INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

sighpc computing continuum

FROM THE EDGE TO THE CLOUD

Computing continuum challenges and opportunities

Carlos Jaime BARRIOS HERNANDEZ, PhD, HDR

Oscar Alberto CARRILLO ROZO, PhD.

SCALAC

EL-BONGÓ physics

Co-funded by the European Union

Universidad Industrial de Santander

Super Computación y Cálculo Científico UIS

LIG

Inria

citi lab

CPE LYON

UNIVERSITÉ DE LYON

OSCAR CARRILLO

- Associate Professor at CPE Lyon (University of Lyon), leading the Networks & Telecommunications domain.
- Researcher at CITI Lab (INSA Lyon / INRIA), DynaMid team.
- Founding member of the CATAI collaboration.
- Research: software/model verification, middleware, distributed systems, IoT & smart cities.
- Degrees: PhD in Computer Science (Univ. of Franche-Comté), MSc (Polytech Nice-Sophia), BSc (UIS, Colombia).

CARLOS J. BARRIOS

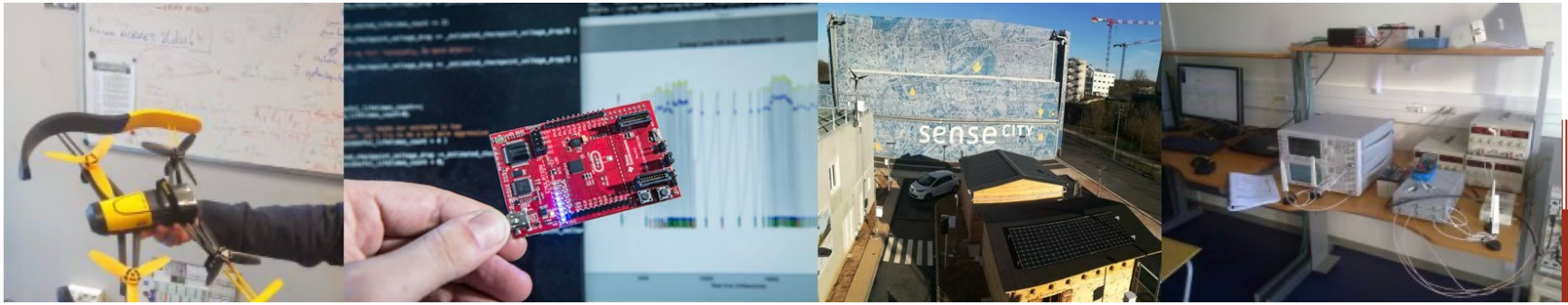
- Director of SC3UIS (High Performance & Scientific Computing Center), UIS – Colombia.
- Full Professor at UIS, School of Informatics and Systems Engineering.
- Chair of SCALAC (Advanced Computing Systems for Latin America & the Caribbean).
- Research: advanced & high-performance computing, scalable architectures, hybrid/heterogeneous systems, computing continuum, performance evaluation.
- Degrees: PhD (Université Nice-Sophia Antipolis), MSc (Université Grenoble-Alpes), BSc (UIS). HDR from INSA Lyon (2025).
- Invited researcher at LIG/INRIA Grenoble and CITI/INRIA Lyon.

LYON, FRANCE

- One of the first capitals of France
- Played a very important role in the trade between the oriental world and Europe
- Founded over an ancient Roman city (ruins still exist)
- 2nd (or 3rd largest city in France)
- The number one bio- and chemical-industrial pole in France
- Home of IN2P3

CITI LAB - ID CARD

- 9 teams in 2025, including 5 Inria teams - ~150 people (30Prof/Ass, 7 Inria DR/CR, ~50 phd students, ~10 inges)
- Common Labs : Nokia Bell, Orange - Chairs : SPIE ICS
- 12 hardware and software platforms (EquipEx FIT/CorteXLab, EquipEx+ TIRREX), 1 spinoff
- ~10.5M€ annual consolidated budget, ~4.5M€ net revenue (20% Europe, 60% France, 20% industrial)

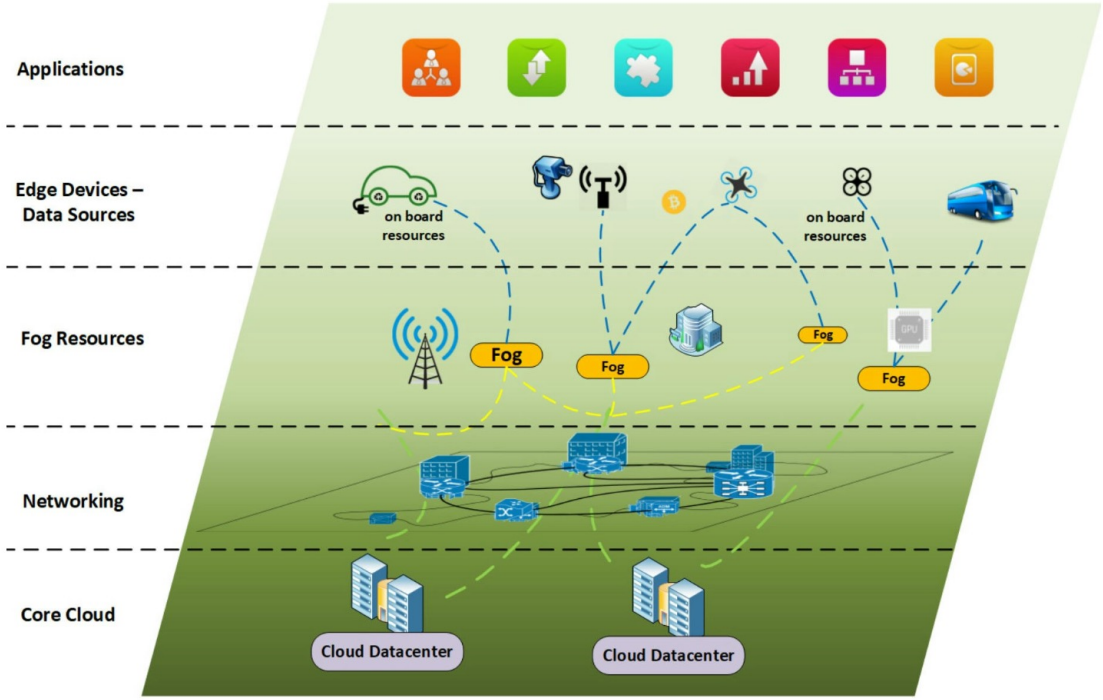


CITI LAB – SCIENTIFIC PROJECT

« Reconnected Society »

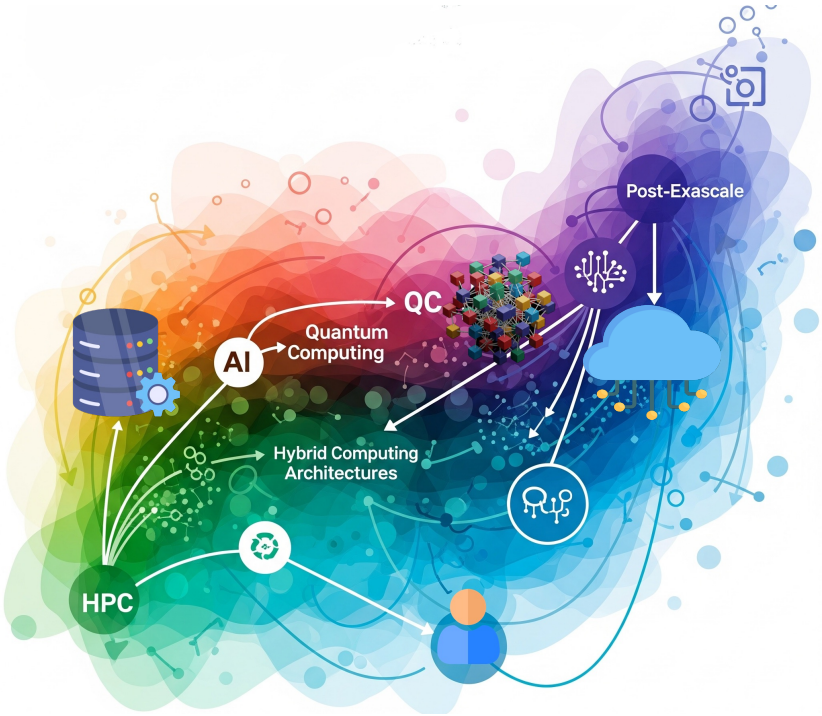
Connected Objects, Connected Society as Digital Mediator Tools for Human and Societal Challenges

THE COMPUTING CONTINUUM



"Distributed computing model allowing end-to-end resource orchestration that considers computation, storage and network for applications executed across the Edge, Fog, and Cloud computing tiers. It ensures communication and open availability between geographically distributed computing resources."

