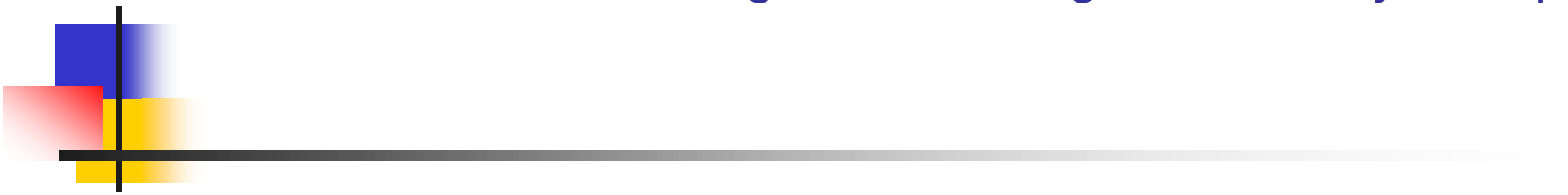
IEEE Congestion Management

Presentation for IEEE Congestion Management Study Group



Contributors

- Jeff Lynch – IBM
- Gopal Hegde -- Intel

Outline

- Problem Statement
- Types of Traffic & Typical Usage Models
- Traffic Characterization and Needs
- How IEEE 802.3 can help
- Summary

Outline

- Problem Statement
- Types of Traffic & Typical Usage Models
- Traffic Characterization and Needs
- How IEEE 802.3 can help
- Summary

Background and Problem Statement

- Ethernet treats all traffic equally
  - Designed decades ago to handle only LAN traffic in the Enterprise
- Today Ethernet carries a wide variety of traffic and is used in a wide variety of applications
- Various traffic types have different characteristics and require differential treatment to coexist on the same network
- So the problem statement could be: "How could 802.3 enhance Ethernet capabilities to provide differentiated service for IPC, storage and network traffic during congestion"

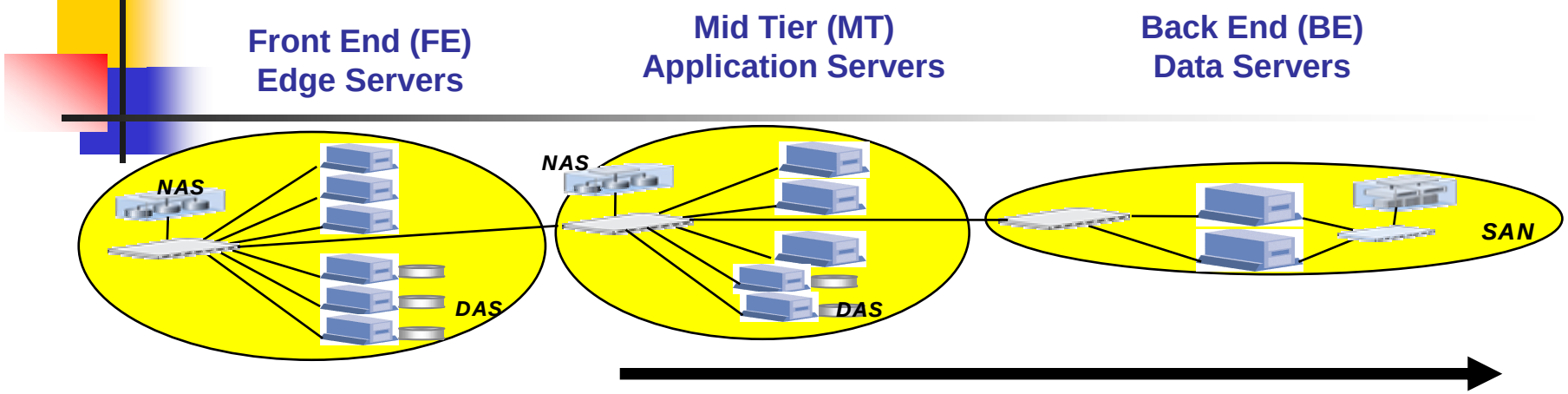
Outline

- Problem Statement
- Types of Traffic & Typical Usage Models
- Traffic Characterization and Needs
- How IEEE 802.3 can help
- Summary

Traffic Types & Usage Models

- Ethernet is being used to carry various types of traffic like
  - Enterprise LAN Traffic
  - Storage Traffic (e.g. iSCSI)
  - Inter-processor Communication (IPC) Traffic
  - Voice Over IP Traffic
  - Video Over IP Traffic
- Example new usage includes
  - Backplane for
    - Enterprise Blade Servers
    - Telecom Blades (advanced TCA)

Emerging Blade Usage Models



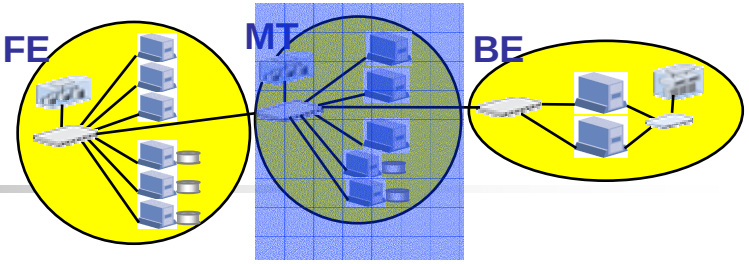
- Blades are increasingly being deployed in BE & MT applications
- Ethernet is the default fabric of choice
  - In addition, today's blades use Fiber Channel and Infiniband® for supporting storage and IPC traffic
- Ethernet traffic differentiation will improve applicability to storage and IPC traffic
- Ethernet blades are a growing part of the server market
  - ~ 26% of Telco servers by '07 – In-Stat/MDR
  - ~ 27% of Enterprise servers by '07 – IDC

