iSCSI Requirements

draft-haagens-ips-iscsireqs-00.txt

Randy Haagens
Director, Networked Storage Architecture
Hewlett-Packard Co.
Randy_Haagens@hp.com

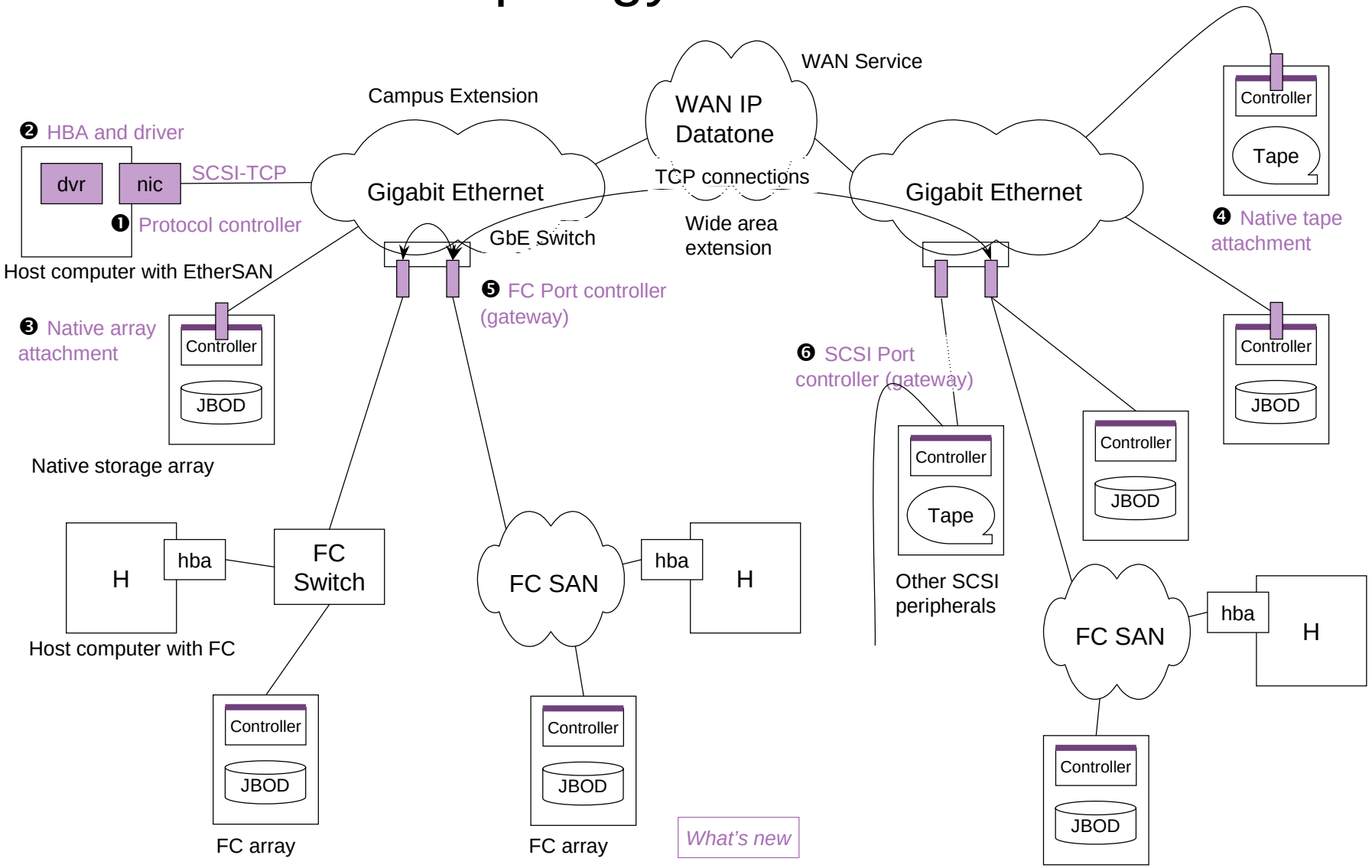
Applicability (Scope)

