DeIC HPC TekRef group report on the

Supercomputing 2018 conference

in Dallas, Texas, USA, in November 2018.

Participants:

- Erik B. Madsen, SDU, erikm@sdu.dk
- Jannick Visling, SDU, janvi@sdu.dk
- Kurt Gammelgaard Nielsen, SDU, kgn@sdu.dk
- Rune Kildetoft, KU, kildetoft@nbi.dk
- Ian Heide Godtliebsen, AU, ian@chem.au.dk
- Ole Holm Nielsen, DTU, ole.h.nielsen@fysik.dtu.dk
- René Jacobsen, DeIC, rene.jacobsen@deic.dk
- Lottie Greenwood, ESS, Lottie.Greenwood@esss.se
- Pietro Antonio Bortolozzo, DTU, pbor@dtu.dk

Preface

The DeIC **HPC TekRef** group, represented by a delegation of *High Performance Computing* (HPC) system administrators from Danish universities (KU, AU, SDU, and DTU), DeIC, and the *European Spallation Source* (ESS), participated in the *SC18 International Conference for High Performance Computing, Networking, Storage, and Analysis*[1] annual conference in November 2018 in Dallas, Texas, USA. The delegation also participated in the satellite conferences *High Performance Consortium for Advanced Scientific and Technical* (HP-CAST) held by HPE, as well as *Intel HPC Forum*. The satellite conference *Dell EMC HPC Community* could not be attended due to overlap with the other conferences.

This report summarizes the fact-finding investigations carried out by the delegation during the conferences for the purpose of obtaining and accumulating knowledge about the latest HPC systems, technology, and software for use by DeIC as well as the HPC community in Denmark.

The delegation attended a large number of conference sessions, and in addition held a number of prearranged one-to-one meetings with key technology vendors under *Non-Disclosure Agreements* (NDA). The vendor list was Intel, AMD, Nvidia, Mellanox, DellEMC, HPE, and Lenovo. The information obtained under NDA cannot be disclosed in the present public report, but any publicly available information is reported.

The topics in the following sections represent the most important trends in HPC, with a particular emphasis on HPC in Denmark.

1 SC18: https://sc18.supercomputing.org/

CPUs for HPC

Intel

Intel celebrates the 20th anniversary of the Intel *Xeon* processor line in 2018. At the *Intel HPC Forum* event, Intel announced a new processor named *Cascade Lake-AP (Advanced Performance).* However, it seems as a quick fix for the AMD EPYC challenge.
The *Cascade Lake-AP* is a *Multi-Chip Package* design, which is two chips mounted in one processor package. It will contain 48 cores and 12 memory channels along with cache improvements. It is still on 14nm technology, so it is large and uses a whopping 350 Watts!
The end of life for this product is set to around only one year from launch, when Intel introduces their next generation processor. The new *Cascade Lake-AP* is set to be released Q1/Q2-2019.

The *Cascade Lake family* is already available for early order and will support 2, 4 and 8 sockets with up to 28 cores per socket, memory speed at 2933 MHz, and *Apache Pass* memory modules. New *VNNI* instructions for use with AI/ML are included. The *Cascade Lake family* still uses PCIe-Gen3, so AMD will have an advantage with their newest lineup *Rome* at least for some time.

The delegation also received NDA information about the future Intel processor lines code-named *Cooper Lake* and *Ice Lake*.

AMD

During 2018, AMD has delivered a series of AMD *EPYC* processors code-named *Naples*, which is now competing in the HPC arena, but so far with no entries on the current TOP500 list.

However, in November 2018 AMD announced[2] the next-generation 7nm *EPYC* server processor code-named *Rome*, which uses the *Zen 2* micro-architecture with multiple chiplets per package. The *Rome* processor is socket-compatible with the current *Naples* and will have up to 64 CPU cores with 256-bit floating-point width and other improvements[3].

Importantly, *Rome* will have PCIe 4.0 interfaces supporting 200 Gbit/s network fabrics and faster GPU communication. Several supercomputer sites are planning to install *Rome* based systems in 2019.

