Evaluation of Congestion Detection Mechanisms for InfiniBand Switches

Jose Renato Santos, Yoshio Turner, John Janakiraman

HP Labs

Outline

- **Introduction**
  - **InfiniBand Characteristics**
- **The problem: Congestion Spreading**
- **Congestion Control Approach**
- **Congestion Detection Mechanisms and Simulation Results**
  - **Naive**
  - **Input-Triggered**
  - **Input-Output-Triggered**
- **Conclusion**

InfiniBand

- **Industry Standard for System Area Network**
- **High Performance Server Interconnect**
  - **High Bandwidth: 2Gb/s(1x) to 24Gb/s(12x)**
  - **Low latency: Cut through switching**
    - tens of nanoseconds - switch forwarding delay (no traffic)
- **Current Version 1.0 : Oct 2000**
  - **Does not address congestion control**
  - **Congestion Management Working Group**
    - Defining Congestion Control mechanisms

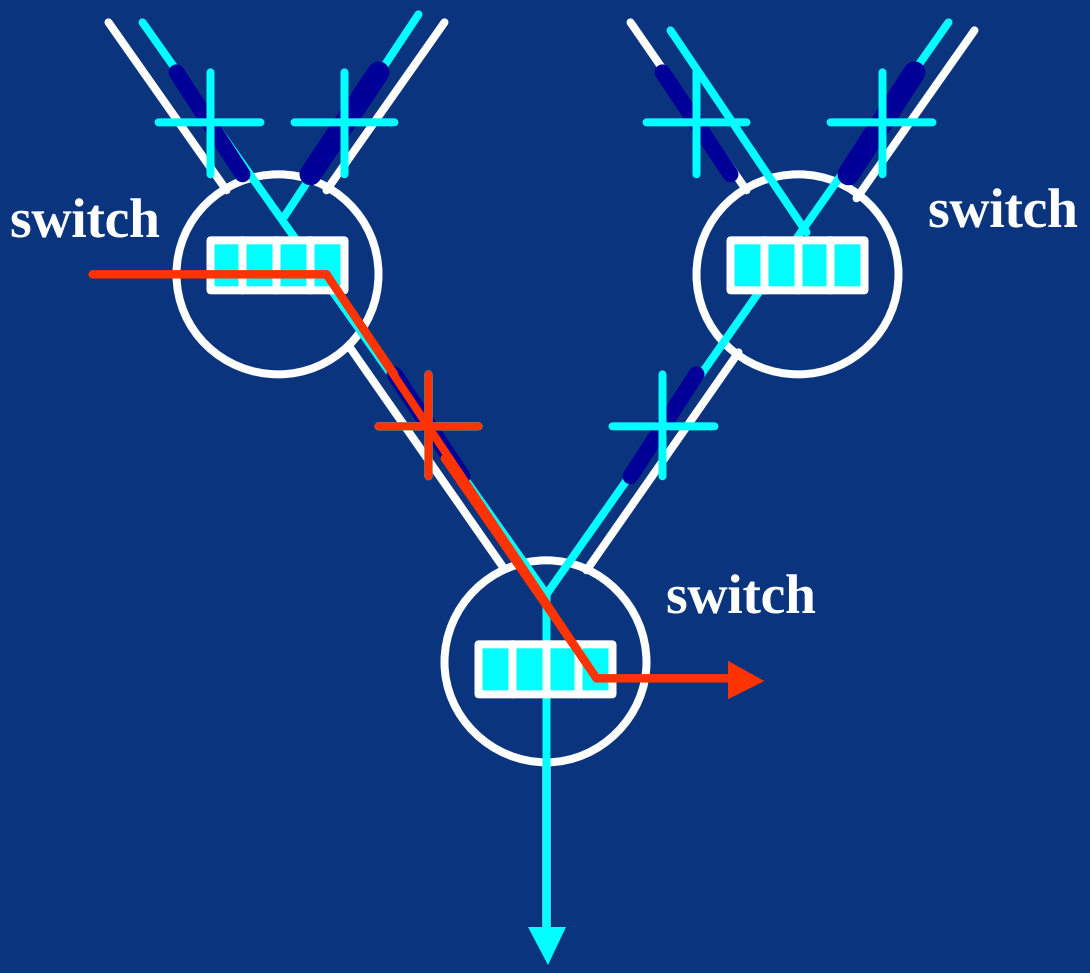
Can't we just adopt TCP traditional congestion control?

