

Harnessing the grid for HERA experiments



Sanjay Padhi

Deutsches Elektronen - Synchrotron



Outline

- Introduction
- Grid developments
- Grid evolution and its evolution@desy
- ZEUS setups and grid@zeus
- Summary and conclusions



Introduction

The term Grid originated in the mid1990's:

Distributed computing infrastructure for advanced science and engineering

(First embraced by the HEP community)

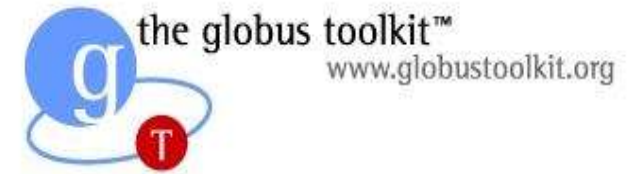
A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to high-end computational capabilities

Carl Kesselman, Ian Foster (1998)



Globus:

- Toolkit (1998 - ...)
- Argonne, U. Chicago, ISI California, University of Edinburgh and Swedish Center for Parallel computers.



EDG (European Data Grid)

- Leading Vehicle for Grid research and deployment
- Assembled large scale grid testbed and defined middleware architecture
- Funded for 3 years (April 2001 – March 2004)



LCG (LHC Computing Grid):

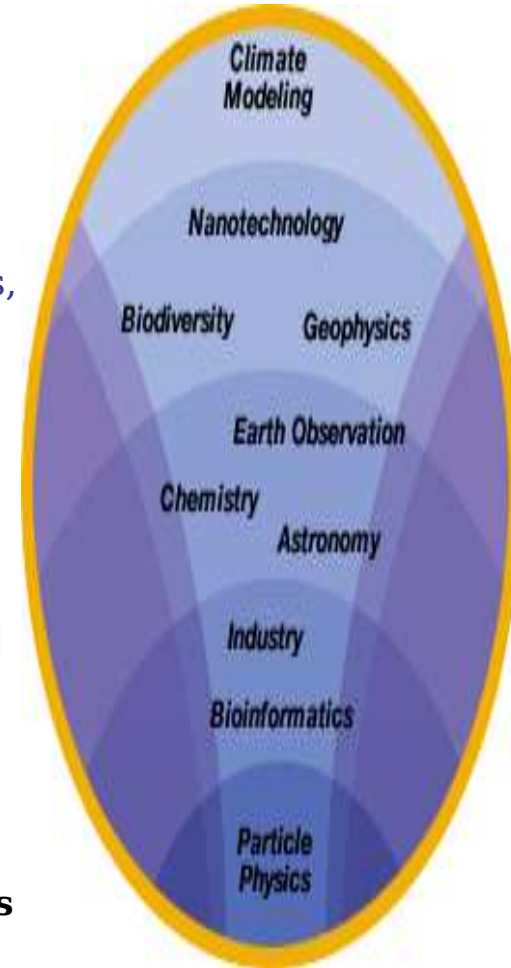
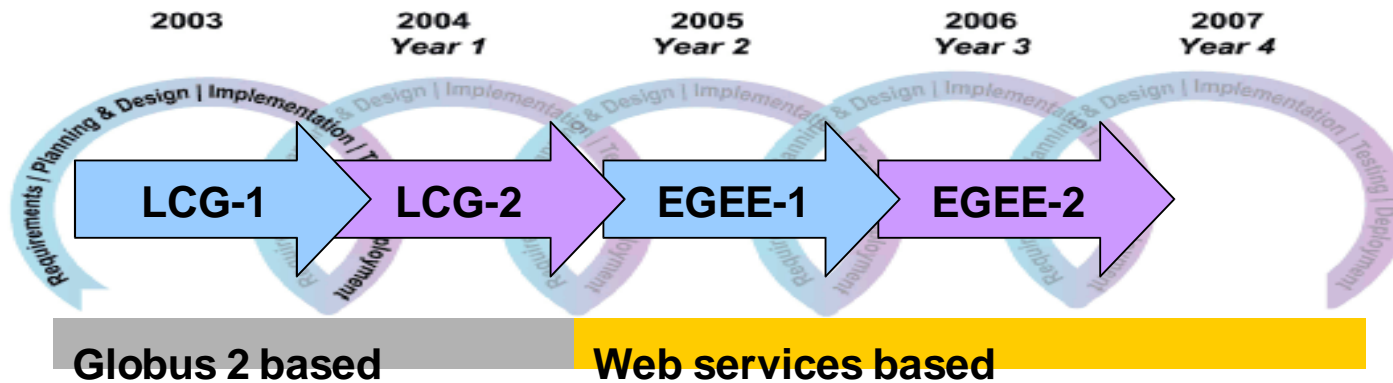
- Grid infrastructure for LHC production
- Based on stable EDG versions
- LCG2 for data challenges in 2004 (EDG 2.0)
- LCG is NOT a development project for middleware



Grid Technologies

EGEE (Enabling Grids for E-Science in Europe)

- ◆ Starts April 2004 for 2 + 2 years
- ◆ Builds on the existing LCG infrastructure to provide expanded grid facility for many application domains
- ◆ Create & operate a production quality infrastructure
- ◆ Identify and support a broad range of applications from diverse domains, starting with the pilot domains: HEP and Biomedical



Web services: a software application accessible via Internet protocols

Who is using the grid at DESY ?

Is it possible to run grid under DESY linux 5 (SuSE 8.2) ?

What are the service/expectation from DESY -IT ...(Development Vs Production env.)

