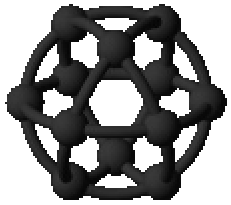
Tutorial on Distributed Disk Technologies



Jaqui Lynch
Circle4 Consulting
jaqui@circle4.com

UKCMG 2002
Session 2F1 and 2F3

Agenda

- Acronyms 101
- NAS and DAS and SANs
- Disk Technologies and Protocols
- Fibre Technologies
- SSA Technologies
- Raid Levels
- I/O Performance Notes
- Issues

Acronyms 101

- DAS Direct Attached Storage
- NAS Network Attached Storage
- SAN Storage Area Network
- SCSI Small Computer Systems Interface
- iSCSI internet SCSI
- FC-AL Fibre Channel Arbitrated Loop
- FC-SW Fibre Channel Switched
- SSA Serial Storage Architecture

NAS

- NAS
  - Storage device attached to TCP/IP network
  - Accessed using CIFS or NFS
  - NAS translates file requests to device requests
  - Consists of processor and internal disk
  - Low cost
  - Simple to deploy
  - Not very scalable

NAS Gateways

- NAS Gateway
  - NAS device with no internal disk
  - Attached to an external disk subsystem using DAS or SAN connections
  - Gateway sends file requests on to the disk subsystem
  - Accessed using CIFS or NFS

DAS

- DAS
  - Disk or Tape storage directly attached
  - I/O requests access the devices directly
  - Easy to deploy
  - Not very scalable

I/O Types

- Block
  - Application accesses a specific block
  - RDBMS
  - OLTP
- File
  - Application accesses a file
  - FTP, SCP
  - NFS, CIFS, Samba
  - HTTP
  - Word, Excel, Notepad, PowerPoint

new

SANs

- SANs
  - Storage devices reside on private dedicated network
  - I/O requests access devices directly using Fibre media
  - Can use multiple Fibre channel ports to increase aggregate transfer speeds
  - Cable lengths up to 10km
  - Usually multiple options for RAID level
  - Can be simple or extremely complex
  - Based of a fabric of switched, hubs, and gateways connecting devices and servers on a many to many basis
  - Can also enable storage to storage direct connections
    - allows mirroring, backup & archiving to be done independently

Network Storage

- SAN
  - Block I/O
  - Typically fibre channel
- NAS
  - File I/O
  - Typically Ethernet
- Distinctions are blurring
- Initiatives
  - iSCSI -- SCSI over IP
  - iFCP -- Internet Fibre Channel Protcol
  - FCIP -- Fibre Channel over Internet Protocol

new

Issues relating to SANs

- ANSI standard for storage networking still being finalized
  - Interoperability amongst vendors
- FC-AL versus FC-SW makes a huge performance difference
- Security and management of data that is now centralized and shared by multiple hosts
- Issues with disparate systems sharing central tape systems
- Understanding the various design points and how they affect performance

Which one should I use?

- DAS
  - Designed for single isolated processors and low initial costs
  - High performance transaction oriented applications
- SAN
  - Optimized for performance and scalability
  - Designed to be accessed by multiple systems concurrently
- NAS
  - Optimized for file-sharing and ease of use
- NAS Gateways
  - Designed to provide the benefits of NAS and the flexibility of the SAN

Disk Technologies

- Arbitrated
  - SCSI 20 to 160 MB/sec
  - FC-AL 100MB/sec
  - Devices arbitrate for exclusive control
  - SCSI priority based on address
- Non-Arbitrated
  - SSA 80 or 160MB/sec
  - Devices on loop all treated equally
  - Devices drop packets of data on loop

Protocols

Arbitrated

- Non-Arbitrated
- Escon
- SCSI
- iSCSI
- Ethernet
- Fibre

Protocols (cont)

- Arbitrated
  - i.e. SCSI or FC-AL
  - Devices gain exclusive control of the bus or loop by arbitration
  - Priority is based on address
  - One device can take over the bus
    - Especially if a streaming protocol (I.e. backup) is being used

Protocols (cont)

- Non-Arbitrated
  - i.e. SSA
  - All devices on loop or bus are treated equally
  - Devices share concurrent use of loop by passing packets
    - (SSSA is 128 byte + 10byte header)
  - Fairness mechanism used to protect from takeovers

Protocols (cont)

