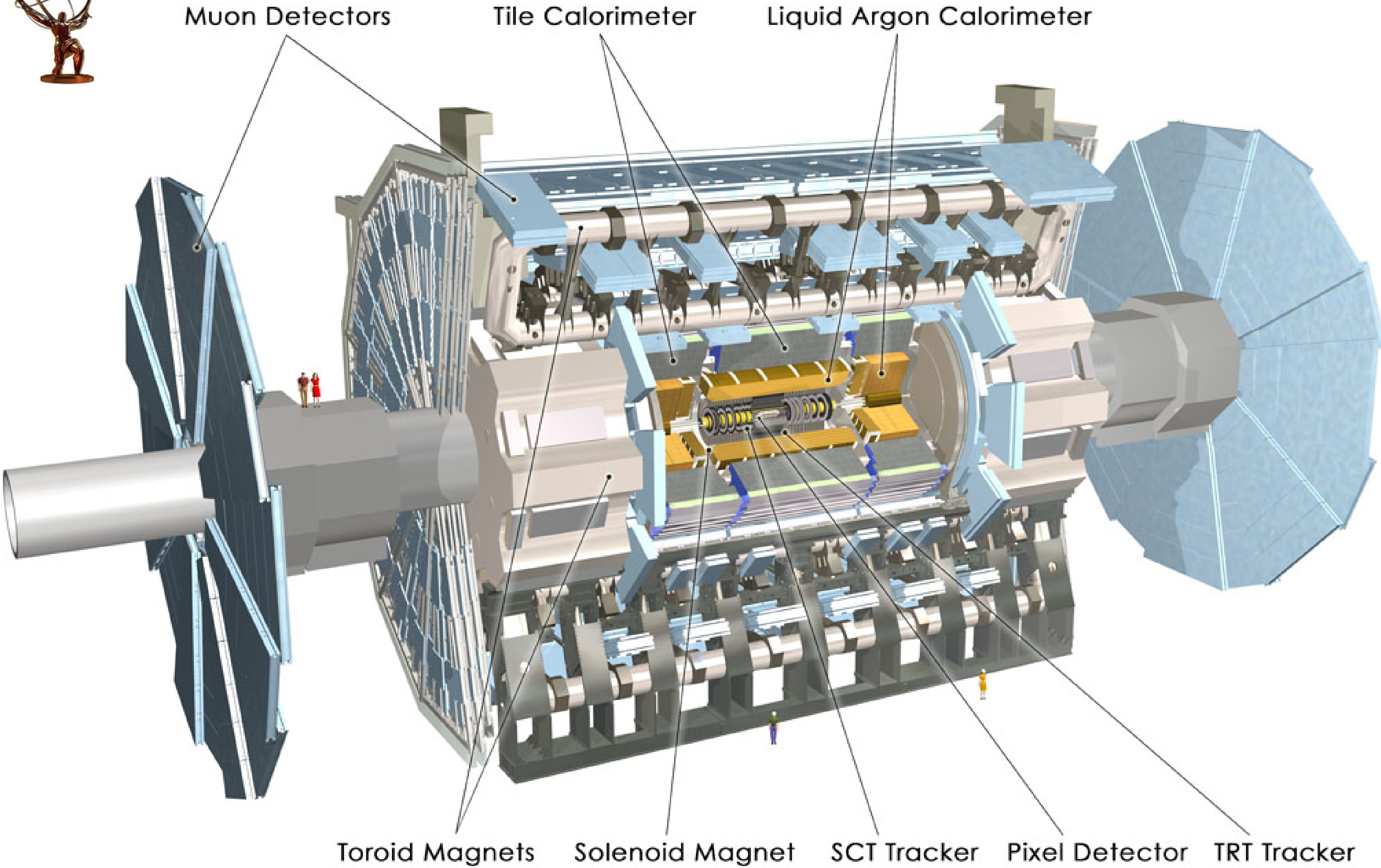
ATLAS実験コンピューティングの現状と将来 －エクサバイトへの挑戦

坂本　宏

東大ICEPP

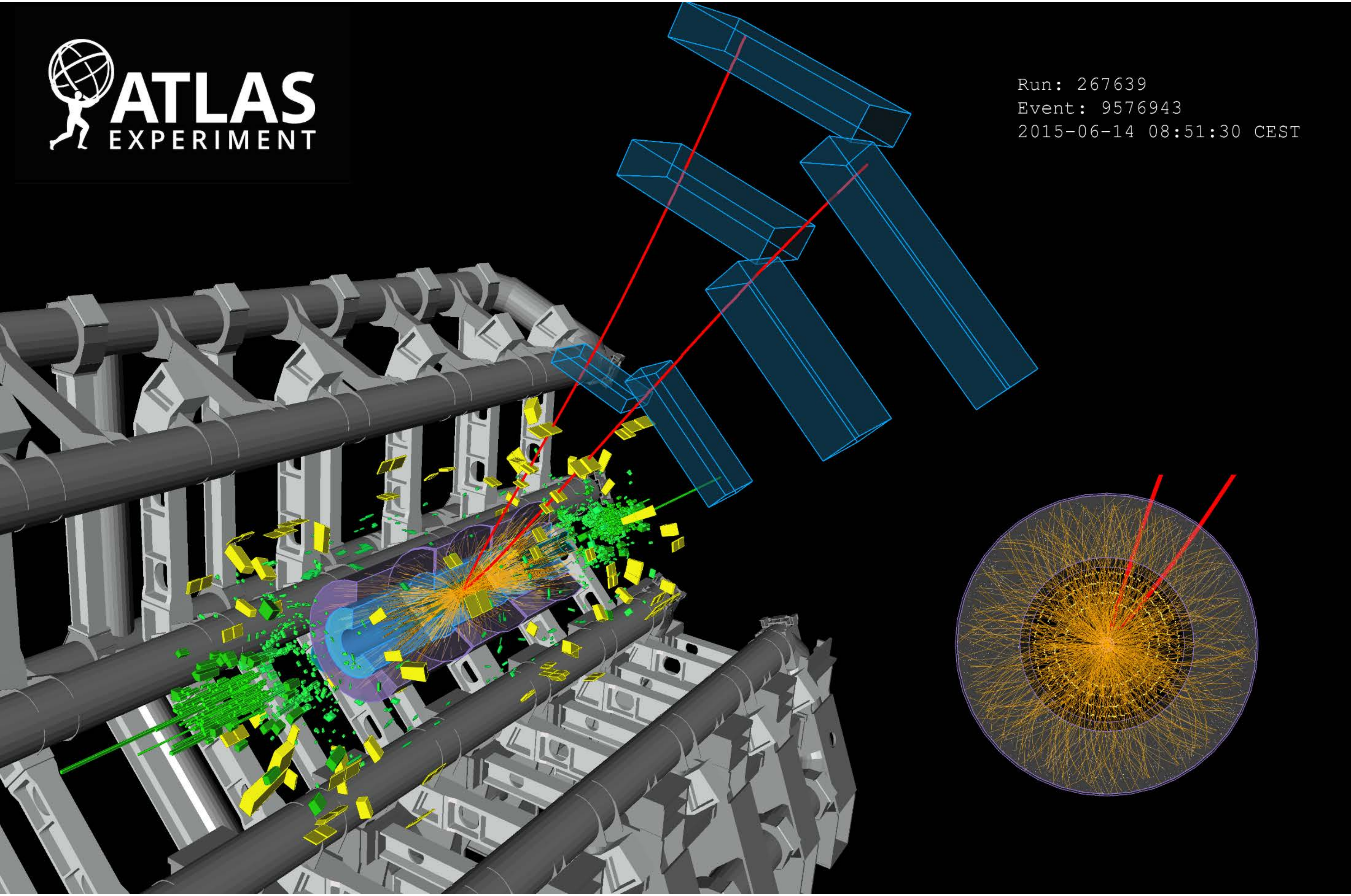
Contents

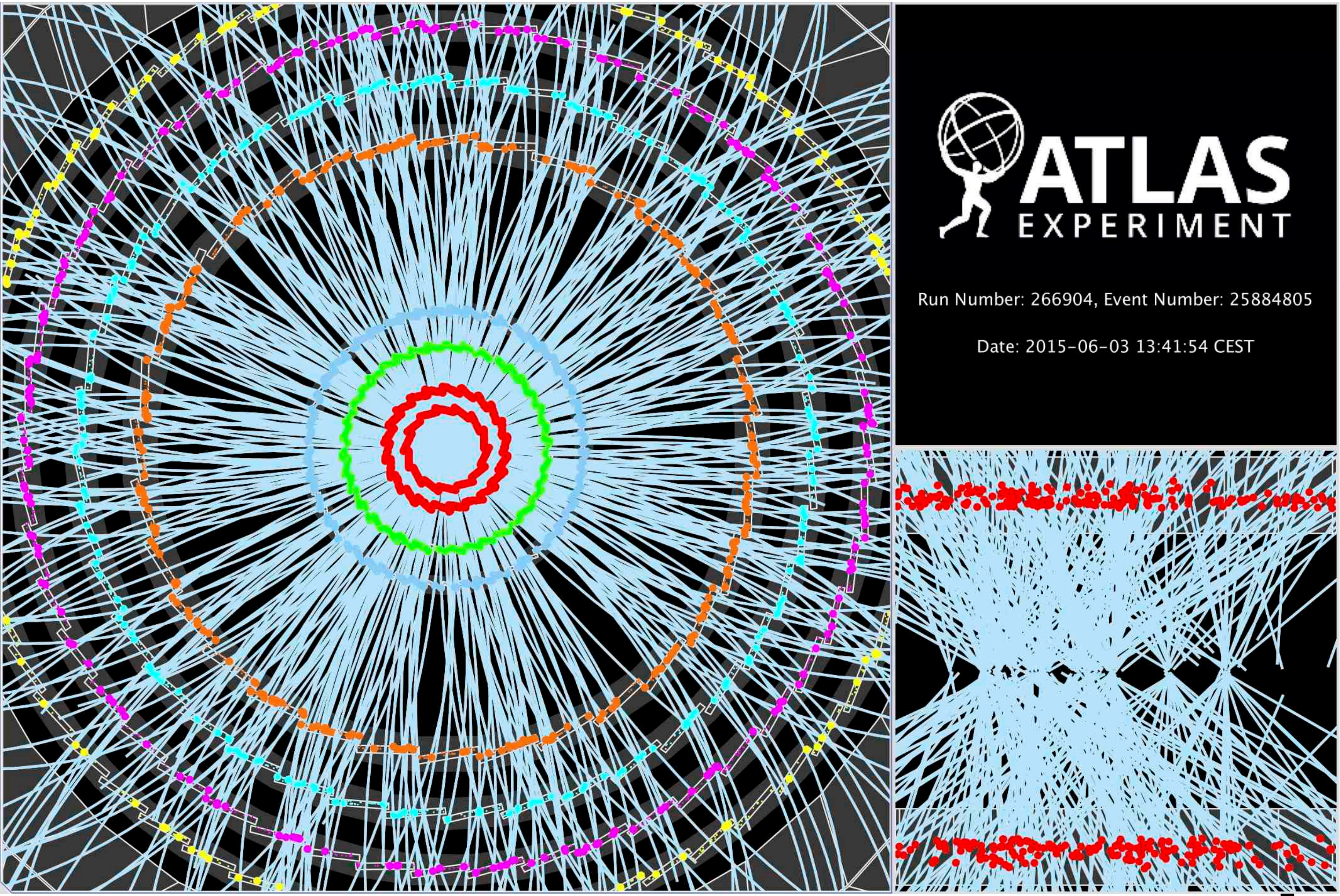
- Energy Frontier Particle Physics
  - Large Hadron Collider (LHC)
  - LHC Experiments: mainly ATLAS
  - Requirements on computing
- Worldwide LHC Computing Grid (WLCG)
  - Globally distributed data analysis infrastructure
  - Middleware
  - Operation
- Toward Exabyte
  - LHC Upgrade plans in 10 years
  - Strategy to handle 100 times more data


Muon Detectors
Tile Calorimeter
Liquid Argon Calorimeter
Toroid Magnets
Solenoid Magnet
SCT Tracker
Pixel Detector
TRT Tracker

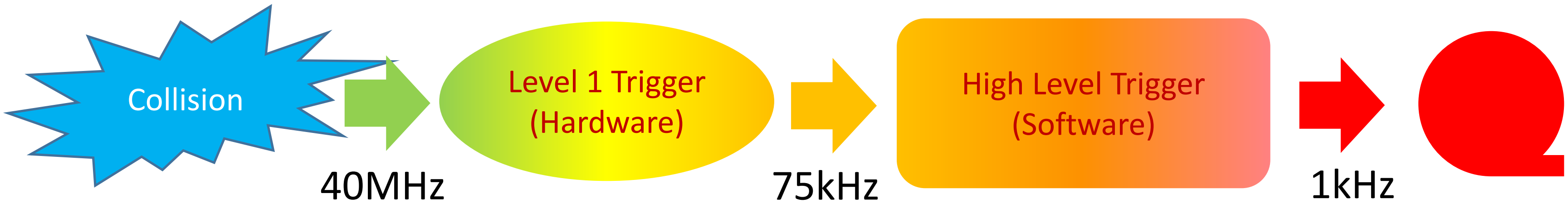
ATLAS Run-2 Detector Status (from July 2017)