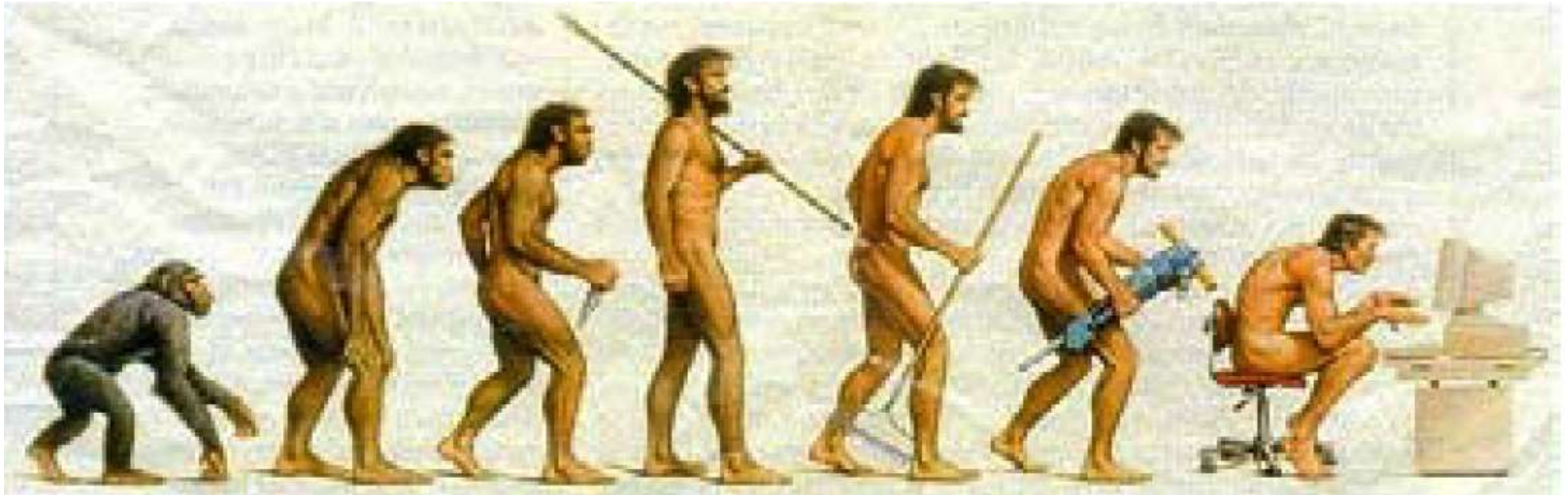
How one can distinguish the individual experiment specific resources ? VOs?

Requirements and authorization for Non LHC experiments

Is it really possible to do the production using Grid ?

- HERA-II programme leads to a sizable increase in MC simulation demand
 - increased luminosity
 - increased event complexity (MVD, STT, ...)
 - distributed computing shifts worldwide to grid technology

Grid Evolution



But which direction ?

Grid Testbed based on EDG software version 1.4 was initiated by IT and H1

- ◆ 10 SuSE linux PCs donated by various groups.
- ◆ The main goal of the Grid Testbed is to study the feasibility to deploy EDG software at DESY.
- ◆ Gain experience of how to operate a Grid site.
- ◆ The Grid Testbed set-up is independent of any experiment specific software settings.
- ◆ Present participants are from IT, H1, HERA-B and ZEUS.
- ◆ Simple tests “hello world” performed by IT (See A. Gellrich talk, computing seminar 11th Nov. 2003)

ZEUS joined the grid efforts more recently than others (Oct. 2003)

ZEUS members: Rainer Mankel, Sanjay Padhi

- ◆ ZEUS integrated into EDG-DESY based testbed
- ◆ ZEUS for the first time managed successfully in running the generator level MCs (CASCADE & HERWIG)



Grid Evolution at DESY

Issues/problems: [See the talk by Max Vorobiev]

1. Centralized information services GIIS or top level MDS, was needed for full operation with CEs and SE
2. EDG 1.4 was not the most recent/advanced grid middleware
3. Firewall problems to/from outside DESY sites not clear (ipchains or tunneling to a given port ??)
4. Huge gap between evaluation and real production .. **rather poor test environment**
5. Integration into new storage systems – dCache, lustre in question

ZEUS initiated LCG-2 (LHC computing grid), with the core services under HEP-grid (non LHC) within the DESY Grid group.

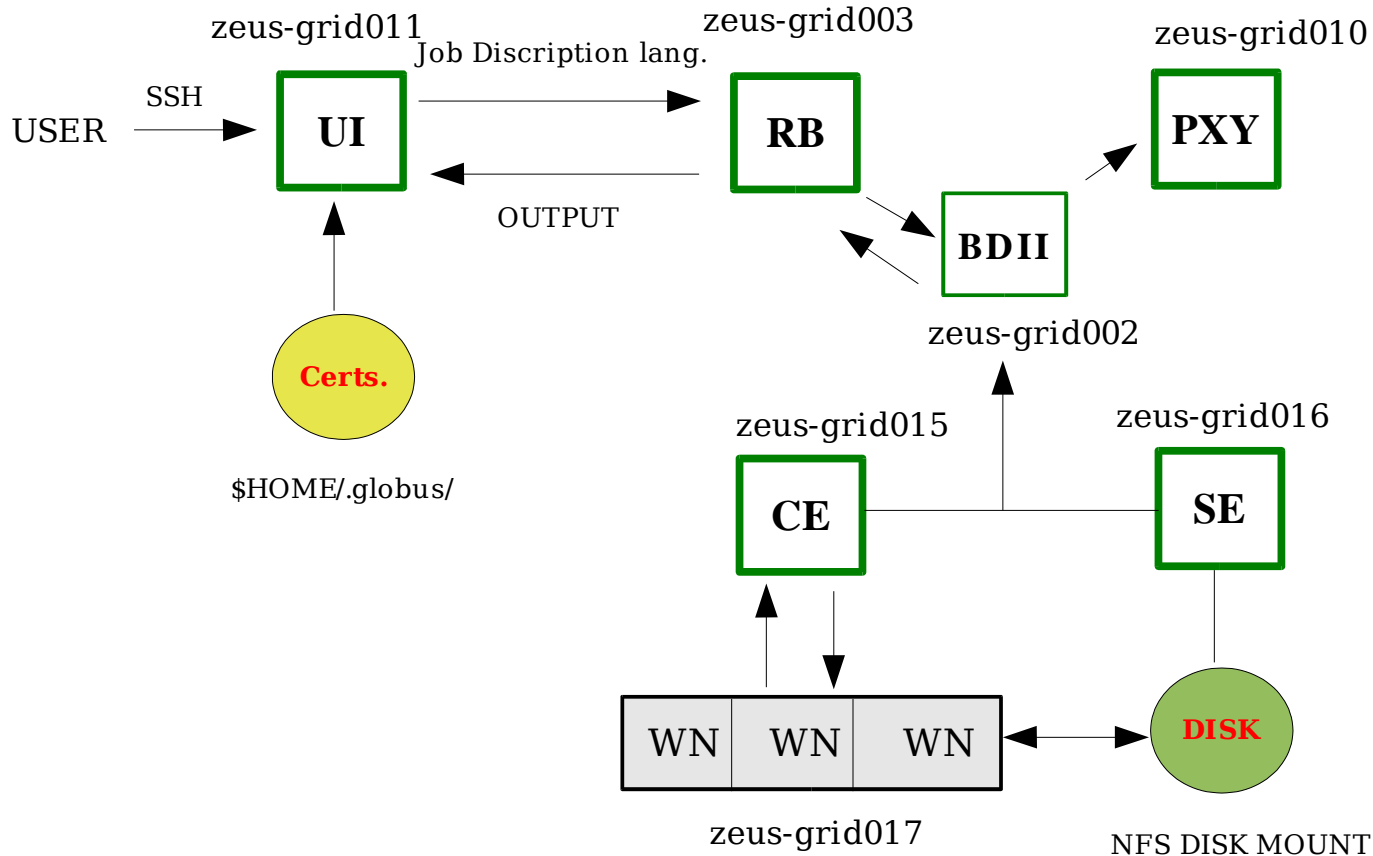
-- Compatible with EGEE (Enabling Grids for E-science in Europe) philosophy.

