FRITZ-HABER-INSTITUT

MAX-PLANCK-GESELLSCHAFT

Q

THE HPC CLUSTER

Heinz Junkes

Short introduction

2. Dec. 2021

Hardware

1x Frontend: Xeon(R) Silver 4116 CPU @ 2.10GHz, 12 cores
10 TByte /home

38 Standard Compute Nodes: Xeon(R) Gold 6130 CPU @ 2.10GHz , 16 cores
196 Gbyte memory, 850 GByte /scratch,

2 x Compute Nodes with GPUs: Xeon(R) Gold 6130 CPU @ 2.10GHz , 16 cores
196 Gbyte memory, 850 GByte /scratch, 2 x Tesla P100

Currently 18 nodes, old yfhix, CPU0: Intel(R) Xeon(R) CPU E5-2660 v2 @ 2.20GHz
128 Gbyte memory, 400 GByte /scratch
Another 16 nodes are still to be integrated (if required)

Rocks cluster system

Welcome to the *q* cluster home page

Cluster Information

- Local ganglia monitoring instance (if ganglia roll is installed)
- Cluster kickstart graph
- Cluster contact: ppb@fhi-berlin.mpg.de

Visit Rocks Website for latest news and updates

Installed Rolls

CentOS	(v.7.4.1708)		hpc	(v.7.0)	roll usersguide
Updates-CentOS-7.4.1708	(v.2017-12-01)		kernel	(v.7.0)	
area51	(v.7.0)	roll usersguide	perl	(v.7.0)	roll usersguide
base	(v.7.0)	roll usersguide	python	(v.7.0)	roll usersguide
core	(v.7.0)		sge	(v.7.0)	roll usersguide
ganglia	(v.7.0)	roll usersguide	zfs-linux	(v.0.7.3)	roll usersguide

http://www.rocksclusters.org

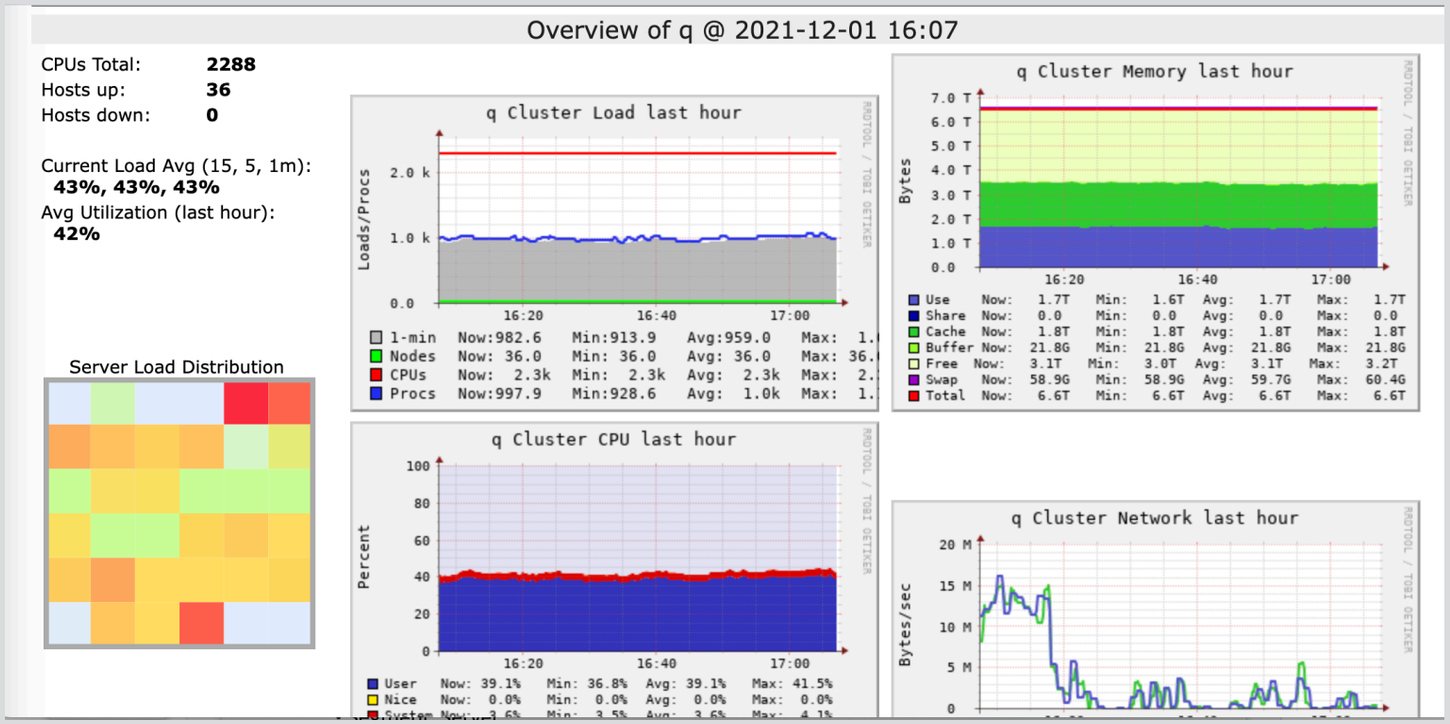
Installed hardware

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

Codename: Skylake-SP (Plattform "Purley")

