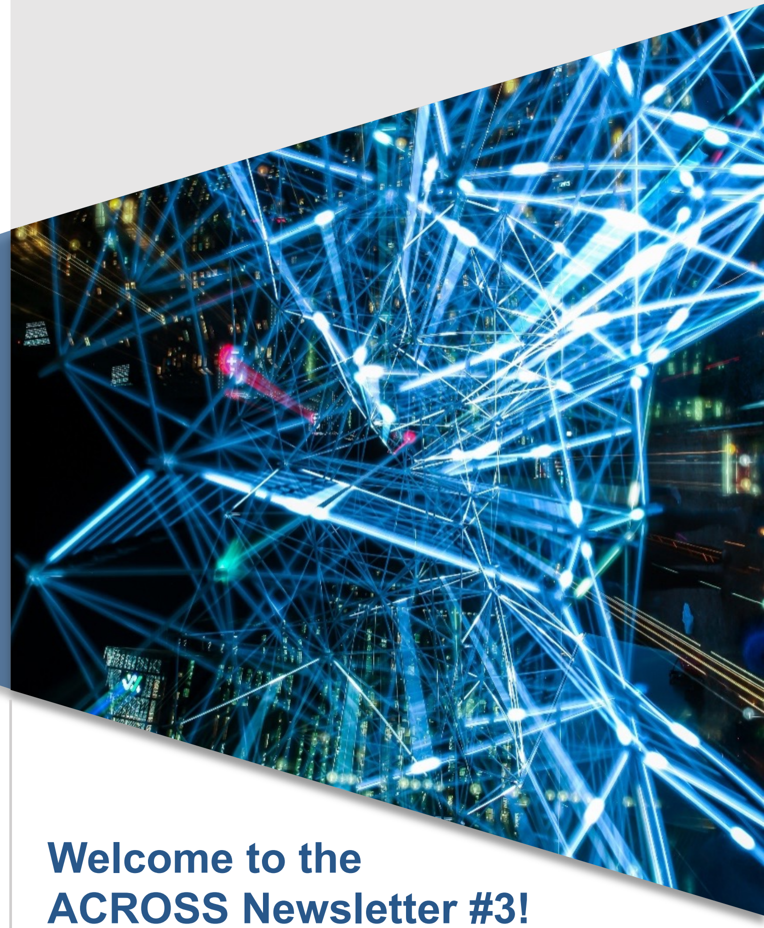
NEWSLETTER #3

NOVEMBER 2022

ACROSS

HPC Big DAta ArtifiCial intelligence cross Stack PlatfoRm TOwardS ExaScale



INSIDE THIS ISSUE

Welcome to the ACROSS Newsletter #3!

The third edition of the ACROSS Newsletter reflects the Milestone 3 of the ACROSS project. It describes the Alpha version of the ACROSS System and technologies.

This newsletter intends to depict the overall ACROSS architecture, including architecture principles and the general architecture view. It also describes the hardware and software technologies that compose ACROSS System.

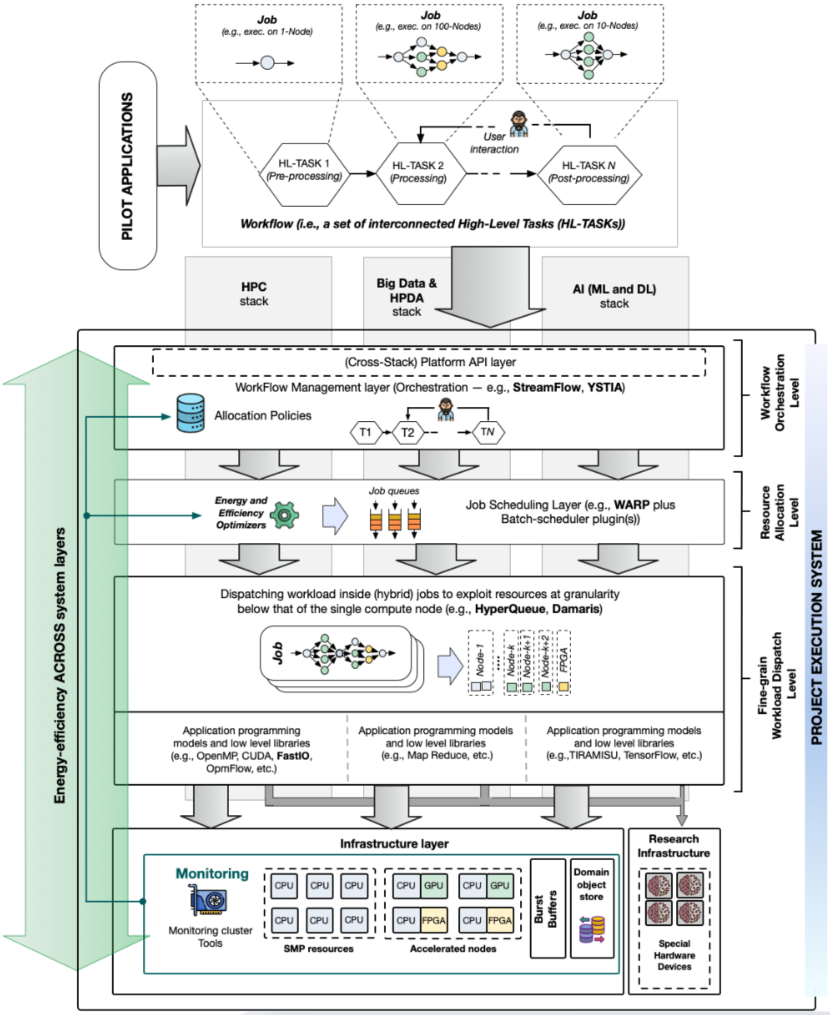
Want to know more? Keep up to date on all project news @across_project and sign up for the Newsletter so you never miss out.

