NEXT GENERATION SUPERCOMPUTING –
BOOSTING SCIENCE IN EUROPE
3  From Vision to Reality

4  A DEEP Look
5   DEEP Hardware Architecture
8   DEEP Software Stack
10  DEEP Applications & Exascale Projections

14  A DEEP Impact

16  A DEEP Look Behind the Scenes

18  A DEEP Connection
19   DEEP Consortium
20   DEEP Continued
22   DEEP in the Exascale Community

24  DEEP Moments
25   DEEP Involvement

27  Imprint
Challenging fundamental laws of nature by experiment is crucial for progress in science. In parallel computing, we are familiar with Amdahl’s Law, considered as fundamental for strong scaling. It states that the fastest speedup achievable through parallelization is limited by the scalar part of the program, or, as a generalization, the part that is least scalable.

The DEEP Project* is an experiment in parallel computing. Its goal is to demonstrate that the limits imposed by Amdahl’s generalized law can indeed be alleviated.

To push the scalability of applications to the limits, DEEP has proposed a new approach to heterogeneous computing that best matches the different intrinsic concurrency levels in large simulation codes. The Cluster-Booster architecture combines two distinct hardware components in a single platform: the Cluster – equipped with fast, general-purpose processors that show highest single thread performance but with a limited total number of (expensive) cores and being less energy efficient – and the Booster – composed of many-core Intel® Xeon Phi™ processors connected by the EXTOLL network, all together most energy efficient, highly scalable and massively parallel. Code parts of a simulation that can only be parallelized up to a limited concurrency level stay on the Cluster – equipped with faster general-purpose processor cores – while the highly parallelizable parts of the simulation are to run on the weaker Booster cores but at much higher concurrency.

To enable application developers to best exploit the machine, DEEP developed the DEEP system software. It allows tasks to be distributed dynamically to the most appropriate parts of the hardware to achieve highest computational efficiency. The MPI programming paradigm in combination with an improved version of the OmpSs task-based programming environment enables application programmers to abstract from the system software by simply requesting the necessary resources. The rest is done transparently and dynamically by the system.

Within less than four years, the DEEP project has realized substantial technological innovation: joining a Eurotech Aurora Cluster system, the European hardware team designed and constructed the entire DEEP Booster from scratch. With 384 first-generation Intel® Xeon Phi™ processors, the DEEP Booster is the largest Xeon Phi based system in Europe, with a peak performance of up to 400 TFLOP/s. But even more importantly, this prototype is different from anything seen in the HPC landscape until now: it is the only platform world-wide in which the Xeon Phi processors do operate autonomously without being attached to a host. This provides full flexibility in configuring the right combination of Cluster and Booster nodes, to optimize the use of the hardware for each application.

The software environment developed by DEEP will eventually become the most important legacy of the project. Communication protocols were developed to efficiently transfer data between different network technologies; programming models were extended to support a new level of hardware heterogeneity, and performance analysis tools have been adapted to study and model next generation HPC platforms. The DEEP software is ready for the heterogeneity expected at Exascale.

Six pilot applications were selected to investigate and demonstrate the benefits of combining hardware, system software and the programming model to leap beyond the limits of Amdahl’s Law. During the project they were highly optimized, acted as drivers for co-design leading to the final realization of hardware and software in the project, and served to identify the main features of applications that most benefit from the DEEP concept.

The DEEP prototype is in operation, the amazing adventure continues in the DEEP-ER project, and we are very much encouraged to venture on the DEEP-ESD project, realizing an Exascale Demonstrator.

*co-funded by the European Commission in the 7th Framework Programme under Grant Agreement number 287530
A DEEP LOOK
The DEEP project has developed innovative HPC system prototypes that validate a promising architectural concept for building Exascale-class systems.

The prototypes validate the Cluster-Booster architecture, which takes the concept of heterogeneous computing to a new level. They combine:

- A standard, InfiniBand®-based Cluster using Intel® Xeon® nodes (Cluster Nodes)
- An innovative, highly scalable Booster constructed of Intel® Xeon Phi™ co-processors (Booster Nodes)

This architectural concept accommodates the fact that most HPC codes are characterized by different concurrency levels. The highly scalable code parts run best on the Booster side of the machine, while code parts with limited scalability benefit from the flexibility and the high single-thread performance of the Cluster.

To put this conceptual idea into practice, the DEEP project has defined a system architecture that fully leverages leading-edge multi-core and many-core processors, interconnects, packaging and cooling methods and monitoring/control approaches and thus addresses key challenges on the way to Exascale.

The Cluster uses Eurotech’s proven and highly efficient off-the-shelf Aurora technology. It consists of 128 dual-socket blade nodes (with Intel Xeon E-5 CPUs) in eight 19-inch chassis and interconnects them with QDR InfiniBand®. Eurotech’s direct liquid cooling technology enables year-round free cooling of the system.

For the Booster, the project has developed two distinct prototypes:

- A 384-node system built by Eurotech from custom-engineered dual-node cards in the Aurora blade form factor – the Aurora Booster with an aggregated performance of around 500 TFLOP/s
- A smaller 32-node prototype built by University of Heidelberg and Megware, based on the latest ASIC implementation of EXTOLL (GreenICE Booster)

Both Booster prototypes profit from the high throughput of Intel Xeon Phi co-processors and the performance of the novel, direct-switched 3D torus interconnect developed by EXTOLL. The Booster interconnect was selected to ensure scalability up to Exascale levels and to best match the spatial domain decomposition schemes commonly used by scalable HPC codes. Whereas the former prototype uses an FPGA implementation of the EXTOLL interconnect and is in 24×7 productive use, the latter leverages the brand-new ASIC implementation of EXTOLL and experiments with immersive liquid cooling technology.

![DEEP architecture diagramme](image-url)
DEEP HARDWARE ARCHITECTURE

Aurora Booster
Eurotech developed the liquid-cooled Booster machine for the DEEP project’s Cluster-Booster concept. While the Cluster is the Eurotech proven Aurora technology, the Booster is a challenging new derivative design.

