Drivers for a paradigm change in HPC deployment

EXDCI-2 – Technical meeting
Barcelona, Dec. 2nd 2019

Gabriel Antoniu, INRIA
Marc Duranton, CEA
Jens Krueger, Fraunhofer
Erwin Laure, KTH
Michael Malms, ETP4HPC
Maria S. Perez, UPM

The EXDCI-2 project has received funding from the European Unions Horizon 2020 research and innovation programmed under grant agreement no. 800957.

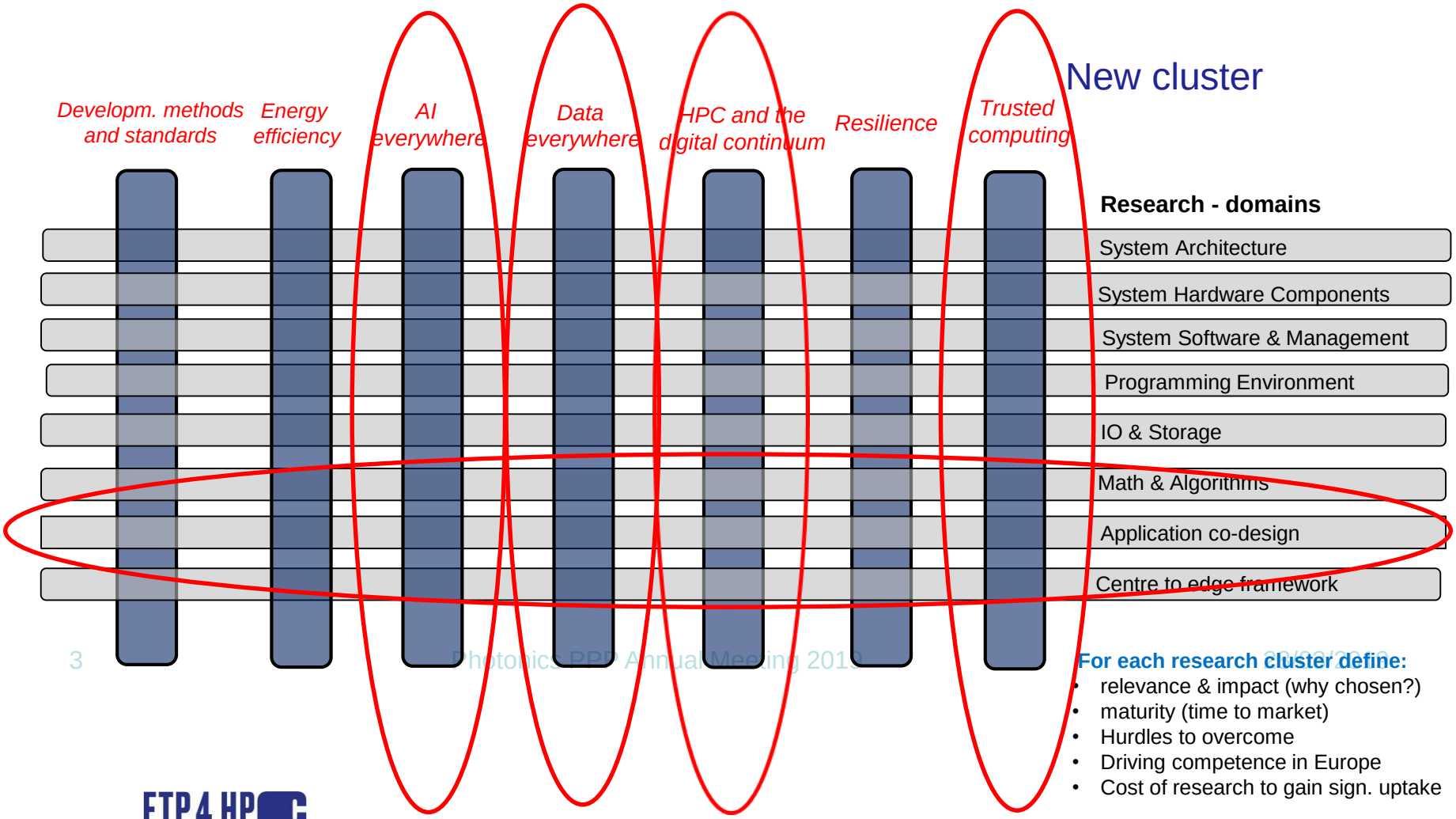
The starting point (M. Malms) -1

- A few starting remarks:
  - Still today, all the focus in HPC R&D is on the "race to exascale – who is first"

