Supercomputer Architectures: Current State and Future Trends

Dirk Pleiter (KTH)

AQTIVATE Kick-off, September 2023

September 2023

Funded by the European Union

AQTIVATE

The *AQTIVATE* project receives funding from the European Union's HORIZON MSCA Doctoral Networks programme, under Grant Agreement No. 101072344

Overview

HPC Systems: Current Challenges

Current (and near-future) HPC Technologies and Architectures

Towards New HPC Technologies and Architectures

Future HPC Infrastructures

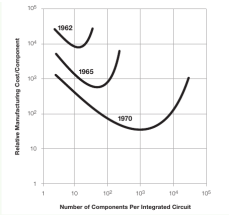
Summary and Conclusions

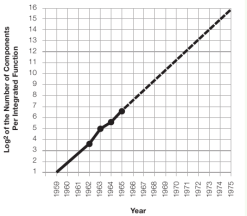
Content

CMOS: “Moore’s Law”

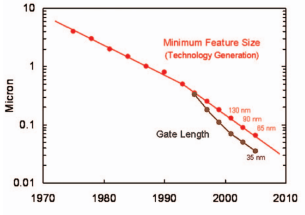
Observation [Moore, 1965+1975]

- Time evolution of optimal manufacturing costs for integrated circuits results in exponential increase of number of components per circuit
- For a longer time period transistor count doubled every 24 month





[Moore, 1965]

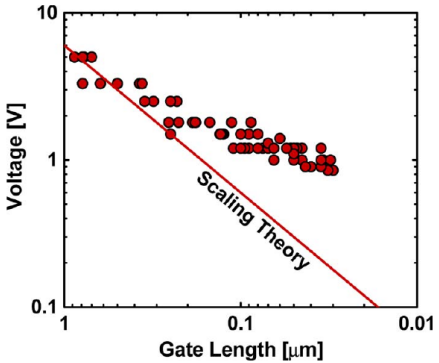


[Bohr, 2007]

CMOS: Dennard Scaling

- **Dennard scaling** allowed to change the following parameters at constant power:
  - Increase of transistor density (Moore's law)
  - Increase clock frequency
  - Reduce supply voltage
- Only remaining option to improve performance: Increase transistor density

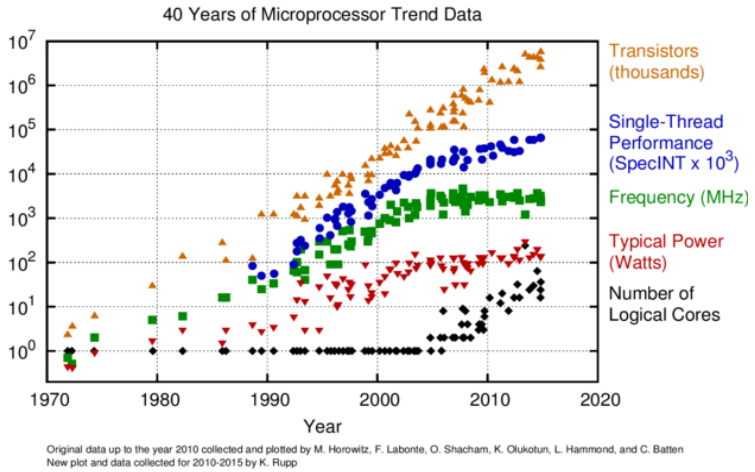
[L. Chang et al., 2010]



☞ **Trend towards increasing parallelism**

CPU Products Trends

[Karl Rupp, 2015]



Digression: Multi-Level Parallelism

A Hardware Perspective

- **Instruction-level parallelism**: Multiple, independent instructions being executed concurrently in superscalar processing cores
- **SIMD parallelism**: Certain Single Instructions can perform the same operations on Multiple Data
- **Device-level parallelism**: CPUs with multiple cores, GPUs with multiple streaming multi-processors
- **Node-level parallelism**: Multiple CPUs and GPUs per node
- **System-level parallelism**: Multiple compute nodes

End of CMOS?

