Congestion Notification Mechanisms in 802 networks

Manoj Wadekar

IEEE Interim Meeting
January 12, 2005

Agenda

- Market Potential
- Requirements and Scope
- Congestion Notification mechanisms
- Proposal for L2 mechanism – L2-CI
- Summary

Summary of request

- In order to enable accelerated deployment of Ethernet into emerging limited-topology applications (clustering, backplanes, storage, data centers, etc.), IEEE 802.1 should specify a standard mechanism for MAC Clients to provide congestion information to L2 edge devices, using wadekar_1_0501.pdf as a basis

Congestion Control Elements

- Detection
  - Could be an AQM like RED (Random Early Detection) – Does not need to be specified by IEEE 802
- Notification
  - Need a standard way to notify congestion between L2 devices
    - Request to IEEE 802.1 to consider
- Action
  - Rate control/reduction done by source in response to congestion notification
  - Left to ULPs (L3 and above) e.g. TCP
    - IETF Domain

Market Potential

Market Opportunities for Ethernet

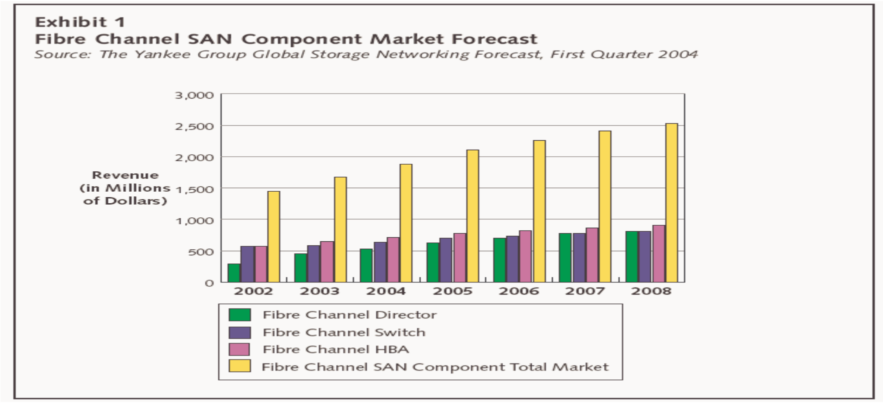


Exhibit 1
Fibre Channel SAN Component Market Forecast
Source: The Yankee Group Global Storage Networking Forecast, First Quarter 2004

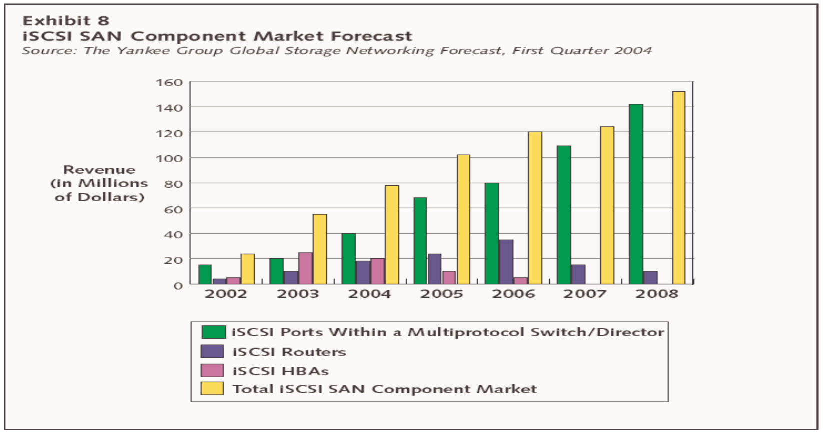
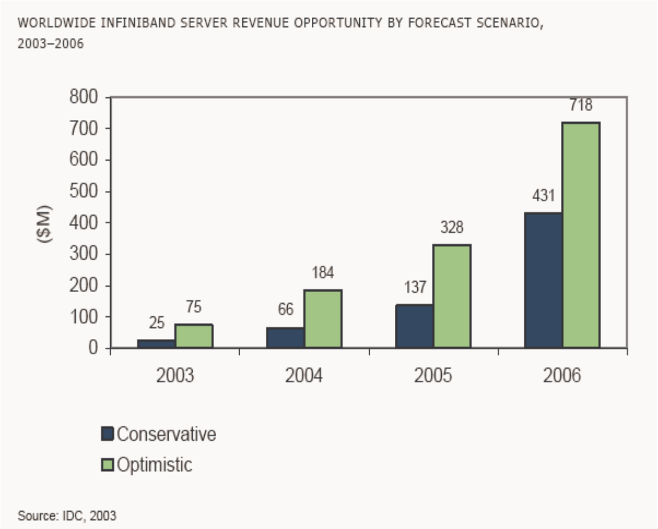


