

Batch Infrastructure Resource at DESY

 A Sun N1 Grid Engine System



Thomas Finnern (DESY/IT)

Batch Infrastructure Resource at DESY

LUM, 06/17/2009



> 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*

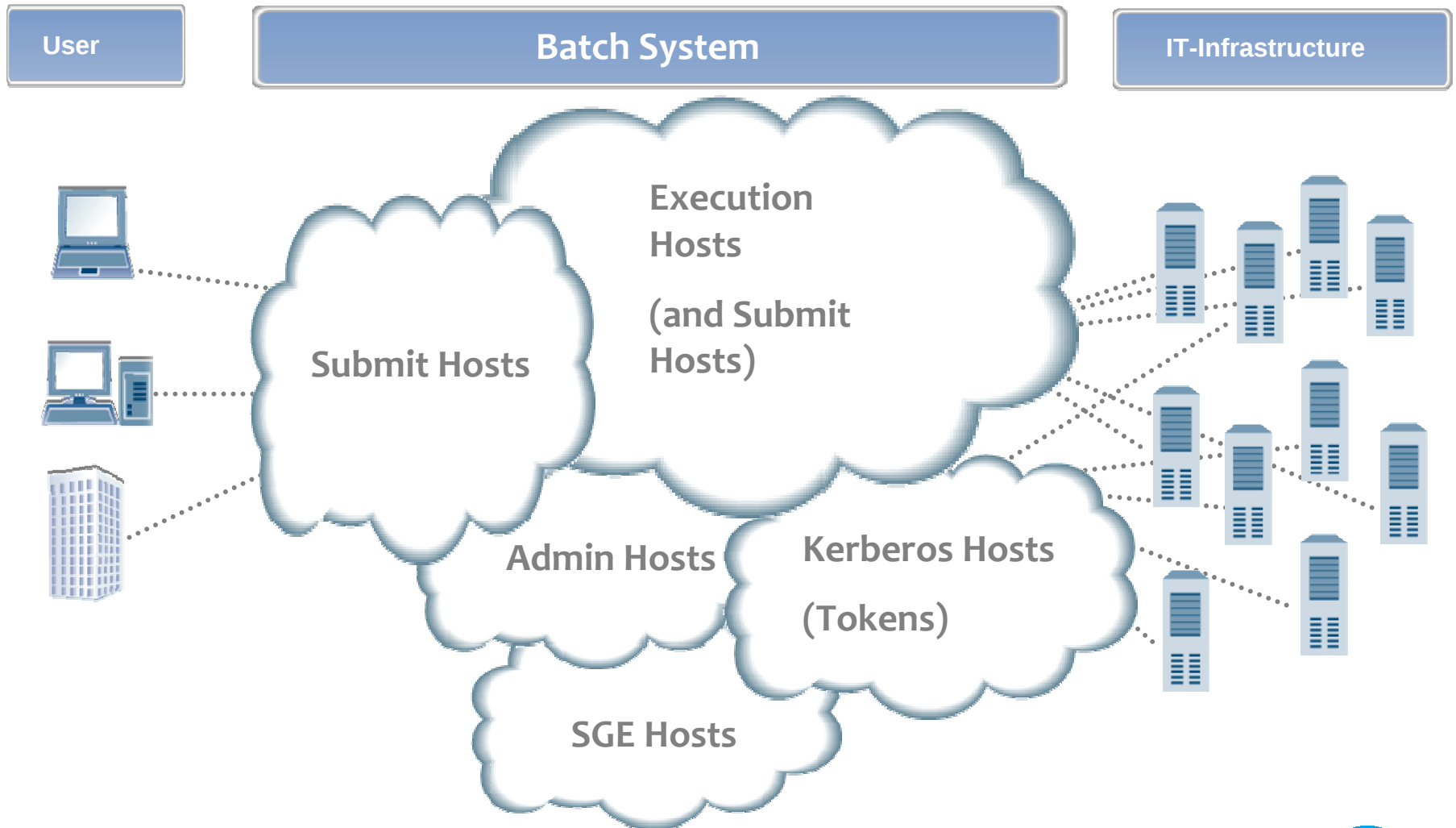




- > Runs on Sun Project N1 Grid Engine Version 6 (N1GE)
- > More than 350 CPU Cores for Interactive and Batch Processing
- > 8-64 GByte Memory + 32-250 GB Scratch Disk per Host
- > sld3, sld4, sld5 OS in 32/64 Bit with Minimum 2GByte Memory / Core
- > More than 180 Users from 20 Different Groups/Projects
- > Group Specific Software and Storage (in AFS, dcache, ...)
- > Submit and Control Facilities from PAL, BIRD and Group Specific Hosts
- > Select your Resources and we define the Queue



The Topology



Queues (selected on your Resource Demands ...)

