[disclaimer: this is a personal view any resemblance to reality is pure coincidence]

[2nd disclaimer: this presentation is slightly biased on storage]

NorduGrid 2017

27-30 June 2017
Europe/Oslo timezone

CERN-IT challenges

the byte, the core and the bit

Xavier Espinal (CERN-IT/ST)

with input from

Arne Wiebalck (IT/CM), Carles Kishimoto (IT/CS) and Ben Jones (IT/CM)

IT-ST

Provide the computing technologies needed by our scientific communities

Local (Users,+)
Experiment (LHC,+)
Global (WLCG,+)

Run computing services at high efficiency with reduced costs

Data resilience
CPU optimization (scheduler)
Efficient network topography

Optimize human resources on operations and maintenance

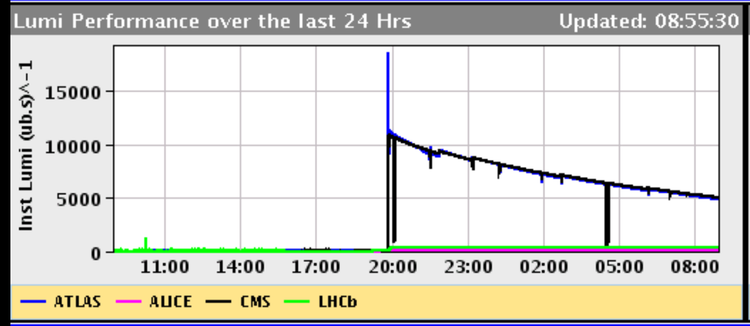
Deployment
Maintenance
Update

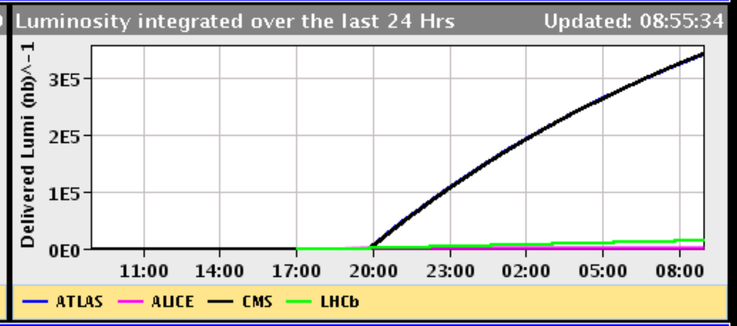
CERN-IT *now*

IT Overview ☆ Zoom Out Last 24 hours

COMPUTING		STORAGE		NETWORK	
Servers (Meyrin)	Cores (Meyrin)	Disks (Meyrin)	Tape Drives	Routers	Star Points
10.3 K	147.1 K	59.1 K	104	221	667
Servers (Wigner)	Cores (Wigner)	Disks (Wigner)	Tape Cartridges	Switches	Wifi Points
3.5 K	56.0 K	29.7 K	24.7 K	3.8 K	2.0 K

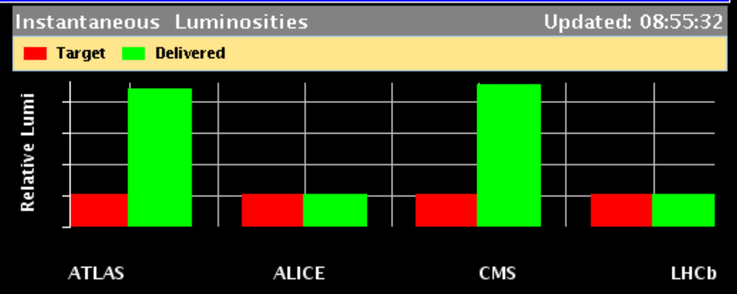
20-Jun-2017 08:55:34 Fill #: 5848 Energy: 6499 GeV I(B1): 1.33e+14 I(B2): 1.37e+14





STABLE BEAMS

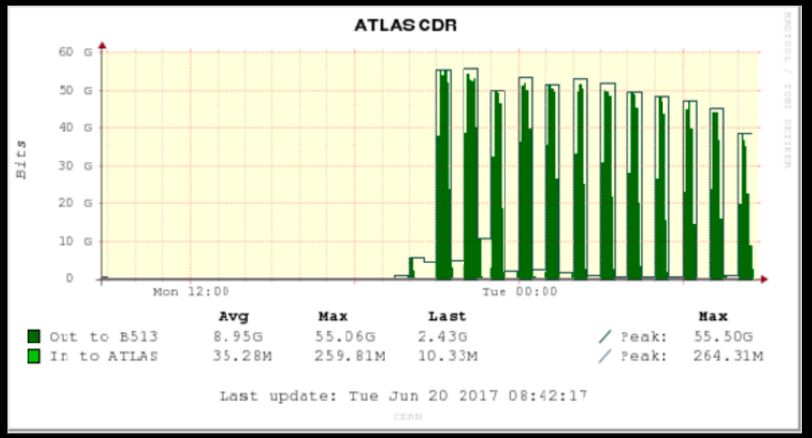
	Luminosity [(ub.s)^-1]	Fill Lumi (nb)^-1
ATLAS	4864.98	341007.3
ALICE	2.59	121.3
CMS	5013.19	342430.7
LHCb	292.11	13728.9

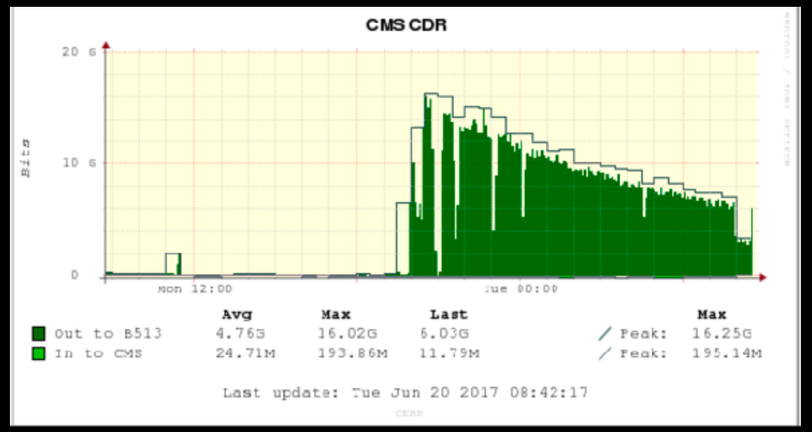


ALICE Target Instantaneous Lumi = 2.6 Hz/ub

LHCb Target Instantaneous Lumi = 300.55475 Hz/ub

20/06/2017 @8:50





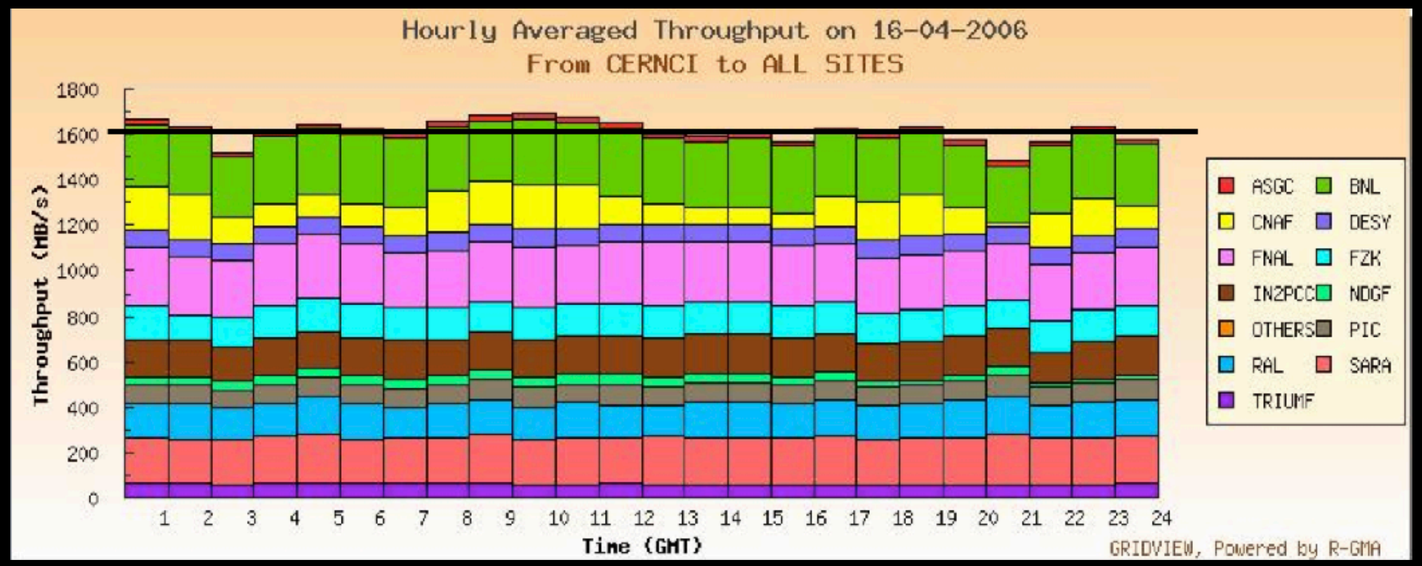
IT-ST

The challenge continues: the goal is unchanged

circa **2006**

Distributed computing (DC) exploration. WLCG Service Challenges. 1PB fit in 8 Racks. Clocks 1.86G/dualcore.10GE is a dream.

Physical space is an issue (commodity PCs as worker nodes). PUE not yet a figure. Network is scaling. 1000km of cables (1 CPU=1eth)



Service Challenge 4 - Goal: 1.6GB/s out of CERN

IT-ST

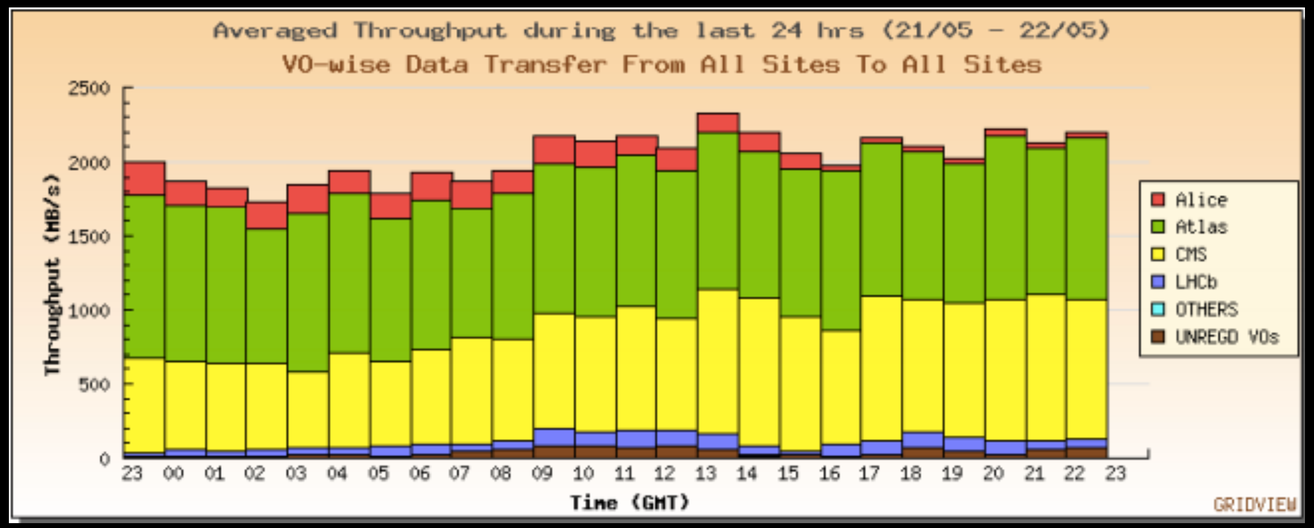
The goal: to provide a computing infrastructure to the experiments and the community to store and analyze data

The challenge continues: the goal is unchanged

circa **2009**

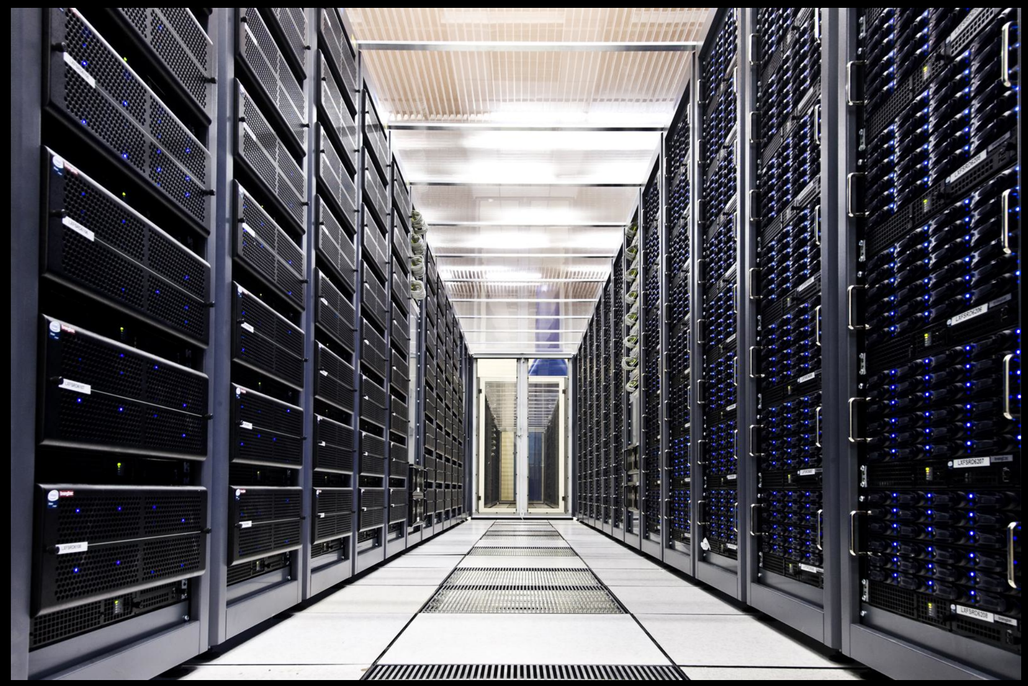
Phasing Run-I. CCRC&FDRs: DC consolidated. 1PB fit in 3 Racks. Clocks at 2.67G/quadcore. 10GE is luxury, 100Gbps on the horizon.

Power is an issue. Hot/cold corridors. Compact diskservers, compact-*pizza* nodes. Heat. PUE is a figure. LAN struggle to scale. 500km of cables.



