

High Performance Computing

Infrastructures at DESY Hamburg

Sven Sternberger & Frank Schlünzen

Compute Infrastructures at DESY - Outline

- BIRD
 - Quick overview
- GRID, NAF, PAL
- HPC resources in Zeuthen
- HPC cluster
 - Hardware & Layout
 - Setup & Numa
 - Getting started
 - Monitoring & Operational model
 - Storage
 - FhGFS
 - Examples: Mathematica & Matlab



Compute Infrastructures at DESY - BIRD

➤ BIRD - Batch Infrastructure Resource at DESY



- See <http://bird.desy.de/info> (being updated)
- General purpose batch farm
- Getting access: via registry batch resource → UCO@desy.de
- Problems: UCO@desy.de; Bird.Service@desy.de

OS	# Hosts	Cores/Host	Total cores	RAM/Core	Total RAM
SLD6-64	17	4-12	168	8-48GB	0.5TB
SLD5-64	157	8-48	1520	8-126GB	4.1TB
Total	174	4-48	1688	8-126GB	4.6TB

- Environments for mpi or GPU jobs
 - High latency 1GE backbone
 - Jobs with heavy inter-process communication will suffer



Compute Infrastructures at DESY - BIRD



- GPUs are embedded into the BIRD farm

Host	OS	Hardware	GPUs	Drv / Cuda	Bird Resource	
It-gpu01	SL6-64	Intel Xeon 2.5GHz 16 cores 24GB ram	Fermi 2*M2050 448 cores, 3GB	Nvidia 304.51 Cuda 5.0.35	-l h=it-gpu01 -l gput=nvidiaM2050	
It-gpu02	SL5-64	AMD Opteron 870 2GHz 8 cores 16GB ram	Tesla 1*C870	Nvidia 270.41.19 Cuda 4.0	-l h=it-gpu02 -l gput=nvidiaC870	Will retire soon
It-gpu03	SL6-64	AMD Opteron 6128 2GHz 16 Cores 24GB ram	Fermi 2*C2070 448 cores, 6GB	Nvidia 304.51 Cuda 5.0.35	-l h=it-gpu03 -l gput=nvidiaC2070	Accelereys Jacket
It-gpu04	SL6-64	Intel Xeon 2.4GHz 24 cores 24GB ram	Fermi 2*M2070 448 cores, 6GB	Nvidia 304.51 Cuda 5.0.35	-l h=it-gpu04 -l gput=nvidiaC2070	
It-gpu05	Coming soon		Kepler K20 2496 cores, 5GB			

- Interactive login via qysh
 - ssh -XY pal; ini bird; qysh -l h=it-gpu03 -l gput=nvidiaC2070 ...
 - Graphical apps like matlab possible (xhost; DISPLAY)
- Non-exclusive access on dual-GPU nodes
 - Unless two gpus requested



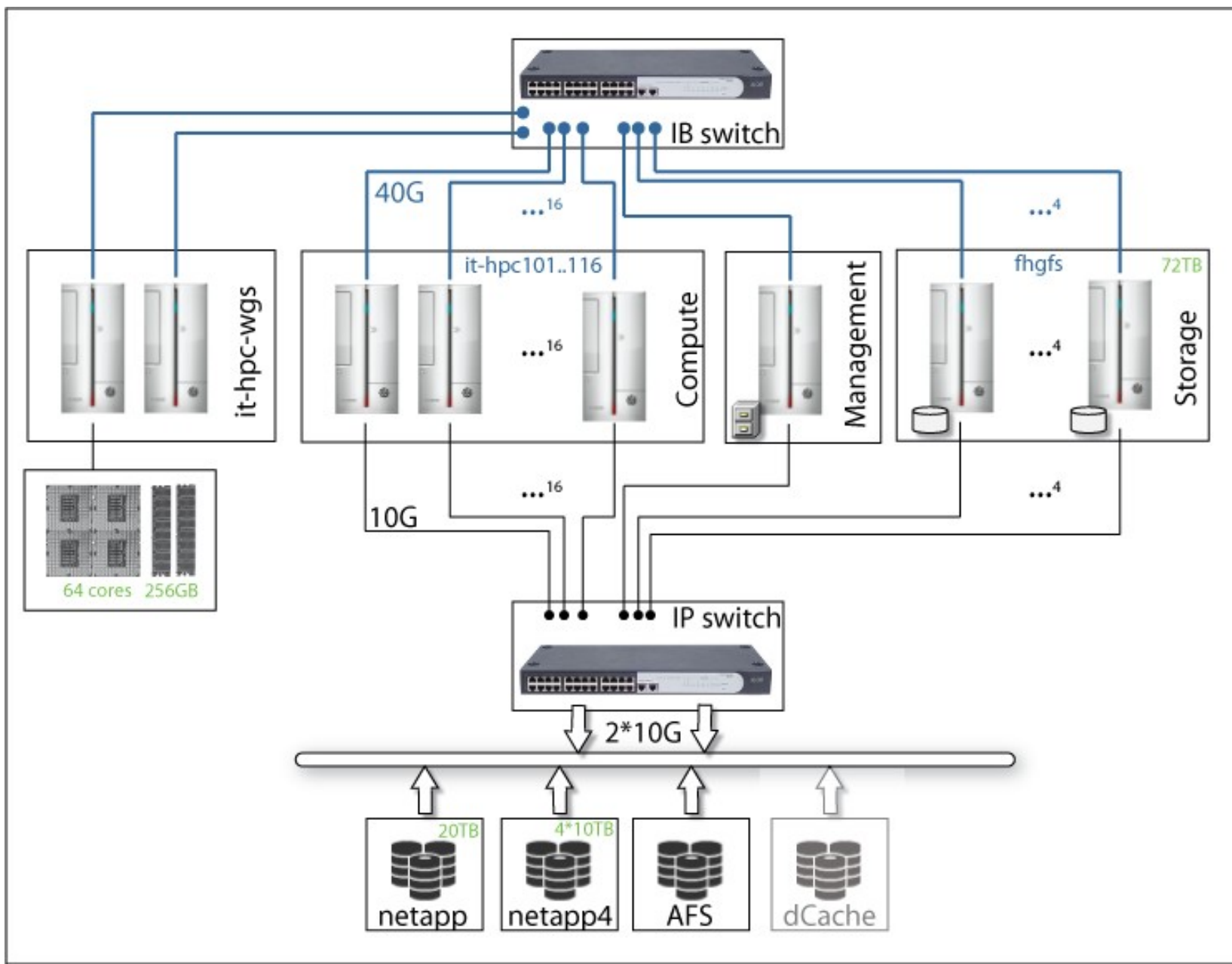
HPC cluster - History

- Zeuthen has a long HPC history
 - APE1 in 1987 → APEnext in 2005 (array processor experiment)
 - QPACE for lattice QCD simulations based on IBM PowerXCell 8i processors
 - GPE - GPU Processing Experiment (HadronPhysics2 project).
- DESY Hamburg
 - Has no history in HPC
 - Started in 2011 with retired (NAF) hardware.
 - 64*8 cores and distributed Fraunhofer filesystem (fhgfs)
 - 20GE Infiniband backbone (on an old and fragile voltaire switch)
 - Proofed to be rather useful, just the old hardware was dying rapidly.
 - Retired the old test-cluster last year and migrated to a new system → it-hpc cluster



HPC Cluster – Layout

1024 (+128) cores – 4TB (+0.38TB) memory – >100TB temporary storage
56G infiniband backbone – interlago PCI-E2



HPC Cluster – Hardware Decision

- Big Boxes vs. Pizza boxes
 - Cores per bucks
 - Scales also for non MPI jobs
 - High density in the rack
 - Simple cabling
- QDR vs. FDR
 - At the time no PCI3 systems on the market
 - Performance only slightly better
- QDR switch vs. FDR switch!
 - Better QDR performance
 - Sustainable
 - economically priced



Big Boxes w. QDR
Infiniband and FDR Switch

➤ Compute Nodes:

▪ **16 Dell PowerEdge R815**

- 4 CPU AMD Opteron(TM) Processor 6276 2.3GHz
 - Per CPU 16 Bulldozer Cores with 64 GB RAM
- 256 GB RAM
- Numa Architecture
- 10GE Ethernet
- QDR Infiniband (40GB/s)

▪ **2 PowerEdge C6145**

- 4 CPU AMD Opteron(TM) Processor 6276 2.3GHz
 - Per CPU 16 Bulldozer Cores with 48 GB RAM
- 192 GB RAM
- Numa Architecture
- 10GE Ethernet
- QDR Infiniband (40GB/s)



> Storage Nodes:

▪ 4 Dell PowerEdge R510

- 2 CPU Intel(R) Xeon(R) CPU E5503 @ 2.00GHz
 - Per CPU 2 Cores with 6 GB RAM
- 12 GB RAM
- QDR Infiniband (40GB/s)
- 20 TB / 12 Disc / Raid 6

▪ 1 PowerEdge 2950

- 2 CPU Intel(R) Xeon(R) CPU E5345 @ 2.33GHz
 - Per CPU 4 Cores with 4 GB RAM
- 8 GB RAM
- Numa Architecture
- 1 GE Ethernet
- QDR Infiniband (40GB/s)

HPC Cluster – Hardware

InfiniBand is a switched fabric communications link used in high-performance computing and enterprise data centers.

