Database Deployment @ CNAF

Barbara Martelli

Rome, April 4st 2006

Outline

- DB service @ CNAF and 3D collaboration
- Overview of deployed technologies:
  - Streams for data propagation
  - Oracle Real Application Clusters
  - Shared storage management technologies
- Deployment status @ CNAF

DB Service @ CNAF and 3D collaboration

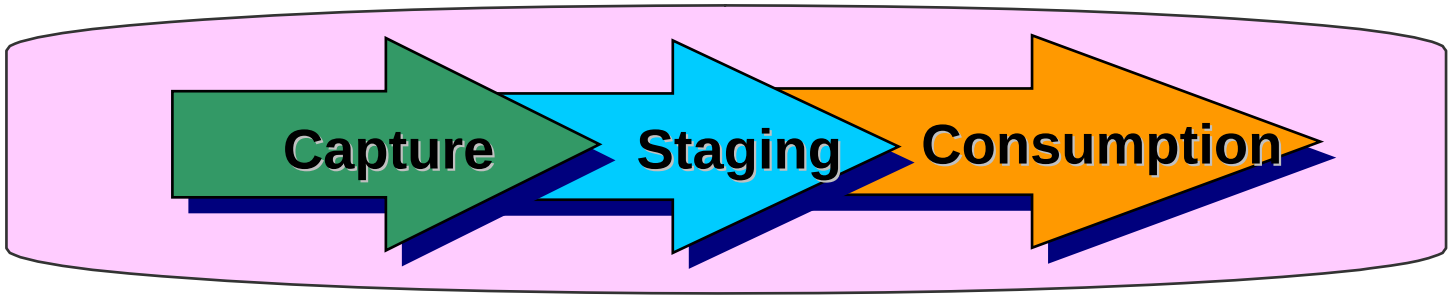
CNAF actively collaborates with LCG3D group, DB service structure follows the guidelines of 3D providing 2 different environments separated on service level basis:

- Development environment:
  - Shared HW setup
  - DBA limited support (via email)
  - 8/5 monitoring and availability
- Production environment:
  - Dedicated HW setup (to be agreed two months in advance)
  - DBA support via email and phone
  - 8/5 monitoring and availability
  - Backups every 10 minutes
  - Limited number and scheduled number interventions

