

A Localized Congestion Control Mechanism for PCI Express Advanced Switching Fabrics

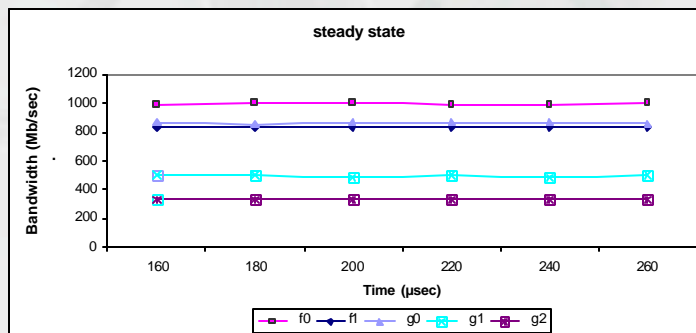
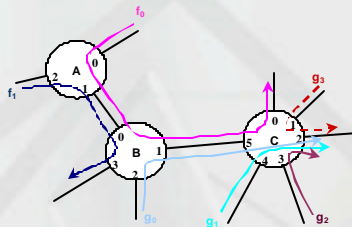
Venkata Krishnan & David Mayhew

StarGen Inc.
www.stargen.com

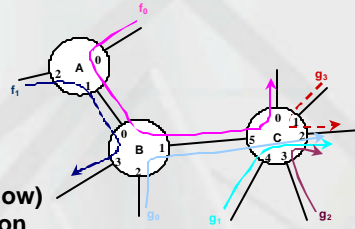
August 24, 2004



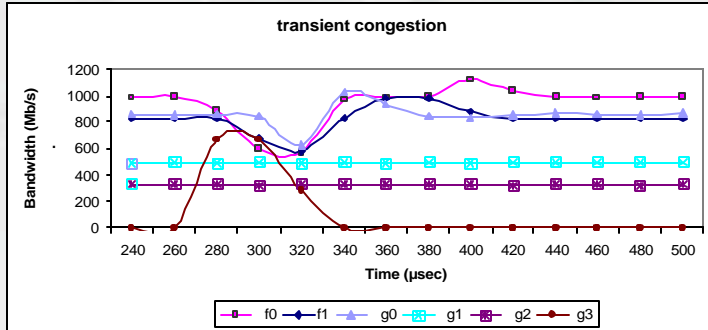
Transient Congestion Problem



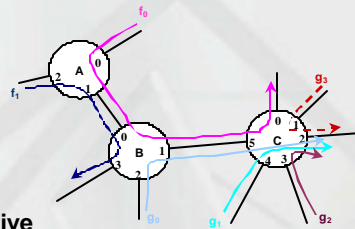
Transient Congestion Problem



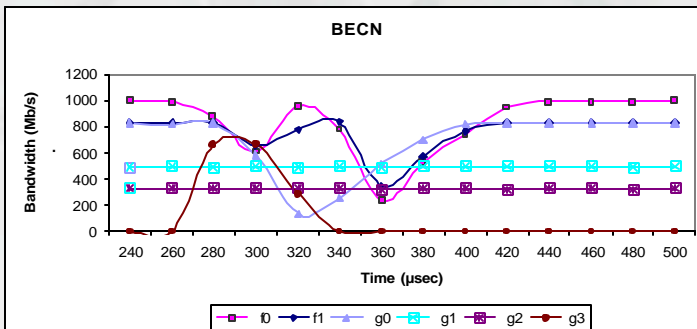
g₃ (intermittent flow) causes congestion



Transient Congestion Problem



ECN is ineffective



Possible Solutions

1. Increase size of (input) buffers

- Use on-board memory for packet buffers
- Inappropriate for single-chip switch solutions
 - Size of on-chip memory is an issue
 - Credit-based flow control exacerbates problem
 - Guaranteed storage for different VCs or VLs

2. Drop packets

- Rely upon upper level protocols to provide reliability (*Ethernet*)
- Inappropriate for protocols that provide reliability at the transaction layer (*Advanced Switching*)

3. Scheduling



5

Hot Interconnects 2004

Outline

- **Problem Statement**
- **PCI Express/Advanced Switching (ASI) Features**
 - Virtual Channels
 - Credit based Flow Control
 - ASI Packet Routing
- **Localized Congestion Management**
- **Results**
- **Closing Remarks**



6

Hot Interconnects 2004

PCI Express & Advanced Switching



Host Based Model

- Single OS domain
- Tree topology
- Best availability/ease of use
- Single host



Distributed Model

- Multiple protocols & topologies
- Peer to peer
- Priority for QoS mechanisms
- Priority for high availability & redundancy

