Scalla/xrootd WAN globalization tools: where we are.

Now my WAN is well tuned!

The network is there. Can we use it?

So what?

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

Outline

- Wan specifics
- What can be done, what can be desired
  - Two simple use cases
  - What improved recently
- Reference test on a 10Gb WAN
  - How does direct access behave?
- An interesting usage example
- About storage globalization
- Conclusions

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

Buy it, tune it… use it !

CERN IT Department

- We could have a very powerful WAN system connecting all the sites
  - And there is. We learnt also how to tune it.
  - But then?
- Over WAN I can use my fancy HTML browser
  - I have to read my CHEP09 slides in INDICO
  - What about first transferring the whole INDICO to my HD? Disk space is cheap now…
    - Is it practical/robust/reasonable to do that?
  - Is it true that for my analysis I just can transfer files around among N fancy storage pools?

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

WANs are difficult

CERN IT Department

- In WANs each client/server response comes much later
  - E.g. 180ms later
- With well tuned WANs one needs apps and tools built with WANs in mind
  - Otherwise they are walls impossible to climb
    - I.e. VERY bad performance… unusable
  - Bulk xfer apps are easy (gridftp, xrdcp, fdt, etc.)
  - There are more interesting use cases, and much more benefit to get
- ROOT has the right things in it
    - If used in the right way
- With XROOTD … OK!, CASTOR too (2.1.8-6)

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

CERN

What can we do

CERN IT Department

- Basically, with an XROOTD-based frontend we can do 2 things via WAN:
  - Access remote data
  - Aggregate remote storages
    - Build an unique storage pool with subclusters in different sites
    - No practical size limits, up to 262K servers in theory
    - No third-party SW needed
    - So, we don't need to know in advance where a file is…
    - We just need to know which is the file we need
- There are pitfalls and things to consider
  - But a great benefit to get as well
  - Let's see what's possible and some of the new ideas

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

A simple use case

CERN IT Department

- I am a physicist, waiting for the results of my analysis jobs
  - Many bunches, several outputs
    - Will be saved e.g. to an SE at CER
  - My laptop is configured t… n ROOT
  - I leave for… pla…
    …results from
    …ave my new histos in the
    …time to loose in tweaking to get a copy of …erything. I loose copies into the confusion.
  - I want to leave the things where they are.

I know nothing about things to tweak.
What can I expect? Can I do it?

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

Another use case

- ALICE analysis on the GRID
- Each job reads ~100-150MB from ALICE::CERN::SE
  - These are cond data accessed directly, not file copies
    - I.e. VERY efficient, one job reads only what it needs.
  - It would be nice to speed it up
    - At 10MB/s it takes 10 secs
    - At 5MB/s it takes 20secs
    - At 1MB/s it takes 100
- Sometimes data are accessed elsewhere
  - It would be nice if it was more efficient
    - Better usage of resources, more processed jobs/day
    - After all, ROOT/ALIROOT is not able to r/w data at more than 20MB/s with 100% usage of 1 core
      - Probably will do better in the future
    - This fits perfectly with the current WAN status

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

Up to now

- Up to now the WAN speedup was possible with ROOT+XrdClient+XROOTD
  - Up to 100-150x with respect to basic client-server protocols
  - But it needed a tweak to enable/disable the WAN mode of XrdClient
    - When to switch it on?
    - So, difficult to automatize. Very technical, hence nobody cares!
- Now the things went further
  - The good old WAN mode is OK for bulk xfers
  - The new internal improvements use the efficiency of the newer kernels and TCP stacks
  - Interactive things should need nothing now
- So, if you have:
  - The new client (bundled in ROOT!)
  - A new server (available through xrd-installer)
    - With the new fixed configuration!
  - You can expect a good improvement over the past
    - Without doing nothing special, no user tweaks

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

Exercise

- Caltech machinery: 10Gb network
- Client and server (super-well tuned)
  - Selectable latency:
    - ~0.1ms = super-fast LAN
    - ~180ms = client here, server in California
      - (almost a worst case for WAN access)
- Various tests:
  - Populate a 30GB repo, read it back
  - Draw various histograms
    - Much heavier than the normal, to make it measurable
    - From a minimal access to the whole files
    - Putting heavy calcs on the read data
    - Up to reading and computing everything
      - Analysis-like behaviour
  - Write a big output (~600M) from ROOT

