GPU-accelerated Computing at Scale

Dirk Pleiter I GTC Europe | 10 October 2018

Outline

- **Supercomputers at JSC**
- **Future science challenges**
- **Outlook and conclusions**

Supercomputers at JSC

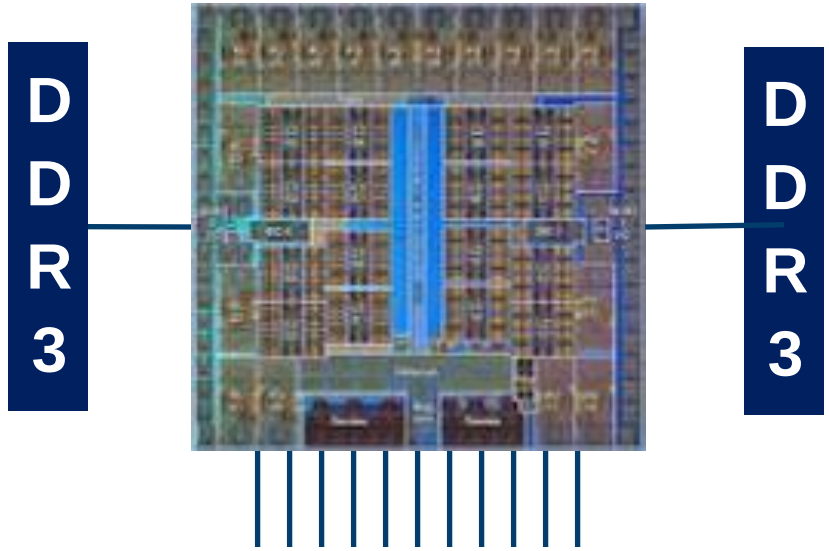
JUQUEEN (until 2018)

Key node features

- Custom processor with 16 A2 compute core
  - 1.6 GHz core clock, 1x 256-bit wide SIMD
  - 32 MByte L2 cache
- 16 GByte DDR3-1300, 2x 128-bit channels
- 10+1 network links (on-chip)
- Small and compact form factor

