Optimal Topology Exploration for Application-Specific 3D Architectures

O. Ozturk, F. Wang, M. Kandemir, and Y. Xie
Pennsylvania State University
Outline

- Introduction
- 3D Thermal Model
- ILP Formulation of Application-Specific 3D Placement
- An Example
- Experimental Evaluation
- Conclusion
Introduction

- 3D ICs: multiple device layers stacked together with direct vertical interconnects tunneling through them

- Advantages:
  - Reduction on global interconnect
  - Higher packing density and smaller footprint
  - Lower interconnect power due to reduction in total wiring length
  - Support for realization of mixed-technology chips
3D ICs

- Thermal Issues:
  - Higher cooling/packaging costs
  - Acceleration of failure mechanisms
  - Performance degradation.
- Thermal issues even more pronounced for 3D
  - Higher packing density
  - Especially for the inner layer of the die
  - A major hindrance for 3D integration
- **3D integration**: Need to be a thermal-aware design
Chip Multiprocessors

- Chip multiprocessor (CMP):
  - AMD Opteron, IBM Power5 and Intel Yonah
  - 2 cores now, soon will have 4, 8, 16, or 32
- Promising for embedded systems:
  - Performance: increasingly difficult to obtain more performance out of single-processor
  - Power consumption: lower frequency
  - Scalability: both loop-level and instruction-level parallelism
  - Cost: simpler design and verification
  - Area: better utilization of the available silicon area
3D CMPs: Placement of processors and storage blocks

- Placement of processors and storage blocks
  - Determine the data communication distances
  - Both power and performance depend on data communication distances
  - Frequently accessed data storage blocks should be placed close to the processor
  - Data block shared between two processors should be put close to both
Application-specific Placement in a Customized 3D Design

- Application-specific
  - Each embedded application can require a different placement for achieving the minimum data communication distances

- Our approach:
  - Integer linear programming (ILP) based placement
  - Constraints: thermal bounds
  - Objective: minimize data communication distances
3D Thermal model

- An 3D resistor mesh model
  - Based on Skadron’s Hotspot thermal model (lumped thermal resistances and thermal capacitances)
  - Employs thermal-electrical duality to enable efficient computation of thermal effects at the functional block level
- Transfer thermal resistance $R_{i,j}$ of block $i$ with respect to block $j$

\[
R_{i,j} = \frac{\Delta T_{i,j}}{\Delta P_{i,j}}.
\]

- Temperature rise for each block

\[
T = R \times P.
\]
ILP Formulation of Application-Specific 3D Placement

- Problem: Minimize data communication cost of a given application by determining the optimal placement of storage blocks and processor cores under a temperature bound

- A storage block corresponds to a set of consecutive cache lines
  - Data cache assumed to be divided into storage blocks of equal size

- In ILP formulation, we view the chip area as a 3D grid and assign processor cores and storage blocks into this grid
ILP Formulation

- ILP provides a set of techniques that solve optimization problems:
  - Objective function and constraints are linear functions
  - Solution variables restricted to be integers.
- In 0-1 ILP
  - Each (solution) variable is restricted to be 0 or 1.
- 0-1 ILP is used in this work for determining:
  - Storage block placements
  - Processor core placements
  - Under temperature bounds
ILP Formulation of Application-Specific 3D Placement

- Constant terms definition

<table>
<thead>
<tr>
<th>Constant</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>Number of processor cores</td>
</tr>
<tr>
<td>$M$</td>
<td>Number of storage blocks</td>
</tr>
<tr>
<td>$C_X$, $C_Y$, $C_Z$</td>
<td>Dimensions of the chip</td>
</tr>
<tr>
<td>$P_X$, $P_Y$</td>
<td>Dimensions of a processor core</td>
</tr>
<tr>
<td>$SIZE_m$</td>
<td>Size of a storage block $m$</td>
</tr>
<tr>
<td>$FREQ_{p,m}$</td>
<td>Number of accesses to storage block $m$ by processor $p$</td>
</tr>
<tr>
<td>$R_{l,v}$</td>
<td>Thermal resistance network</td>
</tr>
<tr>
<td>$T_B$</td>
<td>Temperature bound</td>
</tr>
</tbody>
</table>
Major Constraint Functions

- $MC_{m,x,y,z}$: indicates whether storage block $m$ is in $(x,y,z)$
- $MD_{m,x,y}$: indicates whether storage block $m$ has dimensions of $(x,y)$

