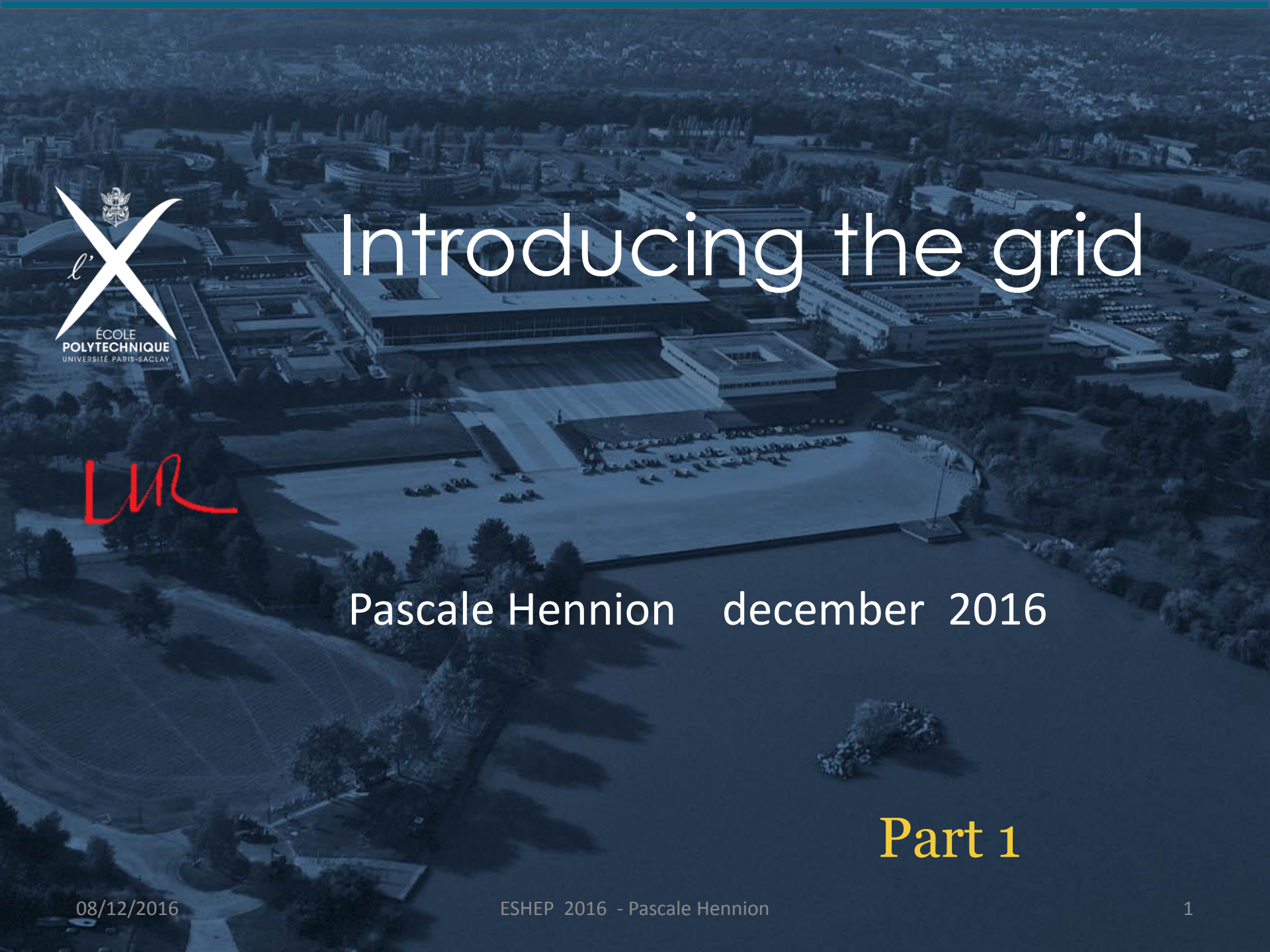


Introducing the grid

Pascale Hennion december 2016

Part 1

Outline of part 1

1. Introducing the grid
    1. What is it;
    2. Difference between cluster, cloud and grid
    3. Hourglass model;
    4. Examples;
2. Introducing WLCG
    1. LHC computing
    2. Activity
3. WLCG Architecture
    1. Tiers
    2. Information system
    3. Authentication
    4. Data management
    5. Jobs workflow
    6. VO's applications

Outline of part 1

1. Introducing the grid
    1. What is it;
    2. Difference between cluster, cloud and grid
    3. Hourglass model;
    4. Examples;
2. Introducing WLCG
3. WLCG Architecture

Introducing the GRID

‘‘ The term «The Grid» was coined in the **mid-1990s** to denote a (then) proposed **distributed computing infrastructure** for advanced science and engineering. ’’

The Anatomy of the Grid. Ian Foster, Carl Kesselman, Steven Tuecke.

- Born during «Building a Computational Grid» workshop at ANL (8-10 Sep 1997)
  - http://www.crpc.rice.edu/newsletters/fal97/news_grid.html;

Ian Foster

- some **fundamental texts**:
  - The Grid: Blueprint for a New Computing Infrastructure. I.F, C.K.
  - Grid Services for Distributed System Integration. I.F., C.K., J. Nick, S. Tuecke. Computer, 35(6), 2002.
  - The Anatomy of the Grid: Enabling Scalable Virtual Organizations. I.F., C. K.C., S. Tuecke. International J. Supercomputer Applications

Carl Kesselman

What is it ?

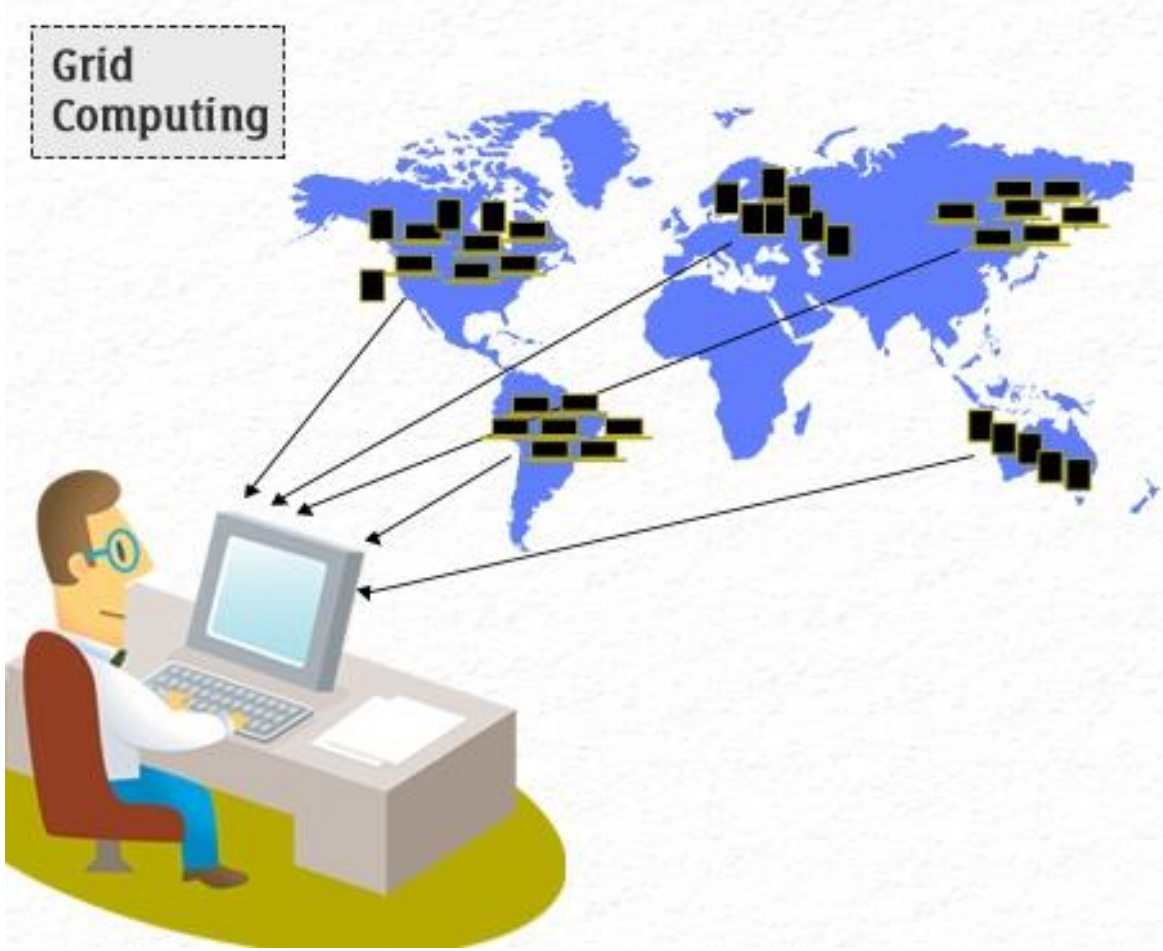
Coordinated resource sharing in dynamic, multi-institutional virtual organizations (VO's)

- sharing is not only file exchange but rather **direct access to computers, software, data**, and other resources;
- it is, necessarily, **highly controlled**, with resource providers and consumers defining clearly and carefully just what is shared, who is allowed to share, and the conditions under which sharing occurs.
- All of which may **vary dynamically**;

Virtual Organizations (VO's)

- a **group** of mutually distrustful participants with **varying degrees of prior relationship** (perhaps none at all) that **want to share resources** in order to perform some tasks.

A vision of the grid



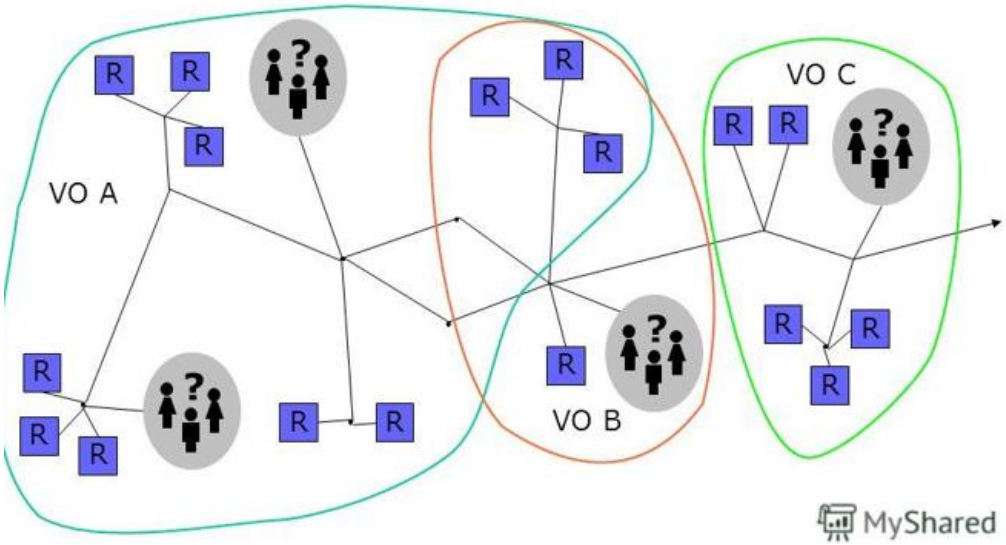
The VO's

A group of participants that **want to share resources**

The users of a VO can belong to any site

One user can belong to more than one VO

Each site choose which VO it wants to support



A grid is :

- Many sites
  - Filled with any kind of servers, for calculation as well as data storage
- All the sites connected by a fast network
- A few servers/softwares that make the grid work
- Tons of users and different applications

To achieve that, we need:

- Network
  - link together geographily distributed resources and allow then to be used collectively
- Software
- Standards
  - Grid standards for developpers and users
  - Mandatory
- Coordination and partnerships
  - among the different actors of the Grid