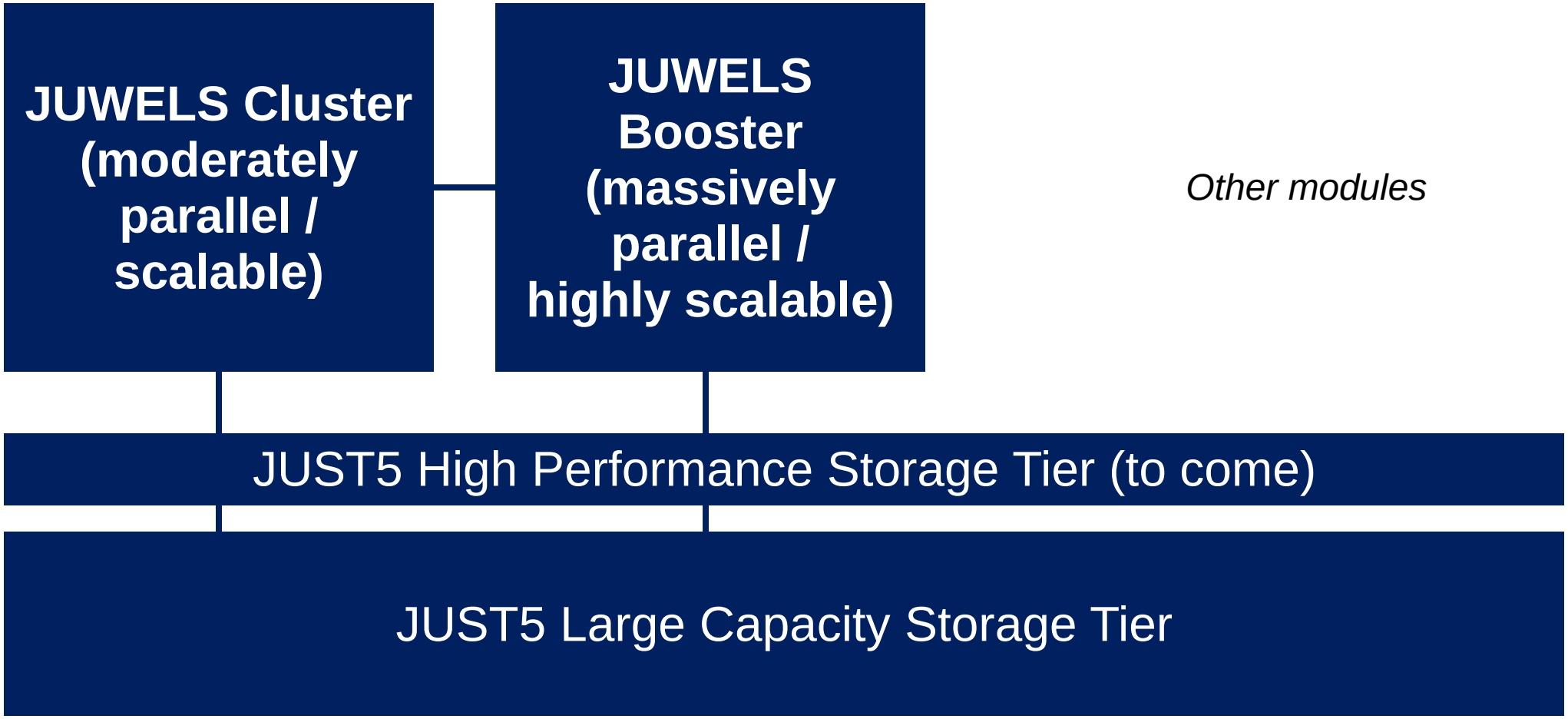
Key system features

- 28,672 nodes
  - 28 racks, 1024 nodes per rack
- 5-dimensional torus network

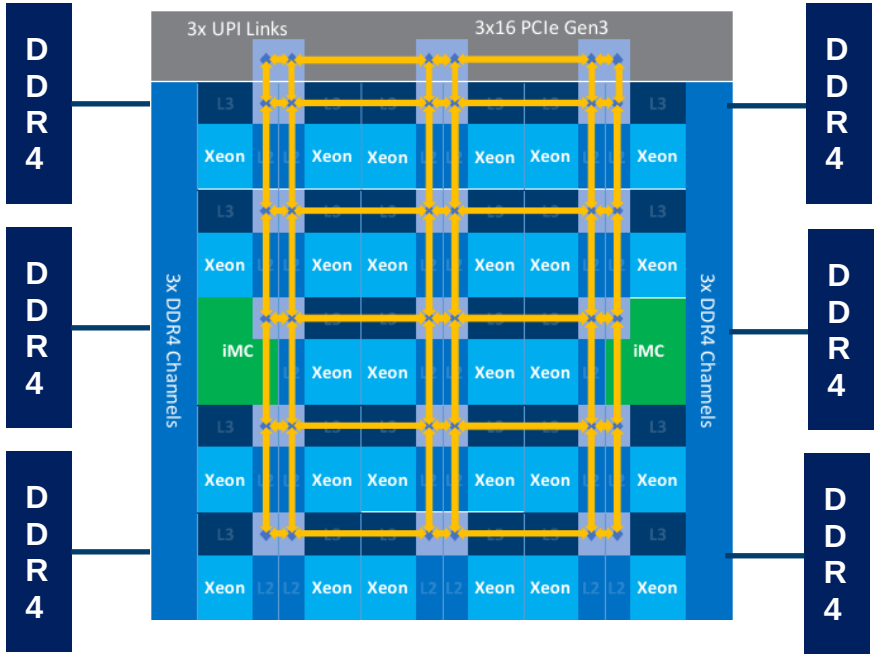


torus network

JUWELS Modules (Starting 2018)



JUWELS Cluster (starting 2018)



Key standard node features

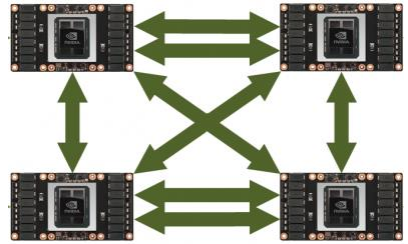
- Dual-socket Xeon Skylake 8168
  - 24 cores at 2.6 GHz, 2x 512-bit wide SIMD
  - 33 MByte L3 cache
- 96 or 192 GByte DDR4-2666, 6 channels
- EDR Connect-X5 port (off-chip)

Key accelerated node features

- 4x V100 GPU island

Key system features

- 2,511 compute nodes, 48 accelerated nodes
- Fat-tree network topology



JÜLICH Forschungszentrum

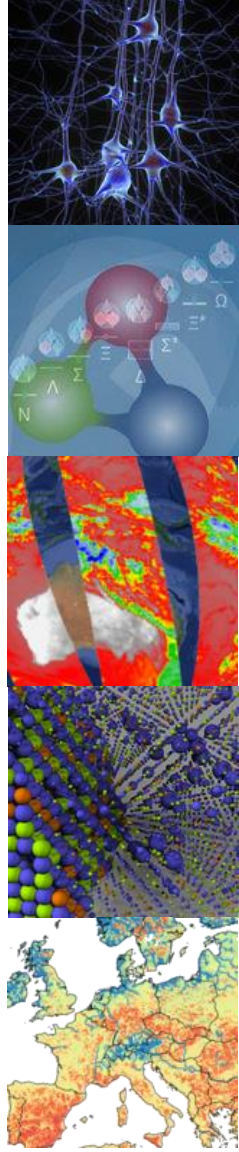
Comparison

	JUQUEEN	JUWELS Cluster	JUWELS Booster
Compute B_{fp} [TFlop/s]	5,872	10,414 (CPU) 1,440 (GPU)	TBA
Compute B_{fp} [Flop/cycle]	$3.7 * 10^6$	$3.9 * 10^6$ (CPU) $1.0 * 10^6$ (GPU)	TBA
Memory bandwidth B_{mem} [TByte/s]	1,223	321 (CPU) 173 (GPU)	TBA
Memory capacity C_{mem} [TiByte]	448	262 (CPU) 3 (GPU)	TBA
Aggregate node injection bandwidth B_{net} [TByte/s]	573	32	TBA

Extreme Scaling: Members of the High-Q Club

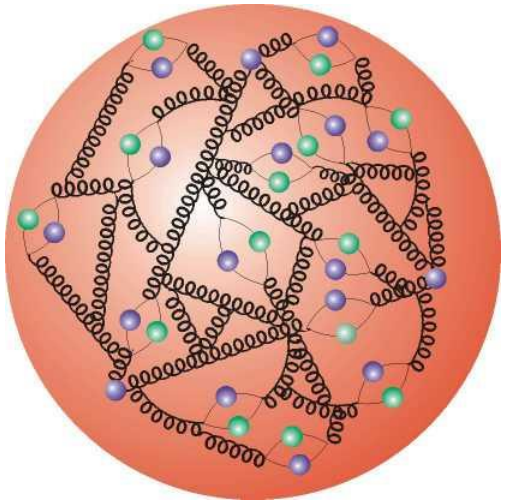
Member = application that scaled to full JUQUEEN

