JÜLICH
FORSCHUNGSZENTRUM

Site Report: Jülich Supercomputing Centre

2015-09-16 | SLUG 2015 |

Jülich Supercomputing Centre (JSC)

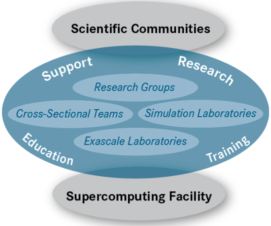
Jülich Supercomputing Centre

Supercomputer operation for:

- Centre – FZJ
- Region – RWTH Aachen University
- Germany – Gauss Centre for Supercomputing
  John von Neumann Institute for Computing
- Europe – PRACE, EU projects

Application support

- Unique support & research environment at JSC
- Peer review support and coordination

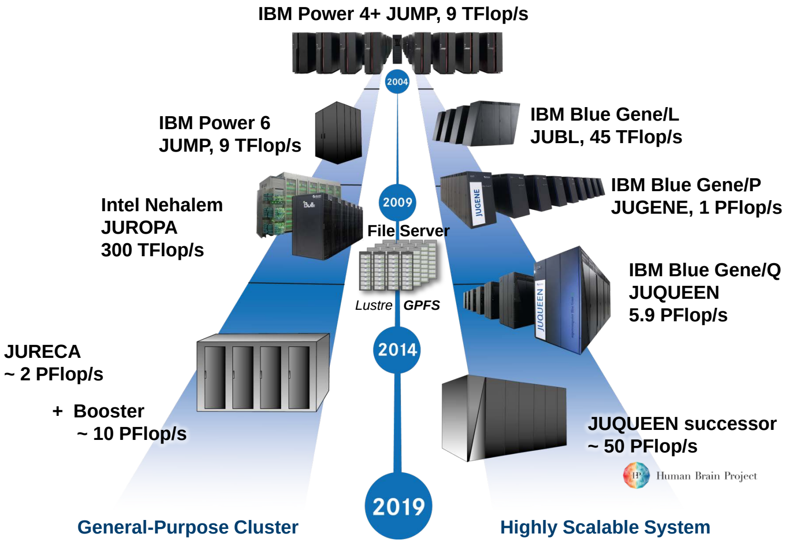

Scientific Communities
Support
Research
Research Groups
Cross-Sectional Teams
Simulation Laboratories
Education
Exascale Laboratories
Training
Supercomputing Facility

R&D work

- Methods and algorithms, computational science, performance analysis and tools
- Scientific Big Data Analytics
- Computer architectures, Co-Design
  Exascale Laboratories: EIC, ECL, NVIDIA

