Experience and Lessons learnt from running High Availability Databases on Network Attached Storage

Ruben Gaspar

Manuel Guijarro et al

IT/DES

Outline

- Why NAS system for DB Infrastructure
- Topology
- Bonding
- NAS Configuration
- Server Node Configuration
- Oracle Software installation
- Performance
- Maintenance & Monitoring
- Conclusion

What is a NAS ?

- Network-attached storage (NAS) **is the name given to dedicated Data Storage technology that can be connected directly to a computer Network to provide centralized data access and storage to heterogeneous network clients.**
- **Operating System and other software on the NAS unit provide only the functionality of data storage, data access and the management of these functionalities .**
- **Several file transfer protocols supported (NFS, SMB, etc)**
- **By contrast to a SAN (Storage Area Network), remote devices do not appear as locally attached. Client requests portions of a file rather than doing block access.**

Why NAS for DB infrastructure ?

- To ease file sharing needed for Oracle RAC: it is much simpler than using raw devices. The complexity is managed by the appliance instead of by the server
- To ease relocation of services within Server nodes
- To use NAS specific features: Snapshots, RAID, Failover based on NAS partnership, Dynamic re-Sizing of file systems, Remote Sync to offsite NAS, etc
- Ethernet is much simpler to manage than Fiber Channel
- To ease File Server Management: automatic failure reporting to vendor, etc
- To reduce cost (no HBA no FC switch)
- To simplify administration of Data Base

NAS Based Oracle DB infrastructure for Castor 2

2 * Network Switches
J8693A
S3500

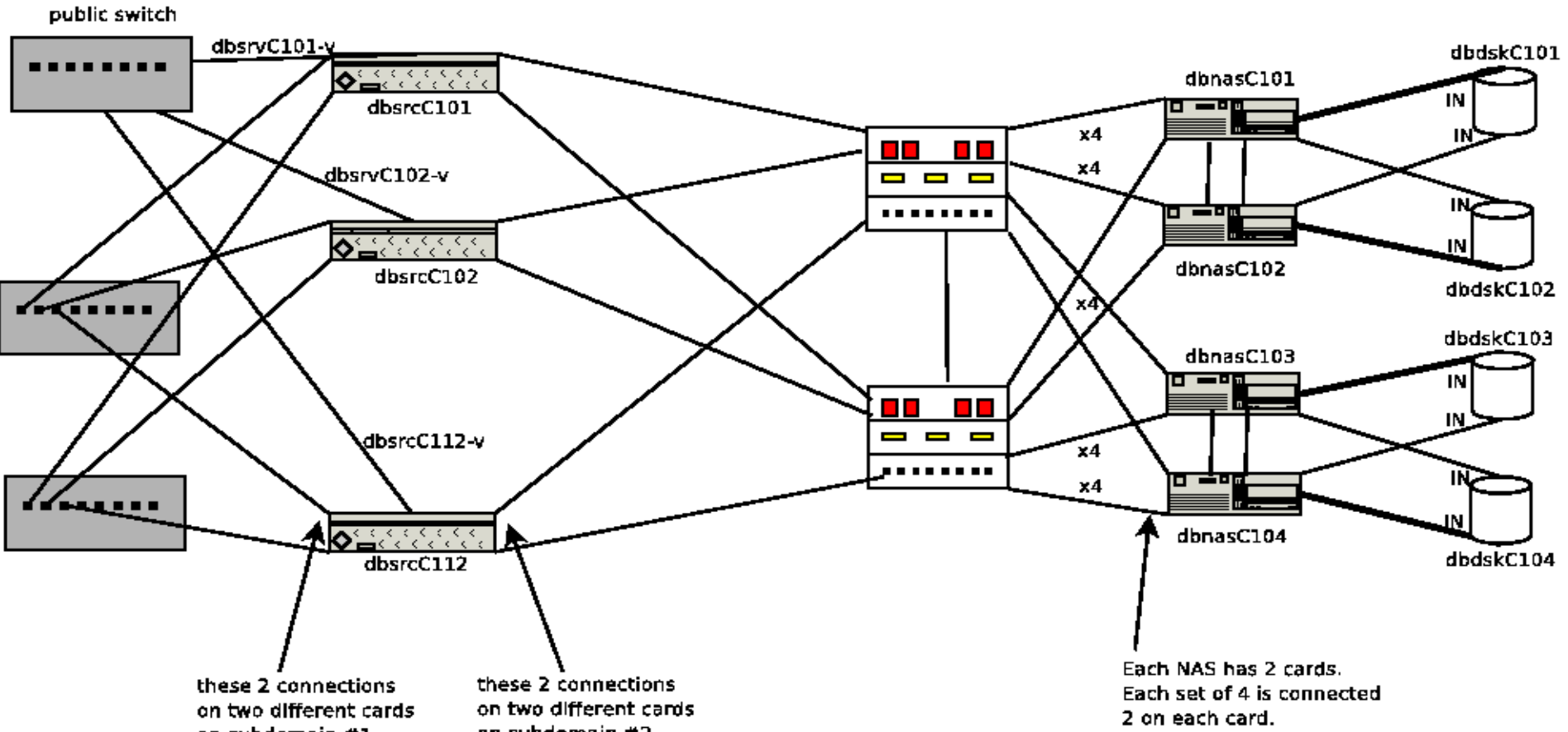
12 * Sever Nodes
HPDL380G5(2u)