HETEROGENEITY, PERFORMANCE, SUSTAINABILITY AND ACCURACY



EDGE THROUGTH ALL THE COMPUTING

NETWORK CONCERNS

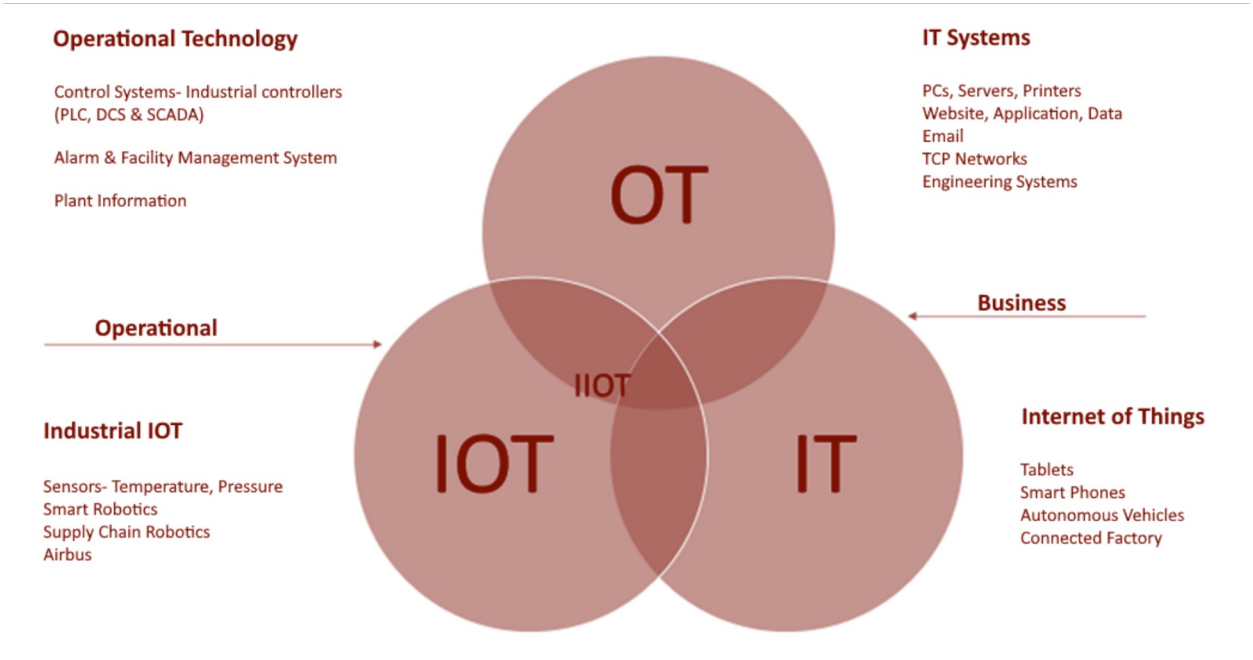
COMPUTER ARCHITECTURE CONCERNS

HPC TO ADVANCED COMPUTING OVER MULTISCALE HPC

MULTISCALE HPC TO COMPUTING CONTINUUM

OPERATION TECHNOLOGY AND DIGITAL TRANSFORMATION

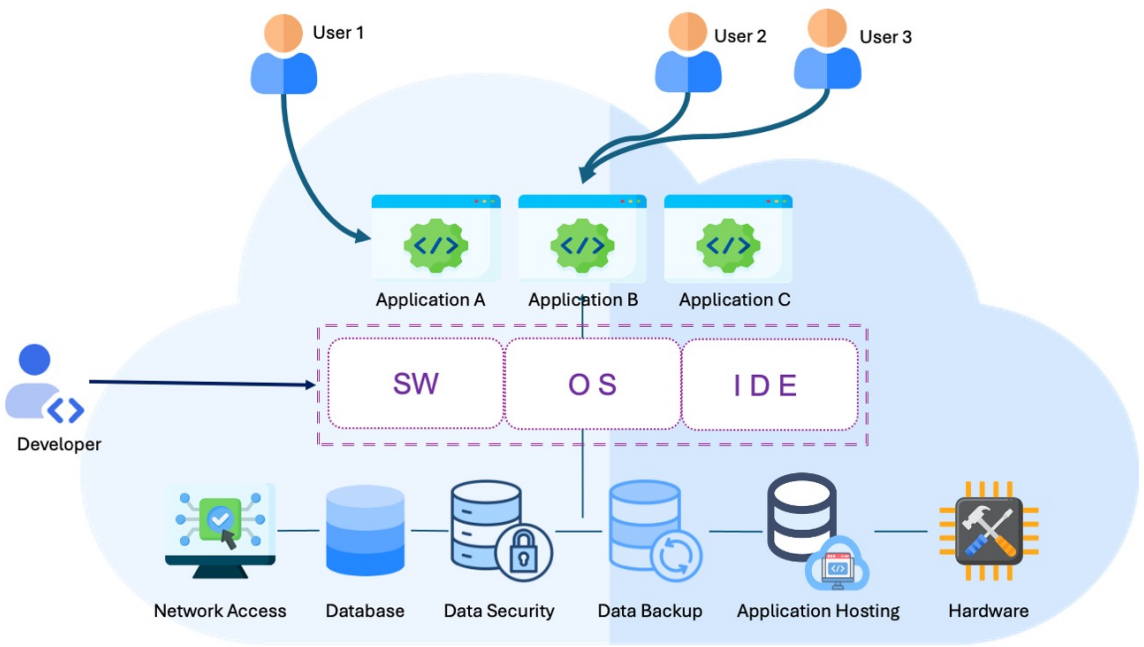
- Digital Sovereignty and Confidence
- Digital Ecosystem Challenges
  - Academia, Enterprises, and Government Interaction
  - Sustainability



From: https://eroglumit.medium.com/the-convergence-of-it-ot-and-iot-2c4d5d22720a

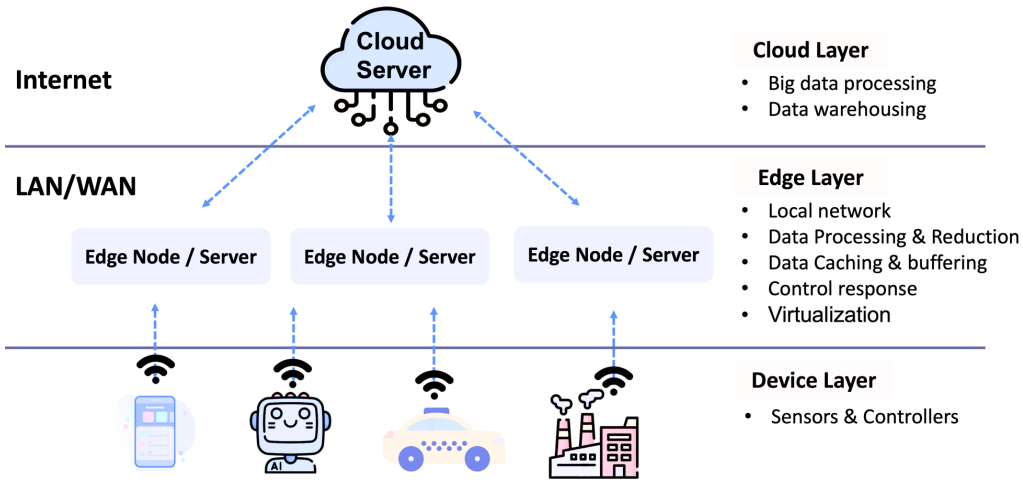
PLATFORM AS A SERVICE

- Affordable Infrastructure Management
  - Operational Support
- Flexible and Quick Development
- Performance Tradeoffs
- Target and Distributed Users

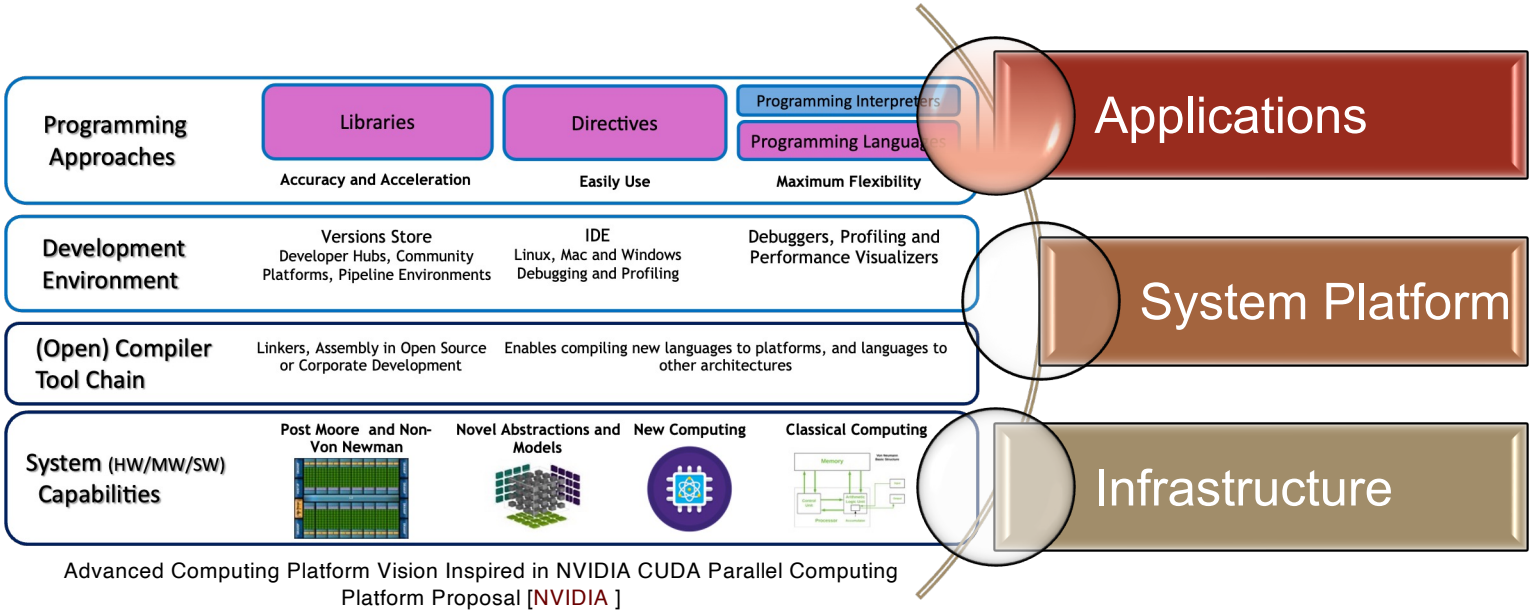


EDGE TO CLOUD ARCHITECTURE

- Edge Devices Integrated into a Cloud
- Data Generation for Edge Devices
- Computing Efficiency
  - Energy Concerns
  - Scalability
  - Real Time
  - Processing Efficiency
  - Data Generation
  - Privacy
- Distributed and Multilayer Deployment and Execution.



PLATFORM AT SCALE



Advanced Computing Platform Vision Inspired in NVIDIA CUDA Parallel Computing Platform Proposal [NVIDIA]

HPC TO ADVANCED COMPUTING

- **HPC** platform and framework create an **ecosystem** to tackle large-scale challenges.
- HPC requires a **multi-dimensional**, and **multi-scale** approach to research and development.
- **HPC is the foundation of advanced computing**, enabling cutting-edge solutions and expanding the limits of computational systems.