Its features include high throughput, low latency, quality of service and failover, and it is designed to be scalable. The InfiniBand architecture specification defines a connection between processor nodes and high performance I/O nodes such as storage devices.
(wikipedia)

> Infiniband Infrastructure:

- Mellanox SX6036 InfiniBand/VPI Switch Systems
 - 36 Ports
 - 56 GBIT/s FDR
 - Managed
- Mellanox Connect-X3 QDR Cards
 - Unfortunately the interlagos does not support PCI gen 3
 - And the PCI gen 2 bus is a slow one ...



HPC Cluster – Setup/Optimization

> Setup

- Scientific Linux 6.3 (always updated)
 - No automatic updates → maintenance
 - No salad/yum/cron jobs
 - Xymon monitoring
- AFS/Kerberos
- IPoIB configuration

> Only some optimizations

- Bound infiniband card to core
- Switch off timer interrupt
- Unlimited locked-in-memory address space



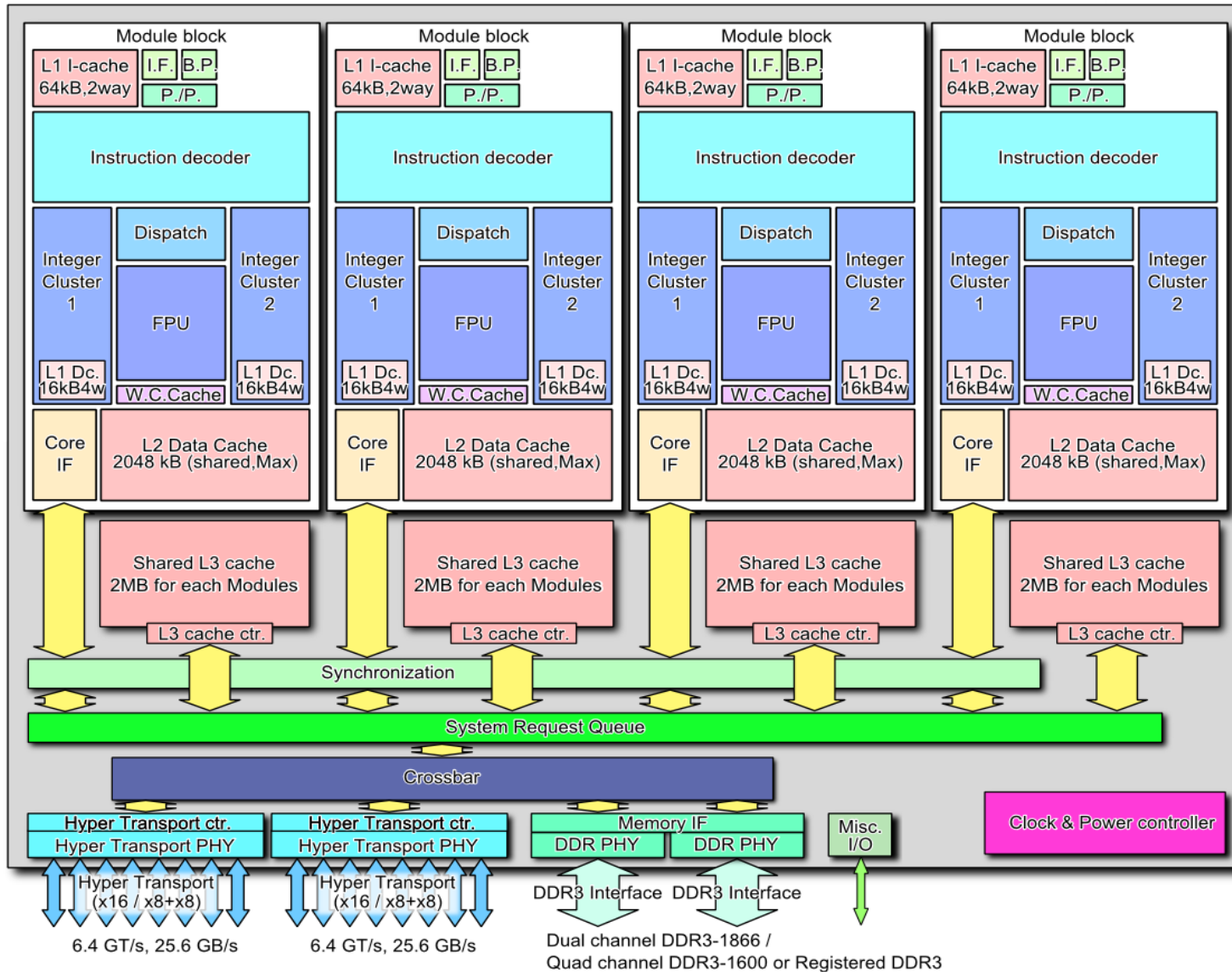
HPC Cluster – Numa Hardware

Non-Uniform Memory Access (NUMA) is a computer memory design used in multiprocessing, where the memory access time depends on the memory location relative to a processor. Under NUMA, a processor can access its own local memory faster than non-local memory (memory local to another processor or memory shared between processors). (wikipedia)

- All compute nodes are numa machines
- To gain the full potential of the nodes you should consider the architecture
- Tools like likwid are helpful (necessary?)
 - Examine Numa topology
 - Profile software on Numaarchitecture
 - Bind processes to cores
- Multicore CPU make it even more complicated
 - Shared caches
 - Shared FPU
 - Different Levels of aggregation
- Commercial software is not always numa aware



HPC Cluster – Numa Hardware



HPC Cluster – Numa Hardware

Socket 0:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB
2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB
6MB								6MB							

Socket 1:

32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB
2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB
6MB								6MB							

Socket 2:

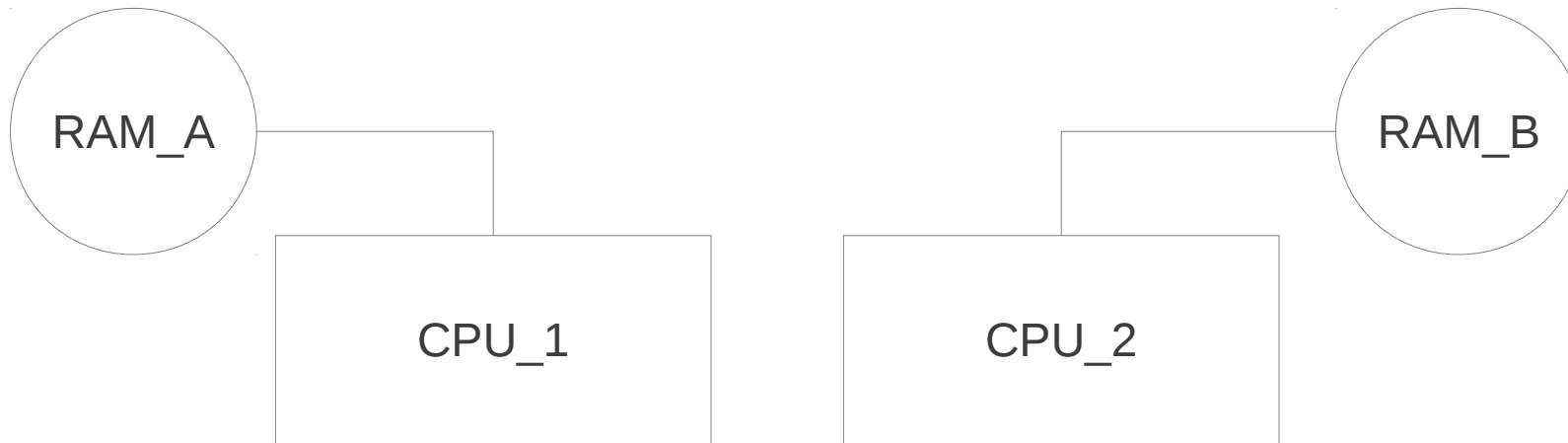
48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB
2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB
6MB								6MB							

