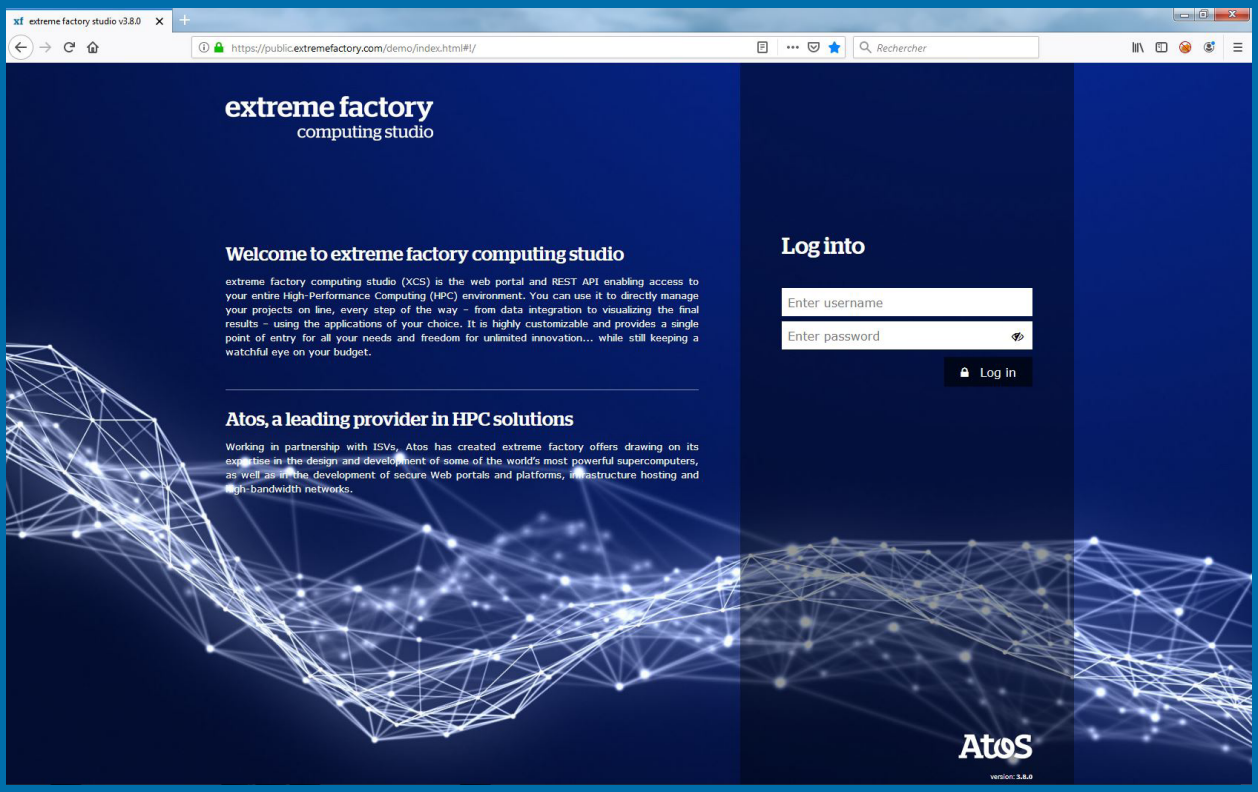
HPC & AI-as-a-service

The critical role of web portals

Trusted partner for your **Digital Journey**

Atos


extreme factory studio v3.8.0
https://public.extremefactory.com/demo/index.html#!/
Rechercher
extreme factory
computing studio
Welcome to extreme factory computing studio
extreme factory computing studio (XCS) is the web portal and REST API enabling access to your entire High-Performance Computing (HPC) environment. You can use it to directly manage your projects on line, every step of the way – from data integration to visualizing the final results – using the applications of your choice. It is highly customizable and provides a single point of entry for all your needs and freedom for unlimited innovation... while still keeping a watchful eye on your budget.
Atos, a leading provider in HPC solutions
Working in partnership with ISVs, Atos has created extreme factory offers drawing on its expertise in the design and development of some of the world's most powerful supercomputers, as well as in the development of secure Web portals and platforms, infrastructure hosting and high-bandwidth networks.
Log into
Enter username
Enter password
Log in
Atos
version: 3.8.0

Summary

Preface

With the democratization of High-Performance Computing (HPC) and the expansion of Artificial Intelligence (AI) in research and industries, organizations require more flexibility and more simplicity in the way they provide end-users with computing resources.

