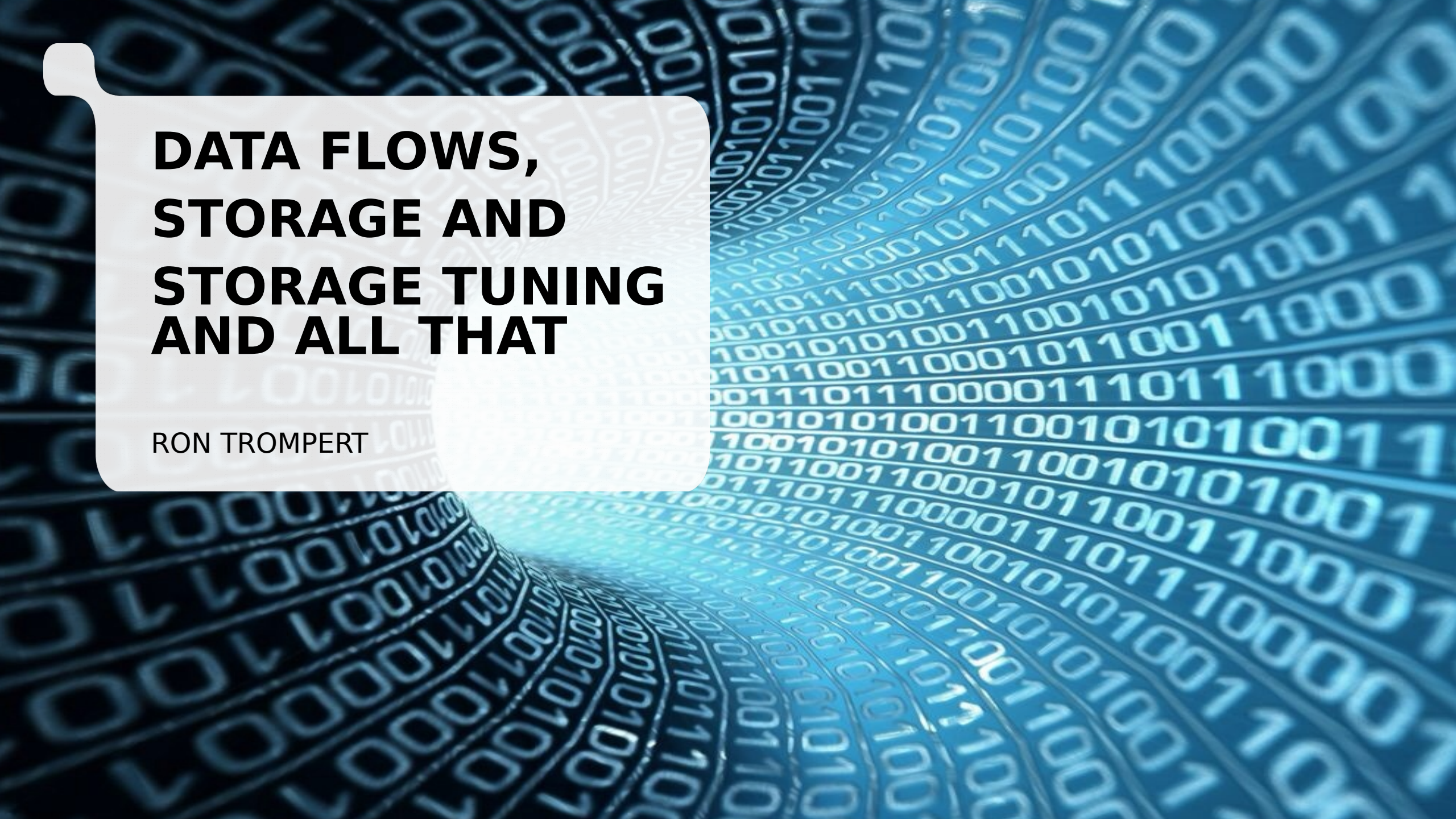


DATA FLOWS, STORAGE AND STORAGE TUNING AND ALL THAT

RON TROMPERT

Contents

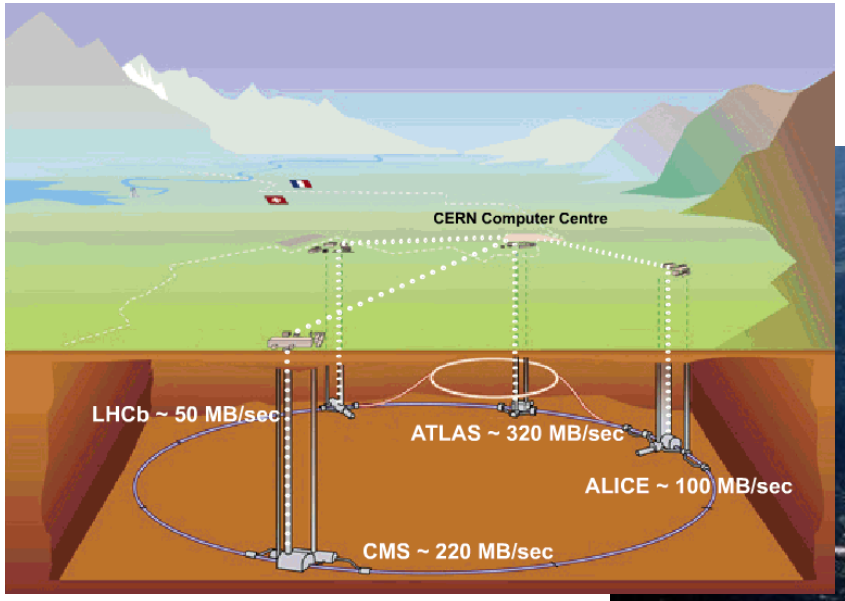
- A bit of history

- Things that we do at SURFsara

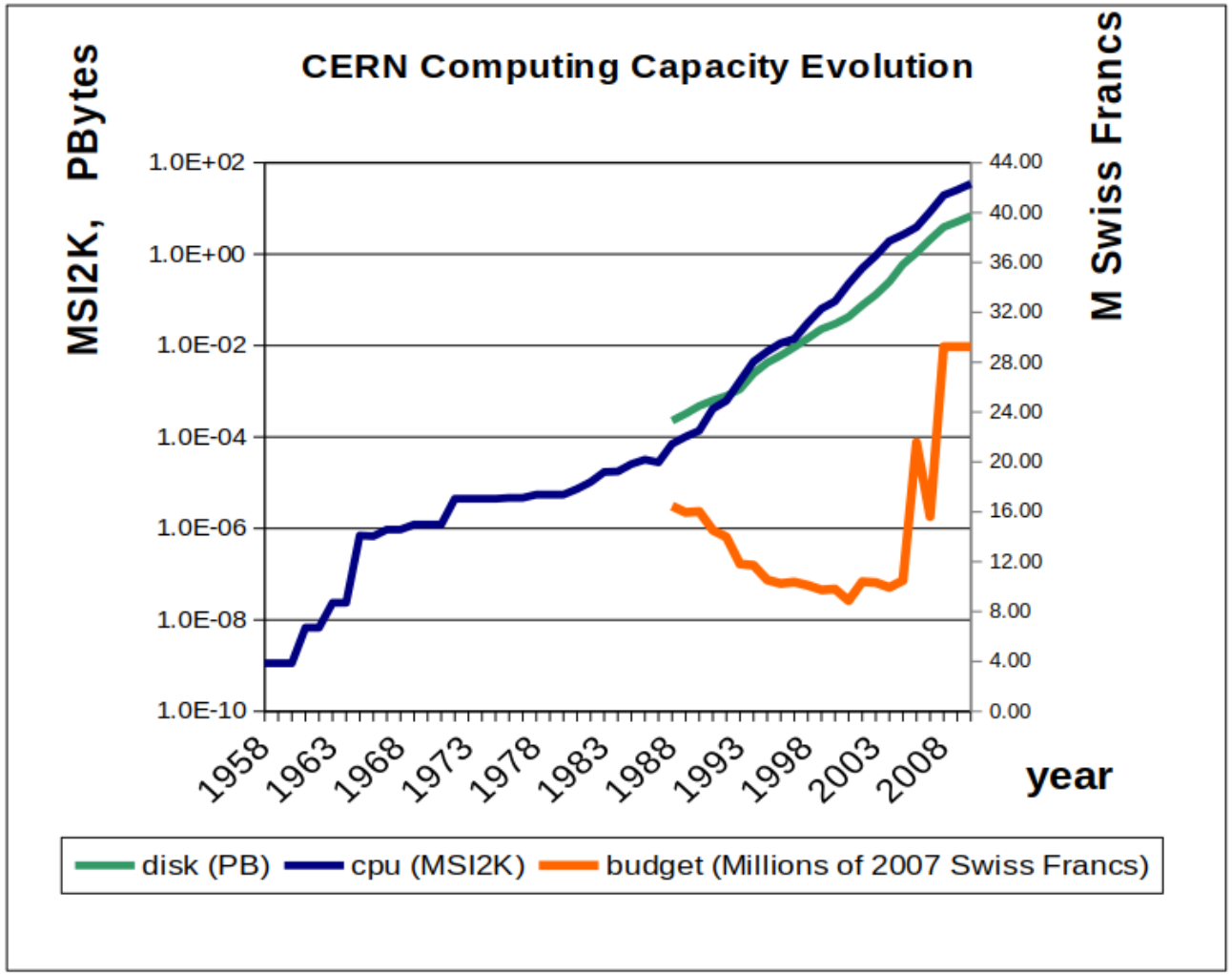
- Dive in the mud

A bit of history

Once upon a time.....

