

Network Storage and Cluster File Systems

Jeff Chase
CPS 212, Fall 2000

DUKE University

“NAS vs. SAN”

In the commercial sector there is a raging debate today about “NAS vs. SAN”.

- Network-Attached Storage has been the dominant approach to shared network storage since NFS.
NAS == NFS or CIFS: named files over Ethernet/Internet.
Network Appliance filers dominate, EMC coming on strong
- Proponents of FibreChannel SANs market them as a fundamentally faster way to access shared storage.
no “indirection through a file server” (“SAD”)
lower overhead on clients
network is better/faster (if not cheaper) and dedicated/trusted
Brocade, HP, Emulex are some big players.

DUKE University

Network Block Storage

One approach to scalable storage is to attach raw block storage (“disks”) to a network.

- dedicated Storage Area Network or general-purpose network
FibreChannel (FC) vs. Ethernet....IETF working group is currently defining SCSI-over-IP (iSCSI).
- Each network-attached “disk” may aggregate multiple disks in a box, with basic volume management.
e.g., EMC, others sell RAID boxes attached through FC or SCSI.
- Access control is a key issue if the network is untrusted.
CMU NASD: Network-Attached Secure Disks suggests cryptographic capabilities to protect logical block spaces (“objects”) on disks.

DUKE University

NAS vs. SAN: Cutting through the BS

- FibreChannel is a high-end technology incorporating NIC enhancements to reduce host overhead...
...but bogged down in interoperability problems.
- Ethernet is getting faster faster than FibreChannel.
gigabit, 10-gigabit, + smarter NICs, + smarter/faster switches
- Future battleground is Ethernet vs. Infiniband.
- The choice of network is fundamentally orthogonal to storage service design.
Well, almost: flow control, RDMA, user-level access (DAFS/VI)
- The fundamental questions are really about *abstractions*.
shared block volume (“SAN”) vs. shared file volume (NAS) vs. private disks (shared-nothing)

DUKE University

Network Storage Volumes

A *volume manager* can knit multiple network disks (or partitions) into a unified logical block space (*virtual disk*).

- e.g., Petal, + Veritas and other companies sell volume managers and related “SAN management” products.
- abstraction: OS addresses storage by $\langle \text{volume}, \text{sector} \rangle$.
Petal: access through souped-up device driver
- shared access with scalable bandwidth and capacity
- volume-based administrative tools
backup, volume replication, remote sharing
- volume manager may incorporate *ility features
e.g., Petal is mirrored

DUKE University

Storage Abstractions

- relational database (IBM and Oracle)
tables, transactions, query language
- file system
hierarchical name space with ACLs
- block storage/volumes
SAN, Petal, RAID-in-a-box (e.g., EMC)
- object storage
object == file, with a flat name space: NASD, DDS
- persistent objects
pointer structures, requires transactions: OODB, ObjectStore

DUKE University

Storage Architecture

Any of these abstractions can be built using any, some, or all of the others.

Use the "right" abstraction for your application.

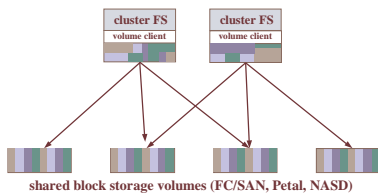
The fundamental questions are:

- What is the best way to build the abstraction you want?
division of function between device, network, server, and client
- What level of the system should implement the features and properties you want?

How does Frangipani answer them?

DUKE

Cluster File Systems

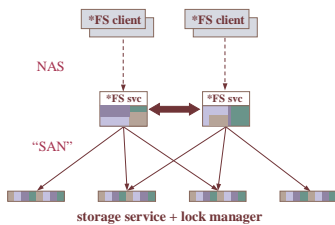


xFS [Dahlin95]
Petal/Frangipani [Lee/Thekkath]
GFS
Veritas cluster file server
EMC Celerra

issues
trust
compatibility with NAS protocols
sharing, coordination, recovery

DUKE

Sharing and Coordination



block allocation and layout
locking/leases, granularity
shared access
separate lock service
logging and recovery
network partitions
reconfiguration

What does Frangipani need from Petal?
How does Petal contribute to F's *ility?
Could we build Frangipani without Petal?

DUKE