Xrootd setup

An introduction/tutorial for ALICE sysadmins

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Outline

- Intro
  - Basic notions
  - Make sure we know what we want
- Setup of an xrootd cluster
- Configuration for ALICE usage
- A few words on data migrations

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Introduction

We must know what we want and the possibilities

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The historical Problem: data access

- Physics experiments rely on rare events and statistics
  - Huge amount of data to get a significant number of events
    - The typical data store can reach 5-10 PB… now
    - Millions of files, thousands of concurrent clients
  - The transaction rate is very high
    - Not uncommon O(10^3) file opens/sec per cluster
      - Average, not peak
      - Traffic sources: local GRID site, local batch system, WAN
    - Up to O(10^3) clients per server!
    - If not met then the outcome is:
      - Crashes, instability, workarounds, “need” for crazy things
- Scalable high performance direct data access
  - No imposed limits on performance and size, connectivity
  - Higher performance, supports WAN direct data access
  - Avoids WN under-utilization
  - No need to do inefficient local copies if not needed
    - Do we fetch entire websites to browse one page?

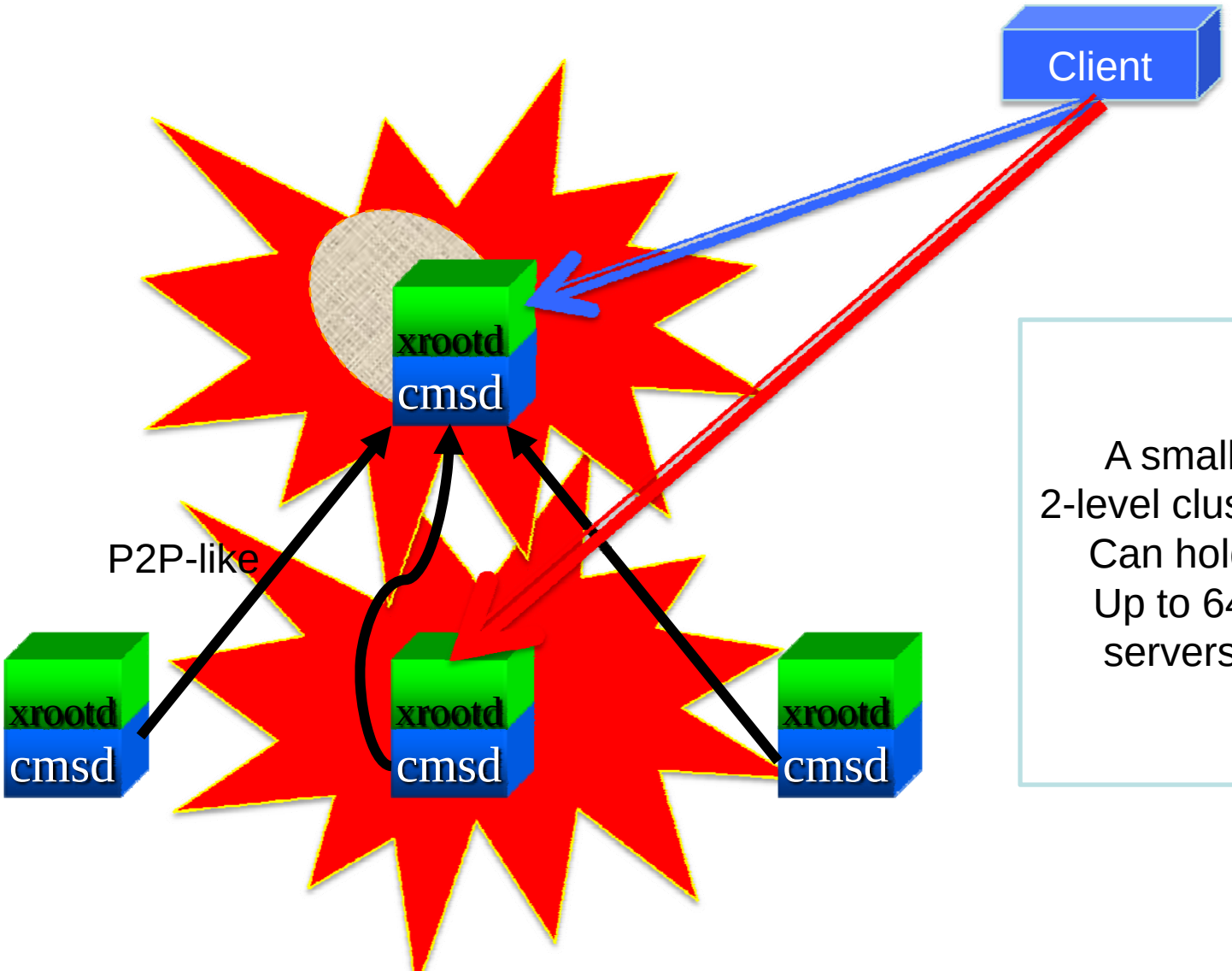
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Main requirement

- Data access has to work reliably at the desired scale
  - This also means:
    - It has to work, simpler is better
    - It has not to waste resources
  - This includes the cluster's dimensioning
    - Very tricky topic for any kind of system
    - With xrootd you are allowed to focus at 99% at the hw only (disks + network)
    - A 2-4 core CPU per server usually is fine these times

