A New Mechanism for Congestion and Deadlock Resolution

Yong Ho Song and Timothy Mark Pinkston

SMART Interconnects Group
Electrical Engineering - Systems
University of Southern California
Outline

- Introduction
- Congestion Detection using Ping Mechanism
- Congestion Resolution using Bubble Mechanism
- Performance Evaluation
- Conclusion
Obstacles to High Performance Communication

- Increasing demand on high performance communication in parallel computing systems
  - Parallel applications
  - Network applications (Multimedia, On-line Transaction, Internet, etc.)

- Contention due to limited amount of network resources and unpredictable traffic behavior
  - Congestion
    - Temporal Blockage in Packet Communication
  - Deadlock
    - Network Congestion with Cyclic Resource Dependency
    - Permanent Blockage in Packet Communication unless Properly Handled
Previous Work

- **Injection limitation for congestion handling**
  - **Local Detection**
    - Detects network congestion at end nodes with no global consensus
    - Restricts the injection of new packets
  - **Global Detection**
    - Detects network congestion inside the network
    - Propagates the detection toward end nodes
    - Restricts the injection of new packets

- **Flow control for congestion handling**
  - Bubble Router: Bubble flow control mechanism that guarantees enough free buffer space in the network to avoid deadlock.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Detection Criteria</th>
<th>Dynamic Threshold</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static Threshold</td>
<td># Busy Virtual Channels</td>
<td></td>
<td>Local</td>
</tr>
<tr>
<td>History Table</td>
<td># Busy Virtual Channels</td>
<td>✓</td>
<td>Local</td>
</tr>
<tr>
<td>LIFE</td>
<td>Alignment betw’n Src &amp; Dest</td>
<td>✓</td>
<td>Local</td>
</tr>
<tr>
<td>DRIL</td>
<td>Status of Pending Msg Queue</td>
<td>✓</td>
<td>Local</td>
</tr>
<tr>
<td>Reactive</td>
<td>Timeout</td>
<td></td>
<td>Global</td>
</tr>
<tr>
<td>ALO</td>
<td>Status of Virtual &amp; Physical Channels</td>
<td></td>
<td>Local</td>
</tr>
<tr>
<td>GCC</td>
<td># Full Buffers and Throughput</td>
<td></td>
<td>Global</td>
</tr>
</tbody>
</table>
Resolution of Network Congestion

- Needs Intelligence in Packet Scheduling
Motivations

- To achieve high performance communication
  - Efficient packet delivery
  - Efficient handling of network congestion

- Need for effective congestion detection and resolution techniques
  - To prevent the network from developing deep congestion
  - Precise detection of congestion configuration
  - Precise control over the movement of bubbles to resolve network congestion
In a Nutshell: Detection and Resolution of Cyclic Network Congestion

Detection
- Triggered on timeout
- Generates / propagates a special control packet along the congestion
- Cyclic congestion path

Resolution
- Shift operation along the detected congestion path
- No bubble along the path
- Need to create a temporary bubble
- Need to control over the bubble movement
In a Nutshell: Detection and Resolution of Acyclic Network Congestion

- **Detection**
  - Triggered on timeout
  - Generates / propagates a special control packet along the congestion
  - Acyclic congestion path

- **Resolution**
  - Shift operation along the detected congestion path
  - At least one bubble exists along the path.
  - Need to control over the bubble movement
Challenges

- How to precisely detect network congestion?
  
  **Ping Mechanism**

- How to make bubbles travel along detected congestion path?
  
  **Resource Reservation (Scheduling)**

- How to make sure that resource reservation does not cause deadlocks?
  
  **Cancellation Mechanism**
**Congestion Detection**

- **Ping**: a vehicle for detecting network congestion.
  - A small control packet
  - Traverses network resources through control network
    - In-Band or Out-of-Band implementation of control network
    - Even in in-band implementation, ping propagation consumes none or little of data transmission bandwidth.

- **Congestion**
  - Non-availability of network resources at input and output ports.

---

**Diagram Description**

- **Switch 1**
  - Input Port
  - Output Port

- **Switch 2**
  - Timeout
  - In Packet Delivery
  - Timeout

- **Switch 3**
  - Congestion Detected

- **Switch 4**
  - Congestion Continues
  - No More Congestion

---

**Additional Information**

- [SMART](http://www.usc.edu/dept/ceng/pinkston.SMART.html)
Ping Generation

- **Ping generation criteria:**
  - The occupancy of the input port > $O_{threshold}$
  - The packet at the head of the queue cannot advance for more than $T_{threshold}$.
  - The output port needed for forwarding the packet at the head of the queue has no bubbles.
  - The input port is not an injection port.

- The output port needed for forwarding the packet at the head of the queue is not reserved by any other ping arriving through a different input port.
Ping Propagation

- Ping propagation criteria:
  - The occupancy of the input port > $O_{threshold}$
  - The packet at the head of the queue cannot advance for more than $T_{threshold}$.
  - The output port needed for forwarding the packet at the head of the queue has no bubbles.
  - The input port is not an injection port.
  - The output port needed for forwarding the packet at the head of the queue is not reserved by any other ping arriving through a different input port.
Resource Reservation

- The resources of associated input and output ports are reserved for future congestion dispersion activities.

Ping Table
- Maintains reservation status in each router.
- An output port can be reserved by only one input port.
- Updated whenever the router generates or propagates a ping.
- Cleared on packet progress in the reserved path between input and output port, or on cancellation request.