Almost parallel development of LCG2 testbeds using DL5-SuSE-8.2 (ZEUS) and RedHat-7.3(IT) OS

Grid Elements

- * **Resource Broker (RB)**: the module that receives users' requests and queries the Information Index to find suitable resources.
- * **Information Index (II OR BDII)**, which can reside on the same machine as the Resource Broker, keeps information about the available resources.
- * **Replica Manager (RM)**, used to coordinate file replication across the testbed from one Storage Element to another. This is useful for data redundancy but also to move data closer to the machines which will perform computation.
- * **Replica Catalog (RC)**, which can reside on the same machine as the Replica Manager, keeps information about file replicas. A logical file can be associated to one or more physical files which are replicas of the same data. Thus a logical file name can refer to one or more physical file names.
- * **Computing Element (CE)**, the module that receives job requests and delivers them to the Worker Nodes, which will perform the real work. The Computing Element provides an interface to the local batch queuing systems. A Computing Element can manage one or more Worker Nodes. A Worker Node can also be installed on the same machine as the Computing Element.
- * **Worker Node (WN)**, the machine that will process input data.
- * **Storage Element (SE)**, the machine that provides storage space to the testbed. It provides a uniform interface to different Storage Systems.
- * **User Interface (UI)**, the machine that allows users to access the testbed.

The LCG2 testbed



Grid Setup:

Core elements:

- RB, BDII, PXY

Grid Supports:

- UI, CE, SE, WNs

zeus-grid014 (CE) and zeus-grid012(RB) for development purpose

Authentication mechanism - Standard Passwd Vs GSI:

Grid Security Infrastructure (GSI) is used for enabling secure authentication and communication over an open network.

Primary motivations:

- ◆ The need for secure communication (authenticated) between elements of computational grid.
- ◆ The need to support security across organizational boundaries (centrally-managed security).
- ◆ The need to support “single sign-on” for users of the grid, including delegation of credentials for computations that involve multiple resources and/or sites.

GSI is based on public key encryption, X.509 certificates and the Secure Sockets Layer (SSL) communication protocol

The central concept in GSI authentication is the **certificate**
Every user and service on the grid is identified via a certificate

The LCG2 Testbed

Authentication Checks:

```
zeus-grid011:spadhi> grid-proxy-init
Your identity: /O=GermanGrid/OU=DESY/CN=Sanjay Padhi
Enter GRID pass phrase for this identity:
Creating proxy ..... Done
Your proxy is valid until: Fri Apr 16 02:09:02 2004
```

```
zeus-grid011:spadhi> grid-proxy-info
subject : /O=GermanGrid/OU=DESY/CN=Sanjay Padhi/CN=proxy
issuer  : /O=GermanGrid/OU=DESY/CN=Sanjay Padhi
type    : full
strength : 512 bits
path    : /tmp/x509up_u2371
timeleft : 11:58:50
```

```
zeus-grid011:spadhi> myproxy-info -s zeus-grid010 -d
username: /O=GermanGrid/OU=DESY/CN=Sanjay Padhi
owner: /O=GermanGrid/OU=DESY/CN=Sanjay Padhi
timeleft: 167:56:27 (7.0 days)
```

```
zeus-grid011:spadhi> myproxy-info -s zeus-grid010 -d
username: /O=GermanGrid/OU=DESY/CN=Sanjay Padhi
owner: /O=GermanGrid/OU=DESY/CN=Sanjay Padhi
timeleft: 167:56:19 (7.0 days)
```

