IEEE 802 Tutorial:
Congestion Notification

17 July 2006

San Diego, CA

Agenda

- Overview
  - Presenter: Pat Thaler; Broadcom
    Chair IEEE 802.1 Congestion management subgroup
- Market
  - Presenter: Manoj Wadekar; Intel
- Example mechanism description and simulation
  - Davide Bergamasco; Cisco
- Document structure
  - Presenter: Norm Finn; Cisco
- Summary and questions
  - Presenter: Pat Thaler

Overview

Pat Thaler

Congestion Notification

- Congestion Notification (CN) provides a means for a bridge to notify a source of congestion allowing the source to reduce the flow rate.
- CN is targeted at networks with low bandwidth delay products: e.g. data center and backplane networks
- Benefits: avoid frame loss; reduce latency; improve performance
- Amendment to IEEE Std 802.1Q

PAR scope*

- Specify protocols, procedures and managed objects for Congestion management of
  - long-lived data flows
  - In network domains of limited bandwidth delay product
  - Bridges signal congestion to end stations
  - VLAN tag priority value segregates congestion controlled traffic
  - Allows simultaneous support of congestion controlled and non-controlled domains

PAR scope, purpose and need are summarized on these slides. For full text see backup slides

PAR purpose

- Data center network and backplane fabrics that
  - with applications that depend on
    - Lower latency
    - Lower probability of packet loss
  - Allowing these applications to share the network with traditional LAN applications

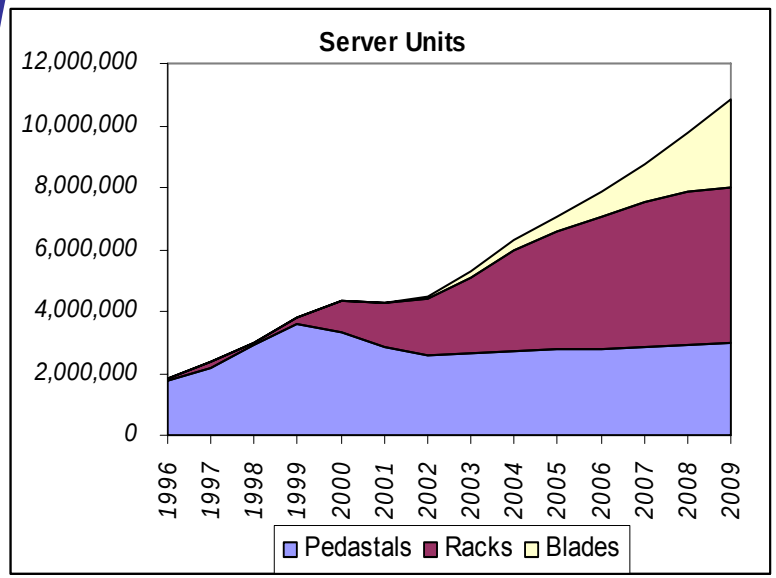
PAR Need

- Opportunity for Ethernet as a consolidated Layer 2 solution in high-speed, short-range networks to support
  - Traffic that uses specialized layer 2 networks today:
    - data centers,
    - backplane fabrics,
    - single and multi-chassis interconnects,
    - computing clusters,
    - storage networks.
  - Network consolidation to provide operational and equipment cost benefits

Market Requirements

Manoj Wadekar

Datacenter : Blade Servers

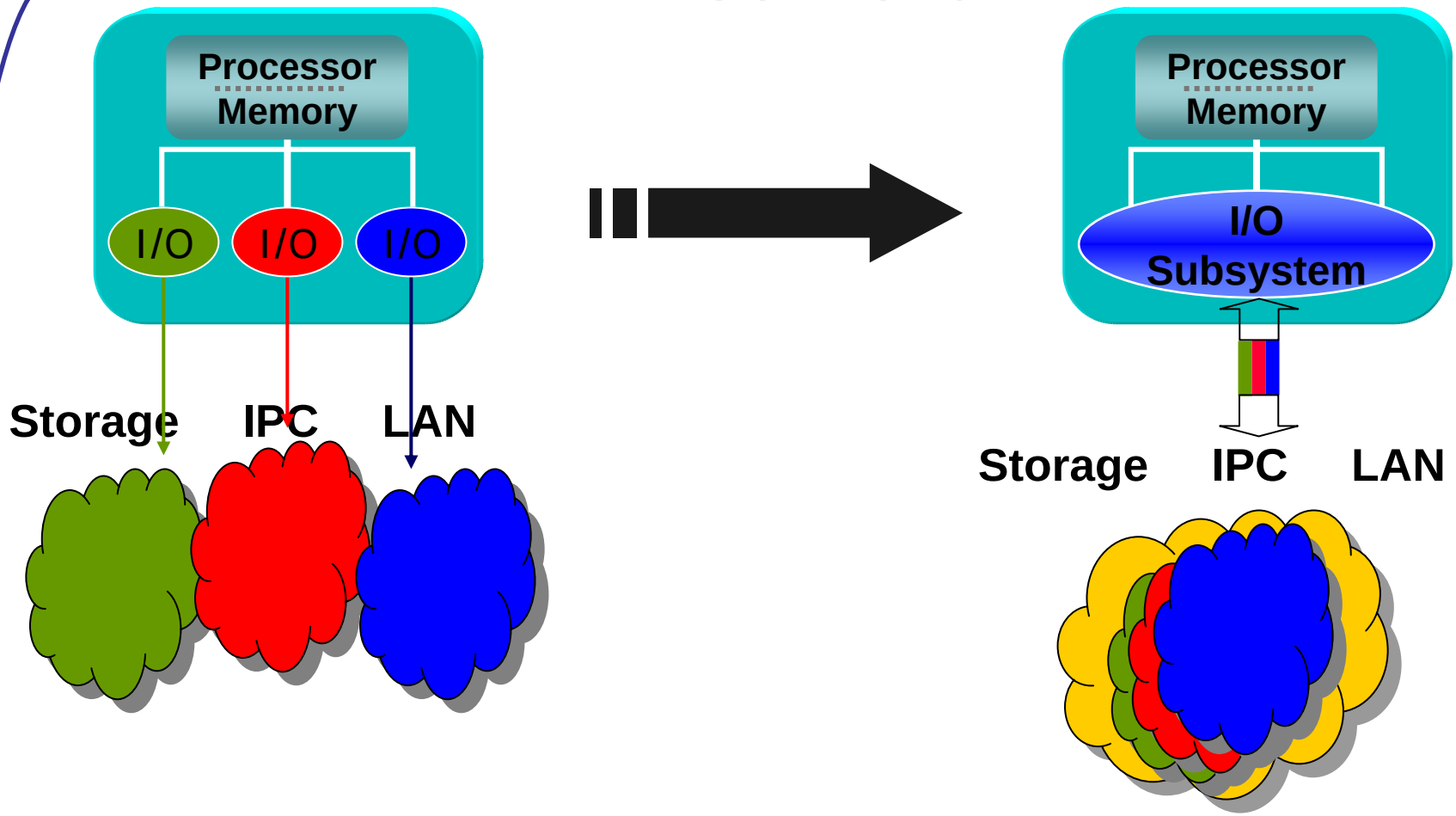


Source: IDC, 2005

- Blade Servers have three types of network traffic:
  - LAN: Ethernet
  - SAN: Fiber Channel
  - IPC: Myrinet, Quadrics, Infiniband etc.
- Each traffic type may require dedicated interconnects:
  - Switches, IO Cards, cables and management
  - Different management tools
- Significant future challenges:
  - Cost, Power, Thermal etc.
  - Overall, high TCO

Scaling Multiple IO Fabrics within the blade system is a challenge!