- iSCSI is a mapping of SCSI-3 to TCP, a "SCSI transport"
- Volume/Block storage on IP Networks (LAN, MAN and WAN)
  - Analogous to today's SAN architectures
    - Typically using Ethernet instead of Fibre Channel
  - Using SCSI protocol
    - SCSI for volume/block storage (NFS and CIFS for file storage)
  - Gateways to other SCSI interconnects
    - Fibre Channel, Parallel-bus, potentially others
- Benefit from IP/Ethernet infrastructure
  - Increasing performance and reduced cost
  - Seamless conversion from local to wide area using IP routers
  - Emerging availability of "IP datatone" services
  - Protocols and middleware for management, security and QoS
  - Economics arising from a single type of network

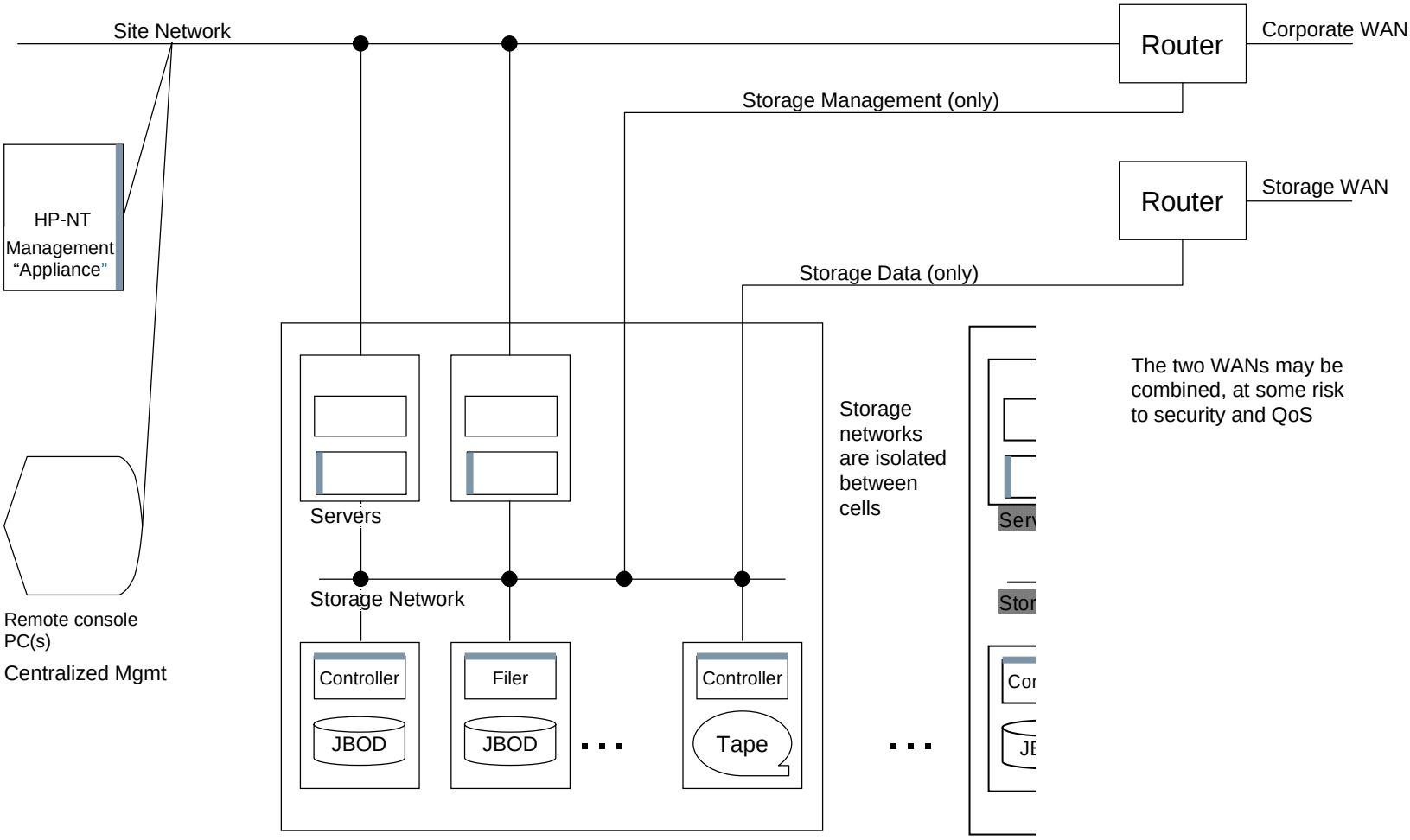
Applicability (Scope)