BUT:

  - What is the next big thing afterwards, what do we need to get prepared for?
  - Which are the trends for HPC technology, skills, expertise deployment ?
  - What are the most important disciplines and their impacts ?
    - AI, Big Data, IOT, Cyber Security, Upstream Technologies, 5G
  - How can we in Europe take an active role in the upcoming change process?
    - ..to the benefit of technology and application providers,
    - ..growing the expertise and skills level in the relevant domains

Research clusters and research domains for SRA-4 (M. Malms) -2



For each research cluster define:

- relevance & impact (why chosen?)
- maturity (time to market)
- Hurdles to overcome
- Driving competence in Europe
- Cost of research to gain sign. uptake

ETP4HPC

exdci European Extreme Data & Computing Initiative

Application co-design (Erwin Laure) - 1

- Application priorities and influences on HPC paradigm change
  - E.g. fundamental sciences; climate, weather, and earch sciences; life sciences & health; energy; engineering & manufacturing; chemistry & materials sciences; ect.
  - Increasing uptake from industry
    - Although at lower level than academia
  - AI plays an increasing role in established and new HPC flieds
    - E.g. humanities, social sciences, finance, etc.
  - Increased focus on data
    - Also increased need for cyber security

Application co-design (Erwin Laure) - 2

- Main Challenges:
  - Porting, adapting, optimizing for new architectures
    - Requires co-design, stable programming environments
  - New opportunities through convergence of HPC, HPDA, and AI
    - More complex workflows, data handling
  - Memory bandwidth and communication latency becomes more important than FLOPS
  - Long term maintenance of applications and codes
    - Continued funding and stable (standardized) environments

Application co-design (Erwin Laure) - 3

Intersections – the “heat map”

	Fundamental Sciences	Climate/Weather/ Earth Sciences	Life Sciences and Medicine	Engineering & Manufacturing	Chemistry & Materials Science	Industrial Applications
Development methods and standards						
Energy efficiency						
AI everywhere						
Data everywhere						
HPC & Digital Continuum						
Resilience						

AI everywhere – research cluster (María S. Pérez) - 1

- Main influences on HPC in this cluster:
  - Deep Learning requires HPC approaches
  - Convergence of HPC, Big Data and AI:
    - New applications combining simulations and data analytics
    - HPC for AI: HPC supporting the efficient execution of AI approaches
      - Scalable and high-performance AI solutions
      - Learning across the digital continuum: Distributed AI, Edge Analytics
    - AI for HPC: AI improving and enabling new HPC solutions
      - Neuromorphic architectures
      - AI helping to write software and build hardware
  - Applications of AI and HPC not only to scientific applications, but also to industrial applications

AI everywhere – research cluster (María S. Pérez) - 2

- Relevance and impact - why chosen:
  - AI is one of the pillars of the 4th industrial revolution
  - Importance of AI in the computing continuum
  - Importance of AI in industrial applications (manufacturing, automotive, finance, communication, ..)
  - New markets
  - SW writing SW
  - Building of new hardware

AI everywhere – research cluster (María S. Pérez) - 3

- Hurdles to overcome:
  - Scalability of AI systems and algorithms:
    - Mainly linked to Big Data and demanding AI algorithms
  - Interoperability of tools and software stack
  - Ethical aspects (1): Human-centric design:
    - Lawful
    - Ethical
    - Robust
  - Liability of AI systems:
    - Complex decision-making processes, involving different parties
  - Explainable AI:
    - Trust and acceptance of AI applications

(1) High-Level Expert Group on Artificial Intelligence, Ethics guidelines for Trustworthy AI, European Commission, 2019