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

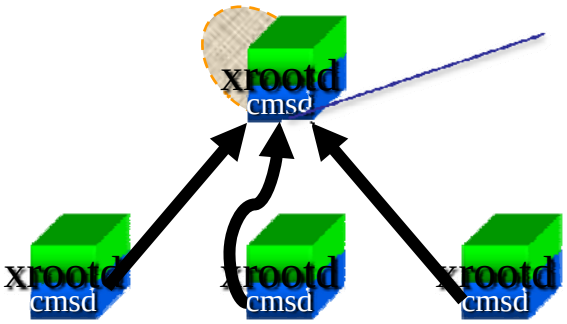
Basic working principle



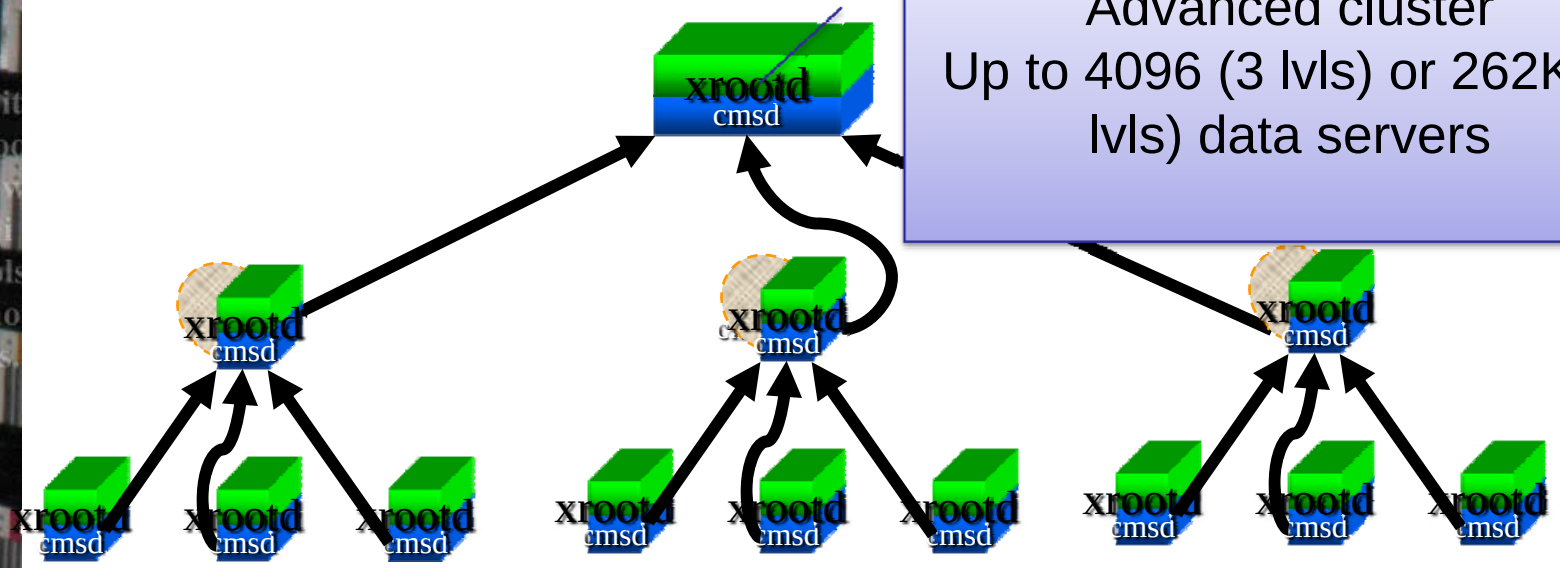
A small
2-level cluster.
Can hold
Up to 64
servers

