DeiC HPC TekRef group Report on the

Supercomputing 2023 conference

Denver, Colorado, USA, November 2023.

Table of Contents

The DeiC HPC TekRef delegation SC2023

- Erik B. Madsen, SDU, erikm@sdu.dk
- Pietro Bortolozzo, DTU, pbor@dtu.dk
- Per Aa. Ankerstjerne, DTU, peraaa@dtu.dk
- Rainer Bohm, AAU, rb@its.aau.dk
- Rasmus Jensen, AAU, rj@its.aau.dk
- Nina Simone Marstrand Sander Aagaard, AAU, nsmsa@its.aau.dk
- Lottie Greenwood, ITU, logr@itu.dk
- Jonas Bardino, KU, bardino@science.ku.dk
- Rasmus Munk, KU, rasmus.munk@nbi.ku.dk
- Rune Kildetoft, KU, kildetoft@nbi.dk

Preface

In November 2023, a delegation of High-Performance Computing (HPC) System Administrators and collaborators representing various Danish universities participated in the Supercomputing Conference 2023 (SC 23) held in Denver, Colorado. The primary objective of their attendance was to acquire comprehensive updates on prevailing trends and prospective advancements within the field of High-Performance Computing. Facilitating this trip, the Danish e-infrastructure Consortium (DeiC) provided essential support for the delegation, thereby enabling the extraction of pivotal insights subsequently encapsulated in this report.

In adherence to customary practice, the group mentioned above engaged in a Non-Disclosure Agreement (NDA) and secured meetings with eminent entities in the chip, network, and server industries. However, owing to the confidential nature inherent in these interactions, the scope of information that can be divulged is constrained. Subsequent sections of this report will endeavor to articulate discernible trends and noteworthy developments from the conference within the bounds of permissible disclosure.

Piggybacking on the hyperscalers?

At this time, the needs of the traditional HPC Community are outpaced by those of the hyperscalers - they need the latest CPU/GPU or accelerator model in their data center *yesterday*. The next generation of GPUs and, more importantly, the APUs and CPU/GPU integrated accelerators oriented towards AI and HPC worlds are expected to grow significantly in power consumption and, to some extent, in sheer physical size. The chiplet design of integrating CPUs with accelerator-oriented units will, in particular, affect the physical layout of the Printed Circuit Boards (PCB) in the future. This change in design is due to several factors, for one, the reorientation toward memory-bound applications such as training and testing deep neural networks; secondly, the need to reduce the latency penalty that such workloads would otherwise encounter in a classical CPU/GPU architecture. This latency is partly due to the physical distance from the CPU to PCIe-attached units such as memory and GPUs. Therefore, to accelerate AI workloads for tasks such as ML and training and inferring on neural network models, the chip vendors are increasingly stacking classical general-purpose CPUs with APU-like units with large caches and HBM3 memory on a single die to, in part, reduce the classic latency penalty for compute and thereby accelerate workloads. Examples include the Instinct MI300X AI accelerator from AMD and Intel's Gaudi 2 processor, both on display on the show floor. This increase in physical chip and chiplet size is expected to challenge the traditional PCB layout inside each rack unit. Therefore, the traditional 19-inch rack might not be suitable for future compute installations that plan to utilize the ML/AI-oriented accelerators that the Hyperscalers are leaning heavily into. Because of this, the traditional HPC could benefit from being on the lookout for new form factors, like the Open Rack v3[1], where the width of the rack is 21 inches. This would allow more space for the coming generations of accelerated CPUs, APUs, and GPUs, thereby allowing more computing power per rack unit. However, there is currently no accepted standard for the future regarding rack size and layout, which is also caused by the expected massive increase in power consumption; the MI300X, for one, is listed with a Thermal Design Power (TDP) of 750W at peak[2]. This means that traditional air cooling is no longer a viable option for high-end AI-oriented rack installation, which will be expanded upon in the Liquid Cooling section.

This additional complication means that many vendors are redesigning their racks to accommodate both the increased and increasing die size and the subsequent liquid cooling equipment that also has to fit inside the already cramped space of a rack. In the early stage of this transition, each vendor is expected to pursue their designs, making interoperability between vendor racks a challenge for HPC customers like universities that often have a mix of different vendors to achieve the lowest price in tenders. The hope is that a common standard will emerge in the future, but at this time, it needs to be clarified what that standard will be; in the meantime, looking toward what Hyperscalers have done and are doing can be a path forward.

The world is AI.

