Proposal for Traffic Differentiation in Ethernet Networks

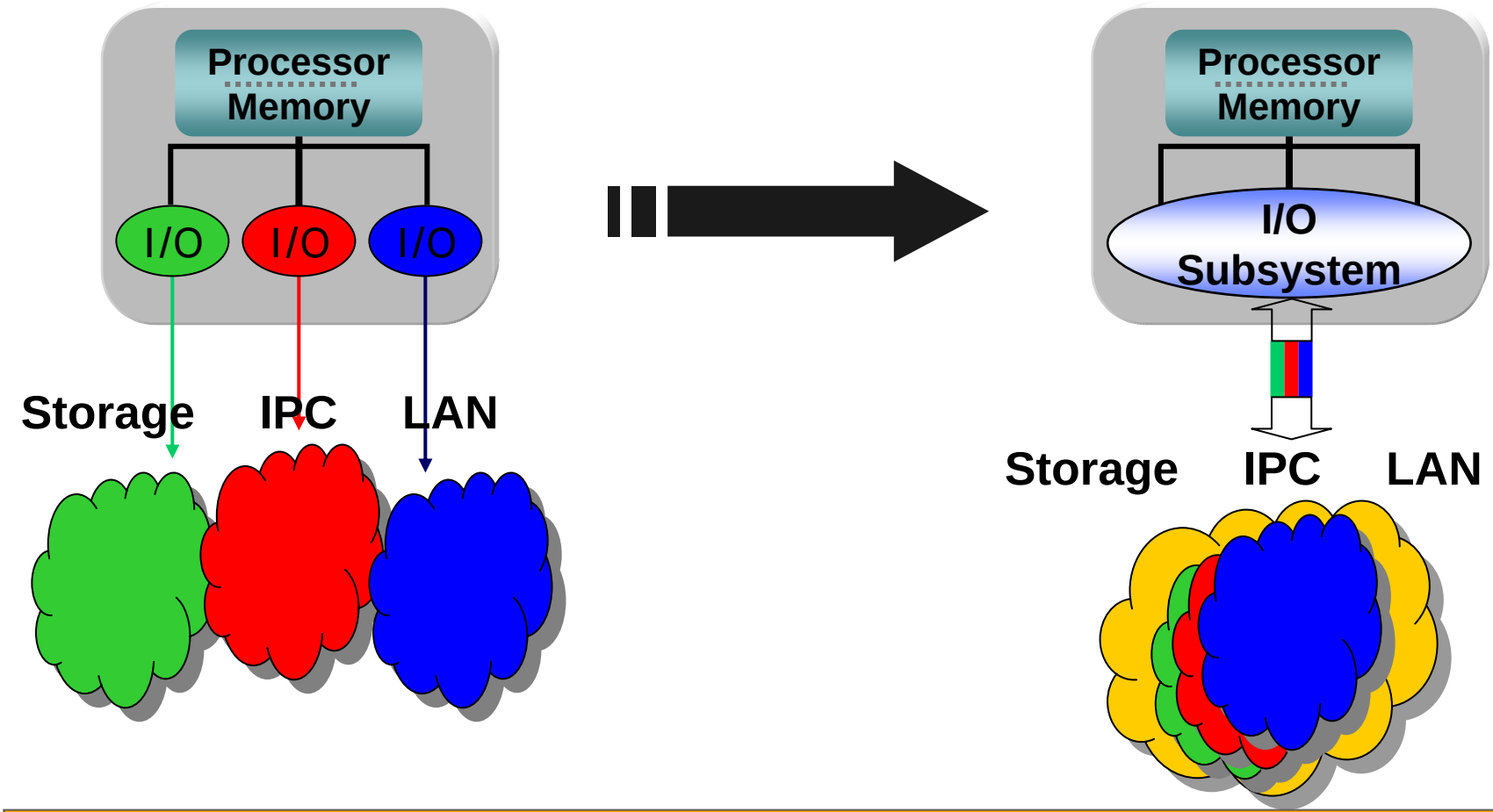
Intel : Manoj Wadekar
Gopal Hegde
Cisco: Norm Finn
Hugh Barrass
IBM: Jeff Lynch

March 14, 2005

Outline

- I/O consolidation in Datacenter
- Traffic types and requirements
- I/O Consolidation options
- Proposal for Virtual Links
- Summary