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Simple LAN clusters



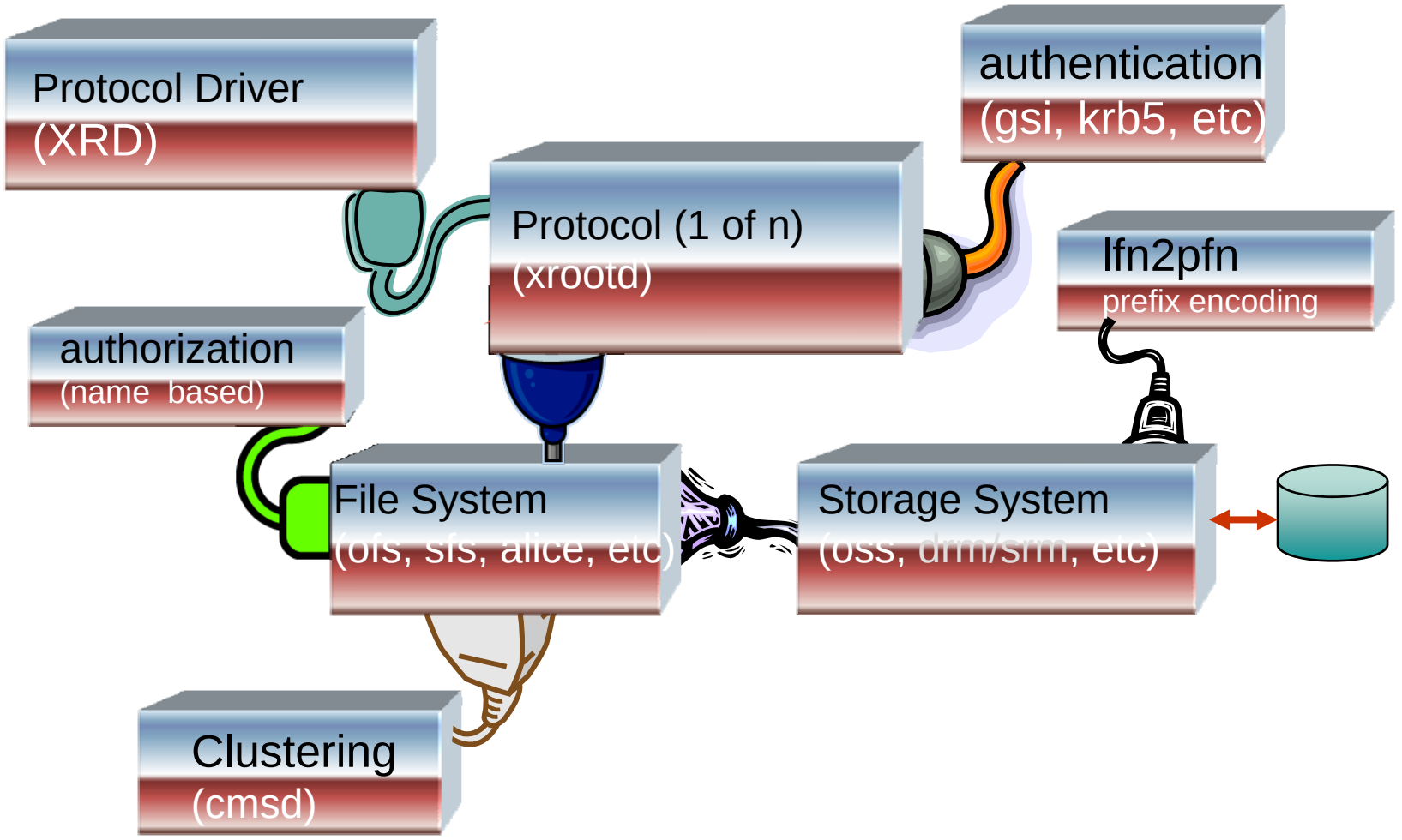
Simple cluster
Up to 64 data servers
1-2 mgr redirectors



Advanced cluster
Up to 4096 (3 lvls) or 262K (4 lvls) data servers

Everything can have hot spares

xrootd Plugin Architecture



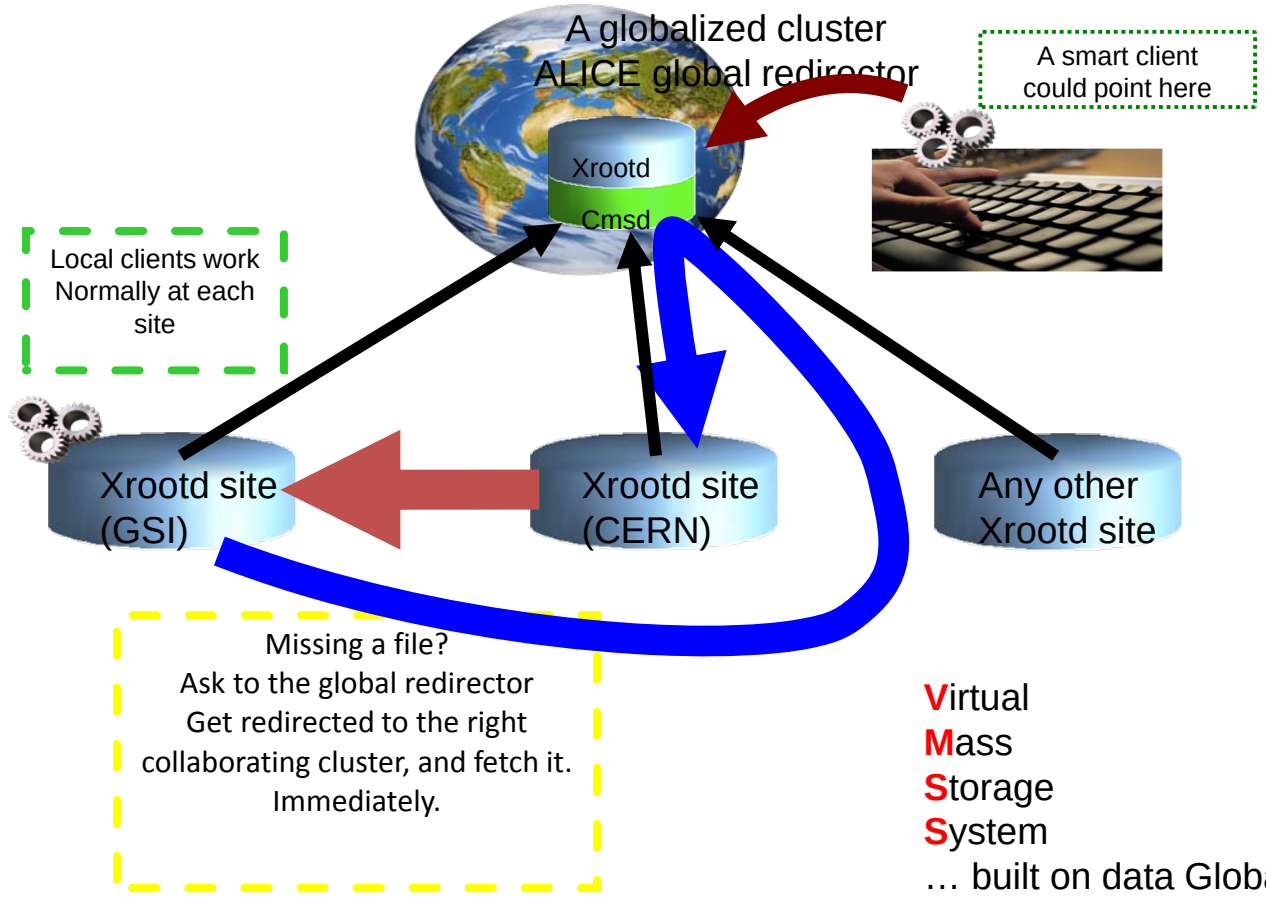
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Pure xrootd flavors

