Comparison of different distributed file systems for use with Samba/CTDB

@SambaXP'09

Henning Henkel

April 23, 2009

Agenda

Agenda

Agenda

Agenda

Agenda

Introduction

- Diploma study in Computer Networking at the Furtwangen University (HFU) for applied science
- Diploma thesis at the science + computing ag in Tübingen
  Supervising tutors:
  → Prof. Dr. Christoph Reich (Furtwangen University)
  → Dipl.-Phys. Daniel Kobras (science + computing ag)

What were the goals of the diploma thesis?

In the context of the diploma thesis was tested . . .

- . . . which features should be provided by a distributed file system to use it with Samba/CTDB
- . . . what the differences between IBM's GPFS, RedHat's GFS and Sun's Lustre are when used with Samba/CTDB

Not tested in the context of the diploma thesis are . . .

- . . . the fencing mechanisms provided by Samba/CTDB
- . . . the cluster management provided by Samba/CTDB

What is pCIFS?

In the Samba/CTDB context

- Parallel CIFS servers as a CTDB layer between CIFS Clients and distributed file systems
- One Client is connected to only one CIFS Server.
- There is no need for modifications on the client side.

What is pCIFS?

In the lustre context

- A set of parallel CIFS servers provied access to the lustre file system.
- One client can connect to multiple CIFS Servers.
  - Advantage: A single client might reach the maximum throughput.
- But there are also major disadvantages:
  - There is the need for a special CIFS client software.
  - The client software is only for one specially picked file system.
- I'm not aware of a product ready implementation.

What is a distributed file system?

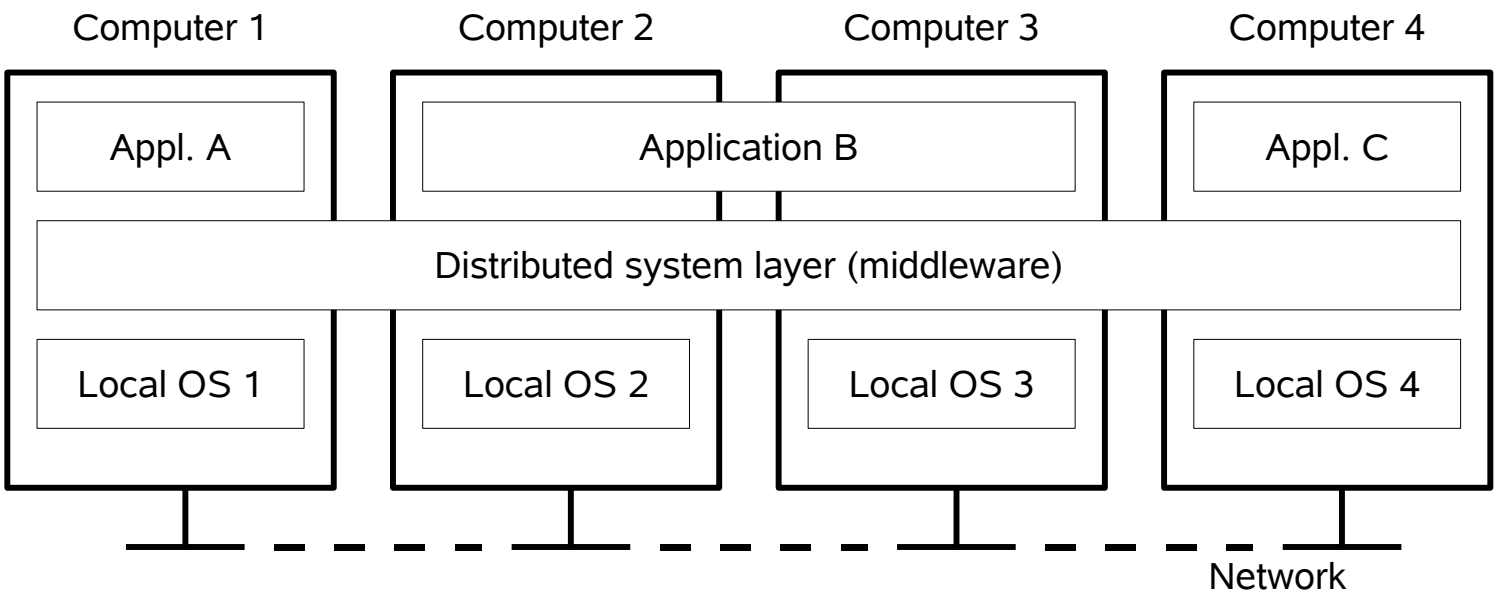
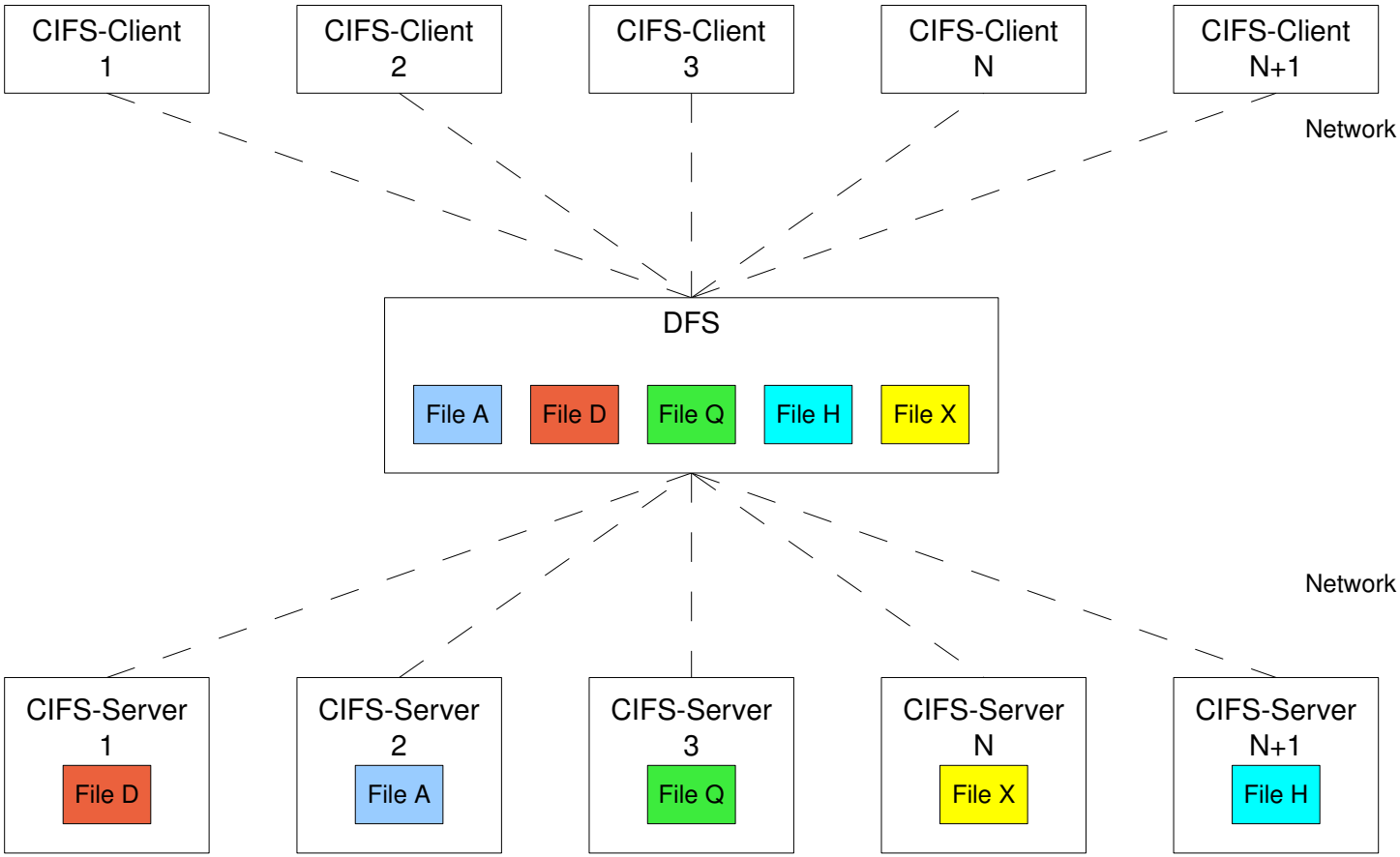