The DEEP Booster is a tight integration of the Intel Xeon Phi 7120X cards with Altera Stratix V FPGAs running EXTOLL communication. Each FPGA connects to one Intel Xeon Phi via PCI Express and provides seven external links to form a 3D Torus. The Eurotech design challenges the limits of the copper communication technology with 1.5 Tb/s per board with a pitch of 25mm in a 6U rack space. The communication between the nodes can run at a maximum copper link speed and provide 120 Gbps node-to-node over the backplane in a chassis and the 12-pair Molex cables between the chassis.

Aurora packs two Booster nodes on a single liquid-cooled blade. A chassis has eight blades with a single Booster Interface blade. The Booster Interface blade uses Intel Xeon E3 CPU with Avago PCIE switch routing signals to Altera FPGA EXTOLL NIC and one Mellanox ConnectX 3 InfiniBand® NIC.

The full Booster system has 384 Xeon Phis packed in 42U by 23-inch rack space, giving 500 TFLOP/s. All blades use Eurotech’s Aurora direct liquid cooling technology that separates electrical and signalling backplanes from the water distribution system.

GreenICE Booster
To demonstrate the performance impact of the new ASIC implementation of the EXTOLL interconnect, University of Heidelberg created an alternative Booster prototype based on the latest GreenICE technology. In lieu of a tightly integrated Booster board, a passive PCI Express backplane pairs eight Intel Xeon Phi 7120D cards with eight EXTOLL TOURMALET NICs. Thanks to EXTOLL technology the Booster nodes can be scaled independently of the Cluster nodes. This system of 32 Intel Xeon Phi nodes yields 38.4 TFLOP/s peak performance.

To ease integration, four assembled backplanes are completely immersed in a basin with 3M Novec®-649 fluid, which also contains the required power supplies and the management CPU. In operation, the heat produced by these components evaporates the Novec fluid (which has a boiling point of 49°C). The Novec vapour is then cooled by loops of special copper pipes (with maximized surface area) with water as the cooling liquid. The condensed vapour drops back into the basin.

System management is performed by a Raspberry Pi system via I2C connections to the Backplane and power supplies. The Booster Interface is implemented with standard Intel Xeon server boards in air-cooled chassis.

The EXTOLL links are carried via copper cables that attach to the NICs using standard HDI6 connectors.

Paul Arts
Technical Director & Head of Operations
HPC Business Unit, Eurotech

"Design challenges of the DEEP project inspired Eurotech innovation. Cooperation within the team and support of best scientists helped to build this extraordinary machine."

Prof. Ulrich Brüning
Professor for Computer Architecture,
University of Heidelberg

"Within the DEEP project we are proud to have developed a highly dense Booster implementation with immersion cooling and the EXTOLL ASIC."
Energy Efficiency
The computational power of an Exascale machine is roughly equivalent to that of 10 million state-of-the-art laptop computers. Yet if we were to assemble the hardware of these laptops into a single system, the power consumption would be in the order of what one large nuclear power plant can produce. Thus, building an Exascale-ready supercomputer first and foremost entails addressing the energy efficiency challenge. The DEEP prototype is based on technologies that reduce the system’s energy consumption, and at the same time help users optimize and tune the system according to their needs.

Besides the efficient cooling technologies implemented in DEEP, the comprehensive monitoring capabilities of the system enable in-depth analysis of the machine’s operating conditions. In addition, the availability of fine-grained and high-frequency power usage traces supports application developers in identifying bottlenecks and tuning the energy efficiency of their codes.

Finally, flexible integration of the monitoring system with external data sources, such as the data centre infrastructure, provides system administrators with a complete picture of the operating condition of their system without having to manually query several different data sources.

Installation
The DEEP Cluster and both Booster systems were installed at Jülich Supercomputing Centre as part of a production HPC environment. All three systems are fully integrated, yet the two Booster systems can be operated completely independently.

To ensure a safe and reliable 24×7 production environment, the qualification and installation of critical system software layers was essential. Logically, also the infrastructure for hot water cooling was developed, put into operation and integrated with the DEEP systems at JSC.

The cooling infrastructure was designed to enable the use of year-round chiller-less cooling on a newly built cooling loop.

Experiments with active cooling using an existing cold water supply are possible as well. Electrically controlled valves permit rapid reconfiguration, and special filters ensure the water quality is up to standard. The maximum coolant temperature at Jülich is 40°C.

Safe unattended operation 24×7 was a top priority. The additional element of risk imposed by direct liquid cooling was addressed by including sensors, which will detect even very minor leaks. An always-on monitoring system registers any critical excursions and mitigates the impact of failures, in the worst case by switching off all power to the affected parts of the systems using web-relays in the 230 V lines.

Axel Auweter
Research Associate
Energy Efficiency,
Leibniz Supercomputing Centre

“Working with the latest hardware technologies and watching a leading-edge HPC system coming together was an amazing experience.”

Jochen Kreutz
Research Associate
System Installation,
Jülich Supercomputing Centre

“We have never worked closer with our partners than in DEEP. Only through this close collaboration were we able to implement all our ideas for a truly energy-efficient system.”
Programming a heterogeneous system like DEEP is a challenging task for developers of HPC applications. To minimize the effort of porting existing applications to the Cluster-Booster architecture, special emphasis was placed in the DEEP project on developing a programming model that gives as much support as possible to Cluster-Booster machine users. Additionally, the DEEP software stack ensures that HPC applications can make the most of the underlying innovative heterogeneous hardware architecture. The team performed a tremendous co-design effort in collaborating with the hardware and the applications team to achieve this.

Managing a truly heterogeneous platform
While traditional supercomputers are either totally homogeneous or heterogeneous only at the node level, the DEEP system is heterogeneous at the cluster level; it mixes two different sets of compute nodes — Xeon and Xeon Phi — and two different types of interconnect networks — Infiniband on the Cluster side of the system, EXTOLL on the Booster side. In this sense, the DEEP architecture radically departs from existing approaches and is heterogeneous in a twofold way. In order to hide this complexity from the application developers, the software stack implements two abstraction layers:

- ParTec’s ParaStation MPI serves as the basic parallelisation layer and was extended into a global MPI covering both Cluster and Booster.
- OmpSs was chosen as programming model and extended to provide flexible and powerful offload features.

