DeIC HPC TekRef group report on the

Supercomputing 2022 conference

Dallas, Texas, USA, November 2022.

Participants:

- Rune Kildetoft, KU, kildetoft@science.ku.dk
- Martin Rehr, KU, rehr@science.ku.dk
- Lars Melwyn, DTU, melwyn@dtu.dk
- Per Aa. Ankerstjerne, DTU, peraaa@dtu.dk
- Ole Holm Nielsen, DTU, ole.h.nielsen@fysik.dtu.dk
- Lottie Greenwood, ITU, logr@itu.dk
- Pietro Bortolozzo, DTU, pbor@dtu.dk
- Rainer Bohm, AAU, rb@its.aau.dk
- Rasmus Jensen, AAU, rj@its.aau.dk

Preface

The DeiC **HPC TekRef** group, represented by a delegation of *High Performance Computing* (HPC) system administrators from several Danish universities (KU, AAU, ITU, and DTU), DeiC participated in the *SC22 International Conference for High Performance Computing, Networking, Storage, and Analysis*[1] annual conference in November 2022 in Dallas, Texas, USA. The delegation also participated in the satellite conference *HPE ForCAST* Fall 2022.

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 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, Lenovo, and Cornelis Networks. The information obtained under NDA cannot be disclosed in the present public report, whereas any publicly available information is reported.

The topics in the following sections represent the delegation's view of the most important trends in HPC, with a particular emphasis on HPC in Denmark. The sections are divided into a number of separate technology topics.

[1] SC22 https://sc22.supercomputing.org/

Processors for HPC

High-performance processors (CPUs) are crucial for most types of HPC applications. The delegation learned about current products as well as roadmaps for future products. Meetings were held with the following vendors.

Intel

Intel announced just before SC22 the *Intel® Xeon® CPU Max Series* (code-named *Sapphire Rapids HBM*) as shown in this announcement: https://www.intel.com/content/www/us/en/products/docs/processors/max-series/overview.html

A *Xeon Max* processor socket may contain up to 56 CPU cores, optionally with up to 64 GB of closely coupled *High-Bandwidth Memory* (HBM2e), and up to 16 DDR5 4800 MHz memory DIMMs per socket (up to 4 TB using 128 GB DIMMs in a 2-socket server with 2 DIMMs per channel). The TDP power may be up to 350 W per processor.
The PCIe Gen5 bus offers twice the data transfer rate of PCIe Gen4, and will enable the use of 200 Gbit/s network fabrics, and faster GPU data transfer.
New CPU instructions include the *Advanced Matrix Extensions* (AMX) tiled matrix multiplication accelerator.

The delegation from DeiC was informed about the roadmap for future CPU and GPU products under NDA.

AMD

AMD has announced the 4th Generation AMD EPYC™ Processors, Code-Named “Genoa” including up to 96 “Zen 4” microarchitecture-based cores, built on 5 nm process technology: https://www.amd.com/en/processors/epyc-9004-series
The memory bandwidth and capacity has been expanded to 12 DDR5 channels (up to 6 TB using 128 GB DIMMs in a 2-socket server with 2 DIMMs per channel), as well as next-gen I/O with PCIe® 5.0 and memory expansion with CXL™. The TDP power may be up to 360 W per processor.

The delegation from DeiC was informed about the roadmap for future CPU products under NDA.

ARM

ARM processors are already everywhere in our phones, tablets, televisions, single board PCs and a wealth of other devices. In fact also in high-performance systems like in recent Apple products, but just not particularly visible in the HPC ecosystem as a source of primary compute power. That is, however, going to change for a number of reasons:

- At first the ARM product is a license and not a physical chip, and the licensees are allowed to customize the design to include low level integration with proprietary functional modules (read accelerators) on the chip.

- Secondly, the speed, memory bandwidth and general performance of high-end ARM CPUs are approaching HPC levels.
- Thirdly, the ARM processor is well known for being very power efficient which will be of increasing importance for future hyperscalers and HPC installations.
- Fourthly, as intense computations move to the GPU, the main CPU will become a “co-processor” to the GPU in integrated superchip or XPU designs, which could provide opportunities for less costly ARM solutions.

For those wanting a headstart on doing serious computing on ARM systems, several cloud computing companies recently announced and already deliver compute cloud instances based on high-end ARM systems, *e.g.*

- Ampere Altra
- AWS Graviton 3
- Azure
- Google cloud
- Oracle cloud

Standard server ARM systems are already available from reputable vendors such as HPE and Gigabyte. But more seriously, we are starting to see high-end NVIDIA GPUs combined with ARM systems, like the more HPC/AI oriented G242-P32 from Gigabyte. The idea of running high-end multi GPU workloads off power efficient ARM based systems is also on several roadmaps.

The main news from ARM is discussed in the NVIDIA section below.

Accelerators

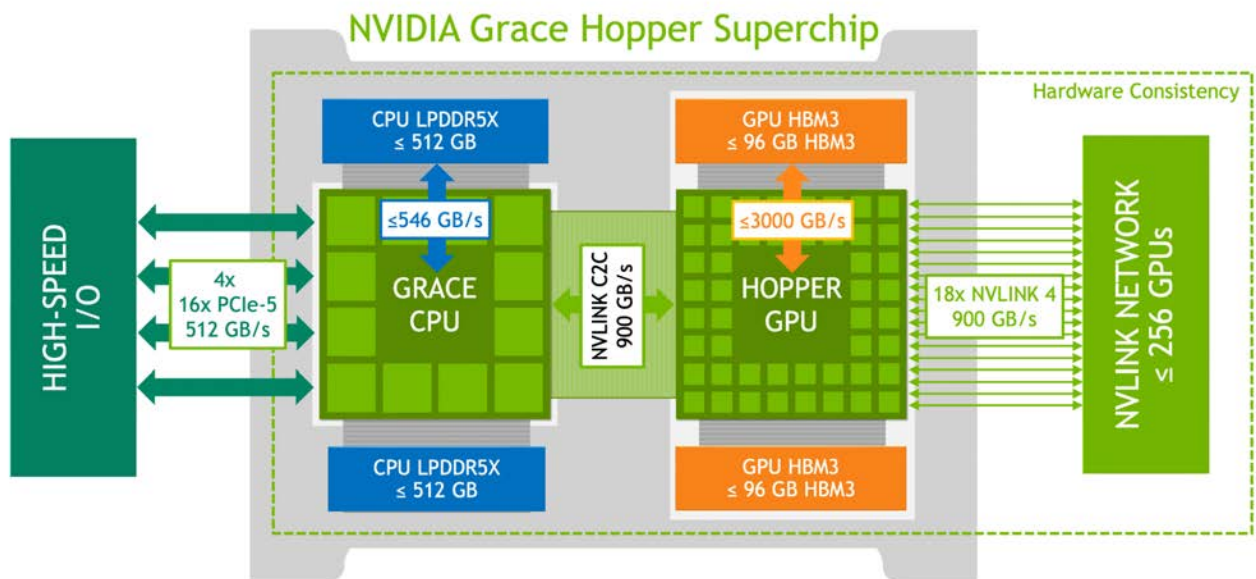
A variety of different computing accelerators are available for consideration, the utility being entirely dependent on the application software being used.

