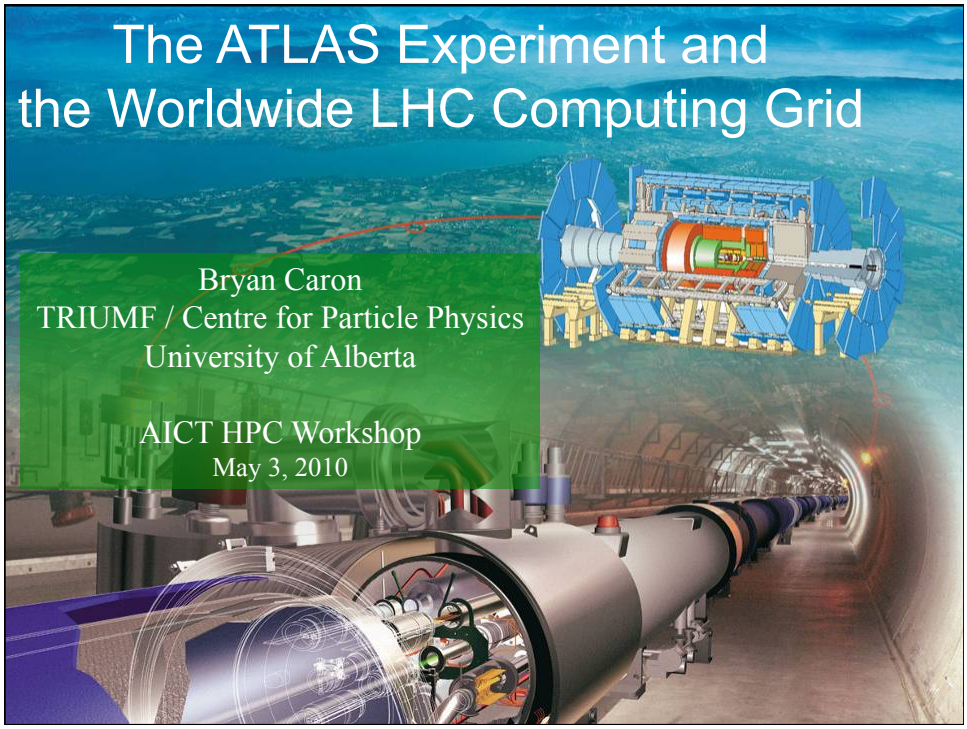
The ATLAS Experiment and the Worldwide LHC Computing Grid



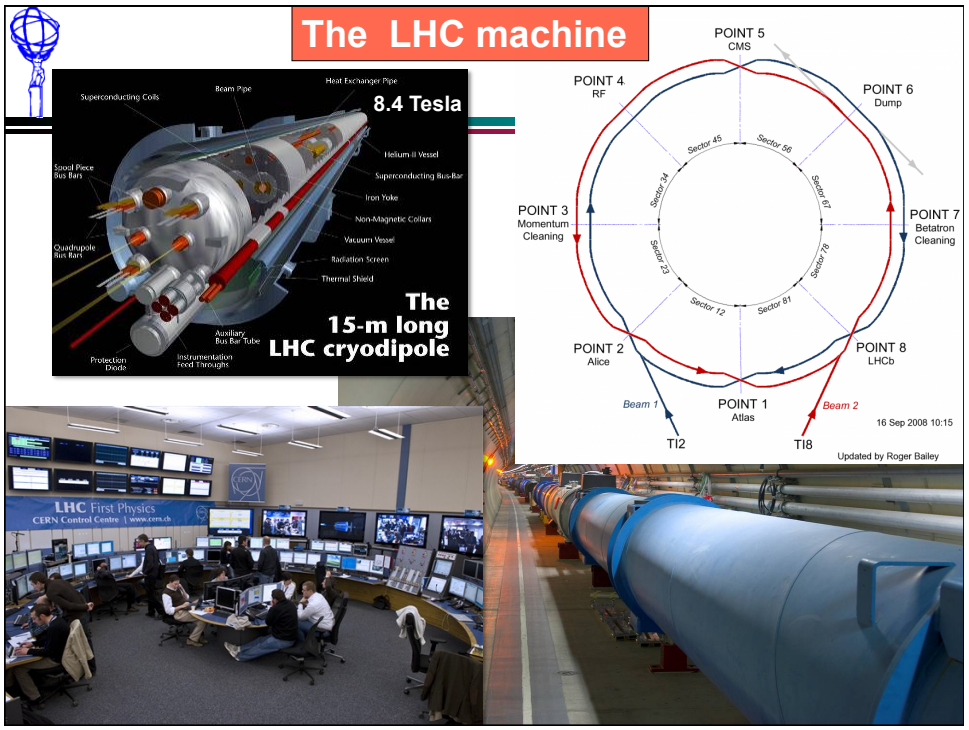
Bryan Caron
TRIUMF / Centre for Particle Physics
University of Alberta

AICT HPC Workshop
May 3, 2010

Outline

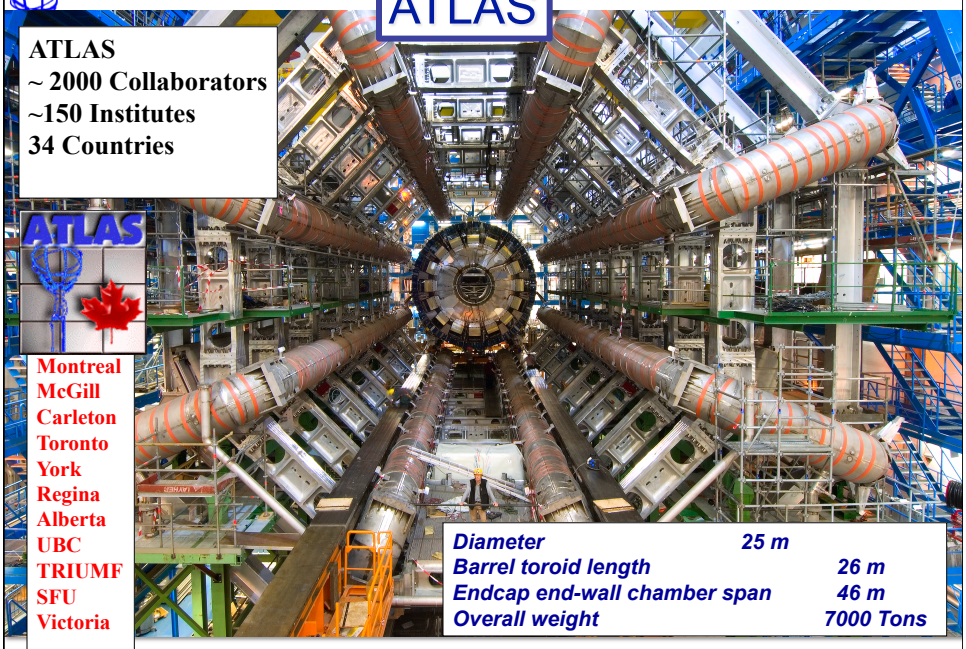
- The Physics
- The Challenge
- The Tools : The World-wide LHC Computing Grid
  - LHC and ATLAS-Canada Computing Model
  - Distributed Data Management
  - Large Scale Distributed Analysis
  - Operations and Monitoring
- Summary and Conclusions
  - LHC and ATLAS Status
  - Outlook for 2010/2011