Ping Path

<table>
<thead>
<tr>
<th>Ping Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ping ID (hex)</td>
</tr>
<tr>
<td>ABCD0001</td>
</tr>
</tbody>
</table>
Ping Termination

- Pings are terminated when one or more following conditions are met:
  - **Acyclic congestion**
    - Bubbles in either input or output ports
  - **Cyclic congestion**

- The ping-terminating router is responsible for providing a bubble for the detected congestion.
Challenges

- How to precisely detect network congestion?
  - Ping Mechanism

- How to make bubbles travel along detected congestion path?
  - Resource Reservation (Scheduling)

- How to make sure that resource reservation does not cause deadlocks?
  - Cancellation Mechanism
Congestion Resolution using External Bubble

- For acyclic congestion, the terminating router should provide an external bubble for the chained reservation.
For cyclic congestion, the detected cycle could be involved in forming deadlocks. Therefore, no external bubble may be available for the chained reservation.

- The terminating router provides an internal bubble by sheltering a packet of an associated input port into the acceleration buffer.
Scheduling Bubbles

- When a bubble appears on a ping path, it should be forwarded within the path.
- The successful forwarding of bubbles clear (or invalidate) the corresponding entry of the ping table.
- When no reservation is found, the scheduler operates on normal scheduling policy.
Challenges

- How to precisely detect network congestion?
  - Ping Mechanism

- How to make bubbles travel along detected congestion path?
  - Resource Reservation (Scheduling)

- How to make sure that resource reservation does not cause deadlocks?
  - Cancellation Mechanism
Reservation Cancellation

- A cancellation packet tracks the reverse path of the ping and clears corresponding reservation up to the ping-generating router.
- Cancellation Packet = Ping Packet with Cancellation bit set
- The ping generator may retry the congestion detection after an arbitrary amount of idle time.
Router Architecture

- Complexity of Ping Table
  - $O(P)$; where $P = \#oports$
- Complexity of Routing Failure Table
  - $O(P)$, where $P = \#iports$
Performance Evaluation

- FlexSim 1.3 (A flit-level network simulator, USC)
  - Augmentations
    - Ping-based resource reservation
    - Bubble mechanism
    - Statistics

- Network Configuration
  - 8 x 8 torus
  - Virtual cut-through
  - Message size of 32 flits
  - 4 virtual channels per physical link
  - Random traffic pattern
Effect of Congestion Handling

- The ping and bubble technique implemented on networks with deadlock-free routing.
The ping and bubble technique implemented on networks with the true fully adaptive routing (TFAR)

- Results of ping operations (per 100 pings generated)
- Effect of shift operations on packet dependency (per 100 packets shifted)
Effect of Deadlock Handling

- Handling techniques used for comparison.
  - Duato’s adaptive routing (Deadlock Avoidance)
  - Disha routing (Deadlock Recovery)
  - Ping and Bubble routing (Deadlock Recovery)
Effect of Cooperation with a Local Congestion Control

- Congestion detection and resolution through Ping and Bubble
- Local congestion control through ALO
  - At least one free virtual channel on all candidate physical links or
  - All virtual channels free on a physical link
- The network performance improves further via the cooperation of PnB and ALO.
Conclusion

- This work proposes an efficient handling mechanism for handling network congestion.
  - Precise detection of congestion configuration
  - Accurate resolution of detected congestion
  - Can be used as a recovery technique in True Fully Adaptive Routing (TFAR)

- Ping and Bubble mechanism
  - Distributed detection and resolution
  - Applicable to all network topologies
  - Harmless false detection
  - Implicit throttling of packet injection at end nodes
Introduction to Bubble

- The unit of an empty buffer space needed for holding a packet is defined as a bubble.

- Each forward movement of a packet in one direction is equivalent to the backward propagation of a bubble in the opposite direction.
Ping Interlocking

Output Port is reserved by other ping

Output Port is reserved by other ping
Overhead of Ping Transmission

- **Implementation of Control Channels** which is used for ping propagation
  - **In-Band**
    - Control packets are transmitted through data channels
  - **Out-of-Band**
    - Control packets are transmitted through separate control channels
    - No consumption of channel bandwidth for data transmission

- Little overhead due to in-band ping transmission
  - Pings are generated and propagated only when network links are congested.
Performance in Large-Scale Networks

- Both recovery techniques suffer from performance degradation.
  - Disha
    - Insufficient recovery bandwidth
  - Ping and Bubble
    - Long ping traces interfere with other ping propagation

- One Solution
  - Each router limits the injection of new packets when any of its resources has been reserved (congested) for the last $t$ network cycles.
True Fully Adaptive Routing (TFAR)

- Provides maximum freedom for routing function.
  - Increases Resource Availability
  - Decreases Resource Contention
  - Maximizes Network Utilization

- Not completely free from deadlock anomalies
  - Due to fully relaxed routing restriction
  - Requires detection and resolution mechanisms for deadlocks

- Requirements to deploy TFAR in networks
  - Accuracy in Deadlock Detection
  - Efficiency in Deadlock Resolution
Deadlock Handling Techniques for TFAR

- **Disha**
  - Time-out based detection
  - Recovery through Deadlock Buffers
  - No control over new packet injection

- **Software Recovery**
  - Detection by monitoring packet activities in physical links
  - Sink and re-inject
  - Injection limitation near network saturation