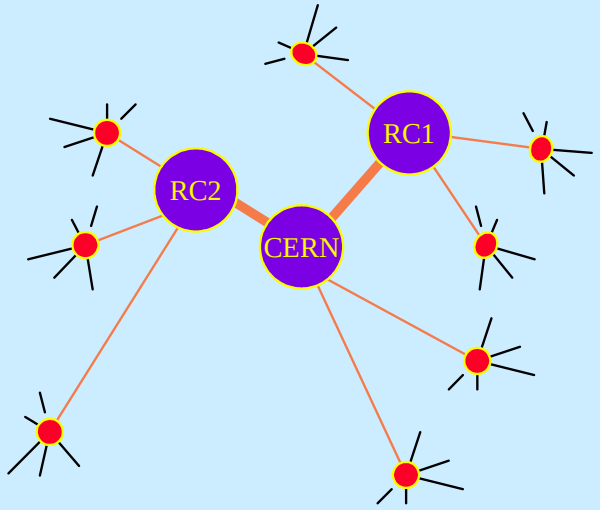


The LHC challenge

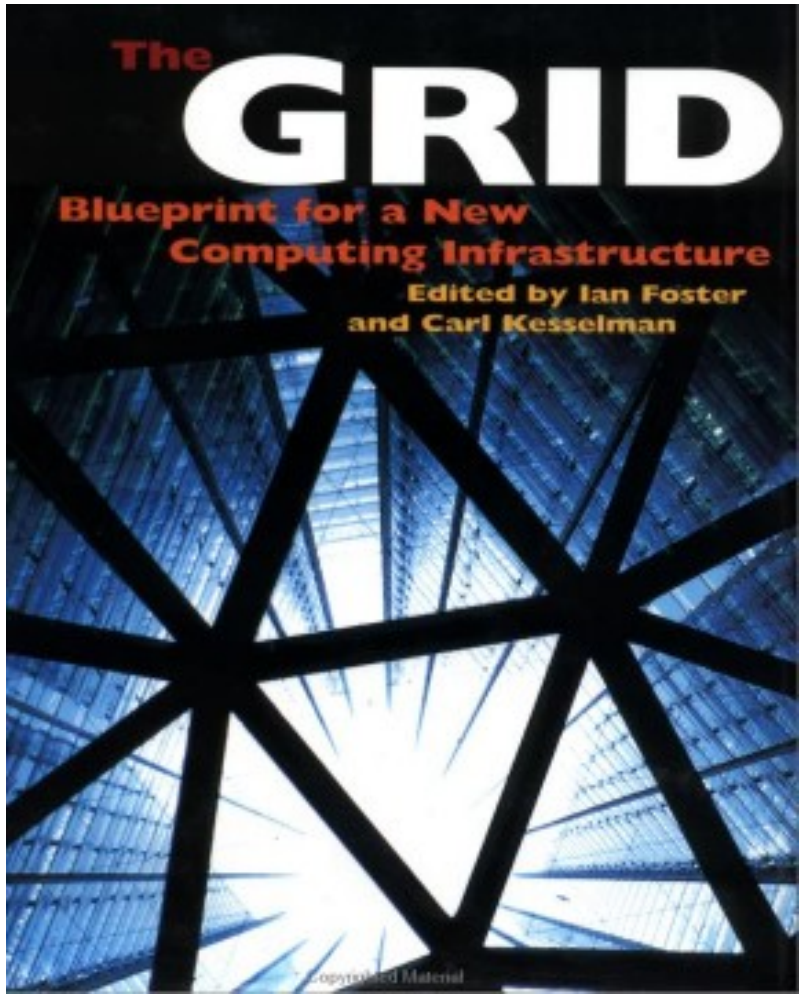


How to solve this?

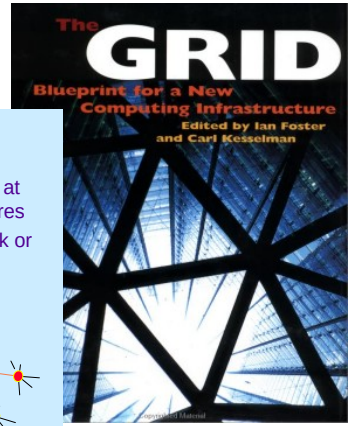
- Partially decentralized model
  - replicate the event data at about five regional centres
  - data transfer via network or movable media



