

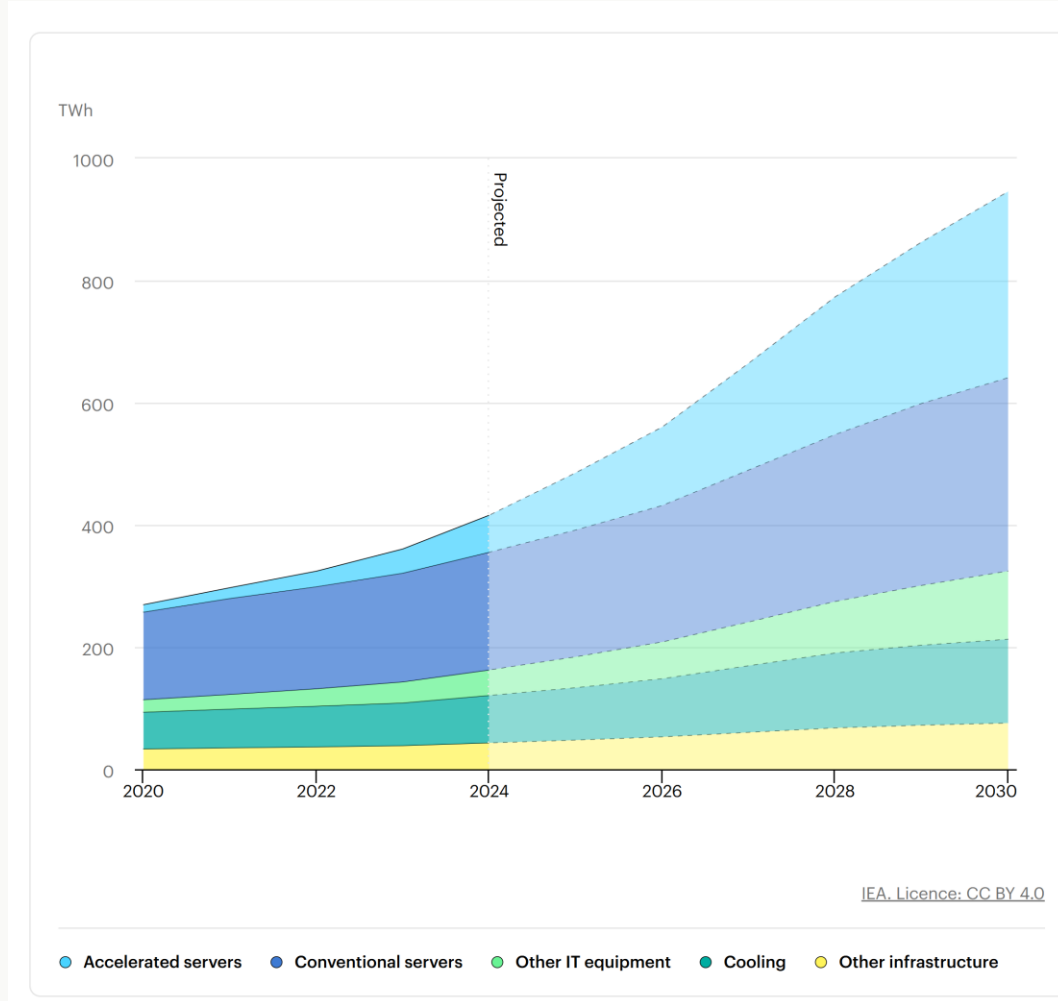
The Next Watt saved won't be from a CPU or a GPU

Devashish Purandare (dpurandare@microsoft.com)

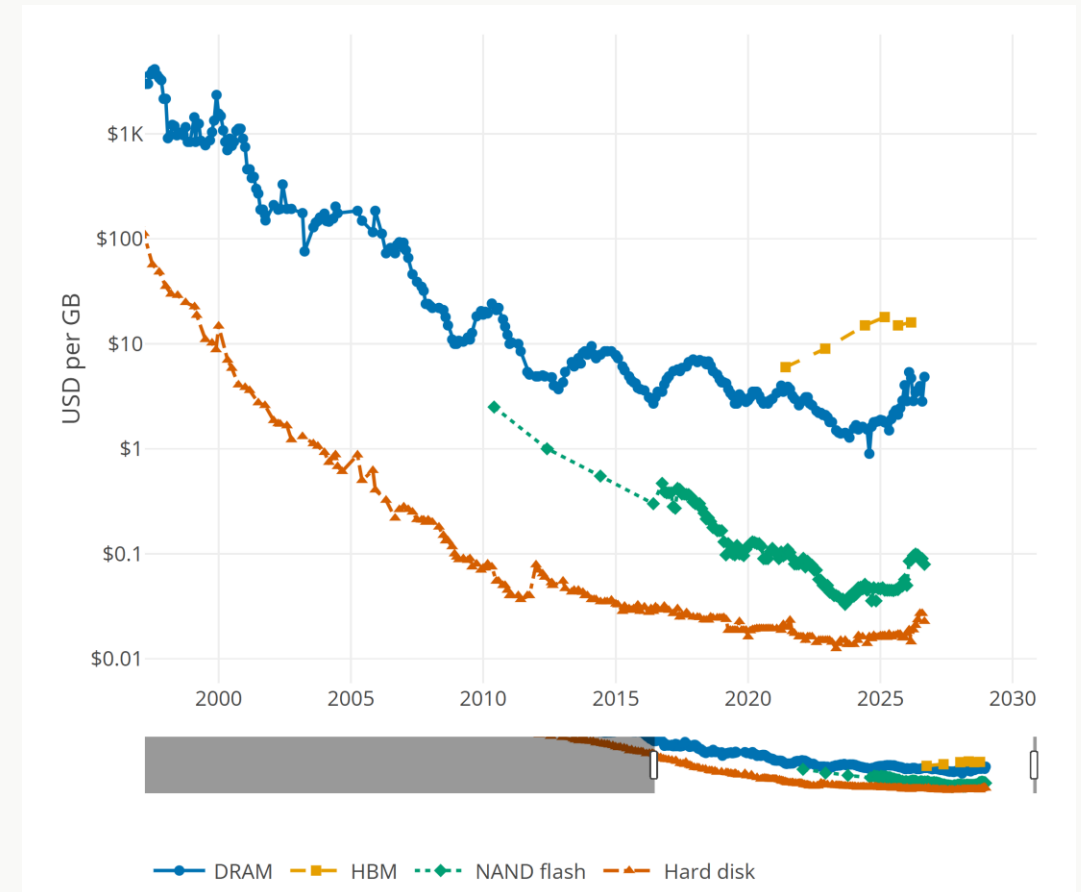
Azure Storage



THE WORLD IS ENDING



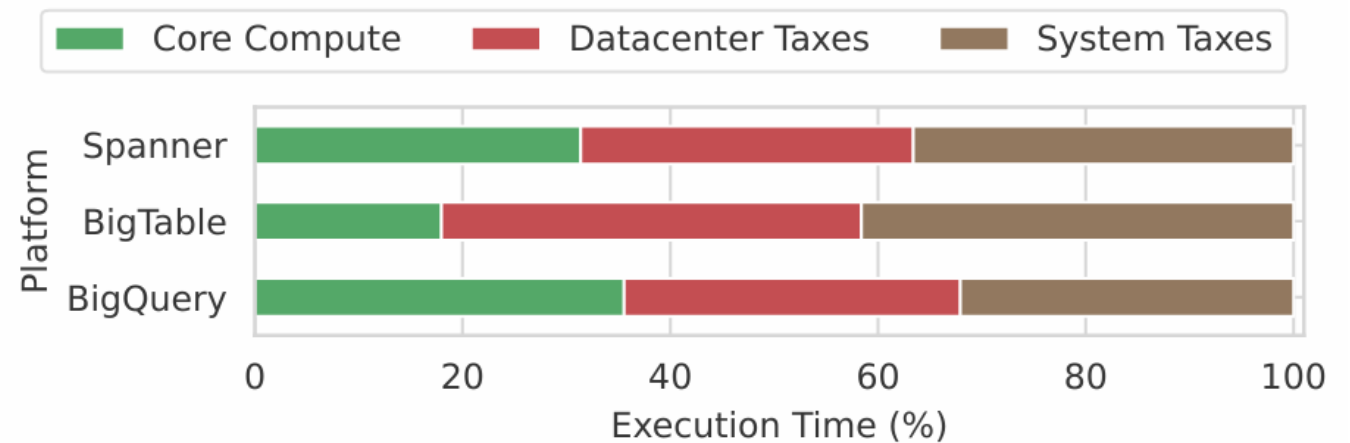
IEA (2025), Global data centre electricity consumption, by equipment, Base Case, 2020-2030, IEA, Paris <https://www.iea.org/data-and-statistics/charts/global-data-centre-electricity-consumption-by-equipment-base-case-2020-2030>, Licence: CC BY 4.0



