# Distributed File Systems

## An Overview

Anant Narayanan

Cluster and Grid Computing
Vrije Universiteit

11 March 2009

# Outline

# Definition

- Allows access to files located on a remote host
  - In a transparent manner, as though the client is actually working on the host
- Typically, clients do not have access to the underlying block storage
  - They interact over the network using a protocol

# Why do we need them?

- Distributed applications usually require a common data store
- Eases ability to keep data consistent
- Access control is possible both on the server and client
  - Depending on how the protocol is designed
- Allows for implementation of
  - Replication
  - Fault tolerance

# Block Oriented

- “Usual” meaning of a file system
- Deal with storing data on a block basis
- Most distributed file systems are based on this at the lowest level
- Examples: ext3, NTFS, HFS+

# Record Oriented

- Were used on Mainframes and Minicomputers
- Fetch and put whole records, seek to boundaries
- Have a lot in common with today's databases
- Examples: Files-11, Virtual Storage Access Method (VSAM)

# Object Oriented

- Splits file metadata from file data
- File data is further split into objects
- Objects stored on object storage servers
- May or may not have a block oriented FS at the lowest layer
- Examples: Lustre, XtreemFS

# High Availability

- Replication
- Parallel Striping
- Examples: Brtfs, Coda, GlusterFS

# Clusters

- Shared disk systems (GFS)
- Distributed disk systems (Lustre)