Exhibit 8
iSCSI SAN Component Market Forecast
Source: The Yankee Group Global Storage Networking Forecast, First Quarter 2004



WORLDWIDE INFINIBAND SERVER REVENUE OPPORTUNITY BY FORECAST SCENARIO, 2003–2006

Source: IDC, 2003

IT Perceptions about Ethernet:

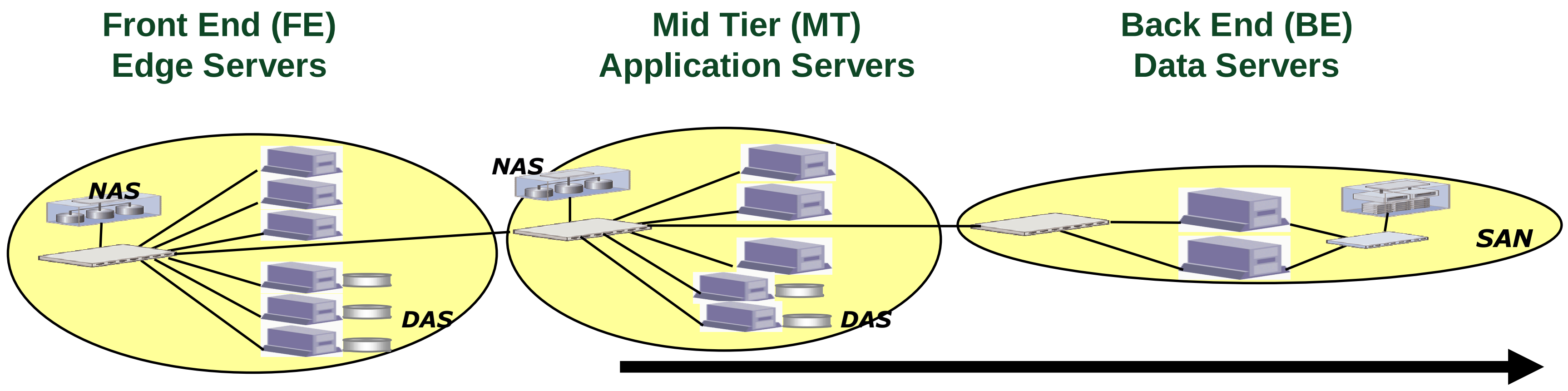
- “Ethernet not adequate for low latency apps”
- “Ethernet frame loss is inefficient for storage”

Market Opportunity

- Clustering & Grid computing (RDMA, iWARP)
- Storage (iSCSI)
- Telco Backplanes

Extend Ethernet Reach by improving congestion management capabilities

Emerging Blade Usage Models



- Blades are increasingly being deployed in BE & MT applications
- Ethernet is the default fabric of choice for LAN
  - In addition to Ethernet, Blades use Fiber Channel and Infiniband® for supporting Storage and Inter-processor communication traffic today
- Ethernet Blades are a growing piece of Telco pie ~ 26% of Telco servers by '07 – In-Stat/MDR