I/O Consolidation in Blade Servers



I/O Consolidation reduces Capital Expense and Operational Expense

Storage Components Market

- iSCSI adoption has been slow despite being more cost effective
- FC continues to be the dominant SAN technology
- F500 IT concerns include
  - Security
  - Performance -- Ethernet behaves poorly in congested environments, packet drops significant, adversely affects storage traffic

Improving Ethernet congestion management can accelerate iSCSI adoption – addresses IT perception & reality

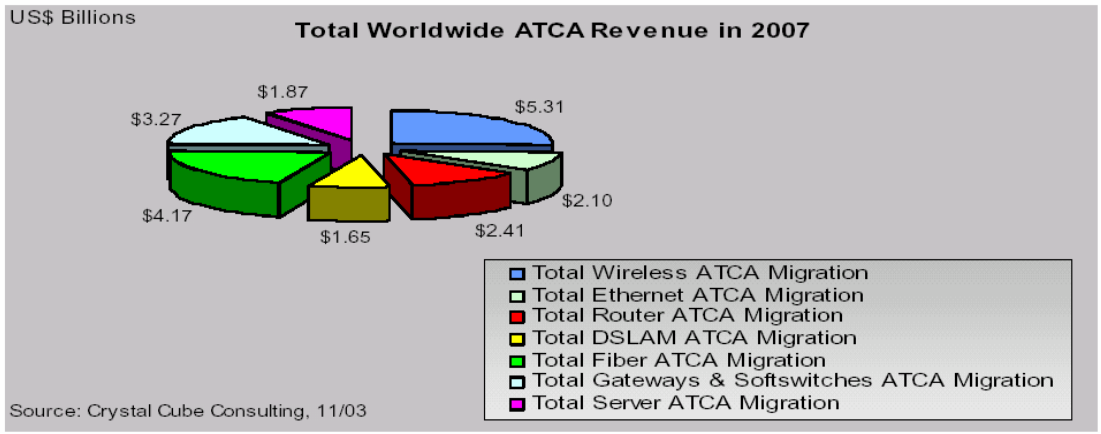
Ethernet Opportunity for Clustering and IPC

- Highest growth in the "Technical Capacity" Servers ~ 20% of High Performance Computing (HPC) market by 2007
  - Clusters built using low cost servers connected by a high performance, low latency fabric
- Users like the cost structure and availability of Ethernet
  - However latency and congestion management are key issues
- Myrinet and Quadrics based fabrics are being deployed to address this need
- Infiniband emerging as fabric of choice for clustering

Addressing latency and packet loss opens up the cluster market for Ethernet

Telco Backplane Opportunity for Ethernet

Figure 15. Worldwide ATCA Projection of Revenue in 2007 by Market Segment



- Ethernet is leading backplane choice for Telco
- However, Layer 2 enhancements are required to ensure no drop in the backplane
- Improving Ethernet congestion management capabilities can accelerate its adoption in the expanding ATCA (Advanced Telecom Computing Architecture) market

Datacenter Requirements

- Address IT perceptions:
  - "Ethernet not adequate for low latency apps"
  - "Ethernet frame loss is inefficient for storage"
- 802.3x does not help
  - Reduces throughput
  - Congestion spreading
  - Increases latency jitter
- Improve Ethernet Congestion Management capabilities that will:
  - Reduce frame loss significantly
  - Reduce end-to-end latency and latency jitter
  - Achieve above without compromising throughput

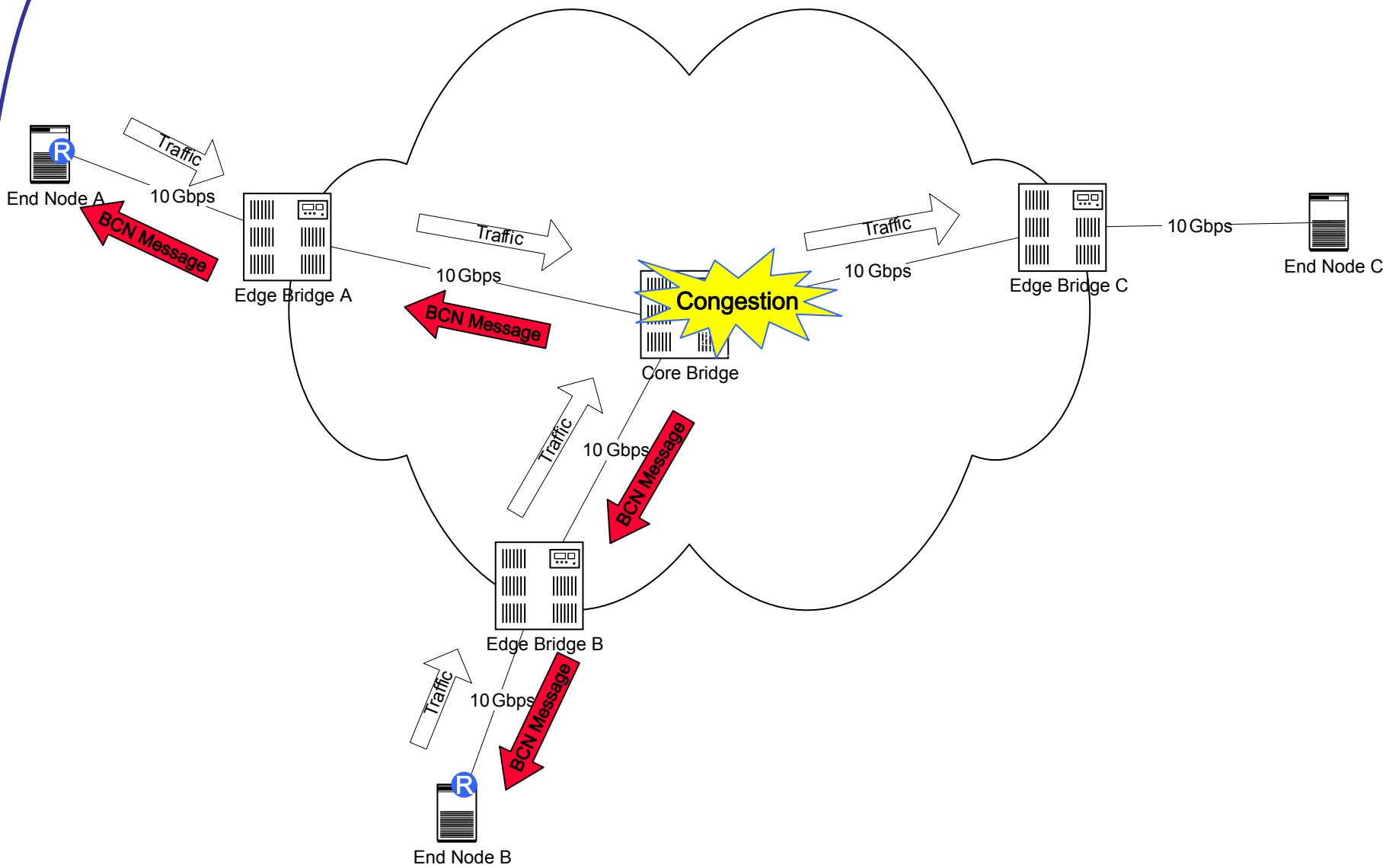
Backward Congestion Notification
An Example of CM Mechanism

Davide Bergamasco (davide@cisco.com)

What is BCN?

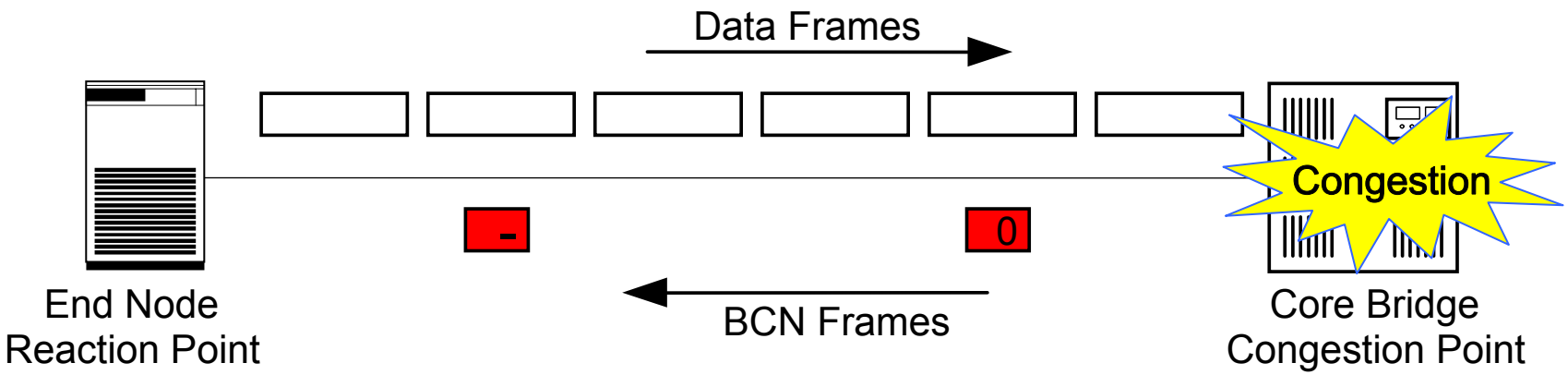
- BCN is a Layer 2 Congestion Management Mechanism
- Principles
  - Push congestion from the core towards the edge of the network
  - Use rate-limiters at the edge to “shape” flows causing congestion
  - Control injection rate based on feedback coming from congestion points
- Inspired by TCP
  - AIMD rate control
    - TCP window increases linearly in absence of congestion
    - Decreases exponentially (gets halved) at every congestion indication (either implicit or explicit)
  - Self-Clocking Control loop (acknowledgements)

