Storage classes in Castor at Cern

Jan van Eldik
Castor operations team
CERN/IT/FIO

Outline

- Deployment overview
  - Castor concepts
  - Castor stagers with their storage classes for the LHC experiments
- Storage Classes
  - disk0tape1
  - disk1tape0
  - disk1tape1

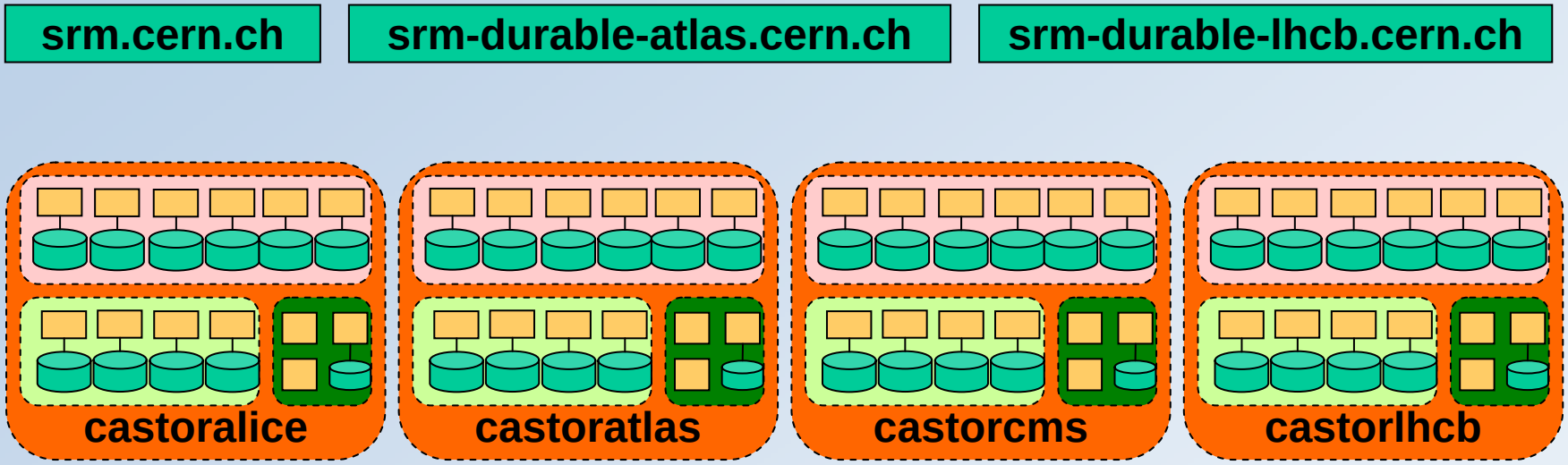
Storage Classes in Castor

- Service Class == Storage Class
- Describes activity: *CDR, Analysis, Reconstruction, …*
- Has attributes:
  - Tape migration policies
    - # copies on tape (zero or more) determined by fileclass, ie filename
  - Garbage collection policy
    - Combination of last acces time, file size
  - Access protocols: rfio, rootd
    - Gsiftp not a native protocol yet, currently on top of rfio
- Combines diskpools
  - Groups of filesystems
  - Castor can recall files from other diskpools
    faster than recalling from tape
  - Also internal replication for hot files
  - CERN deployment choices *(for reasons of simplicity):*
    - Diskpool == Service Class == Storage Class
    - All filesystems of a server belong to same Diskpool

Castor-2 deployment at Cern

- Common tape infrastructure
  with tape drive allocations
- Common Castor nameserver
  to manage namespace and tape segments
- Dedicated independent diskcaches per LHC VO
  - ensures that tapes have data from one VO only ☺
  - …and a public stager for others
- Service classes sized and configured for needs of experiments *(we hope…)*
  - *svcClass names will be SRM v22 spaceTokens*
- All service classes support all access protocols
  incl gsiftp: all service classes are WAN-enabled
- Shared SRM v1 endpoint srm.cern.ch
  …and 2 dedicated endpoints for Atlas and LHCb