Requirements and Scope

CM Requirements for Datacenter

- Address IT perceptions:
  - “Ethernet not adequate for low latency apps”
  - “Ethernet frame loss is inefficient for storage”
- Improve Ethernet Congestion Management capabilities that will:
  - Reduce frame loss significantly
  - Reduce end-to-end latency and latency jitter
  - Achieve above without compromising throughput
- Address needs of Short Range Networks
  - Backplanes
  - Clusters
- BUT “Do No harm” if enabled in other topologies

CMSG Discussions - Recap

- Existing Link level mechanisms for congestion control do not improve network throughput
  - Head of line blocking
  - Congestion spreading
  - Increase jitter for high-priority traffic
  - Sacrifices throughput for avoiding frame loss
- Congestion control can be done at data source that is causing congestion
  - However, congestion happens somewhere else (bridges, destination nodes etc.) Congested devices need to provide information finally to source
  - Data sources can respond by reducing traffic into congested paths

Applicability of CN from Bridges

- Congestion Management is achieved by:
  - 802.1 Bridges providing congestion information
  - Data Sources (ULP) providing Rate Control mechanisms
- Remaining presentation focuses on Ethernet (802.3) networks
- However, 802.1 enhancements may be viable for other networks as well
  - 802.17, 802.11 etc.

Congestion Control Elements

- Detection
  - Could be an AQM like RED – Does not need to be specified by IEEE 802
- Notification
  - Need a standard way to notify congestion between L2 devices
    - Request to IEEE 802.1 to consider
- Action
  - Rate control/reduction done by source in response to congestion notification
  - Left to ULPs (L3 and above) e.g. TCP
    - IETF Domain

Congestion Notification Mechanisms

Congestion Indication mechanisms

- Packet Marking (triggered by congestion event)
  - Forward Marking of the packet experiencing congestion
    - Leave it to upper protocol for getting information back to the source
  - Or Backward Marking of packets going to congestion source
    - Which source (L2, Upper Protocol, what granularity)?
- Control Message
  - Send control packet to congestion source triggered by congestion
    - Which source? Granularity - L2, Upper Protocol, Socket,??
    - Should be in fast-path
  - Periodic Control messages carrying congestion information

More discussion on Backward Notification

- Faster turnaround, support for asymmetric traffic sources (e.g. non-TCP flows)
- Backward Notification creates traffic in congested networks
  - Can argue that transient congestions may not affect same paths simultaneously
- How to define granularity
  - Is L2 information sufficient?