I/O Consolidation in Datacenter



I/O Consolidation simplifies platform architectures, reduces overall platform costs

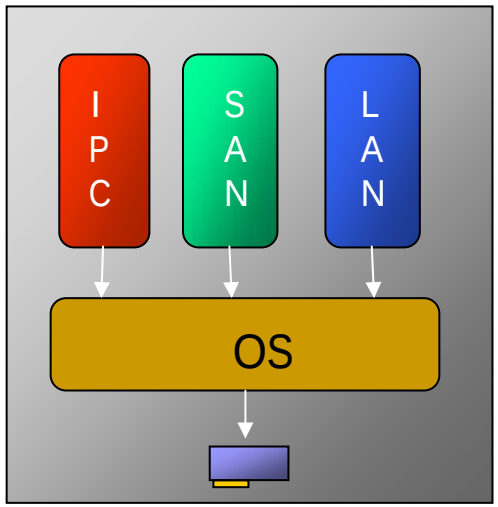
Traffic Types and Requirements

- Datacenter Ethernet to carry LAN, SAN and IPC traffic : I/O consolidation
  - Eliminates multiple backplanes (Blade Server application)
  - Should support appropriate characteristics for each traffic type
- LAN:
  - Large number of flows, not very sensitive to latency
  - E.g. dominant traffic type in Front End Servers
- SAN:
  - Large packet sizes, sensitive to packet drops
  - E.g. MT and BE servers
- IPC:
  - Mix of large & small messages, small messages latency sensitive
  - E.g. BE Servers, HPC Applications

Challenges in traffic differentiation

- Link Sharing (Transmit)
  - Different traffic types may share same queues/links
  - Large burst from one traffic should not affect other traffic types
- Resource Sharing
  - Different traffic types may share same resources (e.g. buffers)
  - Large queued traffic for one traffic type should not starve other traffic types out of resources
- Receive Handling
  - Different traffic types may need different Receive handling (e.g. interrupt moderation)
  - Optimization for CPU utilization for one traffic type should not create large latency for small message for other traffic types