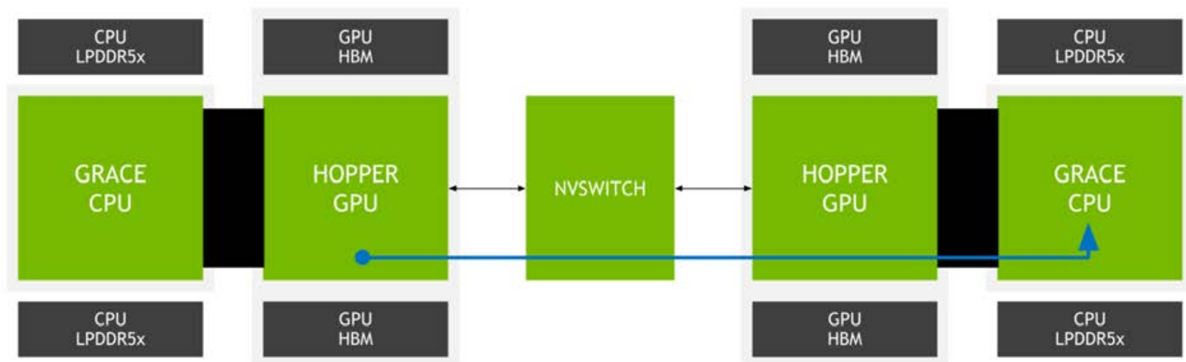
NVIDIA

The *Grace Hopper Superchip*[2] public information was described in a meeting with NVIDIA:
https://www.nvidia.com/en-us/data-center/grace-hopper-superchip/
The NVIDIA *Grace Hopper Superchip* combines the Grace ARM CPU as well as Hopper GPU architectures using NVIDIA® NVLink®-C2C[3] to deliver a CPU+GPU **coherent memory model** for accelerated AI and HPC applications.

---

[2] https://www.nvidia.com/en-us/data-center/grace-hopper-superchip/
[3] https://www.nvidia.com/en-us/data-center/nvlink-c2c/





AMD

AMD is working on adding GPU cores to their chips - as it was noted from the group: “so one vendor is adding CPUs to a GPU, whereas the other is adding GPUs to a CPU - interesting”. Either way, we are going to get heterogeneous chips with the gain (and challenges!) in programming.

The DeiC delegation received NDA information about the AMD *Instinct* accelerators roadmap.

Intel

Intel announced just before SC22 the *Intel® Data Center GPU Max Series* (code-named *Ponte Vecchio*): https://www.intel.com/content/www/us/en/products/docs/processors/max-series/overview.html

The *Data Center GPU Max* is Intel's new GPU series with up to 128 Xe HPC Cores Compute Units which will compete with GPUs from NVIDIA and AMD. The TDP power may be 300-600 W per GPU, so direct liquid cooling needs to be considered.

An accelerator overview

Tom's Hardware has a comparison of Intel, AMD, and NVIDIA GPUs/accelerators in https://www.tomshardware.com/news/intel-fires-up-xeon-max-cpus-gpus-to-rival-amd-nvidia

Cerebras

Hardware overview

Cerebras[4] is taking AI Computing and Computer Simulation to the next level.
The number of cores on a single Cerebras chip equals the core count of
123 nVIDIA A100 boards.

The Cerebras system has memory equivalent to a **1000 nVIDIA A100** (40GB) boards and the system-on-a-chip-design delivers a theoretical max bandwidth 12733 times the nVIDIA A100 board and a fabric bandwidth that is 45833 higher.

A complete Cerebras system[5] consumes 23 kW of power where the chip itself consumes 20 kW. This equals the power consumption of 80 nVIDIA A100 40GB boards.
With a core count of x123, a theoretical max bandwidth of x12733 and a fabric bandwidth of x43833 the Cerebras system has a huge potential to provide more energy efficient AI solutions compared to the traditional GPU-based AI systems that are used today.

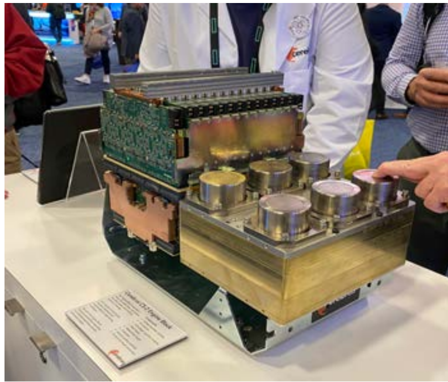
[4] https://www.cerebras.net
[5] https://f.hubspotusercontent30.net/hubfs/8968533/CS-2%20Data%20Sheet.pdf

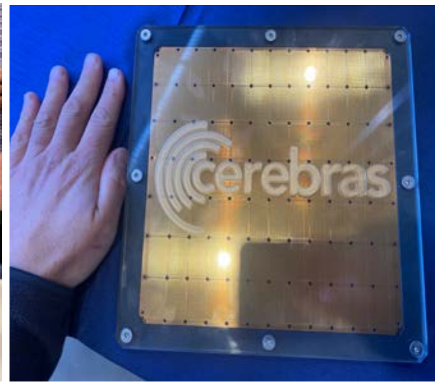
Maximum Power Requirement
23 kW

System IO
12x 100 Gb Ethernet

Cooling
Air- or water-cooled

Dimensions
15 Rack Units (26.25")





Multiple Cerebras nodes are interconnected into a cluster setup where 12x100 Gb ethernet connects deliver an aggregated interconnect bandwidth of 1200 Gb/s. Each node is cased in a standard 15u rack.