Geometric:

$$
\sum_{i=0}^{C_X-1} \sum_{j=0}^{C_Y-1} \sum_{k=0}^{C_Z-1} MC_{m,i,j,k} = 1, \ \forall m.
$$

$$
\sum_{i=1}^{C_X} \sum_{j=1}^{C_Y} MD_{m,i,j} = 1, \ \forall m.
$$

Thermal:

$$
Temp_m = \sum_{j=1}^{P+M+1} R_{m,j} \times Power_j, \ \forall m.
$$

$$
Temp_m \leq T_B, \ \forall m.
$$
Objective Function

- $X_{dist_{p,m,x}}$: indicates whether the distance between processor $p$ and storage block $m$ is equal to $x$ on the $x$-axis

\[
X_{Cost} = \sum_{i=1}^{P} \sum_{j=1}^{M} \sum_{k=1}^{C_x-1} FREQ_{i,j} \times X_{dist_{i,j,k}} \times k.
\]

\[
Y_{Cost} = \sum_{i=1}^{P} \sum_{j=1}^{M} \sum_{k=1}^{C_y-1} FREQ_{i,j} \times Y_{dist_{i,j,k}} \times k.
\]

\[
Z_{Cost} = \sum_{i=1}^{P} \sum_{j=1}^{M} \sum_{k=1}^{C_z-1} FREQ_{i,j} \times Z_{dist_{i,j,k}} \times k.
\]

\[\min \ (\alpha \times (X_{Cost} + Y_{Cost}) + \beta \times Z_{Cost}).\]
An Example

- 4 processors and 20 storage blocks

<table>
<thead>
<tr>
<th>Processor</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>B5</th>
<th>B6</th>
<th>B7</th>
<th>B8</th>
<th>B9</th>
<th>B10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.20%</td>
<td>0.18%</td>
<td>61.76%</td>
<td>0.19%</td>
<td>0.17%</td>
<td>0.20%</td>
<td>3.48%</td>
<td>2.86%</td>
<td>0.20%</td>
<td>0.20%</td>
</tr>
<tr>
<td>2</td>
<td>0.20%</td>
<td>0.19%</td>
<td>61.86%</td>
<td>0.19%</td>
<td>0.22%</td>
<td>4.07%</td>
<td>2.30%</td>
<td>0.18%</td>
<td>0.19%</td>
<td>0.18%</td>
</tr>
<tr>
<td>3</td>
<td>0.18%</td>
<td>0.18%</td>
<td>61.83%</td>
<td>0.18%</td>
<td>0.18%</td>
<td>4.05%</td>
<td>2.34%</td>
<td>0.19%</td>
<td>0.21%</td>
<td>0.19%</td>
</tr>
<tr>
<td>4</td>
<td>0.18%</td>
<td>0.22%</td>
<td>61.76%</td>
<td>0.19%</td>
<td>0.18%</td>
<td>4.05%</td>
<td>2.32%</td>
<td>0.18%</td>
<td>0.20%</td>
<td>0.18%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Processor</th>
<th>B11</th>
<th>B12</th>
<th>B13</th>
<th>B14</th>
<th>B15</th>
<th>B16</th>
<th>B17</th>
<th>B18</th>
<th>B19</th>
<th>B20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.83%</td>
<td>0.22%</td>
<td>0.20%</td>
<td>0.18%</td>
<td>0.20%</td>
<td>0.19%</td>
<td>0.18%</td>
<td>0.19%</td>
<td>1.83%</td>
<td>4.55%</td>
</tr>
<tr>
<td>2</td>
<td>22.64%</td>
<td>0.20%</td>
<td>0.18%</td>
<td>0.18%</td>
<td>0.17%</td>
<td>0.19%</td>
<td>0.18%</td>
<td>1.91%</td>
<td>4.49%</td>
<td>0.28%</td>
</tr>
<tr>
<td>3</td>
<td>22.65%</td>
<td>0.20%</td>
<td>0.20%</td>
<td>0.22%</td>
<td>0.19%</td>
<td>0.20%</td>
<td>0.18%</td>
<td>1.89%</td>
<td>4.47%</td>
<td>0.29%</td>
</tr>
<tr>
<td>4</td>
<td>22.59%</td>
<td>0.17%</td>
<td>0.18%</td>
<td>0.18%</td>
<td>0.19%</td>
<td>0.21%</td>
<td>0.18%</td>
<td>1.89%</td>
<td>4.49%</td>
<td>0.31%</td>
</tr>
</tbody>
</table>
An Example