At Supercomputing 2023, unsurprisingly, AI was a major presence. This was clear across every vendor booth and talk that we attended. The priority for the vendors was to exhibit how they performed/benchmarked regarding AI workloads.
This was not unexpected since the AI field has been omnipresent over the last few years. However, with 2023 capping, it seemed to dominate everything.
A shift, if any, was the transition from highlighting training performance to inference performance on the hardware front.
This is likely to respond to the fact that training the emerging large models currently requires a hefty investment, which seems beyond the largest hyperscalers.
Because of this, training any of the big LLMs seems like an unviable option for most. Instead, most other commercial-scale players will rely on adjusting the hyperscaler-provided models with less training and optimizing the continuous inference performance when serving the models.

---

[1] https://www.opencompute.org/documents/google-implementation-orv3-spec-1-pdf
[2] https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html

In addition, the current limitation on high-end training hardware such as the various DGX systems, H100, or A100 GPUs from Nvidia makes inference performance optimization more attractive for most in the short term. Instead, hardware like the Intel Gaudi 2 seemed widely available if there was an urge to conduct price-competitive inference.
However, vendors still sent optimistic messages to institutions such as universities. They committed to making high-end hardware like the mentioned systems available for research purposes, a general priority for the vendors.

In terms of future platforms, as mentioned in the Piggybacking section, the future is likely to be a mix of CPU and APU or APU-like integrated CPU/GPU accelerators, with the latter being the primary. Here, the tight integration and increased die size only seem to continue to increase the throughput partially via reduced latency and increased bandwidth between memory, cache, and compute units.
Current training and inference perform adequately with lower precision regarding data type size. Intel’s Gaudi 2, for example, seems to abandon FP64 for their most recent release and instead focus more on optimizing FP16 and FP8 performance for AI workloads[3] AMDs Instinct retains FP64, in particular for HPC-oriented applications and FP32 in general, but also emphasizes FP16 and FP8 performance in AI-related workloads[4]. This makes sense in maximizing throughput and increasing efficiency if enough information can be retained with lower precisions[5]. Whether this is also enough for the future multi-modal models that are expected to be the next step in LLM development will be interesting to follow. Nevertheless, AI developments, including LLMs and beyond, define the future of HPC accelerator-oriented hardware.

AI is already significantly impacting our world and is expected to become even more influential. It can bring about several positive societal changes, such as increased productivity, better healthcare, and improved access to education. However, AI also has significant downsides and high risks associated with it. The impact of AI on the workforce, ethical considerations, and the potential for misuse or abuse are subjects of ongoing concern.

Artificial Intelligence (AI) is constantly evolving and is predicted to become increasingly pervasive in various industries, such as healthcare, banking, and transportation. AI's impact on society is exciting and challenging, and ensuring its responsible and ethical development and use is crucial.

As AI progresses and gains importance in our world, it is essential to invest in its development and advancement through research, development, and the creation of policies and regulations that promote its responsible use.

The future of AI is unpredictable, and while it holds great promise, it also presents several difficulties and uncertainties. The impact of AI on the workforce, the potential for job displacement, and the need for workers to upskill and reskill are crucial considerations for the future. Additionally, the ethical, legal, and regulatory implications of AI's transformative impact on society must be carefully discussed and prepared for. The development and deployment of AI require the establishment of ethical frameworks and guidelines to ensure

---

[3] https://www.amd.com/en/products/accelerators/instinct/mi300.html

[4] https://habana.ai/products/gaudi2/

[5] https://arxiv.org/abs/2309.17224

that AI remains a positive force in our world, driving progress and improving lives for years to come.

The field of AI is advancing at a rapid pace, which has the potential to generate significant positive changes in various industries and aspects of our lives. However, it also poses challenges and risks that must be carefully addressed. The future of AI holds great promise, but responsible development, ethical considerations, and proactive investment are necessary to guarantee that its impact remains beneficial for society.

Quantum Computing Fundamentals

Another prominent topic at the conference was Quantum, how research uses different techniques to implement Qubits, and how developments of Quantum algorithms can improve software today.

But before getting into the future, there’s a need to explain the fundamentals of Quantum Computing.

Introduction

The history of quantum computing is a captivating narrative that unfolds at the intersection of physics, computer science, and information theory. Rooted in the principles of quantum mechanics, the journey toward developing quantum computers has been marked by scientific breakthroughs and persistent challenges. This exploration begins in the early 20th century with the inception of quantum theory. It follows the evolution of quantum computing to its present-day status as a frontier technology.

