HPC SYSTEMS IN THE NEXT DECADE - WHAT TO EXPECT, WHEN, WHERE

Dirk Pleiter | CHEP 2019, Adelaide | 06.11.2019

JÜLICH
Forschungszentrum

Disclaimer

[wikimedia]

- Long-term technology trends provide guidance
- Budget constraints can change
- (Non-HPC) Market trends are of critical importance and can change

Outline

- **Introduction**
- **Exascale hardware technologies**
- **Exascale programming**
- **Future HPC infrastructures**
- **Summary and conclusions**

INTRODUCTION

What Is Exascale Computing?

Introduction

- Top500 continues to be important metric for governments and funding agencies

Answer 2: HPC infrastructure allow to address new science and engineering challenges providing 10-100x more performance compared to today's systems

- Focus on purpose of supercomputers: enable science and engineering
- No performance metric fits all
- Monolithic supercomputers may in the future become less relevant

Canonical Exascale Computing Challenges

Reducing power requirements

- Few sites can afford >10-15 MW

Exploiting massive parallelism

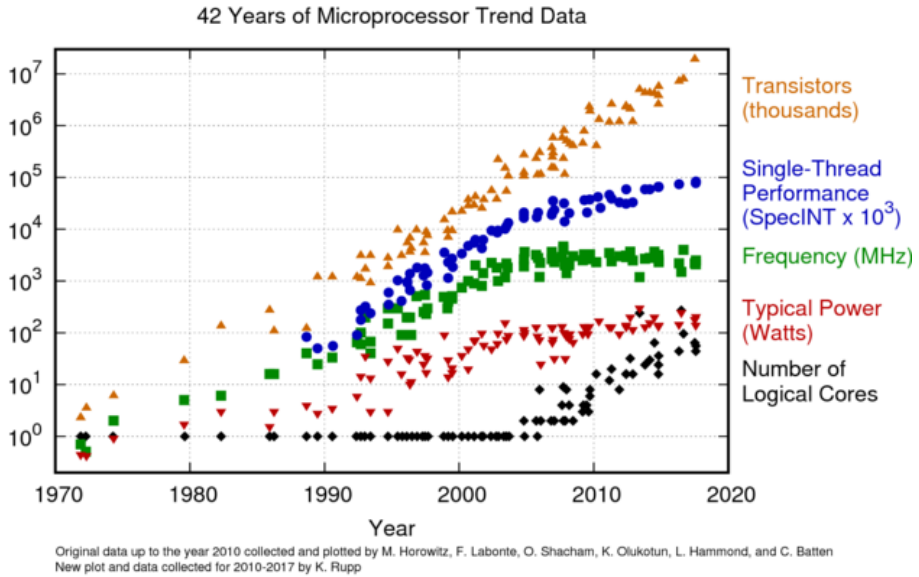
- More computational performance
  → more parallelism
  → scalability challenge

Maintaining a balanced system

- Improved data transport capabilities
- Flops are cheap, but data transport is expensive

Coping with run-time errors

- More components → higher risk of system failures



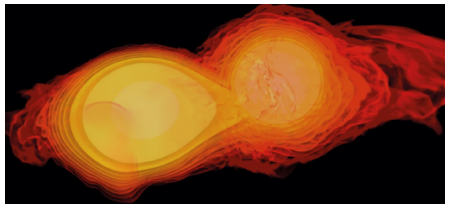
JÜLICH Forschungszentrum

Exascale Science Cases

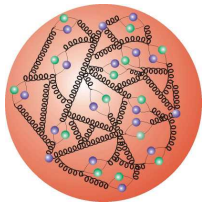
E. Lindahl, S. Ryan et al., “The Scientific Case for Computing in Europe 2018-2026”, October 2018

Fundamental sciences

- Astrophysics, cosmology
- Particle physics

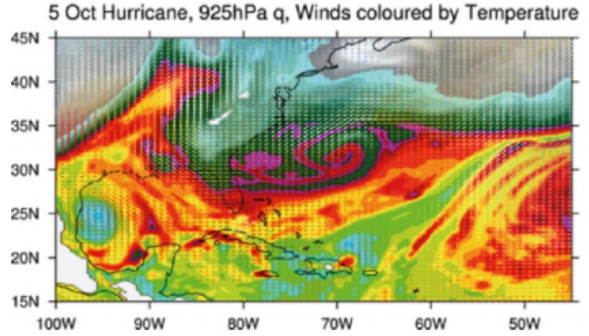


[Luciano Rezzolla]



Climate, weather and earth sciences

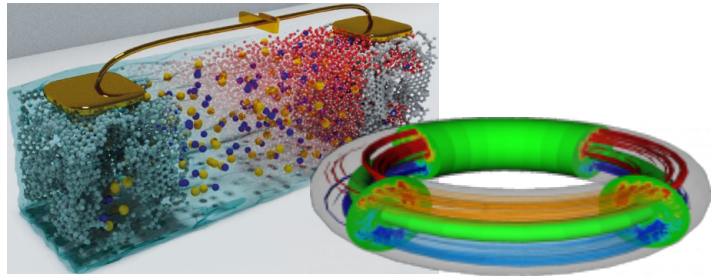
- Understanding and Predicting a Changing Climate
- Accurate Weather Forecasting and Meteorology



[Luigi Vidale/University of Reading]

Materials science and energy research

- Material properties from atomic and electronic structure simulations
- Improving lower-CO2-impact energy source

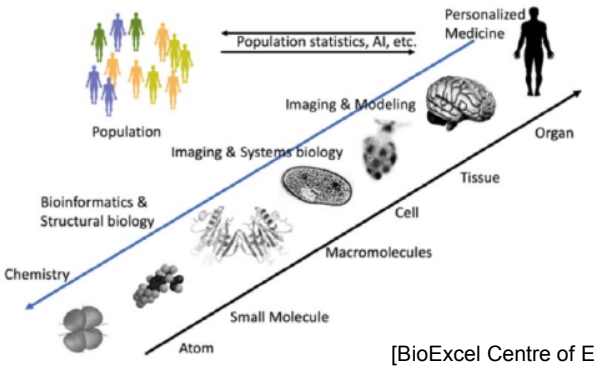


[Mathieu Salanne, EPFL]

[J. Bigot et al., CEA]