CCRC-08

https://indico.cern.ch/event/23563/timetable/#20080613



IT-ST

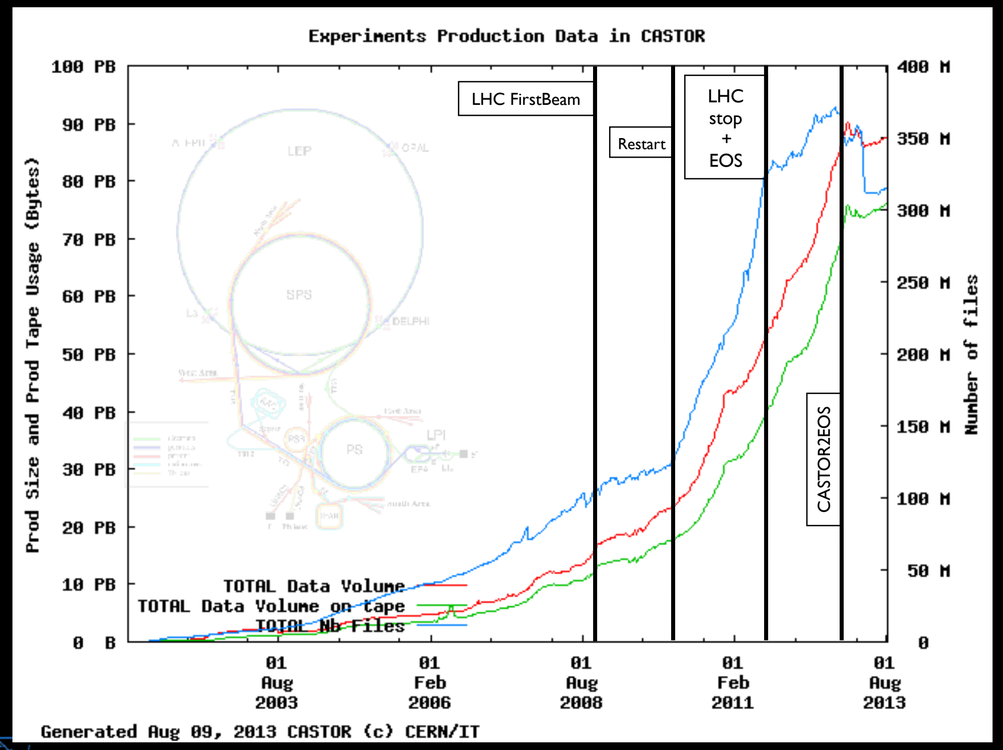
The goal: to provide a computing infrastructure to the experiments and the community to store and analyze data

The challenge continues: the goal is unchanged

circa **2012**

Phasing Run-II. DC paradigms shifting. 1PB fit in one Rack. Clocks at 2.4G/multicore. 10GE is the standard and 100Gbps in place (backbones, WAN)

Power consumption is a figure on tenders. Physical space freed. Networks upgraded. PUE “controlled”. 100km of cables.



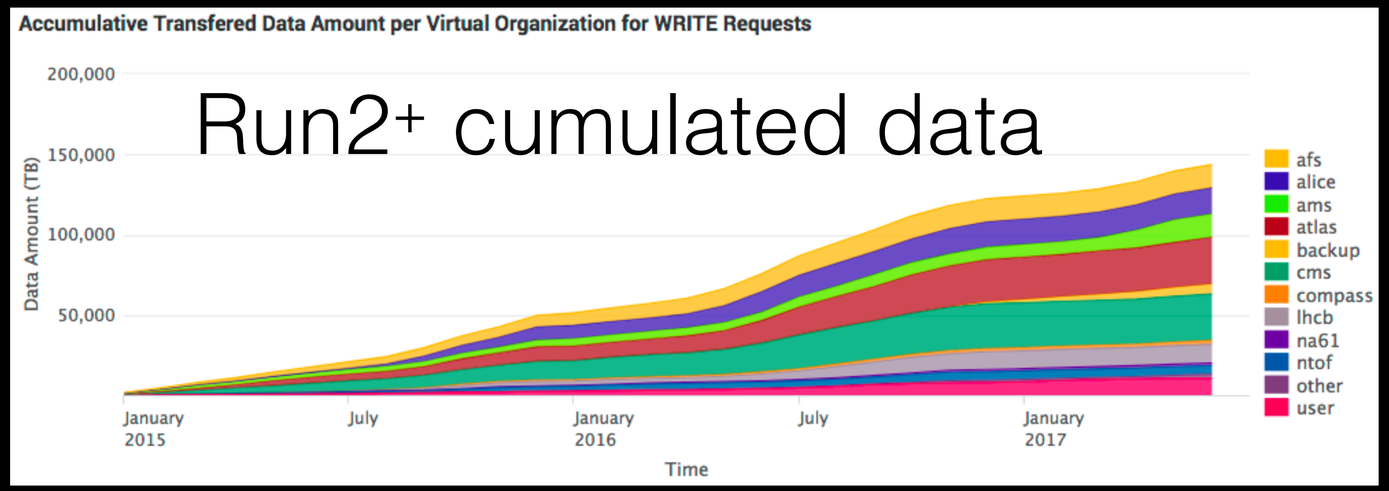
The goal: to provide a computing infrastructure to the experiments and the community to store and analyze data

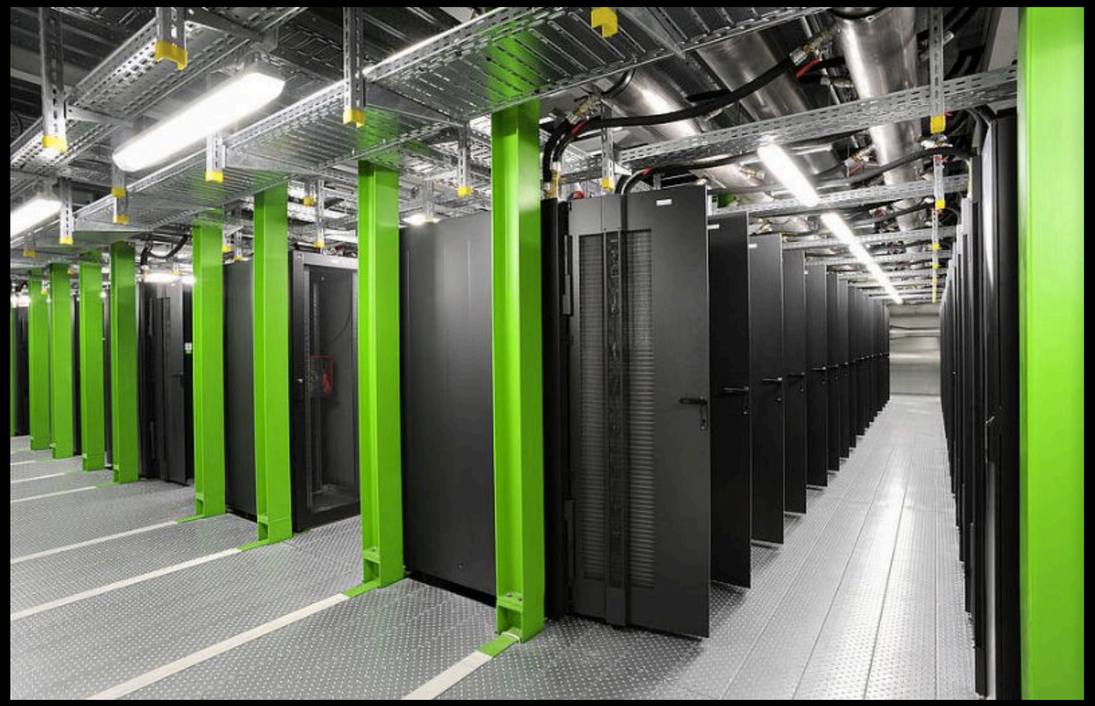
The challenge continues: the goal is unchanged

circa **2017**

Ending Run-II. DC model redesign. 1PB fit in single server (5U). Clocks at 2.4G/multicore. 10GE at the limit, 40GE next standard (~2018).

CCs getting "*empty*". Super racks: +kW, internal cabling. Super-compact servers. Green-IT. $$$ is the limit. 50km of cables.





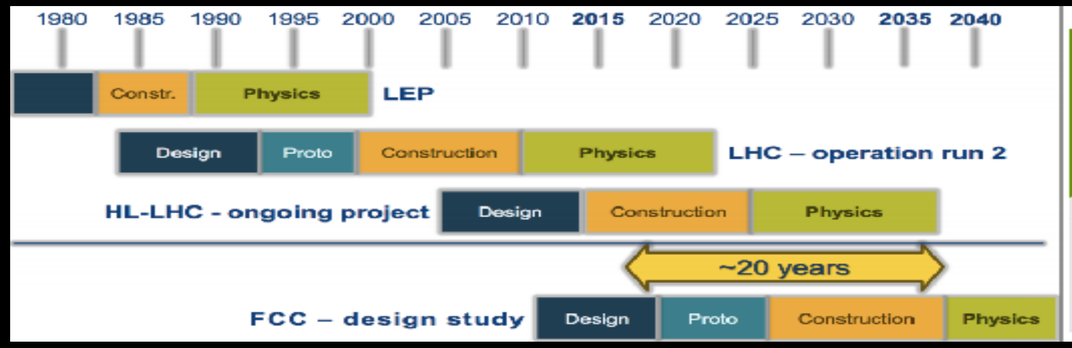
IT-ST

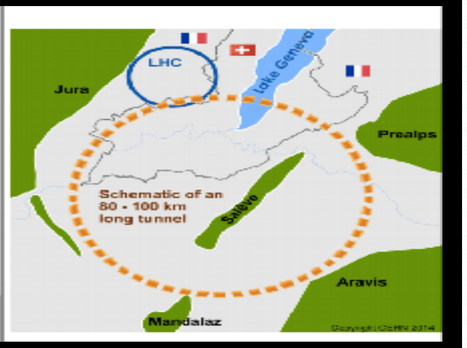
The goal: to provide a computing infrastructure to the experiments and the community to store and analyze data

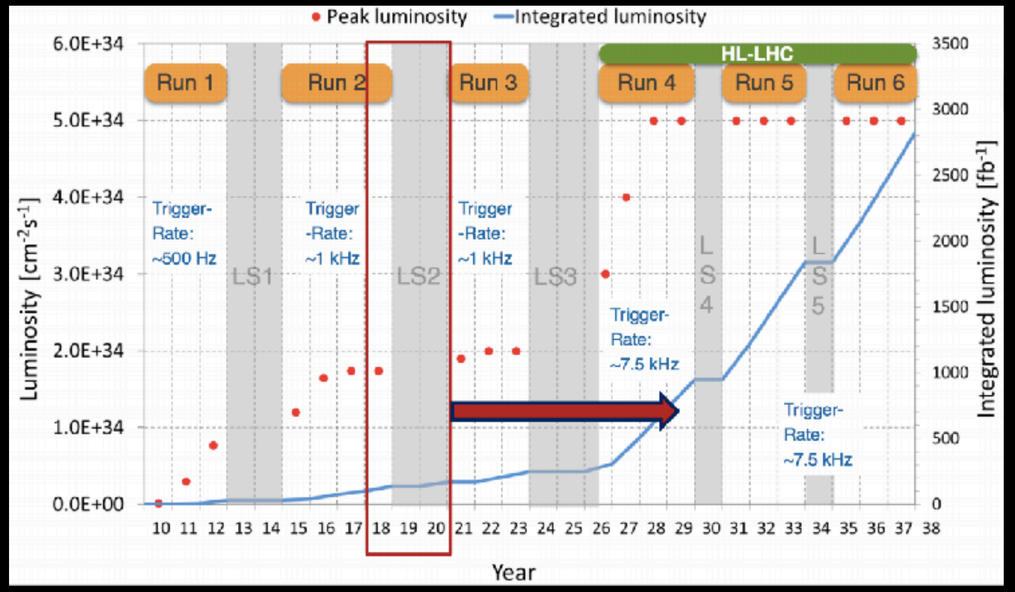
The challenge continues: the goal is unchanged

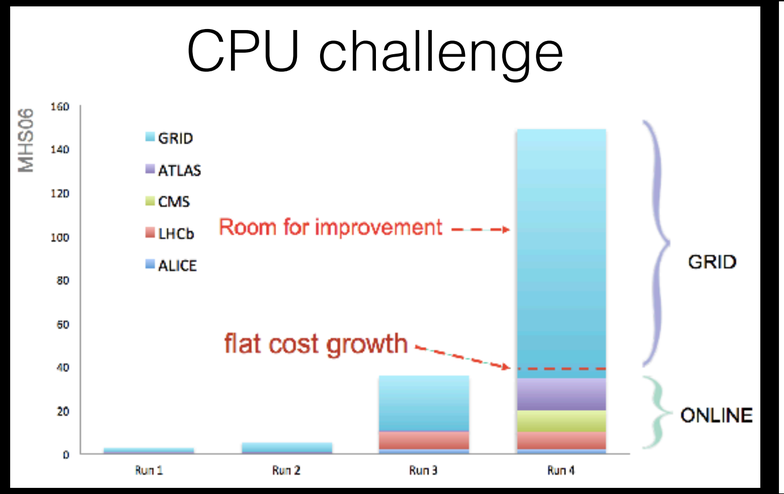
2019+

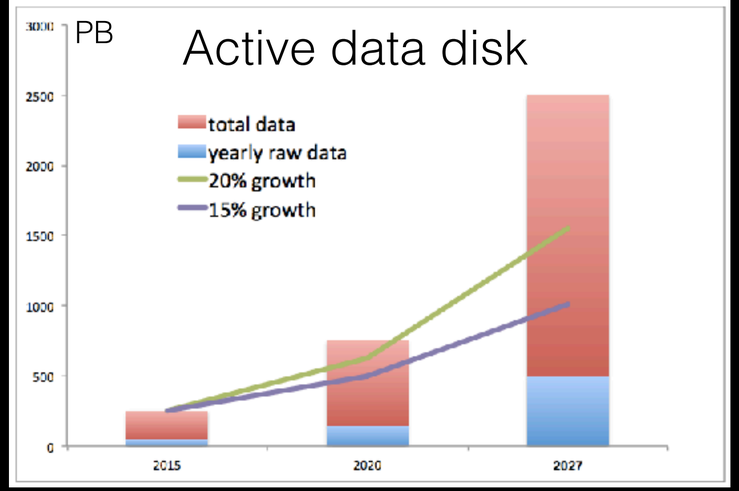
Prearing Run-III Don't dare to make predictions but need to address:

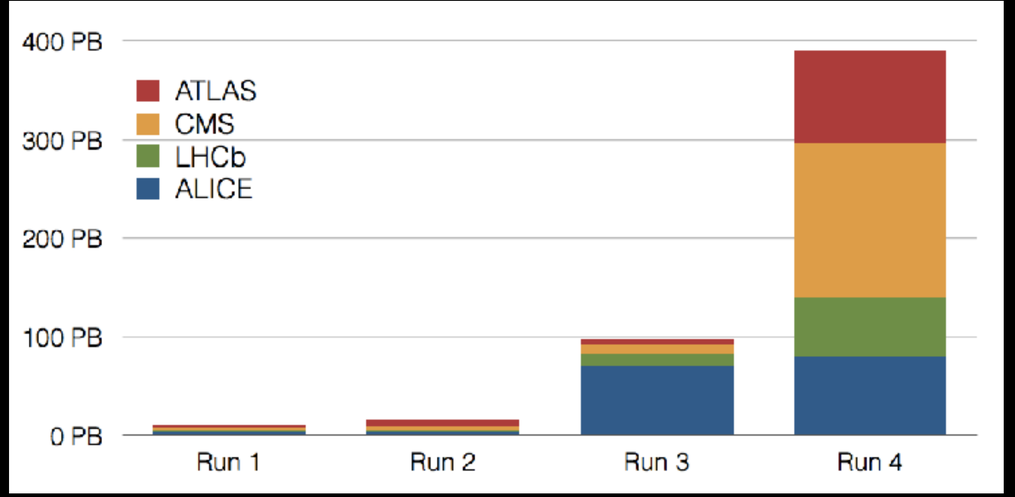












IT-ST

The goal: to provide a computing infrastructure to the experiments and the community to store and analyze data