- **NO: InfiniBand has unique characteristics that requires a different solution:**
  - **No packet dropping**
  - **Low network latency**
  - **Low buffer capacity at switches**
  - **Switch buffers at input ports**
- **Therefore:**
  - **Need network support for detecting congestion**
  - **Simple Logic for Hardware implementation**
  - **TCP window mechanism inadequate (narrow operational range)**
  - **Alternative congestion detection mechanisms**

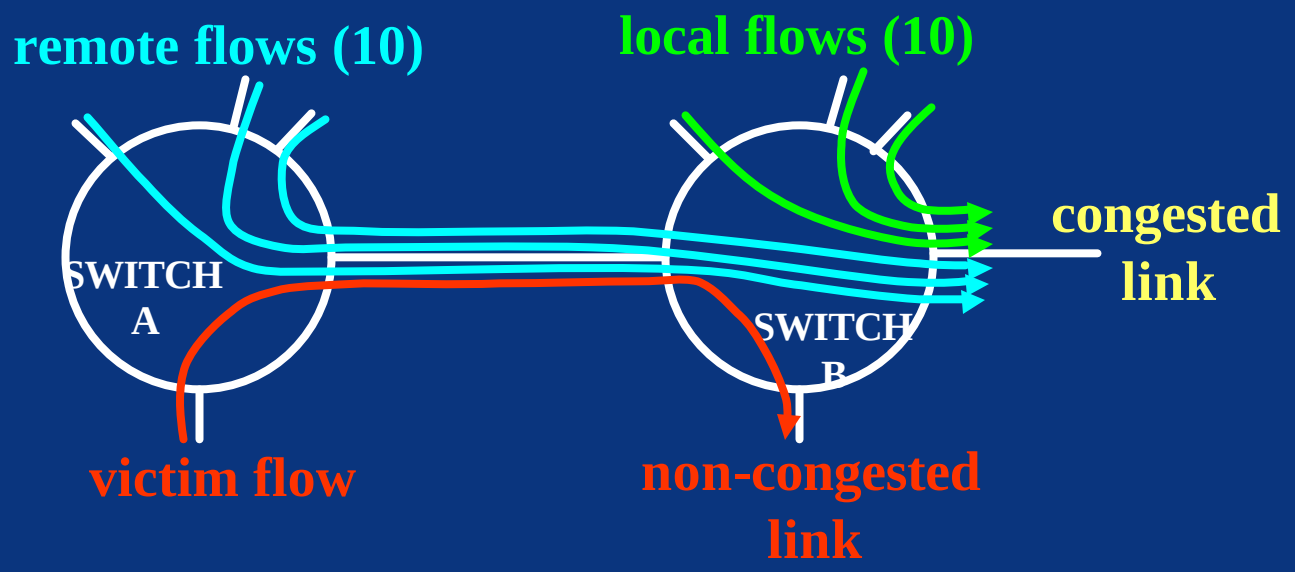
What is the Congestion problem ?