- Applications
  - Local storage access, consolidation and pooling
  - Remote disk access (as for a storage utility)
  - Local and remote synch and asynch mirroring between controllers
  - Local and remote backup and restore
  - Evolution with SCSI to support emerging object storage model
- Topologies
  - Point-to-point direct connection
  - Dedicated storage LAN, consisting of one or more LAN segments
  - Shared LAN, carrying a mix of traditional LAN plus storage traffic
  - LAN-to-WAN extension using IP routers or carrier “IP datatone”
  - Private networks and the public Internet

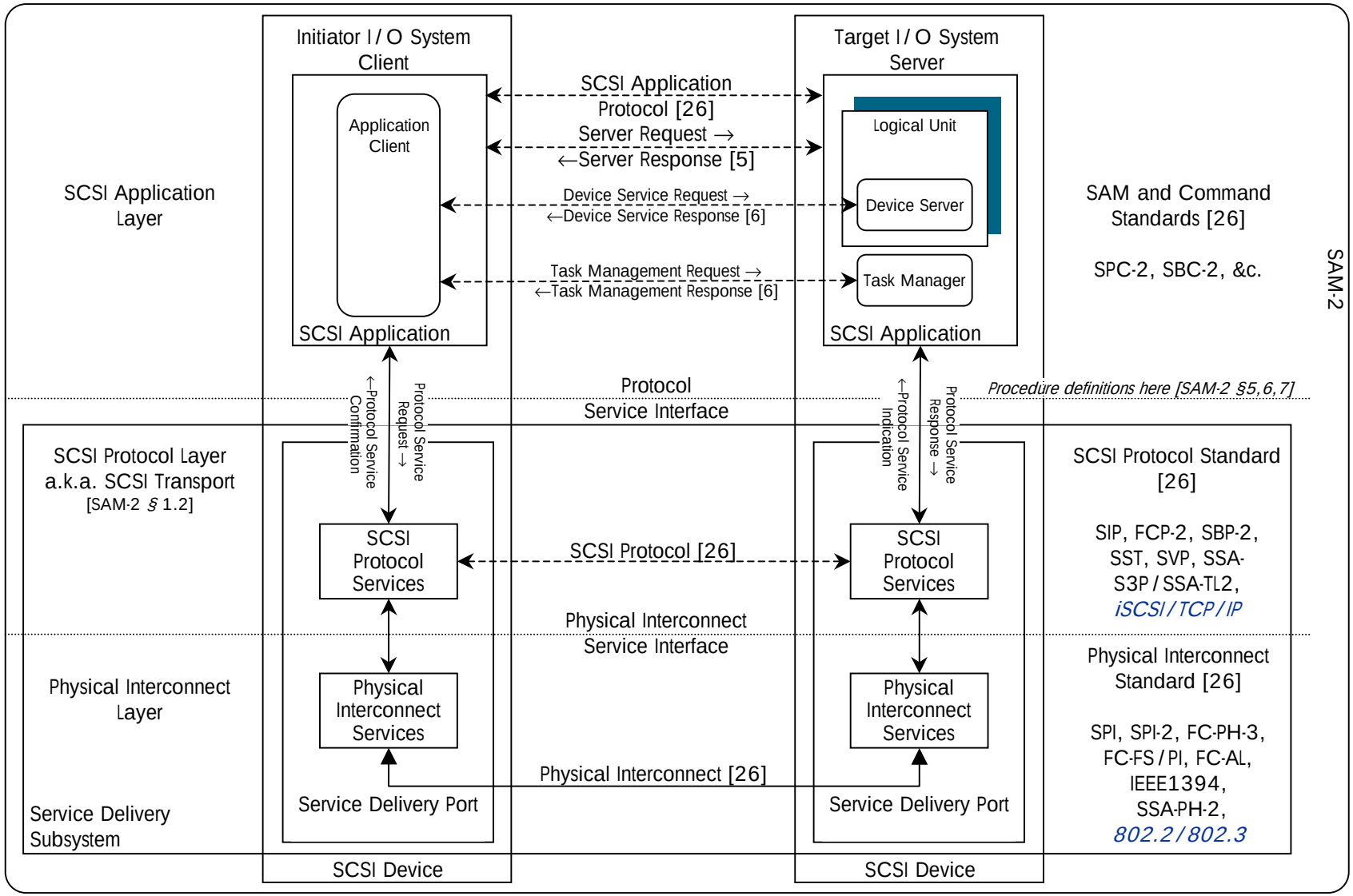
iSCSI Solution Topology



iSCSI Solution Topology

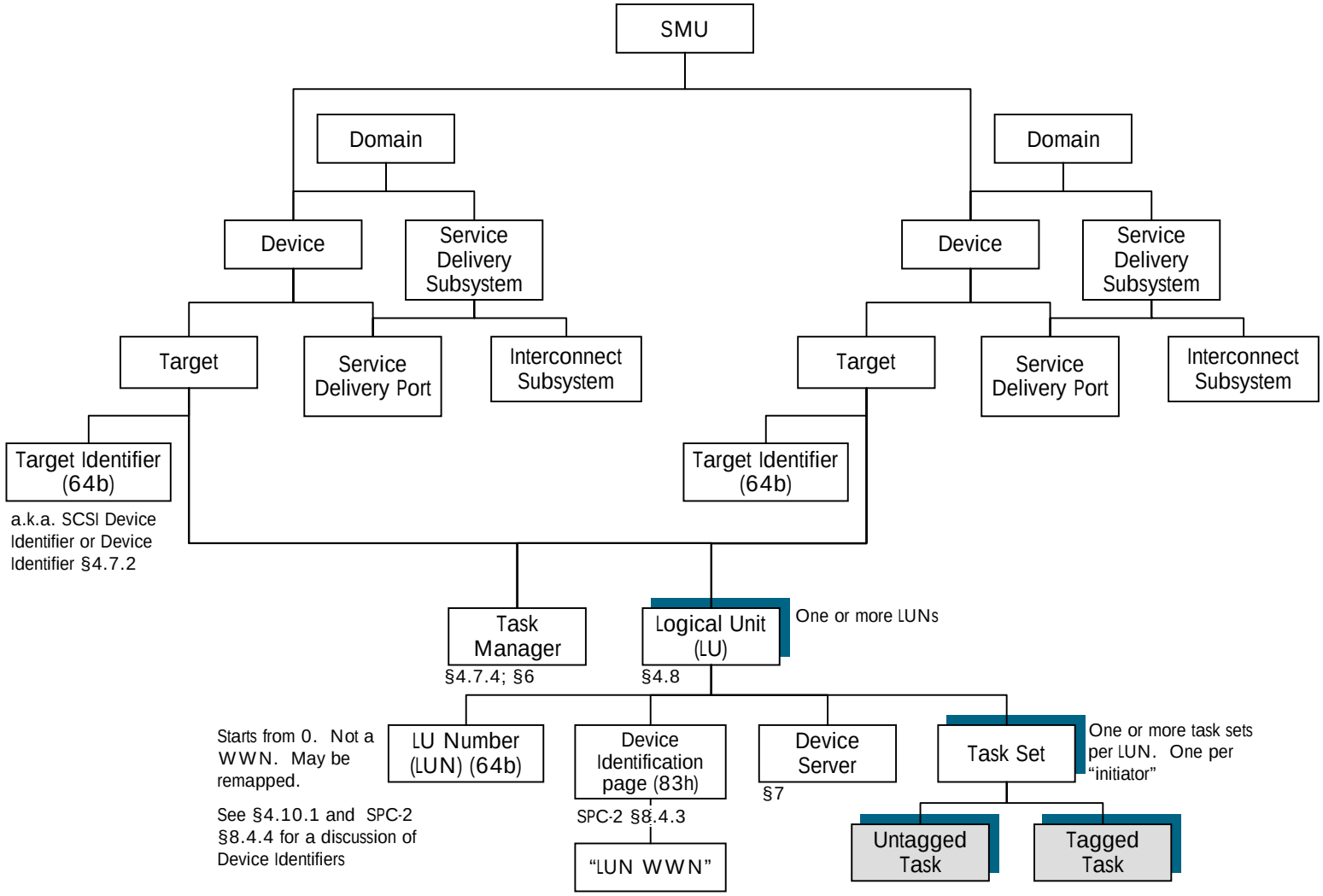


4.12 The SCSI model for distributed communications



Composite of SAM-2 Fig.s 2, 5, 6, 7, 9, 26, 28

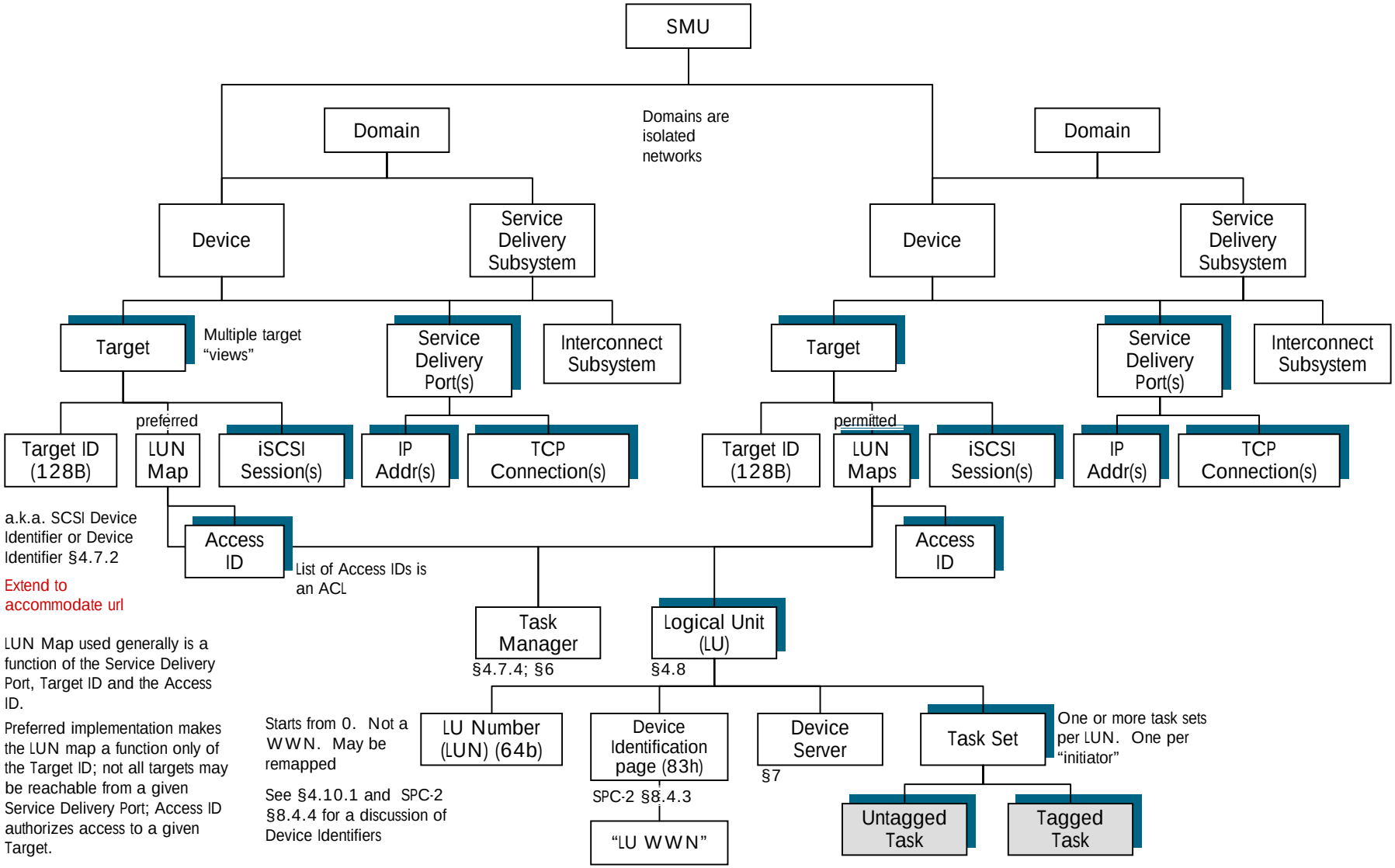
SCSI Multiport Target Unit



SCSI-layer Issues

- Naming of SCSI targets and LUs
  - 64b Target ID limitation imposed by SAM-2
  - Names vs. addresses of SCSI LUs
  - 3rd party copy (reference to LU)
  - Compatibility with new Access Controls model [T10/99-245 rev 8]
- Multi-port device model
  - What exactly is a SCSI Service Delivery Port in the iSCSI session model?
- In-order delivery of Task requests (commands)
  - SCSI attributes that control ordering of task execution depend on in-order task delivery
  - iSCSI layer is complicated by need to deliver tasks in order
    - Command numbering
- Gateway architecture
  - Gateways to parallel SCSI and SCSI-FCP are contemplated

