Traffic- and Thermal-Aware Run-Time Thermal Management Scheme for 3D NoC Systems

Chih-Hao Chao, Kai-Yuan Jheng, Hao-Yu Wang, Jia-Cheng Wu, and An-Yeu (Andy) Wu

Presentation for NOCS2010 in Grenoble, France
May 6, 2010
Background and Motivation

- Die stacking 3D IC Technology + NoC → 3D NoC

- Advantage of 3D NoC System:
  - Small form factor
  - Small p2p latency
  - Large bandwidth

- Worse conditions of 3D IC/NoC
  - Larger power density
  - Longer heat conduction path
    - Different cooling efficiency

- Thermal issues of 3D NoC System
  - Higher possibility of thermal run-away
  - Vulnerability of performance, power, reliability, packaging cost for heat dissipation

Both performance and temperature have to be considered for 3D NoC Design
Motivational Example: Temperature and Performance of a 4-Layer 3D NoC

- Temperature is proportional to PIR and power density
  - Saturation throughput = PIR which has 2x latency of zero load (~15)
  - Assume thermal limit = 125°C

Network performance is prone to be limited by temperature in 3D NoC
Problem Description

- **Steady state optimization**
  - Maximize performance
    - Balance network loading
    - Balanced power distribution
    - Top layer has worse cooling efficiency than bottom layer
    - Prone to overheat at top layer
  - Maximize heat conduction
    \[
    \frac{\Delta Q}{\Delta t} = -k A_{\text{cross}} \frac{\Delta T}{\Delta x}
    \]
    - Maximize temperature on bottom layer
    - Concentrated network loading
    - Prone to congest at bottom layer

- **Transient state temperature control**
  - Prevent overheat occurrence
    on busy network, total injected power > thermal design power (TDP)
    - Temporary reduce heat injection by slowing down or turning off circuits
    - Suffer from performance loss
Overview of the Design Goal

- **Given** network topology (3D mesh), power model (80-core [8])
- **Find** a framework and policy for run-time thermal management (RTM) such that optimize performance and control temperature < thermal limit

![Diagram showing proactive and reactive thermal management strategies](image)

**Proactive performance optimization with traffic-awareness under thermal limit**

**1. Traffic-Aware Downward Routing**

**2. Thermal-Aware Vertical Throttling**
Related Works

- **RTM for 2D NoC**
  - Heat of NoC cannot be ignored
  - ThermalHerd [7]
  - Reactive Distributed traffic Throttling (DT) for emergency cooling

- **Power migration for 3D IC thermal optimization**
  - Thermal Herding [9]
  - Steer and herd the majority of switching activity to the heat sink

- **3D NoC Architecture**
  - Dimension-decomposed (DimDe) router [14]
  - Any vertical transmission within single cycle

---

**Partition the datapath of a 64-bit processor, LSB into 4 parts.**

**Overheated router**

Throttle: partially turn off
Proposed Downward Routing for Run-Time Power Migration

Downward Power Migration

\[ P = P_{src} + P_Z + P_X + P_Y + P_{dst} \]

Migrate horizontal routing power toward heat sink

Downward level = 0 (XYZ)  Downward level = 1  Downward level = 2  Downward level = 3
Traffic-Aware Downward Routing

- Prevent congestion at bottom layer by traffic-aware level selection

<table>
<thead>
<tr>
<th>Downward level (DL)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 0 load</td>
<td>$T_0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Layer 1 load</td>
<td>$T_1$</td>
<td>$T_0$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Layer 2 load</td>
<td>$T_2$</td>
<td>$T_1$</td>
<td>$T_0$</td>
<td>0</td>
</tr>
<tr>
<td>Layer 3 load</td>
<td>$T_3$</td>
<td>$T_2+T_3$</td>
<td>$T_1+T_2+T_3$</td>
<td>$T_0+T_1+T_2+T_3$</td>
</tr>
</tbody>
</table>

- Determine downward level for each pillar
  - Information exchange within a pillar can be done within a relative short time
- Dead-lock free (Proof in paper) as we adopt:
  - **Vertical routing**: downward or traffic-aware downward routing
  - **Horizontal routing**: arbitrary 2D routing

if \(\text{cnt}_3 \geq \text{threshold}\)

\[\text{DW}_{\text{level}} = 0;\]

else if \(\text{cnt}_3+\text{cnt}_2 < \text{threshold}\)

\[\text{DW}_{\text{level}} = 1;\]

else if \(\text{cnt}_3+\text{cnt}_2+\text{cnt}_1 < \text{threshold}\)

\[\text{DW}_{\text{level}} = 2;\]

else if \(\text{cnt}_3+\text{cnt}_2+\text{cnt}_1+\text{cnt}_0 < \text{threshold}\)

\[\text{DW}_{\text{level}} = 3;\]
Proposed Vertical Throttling for Emergency Cooling

