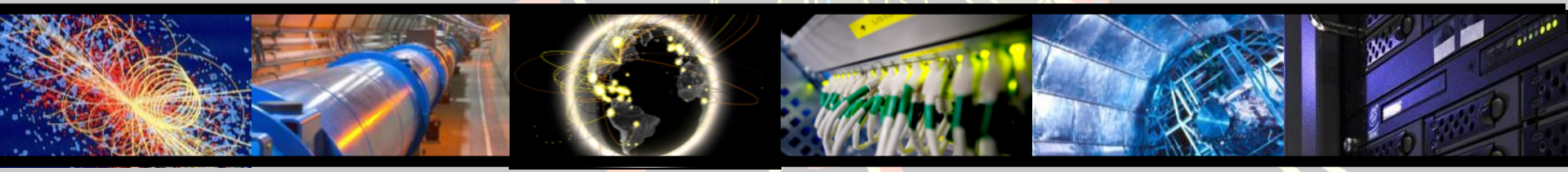
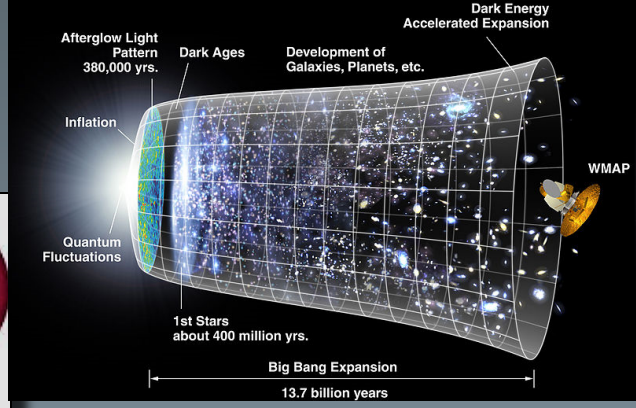
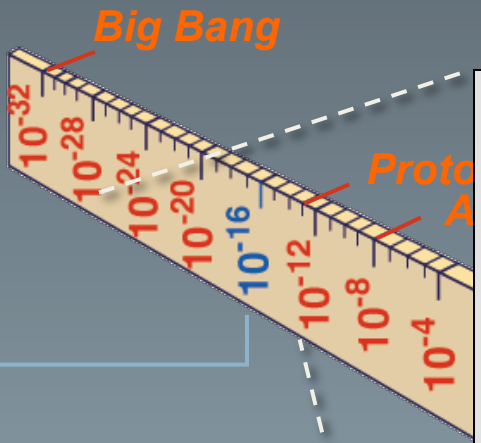
Experience with 30PB of Data from the LHC

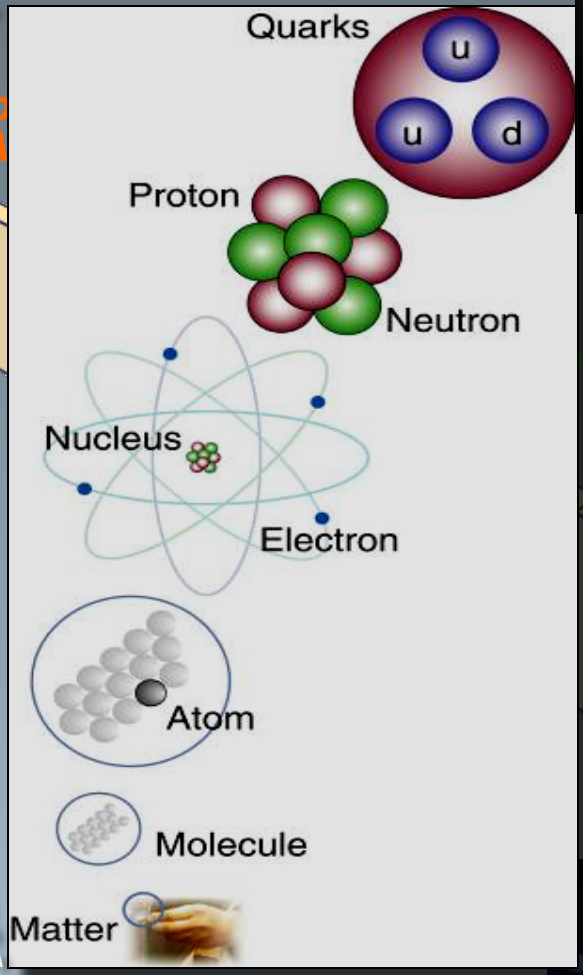
Michael Ernst
Brookhaven National Laboratory

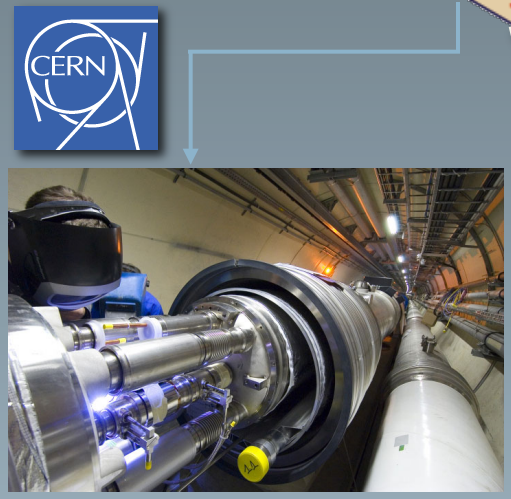
27th IEEE Symposium on Massive Storage Systems
24 May 2011, Denver





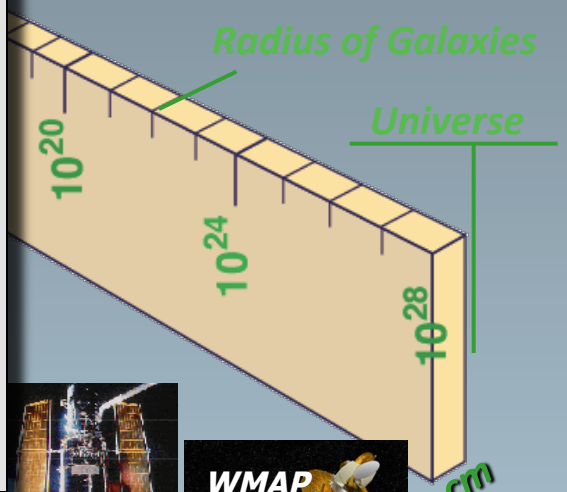


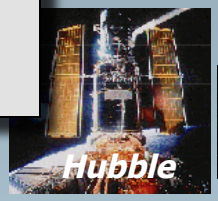




LHC

Super-Microscope



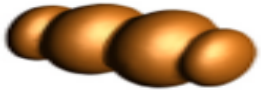


Study physics laws of first moments after Big Bang increasing Symbiosis between Particle Physics, Astrophysics and Cosmology

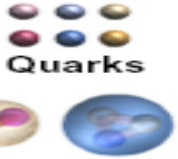
The Four Forces in Nature

Strong

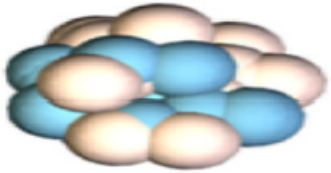
Gluons (8)



Quarks



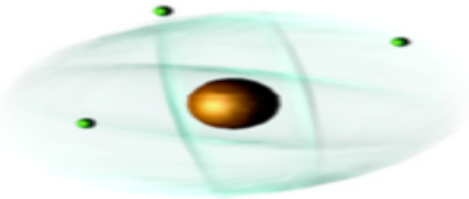
Mesons
Baryons



Nuclei

Electromagnetic

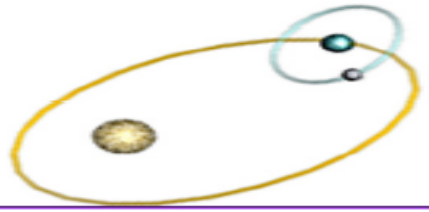
Photon



Atoms
Light
Chemistry
Electronics

Gravitational

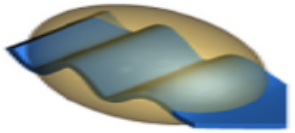
Graviton ?

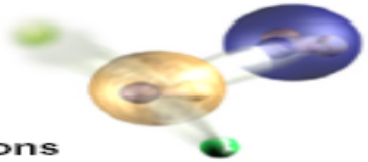


Solar system
Galaxies
Black holes

Weak

Bosons (W,Z)

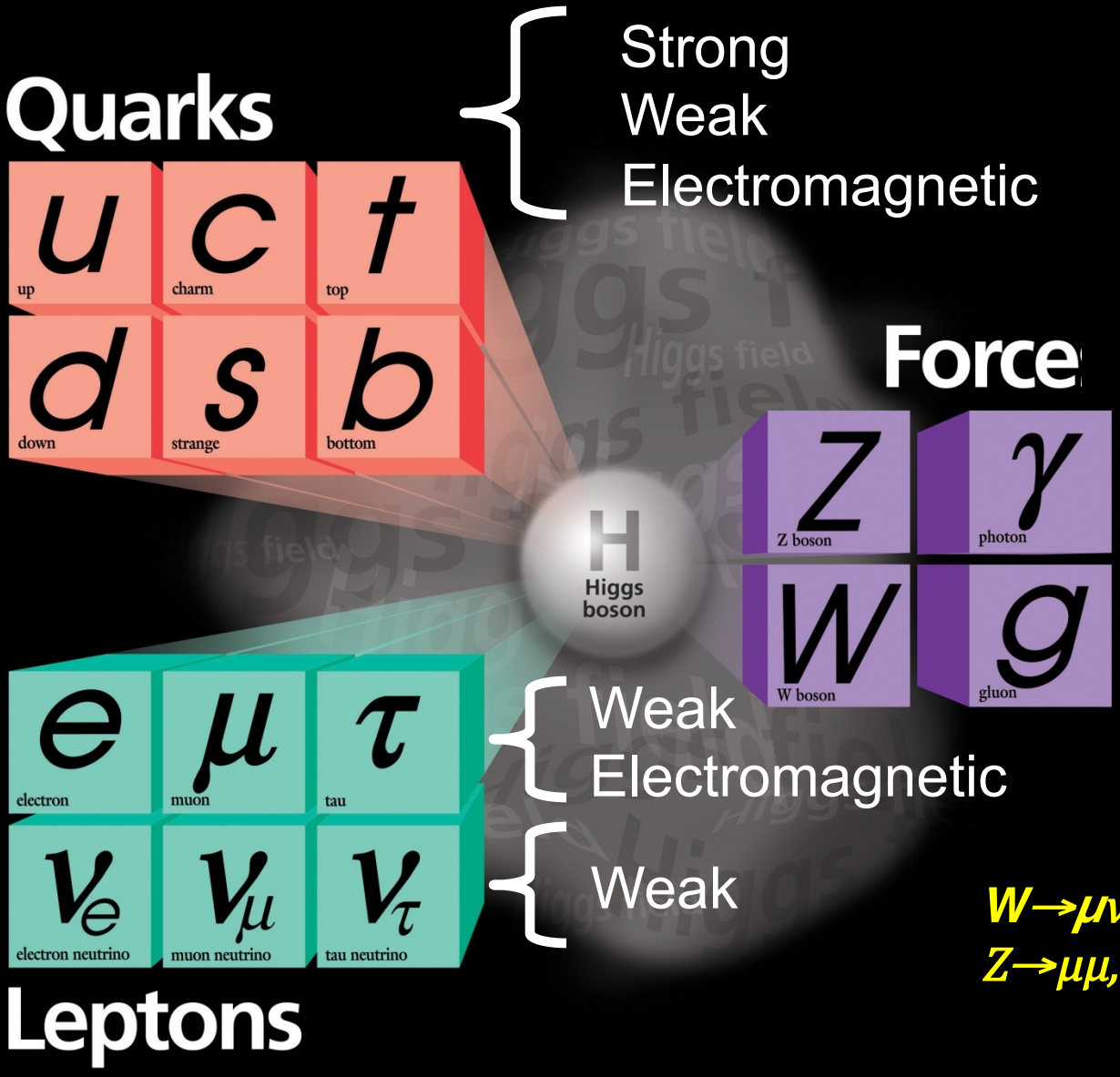


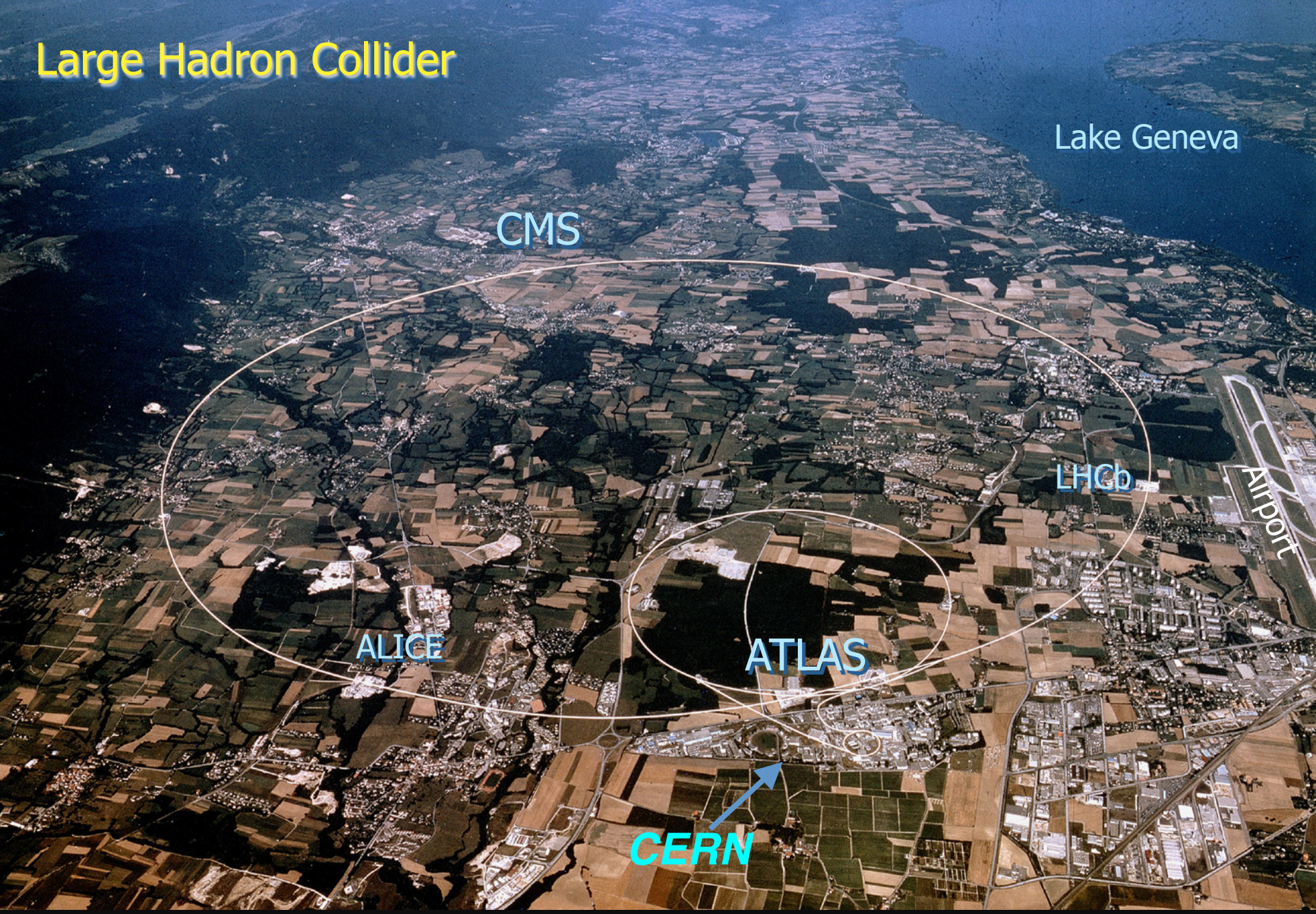


Neutron decay
Beta radioactivity
Neutrino interactions
Burning of the sun

The particle drawings are simple artistic representations

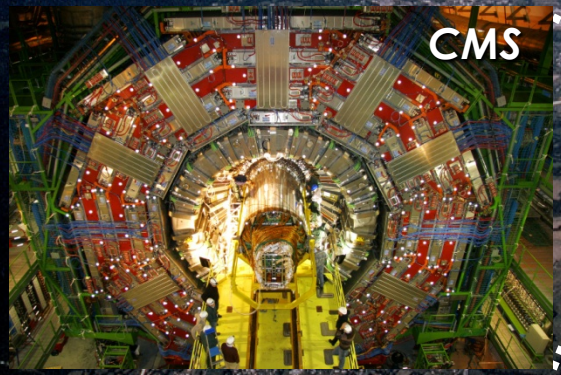
Standard Model of Particle Physics



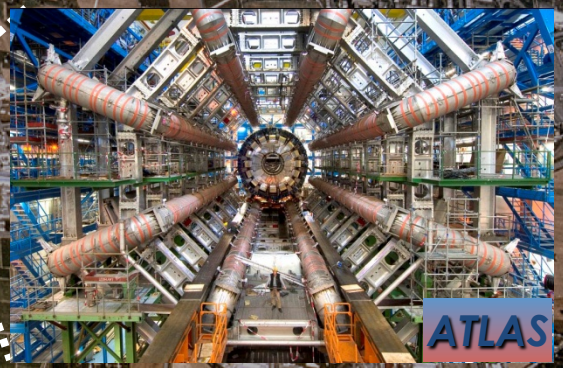

Large Hadron Collider
Lake Geneva
CMS
LHCb
Airport
ALICE
ATLAS
CERN

Enter a New Era in Fundamental Science

Start-up of the Large Hadron Collider (LHC), one of the largest and truly global scientific projects ever, is the most exciting turning point in particle physics.

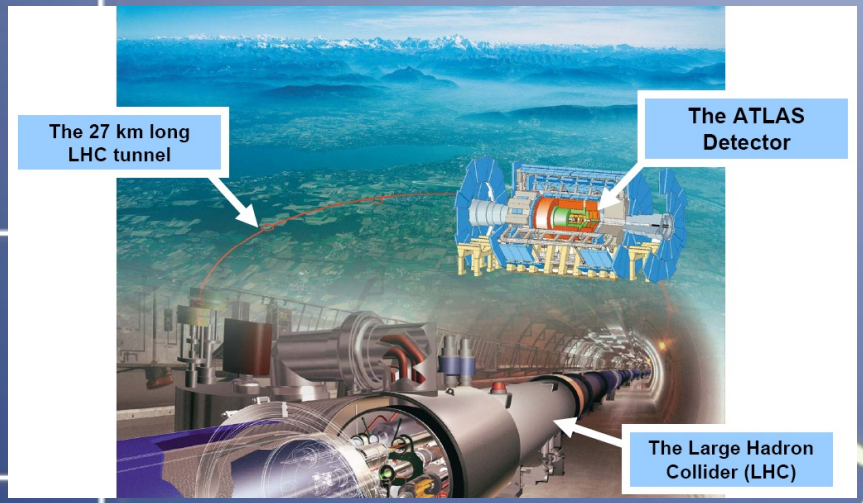


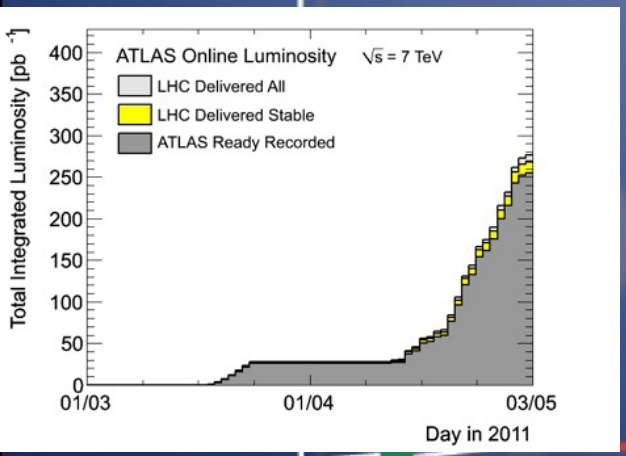
Exploration of a new energy frontier in proton-proton and heavy ion collisions

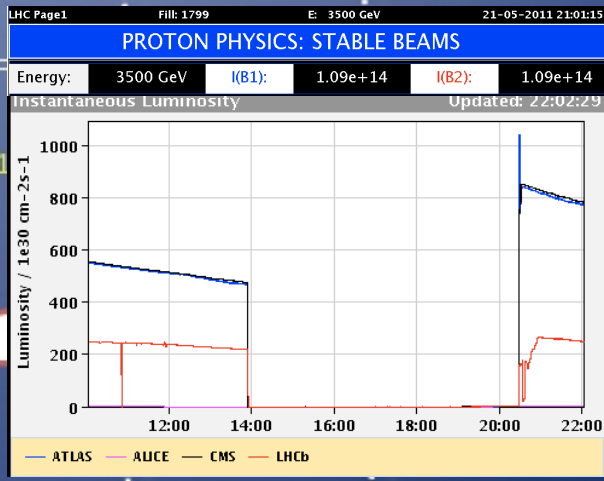


A multi-decade Program

A possible luminosity evolution with 2 years cycle with 2 major shutdowns?







~ 300 fb^{-1} by ~2020 ?

ions

New ID, LAr? ... sLHC

100 fb^{-1}

L= 1-2 * 10^{34}

LINAC 4 + IR

+ .....

