Introduction to Distributed 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
• **Distributed Computing**
Event Collection in ATLAS

Proton-Proton 2835 bunch/beam
Protons/bunch $10^{11}$
Beam energy 7 TeV ($7 \times 10^{12}$ eV)
Luminosity $10^{34}$ cm$^{-2}$ s$^{-1}$

Crossing rate 40 MHz
Collisions rate $\approx 10^7 - 10^8$ Hz

New physics rate $\approx 0.0001$ Hz

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

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 -> DAOD

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

Skimming: removal of whole events based on pre-set criteria

Thinning: removal of whole objects within events based on pre-set criteria

Slimming: removal of variables within objects uniformly across events
**Big Data?**

Large Hadron Collider Data/Year

- 25,000,000 GB
- 15,000,000 GB
- 2,500,000 GB

Tweets/Year

- All numbers approximate.
- * Binary Data

Human Memory

World of Warcraft Servers

US Library of Congress

Wikipedia*

Business emails per year: 3000 PB

Google search index (2013): 98 PB

LHC data output (2012): 15 PB

Total ATLAS data (2017): 250PB

YouTube uploads per year: 15 PB

Content uploaded to Facebook per year: 182 PB

Content uploaded to Twitter per year: 3000 PB

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)
  – 250,000 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 (digital certificates)
Computing Element in more detail

- X509 certificate → local account

- Job submission interface
  - Jobs
  - Info
  - Files

- Local batch system interface

- Info query interface

- Information provider

- File access interface

- Input/output file staging

Client tools
Storage Element in more detail

SRM Server

File Server 1
FS1  FS2  FS3

File Server 2
FS1  FS2  FS3

File Server 3
FS1  FS2  FS3

Graphics by G. Stewart
Copying data around

File Transfer Service

Replica Catalog

SRM Server

File Server 1
FS1  FS2  FS3

File Server 2
FS1  FS2  FS3

File Server 3
FS1  FS2  FS3

SRM Server

File Server 1
FS1  FS2  FS3

File Server 2
FS1  FS2  FS3

File Server 3
FS1  FS2  FS3

data flow
Some Grid projects; originally by Vicky 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:
  - 364,389,470 GB (disk)
  - 330,703,439 GB (tape)
- Total processors available: 644,129

NorduGrid

- Conceived in 2001 as Scandinavian Grid
  - UiO heavily involved in coordination and development
- Now 93 sites in 20 countries
- Software: Advanced Resource Connector (ARC)
  - Computing Element
  - (Basic) Storage Element
  - Information System
- Scandinavian design principles: clean and simple!
# NorduGrid Monitor

## Overview

<table>
<thead>
<tr>
<th>Country</th>
<th>Site</th>
<th>CPUs</th>
<th>Load (processes: Grid+Local)</th>
<th>Queueing</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>BOINC Cluster</td>
<td>20</td>
<td>8393+378</td>
<td>0+0</td>
</tr>
<tr>
<td>Denmark</td>
<td>Steno Tier 1 (DCSC/KU)</td>
<td>6088</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td>Europe</td>
<td>EENet</td>
<td>392</td>
<td>704</td>
<td>0+0</td>
</tr>
<tr>
<td>Estonia</td>
<td>AceYLE (FGI)</td>
<td>72</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>Alcyone (CMS)</td>
<td>892</td>
<td>24621</td>
<td>27+0</td>
</tr>
<tr>
<td></td>
<td>Celaeno (FGI)</td>
<td>448</td>
<td>16+4+69</td>
<td>7+0</td>
</tr>
<tr>
<td></td>
<td>Asterope (FGI)</td>
<td>192</td>
<td>13416</td>
<td>60+0</td>
</tr>
<tr>
<td></td>
<td>DEA (CMS)</td>
<td>200</td>
<td>2035+73</td>
<td>163+0</td>
</tr>
<tr>
<td></td>
<td>Electra (FGI)</td>
<td>672</td>
<td>1784+183</td>
<td>278+56</td>
</tr>
<tr>
<td></td>
<td>Jade (HIP)</td>
<td>768</td>
<td>160+500</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>Malia (FGI)</td>
<td>768</td>
<td>156+912</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>Meropeno (FGI)</td>
<td>1812</td>
<td>10+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>Pielono (FGI)</td>
<td>288</td>
<td>23+8+4</td>
<td>46+0</td>
</tr>
<tr>
<td></td>
<td>Tegeta (FGI)</td>
<td>360</td>
<td>324+284</td>
<td>25+0</td>
</tr>
<tr>
<td></td>
<td>Trion (FGI)</td>
<td>3816</td>
<td>5613+139</td>
<td>10+0</td>
</tr>
<tr>
<td></td>
<td>Usva (CSC/FI/est)</td>
<td>144</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>LRZ-C2PAP</td>
<td>4032</td>
<td>30+816</td>
<td>272+0</td>
</tr>
<tr>
<td></td>
<td>LRZ-LMU</td>
<td>1484</td>
<td>153+376</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>LRZ-LMU-lgc-lrz-ce0</td>
<td>1492</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>RZG ATLAS HYDRA</td>
<td>16784</td>
<td>1024+1165</td>
<td>237+1191</td>
</tr>
<tr>
<td></td>
<td>wuppertalprod</td>
<td>3320</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>NIFi SC</td>
<td>768</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>IMC SUL</td>
<td>1</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>RTUETF</td>
<td>180</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>VU-MIF-LG2</td>
<td>1532</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>Abel C1(UO/IUSIT)</td>
<td>10872</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>Abel C2(UO/IUSIT)</td>
<td>10872</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>Abel C3(UO/IUSIT)</td>
<td>10872</td>
<td>222+979</td>
<td>2+0</td>
</tr>
<tr>
<td></td>
<td>EPF (UO/IF)</td>
<td>106</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>fimm (BCCS/UJB)</td>
<td>929</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>Arctus-1</td>
<td>432</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>Arnes</td>
<td>2244</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>atos</td>
<td>1417</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>CIPK e/BP</td>
<td>984</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>SIGNET</td>
<td>2834</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td></td>
<td>UNG</td>
<td>112</td>
<td>0+0</td>
<td>0+0</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td></td>
<td>327692</td>
<td>23893</td>
</tr>
</tbody>
</table>

