Status of DESY Batch Infrastructures

This presentation will provide information on the status of the batch systems at DESY Hamburg. This includes the clusters for GRID, HPC and local batch purposes showing the current state and the activities for upcoming enhancements.

Thomas Finnern
DESY/IT-Systems

HELMHOLTZ GEMEINSCHAFT

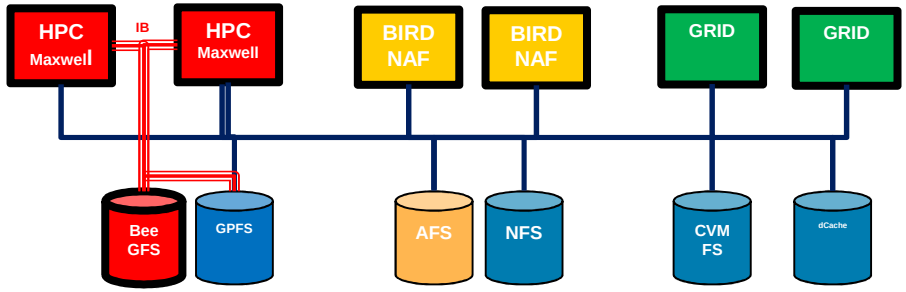
HEPiX BROOKHAVEN NATIONAL LABORATORY

BROOKHAVEN NATIONAL LABORATORY

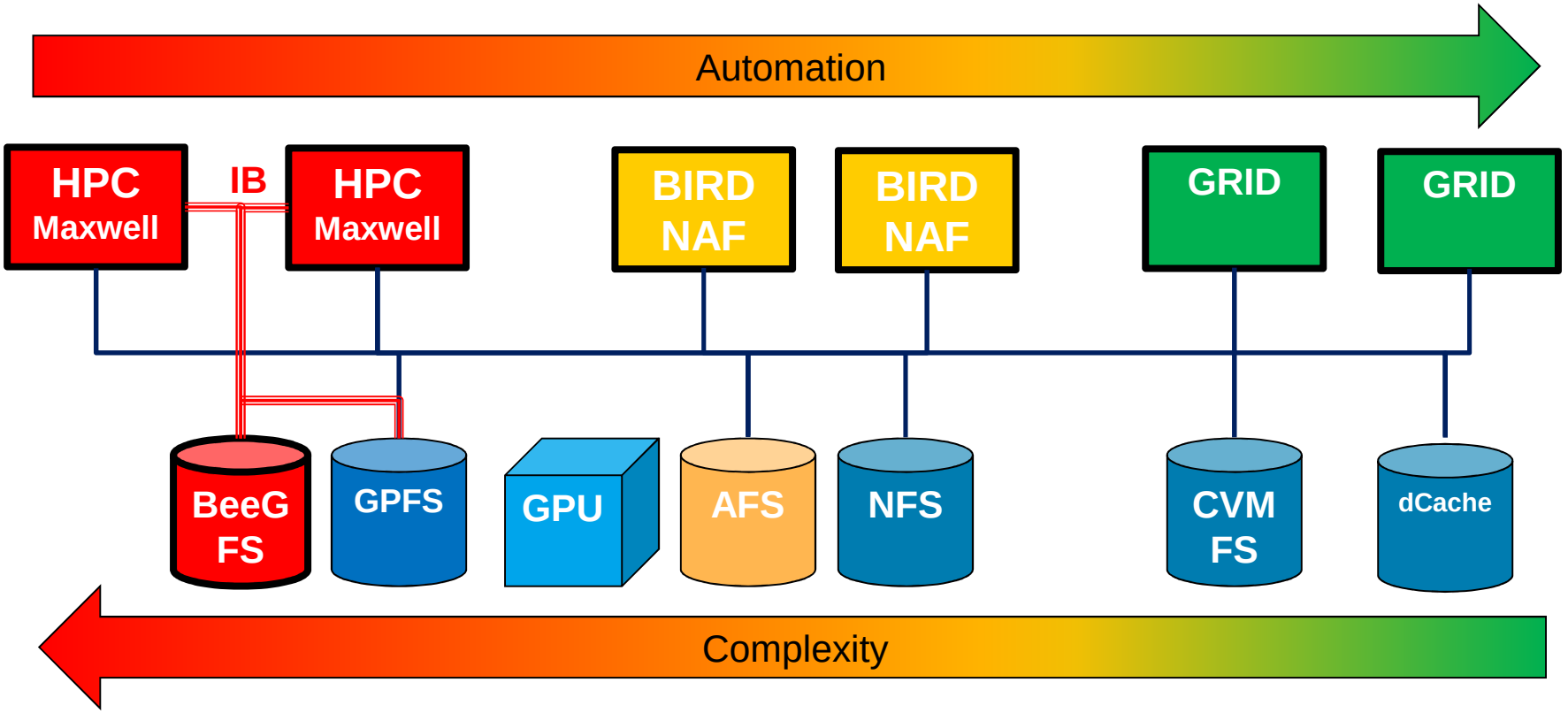
DESY

The Team and the Outline

- The Team
  - Christoph Beyer
  - Jan Engels
  - Thomas Finnern
  - Martin Flemming
  - Andreas Gellrich
  - Yves Kemp
  - Frank Schlünzen
  - Sven Sternberger
- Outline of Talk
  - Topologies
  - Resources and Policies
  - Status and Numbers
  - Plans



