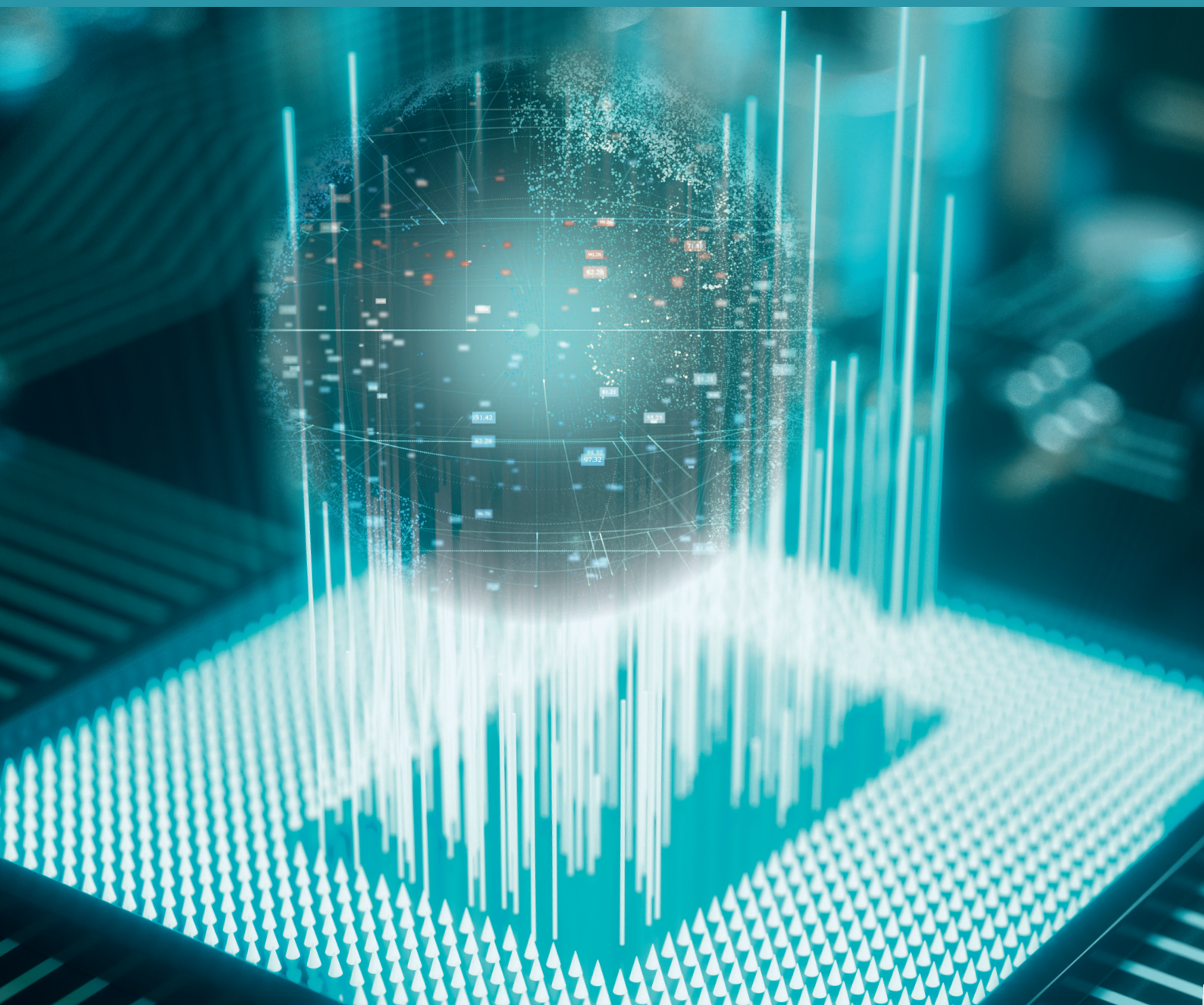
Fraunhofer
ITWM

FRAUNHOFER INSTITUTE FOR INDUSTRIAL MATHEMATICS ITWM

COMPETENCE CENTER HIGH PERFORMANCE COMPUTING



COMPETENCE CENTER HIGH PERFORMANCE COMPUTING

DR. FRANZ-JOSEF PFREUNDT
HEAD OF DEPARTMENT

ENERGY EFFICIENT COMPUTING IS HIGH PERFORMANCE COMPUTING

Our IT systems are among the largest single consumers of energy and emitters of CO2, with energy consumption continuing to rise sharply. Currently, they account for five to ten percent of electricity consumption, but this figure is expected to grow to twenty percent. This is then the order of magnitude that completely electrified passenger car traffic would take. In most cases, the focus is then on "green" data centers, which are powered by electricity from renewable sources and where attention is paid to efficient cooling technology. However, the much greater potential lies in software and how software is used on which processors, and thus in its impact on the vast rest of IT systems.

In high-performance computing, energy costs are already a decisive factor in hardware procurement; however, the efficiency of the software used is even more important. Software that does not take advantage of the parallelism of modern processors and their architecture quickly loses an order of magnitude in energy efficiency here as well. For us at CC HPC, High Performance Computing means the use and development of highly optimized software on suitable hardware. Already in 2008, the Pegasus system of the ITWM was not only among the TOP 500 of the fastest HPC systems, but also number 1 in the Green 500! The software we developed for it was orders of magnitude more efficient than the original customer software. In that case, the saved energy costs of one year alone would have financed the Pegasus computer and the development of the software. Our experience in many industrial projects since then shows that unoptimized software can often be improved by at least one order of magnitude.

