DEEP-EST: A Modular Supercomputer for HPC and High Performance Data Analytics

How does one cover the needs of both HPC and HPDA (high performance data analytics) applications? Which hardware and software technologies are needed? And how should these technologies be combined so that very different kinds of applications are able to efficiently exploit them? These are the questions that the recently started EU-funded project DEEP-EST addresses with the Modular Supercomputing architecture.

Scientists and engineers run large simulations on supercomputers to describe and understand problems too complex to be reproduced experimentally. The codes that they use for this purpose, the kind of data they generate and analyse, and the algorithms they employ are very diverse. As a consequence, some applications run better (faster, more cost- and more energy-efficient) on certain supercomputers and some run better on others.

The better the hardware fits the applications (and vice-versa), the more results can be achieved in the lifetime of a supercomputer. But finding the best match between hardware technology and the application portfolio of HPC centres is getting harder. Computational science and engineering keep advancing and increasingly address ever-more complex problems. To solve these problems, research teams frequently combine multiple algorithms, or even completely

Fig. 1: DEEP-EST collaboration at the kick-off meeting in Jülich, July 13th.

different codes, that reproduce different aspects of the given topic. Furthermore, new user communities of HPC systems are emerging, bringing new requirements. This is the case for large-scale data analytics or big data applications: They require huge amounts of computing power to process the data deluge they are dealing with. Both complex HPC workflows and HPDA applications increase the variety of requirements that need to be properly addressed by a supercomputer centre when choosing its production systems. These challenges add to additional constraints related to the total cost of the machine, its power consumption, the maintenance and operational efforts, and the programmability of the system.

The modular supercomputing architecture

Creating a modular supercomputer that best fits the requirements of these diverse, increasingly complex, and newly emerging applications is the aim of DEEP-EST, an EU project launched on July 1, 2017 (see Fig. 1). It is the third member of the DEEP Projects family, and builds upon the results of its predecessors DEEP[1] and DEEP-ER[2], which ran from December 2011 to March 2017.

DEEP and DEEP-ER established the Cluster-Booster concept, which is the first incarnation of a more general idea to be realised in DEEP-EST: the **Modular Supercomputing Architecture**. This innovative architecture creates a unique HPC system by coupling various compute modules according to the building-block principle. Each module is tailored to the needs of a specific group of applications, and all modules together behave as a single machine. This is guaranteed by connecting them through a high-speed network and, most importantly, operating them with a uniform system software and programming environment. In this way, one application can be distributed over several modules, running each part of its code onto the best suited hardware.

The hardware prototype

The DEEP-EST prototype (see Fig. 2) to be installed in summer 2019, will contain the following main components:

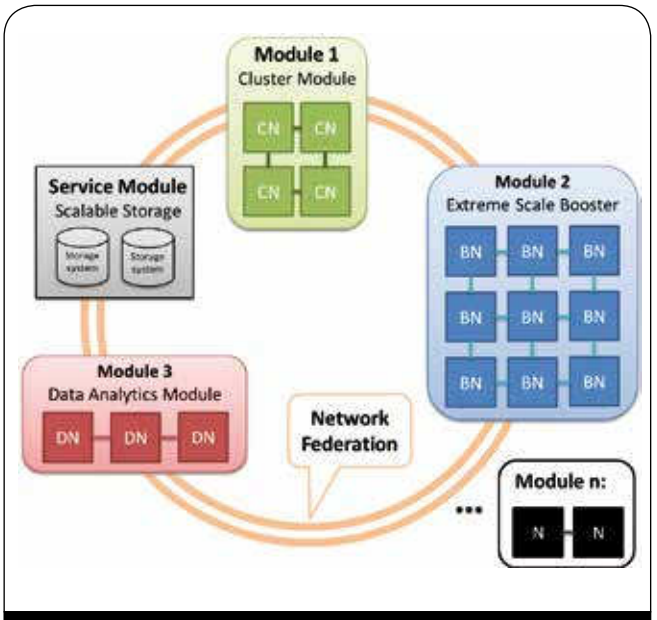


Fig. 2: Modular Supercomputing Architecture as implemented in DEEP-EST. (CN: Cluster Node; BN: Booster Node; DN: Data Analytics Node). Each compute module addresses the requirements of specific parts of or kinds of applications, and all together they behave as a single machine. Extensions with further modules (n) can be done at any time.

- Cluster Module: to run codes (or parts of them) requiring high single-thread performance
- Extreme Scale Booster: for the highly-scalable parts of the applications
- Data Analytics Module: supporting HPDA requirements

The three mentioned compute modules will be connected with each other through a "Network Federation" to efficiently bridge between the (potentially different) network technologies of the various modules. Attached to the "Network Federation," two innovative memory technologies will be included:

- Network Attached Memory: providing a large-size memory pool globally accessible to all nodes
- Global Collective Engine: a processing element at the network to accelerate MPI collective operations

In addition to the three abovementioned compute modules, a service module will provide the prototype with the required scalable storage.

One important aspect to be considered in the design and construction of the DEEP-EST prototype is energy efficiency. It will influence the choice of the specific components and how they are integrated and cooled. An advanced monitoring infrastructure will be included to precisely quantify the power consumption of the most important components of the machine, and modelling tools will be applied to predict the consumption of a large scale system built under the same principles.