L= 3 * 10^{33}

L_{int} ~ 50 fb^{-1}

L_{int}=10-25 fb^{-1}

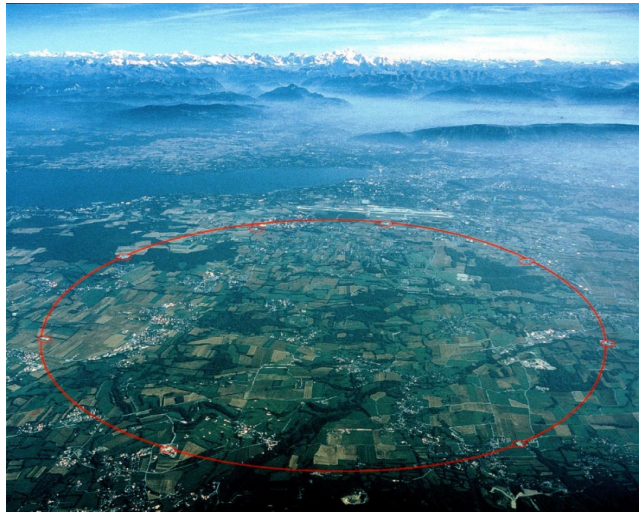
L_{int}=1 fb^{-1}

Year

12 14 16 18 20 22 24

Fastest

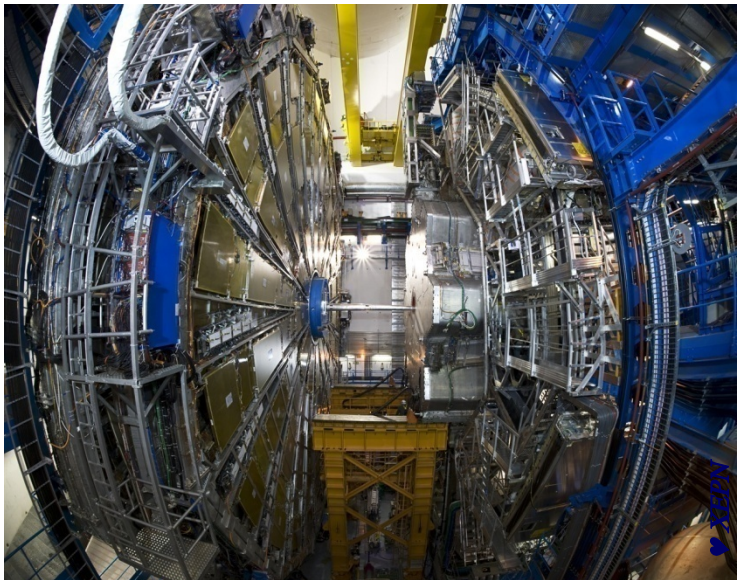
Trillions of protons travel the 16.5-mile-long tunnel



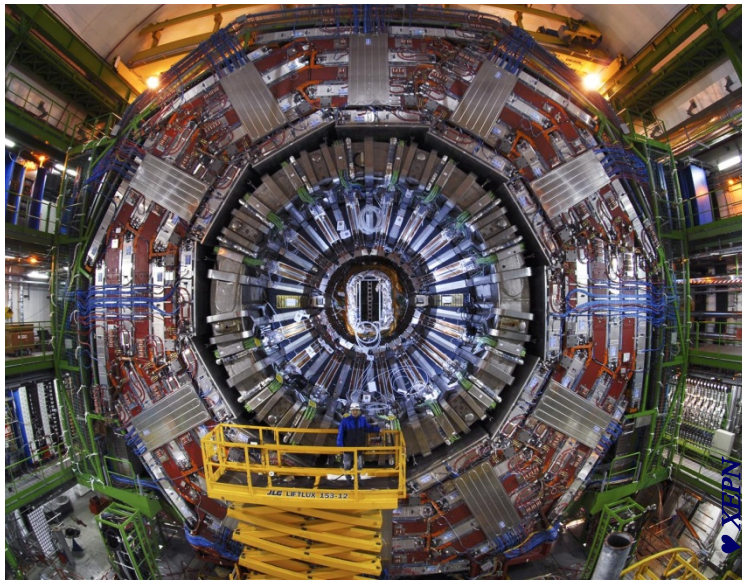
11,000 times a second (that's 670,626,025 mph)

Biggest

Largest, most complex detectors ever built

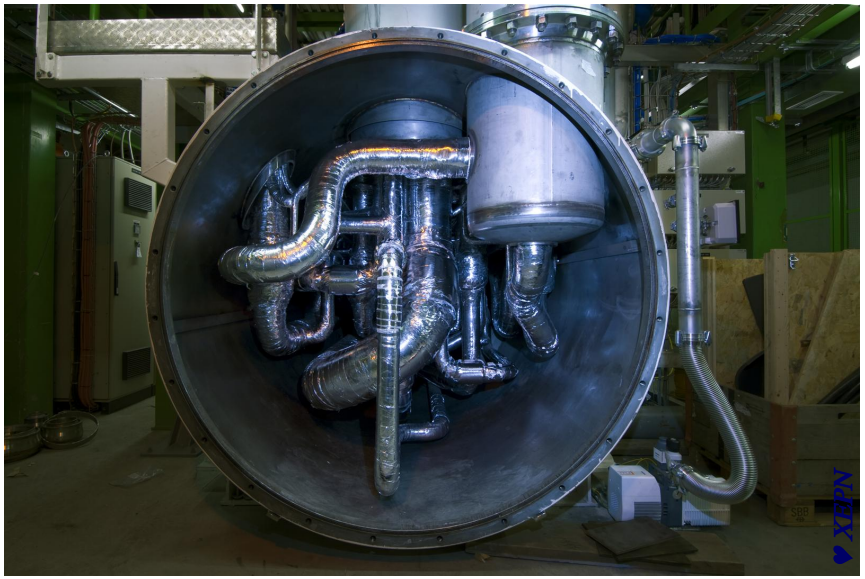


Study the tiniest particles with incredible precision



Coldest

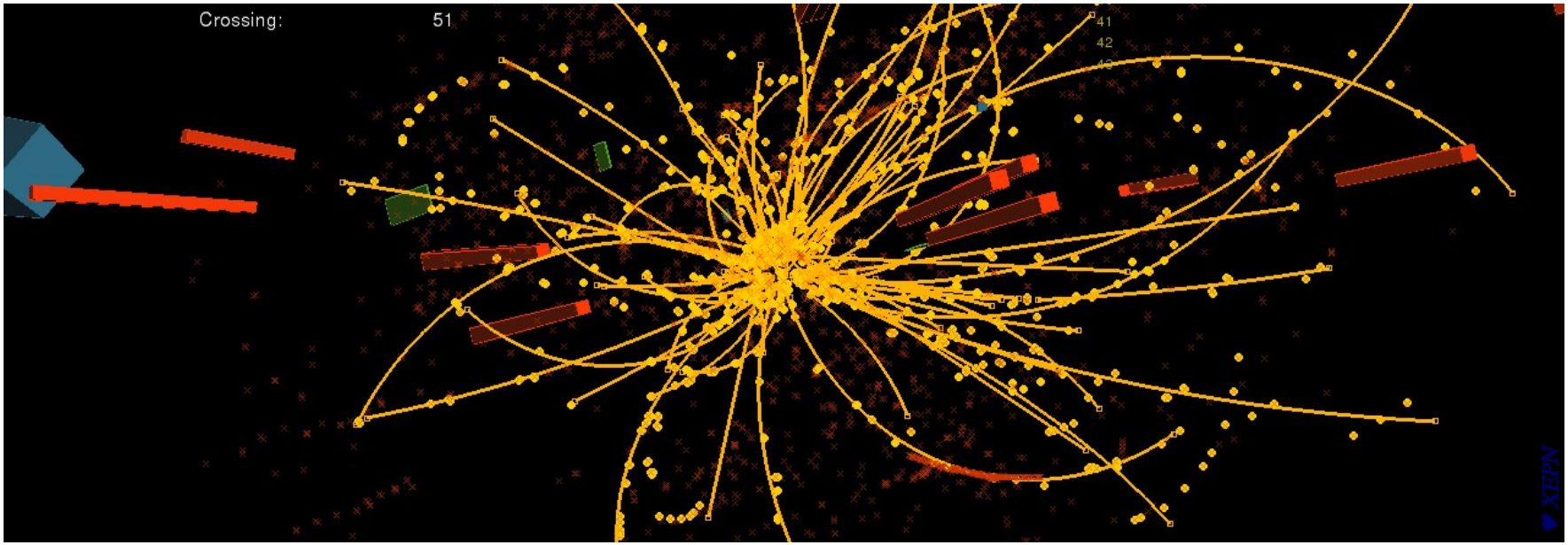
LHC's superconducting magnets operate at -456°F



Colder than the vacuum of outer space

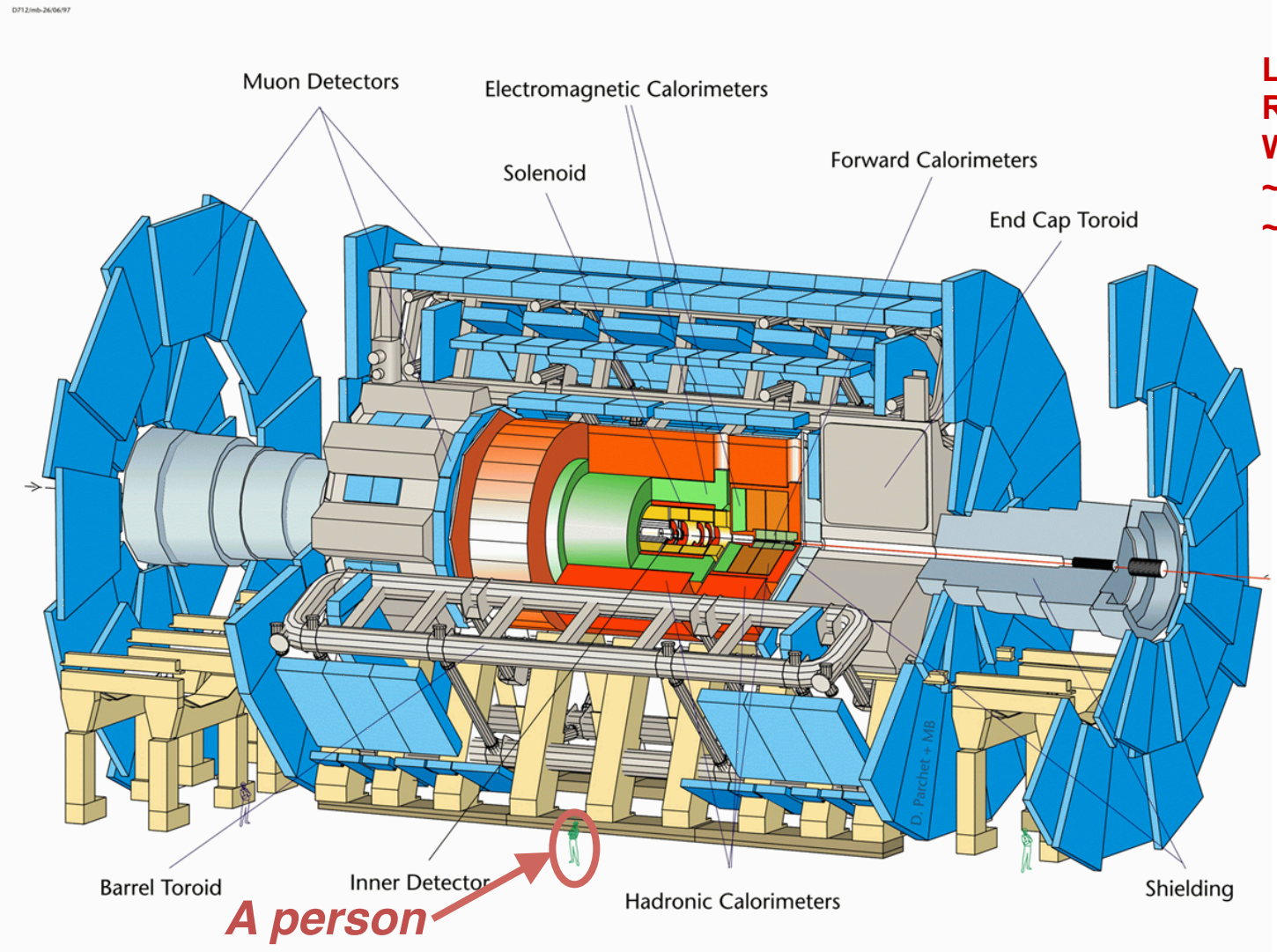
Hottest

Colliding protons generate temperatures 1 billion times hotter than the center of the sun

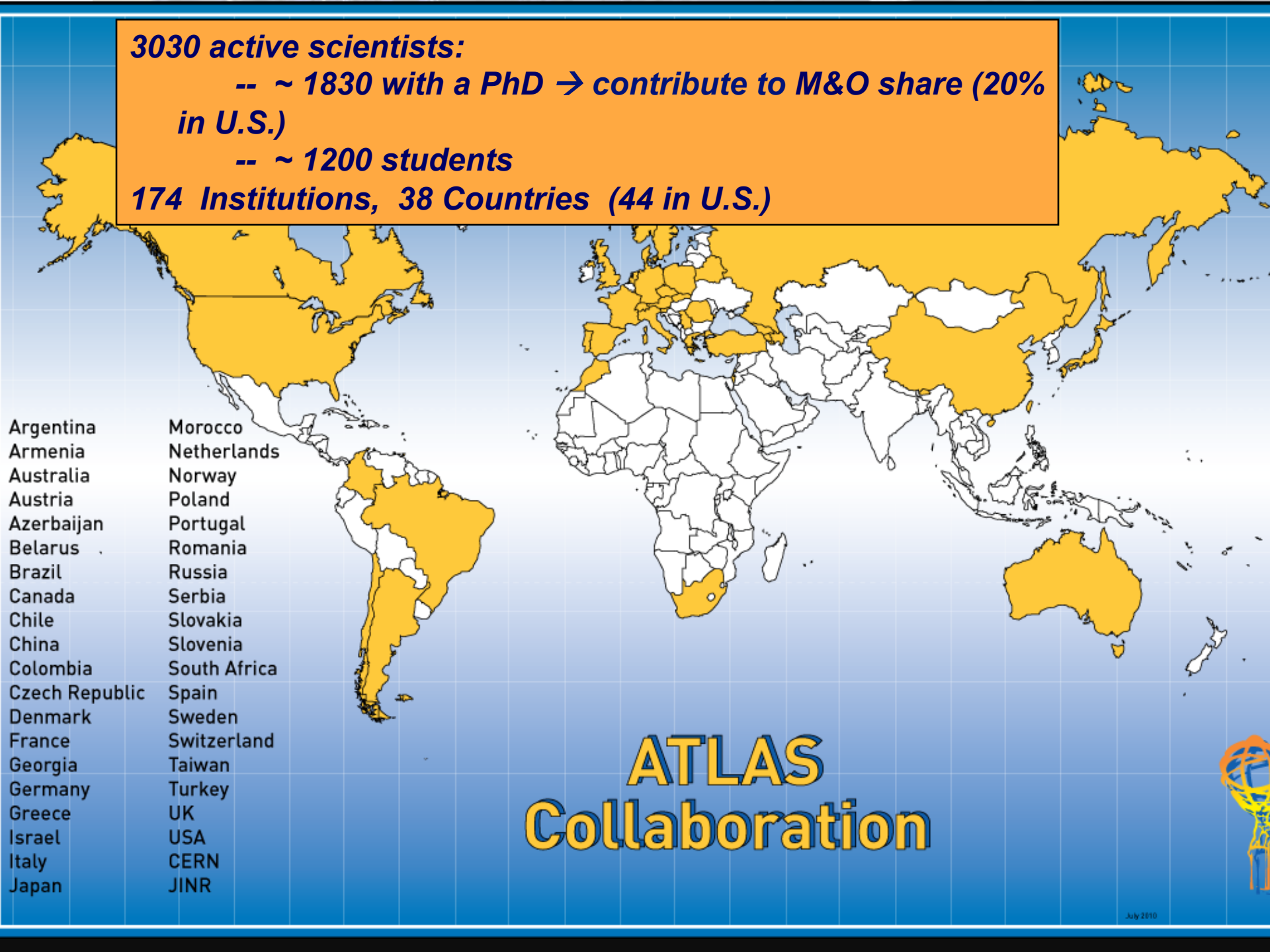


Detectors at the LHC are Huge

(Example: ATLAS)



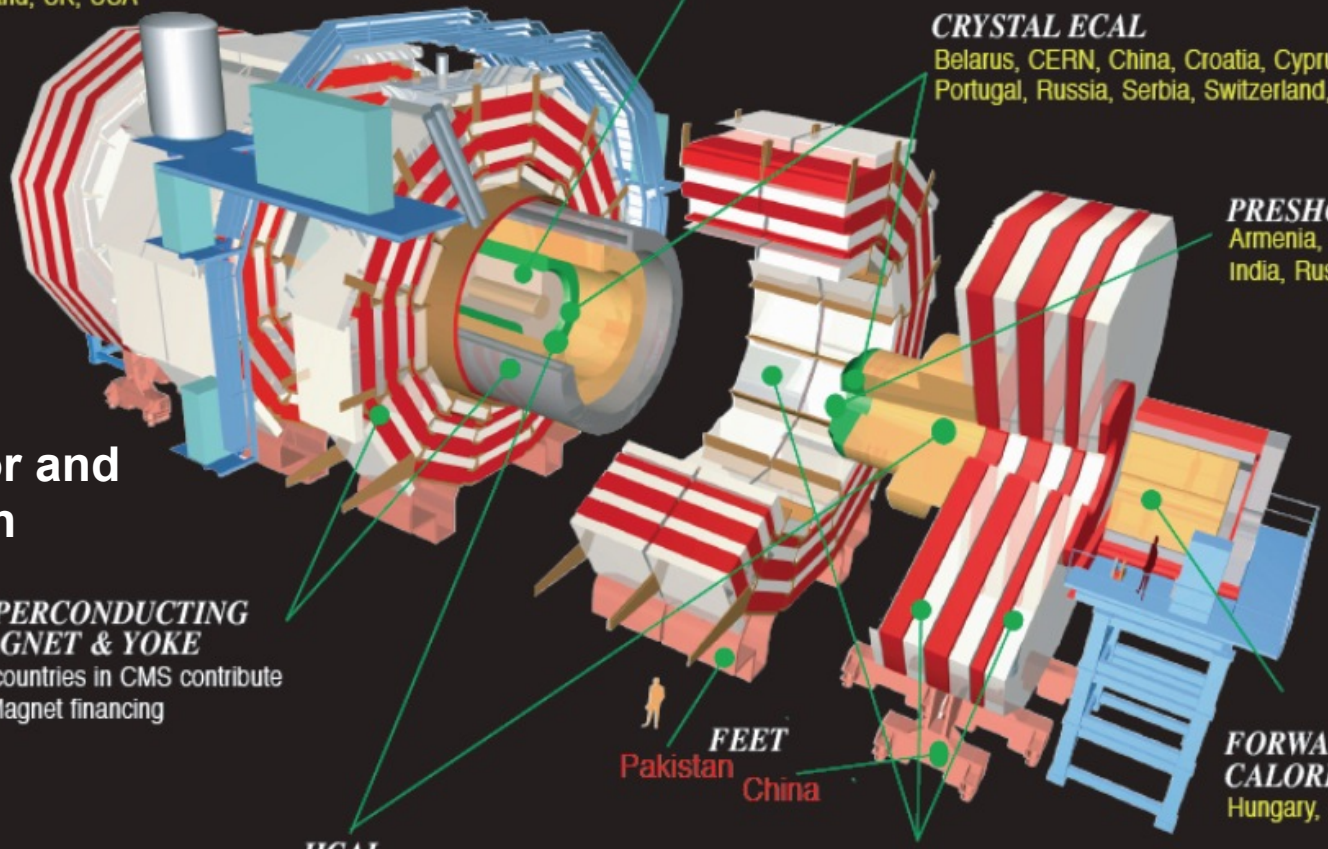
Length : ~ 46 m (150 ft)
Radius : ~ 12 m (40 ft)
Weight : ~ 7000 tons
~ 10^8 electronic channels
~ 1800 miles of cables


3030 active scientists:
-- ~ 1830 with a PhD → contribute to M&O share (20% in U.S.)
-- ~ 1200 students
174 Institutions, 38 Countries (44 in U.S.)
Argentina
Armenia
Australia
Austria
Azerbaijan
Belarus
Brazil
Canada
Chile
China
Colombia
Czech Republic
Denmark
France
Georgia
Germany
Greece
Israel
Italy
Japan
Morocco
Netherlands
Norway
Poland
Portugal
Romania
Russia
Serbia
Slovakia
Slovenia
South Africa
Spain
Sweden
Switzerland
Taiwan
Turkey
UK
USA
CERN
JINR
ATLAS
Collaboration
July 2010

July 2010

39 Countries, 169 Institutes, 3170 scientists and engineers (including about 800 students)

CMS Detector and Collaboration



TRIGGER, DATA ACQUISITION & OFFLINE COMPUTING
Austria, Brazil, CERN, Finland, France, Greece, Hungary, Ireland, Italy, Korea, Lithuania, New Zealand, Poland, Portugal, Switzerland, UK, USA

TRACKER
Austria, Belgium, CERN, Finland, France, Germany, Italy, Mexico, New Zealand, Switzerland, UK, USA

CRYSTAL ECAL
Belarus, CERN, China, Croatia, Cyprus, France, Italy, Portugal, Russia, Serbia, Switzerland, UK, USA

PRESHOWER
Armenia, CERN, Greece, India, Russia, Taiwan

SUPERCONDUCTING MAGNET & YOKE
All countries in CMS contribute to Magnet financing

FEET
Pakistan
China

FORWARD CALORIMETER
Hungary, Iran, Russia, Turkey, USA

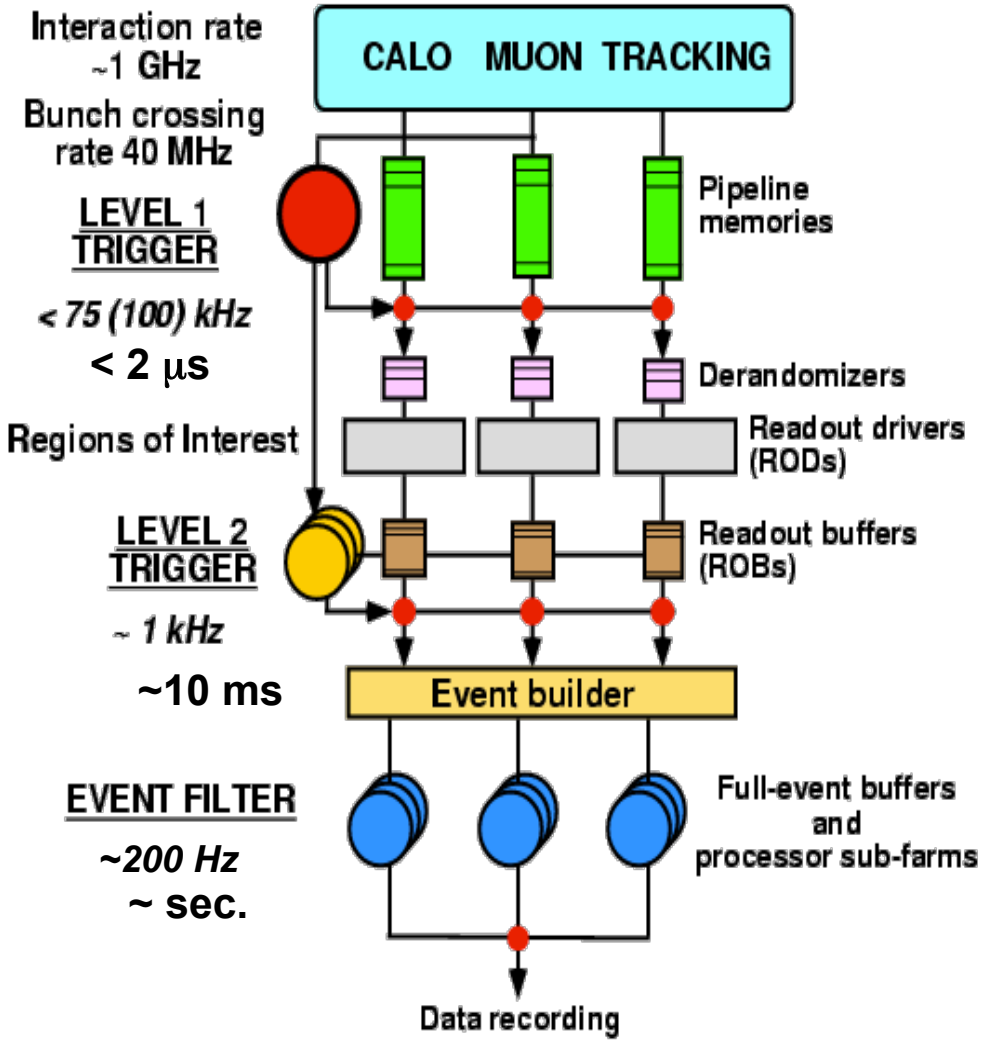
HCAL
Barrel: Bulgaria, India, USA
Endcap: Belarus, Bulgaria, Georgia, Russia, Ukraine, Uzbekistan
HO: India