- Standardized all-in one setup
  - Supports up to 64 servers
  - ALICE xrootd-based sites use this
    - Except NIHAM
- Manual setup
  - Always possible
  - Need for deeper knowledge
    - Used for PROOF
    - Other experiments (BaBar, Fermi [GLAST], Some Atlas US sites, …)
- Integrations/bundles (not treated here)
  - DPM, CASTOR
    - Described in their own documentation
  - GPFS+STORM
    - Nothing special is really needed from the xrootd point of view

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The ALICE way with XROOTD



Virtual
Mass
Storage
System
… built on data Globalization

- Pure Xrootd + ALICE strong authz plugin. No difference among T1/T2 (only size and QOS)
- WAN-wide globalized deployment, very efficient direct data access
- CASTOR at Tier-0 serving data, Pure Xrootd serving conditions to the GRID jobs
- “Old” DPM+Xrootd in several tier2s

More details and complete info in “Scalla/Xrootd WAN globalization tools: where we are.” @ CHEP09

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

LFNs and PFNs in ALICE

- Not so trivial… here's the path:
  - Alien LFN … fed to the Alien FC →
  - Alien PFN, in two flavors
    - `<PFN prefix>/<pathname made of ugly numbers>`
    - `<pathname made of ugly numbers>`
    - The Alien PFN is what is asked to the storage
  - The Alien PFN is given as an Xrootd LFN
  - Xrootd name translation
    - If the site has a local Alien PFN prefix (DEPRECATED), add it (the ALICE GLOBAL name xlation)
    - Regular Xrootd translation (adds a prefix, i.e. the mountpoint name)
  - Xrootd PFN
    - When you migrate/maintain you deal with these!

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The name translation

- The xrootd name translation is very simple
  - It adds a prefix in front of everything
    - We call it “localroot”
  - In practice, it is the name of the directory where the files are stored
  - An empty prefix is not admitted
    - Admins are supposed to choose safe places

