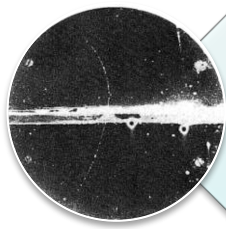
UiO : Department of Physics
University of Oslo

David Cameron, University of Oslo, ATLAS Experiment and NorduGrid Collaboration

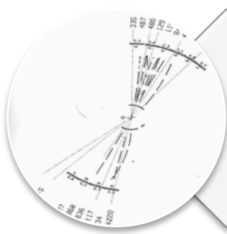
Grid Computing

The Changing Scale of Particle Physics



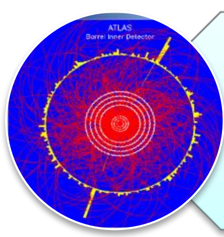
A discovery in 1930s

- ~2 scientists in 1 country
- pen-and-paper



A discovery in 1970s

- ~200 scientists in ~10 countries
- mainframes

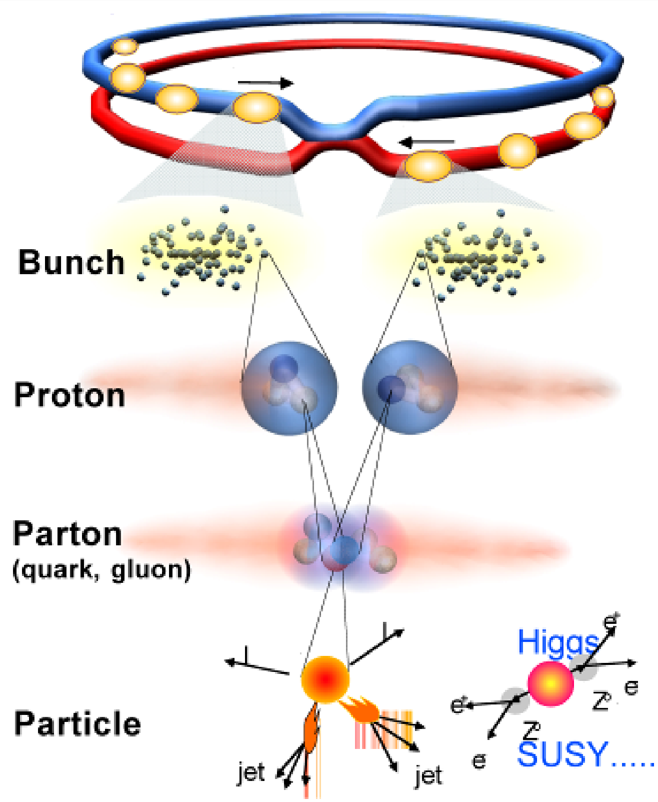


A discovery today

- ~2000 scientists in ~100 countries
- **<u>Distributed Computing</u>**

Graphics by O. Smirnova

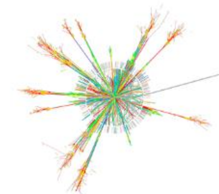
Event Collection in ATLAS

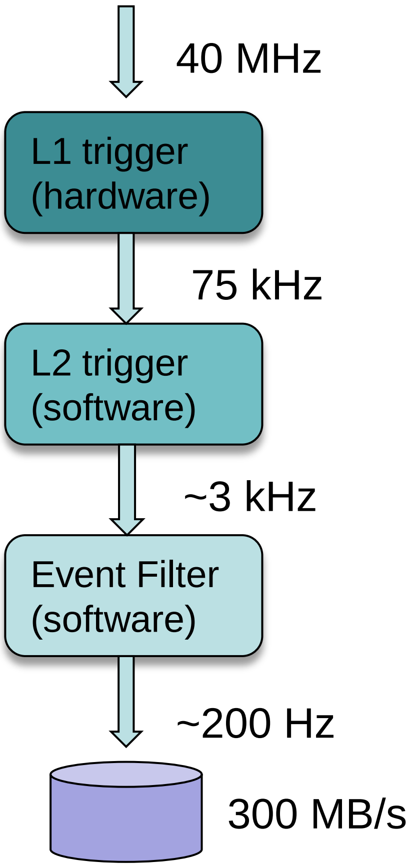


Proton-Proton	2835 bunch/beam
Protons/bunch	10^{11}
Beam energy	7 TeV (7×10^{12} eV)
Luminosity	10^{34} $cm^{-2} s^{-1}$
Crossing rate	40 MHz
Collisions rate ≈	10^{7} - 10^{9} Hz
New physics rate ≈	.00001 Hz

Event selection:
1 in 10,000,000,000,000

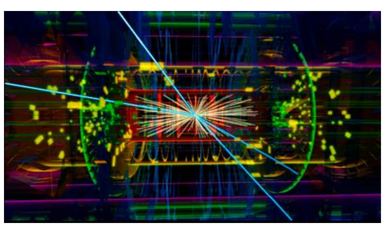
Graphic by CERN

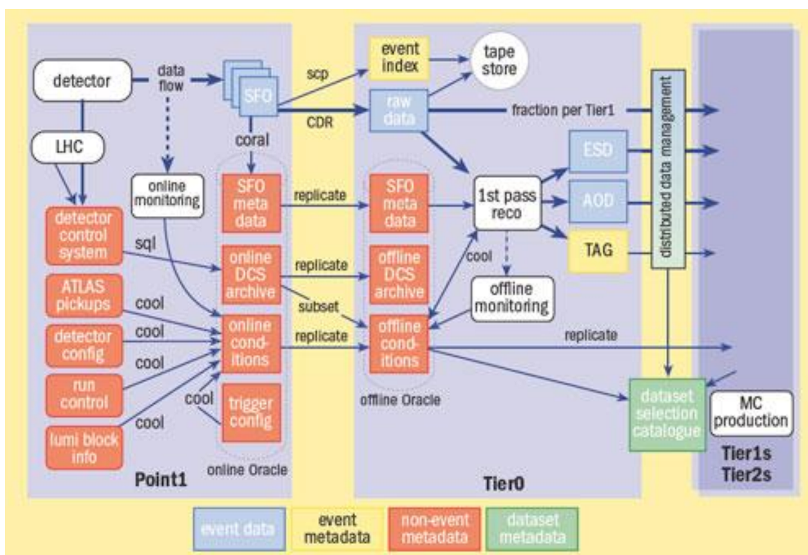




What is the data?

- C++ objects representing tracks, parts of detector etc, saved in files. Some geometry information in databases
- Data is reconstructed and reduced through various formats
  - RAW -> ESD -> AOD -> NTUP





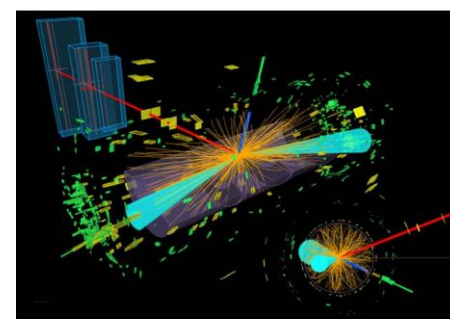


Figure from http://cerncourier.com/cws/article/cnl/34054

Big Data?

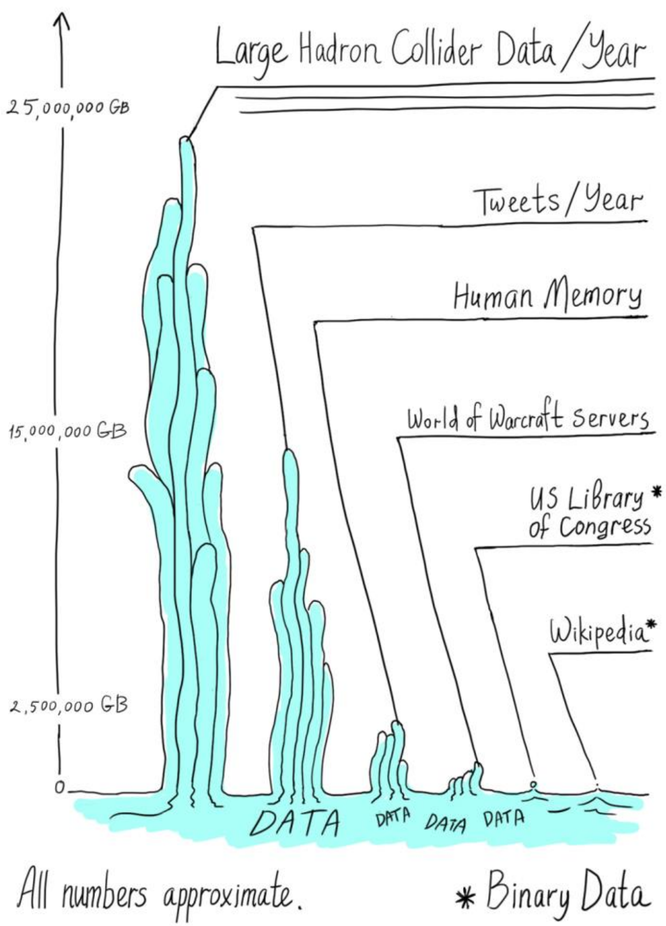
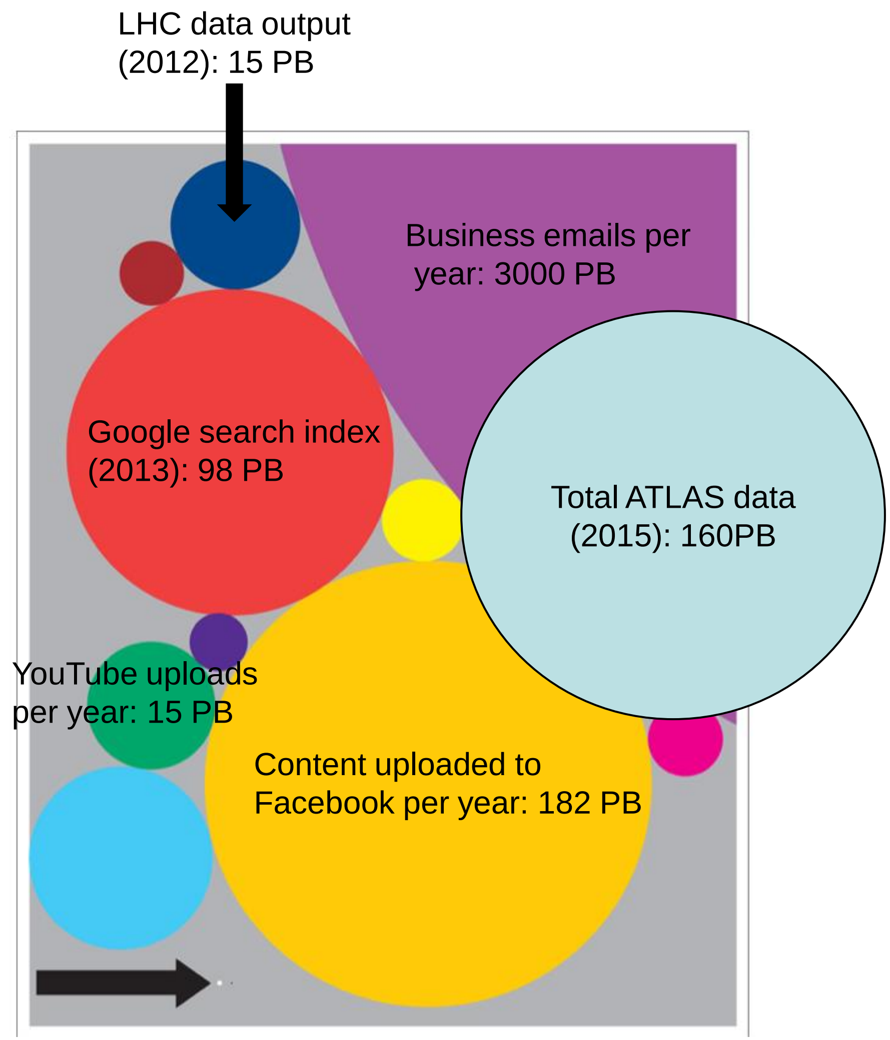


Illustration by Sandbox Studio, Chicago
Taken from http://www.symmetrymagazine.org/image/august-2012-big-data



WIRED.com © 2014 Condé Nast.
Taken from http://www.wired.com/2013/04/bigdata/

Do everything at CERN?