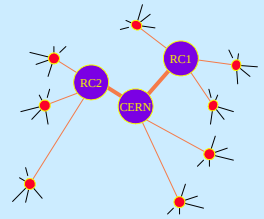
Then this happened

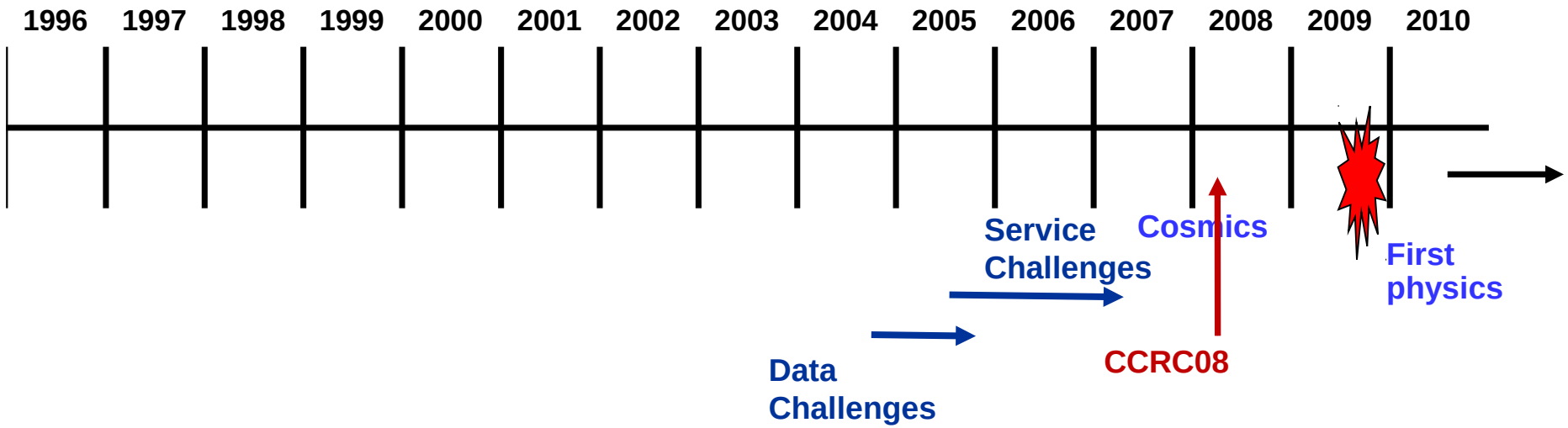


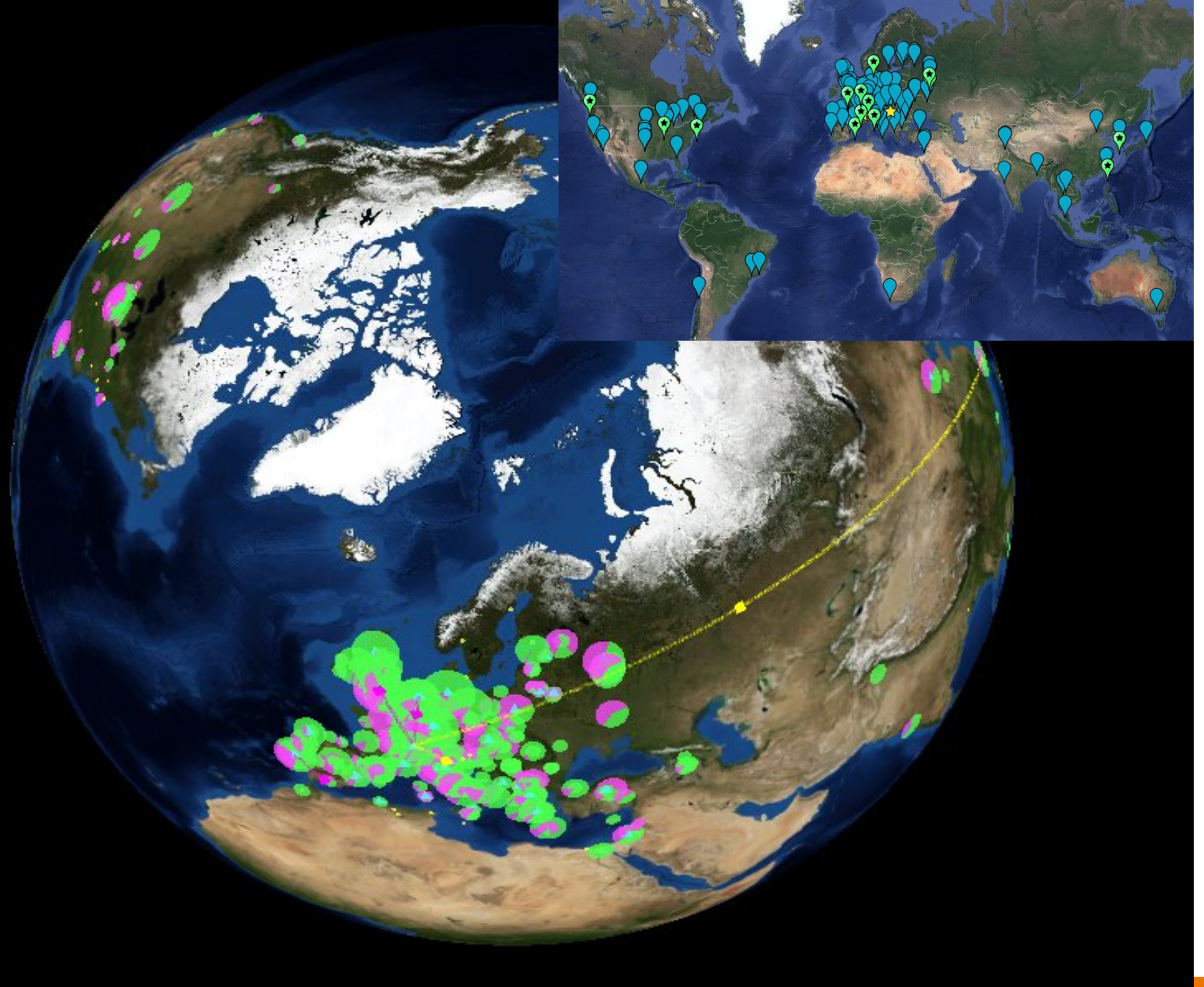
How did we get there?



- Partially decentralized model
  - replicate the event data at about five regional centres
  - data transfer via network or movable media





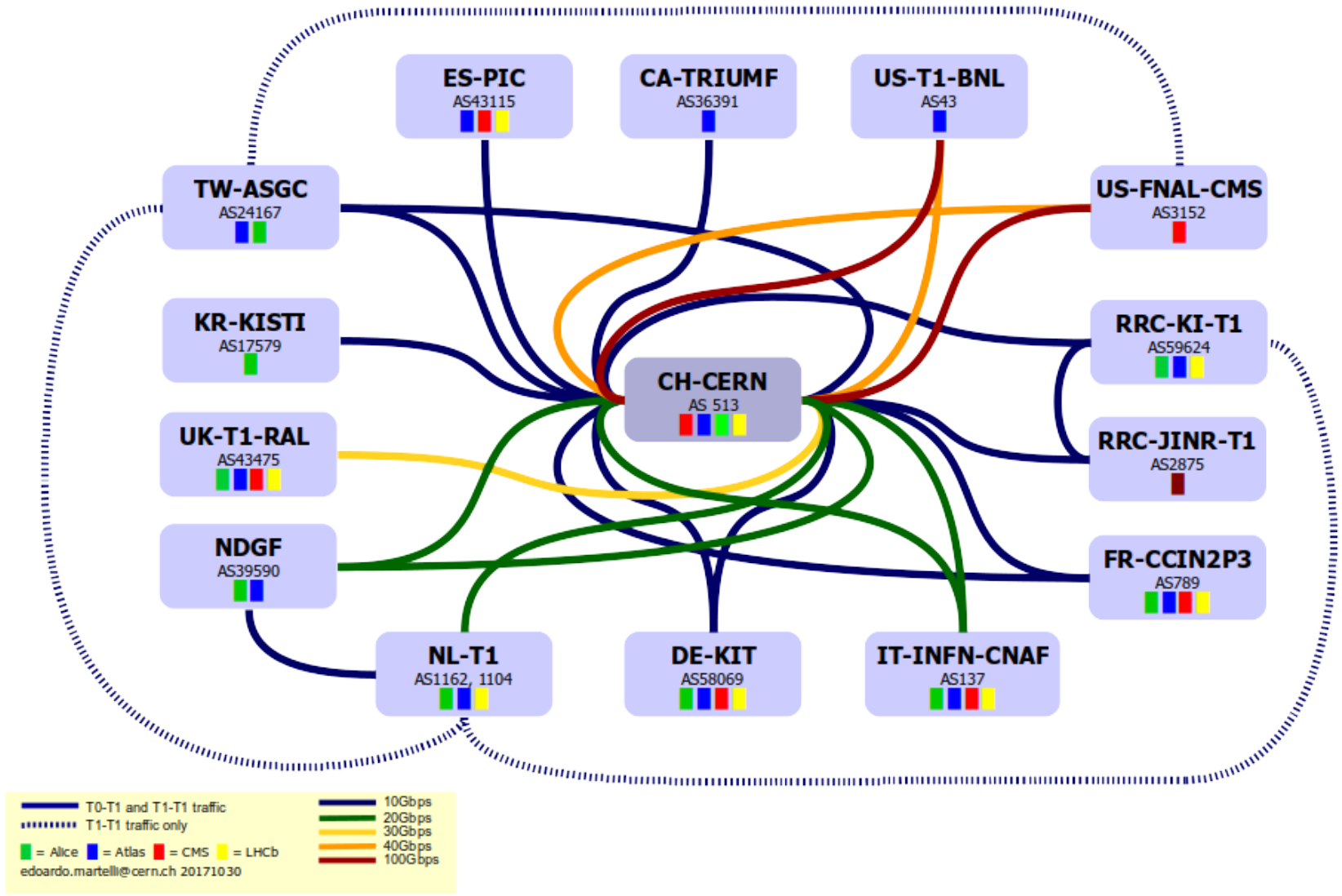


WLCG
Worldwide LHC Computing Grid

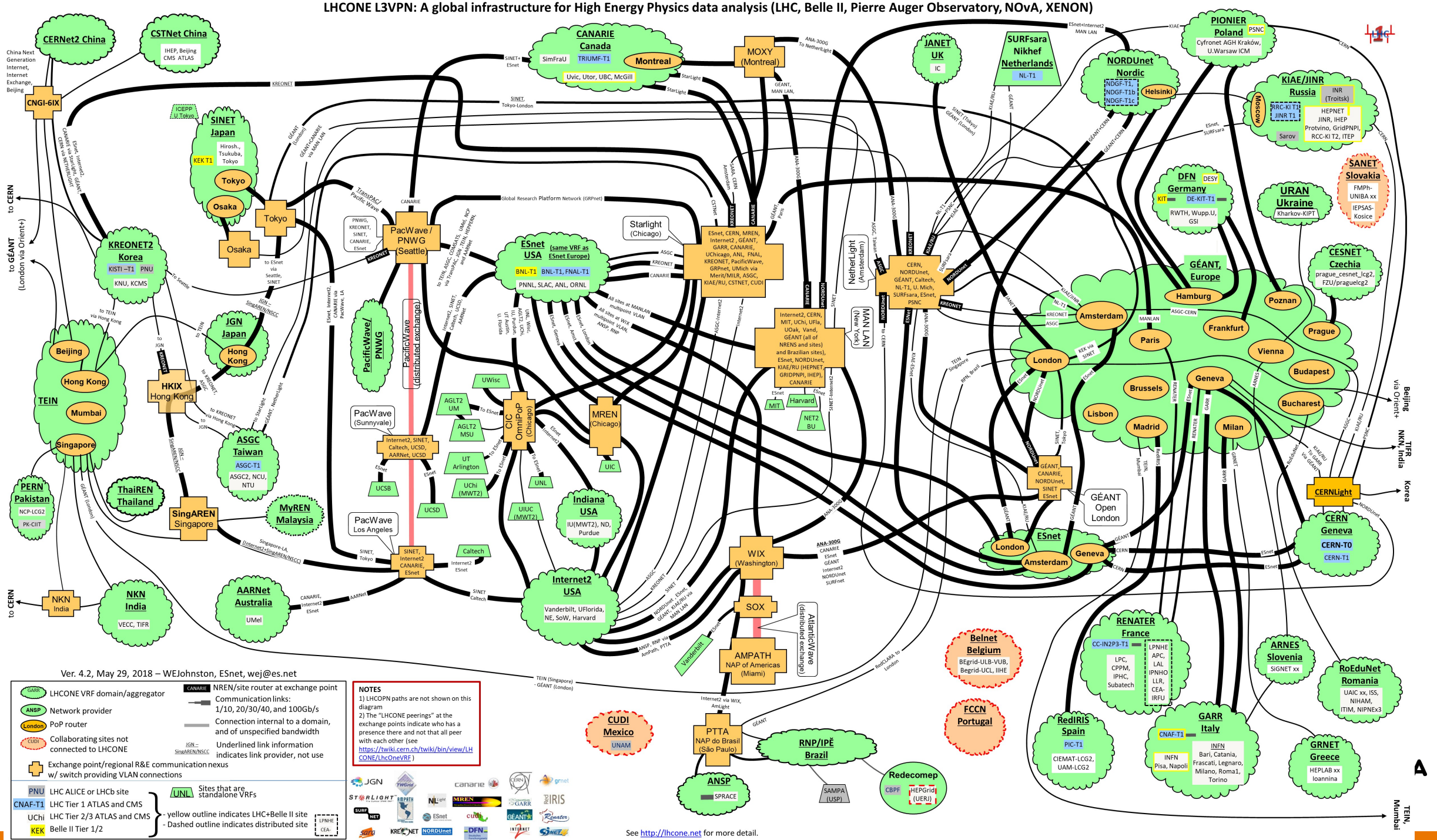
- 42 countries
- 170 datacenters
- 50-70PB of data per year
- 800000 jobs per day
- 500PB disk
- 750PB tape