There are three main actors ruling LHC computing

The byte: “byte’em and smile”

The core:
“I couldn’t
core less
about speed”

The byte:
“byte’em
and smile”

The core:
“I couldn’t core less about speed”

The byte:
“byte’em and smile”

The bit:
”that bitter feeling of miscommunication”

Present challenges: bytes, cores and bits

- **Data** store and data accessibility: tapes, disks, s3, fuse mounts, shared filesystems, clouds, globalized data access

- **Computing** resources: shares, schedulers vs. metaschedulers, pluggability, cloud computing, VMs, auth/authz, accounting

- **Networking**: simplificiation of Distributed Computing model is bound to networking evolution, LAN scaling (fat storage nodes), IPv6, WAN to 400Gbps(Tbps soon?), WAN to the node bottlenecks

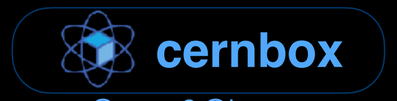
CERN-IT Storage Services

EOS - Main Storage Platform: elastic, adaptable, scalable

1.5B files
200PB

+1.2k
+50k

Data Recording
User Analysis
Data Processing



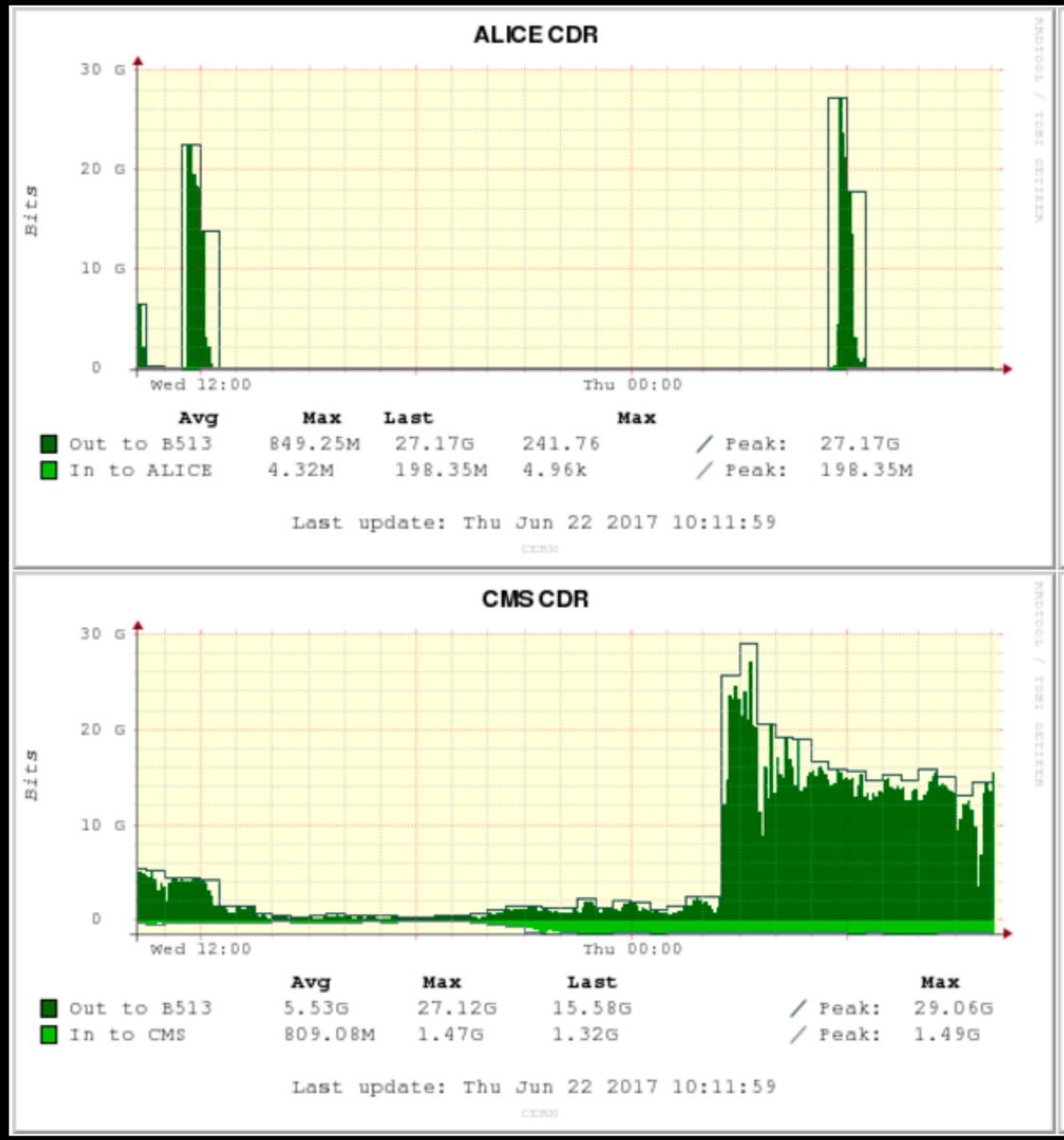
Sync&Share

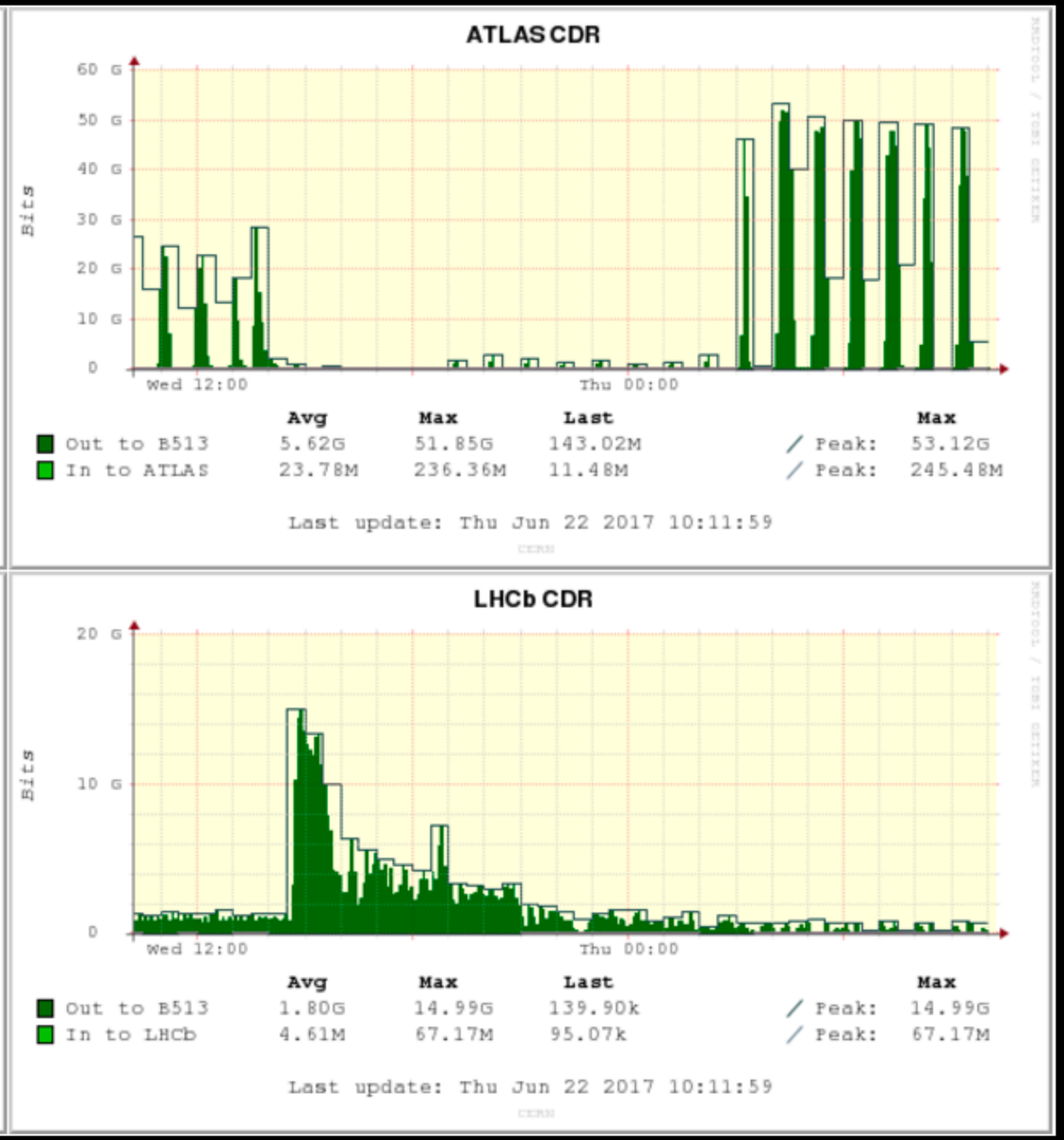
LHC Data in a shell
FUSE/batch

Quality on Demand provided by CEPH: Openstack, HPC, S3, CVMFS, NFS

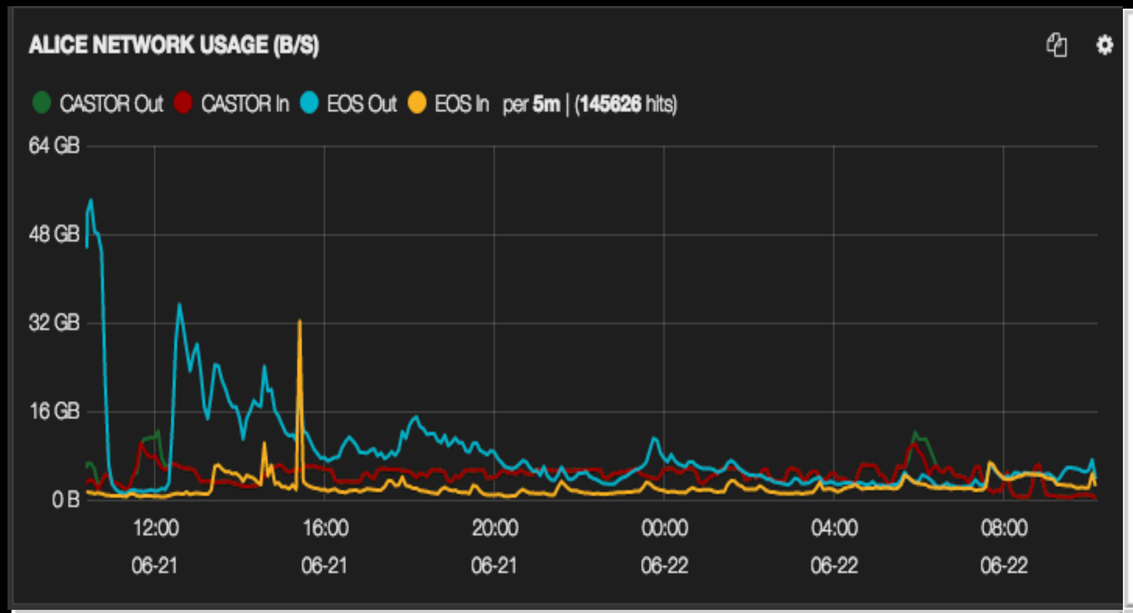
Openstack: VI+cinder
CVMFS
NFS/Filers and S3

