

Grid-Computing

Ingo Strauch

Mottenlinux-Treffen – 10.05.2005

Übersicht:

- Was versteht man unter dem “Grid”
- Das LHC Computing Grid (LCG)
- Anwendungsgebiete
- Beispiele
- Middleware
- Referenzen

Was versteht man unter dem “Grid”

Begriff angelehnt an “Power-Grid” (d.h. Stromnetz)

Der Verbraucher von Rechenleistung stellt eine Verbindung zum Rechnernetz her, so wie der Stromverbraucher zum Stromversorgungsnetz. Alles was hinter der Steckdose passiert ist für den Konsumenten verborgen, er verbraucht einfach Leistung.

(<http://de.wikipedia.org/wiki/Grid-Computing>)

Aber. . . . es gibt nicht “das Grid” ☹

Unterschiedliche Auffassung von "Grid"

"Grid" ist Modewort mit z.T. sehr unterschiedlicher Bedeutung

- Generelle Typen

- Computational-Grids: Zusammenfassung verteilter Rechenkapazitäten
- Data-Grids: transparenter Zugriff auf Daten an verschiedenen Standorten
- Kombination aus beidem

- Beispiele in der Industrie

- **IBM**: Rechenleistung on demand
- **Oracle**: haupts. hochverfügbarer Cluster
- **Sun Grid Engine**: lokale Rechner als Farm

Wer braucht Grid-Technologie?

- Physik
 - Elementarteilchenphysik (Large Hadron Collider Experimente, . . .)
 - Astronomie (Radioteleskope¹, . . .)
- andere Felder
 - Medizin (Visualisierung, Therapie, . . .)
 - Bio-Informatik (Genom-Analyse, . . .)
 - Ingenieurwesen (Automobil- u. Flugzeugbau, Simulationen, . . .)
 - Klimaforschung (Wettervorhersage, Hochwasser-Simulation, . . .)

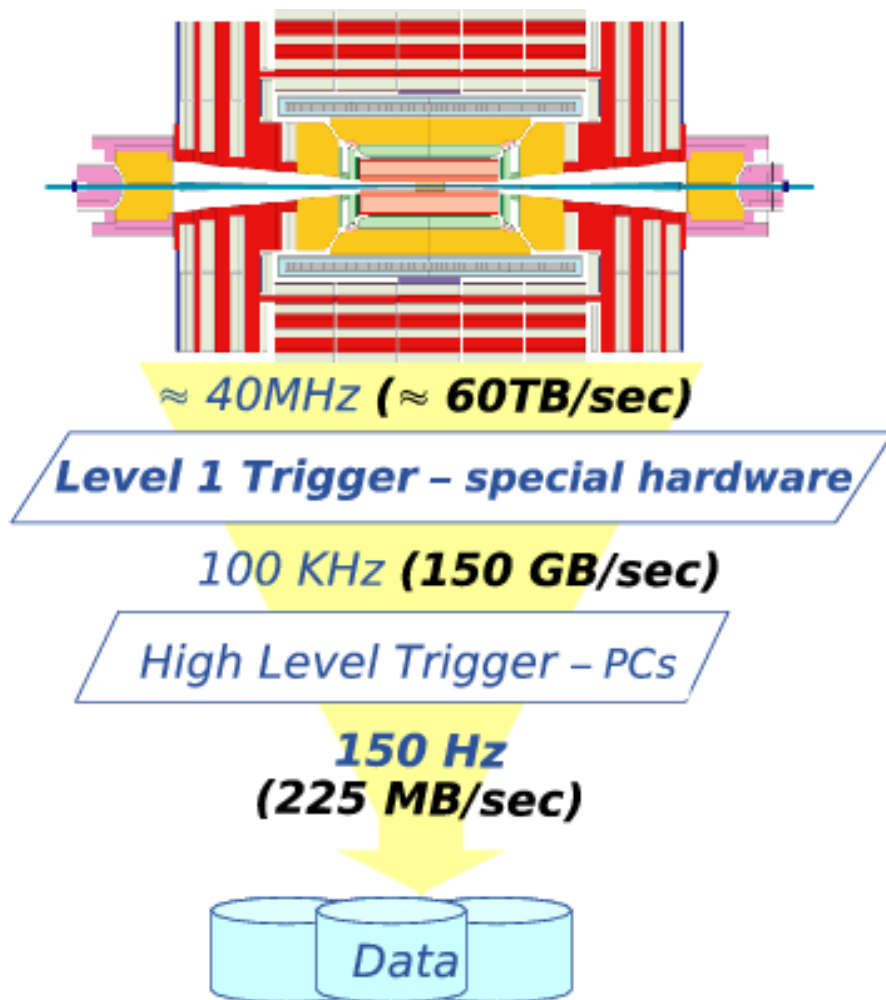
Heute: Beschränkung auf Elementarteilchenphysik; dort wichtigste Grids