- Architektur: Skylake
- Kerne: 16
- Threads: 32
- Basistakt: 2.10GHz
- Turbotakt: 3.70GHz
- TDP: 125W
- Fertigung: 14nm
- Interface: UPI (3 Links), 10.4GT/s
- L2-Cache: 16MB (16x 1MB)
- L3-Cache: 22MB
- PCIe-Lanes: 48x (PCIe 3.0)
- Speicher max.: 768GB
- Speichercontroller: Hexa Channel PC4-21300U (DDR4-2666)
- Speicherbandbreite: 128.0GB/s
- Stepping: H0
- Einführung: 2017/ Q3
- Segment: Server



Installed hardware

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

```
[root@q ~]# qhost
HOSTNAME                ARCH         NCPU NSOC NCOR NTHR  LOAD  MEMTOT  MEMUSE  SWAPTO  SWAPUS
----------------------------------------------------------------------------------------------
global                  -               -    -    -    -     -       -       -       -       -
compute-0-0             lx-amd64       64    2   32   64  3.84  187.3G    3.0G    3.9G     0.0
compute-0-1             lx-amd64       64    2   32   64  6.46  187.3G    1.9G    3.9G  210.6M
compute-0-10            lx-amd64       64    2   32   64 33.86  187.3G  102.7G    3.9G    2.2G
compute-0-11            lx-amd64       64    2   32   64 38.13  187.3G  151.2G    3.9G    1.7G
compute-0-12            lx-amd64       64    2   32   64  8.01  187.3G   15.4G    3.9G    1.7G
compute-0-13            lx-amd64       64    2   32   64 24.01  187.3G   31.1G    3.9G    1.6G
compute-0-14            lx-amd64       64    2   32   64 16.01  187.3G   15.0G    3.9G    1.7G
compute-0-15            lx-amd64       64    2   32   64 29.76  187.3G   13.4G    3.9G    1.7G
compute-0-8             lx-amd64       64    2   32   64 44.81  187.3G  103.7G    3.9G    3.9G
compute-0-9             lx-amd64       64    2   32   64 38.90  187.3G   75.0G    3.9G    1.7G
compute-1-0             lx-amd64       64    2   32   64  0.02  187.4G    1.9G    7.8G   83.4M
compute-1-1             lx-amd64       64    2   32   64  0.01  187.4G    2.2G    7.8G     0.0
compute-2-0             lx-amd64       20    2   20   20  0.03  125.7G    1.3G    2.0G     0.0
compute-2-1             lx-amd64       40    2   20   40  0.01  125.7G    1.4G    2.0G     0.0
compute-2-10            lx-amd64       20    2   20   20  0.01  125.7G    1.3G    2.0G     0.0
compute-2-3             lx-amd64       20    2   20   20 23.43  125.7G    8.3G    2.0G     0.0
compute-2-4             lx-amd64       20    2   20   20  4.25  125.7G   23.3G    2.0G  143.1M
compute-2-5             lx-amd64       20    2   20   20  0.01  125.7G    1.3G    2.0G     0.0
compute-2-6             lx-amd64       20    2   20   20  3.76  125.7G    4.1G    2.0G     0.0
compute-2-7             lx-amd64       20    2   20   20  0.01  125.7G    1.5G    2.0G     0.0
compute-2-8             lx-amd64       20    2   20   20  0.01  125.7G    1.2G    2.0G     0.0
compute-2-9             lx-amd64       20    2   20   20  0.01  125.7G    1.2G    2.0G     0.0
```

How to use the system (best)?

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

```
[root@q apps]# rocks list host partition compute-0-37
DEVICE MOUNTPOINT        START  SIZE   ID TYPE           FLAGS FORMATFLAGS
sda1   /                 1049kB 33.6GB -- ext4           boot  ------------
sda2   /var               33.6GB 16.8GB -- ext4           ----- ------------
sda3   swap               50.3GB 4194MB -- linux-swap(v1) ----- ------------
sda4   ----------------- 54.5GB 65.5GB -- -------------- ----- ------------
sda5   /state/partition1 54.5GB 65.5GB -- ext4           ----- ------------
sdb1   /state/partition2 1049kB 33.6GB -- ext4           ----- ------------
sdb2   /scratch           33.6GB 927GB  -- ext4           ----- ------------
```

