Minimizing Thermal Variation in Heterogeneous HPC System with FPGA Nodes

Yingyi Luo, Xiaoyang Wang, Seda OGRENCI-MEMIK, Gokhan MEMIK, Kazutomo YOSHII, Pete Beckman

@ICCD 2018
Motivation

- FPGAs in data centers
  - Energy efficient
  - Reconfigurable
  - Emerging high level synthesis tools
    - Easier programmability
Motivation

• Less explored FPGA thermal dynamics
  ▪ Cooling power
  ▪ Reliability
    ➢ Mean Time to Failure: exponentially dependent on temperature
    ➢ Negative Bias Temperature Instability: affects PMOS threshold voltage
Goal

- Study thermal characteristics of multi-FPGA systems
- Minimize system peak temperature
Outline

• Thermal Model
• Task Placement
• Experimental Results
Thermal Model

- Given a set of HPC tasks to be assigned to a multi-FPGA system, predict system peak temperature
  - HPC Task: implemented in OpenCL
    - Easier programming
  - Model: utilizes machine learning techniques
    - No need for domain specific knowledge (e.g., FPGA model, system environment, etc.)
Features

• Temperature
  ▪ On chip temperature sensor
  ▪ Read @1Hz
  ▪ Represents the peak temperature (stable)
Features

• Task properties
  ▪ Intel FPGA SDK for OpenCL compilation information
    ➢ Logic Utilization (LU)
    ➢ RAM Blocks (RB)
    ➢ Frequency (F)
    ➢ DSP Blocks
    ➢ Memory Bits
    ➢ I/O Pins
Model Construction

Five steps
1. Map a series of benchmarks onto each FPGA in the system
2. Gather benchmark properties
3. Record peak temperature at runtime
4. Construct training data
5. Feed data into a machine learning algorithm
4. Construct training data

- Suppose N FPGAs in the system N tasks, a training sample is as follows:

<table>
<thead>
<tr>
<th>Task1</th>
<th>Task2</th>
<th>...</th>
<th>TaskN</th>
</tr>
</thead>
<tbody>
<tr>
<td>LU</td>
<td>RB</td>
<td>F</td>
<td>LU</td>
</tr>
</tbody>
</table>

LU: Left Upper; RB: Right Bottom; F: Feature; Peak Temp: Peak Temperature
5. Feed data into a machine learning algorithm
   - Candidate algorithms
     - Linear Regression (LR)
     - Multilayer Perceptron (MLP)
     - Random Forest (RF)
   - Utilize scikit-learn
Outline

• Thermal Model

• Task Placement

• Experimental Results
Experimental Setup

- Multi-FPGA system
  - Resides in the Chameleon Cloud
  - Four 2U nodes in the same rack
  - Nallatech 385A board
    - Arria10 1150GX FPGA chip
    - 8GB DDR3 memory
Experimental Setup

• Benchmarks
  ▪ 10 HPC tasks from Intel
  ▪ 10 minutes run for each

• Tool
  ▪ Intel Quartus Prime:
    Intel FPGA SDK for OpenCL
Temperature Variations

• Observations
  ▪ 4.55ºC between max and average
  ▪ 10.02ºC between max and min
  ▪ Performance is not affected

• Possible causes
  ▪ Process variation
  ▪ Air flow
  ▪ Node location
Thermal-Aware Task Placement

- An illustrative example

<table>
<thead>
<tr>
<th>Task</th>
<th>Initial Temperature</th>
<th>New Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>sobel_filter</td>
<td>49.69°C</td>
<td>69.31°C</td>
</tr>
<tr>
<td>asian_option</td>
<td>80.82°C</td>
<td>57.81°C</td>
</tr>
</tbody>
</table>
Thermal-Aware Task Placement

- Static
- When a task set arrives for scheduling
  1. Predict system peak temperatures for all possible task placement scenarios
  2. Suggest the placement with lowest predicted peak temperature
  3. Tasks reside on the nodes for a sufficiently long time
    - Typical of long-running computation scenarios in HPC systems
Thermal-Aware Task Placement