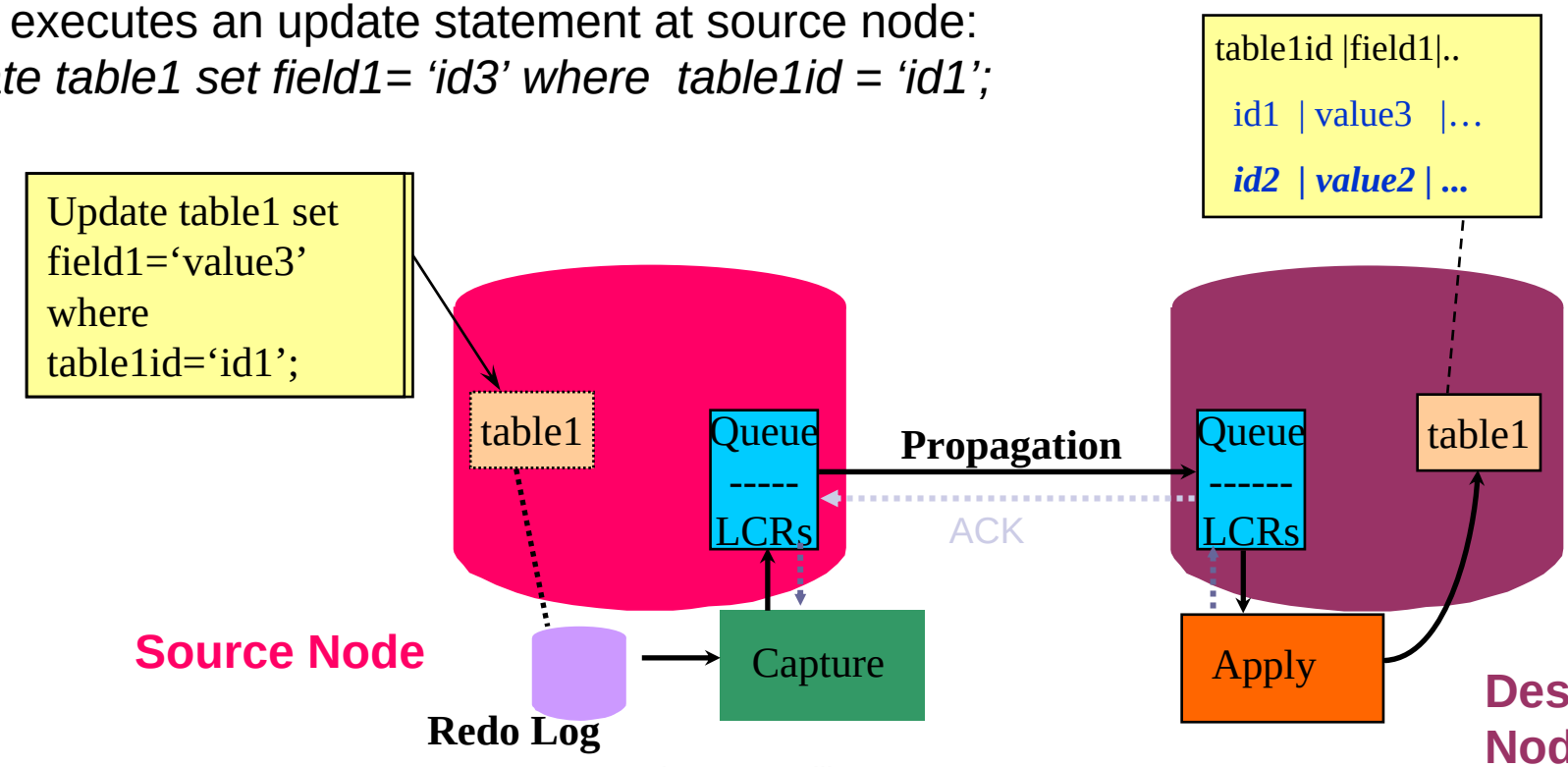
HW and Human resources

- Development environment:
  - One 4-nodes RAC (OCFS2), 1TB shared storage on IBM FastT900.
- Production environment:
  - 2 2-nodes RACs, 2 TB shared storage on 2 JBOD Dell PowerVault 224F. Allocated to LHCb and ATLAS.
  - 1 HP Proliant DL380G4 for Service instances such as Castor2 stager and FTS. 2 Xeon 2,4GHz for DLF (Castor2) and FTS catalog.
- 2 people involved (almost 1 FTE)

Oracle Streams (Data replication)

