INFN CNAF

Il Tier-1

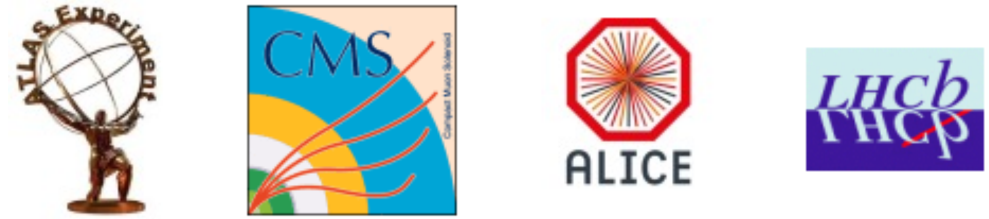
(*chi siamo, dove siamo, dove andremo….)*

Luca dell’Agnello

May 27 2015

The INFN Tier-1

- CNAF is officially supporting ~30 experiments
  - 4 LHC



  - 27 non-LHC

AMS-02
PIERRE AUGER OBSERVATORY
Belle II
ICARUS
JUNO
BOREXINO
CDF
cta
cherenkov telescope array
CUORE
LHCf
MAGIC
DARKSIDE
Fermi
Gamma-ray Space Telescope
GERDA
OPERA
panda
VIRGO
Xe
XENON
Dark Matter Project
KM3NeT

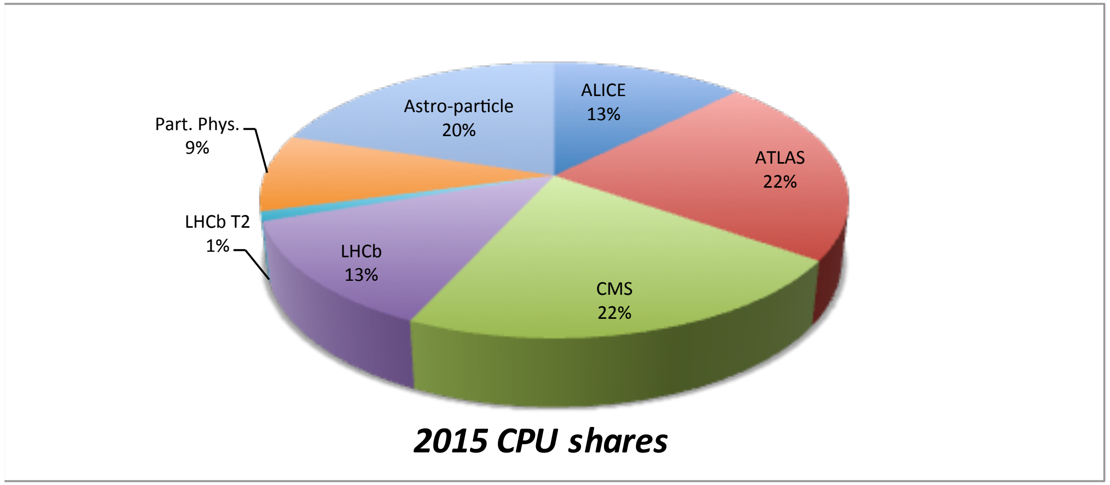
- Ten Virtual Organizations in opportunistic usage via Grid services
- LSPE and EUCLID in a near future?

Organization

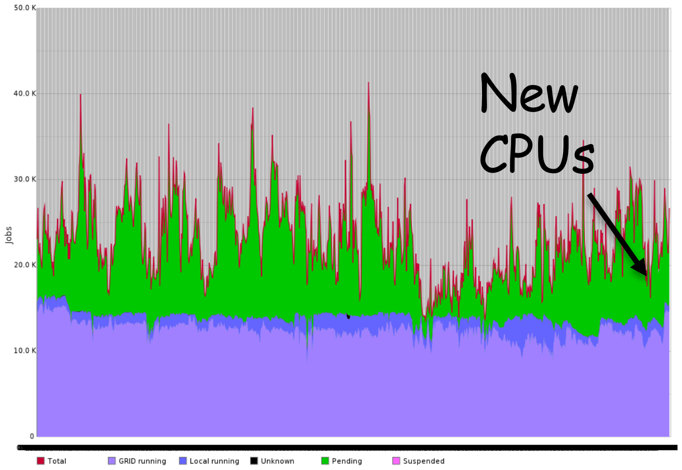
- The Tier-1 staff is composed by 23 people structured in 5 groups:
  - Farming unit (farm, CEs, UIs…)
  - Data Management unit (storage, srm and dbs)
  - Network (CNAF LAN and WAN connections)
  - Facility management group
  - User support (interface to experiments)
- Our main challenge is to guarantee H24 support
  - To avoid H24 manpower requirements, all services are completely redundant and based on enterprise hw

Services and resources: HTC

- WLCG Tier-1 standard services offered to all users/scientific collaborations
  - CPU resources assigned according to fair share mechanism
  - Non grid access supported
  - Cloud access under evaluation/test
- 1 general purpose farm
  - Currently ~ 180 KHS06 (~17K job slots)
  - ~100K jobs/day
  - Whole farm rebooted twice in the past 12 months (2 critical upgrades)
  - Dynamic partitioning supported (Atlas and CMS)



2015 CPU shares

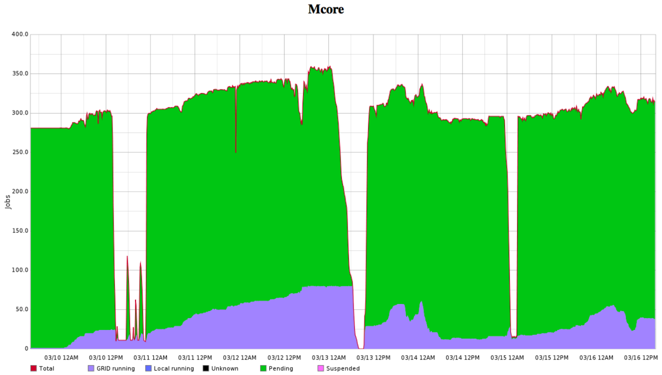


New CPUs

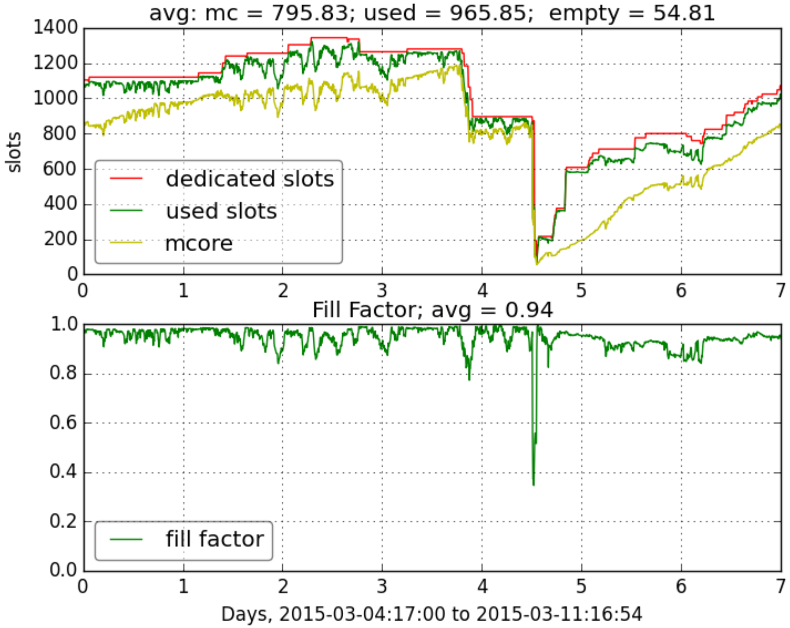
May 2014 – May 2015

Dynamic partitioning: Multicore support

- Dynamic partitioning on LSF farm
  - Allows to dynamically move WNs to a dedicated multicore queue
  - In production since last August
  - Minimization of waste of resources due to the draining phase
- CMS is now 100% multicore
- Same mechanism can be used to allocate WNs to a cloud controller (test phase)



