RECHENZENTRUM der UNIVERSITÄT MANNHEIM

INSTITUT FÜR INFORMATIK

Programmierung und Leistungsanalyse des bwGRiD-Clusters

Dr. Heinz Kredel und Dr. Hans-Günther Kruse

Inhalt

- bwGRiD Cluster
- Programmierung
  - Threads, OpenMP,
  - Kommunikation, OpenMPI, Strategien
- Leistungsanalyse
  - Gesetze von Amdahl und Gustafson
  - Skalierungsgesetz, Speed-Up

Termine (1)

18.2.: Vorbesprechung

4.3.: Einleitung bwGRiD Cluster

11.3.: Leistung 1: Allgemeine Einführung

18.3.: Leistung 2: Amdahl und Gustafson

25.3.: Thread Programmierung

1.4.: OpenMP Programmierung

8.4.: Leistung 3: Skalierungsgesetze

15.4.: Leistung 4: Speed-Up bei Clustern

Termine (2)

22.4.: Verteilte Programmierung

29.4.: OpenMPI Programmierung

6.5.: Leistung 5: Kommunikation

13.5.: Strategien zur Parallelisierung

20.5.: Hybrid-Programmierung

27.5.: letzte Vorlesung

3.6.: Pfingstwoche

Grid

The Grid is a system that:

- coordinates resources that are not subject to centralized control …
- … using standard, open, general-purpose protocols and interfaces …
- … to deliver nontrivial qualities of service.

Ian Foster, 2002

War MACH ein Grid ?

- Rechnerkopplung Mannheim - Heidelberg
- ca. 1977 - 1985
- synchrone Modem Standleitung mit 64Kbit
- zwischen Siemens und IBM Mainframes
- dezentral, ok
- offene Standards, BC, X-25 Protokoll ?
- nicht triviale Dienste, ok
  - Plotausgabe, Filetransfer
  - Remote Job Execution

