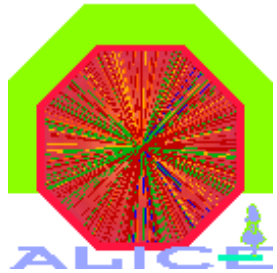
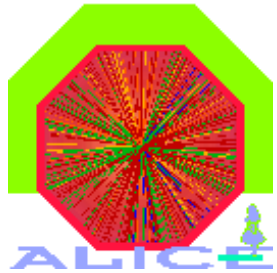


Staging to CAF
+
User groups + fairshare

Jan Fiete Grosse-Oetringhaus, CERN PH/ALICE

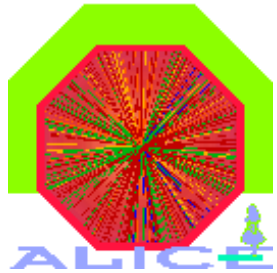
Offline week, 08.04.08

Datasets in Practice



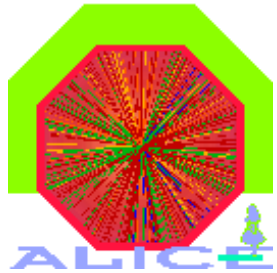
- Create a AliEn collection (in aliensh)
  - find -c myCollection /alice/sim/2007/LHC07c/pp_minbias/8051 root_archive.zip
  - find -c myCollection /alice/sim/2007/LHC07c/pp_minbias/8051 AliESDs.root
- Use a ROOT version that supports datasets
  - LXPLUS: source /afs/cern.ch/alice/caf/caf-lxplus-datasets.sh
  - OR: Check out from ROOT SVN: branches/dev/proof
- Create DS from AliEn collection
  - Connect to AliEn
    - TGrid::Connect("alien://")
  - gridColl = gGrid->OpenCollection("alien:///alice/cern.ch/user/j/jgrosseo/myCollection")
  - proofColl = gridColl->GetFileCollection();
  - proofColl->SetAnchor("AliESDs.root"); // collection of root_achive.zip

Datasets in Practice (2)



- Upload to PROOF cluster
  - Connect to PROOF
  - TProof::Mgr("lxb6046")->SetROOTVersion("vPROOFDSMGR");
  - TProof::Open("lxb6046");
  - gProof->RegisterDataSet("myDataSet", proofColl);
- Check status
  - gProof->ShowDataSets();
- Use it
  - mgr->StartAnalysis("proof", "myDataSet");

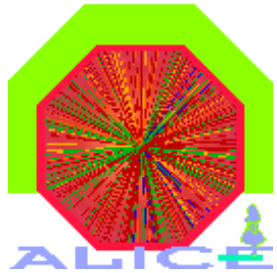
Dataset in Practice (3)



- List available datasets: gProof->ShowDataSets()
- You always see common datasets and datasets of your group