Socket 3:

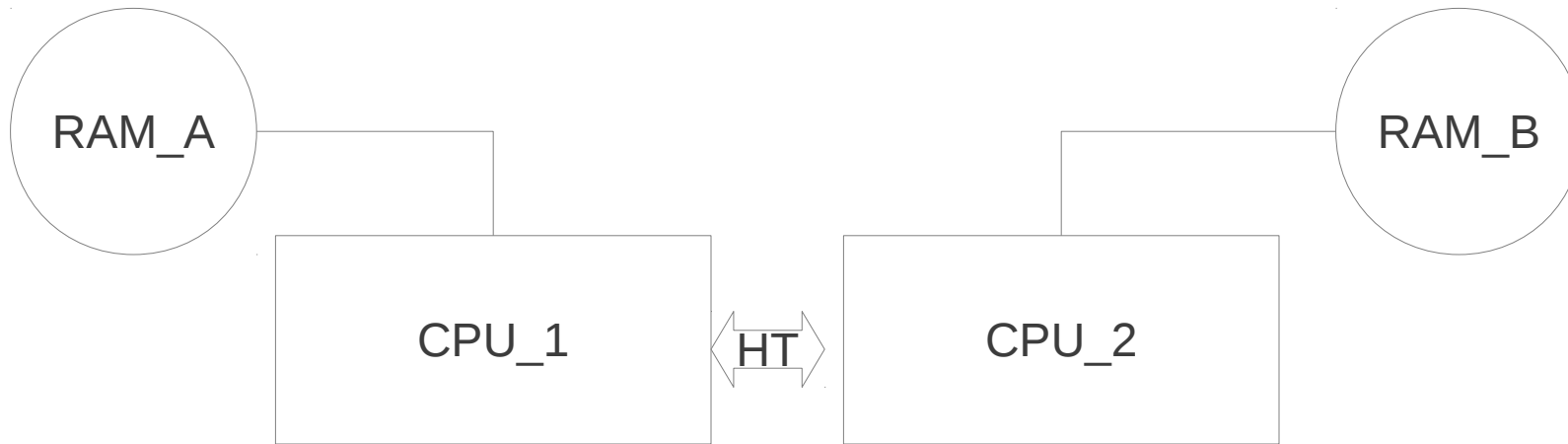
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB	16kB
2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB	2MB
6MB								6MB							