LCG, NorduGrid, Grid 3, SAM

¹Artikel in iX 11/04

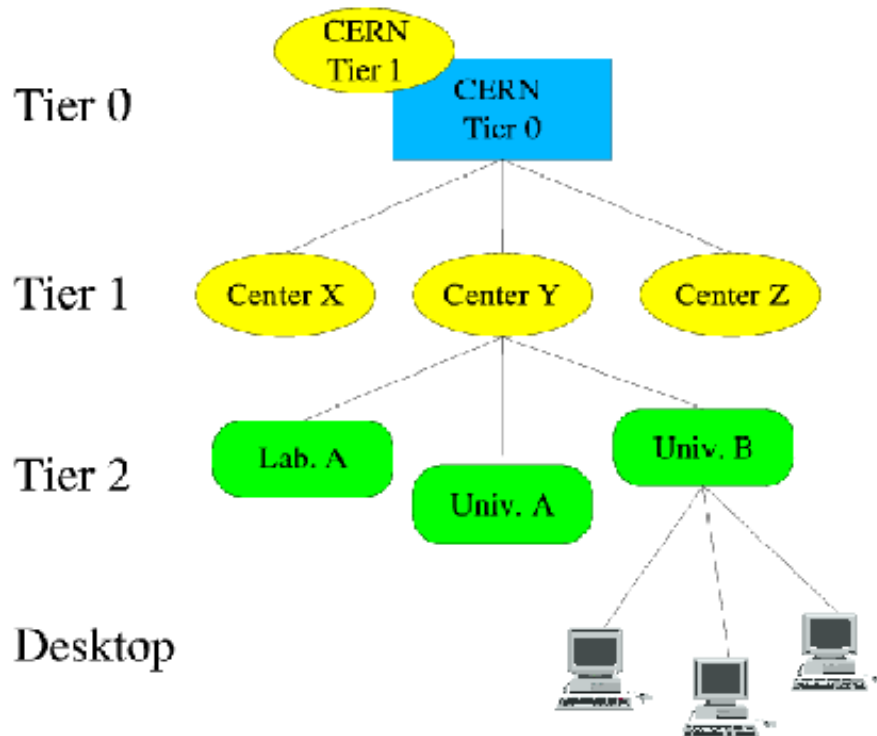
Grids in der Elementarteilchenphysik

Beispiel: das CMS-Experiment am CERN



- eines von 4 Experimenten am Large Hadron Collider
 - Messung von Proton-Proton Kollisionen bei Energie von 14TeV
 - alle 4 Experimente zusammen weltweit
 - ≈ 470 Institute
 - ≈ 6200 Wissenschaftler
 - erwartete Datenrate: ≈ 15 PetaByte/Jahr
- Daten und Rechenkapazität verteilt vorhalten
- LHC-Computing-Grid (LCG)

Schalenstruktur



- hierarchische Datenverteilung ("Tiers")
 - Tier-0: Rekonstruktion, Rohdaten
 - Tier-1: rekonstr. Daten, Reprocessing
zentrale Analysen
 - Tier-2: Ereignis-Simulation, Nutzer-Analysen
 - Tier-3: optional, kleine Gruppen/Unis
- deutsches Tier-1 Zentrum: [GridKa](#)
 - ≈ 500 Rechenknoten mit je 2 CPUs
 - $\approx 200\text{TB}$ Plattenplatz, $\approx 400\text{TB}$ Tape