MUON CHAMBERS
Barrel: Austria, Bulgaria, CERN, China, Germany, Hungary, Italy, Spain
Endcap: Belarus, Bulgaria, China, Colombia, Egypt, Korea, Pakistan, Russia, USA

Total weight	: 14000 tonnes
Overall diameter	: 15.0 m
Overall length	: 28.7 m
Magnetic field	: 3.8 T

Detectors produce a huge amount of Data

(Online Data reduction at ATLAS)



Physics selection of the 200 'best' events/sec:

40 MHz, 1 PB/sec

Level 1: Coarse calorimeter data and muon trigger chambers

75 kHz, 75 GB/sec

Level 2: Full information from all detectors in regions of interest

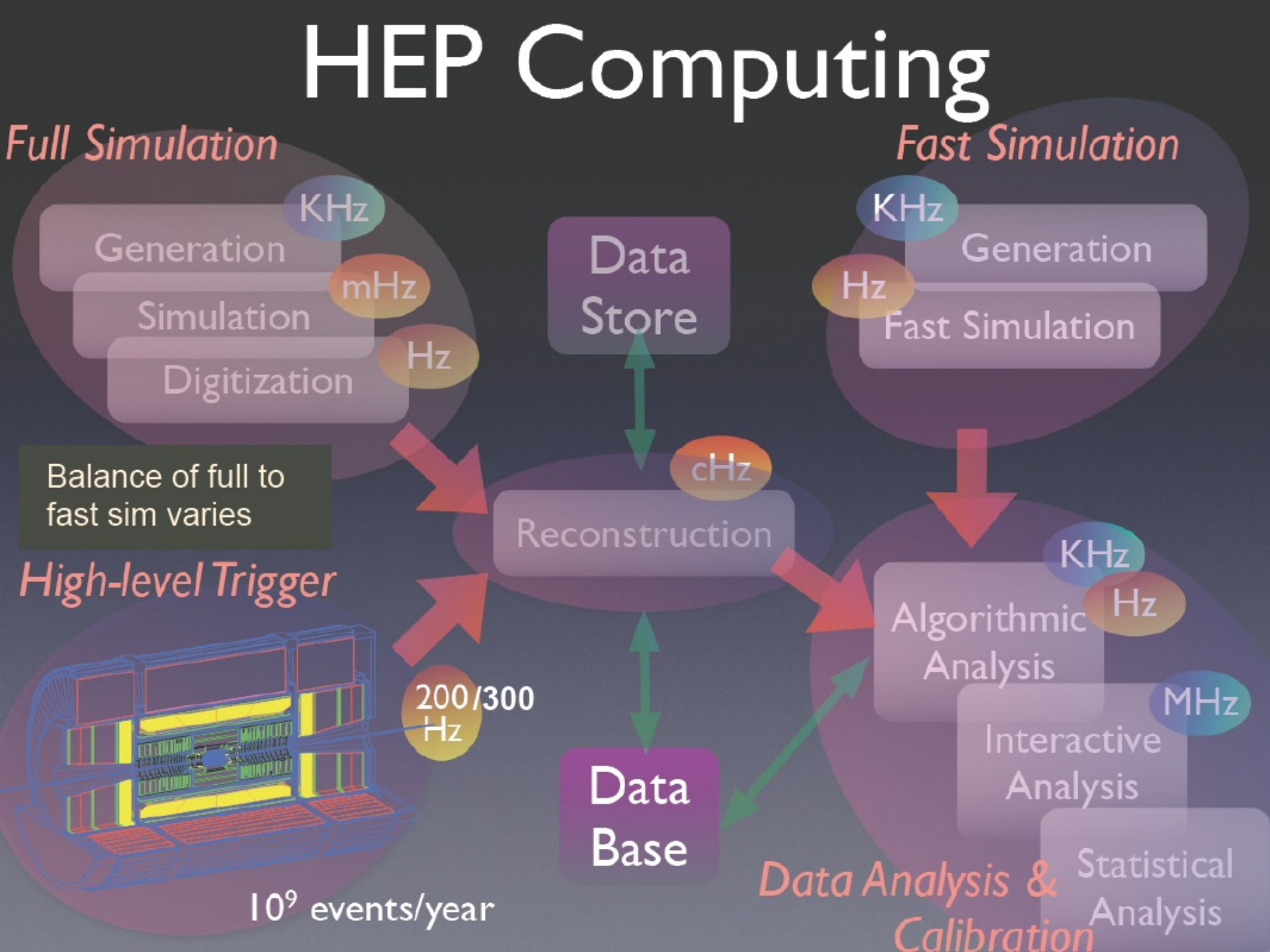
1 kHz, 1 GB/sec

Event Filter: Reconstruction of complete event using latest alignment and calibration data

200 Hz, ~320 MB/sec

~20 TB/day, 2 Petabyte/year of recorded raw data

HEP Computing



Computing Model at the Beginning

- Resources Spread Around the GRID

Data Reprocessed potentially regularly
Archive RAW and RECO
Synchronize RECO and AOD to T1 Centers

•Interactive Analysis
•Plots, Fits, Toy MC, Studies, ...

Tier 3

DPD

100 Sites Worldwide

Tier 2

AOD

RECO

•Derive 1st pass calibrations within 24 hours.
•Reconstruct rest of the data keeping up with data taking.

Tier 1

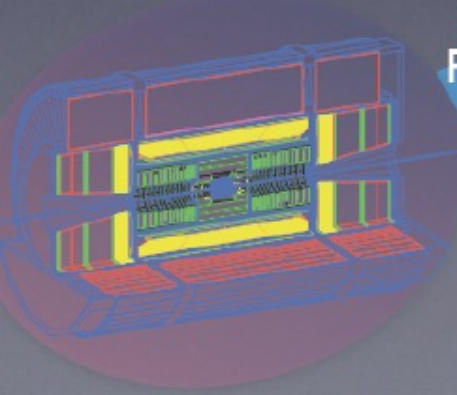
11 Sites Worldwide

RAW/ AOD/ ESD

Tier 0

RAW

•Production of simulated events.
•User Analysis: 12 CPU/ Analyzer
•Disk Store: AOD

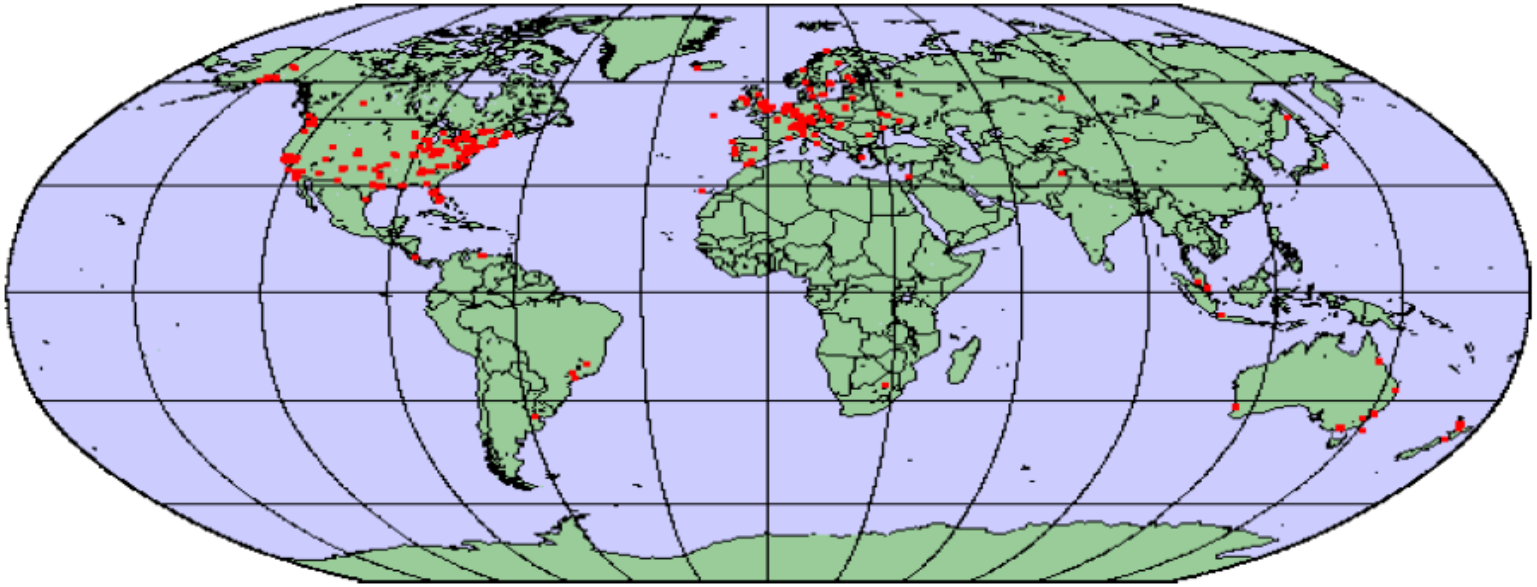


CERN Analysis Facility

•Primary purpose: calibrations
•Small subset of collaboration will have access to full ESD.
•Limited Access to RAW Data.

Worldwide Collaboration

- More than 6000 users at ~450 institutions from around the world are participating in the LHC Experiments

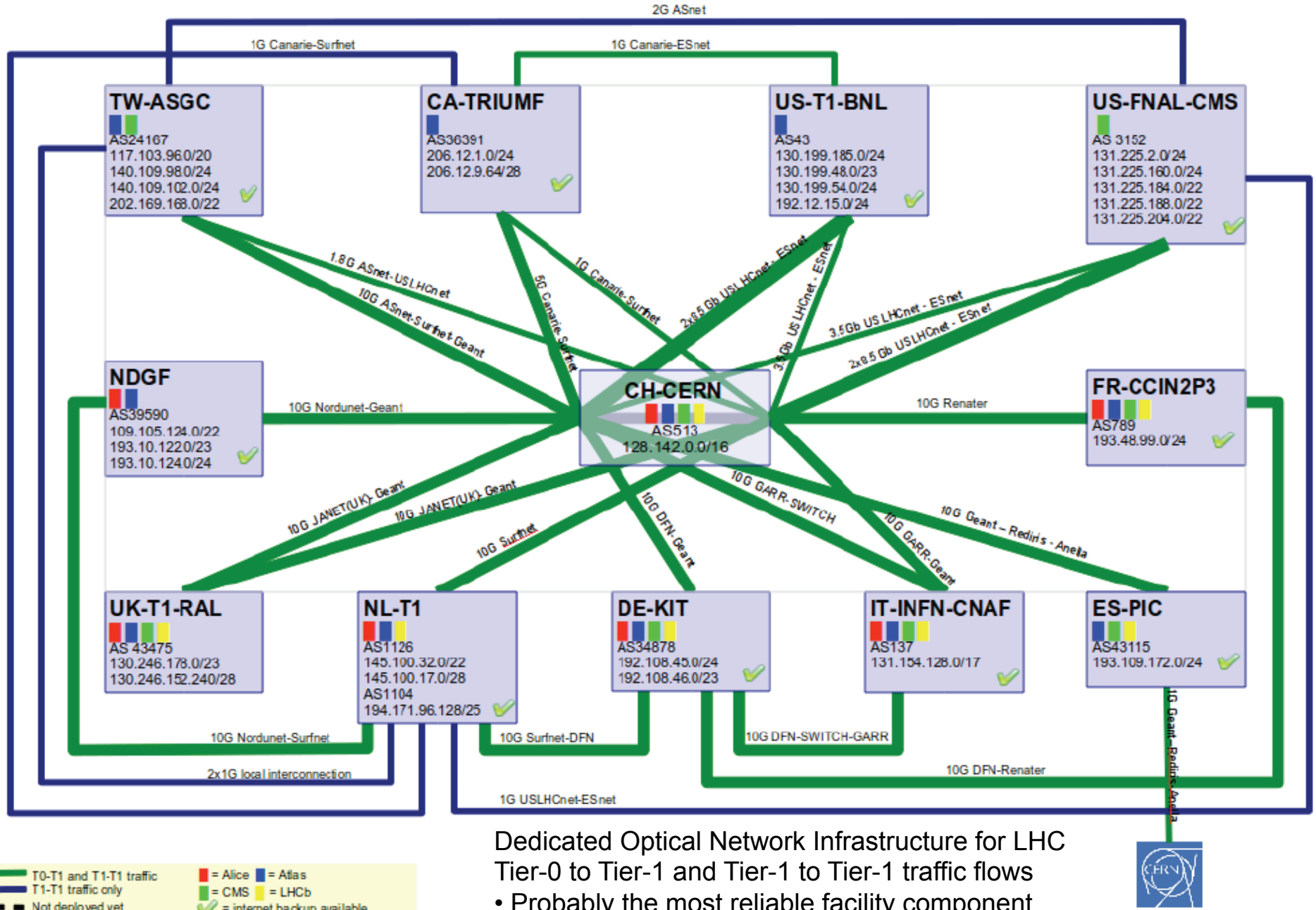


- LHC Computing unites the computing resources for particle physicists in the world

The Worldwide LHC Computing Grid Project (WLCG)

- **Approach**
  - To prepare, deploy and operate the computing environment for the experiments to analyze the data from the LHC detectors
    - HEP community very experienced for decades in consent-based collaborations
  - Applications development environment, common tools and frameworks
  - Build and operate the LHC computing service
- The Grid is just a tool towards achieving this
- **A Collaboration between**
  - The physicists and computing specialists from the LHC experiments
  - The projects in Europe and the US that have been developing Grid middleware
  - The regional and national computing centers that provide resources for LHC
  - The research networks

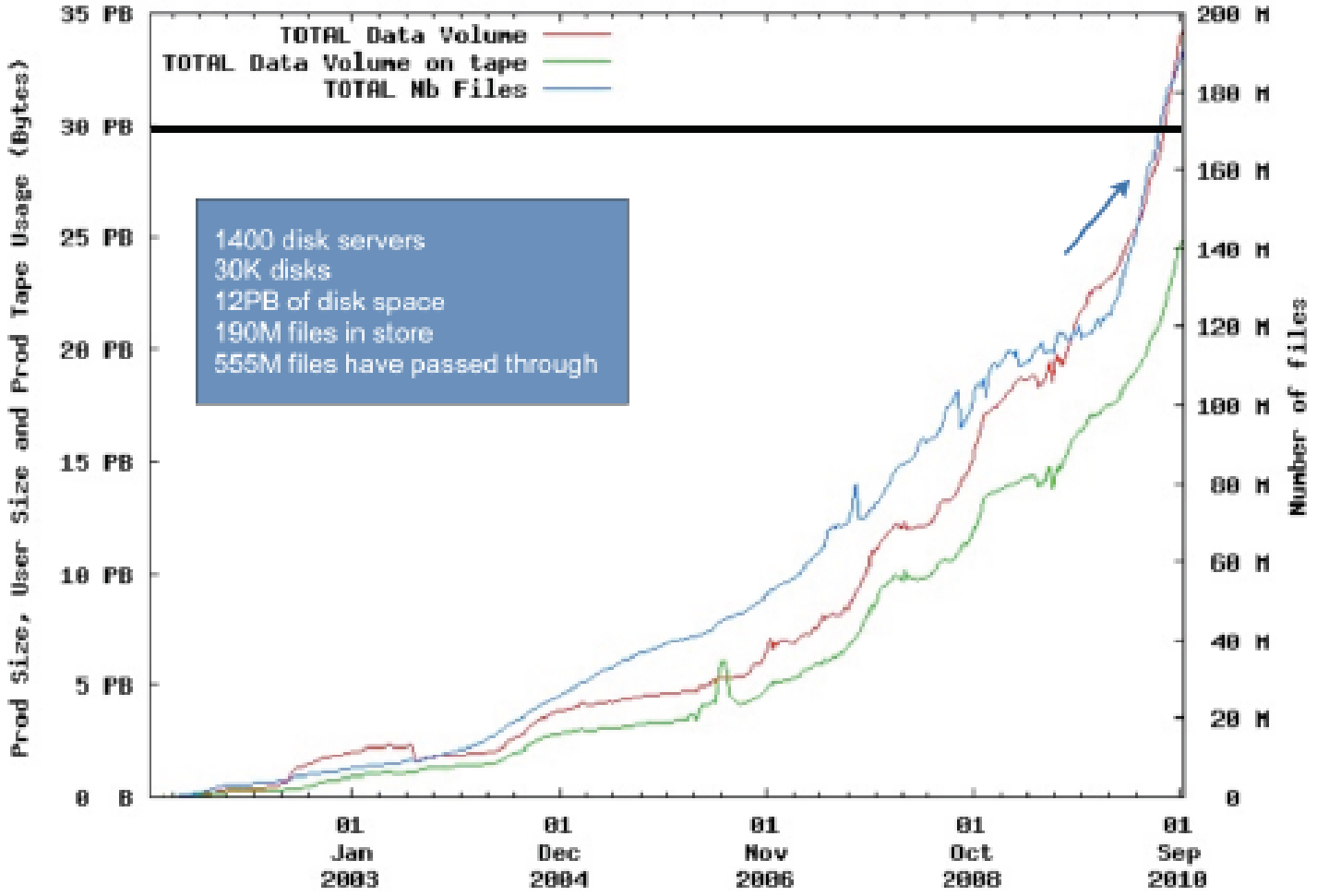
LHCOPN



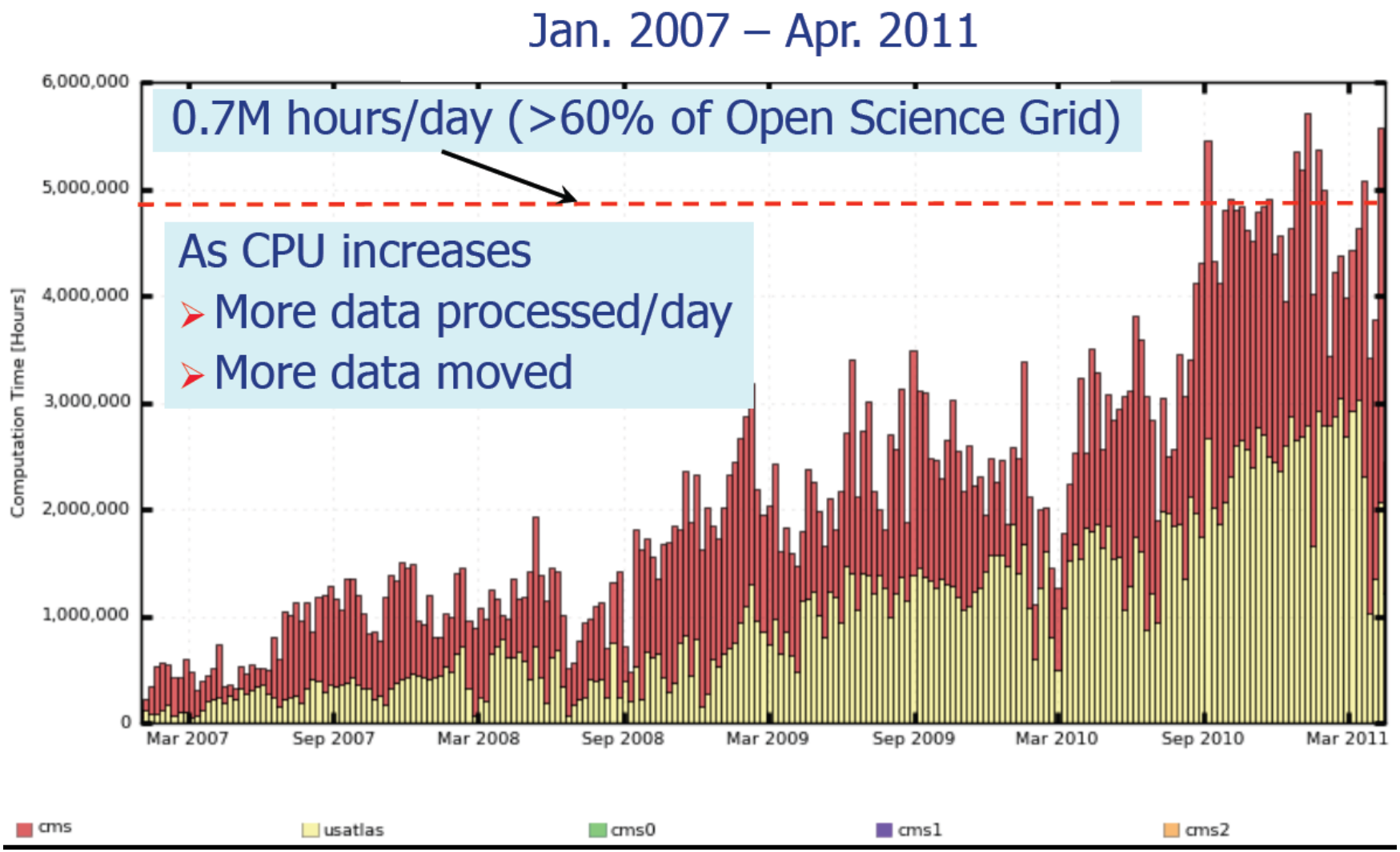
Dedicated Optical Network Infrastructure for LHC Tier-0 to Tier-1 and Tier-1 to Tier-1 traffic flows

- Probably the most reliable facility component

Evolution of Storage Capacity at CERN

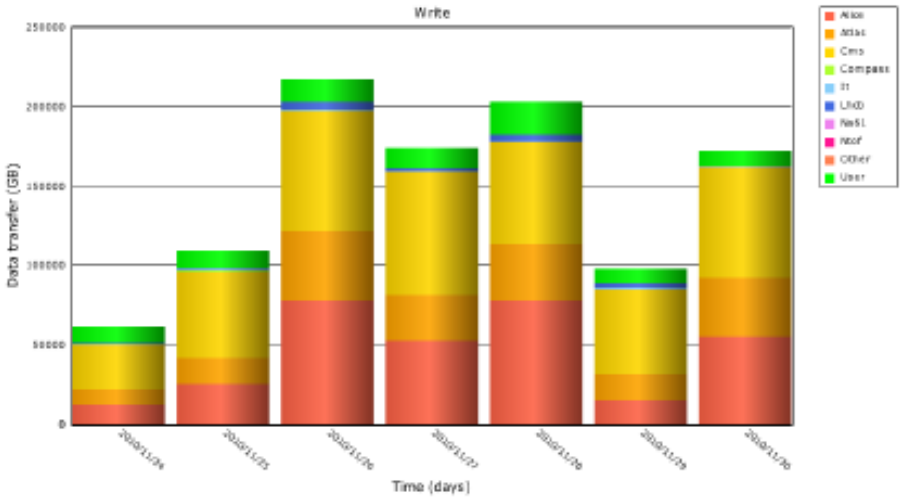


Computing Growth: ATLAS and CMS in the U.S.