• In the experimented multi-FPGA system, for a task set
  ▪ \[4! = 24\] placements need to be evaluated
Outline

• Thermal Model

• Task Placement

• Experimental Results
Task-Dependent Model Performance

- Tasks used for training may appear in the testing set
- Accuracy
  - Split 80% for training, 20% for testing randomly
  - Repeat 10 times

<table>
<thead>
<tr>
<th>Methods</th>
<th>LR</th>
<th>MLP</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>3.11°C</td>
<td>2.40°C</td>
<td>1.43°C</td>
</tr>
</tbody>
</table>
Task-Dependent Model Performance

- Task placement
  - Optimal: minimal peak $T$ of all possible placements
  - Thermal oblivious (average behavior): average peak $T$ of all possible placements
Task-Independent Model Performance

- Tasks used for training are not in the testing set
- Accuracy
  - Select 6 tasks for training, 4 tasks for testing
  - Repeat over all possible combinations: $\binom{10}{4} = 210$

<table>
<thead>
<tr>
<th>Methods</th>
<th>LR</th>
<th>MLP</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>7.31°C</td>
<td>12.87°C</td>
<td>7.66°C</td>
</tr>
</tbody>
</table>
Task-Independent Model Performance

- For (mandelbrot, matrix_mult, sobel_filter, video_downscaling)
  - All prediction models can pick one of the lowest T placement

![Graph showing performance comparison between Actual, Linear Regression, Multilayer Perceptron, and Random Forest models for different placement indices. The graph indicates that for the best cases, the actual performance is close to the predictions made by the models, while for the worst cases, there is a significant deviation.]
Task-Independent Model Performance

- Task placement
  - Optimal
  - Average behavior

![Bar chart showing average peak temperature in degrees Celsius for different task placement models. The models include Optimal, Thermal Oblivious (Average Behavior), Linear Regression, Multilayer Perceptron, and Random Forest. The Optimal model has an average peak temperature of 65.30°C, the Thermal Oblivious model has 69.85°C, the Linear Regression model has 66.02°C, the Multilayer Perceptron model has 67.62°C, and the Random Forest model has 66.50°C.]
Task-Independent Model Performance

- Task placement
  - Optimal
  - Average behavior
  - Worst case behavior: max peak T of all possible placements
    - Represents maximum thermal stress
    - Up to 11.50°C
    - On average, 9.31°C, 8.20°C, 9.07°C for LR, MLP, RF respectively
Model Comparison

- Train offline
- Prediction time per placement

<table>
<thead>
<tr>
<th>Methods</th>
<th>LR</th>
<th>MLP</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>0.07ms</td>
<td>0.08ms</td>
<td>0.67ms</td>
</tr>
</tbody>
</table>

- Strengths
  - LR: many unknown tasks
  - MLP: all known tasks
  - RF: non-linear system environment
Scalability

• Partition nodes into S sets
• For each set, pick the hottest node to represent the set
• Perform predictions over partitions
• Do finer grain partitioning within the set recursively as needed

Applicability

• Logic utilization, RAM blocks and frequency are easily accessible metrics
Conclusion

• Temperature of the same task varies significantly across different FPGAs of the same generation
  ▪ Up to 11.50°C
• Propose thermal models
• Reduce peak temperature with our proposed static thermal-aware task placement algorithm
  ▪ On average 4.21°C
  ▪ Up to 11.50°C
Future Work

• Apply the framework at run-time
  ▪ Re-assign all tasks when a new task coming
  ▪ Calibrate models online

• Investigate other platforms
  ▪ Tightly coupled heterogeneous systems
  ▪ System-on-chips
Thank you!
Backup
Temperature & Power

- Linear relationship
### Peak Power Reduction

- Limiting factor for systems that operate under stringent power caps
- Reduction results

<table>
<thead>
<tr>
<th>Methods</th>
<th>LR</th>
<th>MLP</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1.48W</td>
<td>1.50W</td>
<td>1.49W</td>
</tr>
<tr>
<td>Max</td>
<td>3.45W</td>
<td>3.44W</td>
<td>3.40W</td>
</tr>
</tbody>
</table>