<https://dam.stanford.edu/memory-prices.html>

Data Center Taxes

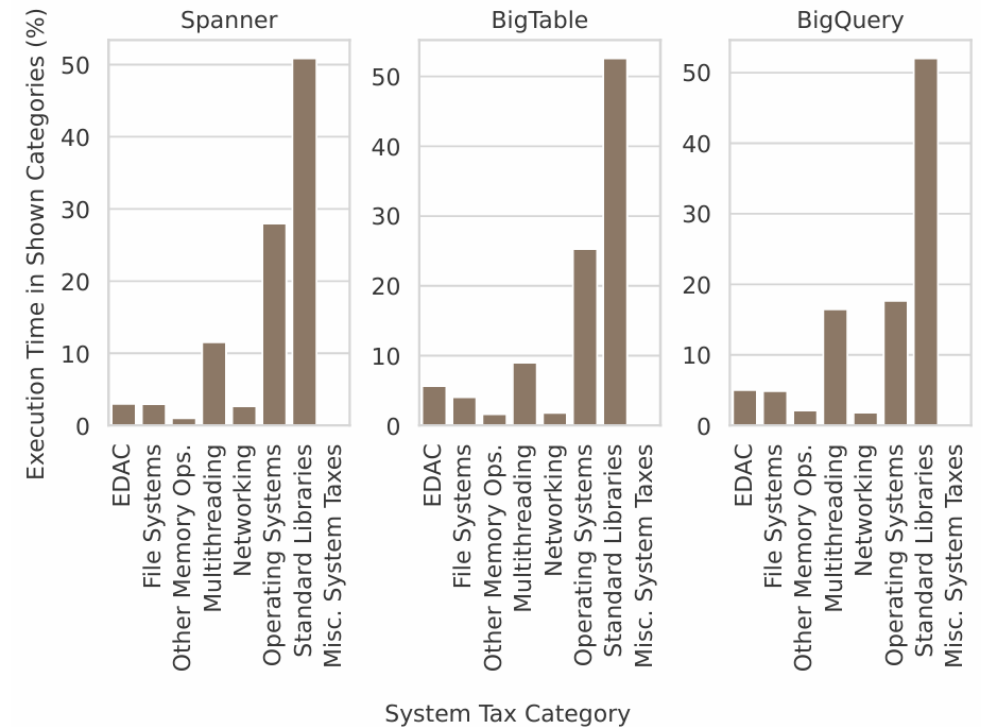
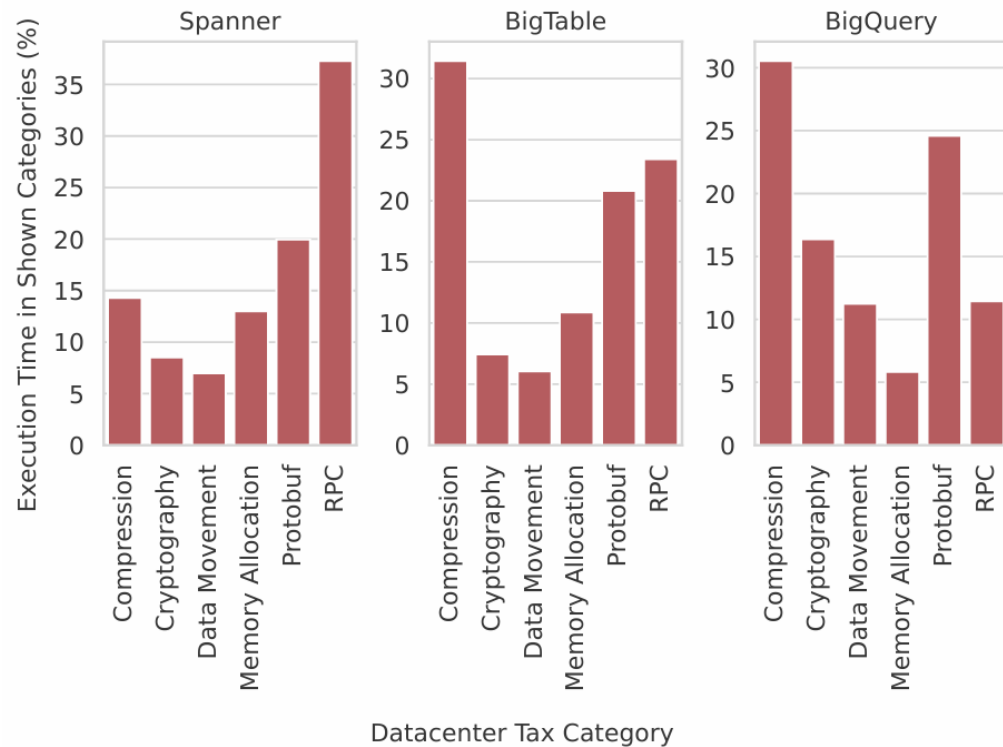
- “Tax cycles” consistently comprise 60-80% of all execution.
- **Data Center Tax**
 - Compression, Crypto
 - Data Movement
 - Memory Allocation
 - Protocol and RPC
- **System Taxes**
 - Errors, Checksums
 - Filesystems, Kernel
 - STL,
 - Networking and Multithreading



Abraham Gonzalez, Aasheesh Kolli, Samira Khan, Sihang Liu, Vidushi Dadu, Sagar Karandikar, Jichuan Chang, Krste Asanovic, and Parthasarathy Ranganathan. 2023. Profiling Hyperscale Big Data Processing. In Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23). Association for Computing Machinery, New York, NY, USA, Article 47, 1–16. <https://doi.org/10.1145/3579371.3589082>

Wait, it gets worse!

- No single category dominates



Abraham Gonzalez, Aasheesh Kolli, Samira Khan, Sihang Liu, Vidushi Dadu, Sagar Karandikar, Jichuan Chang, Krste Asanovic, and Parthasarathy Ranganathan. 2023. Profiling Hyperscale Big Data Processing. In Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23). Association for Computing Machinery, New York, NY, USA, Article 47, 1–16. <https://doi.org/10.1145/3579371.3589082>

Scaling Storage is Hard



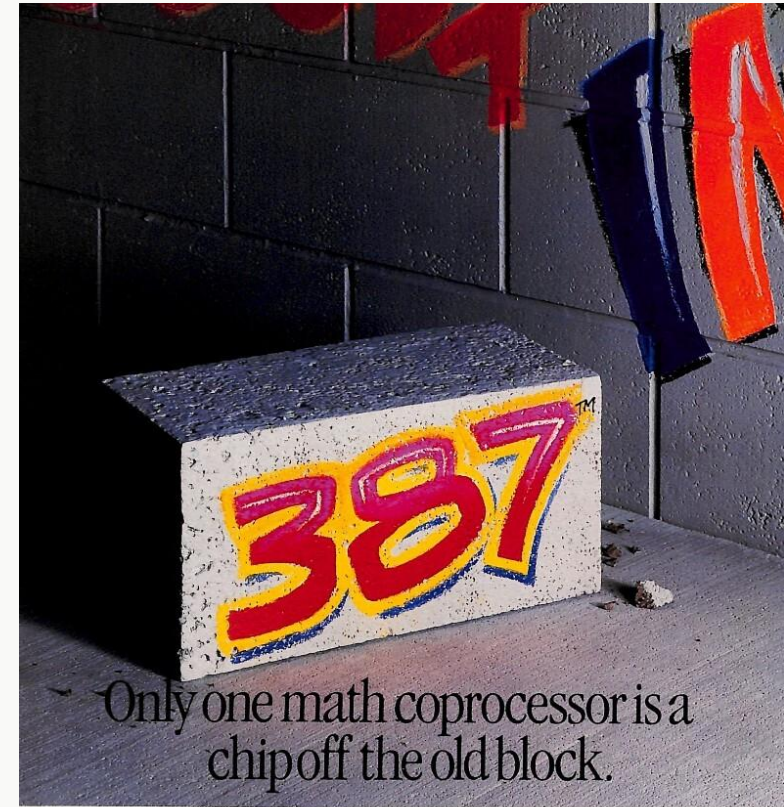
- Storage expansion means adding compute
- Per-node metrics need to be met
 - IOPS
 - Throughput
 - Latency
- **Efficiently**
 - Minimize the power needed to hit per-node targets
- **Compute power consumption is going up**
 - GPUs, TPUs, DPUs, ...

So, how do we fix it?

(or: How I Learned to Stop Worrying about Amdahl's law and Love the Accelerator)

Accelerate?

- On-chip Acceleration
 - CPUs have accelerators for erasure coding, encryption, etc.
 - Efficient execution
- Off-chip Acceleration
 - Dedicated ASICs as expansion cards
 - Remote accelerators can be used



In fact, working side by side with the Intel microprocessor already inside your computer, an Intel Math CoProcessor can increase the speed of your spreadsheet, graphics, CAD and

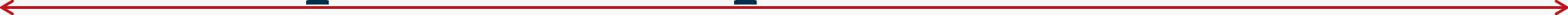
That's Intel's. And our new family of Math CoProcessors is faster—up to 50% for the 287XL.

database programs by as much as 500%. That's good to know.