- **Global throttling (GT)**
  - Large impact
  - Fast cooling

- **Distributed throttling (DT)**
  - Small impact
  - Slow cooling

- **Vertical throttling (VT)**
  - Middle impact
  - Middle cooling

**Key idea of vertical throttling on overheat**
- Create a vertical path toward heat sink for heat conduction
- Utilize the larger thermal conductance among the vertically aligned routers

**Trade off between cooling efficiency and performance impact**
Thermal-Aware Vertical Throttling

- Prevent large performance impact on vertical throttling
  - Gradually increase the level of throttling of the pillar which has overheat router if the temperature is still high
  - Always preserve fully bandwidth on bottom layer
  - Clear throttling level when temperature is smaller than thermal limit.
Simulation Environment

**Simulation Tools:**
- Traffic-Thermal Mutual-Coupling Co-Simulation Platform for 3D NoC [15]

**Configuration**
- 4x4x4 mesh
- \textit{DimDe} router architecture [13]
- Power Model: Intel 80-core [8]
  - Synthetic traffic
  - PE power in proportion to local traffic
- Buffer depth = 4, 6 flits per packet
- Wormhole flow control, no VC
- Round-robin arbitration
- Horizontal routing: DOR (XY)
- Thermal sensor sensing interval = 1ms
Proactive Technique: Achievable Throughput

Achievable throughput
≡ maximal PIR that:
  • not making temperature ≥ thermal limit
  • not saturate network (latency ≤ 2X zero load latency)

Uniform Traffic

<table>
<thead>
<tr>
<th>Thermal limit</th>
<th>60°C</th>
<th>80°C</th>
<th>100°C</th>
<th>120°C</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ (DW_level = 0)</td>
<td>0.0020</td>
<td>0.0059</td>
<td>0.0096</td>
<td>0.0133</td>
<td>0.0230</td>
</tr>
<tr>
<td>Fixed DW_level = 1</td>
<td>0.0023</td>
<td>0.0064</td>
<td>0.0104</td>
<td>0.0137</td>
<td>0.0160</td>
</tr>
<tr>
<td>Fixed DW_level = 2</td>
<td>0.0024</td>
<td>0.0065</td>
<td>0.0108</td>
<td>0.0112</td>
<td>0.0112</td>
</tr>
<tr>
<td>Fixed DW_level = 3</td>
<td>0.0025</td>
<td>0.0067</td>
<td>0.0083</td>
<td>0.0083</td>
<td>0.0083</td>
</tr>
<tr>
<td>Traffic-aware</td>
<td>0.0023</td>
<td>0.0063</td>
<td>0.0101</td>
<td>0.0140</td>
<td>0.0195</td>
</tr>
<tr>
<td>Improvement</td>
<td>15.00%</td>
<td>6.78%</td>
<td>5.21%</td>
<td>5.26%</td>
<td>-15.22%</td>
</tr>
</tbody>
</table>

Transpose Traffic

<table>
<thead>
<tr>
<th>Thermal limit</th>
<th>60°C</th>
<th>80°C</th>
<th>100°C</th>
<th>120°C</th>
<th>Infinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ (DW_level = 0)</td>
<td>0.0021</td>
<td>0.0058</td>
<td>0.0094</td>
<td>0.0131</td>
<td>0.0186</td>
</tr>
<tr>
<td>Fixed DW_level = 1</td>
<td>0.0024</td>
<td>0.0064</td>
<td>0.0103</td>
<td>0.0104</td>
<td>0.0104</td>
</tr>
<tr>
<td>Fixed DW_level = 2</td>
<td>0.0024</td>
<td>0.0066</td>
<td>0.0067</td>
<td>0.0067</td>
<td>0.0067</td>
</tr>
<tr>
<td>Fixed DW_level = 3</td>
<td>0.0025</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0050</td>
</tr>
<tr>
<td>Traffic-aware</td>
<td>0.0023</td>
<td>0.0062</td>
<td>0.0096</td>
<td>0.0121</td>
<td>0.0121</td>
</tr>
<tr>
<td>Improvement</td>
<td>9.52%</td>
<td>6.90%</td>
<td>2.13%</td>
<td>-7.63%</td>
<td>-34.95%</td>
</tr>
</tbody>
</table>
Reactive Technique: Number of Throttled Router and Throttling Time

- Number of throttled router

<table>
<thead>
<tr>
<th># of throttled router</th>
<th>Global Throttling (GT)</th>
<th>Vertical Throttling (VT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GT</td>
<td>DTT</td>
</tr>
</tbody>
</table>