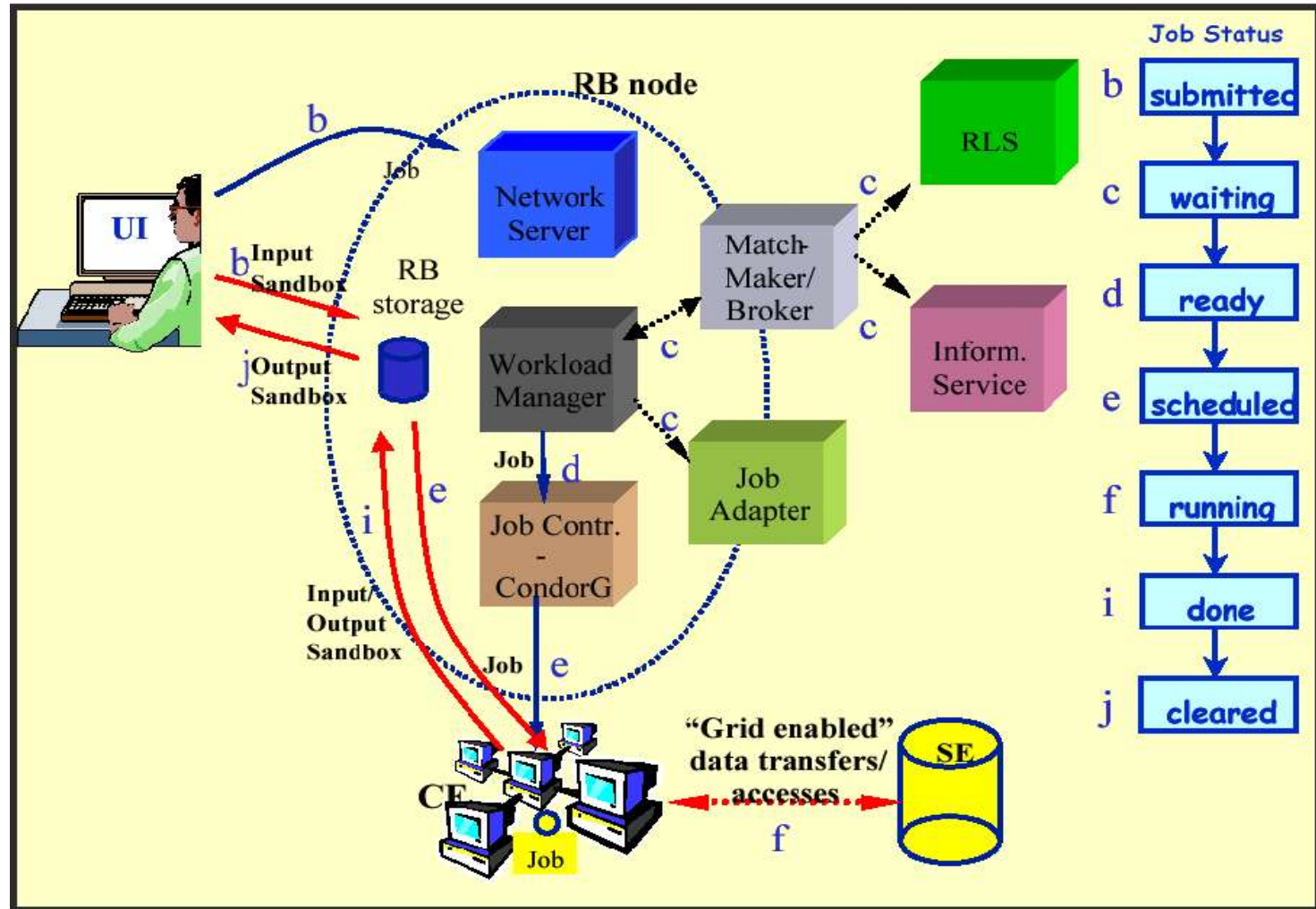
The LCG2 Testbed

LCG Structure is rather complex: but has several advantages

Input Sandbox is what you take with you to the node

Output Sandbox is what you get back

Failed jobs are resubmitted 3 times



The LCG2 Testbed

Before submitting a job one might want to see where can one run

```
zeus-grid011:spadhi> edg-job-list-match --vo dteam testjob.jdl
```

VO : individuals, institutions, and organizations that share a common goal

```
Selected Virtual Organisation name (from --vo option): dteam  
Connecting to host zeus-grid003.desy.de, port 7772
```

```
*****
```

COMPUTING ELEMENT IDs LIST

The following CE(s) matching your job requirements have been found:

CEId

```
grid-ce1.desy.de:2119/jobmanager-lcgpbs-infinite  
grid-ce1.desy.de:2119/jobmanager-lcgpbs-long  
grid-ce1.desy.de:2119/jobmanager-lcgpbs-medium  
grid-ce1.desy.de:2119/jobmanager-lcgpbs-short  
hik-lcg-ce.fzk.de:2119/jobmanager-pbspro-lcg  
lcg06.sinp.msu.ru:2119/jobmanager-lcgpbs-infinite  
lcg06.sinp.msu.ru:2119/jobmanager-lcgpbs-long  
lcg06.sinp.msu.ru:2119/jobmanager-lcgpbs-short  
wipp-ce.weizmann.ac.il:2119/jobmanager-lcgpbs-infinite  
wipp-ce.weizmann.ac.il:2119/jobmanager-lcgpbs-long  
wipp-ce.weizmann.ac.il:2119/jobmanager-lcgpbs-short  
zeus-grid015.desy.de:2119/jobmanager-lcgpbs-infinite  
zeus-grid015.desy.de:2119/jobmanager-lcgpbs-long  
zeus-grid015.desy.de:2119/jobmanager-lcgpbs-short  
wn-04-07-02-a.cr.cnaf.infn.it:2119/jobmanager-lcgpbs-dteam
```

DESY -IT

FZK

MSU

Weizmann

DESY -ZEUS

CNAF, Italy

Switching RBs:

- Use the --config-vo <vo config file >
- and --config <conf file>
- find out which RB/CEs you can use

Clearly a need for various VOs within HERA experiments

Need for Grid

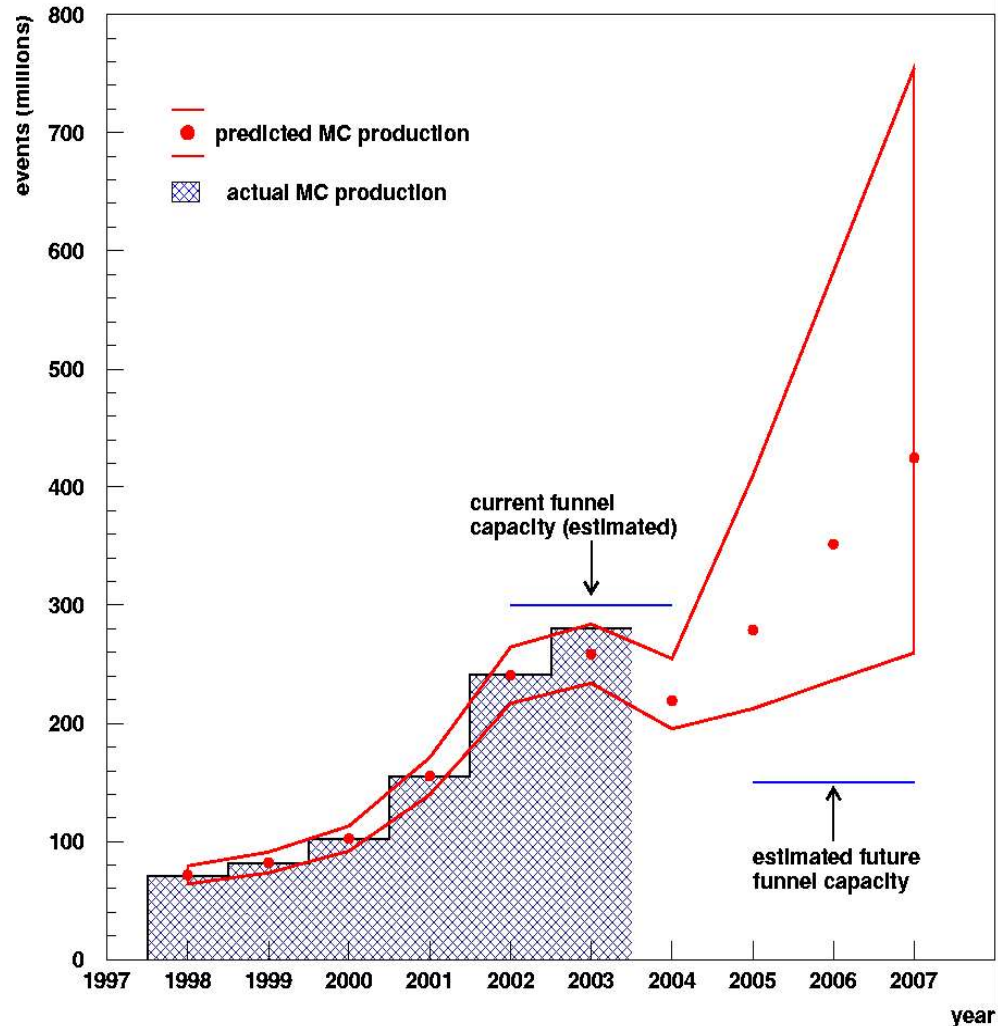
Monte Carlo per year $> 2 \times$ real data events