- All this requires (just for ATLAS)
  - 150,000 CPU constantly processing data
  - Storing 10s of PetaBytes (million GB) of data per year

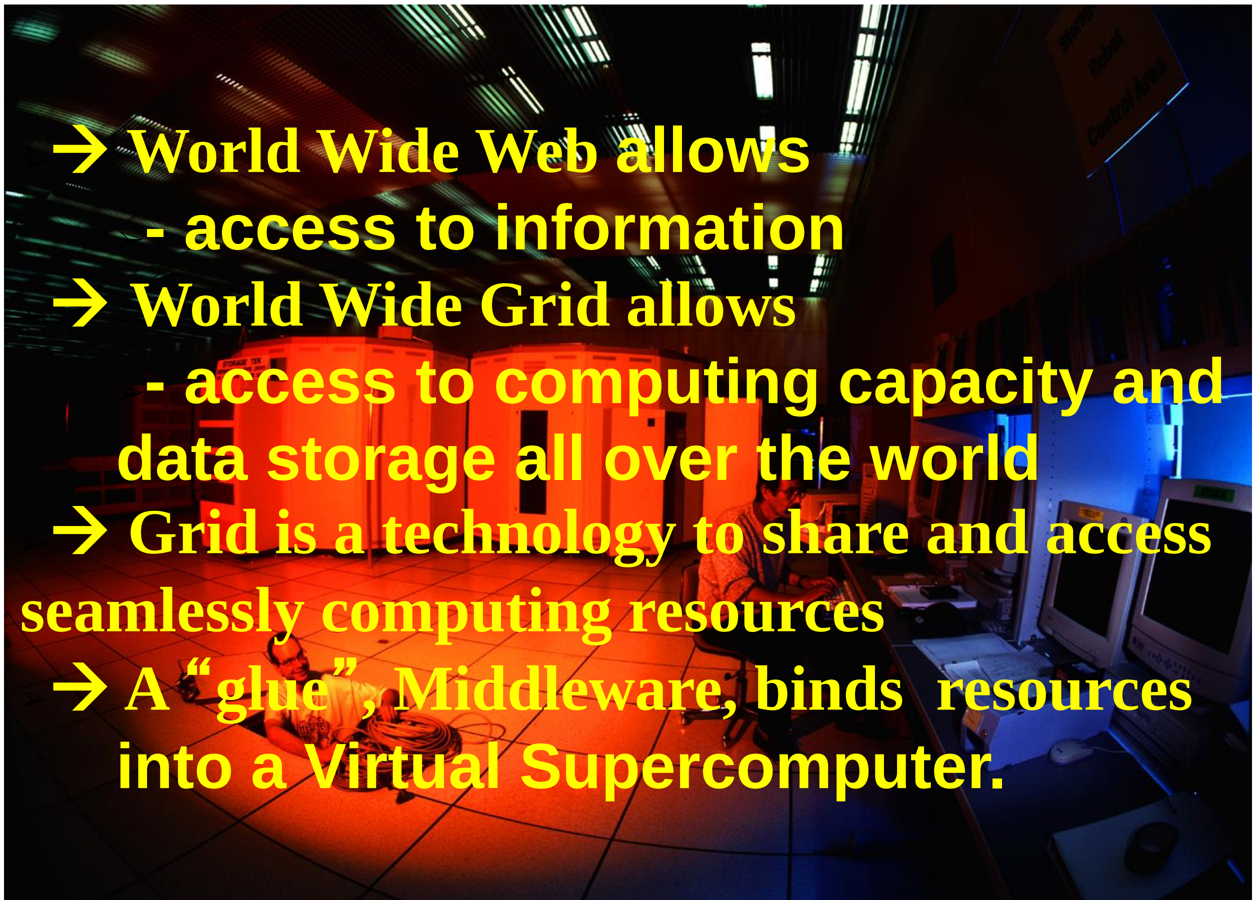
- CERN cannot physically handle this

→ **Grid Computing!**

Grid Computing

- Idea started in late ‘90s
- Like the electricity Grid
- Grid is a **technology** that enables optimized and secure access to widely distributed heterogeneous computing and storage facilities of different ownership

From WWWeb to WWGrid

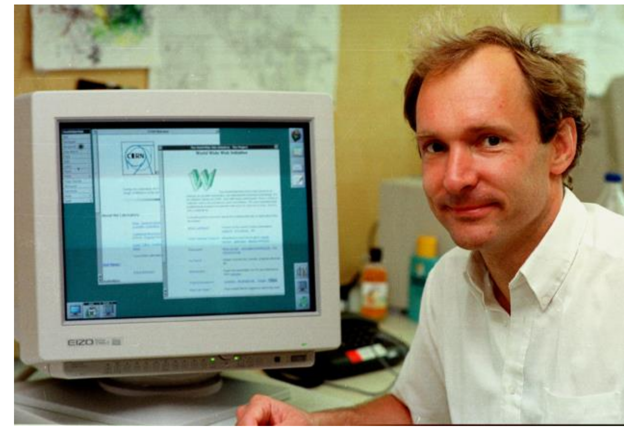


→ World Wide Web allows
- access to information

→ World Wide Grid allows
- access to computing capacity and data storage all over the world

→ Grid is a technology to share and access seamlessly computing resources

→ A "glue", Middleware, binds resources into a Virtual Supercomputer.

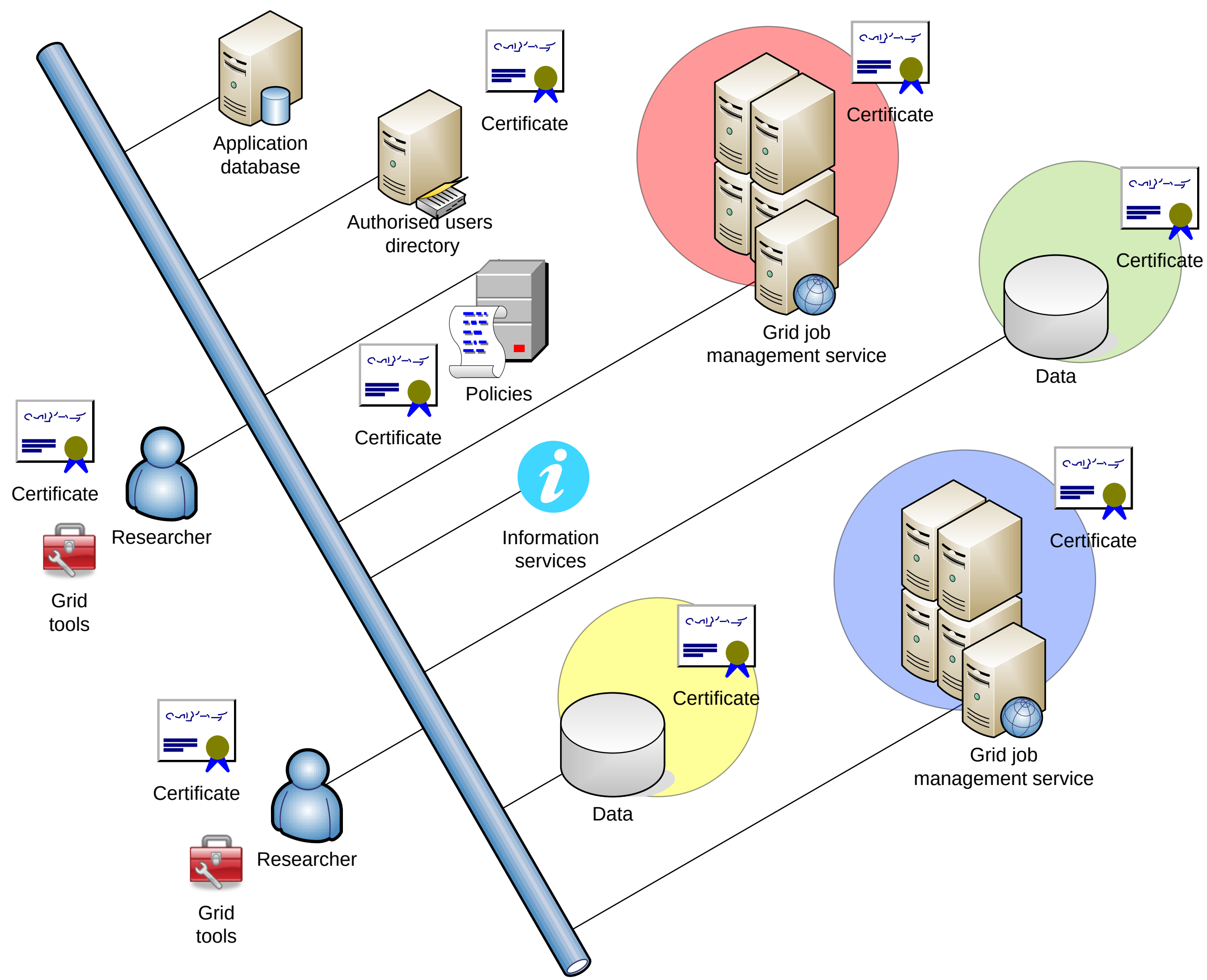


Slide by F. Ould-Saada

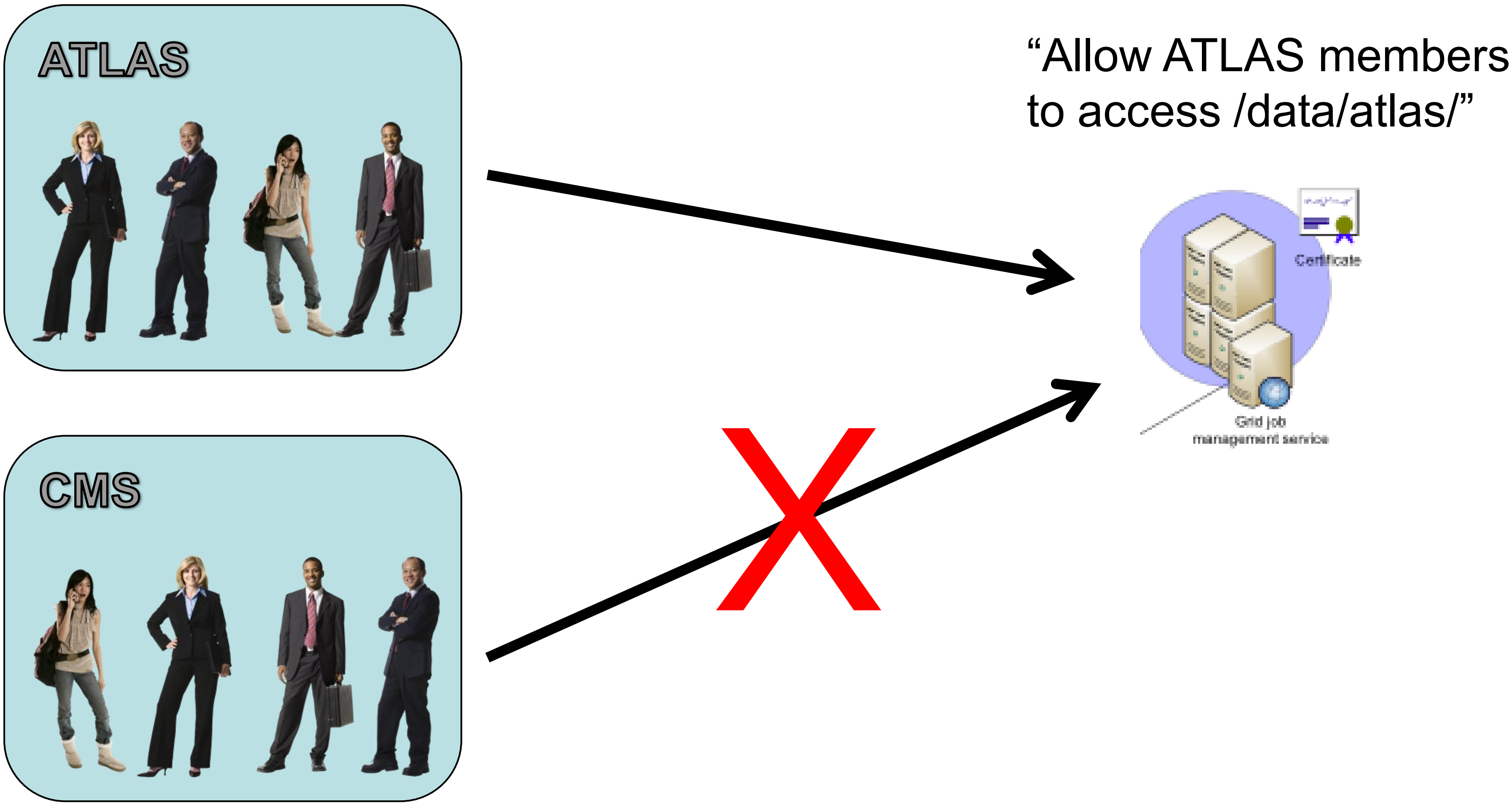
How to make a Grid

- The “Grid middleware” exposes heterogeneous resources to the Grid in a uniform interface
  - Computing Elements give access to CPUs
  - Storage Elements give access to data
  - Information systems describe the Grid
- How to allow access to resources?
  - Cannot give usernames and passwords for hundreds of sites to thousands of people!
  - Fundamental basis is X509-based cryptography