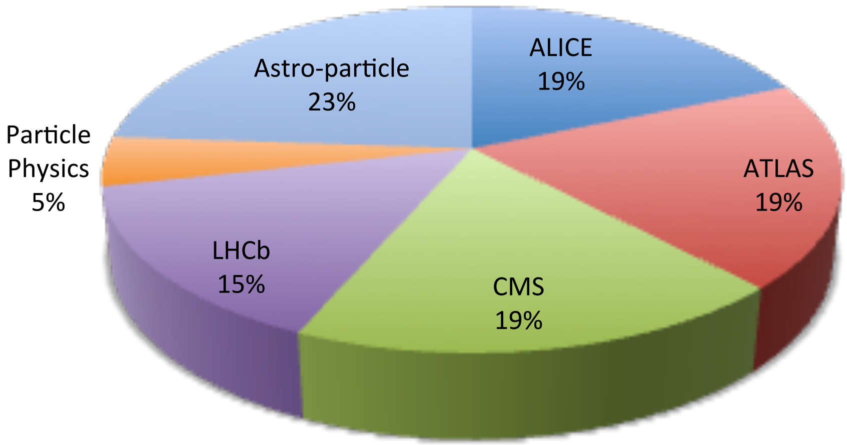
Multicore queue



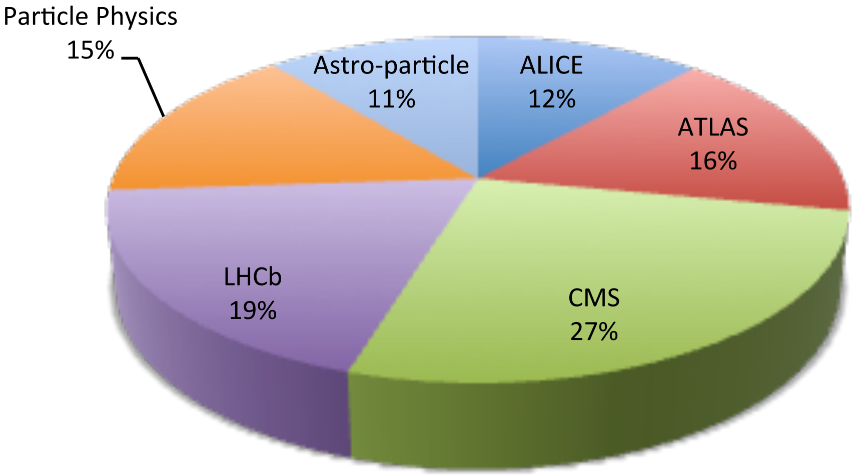
Multicore queue efficiency

Services and resources: storage

- Standard HSM service for all experiments
  - GEMSS (Grid Enabled Mass Storage System)
  - Both local and grid access
  - Standard protocol set (file, GridFTP, XrootD, http/webdav)
- Currently ~17.4 net PB of disk and ~21 PB of tapes
  - 1 tape library with 10000 slots (currently up to 85 PB capacity)
  - 5+3 PB of disk to be installed in Q3 2015
- Oracle Database services
  - Atlas calibration database and (near future) CDF databases for LTDP
  - Lemon. Grid-console. VOMS devel. (FTS)

