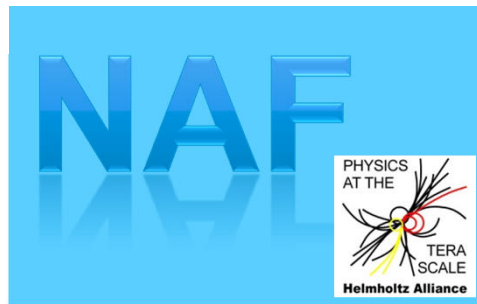


Large Public Computing Infrastructures at DESY



Thomas Finnern (DESY/IT Systems) 
Yves Kemp (DESY/IT Systems)
Andreas Gellrich (DESY/IT Scientific Computing) 

39th Linux User Meeting
April 13th, 2012 @ DESY

Overview / Outline / Outlook

> IT Service for You

- XaaS
- Overview

> 3 Systems for different Purposes

- BIRD: Local Resources: Large Batch and Special Purpose
- NAF: National Analysis Facility: Batch and Interactive
- GRID: VOs Worldwide: Batch

> Activities 2012

- BIRD: SGE to SoGE, GPUs
- NAF: New Hardware and closer Integration
- GRID: New Scheduler, Test SoGE with CreamCE
- Cloud Integration Tests



Comparison

	BIRD	NAF	GRID
Name	Batch Infrastructure Resource DESY	National Analysis Facility	Global Resource Information Database
Community	Local Groups	German LHC	Global VOs
Cores	900	3000	5000
Login	Local Registry	Grid Cert	Grid Cert
Batch System	Sun Grid Engine	Sun Grid Engine	Torque/Maui?
Min. Mem./Core	2GByte/Core	2GByte/Core	2GByte/Core
Max. Job Length	2 weeks	1 week	3 days
Max. Mem Job	32 GByte		2 GByte?
Fairshare	X	X	X
Shared Filesystems	AFS, NFS, FH?	AFS, Lustre	None



BIRD:

Batch Infrastructure Resource at DESY

A Grid Engine System

BIRD





> Integration of DESY Wide Batch Resources

- Support and Know-how
- IT and Project Hardware
- Fairshare for the *Rich* and for the *Poor*
- Complements GRID and NAF

> Fairshare

- IT Hardware + Project Specific Hardware = Valuable Resources
- Contributing Projects will be granted Fairshare Points
- Guaranteed Access to Project Resources
- IT gives own Share to the Community
- Batch Resource Requirements are not continuous over Time

> Usage Policy

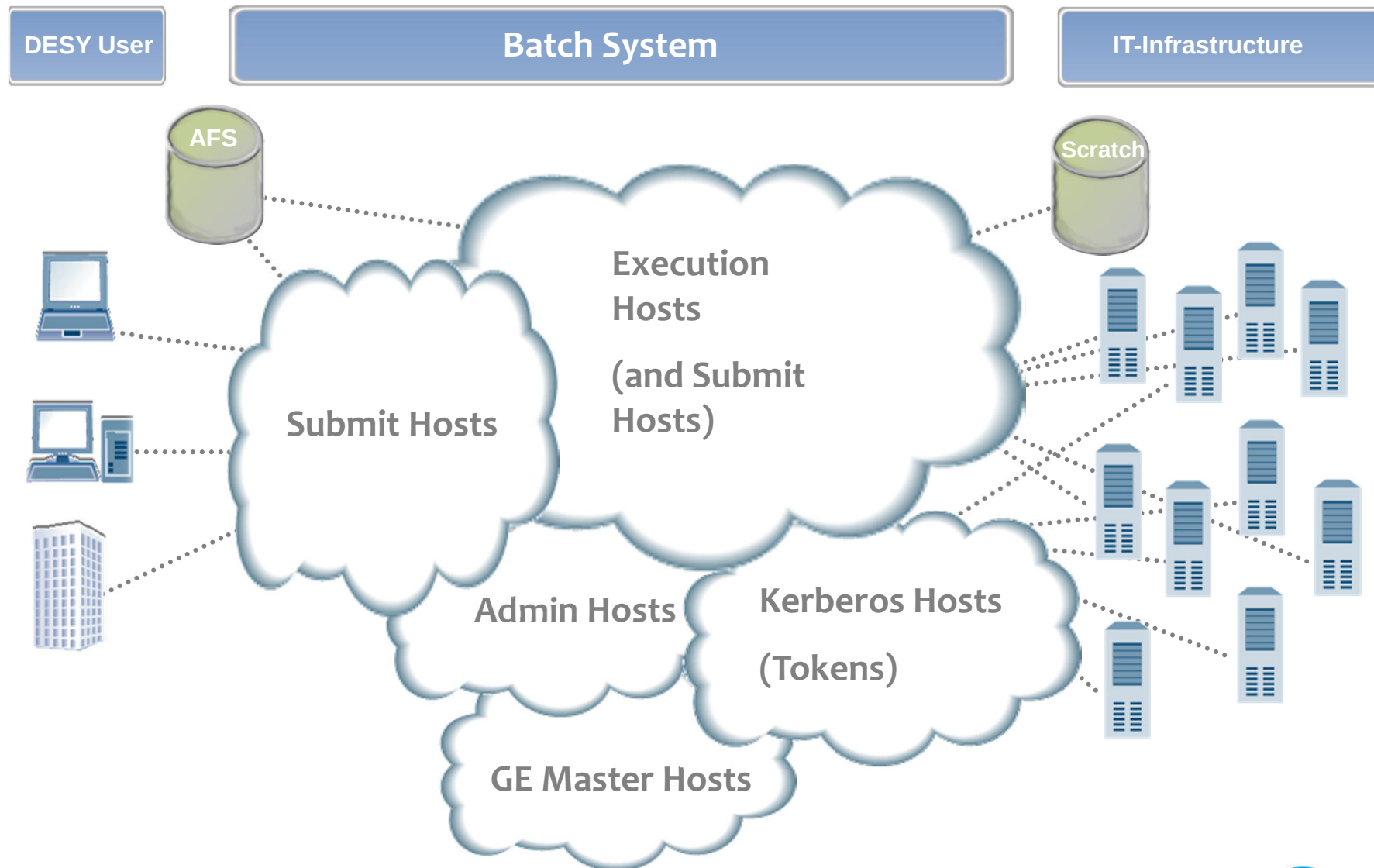
- *Lightweight* Resources should be available within a Workday
- Every Project should be capable of using its dedicated Resource Share within Week Times
- No User/Project can use the complete System on it's own
- To ensure this we use Quota and Fairshare Settings to keep the Batch Cluster in a State where all Resources can be shared in a Fair Manner

