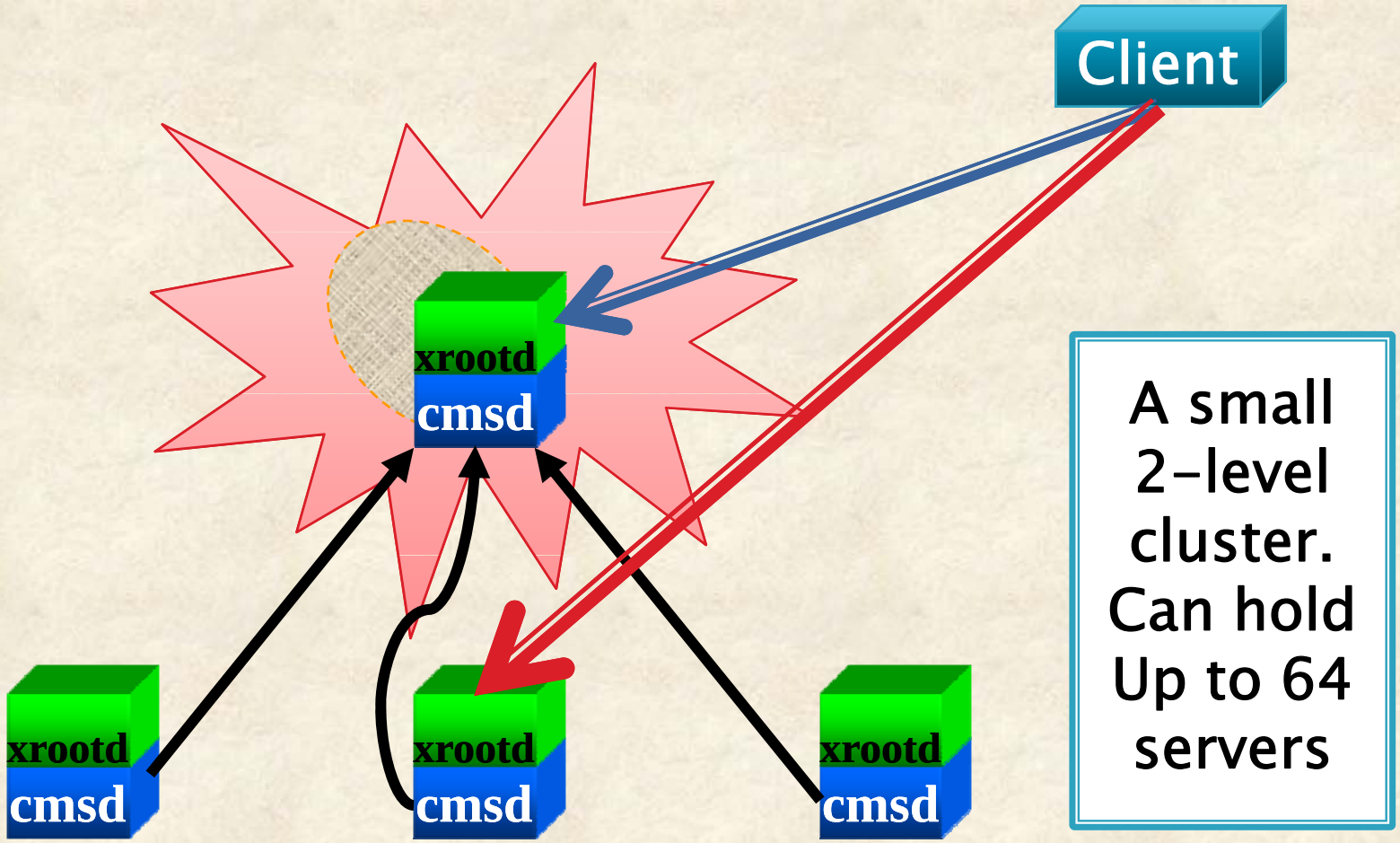
Virtual MSS
The first steps

Some new directions
About the new ALICE::CERN::SE

FabrizioFurano
CERN IT/GS
22-July-08
ALICE TF Meeting

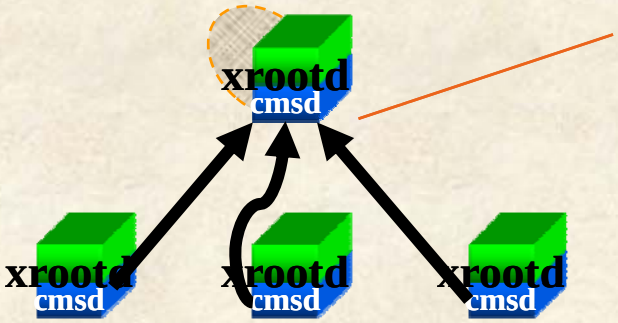
http://savannah.cern.ch/projects/xrootd
http://xrootd.slac.stanford.edu