International Roadmap for Devices and Systems (2022 Edition): [IRDS, 2022]

2022 IRDS ORTC							
YEAR OF PRODUCTION	*2021*	*2022*	*2025*	*2028*	*2031*	*2034*	*2037*
Logic device technology naming note definition [1a]	G51M29	G48M24	G45M20	G42M16	G40M16T2	G38M16T4	G38M16T6
Logic industry "Node Range" Labeling (nm) [2]	"5"	"3"	"2"	"1.5"	"1.0-eq"	"0.7nm-eq"	"0.5nm-eq"
Fine-pitch 3D integration scheme		Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Platform device for logic [1b]	FinFET	FinFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
LOGIC CELL AND FUNCTIONAL FABRIC TARGETS							
Digital block area scaling	1.00	1.00	0.74	0.55	0.26	0.13	0.08
LOGIC DEVICE GROUND RULES							
MPU/SoC M0 1/2 Pitch (nm) [3]	15	12	10	8	8	8	8
Gate length (nm) [4]	17	16	14	12	12	12	12
Lateral GAA (nanosheet) Minimum Thickness (nm)		1	3	3	4	4	4
Number of stacked tiers [5]		1	1	1	2	4	6
Number of stacked nanosheets in one device [5]		1	3	3	4	4	4
LOGIC DEVICE Electrical							
Vdd (V) [6]	0.75	0.70	0.65	0.65	0.60	0.60	0.60

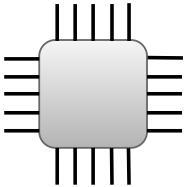
Answer: The industry believes in further CMOS scaling thanks to 3-d stacking

Rent's Rule

- **Rent's rule**:

$$T = k\, G^p$$

  - G ... Number of logic elements (gates)
  - T ... Number of edge connections (terminals)
  - k ... Rent's coefficient
  - p ... Rent's exponent



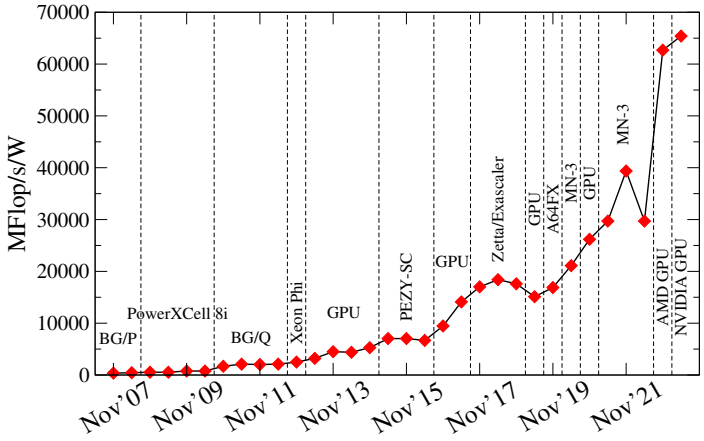
- Problem: typically $p \ll 1$ ☞ Difficult to balance communication and compute
- Selected empirical results for Rent's rule (data from Bakoglu, 1990): [Lanzerotti et al., 2005]

Design type	Rent coefficient	Rent exponent
SRAM	6	0.12
Gate arrays	1.9	0.50
Chip and module	1.4	0.63
Board and system	82	0.25

Power Efficiency: The Green500 List

- Green500: Rank supercomputers according to power efficiency
  - Metric = floating-point performance vs. power consumption
  - Supercomputer = system listed in TOP500
  - Performance = HPL performance (like for TOP500)
- Current number #1 (Jun'23): 65.4 GFlop/s/W (or 15.3 pJ/Flop)
- Exascale goal: keep below 20 MW (or 20 pJ/Flop)
- Exascale system Frontier: 52.6 GFlop/s/W (or 19pJ/Flop)
- Criticism
  - The High-Performance LINPACK (HPL) benchmark load is not representative
  - Green500 does not cover full system

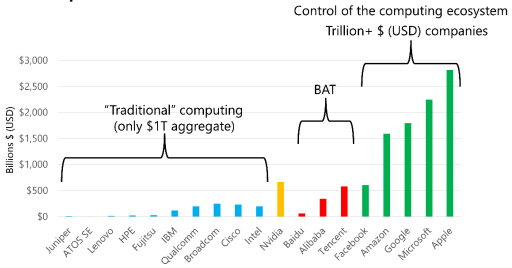
The Green500 List (cont.)



