Interview with Utz-Uwe Haus on "High Performance Computing in Economic Environments"

This interview is part of the special issue (01/2020) on "High Performance Business Computing" to be published in the journal *Business & Information Systems Engineering*. The interviewee Utz-Uwe Haus is Senior Research Engineer @ CRAY European Research Lab (CERL)). A bio of him is included at the end of the interview.

Citation: *Schryen, G., Kliewer, N., Fink, A. (2020), Interview with Utz-Uwe Haus on "High Performance Computing in Economic Environments", Interview, Business & Information Systems Engineering, forthcoming.*

Question: High Performance Computing (HPC) has become established and indispensable in many application and research fields, supporting weather forecasting, simulation in computer-aided engineering, Alzheimer's research, molecular modeling, and many more fields. How can HPC add value to solving problems in business administration and economics?

Answer: While "Big Data" has become a topic of interest in the digital economy in recent years, HPC has been doing Big Data since its inception half a century ago: The data deluge is not a new phenomenon, and has been dealt with in scientific computing very successfully: Classic HPC customers, such as the European Center for Medium-Range Weather Forecast (ECMWF), have been running forecasts round-the-clock and every day for 40 years, and all data sets ever produced (around 330 of Petabytes in 2019, with an additional 200 Terabytes added each day) are available at the touch of a fingertip. As business administration and economics move beyond the database- and ERP-system driven IT infrastructure into an age of cross-silo and unstructured data, with product data coming back to the producer or distributor from IoT- and sensor-enabled devices, the computing and storage capabilities of HPC systems are becoming a necessity – if not for the business itself then for dedicated service providers offering dedicated business analytics services.

Question: HPC may help to solve economic problems by parallelizing methods of simulation, data analytics, machine learning, artificial intelligence, and optimization. How do you see the potential of HPC for these areas?

Answer: HPC has potential in all of these areas. Data analytics and machine learning methods are often iterative, with phases of obviously (data-)parallel operations interleaved with steps of global communication, such as broadcasts or global reductions. Such algorithms profit from the advanced communication infrastructure available in HPC systems, which cluster or cloud systems are lacking. In particular, the completion time of the training phase of deep learning networks can be directly scaled down by using a large HPC system with GPUs, but only if the gradient exchange is performed efficiently using an HPC interconnect. At the same time, the amount of training data that can be consumed scales up with the size of the system due to the high performance and high capacity global file system directly accessible from all compute nodes. Optimization problems arise on their own in specific Operations Research applications, but are also core components of the data analytic, machine learning, and AI toolchain. However, there is a lack of well-scaling optimization software, in particular discrete optimization and integer programming tools. This is an area where further research and development, ideally driven by business computing cases, is needed. Last but not least, there is the classic field of econometrics, where HPC systems enable the use of both higher-resolution models (e.g., for the economic sectors, or their interaction behavior, or with more scenarios/longer horizons) and the study of larger and more complicated interconnected systems.

Question: HPC in business computing seems to be used rarely, in contrast to several other disciplines. What are the reasons, and what are success factors for overcoming limited deployment?

Answer: Business computing is traditionally focusing more on enterprise software stack integration and using a well-defined but limited amount of hardware. It is often not considered an option to invest in dedicated hardware for specific tasks, and the software ecosystem for business tasks is often not able to exploit HPC resources. This contrasts with R&D or engineering usage, where getting the best dedicated hardware and software combination for a particular task is considered essential, even if it incurs complications in the integration, and even if these environments change in comparatively short cycles. This indicates that programmability and availability of standard software are key to better adoption of HPC in business computing, and I believe the advent of data-driven models which requires cross-cutting analytics on the (currently often siloed) data, a core asset in a digital economy, will be a significant driver to drive such interfacing efforts.

Question: Since a few years we observe a substantially slowing growth of computing performance (of single CPU cores), resulting in the phenomenon that sequential programs will stop running faster on successive generations of hardware. This technological development has been considered economically disruptive for the IT economic cycle, constituting that "software and hardware advances fed each other". Is "thinking parallel" (in terms of problem-solving approaches, algorithmic design, and program execution in HPC environments) a promising, or even the only promising way out of this dilemma?

Answer: Until around 2006 Dennard's scaling law allowed increasing the transistor density and operating frequency with each chip technology generation at constant energy consumption by lowering voltage. The breakdown of this scaling law resulted in an end of ever increasing single-core performance. [https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/] The manufacturer's reaction of increasing the core count – both on CPUs and accerators, such as GPUs – has already made parallelization a necessity. Some "embarrassingly parallel" algorithms and usage scenarios have hidden this situation from users: client-server REST architectures where servers have improved performance over traditional mainframe services, and environments like the Java Virtual Machine. But such opaque parallelization gains are limited in scope, and we truly need explicitly parallel approaches throughout the software stack. Programming languages, and more generally programming environments, are the most critical components that influence adoption here: A successful programming language needs to both force the programmer to consider concurrent behavior of programs and also to support the user at doing so. Functional languages are on the rise (again), as are domain-specific languages, and classic low-level envrionments, such as MPI, are keeping up to abstract the implementations from the rapidly changing underlying architectures.