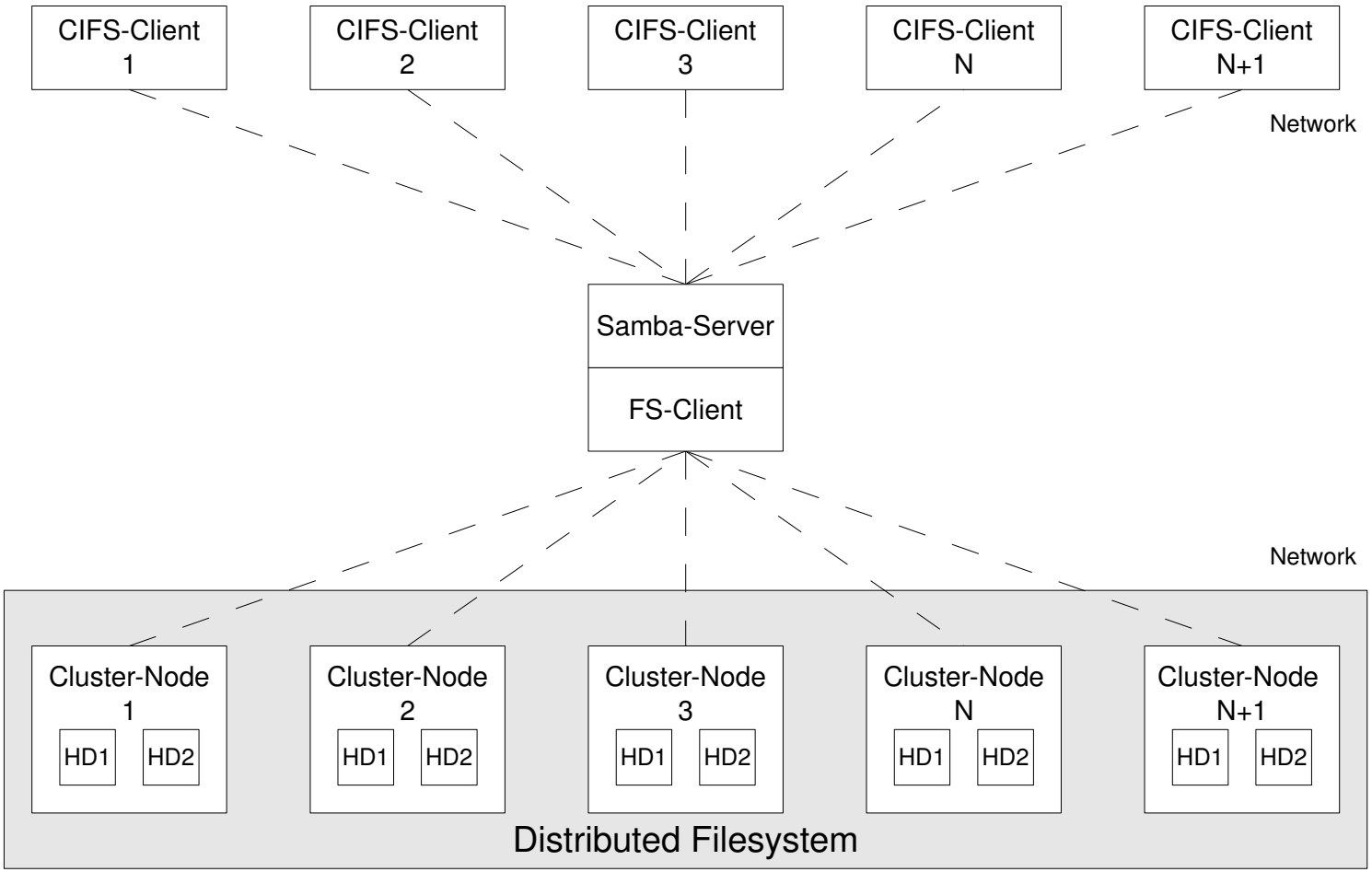
Figure: Distributed file systems are a middelware

