
SIMUTOOLS 2015
Athens, Greece

Mario Bielert, Florina M. Ciorba, Kim Feldhoff, Thomas Ilsche, Wolfgang E. Nagel
Outline

- Motivation

- Design and Architecture
  - Tracing
  - Overview
  - Control-Flow
  - Parallelism
  - Time Synchronization

- Use Cases
  - Communication Performance Modeling
  - Computation Power Modeling

- Conclusions and Future Work
Motivation

HAEC Project

- Challenge: Energy-Efficiency
  - Information and communication technology portion of total energy consumption: 10% in 2010
  - Current trend: 10x in next 20 years > 50% in 2030

- Challenge: Communication bottleneck in parallel computing
  - Parallelism = 2: #{communication links} = 1
  - Parallelism = 64: #{communication links} = 2016
  - Parallelism = n: #{communication links} = \( \frac{n}{2}(n - 1) \)
Motivation

Highly Adaptive Energy-efficient Computing

Challenges of HAEC:
- Build a computer...
  - which is energy-efficient (energy should be proportional to computational load)
  - with a very good performance (mitigates the computational bottleneck of parallel computing)
  - can be used / embedded almost everywhere (e.g., in a car)
Motivation
Target Platform - HAEC Box

- 4 boards with $4 \times 4$ massive parallel computing nodes
- Optical intra-board connections with 250 Gbit/s, 10 ns
- Wireless inter-board connections with 100 Gbit/s, 100 ns
- Bit error rate: $10^{-12}$, $10^{-8}$ respectively
Motivation

HAEC-SIM

- HAEC Box not available yet
- Need for design space exploration
  ⇒ Simulation required
- Design of HAEC Box requires a simulation environment that combines
  - hardware models,
  - software models, and
  - architecture models.
Motivation

Approach

- Trace-based simulation (TBS) using performance and energy models
Design and Architecture

Tracing

- Record the behavior of the application during its execution
  - called functions
  - communications and I/O
  - metrics
Design and Architecture
Tracing

- Record the behavior of the application during its execution
  - called functions
  - communications and I/O
  - metrics

- Diagram of events in different locations

![Diagram of events in different locations](image-url)
Overview of the HAEC-SIM framework.
Software control flow in HAEC-SIM.
Design and Architecture
Parallelism in HAEC-SIM

- Each location in input trace is simulated with its own process
- Uses MPI for communication and synchronization
- Information exchange can be done directly between processes and via resource managers
- State of shared resources are handled within resource managers (implemented as additional MPI processes)
Communicators for the parallel simulation ranks of HAEC-SIM.
Design and Architecture

Time Synchronization

- No synchronization needed, unless using shared resources
- Function enter or leave events are denoted as $e_{\text{func}}$ and $l_{\text{func}}$
- Send and receive message events are denoted as $s_{\text{msg}}$ and $r_{\text{msg}}$

- event
- message
- simulation time of simulation process 0
- simulation time of simulation process 1
- time synchronization between simulation processes
Use Cases
Communication Performance Modeling

- Running NAS Parallel Benchmark LU.C.64.1
- xyz mapping, 100 Gbit/s bandwidth, 100 ns latency, PNC model
- Explore the impact of the topology on the communication
  - $4 \times 4 \times 4$,
  - $16 \times 4 \times 1$,
  - $64 \times 1 \times 1$
Use Cases
Communication Performance Modeling

- Running NAS Parallel Benchmark LU.C.64.1
- xyz mapping, 100 Gbit/s bandwidth, 100 ns latency, PNC model
- Explore the impact of the topology on the communication
  - $4 \times 4 \times 4$
  - $16 \times 4 \times 1$
  - $64 \times 1 \times 1$

- Simulation results for three topologies

<table>
<thead>
<tr>
<th>Topology</th>
<th>Total time</th>
<th>Accumulated time for appl.</th>
<th>Accumulated time for MPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$4 \times 4 \times 4$</td>
<td>23.096 s</td>
<td>961.729 s</td>
<td>484.939 s</td>
</tr>
<tr>
<td>$16 \times 4 \times 1$</td>
<td>23.548 s</td>
<td>961.729 s</td>
<td>513.862 s</td>
</tr>
<tr>
<td>$64 \times 1 \times 1$</td>
<td>23.509 s</td>
<td>961.729 s</td>
<td>511.344 s</td>
</tr>
</tbody>
</table>
Use Cases
Computation Power Modeling

- Power estimation based on energy model utilizing PAPI counters
- Linear model for OFFCORERESPONSES and UOPS_DISPATCHED
- Measurement test system “artemis” with two Xeon E5-2690 processors
- Assume computing nodes behavior equal to artemis’ processor
- Mapping 16 processes to one node ⇒ shared resource
- Utilizes nodes resource manager
Use Cases
Computation Power Modeling

- Results of the power estimation for Parallel NAS benchmark LU.C.1.32

<table>
<thead>
<tr>
<th>Master thread</th>
<th>OMP thread 2</th>
<th>OMP thread 4</th>
<th>OMP thread 6</th>
<th>OMP thread 16</th>
<th>OMP thread 18</th>
<th>OMP thread 20</th>
<th>OMP thread 22</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.491054 s</td>
<td>+20 ms</td>
<td>+40 ms</td>
<td>+60 ms</td>
<td>+80 ms</td>
<td>20.491054 s</td>
<td>+20 ms</td>
<td>+40 ms</td>
</tr>
</tbody>
</table>

Values of Counter "power estimate node (0,0,0)" over Time

- W
  - ~2000 estimates
  - no estimates

Values of Counter "power measurement socket 0" over Time

- W
  - ~75 measurements
  - ~150 measurements
Conclusions

- Need for a highly adaptive energy-efficient computing platform
- Simulation framework supporting the integration of abstract models
- Demonstrate two simulation scenarios
  - Communication models allow to evaluate the impact design decisions, e.g., the topology
  - Energy models enhance the simulation to put a focus on energy optimization
- Source code available: http://tu-dresden.de/zih/haec_sim
Future Work

- Develop mapping strategies that consider the communication patterns of the application
- Verify the implemented model for unicast communication in the presence of errors
- Modeling of multicast communication
- Extend the energy models for communication operations
- Support for migration of tasks across computing nodes
Thank you for your attention

Questions?