LHC Computing Grid today

*Did it work?*

Sept. 9th 2011, Günter Quast
The LHC is addressing fundamental questions of contemporary particle physics and cosmology:
- Origin of mass – the Higgs mechanism
- What is dark matter?
- Matter – antimatter asymmetry?
- Matter in the early universe ?
- hints for Quantum Gravity ??
- Dark Energy ???
Particle Physics is international teamwork

Working @Cern: 290 institutes from Europe, ~ 6349 Users
318 institutes elsewhere, ~ 3766 Users
Particle Physics international

Status: 2008

MEMBER STATES
- AUSTRIA: 68
- BELGIUM: 100
- BULGARIA: 39
- CZECH REPUBLIC: 174
- DENMARK: 57
- FINLAND: 97
- FRANCE: 838
- GERMANY: 962
- GREECE: 87
- HUNGARY: 43
- ITALY: 1475
- NETHERLANDS: 171
- NORWAY: 74
- POLAND: 168
- PORTUGAL: 104
- SLOVAKIA: 46
- SPAIN: 278
- SWEDEN: 72
- SWITZERLAND: 341
- UNITED KINGDOM: 666

5860

OBSERVER STATES
- INDIA: 94
- ISRAEL: 57
- JAPAN: 192
- RUSSIA: 905
- TURKEY: 61
- USA: 1447

2756

OTHER STATES
- ARGENTINA: 10
- ARMENIA: 17
- AUSTRALIA: 15
- AZERBAIJAN: 4
- BELARUS: 18
- BRAZIL: 68
- CANADA: 127
- CHILE: 8
- CHINA: 68
- COLOMBIA: 12
- CROATIA: 22
- CUBA: 6
- CYPRUS: 7
- ESTONIA: 12
- GEORGIA: 11
- GHANA: 1
- ICELAND: 2
- IRAN: 13
- IRELAND: 10
- KOREA: 51
- LITHUANIA: 10
- MADAGASCAR: 1
- MALAYSIA: 1
- MALTA: 4
- MEXICO: 36
- MONTENEGRO: 1
- MOROCCO: 5
- NEW ZEALAND: 8
- PAKISTAN: 29
- PALESTINIAN TERR.: 1
- ROMANIA: 49
- SAUDI ARABIA: 2
- SERBIA: 15
- SLOVENIA: 17
- SOUTH AFRICA: 8

744

SRI LANKA: 1
TAIWAN: 46
THAILAND: 1
UKRAINE: 20
U.A.E.: 4
VIETNAM: 3

Fakultät für Physik
Institut für Experimentelle Kernphysik

GridKa School
Sept. 2011

Günter Quast
LHC Computing Grid today
Grid for LHC

Given the international and collaborative nature of HEP Computing must be distributed

→ harvest intellectual contributions from all partners, also funding issues

Early studies in 1999 (Monarc Study group) suggested a hierarcical approach, following the typical data reduction schemes usually adopted in data analysis in high energy physics

Grid paradigm came at the right time and was adopted by LHC physicists as the base line for distributed computing

Major contributions by physicists to developments in Grid computing

Other HEP communities also benefit and contributed
WLHC Computing Grid in action

A truly international, world-spanning Grid for LHC data processing, simulation and analysis

Source: http://rtm.hep.ph.ic.ac.uk/webstart.php
WLHC Computing Grid in action

A truly international, world-spanning Grid for LHC data processing, simulation and analysis

Source: http://rtm.hep.ph.ic.ac.uk/webstart.php

Scheduled = 21539
Running = 25374

21:13:50 UTC
Organisation of the World-wide LHC computing Grid

Grids can’t work without an organisational structure representing all parties involved
Structure of the LHC Grid

A grid with hierarchies and different tasks at different levels

In addition, it is a “Grid of Grids” with interoperability between different middlewares:
- EGEE middleware in most of Europe
- Open Science Grid in USA
- NorduGrid in Northern Europe
- Alien by the Alice collaboration
LHC Experiments are huge data sources

- ~ 100 Million detector cells
- LHC collision rate: 40 MHz
- 10-12 bit/cell
  → ~1000 Tbyte/s raw data

- Zero-suppression and Trigger reduce this to "only" some 100 Mbyte/s
  i.e. 1 /sec

