DeiC

DANISH eINFRASTRUCTURE CONSORTIUM

ISC 2025 conference

DeiC TekRef report

Joachim Sødequist - joachim.sodequist@deic.dk
Nina Simone Marstrand Sander Aagaard - nsmsa@its.aau.dk
Rainer Bohm - rb@its.aau.dk
Rasmus D. Jensen - rj@its.aau.dk
Rune Gamborg Ørum - rune.oerum@deic.dk
Tor Skovsgaard - tor.skovsgaard@deic.dk
Julius Roeder - julius.roeder@deic.dk

June 2025

Contents

Preface

This report summarizes the main takeaways collected and by the delegation. With the purpose of disseminating the knowledge about the latest HPC systems, technology, and software for use by DeiC as well as for the broader HPC community in Denmark.

This year was the first time at ISC for the majority of the delegation, and we managed to cover a broad range of the activities at ISC, with the main focus on Birds of a Feather sessions, invited talks, vendor meetings and networking.

1. Overall themes from sessions and talks

1.1. Sustainability

Energy efficiency is becoming a primary constraint for modern HPC systems. The traditional bottlenecks are still there, but the focus is shifting from sustainable systems to 'viable' systems. As the jobs increase in size, delivering performance with a lower power draw becomes essential and requires scaling across the entire stack, not just the HW.

The sustainability theme permeated the conference's opening keynotes, with energy efficiency discussions spanning from Papermaster's technology presentation alongside Oak Ridge examples on Frontier, to an entire keynote devoted to climate modeling. Sustainability is not just environmentally responsible but economically essential.

In the same vein, the industry is transitioning from a pure scale-focused approach to one emphasizing efficiency, usability, and trustworthiness.

Instead of competing primarily on model size or computational capacity, the focus is shifting toward "smart orchestration, hybrid workflows, and many more things that come into play." The emphasis on trustworthiness reflects growing concerns about AI safety, reliability, and governance in production systems.
The next wave of innovation will focus on elegant solutions that optimize across multiple dimensions rather than maximizing any single metric.

1.2. HPC and AI

While HPC and AI are technically converging at the infrastructure level, significant organizational and cultural barriers persist.

From a session: "AI is just a different HPC workload… but the point is that it's this one workload that has become so dominant that it's driving all of the investment in the space right now." The challenge is that while HPC practitioners view AI as another workload, the broader AI community doesn't reciprocate this perspective, creating an asymmetrical relationship that affects resource allocation and strategic priorities.

1.3. Orchestration

At the conference floor there were many emerging vendors in the orchestration space. Alongside these developments were emerging collaborative groups and coordination initiatives. There is a strong consensus that this new market is developing fast and the need for standards and slowing down the fragmentation is strong.

Runes thoughts: But there still seems to be no clear market leaders, and getting a fuller overview of all the vendors and how they integrate in the stack seems to be a task on its own, that could perhaps be interesting to tech-ref?

The CIS Sage software framework represents a practical breakthrough in quantum-classical integration, enabling dynamic resource allocation across CPUs, GPUs, and QPUs within the same system. This addresses the complex computer science challenge of "balancing three kinds of shared resources" where calculations may be suspended while waiting for quantum devices, similar to early GPU integration challenges.
Orchestration of the storage layer is a huge part of these developments.

1.4. Storage

The resurgence of tape storage is driven by practical considerations in scientific computing. Climate researchers and other simulation communities are "unwilling to compress their data and want to keep centuries of simulated data," creating enormous storage requirements. As data doubles annually, the need for cost-effective cold storage becomes critical.

Active archive capabilities have matured significantly, with LTFS (Linear Tape File System) enabling users to "mount your tape and get a directory listing of what's there," making tape storage more accessible for hybrid storage architectures that combine hot, warm, and cold storage tiers.

1.5. New silicon

The semiconductor landscape is diversifying beyond the traditional x86-NVIDIA duopoly.
RISC-V is gaining commercial traction, with David Keys noting that "RISC-V is now on the market with Tense Torrent… It's open source, cost-effective computing." While still emerging, this represents a significant shift toward open hardware architectures that could reduce vendor lock-in.
Wafer-scale computing solutions also gained attention, with companies like Cerebras, Groq, and SambaNova demonstrating alternative approaches to traditional chip architectures. As Addison observed, the AI market's dominance has "cleared the lane a little bit" for HPC-focused companies like "Tenstor or Cypl or Next Silicon" to establish their positions in specialized computing environments.
The emergence of these alternatives is particularly significant in energy-sensitive environments where efficiency trumps raw performance, creating new competitive dynamics in the semiconductor space.