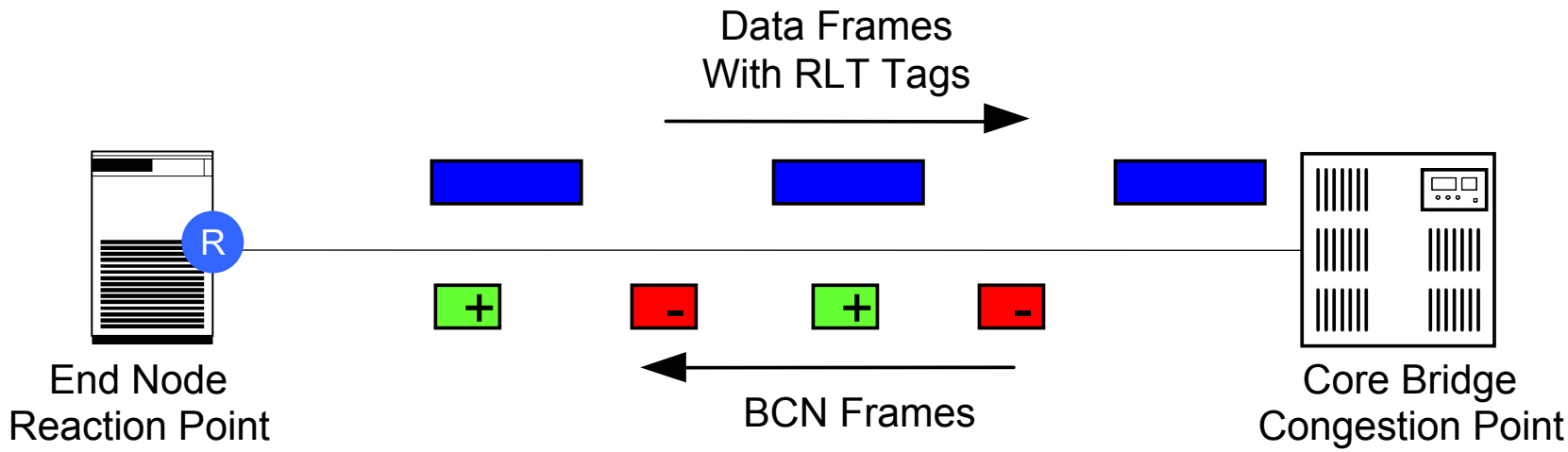
BCN Concepts (1)



BCN Concepts (2)

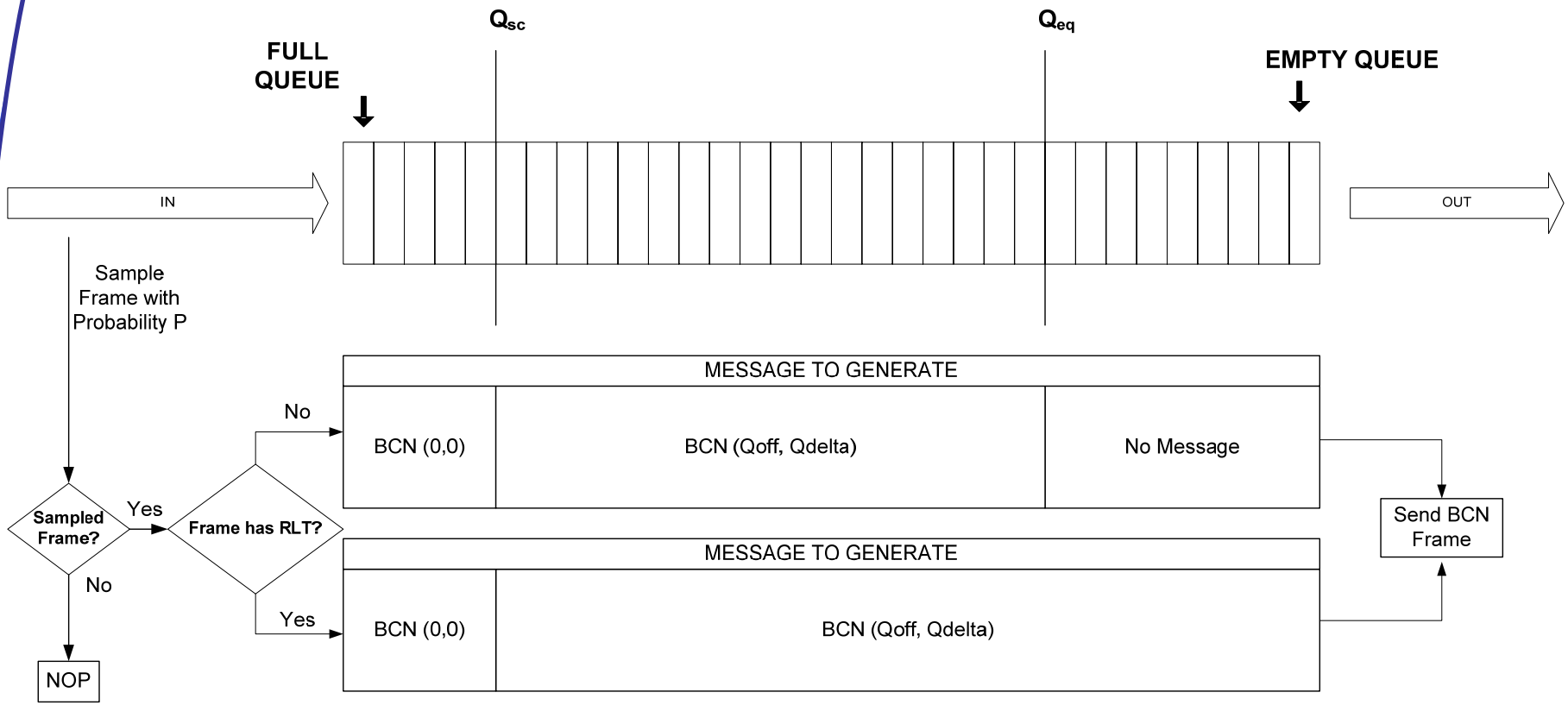
- **Signaling (w/o animation)**





BCN Concepts (3)

Detection



Qoff = Qlen - Qeq [-Qeq, +Qeq]

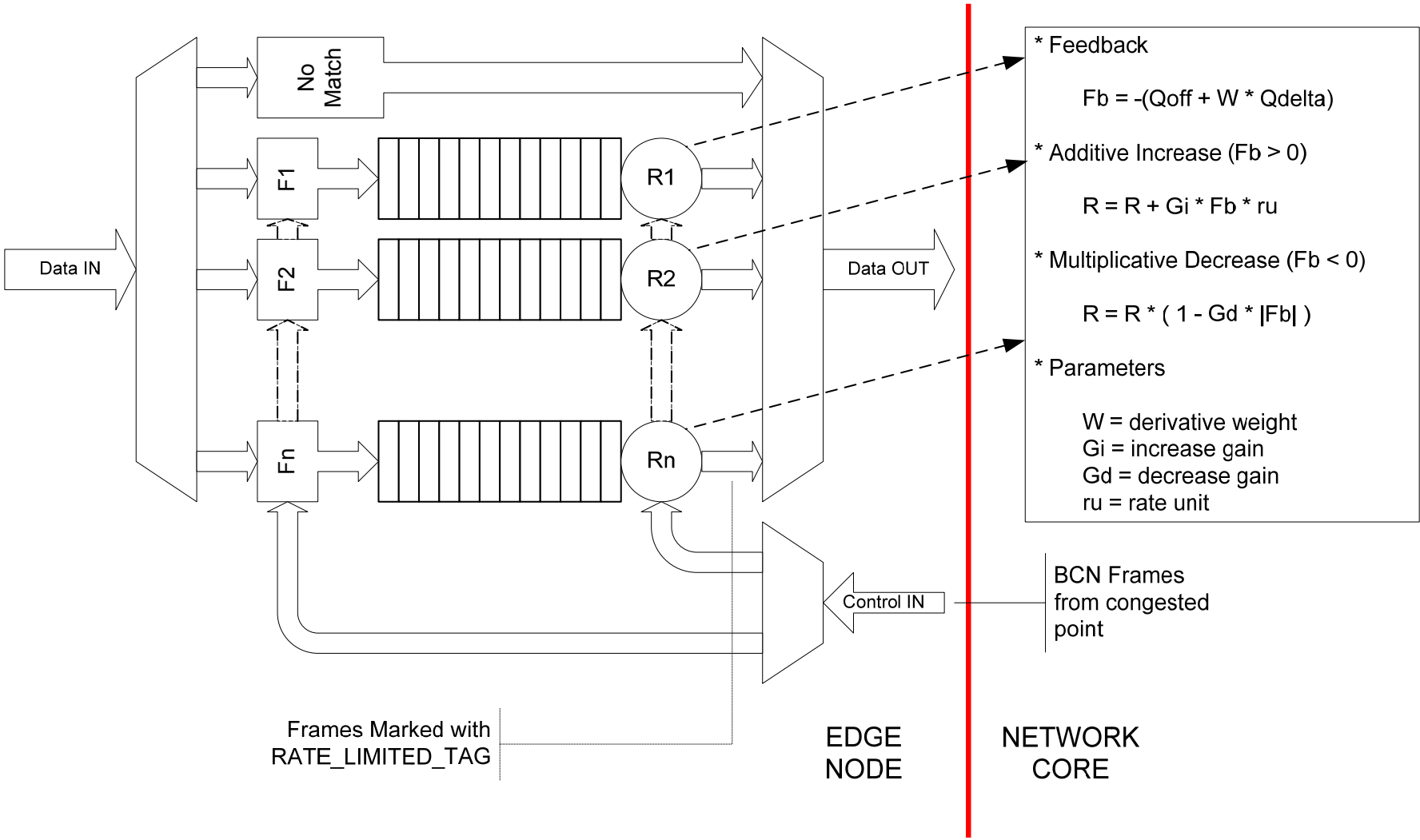
Qdelta = Qlen - Qold [-2Qeq, +2Qeq]