Thanks to Iosif Legrand and Ramiro Voicu

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

Exercise

CERN IT Department

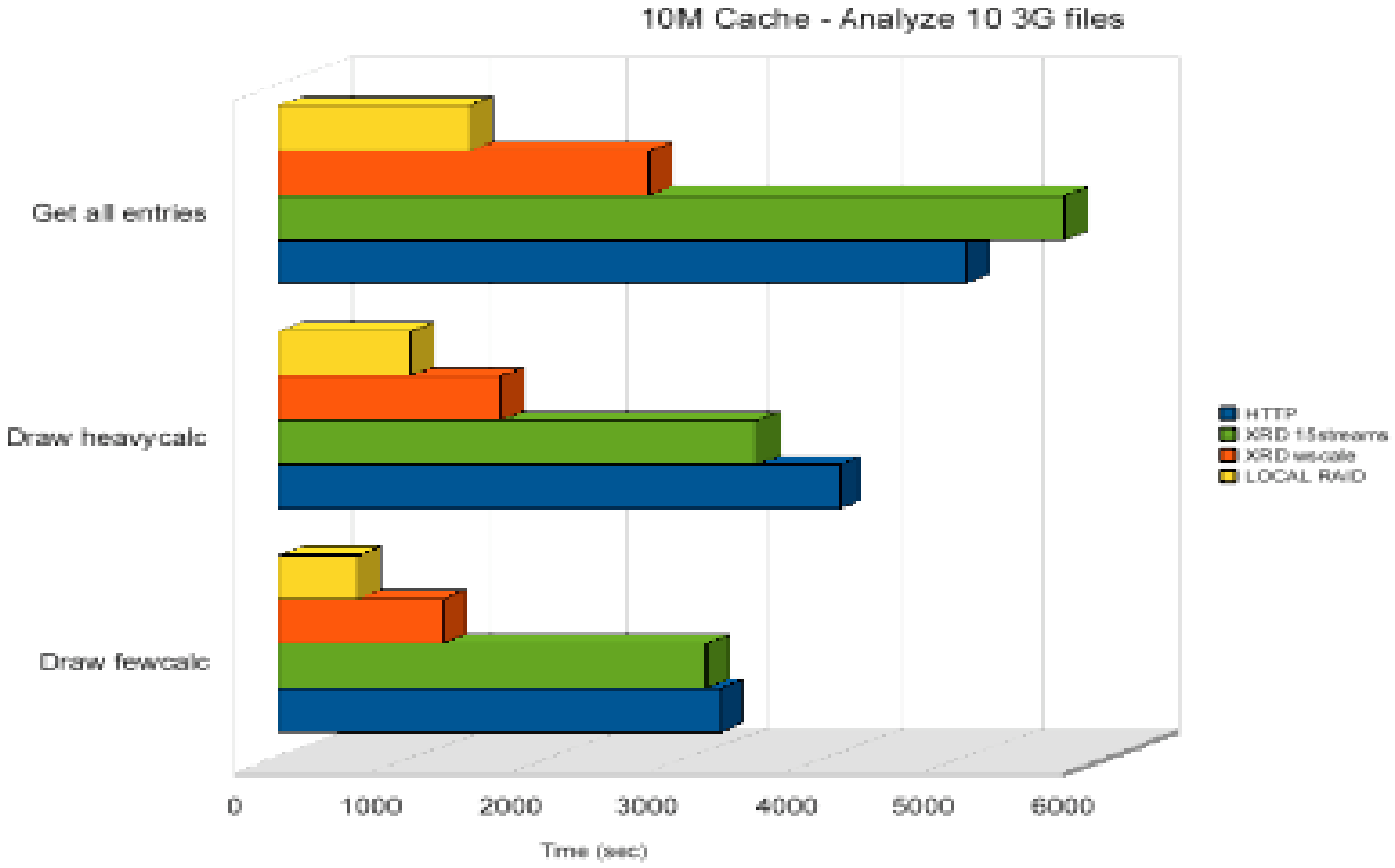
- This is not a “Bandwidth race”
  - The goal is not to fill the 10Gb bandwidth
    - Others are interested in that, and do it very well
- We wanted to see:
  - Can we use all this to live better with data?
  - How does a normal task perform in LAN/WAN?
    - In a measurable and stable WAN environment
- Local disk vs XROOTD vs HTTP (Apache2)
  - Why HTTP? Because it is just the most difficult opponent:
    - Efficient (LAN+WAN) and lightweight
    - No bandwidth waste *
    - Very robust server (but not enough OK for HEP data mgmt)
    - Well integrated in ROOT, works well (except writes, not supported)

