DRAFT

Industrial Users: On the usage of EuroHPC machines

Draft Chapter Strategic Research Agenda 6

Edited by: the ETP4HPC Office (office@etp4hpc.eu)

Executive Summary

This document discusses the requirements of European Industrial users of HPC in light of the availability of EuroHPC computing infrastructure. EuroHPC machines will predominantly used by scientific users. Bringing in industrial users into the same infrastructure is bound to introduce unforeseen challenges both for the users as well as for the infrastructure providers/EuroHPC. Having computing infrastructure that is capable and adaptable for both scientific and industrials will help to propel Europe's societal and technological advancement. We summarise below some of the perspectives from the industrials.

The ETP4HPC has put together an Industrial User working group consisting of some of the major industries that have been leveraging HPC resources for industrial innovation for the last many years/ decades. Many of the HPC systems used to be in-house and there has been a lack of availability of public infrastructures that cater to their specific needs. However the new EuroHPC machines offer the possibility for incumbent industrial users to expand their computing capabilities, as well as for new industries to explore and exploit the power of advanced computing.

Introduction

Industrial HPC end-users rely on a large set of core applications, which perform simulations, optimisations, complex data analysis, and increasingly AI/ML processing. Such core applications continue to cover a huge scope - such as simulation of internal and external flows (CFD), combustion and other chemical reactions, structural analysis (often using finite element codes), materials sciences, molecular dynamics, quantum effects, circuit simulations, financial risk analysis, docking and protein folding, *-omics, traffic and crowd simulations, weather, large-scale graph analytics, discrete optimisation and many others. While the algorithms used in these do not differ widely from scientific applications, differences are caused by the different scale and complexity of problems, the frequent need for ensemble scaling to drive optimisations, and the role of closed-source, licensed applications. They can also influence the choice of system technology & design.

The use of HPC for emerging Machine Learning (ML) and Artificial Intelligence (AI) applications in the various fields of engineering allow, among other things, to develop Digital Twins and model-based design methodologies. The combined power of new methods and new computing architectures facilitate innovation, increase the efficiency and safety in many critical engineering areas, ultimately contributing to a sustainable digital transition.

The following are some of the areas that needs attention.

Data Protection

Commercial users store and control significant amounts of data with high commercial value and highly sensitive data, and personal data in light of GDPR considerations. Such data commonly resides in the protected IT infrastructures of the customer, and it has to be transferred onto the IT infrastructure of a HPC provider, accessed and modified via HPC applications and workflows there, and transferred back (results) or purged (input data which is no longer useful). It is critical that technical measures to protect data from unauthorised access and modification are in place throughout the complete data processing chain. In addition, effort and costs for data movement have to be kept under control. Data storage and appropriate data segregation should be thoroughly considered.

Security considerations

Industrial companies operate their own IT infrastructure. The industrial work environment relies on high-level frameworks and workflows, which match the business or engineering processes in place, and provide end-users with easy-to-use interfaces. Such frameworks and the corresponding high-level workflows have to be supported by any HPC usage model. Authentication, authorisation and accounting support to non-personal accounts may be required, and strong monitoring and security mechanisms are required for any operations that touches confidential data. The EuroHPC machines (& the personnel operating these) must thus ensure that industry standards for Cyber Security and Data Security are met. This has to be explained and potentially audited from/to the industry users including mandatory company conducted risk assessments.

Financial considerations and pricing

From a contractual point of view, the financial engagements linked to the requested capacity should be clear from the beginning, for example, the rental costs by Core/CPU/Node hour, etc. Larger upfront investment is difficult to handle from industry side. Pricing options should reconcile with the use of public funds in the TCO of these systems.

Accounting and billing must comply with the business requirements of the end users and accommodate paying per use. Corporate rules such as payment terms, for example, might be non-negotiable. Pricing models should consider dedicated queues or pseudo-idle resources ready to be used by industry.

Legal considerations

Contracts tend to become very complicated due to existing corporate and legal obligations (SLAs, DPA, IP, liabilities , etc). Also corporate rules such as payment terms may be non-negotiable. Forming

legal consortia to access the machines might thus be very involved and complicated. Such legal issues need to be considered.