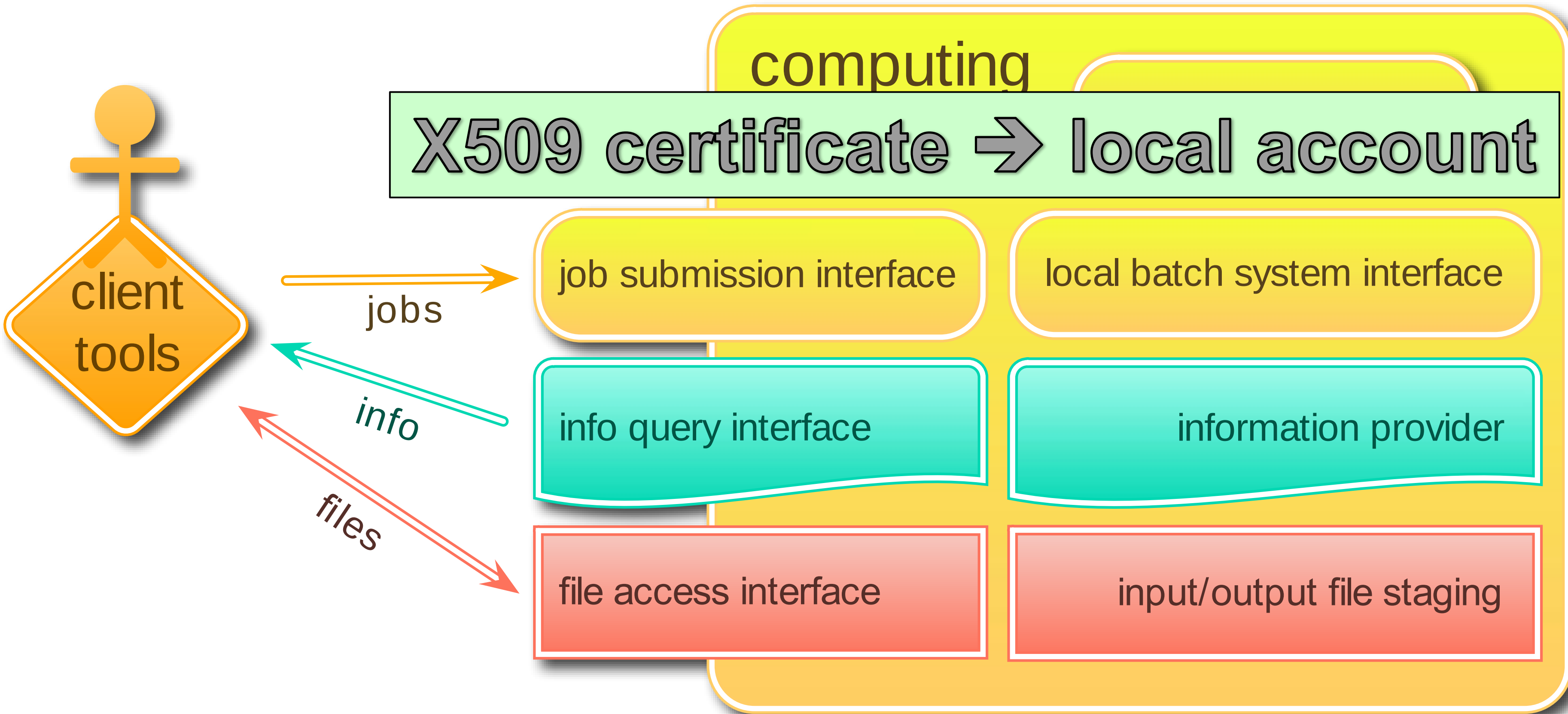
Grid Security (your "passport")



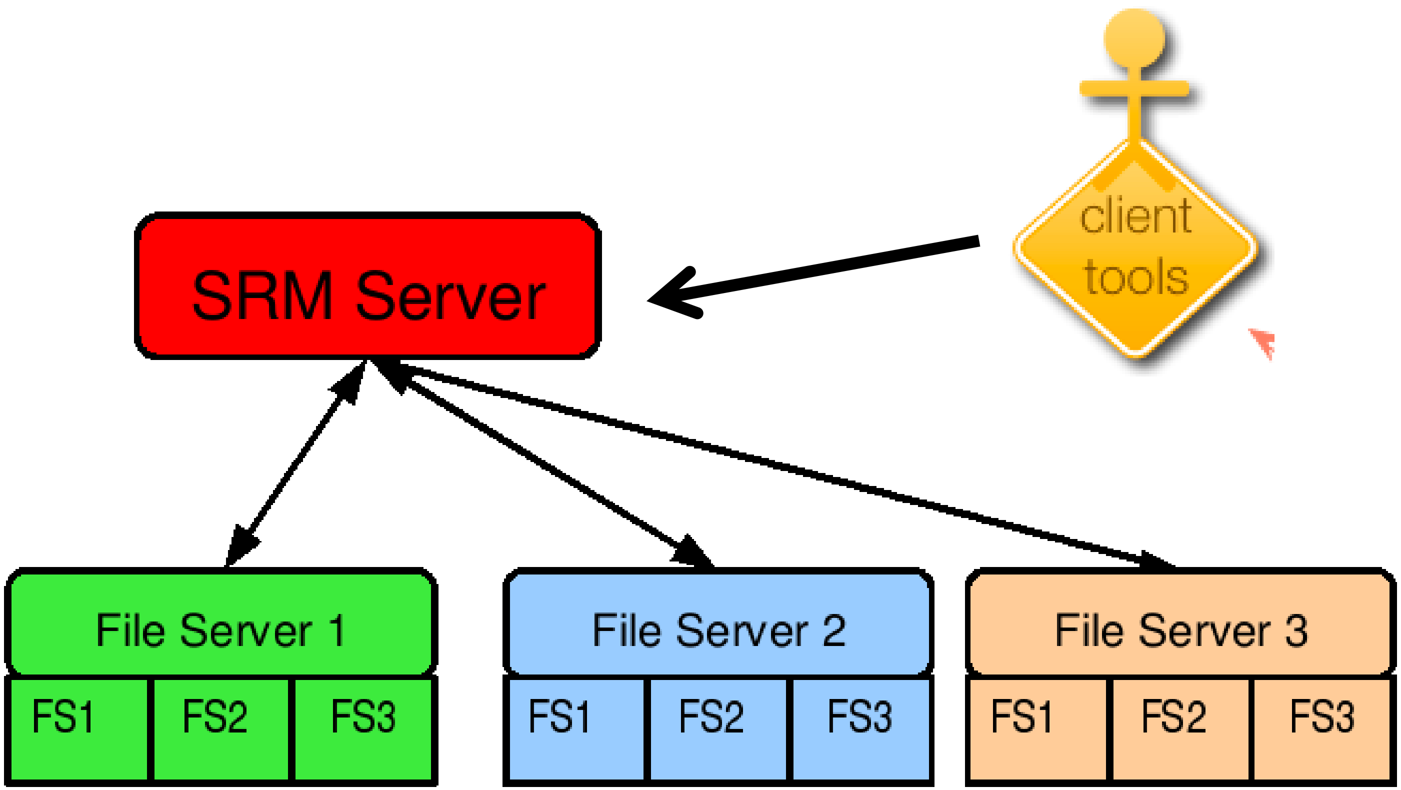
Virtual Organisations (your “visa”)



Computing Element in more detail

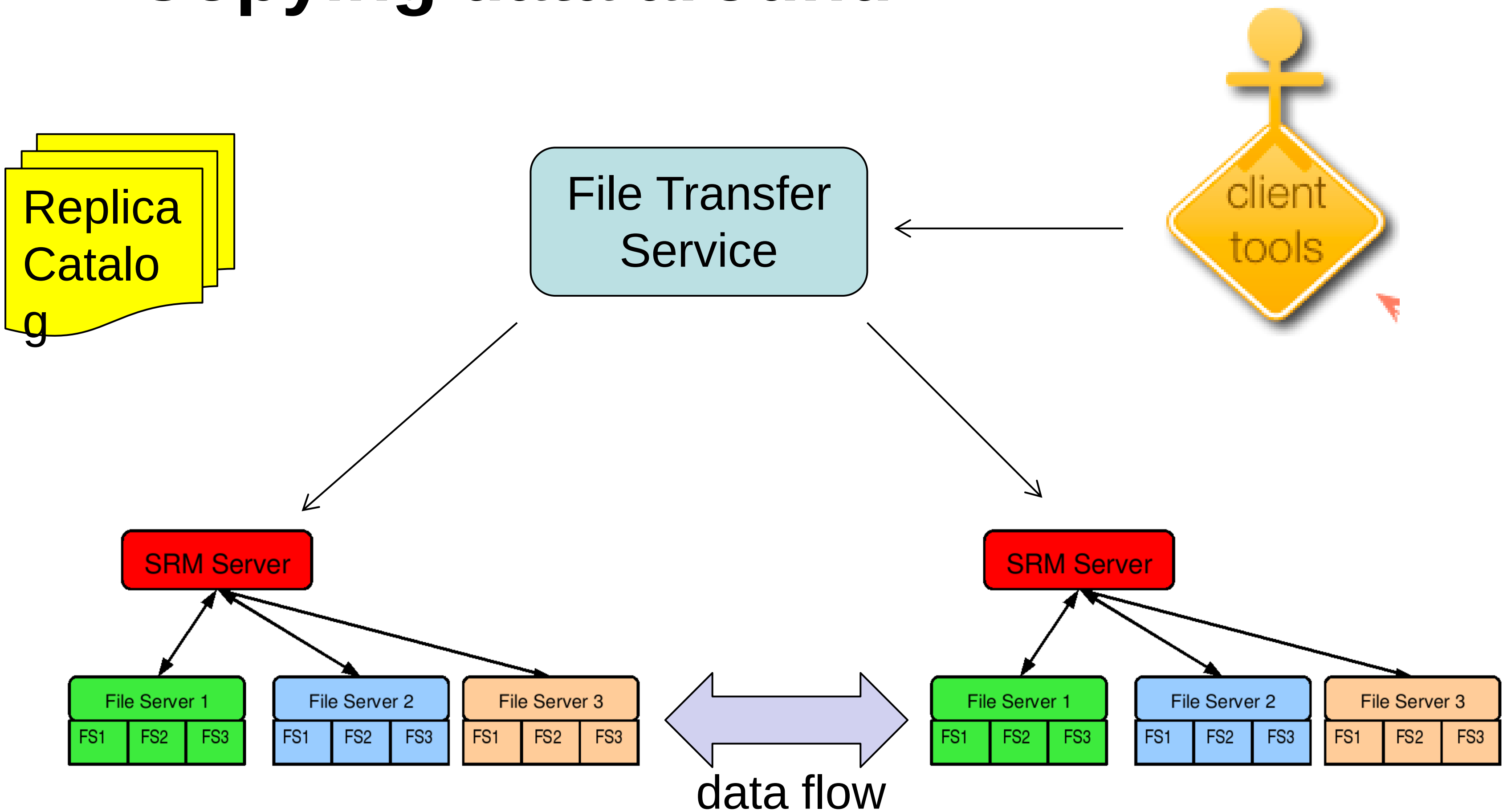


Storage Element in more detail



Graphics by G. Stewart

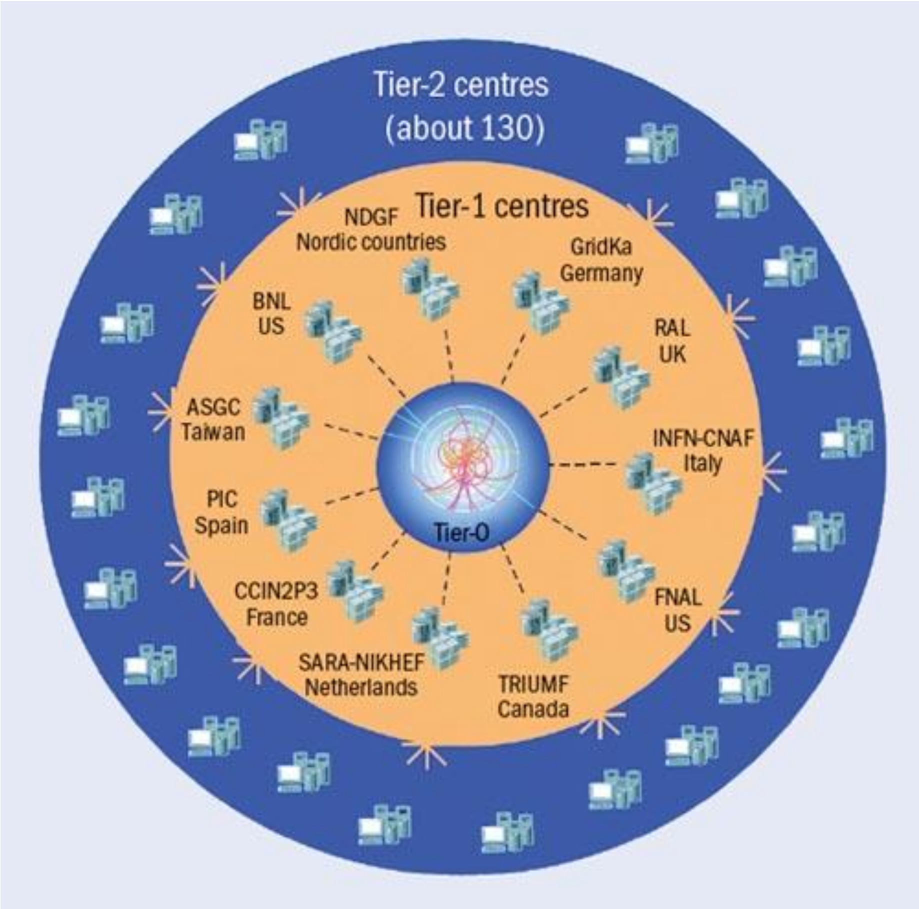
Copying data around



Many Grids!

