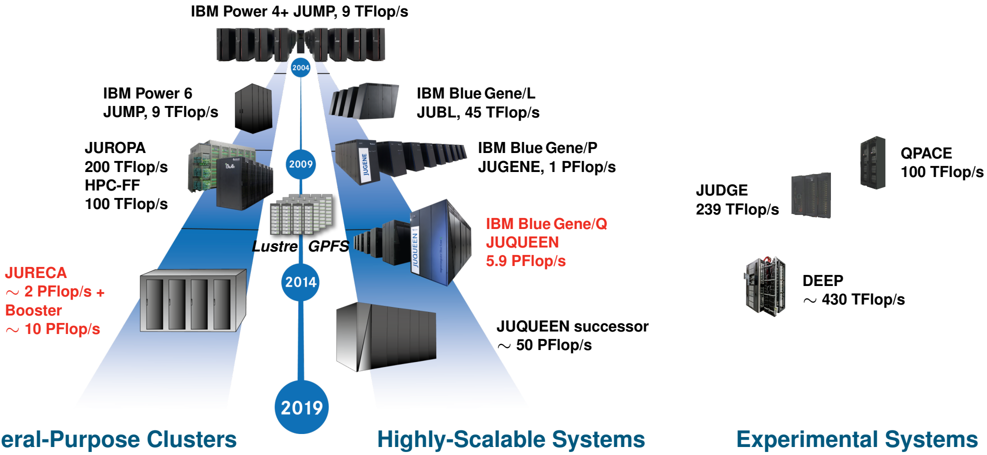
JÜLICH
FORSCHUNGSZENTRUM

Extreme-scaling applications 24/7 on JUQUEEN Blue Gene/Q

Experience from our latest scaling workshop

03.09.2015 | D. Brömmel, W. Frings, and B. J. N. Wylie | Jülich Supercomputing Centre | Forschungszentrum Jülich GmbH

Computer systems at JSC



IBM Power 4+ JUMP, 9 TFlop/s

2004

IBM Power 6
JUMP, 9 TFlop/s

IBM Blue Gene/L
JUBL, 45 TFlop/s

JUROPA
200 TFlop/s
HPC-FF
100 TFlop/s

2009

IBM Blue Gene/P
JUGENE, 1 PFlop/s

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

2019

JUDGE
239 TFlop/s

QPACE
100 TFlop/s

DEEP
∼ 430 TFlop/s

General-Purpose Clusters

Highly-Scalable Systems

Experimental Systems

User support and scaling activities

A continuous 24/7 effort

Established support levels at JSC provide help scaling application codes.
This includes:

- Application support (initial contact point)
- Cross-sectional teams (Performance Analysis and Mathematical Modelling)
- Simulation Laboratories (part of Computational Science Division at JSC)

User support and scaling activities

A continuous 24/7 effort

Established support levels at JSC provide help scaling application codes.
This includes:

- Application support (initial contact point)
- Cross-sectional teams (Performance Analysis and Mathematical Modelling)
- Simulation Laboratories (part of Computational Science Division at JSC)

In addition:
Workshops on Porting and Tuning on JUQUEEN and Extreme Scaling on JUQUEEN with dedicated or even exclusive access to the system and direct support during hands-on sessions.

Extreme Scaling Workshop on JUQUEEN

Extreme Scaling 24/7

- This latest edition of Extreme Scaling Workshops invited 7 applications teams and was extremely successful: all teams had their codes running on the full system within 24 hours.
- The workshop provided exclusive access to JUQUEEN with close support by JSC Simulation Laboratories for Climate Science, Fluids & Solids Engineering and Neuroscience assisted the code-teams, along with JSC Cross-sectional Teams, JUQUEEN and IBM technical support.
- 5 new codes entered the High-Q Club as a result.
- A detailed report with user contributions is available as technical report FZJ-JSC-IB-2015-01 http://juser.fz-juelich.de/record/188191

The High-Q Club idea

Start a collection of codes to showcase running on all 28 racks of Blue Gene/Q at JSC, effectively using all 458 752 cores with up to 1.8M hardware threads

⟶ Promote the idea of exascale capability computing

⟶ Spark interest in tuning and scaling codes

Goal

- Encourage our users to try and reach exascale readiness
- Establish milestones in application development towards future systems
- Identify and understand bottlenecks in trying to reach millions of threads/processes and learn how to transition to exascale systems

Current status of the High-Q Club

Diverse membership of ~~24~~ 25 codes from fundamental physics, neuroscience, plasma physics, molecular dynamics, engineering and climate and earth science.