Questions? Contact info@acrossproject.eu

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955648. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, France, Czech Republic, United Kingdom, Greece, Netherlands, Germany, Norway.

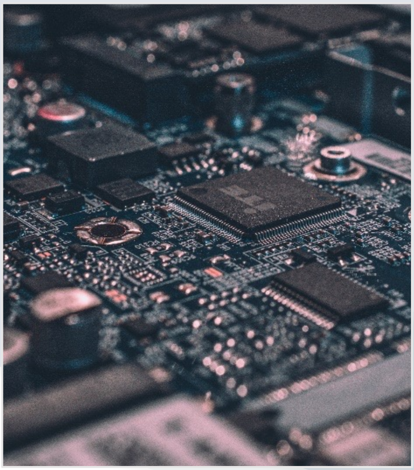
ACROSS System Overview (Alpha version)

The ACROSS architecture contains two main pillars. The bottom layer of the architecture called the infrastructure layer relies on heterogeneous hardware that enables support of hardware acceleration. The second layer also called the system layer is built to support a multi-level orchestration while acting as the workflow orchestration layer. The interconnection between the infrastructure layer and orchestration layer acts as core of the entire ACROSS system that will provide access to heterogeneous hardware and advanced orchestration services. Heterogeneous workflows are described at the user level through the Common Workflow Language (CWL) standard augmented with the information of resources required by each step to execute. Each step, can be mapped on one or multiple HPC jobs. To better exploit infrastructural resources (which are heterogeneous in nature), the orchestrator relies on a smart workload scheduling based on the HyperQueue solution.



ACROSS PARTNERS

This newsletter has been prepared by Bull an Atos Technology. Atos is a global leader in digital transformation. European number one in cybersecurity, cloud and high-performance computing, the Group provides tailored end-to-end solutions for all industries in 71 countries. With more than 80 years of technology innovation expertise, Bull is the hub of technology development at the heart of Atos Business Technologist family.



LEARN MORE

Key Infrastructure Technologies

The co-design activities within the project has envisioned a hardware architecture based on multi-core nodes enriched with a wide range of accelerators varying from general-purpose Graphics Processing Units (GPUs), Field Programmable Gate Arrays (FPGAs) to more specific AI-acceleration devices like Neural Network Processors (NNPs), Visualization Processing Units (VPUs) and even neuromorphic emulators/simulators, while emphasizing the communication between these computing devices as well as their potential for further extension.

The hardware will be abstracted by layers of software, which will have to be combined judiciously by the orchestrator for an optimal use of these computing technologies, according to their availability and their configuration.

The infrastructure technologies and platforms described below are provided by ATOS/BULL, CINECA and IT4I.