The delegation received NDA information about *Rome* and the future *Milan* and *Genoa* processor lines.

ARM

Sandia National Lab has installed[4] an ARM processor based supercomputer, which is number 204 on the current TOP500 list. The French CEA has announced plans to install an ARM-based supercomputer in 2019.

The delegation received NDA information from vendors planning to launch ARM based products.

---

2 AMD Rome: http://ir.amd.com/index.php/news-releases/news-release-details/amd-takes-high-performance-datacenter-computing-next-horizon
3 AMD Rome: https://www.top500.org/news/amd-takes-aim-at-performance-leadership-with-next-generation-epyc-processor/
4 ARM: https://www.top500.org/news/sandia-to-install-first-petascale-supercomputer-powered-by-arm-processors/

Accelerators

NVIDIA

In September 2018 NVIDIA announced the **Tesla T4**, a 70W low profile PCI-e GPGPU card focused on machine learning (ML). It uses NVIDIA's *Turing* architecture including GDDR6 memory, instead of the HBM2 used on the Tesla V100. It has the following specifications:

CUDA cores	3840
Tensor cores	320
Memory	16GB GDDR6
FP32 Peak	8.1 TFLOPs
FP16 Peak	65 TFLOPs
INT8 Peak	130 TOPs

NVIDIA also showcased their **HGX-2**, a reference design for a server with 16 Tesla V100 GPUs, up from 8 GPUs in the previous models and doubling the theoretical max performance. The communication channel is changed from a hybrid cube mesh to using an *NVSwitch*, which should improve bandwidth for inter-GPU communication.

AMD

In the days up to the conference, AMD announced two new GPGPUs, **Radeon Instinct** MI50 and MI60.[5] They are both based on their new 7nm Vega GPU. The specifications are as follows:

	Radeon Instinct MI50	Radeon Instinct MI60
Stream processors	3840	4096
Memory	16GB HBM2	32GB HBM2
FP64 Peak	6.7 TFLOPs	7.4 TFLOPs
FP32 Peak	13.4 TFLOPs	14.7 TFLOPs
FP16 Peak	26.8 TFLOPs	29.5 TFLOPs
INT8 Peak	53.6 TFLOPs	589 TFLOPs
TDP	300W	300W

Intel

In July 2018, Intel announced[6] its plans to discontinue their entire **Xeon Phi** *Knights Landing* line of many-core processors. No announcement of new Xeon Phi products were made.

No announcements were made in the area of **FPGAs**, but the delegation held an NDA meeting where the future of their FPGA lineup was discussed.

Intel did not make any announcements regarding their **Nervana neural chips**, but they did so earlier this year when they announced[7] that they plan to bring a product to market in 2019.

---

[5] Radeon Instinct: https://www.anandtech.com/show/13562/amd-announces-radeon-instinct-mi60-mi50-accelerators-powered-by-7nm-vega
6 Xeon Phi: https://qdms.intel.com/dm/i.aspx/9C54A9A7-BF37-4496-B268-BD2746EA54D3/PCN116378-00.pdf
7 Nervana: https://www.hpcwire.com/2018/05/24/intel-pledges-first-commercial-nervana-product-spring-crest-in-2019/

Field Programmable Gate Array (FPGA)

Even though *Field Programmable Gate Arrays* (**FPGAs**) have been available for many years, they have recently come into the spotlight for HPC purposes. In essence, a FPGA is a (re-)programmable unit to which computation is off-loaded from the CPU. FPGAs can be used instead of *Application Specific Integrated Circuits* (**ASICs**) if one needs more flexibility than with chips soldered directly onto the system board.

FPGAs are used mainly in the field of *High Frequency Trading* (**HFT**) or for data streaming in fast **AI** inference, but in principle could be applied with advantage in many domains due to their programmability. For example, one could off-load computational "hot loops" to an FPGA, in which case the FPGA can be thought of as a form of general-purpose accelerator just like a GPU.