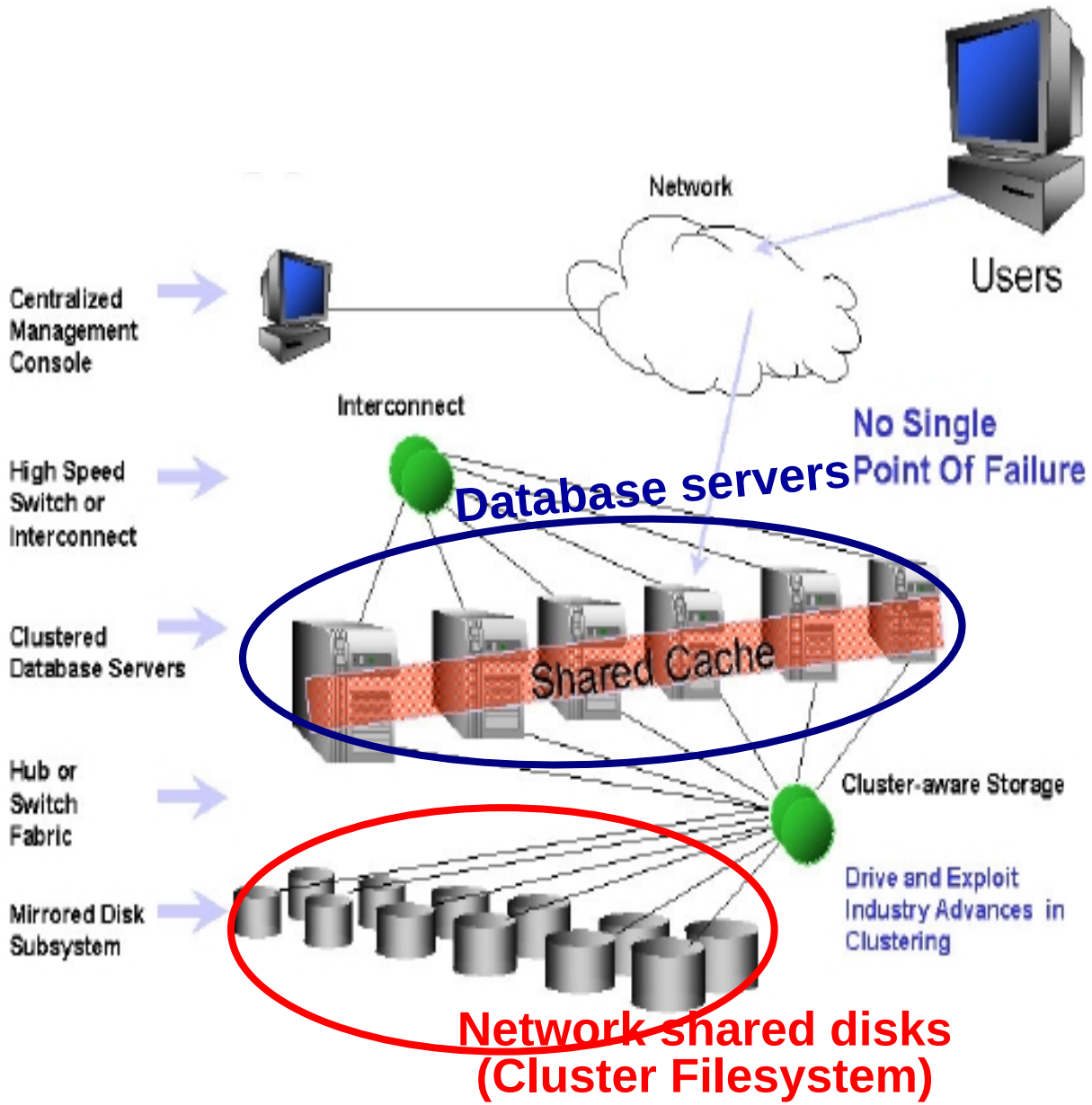


User executes an update statement at source node:
update table1 set field1= 'id3' where table1id = 'id1';



Oracle Real Application Clusters

- The Oracle Real Application Cluster technology allows to share a database amongst several database servers
- All datafiles, control files, PFILEs, and redo log files in RAC environments must reside on cluster-aware shared disks so that all of the cluster database instances can access them.
- RAC aims to provide highly available, fault tolerant and scalable database services



