XROOTD news

Status and strategic directions

ALICE-GridKa operations meeting

03 July 2009

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

The historical Problem: data access

CERN IT Department

- 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
      - Each one opening many files (about 100-150 in Alice, up to 1000 in GLAST/Fermi)
      - Each one keeping many open files
  - 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
- Need scalable high performance data access
  - Need to squeeze every byte/s from the hw
  - No imposed limits on performance and size, connectivity
    - Do we like clusters made of 5000 servers? We MUST be able to do it.
  - Need a way to avoid WN under-utilization

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

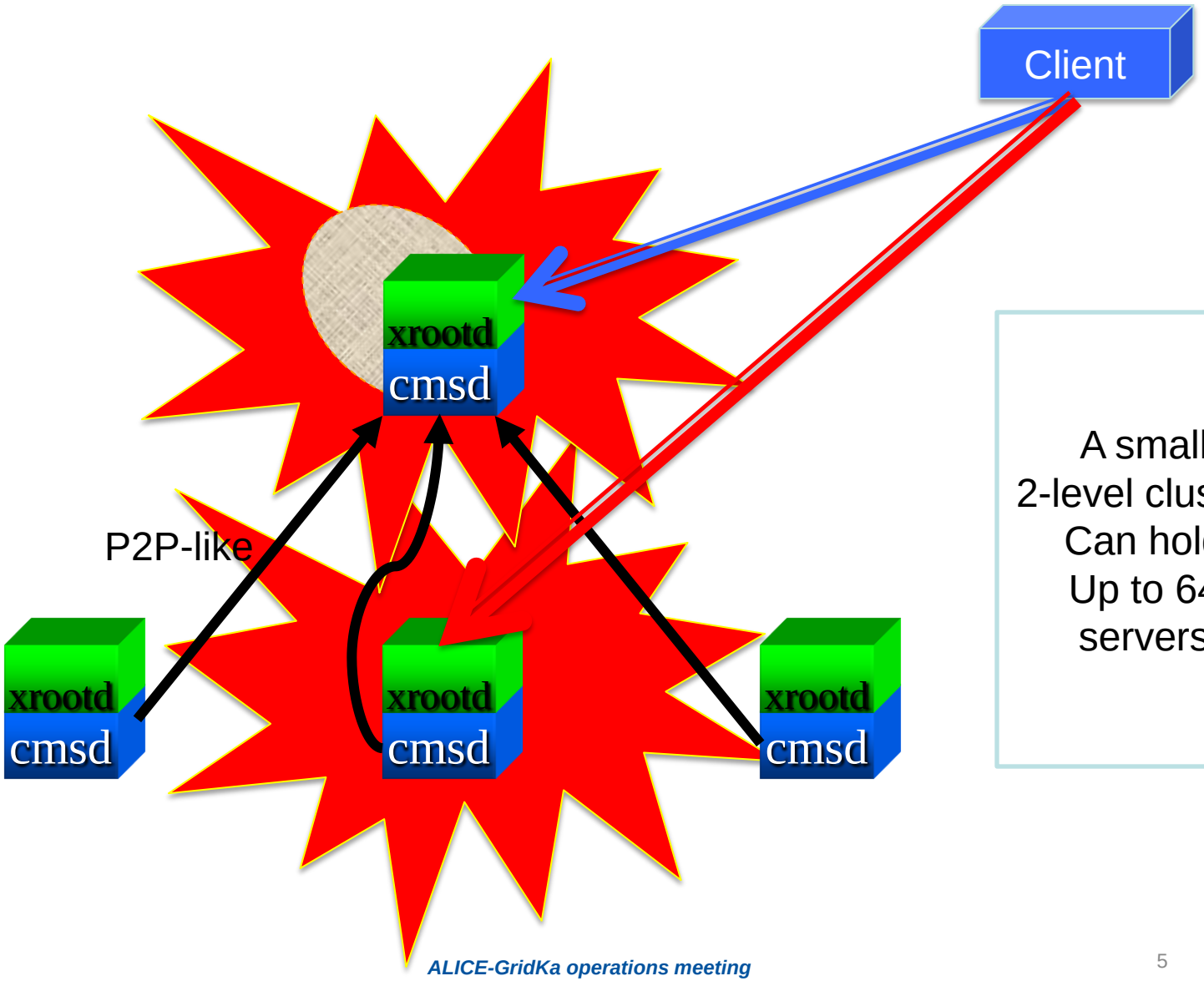
File serving with Xrootd

- Very open platform for file serving
  - Can be used in many many ways, even crazy
    - An example? Xrootd-over-NFS-over-XFS data serving
      - BTW Do your best to avoid this kind of things!
      - In general, the really best option is ‘local disks’ and redundant cheap servers (if you want) + some form of data redundancy/MSS
      - Additional complexity can impact performance and robustness
- Xrootd [Scalla] is only one, always up-to-date!
    - http://savannah.cern.ch/projects/xrootd and http://xrootd.slac.stanford.edu
  - Many sites historically set up everything manually from a CVS snapshot
    - Or wrote new plugins to accommodate their reqs
      - Careful manual config (e.g. BNL-STAR)
  - Many others rely on a standardized setup (e.g. the Alice sites)
  - Others take it from the ROOT bundle (e.g. for PROOF)
      - … which comes from a CVS snapshot
      - Again, careful and sometimes very delicate manual config

Most famous basic features

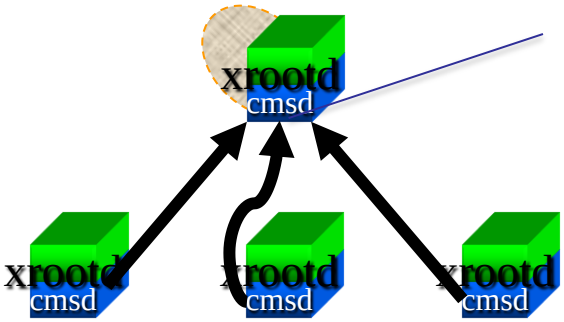
- No weird configuration requirements
  - Scale setup complexity with the requirements' complexity. No strange SW dependencies.
- Highly customizable
- Fault tolerance
- High, scalable transaction rate
  - Open many files per second. Double the system and double the rate.
  - NO DBs for filesystem-like funcs! Would you put one in front of your laptop's file system? How long would the boot take?
  - No known limitations in size and total global throughput for the repo
- Very low CPU usage on servers
- Happy with many clients per server
  - Thousands. But check their bw consumption vs the disk/net performance!
- WAN friendly (client+protocol+server)
  - Enable efficient remote POSIX-like direct data access through WAN
- WAN friendly (server clusters)
  - Can set up WAN-wide huge repositories by aggregating remote clusters
  - Or making them cooperate