Dataset URI	# Files	Default tree	# Events	Disk	Staged
/COMMON/COMMON/ESD5000	21899	/esdTree	2121800	1 TB	96 %
/COMMON/COMMON/ESD5000_small	100	/esdTree	9700	4 GB	97 %
/COMMON/COMMON/ESD600	1844	/esdTree	178600	51 GB	96 %
/COMMON/COMMON/run15034_PbPb	967	/esdTree	639	468 GB	66 %
/COMMON/COMMON/run15035_PbPb	962	/esdTree	579	454 GB	60 %
/COMMON/COMMON/run15036_PbPb	961	/esdTree	588	462 GB	61 %
/COMMON/COMMON/run15037_PbPb	965	/esdTree	609	479 GB	63 %
/COMMON/COMMON/run82XX_part1	10000	/esdTree	759000	289 GB	75 %
/COMMON/COMMON/run82XX_part2	10000	/esdTree	652000	288 GB	65 %
/COMMON/COMMON/run82XX_part3	10000	/esdTree	713400	288 GB	71 %
/PWG2/COMMON/run82XX_test4	10	N/A		297 MB	0 %
/PWG2/COMMON/run82XX_test5	10	N/A		297 MB	0 %
/PWG2/hricaud/LHC07f_160033DataSet	915	/esdTree	72800	2 GB	79 %
/PWG2/hricaud/LHC07f_160038_root_archiveDataSet	862	/esdTree	36900	433 GB	42 %
/PWG2/hricaud/LHC07f_16004xDataSet	4643	/esdTree	74900	12 GB	16 %
/PWG2/jgrosseo/ESD5000_small	100	/esdTree	9800	4 GB	98 %
/PWG2/jgrosseo/ESDBlind01	364	/esdTree	10325	817 MB	30 %
/PWG2/jgrosseo/run12000	62	/esdTree	50	5 GB	80 %
/PWG2/jgrosseo/run12001	967	N/A		104 GB	0 %
/PWG2/jgrosseo/run2003XX	12600	/HLTesdTree	297600	124 GB	29 %
/PWG2/mvala/RSNMV_PDC07_09_part1	326	/rsnMVTree	1715829	5 GB	58 %
/PWG2/mvala/RSNMV_PDC07_09_part1_new	326	/rsnMVTree	329737	5 GB	11 %

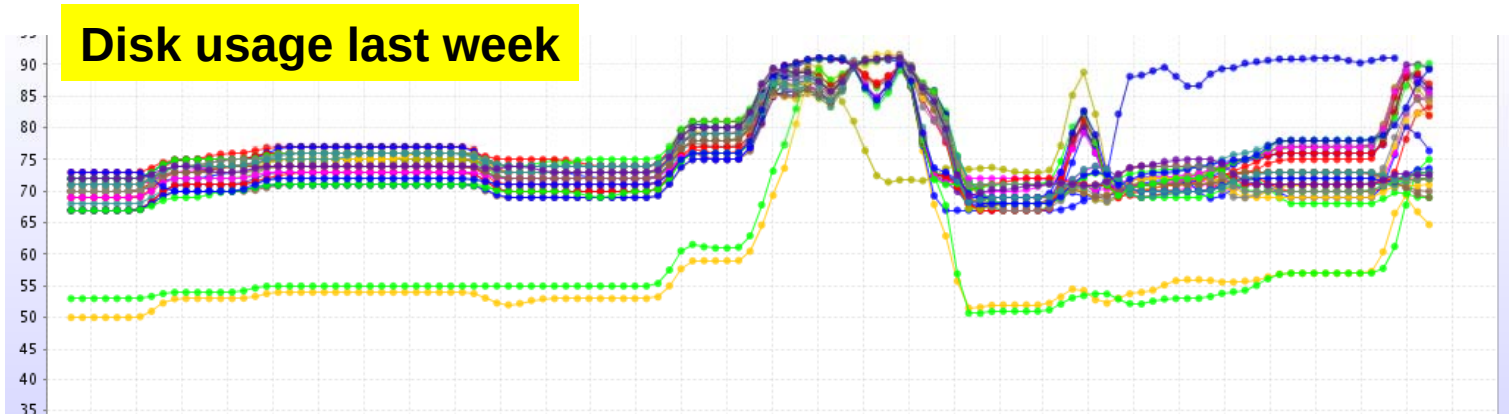
Dataset in Practice (4)