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

** See the talk about gpfs/xrootd (on Thu) and the Lustre analysis by A.Peters (ACAT08)*

CERN

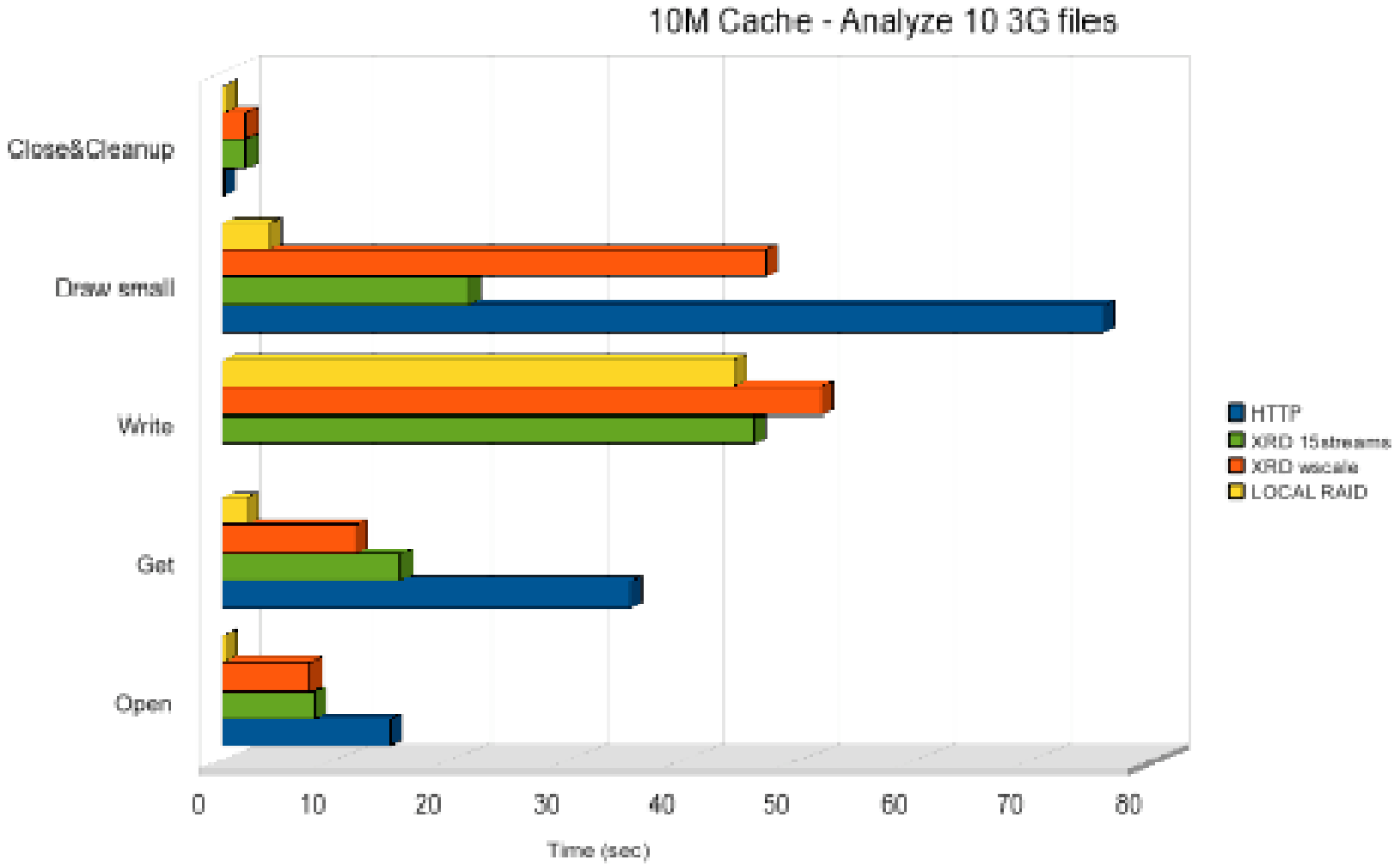
10Gb WAN 180ms Analysis



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

F. Furano, A. Hanushevsky - Scalla/xrootd WAN globalization tools: where we are. (CHEP09)

10Gb WAN 180ms Analysis



An estimation of Overheads and write performance

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

F. Furano, A. Hanushevsky - Scalla/xrootd WAN globalization tools: where we are. (CHEP09)

Comments

- Things look quite interesting
  - BTW same order of magnitude than a local RAID disk (and who has a RAID in the laptop?)
  - Writing gets really a boost
    - Aren't job outputs written that way sometimes?
    - Even with Tfile::Cp
- We have to remember that it's a worst-case
  - Very far repository
  - Much more data than a personal histo or an analysis debug (who's drawing 30GB personal histograms? If you do, then the grid is probably a better choice.)