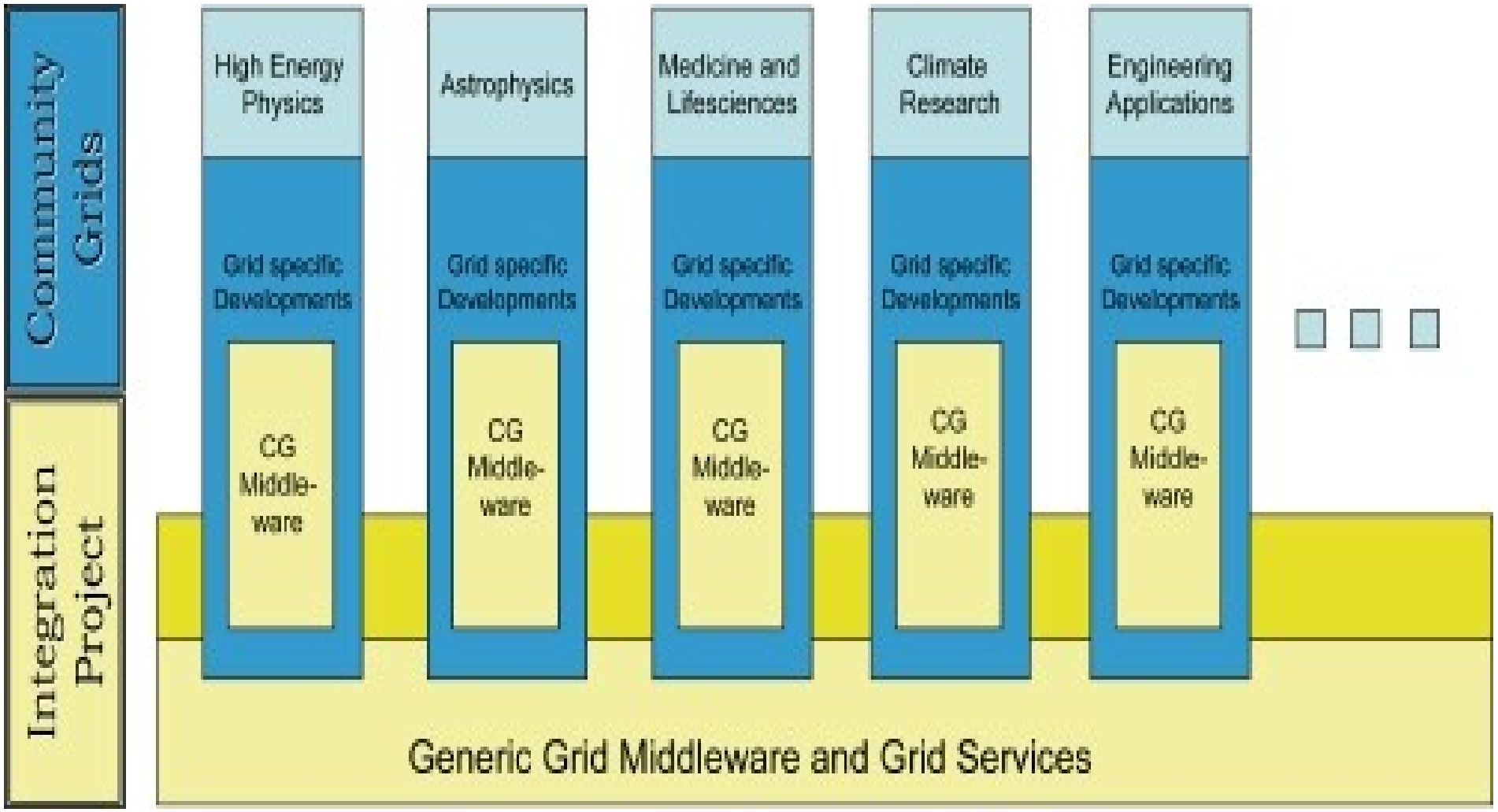
D-Grid

Das Grid

- ermöglicht den direkten Zugriff auf Ressourcen, wie Rechner, Speicher, wissenschaftliche Instrumente und Experimente, Anwendungen und Daten, Sensoren und sogenannte Grid-Middleware Dienste
- basierend auf weit verbreiteten Grid- und Web-Services-Standards.

D-Grid Web-Seite, 2008

D-Grid Projekte



D-Grid Ressourcen

Dynamisch erstellte Karte von Ressourcenanbietern / Wartungszustand

bw-Grid Cluster

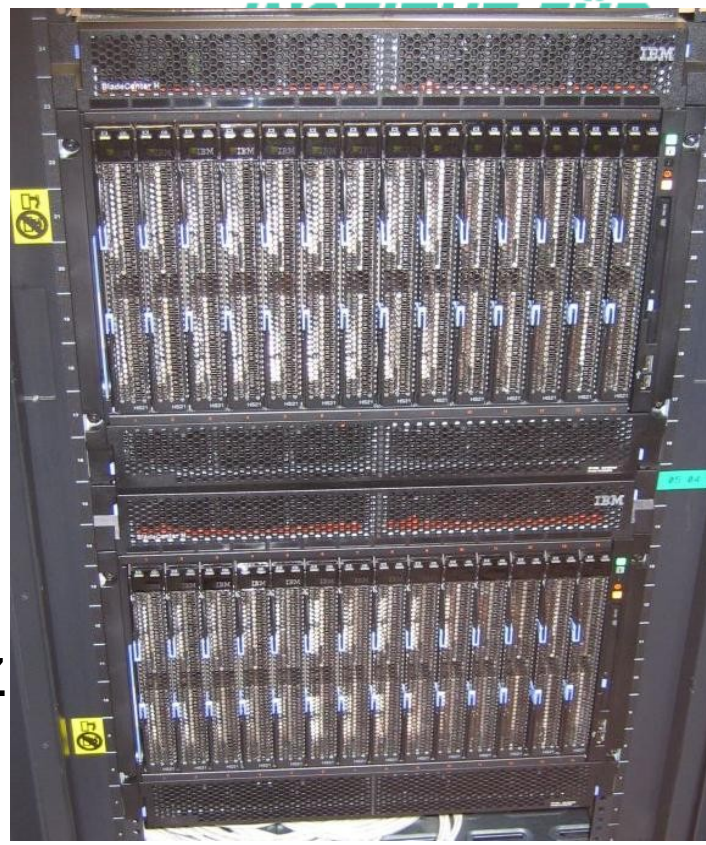
- Projektantrag vom HLRS an BMFT in 2007
- für D-Grid Infrastruktur an den Universitäten in Baden-Württemberg
- explizit als verteiltes System
- mit dezentraler Verwaltung
- an den Standorten
  - Stuttgart, Ulm (mit Konstanz), Freiburg, Tübingen, Karlsruhe, Heidelberg, Mannheim
- für die nächsten 5 Jahre

bw-Grid Ziele

- Nachweis der Funktionalität und des Nutzens von Gridkonzepten im HPC Umfeld
- Überwindung von bestehenden Organisations- und Sicherheitsproblemen
- Entwicklung von neuen Cluster- und Grid-Anwendungen
- Lösung der Lizenzproblematik
- Ermöglichung der Spezialisierung von Rechenzentren