Intel became a player on the FPGA market with its 2015 purchase of the FPGA producer **Altera**, continuing production and development of the **Stratix** and **Arria** lines[8] of FPGA's. Intel and Altera are, however, not the only producers of FPGAs, and they receive strong competition from long-time manufacturer **Xilinx[9].**

Illustration 1: Virtex FPGA systems from Xilinx

System on a Chip (**SoC**) cards providing FPGAs are probably of most interest to the HPC community for use as general-purpose accelerators. As an example we mention Intel's **Arria** 10 GX FPGA, which is a half-length PCI card with a PCIe 3.0 x8 interface, boasting 8GB of DDR4 memory and 128MB of flash for storage, operating at 60 watts meaning that they are well below the needs of a power hungry GPU.

We also note that Intel has plans to produce a CPU-FPGA hybrid chip[10]. Chips like these will likely not replace the conventional CPU, but will certainly be highly applicable for specific domain problems.

8 Intel FPGAs: https://www.intel.com/content/www/us/en/fpga/devices.html
9 Xilinx: https://www.xilinx.com/products/silicon-devices/fpga.html
10 CPU-FPGA: https://www.nextplatform.com/2018/05/24/a-peek-inside-that-intel-xeon-fpga-hybrid-chip/

Many of the big server manufacturers either do or have plans to support Intel FPGAs as general-purpose accelerators in some of their products. This includes the HPE Proliant DL360 and DL380[11] as well as the DellEMC PowerEdge R840 and R940xa[12], just to name a few.

It seems that FPGAs, at least in the near future, will be part of the HPC landscape, as Intel and the big vendors are currently cooperating in producing servers with FPGA capabilities. Due to the programmability, the application range of FPGAs is in principle infinite, but this feature also makes it harder to use for non-specialists and more research and development is needed for them to become more generally applicable to HPC.

Quantum computing

Quantum computing for HPC is still far off into the future, and the talk on *A Quantum Future of Computation*[13] outlined how quantum computers may someday be integrated with normal HPC systems as special purpose accelerators. Take home messages were:

- Quantum computers are getting closer to reality, but will not replace classical computers.
- They will be extremely powerful accelerators for specific **big compute/small data** problems.
- "Dequantizing" quantum algorithms gives new "quantum inspired" classical algorithms.

Illustration 2: IBM Q quantum computer

IBM showcased the IBM Q[14] publicly available quantum computer system operating at a temperature of 0.015 Kelvin and featuring 5 to 14 qubits, as well as a simulator environment.

11 HPE FPGA: https://community.hpe.com/t5/Servers-The-Right-Compute/Why-Field-Programmable-Gate-Arrays-FPGAs-are-the-versatile/ba-p/7016281#.XA48ARAnbCI
12 Dell FPGA: https://www.top500.org/news/dell-emc-offers-fpga-and-gpu-options-on-new-servers/
13 Quantum: https://sc18.supercomputing.org/presentation/?id=inv102&sess=sess229
14 IBM Q: https://www.research.ibm.com/ibm-q/

Server products

The most popular HPC server unit seems to be the compact *4 servers in 2U* ("*½U*") form factor, which is available from most server vendors. Liquid cooling is also possible for such compact systems. The number of DIMM memory modules is limited to 8 per processor with the current ½U designs, but future processors are going to require more than 8 memory channels for high-performance systems.

An important new development is that most of the main server vendors are planning to deliver products based upon the upcoming AMD *Rome* processor in 2019, thereby giving Intel-based products a strong competition.

The delegation received NDA information about future server products from vendors HPE, Lenovo, and Dell.

Interconnects for HPC

Mellanox Infiniband

Last year Mellanox announced 200 Gbit/s (200G) **HDR** fabrics, and the past year has seen adoption in supercomputers (e.g., at TACC's "Frontera"[15],[16]). It is worth noting that since 200G requires the system to have a PCIe Gen4 interface this will give AMD a temporary advantage with the *Rome* processor that includes PCIe Gen4.