- CoreNeuron
  - Simulating electrical activity of neuronal networks with morphologically-detailed neurons
- dynQCD
  - Lattice Quantum Chromodynamics (QCD) with dynamical fermions
- JURASSIC
  - A fast solver for infrared radiative transfer in the Earth's atmosphere
- KKRnano
  - Korringa-Kohn-Rostoker Green function code for quantum description of nano-materials
- NEST
  - A simulator for spiking neural network models that focus on the dynamics, size and structure of neural systems
- ParFlow+p4est
  - An open-source parallel watershed flow model
- waLBerla
  - A widely applicable Lattice Boltzmann solver from the University of Erlangen
- *… (25 other applications)*



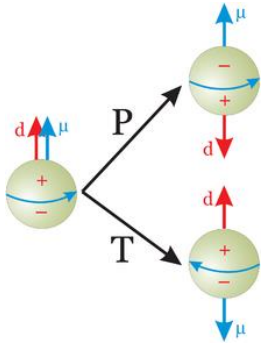
Future Science Challenges

Science Challenge: Fundamental Structure of Matter



Theory of strong interactions: Quantum Chromodynamics

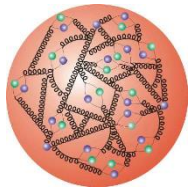
- Quarks are the constituents of matter which strongly interact exchanging gluons
- Particular phenomena:
  - Confinement
  - Asymptotic freedom (Nobel Prize 2004)



Selected science challenge: Upper bound for the neutron electric dipole moment

- Search for evidence for violations of charge and parity symmetry (CP symmetry) violations
- Piece in puzzle to understand the asymmetry of matter and anti-matter observed in the universe

Computational Challenge

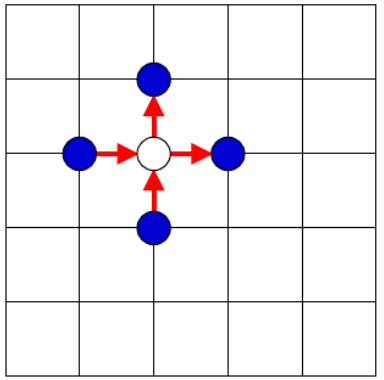


Lattice Quantum Chromodynamics (LQCD)

- Discretised version of the theory

Workflow for computation of physical observables

- Generate chain of gauge field configurations using Monte-Carlo methods
  U0 → U1 → … → UN
- Compute physical observables on the configurations



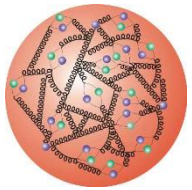
Key kernel: Iterative solver for linear set of equations: $M\chi = s$

- Matrix *M* huge, but sparse
- Typical size for size for state-of-the-art projects: O(10^8)

Key performance signatures

- High level of parallelism and regular control flow
- Low Arithmetic Intensity $I_{fp} / I_{mem} < 1$

Why GPUs?



Performance signatures are good match for GPUs

- High level of parallelism to be exploited with regular control flow
- Need for high memory bandwidth

LQCD one of the first scientific applications to exploit GPUs

- Egri et al., “Lattice QCD as a video game”, 2006 (arXiv:hep-lat/0611022)

Challenges

- Acceleration of non-kernel codes
- Multi-node scalability

Data Transport Challenge

Characterising applications in terms of Information Exchange

- A computation implies that information is transferred from a storage device *x* to *y*
- Information Exchange function: data transferred between storage devices for specific computation *k* as a function of the workload *W*

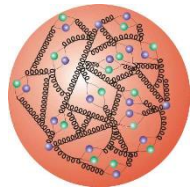
$$I^{k}_{x,y}(W)$$

Examples

- I_{fp} …….. number of floating-point operations
- I_{mem} …….. Amount of data to be loaded/stored
- I_{net} ……... Amount of data to be communicated over the network

Arithmetic Intensity: AI = I_{fp} / I_{mem}

Data Transport Challenge: Network



Model analysis of block based solver

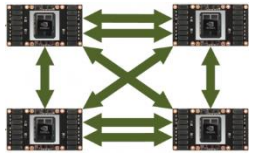
- Problem size $V=96^3 \times 192$
- Aim for communicating data of “white” blocks while ”black” blocks are being communicated

