DATA STORAGE

Requirements

- Host-level redundancy
- Fault tolerance
- Disaster tolerance
- High speed access
- Low latency
- Interoperability
- High scalability
- Ease of management

Storage access modes

- DAS - direct attached storage
- NAS - network attached storage
- SAN - storage area network
- iSCSI - internet SCSI
- FCIP - Fibre channel over IP

DAS vs SAN

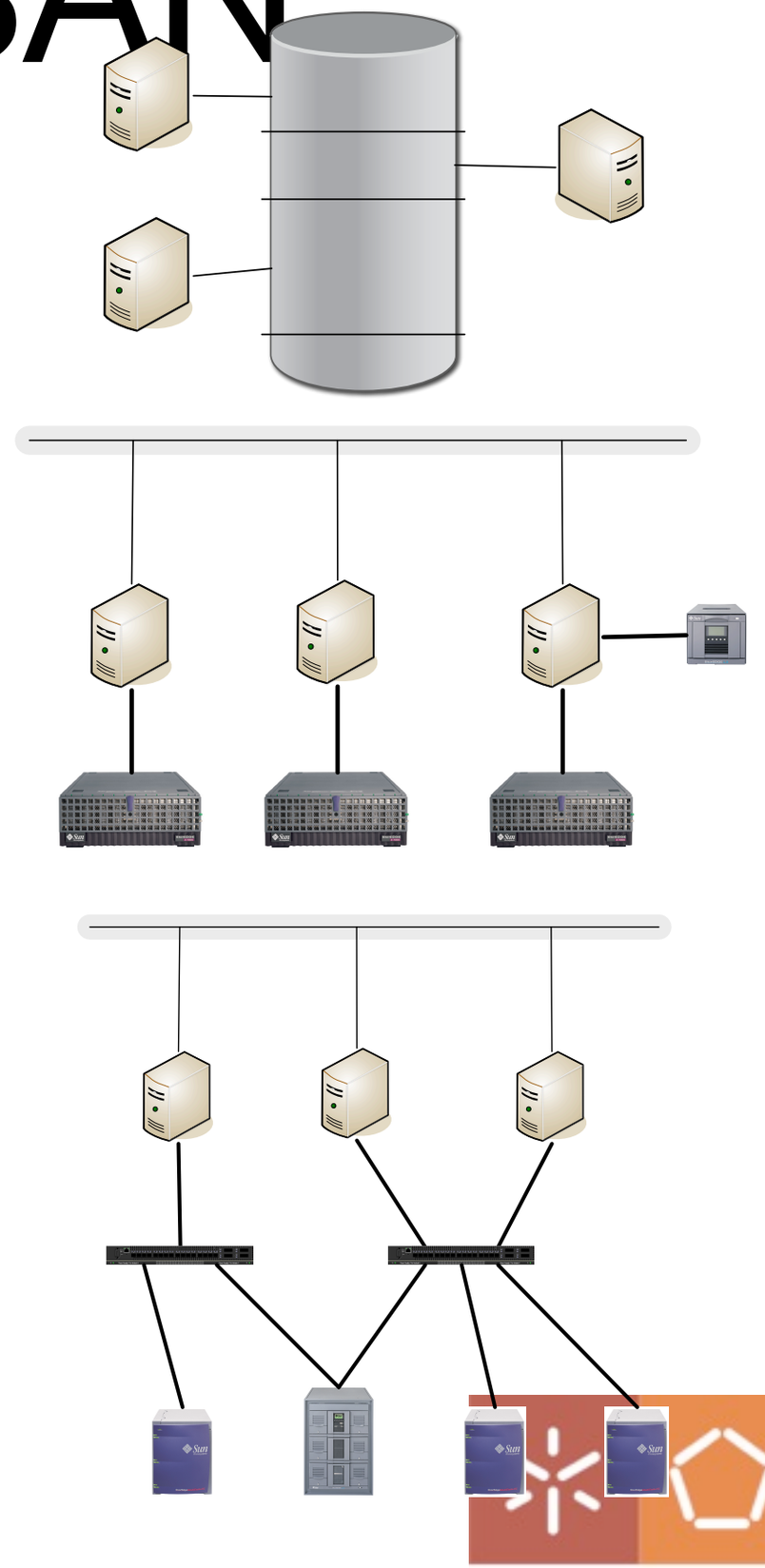
Consideration	DAS	SAN
Scalability	Complex and costly to add/remove storage devices	Servers & storage easily added/removed. easy to reallocate storage
Use of host resources	Host resources used for applications, I/O, data transfers, backups	Storage & I/O functions done by SAN
Availability of data	Storage is cabled to 1 or 2 servers.	Multiple serves can access the data. Several data paths for each server
Storage consolidation	None. each server has its dedicated storage	there can be a large, shared storage, portions of which can be allocated to different servers
Cost of storage adm	Expensive & time consuming. Each storage must be managed individually	Servers & storage are grouped into one network. Simplifies adm.
Distance	Limited to cable length (SCSI3 is 25m)	Single mode fiber 10 Km.
Reliability	Copper SCSI suffers EMI	Fiber is less impacted by EMI
Number of devices	SCSI has 7 devices per target	126 nodes for FC-AL, and up to 16M for SAN fabrics

