers
### Research: Higher Education and Supercomputing

**COMPUTATIONAL CHEMISTRY AND BIOLOGY**

#### Molecular Dynamics

<table>
<thead>
<tr>
<th>Application</th>
<th>Brief Description</th>
<th>Time per iteration</th>
<th>GPU(s)</th>
<th>CPU(s)</th>
<th>Available for</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHARMM</td>
<td>Molecular dynamics of flexible molecules for simulations of proteins, DNA and RNA</td>
<td>1-2h</td>
<td>4094</td>
<td>4094</td>
<td>Yes</td>
<td>Available for Linux only.</td>
</tr>
<tr>
<td>GROMACS</td>
<td>Molecular dynamics for flexible macromolecules for simulations of proteins, DNA and RNA</td>
<td>3-5h</td>
<td>4094</td>
<td>4094</td>
<td>Yes</td>
<td>Available for Linux only.</td>
</tr>
<tr>
<td>NAMD</td>
<td>Molecular dynamics of flexible macromolecules for simulations of proteins, DNA and RNA</td>
<td>1-2h</td>
<td>4094</td>
<td>4094</td>
<td>Yes</td>
<td>Available for Linux only.</td>
</tr>
<tr>
<td>ALAMO</td>
<td>Molecular dynamics of flexible macromolecules for simulations of proteins, DNA and RNA</td>
<td>1-2h</td>
<td>4094</td>
<td>4094</td>
<td>Yes</td>
<td>Available for Linux only.</td>
</tr>
</tbody>
</table>

#### Computational Chemistry and Biology

<table>
<thead>
<tr>
<th>Application</th>
<th>Brief Description</th>
<th>Time per iteration</th>
<th>GPU(s)</th>
<th>CPU(s)</th>
<th>Available for</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECONS</td>
<td>Reference sequence analysis</td>
<td>5-10s</td>
<td>4094</td>
<td>4094</td>
<td>Yes</td>
<td>Available for Linux only.</td>
</tr>
<tr>
<td>UGENE</td>
<td>Universal gene finding and analysis</td>
<td>5-10s</td>
<td>4094</td>
<td>4094</td>
<td>Yes</td>
<td>Available for Linux only.</td>
</tr>
<tr>
<td>SOPMA</td>
<td>Structural genomics</td>
<td>5-10s</td>
<td>4094</td>
<td>4094</td>
<td>Yes</td>
<td>Available for Linux only.</td>
</tr>
<tr>
<td>RECON</td>
<td>Reference genome analysis</td>
<td>5-10s</td>
<td>4094</td>
<td>4094</td>
<td>Yes</td>
<td>Available for Linux only.</td>
</tr>
</tbody>
</table>

**References**

- [www.nvidia.com/appscatalog](www.nvidia.com/appscatalog)

---

**HUNDREDS OF GPU-ACCELERATED APPLICATIONS**

- [www.nvidia.com/appscatalog](www.nvidia.com/appscatalog)
### The Green500 List

Listed below are the June 2014 The Green500’s energy-efficient supercomputers ranked from 1 to 100.