ASM

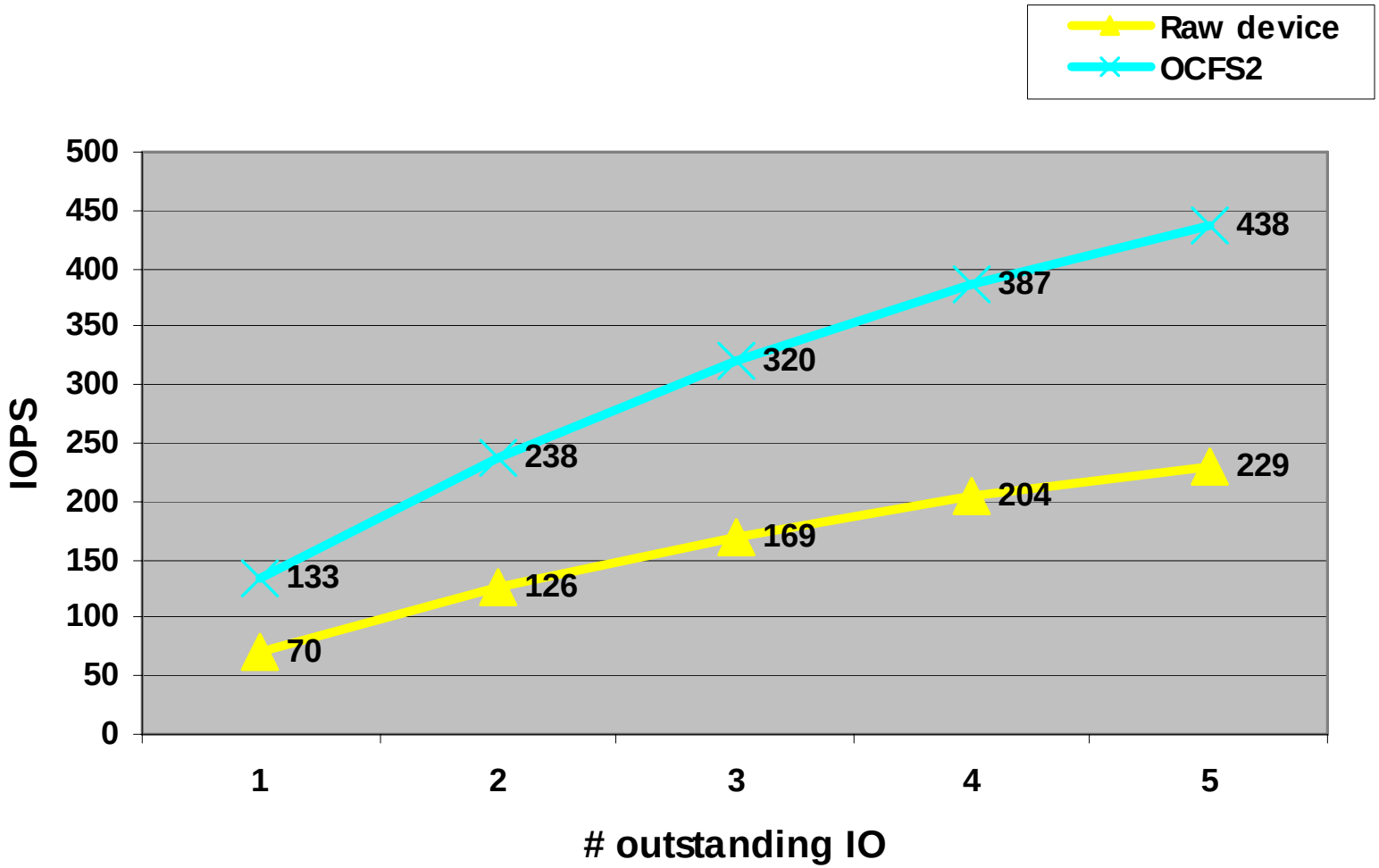
Automatic Storage Management (ASM) is a database service that allows the efficient management of disk drives. ASM can provide management for single SMP machines, or across multiple nodes of a RAC.

- ASM has the following characteristics:
  - It automatically does **load balancing** in parallel across all available disk drives to prevent hot spots and maximize performance, even with rapidly changing data usage patterns.
  - It **prevents fragmentation** so that there is never a need to relocate data to reclaim space.
  - It does **automatic online disk space reorganization** for the incremental addition or removal of storage capacity.
  - It can maintain **redundant copies** of data to provide fault tolerance, or it can be built on top of vendor-supplied, reliable storage mechanisms.
  - Data management is done by selecting the desired reliability and performance characteristics **for classes of data** rather than with human interaction on a per file basis.