Grid computing
ideal for the job!
CPU-intense simulation of particle physics reactions and detector response

processing (=reconstruction) of large data volumes

I/O-intense filtering and distribution of data

transfer to local clusters and workstations for final physics interpretation
Why Grid is well suited for HEP

Experimental HEP codes - key characteristics:

- modest memory requirement (~2GB) & modest floating point
  → perform well on PCs
- independent events
  → easy parallelism
- large data collections (TB → PB)
- shared by very large collaborations
WLCG: Grid with hierarchical structure

**Tier-0**
- **the accelerator centre**
  - Data acquisition & initial processing
  - Long-term data curation
  - Distribution of data to T1/T2

**Tier-1**
- 11 Tier-1 Centres
  - “online” to the data acquisition process → high availability
  - Managed Mass Storage → grid-enabled data service
  - Data-intensive analysis
  - National, regional support

**Tier-2**
- 150 Centres in 60 Federations in 35 countries
  - End-user (physicist, research group) analysis & collaboration with T3 (= institute resources) – where the discoveries are made
  - Monte Carlo Simulation

**Tier-3**
- several 100 grid-enabled PC clusters @ institutes
Grid Structure

Example: CMS computing model

LHC-Experiments
- typically share big Tier-1s, take responsibility for experiment-specific services
- have a large number of Tier2s, usually supporting only one experiment
- have an even larger number of Tier-3s without any obligations towards WLCG
Typical workflows on the Grid

User-Interface

1. Input Sandbox

WMS

1. Rechenanforderung

Workload Manager

Informations-System

2. Discovery

Sammelt Informationen über alle Grid-Zentren

Grid-Zentrum

3. Weitergabe an CE

Computing Element

4. Daten-Transfer

Storage Element

5. Output Sandbox

SiteBDII

Stellt Ressourcen-Informationen bereit

Lokaler PC

Benutzer schickt Grid-Job an WMS

Günter Quast

LHC Computing Grid today
How big is WLCG today?

**CPU (kHS06) T0 - T1 - T2**

- **Total CPU 2011:**
  - 1500 kHS06
  - approx. equiv.
  - 150'000 CPU cores

- **Disk (PB) T0 - T1 - T2**

  - **Total Disk 2011:**
    - 130 PB,
    - same amount as
  - **Tape Storage 2011:**
    - 130 PB
    - (40 PB at CERN, none @ T2s)

2012 numbers still being negotiated!

**The largest Science Grid in the World**
A closer look to the surroundings of GridKa

GridKa supports >20 T2s in 6 countries, provides ~15% of WLCG T1 resources

Alice T2 sites in Russia

Most complex “T2 cloud” of any T1 in WLCG
After almost 2 years of experience with LHC operation:

How well did it work?
Almost 2 years of experience - Did it work?

Up to the users to give feedback: D. Charlton, ATLAS, EPS HEP 2011

Computing Grid Delivers Physics

Data preparation:
- First-pass reco. at Tier-0 within ~2 days
- Calibration/DQ good for physics analysis
- Data analysable on Grid within ~1 week

Tier-1 and Tier-2's process ~3/2 M jobs per day
- simulation
- re-reconstruction (campaigns)
- group production (ntuples...)
- physics analysis

The high quality computing system allows us to show results on data taken until the end of June

Payback for the years of investment and hard work
Did it work (2)

G. Tonelli, CMS, EPS HEP 2011

Offline and Computing running smoothly

- Smooth Tier-0 operation, keeping up with the data taking. Increase in Tier-1 utilization, for reprocessing and skimming jobs; High usage of Tier-2 for analysis, >400 (800) individual users per week (month). More than 5.4 Billions MC events.

- CMS PhEDEx - Transfer Rate
  - T1→T2, last 12 months
    - (each color is a T1-T2 link, >400 links)

- Efficiency based on success/failures
  - 67 Weeks from Week 13 of 2010 to Week 26 of 2011

- T2 analysis and simulation jobs
  - (each color is a different T2 site)

- Events, total ~ 5.4 Billion MC events

G. Tonelli, CERN/INFN/UNIPI

HEP_2011_GRENoble

July 25 2011
Did it work?

Obviously it did!

- Grid infrastructure for the LHC performed extremely well
- physics results from freshly recorded data
- but: effort for running computing infrastructure is high!

There are challenges ahead!