Service level agreements

Due to (very) expensive licenses, queuing times and time to access need to be reduced to a minimum. To become (more) attractive, public infrastructure needs to outperform existing cloud offerings. There should be a consistent ecosystem across all EuroHPC sites and "paper work" should be kept minimum and simple. The big question is whether SLAs for scientific users can be adapted for meeting those of industry specific needs?

It is to be noted that there could be a need for tailored architecture for different workflows. In the industry there are typically two kinds of user groups - "computing intensive group" that needs high memory bandwidth processing where the filesystem use is not very intensive (eg: CFD users) and high throughout users that have small workloads and need fewer core and small wall clock times but they stress a lot, the filesystem (typically, users are computational mechanics, AI, and Big Data). These two different classes require different architectural setup, configuration and tuning that often are in opposition. The SLAs need to reconcile with these kinds of contradictions.

Experienced centres such as HLRS in Europe, catering to industrial users can inform and advice.

Interoperability with Cloud infrastructures

In order to enhance the capabilities of the company to "follow the data", it is important to establish a common technological framework including a HPC for Edge laboratory facility, and, software platforms and services to enable products and services to evolve toward the edge, and leverage cloud technology.

As digital ecosystems expand, the ability of systems, devices, and applications to work together becomes critical. Edge, fog, and cloud computing can support interoperability by standardizing protocols and interfaces, facilitating data exchange and process coordination, targeting both HPC and cloud resources in a distributed federated approach, via user-friendly tools and interfaces.

Software management

It's essential to implement a robust and efficient management system to simplify the use of software licenses on private clouds and to work together with the software vendor to optimize licenses usage in HPC environment. It should also include features for tracking and monitoring the usage of licenses, which can help in optimizing resource allocation and reducing costs.

Additionally, integrating automated tools for deploying and configuring HPC applications can significantly simplify the process and reduce the time required to get an application up and running.

Potential for new pricing models

Finally we explore the possibility of novel pricing models.

Manufacturers of tangible goods measure their efficiency and competitiveness against their competitors using metrics (key performance indicators) linked to the goods they produce. For example, a car manufacturer measures the efficiency of its factories according to three fundamental criteria: cost per car produced, number of cars produced per day and production time per car.

This method can also be applied to intangible assets. For example, a company that designs molecules, such as a pharmaceutical laboratory, would like to know in advance the calculation cost per molecule, the number of molecules simulated per 24 hours and the calculation time per molecule, in a predictable and fixed manner. A company that relies on inference queries wants to know in advance the cost of responding to a query, the number of queries processed per second and the latency to obtain the response. This enables them to build their business model on stable production and financial bases, giving them greater visibility over the evolution of their sales and profits.

All manufacturers and companies that use or will use "digital factories" for their business have switched all or part of their production to hyperscalers. Their operational (COO) and financial (CFO) managements are rapidly facing two critical problems: (1) the unpredictability of production costs (with the big hyperscalers, the number of parameters involved in the billing model is very large, they are linked to each other, and are difficult to understand and anticipate, making the billing model unreadable and unpredictable from one month to the next); (2) the monthly cost, very often much higher than before the switch of production to the cloud.

Based on this analysis, for companies that switch their production to the cloud or a HPC service model, it is necessary to develop an offering based on a commitment to results, also known as OaaS ("Outcome as a Service"), rather than a simple commitment to means, however sophisticated, offered by traditional hyperscalers. The OaaS model will become the preferred choice of financial and operational managers, as it provides a simple answer to both these problems, offering a predictable billing model, guaranteed per unit of production and aligned with business needs.

The OaaS approach requires a great deal of research, industrialization and production work to meet all needs across all verticals.

Furthermore, the OaaS approach encourages sobriety, as all stakeholders benefit from optimization. If the plant is optimized, production costs come down and production criteria (lead times, etc.) are improved. Thus, the integration of issues related to energy savings, eco-responsibility, carbon footprint calculation, water, ozone as required by the European green taxonomy naturally becomes addressed.