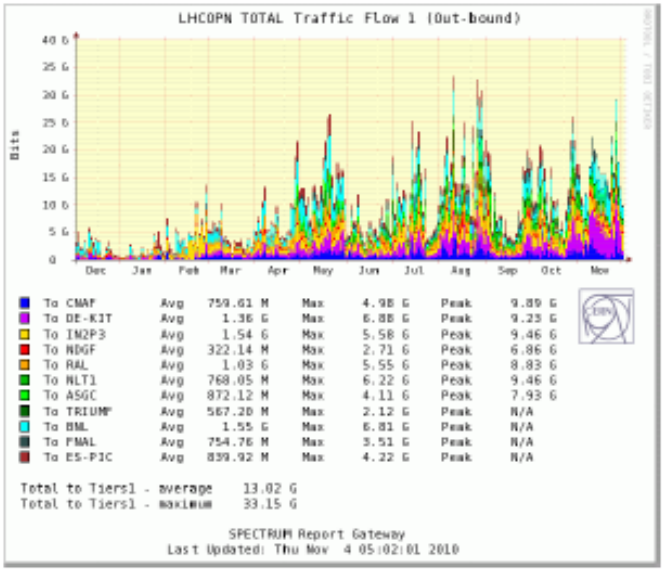


2010 Data Taking at Tier-0

Tape recording: 220TB/day



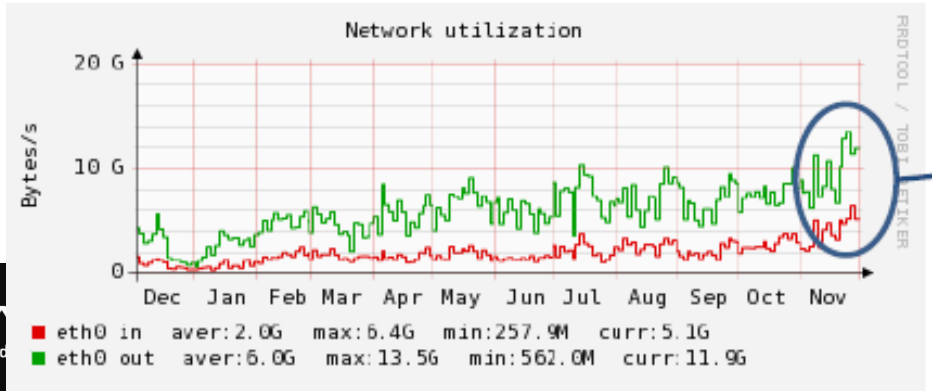
LHCOPN External Networking: Avg(year): 13 Gb/s with peaks at 70 Gb/s

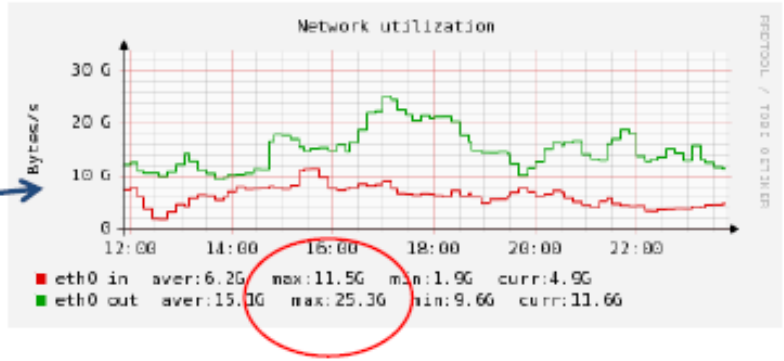


Tier-0 Bandwidth

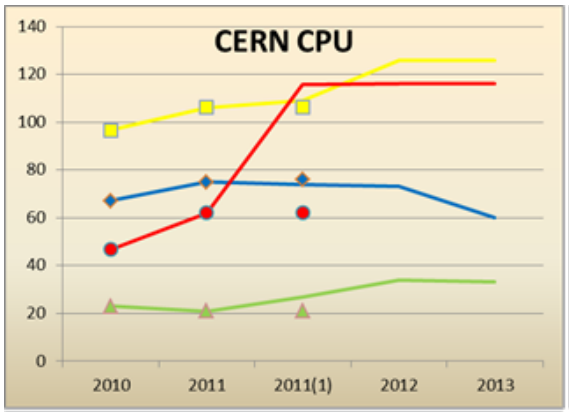
Average in: 2 GB/s with peaks at 11.5 GB/s

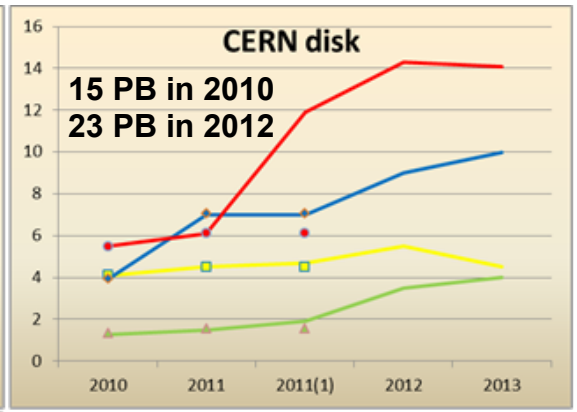
Average out: 6 GB/s with peaks at 25 GB/s

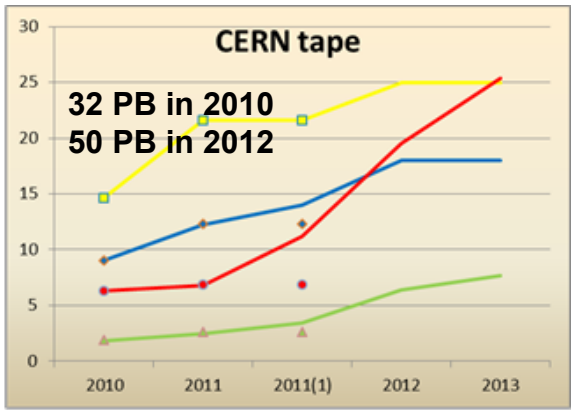


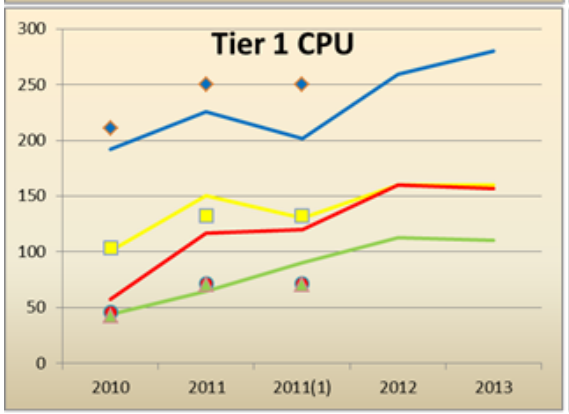


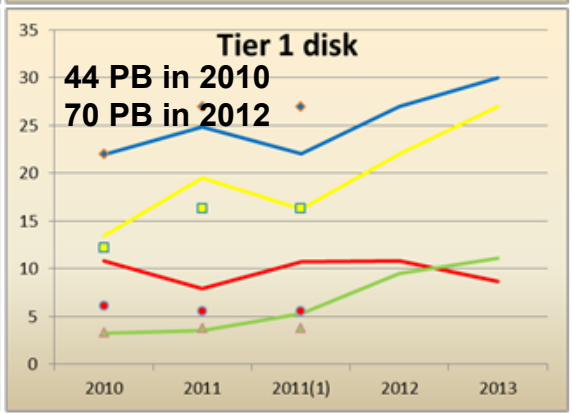
Evolution of the Worldwide Distributed LHC Computing Facility

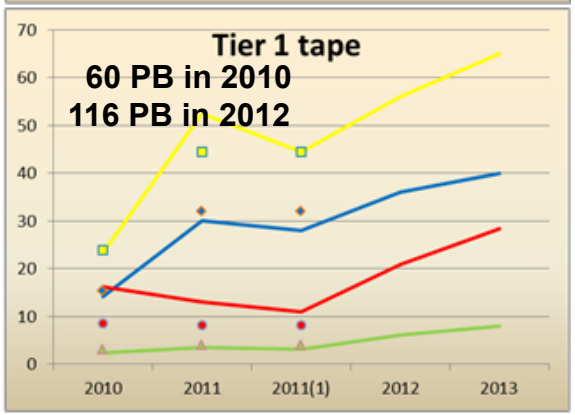


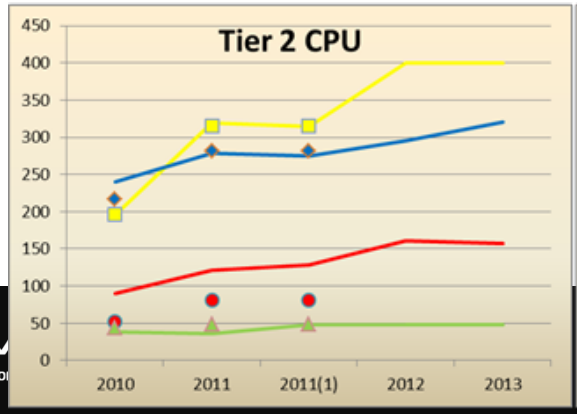


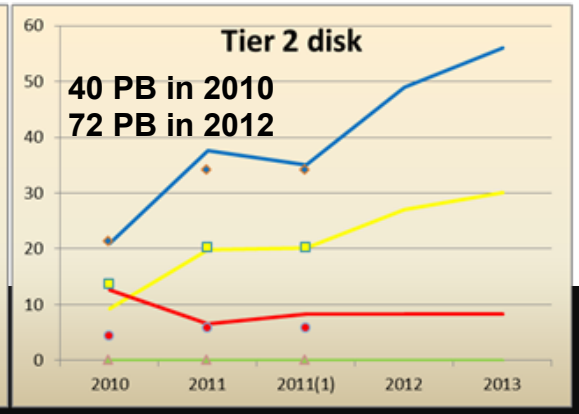


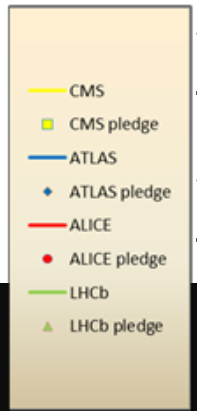












Storage Capacity in 2010
Disk: 99 PB
Tape: 92 PB

Storage Capacity in 2012
Disk: 165 PB
Tape: 166 PB

Tier-1 Facility Architecture and Components

(example BNL)

9 PB Disk

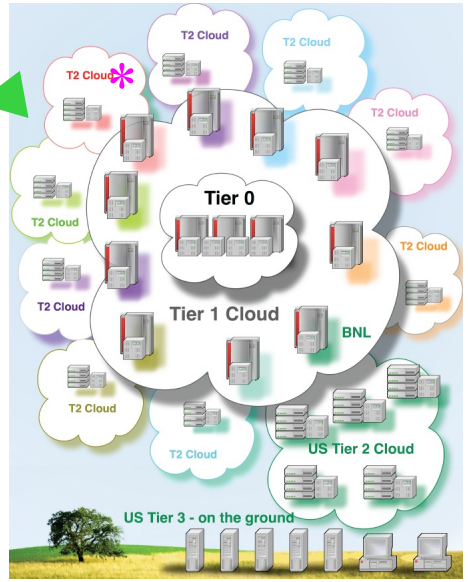
Storage Element (Disk & Tape)

12 PB Tape

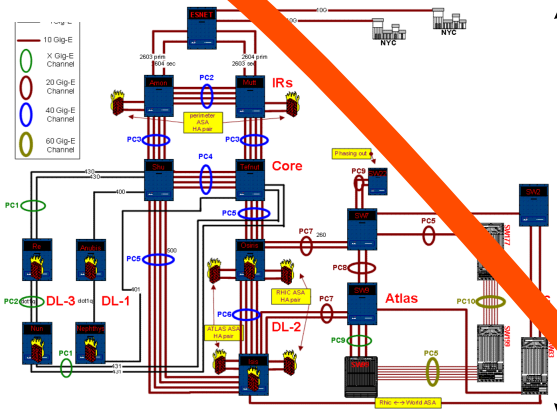
6 Libraries with ~50k slots, 100 Tape Drives, 40 PB (with 2009) capacity

The "Grid"

Based on technology used by WLCG



60 Gbps WAN



160 Gbps LAN

2600 Compute Servers

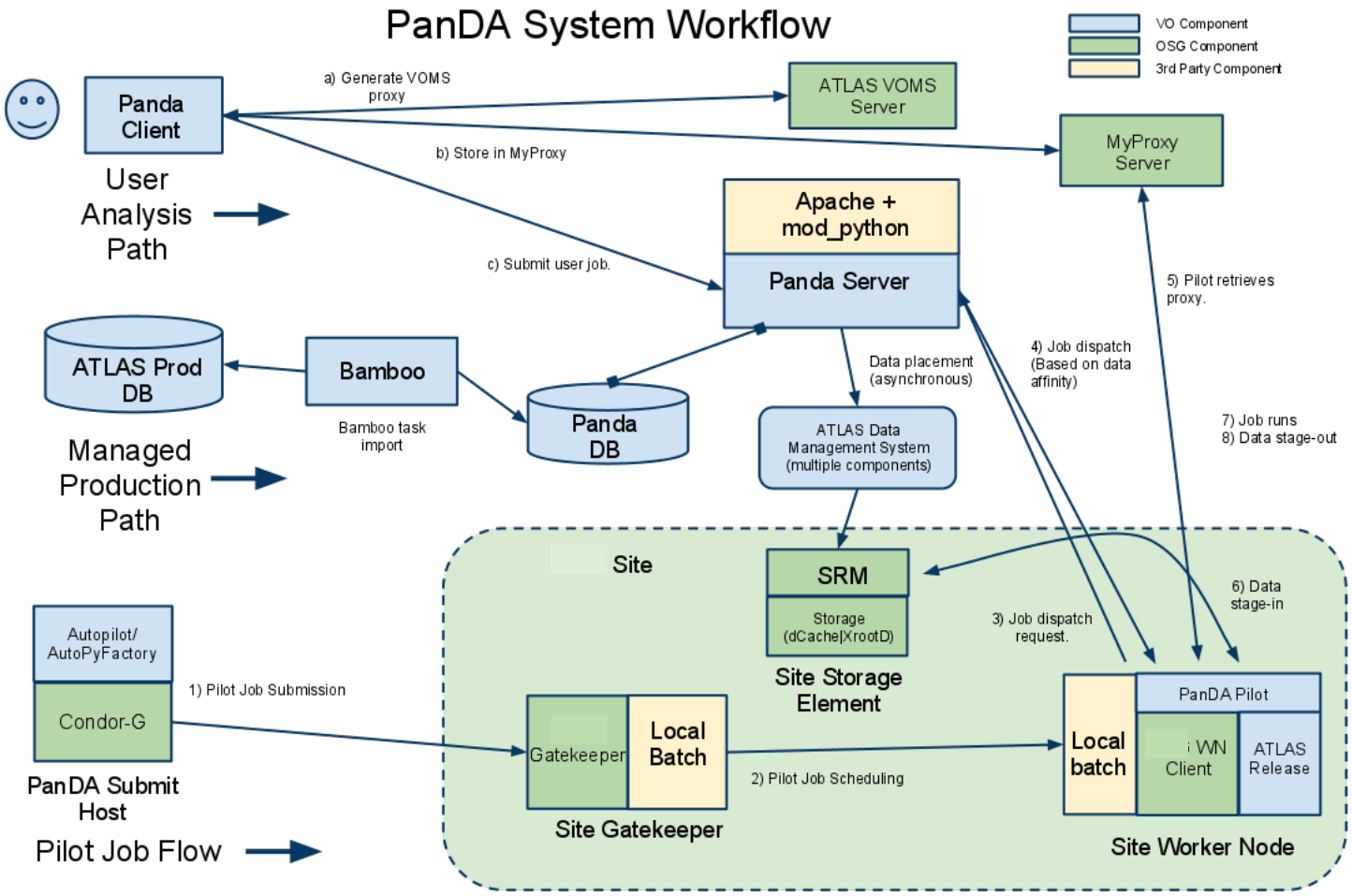
Network

Compute Element

- The term "cloud" is used here in the context of ATLAS regional resources (not cloud computing)

WLCG Worldwide LHC Computing Grid

ATLAS Workload Management



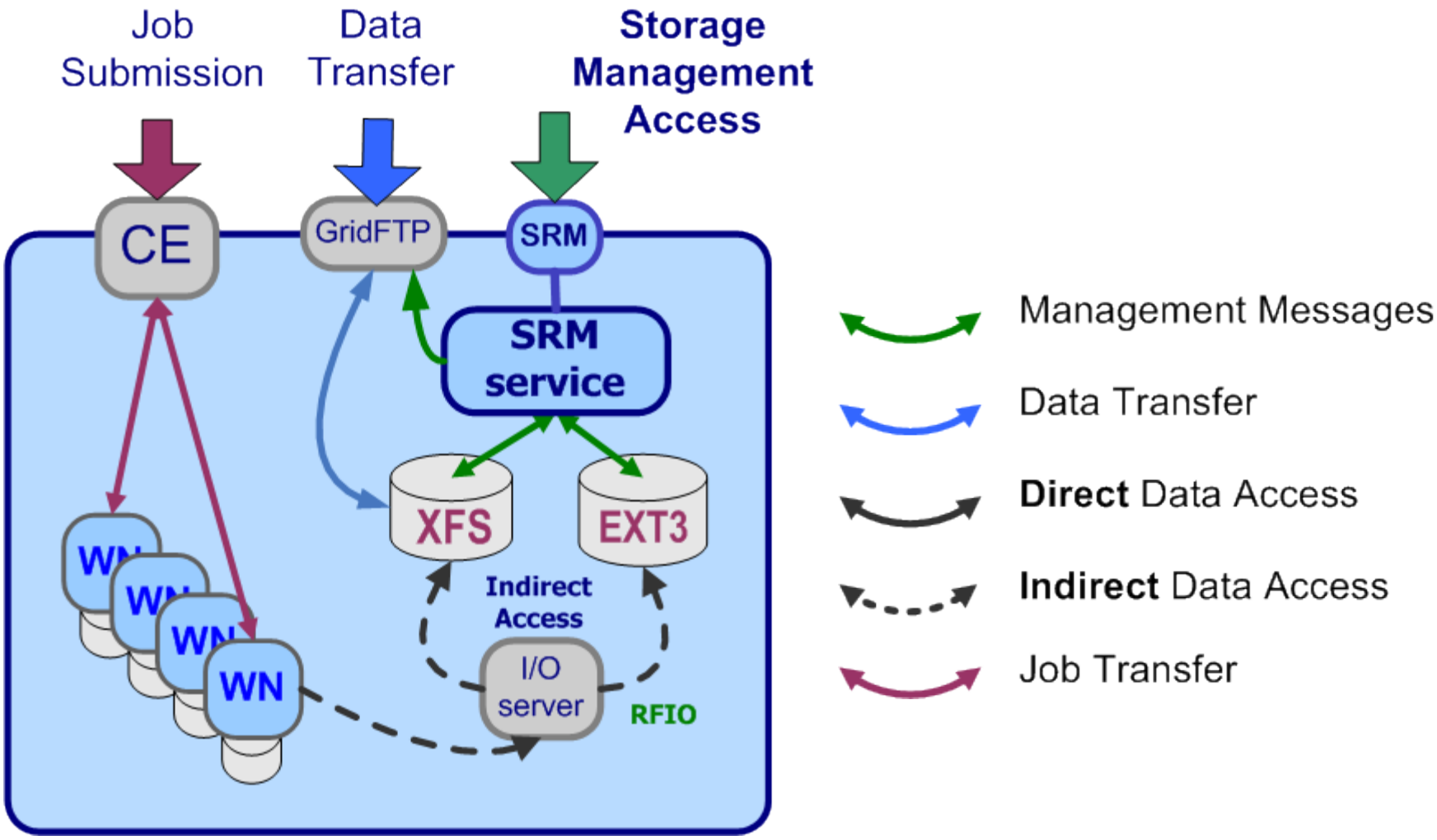
Manages 300,000 - 500,000 production and analysis jobs per Day (up to 2M CPU hours/day)

SRM – A common storage management interface for storage systems

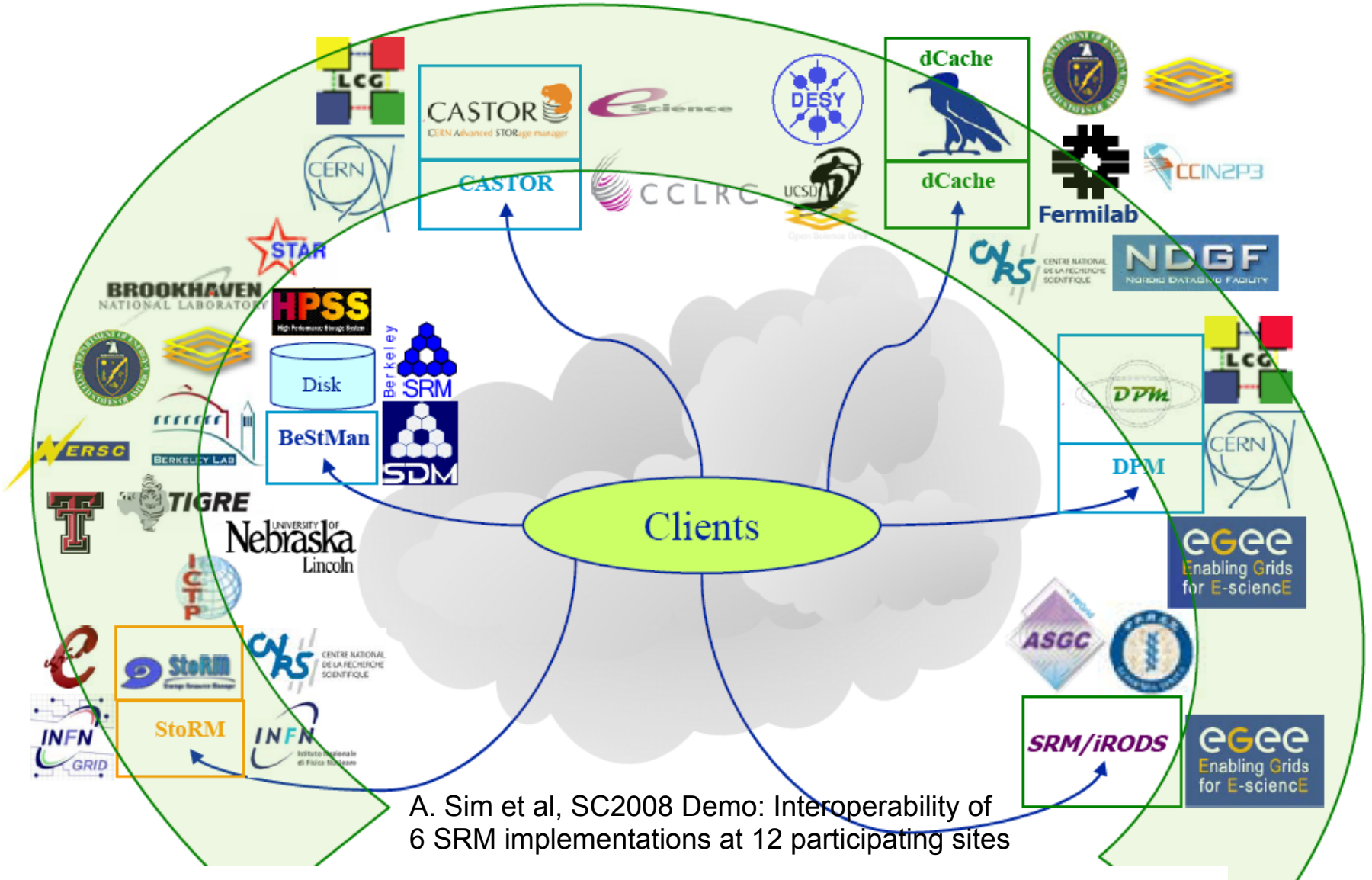
- *SRM approach is to have uniform interface specifications allowing multiple implementations to interoperate. This became crucial to the interoperation of storage systems for the experiments that have to manage and distribute massive amounts of data efficiently and securely.*

- *Data transfer functions* to get files into SRM spaces from the client's local system or from other remote storage systems, and to retrieve them from MSS (e.g. tape)
  - srmPrepareToGet, srmPrepareToPut, srmBringOnline, srmCopy
- *Space management functions* to reserve, release, and manage spaces, their types and lifetimes.
  - srmReserveSpace, srmReleaseSpace, srmUpdateSpace, srmGetSpaceTokens
- *Lifetime management functions* to manage lifetimes of space and files.
  - srmReleaseFiles, srmPutDone, srmExtendFileLifeTime
- *Directory management functions* to create/remove directories, rename files, remove files and retrieve file information.
  - srmMkdir, srmRmdir, srmMv, srmRm, srmLs
- *Request management functions* to query status of requests and manage requests
  - srmStatusOf{Get,Put,Copy,BringOnline}Request, srmGetRequestSummary, srmGetRequestTokens, srmAbortRequest, srmAbortFiles, srmSuspendRequest, srmResumeRequest
- Other functions include Discovery and Permission functions
  - srmPing, srmGetTransferProtocols, srmCheckPermission, srmSetPermission, etc.