Installation

Leibniz Supercomputing Centre[6] (Lrz) of the Bavarian Academy of Science and Humanities announced on May 22, 2022 that they are installing a Cerebras system[7].

Software overview

The greatest challenge with new hardware is always porting the existing software stack. According to Cerebras their ML Software integrates with the popular machine learning

---

[6] https://www.cerebras.net/press-release/leibniz-supercomputing-centre-accelerates-ai-innovation-in-bavaria-with-next-generation-ai-system-from-cerebras-systems-and-hewlett-packard-enterprise/
[7] https://www.theregister.com/2022/05/25/hpe_cerebras_lrz/

frameworks TensorFlow and PyTorch which should ensure that researchers can effortlessly bring their models to the CS-2 system[8].

Server products

The delegation met with several of the major server vendors to learn about the latest products as well as roadmaps for future products.

HPE

Following the acquisition of CRAY, HPE is now rebranding their systems to include the CRAY name - e.g. the Apollo servers will be known as *HPE CRAY XD2000*. Upcoming and future server products from HPE were described under NDA.

Dell

On the 14th of November Dell Technologies announced updates to the *PowerEdge* series servers and expanded the XE line of specialty servers with new air-cooled and dense liquid-cooled systems, containing the latest NVIDIA high-number multi-GPU servers dedicated to intense HPC and AI workloads.

The updated list of Dell server offerings:

- Dell PowerEdge
    - 7625 Two-socket, 2U server, HPC oriented with GPUs
    - 7615 One-socket, 2U server,
    - 6625 Two-socket, 1U server, HPC oriented, high-density
    - 6615 One-socket, 1U server
- Dell PowerEdge XE
    - XE9680 6U, two-socket, 4th Gen Intel Xeon, 8 x NVIDIA H100 or A100 GPU, PCIe Gen5
    - XE9640 2U, two-socket, 4th Gen Intel Xeon, 4 x Intel GPU Max, direct liquid cooling, PCIe Gen5
    - XE8640 4U, two-socket, 4th Gen Intel Xeon, 4 x NVIDIA H100, PCIe 5
    - XE8545 4U, two-socket, 3rd Gen AMD Milan, 4 x NVIDIA A100 GPU, PCIe Gen4

Future server products based upon Intel *Sapphire Rapids* as well as AMD *Genoa* were presented under NDA. See the Intel and AMD sections in this document.

Lenovo

[8] https://8968533.fs1.hubspotusercontent-na1.net/hubfs/8968533/Whitepapers/Cerebras-Whitepaper-ProgrammingAtScale.pdf

Future server products based upon Intel *Sapphire Rapids* as well as AMD *Genoa* were presented under NDA. See the Intel and AMD sections in this document. Existing liquid cooling solutions from Lenovo were presented.

Interconnects for HPC

The delegation met with a number of network interconnect vendors of interest for parallelization of HPC codes.

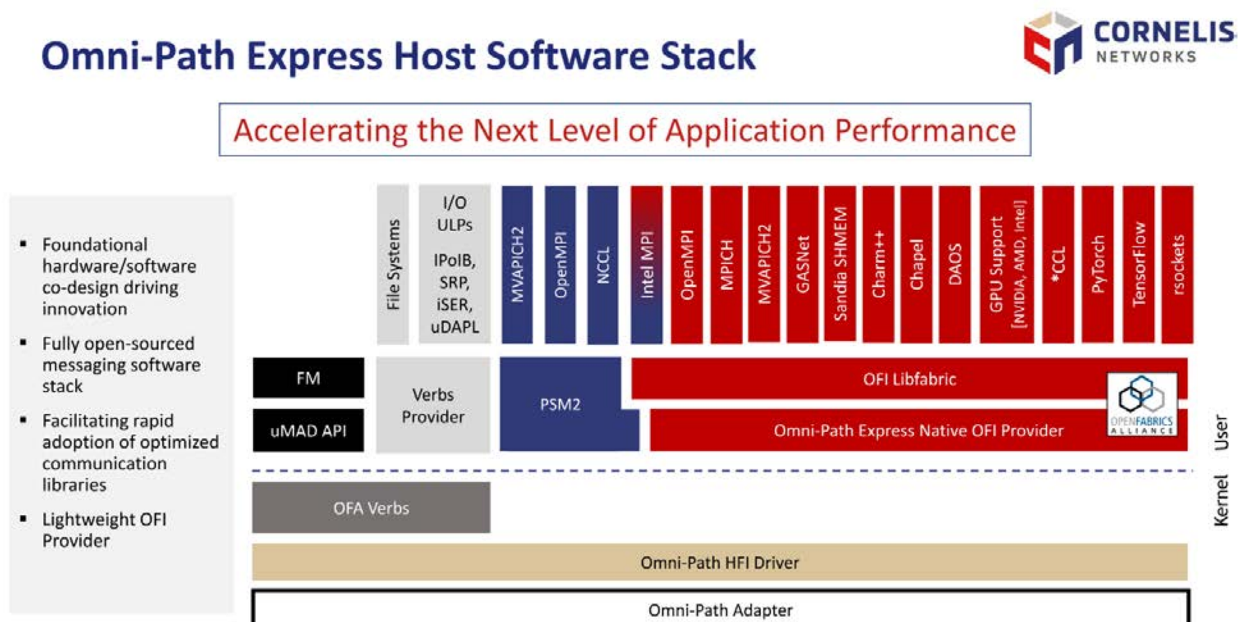
Mellanox Infiniband

The delegation from DeiC was informed about the roadmap for future Infiniband networking products from NVIDIA/Mellanox under NDA.