Mellanox currently offers products which split the 4 communication lanes in HDR into cable pairs providing **HDR100** 100G links to two nodes, but between the switches the speed is 200G HDR. With this technique a 40-port 200G HDR switch will match an 80-port 100G EDR switch. It is important to note that HDR100 is **2 lanes at 50G**, whereas EDR is **4 lanes at 25G**!

Mellanox supports 200G for both Infiniband and Ethernet with their *ConnectX-6* adapter cards, and with different interconnect cables one can support both technologies on the same adapter. For 200G speed, copper cables can be up to 2m long and optical cables up to 100m long.

Intel Omni-Path

Intel is repeating the message "Omni-Path has 48 ports per 1U and this reduces fabric costs", but until the next 200 Gbit/s version of Omni-Path comes to market (expected in 2019), Mellanox will have an advantage on network capacity per U.

A Mellanox strength versus Omni-Path is that Omni-Path relies on the CPU for network processing, whereas Mellanox Infiniband offloads the network processing to adapters and switches.

HPE Composable Fabric (Plexxi)

HPE recently acquired the company *Plexxi*[17] that aims to create a new *Intelligent Network Fabric*. This will be an HPE-only technology, where the idea is to have a dynamic network using software and very few cables, making it easy to scale and easy to connect using breakout cables.

---

15 Mellanox: http://www.mellanox.com/page/press_release_item?id=2094

16 HPCwire: https://www.hpcwire.com/2018/08/29/taccs-frontera-supercomputer-expands-horizon-for-extreme-scale-science/

17 Plexxi: https://www.plexxi.com/

Data centers and liquid cooling

At the HP-CAST conference, **HPE** described their current liquid-cooled SGI 8600 system, as well as the company's path towards future 100% liquid-cooled systems.

At an NDA meeting with **Lenovo** it was emphasized that future power consumption of 300-400W per processor or GPU will almost certainly require liquid cooling technology. Current HPC systems from Lenovo can use heat pipes to cool processors, memory and PCIe adapters (see photo).

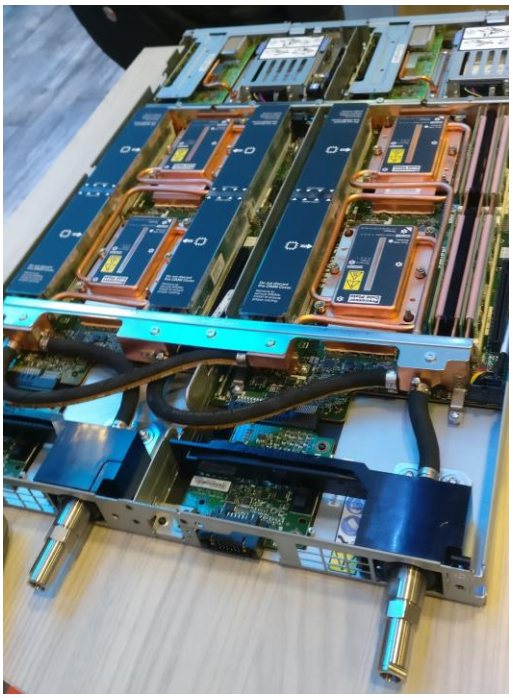

Illustration 3: Lenovo tray with two ½U servers containing all-copper cooling pipes.

Illustration 4: CoolIT rack Cooling Distribution Unit (CDU)

Liquid cooling vendors **CoolIT** and **Asetek** (a Danish company) showcased their liquid cooling technology on the show floor. CoolIT has changed the water tubes inside the servers to a new technology that should better prevent water leakage compared to previous types of tubes.

Cooling Distribution Units (*CDUs*, see photo) are crucial infrastructure components in liquid cooled data centers, as are water quality control and monitoring. A number of HPC centers are already using water-cooling technology.

Middleware and software for HPC

Slurm

The *Slurm*[18] open source resource manager software is adopted increasingly by HPC centers worldwide, including a few known sites in China. On the TOP500 list, *Slurm* is used on 6 of the top 10, and is used on about 40% of the entire TOP500 list.