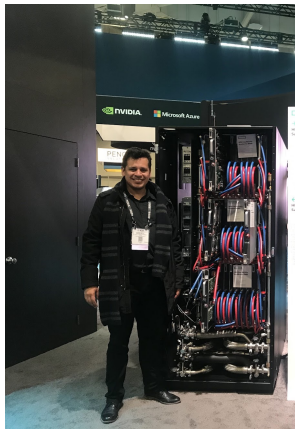
FROM HPC TO MULTISCALE HPC

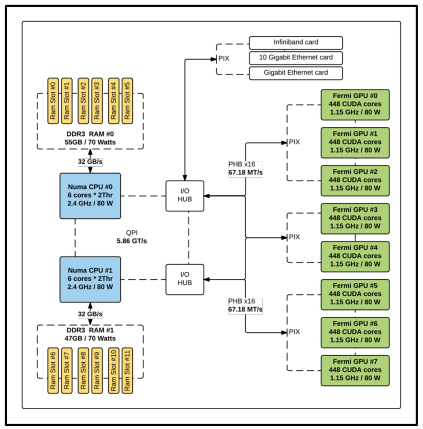
Hybrid, Heterogeneous, and Reconfigurable Large Scale HPC Systems

HPC/AI Convergence

HPC as a Service

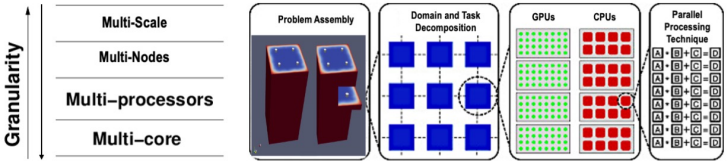
HYBRID, HETEROGENEOUS, AND RECONFIGURABLE HPC

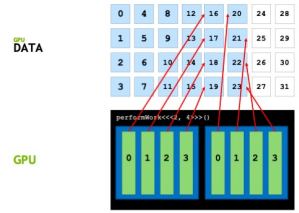




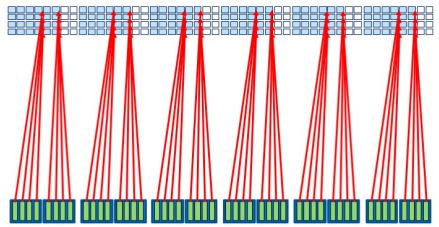
- Heterogeneous hybrid systems integrate diverse architectures and support Massively Parallel Processing (MPP).
- These systems pose multiscale challenges: parallelism optimization and efficient load-balancing techniques.

- Multilevel parallelism boosts computational efficiency and speed via coarse, fine-grained, and instruction-level parallelism.
- Multi-Scale HPC systems support efficient parallelism according to requirements and enable more complex applications.



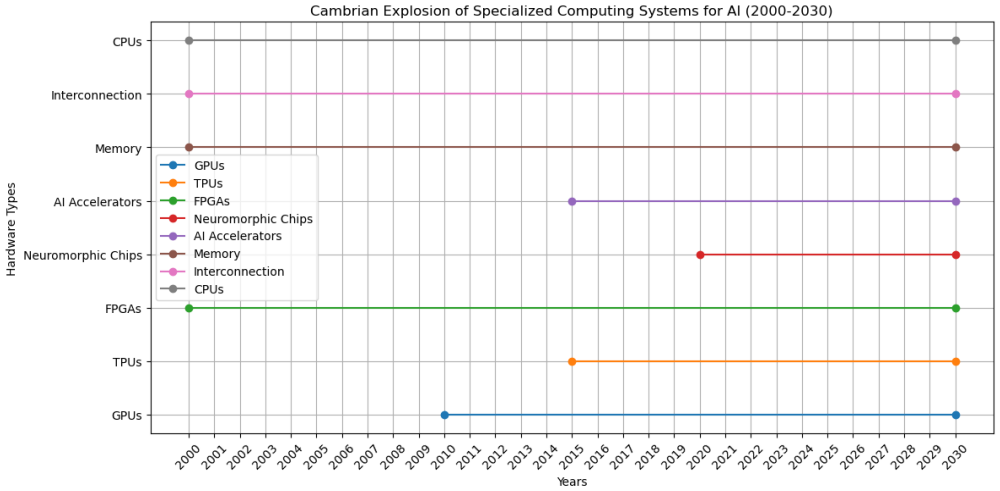


a. In a grid-stride loop, the thread's first element is calculated as usual, with **threadIdx.x + blockIdx.x * blockDim.x**

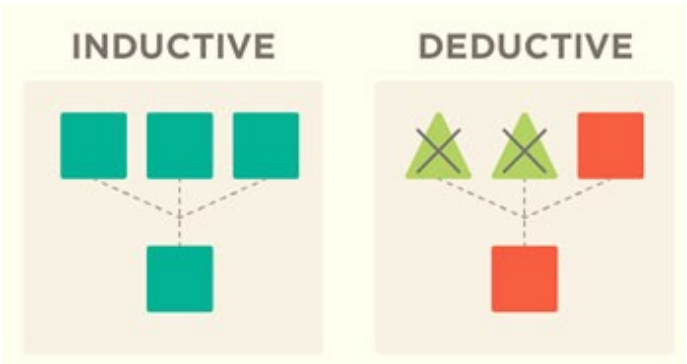


b. In a multi-GPU setup, if one GPU is overloaded while others are underutilized, **Effective load balancing ensures that all GPUs are working at optimal capacity.**

HPC/AI CONVERGENCE



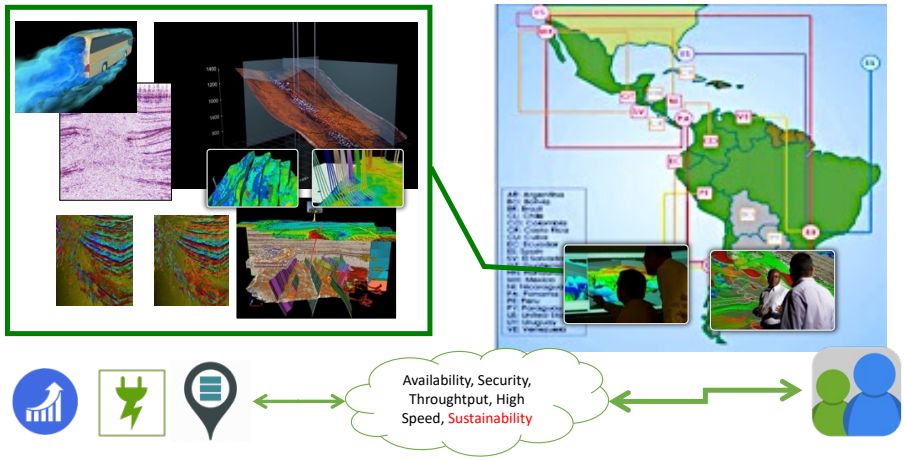
Analysis from: S. Matsouka, Cambrian Explosion of AI Systems Vision (2012)



From: Kadir, Bzhwen. (2020). Designing new ways of working in Industry 4.0. 10.13140/RG.2.2.33234.79041.

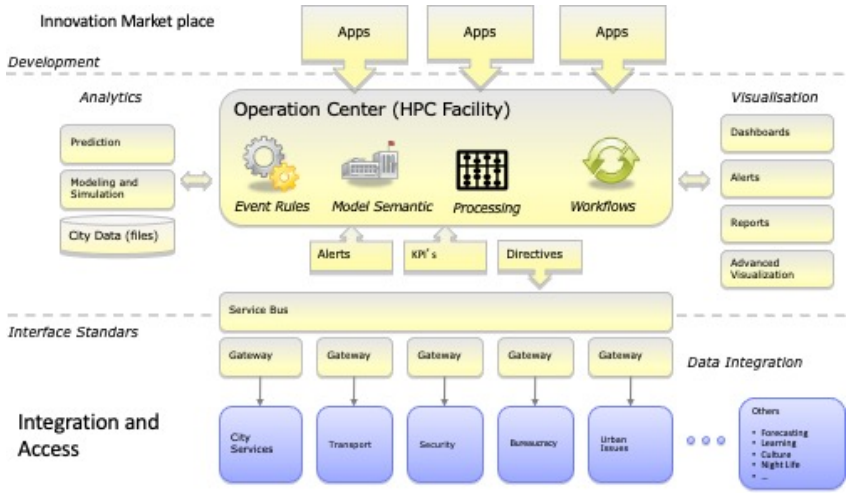
- AI needs HPC for algorithm implementation due to data explosion and intensive workloads.
  - Massive Parallel Processing (Induction)
  - Intensive Parallel Processing (Induction* / Deduction)

HPC AS A SERVICE



- Hybrid and heterogeneous HPC systems support diverse use cases at multiple scales.
- Interaction, integration, security, and compliance are essential in HPC systems, along with performance and usability.
- HPCaaS provides a scalable, sustainable, user-friendly, and reliable service model based on the Cloud Computing visibility model.
- Access to diverse computing technologies complicates HPCaaS and its interaction with multi-scale HPC systems.

SOME CHALLENGES



- Energy consumption and computing performance prediction
- Data hierarchy
- Cloud visibility model
  - In-Situ Cloud/HPC Support
- Data visualisation
- Link between nano-clouds

From www.catai.fr

COLLABORATION WITH COLOMBIA - “CATAÏ”

- Colombia
  - UniAndes, Bogota
    - Imagine, Commit, Colivri
  - UIS, Bucaramanga
    - GOTS, SC3UIS, SIMON
  - UNAL, Bogota
  - Alliance Française de Bucaramanga
- Government
  - French embassy
  - COLCIENCIAS
  - Grenoble
  - Bucaramanga
- Companies
  - Atos
  - Ecopetrol
- France
  - Université de Grenoble-Alpes, Grenoble
    - LIG
  - Université de Côte d’Azur, Nice-Sophia Antipolis
    - I3S
  - INRIA
  - INSA-Lyon
  - CPE- Lyon
  - Arts et Métiers, Paris
  - La Fabrique de la Cité

https://www.catai.fr

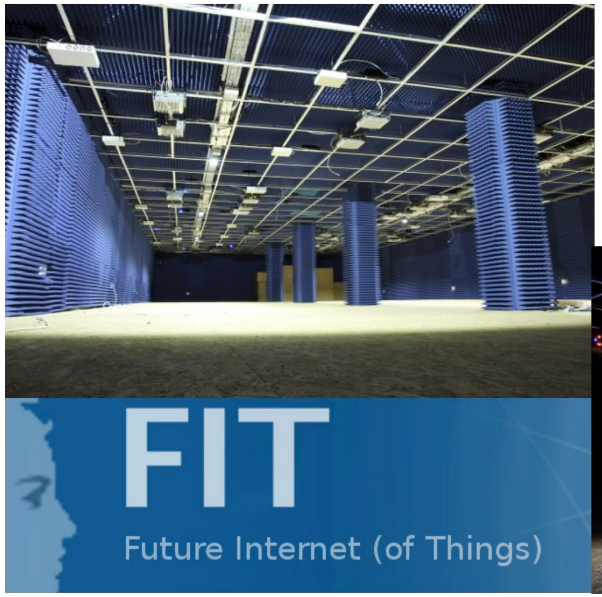
CATAI

EXPERIMENTAL PLATEFORMS

- Building a continuum IoT (CortexLab) – Edge, Fog (YOUPI) - Cloud (Grid5k) @Lyon, France
- Reproducible experiences
- Open access
- Heterogenous hardware (ARM, GPU, TPU, FPGA)