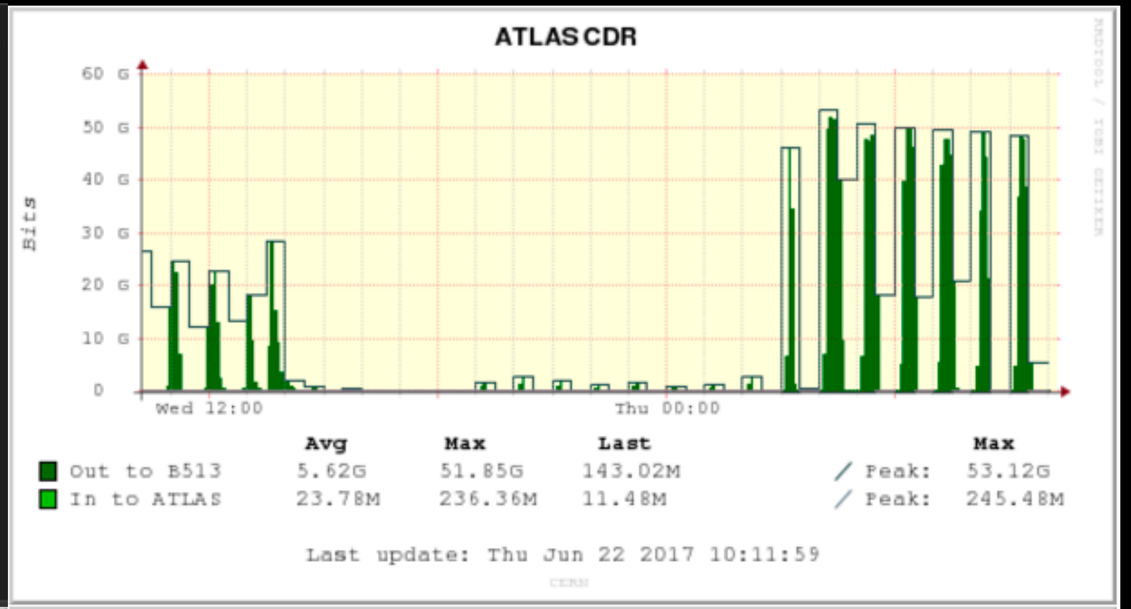
CERN-IT Storage Services: DAQ

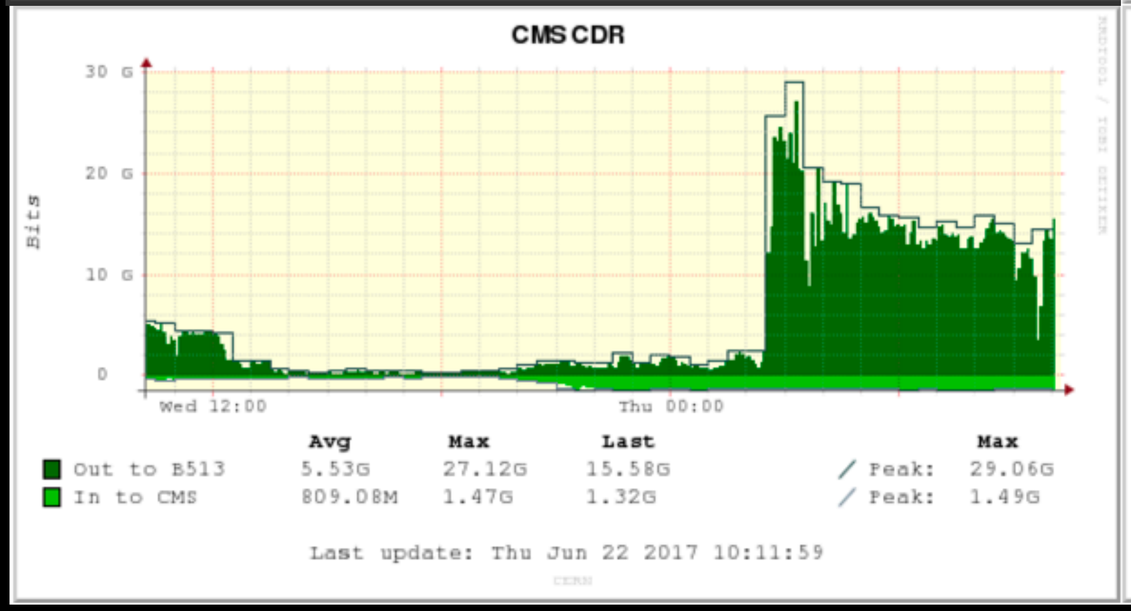


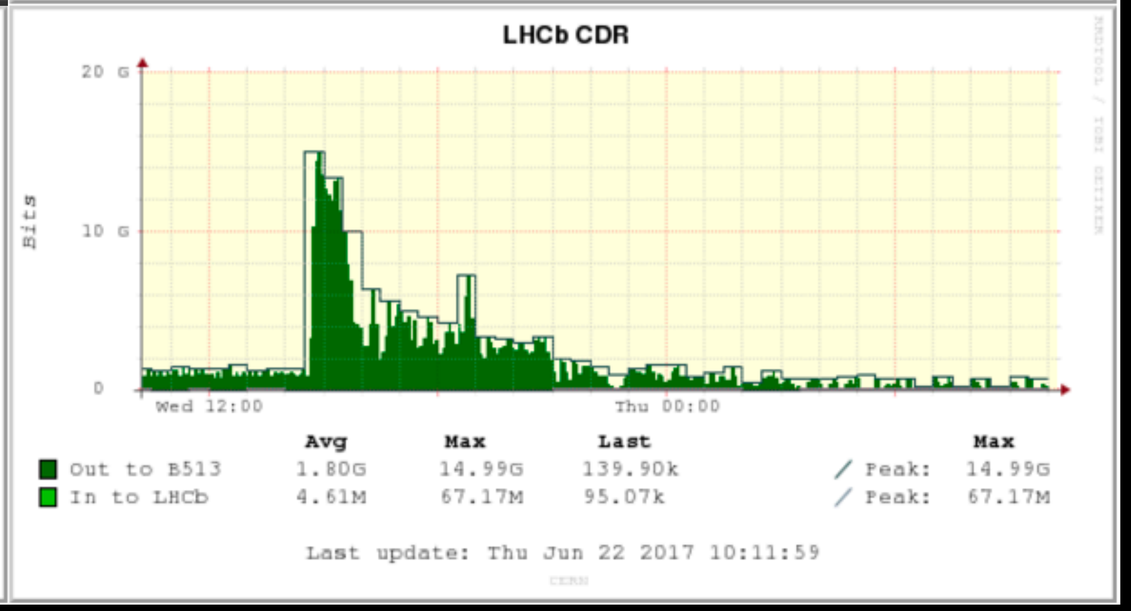


CERN-IT Storage Services: DAQ

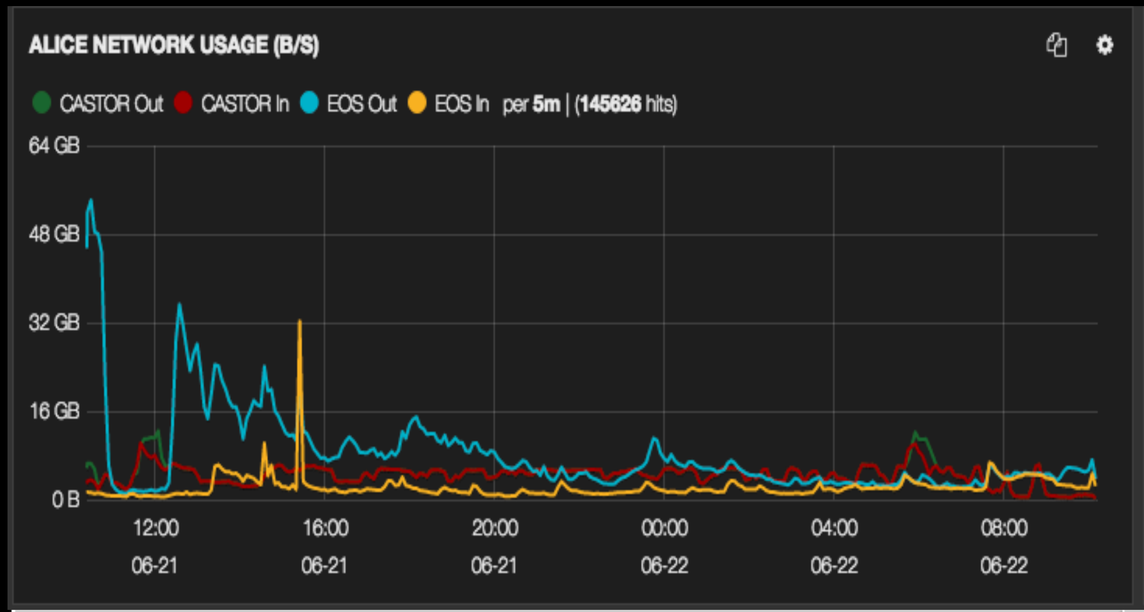


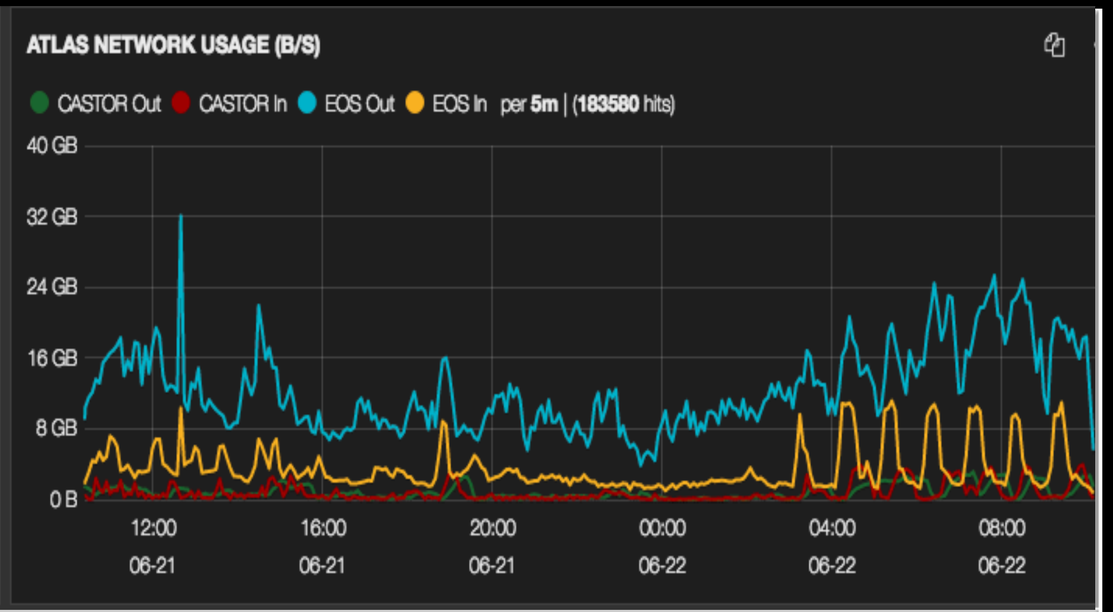


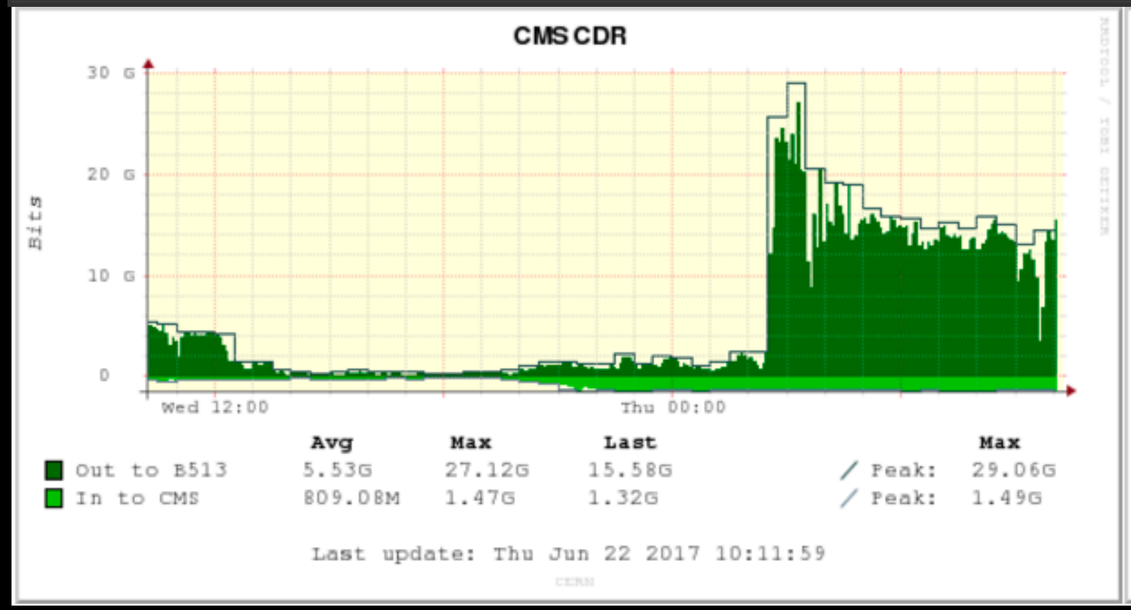


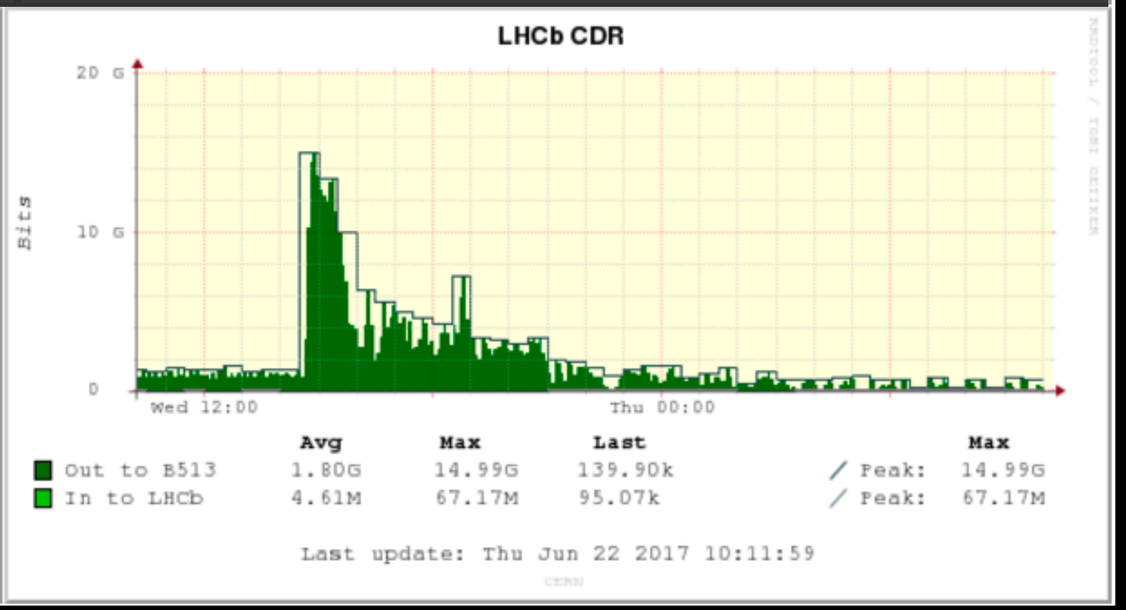


CERN-IT Storage Services: DAQ

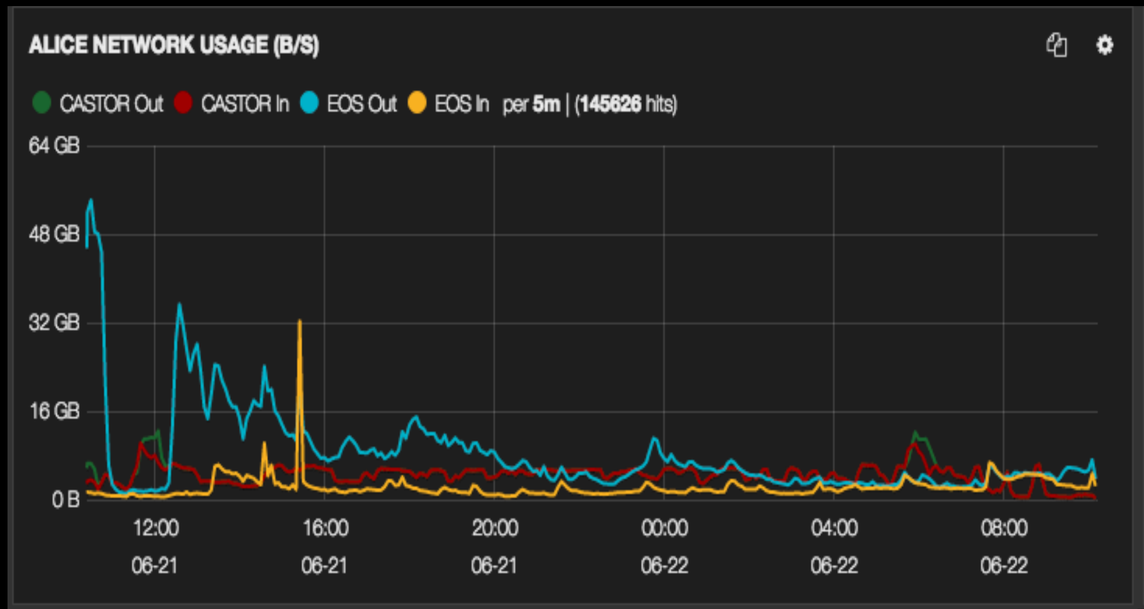


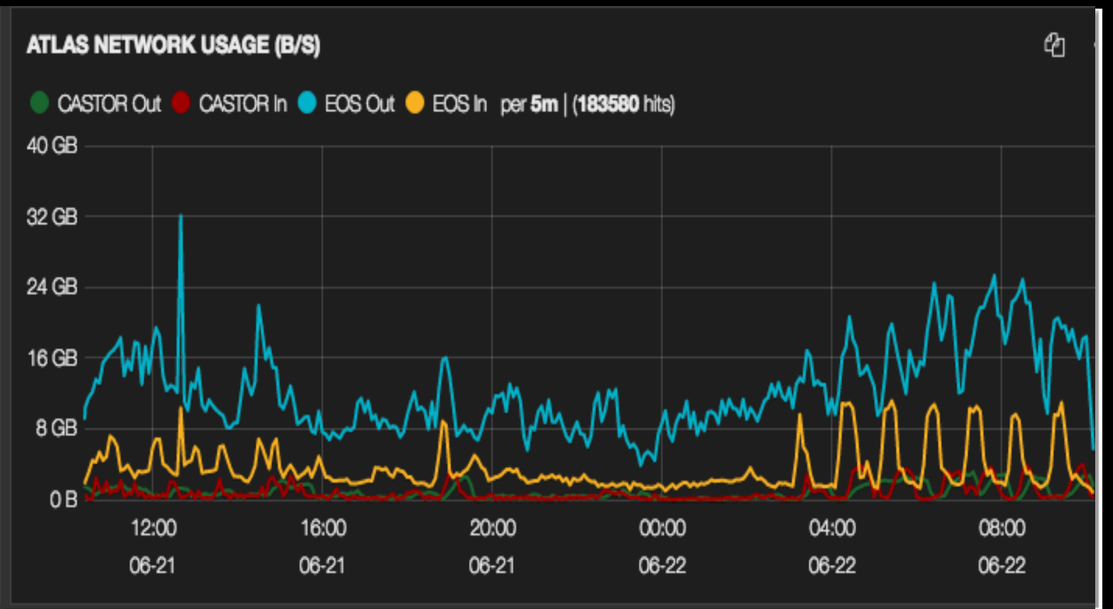


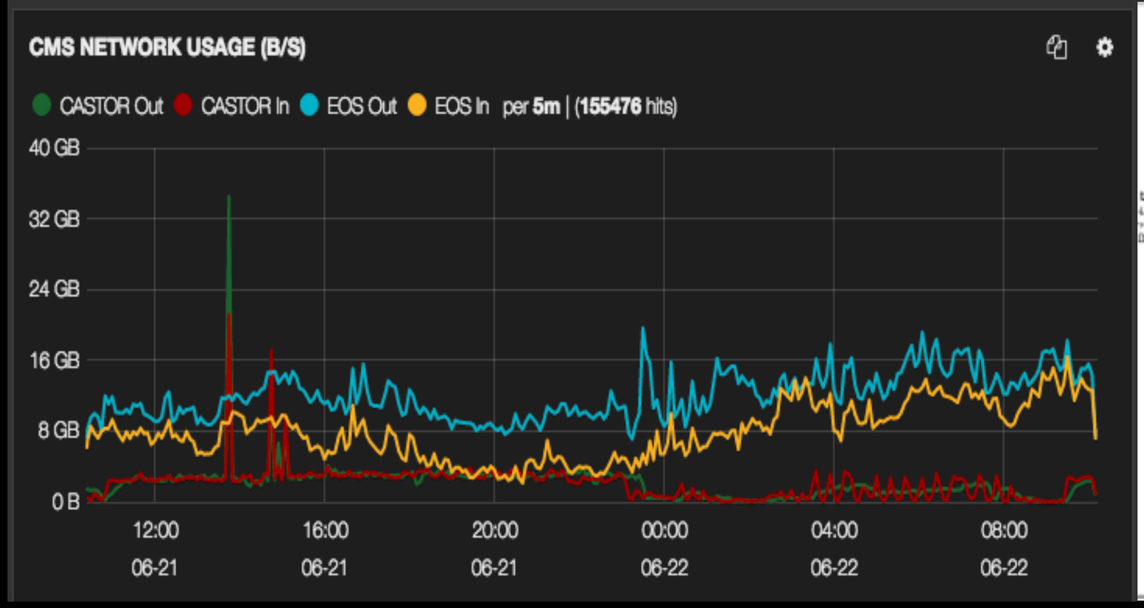


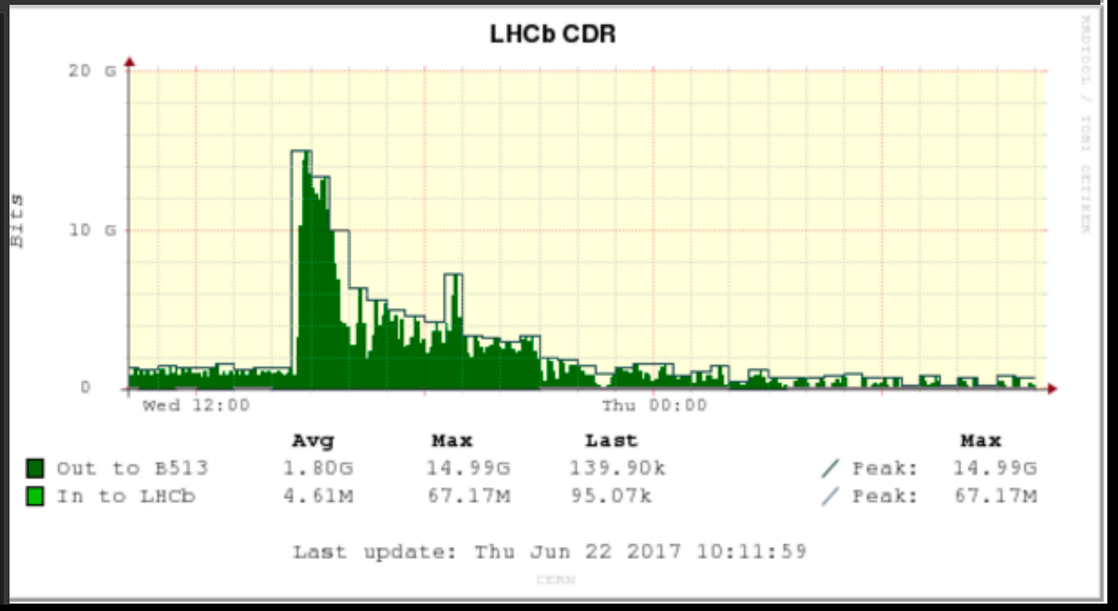


CERN-IT Storage Services: DAQ

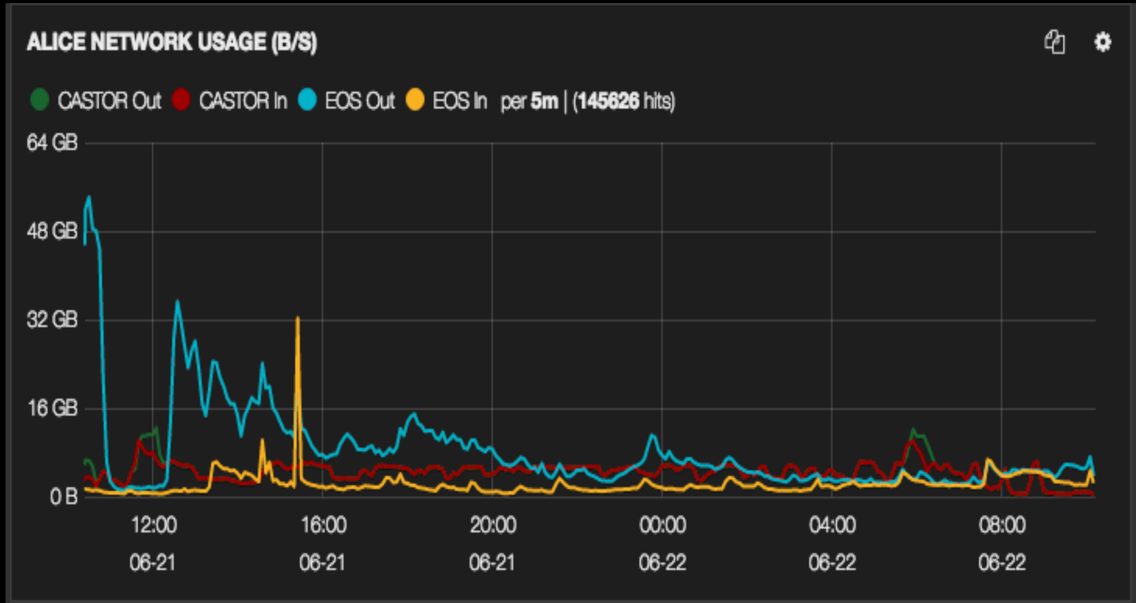


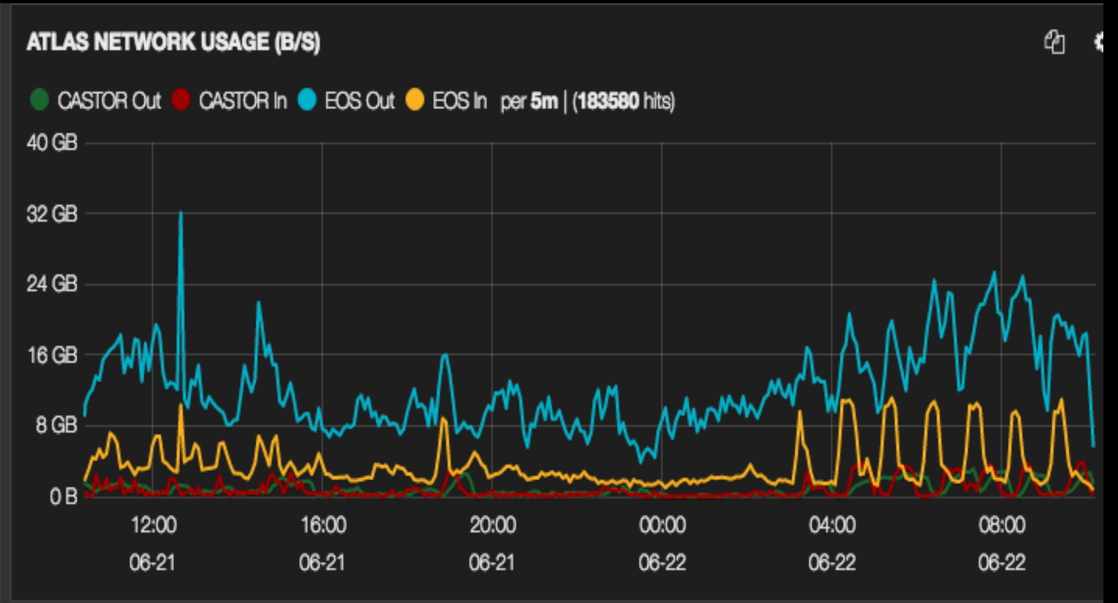


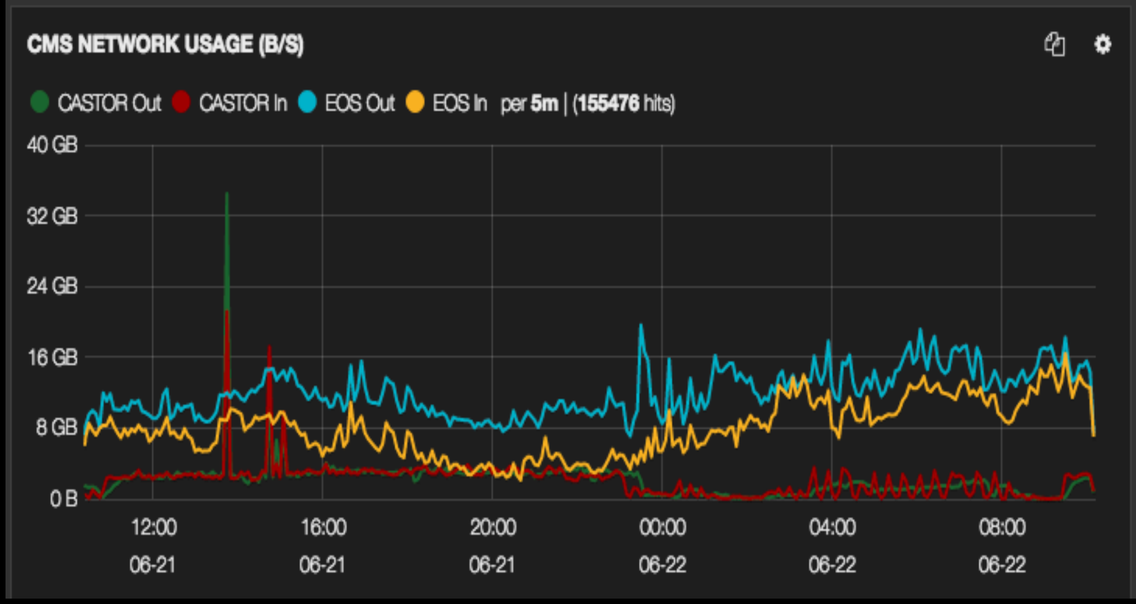




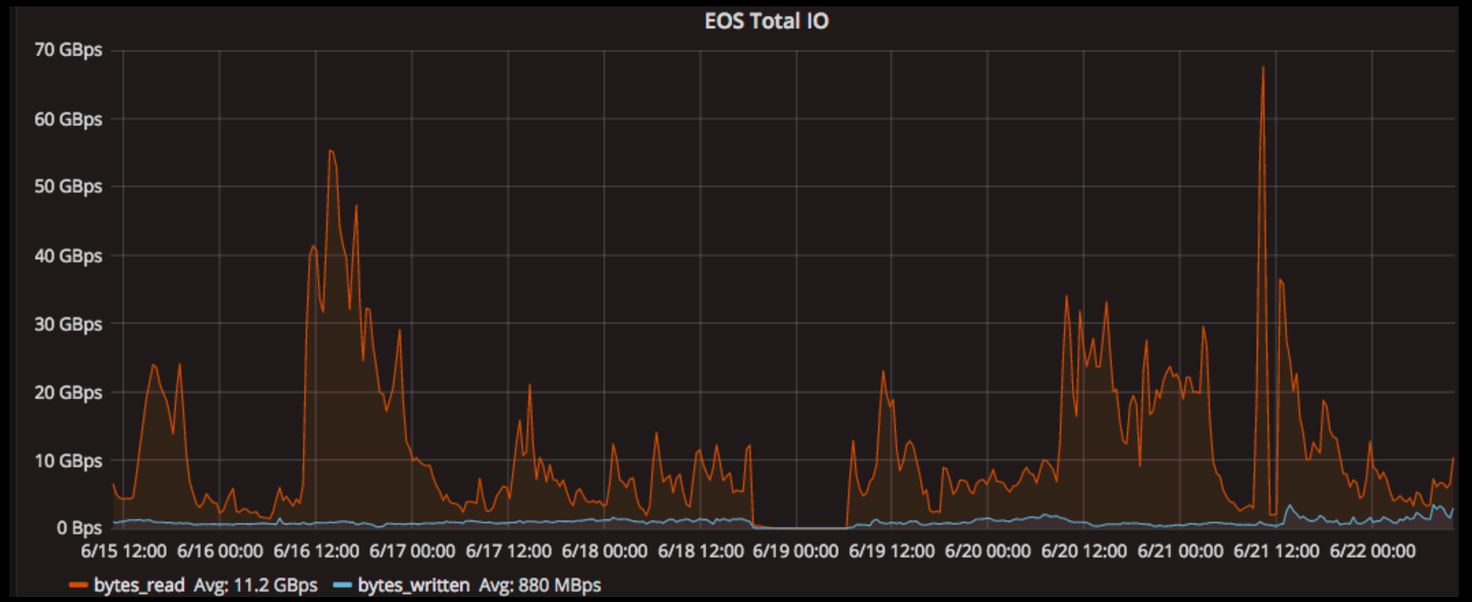
CERN-IT Storage Services: DAQ

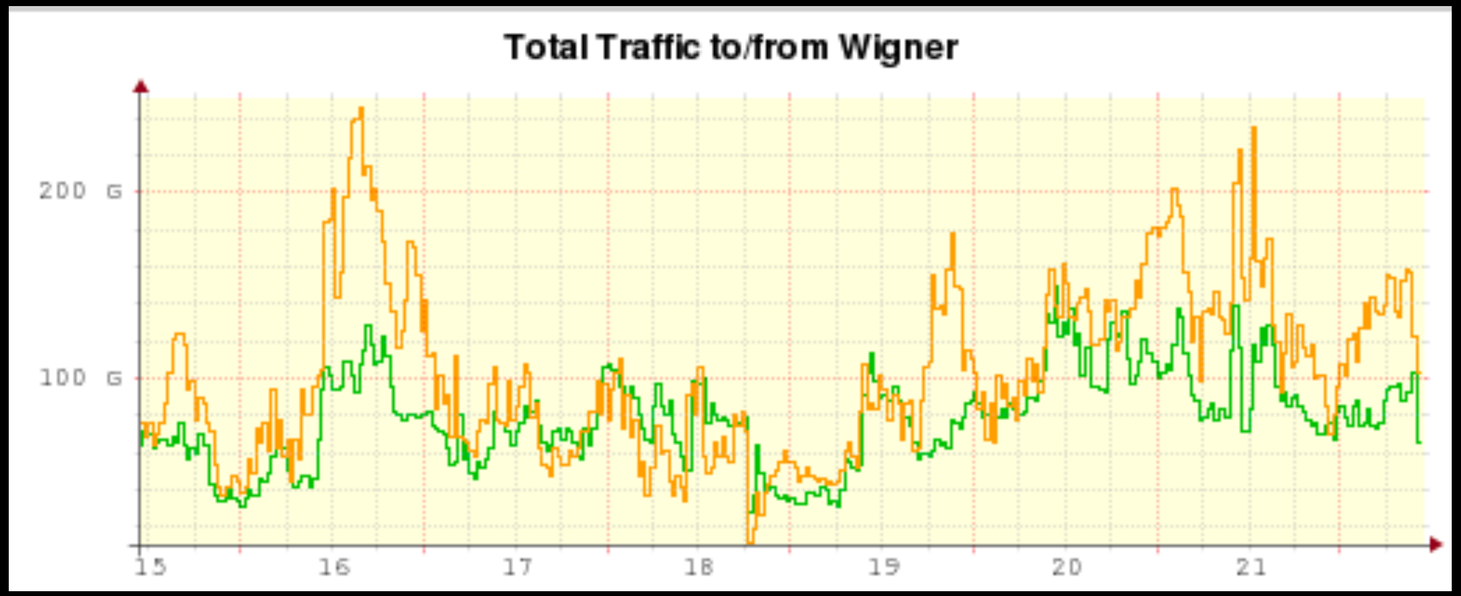


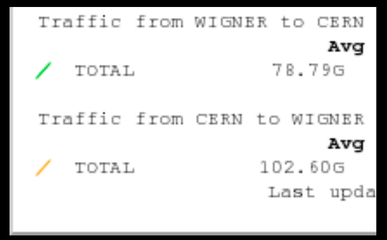




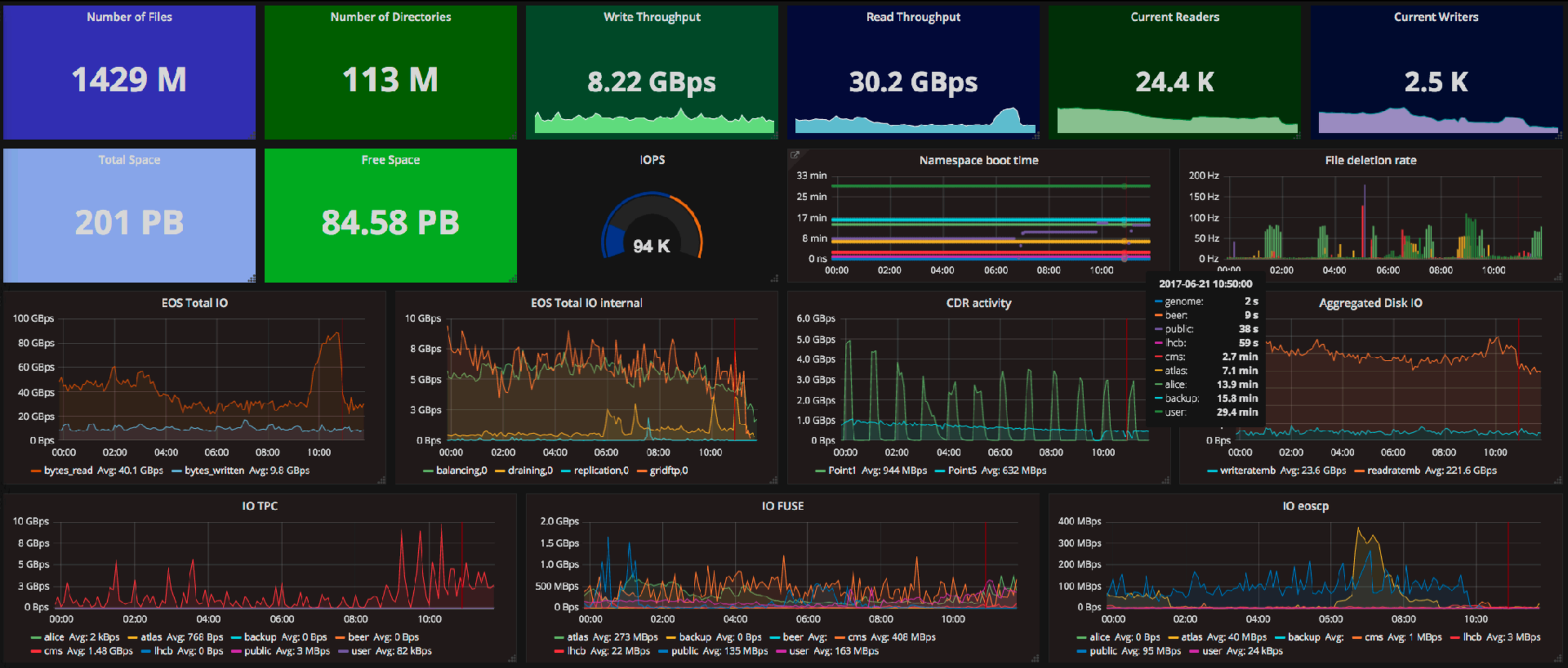
CERN-IT Storage Services: WAN







