Unraveling Grid Computing:
From Basics to WLCG

By Robin Hofsaess (Robin.Hofsaess@cern.ch)

"A complex, chaotic scene depicting a globe with numerous computing resources such as servers, data centers, and computers distributed unevenly around it. These resources are interconnected by a tangled, knotted thread, symbolizing a complicated and disorganized network." (DALL·E)
Hi there!

I am Robin Hofsaess, a PhD student in Physics at KIT, Germany. Currently, I am working on efficiency and workflow optimizations for HEP jobs on Opportunistic Resources, like HPC centers, integrated in GridKa, the German T1 center. My main interests are computers, music (guitar), biking, gaming, and also a little bit of physics :D
2023!

~200,000,000 GB of physics data

>1.710.000.000 GB transferred via FTS

367,176,289 Number of compute jobs
The data rates of the HL-LHC will expectedly increase the resource demand drastically.

With constant research and development (and funding), we expect to stay within the projections.

Source
The (20 years old) solution:

Grid Computing
01 What is Grid Computing?
A historical overview: from basics to grid computing

02 How Does a Computing Grid Look Like?
Learning the concepts behind a modern computing grid

03 The Worldwide LHC Computing Grid (WLCG)
Let’s have a look at one of world’s biggest grids!
Computing is the process of utilizing computer technology to complete tasks or solve problems through the execution of algorithms and manipulation of data.
What is Grid Computing?

A small (historical) overview and definition
“A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities.”

—Ian Foster and Carl Kesselman (The Grid: Blueprint for a New Computing Infrastructure, 1998)
The Journey to Grid Computing

- ** PCs made computing possible and popular **
- ** Sharing is caring **

**Personal Computers**
- ‘50s–’60s
- ‘70s–’80s
- ‘80s–’90s
- ‘90s–today

**Network**
- LAN and WAN enabled distributed computing

**Mainframes**
- It’s the size that matters!

**Distributed**
- Grid
- Cluster
- Cloud
- …
1950s–1960s: Mainframe Computing

- Computing dominated by centralized, large mainframe computers
- Very rare and expensive
- Terminal access for multiple users on site
- 1958: CERN’s first computer

In this era, computing was very exclusive. Particle physics typically wasn’t that Privileged... Analysis was mainly done by hand!

https://cerncourier.com/a/in-the-tracks-of-the-bubble-chamber/
The Journey to Grid Computing

1960s: Mainframe Computing

- 1960: IBM 709 as second Computer

“With Mercury and the 709 operating together, CERN had its first experience of compatibility problems. This was a continuing source of difficulty as various different computers came into operation at CERN.”

⇒ The first heterogeneous IT infrastructure at CERN!
(Already with its merits and pitfalls...)

- With the 709: Introduction of FORTRAN
The Journey to Grid Computing


- PCs made computing power broader accessible
- Shift to decentralized computing
- Computer Science strongly gained in popularity

Businesses, laboratories, and even individuals could own and operate their own computing resources.
1980s–1990s: Network Computing

- LAN and Internet enabled communication between computers
- Resource (and data) sharing over the network became possible!
- First client-server architectures:
  - multiple clients could request services from centralized servers
- First computing clusters (multiple computers interconnected)
- Beginning of the *Internet* at CERN
The Journey to Grid Computing

Also 1990s: Parallel and Distributed Computing

- Start of parallel usage of multiple (distributed) computers for the same problems (not necessarily at the same place!)
- Many new concepts in software development (and collaboration)
  - E.g. breaking down tasks into smaller sub-tasks that could be processed concurrently across different processors or computers
- First public distributed volunteer computing projects:
  - GIMPS
  - SETI@home
  - distributed.net

Folding@home was the first exa-scale computing system during Covid in 2020!
The Journey to Grid Computing

2000s: Grid Computing

- Grid computing built upon the earlier stages by focusing on connecting computers across multiple organizations
- 2001: The European DataGrid project was introduced to develop a production quality computing grid to pave the way for the LHC Computing Grid (LCG)
- 2003: First prototype with 25 sites worldwide (LCG-Phase 1)
- Today: widely established concept in science and industry
What is Grid Computing?