The Large Hadron Collider
LHC accelerator ring 27 km circumference
France
ALICE
CMS
CERN
ATLAS
LHC-B
7 TeV protons
Switzerland


The LHC machine
8.4 Tesla
The 15-m long LHC cryodipole
POINT 5 CMS
POINT 6 Dump
POINT 7 Betatron Cleaning
POINT 8 LHCb
POINT 1 Atlas
POINT 2 Alice
POINT 3 Momentum Cleaning
POINT 4 RF
Beam 1
Beam 2
TI2
TI8
16 Sep 2008 10:15
Updated by Roger Bailey
LHC First Physics
CERN Control Centre

ATLAS

ATLAS
~ 2000 Collaborators
~150 Institutes
34 Countries

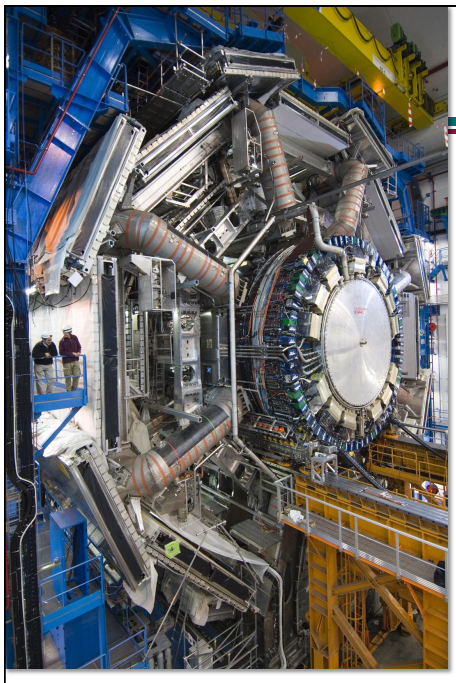


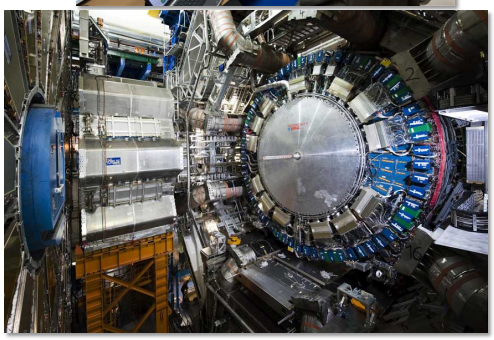
Montreal
McGill
Carleton
Toronto
York
Regina
Alberta
UBC
TRIUMF
SFU
Victoria

Diameter	*25 m*
Barrel toroid length	*26 m*
Endcap end-wall chamber span	*46 m*
Overall weight	*7000 Tons*

Bryan Caron – ATLAS Computing Grid 5

ATLAS





Bryan Caron – ATLAS Computing Grid 6

Objectives of ATLAS and LHC

- One of the principle objectives: Discover the **Higgs** particle
  - Existence is fundamental to the Standard Model and many particle theories
- Some of the most fundamental questions in physics still to be answered:
  - What gives particles masses?
  - Why are there 3 types of quarks and leptons?
  - Are there more types of particles and forces to be discovered at higher energies?
  - How can gravity be incorporated into the Standard Model?

LHC Data Challenge

• A particle collision = an event
• 40×10^6 collisions/second
• Our goal is to count, trace and characterize all the particles produced and fully reconstruct the process.

• After filtering, 100 collisions of interest per second
• > 1 Megabyte of data digitised per collision
recording rate > 1 Gigabyte/sec
• 10^{10} collisions recorded each year
stored data > 10 Petabytes/year of data

1 Megabyte (1MB)
A digital photo

1 Gigabyte (1GB)
= 1000MB
5GB = A DVD movie

1 Terabyte (1TB)
= 1000GB
World annual book production

1 Petabyte (1PB)
= 1000TB
Annual production of one LHC experiment

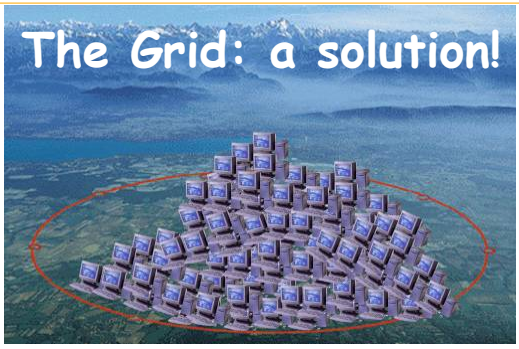
1 Exabyte (1EB)
= 1000 PB
3EB = World annual information production

LHC Processing Challenge

LHC data analysis requires a computing power equivalent to ~ 100,000 of today's fastest PC processors!

Where will the experiments find such a computing power?

Nowhere near enough at CERN!



What is the Grid?

- The name Grid is chosen by analogy with the electric power grid:
  - plug-in to computing power without worrying where it comes from.
- The Grid is an emerging infrastructure that provides:
  - seamless access to computing power
  - data storage capacity distributed over the globe.