2 * Network Switches
J8697A
5406zl 3u

4 * IBM NAS
N5200 3u
(AKA NetApp 3020)

4* Disk Shelfs
EXN2000 3u

14 disks each



Bonding modes used

- Bonding (AKA Trunking): Aggregate multiple network interfaces into a single logical bonded interface. Using load balancing, i.e. using multiple network ports in parallel to increase the link speed beyond the limits of a single port or for automatic failover

  vif create multi -b ip dbnasc101_vif1 e0a e0b e0c e0d
- Each NAS has 2nd level VIF (Virtual Interface) made of two 1st level VIF's of 4 NIC's each. Only one of the 1st level VIF is active.
- Use active-backup mode for bonding with monitoring frequency of 100 ms on the Node server side (dbsrvdXXX and dbsrvcXXX machines)
- Use IEEE 802.3ad Dynamic link aggregation mode for trunking between NAS and switches and connection between switches

NAS Configuration

- Data ONTAP 7.2.2
- 13+1 disks in each Disk Shelve (all in a single disk aggregate)
- NetApps' RAID_DP (Double Parity): Improved RAID 6 implementation
- Several FlexVol's in each aggregate
- AutoSize Volume option
- SnapShots: Only used before DB maintenance operations for the time being
- Quattor can not be used since no agent can be installed in Data ONTAP 7.2.2

NAS Configuration

- **Aggregate (13 disks + 1 spare disk)**

  dbnasc101> sysconfig -r

  Aggregate dbdskc101 (online, raid_dp) (block checksums)

  Plex /dbdskc101/plex0 (online, normal, active)

  RAID group /dbdskc101/plex0/rg0 (normal)

  RAID Disk Device HA SHELF BAY CHAN Pool Type RPM Used (MB/blks) Phys (MB/blks)

  --------- ------ ------------- ---- ---- ---- ----- -------------- --------------

  dparity 0a.16 0a 1 0 FC:A - FCAL 10000 136000/278528000 137104/280790184

  parity 0a.17 0a 1 1 FC:A - FCAL 10000 136000/278528000 137104/280790184

  data 0a.18 0a 1 2 FC:A - FCAL 10000 136000/278528000 137104/280790184

  – Raid_dp: two failing disks protection

NAS Configuration

- **Flexible volumes**

  vol create volname aggregate 1g
  vol autosize volname -m 20g -i 1g on
  vol options volname nosnap on
  vol options volname no_atime_update on
  snap delete -a volname

- **Redundancy accessing disks in the storage system, dual loop:**

  dbnasc101> storage show disk -p

PRIMARY PORT		*SECONDARY PORT*		*SHELF*	*BAY*
0d.32	*B*	*0b.32*	*A*	*2*	*0*
0b.33	*A*	*0d.33*	*B*	*2*	*1*

Server Node Configuration

- Quattor based installation and configuration covering:
  - Basic RHE4 x86_64 installation
  - Network topology setup (Bonding): modprobe and network NCM components
  - Squid, Sendmail setup (http proxy for NAS report system) using NCM components
  - All hardware components definition in CDB (including NAS boxes and disk shelves)
  - RPM's for: Oracle-RDBMS, Oracle-RACRDBMS and Oracle-CRS
  - NCM RAC component
  - TDPO and TSM components used to configure Tivoli
- Lemon based monitoring for all Server node resources
- Always using CERN's CC infrastructure (Installations performed by SysAdmin support team, CDB)

Server Node Configuration: standard DB set-up

- Four volumes created, critical files split among volumes (RAC databases)
  - /ORA/dbs00: logging and archive redo log files and a copy of the control file
  - /ORA/dbs02: a copy of the control file and the voting disk
  - /ORA/dbs03: spfile and datafiles, and a copy of the control file, the voting disk and registry file
  - /ORA/dbs04:a copy of the control file, the voting disk and the registry - different aggregate (i.e. different disk shelf and different NAS filer).
- Accessed via NFS from the sever machines
- Mount options rw,bg,hard,nointr,tcp,vers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768
- DATABASES in production on NAS. RAC instances in production:
  - Castor Name server, various Castor 2 Stagers (for Atlas, CMS LHCb, Alice..), Lemon, etc

