Aging-Aware Routing Algorithms for Network-on-Chips (NoCs)

Koushik Chakraborty
BRIDGE Lab
Electrical and Computer Engineering
Utah State University
Outline

• Background
  – NoC basics
  – Routing Algorithms
  – Reliability concerns
• Proactive Aging Mitigation in NoCs
• MILP-based Aging-aware Oblivious routing
• Aging-aware adaptive routing
• Conclusion
Why NoC

• Proliferation of on-chip cores
  – Raises the importance of communication arch.

• Need a design that is:
  – Scalable
  – Composable (lowers verification cost)

• Network-on-chip most promising
  – Many-core implementation exists (Tilera, Tera-flop)
NoC Architecture

- Components: Routers (r0, r1) and Links (l).
Routing Algorithms

• *Oblivious Routing*: Path determined by source and destination addresses (static).
  – Pros: Simpler router design
  – Cons: What about congestion?
• *Adaptive Routing*: Path decided based on runtime conditions (dynamic).
  – Pros: Tackles congestion.
  – Cons: complex router design.
Reliability Concerns: NBTI

• Negative Bias Temperature Instability (NBTI): Affect the PMOS transistors under
  – Negative bias
  – High temperature

• NBTI causes:
  – Increase in PMOS threshold voltage.
  – *Delay degradation in combinational/sequential circuit*
Reliability concern: Electromigration

• Electromigration: Reduction in interconnect widths causes:
  – Metal ions drift in the direction of electron flow causing voids.

• EM causes:
  – Reduction in effective conducting area of wire and increase in wire resistance.
  – *Increase in wire delay of interconnect.*
EFFECT OF AGING ON NOC FAULT-TOLERANCE
Effect of Aging on NoC

• Aging effects both NoC routers and links.
• Different Aging Scenarios:

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Degradation in Routers</th>
<th>Degradation in Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>r(N) [Fu, 2010]</td>
<td>NBTI</td>
<td>NONE</td>
</tr>
<tr>
<td>I(N)r(N)</td>
<td>NBTI</td>
<td>NBTI</td>
</tr>
<tr>
<td>I(E)r(N)</td>
<td>NBTI</td>
<td>Electromigration</td>
</tr>
<tr>
<td>I(NE)r(N)</td>
<td>NBTI</td>
<td>NBTI and electromigration</td>
</tr>
</tbody>
</table>
Aging in Links: Effect on Network Latency

- \text{l(NE)r(N)}: highest latency, \text{r(N)}: least latency

Consideration of NBTI and electromigration in links gives more realistic aging estimation.
Aging in Links: Effect on Fault-tolerance of NoC

- \( l(NE)r(N) \): faulty in 3 years, \( r(N) \): faulty in 8 years.
- Threshold: 66.6 % used in this study.

Consideration of NBTI and electromigration in links gives more realistic time-to-failure estimation.
Motivational Study: Summary

• Aging mechanisms affecting NoCs: NBTI (routers) and NBTI and Electromigration (links)
• Aging in links should not be ignored.
• Evaluation using a simple NoC model under moderately heavy traffic.
SYSTEM-LEVEL IMPACT OF AGING
System Level Aging in NoCs

• Utilization -> Thermal Profile -> Aging
• Utilization asymmetry in NoC components
  – Driven by routing policies, application characteristics, and topology
• Two types of asymmetry
  – Spatial Asymmetry: Different routers/links have different utilization.
  – Temporal Asymmetry: Same router/link exhibits variation in utilization during runtime.
NoC Mesh: Spatial Asymmetry

- Buffer utilization of different routers for Canneal benchmark run using GARNET [Agarwal, 2009].
- Aging order: $r1 > r0 > r2 > r5 > r4 > r6 > r3 > r7$
Temporal Asymmetry: Buffer Utilization for Router R1 for Different Epochs
Aging Mitigation

• Reactive Approach
  – Once components fail, adapt routing
  – Typically use some form of probabilistic flooding
  – Design goal: limit overhead while maintaining error free communication

• Proactive Approach
  – Take action before failure happens
  – Understand the impact of current usage on long-term performance under wearout/aging
TAC: Reliability Metric for NoCs

• Metric to connect spatial asymmetric utilization with aging.
• Traffic Acceptance Capacity (TAC): fraction of nominal traffic that a stressed link/router should accept to maintain a target long-term performance
• TAC system-level modeling captures:
  – Differential router/link utilization.
  – Effect of NBTI on routers.
  – Effect of NBTI and electromigration on links.
TAC: A Case Study

- 4 Parsec benchmarks generate traffic in GARNET NoC simulator.
- TAC for stressed routers (TAC_r) (%):