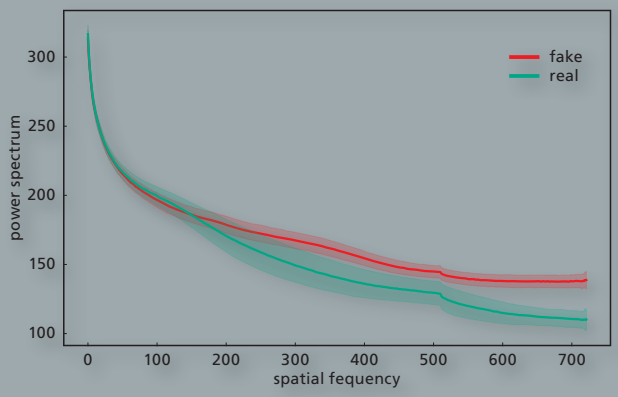
We are now the specialists in energy-efficient programming, green computing. However, it must become easier for everyone to write efficient software. The STX processor, which we developed as part of the EPI project, brings us a great deal closer to this goal. Its design is such that it is easy for a large class of algorithms to achieve high energy efficiency and thus cost efficiency with the support of the compiler. Energy Efficient Computing today means the perfect exploitation of parallelism, optimal data transport as well as suitable algorithms in combination with the right hardware. This holistic approach to Green Computing is the core of our self-image and motivation for the employees of the Competence Center for High Performance Computing.

Contact
franz-josef.pfreundt@itwm.fraunhofer.de
www.itwm.fraunhofer.de/en/hpc

MAIN TOPICS

- Green by IT
- Fraunhofer Parallel File System (BeeGFS)
- Visualization
- Seismic Imaging
- Data Analysis and Machine Learning
- Scalable Parallel Programming

1



2

DEEP FAKE DETECTION: MATHEMATICS DEBUNKS IMAGE MANIPULATION

1 *Synthetic faces; generated using the approach of Karras, Tero, Samuli Laine, and Timo Aila, presented in "A Style- Based Generator Architecture for Generative Adversarial Networks".*

2 *ITWM Deep Fake Detector: according to the mathematic mapping, real and fake images can be easily distinguished.*

In recent years, AI research has made impressive progress. Particularly in the area of generative models, which not only evaluate data, but can also generate realistic-looking synthetic data. Unfortunately, these also carry the risk of misuse. The Deep Fake Detector, which we co-developed, shows how these visual fakes can be detected.

Generative models, also called Generative Adversarial Networks (GANs), have been a breakthrough in AI research based on deep learning algorithms. Unfortunately, GAN technology also has its negative sides: Shortly after their introduction, GANs were used to manipulate image and audio data. Fake but deceptively real images and videos of celebrities and politicians circulated on the Internet. These manipulations became known as "deep fakes." Our illustrations show some examples of synthetic faces that are almost indistinguishable from real faces for human observers.

Collaborative research against loss of trust

Due to the expected societal impact that would come with a complete loss of trust in potentially manipulated image and audio data, teams of researchers around the world have set out to develop algorithms that can automatically detect "deep fakes." While most current approaches try to revert to using learning algorithms to detect the manipulations, we took a different approach with a group of researchers from the University of Mannheim and the University of Offenburg: Analyses revealed that GANs make inherent errors when generating images. Although these are barely visible to the human eye, they are very easy to map mathematically in Fourier space.

The Deep Fake Detector determines

The new method developed at the ITWM has several advantages over existing methods:

- The detected GAN error is systematically conditioned; therefore, it is theoretically impossible for current GAN architectures to learn to bypass the detector.
- Very little sample data is needed to reliably detect deep fakes.
- The method is easy to implement and requires comparatively little computing power.

In an initial evaluation on public test data, the new approach achieved an accuracy of almost one hundred percent.

DEEP TOPOLOGY LEARNING: AUTOMATIC DESIGN OF DEEP NEURAL NETWORKS

The development of the last years shows that machine learning and especially the sub-area of deep learning will be a significant building block in the future, both in the scientific as well as in the industrial field. In the BMBF project "Deep Topology Learning", we are working together with several universities to accelerate design algorithms.

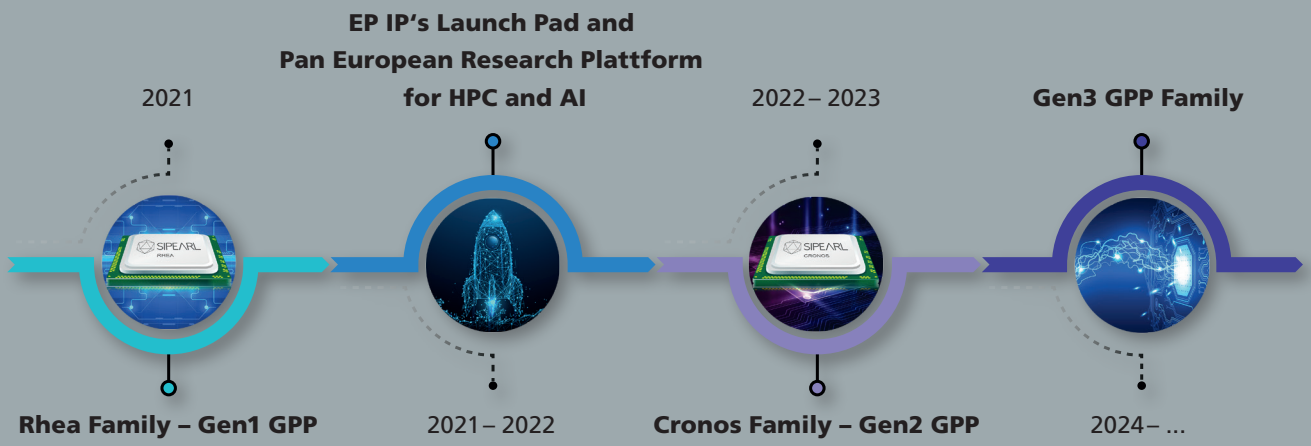
From speech recognition and automatic image analysis to prototypes of autonomously driving cars or "Go"-playing algorithms at world champion level: so-called deep learning methods are almost always behind the success stories. This family of learning methods typically uses over parameterized and usually very large artificial neural networks (DNN) to model the learning problems. Training such networks requires not only very large amounts of data, but also enormous computing power. Despite the sometimes impressive results achieved with DNNs, they still have some disadvantages that currently often hinder their widespread use in practice. In addition to the typically very large amounts of data required, this is primarily the time-consuming development process.