RECHENZENTRUM der UNIVERSITÄT MANNHEIM

Hardware

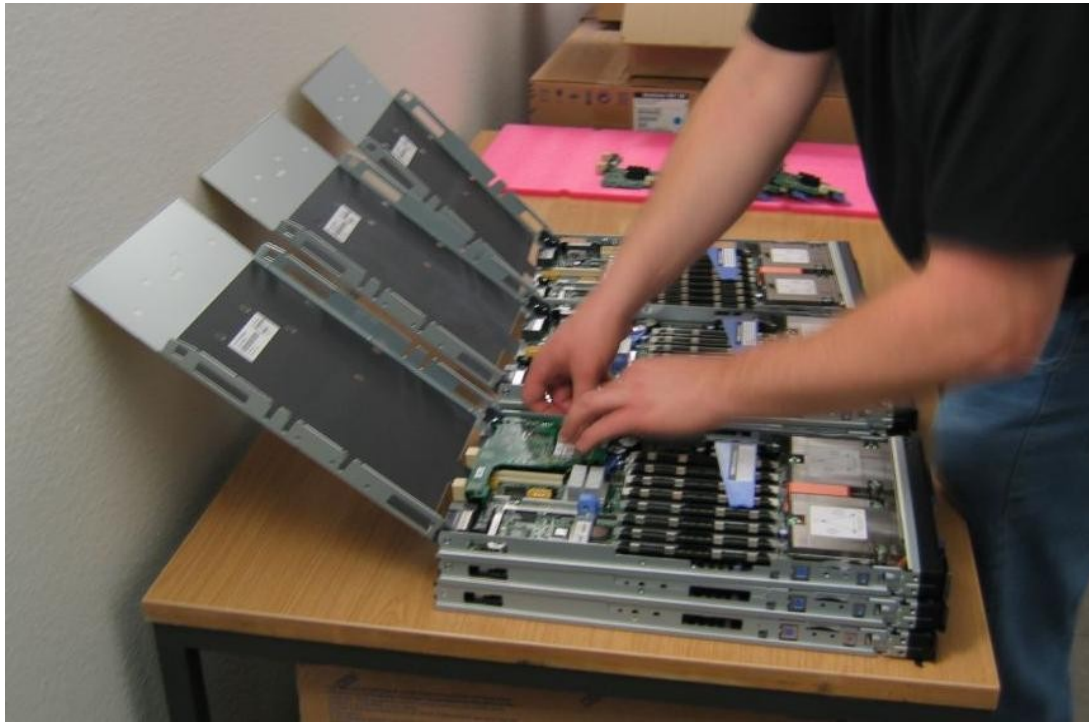


- 101 IBM Blade-Center Gehäuse
- mit je 14 IBM HS21 XM Blades
  - mit je 2 Intel Xeon CPUs, 2.8 GHz
    - mit je 4 Cores
  - 16 GB Hauptspeicher
  - keine Festplatten
  - Gigabit Ethernet
  - Infiniband Netzwerk