Ende April: 10 Tage Dauertransfer von Tier-0 zu sieben Tier-1 Zentren @ 600 MB/s
<http://www.heise.de/newsticker/meldung/59027>

Grid-Middleware

Benötige Software-Komponenten, die Komplexität verstecken → **Middleware**
Verschiedene Projekte, hier Beschränkung auf

- **Globus**
 - grundlegendes Toolkit
 - wird von vielen Middlewares benutzt
- **EDG** (EU DataGrid)
 - EU-gefördertes Projekt (04/2001 - 03/2004)
- **LCG** (LHC Computing Grid)
 - produktionsreife Infrastruktur für die LHC Experimente (und andere)
 - basiert auf stabilen EDG-Releases
- **EGEE** (Enabling Grids for E-Science in Europe)
 - EU-gefördertes Projekt (04/2004 - 03/2006, geplant bis 2008)

Acronym-Overkill

BDII: Berkeley Database Information Index

CE: Computing Element

ClassAd: Classified advertisement

CLI: Command Line Interface

DIT: Directory Information Tree

DN: Distinguished Name (LDAP)

EDG: European DataGrid

EDT: European DataTag

EGEE: Enabling Grids for E-science

ESM: Experiment Software Manager

GFAL: Grid File Access Library

GGUS: Global Grid User Support

GIIS: Grid Index Information Server

GLUE: Grid Laboratory for a Uniform Environment

GOC: Grid Operations Centre

GRAM: Globus Resource Allocation Manager

GRIS: Grid Resource Information Service

GSI: Grid Security Infrastructure

GUID: Grid Unique ID

ID: Identifier

IS: Information Service

JDL: Job Description Language

LB: Logging and Bookkeeping Service

LDAP: Lightweight Directory Access Protocol

LFC: LCG File Catalog

LFN: Local File Name

LRC: Local Replica Catalog

LRMS: Local Resource Management System

LSF: Load Sharing Facility

MDS: Monitoring and Discovery Service

MPI: Message Passing Interface

PFN: Physical File name

RB: Resource Broker

RLI: Replica Location Index

RLS: Replica Location Service

RM: Replica Manager

RMC: Replica Metadata Catalog

RMS: Replica Management System

ROS: Replica Optimization Service

SE: Storage Element

SRM: Storage Resource Manager

SURL: Storage URL

TURL: Transport URL

UI: User Interface

VDT: Virtual Data Toolkit

VO: Virtual Organization

WMS: Workload Management System

WN: Worker Node

(aus [LCG-2 User Guide](#))

GSI: Grid Security Infrastructure

Zugang zum Grid erfordert

Authentifizierung: wer bin ich?

Authorisierung: was darf ich?

- **Public Key Infrastructure (PKI)**
 - basiert auf X509 Zertifikaten
 - ausgestellt von Certificate Authorities (CAs), für Deutschland: GridKa
 - gesamte Authentifizierung kommt aus dem Globus Toolkit
 - wer ich bin
- **Virtual Organizations**
 - Zusammenschluß von z.B. Mitarbeitern des gleichen Experimentes
 - Zugriff auf gemeinsame Ressourcen
 - was ich darf

Was kann ich jetzt tun?

Aufteilung in zwei Schwerpunkte

- **Job/Workload Management**

- Jobs starten/abbrechen
 - Programme/Parameter spezifizieren
 - einem Job Dateien mitgeben/Output abholen
- Job Description Language (JDL)