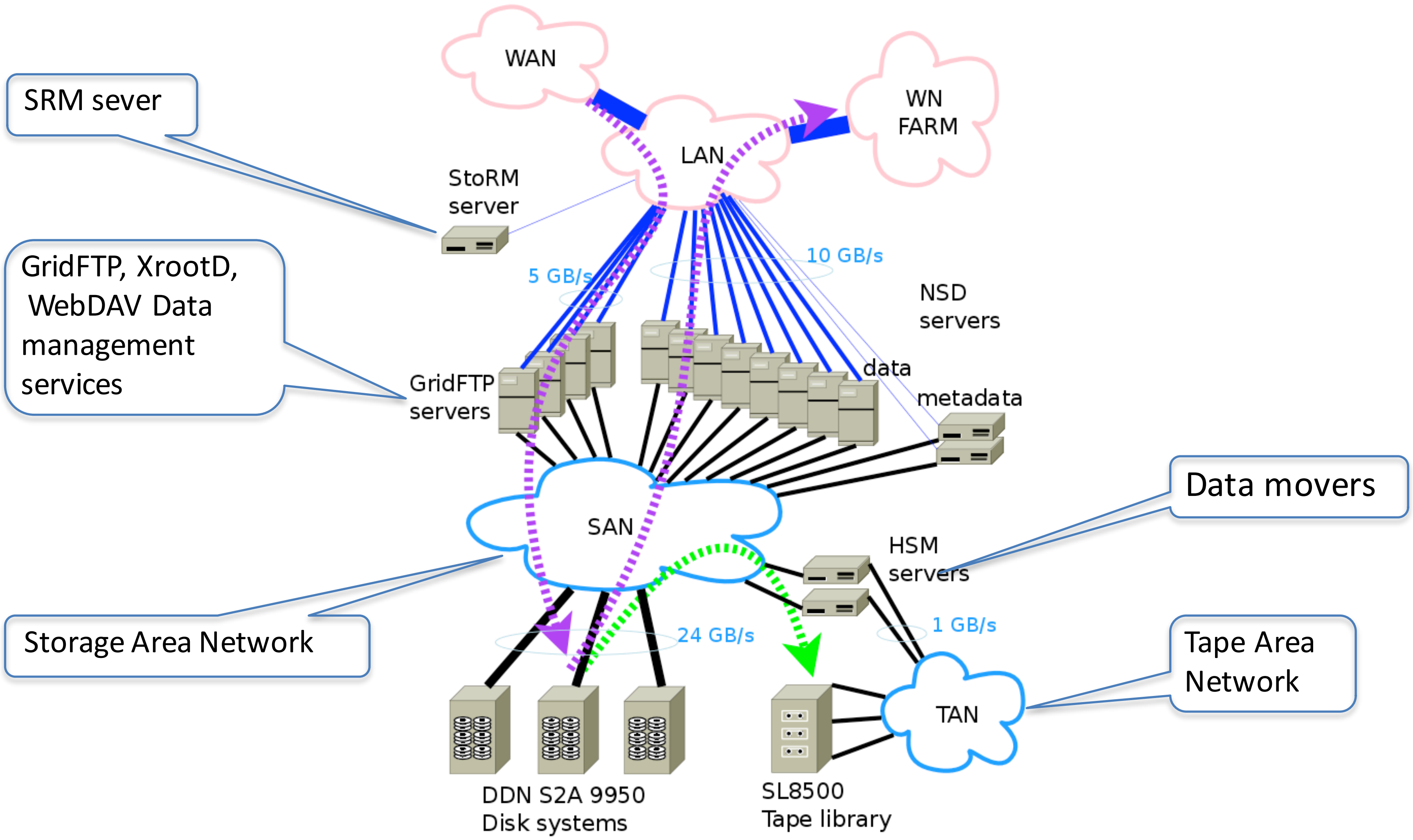


2015 Disk pledges



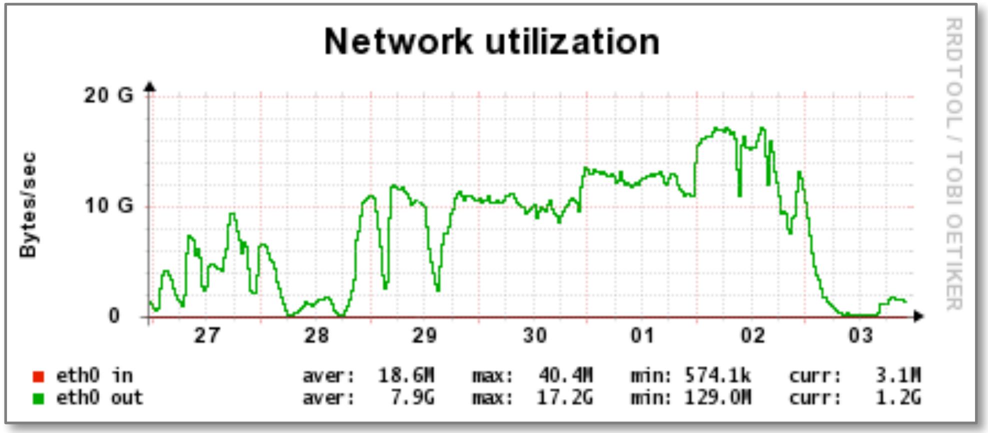
2015 tape pledges

Data flow in a single experiment cluster



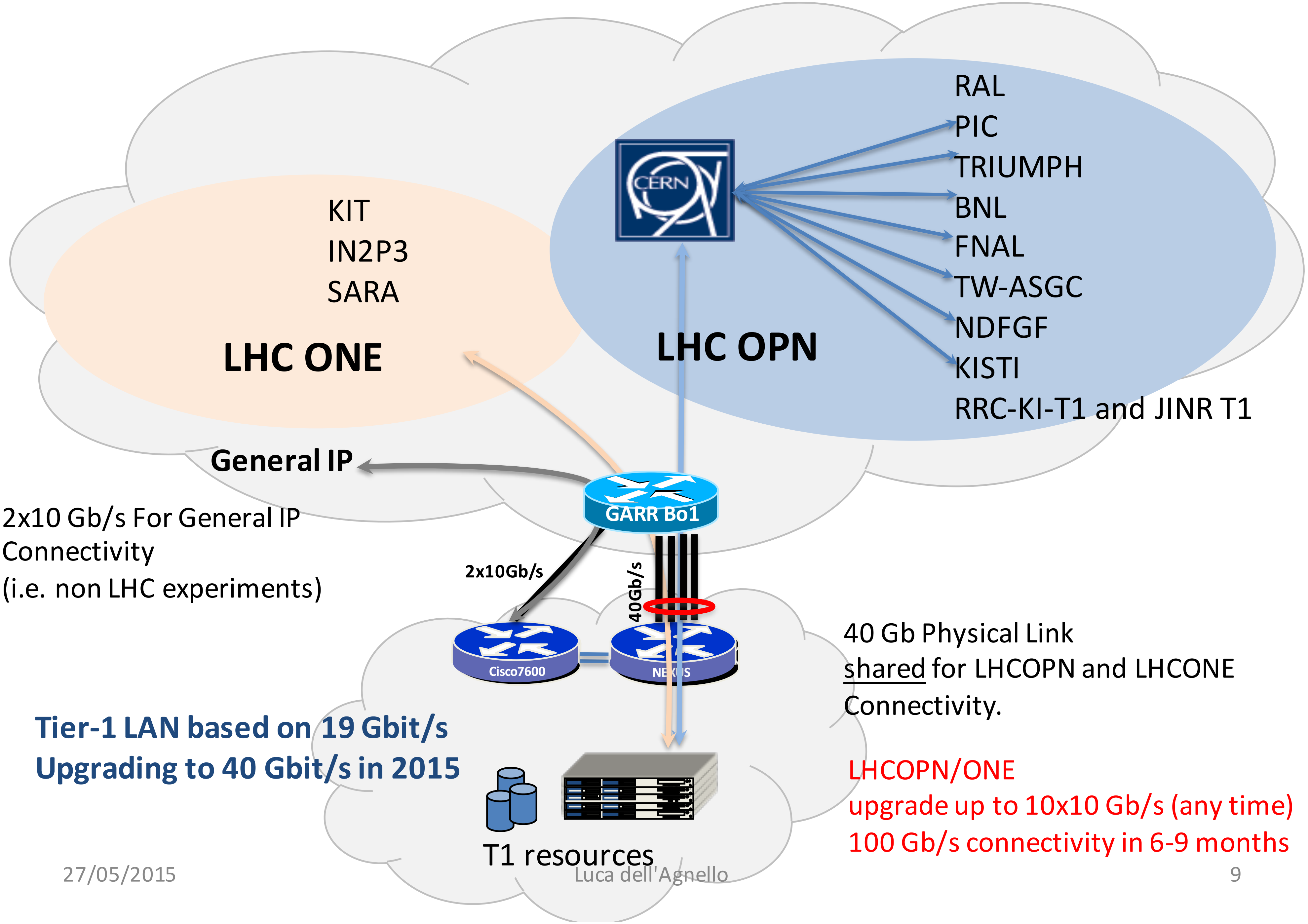
Storage model

- GPFS as POSIX interface and back-end for all data management services
  - Flexibility in management
  - Failure resilience
  - Performance
- Storage accessed from 17k concurrent processes
  - With frequent configuration changes (new installations, data migrations) too
  - 5 MB/s/TB-N guaranteed
  - 5 MB/s bandwidth for each job
- Aggregated data bandwidth to storage ~ 80 GB/s
  - Peaks of ~ 20 GB/s on LAN
  - Some occasional saturation of LHCONE links (~ 4 GB/s)



Access to CMS file-system

Connectivity



2x10 Gb/s For General IP Connectivity
(i.e. non LHC experiments)

