Future Computing Architecture

Lessons 7th & 8th

Marco Briscolini, PhD

marco.briscolini@gmail.com

Cell: 3357693820

04/16/2025

Piano del Corso – 16 ore in 8 moduli

Descrizione generale delle architetture HPC e AI e loro componenti di base
- Le previsioni di mercato AI&HPC nel mondo
- Componenti principali: parte computazionale, rete di interconnessione, sottosistema storage
- Concetti di metrica delle varie componenti (misurazione della capacita' computazionale, trasmissione dati, lettura/scrittura dati)
- Metriche riconosciute a livello mondiale (Top500, Green500, IO500)
- Concetti introduttivi sull'analisi della complessita' computazionale di un ambito applicativo

Architetture di calcolo e loro evoluzione
- Architetture omogenee e accelerate
- Concetti generali sui microprocessori (CPU)
- Concetti generali sugli acceleratori (Graphical Processor Unit)
- Integrazione CPU-GPU e trasmissione dati

Reti a alte prestazioni per architetture HPC e AI e loro evoluzione
- Reti con protocollo Infiniband e alcune topologie correlate
- Reti di tipo Ethernet a alte prestazioni
- Protocolli RDMA e RoCE

Sottosistemi storage a alte prestazioni e loro evoluzione
- Concetti generali sulla gerarchia dei sottosistemi storage
- Sistemi a disco magnetico e a stato solido
- Connessione di sistemi storage su SAN, Infiniband, Ethernet, nVME over Fabric, e altro

Architetture storage a alte prestazioni
- Architetture di sottositemi storage
- Filesystem paralleli per lettura/scrittura a alte prestazioni

06

Problematiche di efficientamento energetico per sistemi HPC a grande scala (architetture pre e exascale)
- Il concetto di PUE e di efficienza energetica a parita' di potenza computazionale
- Come le varie architetture si caratterizzano in termini di "Potenza di Calcolo"/Watt
- Utilizzo di tecniche di gestione del carico di lavoro per ottimizzare l'efficienza energetica
- Soluzioni di raffreddamento a aria, a acqua diretta e immersivo
- Concetti generali sul disegno e la realizzazione di Data Center efficienti

07

Accenni sulle architetture innovative in ambito AI&HPC
- Architetture AI scalabili
- Interconnessione tra sistemi AI
- AI/HPC/Q-C architettura integrata per carichi computazionali complessi

08

Accenni al disegno e alla progettazione di un'architettura HPC
- Definizione di specifiche di progetto
- Valutazione preliminare dell'architettura ottimale
- Disegno di massima dell'architettura

- Concetto di rispondenza e verifica alle specifiche di progetto

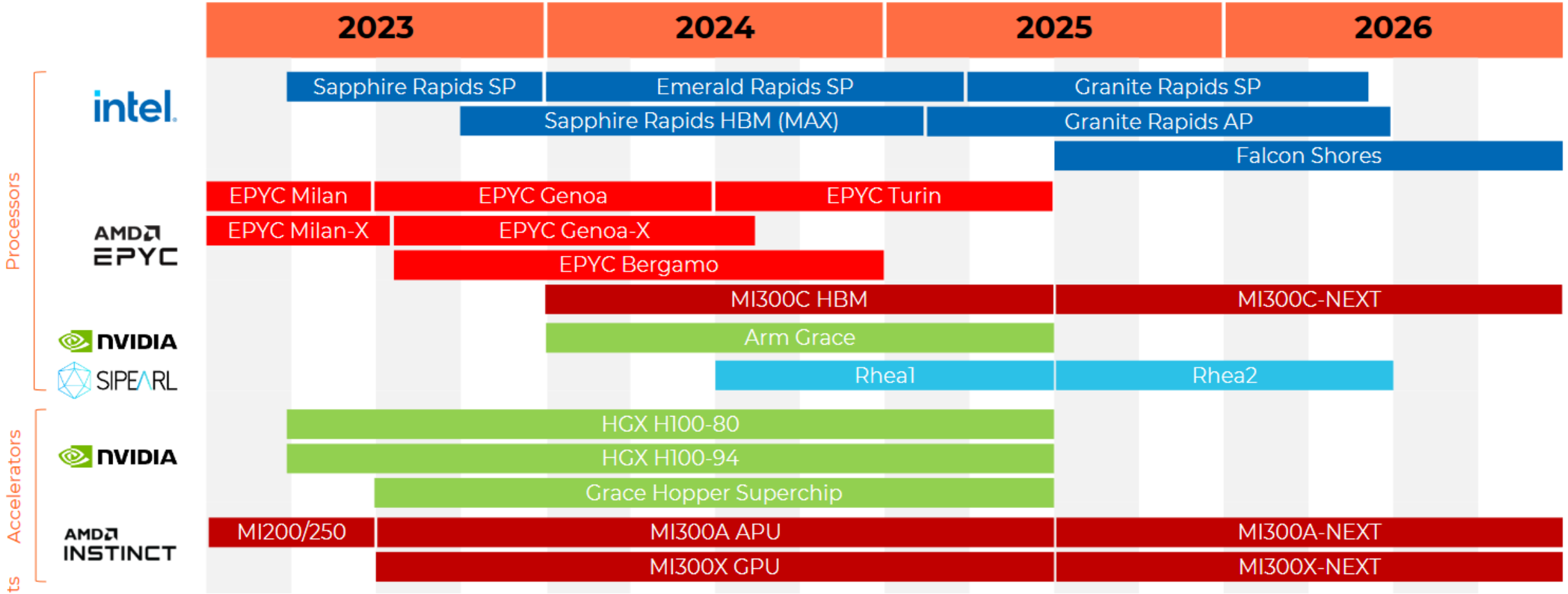
Piano del Corso – Lesson 7th

Accenni sulle architetture innovative in ambito AI&HPC

- Architetture AI scalabili
- Interconnessione tra sistemi AI
- Architetture disaggregate
- AI/HPC/Q-C architettura integrata per carichi computazionali complessi

Technology roadmap

As known in May 2023 (subject to change)



CPU
Aumento del numero di cores per CPU: 64 → 128 → 256
Inserimento memoria HBM (High Bandwidth Memory)
Consumi in crescita -→ 500W/CPU

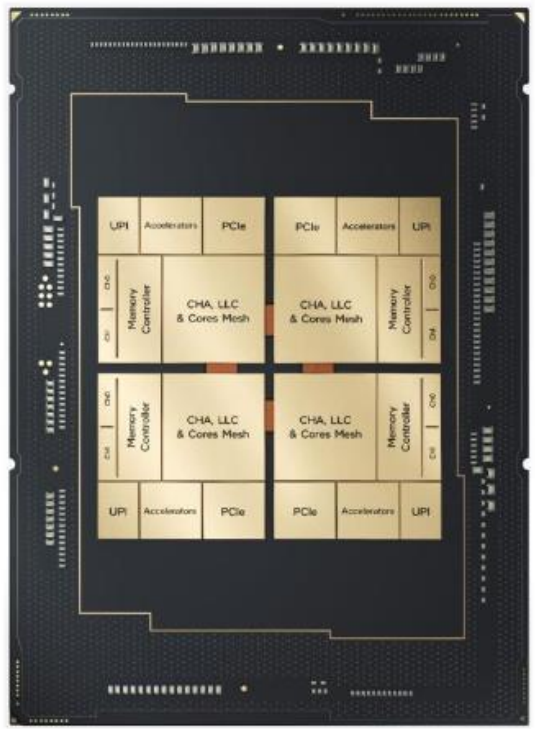
GPU
CPU-GPU in a single Socket
HBM (High Bandwidth Memory) ~ 3TBs
Consumi in crescita -→ 1000W

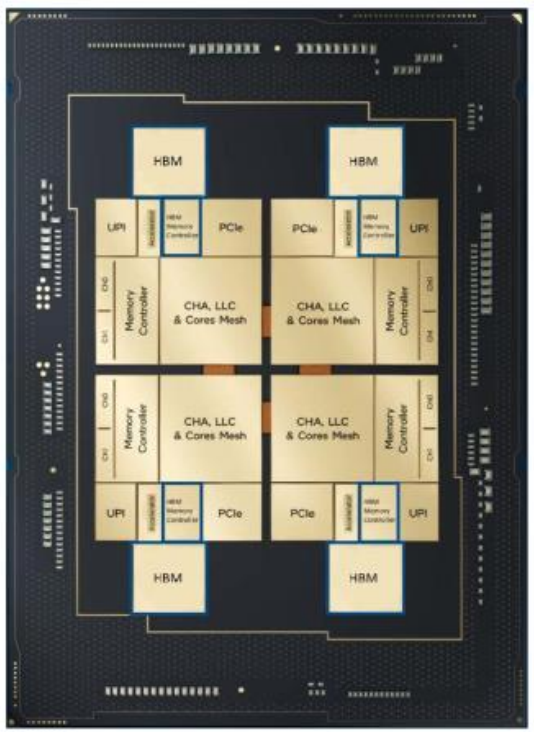
Network
Aumento BW per link ~ 400Gbs
RoCE in High Fast Ethernet
Interconnessione delle GPUs nel nodo ~ 1TBs

Intel Sapphire Rapids

4th Gen Intel® Xeon® Scalable Processors

- Up to 56x Cores
- 64GB HBM2e (MAX Series)
- 8x Channels of DDR5 4800MT/s
- Up to 4x UPI link at 16GT/s
- Increased I/O Bandwidth with PCIe Gen 5.0
- Intel® Speed select technology
- Intel® Advance Matrix Extension (AMX) and AVX-512
- Foundational support for CXL 1.1
- TDP up to 350W

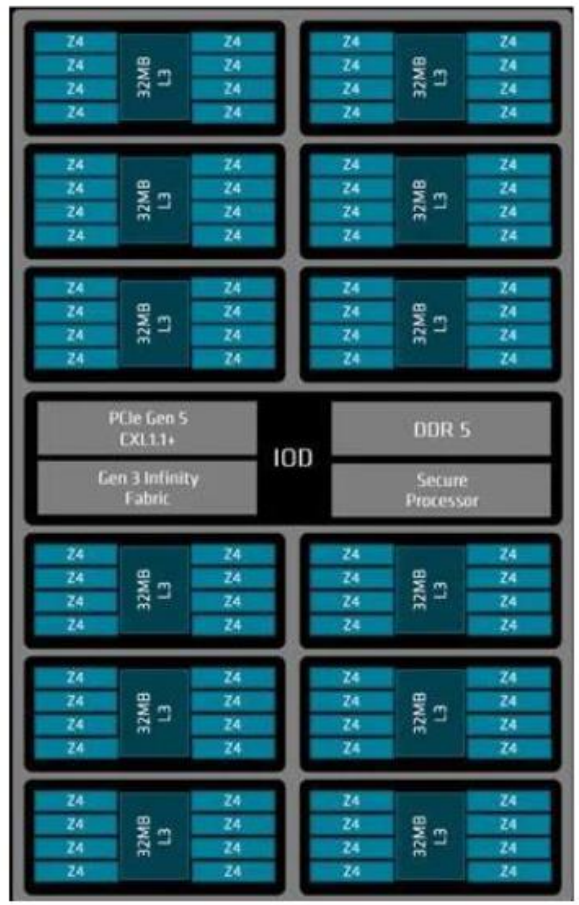




AMD EPYC Genoa & Bergamo

4th Gen AMD EPYC™ Processors

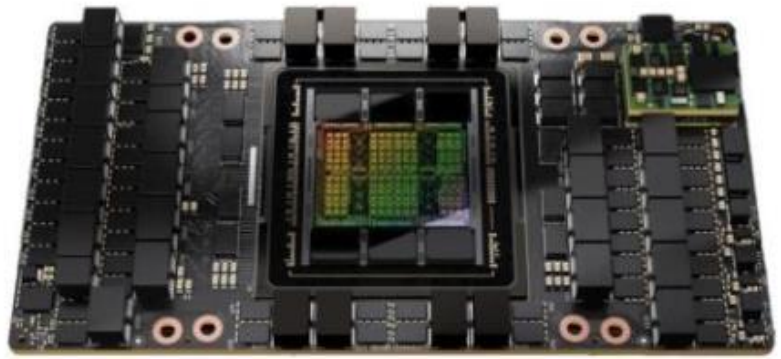
- Up to 96x – 128x Zen4 cores
- Maximum boost clock up to 4.4GHz
- 5nm technology
- Up to 12TB of memory per socket
- 12x memory channels of DDR5 4800MT/s
- Up to 160x PCIe 5.0 lanes
- Foundational support for CXL 1.1+
- AVX-512 support
- TDP up to 400W

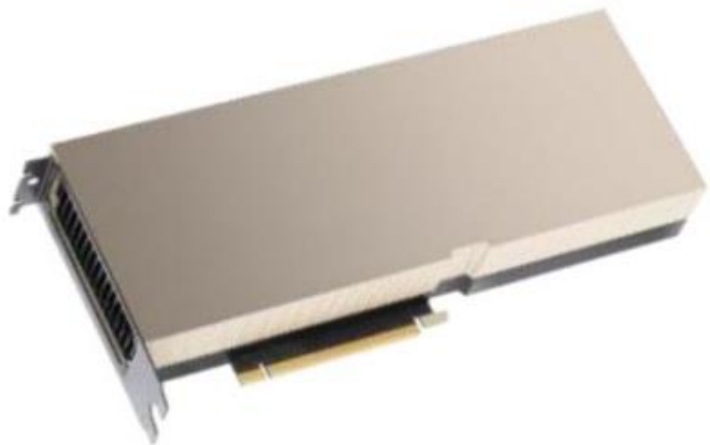


NVIDIA Hopper H100

Specifications

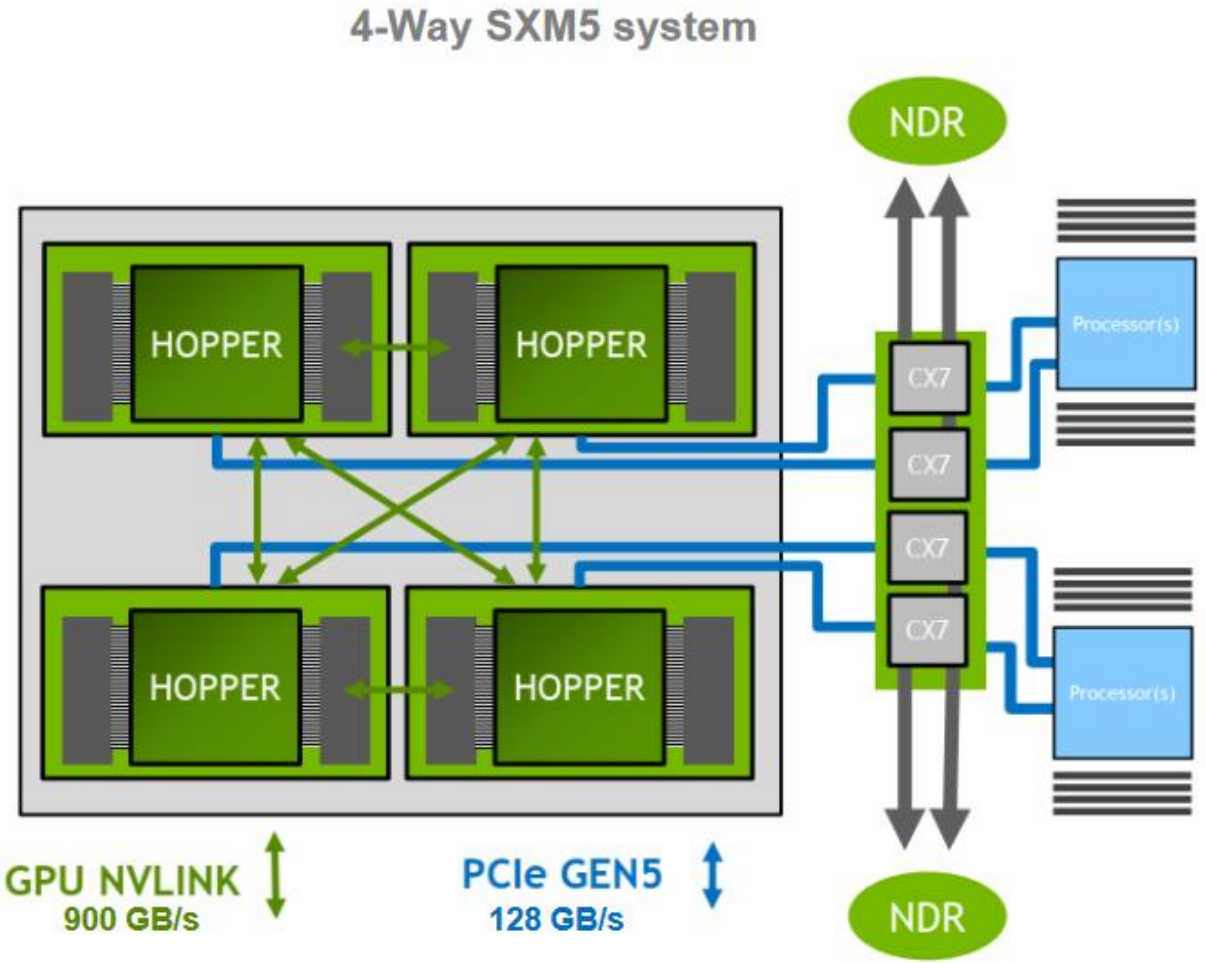
Technical Specifications		
	H100 SXM	**H100 PCIe**
FP64	34 teraFLOPS	26 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	51 teraFLOPS
FP32	67 teraFLOPS	51 teraFLOPS
TF32 Tensor Core	989 teraFLOPS[2]	756 teraFLOPS[2]
BFLOAT16 Tensor Core	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]
FP16 Tensor Core	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]
FP8 Tensor Core	3,958 teraFLOPS[2]	3,026 teraFLOPS[2]
INT8 Tensor Core	3,958 TOPS[2]	3,026 TOPS[2]
GPU memory	80GB	80GB
GPU memory bandwidth	3.35TB/s	2TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Max thermal design power (TDP)	Up to 700W (configurable)	300-350W (configurable)
Multi-instance GPUs	Up to 7 MIGs @ 10GB each	Up to 7 MIGs @ 10GB each
Form factor	SXM	PCIe > dual-slot > air-cooled
Interconnect	NVLink: > 900GB/s PCIe > Gen5: 128GB/s	NVLink: > 600GB/s PCIe > Gen5: 128GB/s





