BIRD:
Batch Infrastructure Resource at DESY

A Grid Engine System

Thomas Finnern (DESY/IT Systems and Operations)

Batch Infrastructure Resource at DESY

HEPiX Fall 2010, Ithaca, NY
November 1st - 5th, 2010 @ Cornell University

HELMHOLTZ
ASSOCIATION

HEPiX

CORNELL UNIVERSITY FOUNDED A.D. 1865

DESY

The Team and the Mission

> The Team

- Christoph Beyer
- Thomas Finnern
- Martin Flemming
- Frank Schlünzen
- Jan Westendorf
- Knut Woller

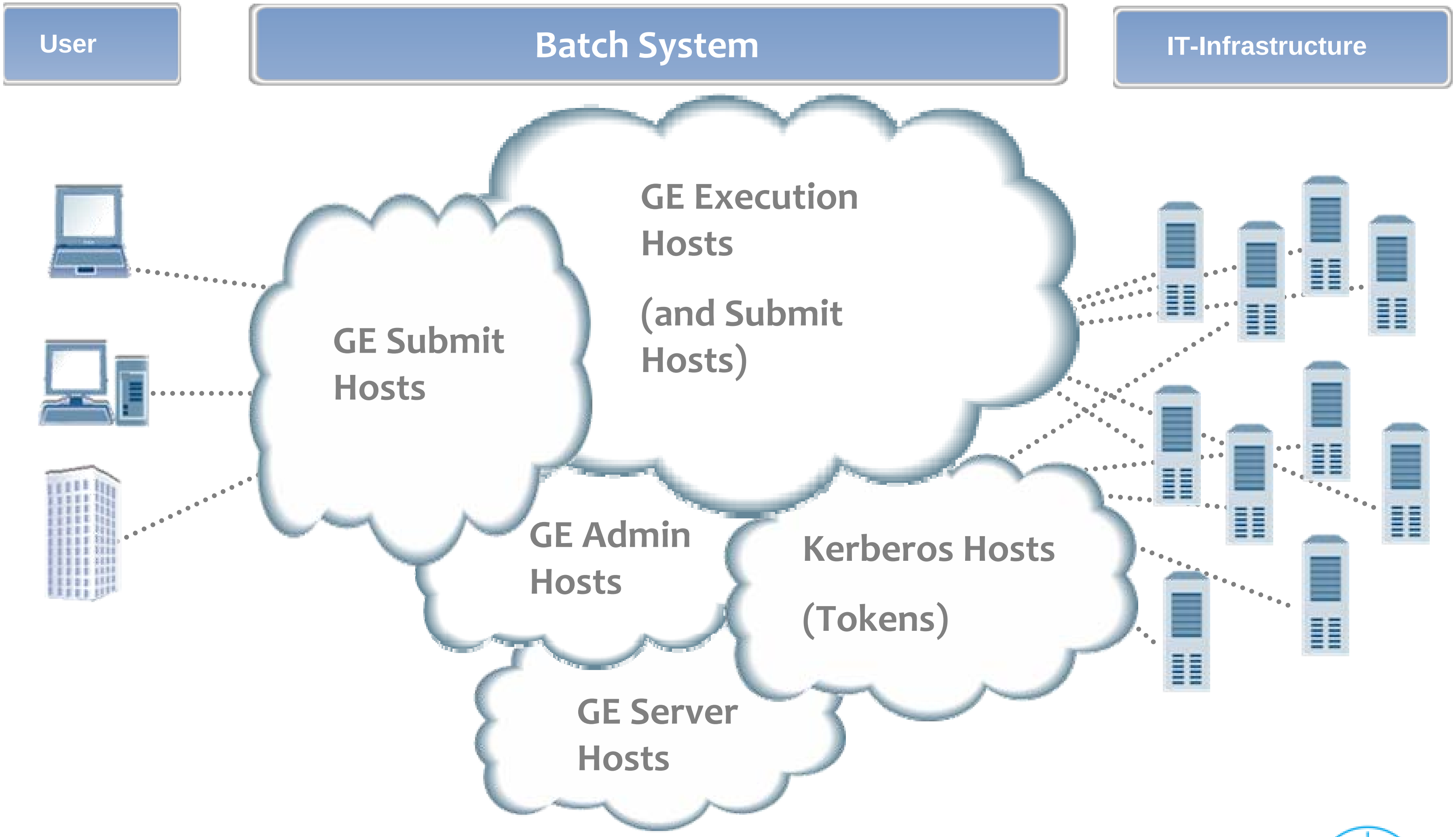
> Integration of DESY Wide Batch Resources