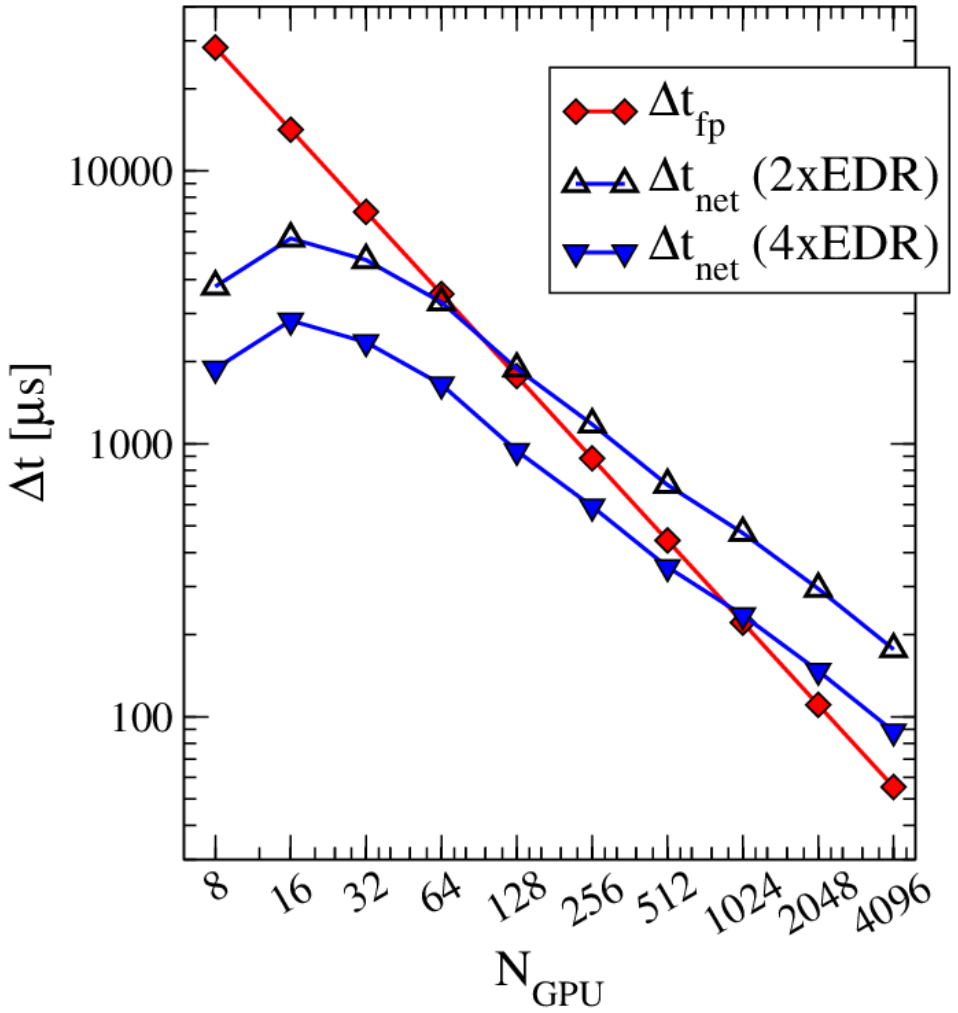
Model Ansatz for single block update

- $\Delta t_{fp} \sim I_{fp} / (\varepsilon_{fp} B_{fp})$
- $\Delta t_{net} \sim I_{net} / (\varepsilon_{net} B_{net})$

Architecture model

- 4 accelerators with B_{fp}=7.5 TFlop/s each
  - ε_{fp} = 10%
- 0.5-1 network ports per GPU with B_{net}=100 Gbit/s
  - ε_{net} = 90%



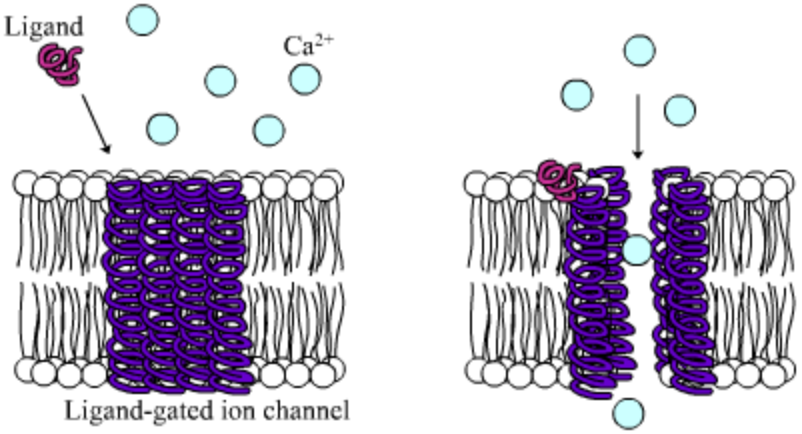


Science Challenge: Bio-Physics

Science challenges

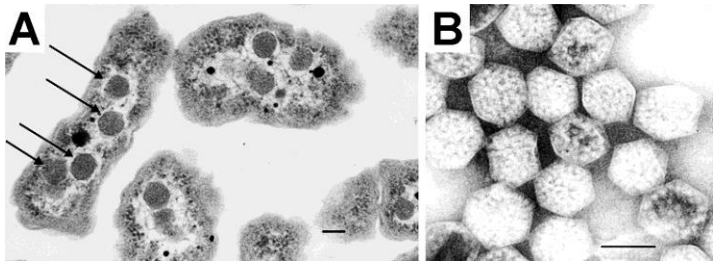
- Investigate slow ligand-controlled ion channels, i.e. enable numerical simulation of biological processes at time-scales the millisecond to second range
- Investigate numerically sub-cellular regions (organelles) involving O(10^8) atoms

[Wikipedia]



Needs

- Extremely large number of simulations to feed a Markov state model ☞ ensemble simulation
- Scalable simulations to accommodate very large number of atoms



[Wikipedia]

Application of choice: AMBER

Computational Challenge and Why GPUs?