- Typically used on local networks
- Fast network access

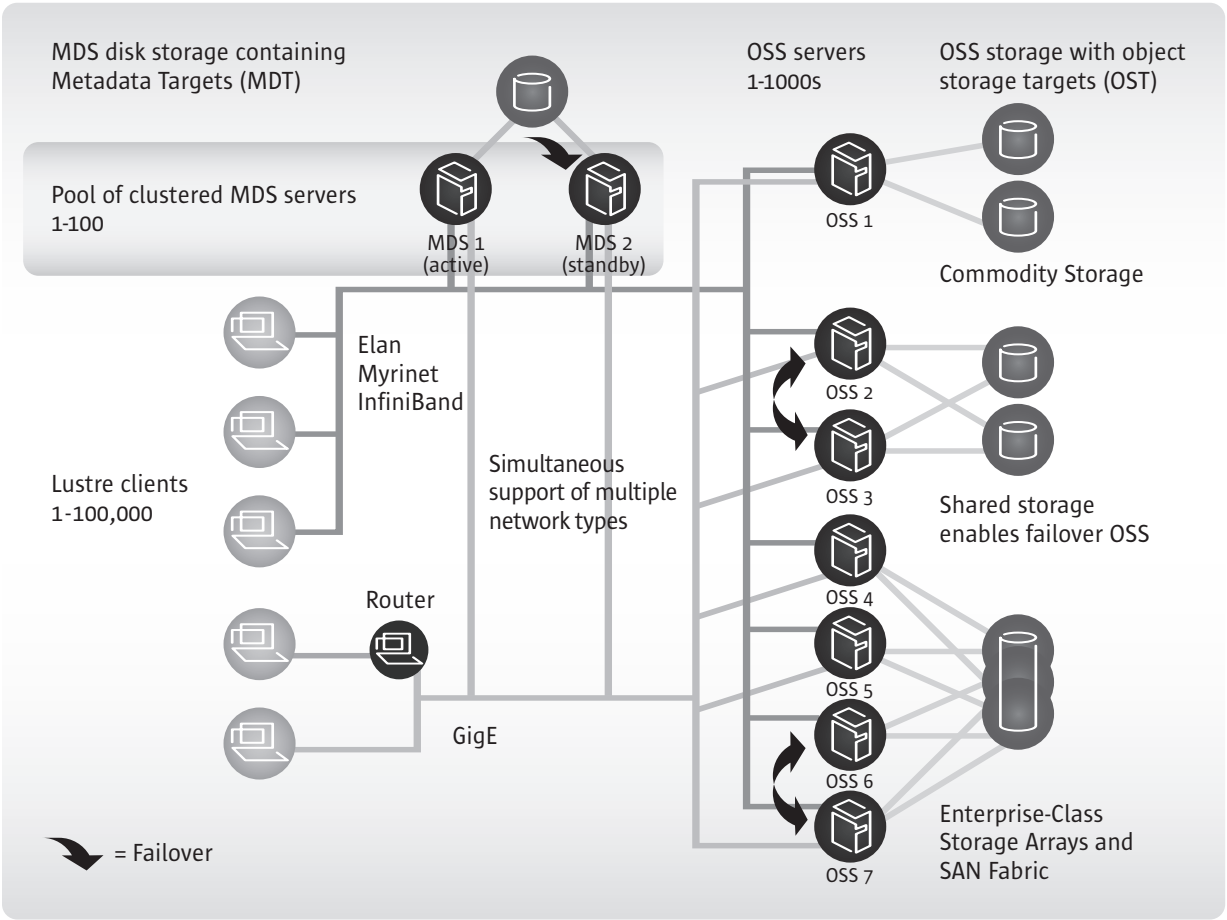
# Grids or Clouds

- Dynamic nature
- Deal with heterogeneity
- Deal with VOs (Grids) and SLAs (Clouds)
- Examples: XtreemFS, Dynamo

# Linux + Cluster

- Distributed, parallel, fault tolerant, object based file system
- Tens of thousands of nodes
- Petabytes of storage capacity
- Hundreds of Gigabytes / second of throughput
- Without compromising on speed or security
- Small workgroup clusters, to large-scale, multi-site clusters, to super-computers

# Layout


MDS disk storage containing
Metadata Targets (MDT)
Pool of clustered MDS servers
1-100
MDS 1
(active)
MDS 2
(standby)
OSS servers
1-1000s
OSS storage with object
storage targets (OST)
OSS 1
Commodity Storage
OSS 2
OSS 3
Shared storage
enables failover OSS
OSS 4
OSS 5
OSS 6
OSS 7
Enterprise-Class
Storage Arrays and
SAN Fabric
Elan
Myrinet
InfiniBand
Lustre clients
1-100,000
Simultaneous
support of multiple
network types
Router
GigE
= Failover

# Components

- File system clients: used to access the file system
- Object storage servers (OSS): provide file I/O service, deals with block storage
- Metadata servers (MDS): manage the names and directories in the file system, deals with authentication

# Characteristics

| | Typical number of systems | Performance | Required attached storage | Desirable hardware characteristics |
|---|---|---|---|---|
| Clients | 1–100,000 | 1 GB/sec I/O, 1000 metadata ops | None | None |
| OSS | 1–1000 | 500 MB/sec — 2.5 GB/sec | File system capacity/OSS count | Good bus bandwidth |
| MDS | 2 (in the future 2–100) | 3000–15,000 metadata ops/sec (operations) | 1–2% of file system capacity | Adequate CPU power, plenty of memory |

# Heterogeneous?

- MDS and OSS may store actual data on ext3 or ZFS block file systems
- Infiniband, TCP/IP over Ethernet and Myrinet are supported network types
- Multiple CPU architectures: x86, x86_64, PPC
- Requires patched Linux kernel

# Setup

- MDS

```
mkfs.lustre -mdt -mgs -fsname=large-fs
/dev/sdamount -t lustre /dev/sda /mnt/mdt
```

- OSS1

```
mkfs.lustre -ost -fsname=large-fs
-mgsnode=mds@tcp0 /dev/sdb mount -t lustre
/dev/sdb /mnt/ost1
```