We are witnessing a progressive shift from a full-CAPEX (Capital Expenditure / investment) on-premises model to a more hybrid way of consuming computing power: through a mix of CAPEX and OPEX (Operational Expenditure / services). HPC & AI as-a-Service models are increasingly popular. Bull had sensed it very early by launching its Extreme Factory offers (private and ondemand) in 2010. Microsoft Azure, AWS, and Google Cloud have followed this shift and included HPC resources in their Cloud offerings, and they are becoming the references for AI-as-a-Service. There are also HPCaaS pure players such as Nimbix, and companies that provide web portals for HPC & AI-as-a-Service such as ActiveEon, Altair, Rescale, for example.

As more and more users have access to remote and heterogeneous resources, it becomes evident that not all of them have the skills to master this complexity. They have to deal with: energy-efficient datacenters, high-density hardware, parallel applications, security, and remote user experience. Also, the increasing variety and number of components in such systems can be a real challenge for IT administrators.

In this white paper, we will give you an overview of how HPC & AI web portals can help several stakeholders master the complexity of these hybrid computing resources. Indeed, HPC end-users and their IT management often underestimate how accessing, managing and using HPC & AI resources through a web portal could make their lives easier. Among other benefits, we will focus on:

- Flexibility
- Cost optimization
- Focus on work
- Security & Confidentiality
- Convenience & Usability

About the authors

Dr. Patrice Calegari is the Technical Domain Leader of GUIs for the HPC & AI software R&D at Atos. He is also the Product Manager of extreme factory computing studio (XCS), Atos HPCaaS web portal product. He joined Bull, now an Atos company, in 2005 as an HPC Application Expert to help start the HPC business at Bull. He was a key contributor to the development of Bull's Extreme Factory HPC cloud offer. Before joining Bull, he held various positions as HPC System software engineer and developer at CERN, Compaq and HP. Patrice holds a Magistère in Computer Science from ENS Lyon in France and obtained a Ph.D. in Computer science from EPFL in Switzerland.

Marc Levrier is the Technical Domain Leader of Hybrid HPC & AI software R&D at Atos, "hybrid" meaning both on-premises and Cloud targets. He has also been extreme factory's HPC-as-a-Service portfolio main architect for 10 years. He joined Serviware (French HPC leading system integrator) in 2002. That company was acquired by Bull, now an Atos company, in 2007. Before joining Serviware, Marc held several positions as scientific computing, visualization and system software developer / architect for various industries (Radiosurgery, Medical imaging as well as various real time and plant automation systems).

1. What is HPC & AI as-a-Service?

"as-a-Service" refers to any offer, solution, resource, or application being made available to a user as a web service. HPC & AI-as-a-Service solutions can include public Cloud as well as private Cloud usage, on-premises as well as hosted clusters, and rented as well as acquired hardware.

Although on-premises systems used to be the rule in the HPC world, there is a fast growing HPCaaS market as organizations tend to optimize their investments in computing resources. They now use a mix of on-premises and private or public Cloud resources, paving the way to a new hybrid computing era. HPCaaS is thus expected to grow at a CAGR of 11.9% between 2018 and 2023[1].

The Artificial Intelligence as-a-service market is soaring: organizations are punctually looking for more computing resources to train their models, exploit their data and thus burst to external resources. Driven by Machine Learning, Computer Vision and Natural Language Processing, AIaaS is expected to grow at a CAGR of 36% between 2018 and 2023[2].