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

Base Features

- Runs on Grid Engine (GE) Version 6
  - AFS as shared file system
  - Availability through Master and Slave Server
- 500 CPU Cores for Interactive and Batch Processing
- 8-64 GByte Memory + 32-250 GB Scratch Disk per Host
- sld4, sld5 OS in 32/64 Bit with Minimum 2GByte Memory / Core
- More than 250 Users from 25 Different Groups/Projects
- Group Specific Software and Storage (in AFS, dcache, …)
- Submit and Control Facilities from PAL, BIRD and Group Specific Hosts
- Everybody at DESY may become BIRD User
  - Access by Setting Registry Resource Batch
  - Delegated to UCO / Groups Admins
  - Active in one Hour (GE, Netgroup, Mailing List)
- "*Select your Resources and we define the Queue*"

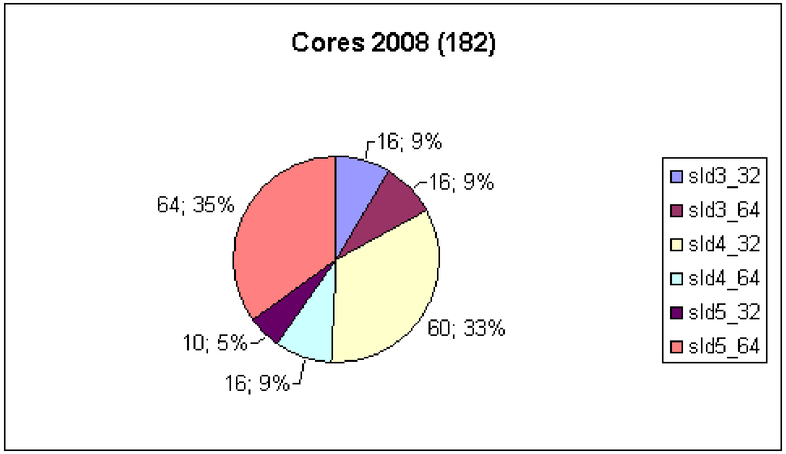
The Topology

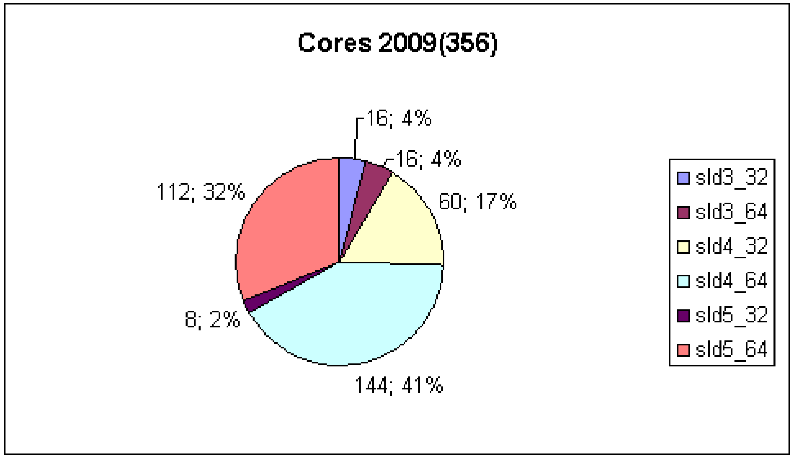


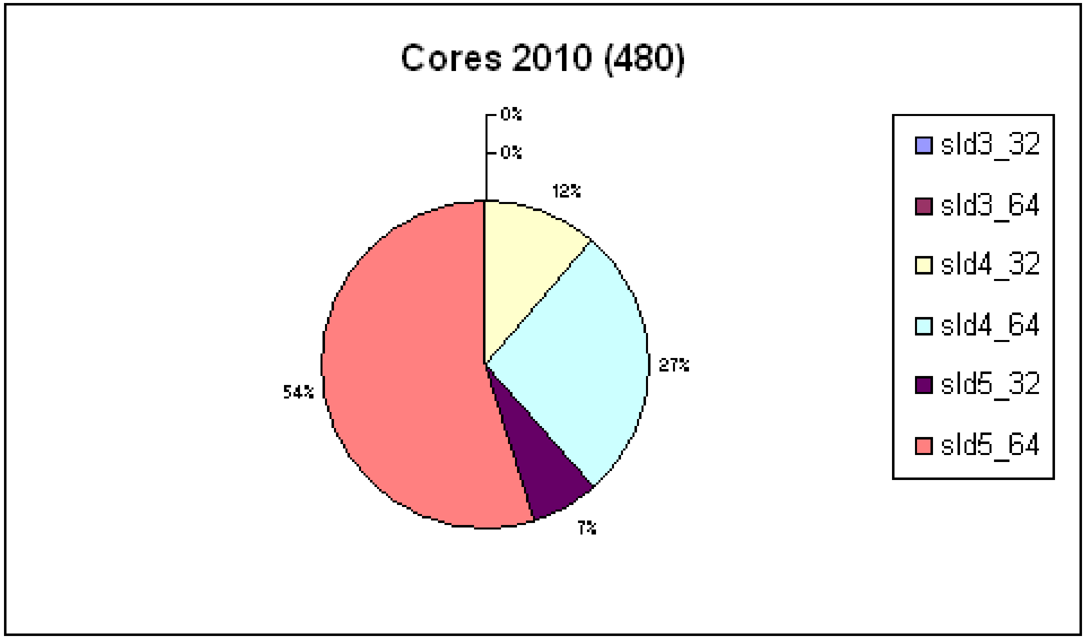
Queues (selected on your Resource Demands …)

Queue	Time Limit	Slots	Comment
default.q	3 hours	100 % (@anyhost)	available as default (h_rt < 3:00:00, h_vmem < 2G)
short.q	1 day	85 % (/platform) (incl. 65 % for long.q)	available for medium sized jobs (h_rt < 24:00:00, h_vmem < 2G)
long.q	1 week		available for long runner and high memory usage (24:00:00 < h_rt < 168:00:00, h_vmem < 4G)
login.q	1 day	50 % (@anyhost)	For preparing jobs interactively in a qterm and/or qshell