L3 Marking Mechanisms : IP-CE

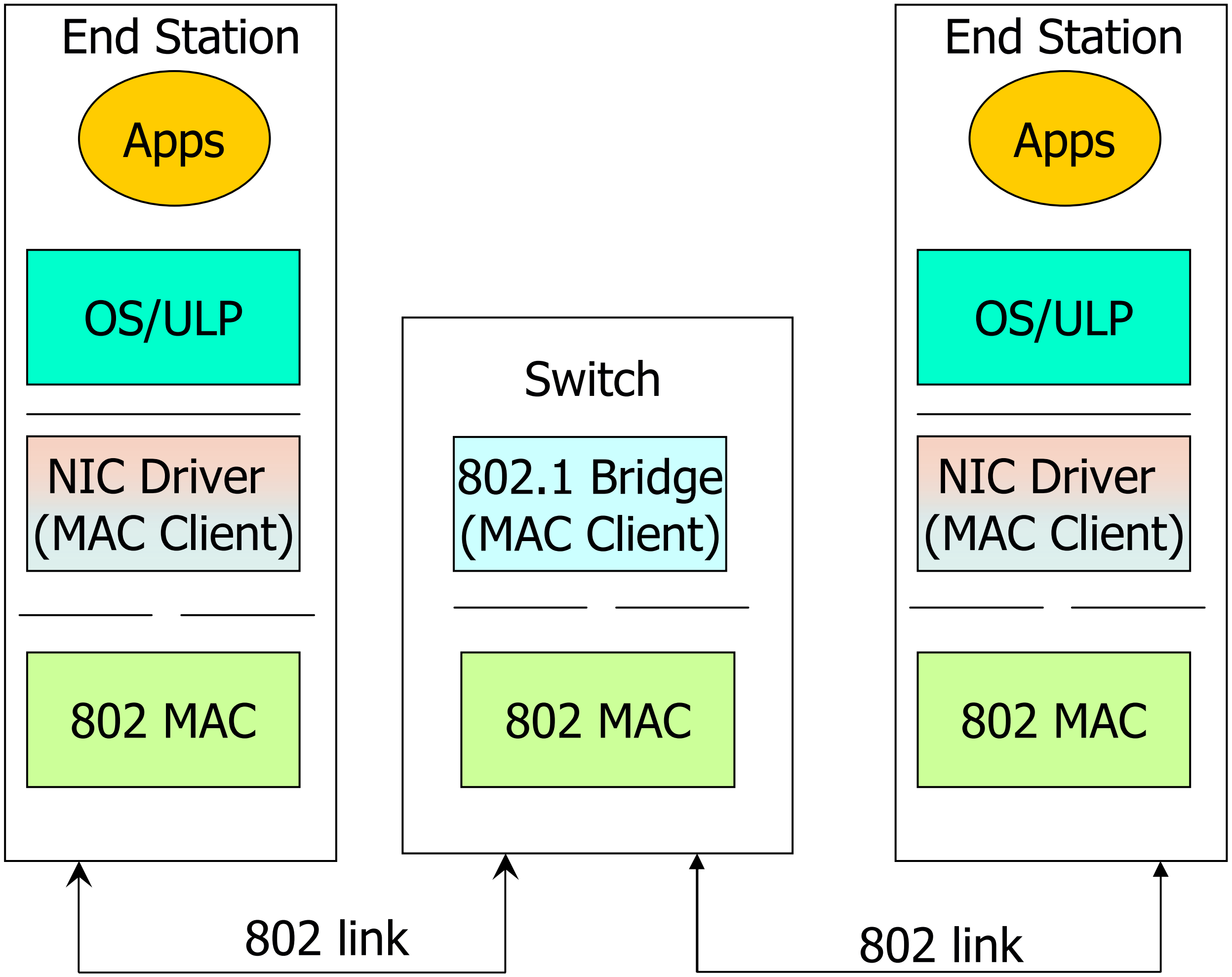
- IP – CE (Congestion Experienced)
  - IP-CE marking by routers or L2+ Switches when congestion is experienced
- Pros:
  - Will provide ECN capability within L2 Subnet
  - No change required in end-station implementations
- Cons:
  - Enables only IP (TCP) applications
  - Can not support asymmetric traffic
    - Backward notification
  - How does one standardize this mechanism for L2 Bridges?
    - Layer violations can make maintenance difficult (Support future changes in Upper Layers (IPv4, IPv6 etc.)
    - Security challenges?

L2 Marking Mechanism proposal : L2-CI

- L2-CI (Congestion Indication)
    - Marking by bridges in L2 header during congestion
- Pros:
    - Standardized congestion notification mechanism in L2 networks
    - Clean layering, ULP-agnostic
    - L2-CI and TCP-ECN together provide hierarchical mechanism
        - Equivalent to 802.1p and DSCP for CoS
- Cons:
    - Requires L2 header modification/extension for data frames
    - Requires End Stations to copy L2-CI information to ULP
        - E.g. to IP-CE code-point for TCP flows to benefit

