Introduction to Storage Area Network (SAN)

Fulvio Risso

http://staff.polito.it/fulvio.risso/
The path toward datacenters

**Historic trends**
- From mainframe to client-server (1980 – 2000)
- From client-server to ... (2000 - )
  - Peer-to-peer
  - Datacenters
  - Not clear who will be the winner (if there will be a single one)

**Datacenter consolidation**
- Storage
  - More flexibility in using storage space
  - Disk not tied to computational resources
- Computational resources
  - More flexibility
  - Power consumption
The storage evolution: outline

- Direct Attached Storage (DAS)
- Small Computer System Interface (SCSI)
- Network Attached Storage (NAS)
- Storage Area Network (SAN)
  - SAN architectures
    - Ethernet – TCP/IP
      - I-SCSI
      - FC-IP
    - Fibre Channel
  - Complexities and performances
    - Fibre Channel vs. Ethernet
The storage evolution

- Traditionally each server has exclusive access to storage devices:
  - Directly Attached Storage (DAS)
- Small Computer System Interface (SCSI)
  - Standard which define:
    - Set of commands
    - Protocol for transactions
    - Physical interface
  - Block-oriented, i.e. host’s O.S. see the storage devices like a contiguous sets of fixed data blocks
The storage evolution

- Max Length 25m
- Max 16 devices
- Parallel and shared bus

Windows, Unix, Novell, MacOS, ecc.

- Low latency
  - through disk (~ms) and through cache (~µs)
- Very low error rate
  - Inefficient error recovery
The storage evolution

- Great difficulties on most important corporate richness’ management: the data
  - Resource administration must be done on each server
  - No optimization
  - Scalability
  - Performance
  - Limited maximum distance between devices
  - Inaccessibility to data during maintenance or in case of server fault
  - Difficult backup management
  - Difficult sharing data among different servers
  - Difficult migrating a server to a more powerful one (either in terms of CPU or storage)

- Requirements:
  - Consolidation of storage resources
  - Centralized management
  - Remote replication of data (disaster recovery)
  - Centralized and transparent backup to LAN and computers
The storage evolution: a two-tier model

- Solution:
  - Separation between storage devices and computing resource
  - Connection realized using network’s technologies
  - Different implementations:
    - Costs
    - Performance

Diagram:
- Servers / clients
- Local Network or Geographic
- Storage
Network Attached Storage

- Characteristics
  - Usually a dedicated appliance, with proprietary or heavily optimized operating system
  - High storage capabilities
  - Use of RAID and Hot-Swap to protect data and guarantee continuity of service
  - Reasonably low cost
  - LAN-oriented
- NAS virtualizes shared disks
  - It serves *files* over the network, usually LAN (not WAN)
  - Raw file system invisible to client
  - It exports either Microsoft SMB or UNIX NFS data

*Single network connection*
NAS: the protocol stack

- Tipically TCP/IP over Ethernet
- TCP may introduce a non-negligible performance overhead
  - UDP may also be used
- File exported via NFS, CIFS (or both)
NAS: pro and cons

- Clients do not have full control of the disk
  - Cannot format the disk as they want
  - Cannot manage the disk at the block size (some applications, e.g. Oracle databases, do that for performance reasons)
  - Some features (e.g., checking for concurrent accesses) are always enabled, even if these are not necessary
  - Some applications require local disks for working
    - Cannot boot from a shared disk
    - Oracle DB
    - Swap file

- User manager
  - Controlling accesses means that the NAS must manage user rights and credentials
  - What about is the NAS is in outsourcing? Do I want the SSP to manage my accounts?
**NAS: pro and cons**

- Computational power required by the NAS appliance
  - Receives file-related request, which have to be remapped in block-related requests
  - User-rights management
- NAS protocol stack not optimized for performance
  - Ethernet and, most important, TCP
- High compatibility
  - Minimal impact on the existing infrastructure
  - All OS are able to mount a shared disk without additional drivers
  - Works also in WAN
Next step: Storage Area Network

- Virtualize physical disks, not logical volumes
  - Access to data through logical blocks and not to file
- Includes a network dedicated to the storage
  - Two-tier model; it may even be a three-tier model
**Storage Area Network**

- SCSI protocol for end-to-end communication
  - Maintains only the upper layers of the SCSI stack
  - This guarantees compatibility with all the existing SCSI-based applications
  - Minimum impact for DAS to SAN migration
  - Interaction with disks must be similar to the one we had in DAS
    - High speed
    - Low latency
    - Very low error rate
  - We cannot use Ethernet (alone) to build a SAN

- SAN features
  - Compatibility with an high number of nodes
  - Metropolitan distance coverage
  - High reliability and ability to react to failures
**JBOD**