> Win-Win Situation for All

- For Those without Share Points who are allowed to use the Idle Times and/or Idle Resources (The Poor)
- Unused Project Shares even enhance Job Priorities for the Future Weeks, so typically a Project may use more Resources than it could do in a Stand-Alone Facility of it's Own (The Rich)



The **BIRD** Topology





> Base

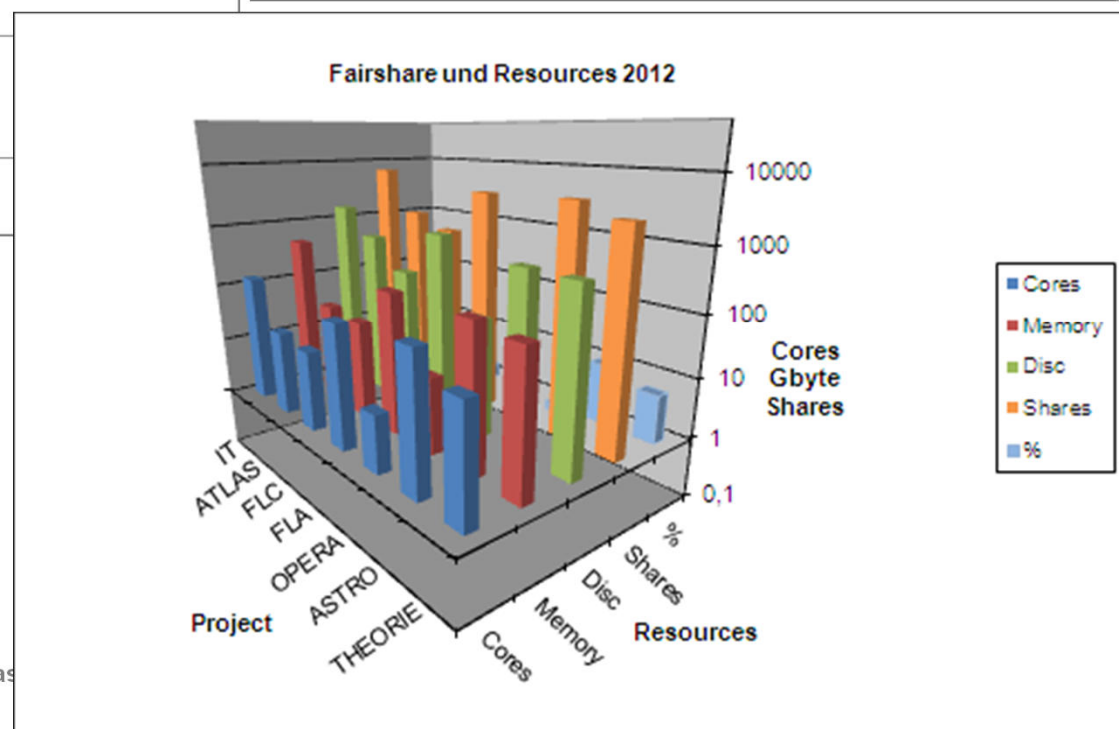
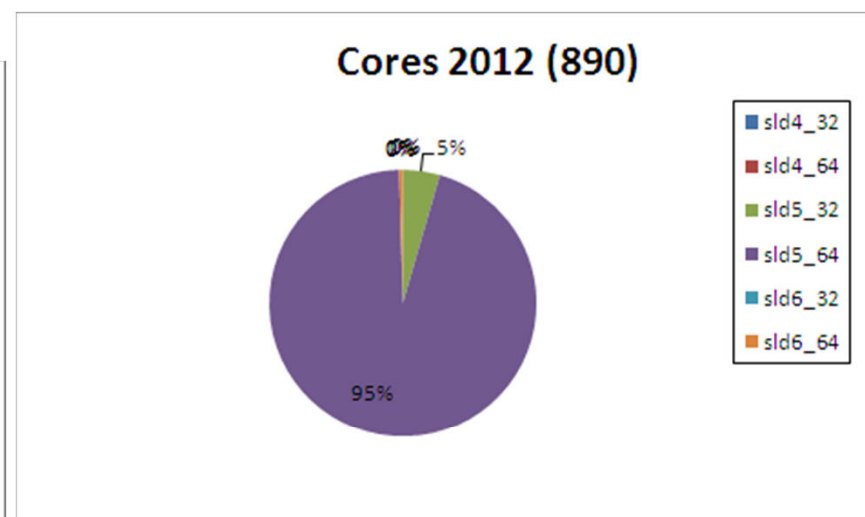
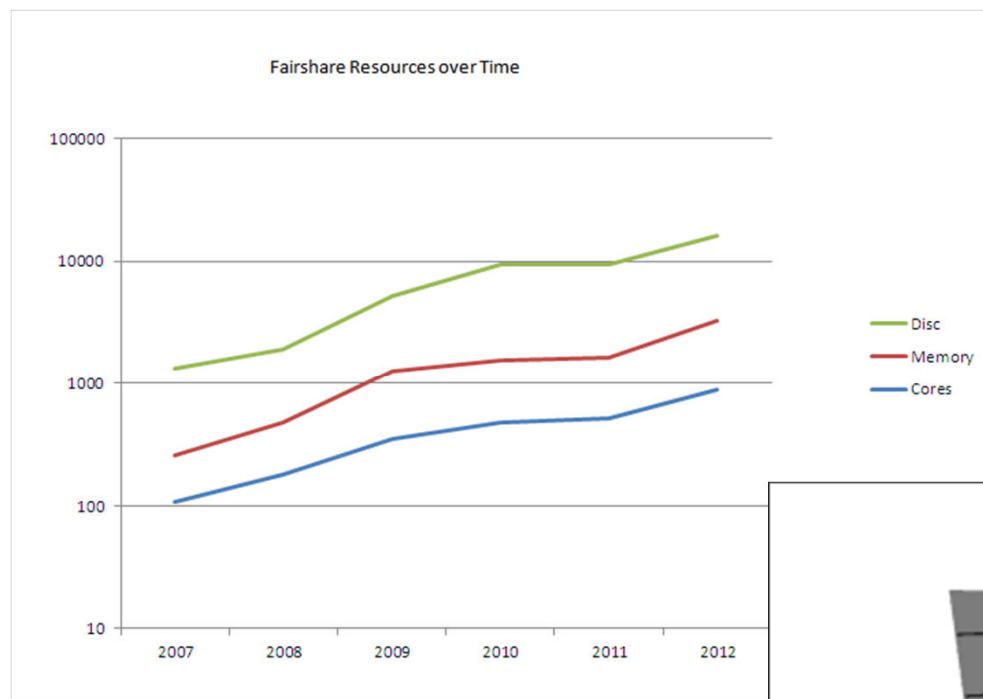
- Runs on Sun Grid Engine Version 6
- ~900 Cores for Interactive and Batch Jobs
- 8-192 GByte Memory per Host
- Minimum 2GByte Memory / Core
- 32-1600 GB Scratch Disk per Host
- Operating System sld5 + sld6 OS in 64 Bit
- 400+ Users from 35+ Different Projects
- Group Specific Software and Storage
- Submit and Control Facilities from PAL, BIRD and Group Specific Hosts
- Select your Resources and we define the Queue