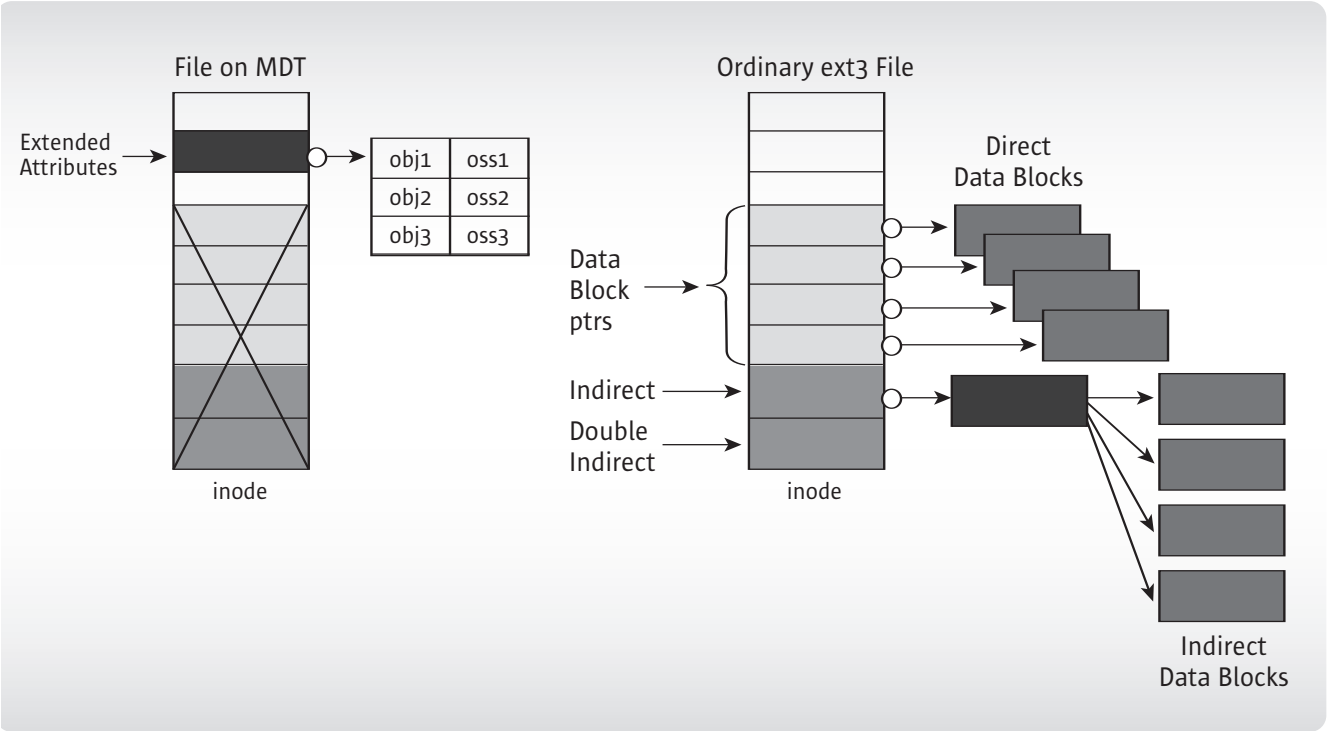
- Client

```
mount -t lustre mds.your.org:/large-fs
/mnt/lustre-client
```

# Networking

- *LNET* abstracts over multiple supported networks
- Provides the communication infrastructure required by Lustre
- Takes care of abstracting over fail-over servers, load balancing
- Provides support for Remote Direct Memory Access (RDMA)
- Provides an end-to-end throughput of 100MB per sec on Gigabit Ethernet networks
- Upto 1.5GB per sec on Infiniband

# Where are the files?



# Striping and Replication

- One object per MDS inode implies “unstriped” data
- Multiple objects per MDS inode implies that the file has been split, similar to RAID 0

- These stripes may be duplicated across several OSS
- Provides fault tolerance and high availability

# Conclusion

- Reliable, Scalable and Performant filesystem
- Open Architecture and Protocols

BUT

- Does not handle dynamic addition / removal of servers
- Does not provide the kind of access control and security that a VO might need

# What next?

- Other distributed file systems and implementation details
- Grids (XtreemFS), Clouds (Dynamo)
- Questions?

# References

You may be required to register on the Sun Website to access these documents!

Datasheet:

`http://www.sun.com/software/products/lustre/datasheet.pdf`

Scalable Cluster Filesystem Whitepaper:

`http://www.sun.com/offers/docs/LustreFileSystem.pdf`

LNET:

`http://www.sun.com/offers/docs/lustre_networking.pdf`

Lustre Documentation Index:

`http://manual.lustre.org/index.php?title=Main_Page`

Lustre Publications Index:

`http://wiki.lustre.org/index.php?title=Lustre_Publications`