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

Comments

- As always, this is not supposed to substitute a good local storage cluster
  - But can be a good thing for:
    - Interactive life, multicore laptops
    - Saving the life of a job landed in a place where its input is not present
    - Federating relatively close sites…
      - E.g. one has the storage, the other has the WNs
    - An user willing to debug its analysis code locally
      - Without copying all the repo locally
    - Whatever could come to the mind of a WWW user

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

A nice example

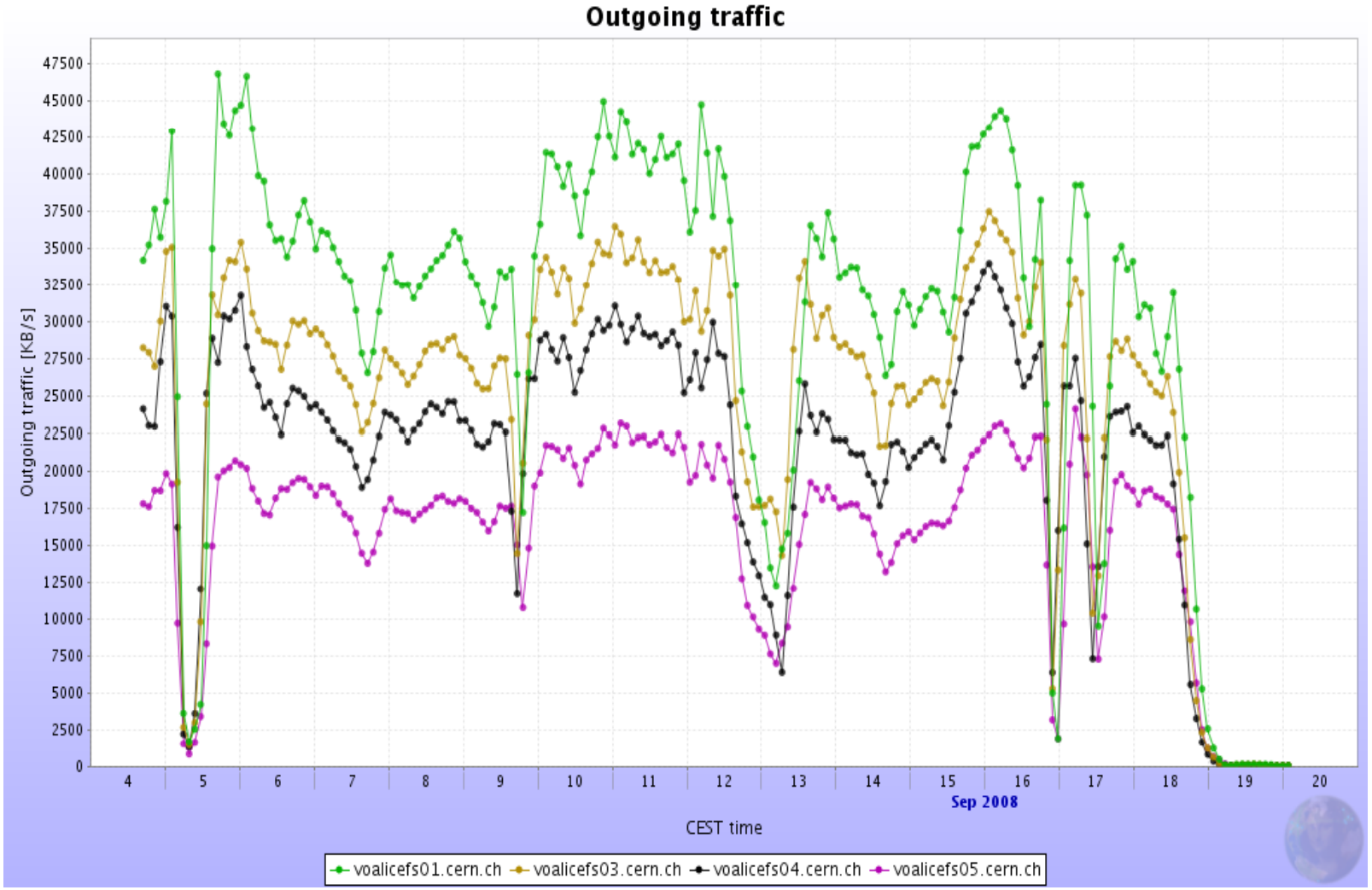
CERN IT Department

- ALICE conditions data repository
  - Regular ROOT files annotated in the AliEn catalogue
    - Populated from various online DBs and runtime detector tasks
    - Nothing strange
  - Primary copy on xrootd storage servers at CERN (5x, 30 TB total)
  - Accessed directly by all MC and reconstruction jobs on the Grid
    - Up to 12Kjobs, Up to 6-8K connections
    - Directly means no pre-copy, i.e. very byte-efficient