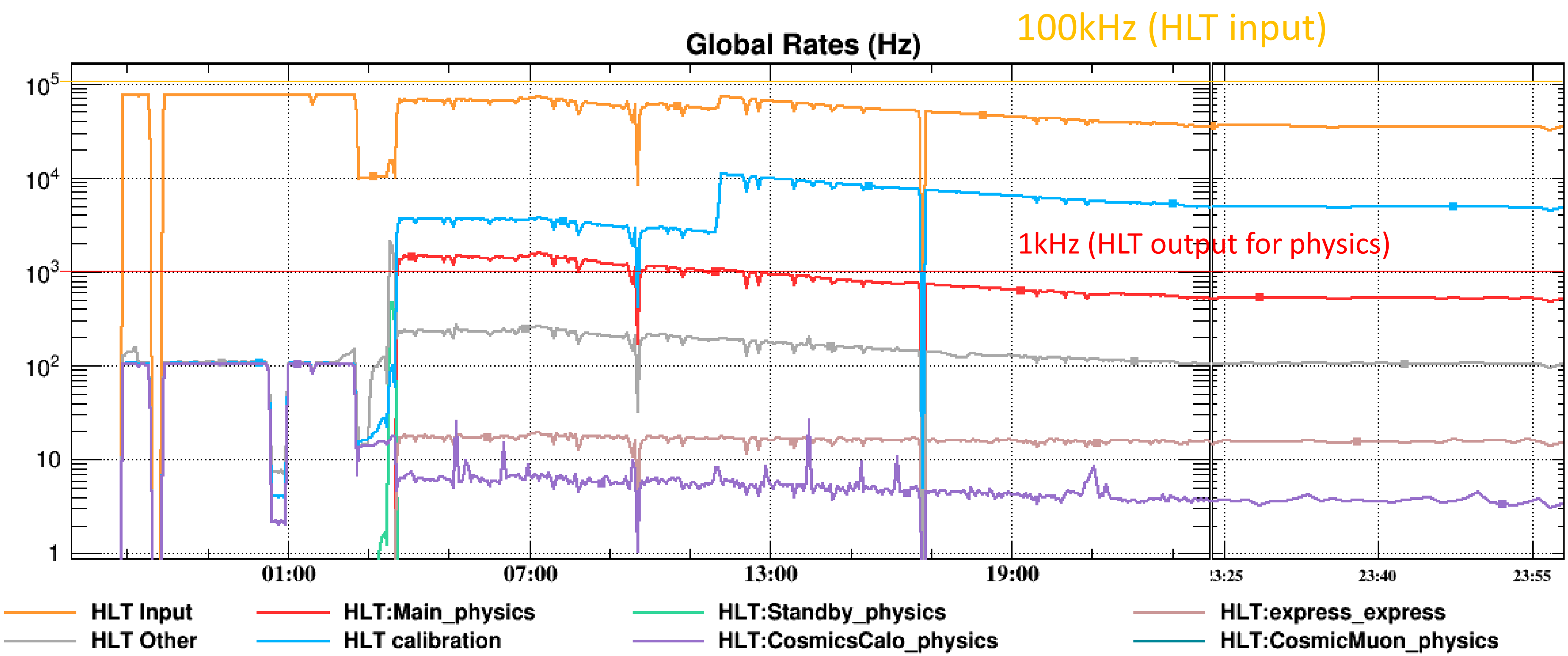
Subdetector	Number of Channels	Approximate Operational Fraction
Pixels	92 M	97.8%
SCT Silicon Strips	6.3 M	98.7%
TRT Transition Radiation Tracker	350 k	97.2%
LAr EM Calorimeter	170 k	100 %
Tile Calorimeter	5200	99.2%
Hadronic End-Cap LAr Calorimeter	5600	99.5%
Forward LAr Calorimeter	3500	99.7%
LVL1 Calo Trigger	7160	99.9%
LVL1 Muon RPC Trigger	383 k	99.8%
LVL1 Muon TGC Trigger	320 k	99.9%
MDT Muon Drift Tubes	357 k	99.7%
CSC Cathode Strip Chambers	31 k	95.3%
RPC Barrel Muon Chambers	383 k	94.4%
TGC End-Cap Muon Chambers	320 k	99.5%
ALFA	10 k	99.9%
AFP	430 k	93.8%

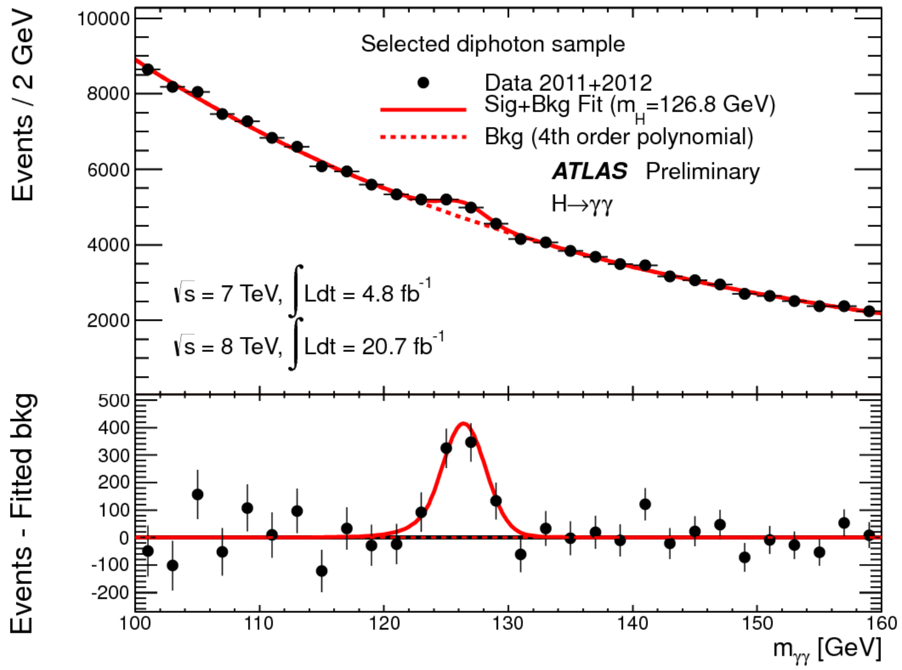

ATLAS
EXPERIMENT
Run: 267639
Event: 9576943
2015-06-14 08:51:30 CEST

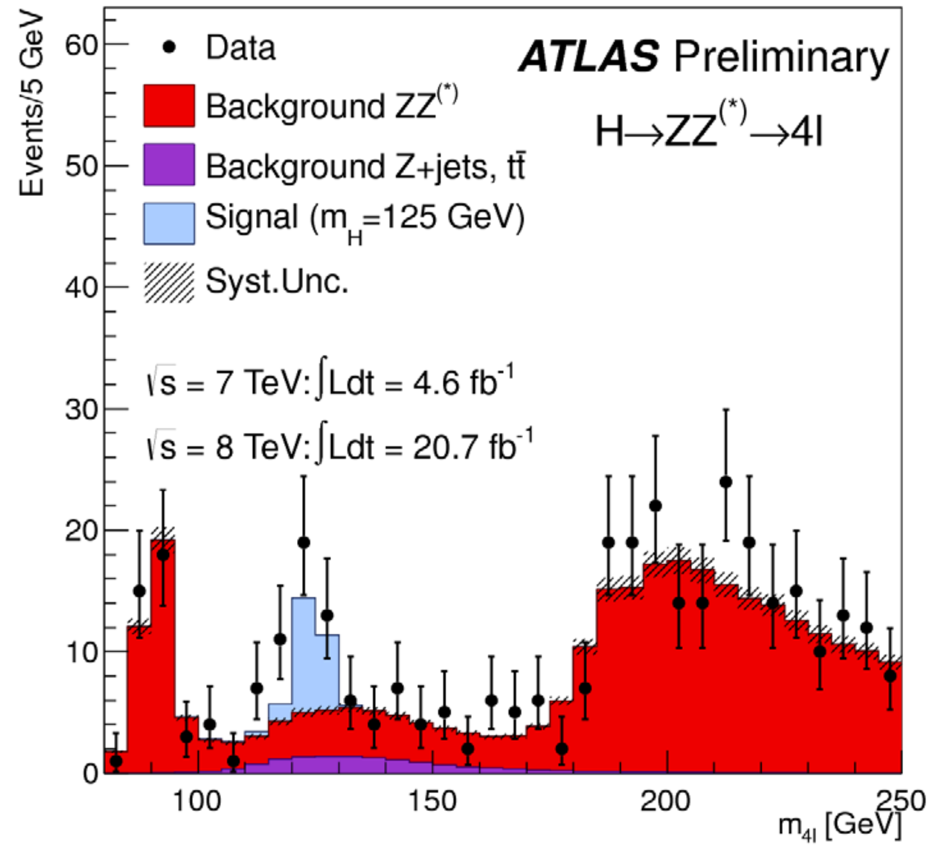

ATLAS
EXPERIMENT
Run Number: 266904, Event Number: 25884805
Date: 2015-06-03 13:41:54 CEST

Trigger System









- Analysis based on:
  - 2011 *pp* data: 3,365,473,349 events
  - 2012 *pp* data: 8,445,206,327 events
- Distributed computing is really working extremely well

CERN Seminar
"Latest update in the search for the Higgs boson"
July 4th, 2012


Global Effort → Global Success
Results today only possible due to extraordinary performance of accelerators – experiments – Grid computing
Observation of a new particle consistent with a Higgs Boson (but which one…?)
Historic Milestone but only the beginning
Global Implications for the future
CERN
R-D Heuer

http://www.wired.com/2013/04/bigdata/

Information Revolution: Big Data Has Arrived at an Almost Unimaginable Scale

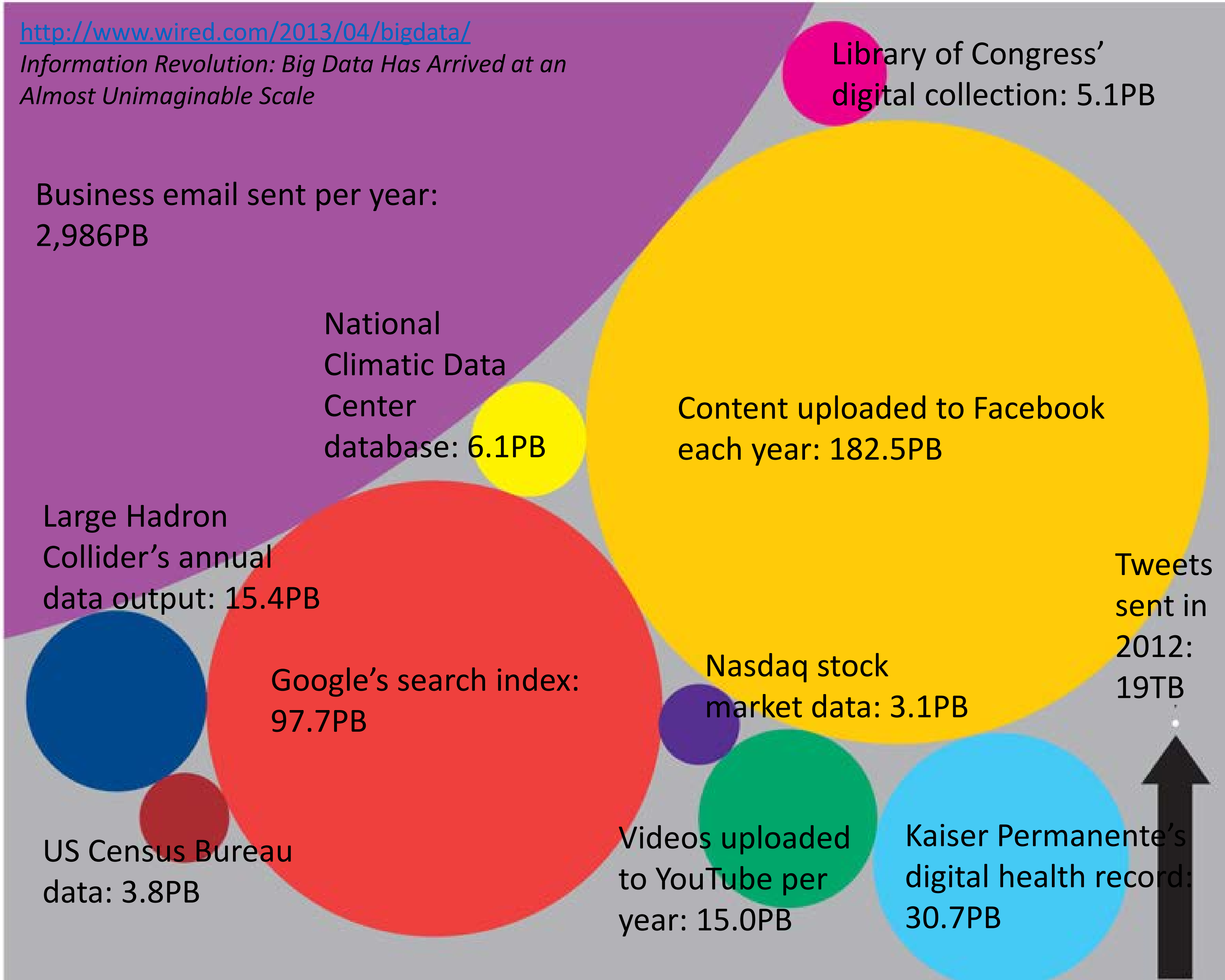


Table 1: Input parameters for ATLAS resource calculations.

LHC and data taking parameters		2012 pp actual	2015 pp μ=25 @ 25 ns	2016 pp μ=40 @ 25 ns	2017 pp μ=40 @ 25 ns
Rate [Hz]	Hz	400 + 150 (delayed)	1000	1000	1000
Time [sec]	MSeconds	6.6	3.0	5.0	7.0
Real data	B Events	3.0 + 0.9 (delayed)	3.0	5.0	7.0
Full Simulation	B Events	2.6 (8 TeV) + 0.8 (7 TeV)	2	2	2
Fast Simulation	B Events	1.9 (8TeV) + 1 (7 TeV)	5	5	5
Simulated Data					
Event sizes					
Real RAW	MB	0.8	0.8	1	1
Real ESD	MB	2.4	2.5	2.7	2.7
Real AOD	MB	0.24	0.25	0.35	0.35
Sim HITS	MB	0.9	1	1	1
Sim ESD	MB	3.3	3.5	3.7	3.7
Sim AOD	MB	0.4	0.4	0.55	0.55
Sim RDO	MB	3.3	3.5	3.7	3.7
CPU times per event					
Full sim	HS06 sec	3100	3500	3500	3500
Fast sim	HS06 sec	260	300	300	300
Real recon	HS06 sec	190	190	250	250
Sim recon	HS06 sec	770	500	600	600
AOD2AOD data	HS06 sec	0	19	25	25
AOD2AOD sim	HS06 sec	0	50	60	60
Group analysis	HS06 sec	40	2	3	3
User analysis	HS06 sec	0.4	0.4	0.4	0.4