Some Grid projects; originally byVicky White, FNAL

The (Worldwide) LHC Computing Grid



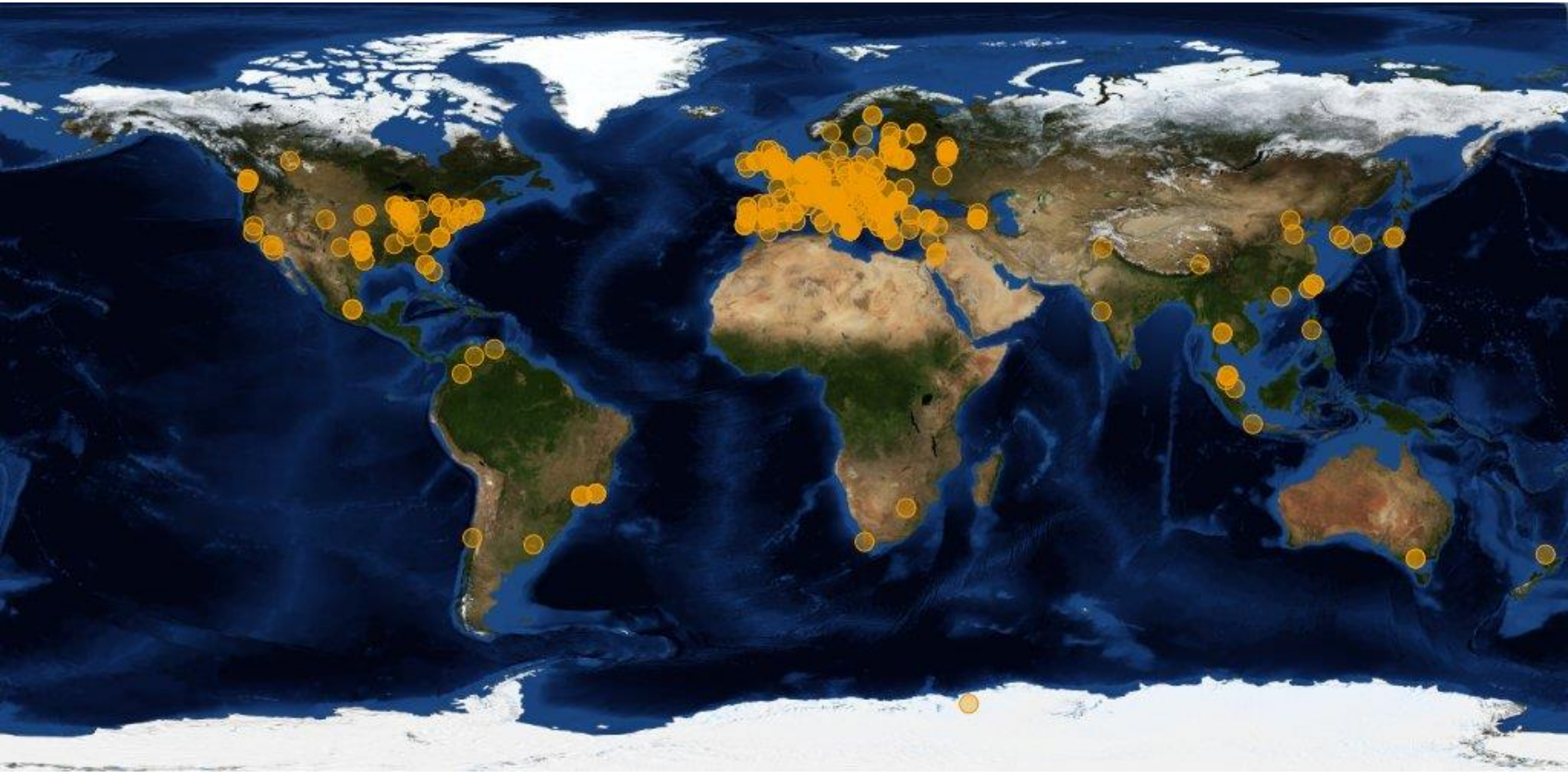
- 1 Tier 0: CERN
  - Data processing
- 11 Tier 1s
  - Simulation
  - Reprocessing
- ~130 Tier 2s
  - Simulation
  - User Analysis

- Total storage space: 539,357,056 GB
- Total processors available: 494,118

WLCG
Worldwide LHC Computing Grid

Numbers taken from http://wlcg-rebus.cern.ch/apps/capacities/pledge_comparison/ March 2015

WLCG Sites



NorduGrid

- Conceived in 2001 as Scandinavian Grid
  - UiO heavily involved in coordination and development
- Now 81 sites in 13 countries
- Software: Advanced Resource Connector (ARC)
  - Computing Element
  - (Basic) Storage Element
  - Information System
- Scandinavian design principles: clean and simple!

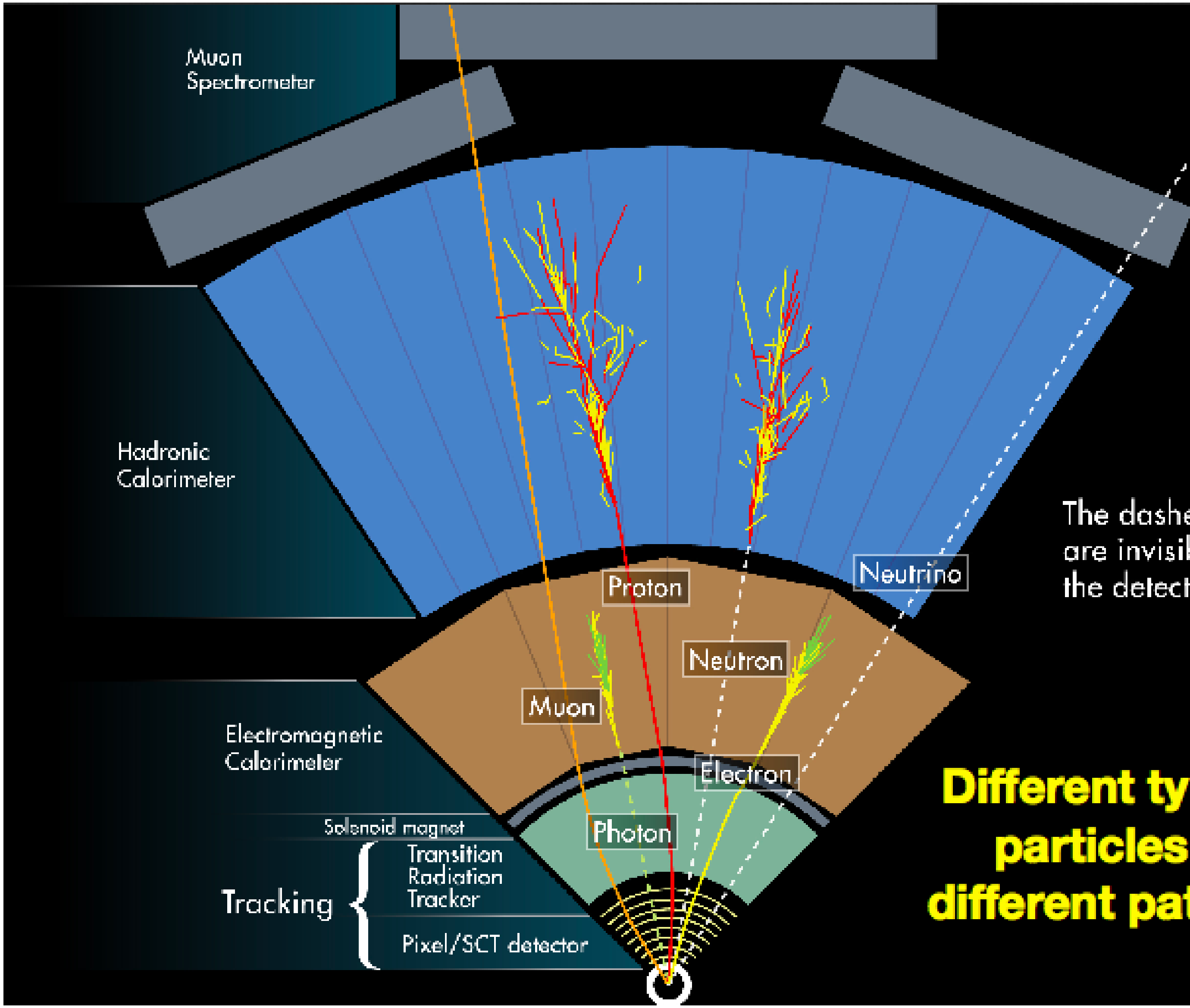
NorduGrid Monitor

Country	Site	CPUs	Load (processes: Grid+local)	Queueing
China	BOINC Cluster	20	0+0	0+0
Denmark	Steno Tier 1 (DCSC/KU)	6088	832+3703	450+0
	Steno Tier 3 (DCSC/KU)	6088	0+4533	0+0
Estonia	cream3 (T2_Estonia)	5076	27+508	0+0
	cream4 (T2_Estonia)	5076	52+733	4+0
	EENet	392	0+78	0+0
Finland	Aesyle (FGI)	72	72+0	7+0
	Alcyone (CMS)	892	2+621	382+0
	Alcyone (FGI)	892	164+459	27+0
	Asterope (FGI)	192	144+0	7+0
	Celaeno (FGI)	448	424+10	41+0
	Dll HEP (CMS)	200	200+0	163+0
	Electra (FGI)	672	178+433	60+0
	Jade (HIP)	768	168+600	278+56
	Maia (FGI)	768	156+612	68+0
	Merope (FGI)	1612	100+979	10+0
	Pleione (FGI)	288	144+24	46+0
	Taygeta (FGI)	360	32+284	25+0
	Triton (FGI)	3816	563+1439	10+0
	Usva (CSC/FGI/test)	144	0+0	0+0
Germany	LRZ-C2PAP	4032	3038+0	272+0
	LRZ-LMU	800	535+150	65+0
	LRZ-LMU lcg-lrz-ce0	1484	1066+2	153+124
	LRZ-LMU lcg-lrz-ce3	1492	0+1376	0+446
	RZG ATLAS HYDRA	167848	0+152674	110+0
	wuppertalprod	3320	1829+1165	237+1191
Hungary	NIIFI SC	768	0+655	0+5
Latvia	IMCSUL	1	0+0	0+0
	RTUETF	160	0+0	0+0
Lithuania	VU-MIF-LCG2	1532	0+107	0+0
Norway	Abel C1(UiO/USIT)	10872	98+9175	219+1004
	Abel C2(UiO/USIT)	10872	0+9274	0+1201
	Abel C3(UiO/USIT)	10872	0+9274	222+979
	EPF (UiO/FI)	106	0+0	0+0
	fimm (BCCS/UiB)	928	0+0	2+0
Slovenia	Arctur-1	432	0+0 (queue inactive)	0+0
	Arnes	2244	1632+0	630+0
	atos	1417	0+1039	0+28
	CIPKeBiP	984	895+0	0+0
	SiGNET	2834	2202+0	225+0
	UNG	112	0+0	0+0
Sweden	Abisko (HPC2N)	15936	341+14736	57+0
	Alarik (SweGrid, Luna>	3776	314+2529	304+1
	Glenn (C3SE)	6112	0+5616	0+226
	Tintin (SweGrid, UPPM>	2624	128+2399	184+4075
Switzerland	Bern ce01 (UNIBE-LHEP)	1368	790+27	74+18
	Bern ce02 (UNIBE-LHEP)	752	464+0	67+0
	Bern UBELIX T3	2592	528+1560	51+13451
	Gordias at hepia	224	0+0	0+0
	Lugano PHOENIX T2	2520	7+2337	74+184
	Lugano PHOENIX T2	2520	5+2340	115+142
	WSL Grid Cluster	408	0+356	0+9839
UK	arc-ce01 (RAL-LCG2)	9262	3265+5746	596+0
	arc-ce02 (RAL-LCG2)	9262	1561+7448	650+0
	arc-ce03 (RAL-LCG2)	9262	1659+7353	563+0
	cetest01 (UKI-LT2-IC->	4	165+2989	29+3273
Ukraine	BITP ARC Training	384	0+49	0+0
	BITP Cluster	384	2+46	0+0
	CHIMERA	192	43+72	16+0
	DFTI Cluster	136	0+96	0+1
	IAP Cluster	12	0+1	0+0
	IAPMM Cluster	52	0+0	0+0
	ICMP Cluster	268	60+80	0+0
	ICYB SCIT-3	1176	0+338	23+-4
	IFBG Cluster	64	0+24	0+0
	ILTPE ARC UA	112	4+0	0+0
	IMATH Cluster	8	0+1	0+0
	IMBG ARC	24	0+0	36+0
	IMMSP Cluster	40	0+0	3+0
	IMP ARC CE	84	0+64	0+0
	IOP Cluster	80	0+66	1+0
	IPMS Cluster	24	0+0	0+0
	IRE Cluster	64	0+0	0+1
	ISMA cluster	516	0+373	14+112
	ISOFTS Cluster	8	0+0	0+7605
	KNU ARC	216	4+95	895+0
	KPI training cluster	72	0+0	0+0
	LNU Training Cluster	32	0+28	0+0
	MHI Cluster	120	0+0	0+0
	PIMEE ARC	24	0+0	0+0
	SRI cluster	4	0+0	0+0
TOTAL	*81 sites*	*327692*	*23893 + 256676*	*7465 + 43958*