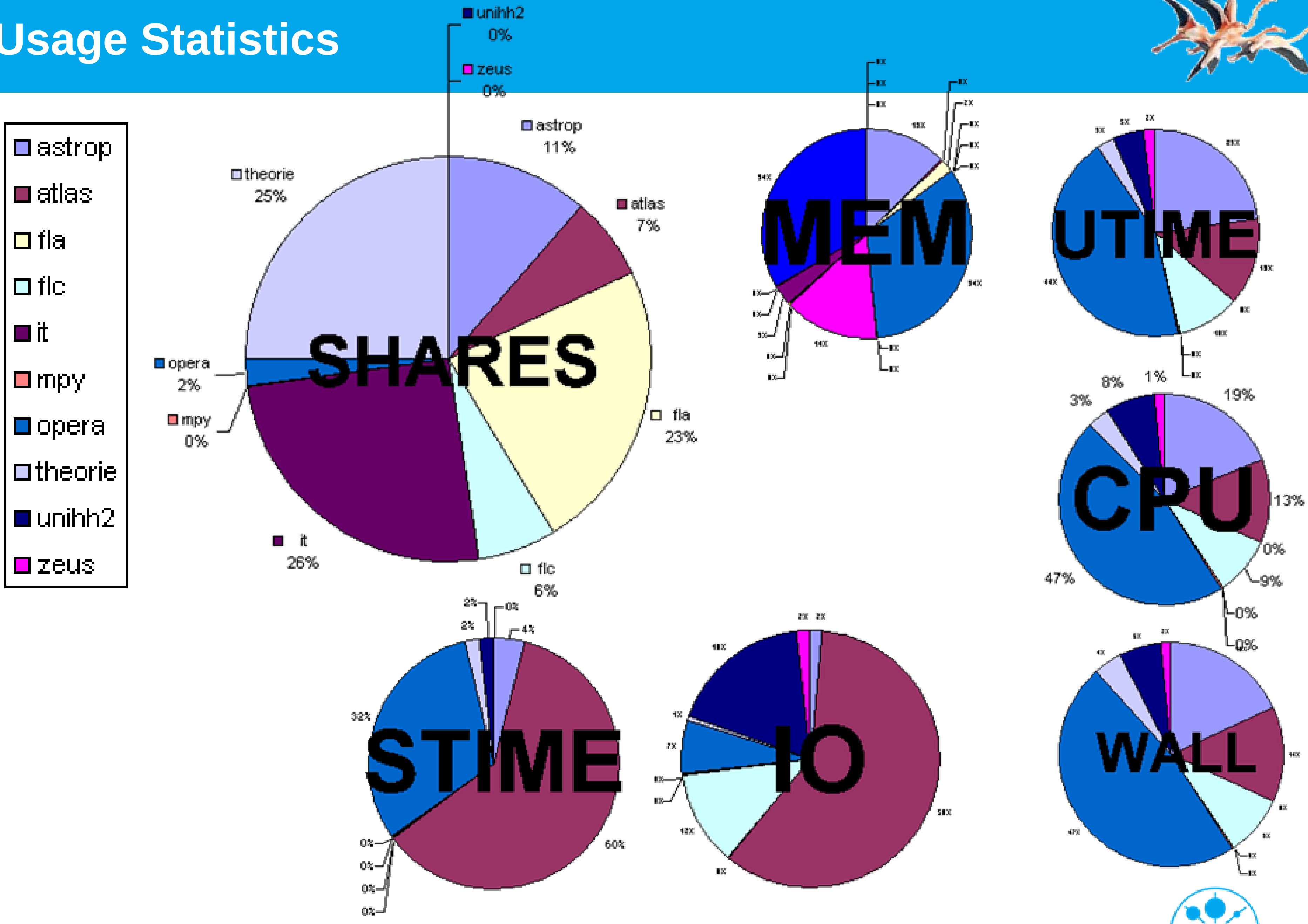
Core Statistics


Cores 2008 (182)


Cores 2009(356)


Cores 2010 (480)