Design algorithms replace trial and error

Designing new, problem-specific network topologies is a very time-consuming and computationally expensive process. Until now, the development of new deep learning solutions has been done in a heuristic and experience-driven "trial and error" approach. The goal of the BMBF project "Deep Topology Learning" (DeToL) is to decisively accelerate and simplify this design process for deep learning solutions using automated, data-driven design algorithms.

Next step: automation

The application area of Deep Learning is broad. From machine vision and autonomous driving to speech recognition, music generation or art, Deep Learning is increasingly making significant contributions to development. At the same time, the amount of experts who could design deep neural networks for a given application area is limited. A logical next step is to maximize the degree of automation in the development of network architectures. Since this automation is very computationally expensive, this is where HPC systems come in. Human interaction is limited to designing a search space of all possible topologies for a given problem. Based on a given search strategy, the architecture is then optimized in the next step.



1

EUROPEAN PROCESSOR INITIATIVE EPI

1 *The EPI Roadmap*

High-performance computing is one of the key technologies for future value creation in many areas. For example, it is used for simulations of weather and climate as well as new materials and in the use of artificial intelligence. Developments from high-performance computing can be found in autonomous vehicles, in the context of Industry 4.0 and in cloud computing.

In all these areas, powerful processors are needed to process the enormous volumes of data. Only a few manufacturers are successfully producing such processors. The international market is dominated by companies from the USA and recently also from China.

Strengthening European independence

To ensure European independence in this key technology, the European Processor Initiative (EPI) brings together 26 partners from industry and academia to design European processors and build a hardware and software ecosystem in Europe. The goal is to develop a general-purpose processor as well as accelerator processors optimized for specific application areas. In this area, we are engaged in collaboration with Fraunhofer IIS and ETH Zurich to design an accelerator for stencil and tensor computations (STX). Stencil-based applications are very common in HPC simulations, tensor computations are the basis of Deep Learning.

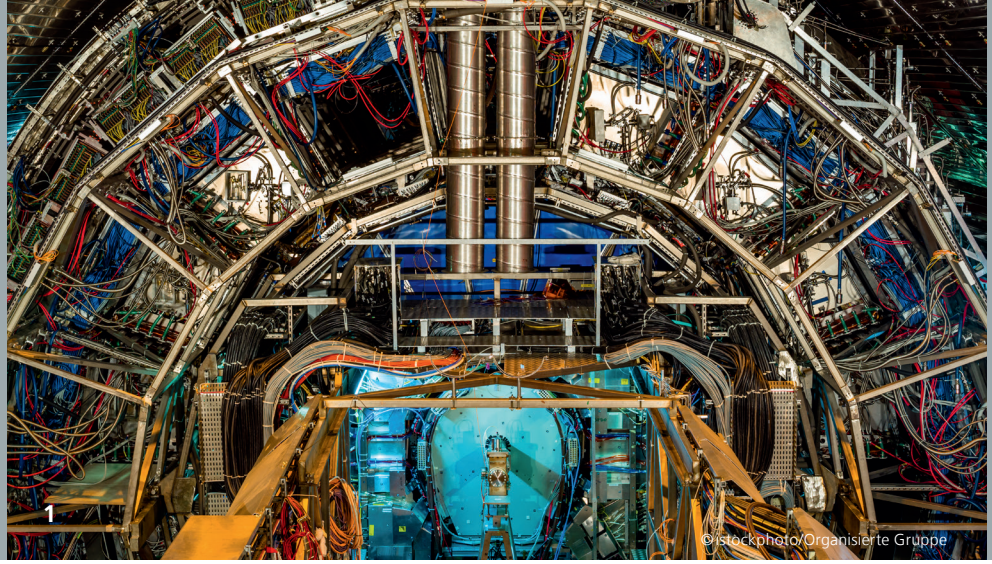
Developing hardware and software simultaneously

We want to build highly energy-efficient hardware that can be programmed in a comfortable way for end-users and is not limited to one application area, e. g. Deep Learning. To achieve this, we use what is known as hardware software co-design. Here, the software is developed and optimized in parallel to the hardware. During this process, requirements of the software can be incorporated into the hardware design and the software can be optimized to the hardware. The selection of the software consists of important mathematical calculation kernels such as matrix-matrix multiplication, convolution filters and various 2D and 3D stencils. For development, we are using the open RISCV Instruction Set Architecture, so there are no licensing costs or dependencies.

The first chip is scheduled for production in early 2021 to test energy efficiency and performance. Furthermore, we are optimizing our HPC software stack with the communication library GPI and our parallel file system BeeGFS for the EPI hardware and support the development of the general-purpose processor with another co-design process.