Education and Training

Supercomputer Systems: Dual Architecture Strategy



IBM Power 4+ JUMP, 9 TFlop/s

2004

IBM Power 6
JUMP, 9 TFlop/s

IBM Blue Gene/L
JUBL, 45 TFlop/s

Intel Nehalem
JUROPA
300 TFlop/s

2009

IBM Blue Gene/P
JUGENE, 1 PFlop/s

File Server

Lustre ***GPFS***

IBM Blue Gene/Q
JUQUEEN
5.9 PFlop/s

2014

JURECA
~ 2 PFlop/s

+ Booster
~ 10 PFlop/s

JUQUEEN successor
~ 50 PFlop/s

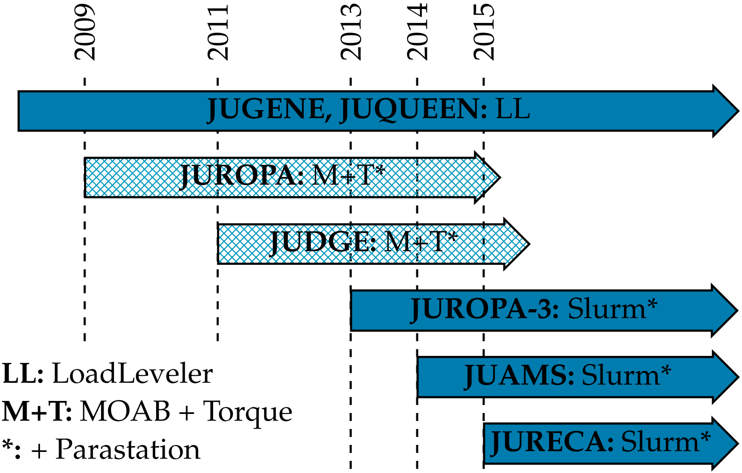
Human Brain Project

2019

General-Purpose Cluster

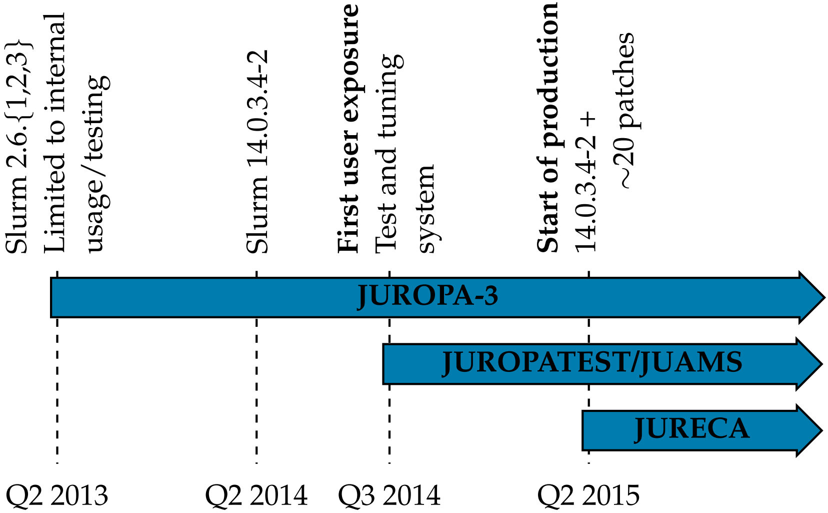
Highly Scalable System

(A subset of) workload managers at JSC



LL: LoadLeveler
M+T: MOAB + Torque
***:** + Parastation

Slurm at JSC



Q2 2013
Slurm 2.6.{1,2,3}
Limited to internal usage/testing

Q2 2014
Slurm 14.0.3.4-2

Q3 2014
First user exposure
Test and tuning system

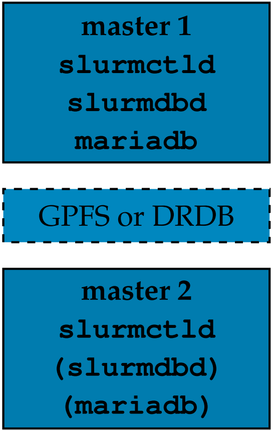
Q2 2015
Start of production
14.0.3.4-2 + ~20 patches