SURF SARA

LHCOPN


ES-PIC
AS43115
CA-TRIUMF
AS36391
US-T1-BNL
AS43
TW-ASGC
AS24167
US-FNAL-CMS
AS3152
KR-KISTI
AS17579
RRC-KI-T1
AS59624
CH-CERN
AS 513
UK-T1-RAL
AS43475
RRC-JINR-T1
AS2875
NDGF
AS39590
FR-CCIN2P3
AS789
NL-T1
AS1162, 1104
DE-KIT
AS58069
IT-INFN-CNAF
AS137
T0-T1 and T1-T1 traffic
T1-T1 traffic only
= Alice = Atlas = CMS = LHCb
edoardo.martelli@cern.ch 20171030
10Gbps
20Gbps
30Gbps
40Gbps
100Gbps

LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON)



Ver. 4.2, May 29, 2018 – WEJohnston, ESnet, wej@es.net

- LHCONE VRF domain/aggregator
- Network provider
- PoP router
- Collaborating sites not connected to LHCONE
- Exchange point/regional R&E communication nexus w/ switch providing VLAN connections
- NREN/site router at exchange point
- Communication links: 1/10, 20/30/40, and 100Gb/s
- Connection internal to a domain, and of unspecified bandwidth
- Underlined link information indicates link provider, not use
- Sites that are standalone VRFs

PNU LHC ALICE or LHCb site
CNAF-T1 LHC Tier 1 ATLAS and CMS
UChi LHC Tier 2/3 ATLAS and CMS
KEK Belle II Tier 1/2
- yellow outline indicates LHC+Belle II site
- Dashed outline indicates distributed site

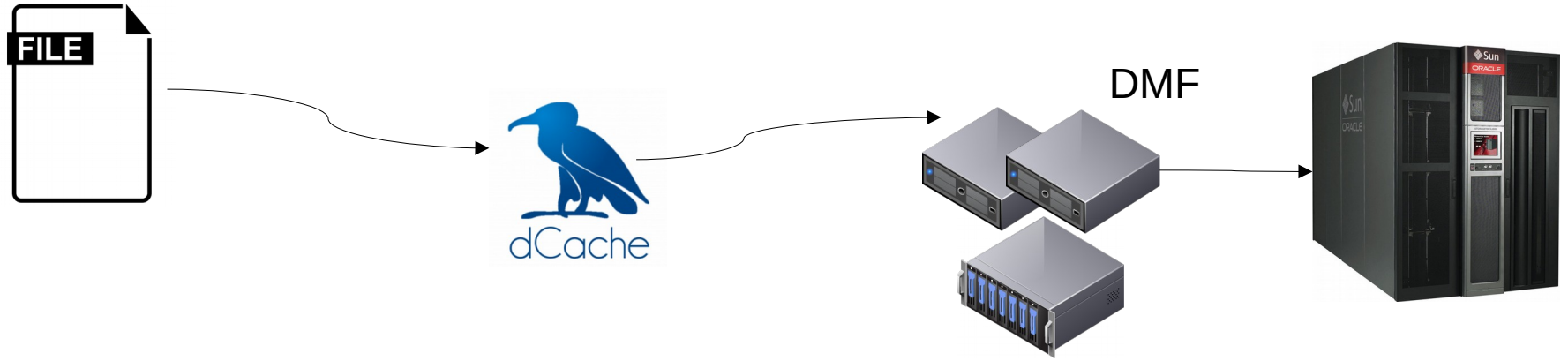
NOTES
1) LHCOPN paths are not shown on this diagram
2) The "LHCONE peerings" at the exchange points indicate who has a presence there and not that all peer with each other (see https://twiki.cern.ch/twiki/bin/view/LHCONE/LhcOneVRF)

See http://lhcone.net for more detail.