A new challenge, however, is data movement cost, which traditionally has been considered secondary to computing cost: even across a single chip with non-uniform memory access speeds, but more so in a distributed HPC system, is data movement the limiting factor. Communication-avoiding algorithms and system middlewares that are data-centric, i.e., take data movement cost in the system memory hierarchy into account, will be central to performance scaling.

Question: Which type(s) of HPC architecture are the most promising for solving problems in business administration and economics? What are the most important skills for unleashing the power of HPC in companies?

Answer: Current and upcoming HPC architectures are shaped by the will to build exascale systems, where one has to remember that this metric is a classic FLOP rating, which does not represent the typical business administration or economic modeling compute mix. They will consist of a much more diverse set of hardware, from CPUs and accelerators, to diverse memories, including network-attached, non-volatile and native object storage. The convergence of classic HPC usage and data analytics, including various kinds of machine learning, however, will help making these resources available for business computing needs.

With regard to required skills, we need to better train users to think parallel, and move from an imperative do-this-after-that paradigm to data-driven designs: Performing operations on certain data sets makes processes naturally data-parallel and thus exhibits concurrency potential without starting from complicated thread-of-work locking designs.

Question: What do you think about the promises of quantum computing to establish an alternative line of HPC systems?

Answer: Quantum computing will not directly – or in the near future – provide an alternative HPC architecture. However, we may see "quantum accelerators" entering the market, much like GPUs did a decade ago. Given that heterogeneous HPC system architectures are already a reality, this should not be surprising or considered difficult to deal with. However, programmability of quantum computing devices is still in its infancy, and it will likely take a significant amount of time to reach a level – or abstraction – that makes it appealing to this audience.

Question: In the European Union, the public-private-partnership EuroHPC has recently started as "Europe's journey to exascale HPC". Do you see any benefits from this development for the deployment of HPC in business computing?

Answer: EuroHPC is an important step to institutionalize the transeuropean network of HPC infrastructure. It offers European research and development infrastructure and will be the critical component in realizing a European exascale computing in an internationally competitive timeframe, which is a tremendous chance for the businesses operating in the common market. At the same time one has to realize that its founding document [council regulation (EU) 2018/1488, http://data.europa.eu/eli/reg/2018/1488/oj] explicitly limits commercial use to 20% of the total access time per system (and at market price) [idem, Art. 14]. Given the expected commercial interest from classic commercial HPC users, such as mechanical engineering, chemical and pharmaceutical industry and industrial R&D in general, it seems challenging to envision business computing to obtain a significant share of these resources. But it is a challenge that can be answered by showing the benefits of highly scalable business computing tasks in a research project context while exascale systems are still a flagship technology, so that they are readily recognized among the standard workload when exascale resources are considered mainstay. For this purpose, EuroHPC does in fact form an excellent basis.

Question: Which kinds of business models are viable for bringing HPC to industries? Will (large) companies invest in their own dedicated HPC systems and/or will HPC mainly be used via cloud computing providers?

Answer: We are already seeing the convergence of on-site HPC infrastructure, private cloud – on-premise and off-premise – and public cloud for many industry users. HPC systems that are ready for containerized application deployment – without sacrificing HPC specific hardware advances, in particular the low-latency and high-throughput interconnect – make it easy to move workloads from experimental state (even a laptop) to make full use of the scaling capabilities of the high density systems. On the other hand, cloud providers realize that offering HPC access inside their data centers will help customers that already hold a lot of data in their cloud storage to perform compute-intensive tasks that require tens of thousands of compute nodes.

Some industrial customers – maybe even more than currently expected by the market – will avoid moving their compute and data storage to actual cloud destinations if they can use on-premise resources in a cloud-like manner: scaling resources as needed, redistributing load across different business unit's hardware assets, with service and security guarantees, and a 'pay as you go' model. With data being the crucial asset of a company, its physical storage location and accessibility guarantees are key; the fact that it simplifies compliance and protection is an added benefit.

Bio of the interviewee: Utz-Uwe Haus is a Senior Research Engineer at CRAY. He studied Mathematics and Computer Science at TU Berlin. After obtaining a Doctorate in Mathematics at University of Magdeburg he worked on nonstandard applications of Mathematical Optimization in Chemical Engineering, Material Science and Systems Biology. After 5 years as Senior Researcher at the Department of Mathematics at ETH Zürich he is now leading the Cray European Research Lab in Basel, developing the Mathematical Optimization and Operations Research group, working on data-dependency driven workflow optimization on future HPC architectures.