Consolidation Options



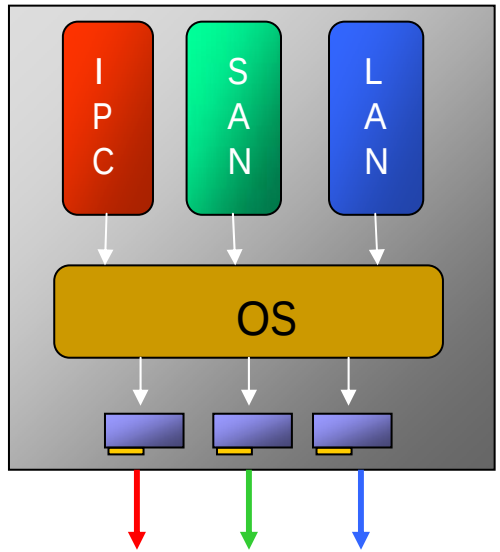
LAN: TCP/IP, UDP

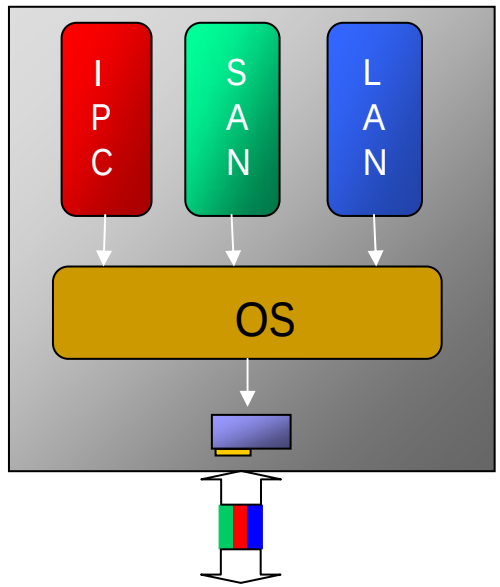
SAN: iSCSI

IPC: RDMA, iWARP

Physical Partitioning
-Consolidation on
Ethernet

Partitioning with
VLAN or .1P



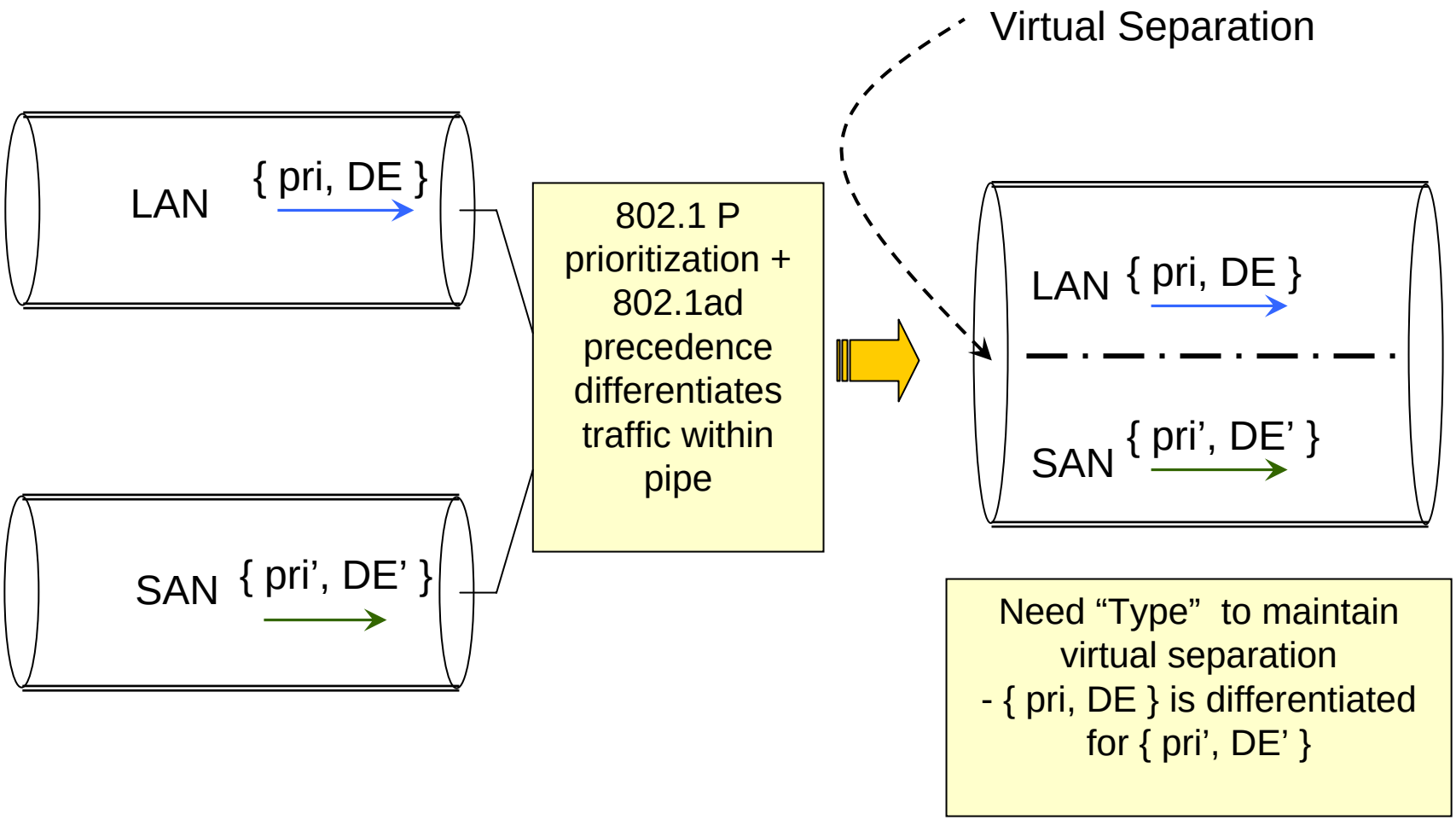


Limitations of current options

- Physical Partitioning
  - Does not reduce cost and complexity of Fabric Interconnects
- VLAN Partitioning
  - VLANs = Broadcast Domain, Subnets
  - SAN (iSCSI) and LAN traffic may belong to same subnet (VLAN)
    - Can not use VLAN as “partition”
- Priority Partitioning
  - Simplest alternative. Current 802.1p specifies only scheduling algorithm, no resource association
  - Standard .1p queue draining algorithms that allocate bandwidth resources are needed
  - This does not address the need to throttle sources