BasicScalla/XRootD working principle



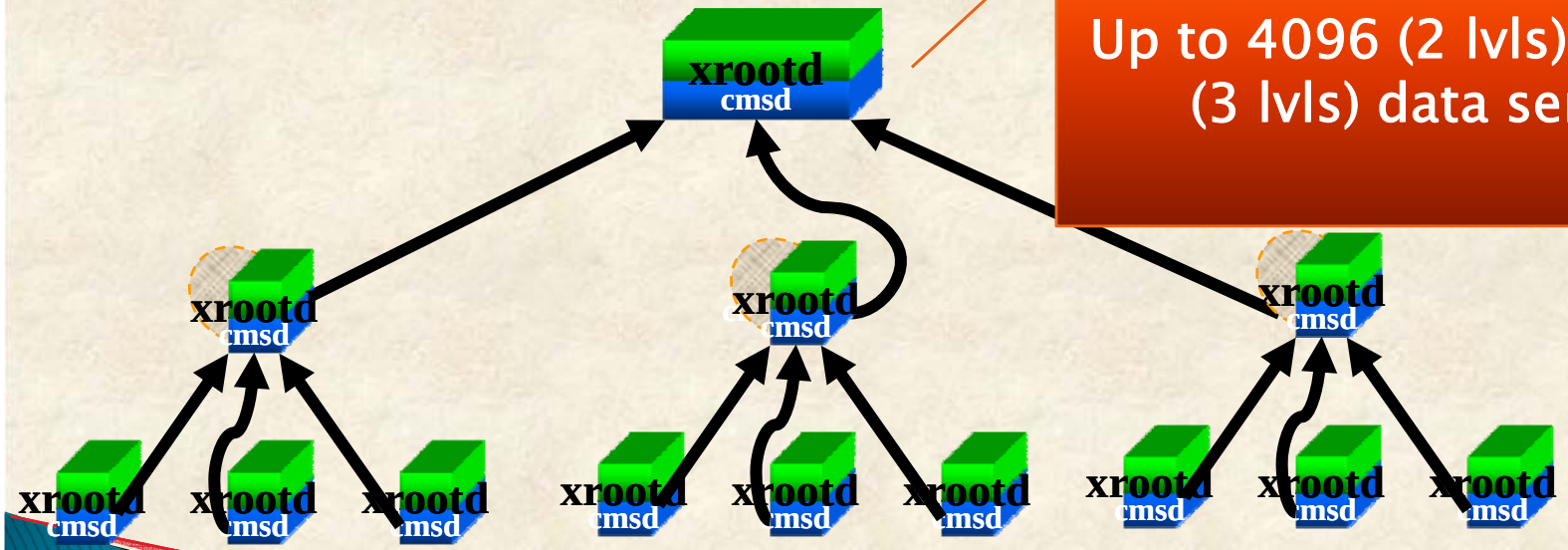
A small 2-level cluster. Can hold Up to 64 servers

