Assessing Energy Efficiency of Fault Tolerance Protocols for HPC Systems

Esteban Meneses, Osman Sarood and Sanjay Kalé

Parallel Programming Laboratory
University of Illinois at Urbana-Champaign

SBAC-PAD 2012
Exascale

Energy
- Power management (20MW budget)
- Administrative considerations (1MW → $1M/year)
- System codesign (architectural features)

Fault Tolerance
- Size of the machine (200,000 sockets → MTBF)
- Types of failures (memory, accelerator, network)
- Different strategies

Energy Efficiency of Fault Tolerance Protocols
Agenda

1. Fault Tolerance Protocols
2. Experimental Setup
3. Experimental Results
4. Analytical Model
5. Discussion
6. Conclusions and Future Work
Fault Tolerance Protocols

- **Checkpoint/Restart**
  - State is saved periodically
  - Coordinated global checkpoint
  - Checkpoint stored locally
  - Failure $\rightarrow$ global rollback

- **Message-Logging**
  - Messages are stored at sender
  - Non-determinism logged
  - Determinants in causal path
  - Failure $\rightarrow$ local rollback

- **Parallel Recovery**
  - Tasks are migratable
  - Failure $\rightarrow$ recovery in parallel

**Parallel Recovery**

- **Message-Logging**
- **Checkpoint/Restart**

**Caveat**

- Many variants of checkpoint/restart
- Several message-logging protocols
- Hybrid schemes
Optimum Checkpoint Period

Daly’s modified model:

\[ \tau = \sqrt{2\delta(M + R)} - \delta \]

Questions

- Optimum \( \tau \) for message-logging and parallel recovery?
- Optimum \( \tau \) to minimize energy?
- Execution time vs energy consumption?
Charm++ Runtime System

- Migratable Objects Model
- Asynchronous Method Invocation
- Adaptive MPI → each rank becomes an object
- Application-level checkpoint

- One process per *logical* node
- Failure injection: `kill -9 pid`
- Failure detection → automatic restart on replacement node
- Fault tolerance protocols at object-level

Parallel Recovery

Node A
Node B
Node B'
Node C
Node D

Time
Energy Cluster

- **General Features**
  - 40 single-socket nodes
  - Each node has a four-core Intel Xeon and 4GB of main memory
  - Gigabit ethernet switch

- **Power Measuring**
  - Liebert power distribution unit (PDU)
  - Power measurement per-node
  - 1-second interval frequency
Checkpoint/Restart

- Test program
  - 7-point stencil
  - Nearest neighbor in 3D
  - Barrier after each step
  - Virtualization ratio = 32
  - 200 steps (checkpoints at 50 and 150)

- Local disk checkpoint
Total Energy Consumed

Checkpoint/Restart

Message-Logging

Parallel Recovery

Meneses, Sarood and Kalé (UIUC)
Test programs

- NAS Parallel Benchmarks
- Block Tridiagonal (BT) and Scalar Pentadiagonal (SP)
- Virtualization ratio = 4
<table>
<thead>
<tr>
<th></th>
<th>Jacobi3D</th>
<th>NPB-BT</th>
<th>NPB-SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Charm++</td>
<td>MPI</td>
<td>MPI</td>
</tr>
<tr>
<td>Problem size</td>
<td>$1024^3$</td>
<td>class C</td>
<td>class C</td>
</tr>
<tr>
<td>Number of cores</td>
<td>128</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Virtualization ratio</td>
<td>32</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Recovery parallelism</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Message-logging overhead</td>
<td>1.0%</td>
<td>3.6%</td>
<td>4.1%</td>
</tr>
<tr>
<td>Max power (C)</td>
<td>106</td>
<td>102</td>
<td>95</td>
</tr>
<tr>
<td>Max power (M)</td>
<td>106</td>
<td>102</td>
<td>96</td>
</tr>
<tr>
<td>Max power (P)</td>
<td>106</td>
<td>102</td>
<td>96</td>
</tr>
</tbody>
</table>

Message-logging does NOT increase power draw
# Execution Time and Energy Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{V}$</td>
<td>Optimal virtualization ratio</td>
<td>$&gt; 8$</td>
</tr>
<tr>
<td>$W$</td>
<td>Time to solution with $\mathcal{V}$</td>
<td>25 h</td>
</tr>
<tr>
<td>$M$</td>
<td>Mean-time-to-interrupt of the system</td>
<td>-</td>
</tr>
<tr>
<td>$S$</td>
<td>Total number of sockets in the system</td>
<td>-</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Checkpoint time</td>
<td>180 s</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Optimum checkpoint period</td>
<td>-</td>
</tr>
<tr>
<td>$R$</td>
<td>Restart time</td>
<td>30 s</td>
</tr>
<tr>
<td>$T$</td>
<td>Total execution time</td>
<td>-</td>
</tr>
<tr>
<td>$E$</td>
<td>Total energy consumption</td>
<td>-</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Message-logging slowdown</td>
<td>1.02</td>
</tr>
<tr>
<td>$P$</td>
<td>Available parallelism during recovery</td>
<td>8</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Message-logging recovery speedup</td>
<td>1.2</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Parallel recovery speedup</td>
<td>$P$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Parallel recovery slowdown</td>
<td>$\frac{P+1}{P}$</td>
</tr>
<tr>
<td>$H$</td>
<td>Max power of each socket</td>
<td>100 W</td>
</tr>
<tr>
<td>$L$</td>
<td>Base power of each socket</td>
<td>40 W</td>
</tr>
</tbody>
</table>
Execution Time and Energy Formulas

