The L-CSC cluster: An AMD-GPU-based cost- and power-efficient multi-GPU system for Lattice-QCD calculations at GSI

Dr. David Rohr
Frankfurt Institute for Advanced Studies
SC14, New Orleans
Green500 BoF Session, 20.11.2014
The Lattice-CSC Cluster at GSI

Lattice-CSC (at GSI):
• Built for Lattice-QCD simulations.
• Quantum Chromo Dynamics (QCD) is the physical theory describing the strong force.
• Very memory intensive.

GSI:
Helmholtz-Center for Heavy Ion Research
Darmstadt, Germany

Currently building a new particle accelerator for the FAIR project.

- Large New Datacenter (Green Cube)
  - 700+ Racks, 15 MW Power
  - PUE: approx. 1.05

Green DataCenter at GSI, Darmstat, Germany
Lattice-CSC (at GSI):
- Built for Lattice-QCD simulations.
- Quantum Chromo Dynamics (QCD) is the physical theory describing the strong force.
- Very memory intensive.

160 Compute nodes:
- 4 * AMD FirePro S9150 GPU
- ASUS ESC4000 G2S Server
- 2 * Intel 10-core Ivy-Bridge CPU
- 256 GB DDR3-1600 1.35V
- FDR Infiniband
- 1.7 PFLOPS Peak

Green DateCenter at GSI, Darmstat, Germany
The Lattice-CSC Cluster at GSI

Lattice-CSC (at GSI):
• Built for Lattice-QCD simulations.
• Quantum Chromo Dynamics (QCD) is the physical theory describing the strong force.
• Very memory intensive.

160 Compute nodes:
• 4 * AMD FirePro S9150 GPU
• ASUS ESC4000 G2S Server
• 2 * Intel 10-core Ivy-Bridge CPU
• 256 GB DDR3-1600 1.35V
• FDR Infiniband
• 1.7 PFLOPS Peak

Installation ongoing, 56 nodes ready

Green DateCenter at GSI, Darmstat, Germany
Custom Open-Source DGEMM & HPL

CALDGEMM Library and HPL-GPU, available as Open-Source under (L)GPL license.
• Optimized for multi-GPU with OpenCL (exchangeable GPU backend – vendor independent).
• Dynamic workload balancing among CPUs / GPUs.
• Optimized for power efficiency.
Custom Open-Source DGEMM & HPL

CALDGEMM Library and HPL-GPU, available as Open-Source under (L)GPL license.

- Optimized for multi-GPU with OpenCL (exchangeable GPU backend – vendor independent).
- Dynamic workload balancing among CPUs / GPUs.
- Optimized for power efficiency.

→ Perfect scaling up to four GPUs.
Custom Open-Source DGEMM & HPL

CALDGEMM Library and HPL-GPU, available as Open-Source under (L)GPL license.

- Optimized for multi-GPU with OpenCL (exchangeable GPU backend – vendor independent).
- Dynamic workload balancing among CPUs / GPUs.
- Optimized for power efficiency.

→ Perfect scaling up to four GPUs.

Our approach for HPL:

<table>
<thead>
<tr>
<th>HPL Steps</th>
<th>Offload</th>
<th>GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGEMM</td>
<td>Dynamic Offload</td>
<td></td>
</tr>
<tr>
<td>DTRSM</td>
<td>Dynamic Pipeline</td>
<td></td>
</tr>
<tr>
<td>Factorization</td>
<td>Pipelined /</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hidden behind GPU</td>
<td></td>
</tr>
<tr>
<td>LASWP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broadcast</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dynamic Work Balancing & Optimal Configuration

Pipeline works well, as long as CPU tasks (solid line) finish before GPU tasks (dashed line).

→ Optimal GPU usage 95% of time

Combined CPU / GPU DGEMM:
  • Better Performance (2-5%)

GPU Only DGEMM:
  • Better Efficiency (3-4%)

![Graph showing dynamic work balancing and optimal configuration.](image-url)
Dynamic Work Balancing & Optimal Configuration

Pipeline works well, as long as CPU tasks (solid line) finish before GPU tasks (dashed line).
→ Optimal 95% of time

Combined CPU / GPU DGEMM:
• Better Performance (2-5%)

GPU Only DGEMM:
• Better Efficiency (3-4%)

→ We have two software versions:
• A performance optimized version
• An efficiency optimized version
Dynamic Parameter Tuning for Best Performance

At different point in time during Linpack run, different parameters are optimal.