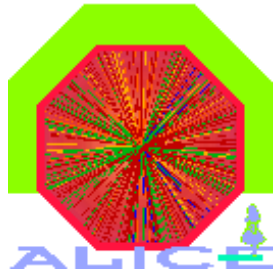
Dataset URI	# Files	Default tree	# Events	Disk	Staged
/COMMON/COMMON/ESD5000	21899	/esdTree	2121800	1 TB	96 %
/COMMON/COMMON/ESD5000_small	100	/esdTree	9700	4 GB	97 %
/COMMON/COMMON/ESD600	1844	/esdTree	178600	51 GB	96 %
/COMMON/COMMON/run15034_PbPb	967	/esdTree	639	468 GB	66 %
/COMMON/COMMON/run15035_PbPb	962	/esdTree	579	454 GB	60 %
/COMMON/COMMON/run15036_PbPb	961	/esdTree	588	462 GB	61 %
/COMMON/COMMON/run15037_PbPb	965	/esdTree	609	479 GB	63 %
/COMMON/COMMON/run82XX_part1	10000	/esdTree	759000	289 GB	75 %
/COMMON/COMMON/run82XX_part2	10000	/esdTree	652000	288 GB	65 %
/COMMON/COMMON/run82XX_part3	10000	/esdTree	713400	288 GB	71 %
/PWG2/COMMON/run82XX_test4	10	N/A		297 MB	0 %
/PWG2/COMMON/run82XX_test5	10	N/A		297 MB	0 %
/PWG2/hricaud/LHC07f_160033DataSet	915	/esdTree	72800	2 GB	79 %
/PWG2/hricaud/LHC07f_160038_root_archiveDataSet	862	/esdTree	36900	433 GB	42 %
/PWG2/hricaud/LHC07f_16004xDataSet	4643	/esdTree	74900	12 GB	16 %
/PWG2/jgrosseo/ESD5000_small	100	/esdTree	9800	4 GB	98 %
/PWG2/jgrosseo/ESDBlind01	364	/esdTree	10325	817 MB	30 %
/PWG2/jgrosseo/run12000	62	/esdTree	50	5 GB	80 %
/PWG2/jgrosseo/run12001	967	N/A		104 GB	0 %
/PWG2/jgrosseo/run2003XX	12600	/HLTesdTree	297600	124 GB	29 %
/PWG2/mvala/RSNMV_PDC07_09_part1	326	/rsnMVTree	1715829	5 GB	58 %
/PWG2/mvala/RSNMV_PDC07_09_part1_new	326	/rsnMVTree	329737	5 GB	11 %

Status

- First users started staging
- Bug fix in CASTOR needed → deployed this Monday
- Problem with files contained in zip archives → in progress



User Groups



- The available disk is shared between groups using quotas
- The priority of concurrent processes is governed by a CPU fairshare mechanism
- Groups are e.g. PWGx, TPC, ACO

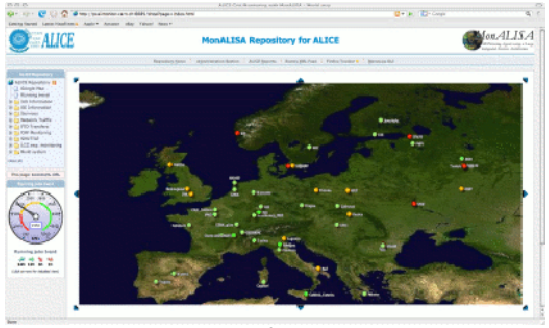
→It is needed that you provide your group (private mail to me or ATF list)

  – Otherwise you cannot stage data + your relative priority will be low

- ~40 users in 14 groups

CPU Fairshare

Get groups' usage. Interval defined per each one: [α*quota..β*quota]



