WHITE PAPER

Intel®, Jülich Supercomputing Centre (JSC), Megware

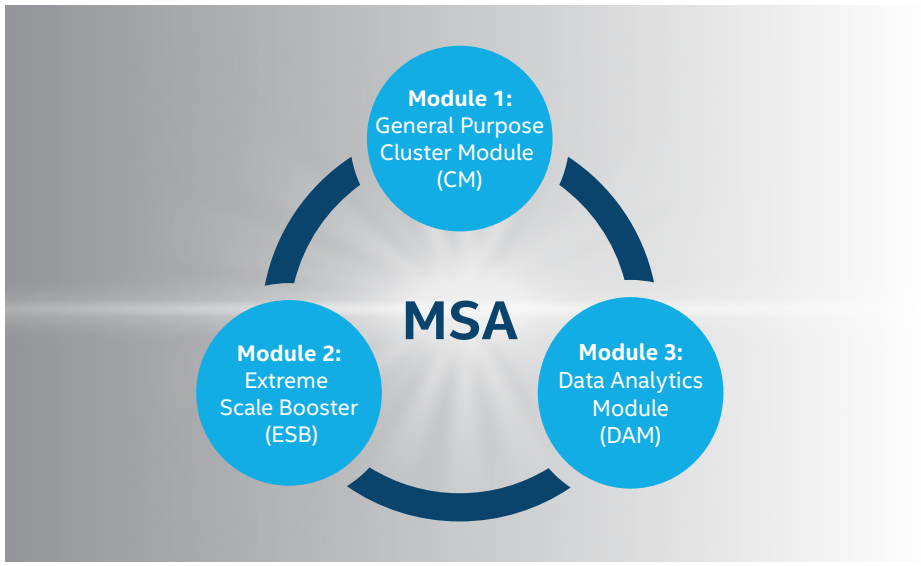
A blueprint for success

Co-design approach for the Modular Supercomputing Architecture (MSA)

Research Institutes like the Jülich Supercomputing Centre (JSC), one of the leading high performance computing (HPC) research centers in Europe, are redefining the future of supercomputing in an European Union (EU)-funded project. The HPC solution which is being developed currently, is part of the EU-funded, collaborative R&D project DEEP-EST (Dynamical Exascale Entry Platform – Extreme Scale Technologies). With a budget of 15 Million Euro (17.6 USD), DEEP-EST develops and builds a fully functional energy-efficient prototype of the Modular Supercomputing Architecture (MSA).

The Modular Supercomputing Architecture (MSA)

The DEEP-EST MSA system consists of three processing modules: The General-Purpose Cluster Module (CM), the Extreme Scale Booster (ESB) Module as well as the Data Analytics Module (DAM). All modules make use of Intel® Xeon® scalable processors. The DAM is utilizing the new Intel® Optane™ technology as a DRAM alternative and as an ultra-fast SSD. Additionally, the DAM employs specialized graphics adapters (GPU) and Intel® FPGA Programmable Accelerator Cards (PACs), which can be programmed individually per application. All components were combined in an efficient way by the HPC specialized Intel partner Megware.



In a nutshell, the main advantage of the DAM is providing very large amounts of memory and processing power, targeted exactly at the matching parts of the relevant applications. In particular, applications in powerful data analytics (high performance

data analytics: HPDA) and machine learning (ML) utilize large memory arrays which, contrasting traditional HPC applications, are addressed in a random, non-linear way. The modular approach, which connects all components by a network federation, makes it unnecessary to fit all nodes with a lot of RAM and acceleration modules like GPUs and FPGAs. This way, both power consumption and costs are reduced. While a traditional monolithic architecture processes all possible application scenarios on the same target hardware, a modular supercomputer will use individually assigned nodes and modules with specific skills – exactly like the DAM for HPDA- and ML-workloads.

Requirements analysis using a comprehensive co-design approach

The DEEP-EST project started mid-2017 with comprehensive requirements analysis and technology reviews. As of today, both the Cluster Module as well as the Data Analytics Module have been finalized and put into operation. Before mid-2020, the Extreme Scale Booster will follow.

Research and development in the field of high-performance computing is teamwork. DEEP-EST combines 16 partners within the EU under the project lead of the Jülich Supercomputing Centre. Among them are six scientific institutes which cover the application part of the project as well as, next to Intel, supercomputer expert Megware, who is building the actual hardware modules. Among the additional partners are well-known names like Astron, the CERN in Geneva, the University of Edinburgh, the Barcelona Supercomputing Center (BSC) and the Leibniz-Rechenzentrum (LRZ).

Solid foundation for computational success: General Cluster Module

The Cluster Module provides the base for the whole Modular Supercomputer Architecture. When designing the module, developers intended providing reliable computational power for as diverse workloads as possible. Each application must be able to achieve sufficiently good results without needing to be specially adapted or optimized. Maximum per-thread performance was especially important for floating point- and integer calculations because they are heavily used by all applications, no matter how well they can be parallelized.