<table>
<thead>
<tr>
<th>Green500 Rank</th>
<th>MFLOPS/W</th>
<th>Site*</th>
<th>Computer*</th>
<th>Total Power (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4,389.82</td>
<td>GSIC Center, Tokyo Institute of Technology</td>
<td>Tsubame-KFC - LX1 U-4GPU/104Ra-1G Cluster, Intel Xeon E5-2690v2 6C 2.10GHz, Infiniband FDR, NVIDIA Keio</td>
<td>34.58</td>
</tr>
<tr>
<td>2</td>
<td>3,631.70</td>
<td>Cambridge University</td>
<td>Wilkes - Dell T5620 Cluster, Intel Xeon E5-2630v2 6G 2.60GHz, Infiniband FDR, NVIDIA Keio</td>
<td>52.62</td>
</tr>
<tr>
<td>3</td>
<td>3,517.84</td>
<td>Center for Computational Sciences, University of Tsukuba</td>
<td>HA-PACS TCA - Cray 3822G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.80GHz, Infiniband QDR, NVIDIA Keio</td>
<td>78.77</td>
</tr>
<tr>
<td>4</td>
<td>3,459.46</td>
<td>SURFsara</td>
<td>Cartesius Acoelerator Island - Bullx BS5 Cluster, Intel Xeon E5-2450v2 8C 2.5GHz, Infiniband 4x FDR, Nvidia K40m</td>
<td>44.40</td>
</tr>
<tr>
<td>5</td>
<td>3,185.91</td>
<td>Swiss National Supercomputing Centre (CSCS)</td>
<td>Piz Daint - Cray XC30, Xeon E5-2670 8C 2.60GHz, Ares interconnect, NVIDIA K20x</td>
<td>1,233.66</td>
</tr>
<tr>
<td>6</td>
<td>3,131.06</td>
<td>ROMEO-HP Center - Champagne-Ardenne</td>
<td>Bar 8x R821-E3 Cluster, Intel Xeon E5-2660v2 8C 2.70GHz, Infiniband FDR, NVIDIA Keio</td>
<td>81.41</td>
</tr>
<tr>
<td>7</td>
<td>3,019.72</td>
<td>CSIRO</td>
<td>CSIRO GPU Cluster - NIVO G16 2GPU, Xeon E5-2660 8C 2.20GHz, Infiniband FDR, NVIDIA Keio</td>
<td>80.20</td>
</tr>
<tr>
<td>8</td>
<td>2,951.95</td>
<td>GSIC Center, Tokyo Institute of Technology</td>
<td>Tsubame KFC - 2.5 Cluster Platform SL3WxG7, Xeon X5670 6C 2.83GHz, Infiniband QDR, NVIDIA K20x</td>
<td>927.86</td>
</tr>
<tr>
<td>9</td>
<td>2,813.14</td>
<td>Exploration &amp; Production - Eni S.p.A.</td>
<td>HPC2 - DataPlex DX560M4, Intel Xeon E5-2660v2 10C 2.8GHz, Infiniband FDR, NVIDIA K20x</td>
<td>1,087.49</td>
</tr>
<tr>
<td>10</td>
<td>2,570.41</td>
<td>Financial Institution</td>
<td>DataPlex DX560M4, Intel Xeon E5-2660v2 10C 2.80GHz, Infiniband, NVIDIA K20x</td>
<td>54.60</td>
</tr>
<tr>
<td>11</td>
<td>2,529.42</td>
<td>Financial Institution</td>
<td>DataPlex DX560M4, Intel Xeon E5-2660v2 10C 2.80GHz, Infiniband FDR, NVIDIA K20x</td>
<td>60.25</td>
</tr>
<tr>
<td>12</td>
<td>2,529.42</td>
<td>Financial Institution</td>
<td>DataPlex DX560M4, Intel Xeon E5-2660v2 10C 2.80GHz, Infiniband FDR, NVIDIA K20x</td>
<td>60.25</td>
</tr>
<tr>
<td>13</td>
<td>2,529.42</td>
<td>Financial Institution</td>
<td>DataPlex DX560M4, Intel Xeon E5-2660v2 10C 2.80GHz, Infiniband FDR, NVIDIA K20x</td>
<td>60.25</td>
</tr>
<tr>
<td>14</td>
<td>2,529.42</td>
<td>Financial Institution</td>
<td>DataPlex DX560M4, Intel Xeon E5-2660v2 10C 2.80GHz, Infiniband FDR, NVIDIA K20x</td>
<td>60.25</td>
</tr>
<tr>
<td>15</td>
<td>2,529.10</td>
<td>Max-Planck-Gesellschaft MBI</td>
<td>DataPlex DX560M4, Intel Xeon E5-2660v2 10C 2.80GHz, Infiniband, NVIDIA K20x</td>
<td>269.94</td>
</tr>
</tbody>
</table>

**June 2014 Green 500 list, Tesla powers 15 of the most energy-efficient supercomputers**

**First sweep since IBM BlueGene**

**Tsubame-KFC: 4.3 GFLOPS / Watt**
OVERARCHING GOALS FOR TESLA

Power Efficiency

Ease of Programming And Portability

Application Space Coverage
KEPLER
THE WORLD’S FASTEST, MOST EFFICIENT HPC ACCELERATOR

- SMX (power efficiency)
- Hyper-Q (programmability and application coverage)
- Dynamic Parallelism
TESLA K80
WORLD’S FASTEST ACCELERATOR FOR DATA ANALYTICS AND SCIENTIFIC COMPUTING

2x Faster
2.9 TF | 4992 Cores | 480 GB/s

Deep Learning: Caffe

Double the Memory
Designed for Big Data Apps

24GB

Maximum Performance
Dynamically Maximize Perf for Every Application

Caffe Benchmark: AlexNet training throughput based on 20 iterations, CPU: E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2
PERFORMANCE LEAD CONTINUES TO GROW

**Peak Double Precision FLOPS**

- **NVIDIA GPU**
  - M1060
  - M2090
  - K20
  - K40
  - K80

- **x86 CPU**
  - Westmere
  - Sandy Bridge
  - Ivy Bridge
  - Haswell

**Peak Memory Bandwidth**

- **NVIDIA GPU**
  - M1060
  - M2090
  - K20
  - K40
  - K80

- **x86 CPU**
  - Westmere
  - Sandy Bridge
  - Ivy Bridge
  - Haswell
TESLA K80: 10X FASTER ON REAL-WORLD APPS