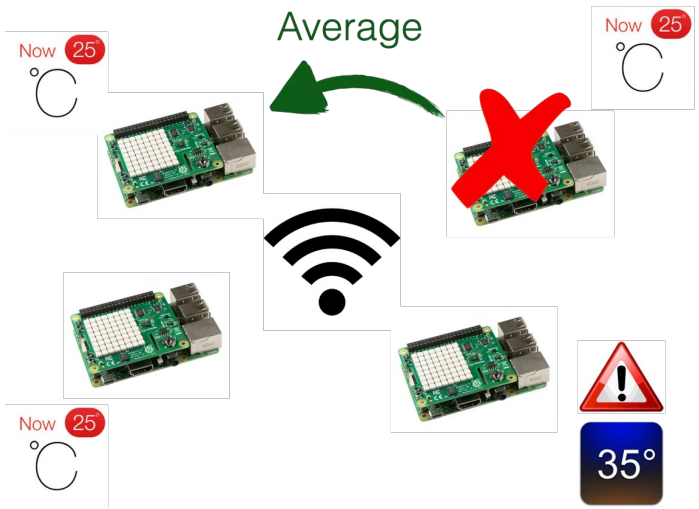
FIT/CORTEXLAB

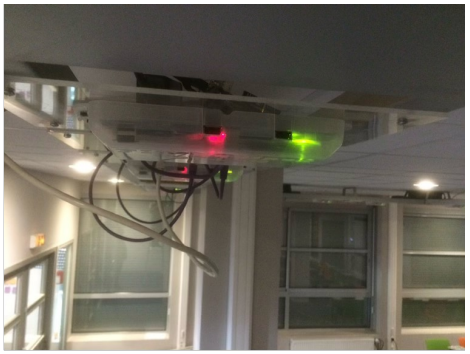
- Future wireless systems (5G, 6G, etc…)
- New waveform design
- Internet of Things
- and many more…



YOUPI

- A testbed for edge and fog computing
- Edge Nodes : Raspberry
- Fog Nodes : NUC, Jetson Nano
- Dedicated PoE Network



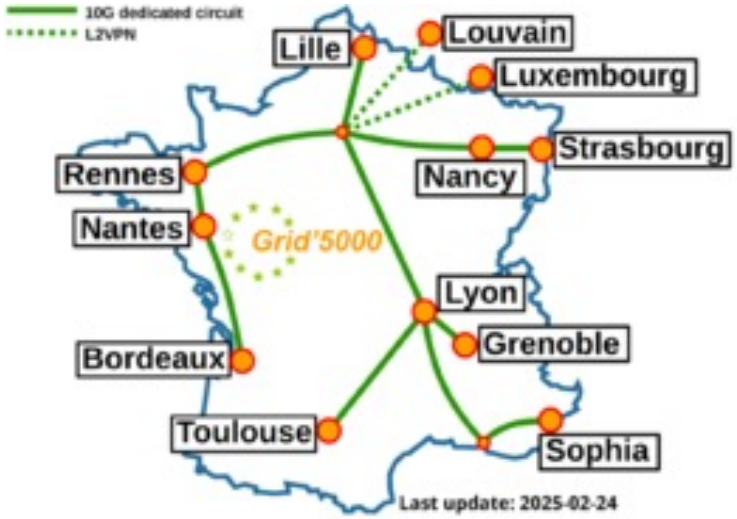


25+ RPI
3 NUC
3 Nvidia

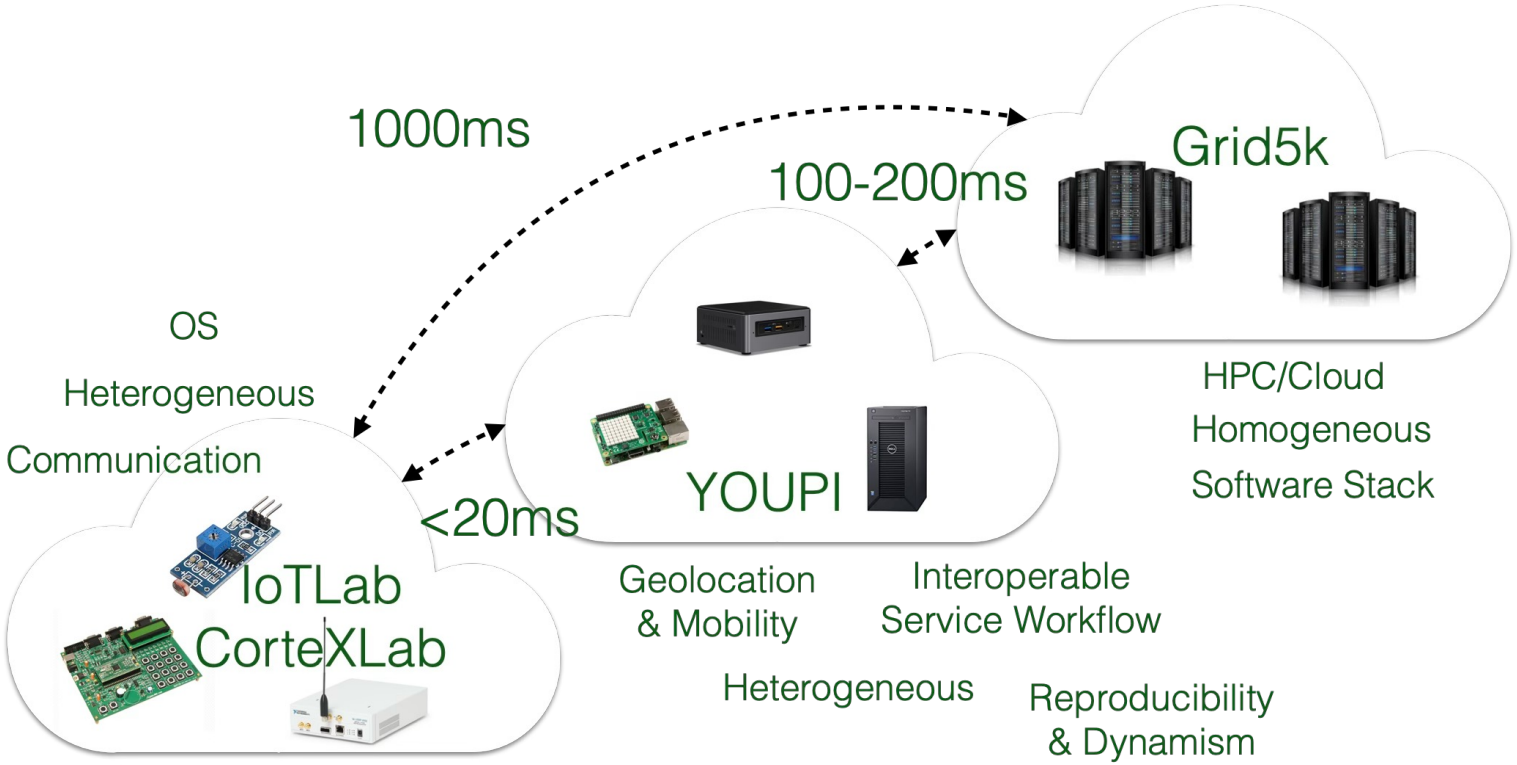
citi lab
SPIE
Red Hat

GRID 5000

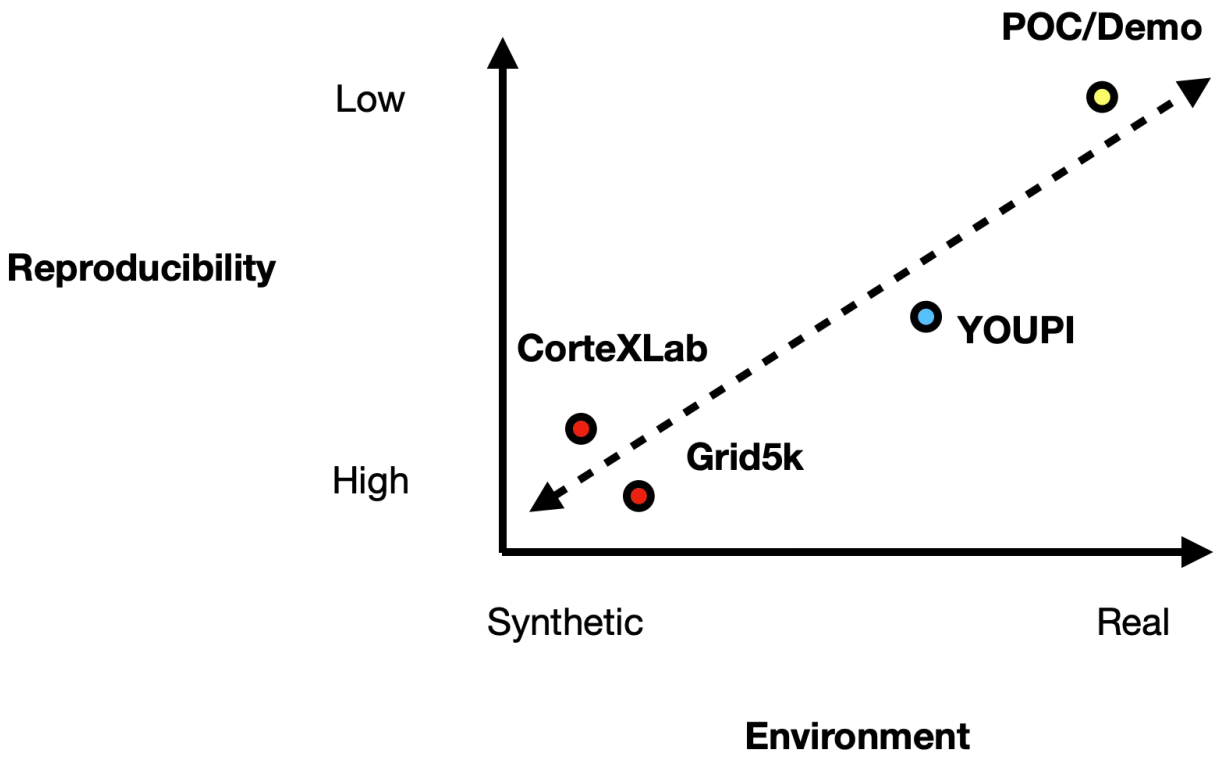
- A large-scale testbed for distributed computing
- 12 sites, 828 nodes, 15000 cores
- Dedicated 10-Gbps backbone network
- 550 users
- Advanced monitoring and measurement features for traces collection of networking and power consumption
- Highly reconfigurable and controllable