7 billion events of real data

2 billion events of full simulation

1MB/event

350 sec to simulate 1 event

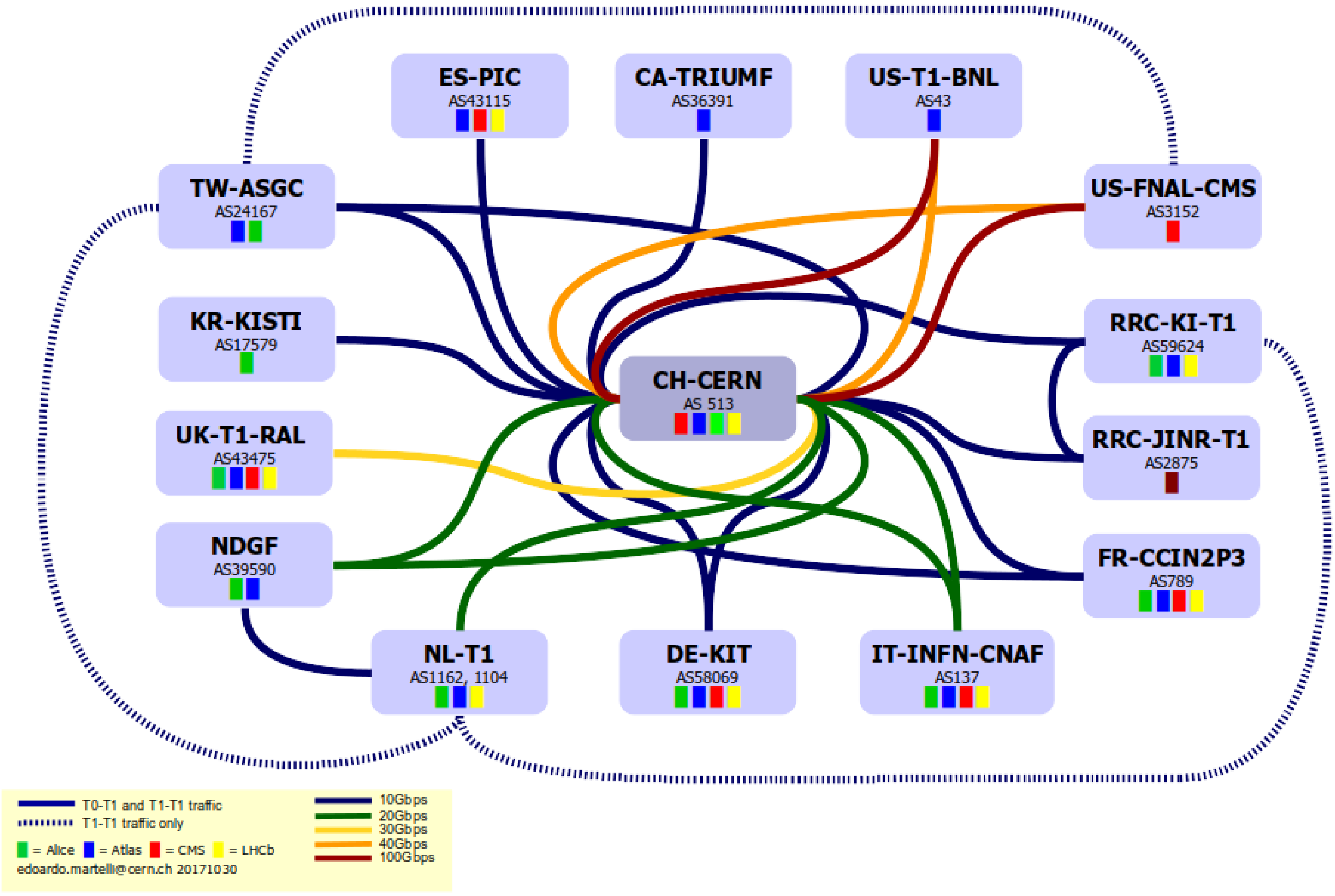
25 sec to reconstruct 1 event

WLCG Collaboration

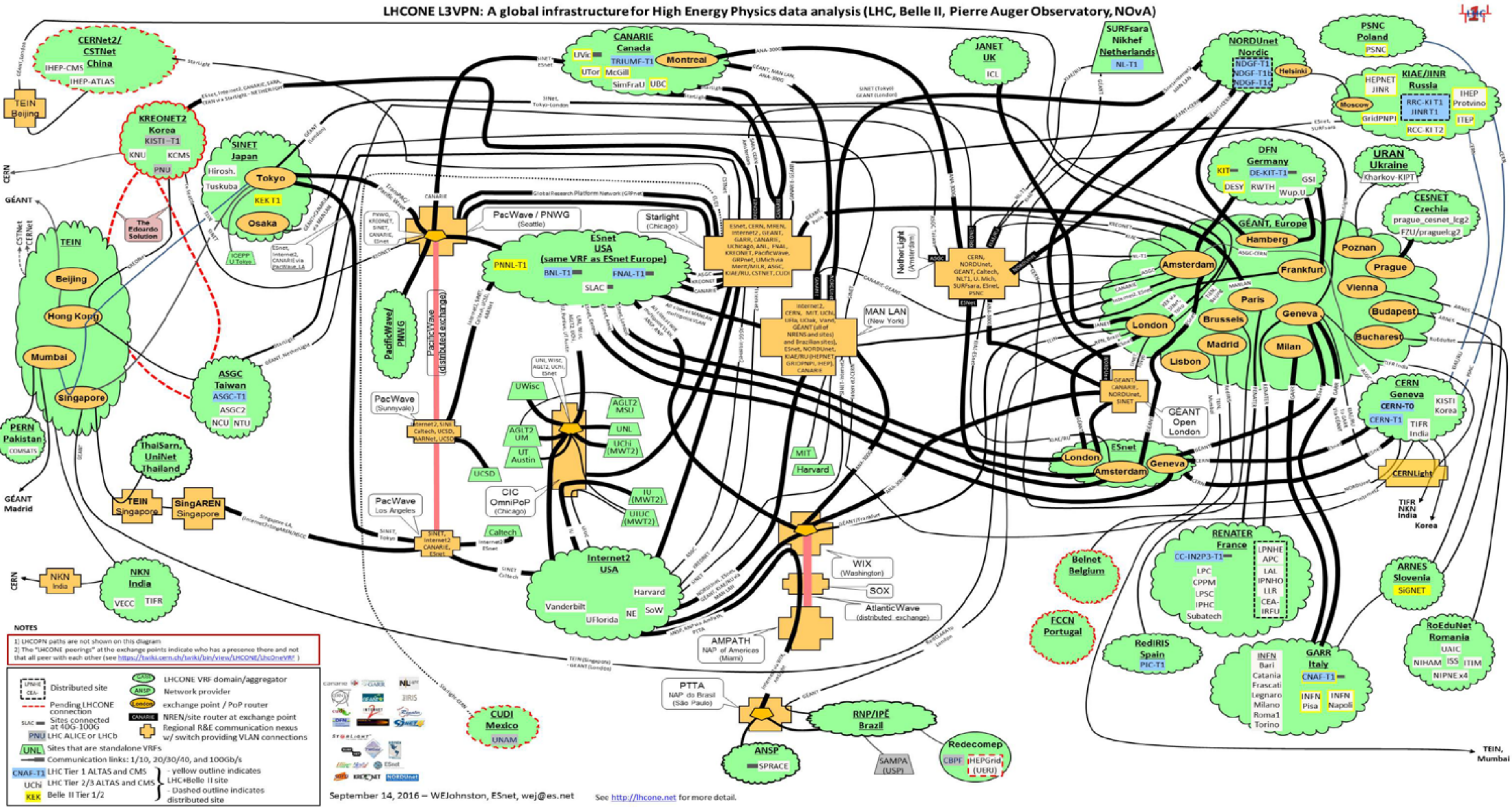
October 2017:
- 63 MoU's
- 167 sites; 42 countries