- Avg. throttling time of the throttled router (ms)

<table>
<thead>
<tr>
<th></th>
<th>GT</th>
<th>DTT</th>
<th>VT</th>
<th>TAVT</th>
<th>Reduction (VT vs. DTT)</th>
<th>Reduction (TAVT vs. DTT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.147</td>
<td>14.333</td>
<td>2.481</td>
<td>4.353</td>
<td>82.69%</td>
<td>69.63%</td>
</tr>
<tr>
<td>Var.</td>
<td>0.125</td>
<td>2.222</td>
<td>0.324</td>
<td>0.581</td>
<td>85.42%</td>
<td>73.85%</td>
</tr>
</tbody>
</table>
Reactive Technique: Throttling Ratio and Network Availability

Throttling Ratio $\equiv \frac{\text{Inactive time slot of all routers}}{\text{All time slot of all routers}}$

<table>
<thead>
<tr>
<th>PIR</th>
<th>GT</th>
<th>DTT</th>
<th>VT</th>
<th>TAVT</th>
<th>Reduction (TAVT vs. DTT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.016</td>
<td>0.0160</td>
<td>0.0085</td>
<td>0.0105</td>
<td>0.0077</td>
<td>9.12%</td>
</tr>
<tr>
<td>0.022</td>
<td>0.0256</td>
<td>0.0126</td>
<td>0.0155</td>
<td>0.0108</td>
<td>14.68%</td>
</tr>
<tr>
<td>0.030</td>
<td>0.0320</td>
<td>0.0169</td>
<td>0.0175</td>
<td>0.0143</td>
<td>15.28%</td>
</tr>
</tbody>
</table>

Network Availability $\equiv 1 - \text{Throttling Ratio}$

<table>
<thead>
<tr>
<th>PIR</th>
<th>GT</th>
<th>DTT</th>
<th>VT</th>
<th>TAVT</th>
<th>Improvement (TAVT vs. DTT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.016</td>
<td>0.9840</td>
<td>0.9915</td>
<td>0.9895</td>
<td>0.9923</td>
<td>$1.0008 \times$</td>
</tr>
<tr>
<td>0.022</td>
<td>0.9744</td>
<td>0.9874</td>
<td>0.9846</td>
<td>0.9893</td>
<td>$1.0019 \times$</td>
</tr>
<tr>
<td>0.030</td>
<td>0.9680</td>
<td>0.9832</td>
<td>0.9825</td>
<td>0.9857</td>
<td>$1.0026 \times$</td>
</tr>
</tbody>
</table>
The thermal issues of 3D NoC are important and different to 2D NoC.

- **The optimization goal varies as the temperature change**
  - Optimize for performance and avoid overheat for steady state optimization
  - Emergency cooling to prevent overheat for transient state temperature regulation

- **We redefine the goal of RTM for 3D NoC systems:**
  - Maximize achievable throughput for steady state
  - Minimize performance impact for throttling

The proposed traffic- and thermal-aware run-time thermal management scheme for 3D NoC systems contains:

- **Traffic-aware downward routing**
  - Improve achievable throughput for temperature-limiting case

- **Thermal-aware vertical throttling**
  - Reduce the average throttling time than distributed throttling
  - Reduce the average throttling ratio and improves network availability
Thank you for your attention!


Improve Accuracy of Power and Thermal Modeling

- **Current approach:**
  - Traffic-thermal mutual-coupling co-simulation
  - Power model: based on Intel 80-core power breakdown
  - Thermal model: homogeneous intra-layer parameter

- **To be improved**
  - Include **leakage power** (correlated to temperature) and **clock tree power**
  - **Power ratio of uP/memory/router**: add Intel’s Single Chip Cloud Computer (SCCC) power model for CMP
  - **Heterogeneous intra-layer parameter** for TSV modeling
• Current status
  – Dimension-ordered (XYZ) router architecture, 3D mesh*
  – Dimension-Decomposed (DimDe) router, parallel single-hop vertical transmission, reduced vertical crossbar
  – Wormhole flow control
  – No virtual channel (VC) and small depth of input buffer (4 flits): deadlock prevention by turn model
  • Very low injection rate and throughput

• To be improved
  – Analyze performance variation with different router parameters and architectures
  – Realistic router design and implement for validation of performance/power/cost model
  – Include VC flow control/fully adaptive routing
  – Non-symmetric 3D network architecture and resource allocation: buffering channel, link bandwidth
Parameter Determination for RTM