since packets are not dropped

Problem: Congestion Spreading

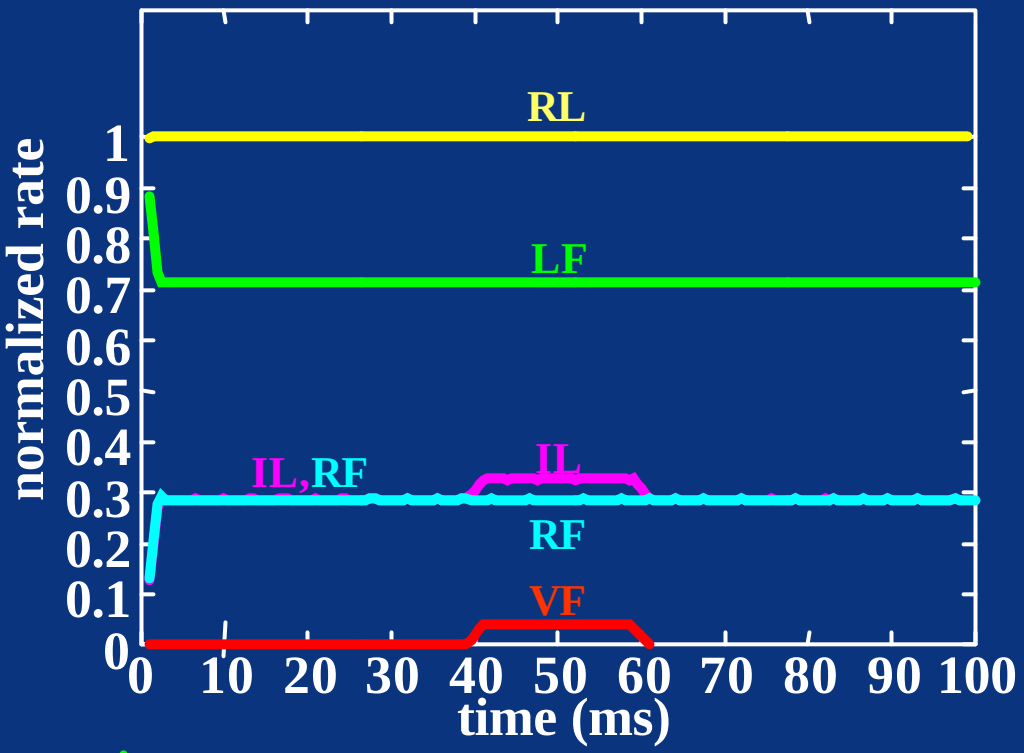


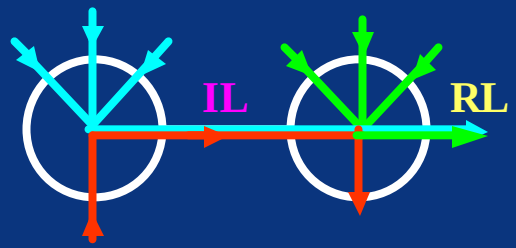
Simulation Scenario



Link BW: 8 Gb/s (4x link)
Packet size: 2 KB
Buffer Size: 4 packets/port (8 KB)
Buffer Org.: Input port

Simulation Results: Congestion Spreading

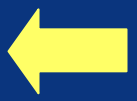




root link (RL)
inter-switch link (IL)
local flows (LF)
remote flows (RF)
victim flow (VF)

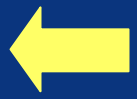
Our approach to Congestion Control

- **Explicit Congestion Notification (ECN)**
  - **Switch detect congestion**
  - **Set single bit ECN field in packet header**
  - **Destination copy packet ECN field in ACK packet**



This Paper

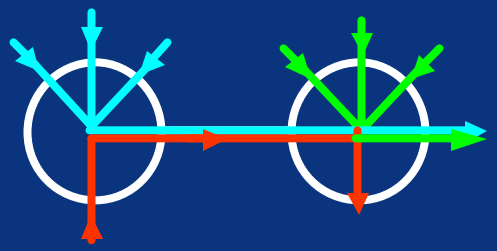
- **Source adjust packet injection according to network feedback encoded in ACK ECN field**
  - **Hybrid source response mechanism:**
    - Combines window and explicit rate control
  - **New Alternative source response functions more efficient than AIMD**