Let`s have a look in detail ...
Data Export Rate from CERN

Averaged Throughput From 03/04/11 To 03/09/11
VO-wise Data Transfer From CERN-PROD To All Sites

Throughput (MB/s)

Date (dd/mm)

03/04 08/04 13/04 18/04 23/04 28/04 03/05 08/05 13/05 18/05 23/05 28/05 02/06 07/06 12/06 17/06 22/06 27/06 02/07 07/07 12/07 17/07 22/07 27/07 01/08 06/08 11/08 16/08 21/08 26/08 31/08

Throughput (MB/s)

0 500 1000 1500 2000

Data Export Rate from CERN

GridKa School
Sept. 2011

Günter Quast

LHC Computing Grid today

Fakultät für Physik
Institut für Experimentelle Kernphysik
Daily data export from CERN

Aggregate Data Movement From 03/04/11 To 03/09/11
VO-wise Data Transfer From CERN-PROD To All Sites

- Atlas
- CMS
- LHCb
- OTHERS
- UNREGD VOs
- compass

Date (dd/mm)
Aggregate Data Mov (Bytes)
200T
150T
100T
50T
03/04
06/04
13/04
18/04
23/04
03/05
08/05
13/05
18/05
23/05
03/06
08/06
13/06
18/06
23/06
03/07
08/07
13/07
18/07
23/07
03/08
08/08
13/08
18/08
23/08
31/08
Data Export from T1: example: CMS @ KIT

CMS PhEDEx - Transfer Rate
26 Weeks from Week 10 of 2011 to Week 36 of 2011

Maximum: 117.55 MB/s, Minimum: 2.32 MB/s, Average: 40.60 MB/s, Current: 11.49 MB/s
Data Import T2->T1 (MC): example CMS@KIT
Processed Jobs  ATLAS

Running jobs
30 Days from 2011-08-04 to 2011-09-03

Processed Jobs ATLAS

BNL-ATLAS
IN2P3-CC
NDGF-T1
DESY-HH
SWT2_CPB
SARA-MATRIX
PIC
PRAGUELCG2
CSCS-LCG2
DESY-ZN
MWT2_UC
TAIWAN-LCG2
LRZ-LMU
SIGNET
UKI-SCOTGRID-GLASGOW
MPPMU
UNI-DORTMUND
UKI-NORTHGRID-LANCS
UKI-FREIBURG
IN2P3-LAPP
FZK-LCG2
RAL-LCG2
CYFRONET-LCG2
UKI-LT2-QMUL
UKI-LT2-RHUL
WT2
AGLT2
CERN-PROD
UKI-NORTHGRID-MANHEM
HEP2P3-CC-T2
TOKYO-LCG2
INFN-NAPOLI-ATLAS
IFIC-LCG2
TRIUMF-LCG2
IN2P3-LPC
INFN-MILANO-ATLASC
OU_OCHEP_SWT2
INFN-T1
GRIF
SE-SNIC-T2
IFAE
GOEGRID
WUPPERTALPROD
UTA_SWT2
UKI-SOUTHGRID-RALPP
... plus 54 more

Processed jobs success rates (ATLAS)

Terminated Jobs per site

- BNL-ATLAS
- FZK-LCG2
- RAL-LCG2
- CERN-PROD
- IN2P3-CC
- INFN-T1
- AGLT2
- MWT2_UC
- GRIF
- TAIWAN-LCG2
- LRZ-LMU
- SARA-MATRIX
- TOKYO-LCG2
- PIC
- NIKHEF-ELFRED
- DESY-HH
- IN2P3.CC-T2
- UKI-SCOTGRID-GLASGOW
- TRIUMF-LCG2
- UKI-LT2-QMUL

Legend:
- terminated
- app-successful
- app-failed
- site-failed
- aborted
- canceled
- app-unknown

Sept. 2011

Günter Quast
Processed Jobs success rates (CMS)

Terminated Jobs per site

- T2_US_Wisconsin
- T1_US_FNAL
- T0_CH_CERN
- T3_US_FNALLPC
- T2_US_Purdue
- T2_DE_DESY
- T2_US_Nebraska
- T1_DE_KIT
- T2_US_MIT
- T2_UK_London_IC
- T2_UK_SGrid_RALPP
- T1_IT_CNAF
- T2_US_Florida
- T2_DE_RWTH
- T2_US_UCSD
- T1_UK_RAL
- T2_US_Caltech
- T2_ES_CIEMAT
- T1_TW_ASGC
- T2_IT_Legnaro

- terminated
- app-successful
- app-failed
- site-failed
- aborted
- canceled
- app-unknown
Job success rates @ T2s

(CMS T2s shown here
ATLAS looks similar)

T2s see a mixture of
- MC production jobs
- user analysis and skimming jobs

Job success rates not in all cases above 90%

90% success rate would be considered very low for a classical computer centre

This must improve …

Not easy to disentangle failures of the system from “user errors”
Site Reliability (examples)
T2 performance

“Availability” of CMS T2 sites

There are sites with performance issues!