Component Topology



Calendar Reservation

Son Of Grid Engine SoGE

Torque +MySched

Cream CE

Batch Infrastructure Resource at DESY

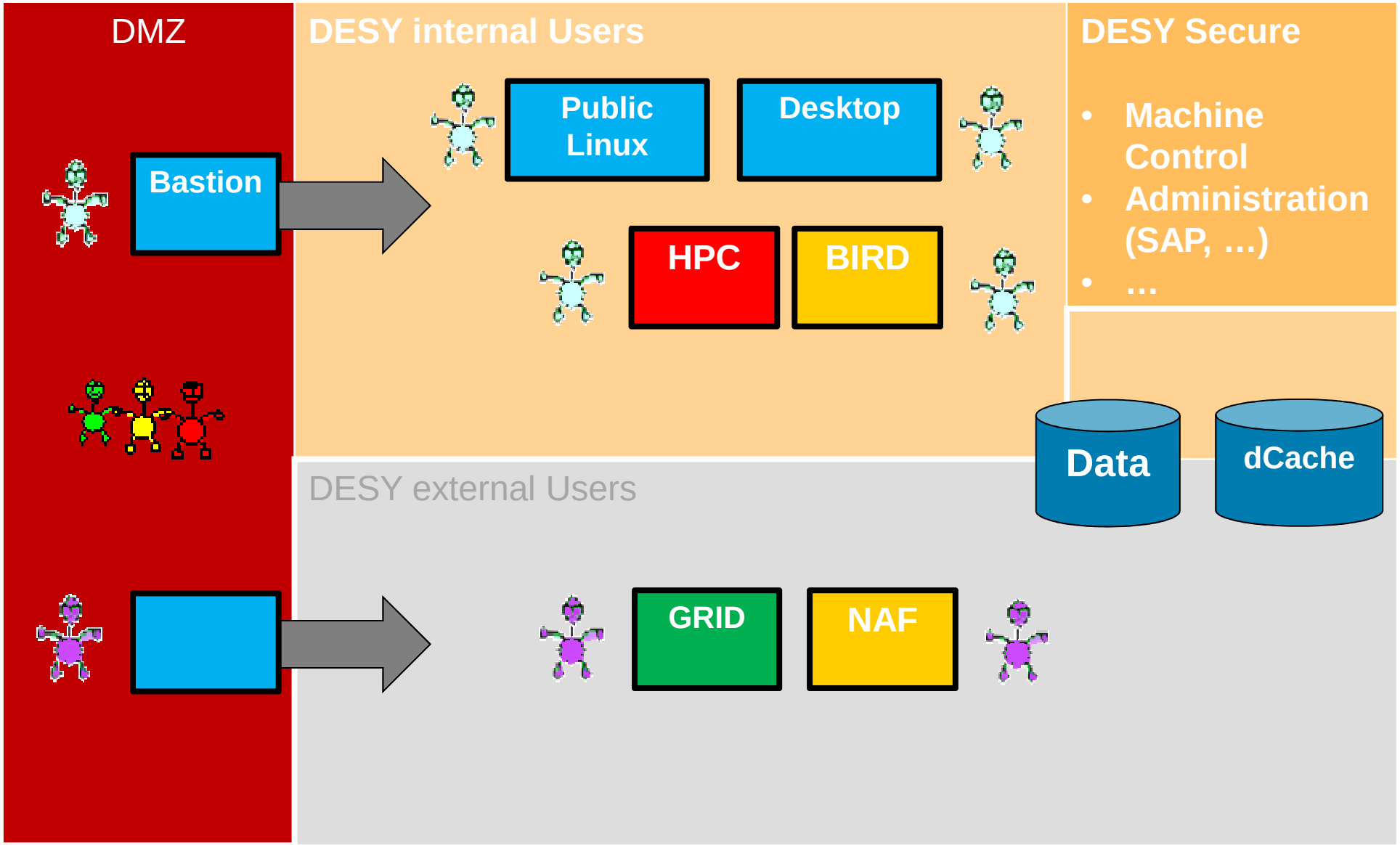