Infocom 2003

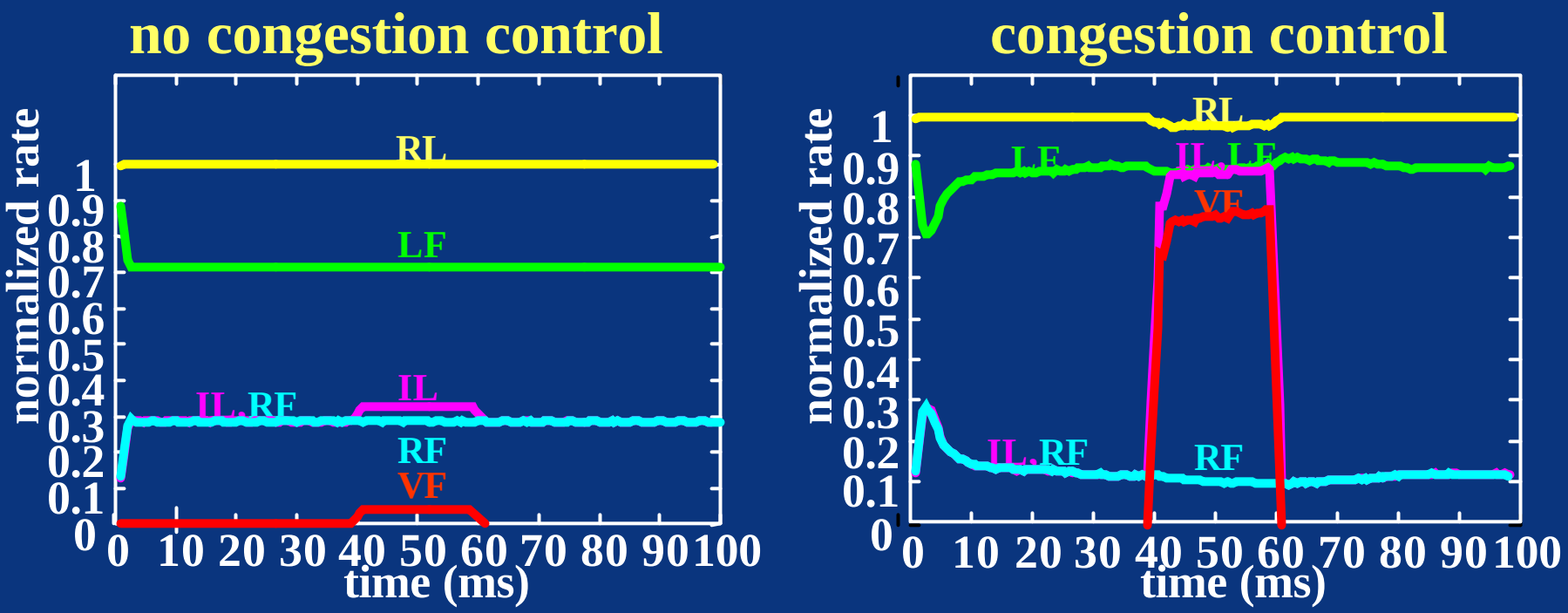
Simulation Results

- **Same scenario used to show congestion spreading:**

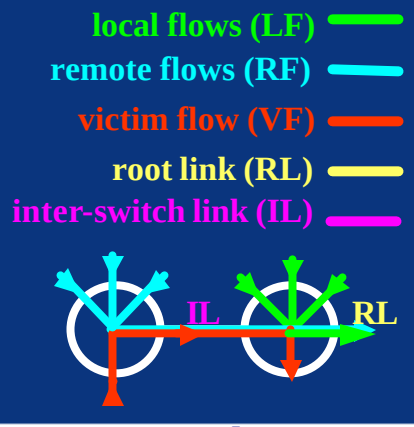


- **It uses a source response function that adjust inter-packet-delay (rate) combined with a fixed window of 1 packet**

Marking Packets in Full Input Buffers (traditional approach)



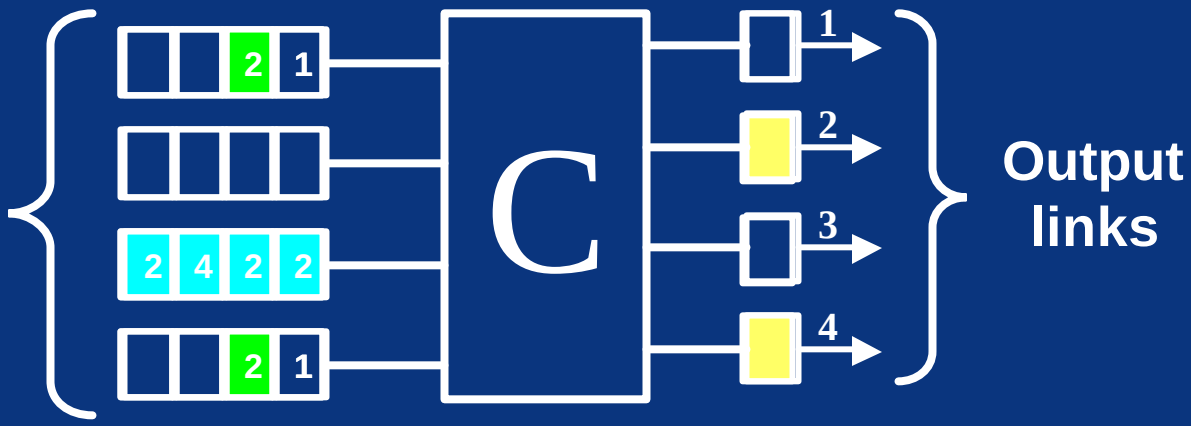
- Effectively avoiding congestion
- Unfairness (remote vs. local flows)