JUROPA-3

JUROPATEST/JUAMS

JURECA

Slurm on JSC clusters



master 1

`slurmctld`

`slurmdbd`

`mariadb`

GPFS or DRDB

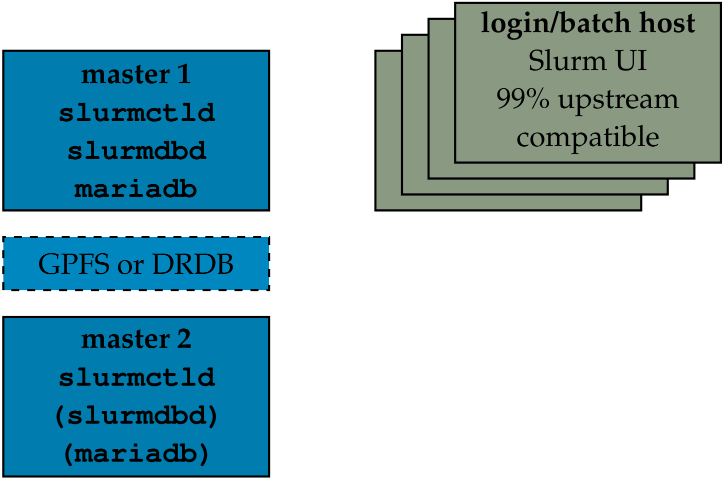
master 2

`slurmctld`

`(slurmdbd)`

`(mariadb)`

Slurm on JSC clusters



master 1

`slurmctld`

`slurmdbd`

`mariadb`

GPFS or DRDB

master 2

`slurmctld`

`(slurmdbd)`

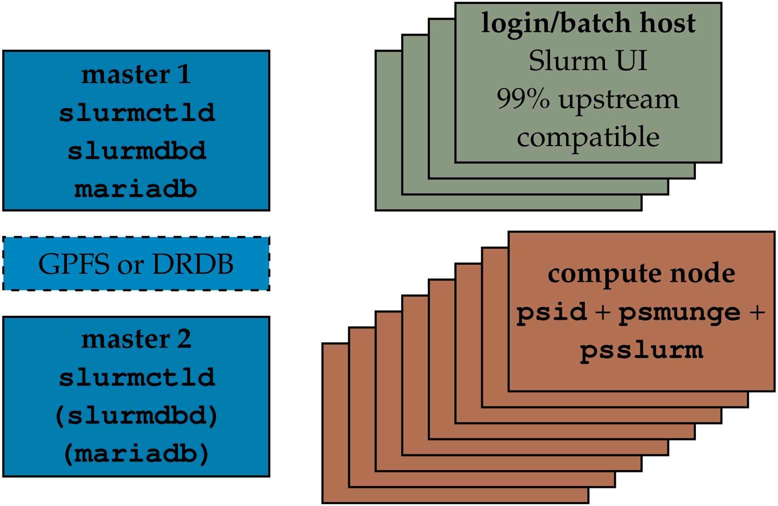
`(mariadb)`

login/batch host

Slurm UI

99% upstream compatible

Slurm on JSC clusters



master 1
`slurmctld`
`slurmdbd`
`mariadb`

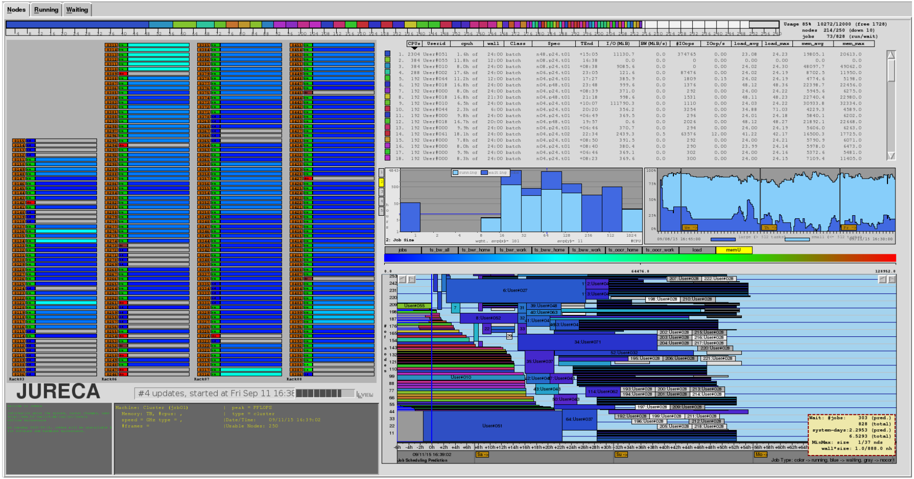
GPFS or DRDB

master 2
`slurmctld`
`(slurmdbd)`
`(mariadb)`

login/batch host
Slurm UI
99% upstream compatible

compute node
`psid` + `psmunge` + `psslurm`

Llview: WM-centric system monitoring



```
Usage 85%  10272/12000 (free 1728)
     nodes  214/250 (down 10)
     jobs    73/828 (run/wait)
```