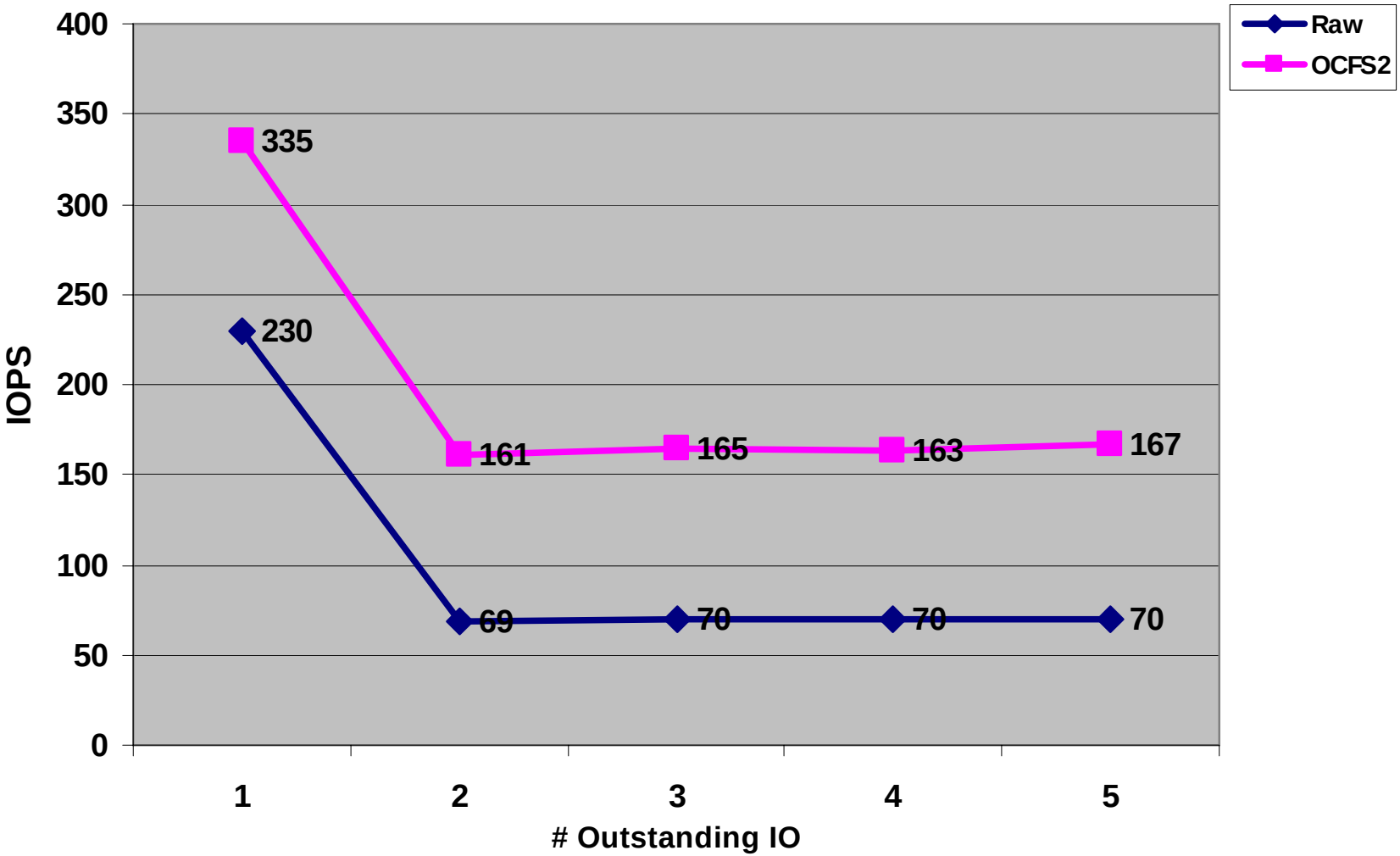
OCFS2

- It is an extent based, POSIX compliant file system.
- OCFS2 is based on a cluster suite with hearthbeat to control the state of the members and a configuration tool which helps to configure and propagate FS configuration to all the nodes
- init.d script which loads the needed modules, mounts the file system and starts the ocfs2 service.
- Some problems due to HW incompatibilities arose during the deployment.
- Data block corruption in system tablespace at each DB installation even if the database appeared to work properly at the beginning.
- We have found that the corruption was due to a node with a QLogic 1210 board slightly different from the others.