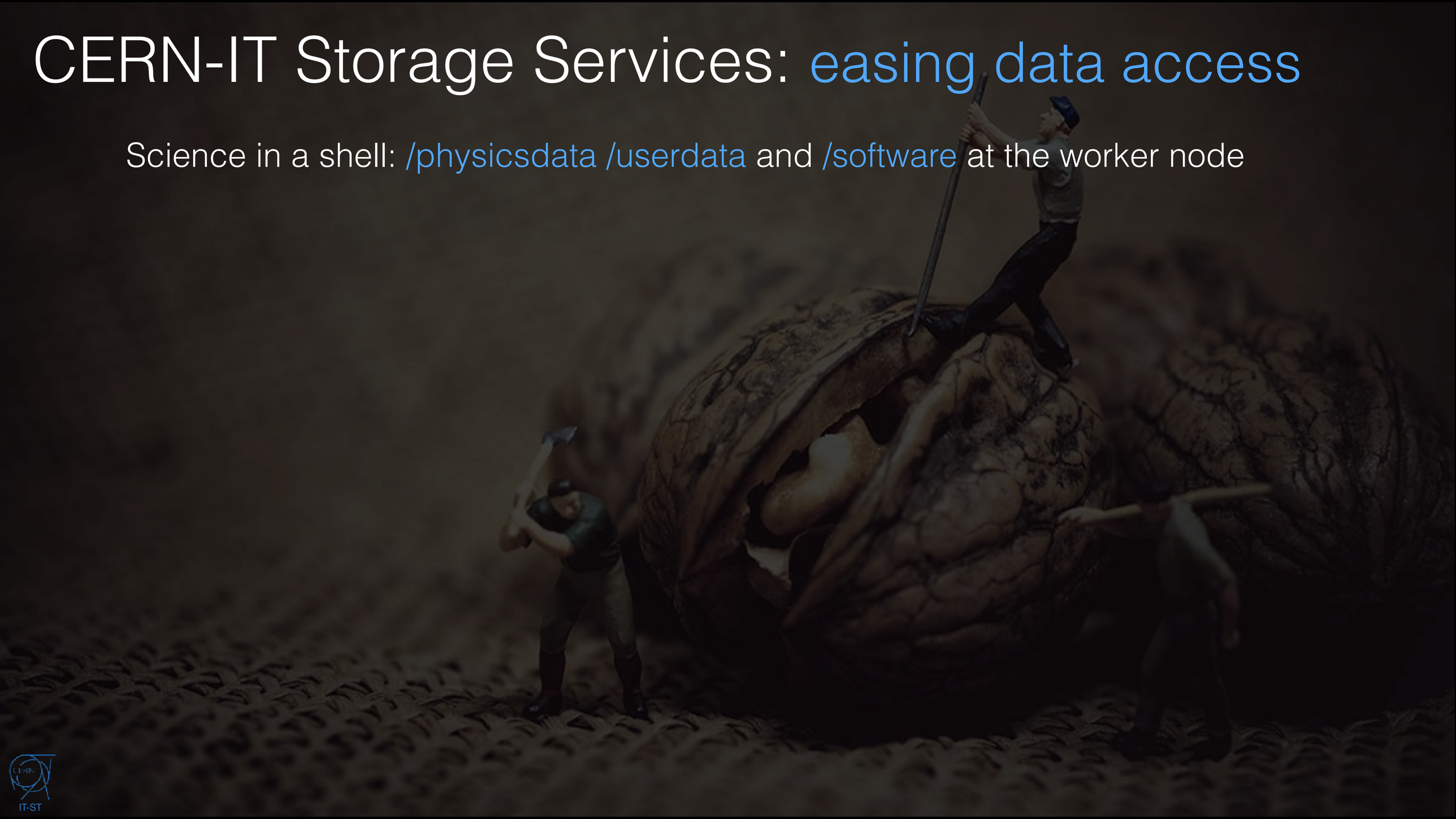
CERN-IT Storage Services: an ordinary day



IT-ST

CERN-IT Storage Services: easing data access

Science in a shell: /physicsdata /userdata and /software at the worker node



CERN-IT Storage Services: easing data access

Science in a shell: /bigdata /userdata and /software mounted on the worker node

My code ***Htozz.kumac*** is on my laptop and **synced** to **cernbox**: **/eos/user/xavi/goldench/**

CERN-IT Storage Services: easing data access

Science in a shell: /bigdata /userdata and /software mounted on the worker node

My code ***Htozz.kumac*** is on my laptop and **synced** to **cernbox**:
/eos/user/xavi/goldench/

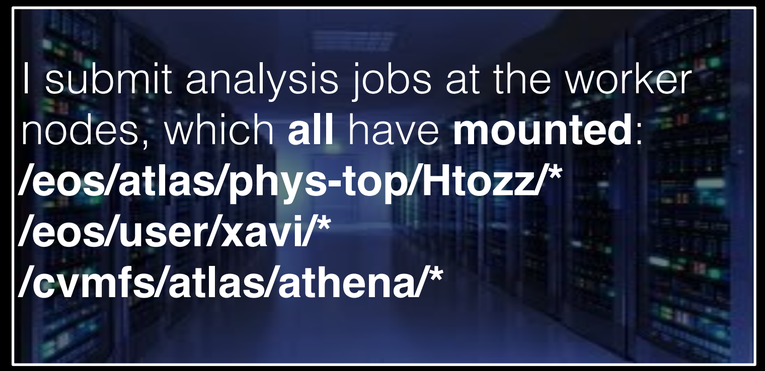
I'm interested in running my analysis on the full **HtoZZ** dataset:
/eos/atlas/phys-higgs/htozz

CERN-IT Storage Services: easing data access

Science in a shell: /bigdata /userdata and /software mounted on the worker node

My code ***Htozz.kumac*** is on my laptop and **synced** to **cernbox**:
/eos/user/xavi/goldench/

I'm interested in running my analysis on the full **HtoZZ** dataset:
/eos/atlas/phys-higgs/htozz



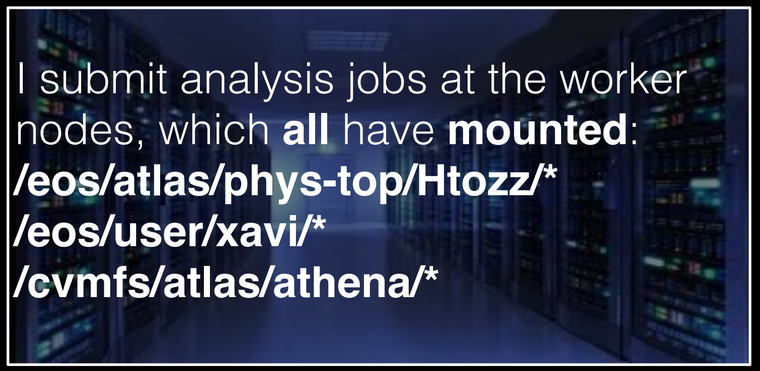
I submit analysis jobs at the worker nodes, which **all** have **mounted**:
/eos/atlas/phys-top/Htozz/*
/eos/user/xavi/*
/cvmfs/atlas/athena/*

CERN-IT Storage Services: easing data access

Science in a shell: /bigdata /userdata and /software mounted on the worker node

My code ***Htozz.kumac*** is on my laptop and **synced** to **cernbox**:
/eos/user/xavi/goldench/

I'm interested in running my analysis on the full **HtoZZ** dataset:
/eos/atlas/phys-higgs/htozz



I submit analysis jobs at the worker nodes, which **all** have **mounted**:
/eos/atlas/phys-top/Htozz/*
/eos/user/xavi/*
/cvmfs/atlas/athena/*

The job results aggregated on **cernbox**:

/eos/user/xavi/goldench/htozz/

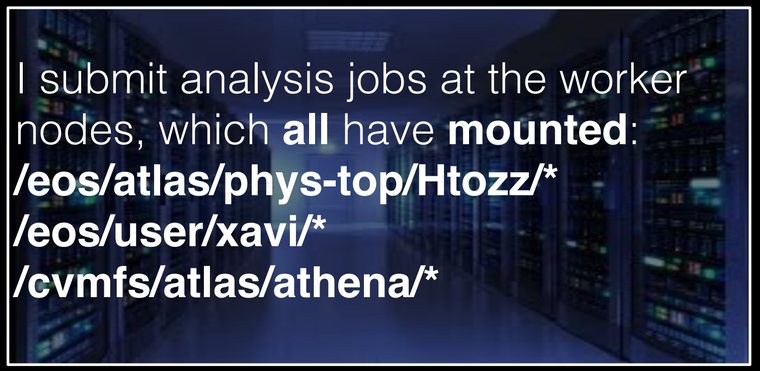
And **synced** on my laptop as the jobs finished