- The Grid is, for the time being, many Grids
- Grids are a framework for Resource Virtualisation
- Grids will rely on new models of Inter-Organisational Security
- Grids will be used by Virtual Organisations
- Grids are a platform for Resource Discovery
- Grid Services will be based on Web Services

Grid development has been driven by the academic community

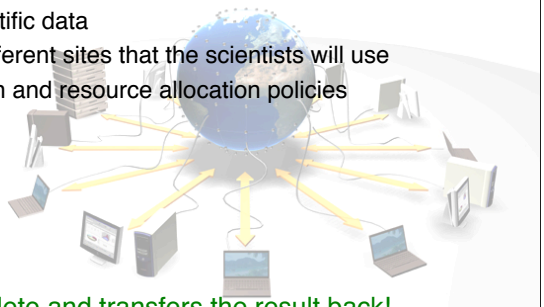
Industrial variants are on-demand computing, Utility computing

How does (should) it work?

• The Grid relies on advanced software, called middleware, which ensures seamless communication between different computers and different parts of the world

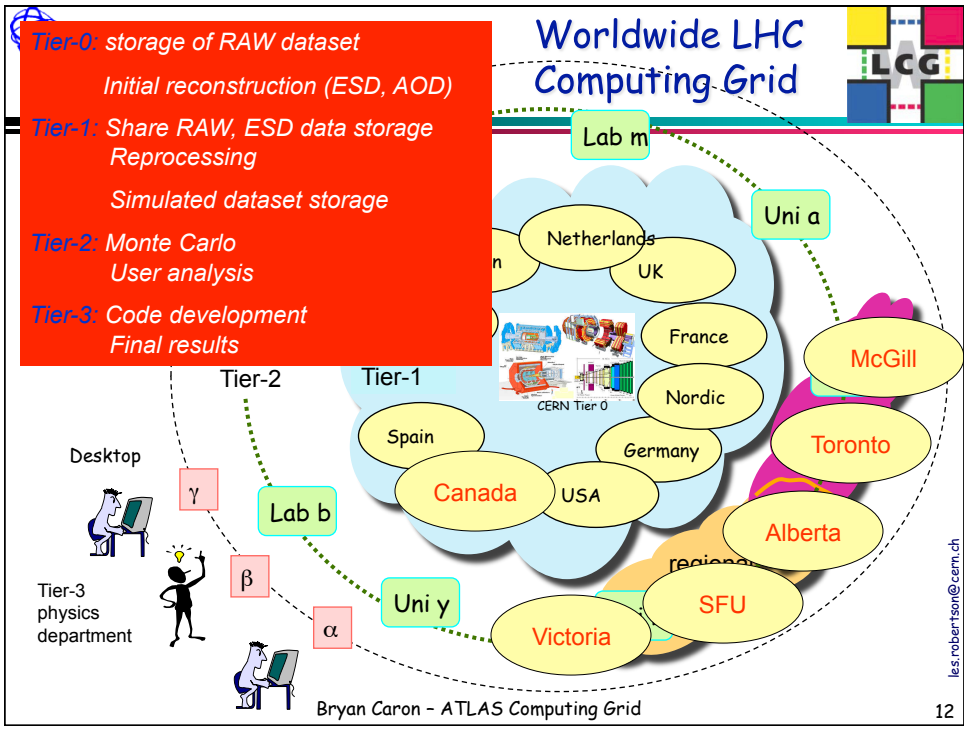
The GRID middleware:

- Finds convenient places for the scientist's computing task to be run
- Optimises use of the widely dispersed resources
- Organises efficient access to scientific data
- Deals with authentication to the different sites that the scientists will use
- Interfaces to local site authorisation and resource allocation policies
- Runs the jobs
- Monitors progress
- Recovers from problems



… and ….

Tells you when the work is complete and transfers the result back!



The Canadian Computing Model

- Tier-1 at TRIUMF:
  - Raw data storage; event reconstruction; expert personnel
  - Dedicated facility funded through CFI-EOF (2007-2011)
- Tier-2 Centres in the Universities:
  - East (Toronto, McGill) and West (UVic, SFU, Alberta) Tier-2 Federations
  - Monte Carlo simulation & physics analysis
  - Organized and independent user access patterns will have peaks and valleys
  - Will use shared facilities in the HPC Consortia (CFI-NPF)
- Resources for common ATLAS contributions and additional Canadian-only use
  - Grid-accessible cpu and storage for Canadian ATLAS users

3 – 4 PB of raw data + 2.5 PB of secondary data generated each year
(200 Hz, 14 hours/day, 200 days/year for nominal operations)
Canadian share/responsibility is 5%

Keys to Grid Usability

- Building a production level infrastructure
- Building robust services
- Building a good operations strategy
- Providing support
  - Grid Computing Infrastructure
  - Distributed Data Management
  - Large Scale Distributed Analysis
  - Operations and Monitoring

Grid Computing Infrastructure

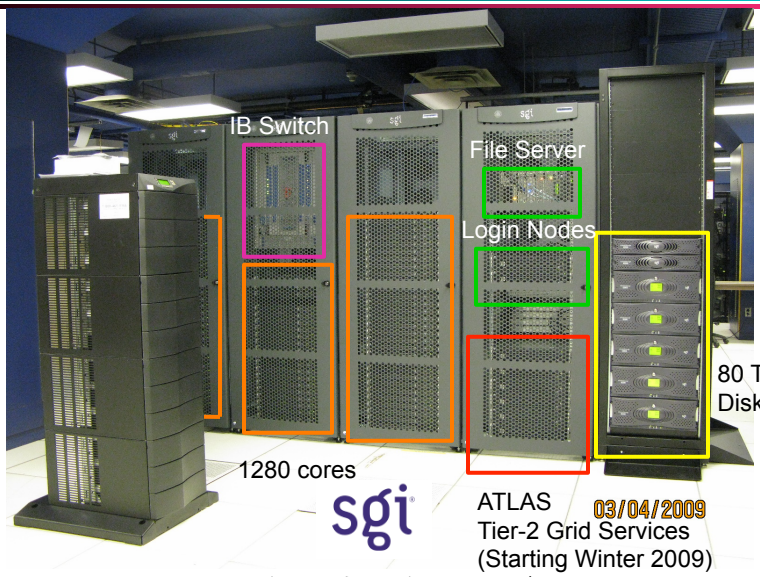
ATLAS Resource Requirements

- TRIUMF Tier-1 Centre
  - 1210 cpu slots,
    2.1 PB disk (usable),
    1.6 PB tape
- Tier-2 Centres (Sum)
  - ATLAS Canada
    WLCG commitment is
    approx. 5% of totals

