Energy Efficient High Performance Computing

Michael Gerndt
Technische Universität München
gerndt@in.tum.de
Faculty of Informatics

• Founded in 1975
  • 3500 students
  • 30% foreign students
• 40 Professors (22 Chairs)
HPC Applications
## TOP 500 – November 2014

<table>
<thead>
<tr>
<th>Name</th>
<th>Site</th>
<th>Architecture</th>
<th>Peak PFlop/s</th>
<th>Power MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tianhe/2</td>
<td>NUDT</td>
<td>NUDT, Cluster, Intel Xeon + Xeon Phi</td>
<td>54.9</td>
<td>17.8</td>
</tr>
<tr>
<td>Titan</td>
<td>ONL, USA</td>
<td>Cray, 560640 cores, Opteron+Nvidia K20</td>
<td>27.1</td>
<td>8.2</td>
</tr>
<tr>
<td>Sequoia</td>
<td>LLNL, USA</td>
<td>IBM BlueGene/Q, 1.6 M cores</td>
<td>20.1</td>
<td>7.9</td>
</tr>
<tr>
<td>K Computer</td>
<td>RIKEN, Japan</td>
<td>Fujitsu, 705 K cores SPARC 64</td>
<td>11.3</td>
<td>12.6</td>
</tr>
<tr>
<td>Mira</td>
<td>ANL, USA</td>
<td>IBM BlueGene/Q, 786 K cores</td>
<td>10.1</td>
<td>3.9</td>
</tr>
<tr>
<td>Piz Daint</td>
<td>CSCS</td>
<td>Cray Xeon + K20</td>
<td>7.7</td>
<td>2.3</td>
</tr>
<tr>
<td>Stampede</td>
<td>Univ. Texas</td>
<td>Dell Power + Xeon Phi</td>
<td>5.1</td>
<td>4.5</td>
</tr>
</tbody>
</table>
## Energy Efficiency (Linpack Peak/W)

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Architecture</th>
<th>Gflops/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tianhe/2</td>
<td>NUDT</td>
<td>Cluster, Intel Xeon +Xeon Phi</td>
<td>1.9</td>
</tr>
<tr>
<td>Titan</td>
<td>ONL, USA</td>
<td>Cray, 560640 cores, Opteron+Nvidia K20</td>
<td>2.1</td>
</tr>
<tr>
<td>Sequoia</td>
<td>LLNL, USA</td>
<td>IBM BlueGene/Q, 1.6 M cores</td>
<td>2.2</td>
</tr>
<tr>
<td>K Computer</td>
<td>RIKEN, Japan</td>
<td>Fujitsu, 705 K cores SPARC 64</td>
<td>0.8</td>
</tr>
</tbody>
</table>
Pace to Exascale

- Exascale: $10^{18}$ Flops/s
- Expected to be reached after 2018
- Massive parallelism and energy consumption

<table>
<thead>
<tr>
<th>Name</th>
<th>Cores</th>
<th>Power MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titan</td>
<td>20 M</td>
<td>303</td>
</tr>
<tr>
<td>Sequoia</td>
<td>80 M</td>
<td>395</td>
</tr>
<tr>
<td>K Computer</td>
<td>64 M</td>
<td>1145</td>
</tr>
<tr>
<td>SuperMUC</td>
<td>49 M</td>
<td>1000</td>
</tr>
</tbody>
</table>
European Exascale Hardware Projects

- **DEEP (2011-2014)**
  - Cluster of accelerators
  - Intel Xeon Phi
  - Extoll network
  - Dynamic distribution of cluster-nodes and accelerators

- **Mont Blanc**
  - Cortex A-15 SoCs
  - Blue Gene approach but
  - GPU on each SoC
  - Critical: network bandwidth
Energy Efficient HPC

- Reduce the power losses in the power supply chain
- Exploit your possibilities for using compressor-less cooling and use energy-efficient cooling technologies (e.g., direct liquid cooling)
- Re-use waste heat of IT systems
- Use newest semiconductor technology
- Use of energy saving processor and memory technologies
- Consider using special hardware or accelerators tailored for solving specific scientific problems or numerical algorithms
- Monitor the energy consumption of the compute systems and the cooling infrastructure
- Use energy aware system software to exploit the energy saving features of your target platform
- Monitor and optimize the performance of your scientific applications
- Use most efficient algorithms
- Use best libraries
- Use most efficient programming paradigm
- Use all performance tuning techniques

Energy efficient infrastructure
Energy efficient hardware
Energy aware software environment
Energy efficient applications
Leibniz Supercomputer Center
SuperMUC, 3 Petaflop Supercomputer
Sandy Bridge Processor

**Latency:**
- 4 cycles
- 12 cycles
- 31 cycles

**Bandwidth:**
- 2*16/cycle
- 32 / cycle
- 32 / cycle

**Network frequency equal to core frequency**

**L3 cache**
- Partitioned with cache coherence based on core valid bits
- Physical addresses distributed by a hash function
• 2 processors with 32 GB of memory
• Aggregate memory bandwidth per node 102.4 GB/s
• Latency
  • local ~50ns (~135 cycles @2.7 GHz)
  • remote ~90ns (~240 cycles)
SuperMUC at Leibniz Computing Centre

- **SHOME**: 1.5 PB / 10 GB/s
- **Archive and Backup**: ~30 PB
- **Visualization Internet**

**Network Infrastructure**

- **80 Gbit/s**
- **Spine Infiniband switches**
- **10GbE access**
- **Pruned tree (4:1)**