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

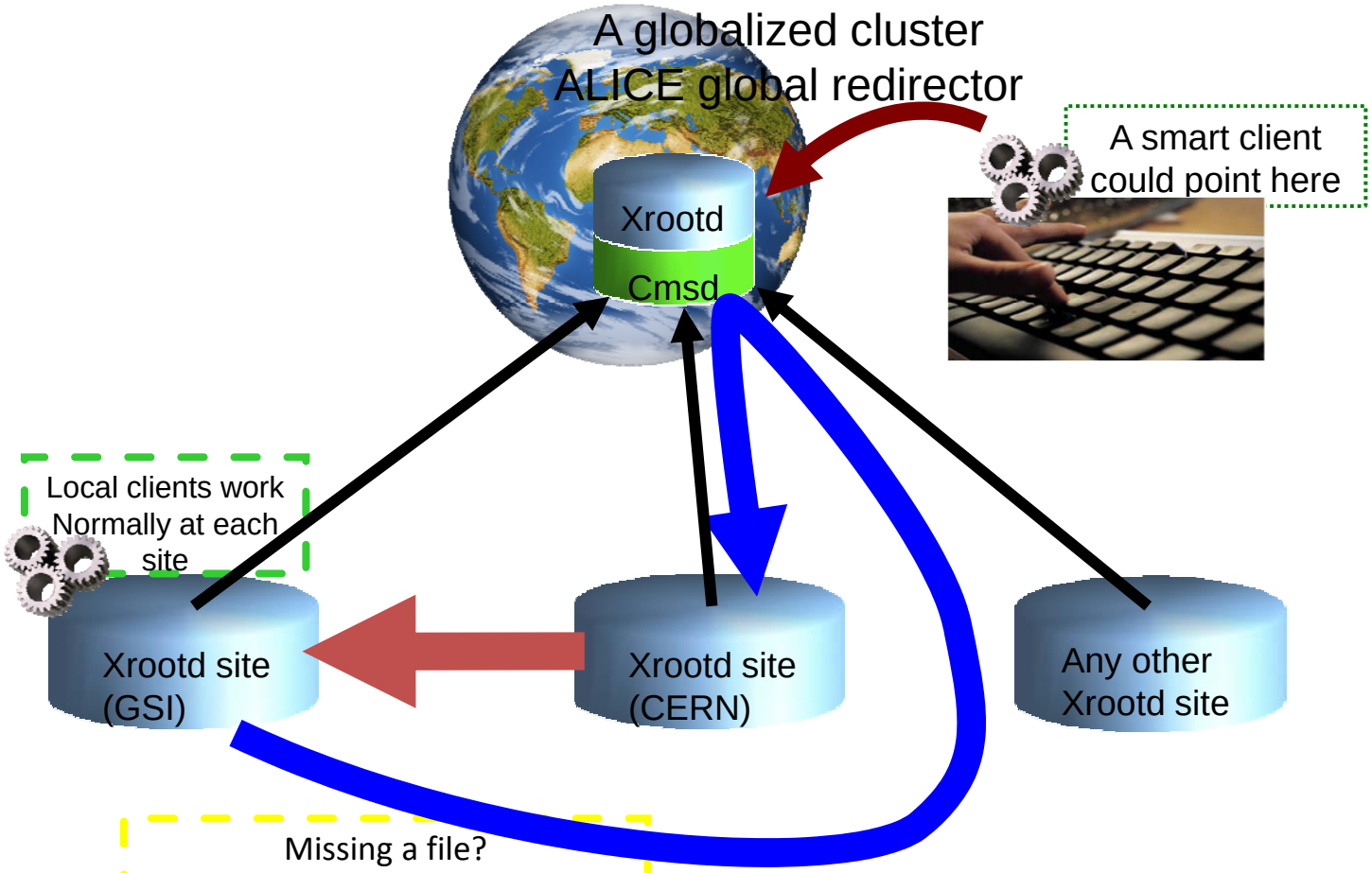
CERN

A nice example



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

More than Globalization: The VMSS



Virtual
Mass
Storage
System
… built on data Globalization

F. Furano, A. Hanushevsky - Scalla/xrootd WAN globalization tools: where we are. (CHEP09)

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

Conclusion

CERN IT Department

- Many things are possible
  - E.g. solving the cond data problem was a breakthrough for Alice
    - It would be nice to use the globalization to lighten the File Catalog
      - And use it as a metadata catalog, as it should
- Technologically it's satisfactory
  - But not ended here, there are new possible things, like e.g.
    - Torrent-like *Extreme copy*
      - To boost data movements also in difficult sites
    - Mounting locally a globalized WAN Xrootd metacluster
      - As a local file system using the XCFS tools