	CPUs	Userid	cpuh	wall	Class	Spec	TEnd	I/O (MiB)	BW (MiB/s)	#IOops	IOop/s	load_avg	load_max	mem_avg	mem_max
1.	2304	User#051	1.6h of	24:00	batch	n48.p24.t01	+15:05	11130.7	0.0	374765	0.00	23.08	24.23	19805.1	20613.0
2.	384	User#055	11.8h of	12:00	batch	n08.p24.t01	16:38	0.0	0.0	0	0.00	0.00	0.00	0.0	0.0
3.	384	User#010	8.0h of	24:00	batch	n08.p24.t01	+08:38	9085.6	0.0	0	0.00	24.02	24.30	48097.7	49062.0
4.	288	User#002	17.6h of	24:00	batch	n06.p24.t01	23:05	121.6	0.0	87476	0.00	24.02	24.19	8702.5	11950.0
5.	192	User#064	11.2h of	12:00	batch	n04.p24.t01	17:27	385.9	0.0	1809	0.15	24.02	24.19	4774.6	5198.0
6.	192	User#018	16.8h of	24:00	batch	n04.p48.t01	23:48	999.6	0.0	1376	0.00	48.12	48.36	22398.7	22456.0
7.	192	User#000	8.0h of	24:00	batch	n04.p24.t01	+08:39	371.0	0.0	292	0.00	24.00	24.22	5945.6	6275.0
8.	192	User#018	16.8h of	21:30	batch	n04.p48.t01	21:18	998.6	0.0	1531	0.00	48.11	48.23	22740.4	22980.0
9.	192	User#010	6.5h of	24:00	batch	n04.p24.t01	+10:07	111790.3	0.0	1110	0.00	24.03	24.22	30933.8	32334.0
10.	192	User#044	2.3h of	6:00	batch	n04.p24.t01	20:20	356.2	0.0	3254	0.00	34.88	71.03	4229.3	4589.0
11.	192	User#000	9.8h of	24:00	batch	n04.p24.t01	+06:49	369.5	0.0	296	0.00	24.01	24.18	5840.1	6202.0
12.	192	User#018	16.7h of	20:00	batch	n04.p48.t01	19:57	0.6	0.0	2026	0.00	48.12	48.27	21892.1	22668.0
13.	192	User#000	9.9h of	24:00	batch	n04.p24.t01	+06:46	370.7	0.0	294	0.00	24.00	24.19	5606.0	6263.0
14.	192	User#061	18.1h of	24:00	batch	n04.p24.t02	22:34	2439.3	0.5	63576	12.00	41.22	42.17	16500.3	17725.0
15.	192	User#000	7.8h of	24:00	batch	n04.p24.t01	+08:50	391.5	0.0	292	0.00	24.00	24.21	5790.9	6071.0
16.	192	User#000	8.0h of	24:00	batch	n04.p24.t01	+08:40	380.4	0.0	290	0.00	23.99	24.14	5978.0	6473.0
17.	192	User#000	9.9h of	24:00	batch	n04.p24.t01	+06:46	369.1	0.0	302	0.00	24.00	24.16	5372.6	5481.0
18.	192	User#000	8.3h of	24:00	batch	n04.p24.t01	+08:23	369.6	0.0	300	0.00	24.00	24.15	7109.4	11405.0

#4 updates, started at Fri Sep 11 16:38

```
welcome to JURECA
*
information about the system, latest changes, user
http://www.fz-juelich.de/ias/jsc/jureca
offline maintenance
*
on tuesday 2015-09-15, JURECA will be unavailable
for software and hardware maintenance.
```

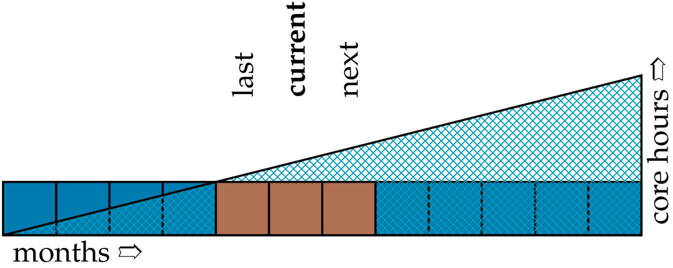
```
Machine: Cluster (jrb01)              |  peak = PFLOPS
  Memory: TB, #cpus: ,                |  type = cluster
  speed = GHz type = ,                |Date/Time:    09/11/15 16:39:02
  #frames =                           |Usable Nodes: 250
```

```
Wait: #jobs:     303 (pred.)
                 828 (total)
 system-days:2.2953 (pred.)
             6.5293 (total)
 MinMax: size   1/37 nds
    wall*size: 1.0/888.0 nh
```