European Processor Initiative

www.itwm.fraunhofer.de/EPI_en



ENORMOUS ACCELERATION WITH GPI-SPACE

The increasingly accurate interpretation of the very large amounts of data produced by experiments at the Large Hadron Collider (LHC) requires the use of new mathematical tools, which are the subject of current research. A successful approach also relies on symbolic computations for the analytical reduction of complex structures. In doing so, mathematics is entering the innermost realm of scientific computing, both in terms of the sheer size of the tasks and in terms of issues of comprehensibility, reproducibility, and automation.

1 *ATLAS particle detector at the Large Hadron Collider of the European nuclear research center CERN*

The "GPI-Space" tool developed in our department is designed to meet these requirements. It allows to efficiently manage and process data on very large machines. Virtualization of memory, proven schedulers, robust error handling, scalable resource management and Petri net based dependency graphs are integrated in GPI-Space.

Answers to open questions

Together with the AG Algebra, Geometry and Computer Algebra of the Department of Mathematics at the Technical University of Kaiserslautern, the worldwide leading computer algebra system "Singular", which was developed there and is the world's leading computer algebra system, was connected to GPI-Space and the resulting conglomerate was used to obtain answers to previously unanswered questions. For this purpose, the TU group has represented the structure of its mathematical methods in Petri nets in several projects and used this representation to scale the concrete calculations to large machines with GPI-Space. Speed-ups by a factor of 400 with simultaneous parallel efficiency of 80 percent are results that were previously considered unattainable.

We were able to generate new requirements for GPI-Space from this collaboration and in part already provide solutions in GPI-Space. For example, many subtasks are based on calculations of Gröbner bases, whose runtime can only be predicted very poorly for concrete inputs. Taking different paths to possible solutions at the same time helps to find solutions faster. Afterwards, the results of some of the already started calculations are no longer needed; GPI-Space supports the controlled abort of these calculations in the latest version.

Based on the very positive experiences, we have planned the next steps for a deeper integration of the two systems GPI-Space and Singular together with the WG Algebra, Geometry and Computer Algebra and started the implementation.

GPI-SPACE

1

2

ENERGY COOPERATIVE OPERATES ITS OWN MICROGRID WITH AMPERIX

1 *Peak shaving in operation. Due to the use of distributed storage, the grid consumption of the energy community (orange) does not exceed the given limit.*

2 *The floating housing estate Schoonschip in Amsterdam North.*

A new form of energy supply is being realized in the north of Amsterdam: Thirty floating houses are almost energy self-sufficient and together form the most sustainable floating residential quarter in Europe. Our Amperix® energy management system is also being used there.

The idea: 30 houses with 47 apartments form an energy unit that generates most of its own electricity using solar power, stores it using heat pumps and batteries, and makes it available to the residents. The houses are thus interconnected and require only a single connection to the municipal power grid for the entire residential project. This grid node is used to back up Schoonschip during peak demand periods, including when battery storage is empty and there is little sunshine. With a maximum power of 175 kW, the grid connection point is relatively small compared to the number of fully electric residential units.

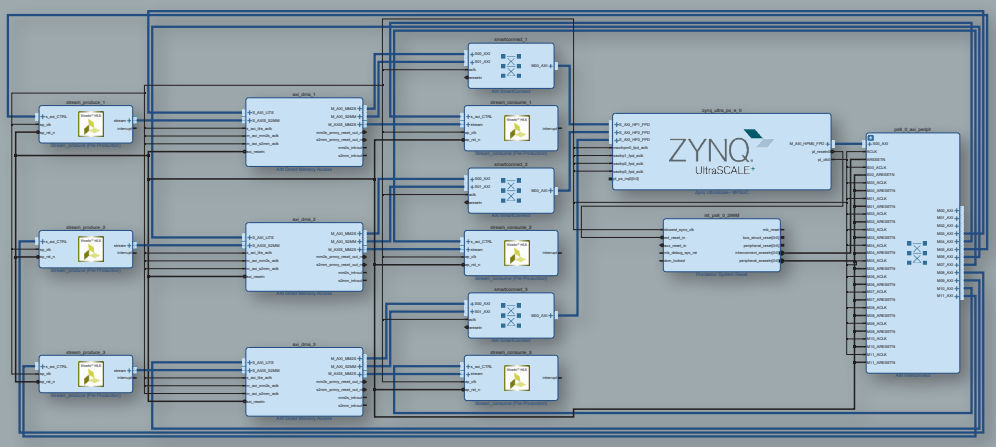
Amperix enables sector coupling

This is where the Amperix® energy management system developed at the Competence Center High Performance Computing comes into play. The Amperix® is a tool for microgrids and energy communities. In addition to controlling electricity storage, the Amperix® also implements sector coupling here. This means that heat storage in combination with heat pumps and charging stations for electric vehicles can be flexibly controlled according to the availability of renewable energies. This way, load shifting within the residential community is optimized and consumption peaks are flattened (so-called peak shaving).

The joint control of the energy community was already able to prove itself in the winter of 2019/2020. Simultaneously occurring load peaks of the heat supply and the electricity supply could be reduced by means of the coordinated feed-out of the battery storage systems and thus a stable energy supply could be guaranteed.

Continuation as H2020 lighthouse project

Due to the great success, the research and development around the innovative housing and energy concept will be continued in the ATELIER Smart City project. The aim of the project is to implement citizen-oriented "Positive Energy Districts" in the lighthouse cities of Amsterdam and Bilbao. The European Commission under Horizon 2020 funds the ATELIER project.



1

FPGAs AS ACCELERATORS: THE ALGORITHMUS AS DIGITAL CIRCUIT

Field Programmable Gate Arrays (FPGAs) are currently experiencing a renaissance as accelerator hardware. On an FPGA, complex functions can be realized directly as a tailor-made digital circuit. FPGAs thus combine the performance of specialized hardware with the flexibility of software and are increasingly being used as accelerators in high-performance computing systems.