NAS vs SAN

Criteria	NAS	SAN
Network	TCP/IP	Dedicated fibre channel network
Design	Defines a product or appliance	Defines an architecture
Protocols	HTTP, FTP, CIFS, NFS	Encapsulated SCSI
Performance	High latency because of TCP/IP overhead	Thin protocol, low latency
Connection	Files shared via an indirect connection	Server accesses a volume via a direct connection
Level of access & heterogenous data sharing	File level access to UNIX and Windows servers and identifies files by filenames. NAS handles data security, user authentication and file locking	provides block level access to the servers connected to the SAN. Data transferred as raw disk blocks
File system maintenance	NAS head unit manages the file systems	File systems are created and managed by servers
Compatibility among platforms	NAS allows simultaneous sharing of files between disparate operating systems and platforms	File sharing is dependent on server operating system. There is little or no cross-platform compatibility

Benefits of SAN

- Consolidation of storage resources
- Concurrent access by multiple hosts
- Reduced TCO
- LAN-free and Server-free data transfers
- Max distance between nodes and use in DR
- High performance
- Scalability and flexibility
- Server clustering



SAN Topologies

- Point-to-point
- Fibre channel arbitrated loop (FC-AL)
- Switched fabric

SAN Topologies

Characteristic	Point-to-Point	FC-AL	Switched Fabric
Number of devices	Direct link between 2 nodes	Max 127 nodes per loop	16M
Required components	Fiber cables, HBAs	Hub connected to server HBAs and storage devices	Multiple interconnected switches, cables and HBAs
Benefits	Simple to set up	Low cost	Any-to-any communication between nodes. Throughput increases with the number of nodes
Disadvantages	Storage accessible by only 1 server. No storage sharing	Bandwidth is shared. Decreases with the number of nodes. loop breaks on devices reboot	A switch is expensive. Most SAN hardware is proprietary.

Switch fabric topology

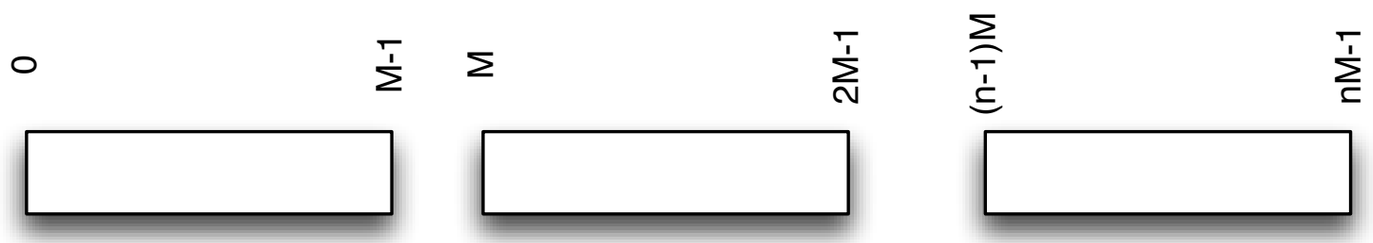
- Star fabrics
- Cascaded fabrics
- Ring fabrics
- Mesh fabrics
- Tree fabrics