Tier-2 CPU (kHS06)	*Old 2010*	*New 2010*	*2011*	*2012*	*Main components*
Simulation production	114	65	65	65	Tier-2 fraction of simulation
Group activities	32	38	49	52	Large scale skimming and slimming jobs
User activities	94	123	164	178	User analysis
Total	***240***	***226***	***278***	***295***	To summary table 3

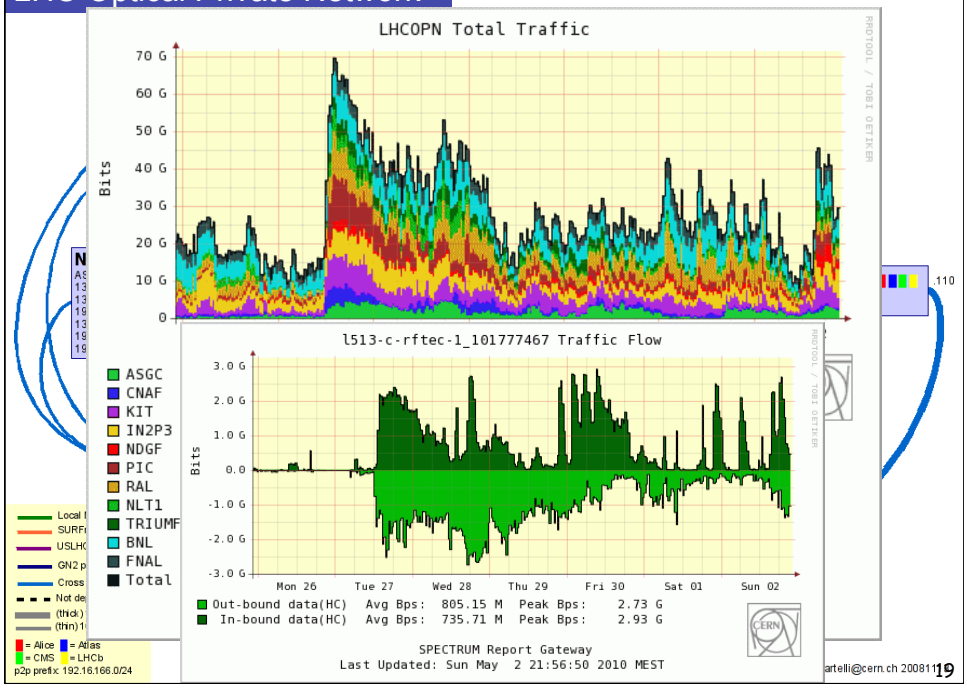
Tier-2 Disk (PB)	*Old 2010*	*New 2010*	*2011*	*2012*	*Main components*
Current RAW data	0.3	0.4	0.3	0.1	On request for debugging
Real data	7.0	11.3	19.7	19.7	AOD+DESD for analysis
Simulated data	9.2	7.7	11.1	15.2	AOD+DESD for analysis
Calibration and alignment inputs	0.2	0.2	0.3	0.3	Calibration data
Group data	1.9	2.8	4.6	5.5	dESD & dAOD from group analysis
User data	1.8	1.2	1.8	2.4	Scratch space
Buffers	0.6	0.6	0.6	0.6	For simulation
Total	***21***	***24***	***38***	***44***	*To summary Table 3*

WestGrid–Alberta Cluster "Checkers"

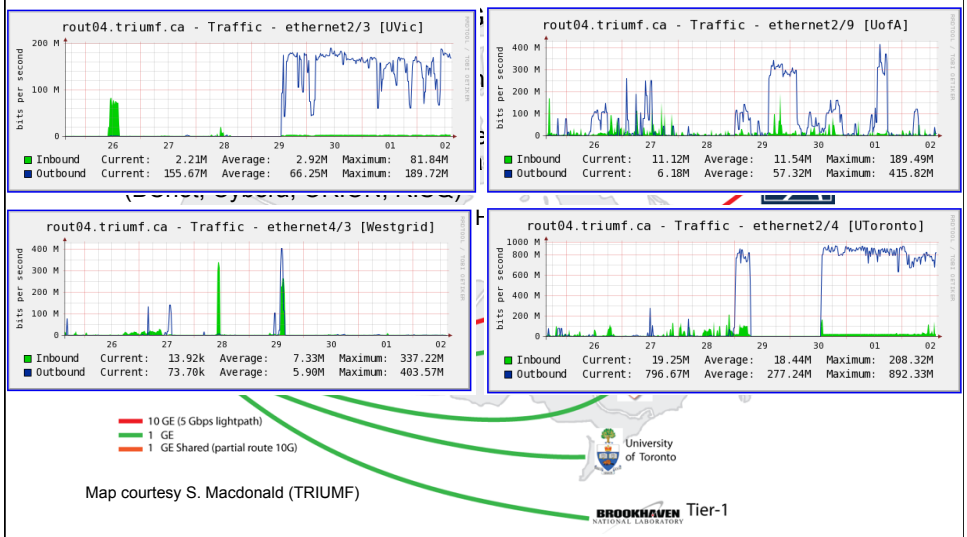


LHC Data Network

LHC Optical Private Network



ATLAS-Canada Network



Map courtesy S. Macdonald (TRIUMF)

Bryan Caron – ATLAS Computing Grid

20

Grid Infrastructure

- Common grid services installed at CA Tier-1 and Tier-2s
  - LCG Compute Element via GLite middleware
    - Computing resources
  - Storage Element based upon dCache
    - Disk based (tape additionally required at the Tier-1)
  - LCG site BDII and MONitoring services
    - Information System
  - LCG WorkerNode client interface applications
    - Data file staging to/from WorkerNode, …
  - Central distribution of ATLAS software releases and conditions databases
  - Additional highly specialized services required at TRIUMF Tier-1
    - Oracle (ATLAS Conditions and Tag data),
    - File Transfer Service (data movement between sites),
    - ATLAS VO-Box (ATLAS Data Management),
    - LHC File Catalogue (file location and meta-data),
    - TOP-BDII (information system), …