Need approx 300M/year MC events – HERA II

At least a factor of two or more CPU resources
are required for MC production

Several years of “distributed” MC production
made FUNNEL (ZEUS MC production facility)
compliance with various different platforms

Funnel Production 1998 - 2007



Issues:

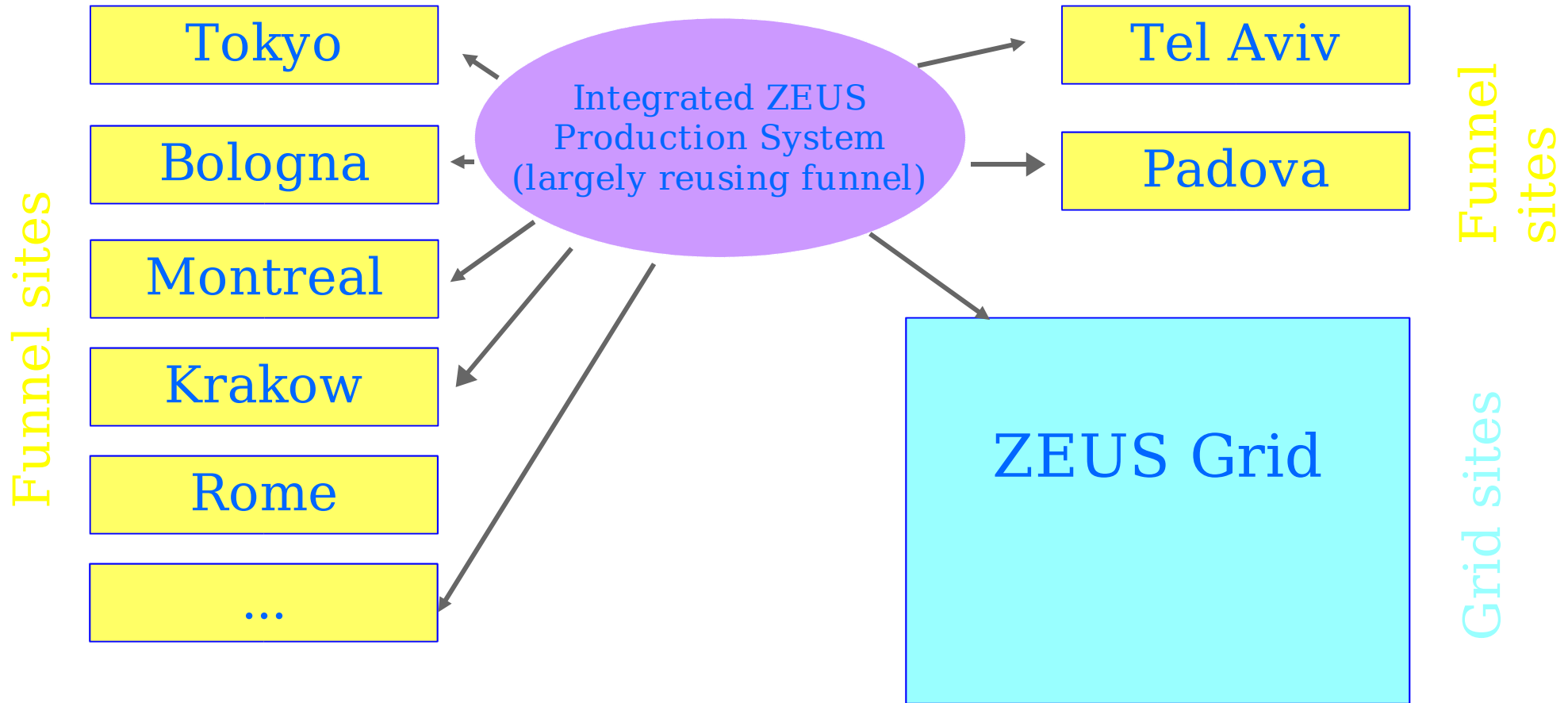
- Need to gain **new resources** through grid
- Update of ZEUS computing strategy in fall 2003 identified enabling MC production for the grid as a major goal
- Need to **keep existing** resources
 - almost 300 M events/year
- User should not have to worry where his jobs runs
 - need **transparent** production system

ZEUS min-term goals:

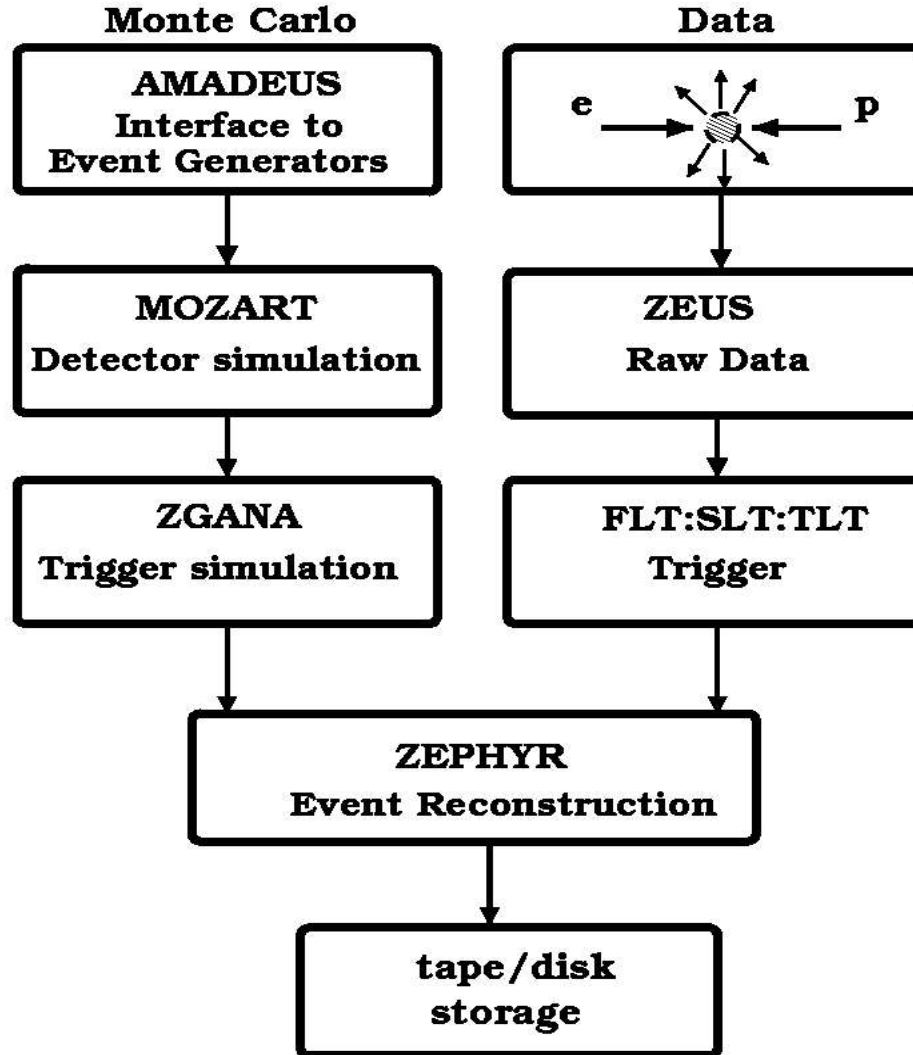
Establish operation of **ZEUS simulation suite** in grid environment (LCG2)

- Establish **interoperation** with external grid sites
- Add **steering** for production mode
- Establish **data flow** for persistent & transient data
- Start **production**

The integrated production concept



ZEUS Environment



1. Start with an event generator:

```
zeus-grid011:herwig> cat herwig.jdl
Executable = "her9800_had.sh";
Arguments = "-vd";
InputSandbox = {"her9800_had.sh","herarun_hrw_9800_had.exe","test2.input"};
Stdoutput = "logfile1";
StdError = "stderr";
OutputSandbox = {"output.rz","logfile1","stderr"};
```

```
zeus-grid011:herwig> edg-job-submit --vo dteam -r zeus-grid015.desy.de:2119/jobmanager-lcgpbs-infinite herwig.jdl
Selected Virtual Organisation name (from --vo option): dteam
Connecting to host zeus-grid003.desy.de, port 7772
Logging to host zeus-grid003.desy.de, port 9002
```

JOB SUBMIT OUTCOME

The job has been successfully submitted to the Network Server.

Use edg-job-status command to check job current status. Your job identifier (edg_jobId) is:

- https://zeus-grid003.desy.de:9000/yr3eQ24FI5bKr4_vBLjPrA

Grid @ ZEUS

```
zeus-grid011:herwig> edg-job-status https://zeus-grid003.desy.de:9000/yr3eQ24FI5bKr4_vBLjPrA
```

```
*****
```

```
BOOKKEEPING INFORMATION:
```

```
Status info for the Job : https://zeus-grid003.desy.de:9000/yr3eQ24FI5bKr4_vBLjPrA
```

```
Current Status: Running
```

```
Status Reason: Job successfully submitted to Globus
```

```
Destination: zeus-grid015.desy.de:2119/jobmanager-lcgpbs-infinite
```

```
reached on: Thu May 13 23:48:40 2004
```

```
*****
```

```
zeus-grid011:herwig> edg-job-status https://zeus-grid003.desy.de:9000/yr3eQ24FI5bKr4_vBLjPrA
```

```
*****
```

```
BOOKKEEPING INFORMATION:
```

```
Status info for the Job : https://zeus-grid003.desy.de:9000/yr3eQ24FI5bKr4_vBLjPrA
```

```
Current Status: Done (Success)
```

```
Exit code: 0
```

```
Status Reason: Job terminated successfully
```

```
Destination: zeus-grid015.desy.de:2119/jobmanager-lcgpbs-infinite
```

```
reached on: Fri May 14 00:44:28 2004
```

```
*****
```

```
zeus-grid011:herwig> edg-job-get-output https://zeus-grid003.desy.de:9000/yr3eQ24FI5bKr4_vBLjPrA
```

```
Retrieving files from host: zeus-grid003.desy.de ( for https://zeus-grid003.desy.de:9000/yr3eQ24FI5bKr4_vBLjPrA )
```

```
*****
```

```
JOB GET OUTPUT OUTCOME
```

```
Output sandbox files for the job:
```

```
- https://zeus-grid003.desy.de:9000/yr3eQ24FI5bKr4_vBLjPrA
```

```
have been successfully retrieved and stored in the directory:
```

```
/tmp/jobOutput/spadhi_yr3eQ24FI5bKr4_vBLjPrA
```

```
*****
```

Grid @ ZEUS

```
zeus-grid011:herwig> ls -l /tmp/jobOutput/spadhi_yr3eQ24FI5bKr4_vBLjPrA
total 816
-rw-r--r--  1 spadhi  zeus    654931 May 14 02:47 logfile1
-rw-r--r--  1 spadhi  zeus    172032 May 14 02:47 output.rz
-rw-r--r--  1 spadhi  zeus         0 May 14 02:47 stderr
```

Charm + Dijets

```
zeus-grid011:herwig> less /tmp/jobOutput/spadhi_yr3eQ24FI5bKr4_vBLjPrA/logfile1
```

