A Novel Dimensionally Decomposed Router for On-Chip Communication in 3D Architectures

J. Kim, et.al.
ISCA 2007

Goal

- Explore interconnection networks for packaging in three dimensions
  - Die stacks

- Distinguish from chip-to-chip networks

- Opportunities/Challenges
  - Non-uniformities in channel characteristics
    - Wire length, area
  - Arbitration extent
The Obvious Extension

- Does not exploit the low inter-layer distances
  - All hops are created equal
- Requires larger crossbars
- Using vertical buses reduces Xbar to 6x6


Extension Based on Full Connectivity

- A “true” 3D crossbar has complexity $O(N^3)$
- Proposal: multistage switching
  - Path from input at layer $i$ to switch output at layer $j$
- Routing complexity

Four layers, 5x5 xbar/layer, 6T/point = 600 control signals

Key Idea: Guided Flit Queuing

- Reduce crossbar size by *pre-routing* flows
  - Really moving wiring around
  - Reduces the arbitration complexity/cost

- Extensions to flows in three dimensions
  - Extant of arbiters is reduced

Guided Flit Queuing in Two Dimensions

- Based on isolation of flows
- Pre-routing at previous router or at the link controller

**Guided Flit Queuing in Three Dimensions**

Impact of vertically segmented design

<table>
<thead>
<tr>
<th>Inter-Layer</th>
<th>Number of Repeater</th>
<th>Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 μm (Layer 1 to 2)</td>
<td>0</td>
<td>7.86 ps</td>
</tr>
<tr>
<td>100 μm (Layer 1 to 3)</td>
<td>0</td>
<td>19.05 ps</td>
</tr>
<tr>
<td>150 μm (Layer 1 to 4)</td>
<td>0</td>
<td>36.12 ps</td>
</tr>
<tr>
<td>150 μm (Layer 1 to 4)</td>
<td>1 (layer 3)</td>
<td>105.14 ps</td>
</tr>
</tbody>
</table>

**Features**

- Vertical Xbar is not partitioned
  - Integrated into the row/column switches via single hop vertical traversals
- Two stage arbitration in vertical dimension for XYZ routing
  - Intra-layer → inter-layer
  - Assumes vertical wire delays are negligible
  - Maximize concurrency
- Additional arbitration @layer (exit) is required for other routing algorithms
  - Note extent of content of arbitration requests

Arbiter Complexity

\[ w = b + 2(n-1)^2 + 5(n-1) \quad \text{For non-XYZ} \]
\[ w = b + 2(n-1)^2 + 6(n-1) + 5(n-1) \quad \text{For other} \]

\[ b = \# \text{data bits/wires} \]
\[ 2(n-1)^2 = \# \text{rq/ack signals to/from arbiter} \]
\[ 6(n-1) = \# \text{additional signals for non-XYZ} \]
\[ 5(n-1) = \# \text{enable signals} \]

Performance


Figure 17: Average Latency with various Synthetic Traffic Patterns (XYZ routing)
Summary

- Exploiting physical properties
  - Non-uniformities in three dimensions

- Importance of arbitration complexity
  - Proportional to routing freedom

- Look for and eliminate redundancy in the design
  - Sensitivity to traffic patterns