Distributed Data Management

Distributed Data Management

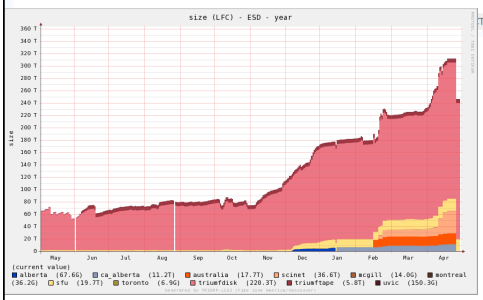
- Distributed Data Management System
  - Baseline Service for Data Distribution and Consolidation, MonteCarlo production and Reprocessing, Data Analysis
  - Provides functionalities for Data Organization (bookkeeping) and Data Transfers
    - Data Organization based on Datasets
      - Collection of Files
      - Unit of data location and data replication, data deletion
    - Data Transfers based on Subscriptions
      - Placement policy, which DDM tries to enforce
- DDM is a Distributed Service
  - Central Catalogs, Accounting Service
    - Hold dataset definitions, subscription requests, Dataset Locations
  - Site Services, Deletion Service
    - Triggering actions to enforce subscriptions
  - Other WLCG services at sites
    - FileTransferService for data movement (at T1s), StorageElements (at T1/T2s), File Catalogs (at T1s) hold information about single files placement
  - Command Line Interfaces

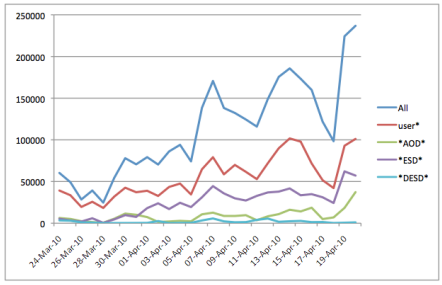
Data Management Dashboard

Activity Summary ('2010-05-02 20:10' to '2010-05-03 00:10' UTC)

Click on the cloud name to view list of sites

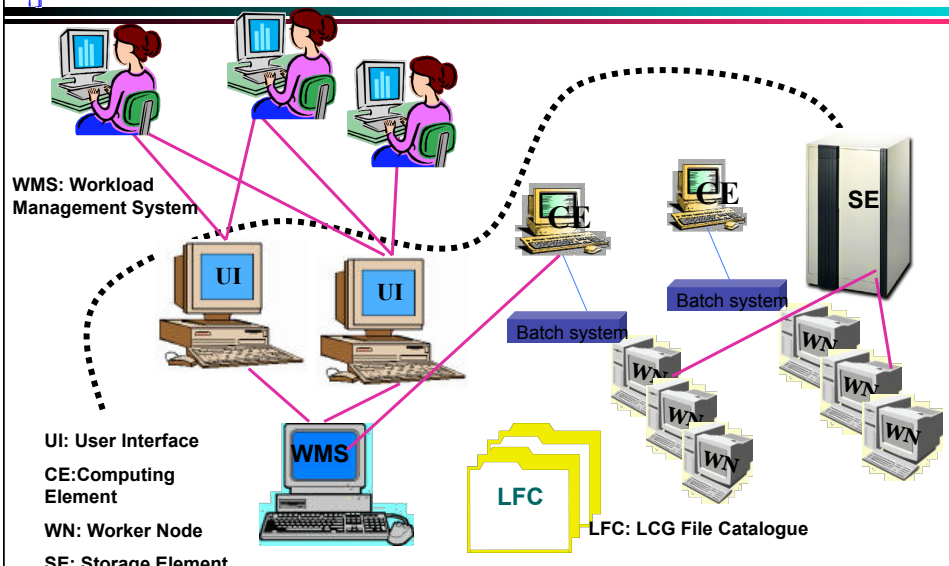
	Transfers			Registrations		Errors			Services
Cloud	Efficiency	Throughput	Successes	Datasets	Files	Transfer	Registration	Services	Grid
ASGC	97%	213 MB/s	10240	131	10240	282	0	0	
BNL	98%	1039 MB/s	22059	536	22329	511	0	0	
CERN	90%	92 MB/s	16876	117	17017	1792	0	0	
CNAF	58%	45 MB/s	2750	256	2766	1994	0	0	
FZK	99%	521 MB/s	16014	768	16020	224	0	0	
LYON	96%	384 MB/s	19698	513	19768	777	0	0	
NDGF	92%	48 MB/s	3273	163	3271	269	0	0	
PIC	91%	20 MB/s	1301	181	1309	130	0	0	
RAL	90%	756 MB/s	20309	433	20303	2133	0	0	
SARA	88%	266 MB/s	10201	219	10215	1389	0	0	
TRIUMF	99%	387 MB/s	16745	260	16763	152	0	0	



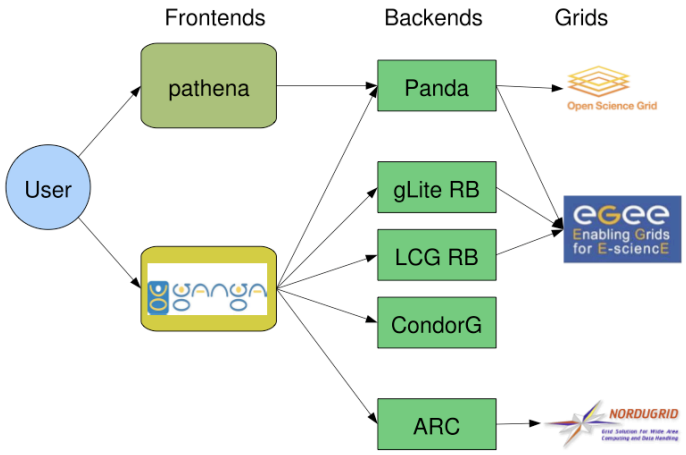


Large Scale Distributed Analysis

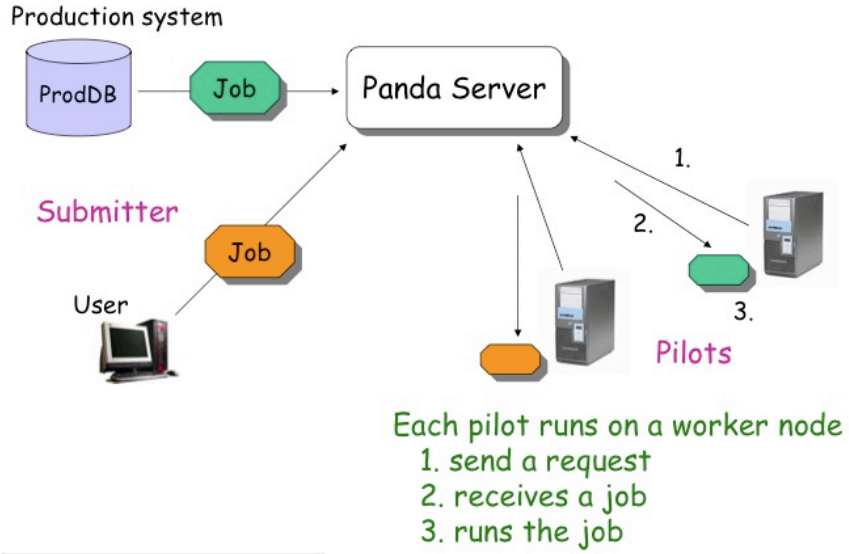
How does a Physicist use the LHC Grid?