Omni-Path

The Omni-Path network fabric products were spun out from Intel in 2020 to a new company *Cornelis Networks*.

On Nov. 11, 2022 *Cornelis Networks* introduced a new *Omni-Path Express software stack* (OPX) based upon the *OpenFabrics Interfaces* (OFI) industry standard. The OPX software is Open Source and will be included in major Linux distributions starting with the libfabric v1.16 package. The portfolio of Omni-Path products can be seen on the page https://www.cornelisnetworks.com/products/ and the OPX software stack is illustrated in this figure:



OPX offers significantly improved performance compared to the previous PSM2 software, and fully supports all popular MPIs, PGAS & AI frameworks, Object Storage file systems, and all popular GPUs.

NVIDIA GPU communication will be supported within *libfabric* in Q2 2023.

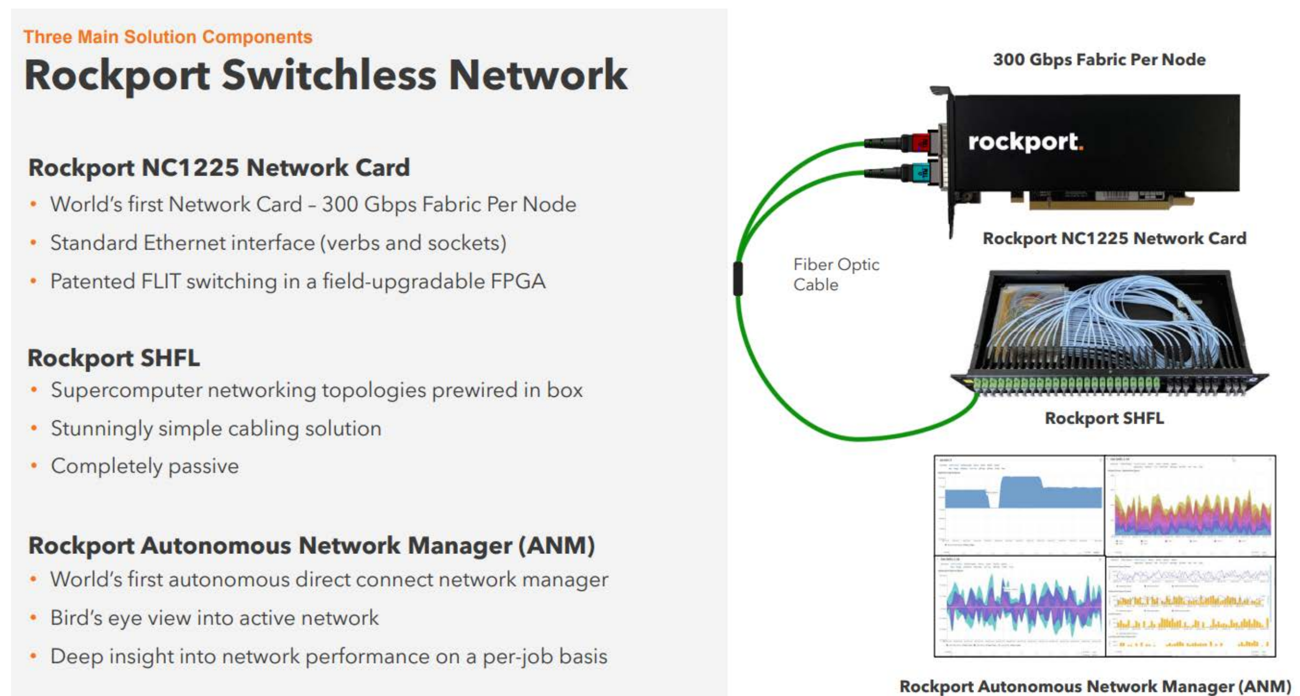
The delegation from DeiC was informed about the roadmap for future Omni-Path products under NDA.

HPE/Cray Slingshot

The HPE/Cray *Slingshot* HPC network fabric was presented by HPE, see https://www.hpe.com/us/en/compute/hpc/slingshot-interconnect.html
Special switches and adapters are offered, but standard Ethernet servers are also interoperable with *Slingshot*. The currently offered speeds are 100 and 200 Gbit/s.

Rockport Networks

The *Rockport* switchless network solution (https://rockportnetworks.com/) is a new take on switchless (but with passive optical Rockport SHFL patch box) interconnects based on custom-designed network cards and cables of fiber bundles used to make point-to-point connections between nodes. Depending on the combined price of the network cards, fiber cables, and corresponding optimizing control software, the Rockport solution may cut the cost of the cluster fabric down to a 1/2 (compared to *Infiniband*) at the expense of specific topology choices and limitation on parallel high-performance low-latency connections. The Rockport N1225 network card boasts 300 Gpbs bandwidth distributed on 12 fiberoptic connections each at 25 Gbps.



Compute Express Link

In the HPC TekRef report from supercomputing ‘16[9] and ‘17[10], the group wrote about Gen-Z - it has since been merged into *Compute Express Link* (CXL) along with e.g. OpenCAPI. Further information is in the CXL homepage at https://www.computeexpresslink.org/

With the release of Zen 4 CPUs from AMD, both Intel and AMD now have CPUs with support for CXL - which was demonstrated on the show floor.

---

[9] https://www.deic.dk/sites/default/files/documents/PDF/SC2016_Rapport.pdf
[10] https://www.deic.dk/sites/default/files/documents/PDF/SC2017_Rapport.pdf

Data centers and liquid cooling

As the CPU and GPU *Thermal Design Power* (TDP) keeps increasing, vendors either have to use ever bigger heatsinks for air cooling, or look for other ways to remove the heat from the servers. This has been a recurring topic in the DeiC HPC TekRef reports, and planners need to consider cooling solutions when designing new server rooms.