2. Specific vendor and product developments

2.1. VAST

Hardware warranty of 10 years. Uses QLC Flash which is cheaper than QRL. Only platform that has full auditlog incorporated. Can be made custom and flexible. Full auditlog takes less than 5% on the performance.
Encryption storage, multi-tenant data access, phallos cypertrust integration.
99,998% uptime for system *"The VAST DASE architecture is designed to deliver 99.99999999% uptime in data centers on systems that can scale*
into the exabyte range.(VAST Data Platform: Unified Authorization and Access Control with ABAC)

The VAST Catalog's metadata indexing is fully integrated into the VAST Data Platform, supporting NIST controls in several critical areas:

Access Control (AC)
The cataloging and indexing of data assets aid in access control by providing a comprehensive inventoryof data that needs protection. This helps in determining appropriate access levels for different users and ensuring that access controls are effectively implemented.

Audit and Accountability (AU)
By maintaining a detailed record of data access, including who accessed what data and when, the VAST Catalog supports audit and accountability. This level of transparency is essential for identifying potential security incidents and verifying appropriate data usage.

Configuration Management (CM)
The VAST Catalog assists in configuration management by offering a clear inventory of data assets and their attributes. This ensures that data management aligns with security policies and procedures.

Identification and Authentication (IA)
The VAST Catalog enhances identification and authentication processes by providing a quick and reliable method to locate and verify the authenticity of data, helping to prevent unauthorized alterations.

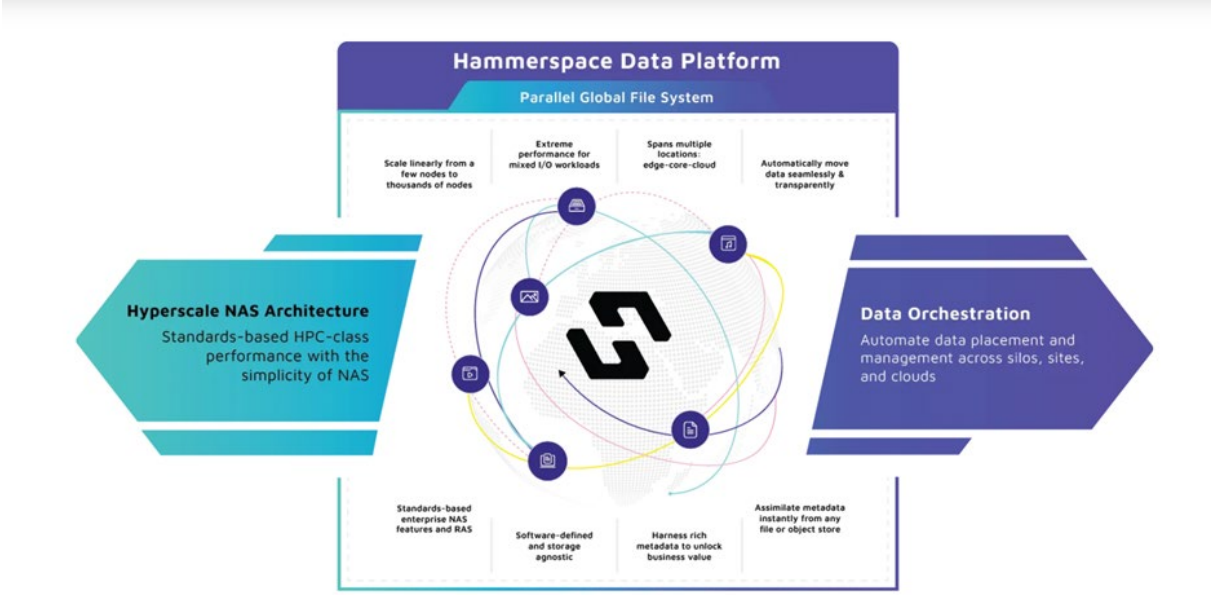
(VAST Data Platform: Unified Authorization and Access Control with ABAC)

2.2. NVIDIA

NVIDIA Deep Learning Institute is a series of courses that NVIDIA provides. These courses seem heavily focused on NVIDIA hardware and software, it would be relevant to check how general purpose they are before investing heavily in them.
NVIDIA has an ambassador program where ambassadors (focused on one field) can then teach the NVIDIA courses to others for "free".
NVIDIA has academic access grants that mainly give out compute time on their clusters.