BCN Concepts (4)

- **Detection**
  - Performed by Congestion Points located in Bridges
    - Usually output [port, class] queues
  - Very simple
    - Two thresholds
    - Minimal state
    - Machinery to generate BCN messages
    - Parser to identify RLT tagged frames

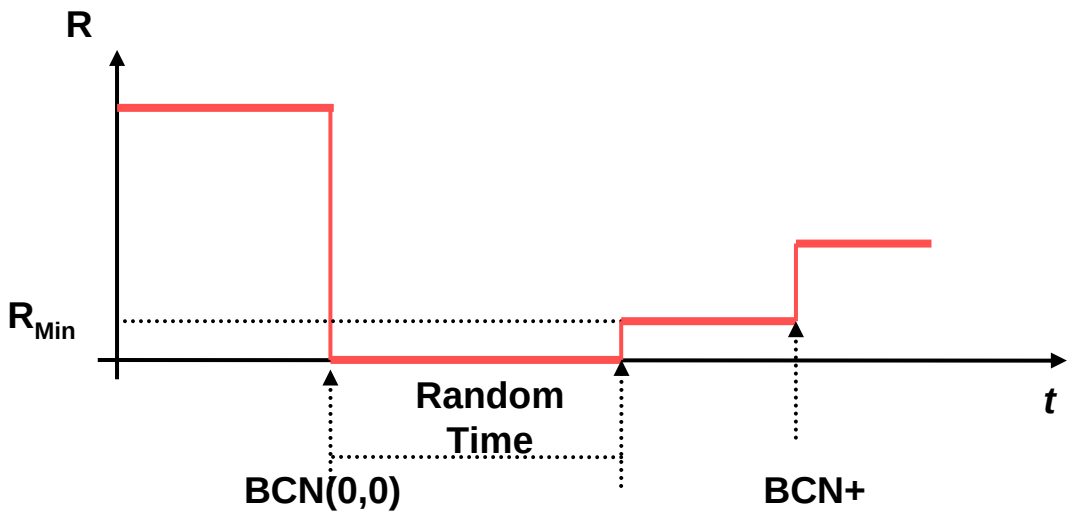
BCN Concepts (5)

- **Reaction**



BCN Concepts (6)

- **Reaction**
  - BCN(0,0): Special feedback message
  - Current rate *R* is set to 0
  - Random timer $[0, T_{Max}]$: when timer expires Current rate *R* is set to R_{Min}



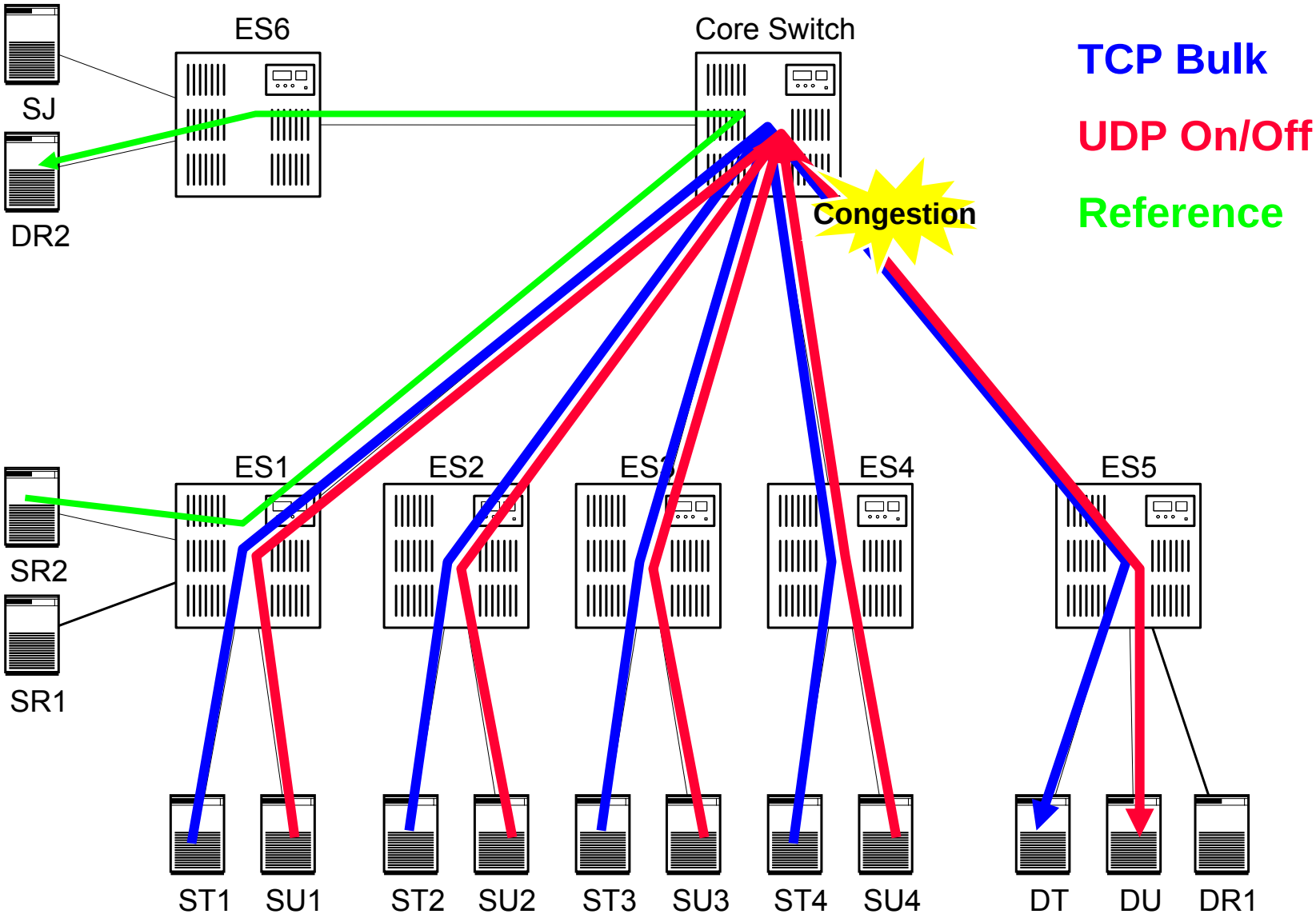
BCN Concepts (7)

- **Reaction**
  - Performed by Reaction Points located in End Nodes
  - More complex
    - Traffic filters
    - Queues
    - Rate limiters
    - More state
  - Arbitrary granularity
    - Example: SA/DA/PRI, DA/PRI, PRI, Entire link
  - Automatic fall-back
    - When finer rate limiters are exhausted, aggregate flows in coarser rate limiters: Eg. SA/DA/PRI → DA/PRI

Validation

- BCN has been validated
  - Analytically
    - http://www.ieee802.org/1/files/public/docs2005/new-bergamasco-bcn-september-interim-rev-final-0905.ppt
  - By Simulation
    - http://www.ieee802.org/1/files/public/docs2005/new-bergamasco-backward-congestion-notification-0505.pdf
    - http://www.ieee802.org/1/files/public/docs2005/new-bergamasco-bcn-july-plenary-0705.ppt
    - http://www.ieee802.org/1/files/public/docs2006/new-cm-jain-bcn-v2-simulation_0306.pdf