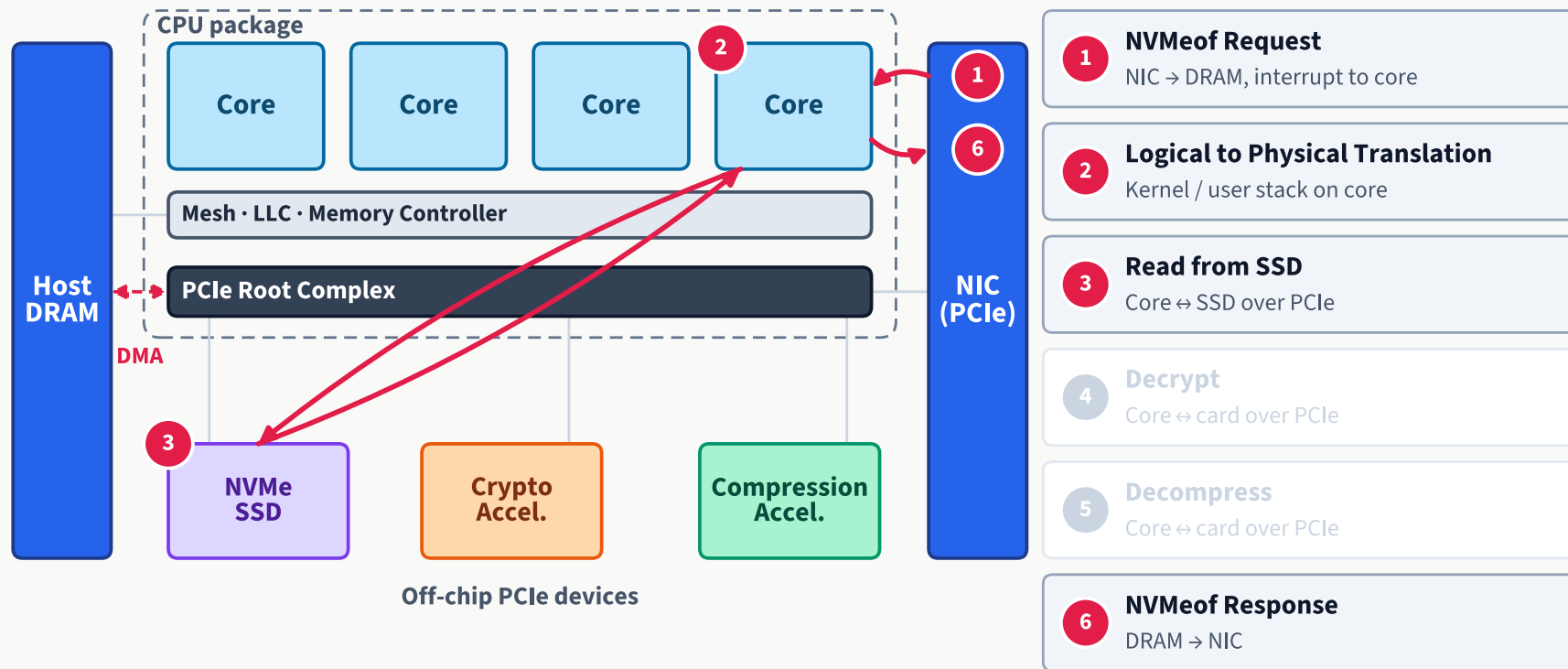
And the fact that it's made by Intel is also good to know.

Because Intel developed the first Math CoProcessor in 1982, and we've shipped millions since then. Each one is manufactured by Intel in the world's most advanced logic

On-Chip or Off-Chip Acceleration?

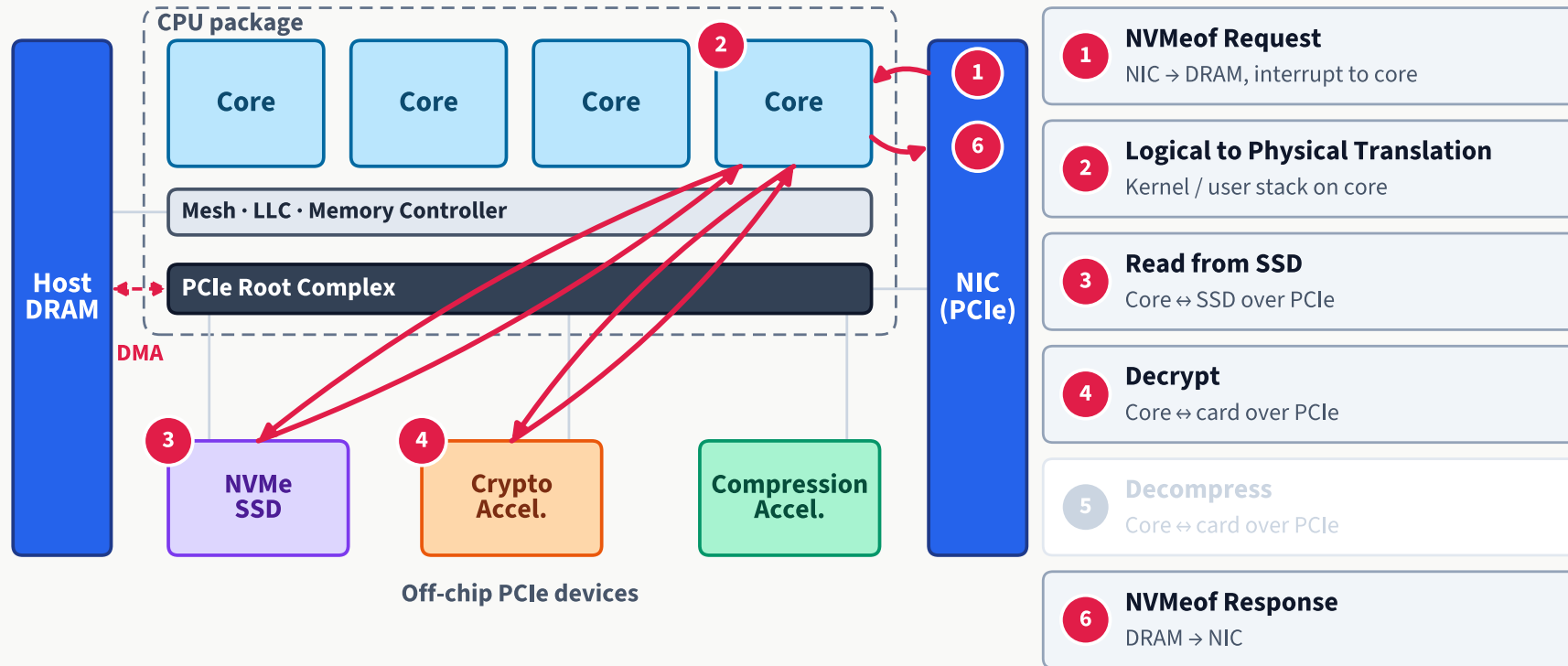


CPU with Accelerators



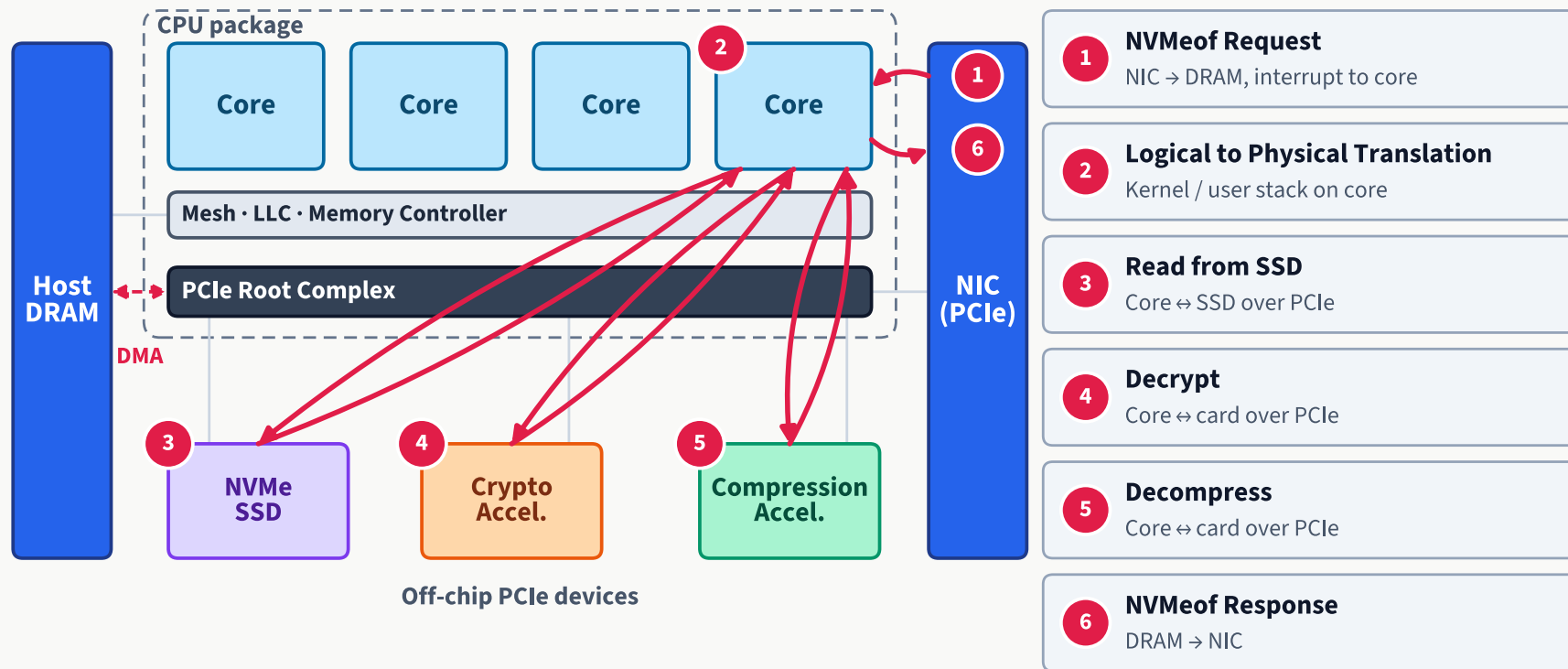
Core in the loop for every step: 3 PCIe hand-offs · 3 DMA passes through DRAM