iSCSI Multiport Target Unit



iSCSI-layer Issues

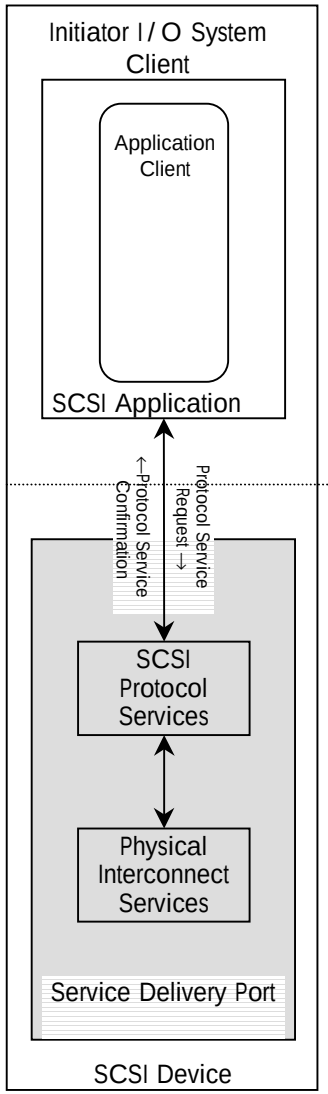
- Naming
  - URL syntax proposed: scsi://<domain-name>[/modifier]
  - Include SCSI “target” in name? Views, mapping
  - URL syntax: length problem (SCSI Target ID 64b limit)
- Connection allegiance
  - SCSI task command/data/status in same TCP connection
- Session Concept
  - A group of TCP connections
  - Supports ordered command striping for bandwidth aggregation
  - Recovery from TCP connection failure
    - SCSI task retry
    - “Replay buffer” may be required
- Possibly need an iSCSI layer CRC
  - Concern about TCP’s checksum robustness
  - More end-to-end even than TCP

SAM-2 Service Delivery Port

3.1.89 **service delivery port:** A device-resident interface used by the application client, device server or task manager to enter and retrieve requests and responses from the service delivery subsystem. Synonymous with "port" (3.1.61)

4.6 …the Service Delivery Port object represents the hardware and software that implements the protocols and interfaces between servers or clients in the SCSI Device and the Interconnect Subsystem.

3.1.81 **SCSI Multi-port unit:** A device that has multiple service delivery ports (see 3.1.89) or responds to multiple SCSI device identifiers (see 3.1.79)…