Next-generation CPUs and GPUs, and the forthcoming HPC APUs and XPUs combining CPUs with GPUs, will result in compute nodes with formidable computational performance, but also with outrageous power requirements. The scale may be up to 4 kW per node, leading to correspondingly extreme cooling requirements.

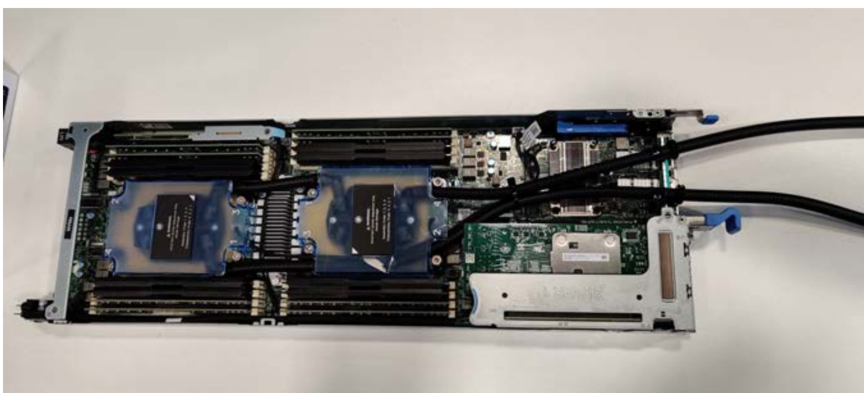
HPC computing may be nearing the end of the line of Moore's law and also of the air-cooled HPC center. Increases in computational performance will scale with power consumption, and thereby also the cooling requirements.

The next-generation dense compute racks presented to the delegation may have power requirements of up to 100 kW, and the following generations on the order of 200+ kW per rack. Liquid cooling in the server room is logically inevitable. At this year's SC22 conference, all server vendors presented a suite of liquid cooling solutions spanning in complexity from retrofitting existing solutions all the way to liquid immersion of equipment.

The presented liquid cooling solutions were of varying degrees of propriety and complexity. The solutions reviewed by the delegation included a) rear door heat exchanger, b) in-chassis liquid cooling, c) combining air + liquid cooling, and d) direct liquid cooling solutions with in-rack, in-node piping and external *Cooling Distribution Units* (CDUs), e) the presence of even more data center cooling infrastructure, and f) liquid immersion cooling. In our opinion, liquid immersion cooling may come to play a role in cost-efficient high-density HPC in the longer term.

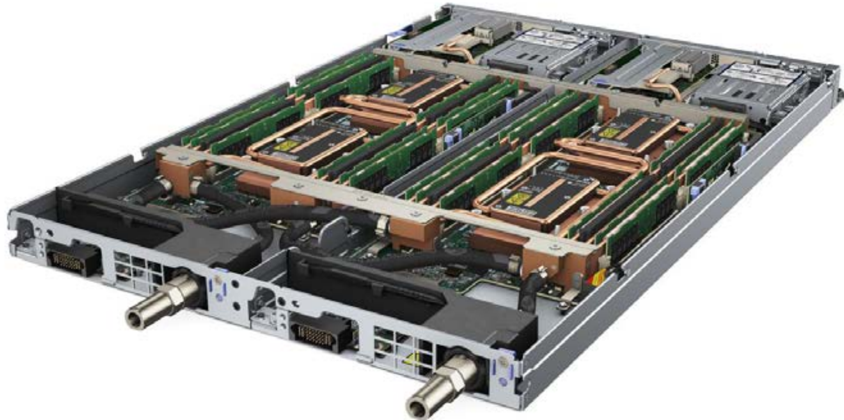
Direct Liquid Cooling of servers

Many, if not all, server vendors now offer **direct liquid cooling** solutions of server CPUs using liquid-cooled cold plates with an external liquid distribution manifold and central *Cooling Distribution Units* (CDU) for connection to facility cooling systems. An example is the server from Dell in this photo:



Notably, some vendors offer cooling solutions where not only the CPUs are cooled, but also hot components such as DIMM memory modules, network adapters, and power supplies.

For example, the Lenovo vendor uses copper-based liquid piping inside the server as in this image:



Liquid cooling solutions can remove from about 60% and up to 90% of all generated heat from a server, making the best solutions almost heat-neutral to the server room.

An important advantage of cooling CPUs by liquid is that the processor’s *Case Temperature* can be kept rather low compared to using air cooling with heatsinks. The latest as well as upcoming generations of CPUs have more restrictive temperature upper limits than previous generations, which causes performance degradations if insufficient cooling is provided. A liquid cooled system should therefore perform better than an otherwise identical air cooled system.

Cooling Distribution Units

Cooling Distributions Units (CDU) as well as *heat exchangers* are key components for a data center installation using liquid cooling: direct liquid cooling, rear-door heat exchangers, or in-row cooling units. External facility cold water is used to provide cooling to IT equipment which usually has higher water quality requirements.

The *American Society of Heating, Refrigerating and Air-Conditioning Engineers* (ASHRAE) recommendations for water-cooled servers is described in a very detailed White Paper available at
https://www.ashrae.org/File%20Library/Technical%20Resources/Bookstore/WhitePaper_TC099-WaterCooledServers.pdf

A typical small CDU unit is shown in this photo:

A SC22 BOF Session on “*Liquid Cooling Adoption: Roadblocks and Key Learnings*” included detailed discussions of water quality issues, cooling pipe materials and quality requirements, and the carbon footprint savings that can be obtained. Several large HPC sites emphasized potential problems with biofilm growth in cooling water loops.

Liquid immersion cooling

Rising energy prices and a focus on the environment and minimizing costs have forced the industry to think alternatively, opening up the market for several new exciting technologies. Since 2015, the "*Immersion Cooling*" concept has been successfully spread in the USA, and the concept is now well underway in Europe.

Immersion cooling is typically used in facilities where traditional air cooling is insufficient to keep electronic components within safe operating temperatures. This includes high-performance computing facilities, such as data centers and supercomputers where the hardware components generate a large amount of heat and also require efficient cooling. Immersion cooling is generally used when high cooling efficiency, compact design, and reliability are essential.