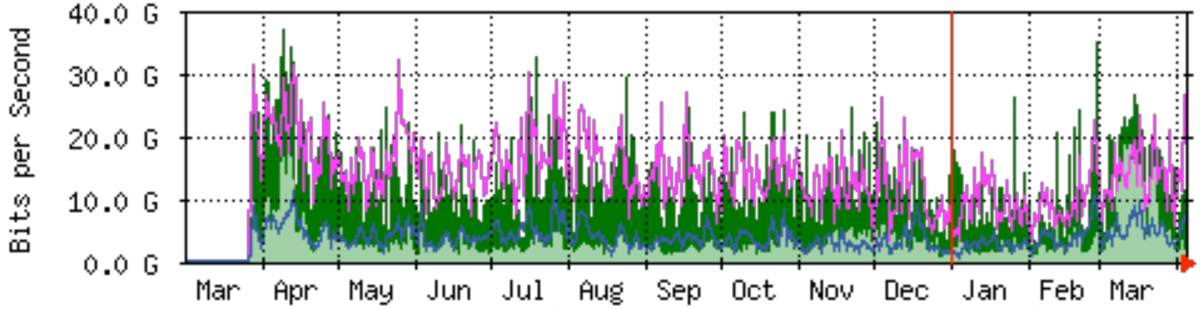
40 Gb Physical Link shared for LHCOPN and LHCONE Connectivity.

Tier-1 LAN based on 19 Gbit/s
Upgrading to 40 Gbit/s in 2015

LHCOPN/ONE
upgrade up to 10x10 Gb/s (any time)
100 Gb/s connectivity in 6-9 months

WAN utilization (last 12 months)

