HPC @ RRZE

Georg Hager

Regionales Rechenzentrum Erlangen

CCC, 24.04.2007

Overview

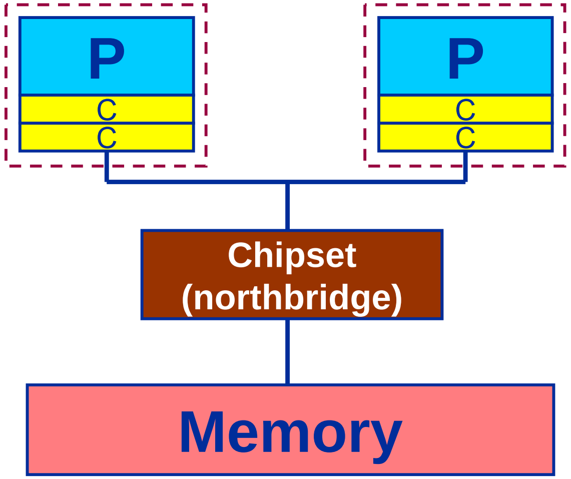
- Common HPC system layouts
  - Clusters
  - Shared-memory nodes
  - Large shared-memory systems
- Systems at RRZE
  - General description, modules system
  - Performance comparison of cluster systems
- File systems
  - Local, NFS, parallel
  - File systems at RRZE
- Batch processing
  - Basics
  - Situation at RRZE
  - Suggested changes
- Which system?

HPC system layouts

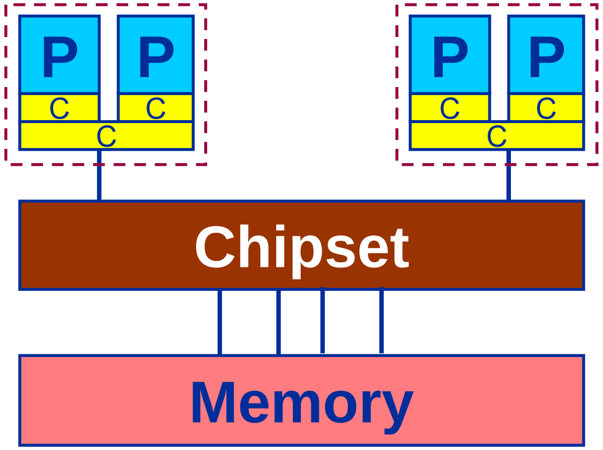
- Today: Clusters are dominant in mid-range HPC centers
  - “Node”: PC-like machine with (usually) up to 2 sockets, i.e. a shared-memory system (see next slide)
  - Several networking options to connect nodes
    - GBit Ethernet (125 MByte/s per direction, latency 30-80 µs)
    - InfiniBand (1-2 GByte/sec per direction, latency 3-5 µs)
    - Quadrics
    - Myrinet
    - …
  - “Head nodes” for compiling, testing, submitting jobs etc.
  - Programming paradigm: Message Passing (MPI)
  - File systems available for short-term and long-term data storage
    - Node-local disks
    - Central NFS store
    - Parallel file system

Shared memory nodes: Some examples

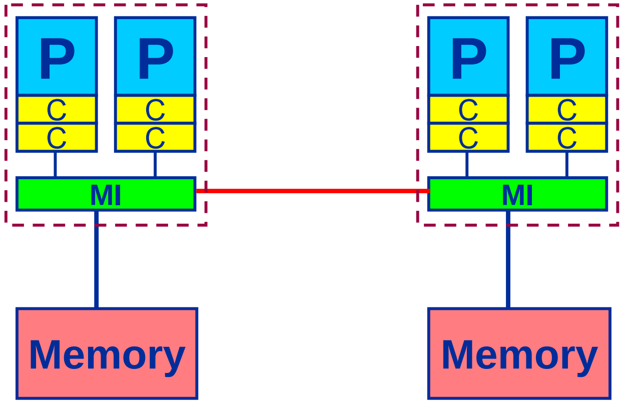
- Dual CPU Intel Xeon node



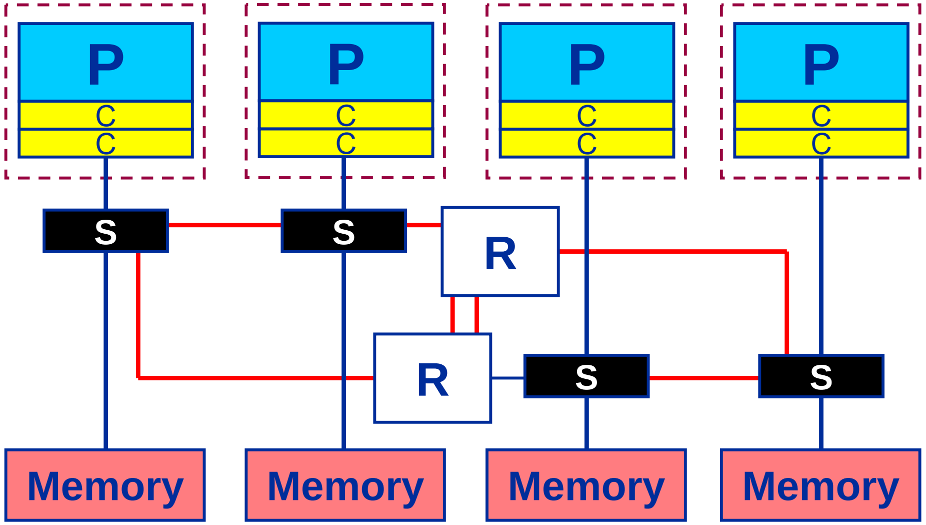
- Dual Intel “Core 2” node



- Dual AMD Opteron node

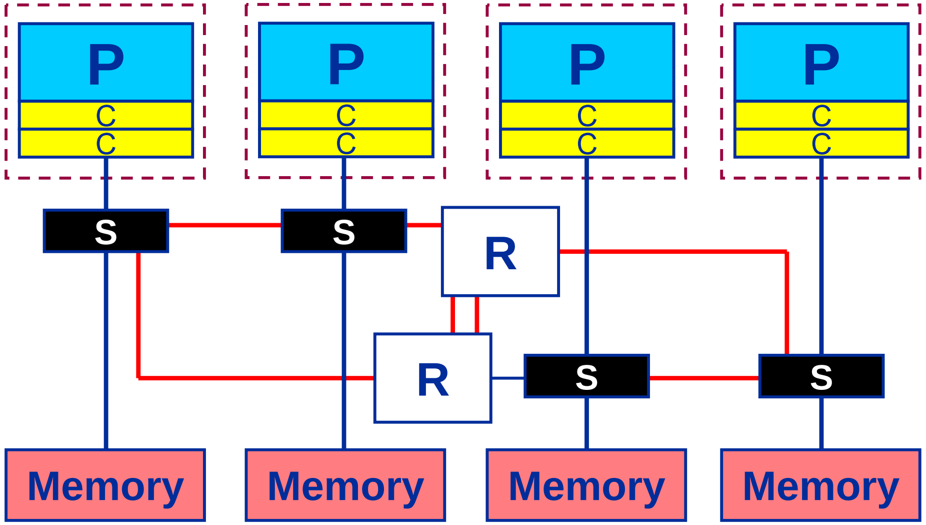


- SGI Altix (HLRB2 @ LRZ)



HPC system layouts

- Large shared-memory systems
  - Non-uniform memory access (ccNUMA)
  - Memory physically distributed but logically shared
  - Very fast network (SGI Altix: NumaLink)
  - Programming paradigms
    - Message Passing (MPI)
    - Shared-Memory programming (OpenMP, pthreads)
  - Problem: Locality of access must be enforced by explicit programming or system tools
  - Suitable for large-memory applications



HPC Systems at RRZE

SGI Altix3700/330

- 32+16 Intel Itanium2 processors (1.3/1.5 GHz; 5.2/6.0 GFlop/s)
- Peak performance: 262 GFlop/s
- 128+32 GB shared memory (ccNUMA)
- Supports all common programming models
- Fast NumaLink3 network
- 2700 GB hard disk space
- Linux OS
- PBSPro batch system