Basic working principle

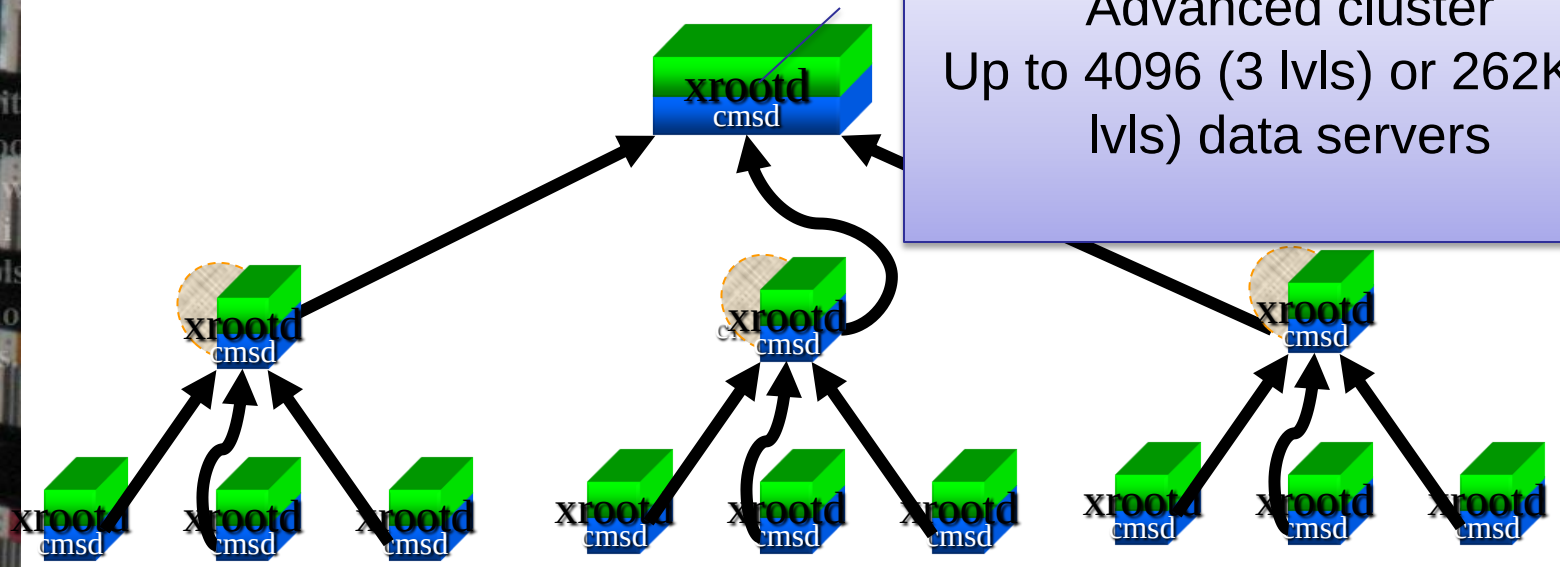


A small
2-level cluster.
Can hold
Up to 64
servers

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

Tech “strategic direction”

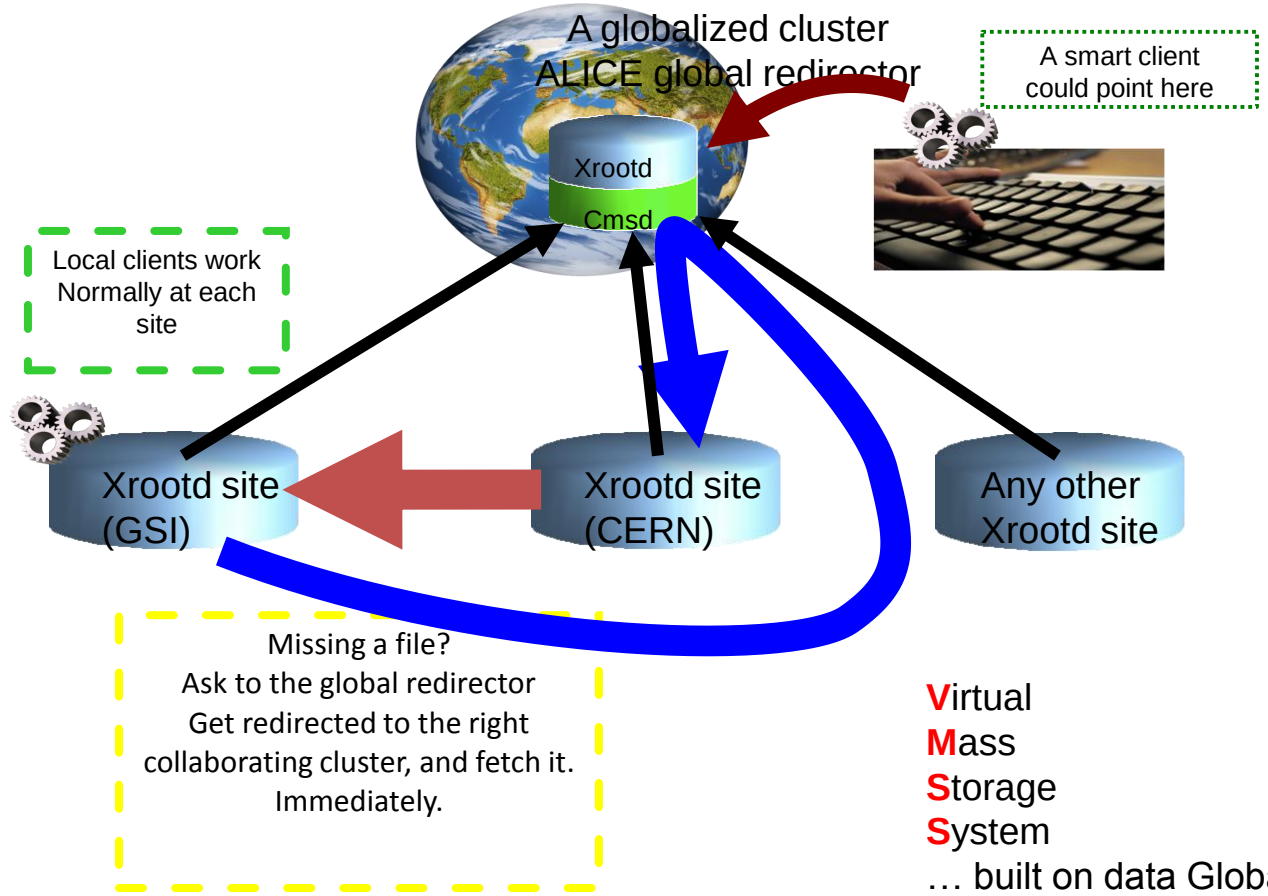
- To give tools able to deploy an “unified worldwide storage”
  - Valid also locally at the site
  - Fast, scalable and efficient
  - WAN friendly
  - Fault tolerance as per xrootd protocol+client
  - Extended fault tolerance by means of a limited ‘self healing’ of an SE’s content
  - Completely self-contained, no external systems
  - Based uniquely on base features of the xrootd platform
- ALICE is moving in this direction
  - Refining history…