Role of SRM at a (StoRM) Site

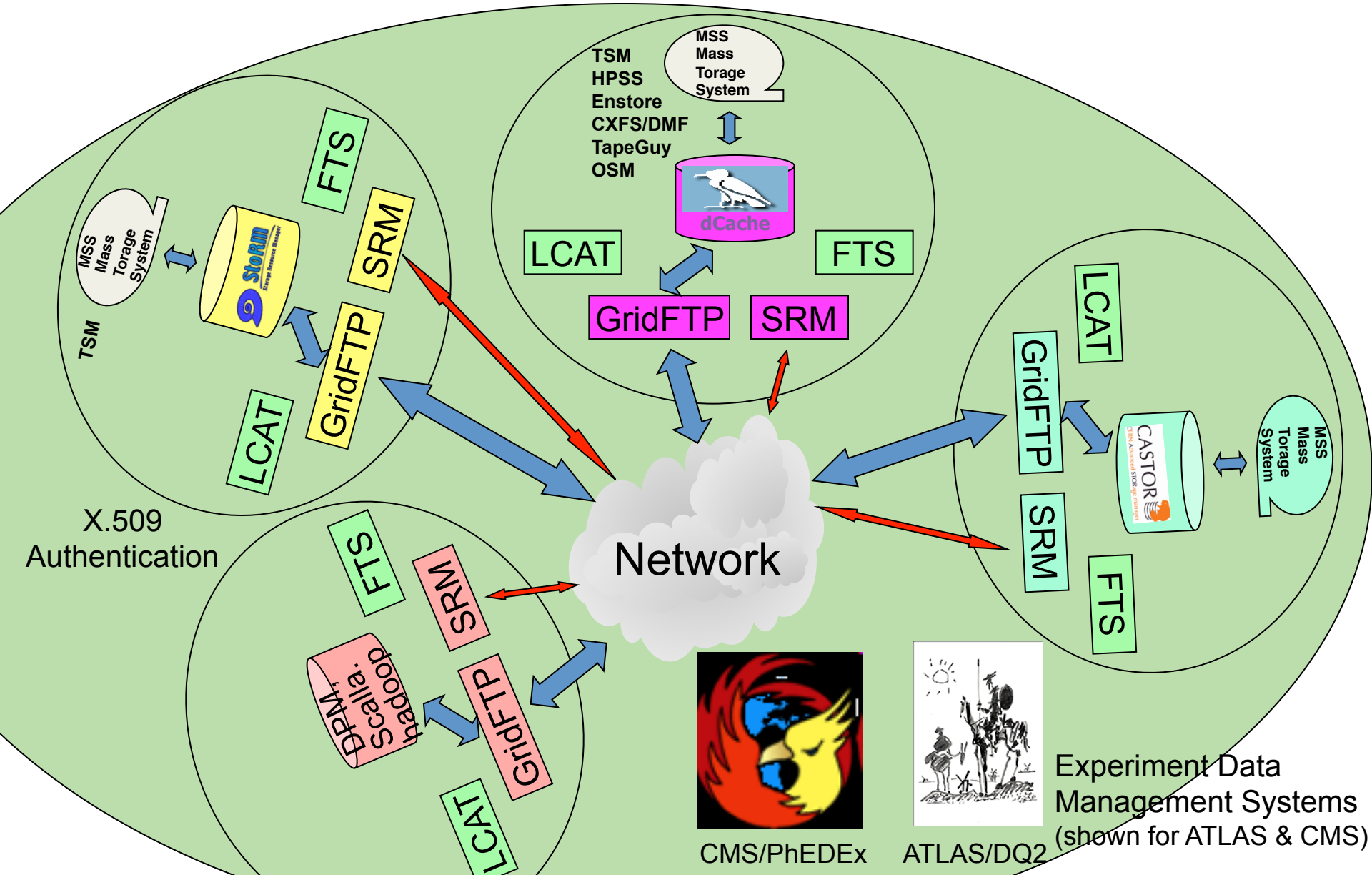


Interoperability in SRM



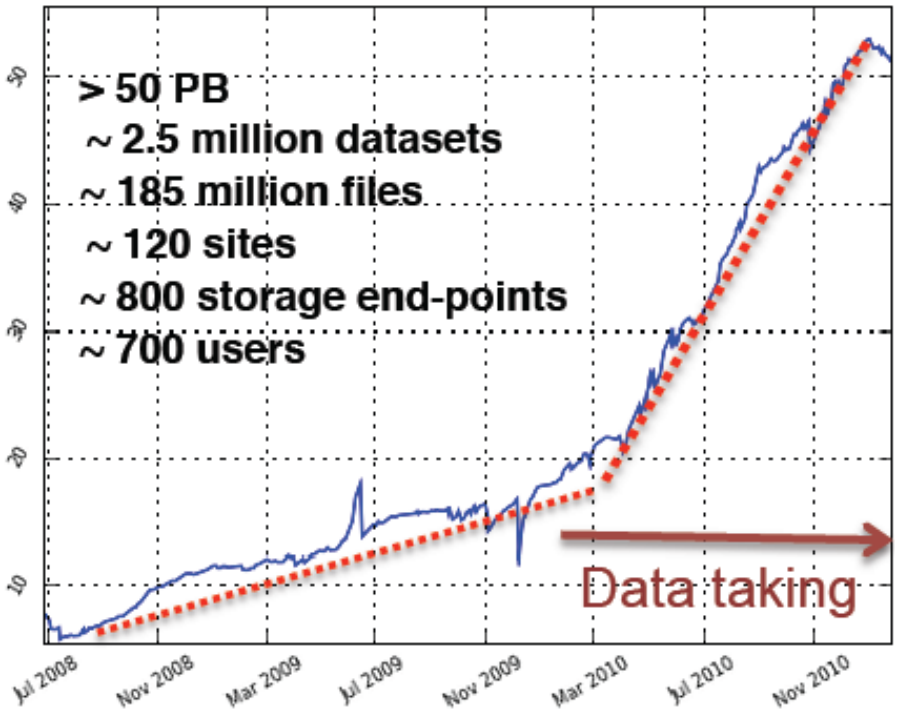
A. Sim et al, SC2008 Demo: Interoperability of 6 SRM implementations at 12 participating sites

Global Data Access & Data Management

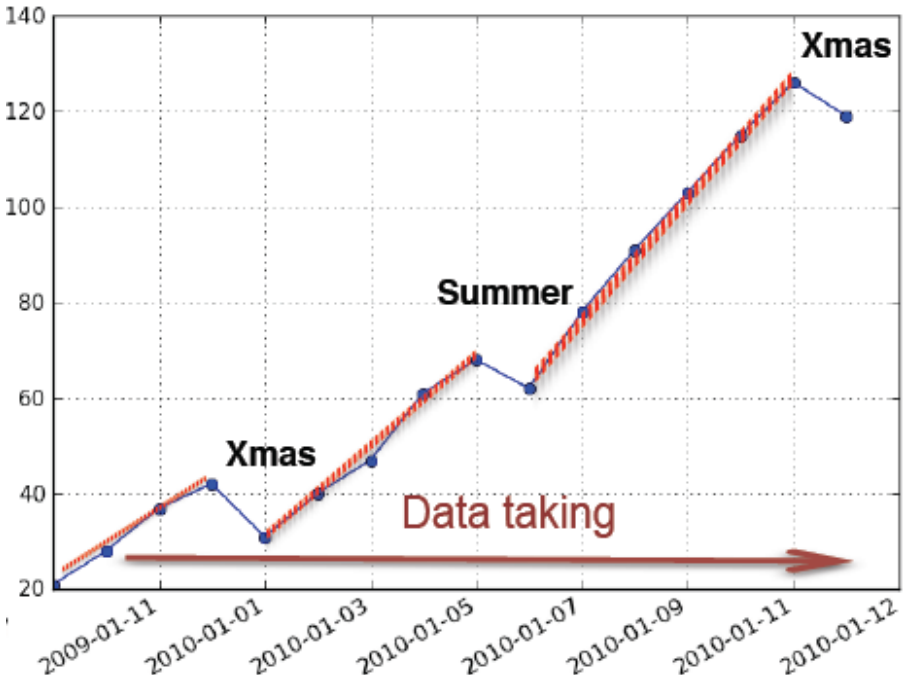


Scale of ATLAS Data in 2010

7PB Primary Dataset (1.6PB RAW Data), n copies worldwide + Simulated Data

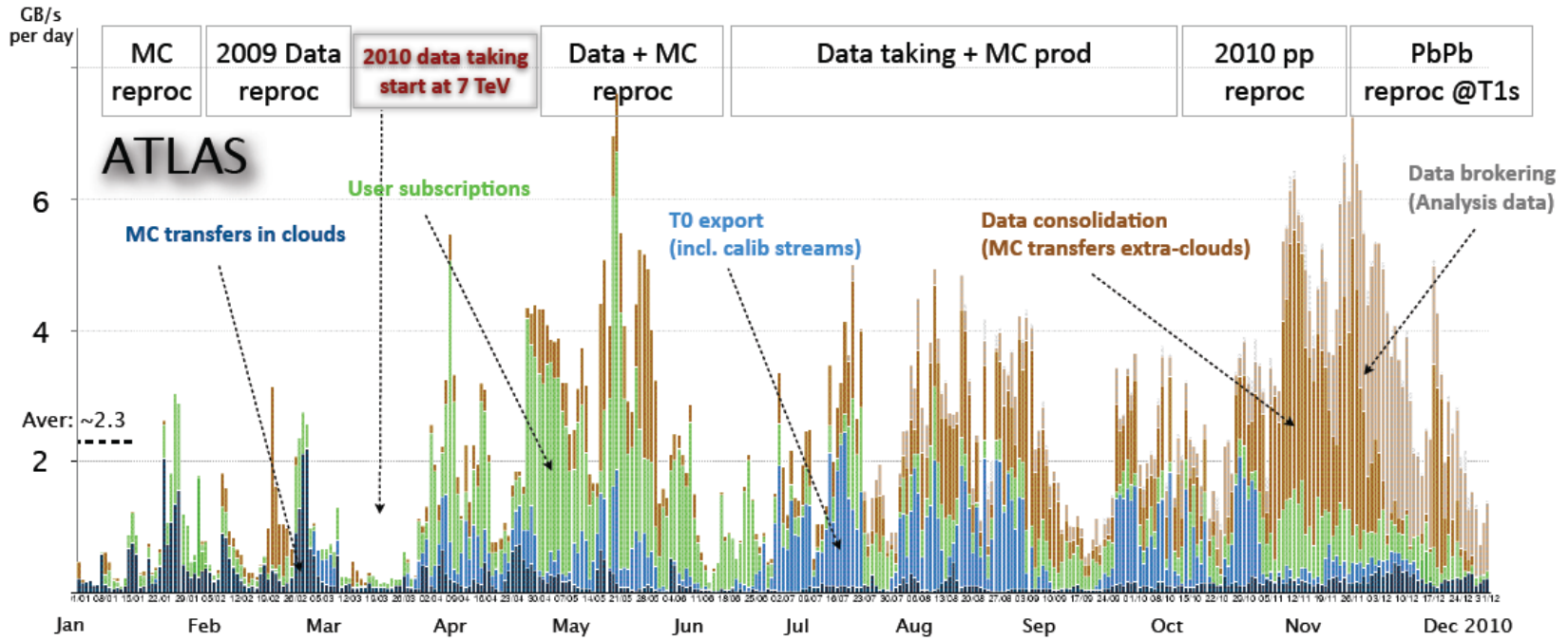


Evolution of Total Space (PB)



Monthly dataset access rate (M)

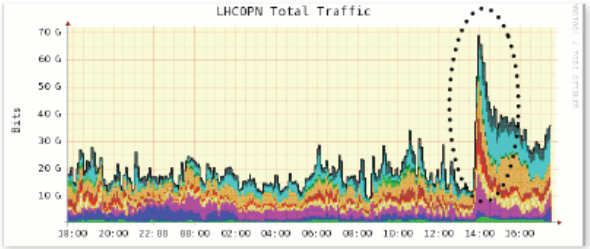
Data Movement



Transfers on all routes (among all Tier levels)

- Average: **~2.3 GB/s** (daily average)
- Peak: **~7 GB/s** (daily average)

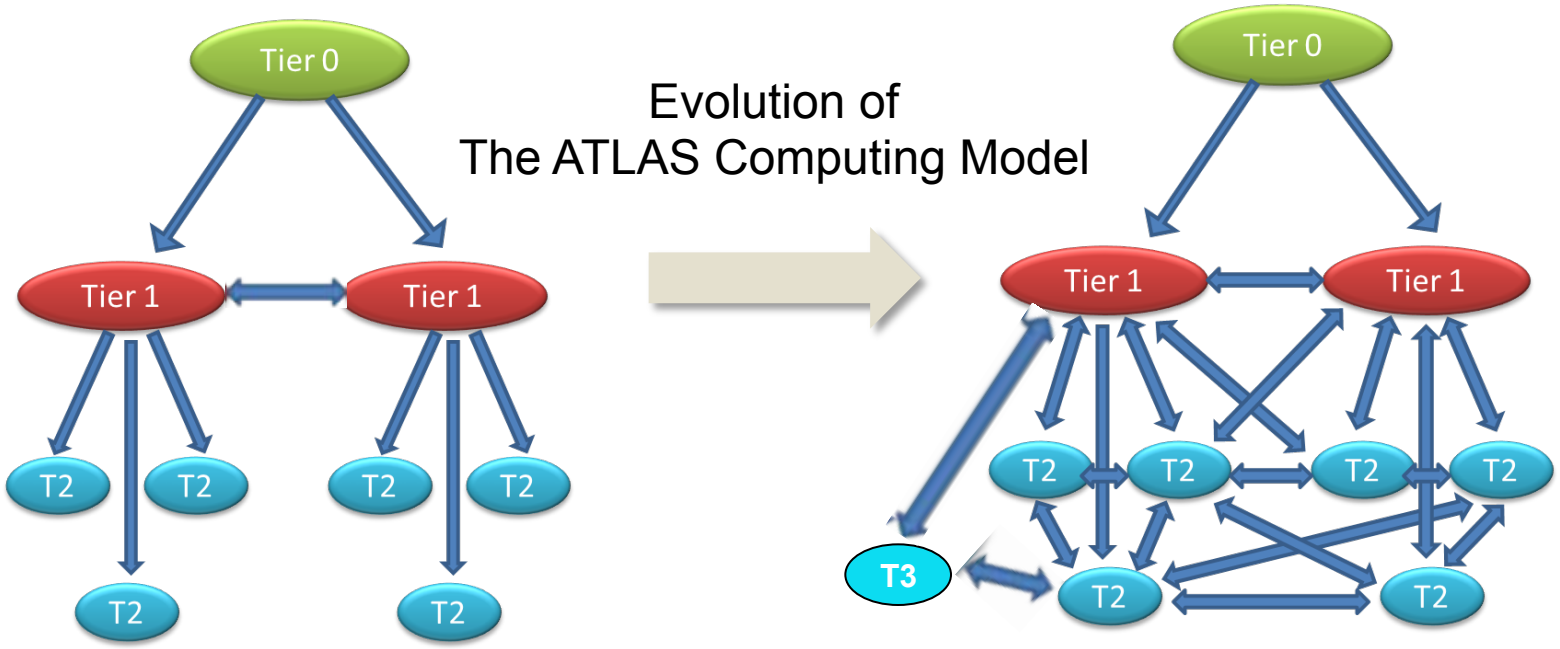
Data available on-site after few hrs.



Traffic on OPN measured up to 70 Gbps

- ATLAS massive reprocessing campaigns

Network Evolution in response to changes to the ATLAS CM

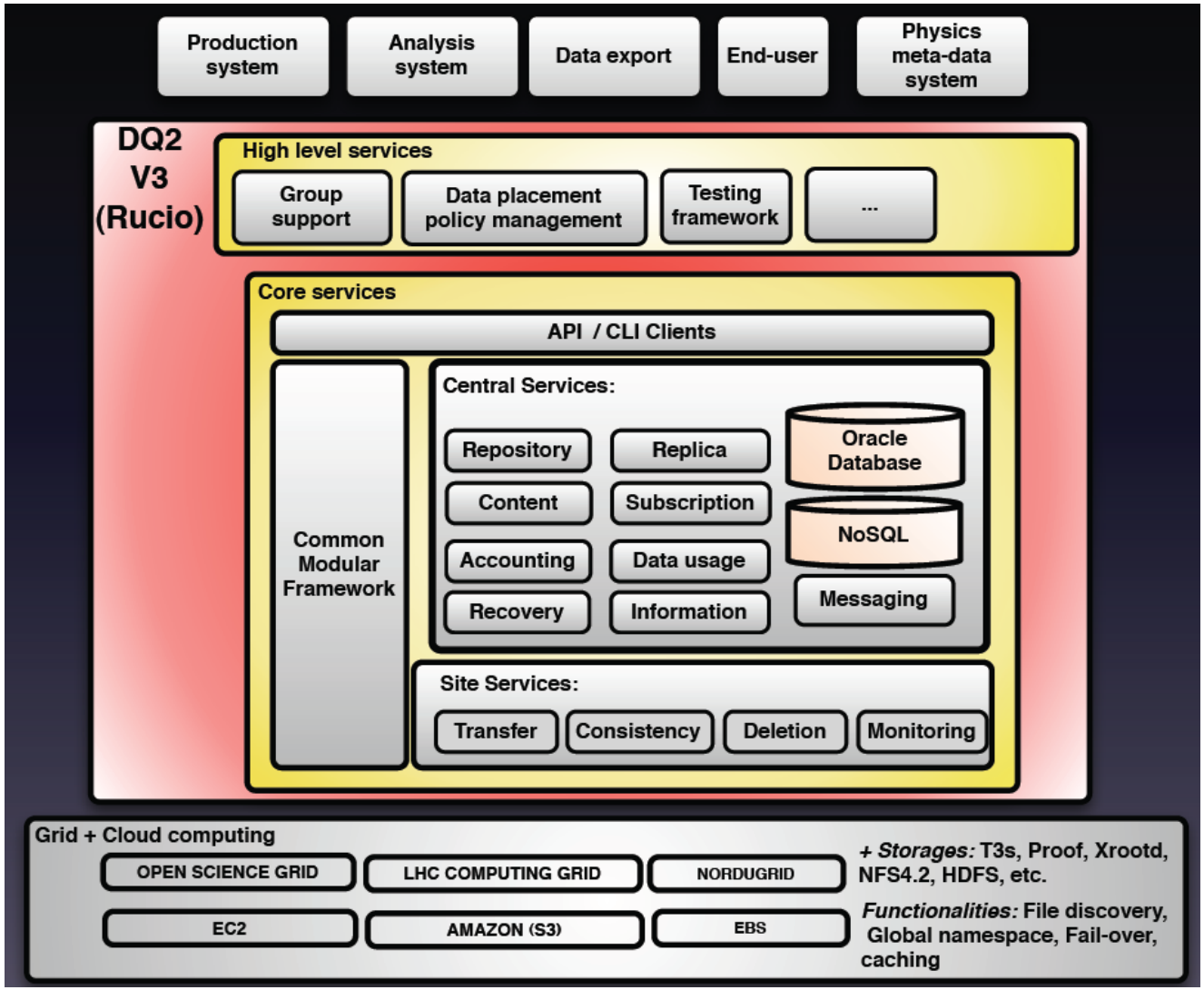


On-demand will augment/replace massive data pre-placement ⟹ Network usage will be more dynamic and less predictable

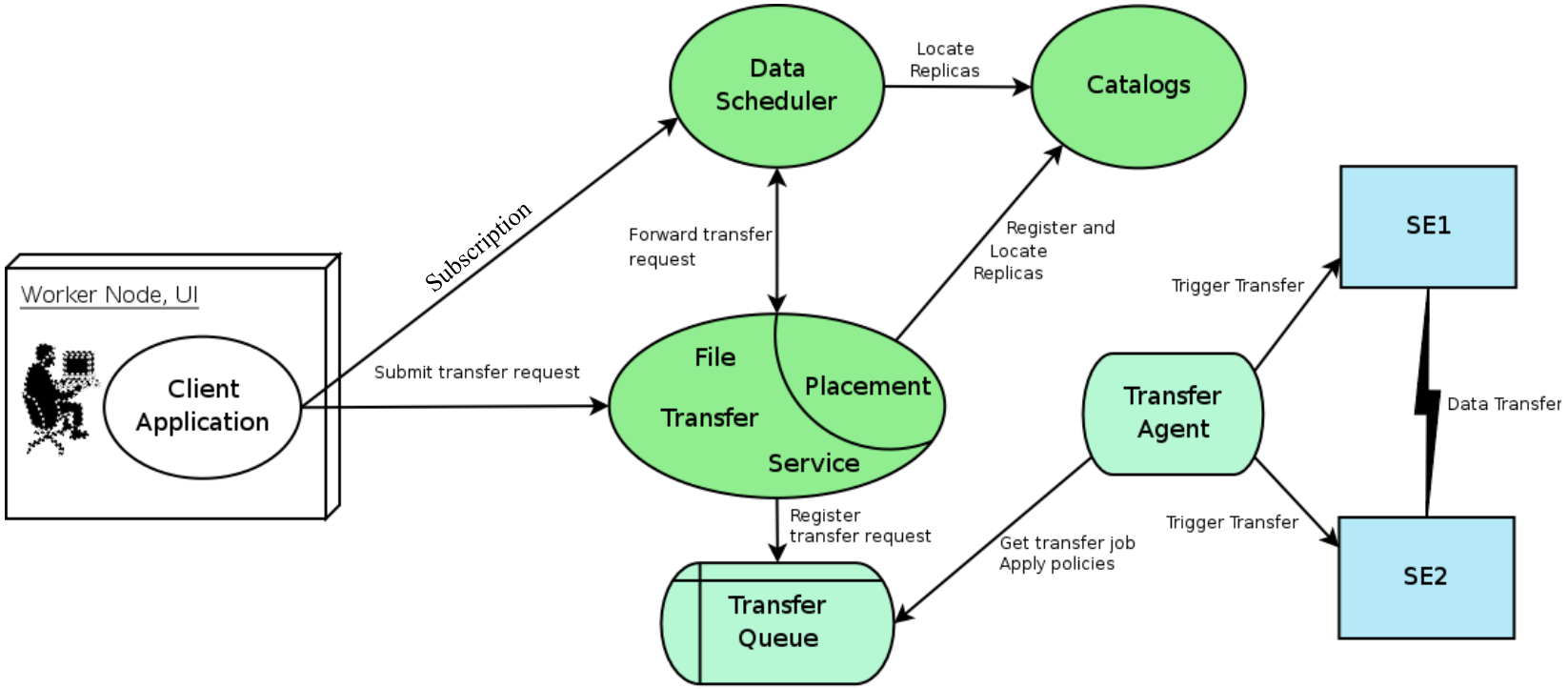
need to enable high-volume data transport between any T1s, T2s, and T3s.

