Performance Improvement of Stencil Computations for Multi-core Architectures based on Machine Learning

Víctor Martínez, Fabrice Dupros, Márcio Castro, and Philippe O. A. Navaux

Federal University of Rio Grande do Sul – BR
Federal University of Santa Catarina – BR
Bureau de Recherches Géologiques et Minières – FR
Outline

• Motivation
  • Multicore Architectures.
  • Stencil Models.
• Machine Learning Methodology.
  • Input/Output vectors.
  • Hardware Counters Behavior.
  • ML Model.
• Experiments.
  • Testbed
  • Training and Validation sets.
• Results.
• Conclusions.
Motivation
Multicore Architectures

- Complexity:
  - Cache levels.
  - Cores.
  - Sockets.

- Optimization:
  - Non-uniform memory access.
  - Vectorization.
  - Compiler flags.
  - Memory policies.
Motivation
Stencil Models

• Heat
  • Arithmetic Intensity: 0.18

• Seismic
  • Velocity and Stress.
  • Arithmetic Intensity: 1.3

• Algorithms
  • Naive (Space domain)
  • Blocking (Space/Time domain)
Machine Learning Methodology
Hardware Counters Behavior

- **Input Vector**
  - Policy Scheduling.
  - Chunk Size.
  - Executed Threads.

- **Hardware Counters**
  - Last Level Cache Misses (L3)
  - Total Cycles.
  - Data Translation Lookaside Buffer Misses.
Machine Learning Methodology
Supported Vector Machines Model

Input Layer
- Input vector

HWC Layer
- SVM L3 CM
- SVM TLB DM
- SVM Cycles

Output Layer
- SVM Gflops
- Time

Performance
- Maximum
- Minimum

Training Set

Test Set

Searching

Searching
# Experiments

## Optimization, Training and Validation Sets

<table>
<thead>
<tr>
<th>Optimization</th>
<th>Parameters</th>
<th>Total of Configurations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Node 1</td>
</tr>
<tr>
<td>Number of threads</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Scheduling policy</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Chunk size</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>Block size X</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>Block size Y</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>Total for Naive</td>
<td>3</td>
<td>512</td>
</tr>
<tr>
<td>Total for blocking</td>
<td>5</td>
<td>2,097,152</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stencil</th>
<th>Set</th>
<th>Node 1</th>
<th>Node 2</th>
<th>Node 1</th>
<th>Node 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-point Jacobi</td>
<td>Training</td>
<td>44</td>
<td>38</td>
<td>2,355</td>
<td>4,054</td>
</tr>
<tr>
<td></td>
<td>Testing</td>
<td>11</td>
<td>10</td>
<td>589</td>
<td>1,014</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>55</td>
<td>48</td>
<td>44,794</td>
<td>49,152</td>
</tr>
<tr>
<td>Seismic wave</td>
<td>Training</td>
<td>211</td>
<td>237</td>
<td>2,176</td>
<td>371</td>
</tr>
<tr>
<td></td>
<td>Testing</td>
<td>53</td>
<td>60</td>
<td>544</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>264</td>
<td>297</td>
<td>6,849</td>
<td>1,020</td>
</tr>
</tbody>
</table>
## Results

### Regression Model Accuracy

- **Testing set**

![Input Output Diagram](Image)

<table>
<thead>
<tr>
<th>Stencil</th>
<th>Set</th>
<th>Naive</th>
<th>Blocking</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-point Jacobi</td>
<td>RMSE GFLOPS</td>
<td>0.7941</td>
<td>1.4185</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>0.6642</td>
<td>2.2537</td>
</tr>
<tr>
<td>7-point Jacobi</td>
<td>R-square GFLOPS</td>
<td>0.9782</td>
<td>0.9627</td>
</tr>
<tr>
<td></td>
<td>R-square Time</td>
<td>0.9879</td>
<td>0.8881</td>
</tr>
<tr>
<td>Seismic wave</td>
<td>RMSE GFLOPS</td>
<td>0.2273</td>
<td>0.3158</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>13.5391</td>
<td>15.6548</td>
</tr>
<tr>
<td>Seismic wave</td>
<td>R-square GFLOPS</td>
<td>0.9970</td>
<td>0.9822</td>
</tr>
<tr>
<td></td>
<td>R-square Time</td>
<td>0.9987</td>
<td>0.9971</td>
</tr>
</tbody>
</table>
Results

Best Performance Prediction

• Execution set
Conclusions

• A model to predict and to simulate performance of stencil computations on multi-core architectures was presented:
  • Performance of stencil can be predicted with a high accuracy.
  • Model can be integrated into an auto-tuning framework to find the best performance.

• Next step: a model based on unsupervised ML algorithms.
Performance Improvement of Stencil Computations for Multi-core Architectures based on Machine Learning

Thanks!

Acknowledgment

Research has received funding from the EU H2020 Programme and from MCTI/RNP-Brazil under the HPC4E Project, grant agreement n° 689772.

Victor Martinez, Fabrice Dupros, Márcio Castro, and Philippe O. A. Navaux

Federal University of Rio Grande do Sul – BR
Federal University of Santa Catarina – BR
Bureau de Recherches Géologiques et Minières – FR