Immersion cooling is a method of cooling electronic equipment where the hardware components are submerged in a dielectric fluid, such as mineral oil or a specialized liquid cooling solution. This type of cooling is highly efficient because the fluid can absorb heat directly from the hardware components, allowing them to operate at higher temperatures as well as power levels without overheating. Compared to air cooling, you will gain performance of approx. —20% according to a Data Centre which has installed immersion cooling.

Comparing immersion cooling to water cooling, the answer is more diffuse because hardware suppliers have difficulty comparing their solutions. Both types of cooling have advantages and disadvantages, but the main reason for choosing water cooling is pricing, hardware warranty and support issues.

Several technologies are used in immersion cooling systems, including the following:

- Dielectric fluids: The main component of an immersion cooling system is the dielectric fluid, which does not conduct electricity. This is important because it allows the liquid to be in direct contact with the electronic components without causing damage. Dielectric fluids are typically chosen for their ability to absorb heat, their low toxicity, and environmental impact.

- Cooling systems: To remove the heat absorbed by the dielectric fluid, immersion cooling systems typically use a closed-loop cooling system. This may include pumps, radiators, and fans to circulate the liquid and transfer the heat to the ambient air or other cooling mechanisms.

- Enclosures: To prevent the fluid from spilling or evaporating, immersion cooling systems often use pens or tanks to contain the liquid and the electronic components. These enclosures are typically made of materials resistant to corrosion and compatible with the dielectric fluid.

- Monitoring and control: To ensure the operation of the immersion cooling system properly, it is essential to monitor and control the temperature and flow of the fluid, as well as the overall health of the electronic components.

Overall, immersion cooling systems combine several technologies to provide efficient and reliable cooling for electronic equipment. There are several benefits to using immersion cooling, including the following:

- High cooling efficiency: Immersion cooling is highly efficient because the dielectric fluid can directly absorb heat from the electronic components. This is in contrast to traditional air cooling which relies on convection and conduction to remove heat, and is therefore less efficient at high power densities.

- Compact design: Because the electronic components are submerged in the dielectric fluid, immersion cooling systems can be more compact than air cooling systems, which require space for airflow and the cooling fans. This makes immersion cooling well-suited for applications where space is limited, such as in high-performance computing systems.

- Reliability and durability: Immersion cooling systems are generally more reliable and durable than air cooling systems because the fluid is less susceptible to contamination than cooling air, and the components are protected from dust and other debris. This can help extend the life of the electronic equipment and reduce the need for maintenance.

- Reduced noise: Immersion cooling systems are generally quieter than air cooling systems because the fluid can absorb and dissipate any sound generated by the electronic components. This can be beneficial in facilities where noise is a concern, such as in offices or residential settings.

While immersion cooling offers many benefits, there are also some downsides to consider. These include the following:

- Cost: Immersion cooling systems can be more expensive than traditional air cooling systems because they require specialized fluids, enclosures, and cooling

systems. This may make immersion cooling less cost-effective for some applications such as in small or low-power systems.

- Compatibility: Not all electronic components are compatible with immersion cooling because the fluid can damage or corrode certain materials. For example, certain plastics and metals may be incompatible with the dielectric fluid, and the fluid itself can damage some electronic components.

- Safety: Immersion cooling systems must be carefully designed and operated to avoid electrical hazards and other safety, health and toxicity issues. For example, the fluid must be adequately contained to prevent spills or leaks, and the system must be designed to avoid the build-up of static electricity.

- Maintenance: Immersion cooling systems require regular maintenance to ensure that the fluid is clean and free of contaminants and that the components operate within safe temperatures. This can be time-consuming and may require specialized equipment and expertise.

- Hardware warranty and support: The question surrounding hardware warranty when used with immersion cooling is very uncertain, and hardware suppliers avoid giving an answer. The same challenges are present in support agreements.

The cooling of high-end compute nodes by full immersion into engineered non-conducting fluids is provenly not only the most energy and cost efficient way of cooling dense GPU nodes, it is also going to be the default commodity way of liquid cooling the coming generations of high core count power-intense standard servers.

Standardized immersion cooling will allow for cooling and compute technologies to continue to evolve along independent paths and will foster proper market competition to prevent costly vendor lock-in through custom vendor specific cooling components. Intel is spearheading the commoditization of immersion cooling through industrial partnerships and open reference designs in the context of the open compute project. By now all leading server vendors appear to support (at some level) immersion cooling of standard servers by removal of heatsinks and fans.

The liquid immersion cooling solutions are either single phase with fluid + CDU + secondary heat exchanger or two-phase (boilers) with a condenser + secondary heat exchanger. where the latter solution is the most energy efficient.

Several immersion cooling companies and types of immersion cooling solutions were presented at SC22:

- Single phase, in-rack
    - Submer, e.g. BSC
    - GRC, e.g. TACC
- single phase, node
    - www.iceotope.com
- Two phase
    - Gigabyte/LiquidStack

The take-away for immersion cooling is that in order to avoid a sequence of costly data center HVAC/CRAC renovations and vendor lock-in, schemes that will fix a percentage of a node budget to non-reusable 5 year life-time in-node plumbing, and at the same time to

harvest the benefits of market innovation and competition in addition to cost and energy efficiencies, the time to transition the data center to immersion cooling is now. For greenfield server room deployment total cost considerations are even more relevant.

Additional benefits of immersion cooling:

- extended component life-time due to stable thermal environment
- space efficient, simple server room
- fewer in-node components, fans, and heat sinks

The drawbacks of both direct-liquid-to-chip and immersion cooling are:

- oil blooms and messy environment due to maintenance and leaks
- potentially hazardous materials or environments
- longer node maintenance cycle
- no industry data center standard
- no commodity component ecosystem

For all the reasons above Intel recently launched an immersion cooling research initiative and released an open IP OCP reference design:

- Intel Lab
- Whitepaper

in an attempt to guide industry towards immersion cooling standards.

Quantum computing

The field of *Quantum Computing* (QC) may promise to disrupt the world of high-performance computing and supercomputing centers in the coming years. QC presents a radically different paradigm to computing and promises a polynomial to exponential speed-up for specific algorithms compared to current technologies. QC is not next-generation but next-level computing.