measure difference between real usages and quotas

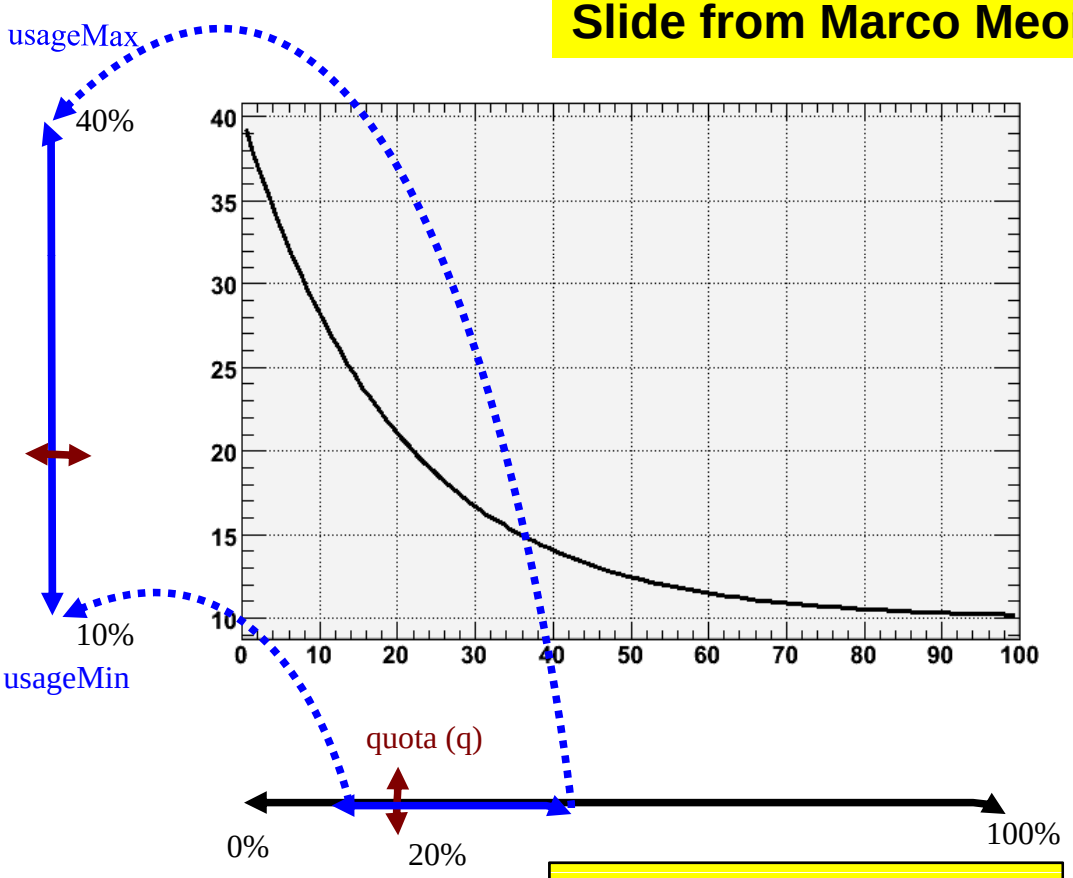
Compute new usages applying a correction formula