CERN-IT Storage Services: easing data access

Science in a shell: /bigdata /userdata and /software mounted on the worker node

My code ***Htozz.kumac*** is on my laptop and **synced** to **cernbox**:
/eos/user/xavi/goldench/

I'm interested in running my analysis on the full **HtoZZ** dataset:
/eos/atlas/phys-higgs/htozz



I submit analysis jobs at the worker nodes, which **all** have **mounted**:
/eos/atlas/phys-top/Htozz/*
/eos/user/xavi/*
/cvmfs/atlas/athena/*

The job results aggregated on **cernbox**:

/eos/user/xavi/goldench/htozz/

And **synced** on my laptop as the jobs finished

Work on **final plots** on the **laptop** and latex-ing the paper directly on

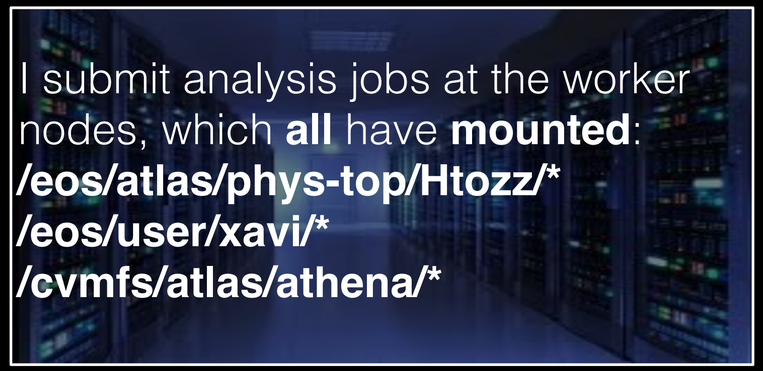
/eos/user/xavi/goldench/htozz/paper/

CERN-IT Storage Services: easing data access

Science in a shell: /bigdata /userdata and /software mounted on the worker node

My code ***Htozz.kumac*** is on my laptop and **synced** to **cernbox**:
/eos/user/xavi/goldench/

I'm interested in running my analysis on the full **HtoZZ** dataset:
/eos/atlas/phys-higgs/htozz

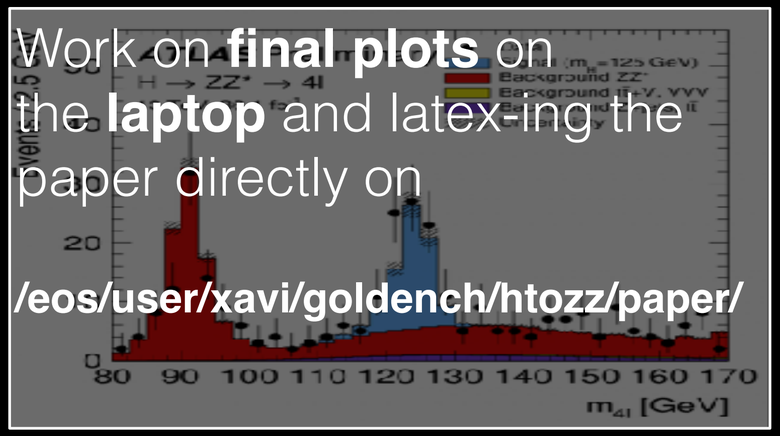


I submit analysis jobs at the worker nodes, which **all** have **mounted**:
/eos/atlas/phys-top/Htozz/*
/eos/user/xavi/*
/cvmfs/atlas/athena/*

The job results aggregated on **cernbox**:

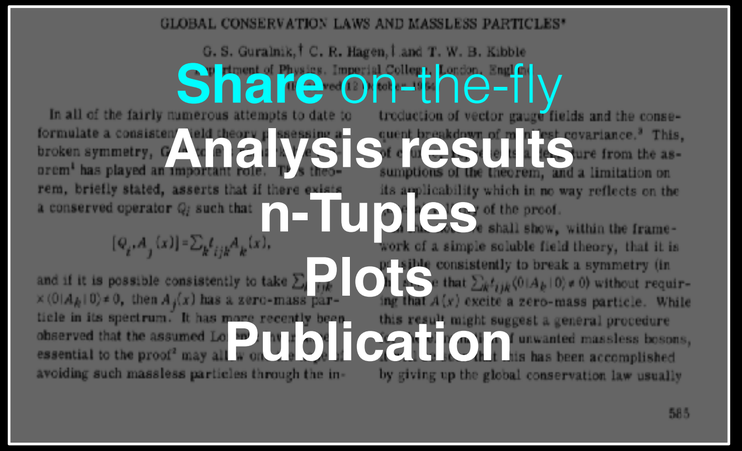
/eos/user/xavi/goldench/htozz/

And **synced** on my laptop as the jobs finished



Work on **final plots** on the **laptop** and latex-ing the paper directly on

/eos/user/xavi/goldench/htozz/paper/


GLOBAL CONSERVATION LAWS AND MASSLESS PARTICLES*

G. S. Guralnik,† C. R. Hagen,‡ and T. W. B. Kibble

In all of the fairly numerous attempts to date to formulate a consistent field theory possessing ... broken symmetry, ... orem[1] has played an important ... rem, briefly stated, asserts that if there exists a conserved operator Q_i such that

$$[Q_i, A_j(x)] = \sum_k c_{ijk} A_k(x),$$

and if it is possible consistently to take $\sum_k c_{ijk}\langle 0|A_k|0\rangle \neq 0$, then $A_j(x)$ has a zero-mass particle in its spectrum. It has more recently been observed that the assumed Lo... essential to the proof[2] may al... avoiding such massless particles through the in-

troduction of vector gauge fields and the consequent invariance.[3] This, ... ure from the assumptions of the theorem, and a limitation on its applicability which in no way reflects on the ... of the proof.

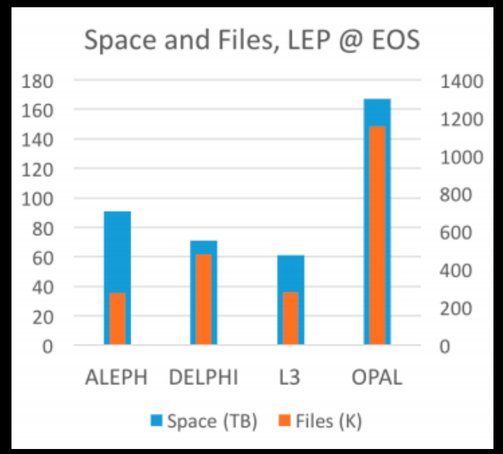
... shall show, within the framework of a simple soluble field theory, that it is ... ible consistently to break a symmetry (in ... e that $\sum_k c_{ijk}\langle 0|A_k|0\rangle \neq 0$ without requiring that $A_j(x)$ excite a zero-mass particle. While this result might suggest a general procedure ... unwanted massless bosons, ... is has been accomplished by giving up the global conservation law usually

585

Share on-the-fly
Analysis results
n-Tuples
Plots
Publication

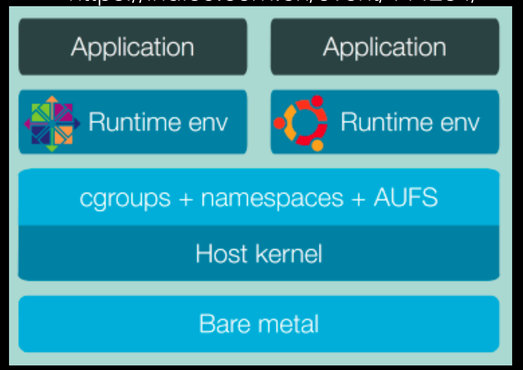
CERN-IT Storage Services: Data ages, preservation!

Keep the data
Keep the data safe (corruption)
Keep the data clean (dust)
Keep the data readable (tape and tapedrive technologies)
Keep the data usable (useful for analyses -> sw, os, compatibility)



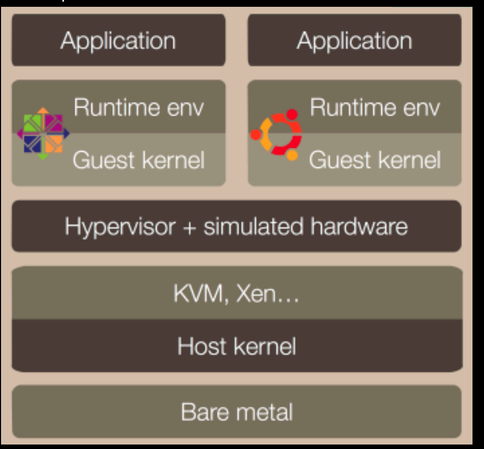
containers

https://indico.cern.ch/event/444264/



VMs

https://indico.cern.ch/event/444264/



Storage Systems: scenarios

Hot storage: Hybrid HDD and SSD tiered storage? SSD ideal for caching on predictive patterns (but this is not our case). On the other hand, indications that 70% of our data is WORN…so?

Cold storage: long term archival. Easy to write, hard to read. What will replace magnetic tapes in 10yr time? 1 PB of SSD in 2U! Power-wake-on-access?

Fractal storage: future of shared file systems and home directories.
(warning: self coined buzz word)

Storage technology: disk, tapes and solid state(s)

HDD old technology. Still evolving but market **shrinking** as SDD is taking over as the solution for commodity hardware. Uncertainty on long term evolution, pricing… HDD #units production declining: -10%(2016), -7%(2017 expected)

https://www.forbes.com/sites/tomcoughlin/2017/01/28/20-tb-hard-disk-drives-the-future-of-hdds/#7f60c5381f88

Tape market under **shockwave** after one of the market leaders announcement. Market soon owned by single manufacturer.

Lot of **gossips** about fat SSDs on new technologies, but $$$ and little data about stability/duration.

Last diskservers at CERN: 2x24x8TB, 10Gbps,12Gbps interlinks, 2xSSD (OS)

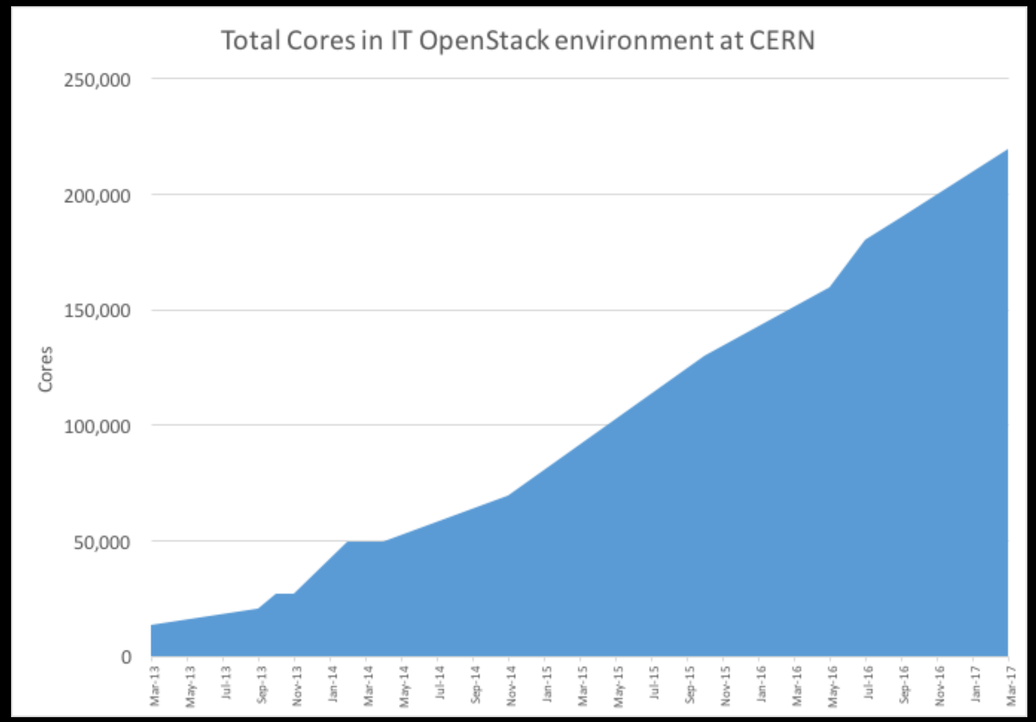
Computing Services and Cloud Infrastructure

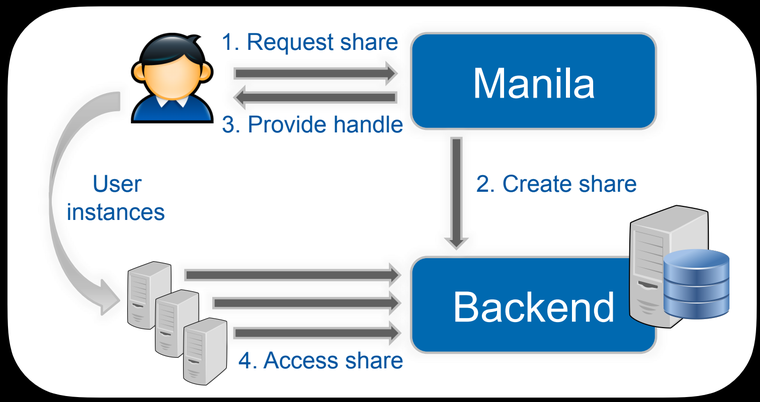
Present: full virtualization of computing servers

~9000 hypervisors in production
~220K cores
~4K volumes with 1.2 PB allocated (Cinder)
~4K images/snapshots (Glance)
27 fileshares with 18 TB allocated (Manila)
71 container clusters (Magnum) (new)

Future:

Steady **growth** expected, soon 300k core
Nova to **Neutron** transition
Cells-V1 to **Cells-V2** (tennant pooling 'enforced' soon)
New services for users:
Manila - Provisioning of **Shared File Systems to VMs**
Ironic - **Baremetal** Service
Magnum - **Containers** as a Service
Mistral - **Workflows** Service
SDNs:
Openstack SDN*'aware'*-neutron: openvswitch (L2/L3), opendaylight
Floating IPs -> live migrations across IP services





IT-ST

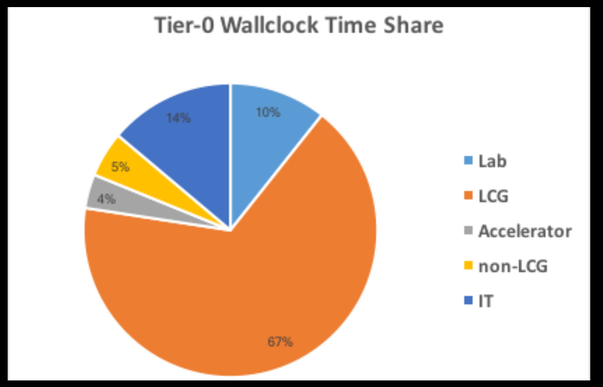
Computing Services and Cloud Infrastructure