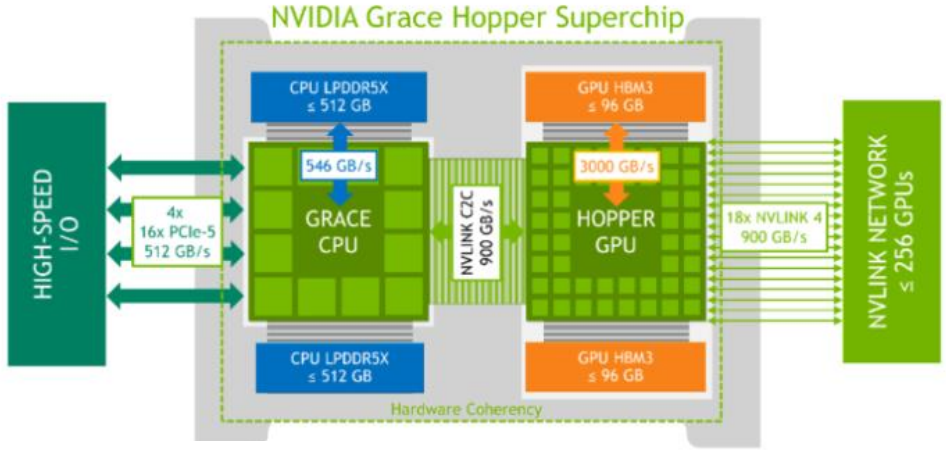
NVIDIA Hopper H100

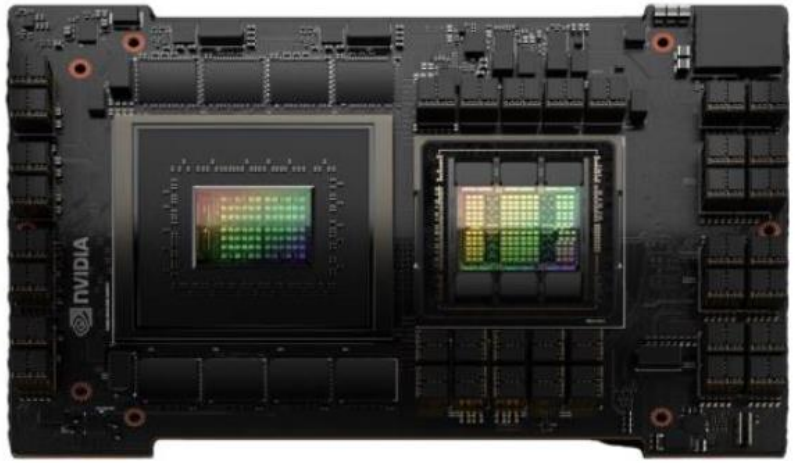
Node design

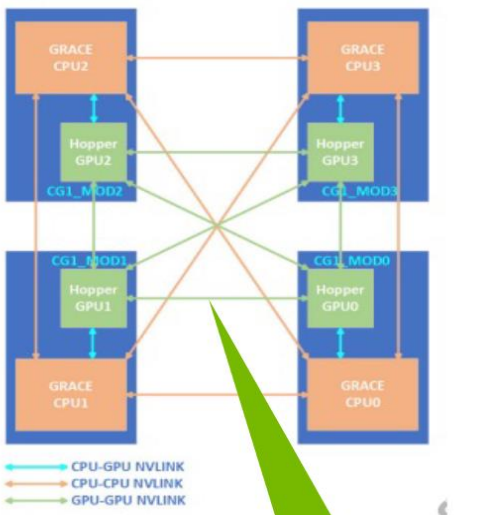


NVIDIA Grace Hopper Superchip

First NVIDIA APU



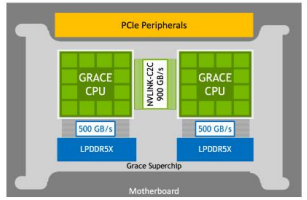




1/3 of the NVLINK bandwidth

NVIDIA Grace CPU Superchip

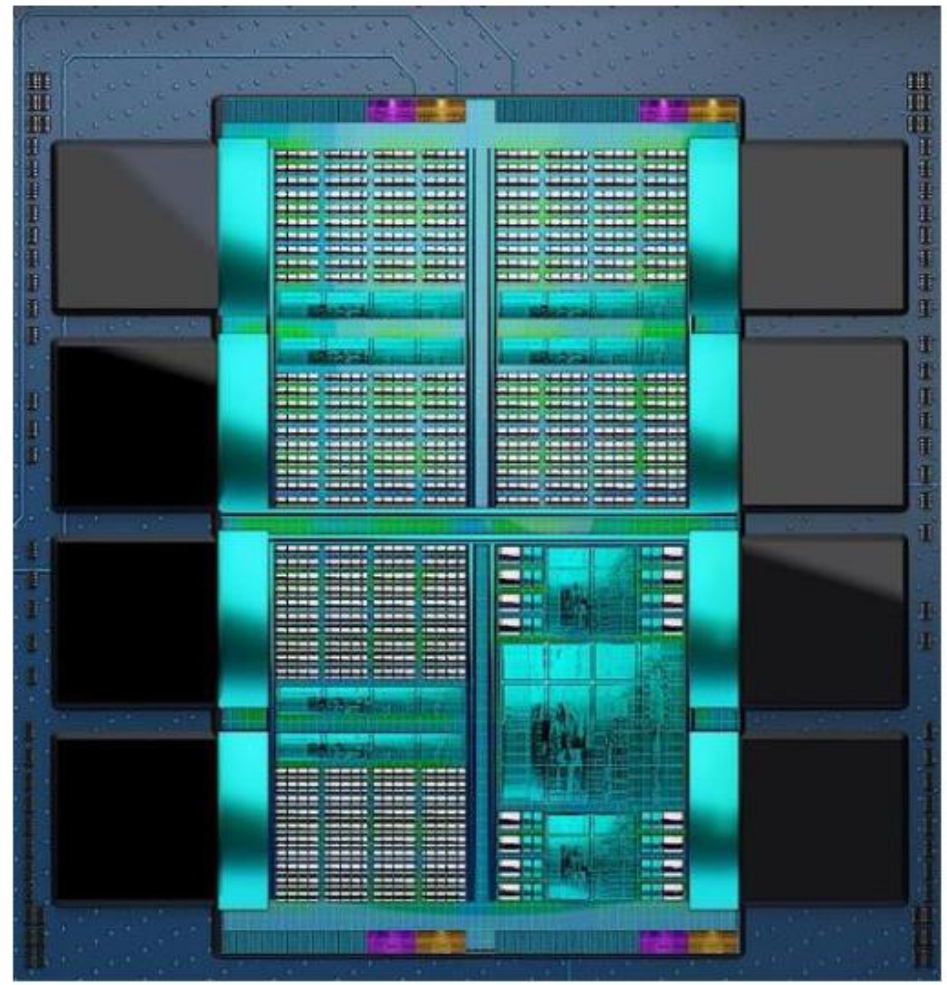
Grace CPU	Feature
CPU core count	72 Arm Neoverse V2 cores
L1 cache	64KB i-cache + 64KB d-cache
L2 cache	1MB per core
L3 cache	117MB
LPDDR5X size	Up to 480GB
Memory bandwidth	Up to 512GB/s
PCIe links	Up to 4x PCIe x16 (Gen5)



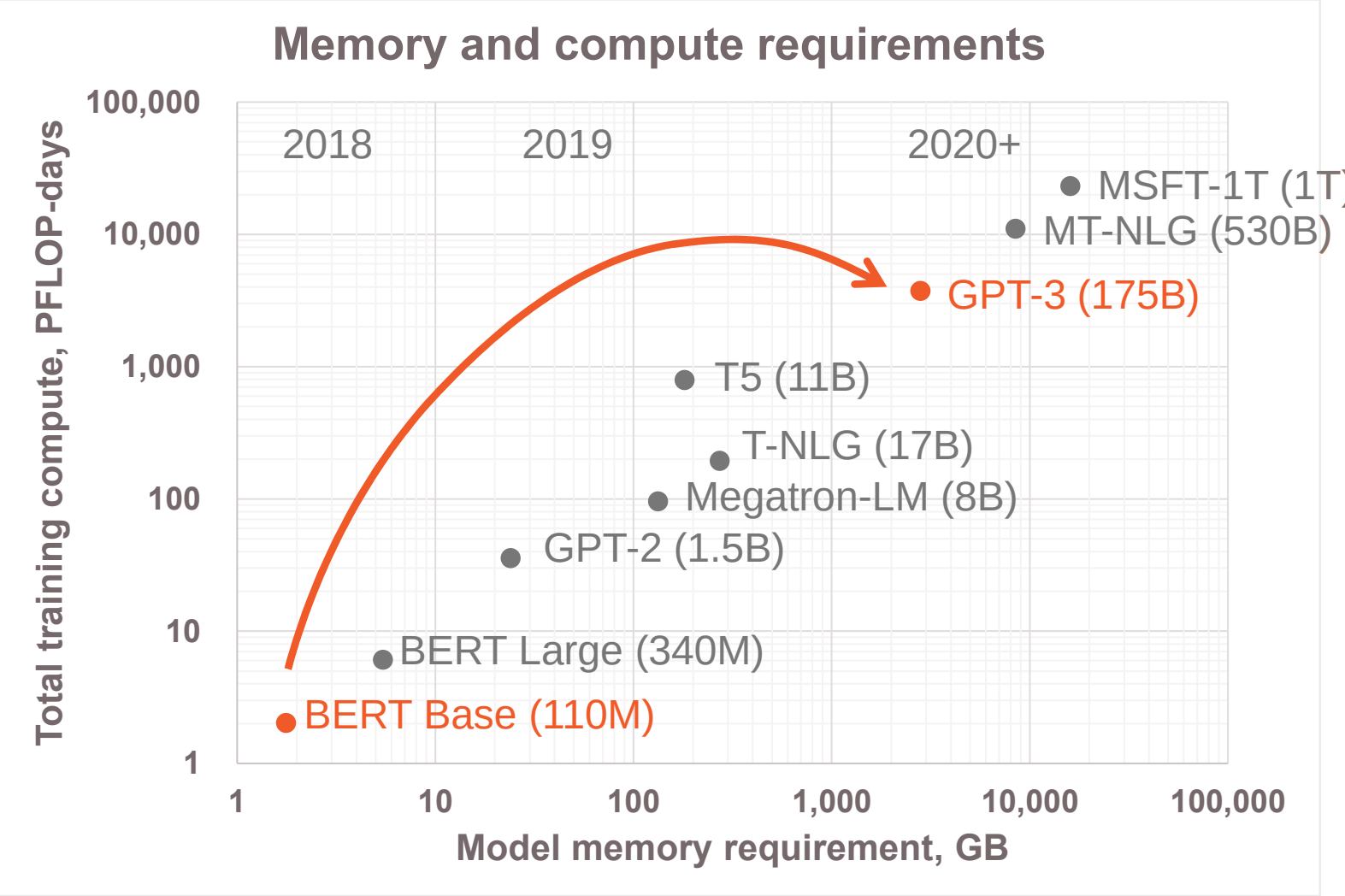
AMD MI300

AMD First APU

- CPU + GPU socket
- 24 EPYC Zen4 cores
- XX GPU Compute Units
- 128GB HBM3
- TDP 550W – 850W



Exponential Growth of Neural Networks

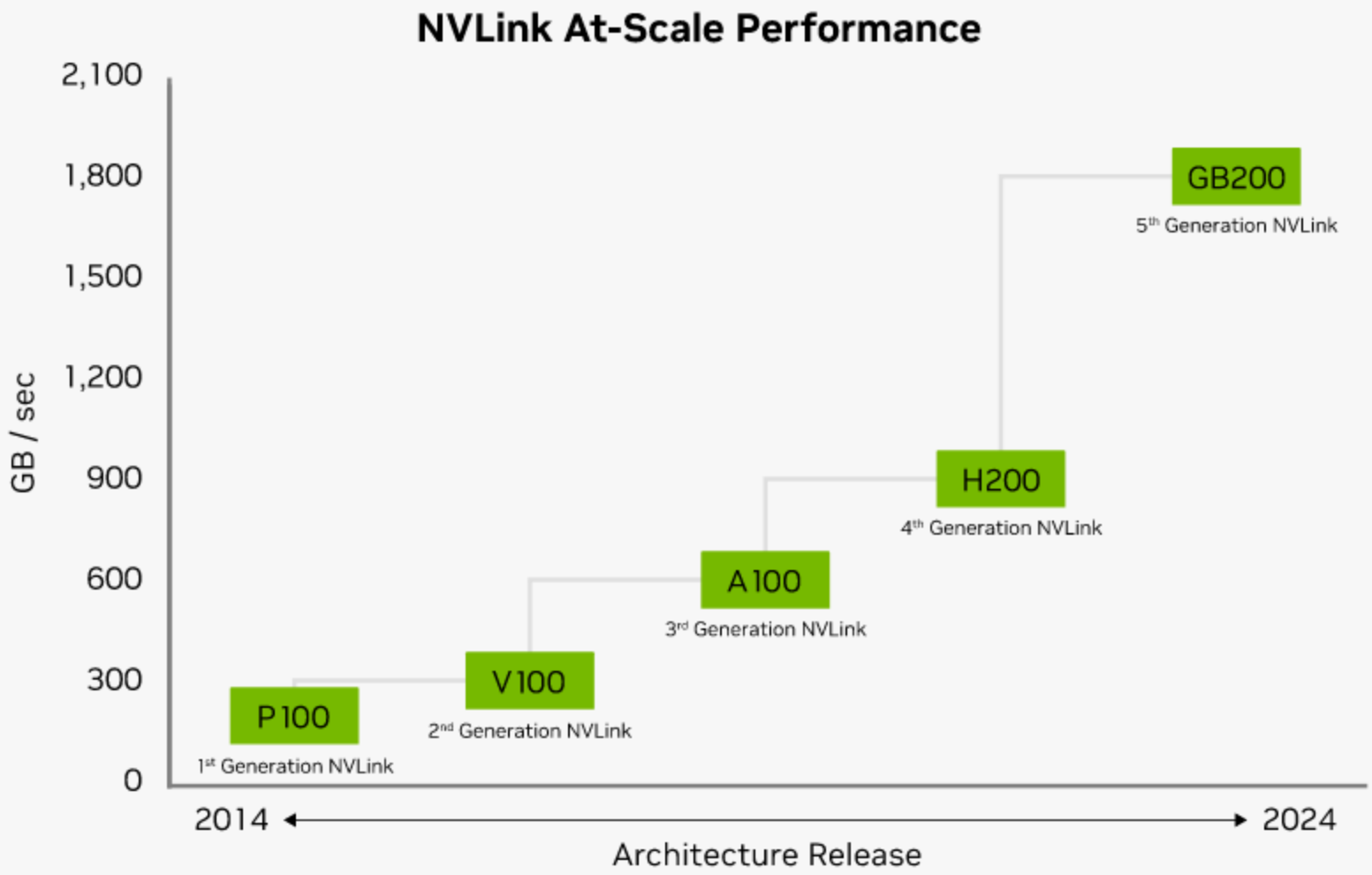


Over 1000x increase

In just **2 years**

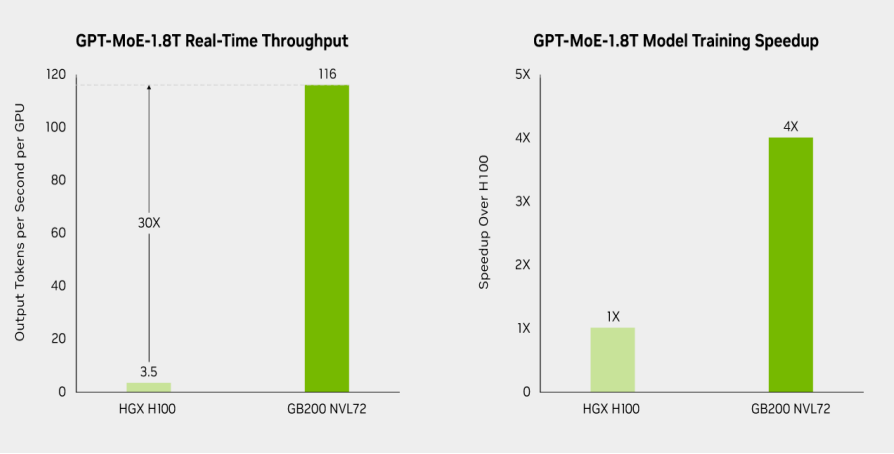
Tomorrow, **multi-trillion** parameter models

NVLink Performance

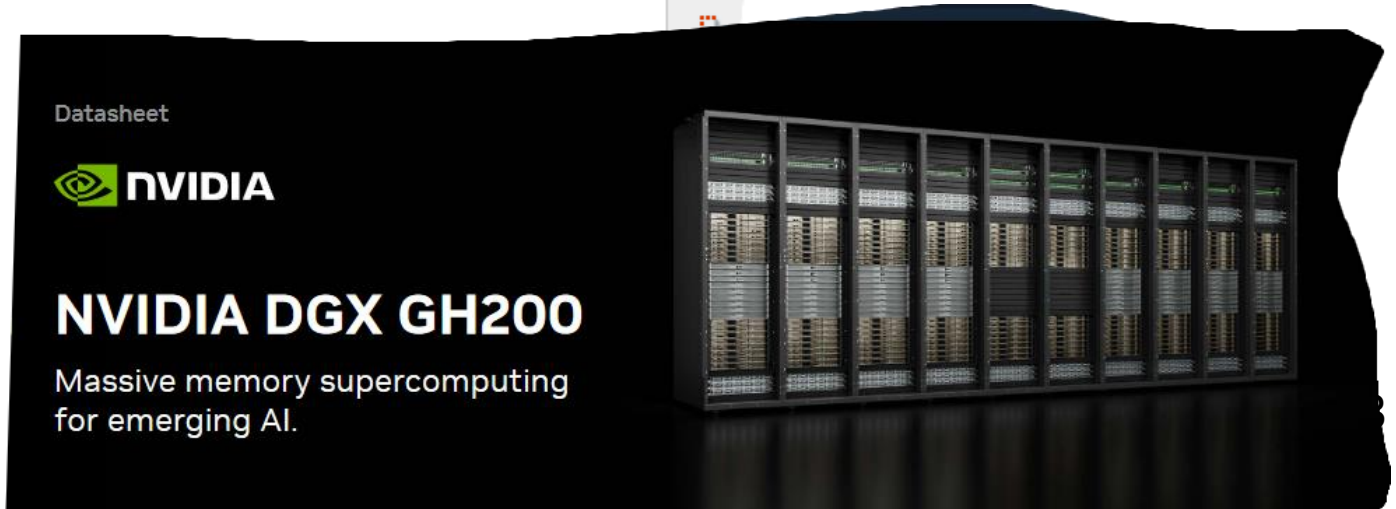


https://www.nvidia.com/en-gb/data-center/nvlink/

DGX GH200 Technical Specifications	
CPU and GPU	32x NVIDIA Grace Hopper Superchips
CPU Cores	2,304 Arm® Neoverse V2 Cores with SVE2 4X 128b
Shared Memory	19.5 TB
Performance	128 petaFLOPS of FP8 AI performance
Networking	32x OSFP single-port NVIDIA ConnectX-7 VPI with 400Gb/s InfiniBand 16x dual-port NVIDIA BlueField®-3 VPI with 200Gb/s InfiniBand and Ethernet
NVIDIA NVLink Switch System	9x L1 NVIDIA NVLink Switches
Management Network	Host baseboard management controller (BMC) with RJ45
Software	NVIDIA AI Enterprise (optimized AI software) NVIDIA Base Command (orchestration, scheduling, and cluster management) DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky (operating system)
Support	Three-year business-standard hardware and software support