Petra III Data Analysis

National Analysis Facility

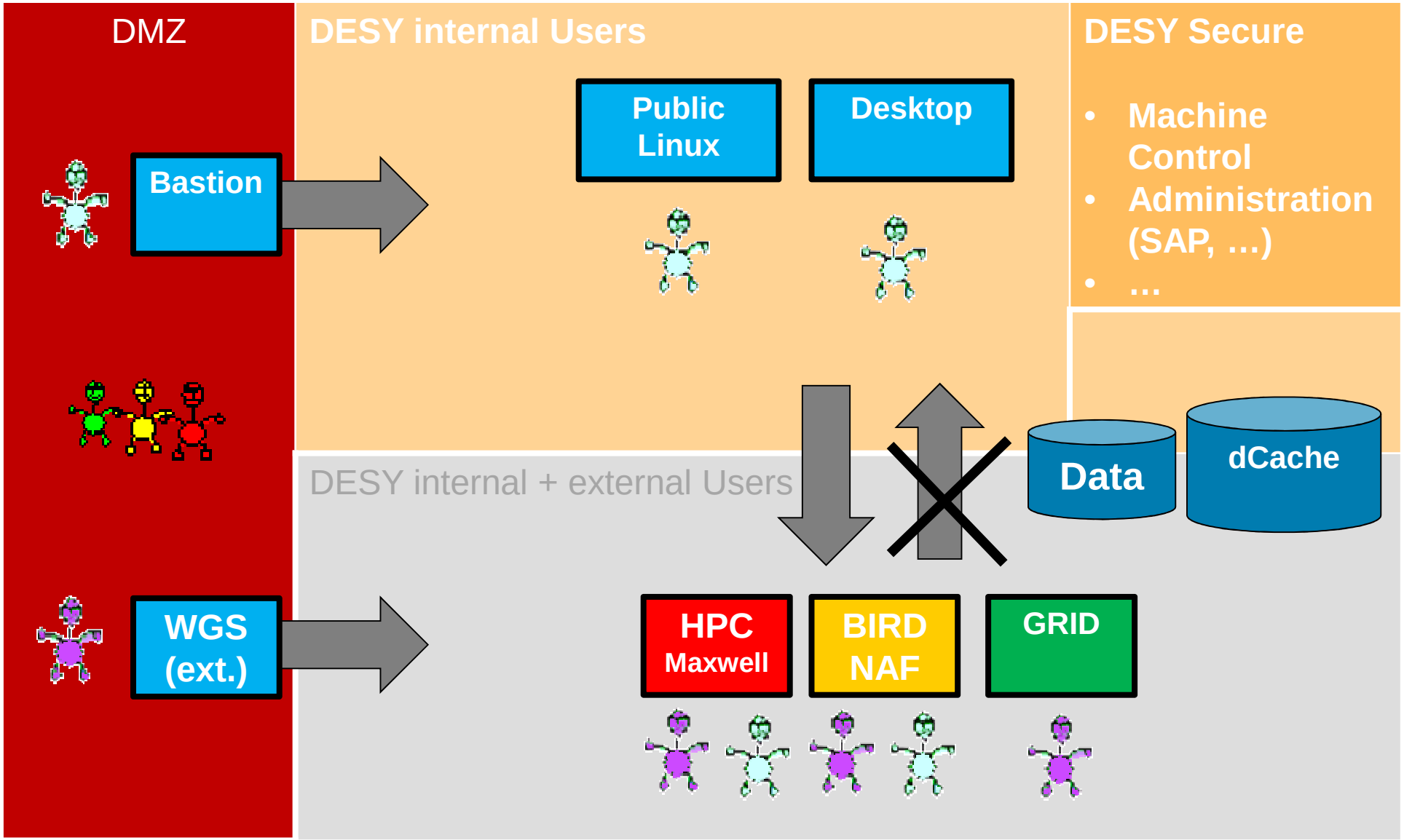
Robots

People **Scaling** **Support**