1 *Block representation of an FPGA design*

As a kind of modular system for digital circuits, FPGAs enable basic electrical components to be combined with each other to create the desired functionality. The interconnection of the physical elements on the FPGA is loaded as a configuration and can thus be changed at will in seconds. While application development for FPGAs was long the preserve of hardware specialists, today's tools make it easier for a broader range of software developers to get started.

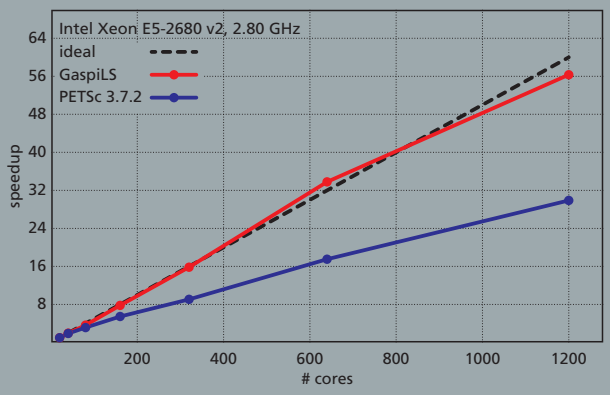
Energy-efficient processing of seismic data

As part of the EuroEXA project, we are porting seismic data processing software to FPGAs. Seismic methods compute an image of the Earth's subsurface from data measured on the surface, allowing geoscientists to discover oil/natural gas deposits. We expect computation on FPGAs to improve performance or energy efficiency compared to CPUs or GPUs. To port the computationally intensive parts of the program, we tune algorithms and data structures to the specific capabilities of the hardware. Performance can be optimized by maximum parallelization within one FPGA and by parallel computation on multiple FPGAs. Parallelization across multiple FPGAs requires efficient data communication, which is implemented using the GPI programming model we developed at the Competence Center High Performance Computing.

FPGAs enable specific circuit design for arbitrary number formats

Another application of interest is deep neural networks (deep learning), as used in machine learning. Deep learning algorithms are particularly well suited for accelerators because the computational patterns are very homogeneous. They are also robust to low precision in number representation. While the hardware in CPUs and GPUs can directly perform computations only with predefined number formats, the circuits on the FPGA can be designed specifically for arbitrary number formats. Usually FPGAs are used to execute already trained networks, we deal with the training of deep neural networks on FPGAs.

1



2

GaspiLS AND THE GPI-2 ECOSYSTEM: GPI-2 SCALABILITY AND PERFORMANCE MADE EASY

1 *GaspiLS provides scalability for FEM and CFD simulations*

2 *Performance plots: Scalability advantage of GaspiLS compared to PetSC. Jacobi preconditioned Richardson Method; 3-D Poisson equation (2d order FD discretization), cubic matrix (359^3)*

The distributed systems used in high performance computing require highly efficient and scalable applications. Scalability is a measure of the efficiency of a parallel implementation and ultimately indicates whether the available resources – for example, CPUs – are being efficiently used. The Competence Center High Performance Computing develops GPI-2, a parallel programming model that is ideal for implementing such applications.

GaspiLS is a library of scalable, iterative linear solvers developed to easily exploit the benefits generated by GPI-2 and make them available for immediate practical use in a multitude of applications. GaspiLS is ready for direct use with a variety of new or existing simulation programs ultimately solving linear systems.

HySCALA explores new areas of application and new markets for GaspiLS

GaspiLS has already proven itself in several industry projects. Its further distribution is currently being promoted as part of the EU project HySCALA (Hybrid Scalable sparse matrix linear algebra for industrial applications).

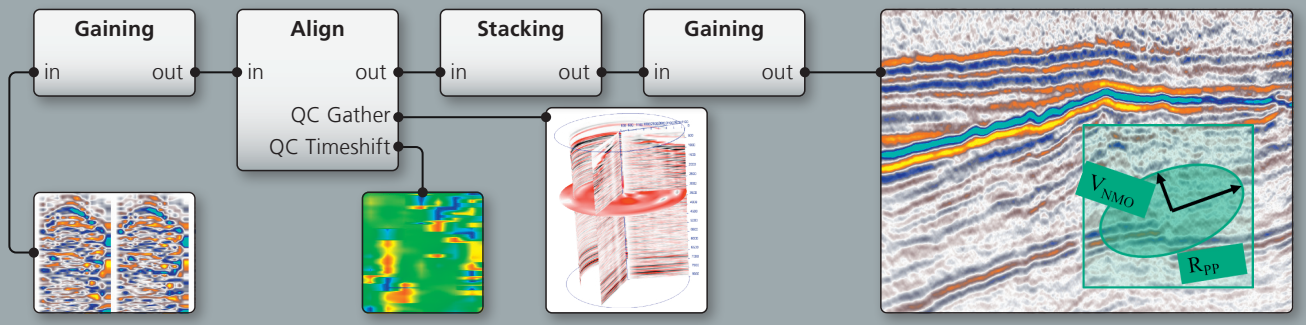
The aim is to analyze various, potential new market segments and fields of application for GaspiLS and to identify specific requirements for a competitive linear solver library. We are looking primarily for generic yet efficient preconditioners that allow us to reduce the number of iterations required for convergence of the iterative process and minimize the total run times. Presently, the focus is on the scalable implementation of efficient preconditioners that can be applied to a broad class of problems.

GaspiCxx for increased productivity

Within GaspiLS, we have factorized the implementation for the explicit management of communication resources required by the GPI-2 data transfer and used GaspiCxx to supply it to other applications. GaspiCxx defines an easy to use C++ interface. It delivers the full native GPI-2 performance. At the same time, the management of GPI-2 communication resources is fully transparent to the application.