Job Type: color -> running. blue -> waiting. gray -> nocont

Batch model

- Associations **==** project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!

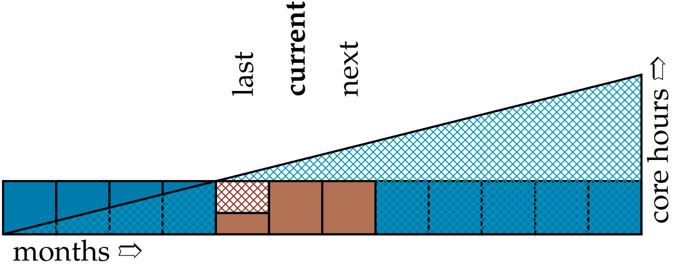


account: +200%

qos: **`normal`**

Batch model

- Associations **==** project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!

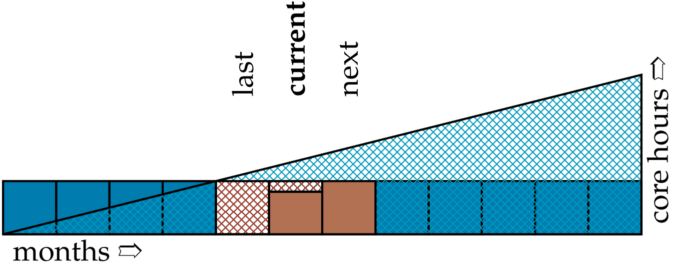


account: +140%

qos: **`normal`**

Batch model

- Associations **==** project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!

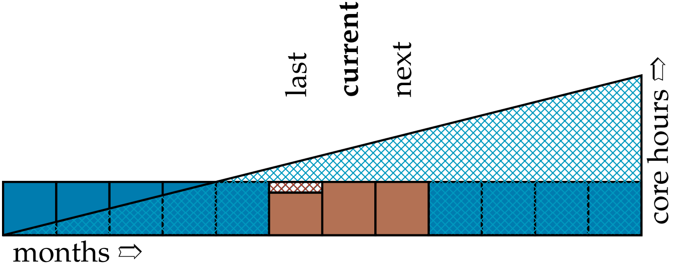


account: +80%

qos: **`normal`**

Batch model

- Associations == project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!

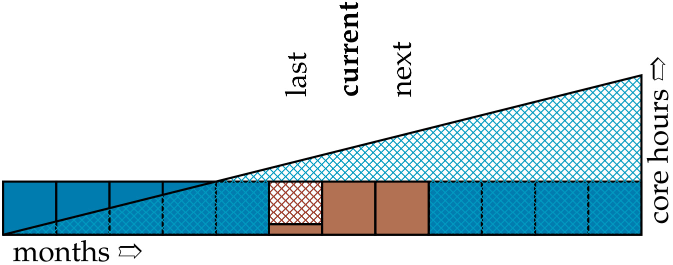


account: +180%

qos: **`normal`**

Batch model

- Associations **==** project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!

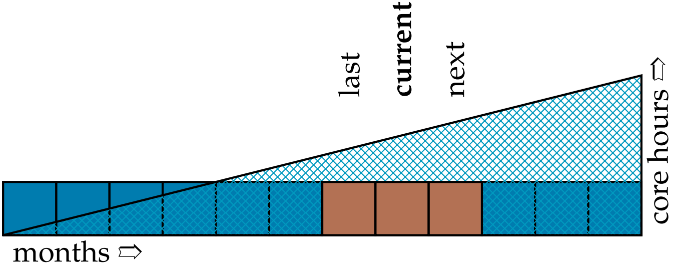


account: +120%

qos: **`normal`**

Batch model

- Associations **==** project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!