Life sciences and health

- Molecular life science and structural biology
- Neuroscience



[BioExcel Centre of Excellence]

JÜLICH Forschungszentrum

Exascale Programs

US

- 3 exascale systems through CORAL2 procurement
- ECP project: application development, software technology, integration

Japan

- 1 exascale system at RIKEN
- Application-driven co-design of hardware

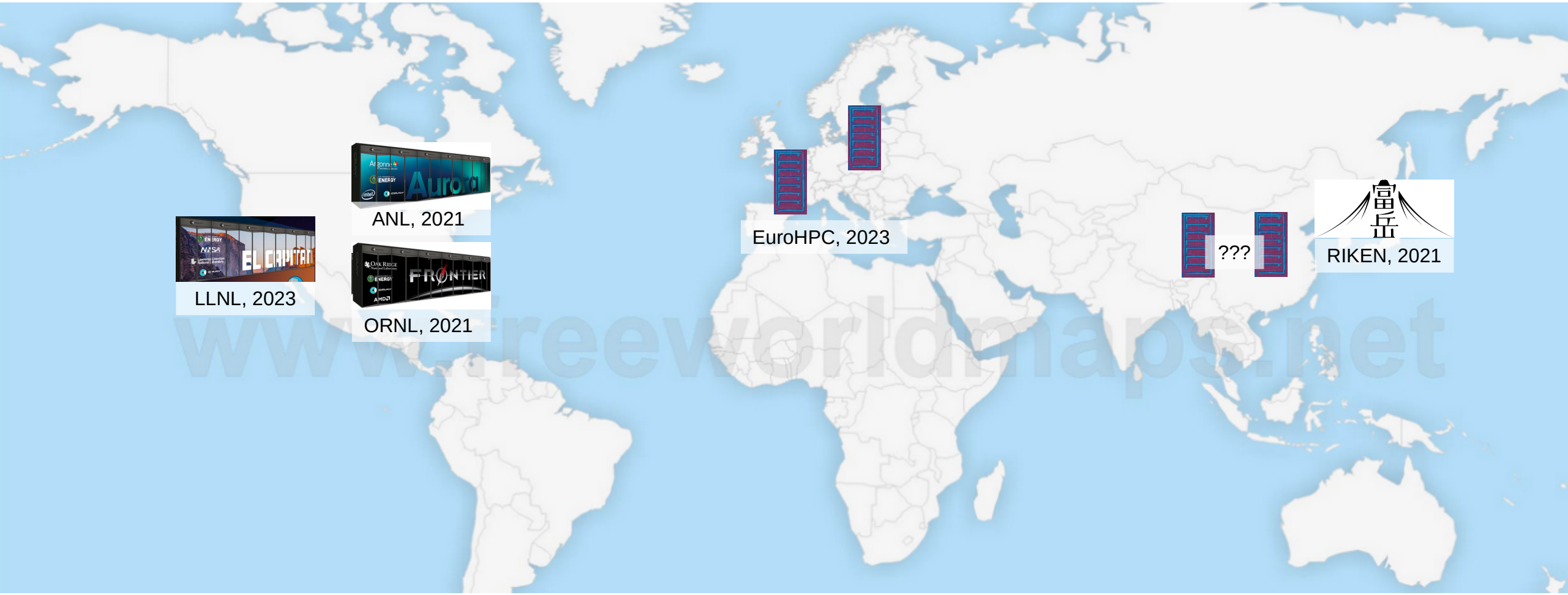
Europe

- 2 exascale systems through Joint Undertaking EuroHPC
- Research and innovation program, Centres of Excellence

China

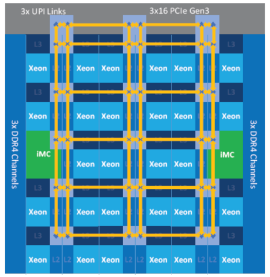
- Details on exascale system deployments less clear
- Research program addressing development of applications and HPC environment

Exascale Systems Planning



EXASCALE HARDWARE TECHNOLOGIES

Diversifying Landscape of CPUs for HPC

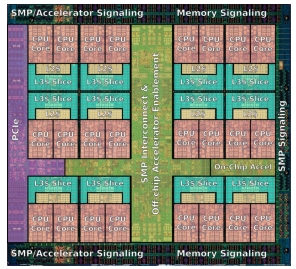


Intel Xeon

- High compute capabilities, relatively low memory bandwidth
- Dominating market position

AMD EPYC

- High compute capabilities, slightly better memory bandwidth
- Emerging market position

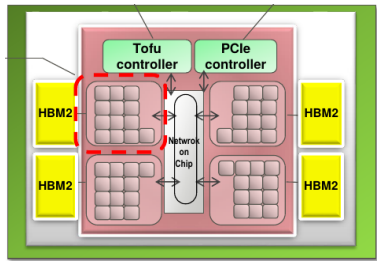


IBM POWER9

- Low compute capabilities, relatively good data transport capabilities
- Used for current Top500 #1 and #2, lacking HPC market uptake

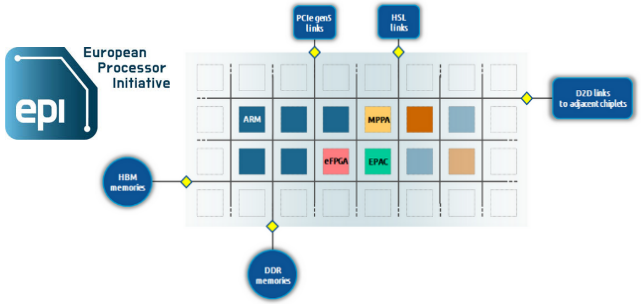
Marvell ThunderX2

- Moderate compute capabilities, relatively good memory bandwith
- Still low overall market uptake



Fujitsu A64FX

- High compute capability and memory bandwidth
- Used for Japan's exascale system