Europe Sites: 81 Running jobs: 23893
ATLAS Data Processing

- Three main kinds of processing
  - Analysis of data
    - Plots and papers!
  - 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
  - Reprocessing data
    - Re-reconstruct the RAW data using improved software or knowledge of detector conditions
- Simulation is the most CPU-intensive process in LHC experiments
The dashed tracks are invisible to the detector.

Different type of particles -> different patterns.
Simulation steps:
- Event generation
- Detector simulation
- Track reconstruction
Current ATLAS Computing Model

Network proved better than anyone imagined
Any job can run anywhere
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

Average ATLAS traffic: 10GB/s

200 PB

Number of Physical Byte

273 Weeks from Week 01 of 2012 to Week 13 of 2017

2012 data

Simulations of 2012 data

2015 data

Simulations of 2015/6 data

2016 data

1 PB/day
Grid job management

Classic “push” model

Resource Broker

Job

Job

Pilot “pull” model

VO Job Management System

Get job

Pilot Factory

Pilot
Job stats

250,000 jobs running continuously

Pledged resources
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

- Computing architecture changes
  - More cores per CPU, less memory

- Need more CPU and disk but with flat budget -> opportunistic resources
  - High Performance Computing (supercomputers)
  - Volunteer Computing (general public)
Current Challenges

• … and getting more data than expected!
ATLAS resource needs at T1s & T2s

Extrapolation from 2014 (Moore law)
Software: from multi-process to multi-threading

AthenaMP (multiprocessing) will not be sufficient anymore
We will need (and we are developing) AthenaMT (multithreading)

Parallel processing in a multithreaded environment will come with its challenges both for developers, operations and infrastructures
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 environment
  – Multiple vs single CPU/OS flavours
  – username/password vs x509 certificate
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,
ATLAS and Chinese HPC Grid

**CNGrid environment**
- 14 sites
  - SCCAS (Beijing, major site)
  - SSC (Shanghai, major site)
  - NSCTJ (Tianjin)
  - NSCSZ (Shenzhen)
  - NSCJN (Jinan)
  - THU (Beijing)
  - IAIPCM (Beijing)
  - USTC (Hefei)
  - XJTU (Xi’an)
  - SIAT (Shenzhen)
  - HKU (Hong Kong)
  - SDU (Jinan)
  - HUST (Wuhan)
  - GSCC (Lanzhou)
- CNGrid Operation Center (based in Beijing)

**CNGrid Resources**
- 14 sites
- >3PF aggregated computing power
- >15PB storage

**CNGrid HPC**
- Tianhe-3A
  - #1 TOP 500, 2010
  - 4701 TFlop/s, 186,368 cores
  - Tianjin
- Sunway Blue Light
  - #14 TOP 500, 2011
  - ShenWei processor
  - 1076.2 TFlop/s, 137,200 cores
  - Jinan
- Nebulae
  - #2 TOP 500, 2010
  - 2984.3 TFlop/s, 120,840 cores
  - Shenzhen
- Dawning 5000A
  - #11 TOP 500, 2008
  - 233.5 TFlop/s, 30,720 cores
  - Shanghai
- DeepComp 7000
  - #19 TOP 500, 2008
  - 146.0 TFlop/s, 12,216 cores
  - Beijing
Volunteer Computing