Input-triggered packet marking

- **Goal: Improve fairness**
  - **Mark all packets using congested link**
  - **Not only packets in full buffer**

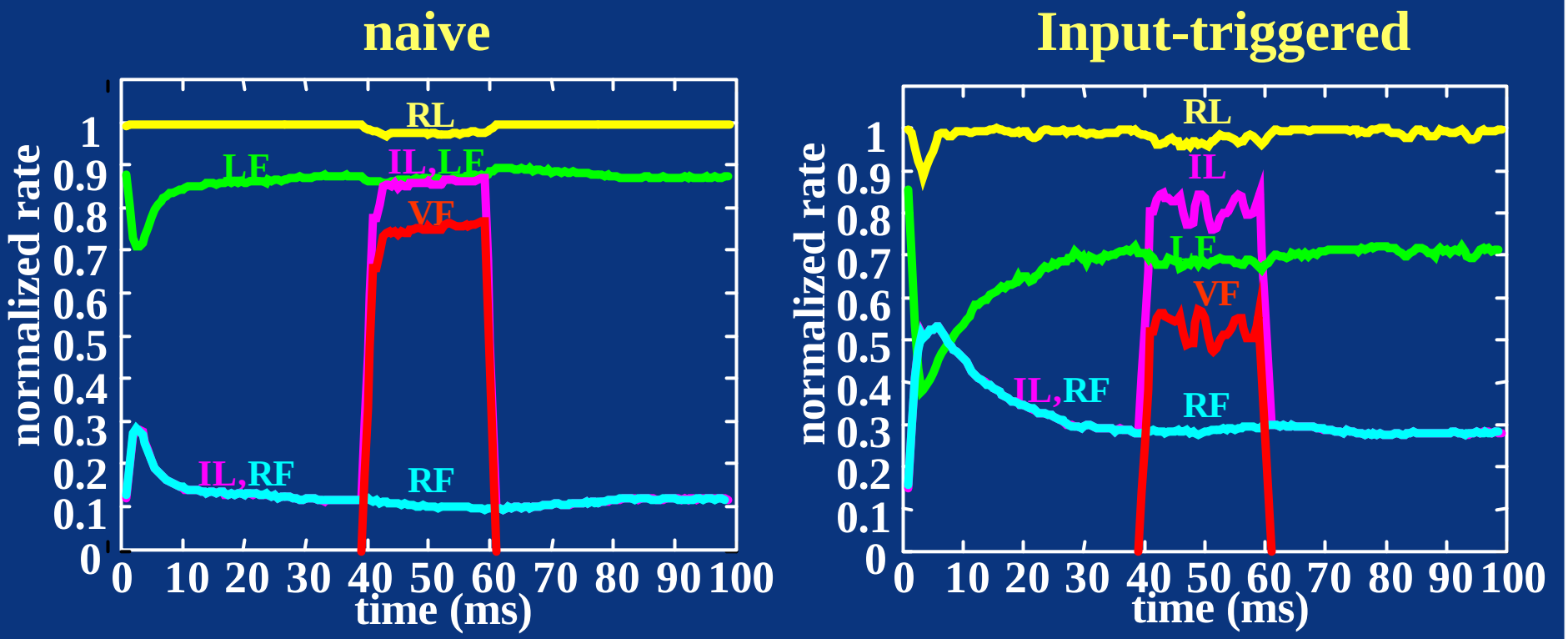
  - Marking triggered by a full input buffer
  - Mark all packets in input buffer (propagating packets)
  - Identify root (congested) links:
    - Destination of packets at full buffer
  - Mark any packet destined to root links (generating packets)



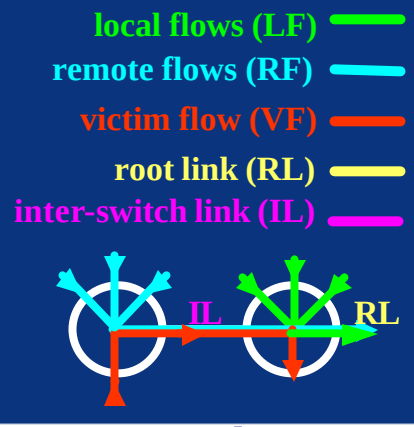
Efficient implementation

- **Use counters to avoid expensive scan of all switch packets (when searching for “generating” packets destined to a congested link)**