- **Current Status**
  - Traffic-aware downward routing
    - Manually determining fixed triggering threshold for each pillar, according to prior traffic distribution
  - Thermal-aware vertical throttling
    - For HW implementation $2^x$
    - Throttling ratio (none, half, and full)
    - Reactively select throttling level

- **To be improved**
  - Determine triggering threshold according to network state
  - Analysis for performance, cost, and cooling efficiency of different throttling ratio and level → thermal-aware scheduling
Current Status

- No assumption, experiments on uniform and transpose traffic
  - Heavy loading and lots of long range traffic
- Local power in proportion to local transfer
  - $P_L = \alpha P_R$, where $\alpha$ is constant

To be improved

- Apply traffic of Rent’s rule and thermal-aware mapped traffic model
  - More realistic to mapped traffic
  - Less long range traffic
- Characterize local power and data transfer
  - $P_L = \alpha P_R$, where $\alpha$ is partition dependent
- Intra-inter layer traffic characteristics
  - Short of related works
  - Needs to consider TSV cost and network topology
Problem Simplification for 3D

Fourier’s law of heat conduction:
\[
\frac{\Delta Q}{\Delta t} = -kA_{cross} \frac{\Delta T}{\Delta x}
\]

- Assumption:
  - Power density is evenly distributed on each layer
  - Single heat sink at bottom

Steady state temperature: \( T_1, T_2, T_3, T_4 \)

For each node, equal flux:
\[
\sum \text{heat}_{in} = \sum \text{heat}_{out}
\]

For system, subject to:
\[
\text{heat}_{in, total} \leq \max(\text{heat}_{dis, sink})
\]
Penn State University’s Dimension Decomposed (DimDe) Router

- Utilize the beneficial attribute of a negligible inter-die distance in 3D chips.

1. Single hop count on vertical path
2. Reduce crossbar area
Proactive Technique: Temperature Distribution and Network Performance

Downward level = 0 (XYZ)  Downward level = 1  Downward level = 2  Downward level = 3

Uniform Traffic  Transposed Traffic
Problem Description

- Heat of NoC cannot be ignored

- **Similar temperature for processor and network**
  - **High switching activity**

- **Balance heat injection ≠ thermal safe**
  - Different cooling efficiency
  - **Policy of control changes as temperature change**
    - Low temperature: optimize for performance
    - High temperature: prevent overheat (temperature > thermal limit)

- Existing non-proprietary NoC simulators cannot support both traffic and temperature for coupling simulation
Related Simulation Platforms for Thermal-Aware NoC

- Simulation flow for thermal-aware mapping in 3D NoC [10]
  - 3D CTM-based thermal model
    - Open-source HotSpot [9]
  - Unidirectional coupling
    1. Synthesize a NoC model
    2. Power estimation from traffic model
    3. Run thermal simulation and examine if temperature fit spec.

- Run-time thermal management for 2D NoC, Sirius [6]
  - No release to the public
  - Mutual-coupling co-simulation
    A. Network model
    B. Module-level power model
    C. CTM-based thermal model
Validation with CFD-RC

- Follow the 2D IC validation approach of the CTM tool:
  - *HotSpot* [9] by *ANSYS* [8]
  - Temperature error within 4K
  - Relative correctness: if $T_1 > T_2$ in FEA tool, $T_1 > T_2$ in our simulator

- Validation of vertical temperature distribution and heat conduction
  - Reference tool: *CFD-RC* [7]

<table>
<thead>
<tr>
<th>Layer</th>
<th>CFD-RC (K)</th>
<th>Ours (K)</th>
<th>Error (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 0</td>
<td>340.98</td>
<td>344.92</td>
<td>+3.94</td>
</tr>
<tr>
<td>Layer 1</td>
<td>339.65</td>
<td>342.78</td>
<td>+3.13</td>
</tr>
<tr>
<td>Layer 2</td>
<td>337.09</td>
<td>338.48</td>
<td>+1.39</td>
</tr>
<tr>
<td>Layer 3</td>
<td>333.10</td>
<td>332.04</td>
<td>-1.06</td>
</tr>
</tbody>
</table>

- Temperature error is similar to [9]
- Relative correctness is assured
Configuration for Simulation of Typical Traffic Patterns

- **Experimental setup**
  - Topology: 4x4x4 mesh
  - Routing: XYZ routing
  - Packet size: 6 flits
  - Buffer depth: 4 flits
  - Power model and floorplan:
    - Intel 80-core tera-flops research processor [3]
    - 2D mesh-based NoC system

- **Offered traffic patterns**
  - Uniform traffic
  - Hotspot traffic
  - Transpose traffic

- **Avg. power density:**
  - Router: 2.61×10^6 W/m²
  - Memory: 5.13×10^5 W/m²
  - FPMAC: 3.78×10^5 W/m²