First abstraction layer: MPI
The choice of MPI supports the fact that the guiding applications of the DEEP project are all based on the MPI programming paradigm. Plus, MPI is the most widely used basic communication layer in HPC in general. The architecture’s twofold heterogeneity needs to be reflected already on the MPI level.

Therefore, the DEEP programming environment provides a global MPI layer. It helps the developers to decompose their applications so as to create efficient overlaps between computations done on the Cluster side and on the Booster side of the system. The parts of an application with a complex communication pattern or needing high single-thread performance (low to medium scalability) run on the DEEP Cluster, while the (highly scalable) code parts with regular communication patterns benefit from the Booster. This mapping of different code parts onto the two sides of the DEEP system is achieved by offering an MPI-compliant process spawning mechanism that
DEEP implements a first incarnation of the heterogeneous Cluster-Booster architecture. It aims to pursue the successful concept of cluster computing into the many-core era, carrying the potential to reach Exascale. We expect this goal to be barely reachable with standard HPC clusters as they are in use today.

allows for the creation of new MPI processes at application runtime in a dynamic and heterogeneity-aware manner. This way, MPI applications can easily offload further MPI kernels from the Cluster to the Booster and vice versa.

Because Cluster and Booster use different interconnect networks, dedicated development and runtime environments are provided: while on the Cluster side an MPI library driver specifically optimized for the InfiniBand fabric is provided, on the Booster side a corresponding driver supports MPI communication over EXTOLL and is used by the highly scalable code parts for intra-Booster communication. In addition, a dedicated Cluster-Booster protocol ensures seamless, transparent communication between both parts of the system. It uses InfiniBand to access memory exported by all Booster nodes via EXTOLL in a PGAS-like manner. The connecting edge is the Booster Interface housing both networks and hosting a special communication daemon to support this functionality also from the software point of view.

Finally, all these features as offered by the ParaStation Global MPI layer for the DEEP system are also being used by the OmpSs runtime environment for offloading code parts according to the OmpSs programming model.

Second abstraction layer: OmpSs and the DEEP Offload

Many HPC applications are developed for homogeneous clusters and typically use a Single Program Multiple Data (SPMD) execution model. Hence per se they cannot directly make the most of the highly scalable and extremely flexible DEEP platform. To address this situation, OmpSs, a data-flow programming model based on user annotation, was implemented into the DEEP software stack and extended with offload capabilities. This is an essential feature to make the DEEP system even easier for application developers to use.

The DEEP Offload – A powerful tool

The DEEP Offload extends the OmpSs programming model with two key features: the dynamic allocation of nodes and the collective offload of unmodified MPI kernels. The first feature enables applications to dynamically use the resources they require for each computational phase (for instance data pre-processing, main computation and data post-processing). The second feature enables the offload of arbitrary computations (MPI kernels) to the dynamically allocated resources.

With these two features it is much easier to switch an application from an SPMD execution model to a Multiple Program Multiple Data (MPMD) execution model that perfectly fits the DEEP architecture. The ability to offload unmodified MPI kernels avoids the need to rewrite applications from scratch, something that besides being time-consuming and error-prone is not even feasible for the large and complex applications used on DEEP. The DEEP offload is the result of a strong co-design effort between application developers and programming model experts, which has led to a model that will be useful beyond the DEEP project.

Implementing the DEEP offload features on the OmpSs programming model ensures that they will be supported and available on a wide range of systems. In fact, these offloading features have already been successfully used on other systems such as Marenostrum3 (Xeon) and Stampede (Xeon & Xeon Phi) supercomputers or the Mont-Blanc prototype (ARM).
Making applications ready for Exascale is a necessary, yet challenging undertaking. In DEEP, six real-world HPC codes have been tuned to the system. The optimisation also delivered performance increases in existing heterogeneous systems and proves the fruitfulness of the general code modernisation achieved in the project.

The DEEP architecture introduces heterogeneity in an innovative way. Programming an HPC system like this and taking advantage of it might sound complicated. However, across all applications it could be shown that only a limited amount of change is necessary to benefit from the Cluster-Booster architecture.

During the project, the applications team worked together effectively with the OmpSs developers to better understand the requirements and constraints imposed by the nature of the applications and by the design of the runtime environment. As a result, both parts of the software were modified to better fit together, and permit easy use of the DEEP architecture.

Every application is different and therefore needs to be considered as a different use case. However, the project delivered impressive evidence of the number of ways HPC applications can benefit from the flexibility of the DEEP hardware and software architecture. For instance, reverse offloading (Booster to Cluster), I/O offloading and dynamic offloading of discrete tasks are all possible on a DEEP machine, and can easily be ported to other systems.

The following use cases show how each application makes use of the unique DEEP architecture. Additionally, this section highlights further code improvements achieved within the project time frame – something the application developers will benefit from on other future systems as well.

### Brain simulation (EPFL)

Brain simulation is making giant leaps towards a better understanding of the inner workings of the human brain. In DEEP, partner EPFL adapted CoreNeuron, an advanced brain simulation application. This simulation requires a model to be built. This, however, poses a significant challenge when making the application future-proof, as the model-building and the actual simulation need to be separated. This separation makes it possible to have more neurons per node, thanks to the reduced memory footprint.

Another important challenge is the implementation of efficient threading, which is more suited for modern processors than a traditional MPI model. The threaded parallelization has very good parallel efficiency, thanks to the introduction of a clever static load balancing scheme. It ensures – prior to the simulation – that all the threads perform the same number of operations, taking into account the complexity of different neurons.

A significant change, required in order to benefit from modern processors, is to make the code vectorisation-friendly. The data structures were changed to favour vectorisation. The loops were carefully rewritten, and hints for the compilers were introduced. In the computation-bound kernels the impact has been dramatic, and in the memory-bound kernels the performance improvement has been very noticeable.