- Escon
  - Fibre technology used mainly for mainframes
  - Allows 17 MB/sec throughput
- SCSI
  - Arbitrated protocol
  - 25 meter limitation (distance from processor)
  - Speeds
    - SCSI FWD 20MB/sec SCSI Ultra 40MB/sec
    - SCSI Ultra2 80 MB/sec SCSI Ultra160 160 MB/sec

SCSI History

- Minicomputer interfaces
  - Defined specifically for one device
  - Not intelligent
  - Hardware and software changes required to support new devices
- Small Computer System Interface – SCSI
  - Began as Shugart Associates Systems Interface (SASI) in 1979
  - Based on IBM System/370 Bus and Tag channel
  - ANSI SCSI Standard published in 1986
  - Supports multiple devices at different rates
  - Logical command set hides device implementation details
  - "Small" Computer System Interface scales to large servers

new

Protocols (cont)

- iSCSI
  - Storage is attached to a TCP/IP based network
  - Accessed by block I/O SCSI commands
  - Can be DAS or network attached using SAN
  - Still evolving as a standard
  - Can be used instead of Fibre channel in a SAN
- Ethernet
  - 10 MB/sec, 100 MB/sec or Gigabit Ethernet
  - Primarily used by NAS devices and iSCSI

Fibre Technologies

- FC-AL - Fibre Channel Arbitrated Loop
  - Loop or hub based Fibre
  - Arbitrated protocol
  - Runs over 100 MB/sec Fibre
  - Bandwidth limited to speed of devices on loop as only one can speak at a time
  - Allows up to 126 devices on the loop using hubs
  - More than one system can share the loop
  - LIP when loop goes down
    - Loop is stopped and relative addresses reassigned
    - Adapter and I/O errors will be seen
  - Newer hubs allow subloops to avoid the LIP problem when a node is rebooted
  - Cannot mix NT and non NT systems on the same loop

Fibre Technologies

- Fibre Point to Point
  - Simplest configuration
  - Dedicated direct Fibre link from system to the adapter in the disk subsystem
  - Limits scalability of disk subsystem (most have between 16 to 32 max FAs)
  - Provides best performance and error recovery
  - Can have multiple Fibre connections
  - Can be run in parallel for redundancy and performance (load balancing)
  - Special software needed to run in parallel

Fibre Technologies

- FC-SW Fibre Channel Fabric Switch
  - Better performance and redundancy than loops
  - More scalable than point to point
  - Can dynamically add and drop nodes
    - Switch assigns an address at login so no need to walk the loop to assign addresses
  - Non blocking switch technology so multiple I/Os can happen at the same time
  - Can cascade switches to have up to 16 million connections

Fibre Distances

- Short wave laser:
  - 500 meters (50 micron Fibre), 175 meters (62.5 micron Fibre).
- Long wave laser:
  - 10 kilometers (9.5 micron single mode Fibre). (Up to 20 kilometers between E_Ports for some directors).
- Longer distances:
  - Extenders, protocol converters, or Dense Wave Division Multiplexors (DWDM)
  - Selection will depend on the available links between the two locations, distance and budget.

Fibre Classes of Service

- Class 1 – dedicated connection
- Class 2 – connectionless, frames acknowledged
- Class 3 – datagram connectionless, no acks
- Class 4 – similar to class 1 but uses less bandwidth – designed for multimedia apps such as video
- Class 5 – isochronous service, no buffering
- Class 6 – similar to class 1 but uses multicast connections
- Class F – for use by switches communicating through ISLs

Design Points for Fibre

- Can't mix NT and non NT systems on the same loop or FA (Fibre adaptor in the disk subsystem)
- Can't mix different vendors Fibre cards on the same loop or FA (I.e. IBM 6227 or 6228 and an Emulex cannot coexist)
- For small number of systems point to point is most cost effective
- For growing or large number of systems use Fibre Switch
- Many problems with Hub technologies so avoid them
- Microsoft Bug randomly on high numbered LUNs – make sure you install the patch

Fibre Port Types – E_Port

- E_Port
  - Expansion port –used as an interswitch expansion port to connect to the E_Port of another switch, to build a larger switched fabric.
  - Found in Fibre Channel switched fabrics and are used to interconnect the individual switch or routing elements.
  - They are not the source or destination of IUs, but instead function like the F_Ports and FL_Ports to relay the IUs from one switch or routing elements to another.
  - E_Ports can only attach to other E_Ports.