- We choose optimal settings dynamically at every point in time.

![Graph showing performance in iteration](image.png)

- Performance in one iteration [GFlop/s]
- Iteration (Remaining Matrix Size)
- Without DTRSM Offload (5458 GFlop/s)
- With DTRSM Offload (5386 GFlop/s)
Dynamic Parameter Tuning for Best Performance

At different point in time during Linpack run, different parameters are optimal.

- We choose optimal settings dynamically at every point in time.
- Take care: Settings yielding optimal performance and settings yielding optimal efficiency may be different!
Dynamic Parameter Tuning for Optimal Efficiency

Using high-resolution power measurement, we plot the efficiency over time.
(Number of Operations per timebin / Energy consumption per timebin)

- Optimal CPU Frequency changes over time.

![Graph showing the efficiency of CPU operations over time. The graph compares two CPU frequencies, 1.8 GHz and 2.7 GHz, with corresponding power efficiency in GFLOPS/W. The graph shows a decrease in efficiency as the progress increases.]
Dynamic Parameter Tuning for Optimal Efficiency

Using high-resolution power measurement, we plot the efficiency over time. (Number of Operations per timebin / Energy consumption per timebin)

- Optimal CPU Frequency changes over time.
- We use dynamic frequency scaling to achieve optimal efficiency at every point in time.
Optimization Summary

Hardware tuning:
- Infiniband Network Root Filesystem – No Hard Disks / Ethernet / USB / etc.
- Optimal Fan Settings – Temperature v.s. Fan Power Consumption

Software optimizations:
- Custom Open-Source DGEMM / HPL Software based on OpenCL.
- Dynamic workload distribution among CPUs / GPUs.
- Dynamic parameter adaption for best performance or best efficiency at every point in time.
- Two settings of parameters – optimized for performance or for efficiency.
- Dynamic voltage / frequency scaling for CPU and GPU.
- For best efficiency, we leave some devices unloaded by intent: CPU at beginning, GPU at end.
Results

Power consumption over time

56 Nodes:
• 301300 GFLOPS
• 1016 W per Node
→ 5295 MFLOPS/W
Results

Power consumption over time

56 Nodes:
- 301300 GFLOPS
- 1016 W per Node
→ 5295 MFLOPS/W

Infiniband Switches:
- 257 W
→ Including the network: 5.27 GFLOPS/W
Results

Perfect scaling to many nodes:

Efficiency:
1 Node: 5378 MFLOPS/W
4 Nodes: 5250 MFLOPS/W
56 Nodes: 5270 MFLOPS/W

Performance (per node):
1 Node: 5791 GFLOPS
4 Nodes: 5380 GFLOPS
56 Nodes: 5380 GFLOPS

Performance optimized version:
6800 GFLOPS (single node)
Playing with the rules

The Green500 rules state that the power measurement interval must at least cover 20% of the middle 80% of the core phase.

GFLOPS/W:
Full measurement: 5296
Playing with the rules

The Green500 rules state that the power measurement interval must at least cover 20% of the middle 80% of the core phase: For instance the period 70%-90%.

GFLOPS/W:

- Full measurement: 5296
- 70%-90%: 6010
Playing with the rules

The Green500 rules state that the power measurement interval must at least cover 20% of the middle 80% of the core phase: For instance the period 70%-90%.

GFLOPS/W:

- Full measurement: 5296
- 70%-90%: 6010
- Short Run: 4907
- Short Run, 70%-90%: 6900
Suggestions

- All power measurements should cover 100% of the core phase.
- Do we want to measure 100% of the cluster?
  - Yes! Otherwise one could screen the nodes and measure the best one.
  - No! Measuring 100 kW and above at high accuracy can be very challenging.
Questions

Q & A