Source: Distributed Systems - Principles and Paradigms, Tanenbaum 2007

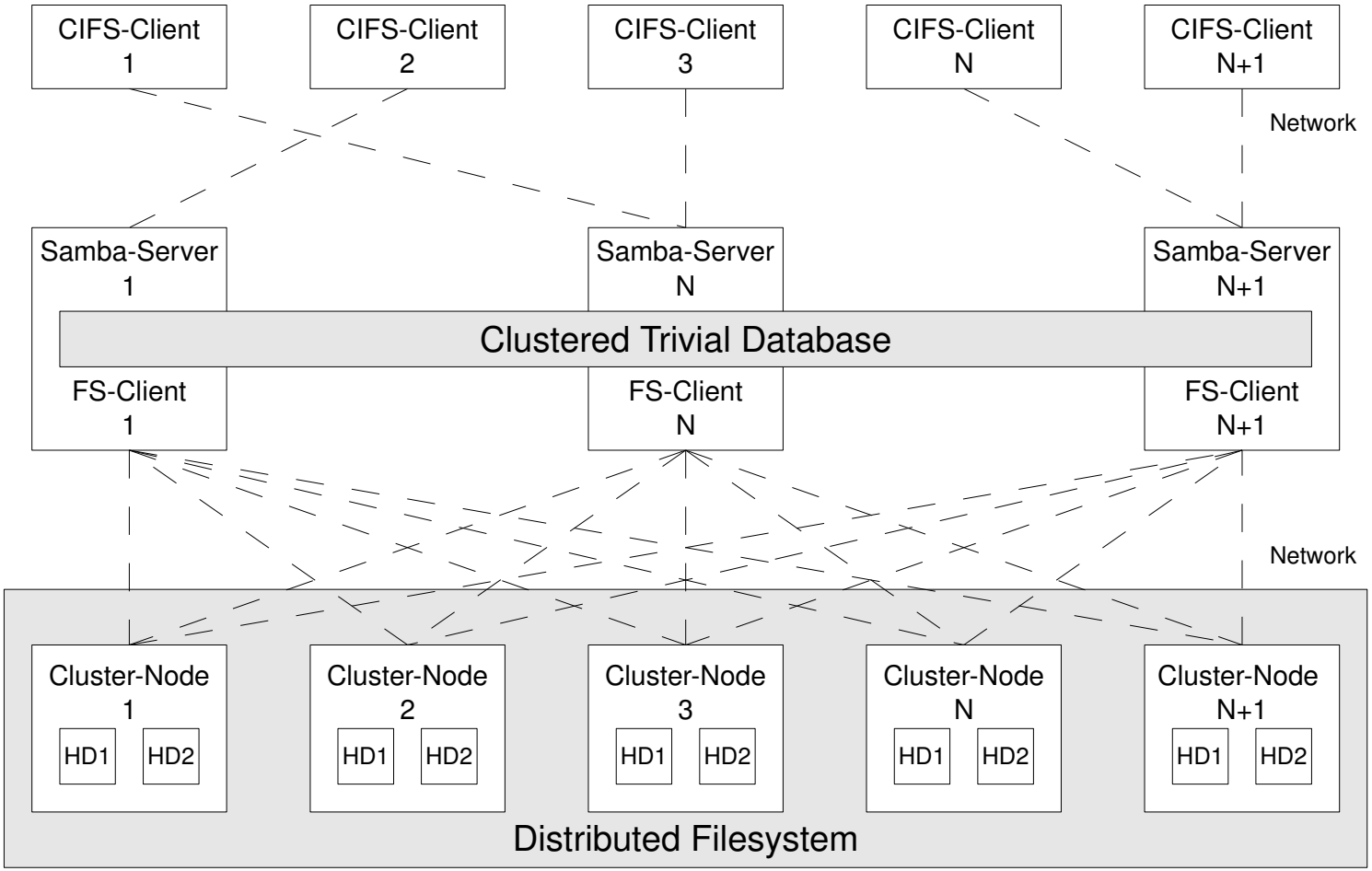
Microsoft's DFS



Access without CTDB



Access with CTDB



The test candidates

- (FrauenhoferFS (FhGFS))
- IBM's General Parallel File System (GPFS)
- Sun's Lustre
- Red Hat's Global File System (GFS)

FhGFS

- Project at the Frauenhofer Instituts für Techno- und Wirtschaftsmathematik (ITWM), Competence Center for High Performance Computing.
- It is a quite young distributed file system.
- Easy to install and configure.
- According to the specifcations of the producer it scales as good as Sun's Lustre and reaches a higher throughput.

FhGFS

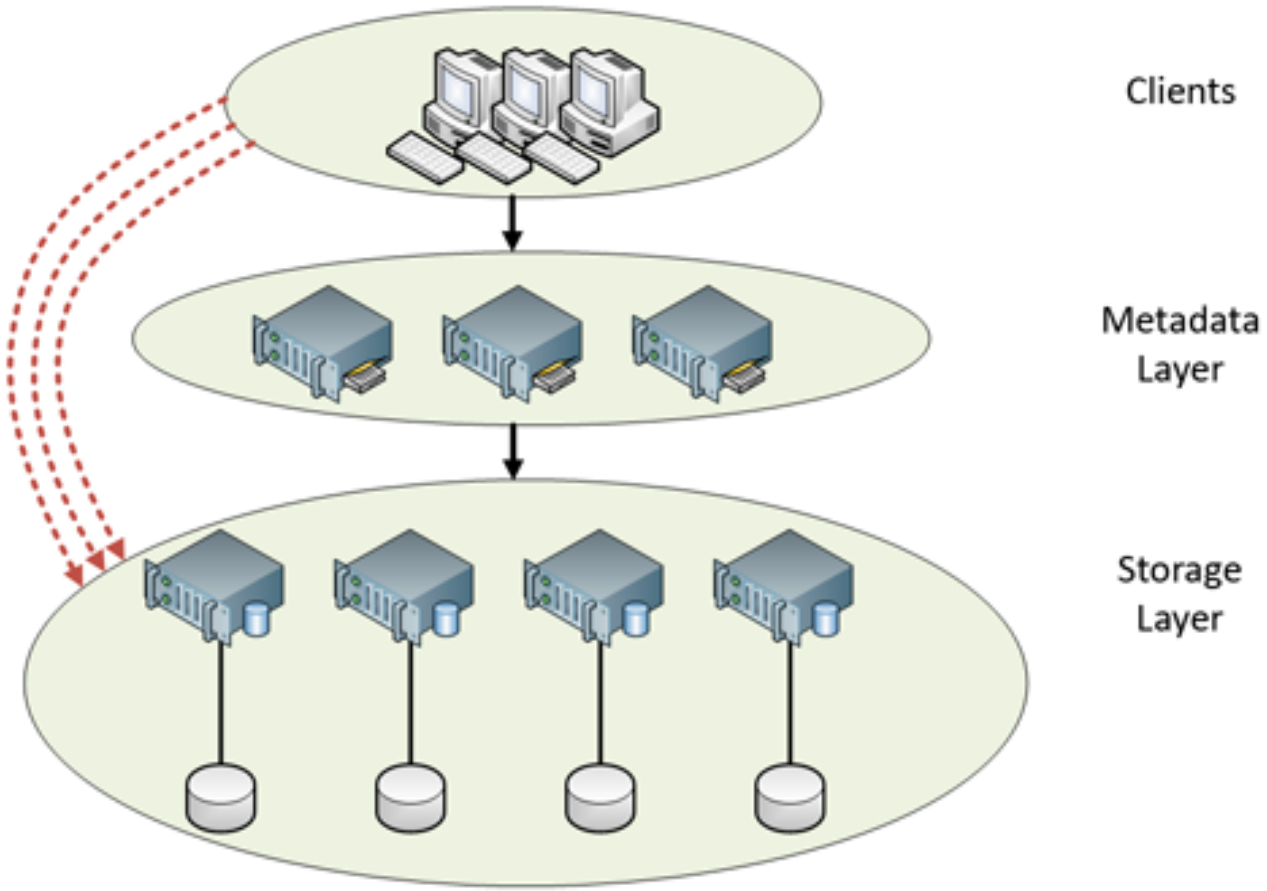


Figure: The FhGFS Architecture

Source: FraunhoferFS User Guide, online

GPFS

- available since 1998 for AIX
- file management infrastructure
- proprietary software
- most tested with Samba/CTDB by the Samba team