Fibre Port Types – F_Port

- F_Port
  - Fabric port that is not loop capable.
  - Used to connect an N_Port to a switch.
  - Found in Fibre Channel switched fabrics.
  - Not the source or destination of Ius
  - Function only as a “middle-man” to relay the IUs from the sender to the receiver.
  - F_Ports can only be attached to N_Ports.

Fibre Ports Types – FL_Port

- FL_Port
  - Fabric port that is loop capable.
  - Used to connect NL_Ports to the switch in a loop configuration.
  - These ports are just like the F_Ports except that they connect to an FC-AL topology.
  - FL_Ports can only attach to NL_Ports.

Fibre Ports Types – G_Port & Isolated E_Port

- G_Port
  - Generic port that can operate as either an E_Port or an F_Port.
  - A port is defined as a G_Port when it is not yet connected or has not yet assumed a specific function in the fabric.
- Isolated E_Port
  - Port that is online but not operational between switches due to overlapping domain ID or non-identical parameters such as E_D_TOVs.

Fibre Ports Types – L_Port

- L_Port
  - Loop capable fabric port or node.
  - Basic port in a Fibre Channel Arbitrated Loop (FC-AL) topology.
  - If an N_Port is operating on a loop it as an NL_Port.
  - If a fabric port is on a loop it is known as an FL_Port.

Fibre Channel Ports – N_Port

- N_Port
  - **N**ot loop capable.
  - Connects an equipment port to the fabric.
  - Found in Fibre Channel nodes, which are defined to be the source or destination of information units (IU).
  - I/O devices and host systems interconnected in point-to-point or switched topologies use N_Ports for their connection.
  - N_Ports can only attach to other N_Ports or to F_Ports.

Fibre Channel Ports – NL_Port & U_Port

- NL_Port
  - Node port that is loop capable.
  - Connects an equipment port to the fabric in a loop configuration through an FL_Port.
  - Just like the N_Port, except that they connect to a Fibre Channel arbitrated loop (FC-AL) topology.
  - NL_Ports can only attach to other NL_Ports or to FL_Ports.
- U_Port
  - Universal port.
  - Generic switch port that can operate as either an E_Port, F_Port, or FL_Port.
  - A port is defined as a U_Port when it is not connected or has not yet assumed a specific function in the fabric.

Blocking vs Non-Blocking

- Non-blocking
  - Any two pairs of ports can be active and transferring data without blocking the transfer of data from another pair of ports. Each port is allocated a time slice to transfer data, and cut through routing occurs that allows for immediate transfer of data from an input port to an output port if that port is free.
- Blocking
  - Occurs in a fabric design with multiple switches when data from multiple sources must be sent to a single destination port, or when data is required to be sent across an inter-switch link from multiple input ports. Data is blocked, that is to say, buffered in the switch, and sent to the destination port based on the priority set of the data (default priority for data based on

Priority

- Virtual channels give greater priority to F_Port traffic on inter-switch links than data traffic.
- Data is transferred based on buffer credits assigned to ports
- Sending and receiving devices manage the credits so that there is never an overrun of data in the switch.

SSA Technology

- Arbitrated protocol using full duplex bi-directional loop
  - If loop broken one way then data can still be moved the other direction
- Was 80 MB/sec, now 160 MB/sec
- Open ANSI standard X3T10.1/1145D
- Allows all disks and hosts to use the loop at the same time (non-arbitrated)
- All have equal priority
- Fairness mechanism used to protect from takeovers by one device

SSA Technology

- Dual Adapter Loop – both adapters can read and write to all disk arrays
  - Provides reliability and redundancy
- Hot pluggable disks
- 160 MB/sec is only obtained when all 4 pieces of Fibre in the loop are being used
- Each piece of Fibre is really 40 MB/sec
- You need a number of disks in the loop to really obtain the 160MB/sec as the disks are limited to around 22 MB/sec

SSA Technology

- Fast Write
  - Allows writes to go to 32MB nonvolatile cache
  - CE/DE returned immediately
  - Improves response for small records being written and random writes
  - Cannot be used if data mirrored across adapters
- Express Raid Write
  - Turns sequential writes into raid-3 operations generating parity on the fly
  - Avoids the raid-5 penalty
- Sequential Prestage
  - Disks pre-stage data into a disk buffer to improve performance for sequential applications