PCI Express*
Transaction Layer

Advanced Switching
Transaction Layer

PCI Express Architecture
2.5Gbs PHY & Reliable Link Layers

Two Interconnects, Two Complementary Solutions



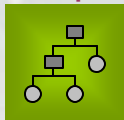
7

Hot Interconnects 2004

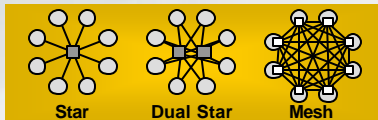
ASI Transaction Layer Adding to PCI Express Capability

- **Peer-to-Peer communications**
- **Multiple Data Transport models**
 - PCI Express (PI-8); Load Store (SLS); Queuing (SQ); RDMA (SDT)
- **Encapsulation**
 - PCI Express, Ethernet & future defined PI's (i.e. FC, ATM, etc.)
- **QoS Mechanisms**
 - Path and Multicast Routing
 - *Congestion Management*
 - 20 VC's (16 Unicast; 4 Multicast), 8 TC's + Token Buckets
 - SAR'ing
- **Reliability and High Availability Features**

PCI Express



Advanced Switching

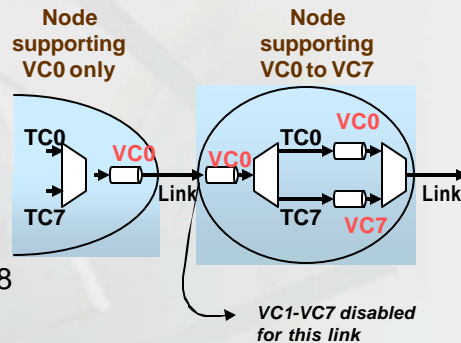


8

Hot Interconnects 2004

Virtual Channels

- Independent entities multiplexed onto a link
- Supports QoS
- Ordered, Bypass-capable, Multicast VCs
- E.g. Bypass-capable VCs
 - Header specifies 1 of 8 Traffic Classes (TC)
 - Hardware implements 1 to 8 VCs (TCs \geq VCs)
 - Link-level VC compatibility
 - Switch may use VC0 on 1 link and VC0-VC7 on another link



9

Hot Interconnects 2004

Credit based Flow Control (CBFC)

- Used in PCI Express and ASI
- Works on a per-VC basis
 - VC Buffer management using credits
 - Sender must have sufficient credits for transmitting a packet
 - No packet drops allowed
 - Packet always finds room in receiver
 - Credit updates managed using Link layer packets (DLLPs)
 - Cumulative *consumed* and *received* credit counters accommodates lossy nature of DLLPs
 - » DLLPs are lossy but do not consume credits
 - » TLPs (transaction layer packets) are not lossy but use flow credits



10

Hot Interconnects 2004

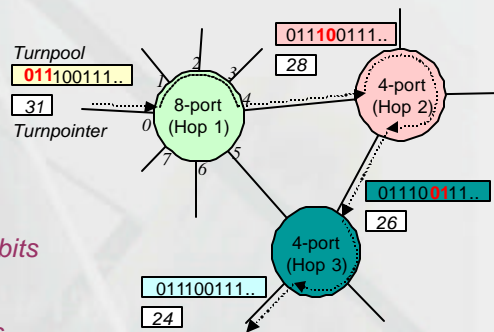
Packet Routing

- **Destination based**
 - Destination address specified in packet
 - Table lookup at each hop
 - Switch makes forwarding decision
 - Source address also specified
 - For responses
 - Appropriate when
 - Path can vary dynamically
 - Environments with several 1000s of nodes (e.g. Ethernet)
- **Source based**
 - No destination or source address
 - Source specifies path to destination
 - Switch plays minor role in forwarding
 - Reduced complexity
 - Source address not required
 - Reverse path generation quite simple
 - Appropriate when
 - path is deterministic
 - Environments with 10s to few 1000s of nodes (e.g. AS)



Packet Routing in ASI

- **Unicast Routing**
 - Source-based routing
 - Path Components
 - Turnpool
 - sequence of turns taken by a packet
 - Turn pointer
 - starting position of bits to choose from Turnpool
 - N port switch consumes $\lceil \log_2 N \rceil$ bits from Turnpool
- **Multicast Routing**
 - Destination-based routing
 - Table lookup for multicast group id



Forward Routing



Outline

- **Problem Statement**
- **PCI Express/Advanced Switching Features**
- **Localized Congestion Management**
 - Status Based Flow Control (SBFC)
- **Results**
- **Closing Remarks**

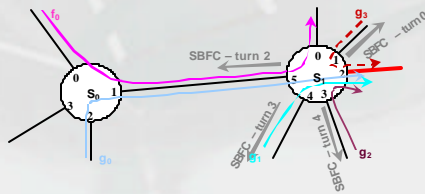


13

Hot Interconnects 2004

Status Based Flow Control (SBFC)