  Moderately parallel applications with high memory and communication demands

http://www.hpc.rrze.uni-erlangen.de/systeme/sgi-altix-3700.shtml

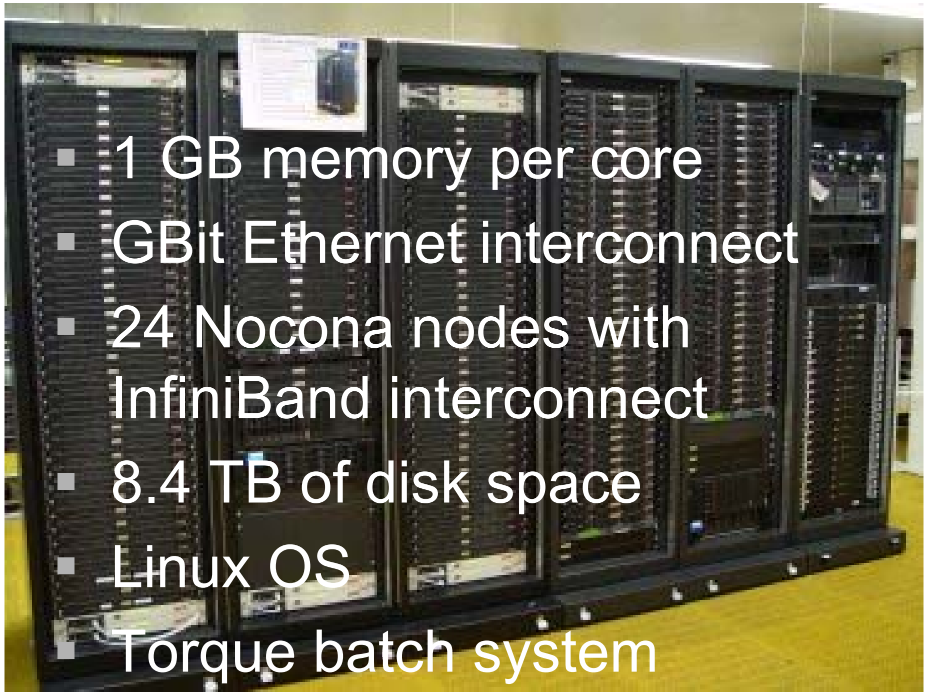
Altix @ RRZE

- **“Frontend”**
  - altix.rrze.uni-erlangen.de
  - 4 CPUs (with 8 GB of memory) for compiling and tests
  - Access for all HPC-enabled accounts
- **Special software**
  - SGI’s own MPI implementation: MPT
  - Some special CFD packages
  - Chemistry packages under **`/opt/bcosw`**
- **Modules system for software packages (see below)**
- **File systems: see below**

HPC @ RRZE

IA32/EM64T/AMD64 Cluster

- 172 Intel Xeon 2.66 GHz CPUs (Prestonia, 32 Bit)
- 128 Intel Xeon 3.2 GHz CPUs (Nocona, 64 Bit)
- 100 AMD Opteron 2.0 GHz cores (64 Bit, ccNUMA)
- Overall peak performance: 2134 GFlop/s



- 1 GB memory per core
- GBit Ethernet interconnect
- 24 Nocona nodes with InfiniBand interconnect
- 8.4 TB of disk space
- Linux OS
- Torque batch system

Massive parameter studies and parallel codes with low communication demands

http://www.hpc.rrze.uni-erlangen.de/systeme/ia32-cluster.shtml

IA32 Cluster @ RRZE

- **Frontends (access for all HPC-enabled accounts)**
  - sfront01/2.rrze.uni-erlangen.de
    - IA32 architecture, 2 CPUs, 4GB memory
    - For compiling and short (serial and parallel) test runs
  - sfront03
    - x86-64 architecture (called EM64T by Intel)
    - 2 CPUs, 4GB memory
    - For compiling and short (serial and parallel) test runs
  - CPUtime-killer
    - Long running jobs are killed automatically
  - Memory killer
    - RSS memory limit per process limited to 1GB
- **Modules system for SW packages**
  - Chemistry SW under /opt/bcosw
- File systems: see below

Woodcrest Parallel Computer „Woody“

- 752 Intel Xeon/Woodcrest (Core2) 3.0 GHz Cores
- Peak Performance: 9024 GFlop/s
- 2 GB memory per core
- InfiniBand interconnect (10 GBit/s per direction, <4µs latency)
- Parallel file system with 15 TB
- NFS file server with 15 TB
- Linux OS
- Torque batch system

Applications with high parallelism and high demand for communication

Top500 rank 124

(Nov 2006)

http://www.hpc.rrze.uni-erlangen.de/systeme/woodcrest-cluster.shtml

Woody @ RRZE

- **Frontends**
  - woody.rrze.uni-erlangen.de
    is aliased to woody1/2
  - Automatic distribution of logins to one of the frontends
  - 4 CPUs, 8GB of memory
  - For compiling and serial test runs
  - For parallel tests submit interactive jobs (`qsub –I -X`)
  - CPUtime killer
- **Special software**
  - DDT parallel debugger, Intel Trace Analyzer
- **Modules system**
  - Chemistry SW under /opt/bcosw
- **File systems: see below**

Modules system

- **“module” command**
  - Provides presets for PATH, MANPATH, LD_LIBRARY_PATH, …
  - Module = collection of variable settings for a specific application or group of applications
  - Can be loaded and unloaded
  - Mutual dependencies and exclusions can be implemented
- **How to use the module command**
  - **`module load <name>`**
    - Loads the module
    - Example: **`module load intel64`**
      - after that can use icc, icpc, ifort
  - **`module rm <name>`**
    - Unloads the module
    - Undoes all changes to env variables
  - **`module avail`**
    - Lists all available modules
- **See**

  `http://www.hpc.rrze.uni-erlangen.de/systeme/software-umgebung.shtml`

Modules system

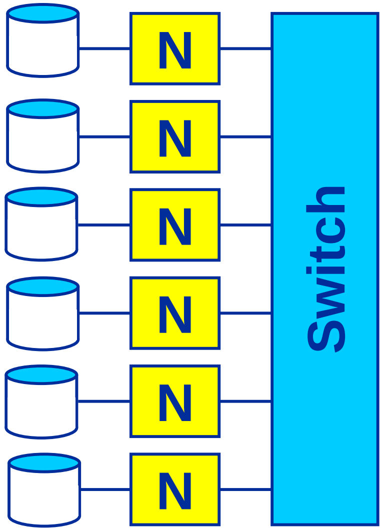
- **Using modules in batch scripts**
  - Often not required
  - If you need modules in a batch script you can
    - Use the module command right away if you use a csh-based script
    - Use the module command in a sh-based script if the script shell is a login shell:

```
#!/bin/bash --login
...

module load turbomole/5.8
```

- **Using modules in your .cshrc**
  - Often not a good idea
  - Tends to lead to problems with new systems

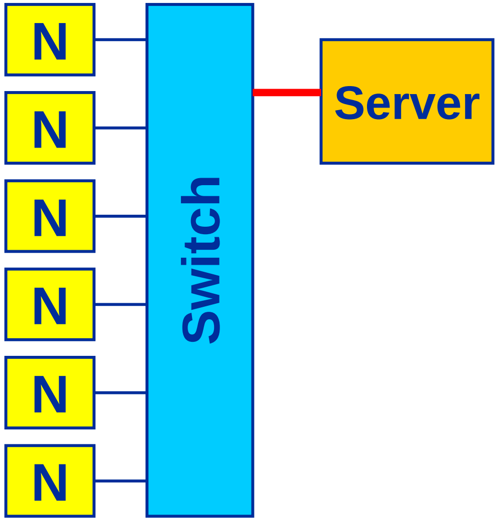
File Systems

Options for file systems: Local disks



- **Local disks at cluster nodes**
  - EIDE/SATA/SCSI disk, local controller
    - Like in your own PC…
  - Suitable for job-local data
  - Only visible on same node
  - Sliding window data deletion (data older than N days will get deleted automatically)
  - No prerequisites to prevent file system fill-up
    - Admins will be notified, but by then it is usually too late
  - Speed ("bandwidth"): 30-80 MByte/s per node
  - No backup
  - Often: job-specific data directory made at job start, deleted at job termination
  - Use this for scratch files and process-specific data