SSA Technology

- Array Sizes
  - From 2+1P to 15+1P
  - Hot spares can be shared across the loop
- Other
  - 48 drives per loop
  - Up to 8 hosts (nonraid) or 2 hosts (raid-5) per loop
  - 96 drives per adapter
  - SSA Optical extender allows distances up to 10km

SSA Technology

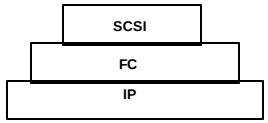
- Spatial Reuse
  - Data routed over one link so multiple transfers can occur as there are up to 4 links, provided the transfers use different links
- Addressing Scheme
  - Network topology consisting of:
    - A Web – a string or loop or collection of strings or loops interconnected by switches
    - Node – disk, controller or other device
  - UID used to identify every node in the SSA web
    - 2 reserved bytes, 3 byte OUI (org. unique id), 3 byte product id
  - UID allows dynamic unobtrusive add/drop of nodes
  - UID protects from device address conflicts common with SCSI

SSA Technology

- Fairness Algorithm
  - Tokens (called SAT tokens)circulated in both directions
  - Each node accumulates a queue of I/O to transmit
  - When node receives the SAT token it is allowed to transmit
  - Then it waits till the link is quiet or it gets another SAT token

iFCP

- Gateway protocol – more complex
  - Cracks FC frame
  - Provides fibre fabric services, e.g.,
    - FLOGI – fabric login
    - SNS – simple name server
    - RSCN – registered state change notification
- Gateway in front of each FC HBA
- Gateway in front of each FC storage device
- Allows use of FC devices on IP network
- Gateways provide fibre fabric function
- This is an IP SAN



new

InfiniBand™ Genesis

- Originated from:
  - Next Generation IO (NGIO)
    - Intel, Microsoft, Sun
  - Future IO
    - IBM, Compaq, HP
  - Consolidated group includes Dell, total 7
- Goal: to obsolete the I/O bus and replace it with a fabric

Circle4 Consulting

new

InfiniBand™

- InfiniBand Trade Association
  - www.infinibandta.org
- High speed, networked interconnect
  - Server Clusters
  - Storage connection
  - Remote DMA
- Board and Chassis interconnect
  - Replaces PCI
- Support for IPV6 semantics
  - Access to MAN/WAN

Circle4 Consulting

new

Physical Protocols and Media

	Parallel S/390	**ESCON S/390**	**SCSI**	**Fibre Channel**	**Gb Ethernet**	**InfiniBand**
Type	Channel	Channel	Channel	Network	Network	Network
Mode	Parallel	Serial	Parallel	Serial	Serial	Serial
ULP				SCSI, IP	SCSI, IP	SCSI, IP
Protocol	Parallel Channel	ESCON	SCSI	FC	Ethernet	IB
HBA	Parallel Channel Card	ESCON Channel Card	SCSI HBA	Fibre HBA	NIC	IB HCA
Medium	Bus and Tag cables, connectors	ESCON Fiber, connectors	Parallel SCSI cable, connectors	Copper or Fiber, GBICs	Fiber, GBICs	Copper or Fiber, connectors
Band width	4 MB/sec	17 MB/sec	320 MB/sec	2 Gb/sec	1 Gb/sec	2.5 Gb/sec

Circle4 Consulting

new

RAID Levels

- Raid-0
  - Disks combined into single volume stripeset
  - Data striped across the disks
  - No redundant data stored
  - Good performance
  - Failure of a disk results in data loss
- Raid-1 or mirroring
  - Every disk mirrored to at least one other disk
  - Full redundancy of data but needs extra disks
  - At least 2 I/Os per random write
  - Faster on reads and slower on writes than a single drive
  - Failures are transparent but no automatic rebuild

Circle4 Consulting

RAID Levels

- Raid-0+1 – striped mirroring
  - Striped mirroring
  - Combines redundancy of raid-1 and performance of raid-0
  - Raid-1 mirror set created and seen as if it was one disk
  - That disk is then turned into a stripe set
  - Full protection against a single disk failure
  - Can lose up to half the disks provided no two members of the same mirror set fail

RAID Levels

- These 3 levels are not in common use today
- Raid-2
  - Intended for use with drives with no built-in error protection
- Raid-3
  - Stripes at a byte level across several drives with parity stored on one drive
  - Requires hardware support for efficient use
- Raid-4
  - Stripes data at a block level across several disks with parity stored on one drive
  - Reads are fast, writes are slower due to writing parity