Quantum Mechanics

The foundation of quantum computing lies in the work of physicists such as Max Planck, Albert Einstein, Niels Bohr, and Erwin Schrödinger in the early 20th century. Planck's quantization of energy and Einstein's explanation of the photoelectric effect laid the groundwork for quantizing physical phenomena. Bohr's model of the atom introduced quantized electron orbits, and Schrödinger's wave equation provided a mathematical description of quantum states. These foundational concepts, coupled with Werner Heisenberg's uncertainty principle, formed the bedrock of quantum mechanics.

Quantum Computing Emerging

In 1980, physicist Richard Feynman proposed the idea of simulating quantum systems with a quantum computer, recognizing that classical computers faced insurmountable challenges in this regard. Concurrently, David Deutsch developed the concept of a quantum Turing machine, suggesting that quantum computers could solve problems in polynomial time that were intractable for classical computers. The theoretical groundwork was laid, but it wasn't until the 1990s that the term "quantum computing" was coined, and experimental progress began.

Milestones in Quantum Computing

1994 witnessed a breakthrough when Peter Shor devised a quantum algorithm for factoring large numbers exponentially faster than the best-known classical algorithms. Simultaneously, Lov Grover developed a quantum search algorithm with quadratic speedup, showcasing the potential for quantum computers to outperform classical counterparts in specific tasks. These algorithms ignited interest in the broader scientific community, attracting researchers, engineers, and investors.

In 2001, IBM and Stanford University demonstrated Shor's algorithm on a small-scale quantum computer, factoring 15 into 3 and 5 using a 7-qubit, quantum-bit system. This marked the first practical implementation of a quantum algorithm and highlighted the potential for scalability. However, building and maintaining stable qubits posed substantial challenges, leading to the exploration of various quantum computing architectures.

Quantum Computing Fundamentals

The Quantum computing area is an intersection of mathematics, physics, electrical engineering, computer science, and computer engineering. It leverages the principles of quantum theory to perform calculations.
Quantum theory explains the behavior of energy and material at the atomic and subatomic levels, and quantum computing uses physical quantum effects to process information.

The fundamental concepts of quantum computing include:

Qubits: Quantum bits, or qubits, are the fundamental data units in quantum computing. Unlike classical bits, which can only be 0 or 1, qubits can exist in a superposition of states, allowing them to represent multiple values simultaneously. Once observed, the Qubit state collapses into a 0 or 1 value. In other words, the outcome of a quantum calculation is only 0 or 1.
A Qubit can be a logical or a physical qubit with distinct roles and properties.

Physical qubits represent the actual physical components of a quantum computer, such as superconducting circuits, trapped ions, or photons. They are the building blocks from which logical qubits are constructed. Due to their delicate nature, physical qubits are susceptible to errors and noise. At the time of this writing, existing quantum computers leverage physical qubits, so a large volume of work is dedicated to leveraging physical qubits constructively with all their flaws. Furthermore, the number of physical qubits is relatively small, which has led to the term Noisy Intermediate-Scale Quantum, NISQ, being used to describe the current era systems.

Logical qubits are higher-level abstractions representing the manipulated information units in quantum computation. They are expected to be formed by combining many physical qubits into a single logical qubit using techniques like quantum error correction. Logical qubits will be more resilient to errors and noise than individual physical qubits, enabling reliable quantum computation. Logical qubits are an active research area where significant challenges remain. Multiple companies and research groups are working towards systems with reliable logical qubits. It is expected to take several years until such systems are available.
Qubits leverages several physical phenomena to allow for computations.

Superposition: Superposition is a quantum phenomenon that allows qubits to exist in multiple simultaneous states. This property enables quantum computers to, for example, explore a large design space in a single computational step, which is impossible with classical computers.

Entanglement: Entanglement is another quantum phenomenon that allows qubits to be linked so that the state of one qubit depends on the state of another, even when long distances separate them. This interdependency can form complex operations graphs, allowing quantum computers to perform arbitrary calculations.

Quantum algorithms can be designed based on superposition and entanglement. Several algorithms have been proposed as theoretically more efficient than classical algorithms.

Quantum computing can revolutionize various fields, including healthcare, finance, and scientific research, by solving complex problems more efficiently than classical computers. However, practical quantum computers are still being developed, and many challenges remain regarding technology and implementation.