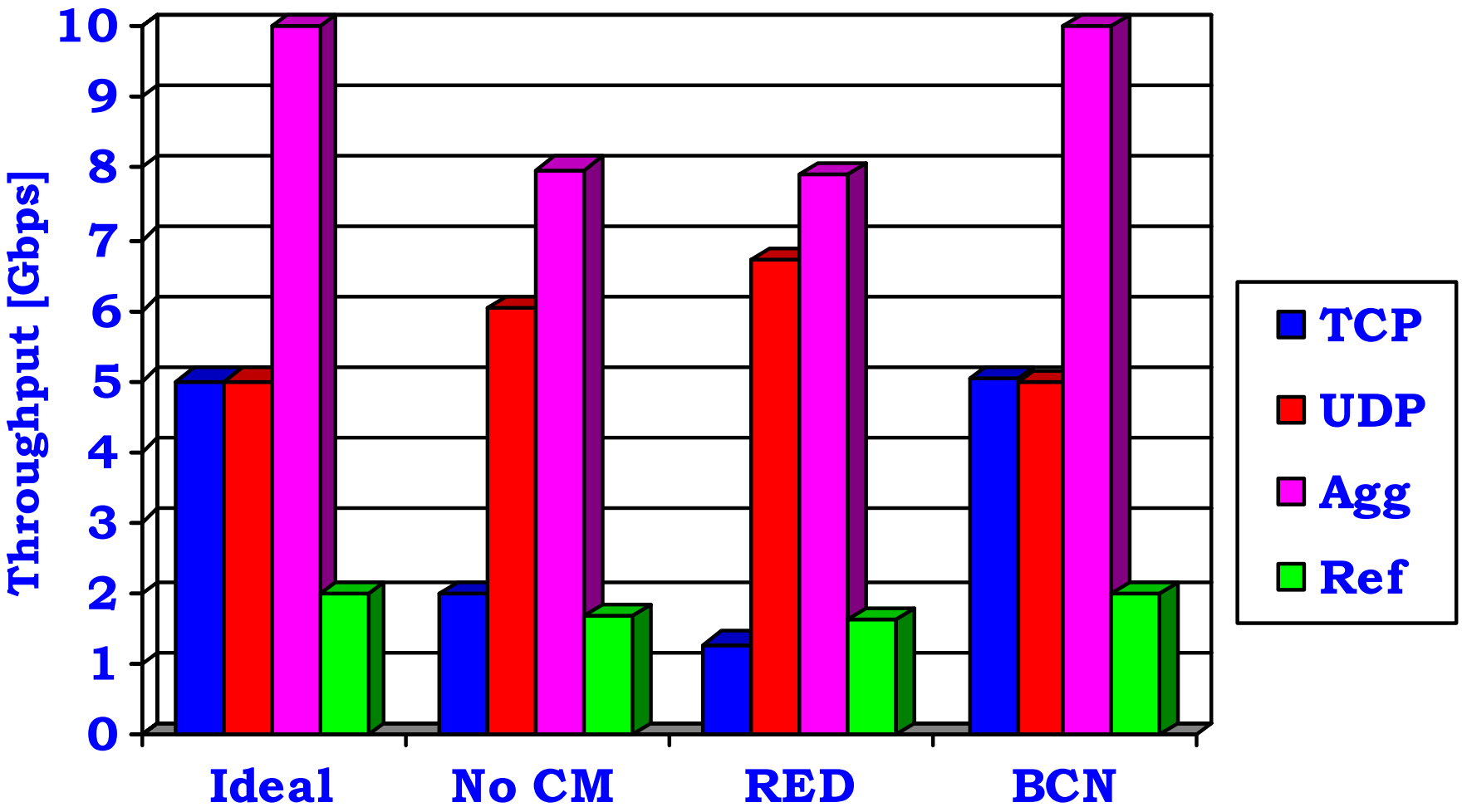
Simulation (1)



Simulation (2)

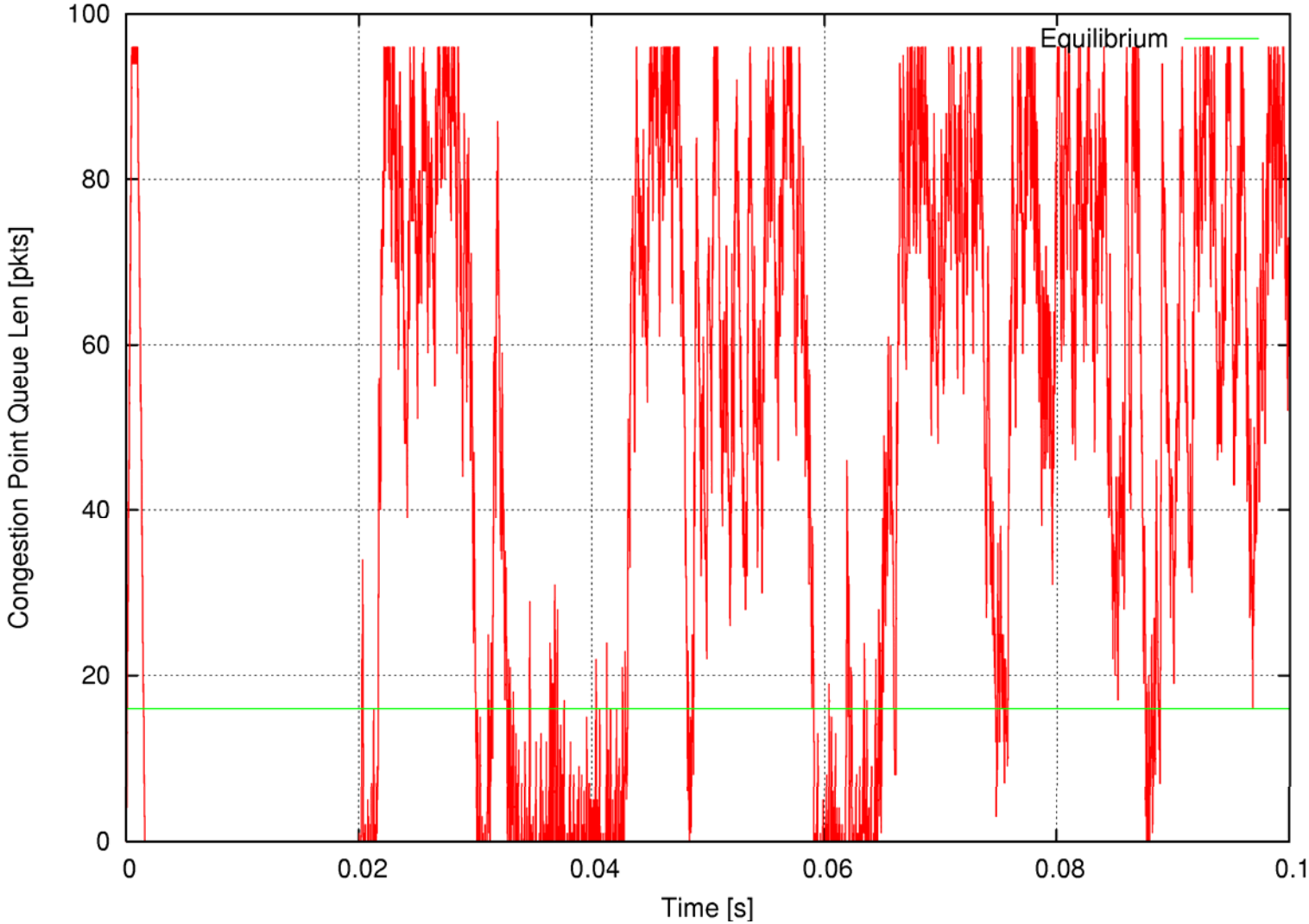
- Short Range, High-Speed Datacenter-like Network
  - Link Capacity = 10 Gbps
  - Buffer Size = 150 KB
  - Switch latency = 1 μs
  - Link Length = 100 m (.5 μ s propagation delay)
- Control loop
  - Delay ~ 3 μs
  - Parameters
    - W = 2
    - Gi = 16
    - Gd = 1/128
    - Ru = 1 Mbps
- Workload
  - 80% TCP + 20% UDP
    - ST1-ST4: 10 parallel connections transferring 1 MB each (t=0 ms)
    - SU1-SU4: variable length bursts with average offered load of 2 Gbps (t=10 ms)
    - SR2: same as above

Simulation (3)