Castor-2 setup at Cern



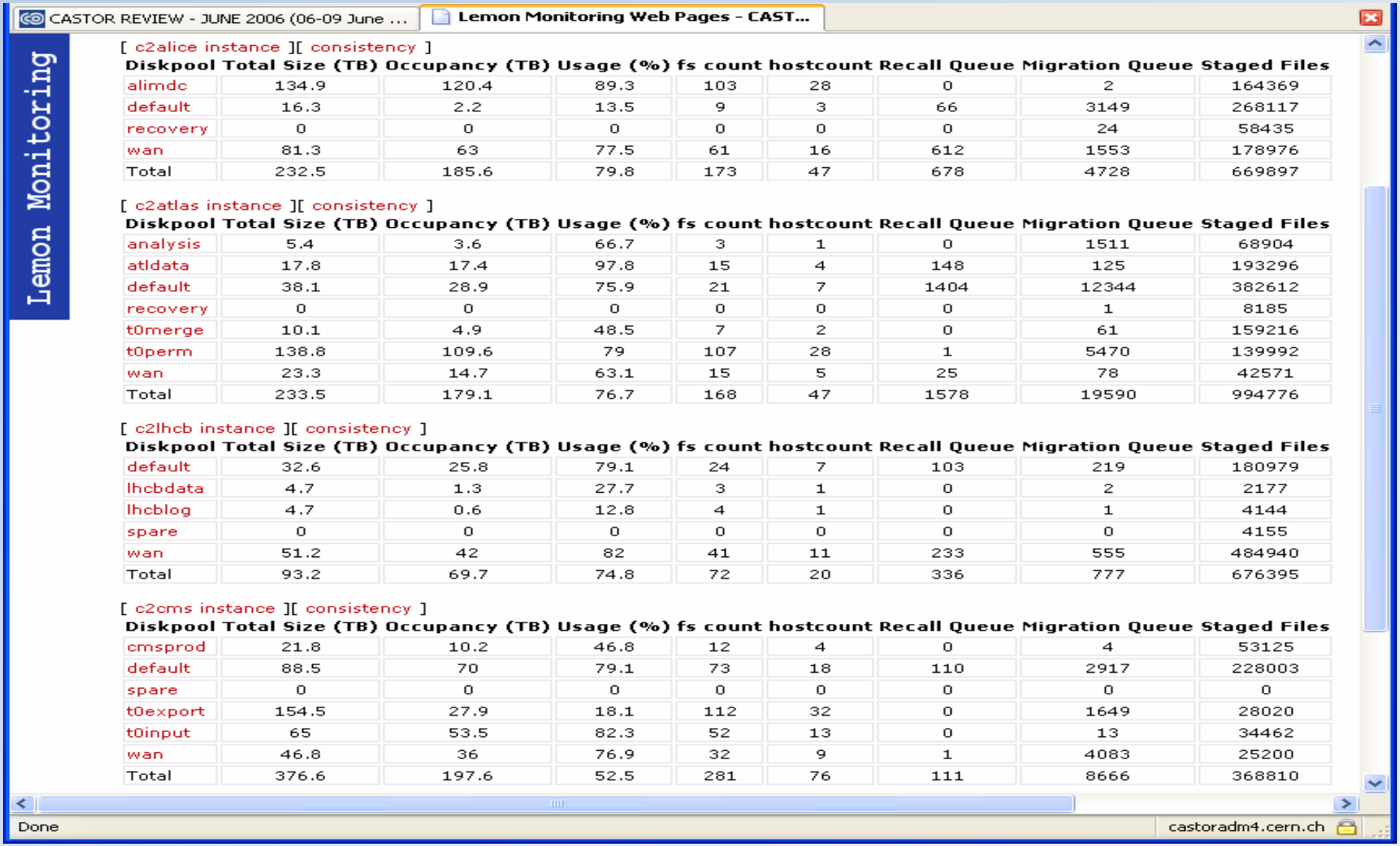
shared tape infrastructure, including

STK T10000 and IBM 3592B drives

shared nameserver

database

Configured SvcClasses



[c2alice instance][consistency]