Upcoming: European Processor Initiative

CPU: Core vs. Vector Parallelism

Strategies for increasing CPU-level parallelism of floating-point operations

- Increase number of cores
- Increase width of SIMD/vector instruction

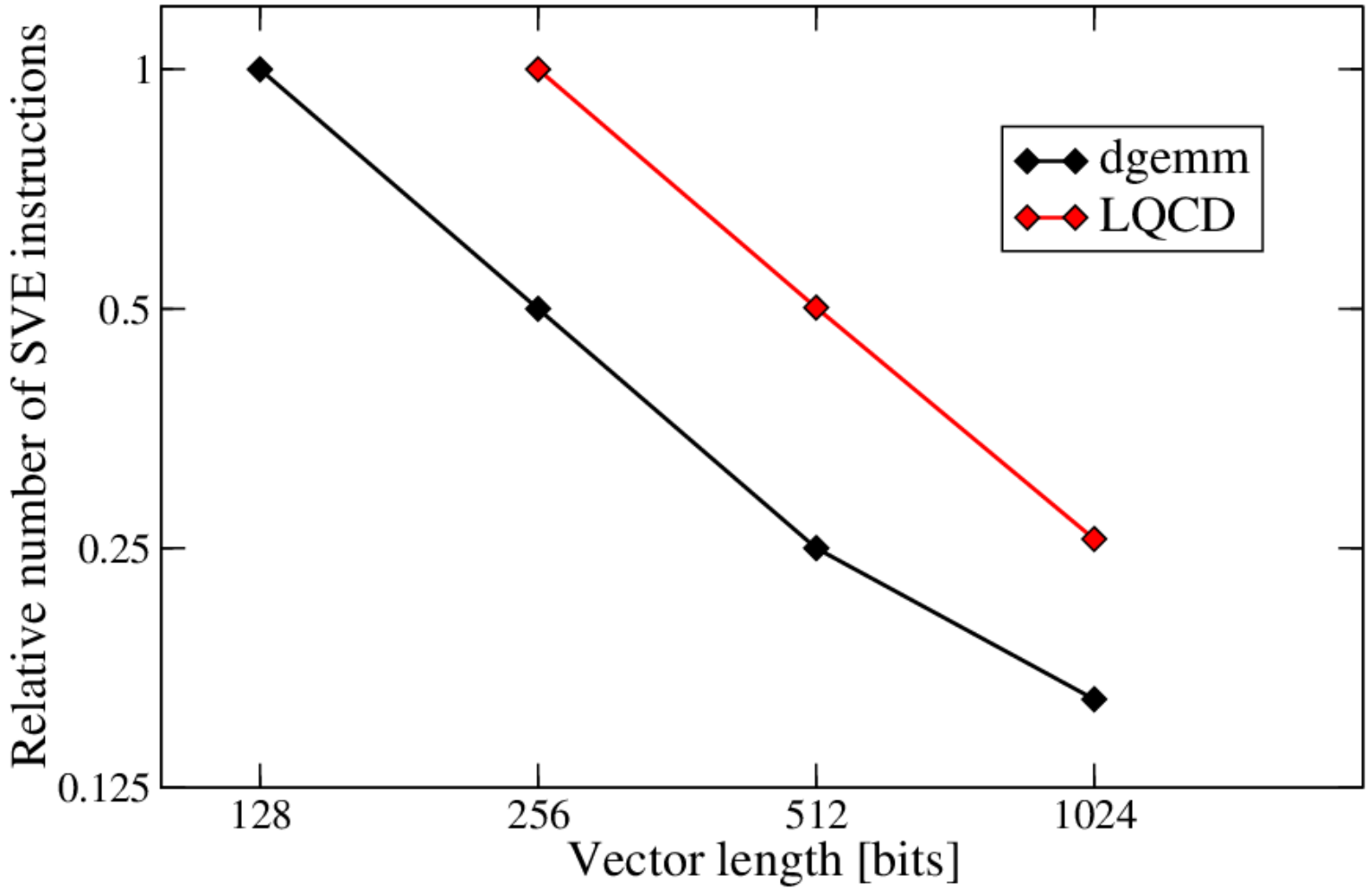
	Increase core count	Increased SIMD/vector width
Pro	More flexible thread-parallelism	Simplified hardware architecture
Con	Replication of instruction front-end	Increased power-consumption per node; Vector/SIMD parallelism more difficult to exploit

→ **Need for trade-off decisions**

New vector ISA: Arm Scalable Vector Extension (SVE)

- Vector length agnostic → multiple lengths supported by ISA: 128, 256, …, 2048 bit