Input-triggered Packet Marking



- **Fairness Improved (still some unfairness)**
- **Marking still triggered by remote packets (bias marking towards remote packets)**

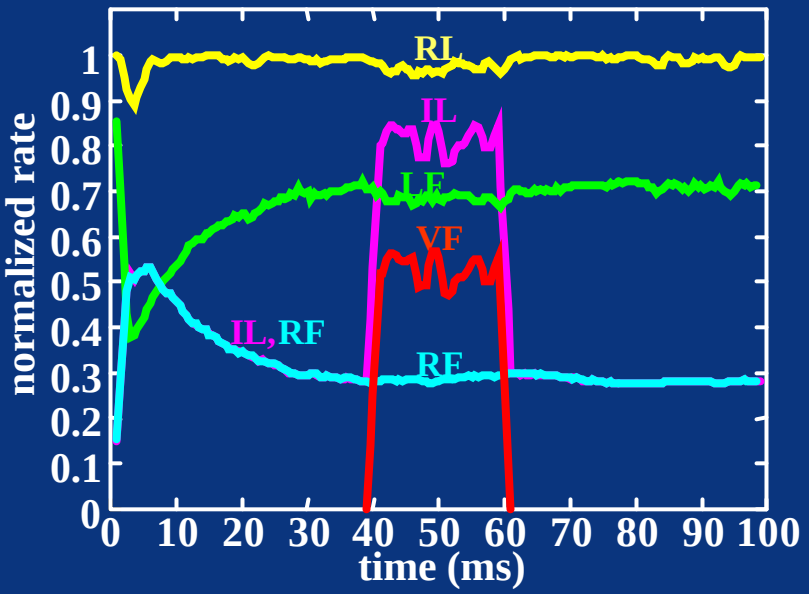


Input-Output-Triggered Packet Marking

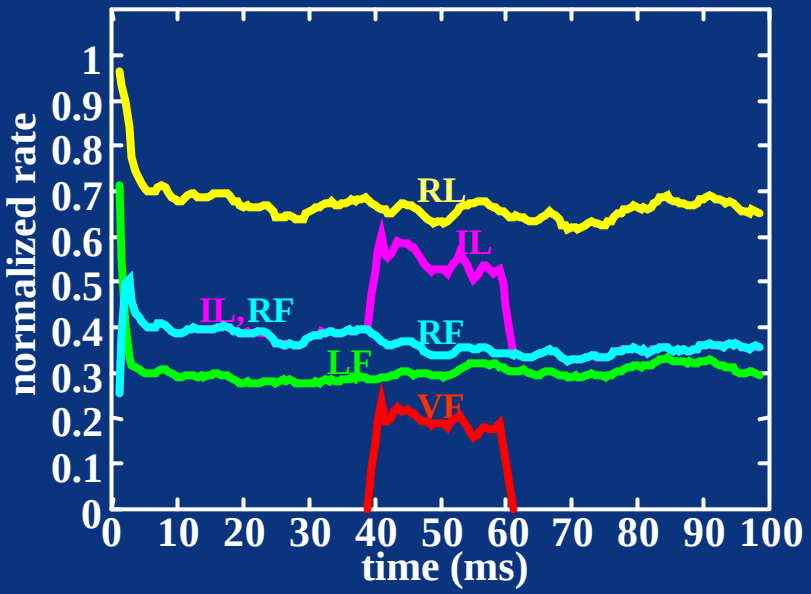
- **Still mark packets when input buffer is full (input triggered)**
  - **To avoid link blocking and congestion spreading**
- **Additional output triggered mechanism**
  - **Mark packets when total number of packets destined to an output port exceeds a threshold**

Input-Output-Triggered Packet Marking

Input-Triggered

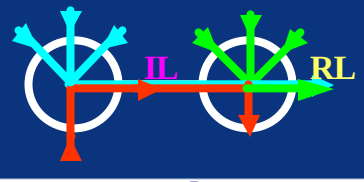


Input-Output-Triggered (threshold: 4 packets)