A *Slurm* BOF session described recent and future feature developments, and a lively discussion with 200+ *Slurm* users showed how widely the product is being used.

Within the BOF session, there was also a presentation on what features could be coming to *Slurm* in the future. Highlights included adding GPU submission options for next year's release of *Slurm* v19.05, and talk of developing a restful API to allow for both job submission via web forms, and providing better integration for emerging tools such as **Jupyter** notebooks.

Software modules

A BOF session *Getting Scientific Software Installed* gave an overview of HPC software tools such as *EasyBuild*, *Lmod*, *Spack*, *Singularity*, *Docker*, etc.

Containers

Software containers in HPC represent challenges due to multi-user security and scalability for parallel applications. Many HPC sites provide container tools such as *Singularity*, Shifter, *uDocker*, and *Charliecloud*.

Whilst there was discussion on how containers could be useful in a HPC context, the scheduling systems typically used for containers in other types of IT services, such as *Kubernetes* or *Docker Swarm*, do not seem to be considered beneficial yet. This is because those schedulers tend to handle cases where resources are available on demand, whereas HPC usually has the demand outstrip the resources available.

Cloud services for HPC

A number of vendors were advertising cloud services available for HPC at SC18. There seemed to be especially focus on '*cloud bursting*', the practice of using the public cloud only during spikes for computing capacity, and otherwise focusing on the private cloud. However, there still did not seem to be a strong solution for how to move large amounts of data between the public and private cloud.

NASA had made a study in order to evaluate the cost-effectiveness of running HPC workloads in the cloud[19], and they found that when the jobs rely on HPC-level interconnects, then cloud computing cannot match the performance (and price) for at tailored HPC system. However, for jobs that do not depend on the high speed interconnects, HPC can benefit from cloud computing[20], so in the future NASA will offer this as a service to their users.

18 Slurm: https://www.schedmd.com/
19 NASA: https://www.nas.nasa.gov/SC18/demos/demo5.html
20 NASA: https://www.hec.nasa.gov/news/features/2018/cloud_computing_services.html

Red Hat, the provider of open source *Red Hat Enterprise Linux* and *CentOS Linux* as well as enterprise *Kubernetes* and hybrid cloud solutions, announced on October 28 that **IBM** is going to buy the company.

Artificial Intelligence (AI) and Machine Learning (ML)

Recently *Artificial Intelligence* (**AI**) has emerged as a new technology in the HPC and *High Performance Data Analytics* (**HPDA**) scene. The definition of AI is somewhat diffuse, but can be loosely interpreted as intelligence or human-like behavior exhibited by machines. *Machine Learning* (**ML**), i.e., machines learning from data, is a broad collective term for models, algorithms, and methods using sample data to make predictions for still unknown data points.

Even though AI and ML describe slightly different, arguably overlapping, concepts, they are currently used interchangeably in many cases to describe the same idea, namely using (mostly) problem-agnostics algorithms to make predictions or decisions through the use of sample data. Usually when applying the terms AI and ML in a broad forum, what is most often meant is the application of **Deep Neural Network** (DNNs) which is also known as **Deep Learning**. DNNs are thus still the predominant method applied to learning problems, but we note that recently other ML approaches have also started to gain momentum for solving problems in many different domains.

The dominance of DNNs has also become a driver for the hardware business, as every major manufacturer is currently providing hardware specifically designed for speeding up the optimization and usage of DNNs. Currently purpose-built machines utilizing the power of multiple GPUs or other accelerator cards can be purchased from most of the big vendors, and attending NDAs with Intel, HPE, Dell EMC, and Lenovo, the delegation has learned that more is to come in the future. The development towards AI centric hardware is already seen today by processors introducing new instructions into the *Instruction Set Architectures* (ISAs), as well as server manufacturers producing servers with massive numbers of accelerator cards.

In Intel's processor portfolio, we see the impact of AI by the introduction of the *Vector Neural Network Instructions* (VNNI) in the new Cascade Lake[21] and the *bfloat16* instruction added to the future Cooper Lake[22].