Exploring Arm's Scalable Vector Extension



Change of number of SVE instructions as a function of the vector length

Power Efficiency: State of Affair

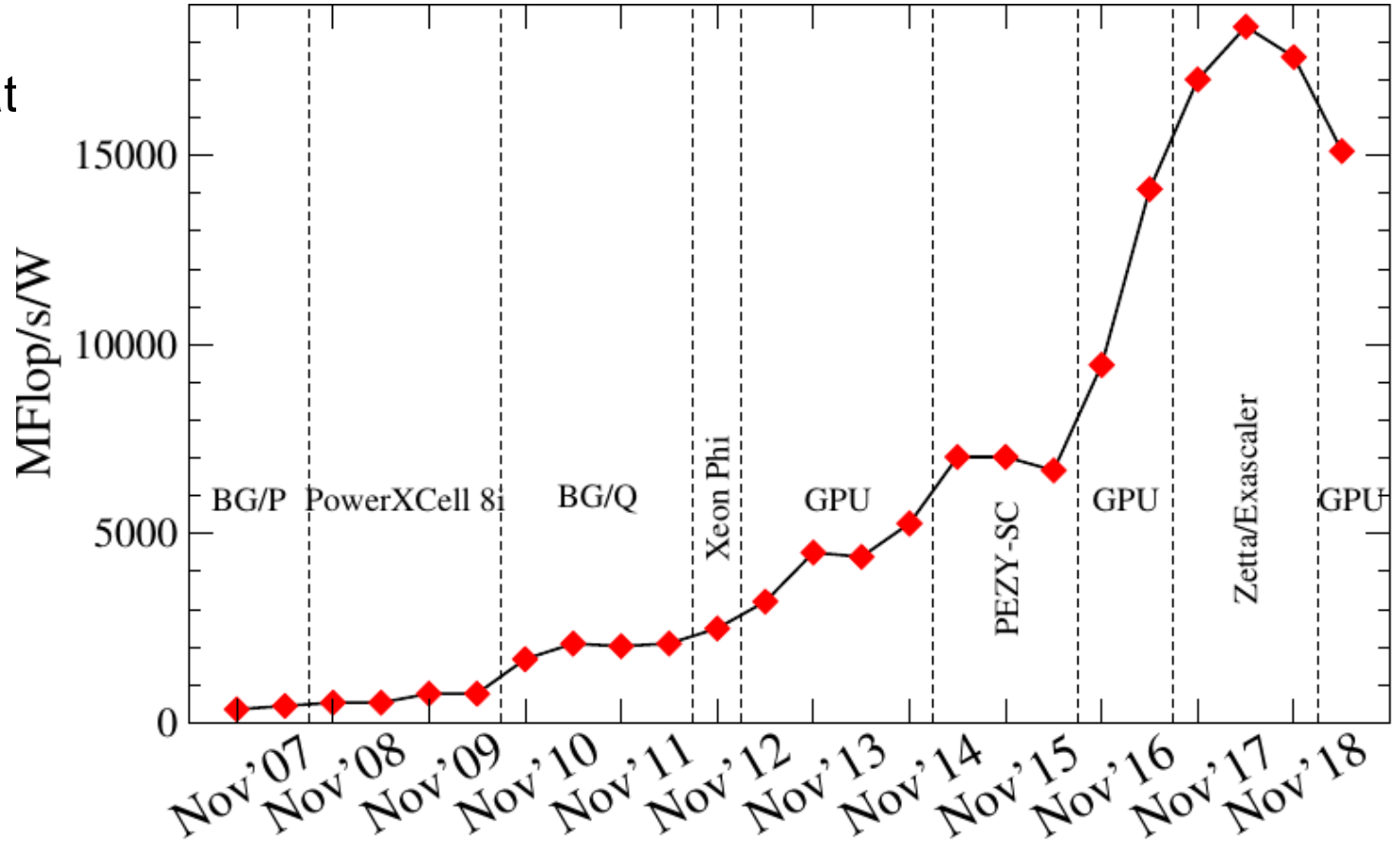
Exascale targets

- US labs: 30 MWatt → 33 GFlop/s/Wat

Green500 list (June 2019)

- #1: 15 GFlop/s/Watt
- Best Xeon-based system at #27: 6 GFlop/s/Watt

→ **Use of compute accelerators unavoidable**

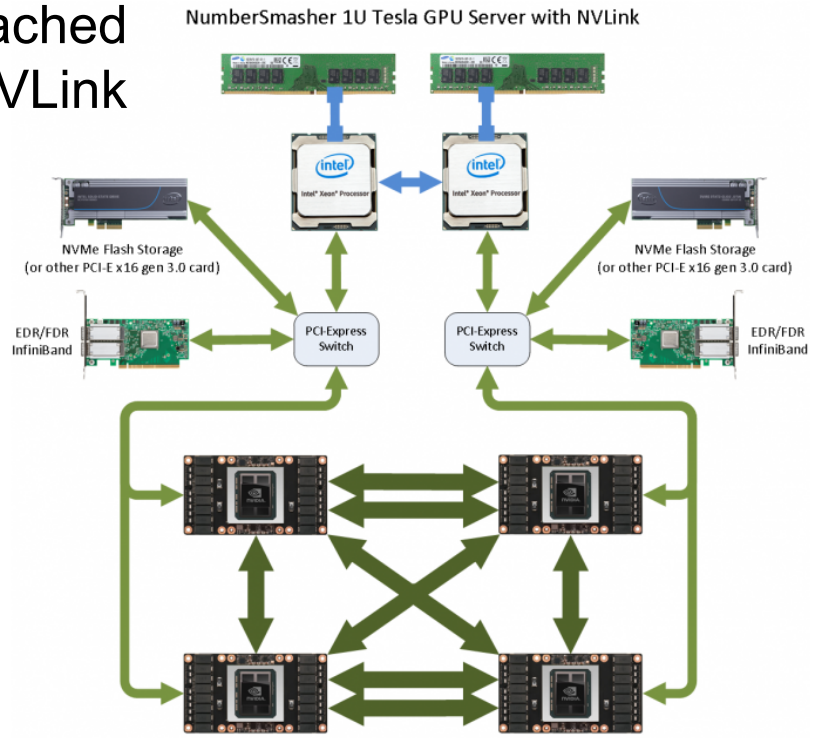


JÜLICH Forschungszentrum

Compute Accelerators for HPC: Today and Near Future

Today: NVIDIA GPUs

- Group of 3-4 Tesla GPUs interconnected via NVLink
  - Up to ~20 TFlop/s per group
- 1-2 CPUs attached via PCIe or NVLink



Future competitors: AMD and Intel GPUs

- Intel Xe GPU set for Aurora at ANL
- AMD GPUs set for Frontier at ORNL and El Capitan at LLNL
  - Similar integration approach as for NVIDIA today

Common future trend: Improve integration of discrete accelerator and CPU through coherent I/O interfaces

JÜLICH Forschungszentrum

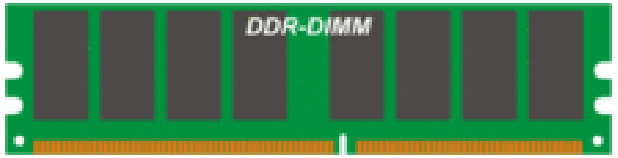
Memory Technologies (1/2)

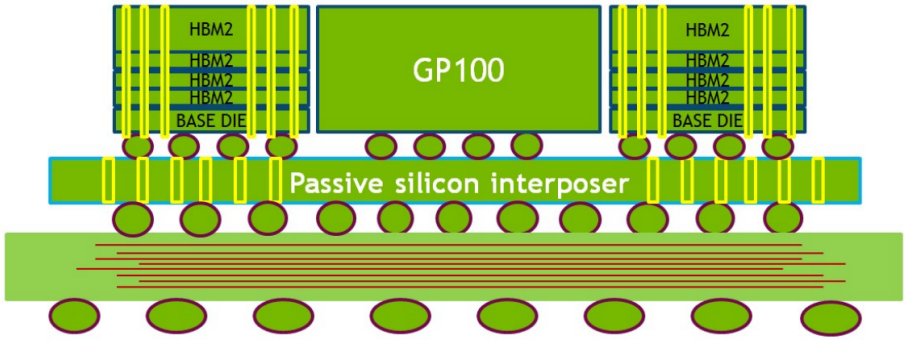
Desirable memory performance features

- Large memory capacity C_{mem}
- High memory bandwidth B_{mem}

Main types of memory

- DDRx
  - DDR4: Up to ~200 GByte/s per CPU using 8 channels
  - DDR5: Aims on doubling performance
  - Capacity of O(100) GiByte
- High-Bandwidth Memory (HBM)
  - Today: ~250 GByte/s per stack
  - Near future: ~400 GByte/s per stack
  - Capacity of O(10) GiByte
    - Capacity per stack announced to double
  - More power efficient (lower data rate per line, shorter traces)