Data everywhere – research cluster (Gabriel Antoniu) - 1

- Main influences on HPC in this cluster:
  - From model-driven and simulation-driven science to data science
    - We are now in an era of "Data Centric" computing
    - The data sphere: 175 ZB by 2025 (Source: IDC)
      - Sensors and IoT becoming a major part of science and innovation
      - Societal trends such as driverless cars and large experiments such as SKA
      - Need to process data at the "edge" as well as at the data center
  - HPC will increasingly start to play a role here
    - Workflows across the whole digital continuum (edge - "fog" cloud data centers – HPC systems) will exhibit specific HPC needs at each level
      - Embedded HPC (fog-level)
      - HPC clouds
      - Supercomputers

Data everywhere – research cluster (Gabriel Antoniu) - 2

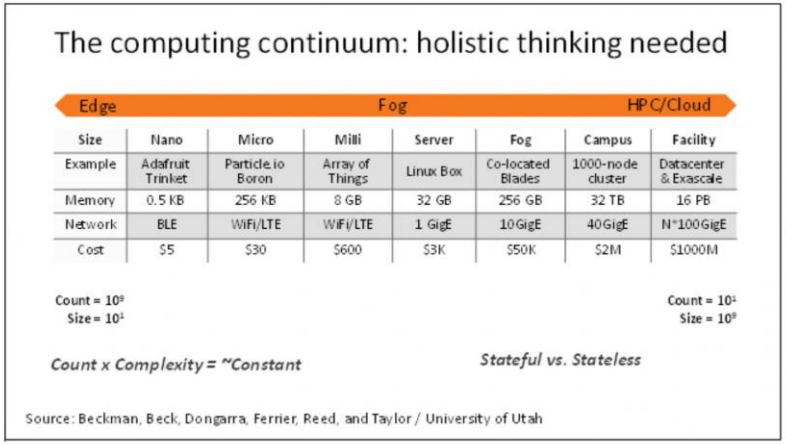
- Relevance and impact - why chosen:
  - Scientific simulations and workflows will also generate and work with enormous volumes of data
  - Data needed by scientific workflows is now highly distributed geographically
    - Edge + Distributed "Fog" nodes + Clouds + HPC Centers
    - Accumulated historical data + real-time data
  - Various data processing techniques needed
    - HPC data processing techniques: in-transit/in situ processing
    - Big Data Processing Techniques (batch + real time/streaming)
    - Trend: support decentralization of (AI-based) analytics towards the edge
    - Cloud based data access & associated APIs (S3, Swift, etc)

Data everywhere – research cluster (Gabriel Antoniu) - 3

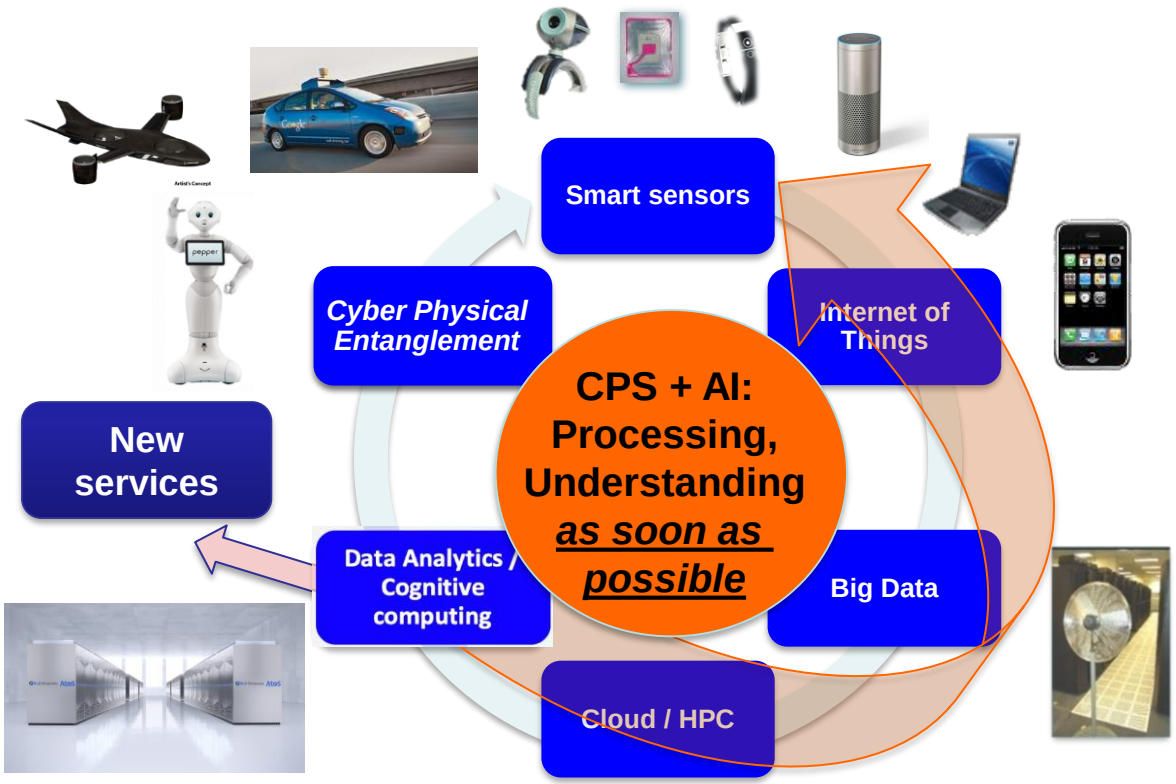
- Hurdles to overcome:
  - Requirement for high bandwidth between the various infrastructure entities (Edge, Fog, Cloud/HPC system)
    - Still lacking – inspite of in-transit and in-situ processing in HPC systems
    - 5G networks may play a role here
    - However data generation continues to outstrip the evolution of N/W bandwidths
  - Data logistics and data life-cycle management
    - When to move data, how long to retain data, etc
    - Collecting provenance metadata becomes important
  - Making data infrastructures lot more resilient/performant at scale
    - Usage of new tech.s like NVRAM still not very clear
  - Data consistency across these infras still a big problem
  - Data federation: lack of unified APIs, need to cope with sensitive data