HPCaaS & AIaaS have introduced more diversity in the resources that organizations have access to. This also means that end-users and IT administrators have to deal with more complexity. If all business cases came with their own user interfaces and system management tools, then HPC & AIaaS would certainly become counterproductive for these actors. Web portals help solving these issues in blurring underlying heterogeneities.

HPC & AI-as-a-Service = HPC & AI web portal + resources

2. What is an HPC & AI web portal?

First let's define what we mean by "HPC & AI web portal". It is a science gateway that provides access, management, and usage of High-Performance Computing & Artificial Intelligence resources. It usually includes web services (typically a web server exposed through a RESTful API), and a web User Interface to ease the services access (typically a web application that runs in a web browser).

HPC & AI web portal = HPC & AI web services + web UI

HPC & AI web portals are used with HPC & AI resources (hardware and middleware) to provide HPC & AI-as-a-Service solutions. HPC & AI web portals can be deployed as a front-end to OPEX and CAPEX infrastructures. There is no adherence with the financial model.

Several public and private Cloud providers have developed their own HPC & AI web portals to help their customers master their hybrid resources complexity. For instance, Atos has 2 Cloud offerings – extreme factory private and on-demand – that both come with the extreme factory computing studio (XCS) web portal, AWS and Nimbix also offer their own HPC web portal in their Cloud offers, respectively named EnginFrame and Jarvice. Many other actors provide web portals and gateways to ease HPC & AI as-a-Service usage: ActiveEon, Adaptive Computing, Altair, Apache AIRAVATA community, Fujitsu, IBM, Indiana University, Lenovo, Ohio Supercomputer Center, Rescale, RStor, Unicore project community, University of Colorado Boulder, to name a few.

If you want to know more about HPC web portals (history, concepts, technologies, etc.), we recommend to read the scientific article [1] published by ACM and that can be downloaded for free.

[1] https://www.reportlinker.com/p05727669/High-Performance-Computing-as-a-Service-Market-by-Verticals-Deployment-Type-Component-Region-Global-Forecast-to.html

[2] https://www.marketresearchfuture.com/reports/ai-as-a-service-market-7059

3. Who can benefit from an HPC & AI web portal?

Everybody, from HPC & AI end-users to the head of IT infrastructures, can benefit from solutions based on an HPC & AI web portal:

3.1 HPC & AI end-users' benefits

This Section lays down the benefits brought to end-users by HPC & AI portals.

1. Focus on work

An HPC & AI portal hides HPC cluster and AI platform complexity. It presents a user-friendly interface dedicated to the domain and/or the project of a user. This helps users stay focused on their important tasks (using their software applications) without worrying about HPC aspects, while taking advantage of it. This ability to avoid end-users to be disturbed by computer topics, while facilitating their scientific work, relates to portal usability and performance.

2. Time savings

An HPC portal promotes the remote execution of complete digital workflow (Computer Aided Design, pre-processing, simulation, and post-processing), keeping data close to compute power (and vice versa). This saves time otherwise wasted in data transfers. Remote job and data management as well as remote visualization features contribute to significantly accelerate Scientist or Engineer work.

3. Possible automation

The RESTful API allows to access remote HPC clusters by executing commands interactively or in a program. This gives the possibility to integrate remote computation work into a workflow controlled locally.

4. Customized workspaces

Each user has unique preferences and works within a standardized workflow that is often specific to a given scientific or technical domain. A highly customizable web user interface with application templates allows all users to benefit from rich and personalized experience without requiring any portal code change. A nice-to-have feature like RWD (Responsive Web Design) can help improve the customization when mobile devices or screens with vertical orientation need to be used.

5. Security

It is common knowledge that laptops or local workstations are particularly exposed to failures, theft and security threats. By working remotely in data centers, users can benefit from several security layers: the physical one that protects the data center itself, and the web services and HPC portal's integrated security features.

3.2 HPC & AI project managers' benefits

This Section lays down the benefits brought to project managers by HPC & AI portals.

1. Monitoring

An HPC & AI web portal can also be used to consolidate and show information about activities of project teams. This can be helpful and save time for project management by providing this information in real time together with resource usage. Statistics and instant status can be exported and in standard formats and shown in reports, table, graphs, dashboards, etc.