CPU with Accelerators



Core in the loop for every step: 4 PCIe hand-offs · 5 DMA passes through DRAM

CPU with Accelerators



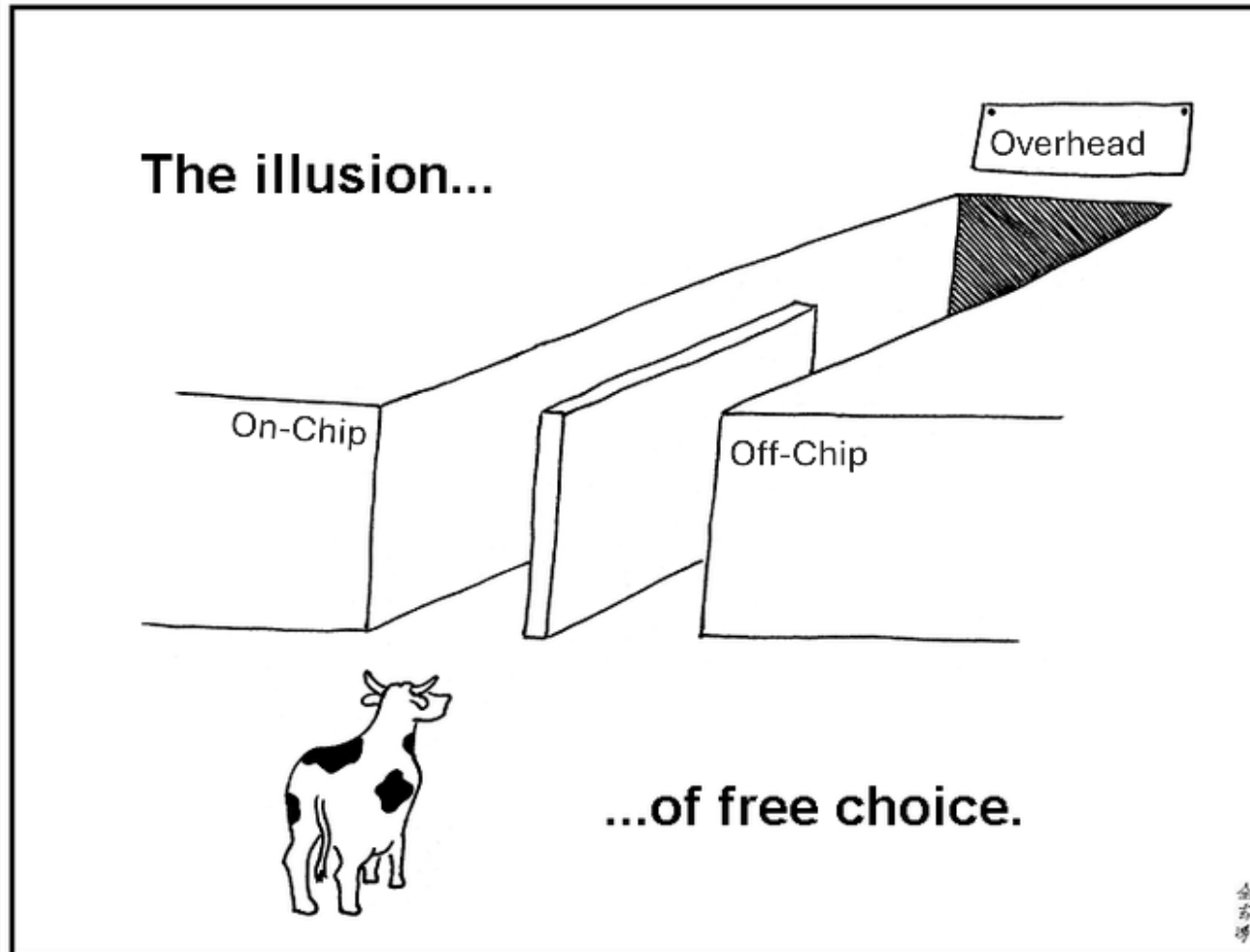
Core in the loop for every step: 5 PCIe hand-offs · 7 DMA passes through DRAM

Now You Have Two Problems!

- Now you need to co-ordinate
 - Offloading logic can be more expensive
- Adding more chips uses more power not less!
- Why is your number-cruncher thinking about infrastructure?



On-Chip or Off-Chip Acceleration?



“Overall, we find that off-chip acceleration is **beneficial when it significantly accelerates a large portion of program** execution to amortize any offload penalties. Furthermore, while full asynchronicity among accelerators is untenable, **accelerator chaining is a practical way** to realize much of the asynchronous performance benefits.”

“**Profiling Hyperscale Big Data Processing**”
Gonzalez et al.

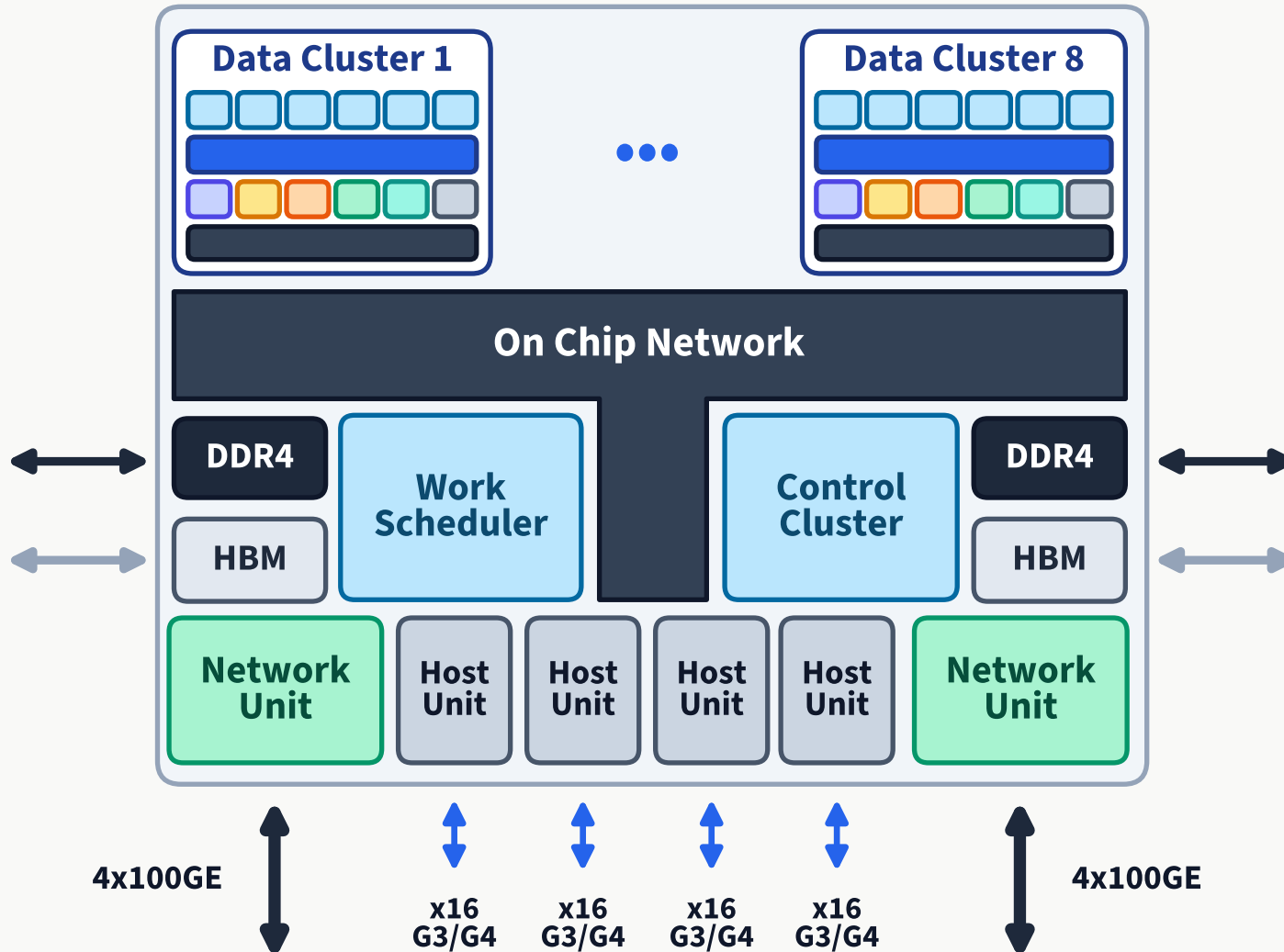
“Finally, our study suggests that **clustering smaller cross-category accelerators** together into a data processing **shared accelerator complex** can provide significant acceleration for hyperscale data processing.”