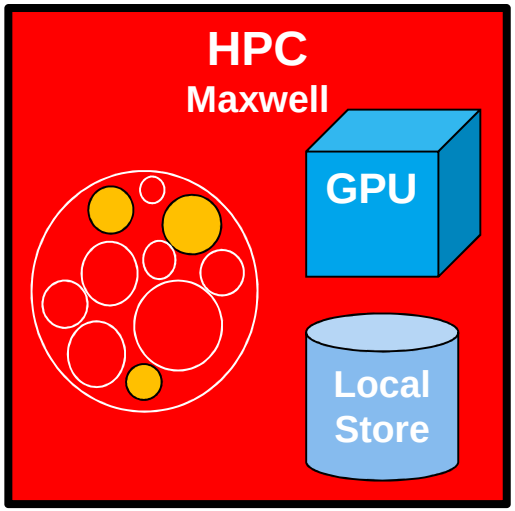
Access Topology (Before)

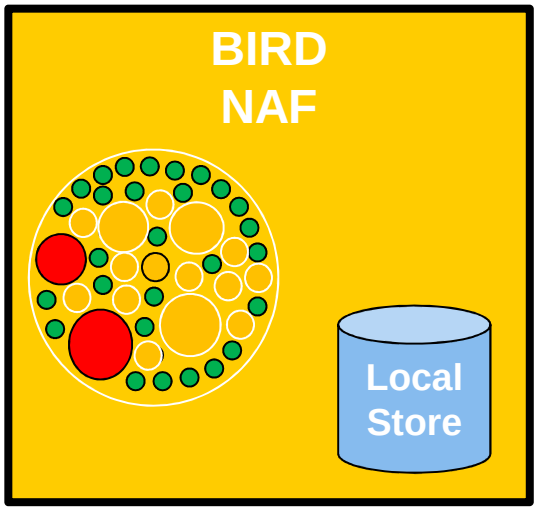


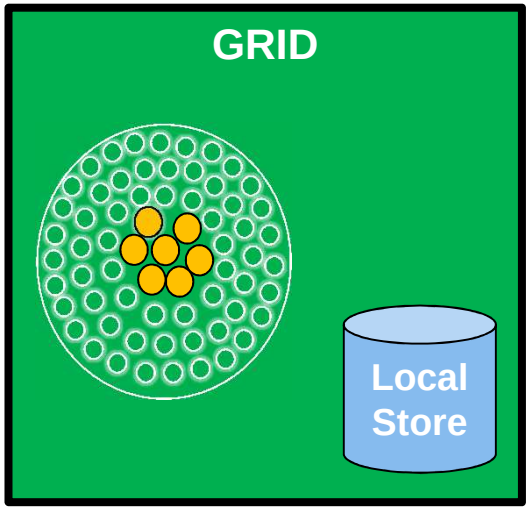
Access Topology (Now)