Distributed Analysis



Panda Job Flow

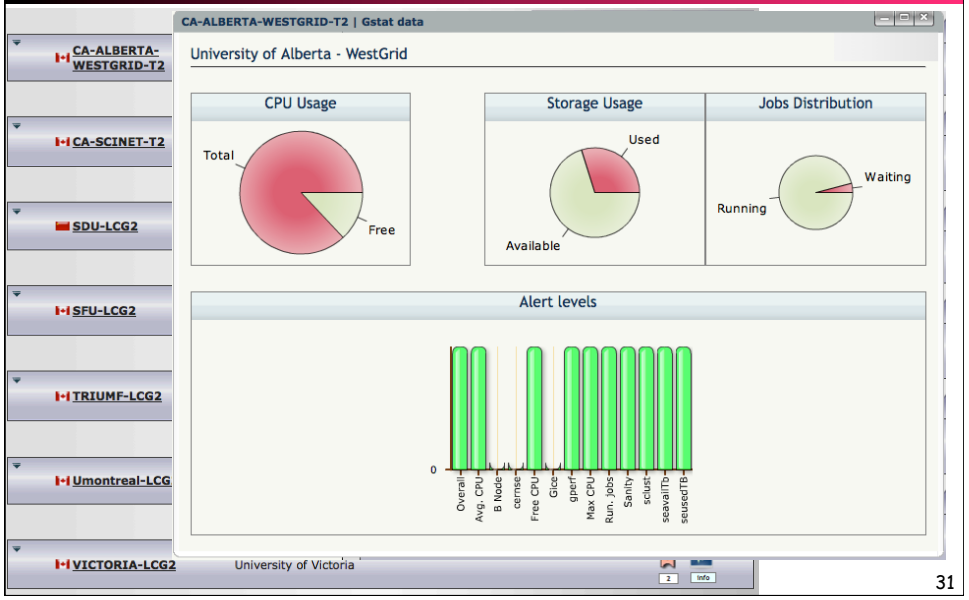


Operations and Monitoring

Distributed Computing Monitoring

- A wide range of tools aimed at **users, shifters, operators and managers**
  - Distributed Data Management dashboard
    - Dataset and file transfer / placement
  - Production System dashboard
    - Production tasks and associated jobs execution
  - Analysis dashboard
    - User jobs on the grid
  - Panda dashboard
    - Activity of Panda on the different clouds
  - Service monitoring dashboard (SLS)
    - Detailed information on the status of site services, central catalogs, …
  - Central Data Replication monitoring
    - A complement to the DDM dashboard
  - Software installations monitoring

Services Monitoring and Metrics



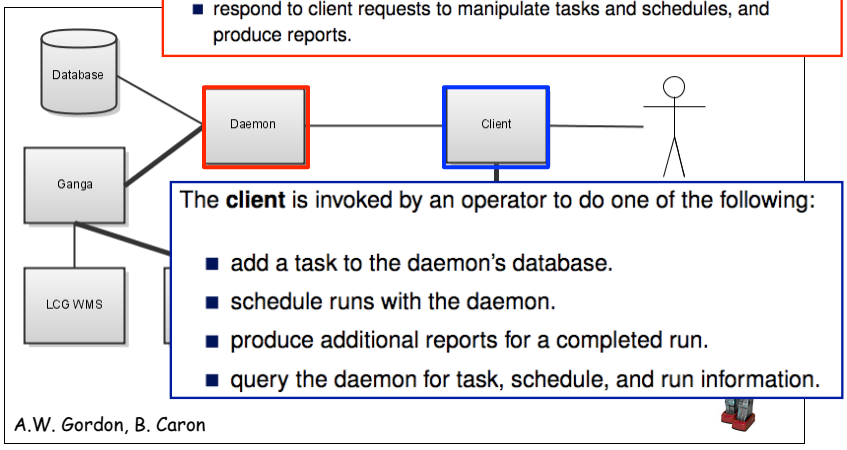
31

Grid

- HammerClou
  - Automa

The **daemon** runs persistently in the background on some server. It is responsible for:

- keeping track of tasks, schedules, runs, and extracted data in a database.
- using Ganga to submit jobs, monitor their status while they run, and resume monitoring if the daemon gets restarted.
- respond to client requests to manipulate tasks and schedules, and produce reports.



The **client** is invoked by an operator to do one of the following:

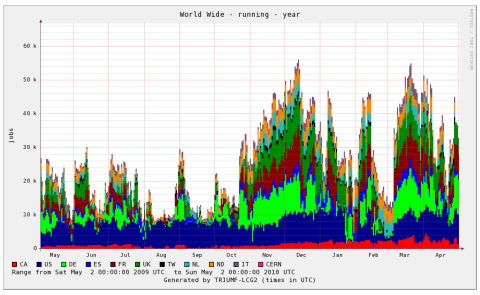
- add a task to the daemon's database.
- schedule runs with the daemon.
- produce additional reports for a completed run.
- query the daemon for task, schedule, and run information.