> Advanced

- AFS and Kerberos Support for Authentication and Resource Access
 - Valid Tokens during complete Job Execution
 - Take Cluster
- Cores, Time, Memory and Scratch under full Control of Scheduler
 - “Select your Resources and we guarantee for it”
- MPICH Parallel Environments
 - Single Host, Multi Host, Job Integration
- Big Birds for High Resource Demands
 - Up to 64 Cores / Host
 - Up to 192 GBytes Memory / Host
 - Up to 1600 GByte Scratch / Host
- GPU Support
 - 4 Hosts with 6 GPUs
 - Documentation started
 - SL5 -> SL6

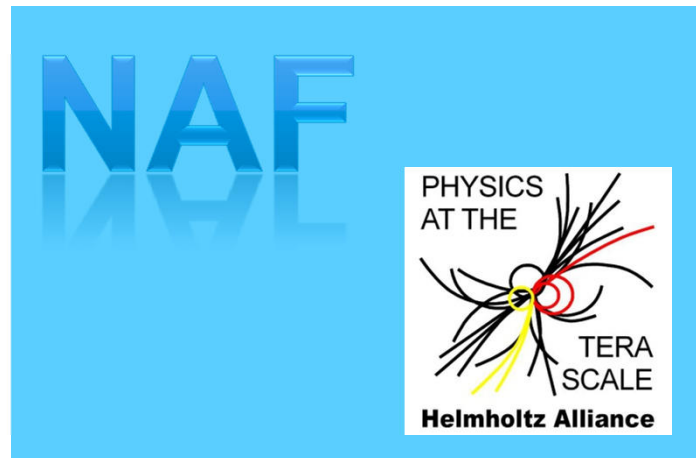


BIRD Graphs

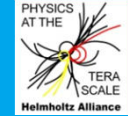


NAF: National Analysis Facility

A Grid Engine System



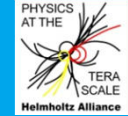
The NAF Mission



- > Facility set up in the Helmholtz Framework “Physics at the Terascale”
- > Provide users of German HEP institutes working on ATLAS, CMS, LHCb and ILC with additional and complementary resources
 - Aim 1: Add more resources to Grid CPU and Grid Storage @ DESY for dedicated analysis
 - Aim 2: Offer complementary resources: Local batch, WGS, fast file system @ DESY
 - (Aim 3: Research on infrastructure dedicated for analysis of data @ DESY)
- > Joint effort between HH and ZN. Current facility designed in 2007 – currently starting NAF evolution and redesign
- > Current hardware: (Aim 2)
 - CPU: 3000 CPU cores in 256 boxes. Currently SL5.7 64bit, managed by SGE batch. (Large hardware part purchased by Uni-HH CMS group!)
 - Storage: separate AFS cell (NAF.DESY.DE), currently Lustre FS, partially migrating to IBM Sonas. Total ~500 TB