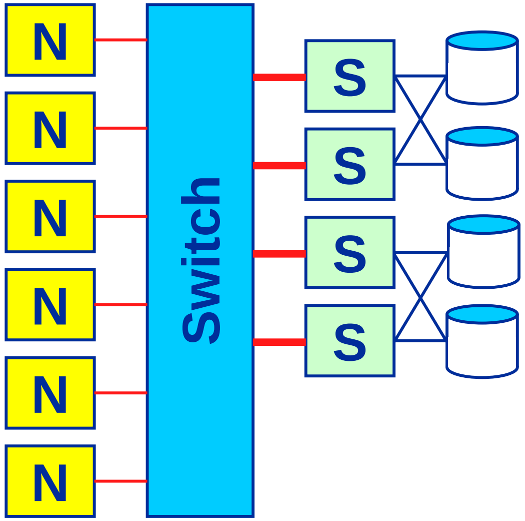
Options for file systems: NFS



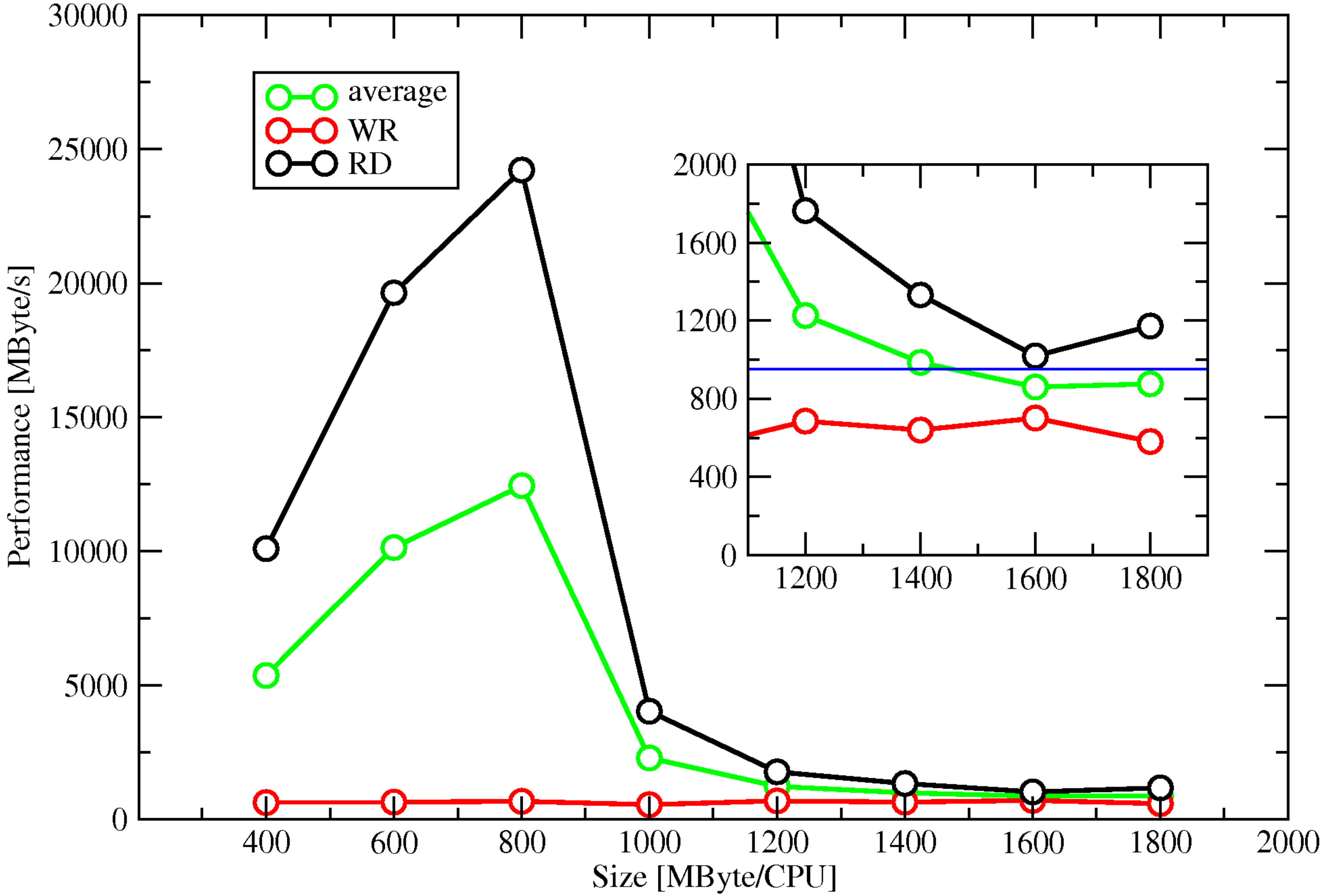
- **Global NFS storage**
  - Long-term data store
  - Every large HPC system has at least one local (= well-connected) NFS volume
  - Available under /home/<SYSTEM>/
  - Beware the usual NFS bottleneck
    - Bandwidth inherently limited by network speed, even if disks could do more
    - All /home file systems are available via GBit Ethernet only
    - If possible, avoid NFS usage from batch jobs
  - No backup for large HPC volumes, but…
  - $HOME is under backup
  - Everything that starts with /home/… is visible on all RRZE machines
  - Try to keep up some sensible data management, i.e. avoid multiple copies of data

Options for file systems:
Parallel file system

- **Parallel file system SFS**
  - Only available (and visible) on Woody cluster
  - Storage cluster of 4 nodes (+ 2 for metadata) with InfiniBand connection to the central cluster switch
  - Aggregate bandwidth of ≈ 1 GByte/s for parallel I/O (MPI-I/O)
    - Much higher bandwidth possible if caching effects can be used (see next slide)
    - ≈ 600 MByte/s max bandwidth for a single compute node
  - Optimized towards large files, contiguous access
  - Suitable for fast global data storage
    - E.g. large restart files
  - High watermark deletion algorithm
    - Fill level > 85% → delete old files until fill level is below 60%



SFS performance on Woody (PIO benchmark) on 16 nodes / 64 CPUs



Overview on file systems: Woody

Mount point	Access via	Purpose	Technology, size	Data lifetime	Quota
/home/ rz[sun]home	$HOME	source, input, imp. results	NFS on RRZE servers, small	Account lifetime	**YES**, restrictive
/home/ woody	$WOODYHOME	cluster-local large vol. store	NFS, 15 TB	Account lifetime	**YES**
/wsfs	$FASTTMP	High perf. parallel I/O; short-term store	SFS parallel FS via IB, 15 TB	**High watermark deletion**	NO
/tmp/…	/tmp, $TMPDIR	temp. job data directory	Node-local RAID0, 130 GB	**Auto-delete, Job runtime**	NO

Overview on file systems: IA32 Cluster

Mount point	Access via	Purpose	Technology, size	Data lifetime	Quota
/home/ rz[sun]home	$HOME	source, input, imp. results	NFS on RRZE servers, small	Account lifetime	**YES**, restrictive
/home/ cluster32	/home/ cluster32/ <GROUP>/ <USER>	cluster-local large vol. store	NFS, 6 TB	Account lifetime	**YES**
/home/ cluster64	/home/ cluster64/ <GROUP>/ <USER>/	cluster-local large vol. store (SFB)	NFS, 2 TB	Account lifetime	**YES**
/tmp	/tmp	temp. job data directory	Single ATA disk, 70GB	Auto-delete	NO