Galileo 100 cluster

The infrastructure GALILEO100 provided by CINECA, is co-funded by the European ICEI (Interactive Computing e-Infrastructure). It's a very powerful infrastructure that provides:

- 554 computing nodes each equipped with 2 CPU Intel CascadeLake 8260, sporting 24 cores each and running at 2.4 GHz, 384GB RAM, subdivided in:
  - 340 standard nodes ("thin nodes") with 480 GB of SSD storage
  - 180 data processing nodes ("fat nodes") with 2TB of SSD storage and 3TB of Intel Optane memory
- 34 GPU nodes with 2x NVIDIA GPU V100 with 100Gbs Infiniband interconnection and 2TB SSD.
- 77 nodes for cloud computing (ADA CLOUD based on OpenStack), 2x CPU 8260 Intel CascadeLake, 24 cores running at 2.4 GHz, 768 GB RAM, with 100Gbs Ethernet interconnection.
- 20 PB of active storage accessible from both Cloud and HPC nodes.
- 5 PB of fast storage for HPC system.
- 1 PB Ceph storage for Cloud (full NVMe/SSD)
- 720 TB fast storage (IME DDN solution)

At the present time, two pilots have run their numerical simulations on CINECA Galileo100 and IT4I Karolina clusters, achieving after 18 months the relevant figure of circa 5 Million-core hours sponsored by PRACE-ICEI calls.

Key Infrastructure Technologies

Leonardo supercomputer

CINECA will host this pre-Exascale class supercomputer, which has been funded by the European Commission in the context of the EuroHPC-JU, and located in Bologna (Italy).

This machine is based on the Atos BullSequana XH2000 architecture, and equipped with nearly 14,000 next generation NVIDIA Ampere architecture-based GPUs, NVIDIA Mellanox HDR InfiniBand, and over 100 petabytes of state-of-the-art storage capacity, which will provide 10 Exaflops (10 EFlop/s) of FP16 AI performance.

This infrastructure will be capable of an aggregated HPL performance of 250 PFlop/s (HPL Linpack Performance (Rmax)), enabling the researchers and scientists to make new discoveries, and contribute to the management and mitigation of critical situations due to extreme events.

The main features of Leonardo supercomputer are:

- 3 Modules: more than 136 BullSequana XH2000 Direct Liquid cooling racks
- 5000 computing nodes:
    - 3456 servers equipped with Intel Xeon Ice Lake and NVIDIA Ampere architecture GPUs
    - 1536 servers with Intel Xeon Sapphire processors
- 3+PB RAM
- 5PB of High-Performance storage
- 100PB of Large Capacity Storage
- 1TB/s bandwidth
- 200Gb/s interconnection bandwidth
- 9MW
- PUE 1,08
- 1500+ m2 footprint

Leonardo system resources will be available as part of the infrastructural support for the pilot workflows execution, when the system will be online.

Key Infrastructure Technologies

Barbora cluster

IT4I provides Barbora cluster, which provides a theoretical peak performance of 849 TFlop/s.

The computing system consists of:

- 189 standard computational nodes; each node is equipped with two 18-core Intel processors and 192 GB RAM.
- 8 compute nodes with GPU accelerators; each node is equipped with two 12-core Intel processors, four NVIDIA Tesla V100 GPU accelerators with 16 GB of HBM2 and 192 GB of RAM.
- 1 fat node is equipped with eight 16-core Intel processors and 6 TB RAM.

The supercomputer is built on the Bull Sequana X architecture and for cooling its standard compute nodes the direct liquid cooling technology is used.

The computing network is built on the latest Infiniband HDR technology.

The SCRATCH computing data storage capacity is 310 TB with 28 GB/s throughput using Burst Buffer acceleration. Another computing data storage is based on NVMe over Fabric with a total capacity of 22.4 TB dynamically allocated to compute nodes.

It is also equipped with the ATOS/BULL SuperComputer Suite cluster operation and management software solution as well as PBS PRO scheduler and resource manager.

NVIDIA DGX-2

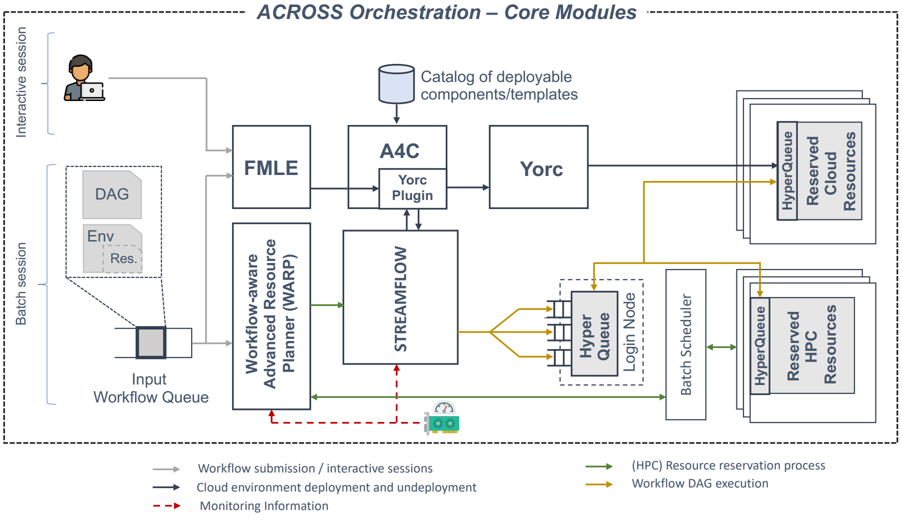
NVIDIA DGX-2, was installed in the spring 2019 with 2 PFlop/s peak performance in AI. It is equipped with 16 powerful data centre accelerators – the NVIDIA Tesla V100 GPU. They are interconnected with revolutionary NVSwitch 2.0 technology that delivers a total bandwidth of 2.4 TB/s.