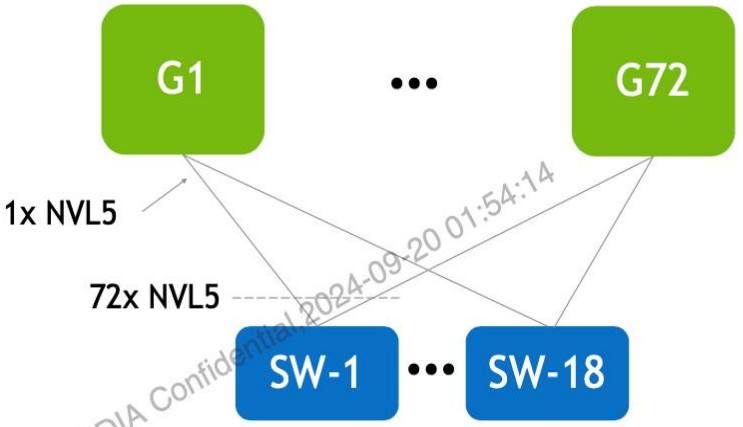
https://www.nvidia.com/en-gb/data-center/gb200-nvl72/



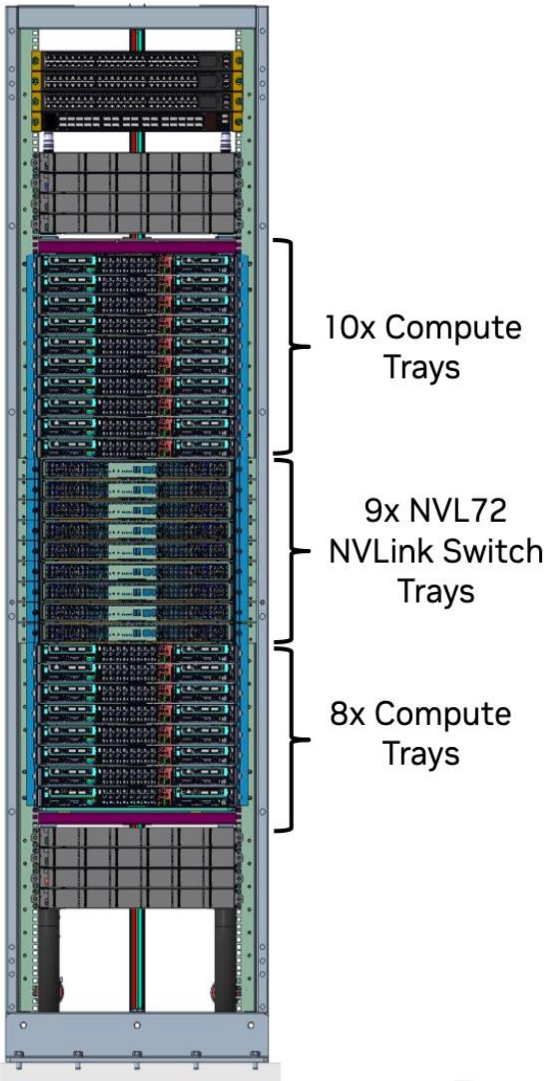
NVLink Network

72 GPU Single Rack NVL Domain

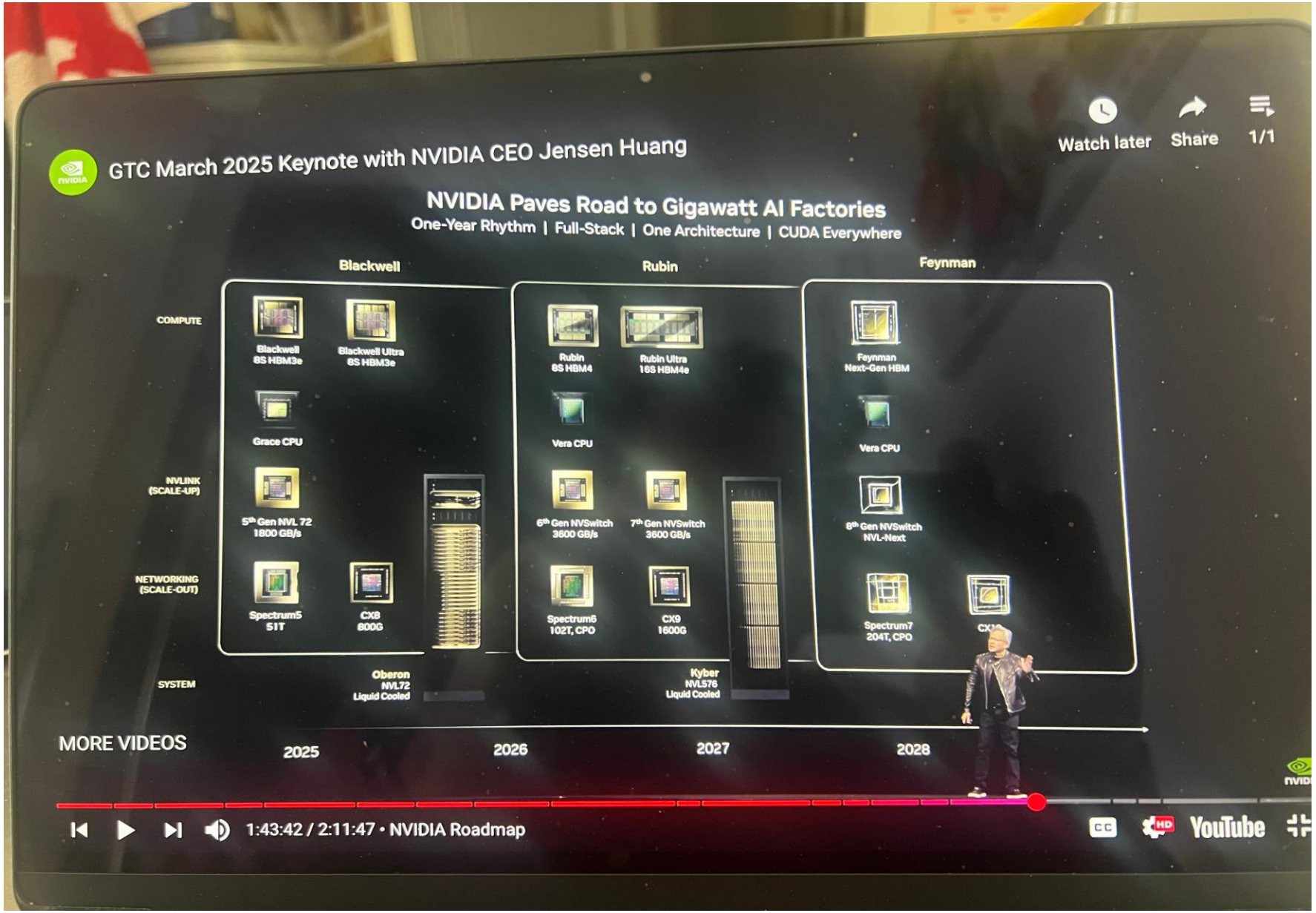
- 9x NVLink Switch trays per rack
- NVL72 NVLink Switch tray
  - NVLink connectivity to all GPUs
  - Each switch tray has two NVSwitch ASICs with all the ports facing the cable backplane
- Single 72-GPU L1 domain
- L1 Domain
  - 72x Blackwell GPUs with 18 ports of NVL5
  - 18x NVL5 Switches



72 GPU



NVIDIA road-map



GPU TDP specification

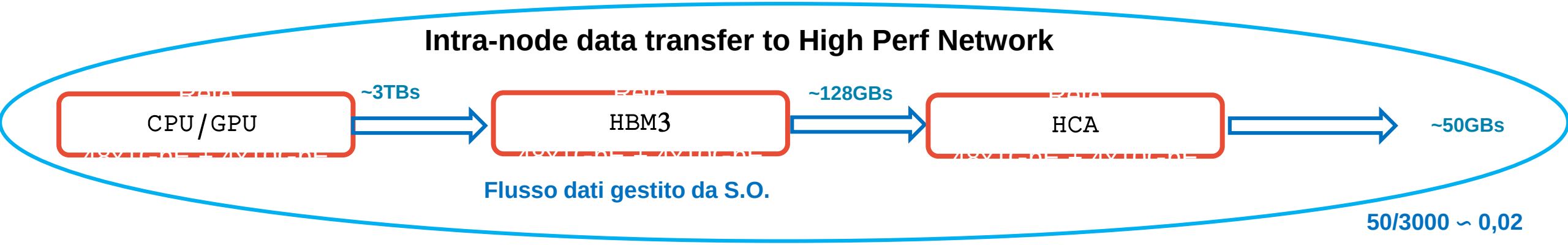
Blackwell: 1200W max power

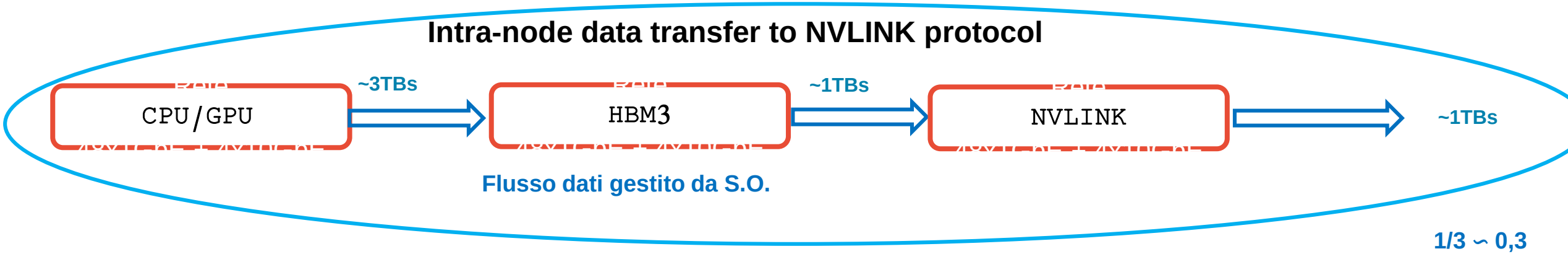
Rubin: 1500W max power est

https://www.nvidia.com/gtc/keynote/

https://www.theregister.com/2024/03/18/nvidia_turns_up_the_ai/

Il flusso dati: dal nodo al sottosistema storage e viceversa





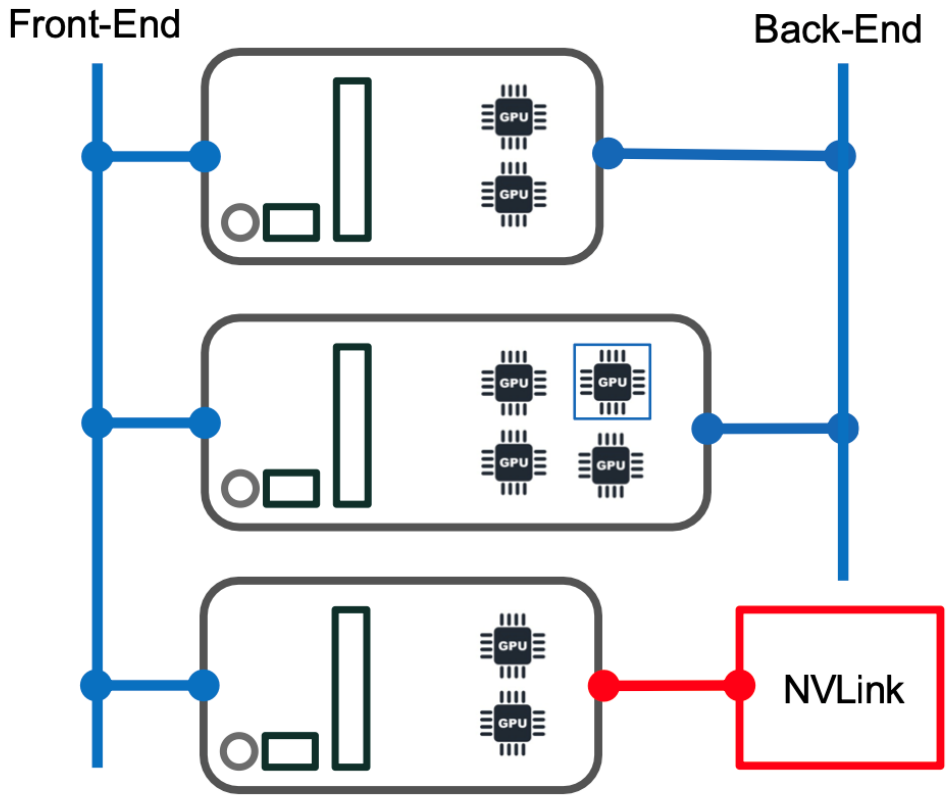
I dati di transfer rate sono teorici, nella realta' i valori misurati sono circa 80% dei valori teorici