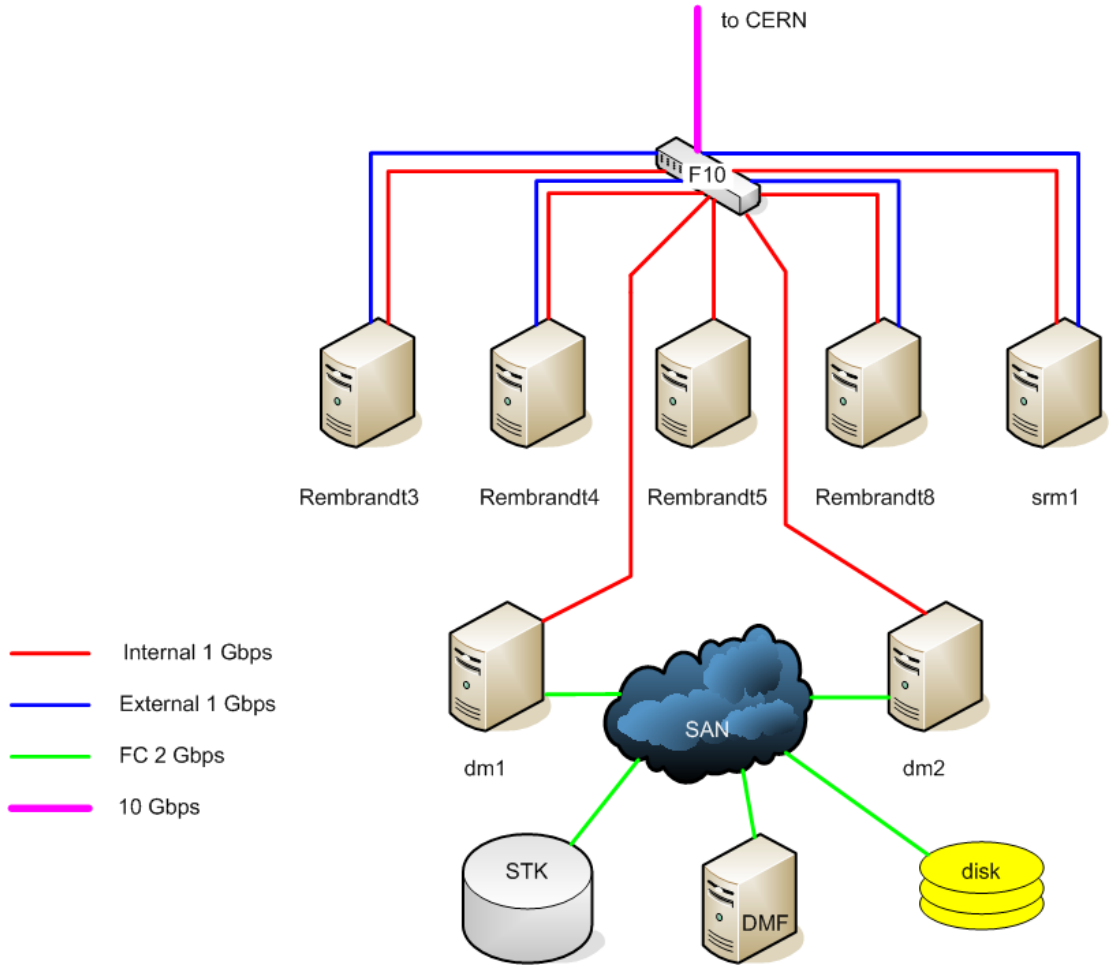

Things that we do at SURFsara

dCache

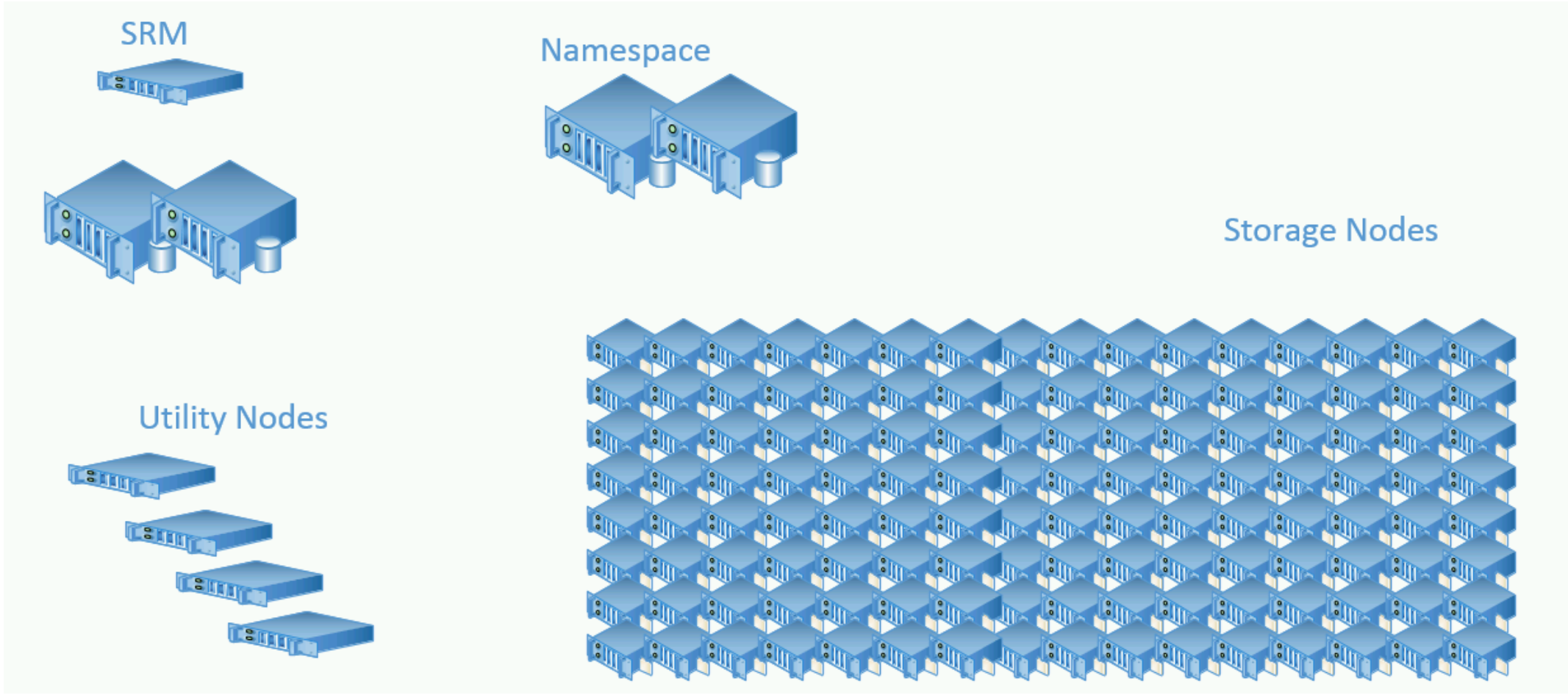
- Manages a large number of disk servers (NAS)
- Put, Get, Delete
- Data is immutable
- Seemless integration with tape systems



2005


to CERN
F10
Rembrandt3
Rembrandt4
Rembrandt5
Rembrandt8
srm1
dm1
SAN
dm2
STK
DMF
disk
Internal 1 Gbps
External 1 Gbps
FC 2 Gbps
10 Gbps

2018

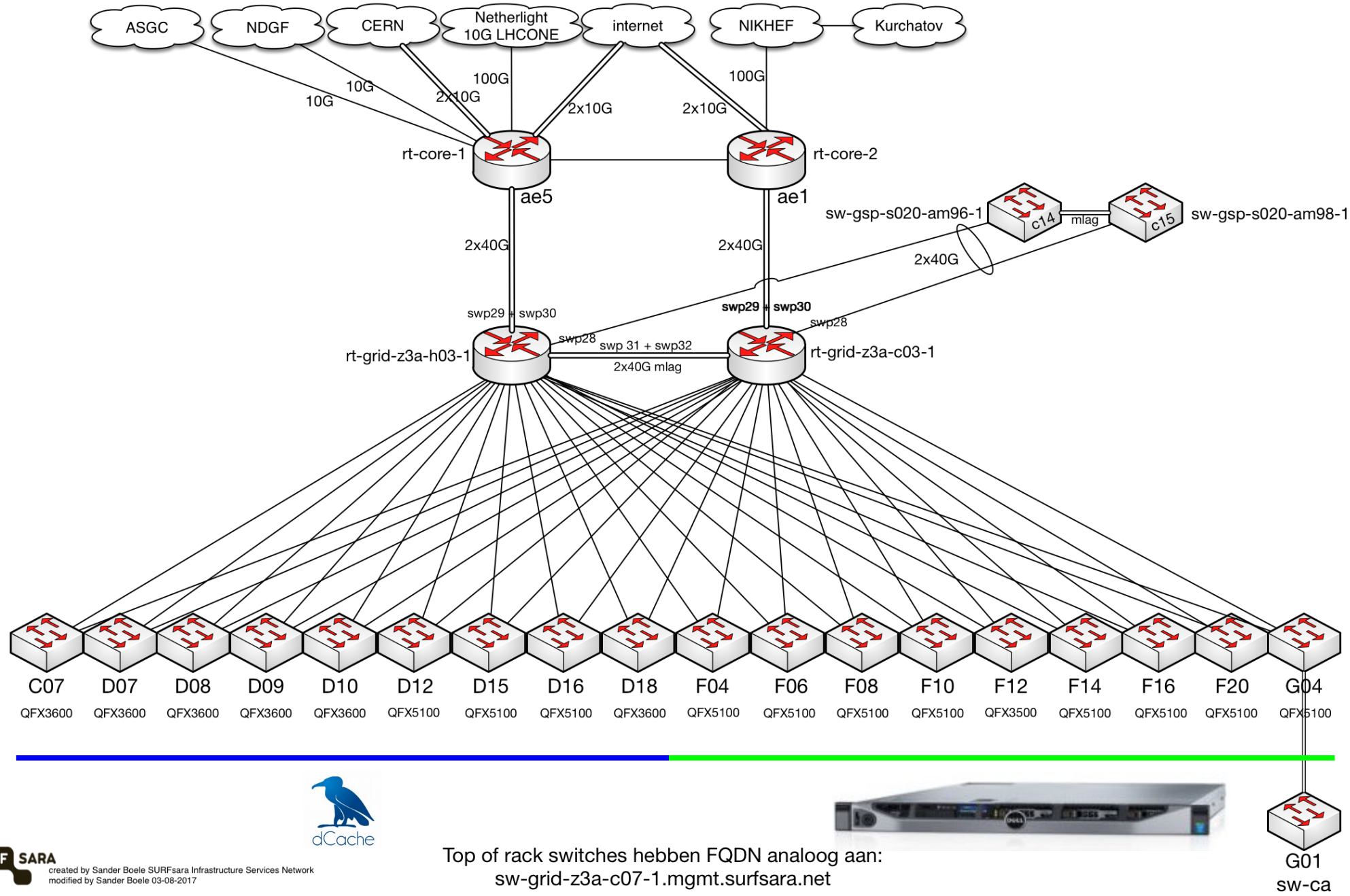


dCache in numbers

- 14PB storage capacity
- Manages 49PB on disk and tape
- Last year: 117 PB read, 22PB written, 7PB stored to tape and 8PB restored from tape
- 128 disk pool nodes with 80-400TB each