\[ T = T_{\text{Solve}} + T_{\text{Checkpoint}} + T_{\text{Recover}} + T_{\text{Restart}} \]

\[ E = E_{\text{Solve}} + E_{\text{Checkpoint}} + E_{\text{Recover}} + E_{\text{Restart}} \]

Execution Time (Parallel Recovery)

\[ T = W\mu + \left( \frac{W\mu}{\tau} - 1 \right) \delta + \frac{T}{M} \left( \delta + \frac{\tau - \delta}{2\sigma} + \frac{\tau + \delta}{2} (\lambda - 1) \right) + \frac{T}{M} R \]

Energy (Parallel Recovery)

\[ E = W\mu SH + \left( \frac{W\mu}{\tau} - 1 \right) \delta SL + \]
\[ \frac{T}{M} \left( \delta SL + \frac{\tau - \delta}{2\sigma} (PH + (S - P)L) + \frac{\tau + \delta}{2} (\lambda - 1) SH \right) + \frac{T}{M} RSL \]

Time-optimum \( \tau \)

Energy-optimum \( \tau \)
Improvement in Execution Time

Up to 17% improvement

Parallel Recovery
Message-Logging

Speedup
Number of Sockets (thousands)

Meneses, Sarood and Kalé (UIUC)  
Energy & Fault Tolerance in HPC  
SBAC-PAD 2012
Improvement in Energy

**Time-optimum** $\tau$

Improvement

**Energy-optimum** $\tau$

Improvement

Up to 13% improvement

## Discussion

- **Trend in ratio of base to maximum power**

<table>
<thead>
<tr>
<th>Processor</th>
<th>Release Date</th>
<th>Max Power</th>
<th>Base Power</th>
<th>Base/Max Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Xeon (E5520)</td>
<td>Q1,09</td>
<td>125</td>
<td>60</td>
<td>0.48</td>
</tr>
<tr>
<td>Intel Nehalem (i7 860)</td>
<td>Q3,09</td>
<td>151</td>
<td>52</td>
<td>0.34</td>
</tr>
<tr>
<td>Intel Sandy Bridge (i7 2600)</td>
<td>Q1,11</td>
<td>101</td>
<td>21</td>
<td>0.21</td>
</tr>
</tbody>
</table>

- **Migratability and over-decomposition in scientific applications**
Conclusions

“Minimize execution time $\implies$ minimize energy” (not true)
- Increase checkpoint frequency
- Recovery is more energy-efficient with message logging

Energy overhead of message-logging
- It does not increase power draw
- It increases energy consumption on the forward path

Parallel recovery leverages message-logging
- It provides the minimum execution time (users happy)
- It offers the minimum energy consumed (administrators happy)
- The model predicts 17% reduction in execution time, 13% reduction in energy consumed
Future Work

- Particle-simulation applications:
  - Molecular Dynamics
  - Quantum Chemistry
  - Cosmology

- Enhancements to analytical model:
  - Different failure distributions: Weibull, log-normal
  - No upper bound for checkpoint period
  - Energy-aware fault tolerance protocols
Acknowledgements

- **HPC Colony II Project.** This work is partially supported by the US Department of Energy under grant DOE DE-SC0001845 and by a machine allocation on XSEDE under award ASC050039N.

- **Prof. Tarek F. Abdelzaher.** The experimental results of this work come from the *Energy Cluster* in the University of Illinois at Urbana-Champaign.
Obrigado!

Q&A
Progress Diagram

- No Fault Tolerance Support
- Fault Tolerance Support

Performance Overhead

Progress Diagram for Energy Efficient Fault Tolerance

- No Fault Tolerance Support
- Fault Tolerance Support

- Slowdown
- Checkpoint
- Recovery

Progress

Power

Time

Failure

100%

Effect of Higher Parallelism During Recovery

![Graph showing improvement in energy efficiency with increased parallelism during recovery, with different lines for P=20, P=16, P=12, P=8, and P=8 (dotted line). The x-axis represents the number of sockets in thousands, and the y-axis represents the improvement factor.]
Optimum Checkpoint Period

- Optimum checkpoint period ($\tau$) vs MTBF

**Time-optimum $\tau$**

**Energy-optimum $\tau$**