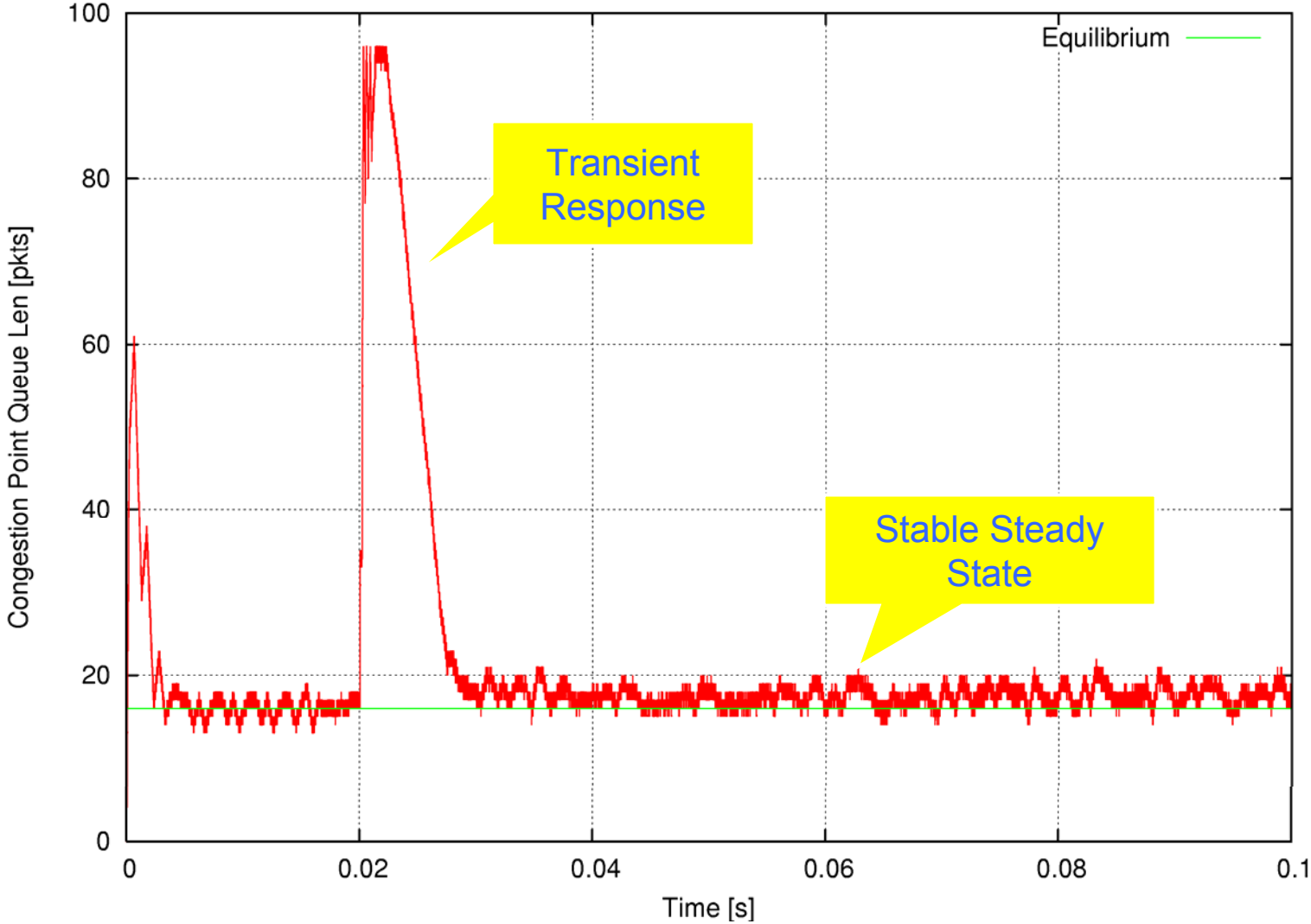
Simulation (4)

- No CM / RED



Simulation (5)

- BCN



Summary

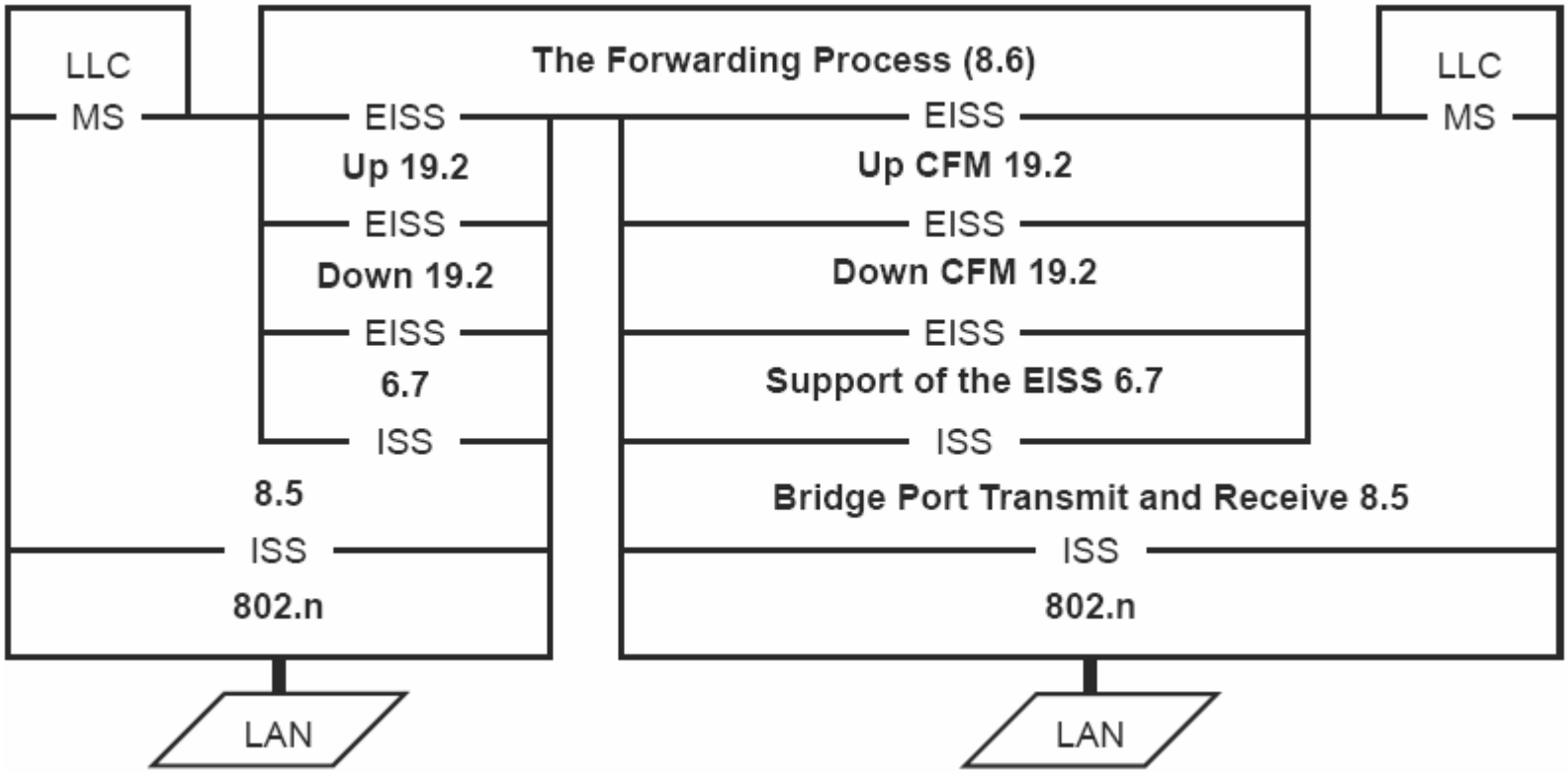
- BCN has a number of advantages …
  - Effectiveness
  - L3/L4 Protocol Agnosticism
  - Fairness
  - Good protection of TCP flows in mixed TCP and UDP traffic scenarios
  - Simple Detection Algorithm
    - Minimal per-queue state
    - No per-flow state
- … and a some of disadvantages
  - Traffic overhead in reverse direction
  - Ideal behavior requires per-flow queuing
  - Flow duration >> network RTT

Document Structure

Norm Finn

An 802.1Q VLAN-aware Bridge

P802.1ag Connectivity Fault Management has affected the presentation, but not the meaning, of the 802.1Q Bridge “baggy pants”.