These changes make it possible to achieve extremely good scalability for simulations with a large number of neurons. This makes the bulk of the application run efficiently on the Booster. Given that I/O is done more efficiently on the Cluster, EPFL decided to start the simulation on the Booster, and offload the I/O to the Cluster. In this way, I/O has been improved by more than an order of magnitude (by executing it directly from the Booster). More importantly, this lays the foundation for interactive supercomputing in the future, showing that the DEEP architecture is suitable and aligned with coming changes needed in the brain simulation community.
Space weather (KU Leuven)
Simulating and understanding space weather phenomena is of critical importance for the safety of our electrical, telecommunications, and space infrastructure. KU Leuven is at the forefront of space weather simulation. Within the DEEP project, a Particle-in-Cell (PIC) code is used to understand the effects of the plasma that constantly flows from the Sun to the Earth. The software employs two separated solvers, one that calculates the electromagnetic fields in a Cartesian grid, and a second that traces individual electrons and ions across the simulated domain.

Particle-in-Cell codes are fundamentally massively parallel, since the very large number of charged particles required for the simulation can be processed independently. The field solver, on the other hand, is parallelized using a subdomain decomposition, which requires constant communications, limiting its scalability. As a result, the particle solver is placed in the Booster, whereas the field solver runs on the Cluster. The amount of data transferred between the solvers is small and limited to the field information stored in the Cartesian grid from the field and particle solvers. This unlocks better scalability, and reflects the spirit of the DEEP architecture.

For optimal performance the particle processing had to be optimised for Xeon Phi processors. The code evolved from a pure MPI implementation to a hybrid MPI+OpenMP implementation, focusing on the computation of the particle movement and its associated fields. Physical requirements of the algorithm push towards an array of structs (AoS) data layout, to allow for easy sorting of particles. However, to achieve efficient vectorization, a fast on-the-fly transposition from AoS to SoA was implemented using intrinsics. The overall result is an extra edge of performance using the DEEP architecture.

Climate simulation (CYI)
Understanding the evolution and changes of the global climate is of utmost importance in the 21st century. The complexity of climate simulation is reflected in the structure of codes in the field. In this case, the application consists of two coupled models. The atmospheric model represents pressures, currents, temperatures and related magnitudes of Earth’s atmosphere. Coupled to this base model, a chemical simulation package analyses fine-grained interactions between chemical elements.

The atmospheric model requires a significant number of transformations and data transpositions, resulting in constant global communication and lack of overall scalability. On top of that, photochemical effects caused by changes in sunlight over the Earth result in a very significant load imbalance, worsening the already suboptimal scalability. Processing these local photochemical effects is what consumes most of the time in these simulations, due to the synchronicity of the model and its heavy computation requirements. In DEEP, individual tasks are offloaded to the Booster dynamically, effectively reducing the load imbalance and allowing the code to scale further than before, due to two reasons: 1) The atmospheric model can be kept as small as possible to avoid excessive communication, without hindering the heavy computing parts of the code; and 2) the load imbalance, the main concern to scale the code, is effectively eliminated.

“Applications tuned to the DEEP system also show better performance on other platforms. If you use MPI you don’t even have to change a line of code. This means, application developers outside of the project can benefit tremendously from our work.”

Damian Alvarez Mallon
Application Support Engineer,
Jülich Supercomputing Centre
Computational fluid dynamics (CERFACS)
The development of more efficient combustion engines is unthinkable without the use of CFD applications. AVBP, developed at CERFACS, is one of the most important applications in this field in Europe and has been ported to DEEP. During the project, the application went through a series of transformations that represent a very significant step forward in terms of scalability and performance.

The excellent scalability of the simulation was limited by the serial nature of I/O and mesh partitioning at the beginning of the project. As a first step these two bottlenecks were removed. The next step was to migrate from a pure MPI approach to a hybrid approach of MPI+OmpSs. The OmpSs model makes it possible to expose additional parallelism, and by using OmpSs, we were able to implement a version of the application that outperforms and outscales the previous one. Loop refactoring and compiler hints gave an extra edge in performance, as now the vector units are used more efficiently.

The application scalability behaviour and structure makes it suitable for running on the Booster. However, I/O, as noted above, is a challenge in certain cases. In DEEP, the application is started on the Booster, and I/O is offloaded to the Cluster, together with costly reductions. These reductions become a burden at large scale, and are needed as a preparatory step for I/O. With them running on the Cluster, with the simulation progressing in parallel on the Booster, this burden is now heavily reduced.

High-temperature superconductivity (CINECA)
Quantum Monte Carlo applications are used for research on high-temperature superconductivity. Like all Monte Carlo applications, this one is embarrassingly parallel, as each walker explores its space with little to no communication. The most intensive part of the application consists of basic BLAS2 and BLAS3 operations, and therefore a good implementation of them is mandatory. Given this structure, the application is in principle suited for massively parallel systems, and in this case the Booster.

However, this is a good example of the need for code modernization. With this being a pure MPI application, threading is limited to the BLAS library, which, depending on the size of the system simulated, might or might not be enough. Attempts to improve this met with limited success due to the structure of the code. At the same time, having a very large number of MPI processes per node increases the burden of collective communications.

In DEEP, OmpSs proved its flexibility by allowing a single process running in the Cluster to act as the driver for the whole simulation, offloaded to the Booster.

Seismic imaging (CGG)
Seismic imaging is a set of key techniques for efficient oil and gas exploration. There are different ways to achieve an accurate representation of the subsoil layers, but all of them share a master-slave approach: different seismic shots are processed by the slaves in parallel with complete independence, whereas the masters dispatch the shots to the slaves and accumulate their results. A master-slave structure maps nicely to the DEEP architecture, as the slaves’ work is basically completely parallel, and can scale in the Booster as long as there are enough shots to be processed. This is the approach followed by CGG, where the master processes take care of the I/O, which is performed efficiently on the Cluster, whereas the larger memory bandwidth of the Booster benefits the kernels of the slave processes. The key benefit in this case is the dynamic ratio of masters and slaves, as opposed to a static ratio in a traditional cluster with attached coprocessors.
LOOKING INTO THE DEEP EXASCALE FUTURE

When imagining a full DEEP Exascale system – a system 2,000 times larger than the actual DEEP prototype – the question arises how the six scientific applications ported to the DEEP system would perform. Performance models and predictions are very helpful to find the answer.

The Barcelona Supercomputing Center has therefore developed a methodology based on their performance tools. Using traces from application runs on a few core counts (from 512 up to 4,096 MPI ranks) predicts the behaviour for half a million ranks. BSC applied it to three of the six DEEP applications and the predictions were successfully validated up to 256K ranks with non-instrumented executions on Jülich BG/Q system.

