Announcements

• Put your name on HW#1 before turning in!
Top500 List – November 2018
<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Country</th>
<th>Arch /Proc</th>
<th>Cores</th>
<th>Max/Peak PFLOPS</th>
<th>Accel</th>
<th>Power kW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summit (IBM)</td>
<td>US/ORNL</td>
<td>Power9</td>
<td>2,397,824</td>
<td>143/200</td>
<td>NVD Volta</td>
<td>9,783</td>
</tr>
<tr>
<td>2</td>
<td>Sierra (IBM)</td>
<td>US/LLNL</td>
<td>Power9</td>
<td>1,572,480</td>
<td>94/125</td>
<td>NVD Volta</td>
<td>7,438</td>
</tr>
<tr>
<td>3</td>
<td>TaihuLight</td>
<td>China</td>
<td>Sunway</td>
<td>10,649,600</td>
<td>93/125</td>
<td>?</td>
<td>15,371</td>
</tr>
<tr>
<td>4</td>
<td>Tianhe-2A</td>
<td>China</td>
<td>x86/IVB</td>
<td>4,981,760</td>
<td>61/101</td>
<td>MatrixDSP</td>
<td>18,482</td>
</tr>
<tr>
<td>5</td>
<td>Piz Daint (Cray)</td>
<td>Switzerland</td>
<td>x86/SNB</td>
<td>387,872</td>
<td>21/27</td>
<td>NVD Tesla</td>
<td>2,384</td>
</tr>
<tr>
<td>6</td>
<td>Trinity (Cray)</td>
<td>US/LANL</td>
<td>x86/HSW</td>
<td>979,072</td>
<td>20/158</td>
<td>XeonPhi</td>
<td>7,578</td>
</tr>
<tr>
<td>7</td>
<td>ABCI (Fujitsu)</td>
<td>Japan</td>
<td>x86/SKL</td>
<td>391,680</td>
<td>20/33</td>
<td>NVD Tesla</td>
<td>1,649</td>
</tr>
<tr>
<td>8</td>
<td>SuperMUC-NG (Lenovo)</td>
<td>Germany</td>
<td>x86/SKL</td>
<td>305,856</td>
<td>19/26</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>9</td>
<td>Titan (Cray)</td>
<td>USA/ORNL</td>
<td>x86/Opteron</td>
<td>560,640</td>
<td>17/27</td>
<td>NVD K20</td>
<td>8,209</td>
</tr>
<tr>
<td>10</td>
<td>Sequoia (IBM)</td>
<td>USA/LLNL</td>
<td>Power BGQ</td>
<td>1,572,864</td>
<td>17/20</td>
<td>?</td>
<td>7,890</td>
</tr>
<tr>
<td>11</td>
<td>Lassen (IBM)</td>
<td>USA/LLNL</td>
<td>Power9</td>
<td>248,976</td>
<td>15/19</td>
<td>NVD Tesla</td>
<td>?</td>
</tr>
<tr>
<td>12</td>
<td>Cori (Cray)</td>
<td>USA/LBNL</td>
<td>x86/HSW</td>
<td>622,336</td>
<td>14/27</td>
<td>Xeon Phi</td>
<td>3,939</td>
</tr>
<tr>
<td>13</td>
<td>Nurion (Cray)</td>
<td>Korea</td>
<td>x86/??</td>
<td>570,020</td>
<td>14/25</td>
<td>Xeon Phi</td>
<td>?</td>
</tr>
<tr>
<td>14</td>
<td>Oakforest (Fujitsu)</td>
<td>Japan</td>
<td>x86/??</td>
<td>556,104</td>
<td>13/24</td>
<td>Xeon Phi</td>
<td>2,719</td>
</tr>
<tr>
<td>15</td>
<td>HPC4 (HPE)</td>
<td>Italy</td>
<td>x86/SNB</td>
<td>253,600</td>
<td>12/18</td>
<td>NVD Tesla</td>
<td>1,320</td>
</tr>
<tr>
<td>16</td>
<td>Tera-1000-2 (Bull)</td>
<td>France</td>
<td>x86/??</td>
<td>561,408</td>
<td>12/23</td>
<td>Xeon Phi</td>
<td>3,178</td>
</tr>
<tr>
<td>17</td>
<td>Stampede2 (Dell)</td>
<td>US</td>
<td>x86/SNB</td>
<td>367,024</td>
<td>12/28</td>
<td>Xeon Phi</td>
<td>18,309</td>
</tr>
<tr>
<td>18</td>
<td>K computer (Fujitsu)</td>
<td>Japan</td>
<td>SPARC VIIIfx</td>
<td>705,024</td>
<td>10/11</td>
<td>?</td>
<td>12,660</td>
</tr>
<tr>
<td>19</td>
<td>Marconi (Lenovo)</td>
<td>Italy</td>
<td>x86/SNB</td>
<td>348,000</td>
<td>10/18</td>
<td>Xeon Phi</td>
<td>18,816</td>
</tr>
<tr>
<td>20</td>
<td>Taiwania-2 (Quanta)</td>
<td>Taiwan</td>
<td>x86/SKL</td>
<td>170,352</td>
<td>9/15</td>
<td>NVD Tesla</td>
<td>798</td>
</tr>
<tr>
<td>21</td>
<td>Mira (IBM)</td>
<td>US/ANL</td>
<td>Power/BGQ</td>
<td>786,432</td>
<td>8/10</td>
<td>NVD Tesla</td>
<td>3,954</td>
</tr>
<tr>
<td>22</td>
<td>Tsubame3.0 (HPE)</td>
<td>Japan</td>
<td>x86/SNB</td>
<td>135,828</td>
<td>8/12</td>
<td>NVD Tesla</td>
<td>792</td>
</tr>
<tr>
<td>23</td>
<td>UK Meteor (Cray)</td>
<td>UK</td>
<td>x86/IVB</td>
<td>241,920</td>
<td>7/8</td>
<td>NVD Tesla</td>
<td>8,128</td>
</tr>
<tr>
<td>24</td>
<td>Theta (Cray)</td>
<td>US/ANL</td>
<td>x86/??</td>
<td>280,320</td>
<td>7/11</td>
<td>Xeon Phi</td>
<td>11,661</td>
</tr>
<tr>
<td>25</td>
<td>MareNostrum (Lenovo)</td>
<td>Spain</td>
<td>x86/SKL</td>
<td>153,216</td>
<td>6/10</td>
<td>Xeon Phi</td>
<td>1,632</td>
</tr>
</tbody>
</table>
Top500 List Notes