Quantum Hardware

Advancements in quantum hardware became a focal point in the early 21st century. Diverse approaches emerged, including superconducting circuits, trapped ions, and topological qubits. Companies have made significant strides in developing increasingly sophisticated and stable qubit architectures. Such companies include IBM, Google, Rigetti, and IonQ in the US and IQM and AQT in Europe. In addition, several different technologies are being developed, with many companies investing in research and development.

Summary

The history of quantum computing is a saga of scientific ingenuity, theoretical breakthroughs, and technological innovation. From the early days of quantum mechanics to the recent strides in quantum hardware, the journey reflects the collaborative efforts of physicists, mathematicians, computer scientists, and engineers. As the field advances, the promise of quantum computing to revolutionize computation and solve problems currently deemed intractable looms ever larger. The quantum domain is dynamic, with active developments in both industry and academia. Significant achievements have led to the current "noisy intermediate-scale quantum" NISQ era with commercial offerings available.

The Future of Quantum

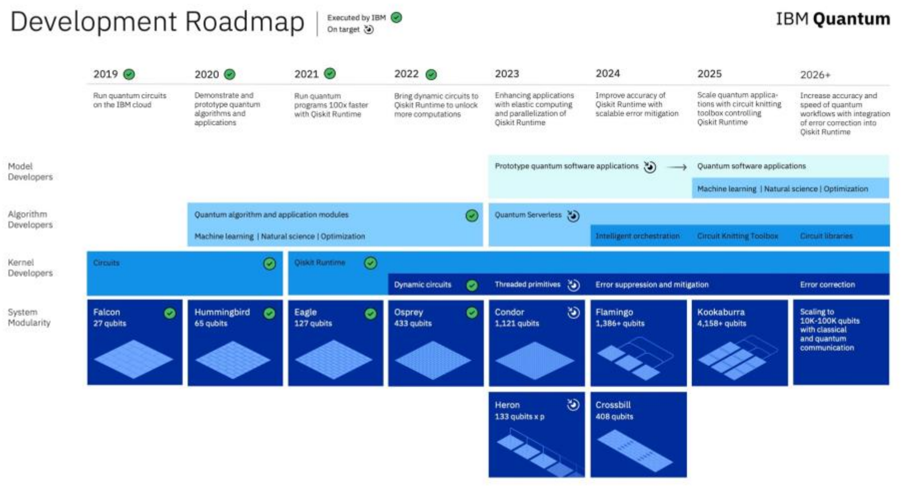
People in the science computational landscape probably know we are far from valuable quantum computers, and the timeline is still being determined. But that is not the same as the time isn’t right to start investing and building the things needed to make quantum computing successful. The long-term potential is incredibly high.

At the same time, there are fundamental questions in the algorithms and application space around what quantum computers will be most beneficial for, where they'll fit into today's scientific computing ecosystem, what the requirements are in terms of scale, performance, and speed for the quantum computer to have an advantage over the best classical methods, and maybe what type of architecture or what type of connectivity is best suited to quantum computing for a particular problem.

At the SC23 conference, a diverse range of quantum computing service providers showcased their capabilities, emphasizing promising initiatives with potential for future advancements.

Among these providers, IBM emerged as one of the figureheads in quantum computing development, demonstrating significant progress across hardware, software development, and various frameworks. While their innovative endeavors were evident, IBM presented their visions and strategies for the field through private engagements rather than showcasing them in the open exhibition.

In an NDA meeting with IBM, the extent of their technological advancements became apparent. They unveiled their five-year roadmap, providing a glimpse into their future endeavors. We will use their roadmap as an example of a roadmap of an innovative company. There are several competing companies also present at SC23. The IBM roadmap highlights the challenges in the quantum computing landscape, particularly calibration and error correction, which are anticipated to be addressed within the next two to three years. These limitations currently hinder the widespread adoption of quantum computing applications. Five-year detailed development Roadmap[6]