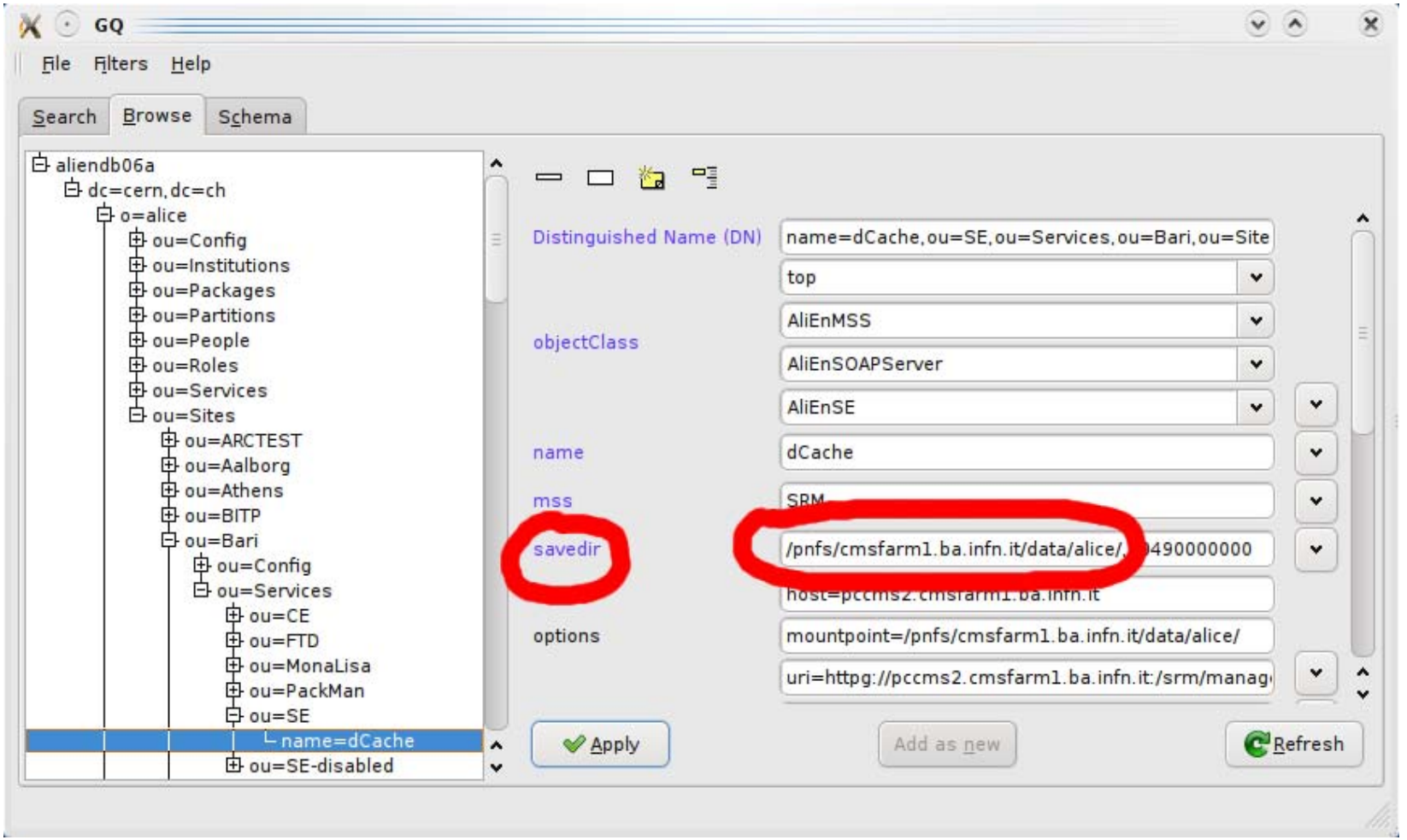
The ALICE Global name translation

- A complex thing which translates in a super-simple rule
  - Look in the ALICE LDAP for your SE
  - If there is a prefix for the data local to your cluster, then your cluster cannot be ok for the global redirector
    - Copy the prefix
    - We well call this “local path prefix” and put it in the xrootd config
  - If xrootd sees a LFN which does not have the “local path prefix”, it will internally add it.
    - This is because any cluster has to expose a “global namespace” (so, without exposing weird local prefixes to other sites)
      - But the old stuff has still to work
    - Stay tuned, more info later

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The ALICE Global name translation

CERN IT Department



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The ALICE Global name translation

- That prefix was a relic of the past
  - When the concepts of “local”, “physical”, “name translation”, were not very clear
- New ALICE SEs must have it empty
  - Because they export a coherent name space to the central services
  - In other words: An Alien PFN must look the same in every site
- The purpose of this additional tricky translation (LOCALPATHPFX) is to accommodate this also for older sites

The cache file system

- A lot of people seem to choke in front of this
- With the xrootd terminology we call this “oss.cache”
- Purpose: aggregate several mountpoints in a data server
  - They MUST be exported as an unique thing
  - The rest of the computing MUST NOT deal with local choices
    - E.g. the name of the disks/directories in a machine
    - These are information which MUST remain local
- Common errors
  - Putting data files in /
  - Putting together data and namespace (i.e. localroot)
    - It works, but it’s messy and difficult to manage when you need it

The cache file system

- A unique directory contains the “name space”
  - A directory tree which contains symlinks
  - Meaningful name -> /mydisk1/data/xrdnamespace
  - This is the xrootd internal prefix (localroot)
- The other N directories contain the data files
  - Slightly renamed (not our business)
  - Meaningful names:
    - /mydisk1/data/xrddata
    - /mydisk2/data/xrddata
- Example:
  - /mydisk1/data/xrdnamespace/my/little/file.txt →
  - /mydisk1/data/xrddata/my%little%file.txt
    - .. And the picture should be more clear now

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Setup

A view on the procedure and the options
Rules of thumb where possible

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Root or xrootd user?

- It depends on the site requirements
  - Internally xrootd runs ALWAYS as a normal user
- If you need an RPM you need root access
  - You may have Quattor
- The configuration does not need to reflect this
  - There are the usual pros and cons. Typically:
    - Normal user setup: easier to maintain for small installations. Less downtime needed for operations.
    - Root setup: very nice for RPM-based automated setups

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

A note on the file descriptors

- The purpose of a server is to handle clients
  - Otherwise it is useless
- Hence the machine has to be configured as a server, meaning typically:
  - Good TCP configuration
    - The SLC defaults are very bad
    - Look here: http://monalisa.cern.ch/FDT/documentation_syssettings.html
    - Ask Costin Grigoras for details/moral support
  - All the OS resources are available
    - Max overall number of file descriptors (per machine, not per user)
  - The max # of file descriptors per user is >65000
    - In the Alien Howtos there is a recipe for that
    - http://alien.cern.ch/twiki/bin/view/AliEnHowToInstallXrootdNew
    - Xrootd will not start if this is not met.