= 1414 Blades

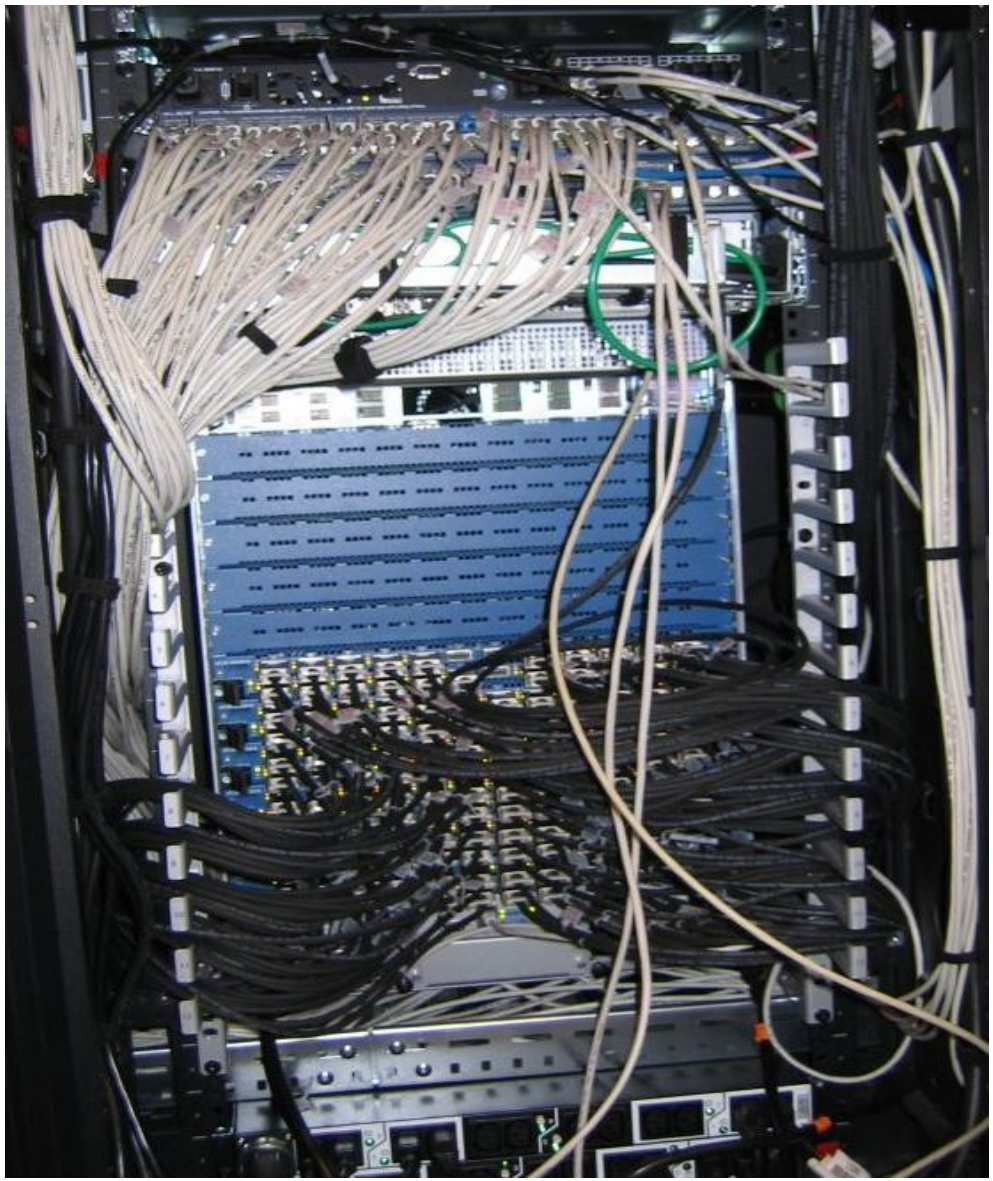
= 11312 CPU Cores,

1120 Cores in Mannheim



Racks mit Blade Center

Netz



Software

- Scientific Linux 5.0
  - basierend auf RedHat 5.0 Quellcode
  - gepflegt u.A. vom CERN
    - dort sind ca. 140.000 Rechner für LHC im Einsatz
  - ähnlich CentOS, wie bei Mailbox-Servern
- booten über das Netz mit DHCP und TFTP
- Festplatten über NFS
- minimaler Swap-Bereich über NBD
- Kernel und System Image vom HLRS vorbereitet

Anwendungen

	Mathematik	Biowissenschaften	Ingenieurwiss.	Wirtschaftswiss.	Chemie	Physik	Informatik
Freiburg	25% CFD		20% Mikrosystemt.				
Heidelberg	25% CAS, Ug	20% Comp. NeuroSc.					
Karlsruhe	10 % LinAlg		30% CFD, CSM				
Konstanz	x SPH	x Biochemie				x Theor. Ph, QM	x DataMining
Mannheim	15% CAS			30% CAS, Simulation			
Stuttgart			35% CFD, Comp.Mech.				
Tübingen		20% BioInfo				25% Astrophysik	
Ulm		5% BioInfo, Medizin			25% Theor. Ch, MD		

Alle: ca. 55% Betrieb, Middleware, Compiler, Tools

CFD = Computational Fuid Dynamics
CAS = Computer Algebra Systems
MD = Molecular Dynamics
QM = Quantum Mechanics