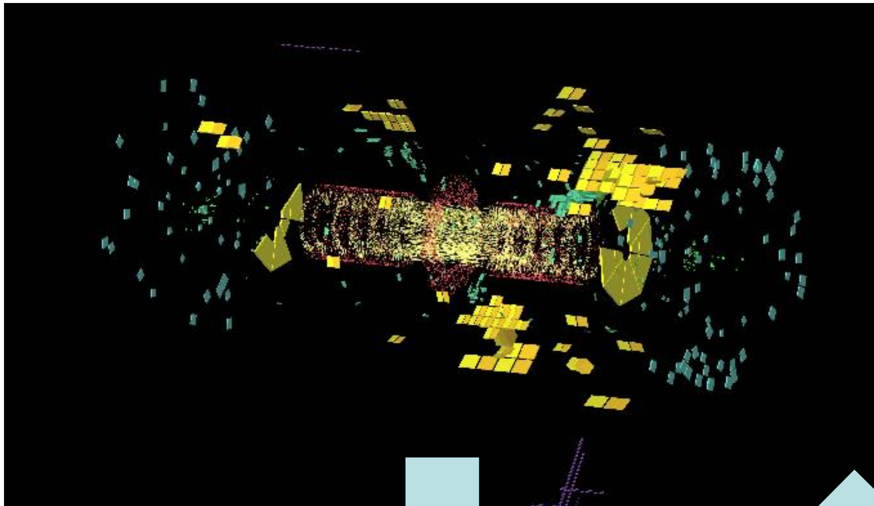
Sites: 81 Running jobs: 23893

ATLAS Data Processing

- Two main kinds of processing
  - Analysis of data
  - Simulation of data
    - Why?
    - At design phase to optimise the detector layout
    - In running phase to validate real data
    - The only way to know we have discovered something new
  - Simulation is the most CPU-intensive process in LHC experiments

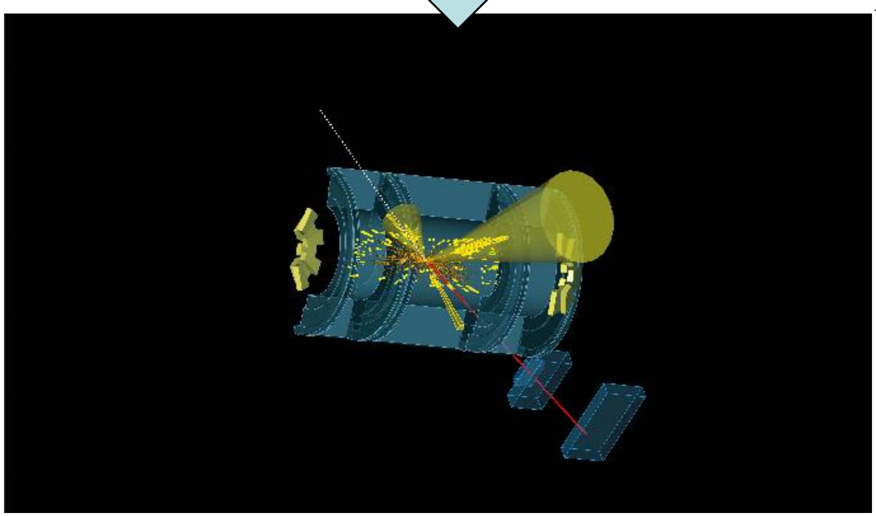


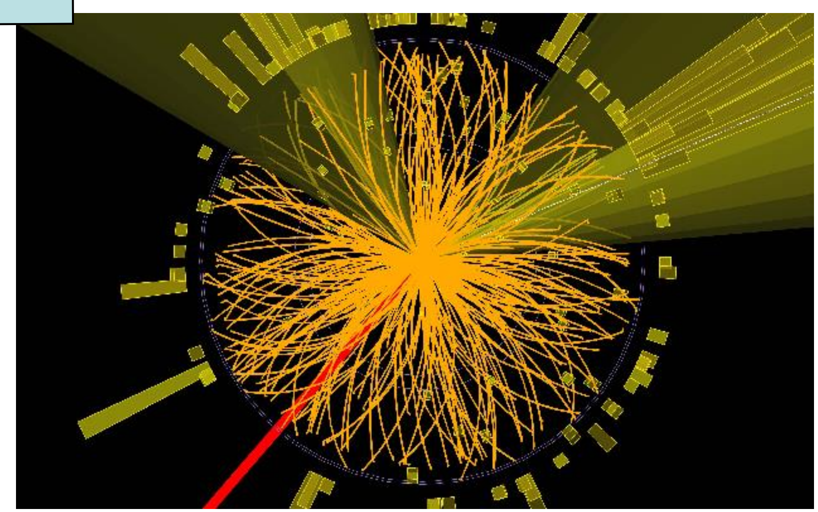
Different type of particles -> different patterns



Simulation steps:

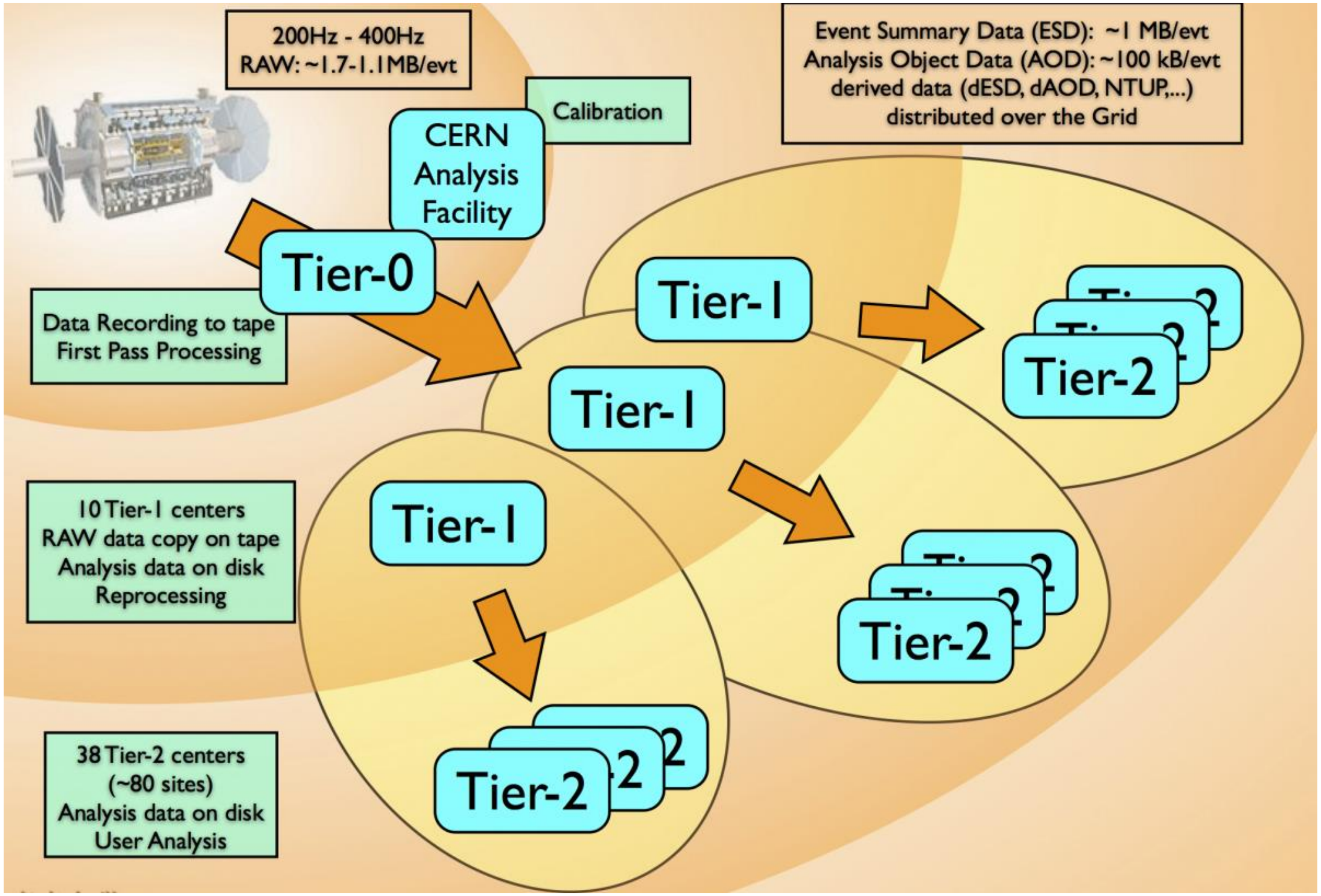
- Event generation
- Detector simulation
- Track reconstruction





UiO : Department of Physics
University of Oslo

ATLAS Computing Model



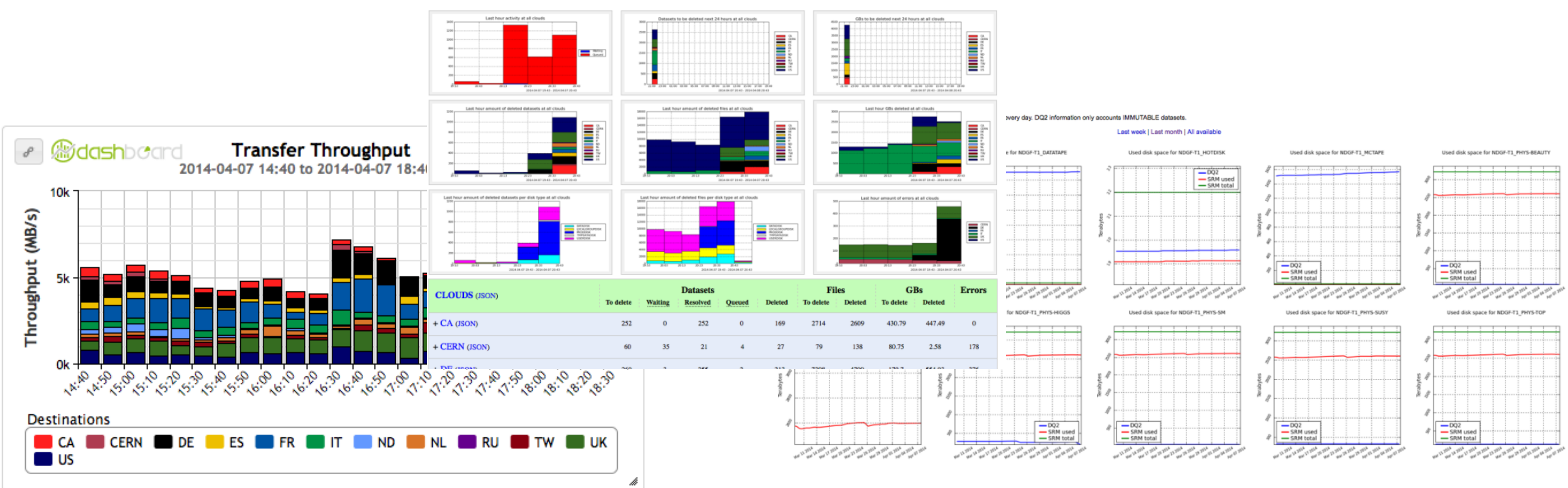
UiO : Department of Physics
University of Oslo