HPC Market Size

- Non-recurring engineering (NRE) costs for developing new technologies and architectures are huge
- NRE costs funding challenge due to th HPC market being small
  - Market for HPC technologies is small
  - Need for significant public funding for new HPC architectures and technologies

Computing company market capitalization:



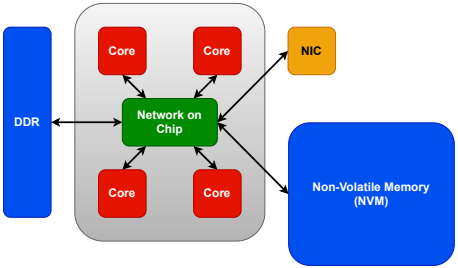
[Reed et al., 2022]

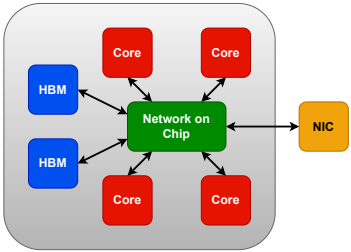
Content

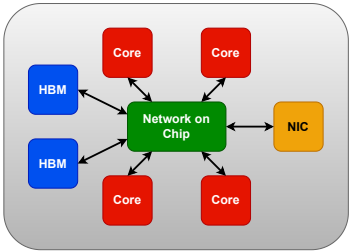
Abstract Machine Model

[Ang et al., 2014]

Single Socket, Homogeneous Cores, DDR+NVM or HBM

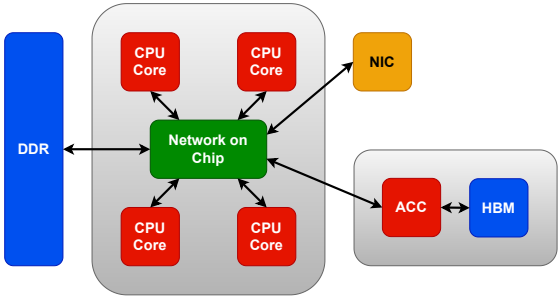


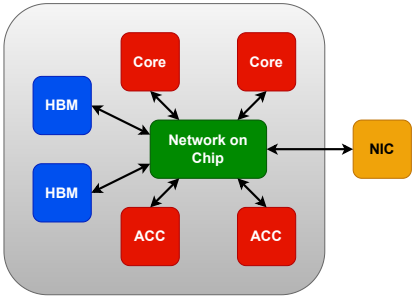




Abstract Machine Model

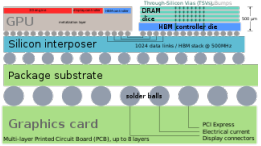
Discrete versus Integrated Accelerator





Memory Technologies

- HBM



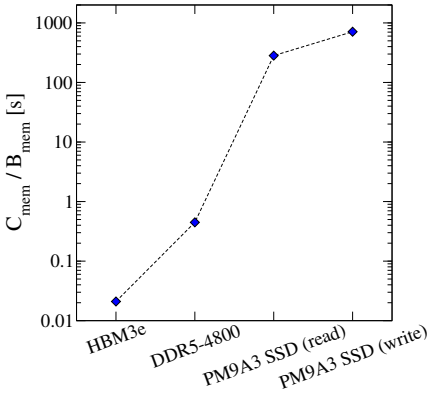
- DDR

- SSD

- Significant differences in terms of

$$\Delta t = \frac{C_{\text{mem}}}{B_{\text{mem}}}$$

Data for selected currently available Samsung products:

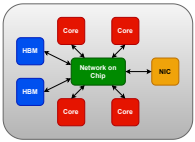


CPU Technologies

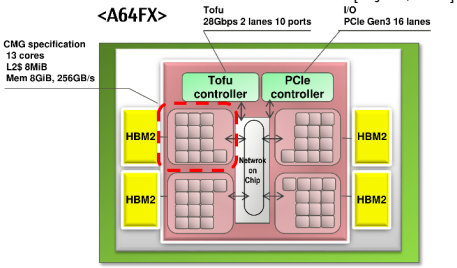
Feature	Xeon	EPYC	A64FX	Grace	POWER9
Model	Max 9480	7763	–	–	–
ISA	x86	x86	Armv8	Armv8	Power v3
SIMD ISA	AVX512	AVX2	SVE (512)	SVE2 (128)	VMX
Number of cores	56	64	48	72	22
Base clock frequency [GHz]	1.9	2.45	2.2	3*	3.07
Perf. B_{fp} [10^{12} FP64/s]	3.4	2.5	3.4	3.1	0.5
Memory technology	HBM2e + DDR5	DDR4	HBM2	LPDDR5X	DDR4
Bandwidth B_{mem} [GByte/s]	> 1000	204.8	1000	~ 500	170
TDP [W]	300*	280	210*	250	300
B_{fp}/TDP [GFlop/s/W]	11*	9	16*	12	2

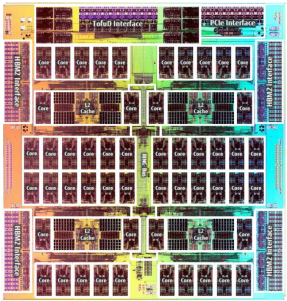
Spotlight: Fujitsu A64FX

[M. Sato, 2019]



[Fujitsu, 2018]





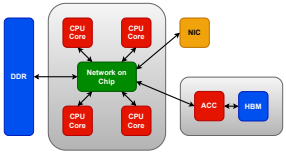
Accelerator Technologies

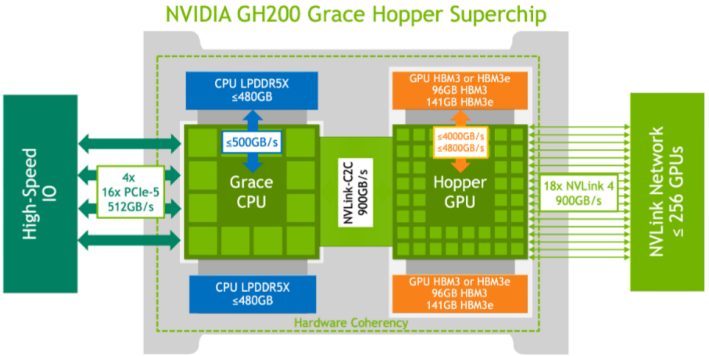
Feature	NVIDIA			AMD	Intel
Model	V100	A100	H100	MI250x	X^e-HPC
Base clock frequency [GHz]	1.3	1.1	?	1.7*	0.9
Performance B_{fp} [10^{12} FP64/s]	6.7	7.6	30	47.9	29.5
Memory technology	HBM2	HBM2e	HBM3	HBM2e	HBM2e
Bandwidth B_{mem} [GByte/s]	900	1555	4000	3277	3277
TDP [W]	300	400	700	560	600
B_{fp}/TDP [GFlop/s/W]	22	19	43	85	20

Caveat: Clocks (and B_{fp}) can be highly variable making comparisons difficult

Spotlight: NVIDIA Grace-Hopper

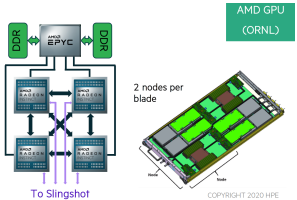
[NVIDIA, 2023]

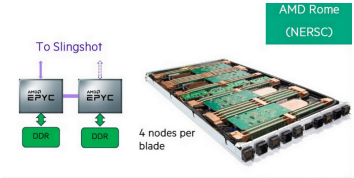


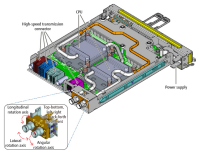


Compute Node Designs

- Ultra-fat nodes
  - Chassis height $\geq 2\,\mathrm{U}$
  - $\gtrsim 10$ CPUs and/or ACCs
- Fat nodes
  - Chassis height $= 1 - 2\,\mathrm{U}$
  - 1-4 CPUs and 4-6 ACCs
- Thin nodes
  - Chassis height $= 1\,\mathrm{U}$
  - 1-2 CPUs
- Ultra-thin nodes
  - Special form-factor
  - 1 CPU