The NAF Topology

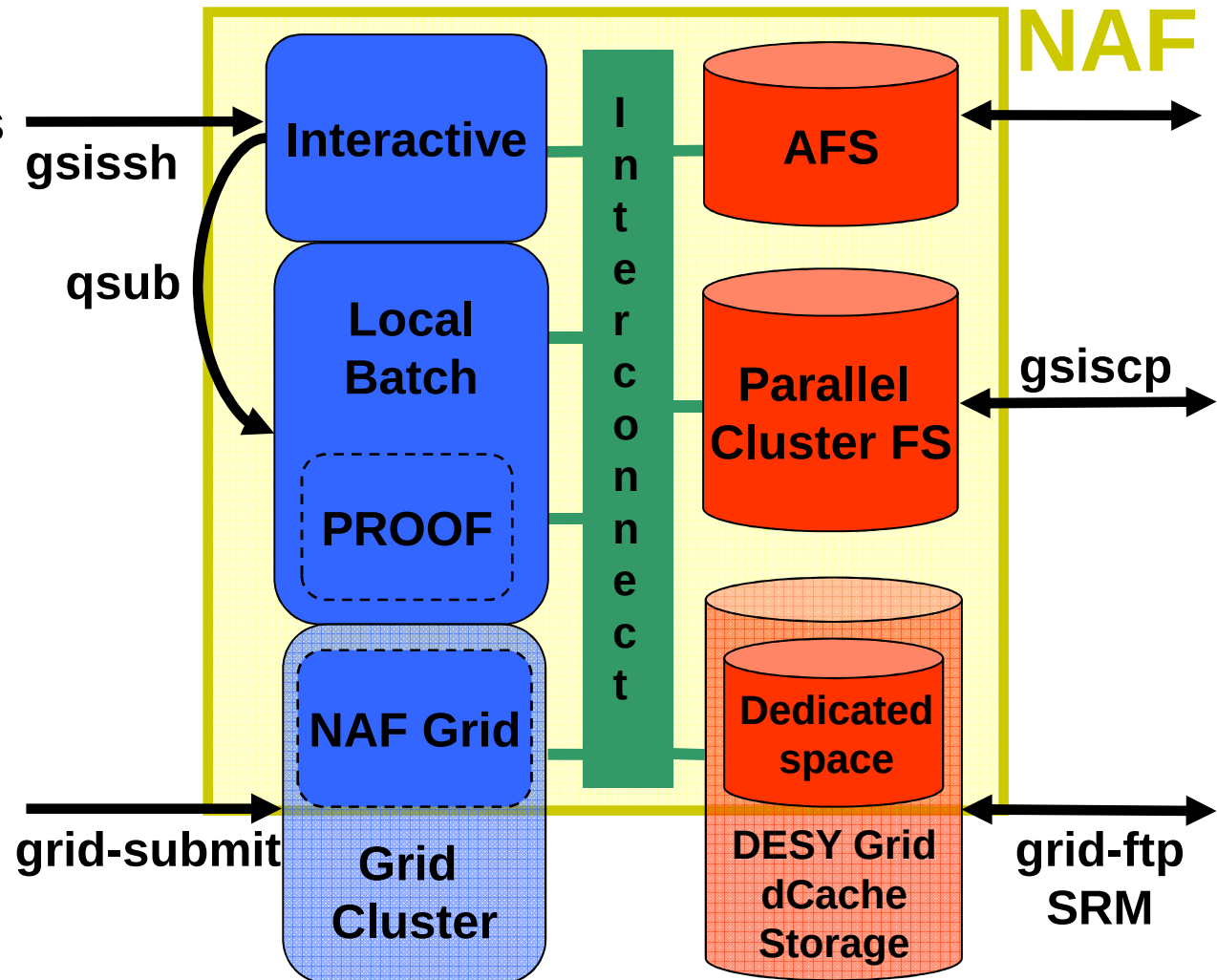


> Login via Grid methods

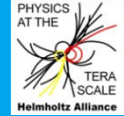
- Subject to change

> Integration with Grid

- NAF share



Lessons learned



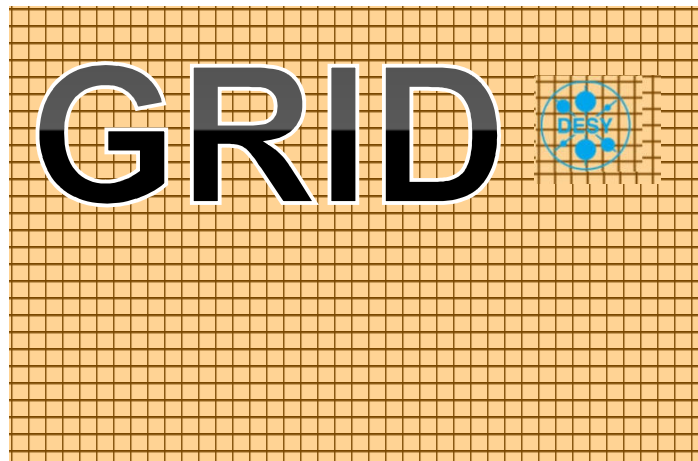
- > Sharing resources among different user groups leads to best usage
- > Bringing in own resources benefits three times:
 - The one that brings in the resources
 - Other users
 - Administrators
- > Know-How: What is needed for user data analysis?

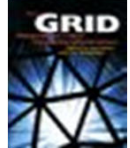


GRID:

Global Resource Information Database

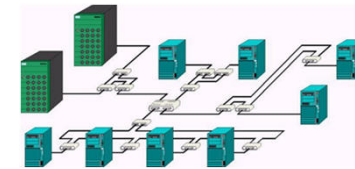
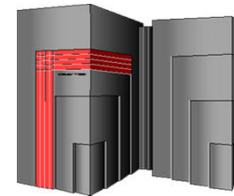
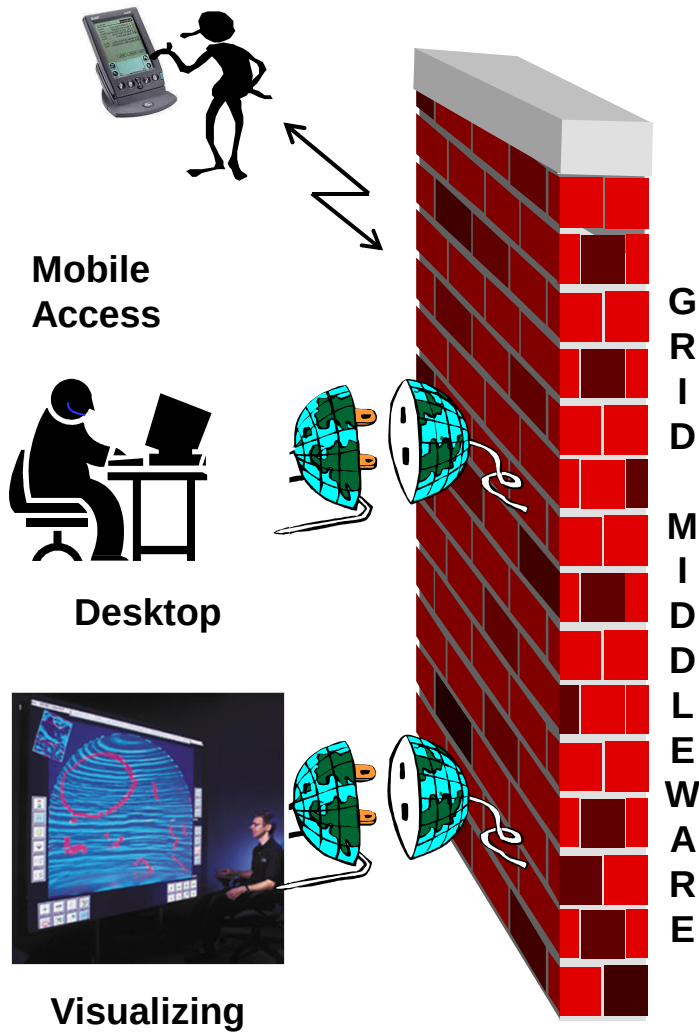
A Torque/Maui System