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Max TAC_r</th>
<th>Min TAC_r</th>
<th>Avg TAC_r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canneal</td>
<td>82.3 (R7)</td>
<td>68.5 (R1)</td>
<td>73.4</td>
</tr>
<tr>
<td>Dedup</td>
<td>73.13</td>
<td>43.75</td>
<td>53.6</td>
</tr>
<tr>
<td>Facesim</td>
<td>63.5</td>
<td>29.24</td>
<td>39.67</td>
</tr>
<tr>
<td>Ferret</td>
<td>65.06</td>
<td>31.35</td>
<td>40.24</td>
</tr>
</tbody>
</table>
TTpE: A runtime variant of TAC

• Metric to connect BOTH spatial and temporal asymmetric utilization with aging.
• Traffic Threshold per Epoch (TTpE): fraction of nominal traffic that a stressed link/router should accept per epoch.
• Changes over the lifetime: useful for adaptive routing techniques
MILP-BASED AGING-AWARE OBLIVIOUS ROUTING
Why MILP?

• Design Goals:
  – A routing algorithm that considers:
    • Aging constraints.
    • Power-performance constraints.

A multi-objective optimization using an MILP-based routing technique.
MILP-Based Aging-Aware Routing

- Link/router delay variation
- TAC₀ and TAC₁
- Maximum hop count

Aging Constraints

Energy and latency constraints

Multi-objective Optimization using CPLEX

Aging and power-performance aware routes
Aging Constraints

- **Router/link delay variation:**
  - Delay experienced by a flit due to stressed router:
    \[ dr_s = dr_{us} + \Delta_{NBTI} \]
  - Delay experienced by a flit due to stressed link:
    \[ dl_s = dl_{us} + \Delta_{NBTI}, EM \]

- **Stressed router/link utilization must be under its threshold.**
  - Given by \( TAC_r \) and \( TAC_l \).
MILP-Based Aging-Aware Routing

- Link/router delay variation
- $TAC_r$ and $TAC_l$
- Maximum hop count

Aging Constraints

- Energy and latency constraints

Multi-objective Optimization using CPLEX

Aging and power-performance aware routes
Power-performance Constraints

• For each source, destination pair, k:

\[ hp_k \leq H \]

• Energy and delay modeling based on no. of hops.
• Flow conservation and deadlock avoidance constraints.
Multi-objective optimization

• Minimize:
  – Total no. of links utilized for all flows.
  – Total delay due to all routers and links.
    • Both stressed and unstressed
  – Energy-Delay-Product-Per-Flit (EDPPF) [Li, 2008]
    • A metric to provide trade-off between power and delay.
Experimental Methodology

• Combination of:
  – SPICE level analysis and statistical timing analysis for PV and NBTI aging [Bhardwaj, 2006].
  – wire resistance modeling of EM [Sun, 2002].
  – full system architectural simulation.

• GARNET NoC simulator embedded inside GEMS.

• 4 X 4 mesh topology.

• Traffic generation by PARSEC benchmarks.
Comparative Schemes

• **TAC-AGE**: Aging effects modeled in routers and links. XY routing used.

• **MIP-ROUT**: Uses the proposed MILP-based aging aware routing while including aging effects from TAC-AGE.
  
  – Overhead calculation: w.r.t. an aging-unaware routing without any aging in routers/links.
Overhead analysis: Network Latency

Lower is Better
Overhead analysis: Energy-Delay-Product-Per-Flit (EDPPF)

Lower is Better
AGING-AWARE ADAPTIVE ROUTING
Aging-aware Adaptive Routing

• **Congestion and aging-aware routing:** Select the shortest path that is:
  – Least affected by aging (minimum aging score)
  – Least congested (minimum congestion score)

• **Honoring TTpE by recovery cycles:** Idle cycles added for an epoch when TTpE reached:
  – Physical Significance: Additional time to recover from stress.
  – Long term performance assurance
Comparative Schemes

- **RCA-1D**: Congestion-aware routing without aging-awareness.
- **AGE-ADAP-REC**: AGE-ADAP scheme that inserts recovery cycles when TTpE reached.
Network Latency

Lower is Better

Normalized Latency

- RCA-1D
- AGE-ADAP
- AGE-ADAP-REC

Lower is Better
Conclusion

- Model impact of aging on both NoC routers and links.
- Develop a novel metric (TAC) to measure the effects of aging on NoC architecture.
- Minimize the aging impact using two aging-aware routing algorithms:
  - MILP-based aging-aware oblivious routing.
  - Aging-aware adaptive routing algorithm.
Thank You!