GPFS

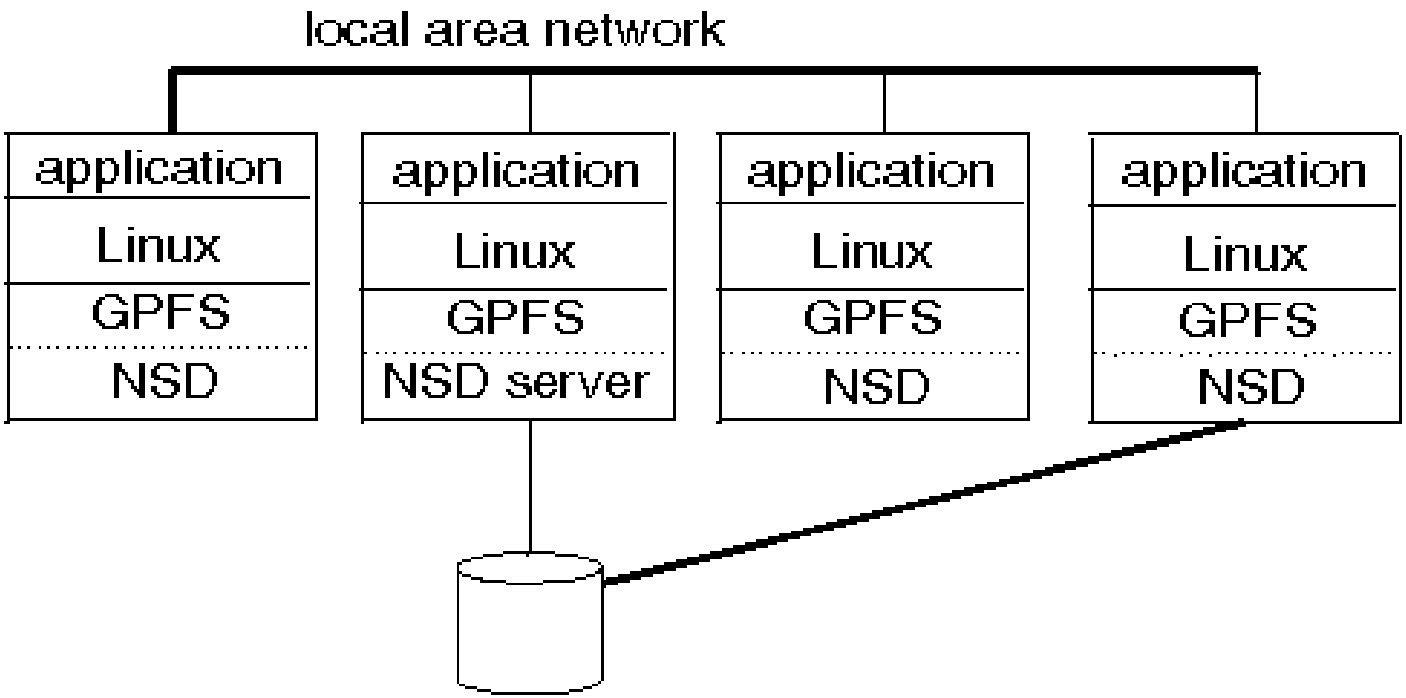
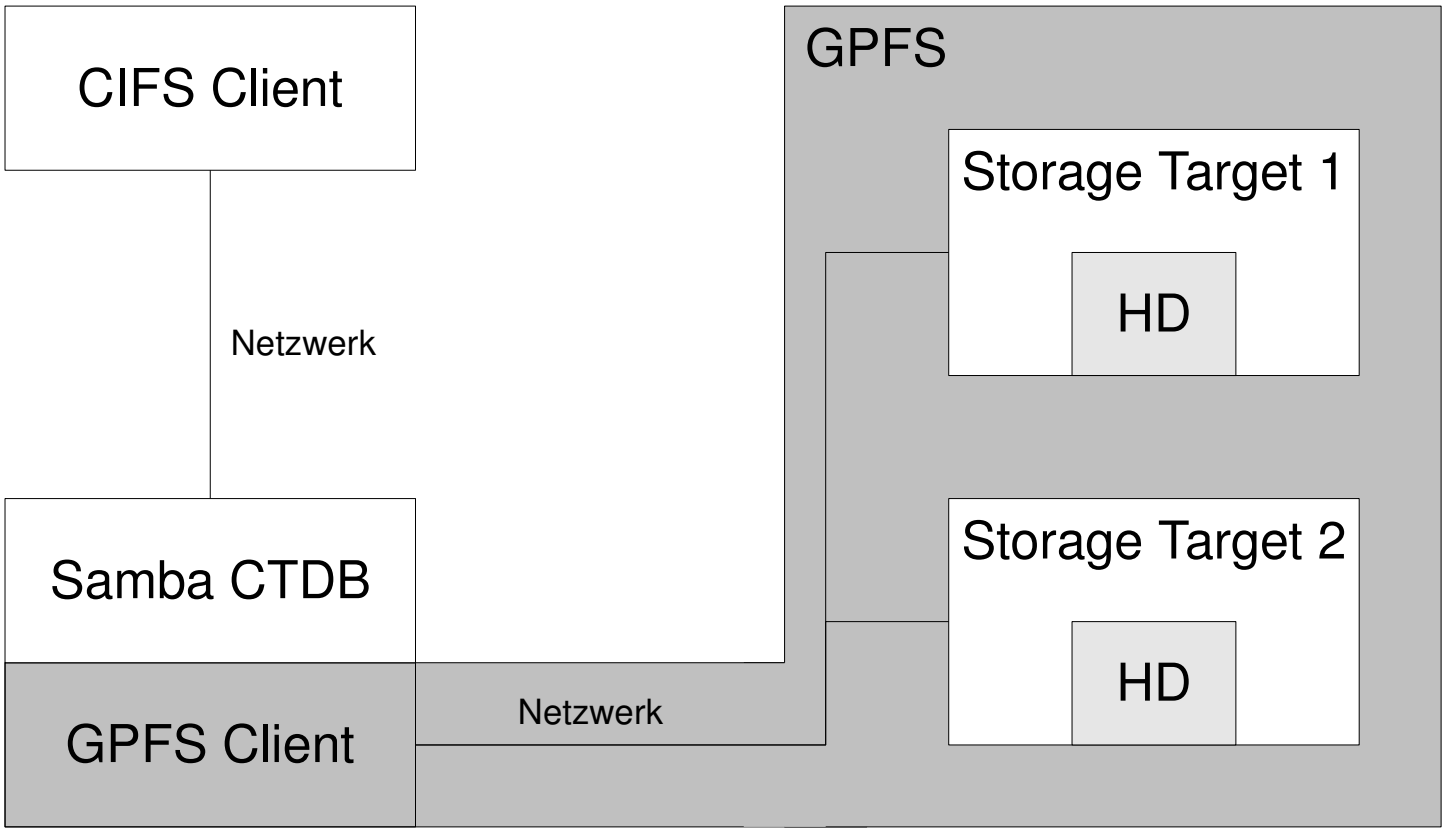


Figure: Accessing a NSD in GPFS

Source: GPFS cluster configurations, online

GPFS - test assembly



GFS

- developed as part of a thesis at the university of minnesota
- licensed under GPL since 2004
- GFS2 as the future successor