HPC and the digital continuum – research cluster (Marc Duranton) - 1

- Main influences on HPC in this cluster:
  - Convergence of HPC, Big Data and AI
    - Current AI and Auto-ML requires HPC capabilities for learning, but it is nourish by data provided "at the edge"
    - "Digital twin" or Cyber Physical Systems will need that HPC will be "in the loop" with real world, coping with its (timing and security) constraints, stream processing, security and safety requirements
    - Edge data will drive/tune numerical simulations
    - HPC in the box or Embedded HPC



COMPUTING NOW FORMS A CONTINUUM



Enabling Intelligent data processing where it is required

- ***Fog computing***
- ***Edge computing***
- ***Stream analytics***
- ***Fast data…***

True collaboration between edge devices and the HPC/cloud ⇨ creating a

continuum

ENABLING EDGE INTELLIGENCE

Transforming **data** into ***information*** ***as early as possible***

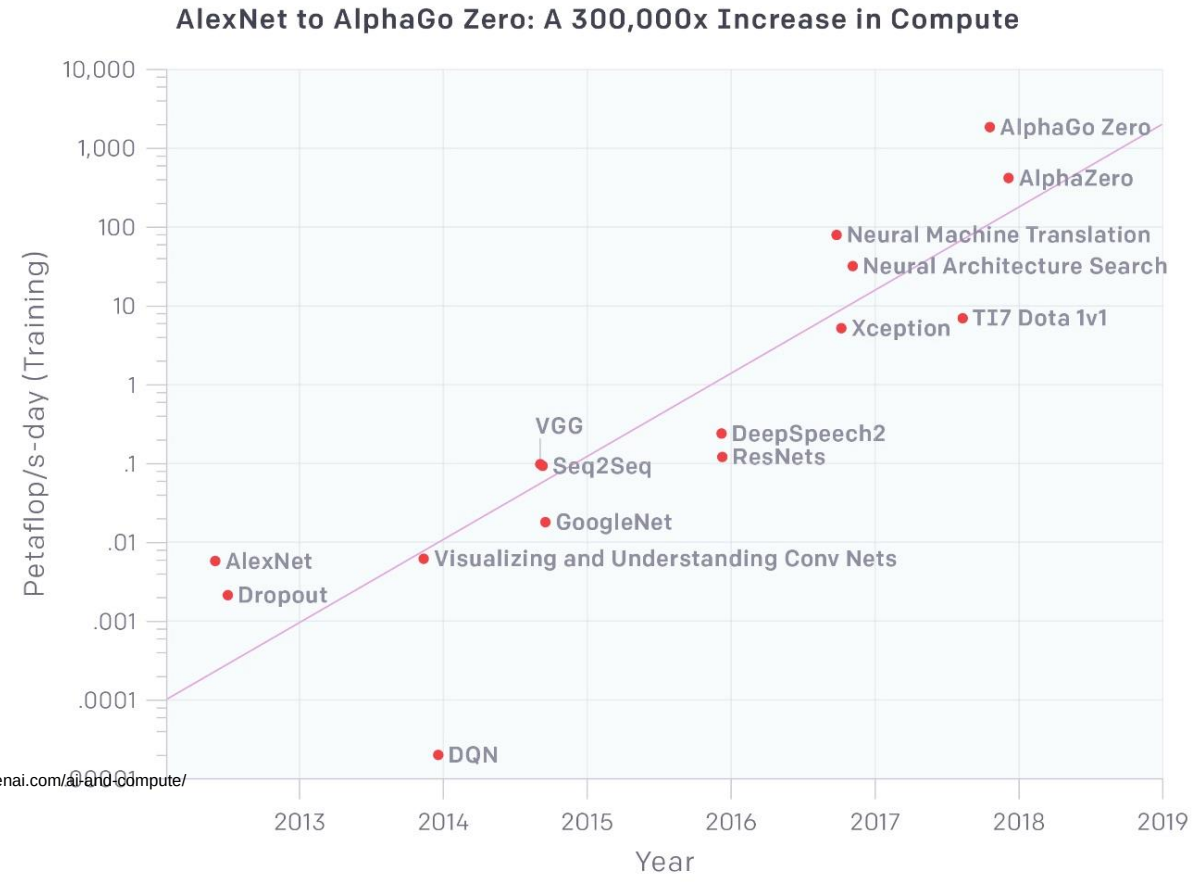
HPC and the digital continuum – research cluster (Marc Duranton) - 2

- Relevance and impact - why chosen:
  - More and more applications are distributed
  - Global energy efficiency requires reducing the amount of data transferred
  - Importance of AI in the computing continuum
    - Deep Learning requires more and more Pflops
    - Auto-ML to compute the meta parameters of Deep Learning systems
  - Importance of industrial applications (manufacturing, automotive, finance, communication, ..) where HPC is more and more "in the loop"
  - HPC capabilities more and more required to improve the efficiency of our digital world

EXPONENTIAL INCREASE OF COMPUTING POWER FOR AI TRAINING

*"Since 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a **3.5 month-doubling time***

