INFN CNAF Tier1 Castor Experience

Giuseppe Lo Presti
on behalf of Castor Operation Team at CNAF

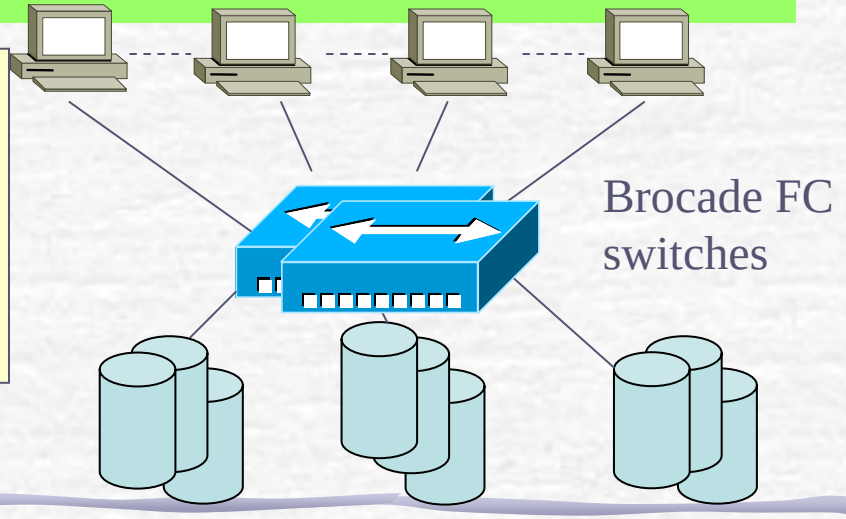
Castor Delta Review, CERN, December 7, 2006

Hardware

Sun Blade v100 with 2 internal ide disks with software raid-1 running ACSLS 7.0 OS Solaris 9.0

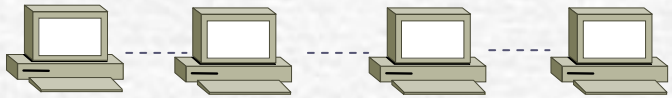
- Core services are on machines with scsi disks, hardware raid1, redundant power supplies
- tape servers and disk servers have lower level hardware, like WNs

~25 disk servers attached to a SAN
full redundancy FC 2Gb/s
connections (dual controller HW and Qlogic SANsurfer Path Failover SW)



Brocade FC switches

- STK L5500 silos (5500 slots, partitioned wth 2 form-factor slots, about 2000 LTO2 for and 3500 9940B, 200GB cartridges, tot capacity ~1.1PB tot non compressed)
- 6 LTO2 + 7 9940B drives, 2 Gbit/s FC interface, 20-30 MB/s rate (some more 9940B going to be acquired in next months).



13 tape servers

STK FlexLine 600, IBM FastT900 …

Core services

CASTOR core services v2.1.0-6 on 4 machines (v2.1.0-8 for clients).

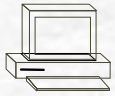
castor-6: rhserver, stager, rtcpclientd, MigHunter, cleaningDaemon

castorlsf01: Master LSF, rmmaster, expert

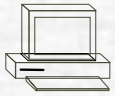
dlf01: dlfserver, Cmonit, Oracle for DLF

castor-8: nsdaemon, vmgr, vdqm, msgd, cupvd

2 more machines for the Name Server and Stager DB



Name server Oracle DB (Oracle 9.2)



Stager Oracle DB (Oracle 10.2)

2 SRMv1 endpoints, DNS balanced:

- srm://castorsrm.cr.cnaf.infn.it:8443 (used for “tape” svc classes)
- srm://sc.cr.cnaf.infn.it:8443 (used for disk-only svc classes for lhcb, cms, atlas and “tape” svc class for the others)