• Can watch video presentation on it here?
• Left off my summary: RAM? (#1 is 3PB) Interconnect?
• Power: does this include cooling or not? Cost of power over lifetime of use is often higher than the cost to build it.
• Power comparison: small town? 1MW around 1000 homes? (this varies)
• How long does it take to run LINPACK? How much money does it cost to run LINPACK?
• Lots of turnover since last time I taught the class?
• Operating system. Cost to run computer more than cost to build it?
• Tiahne-2 was Xeon Phi, but US banned Intel from exporting anymore, so upgraded and using own custom DSP boards now.
• Need to be 10 PFlops to be near top these days? 100k cores at least?
• First ARM system, Cavium ThunderX in Astra (US/LANL) at 204
What goes into a top supercomputer?

- Commodity or custom
  embedded vs high-speed?
- Memory
- Storage
  How much?
  Large hadron collider one petabyte of data every day
  Shared? If each node wants same data, do you need to
  replicate it, have a network filesystem, copy it around
with jobs, etc? Cluster filesystems?

- Reliability. How long can it stay up without crashing? Can you checkpoint/restart jobs?
  - Sequoia MTBF 1 day.
  - Blue Waters 2 nodes failure per day.
  - Titan MTBF less than 1 day

- Power / Cooling
  - Big river nearby?

- Accelerator cards / Heterogeneous Systems

- Network
  - How fast? Latency? Interconnect? (torus, cube,
hypercube, etc)
Ethernet? Infiniband? Custom?

- Operating System
  Linux? Custom? If just doing FP, do you need overhead of an OS? Job submission software, Authentication

- Software – how to program?
  Too hard to program can doom you. A lot of interest in the Cell processor. Great performance if programmed well, but hard to do.