2. Accountability

When a web portal is used as the only entry point to HPC & AI resources, it allows tracking and logging all actions executed. This can be key in case of contractual or legal issues about some work done in a project.

3. Billing

Because of its monitoring and accountability features, a web portal can collect the solution's global activity (computations, data transfers, applications, etc.), Therefore, it allows to precisely control the budget spent for HPC & AI tasks by projects or users. In some use cases (e.g., collaborations, subcontracting, or partnership) it can also be used as the central point for billing or one entry point for third-party billing software.

4. Resource allocation

An HPC & AI web portal allows project managers to precisely adjust the resources allocated to each project according to their own priorities at a given time: hardware (types, quantities, etc.) and software (applications, licenses, etc.).

5. Resource limitation

An HPC & AI web portal allows project managers to pre-allocate credits to projects in order to limit the amount of resources that they are allowed to use.

3.3 HPC & AI cluster administrators' benefits

This Section lays down the benefits brought to cluster administrators by HPC & AI portals.

3.3.1 Cluster administrators challenges

The 5 main challenges for HPC IT administrators are:

1. Control

The challenge is to securely control the access to the resources, taking into account the specific rules that define the access rights depending on user roles in the organization.

2. Quality of service

The challenge is to provide efficient HPC services through an ergonomic, efficient and secured web UI that gives access to user applications, data and all related information from any location at any time.

3. Costs

The challenge is to reduce global costs of resources that are still too often distributed in workstations (e.g. compute power, GPUs, large memory for pre-post processing, licenses) or as compute clusters on several locations.

4. Complexity

The challenge is to simplify maintenance and end-user support for the many applications, each with dozens of versions, that are usually installed on HPC clusters.

5. Confidentiality

The challenge is to protect data from forbidden download and copy attempts. Usually, extremely sensitive data with high intellectual property value is stored on HPC clusters.

3.3.2 Cluster administrators' solutions

An HPC & AI portal can answer the 5 IT administrator challenges listed in previous Section as explained below.

1. Resource access control

An HPC portal is a secured and flexible tool that controls access to the resources. It helps better manage and consolidate compute and data usage. Resource access is controlled by roles that can be defined at a fine-grained level with scopes assigned to privileges, in accordance with the RBAC permission model (Role-Based Access Control).

2. Provision of high level web services to users

Thanks to HPC portals, IT administrators can provide high level web services to their users who can enjoy all the benefits listed in Section 3.1. Moreover, an HPC portal that supports multi-clusters, multi-schedulers, and multi-directory services allows the administrators to change each or all these services transparently for the end-users who can keep using the very same web UI, even if major changes are done on the HPC cluster(s)' side. This results in a stable user experience.

3. Global cost minimization

Global cost is reduced by facilitating the centralization of all resources (applications, middleware, licenses and license servers, Data, servers, GPUs, etc.). All resources can be installed in a unique centralized location, resulting in a global cost much lower than the sum of the costs of equivalent distributed resources. Moreover, a frequent side effect of the seamless access to the HPC resources is a quicker adoption of HPC computation by new users, and an increased number of submitted jobs by regular users, resulting in a higher cluster load rate. All together these elements contribute to raising the Return On Investment (ROI) of the HPC resources.

The use of a portal is one enabler to transform CAPEX costs into incomes by using a new business model for reselling compute power on-demand to new users external to the organization who invested in the resources. The infrastructure cost can then be optimized thanks to the use of a portal.

4. Low end-user training and application maintenance

Application templates of HPC portals provide a unified job submission method and a unified scientific application management model. They comply with the same publication framework, which dramatically decreases the amount of end-user training and HPC application environment maintenance efforts. The maintenance of application software and application license can both be made simpler from a portal by the use of unified templates, methods, and views.