SAM-2, SCSI-3 Commands

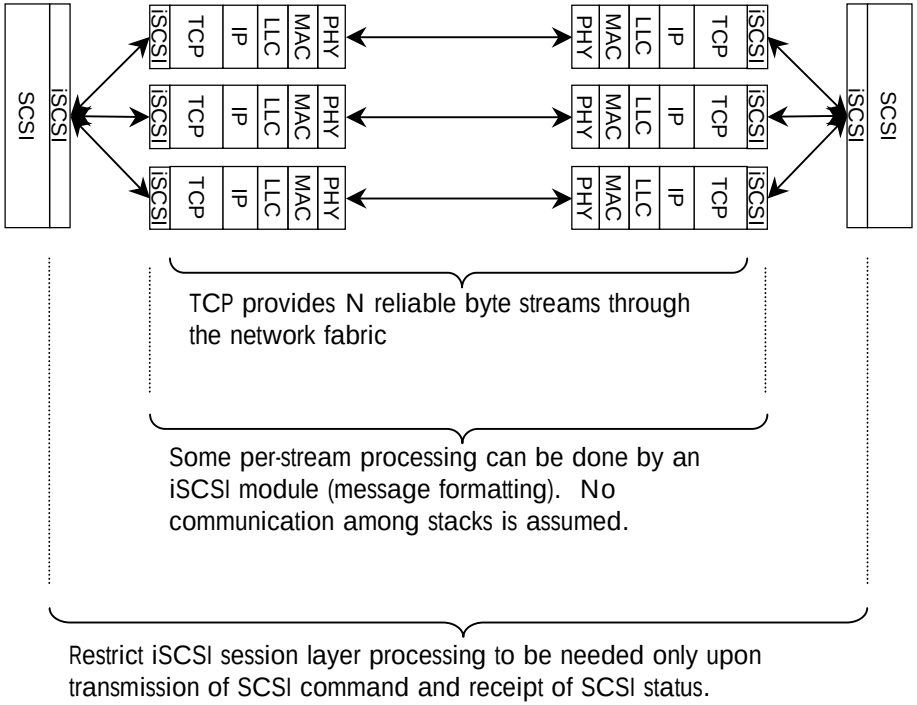
Protocol
Service Interface

FC-4 SCSI-FCP	iSCSI	TCP	iSCSI / iSCSI / iSCSI	TCP / TCP / TCP
FC-3 Common Services				
FC-2 Framing FC-1 Coding (FC-FS)	IP	802.2 LLC / Ethernet Framing	802.3 Media Access	Physical
FC-0 Physical Interface (FC-PI)				

IP	IP	IP	IP
802.2 LLC / Ethernet Framing	LLC	LLC	LLC
802.3 Media Access	MAC	MAC	MAC
Physical	PHY	PHY	PHY

With channel bonding / port aggregation

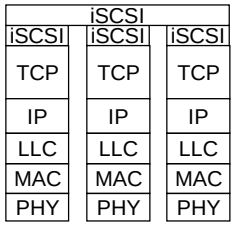
iSCSI Session Concept



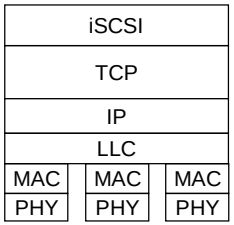
TCP-layer Issues

- Recovery of data stream processing following segment drop
  - Segment drop may result in loss of iSCSI framing
    - Unable to move data to final location until framing is recovered
    - Pipe may contain 250 MB of data (at 10 Gbps)
  - RDMA or a framing mechanism may solve the problem
- Error detection
  - Link layer is not end-to-end in IP networks
  - TCP checksum strength possibly inadequate
  - IPsec message digest could be used for increased strength
  - Alternatively, a CRC for TCP?
- Selective retransmission desirable
- Possible use of SSL/TSL in security architecture

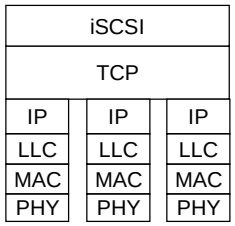
Aggregation Alternatives



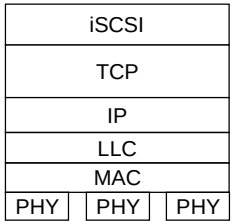
Proposed for iSCSI. Commands and status iSCSI messages are sequenced independently, in a central iSCSI module. Other iSCSI functions can be delegated to the individual protocol stacks. Multiple TCP/IP engines operate independently.



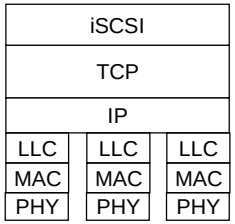
Effectively the same as above, with the additional problem that it adds a link dependency.



TCP is modified to aggregate over multiple IP addresses. That means that an end node can have multiple IP addresses, and the TCP implementation is able to load balance across them. Segments for the TCP connection frequently arrive out of order at the several interfaces, but TCP puts them back in order using its sequence numbers. Problem: TCP connections are currently defined by the (IPaddr, Port, IPaddr, Port) 4-tuple. There is no TCP-layer connection ID to relate segments arriving on different IP addresses. Potential problem: One TCP engine must service all links, and could become a bottleneck.



As specified by 802.3ad. Problem: frames for the same TCP connection will take the same link in a link bundle (so that they will arrive in order, which is not what's desired here).



IP does the aggregation, balancing traffic over multiple links. Problem: current routers would have difficulty preserving parallel flows in the last hop, as they would tend to discover (through ARP) only one destination MAC address for a given IP address.

Other Issues

- Topology discovery
  - Uses conventional IP endpoint discovery techniques
  - A means for discovering that an IP end point is an iSCSI node
  - A means of determining the IP connection topology within the end node
  - A means for acquiring a list of valid targets
  - SCSI protocol-dependent means for discovering LU topology
- Security
  - Security requirements are discussed by Steve Bellovin in this session