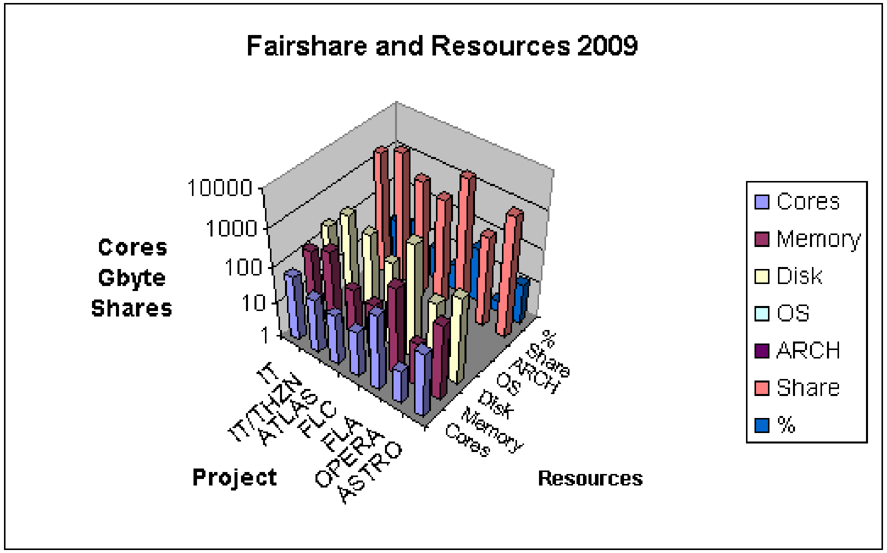
Usage Statistics

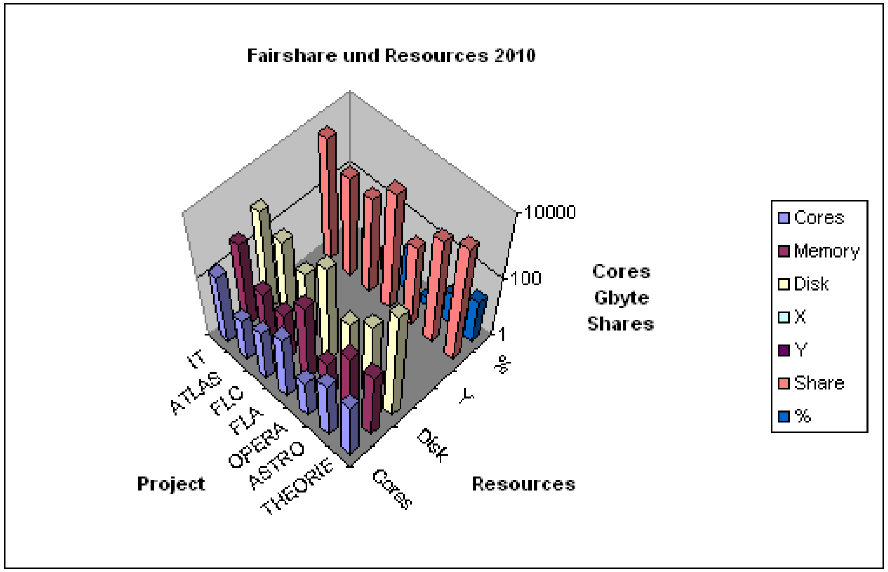


Advanced Features

- AFS and Kerberos Support for Authentication and Resource Access
  - Valid Tokens during Complete Job Execution
  - Take Cluster
- Cores, h_rt, h_vmem, h_fsize Under Full Control of Scheduler
  - *“Select your Resources and we guarantee for it”*
- MPICH2 Parallel Environments
  - mpich2-1 (Single Host)
  - mpich2 (Multi Host …)
    - Enhanced by changing from ssh to GE internal interconnect
- 4 Big Birds for High Resource Demands
  - 8 Cores / Host
  - 64 Gbytes Memory / Host
  - 250 GByte Scratch / Host
  - Modified Queue Settings: e.g. 32 GByte Memory / Job
- Fair Share Load Distribution and Quota Handling

Shares


Fairshare and Resources 2009
Cores
Gbyte
Shares
10000
1000
100
10
1
IT
IT/THZN
ATLAS
FLC
FLA
OPERA
ASTRO
Project
Cores
Memory
Disk
OS
ARCH
Share
%
Resources
Cores
Memory
Disk
OS
ARCH
Share
%


Fairshare und Resources 2010
10000
100
1
Cores
Gbyte
Shares
IT
ATLAS
FLC
FLA
OPERA
ASTRO
THEORIE
Project
Cores
Disk
Y
%
Resources
Cores
Memory
Disk
X
Y
Share
%

Usage Policy

- Lightweight Resources should be a 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

Quotas

- Quotas are set up for each OS Flavor separately
  - As some Projects depend on one OS
- People Related Quota Settings
  - A Single User must not use more than 65 % of the Cores of an OS Flavor
  - Projects are limited to 75 % of a core set
- Queue Related Quota Settings
  - Allowing 100 % Jobs in the Default Queue (ensures Scheduling at Least Every 3 Hours
  - Longer Queues (long, short and long+short) are limited to 65, 75 and 85 % respectively.
  - The Interactive Login Queue is limited to 50 %

Fairshare