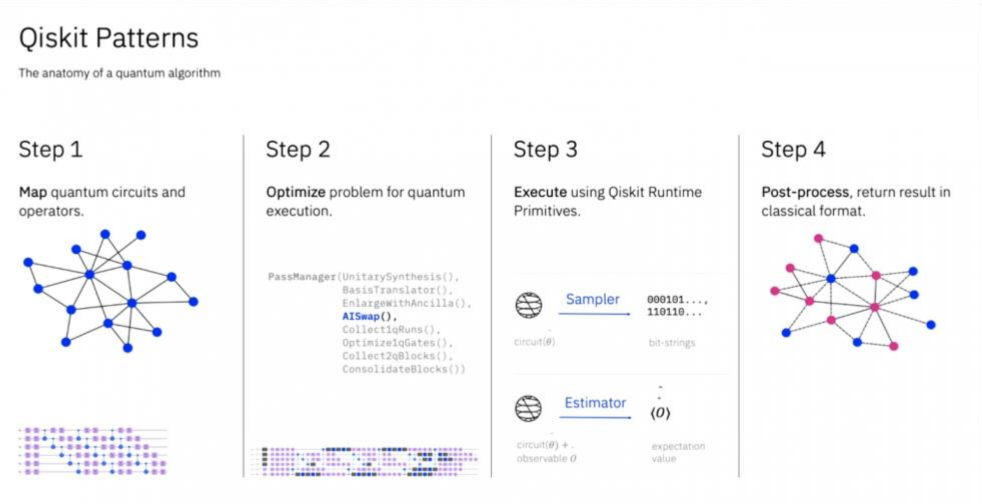


Alongside hardware advancements, IBM emphasizes the importance of software development for quantum computing. While the Python programming language has been the primary tool for quantum computing software, IBM acknowledges the need for further development and growth in this area. The company assures that robust and comprehensive software tools are in the pipeline to address the growing demands of the quantum computing ecosystem.

Within Qiskit, IBM will introduce the concept of "Qiskit Patterns," a simple algorithm development framework. It is a way of mapping a problem to quantum circuits and operators, optimizing the problems for quantum execution, executing them on the runtime, powered by the System Two infrastructure, and then post-processing results to get a simple output.

---

[6] https://www.hpcwire.com/2022/12/01/ibm-quantum-summit-osprey-flies-error-handling-progress-quantum-centric-supercomputing/



Qiskit Patterns

The anatomy of a quantum algorithm

Step 1

Map quantum circuits and operators.

Step 2

Optimize problem for quantum execution.

```
PassManager(UnitarySynthesis(),
            BasisTranslator(),
            EnlargeWithAncilla(),
            AISwap(),
            Collect1qRuns(),
            Optimize1qGates(),
            Collect2qBlocks(),
            ConsolidateBlocks())
```

Step 3

Execute using Qiskit Runtime Primitives.

Step 4

Post-process, return result in classical format.

In 2021, when the 127-qubit Eagle quantum processor was introduced, IBM worked on the "Quantum Serverless," a new programming model for leveraging quantum and classical resources.[7] Their visionary programming model will interact seamlessly with the users' workflows, as they can focus on coding without worrying about the deployment and infrastructure. That equals a serverless architecture, and from an IBM perspective, that indicates four key attributes:

- Coding only, no need for infrastructure management consideration from a developer's point of view.
- Everything is a cloud service.
- No capacity or life cycle management considerations and scales seamlessly.
- Users pay only for consumption, never for idle time.

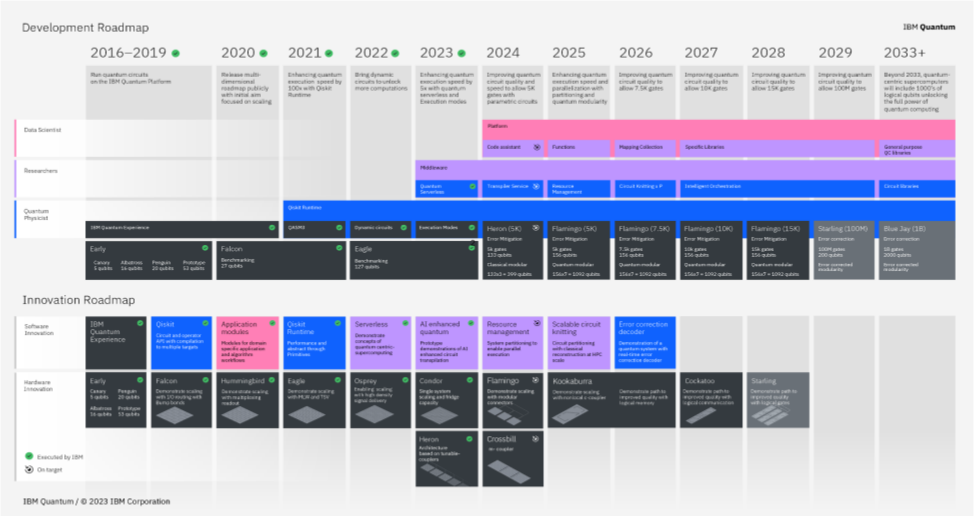
This architecture allows quantum and classical to work in parallel: Quantum to call classical and classical to call Quantum. Capabilities such as circuit knitting, error mitigation, and circuit embedding to become part of the standard development workflow by working with system parallelism.
With Qiskit patterns and Quantum Serverless, users can build, deploy, run, and share for other users to use in the future.