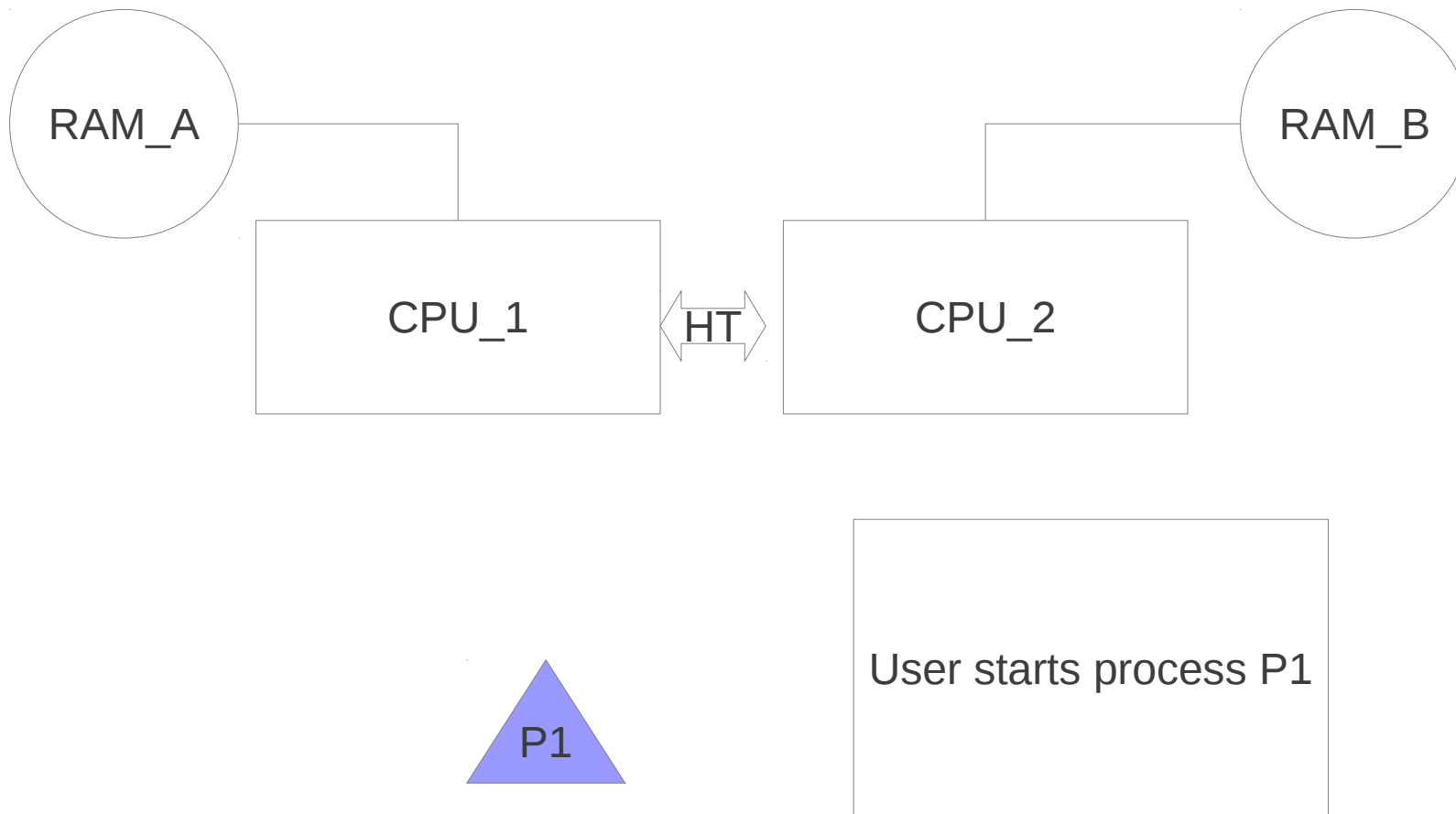
HPC Cluster – Numa Hardware



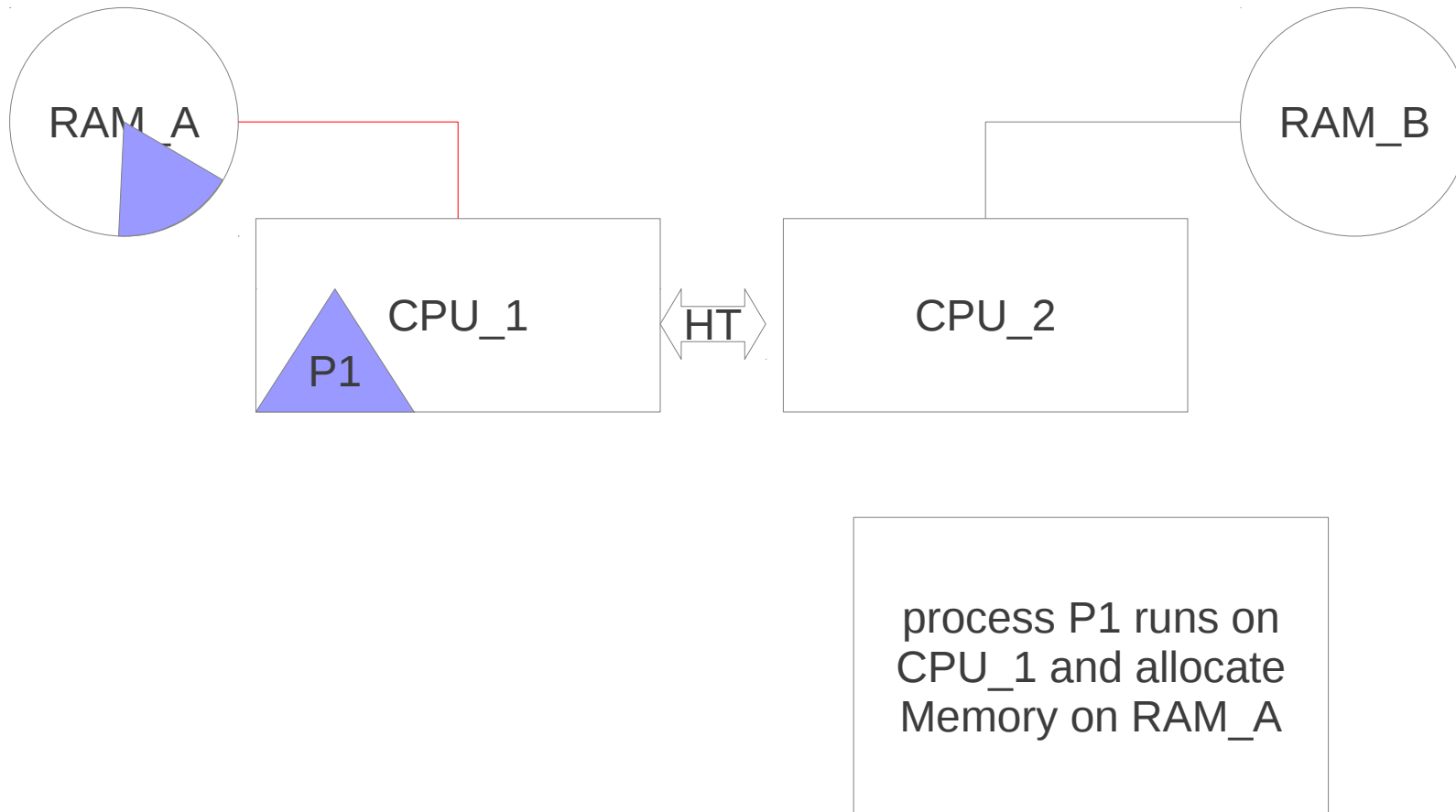
HPC Cluster – Numa Hardware



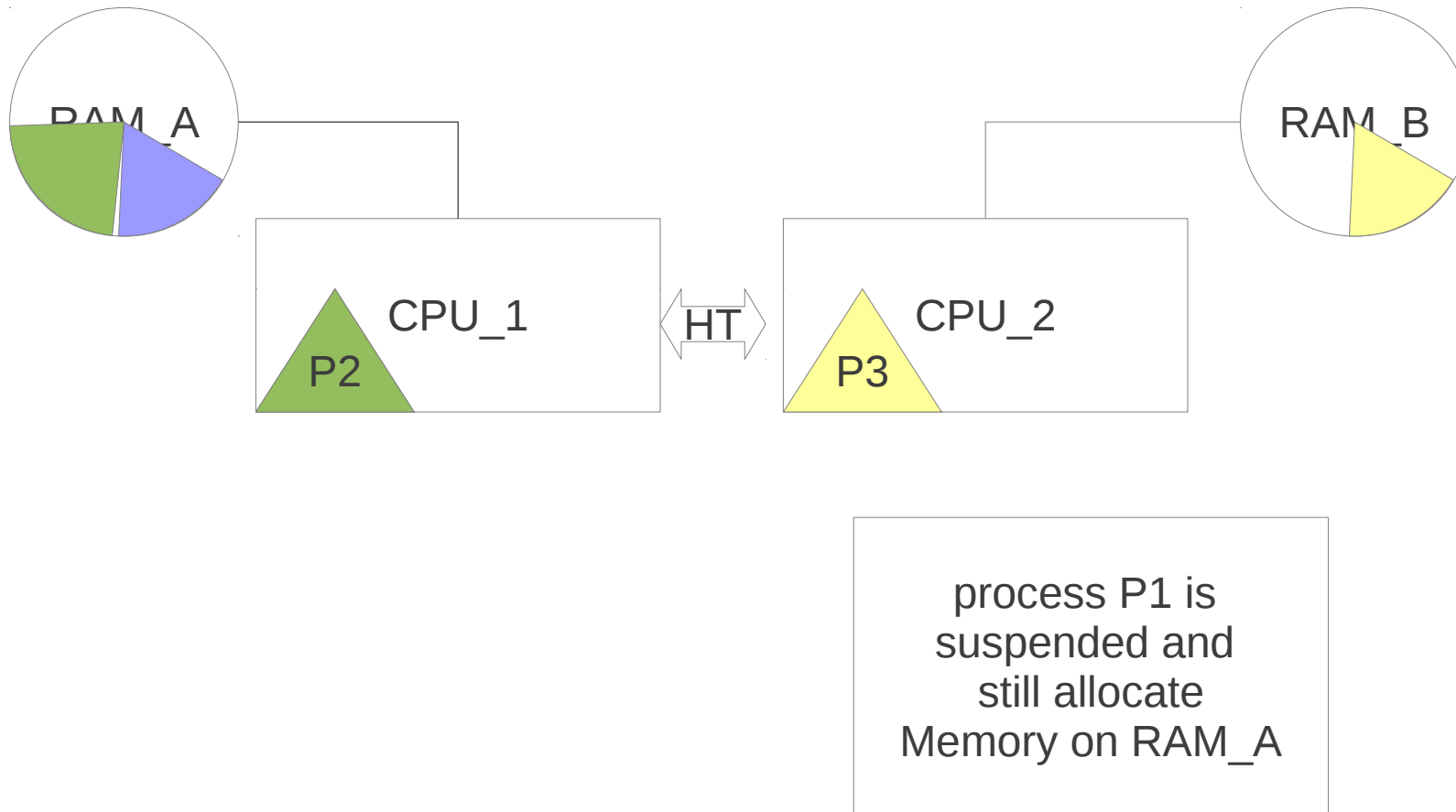
HPC Cluster – Numa Hardware



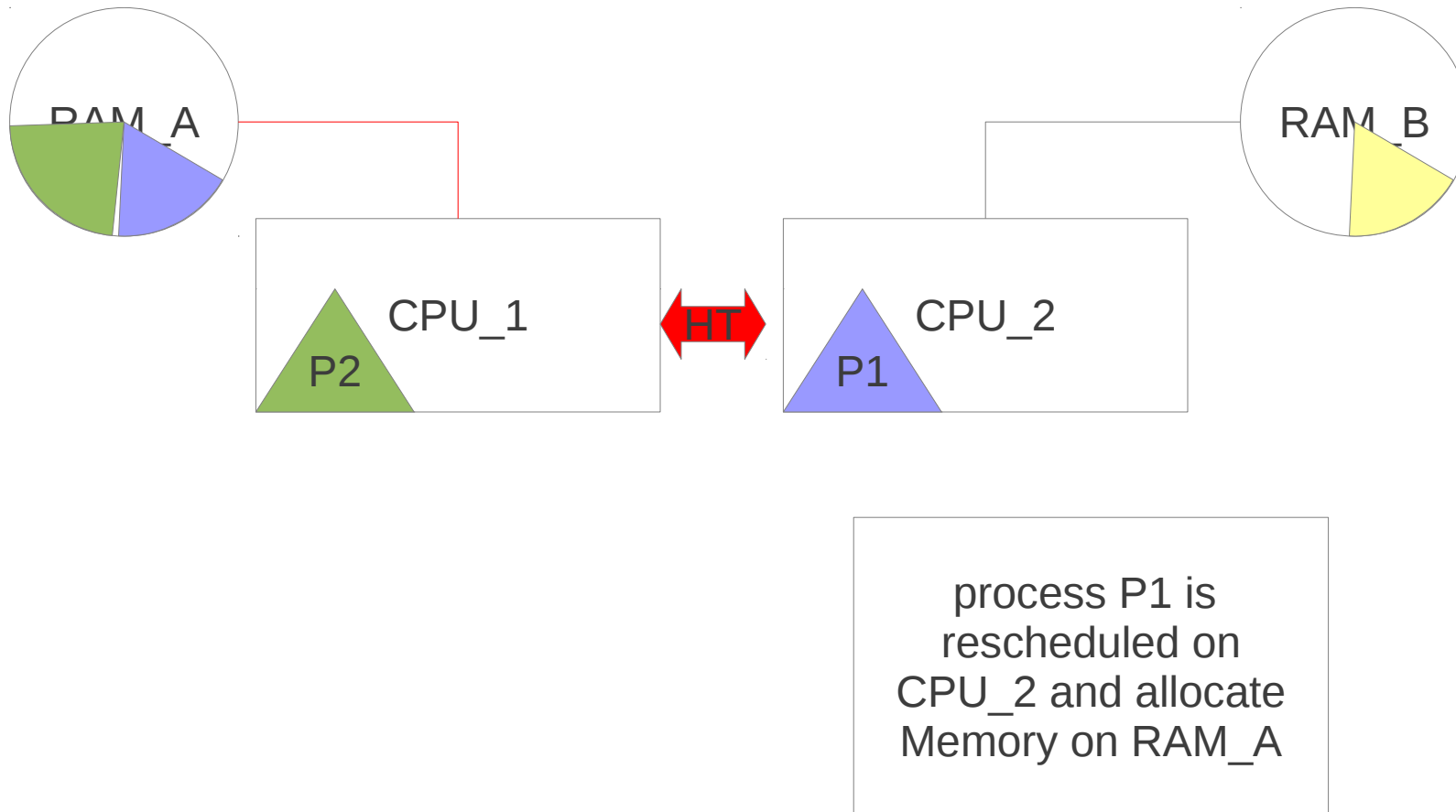
HPC Cluster – Numa Hardware



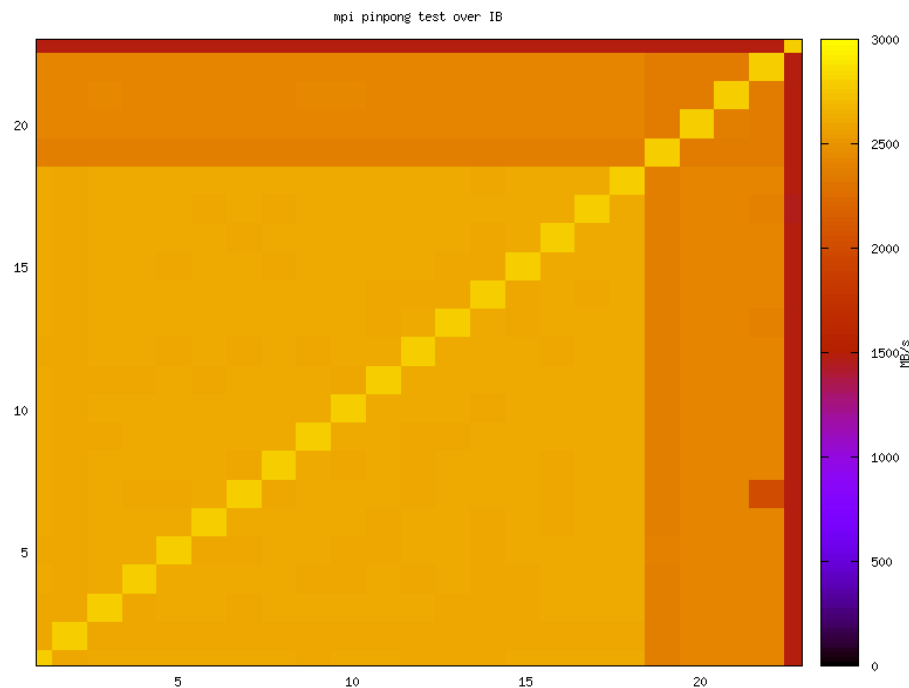
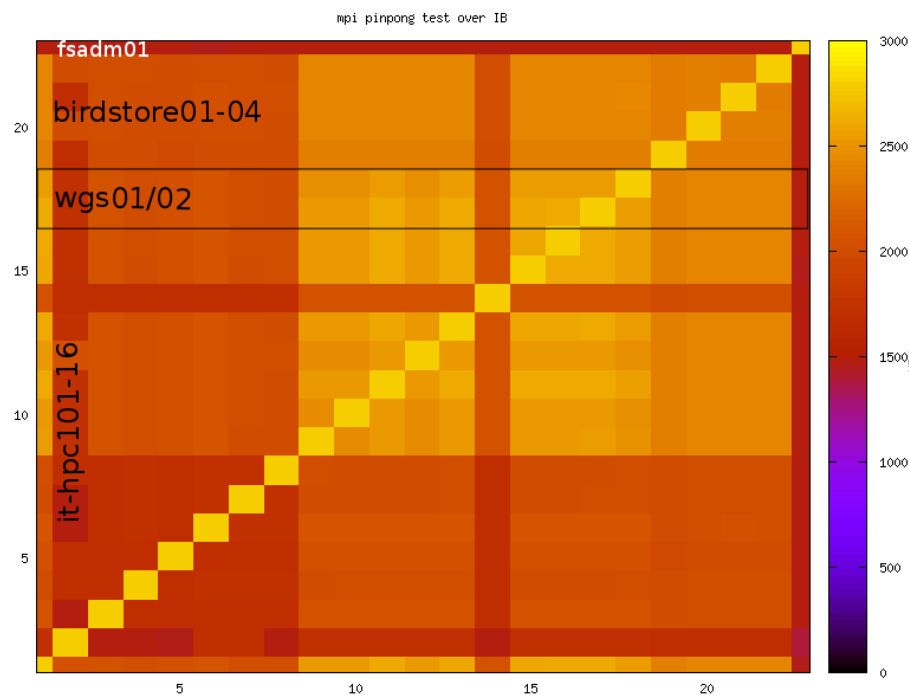
HPC Cluster – Numa Hardware



HPC Cluster – Numa Hardware



Simple example - binding sockets and IB adapter



Using the HPC cluster



HPC Cluster – First steps

> Pre-requisites:

- Multi-core capable applications
- embarrassingly parallel problems → Bird or Grid (if an option)
- 32/64 cores per node with 8/4GB per core → prefer $\leq 8\text{GB/core}$
 - Single core performance of AMD interlagos not so great
 - If you need lots of memory/core different architectures might be favourable
 - Could be added to it-hpc if that's a common requirement

> Uncertain your application is suitable?

- Get in touch!
- Run tests on it-hpc-wgs!
 - it-hpc-wgs are freely accessible.
 - Identical configuration like it-hpc(xxx), but slightly less memory (192GB)



HPC Cluster – Typical Applications

- Laser Wakefield simulations (e.g. osiris, genesis)