The ATLAS Grid(s)

- ATLAS has its own systems on top of the Grids
  - PanDA (Production and Data Analysis) for job management
  - Rucio for data management

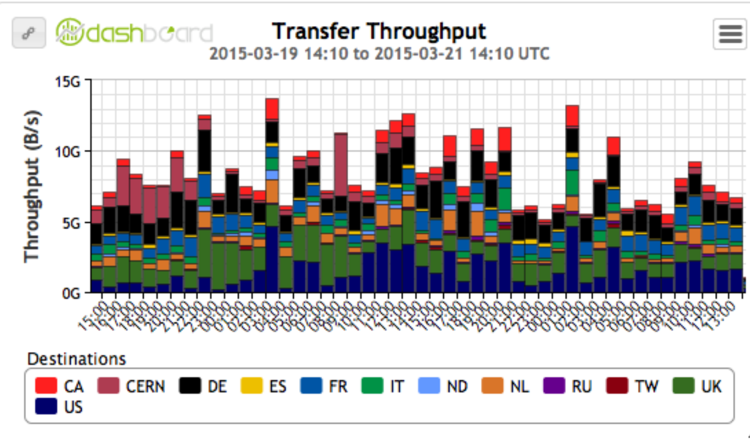
Rucio

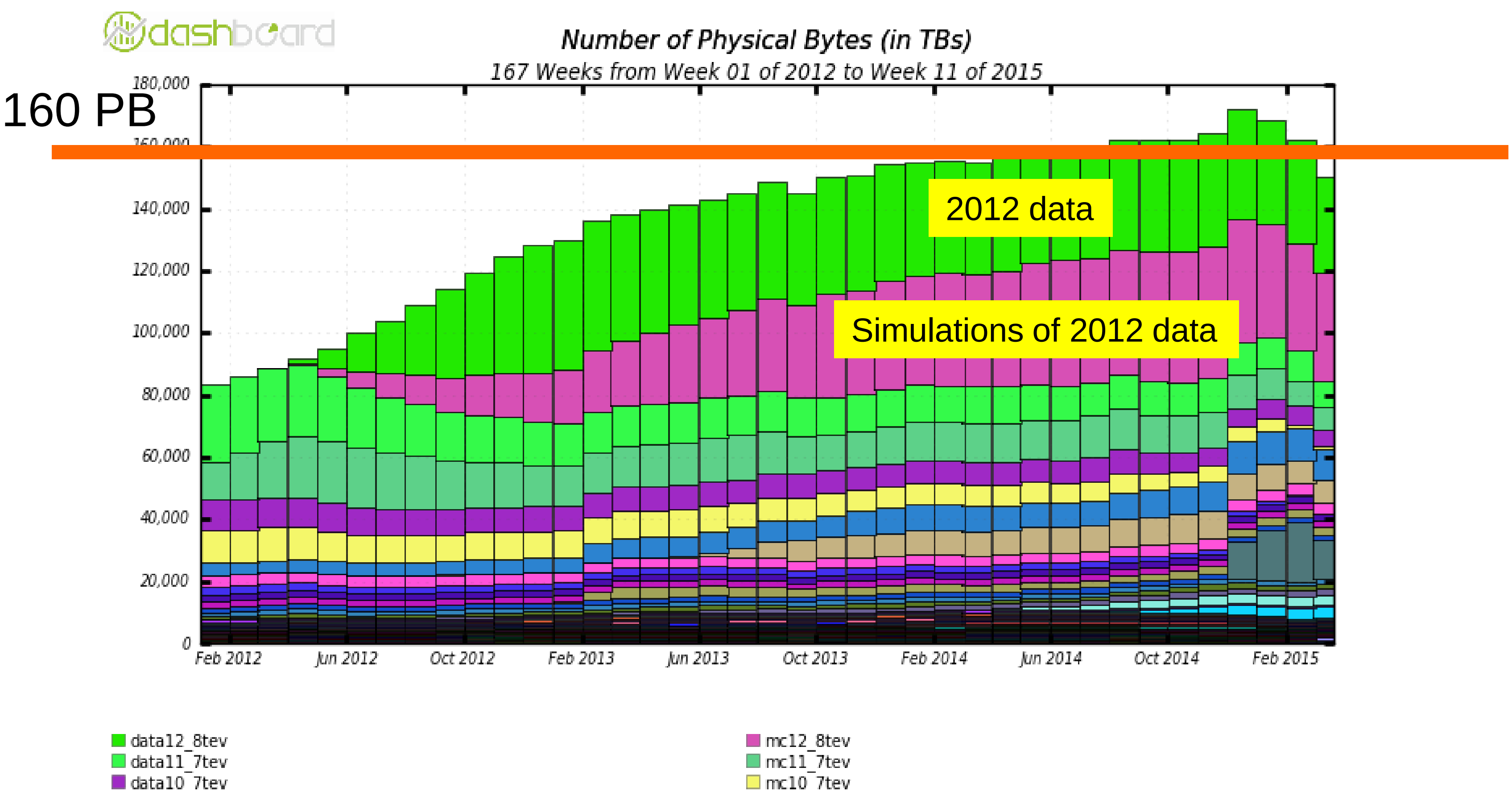
- A data management system to implement the ATLAS computing model
  - A dataset catalog and transfer system, and more
  - deletion, quota management, consistency, accounting, monitoring, end-user tools, …



It's a lot of data

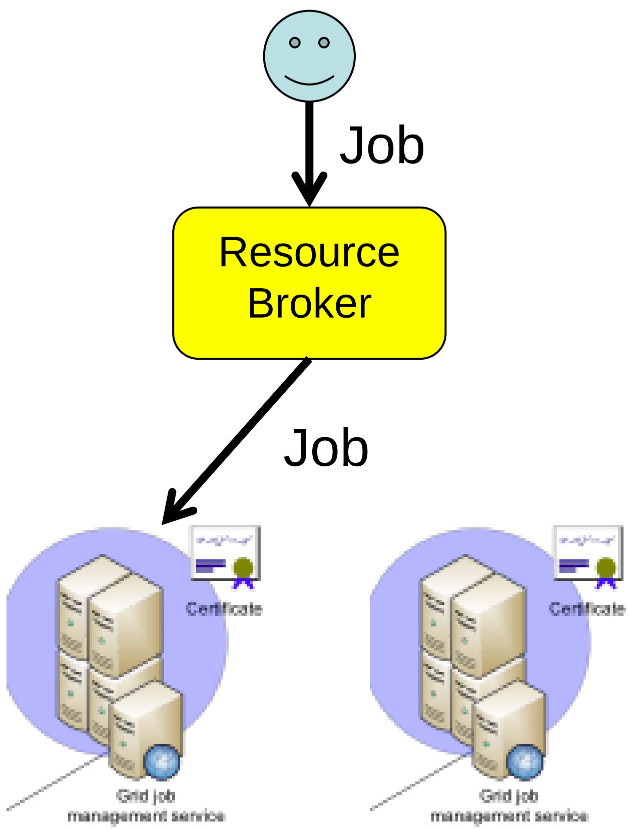
Max Telenor broadband speed: 6MB/s
Average ATLAS traffic: 10GB/s