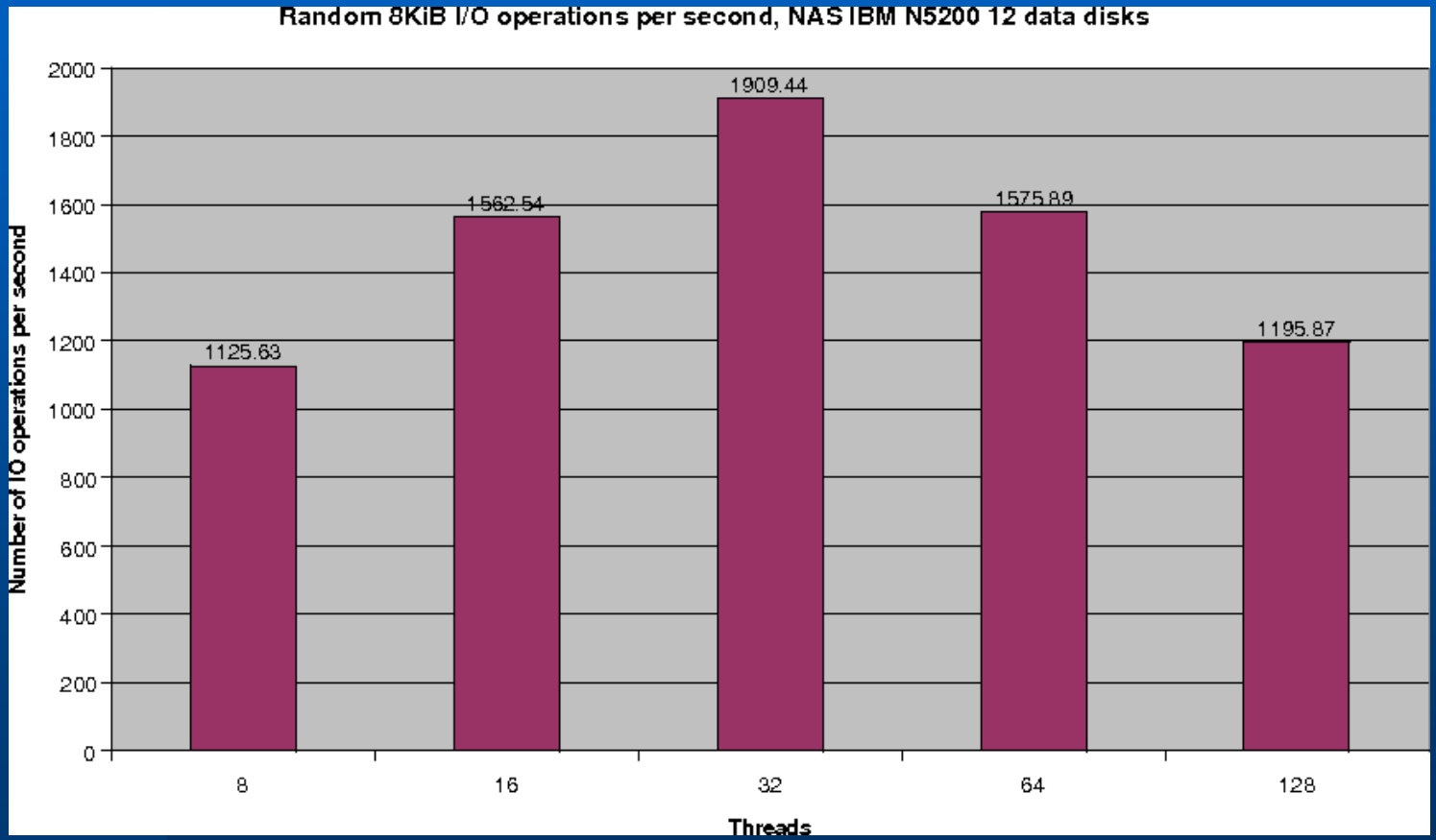
Oracle Software Installation

- CRS, RDBMS are packaged in RPMs. (Oracle Universal Installer + Cloning)
- BUT dependencies in RPMs are not usable
  - Inclusion of strange packages (perl for kernel 2.2!)
  - Reference in some libraries to SUN (?!) packages
- BUT package verification cannot be used as the cloning procedure does a re-link (checksum)
- THUS, RPMs can only be used for installation of files
- A Quattor component configures RAC accordingly and starts daemons successfully