Stand des Aufbaus

- Hardware
  - geliefert und aufgebaut, Januar bis März
  - Schwierigkeiten die Racks mit dem Fahrstuhl ins 11. OG zu transportieren
    - Demontage notwendig
  - funktioniert nach Austausch defekter Komponenten
    - Infiniband HCA defekt
    - Ethernet Switch in Bladecenter defekt
    - lose Kabel und Module ein-/ausgesteckt
  - Luft-Kühlung über Klimaanlage funktioniert
  - Stromanschlüsse werden noch überarbeitet

Stand der Installation

- Software
  - Zugangsknoten sind installiert
  - Betriebssystem ist installiert und läuft auf Blades
  - Netzwerk funktioniert, Ethernet und auch Infiniband
  - Festplatten Platz für Home-Dirs über NFS-Server
  - Benutzerverwaltung ist konfiguriert, über LDAP
  - Batchsystem ist installiert und teilweise konfiguriert
- Zugang:

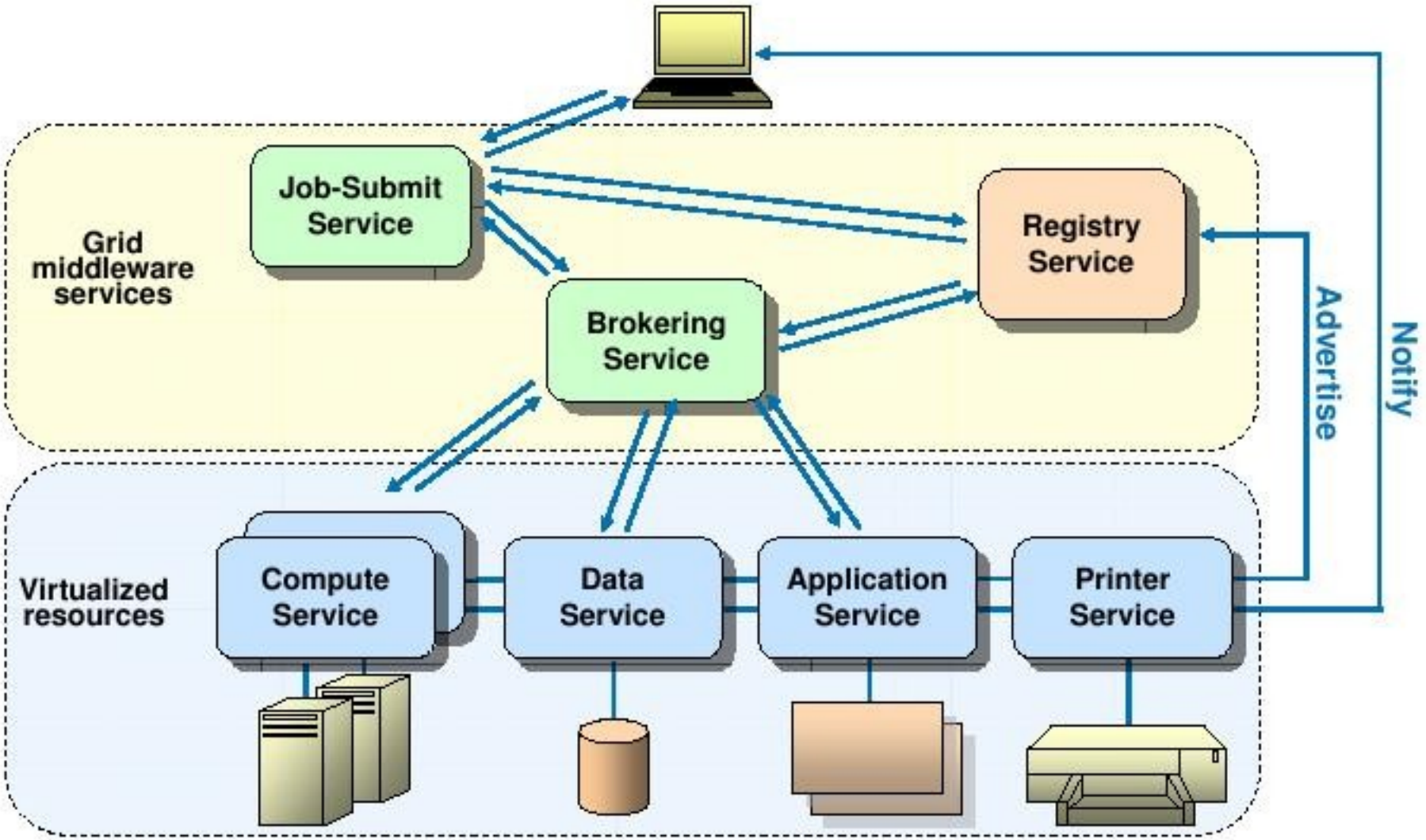
```
ssh -XY user@grid.uni-mannheim.de
ssh n010301
```