Member • Member • Member • Member • Member • Member • Member • Member
HiQ
2013

10 codes

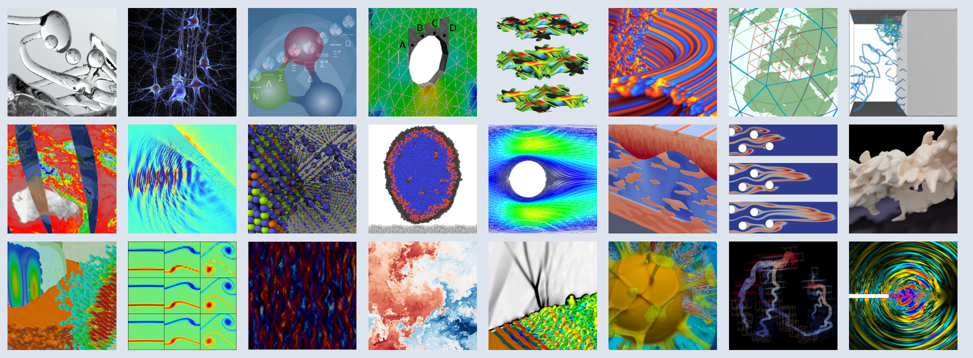
Member • Member • Member • Member • Member • Member • Member • Member
HiQ
2014

7 codes

Member • Member • Member • Member • Member • Member • Member • Member
HiQ
2015

~~7 codes~~
8 codes

Current status of the High-Q Club



CIAO, CoreNeuron, dynQCD, FE2TI, FEMPAR, Gysela, ICON, IMD, JURASSIC,
JuSPIC, KKRnano, LAMMPS(DCM), MP2C, $\mu\varphi$ (muPhi), Musubi, NEST, OpenTBL,
PEPC, PMG+PFASST, PP-Code, psOpen, SHOCK, Terra-Neo, waLBerla, ZFS

Becoming a member

Or: how to compare and judge applications

- Wide range of applications ⟶ no common set of criteria
- Selection criteria are flexible (open for discussion!)
  - We try to collect as much information as possible (not all is made public)
  - Discussions with developers and within JSC
- Run a non-trivial example, ideally very close to production runs
- Submit evidence of strong and/or weak scalability to all available cores
- Preference on multi-threading (at least use HWTs)
- Include I/O if possible
- Possibly provide peak performance numbers

Extreme Scaling Workshop on JUQUEEN

Code characteristics

	Programming			Parallelisation			
Code	Languages			MPI	OMP	Concurrency	File I/O
CoreNeuron	C	C++		1	64	64: 1 835 008	MPI-IO
FE^{2}TI	C	C++		16	4	64: 1 835 008	
FEMPAR			F08	64		64: 1 756 001	
ICON	C		Ftn	1	64	64: 1 835 008	(netCDF)
MPAS-A	C		Ftn	16		16: 458 752	PIO,pNetCDF
psOpen			F90	32	2	64: 1 835 008	pHDF5
SHOCK	C			64		64: 1 835 008	(cgns/HDF5)

Extreme Scaling Workshop on JUQUEEN

Code characteristics

Code
CoreNeuron
FE^{2}TI
FEMPAR
ICON
MPAS-A
psOpen
SHOCK

EPFL Blue Brain Project

simulation of electrical activity of neuronal networks including morphologically detailed neurons

Extreme Scaling Workshop on JUQUEEN

Code characteristics

Code
CoreNeuron
FE^{2}TI
FEMPAR
ICON
MPAS-A
psOpen
SHOCK

Universität Köln & TUB Freiberg

scale-bridging incorporating micro-mechanics in macroscopic simulations of multi-phase steels

UPC-CIMNE

massively-parallel finite-element simulation of multi-physics problems governed by PDEs