Reinventing the DC

From Traditional to Ultimate Mode of Ops

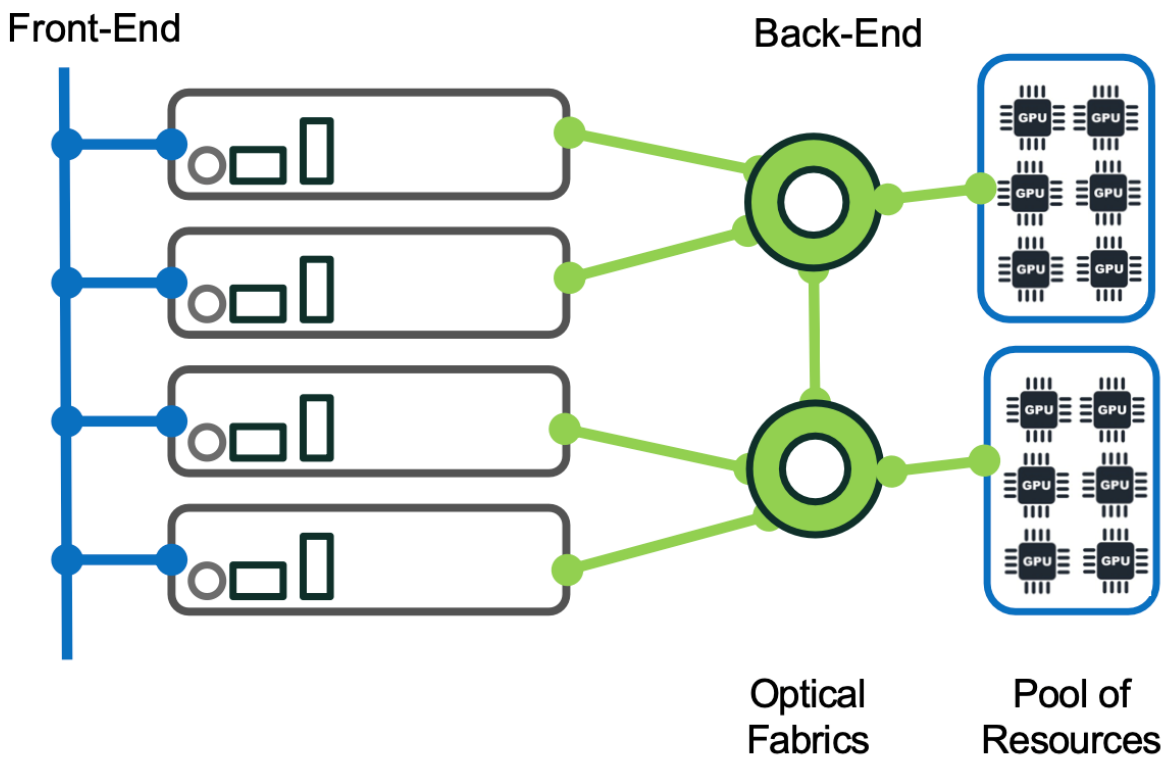
TMO

The Network is the Clustering Backbone
Proprietary multi-GPU interconnect



UMO

Optical Patch Panel to PCIe resources

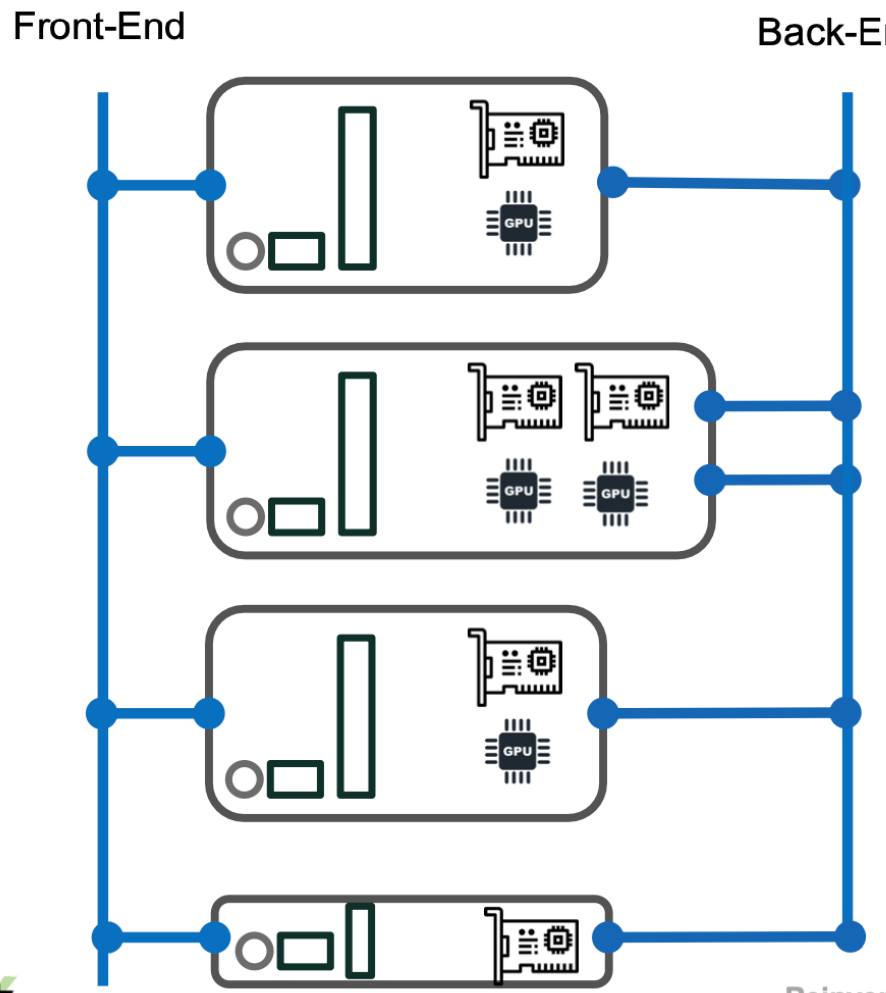


Drut

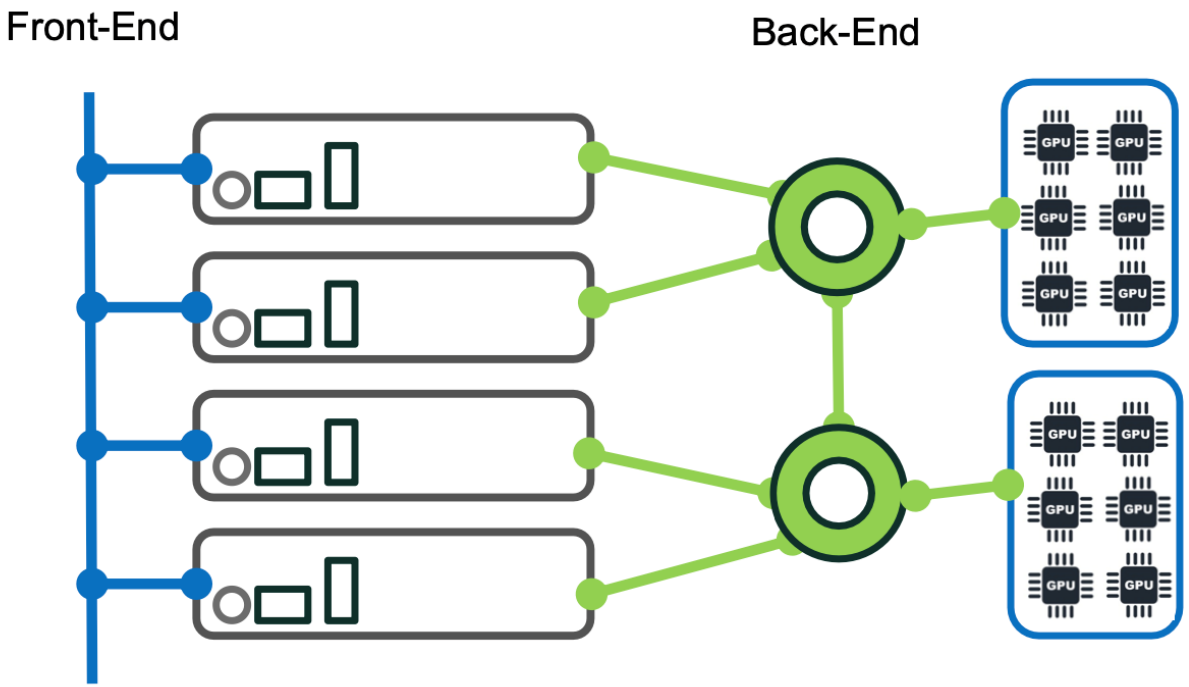
Better use of Resources, more Re-use, Less Capex

Systems Approach to New Solutions

RoCE

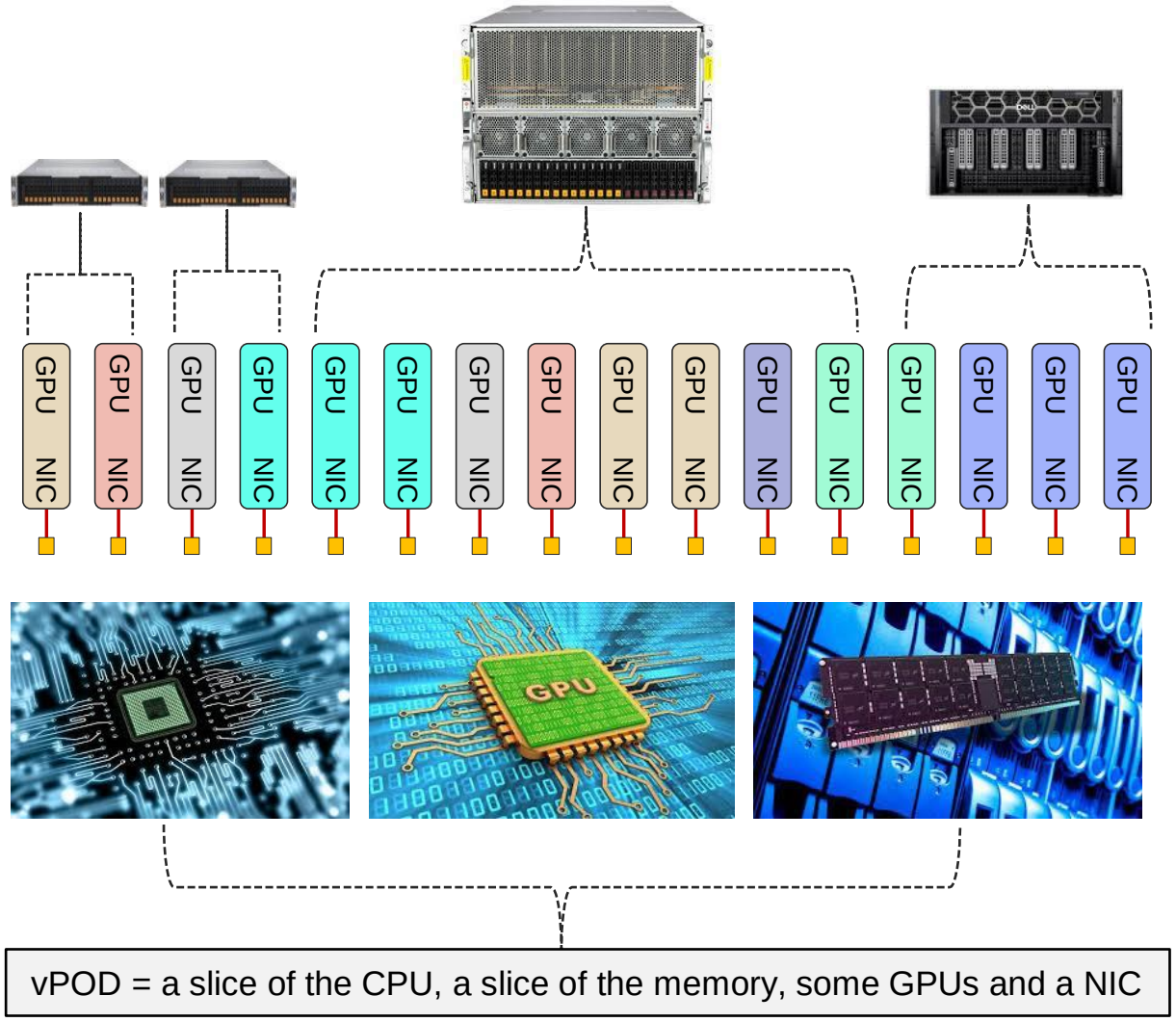


Disaggregated Design



INTRODUCING vPODs

- Next evolution of our Cloud Software Suite is called DX 3.0
- vPODs – A vPOD is a grouping of shared resources (e.g. CPU, GPU(s), Memory, NICs) in a secure user/workload entity
- vPODs are reconfigurable and begin the process of bringing dynamic configurability to what are difficult to configure static siloes
- vPODs can span different physical machines
- vPODs utilize a GPU back-end network - using off the shelf RoCE NICs 100/200/400G
- Designed for enterprises and service providers deploying GPUs with need to isolate resources to improve utilization and create secure user defined resource groups



Systems Approach in More Details

RoCE and vPODs

- DX 3.0 Release
- vPODs for GPU isolation / Containers
- Supports RoCE
- Supports RDMA over low-latency, all photonic fabric (PXCs)
- Creates the highest performing bare metal GPU machine, in a virtual machine
- Uses off the shelf server systems and RoCE NICs
- Works with any server platform
- Works with any GPU vendor
- Works with existing Ethernet switches
- Works with Drut PXC (Photonic switch)
- Server / Fabric integration as a software-controlled resource grouping that provides an isolated GPU instance

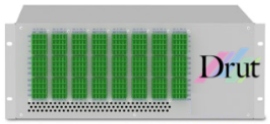
Optical Back-end for a powerful vPod data & control plane

Optical Back-end together with full resources pooling

Disaggregated Design

- Combines Cloud Suite Software and DX 3.0 into a complete system
- Full-server disaggregation
- PCIe over Photonics (PoPH)
- Fabric Interface Cards (FICs)
- Photonic Resource Units (PRUs)
- Photonic Switches (PXCs)
- Drut Cloud Suite Software
- vPODs at scale
- Large resource cluster designs

Your favorite Ethernet Switch Vendor



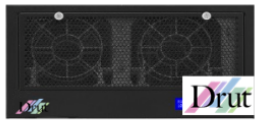
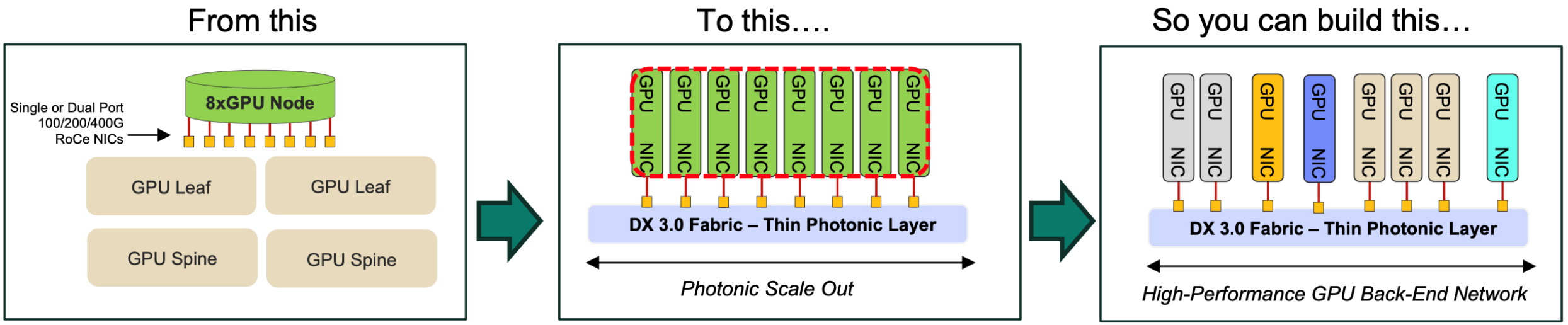


Illustration: 8xGPU Server Transition



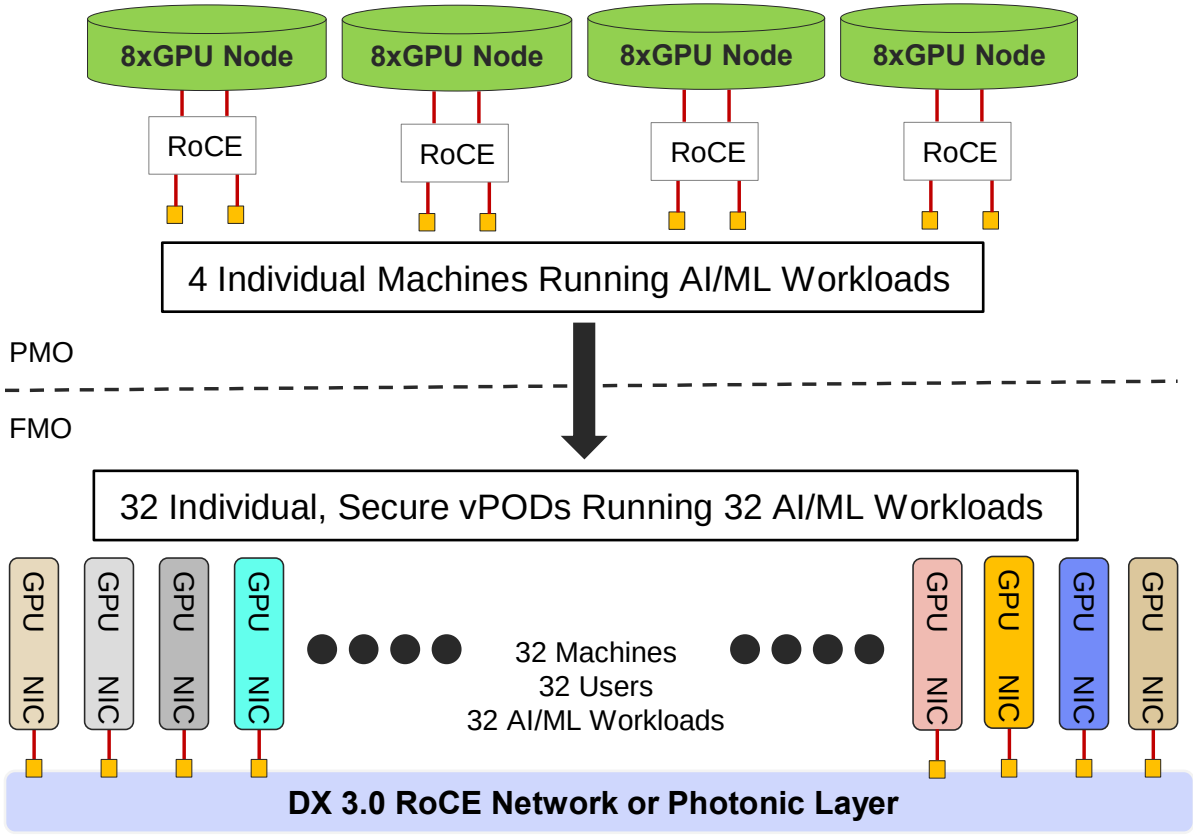
- DX 3.0 turns machine silos and network silos into user definable secure spaces called a vPOD
- Uses off the shelf servers and NIC cards
- Have 16x8-GPU servers and want to turn them into 128 individual machines? Or maybe 64 dual GPU machines for 64 users? DX 3.0 can do that for you.

- DX 3.0 delivers a secure, private resource grouping over a low-latency photonic fabric under software control.

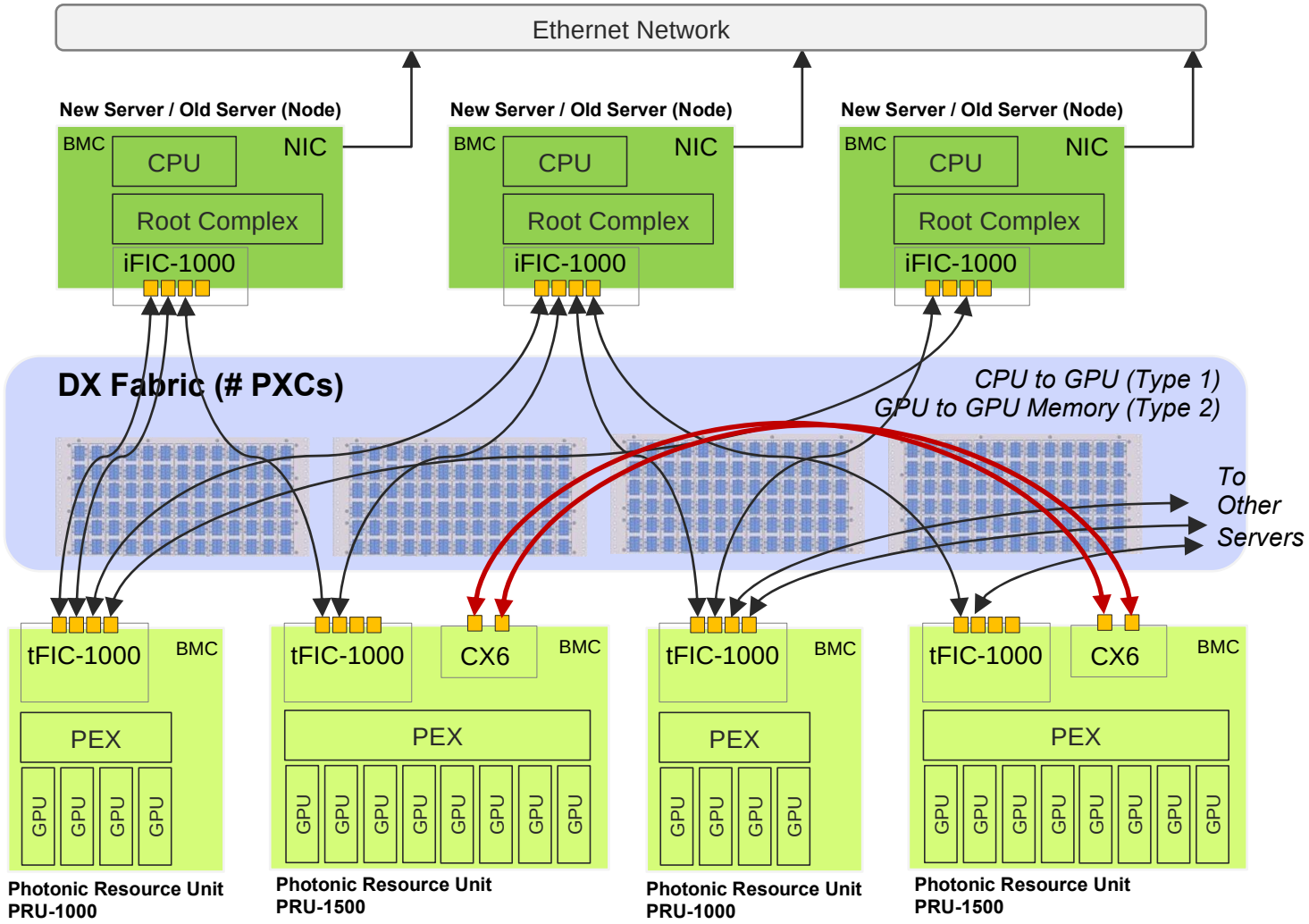
=Any Combination of Secure vPODs

SMALL MACHINES FROM BIG MACHINES

- User has four 8xGPU Machines
- The resources in these machines can be separated into vPODs to create 32 individual machines
- Or they can be reconfigured to 4xGPU or 2xGPU machines
- Diurnal Resource Grouping - vPODs can be created for AI/ML workers during the day and night if GPU resources are not utilized, they can be right-sized for evening workloads



DX Fabric 2.0 – Combine with RDMA



- Using a PRU-1500 with one tFIC and one CX-5, CX-6 or CX-7 allows for Direct Memory Access between GPUs with out involving the CPU
- GPUs communicate directly over the DX Fabric
- CX-6 has two ports and CX-7 has four ports that can be used for GPU memory sharing
- Lowest possible latency
- Direct connect fabric, no switch hops
- Private, secure memory sharing isolated from other workloads

From cluster to single image architecture for training – Cerebras WSE2

Our customers can easily train and reconfigure GPT-3 and GPT-J language models with up to 20 billion parameters on a single CS-2 system

The last few years have shown unprecedented growth in the size and complexity of natural language processing (NLP) language models. The result is that the models are so large that they must be trained using hundreds or thousands of conventional processors in super-computer-style clusters. So complex are the models, and so difficult are the compute clusters to set up, that only a very small portion of the artificial intelligence (AI) community can train them. Only a handful of companies around the world have this kind of capability.