- Just a Bunch Of Disks
  - A way to reduce costs
- Multiple disks are placed in a special cabinet, which exports a disk whose size is the sum of comprising disks
- No RAID used, therefore no data protection
  - Protection can be obtained through data redundancy
- Very cheap
- Common in SAN
SAN: the protocol stack
**SAN on Ethernet – TCP/IP**

- **Pro**
  - Network simplicity
  - Infrastructural and training costs are very low
    - May not be true in some environments
  - Prospective evolution uncomparable with respect to whatever rival technology
  - May have a single network (instead of LAN + SAN)

- **Cons**
  - No guarantees to receive transferred informations
    - Frame loss is, today, a feature of Ethernet functioning
    - Error recovery relies on TCP
      - Timeouts in the order of hundred/thousand of milliseconds
    - TCP hardware implementation is difficult
  - No guarantees on latency
Fibre Channel

- Born from the need of a reliable support for serial Ultra3 SCSI
  - Basically, a new physical layer for SCSI
  - Simple data plane (in fact, still SCSI)
- Support high transfer rate
  - 1Gb/s, 2Gb/s, 10Gb/s
- Include a lossless mode
- The control plane is complex
  - New features for managing disks
- Three possible working modes
  - Direct connection (hystoric)
  - Arbitrated loop (hystoric)
  - Mesh network
Fibre Channel: connection modes

- Direct Connection
  - Still used as SCSI replacement
- Ring (Arbitrated Loop)
  - Up to 127 nodes connected in ring topology physically linked or through hub (better reliability)
  - Hystoric
- Meshed network (Switched Fabric)
  - Switches are linked to nodes and between themself
  - Full duplex links
Fibre Channel: protocol stack

SCSI

FC-4: ULP Mapping
FC-3: Common services
FC-2: Signalling
FC-1: Transmission
FC-0: Physical layer

IP
Fibre Channel

- FC-0: Physical interface definition
- FC-1: Encoding and link’s low level control
- FC-2: End to end data transfer protocol
  - Frame format
  - Addressing
  - Segmentation
  - Flow control
  - Error detection/correction
- FC-3: Services common to every port
  - Cryptography
  - Compression
  - Channel bonding
- FC-4: Protocol mapping
  - Mapping between upper layer protocols and the transport layer for the delivery through the the fabric
Fibre Channel: ports

- Exist several kind of ports with specific functioning
  - N_port: HBA (Host Bus Adapter)
  - F_port: switches through HBA
  - E_port: connection between switches (ISL)
  - NL_port, FL_port: loop functioning
Fibre Channel: Addressing

- Address structure
  - Nodes, ports and switches have a unique 64 bit (“World Wide Name”) “Name identifier” which is assigned in factory
  - Dynamic assignment of 24 bit address for data exchanges

- Domain_ID from 00h to EFh.
  - Usually one per switch. Assignment managed by a main switch
  - 239 switches supported
  - From F0h to FFh “Well Known Address”
    - Services offered by Fabric
    - Implemented with distributed protocols in the switch internals

- Area_ID e Port_ID assigned to nodes, 65536 nodes per switch.

<table>
<thead>
<tr>
<th>Domain_ID</th>
<th>Area_ID</th>
<th>Port_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>16</td>
<td>8</td>
</tr>
</tbody>
</table>
Fibre Channel: Routing

- Need to propagate reachability of domains/areas across the fabric
- Protocols
  - FSPF: Fabric Shortest Path First
    - Link-state protocol similar to OSPF
- Routing and Loops
  - FC doesn’t have TTL mechanism: infinite loop of packets are possible
  - The network convergence must be as fast as possible
Fibre Channel: communication between nodes

- The communications between two nodes expects the an “exchange” opening
- Each exchange expects half duplex frame “sequences”
- Several kinds of communication
  - Flow control
  - Reservation of resources
  - Guarantee on ordered frame delivery
Fibre Channel

- Flow control
  - End to end
  - Buffer to buffer
  - Credits mechanism

- Problems
  - Deadlock
  - Traffic is blocked on the whole link due to lack of credits
Advanced aspects

- VSAN
  - Like VLAN, but on SAN
  - Interesting for Storage Providers
- Link Aggregation
- Load Balancing
Complexities and performances

- Frame size
  - Fibre Channel: 36 byte overhead
  - Ethernet - TCP/IP: 18 Ethernet + 20 IP + 20 TCP + iSCSI or FC-IP
- Fibre Channel vs. Ethernet
  - Performances
  - Guarantees
    - Investments protection
    - Developing
  - Prices
    - Different market segments

<table>
<thead>
<tr>
<th>SOF</th>
<th>Header</th>
<th>Opt Hdr</th>
<th>Data</th>
<th>CRC</th>
<th>EOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4)</td>
<td>(24)</td>
<td></td>
<td>(from 4 to 2112)</td>
<td>(4)</td>
<td>(4)</td>
</tr>
</tbody>
</table>
Using scenarios