No interactive jobs. Only per gridengine (job submission)

/scratch on each node (fast ssd, except compute-2-x), 192 GB Memory (128 GB on old system)

home-dir:
/dev/mapper/rocks_q-home 10T 6.9T 2.6T 73% /export

Installed software

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

```
[root@q ~]# cd /share/apps
[root@q apps]# ls -l
total 68
drwxr-xr-x  2 root       root      4096 Mar 22  2018 exampleScripts
drwxr-xr-x  2 rossi      FHI_COMP  4096 May 25  2018 fhiaims
drwxr-xr-x  3       350  FHI_COMP  4096 Mar 11  2018 Gaussian
drwxr-xr-x  3 root       FHI_COMP  4096 Mar 28  2021 Gaussian_c_2
-rw-r--r--  1 root       root      3693 Sep  9  2020 gmond.conf
-rwxr-xr-x  1 root       root       543 Mar 24  2021 gpu_load_sensor
-rw-r--r--  1 root       root      1939 Sep  9  2020 hosts
drwxr-xr-x 43 root       root      4096 Feb 19  2020 intel
drwxr-xr-x  4 root       root      4096 Aug  4  2020 MATLAB
drwxr-xr-x  8 root       FHI_COMP  4096 Jul 29  2019 molpro
drwxr-xr-x  7 root       FHI_COMP  4096 Mar 18  2018 openmpi2
drwxr-xr-x  7 root       FHI_COMP  4096 Feb 19  2019 openmpi2.1.5
drwxr-xr-x  7 root       root      4096 Nov 14  2019 openmpi2.1.6
drwxr-xr-x  7 root       FHI_COMP  4096 Nov 11  2019 openmpi3.1.4
drwxr-xr-x  6 root       root      4096 Mar 19  2018 python
-rwxr-xr-x  1 root       root       124 Jan 22  2019 sgeUpdate
drwxr-xr-x 21 turbomole  FHI_COMP  4096 Mar 19  2018 TURBOMOLE
```

gridengine

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

```
[root@q ~]# qhost
HOSTNAME                ARCH         NCPU NSOC NCOR NTHR  LOAD  MEMTOT  MEMUSE  SWAPTO  SWAPUS
----------------------------------------------------------------------------------------------
global                  -               -    -    -    -     -       -       -       -       -
compute-0-0             lx-amd64       64    2   32   64  3.84  187.3G    3.0G    3.9G     0.0
compute-0-1             lx-amd64       64    2   32   64  6.46  187.3G    1.9G    3.9G  210.6M
compute-0-10            lx-amd64       64    2   32   64 33.86  187.3G  102.7G    3.9G    2.2G
compute-0-11            lx-amd64       64    2   32   64 38.13  187.3G  151.2G    3.9G    1.7G
compute-0-12            lx-amd64       64    2   32   64  8.01  187.3G   15.4G    3.9G    1.7G
compute-0-13            lx-amd64       64    2   32   64 24.01  187.3G   31.1G    3.9G    1.6G
compute-0-14            lx-amd64       64    2   32   64 16.01  187.3G   15.0G    3.9G    1.7G
compute-0-15            lx-amd64       64    2   32   64 29.76  187.3G   13.4G    3.9G    1.7G
```

http://wiki.gridengine.info/wiki/index.php/Main_Page

Matlab parallel server / Turbomole

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

Matlab parallel server can be used to run jobs on q via gridenengine

see:
https://de.mathworks.com/help/matlab-parallel-server/configure-using-the-generic-scheduler-interface.html

-> Manually Configure a Cluster Profile -> SGE

INSTALLATION OF TURBOMOLE 7.2.1 in /share/apps/TUBOMOLE

Use of /scratch on compute nodes

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

```
[root@q exampleScripts]# pwd
/share/apps/exampleScripts
[root@q exampleScripts]# ls -l
total 8
-rw-r--r-- 1 root root 1547 Mar 18  2018 gpawsub
-rw-r--r-- 1 root root 1434 Dec  1 18:23 subcfour
-rwxr-xr-x 1 root root 1054 Mar 22  2018 subg16
```

Gratefully provided by Gert.

```
cd $cwd
for host in \`cat \$PE_HOSTFILE | awk '{print \$1}'\`; do
echo ” create scratch on \$host ”
ssh \$host mkdir \${WORKDIR}
scp -p  * \$host:\${WORKDIR}/.
done
```