Simple LAN clusters



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

Advanced cluster
Up to 4096 (2 lvls) or 262K
(3 lvls) data servers

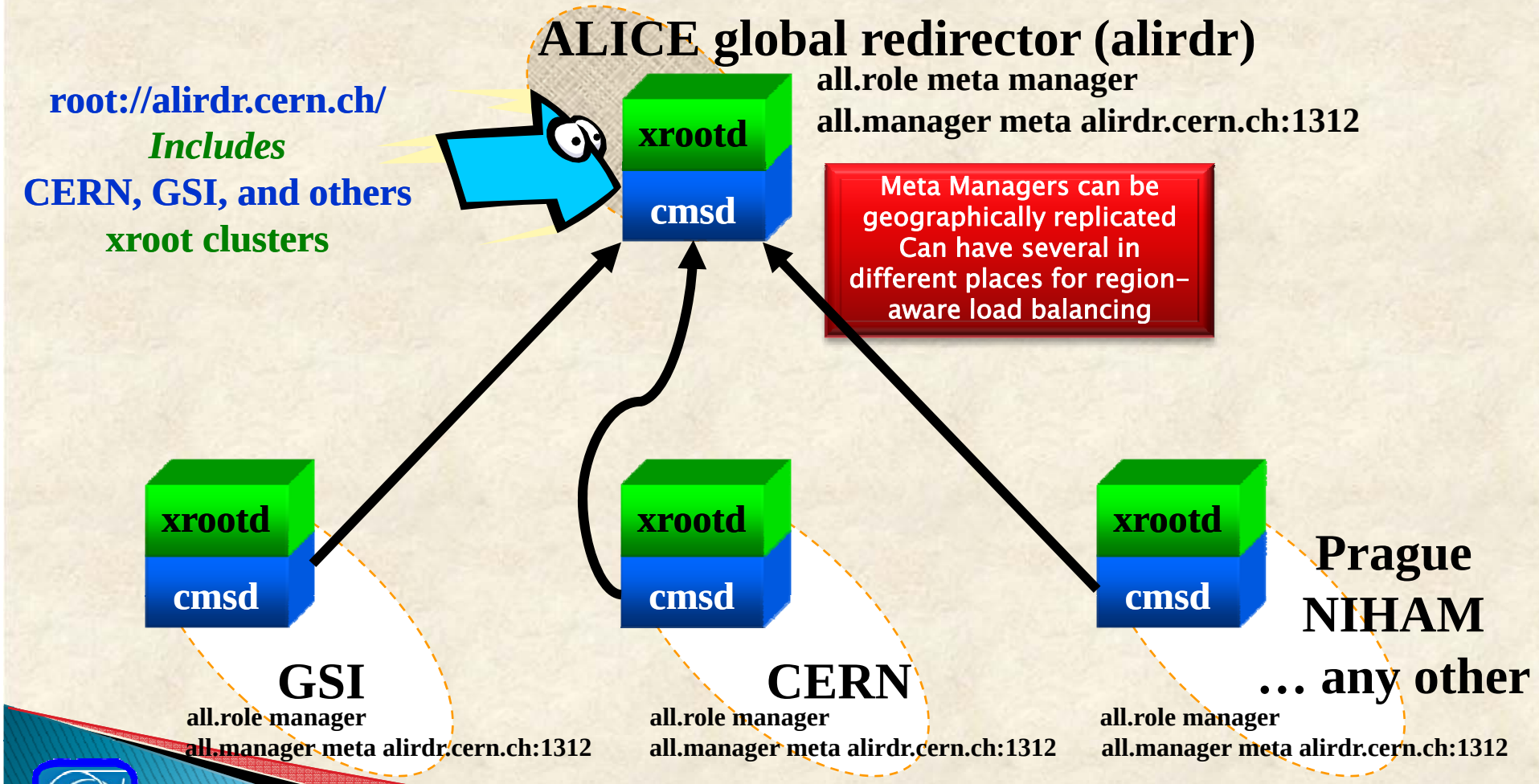


Everything can have hot spares

Virtual MSS

- Purpose:
  - A request for a missing file comes at cluster X,
  - X assumes that the file ought to be there
    - And tries to get it from the collaborating clusters, from the fastest one
- Note that X itself is part of the game
  - And it's composed by many servers
- The idea is that
  - Each cluster considers the set of ALL the others like a very big online MSS
  - This is much easier than what it seems
    - *Slowly Into production for ALICE*