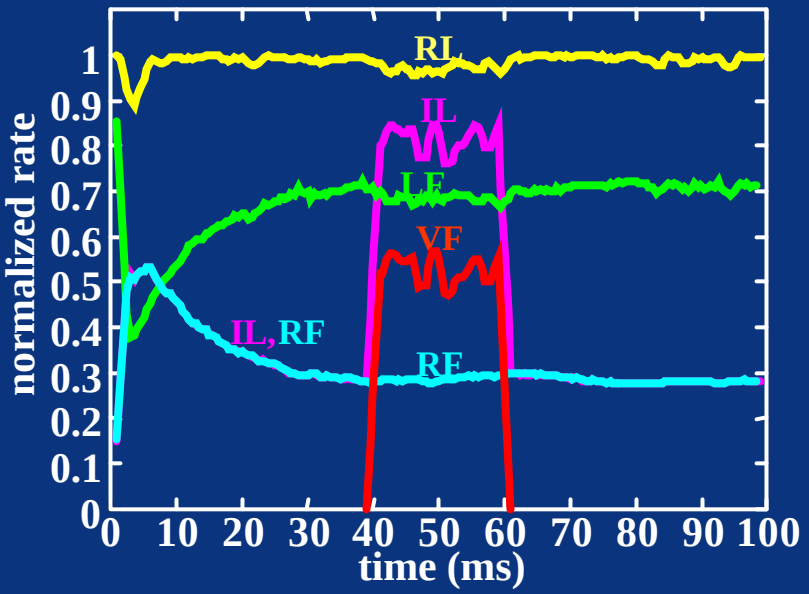
- **Fairness Improved**
- **Under-utilization (aggressive marking)**

local flows (LF)
remote flows (RF)
victim flow (VF)
root link (RL)
inter-switch link (IL)