SURF SARA

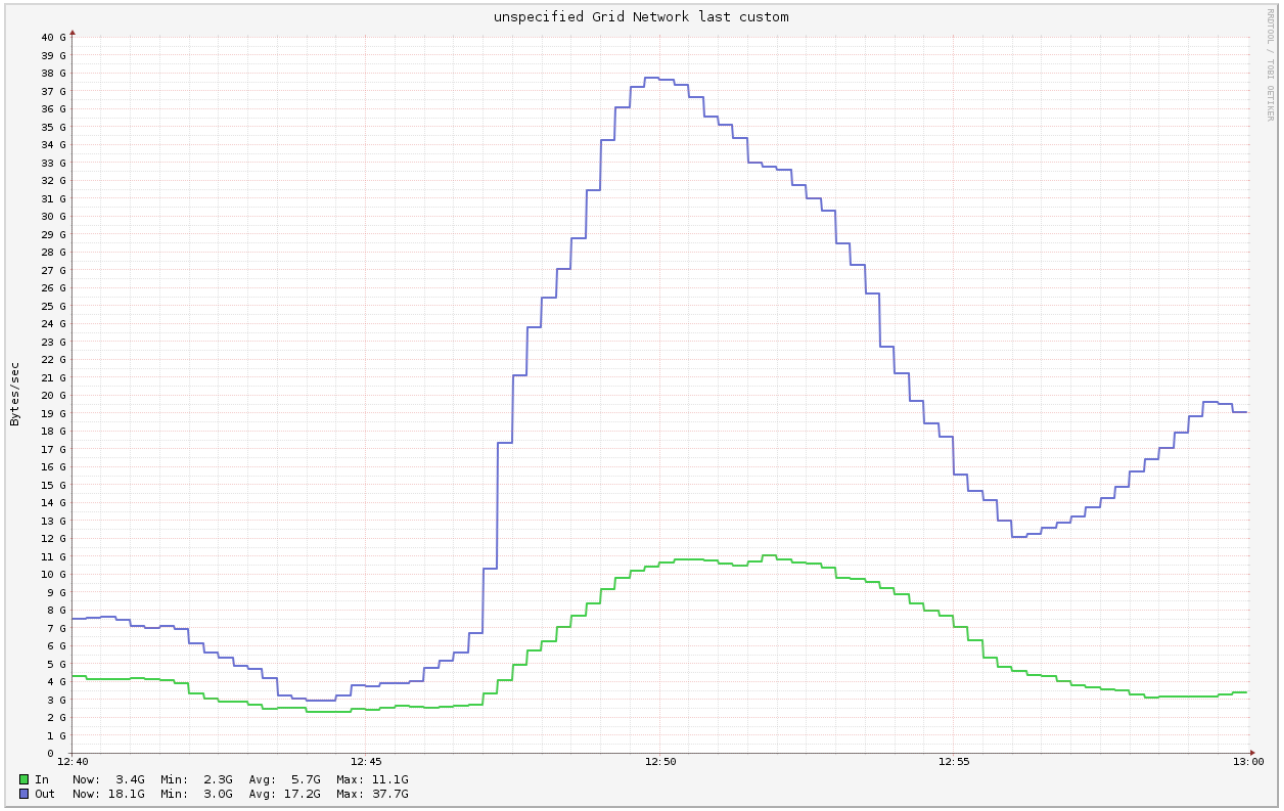
GRID netwerk

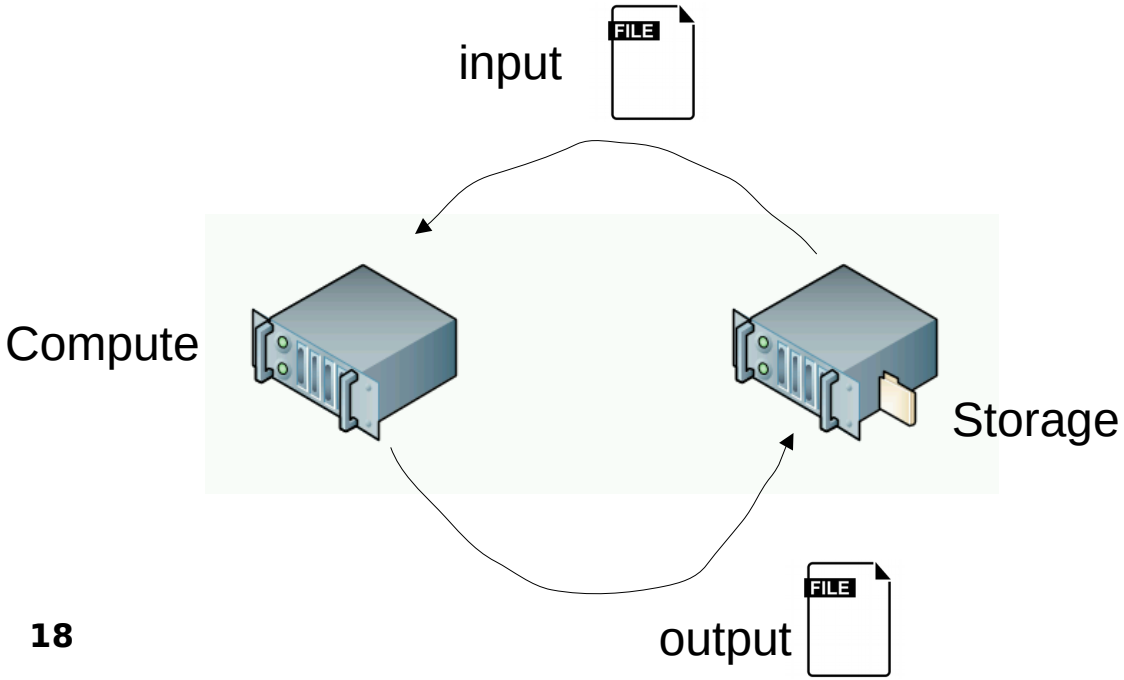


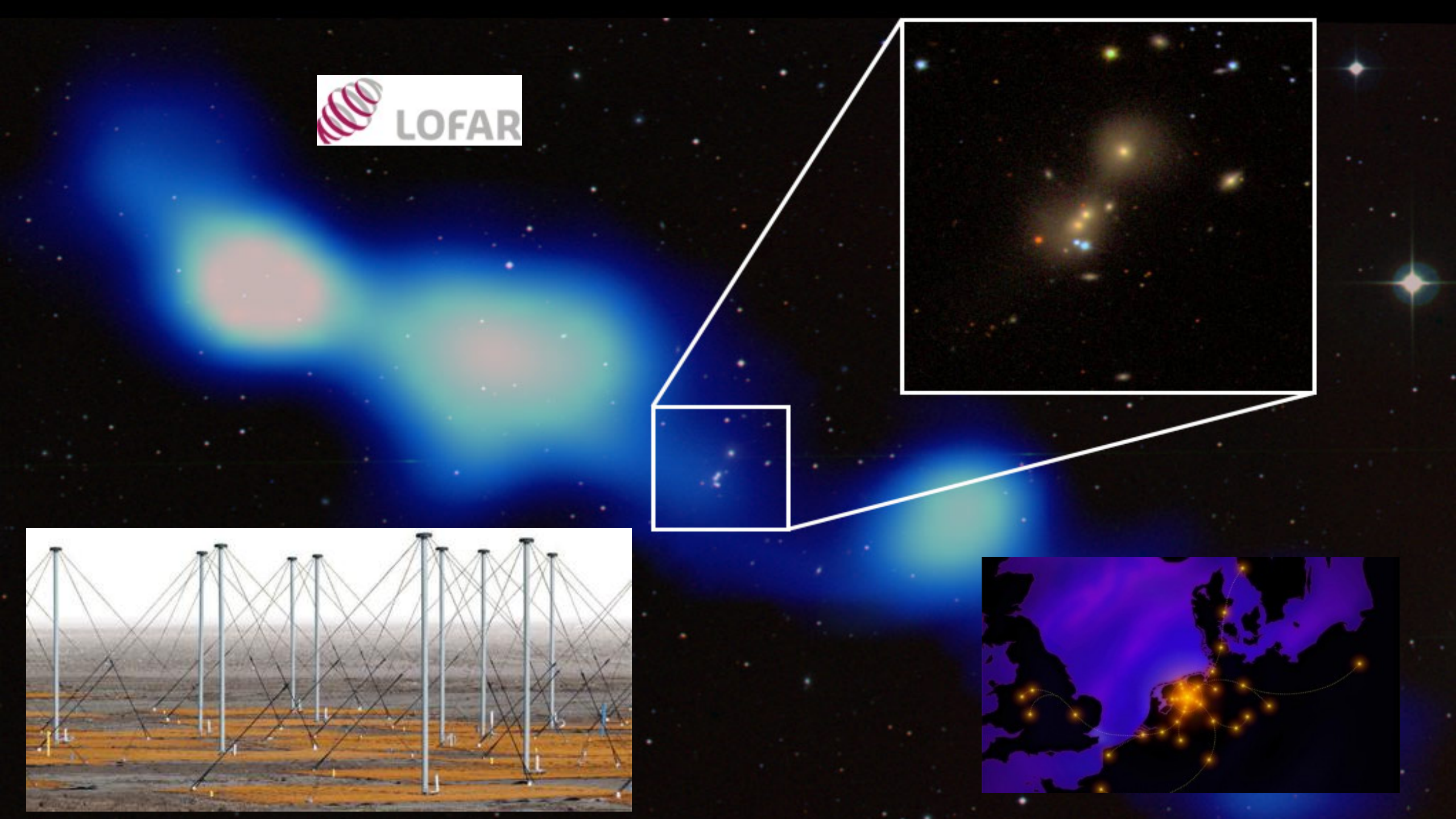
Top of rack switches hebben FQDN analoog aan:
sw-grid-z3a-c07-1.mgmt.surfsara.net

Data processing

- Thousands of running jobs processing data






LOFAR

TROPOMI

NO_2 Concentration [molec/cm^2]