GFS

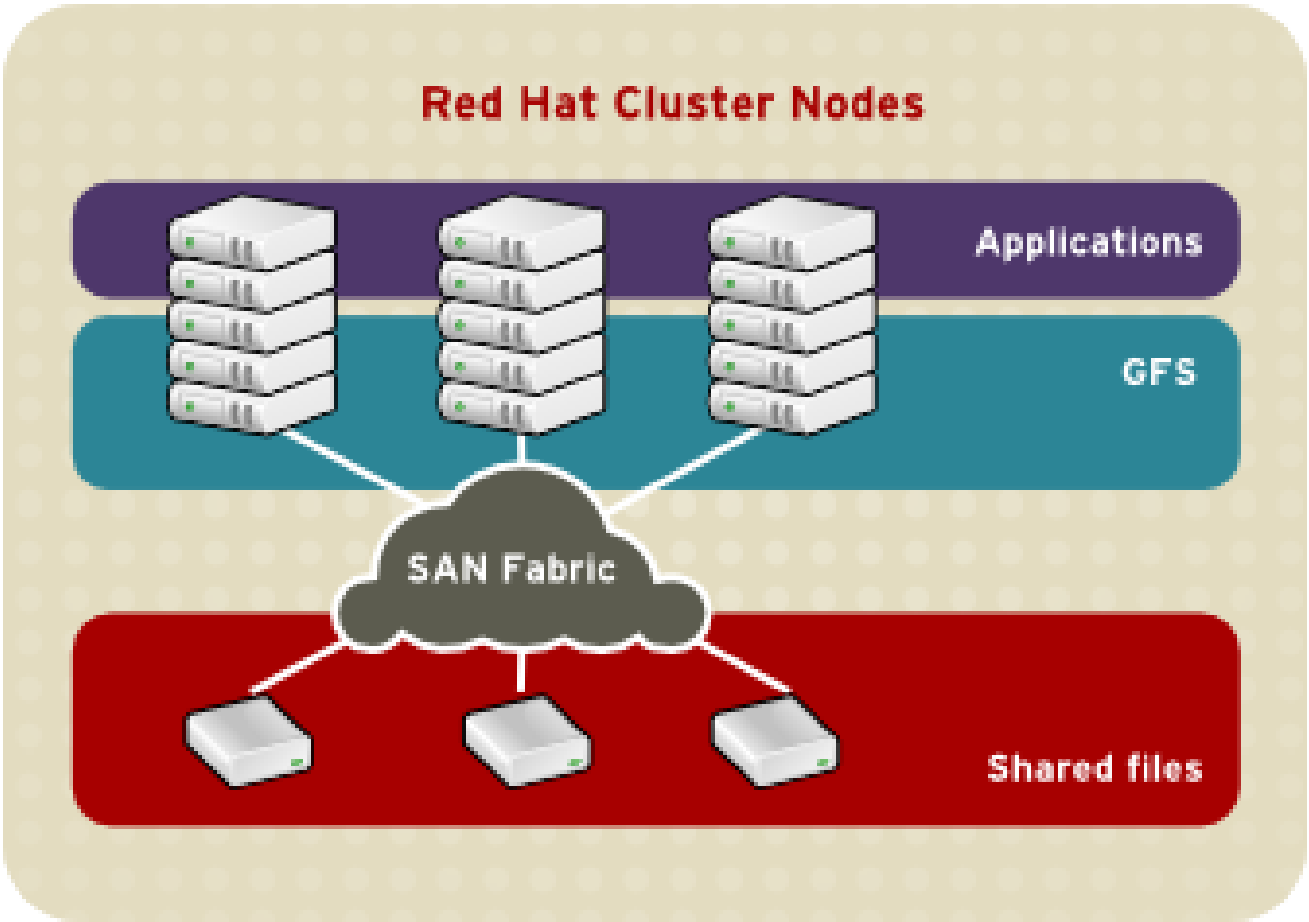
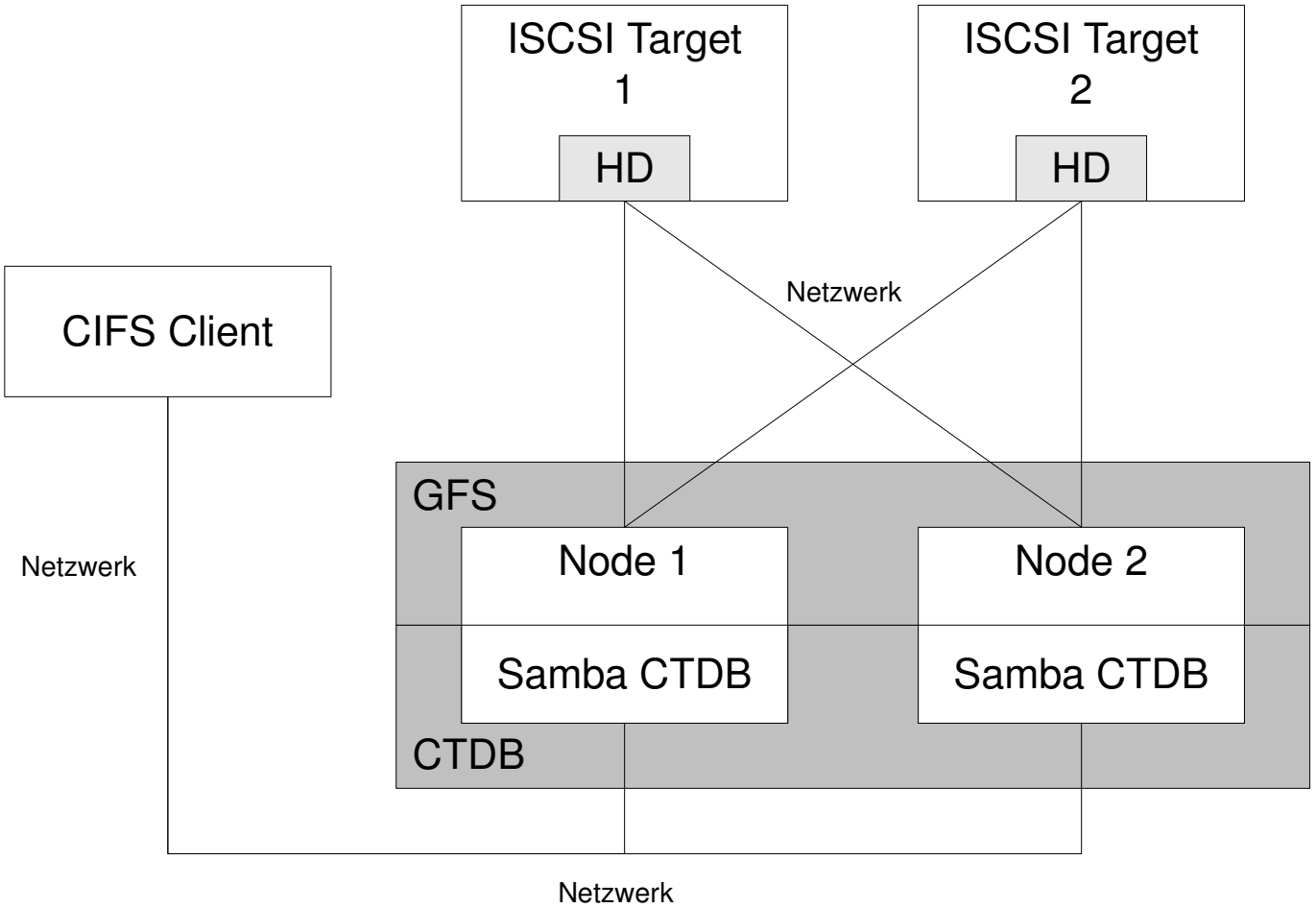


Figure: Global File System used with a SAN

Source: Red Hat Cluster Suite Overview: Red Hat Cluster Suite for Red Hat Enterprise Linux, online

GFS - test assembly



Lustre

- developed as part of a research project at the Carnegie Mellon University in 1999
- since October 2007 part of sun's portfolio
- licensed under GPL