Ten-year Development and Innovation Roadmap[8]

---

[7] https://research.ibm.com/blog/quantum-serverless-programming

[8] https://www.hpcwire.com/2023/12/04/ibm-quantum-summit-two-new-qpus-upgraded-qiskit-10-year-roadmap-and-more/?utm_source=HPCwire+Newsletter&utm_medium=email&utm_campaign=&utm_term=0795B9355467H5T&oly_enc_id=0795B9355467H5T



Observations

We now reflect on IBM's input and input from other companies and researchers. As indicated in the “five-year roadmap” (page 1), we are entering the state of “Error suppression and mitigation,” and the “Ten-year development and innovation roadmap” shows the supporting strategy technologies to reach a stage where we can use logical qubits. As explained earlier, physical qubits are the underlying hardware components of quantum computers.
In contrast, logical qubits are abstract units of information that are manipulated and processed. They are constructed using error-correcting codes to enhance their robustness against errors, enabling the practical implementation of quantum computations.
The advent of quantum computing holds immense transformative potential and is poised to revolutionize various domains, including medicine, finance, and material science. However, the path to realizing this transformative power has its challenges. Researchers grapple with critical challenges that must be surmounted before quantum computing can be fully harnessed for practical applications.

One of the most formidable obstacles is quantum decoherence, a phenomenon that erodes the unique properties of qubits, the fundamental units of quantum information. Unlike their classical counterparts, Qubits can exist in a superposition, simultaneously representing multiple states. However, this delicate superposition is susceptible to environmental interactions, causing qubits to lose their quantum coherence and collapse into classical bits. Addressing this issue requires innovative techniques to isolate qubits from external noise and maintain their fragile quantum states.

Another major challenge is quantum error correction. Despite efforts to safeguard qubits, errors inevitably arise during quantum computations. These errors can accumulate, leading to unreliable and erroneous results. Quantum error correction, a complex and burgeoning field, aims to mitigate these errors by introducing redundancy into quantum systems, enabling the detection and correction of errors before they compromise the integrity of computations.
Scalability is another critical challenge. Quantum computers, to be truly impactful, necessitate a large number of qubits. However, constructing and maintaining large-scale quantum systems takes time and effort, demanding significant resources and technological advancements. Researchers are devising more efficient and scalable quantum computer architectures to overcome this challenge and foster the development of practical quantum

computing platforms.

In addition to hardware advancements, software development poses a significant challenge. Quantum computers operate fundamentally differently from classical computers, necessitating the development of specialized software tools and programming languages. Researchers are working on creating a comprehensive software ecosystem that aligns with quantum computing's unique characteristics, enabling efficient programming and harnessing the full potential of these powerful machines.

The field of quantum computing stands at the precipice of transformative breakthroughs, yet its realization hinges on overcoming a range of critical challenges. Quantum decoherence, quantum error correction, scalability, and software development demand concerted efforts from researchers across various disciplines. As advancements in these areas continue, quantum computing is poised to revolutionize industries and reshape our understanding of computation.

Finally, Quantum computers will likely retain classical computers. For the foreseeable future, they will be used as accelerators for HPC systems. This poses additional challenges. How do we go from the current Quantum computers, built as standalone systems, to having quantum computers integrated into HPC systems? How do we merge the currently separated software stacks, quantum, and HPC, into a single software stack with single source programming and high-level abstractions? How do we prepare the quantum computing technology to be integrated into a data center with its cooling, power, and monitoring infrastructure? These questions were addressed at Supercomputing with directed sessions.
Denmark and DTU are leading in this area, with Sven Karlsson organizing the session as part of the HPCQC.org initiatives he co-founded with Martin and Laura Schulz from TUM and LRZ, respectively.

But before we rush into “the Quantum Era,” it is the time to ask some fundamental questions and learn from the previous years with HPC technology and its anchoring within our research community and society.

As a research community, we need to approach this ”Quantum Era” more holistically and ask ourselves: How do we create the necessary ecosystem, community, and infrastructure to support the sustainable development of the technical advancements within Quantum and the infrastructure surrounding these advancements? How do we form a single community from the different mathematics, physics, computer science, and electrical engineering fields? What is a groundbreaking technology like quantum worth if we cannot explain the benefits to our researchers? How will our research community and, thus, society ever harvest the potential of quantum technology if we fail to enable our researchers to understand and utilize this technology?