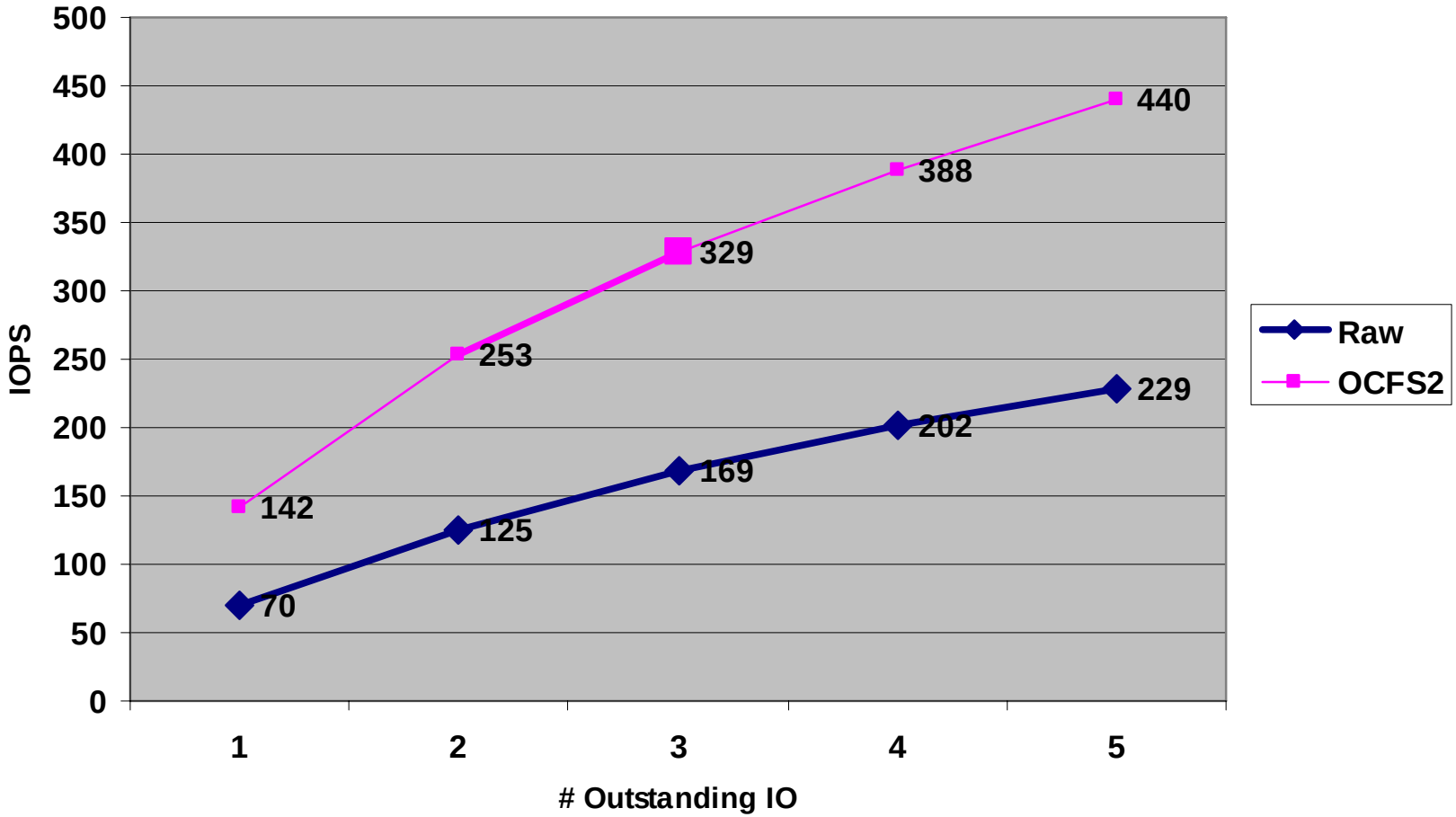
Raw vs OCFS2 configuration Random Reads