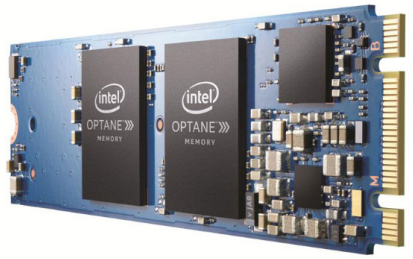




Memory Technologies (2/2)

Main types of memory (cont.)

- Non-volatile memory
  - Technologies: NAND Flash, 3D-Xpoint
  - Attachment via I/O interface or memory bus
  - Large variety of software interfaces



Significant differences in technology for $\Delta\tau = C_{mem} / B_{mem}$

NVIDIA V100 GPU	HBM2	$\Delta T \approx$ 20-40 ms
JUWELS compute node	DDR4	$\Delta T \approx$ 0.4 s
Intel Optane DC	Memory attached 3D XPoint	ΔT = O(35-140s)
Intel DC P4511 2 TByte	PCIe attached NAND Flash	$\Delta T \approx$ 1300 s

Deeper Memory Hierarchies

Rationale

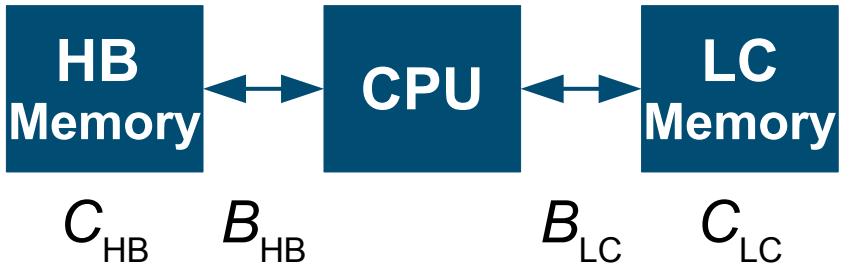
- High-bandwidth memory tier based on technologies like HBM (maximise performance/€)
- Large-capacity memory tier based on technologies like DDR4/DDR5 or new high-density memory technologies (maximise capacity/€)

Need for hardware parameter trade-off decisions

- Bandwidth and capacity ratio high-bandwidth vs. large-capacity memory tier
- Hardware support for data transport between both memory tiers

Use cases

- Separation of hot and cold data objects
- Staging/double buffering of kernel data



Brief Outlook on Network Technologies

Main high-end technologies

- Infiniband
- Slingshot

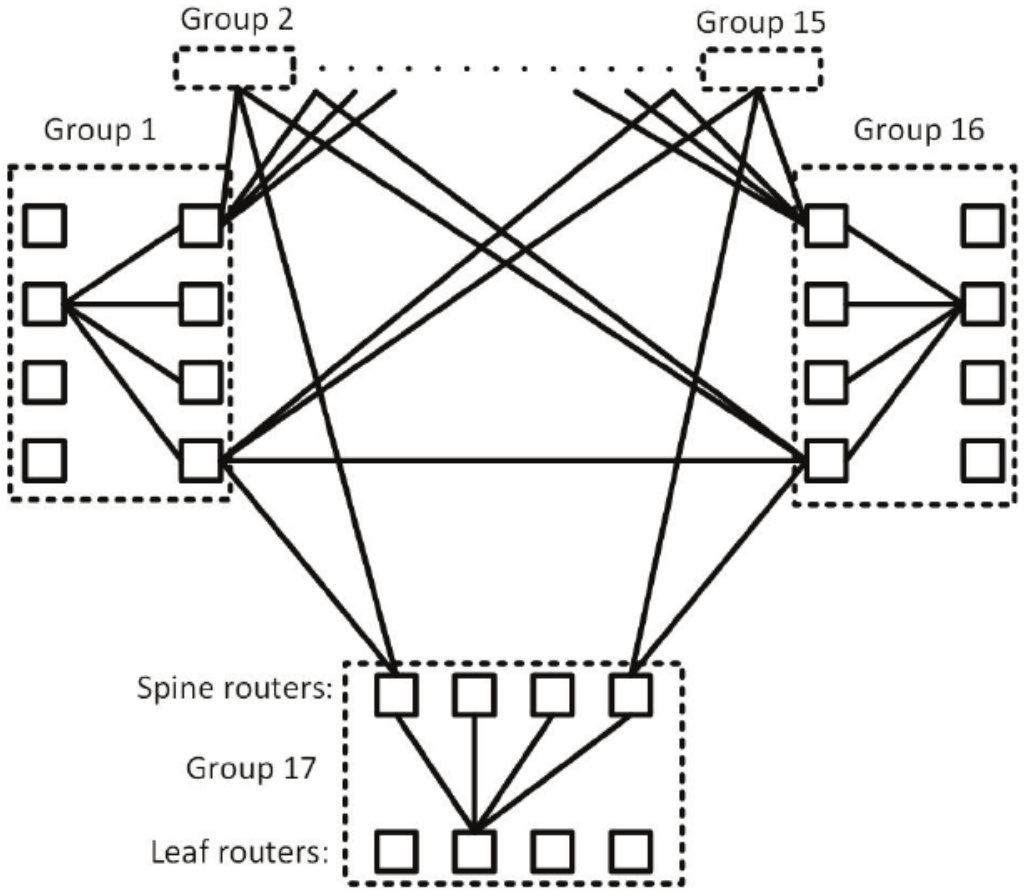
Linkspeed

- 200 Gbps during the next few years

Topologies

- Dragonfly topology likely prime choice for largest systems
  - Cray Slingshot
  - Mellanox Dragonfly+

[Alexander Shpiner et al., 2017]