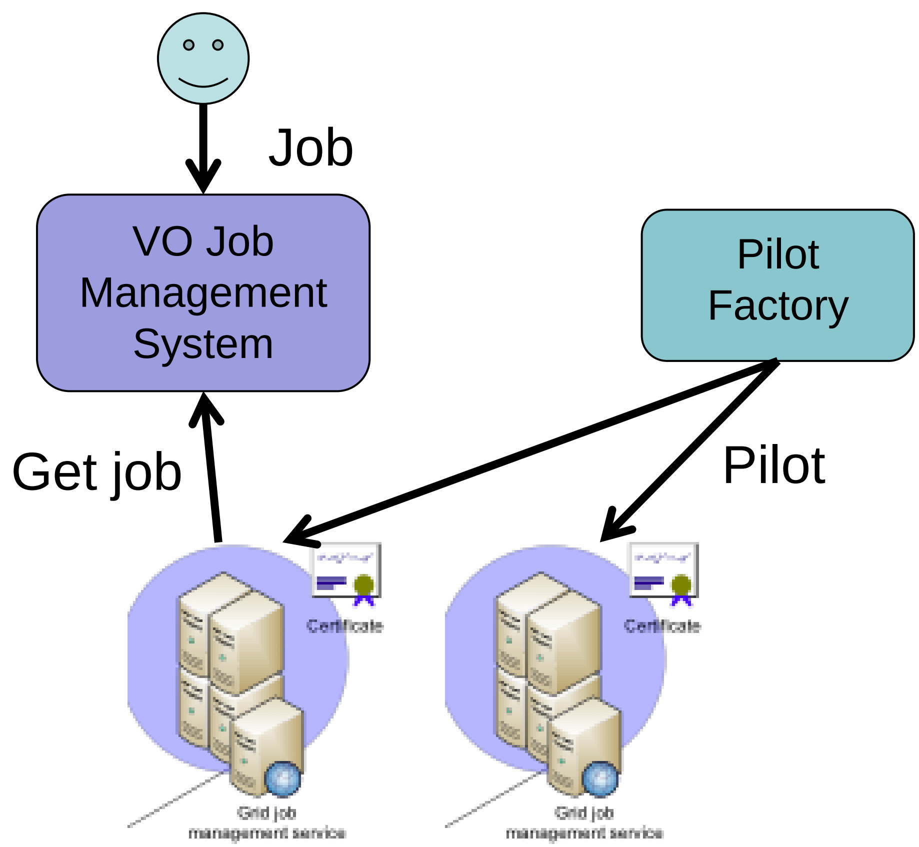


Grid job management

Classic "push" model



Pilot "pull" model



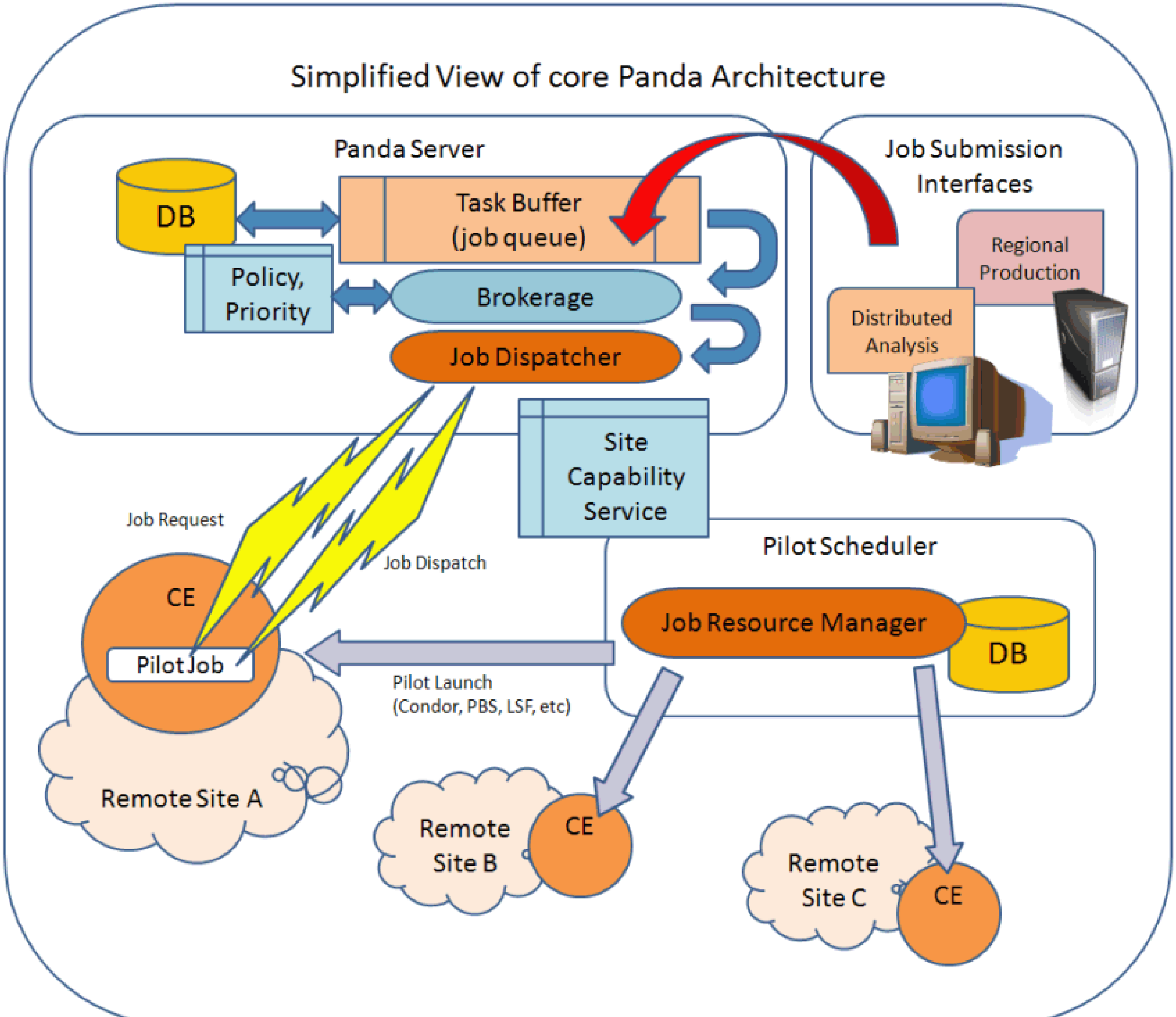
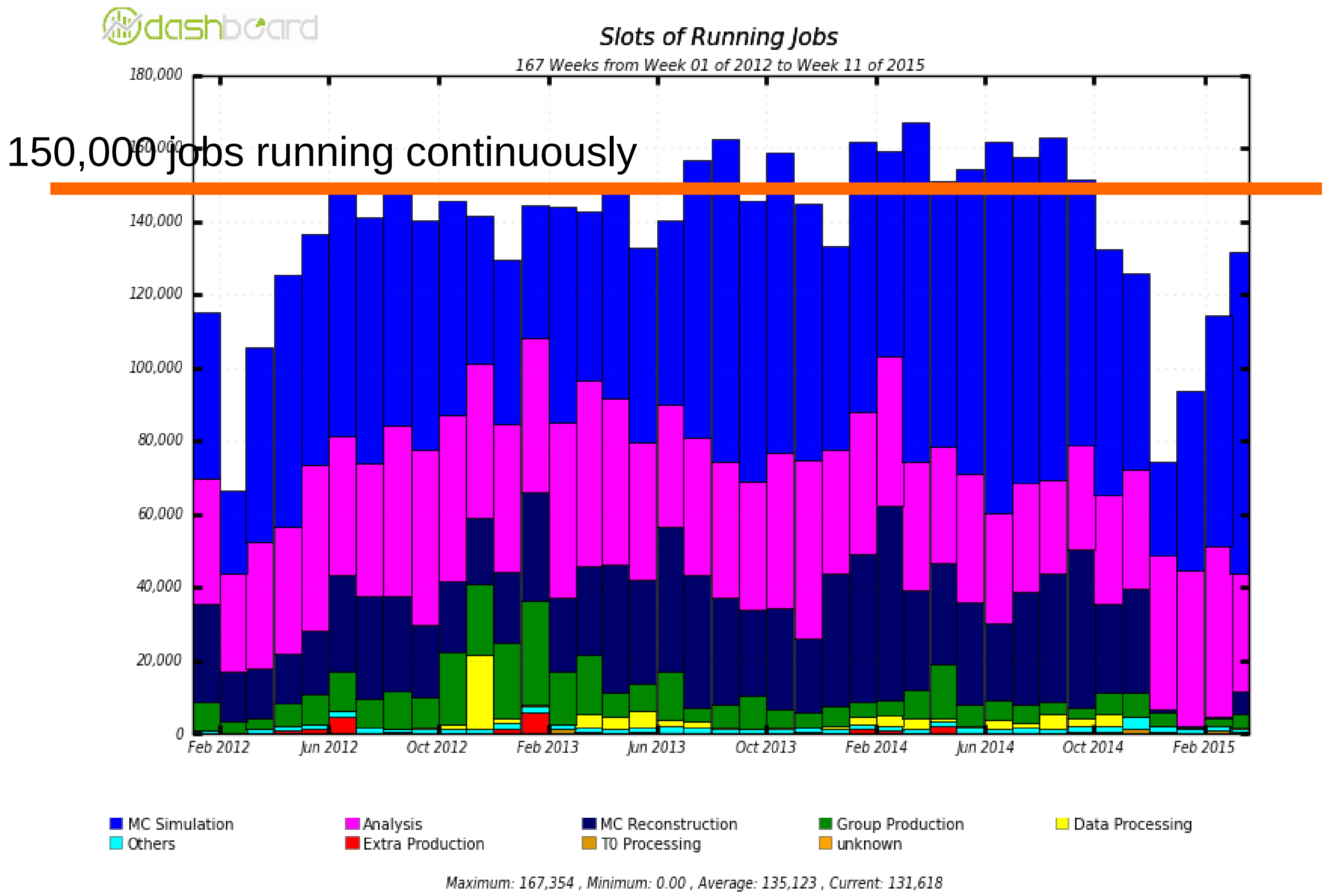
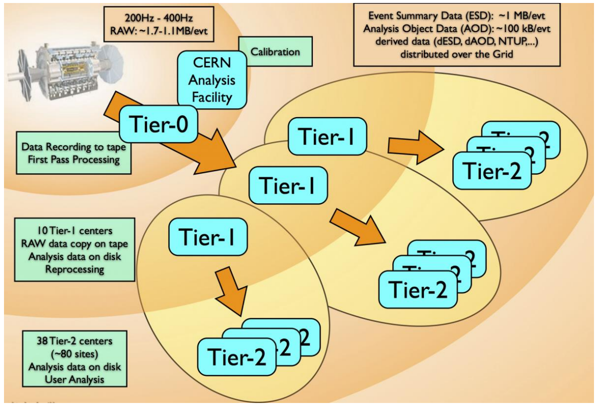

Simplified View of core Panda Architecture
Panda Server
DB
Task Buffer
(job queue)
Policy,
Priority
Brokerage
Job Dispatcher
Job Submission
Interfaces
Regional
Production
Distributed
Analysis
Site
Capability
Service
Job Request
Job Dispatch
CE
Pilot Job
Pilot Scheduler
Job Resource Manager
DB
Pilot Launch
(Condor, PBS, LSF, etc)
Remote Site A
Remote
Site B
CE
Remote
Site C
CE

Image by BNL

Job stats


dashboard
Slots of Running Jobs
167 Weeks from Week 01 of 2012 to Week 11 of 2015
150,000 jobs running continuously
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
Feb 2012
Jun 2012
Oct 2012
Feb 2013
Jun 2013
Oct 2013
Feb 2014
Jun 2014
Oct 2014
Feb 2015
MC Simulation
Analysis
MC Reconstruction
Group Production
Data Processing
Others
Extra Production
T0 Processing
unknown
Maximum: 167,354 , Minimum: 0.00 , Average: 135,123 , Current: 131,618

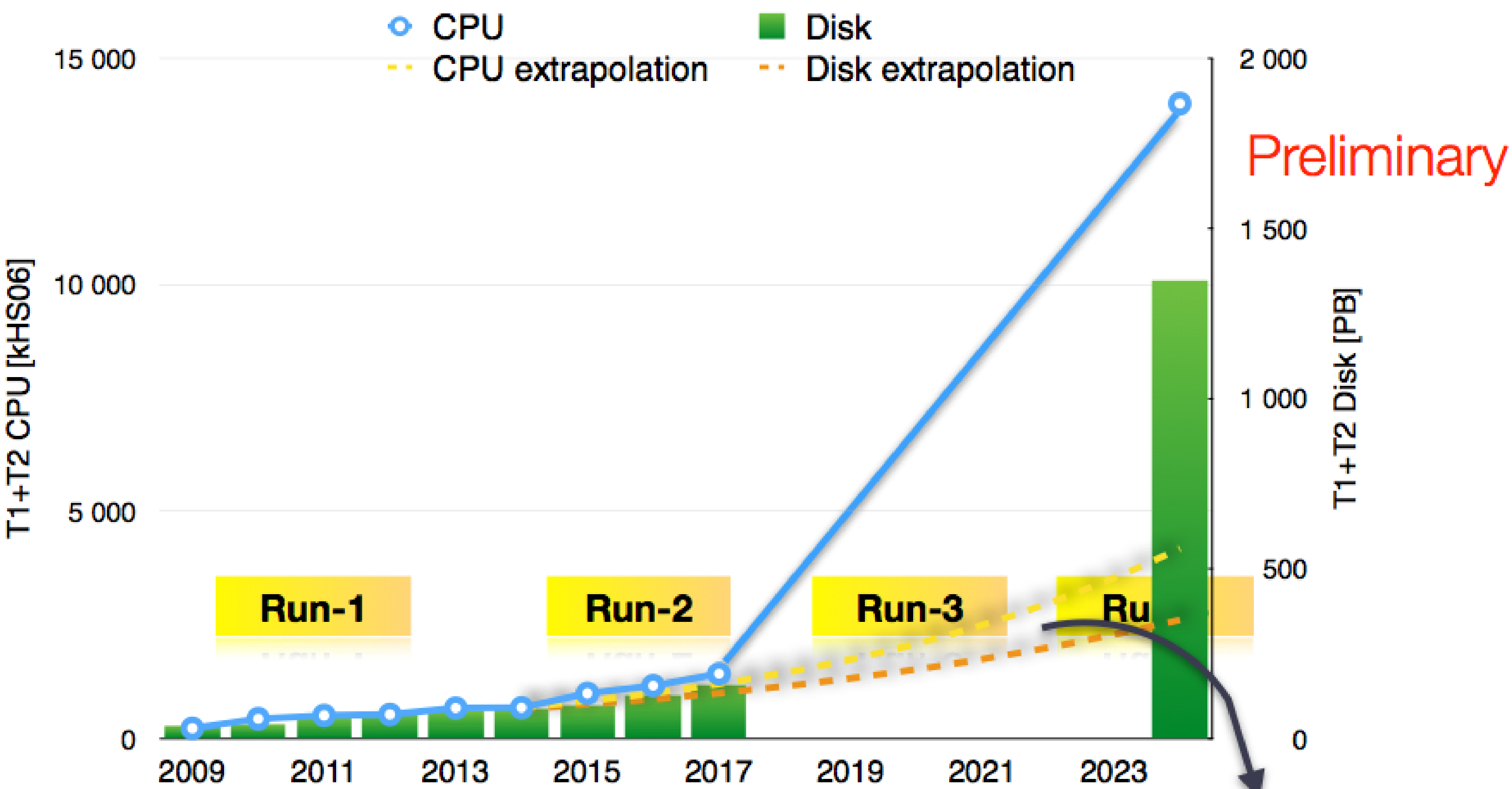
UiO : Department of Physics
University of Oslo



Current Challenges

- New trends in data management
  - Original model was based on network being the weak point
  - But network has proven to be cheaper and better than expected
  - Break the rigid hierarchical model of data flow and sending jobs to data
    - Dynamic data placement
    - Remote data access over wide area network
- Event-level workflow instead of file-level
- Need more CPU and disk but with flat budget -> opportunistic resources
  - High Performance Computing (supercomputers)
  - Volunteer Computing (general public)

ATLAS resource needs at T1s & T2s



Extrapolation from 2014 (Moore law)