L2-CI: details

Layered view of network



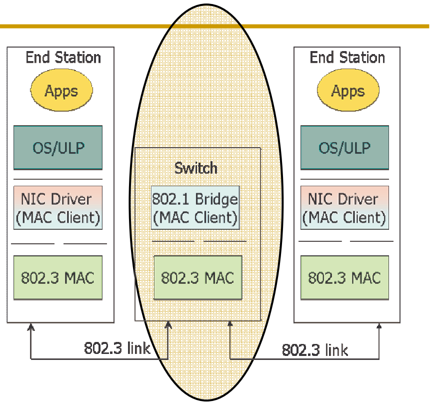
L2-CI : What it is and is not

- Is:
  - Mechanism for MAC Clients to provide congestion information
  - Enables MAC Clients to pass this information to upper layers (in end-systems typically) – API enhancements
    - Enables triggering Rate Controllers in upper layers
- Is Not:
  - Does not define congestion detection mechanism for MAC Clients
  - Does not define Rate Controllers in MAC Client
- How to achieve:
  - Use CFI bit in Tag Header
    - DE for Provider Bridge applications, CI for short-range networks
  - Definition of new L2 header (FESG can be leveraged)

DE and CI bit considerations

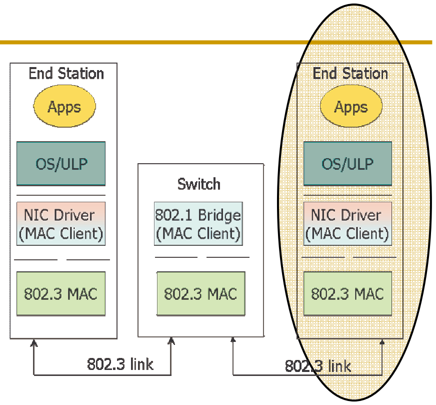
- Both mechanisms impact packets that “exceed traffic policy”
- DE: Packet is marked down making it eligible for drop in downstream switches
  - Primary target: Provider Bridge networks
- CI: Packet is marked so that sources can reduce injection rate
  - Primary target: Short range networks

Bridge Role:



- AQM to detect congestion
- When AQM threshold is exceeded, mark the packets (e.g. with probability for RED) on L2 header to indicate that “this” packet experienced congestion
  - Actual position/s in header TBD

End - Station Role:



- Copy L2-CI information from L2 header
- Pass it to Upper Layer through API (enhanced)
  - E.g. NDIS API may need to be enhanced to carry additional information
  - Should be easier to handle in Chimney architecture for offload engines
- ULP = TCP/IP
  - IP to copy L2-CI information received via enhanced-API to IP-CE bit before handing to TCP flow
  - TCP remains unchanged (Sends ECN-response back etc.)
- ULP != TCP/IP
  - Use L2-CI information to propagate backwards towards the source
    - Source can take appropriate Rate Controlling decisions
- End Node – MAC Client could also generate L2-CI

L2-CI Considerations

- More than 1 bit congestion information
  - Congestion levels in the path (e.g. XCP)
  - Hook for reverse congestion notification (to be used by non-TCP protocols?)
- Additional information about “capabilities” of flow
  - Equivalent to “ECT” bit in IP – ECN
  - At congested devices, “non-capable” flows get packets dropped instead of marked

Summary

- In order to enable accelerated deployment of Ethernet into emerging limited-topology applications (clustering, backplanes, storage, data centers, etc.), IEEE 802.1 should specify a standard mechanism for MAC Clients to provide congestion information to L2 edge devices, using wadekar_1_0501.pdf as a basis
- Any congestion notification mechanism defined by IEEE 802.1 should be agnostic to L3-protocols
  - IP-CE is not agnostic to L3 protocols
- L2-CI mechanism provides ULP agnostic Congestion Notification for short range LAN topologies
- Modeling data for L2-CI with TCP-ECN shows that L2-CI can provide significant improvement in throughput and latency reduction for short-range networks

Ref: http://grouper.ieee.org/groups/802/3/cm_study/public/september04/wadekar_03_0904.pdf