RAID

- Standard for arranging groups of disks
- Addresses:
  - Capacity
  - Redundancy
  - Performance
- Different RAID levels balance those factors and price

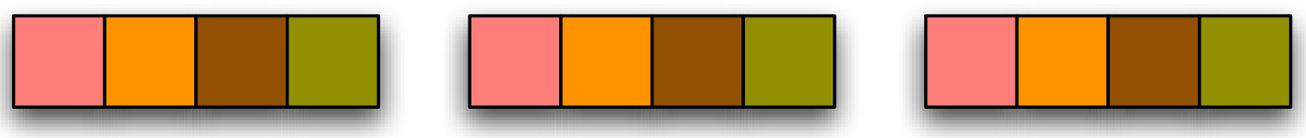
Linear

- Linear concatenation of disks:
- N disks with M blocks
- We get a disk with N*M blocks
- sectors 0 to M-1 are on disk 0
- sectors M to 2*M-1 are on disk 1
- ...
- Addresses only disk capacity
- vulnerable to the failure of a single disk



RAID 0 (Striping)

- N disks with M blocks
- We get a disk with N*M blocks
- Each stripe with F blocks
- sectors 0 to F-1 on disk 0
- sectors F to 2*F-1 on disk 1
- ...
- Addresses performance and capacity
- Extremely vulnerable to the failure of a single disk

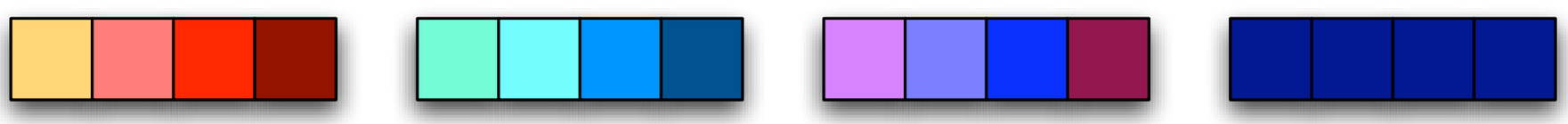


RAID 1 (Mirroring)

- 2*N disks with M blocks
- We get a disk with N*M blocks
- The sector ***i*** is read from any disks
- Every write on sector ***i*** must be duplicated on every disk
- Addresses only redundancy

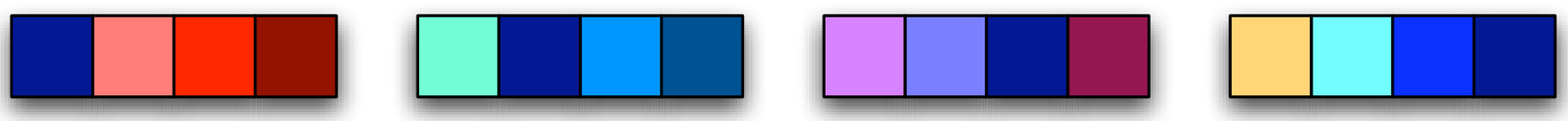
RAID 4

- Reserves a disk for parity (XOR):
- N disks with M blocks
- We get a disk with (N-1)*M blocks
- addressing similar to stripping
- Each writes forces the recalculation of the sector in the parity disk
- Addresses redundancy with little impact on capacity
- Tolerates the failure of a disk
- The parity disk limits the performance



RAID 5

- Parity in all disks:
- N disks with M blocks
- We get a disk with (N-1)*M blocks
- addressing similar to stripping
- Each stripe has a different parity disk
- Better redundancy with a small impact on capacity
- Tolerates the failure of a disk