The software stack

The DEEP-EST system software, and in particular its specially adapted resource manager and scheduler, enable running concurrently a mix of diverse applications, best exploiting the resources of a modular supercomputer. In a way, the scheduler and resource manager act similar to a Tetris player, arranging the differently shaped codes onto the hardware so, that no holes (i.e. empty/idle resources) are left between them (see Fig. 3). When an application

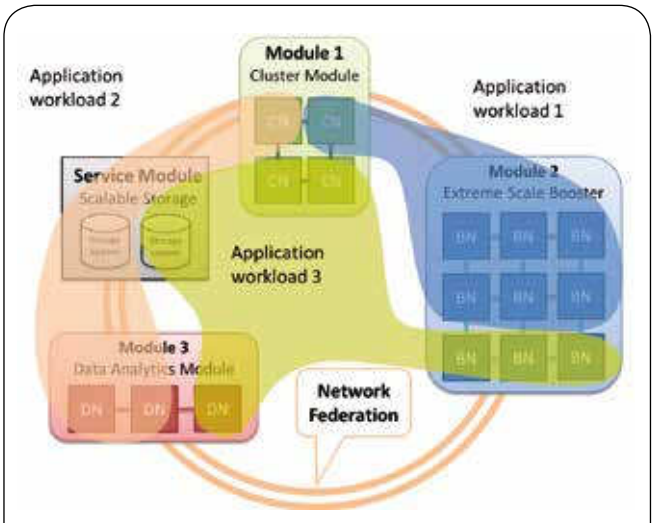
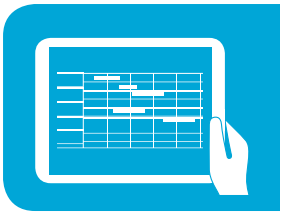


Fig. 3: Three example applications running on a Modular Supercomputer, distributed according to their needs. In this example, workload 1 would be a typical HPC code, workload 2 a typical HPDA application, and workload 3 a code combining both fields.



finishes using some nodes, these are immediately freed and assigned to others. This reservation and release of resources can be done also dynamically, what is particularly interesting when the workloads have different kinds of resource requirements along their runtime.

In DEEP-EST, the particularities and complexity of the underlying hardware are hidden from the users, which face the same kind of programming environment (based on MPI and OpenMP) that exists in most HPC systems. The key components of the programming model used in DEEP-EST have been in fact developed already DEEP. Employing ParaStation MPI and the programming model OmpSs, users mark the parts of the applications to run on each compute module and let the runtime take care of the code-offload and data communication between modules. Further resiliency capabilities were later developed in DEEP-ER. In DEEP-EST, ParaStation MPI and OmpSs will be, when needed, adapted to support the newly introduced Data Analytics Module and combined with the programming tools required by HPDA codes.

The DEEP-EST software stack is completed with compilers, the file system software (BeeGFS), I/O libraries (SIONlib), and tools for application performance analysis (Extrae/Paraver), benchmarking (JUBE) and modelling (Dimemas).

Co-design applications

The full DEEP-EST system (both its hardware and software components) is developed in co-design with a group of six scientific applications from diverse fields. They come from neuroscience, molecular dynamics, radio astronomy, space weather, earth sciences and high-energy physics. The codes have been chosen to cover a wide spectrum of application fields with significantly different needs, and include traditional HPC codes (e.g. GROMACS), HPDA applications (e.g. HPDBSCAN), and very data intensive codes (e.g. the SKA and the CMS data analysis pipelines).

The requirements of all of these codes will shape the design of the hardware modules and their software stack. Once the prototype is installed and the software is in operation, the application codes will run on the platform, demonstrating the advantages that the Modular Supercomputing Architecture provides to real scientific codes.

Project numbers and GCS contribution

The DEEP-EST project will run for three years, from July 2017 to June 2020. It was selected under call FETHPC-01-2016 ("Co-design of HPC systems and applications") and receives a total EU funding of almost €15 million from the H2020 program. The consortium, led by JSC, includes LRZ within its 16 partners comprising computing centres, research institutions, industrial companies, and universities.

LRZ leads the energy efficiency tasks and the public relations and dissemination activities. It also chairs the project's Innovation Council (IC): a management body responsible to identify innovation opportunities outside the project.

Beyond the management and coordination of the project, JSC leads the application work package and the user-support activities. It will also contribute to benchmarking, and I/O tasks. Furthermore, in collaboration with partners Barcelona Supercomputing Centre and Intel, JSC will adapt the SLURM scheduler to the needs of a modular supercomputer. Last but not least, JSC drives the overall technical definition of the hardware and software designs in the DEEP-EST project as the leader of the Design and Development Group (DDG).

Acknowledgements

The research leading to these results has received funding from the European Community's Horizon 2020 (H2020) Funding Programme under Grant Agreement n° 754304 (Project "DEEP-EST").

References

[1] Suarez, E., Eicker, N., Gürich, W.:
"Dynamical Exascale Entry Platform: the DEEP Project", inSiDE Vol. 9 No.2, Autumn 2011, http://inside.hlrs.de/htm/Edition_02_11/article_12.html

[2] Suarez, E. and Eicker, N:
"Going DEEP-ER to Exascale", inSiDE Vol. 9 No.2, Spring 2014, http://inside.hlrs.de/htm/Edition_02_11/article_12.html

[3] www.deep-projects.eu

Written by Estela Suarez
Jülich Supercomputing Centre (JSC)
Contact: e.suarez@fz-juelich.de