The most important dimension to be evaluated was parallel efficiency. With the help of BSC tools, the results obtained were further analysed to determine the limiting factors. These insights were provided to the application developers to further optimize their codes.

CoreNeuron from EPFL was executed with a large input case dominated by the computation, which can run up to 64K MPI ranks. With such a scenario, the parallel efficiency of a half million cores is 92%. The main factor that influences this small inefficiency is load balance. The code does not allow to split neurons so the scalability is limited to the number of neurons being simulated.

TURBORVB from CINECA reported parallel efficiency of 60% when running with half a million cores. The study found that the main factor limiting the efficiency is the transfer of data due to node contention. A simple modification that limits the random selection of processes per node would reduce the traffic from the nodes to the network, achieving better scalability.

AVBP from CERFACS usually runs in strong scale mode. Therefore, the predicted parallel efficiency for a half-million core run goes down to 38%. The model indicates that code dependencies are the main factor limiting efficiency. This is due to the limited scaling of the small computations between MPI calls that at larger scale become a higher percentage of the iteration time.

BSC performance tools and prediction techniques allowed us not only to estimate the performance that the applications will get in system sizes far beyond our current prototypes but also helped identifying code/algorithmic refactorings that will result in overcoming bottlenecks that current programs will be exposed at Exascale.

Prof. Jesus Labarta
Director Computational Sciences Department, Barcelona Supercomputing Center

Comparison of Predicted and Measured Fundamental Factors
{timeBased_neuron
Comparison of Predicted and Measured Fundamental Factors (turborvb.T,S,LB=cubic_ahmdahl)
Comparison of Predicted and Measured Fundamental Factors
{avbp_juqueen.linear_amdahl}
DEEP LEGACY

DEEP IMPACT

With its key achievements and the large body of expertise created, the DEEP project is poised to have a significant and lasting impact along four vectors. Besides opening up new avenues for the architecture of efficient HPC systems, it has materially increased Europe’s indigenous capabilities in HPC system design and production, and has produced a complete system software stack together with a programming environment for heterogeneous platforms. Six relevant applications in critical fields of the European Research Arena have been remodelled and adapted, and what is more, best-known methods have been established that will enable many more codes to reap the benefits of the DEEP software and hardware architecture.

Novel approach to heterogeneous cluster computing
The DEEP system has proven that the Cluster-Booster architecture concept of dynamically associating different kinds of computing resources to best match workload needs can be implemented with state-of-the-art multi-core and many-core technology, and that such a system can indeed provide a superior combination of scalability and efficiency. It has thereby opened up a new avenue towards affordable, highly efficient and adaptable extreme scale systems (up to Exascale-class), merging the hitherto separate lines of massively parallel and commodity Cluster systems. The sibling project DEEP-ER is already carrying the flag further by integrating novel memory and storage concepts and providing scalable I/O and resiliency capabilities.

Leading-edge European system development
With its unprecedented integration of sensors, the DEEP system delivers a wealth of voltage, current and temperature data for all system components at high frequency, and uses this data for good measure to optimize operating parameters and safeguard operation. This example will influence future HPC system design and create opportunities for advanced analysis of monitoring data and data-driven system management.

"The companies, research institutes and universities behind DEEP can all be proud of having created a unique system, which is both most generally applicable and scalable beyond imagination. The DEEP Cluster-Booster concept will become part of the future of supercomputing."

Thomas Lippert,
Head of Jülich Supercomputing Centre

DEEP IMPACT

The companies, research institutes and universities behind DEEP can all be proud of having created a unique system, which is both most generally applicable and scalable beyond imagination. The DEEP Cluster-Booster concept will become part of the future of supercomputing.

Thomas Lippert,
Head of Jülich Supercomputing Centre

Eurotech is one of the world-wide pioneers of direct liquid cooling for HPC. The DEEP project is the proof that hot water cooling can be safely operated, is compatible with modern system technology and can indeed provide free cooling year-round. These results will most importantly shape the expectations of HPC customers, who now know that they can eliminate a significant part of operating costs, and in turn materially increase the take-up of hot water, direct liquid cooling by future HPC systems.

Equally remarkable is the achievement of University of Heidelberg and their spin-off EXTOLL GmbH. They have brought a completely new and highly competitive interconnect into the HPC landscape: the direct-switched EXTOLL network has shown excellent scalability in the DEEP Booster system, and the performance improvement brought by the new ASIC-based TOURMALET implementation is a clear demonstration of the capability of this European technology.
The Munich-based HPC software house ParTec has contributed key expertise to design and build critical parts of the system software stack. The proven ParaStation product was the basis to master the major software challenge of DEEP’s Cluster-Booster concept: the seamless co-operation of two physically autonomous parallel systems, bridging dissimilar networks efficiently, without requiring substantial CPU involvement on the Booster Interface nodes while being transparent for the application. ParaStation ensures the future adaptation of the Cluster-Booster concept to a much wider range of heterogeneous platforms.

Collaborating in DEEP, the three European HPC companies, Eurotech, EXTOLL and ParTec have gathered a large body of invaluable expertise in designing, integrating and manufacturing the system together with Intel. The DEEP system is testament that these European technology companies are able to create innovative HPC solutions which deliver highest density and leading efficiency while fully mastering system complexity. DEEP’s success will boost the market position of the three companies and will provide them with new opportunities.

Managing such a large-scale supercomputer project and driving the co-design between applications experts, system software developers and hardware architects is in truth no small task. JSC (for the project as a whole), LRZ (for the critical energy efficiency area) and BSC (for the programming model co-design) have amply demonstrated their capability to rise to the challenge. This will set them up as prime partners for the next round of system-centric co-design projects in Europe – in particular considering the impact discussed here.

Software innovation towards Exascale
The DEEP system software and programming model were carefully architected to be based on existing standards and product-quality solutions. Extensions were made where necessary to make the unique DEEP features available or enhance the ease of programming. Supported by the application proof points, the resulting software stack will certainly and substantially influence the direction of Exascale software architecture, with ParTec as a European HPC software house in a key role. Today, it provides a solid base for increasing the circle of applications optimized for heterogeneous architectures in general, and in particular for the DEEP-ER project.