Particle Mesh Ewald Molecular Dynamics

- Most time spent in computing potential energy, e.g. Coulomb term

$$V_{Coulomb} = V_{direct} + V_{reciprocal} + V_{self}$$

$$V_{direct} = \frac{1}{2} \sum_{\mathbf{n}} \sum_{i,j}^{N} q_i q_j \frac{\mathrm{erfc}(\alpha r_{ij,\mathbf{n}})}{r_{ij,\mathbf{n}}}$$

$$V_{self} = \frac{-\alpha}{\sqrt{\pi}} \sum_{i=1}^{N} q_i^2$$

$$V_{reciprocal} = \frac{1}{2\pi\nu} = \sum_{\mathbf{m}} \frac{\exp(-(\pi m/\alpha)^2)}{\mathbf{m}^2} S(\mathbf{m}) S(-\mathbf{m})$$

Implicit solvant simulations

- Compute potential energy in the presence of a solvant using the generalized Born algorithm

Performance signatures are good match for GPUs

- High level of parallelism and reasonably regular control flow
- Typically higher arithmetic intensity

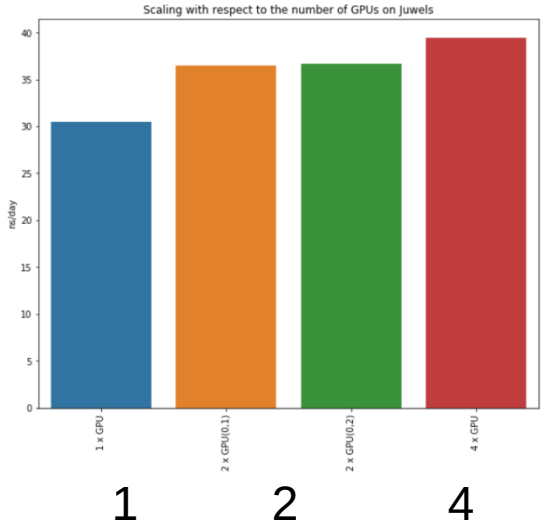
Scalability Challenge

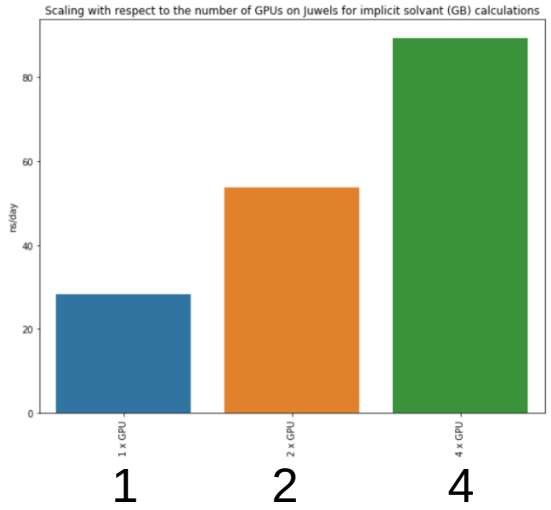
Explicit solvent calculation with 1,067,095 atoms

- Unbalanced distribution of kernels over GPUs
- Some computations repeated on all involved GPUs
  ☞ No scaling

Implicit solvent: Nucleosome benchmark with 25,095 atoms

- Balanced distribution of all kernels over GPUs
- Good scaling behaviour for all main kernels





Number of GPUs on 1 JUWELS node

[K.Mood, 2018]

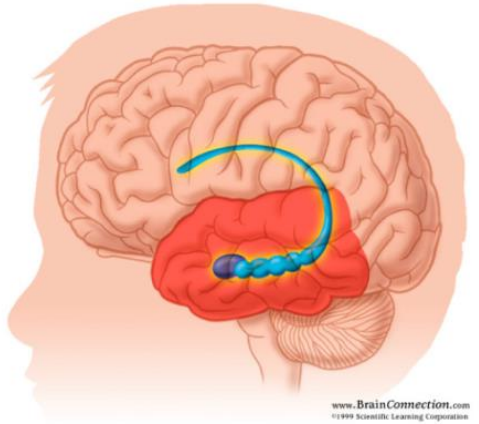
Science Challenge: Decoding the Human Brain

Overall research challenge

- Create an understanding of brain at different spatial and temporal scales
- Help to address dysfunctions of the brain causing mental diseases including Alzheimer

Create high-resolution atlases of the human brain

- Uncovering details of the fiber architecture

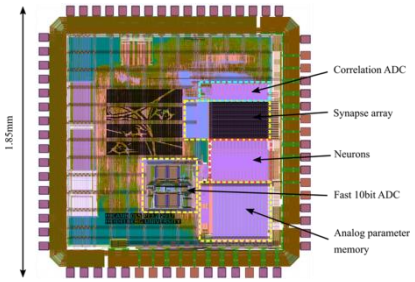


Create realistic models of the human brain

- E.g., create models of synaptic plasticity of hippocampal synapses
- Create understanding of higher brain functions (learning, memory, spatial navigation)