- **Detection & Notification**
 - Applicable only to ASI switches
 - Detection based on thresholds (e.g. VOQ-occupancy)
 - Notification sent only on link on which new packet received (i.e. not broadcast on all ports)
- **Response**
 - Applicable to any ASI node upstream (switches & endpoints)
 - Scheduler throttles congested flow allowing high priority for flows targeting non-congested ports (for a specific duration)



14

Hot Interconnects 2004

Status information

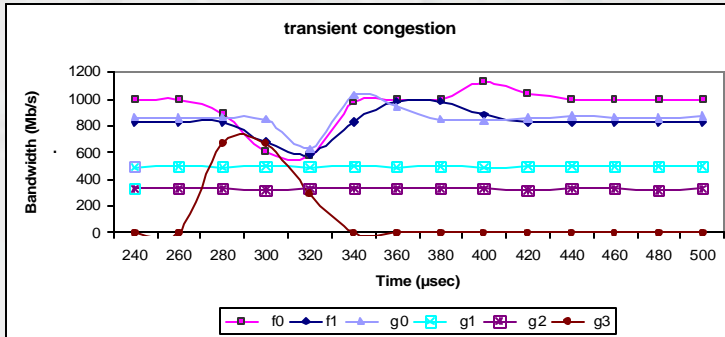
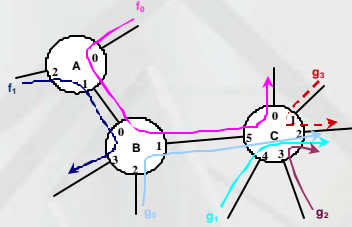
Value	Description
0	No Congestion (clear - Equivalent to XON)
1	Low congestion Threshold (short timer)
2	High congestion Threshold (long timer)
3	Severe Congestion (very long timer - Equivalent to a traditional XOFF)

Outline

- Problem Statement
- PCI Express/Advanced Switching Features
- Localized Congestion Management
- Results
- Closing Remarks

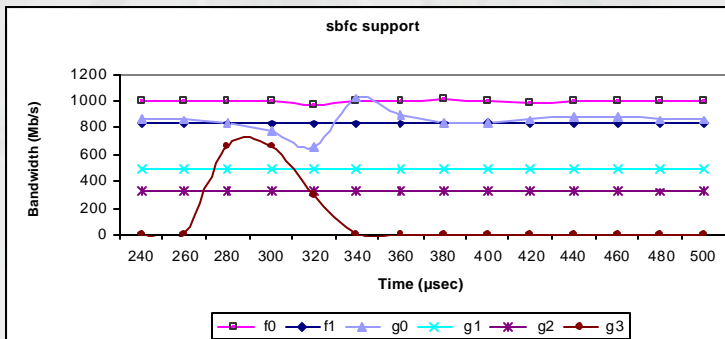
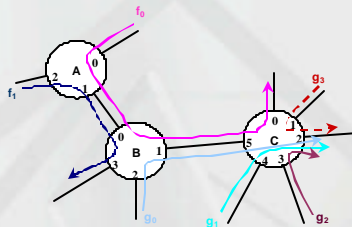
SBFC Effectiveness

Without SBFC



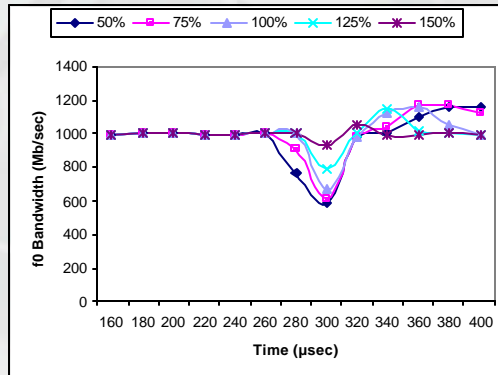
SBFC Effectiveness

With SBFC



Impact on Buffer Size

- **Effectiveness of a larger buffer**
 - Packets spread across multiple nodes



Outline

- **Problem Statement**
- **PCI Express/Advanced Switching Features**
- **Localized Congestion Management**
- **Results**
- **Closing Remarks**



Closing Remarks

- **SBFC - an effective mechanism for handling transient congestion**
- **For handling persistent congestion i.e. as a replacement for an end-to-end (ECN) solution**
 - SBFC applicable only for 1- and 2-stage switching
 - SBFC DLLPs cannot carry information spanning multiple switches
 - For large scale switch fabrics, SBFC complements ECN
- **Future work**
 - Focus on an integrated SBFC-ECN model