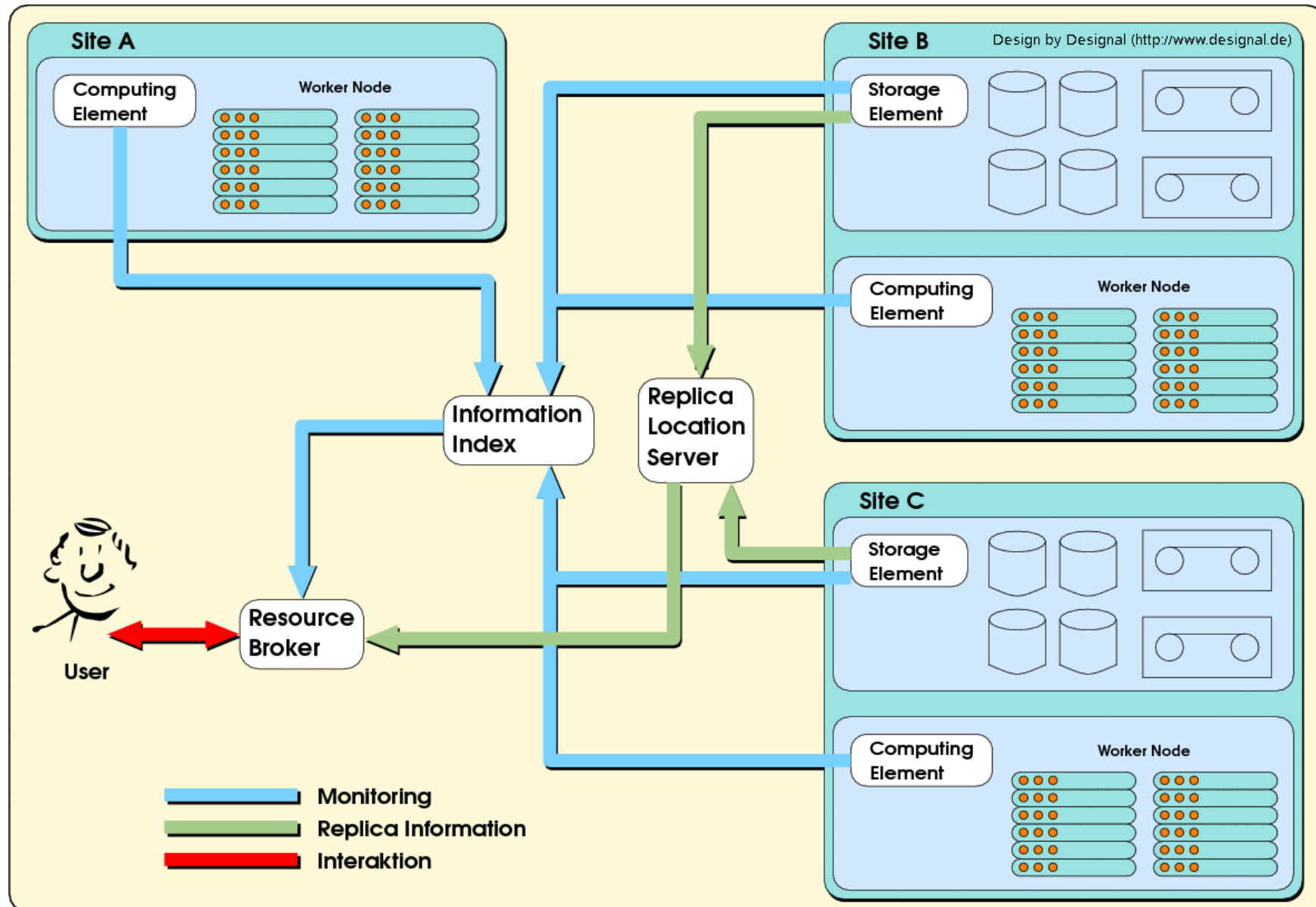
- **Data Management**

- Dateien dem Grid bekannt machen
 - Dateien im Grid von einer Site zu anderen replizieren
 - Dateien ins Grid/aus dem Grid kopieren
- Datenkataloge

Wichtige Komponenten (Rechner/Dienste)

- User Interface (UI)
 - Benutzerinteraktion
- Computing Element (CE)
 - Zielsystem, an dem gerechnet wird
 - i.a. ganze Rechnerfarm
- Worker Node (WN)
 - eigentliche Rechenknoten
 - viele WNs zu einem CE zusammengefaßt
- Storage Element (SE)
 - wo die Daten liegen
- Replica Location Server (RLS)
 - verwaltet Replicas und deren Speicherorte
- Grid Resource Information Service (local GRIS)
 - Infodienst auf CEs & SEs
 - welche Info durch sog. GLUE Schema definiert
- Grid Index Information Server (site GIIS)
 - sammelt Informationen aller lokalen GRISes
 - wie GRISes per LDAP abfragbar
- Berkeley Database Information Index (BDII)
 - zentr. Anlaufstelle für Information System (IS)
 - sammelt Informationen aller site GIISes
- Resource Broker (RB)
 - entscheidet, wo gerechnet wird