YOUPI VS FIT(IOTLAB/CORTEXLAB)/GRID5000



REPRODUCIBILITY



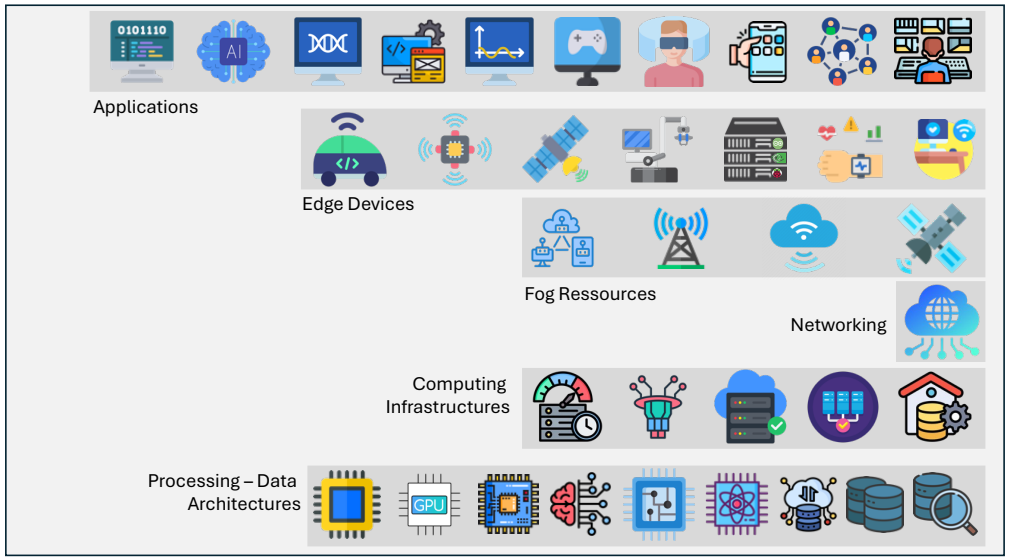
FROM MULTISCALE HPC TO COMPUTING CONTINUUM

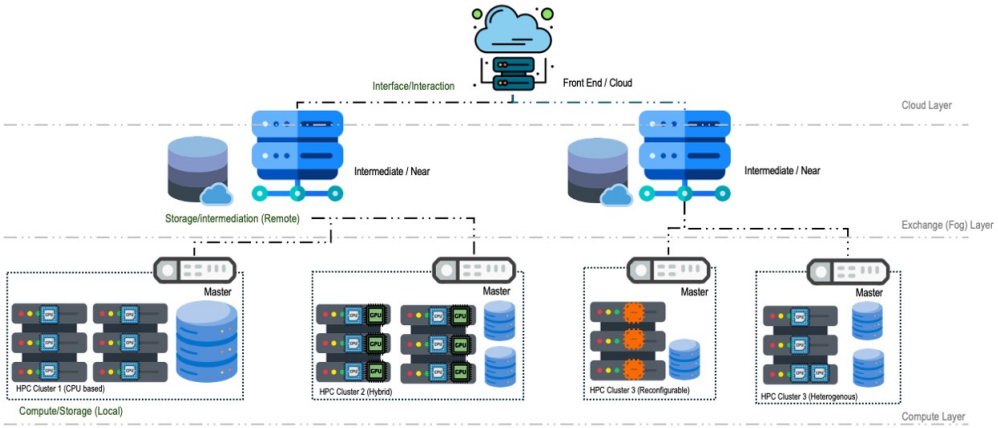
Ecosystem Approach

Massive and Intensive Workloads

Sustainable and Efficient HPC

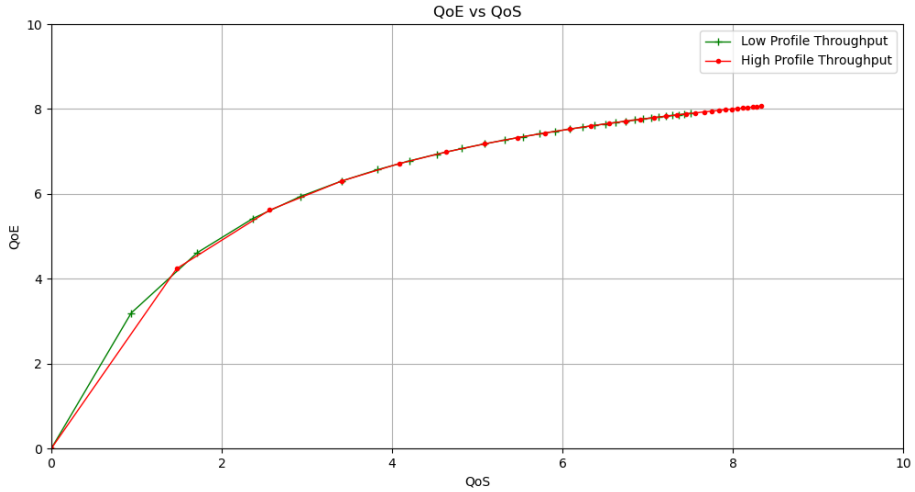
THE ECOSYSTEM APPROACH



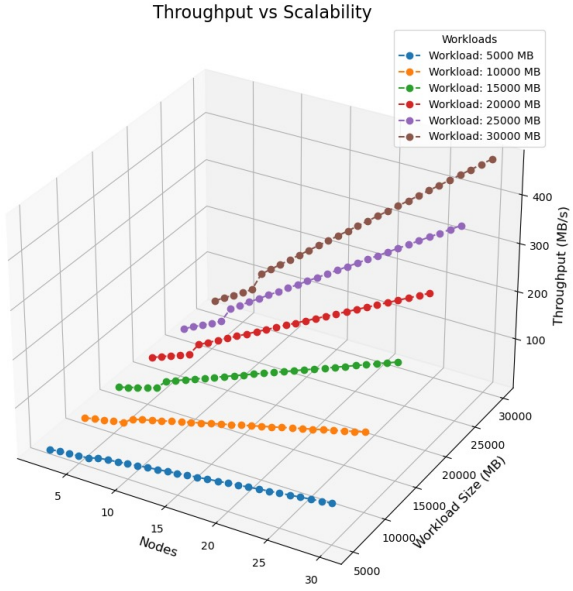


- A multi-scale HPC system handles diverse computational workloads in an HPC ecosystem, using advanced computing elements for large data processing and calculations.
- An HPC ecosystem includes components that interact to create workflows and form a computing continuum.
- Computing continuum activities manage massive workloads in data processing, transfer, and storage.

CHALLENGE: MULTIDIMENSIONAL GUIDELINES



- Analyzing performance in computing environments, especially resource allocation, offers insights into Computing Continuum systems through multiscale guidelines.
  - Performance and Sustainability Metrics
  - Workloads and Workflows
  - QoS, SLAs and QoE* Analysis



Scientific Advising Related Works:
Efficient Orchestration and Deployment for Multi-scale Computing Systems. (Pablo Rojas)
Optimizing Scientific Workflows for Massive and Intensive Data Transfers in Large Scale Scientific Applications (Alexander Martinez)

OPEN QUESTIONS

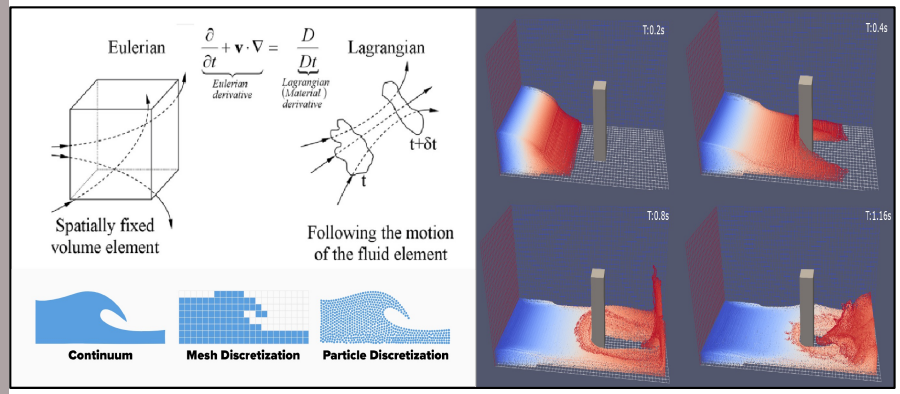
Ref.	Objectives	Bare metal	VMs	CTs	Contention	Evaluation
[23]	Energy	-	•	•	-	Simulation
[24]	Energy	-	•	•	-	Simulation
[25]	Energy, QoS	-	-	•	-	Real
[26]	Energy, QoS	-	-	•	-	Real
[27]	Energy, QoS	-	•	•	-	Simulation
[28]	Energy, SLA/QoS	-	•	•	-	Simulation
[29]	Energy, performance	-	•	•		Simulation
[30]	Response time	-	-	•	-	Real
[31]	Energy	-	-	•	-	Simulation
[32]	Deadline, cost	-	-	•	-	Simulation
[33]	Cost, QoS	-	•	•	-	Simulation
[12]	Energy	•	•	-	•	Simulation
[14]	Utilization	-	-	-	•	Simulation
[15]	Energy	•	-	-	•	Real
[16]	Accuracy	-	•	-	•	Simulation
[34]	Energy, Utilization	-	•	-	•	Simulation
[35]	Energy, Utilization	-	-	-	•	Real