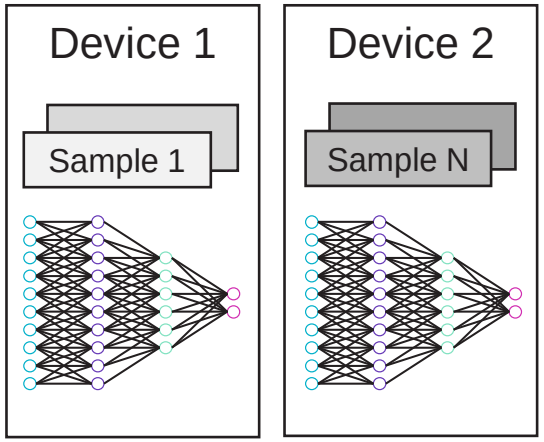
This is painstaking work, and often takes months. And this work is not ML! It's distributed computing work. Interestingly, what is most challenging about very large ML is often not the ML. It's the distributed computing.

https://medium.com/@cerebras/cerebras-sets-record-for-largest-artificial-intelligence-models-ever-trained-on-single-device-d1cadf1ce875

Challenges to Scaling on GPU Clusters

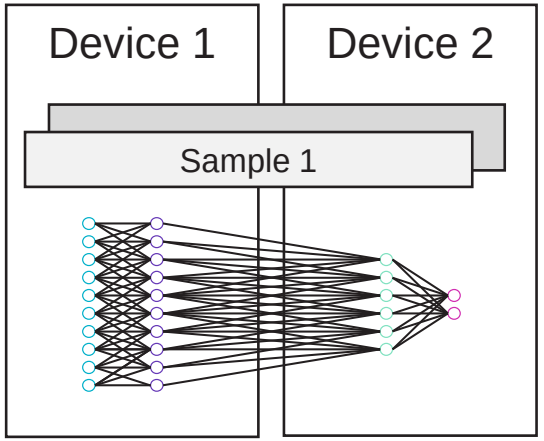
Hybrid parallelism on traditional devices

Data Parallel



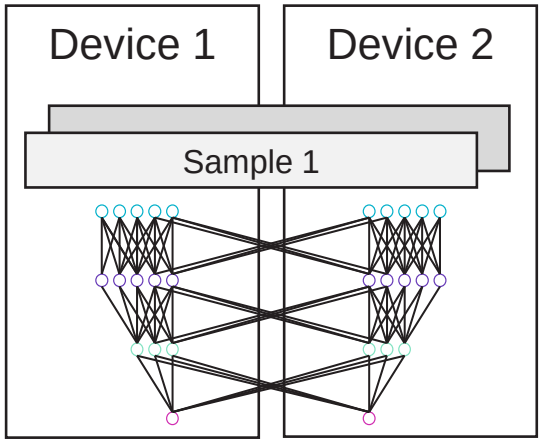
Multiple samples at a time
Parameter memory limits

Pipelined Model Parallel



Multiple layers at a time
Communication overhead
N^2 activation memory

Tensor Model Parallel

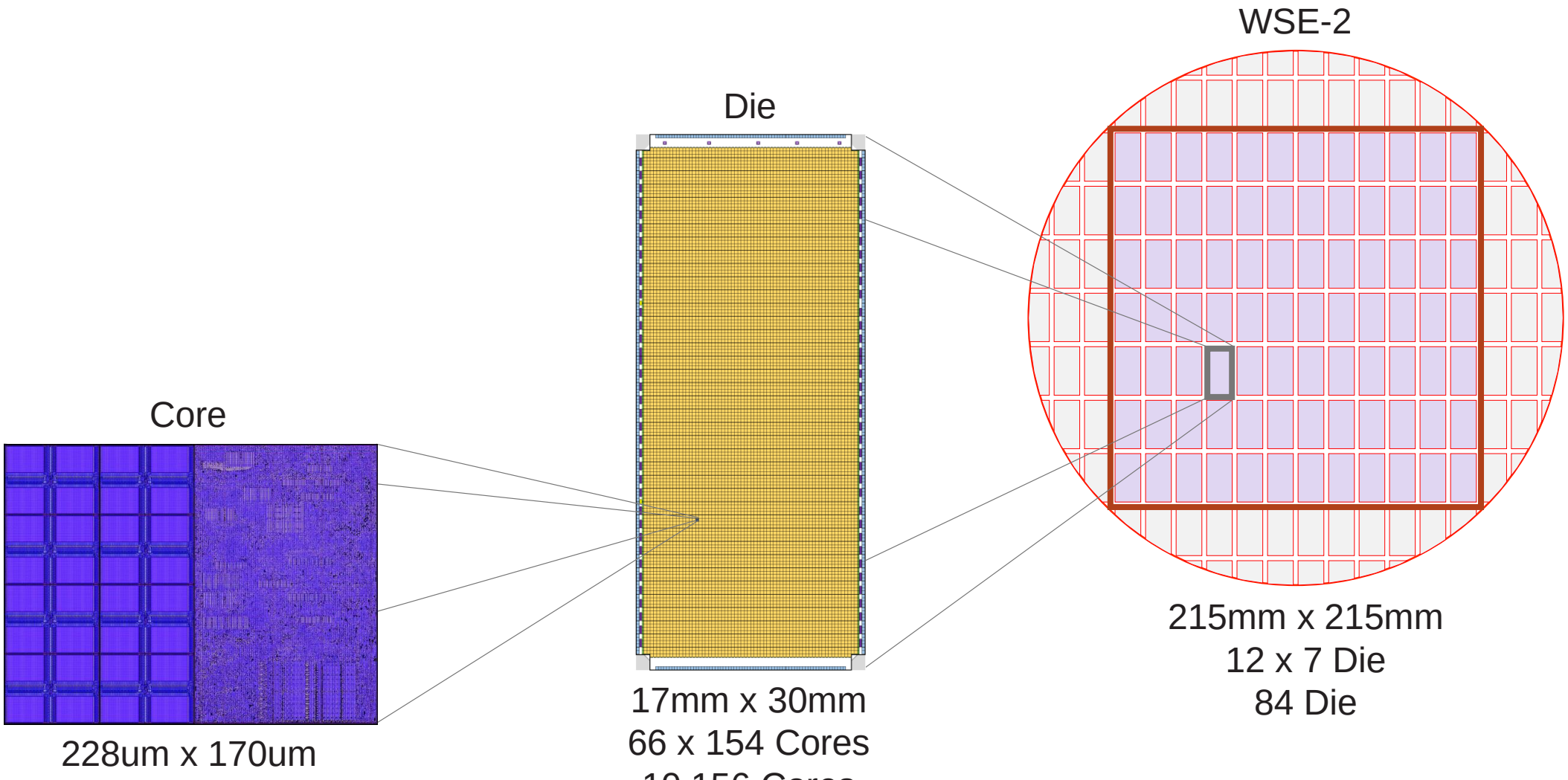


Multiple splits at a time
Communication overhead
Complex partitioning

Distribution complexity scales dramatically with cluster size

https://medium.com/@cerebras/cerebras-sets-record-for-largest-artificial-intelligence-models-ever-trained-on-single-device-d1cadf1ce875

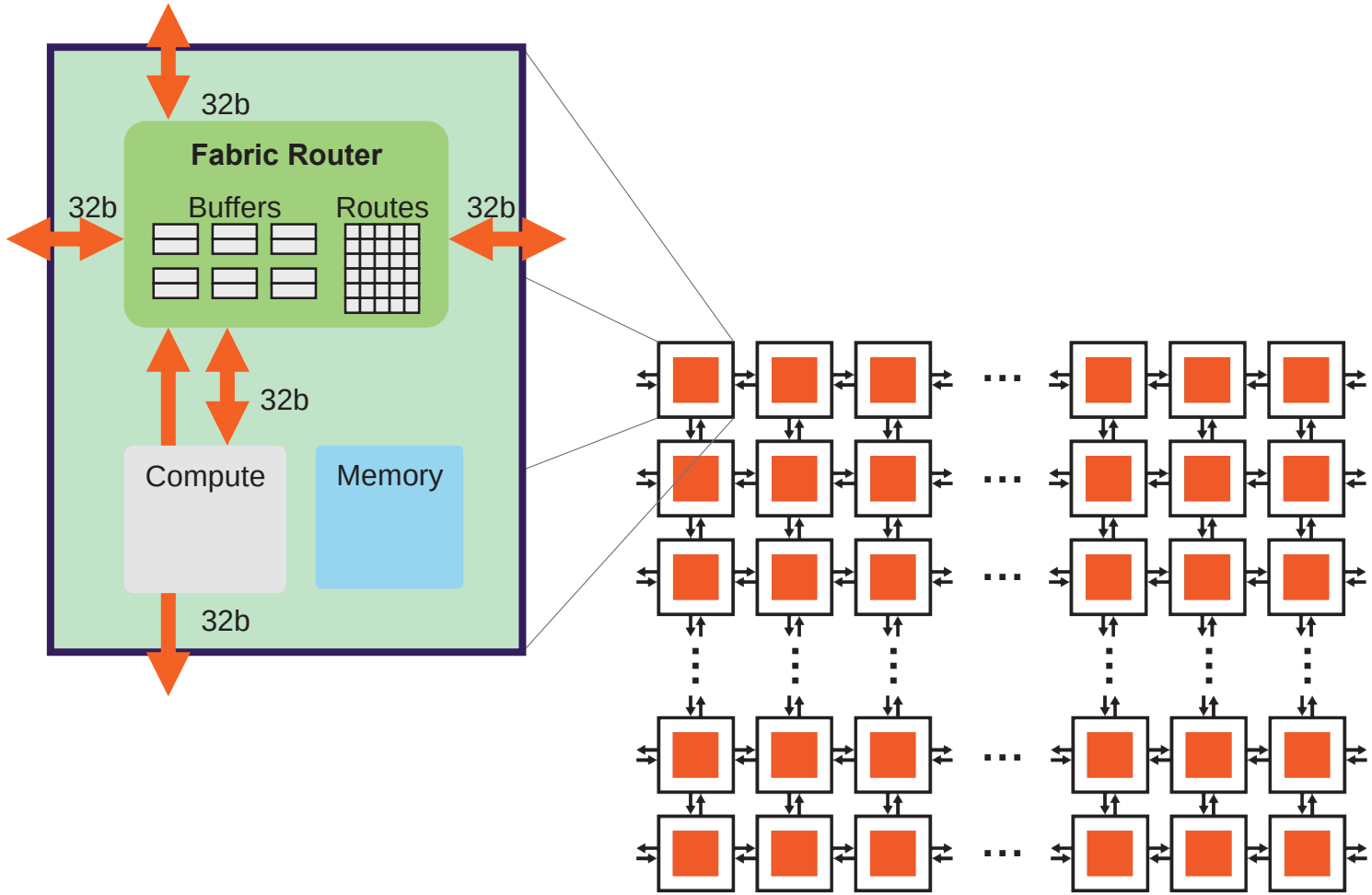
From Small Core to Massive Wafer



High Bandwidth Low Latency Fabric

Efficient high performance

- 2D mesh topology with low overheads
- 5-port router to 4 neighbors and core
- 32b/cycle bidirectional data transfer
  - Individual packages are 32b
  - Payload carries data (16b) and index (16b)
- Single cycle latency between cores
  - Flow controlled with low buffering
- 24 configurable static routing (colors)
  - Each color has dedicated buffering, is non-blocking
  - All colors are time-multiplexed onto same physical link
- Hardware broadcast/multicast



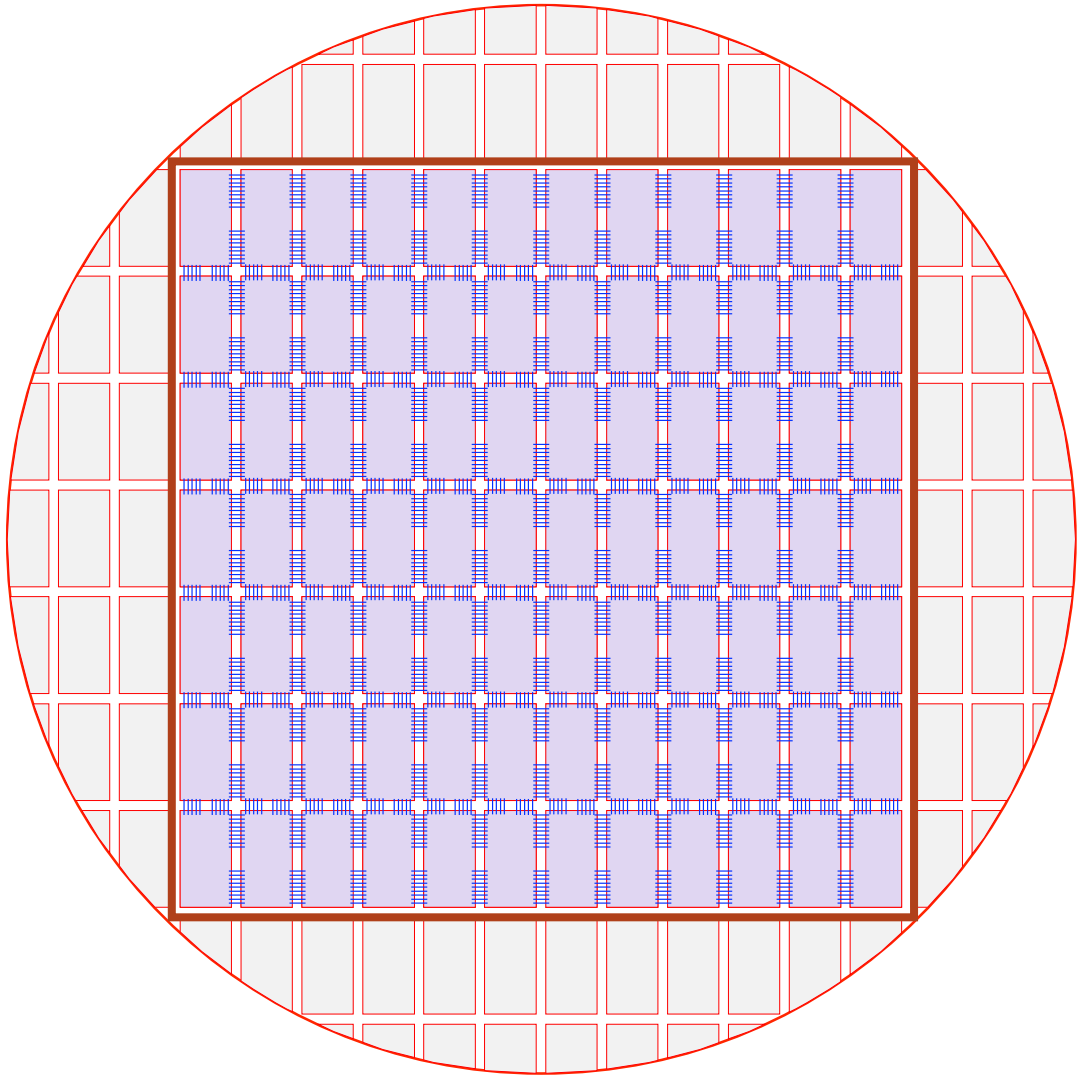
Uniform Fabric Across Entire Wafer

Designed to scale beyond individual die

- Bridge <1mm across scribe lines between die
- Source synchronous parallel interface
- Redundancy with training and auto-correction state machine

Uniform bandwidth across entire wafer

- The entire wafer is a single chip all with *on-chip* bandwidth
- Full bandwidth within die and between die
- Wafer integration enables ultra short inter-chip links