Another requirement for the Cluster Module was that it can work with all commonly used software development environments and stacks in the HPC field. While HPC usually focusses on a limited amount of programing models and APIs like MPI, OpenMP, OpenCL and CUDA, data analysis applications now rely on a wide variety of programming languages and frameworks. Those range from Python over R, Caffe, Tensorflow Keras and Theano up to Apache Spark. Even though the Data Analytics Module will bear the brunt of the analysis tasks, the developers expect that the HPDA code will be developed on the Cluster Module.

The current specification for the 50 CM nodes utilizes two Intel Xeon Gold 6146 scalable processors containing 12 cores each, operating at 3.2GHz with 24.75MB L3-cache. 192 GB DDR4 RAM together with Intel® NVMe PCIe3 SSDs offer more than enough memory for all processing tasks.

Each cluster node can write to the connected SSDs via four PCIe Generation 3 Links, each delivering about 1.0 GByte/s throughput. This way, applications spread over several nodes can profit from almost linear speed gains for I/O operations. In the prototype system, Megware realized ten nodes each in a six-HU chassis. Considering the space for cooling, power distribution and the InfiniBand switches, up to 50 nodes will fit into a fully populated, efficiently warmwater-cooled rack.

A specialist for machine learning: The Data Analytics Module (DAM)

The DAM's main task is to perform data analysis tasks as fast as possible, making it ideally suitable for machine learning, Artificial Intelligence and deep learning workloads. Even though it is a specialized module, the designers mainly used off-the-shelf products like graphics- and FPGA cards without costly custom developed components. Intel Optane technology is an essential component – PCIe-attached SSDs provide fast mass storage, and each node has up to 3 TB of Intel Optane persistent memory (PMem) in addition to 384GB DDR4 RAM. Intel Optane PMem can be used in two modes: as persistent, fast, byte-addressable memory (AppDirect) or as a volatile & transparent extension of the installed DRAM, providing applications with up to 3 TB of random access memory with close to DRAM performance. Intel Optane PMem offers extremely low latency and a very

good price/capacity-ratio – ideally matching the requirements of the HPC and HPDA field.

Furthermore, each node can utilize two 1.5 TB Intel Optane SSD data center hard-drives. They are configured either as a temporary user partition, a checkpoint/reboot-memory or as a BeeOND (BeeGFS on-demand) parallel file system, depending on the application.

The DAM consists of 16 nodes, each containing two 2nd Gen Intel Xeon Platinum 8260M scalable processors (each with 24 cores, 2.4 GHz clock frequency and 35.75 MB L3-Cache) as well as a 40 GbE cluster fabric for the connection. A 100 Gbps EXTOLL interconnect with Tourmalet PCIe3 add-in-cards tops off the hardware. The DAM has been especially designed to run HPDA codes like unattended clustering (DBSCAN or KMeans), attended clustering with support vector machines (SVMs) or random forests and of course deep learning. One practical example for this kind of application are the co-design workloads for image analysis of KU Leuven and the University of Iceland.

Homogeneous memory areas for optimal machine learning performance

HPC applications usually read and process data sequentially as large arrays. However, machine learning and data analytics applications access the memory randomly. More operations are executed on small data units and smaller data types are used. Due to these specific needs, which are not addressed by traditional HPC architectures, the DAM, which involves FPGAs that can perfectly match them, has become extremely important. A scalable cluster system like the DAM is an energy-efficient solution for parallel applications, since the memory capacity required by a workload can be provided by allocating a matching number of nodes.

For tasks dealing with HPDA and ML, I/O-throughput is much more important than in traditional simulations. Nevertheless, a strong integer-calculation performance is needed to profit from the optimizations built into many machine learning algorithms, for example in the co-design workload of the partner KU Leuven. The DAM design allows memory-intensive applications in the field of bioinformatics and machine learning to run efficiently. The time-consuming modelling process of HPDA applications often reuses large datasets. With persistent memory like Intel Optane, these datasets, which are accessed again and again, can be supplied much faster than with traditional storage or volatile memory technologies.

The Intel FPGA PAC D5005 (PCIe3 x16) cards used in the DAM are freely programmable and constitute the easiest way of using existing ML frameworks. Developers can choose from a wide variety of already optimized libraries and frameworks for Intel FPGA PACs. Having a large amount of existing and tested software at hand can speed up the implementation of workloads on the DAM. The need for custom adaptions and code development is reduced to a minimum. However, users still can create custom code if necessary, for example if the application is not yet covered by existing software. The choices start at high-level programming languages like OpenCL and Data Parallel C++ (as part of Intel® oneAPI) but can also include a very hardware-oriented language like VHDL. Moreover, the Intel FPGA PACs leverage the Intel® Acceleration Stack for Intel Xeon CPU with FPGAs providing optimized and simplified hardware interfaces and software application programming interfaces (APIs), saving developers time so they can focus on the unique value-add of their solution. The workloads that are used for testing the Modular Supercomputing Architecture already employ both approaches – existing frameworks and custom software.