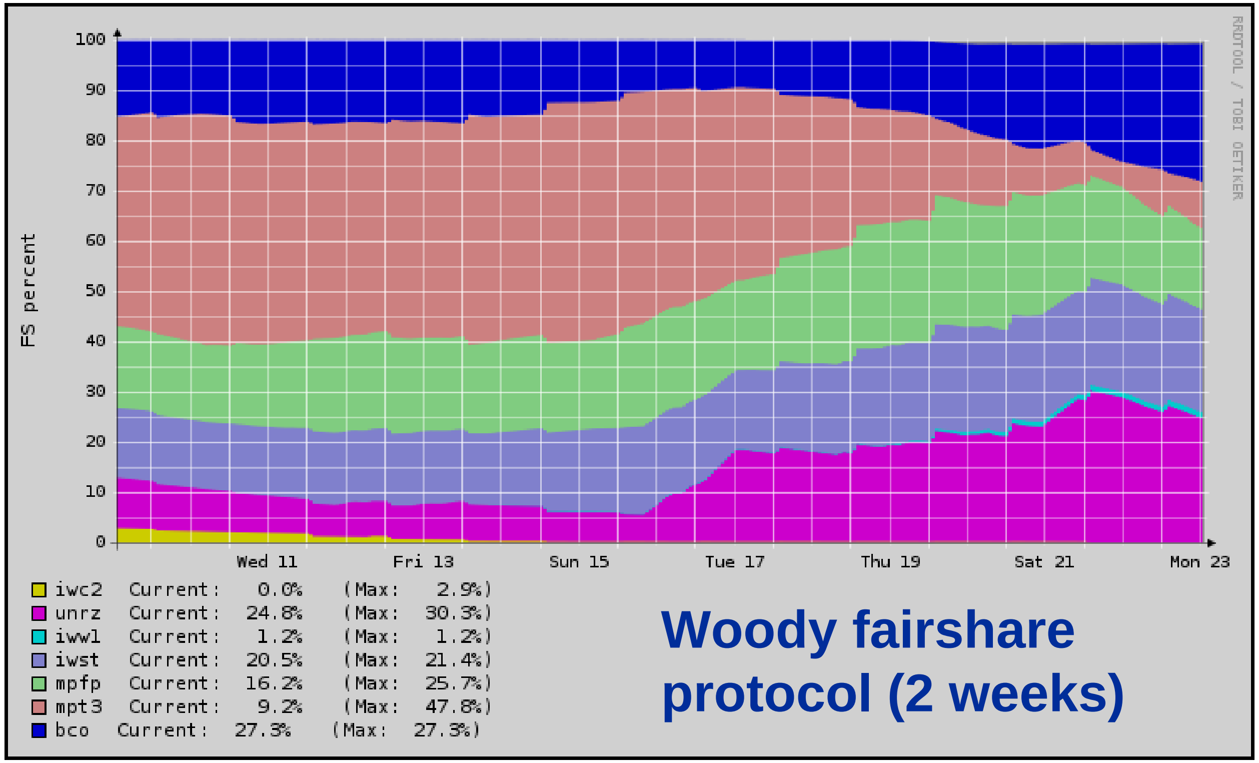
Overview on file systems: Altix

Mount point	Access via	Purpose	Technology, size	Data lifetime	Quota
/home/ rz[sun]home	$HOME	source, input, imp. results	NFS on RRZE servers, small	Account lifetime	**YES**, restrictive
/home/ altix	/home/ altix/ <GROUP>/ <USER>	local large vol. store	Local array (altix-batch) or NFS (altix), 2.7 TB	Account lifetime	**YES**
/scratch	/scratch, $TMPDIR	temp. job data directory	Local array	Auto-delete, job runtime	**NO**

Batch processing

Some general remarks on batch

- **Batch processing tries to ensure some “fair” distribution of resources among people and groups**
- **Main component of job priority: “fairshare”**
  - Accumulated runtime over preceding 10 days per user and group with a damping weight of 0.95 per day
  - FS target: 10% for non-privileged users, more for special groups (bco, mfbi, …)



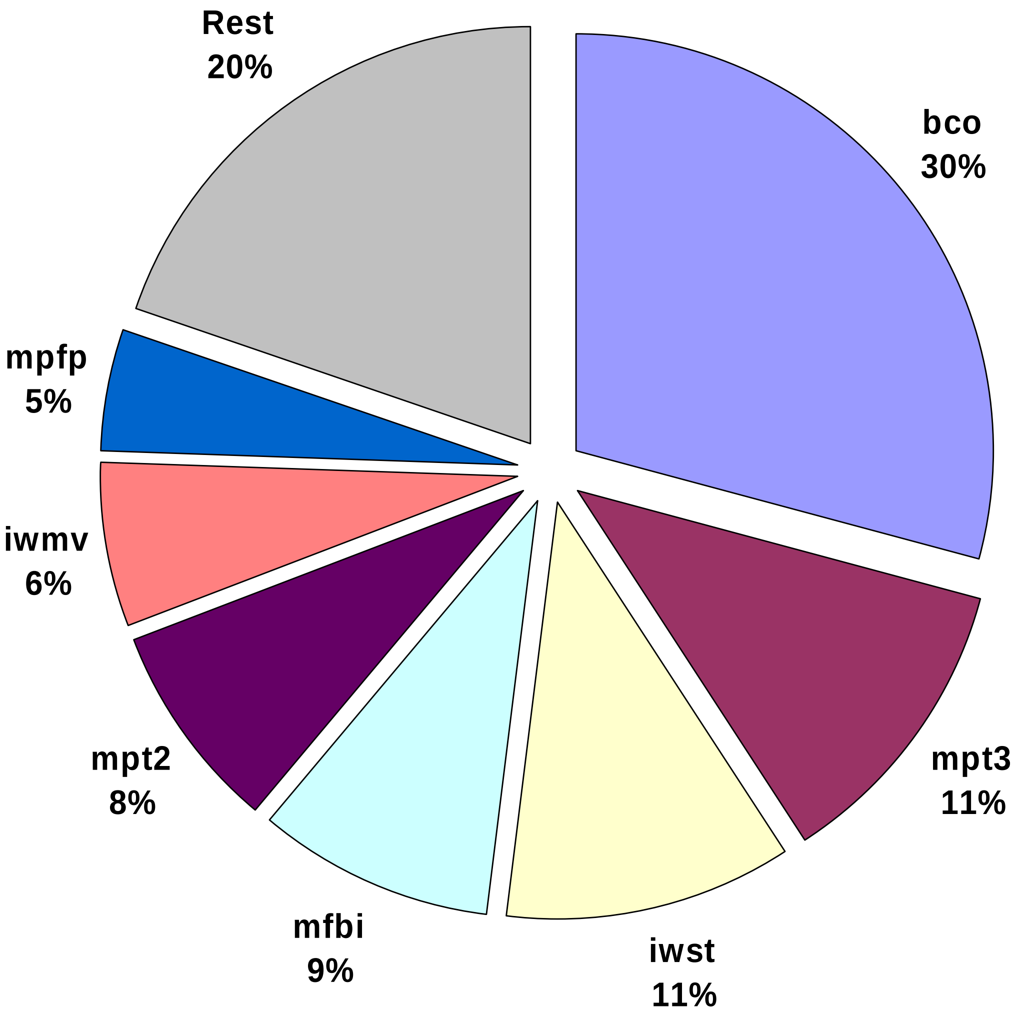
Woody fairshare protocol (2 weeks)

Some general remarks on batch

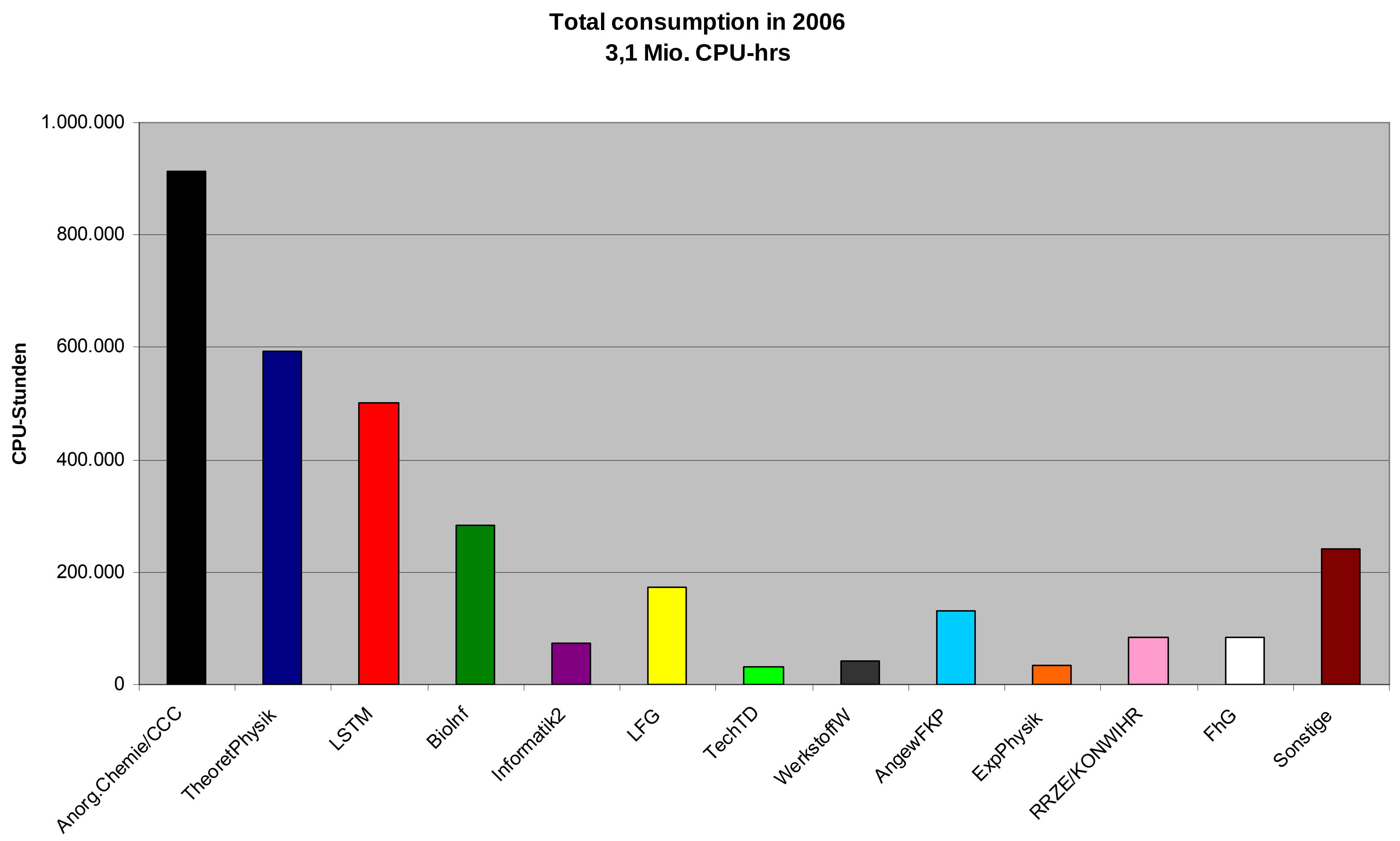
- **Other factors (minor importance)**
  - Queue priority (small but essential!)
  - Job wait time
  - Quality of service

```
Job                    PRIORITY*   Cred(  QOS:Class)    FS( User:Group)  Serv(QTime)
             Weights   --------       1(    1:   10)   100(   10:   10)     1(    1)

12608                       3910   81.9(  0.0:500.0)  18.0(9287.:-1038)   0.1(  8.3)
12592                       2087    9.6(  0.0: 20.0)  85.0(887.1:887.1)   5.4(112.4)
12601                      -1777    8.9(  0.0: 20.0)  89.4(8369.:-1038)   1.7( 39.5)
12590                      -7561    2.4(  0.0: 20.0)  96.1(2508.:-1038)   1.4(116.3)
12604                      -7657    2.5(  0.0: 20.0)  97.3(2508.:-1038)   0.2( 19.8)
12598                      -8341    2.3(  0.0: 20.0)  97.0(1782.:-1038)   0.7( 61.7)
12603                      -8374    2.3(  0.0: 20.0)  97.4(1782.:-1038)   0.3( 29.0)
12594                     -11760    1.6(  0.0: 20.0)  97.6(-5954:-6100)   0.8( 94.1)
12596                     -11771    1.6(  0.0: 20.0)  97.7(-5954:-6100)   0.7( 83.5)
12599                     -11803    1.6(  0.0: 20.0)  98.0(-5954:-6100)   0.4( 51.5)
12602                     -11821    1.6(  0.0: 20.0)  98.1(-5954:-6100)   0.3( 33.6)

Percent Contribution   --------     7.5(  0.0:  7.5)  91.8( 54.4: 93.5)   0.7(  0.7)
```