Known Exascale Architecture Swim Lanes

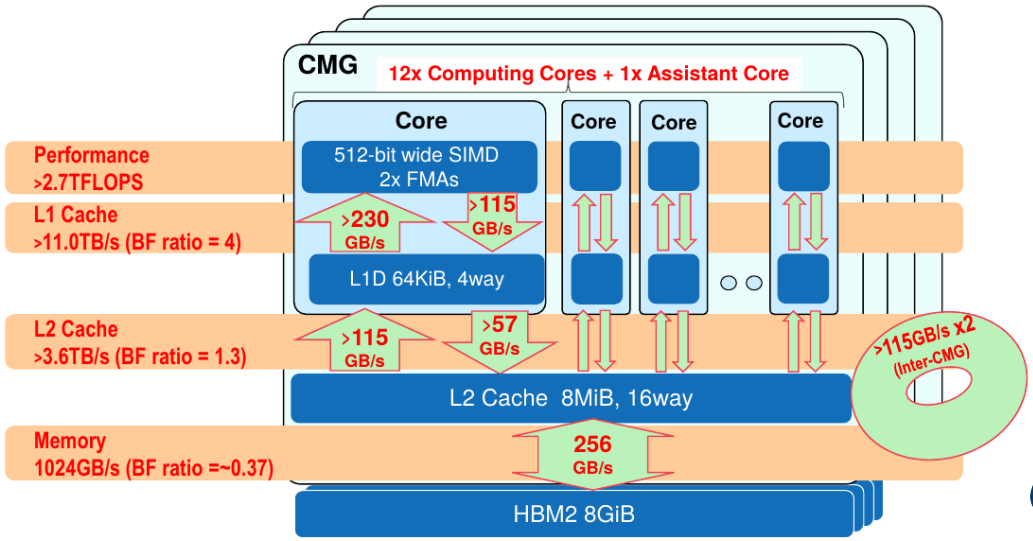
GPU-based: Frontier, Aurora, El Capitan

- Node architecture
  - 1-2 CPU
  - Multiple GPUs
  - ≫40 TFlop/s per node
    → O(10,000) nodes
  - $B_{fp}/B_{mem} \gg 5$, B_{mem} = ?
- Dragonfly-type network

Commonality: Use of high-bandwidth memory technologies

CPU-based: Fugaku

- Node architecture
  - 1 CPU
  - >2.7 TFlop/s per node
  - $B_{fp}/B_{mem} \approx 3$, B_{mem} = 32 GiByte
- 6-dimensional torus network



[T. Yoshida, 2018]

JÜLICH Forschungszentrum

Hierarchical Storage Architectures

Increasing diversity of storage device technologies used for HPC

Upcoming HPC challenges and trends

- Scale-up to ≫O(10) TByte/s bandwidth
  - Mandates storage architectures becoming (more) hierarchical
- Enable near-node storage
  - Drop separation of compute and storage cluster
- Mitigate metadata performance limitations
  - Promote APIs beyond POSIX

Devices currently (or soon) in use at JSC:

Technology	Capacity [TiByte]	Bandwidth [GByte/s]	C [s]
Intel Optane DC	0.125-0.5	~4 (r/w mix)	35-140
Intel Optane SSD	0.4-1.5	~2	210-830
Toshiba CM5 SSD	0.8-6.4	~3	290-2300
Lenovo 01DC407	1.2	~0.30	4,500
Seagate Tatsu series	10	~0.17	65,000

SAGE: Hierarchical Object Store

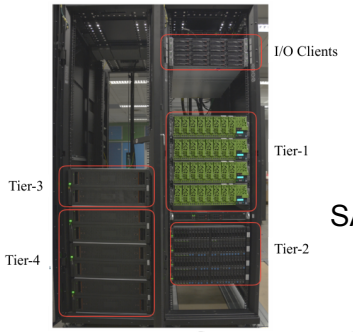
Approach

- Create hierarchical storage architecture based on
  - Advanced object storage technology = MERO
  - Multiple tiers with storage devices with different characteristics
  - Integrated compute capabilities
- Make new architecture usable

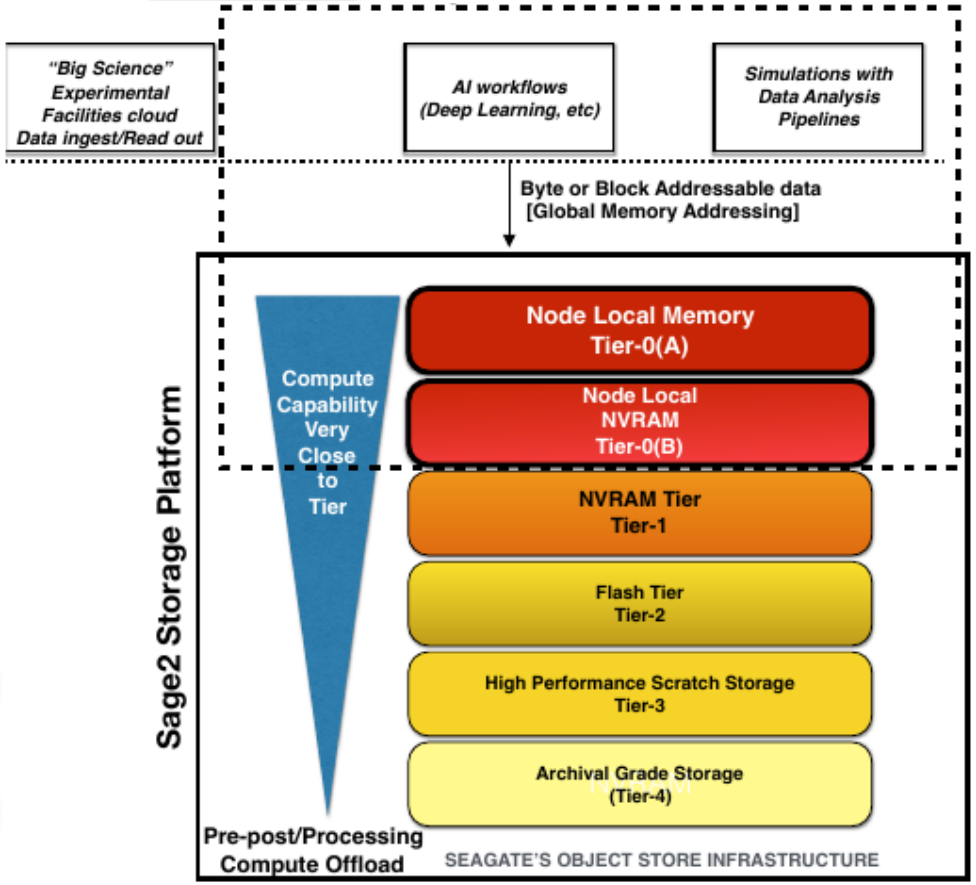
Native object storage platform

- Scalable re-writable fault-tolerant data objects
- Index store with key-value indices
- Support of “composite layouts” with objects distributed over multiple tiers
- Resource management capabilities