We should ask ourselves how we, as a research community, build the correct infrastructure and ensure a robust set-up for our researchers, where we have the competencies to provide the necessary access, knowledge, onboarding, and support to quantum facilities and technology as well as the qualified competencies (early career program, diverse workforce, educations, etc.).

Also, these questions were addressed at Supercomputing with a session organized by the IEEE Quantum-HPCWorking group. Also, here, there is active participation from Denmark and DTU from Sven Karlsson, who, in a related context, has also co-founded the European Quantum Systems and Software Summits, EQS3.

STORAGE

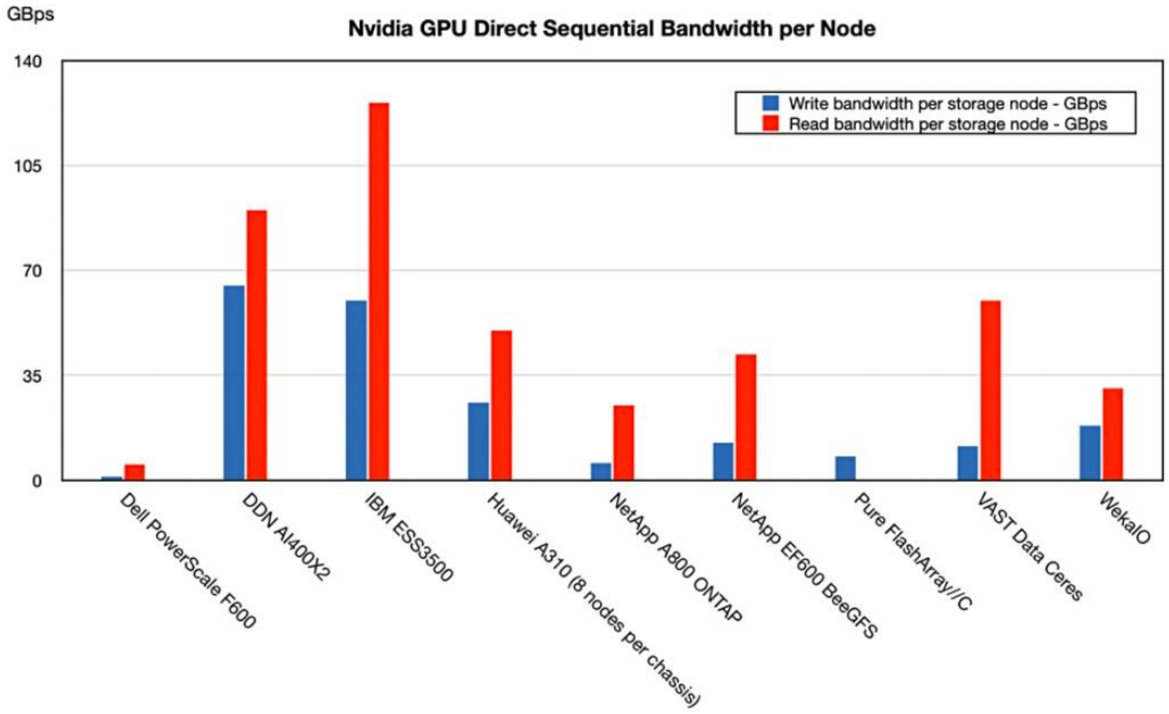
In the university world, we are in an industry where the need for storage constantly increases, with performance, capacity, and security being essential parameters.

IBM

The delegation had arranged a briefing at IBM, an established company with over 100 years of history. IBM strongly focuses on High-Performance Computing (HPC), particularly at IBM Scale.

HIGH-PERFORMANCE PARALLEL FILESYSTEM

We were introduced to IBM Scale (GPFS), their high-performance parallel file system, which can be implemented as pure software and appliances. IBM Scale is a parallel file system known for its extremely high performance and scalability. It provides access to data through various protocols such as NFS, SMB, S3, POSIX, and NVIDIA GPUDirect Storage. In the latest performance tests on an NVIDIA SuperPOD, the IBM Scale platform delivered an extreme performance of 126GB/s on an IBM ESS3500.
The latest IBM Scale system (ESS6000) roadmap was presented and can deliver a system performance of 250GB/s.