Accounting 2006 Cluster32



Accounting all systems at RRZE 2006



Information about jobs and cluster status

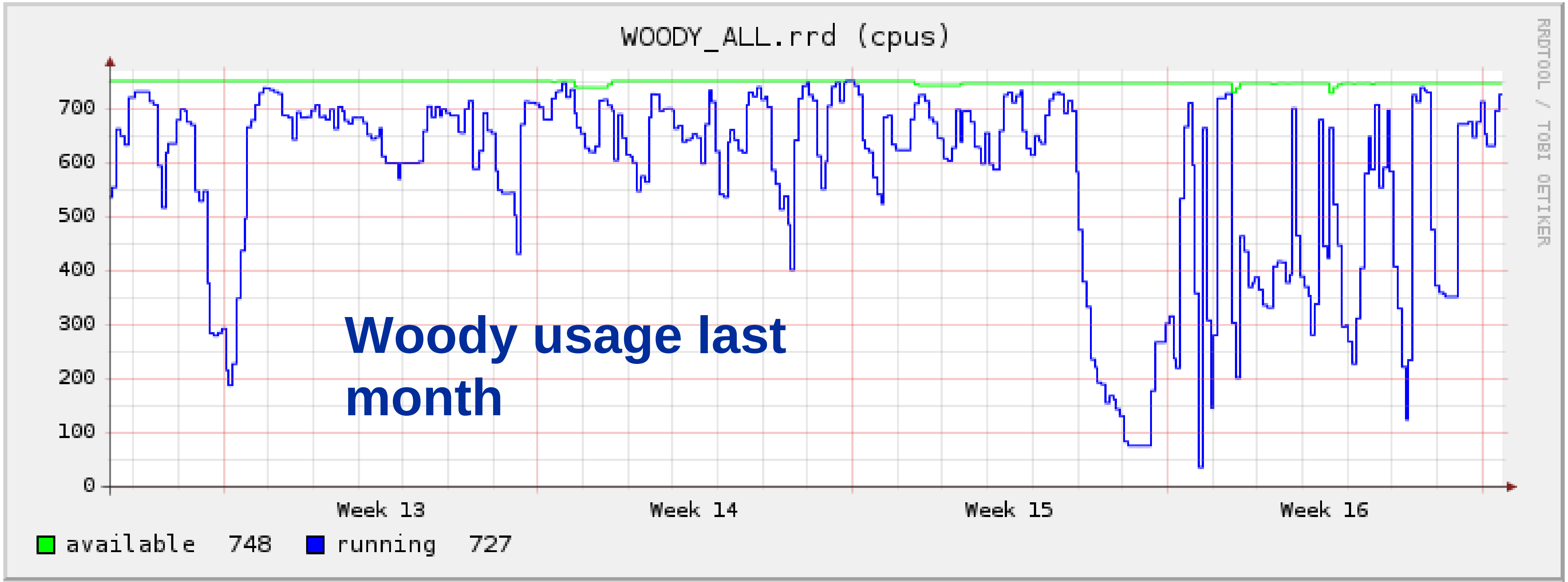
- IA32 cluster
  - `/home/cluster32/hpcop/joblist`
  - `/home/cluster32/hpcop/nodelist`
  - `/home/cluster32/hpcop/downlist`
  - `/home/cluster32/hpcop/reservationlist`
- Woody
  - `/home/woody/STATUS/joblist`
  - `/home/woody/STATUS/nodelist`
  - `/home/woody/STATUS/downlist`
  - `/home/woody/STATUS/reservationlist`

- Even better:
  - `http://www.hpc.rrze.uni-erlangen.de/kundenbereich/`
  - Get user/PW by “docpw” command

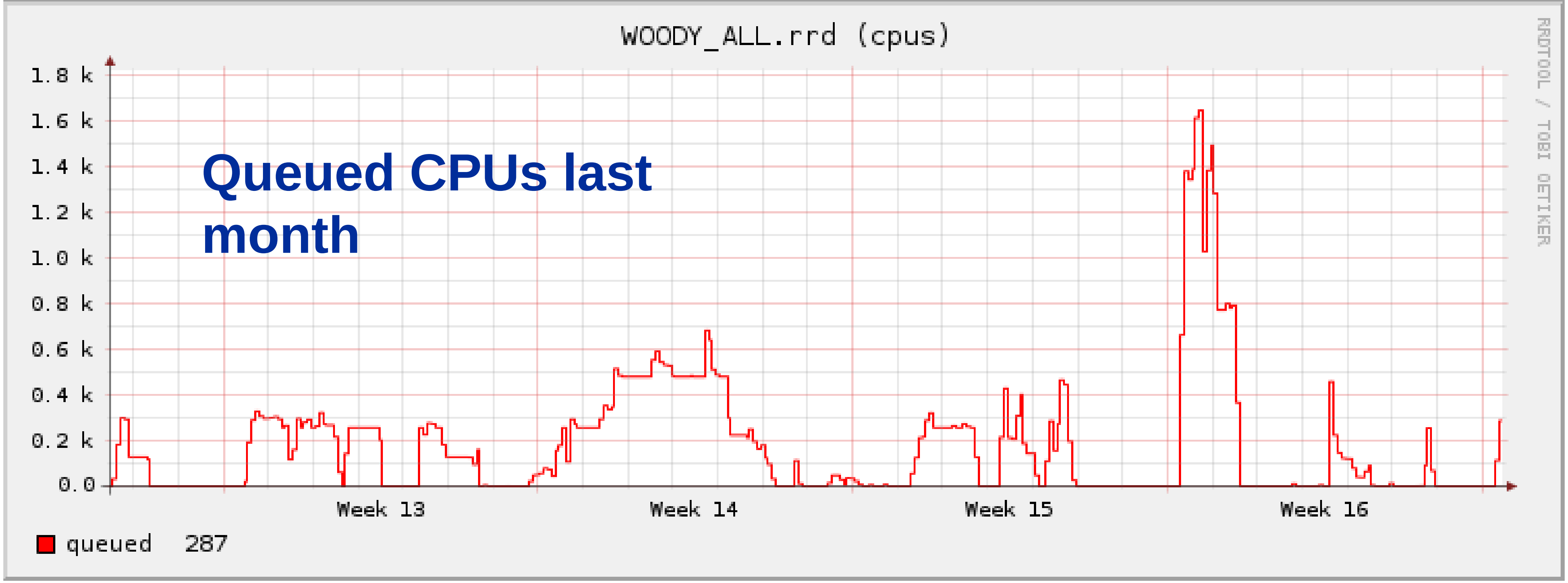
Woody: Queue configuration

- Very simple setup due to homogeneous hardware
- Queue: devel
  - 4 nodes reserved during working hours
  - Runtime 0 – 00:59:59
  - For batch and interactive testing
  - Open for everyone
- Queue: work
  - Runtime 01:00:00 – 24:00:00
  - Open for everyone