Benutzung

- Batchsystem
  - qsub shell-script
  - qstat
  - nstat
- PBS Optionen
  - #PBS -l nodes=30
  - #PBS -q normal
- Ergebnis in home
  - shell-script.o113
  - shell-script.e113

```
[kredel@zug-ma-2 ~]$ nstat

        node/job status on IBM bw-GRID      v1.4

     ID      user      name  nodes state  start [h]  remain [h]
-------------------------------------------------------------------
    113    kredel   versuch     10     R      -0.03       11.97
    114    kredel   versuch     30     R      -0.01       11.99
-------------------------------------------------------------------

  waiting jobs: 0,  requested nodes waiting for running: 0

   Node |      +0      +1      +2      +3      +4      +5      +6
-------------------------------------------------------------------
n010301 |       -       -       -       -       -       -       -
n010308 |       -       -       -       -       -       -       -
n010401 |       -       -       -       -       -       -       -
n010408 |       -       -       -       -       -       -       -
n020301 |       -       -       -       -       -       -       -
n020308 |       -       -       -       -       -       F       F
n020401 |       F       F       F       F       F       F       F
n020408 |       F       F       F       F       F       F       F
n030101 |       F       F       F       F       F       F       F
n030108 |       F       F       F       F       F       F       F
n030201 |       F       F       F       F       F       F       F
n030208 |       F       F       F       F       F       F       F
n040301 |       F       F       F       F       F       F       F
n040308 |       F       F       F       F       F       F       F
n040401 |       F       F       F       F       F       F       F
n040408 |       F       F       F       F       F       F       F
n050301 |       F       F       F       F       F       F       F
n050308 |       F       F       F       F       F       F       F
n050401 |       F       F       F       F       F       F       F
n050408 |       F       F       F       F       F       F       F

    F = free    R = reserved    . = empty    - = offline
    H = Head    I = I/O node                 + = denied

      0/140 nodes reserved  (0.0%)
      0/140 nodes used      (0.0%)
      0/140 nodes offline   (0.0%)
    100/140 nodes free      (71.4%)
```

Middleware und Ressourcen



Aufgaben - Middleware

- Konfiguration des Batchsystems
  - mehrere Queues mit verschiedenen Qualitäten
- Installation der Grid-Middleware
  - Globus-Toolkit, DoE-, NSF-Projekt
  - UNICORE, EU-Projekt
  - LCG/gLite, LHC Computing Projekt
- Konfiguration der Benutzer-Zugangsrechner
- Schulung und Information der Benutzer