Networking

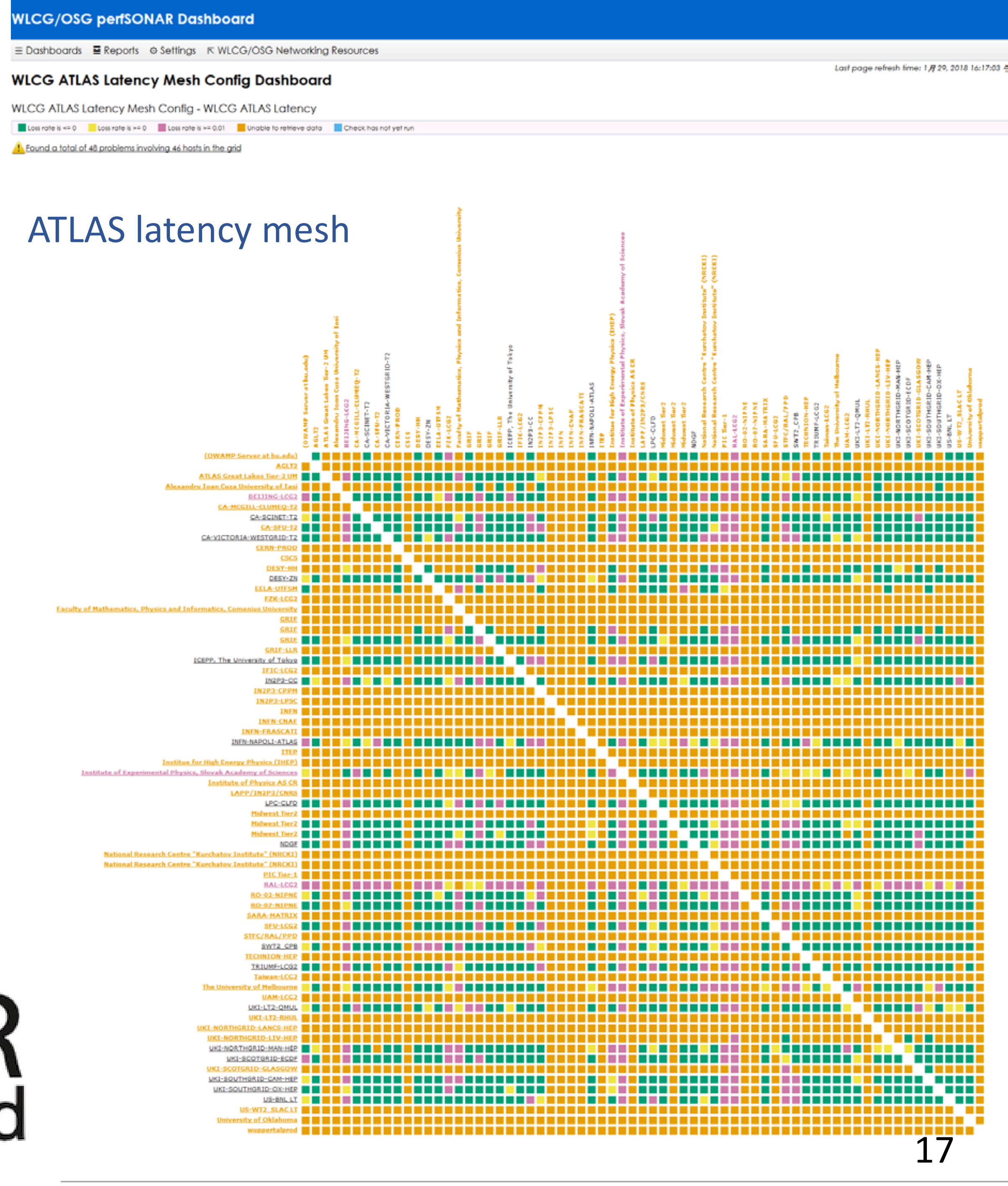
LHCOPN LHC Optical Private Network



LHCONE: LHC Open Network Environment



ATLAS latency mesh



Network monitoring for dynamic resource allocation

perfSONAR monitoring

- Latency (RTT)
- Bandwidth

Grid Middleware

PanDA

Rucio

ami

dCache.org

EOS

XRootD

DPM/LFC

FTS3

VOMS

HTCondor
High Throughput Computing

CernVM
File system

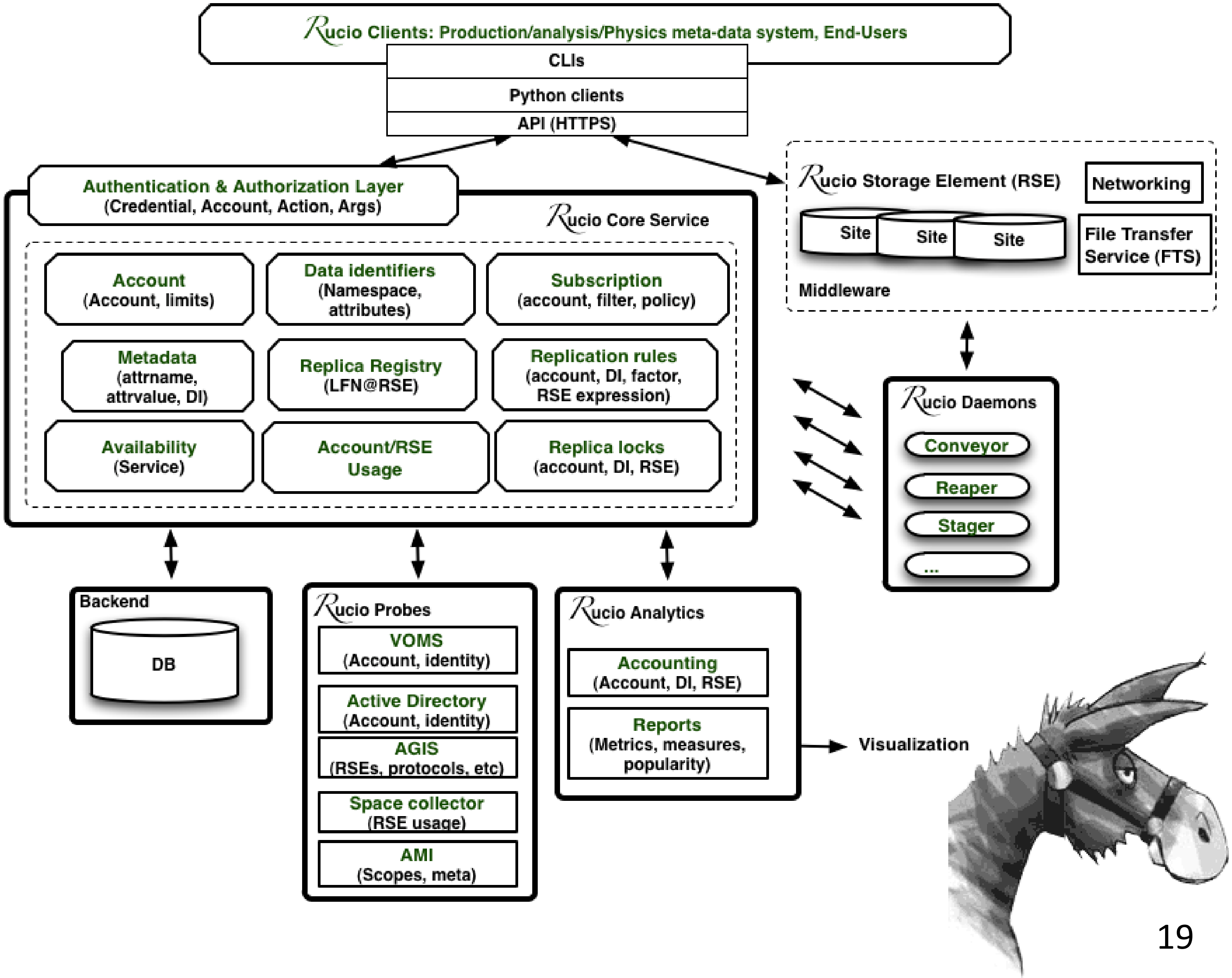
EMI
EUROPEAN MIDDLEWARE INITIATIVE

ARC

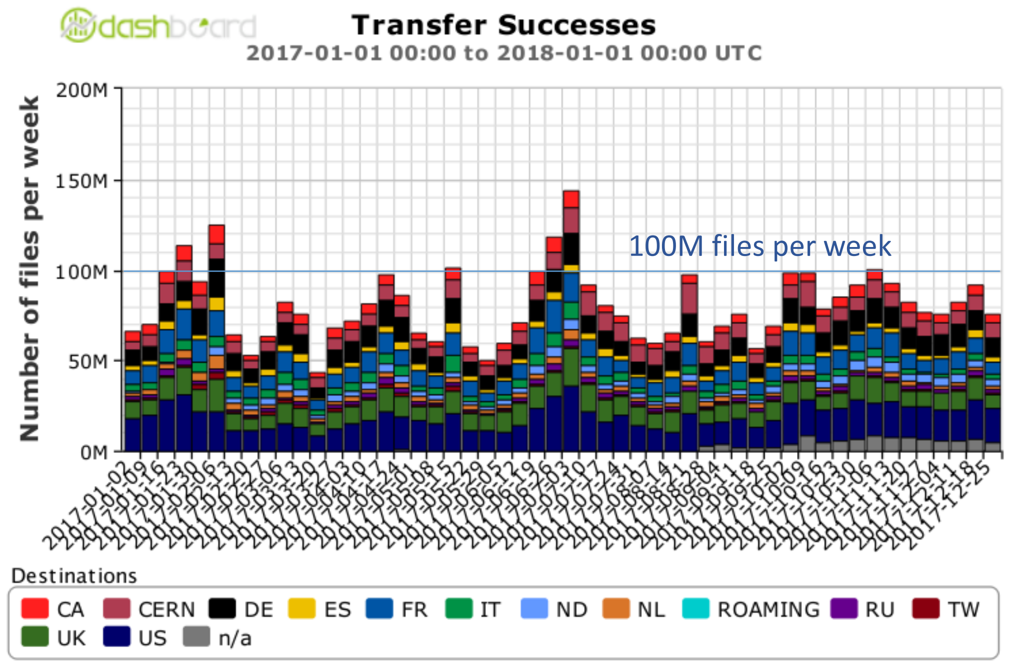
VDT
Virtual Data Toolkit

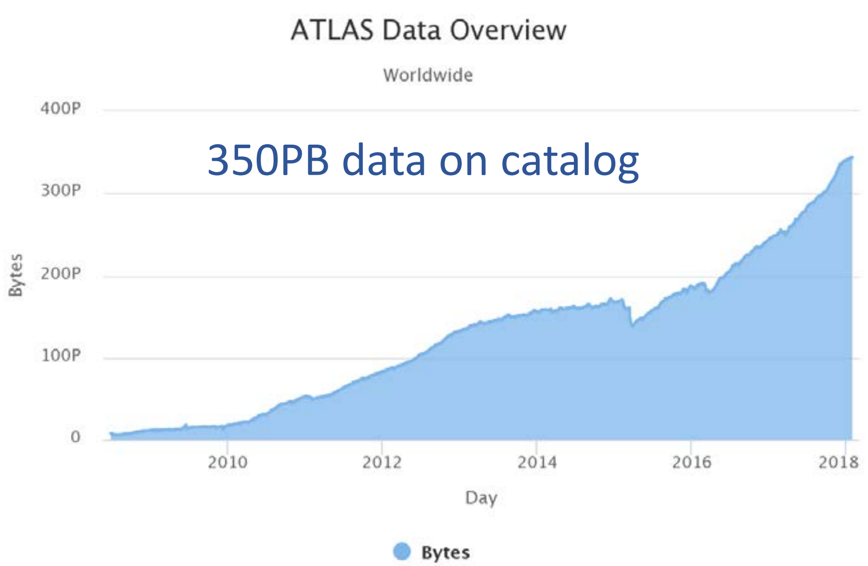
globus toolkit

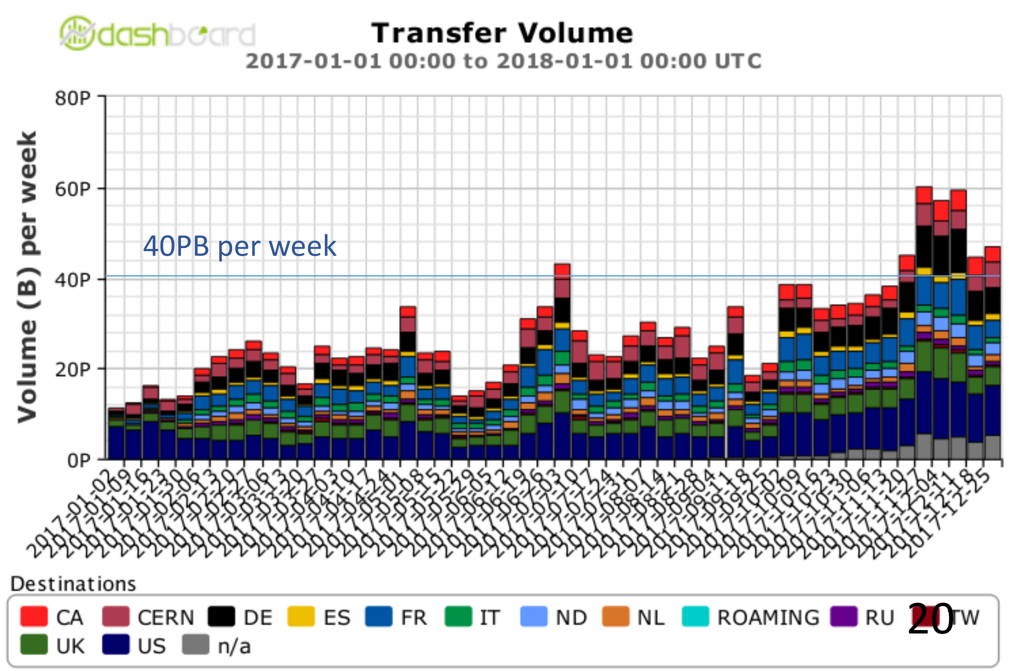
Distributed Data Management



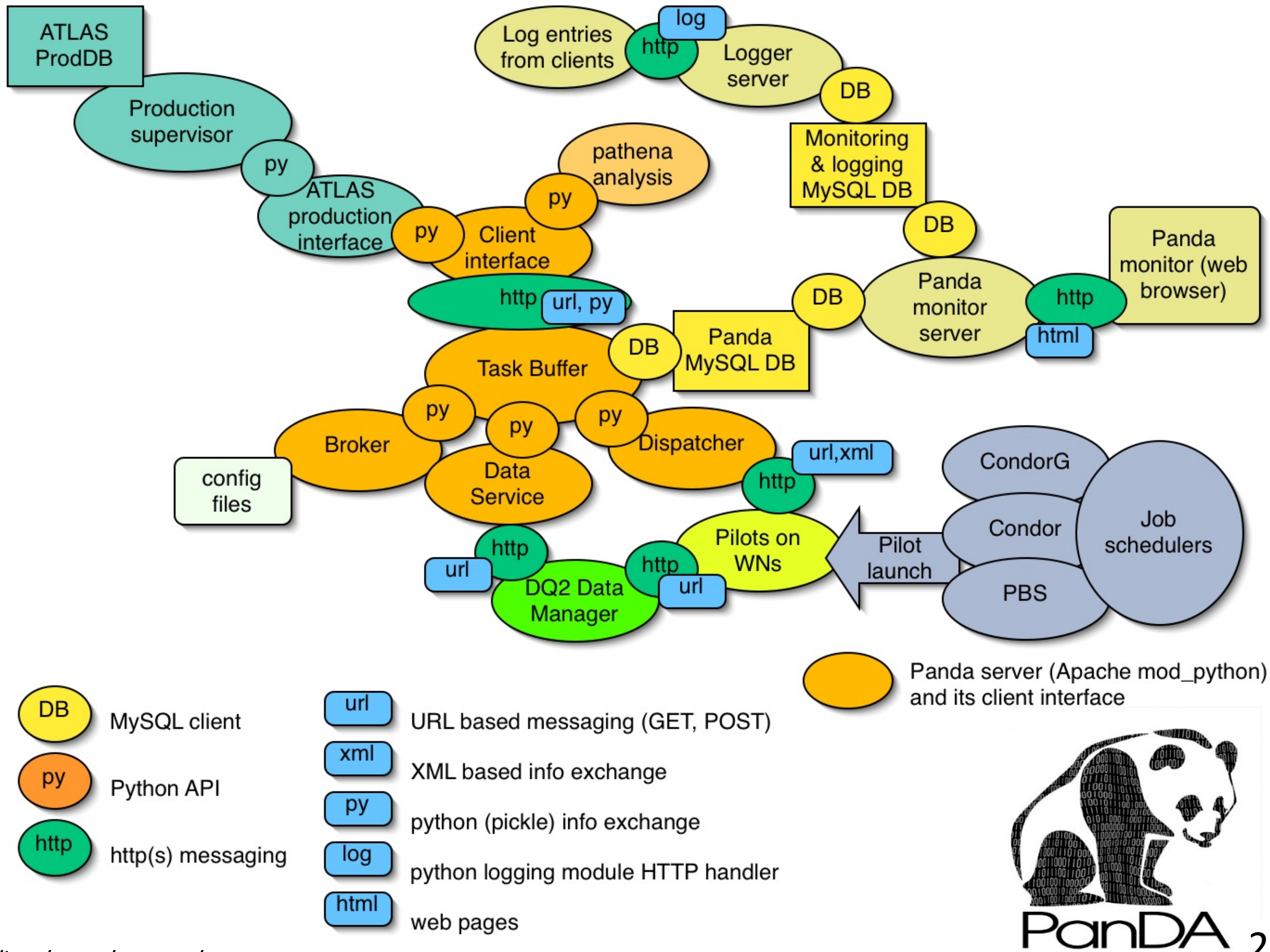
DDM Operation







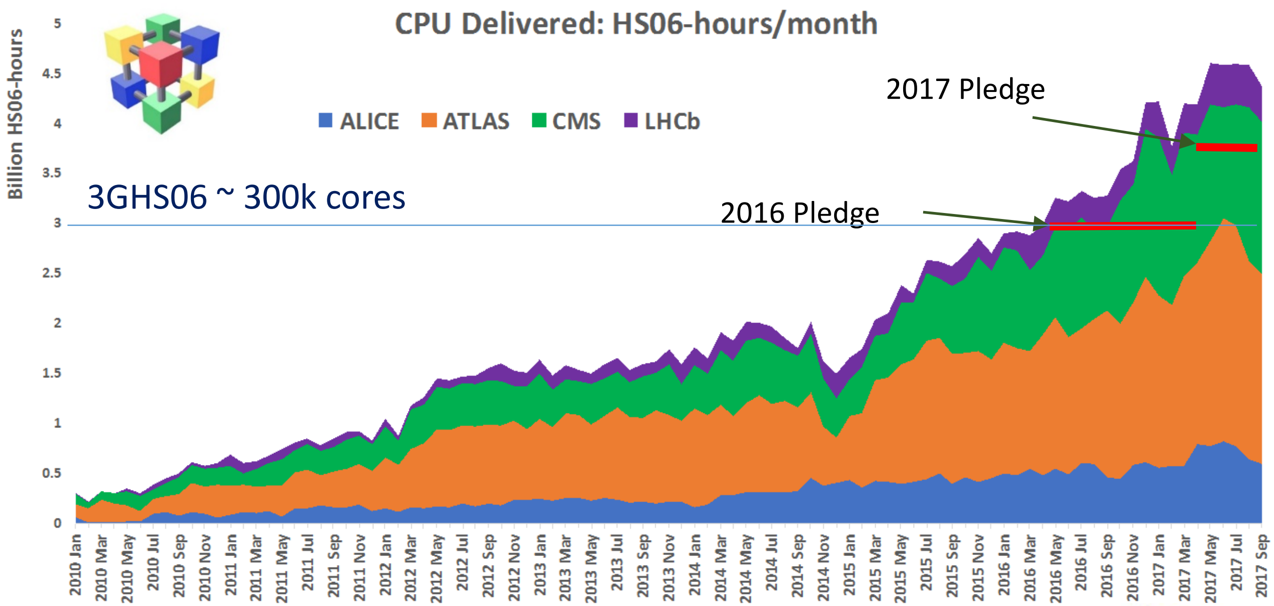
Workload Management



https://twiki.cern.ch/twiki/bin/view/PanDA/PanDA

CPU Delivered

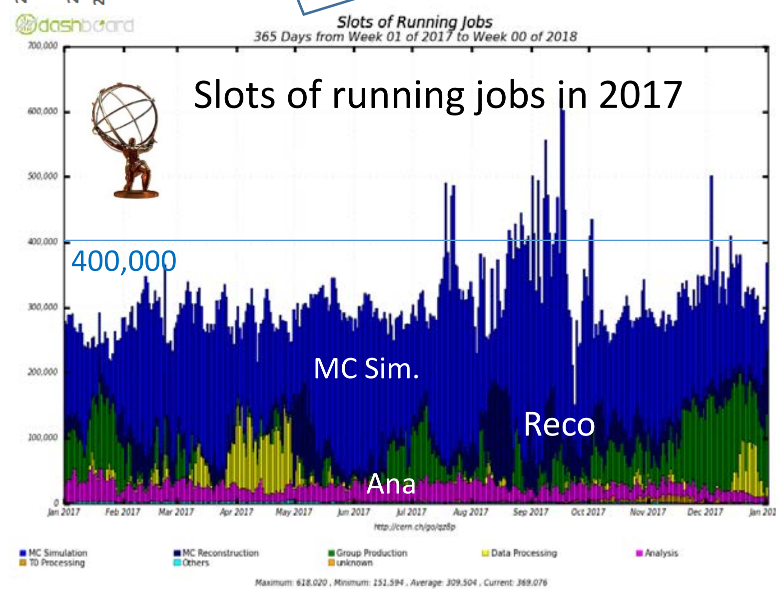
"WLCG Status Report", Ian Bird, Oct.,2017



New peak: ~192 M HS06-days/month
~ 650 k cores continuous

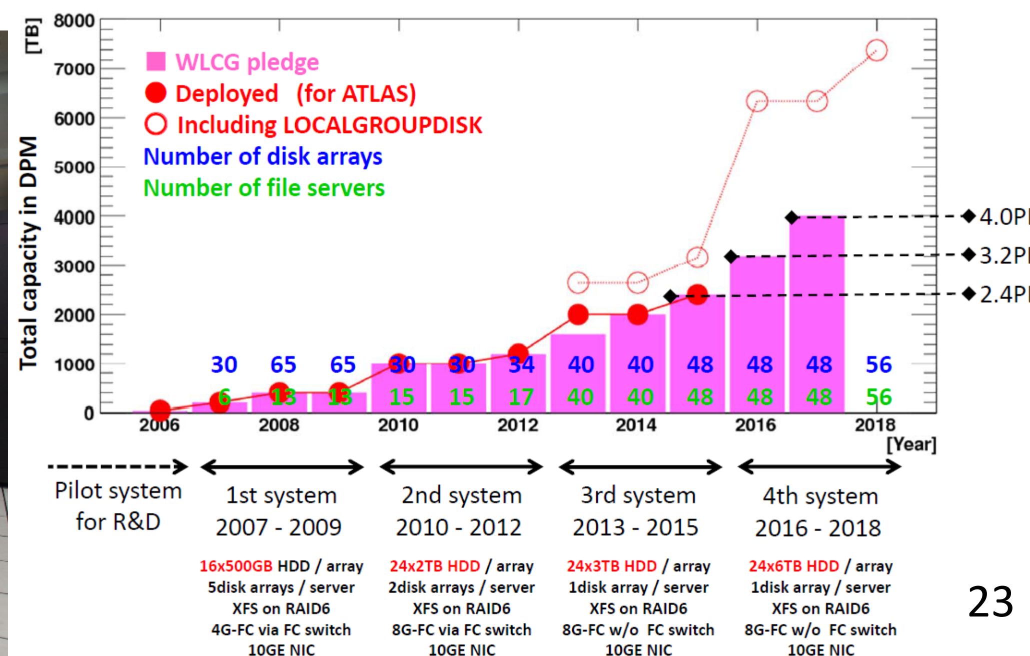
HS06: HEP SPEC 2006 benchmark
(recent core 10 ~ 20 HS06)