Übersicht



Vorbereitung für Arbeiten im Grid

Zum Arbeiten mit dem Grid ist **Proxy-Zertifikat** nötig, d.h. als erstes immer

```
$ grid-proxy-init
Enter GRID pass phrase for this identity:
Creating proxy ..... Done
Your proxy is valid until: Wed May 11 03:23:42 2005
```

Passwort wurde beim Certificate-Request gewählt

Test ob Proxy vorhanden

```
$ grid-proxy-info
[...]
type      : full legacy globus proxy
strength  : 512 bits
timeleft  : 11:14:10
```

Job/Workload Management – Beispiel: Hello World

- erstelle JDL-Datei (hello_world.jdl) mit folgendem Inhalt

```
Executable = "/bin/echo";  
Arguments = "Hello World";  
StdOutput= "stdout";  
StdError = "stderr";  
OutputSandbox = {"stdout","stderr"};
```

- ermittle verfügbare Ressourcen

```
edg-job-list-match --vo <vo_name> hello_world.jdl
```

- Job abschicken

```
edg-job-submit --vo <vo_name> hello_world.jdl
```

dabei: OutputSandbox wird am Job-Ende aus dem Grid zurückkopiert

Job/Workload Management (Fortsetzung)

Jeder Job über eindeutige **Job ID** gekennzeichnet, Form:

```
https://grid-rb.desy.de:9000/WLB3aGGdSNUFhgCeIbcJhg
```

- **Job Status abfragen**

```
edg-job-status https://grid-rb.desy.de:9000/WLB3aGGdSNUFhgCeIbcJhg
```

- mom. Status (z.B. SCHEDULED, RUNNING, DONE)
- bei welchem Computing Element gelandet

- **Job Output abholen**

```
edg-job-get-output https://grid-rb.desy.de:9000/WLB3aGGdSNUFhgCeIbcJhg
```

- holt OutputSandbox vom Resource Broker ab

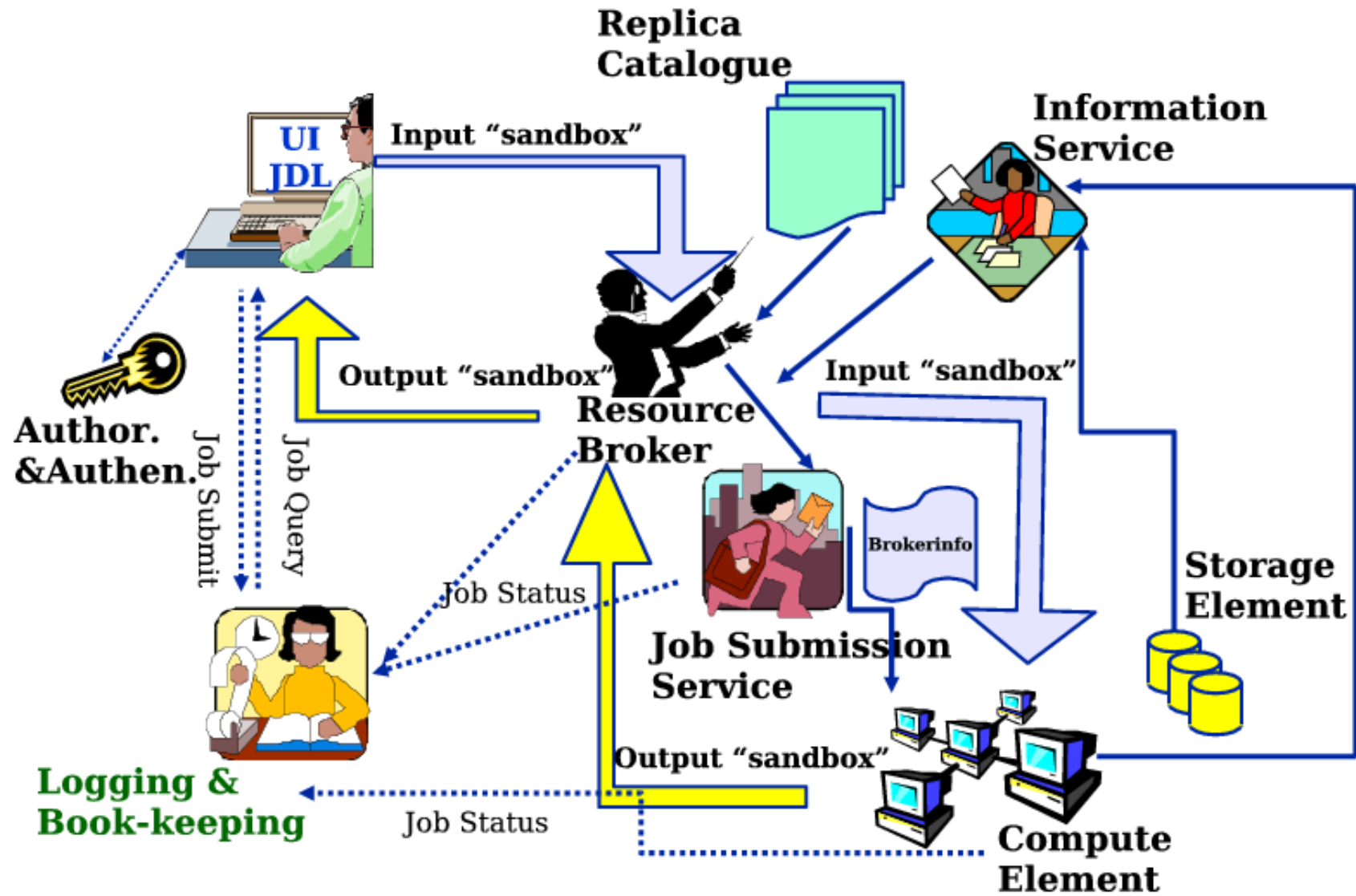
- **Job abbrechen**

```
edg-job-cancel https://grid-rb.desy.de:9000/WLB3aGGdSNUFhgCeIbcJhg
```