Cluster Globalization… an example



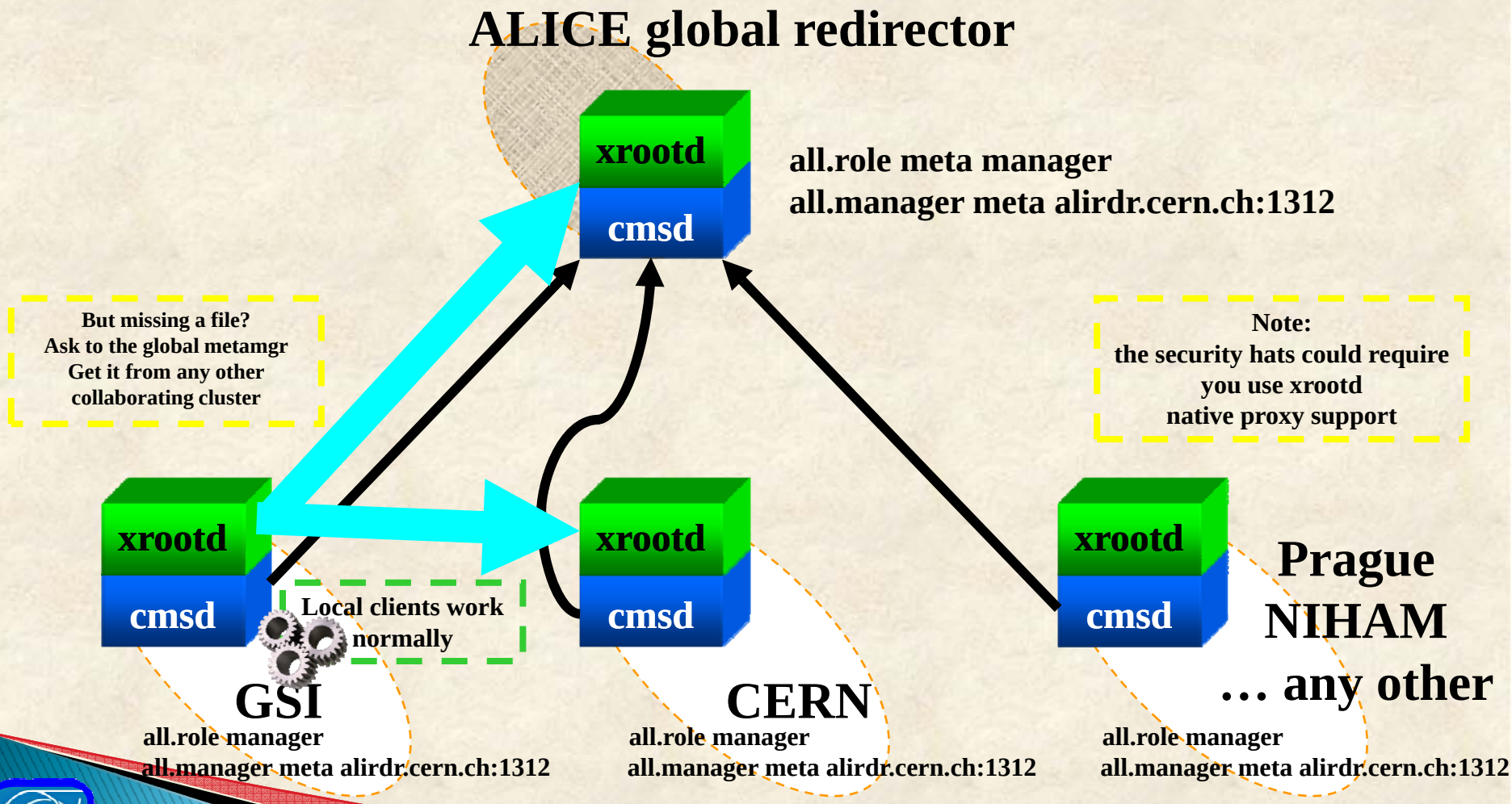
Cluster globalization

- Up to now, xrootd clusters could be populated
  - With xrdcp from an external machine
  - Writing to the backend store (e.g. CASTOR/DPM/HPSS etc.)
- E.g. FTD in ALICE now uses the first. It “works”…
  - *Load and resources problems*
  - *All the external traffic of the site goes through one machine*
    - Close to the dest cluster
- If a file is missing or lost
  - For disk and/or catalogscrewup
  - Job failure
    - *... manual intervention needed*
    - *With 10^7 online files finding the source of a trouble can be VERY tricky*

Many pieces

- Global redirector acts as a WAN xrootd meta-manager
- Local clusters subscribe to it
  - And declare the path prefixes they export
  - Local clusters (without local MSS) treat the globality as a very big MSS
  - Coordinated by the Global redirector
    - *Load balancing, negligible load*
    - *Priority to files which are online somewhere*
    - *Priority to fast, least-loaded sites*
    - *Fast file location*
- True, robust, realtime collaboration between storage elements!
  - Very attractive for tier-2s