→ **Poster A. Davis**



SAGE prototype @ JSC



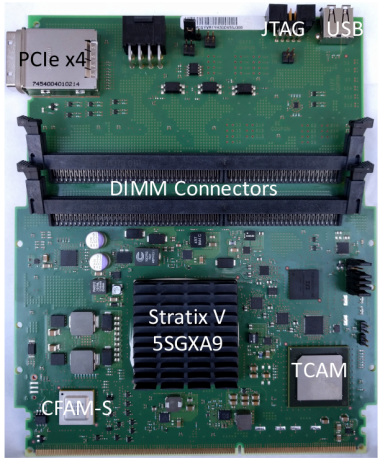
Emerging Hardware Architectures and Technologies

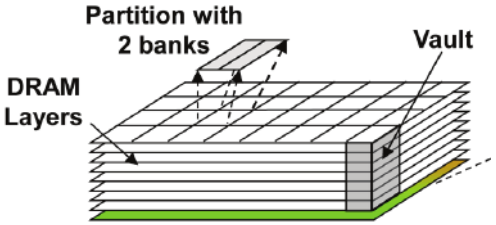
Data-flow architectures

- Promising approach for improving energy efficiency
- Today based on FPGA, in future other options might emerge

In-memory compute / Near-Memory Acceleration

- Rationale: Reduce energy by avoiding data movement by moving computation to data
- Different implementations
  - Compute in stacked memory, e.g. AMC
  - Compute in memory buffer
  - Compute in PCM
- Challenge: Programming model





[R. Nair et al., 2015]

[J.v. Lunteren, 2016]

JÜLICH Forschungszentrum

PROGRAMMING

Programming Models

Programming models and run-time systems critical for coping with increasing hardware complexity, but choice increases

- MPI, OpenMP, OpenACC, TBB, PGAS, GPI, StarPU, OmpSs, CUDA, ROCm, Sycl, HPX, Kokkos, Legion, RAJA, ...

“MPI + X” likely dominates

- No one-fits-all parallel programming model
  - Example: Upcoming diversity of GPUs

Critical short-coming: Most programming models lack data orchestration capabilities

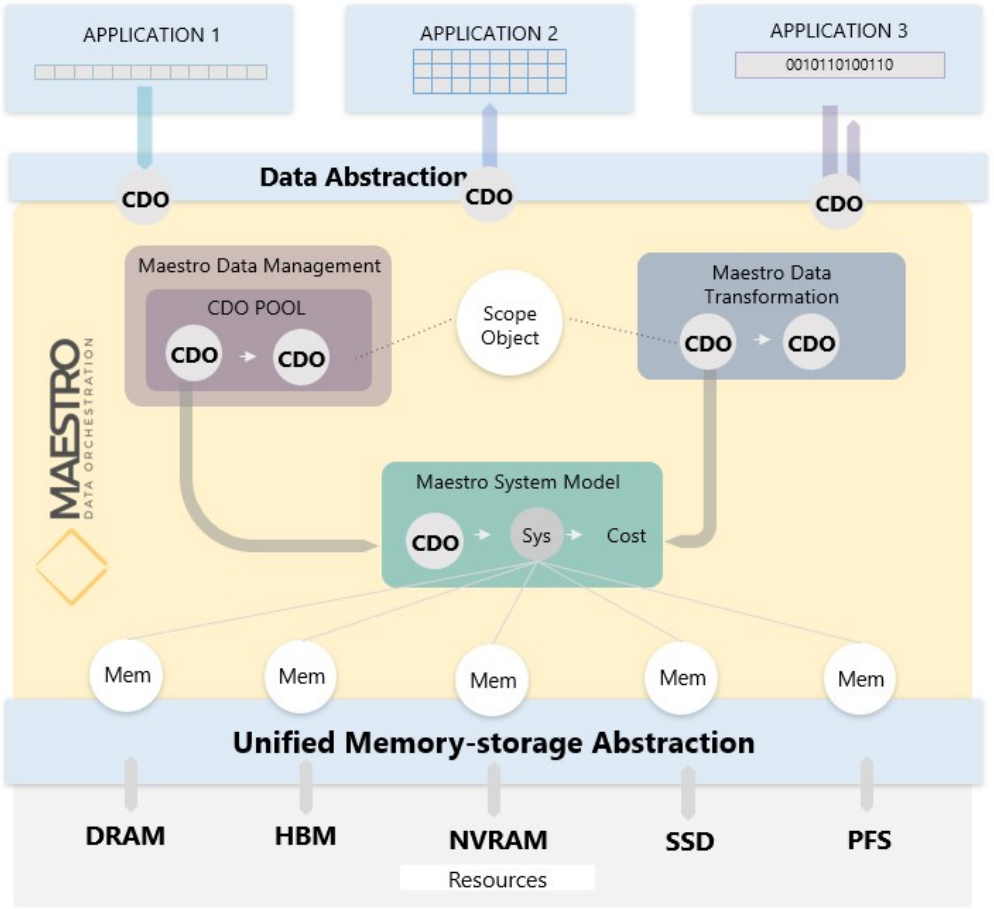
Data Orchestration: Maestro's Approach

Solution concept

- Systems software middleware that intelligently manages data placement and movement
- Object-based approach to encapsulate data with application and Maestro related metadata
- Data movement decision based on workflow annotations and real-time I/O monitoring

Implementation

- Data pool managed by middleware
- Give/take object semantics



Modernising Application Design

Programming means

[Schulthess, 2015]

- Specifying computation
- Managing computer resources

Opportunity for separation of concern

- Front-end: Computation specified by domain scientist using high-level languages
- Back-end: Management of computer resources by HPC experts