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

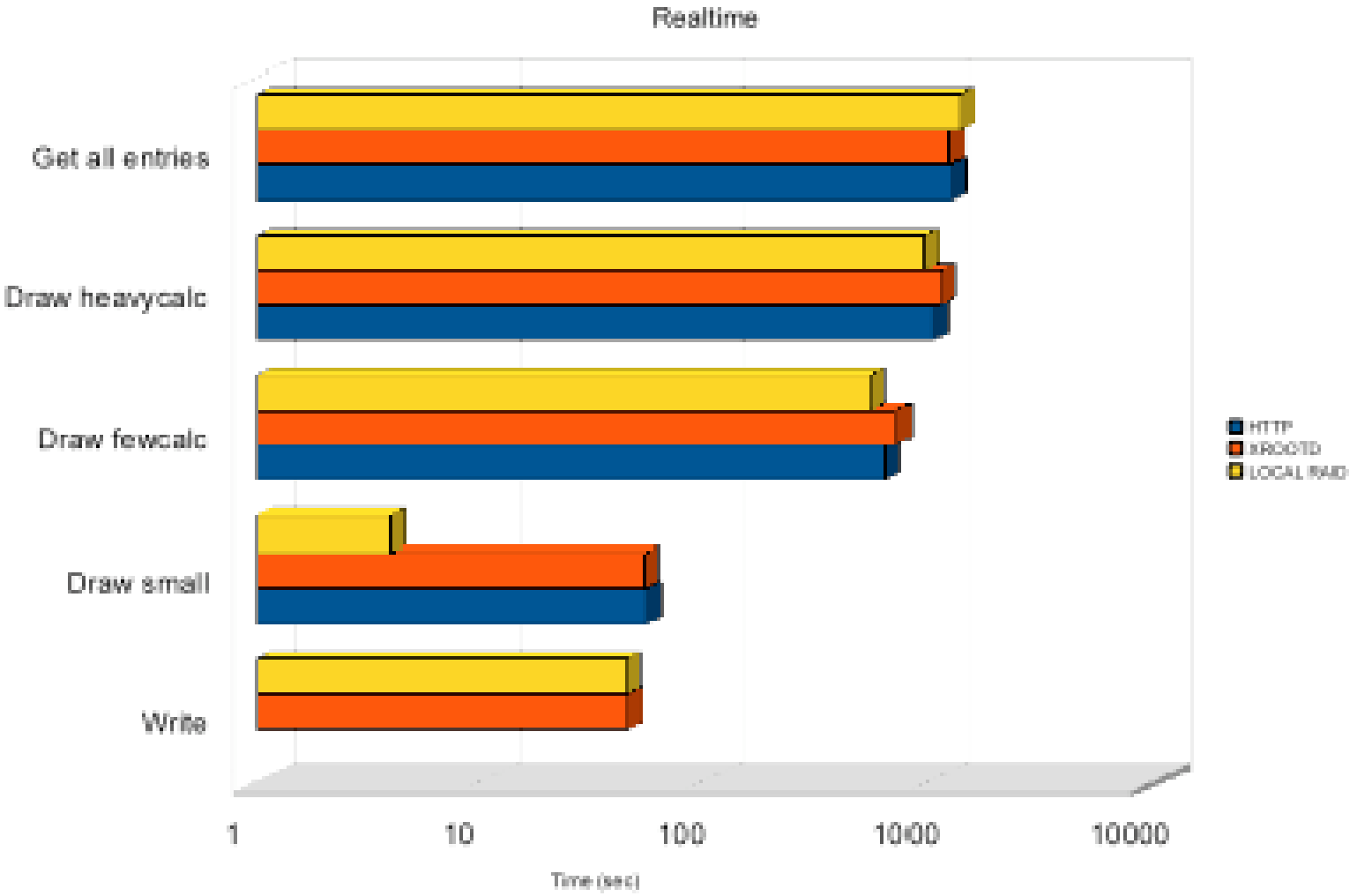
CERN

Thank you!

QUESTIONS?