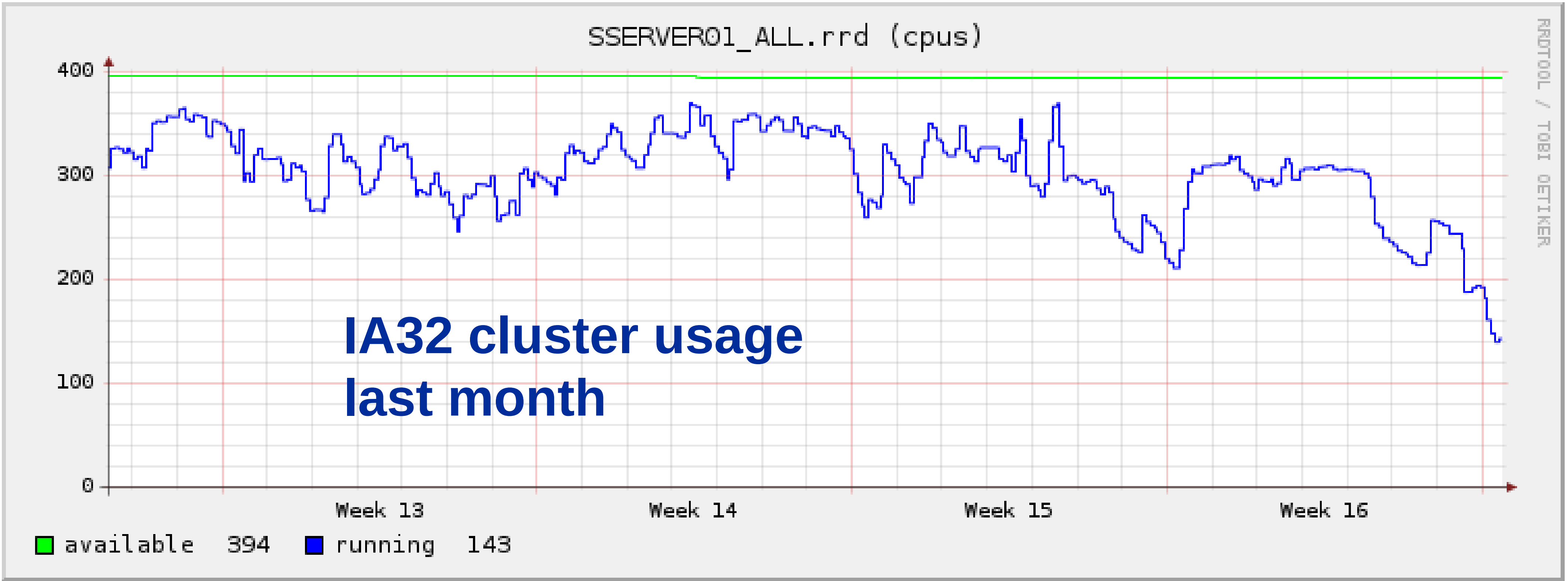
- Policies are enforced via scheduler config
  - E.g., increased fairshare target for special groups