CAF

Store computed usages

- Average every 6 hours
- Retrieved every 5 mins

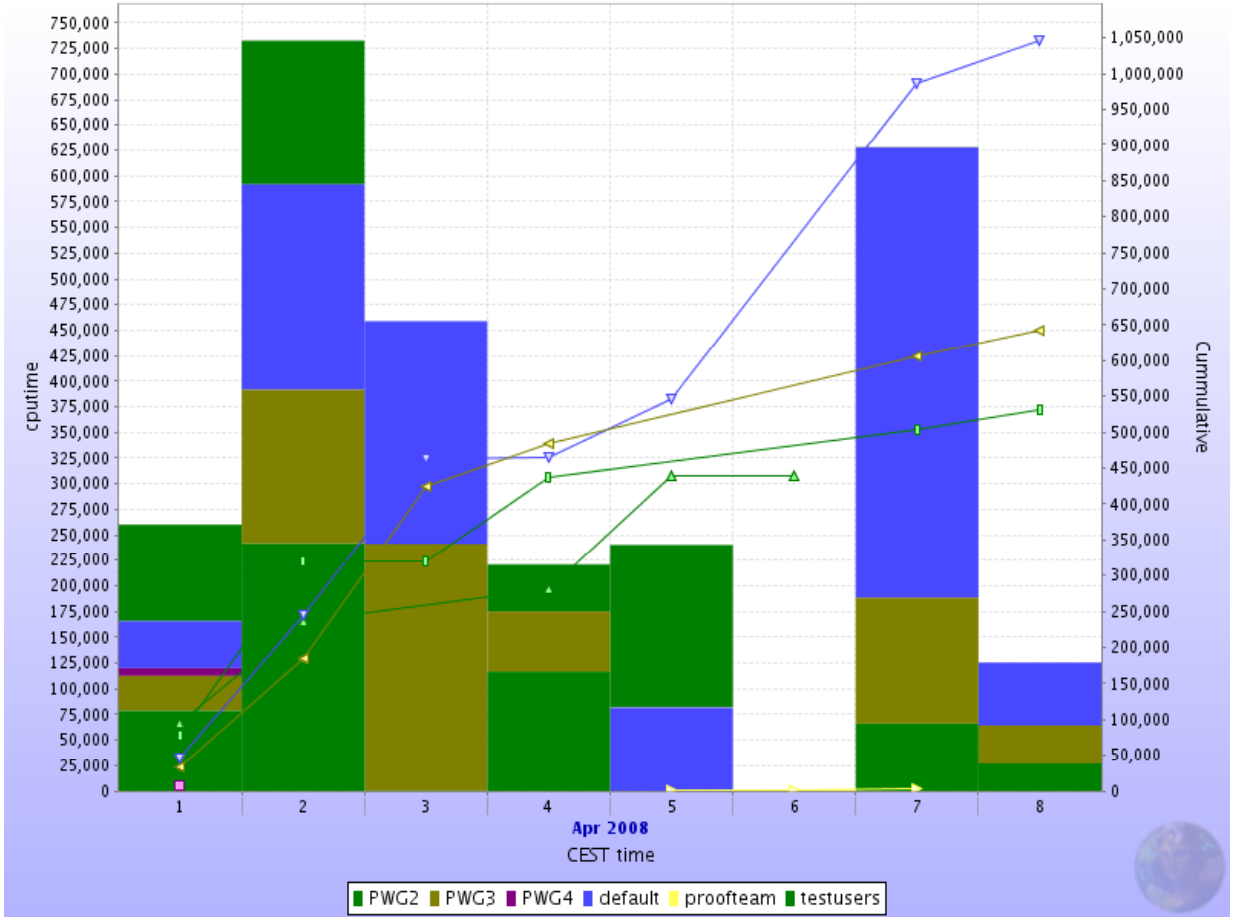
Slide from Marco Meoni



$f(x) = \alpha q + \beta q * \exp(kx)$

$k = 1/q * \mathrm{Ln}(1/4)$

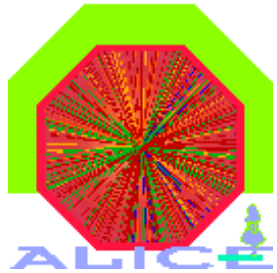
Usage



Input to PROOF
default=5.00045
EMCAL=20
PHOS=20
proofteam=20
ITS=20
MUON=20
PWG0=20
ZDC=20
PWG1=20
PWG2=20
PWG3=18.9864
T0=20
PWG4=20

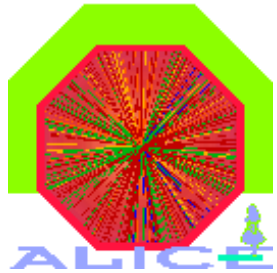
Values are aggregated over one month.
Currently, dominated by usage in group "default" in the last weeks.