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

Storage and metadata

- Here, “Storage” means just “Storage”
  - i.e. you can get the files you put
  - Related with the HEP computing models, which typically need a “metadata DB”
    - To “know” which files are supposed to exist
      - But not necessarily their location, which is handled by the Storage System
      - It does not go out of sync with reality
    - To associate HEP-related metadata to them and do offline queries

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
- Tier-0: CASTOR+Xrd serving data normally.
- Tier-0: Pure Xrootd cluster serving conditions to ALL the GRID jobs via WAN
- "Old" DPM+Xrootd in some tier2s

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

Grid Storage collaboration

- Set of features in the latest ALICE xrootd Ses
- Very generic features, used in a plain way
  - Worldwide storage aggregation
  - “Hole fixing” Virtual Mass Storage System
  - A new entry: The eXtreme Copy (alpha)

Status

- Good!
- The latest releases of the xrootd bundle (1.6+1.6b) are very stable
  - 1.6b to be released now
  - No trace of issues about global cooperation
    - And about data access as well
  - 1.6b fixes a bug in the ALICE token lib
    - A crash which can be triggered only with the ‘future’ tools. It does not hurt the current ALICE usage.
    - So, we are at least 1 step beyond. ☺

Worldwide Storage

- Recipe:
  - Take all the ALICE xrootd Ses
  - Make sure that they export a correct namespace
    - i.e. URLs must not contain site-specific parts
  - Aggregate them into a unique worldwide metacluster
- … and this metacluster appears as an unique storage
- Is it so easy? What can I do then?

Worldwide Storage

- For the sysadmin it's quite easy
  - Eventually, he must just set correctly one parameter
    - LOCALPATHPFX … See the latest "Setup tutorial"
- Why?
  - Historically, each site was given a "VO" prefix
    - Which turned to be a site-dependent mistake
      - Because local choices must remain local
      - A shared service MUST ignore local choices
    - All the sites hosting the SAME file had a different path prefix for it, not VO-related, but site-related
    - Many file entries in the Alice DB contain disk names, local directory names and other bad things
      - What if that disk/machine/local user is changed?
    - If present, that local prefix must be specified in the local config, so that it can be made ***locally optional*** by the xrootd daemon
    - /alice/cern.ch
    - /something/sitename/diskname/

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

Worldwide Storage

- Does it cover the whole ALICE repository by now?
  - Unfortunately not yet
    - "Old" DPM setups do not support this
      - There are chances for the upcoming version
    - dCache-based sites do not run xrootd
      - The xrootd door is a basic emulation of just the basic data access protocol
      - It does not support the xrootd clustering, hence it cannot participate
  - So, what can we do until it grows up?

Worldwide Storage

- The Global redirector ‘sees’ a fraction of the ALICE repository
  - Quite a large one but not complete
  - If a site has a ‘hole’ in the repository this is a good place to fetch this file from
    - The absence of a requested file triggers a “staging” from the Global redirector
    - Born experimental, turning to be a big lifesaver

ALICE Globalization status

- As said, technically very good
  - More sites joining
  - All the new sites are automatically in the game
  - Some older sites have upgraded
  - The new CASTOR is able to join
- ALICE SEs should:
  - Expose a coherent namespace (i.e. a globalizable path)
  - Open access for reading
  - Secured access for writing