Typically, less well performing sites are very small!
Site Availability using WLCG_NAGIOS
30 Days from 2011-08-04 to 2011-09-03
This is an example of a time-resolved measurement of site availability.

Message similar as previously.

This kind of graphs helps the site responsibles to monitor their sites and act on problems.

Provided centrally by WLCG and expriments!
Again: Does it work?  

YES!

- Routinely running ~150'000 jobs simultaneously on the Grid
- Shipping over 100 TB/day to T1 centres
- Data distribution to T2 works well
- Some T2s have performance issues
- Very little is known about T3 usage and success rates - responsibility of the institutes
- Plenty of resources available at LHC start-up, now approaching “resource limited operation”
- Users have adapted to the “GridWorld” - Grid is routinely used as a huge batch system, output is transferred home

but ...
Does it work?

Message:

it worked better than expected by many,
but running such a complex computing infrastructure
as the WLCG is tedious (and expensive!)

Reliability and cost of operation can be improved by
- simplified and more robust middleware
- redundancy of services and sites,
  requires dynamic placement of data and investment in network bandwidth
- automated monitoring and triggering of actions
- use of commercially supported approaches to distributed computing:
  - private clouds are particularly important for shared resources at universities
  - eventually off-load simple tasks (simulation, statistics calculations) to commercial clouds

Many of the new developments were addressed at this School

Let`s have a look at some future developments ...
**LHCONE**

ATLAS and CMS computing models differ slightly

CMS already more “distributed”

**Aim of LHCONE project is**

better trans-regional networking for data analysis,
complementary to **LHCOPN** network connecting LHC T1s

- **flat(er) hierarchy:** any site has access to any other site's data
- **dynamic data caching:** pull data “on demand”
- **remote data access:** jobs may use data remotely

by interconnecting open exchange points between regional networks
A dedicated HEP network infrastructure – what is the cost?
Virtualisation

You have heard a lot about clouds and virtualisation at this school in a nutshell:
- Clouds offer “Infrastructure as a Service”
- easy provision of resources “on demand”
  even by including (private) cloud resources as a classical batch queue
  (e.g. ROCED project developed at EKP, KIT)
- independent of local hardware and operating system
  (Scientific Linux 5 for Grid middleware and experiment software)
**CernVM & CernVM-FS**

**CernVM** is a virtual machine ("Virtual Software Appliance") based on Scientific Linux with CERN software environment, runs on [VirtualBox](https://www.virtualbox.org/), [VMware](https://www.vmware.com/), [KVM](https://www.linux-kvm.org/), [Xen](https://www.xen.org/).

**CernVM-FS** is a client-server file system based on http and implemented as a user-space file system optimized for read-only access to software repositories with a performant caching mechanism. Allows a CernVM instance to efficiently access software installed remotely.

![Diagram showing the components of CernVM-FS and their interactions](image-url)
A recap of this School

This week, you have heard about many of the new developments:

- J. Templon, *Grid and Cloud*
  
  “Grids need Clouds to prosper, Clouds need Grids to scale”

- O. Synge, *Virtualisation*
- P. Millar, *Data Storage*
- C. Witzig, *European Grid Projects*
- T. Beckers, Storage Architectures for Petaflops Computing
- S. Reißer, *Grid User Support*
- N. Abdennadher, Combining Grid, Cloud and Volunteer Computing
- U. Schwickerath, *Cloud Computing*
- S. Maffioletti, *ARC for developers*
- T. Metsch, B. Schott, Sustainable DCI Operations
- A. Aeschlimann, *Grid and Cloud Security*

and a number of Hands-on workshops in parallel sessions.

HEP Grid runs fine in its initial version, but virtualisation and “clouds” offer new possibilities for resource increase, efficiency, cost effectiveness, operation and reliability.

It's up the you, the participants of this school, to shape the future!

Thanks to all speakers, session teams and organizers and also from my side.