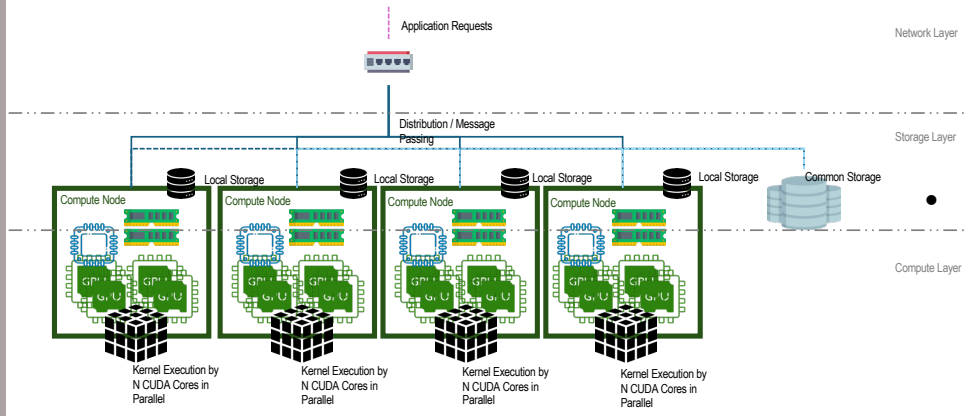
https://doi.org/10.1371/journal.pone.0261856.t001

From: : Dynamic performance–Energy tradeoff consolidation with contention-aware resource provisioning in containerized clouds Rewer M. Canosa-Reyes, Andrei Tchernykh ,Jorge M. Cortés-Mendoza, Bernardo Pulido-Gaytan, Raúl Rivera-Rodriguez, Jose E. Lozano-Rizk, Eduardo R. Concepción-Morales, Harold Enrique Castro Barrera, Carlos J. Barrios-Hernandez, Favio Medrano-Jaimes, Arutyun Avetisyan, Mikhail Babenko, Alexander Yu. Drozdov (2022) PLOS ONE 17(1): e0261856. https://doi.org/10.1371/journal.pone.0261856

- In modular multi-scale HPC systems, separate metrics fall short. How does linking metrics to analyses improve profiling of system components and executions?
- Consolidating services reduces resource waste and energy consumption. How do allocation strategies affect job type analysis, such as network and disk-intensive jobs, in different scenarios?
- The Computing Continuum includes orchestration and DevOps. How should we prioritize secure strategies?

MASSIVE AND INTENSIVE WORKLOADS





- Compute-intensive workloads require important power for scientific computation, visualization, or AI tasks.
- Data-intensive workloads handle large data, such as Big Data analytics or ML techniques.
- Multi-Scale HPC systems efficiently manage dynamic workloads, enhancing parallelism and scalability while reducing resource contention and boosting performance.
- Managing dynamic workloads requires efficient data distribution and memory, regardless of data volume.

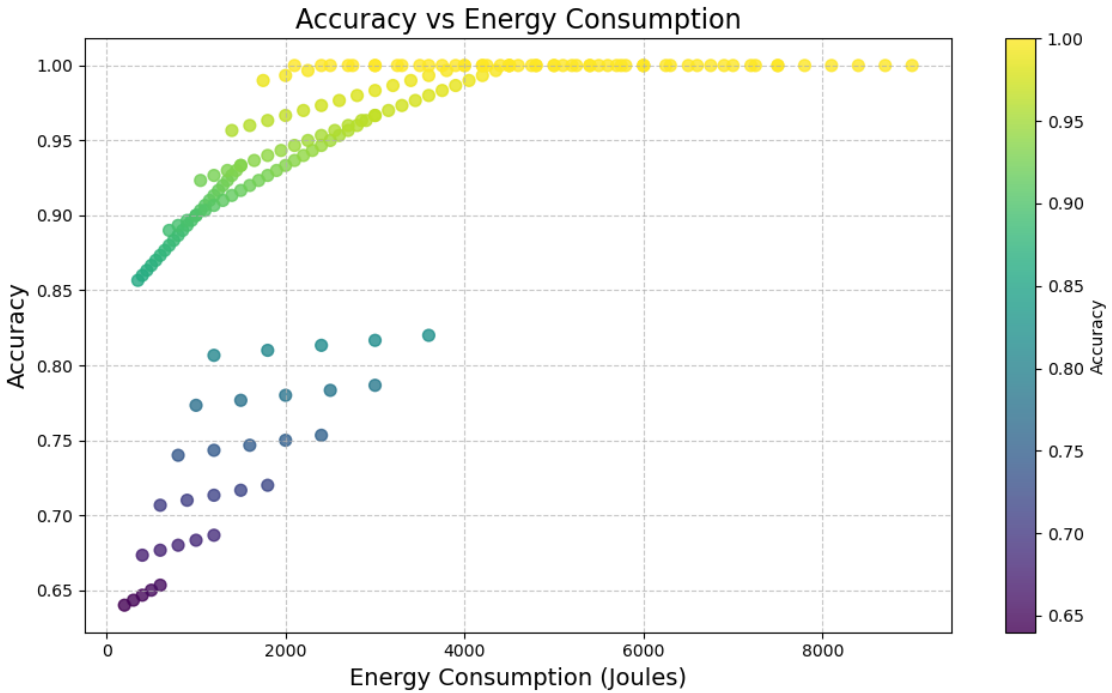
SUSTAINABLE AND EFFICIENT HPC

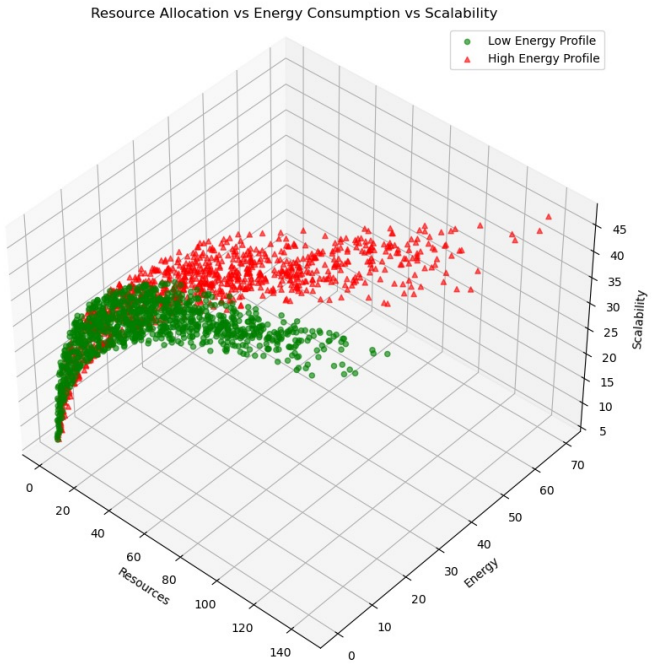
Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	222	**JEDI** - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, ParTec/EVIDEN EuroHPC/FZJ Germany	19,584	4.50	67	72.733
2	122	**ROMEO-2025** - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, Red Hat Enterprise Linux, EVIDEN ROMEO HPC Center - Champagne-Ardenne France	47,328	9.86	160	70.912
3	440	**Adastra 2** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, RHEL, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	16,128	2.53	37	69.098
4	155	**Isambard-AI phase 1** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE University of Bristol United Kingdom	34,272	7.42	117	68.835
5	51	**Capella** - Lenovo ThinkSystem SD665-N V3, AMD EPYC 9334 32C 2.7GHz, Nvidia H100 SXM5 94Gb, Infiniband NDR200, AlmaLinux 9.4, MEGWARE TU Dresden, ZIH Germany	85,248	24.06	445	68.053

Source: Green500 https://top500.org/lists/green500/2024/11/

- Computational efficiency measures system performance through resource use, especially energy per task, varying with workload.
- HPC systems rank by performance per watt, with efficiency declining in larger systems, thus favoring smaller ones with similar technologies.
- HPC systems can balance performance and sustainability by characterizing workloads, optimizing resource usage, and adopting energy-efficient technologies.
- Energy-efficient designs in HPC solutions reduce environmental impact.

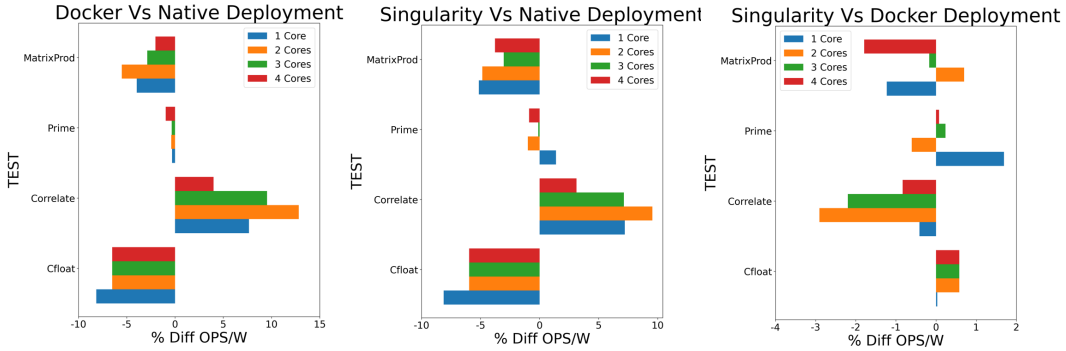
OUR CONTRIBUTION: PERFORMANCE VS SUSTAINABILITY



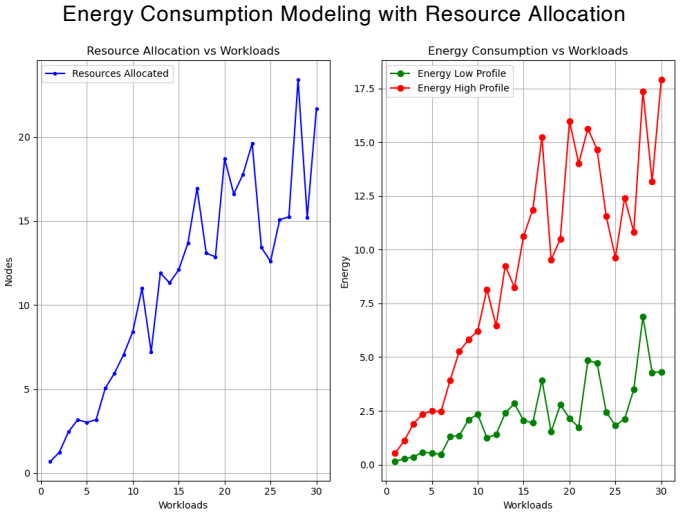


- Evaluating performance techniques for multi-scale HPC systems, focusing on efficiency and resource management.
- Adopt a holistic approach integrating all levels of a multi-scale HPC system, incorporating technologies, applications, and best practices into a unified platform.