Furthermore, the innovative monitoring and control hardware and software infrastructure prototyped in DEEP has created substantial progress in the field, showing how high-frequency sensor data can be collected and processed in a scalable way, and how it can effectively interact with the firmware of the system components to ensure safe and efficient operation.

Last but not least, proven performance analysis and modelling tools from JSC and BSC were extended in the project to fully support the programming models; they were also used to predict the performance of scaled-up systems, establishing a precedent for full system performance projection in the scaling dimension without the need to first create analytical application models. This could prove an extremely valuable innovation, given the need to carefully analyse architectural choices before building expensive Exascale prototypes.

Proof of concept: DEEP applications
Six relevant, real-world applications in important scientific and engineering fields of the European Research Arena have been thoroughly analysed, modernized and adapted to the DEEP architecture. This enables users to make scientific discoveries faster, and engineers to come up with better solutions. Since the DEEP software interfaces are based on standards and backed by a commercial company, they will be adapted to future heterogeneous platforms, enabling the six applications to take advantage of such new systems. In addition, the DEEP-enabled codes continue to run on conventional architectures, sometimes showing surprising performance and efficiency improvements compared to their old formulation.

Even more importantly, the experience gathered in the application analysis and adaptation was distilled into “best-known methods”, resulting in a playbook for tackling a wide range of additional applications and preparing them for DEEP-class systems. It is our hope that this will have a profound beneficial effect on the entire application ecosystem.
Having led the project for the last three and a half years, what do you think is so special about DEEP and what excites you about its potential?

This project presents an architectural concept that addresses the most important challenges we face when working towards Exascale. Hardware, system software, tools, energy efficiency and applications are some of the topics covered. This holistic approach makes it a really challenging project but a very exciting one to be part of as well.

Integrating all these aspects into a medium-sized prototype is the first step on our way to building an Exascale-ready production system that really works. This is not only important within the relatively small HPC world but also for society at large, since the technology developed in the project is also advancing European industry and research, increasing their competitiveness in the future.

What does it take to make such a complex project a success?

One key to success is the research infrastructure provided by the European Union. It helps to bring together the people and project partners with the right skills. Additionally the scientific guidance provided by the EU selected reviewers is extremely valuable.

From a project management point of view, it is challenging to unite all the different partners involved – but it is definitely doable. This might actually not be too different from how it is done in international companies: you turn to the usual tools like telephone or video conferences, face-to-face meetings from time to time, a common working platform and the like. On top of that, cross-cultural competence is key for project management but also for the individuals in the project.

What have been the toughest moments in the project?

Obviously the co-design approach I was just talking about is an extremely complex undertaking. Hence, we had expected to face challenges from the beginning. It is really tough to develop and integrate hardware and software at the same time. If you have delays for instance in the hardware part of the project, this immediately also affects the software part. You have to come up with a mitigation plan and basically re-adjust the whole project plan. For the DEEP project that meant we also had to apply for an extension of
DEEP was one of the first three Exascale research projects funded by the European Commission back in 2011. It had an €18.5 million budget (EU funding: €8.3 million) and 16 partners distributed over eight European countries. Managing a project of this size is certainly no easy task. The ambitious prototyping efforts and the stringent co-design approach required to integrate all parts of the project – hardware, software and applications – made this challenging task even more complex. Project Manager Estela Suarez, Jülich Supercomputing Centre, shares how she has navigated the sometimes stormy waters and led the project to a successful outcome.

DEEP was one of the first three Exascale research projects funded by the European Commission back in 2011. It had an €18.5 million budget (EU funding: €8.3 million) and 16 partners distributed over eight European countries. Managing a project of this size is certainly no easy task. The ambitious prototyping efforts and the stringent co-design approach required to integrate all parts of the project – hardware, software and applications – made this challenging task even more complex. Project Manager Estela Suarez, Jülich Supercomputing Centre, shares how she has navigated the sometimes stormy waters and led the project to a successful outcome.

Now that the DEEP project is over, what are you most proud of?

There are not many projects out there as interesting and challenging as DEEP. I am certainly proud of having been able to contribute to its success. It has been fantastic to experience how the excellent interdisciplinary and international DEEP team realises and demonstrates a totally new architecture starting from the original “blackboard” discussions. Also, sharing our ideas, progress and results with the world and seeing the great interest they receive has reminded me over and over again of how fascinating the project is which I had the opportunity to manage. I will certainly miss it. Fortunately, our Exascale endeavour is not over yet: We have already started the sibling project DEEP-ER and are working on the next challenges.

What do you value most about your Exascale endeavour?

Definitely the holistic approach we pursue – or co-design, as we like to say. DEEP presents an architectural concept that spans all aspects from hardware to system software to tools to applications. This means we are working on quite a few subjects in parallel and there are quite a few researchers involved, everyone being eager to contribute to this bigger picture.
A DEEP CONNECTION
DEEP CONSORTIUM

Represented by the Jülich Supercomputing Centre (JSC); project coordinator and responsible for system software, installation and application optimisation

Provides semiconductor chips and platforms for the world-wide digital economy; leads the hardware prototyping effort

Industry partner, involved in hardware development and responsible for the integration of the prototype

Represented by the Computer Architecture Group, responsible for the development and integration of the EXTOLL interconnect and the GreenICE Booster

German national supercomputing centre, responsible for energy efficiency-related aspects and dissemination

Represented by the Department of Physics; shared its experience in warm water cooling systems design and operation

Leading supplier of end-to-end connectivity solutions for high-performance servers and storage that optimise data centre performance

Involved in developing the system software, especially the resource allocation mechanism

Spanish national supercomputing centre, leading OpenMP-related tasks and developing the DEEP Offload model

German SME further developing its ParaStation Cluster Suite, which forms the base of the DEEP software stack

French research centre involved as an application partner working on computational fluid dynamics

National supercomputing centre in Italy, involved as an application partner with a high temperature superconductivity code

One of two Federal Institutes of Technology in Switzerland, involved as an application partner with a brain simulation code

Contributes an atmospheric chemistry global circulation model as a pilot application for climate science on DEEP

Represented by the Centre for mathematical Plasma Astrophysics (CmPA) and involved as an application partner with a space weather forecast code

Global geophysical company in the oil and gas industry, involved as an application partner with a seismic imaging code
DEEP CONTINUED

**DEEP-ER INTO EXASCALE COMPUTING**

Within the given time frame of the project, DEEP has achieved tremendous success on the way to solving the most pressing challenges towards Exascale. The areas included energy efficiency, scalability, programmability and manageability, to name a few.