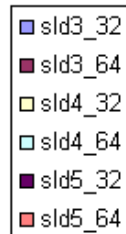
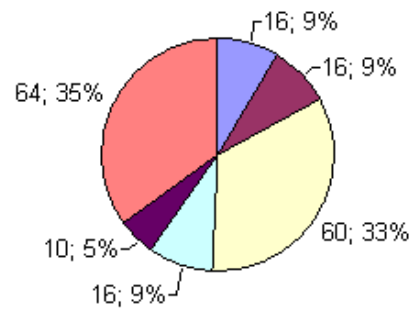


Queue	Time Limit	Slots	Comment
default.q	3 hours	100 % (@anyhost)	available as default (h_rt < 3:00:00, h_vmem < 2G)
short.q	1 day	appr. 85 % (/platform) (incl. 50 % for long.q)	available for medium sized jobs (h_rt < 24:00:00, h_vmem < 2G)
long.q	1 week		available for long runner and high memory usage (24:00:00 < h_rt < 168:00:00, h_vmem < 4G)
login.q	1 day	4 (@somehost)	under evaluation, useful for debugging, startup may be difficult (see advice)
idle.q	3 weeks	1 (@anyhost)	under evaluation, jobs will be suspended while other jobs request same host (h_rt < 504:00:00 , s_rt < 503:00:00) [job is "warned" via the SIGUSR1]

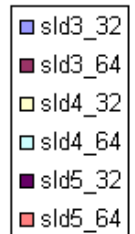
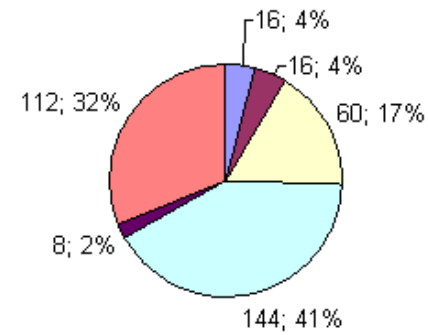




Cores 2008 (182)



Cores 2009(356)