Summary

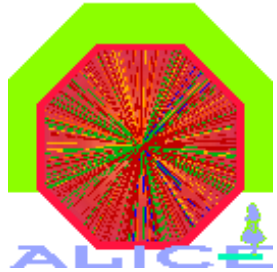


- Staging + Disk quotas
  - User staging slowly starting
  - Still problems with the xrootd client in ROOT (crashes, stalling)
    → causes files to disappear, are restaged automatically
- CPU Fairshare
  - Users assigned to groups
  - Mechanism running
    → Priorities that are fed into PROOF will also be published in MonaLisa

Backup

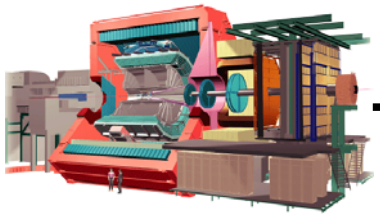


Good User Practices



- Before you start using CAF
  - Subscribe to alice-project-analysis-task-force@cern.ch using CERN SIMBA (http://listboxservices.web.cern.ch/listboxservices)
  - Read http://aliceinfo.cern.ch/Offline/Analysis/CAF
- Code development
  - Try your code on at least 2 files locally
    - 1 file may hide problems when switching to the next file
  - Run your code "as in PROOF"
    - Just change "proof" to "local" in StartAnalysis
  - Run "full PROOF"
- Don't use TProof::Reset if it is not needed (current issue)

CAF Schema



AliEn SE

CASTOR MSS

Disk Buffer

Tape

Tier-1
data export

Staging

CAF computing cluster

Proof master,
xrootd redirector

Proof worker

local disk

Proof worker

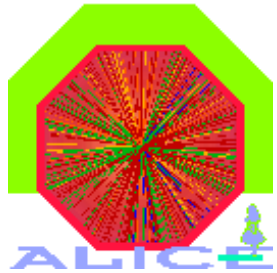
local disk

Proof worker

local disk

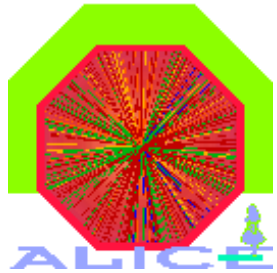
...