(https://blocksandfiles.com/2023/08/15/ibm-nvidia-gpu-data-delivery/)

CONTAINER WORKLOAD

After IBM's acquisition of Red Hat, the IBM Fusion platform has been developed for running container workloads with an underlying architecture based on Red Hat/Openshift, along with specially designed Fusion software. This Fusion software ensures backup and High Availability/Disaster Recovery (HA/DR), which a standard Openshift installation typically cannot offer. IBM Fusion is an HCI (Hyper-Converged Infrastructure) appliance with a validated design, enabling quick and easy implementation and providing simple administration where all components are monitored and managed from the same GUI.

IBM Fusion is a container-native hybrid cloud data platform that offers simplified deployment and data management for Kubernetes applications on the Red Hat OpenShift Container Platform.

USE CASES

- Application modernization
- Databases, analytics, logging and monitoring
- AI/ML, analytics, and data Lakehouse.

Liquid Cooling

Due to the rapid growth in power consumption in HPC systems, especially with accelerators like GPUs, it becomes increasingly difficult to keep systems running at acceptable temperatures with only classic air cooling. For several years, alternative cooling solutions have been investigated to address that.

At SC23, many vendors presented their solutions for liquid cooling. They ranged from indirect rack back-door cooling to direct liquid cooling of individual components with cooling pipes or similar direct cooling of entire systems with full immersion in oily liquids not considered harmful to electronics and enclosures.

Back-door cooling has the appealing strict isolation of liquid from the actual fragile hardware because it only cools the air near the equipment and helps the flow. Vendors already have such fully functional solutions shipping, and a few were on display with significant cooling capabilities in a form that didn't appear to hinder daily maintenance and physical siting of equipment. It is known to work for current installations and should continue to be suitable for many workloads in the near future, including in university data centers. The logical question is how long that solution will suffice as systems grow increasingly compact and power-hungry. Still, even if it can't keep up, it may assist in keeping racks cool in scenarios where direct liquid cooling doesn't completely cut it.

We also saw full immersion systems on display with lots of heat bubbling off the silicon into the oily liquid. One could quickly get associations of a deep fryer in action, but it appears to work.

As several university participants remarked, it is likely very efficient and primarily nice with the full immersion model, at least right up to the point where hardware maintenance is required. Having to lift greasy/wet systems out of the cooling container and do service on them, on the other hand, quickly becomes a serious concern in data centers, especially regarding potential drips, spills, and perhaps any dangerous fumes from the boiling liquid. Such concerns will definitely require rethinking of a number of procedures and daily system administrator tasks. The alternative for integrated cooling is the more traditional cooling tubes mounted on equipment or integrated directly as channels in the hottest components like GPUs and CPUs.

In general, the group was concerned about the lack of standardization of the liquid cooling solutions and the resulting risk of vendor lock-in. Some of those direct liquid cooling systems may be interchangeable. Still, the devil is in the details, and the common impression is that it quickly becomes a nightmare to change anything later because of differences in actual cooling liquids, piping, and handling.

With pipes and channels inside the equipment, another major concern among the participants was the risk of leaks and the damage they could cause in a data center. One novel idea for preventing exactly that was on display at the exhibition. Chilldyne showcased a leak-free liquid cooling technology using a negative pressure gradient. The basic idea behind this is to have negative pressure (suction) at BOTH ends of the cooling tube, but with one end having more negative pressure to keep a (strained) flow running. The result is that even if a cooling pipe or tube breaks or cracks, the negative pressures will pull the liquid back to the source tank on both sides of the leak point. They had a simple but effective hands-on demo on display where we were handed a pair of scissors to cut such a tube while the liquid was flowing and for ourselves to see the liquid getting immediately sucked away from the resulting hole. Similar to the small video at the top of their website.[9]

For further protection, active flow monitoring can detect leaks and trigger automatic suspension or shutdown of the system to avoid overheating when cooling cannot continue. Additionally, upon replacing the damaged tube section after the cut, the cooling flow automatically started again. So, in principle, hot-plug replacement of the cooling tubes would even be possible.

Participants were also concerned that systems with liquid cooling are not 100% cooled that way, leaving significant air-cooling headaches about handling the few remaining percent of excess heat. With large enough power envelopes, those few percent can add up to a lot. Perhaps the rack back-door cooling concept could still help address such concerns.

Without going into any nitty-gritty NDA-covered details, vendors are at least aware that the trend in miniaturization and power envelopes will put a lot of pressure on them both in terms of upscaling rack power capacity to fit the requirements and the resulting cooling challenges it will pose.

---

[9] (https://chilldyne.com/).