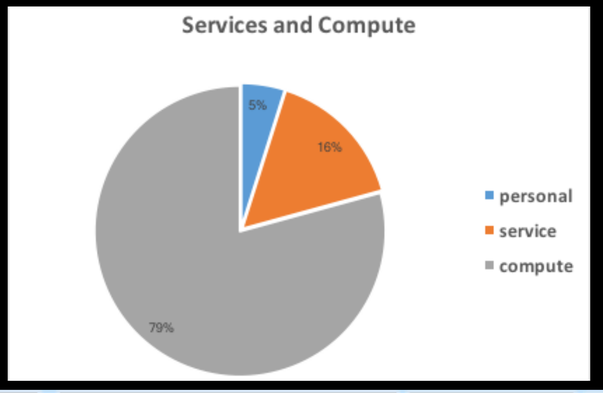
Present:

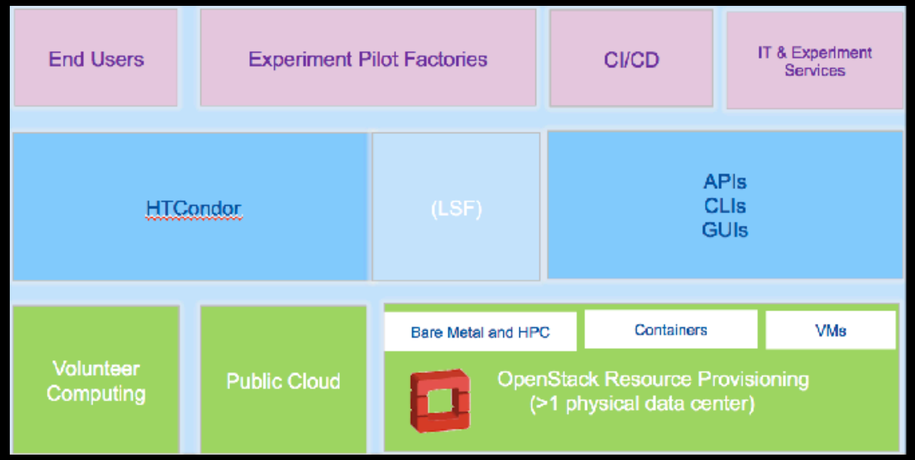
- 50% LSF, 50% Condor
- ~130k cores for batch (200k end of 2017)
- ~650k jobs/day
- Small high-memory (~1TB) facility to be provided this year for special cases.
- Big data local access via FUSE: /eos and experiment software: /cvmfs
- Vast majority deployed as long-lived VMs on Openstack using HTCondor vanilla universe

HPC:

- MPI, shared memory across nodes, infiniband
- Lattice QCD Theory simulations, Beam / plasma, fluid dynamics applications (fire safety, cryo), engineering simulations (civil and electronic)
- Theory cluster, Beams cluster
- SLURM batch system being deployed for this (~5k cores). Backfill via HTCondor / SLURM interface







Computing Services and Cloud Infrastructure

Future:

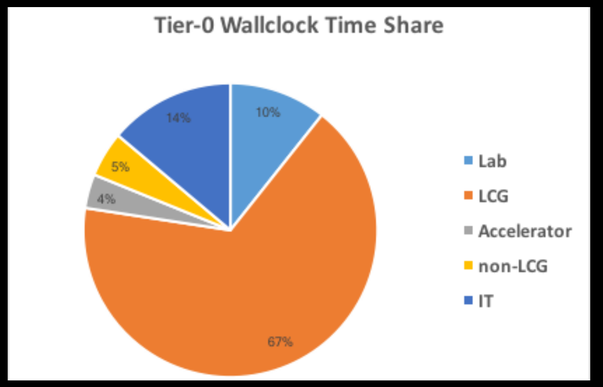
LSF HTCondor

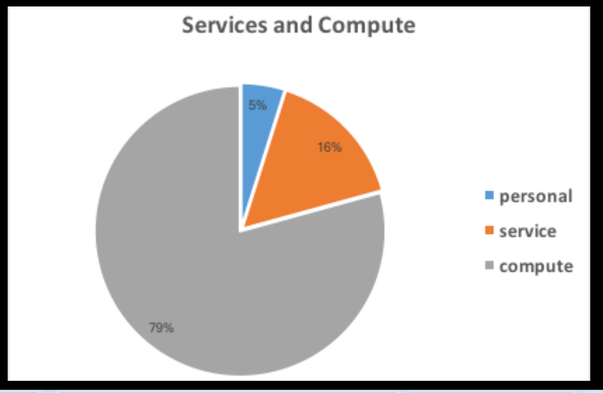
Containers

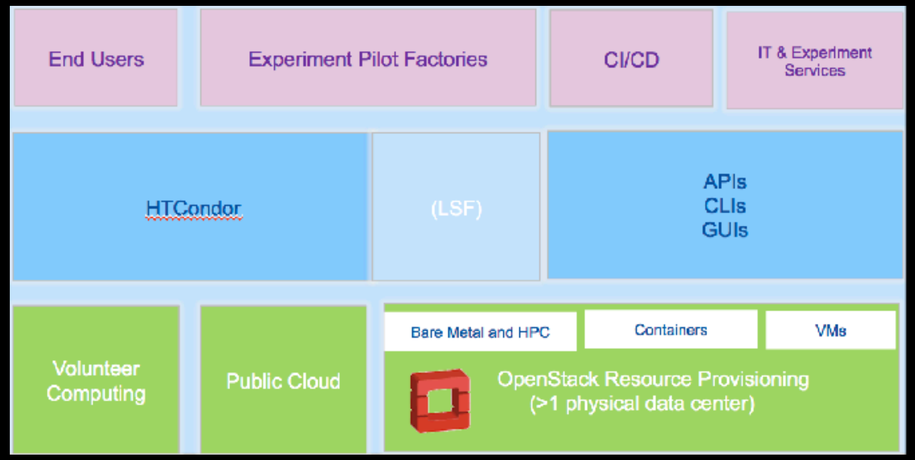
- (pilot isolation) Containers: deploy **singularity** for experiments
- (job isolation) HTCondor Docker universe for job isolation, CVMFS / EOS mounts, no AFS

Making better user

- Making use of disk-server CPUs
- Spare service “headroom” on cloud, choppy cloud compute capacity, external cloud spot
- HPC backfill, pre-empt by prompt work (Tier-0/ CAF)

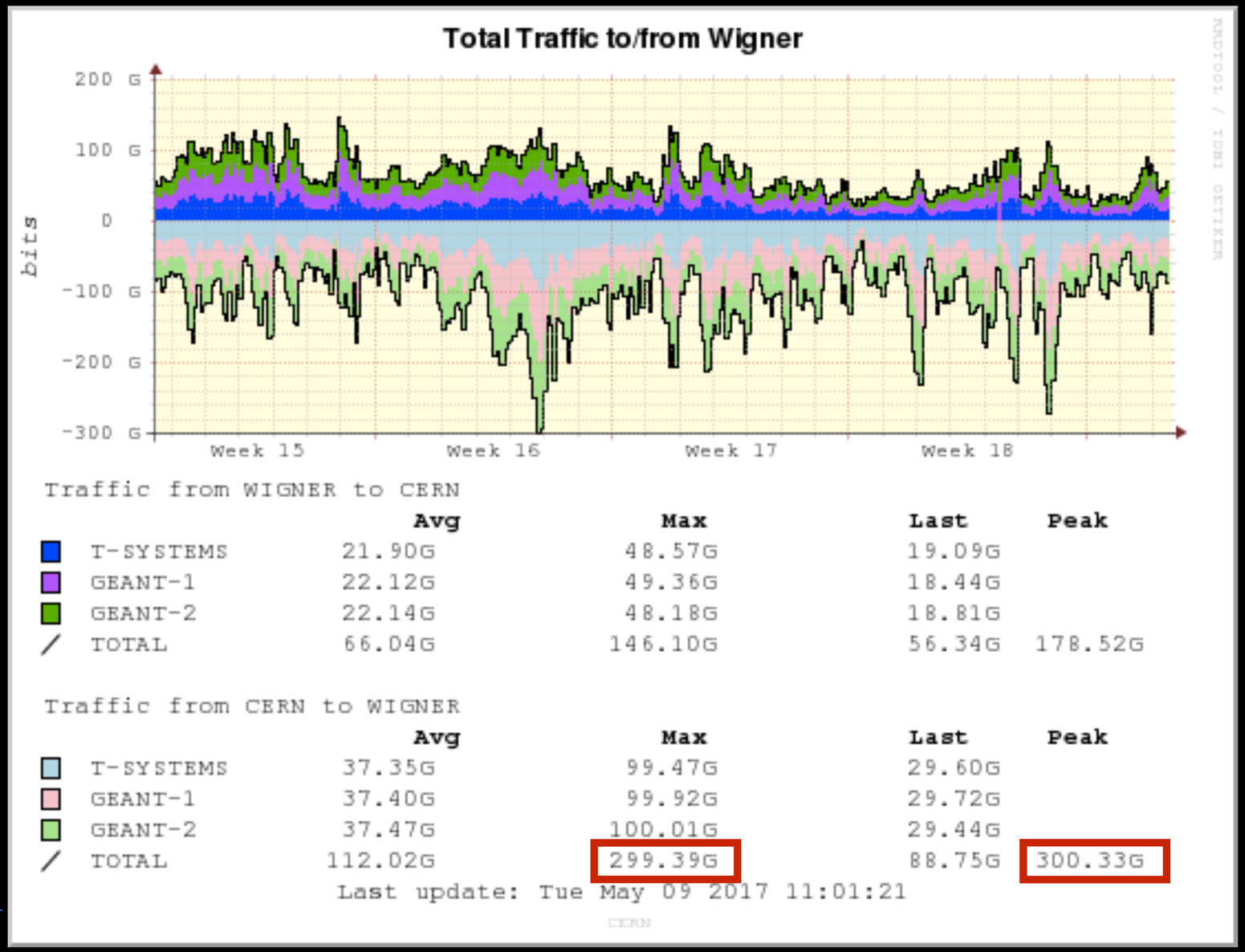






Network

CERN-WIGNER: 3x100Gbps links



Traffic from WIGNER to CERN

	Avg	Max	Last	Peak
T-SYSTEMS	21.90G	48.57G	19.09G	
GEANT-1	22.12G	49.36G	18.44G	
GEANT-2	22.14G	48.18G	18.81G	
TOTAL	66.04G	146.10G	56.34G	178.52G

Traffic from CERN to WIGNER

	Avg	Max	Last	Peak
T-SYSTEMS	37.35G	99.47G	29.60G	
GEANT-1	37.40G	99.92G	29.72G	
GEANT-2	37.47G	100.01G	29.44G	
TOTAL	112.02G	299.39G	88.75G	300.33G

Last update: Tue May 09 2017 11:01:21

CERN

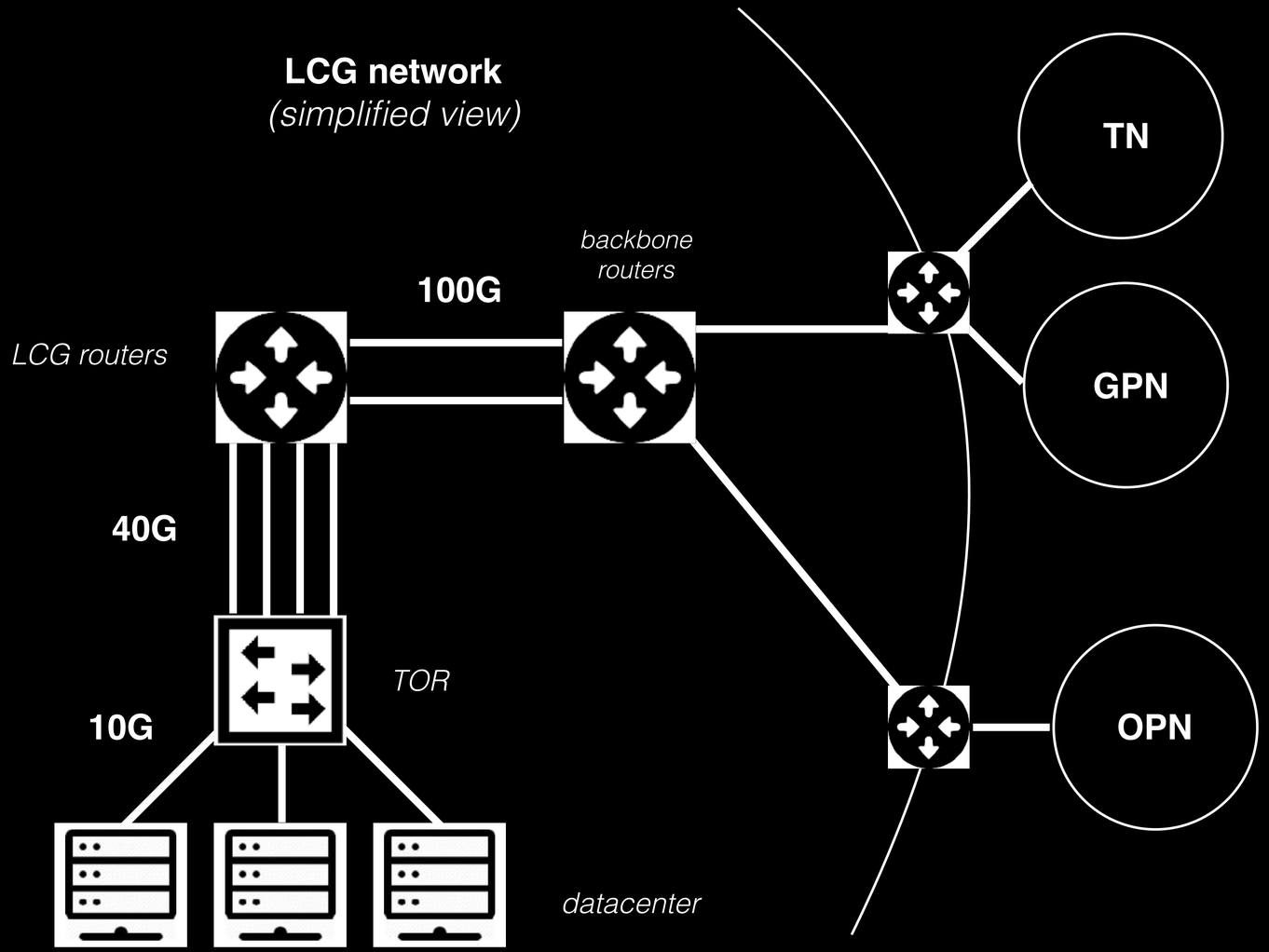
Datacenter in numbers

16315 devices
1331 Switches
39 Routers
7 Star points
29953 IPv4 addresses

CERN-wide in numbers

309902 devices
3832 Switches
233 Routers
667 Star points
2021 wifi access points

Network



Present:

- 10GE for diskservers and Hypervisors
- TOR uplink: 4x40Gbps (BF 1:2 / 1:3)
- TOR switch: 20 (ports) x32 (slots) for 10G or 4x32 40G
- ‘SDN’ since years: landb dynamic config
- IPv6 ready (full dual stack) since 2010

Future:

- High-lumi preparation (2018) -> 2xLAN bandwidth
- Deployment of new routers
- Run-III (2021):
  - 40GE default
  - 400Gbps uplinks to the backbone routers
- Ethernet still the standard for the years to come
- Mitigation automation (detection+solving)

Thanks for your attention!