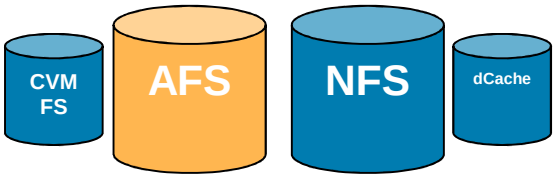


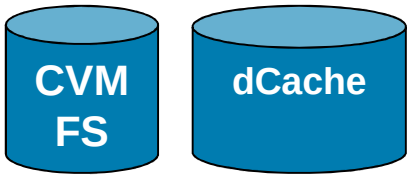
Resource Allocation











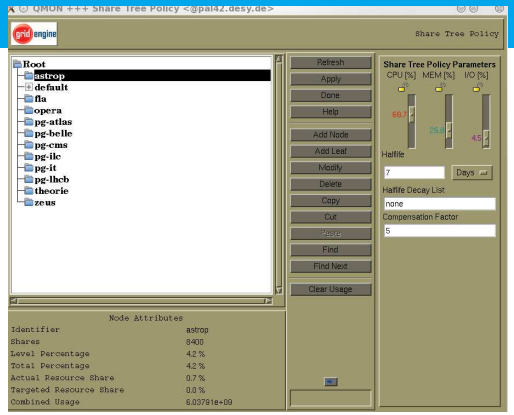
Consume Hosts **People** **Consume Shared Resources** **Robots**

Diversity Handling **Fair Share** **Cluster Fill Level (Core, Memory, Disk, …)**

Dynamic Project Support

Optimizing vs. Securing **Queue Waiting Times**

Beneficial Features and Policies for Batch

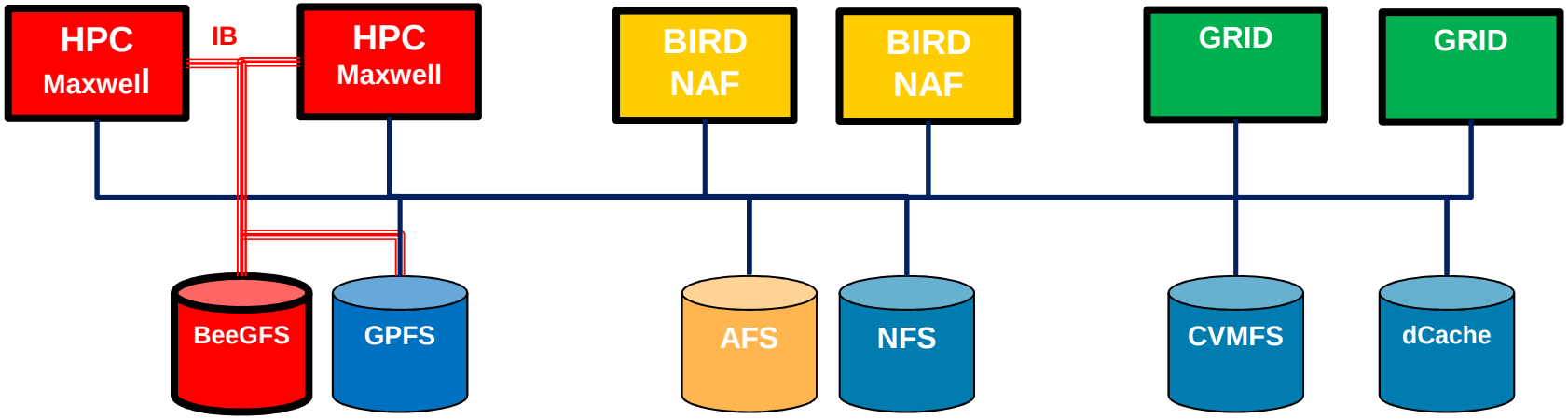


- Project Fairshare
  - Configuration for
    - Relative Share
    - Decay Time
    - Resources
- Resource Groups
  - Light resources with fast access
  - Big resources with slow(er) access
  - Requestable and Guaranteed
    - Hosts, **Cores, Time, Memory, Disk**, I/O, License …
- Kerberos Support
  - Use of personal access tokens
  - Token Live Time / Prolongation according to job times
  - User specific AFS Access
- Multicore Jobs
  - Local to one Host
  - Distributed over several Hosts
- Authentication / Authorization
  - User/Admin Access
  - Dynamic Project Membership
  - Cluster Hosts
- Accounting
- Monitoring
- Other Features
  - Job Data

    Shared File System
    Staging
    Input/Output Files
  - Host Reservation
  - Scheduled Reservation
  - GPU Reservation
  - MPI Support
  - HA Batch Server

We have a Plan !