“Profiling Hyperscale Big Data Processing”
Gonzalez et al.

The Azure Boost DPU

(The Panacea)

The Azure Boost DPU

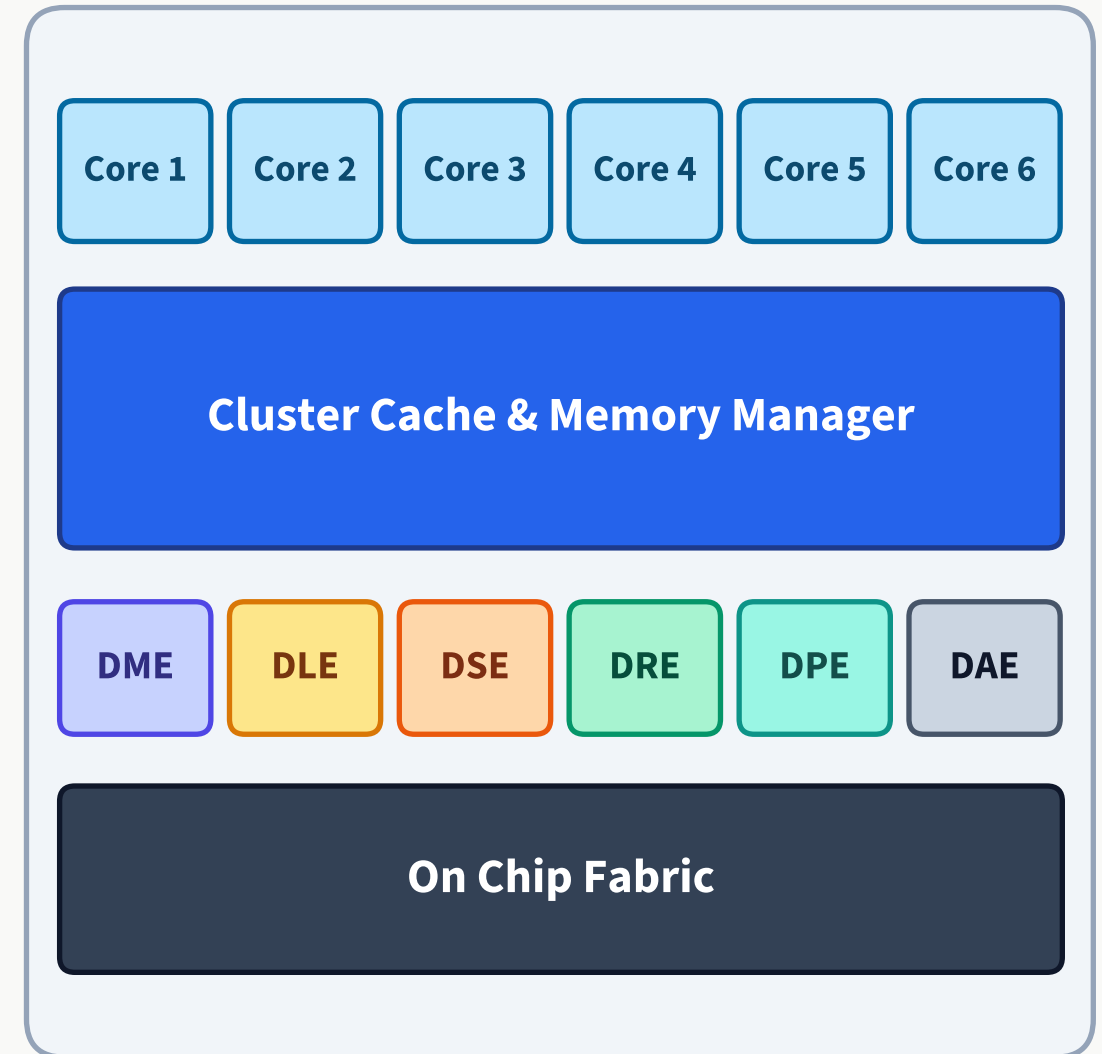


PCIe to NVMe SSDs (root complex) or x86 host (endpoint)

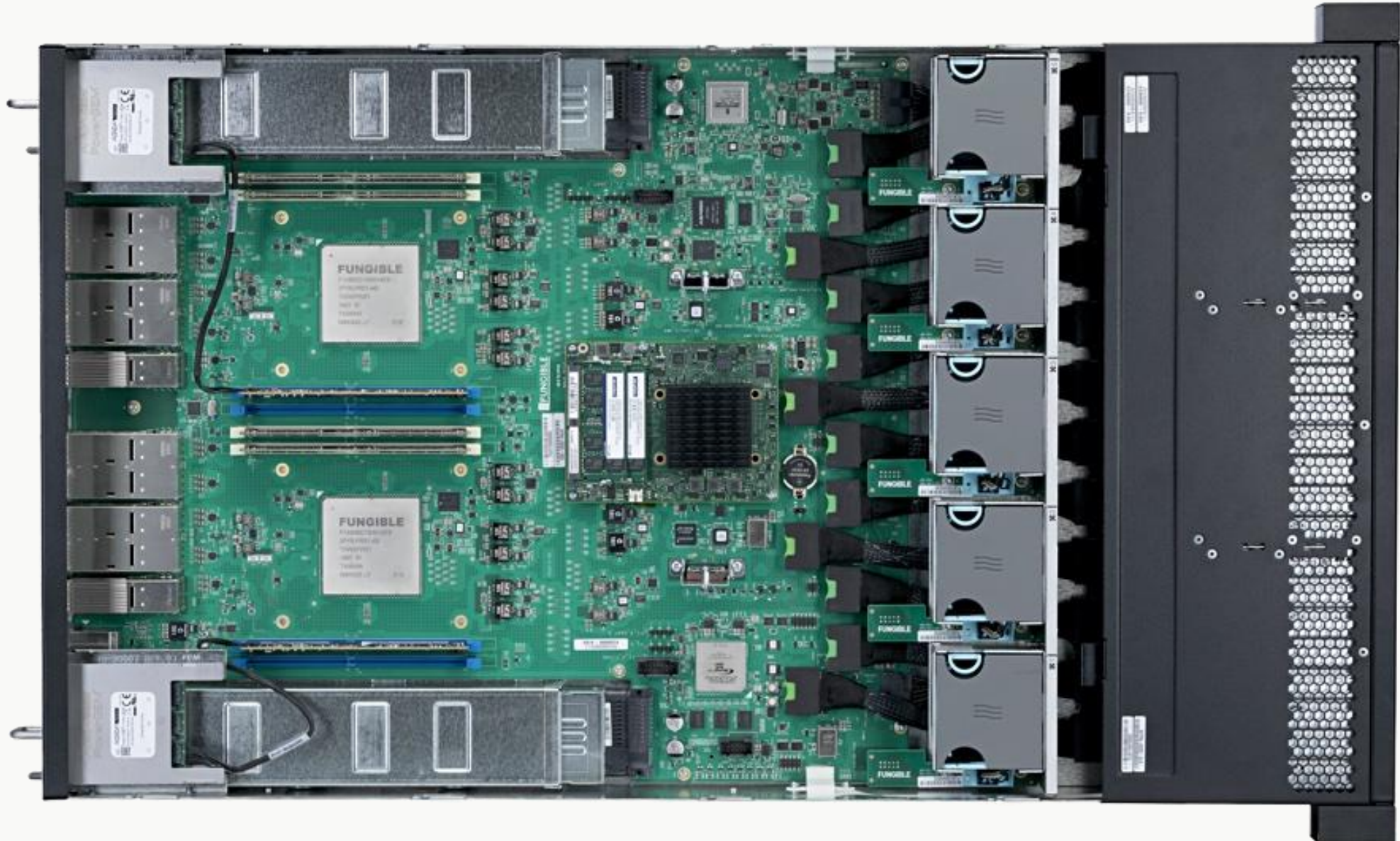
- **8 Data Clusters**
 - 192 processor threads
 - Full cache coherency
 - Tightly integrated accelerators
- **Control Cluster**
 - 8 processor threads
 - Secure complex
 - HSM
 - Root-of-trust
 - Work scheduler
- **Memory & I/O Interfaces**
 - 800 Gbps network unit
 - 4x16 G3/G4 PCIe unit
 - DDR4
 - HBM
- **On Chip Network**
 - High performance
 - Low latency
 - Any unit to any unit

Data Cluster

- 6 Cores × 4 Threads
- Multi-Threaded Accelerators
 - Data movement (DMA)
 - Data lookup
 - Data security (Encryption)
 - Data reduction (Compression)
 - Data protection (Erasure Coding)
 - Data analytics