Outline of part 1

1. Introducing the grid
    1. What is it;
    2. Difference between cluster, cloud and grid
    3. Hourglass model;
    4. Examples;
2. Introducing WLCG
3. WLCG Architecture

Clusters and Super Calculator

- A cluster is :
  - A bunch of computers in a single location
  - All identical
  - Connected by a low latency network (Infiniband)
  - For parallel jobs
  - MPI (message passing interface) to pass informations between process
  - Expensive (IB => 30% extra cost)
- Supercalculator = huge cluster
- In HEP :
  - No need for supercalculator' fast network
    - Process each event individually
    - No need for communication between jobs
  - But can easily use a suprcalculator (Embarrassingly parallel)

Cloud

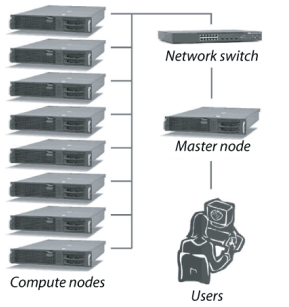
- Recent technology
- Both for data and computers
- Computer = > virtualisation
- 2 possibilities for HEP
  - Cloud federation
    - The user choose the exploitation system, the amount of memory, the number of CPU …
    - Possibility to get a VM from any cloud of the federation
  - Any computer of the grid can be a virtual machine
    - A small amount of CPU is lost because of virtualisation
- Tests ongoing to use commercial cloud for HEP computing

Grid not dead

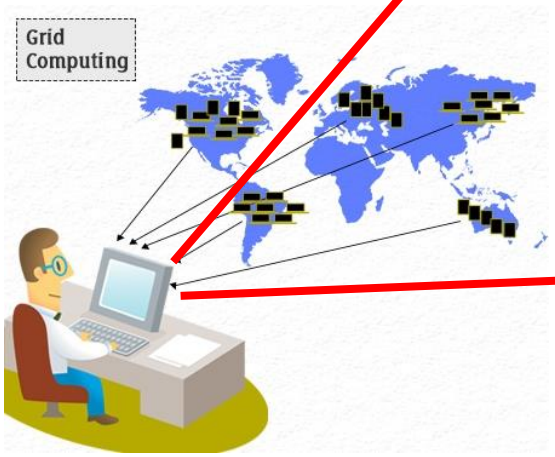
- Grid users like it:
  - HEP tuned
  - Cheaper than cloud or clusters
  - Practice it for ~10 years
  - Strong authentication
- But it has limitations:
  - Very few different exploitation system
  - 2Gb memory/core
  - Strong authentication
- Future: more data to arrive (higher luminosity) . HEP will have to make use of any available computer time

Future ?

HPC cluster



Source: ILRI High Performance Computing



Outline of part 1

1. Introducing the grid
    1. What is it;
    2. Difference between cluster, cloud and grid
    3. Hourglass model;
    4. Examples;
2. Introducing WLCG
3. WLCG Architecture

Hourglass Model

Wide top: various users applications

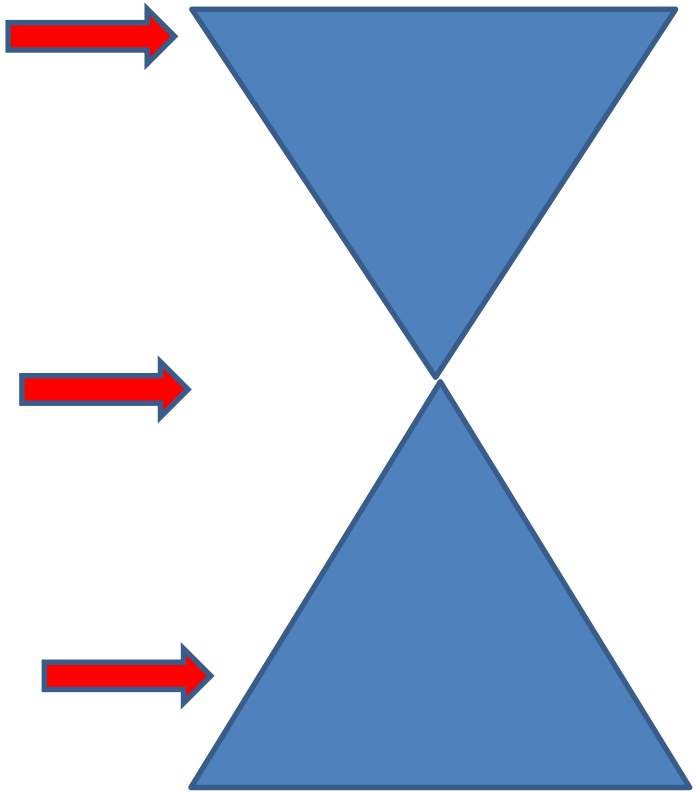
Application layer

Thin center: few standards

Middleware layer

Wide bottom: many types of hardwares

Resource layer



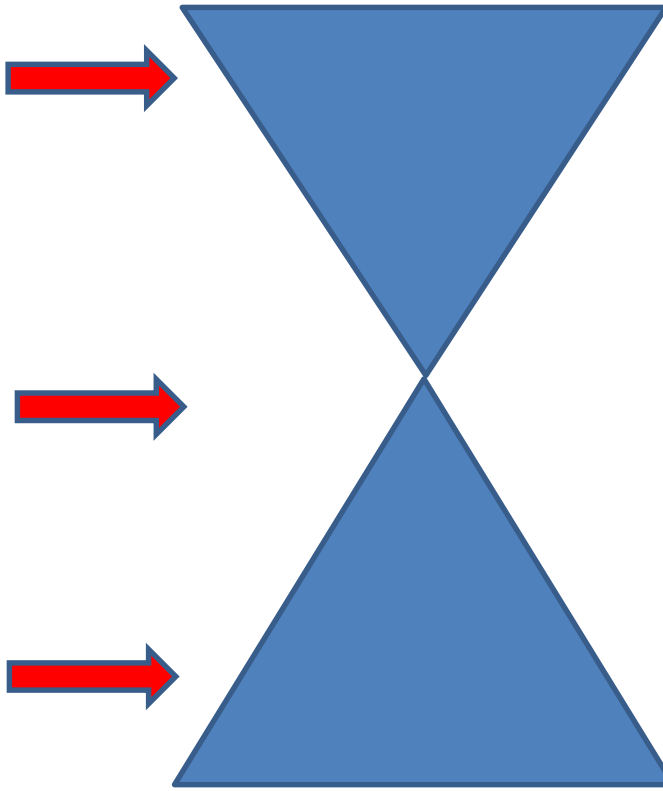
Layers

Application layer : applications in science, engineering, business, finance and more, as well as portals and development toolkits to support the applications.

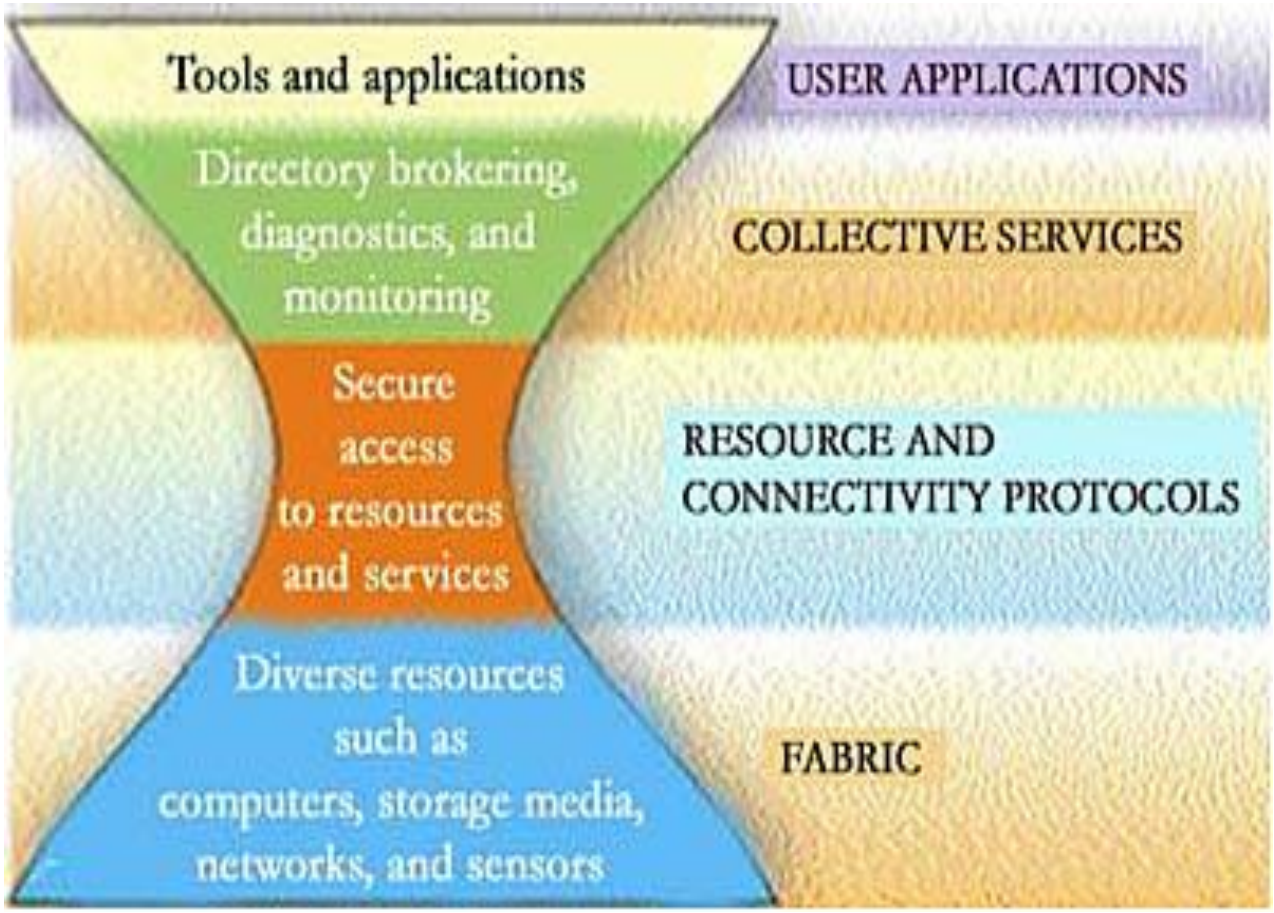
This is the layer that grid users "see" and interact with.

Middleware layer provides the tools that enable the various elements (servers, storage, networks, etc.) to participate in a grid. The middleware layer is sometimes the "brains" behind a computing grid!

Resource layer: actual grid resources, such as computers, storage systems, electronic data catalogues, that are connected to the network.