OPEN QUESTIONS



From: P. J. Rojas Y., C. J. Barrios H., and L. A. Steffenel. 2023. Understanding Energy Performance of Containers Deployment on HPC-Based post-Moore Platforms. In Proceedings of the SC '23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis (SC-W '23). Association for Computing Machinery, New York, NY, USA, 147–154. https://doi.org/10.1145/3624062.3624586



- Examining two profiles in a multi-scale HPC system reveals ways to optimize energy consumption. How can we balance performance, sustainability, and scalability?
- Hierarchical processing and resource use reveal unique characteristics, such as energy profiles. Throughput is the main metric for system characterization; how can we quantitatively integrate QoS and QoE?

DISCUSSION

DISCUSSION: ADVANCING COMPUTING CONTINUUM

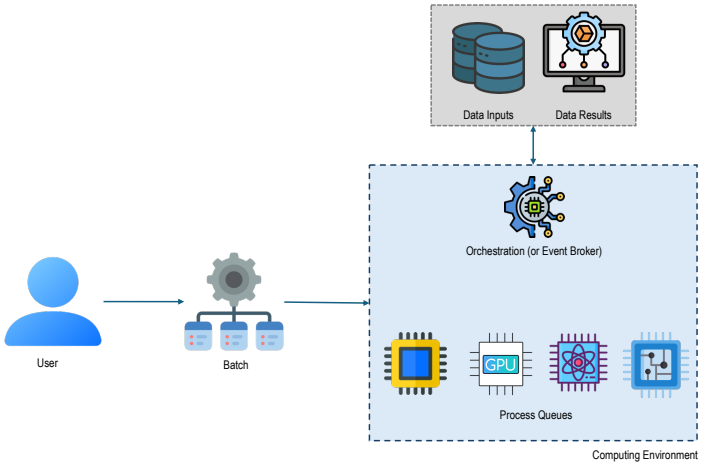
How to Exploit Better Paralleism?

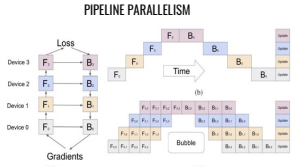
How to Exploit Efficiently Advanced Computing Environments?

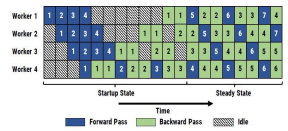
How to Achieve Efficiency and Scalability?

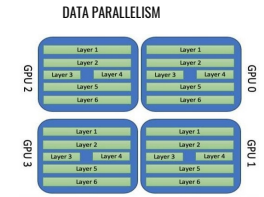
How to support new computing challenges (as Quantum Computing)?

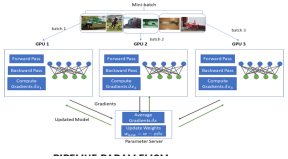
HOW TO EXPLOIT BETTER PARALLELISM?





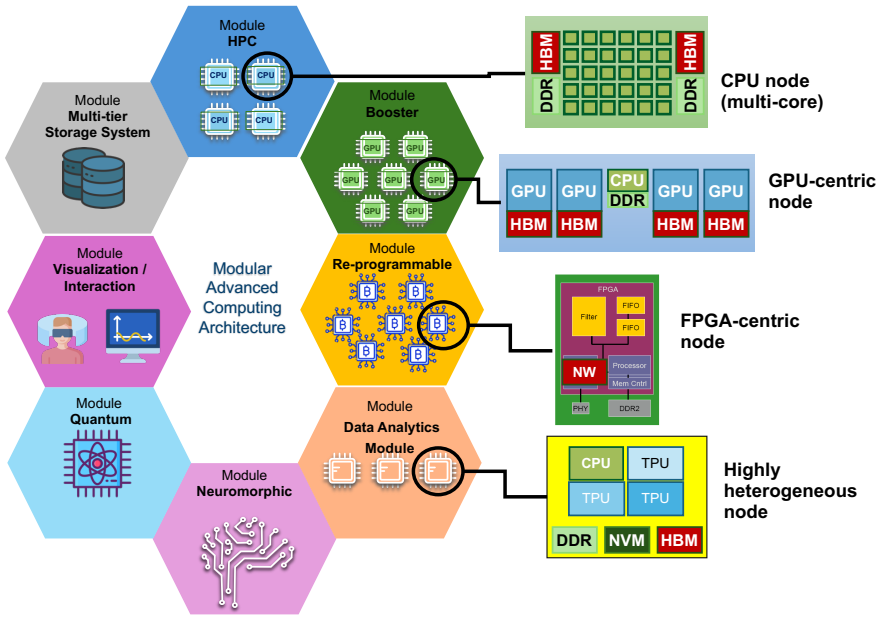






- Managing parallel computing environments through efficient resource allocation and coordinated task execution.
  - Orchestration (or Choreography) strategies
  - Combination of various hardware components and an innovative software stack optimized for specific tasks.
- Understanding the workloads and matching them with the appropriate level of parallelism.
  - Proposing specialized algorithms and implementation mechanisms.
  - Implement multithreading, multiprocessing, loops, and vectorization techniques.

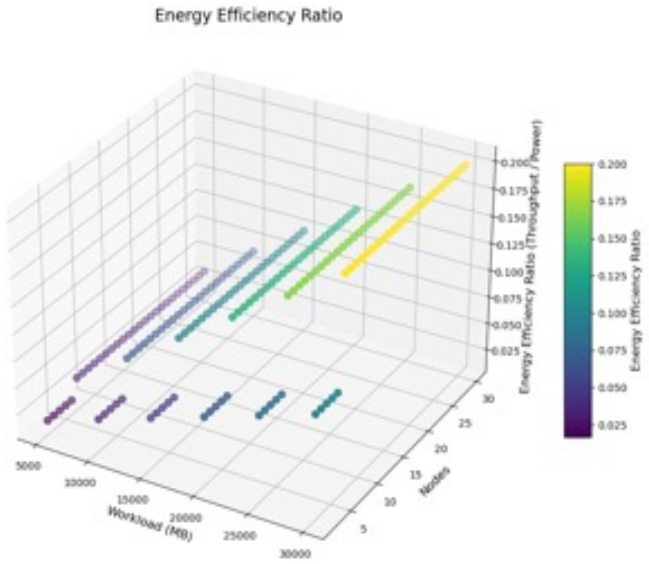
HOW TO EXPLOIT EFFICIENTLY ADVANCED COMPUTING ENVIRONMENTS?



Modular Advanced Computing Architecture inspired in [Suarez et al. 2022]

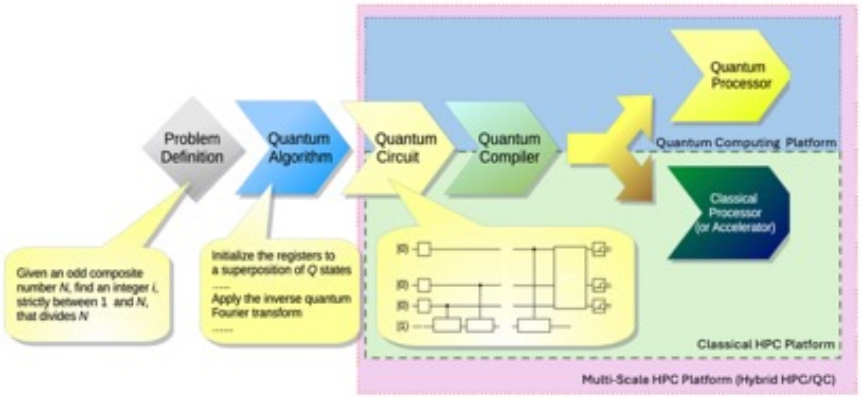
- Analyzing advanced computing architecture modularity, infrastructure (components) integration, applications support, and programming models.
- Optimal integration of key characteristics such as specialization, communication, and scalability amidst technological diversity (and hybrid heterogeneous support).
- Observing advanced computing architectures and workload management complexities.

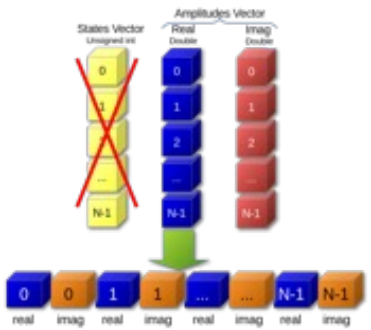
HOW TO ACHIEVE EFFICIENCY AND SCALABILITY?

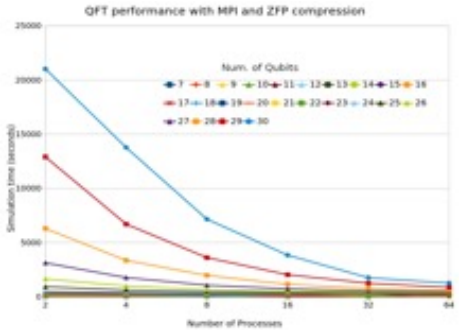


- Confront Accuracy vs Performance vs Sustainability, to propose metrics from a multidimensional perspective.
- Identify behavioral patterns by examining sustainability metrics and characterizing a ratio that outlines multi-scale HPC systems.
- Streamline guidelines to enhance execution and optimize HPC systems with better configurations, boosting computational efficiency while maintaining performance and accuracy.

HOW TO SUPPORT NEW COMPUTING CHALLENGES (AS QUANTUM COMPUTING)?