Networking:
NVIDIA are not selling HDR any longer, NDR (400 Gbp) and moving into the new XDR (800 Gbps) while also still selling ethernet products as spectrum-x
NDR aircooled versions are being finished now. For XDR there are several switch versions: aircooled 4U (144ports, 2ufms on front) and a 2U liquid cooled at the end of the year. They should be able to do up to 10000 nodes and come with ConnectX-8. Ultra ethernet compatibility will take some time and they did not provide any additional information.

2.3. Hammerspace



Sikkerhed:
Ikke fuld auditlog og versionsstyring. Anvender standard protokoller. Mulighed for fuld kryptering (- på NFS det arbejder de på nu) kan se hvilket data, der senest er tilgået bare ikke af hvem.
Role based access rights both on metadata and filesystem access rights can be set on both file and directory level
Drift:
Non-disruptive data orchestration and non-disruptive upgrades.
Needs minimum admin support (vi skal have mere data på dette, hvad ligger der i driftopgaven lokalt)
Besparelser på storage omkostninger, da der ikke bliver lavet "flere kopier af data".
Performance:
500TB per GPU and 2 PTB per server
Integration:
Platformen støtter adskillige leverandører. Både muligt at vælge lokalt og Cloud. Enable Collaboration at Scale Across Edge, Core, and Hybrid-Cloud Environments
Hammerspace har tæt samarbejder og fuld integration til fx. Hitatshi storage system Technology Partners | Hammerspace
Hitachi Vantara | Hammerspace
Forskningsdata/Rigsarkivsopgaven:

Forskningsdata kan deles på tværs af lande og projekter da det styres gennem rollebaserede og rettighedsbaserede adgange med mulighed for selv at sætte rettigheder op som "kun adgang i en uge mv.)
Man har muligheder for at custom tagge og multiple tags på fx metadata niveau.
The platform makes data available and accessible using a feature called Data-in-Place Assimilation. Metadata is copied into the Hammerspace metadata database, while the data itself stays in place, so files and objects are available and accessible quickly, often in a matter of minutes.
Priser

2.4. VDURA

VDURA are technically only the software side of the file-system.
1.2 M iops per node
Flash can provide up to 2.3 Tbps
Up to 26 PB per rack
E2E Encryption available

Flash storage is used to hide the latency of hdd storage in hybrid setups
Raid 6 on large files, smaller files saved in triplicate on SSD to enable fast access

Automatic rebalancing when new hardware is added

Promises good performance with up to 80% HDD storage

HPC vs AI:
Hybrid (flash + HDD) high-performance storage solution, specifically designed for AI and HPC workloads.

Key features include:

Hybrid Approach with Scalable Flash: Supports both large and small files initially on mostly HDDs (starting at ~2% flash), but allows scaling performance by upgrading to an all-flash system or moving data into flash using configurable thresholds (default 1MB, tunable). Files larger than the threshold go to HDDs.

High Performance: Extremely high throughput

AI-Specific Security: Emphasizes security with immutable audit logs, crucial for data provenance tracking in AI environments.

Minimized Downtime: Designed to support fast parallel offline processing (fast parallel system offline) and hide slower HDD speeds via buffering on flash (buffering over flash), aiming for minimal disruption during GPU-heavy AI tasks.

Architecture: Utilizes Director nodes (at least 3, potentially more) running a distributed key-value store to manage metadata and permissions.
The system uses building blocks (plug-and-play) allowing flexibility and easy scale-out by adding storage node types as needed.

In essence, it's a scalable, secure, high-performance hybrid storage solution tailored for AI and HPC workloads, leveraging flash buffers to overcome the speed limitations of HDDs while keeping costs down initially.

2.5. Lenovo

This describes hardware and storage configurations geared towards future AI/HPC environments.

Cooling:

Future systems will use (mostly) liquid cooling. Liquid cooling for memory/CPU but air cooling for interconnects etc. They are also looking at liquid in node and then heat rejection to aircooled system

GPU:
The MI430X generation should provide good FP64 performance. However, MI45xX is probably not well suited for FP64.

Blackwell architecture may need FP64 emulation but the emulation looks promising. (we will see.)

Server & Interconnect:
Lenovo V4 Neptune is a relevant system.
The system will have 3-phase power delivered directly to rack.
16 channel RAM will not fit in a 19 chassi with air cooling. To fit large memory in an air-cooled 19 chassi, SP7 1s concept is being developed to use DIMMs on both side. For very large memory configurations (like 16CH), signal integrity becomes a challenge, which can be addressed by using flipped DIMMs or other techniques. Flipped DIMMs has the potential to save costs.