Hourglass Model



Outline of part 1

1. Introducing the grid
    1. What is it;
    2. Difference between cluster, cloud and grid
    3. Hourglass model;
    4. Examples;
2. Introducing WLCG
3. WLCG Architecture

Example 1

- **World Community Grid** enables anyone with a computer, smartphone or tablet to **donate** their **unused computing power to advance cutting-edge scientific research on topics related to health, poverty and sustainability.** Through the contributions of **over 650,000 individuals and 460 organizations**, World Community Grid has supported **25 research projects** to date, including searches for more effective treatments for cancer, HIV/AIDS and neglected tropical diseases. Other projects are looking for low-cost water filtration systems and new materials for capturing solar energy efficiently.
- …
- Started in 2004, World Community Grid is a **philanthropic initiative of IBM** Corporate Citizenship, the corporate social responsibility and philanthropy division of IBM. Through Corporate Citizenship, IBM donates its technology and talent to address some of the world's most pressing social and environmental issues.

http://www.worldcommunitygrid.org

Example 2

Open Science Grid

- The **Open Science Grid Consortium** is an organization that administers a worldwide grid of technological resources called the Open Science Grid, which facilitates distributed computing for scientific research. Founded in 2004, the consortium is composed of **service and resource providers, researchers from universities and national laboratories, as well as computing centers across the United States**. Members i**ndependently own and manage the resources which make up the distributed facility**, and consortium agreements provide the framework for technological and organizational integration.
- …
- The OSG facilitates **access to distributed high throughput computing for research in the US**. The resources accessible through the OSG are contributed by the community, organized by the OSG, and governed by the OSG consortium. In the last 12 months, we have provided more than 800 million CPU hours to researchers across a wide variety of projects.
- …

http://www.opensciencegrid.org

Example 3

European Grid Initiative (EGI)

- EGI is a federation of computing and storage resource providers united by a mission to support research and development. The federation is governed by the participants represented in the EGI Council and coordinated by the EGI Foundation.
- The EGI federated e-infrastructure is publicly funded and provides computing and storage resources to support research and innovation.
- https://www.egi.eu/

Egypt T3

- EG-ZC-T3 – AfricaArabia is one of EGI Tier3 data center (Zewail city)
  - UI.
  - CE
  - TORQUE server (pbs)
  - 2 Physical WN + 2 virtual WNs.
  - DPM_MYSQL and DPM_DISK on one machine.
  - SQUID-server
  - 8 TB separate storage server (nfs)
- BU : Tier3 working, not EMI register yet
- Other T3 and T2 under construction

European Middleware Initiative

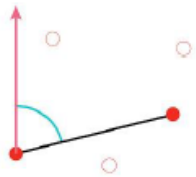
- The **European Middleware Initiative** (EMI) is a **computer software platform for high performance distributed computing**. It is developed and distributed directly by the EMI project. It is the **base** for other **grid middleware** distributions used by scientific research communities and distributed computing infrastructures all over the world especially in **Europe, South America and Asia**. EMI supports broad scientific experiments and initiatives, such as the Worldwide LHC Computing Grid (for the Large Hadron Collider).
- The EMI middleware is a cooperation among three general purpose grid platforms, the **Advanced Resource Connector, gLite and UNICORE and the dCache storage software**.
- …
- EMI **improves the existing middleware services and harmonizes them**, realizing a common framework with the result of rendering the middleware to be simpler and easier to use, reducing at the same time the interoperability problems faced by the distributed computing infrastructure providers. Thanks to EMI, it was possible to get the results in a previously unthinkable time and to contribute to the discovery of one of the most 'wanted' particles: a Higgs boson.

http://www.eu-emi.eu

Outline of part 1

1. Introducing the grid
2. Introducing WLCG
    1. LHC computing
    2. Activity
3. WLCG Architecture

Introducing WLCG.



A key tool to study physics

The most sophisticated data-taking & analysis system ever built for science, providing near real-time access to LHC data.

Global collaboration

42 countries
170 computing centres
2 million jobs run every day

WLCG
Worldwide LHC Computing Grid

The **Worldwide LHC Computing Grid** (WLCG) is a global computing infrastructure whose mission is to **provide computing resources to store, distribute and analyze the data generated by the Large Hadron Collider** (LHC), making the data equally available to all partners, regardless of their physical location.

WLCG is the world's largest computing grid. It is **supported by many associated national and international grids across the world,** such as European Grid Initiative (Europe-based) and Open Science Grid (US-based), as well as many other regional grids.

WLCG is co-ordinated by CERN. It is **managed and operated by a worldwide collaboration** between the experiments (ALICE, ATLAS, CMS and LHCb) and the participating computer centers. It is reviewed by a board of delegates from partner country funding agencies, and scientifically reviewed by the LHC Experiments Committee.

http://wlcg.web.cern.ch/

LHC (CMS) computing.

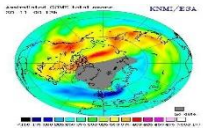
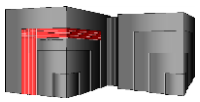
WLCG
Worldwide LHC Computing Grid

- Data pours out of the LHC detectors at a blistering rate. Even after filtering out 99% of it, there is still around **30 petabytes of data per year to deal with**. That's 30 million gigabytes, the equivalent to nearly 9 million high-definition (HD) movies.
- The scale and complexity of data from the LHC is unprecedented. This data **needs to be stored, easily retrieved and analyzed by physicists all over the world**. This requires massive storage facilities, global networking, immense computing power, and, of course, funding.
- **CERN does not have the computing or financial resources to crunch all of the data** on site, so in 2002 it turned to grid computing to **share the burden with computer centres around the world.**
- The result, the **Worldwide LHC Computing Grid** (WLCG), is a distributed computing infrastructure arranged in tiers – giving a community of over 10,000 physicists near real-time access to LHC data. The WLCG builds on the ideas of grid technology initially proposed in 1999 by Ian Foster and Carl Kesselman (link is external).

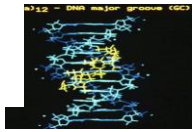
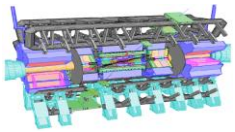
http://wlcg-public.web.cern.ch/about

The characteristics of WLCG

- Hardware
  - Off the shelf
  - Inexpensive
  - Many hardware varieties and suppliers
- Available ressources vary in time
- Grid is open to other communities




A few numbers

- the world's largest computing grid
- 553 611 cores
- 2 000 000 jobs/day
- 300 Petabytes of online disk storage and 230PB of nearline (magnetic tape) storage.
- ~50 Petabytes of data expected in 2016
- funding is ~$100M/year

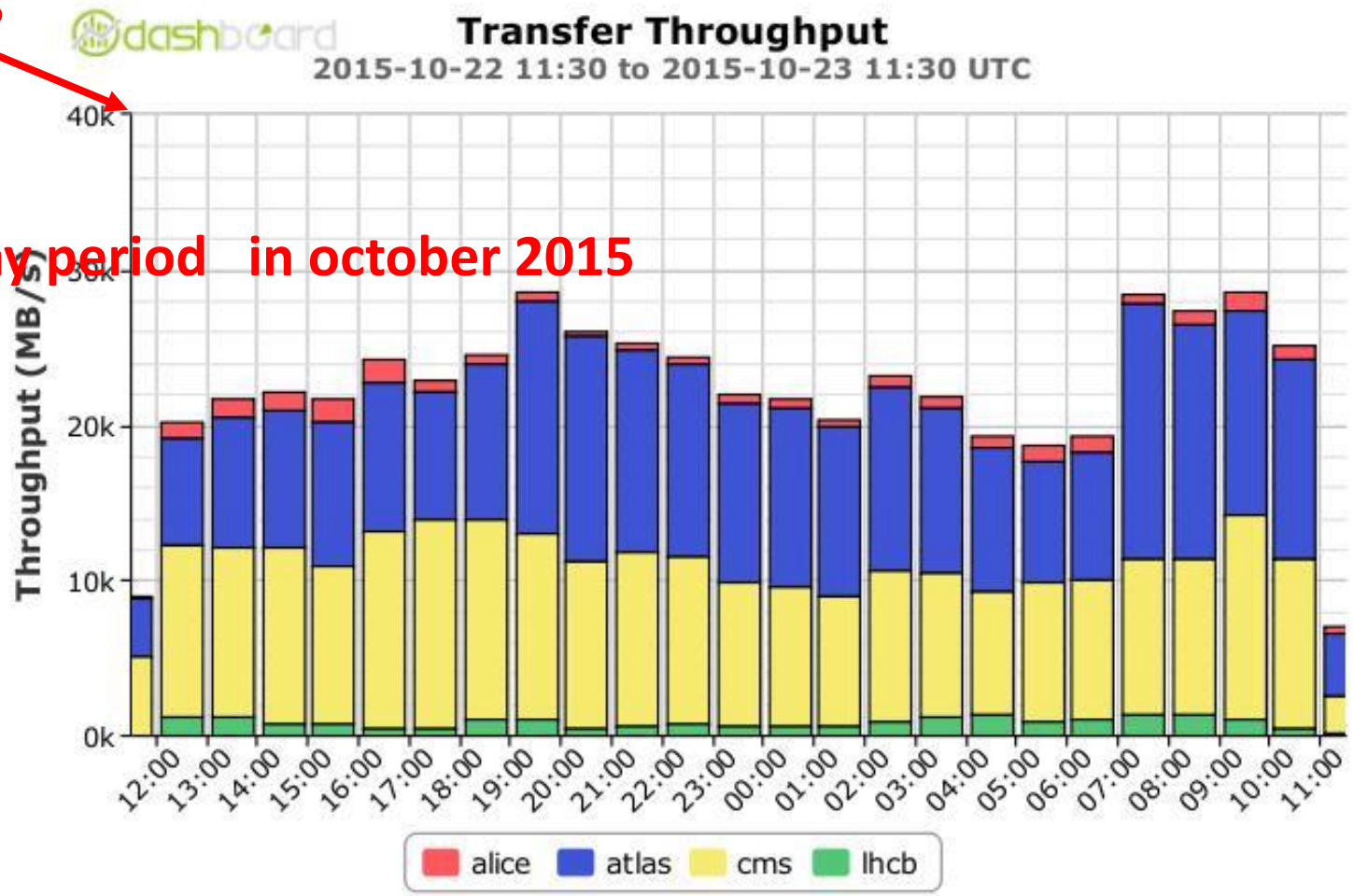
Outline of part 1

1. Introducing the grid
2. Introducing WLCG
    1. LHC computing
    2. Activity
3. WLCG Architecture

WLCG activity 2015

40000 MB/S

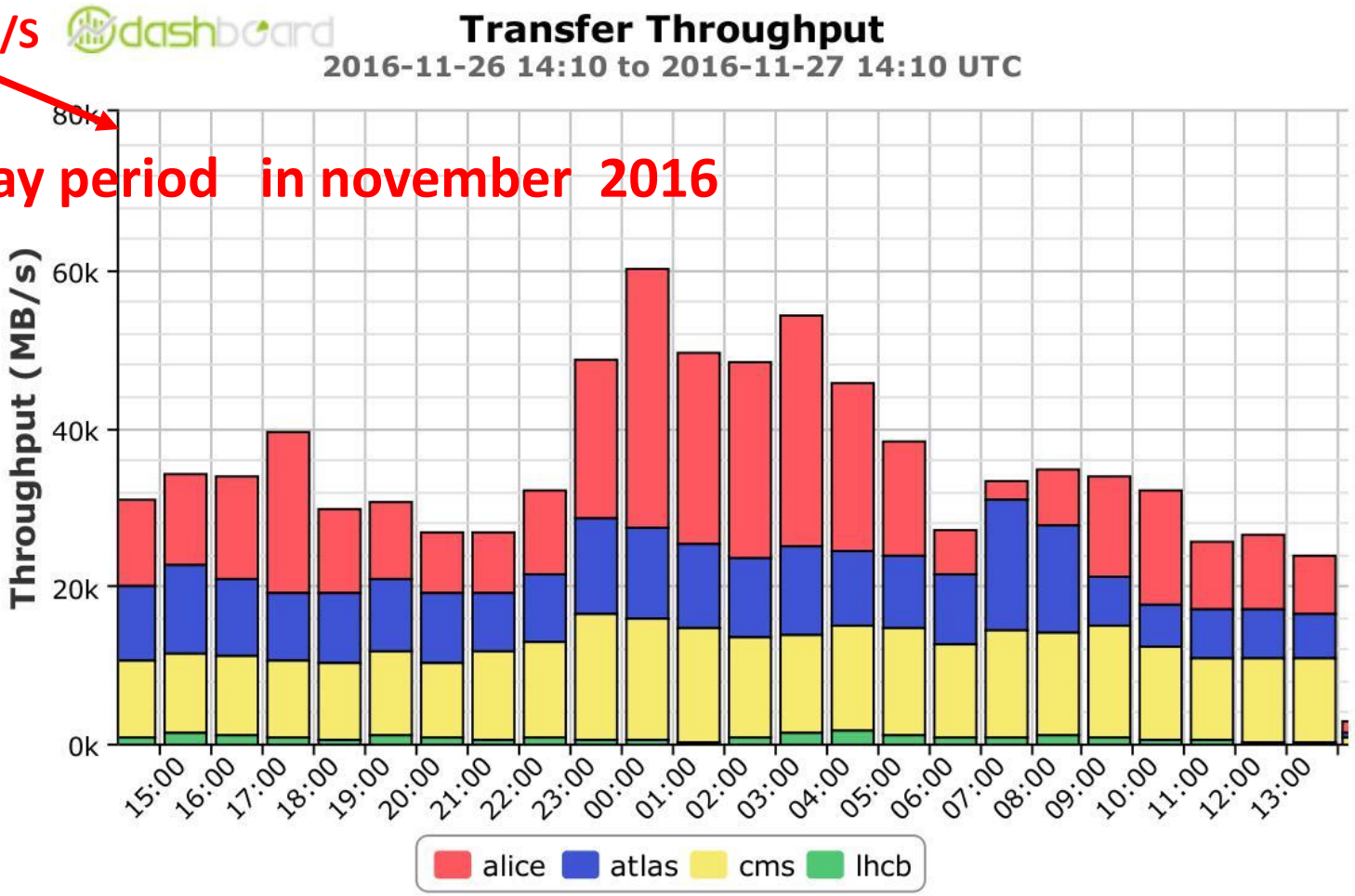
One day period in october 2015



WLCG activity 2016.