Lustre

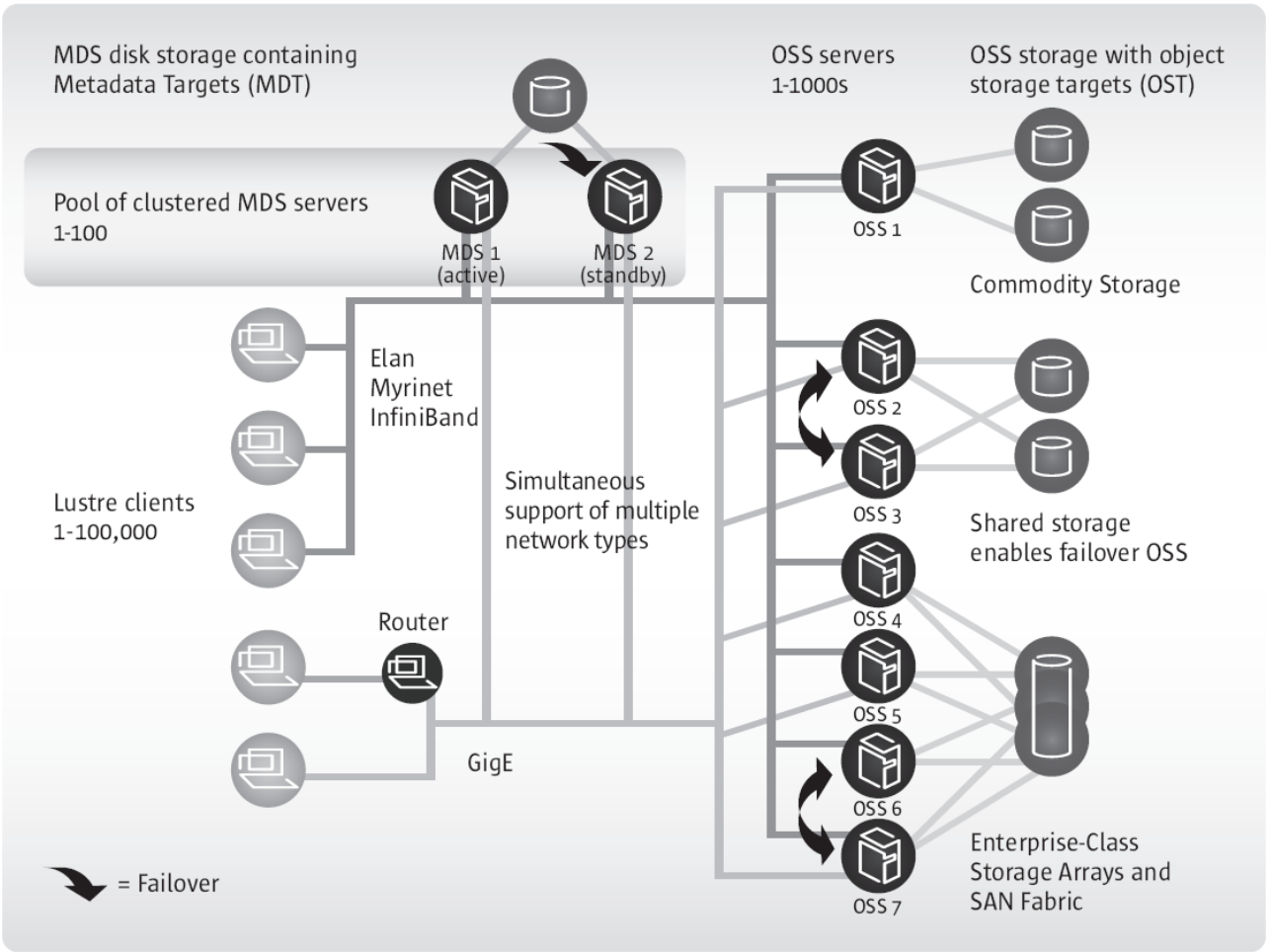
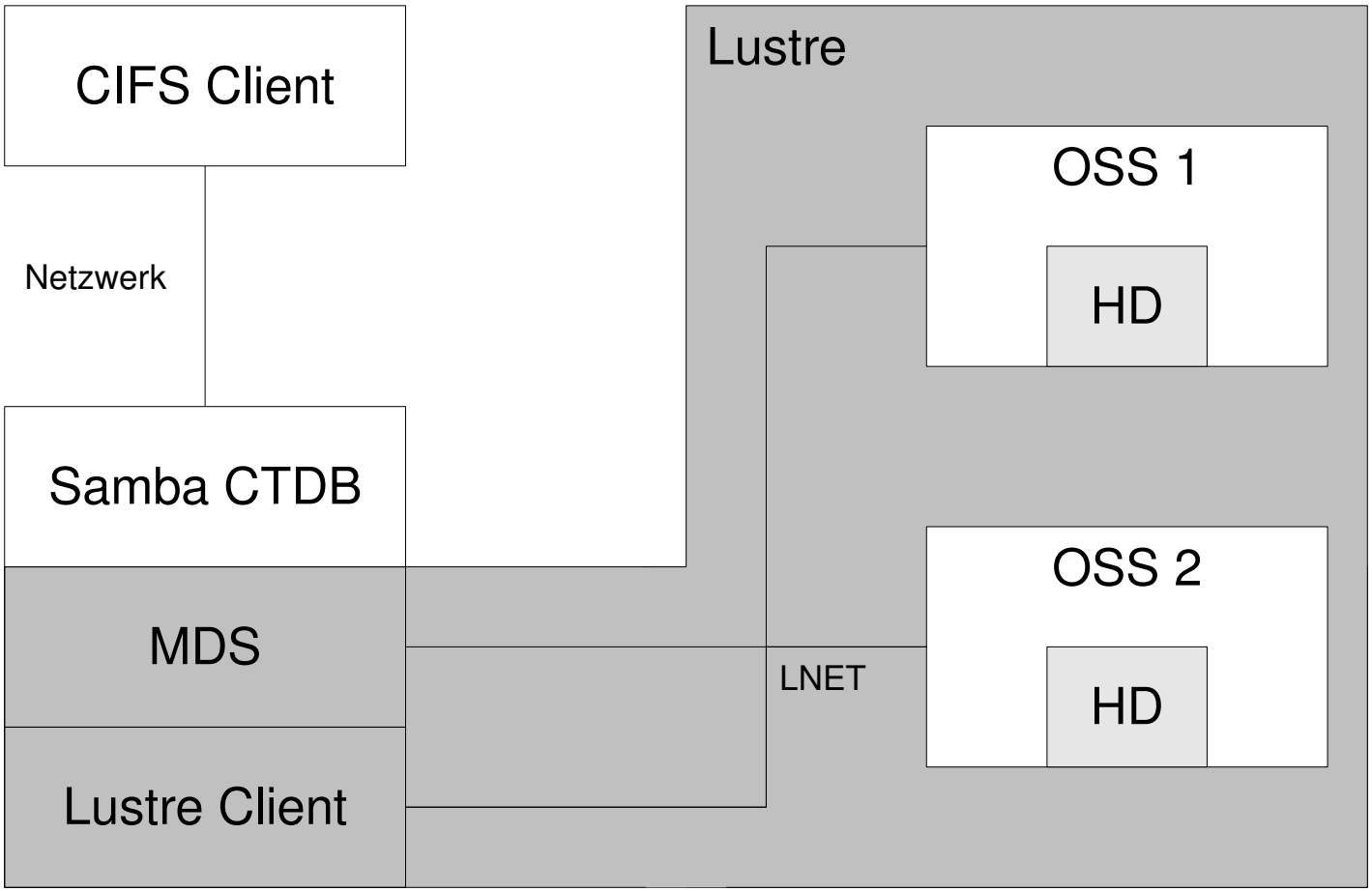