(by comparison, Moore's Law had an 18-month doubling period)"*



* https://blog.openai.com/ai-and-compute/

- Hurdles to overcome:
  - Not anymore floating point simulations
    - More data, more memory
    - Diversity of data types
  - Real-time use, stream processing, less "batches"
  - More open to external data, with the challenges of interoperability, security, data protection, interconnectivity, ….
  - Different "silos" for HPC, Big Data and IA applications and software stacks
    - Interoperability of tools and software stack
  - Interdisciplinarity

Trustworthy Computing – research cluster (Jens Krueger) - 1

- Main influences on HPC in this cluster:
  - Security is becoming an important factor
    - HPC data centres usage model are changing; data science, deep learning
    - HPC in the continuum
  - Trustworthy computing for the complete HW / SW stack
    - chip design + firmware
    - operating systems
    - runtimes
    - applications design
    - interfaces, etc.
  - establishing processes and accountability within data centres

Trustworthy Computing - research cluster (Jens Krueger) - 2

- Relevance and impact - why chosen:
  - Strategic investment
    - trustworthiness of vendor products
      - handling critical data
    - withstand new threat scenarios
  - GDPR complience
    - handling of personal data
  - HPC technologies in critical environments

Trustworthy Computing - research cluster (Jens Krueger) - 3

- Hurdles to overcome:
  - Establishing security within the mindset of HPC
  - Cross-sectorial / cross-disciplinary effort
  - Security vs. Energy Efficiency vs. Performance

Thank you !

The EXDCI-2 project has received funding from the European Unions Horizon 2020 research and innovation programmed under grant agreement no. 800957.