80000 MB/S

One day period in november 2016



CMS activity in oct 2015

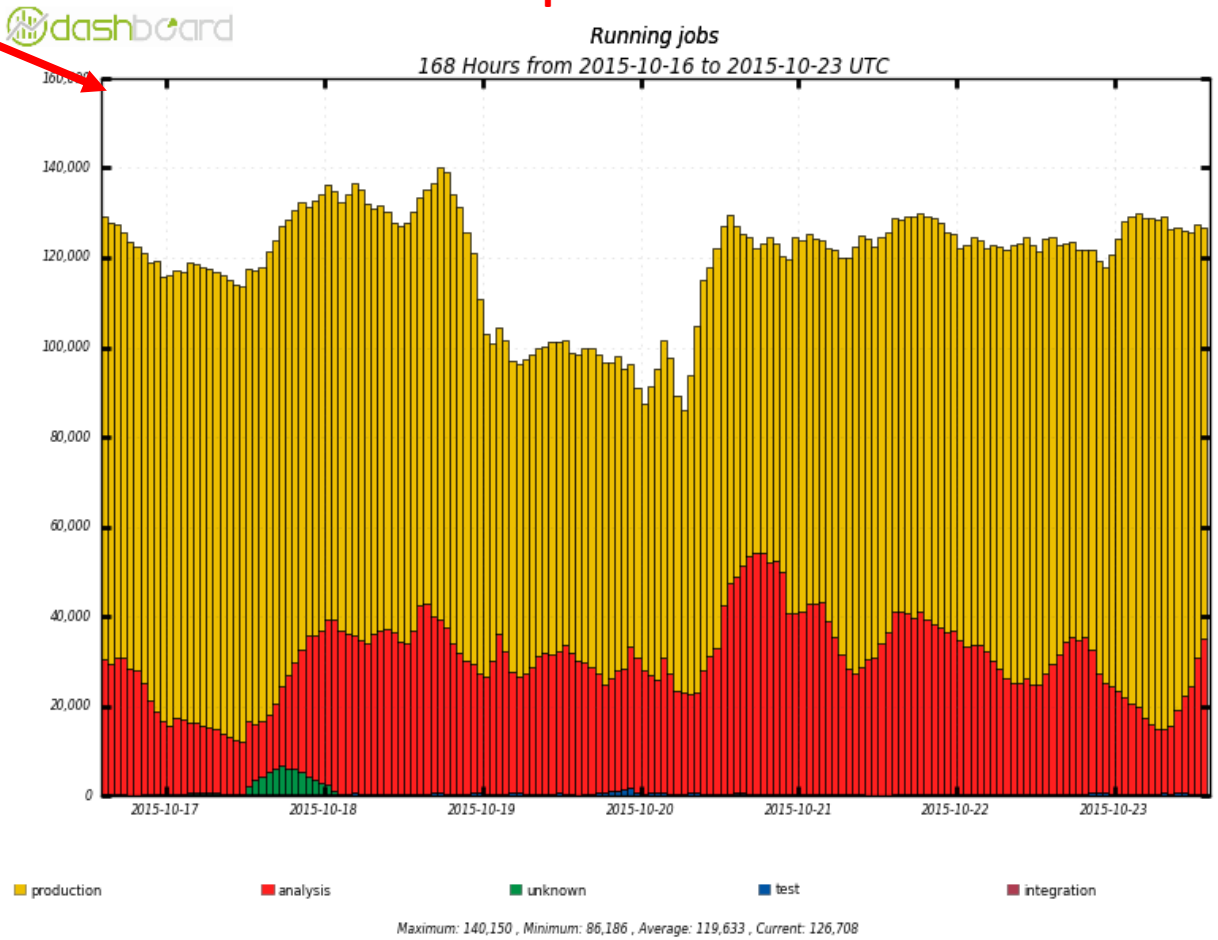
160 000 running jobs

One week period in october 2015

production

analysis

unknown



CMS activity in nov 2016

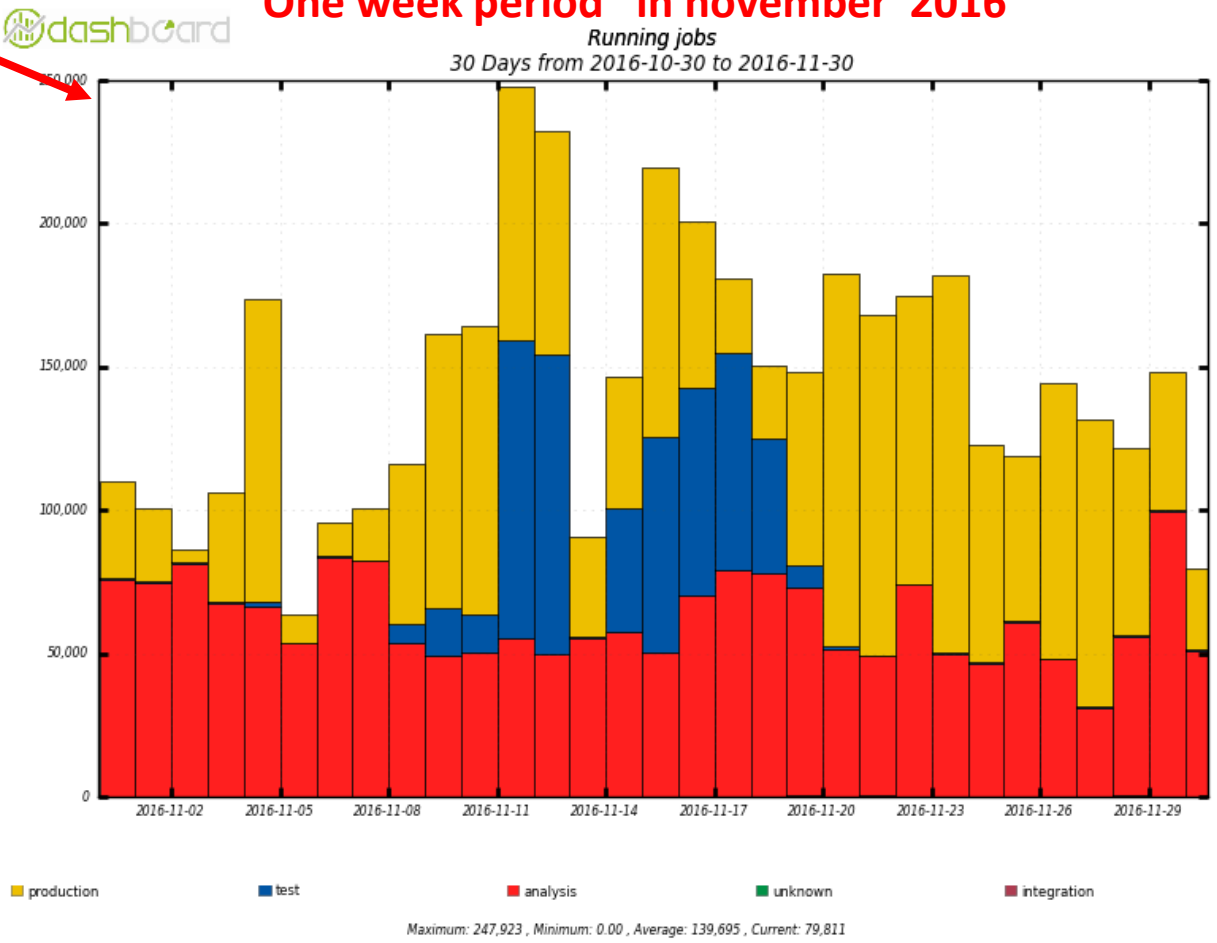
250 000 running jobs

One week period in november 2016

production

analysis

test



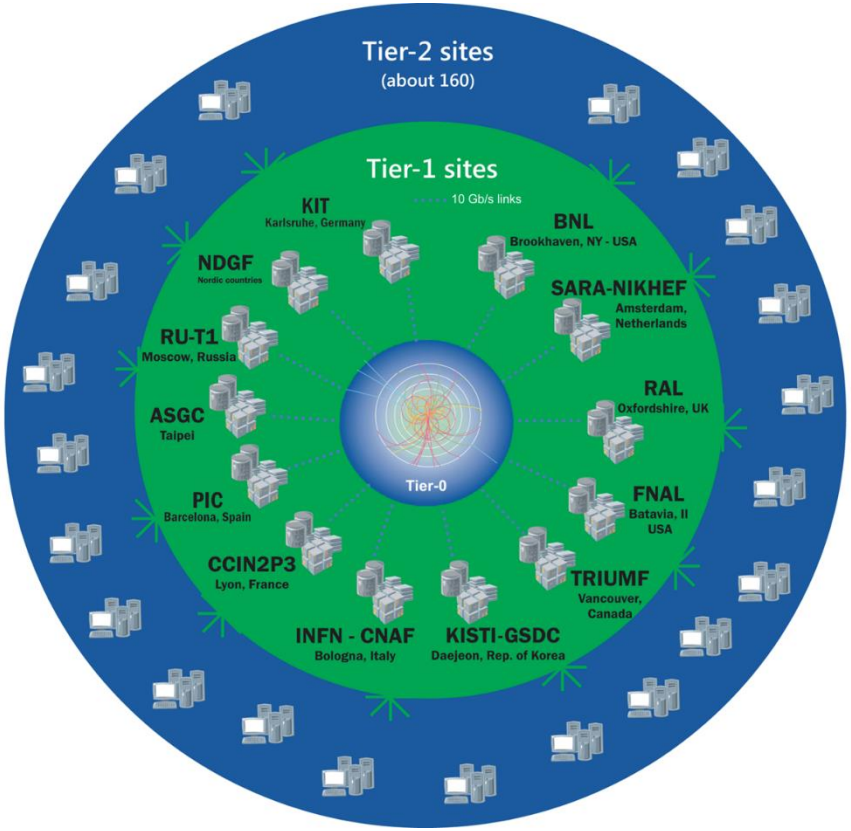
Outline of part 1

1. Introducing the grid
2. Introducing WLCG
3. WLCG Architecture
    1. Tiers
    2. Information system
    3. Authentication
    4. Data management
    5. Jobs workflow
    6. VO's applications

WLCG sites - Tier-0

Tier-0

This is the **CERN Data Centre**, which is located in Geneva, Switzerland and also at the **Wigner Research Centre for Physics** in **Budapest**, Hungary over 1200km away. The two sites are connected by two dedicated 100 Gbit/s data links. All data from the LHC passes through the central CERN hub, but CERN provides less than 20% of the total compute capacity.
Tier 0 is responsible for the **safe-keeping of the raw data** (first copy), **first pass reconstruction**, **distribution** of raw data and reconstruction output to the Tier 1s, and reprocessing of data during LHC down-times.