- Distributed computing paradigm with:
  - Shared heterogeneous computational resources across multiple (geographic disperse) administrative domains loosely coupled over network and controlled centrally (but not managed!)
  - Distributed users of Virtual Organizations (VO) with a common access interface
  - Goal: Collaboration on complex (compute) projects across different geographic and institutional boundaries to solve a common large-scale problem
- The naming “Grid” is inspired by the power grid
The grid

Identical access

Distributed users

Distributed, heterogeneous resources

Adapted from: Source (CC BY 3.0)
What is Grid Computing?

- Key aspects:
  - Coordinated hetero. Resources
  - Standardized software
  - Scalability and flexibility
  - Reliability and redundancy
  - Cost efficiency
  - Collaboration and shared contributions
“Grid Computing is a distributed computing paradigm that involves a coordinated sharing of heterogeneous, flexible, loosely coupled resources across dynamic and geographically dispersed organizations with the goal to create a "virtual supercomputer" that is able to solve complex, large-scale problems.”

– own definition
Further Reading

- The History of the Grid
- Grid History and Standards
How Does a Computing Grid Look Like?

Let’s have a more detailed look at the key concepts and core components of a modern computing grid.
Key Components of a Grid

- Hardware
- Software and Tools
- Organizational
Key Components of a Grid

Servers:
- Compute
- Databases
- Monitoring
- ...

Storage:
- Tape
- Disk

Hardware

Network:
- Switches
- Routers
- Firewalls
- ...

TOpAS

HoreKa

GridKa
Key Components of a Grid

**Grid middleware:**
- Resource management
- Security
- User interfaces
- Backend for further services
- Interoperability

**Workflow management:**
- defining workflows
- job submission and resource allocation
- job monitoring

**Virtualization:**
- mostly containers

**Data management:**
- data placement
- data transfers and replication
- data access

**Software and Tools**

**Compute Element:**
- E.G. HTCondor or ARC

**Frontend:**
- e.g. CRAB + CMSSW

**Monitoring:**
- grid status
- metadata
- jobs
Key Components of a Grid

Grid Initiatives:
- R&D
- Support
- Security
- Standardization
- Accounting

Organizational

Administration:
- Governance
- VOs
- Collaborators
- Policies

Institutions:
- Contribute resources and manpower

Politics:
- e.g. dCMS

Pledges: CRIo
Further Reading

- What is Grid Computing (AWS)
- https://www.egi.eu/
The Worldwide LHC Computing Grid (WLCG)

World’s biggest most sophisticated open source, scientific computing grid!
(Unfortunately, it’s not the biggest – we probably cannot challenge AWS EC2 ...)
WLCG: The Beginning

• End 1994: LHC was approved

Requirements on the IT infrastructure:

○ The storage and efficient processing of PBs of data
○ Distribution of users and copies of the data around the world
○ Collaboration and shared contributions
○ Reliability, redundancy (of services), and sustainability
○ Scalability and cost effectiveness
○ National interests (e.g. of funding agencies)

• 1998: The Models of Networked Analysis at Regional Centers for LHC (MONARC) project was initiated to find a feasible architectural model for the LCG
Instead of a giant computing center, we seek a distribution of our compute and storage resources.
WLCG: Hierarchical Structure

- **“Tiers”** of the WLCG:
  - **T0:**
    - data centre at CERN
    - 20% of total compute
    - first copy of data
  - **T1:**
    - backup share of data
    - dist. of data to T2s
  - **T2:**
    - provide significant compute power
  - **T3:** local analysis groups
WLCG: The Beginning

- End 1994: LHC was approved
- 1998: The MONARC project was initiated to find a feasible architectural model for the LCG
- 2002: The **success** led to the LCG proposal
- 2005: The LCG Technical Design Report was published
- 2008: Beginning of the LHC and LCG operations
WLCG: Today
Interactive map