RAID Levels

- RAID-5
  - Data striped across a set of disks
  - 1 more disk used for parity bits
  - Parity may be striped across the disks also
  - At least 4 I/Os per random write (read/write to data and read/write to parity)
  - Cache and fast-write technologies used to reduce this overhead
  - On failure error message produced and it fails over to the hot spare and rebuilds the lost data from the parity disk
  - If parity disk fails then parity is rebuilt from the info on the data disks
  - Uses hot spare technology

Summary of RAID

- www.uni-mainz.de/~neuffer/scsi/what_is_raid.html
- RAID-0 is fastest and most efficient but no fault tolerance
- RAID-1 is array of choice for performance critical (mostly read), fault tolerant environments.
- RAID-2 seldom used since ECC now embedded in disk drives.
- RAID-3 and RAID-4 offer no advantages over RAID-5 and do not support multiple simultaneous write operations. RAID-3 does not allow multiple I/O operations to be overlapped.
- RAID-5 is the best choice for multi-user environments which are not write performance sensitive.

Raid Disk Q Length

- lsattr –El hdisk0 (scsi attached disk)

```
pvid        000921df3259e6db0000000000000000 Physical volume identifier  False
queue_depth 3                                Queue DEPTH                  False
size_in_mb  9100                             Size in Megabytes            False
```

- Default queue depth is 3
- Raid array is seen as 1 disk
- Set Qlen to #disks*3
- i.e for a 4+1 use 5*3=15

lsattr –El on a Fibre disk

scsi_id	0x614313	SCSI ID	True
lun_id	0x15000000000000	LUN ID	True
location		Location	True
ww_name	0x50060482c094e81d	World Wide Name	True
pvid	000921dfd0ec5beb0000000000000000 Physical Volume ID		False
q_type	simple	Queue TYPE	True
queue_depth	8	Queue DEPTH	True
q_err	no	Use QERR bit	True
reserve_lock	yes	Reserve Device on open	True
clr_q	yes	Clear Queue (RS/6000)	True
start_timeout	180	START UNIT time out	True
rw_timeout	40	READ/WRITE time out	True

Other I/O performance notes

- Mirroring of disks (software versus hardware)
- Mirror Write Consistency
- Write verify
- Logical volume scheduling policy
  - Parallel or sequential
- Buffering and two way searches
- I/O Pacing and Async I/O

Intra and Inter Policies

- Intra Policy
  - How data is laid out on the disk
  - Outer edge, outer middle, center, inner middle, inner edge
- Inter Policy
  - How data is laid out between/across disks

I/O Pacing for a mksysb

```
# Turn on I/O pacing
chdev -l sys0 -a maxpout='17' -a minpout='12'
#run the mksysb to tape
mkszfile -f" && mksysb '/dev/rmt0'
# Turn off I/O pacing
chdev -l sys0 -a minpout='0'
chdev -l sys0 -a maxpout='0'
```

Async I/O on UNIX

- pstat –a | grep aios | wc –l
- Show as kprocs
- Needs reboot to change
- Min is 1, max is 10
- Max should be <=80
- Set to 10x disks accessed concurrently using AIO
- Used by Oracle, DB2 has own AIO

Network Storage Performance

- Considerations
- Link speed
  - GB Ethernet
  - Shared vs. dedicated
- TCP load on server
  - IPSEC
  - Packet assembly/disassembly
  - Checksum processing
- NIC performance
- Frame size
  - Jumbo frames

new

Jumbo Frames

- Normal Ethernet IP frame size is 1500 bytes
- Fragmentation of large SCSI blocks
- Jumbo frames 9000 bytes
  - Good for NFS – 8192 bytes blocks
- Reduces packet overhead
- All components (NIC, switch, router) must support jumbo
- Results:
  - Oak Ridge National Laboratory (ORNL)
  - Dedicated GB Ethernet, jumbo frames
  - 93 MB/sec (93% of link speed) sustained over 33 hours

new

Issues with Jumbo Frames & Network

- Mtu size default is 1500 for a network
- If Mtus don't match then they negotiate using Mssdefault which is normally 512
- Watch for window sizes – do a search on Nagle
- Tweak network tuning parameters
  - Thewall
  - Tcp send and receive buffers
  - Rfc1323

new

References

- SG24-6143 IBM SAN Survival Guide
- www.uni-mainz.de/~neuffer/scsi/what_is_raid.html
- Various Web Searches
- Vendor Web sites

Questions?