Job/Workload Management – Komplexere Beispiele

- weitere JDL Attribute
 - InputSandbox: Dateien dem Job mitgeben (wenige MB)
 - Requirements: z.B. installierte s/w, CPU-Zeit, . . . (Attribute aus GLUE Schema)
 - Rank: wie CEs bevorzugen, die Requirements erfüllen
 - InputData: welche Daten auf “nahem” SE liegen sollen
- nicht-Standard-Software benutzen
 - offizielle s/w der VO auf WNs installieren
 - per InputSandbox mitschicken
 - wie Datenfile betrachten und am Jobanfang vom SE kopieren

Ablaufschema



Data Management – Begriffe

Jede Datei auf einem Storage Element hat

- einen Globally Unique Identifier (GUID)
 - kennzeichnet Datei im Grid eindeutig
 - z.B.: `guid:8fa51ad0-f973-4afa-9607-905f6ee6761b`
- ein oder mehrere Logical Filenames (LFNs)
 - menschenlesbares Alias, mehrere für das gleiche File möglich
 - z.B.: `lfn:mcevents_pythia_ztautau.ntpl.CMKIN_4_3_1.5000Events`
- ein oder mehrere Physical Filenames (PFNs)
 - ein PFN pro Replica
 - z.B.: `sfn://accip41.physik.rwth-aachen.de/storage/dcms/generated/\n2005-03-11/file92679a17-5773-4b6e-9d7b-19312fcc02e4`

Data Management – Kommandos

- Datei auf SE kopieren und im RLS registrieren

```
$ lcg-cr -d grid-se.desy.de --vo dcms \  
-l lfn:hello_world.jdl \  
file:$PWD/hello_world.jdl  
guid:\  
cf03754d-d092-497e-8ecc-a6213b758384
```

- Name der Replica

```
$ lcg-lr --vo dcms lfn:hello_world.jdl  
srm://grid-se.desy.de/pnfs/desy.de/data/\  
dcms/generated/2005-05-10/\  
file5d75ec1c-d5dc-4536-848d-20aec5399899
```