We need partitioning while maintaining prioritization

Virtually Partitioned Traffic

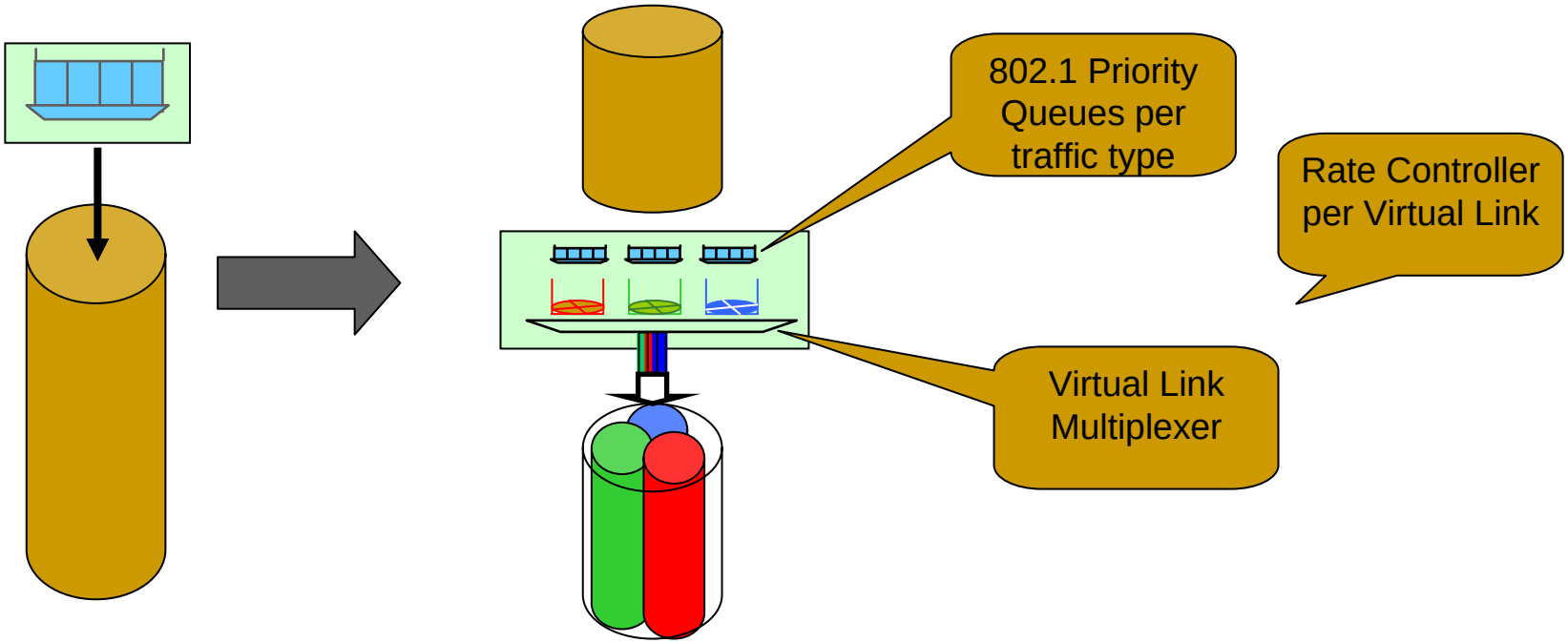


Virtual Links

- "Virtual Links" can provide differentiation among traffic types (LAN, SAN, IPC etc.)
  - BW can be associated with Virtual Links
  - Resources could be associated with Virtual Links at the network nodes (Different traffic profiles)
  - Interrupt moderation/receive handing differently for each Virtual Link
  - Traffic rates can be adapted according to congestion feedback
- Proposed changes to Queue management and resource association
- No contemplated changes to FDB, VLAN membership, etc.