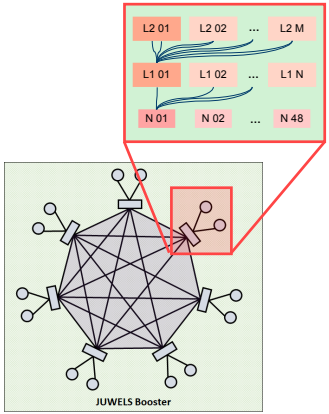




Network Technologies and Architectures

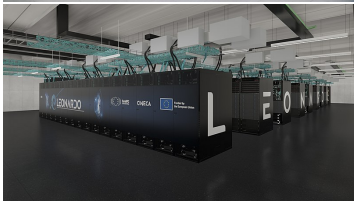
- Popularity of interconnect technologies based on Top100 (June 2023, system count)
  - 62%: Infiniband
  - 16%: Slingshot
  - 6%: TOFU
  - 6%: Aries
  - 5%: Omni-Path
  - 5%: Other (incl. BXI, Ethernet)
- Topologies used for Top5 (June 2023)
  - Dragonfly (e.g. Frontier #1, LUMI #3)
  - Torus (e.g. Fugaku #2)
  - Dragonfly+ (e.g. Leonardo #4)
  - Fat-tree (e.g. Summit #5)

Example: JUWELS Booster's dragonfly+ network



EuroHPC Pre-Exascale Systems

- LUMI at CSC (Finland)
  - AMD EPYC CPUs
  - AMD Instinct MI250x GPUs
  - Slingshot interconnect with dragonfly topology
- Leonardo at CINECA (BSC)
  - Intel Xeon CPUs
  - NVIDIA A100 GPUs
  - Infiniband HDR100 with dragonfly+ topology
- Mare Nostrum 5 at BSC (Spain)
  - Intel Xeon CPUs (and others)
  - NVIDIA H100 GPUs (and others)
  - Infiniband with dragonfly+ topology



Content

Rebooting Computing

IEEE's Rebooting Computing Initiative considers 4 levels of change: [Conte et al., 2017]

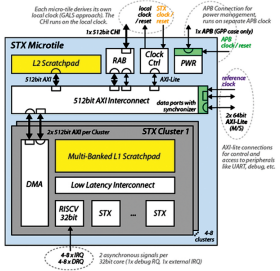
- **Level 1:** Introducing new transistors
  - But: There are no clear candidates beyond CMOS
- **Level 2:** Other hidden hardware changes
  - Examples: New packaging techniques (3-dimensional stacking, interposers+chiplets)
- **Level 3:** Architectural changes that are exposed to the programmer but do not require new algorithms ← **focus in the following**
  - Examples: Reconfigurable computing devices like FPGAs
- **Level 4:** Fundamental changes of the computing stack requiring new algorithms
  - Examples: Quantum or neuromorphic computers

Domain-Specific Accelerators

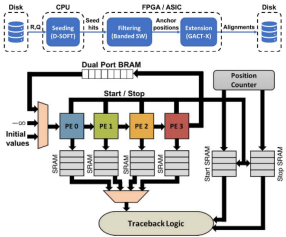
[Dally et al., 2020]

- Opportunities to benefit from specialisation through
  - Specialised instructions that take domain-specific data structures as input
  - Increased parallelism while optimising for data locality
  - Optimised memory hierarchy with local memories
  - Reduced overheads due to simplified instruction processing
- Current examples:
  - STX stencil calculation accelerator (Fraunhofer, DE) [EPI]
  - Darwin-WGA genomics accelerator (Stanford, US) [Turakhia et al., 2019]
  - DOJO (Tesla, US) [Talpes et al., 2023]

STX:



Darwin-WGA:



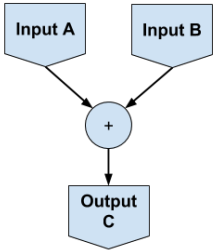
Data-Flow Architectures

[J. Dennis, 1980]

- In data-flow architectures data becomes processed when arriving at actors
  - There may be no instructions
  - Actors can be moved close to the data
- Particular suitable for reconfigurable hardware (e.g. FPGAs)
- Example numerical task: Addition of 2 vectors

$$\vec{c} \leftarrow \vec{a} + \vec{b}$$

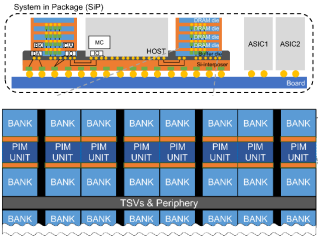
```
1 #define N 128
2 float a[N], b[N], c[N];
3
4 int main() {
5   for (int i = 0; i < N; i++)
6     c[i] = a[i] + b[i];
7   return 0;
8 }
```



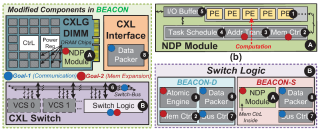
Near- and In-Memory Computing

- Near-memory computing = processing units placed closer to the memory
  - Challenge: Integration with CPU
    - Protocols like CXL will make that easier
  - Challenge: Identification of amenable algorithms
    - Need, e.g., localised operands
- In-memory computing = augmented memory devices supporting computational primitives
  - Challenge: Computational errors
  - Use for scientific computing in the near future less likely

PIM-HBM: [Lee et al., 2021]



BEACON: [Huangfu et al., 2022]



Content

Drivers: Distributed Research Infrastructures

- Changing needs of existing user communities and new needs of emerging new science and engineering domains:
  - Support of collaborative research
  - Wider access to HPC enabled by higher-level services
  - Workflows extending HPC data centre ("computing continuum") and use of geographically dislocated services
  - Ability to deploy domain-specific platform services
- Example: The European brain research community has started to operate EBRAINS

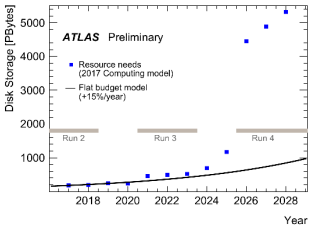
- Selected use cases from the brain research community
  - Simulations at different scales including large-scale, coupled simulations
  - Data processing and machine learning in parts involving extreme-scale data sets
  - Interactive computing including visualisation of extreme-scale data sets
  - Deployment of a science gateway ("collaboratory")

Drivers: Science Instruments

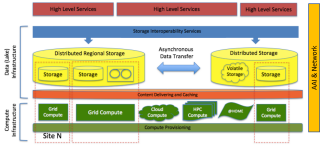
- Large-scale science instruments increasingly have a need for HPC resources
  - High-energy physics experiments
  - Light sources
  - Radio-astronomy
- New initiative of DoE: Integrated Research Infrastructure (IRI) [Miller et al., 2023]
  - Initial identification of “science patterns”
    - Time-sensitive: Time-critical workflows related, e.g., to timely decision making, experiment steering
    - Data integration: Analysis of data from multiple sources
    - Long-term campaigns: Need for sustained access to resources at scale

[Albrecht et al., 2019]

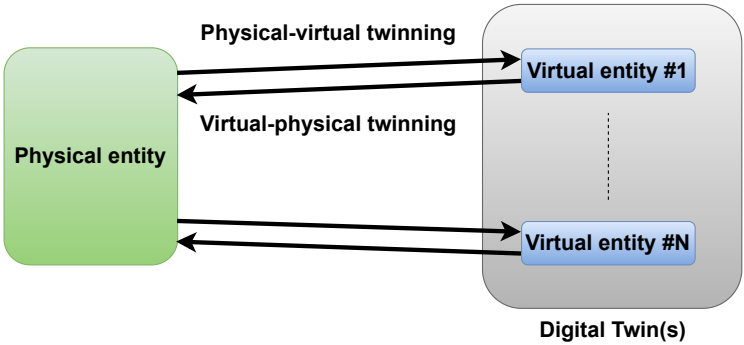
ATLAS disk storage estimates:



ESCAPE Data Lake:



Digression: Digital Twins



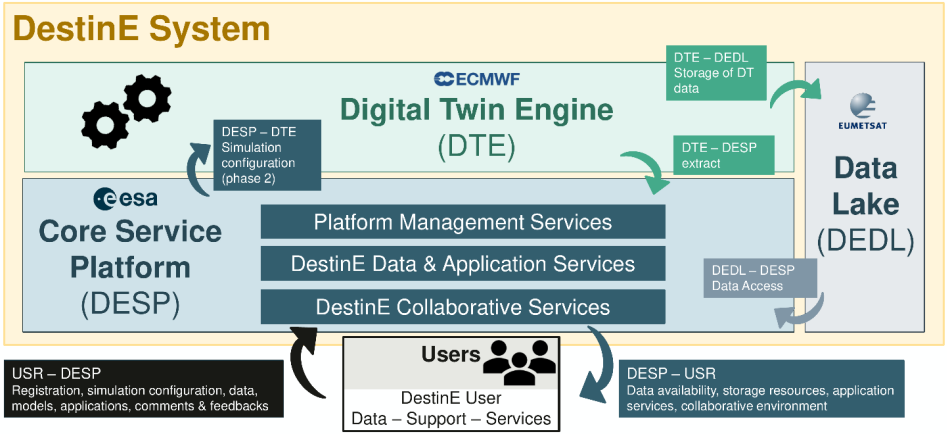
- Virtual entities = models
- Models may require HPC for simulation or ML training

Drivers: Destination Earth

- **Mission**: Mission: Destination Earth aims to develop - on a global scale - a highly accurate digital model of the Earth to monitor and predict the interaction between natural phenomena and human activities
- **Approach**: Distributed infrastructure comprising digital twins implemented, in particular, on HPC systems
- **Key components**
  - Digital twins and a "Digital Twin Engine"
  - Data lake
  - Service platform
- Development of **use cases** as end-to-end solutions
  - Coastal area flooding: Improved forecasting and assessment of climate adaptation measures [ECMWF/Deltares]
  - Air quality: Forecasts based on extreme weather events forecasts for national and regional environmental agencies [ECMWF/FZJ]

Destination Earth System

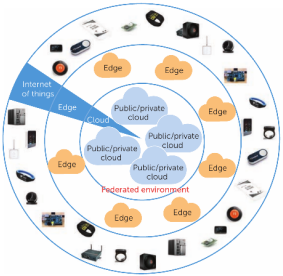
[ESA, 2022]



Computing Continuum

- Emerging paradigm referring to a continuum of resources available from the (cloud) data centre to the edge (and beyond)
  - Edge devices = devices at the edge of the network
  - Motivations for edge computing:
    - Faster response time between device and application server
    - Need for IoT gateways
    - Facilitate data reduction at the edge of the network
- An increasing number of usage scenarios have been identified where HPC resources need to be integrated in a computing continuum
  - Connect observational data from sensor devices to large-scale simulations
  - Facilitate continues training of large-scale models

One of many views on a computing continuum:



[Villari et al, 2016]

Computing Continuum: Research Challenge

- Scheduling within a highly distributed and heterogeneous environment
- Workflow systems
- Software deployment in heterogeneous environments
- Server-less computing
- Data management, protection of sensitive data
- Security and privacy
- Trust

Content

Summary and Conclusions

- Current challenges in HPC will continue to be relevant in the future
  - Technology roadmaps indicate further vast increase of parallelism
  - Data transport limitations might become even more critical
- The HPC architectures and technologies are becoming increasingly diverse
  - More accelerators and non-conventional technologies will become available
  - This diversity cannot be hidden to the users
- Trend towards workflows and use cases extending beyond the data centre
  - There are multiple strong science (and industrial) drivers
  - Transition from HPC as siloed systems to HPC infrastructures