Extreme Scaling Workshop on JUQUEEN

Code characteristics

Code

CoreNeuron
FE^{2}TI
FEMPAR
ICON
MPAS-A
psOpen
SHOCK

DKRZ & JSC SimLab Climate Science
icosahedral non-hydrostatic atmospheric model

KIT & NCAR
multi-scale non-hydrostatic atmospheric model for global, convection-resolving climate simulations

Extreme Scaling Workshop on JUQUEEN

Code characteristics

Code
CoreNeuron
FE^{2}TI
FEMPAR
ICON
MPAS-A
psOpen
SHOCK

RWTH-ITV Inst. for Combustion Technology & JARA

direct numerical simulation of fine-scale turbulence

RWTH Shock Wave Laboratory

structured high-order finite-difference kernel for compressible flows

Extreme Scaling Workshop on JUQUEEN

Code characteristics

	Programming			Parallelisation			
Code	Languages			MPI	OMP	Concurrency	File I/O
CoreNeuron	C	C++		1	64	64: 1 835 008	MPI-IO
FE^{2}TI	C	C++		16	4	64: 1 835 008	
FEMPAR			F08	64		64: 1 756 001	
ICON	C		Ftn	1	64	64: 1 835 008	(netCDF)
MPAS-A	C		Ftn	16		16: 458 752	PIO,pNetCDF
psOpen			F90	32	2	64: 1 835 008	pHDF5
SHOCK	C			64		64: 1 835 008	(cgns/HDF5)

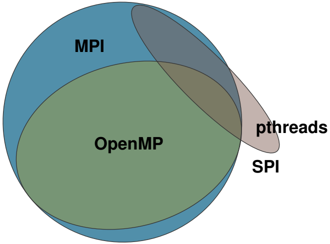
Extreme Scaling Workshop on JUQUEEN

Code characteristics

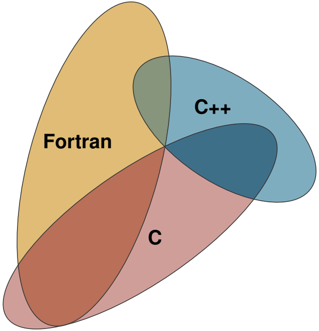
	Programming			Parallelisation			
Code	Languages			MPI	OMP	Concurrency	File I/O
CoreNeuron	C	C++		1	64	64: 1 835 008	MPI-IO
FE^{2}TI	C	C++		16	4	64: 1 835 008	
FEMPAR			F08	64		64: 1 756 001	
ICON	C		Ftn	1	64	64: 1 835 008	(netCDF)
MPAS-A	C		Ftn	16		16: 458 752	PIO,pNetCDF
psOpen			F90	32	2	64: 1 835 008	pHDF5
SHOCK	C			64		64: 1 835 008	(cgns/HDF5)

Extended statistics

Programming models



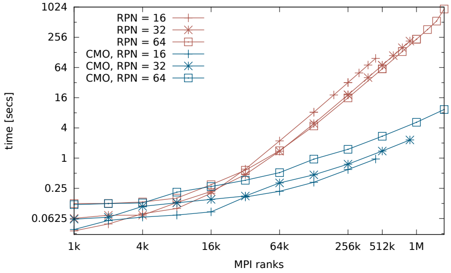
Programming languages



Venn diagrams with areas proportional to absolute numbers.

MPI to the extreme

- MPI-only possible – but only 256 MB available per rank
  sometimes memory is an issue either way
- MPAS-A: model initialisation took 30 mins (grid and neighbourhood set-up)
- ICON: `MPI_THREAD_MULTIPLE` w/ user-defined `MPI_Allreduce` & `MPI_IN_PLACE` prohibitive
- FEMPAR: MPI communicator management gets increasingly costly, trade speed for even more memory w/ `MPI_Comm_Split`



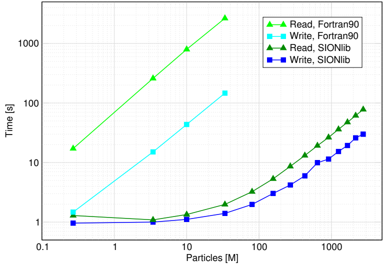
File I/O – too slow, too big, too complicated?