A real example

- Remember: ALICE creates always 3 replicas of production-related files
- During Spring 09 the ISS SE had serious hw troubles
  - Relatively small but very good site
  - They lost a number of files, very difficult to tell which ones
    - Hence, the ALICE DB was out of sync with it
  - When updated and put again in production
    - The users did not notice anything, no jobs crashed
    - >1600 missing files were instantly fetched, fixing the repo as they were accessed there
      - During the first production afternoon
    - Success rate was ~99.9%, the avg throughput from external sites was ~5-6MB/s
      - We must remember that the Global redirector cannot see all the sites
      - The few failures were just files in SEs not seen by the Global Redirector

The eXtreme Copy

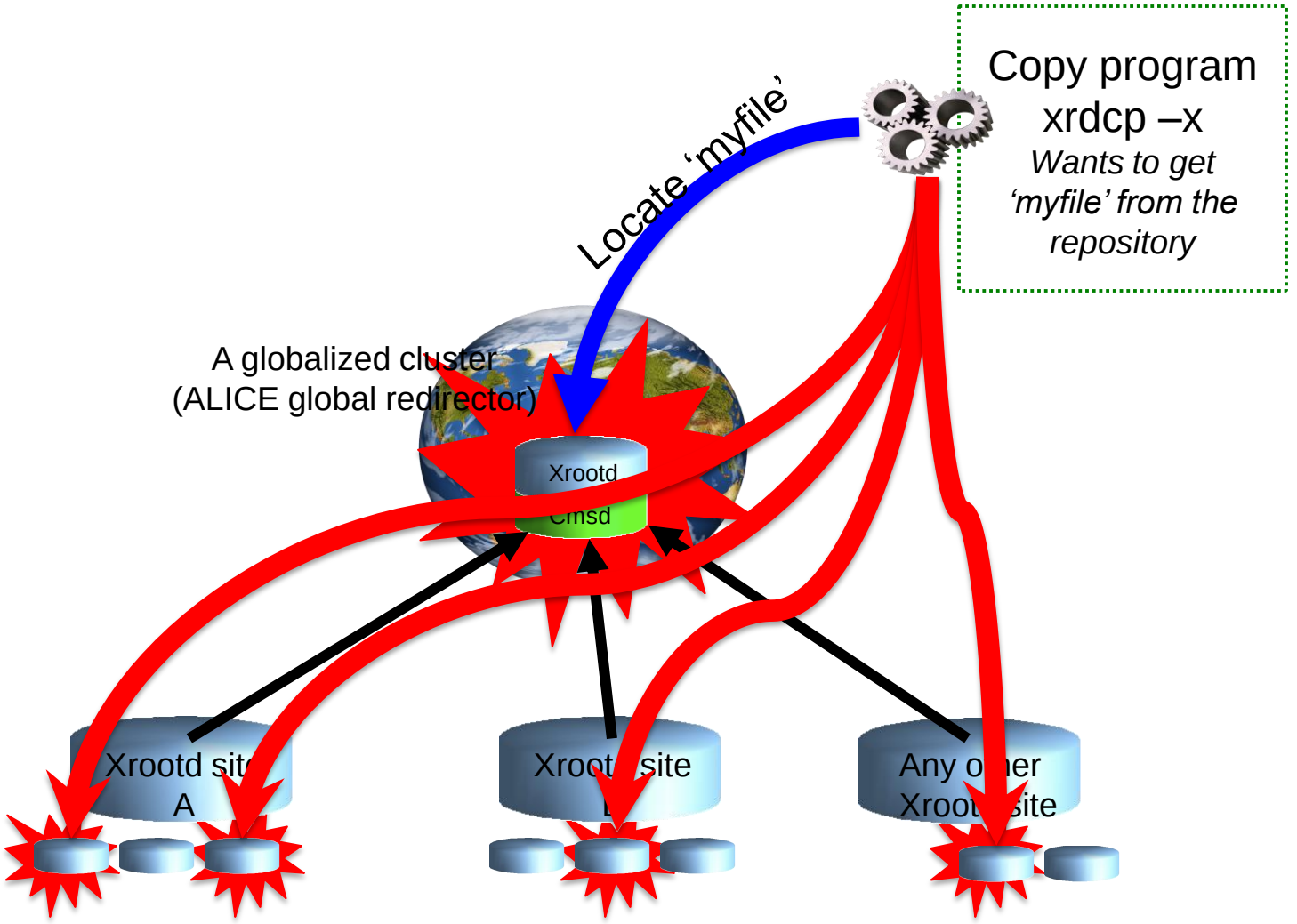
- Let's suppose that I have to get a (big) file
  - And that there are several replicas in different sites
- Big question: where to fetch it from?
  - The closest one?
    - How can I tell if it's the closest? Closest to what? Will it be faster as well?
  - The best connected one?
    - It can always be overloaded or a hoax
  - Whatever I choose, the situation can change over time
    - Instead I want always the max efficiency