2D

3D

Layer 1

Layer 2
## Experimental Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of processor cores (in multi-core designs)</td>
<td>4</td>
</tr>
<tr>
<td>Number of blocks</td>
<td>24</td>
</tr>
<tr>
<td>Number of layers</td>
<td>2</td>
</tr>
<tr>
<td>$\frac{\alpha}{\beta}$</td>
<td>10</td>
</tr>
<tr>
<td>Total storage capacity</td>
<td>128KB</td>
</tr>
<tr>
<td>Set associativity</td>
<td>2 way</td>
</tr>
<tr>
<td>Line size</td>
<td>32 Bytes</td>
</tr>
<tr>
<td>Number of lines per block</td>
<td>90</td>
</tr>
<tr>
<td>Temperature bound</td>
<td>110°C</td>
</tr>
</tbody>
</table>
# Benchmarks

<table>
<thead>
<tr>
<th>Benchmark Name</th>
<th>Source</th>
<th>Description</th>
<th>Number of Data Accesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>ammp</td>
<td>Spec</td>
<td>Computational Chemistry</td>
<td>86967895</td>
</tr>
<tr>
<td>equake</td>
<td>Spec</td>
<td>Seismic Wave Propagation Simulation</td>
<td>83758249</td>
</tr>
<tr>
<td>mcf</td>
<td>Spec</td>
<td>Combinatorial Optimization</td>
<td>114662229</td>
</tr>
<tr>
<td>mesa</td>
<td>Spec</td>
<td>3-D Graphics Library</td>
<td>134791940</td>
</tr>
<tr>
<td>vortex</td>
<td>Spec</td>
<td>Object-oriented Database</td>
<td>163495955</td>
</tr>
<tr>
<td>vpr</td>
<td>Spec</td>
<td>FPGA Circuit Placement and Routing</td>
<td>117239027</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Benchmark Name</th>
<th>Source</th>
<th>Description</th>
<th>Number of Data Accesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>3step-log</td>
<td>DSPstone</td>
<td>Motion Estimation</td>
<td>90646252</td>
</tr>
<tr>
<td>adi</td>
<td>Livermore</td>
<td>Alternate Direction Integration</td>
<td>71021085</td>
</tr>
<tr>
<td>btrix</td>
<td>Spec</td>
<td>Block Tridiagonal Matrix Solution</td>
<td>50055611</td>
</tr>
<tr>
<td>tsf</td>
<td>Perfect Club</td>
<td>Nearest Neighbor Computation</td>
<td>54917732</td>
</tr>
</tbody>
</table>
Experimental Evaluation

- Normalized data communication cost of 2D-Opt and 3D-Opt w.r.t. 2D-Random

Single-Core

2D-Opt: 63% for single-core; 58% for multi-core
3D-Opt: 82% for single-core; 69% for multi-core

Multi-Core

2D-Opt: 63% for single-core; 58% for multi-core
3D-Opt: 82% for single-core; 69% for multi-core
Experimental Evaluation

- Normalized data communication cost of 3D-Opt and 3D-Random w.r.t 2D-Random

Single-Core

Multi-Core
Experimental Evaluation

- Effect of number of 3D layers (ammp)

<table>
<thead>
<tr>
<th>Number of Layers</th>
<th>Normalized Data Communication Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50%</td>
</tr>
<tr>
<td>2</td>
<td>25%</td>
</tr>
<tr>
<td>3</td>
<td>15%</td>
</tr>
<tr>
<td>4</td>
<td>10%</td>
</tr>
</tbody>
</table>
Experimental Evaluation

- Normalized data communication cost with respect to the temperature bound (ammp)
- Default: 110

![Graph showing normalized data communication cost with respect to temperature bound (°C). The graph compares different temperature bounds (110, 105, 100, 95, 90, 85, 80 °C) and their corresponding normalized communication costs.]
Conclusion

- Shrinking process technology and increasing data communication requirements of embedded applications
  - An increasing bottleneck: On-chip interconnects
  - Solution to the global interconnect problem: 3D designs
- Our goal: application-specific placement of processor cores and storage blocks in a customized 3D design
- Formulated using ILP
- Experiments with single-core and multi-core
  - Optimal placement of storage blocks and processor cores is very important in 3D design
Thanks