This eliminates a large part of the implementation work normally required to develop GPI-2 applications. Development of GPI-2 applications and the exploitation of the advantages – like the good scalability – has never been so easy.

www.itwm.fraunhofer.de/GaspiLS_en



1

ALOMA LETS GEOSCIENTISTS FOCUS ON THEIR FIELD OF EXPERTISE

ALOMA lifts the burden of dealing with parallelization, multi-threading, and other challenges in high-performance computing from its users. Instead, the experts for geophysical questions can focus on their area of expertise while ALOMA takes care of efficiently executing their algorithms even on large scale and heterogeneous systems.

1 *Presentation of a simple workflow with ALOMA: Inbound gatherers are corrected and then stacked. Input and results can be visualized interactively.*

The software is a specialized version of GPI-Space which is widely used in fields beyond geophysics such as big data and machine learning.

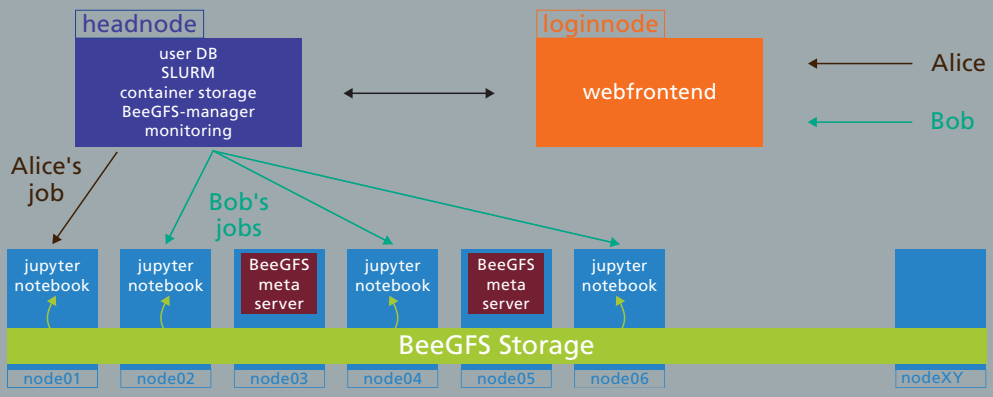
Complex computations on ever growing amounts of data are characteristic for the geosciences and thus geophysicists are forced to learn about HPC techniques in order to make their software run efficiently on large scale systems. We developed a system that sits in between the geophysicist and the HPC expert. Computer scientists and geophysicists together came up with ideal strategies for parallelization, data partition, and failure tolerance in the context of geophysical applications.

The heart of ALOMA, its failure tolerant runtime system to execute workflows on distributed systems, was then developed by the HPC experts of our group. For its users, the geophysicists and geologists, ALOMA is merely a black box in which they can integrate their latest developments via a well-defined interface. The learning curve for the new approach is easy to manage. Once ALOMA is installed, it takes users not more than a day to port their first module to the new system.

Quick prototype development and scaling

The main benefit of ALOMA is to quickly integrate and test newly developed algorithms and prototypes on production scale real-world problems in no time. Furthermore, existing codes and applications – even in different programming languages such as C/C++, Fortran, Matlab etc. – can be integrated as modules in ALOMA. With a graphical editor, users can combine these modules into workflows and let the software deal with the automatic parallelization and execution.

We were able to prove the feasibility of this concept in various projects with partners in the oil and gas industry, where we managed to make a customer software scale within a few days. The concept is so convincing that a Houston based company has commissioned us with switching over their existing processing software to ALOMA.



1

HPC FOR MACHINE LEARNING: CARME

1 *Simplified scheme of the most important system components and their connections*

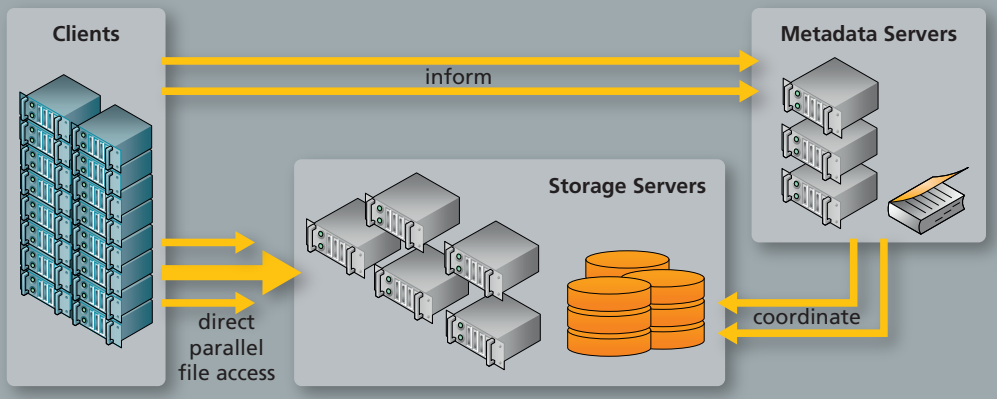
Machine learning has an increasingly higher priority in both scientific and industrial enterprises. This is evident from the investment in new, above all, GPU-based hardware – from simple desktop computers to high performance computing clusters. Computing clusters are used in Data Analysis (DA) and highly complex Machine Learning (ML) systems to process and simulate very large amounts of data – to include even the human brain.

Machine learning in HPC clusters presents certain challenges. The procurement of the individual hardware components is the least of these challenges. The biggest questions arise subsequent to that acquisition:

- How to manage existing resources?
- How to make an application scalable to several GPUs?
- How to solve the challenge of data storage and continuous upload to the program?
- How to train users to effectively use the hardware?