The successor project DEEP-ER (Dynamical Exascale Entry Platform – Extended Reach) builds on the basis laid out by DEEP and further evolves this innovative approach to heterogeneous computing. In focus this time are two significant Exascale computing challenges: highly scalable and efficient parallel I/O as well as resiliency. Co-design is again key to tackling these challenges and jointly developing solutions. All design decisions are guided by the Exascale requirements of the seven real-world HPC applications involved in DEEP-ER.

![Diagram](image)

**Leading-edge system development**

In terms of hardware, DEEP-ER further develops the Cluster-Booster architecture building a prototype that leverages advances in the hardware components, e.g. Intel Xeon Phi second generation processors. The next prototype employs EXTOLL as the unified network for both Cluster and Booster and experiments with the use of new storage technologies to extend the memory hierarchies. These innovative technologies include non-volatile memory (NVM) and network attached memory (NAM).

The ultimate goal is to create a hardware architecture that is utterly flexible and hence allows for easily upgrading or implementing entirely new technologies.

**Focus: I/O and resiliency**

The enhancements in terms of the Cluster-Booster architecture form the basis for and are closely connected to the software improvements geared towards highly scalable I/O and resiliency approaches.

DEEP-ER software experts are developing an efficient, user-friendly parallel I/O system tailored to the specific needs of large-scale HPC applications. Extensions to the Posix I/O API enable applications to efficiently use the file system and the different levels of the memory/storage subsystem. These extensions originate from Fraunhofer’s parallel file system BeeGFS, the parallel I/O library SIONlib, and Exascale10, a novel I/O concept developed by the Exascale10 Workgroup.
On top of the I/O system, a unified user-level checkpointing system with low overhead is being implemented, exploiting multiple levels of storage. The programming model developed in DEEP is extended in DEEP-ER to introduce easy-to-use annotations to control checkpointing. Additionally, traditional user-level checkpointing, which permits recovery of long-running applications, is combined with OmpSs’ capability to automatic re-execute failed remote-tasks, for instance those offloaded to the Booster side of the system.

Driver to Exascale: Applications

The requirements of HPC codes in terms of I/O and resiliency are the prime factor guiding the design of the DEEP-ER hardware and software components. Seven real-world HPC applications have been carefully selected to drive this co-design process. These include simulations on human exposure to electromagnetic fields, space weather, oil exploration, earthquake source dynamics, radio astronomy, high-temperature superconductivity and lattice quantum chromodynamics.

All codes are optimized for the prototype to demonstrate and validate the benefits of the DEEP-ER extensions to the Cluster-Booster Architecture.

DEEP-ER in a Nutshell

CONSORTIUM
- Coordinator: Jülich Supercomputing Centre
- 14 partners
- 7 European countries

BUDGET
- € 10 million (EU funding: € 6.4 million)

TIME PERIOD
- Oct 2013 – Mar 2017
Exascale research in Europe is one of the grand challenges tackled by the Seventh Framework Programme for Research and Technological Development (FP7). The declared aim is to respond to Europe’s needs in terms of jobs and competitiveness, and to maintain leadership in the global knowledge economy.

To date, eight projects represent the Exascale research efforts funded by the European Commission (EC) under the FP7 framework with a total budget of over € 50 million: CRESTA, DEEP and DEEP-ER, EPiGRAM, EXA2CT, Mont-Blanc (part I + II) and Numexas. The challenges they address in their research are manifold: innovative approaches to algorithm and application development, system software, tools and hardware design are at the heart of the EC funded initiatives.

Over the last four years, the projects have joined forces and started to develop a European Exascale Community to increase the visibility of European efforts in HPC on a global scale.

Internal cooperation
The collaboration started by exchanging and discussing ideas and concepts between individual researchers of the various projects. It did not take us very long, however, to realize how much we benefit from this interaction. Lifting these activities to the next level and developing a more formal framework for our collaboration was the logical consequence. Already in 2012, DEEP and Mont-Blanc teamed up to host a training workshop – and this was certainly not the last one to be jointly organised. In 2013, the cooperation was further intensified: the European Exascale Projects (EEP) Workshop series was born. Since then, the growing community has been meeting once a year for a two-day workshop to intensively discuss the status of projects, talk about lessons learned and this way identify synergies as well as cooperation opportunities.

Global outreach
Even though, our internal activities have proven to be extremely fruitful, we knew from the beginning that we wanted to interact with interested audiences on a larger scale. While in 2012 we started smaller activities like arranging Birds-of-a-Feather Sessions at conferences, in 2013 we decided to go big and organised our first joint European Exascale Projects booth at the world’s largest HPC conference and exhibition, Supercomputing (SC). Having been present with our joint booth at four exhibitions now (twice at ISC and twice at SC), we have gained high visibility for our projects on a European and on a global scale. Our cooperation has also enabled us to organise events tailored to more specific audiences, like the PRACEdays15, at which we hosted a satellite event on “Enabling Exascale in Europe for Industry”. There, we showed how industrial users can benefit from the research done in our projects by showcasing the experience from some of the application developers involved.

Shaping Europe’s Exascale Future
Looking back at the last four years, it is undeniable that the European Exascale Projects bring together experts from world-leading companies, Europe’s leading supercomputing institutions and outstanding academics to solve the challenges at Exascale. We have demonstrated vividly that Europe is a frontrunner in the global race to Exascale and how important collaboration within and between projects is in this context. The European community is soon to grow significantly, with 21 new Exascale projects funded via the Horizon 2020 framework starting at the end of 2015. Together we are eager to move Europe towards Exascale!
Over the last four years, the DEEP project has established itself as an integral part of the European HPC and Exascale Community. We asked key players in the field for their opinion on European Exascale efforts and the DEEP contribution to it.