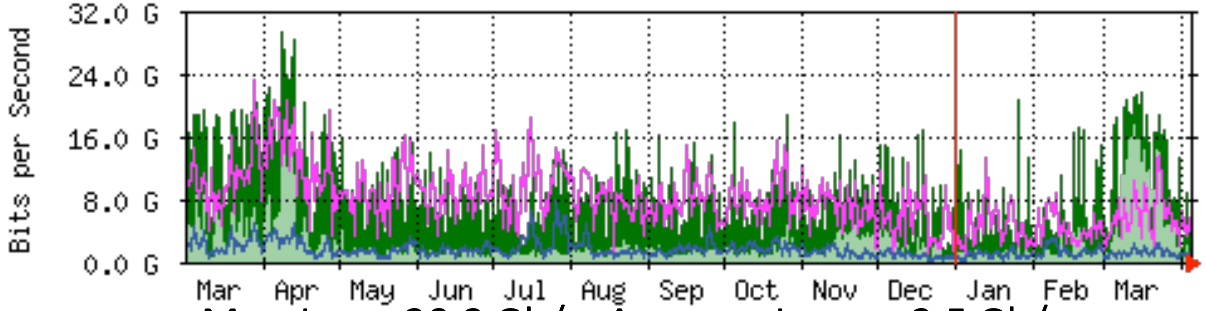
LHC OPN + ONE



Max In: 37.0 Gb/s Average In: 3.5 Gb/s

Max Out: 32.0 Gb/s Average Out: 3.8 Gb/s

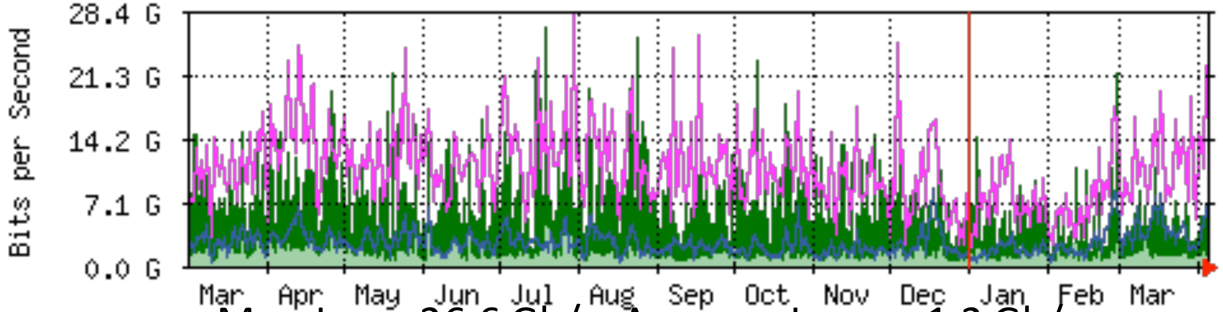
LHCOPN



Max In: 29.2 Gb/s Average In: 2.5 Gb/s

Max Out: 22.8 Gb/s Average Out: 1.4 Gb/s

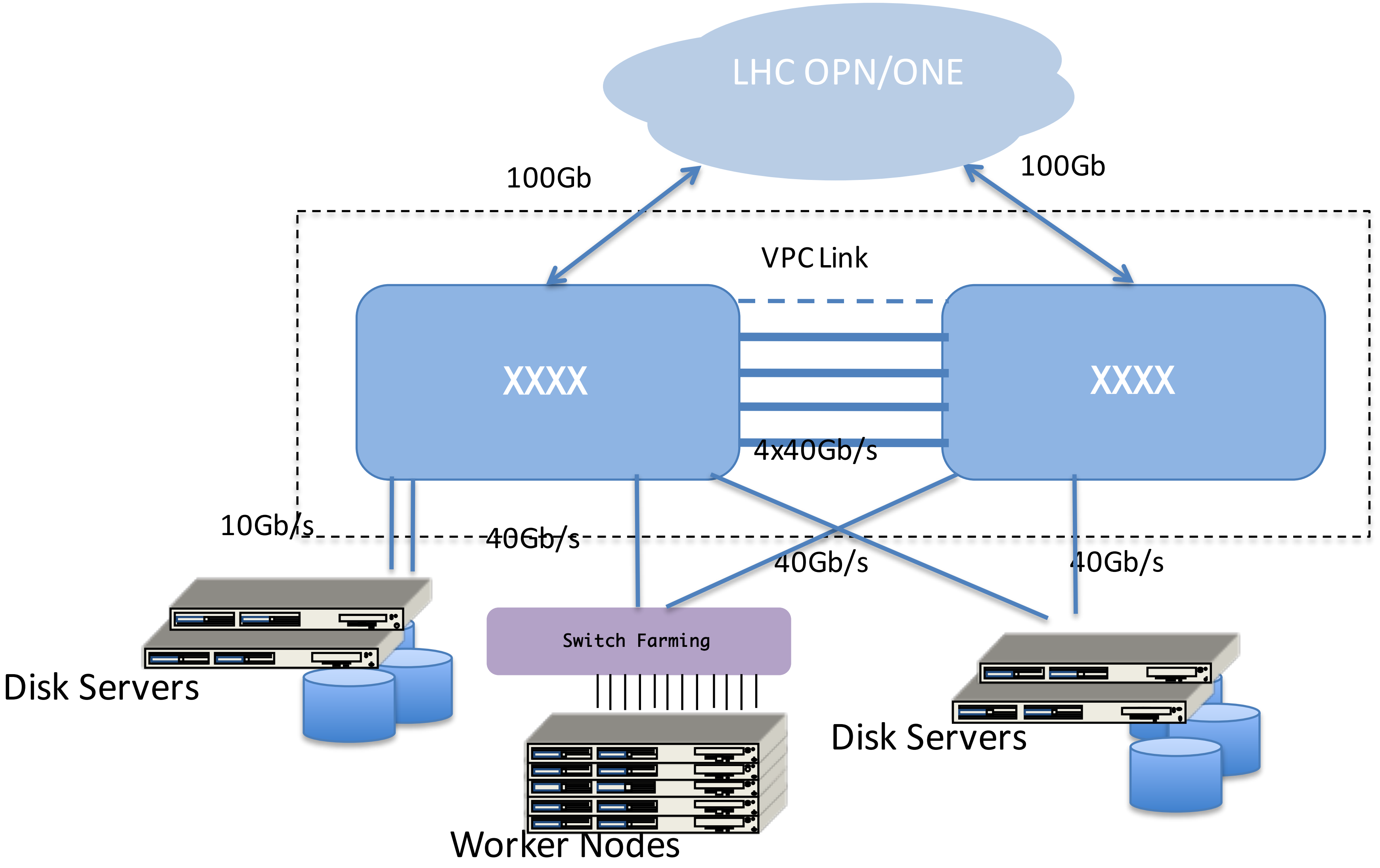
LHC ONE



Max In: 26.6 Gb/s Average In: 1.2 Gb/s

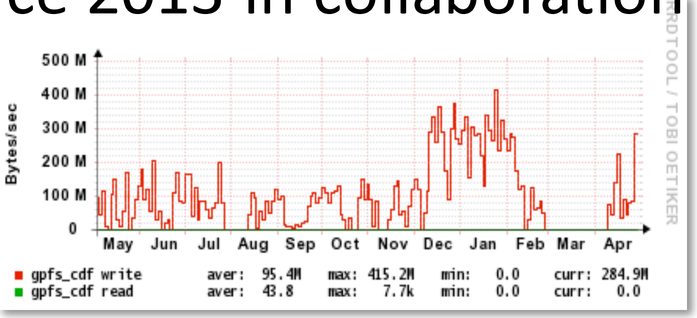
Max Out: 28.1 Gb/s Average Out: 2.4 Gb/s

Network infrastructure Evolution

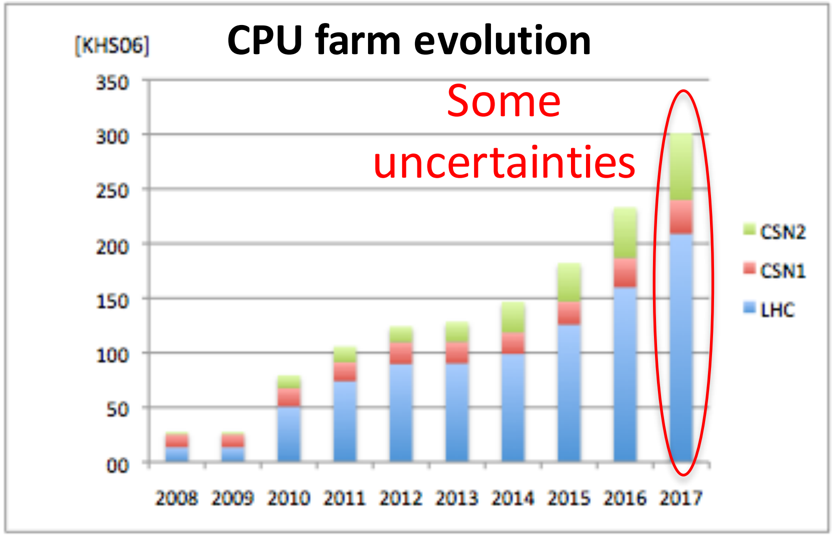


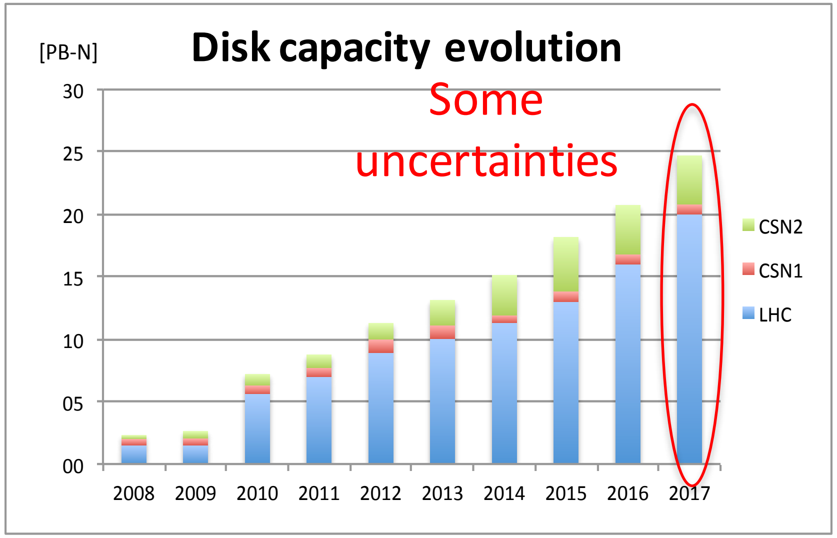
Long Term Data Preservation

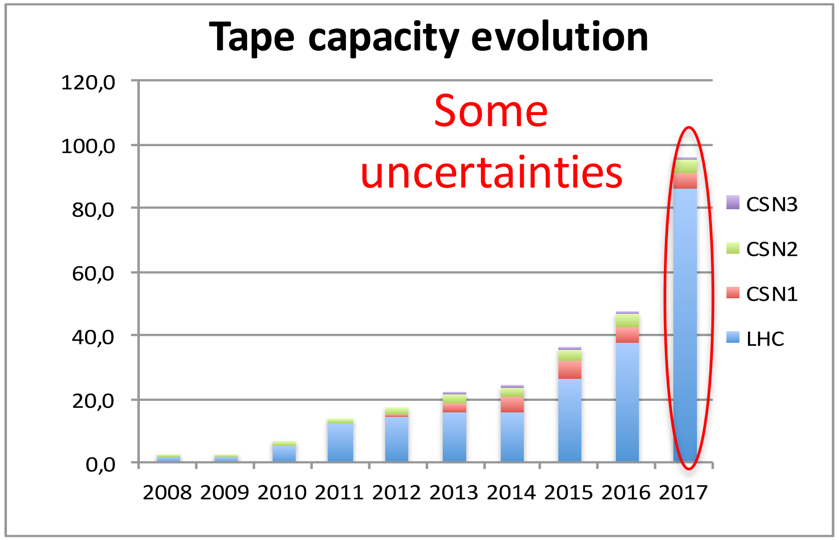
- LTPD activity for CDF is ongoing since 2013 in collaboration with Fermi Lab
- Two main areas of activity
  - Bit preservation
    - ~4 PB of data transferred and archived at CNAF (with a transfer rate up to 500 MB/s)
    - Automated system to perform regular checks of data integrity and copy back fro FNAL corrupted files is under development.
  - Preservation of code and analysis frameworks
    - instantiation on demand services and analysis computing resources on pre-packaged VMs
    - job submission to move from a dedicated portal (*Eurogrid*) to *jobsub*, to permit execution of legacy software on SL6 nodes.
    - The metadata, accessed directly at FNAL through Squid servers, will be copied to a local DB to ensure complete independence from FNAL.