NAS = Network Attached Storage
DAS = Direct Attached Storage
SAN = Storage Area Network

Outline

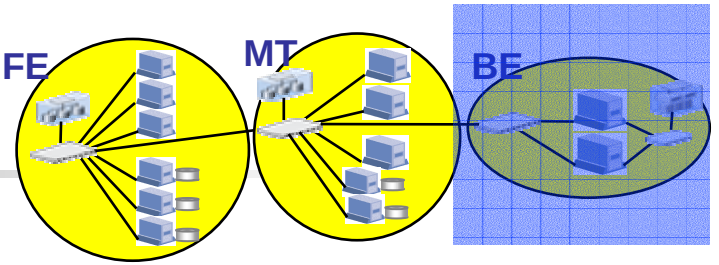
- Problem Statement
- Types of Traffic & Typical Usage Models
- Traffic Characterization and Needs
- How IEEE 802.3 can help
- Summary

Mid Tier (Application) Server Workload Characterization



- Between the FE & MT – <u>100% Networking Traffic</u>
  - Majority of messages < 1 KBytes
  - Connections typically have longer life compared to FE servers
- Between the MT and the BE – <u>100% IPC</u>
  - MT ➔ BE: Lock/Read Requests < 1KBytes
  - BE ➔ MT: Mixed small and large messages (>4KBytes)
- Storage traffic is < 200 Mbps
  - Storage architecture is DAS; trending to NAS
- <u>Platform Requirements:</u>
  - NW throughput is limited by server processing capacity
    - TCP-Acceleration/Offload helpful
  - IPC requires very low latency
  - Storage traffic sensitive to dropped packets

Back End (Database) Servers -- Workload Characterization

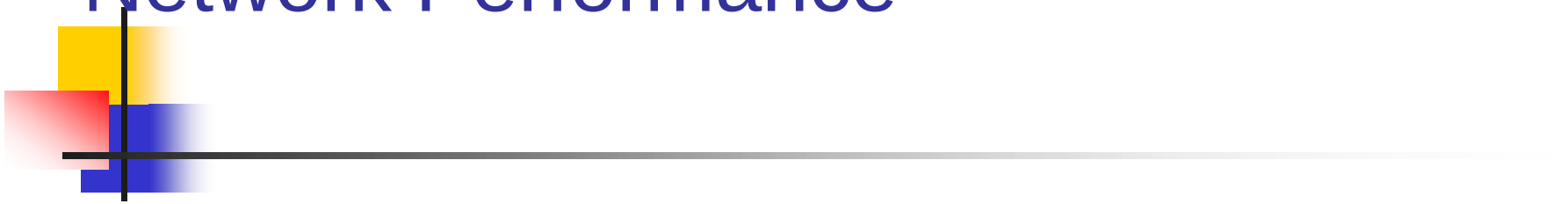


- Limited Networking Traffic (<200 Mbps)
- Storage Traffic is between 1 to 4 Gbps
  - Storage Architectures NAS & SANs
- Within BE – 2-3 Gbps of IPC Traffic
  - 80% Small messages for synchronization, locking, etc.
  - 20% Large messages for exchanging data (often > 4KByte)
- Platform Requirements:
  - Offloading / TCP Acceleration to reduce high Server Processor Utilization
  - Improved NIC and Fabric Latency
  - Improved Storage Subsystem Efficiency
    - Reduce probability of packet drops

Outline

- Problem Statement
- Types of Traffic & Typical Usage Models
- Traffic Characterization and Needs
- How IEEE 802.3 can help
- Summary

Current Standard Efforts to Improve Network Performance



- Storage over IP work (iSCSI) within IETF
- Remote DMA (RDMA) within the RDMA consortium
  - Reduces copies and therefore latencies
- TCP Acceleration Capabilities
  - Reduces TCP/IP handling overhead
- Quality-of-service protocols (IETF)
  - Differentiated Services (“diffserv”), and Integrated Services (“intserv”)
  - Resource Reservation Protocols (RSVP)

Above Efforts will require support of L2 enhancements

Why Enhance 802.3?

- Frame Discard in Response to Congestion
  - TCP timeouts & retransmits = big performance hits
    - Cluster traffic prone to frequent congestion events
    - Storage retransmits are big performance limiters
- Variable LAN latencies (loaded) can be in milliseconds
  - Voice, video, real time apps sensitive to delay & delay variance
  - Additive over many hops
- IPC Requirements
  - Some IPC requires mean latency < 10 uS – loaded
- Un-optimized switch buffers
  - Smaller buffers enable lower per-port cost

Possible 802.3 Areas of Focus

- To improve Ethernet's layer 2 capabilities, 802.3 needs to
  - Minimize latency jitter sensitivity due to transitory congestion
    - IPC Traffic
  - Reduce probability of MAC Client packet drops due to over-subscription
    - Storage traffic
  - Extend the differentiated services of upper layer protocols into 802.3
- Improvements will enable broader and faster Ethernet deployment into new market segments

Summary

- Ethernet is being used to carry a variety of traffic types and in a variety of new usage models
  - These traffic types have different characteristics
- Ethernet treats all traffic equally
  - Needs to evolve to treat various traffic types differently
- Proprietary, non-interoperable solutions address this need today
- Market is better served by addressing with an 802.3 standard
  - Provides interoperability and consistency