0.1 0.5 1.0 1.5

1e16

project MinE

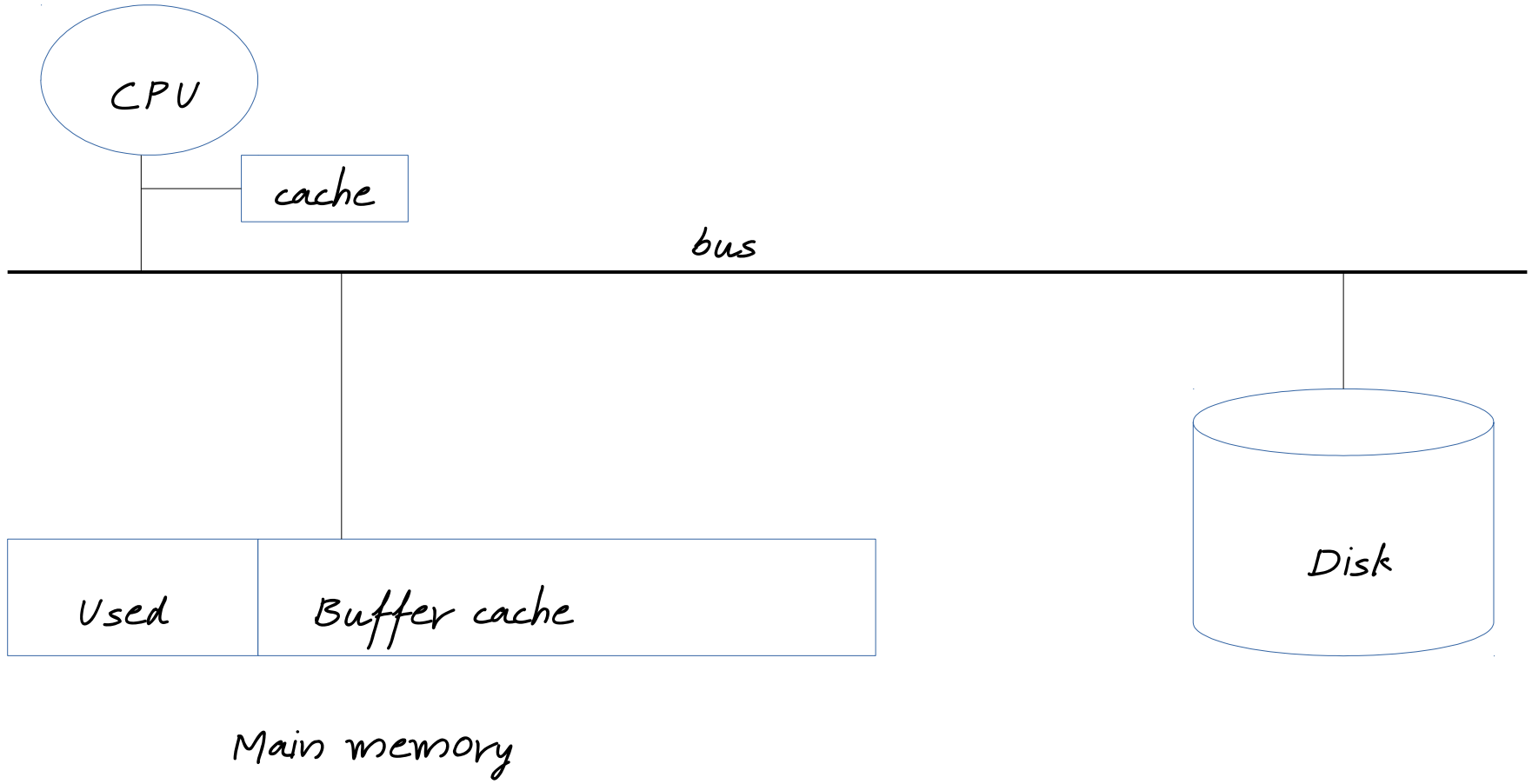
Make it yours

Dive in the mud

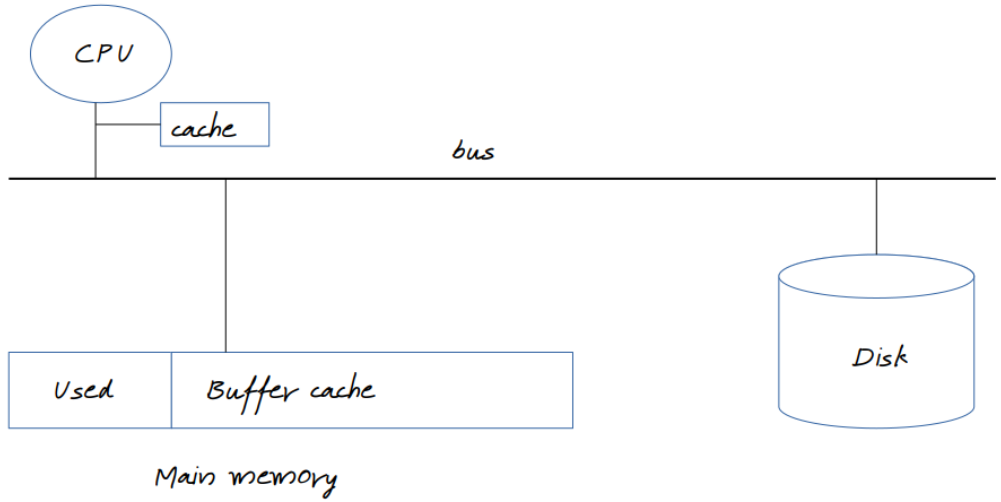
Testing

- Tools like Bonnie++ or iozone are fine but the best test is doing what your users are going to do.
- What applications are going to run to transfer data?
- How big are files typically
- What block sizes are used to read/write the data?
- Do people do random I/O or not?
- There is one other thing,...

The Buffer Cache



Testing



- Reading from disk is a lot slower than from memory
- Caching in memory is great
- Written files are often read back soon after
- Save source code in your editor and start compiling it
- Save a pdf you have got by mail and opening it
- Files are often read more than once in a short time span
- Forward an email
- Reduce the effects of the file system buffer cache as much as possible while testing. You want to test the disk subsystem, network card and CPU, not memory

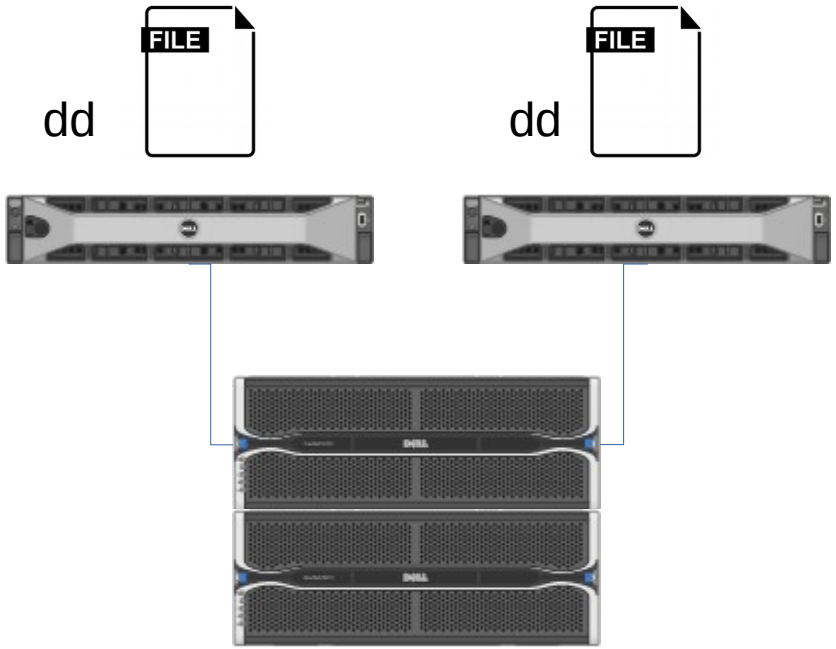
Testing: dd

- Test only the disk system

- Receipe:
- Create a number of testfiles. The size of the testfiles should be what you will expect in production.
- In total, the amount of space taken by the testfiles should be considerably larger than the amount of memory in the machine. We use 25x more than there is memory.
- Pick at random files to read for testing. Chances are slim that you pick a file that is still in memory from a previous test.
- Spawn a bunch of concurrent processes that are continously reading and writing files using dd. Reading and writing at the same time.
- Run dd’s with the same block size as your application would use.
- Let the tests run until the average throughput stabilises.

Testing: dd

- When multiple nodes are attached to the same disk system, then run the tests on all of them