LHC Run 2 Forecast







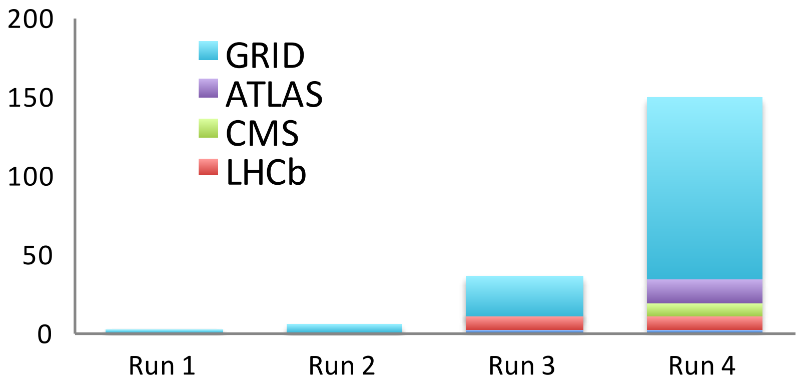
LHC Run 2

- Rapid growth of CPU and storage (disk/tape) driven by LHC
  - But a non-negligible increase of resources for Astro-Particle experiments in the following years foreseen (x4 according to some “rough” estimate)
- Hp.: CPU up to 300 kHS06, disk up to 27 PB-N and tape up to 100 PB
  - Δ CPU ~120 kHS06, Δ Disk ~10 PB-N, Δ tape ~80 PB
  - 2015 tender blades: ~19 kHS06/rack ➔ 120 kHS06 ~ 6 racks
  - High density storage: at least 1 PB/rack ➔ 10 PB-N ~ 10 racks
- Space: ~30 empty racks after 2015 resources installation
  - ~15-20 racks needed to add 2016-2017 resources
  - Space available for a new library if needed
- IT power can be increased up to ~1.2 MW with a safe (n+2) redundancy on the cooling system
  - Current IT load ~640 kW
  - The total IT load should remain under 1 MW

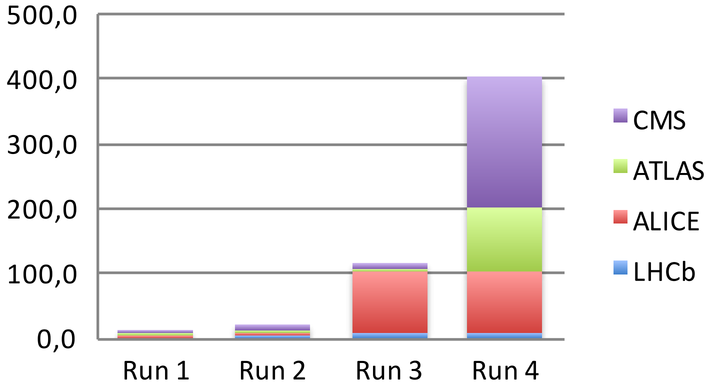
Data Center ready to host resources for LHC Run 2

Beyond Run 2?

- Huge increase of resources foreseen and our Data Center will be unlikely able to support it (budget issues not considered!)
- New technologies (e.g. GPU, low power processors)
- Data Center extension on remote sites?
- Data Center extension on Cloud?
  - Hybrid Cloud?

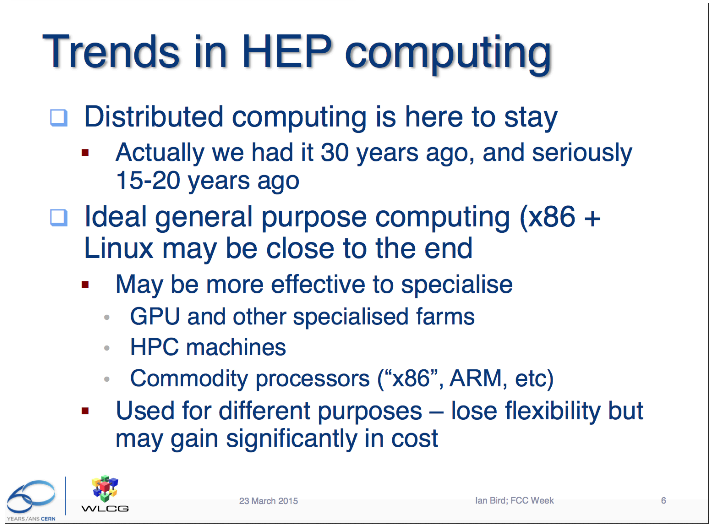


CPU requirements for online and offline processing



Raw data volumes estimations

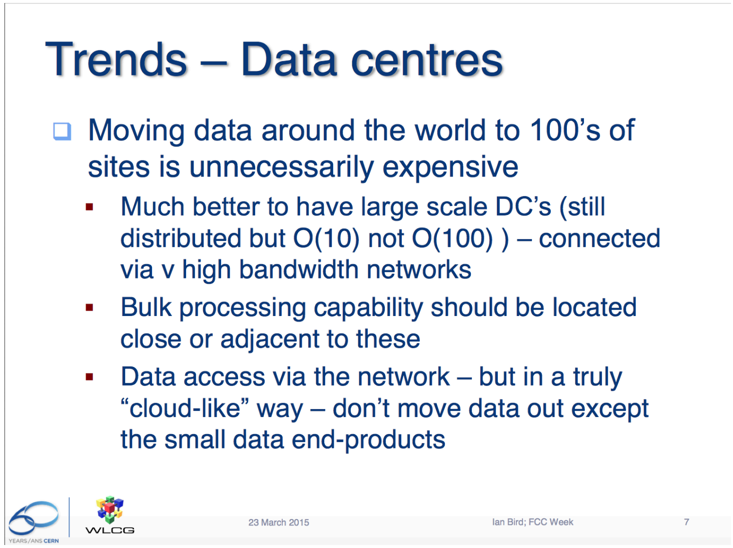
Trends



Trends in HEP computing

- Distributed computing is here to stay
  - Actually we had it 30 years ago, and seriously 15-20 years ago
- Ideal general purpose computing (x86 + Linux may be close to the end
  - May be more effective to specialise
    - GPU and other specialised farms
    - HPC machines
    - Commodity processors ("x86", ARM, etc)
  - Used for different purposes – lose flexibility but may gain significantly in cost