Staging – Technical side



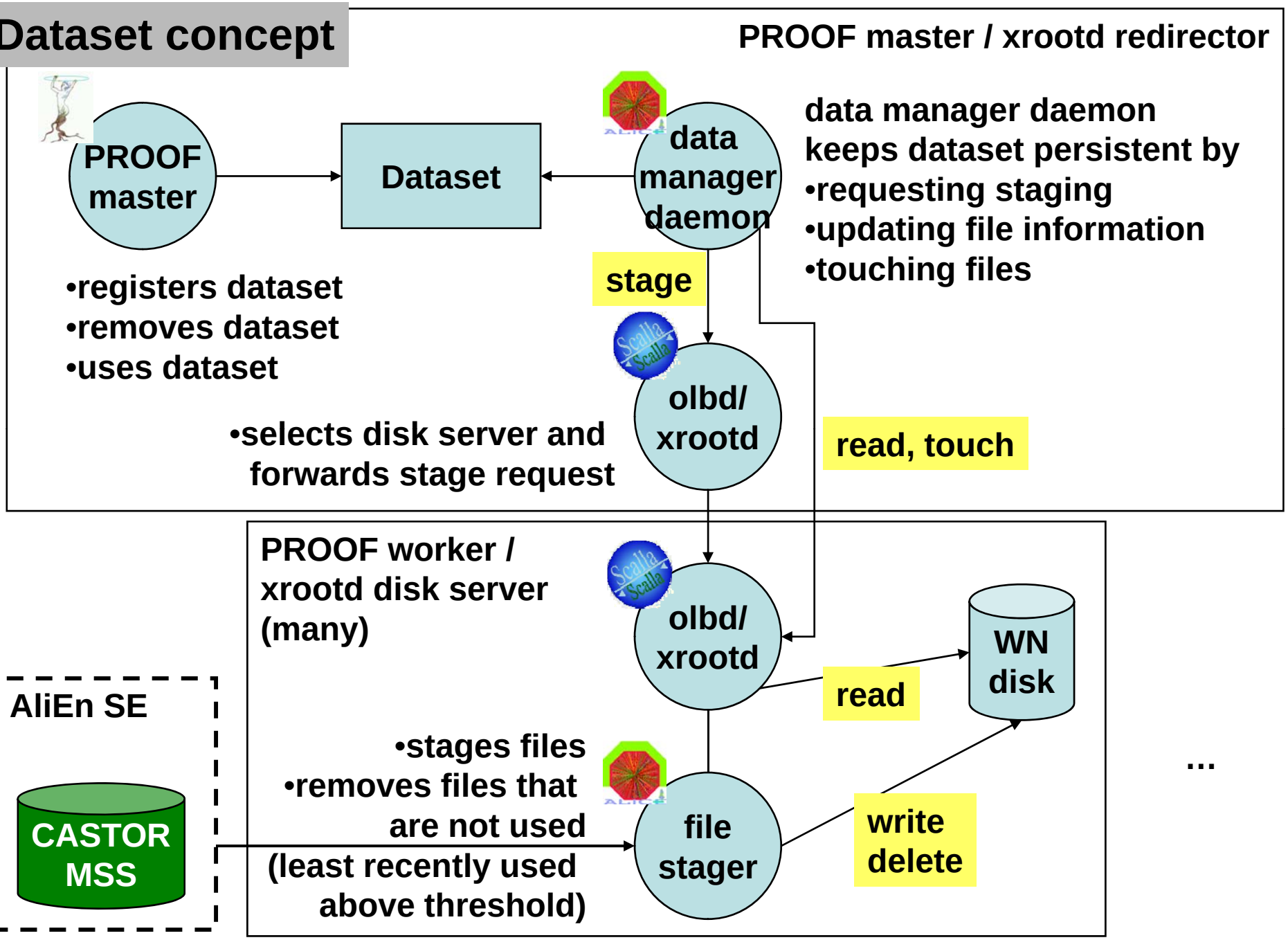
- Step 3 (now): Automatic
  - Staging script plugged into olbd
  - Implementation of PROOF datasets (by ALICE)
  - Staging daemon that runs on the cluster
  - Transparent migration from AliEn collection to PROOF datasets
  - Convenient for users, quota-enabled, garbage collection

Introduction of PROOF datasets

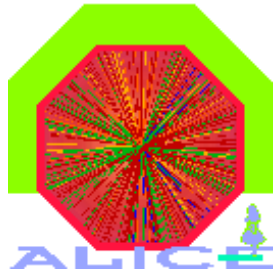


- A dataset represents a list of files (e.g. physics run X)
  - Correspondence between AliEn collection and PROOF dataset
- Users register datasets
  - The files contained in a dataset are automatically staged from AliEn (and kept available)
  - Datasets are used for processing with PROOF
    - Contain all relevant information to start processing (location of files, abstract description of content of files)
- File-level storing by underlying xrootd infrastructure
- Datasets are public for reading (you can use datasets from anybody!)
- There are common datasets (for data of common interest)

Dataset concept

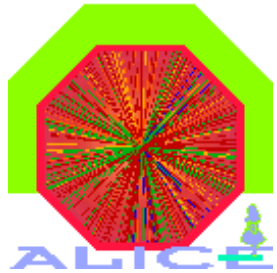


Staging script



- Two directories configured in xrootd/olbd for staging
  - /alien
  - /castor
- Staging script given with olb.prep directive
  - Perl script that consists of 3 threads
  - Front-End: Registers stage request
  - Back-End
    - Checks access privileges
    - Triggers migration from tape (CASTOR, AliEn)
    - Copies files, notifies xrootd
  - Garbage collector: Cleans up following policy file with low/high watermarks (least recently used above threshold)

Data manager daemon



- Keeps content of datasets persistent on disk
- Regularly loops over all datasets
- Sends staging requests for new files
- Extracts meta data from recently staged files
- Verifies that all files are still available on the cluster (by touch, prevents garbage collection)
  - Speed: 100 files / s