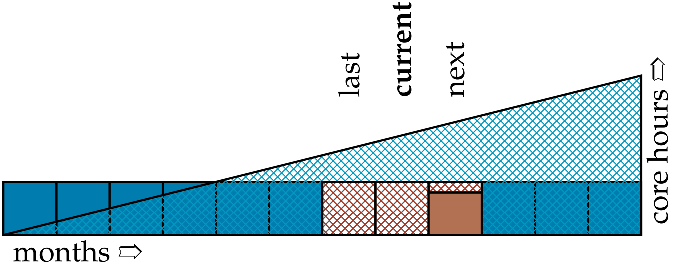


account: +200%

qos: **`normal`**

Batch model

- Associations **==** project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!

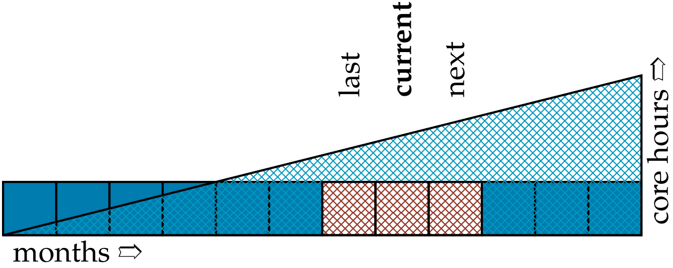


account: -20%

qos: **`normal`**

Batch model

- Associations **==** project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!

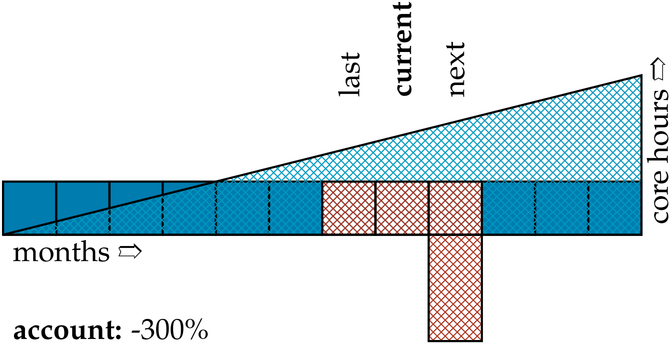


account: -100%

qos: nocont ⇨ Low priority, wallclock limitation

Batch model

- Associations **==** project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!

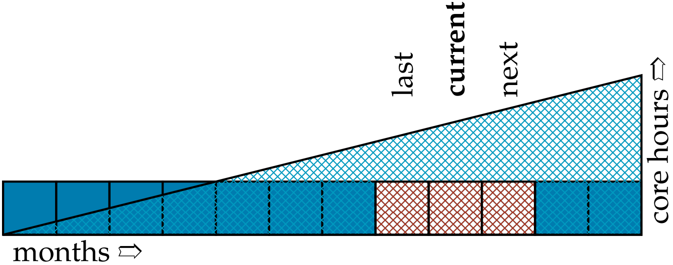


account: -300%

qos: nocont ⇨ Low priority, wallclock limitation

Batch model

- Associations **==** project (primary group), user (user id)
- Scheduling with multifactor priorities ⇨ QoS!



account: -100%

qos: nocont ⇨ Low priority, wallclock limitation

Noteworthy/Exotic settings

- Custom X forwarding mechanism (**`–forward-x`**) for **`srun`**
- Custom CPU binding code
  - Designed for the 99% while leaving freedom for the 1%
  - Example: Do not “pack” **`–exclusive`** job steps
- Non-consumable gres (**`mem{128,256,512,1024}`**) to handle heterogeneities
  - Features are not captured in accounting database
  - **Rule of thumb:** If it is not in the accounting database we cannot use it