- IT Hardware + Project Specific Hardware = Valuable Resources
- Contributing projects will be granted Fairshare Points
  - 10 for each Compute Core,
  - 10 for each 2 GByte Memory
  - 1 for each GByte Disk
- Guaranteed Access to this Relative Amount of Batch Resources to the Project Members
- IT gives own Share to the Community
- Batch Resource Requirements are not continuous over Time
- 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*)

User Information

- http://bird.desy.de/info
- BIRD-Flyer:

HOW TO USE THE BIRD CLUSTER.

Batch computing at DESY

Accelerators | Photon Science | Particle Physics

Deutsches Elektronen-Synchrotron
A Research Centre of the Helmholtz Association

Spotlights

> Updated Kerberos Support

- (Internal) Ticket Prolongation over 2 Weeks
- Kerberos 5 Setup including afs tokens
- Take Cluster for redundant Token Creation
- Take Cluster needs heimdal

> Oracle Grid Engine

- Suns „Red Carpet“ removed by Oracle
- No offers to science and research communities anymore
- Using (last „free“) Gridengine Version 6 or
- "Open Grid Scheduler" forked from 6.2u5:

  http://gridscheduler.sourceforge.net/

> Interactive Batch

- GE Internal Connect for Interactive Environments
- ssh only allowed for admins by pam config
- Qterm and Qshell

Plans and Ideas

- Virtualized Worker Nodes (?)
  - OS Version as a Dynamic Resource
  - Multicore ?
  - Big Memory ?
- Shared Filesystem (?)
  - Fraunhofer File System

    http://bird.desy.de/fhgfs (internal)
- HPC (?)
  - Local Request for High Speed Parallel Environments
- Dynamic Resource Management Application API DRMAA (?)
  - Used e.g. by Mathematica
  - Not Compliant with AFS/Kerberos Version of GE

Last Slide

- Thank you for your attention