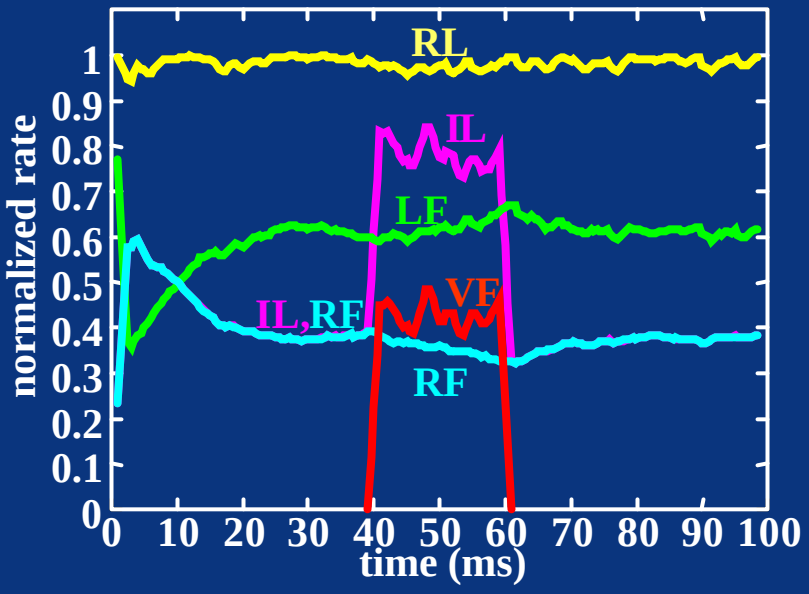


Input-Output-Triggered Packet Marking

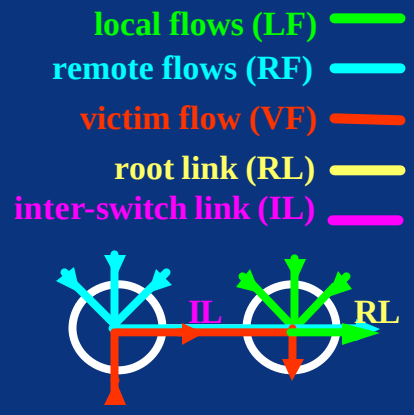
Input-Triggered



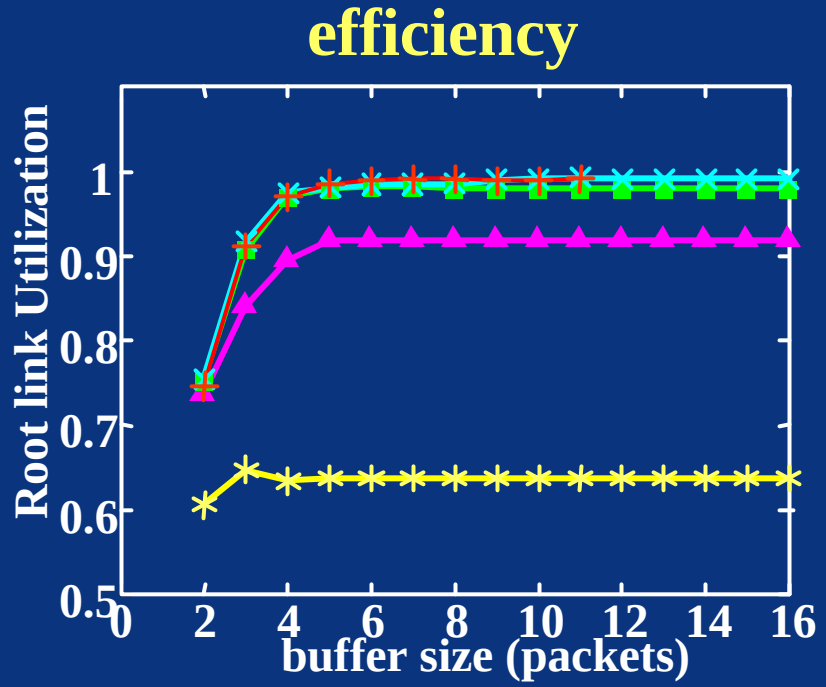
Output-Triggered (threshold: 8 packets)

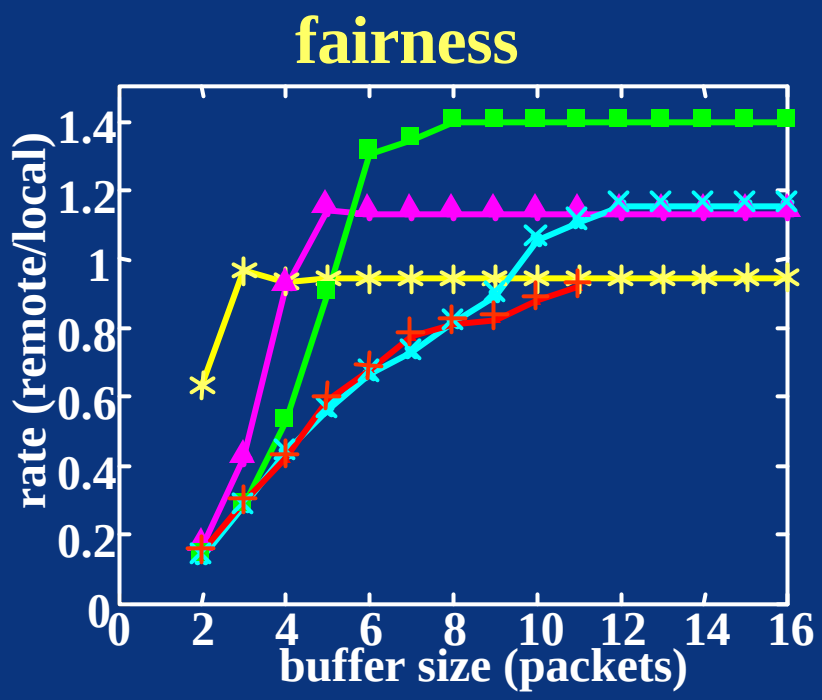


- **High Bandwidth Utilization**
- **Better fairness than input-triggered**



Input-Output-Triggered Packet Marking





Threshold = 4

Threshold = 6

Threshold = 8

Threshold = 16

No output marking

- **Opposing effects**
  - **Input-triggering (bias against remote packets)**
  - **Output-triggering (bias against local packets)**

Conclusion

- **Proposed Congestion Control Mechanism for System Area Networks based on ECN at switches and rate control at end nodes**

- **Proposed and evaluated mechanisms for detecting congestion and marking packets at switches**
  - **Simple mechanisms**
    - **for hardware implementation**
  - **Input-triggered mechanism improves fairness over a naïve full buffer marking scheme**
  - **Input-output-triggered mechanism can improve fairness further**


invent