Vendor & Storage Options:
IBM and Lustre/BeegFS/Ceph/Weka are mentioned asstorage solutions or vendors.
Specific hardware like the Lenovo Spinning disk node exists for certain types of storage.

MPI
MPI 5.0 is out! MPI ABI is now live! Once this gets implemented everywhere you can run any MPI command on any other MPI system!

OpenMP
OpenMP 6.0 is out! Future OpenMP will have a kernel language for complete low-level control

2.6. Cornelis Network

Cornelis network explained their network interconnect roadmap to us.
CN5000, CN6000 and CN7000 were discussed. The highlevel summary is:

- CN5000 --> 400Gb omni path; PCIe Gen5 x16
- CN6000 --> 800G (PCIe 6); ethernet mode (RoCE) instead of omni path; but switching will be omni path; omni path for compute connect; ethernet for storage
- CN7000 --> 1600G (2027); omni path, ultra ethernet, ethernet

CN5000:
- air & liquid cooling;
- switches also air & liquid cooling
- director class switch (air & liquid); unique for them
- 30% lower latency than infiniband;
- use random route diversification to prevent fabric congestion
- fine-grained adaptive routing --> telemetry data informs switch about congestion appearing in the fabric

CN6000
- mid-early 2026
- going to verification at the moment
- Gen6 PCIe x16
- Switch --> dual fabric plane
- CN6000 sits alongside CN5000

CN7000
→ Estimated 2027

- concurrent protocols
- adapters and switch can speak all 3 protocols with each other
- adapter to external switch can do ultra ethernet

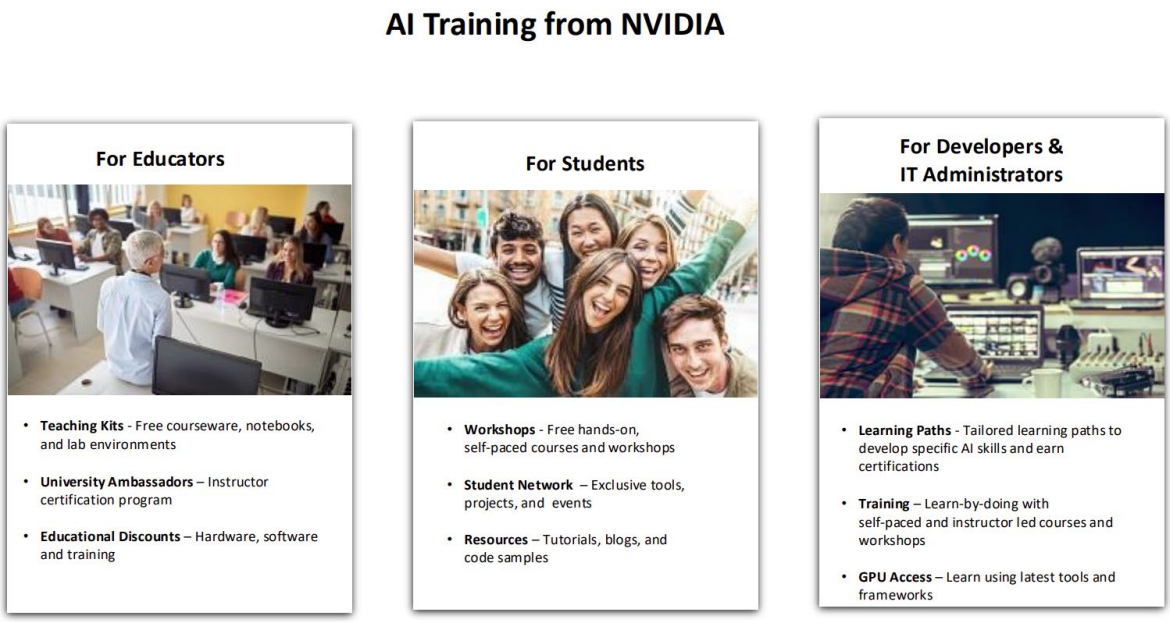
Cornelis believes that they can provide much better performance than infiniband for HPC. For smaller message sizes, they claim 2.7x bandwidth. However, AI message sizes are much larger and the difference in bandwidth is smaller. In comparison to NDR they see lower pinpong latency; better bandwidth and a 2.5x higher message rate. Lastly, Cornelis raw performance is 45% higher and the performance per dollar even better than that.