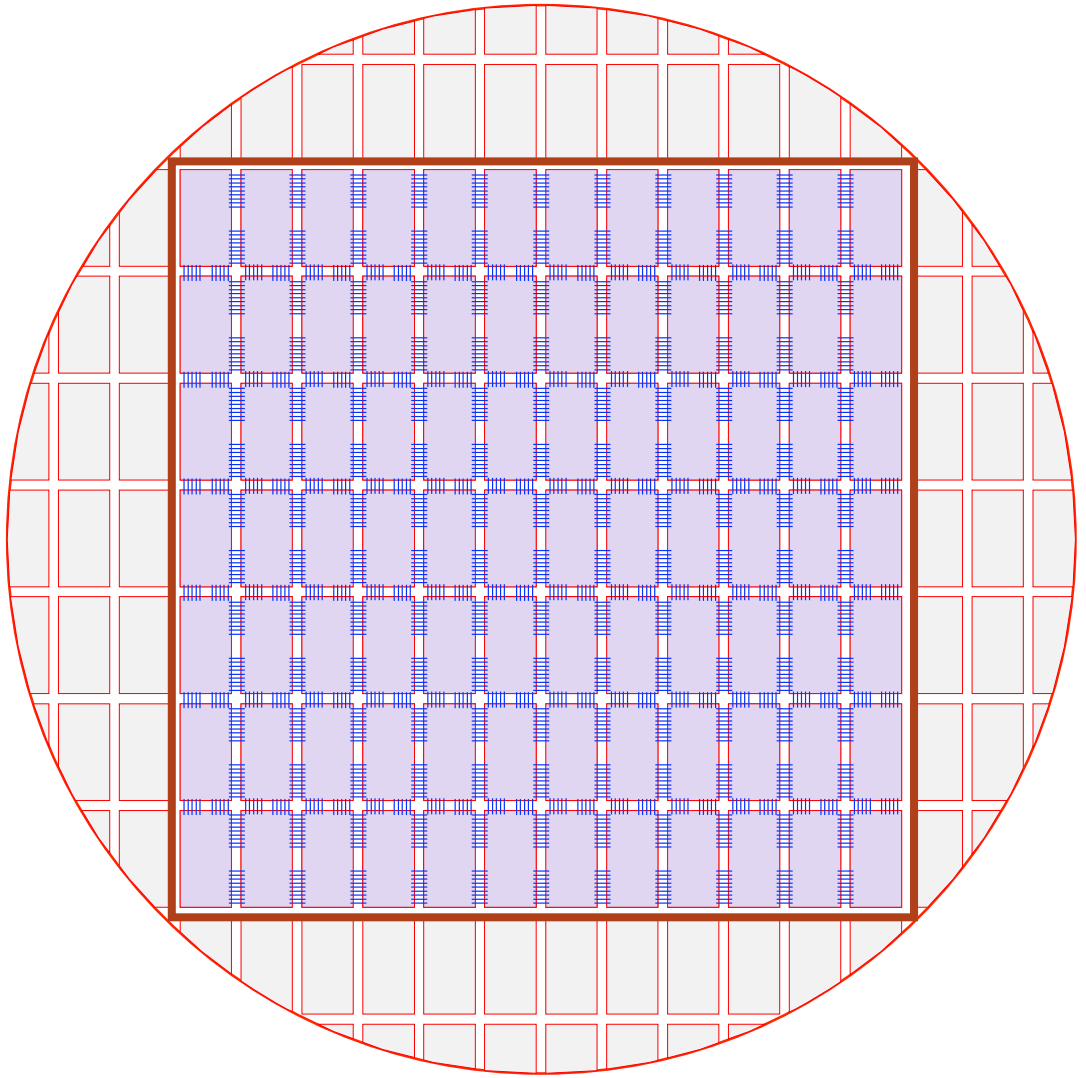


Unprecedented Fabric Performance and Power

	Area	Bandwidth		Power	
	Mm^2	TB/s	$GB/s/mm^2$	pJ/bit	W
GPU Estimate	826	0.6	0.7	10	60
WSE-2 Sub-fabric	826	4.3	5.2	0.15	6
Ratio		**7x**	**7x**	**66x**	**10x**

Wafer-scale fabric architecture gives WSE-2
unprecedented fabric bandwidth and power

7x normalized fabric bandwidth vs. GPU



*GPU estimate uses 7nm PCIe 5.0 serdes with Nvidia A100 NVLink bandwidth
*WSE-2 sub-fabric uses subset of wafer area equivalent to Nvidia A100

All Model Sizes at Extreme Performance on a Single Chip

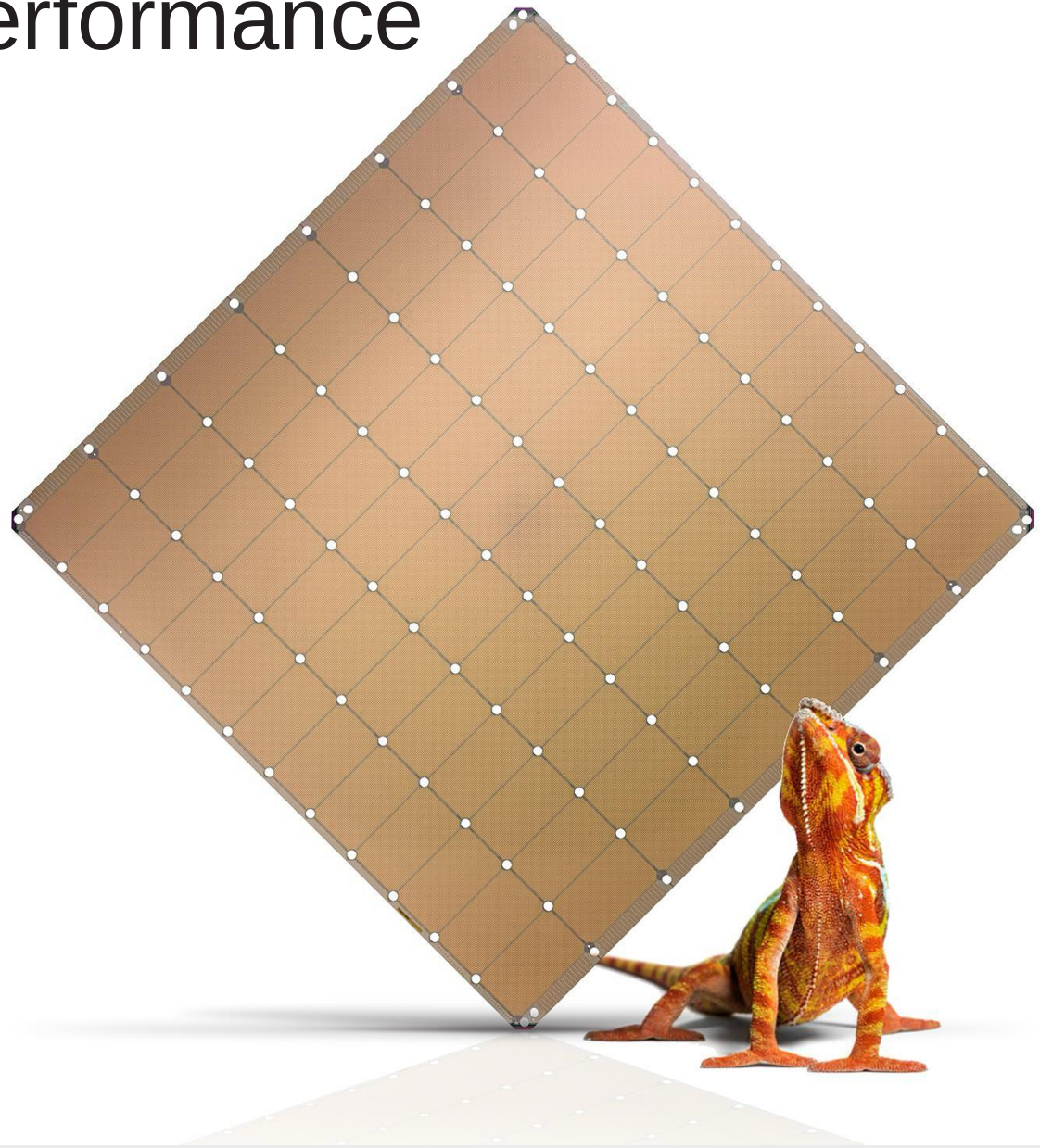
Architecture enables efficient wafer-scale computation

- Full bandwidth memory to datapath
  - AXPY operations for sparsity acceleration
- Dataflow scheduling
  - Unstructured sparsity acceleration by skipping zero weights
  - Massive model support by never storing weight matrix
- High bandwidth wafer-scale fabric
  - Global weight broadcast and reduction across wafer

No matrix blocking or partitioning required

Up to 100k x 100k MatMul
Run **models of all sizes** in a single device with

75 PFLOPS FP16 Sparse
7.5 PFLOPS FP16 Dense



cerebras

*Sparse performance with 10x sparsity acceleration

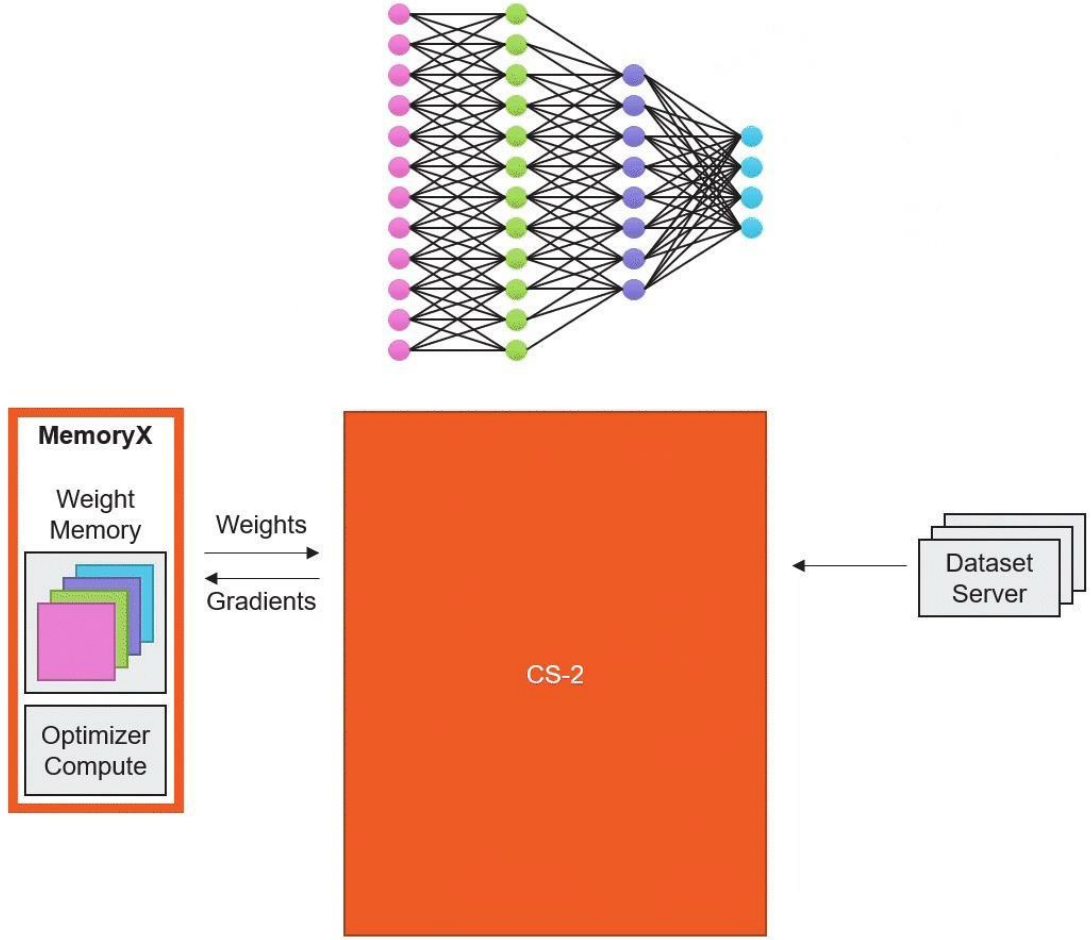
All Model Sizes on a Single Chip

Cluster-scale compute in a single chip

- Train the largest neural networks (e.g. GPT-3)
- On a single chip without partitioning

Built for extreme-scale neural networks

- *Weight Streaming* execution decouples memory from compute
- Weights stored externally off-wafer in MemoryX
- Weights streamed onto wafer to compute layer
- Execute one layer at a time
- Gradients streamed out of wafer
- Weight update occurs in MemoryX



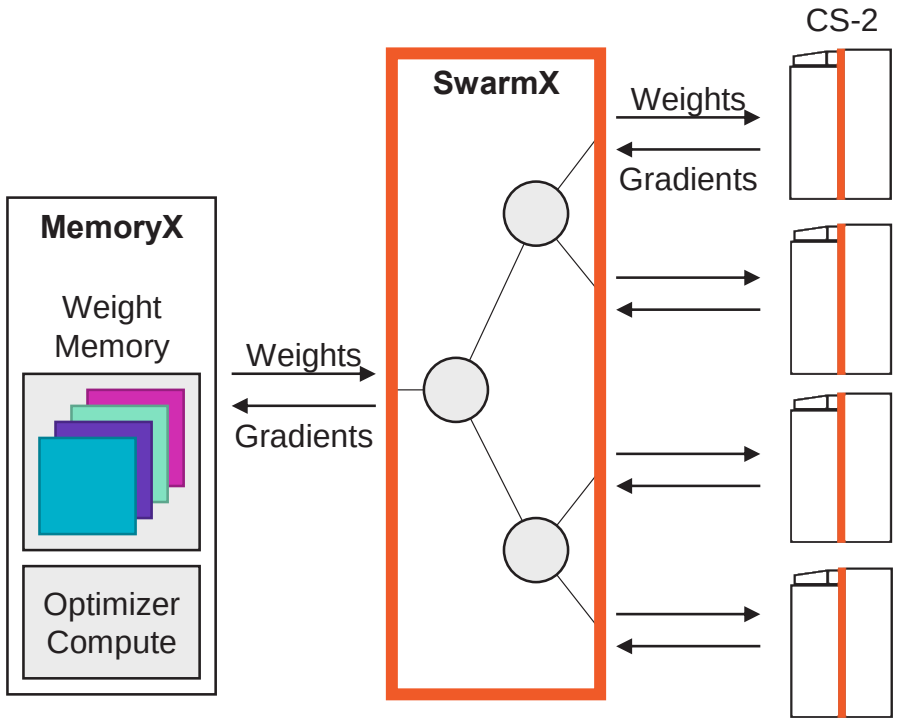
Near-Linear Data Parallel Only Scaling

Specialized interconnect for scale-out

- Data parallel distribution through SwarmX interconnect
- Weights are **broadcast** to all CS-2s
- Gradients are **reduced** on way back

Multi-system scaling with the same execution as single system

- Same system architecture
- Same network execution flow
- Same software user interface



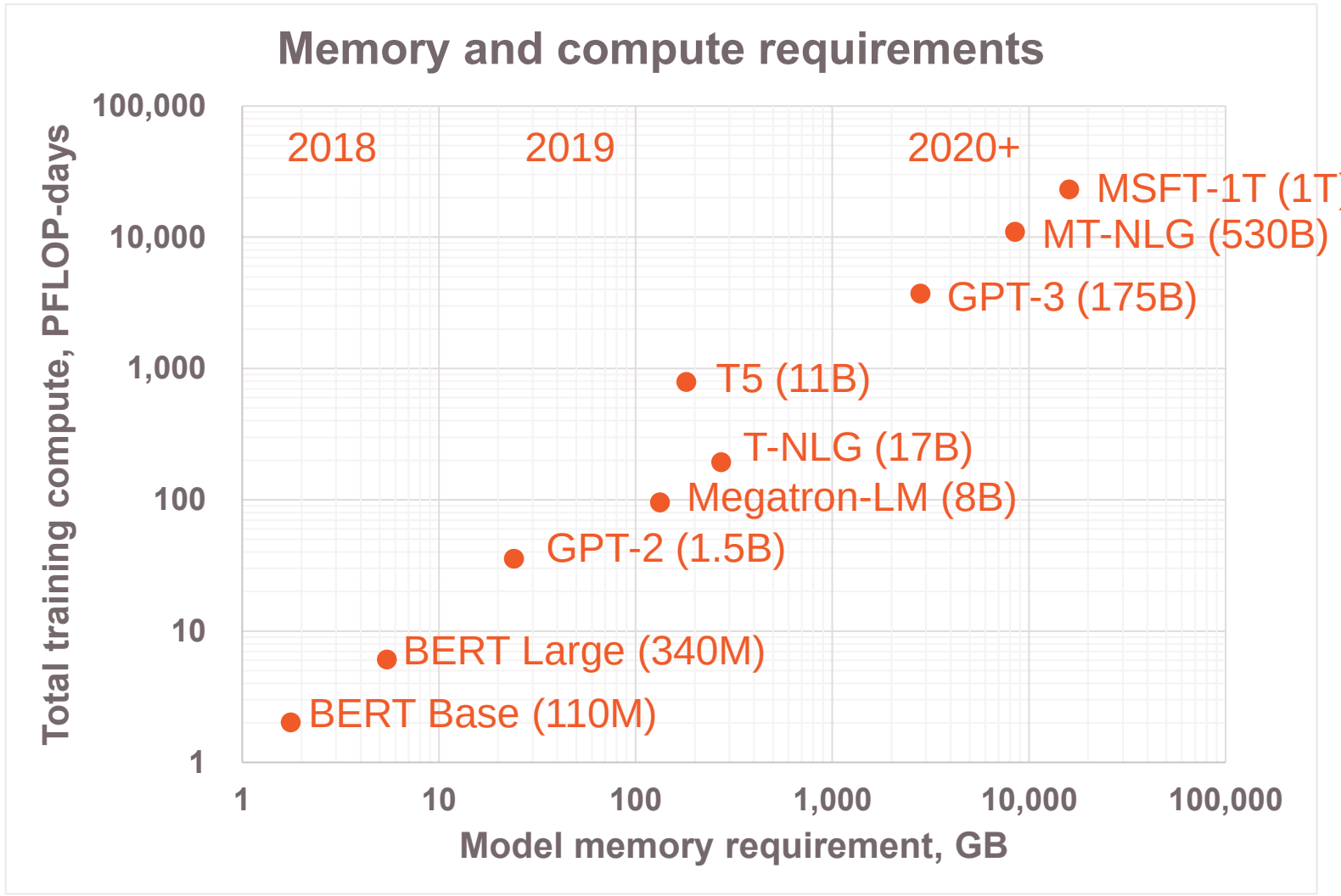
World's most powerful processor chip

First, our Wafer-Scale Engine is physically 56 times larger than the largest GPU. We have 123 times more cores, 1,000 times more on-chip memory, 12,000 times more memory bandwidth, and 45,000 times more fabric (Figure 8). These are the resources that enable us to fit the largest layers of the largest neural networks onto a single wafer. In fact, on the Wafer-Scale Engine, we can fit layers 1000 times larger than the largest layer in the largest NLP network!

	Cerebras WSE-2	Nvidia A100	Cerebras Advantage
Cores	850,000	6912 + 432	123 X
On-chip memory	40 Gigabytes	40 Megabytes	1,000 X
Memory bandwidth	20 Petabytes/sec	1,555 Gigabytes/sec	12,733 X
Fabric bandwidth	220 Petabits/sec	600 Gigabytes/sec	45,833 X

Figure 8. Specification of Cerebras WSE-2 processor vs. Nvidia A100 GPU.

The Grand ML Demand Challenge



Is it possible?

- If you want to change the solution, you must first change how you think about the problem

Technical R&D today: Disruption opportunity

Knowledge

Researchers need to easily access quickly growing and widely diverse information sources.