Accelerators are widely used in AI and there is a rapid development in products tailored towards ML. As an example, the announced NVIDIA *Tesla T4* utilizes *Turing Tensor Cores* providing up to 32 times more throughput than the prior NVIDIA *Pascal* GPUs for AI workloads. Intel has also joined the accelerated AI market and has embraced a so-called holistic approach with the introduction of the *Nervana*[23] product lines code-named *Lake Crest* and the coming *Spring Crest*. Also, FPGAs have recently been introduced as a fast way of streaming and processing real-time data, mainly used for fast inference on image data which is key, for example, when designing self-driving cars where any latency can mean life or death.

Many software frameworks for working with ML exist, easing the practical task of implementing ML models, e.g., **TensorFlow** and **Caffe2** to name a few. These frameworks are still widely developed and highly optimized for many different hardware architectures, making them the go-to choice when implementing ML algorithms efficiently.

---

21 Cascade Lake: https://en.wikichip.org/wiki/intel/microarchitectures/cascade_lake
22 Cooper Lake: https://en.wikichip.org/wiki/intel/microarchitectures/cooper_lake
23 Nervana: https://www.hpcwire.com/2018/05/24/intel-pledges-first-commercial-nervana-product-spring-crest-in-2019/

In conclusion, AI and ML are still major driving factors in business, and thereby in HPC and HPDA, pushing the development of both hardware and software. Development is being performed at all levels and in the future we will continue to see even more hardware designed specifically for fast optimization of DNNs on massive data sets, but we will also start to see the emergence of purpose-built machines designed, for example, for fast inference. Many different paths are currently being explored such as acceleration using GPUs, CPUs utilizing specialized instructions, or integrated approaches like the Intel *Nervana*. Which strategy will prevail only the future will tell. It is certain that AI and ML will heavily influence the HPC landscape in the years to come.

TOP500 supercomputers

In the latest 52nd bi-annual **TOP500** supercomputer list[24] of the world's highest performance HPC systems, the USA now holds positions 1 and 2, whereas long-time leader China now holds positions 3 and 4. Switzerland is the top European country with the number 5 position. Remarkably, China has 45% of the TOP500 systems, while the USA is down to 22%. However, USA leads with 38% of the aggregate system performance versus China's 31%. Intel leads in terms of processor technology with 95% of the TOP500 systems.

Exascale computing

The struggle towards Exascale[25] computing continues, and it is crystal clear that this cannot be achieved using the hardware designs currently available due to the power required per FLOPS. Even though we are unlikely to host a machine in the Exascale regime in Denmark for many years to come, the "Quest for Exascale" will still make a huge impact on the landscape of Danish HPC, as the improved data center designs and hardware advancements gained through this endeavor will benefit facilities that are more modest as well.

What has become apparent is that the capabilities and power consumption of standard x86 CPUs is insufficient for going to Exascale, and we need to save precious CPU cycles by doing off-CPU computing, for example, by offloading computations from CPUs to ASICs and FPGAs. We are already familiar with accelerators such as GPUs, but other types of accelerators are now starting to emerge such as FPGAs or new designs of so-called vector engines[26]. Off-CPU computing is exemplified in HPC by accelerators, but could also in the future materialize as, for example, interconnect PCI boards doing fast on-board compression providing higher fabric throughput, or other specialized ASICs.

Conclusion

In these years, we are seeing a true "Cambrian explosion" of HPC hardware, where the trend is that each unique task needs specialized hardware for maximum performance. This observation underlines the importance of tracking current development in order to keep our HPC centers up to date.

Thus attending the Super Computing (SC) conference series is as important as ever! This is where the Danish HPC community may gather the knowledge and tools to elevate Danish HPC to new heights!

---

24 TOP500 list: https://www.top500.org/lists/2018/11/
25 Exascale: https://en.wikipedia.org/wiki/Exascale_computing
26 Vector engine: https://www.nec.com/en/global/solutions/hpc/sx/vector_engine.html