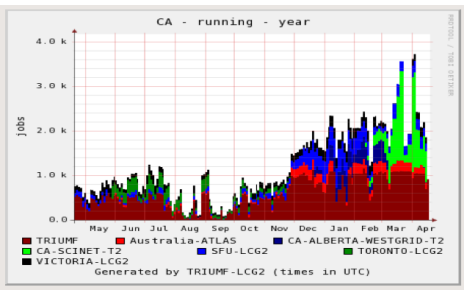
Bryan Caron – ATLAS Computing Grid

32

Analysis and Production Monitoring

CA Sites	Job Nodes	Jobs	Latest	Pilot Nodes	defined	assigned	waiting	activated	sent	running	holding	transferring	finished	failed	tot trf	other
Site Name	704	39	05-02 21:15	999	0	127	0	6883	2	2765	117	2515 / 28	2969	39	1%	0% 1%
Australia-ATLAS	30	0	05-02 21:15	51	0	0	0	80	0	208	3	204 / 0	211	0	0%	0% 0%
CA-ALBERTA-WESTGRID-T2	86	1	05-02 21:15	109	0	0	0	379	0	320	4	253 / 0	553	1	0%	0% 0%
CA-SCINET-T2	328	34	05-02 21:15	445	0	0	0	3302	2	1190	74	2029 / 0	353	34	9%	1% 8%
SFU-LCG2	21	0	05-02 20:52	24	0	45	0	0	0	0	0	28 / 28	0	0		
TORONTO-LCG2	0	0	offline	0	0	0	0	0	0	0	0	0 / 0	0	0		
TRIUMF	236	4	05-02 21:15	344	0	0	0	3122	0	1046	36	0 / 0	1852	4	0%	0% 0%
VICTORIA-LCG2	3	0	05-02 21:13	26	0	82	0	0	0	1	0	1 / 0	0	0		

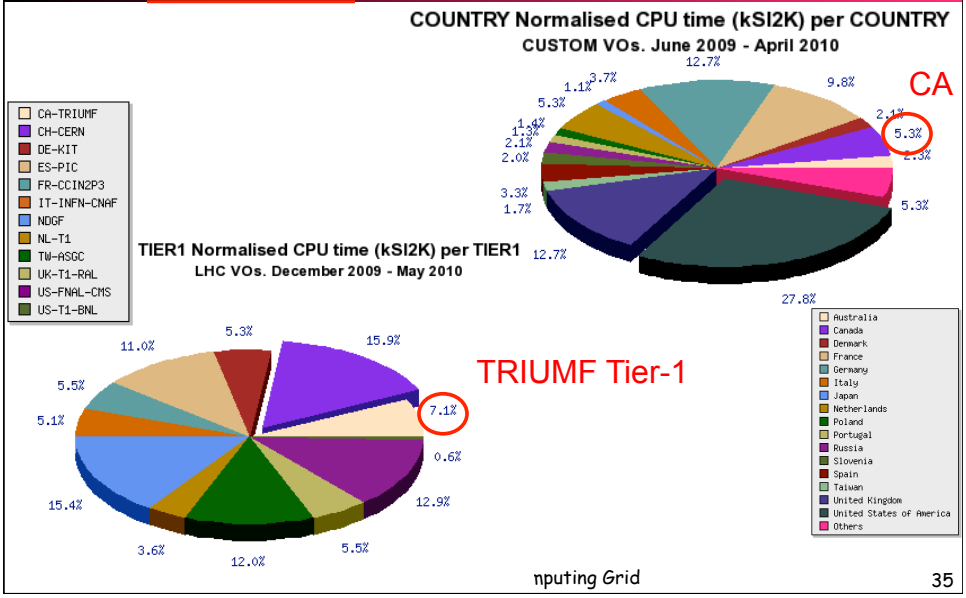




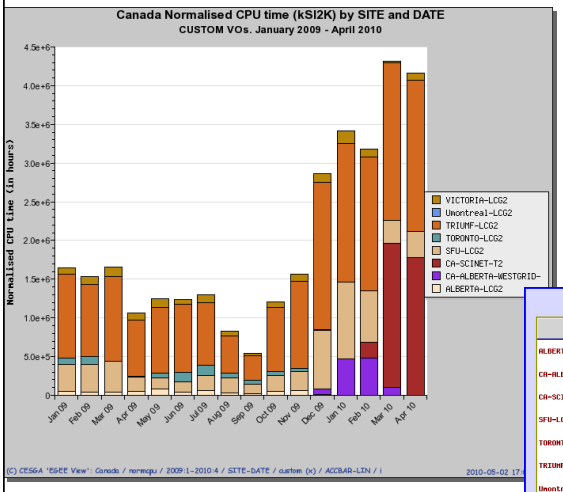
ATLAS-Canada Operations

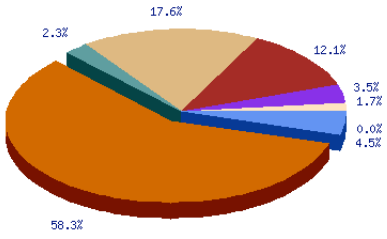
- ATLAS-Canada computing management and technical groups
  - discuss & propose policy to ATLAS investigators (e.g. sharing of resources, interaction with the HPC consortia)
  - share information on hardware and Grid middleware operations
- ATLAS Canada Cloud (Tier-1 + Tier-2s) Operations Twiki
  - Weekly grid services status reports
  - National and institutional production status
  - Local site problem reporting, tracking and solution documentation
- CA ROC (Regional Operations Centre) – est. 2009
- Site Availability and Monitoring Tests and others for diagnostics
- ATLAS Production, Analysis and Data Management Dashboards
- HEPNet Canada teleconferences
- WLCG Workshops
- Central ATLAS Tier-0/1/2/3 Jamborees (coordination)

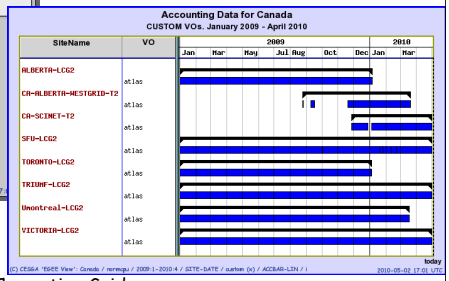
Canadian Contribution to ATLAS Computing