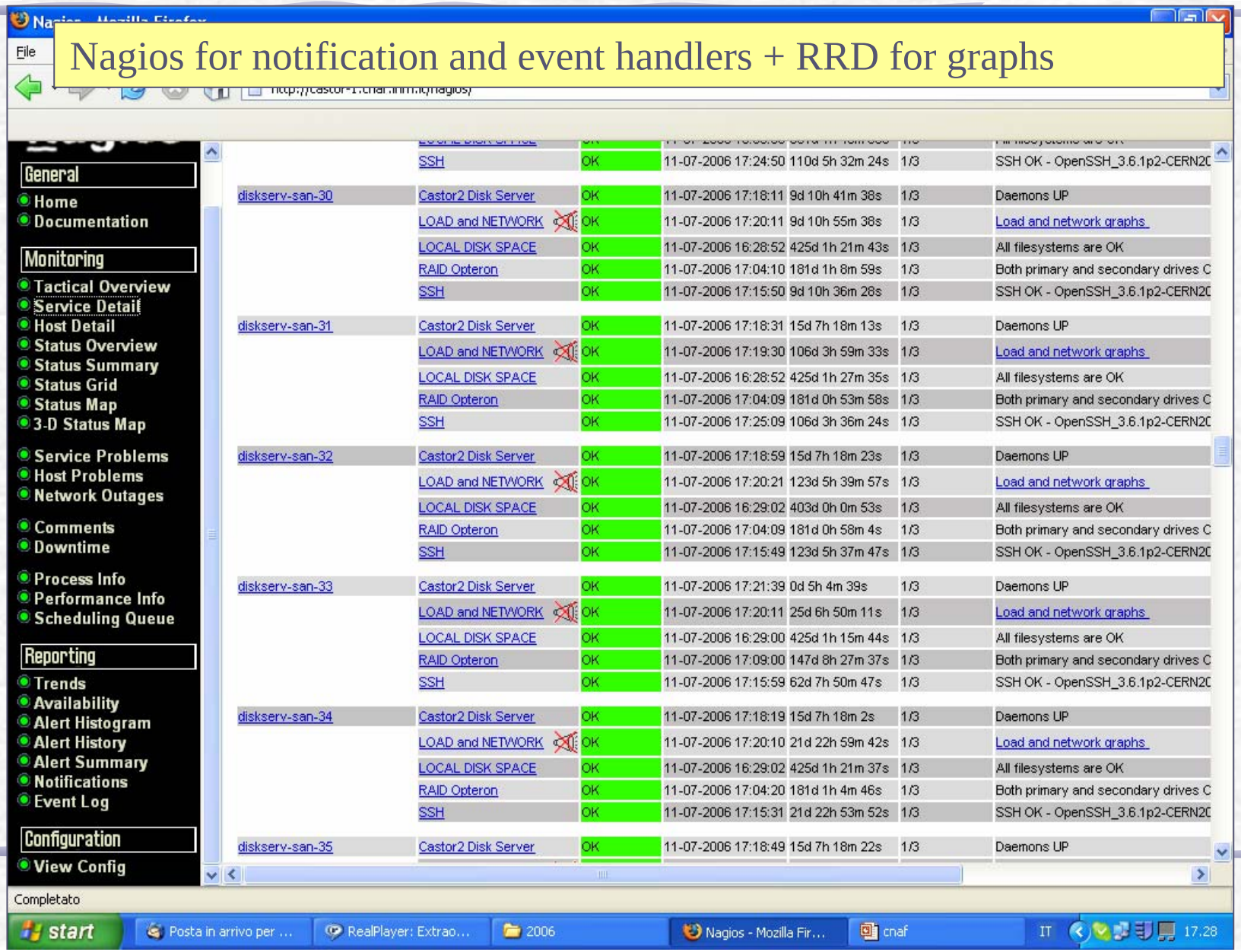
V2.1.1, disk server in TURL

Setup: Supported VOs - Svcclasses - diskpools

Svc class	Exp	Disk pool	Garbage Collct	Size (TB)
alice	ALICE	alice1	yes	19.7
cms	CMS	cms1	yes	28
cmsdisk	CMS	cms1disk	no	29
atlas	ATLAS	atlas1	yes	14
atlasdisk	ATLAS	atlas1disk	no	26
lhcb	LHCb	lhcb1	yes	8.5
lhcbdisk	LHCb	lhcb1disk	no	15.2
dteam		dteam1	yes	0.1
lvd	LVD	archive1	yes	1.6
argo	ARGO	argo1	yes	8.3
argo_download	ARGO	argo2	yes	2.2
virgo	VIRGO	archive1	yes	1.6
ams	AMS	ams1	yes	2.7
pamela	PAMELA	pamela1	yes	3.5
magic	MAGIC	archive1	yes	1.6
babar	BABAR	archive1	yes	1.6
cdf	CDF	archive1	yes	1.6