The answers to these questions begin with our open-source software stack Carme. The basic concept is to combine the world of machine learning and data analysis with the world of HPC systems. We achieve this using established ML and DA tools with HPC back ends. Specifically, we use a variety of HPC and ML technologies. Some of these technologies are developed in this department, for example, the highly reliable parallel file system BeeGFS for fast data links.

Carme combines the worlds of machine learning and HPC clusters. ML is a steady and fast growing field of technology. This new agility challenges data centers to provide very different applications for single users. It is not enough to have one user interface for the user; rather there must also be a guarantee of a seamless integration of this surface in existing and emerging clusters. To make clusters attractive to ML and DA users, an intuitive software environment must be provided to the clusters. Interactive management of the cluster is essential in the development of ML applications. Users must have the opportunity to use tools they are familiar with on a complex HPC cluster, making it easier for them to migrate to and use the cluster.



1

BeeGFS – THE FILE SYSTEM FOR BIG DATA AND AI

The success of current AI technologies such as neural networks is based on the increased power of today's processors – mostly GPU´s – but, above all, on the availability of very large amounts of data. For example, new medical devices, autonomous vehicles, and genome analyses supply ever more fine resolution data in quick succession forming the basis for future AI solutions. Developed at ITWM and distributed by ThinkparQ, the parallel file system BeeGFS (also known as Fraunhofer Parallel File System – FhGFS) helps in mastering the large data volumes with a very flexible software solution.

1 *BeeGFS architekture*

BeeGFS is a parallel file system where storage capacity as well as read and write speeds grow linearly with the number of linked storage units. As a pure software solution, it can be flexibly installed both on existing hardware and on the latest, superfast flash-memory systems. In addition to very good scalability, our system development team attaches great importance on easy handling and a high degree of flexibility for a variety of potential use cases.

BeeGFS on NVMe

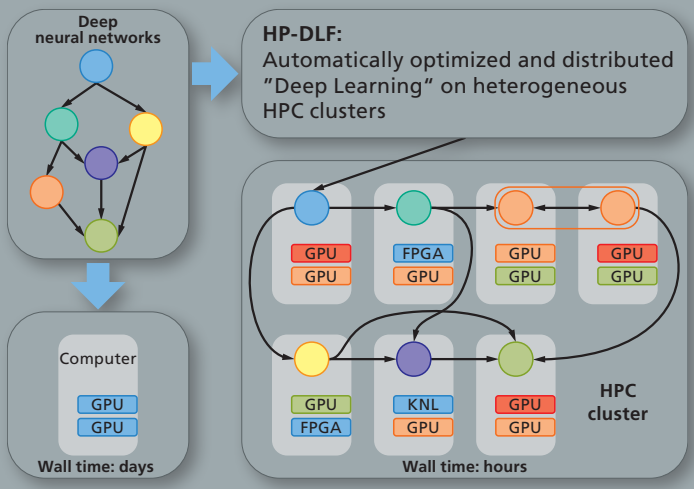
Training deep neural networks (Deep Learning) demands that existing data be provided several times very quickly to the computing units. Most external storage systems are hardly suitable for this task, so the data is cached directly to the computer servers on fast local systems (NVMe). Since these have relatively small capacities, the need arises for data to be distributed on several units in parallel.

The BeeGFS software system is specially optimized for high speed requirements even with a large number of files and this ability is its biggest strength. BeeGFS can be installed directly on the computer servers and is scalable to high I/O rate of 1 TByte/sec and more. Japanese AI researchers were convinced: BeeGFS is now successfully deployed on the two major Japanese AI systems TSUBAME 3.0 (HPE) and AI Bridging Cloud Infrastructure (ABCI, Fujitsu).

Open-source license

The software is distributed with an open-source license and source files are provided on the BeeGFS website. A spin-off of Fraunhofer ITWM, ThinkparQ, supplies worldwide commercial support for BeeGFS and manages further development from a customer perspective. The joint development team also successfully applies its extensive knowledge in several EU funded projects that focus on the use of BeeGFS on future Exascale computing systems.

BeeGFS®
developed by Fraunhofer



1

CALCULATE COMPLEX MODELS FASTER WITH TARANTELLA

1 *Scheme "High Performance Deep Learning Framework"*

Artificial neural networks are becoming more and more important. When using social media or online stores, people already benefit from the latest breakthroughs in speech and image processing, which are the result of the training of these neural networks. The areas of application are diverse: whether in cosmology, climate research or particle physics – science also uses the large neural networks. For the scalable training of Deep Neural Networks on supercomputers, the framework "Tarantella" comes into play.

In recent years, considerable progress has already been made in the field of machine learning through the further development of so-called »deep learning« algorithms. The neural networks required for these achievements are becoming larger and more complex, which is why an enormous amount of computation and training data is required for their development and training. Therefore, in the BMBF project "High Performance Deep Learning Framework - Software Environment for the Efficient Design of Deep Neural Networks on High Performance Computers", the open-source framework Tarantella was developed: it enables the training of artificial neural networks on high-performance computers.

New possibilities through Tarantella

By using high-performance computers, the time-consuming process of training neural networks can be accelerated. More data is sent through the network and the artificial neurons are activated more frequently – in this way, as the logo of the framework suggests, a neural network is woven. In addition, training of arbitrarily large neural networks is supported, allowing more complex mathematical functions to be learned and even more difficult problems to be solved.