Global Data Management is Key

(Example ATLAS Distributed Data Management)



Managing Transfers with FTS

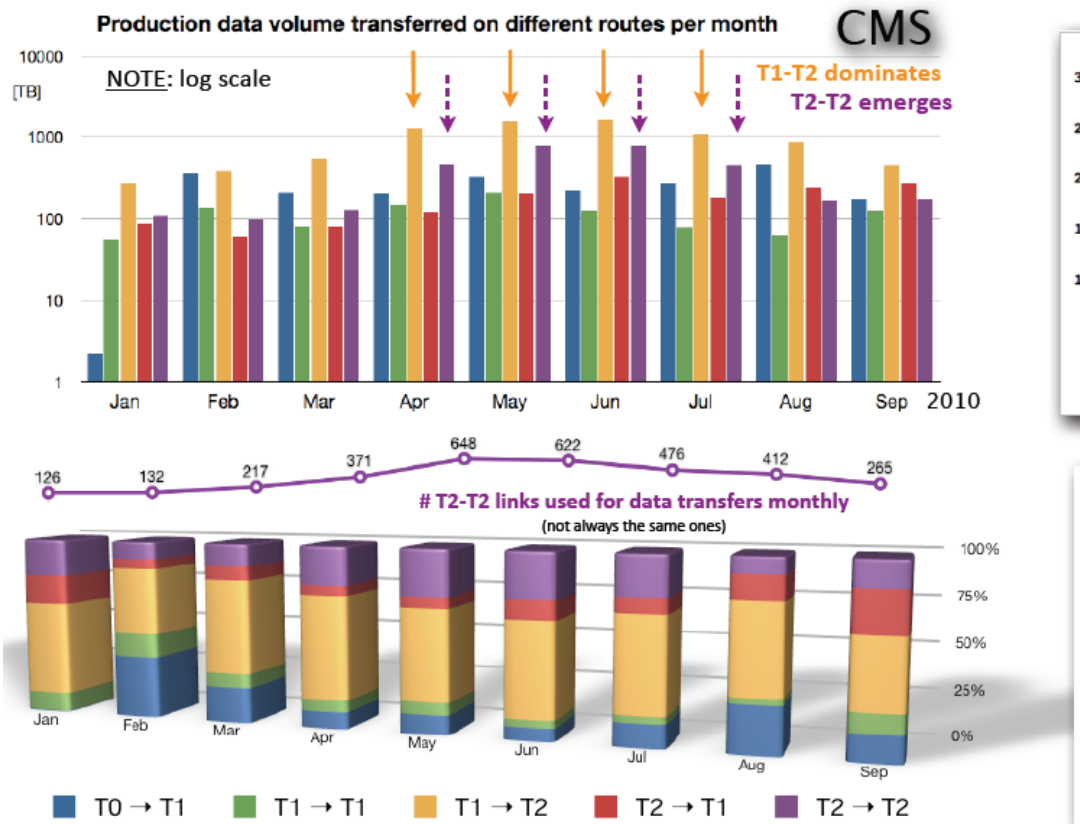


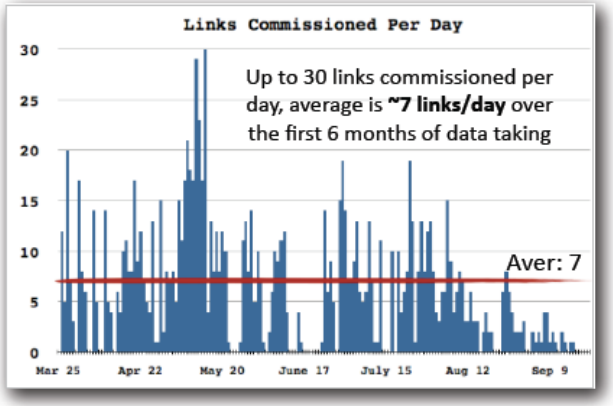
- Transfer “Channels” (Source/Destination pair) over existing Network
  - Transfers only between sites with predefined “channel”
- Transfer Requests are queued & initiated according to Channel configuration
  - TCP Buffer size, # parallel transfers, timeouts
  - Static, no feedback loop between FTS and SE regarding storage system load, available storage space etc
  - FTS throttles transfers to protect SEs from overload and guarantee VO bandwidth shares
    - Caveat: Multiple independent FTS instances not communicating can create load problems

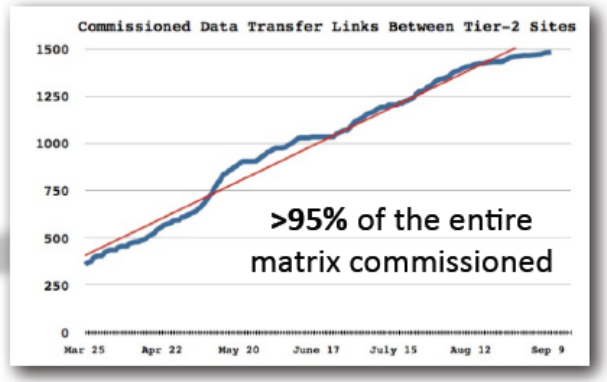
Data Placement for Analysis

Once the data is available on the Grid it must be made accessible to analysis applications

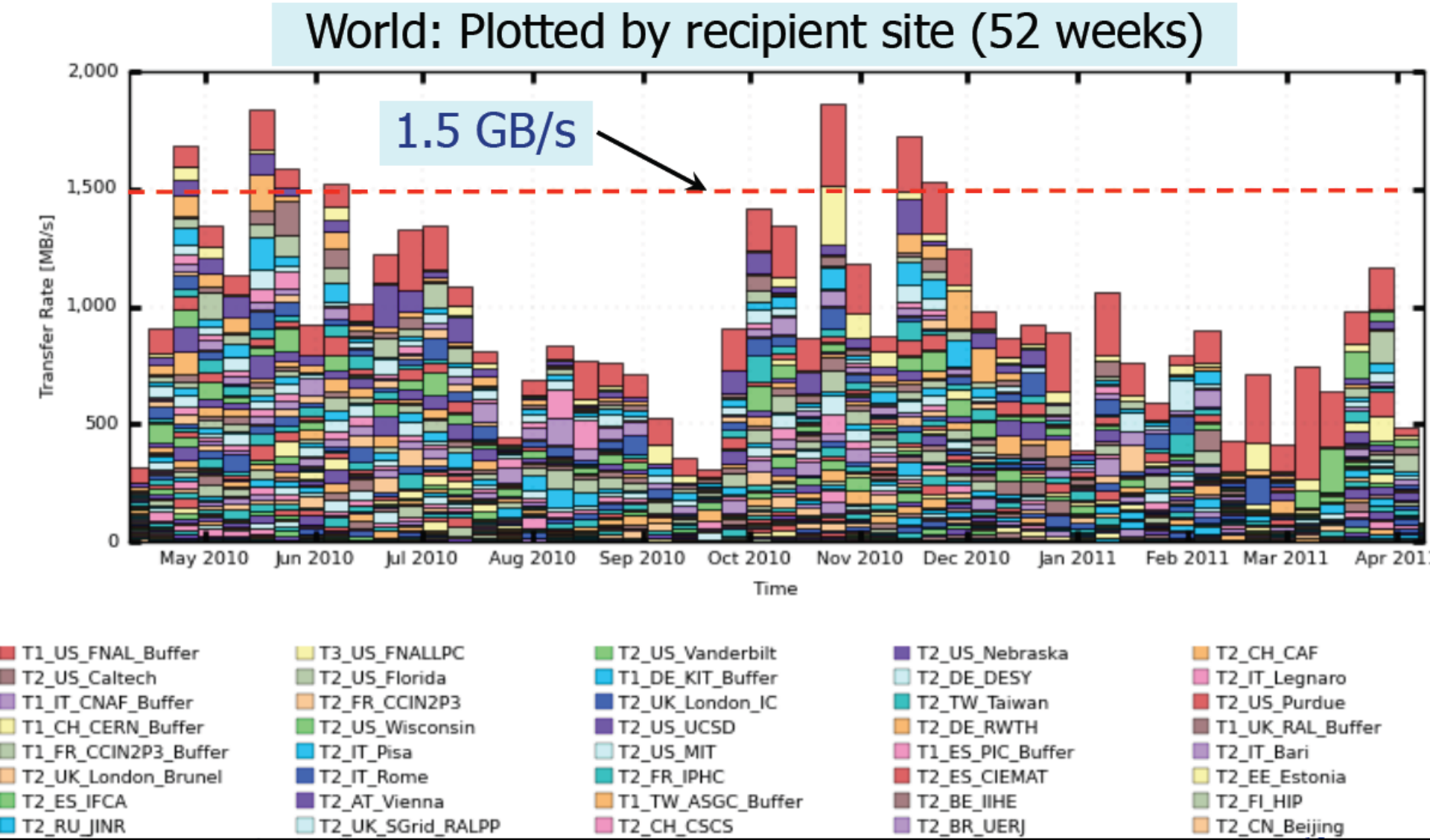
- Largest fraction of analysis computing at LHC is at the T2 level
- Flexibility of the transfer model help to reduce the latency seen by the analysis end-users



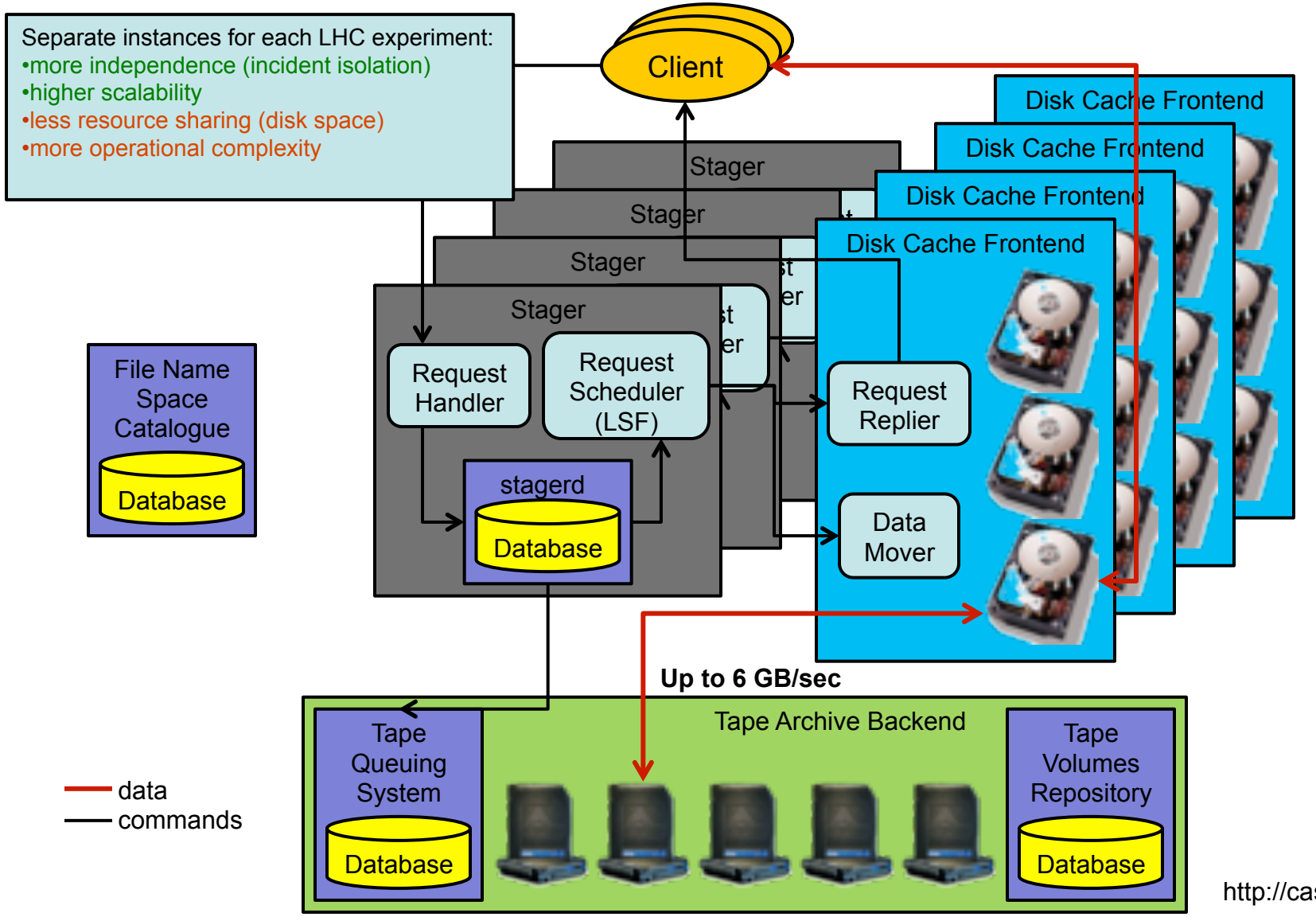




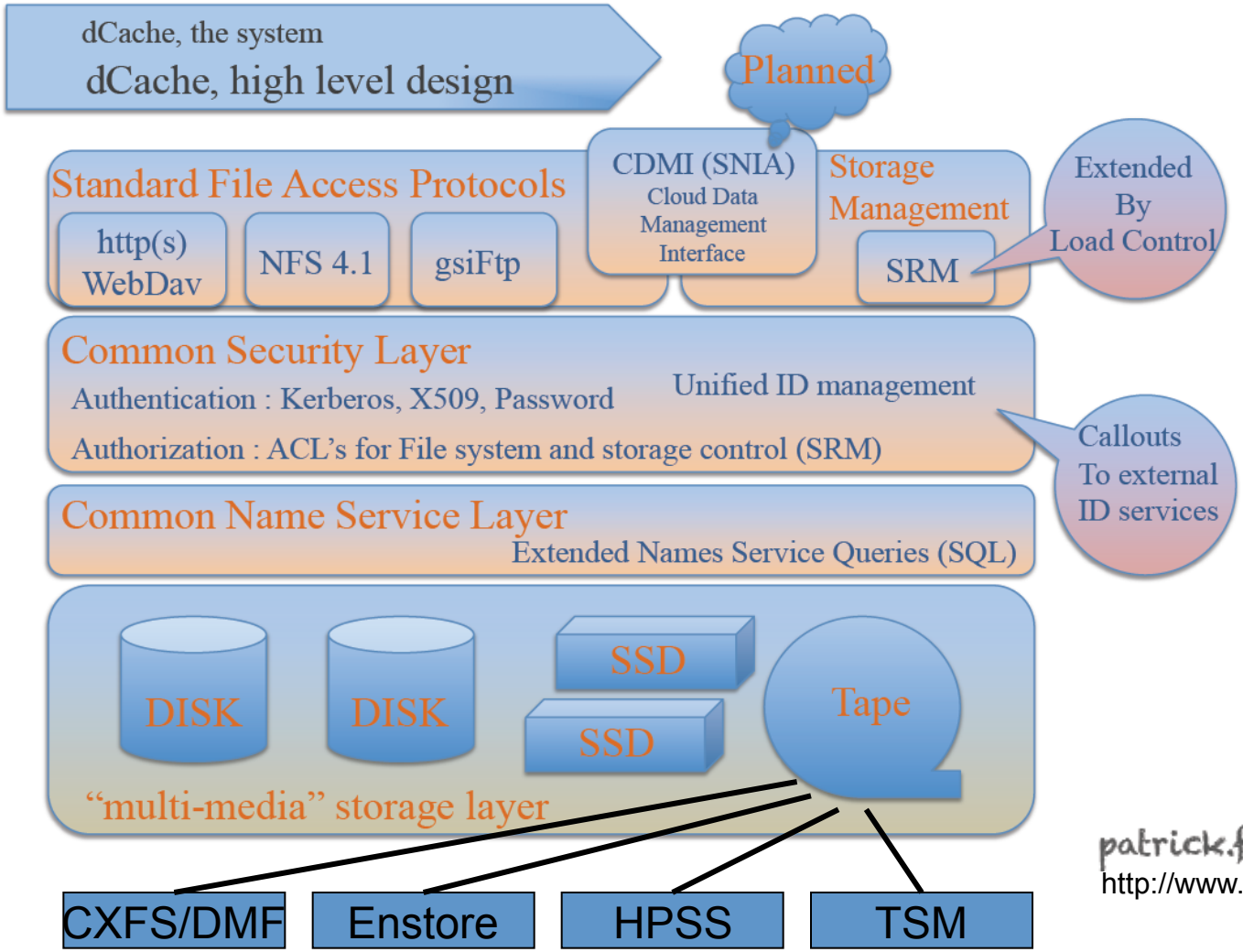
Weekly CMS PhEDEx Data Rates (World)