Performance

- **Oracle uses direct I/O over NFS**
  - **avoids external caching in the OS page cache**
  - **more performing (for typical Oracle I/O database workloads) than buffered I/O (double as much)**
  - **Requires enabling direct I/O at the database level:**
    - ***alter system set FILESYSTEMIO_OPTIONS = directIO scope=spfile sid='*';***
- **Measurements**
  - **Stress test: 8KB read operations by multiple threads in a set of aggregates (with a total of 12 data disks): Max of ~ 150 I/O operations per second and per disk.**

Performance



Maintenance and Monitoring

- Some interventions done (NO downtime):
  - Some failed Disks replacement
  - HA change
  - NAS Filer OS update Data ONTAP upgrade to 7.2.2 (rolling forward)
  - RAC member full automatic reinstallation
  - PSU replacement (both in NAS filer and Server nodes)
  - Ethernet Network cable replacement
- Several Hardware failures successfully tested
  - Powering off: Network switche(s), NAS filer, RAC member, …
  - Unplugging: FC cables connecting NAS filer to Disk Shelf, Active bond members cable(s) of Private interconnect, VIF's,..
- Using Oracle Enterprise Manager monitoring – Lemon sensor on Server Node could be developed

Maintenance and Monitoring

ORACLE Enterprise Manager 10g Grid Control

Setup | Preferenc

Home | Targets | Deployments | Alerts | Policies | Job

Hosts | Databases | Application Servers | Web Applications | Services | Systems | Groups | All Targets

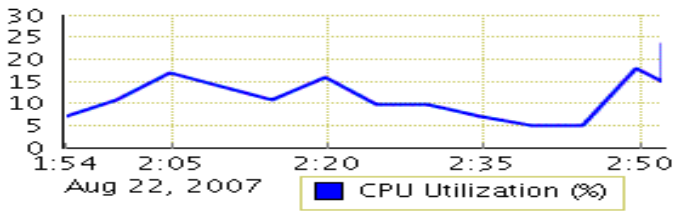
Network Appliance Filer: dbnasc101

Home | Performance | Volumes | Qtrees

Page Refreshed Aug 22, 2007 2:52:13 PM

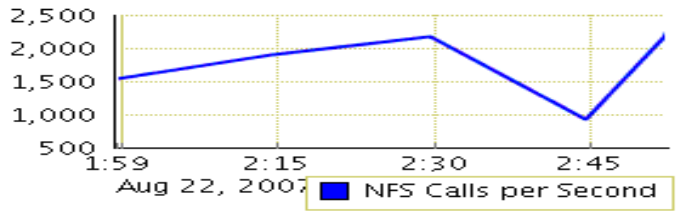
View Data: Real Time: Manual Refresh

CPU Utilization (%)



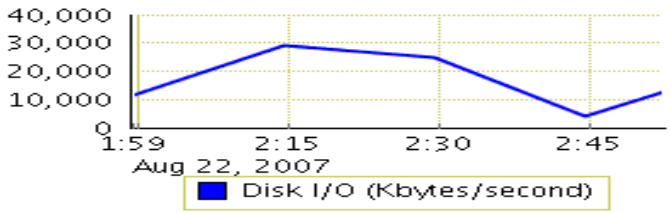
Number of CPUs	2
CPU Utilization (%)	24

NFS Calls per Second



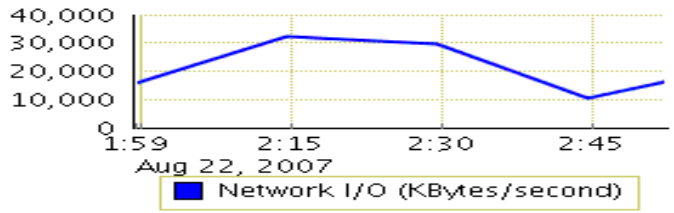
NFS Calls per Second	2249
NFS Bad Calls (%)	0

Disk I/O



Total Disk Read Rate (KBytes/second)	9896
Total Disk Write Rate (KBytes/second)	3025

Network I/O



Total Network Receive Rate (Kbytes/second)	3128
Total Network Send Rate (Kbytes/second)	13766

Conclusion

- NAS based infrastructure is easy to maintain
- All interventions so far done with 0 downtime
- Good Performance for Oracle RAC databases
- Several pieces of infrastructure (NCM components, oracle rpm,…) were developed to automate installation and configuration of databases within the Quattor framework
- Ongoing migration of legacy db's to NAS based infrastructure
- Strategic deployment platform
  - Reduces the number of Storage Technologies to be maintained
  - Avoids complexity of SAN based infrastructure (multipathing, Oracle ASM, etc).

QUESTIONS