Yet, on the other hand, it will not replace current classical HPC centers or supercomputers, but rather augment those with dedicated and possibly cloud-connected quantum processing units (QPU accelerators). HPC and supercomputing centers will play a pivotal role not only concerning the education in and the adoption (integration) of QC, but also in the development and validation of algorithms by using QC emulators and being a part of hybrid algorithms mixing classical and quantum computing.

In essence, quantum processors run specific quantum algorithms as quantum circuits consisting of sequences of quantum logic gates applied to registers of qubits to solve particular problems intractable on even the largest current supercomputers.

Today's two most important aspects of quantum computing are finding reproducible physical platforms of optimal qubits and quantum gates, and developing and implementing high-speed-up algorithms. But with the recent advent of numerous commercially available quantum processors as either in-house or API-driven cloud-based cryolabs, QC is becoming of increasingly practical relevance for HPC centers.

At SC22 the leading quantum processor vendors presented recent, or announced coming, quantum processor generations with updated properties such as qubit numbers, circuit depths, coherence times, gate fidelities, and so on. QPUs and qubit platforms at SC22 included the following:

- Amazon Bracket, multi vendor
- D-Wave, annealing
- Google, Sycamore (54 qubits), superconducting transmons
- IBMQ, Eagle (127 qubit), soon Osprey (433 qubits), transmons
- Intel, Tangle Lake (49 qubit), silicon spin qubits
- IonQ, Aria (21 qubits), trapped ions
- IQM, ? (20 qubits), superconducting transmons, soon unimons
- Microsoft, (?), topological Majorana qubits
- Rigetti, Aspen-M-2 (80 qubits), soon Aspen-M-3 (80 qubits), superconducting transmons

A computer science perspective on current research trends in QC was offered at the one-day 3rd International workshop on QC software held in conjunction with the SC22 technical program and also the concluding SC22 panel discussion.

A sobering remark on the potential role of Quantum Computing was made in Dongarra's Keynote Talk (see below):[11]

> "Today, we have machines that are built on manycore plus GPUs. I would think that in the future, we would see that expand, [and] have other accelerators added to that collection. So think about adding an accelerator that does something specific for AI. Or think about adding an accelerator which does something like neuromorphic computing. We can add accelerators to the collection to help in solving our problems. **Maybe quantum would be another accelerator – I don't see quantum being its own compute**," said Dongarra.

Quantum Computing Links

- QC Introduction
  - Quantum Inspire
- QC related fields of research
  - Quantum algorithms
  - Quantum annealing
  - Quantum computing
  - Quantum cryptography
  - Quantum information
  - Quantum networks
  - Quantum simulator
- Qubit platforms
  - Superconducting transmon
  - Trapped Ion

[11] https://www.hpcwire.com/2022/11/16/jack-dongarra-a-not-so-simple-matter-of-software/

Storage

HPC obviously has to deal with large to huge amounts of data. Therefore storage systems are integral parts of any HPC data center.

WEKA

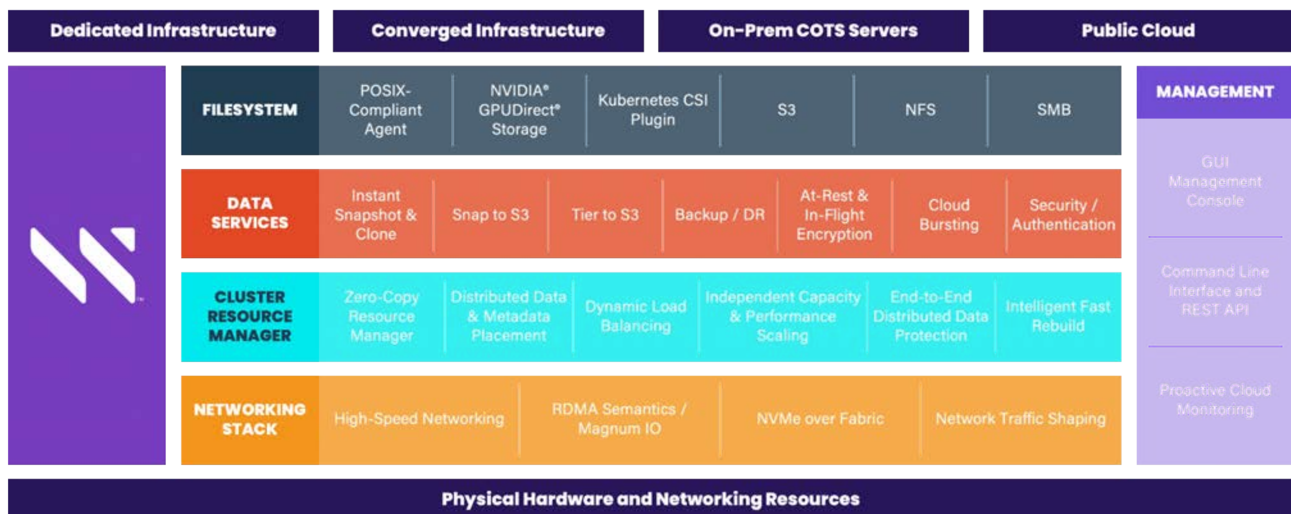
The delegation had an NDA presentation by storage vendor WEKA. WEKA[12] was founded in 2013 and produces software defined storage – a fully distributed parallel file system.

The minimum requirement to get started is 6 servers with 1 NVMe drive in each server. WEKA is meant to be a Tier 1 storage that will support the following protocols for access to data: Native NVIDIA GPUDirect Storage, POSIX, NFS, SMB, and S3.

Tier 2 storage will be an S3 Storage of your choice and it can be on-premise or Cloud.

Licensing is based on TB/PB per year with one price for the NVMe storage tier, and a second and much lower price for tier 2 storage: About one 10th of the Tier 1 license price.

The architecture is shown in this figure:



It seems that WEKA will be a player that we will see more in the future. One may hope that WEKA will make an agreement with a hardware vendor for the purpose of making an integrated all-in-one solution in the future.