Other benefits

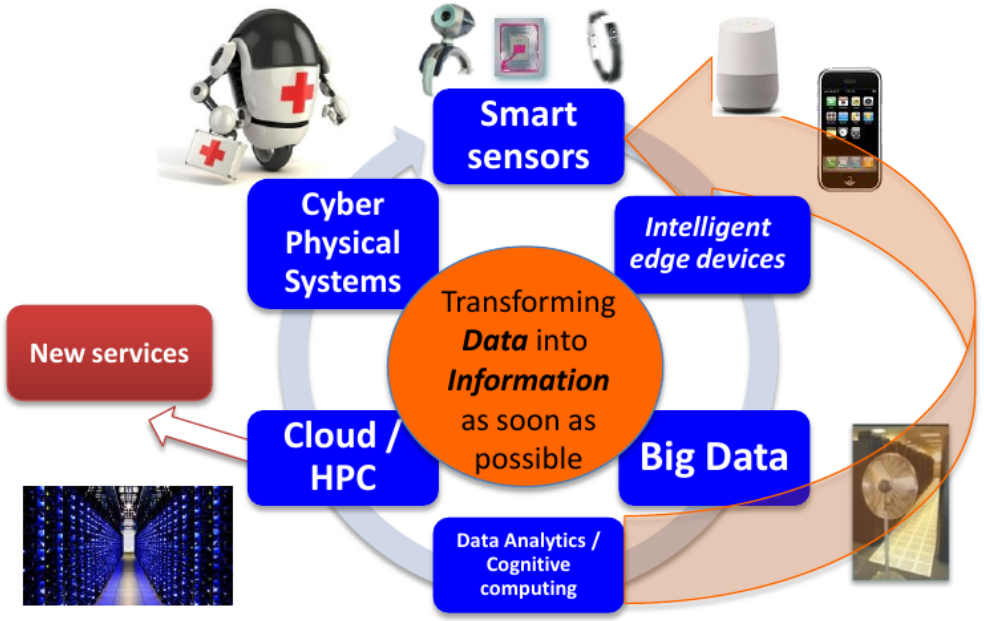
- Descriptive frontend could link to different back-ends → performance portability
- Efforts for implementing back-ends could become community effort
  - Allow domain scientists compete on frontend

FUTURE HPC INFRASTRUCTURES

JÜLICH
Forschungszentrum

New HPC Usage Models Emerging

ENABLING EDGE INTELLIGENCE
C²PS: COGNITIVE (CYBERNETIC* AND PHYSICAL) SYSTEMS



Enabling Intelligent data processing at the edge:
Fog computing
Edge computing
Stream analytics
Fast data…

True collaboration between edge devices and the cloud ensuring:
- **Data security / Privacy**
- **Lower bandwidth**
- **Better use of cloud**

- ***Secure exchanges* between the edge devices and the cloud**
- ***With human in the loop: Centaur era***

[HiPEAC Vision, 2017]

**** As defined by Norbert Wiener: how humans, animals and machines control and communicate with each other.***

Need for designing e-infrastructures including supercomputers

- Supercomputers cannot be designed as silos

JÜLICH Forschungszentrum

Opening HPC Infrastructures: Expected Developments

Federated user management

- Few technical challenges
- Major organisational challenges

Setup clusters for deploying Cloud-type services with path to HPC world

- But: HPC will remain a protected region

Option to deploy services in HPC world

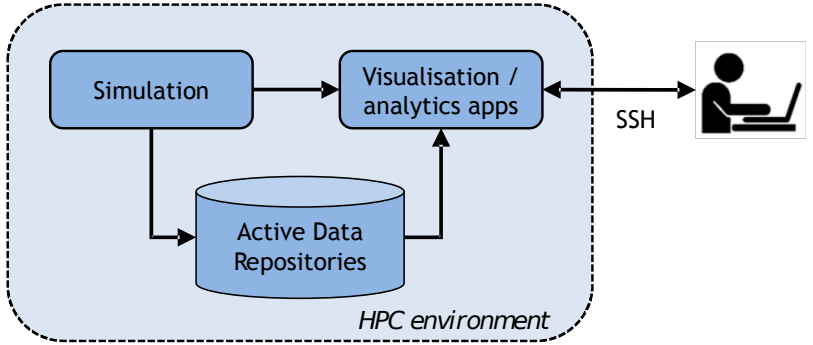
- Examples: databases, workflow schedulers

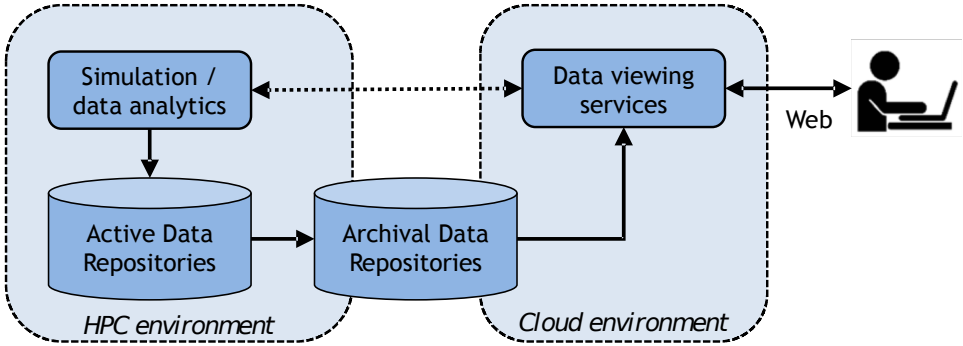
Improve on interactivity

- Support of interactive frameworks like Jupyter notebooks

Allow for new allocation models

- Depends largely on funding agencies





SUMMARY AND CONCLUSIONS

Summary and Conclusions

Paths to exascale architectures established

- Mainly based on compute accelerators
- Use of high-bandwidth memories mandatory for getting to >0.1 Ebyte/s
- First systems planned to become operational in 2021

Major challenges for users

- Cope with heterogeneity and diversity
  - Different processor architectures (x86, Arm)
  - Variety of GPUs
  - Deeper memory hierarchies
- Need for increased efforts in code modernisation aiming for split of concerns between domain scientists and HPC experts

Supercomputers becoming part of wider e-infrastructures

- New approaches to managing boundaries to HPC environment