CASTOR Architecture (simplified)



http://castorweb.cern.ch



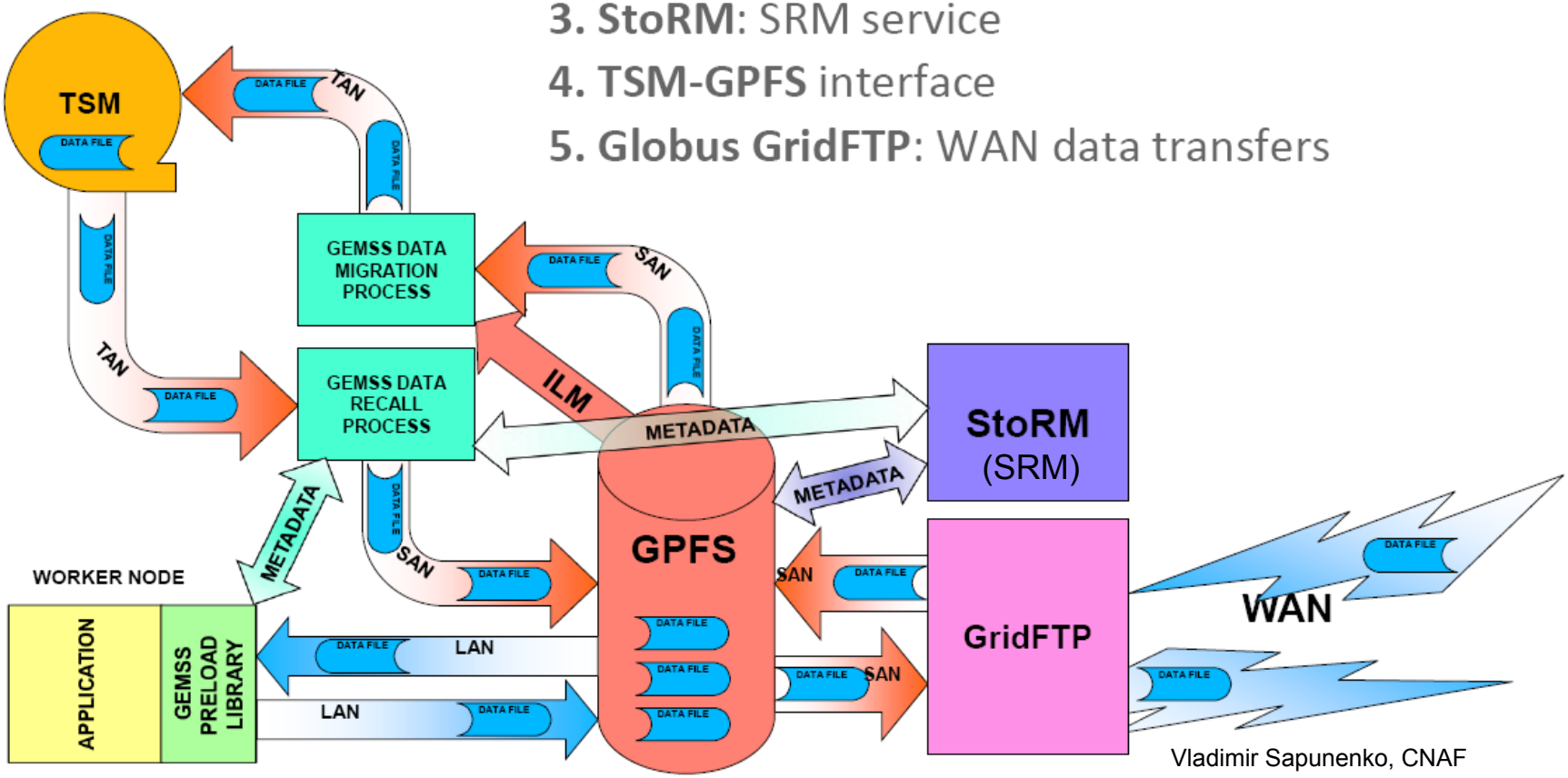
- Used by 8 WLCG Tier-1 and ~40 Tier-2 Centers
- Manages ~50% of WLCG data

patrick.fuhrmann @ dCache.ORG
http://www.dcache.org

GEMSS at CNAF

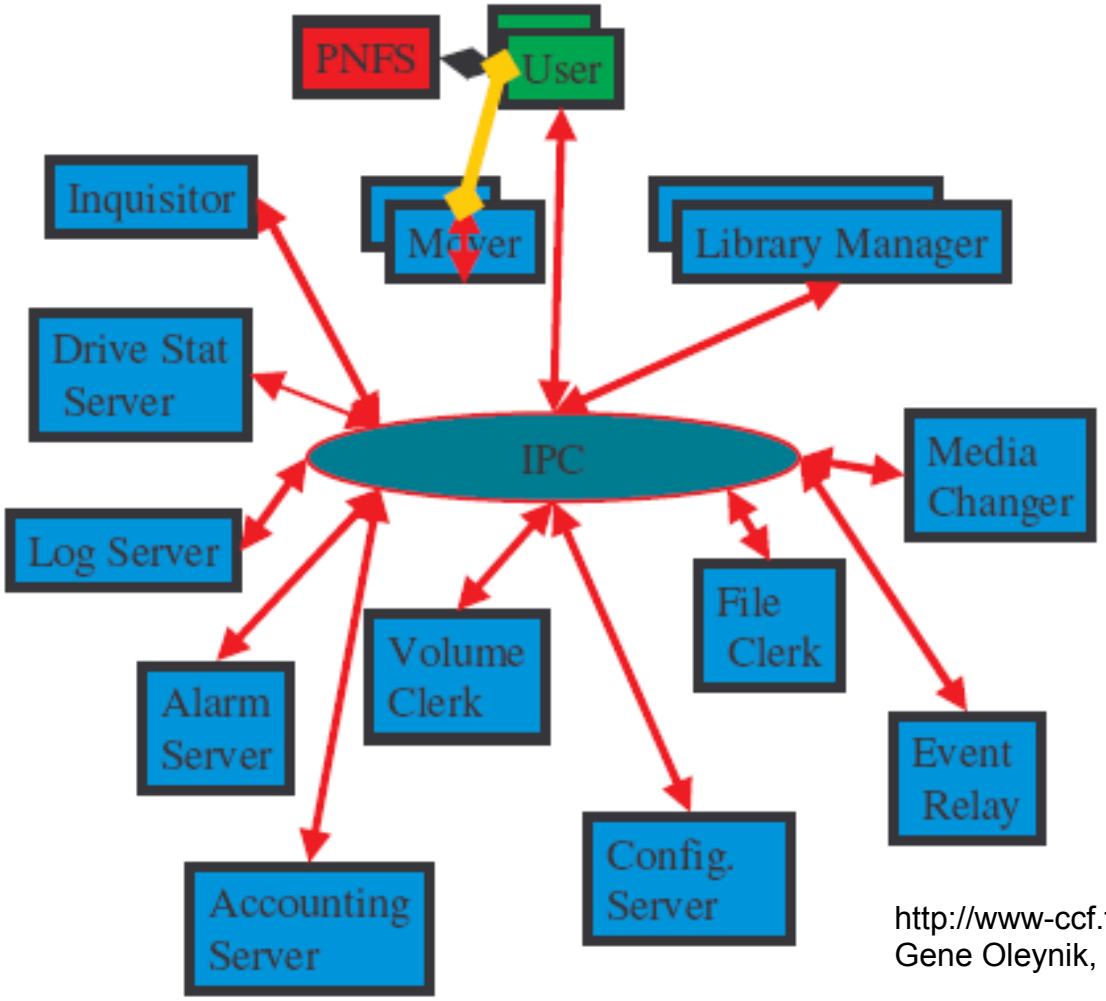
Disk-centric system with five building blocks

1. **GPFS**: disk-storage software infrastructure
2. **TSM**: tape management system
3. **StoRM**: SRM service
4. **TSM-GPFS** interface
5. **Globus GridFTP**: WAN data transfers



Vladimir Sapunenko, CNAF

Enstore Mass Storage System



http://www-ccf.fnal.gov/enstore/
Gene Oleynik, FNAL

WLCG Tier-1 Site Overview

Site	Disk [TB]	Storage Mgmt	Tape [TB]	MSS	Tape library	Tape Drive
ASGC (TW)	2500	DPM	3600	Castor	N/A	N/A
BNL (US)	7300	dCache	4700	HPSS	SL8500	LTO4, LTO5
CERN	17500	Castor	23100	Castor	SL8500,TS3500	TS1130, T10k
CNAF (IT)	6400	STorM/GPFS	1600	TSM	SL8500	T10K
FNAL (US)	9000	dCache	7000	Enstore	SL8500	LTO3, LTO4
CC-IN2P3 (FR)	5100	dCache	5300	HPSS	SL8500	T10K
KIT (DE)	7800	dCache	5000	TSM	SL8500,TS3500	LTO3, LTO4
NDGF (DK,FI,NO,SE)	2800	dCache	1800	N/A	N/A	N/A
PIC (ES)	3500	dCache	1500	Enstore	SL8500,TS3500	LTO4, LTO5
RAL (UK)	6000	Castor	3000	Castor	SL8500	T10K
SARA (NL)	2800	dCache	1700	CXFS/DMF	SL8500	T10k
TRIUMF (CA)	2100	dCache	700	TapeGuy	TS3584	LTO4
Total	72800	N/A	59000	N/A	N/A	N/A

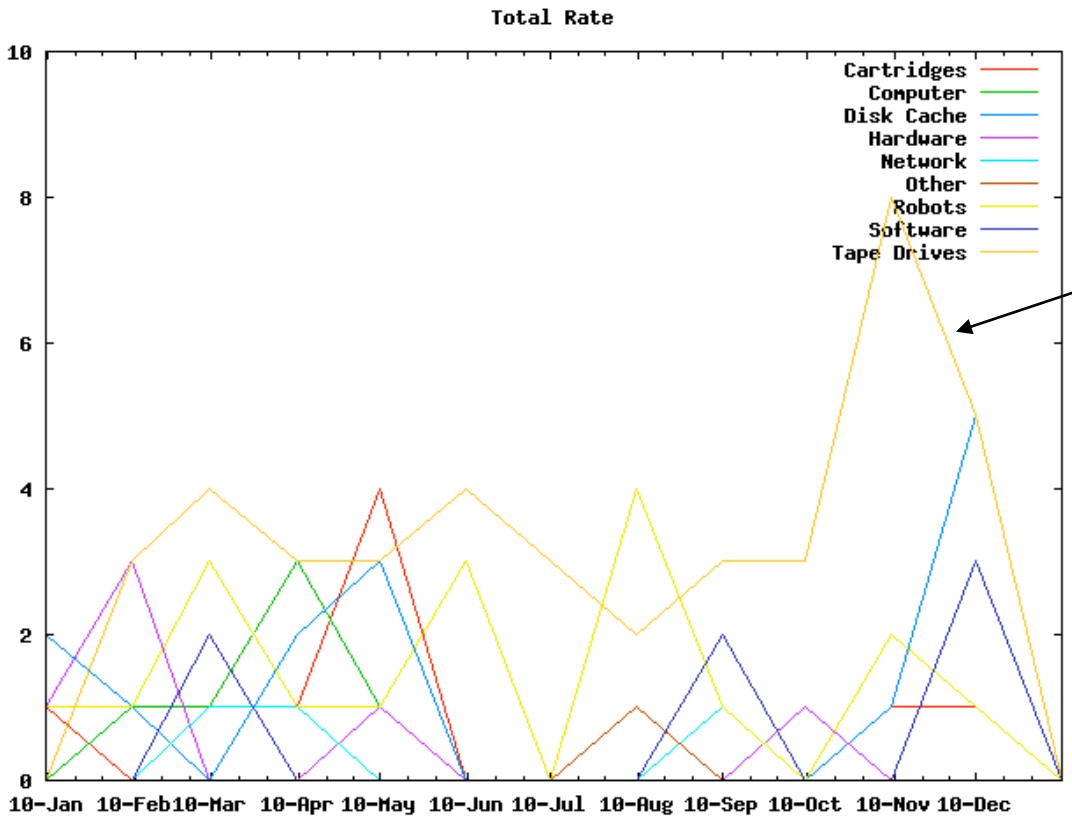
- Hardware reliability
  - Disk
    - Lifetime of 3 - 4 years
    - Most sites have chosen RAID6 for performance and resilience
    - Sites are losing disks at a rate of 5-10/10,000 drives per month – failure/replacement mostly transparent to operations
    - ~30% of Tier-1 sites are checking data integrity on disk on a regular basis (in addition to compute checksum on every xfer)
    - Where applicable, all but one Tier-1 sites are using FC (4 & 8 Gb) as interconnect between disk backend and head nodes
  - Tape
    - Flexible Lifetime with drive and library component replacement on demand
    - Data loss varies from “never lost a file/cartridge” to “one cartridge a month”
    - Several sites are using redundant arms & grippers to improve library availability

Role of Tape in LHC Computing

Assumption in early phase of LCG Project that there would be no Tape by the time LHC data taking starts, but …

- Technical Evolution of Tape Technology leading to unprecedented capacity growth and reduced cost
  - Native capacity of tape cartridge surpassed capacity of biggest disk drive reducing price/GB
    - Expect ~60TB/cartridge by the end of this decade, further improving price/ capacity advantage of tape
  - LTO/LTFS adds an important dimension that could help to improve access times
- Tape drives & media have steadily improved in reliability
  - Less frequent labor-intensive migrations to next gen technology
  - Lower BER and longer useful life than disk making tape better suited for long-term data-retention requirements
  - Cost-effective in terms of operating effort: at BNL ~1 FTE per 5 PB (MSS S/W + tape library and drive H/W)

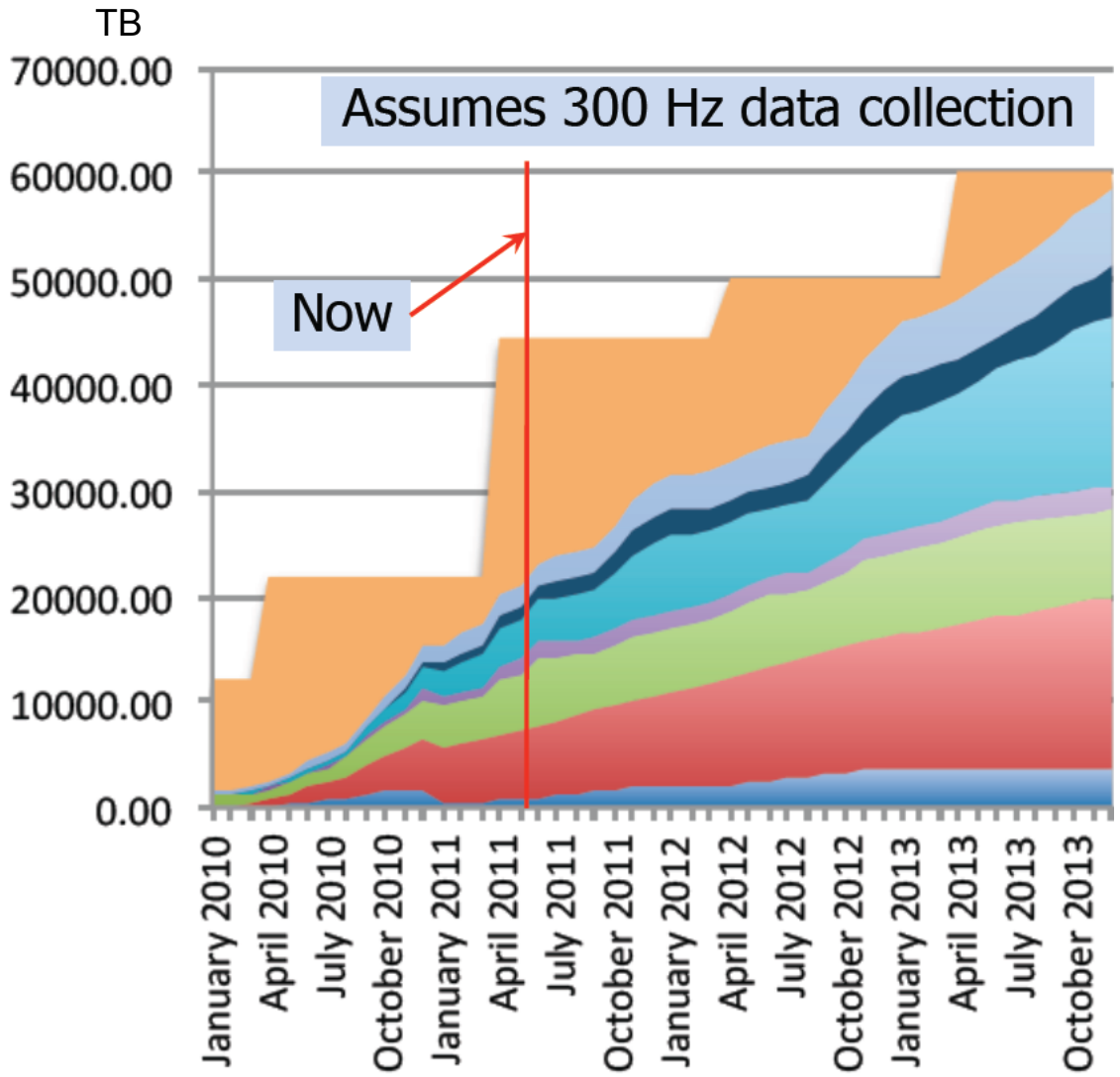
MSS Failure Statistics (Jan – Dec 2010 at BNL)



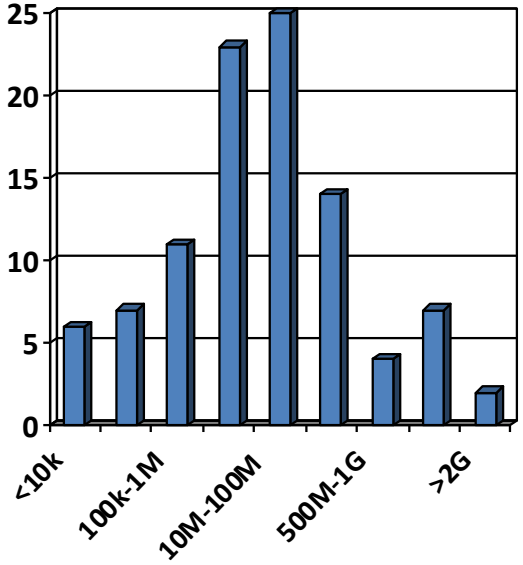
LTO4 uCode problem

- 4 Tape Libraries (~40,000 LTO cartridge slots), 80 LTO4/LTO5 drives
- Includes HPSS disk cache, HPSS core and mover H/W, network etc
- All MSS software components

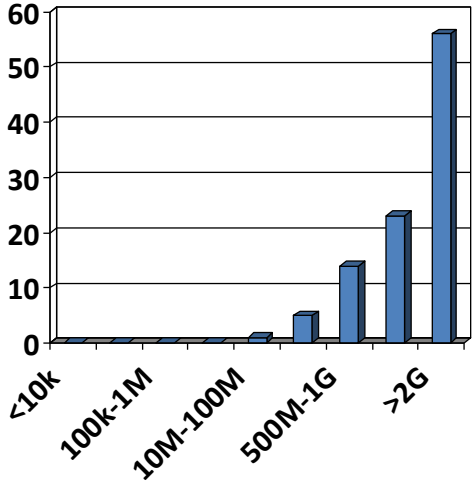
CMS Tier-1 Tape Resources 2010 - 2013



File Size Distribution, Space Occupancy on Tape and Read Performance

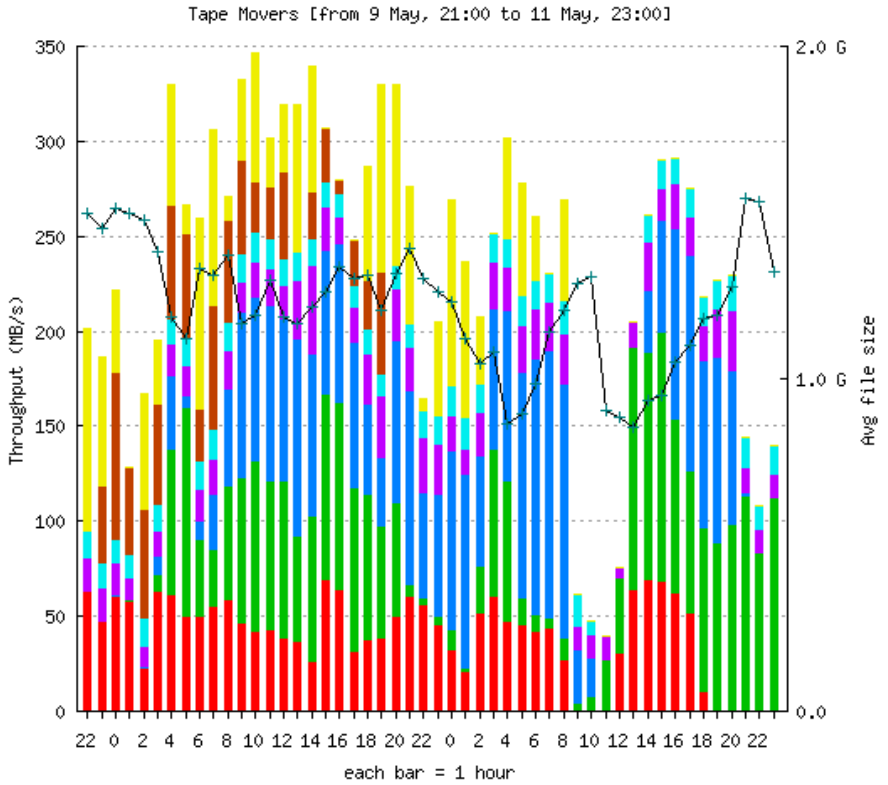


Filesize Distribution (CERN)



Space Distribution (CERN)

Experience at KIT when processing ATLAS data

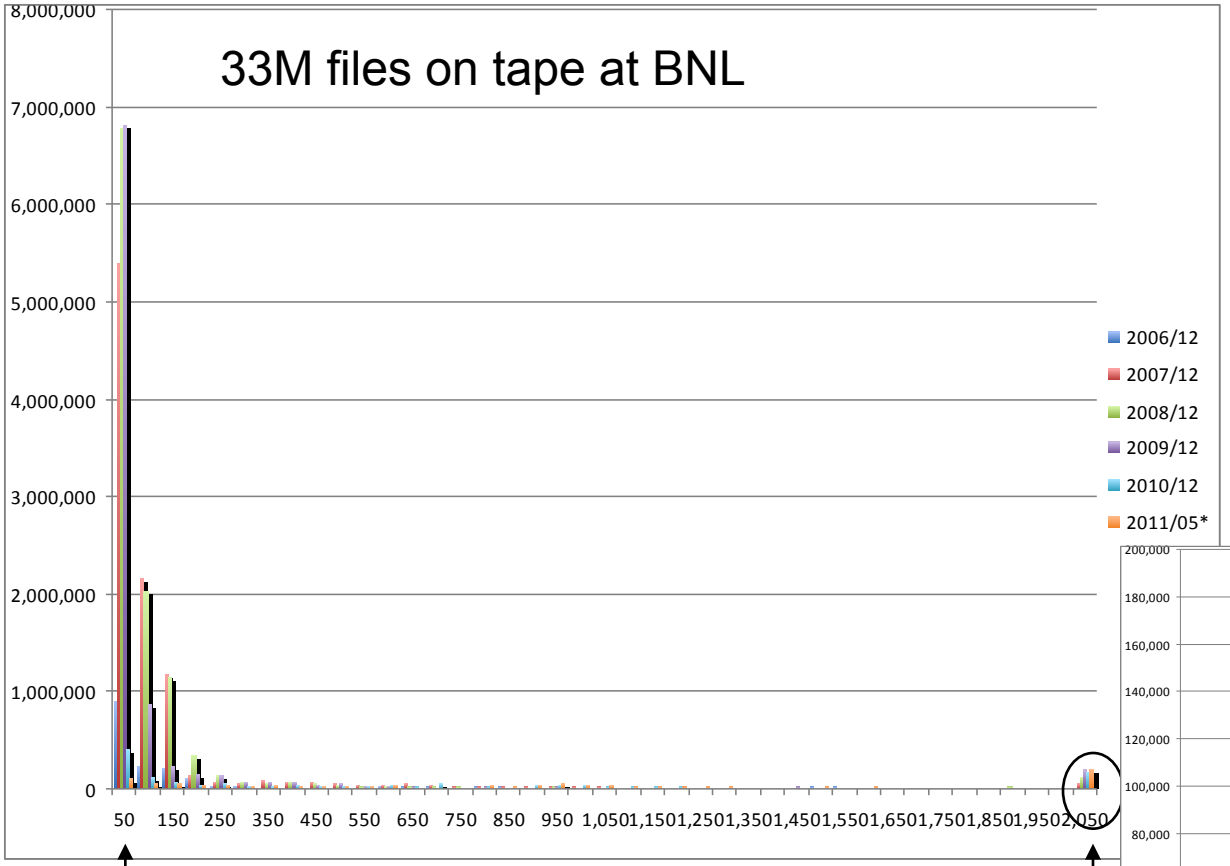


7 LTO4 Drives

Legend: 1 bar/hour, 1 color/drive;
black graph = avg. file size

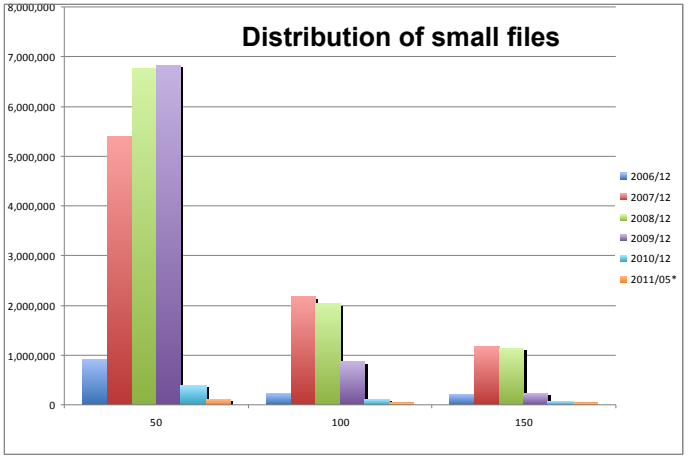
Mover server/disk, tape mounts & positioning limiting transfer rate
• 40-60 MB/s typical transfer rate/drive

