CTO-Vision Panel:
Energy–Efficient Memory, Storage & Networks

FMS (Future of Memory and Storage)

02025-Aug-06. Santa Clara Convention Center

"It costs over 1000× more energy to move a byte than to compute on it."
*— Bill Dally, NVIDIA Chief Scientist**

Theme

As artificial intelligence (AI) workloads grow exponentially, energy consumption has become the paramount challenge in datacenter operations. In April 2025 **International Energy Agency (IEA)** projects that electricity demand from datacenters worldwide will more than double by 2030, reaching approximately **945 terawatt-hours (TWh)** — exceeding Japan's total electricity consumption today.

NVIDIA's Bill Dally highlights* that while a floating-point operation requires about **20 picojoules**, reading 64 bits from mobile DRAM consumes approximately **1,200 picojoules**—a **60×difference**.

Similarly, moving data across racks and interconnects now dominates energy budgets, especially in AI training clusters where over 80% of power may go toward memory and communication rather than compute.

Discussion Points

This panel will bring together CTOs from across the compute, memory, and networking domains to explore:

- The energy cost of data movement vs. computation
- Interconnect bottlenecks and disaggregated chiplet architectures
- Innovations in XPUs/SmartNICs, memory, storage, & CXL fabrics
- What software can do to address these challenges
- Prospects for Reversible and Quantum Computing
- System-level energy accountability: Can we meter what matters?

Conclusion

The stakes are existential. Without a fundamental redesign of how we move and process data, AI infrastructure and cloud scale-out architectures will hit energy ceilings well before reaching their full potential.

Panelists

Panelist	Company
Hannah Earley (CTO)	Vaire Computing
Jason Hardy (CTO)	Hitachi Vantara
Michael Kagan (CTO)	NVIDIA
Rob Lee (CTO)	Pure Storage
Sven Oehme (CTO)	DDN
Alex Veprinsky (CTO)	HPE

Table 1: Alphabetical order of last name

*A 64-bit floating-point operation consumes $\sim 5-20pJ$ per operation in modern silicon—significantly less than the 10s of nJ required to ship a single bit across a data center link.

Bill Dally emphasizes that optimizing for data locality—i.e., keeping data as close to the compute as possible – can save huge amounts of energy compared to frequently pulling data from far-flung locations in a data center.

This guides our architectural decisions. Knowing that data center-scale communication can cost over 1,000x more than on–chip data movement influences how datacenter systems and software are designed.

The scale of a single AI datacenter is constrained by its power plant supply capacity. As demand for training resources grows, hyperscalers are exploring strategies to utilize the compute capacity of multiple datacenters within a single pre-training job [SDR-RDMA.] https://arxiv.org/pdf/2505.05366

Michael Kagan (CTO) NVIDIA

Michael Kagan is the CTO (Chief Technology Officer) at NVIDIA since May 2020. He joined NVIDIA through the Mellanox acquisition. Michael was previously the CTO and co-founder of Mellanox, which was founded in April 1999. From 1983 to April 1999, Michael held a number of architecture and design positions at Intel Corporation. While at Intel, Michael was the architect of the i860XP vector processor, managed Pentium MMX design and managed the architecture team of the Basic PC product group. Michael holds a BSc. in Electrical Engineering from Technion – Israel Institute of Technology.

Figure 1: Michael Kagan – NVIDIA

Hannah Earley (CTO) Vaire

Dr. Hannah Earley is the Chief Technical Officer and Co-Founder of Vaire Computing, a startup pioneering near-zero-energy, reversible-computing chips for AI infrastructure. She earned her PhD in Applied Mathematics & Theoretical Physics from the University of Cambridge, where her thesis explored reversible molecular computation and scaling laws for physical computing systems. Joining Vaire in 2021, she leads hardware architecture and foundational research to dramatically reduce energy dissipation in silicon chips. Dr. Earley also serves as an Affiliate Lecturer at Cambridge, teaching computational biology and biodesign. An innovator and community builder, she co-founded the Molecular Programming Interest Group and has authored multiple publications on reversible computing.

Figure 2: Hannah Earley – Vaire

Rob Lee (CTO) Pure Storage

Rob Lee serves as Chief Technology Officer (CTO) at Pure Storage, focused on global technology strategy, and identifying new innovation and market expansion opportunities for Pure. Rob joined Pure in 2013 as part of the FlashBlade founding team and led the software architecture and development, from concept to launch, for the product which has eclipsed $1B in business. Prior to Pure, Rob was an Architect at Oracle, where he worked on programming language runtimes for the Oracle RDBMS, as well as high-performance distributed transaction processing systems. During his 12-year tenure at Oracle, he focused on improving the performance, reliability, and programming models behind large-scale distributed systems. Rob holds over 75 patents in the areas of distributed systems, language runtimes and storage systems.

Figure 3: Rob Lee – Pure Storage

Sven Oehme (CTO) DDN

Sven began his career at IBM in 1993, and has worked in multiple disciplines over the last 30 years. From Linux Virtualization, and Storage Virtualization to Filesystems, Sven led the team driving most of the performance improvements on Spectrum Scale (GPFS) over the last 10 years. As part of the core architecture team, Sven was responsible for metadata, streaming IO and specific analytics optimizations. Sven joined DDN in 2018 as Chief Research Officer driving innovation across DDN's existing and future product portfolio.

Figure 4: Sven Oehme – DDN

Jason Hardy (CTO) Hitachi Vantara

As Chief Technology Officer for Artificial Intelligence, Jason Hardy is responsible for the creation and curation of Hitachi Vantara's AI strategy and portfolio. He is defining the future and strategic direction of Hitachi iQ, the company's AI Platform, and cultivating a level of trust and credibility across the market by fostering strong working relationships with customers and partners, and leading public facing events. Jason represents the company externally by communicating the company's vision and value proposition for AI and by collaborating with key partners to develop comprehensive go-to-market strategies.

Figure 5: Jason Hardy – Hitachi Vantara

Alex Veprinsky CTO HPE

Alex Veprinsky is the Chief Architect for HPE Storage, responsible for end-to-end architecture and design across the business unit's product portfolio. He joined HPE's 3PAR division in 2016 as a Distinguished Technologist, following a role as Distinguished Engineer at Dell/EMC, where he helped develop and architect major storage platforms. Earlier in his career, Alex worked at Comverse Technology on enterprise telecom services. A prolific inventor, he holds numerous patents in storage and computer science, and is deeply engaged in mentoring technical teams and driving foundational research. Alex holds a BSc in Electrical Engineering from Technion — Israel Institute of Technology.

Figure 6: Alex Veprinsky – HPE