Now	Calendar Reservation Tool		Son Of Grid Engine SoGE	Torque + MySched	Cream CE
Test		☜ ☟	☟	☟	☟
Future		SLURM	HTCondor		ARC CE



- Scaling
- Comply to future HPC

- Scaling and Support
- Common Support Team

- Support

Test Status SLURM

- 1 Management Node
- 6 Test Compute Nodes
- Node Authentification: munge
- Cluster Filesystem: BeeGFS
  - Supports MPI-IO and dgl
- HOME Dirs in BeeGFS
  - Fast low latency data access by default
- Scheduling only complete Nodes
  - Old Scheduling system should be replaced
  - Need also interactive user login
- Logging+Accounting: SlurmDBD
- No Kerberos (No AFS Home access).
- Software Distribution over NFS Share
- Petra III Data Taking
  - GPFS Access
- Different SLURM Partitions
  - Private Group Nodes
  - GPU-Nodes
  - Standard Nodes
- References
  - http://www.llnl.gov/linux/slurm

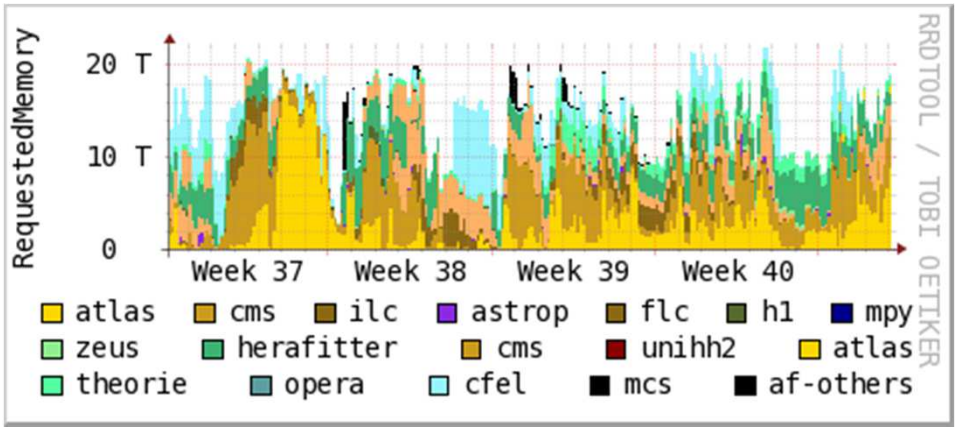
Test Status HTCondor

- Partitionable/Dynamic Slots
- Hierarchical Accounting Groups
- PID Namespaces
- CPU Cgroups
- MOUNT_UNDER_SCRATCH
- Memory Cgroups
- Condor_ssh_to_jobs
- NFS Share
- Monitoring
  - Ganglia (… with a little help from Andrew from RAL)
- Grid Test Running
- For BIRD/NAF
  - Waiting for Kerberos/AFS Feature
- Simple Batch Working
- Fairshare
  - No direct Transformation (=~ hierarchical accounting groups + quotas + priority mechanism)
- Docker ?
  - EL7, Ubuntu, ... (Remote Future)
- References
  - http://research.cs.wisc.edu/htcondor

Numbers

	Ser-ver	Wor-ker	Cores	HT	Mem/ GB	Mem/ Core	Disk / GB	Disk/ Core	Net	GPU
Grid	5	254	9672 (11365)	yes	29000	3 GB	193000	20 GB	1 G	-
BIRD	2	560	6500	no	13000	2 GB	124000	18 GB	1 G	-
Test	2	10	960	yes	320	3 GB	17280	18 GB	1 G	-
HPC	1	38	2432	-	9700	4 GB	70000	3,5 GB	IB + 10 G	6+1
Maxwell		15	288	no	970	16 GB				2+0
Test	%	6	384		1530		%	%	%	-

Batch Infrastructure Resource at DESY



Outlook and Conclusions

- Large Conceptional Differences between Resource Schedulers
- „Proof of Concept“ for each planned feature needed
- GRID and BIRD/NAF
  - Combined Support with HTCondor
  - GRID: HTCondor/ARC CE 2016
  - BIRD/NAF: HTCondor/AFS 2016
    - Waiting for AFS/Kerberos Support
    - More Features to be approved
    - Want smooth Transition
  - GRID and BIRD/NAF
    - Starting with GRID
    - Share of both should be maximized

- Maxwell HPC Cluster
  - SLURM should replace Reservation System
  - Maxwell Cluster with SLURM 2016
  - Petra III Data Analysis

- Questions ?
- Answers !