Figure: The Lustre Clustre Architcture

Source: LUSTRE FILE SYSTEM. – Whitepaper, online

Lustre - test assembly



Provided file system features

Table: Features provided by the distributed file systems for CTDB

file system	Locking (Posix/BSD)	unique Inode-Number	FileId-Mapping (fsid/fsname)
GFS	yes/yes	yes	yes/yes
GPFS	yes/yes	yes	yes/yes
Lustre	yes/yes[a]	yes	yes[b]/yes
FhGFS	-/-	-	-/-

[a] With flock as mount option
[b] With Lustre Version 1.6.2

PingPong

Table: PingPong results - lock coherence

	GFS Locks/Sec	**GPFS** Locks/Sec	**Lustre** Locks/Sec
1 node	98	264.072	5.461
2 nodes	98	2.249	3.655

PingPong

Table: PingPong results - I/O coherence

	GFS Locks/Sec	**GPFS** Locks/Sec	**Lustre** Locks/Sec
1 node	97	117.142	5.177
2 nodes	13	233	83

PingPong

Table: PingPong results - mmap coherence

	GFS Locks/Sec	**GPFS** Locks/Sec	**Lustre** Locks/Sec
1 node	98	195.533	5.559
2 nodes	31	242	124

bonnie++

- Measured speed on distributed file systems is slower then on local devices
- Measured speed on distributed file systems over Samba/CTDB is once again slower
- bonnie++ benchmark failed with lustre over Samba/CTDB

smbclient

Table: Results - reading and writing with smbclient

Dateisystem	Read (MiB/Sec)	Write (MiB/Sec)
GFS	11,78	9,49
GPFS 1HD	16.53	58.51
GPFS 2HD	32,61	61,88
Lustre 1HD	81,45	39,85
Lustre 2HD	67,57	39,18

Microsoft W2k3 - robocopy

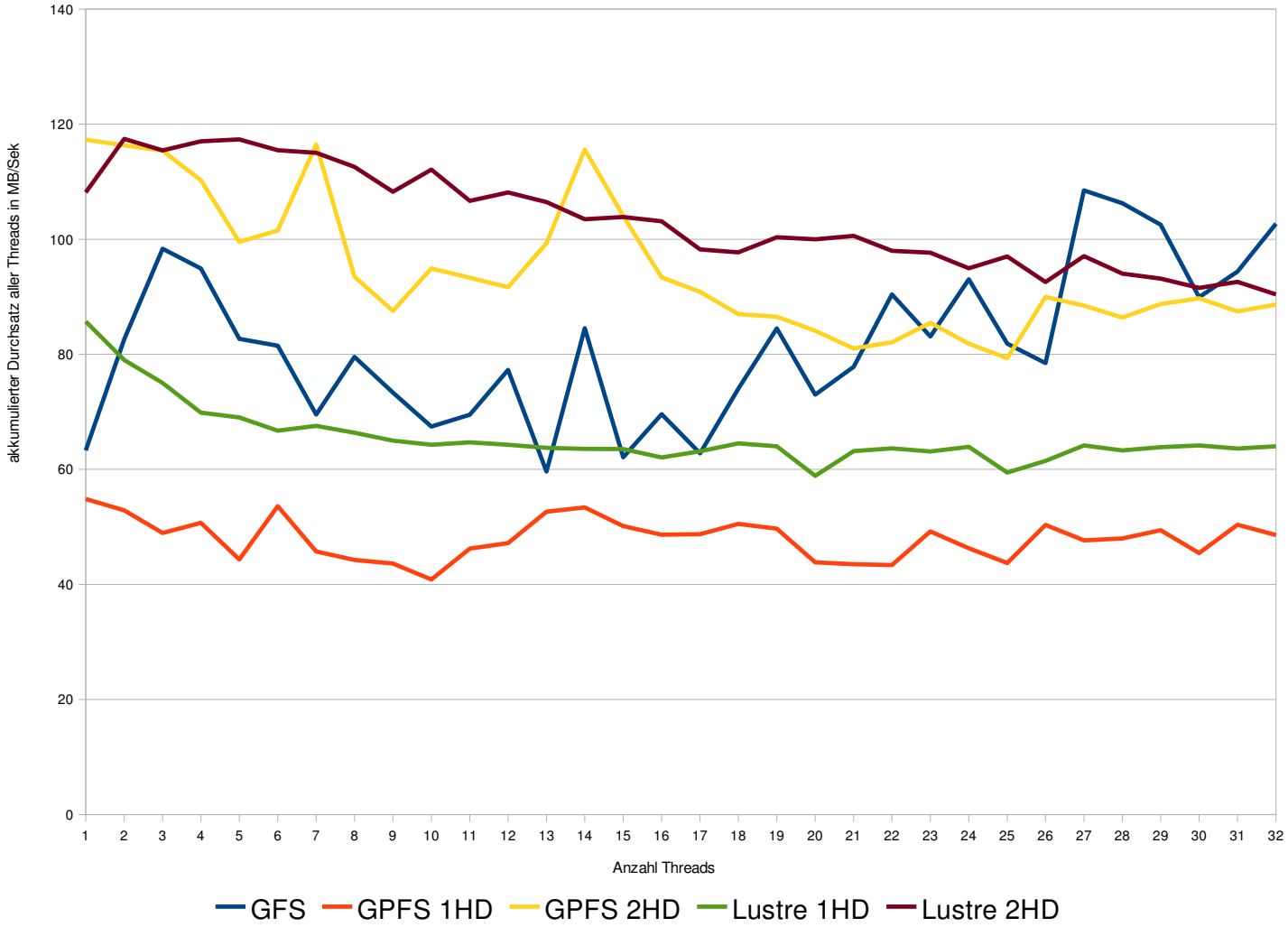
Table: Windows 2003 Server as a client

	Read	Write	Read to write
GFS	21,65 MiB/Sec	21,61 MiB/Sec	7,05 MiB/Sec
GPFS 2HD	14,2 MiB/Sec	35,81 MiB/Sec	5,18 MiB/Sec
Lustre 1HD	22,77 MiB/Sec	20,61 MiB/Sec	5,83 MiB/Sec
Lustre 2HD	23,75 MiB/Sec	20,63 MiB/Sec	2,85 MiB/Sec