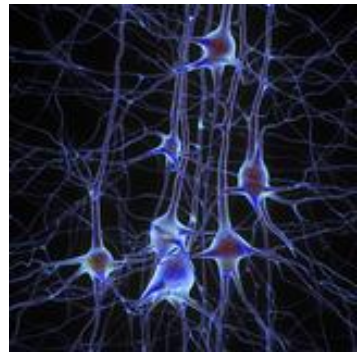
Develop brain-inspired technologies

- Neuromorphic computer architectures
- New approaches to machine learning



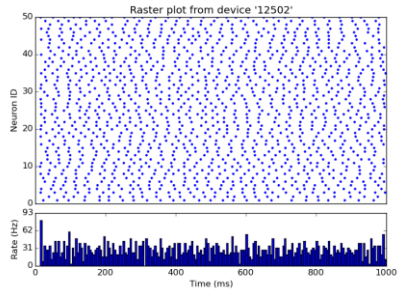
BrainScaleS 2

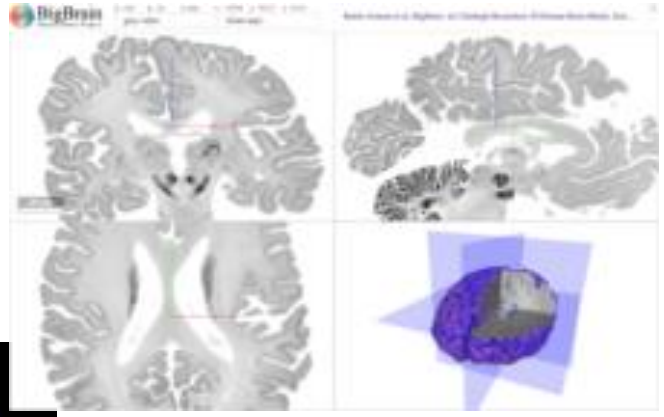
Computational Challenges



Simulations

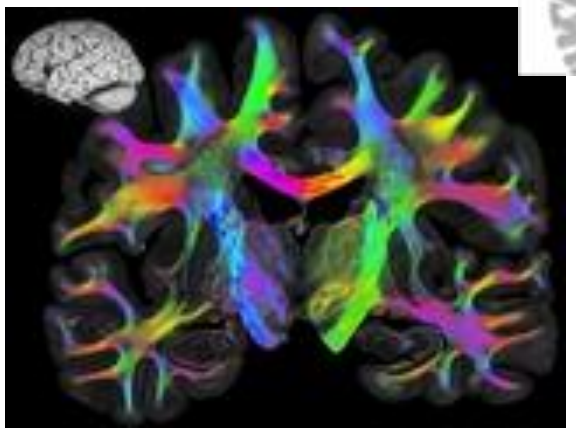
- Challenge of addressing different spatial and temporal scales
- Need for different (and possibly coupled) simulators
  - Molecular level simulations
  - Simulation with morphologically-detailed neurons
  - Simulations with networks of realistic complexity





Histological data analysis

- Image registration
- Automatic image analysis
- Fibre tracking



Why GPUs?

Acceleration of simulation

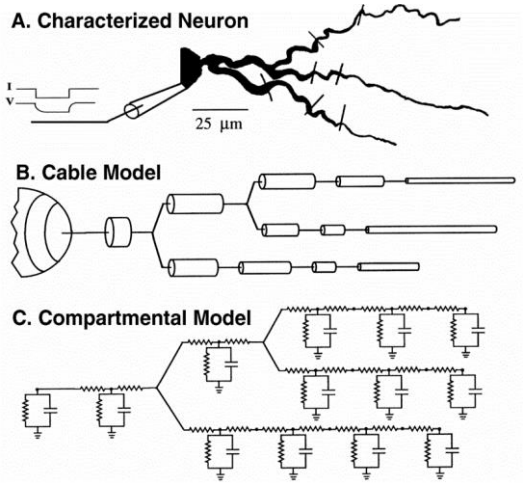
- Simulation of morphologically-detailed neurons

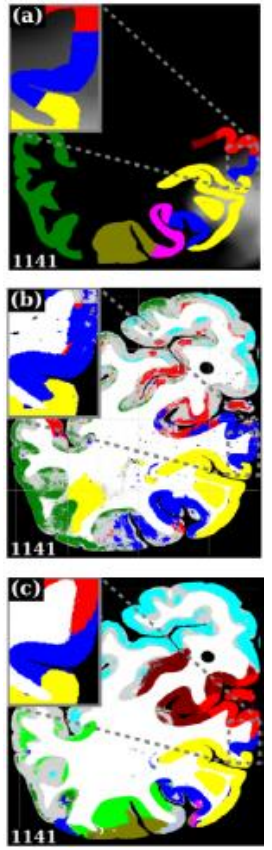
Analysis of data

- Computing mutual information metric during image registration
- Machine learning for automatised image analysis