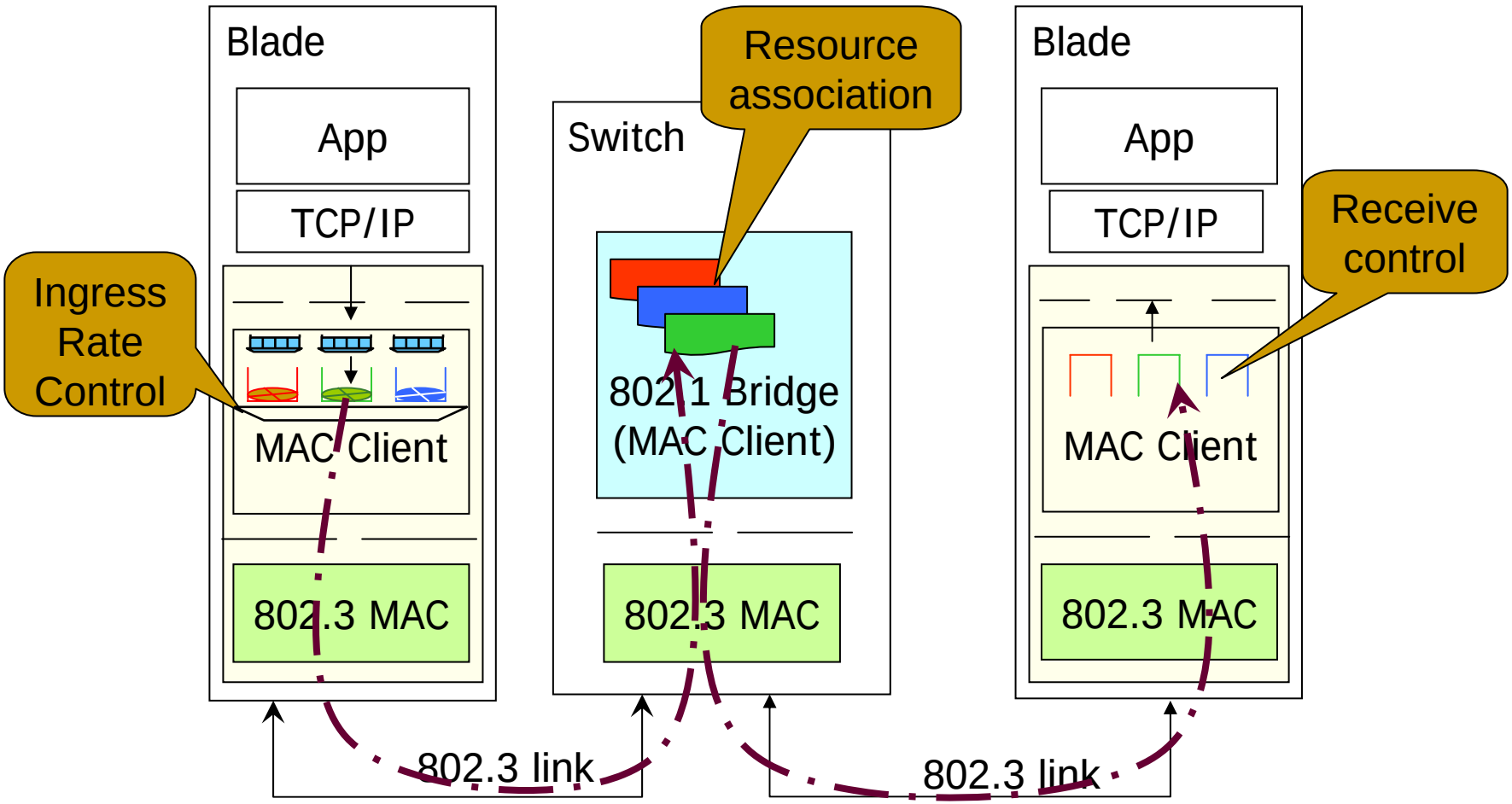
802.1 should consider defining required changes for Virtual Links

Virtual Links and 802.1p



- BW shared across multiple partitions
- Guaranteed access to multiple traffic types: Maintain priority among various flows within a traffic type
- Resources reserved per “Virtual Link”
  - Different profiles for each traffic type
- Need to allow utilization of available BW to compensate for jitter

Packet through the network



Flow Control and Virtual Links

- Link level flow control provides insurance against packet drops during transient congestion
  - Real time effect of end-to-end congestion management
  - Infrequent occurrence of buffer overflow leads to packet loss
    - Remedied by PAUSE
- Link Level PAUSE creates HOL blocking for multiple Virtual Links
  - Oversubscription for one traffic type may create blocking for other traffic types
- Consider per-Virtual-Link flow control
  - Can be defined completely within 802.1

Summary

- I/O Consolidation is important for Datacenter Ethernet
- “Virtual Links” can provide appropriate differentiation allowing various traffic types to share Ethernet network
  - BW, Resources etc.
- 802.1 should consider defining standard mechanism for such differentiation
  - Work towards a proposal for May Interim meeting
  - Requesting discussion/suggestions