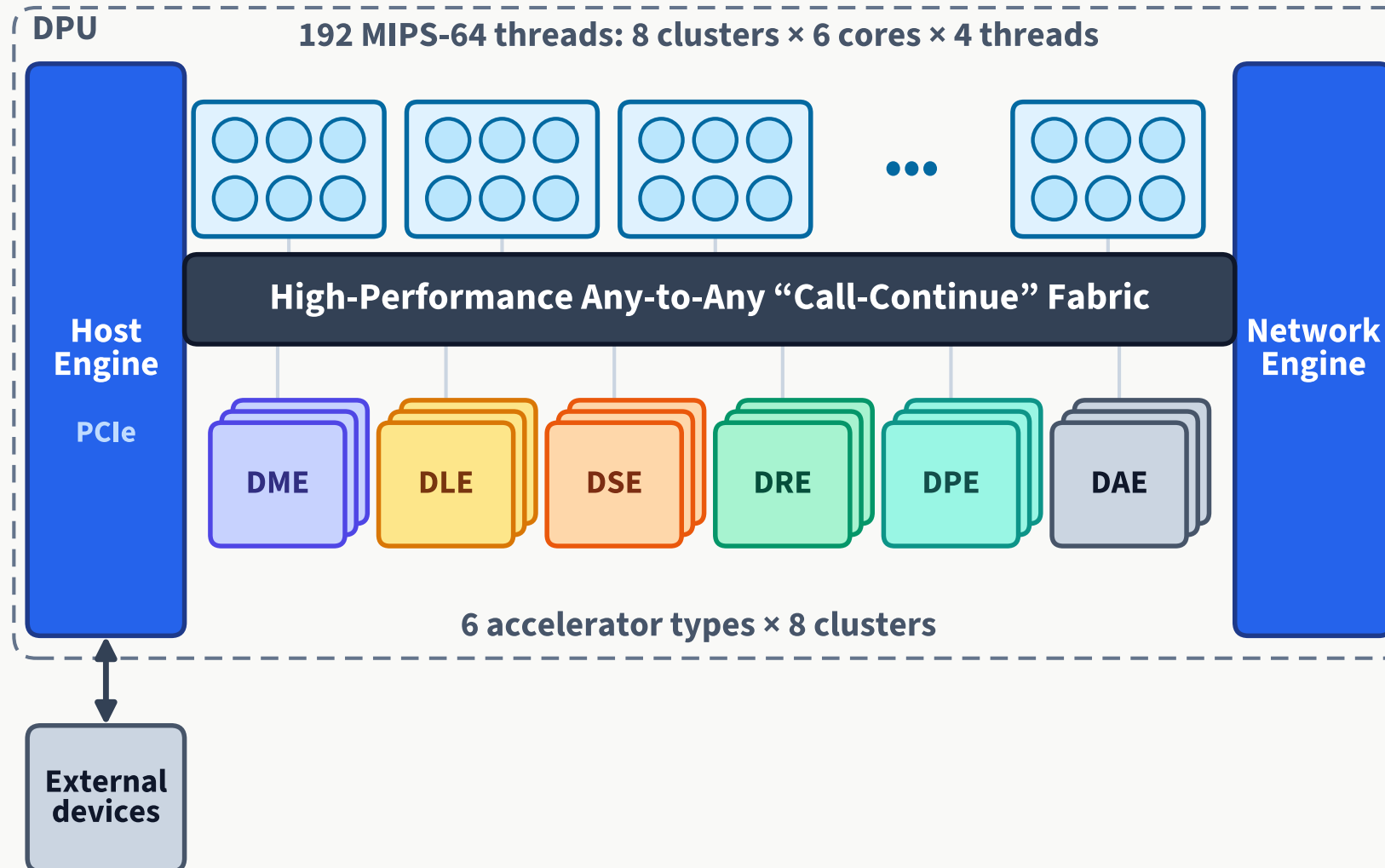
An Absolute Unit.



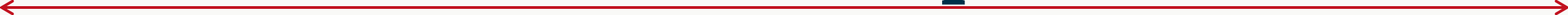
Hardware Won't Save Us

(But Hardware-Software Co-design will)

Data Path Programming Model

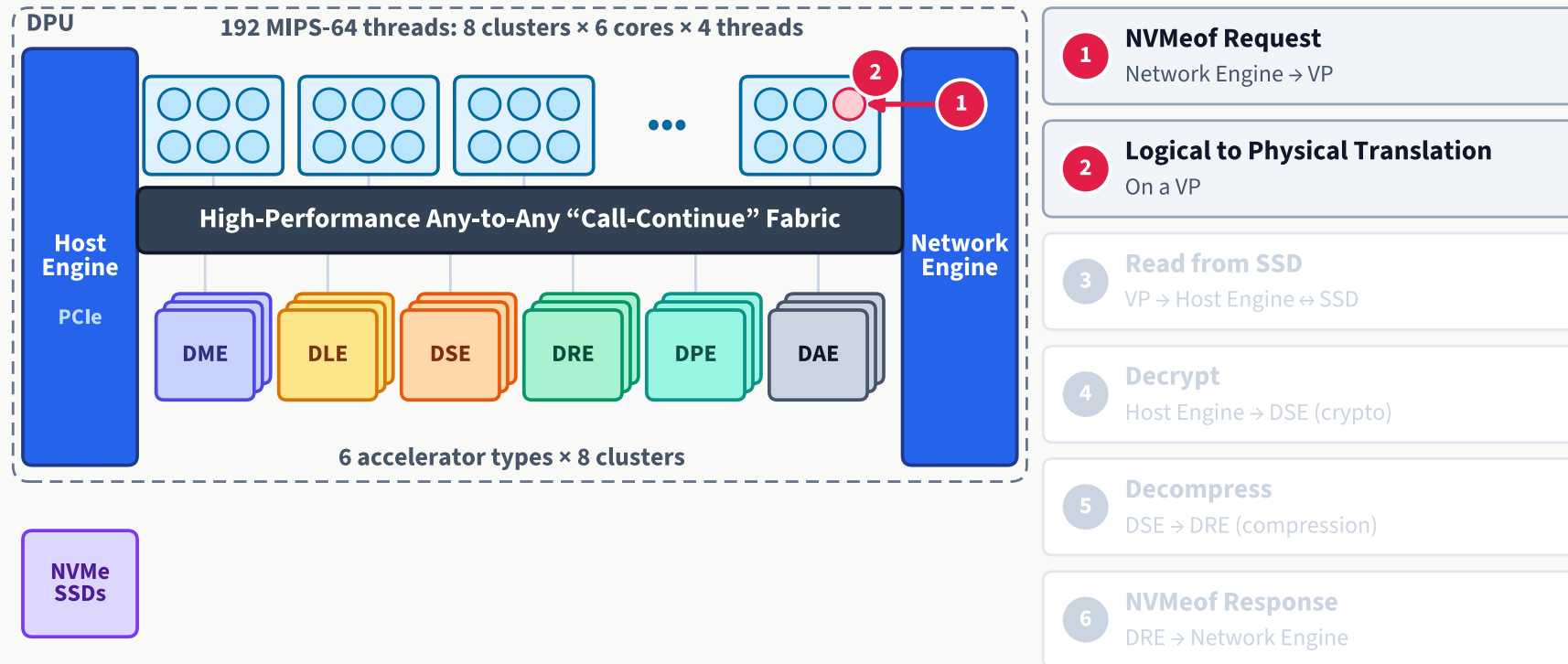


Work-Unit Based Development



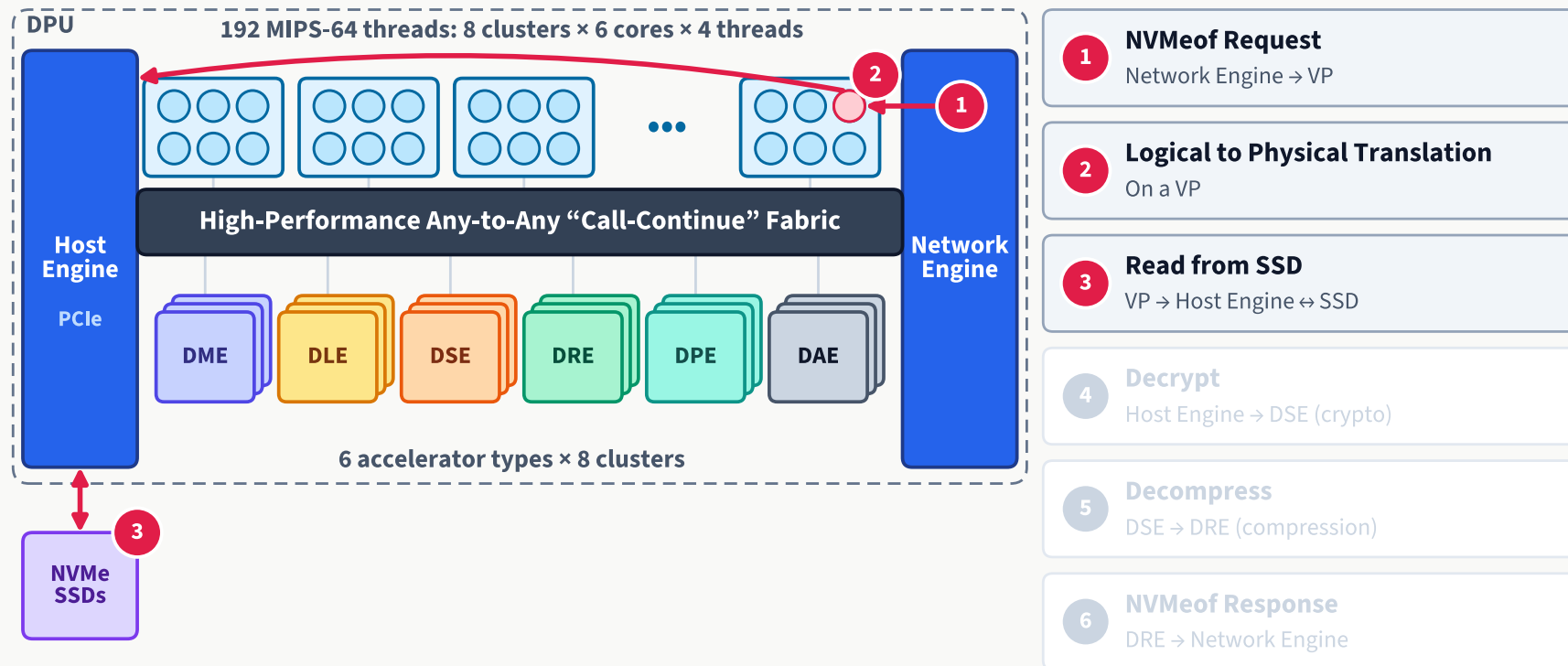
- **Run-to-Completion**
 - Work units are limited to a few thousand instructions
 - No interrupts, no pre-emption, no scheduling
 - Stack-based continuation model
- **Developer needs to break up logic into pipelined stages**
 - Balance between fine and coarse-grained locks
 - Single-level mutual exclusion means no deadlocks
- **Massively-parallelizable**
- **Allows static resource allocation**
 - Processor and accelerator allocation decided ahead-of time