- > The idea to provide computing resources similar to the *electrical power Grid* was formulated in the 1990s by Foster and Kesselmann 
- > Grid computing is about collaborative virtualization of global resources
- > The most essential building block is the *Virtual Organization* (VO) 
- > In VOs, users utilize global resources according to common sharing rules
- > A Grid infrastructure consists of *services* and *resources*
- > The Grid is the main resource provider for HEP where *collaborations* form *VOs*
- > LHC cannot live without the *WLCG* (Worldwide LHC Computing Grid) 
- > The Grid has become a *key technology* for e-science 
- > Projects: *D-Grid* (German Grid Initiative), *EGI* (European Grid Initiative)

Introduction: The Grid dream

GRID



Supercomputer, PC-Cluster



Data Storage, Sensors, Experiments



Internet, Networks



- > Grid computing started at DESY in 2004:
 - DESY is the *home* of 10 VOs (site: DESY-HH)
 - (WLCG-)Tier-2 for ATLAS, CMS, and LHCb in Germany (Tier-1: GridKa)
 - HERMES, H1, ZEUS; ILC, CALICE;
 - BELLE2, IceCube, Biomed, W-ENMR

- > One complete *generic* Grid infrastructure for *all* VOs
 - *Federated* resources w/ *opportunistic* usage (“everybody profits”)
 - Flexible and scalable to support new VOs
 - Roughly 2/3 of the resources are currently used by the Tier-2 VOs

- > Grid is *complemented* by the National Analysis Facility (NAF) [size: ~1 Tier-2]



- > The Grid infrastructure is the largest Linux installation at DESY
- > Grid services: (Core servers)

- Servers: ~50
- OS: SL 5.6/64-bit (x86_64)

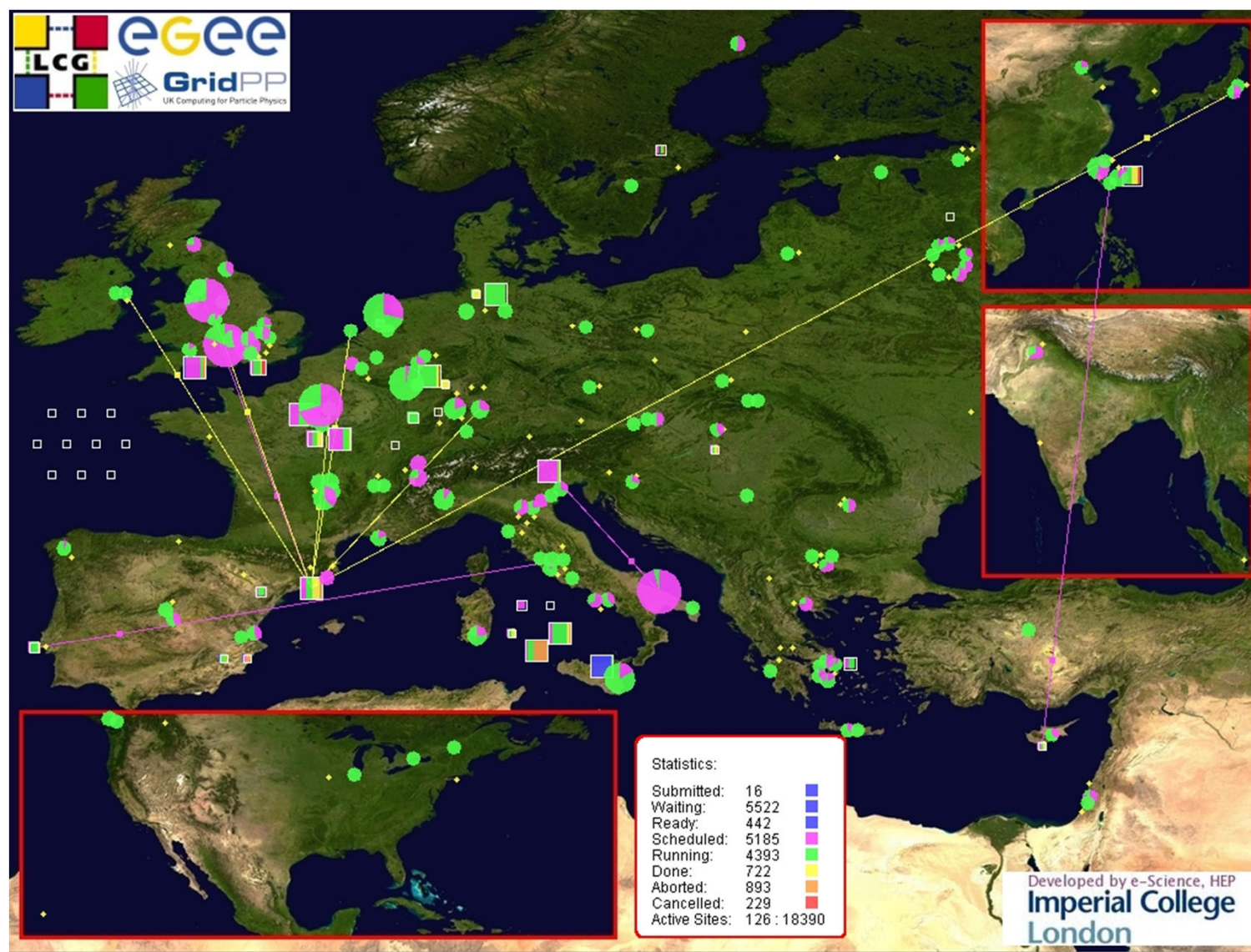


- > Computing Resources: (Computing Elements)