Intel - MPI

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

For Intel MPI the fabric to be use must be defined
In our case shared memory for intra node communication
and tcp for inter node communication like: **export I_MPI_FABRICS=shm:tcp**

submission script example

```
#!/bin/bash
#$ -N Work
#$ -cwd
#$ -pe mpi 128
#$ -S /bin/bash
#$ -q short.q
#$ -e $JOB_NAME.e$JOB_ID
#$ -o $JOB_NAME.o$JOB_ID
impi=/share/apps/intel/impi/2018.1.163/bin64
create scratch .....
myProg=/home/junkes/workshop/a.out
export I_MPI_FABRICS=shm:tcp
$impi/mpirun -machinefile $TMPDIR/machines -np $NSLOTS $myProg
```

drmaa

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

SGE (gridengine) is drmaa compatible, can be used e.g. with drmaa-python

E.g. to run script "sleeper.sh"

```
#!/bin/bash
echo "Hello World, the answer is $1"
sleep 3s
echo "$2 Bye world!"
```

run:

```
python example.py
```

example.py:

```
#!/usr/bin/env python

from __future__ import print_function
import drmaa
import os
def main():
    """Submit a job.
    Note, need file called sleeper.sh in current directory.
    """
    s = drmaa.Session()
    s.initialize()
    print('Creating job template')
    jt = s.createJobTemplate()
    jt.remoteCommand = os.getcwd() + '/sleeper.sh'
    jt.args = ['42','Simon says:']
    jt.joinFiles=True

    jobid = s.runJob(jt)
    print('Your job has been submitted with id ' + jobid)
    print('Cleaning up')
    s.deleteJobTemplate(jt)
    s.exit()

if __name__=='__main__':
    main()
```

open mpi / python / gaussian

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

openmpi 2.1.5 avaliable in /share/apps/openmpi2.1.5
(compiled for skylake Intel 6130)

Numpy scipy and python[23] can be found in /share/apps/intel/intelpython[23]

Gaussian compiled for skylake Intel 6130

Nvidia Toolkit 10 installed on gpu nodes (compute-1-x).
Can be used via gridengine by requesting -l gpu=1

e.g. qsub -l gpu=1 “submitScript"

Production software

All programs that should be used in production should be compiled on compute-0-37 -> skylake Intel 6130

How to apply for an account?

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

Mail to “ppb@fhi.mpg.de"

Note:

Unfortunately, we do not have the capacity at pp&b to provide assistance in using the cluster.
(Unfortunately also no support from the departments/institute).

We have installed the system and are doing our best to keep the system running.
(Info in /etc/motd)

We depend on you as users to support each other and be considerate of each other.

A Rocket Channel is available for the exchange of experiences:
https://chat.fhi-berlin.mpg.de/group/qCluster_Users

Last info on Rocket-channel

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

Theres more to this. @all
We would like to reconfigure the RAID to a different Level to increase security against hardware defects. (this time we were lucky).
Please take a look at your `home` on `q` and do some housekeeping. This would make syncing and restoring easyier and faster for us.
We then want to temporarly move `/export` to another space to clear and reconfigure the harddrives.

heres a list of the biggest homes:

4.0K	./langenhan	2.3M	./grabarics	462M	./spok	4.1G	./arghya
36K	./bagus	2.9M	./doppelbauer	497M	./windsor	4.6G	./intel
36K	./cian	5.4M	./schalsas	883M	./sreekanta	4.6G	./tkropp
36K	./cpratsch	5.6M	./mirahmadi	929M	./omojola	5.0G	./pincelli
36K	./maikelettow	7.4M	./jbischoff	1.2G	./kuhlenbeck	11G	./ober
36K	./morsa	8.1M	./mcretu	1.2G	./wschoell	16G	./liyake
36K	./rossi	8.5M	./juhyeonlee	1.5G	./gewinner	17G	./fmueller
64K	./gpszekeres	14M	./green	2.1G	./zli	22G	./krecinic
64K	./mike	21M	./eibenberger	2.4G	./lee	23G	./dathomas
128K	./fielicke	73M	./lapsanska	2.5G	./junkes	23G	./kmeyer
352K	./neef	209M	./torres	3.5G	./hernandez	30G	./ckirschbaum

Last info on Rocket-channel

MAX-PLANCK-GESELLSCHAFT

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

Theres more to this. @all
We would like to reconfigure the RAID to a different Level to increase security against hardware defects. (this time we were lucky).
Please take a look at your `home` on `q` and do some housekeeping. This would make syncing and restoring easyier and faster for us.
We then want to temporarly move `/export` to another space to clear and reconfigure the harddrives.
heres a list of the biggest homes:

34G	./hetaba	102G	./maklar	336G	./mkarra
35G	./kschatz	110G	./dong	420G	./dendzik
43G	./rayoonchang	122G	./xiaoyan	475G	./ghafari
49G	./baldauf	170G	./helden	480G	./wang
52G	./jennyt	173G	./reu	583G	./xyliu
58G	./twinkleyadav	233G	./trjones	1.1T	./sophia
72G	./greiskim	236G	./xian	1.7T	./woosunjang
81G	./jperezri	244G	./martint		
89G	./kraus	263G	./marianski		