300,000 jobs are running always
Excess comes from opportunistic resources like cloud or HPC

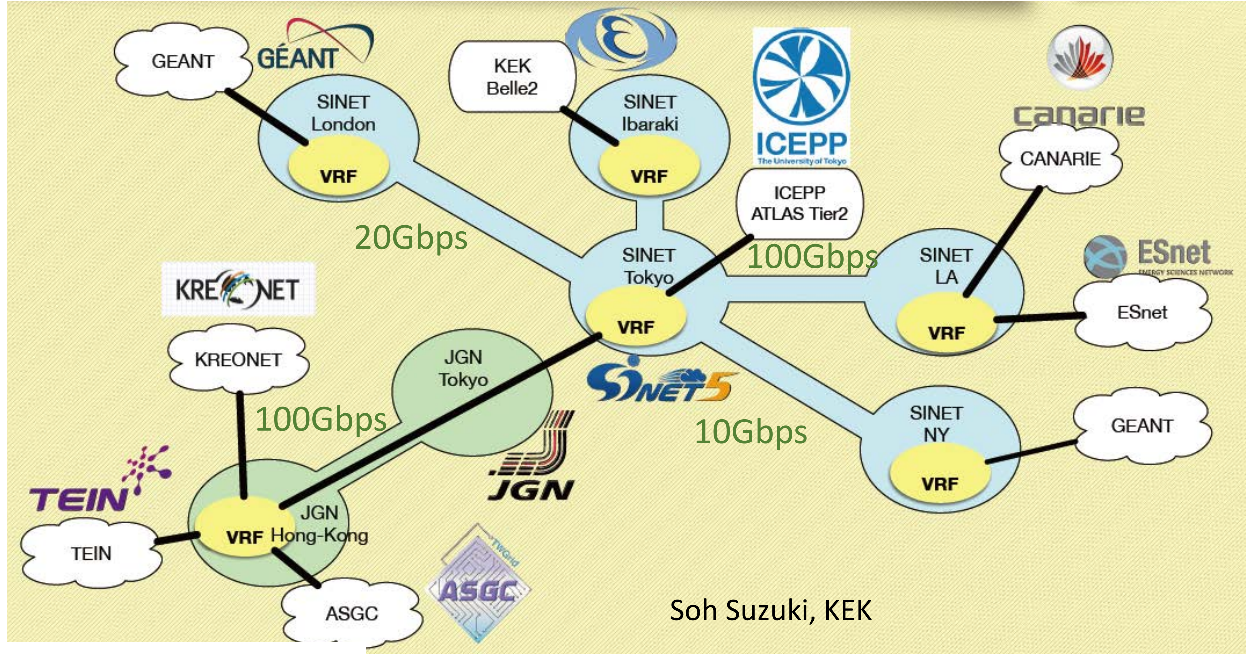


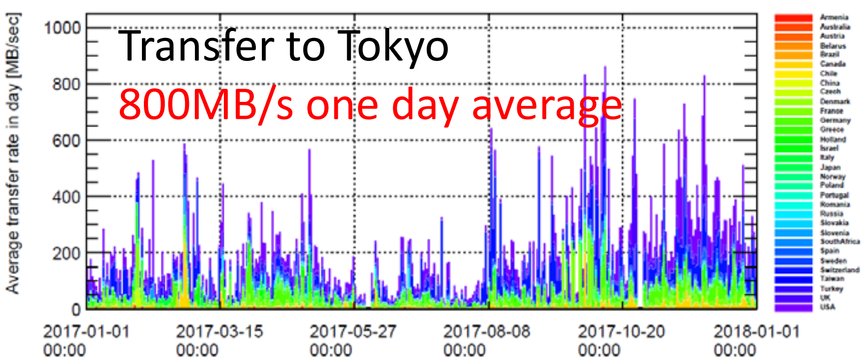
Our Contribution

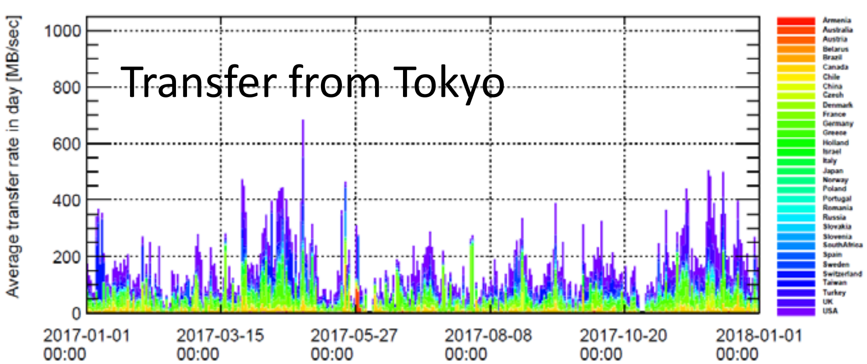
- TOKYO-LCG2
  - Regional Analysis Center in Japan
  - Resources for ATLAS and domestic users
  - 10,000 CPU cores, 10PB disks, 20Gbps network to WAN
- Operational since 2006



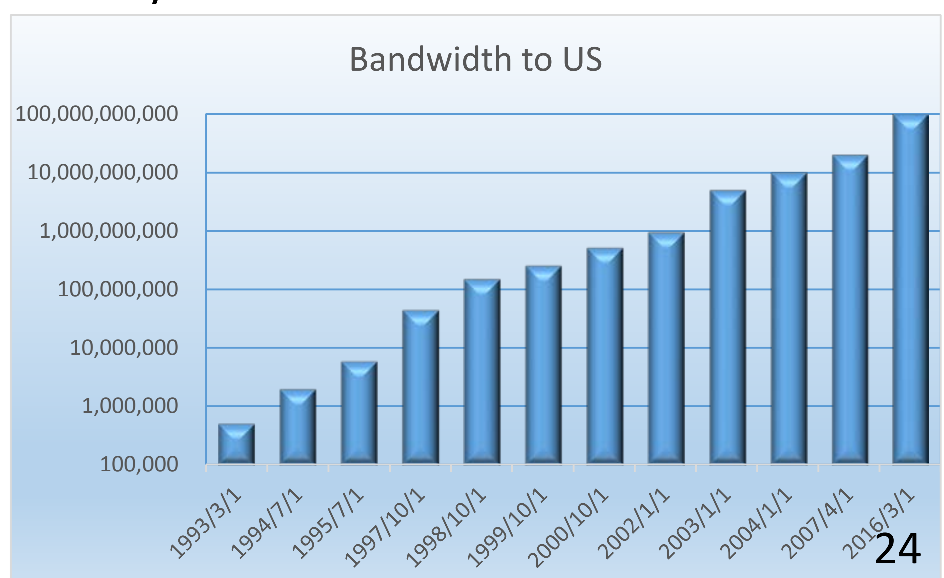
Network Connectivity of Tokyo







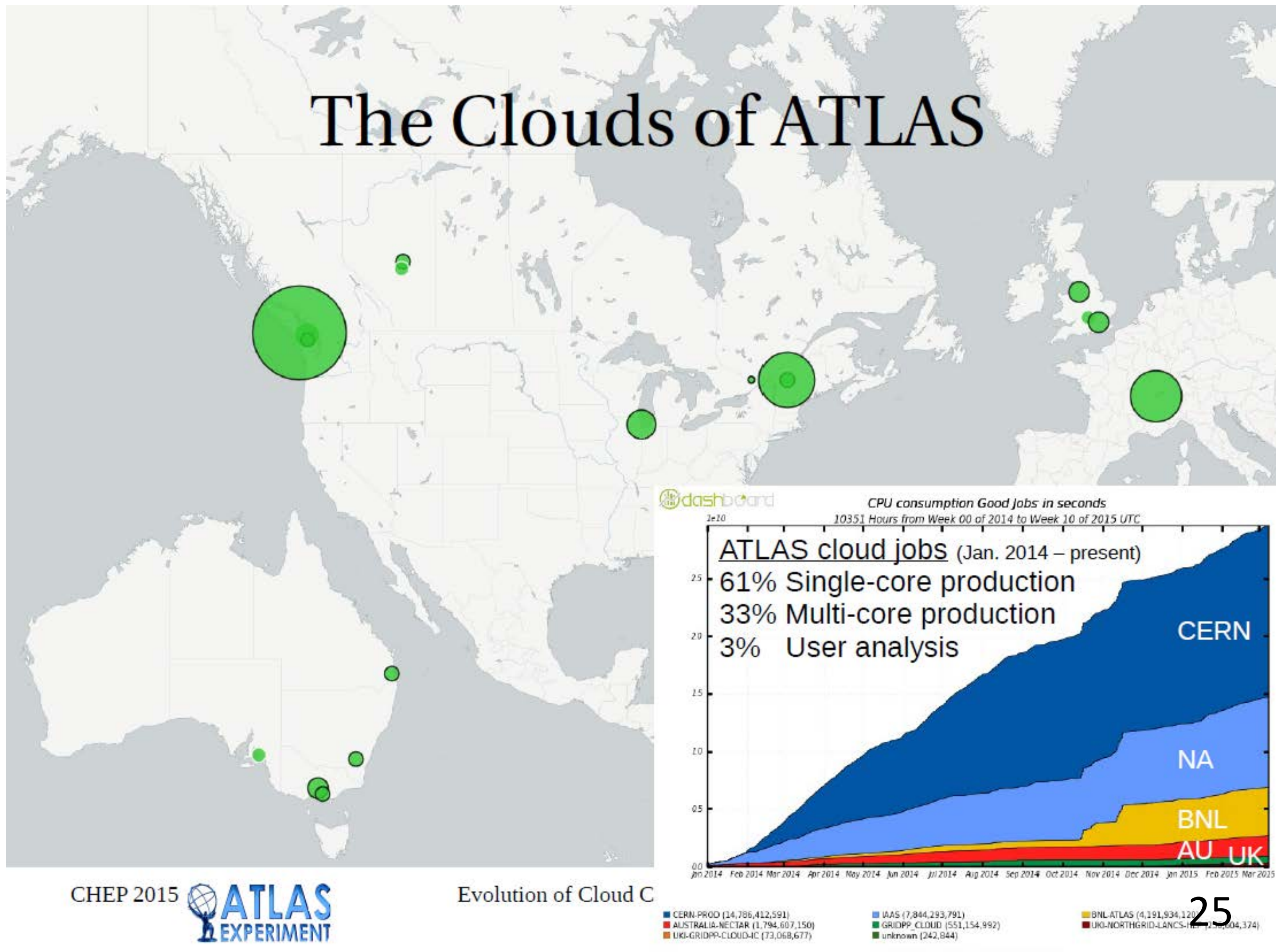
History of SINET international connection to US



Expanding Wings: Cloud Computing

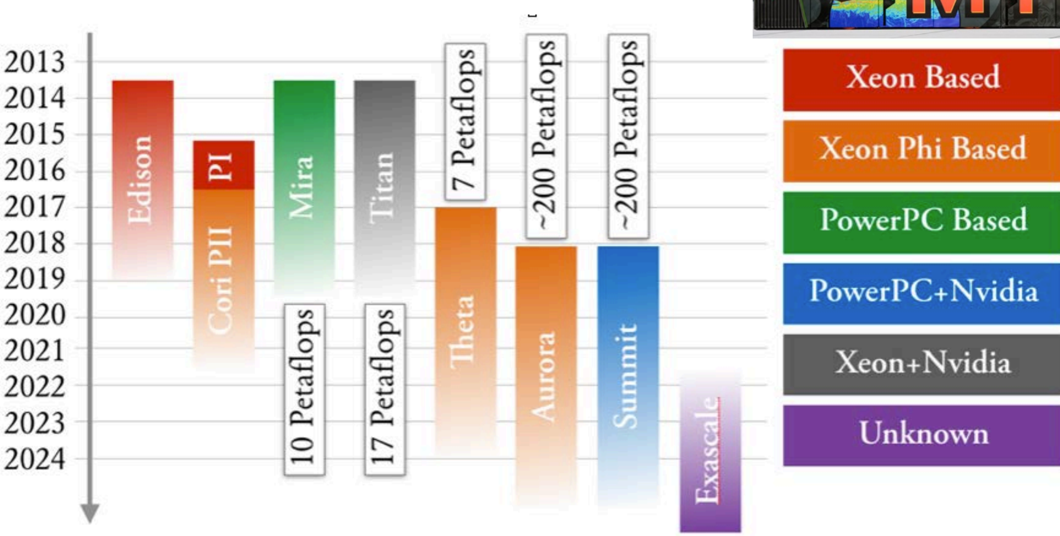
- Private cloud based on OpenStack
- Commercial cloud as opportunistic

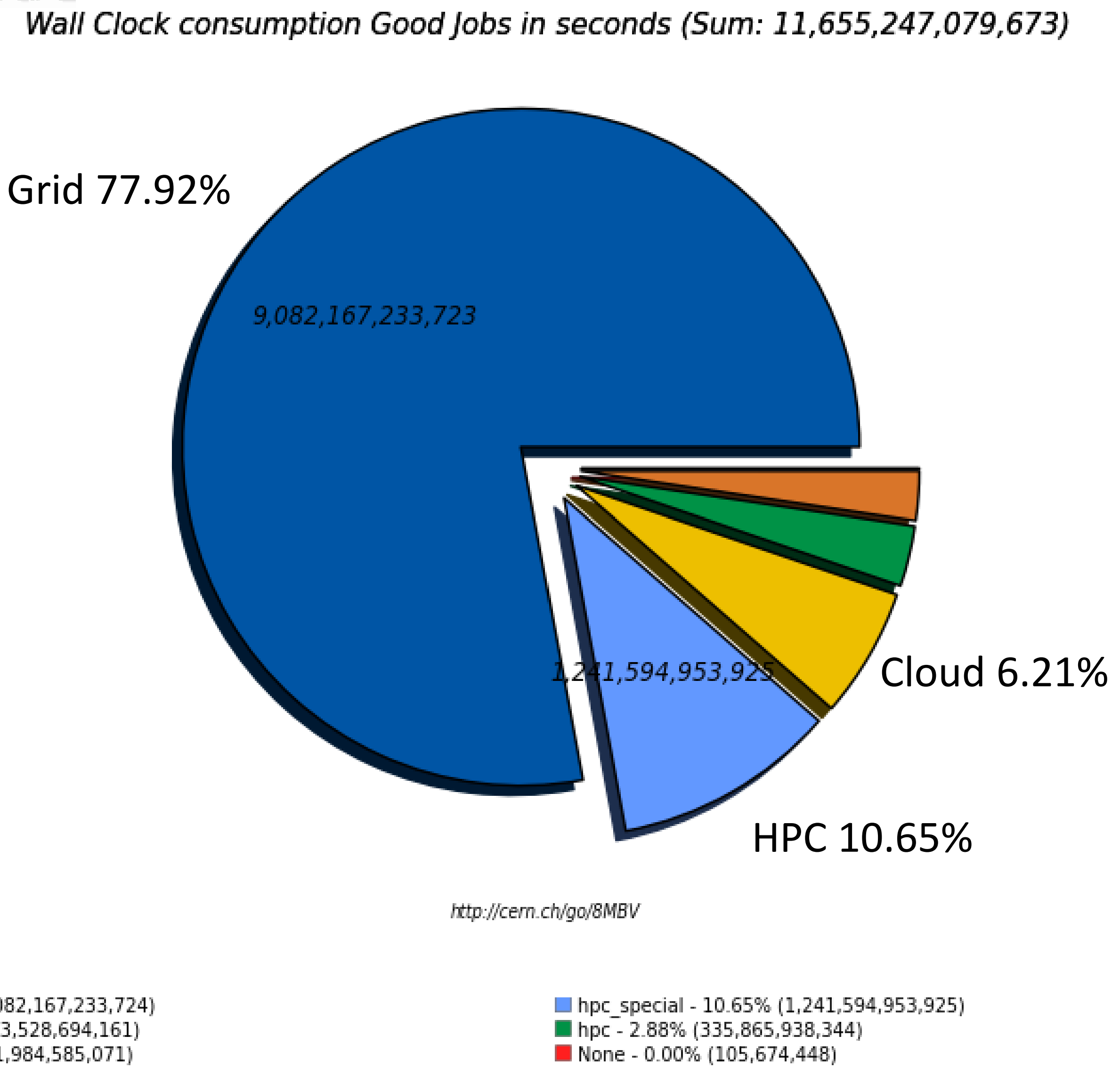
The Evolution of Cloud Computing in ATLAS, Ryan Taylor, CHEP2015 Okinawa Japan, April 13-17, 2015