- Compute nodes: 370 hosts, 808 procs, 3504 cores, 4784 job slots
2GB RAM/slot, 15GB scratch space/slot
- Processing power: ~38 kHEPSPEC
- OS: SL 5.6/64-bit (x86_64)

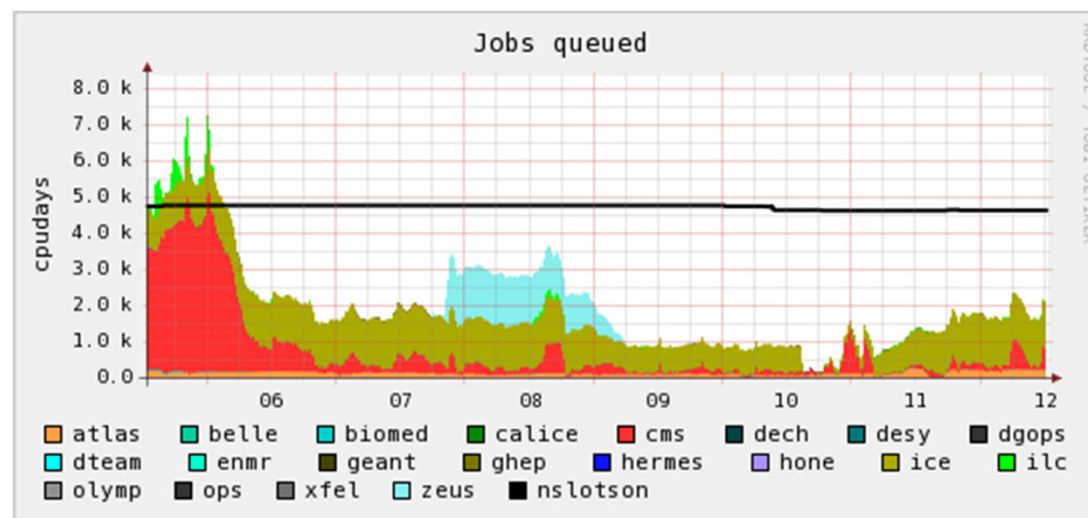
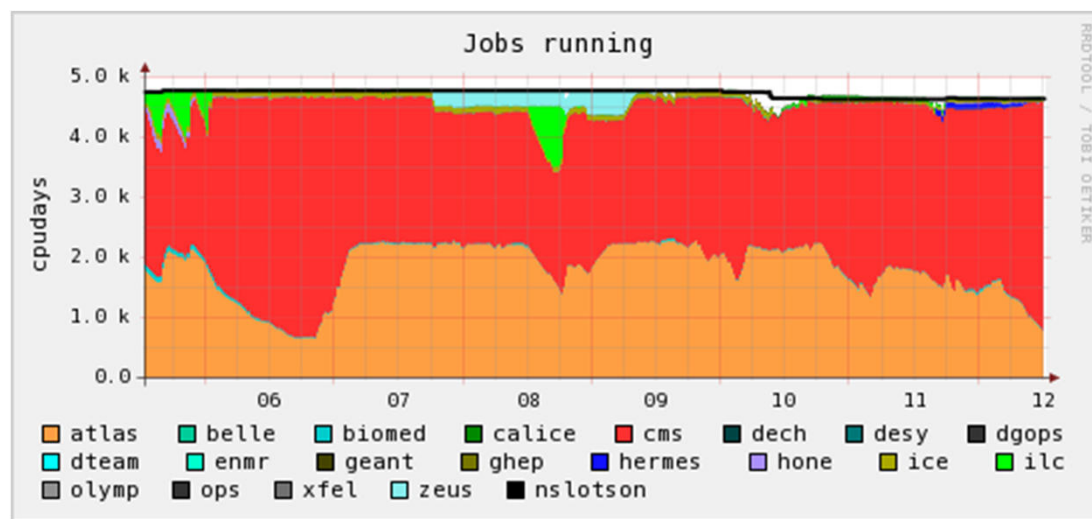
- > (Disk) Storage Resources: (Storage Elements)

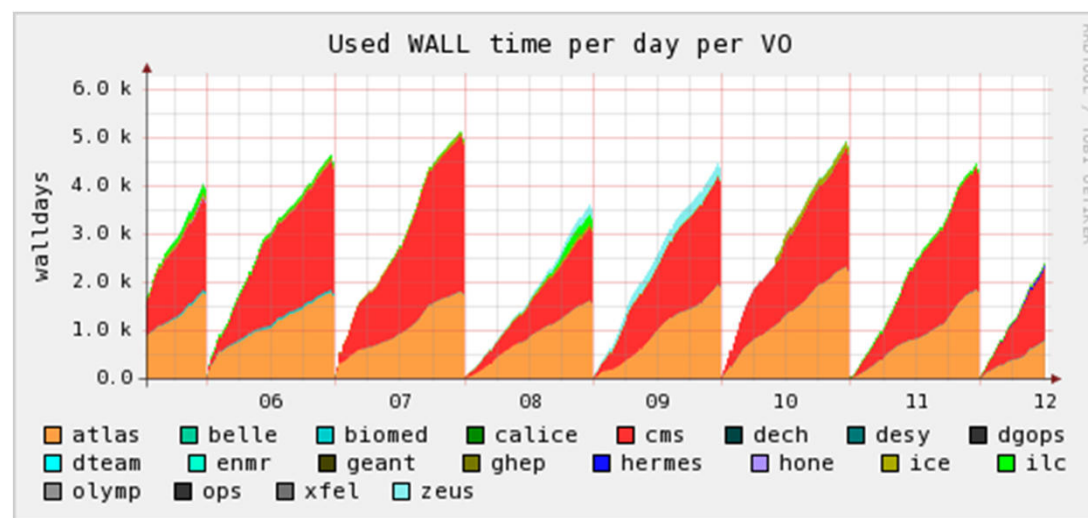
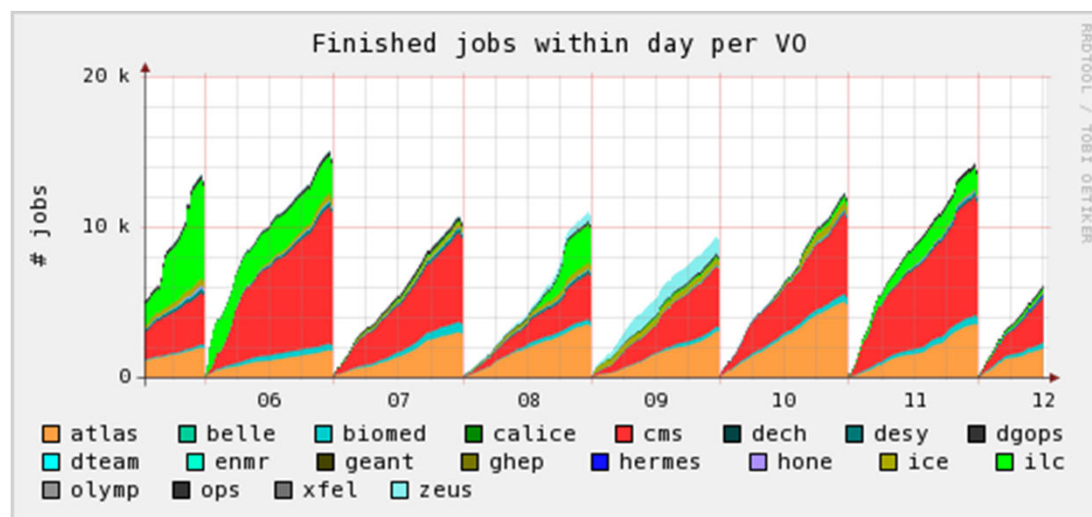
- Total: 4300 TB