HERWIG 6.100 December 1999

OUTPUT ON ELEMENTARY PROCESS

NUMBER OF EVENTS = 100000
NUMBER OF WEIGHTS = 7068511
MEAN VALUE OF WGT = 1.4347E+02
RMS SPREAD IN WGT = 3.7015E+02
ACTUAL MAX WEIGHT = 9.5682E+03
ASSUMED MAX WEIGHT = 1.0258E+04

PROCESS CODE IPROC = 11704
CROSS SECTION (PB) = 1.4347E+05
ERROR IN C-S (PB) = 139.2
EFFICIENCY PERCENT = 1.399

Grid @ ZEUS

2. Use “tarball” for MOZART (Monte carlo for Zeus Analysis, Reconstruction and Trigger) ZGANA (Zeus Geant ANALysis) and ZEPHYR.

```
zeus-grid011:released> cat mc.jdl
Executable = "mc.sh";
Stdoutput = "stdout";
StdError = "stderr";
InputSandbox = {"mc.sh","funnel_run","exe.tgz","funnel.tgz","input.tgz","gaf.tgz"};
OutputSandbox = {"rootdir/funnel/ZEUSMC.SA82P020.E8718.QQ100.ARI.NC99T.Z01",
"rootdir/funnel/mozart.log","rootdir/funnel/mozart.zlog","rootdir/funnel/zephyr.log",
"rootdir/funnel/zephyr.zlog","rootdir/funnel/zgana.log","rootdir/funnel/zgana.zlog",
"stdout","stderr"};
```

```
zeus-grid011:released> edg-job-submit --vo dteam -r \
zeus-grid015.desy.de:2119/jobmanager-lcgpbs-infinite mc.jdl
```

Selected Virtual Organisation name (from --vo option): dteam

Connecting to host zeus-grid003.desy.de, port 7772

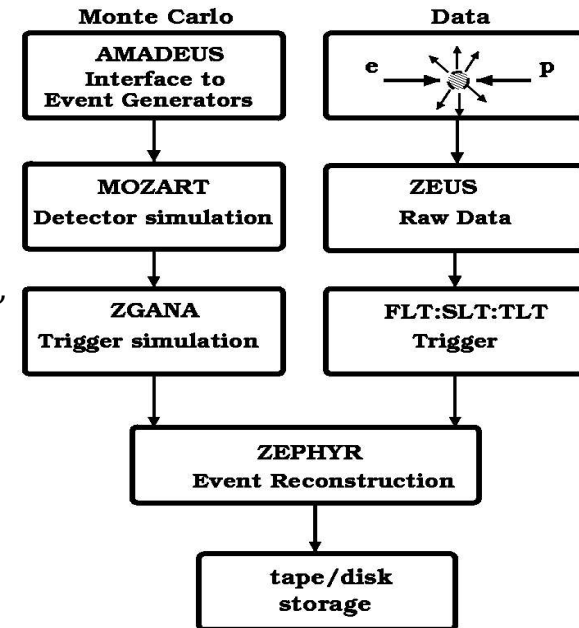
Logging to host zeus-grid003.desy.de, port 9002

JOB SUBMIT OUTCOME

The job has been successfully submitted to the Network Server.

Use edg-job-status command to check job current status. Your job identifier (edg_jobId) is:

- https://zeus-grid003.desy.de:9000/6nSc_vM5LnXxN2TMf79xPQ



Grid @ ZEUS

```
zeus-grid011:released> edg-job-get-output https://zeus-grid003.desy.de:9000/6nSc_vM5LnXxN2TMf79xPQ
Retrieving files from host: zeus-grid003.desy.de ( for
https://zeus-grid003.desy.de:9000/6nSc_vM5LnXxN2TMf79xPQ )
```

```
*****
```

JOB GET OUTPUT OUTCOME

Output sandbox files for the job:

```
- https://zeus-grid003.desy.de:9000/6nSc_vM5LnXxN2TMf79xPQ
have been successfully retrieved and stored in the directory:
/tmp/jobOutput/spadhi_6nSc_vM5LnXxN2TMf79xPQ
```

```
*****
```

```
zeus-grid011:released> ls -l /tmp/jobOutput/spadhi_6nSc_vM5LnXxN2TMf79xPQ
```

```
total 43292
```

```
-rw-r--r--  1 spadhi  zeus  41778000 May 14 01:25 ZEUSMC.SA82P020.E8718.QQ100.ARI.NC99T.Z01
-rw-r--r--  1 spadhi  zeus   752850 May 14 01:25 mozart.log           ← MOZART
-rw-r--r--  1 spadhi  zeus   1145 May 14 01:25 mozart.zlog
-rw-r--r--  1 spadhi  zeus   75087 May 14 01:25 stdout
-rw-r--r--  1 spadhi  zeus  515343 May 14 01:25 zephyr.log           ← ZEPHYR
-rw-r--r--  1 spadhi  zeus   1124 May 14 01:25 zephyr.zlog
-rw-r--r--  1 spadhi  zeus 1126818 May 14 01:25 zgana.log           ← ZGANA
-rw-r--r--  1 spadhi  zeus   1114 May 14 01:25 zgana.zlog
```

~ 500 DIS events simulated

“tarball” has obvious drawbacks:

1. only pre-selected “static” jobs can be submitted
2. use of Noise files, Calibration constants,
- jobs size can be enormous
3. data transfer till the WNs can limit the
bandwidth