The eXtreme Copy

- So let's be Torrent-like
  - And fetch individual chunks from everywhere
    - From ALL the existing replicas of the file
    - Adapting the speed and the load
  - This can boost the performance in doing copies
    - And adapt it to varying conditions
    - But the ALICE computing model does not need to do many copies

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

The eXtreme copy



Main recent fixes

- Replica problem
  - The ALICE xrootd storage is globalized
  - This unique worldwide metacluster was refusing to create replicas (‘File already exists’)
- The hung xrdcp problem
  - Rarely, but it was mainly hung (extremely slow)
  - This also fixes a problem inside (Ali)ROOT, dealing with errors when reading data.
- Stat problem
  - Servers were interacting too much with the global redirector, slowing down a bit the creation of new files
- Drop the token Ofs lib
  - No more authz inside the ofs lib, now it uses the default one
  - Authz now is performed by data servers, not redirectors
  - Authz now is in an XrdAcc plugin by A.Peters (the same as CASTOR)
    - Redirectors now are performant as they should be (cut the CPU consumption by 5-10X)
- Better parameters for servers and partitions balancing
  - Space and load
- Better handling and reporting of server load
  - Aids the load balancing among servers

ALICE URL format

- Now the ALICE internal URL format is ‘as it should have been’
  - root://host[:port]//my/path/to/the/file/as/exported
  - In practice, Alien asks the storage for Alien PFNs (i.e. GUIDs), not LFNs anymore
  - This triggered a hard debugging tour in the central services (thanks Pablo/Andreas)
- As far as I have understood, only the dcache-based sites needed a central fix
  - For some reason the standard ROOT url format was not working
    - Yes, now the central services send a non-standard one to those sites

Beware!

- A different option now forbids us to copy an old system.cnf into a new setup
  - OFSLIB does not exist anymore
  - Replaced by LIBACC
  - Copy by hand the values into the new one
    - 10 simple values, not a big deal
  - Remember to switch the DEBUG mode OFF when you are satisfied
- For the rest, the instructions are the same
  - And the Alien Howto Wiki is up-to-date

Upgrade!

- Due to the importance of the fixes/new features, updating is a very good idea now.
  - And thanks again to the sysadmins of the test sites (ISS::File, Kolkata::SE, CERN::SE, Subatech::SE, Legnaro::SE)

Overview

- Relatively pleasant overall situation
  - New honourful SEs popping up regularly
  - Also small sites give a great contribution to the storage
    - Easy to spot in the SE monitoring
- Path to evolution very alive
  - All the new SEs are in the global domain
  - New xrootd-side developments are making things smoother
  - Means “possibility to do something really good”
- Very good responses from site admins

Recent Q&A

- Frequent question
  - Q: “Hey, I installed everything but I get this warning about alice-gr01:1213”
  - A: Very good! The login in the global redirector is NOT open to every SE. By asking you made the right thing.
- What I usually do is to allow the login for that (new) rdr, and the warning disappears
- Note that locally the new SE is anyway operational

Recent Q&A

- The xrd-installer automated setup is giving very good results
  - However, it still needs a bit of creativity
  - The partitions to aggregate must be specified by the sysadmin
  - In the case there are multiple partitions
    - E.g /disk1 + /disk2
    - It's a very BAD idea putting every data file in the root dir /disk1 or /disk2
      - It works, but it's messy
      - There is also the xrootd namespace (LOCALROOT) to deal with
    - There is a wiser recipe which makes life easier

Subdirectories are easier!

- Don’t put ALL your data in / or /diskX !!!
  - This causes massive headaches
- Create subdirectories instead, with meaningful names
  - /disk1/xrddata
  - /disk2/xrddata
    - These will contain the data (OSSCACHE)
  - /disk1/xrdnamespace
    - Will contain the xrootd namespace
- Doing this, a backup/restore gets trivial to perform
  - And you know what you have

Thank you

Questions?

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