- Quantum Chemistry (e.g. gamess)
- Fluid Dynamics (e.g. OpenFOAM)
- X-ray spectroscopy simulations (e.g. fdmnes)
- X-ray photon correlation spectroscopy (e.g. xpcsgui; matlab)
- Molecular dynamics (e.g. namd)
- Structural and dynamics analysis (e.g. ansys)
- “General purpose” (Mathematica, Matlab)



HPC Cluster – First steps

> Gaining access:

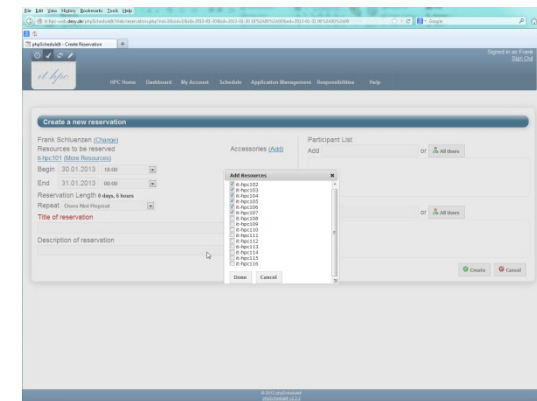
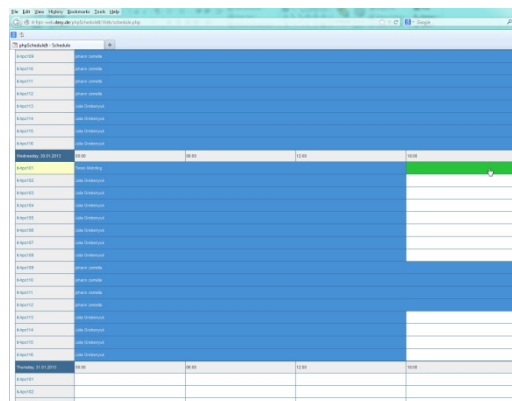
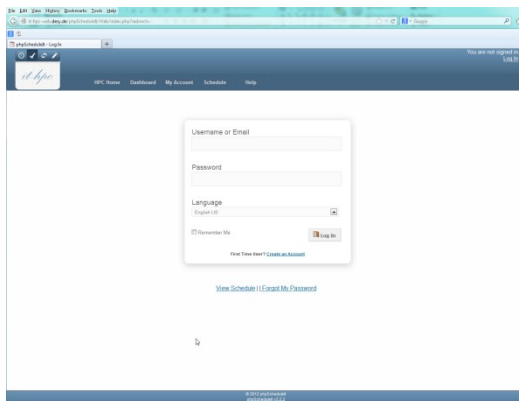
- Access is restricted to „registered“ users
 - No restrictions on it-hpc-wgs!
- Send mail to unix@desy.de or uco@desy.de asking for it-hpc „resource“
 - Let us know what kind of application you intend to run
 - Helps us providing essential dependencies
 - Helps us to plan future expansions



HPC Cluster – First steps

➤ Scheduling:

- Create an account on the web-calendar:
 - Use your DESY account name (not the password!)
 - Currently not kerberized.
- Reserve resources
 - Co-operative model, no fair-share
 - Reservation for full nodes – no sharing of hosts
 - Currently fixed 1h time slots (can be adjusted if needed).
 - Don't reserve all nodes for a long time!
 - We do some accounting, if it gets too much out of balance, the model might be adjusted.