The New IO Pipeline



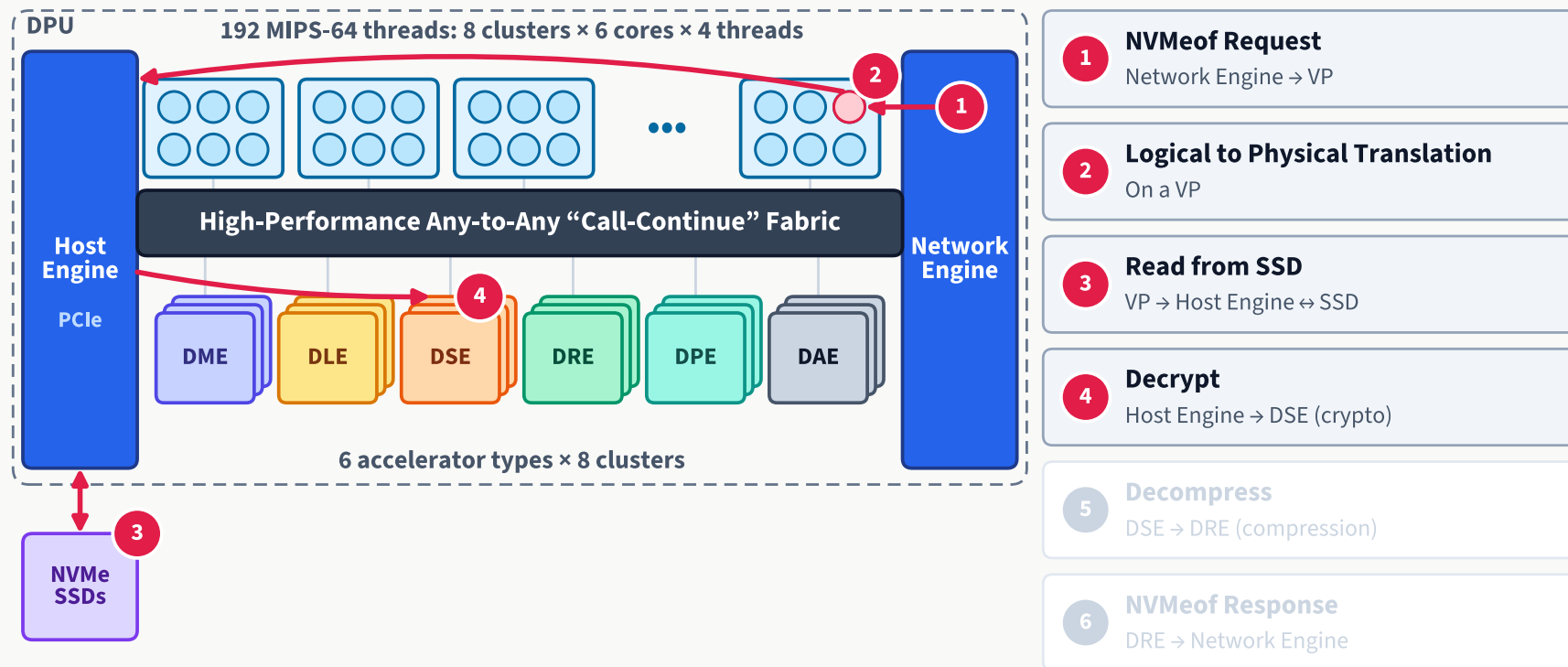
Steps chain without returning to a core: 0 PCIe hand-offs · 1 fabric hand-off

The New IO Pipeline



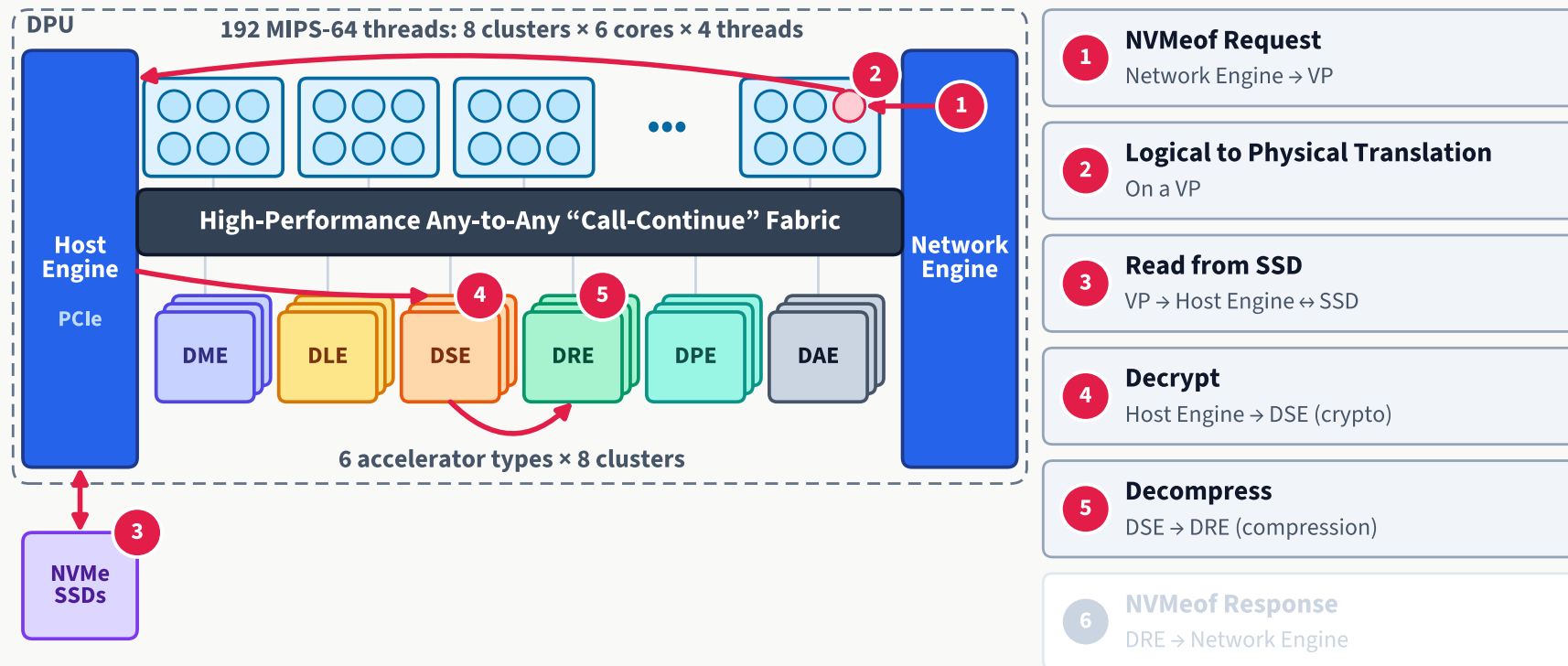
Steps chain without returning to a core: 1 PCIe hand-off · 2 fabric hand-offs