Luis Carlos Busquets Pérez  
DEEP Project Officer, European Commission

“Modern societies increasingly face major challenges that involve processing enormous amounts of data and carrying out complex computations. High Performance Computing (HPC) is a powerful enabling tool helping societies to respond to these challenges in an effective way. Europe needs to stay at the forefront of HPC and keep up with its competitors. To achieve this, a wise investment into HPC research is needed. FP7 Exascale projects like DEEP and Mont-Blanc have demonstrated vividly how successful European HPC science and industry operates. We are on the right track and have to make sure we continue advancing in the right way and speed with the projects to follow.

Prof. Sanzio Bassini  
Chair PRACE Council

“In PRACE, it is our goal to provide European world-class scientists and researchers with a Research Infrastructure of persistent leading high performance computing services. To be able to continue these efforts, we need European research focusing on future computing challenges and on developing concepts and prototypes for cutting-edge HPC system technology. Contributions coming from the DEEP project and its follow-up DEEP-ER are very promising and we are confident that they will be fundamental elements for the evolution of the infrastructure of PRACE towards the Exascale challenge.

Prof. Arndt Bode  
Head of Leibniz Supercomputing Centre

“DEEP is a prime example of Europe’s ability to build leading technology prototypes. We at the Leibniz Supercomputing Centre are proud to have been part and to have contributed to one of the major research topics: energy efficiency. This is for sure one of the most limiting factors on the way to Exascale. It is of special importance to European HPC science and industry — yet also a field where Europe has lots of expertise to offer. Within the project energy efficiency was addressed in a holistic way — which might very well be leading also outside of Europe.”
More than 80 people from 16 different partners, distributed over eight European countries, more than 220,000 hours of work in the last three and a half years: the DEEP project has played a major role in the working life of many of the team members. Yet the project would certainly not have been such a success without the dedication shown by each and every person involved. This section captures our unique DEEP team spirit.

Julián David Morillo Pozo
Research Support Engineer,
Barcelona Supercomputing Center

“The individual commitment to the DEEP group effort was what made the project work. From the beginning everyone had a clear plan to accomplish the project objectives. I’m determined to apply what I have learned in whatever upcoming project I participate in.”

Anna Wolf
Application Support Engineer,
Jülich Supercomputing Centre

“Working in such a fascinating project with so many excellent partners from all across Europe was a great start to my working life after my Master’s degree. I got a good insight into the variety of applications. Supporting the developers in enhancing their code and porting it to the new system was an exciting challenge.”

Michael Ott
Head of Hardware Labs,
Leibniz Supercomputing Centre

“DEEP was like a huge playground: we could come up with an entirely new framework to make supercomputers more energy efficient – and it looks really promising. This would not have been possible without the input of so many skilled people inside and outside of the project.”

Mauro Rossi
Chief Engineer,
Eurotech

“The challenges I had to deal with during the early conceptualization phase of the project were extremely exciting: finding technological solutions to fit all electronics, cooling and interconnections in such a compact volume was an intriguing and intricate puzzle to solve.”
DEEP INVOLVEMENT

Pramod Kumbhar
HPC Engineer, Blue Brain Project,
École Polytechnique Fédérale de Lausanne

“Simulations of morphologically detailed brain models are computationally challenging and will need capabilities at the Exascale. DEEP is an interesting co-design project which helped us to develop and prepare the application for future heterogeneous architectures.”

Andreas Galonska
Software Engineer,
Jülich Supercomputing Centre

“It was very exciting to exploit cutting-edge technologies with innovative software solutions. The great teamwork with our international partners was very fruitful and the basis of the success of our project.”

Thomas Moschny
Chief Technical Officer,
ParTec

“Being part of a team of researchers and HPC specialists from all over the EU developing a radical new architecture on the way to Exascale has been an outstanding experience. The joint effort by all partners led the project to great success.”

Sabrina Eisenreich
PR Manager Research Projects,
Leibniz Supercomputing Centre

“Joining the team only midway, I was immediately struck by the positive spirit and the tremendous dedication to the project my new colleagues showed. Being part of DEEP was of course always guided by the will to make this ambitious Exascale endeavour a success. Yet it was also great fun to work with this skilled and motivated international group of people.”

In Memoriam

The whole DEEP team would like to express our heartfelt, deep, sincere sympathy and condolences on the sudden and unexpected passing of our colleague Dr. Alec Johnson in December 2014.

We all have come to know Alec as an extremely passionate, inspiring and hard-working colleague and scientist. But he was missed in the project most of all for his kind, friendly and caring personality.

Our thoughts are with his family.

The DEEP Project Colleagues
IMPRINT

Editors
Damian Alvarez Mallon
Axel Auweter
Vicenç Beltran
Carsten Clauss
Norbert Eicker
Sabrina Eisenreich
Judit Gimenez
Hans-Christian Hoppe
Sabine Höfler-Thierfeldt
Jochen Kreutz
Thomas Lippert
Thomas Moschny
Ina Schmitz
Estela Suarez
Igor Zacharov

Graphic Design
die maschinerie GmbH

Photos
DEEP Project (p. 5, 6 bottom left, 7, 8, 9, 14, 20, 21, 22 bottom left, 24)
Jülich Supercomputing Centre (p. 3)
Axel Auweter (p. 7 top left)
Dirk Frey (p. 6 top right)
Hans-Christian Hoppe (p. 6 top left, 18, 22 bottom right)
Nils Krebs, NK Film (p. 4)
Mai-Linh Doan (self photo), Wikimedia Commons, licensed under CreativeCommons-license CC-BY-SA-3.0, URL: http://creativecommons.org/licenses/by-sa/3.0/ (p. 12 middle)

The copyright for the application images on pages 10 to 12 and the portrait pictures on pages 23 as well as 25/26 lie with the respective individuals or DEEP partners. The DEEP Project Management would like to thank all colleagues who contributed images to the brochure.

Contact
DEEP Project Management Team
Jülich Supercomputing Centre
Forschungszentrum Jülich
52425 Jülich
Germany
E-mail: pmt@deep-project.eu
Website: www.deep-project.eu

Copyright
DEEP Project
September 2015

The research leading to these results has received funding from the European Commission’s Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n° 287530.