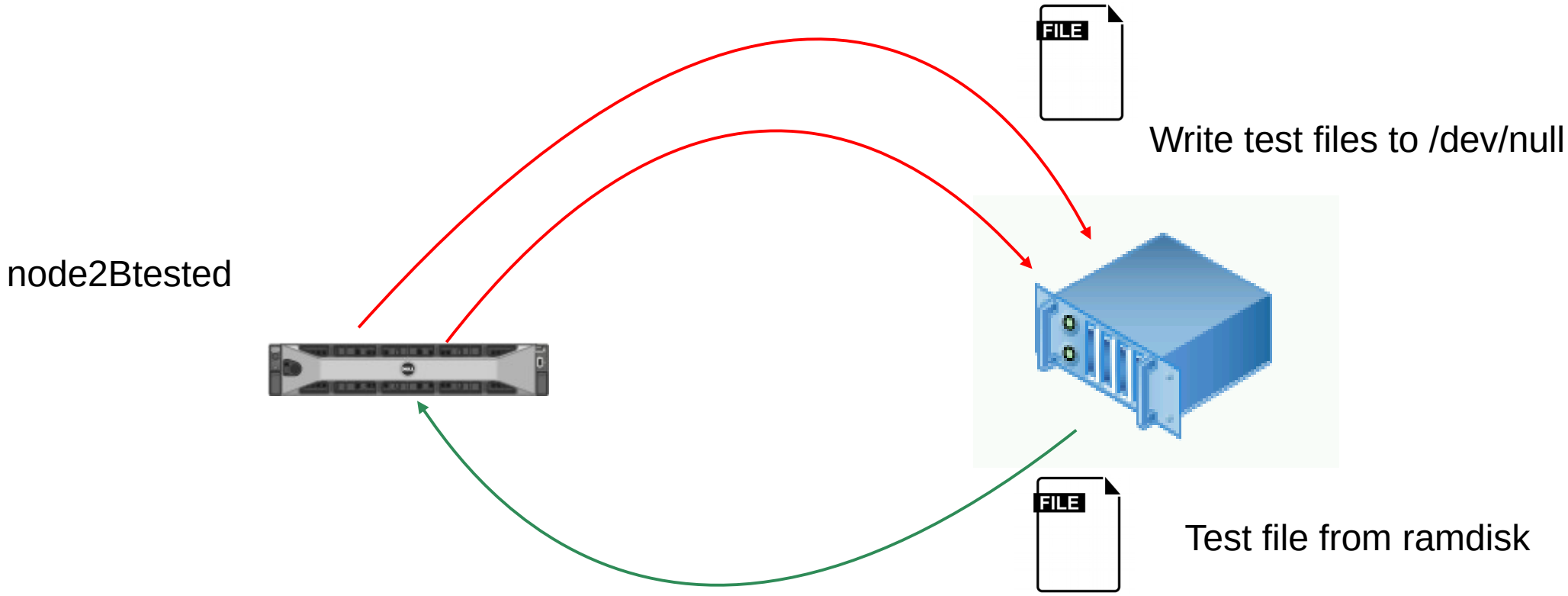


Testing: gridftp

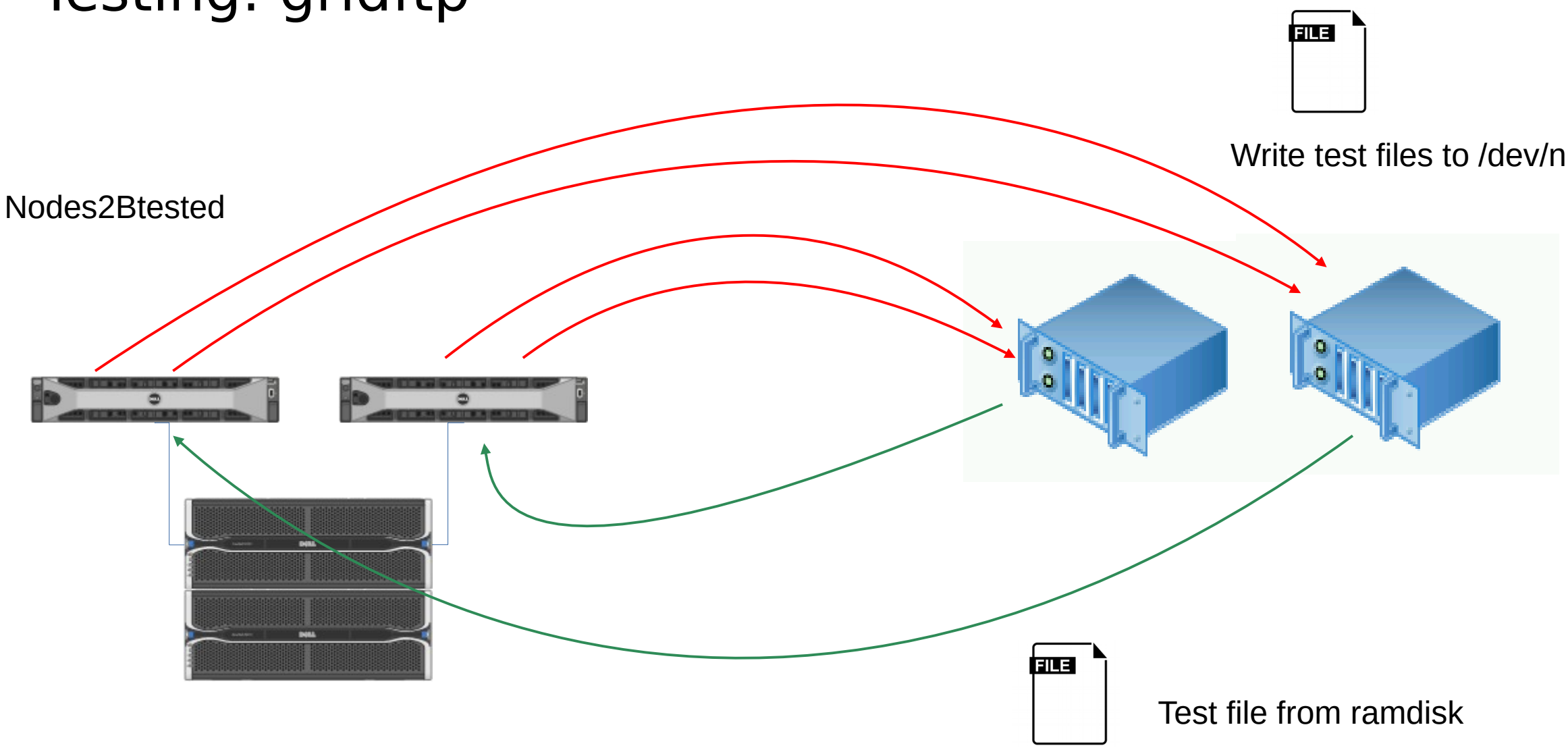
- Test disk and network

- Receipe:
- Same as dd tests.
- Setup gridftp server on the node that is tested and on another node on the same LAN having an equally performant NIC. The another node should not become a bottleneck. Avoid testing the disk system of the other node.
- Spawn multiple concurrent processes that transfer files to and receive files from another machine.
- We are talking about gridftp here but other protocols are fine as well.

Testing: gridftp



Testing: gridftp



Testing

- https://github.com/SURFsara-e-infra/dd_tests.git
- https://github.com/SURFsara-e-infra/gridftp_tests.git

Tuning

- Kernel IO scheduler
- cat /sys/block/sd???/queue/scheduler
- To our experience, when you use a hardware RAID controller, then this controller is probably a lot smarter than you are in controlling what HDDs do. We use noop.
- https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/performance_tuning_guide/chap-red_hat_enterprise_linux-performance_tuning_guide-storage_and_file_systems#sect-Red_Hat_Enterprise_Linux-Performance_Tuning_Guide-Considerations-Generic_tuning_considerations_for_file_systems

Tuning

- Set MTU to 9000 (normally MTU=1500)
- Better performance on high bandwidth networks
- Faster recovery from packet loss
- MTU=9000 all the way, but what if it isn’t?
- Fragmentation leading to overhead and performance degradation
- DF (don’t fragment) bit set, then no fragmentation, packet is dropped and ICMP packet send back so sender can adjust MTU
- Sometimes ICMP is blocked→ MTU black hole
- net.ipv4.tcp_mtu_probing = 1
- Disables MTU probing and turns it on when a black hole is detected
- net.ipv4.tcp_congestion_control = htcp
- http://fasterdata.es.net/host-tuning/linux/

Tuning

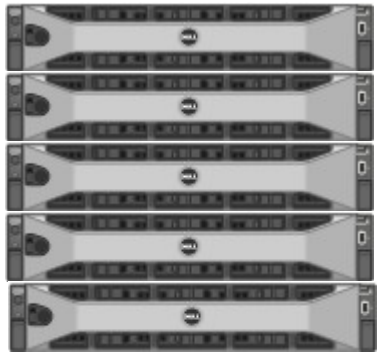
- Send and receive buffer sizes
- net.ipv4.tcp_rmem, net.ipv4.tcp_wmem
- Modern kernels are getting better in auto tuning but it may be worthwhile to look at these autotuning settings
- Send and receive buffer sizes
- Example: net.ipv4.tcp_rmem='4096 87380 8388608'
  - 4096= minimum buffer for each tcp connection
  - 87380=default buffer size
  - 8388608=maximum receive buffer for tcp connection
- https://wwwx.cs.unc.edu/~sparkst/howto/network_tuning.php
- https://paulgrevink.wordpress.com/2017/09/08/about-long-fat-networks-and-tcp-tuning/

Tuning

- Linux vm tuning
- vm.dirty_background_ratio, vm_dirty_ratio, vm.dirty_expire_centisecs,vm.dirty_writeback_centisecs
- Fast disk system with battery backupped RAID controller:
  - vm.dirty_background_ratio = 5
  - vm.dirty_ratio=10
- https://lonesysadmin.net/2013/12/22/better-linux-disk-caching-performance-vm-dirty_ratio/
- Max connections
- net.core.somaxconn (default 128)

Concluding remarks

- Hardware considerations
  - Set of smaller nodes
  - Nodes with separate disk system



- HDDs, SSDs, NVMe
  - For what applications?
  - €€€?
- Scaling up
- Data migration scenario when hardware is decommissioned (outages? And for how long)
- Io requirements
  - We use 12MiB/s/TiB


NOT SURE IF PRESENTATION WAS SO GOOD
NO ONE HAD ANY QUESTIONS
OR NO ONE WAS PAYING ATTENTION

SURF SARA