Diskpool	Total Size (TB)	Occupancy (TB)	Usage (%)	fs count	hostcount	Recall Queue	Migration Queue	Staged Files
alimdc	134.9	120.4	89.3	103	28	0	2	164369
default	16.3	2.2	13.5	9	3	66	3149	268117
recovery	0	0	0	0	0	0	24	58435
wan	81.3	63	77.5	61	16	612	1553	178976
Total	232.5	185.6	79.8	173	47	678	4728	669897

[c2atlas instance][consistency]

Diskpool	Total Size (TB)	Occupancy (TB)	Usage (%)	fs count	hostcount	Recall Queue	Migration Queue	Staged Files
analysis	5.4	3.6	66.7	3	1	0	1511	68904
atldata	17.8	17.4	97.8	15	4	148	125	193296
default	38.1	28.9	75.9	21	7	1404	12344	382612
recovery	0	0	0	0	0	0	1	8185
t0merge	10.1	4.9	48.5	7	2	0	61	159216
t0perm	138.8	109.6	79	107	28	1	5470	139992
wan	23.3	14.7	63.1	15	5	25	78	42571
Total	233.5	179.1	76.7	168	47	1578	19590	994776

[c2lhcb instance][consistency]

Diskpool	Total Size (TB)	Occupancy (TB)	Usage (%)	fs count	hostcount	Recall Queue	Migration Queue	Staged Files
default	32.6	25.8	79.1	24	7	103	219	180979
lhcbdata	4.7	1.3	27.7	3	1	0	2	2177
lhcblog	4.7	0.6	12.8	4	1	0	1	4144
spare	0	0	0	0	0	0	0	4155
wan	51.2	42	82	41	11	233	555	484940
Total	93.2	69.7	74.8	72	20	336	777	676395

[c2cms instance][consistency]

Diskpool	Total Size (TB)	Occupancy (TB)	Usage (%)	fs count	hostcount	Recall Queue	Migration Queue	Staged Files
cmsprod	21.8	10.2	46.8	12	4	0	4	53125
default	88.5	70	79.1	73	18	110	2917	228003
spare	0	0	0	0	0	0	0	0
t0export	154.5	27.9	18.1	112	32	0	1649	28020
t0input	65	53.5	82.3	52	13	0	13	34462
wan	46.8	36	76.9	32	9	1	4083	25200
Total	376.6	197.6	52.5	281	76	111	8666	368810

disk0tape1

- Classic case, Castor is designed for this
- Service Class with
  - Garbage Collector YES
  - Tape Migration YES
- Details of these policies vary between Service Classes
  especially tape migration
- Most of our Service Classes are disk0tape1

disk1tape0

- No Garbage Collector allowed
  - VO's are expected to manage space themselves
- No tape migration required
  - but at Cern, we do it anyway ☺
  - again, for operational simplicity:
    - To operate the setup in a transparent way
    - In case of H/W problems with the diskcache, it is easier to retrieve from tape than to replicate from other sites
    - And it is just a small fraction of tapespace…
      - Today, 5 PB tapespace, with 30TB disk1tape0
  - Castor does not handle NoTape files very well

At CERN, disk1tape0 → disk1tape1

disk1tape1

- Tape Migration YES
- Garbage collection NO
  - VO's are expected to manage space themselves
- Atlas and LHCb have such SvcClasses
  - atldata, lhcbdata
  - VO data managers write data, physicists read-only but Castor does not enforce this
  - Overfull diskpools cause problems…
- Require dedicated SRM-v1 endpoints
  - srm-durable-{atlas,lhcb}.cern.ch
  - Fixed with spaceToken in SRM v22

Conclusion

- Castor2 at Cern has complex configuration
  - Tailored for the experiment activities
- We make deployment choices to avoid trouble
  - VO independent diskcaches
  - Disk1tape0 == disk1tape1 at CERN
- spaceTokens in SRM v2 will simplify SRM deployment
- Main worry: management of diskcache in non-garbaged collected disk1tape1