>160 sites
~ 40 countries

More info and visualization of transfers
WLCG: Network

E.g. monitoring for KIT:
- I
- II
- III
WLCG: Network

Monitoring

Total traffic:
- LINK 1
- LINK 2
WLCG: Overview

~1,400,000

CPU cores within WLCG

1,500,000 > 800,000

TB tape storage space  TB disk storage space

112,596 > 260

Concurrently running jobs  GB/s global transfer rate

Source
## WLCG: Benefits and Challenges

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Massive data handling</td>
<td>● Data management and access</td>
</tr>
<tr>
<td>● Enables worldwide research</td>
<td>● Security concerns</td>
</tr>
<tr>
<td>● Collaboration and sharing</td>
<td>● Interoperability</td>
</tr>
<tr>
<td>● Eff. resource utilization</td>
<td>● Network and latency</td>
</tr>
<tr>
<td>● Cost efficient</td>
<td>● Policies</td>
</tr>
<tr>
<td>● Easy to contribute</td>
<td>● Complexity (management and usage)</td>
</tr>
<tr>
<td>● Reliable and flexible</td>
<td>● Default user...</td>
</tr>
</tbody>
</table>
WLCG: Benefits and Challenges

- Massive data handling
- Enables worldwide research
- Collaboration and sharing
- Efficient resource utilization
- Cost efficient
- Easy to contribute

WLCG with a 2 page pdf manual

WHAT GIVES PEOPLE FEELINGS OF POWER

- Money
- Status
- 10000 Running Jobs

i prefer running on laptop, that's more easy

AVERAGE USER

physics analysis

CA 2024, COLORIZED

What gives people feelings of power:

- Money
- Status
- 10000 Running Jobs

@iamnotnarkid...
Further Reading

- MONARC report
- Good overview: https://wlcg-public.web.cern.ch/structure
- WLCG public side: https://wlcg-public.web.cern.ch/
- https://wlcg-public.web.cern.ch/tiers
Bonus: How about Cloud Computing?

Is a Cloud a Grid?

NO PATRICK

A CLOUD IS NOT A GRID
“A **Cloud** is a model for enabling reliable, convenient, on-demand network access to a shared pool of configurable computing resources (networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.”

– own definition
Cloud Computing is also a form of distributed computing that provides on-demand access to a shared pool of resources over the internet. Fixed definition by NIST. Different types: public, hybrid, private cloud. Different models: IaaS, PaaS, SaaS. Underlying infrastructure not specified. It offers scalable and elastic services with a pay-as-you-go pricing model, mainly focusing on business and consumer applications.
### Comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cloud</th>
<th>Grid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributed computing</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Scalable, flexible</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Cost efficient</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Reliability and redundancy</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Enable large scale data analysis</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Grid vs. Cloud Computing

**Grid**
- Decentralized heterogeneous resources from many organizations
- Centrally controlled
- More complex access mechanisms
- Purpose: common goal (e.g. large scale research applications)
- One task per job

**Cloud**
- Centrally managed homo. resources belonging to one organization
- User controlled
- Easy access over the web
- Purpose: elastic provisioning of non-related ready-to-use services
- Complex services (like CDNs)

⇒ A cloud can be deployed on a grid, a cloud can be a part of a grid, but a grid is not a cloud!
Example: The Tier-0 at CERN

“[...] Over 90% of the resources for computing in the Data Centre are provided through a private cloud based on OpenStack, an open-source project to deliver a massively scalable cloud operating system.”

Source Monitoring
Bonus: How about Cloud Computing?

Could we replace the Grid by a (Commercial) Cloud?
Bonus: How about Cloud Computing?

Sure, Why not?

Well...
**Well ...**  
Example: AWS  
(very rough estimate for 2023)

<table>
<thead>
<tr>
<th>Storage: (Speed=???, prob. too low)</th>
<th>Transfers: (tape only: ~50GB/s avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 900PB disk: 18.900.000 USD</td>
<td>1.577.000.000GB: 94.600.000 USD</td>
</tr>
<tr>
<td>- 1.5EB tape: 6.000.000 USD</td>
<td></td>
</tr>
</tbody>
</table>

Compute: (with 100G WAN)  
1.280.000 cores: 693.786.744 USD

**Does not include:**  
- Other transfers  
- Databases  
- Monitoring  
- Person power  
- ...

**Very rough estimate:**  
**1.000.000.000 USD per year**  
+ Vendor lock-in ...
Thanks!