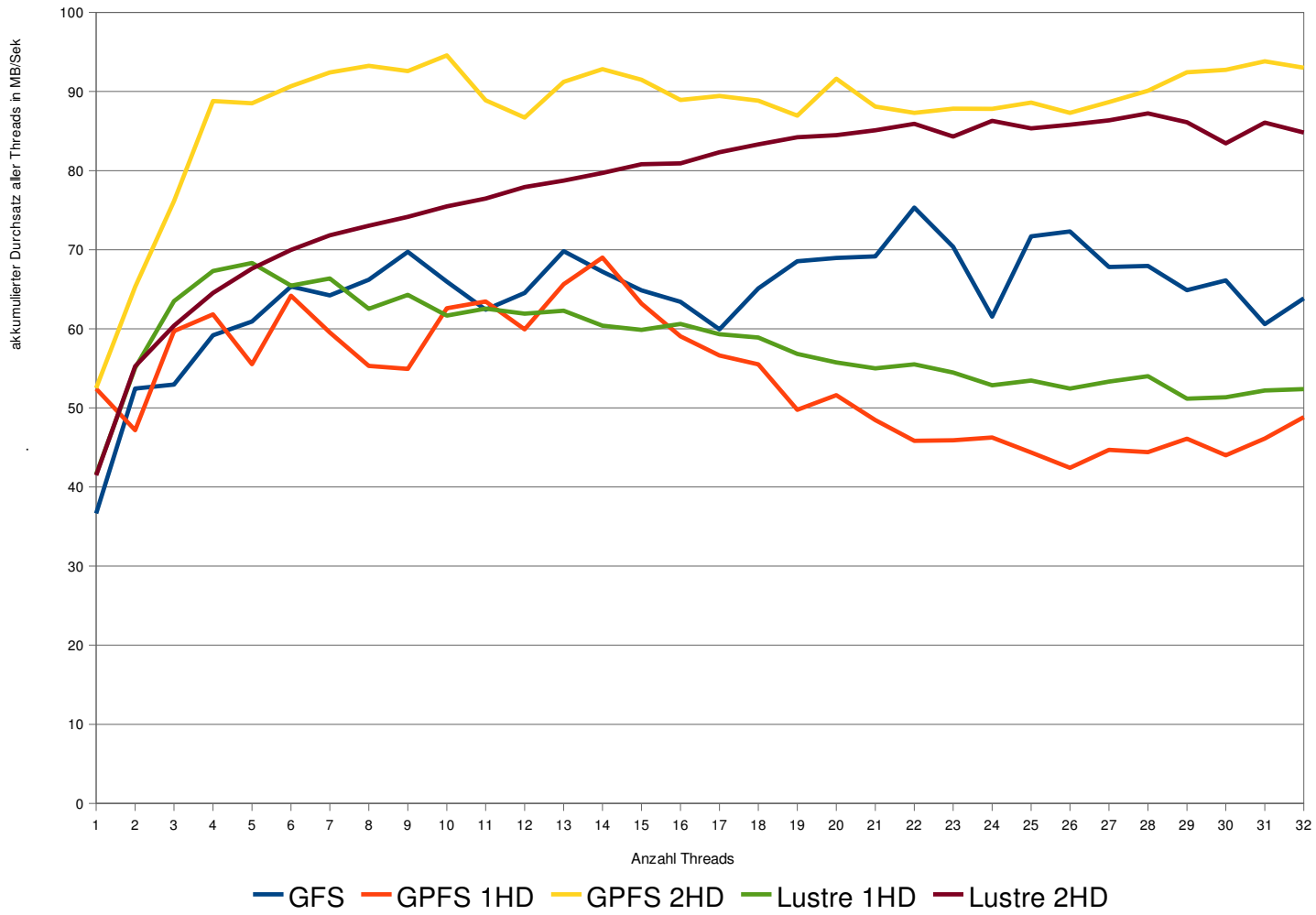
IOZone

- Distributed file system access achieved nearly the theoretical network bandwith
- Access over Samba/CTDB with iozone was limited to ca. 50 MB/Sec reading and writing with cifs-kernel-modul
- Windows version of IOZone was not used due to a bug in lustre

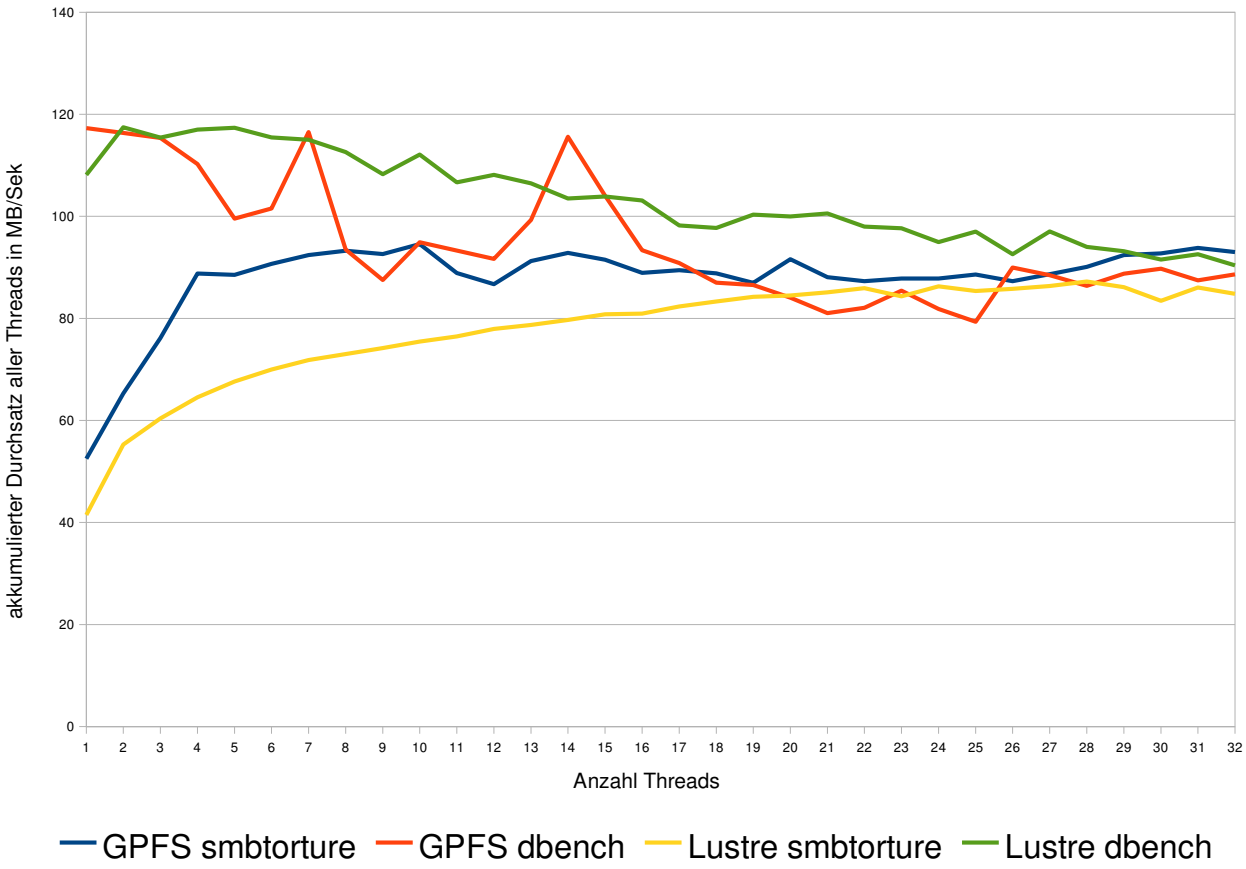
dbench - writing 1 GiB files



smbtorture - writing 1 GiB files



smbtorture & dbench altogether



Conclusion

- The locks/sec are heavily depending on the distributed file system.
- With concurrent access the locks/sec drop.
  $\Rightarrow$ Higher latency
- There could be many reasons for more latency, like higher network latency, seek latencies, more managemet overhead with more clients and so on...
- Throughput with Samba also depends on the cifs implementation of the client
- According to my tests one client alone could not reach the maximum throughput with Samba/CTDB

Questions ?

Questions ? Thanks for your attention!