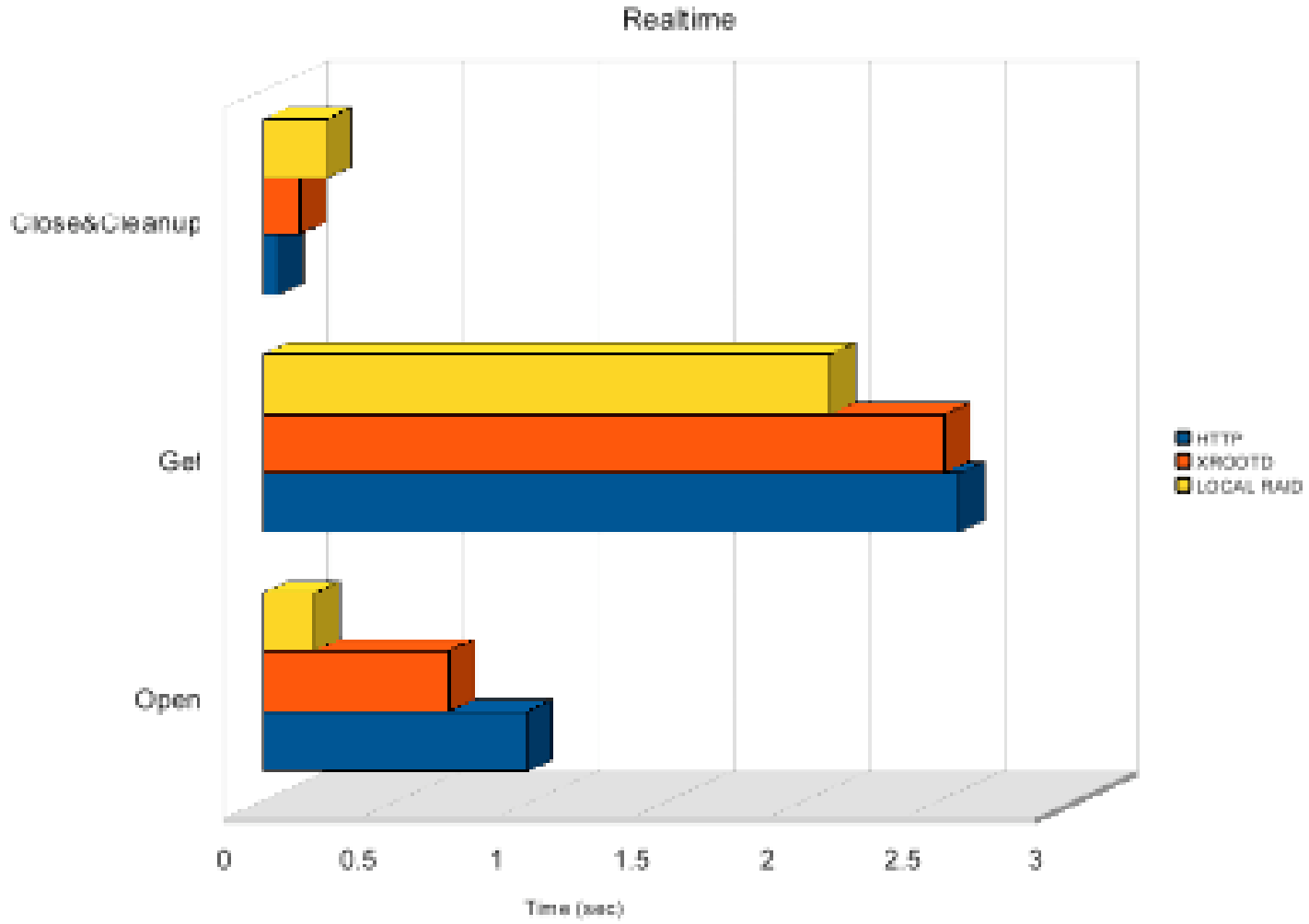
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