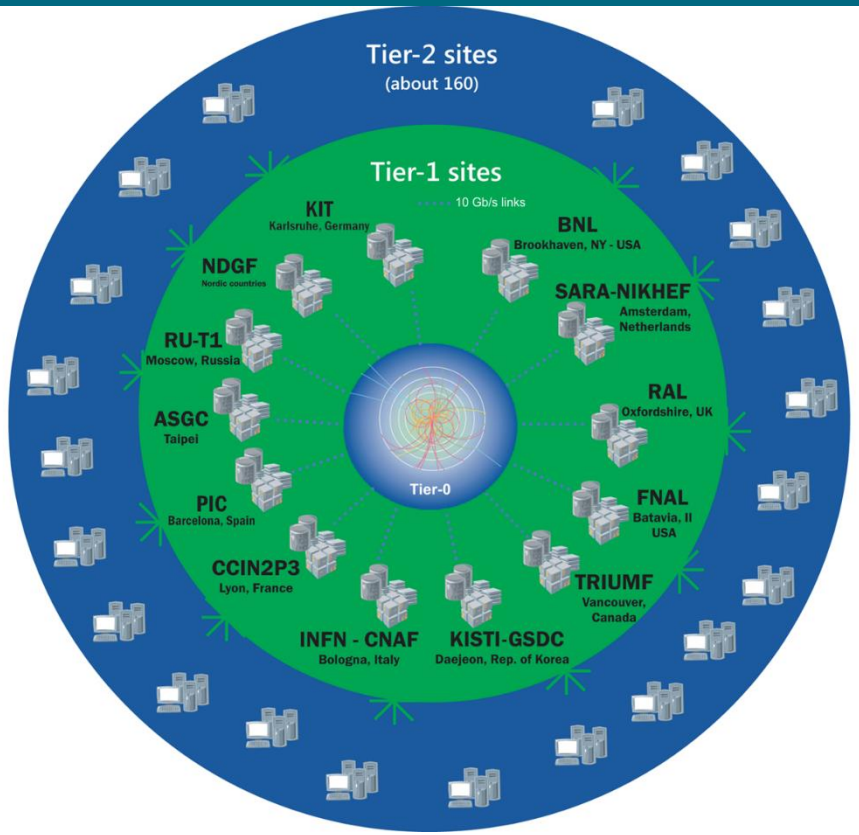


http://wlcg-public.web.cern.ch/tier-centres

WLCG sites - Tier-1's

Tier-1's

These are **thirteen large computer centres** with sufficient **storage capacity** and with **round-the-clock support** for the Grid. They are responsible for the **safe-keeping** of a proportional **share of raw and reconstructed data, large-scale reprocessing** and safe-keeping of corresponding output, **distribution of data to Tier 2s** and **safe-keeping** of a share of **simulated data** produced at these Tier 2s.



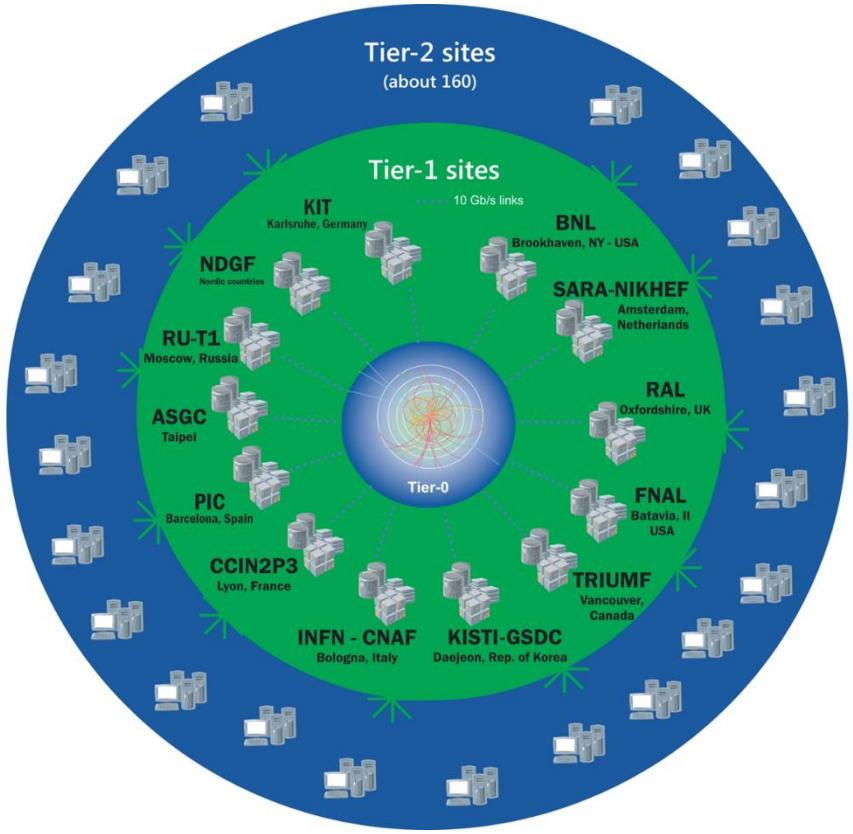
http://wlcg-public.web.cern.ch/tier-centres

WLCG sites - Tier-2's

Tier-2's

The Tier 2s are typically universities and other scientific institutes, which can **store** sufficient **data** and **provide** adequate **computing power** for **specific analysis tasks**. They handle analysis requirements and proportional share of simulated event production and reconstruction.

There are currently around 160 Tier 2 sites covering most of the globe.

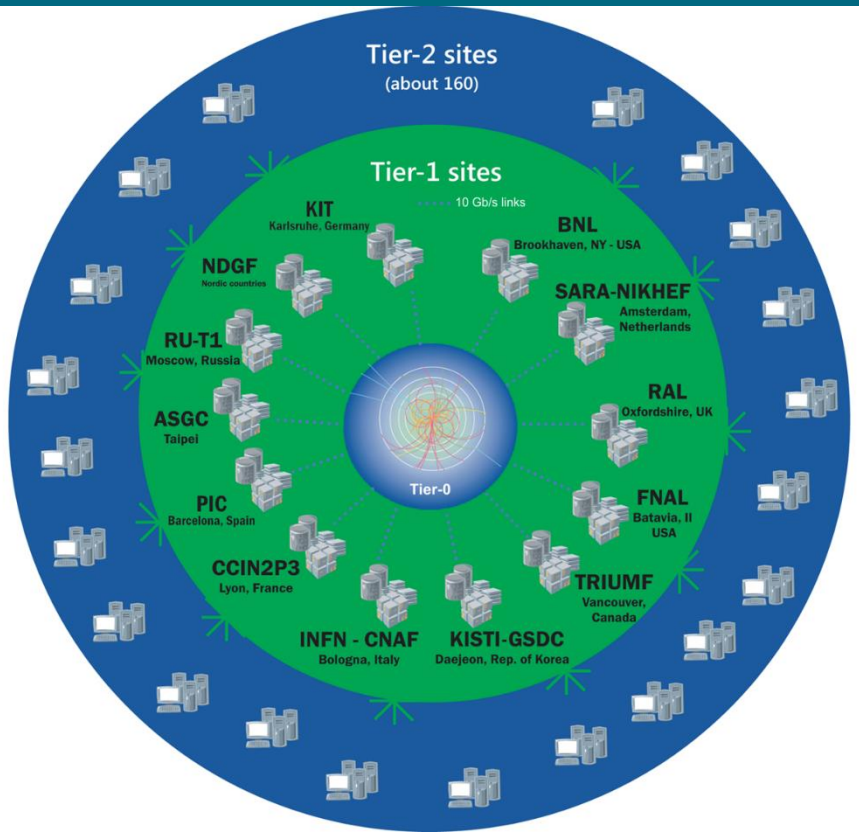


http://wlcg-public.web.cern.ch/tier-centres

WLCG sites – Tier-3's

Tier-3's

Individual scientists will access these facilities through **local** (also sometimes referred to as Tier 3) **computing resources**, which can consist of local clusters in a University Department or even just an individual PC. There is **no formal engagement** between WLCG and Tier 3 resources.



http://wlcg-public.web.cern.ch/tier-centres

Outline of part 1

1. Introducing the grid
2. Introducing WLCG
3. WLCG Architecture
    1. Tiers
    2. Information system
    3. Authentication
    4. Data management
    5. Jobs workflow
    6. VO's applications

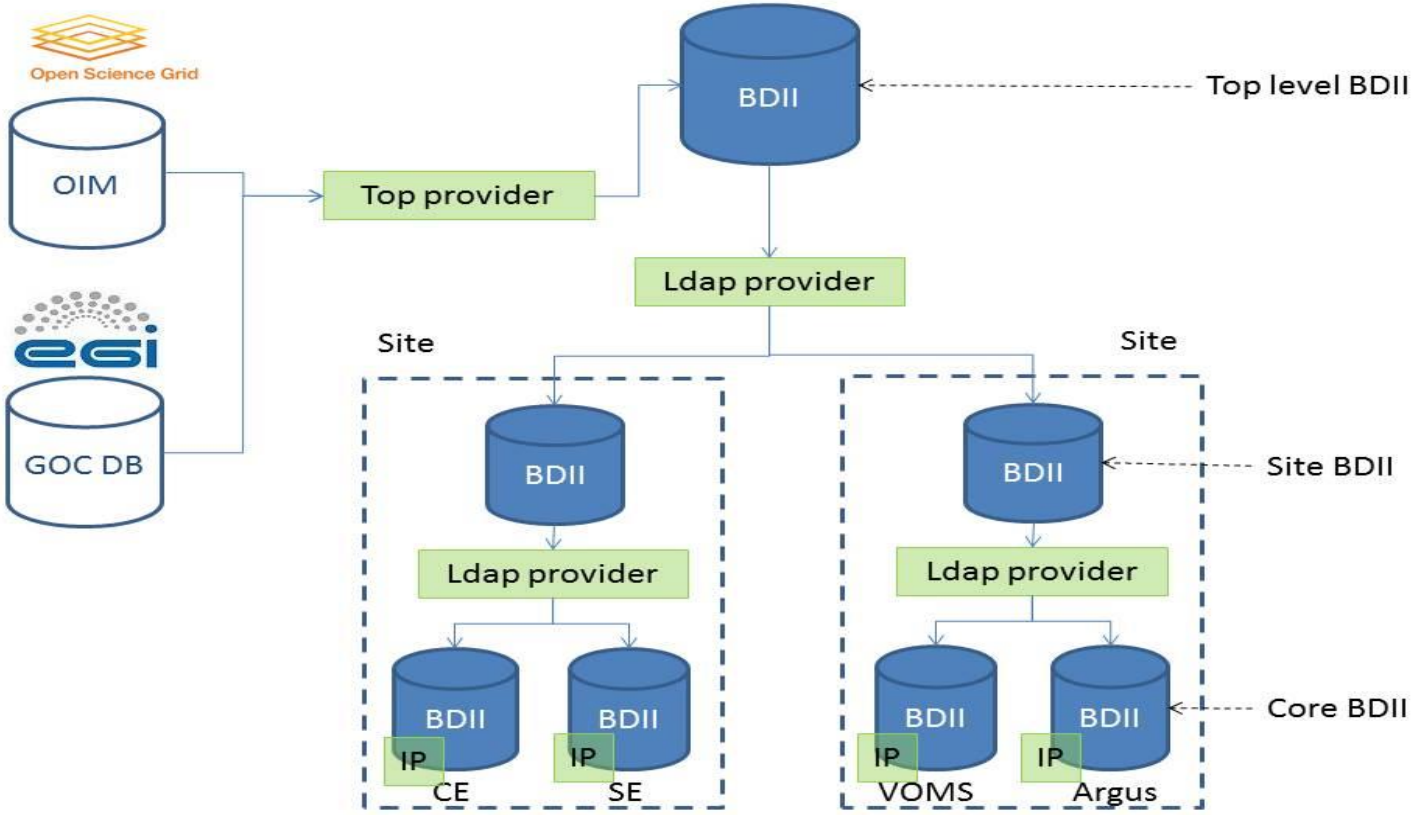
Information system.