- How YOU can help LHC experiments including ATLAS!
- Run simulation of collisions inside the ATLAS detector at home
History of Volunteer Computing at CERN

- 2004: LHC@Home: Sixtrack
  - LHC beam simulations
- 2011: LHC@Home: Test4Theory
  - Pioneered virtualisation within BOINC
- 2014: ATLAS@Home, CMS@Home, Beauty@Home (LHCb)
- 2017: All projects under one umbrella: LHC@Home
Volunteer Computing via BOINC
ATLAS@Home

- Producing ~1% of all ATLAS simulation
- 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
Volunteers

- BOINC has a huge dedicated community
- Many volunteers are active in several projects

http://boincstats.com/en/stats/152/project/detail/country
Who are the volunteers?

- **Credit**
  - Allocated by BOINC for successfully completed jobs
  - Very roughly 3 credits/event
  - No monetary value but a strong motivator

- **Golden rule**: always give credit for work done!
  - If a job crashes but used significant CPU, credit is still given
  - If a job is cancelled by ATLAS, don’t cancel volunteer’s job

---

### Top participants

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Recent average credit</th>
<th>Total credit</th>
<th>Country</th>
<th>Participant since</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MPI für Physik</td>
<td>25,262</td>
<td>55,683,198</td>
<td>Germany</td>
<td>20 Mar 2015, 09:41:04 UTC</td>
</tr>
<tr>
<td>2</td>
<td>WLCG Performance-Test Cluster</td>
<td>19,625</td>
<td>27,701,020</td>
<td>Switzerland</td>
<td>12 Sep 2016, 18:02:13 UTC</td>
</tr>
<tr>
<td>3</td>
<td>Claus Varming Lund</td>
<td>6,150</td>
<td>13,104,473</td>
<td>Denmark</td>
<td>8 Mar 2013, 13:53:43 UTC</td>
</tr>
<tr>
<td>4</td>
<td>poolgridcoin.co</td>
<td>6,036</td>
<td>6,390,160</td>
<td>International</td>
<td>16 Dec 2015, 19:53:54 UTC</td>
</tr>
<tr>
<td>5</td>
<td>kane</td>
<td>5,109</td>
<td>1,642,627</td>
<td>Germany</td>
<td>24 Jan 2017, 08:33:17 UTC</td>
</tr>
<tr>
<td>6</td>
<td>gorinle</td>
<td>4,488</td>
<td>5,310,887</td>
<td>Italy</td>
<td>21 Apr 2015, 14:11:43 UTC</td>
</tr>
<tr>
<td>7</td>
<td>taurec</td>
<td>4,308</td>
<td>1,004,054</td>
<td>International</td>
<td>26 Jun 2014, 09:54:40 UTC</td>
</tr>
<tr>
<td>8</td>
<td>Velodrome0815</td>
<td>4,184</td>
<td>1,947,228</td>
<td>Germany</td>
<td>18 Dec 2014, 14:17:43 UTC</td>
</tr>
<tr>
<td>10</td>
<td>BelgianEnthusiast</td>
<td>3,037</td>
<td>833,744</td>
<td>Belgium</td>
<td>5 Apr 2015, 10:31:40 UTC</td>
</tr>
<tr>
<td>11</td>
<td>Mnemonic</td>
<td>2,540</td>
<td>2,862,673</td>
<td>Switzerland</td>
<td>1 Aug 2014, 23:23:22 UTC</td>
</tr>
<tr>
<td>12</td>
<td>Kosice@CERN</td>
<td>2,351</td>
<td>3,806,834</td>
<td>Switzerland</td>
<td>23 Jul 2014, 11:06:00 UTC</td>
</tr>
</tbody>
</table>

[http://atlasathome.cern.ch/top_users.php](http://atlasathome.cern.ch/top_users.php)
UiO Involvement

• David C in charge of the project
• Last year a bachelor thesis by Giulio Isacchini to make a graphical interface
• Many more potential things to work on!
Join us!

• [http://lhcatlhome.cern.ch](http://lhcatlhome.cern.ch)

• BUT!
  – You need a fairly good machine to run ATLAS software
  – 4GB of RAM per job
  – And download a 500MB file to get started
Why not just use “the cloud”?  

- We do!
  - But on a small scale
  - Or in a way where we don’t notice it (eg CERN’s OpenStack 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 (but falling) using commercial cloud providers
Summary

• Distributed 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, especially looking to HL-LHC
• New technologies may save us: machine learning, analytics, etc