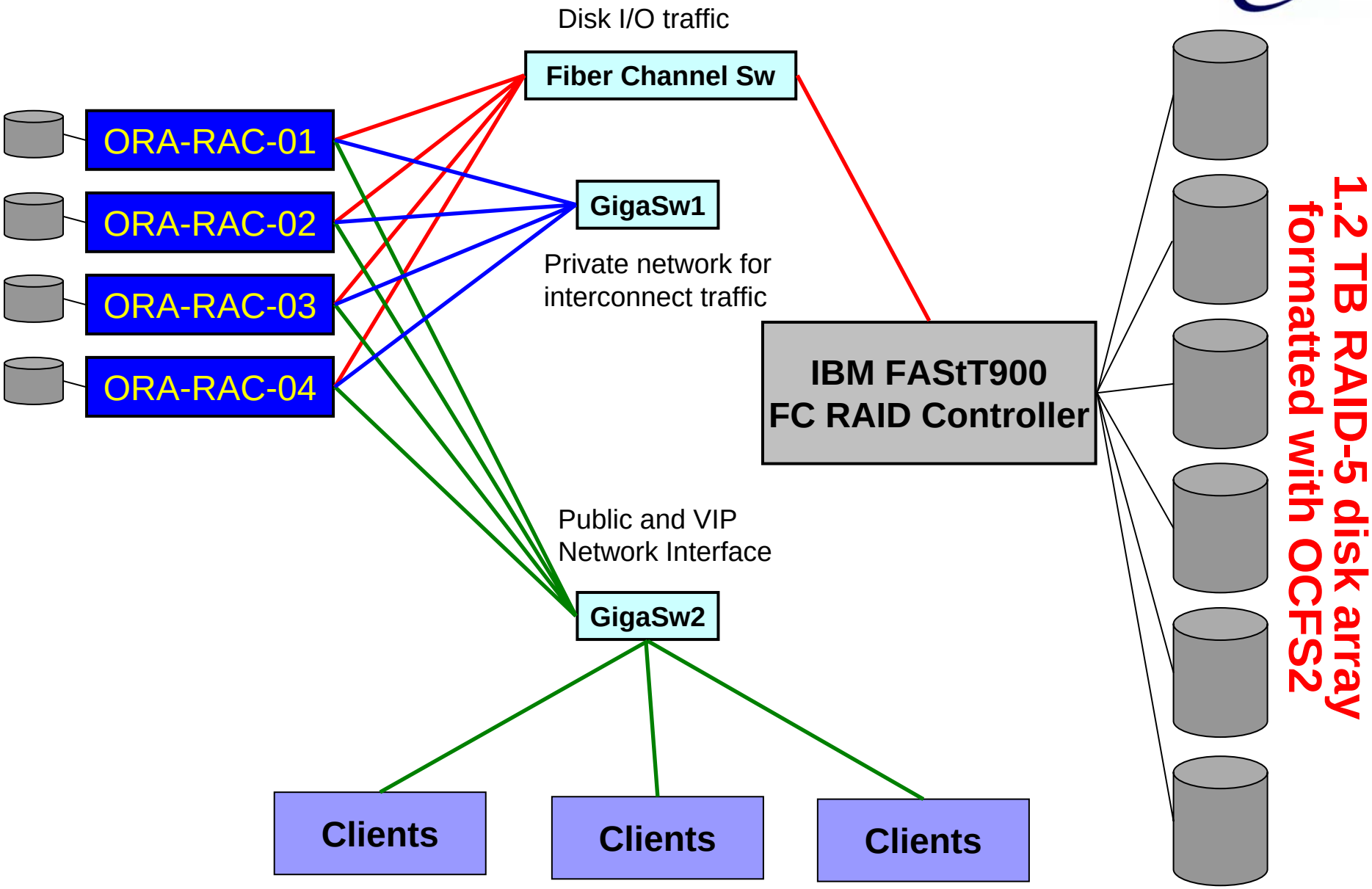
Raw vs OCF2 configuration
IOps with Random Writes



