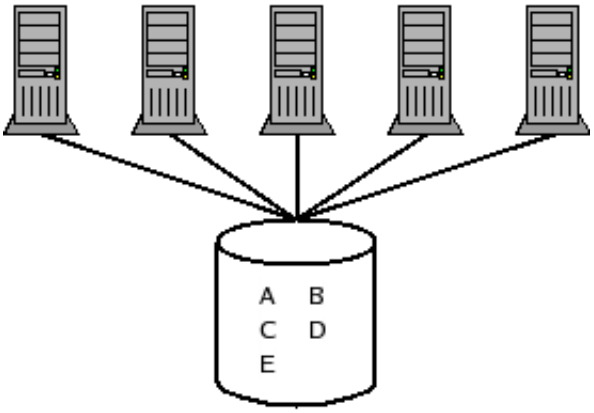
Distributed Parallel Fault Tolerant File Systems

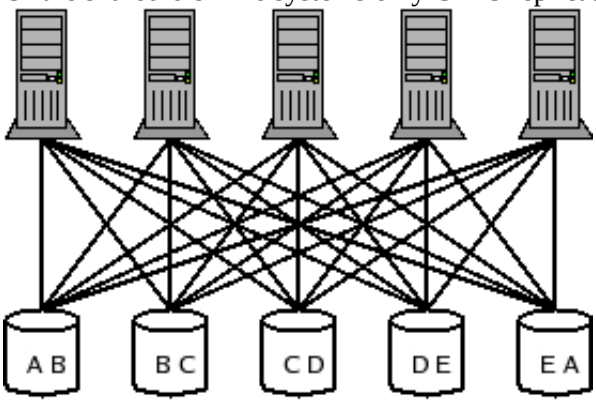
This is an overview of the current (January 2007) available file systems for Linux that may be considered for use in high-availability clusters and high performance computing, with a focus on open source.



Shared-disk file systems locate their data on shared block storage, where nodes have direct access to the data. They are not fault tolerant, unless run on such hardware (external hardware raid etc).

Examples: GPFS, GFS, OCFS.

Of the shared-disk file systems only GPFS replicate data, but it has a proprietary license.



Distributed parallel file systems access their data over the network using a protocol above the block level.

Examples: Ceph, GFarm, GlusterFS, Lustre, PVFS.

Of the distributed parallel file systems several are still in development like Ceph, GFarm and GlusterFS. They all focus on fault tolerance. Lustre and PVFS may be set up to be fault tolerant, but only in the same way as an ordinary distributed file server - with shared storage between the servers and failover.

Summary table of interesting file systems

Name	Features	License	Vendor/distribution	Description
Ceph	distributed, parallel, fault tolerant	GPL	SSRC at UCSC	In development
Gfarm Grid File System	distributed, parallel, fault tolerant	X11	ApGrid	No automatic replication.
GFS	shared	GPL	Linux kernel	
GlusterFS	distributed, parallel, fault tolerant	GPL	Z RESEARCH	In development
Google File System	distributed, parallel, fault tolerant	not available	Google	Only internal use at Google.
GPFS	shared, replication	Proprietary	IBM	Sweet but expensive.
Hadoop Distributed File System	distributed, parallel, fault tolerant	Free	Apache	Development just started.
Lustre	distributed, parallel	GPL	Cluster File Systems	No replication - only fault tolerant with failover and external hardware raid.
NFS	distributed	GPL	Linux kernel	Only fault tolerant with failover and shared storage.
OpenAFS	distributed	IBM Public License	IBM	Does have read-only replication, but only fault tolerant with failover and shared storage.
OCFS	shared	GPL	Linux kernel	
PeerFS	distributed, parallel, fault tolerant	Proprietary	Radiant Data Corp	Only mirroring.

PVFS	distributed, parallel	GPL	Several contributers	
Terragrid	distributed, parallel, fault tolerant	Proprietary	Terrascale	Nice.

Conclusion

With the goals **commodity hardware**, **parallel operation**, **fault tolerance** and a **free software license** there are several interesting file systems in development; **Ceph**, **GlusterFS** and **GFarm**.

However, **Lustre** may be an alternative soon since they have got replication in their roadmap, but whether or not the replication will be free software or proprietary remains to be seen.

This overview was orginally written in July 2006 and updated in January 2007 by Jerker Nyberg, Bitvis Datakonsult.