- Tools – software that can help you find performance problems
Other stuff

• Rmax vs Rpeak – Rmax is max measured, Rpeak is theoretical best
• HPL Linpack
  ○ Embarrassingly parallel linear algebra
  ○ Solves a (random) dense linear system in double precision (64 bits) arithmetic
• HP Conjugate gradient benchmark
  ○ More realistic? Does more memory access, more I/O bound.
• #1 on list is Summit. 3PFLOPS CG whereas 143PFLOPS HPL
• Some things can move around, K-computer 18th in HPL but 3rd with CG
• Green 500
Historical Note

- From the November 2002 list, entry #332
- Location: Orono, ME
- Proc Arch: x86
- Proc Type: Pentium III, 1GHz
- Total cores: 416
- RMax/RPeak: 225/416 GFLOPS
- Power: ???
- Accelerators: None
Introduction to Performance Analysis
What is Performance?

- Getting results as quickly as possible?
- Getting *correct* results as quickly as possible?
- What about Budget?
- What about Development Time?
- What about Hardware Usage?
- What about Power Consumption?
Motivation for HPC Optimization

HPC environments are expensive:

- Procurement costs: $\sim$40 million
- Operational costs: $\sim$5 million/year
- Electricity costs: 1 MW / year $\sim$1 million
- Air Conditioning costs: ??
Know Your Limitation

- CPU Constrained
- Memory Constrained (Memory Wall)
- I/O Constrained
- Thermal Constrained
- Energy Constrained
Performance Optimization Cycle

Develop Code

Functionally Complete/Correct Code

Measure

Analyze

Modify / Tune

Functionally Complete/Correct/Optimized Code

Usage / Production
“We should forget about small efficiencies, say about 97% of the time:

**premature optimization is the root of all evil.**

Yet we should not pass up our opportunities in that critical 3%. A good programmer will not be lulled into complacency by such reasoning, he will be wise to look carefully at the critical code; but only after that code has been identified” — Donald Knuth
Amdahl’s Law

- Original
- Speed up Blue 100x
- Speed up Red 2x

Time
Speedup

- Speedup is the improvement in latency (time to run)

\[ S' = \frac{t_{old}}{t_{new}} \]

So if originally took 10s, new took 5s, then speedup=2.
Scalability

• How a workload behaves as more processors are added

• Parallel efficiency: \( E_p = \frac{S_p}{p} = \frac{T_s}{pT_p} \)
  
  \( p \) = number of processes (threads)
  
  \( T_s \) is execution time of serial code
  
  \( T_p \) is execution time with \( p \) processes

• Linear scaling, ideal: \( S_p = p \)

• Super-linear scaling – possible but unusual
Strong vs Weak Scaling

• Strong Scaling – for fixed program size, how does adding more processors help.

• Weak Scaling – how does adding processors help with the same per-processor workload.
Strong Scaling

- Have a problem of a certain size, want it to get done faster.

- Ideally with problem size $N$, with 2 cores it runs twice as fast as with 1 core (linear speedup)

- Often processor bound; adding more processing helps, as communication doesn’t dominate

- Hard to achieve for large number of nodes, as many
algorithms communication costs get larger the more nodes involved

- Amdahl’s Law limits things, as more cores don’t help serial code

- Strong scaling efficiency: $\frac{t_1}{N \cdot t_N} \times 100\%$

- Improve by throwing CPUs at the problem.
Weak Scaling

• Have a problem, want to increase problem size without slowing down.

• Ideally with problem size $N$ with 1 core, a problem of size $2^n$ just as fast with 2 cores.

• Often memory or communication bound.

• Gustafson’s Law (rough paraphrase)
  No matter how much you parallelize your code, there will be serial sections that just can’t be made parallel.
• Weak scaling efficiency: \( \left( \frac{t_1}{t_N} \right) \times 100\% \)

• Improve by adding memory, or improving communication?
Scaling Example

LINPACK on Rasp-pi cluster. What kind of scaling is here?

![Graph showing scaling example](image)
Weak scaling. To get linear speedup need to increase problem size. If it were strong scaling, the individual colored lines would increase rather than dropping off.