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Normal setup

- From the Wiki (and simplified)
  - Get the script
    - `wget http://project-arda-dev.web.cern.ch/project-arda-dev/xrootd/tarballs/installbox/xrd-installer`
  - Run it
    - `./xrd-installer --install --prefix <install-prefix>`
  - Check the result
    - FAILED means a bad thing

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Making an RPM

- From the Wiki page:
  - Go into a work subdirectory:
    - mkdir work
    - cd work
  - Get the script:
    - wget http://project-arda-dev.web.cern.ch/project-arda-dev/xrootd/tarballs/installbox/xrd-rpmer
    - chmod 755 ./xrd-rpmer
  - Run it as root user (and wait):
    - sudo ./xrd-rpmer
  - Check that there were no compilation failures.

- The RPM will be placed in the usual directory /usr/src/redhat/RPMS/.
- Depending on the Linux distribution, the location could be different.
  - No way to know it in advance

Configuration files

- Two files, but only one needs modifications
  - The default ***authz.cnf*** is already configured for the ALICE security model
- ***system.cnf*** contains the configuration
  - Redirector name, data partition names, etc…

Important config options

VMSS_SOURCE: where this cluster tries to fetch files from, in the case they are absent.

LOCALPATHPFX: the prefix of the namespace which has to be made "facultative" (not needed by everybody)

LOCALROOT is the local (relative to the mounted disks) place where all the data (or the namespace) is put/kept by the xrootd server.

OSSCACHE: probably your server has more than one disk to use to store data. Here you list the mountpoints where the data is stored.

MANAGERHOST: the redirector's name

SERVERONREDIRECTOR: is this machine both a server and redirector?

OFSLIB: the authorization plugin library to be used in xrootd.

METAMGRHOST, METAMGRPORT: host and port number of the meta-manager (global redirector)

XRDUSER

The name of the user which runs the daemons (e.g. xrootd, aliprod)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Config options

- MANAGERHOST
  - The name of the machine hosting the redirector
    - Execute in the redirector machine:
      - `hostname -f` and copy/paste the result
  - The redirector and all the data servers will understand their role from this information
    - Automagically
  - So, the content must be the same for all the machines of the cluster

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Config options

- SERVERONREDIRECTOR
  - Is your cluster small (1-2 machines)?
    - If so, you can use a machine to host BOTH the redirector services and the data server services. If you like this, put “1”
      - You need decent hw to do that
      - Reconsider this when you add new servers
    - Otherwise put “0”

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Config options

- XRDUSER
  - The daemons will be run under a user's credentials
    - Typically everybody uses "xrootd" or "aliprod"
    - Write here the name of this system user
    - MAKE SURE IT EXISTS
    - MAKE SURE IT HAS A HOME DIRECTORY

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Config options

- METAMGRHOST, METAMGRPORT
  - This is the address of the Global redirector
  - Through this all the data is accessible in the global domain
- VMSS_SOURCE
  - Also this is the URL prefix which point to the Global Redirector
  - Through this, the cluster auto-feeds itself from the global domain in case of troubles

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Config options

- OFSLIB
  - The way the ALICE strong auth works
  - Specify `TokenAuthzOfs` to enable the authz
    - Mandatory for ALICE
    - Writing is only allowed via the auth of the central services
  - Specify `Ofs` to make your little tests out of production
    - But the SE will not work without the strong auth