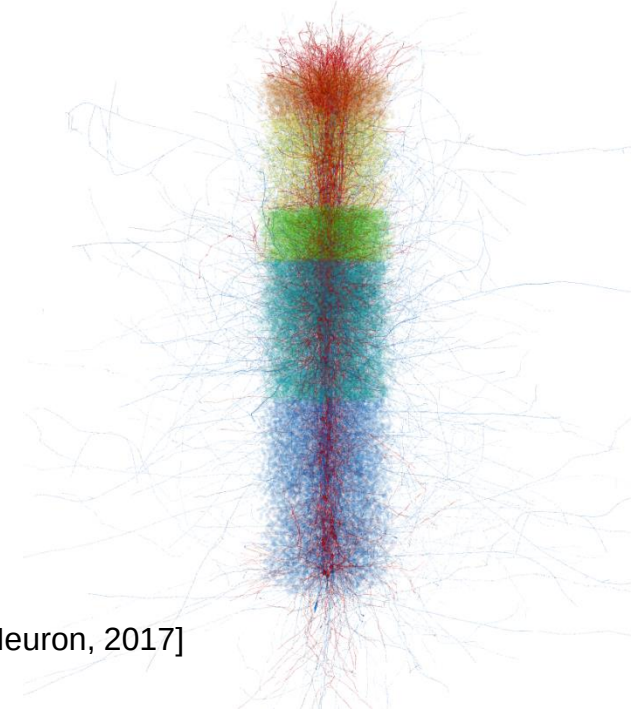
Visualisation

- Navigate through brain image data
- Monitor or explore simulation results



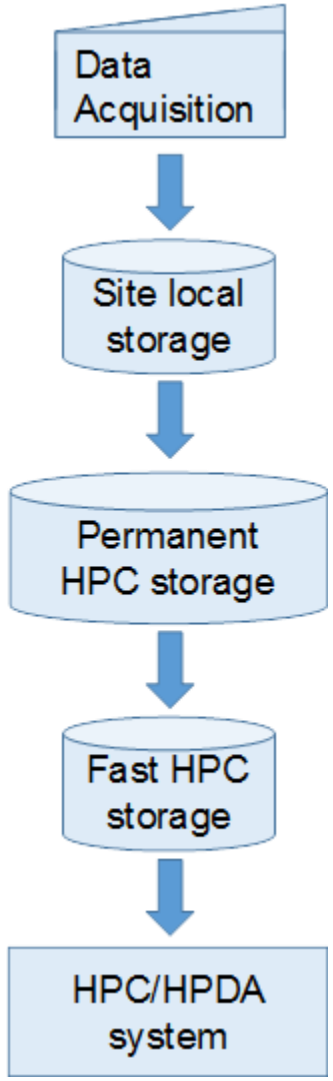


[Spitzer et al., 2017]



[RTNeuron, 2017]

Data Transport Challenge



Data volume challenge

- Data acquisition using 8-9 microscopes at ~15 MByte/s → ~10 TByte/day
- Archival data increase 1-2 PByte per year

Data analytics challenge

- Deep learning methods for automatised cell segmentation
- Massively parallel feature extraction (e.g. cell count)
- Image registration

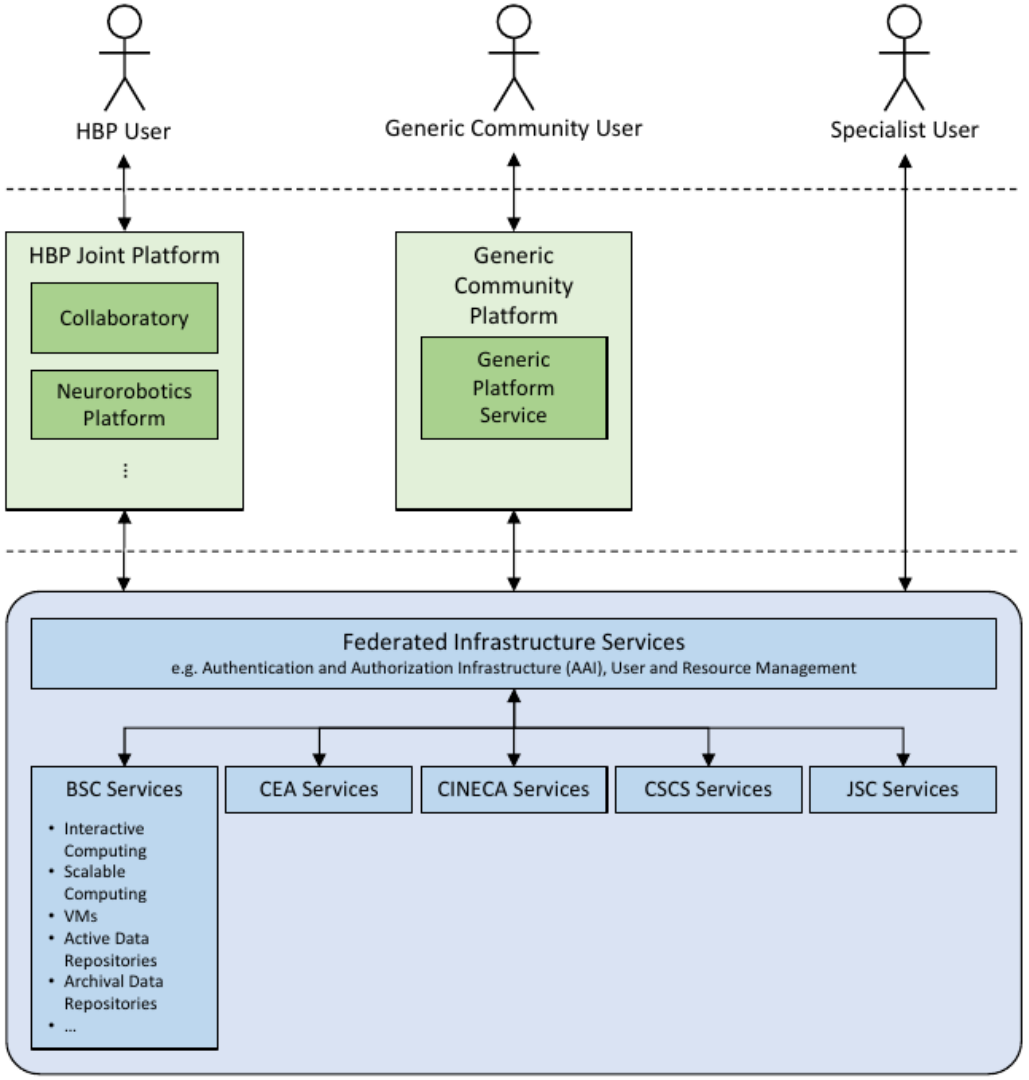
Data access speed challenge

- Learning data set of size ~100 TByte
- Example of multi-GPU training requirements (using P100 GPUs)

$$\Delta t_{IO} < \epsilon \Delta t_{comp} \quad \rightarrow \quad B_{IO} > \frac{N_{GPU}}{\epsilon} \, 0.3 \, GByte/s$$