Service	Provided by	Content	Protocol/format
REBUS	WLCG	List of WLCG federations and sites Repository for pledge information (federation level) Installed capacities (from BDII)	HTTPS, JSON+XML+CSV
GOCDB	STFC, EGI	Administrative info for EGI sites List of all site services Downtime information	HTTPS, XML
OIM	OSG	As above but for OSG	HTTPS, XML
BDII	WLCG, EGI	Detailed information for sites and services	LDAP, LDIF
Info provider	Middleware developers	Generate the information in the BDII	LDIF

Collective

Resource

Information system.



Outline of part 1

1. Introducing the grid
2. Introducing WLCG
3. WLCG Architecture
    1. Tiers
    2. Information system
    3. Authentication
    4. Data management
    5. Jobs workflow
    6. VO's applications

Certificates

- A user certificate is a text file with encrypted data that you install on your PC so that you can secure/encrypt sensitive communications between your site and the grid.
- grid credentials:
  - digital certificate and private key
  - Based on Public Key Infrastructure (PKI). X.509 standard
  - Certification Authority (CA) signs certificates. Trust relationship

Proxy

- Each grid service requires a valid certificate.
- The user certificate, whose private key is protected by a password, cannot be used to access these services;

  -> user has to create a Proxy

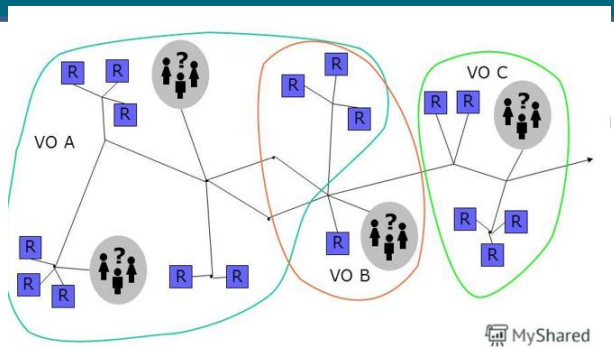
Proxy :

temporary certificate without password.

Proxy = proof of identity -> short lifetime (typically 12 hours) to reduce security risks should it be stolen

VO
(Just in case you have forgotten ;=)

Virtual Organizations (VO's)

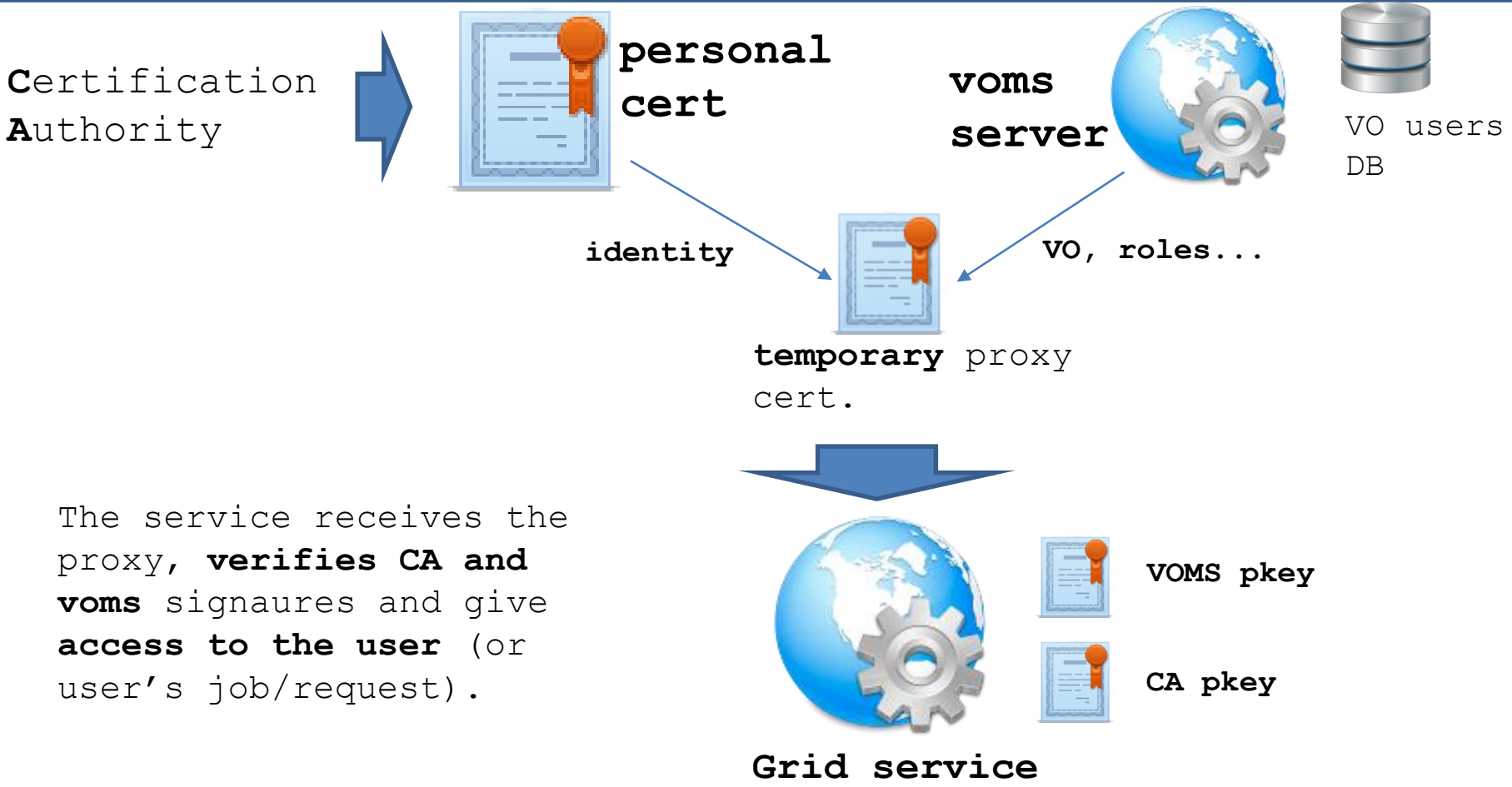


- a **group** of mutually distrustful participants with **varying degrees of prior relationship** (perhaps none at all) that **want to share resources** in order to perform some tasks.
- For instance there is a VO called CMS, but a VO can have any name

Virtual Organization Membership Service (VOMS)

- EGEE/gLite enhancement for VO management
- Provides information on user's relationship with Virtual Organization
  - Membership
  - Group membership
  - Roles of user
- Multiple VO
  - User can register to multiple VOs and create an aggregate proxy
  - Access ressources in every registered VO
- Backward compatibility
  - Extra VO related information in users proxy certificate
  - Users proxy can still be used with non VOMS-aware services

Authentication.



The service receives the proxy, **verifies CA and voms** signaures and give **access to the user** (or user's job/request).

Outline of part 1

1. Introducing the grid
2. Introducing WLCG
3. WLCG Architecture
    1. Tiers
    2. Information system
    3. Authentication
    4. Data management
    5. Jobs workflow
    6. VO's applications

Data

- User and programs produce and require data
  - data can be send from/to jobs, but
    - Input/Output Sandboxes are limited to 10 MB
    - Data has to be copied from/to local filesystems to the Grid
- Solution
  - Storing data in Grid datasets
    - Located in Storage Elements (SE)
    - Several replicas of one file in different sites
    - Accessible by Grid users and applications from “everywhere”
    - data requirements in the job description (jdl)

Data ressource

- Each VO manages its own files
- via a file catalog
- The File Catalog (or Replica Catalog) maps between the Logical File Name and the Storage URL (or URLs) for files managed by the Grid
- lcg-util for operations on the files

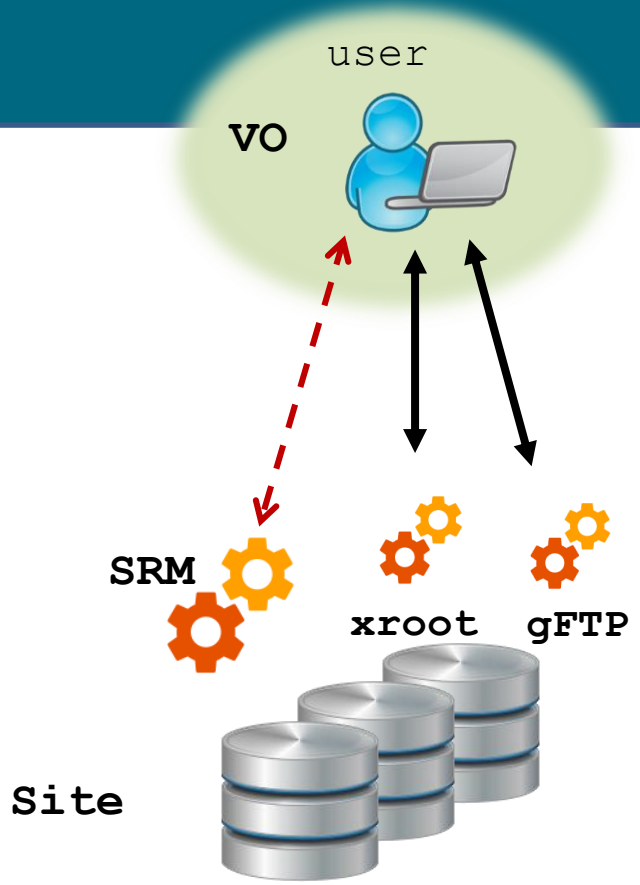
Files names:

- the GridUniqueIDentifier(GUID) identifies a file uniquely
- the LogicalFileName(LFN) is a human-readable name for a file
- the Storage URL (SURL), also known as Physical File Name (PFN), identifies a replica in a SE (storage element)
- the Transport URL (TURL), is a valid URI with the necessary information to access a file in a SE

Files catalog (LFC):

- Maintains mappings between LFN(s), GUID and SURL(s)
- a Grid file must be both physically present in a SE and registered in a file catalogue
- LFC attaches to a file or directory an access control list (ACL), a list of permissions which specify who is allowed to access or modify it
  - LFN : LogicalFileName
  - GUID : GridUniqueIDentifier
  - SURL: Storage URL = PFN : Physical File

Storage Resource Manager (SRM)