5. Intellectual Property (IP) protection

While IT security officers are usually reluctant to provide their users with SSH (Secure SHell), most of them are in favor of HTTPS and related protocols that are easier to monitor and secure through firewalls. Combined with role-based access control (RBAC) approach, remote visualization facilitates companies' IP protection by restricting the access to data: end-users can interact with their graphical software and 3D data via 3D streaming technologies, but they may be deprived of the right to download the highly valuable original data sets (configurable policy).

4 Benefits of accessing HPC & AI Cloud resources through a web portal

A web portal can be used to abstract the notion of HPC & AI resources. HPC clusters, supercomputers, and Deep Learning dedicated machines can be located anywhere (on-premises, on a remote data center, or in the public Cloud). The actual location of these resources is easily made transparent by portals for the end-users.

In the specific case of Cloud resource usage, the intrinsic abstraction qualities of HPC & AI web portals simplify the user experience. They help to benefit more easily from Cloud advantages such as those listed below:

1. Hardware diversity

A compute center cannot include all type of hardware in one place to cover all possible needs of its users. The requirements of the users can be very diverse, including for example:

- various references of CPU, GPU, FPGA server products
- various types of storage hardware technologies and file systems,
- legacy and latest technology material,
- small and large memory servers,
- vendor specific machines,
- interconnect specific technologies.

Accessing several Cloud platforms, it is nowadays possible to fulfill all these requirements from one single point of access: an HPC & AI web portal managed by the IT admins and configured to access the best fitted Cloud platforms.

2. Flexibility

The different requirements listed above can, and usually do, change over time. The access to the resources should then be as flexible as possible to quickly adapt to users' needs. Using Cloud resources allows selecting any type of hardware and changing it at any time. Moreover, it allows to increase and decrease the amount of these resources for given periods according to the needs.

3. Cost

Project managers can translate CAPEX costs (investments) into OPEX costs (services) and reduce their global cost to their exact needs without dealing with large investment costs that are diffcult to amortize during the life of a project. The global cost reduction is usually higher than the costs induced by Cloud services.

For instance, Cloud usage is necessary when on-premises HPC & AI infrastructures are not sufficient or do not exist. It gives an opportunity to exploit large-scale infrastructures, not affordable otherwise. For example, Cloud eases the use of HPC in the following cases:

- when R&D teams need to test new software,
- when R&D teams need to test software on different hardware,
- when performance needs to be validated on a specific hardware (i.e., benchmark),
- when unplanned/unexpected/ unpredictable computation loads need to be run quickly (i.e., burst).

And since Cloud access can be better controlled and managed through a unique HPC portal that centralizes all authentication and authorization mechanisms, the association of both of them (Cloud and portal) offers great and almost infinite solutions. HPC & AI resources (hardware and middleware) are often not 100% in the Cloud: computations are partially executed on-premises and partially in the Cloud. That is what we call HPC & AI hybrid computing.

HPC & AI web portal + HPC & AI hybrid computing resources = Your Best HPC & AI Solution