File I/O remains the most common impediment to scalability

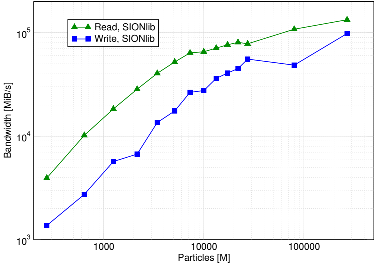
- Times for I/O prevent scaling, on-the-fly analysis necessary
- MPAS-A: 1.2 TB of data not enough to scale, difficult to transfer and read in
- SHOCK: CFD specific library on top of HDF5 lead to non-fixable errors, synthetic data used
- Tools like Darshan or Score-P used at scale to identify problems

File I/O – too slow, too big, too complicated?

- Effective solutions need to be employed, such as SIONlib
- 11 (25) codes use parallel I/O, 5 (25) use SIONlib



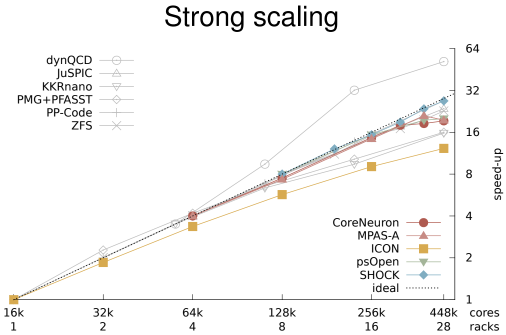
MP2C on one mid-plane

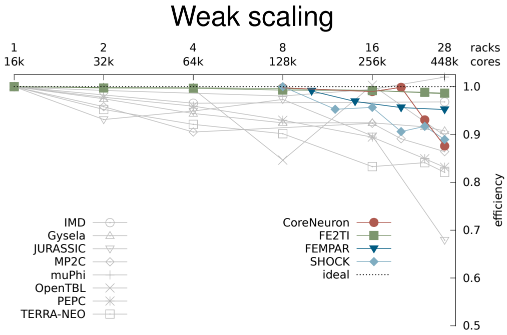


MP2C on 28 racks

Extreme Scaling Workshop on JUQUEEN

Scaling results

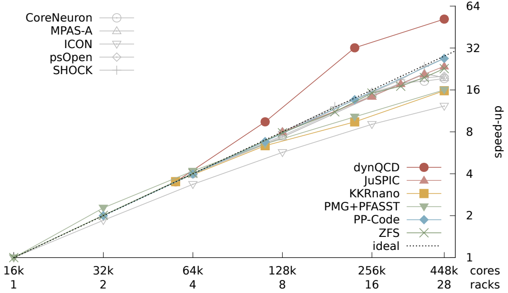




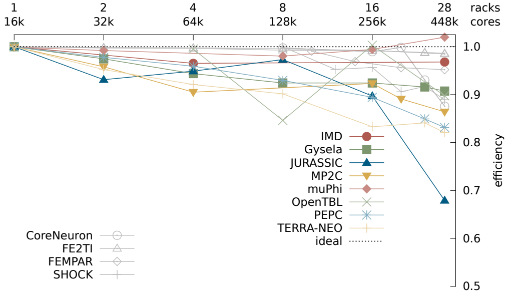
Remaining High-Q codes

Scaling results

Strong scaling



Weak scaling



Workshop and High-Q Club – also for other systems?

Wide range of HPC applications have demonstrated excellent scalability, generally with only modest tuning effort

- Standard languages and MPI+multi-threading are sufficient
- Over-subscription of cores delivers important efficiency benefits
- Use vectorisation/SIMDization & libraries for node performance
- Did not see (m)any disruptive changes

Scaling on BG/Q also delivers benefits for other HPC computer systems

Summary

- Our activities attract a lot of interest: users ask for scaling workshops and strive to join the High-Q Club ⟶ currently 25 codes listed
- Hopefully enable our users to transition from peta to exascale
- Identified bottlenecks, solutions to common issues at hand – not unique
- So far no disruptive changes necessary or chosen

⟶ Performance improvements also not unique

- Browse the High-Q Club webpages: `http://www.fz-juelich.de/ias/jsc/high-q-club`
- Download our technical report: FZJ-JSC-IB-2015-01 `http://juser.fz-juelich.de/record/188191`