Evaluating Autotuning Heuristics for Loop Tiling

Tomoya Yuki¹, Yukinori Sato¹, Toshio Endo¹
¹Tokyo Institute of Technology

**Background**

Loop Tiling is a technique that uses cache hierarchy efficiently:
- It improves data locality of reference in nested loops.
- Tile sizes need to be properly adjusted in order to acquire better performance.
- Polyhedral compilers enable to automate the loop transformation.

We developed PATT to search proper tile sizes (Polyhedral compilation based and Tile size optimizer).
- Iterative building binary & measurement.
- Implemented some heuristics.

**Heuristics**

We have developed LPATT heuristic which is a specialized one for tile size adjustment:
- Load balance is important in the case of many core processors, which is the dominant factor for performance.
  - In the case of triply nested loops:
    - First, fix TJ and TK (inner loop tile sizes) and find fastest TI (outer most loop tile size).
    - Then, search TI and TK using the algorithm below.
  - In the case of doubly nested loops:
    - Search TI and TJ using the algorithm below.
  - 2-dimensional search algorithm: Starting from wide and coarse search space, it gradually focuses on a finer region.
  - In order to be over flat heatmap region and fast convergence:
    - We adopted hill climbing technique to reduce search steps.

Evaluation

- We use Polybench [3] as a benchmark suite, which is familiar with Polly.
  - This time we pick up gemm as a triply nested loop kernel and axet as a doubly one.
  - A few user setting values of each heuristics are optimized based on the result of our preliminary experiments.

**Motivation**

- Investigate the best tile size parameters for performance.
  - Need to search from a huge parameter space.
  - FAST convergence speed and HIGH accuracy.
- Many core processors such as XeonPhi (KNL) are becoming popular.

Heatmap that presents performance for each tile size (blue is fast):

Nelder-Mead (NM) is one of meta heuristics to find minimum:
- Using the concept of “Simplex”, which consists of problem dimension + 1 coordinates (e.g. 2-dimension: 3 coordinates: means an triangle).
- Moving a coordinate of a simplex, gradually reach the minimum

Simulated Annealing (SA) is also one of well-known meta heuristics:
- Using the concept of “Temperature”, which affects the probability below.
  - Selecting a random coordinate from neighbor.
  - Comparing the results of measurement, probabilistically moving the new coordinate.

We can see that the tile sizes of LPATT is nearby the one of Brute-Force.

**Table**

<table>
<thead>
<tr>
<th>Heuristics</th>
<th>Tile Sizes</th>
<th>Time</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brute-Force</td>
<td>16, 1124, 12</td>
<td>2.89</td>
<td>100.0</td>
</tr>
<tr>
<td>LPATT</td>
<td>16, 1392, 12</td>
<td>2.91</td>
<td>99.8</td>
</tr>
<tr>
<td>NM</td>
<td>16, 36, 46</td>
<td>3.36</td>
<td>86.0</td>
</tr>
<tr>
<td>NM (best)</td>
<td>16, 28, 44</td>
<td>3.50</td>
<td>82.6</td>
</tr>
<tr>
<td>NM (worst)</td>
<td>16, 24, 80</td>
<td>3.19</td>
<td>90.6</td>
</tr>
<tr>
<td>SA (best)</td>
<td>36, 36, 44</td>
<td>2.60</td>
<td>64.8</td>
</tr>
<tr>
<td>SA (worst)</td>
<td>36, 36, 44</td>
<td>2.60</td>
<td>64.8</td>
</tr>
</tbody>
</table>

“Brute-Force” is an exhaustive experiment.
We can see that the tile sizes of LPATT is nearby the one of Brute-Force.