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Config options

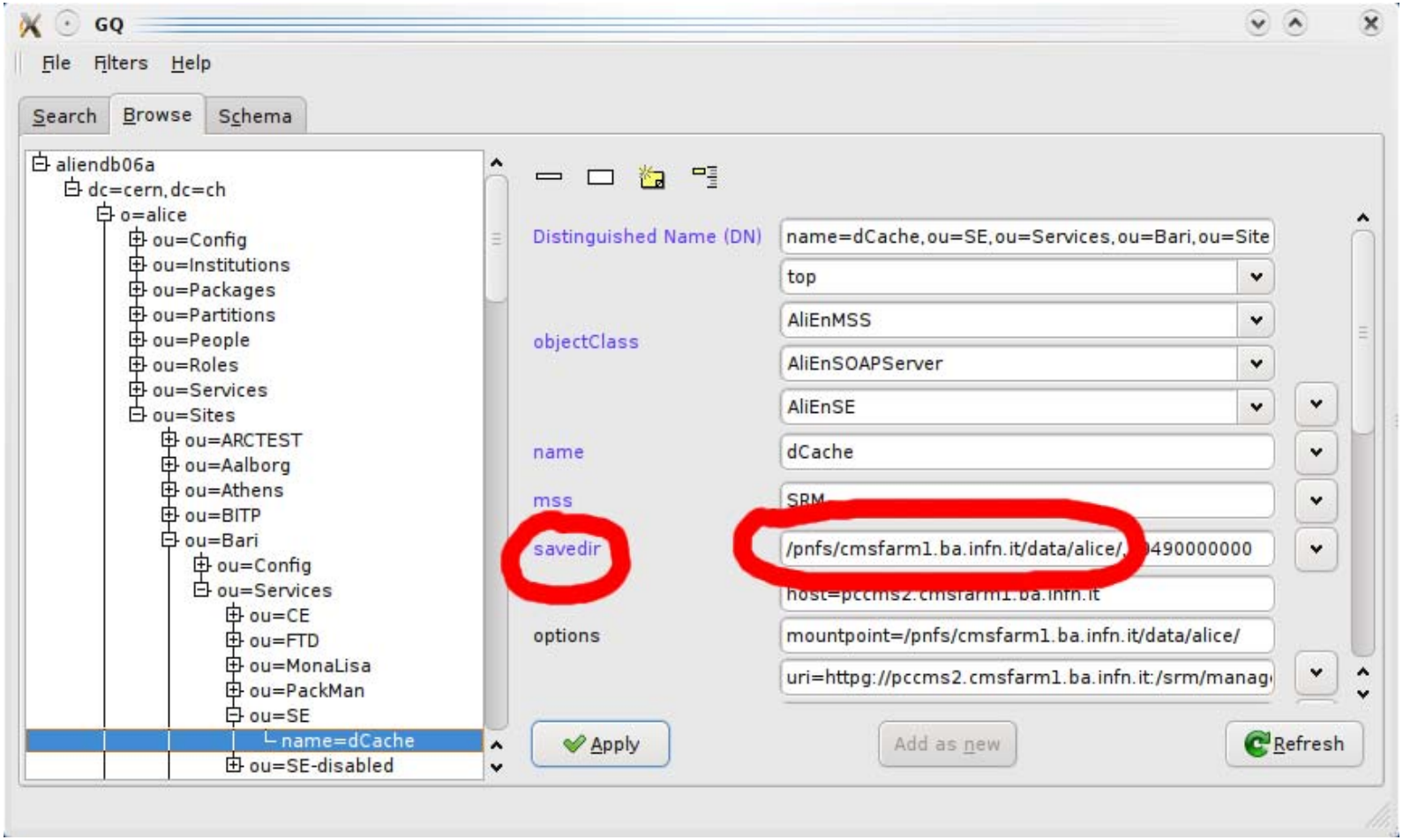
- LOCALROOT
  - The place where the xrootd daemon puts the written data
  - Something like:
    - /mydisk1/xrdfiles/
  - If you are aggregating multiple partitions (OSSCACHE option) then you will put something like this:
    - /mydisk1/data/xrdnamespace
    - This is going to be your most important directory
    - Keep it separated from the others!

Config options

- LOCALPATHPFX
  - The “old style” local prefix to the data
  - Take it from LDAP (next slide)
  - If you are in doubt, please ask

Config options

- LOCALPATHPFX



Config options

- OSSCACHE
  - Here you list (with the xrootd syntax) the xrootd config lines which configure the cache file system, separated by “\n”
    - e.g. if your machine can use as raw storage
      - /data/disk1 and
      - /data/disk2
    - HINT: don’t put data in the root dir of a mountpoint. It works but it’s uncomfortable… then do:
    - “oss.cache public /data/disk1/xrddata\noss.cache public /data/disk2/xrddata”

Setup: start and stop

- Check the status of the daemons

```
xrd.sh
```

- Check the status of the daemons and start the ones which are currently not running (and make sure they are checked every 5 mins). Also do the log rotation.

```
xrd.sh -c
```