Infrastructure for Brain Research: Concept

FENIX RESEARCH INFRASTRUCTURE



User communities

Platform services

Infrastructure services

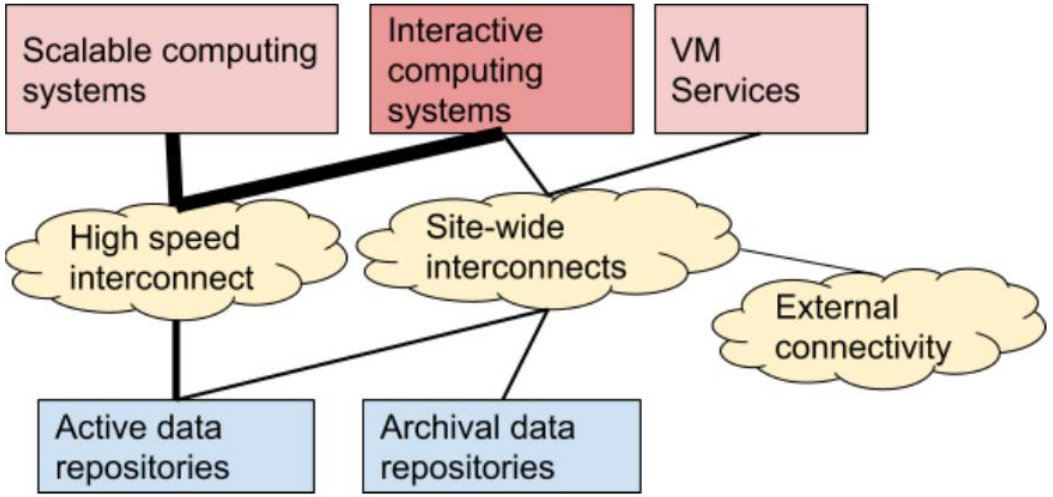
http://fenix-ri.eu

Infrastructure for Brain Research: Services

FENIX RESEARCH INFRASTRUCTURE

Computing services

- Interactive Computing Services
- (Elastic) Scalable Computing Services
- VM Services



Data services

- Active Data Repositories
- Archival Data Repositories
- Data Mover Services, Data Location and Transport Services

Service federation

- Cope with variety of data sources and make data available to the wider community
- Provide access to a diversity of computing capabilities at different sites

Outlook and Conclusions

High-Q Club 2.0

CoreNeuron

- Initial ports to GPU available
- Likely to scale well due to lacking communication challenges

dynQCD

- Suitable for GPUs, but scaling challenge

JURASSIC

- Port to GPUs available
- Will scale well assuming sufficient I/O capabilities

NEST

- Porting to GPUs very challenging

ParFlow+p4est

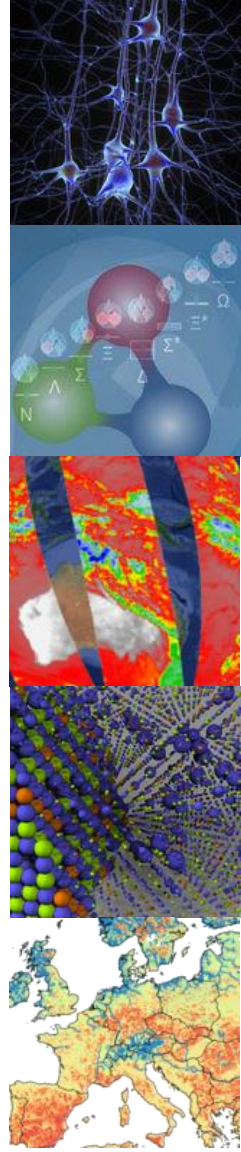
- Suitable for GPUs

waLBerla

- Suitable for GPUs
- Good scaling achievable (easier compared to LQCD)

Other 25 applications

- Porting a challenge in most cases



Addressing the Data Transport Challenge / System Balance

Memory bandwidth B_{mem}

- Memory bandwidth is (given the current compute performance levels) critical for many scientific computing applications
- Use of high-bandwidth memory for GPUs helps pushing aggregate memory performance more compared to what is possible with currently available CPUs

Network bandwidth B_{net}

- Some applications require 100 Gbit/s per V100 class GPU

I/O bandwidth B_{IO}

- Some Machine Learning applications require several GByte/s per GPU

Conclusions

Current technologies result in difficult balance between compute and data transport capabilities

- Bandwidth more often the limiting factor
  - Another way of saying HPL is not relevant for many relevant science applications
- Current accelerators like GPUs are moving in the right direction

Many science challenges need scalable architectures

- Need extreme scale compute and capability of processing extreme scale data volumes
- Significant efforts needed to get applications scale like they did on Blue Gene/Q

But: A lot of exciting science is ahead of us