Member of the Helmholtz-Association

Slurm experience

- Overall very positive experience
- Largely stable
  - Five internal tickets open against **`slurmctld`** (14.03.4-2):
    1. **`assoc_mgr`** lock contention during share updates
    2. **`slurmctld`** crash during assoc update
    3. Split brain occurence
    4. Invalid **`protocol_version`** in **`_pack_cred()`** during job launch
    5. Bogus projected start/end times
- Open source is a big plus
  - Increased insights
  - Ability to push features based on own needs/roadmap

End of presentation

JUQUEEN: Jülich's Scalable Petaflop System

IBM Blue Gene/Q JUQUEEN

- IBM PowerPC® A2 1.6 GHz,
  16 cores per node
- 28 racks, 458,752 cores
- 5,9 Petaflop/s peak
  5,0 Petaflop/s Linpack
- 448 TByte main memory

- connected to a Global Parallel File System (GPFS) with
  O(10) PByte online disk and O(50) PByte offline tape capacity
- 5D network
- Production start: Nov 5, 2012

Jun 2015:
#2 in Europe
#9 worldwide
#51 in Green500

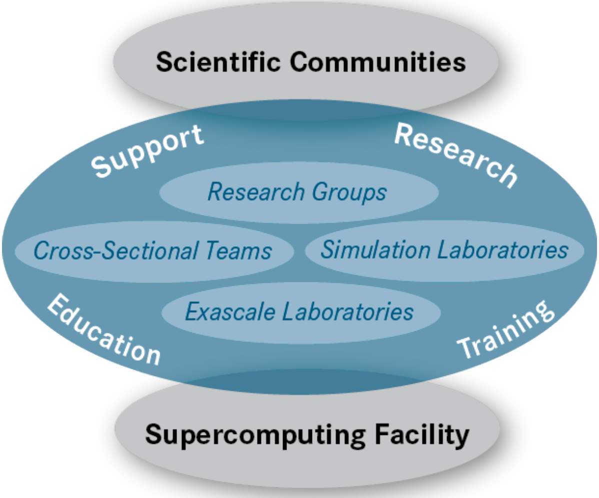
JURECA: Jülich Research on Exascale Cluster Architectures

System integrator:
T-Platforms, Russia

JURECA, an Intel-based cluster

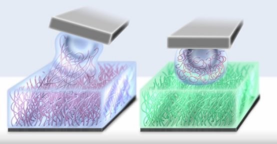
- 2 Intel Haswell 12-core processors,
  2.5 GHz, SMT, 128 GB main memory
- 1,884 compute nodes or 45,216 cores, thereof
  75 nodes with 2 K80 NVIDIA graphics cards each and
  12 nodes with 512 GB main memory and 2 K40 NVIDIA graphics cards each for visualisation
- 2.245 Petaflop/s peak (with K80 graphics cards)
  ?? Petaflop/s Linpack
- 281 TByte memory
- Mellanox Infiniband EDR
- Connected to the GPFS file system on JUST

R&D and Application Support at the Jülich Supercomputing Centre

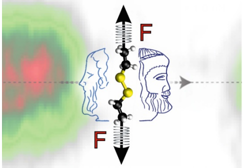


High Impact Publications

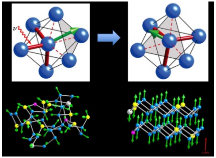
Users of the facility at JSC produce about 250 publications per year



S. de Beer, M. Müser
Nature Communications **5**
(2014) 3781

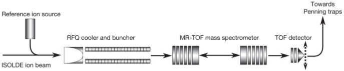


D. Marx et al.,
Nature Chemistry **5**
(2013) 685

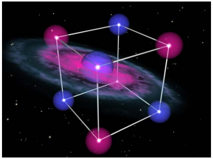


R.O. Jones et al.,
Nature Materials **10**
(2011) 129

Sz. Borsanyi et al.,
Science **347** (2015) 6229



A. Schwenk et al.,
Nature **498** (2013) 346



M. Lezaic et al.,
Nature Materials **9**
(2010) 649