**Node Configurations**

- **18 Thin node islands** (each >8000 cores)
- **1 Fat node island** (8200 cores) also used as Migration System
- **5x-EP**: 16 cores/node, 2 GB/core
- **WM-EX**: 40 cores/node, 6.4 GB/core

**Storage**

- **10 PB**
- **200 GB/s**
- **Parallel Storage**
- **I/O nodes**
- **Login Support nodes**

**System Components**

- **Snapshots/Replica**: 1.5 PB
- **Disaster Recovery Site**

---

Prof. Dr. Michael Gerndt, gerndt@in.tum.de
What’s special about SuperMUC

SuperMUC is the third most powerful pure x86 ISA systems of the world

- General purpose, standard programming interface, „easy“ to port to,
- future-safe for many applications

SuperMUC is the most energy efficient x86 based supercomputer of the world

- Dark Center infrastructure at LRZ
- Warm Water cooling
- Energy aware scheduling with Load Leveler
SuperMUC Power Consumption

**SuperMUC HPL Power Consumption (Infrastructure, Machine Room & PDU Measurements)**

- **Power (Machine Room, kW)**
- **Power (PDU, kW)**
- **Power (infrastructure, kW)**
- **Energy (Machine Room, kWh)**

**Linpack HPL run May 17, 2012 – 2.582 PF**
- **Run Start:** 17.05.2012 20:56, 965.40 kW
- **Run End:** 18.05.2012 08:37, 711.02 kW
- **Duration:** 42045s or 11.68 hours
- **Avg. power:** 2758.87 kW
- **Energy:** 32308.68 kWh

Subsystems included in PDU measurements:
- Computational Nodes
- Interconnect Network

Provided by Torsten Wilde, LRZ

Prof. Dr. Michael Gerndt, gerndt@in.tum.de
CPU frequency and RAM setting

Provided by Andrey Semin, Sr. Engineer HPC, Intel
## SuperMUC - Costs

<table>
<thead>
<tr>
<th>Component</th>
<th>2010-2014</th>
<th>2014-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High End System</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investment Costs (Hardware and Software)</td>
<td>53 Mio €</td>
<td>~ 19 Mio €</td>
</tr>
<tr>
<td>Operating Costs (Electricity costs and maintenance for hardware und software, some additional personnel)</td>
<td>32 Mio €</td>
<td>~ 29 Mio €</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td>85 Mio €</td>
<td>~48 Mio €</td>
</tr>
<tr>
<td><strong>Extension Buildings (construction and Infrastructure)</strong></td>
<td>49 Mio €</td>
<td></td>
</tr>
</tbody>
</table>
Energy Capping in Contract with IBM

• New funding scheme: energy included
• Contract includes energy cost for 5 years
• Power consumption of the system varies between 1 and 3 MW depending on the usage by the applications
• The contract is based on the energy consumed in a benchmark suite agreed between IBM and LRZ
IBM iDataplex dx360 M4: Water cooling

- Heat flux > 90% to water; very low chilled water requirement
- Power advantage over air-cooled node:
  - Warm water cooled ~10%
    (cold water cooled ~15%)
  - due to lower $T_{\text{components}}$ and no fans
- Typical operating conditions: $T_{\text{air}} = 25 – 35^\circ\text{C}$, $T_{\text{water}} = 18 – 45^\circ\text{C}$
Warm Water Cooling

- Air, 23°C: 51°C
- Water, 30°C: 37°C
- Water, 40°C: 48°C
- Water, 50°C: 56°C
Warm Water Cooling

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water, 23°C</td>
<td>290W</td>
</tr>
<tr>
<td>Water, 23°C</td>
<td>248W</td>
</tr>
<tr>
<td>Water, 23°C</td>
<td>254W</td>
</tr>
<tr>
<td>Water, 23°C</td>
<td>260W</td>
</tr>
</tbody>
</table>

Prof. Dr. Michael Gerndt, gerndt@in.tum.de
Warm Water Cooling: CoolMUC
FP7 AutoTune Project Goals

• Tackle complexity of HPC architectures
  – Multicore, multisocket, accelerators, DVFS

• Enable higher productivity via auto-tuning

• Focus on static tuning in pre-production phase
  – Produce tuning recommendations

• Leverage state of the art performance analysis

• Implement an extensible environment
  – Support open and proprietary plugins
Periscope Tuning Framework

- Extension of Periscope
- Online tuning process
  - Application phase-based
- Extensible via tuning plugins
  - Single tuning aspect
  - Combining multiple tuning aspects
- Rich framework for plugin implementation
- Automatic and parallel experiment execution
Energy Efficiency Tuning in AutoTune

• Application tuning for energy efficiency
  – Performance optimization in general
  – Data transfer optimization
  – Concurrency throttling
  – Power gating
  – Dynamic voltage and frequency scaling

• Energy measurement
  – RAPL (Running Average Power Limit)
  – Node power supply
ENOPT library implemented by LRZ
Energy Properties

• **Automatic search for application properties**
  - Energy Inefficient Regions (Flops/J)
  - Energy Tunable Regions
  - Memory/compute bound regions
  - Regions with Load imbalance

• **Tuning plugin**
  - Inspect the energy properties
  - Identifies potential good combinations of governor and frequency
  - Performs an online evaluation of the scenarios
Results

SIP

Matrix Multiply

SeiSol

Energy (kJ)

Time (10s)
Summary

- Energy wall
- Energy reduction on all levels required
  - Technology
  - System
  - Application
- Auto-tuning can help in reducing energy consumption
  - Indirect through performance tuning
  - Direct by tuning DVFS
- Overprovisioning and power capping for exascale
  - Lawrence Livermore National Lab (Sequoia)