The New IO Pipeline



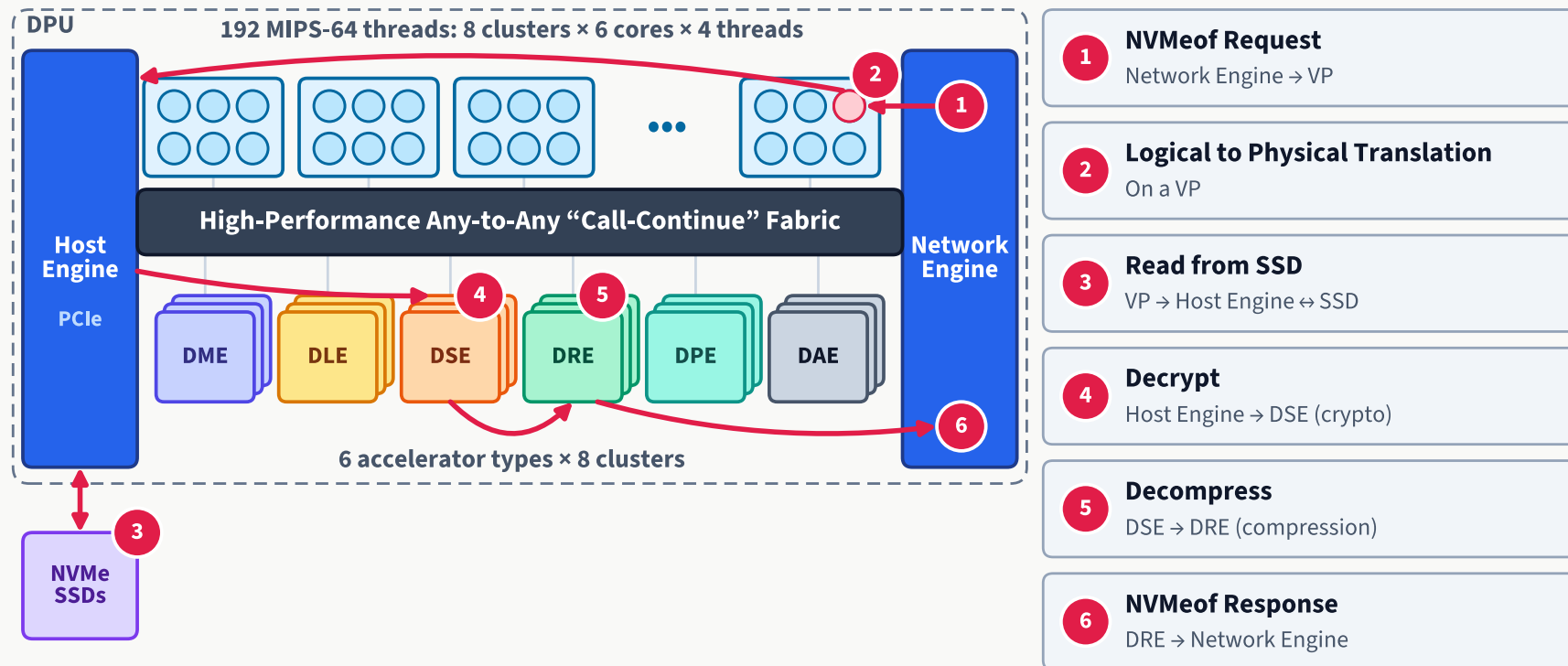
Steps chain without returning to a core: 1 PCIe hand-off · 3 fabric hand-offs

The New IO Pipeline



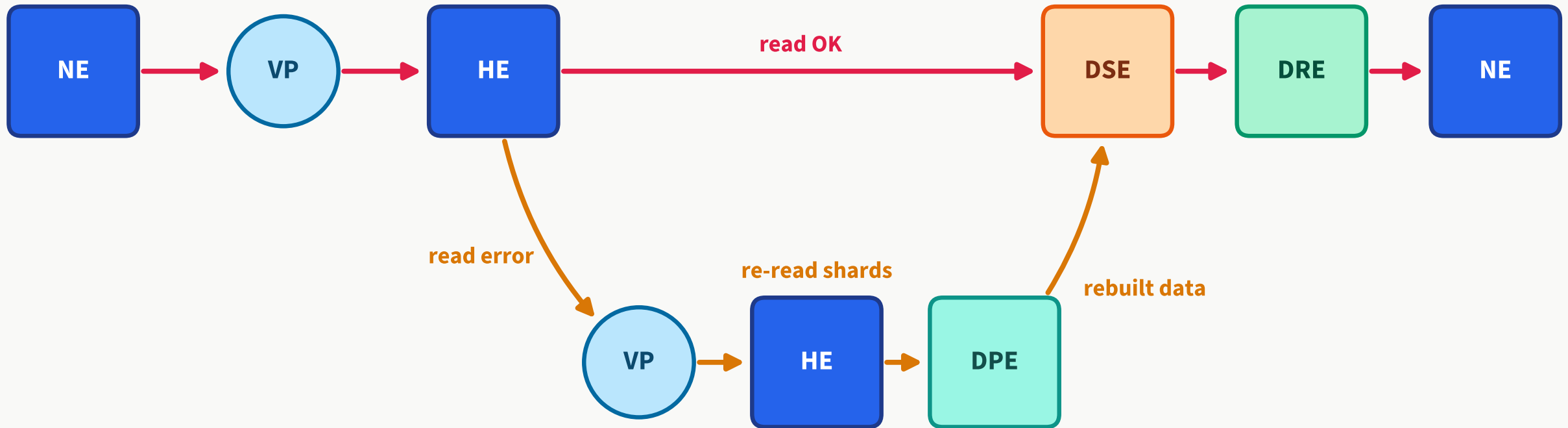
Steps chain without returning to a core: 1 PCIe hand-off · 4 fabric hand-offs

The New IO Pipeline



Steps chain without returning to a core: 1 PCIe hand-off · 5 fabric hand-offs

The Modern Data Pipeline

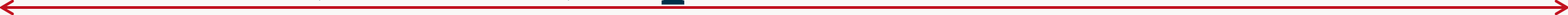


Why do we use a DPU?

- **Storage-centric architecture with a built-in NIC**
 - 1/3rd the power consumption
 - Fewer components → Fewer failure points
- **Performance**
 - 4x Network bandwidth
 - Road to 10M IOPS per node
 - 3-100x improvement in TPC-H
- **Scalability**
 - Independent scaling of storage and compute
 - Higher storage density per rack



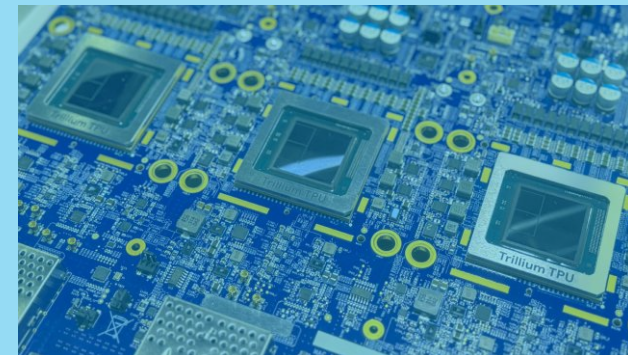
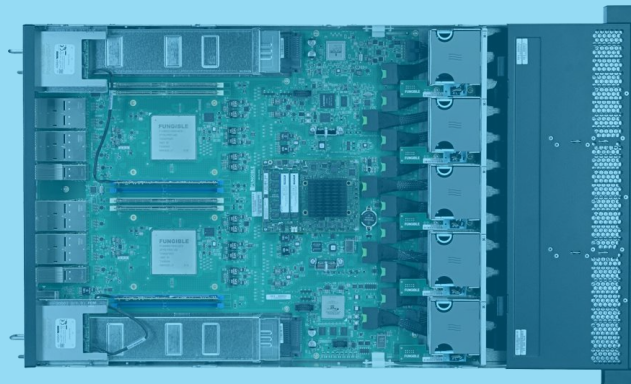
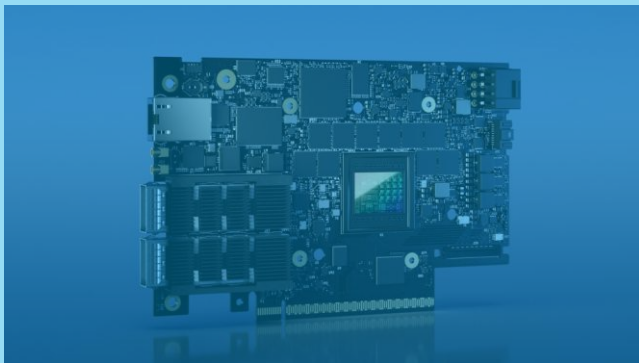
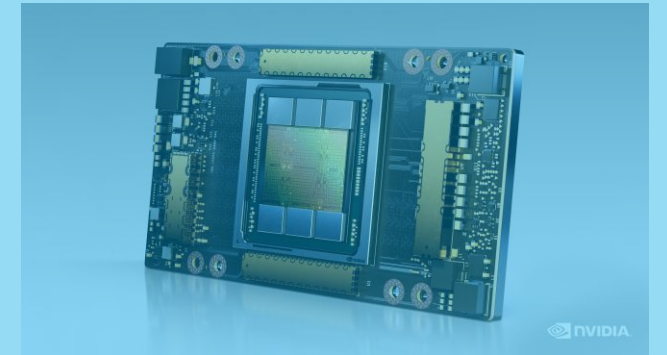
Offload, Avoid, Optimize



- **Offload**
 - Repetitive logic executed by power-intensive units
- **Avoid**
 - Co-ordination
- **Optimize**
 - Everything that you cannot offload or avoid

A Sea of Accelerators

- General Purpose vs Special Purpose
- Fixed Function vs Programmable
- Kernel offload vs dynamic logic
- Host managed vs Independent
- Blades vs Expansion Cards



Thank You!

Contact: dpurandare@microsoft.com

Social: dev@discuss.systems

This work is the joint work of hundreds of people across Fungible and Microsoft.

Special thanks to my Manager,
Nakul Dhotre (ndhotre@microsoft.com)