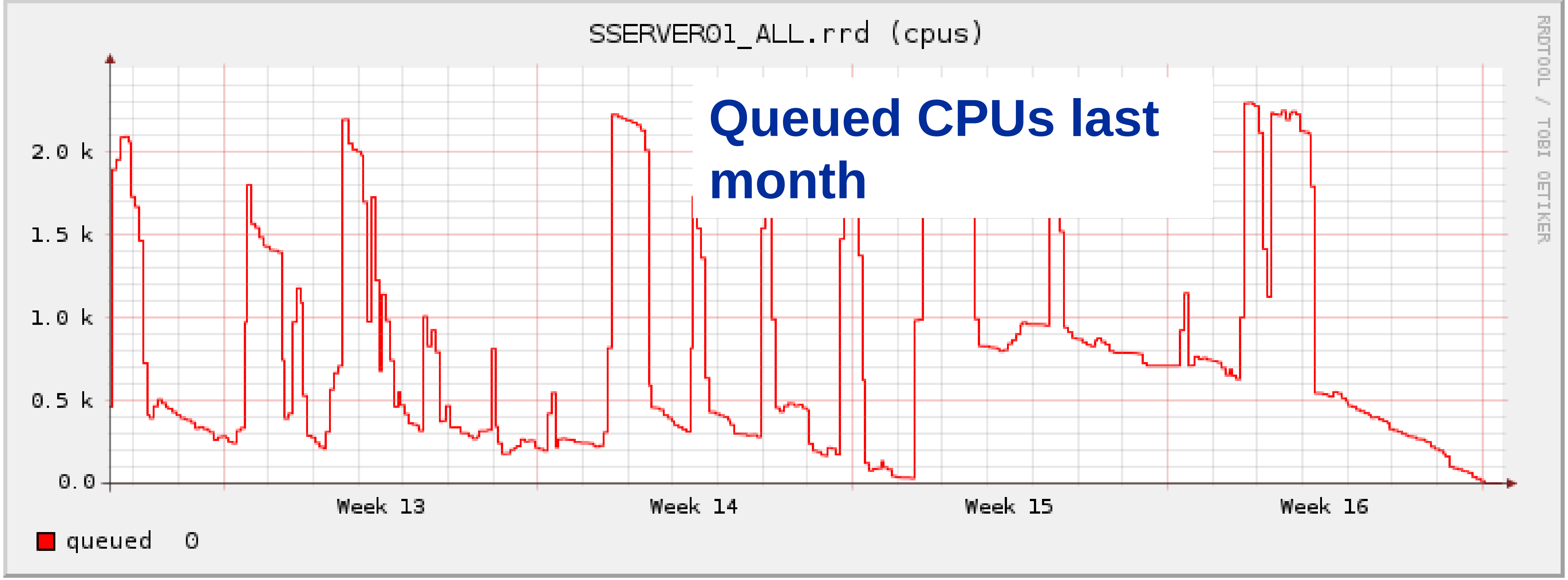
Woody usage last month



Queued CPUs last month

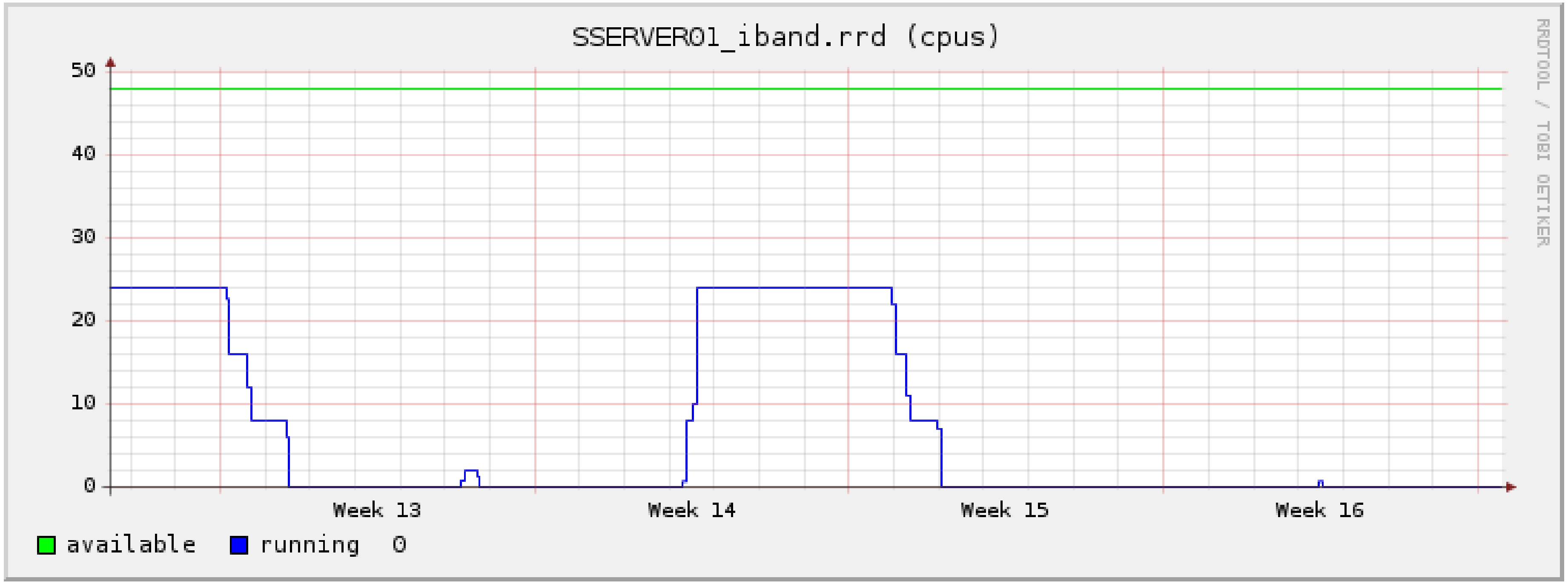


IA32 cluster usage last month



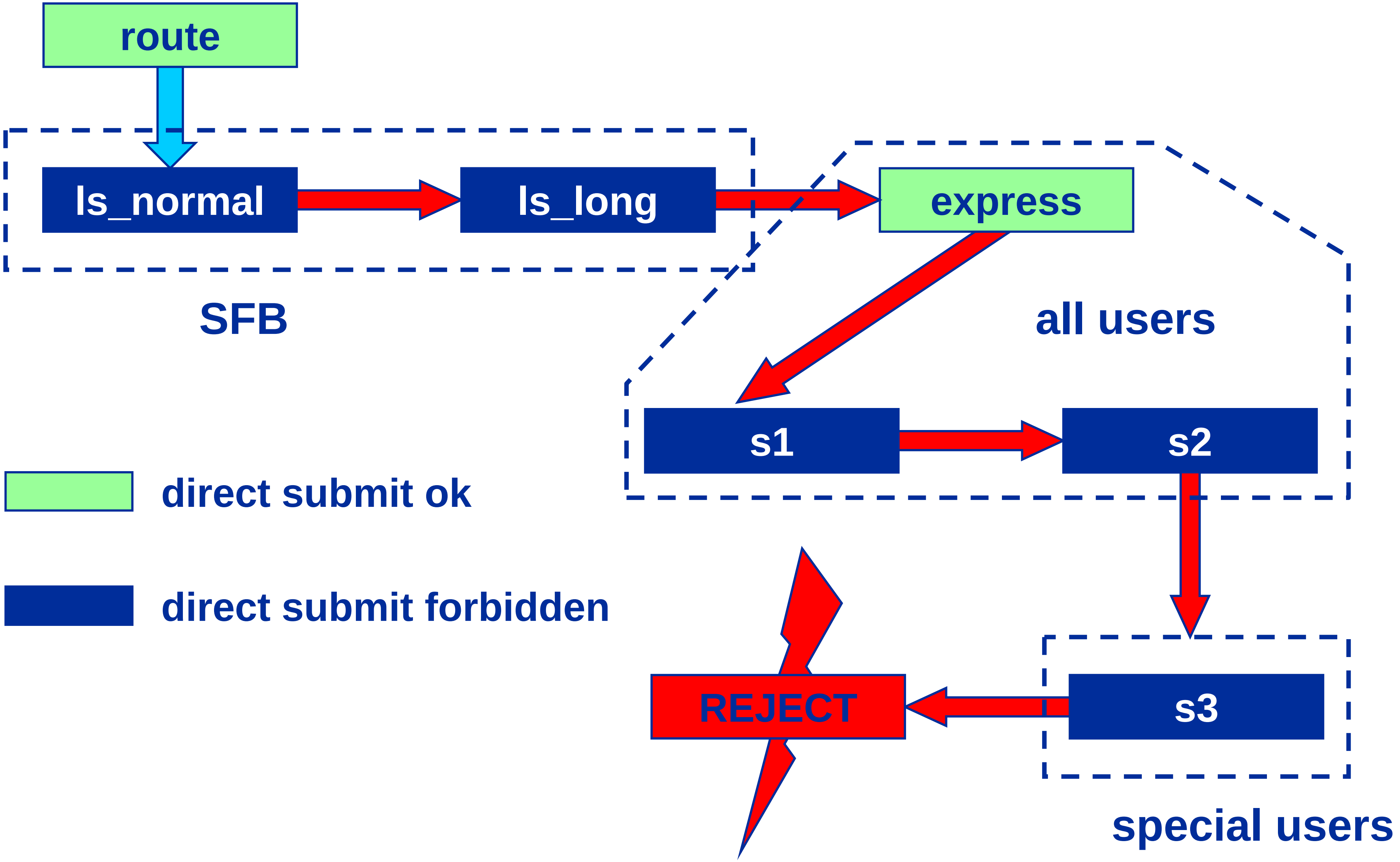
Queued CPUs last month

iband queue utilization (last month)



- **iband queue very thinly populated recently**
  - We suggest some changes in access policies to improve system utilization (see below)

IA32 Cluster: Batch as it was till today (excluding Opterons)



route

ls_normal

ls_long

express

SFB

all users

s1

s2

s3

REJECT

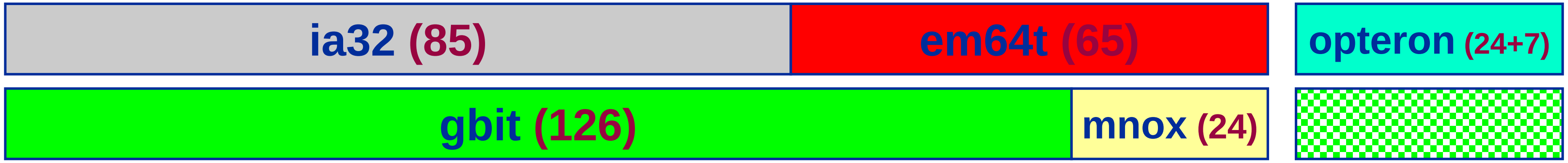
special users

🟩 direct submit ok

🟦 direct submit forbidden

IA32 cluster: Suggested changes for partitions and Queues

"node properties"



ia32 (85)

em64t (65)

opteron (24+7)

gbit (126)

mnox (24)

CPU-property	Network property	Node type	Up to now: access for	In future (starting 2.5.2007)	32-bit Exe	64-bit Exe
ia32	gbit	85x 32-bit Xeon 2 CPUs/node	expres, s1, s2, *s3, ls_normal ls_long*	express, s1, s2, *s3, ls_normal ls_long*	-xW	
em64t	gbit	41x 64-bit Xeon 2 CPUs/node	s1, *ls_normal, ls_long*		-xW, -xP	-xW, -xN, -xP
em64t	mnox	24x 64-bit Xeon 2 CPUs/node Infiniband	iexpress, s1, *iband*	iexpress, s1, iband	-xW, -xP	-xW, -xN, -xP
opteron	▨	Dual-Core Opteron 4 CPUs/node	oexpress, o2, *o3*	oexpress, o2, *o3*	-xW	-xW

→ **iband open for everyone** (alternative for opteron jobs?)

→ **express, s2, s3 can now also acccess 64-bit Xeons; if required, specify ":ia32" explicitly**

→ No changes with queue runtimes and max. processor numbers

IA32 cluster: Standard queues

	Submit-Queue	Run-Queue	Nodes	Runtime [HH:MM:SS]	min-max. CPUs/Job	MAX RUNNING	who?
ppn=1 / ppn=2	route / express	express	gbit ia32+em64t	≤ 01:00:00	1-8		all
	route	s1	gbit + mnox ia32+em64t	01:00:01 ≤ T ≤ 06:00:00	1-64		all
	route	s2	gbit ia32+em64t	06:00:01 ≤ T ≤ 48:00:00	1-64		all
	route	s3	gbit ia32+em64t	48:00:01 ≤ T ≤ 168:00:00	1-16	32 CPUs	on request
	route	ls_normal	gbit ia32+em64t	T ≤ 24:00:00	1-32		SFB
	route	ls_long	gbit ia32+em64t	24:00:01 ≤ T ≤ 240:00:00	1-8	80 CPUs	SFB
ppn=1 ppn=2	iband / iexpress	iexpress	mnox / em64t	≤ 01:00:00	1-32		all
	iband	iband	mnox / em64t	01:00:01 ≤ T ≤ 64:00:00	4-32		*all*
ppn=1 ppn=4	opteron	oexpress	opteron	≤ 01:00:00	1-32		all
	opteron	o2	opteron	01:00:01 ≤ T ≤ 24:00:00	1-32		all
	opteron	o3	opteron	24:00:01 ≤ T ≤ 48:00:00	1-32		WAP

IA32 cluster: Batch details

How do you specify IA32 or EM64T explicitly?