High Performance Computing

- Mainly for Monte Carlo simulation ~ less IO demands
- Backfill of idling nodes



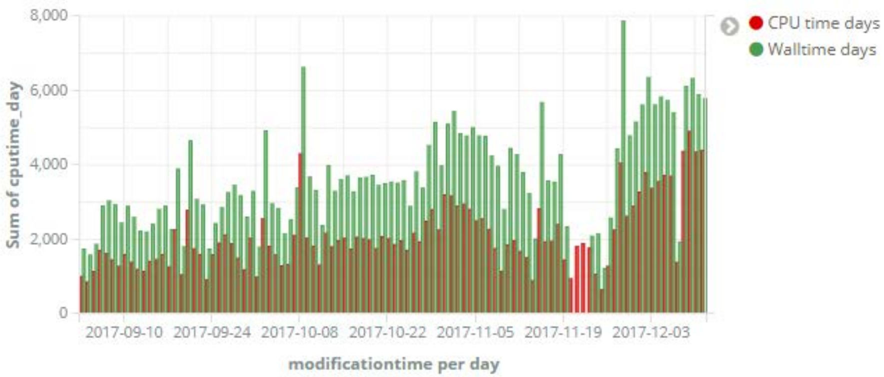


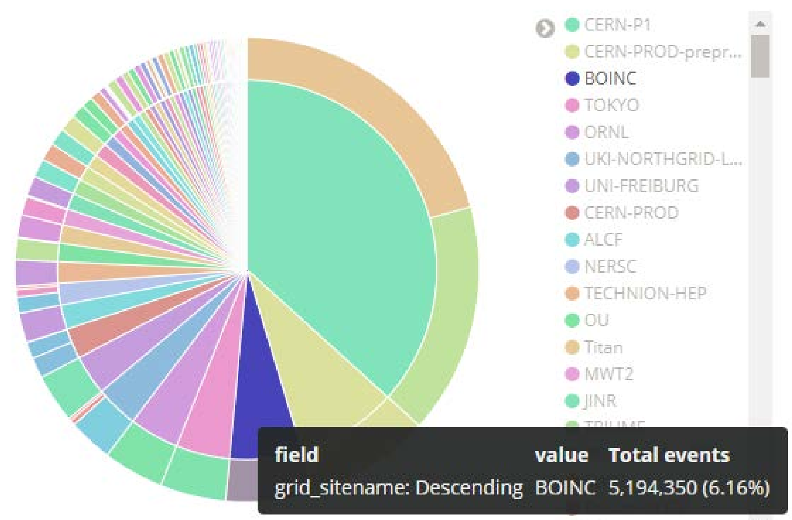
Wall Clock consumption Good Jobs in seconds (Sum: 11,655,247,079,673)
Grid 77.92%
9,082,167,233,723
1,241,594,953,925
Cloud 6.21%
HPC 10.65%
http://cern.ch/go/8MBV
grid - 77.92% (9,082,167,233,724)
cloud - 6.21% (723,528,694,161)
local - 2.33% (271,984,585,071)
hpc_special - 10.65% (1,241,594,953,925)
hpc - 2.88% (335,865,938,344)
None - 0.00% (105,674,448)

ATLAS@HOME: Volunteer Computing

- Framework based on BOINC
- Integrated into WLCG
- Very low cost of operation thanks to virtualization
  - Even for small sites

http://boinc.berkeley.edu/





http://lhcathome.web.cern.ch/projects/atlas

Run 1

Run 2

‘01- NorduGrid

‘02- ARC Advanced Resource Connector

‘06-’08 gLite Lightweight Middleware for Grid Computing

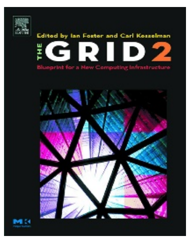
‘08- EMI European Middleware Initiative

‘04-’05 LCG2 Grid Middleware

’11-’14 UMD2 Unified Middleware Distribution

‘13-’17 UMD3 Unified Middleware Distribution

‘16- UMD4 Unified Middleware Distribution



’01- LCG LHC Computing Grid

‘98 “The grid: blueprint for a new computing infrastructure”, I. Foster, C. Kesselman

‘01-’04 EDG European Data Grid

‘04-’10 EGEE Enabling Grids for E-science

‘10- EGI European Grid Infrastructure

‘98-’99 Monarc Document

‘04- OSG Open Science Grid

‘98 ‘99 ‘00 ‘01 ‘02 ‘03 ‘04 ‘05 ‘06 ‘07 ‘08 ‘09 ‘10 ‘11 ‘12 ‘13 ‘14 ‘15 ‘16 ‘17 ‘18

‘16 Machine Learning
‘04 OSG Open Science Grid
’13 UMD2 ‘15 HPC backfill
‘03 US Grid2003 Project
’15 HSF HEP Software Foundation
’01 US ATLAS Testbed
‘14 ATLAS@HOME (BOINC)
’13 ATLAS Prodsys II
‘11 EMI1
’13 ATLAS DDM Rucio
‘12 CERN Agile Infrastructure (Cloud Svc)
‘12 Cloud Computing R&D
’11- LHCONE Open Network Environment
’06 gLite3
‘10 CVMFS CERN VM file system
‘07 Panda production (EGEE)
’05- Panda production system (OSG)
’06- ATLAS DDM DQ2
’04- ATLAS DDM Don Quijote
‘04 LCG-2 Deployment
‘03 LCG-1 Deployment
’02 EDG Testbed 1
‘01 Nordugrid
‘08- ADCoS Distributed Computing Operations Shift
‘05 LCG MoU Memorandum of Understanding
‘05 ATLAS LCG Computing TDR
‘14 ATLAS DC14
‘09 STEP09 Scale Testing for the Experimental Programme
‘08 WLCG CCRC Combined Computing Readiness Challenge
‘08 FDR2
’15 LHC start of run2
‘07 FDR Full Dress Rehearsal
‘13 Nobel prize to Higgs and Englert
’05-’06 LCG Service Challenges
’13 LHC end of run1
’05-’06 ATLAS DC4
‘12 ATLAS CMS discover new particle
’04-’05 ATLAS DC3
’10 First collision at 7 TeV
‘03 ATLAS DC2
’09 LHC beam back
‘02 ATLAS DC1
’08 LHC First beam
‘01 ATLAS DC0 Data Challenge
‘08 ATLAS Installation Completed

‘98 ‘99 ‘00 ‘01 ‘02 ‘03 ‘04 ‘05 ‘06 ‘07 ‘08 ‘09 ‘10 ‘11 ‘12 ‘13 ‘14 ‘15 ‘16 ‘17 ‘18

Evolution of middleware ~ Lessons learned

- Static allocation to Dynamic allocation
- Pre-scheduled operation to On-Demand operation
- Private protocol to Industrial Standard
- Single flavor platform to Virtual machines
- General purpose to Application specific
- Manual operation to Automation

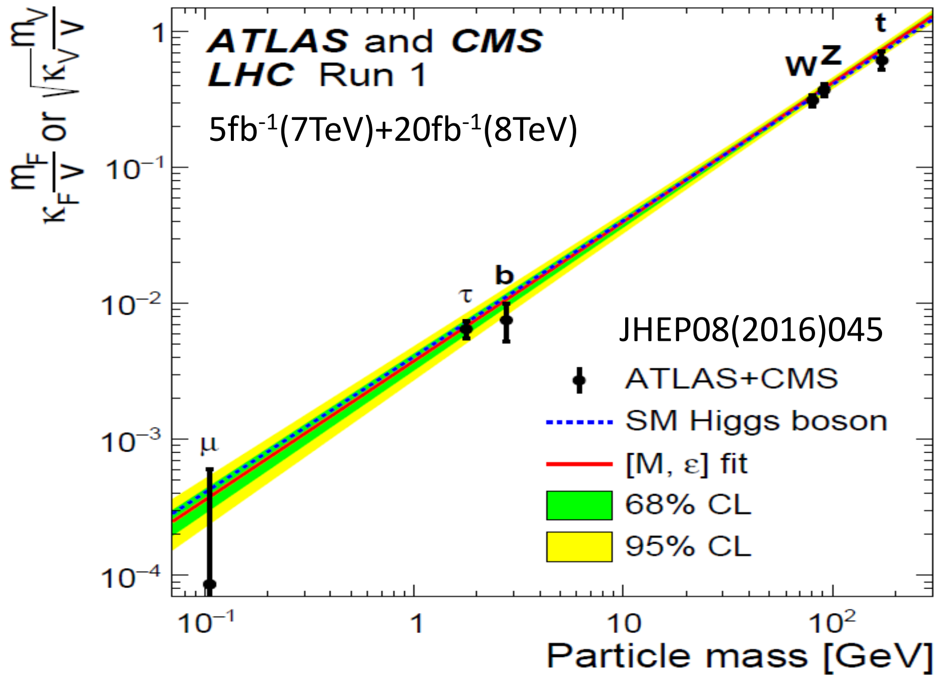
Higher performance
Better resource utilization
Stable/Sustainable operation
Lower maintenance cost

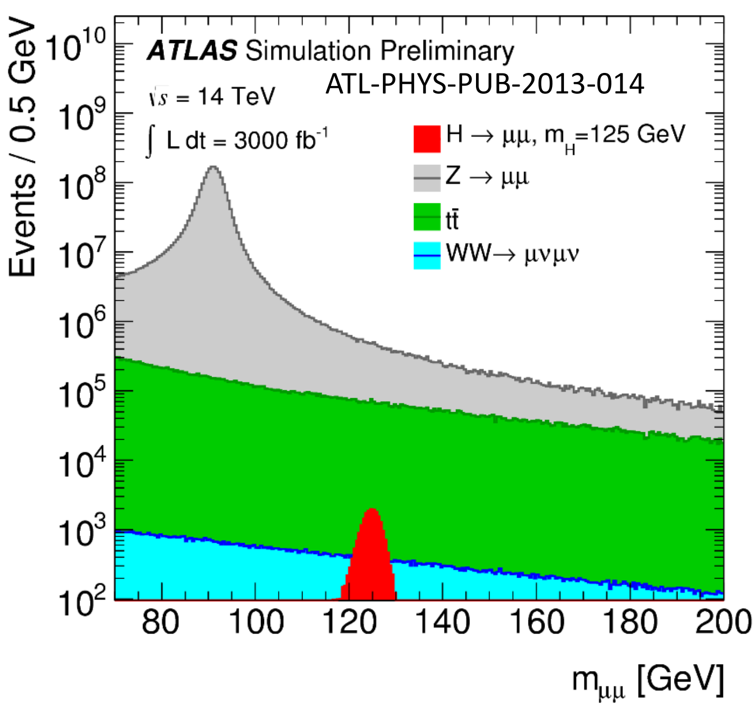
Toward Exabyte

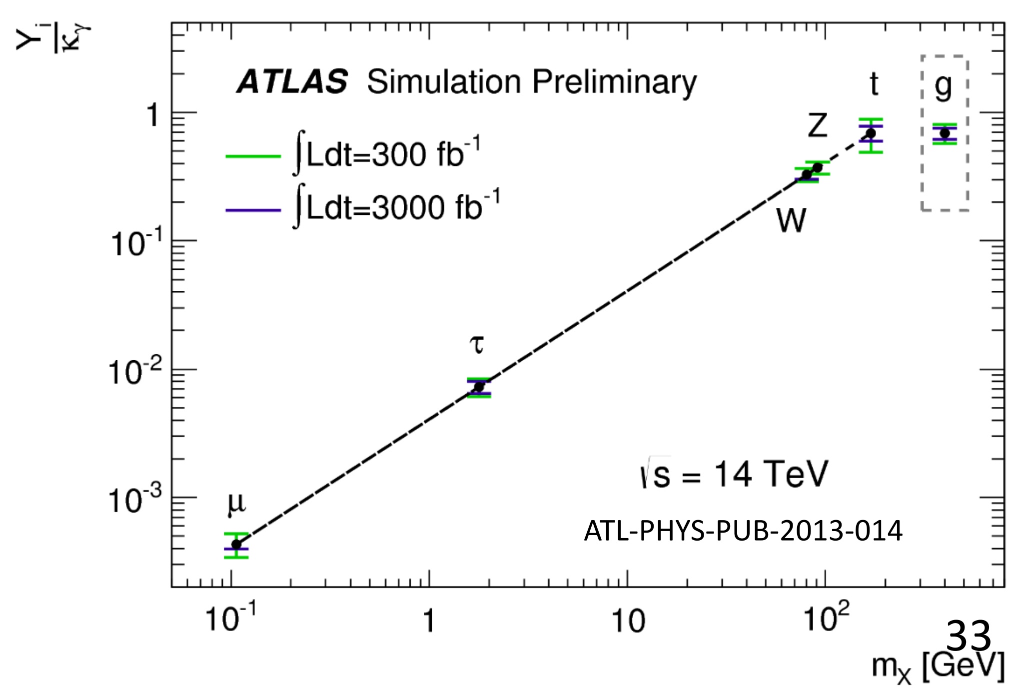
- LHC Upgrade plans in 10 years
  - 10 times higher luminosity LHC (HL-LHC)
  - 10 times more events
  - 10 times more complex data
- Strategy to handle 100 times more data
  - Will ‘Moore’s law’ work?
  - Network will be the key
  - Machine learning helps a lot
  - A new architecture for the distributed analysis infrastructure

LHC Upgrade plans in 10 years

- High Luminosity LHC
- Accumulate 3,000fb^{-1} data (30 times more)