Virtuelle Organisationen

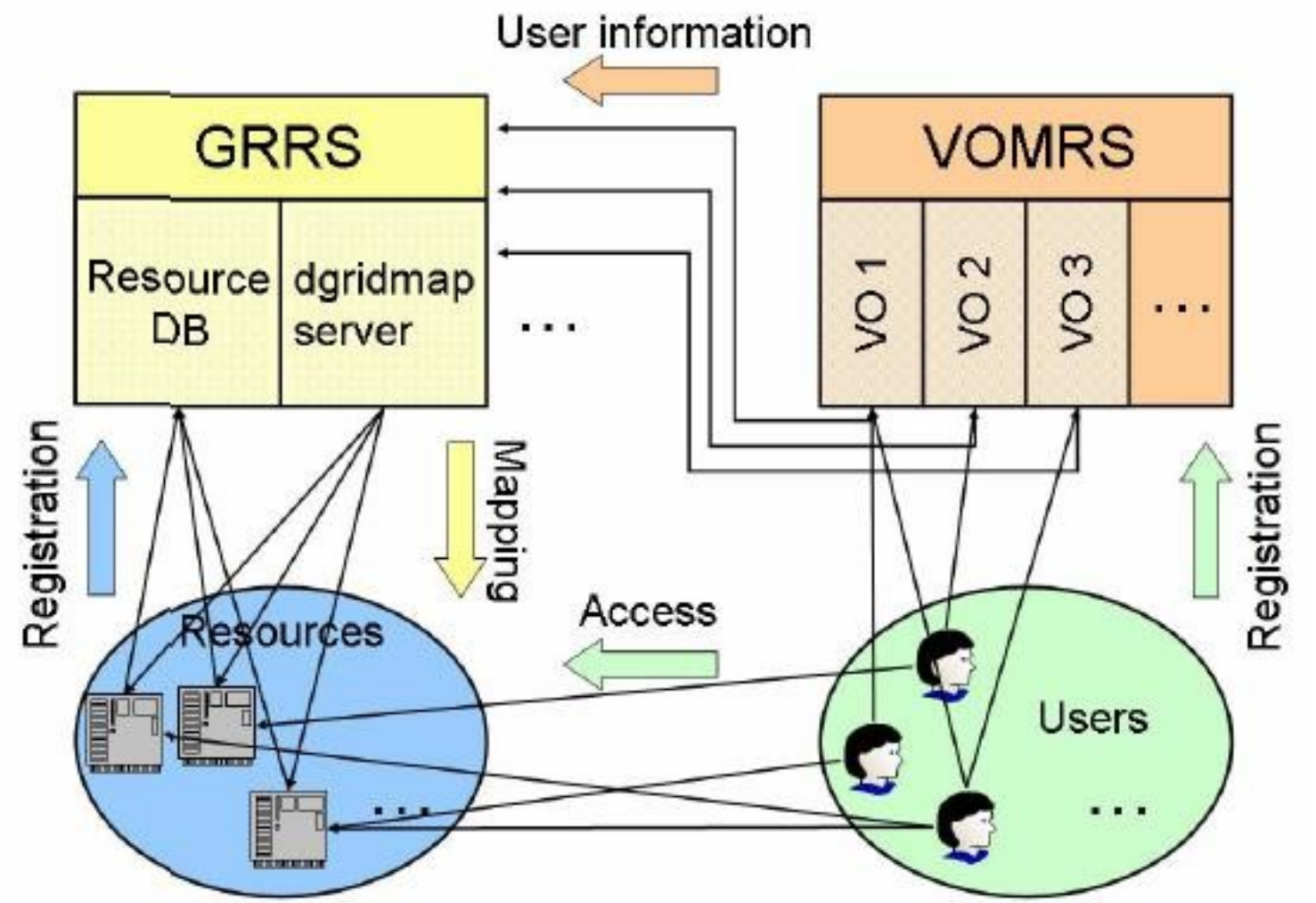


Abbildung 2 Architekturbild Betrieb der VO- und Nutzerdienste und der Ressourcen

Aufgaben - Integration

- D-Grid Virtuelle Organisation (VO) aufsetzen
- Mitglied in VO werden, Grid-Zertifikate
- D-Grid Ressourcen aufsetzen und registrieren
- D-Grid Benutzerverwaltung
- Anbindung an zentralen Storage von bw-Grid
  - geplantes Projekt ab 2008 über HBFG Antrag

Aufgaben - Software

- Cluster Programmierungs Tools
  - MPI
  - OpenMP
  - Intel Compiler
- Installation der Anwendersoftware
  - Matlab
  - Maple
  - Mathematica (Grid)
  - R
  - Octave

Vor- und Nachteile

- ✗ Lernaufwand für Entwicklung paralleler Programme und Batch-System
- ✗ Lernaufwand für die Grid-Middleware und Einrichtung der Grid-PKI Anforderungen
- ✓ State-of-the-art Plattform für Mannheim
- ✓ Zugang zu (fast) allen D-Grid-Ressourcen
- ✓ Auslastung unseres Clusters wird höher
- – seine Programme muss jede(r) immer noch selbst parallelisieren
- – Lizenzen für kommerzielle Software können recht teuer werden

Vielen Dank für die Aufmerksamkeit

- es besteht die Möglichkeit zu einer Führung zum bw-Grid Cluster
- weitere Veranstaltungen folgen
- Links
  - http://www.d-grid.de/
  - http://www.hlrs.de/
  - http://www.unicore.eu/, http://www.globus.org/
  - http://lcg.web.cern.ch/lcg/
  - https://www.scientificlinux.org/