- Force a restart of all daemons

```
xrd.sh -f
```

- Stop all daemons

```
xrd.sh -k
```

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

I want to migrate/clean up

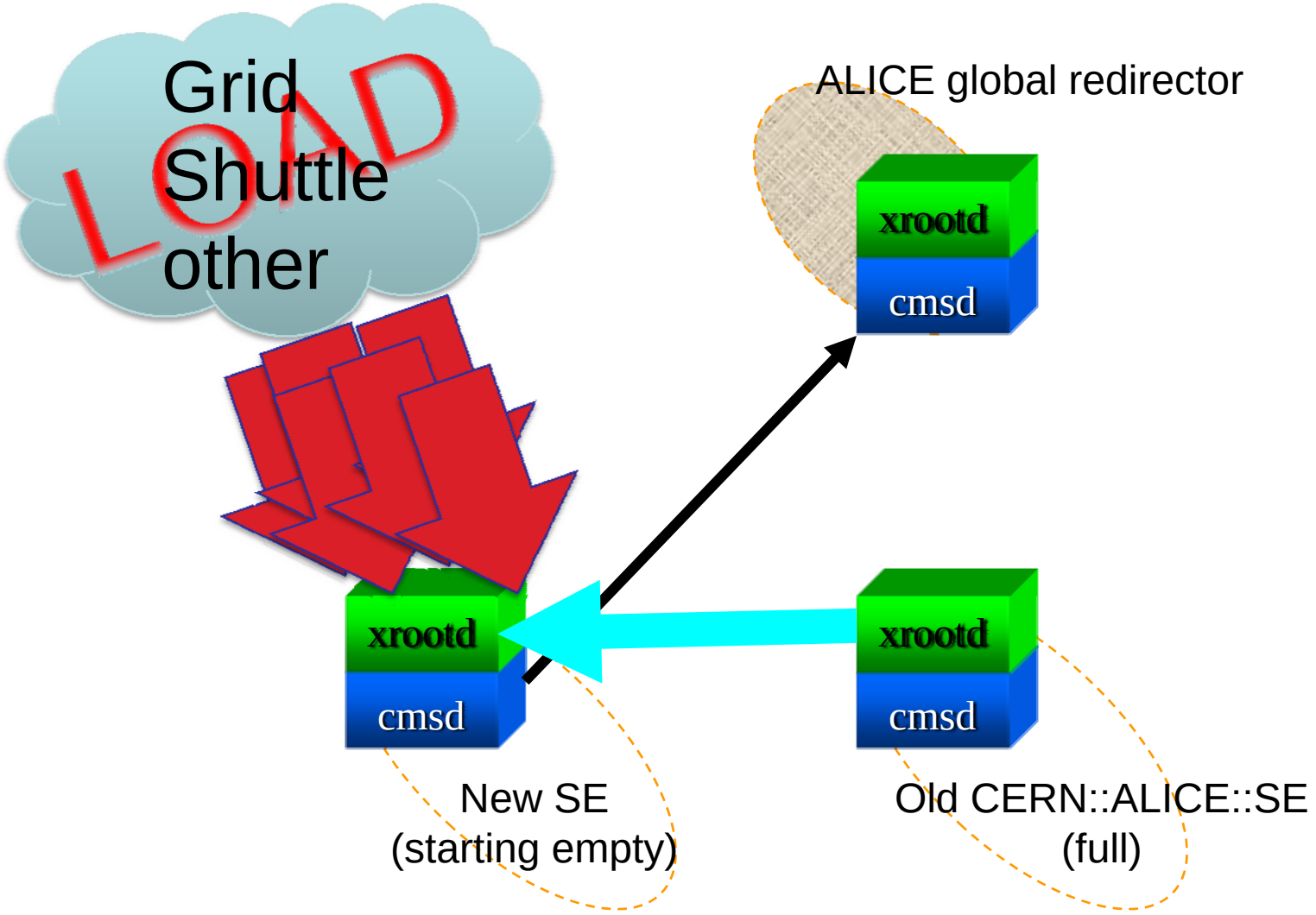
- Too many different situations
  - E.g. Mismatches in the number of servers, in the partition sizes, in the partition numbers, …
  - Hence a single recipe is difficult to invent…
    - Better to understand how it works and be prepared
- BUT … The files are there
    - They are “in clear”
  - Easy with one partition per server
    - Just do it with the normal UNIX tools
  - But, with OSSCACHE we have also the need to migrate/recreate the filesystem part, made of symlinks
    - Every case is different, I never saw the same thing twice
    - In the worst cases a small shell script does the job
- Here we show two relevant cases

The ALICE::CERN::SE July trick

- All of the ALICE cond data is in ALICE:CERN::SE
  - 5 machines pure xrootd, and all the jobs access it, from everywhere
  - There was the need to refurbish that old cluster (2 yrs old) with completely new hw
  - VERY critical and VERY stable service. Stop it and every job stops.
- A particular way to use the same pieces of the vMSS
- In order to phase out an old SE
  - Moving its content to the new one
    - Can be many TBs
    - `rsync` cannot sync 3 servers into 5 or fix the organization of the files
- Advantages
  - Files are spread evenly → load balancing is effective
  - More used files are fetched typically first
  - No service downtime, the jobs did not stop
    - Server downtime of 10-20min… just to be sure
    - The client side fault tolerance made the jobs retry with no troubles
- How to do it:
  - Point the VMSS_SOURCE to the old cluster
  - Point the load to the new cluster

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The ALICE::CERN::SE July trick



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The CNAF09 migration

- Move all the data (17TB)
  - From the phys. dir: /storage/gpfs_alice/alice
  - To the phys dir: /storage/gpfs_alice/storm
  - No downtime wanted (not really needed, but a very nice exercise)

- First we note (from the paths) that there is no cache file system at CNAF
  - They use GPFS as a backend store (+ STORM as SRM implementation)
  - All the files are stored in one directory tree which has to be moved to a different mount point.

Credits: F.Noferini (CNAF)

The CNAF09 migration

- Scenario #1: downtime
  - Announce downtime
  - Stop all the services
  - Start a monster copy with rsync
  - Check the destination
  - Update the xrootd config
    - Fix LOCALROOT to the new path
  - Restart services
  - End of downtime
- Depending on the data volume, it could take even days/weeks (17TB is not much, they could be 1700 or 17000)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The CNAF09 migration

- NO downtime, full service during migration
  - And the new files go in the new place
    - Add a new r/w server to the cluster, on the new mountpoint
    - Make the old server r/o (needs a cfg tweak)
      - So, all the new files will go to the new one
      - And the old ones are still accessible
    - Start the monster rsync
    - Check the destination
    - Switch the old server off
      - The client's fault tolerance will save the jobs, by jumping to the new server. No saving action needed.
    - Done!

Conclusion

- I realize that this might not be enough
  - Impossible to cover all the use cases, even if simple
  - We could try to write a book… would you buy it? :-D
- Most of the complexity is hidden
  - But still there for advanced scenarios
- Basic principle:
  - Know how it works
  - Develop your own ideas
  - Get in touch with the ALICE TF for hints/assistance/moral support
  - The documentation for geeks is public:
    - http://savannah.cern.ch/projects/xrootd
    - Quite advanced but very complete
  - More normal persons will like this:
    - http://alien.cern.ch/twiki/bin/view/AliEn/HowToInstallXrootdNew

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Thank you

Questions?

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it