IEEE Std. 802.1Q-2005

Subclause 8.6 The Forwarding Process

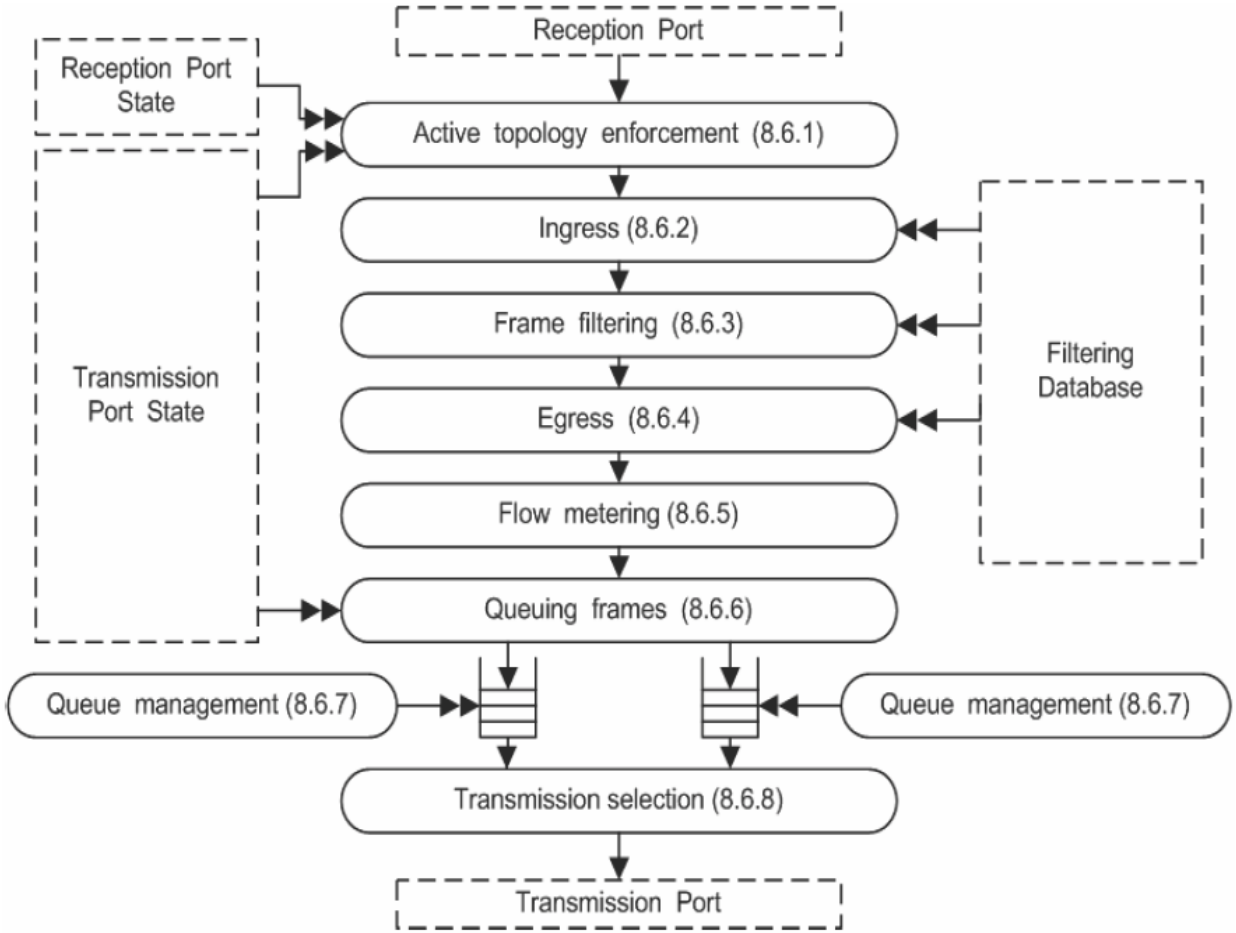


Figure 8-9—Forwarding Process functions

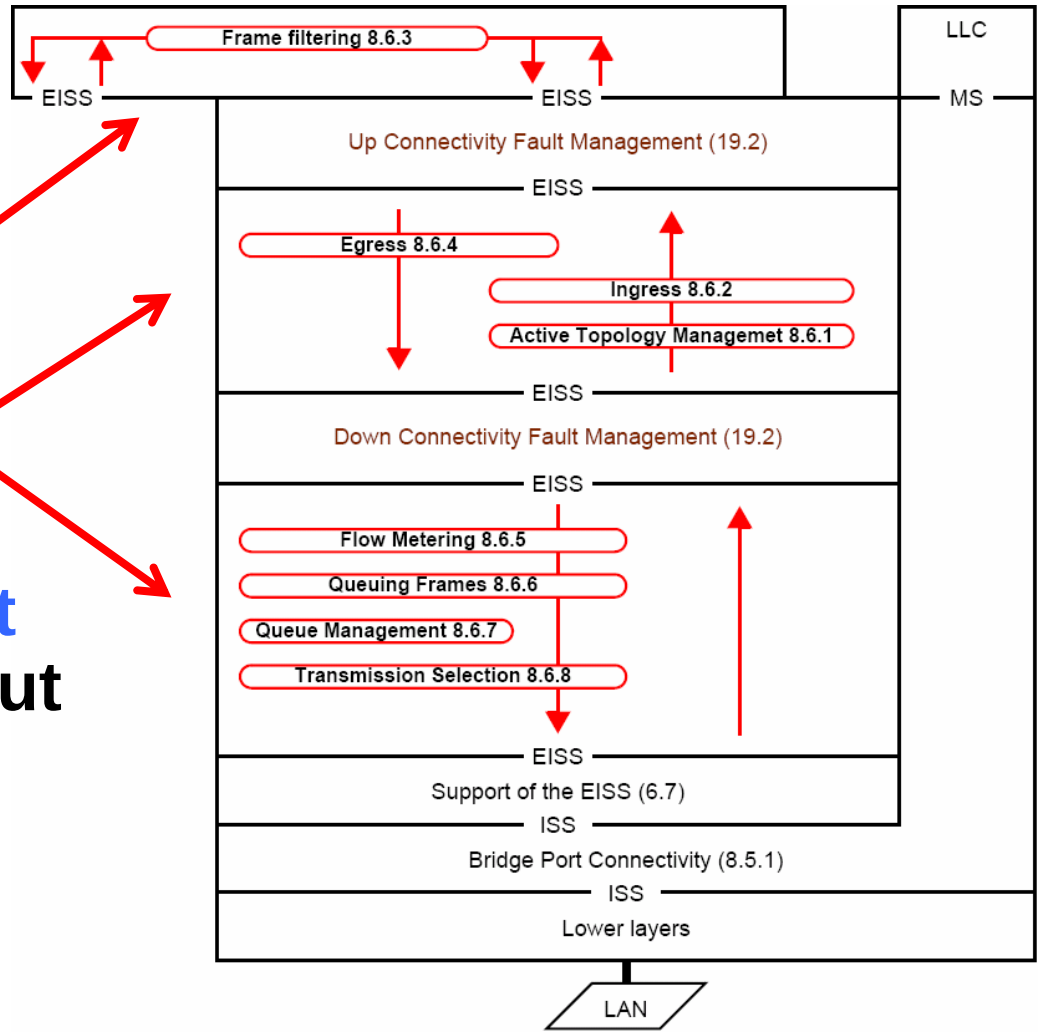
The Forwarding process has been exploded and shifted in Draft 6.1 of P802.1ag CFM.

This does not change any existing functional relationships

This function is per-Bridge.