File Size Distribution (BNL)

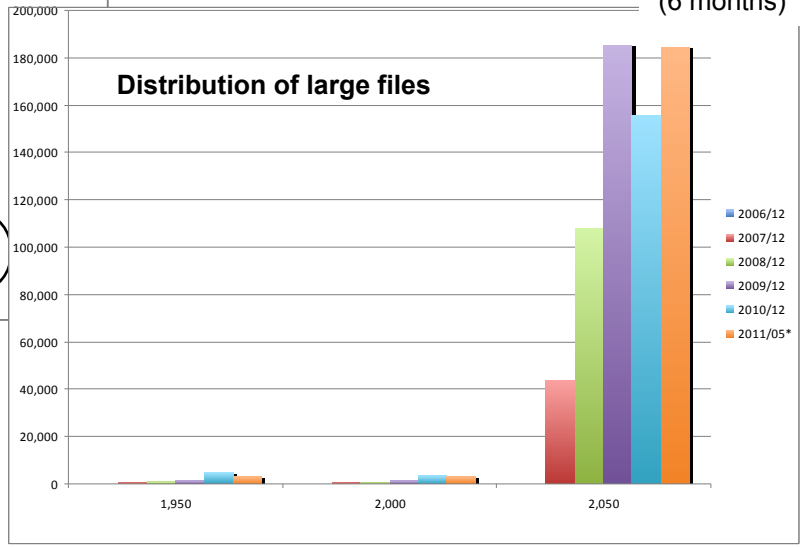


0-50MB

>2GB



2011
(6 months)

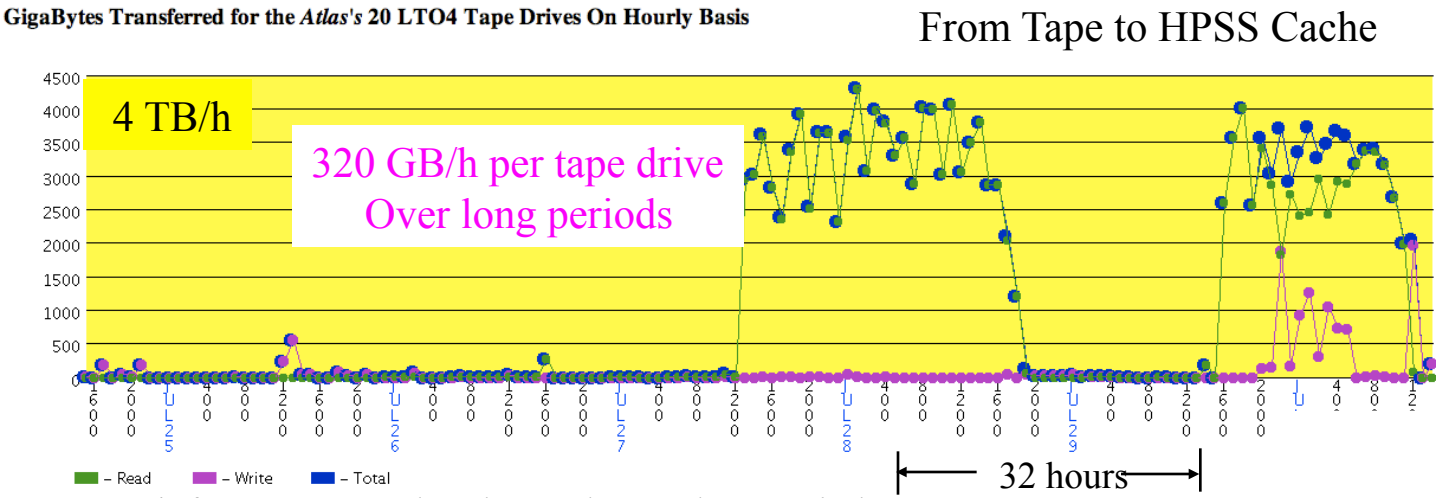


Small File aggregation implemented in HPSS and Castor

- Transparent to application

Experiments are merging small files

Reprocessing (at BNL): Tape Operations – Performance of a vital Storage Component

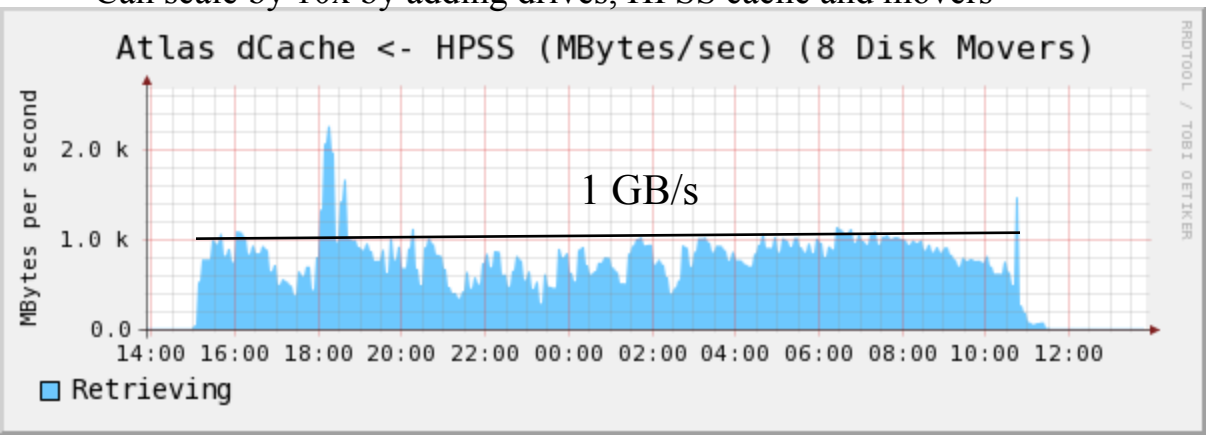


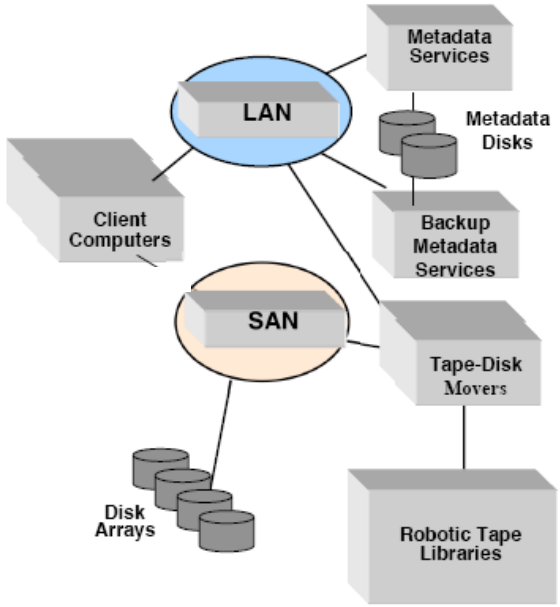
From Tape to HPSS Cache

Average write performance of 60 MB/s per drive
Average read performance of 50 MB/s (incl. mount/position)

Numbers based on avg. filesize of <500MB

1 GB/s from HPSS to dCache Pools over long periods
- Can scale by 10x by adding drives, HPSS cache and movers



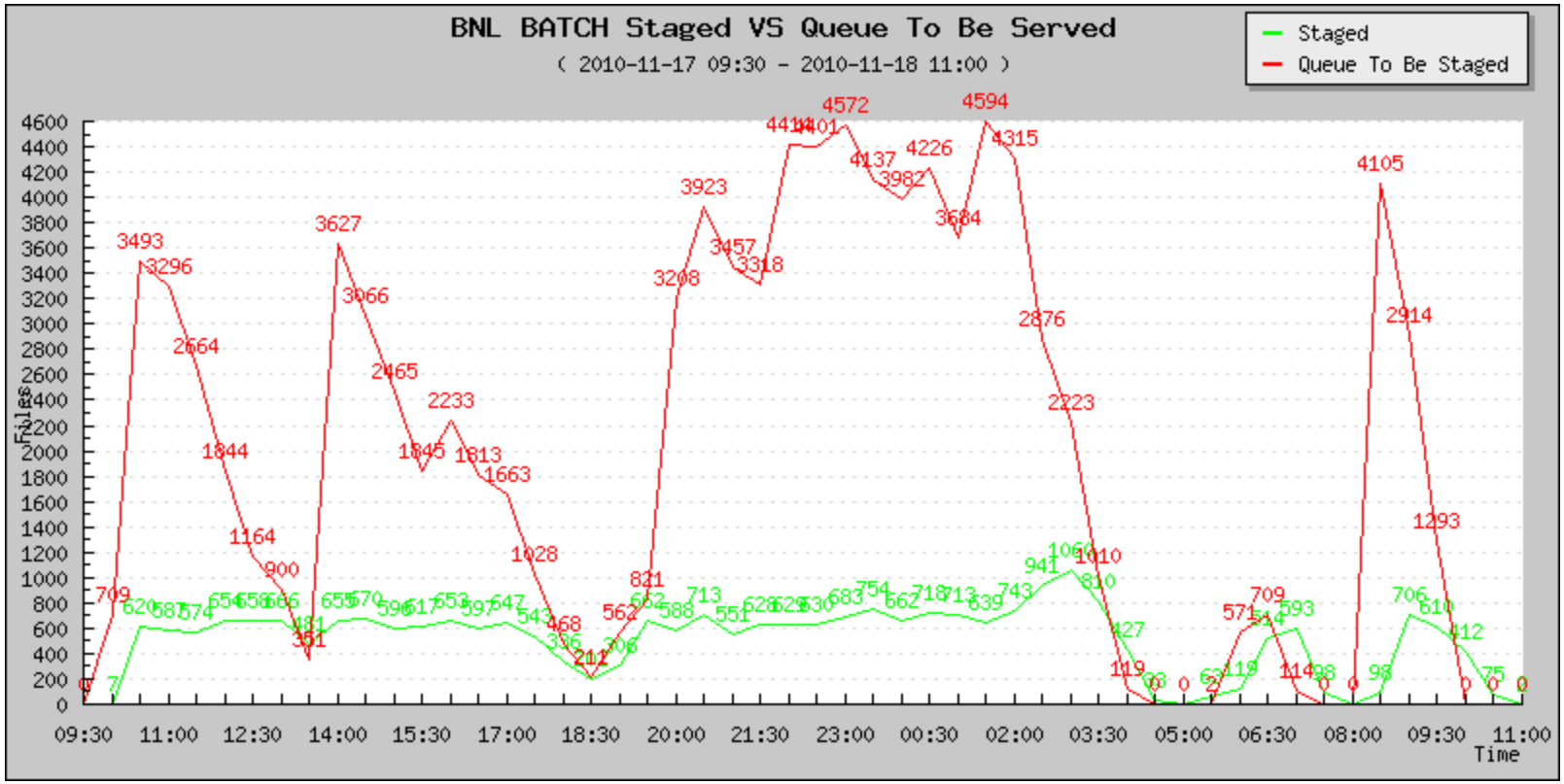


HPSS Cache

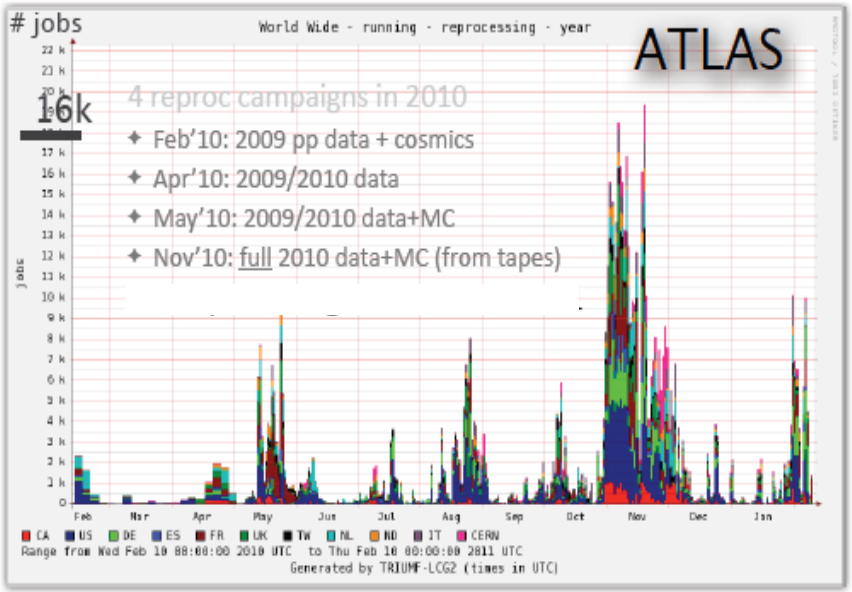
Optimizing Tape Access Performance

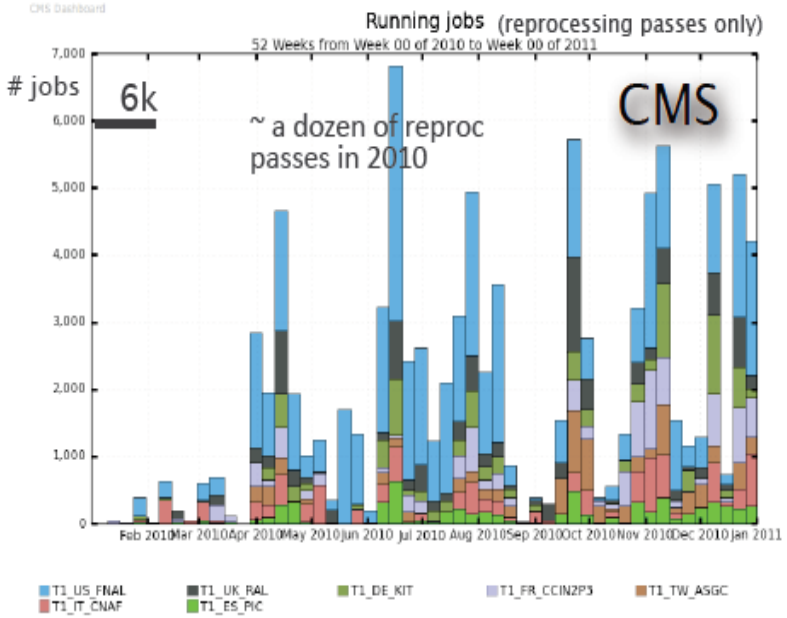
- 1.5 files per tape mount when passing request to system w/o request prioritization and re-ordering
  - Most MSSs initial implementation was based on serving any request at the time it was received and for any user
  - Inefficient (slow) and leading to significant equipment wear and tear
- Directly integrated logic or modules on top of MSSs support user priorities and request re-ordering according to customizable recall/migration policies
  - Improves access performance typically by 10x
    - Requires large batches of requests (1000s)

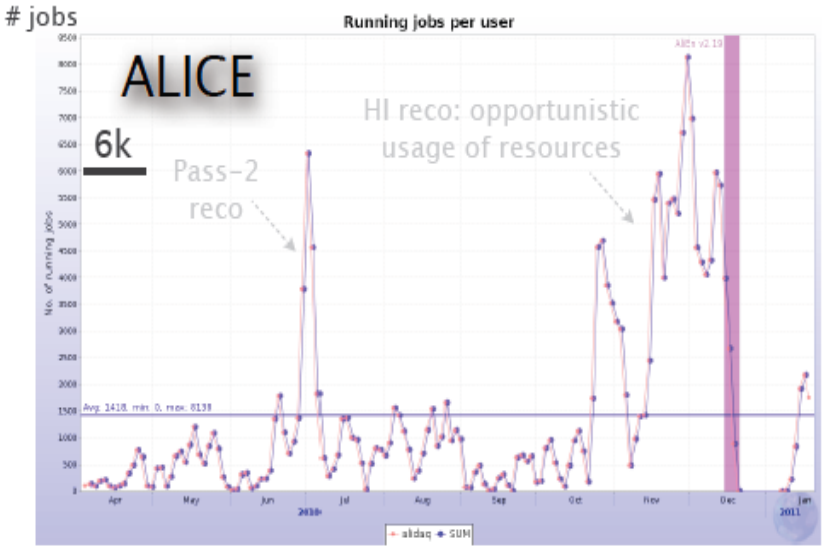
File Staging with Optimization

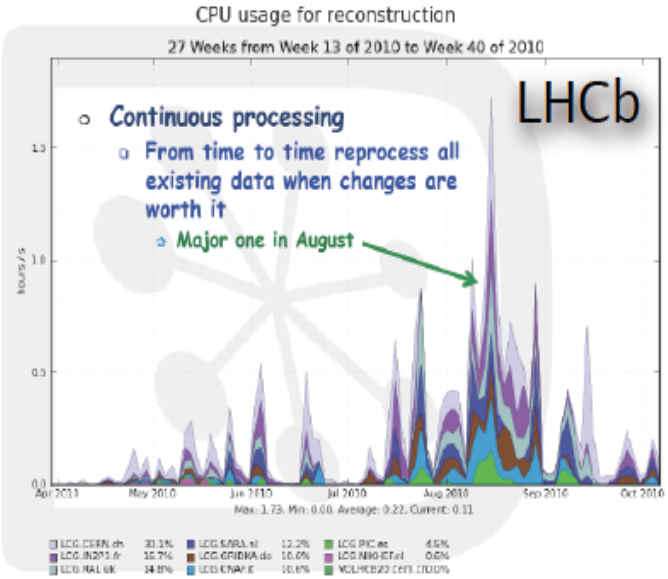


Reprocessing at the Tier-1 Centers



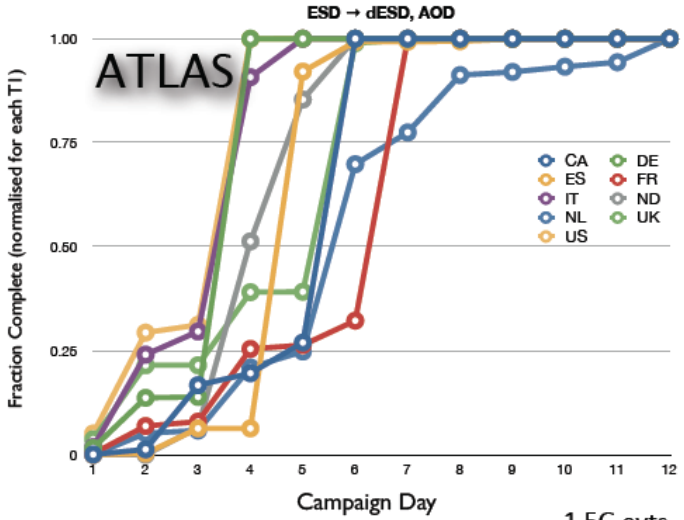






Reprocessing Profile

In 2010, possible to reprocess even more frequently than originally planned

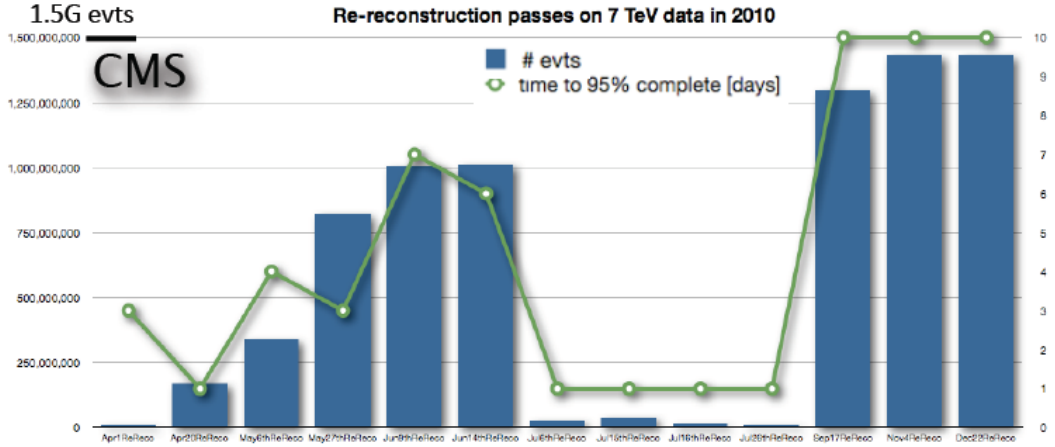


ATLAS reprocessed 100% of data

- RAW→ESD
- ESD merge
- ESD →dESD, AOD
- Grid distribution of derived data

About a dozen of **CMS** reprocessing passes in 2010

Actually, from 7 days onwards mostly dealing with tails



Outlook

The overall Grid infrastructure **is working** for LHC Physics at 7 TeV

- ✓ Distributed storage system able to cope with the amount of data and data access performance requirements so far
  - Computing enabled timely analysis of petascale datasets
  - Challenges ahead with significantly increasing luminosity (100x vs. 2010) and improvements of the LHC machine efficiency during the long 2011/2012 run
- ➢ Is the worldwide distributed Facility with its storage and data management components prepared to scale by 5x ?
  - We are operating systems already at large scale
    - Components limiting scalability (e.g. metadata service, SRM, request scheduler) have been identified and were either already modified/replaced or will be replaced/modified in due time
  - Reduced risk by using different solutions developed for/within the community
  - Evolving Computing Models of the Experiments result in using storage resources more efficiently
    - With changing data access mechanisms/profile may reduce performance requirements
  - ➢ Characteristics of deployed systems have indicated their ability to scale with hardware resources, even at large sites

Acknowledgements

- The following people made significant contributions to this presentation
  - Jon Bakken, Fermilab
  - Luca Dell'Agnello, CNAF
  - Dirk Duellmann, CERN
  - Gonzalo Merion, PIC
  - Andrew Samsun, RAL
  - Reda Tafirout, TRIUMF
  - Ron Trompert, SARA
  - Jos van Wezel, KIT

The Higgs Boson

Professor Peter Higgs proposed that all of space is permeated by a field, the Higgs field.

Quantum theory says that all fields have particles associated with them, so…

in this case…a ***Higgs Boson****.*

The Higgs Boson

Professor Peter Higgs proposed that all of space is permeated by a field, the Higgs field.

Quantum theory says that all fields have particles associated with them, so…

in this case…a ***Higgs Boson****.*

The Higgs has already been discovered at the ATLAS Experiment,

The Higgs Boson

Professor Peter Higgs proposed that all of space is permeated by a field, the Higgs field.

Quantum theory says that all fields have particles associated with them, so…

in this case…a ***Higgs Boson****.*

The Higgs has already been discovered at the ATLAS Experiment, but it was Prof. Higgs, …not the Higgs Boson.

Backup Slides

The Distributed U.S. ATLAS Computing Facility