Resource Usage Accounting



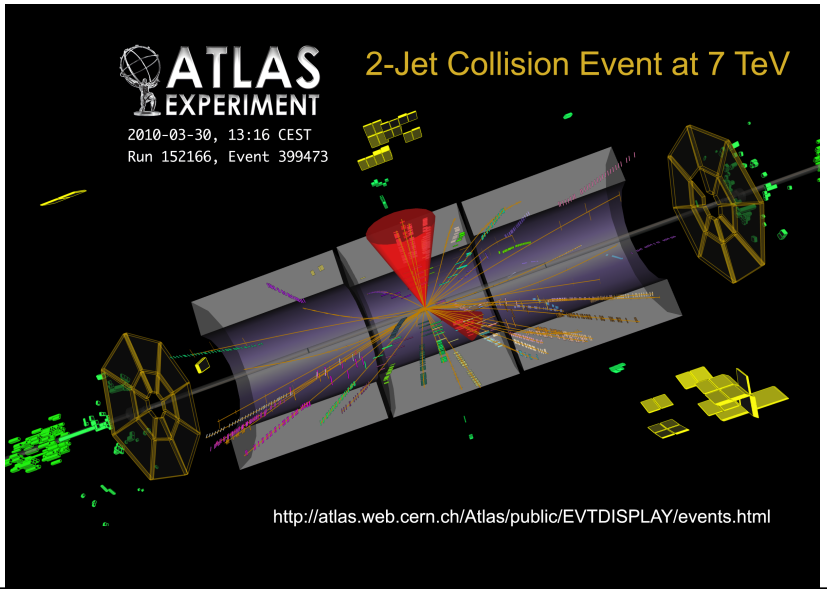




LHC and ATLAS Status

- Well known history of Friday September 19, 2008:
  - Faulty electrical connection between 2 magnets melted at high current during training for 5 TeV in LHC sector 34
    ➔ Large He leak into tunnel
  - Repairs + enhanced detection / protection system installed
- ATLAS resumed global cosmics runs around May 09
- LHC beam physics data-taking period resumes November 2009
  - extended running period of approx. 18–24 months
- Nov 23: colliding beams at 450 GeV per beam
- Nov 30: new record collision energy of 2.36 TeV
  - LHC in technical stop Dec 16 – Feb 2010
- March 19: circulating beams at 3.5 TeV per beam
- March 30: colliding beams at 7 TeV centre-of-mass

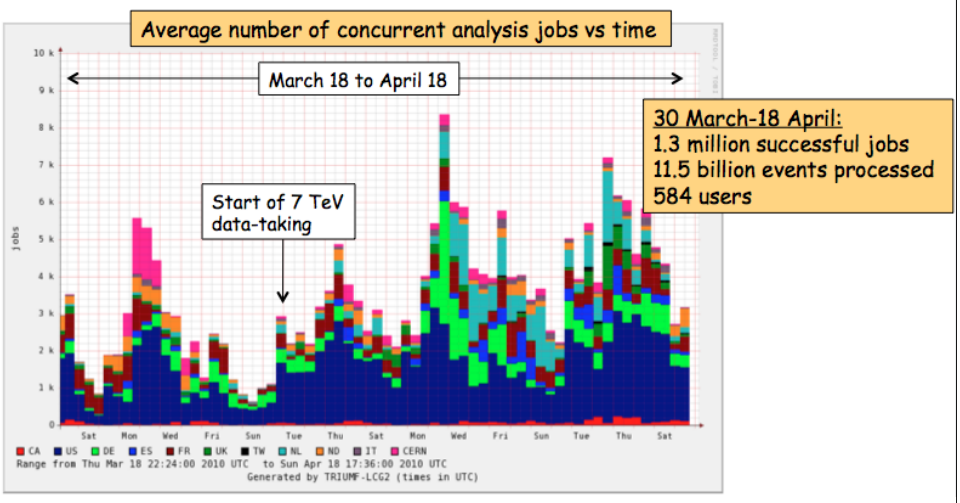
LHC and ATLAS Status



LHC and ATLAS Status

- April 24/25: factor 10 increase in luminosity for 30 hour 'fill' of the accelerator
  - More than double number of collision events recorded during single run compared to all previously collected 7 TeV data
- To Date: total # collision events @ 7 TeV = 66M
- Continuing improvement in LHC beam conditions are targeting 10^4 luminosity increase over 2010 and 2011

User Analysis Activity



Conclusions and Outlook

- Numerous challenges to extract physics from PetaBytes of collected data
- Turn to the World-wide LHC Computing Grid
  - Tiered structure for data distribution and processing
  - Distributed data analysis
  - Distributed data management
  - Operations and monitoring
- ATLAS-Canada Tier-1 at TRIUMF and Tier-2 Centres
  - ATLAS-common and Canadian-specific resource allocations
  - Tier-2 Centres: Analysis and Simulation Facilities
    - Hardware resources for 2010 (and beyond) via HPC Consortia
      - establish smooth stable WLCG operations at the new facilities
      - Strong base of experience in providing the needed ATLAS and LCG services in Canada
- The beam collisions data-taking era has started, and so the future is now!

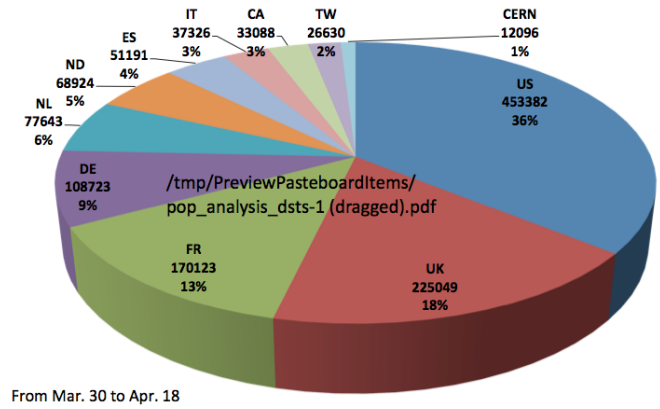
Acknowledgements

- TRIUMF and ATLAS-Canada
  - Tier-2 sites: Alberta, SFU, Victoria, Toronto, McGill
- Our partners in ATLAS Computing
  - WestGrid, SciNet, CLUMEQ
- University of Alberta – AICT, CPP, Physics
- Supporting Organizations
  - NSERC, CFI, Compute Canada, HEPnet Canada, Cybera, CANARIE, …
- The numerous colleagues who contributed material that was included in this presentation

Backup Material

User Analysis Activity

User Analysis Successful Jobs PanDA Backend



From Mar. 30 to Apr. 18
Total: 1.3 million successful jobs, 11.5 billion events processed, 584 users

From Kaushik De