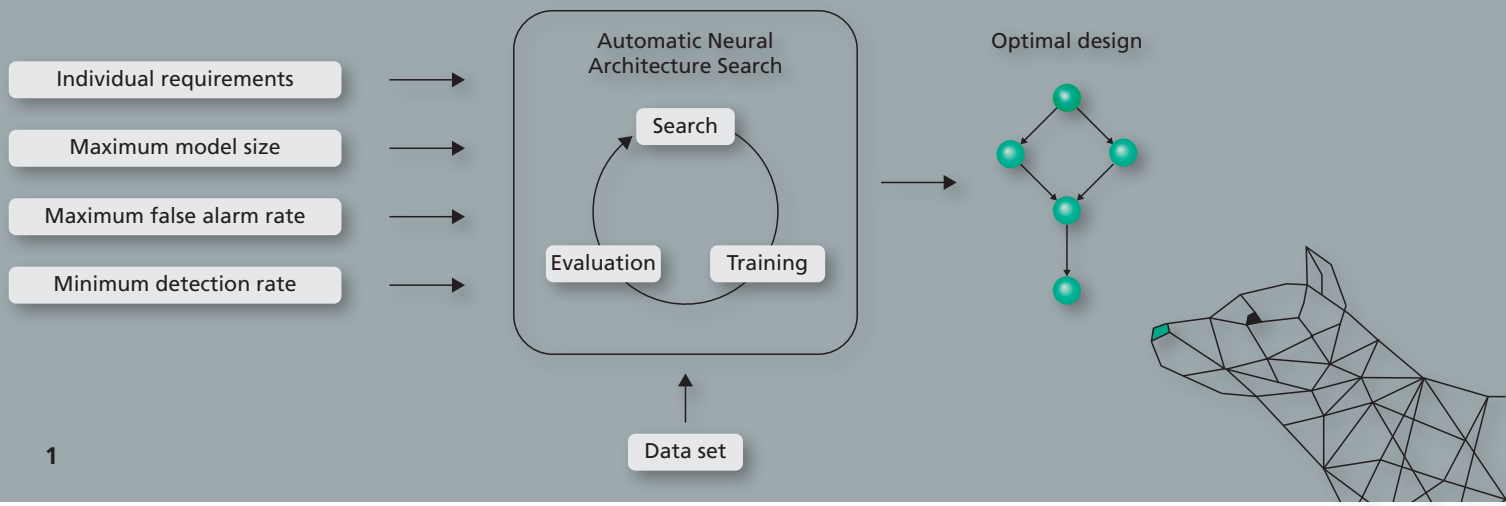
Tarantella builds on advanced technologies from the field of Deep Learning and High Performance Computing. Based on "TensorFlow", the most widely used deep learning platform in research and production, the framework is intuitive to use. Existing AI applications can be ported to run on a high-performance computer with minimal changes thanks to Tarantella.

Support for the development of the framework

The German Research Center for Artificial Intelligence DFKI in Saarbrücken was among those involved in the development of Tarantella; it dealt with the question of dividing or partitioning the neural network among the various servers of the supercomputer. The Center for Information Services and High Performance Computing ZiH in Dresden supported the performance evaluation of the software and the University of Heidelberg dealt with application scenarios for particularly large neural networks.

TARANTELLA

www.tarantella.org



1

EFFICIENT HARDWARE FOR ARTIFICIAL INTELLIGENCE

Machine Learning achieves remarkable performance in tasks such as computer vision and natural language processing. This technology enables novel, disruptive innovations for the edge and the embedded market, for example security, autonomous driving and medicine. However, one of the prevailing problems is the high demand in computational resources of the machine learning algorithms. In order to find the best feasible solution, multiple design constraints and objectives for both the application and hardware must be considered, such as model accuracy, energy consumption, latency, etc.

1 *NASE – Neural Architecture Search Engine*

Hardware aware Neural Architecture Search (NAS) can automatically find Deep Neural Network (DNN) models which are not only optimal for the application, but also for the hardware performance. However, NAS has a major hurdle, which is the extremely vast search space of potential DNN candidates. We combine evolutionary, genetic algorithms with Bayesian selection strategies for Pareto optimal solutions, which traverse the search space intelligently and efficiently. Nevertheless, a considerable amount of DNN models need to be trained and evaluated.

Therefore, on top of the mathematical knowledge, we employ the high-performance computing know-how of our department to leverage the power of computing clusters with our NAS. We use our dynamical runtime scheduler GPI-SPACE together with our Python frontend DART to parallelize not only the training and evaluation of the DNN models, but also more computation intensive hardware-awareness tasks. Our software tool, the neural architecture search engine (NASE), is able to scale with complex and demanding applications, so that hardware efficient neural networks can be found automatically without DNN know-how.

With our solution, we entered in a competition initiated by the German Federal Ministry of Education and Research (BMBF) with the goal of accurately detecting atrial fibrillation in ECG data with as little energy consumption as possible using FPGAs as the underlying hardware technology. We participated in this competition together with our partners, the department of microelectronic system design at the TU Kaiserslautern. A distinguishing feature of our solution is that we incorporate the architecture design for the FPGA hardware into the NAS, so that both the DNN model and the hardware architecture are designed in the same co-design process. Compared to a hand-crafted model, we reduced the energy consumption and model size by many orders of magnitude. Our approach, the holistic automatic machine learning for FPGAs (HALF), achieved the best energy efficiency in the competition and was awarded first prize.

Contact
Fraunhofer-Institut für Techno- und
Wirtschaftsmathematik ITWM

Fraunhofer-Platz 1
67663 Kaiserslautern
Germany

Telefon +49 (0) 631 / 31600-0
Telefax +49 (0) 631 / 31600-1099
E-Mail info@itwm.fraunhofer.de
www.itwm.fraunhofer.de
www.itwm.fraunhofer.de/en/hpc