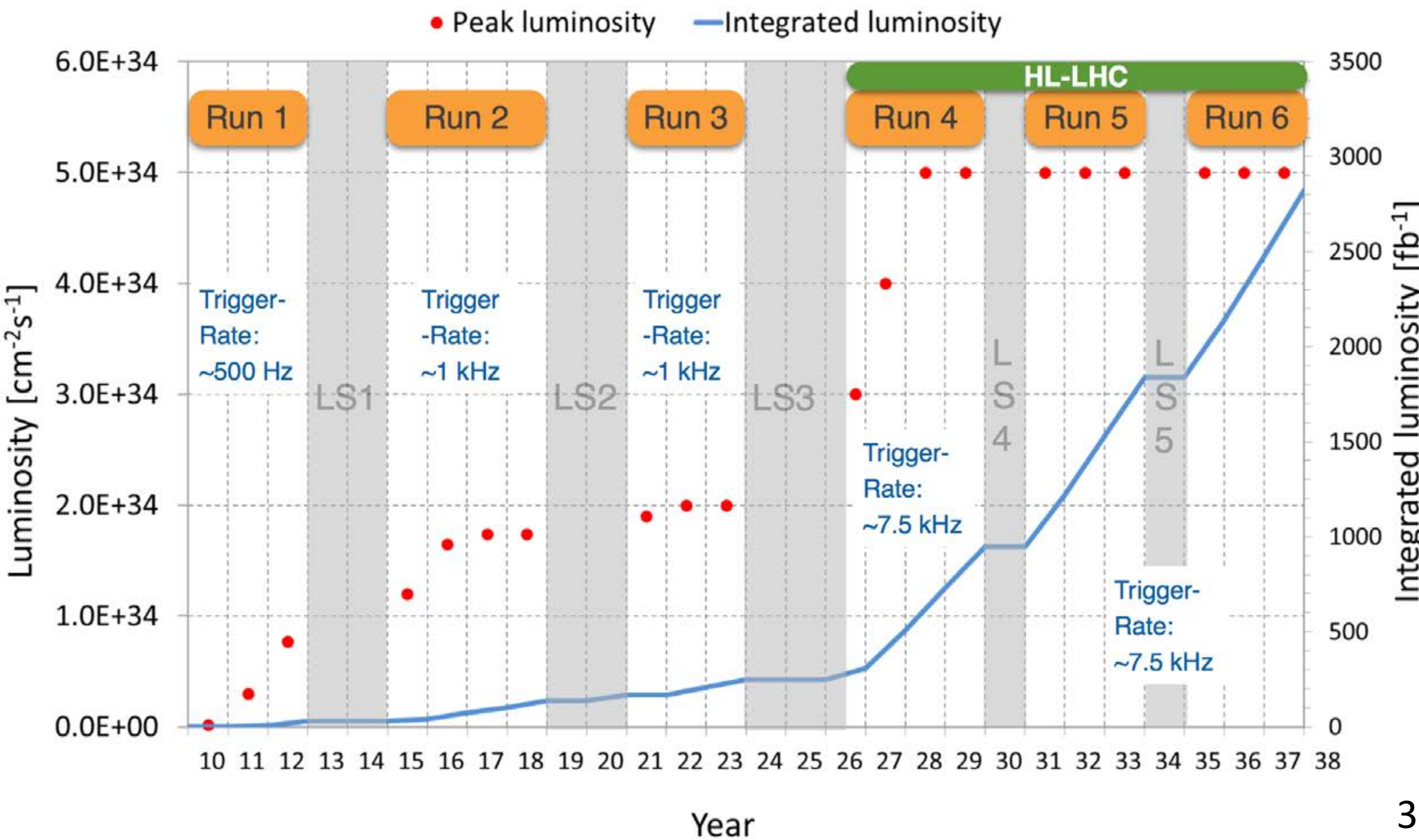






High Luminosity LHC (HL-LHC)

The HiLumi LHC Design Study is included in the High Luminosity LHC project and is partly funded by the European Commission within the Framework Programme 7 Capacities Specific Programme, Grant Agreement 284404

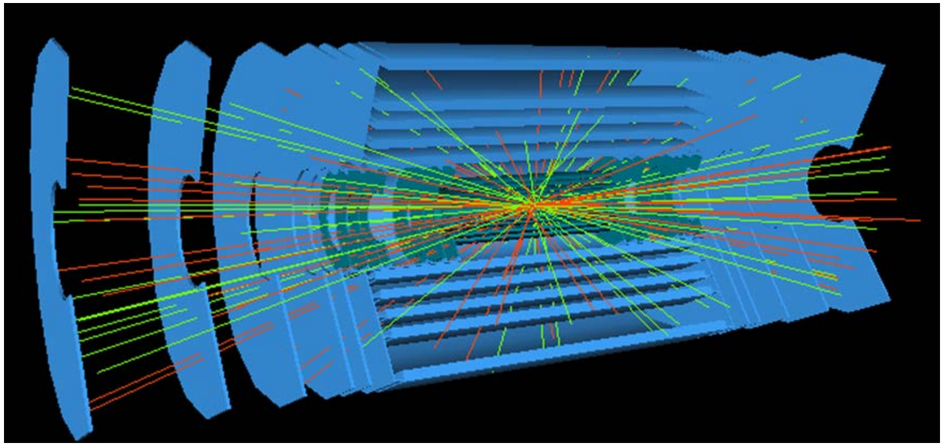


10 times more complex data

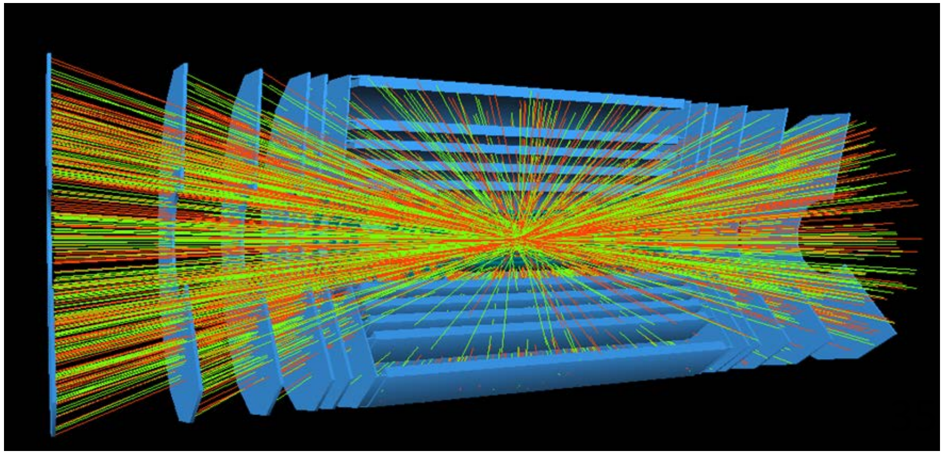
10 times higher luminosity means

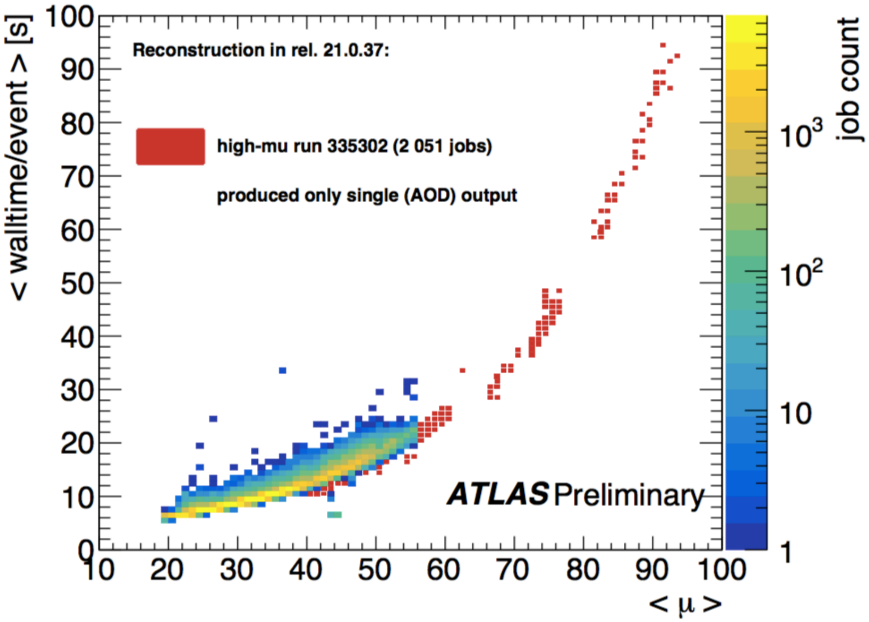
- 10 times more events
- 10 times more complex event data

~23 collisions per crossing ($5x10^{33}cm^{-2}s^{-1}$)



~230 collisions per crossing ($5x10^{34}cm^{-2}s^{-1}$)

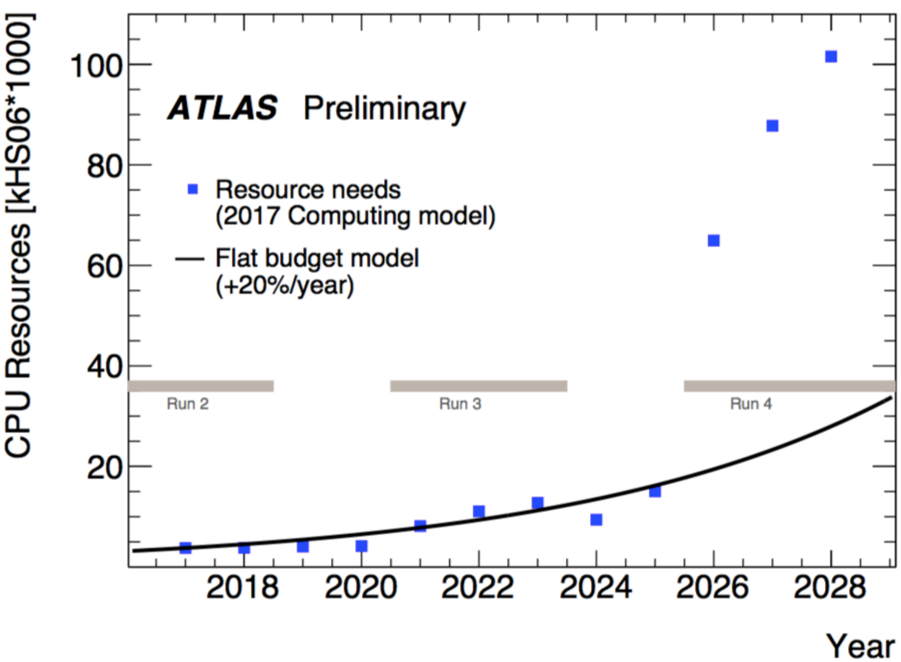


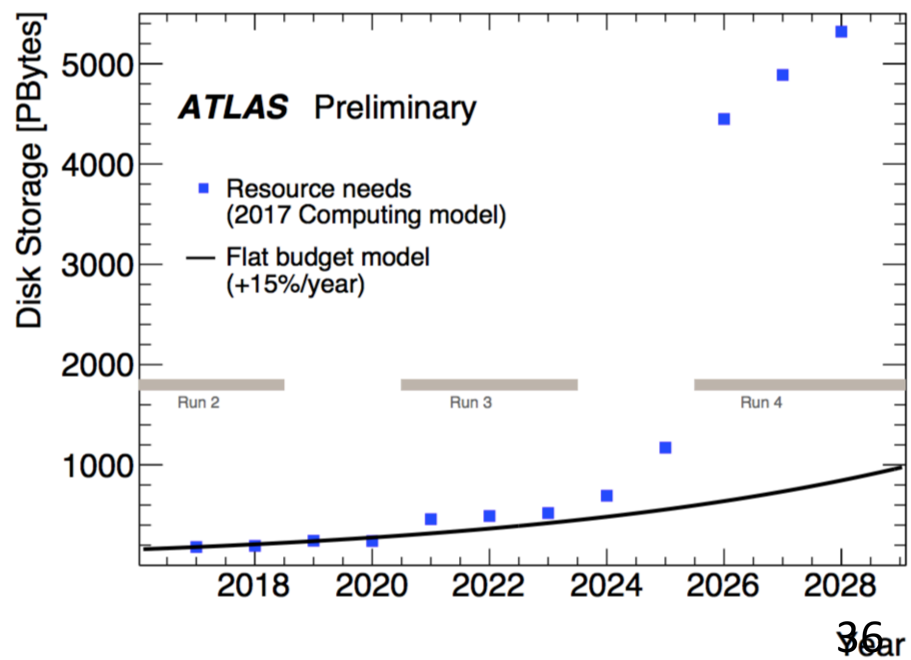


Strategy to handle 100 times more data

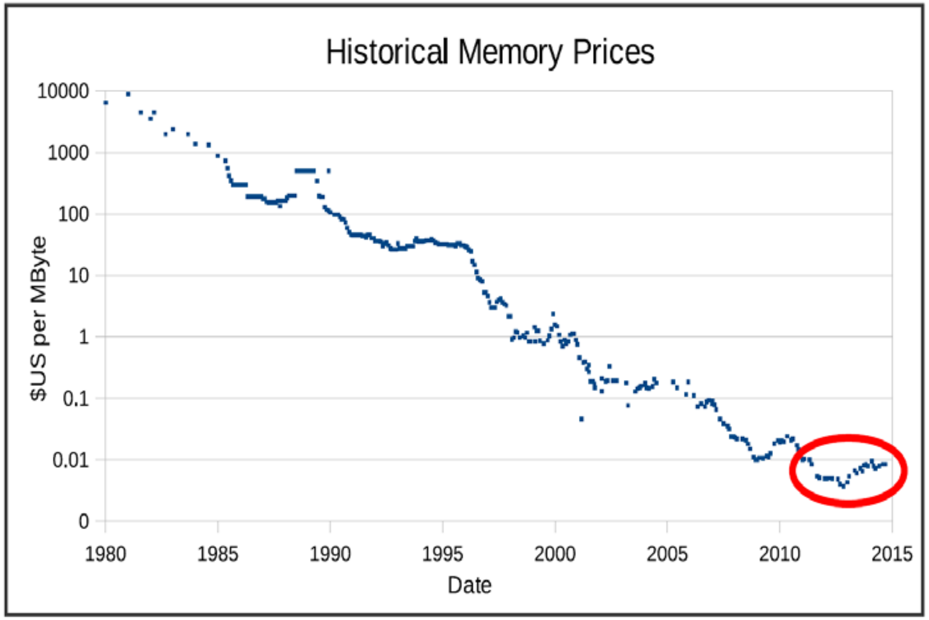
- Flat budget model ~ expected improve of 20%/year
- Around 10 times difference between requirements and flat budget model expectation

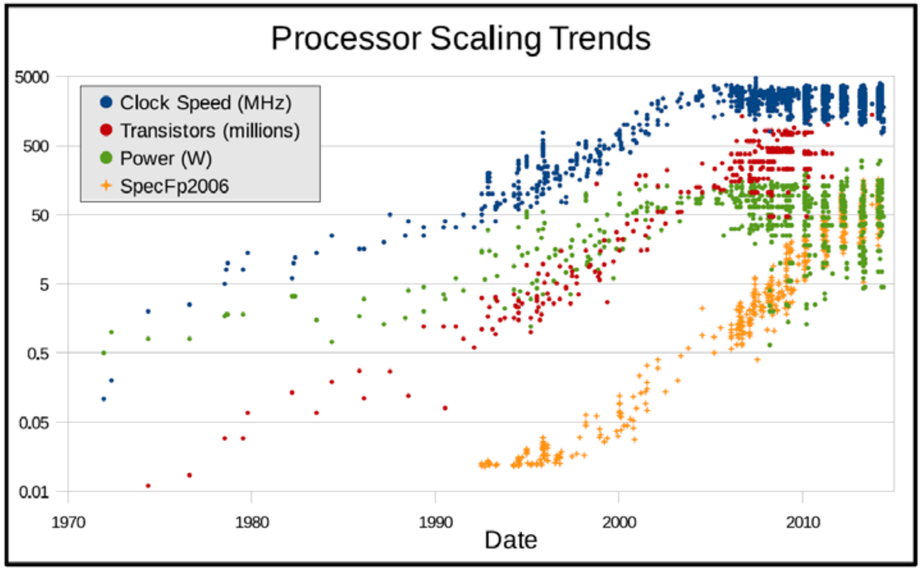
Need a breakthrough!