- Auf anderes SE replizieren

```
$ lcg-rep --vo dcms \  
-d ekp-lcg-se.physik.uni-karlsruhe.de \  
lfn:hello_world.jdl
```

- Neuen LFN hinzufügen

```
$ lcg-aa --vo dcms \  
guid:cf03754d-d092-497e-8ecc-a6213b758384  
lfn:doofer_name
```

- LNFs zur GUID finden

```
$ lcg-la --vo dcms \  
guid:cf03754d-d092-497e-8ecc-a6213b758384  
lfn:doofer_name  
lfn:hello_world.jdl
```

- Alle Replicas finden

```
$ lcg-lr --vo dcms lfn:doofer_name  
sfn://ekp-lcg-se.physik.uni-karlsruhe.de/...  
srm://grid-se.desy.de/...
```

Am Ende alle Test-Dateien löschen:

```
$ lcg-del -a --vo dcms lfn:doofer_name
```

Wo/wie kann ich das alles tun? → User Interface (UI)

Zentraler Punkt: das **User Interface**

- **Einstiegspunkt ins Grid**

- Nutzer loggen sich per ssh ein
- alle wichtigen Grid-Kommandos vorhanden
- prinzipiell jeder Rechner kann UI sein

- **nötige Software**

- globus, edg, lcg, . . .
- \approx 400MB
- als RPMs für RedHat und ScientificLinux verfügbar
- muß anschließend an Site angepasst werden
- Informationen über CAs müssen ständig aktualisiert werden

nicht vergessen: brauche **Grid-Zertifikat** und Mitgliedschaft in einer **VO!**

Probleme/Limitationen

- VO Mitgliedschaft
 - derzeit nur Mitgliedschaft in einer VO möglich
 - wird in Zukunft gelöst durch Virtual Organization Membership Service (VOMS)
 - alle Mitglieder einer VO werden auf WNs/SEs auf dieselbe Unix ID abgebildet
- Information System
 - Resource Broker starker Belastung ausgesetzt
 - Information System propagiert Änderungen erst nach ≈ 2 Minuten
- verfügbare Tools
 - teilweise sehr low-level \rightarrow viel “per Hand”
 - Diagnosemöglichkeiten teilweise beschränkt/Fehlermeldungen nichtssagend

Nachtrag für "Hacker"

- Zertifikat mit OpenSSL betrachten

```
openssl x509 -text -in ~/.globus/usercert.pem
```

- LDAP Abfragen

- Mitglieder einer VO

```
ldapsearch -x -H ldap://<VO-Server> -b 'ou=<VO>,ou=vo,o=<ORG>,c=de'
```

- local GRIS auf SE oder CE

```
ldapsearch -x -H ldap://<CE/SE-Host>:2135 -b 'Mds-Vo-name=local,o=grid'
```

```
ldapsearch -x -H ldap://<CE-Host>:2135 -b 'Mds-Vo-name=local,o=grid' \  
objectclass=GlueCluster
```

```
ldapsearch -x -H ldap://<CE-Host>:2135 -b 'Mds-Vo-name=local,o=grid' \  
objectclass=GlueCE
```

- site GIS (typisch auf CE)

```
ldapsearch -x -H ldap://<CE/SE-Host>:2135 -b 'Mds-Vo-name=<Site>,o=grid'
```

Weitere Informationsquellen

- I. Foster, C. Kesselman (1999):
The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann Publisher Inc.
- I. Foster, C. Kesselman, S. Tuecke (2000):
[The Anatomy of the Grid: Enabling Scalable Virtual Organizations](#)
- I. Foster, C. Kesselman, J. Nick, S. Tuecke (2002):
[The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration](#)
- H. Kornmayer im Linux-Magazin 06/2004:
[Vernetzte Welten - Das Globus-Toolkit, Version 2](#)
- A. Meyer in c't 22/2004:
[Nützliche Abfälle - Vom WWW zum Grid](#)

- [EGEE: Introduction to Grid Computing with the LCG-2 middleware](#), Grid-Kurs am DESY
- [LCG-2 User Guide](#), sehr empfehlenswertes Handbuch
- [Global Grid User Support](#), zentrale Anlaufstelle für Probleme
- [LCG-2 Monitoring Map](#), Karte der Sites in LCG-2