HPC/AI Divergence Discussion Notes

Co-chairs: Mohamed WAHIB (RIKEN), Emmanuel Jeannot (INRIA), Artur Lorenzon (UFRGS), Philippe Navaux (UFRGS), Brian Spears (LLNL)

1- HPC–AI CONVERGENCE

- **Strategic autonomy in HPC–AI**
  - Control: compute + data + software + models + operations
  - Collaborate globally
- **Dependencies**
  - Cloud
  - Chips
  - Software
  - Data
- **Key point**
  - Not isolation
  - Freedom to choose (no lock-in)

2- FIVE LAYERS OF SOVEREIGNTY

- **Compute sovereignty:** control over accelerators, storage, interconnect
  - Others (US, China, Japan?) are investing in alternatives to Nvidia
  - Europe needs focussed effort (small team, long term sufficient funding, sustained focus, no political intervention)
- **Data sovereignty:** control over where data lives and flows
  - At lot on foundational data is done
  - Training sets for open source are available
    - Need to investigate availability for different languages

        - Scientific data collections should be curated - to build domain specific models
            - Need to remove barriers, international collaboration
- **Software sovereignty:** portability and auditability of stacks
    - Very dynamic market - open opportunities, can be done quicker than hardware
- **Model sovereignty:** ability to train, adapt, deploy independently
    - New foundation models need lot of resources (o(100k) GPUs over at least a year) plus small focussed team
    - Increase diversity in models (diverse reasoning abilities, diverse training data, different learning models on data)
- **Operational sovereignty:** control over infrastructure and expertise
    - European cloud market volume is there - but it is not invested in Europe
        - Needs to be business driven, not EC
    - Japan tries to build up cloud market at the moment
    - China is building up (or has it)

3- WHY SOVEREIGNTY IS NOW A RESEARCH ISSUE

- **Frontier AI:**
- depends on large-scale compute and infrastructure

  => sovereignty = access to AI-capable systems
- **Europe's response:** AI Factories (EuroHPC) + AI Continent strategy (including AI GigaFactories soon)

- **Implication:** sovereignty is a technical HPC issue, not only geopolitical

4- Scalability

- **Four topics:** Mixed precision, Memory Limits, Slow Interconnection, Heterogeneous Architectures.
- One challenge is the **heterogeneity** of hardware (accelerators):
  - Becoming hard to program and exploit all available resources (software stack).
  - A language that can help the user to define the locality and how to move data between different accelerators.
  - We are trying to fit HPC to AI machines.
- Regarding mixed precision:
  - What new research do we need to have a more efficient FP64?
  - Applications should be rewritten to better leverage the new AI hardware features provided by vendors.
  - Use profilers to understand what the ideal (and necessary) precision is for the application (RAPTOR: Practical Numerical Profiling of Scientific Applications)
- Regarding Interconnection:
  - network programmability via interface cards (smart switches) to improve performance
  - However, complex to program and use at the user level

5- Sovereignty :

What are the concerns driving sovereignty

* need to keep expertise on software stack
* gives hardware choice
* data
    -> on prem - cost

Cost : cloud v on prem
High utilisation means cheaper costs (token maxing)
Sovereignty and open source -> avoids silos
But
Geopolitical

-> need to be contributor not user
    -> advance technology
    -> create and contribute SMEs to country economy

Sovereignty fairly need discussion point

Around 2016 : more prevalent during Covid with supply chain issues : now geopolitical

6- Scalability :

FP64 bit precision is not featuring in all new hardware in some vendors' new chip (or turned off)
Focusing on AI drivers so with HPC v AI : AI dominating market
Emulation is required and not favoured by all

Ozaki Scheme -> lower power; more INT8 -> can be a misleading selling point

Hardware choice will determine software

HPC will need to change workflows incorporate AI

CRITICAL: Can only buy AI hardware unless there is a big market change
Will require good validation techniques following HPC application code porting

7- Data sovereignty

- Data sovereignty is a global and dominating issue
- The "embedding" of data in the models creates a high level of complexity in regard to ensuring data sovereignty
- This is NOT isolated or specific to the HPC + AI community, but we currently have (or try to have) the necessary infrastructure, we can and should play a role into solving this problem for society
- Related to trust: depending how paranoid your countries/communities/companies is about sharing / leaking data, the more effort you have to spent

8- This subgroup discussed two main topics: mixed precision and AI/HPC resource utilization.

For the mixed-precision topic, we discussed its impact in domains where numerical accuracy is required, particularly in bioinformatics and applications used by oil and gas companies. These areas are still largely dominated by FP64 workloads, as small numerical deviations can affect scientific validity, simulation stability, and decision-making processes. In this context, the discussion focused on how the community can safely exploit lower-precision formats without compromising correctness. Possible alternatives included selective mixed precision, where only numerically tolerant kernels are executed in lower precision; adaptive precision schemes guided by error bounds; validation

workflows comparing reduced-precision outputs against FP64 baselines; and domain-specific criteria to define acceptable accuracy loss. The group also emphasized that mixed precision should not be treated only as a hardware optimization, but as a co-design problem involving numerical methods, application requirements, scientific reproducibility, and domain validation.

For the second topic, which concerns resource utilization and strategies for identifying the most appropriate hardware accelerator for each kernel or application type, the discussion considered that future HPC/AI systems will likely combine CPUs, GPUs, AI accelerators, network-attached processing units, and potentially domain-specific accelerators, making static resource assignment increasingly inefficient. Potential strategies include profiling applications at the kernel level, characterizing computational patterns such as memory bandwidth pressure, communication intensity, arithmetic intensity, precision requirements, and data movement costs, and then mapping each workload to the hardware that provides the best trade-off among performance, energy efficiency, accuracy, and system throughput. Data-driven approaches were also discussed as a possible direction, using historical execution traces, performance counters, runtime telemetry, and application metadata to guide scheduling and placement decisions. In this direction, intelligent runtime systems could learn which accelerators are better suited for specific workload phases and dynamically adapt execution to improve overall facility efficiency while preserving user-level performance objectives.

9- **AI–HPC divergence / convergence**

- Need some form of virtualization to deploy AI/cloud applications
- Containers + elasticity + multi-tenancy
- You want things to be dynamic and elastic
- Hybrid architecture to bring things together
- Micro-heterogeneity: inference is pushing for diversification with disaggregated inference
- No optimization for cost yet, but it will come
- Need to be first, then to have the best, but sooner or later optimize TCO
- Quantum devices: how should they be connected? What kind of latency requirements?
- Elastic computing within a standard HPC system is possible
- Flexible storage: everyone has its own API
- S3 storage how to access them in an abstracted way

10- Long presentation of wahib on converegce/divegence of AI and HPc on differetn aspect:
- Hardware, precision, coding, sofwtare, etc.