- Highly unstructured/dark
- Current human based approach not scalable

Inference & Simulation

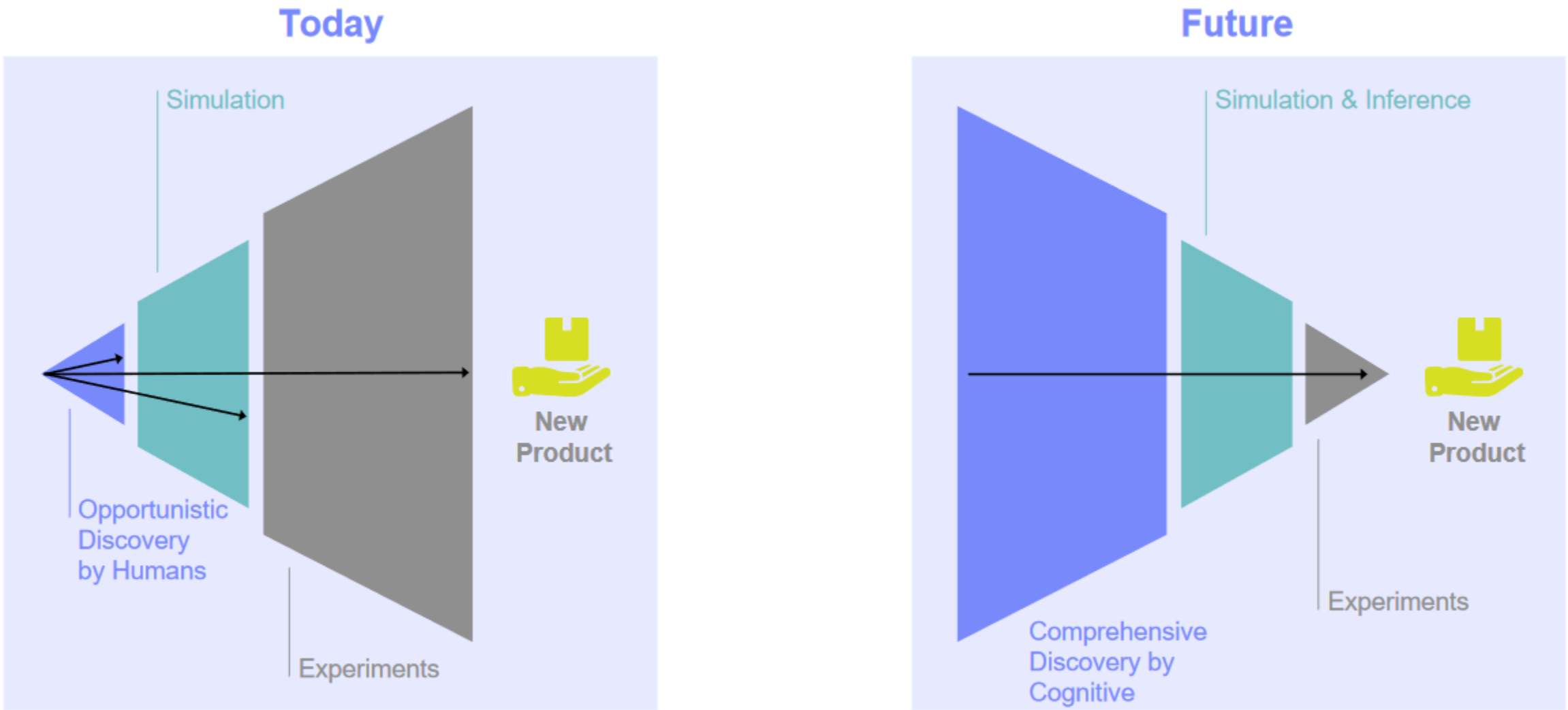
Domain related inference is largely missing. Setting up and deploying the right simulations is very hard.

- Human capital intensive, non scalable

Evidence & Experiments

Internal evidence and experiments are driven primarily empirically, often brute force, and their results are isolated from wider knowledge space.

Technical R&D today: Disruption opportunity



Computing Reimagined: Knowledge Discovery Pipeline

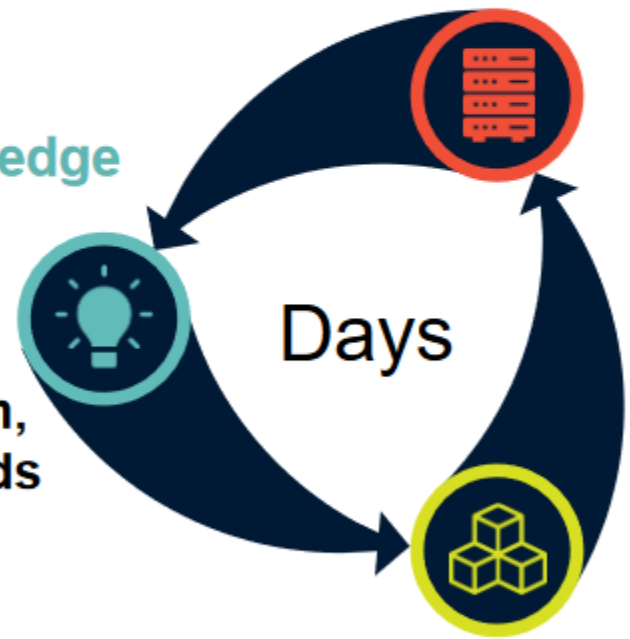
Ingest structured and unstructured data to create massive knowledge spaces

- Complex documents
- Structured DBs
- Simulation/Previous runs
- Public and proprietary

Knowledge Space

Enable contextual based search, based on meaning, not keywords

- Tables, Images, Formulas, Diagrams
- *"What are all the properties of a chemical" -- "What is the cheapest health contract for non smokers younger than 25"*



Deep Search

Inference

Get deep insights by ML/DL on the Knowledge Space

- Identify trends
- Discover gaps in the knowledge
- Explore "what if" scenaria
- **Run the "right" simulations**
- Enrich knowledge space

- *"How has the use of copper in alloys in the auto industry evolved in the last 20 years?"*
- *"What is the most likely use of a certain chemical by-product of an organic synthesis route next year? Which company is the most likely to want to buy it?"*
- *"How do I come up with the most relevant questions to ask a client for an insurance contract based on their answers so far?"*

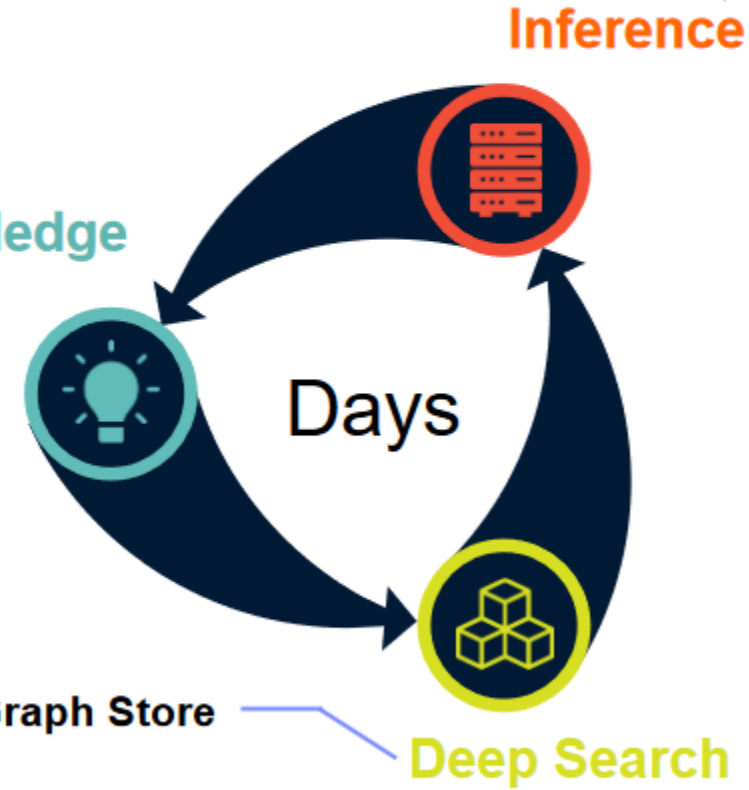
Knowledge Discovery Pipeline: Key APIs

Corpus Conversion

- PDF Smart annotator. Any type of document. PDF is key here. Minimal and trivial annotation needed

NLP: Facts extraction

- Smart Reader: non supervised extraction from text
- Diagram Ingestion: extracts data from diagrams, scientific plots, schematics
- Tables and Forms extraction and semantic representation



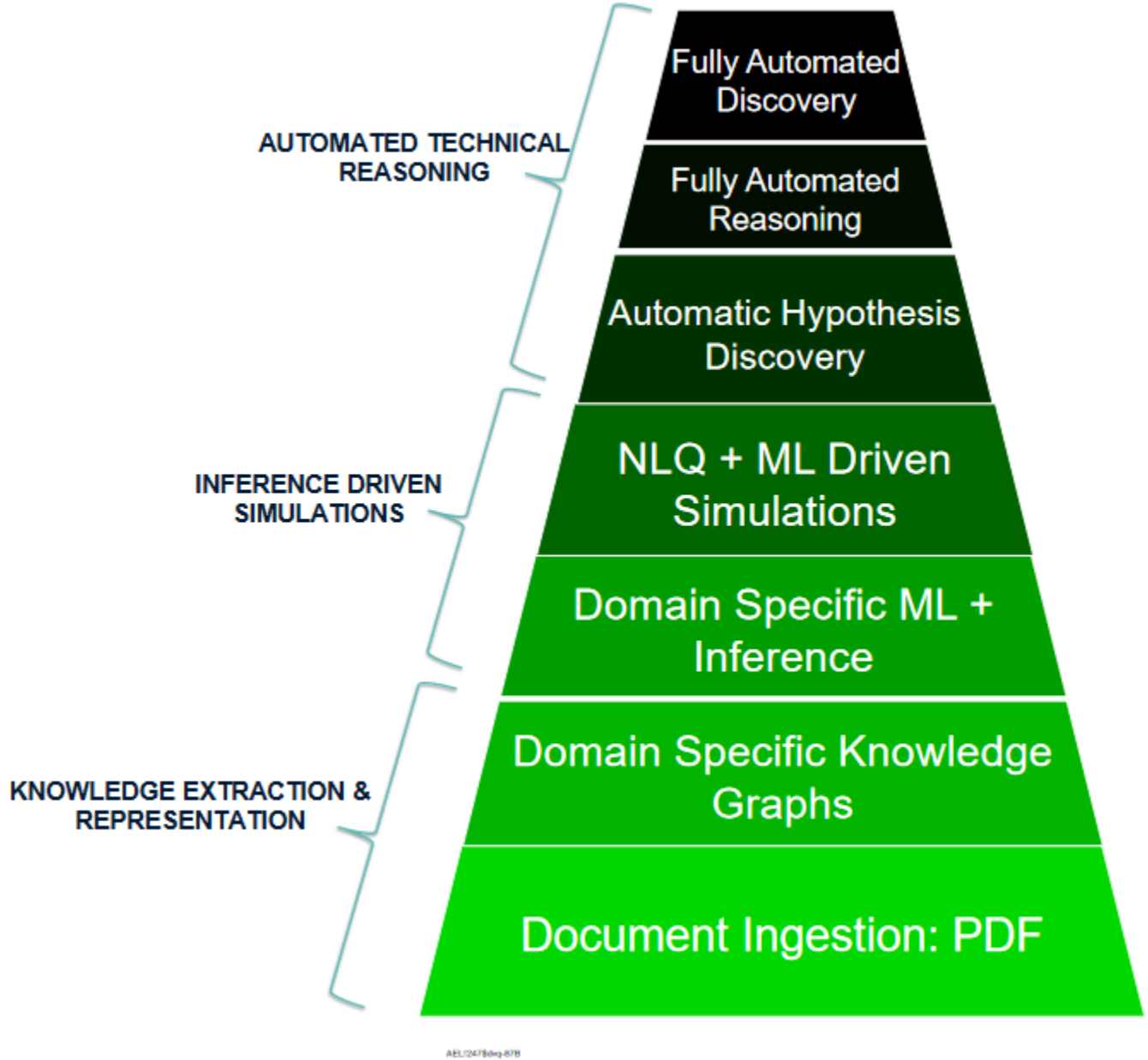
Knowledge Graph Store & Engine

- High performance in memory graph DB and query engine

Analutos

- Low cost & scalable graph analytics and inference algorithms. **Surrogate DL models to physical/mechanistic models.**
  - Node importance (spectral centralities)
  - Graph simplification
  - Graph comparisons
  - Data uncertainty quantification
  - HPC Machine and Deep Learning
  - Autotuning for DL
  - Scalable non supervised text analytics: Word2vec, Node2vec

Technical Computing Reimagined: Holistic View

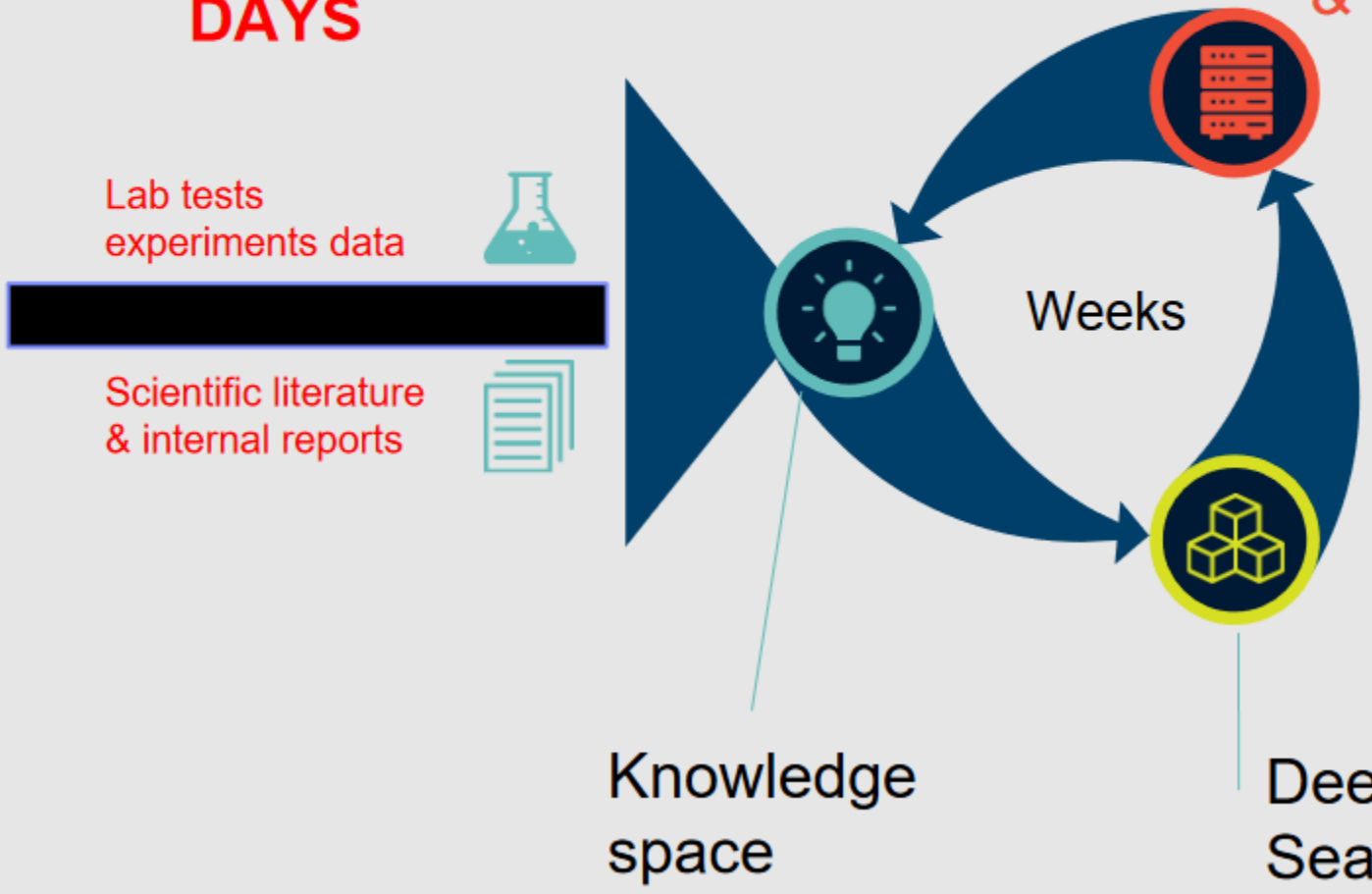


The Materials Discovery Case Study: Alloy design

- Atomistic simulations
- Deep Learning based property prediction

Simulation & Inference

DAYS

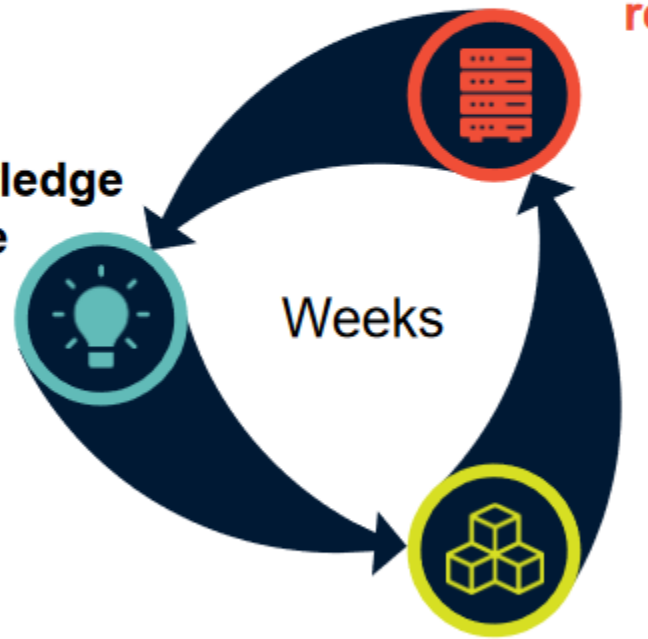


Design alloys to avoid catastrophic failure that can lead to huge liabilities

- Corrosion
- Cracks
- Special environmental and deployment conditions

Cognitive Discovery: An Algorithmic view to computing The Future is here!