- **SRM based storage:**
  - SRM the middleware over the different storage technologies;
  - serves metadata and redirects data request to other services: gFTP, (gsi)http, xrootd;
  - (gsi)http, xrootd and gridFTP are middleware services over data serving;

info/requests

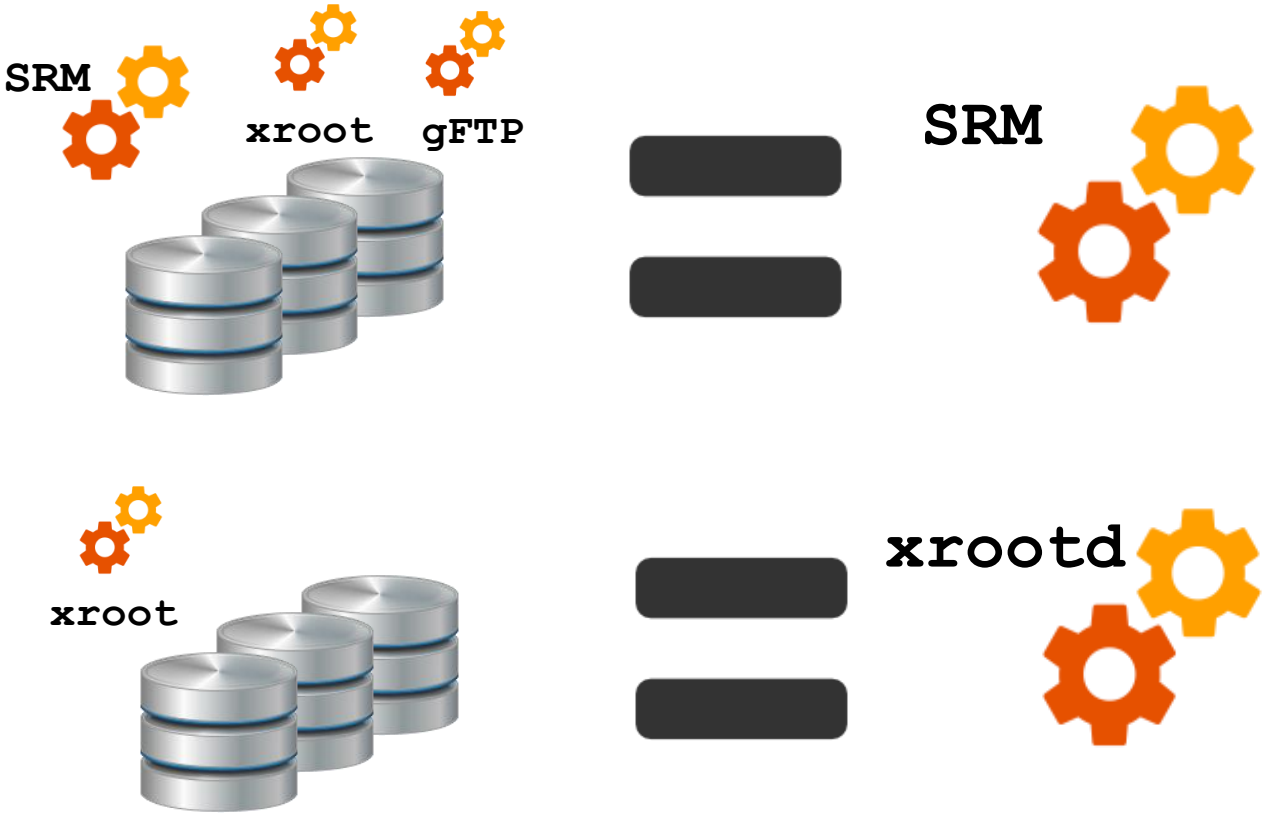
data

(Resource Layer)

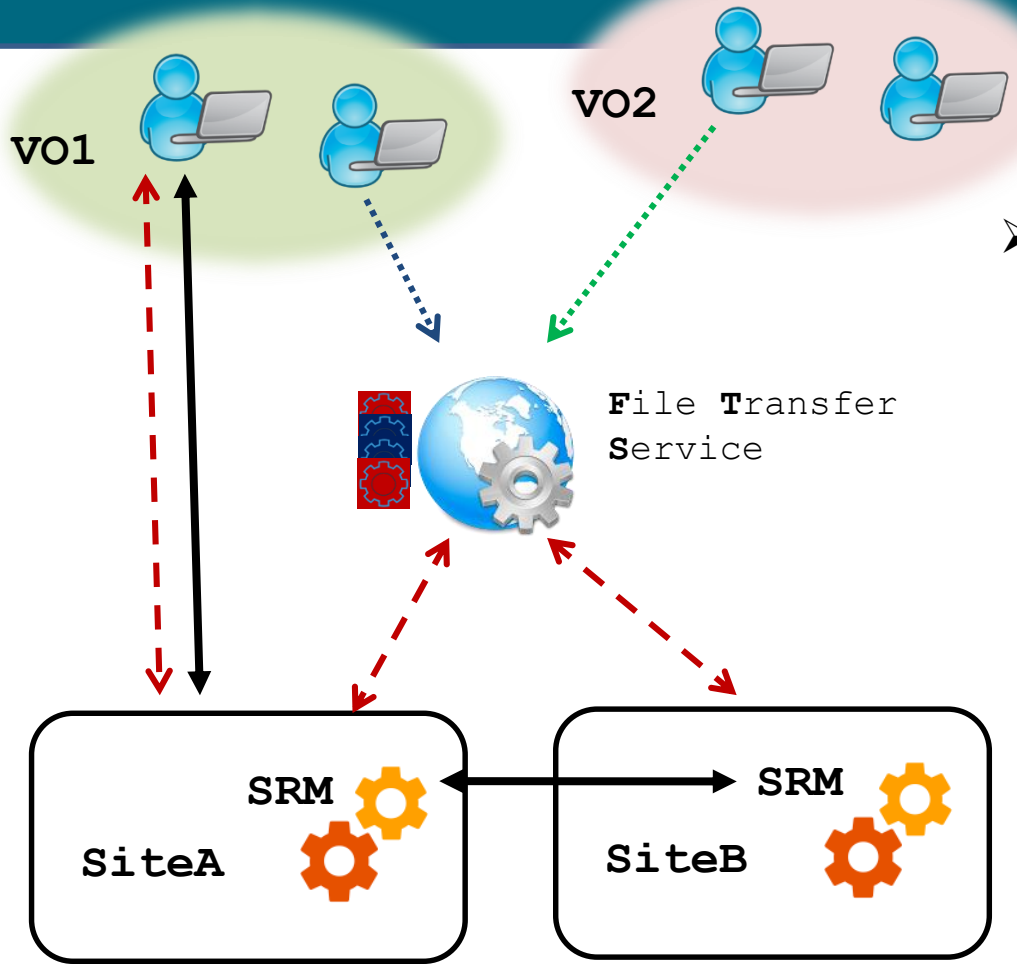
GFAL

Data management.

To ease notation, from now on…



File transfert workflow.



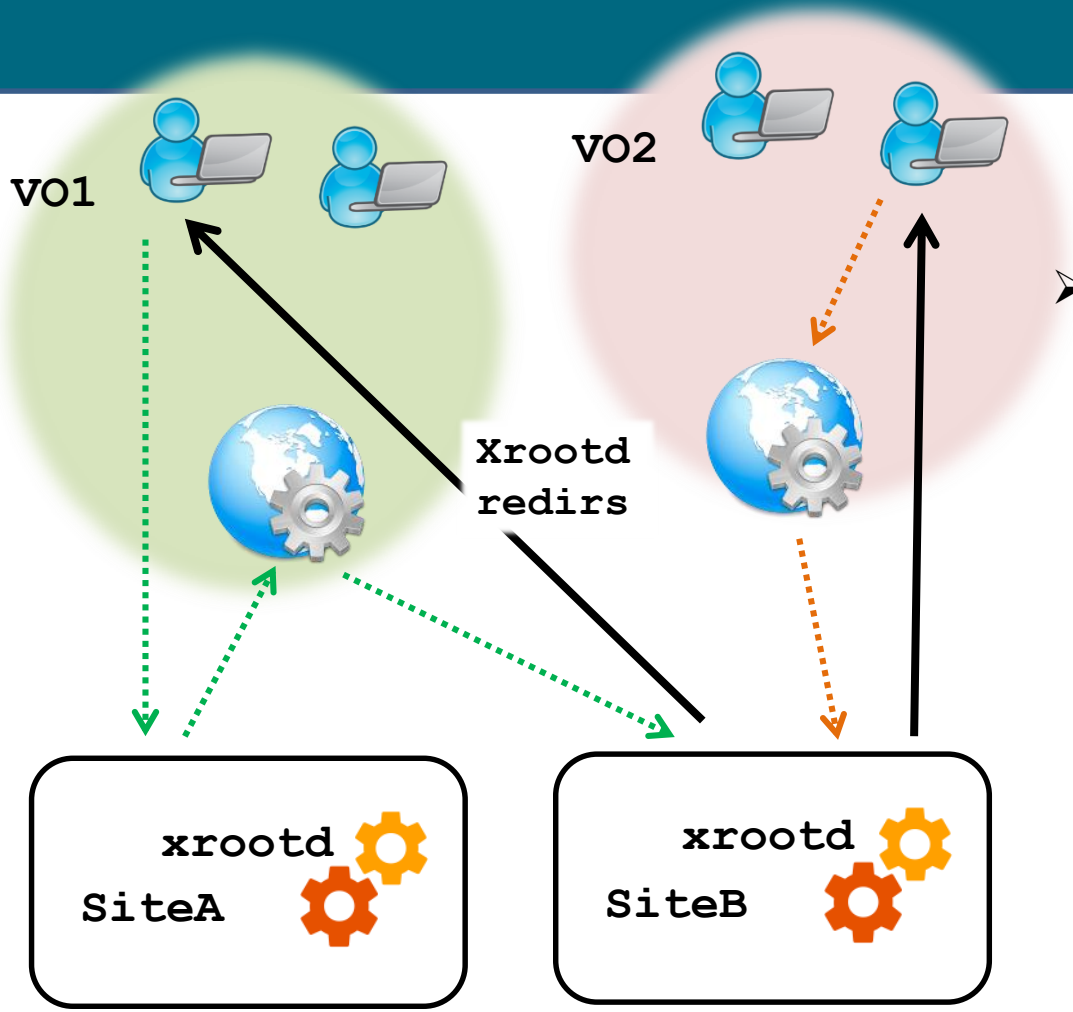
- **File Transfer Service:**
  - manages **site to site** transfers at **file level**;
  - the **user submits** a file **request to the FTS queue** and FTS deals with the SRMs to perform it;

info/requests

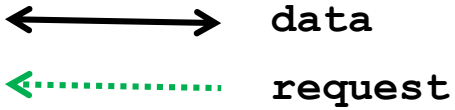
data

transf. req.

File transfert workflow.



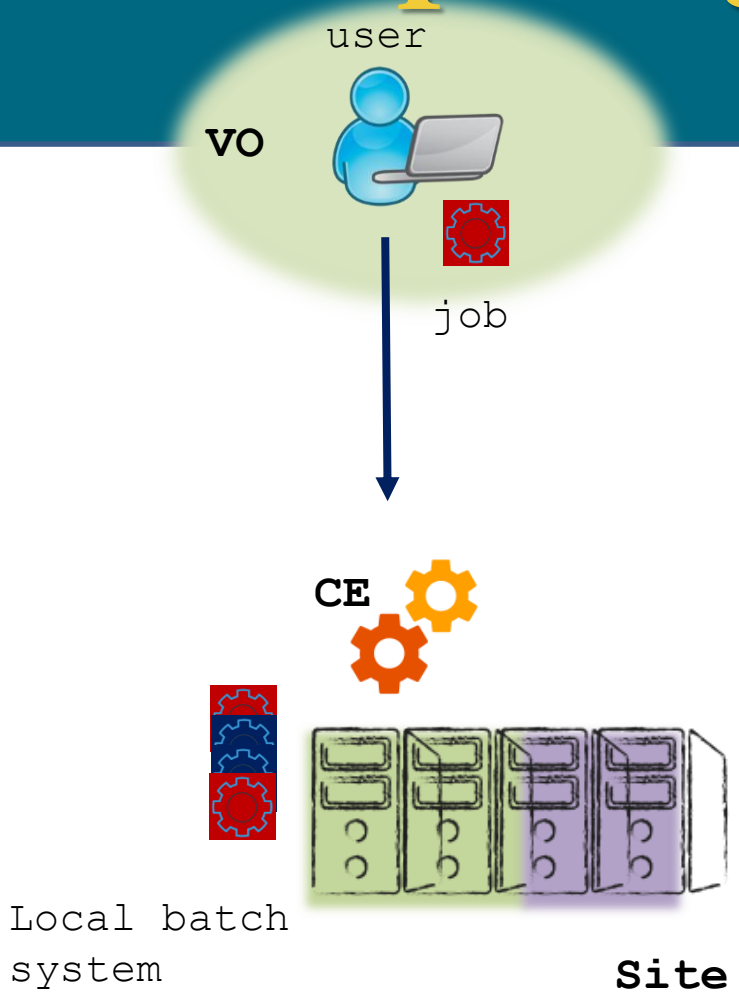
- **Xrootd federations**
  - global (read) **access** to files via a **hierarchy of redirectors;**
  - **integrated** in **root** and in the **exp's software;**
  - soon http and gsiftp federations as well.