DESY Grid Center: Jobs at DESY-HH (weekly)

GRID

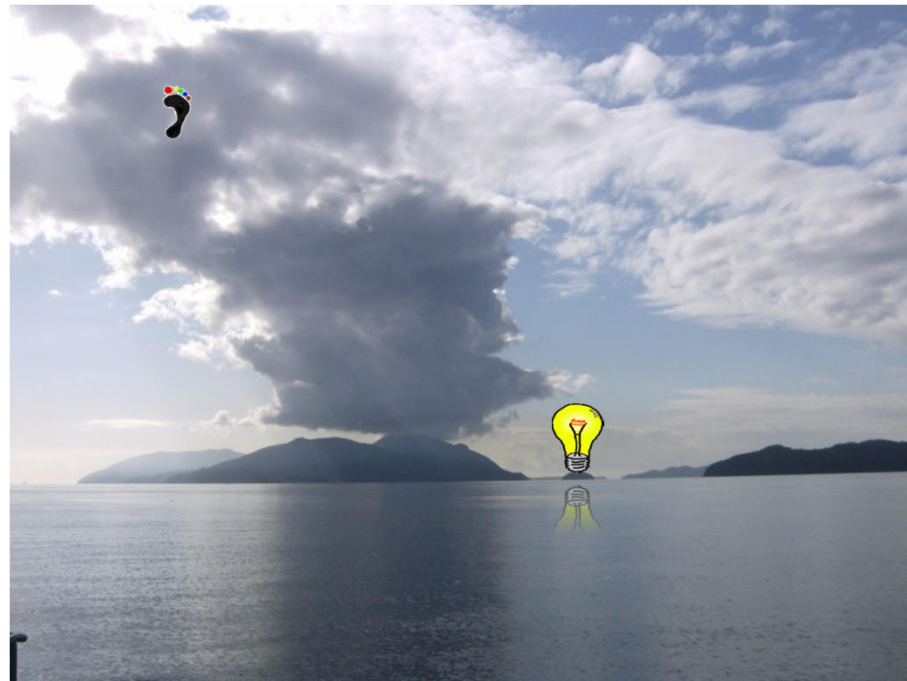




Clouds are on the Track !

Outlook on Clouds and Virtualization

OpenNebula



Wikipedia:
Cloud computing is a marketing term for technologies that provide computation, software, data access, and storage services that do not require end-user knowledge of the physical location and configuration of the system that delivers the services ...