CPU: 12 cores, E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2
GPU: Single Tesla K80, Boost enabled
WHAT DOES THE FUTURE HOLD?
FAST PACED CUDA GPU ROADMAP

- **Tesla**
  - CUDA
  - FP64

- **Fermi**
  - Dynamic Parallelism

- **Kepler**
  - Unified Memory
  - 3D Memory
  - NVLink

- **Maxwell**
  - DX12

- **Pascal**
  - Unified Memory
  - 3D Memory
  - NVLink

<table>
<thead>
<tr>
<th>Year</th>
<th>Technology</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>Tesla</td>
<td>CUDA, FP64</td>
</tr>
<tr>
<td>2010</td>
<td>Fermi</td>
<td>Dynamic Parallelism</td>
</tr>
<tr>
<td>2012</td>
<td>Kepler</td>
<td>Unified Memory, 3D Memory, NVLink</td>
</tr>
<tr>
<td>2014</td>
<td>Maxwell</td>
<td>DX12</td>
</tr>
<tr>
<td>2016</td>
<td>Pascal</td>
<td>Unified Memory, 3D Memory, NVLink</td>
</tr>
</tbody>
</table>
PASCAL GPU FEATURES
NVLINK AND STACKED MEMORY

**NVLINK**
- GPU high speed interconnect
- 80-200 GB/s

**3D Stacked Memory**
- 4x Higher Bandwidth (~1 TB/s)
- 3x Larger Capacity
- 4x More Energy Efficient per bit
NVLINK
HIGH-SPEED GPU INTERCONNECT

KEPLER GPU

PASCAL GPU

X86, ARM64, POWER CPU

2014

X86, ARM64, POWER CPU

2016

NVLink

POWER CPU

PCIe

PCIe

NVLink

NVLink

POWER CPU

2014

2016
3D MEMORY

3D Chip-on-Wafer integration
Many X bandwidth
2.5X capacity
4X energy efficiency
PASCAL

- NVLink: 5 to 12X PCIe 3.0
- 3D Memory: 2 to 4X memory BW & size
- Module: 1/3 size of PCIe card
PARALLELISM IN MAINSTREAM LANGUAGES

- Enable more programmers to write parallel software
- Give programmers the choice of language to use
- GPU support in key languages
C++ PARALLEL ALGORITHMS LIBRARY

- Complete set of parallel primitives: 
  for_each, sort, reduce, scan, etc.

- ISO C++ committee voted unanimously to accept as official tech. specification working draft

std::vector<int> vec = ... 

// previous standard sequential loop
std::for_each(vec.begin(), vec.end(), f);

// explicitly sequential loop
std::for_each(std::seq, vec.begin(), vec.end(), f);

// permitting parallel execution
std::for_each(std::par, vec.begin(), vec.end(), f);

Prototype: https://github.com/n3554/n3554
Incorporating OpenACC into GCC is an excellent example of open source and open standards working together to make accelerated computing broadly accessible to all Linux developers.

Oscar Hernandez
Oak Ridge National Laboratories
NUMBA PYTHON COMPILER

- Free and open source compiler for array-oriented Python
- NEW numba.cuda module integrates CUDA directly into Python

```python
@cuda.jit("void(float32[:], float32, float32[:], float32[:])")
def saxpy(out, a, x, y):
    i = cuda.grid(1)
    out[i] = a * x[i] + y[i]

# Launch saxpy kernel
saxpy[griddim, blockdim](out, a, x, y)
```

http://numba.pydata.org/
Approach: apply a closure to a set of arrays

```java
// vector addition
float[] X = {1.0, 2.0, 3.0, 4.0, ... };
float[] Y = {9.0, 8.1, 7.2, 6.3, ... };
float[] Z = {0.0, 0.0, 0.0, 0.0, ... };
jog.foreach(X, Y, Z, new jogContext(),
    new jogClosureRet<jogContext>(){
        public float execute(float x, float y) {
            return x + y;
        }
    });
```

foreach iterations parallelized over GPU threads

![Graph: Java Black-Scholes Options Pricing Speedup]

Learn More: **S4938**: Vinod Grover: “Accelerating JAVA on GPUs”
Wednesday, 17:30 - 17:55
Room LL20C
THE FUTURE OF HPC IS GREEN

- Power is the constraint
  - Vast majority of work must be done by cores designed for efficiency
- GPU computing has a sustainable model
  - Aligned with technology trends, supported by consumer markets
- Future evolution will focus on:
  - Integration (CPU, network, memory)
  - Increased generality - efficient on any code with high parallelism
- This is simply how computers will be built