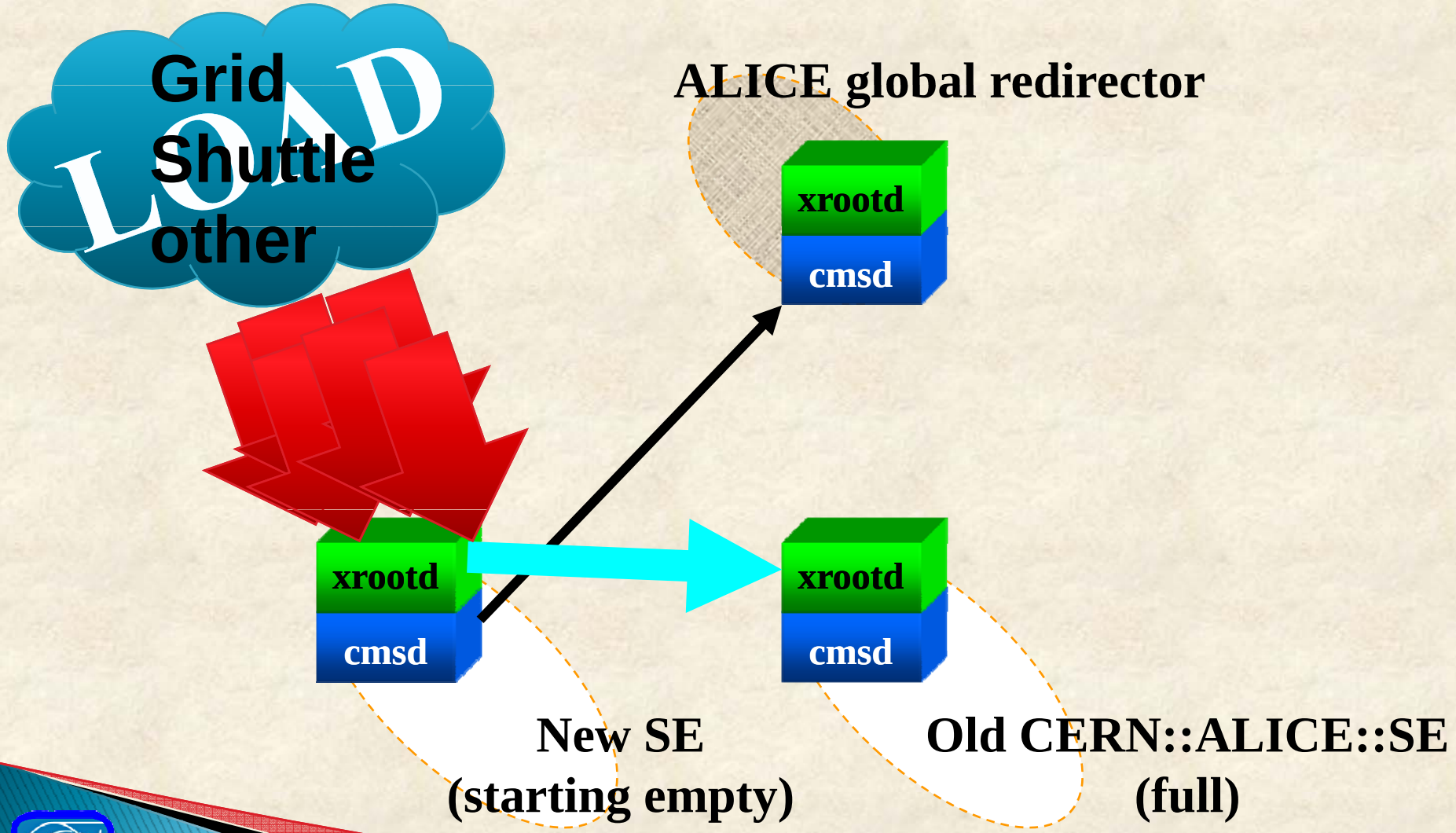
The Virtual MSS Realized



The ALICE::CERN::SE July trick

- A particular way to use the same pieces of the vMSS
- In order to phase out an old SE
  - Keeping its content!
- Advantages
  - Files are spread evenly → load balancing is effective
  - More used files are fetched typically first
- Default vMSSconfig will be restored soon
  - Fetch from the global rdr
- But it's already subscribed to the global rdr

The ALICE::CERN::SE July trick



Virtual MSS

- The mechanism is there, fully "boxed"
  - The new setup does almost everything it's needed
- A (good) side effect:
  - Pointing an app to the "area" global redirector gives complete, load-balanced, low latency view of all the subscribed SEs
  - An app using the "smart" WAN mode can just run
    - *Probably now a full scale production/analysis won't*
      - But what about an interactive small analysis on a laptop?
      - After all, HEP sometimes just copies everything, useful and not
    - *I cannot say that in some years we will not have a more powerful WAN infrastructure*
      - And using it to copy more useless data looks just ugly
      - If a web browser can do it, why not a HEP app? Looks just a little more difficult.
- Better if used with a clear design in mind

What's missing

- XrdCASTOR subscription to the Global redirector
  - Needs a complete xrd refurbishment, very old versions
- The new xrootd packages will be published shortly
  - 1-2 weeks. Just some minor fixes, to avoid troubles
  - Verify that the ML info is there (should be)
- Migration of the tier-2s?
  - They have very old versions too
  - This should be quite easy for pure-xrootd sites
- Xrd-DPM refurbishment?
  - With consequent subscription to the global rdr
  - Needs a completexrd refurbishment, very old versions
- 3rd party fetches development
  - Reduce load on FTD
  - Put the DCaches into the vMSS game in some way