Monitoring

Nagios for notification and event handlers + RRD for graphs



	SSH	OK	11-07-2006 17:24:50	110d 5h 32m 24s	1/3	SSH OK - OpenSSH_3.6.1p2-CERN2C
diskserv-san-30	Castor2 Disk Server	OK	11-07-2006 17:18:11	9d 10h 41m 38s	1/3	Daemons UP
	LOAD and NETWORK	OK	11-07-2006 17:20:11	9d 10h 55m 38s	1/3	Load and network graphs
	LOCAL DISK SPACE	OK	11-07-2006 16:28:52	425d 1h 21m 43s	1/3	All filesystems are OK
	RAID Opteron	OK	11-07-2006 17:04:10	181d 1h 8m 59s	1/3	Both primary and secondary drives C
	SSH	OK	11-07-2006 17:15:50	9d 10h 36m 28s	1/3	SSH OK - OpenSSH_3.6.1p2-CERN2C
diskserv-san-31	Castor2 Disk Server	OK	11-07-2006 17:18:31	15d 7h 18m 13s	1/3	Daemons UP
	LOAD and NETWORK	OK	11-07-2006 17:19:30	106d 3h 59m 33s	1/3	Load and network graphs
	LOCAL DISK SPACE	OK	11-07-2006 16:28:52	425d 1h 27m 35s	1/3	All filesystems are OK
	RAID Opteron	OK	11-07-2006 17:04:09	181d 0h 53m 58s	1/3	Both primary and secondary drives C
	SSH	OK	11-07-2006 17:25:09	106d 3h 36m 24s	1/3	SSH OK - OpenSSH_3.6.1p2-CERN2C
diskserv-san-32	Castor2 Disk Server	OK	11-07-2006 17:18:59	15d 7h 18m 23s	1/3	Daemons UP
	LOAD and NETWORK	OK	11-07-2006 17:20:21	123d 5h 39m 57s	1/3	Load and network graphs
	LOCAL DISK SPACE	OK	11-07-2006 16:29:02	403d 0h 0m 53s	1/3	All filesystems are OK
	RAID Opteron	OK	11-07-2006 17:04:09	181d 0h 58m 4s	1/3	Both primary and secondary drives C
	SSH	OK	11-07-2006 17:15:49	123d 5h 37m 47s	1/3	SSH OK - OpenSSH_3.6.1p2-CERN2C
diskserv-san-33	Castor2 Disk Server	OK	11-07-2006 17:21:39	0d 5h 4m 39s	1/3	Daemons UP
	LOAD and NETWORK	OK	11-07-2006 17:20:11	25d 6h 50m 11s	1/3	Load and network graphs
	LOCAL DISK SPACE	OK	11-07-2006 16:29:00	425d 1h 15m 44s	1/3	All filesystems are OK
	RAID Opteron	OK	11-07-2006 17:09:00	147d 8h 27m 37s	1/3	Both primary and secondary drives C
	SSH	OK	11-07-2006 17:15:59	62d 7h 50m 47s	1/3	SSH OK - OpenSSH_3.6.1p2-CERN2C
diskserv-san-34	Castor2 Disk Server	OK	11-07-2006 17:18:19	15d 7h 18m 2s	1/3	Daemons UP
	LOAD and NETWORK	OK	11-07-2006 17:20:10	21d 22h 59m 42s	1/3	Load and network graphs
	LOCAL DISK SPACE	OK	11-07-2006 16:29:02	425d 1h 21m 37s	1/3	All filesystems are OK
	RAID Opteron	OK	11-07-2006 17:04:20	181d 1h 4m 46s	1/3	Both primary and secondary drives C
	SSH	OK	11-07-2006 17:15:31	21d 22h 53m 52s	1/3	SSH OK - OpenSSH_3.6.1p2-CERN2C
diskserv-san-35	Castor2 Disk Server	OK	11-07-2006 17:18:49	15d 7h 18m 22s	1/3	Daemons UP

First CASTOR2 production experiences

- Migration from CASTOR1 to CASTOR2 started at the beginning of SC4
  - First dteam Jen-March 2006. Good results during throughput phase, 180 MB/s disk-disk and 70 MB/s disk-tape sustained ☺
  - other LHC VOs, Jun-Jul 2006. Their castor1 stagers were already at 200-300k entries... Could not face the service phase.
  - non-LHC exps, Jul-Sep 2006
- When the exp requested to keep their CASTOR1 files on disks we used the script written by G. Lo Presti for the registration of CASTOR1 files into the CASTOR2 stager. No files moved from disks. ☺
- In Sep we had many problems, mostly due to a lack of knowledge of the basic management procedure of the stager tables (more difficult than expected) and to software bugs and packaging error, ex:
  - Garbage collector not working properly (still now we see deadlocks ...☹) and sometime causing a quick filling of our disk pools)
  - CERN customized tmpwatch non present in the CASTOR2 rpm repository, that caused many tape servers failures. ☹
- At mid-Sep the introduction into castor.conf of an alias of the name server caused a complete block, solved at end of Sep, thanks to Giuseppe, Olof, Sebastien.

Now CASTOR2 is up and running by the end of Sep with no long/critical outages. ☺

Open issues

- Monitoring/notification: what we have seems not to be enough, investigating Lemon
- Dedication of drives to an exp (easy to do with CASTOR1 on the GID basis...)
- Handling more than one svcclass (=disk pool for us) is very difficult via GRID with SRMv1, the two endpoint approach (one disk-only and the other with tape backend) is not easy to manage. SRMv2 should help, only one srm endpoint.
- Disk servers load: gridftp transfers vs rfopen... more disk servers and disk pools needed
- Currently the Rmmaster scalability limits and the prepareToGet implementation makes the stage-in operations very difficult. At the moment LHCb recostruction at CNAF is suffering a lot from this.

Near future plans

- Upgrade servers from V2.1.0-6 to V2.1.1-x
- SRM2.2 deployment
- Configure another stager, to be used for testing (ex. dteam transfers), for SRM2 and for tape repack operation.