> Why such a model?

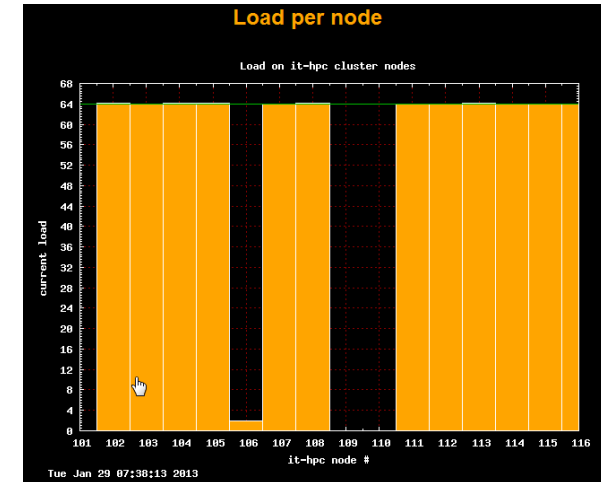
- Most users have at best a very rough idea about resource requirements
 - Usually no idea how an application scales
 - Usually no idea on memory consumption
- Only booking of complete nodes (not single core)
 - Many apps do not support affinity bindings (e.g. matlab on non-intel)
 - Avoid interference between users and applications
- Most users/apps are not capable taking snapshots
 - Start from scratch if a job fails
 - Keep jobs running as long as possible
- Many users know well in advance when they'll need compute resources in the future
 - Scheduled beamtime for example
- Some application are difficult to integrate in an batch system (matlab)
- Easy to maintain compared with a batch system (SGE, slurm ...)



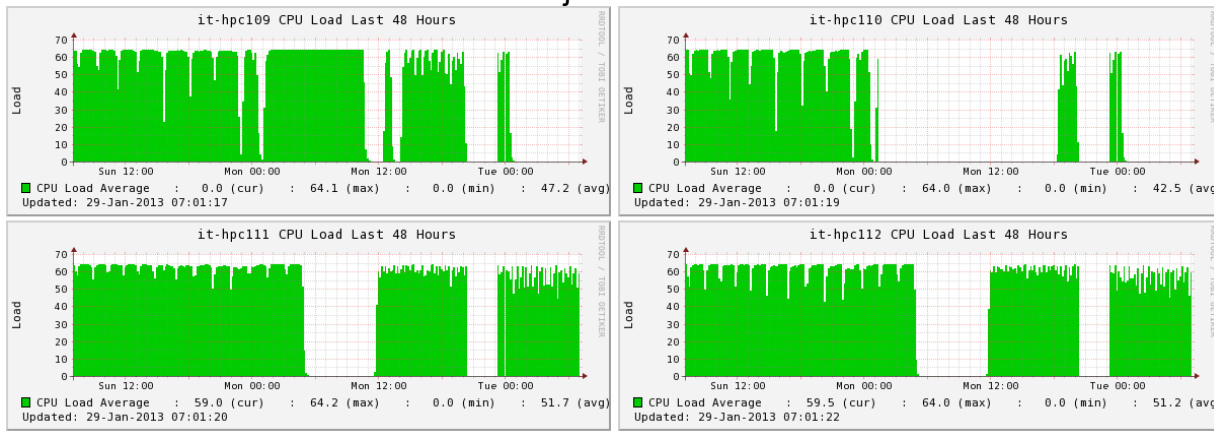
HPC Cluster – Monitoring

> Reservations:

- Reservations are meanwhile enforced
 - No access without reservation
- Jobs running into slots without reservation create an alert
 - But will continue running
- Jobs conflicting with a reservation will be terminated
 - Jobs exceeding the reservation will be terminated as well.
 - IF there is activity from the user with current reservation
- Over-subscription creates an alert
 - Load > 64 will slow everything down
 - No termination of jobs.



5' update cycle from xymon



HPC Cluster – Running jobs

> Scientific SW & Compiler usually installed in /opt/<app>/<version>

- idl, intel, maple, mathematica, matlab, pgi
- AFS-hosted; but no token required
- Can be changed to local installations if needed
- Own software: best on /data/netapp or /afs

• Various MPI implementations

- openMPI-1.5.x and openMPI-intel-1.5.x
- mpich2, mvapich2, mvapich2-intel
 - Thanks to Götz!
- pvm works as well...

> Compiler

- gnu 4.4.x, pgi-2012/13, intel-2011/13
- gnu 4.7.x in /data/netapp/it/tools permits interlagos optimization (`-march=bdver1`)
- Be careful with intel's optimization (seems a common pitfall)
 - Floating point model defaults to „*value unsafe optimization*“
 - Intel recommends `-fp-model precise -fp-model source` to be on the safe side
 - Other apps (e.g. matlab) have similar “issues”



HPC Cluster – Running jobs

> Initialize

- `ini` deprecated on SL6, use `module` instead
 - `module avail` to display
 - `module load intel.2011`
 - `module load mathematica`
 - `module load openmpi-x86_64` etc
- For openmpi propagate the prefix
 - `mpirun --prefix /usr/lib64/openmpi ...`
 - or: `eval `modulecmd bash load openmpi-x86_64`` in e.g. `.zshrc`

• Running

- `/usr/lib64/openmpi/bin/mpirun --prefix /usr/lib64/openmpi -mca orte_base_help_aggregate 0 -mca btl_openib_warn_default_gid_prefix 0 -np 1024 -hostfile mpi.hosts <app>`
- Remember to assemble the hostfile corresponding to your reservation (addon to booking?)
- Start the jobs preferably on it-hpc-wgs
- Where to store data?



HPC Cluster – Using storage

> AFS

- Secure, low volume storage.
- TSM Backup possible
- Global File System. Accessible from everywhere
- Good for final results.
- ACL for fine grade right management
- Keep in mind: no automatic token extension, difficult in long running jobs (> 24h)
- Not optimal for HPC cluster workload
 - Latency
 - Small files
 - Large data volumes



HPC Cluster – Using storage

> NFS (ver. 3)

- /data/netapp
- 20 TB at the moment without quota
- Secure storage
- Good for long term storage
 - larger volumes of final results
 - Store group specific software
 - Fast user/group home directories
- Per default not accessible from outside it-hpc
- Backup possible
- Cooperative storage



HPC Cluster – Using storage

> dCache

- High performance distributed storage
- Possible connection to tertiary Storage Systems (Tape robot)
- Currently not mounted, but could be mounted via NFS4.1
- Let us know if that's a requirement



HPC Cluster – Requirements for cluster storage

- Low latency
- Free to use
- Easy to install and manage
- Robust and community supported
- Infiniband support



HPC Cluster – Using storage

FHGfs is used in the it-hpc cluster as *embedded* storage, it is not a strategic solution for our customer, so will not offer this outside this cluster as a service.

> FHGFS

- /data/fhgfs
- 72TB without quota
- Low latency, fast read and write
- Data transfer via Infiniband
- Good for short term temporary storage.
 - Interprocess communication
 - Staging data files
- No export outside the it-hpc cluster
- No backups, no protection
- Cooperative storage (otherwise we would delete chronology)



HPC Cluster – Using storage

➤ FHGFS Key benefits

- **Distributed File Contents and Metadata**
 - strict avoidance of architectural bottle necks.
 - distribution of file system metadata (e.g. directory information)
- **HPC Technologies**
 - scalable multithreaded core components
 - native Infiniband support
- **Easy to use**
 - FhGFS requires no kernel patches
 - graphical cluster installation and management tools
- **Increased Coherency** (compared to simple remote file systems like NFS)
 - changes to a file are always immediately visible to all clients.