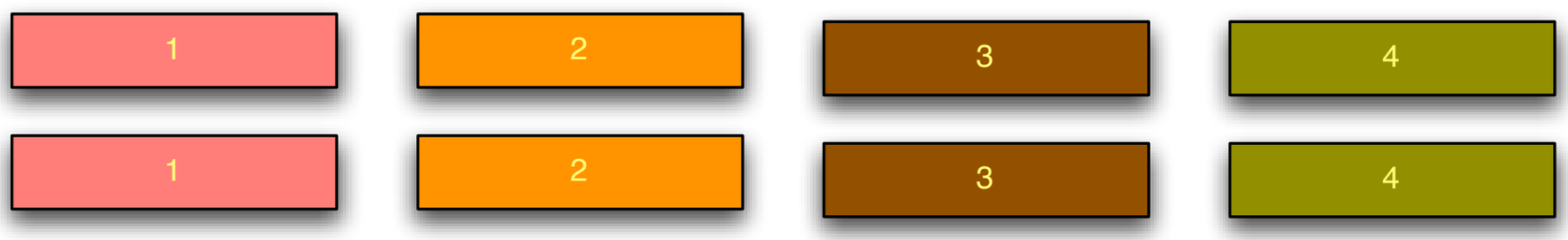


RAID 0+1

- Striping before mirror:
- Stripe the disks in two groups
- Mirrors the two groups
- The failure of a disk shuts down a stripe
- The failure of a disk in the other stripe looses data

RAID 1+0 (10)

- Mirror before striping:
- Mirror the disks
- Stripe the mirrored disks
- The failure of a disk does not shut down the stripe
- Two failures of the mirrored disks loose data



Other RAID levels

- RAID 3: Byte striping
- RAID 6: Two parity blocks (tolerates 2 failures)
- RAID 15 (1+5) and 51 (5+1): Tolerates multiple failures
- ...

RAID Level	Advantage	Disadvantage	When to use
RAID-0 (Striping)	No reduction. All disk space is usable; increases performance	No protection from disk failure	Need high speed access to storage, but do not need data protection
RAID- 1 (Mirroring)	Perf improve for reads; protection from disk failures; No write perf degradation with 1 disk failure	50% loss of disk space	Use RAID-1 when data must be protected from disk failures
RAID 0+1 (stripped and mirrored)	Perf increases; protection from disk failures	50% loss of disk space	For apps requiring high performance and data protection
RAID 1+0 (mirrored and stripped)	Perf increases; high degree of protection from disk failures	50% loss of disk space	For apps where data protection is most critical
RAID-4 (Striping with single parity disk)	Disk redundancy; read perf. enhancement because of striping	The single parity in volume is perf bottleneck	for high read-only services
RAID-5 (Striping with distributed parity)	Disk redundancy; good read perf.	If app has high % writes, decreases I/O rates; if a disk fails, perf degrades in recovery mode	Applications are mostly read intensive; I/O requests are random as in databases; data protection is needed but cannot sacrifice 50% of disks

Which RAID?

- Linear inadequate for the majority of the applications
- RAID 0 extremely vulnerable to failures
- RAID 1 does not take advantage of disks parallelism
- RAID 4 and 5 have limited performance:
  - With applications with large amounts of written data during recovery

Which RAID?

- Applications with high read share (>3/4)
  - RAID 5:
    - File servers
    - Web servers
- Applications with high share of writes
  RAID 1+0:
    - OLTP databases

Partições

- Divisão (estática) de um dispositivo físico em diversos dispositivos lógicos
- Uma partição ocupa sectors contíguos
- Utilidade:
  - diferentes sistemas operativos
  - alocação de espaço para diferentes finalidades dentro do mesmo sistema operativo (quotas, users, log, sistema,swap,...)
  - facilidade de manutenção (backups, actualizações, ...)

Causas de indisponibilidade

- Modificar o tamanho de partições:
  - paragem das aplicações
  - umount das partições em causa
  - na melhor das hipóteses, modificação directa do tamanho da partição (se existir espaço contíguo)
  - caso contrário, cópia de segurança seguida de reinicialização das partições e reposição dos dados
  - mount das partições
  - reinício das aplicações

Causas de indisponibilidade

- Mover uma partição para um novo disco:
  - paragem das aplicações
  - umount das partições em causa
  - cópia dos dados para a nova localização
  - mount das partições
  - reconfiguração do sistema para localizar os dados
  - reinício das aplicações

Causas de indisponibilidade

- Cópias de segurança coerentes:
- paragem das aplicações
- cópias de segurança
- reinício das aplicações

Gestão de volumes lógicos

- Constrói dispositivos lógicos com funcionalidade acrescida:
  - isolamento
  - redundância
  - virtualização
  - Funcionalidade com semelhanças à gestão de memória virtual!