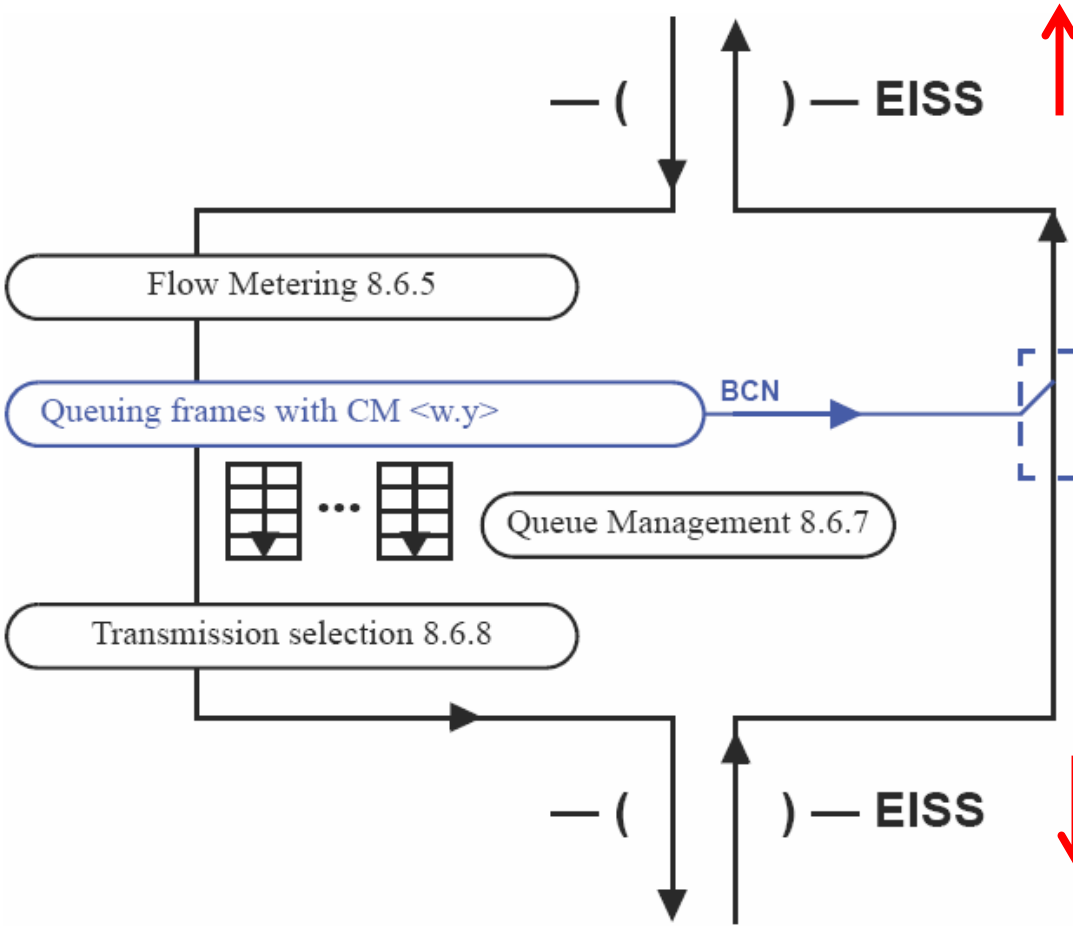
These functions are per-Port.

This breakout is not required for BCN, but it makes both BCN and CFM easier to understand.



Looking at just the Bridge Queuing shim:

Every Port in a CM-capable Bridge requires:



The data is coming from, and the BCNs going to, the Bridge's forwarding function in this direction.

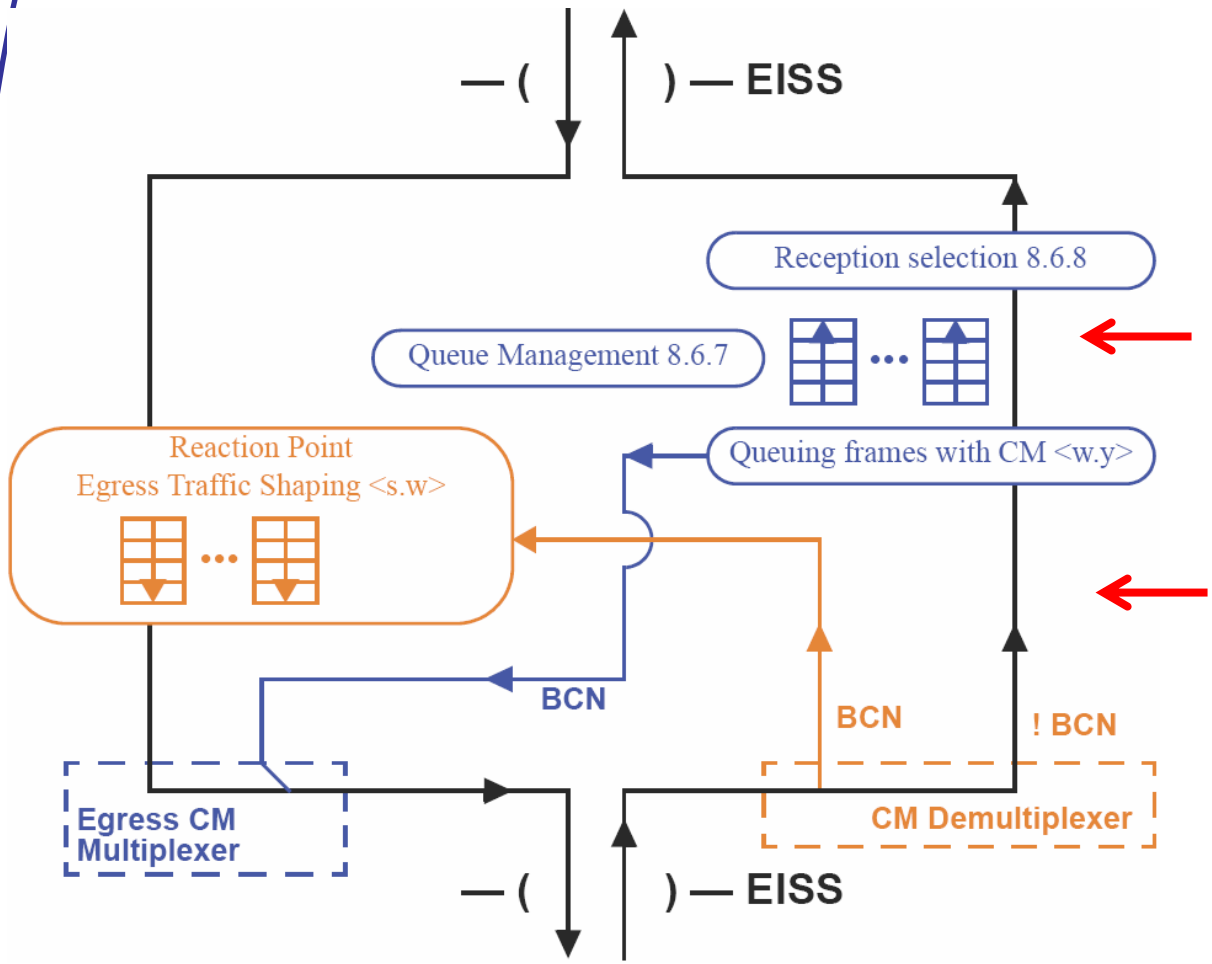
Ingress BCN Multiplexer

New part: BCN generation in blue.

The data is going to the LAN on this side of the port.

Looking at a new shim for the Station:

Every CM-Capable Station requires:



The station's "brain" is in this direction

Station's receiver is a congestion point just like a Bridge Port.
(It's upside down, relative to Bridge.)

The Reaction Point sets up and performs the per-flow queuing.

The LAN is in this direction

Summary

Objectives under consideration

- Independent of upper layer protocol
- Compatible with TCP/IP based protocols
- Support bandwidth delay product of at least 1 Mbit, preferably 5 Mbit
- Coexistence of congestion managed and unmanaged traffic segregated by VLAN tag priority bits
- Full-duplex point-to-point links with a mix of link rates.

Objectives under consideration

- Define messages, congestion point behavior, reaction point behavior and managed objects
- Confine protocol messages to domain of CN capable bridges and end stations
- Consider inclusion of discovery protocol (e.g. LLDP)
- Don’t introduce new bridge transmission selection algorithms or rate controls
- Do not require per flow state or queuing in bridges

Additional references

- Files
  - At http://www.ieee802.org/1/files/public/docs2006
  - PAR
    - new-p802.1au-draft-par-0506-v1.pdf
  - 5 Criteria
    - New-p802.1au-draft-5c-0506-v1.doc
  - First draft of objectives
    - New-cm-thaler-cn-objectives-draft-0506-01.pdf
  - Going forward, CN files will begin “au-”

IEEE 802

Questions?

Background slides

PAR Scope

- This standard specifies protocols, procedures and managed objects that support congestion management of long-lived data flows within network domains of limited bandwidth delay product. This is achieved by enabling bridges to signal congestion information to end stations capable of transmission rate limiting to avoid frame loss. This mechanism enables support for higher layer protocols that are highly loss or latency sensitive. VLAN tag encoded priority values are allocated to segregate frames subject to congestion control, allowing simultaneous support of both congestion controlled and other higher layer protocols. This standard does not specify communication or reception of congestion notification information to or from stations outside the congestion controlled domain or encapsulation of frames from those stations across the domain.

Purpose

- Data center networks and backplane fabrics employ applications that depend on the delivery of data packets with a lower latency and much lower probability of packet loss than is typical of IEEE 802 VLAN bridged networks. This amendment will support the use of a single bridged local area network for these applications as well as traditional LAN applications.

Need and stakeholders

- There is significant customer interest and market opportunity for Ethernet as a consolidated Layer 2 solution in high-speed short-range networks such as data centers, backplane fabrics, single and multi-chassis interconnects, computing clusters, and storage networks. These applications currently use Layer 2 networks that offer very low latency and controlled frame loss due to congestion. Use of a consolidated network will realize operational and equipment cost benefits.
- Developers and users of networking for data center and backplane Ethernet environments including networking IC developers, switch and NIC vendors, and users.