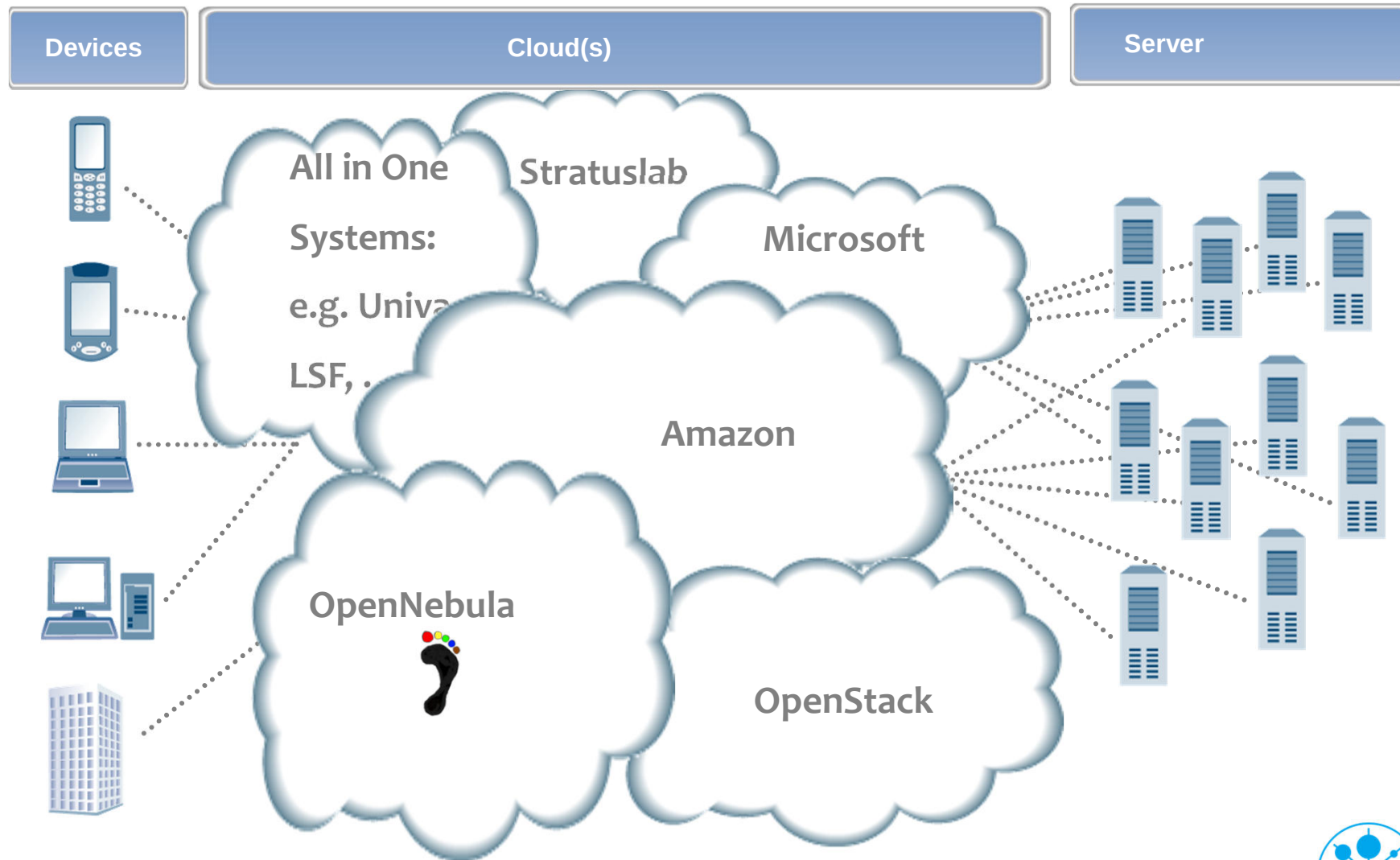
> Public Cloud Promises

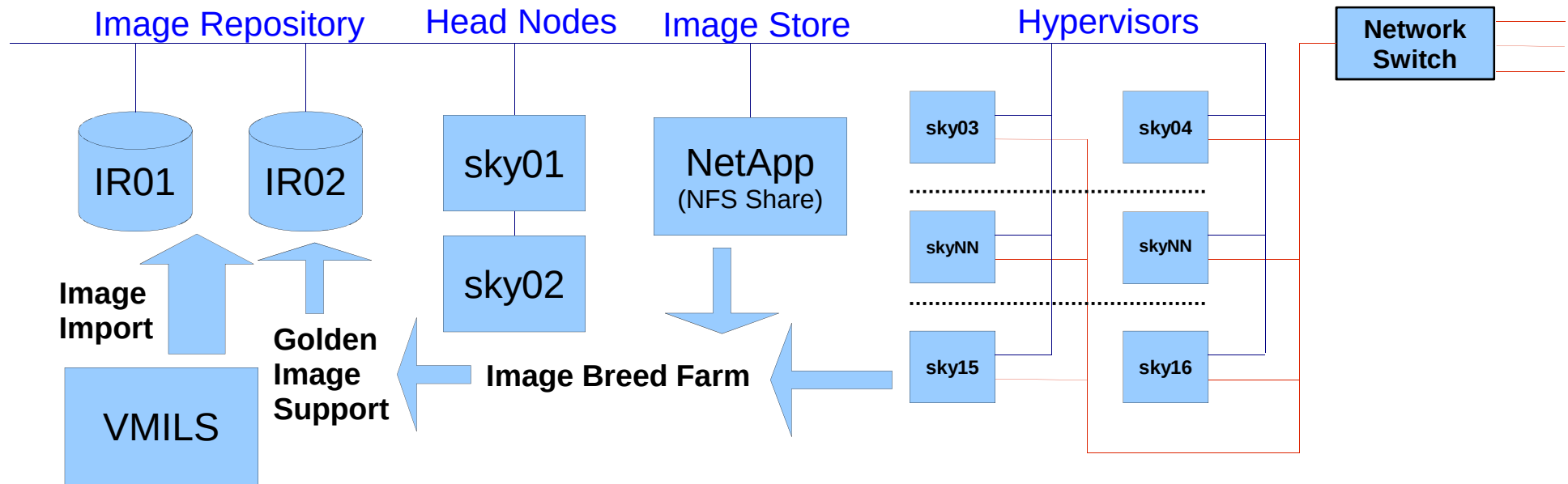
- Unlimited Computing Resources available on Demand
- No up-front Commitment by the User
- Pay for use of computing resources only as needed

> Does it fit to IT or DESY Culture ?



The CLOUD Topologies





> Blue: Management LAN

- OpenNebula 3.0
- WBOOM/QIP Integration
- SL6 KVM Hypervisor
- 256 private IPs

> Red: VM LANs

- Customer VLANs with
- 512 private Cloud IPs
- 512 public Cloud IPs
- nnn Selected DESY IPs

> To Do

- HA Head Nodes
- Enhanced Storage (Local/Shared)
- VLAN Tagging (Open vSwitch)
- ...





> User Interfaces

- EC2
- Occi ?
- GUI ??

> Cloud Basics

- Image Handling
- Contextualisation
- Network Infrastructure
- Storage
- Hypervisors

Type, Resources, Speed, ...

> Usage

- APIs
- Power Admins
- Batch Backend
- Dynamic Batch Resources
- Temporary Resources on Request
- ...

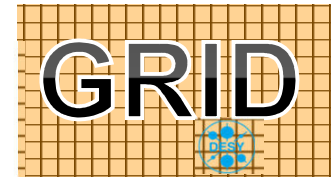
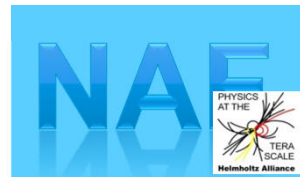
> but (currently) not ...

- Long Term VMs (XEN Cluster)
- ...



Large Public Computing Infrastructures at DESY

Questions ?



OpenNebula