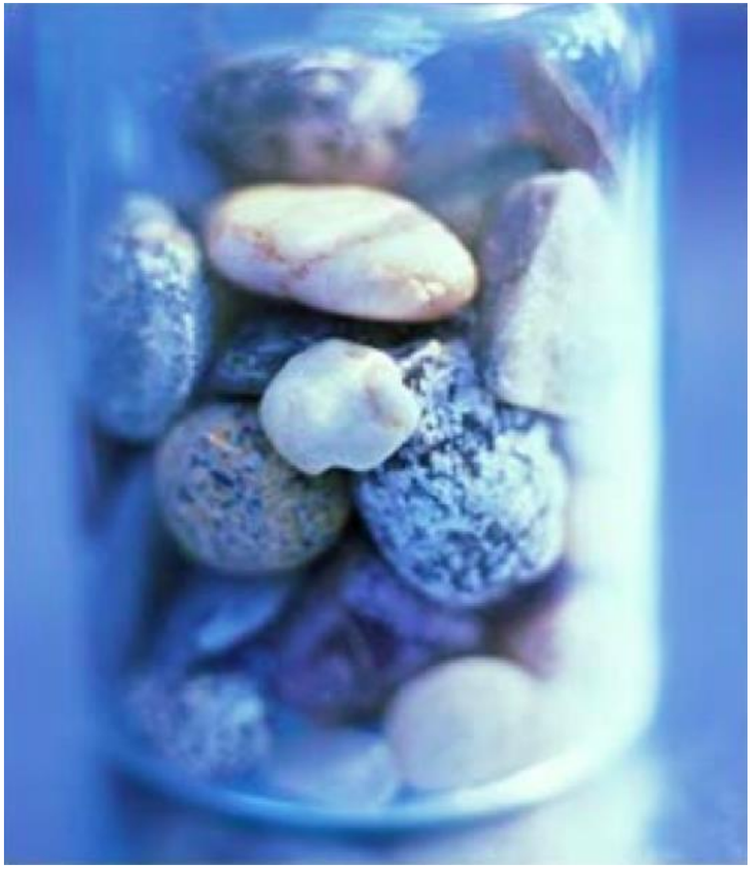
Slide by E. Lancon

High Performance Computing (HPC)

- The Grid is made up of dedicated computing clusters
- Most other scientific computing takes place on HPC
- Differences HPC vs Grid:
  - Massively parallel vs single-node workload
  - Low vs high I/O
  - Restricted vs open enviroment
  - Multiple vs single CPU/OS flavours
  - username/password vs x509 cerfiticate

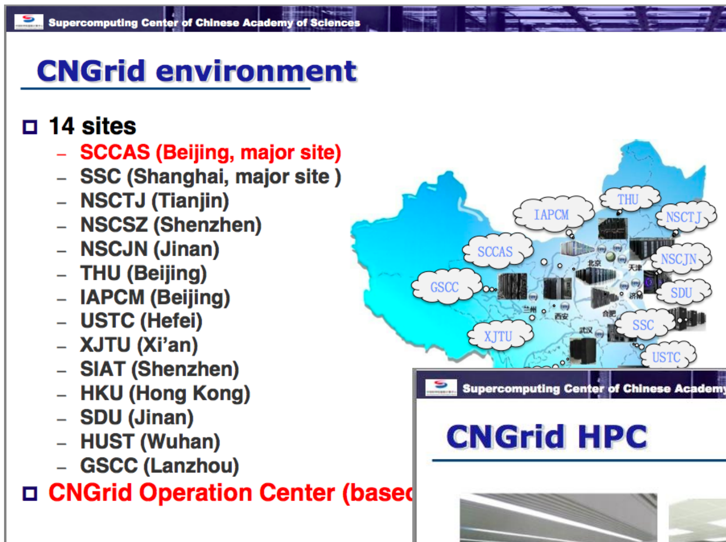
HPC potential - backfilling

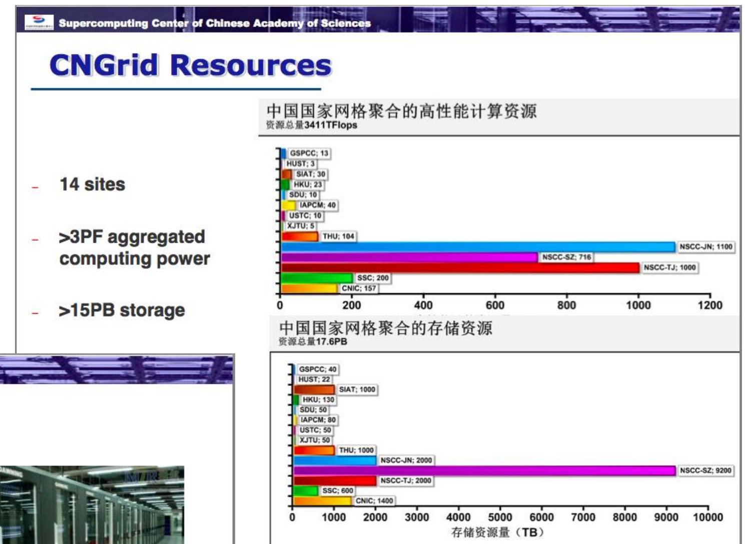
- HPCs are used at 80-90% capacity
- Fill in scheduling holes between big jobs with our small jobs
  - Resources would not be used anyway so we can get them for free
  - The HPC gets higher utilisation and recognition in papers

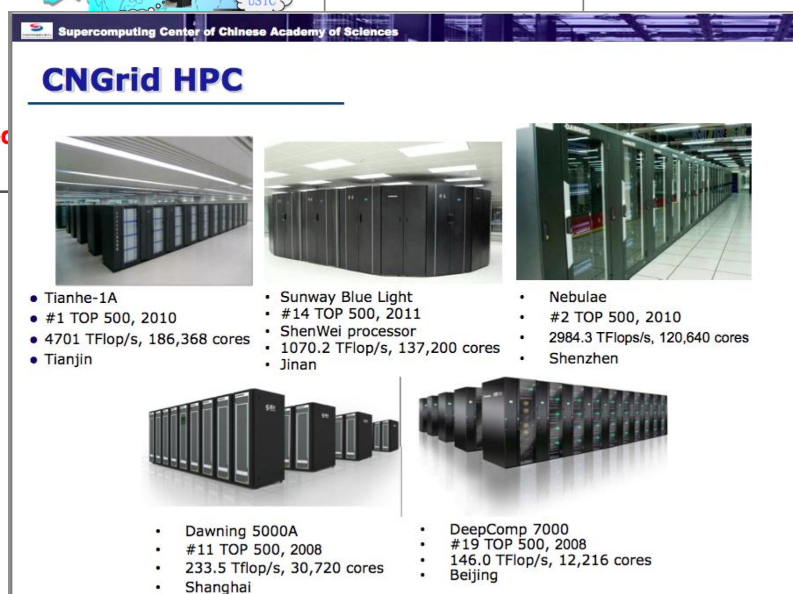


- Targeting HPC centres in Scandinavia, USA, France, Germany, Switzerland, UK, China, …

Future project for ATLAS access to Chinese HPC Grid



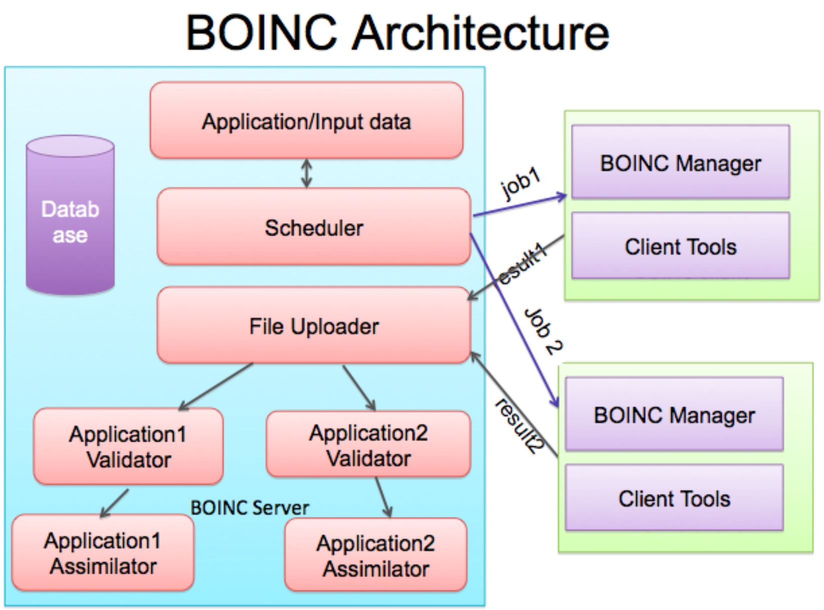


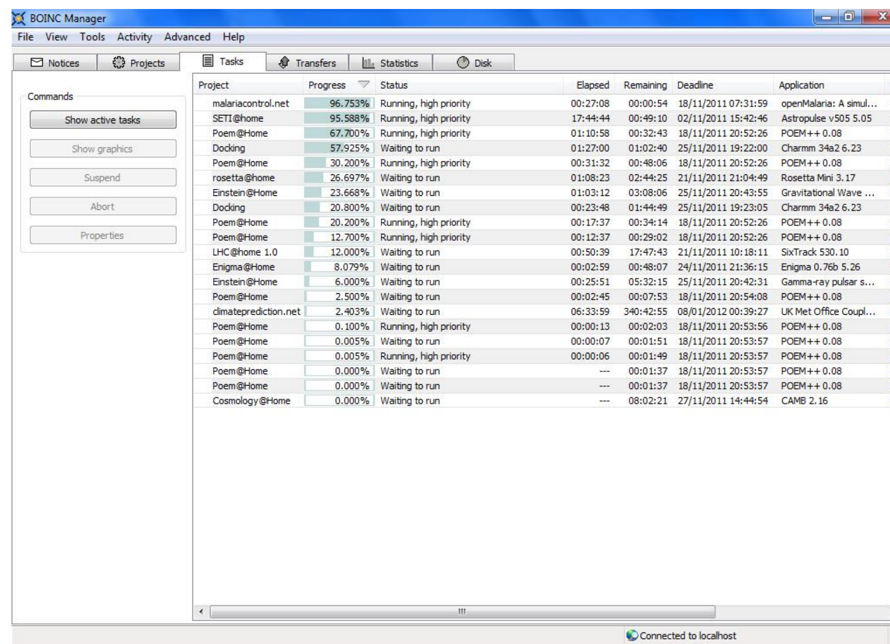


Volunteer Computing

- How YOU can help ATLAS!
- Run simulation of collisions inside the ATLAS detector at home

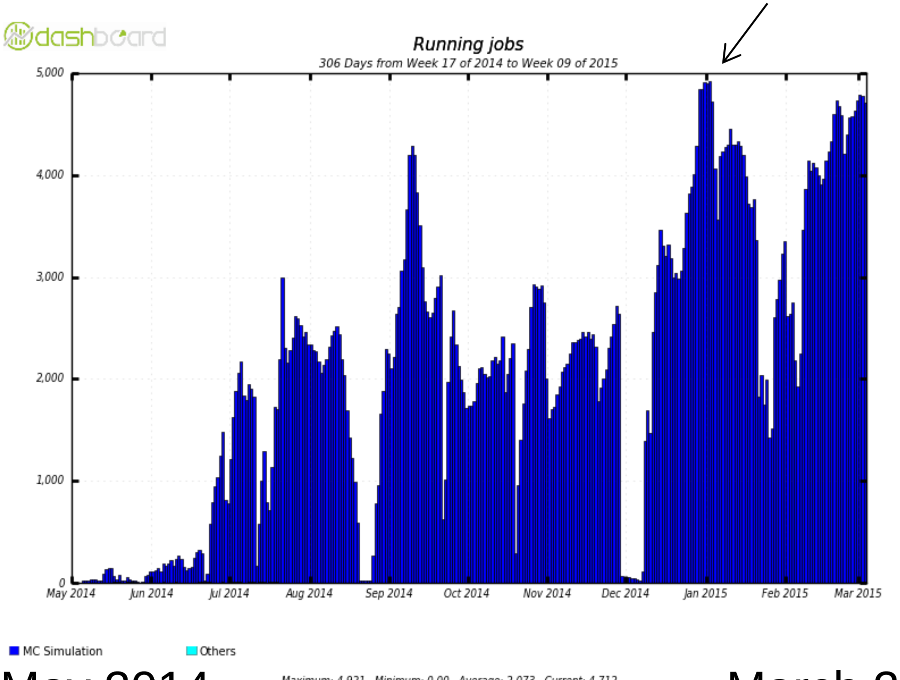
Volunteer Computing via BOINC





ATLAS@Home

5000 running jobs



May 2014 March 2015

- Like getting a large computing centre for free
- Not quite for free, volunteers expect a certain level of support
- Large potential in idle institute desktops
- Join us!

http://atlasathome.cern.ch

Why not just use "the cloud"?

- Historical reasons
  - Grid infrastructure has developed and stabilised over many years
- Funding
  - Research agencies prefer to pay for in-house expertise
- Sustainability
  - LHC will be taking data for the next 20+ years, data must be kept for even longer than that…
- Cost
  - Data-intensive computing 5-10 times more expensive using commercial cloud providers

Summary

- Grid computing is a vital part of LHC physics

- For the average user it is really like the Electric Grid
- UiO plays a strong part at many levels of Grid computing work
- Many interesting challenges ahead

Slide by Rolf Heuer, 4 July 2012