HPC Cluster – Running mathematica

> Licenses available

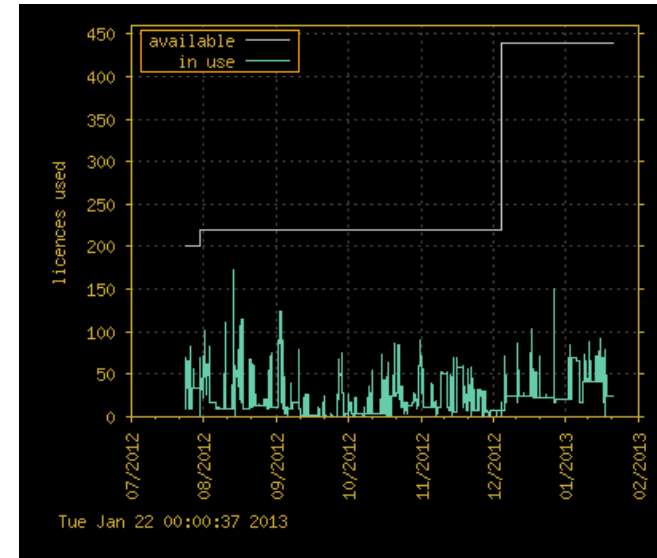
- 440 SubKernel licenses occupying a single main-license
- Usually a substantial number of licenses available

> Mathematica lightweight-grid

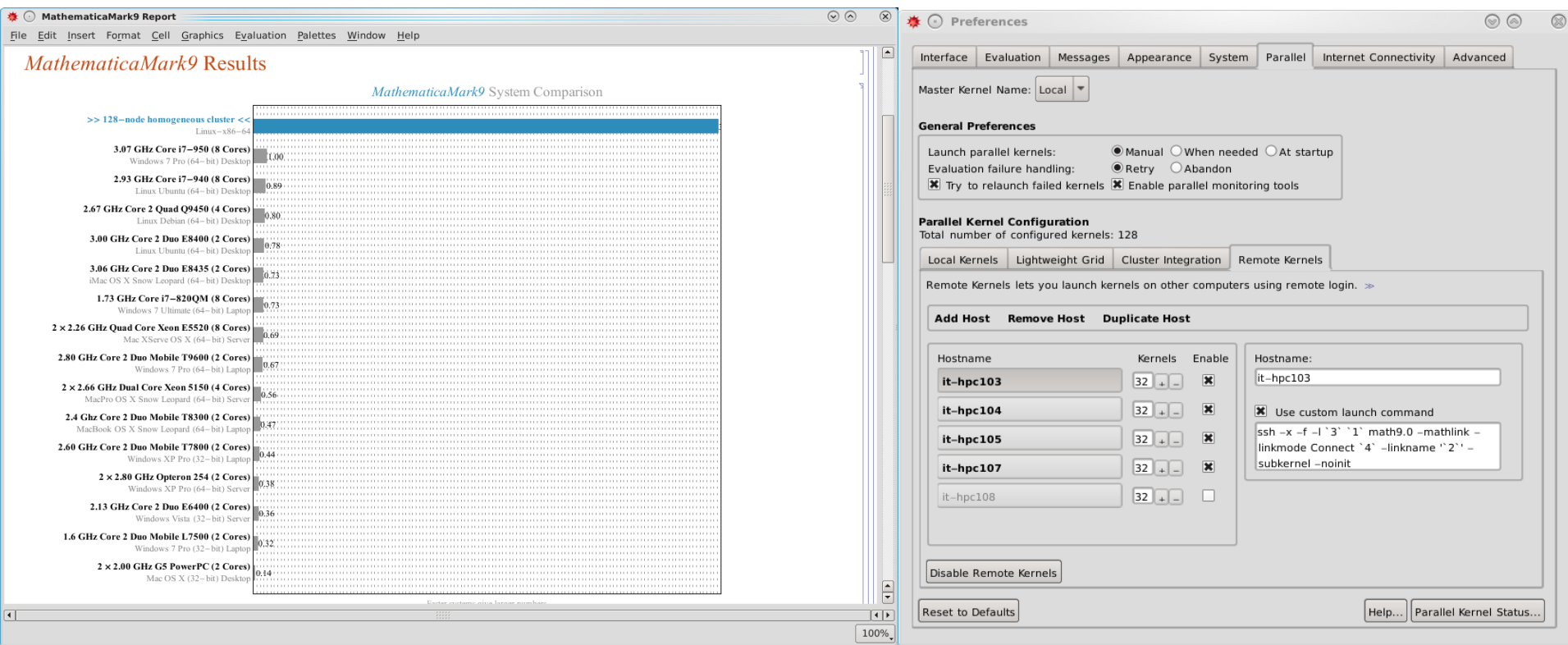
- Requires startup of tomcat server
- Currently not an option on the cluster
 - Could be added to reservation tool

> Parallel kernel configuration

- Easily done from within the application
- Unmanaged
- Requires some care selecting the number of kernels per host.



HPC Cluster – Running mathematica



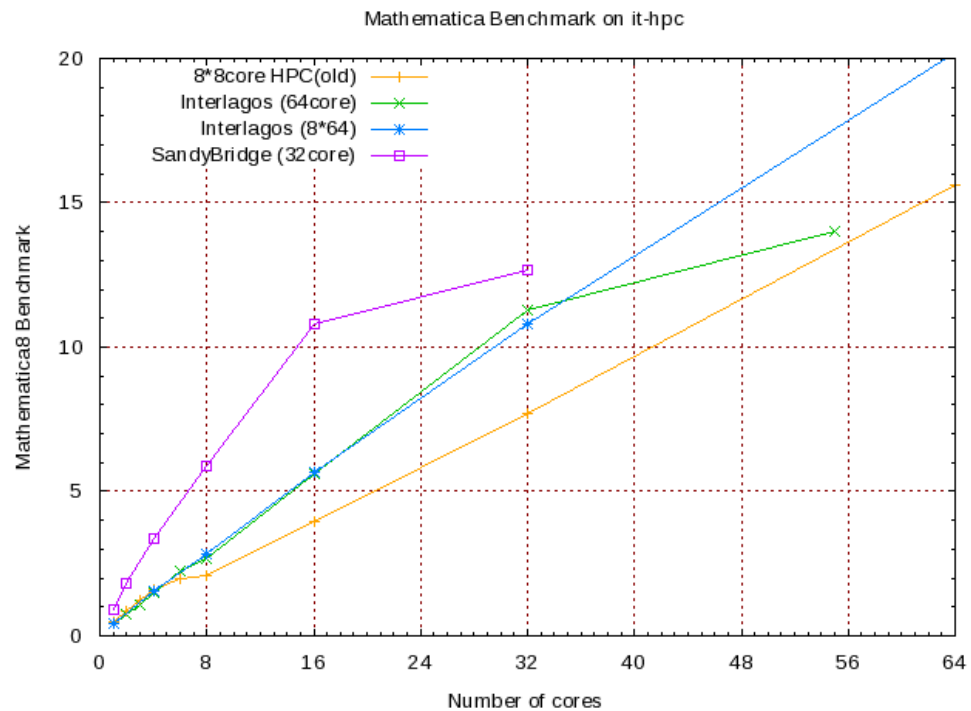
Mathematica benchmark with 128 cores on 4 nodes ...



HPC Cluster – Running mathematica

➤ Mathematica benchmarks

- Scales well up to 32-bit cores, gets worse beyond that
 - But true for other architectures as well
- Mathematica has troubles acquiring much more than 50 cores
- Using up to 32-48 cores per node should be fine.



HPC Cluster – Running matlab

➤ New: Matlab Distributed Computing Server

- **256** kernel licenses → 256 core job with a single matlab / distributed computing toolbox

- Currently need to start the matlab scheduler as root

- No automatic startup yet. If frequently used, could be added that to reservation tool.

Admin Center

File Hosts MJS Workers Help

Hosts

Add or Find ...

	Host		Cores	Status	Up Since	Name	Workers
	Hostname	Reachable					
Start mdce Service...	it-hpc-wgs01	yes	64	running	2013-01-29 15...		32
Stop mdce Service...	it-hpc-wgs02	yes	64	running	2013-01-29 16...	hpc-test	32

Test Connectivity...