[12] https://www.weka.io/how-it-works/

Middleware and software for HPC

Dongarra keynote talk

Jack Dongarra, Distinguished Professor, University of Tennessee gave the **conference opening keynote talk**[13] entitled "*A Not So Simple Matter of Software*".

This was the **ACM A.M. Turing Award Lecture** (also called the "*Nobel Prize of Computing*") given for pioneering contributions to numerical algorithms and libraries that enabled high performance computational software to keep pace with exponential hardware improvements for over four decades.

Dongarra described in the talk his work since the 1970ies on *Open Source* high-performance software libraries such as *Linpack* and *ScaLAPACK*. He also designed the *Message Passing Interface* (**MPI**) for parallel computing. Since 1993 Dongarra and his co-workers have released the bi-annual **TOP500** supercomputer lists (see also below).

Slurm

The *Slurm* resource manager (batch job queue system, see https://slurm.schedmd.com) was represented at the *SchedMD* company's show floor booth. This year there was no *Slurm* BOF session.

A number of presentations[14] was made by *SchedMD* on new features in the current release 22.05 as well as upcoming features in release 23.02. Dynamic addition and removal of nodes, GPU sharding, and Cgroup v2 are the most noticeable new features in *Slurm* 22.05.

OpenHPC

The OpenHPC project (https://openhpc.community/) provides packages and recipes for the deployment of a fully featured HPC cluster. An OpenHPC BoF was held during SC22, where they discussed plans for the project including the decision of moving to Rocky8. In the discussion they highlighted that they plan to release one last patch for the CentOS7 branch in Q1 2023.

There are two branches of the OpenHPC currently, version 1.3.x supporting CentOS7 with either xCAT or Warewulf for deployment/management of the cluster, and version 2.7 including support for Rocky8 with only Warewulf (https://warewulf.org/) for cluster management.

---

[13] https://sc22.supercomputing.org/2022/06/14/acm-a-m-turing-award-lecture-to-be-presented-at-sc22/
View the video of the lecture at https://www.youtube.com/watch?v=cSO0Tc2w5Dg and a transcription at https://www.hpcwire.com/2022/11/16/jack-dongarra-a-not-so-simple-matter-of-software/

[14] *Slurm* roadmap slides are available in https://slurm.schedmd.com/SC22/Roadmap.pdf

During the BoF they explained that this is due to xCAT support being dropped by Lenovo, as well as a lack of maintainers for the deployment system. It is expected that unless someone volunteers to maintain the xCAT deployment recipes, going forward OpenHPC will only provide documentation for installation and management of Warewulf.

They also encouraged more users to join the OpenHPC technical steering committee, so they can gather experiences from a wider range of cluster scales.

Their slides on community engagement shows the number of monthly visitors to the OpenHPC website has more than doubled since last year to over 100 000, which looks positive for the future growth of the OpenHPC community.

TOP500 supercomputers

During SC22 the new edition of the TOP500 list (https://www.top500.org/) was released.

Compared to the previous list from June 2022, the top end of the list remains the same, with Frontier, Supercomputer Fugaku, and LUMI occupying the top 3 spots. The Leonardo system at EuroHPC/CINECA, entering at No. 4, is the only new machine at the top of the list.

The LUMI supercomputer has retained its 3rd position on the TOP500 thanks to a major upgrade that kept it competitive. DeiC visited the LUMI exhibition booth, which is described in the LUMI news at https://www.lumi-supercomputer.eu/a-successful-sc22-week-for-lumi/, and made a photo of the LUMI booth presentation:

In the TOP500 list 75.8% of the systems are using Intel processors, vs 20.2% of systems using AMD. However, AMD systems have 48.7% of the overall performance.

China holds 32.4% of the top machines, in comparison to the United States's 25.4% - this is not a drastic change from June, where China hosted 34.6% of the top machines, and the US 25.6%. The USA has 43.6% of the aggregate system performance on the TOP500, whereas China has 10.6%. When considering continents, Europe has both 26.2% of the systems and aggregate system performance on the Top500.

A minimum of 1.73 PetaFlops is required to have a place in the TOP500, in comparison to June's 1.65 PetaFlops.

There has been some movement at the top of the Green500 list (https://en.wikipedia.org/wiki/Green500) - the Henri system at the Flatiron Institute in the US, whilst being ranked at 405 on the TOP500, is ranked top here. Frontier TDS, which last time topped the list, has dropped by one place. The LUMI System, which previously placed third, has fallen to No. 7. Also of note is the Leonardo system entering the Green500 list at No. 14.

Conclusion

New high-performance CPUs as well as GPUs continue to make an impact on the HPC market. The latest generation chips may comprise of the order of 50 to 100 billion (10^9) transistors, and even though chip features continue to shrink according to Moore's Law, the heat dissipation continues to grow and setting limits for computer performance.

Cooling the new, hot processors has become a crucial challenge for system designers and data center architects. Air cooled systems continue to work reasonably well in larger 1U or 2U chassis form factors with ample air flow, but dense dual-socket dual-servers in 1U have to resort to *Direct Liquid Cooling* as soon as the processor power exceeds 200-250 W.

HPC servers with a power consumption of 1-2 kW and even more are just around the corner, and liquid cooling needs to be considered for all HPC data centers. Emerging technologies such as *Liquid Immersion Cooling* are beginning to appear but are not mainstream at this time.

The TOP500 supercomputer list tracks deployment of the largest HPC computers in the world. While the No. 1 supercomputer performance continues the Moore's Law exponential growth, the No. 500 performance also grows, albeit with a decreasing growth rate during the last 10 years. The entry point to TOP500 currently is at 1.7 PetaFLOPS.

The race towards the Exa-scale HPC supercomputers, which has been ongoing for a number of years, was won by the *Frontier* supercomputer at the *DOE/SC/Oak Ridge National Laboratory* in the USA. At 1.1 ExaFLOPS *Frontier* is currently the only Exa-scale system, but more are expected to arrive in the next couple of years. The LUMI supercomputer maintains its No. 3 TOP500 position with 309 PetaFLOPS.