- iSCSI
  - Low cost connection of hosts (Ethernet) to SAN
- FC
  - Connection between servers and disk arrays
- FCIP
  - Connection of SANs through a geographic link
    - Es. Backup or redundancy
Main vendors

- SAN
  - Brocade + McData
  - Cisco
- NAS
  - NetApp
  - HP, Dell
- “turnkey” solutions
  - IBM, EMC
The path toward 10GbE

- A server potentially needs
  - 1 NIC for LAN (usually Ethernet)
  - 1 HBA for storage (usually Fibre Channel)
  - 1 NIC for clustering (usually Infiniband)
  - x2 if we need redundancy

- This solution lead to
  - Unnecessary power consumption
  - A lot of PCI slots
  - A lot of space in racks
  - A lot of cables in datacenters

- Should be nice to have a single NIC for all
  - Ethernet seems to be the obvious choice
  - But the current speed is not enough
10GbE and I/O consolidation

- Ethernet was never an option in large datacenters
  - Limited speed (1Gbps against 2/4 Gbps FC)
  - No reliable delivery (no longer CSMA/CD, but congestions may happen in the network)
  - Datacenter managers were used to the FC management model
    - Not easy for them to switch to Ethernet (they do not know, nor trust, Ethernet)

- 10GbE delivers the required speed
  - Together with the latest version of the PCI Express bus
  - FCoE (next slide)

- Still the problem of reliability
  - Priority Flow Control (i.e., per-priority PAUSE)
    - Currently submitted at IEEE 802.3
  - Lossless behaviour at the network (not link!) level

⇒ I/O consolidation!
FCoE

- Enables FC frames over Ethernet
  - All FC protocols (e.g. DNS, ...) still active
  - All the tools for provision, manage, etc over FC are still working
    - FC at the logical level
    - Ethernet at the physical level

- Convergent Network Adapter in new hosts
  - Implements NIC + HBA, exported as two different physical adapters at the upper layers
    - Preserves application compatibility

- FCoE and FCIP
  - FCoE is oriented to datacenter environment
  - E.g., FCoE is not routable
The need for virtualization

- Storage virtualization
  - Already addressed by SAN
    - No need to talk more about that
- Computing virtualization
  - Better use of computational resources
  - Energy consumption
Power consumption (1)

- Typically, 5-15% CPU utilization (per server)
- Power consumption vs CPU utilization
Power consumption (2)

- Electrical power
  - Politecnico di Torino (2005): > 3M€

- Power consumption of a single server
  - 10GbE NIC: ~15W (2008)
  - 1TB disk: ~10W idle, ~15W R/W (2008)
  - CPU: may be more than 100W
  - Idle server: ~ 66% of the peak power
    - OS still running
    - Memory, disks, motherboard, PCI slots, fans still active
Virtualization and datacenters

- The “One application per server” rule in datacenters
  - Failure of popular OSes to provide
    - Full configuration isolation
      - E.g. A requires DLL version 1.0, B requires DLL version 2.0
    - Temporal isolation for performance predictability
      - If A is eating all the CPU, performance of B will worsen
  - Strong spatial isolation for security and reliability
    - A crash in A may compromise B
  - True backward app compatibility
    - My Application runs only on OS version X, path Y
    - Or... Sometimes it is certified only in this environment
Virtualization benefits

- Server consolidation
  - Exploit multi-core CPUs
  - Optimize energy consumption
    - Huge savings
- Decoupling physical hardware from logical servers
  - Rapid deployment of new servers
  - Move servers between different hardware
    - Either as image, or with server running (e.g., VMware VMotion)
  - Capability to give more CPU cycles to servers that require more power
    - Dynamic load balancing between server
  - Disaster recovery
    - Either static (move images) or dynamic (replace faulty server)
- Management
  - Secure remote console
  - Reboot / power control
  - Performance monitoring
  - Easier to setup a new server (no need to deal with different physical hardwares)
The Hypervisor

- Hardware is managed by stripped-down OS, the hypervisor
  - Often Linux-based
  - Native drivers manage hardware
  - A virtualization layer exports a set of “standard” devices to the upper-layer OS
    - Usually, we do not virtualize the latest video card
    - However, most important characteristics of the hardware can be exploited “natively”
    - Enable hosted OS to support a limited set of hardware

- The hypervisor may be attacked
  - Although no successful attacks are known right now (2008)
  - Much smaller and more defendable than a conventional OS
Bibliography

- Books
  - C. Beauchamp, J. Judd, Building SANs with Brocade Fabric Switches, Rockland, MA, USA, Syngress Publishing Inc., 2001

- Standard Fibre Channel
  - http://www.t11.org/index.htm

- Tutorials Fibre Channel
  - http://www.fibrechannel.org/OVERVIEW/
  - http://www.iol.unh.edu/knowledgeBase/training/fc

- Storage over IP

- Documentation from the top two vendors of Storage Switch equipments
  - http://www.brocade.com/san/resource.jsp