10Gb LAN – Reference



F. Furano, A. Hanushevsky - Scalla/xrootd WAN globalization tools: where we are. (CHEP09)

10Gb LAN – Reference

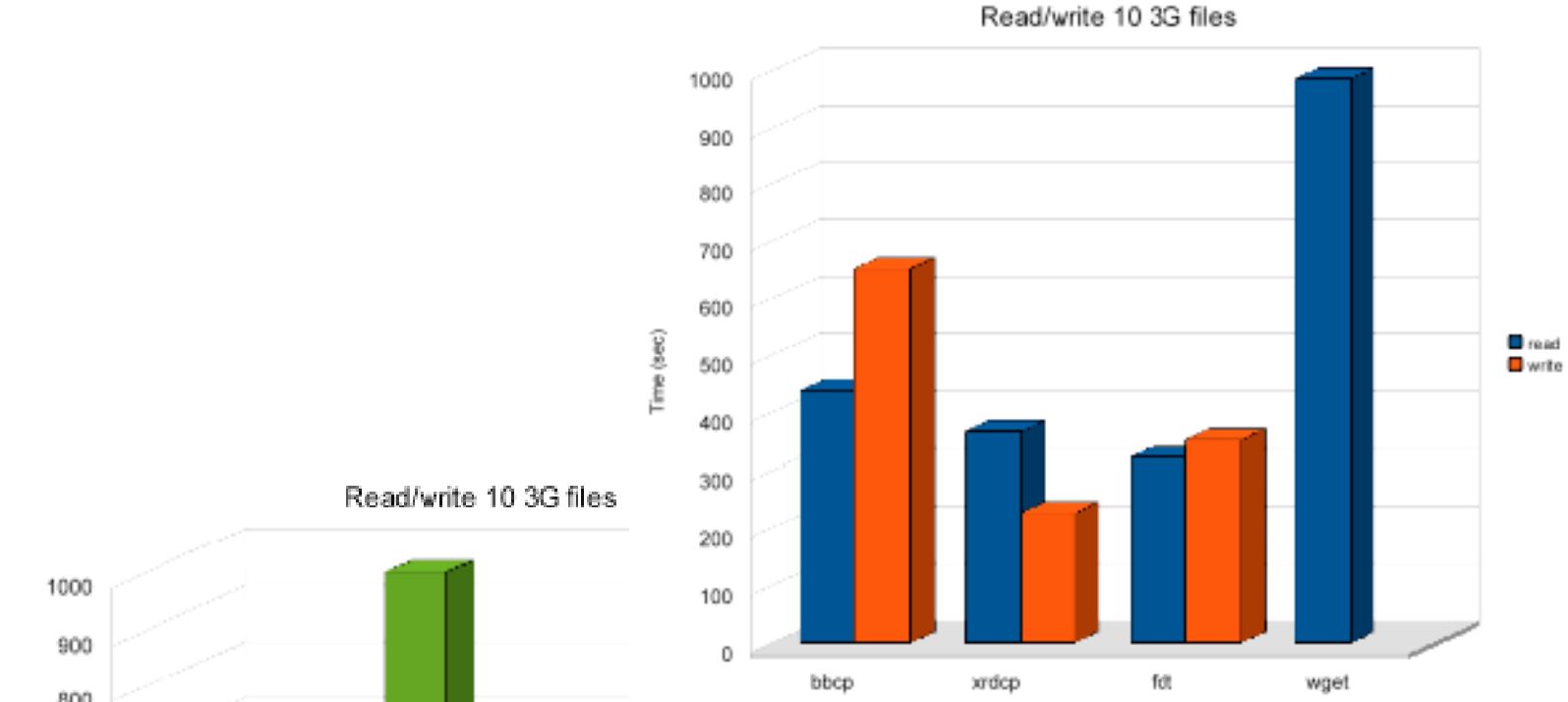


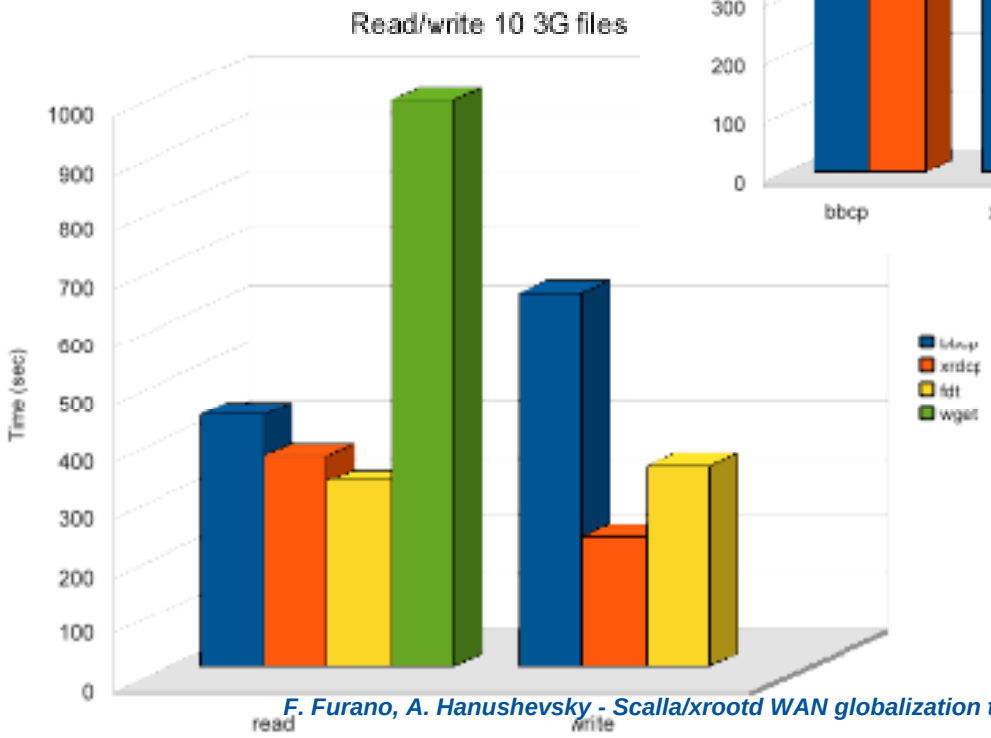
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

F. Furano, A. Hanushevsky - Scalla/xrootd WAN globalization tools: where we are. (CHEP09)

10GbWAN 180ms Bulk Xfer

CERN IT Department





F. Furano, A. Hanushevsky - Scalla/xrootd WAN globalization tools: where we are. (CHEP09)

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