- **"node property" on job submit!**

  **Example: 32 bit executables → only IA32
  (new default: all GBit nodes):**

  ```
  qsub -l walltime=20:00:00,nodes=4:ppn=2:ia32 ...
  ```

  Example: Short job on EM64T only:

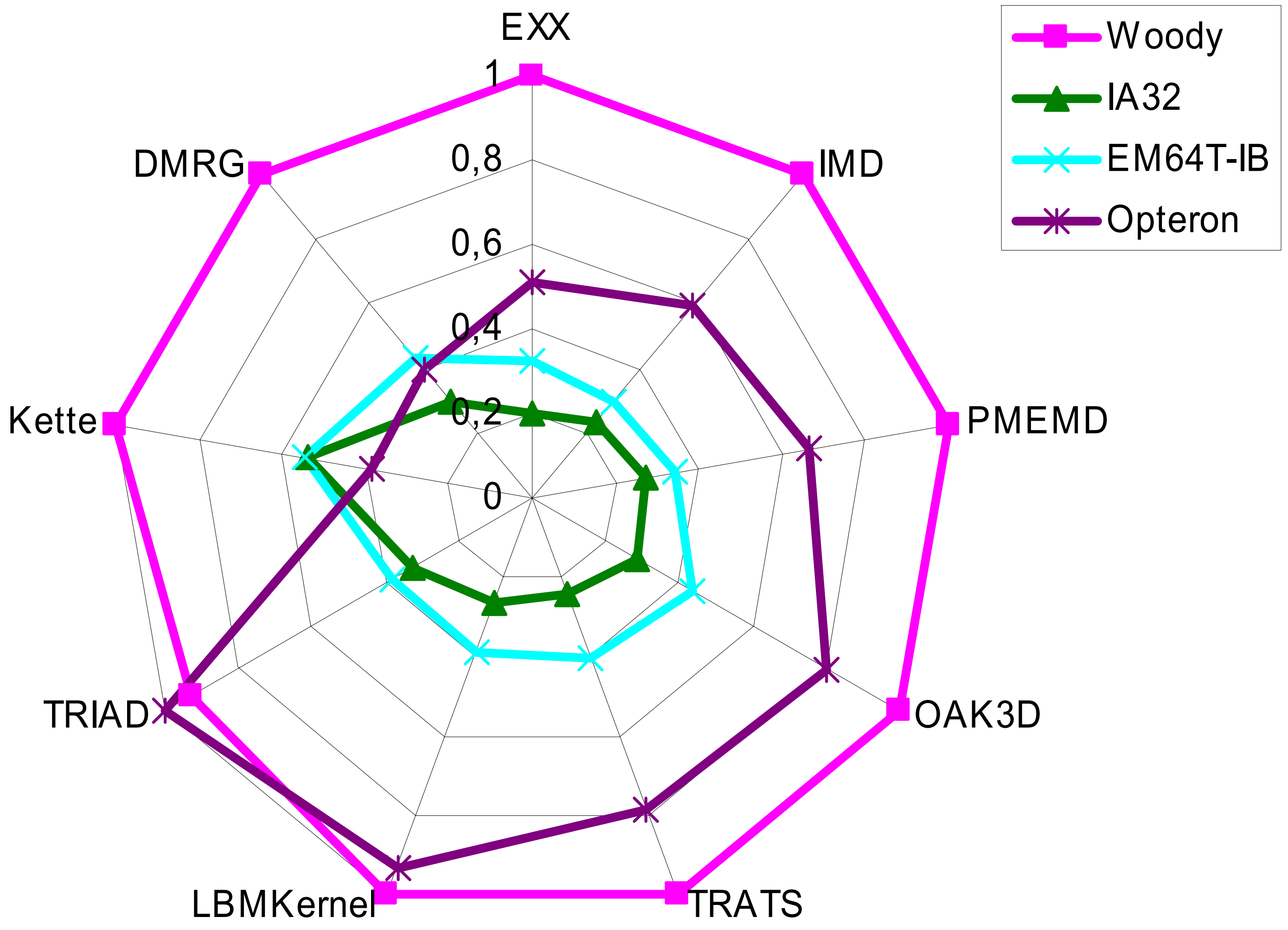
  ```
  qsub –l walltime=04:00:00,nodes=2:ppn=2:em64t ...
  ```

- **Caveat! Combining mutually exclusiuve properties can lead to blocking of jobs:**

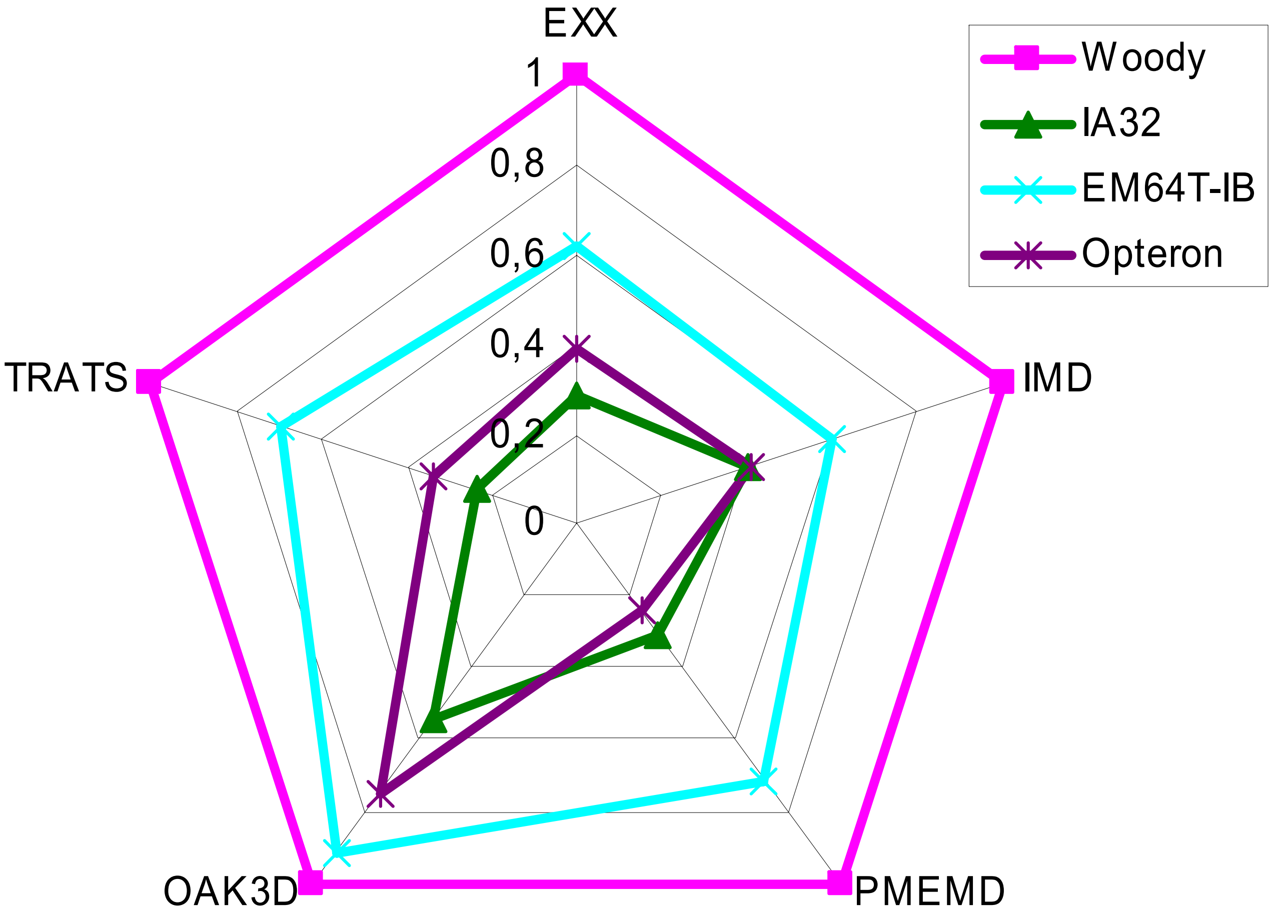
  ```
  qsub -l walltime=12:00:00,nodes=1:ppn=2:mnox …   ⇨   -q iband
  qsub -l walltime=12:00:00,nodes=1:ppn=4 …            -q opteron
  ```

Which System?
Performance and other arguments

Clusters: 1-node shootout



Clusters: Parallel benchmarks



Which system to choose?

	IA32 Cluster	Woody	SGI Altix	SGI Origin
Number of nodes / cores	85 / 170 41+24 / 130 24+7 / 124	188 / 752	1 / 30 1 / 12	1 / 28
Main memory	1 GB per CPU	2 GB per CPU	4 GB per CPU 2 GB per CPU	2 GB per CPU
Interconnect	GBit	Infiniband	Shared Memory	Shared Memory
Sequential throughput	+	— + if "quick" jobs	—	—
Trivial parallel	+	—	—	—
MPI parallel	(+)	+	+	+
Long running	+	—	—	+
Much memory	Opteron nodes	+	+	+
OpenMP	Opteron nodes	(+)	+	+
Development	+	+	+	++
New projects	+	+	+	—