Intra-application Interference on Dragonfly Network

Xin Wang∗, Xu Yang∗, Misbah Mubarak†, Robert B. Ross†, Zhiling Lan∗

∗ Department of Computer Science, Illinois Institute of Technology, Chicago, IL, 60616
† Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 60439
Motivation

Performance Variability

![Graph showing average messaging rates for batch jobs running a laser-plasma interaction code on three architectures: IBM Blue Gene/Q (Mira), Cray XE6 (Hopper), and IBM Blue Gene/P (Intrepid).]

- The execution time on Cray XE systems can range from 28% faster to 41% slower than the average observed performance.

Motivation

Application Interference

- Application interference is a dominant cause of performance variability.
- Application interference can be broadly characterized as intra- and inter-application interference.
  - Intra-application interference is introduced by the processes within an application.
  - Inter- application interference is introduced by concurrently running applications.
Goals

• How intra-application interference affects application performance on Dragonfly network when different placement & routing configurations are in use?

• How to prevent (alleviate) intra-application interference on dragonfly networks?
Overview of Dragonfly Topology

- Theta Architecture
Overview of Dragonfly Topology

- Theta Architecture
Overview of Dragonfly Topology

- Theta Architecture
Job Placement and Routing Policies

- Job placement policies
  - Contiguous (cont)
  - Random (rand)
    - Random-chassis (chas)
    - Random-cabinet (cab)

- Routing policies:
  - Minimal routing (min)
  - Adaptive routing (adp)
Applications

1. **AMG** conforms to stringent 3D nearest neighbor communication pattern.

2. **MultiGrid** conforms to "many-to-many" communication pattern, with intensive communication between neighboring ranks, resembles nearest neighbor communication.

3. **CrystalRouter** conforms to “many-to-many” communication pattern along a hypercube overlay, but with a substantial portion of the communication occurs in small neighborhoods of ranks.

<table>
<thead>
<tr>
<th></th>
<th>Num. Rank</th>
<th>% Comm</th>
<th>Avg. Data/ Rank</th>
<th>Total Data Transferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMG</td>
<td>1728</td>
<td>52.90</td>
<td>0.69 MB</td>
<td>1.2 GB</td>
</tr>
<tr>
<td>MultiGrid</td>
<td>1000</td>
<td>3.72</td>
<td>3.04 MB</td>
<td>2.97 GB</td>
</tr>
<tr>
<td>CrystalRouter</td>
<td>1000</td>
<td>44.93</td>
<td>118.27 MB</td>
<td>115.5 GB</td>
</tr>
</tbody>
</table>

DOE DesignForward Project: [http://portal.nersc.gov/project/CAL/designforward.htm](http://portal.nersc.gov/project/CAL/designforward.htm)
Evaluation Metrics

- Message Load per Rank
  measures the communication behavior of an application

- Communication Time
  time each rank spends in completing all its message exchanges with other ranks.

- Average Hops
  measure the message locality of a packet.

- Link Saturation Time
  the time period when the buffer of a router’s port is full.
Application Analysis

(a) CR Message Load

(b) AMG Message Load

(c) MG Message Load

(d) CR Communication Time

(e) AMG Communication Time

(f) MG Communication Time
Application Analysis

Crystal Router Metrics

• Contiguous placement preserves message locality, resulting in less message transfer time at the cost of more local link saturation.

• Random node placements spread the message load over the whole networks, resulting less local link saturation but suffering more message transfer time.

• Minimal routing help preserve message locality by reducing the number of hops.

• Adaptive routing help balance network traffic to alleviate local link congestion.
Application Study Summary

*Intra-application interference is basically a trade-off between localized message and balanced network.*

- *Localized message* reduces the number of hops, resulting in less message transfer time at the cost of potential local link saturation.

- *Balanced network* distributes message load across all channels, resulting in less link saturation at the cost of higher message transfer time with extra number of hops.

- Applications that constantly exchange messages such as Crystal Router and MultiGrid benefit from balanced network traffic.

- Applications that do not exchange messages often such as AMG benefit from localized message.
Sensitivity Study

• Message Load

• Communication Frequency
Message Load

TABLE II: Message Load Statistics

<table>
<thead>
<tr>
<th>Application</th>
<th>Total Msg Load (MB)</th>
<th>Peak Msg. Load/Rank (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D-S</td>
<td>15.63</td>
<td>0.78</td>
</tr>
<tr>
<td>2D-M</td>
<td>164.06</td>
<td>8.20</td>
</tr>
<tr>
<td>2D-L</td>
<td>13125.00</td>
<td>656.25</td>
</tr>
</tbody>
</table>

(d) 2D-S Local Link Saturation  
(e) 2D-M Local Link Saturation  
(f) 2D-L Local Link Saturation
Communication Frequency
Sensitivity Study Summary

• Applications with low message load or frequency can benefit from localized message

• Applications with high message load or frequency can benefit from balanced network
Mitigation Solution

(a) CR Communication Time

(b) AMG Communication Time

(c) MG Communication Time

(Crystal Router) Metrics

(a) Average Hops

(b) Local Link Saturation

(c) Global Link Saturation
Conclusions

• We have presented a comparative analysis of intra-application interference under various job placement and routing strategies

• Intra-application interference is a trade-off between localized message and balanced network

• Applications with low message load or frequency benefit from localized message; applications with high message load or frequency benefit from balanced network

• Random-cabinet and random-chassis placement policies significantly improve the application performance

• In the future, We plan to investigate task mapping for diversified workloads

• We will also study the joint actions among application and system to minimize performance variability on Dragonfly systems
Thank You!

Q&A