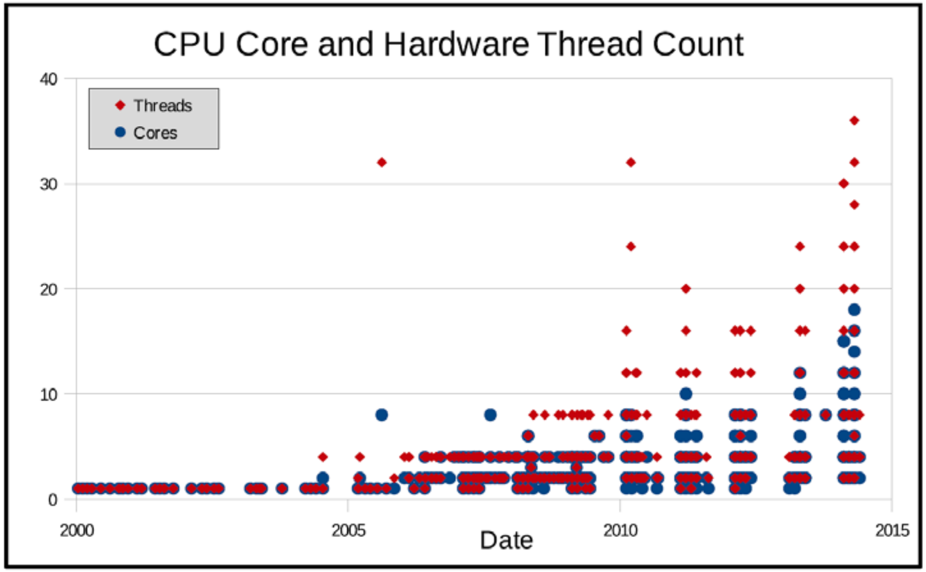




Will ‘Moore’s law’ work?







“Development of a Next Generation Concurrent Framework for the ATLAS Experiment,” P. Calafiura et al 2015 J. Phys.: Conf. Ser. 664 072031

Network will be the key

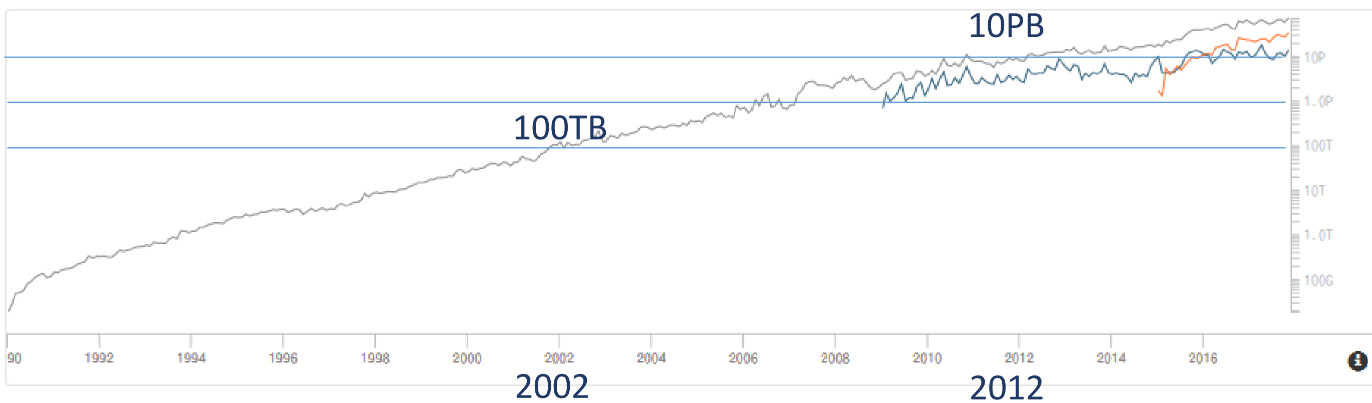
ESnet ENERGY SCIENCES NETWORK

Fasterdata | MyESnet | ESnet

HOME »

Traffic Volume

— Total Traffic — OSCARS — LHCONE



Options

MONTH

December 2017

TIME RANGE

All

VIEW

Overlay Summary

Interface Detail

SCALE TYPE

Log

Linear

December 2017

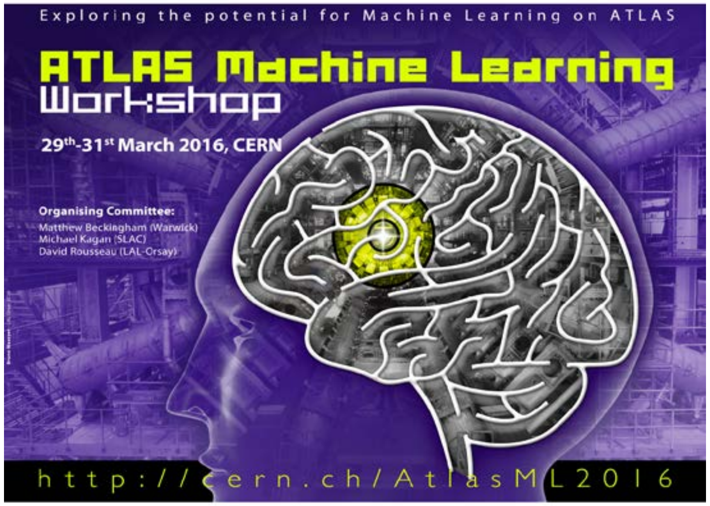
	Bytes	Percent of Total	One Month Change	One Year Change
OSCARS	11.48PB	14.8%	-17.2%	-9.53%
LHCONE	37.33PB	48.0%	+13.3%	+62.9%
Normal traffic	28.91PB	37.2%	+15.9%	+2.37%
Total	77.72PB		+8.30%	+21.7%

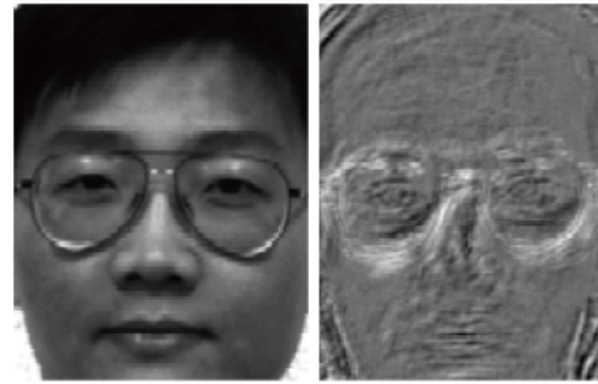
~30GB/s sustained

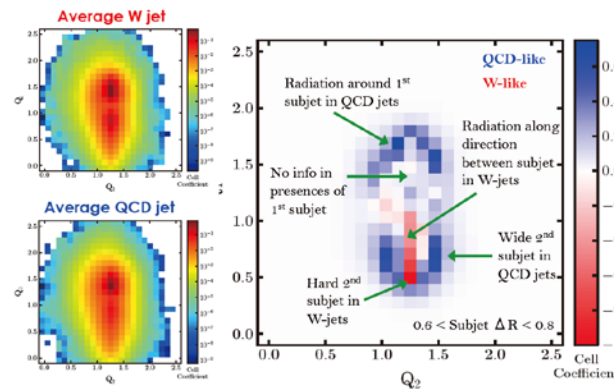
https://my.es.net/traffic-volume

10 times increase every 4.5 years

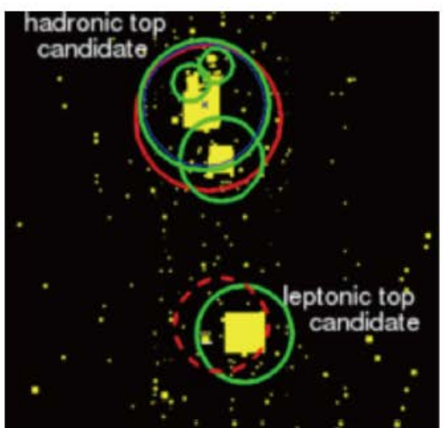
Machine learning will help a lot

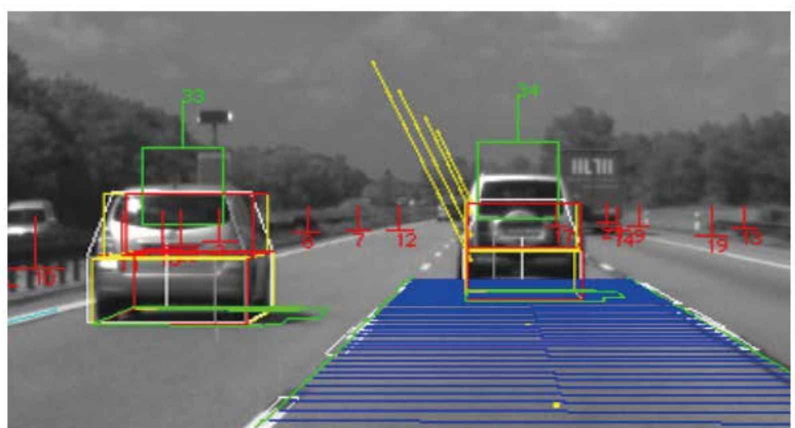




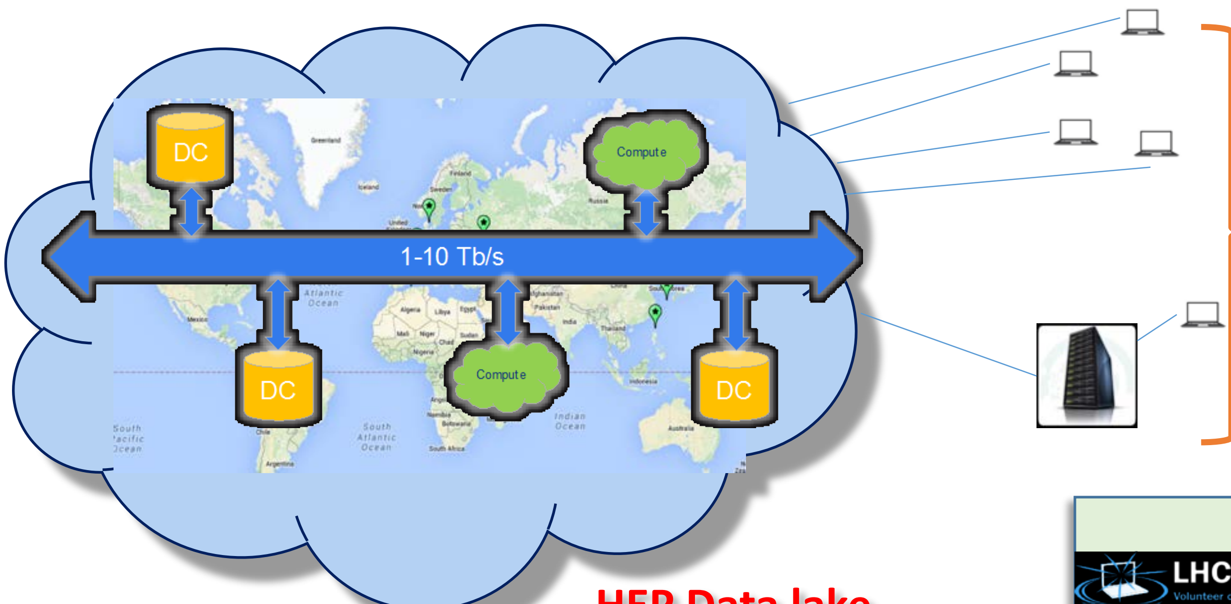


Computer Vision and Jet Physics:
Michael Kagan, Ben Nachman,
Ariel Schwartzman, Luke De Oliveira
SLAC, Stanford University





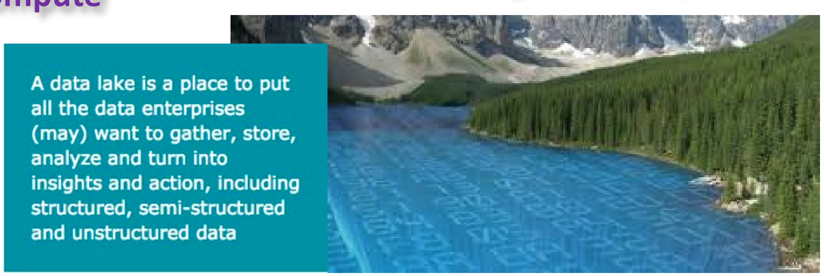
Possible Model for future HEP computing infrastructure

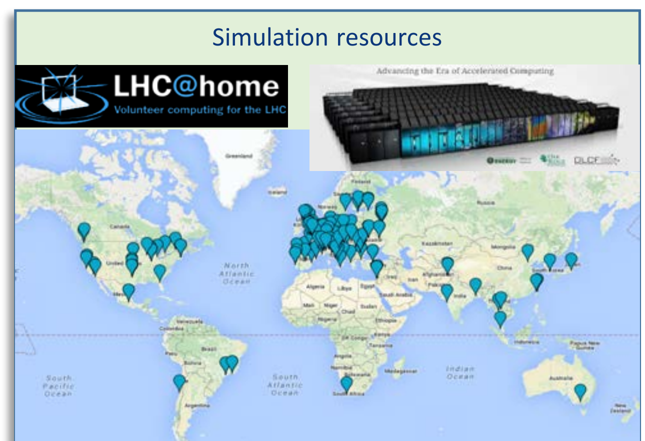


Cloud users:
Analysis

HEP Data cloud
Storage and compute

HEP Data lake
Storage and compute





Summary

- After 10 years of preparation, our computing grid started operation
  - Deployed to 150 institutes from 40 countries
  - Contributed to the discovery of Higgs particles
- The system has been evolving during 10 years' run
  - More scalable, robust, flexible, automatic, user-friendly
  - Expanding to cloud computing, HPC and volunteer computing
- More challenges to come in the next 10 years
  - 100 times more data to be managed
  - Linear extrapolation does not work: a breakthrough is inevitable
  - Your suggestion is very, very welcome