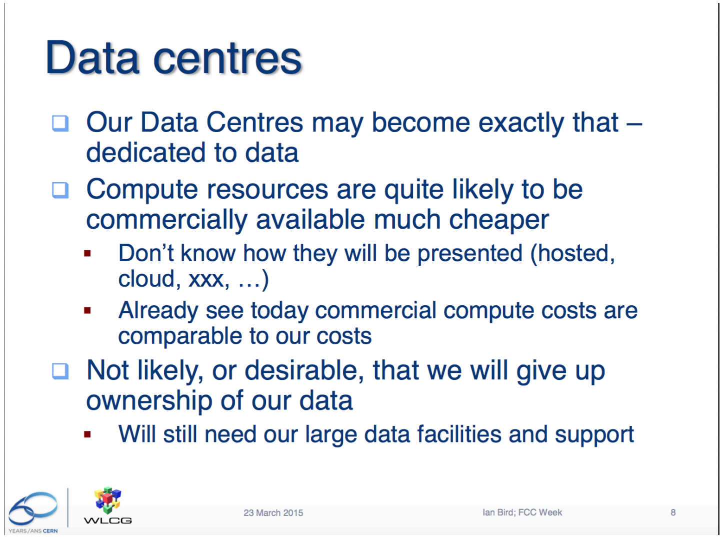
23 March 2015 Ian Bird; FCC Week 6



Trends – Data centres

- Moving data around the world to 100's of sites is unnecessarily expensive
  - Much better to have large scale DC's (still distributed but O(10) not O(100)) – connected via v high bandwidth networks
  - Bulk processing capability should be located close or adjacent to these
  - Data access via the network – but in a truly "cloud-like" way – don't move data out except the small data end-products

23 March 2015 Ian Bird; FCC Week 7



Data centres

- Our Data Centres may become exactly that – dedicated to data
- Compute resources are quite likely to be commercially available much cheaper
  - Don't know how they will be presented (hosted, cloud, xxx, …)
  - Already see today commercial compute costs are comparable to our costs
- Not likely, or desirable, that we will give up ownership of our data
  - Will still need our large data facilities and support

23 March 2015 Ian Bird; FCC Week 8

From Ian Bird's talk at WLCG workshop in Okinawa

Data Center extension and opportunistic use

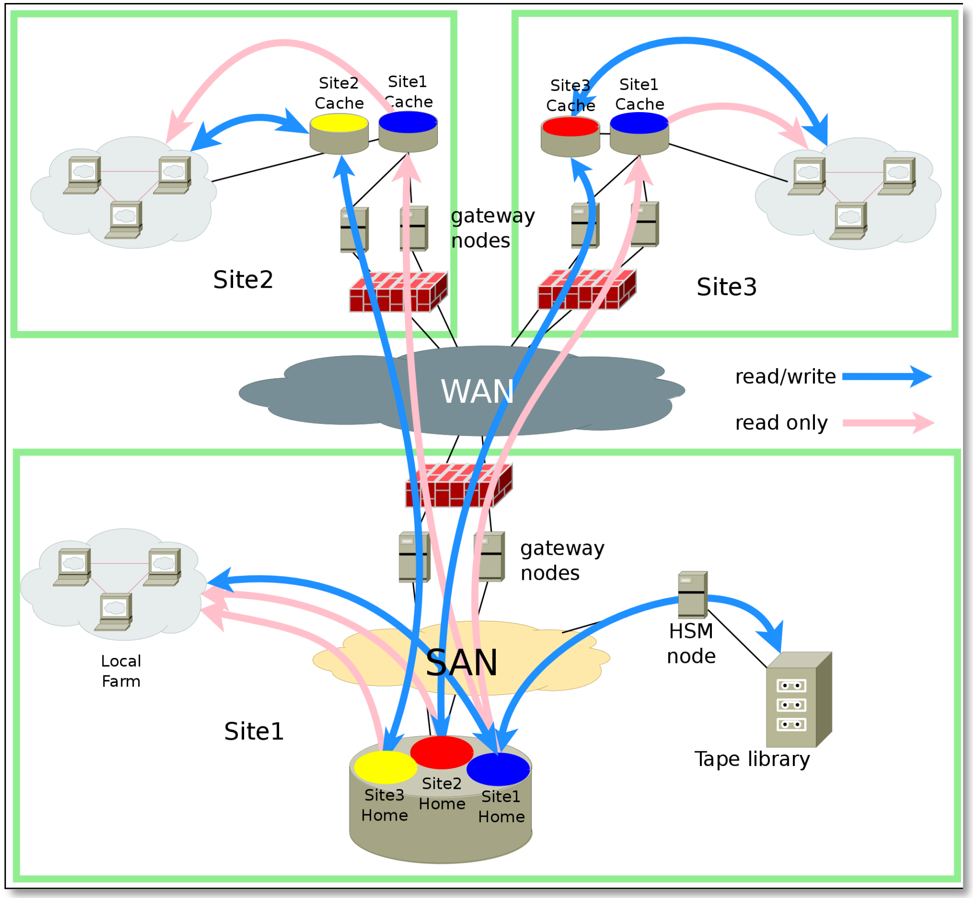
- Remote Data Center Extension under study
  - Functionality tests ongoing with another site on GARR
    - Goal: transparent LSF extension
  - Also pilot setup for transparent remote storage access with AMS and theorists groups
    - GPFS extension based on a new feature
- Opportunistic use
  - Preliminary contacts on going also with one of the main Commercial Cloud Providers and with Unicredit Bank
  - Use of other centers (e.g. GARR, RECAS)?
  - Planning tests with CINECA for HPC system
- HNSciCloud PCP proposal, if approved, will lead to build an hybrid cloud pilot with Commercial providers
  - Hybrid infrastructure as a Service (IaaS) platform
  - 70% funded by EU
  - (If approved and successful!) much larger project in 2 years

Remote data access via GPFS AFM

Cache basics

- Asynchronous updates
- Writes can continue when the WAN is unavailable
- TCP/IP for communication between sites (NFS or GPFS protocol)

- Two sides
  - Home - where the information lives
  - Cache
    - Data written to the cache is copied back to home as quickly as possible
    - Data is copied to the cache when requested
- Communication is done using NFS
- GPFS has it's own NFSv3 client
  - Automatic recovery in case of a communication failure
  - Parallel data transfers (even for a single file)
  - Transfers extended attributes and ACL's



HPC@CNAF

- (Small) HPC cluster also available
  - 24 nodes, 800 cores (~10 Tflops)
  - 17 GPUs
  - 3 Intel Xeon Phi

    ~20 TFlops (dbp)
  - Nodes interconnected via Infiniband
  - Operated with same tools as the generic farm
  - Dedicated GPFS storage (70 TB-N)
- Pilot project started in Jan 2014, now in production phase
  - Cluster used at 80% on average, with a total of about 10k jobs.
- Main users: theoretical physics groups (particle acceleration and laser plasma acceleration simulations)
- Interest expressed also by Virgo, Atlas etc…

Summary

- INFN Tier-1 ready to host resources for LHC Run2
  - But also for the non LHC experiments
- Exploring and testing new technologies
  - HPC, low power processors...
  - … but this is mainly driven by experiments' requirements and choices
- Exploring and starting to test data center extension on remote sites
  - Hybrid cloud?

Low power processors tests

- HP Moonshot with m350 cards and external storage
  - HP probed our WNs in order to determine the best storage solution
  - Providing us a dl380 as an iSCSI server
  - M300 cards with internal storage are too expensive according to HP

- Supermicro microblade
  - Each blade carries 4 motherboards and 4 discs, less compact but with built-in storage

What about Indigo?

- 1 FTE from DC dedicated to Indigo
  - Involvement from networking, farming and Data Management groups
  - 6 people in total
- Goal: to gain expertise on Cloud technology to improve service management and allow remote data center extension (e.g. CINECA, Aruba, PCP….)
- Program still to be detailed
  - Dynamic provisioning for Cloud
  - Virtualization & containers
  - Cloud storage

…….