MATLAB Job Scheduler (MJS)

Start...

	Name	Hostname	Status	Up Since	Workers
Stop...	hpc-test	it-hpc-wgs02	running	2013-01-29 16:02	64

Resume

Workers

Start...

	Worker				Connection	MJS	
	Name	Hostname	Status	Up Since		Name	Hostname
Resume	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02
	it-hpc-wgs01_worke...	it-hpc-wgs01	idle	2013-01-29 1...	connected	hpc-test	it-hpc-wgs02

Last updated: Tue Jan 29 16:19:35 CET 2013

Update

every 2 minutes

Update Now



HPC Cluster – Running matlab

➤ Once the scheduler is running, job submission from a desktop is easy:

The left screenshot shows the 'Cluster Profile Manager' window with the 'HPC' cluster profile selected. The 'Properties' tab is active, showing the following details:

- Description of this cluster: HPC
- Hostname of the machine where MJS is running: it-hpc-wgs02
- The name of the MJS: hpc-test
- User name for MJS access: current user (default)
- License number: <none>

The right screenshot shows the 'Validation Results' tab for the 'HPC' cluster profile. The overall status is 'Passed'. The following table shows the results of various tests:

Stage	Status	Description
Cluster connection test (parcluster)	Passed	
Job test (createJob)	Passed	
SPMD job test (createCommunicatin...	Passed	
Pool job test (createCommunicatingl...	Passed	
MATLAB pool test (matlabpool)	Passed	

```
>> myCluster = parcluster; matlabpool(myCluster);  
>> myrun = matlabpool('open','HPC','64','AttachedFiles',{nb1.m,nb2.m});
```



HPC Cluster – Running matlab

The screenshot shows the MATLAB R2012b interface. The Command Window displays the following output:

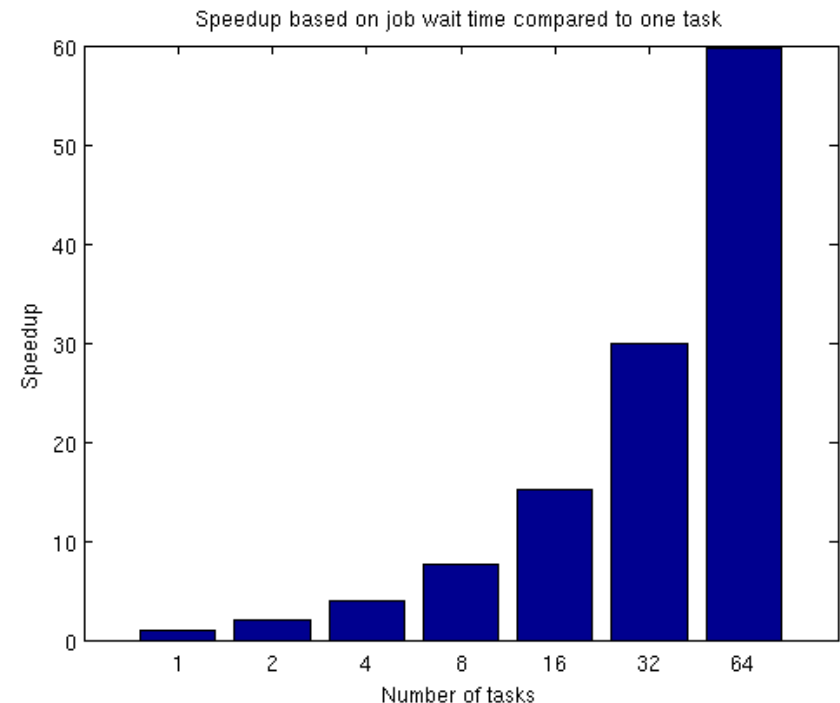
```
>> myCluster = parcluster;  
>> parallel_demo_distribjob_bench;  
Starting weak scaling timing. Submitting a total of 35 jobs.  
Job wait time with 1 task(s): 54.477931 seconds  
Job wait time with 2 task(s): 55.342876 seconds  
Job wait time with 4 task(s): 56.338393 seconds  
Job wait time with 8 task(s): 56.549158 seconds  
Job wait time with 16 task(s): 57.372072 seconds  
Job wait time with 32 task(s): 58.392965 seconds  
Job wait time with 64 task(s): 58.432444 seconds  
Sequential execution time: 43.413496 seconds  
Starting strong scaling timing. Submitting a total of 35 jobs.  
Job wait time with 1 task(s): 54.292447 seconds  
Job wait time with 2 task(s): 28.363432 seconds  
Job wait time with 4 task(s): 15.277171 seconds  
Job wait time with 8 task(s): 8.457225 seconds  
Job wait time with 16 task(s): 5.311071 seconds  
Job wait time with 32 task(s): 5.048173 seconds  
Job wait time with 64 task(s): 3.997336 seconds  
Starting matlabpool using the 'HPC' profile ... connected to 6  
Sending a stop signal to all the workers ... stopped.  
Execution time with parfor using 64 workers: 1.566806 seconds  
Speedup based on strong scaling with parfor using 64 workers:
```

The Workspace window shows a variable `myCluster` of type `<1x1 parallel.cluster.MJS>`.

The Command Window also includes a link: [View the published version of this example](#) to learn more about.

Matlab benchmark with 2*32 cores

Ste



GPU/parallel computing - workshops

- Terascale Alliance Workshop on GPU computing
 - 15-16.04.2013 @ DESY-HH
 - See: <https://indico.desy.de/event/7143>

- Accelerating Matlab Algorithms with System Objects and Parallel Computing
 - 23.04.2013 @ DESY-HH
 - See: <https://indico.desy.de/event/7719>

- Mathematica & Origin day
 - 26.04.2013 @ DESY-HH
 - See: <https://indico.desy.de/event/7749>

- Anything else (mpi, c++)?
 - it-training@desy.de



HPC Cluster – Information

> Web:

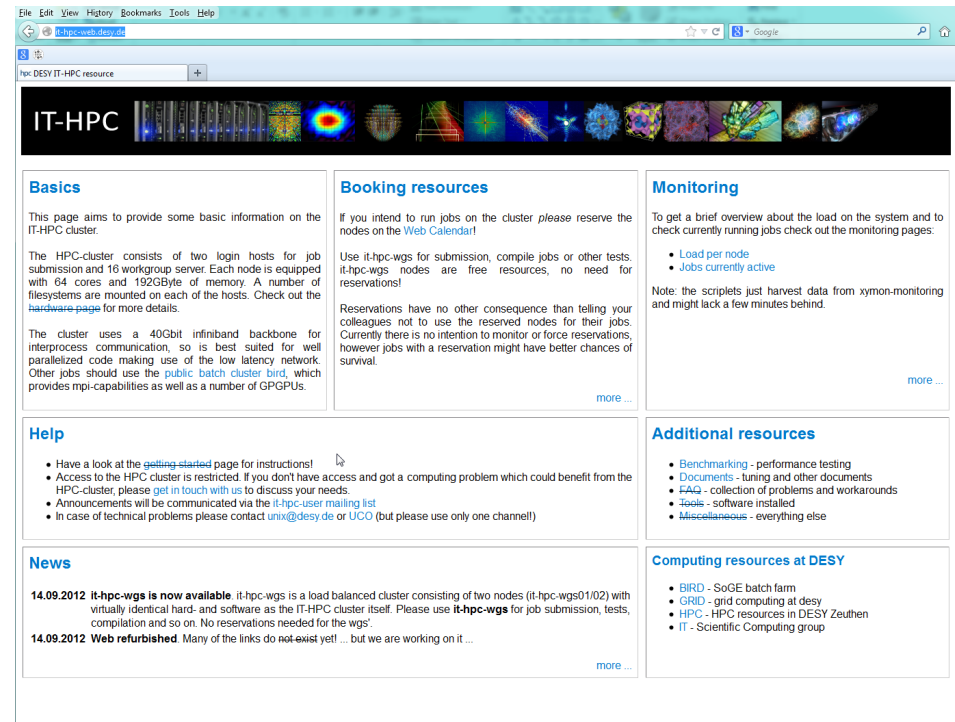
- http://it.desy.de/services/computing_infrastructure
- <http://it-hpc-web.desy.de> (only DESY-internal)

> Contact:

- uco@desy.de
- unix@desy.de

> Announcements

- Subscribe to it-hpc-user@desy.de



Questions? Comments?