Results so far and next steps

Even though the project is still in progress, the experience gained already shows a lot of potential for future applications and for other institutional and private or business research institutes. The project partners expect a significant performance increase for suitable workloads, compared to monolithic systems. This is true for workloads which profit from calculations run on specialized modules. All involved partners will have full access to all measurements and test results so that each institution can put individual variations of the architecture to the test and refine them. The end result should be pre-built software packages which already contain the majority of customizations for specific workloads and can be deployed even faster.

The JSC will align its future procurements along the DEEP-EST concept. Already in mid-2020, the installation of a Booster module is scheduled for the production system JUWELS. Even the current DEEP-EST test platform has convinced users of its suitability for a very good universal blueprint design. While scientists at partner universities and institutes involved are still evaluating the platform, the wide range of applications and their validity shows, that the Modular Supercomputer Architecture is on the right track for a bright HPC future.

Solution highlights

- Combination of technologies for acceleration of complex scientific simulations, data analysis and machine learning applications
- Heterogeneous applications and workloads can run on precisely tailored compute and memory resources
- Intel Xeon Platinum 8260M processors with VNNI instructions and Intel FPGA PAC D5005 for data analysis & ML inference
- Intel Optane persistent memory able to act as RAM extension or as an ultra-fast persistent data buffer as well as Intel Optane SSDs for fast data storage
- Opportunity to quick connect additional system modules, for example specialized Artificial Intelligence platforms
- Full utilization of the whole Intel spectrum of tools (Compiler, Analysis, ML-Frameworks) and Libraries (Intel® Math Kernel Library, Intel® Data Analytics Acceleration Library)
- Ready-to-solution with drastically reduced time- and energy consumption

Efficiently supporting as many scientific fields of application as possible is one of the main requirements of the Modular Supercomputer Architecture (MSA). Several key partners are already using the prototype at the Supercomputing Centre in Jülich for reference projects:

Radio Astronomy, Astron

All around the world, large antenna arrays are observing the universe through radio waves. Data received by the separate antennas needs to be aggregated and analyzed. The processing takes place in a sequential operation. Two of those steps are especially processing intensive: The combination of data (Correlator) and the calculation of the image (Imager). Correlator and Imager have been ported to several platforms (GPUs, Intel Xeon CPUs, DSP and FPGA). With the MSA prototype, the researchers can evaluate the best platform for each task and the most energy-efficient approach.

Space Weather, Katholieke Universiteit Leuven

The sun is a giant sphere of plasma, a pool of molecules so hot that electrons are ripped apart from atom nuclei. These electrically charged particles are trapped in the strong magnetic fields of our star. The magnetic fields expand gradually into the solar system carrying the plasma outwards. When they become too strong on the solar surface, they can produce giant magnetic eruptions, called flares. Magnetic fields and plasma eruptions can cause disruptions in electronic devices and impair satellites. With its new simulation model, the University of Leuven is trying to predict the effects of solar activity on our planet. A machine learning code running on the DAM is used to connect the sun and the earth. A traditional HPC code that runs on the other modules is then used to study the earth plasma environment.

Neuroscience, Norwegian University of Life Sciences, Ås

The NEST software is used to simulate neuronal network models representing networks at the level of simplified neurons and synapses which interact through spikes, stereotypical electrical pulses. These simulations are useful for a wide range of research topics ranging from new learning paradigms for neurorobotics to models of animals and eventually even the human brain. NEST by itself exploits the capabilities of conventional supercomputer architectures very well, but leaves the scientists with huge amounts of output data requiring further processing and analysis. With the MSA, data generated by NEST on the Compute Module can be fed directly into the statistical analysis tool Elephant running on the DAM to extract relevant signals from raw data while a large brain simulation is running.

High-energy Physics, CERN Geneva

CERN's Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator. Particles are accelerated to close to the speed of light and then collided, generating huge amounts of data. Collision events are reconstructed using data from giant detectors. The DAM helps in assigning the respective energies to all newly created particles and in reconstructing the collision by combining all individual data points. This is a task perfectly suited for the DAM because it can process large amounts of data without lengthy setup and pre-formatting.

Data Analytics in Earth Science, University of Iceland

Remote sensing, for example via satellite sensors or laser-based LiDAR (Light Detection and Ranging), is used in numerous fields. One of them is earth science when the earth needs to be observed with respect to land cover in order to survey, for example, environmental changes. Machine learning is used to automate the classification of land cover from satellite images and to cluster LiDAR point clouds to separate objects. Training of neural networks and support-vector machines benefits from the fast and energy-efficient accelerators and huge memory of the DAM.

For more information about Intel solutions in HPC and AI, visit www.intel.com/hpc.