Introduction
Fact: hardware acceleration is common

- 15% of the systems have accelerators or co-processors
- trend is increasing

source: top500.org
Fact: hardware acceleration delivers

- 34.4% of the “total” performance is from accelerators or co-processors

Source: Top500.org
Clear trend towards heterogeneity - for the next quantum leap in performance

There is a clear expectation and proof that substantial performance improvement is possible from heterogeneous architectures.

Kinds of heterogeneity

• General purpose application accelerators
• General purpose coprocessors

Costs involved

• Increased power consumption
• Effort required to modify applications
• In some cases implication on licensing costs
Practical limits – ways to connect CPUs to accelerators

# CPUs

• Vast majority of compute nodes have up to two CPUs

# Accelerators

• Due to space the practical limit is Four per U

Options (about seven due to the limitations above)

- CPU0 -> GPU0
- CPU0 -> GPU1, GPU2
- CPU0 -> GPU1, GPU2, GPU3
- CPU0 -> GPU0, GPU2, GPU3, GPU4
- CPU0 -> GPU0 | CPU1 -> GPU2
- CPU0 -> GPU0, GPU2 | CPU1 -> GPU3, GPU4
- CPU0 -> GPU0, GPU2, GPU3, GPU4 | CPU1 -> NC
Ratios effecting application performance
**CPU:GPU**

Among the seven options:

<table>
<thead>
<tr>
<th>Option</th>
<th>CPU:GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU0 -&gt; GPU0</td>
<td>1:2</td>
</tr>
<tr>
<td>CPU0 -&gt; GPU1,GPU2</td>
<td>1:2</td>
</tr>
<tr>
<td>CPU0 -&gt; GPU1,GPU2,GPU3</td>
<td>1:4</td>
</tr>
<tr>
<td>CPU0 -&gt; GPU0,GPU2,GPU3,GPU4</td>
<td>1:4</td>
</tr>
<tr>
<td>CPU0 -&gt; GPU0</td>
<td>2:2 or 1:1</td>
</tr>
<tr>
<td>CPU0 -&gt; GPU0,GPU2</td>
<td>2:4 or 1:2</td>
</tr>
<tr>
<td>CPU0 -&gt; GPU0,GPU2, GPU3, GPU4</td>
<td>1:4</td>
</tr>
</tbody>
</table>
Goals of the study

Give available system configurations

• performance vs. increased power
  – how much more performance is observed and for how much more power?

• maximizing performance/watt
  – how to make the power consumption more efficient?
A brief introduction HW used for the performance study
K80 is the latest Tesla GPU from NVIDIA

- **K80** has high raw compute power
  - **2.8X - 6.1X** CPU (theoretical peak)

- compare K80 vs. K40
  - **cores**: 4992, 73%
  - **memory**: 24GB, 100%
  - memory BW: 480GB/s, 67%
  - clock: 875 MHz, 17%
  - **power**: 300W max., 28%
  - **SP**: 5.6 TFLOPs, 30%
  - **DP**: 1.87/2.7 TFLOPs, 30%-60%
  - Architecture: GK210 vs. GK110
  - PCIe: Gen 3 (15.7 GB)

Intel Xeon E5-2690 v3 @ 2.6 GHz (12 core) ~ 441.6 GFLOPS
Intel Xeon E5-2699 v3 @ 2.3 GHz (18 core) ~ 662.2 GFLOPS
C4130 Server – 1U for high GPU density server

- Five system configurations to select from – 4 GPUs in 1U

- 1 CPU
- 2 CPUs
Four accelerator configurations

- **A** economical for connecting 4 accelerators, unbalances
- **B** is **A** with an additional CPU, both have SW
- **C** has no switch module, also **C** is balanced
Two accelerator configurations

- E economical for connecting 2 accelerators
- D is the balanced option for 2 accelerators
Application performance characterization with K80 GPUs
• For 4 GPUs “C” is best, “B” is a close second.
• For 2 GPUs “D” is best

N=121344, Auto boost=ON
• For 4 GPUs – “C” is best, “B” is a close second
• For 2 GPUs – “D” is best
• Code performances best with balanced configurations
C4130 Performance – K80 – LAMMPS

- For 4 GPUs – “C” is best
- For 2 GPU – “D” is best
- Code performs best with balanced configurations
C4130 Performance – K80 – GROMACS

- For 4 GPUs – “B” is best
- For 2 GPU – “D” is best
- Code performs best with 1:4 :: CPU:GPU ratio; higher is better.
Design features of the C4130
Features to maximize acceleration for a wide class of applications

- GPUs in the front of the server
- GPUs power given priority over CPU power
- Users can choose the configuration most suitable to their needs!
Thank you