Outline of part 1

1. Introducing the grid
2. Introducing WLCG
3. WLCG Architecture
    1. Tiers
    2. Information system
    3. Authentication
    4. Data management
    5. Jobs workflow
    6. VO's applications

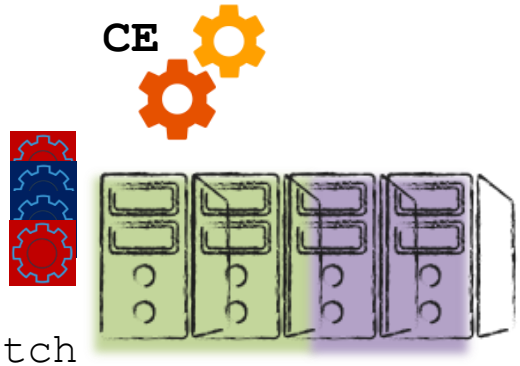
Computing Element (CE)



- **Computing Element (CE):**
  - **middleware** layer on top of a local batch system;
  - standard protocol for **different batch technologies**;
  - deals with **authentication** and **authorization**, map the grid user to local group;
  - **local** batch system implements **shares and priorities**.

To ease notation:

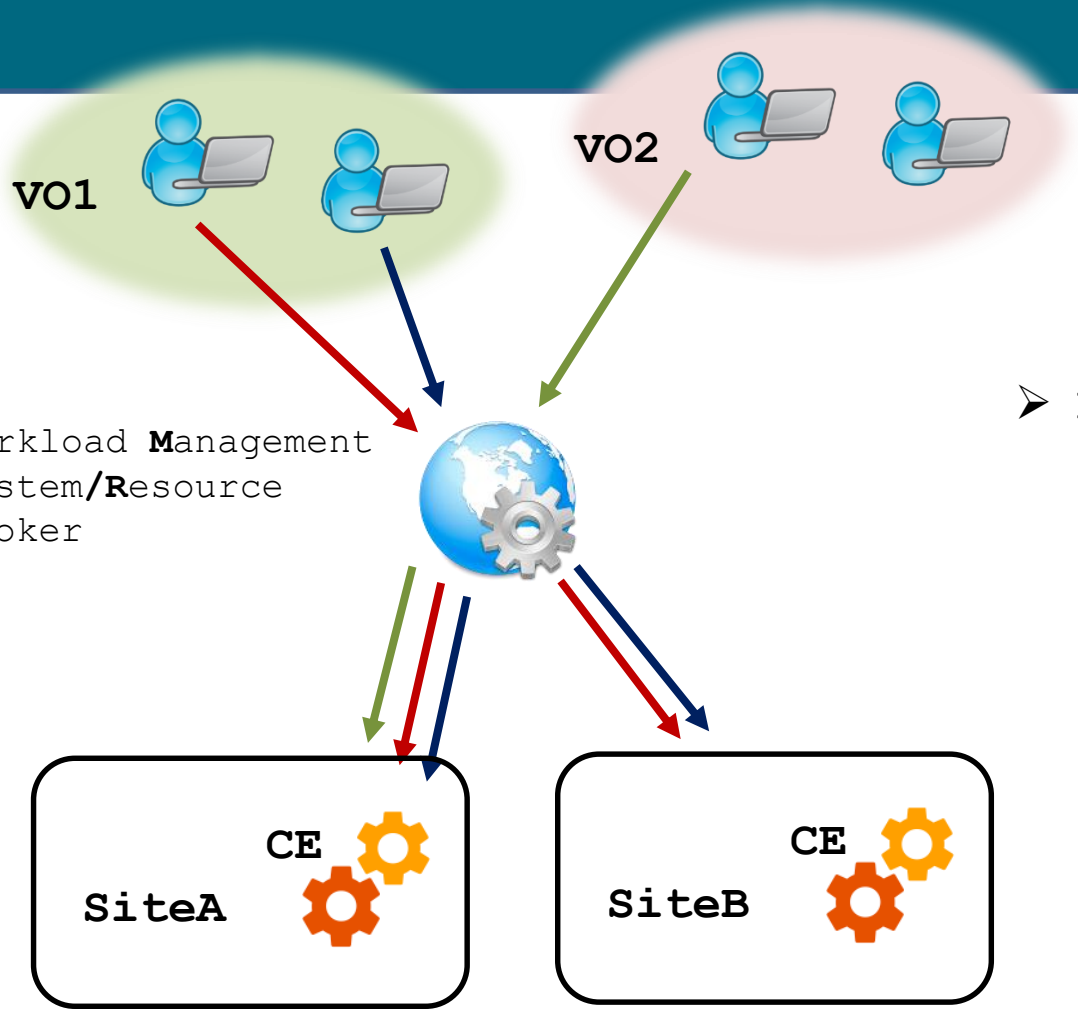
To ease notation, from now on…



Local batch system

=

Grid jobs workflow (Push mode)



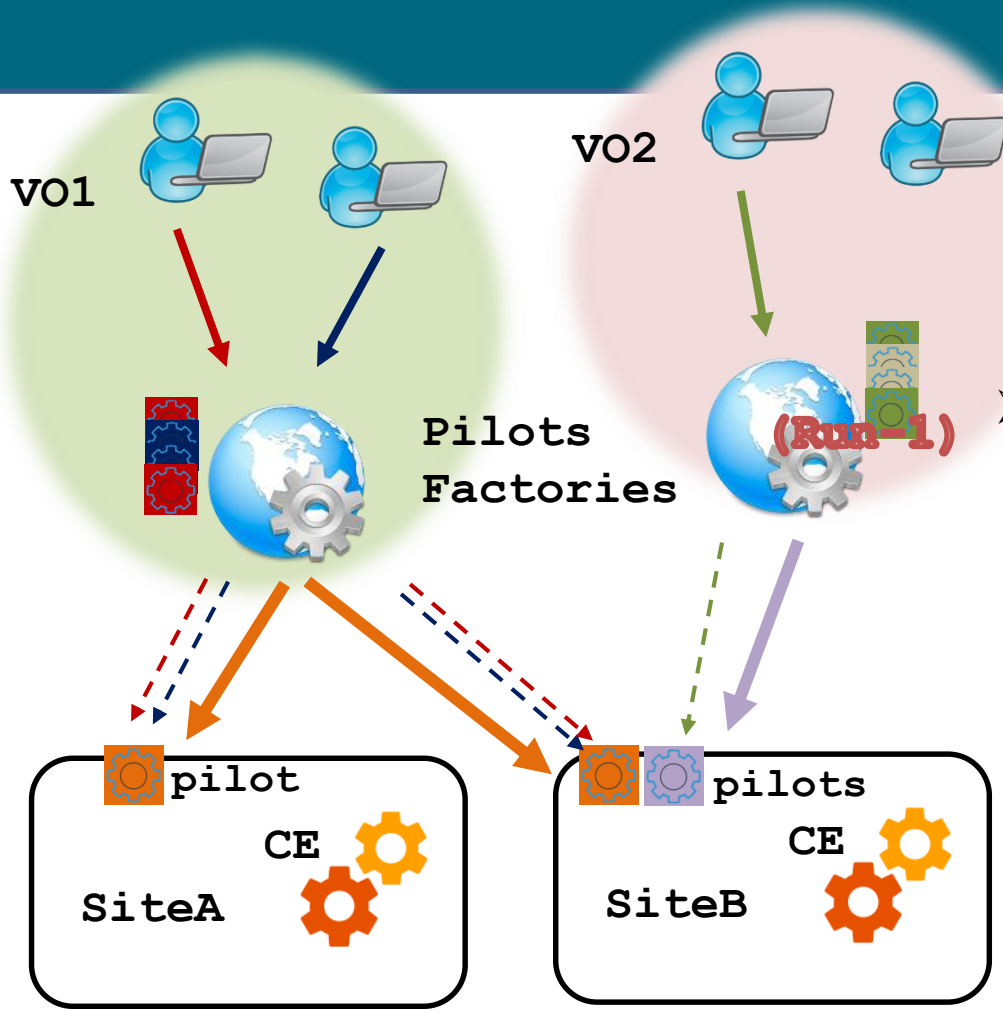
(Run-1)

- **Push mode:**
  - **WMS** does **match-making;**
  - **local sites scheduling** does **fine grained** (intra-VO) policies: not optimal for manage complex policies;
  - users **jobs run directly** on the nodes: non optimal slots usage time.

Workload management System (WMS)

- If the jobs do not specifie where it must be run, it is sent to a WMS (Workload Management System)
- The WMS selects the site where the job will be run.
  - The site has to match the job requirements
  - jobs must be executed as fast as possible
  - jobs must be evenly and efficiently distributed across the entire Grid.
- The WMS interacts with the Data Management service and the Information System
  - They supply all the information required for jobs dispatching

Grid jobs workflow (Pull mode)



(Run-2)

- **Pull mode:**
  - **pilot factory** does **match-making** and **fine-grained** (intra-VO) **scheduling;**
  - submits pilot agents to the site. Local scheduling for coarse grained policies;
  - **pilots pull users jobs** from the pilot factory: usage time optimized.

Outline of part 1

1. Introducing the grid
2. Introducing WLCG
3. WLCG Architecture
    1. Tiers
    2. Information system
    3. Authentication
    4. Data management
    5. Jobs workflow
    6. VO's applications

VO (CMS) applications.

SiteDB: CMS specific informations about sites, people and resources.

DBS, TMDB and DAS: informations about data: location of replicas, organisation in datasets/fileblocks, physics related info (Run, lumi, etc.).

PhEDEx: site to site data transfers at dataset/fileblock level, tracking of sites occupation.

CRAB: user tool for performing analysis on grid. Connects grid submission to CMS specific data management (DBS/TMDB) and site's info (SiteDB).

Questions?

Files names:

- the GridUniqueIDentifier(GUID) identifies a file uniquely, is of the form:

  guid:38ed3f60-c402-11d7-a6b0-f53ee5a37e1d

- the LogicalFileName(LFN) is a human-readable name for a file, has this format: lfn:/grid/<vo>/<directory>/<file>
- the Storage URL (SURL), also known as Physical File Name (PFN), which identifies a replica in a SE, is of the general form:

  srm://srm.cern.ch:8443/srm/managerv2?SFN=castor/cern.ch/dteam/file

- the Transport URL (TURL), which is a valid URI with the necessary information to access a file in a SE, has the following form:

  <protocol>://<some_string>

  gsiftp://tbed0101.cern.ch/data/dteam/doe/file1

  root://$STAGE_HOST//castor/cern.ch/compass/data/2006/oracle_