Arquitectura

- Volumes físicos (PV) são agregados num grupo de volumes (VG):
  - o espaço disponível num VG pode ser usado independentemente do PV originário
- Camada que aloca e de um VG para volumes lógicos (LV):
  - a alocação pode obedecer a diferentes critérios
- Podemos fazer a correspondência entre:
  - VG e discos
  - LV e partição

Capacidade variável

- O espaço alocado a um LV não precisa de ser fisicamente contíguo
- Um LV pode estar espalhado por vários discos físicos
- Modificar o tamanho de uma partição é sempre possível, desde que haja espaço no VG
- É possível acrescentar novos discos a um VG aumentando o espaço disponível para LVs

Migração

- Havendo necessidade de substituir um PV:
  - acrescenta-se ao VG um novo PV with tamanho suficiente
  - mudam-se os sectors do PV original, sendo ocupado o novo PV
  - A troca é invisível para os utilizadores
  - Não é necessário parar as aplicações nem fazer umount

Snapshot

- É possível congelar um LV num determinado instante no tempo, tornando-o num novo LV
- Quaisquer modificações ao LV original alocam novo espaço no VG deixando intacto o LV cópia
- O snapshop pode ser usado para efectuar cópias de segurança

Coordenação com FS

- A gestão de volumes lógicos pode ser coordenada com sistemas de ficheiros:
  - modificar o tamanho do LV sem fazer umount
  - fazer um snapshop coerente sem fazer umount
  - A coordenação tem que ser feita modificando o código do próprio sistema de ficheiros

Cuidado!

- A falha de um PV pode inutilizar os dados contidos em diversos LVs!
- A utilização de VGs com vários discos só é aconselhada se combinada com RAID
- Com RAID em hardware, usa-se a gestão lógica por cima de RAID
- Com RAID em software, são possíveis ambas as combinações

Gestão lógica e RAID

- Gestão lógica antes de RAID:
  - LVs de um mesmo VG não proporcionam falha independente
  - RAID 1, 4 e 5 necessitam de LVs obtidos de VGs distintos
- RAID antes de gestão lógica é a combinação recomendada
- Por vezes é possível obter funcionalidade RAID (linear, 0 e 1) com a gestão lógica

Gestão lógica em Linux

- LVM faz parte da versão 2.4:
  - parecido com LVM da HP
- EVMS 1.x é disponibilizado pela IBM:
  - integra RAID
  - compatível com volumes AIX, OS/2, ...
- Device Mapper no kernel 2.6:
  - LVM2
- EVMS 2.x

Device Mapper

- Conjunto mínimo de funcionalidade oferecida pelo kernel para suportar dispositivos lógicos:
  - criar um dispositivo lógico
  - indicar para cada sector do dispositivo lógico, quais o dispositivo fisico correspondente
- Grande semelhança com tabelas de páginas na gestão de memória virtual!

Regras

- Não é necessário enumerar cada um dos sectores, pois isso consumiria demasiada memória
- Usa-se uma regra para cada conjunto de sectores:
  - Linear
  - Stripping, mirror
  - Disperso, multi-path, erro
- Podem ser combinadas quaisquer regras com quaisquer dispositivos

Gestores lógicos

- Alguns sectores de cada disco são utilizados para guardar informação que descreve os volumes
- Aplicação em modo utilizador:
  - recolhe essa informação
    - inicializa as tabelas necessárias no Device Mapper para disponibilizar os volumes lógicos
- As ferramentas para manipular volumes lógicos podem integrar também a manipulação de RAID e partições tradicionais

Volumes partilhados

- A solução sempre possível consiste em activar um VG apenas em um host
- Em caso de falha:
  - o VG deve ser desactivado no host original (por exemplo, usando STOMITH)
  - o VG pode então ser activado de novo em outro host
- As operações de gestão têm que ser efectuadas a partir do host que está activo