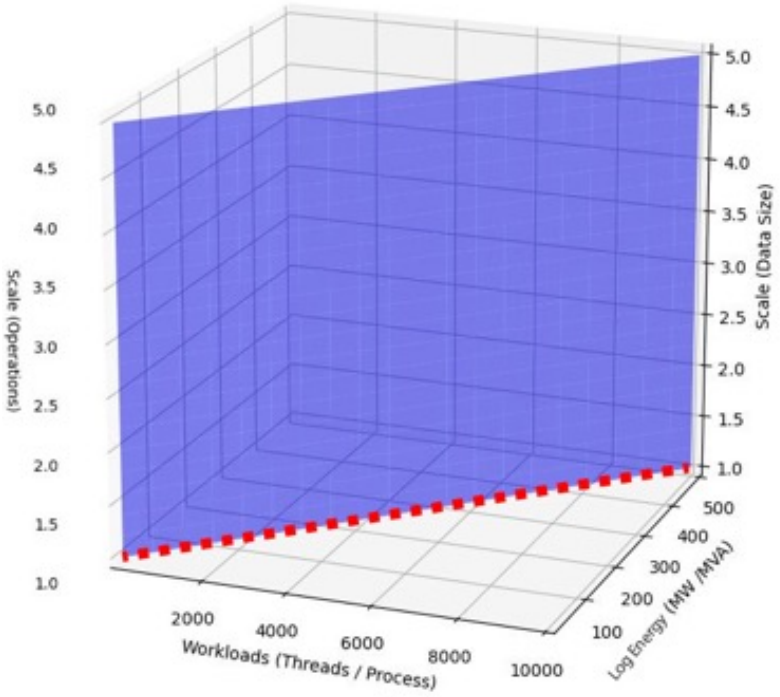


Scientific Advising Related Works:

- *Memory Management in Software Quantum Computing Simulators.*

- Integrating quantum and classical HPC creates hybrid multi-scale (HPC) systems.
- A hybrid model combining quantum steps in multi-scale HPC systems with quantum and classical components enables (real) quantum computing.
- Memory management is essential for scalable hybrid systems. Enhancing data exchange and distributed quantum memory boosts quantum computing in multi-scale HPC.

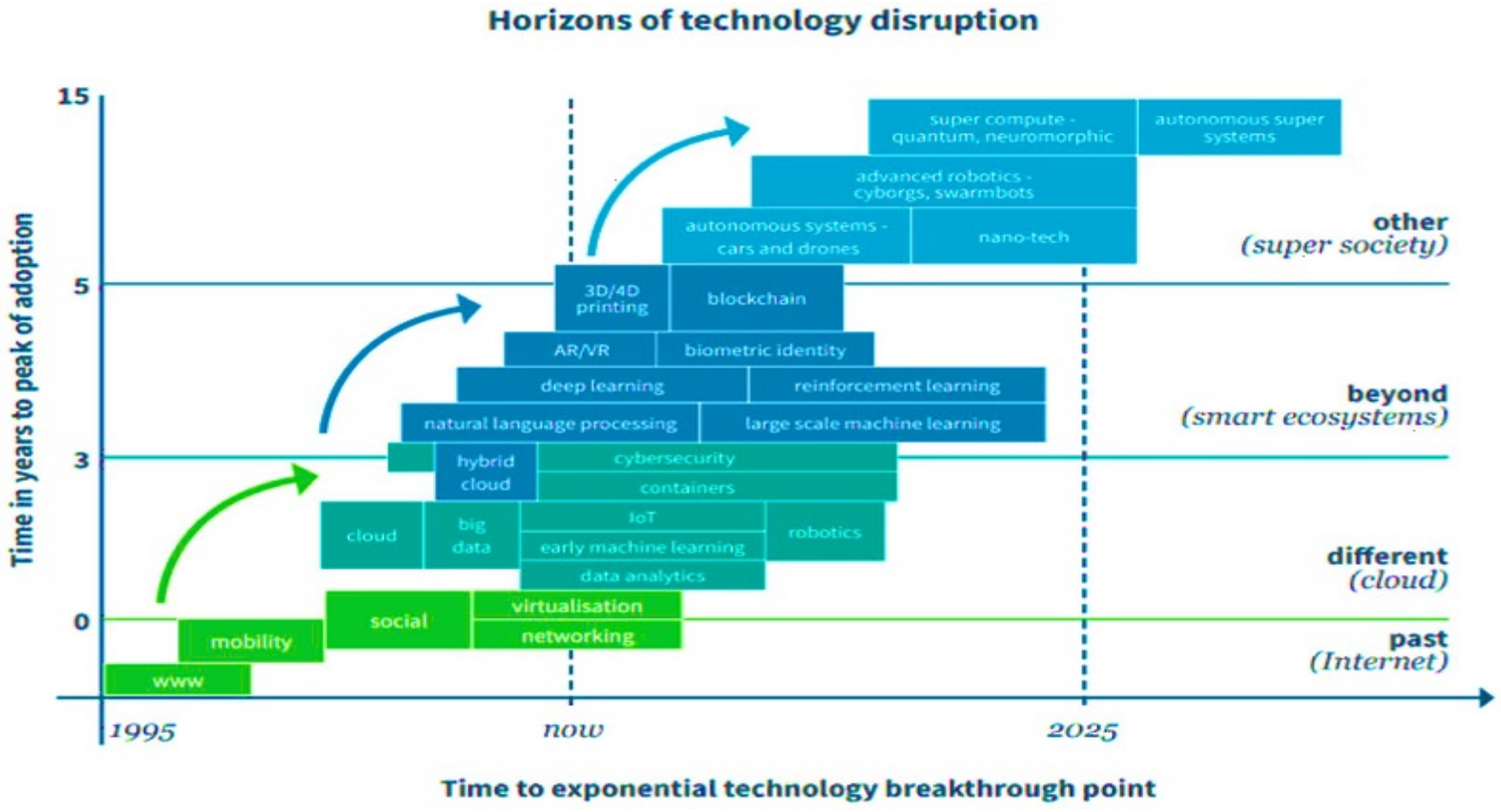
CONCLUSION: TOWARDS A POST-EXASCALE ADVANCED COMPUTING



- High throughput allows efficient computations, and scalability enables systems to adapt to growing demands without sacrificing performance. These aspects improve computational capabilities.

- How does this holistic perspective reshape computational systems for optimal performance, accuracy, and sustainability?
  - Don’t forget Human impact!

TECHNOLOGY DISRUPTION



From: http://asiandatascience.com/horizons-of-technology-disruption/

CONCLUSION AND FURTHER WORK

- Edge to Cloud into the Computing Continuum drive digital transformation through automation, scalability, and data insights.
  - Challenges such as orchestration, energy efficiency, and Edge-to-Cloud sharing must be addressed to realize their full potential and will be included in future developments.
- Integrating OT and PaaS guarantees high availability, performance, consistency, and resilience in contemporary IT and cloud environments.
  - This intersection highlights the need for efficient platform management and search automation in ongoing developments.
- The transition to Edge-to-Cloud sharing and Operational Technology evolving into EaaS raises questions about service models and energy efficiency.
  - These concerns may alter our perspective on the implementation of operational technologies within the Cloud Continuum.
- The computing continuum is not only a technological challenge but also an opportunity for international collaboration and for building bridges across IoT, HPC, and Cloud

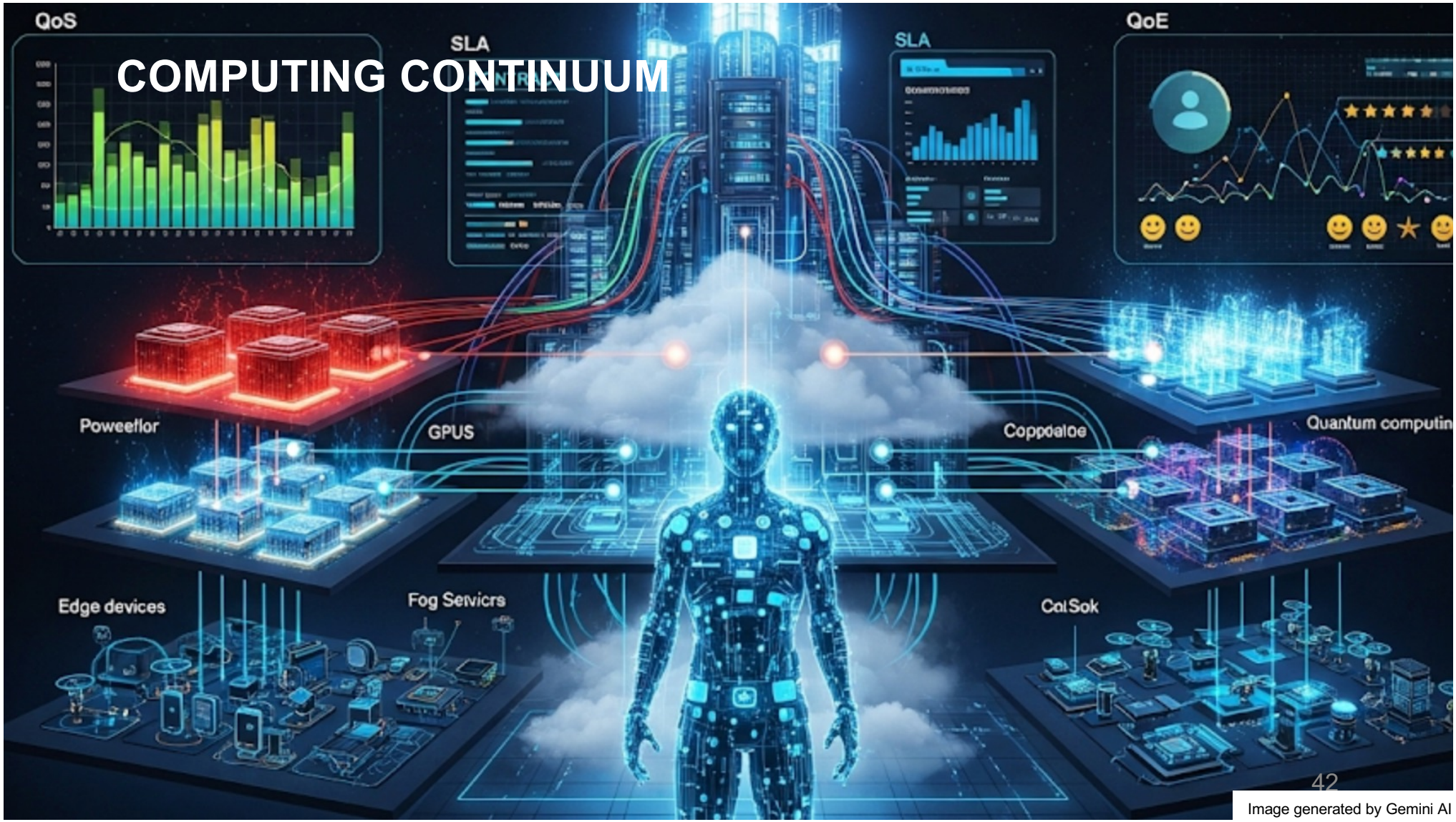


Image generated by Gemini AI

Gracias, Merci, Thanks

Questions, Comments?

EL-BONGÓ physics

Co-funded by the European Union

CATAI Collaboration France-Colombia

SCALAC

Universidad Industrial de Santander

Super Computación y Cálculo Científico UIS

Inria

UGA

LIG

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

citi lab