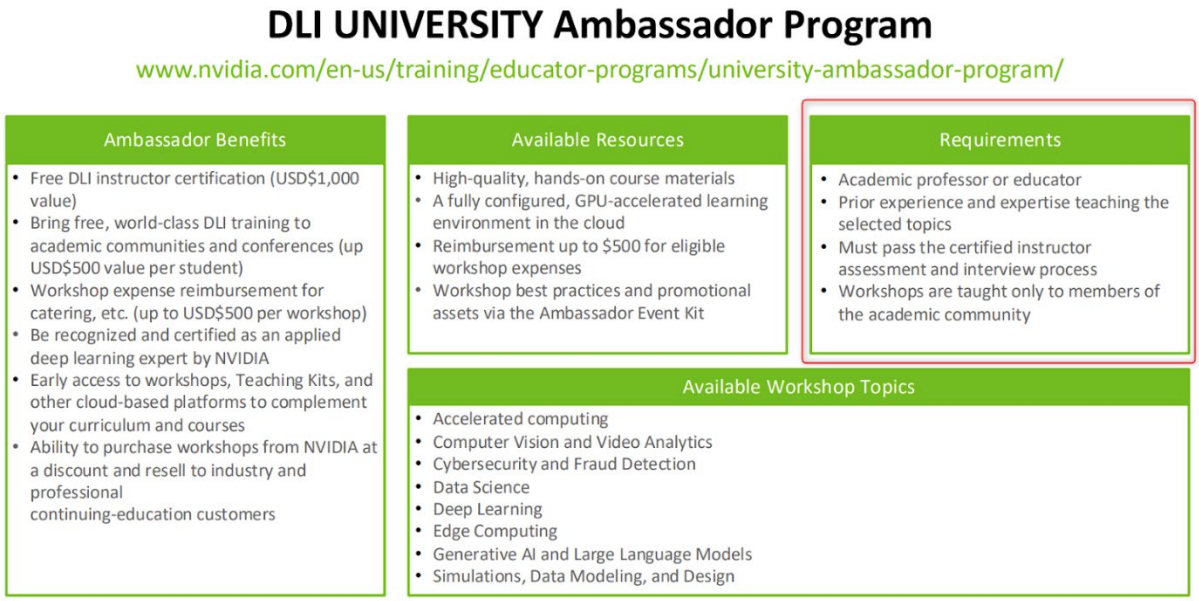
3. Interessante træninger og best practise på HPC området i EU

Træning, onboarding og "self-learning" Hartree: Home Page
NVIDIAs omfattende træningskataloger, som kunder har "gratis" adgang til:'

3.1. NVIDIA

Som Nvidia kunder har vi gratis adgang til et omfattende træningskatalog. Dette kræver vi bliver certificerede nvidia ambassadører (denne træning er også gratis for kunder) og så må ambassadørerne træne resten af organisationen.
Find Training | NVIDIA





4. Europæisk hardware landskab

Det Europæiske Processor Initiativ (EPI) projekt har været undervejs siden 2018, hvor målet har været at designe og producere europæiske RISC-V processorer. Dette inkluderer ikke silicium fabrikation. Projektet har lykkedes med at producere en test chip. Ud fra dette projekt er der udsprunget 3 nye europæiske projekter:

1. DARE. (Nyt stort projekt, Mar 2025 - Feb 2030) Dette projekt er en ligefrem videreudvikling af EPI's RISC-V chip design frem imod en HPC-værdi chip med 2 typer af acceleratorer.
2. EUPEX. (Undervejs, Jan 2022 - Dec 2026) Et pilotprojekt hvis ambition er at integrere noder baseret på den europæiske EPI RISC-V chip i JUPITER HPC anlægget.
3. EUPILOT. (Afsluttet, Dec 2021 - Maj 2025) Også et pilotprojekt som skulle demonstrere brugbarheden af EPI RISC-V chippen.

Danmark deltager ikke i disse initiativer. De store deltagere i disse projekter er Italien, Tyskland, Grækenland og Spanien (i DARE) / Frankrig (i EUPEX). Der er også flere mindre deltagere, især i DARE.

Der er (og har været) flere europæiske projekter inden for hardware området: "NET4EXA" (undervejs), "eProcessor" (Afsluttet), "RED-SEA" (Afsluttet), "MEEP" (Afsluttet) og "TEXTAROSSA" (Afsluttet).