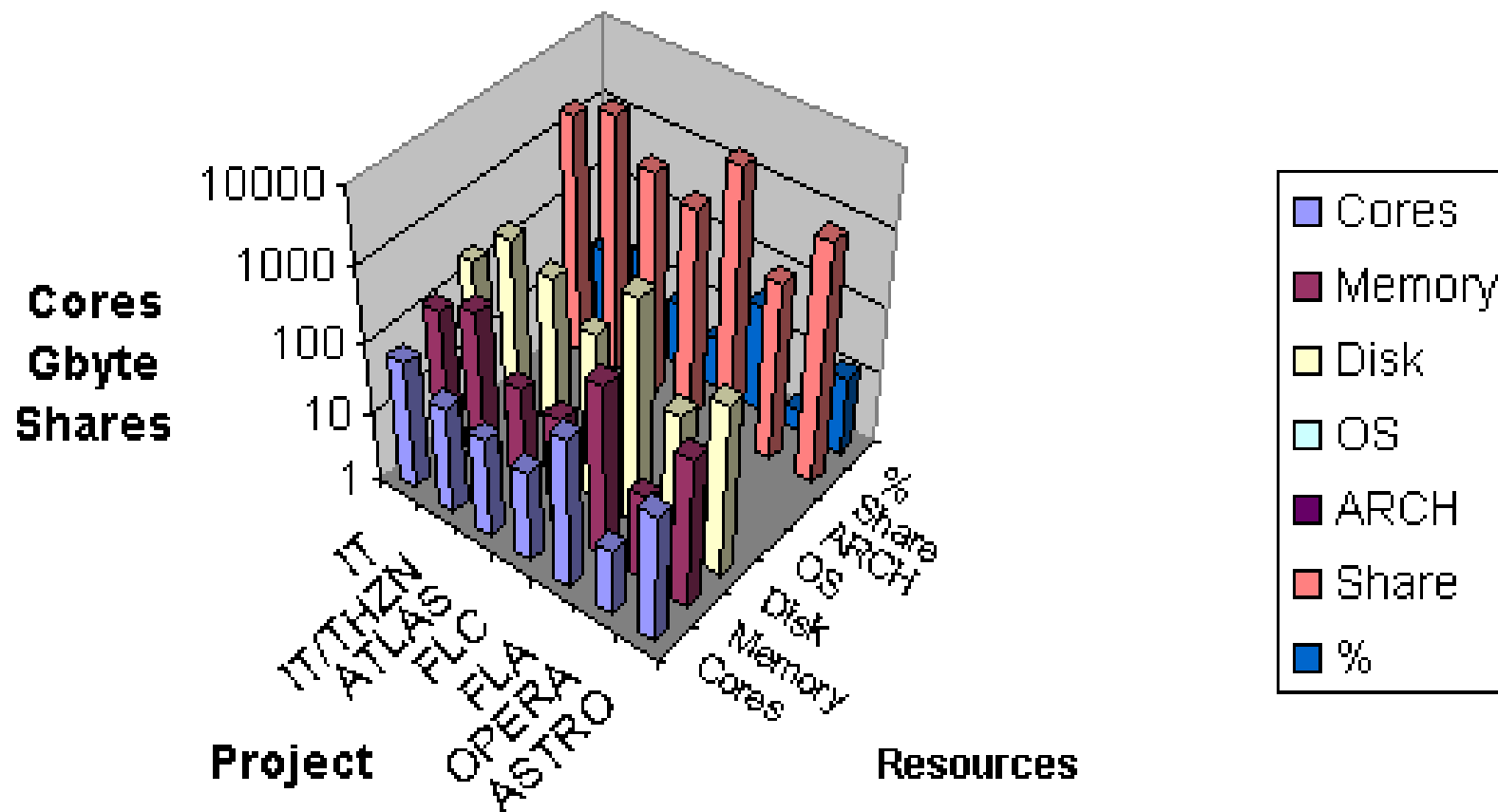


- > AFS and Kerberos Support for Authentication and Resource Access
 - Valid Tokens during Complete Job Execution
- > Cores, h_rt, h_vmem, h_fsize Under Full Control of Scheduler
 - *Select your Resources and we guarantee for it*
- > MPICH2 Parallel Environments
 - mpich2-1 (Single Host)
- > 4 Big Birds 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





Fairshare and Resources 2009





- > 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 its 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 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 %



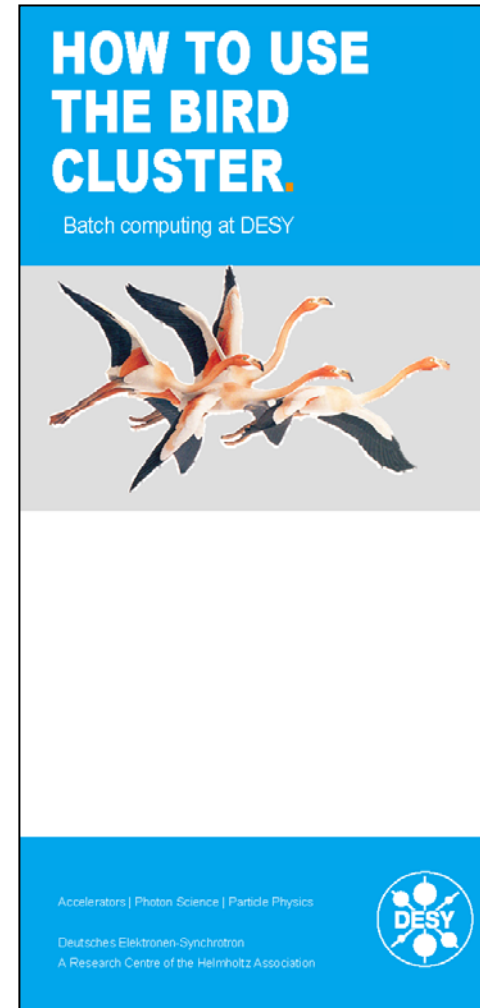


- > IT Hardware + Project Specific Hardware = Valuable Resources
- > Contributing projects will be granted Fairshare Points
 - 10 for each Compute Core,
 - 10 for each 2 GByte Memory
 - 1 for each GByte Disk
- > Guaranteed Access to this Relative Amount of Batch Resources to the Project Members
- > IT 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*)





- > <http://bird.desy.de>
- > BIRD-Flyer:





> Update to N1GE 6.2

- Multi-Clustering With Service Domain Manager
- Improved Scalability and Job Throughput
- Advance Reservations
- New Support for Interactive Jobs
- Multi-Cluster Support for Accounting and Reporting Console
- Ability to Request the Master and Slave Queues for Parallel Jobs
- New Unix Resource Limits Support
- Scalability Improvements
- Memory Foot Print Reductions in Huge HPC Clusters
- Job Submission Verifiers
- Consumable Resources Per Job
- jemalloc Library
- Bug Fixes

> MPICH2 Parallel Environment

- Mpich2 (Multi Host)
- With n1ge 6.2+

> Update Kerberos Support

- (Internal) Ticket Prolongation
- Kerberos 5 Setup

> Virtualized Worker Nodes (?)

- OS Version as a Dynamic Resource