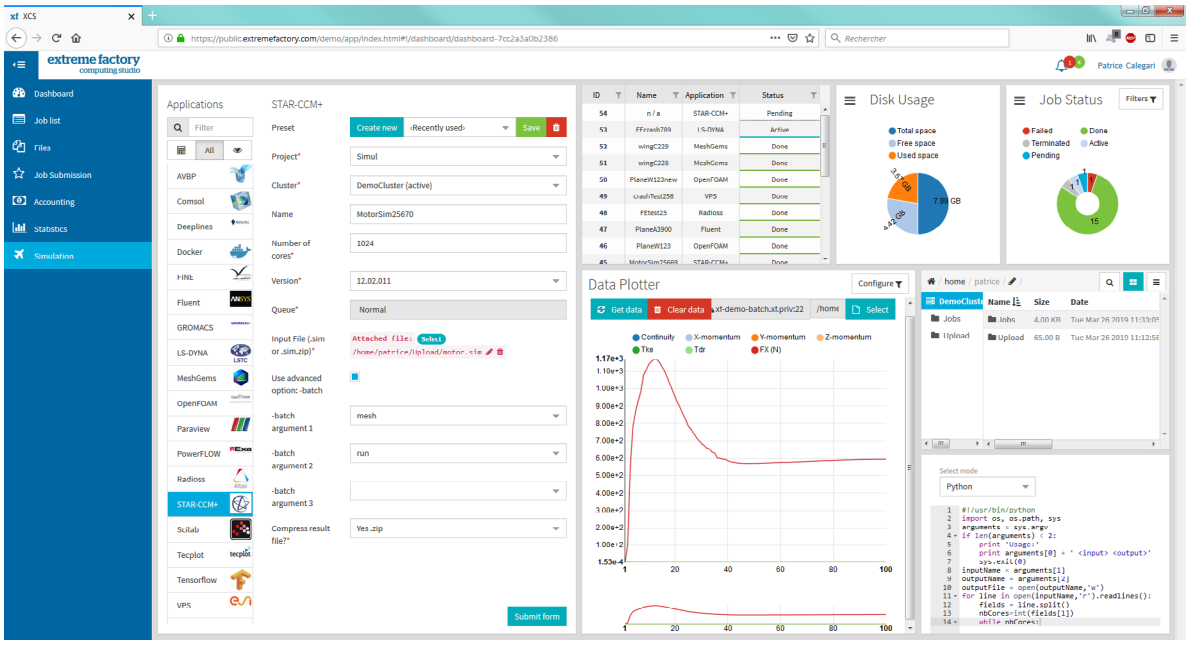
Conclusion

HPC & AI web portals can provide many benefits to both end-users and IT administrators. They are a key enabler to bridge the gap between these various scientific domains and workloads on-premises, on the Cloud or at other service providers' (such as regional/national HPC resource providers). By providing safe and easy access to complex applications and resources, they also make a large contribution to refocusing Scientists and Digital Simulation Engineers on their real job and IT cost optimization.

Atos has long-standing experience of HPC portal software development, acquired in providing end-to-end HPC solutions to various clients in diverse market segments. XCS (extreme factory computing studio), the Atos web portal product, has become a critical asset for Atos customers: it is a flexible, non-intrusive and fully customizable with editable dashboards. Its aim is to federate different kinds of scientific workloads under a single and homogeneous user experience. XCS technologies have recently been extended to Codex AI Suite, Quantum Learning Machine (QLM), Supercomputing Suite Power Efficiency portal, and many more to come.

Atos HPC & AI portals are part of a broader Hybrid computing portfolio. As such, Atos' extreme factory computing studio (XCS) has also been designed to help organizations easily and centrally manage their hybrid compute and visualization workloads as-a-Service.

Get more information from: https://atos.net/en/products/high-performance-computing-hpc/bull-extreme-factory



Reference

[1] Patrice Calegari, Marc Levrier, Paweł Balczyński. "Web Portals for High-performance Computing: A Survey." ACM Transactions on the Web (TWEB), volume 13, Issue 1, Article 5 (February 2019), 36 pages. https://doi.org/10.1145/3197385

White Paper

About Atos

Atos is a global leader in digital transformation with over 110,000 employees in 73 countries and annual revenue of over € 11 billion.

European number one in Cloud, Cybersecurity and High-Performance Computing, the Group provides end-to-end Orchestrated Hybrid Cloud, Big Data, Business Applications and Digital Workplace solutions. The group is the Worldwide Information Technology Partner for the Olympic & Paralympic Games and operates under the brands Atos, Atos Syntel, and Unify. Atos is a SE (Societas Europaea), listed on the CAC40 Paris stock index.

The purpose of Atos is to help design the future of the information technology space. Its expertise and services support the development of knowledge, education as well as multicultural and pluralistic approaches to research that contribute to scientific and technological excellence. Across the world, the group enables its customers, employees and collaborators, and members of societies at large to live, work and develop sustainably and confidently in the information technology space.

Find out more about us
atos.net/en/products/high-performance-computinghpc/bull-extreme-factory

Let's start a discussion together

CT_I2369_190522_RY_WP_HPC&AI