The INFN Tier-1

- The Tier-1 is the main INFN computing centre providing computing and storage services to ~30 scientific collaborations
  - Tier-1 for LHC experiments (ATLAS, CMS, ALICE and LHCb)
  - Particle physics at accelerators
    - Kloe, LHCf, CDF, Agata, NA62, Belle2 (formerly also Babar and SuperB)
  - Astro and Space physics
    - ARGO (Tibet), AMS (Satellite), PAMELA (Satellite), MAGIC (Canary Islands), Auger (Argentina), Fermi/GLAST (Satellite)
  - Neutrino physics
    - Icarus, Borexino, Gerda, Opera, Cuore (Gran Sasso lab.)
    - KM3NeT (underwater)
  - Dark Matter search
    - Xenon, DarkSide (Gran Sasso lab.)
  - Gravitational waves physics
    - Virgo (EGO, Cascina)
  - Gamma Ray Observatory
    - CTA, LHAASO
  - LSPE and EUCLID in a near future?

Some other (r)evolutionary aspects

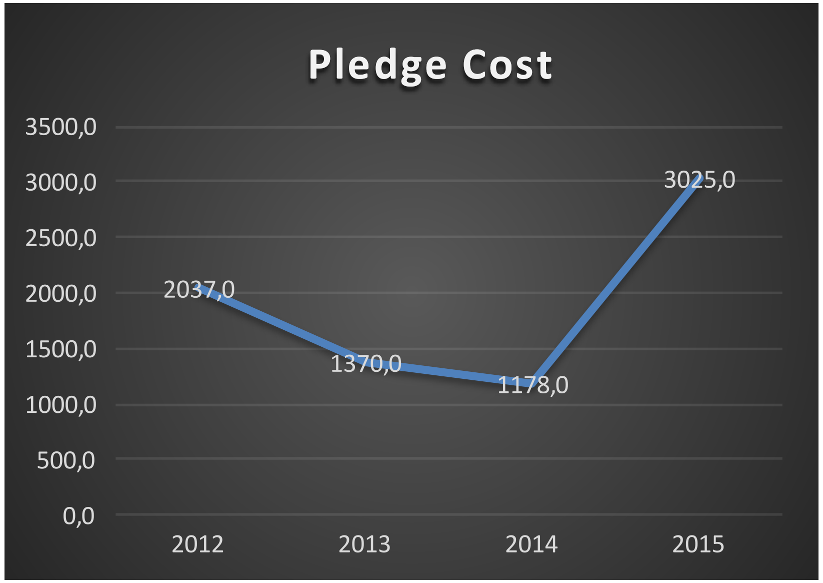
- Behind the scenes:
  - New provisioning system in place
  - (Still) looking for alternatives to LSF
  - Cloud and virtualization
  - Rethinking monitoring system
    - Exploring solutions based on modular components (graphana etc…)

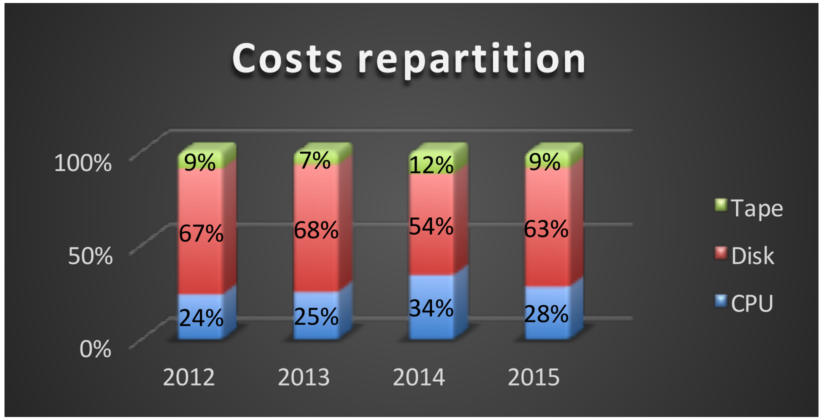
See Andrea's slides

- Network evolution
  - New core
  - 4x10 Gbps Disk-servers in next storage tender

Pledge Co$t

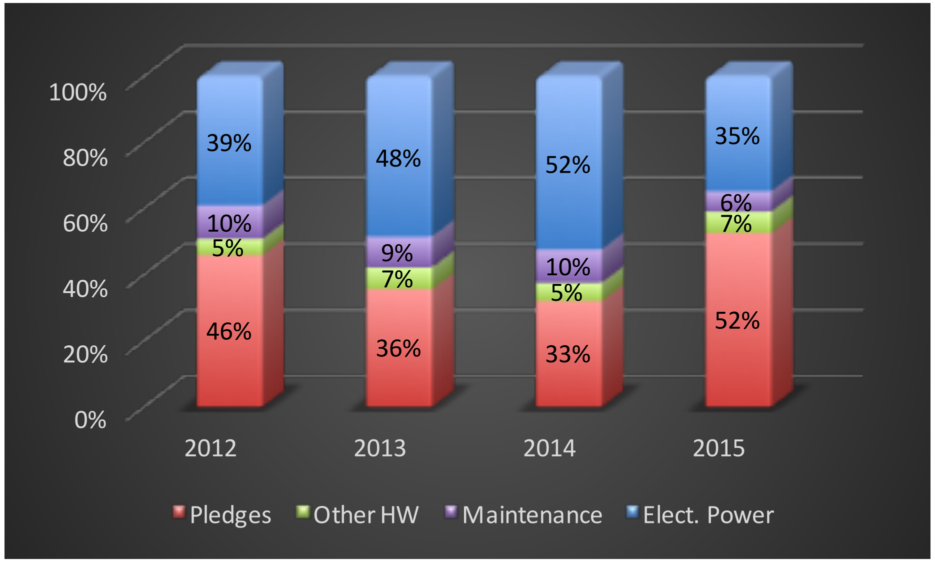
- In 2015 sharp increase of pledge costs due to storage (5 + 3 PB-N)
  - Phase-out of 2010 tender (5 PB-N)
- €/$ rate exchange does not help either (- 20% since May 1 2014)
  - 2015 CPU tender ended with HS06 unit cost 15% higher than expected
- Disk most costly component
- In order to lower disk cost acquired entry-level class storage in 2014 tender
  - 2 PB-N in 4 systems
  - Good price but still to be verified robustness and resiliency (in production since Easter 2015)
  - New interesting feature (Distributed Raid) (see Vladimir's slides)





Other Cot

- Pledges cost accounts usually for less than 50%
  - Higher this year due to large disk tender
- Maintenance costs (6% in 2015)
  - Facility systems including cooling, dynamic UPS's....



- Non-pledge Hw (7% in 2015)
  - E.g. network components, library drives,....
- Important contribution comes from the electricity cost
  - ~2 ME (equivalent to 35%) estimated for 2015 (0.19 E/kWh)
- Room for improvement mostly from reducing IT load
  - Strong interest for low power processors (see Andrea's presentation)