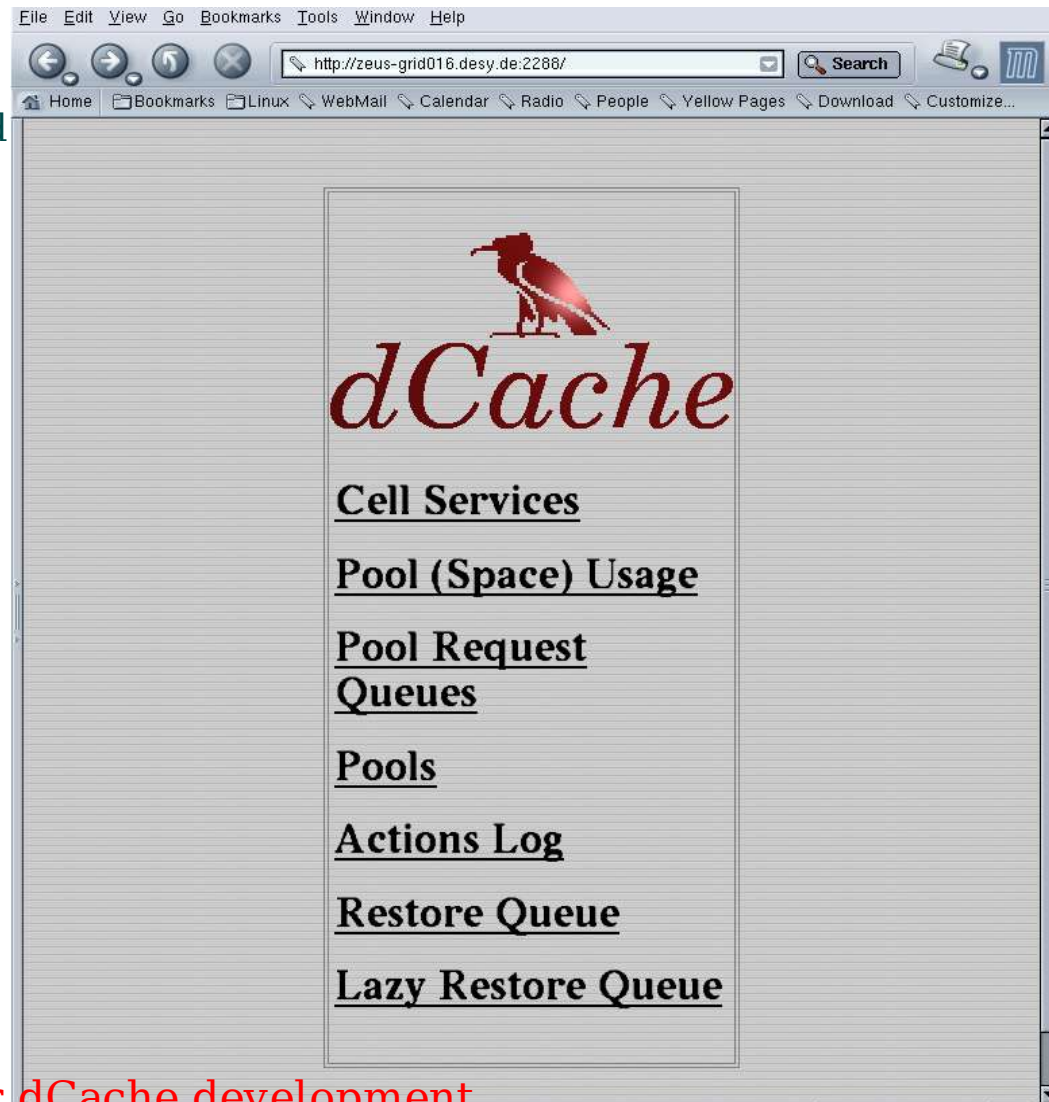
.....

What about the final output storage ?

Thanks to M. Ernst for implementing dCache
on zeus-grid016.desy.de (SE) !!!

- ◆ Based on Java – OS independent
- ◆ Supports GridFTP and SRM

[LCG2 will release the dCache support by the
end of this month]



THANKS !! also to the DESY-FNAL team for dCache development

Grid @ ZEUS



dCache implemented on SE - zeus-grid016.desy.de

Services

CellName	DomainName	Requests Pending	Threads	Ping	Creation Time
PnfsManager	pnfsDomain	0	5	102 msec	05/13 03:49:06
PoolManager	dCacheDomain	0	3	58 msec	05/13 03:48:44
zeus-grid016_1	zeus-grid016Domain	0	9	106 msec	05/13 03:50:53
zeus-grid016_2	zeus-grid016Domain	0	9	114 msec	05/13 03:50:55

There are two dCache pools on zeus-grid016.desy.de

Disk Space Usage

CellName	DomainName	Total Space/MB	Free Space/MB	Precious Space/MB	Layout (precious/used/free)
zeus-grid016_1	zeus-grid016Domain	30720	30673	46	
zeus-grid016_2	zeus-grid016Domain	30720	30678	41	

Grid @ ZEUS

```
zeus-grid011:output> /opt/d-cache/srm/bin/srmcp file:///data/spadhi/released/exe.tgz \  
srm://zeus-grid016.desy.de:8443/pnfs/desy.de/data/dteam/dteam001/zeus/exe.tgz
```

```
copying CopyJob, source = file:///data/spadhi/released/exe.tgz destination = \  
gsiftp://zeus-grid016.desy.de:2811//pnfs/desy.de/data/dteam/dteam001/zeus/exe.tgz
```

```
GridftpClient: connecting to zeus-grid016.desy.de on port 2811
```

```
GridftpClient: gridFTPWrite() wrote 9771956bytes
```

```
zeus-grid011:output> ls -l /pnfs/dteam/dteam001/*
```

```
/pnfs/dteam/dteam001/zeus:
```

```
total 9543
```

```
-rw-r--r--  1 90051  1305  9771956 May 14 08:37 exe.tgz
```

- can be used for reading a file from a remote SE to the local filesystem
- supports true third party transfer, allows batch files
- source and a destination pairs can be used as well

Next step is to use the srm protocols to access the files from the zeus-SE

- store big fz, rz and other compressed/uncompressed files
- access the files using the jdl
- store the jobput (which can be an input for the next) in the SE

Several functionalities exists, some explored, still many needs to be tried ...

Summary

DESY experimental groups are successful

- for the first time in having LCG2 testbed using DL5 -SuSE-8.2 (non RedHat OS)

ZEUS binaries has been successfully used in LCG2 grid environment

MC tests were successfully performed

Established remote execution at external LCG2 sites in progress (INFN, Bonn, MSU ...)

Advanced plans on Integrated ZEUS production system in place using UI as the gateway to the grid

After that long “we started using grid applications” rather than using them in seminar talks ...

Outlook

Consider feasibility of data analysis applications.

Need for DESY -IT to provide and maintain the grid core and information services.

DESY related VOs (H1, FLC, ZEUS ...) needs to be published outside the DESY sites.

Both development and production (stable, transparent, ..) systems are needed.

Issues related OS (RedHat Vs SuSE) for the production environments.

Perfection is reached not when you have added all that can be added,
but when you have removed all that can be removed

– Michelangelo