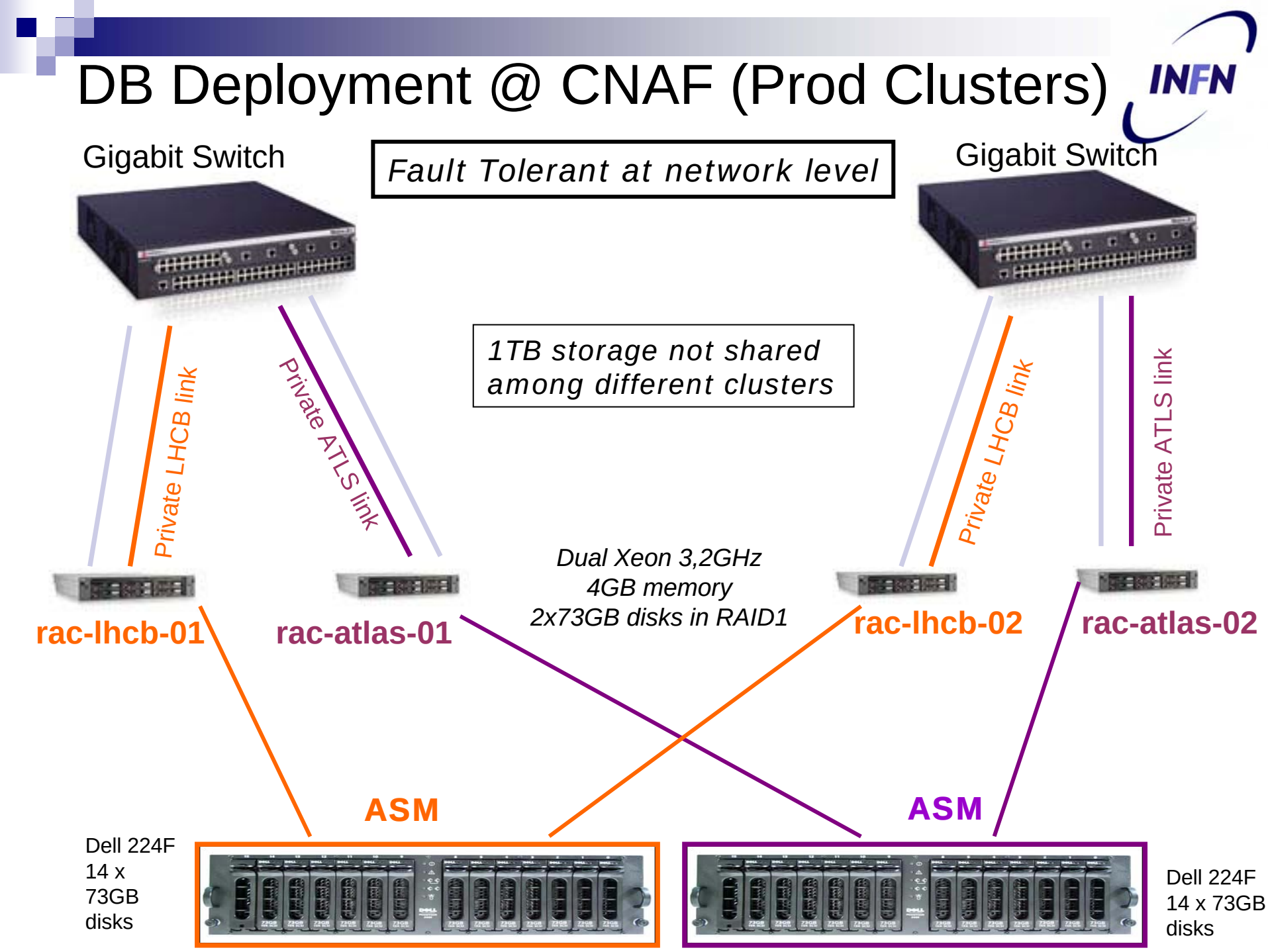
Raw vs OCFS2
IOps with Sequential Reads



DB Deployment @ CNAF (Test Env)



DB Deployment @ CNAF (Prod Clusters)



References

- OCFS2:
  - http://oss.oracle.com/projects/ocfs2
- RAC and High Availability:
  - http://download-east.oracle.com/docs/cd/B19306_01/rac.102/b14197/toc.htm
  - http://www.oracle.com/technology/deploy/availability/pdf/ora_lcs.pdf
- Oracle Streams:
  - http://download-east.oracle.com/docs/cd/B19306_01/server.102/b14229/toc.htm