The systems include 512 GB of HBM2 memory.

The NVIDIA DGX-2 also offers 30 TB of internal capacity on fast NVMe SSDs disks. Interconnection to the surrounding infrastructure is provided via eight 100 Gb/s Infiniband/Ethernet adapters.

One NVIDIA DGX-2 can replace 300 dual-socket servers with Intel Xeon Gold processors for deep neural network training (e.g ResNet-50). NVIDIA DGX-2 is powered by the DGX software stack; NVIDIA-optimized and tuned AI software that runs the most popular machine learning and deep learning frameworks with maximized performance. The NVIDIA DGX-2 can also be used for traditional HPC workloads to deliver a theoretical peak performance of 130 TFlop/s.

Key Software Technologies
Applications Orchestration



Orchestration tools at a glance

ACROSS orchestration toolbox is built on top of a set of technological solutions that allow us to effectively execute workflows mixing numerical simulations, HPDA operations and ML/DL tasks. The whole solution exploit the capability of StreamFlow to parse and execute CWL-based workflows, WARP to deterministically allocate HPC resources, HyperQueue to distribute the workload on reserved resources (also allowing for job co-location), FMLE/YSTIA to address the spinning up of Cloud-based resources as well as to ease the training of ML/DL models.

- **StreamFlow.** StreamFlow is a container-native Workflow Management System based on the CWL standard, and designed for scheduling and coordinating different workflow steps on top of a diverse set of execution environments, taking care of worker nodes' life cycle, data transfers, and fault-tolerance aspects.

- **WARP.** It allows reserving HPC resources (in advance) by applying advanced and workflow-aware planning, which is based on the specific resources needed by each workflow to be executed. It exploits the capability of a batch-scheduler plugin to actually perform the reservation process.

- **HyperQueue.** Part of the HyperTools set developed by IT4Innovations, HyperQueue is a scheduler that transparently schedule tasks through HPC system schedulers like PBS or SLURM.

- **FMLE/YSTIA.** FMLE is a ML/DL toolbox for HPC which hides complexity of HPC jobs management for AI model management. YSTIA is a TOSCA based orchestrator which is exploited by FMLE to manage operations on AI models.

Key Software Technologies

Damaris

In support of Carbon Sequestration pilot, Inria is extending the capabilities of the Damaris asynchronous I/O and visualization library. There is now a newly available Damaris plugin that supports asynchronous, in-situ processing using Python and Dask. This opens up a large number of analytic methods to the pilot simulation via the use of Dask data types, which includes methods from libraries SciKit-Learn and Tensorflow/Keras.

Damaris

GPU accelation

ACROSS also targets to improve GPU acceleration of the Open Porous Media reservoir simulator(OPM Flow). A reservoir simulator solves sets of nonlinear partial differential equations, and an important part of this is solving large, sparse linear systems using iterative Krylov methods such as BiCGStab or GMRES.

OPM Flow uses the DUNE Iterative Solver Template Library (ISTL) to solve linear systems of equations through so-called iterative Krylov methods. These methods typically involve a series of linear algebra operations such as inner products and matrix-vector-multiplications. The implementation of the Krylov solvers in DUNE ISTL is decoupled from the underlying implementation of the linear algebra operations. In the ACROSS project, we took advantage of this decoupling by replacing the standard matrix and vector classes with implementations that use the GPU through the cuSPARSE library. This led to no change in the code of the linear solvers, and minimal change in the overall simulation code but introduced both GPU and multi-GPU paths to the solving of the linear systems. These Krylov solvers were then combined with preconditioners from cuSPARSE.

The software is available as open source under the GNU General Public License, version 3 or later.

Visit our website www.across-project.eu to see all our latest developments.