Ingest data and create massive knowledge spaces



- Supercomputing
- Neuromorphic, Quantum and new computing paradigms
- Inference (ML)

Link evidence with knowledge spaces. Drive deep search

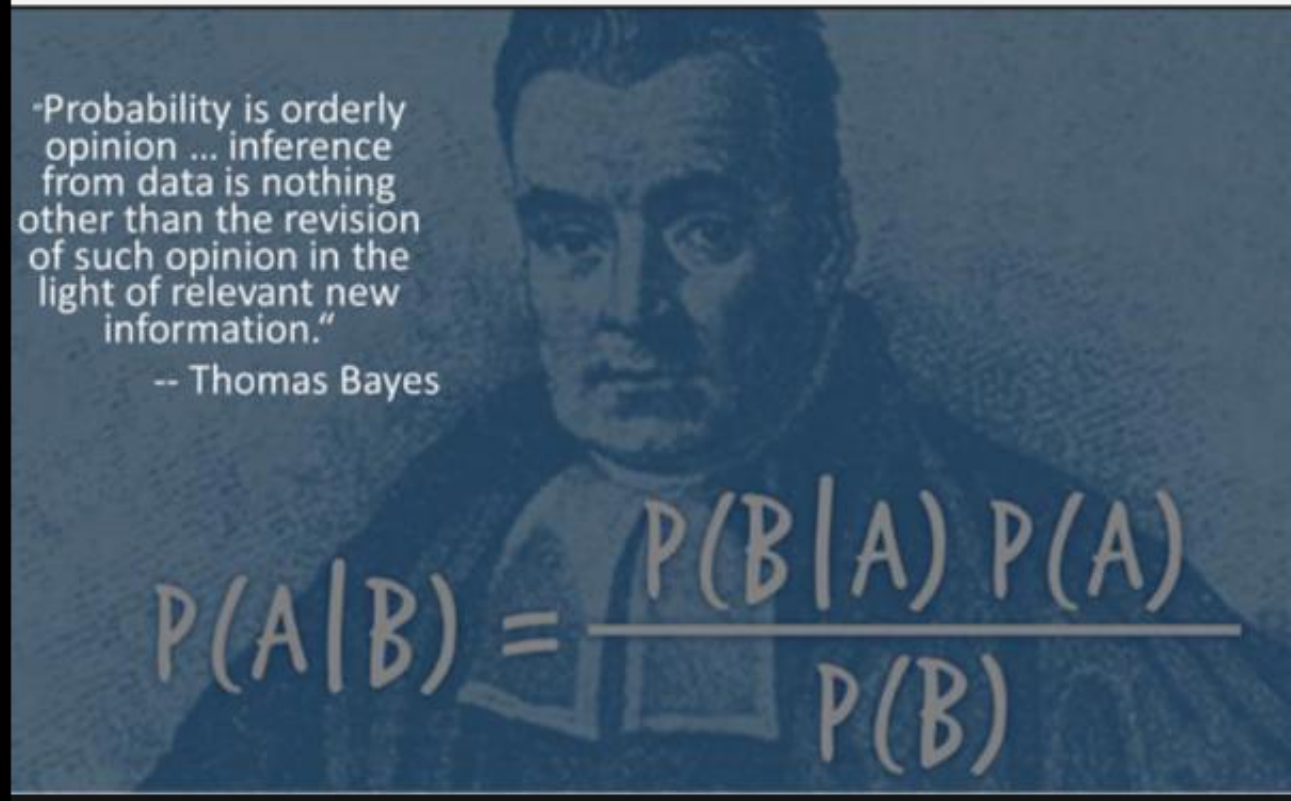
INTRODUCING …

The IBM Bayesian Optimization

- ✓ State of the art scalable Bayesian optimization
- ✓ Broad spectrum of applications
- ✓ Delivered in an easy to consume API

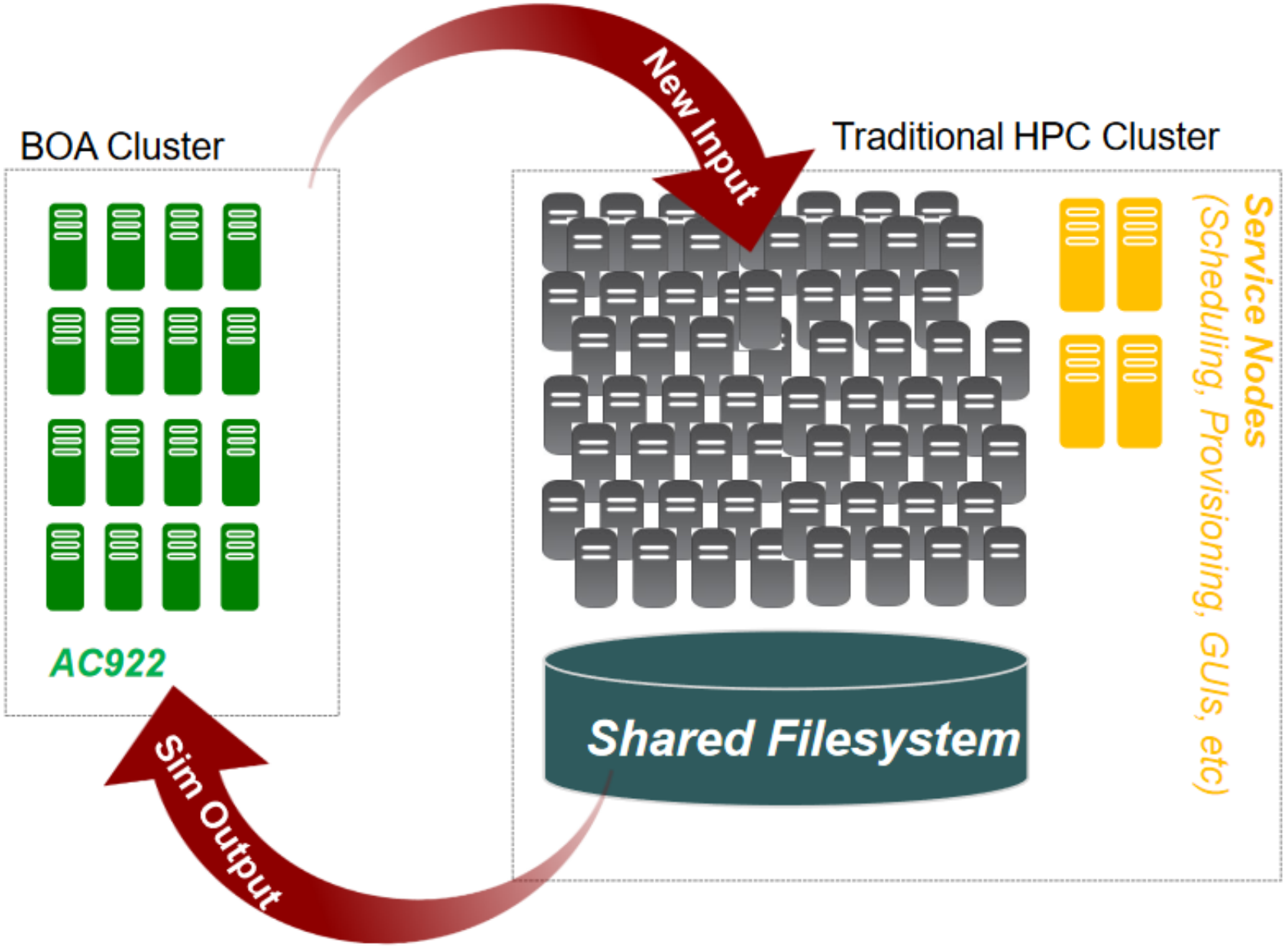
Bayesian optimization is a sequential design strategy for global optimization.

Many workflows require you to find a powerful set of parameters solve a problem. The challenge is finding those parameters robustly in as little time as possible.

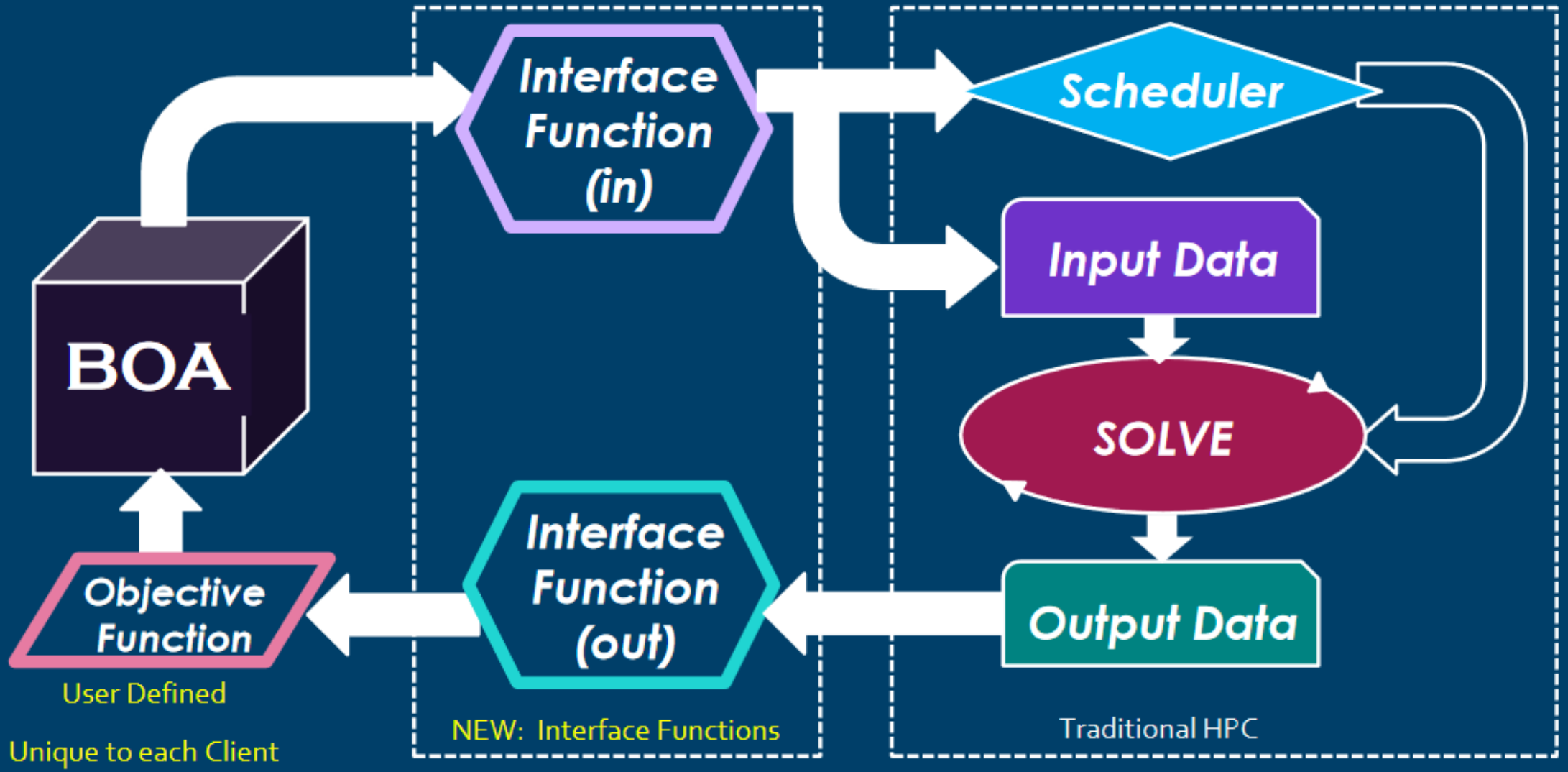


BOA Topology

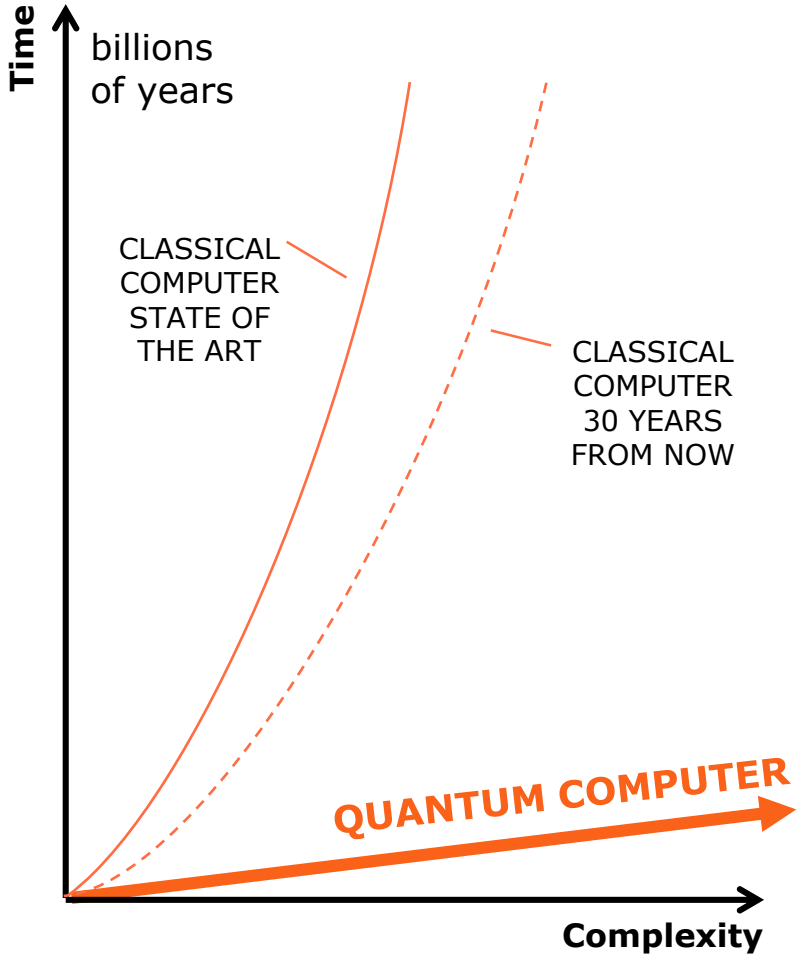
- BOA servers are dedicated to running BOA only
- Physically co-located with the HPC environment
- Large BOA systems include multiple GPU enabled nodes for throughput
- Multi-user, multi simultaneous experiments being simulated in the HPC



Interface Functions



Quantum Computing is moving from promise to reality



It will **massively reduce the complexity** of some problems currently intractable to classical computers…

Quantum computing uses quantum physics to **solve complex issues** that traditional supercomputers cannot.

It exploits the capabilities of ultrafast calculations to accelerate solutions that **tackle critical societal and organizational challenges.**

Industries are adopting quantum computing to stay competitive and innovative, particularly those with optimization needs.

The importance of this leap is so great that companies who fail to act will lose their competitive edge.

Qaptiva™ Partner Ecosystem

Expanding offerings and capabilities to deliver more value

Photonics

2 to 12 optical qubits

Co-design approach or ready-made hardware

Available now on Qaptiva™ as a Service

- Hosted by Quandela
- VQE example
- 10000 shots - 20 seconds execution – few euros to run

Superconducting

Gate-based paradigm

World-class error rates

Co-design approach or ready-made hardware

Use-case-specific hardware design

- As a Service in 2024
- Plan to be hosted by Eviden
- 5 qubits capabilities

Neutral atoms

Analog and gate-based computing paradigm

Up to 200 qubits

Deep integration with myQLM tools

CAT Qubits

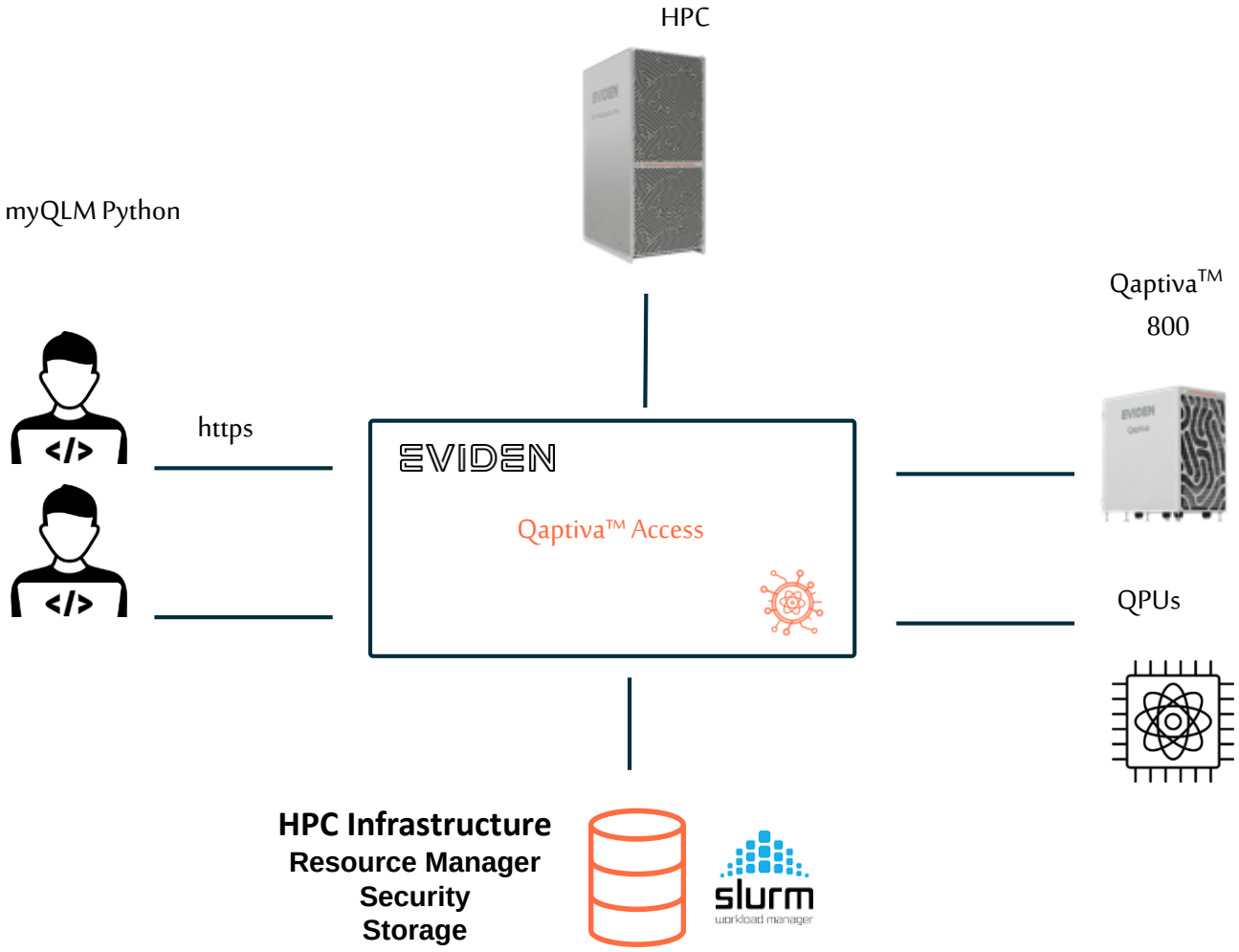
Innovative hardware-efficient design will reduce the hardware requirements for a fault-tolerant quantum computer

FD-SOI

FD-SOI (Fully Depleted Silicon On Insulator) uses an ultra-thin layer of silicon over a buried oxide as a means to reduce leakage and variation in chips.

Qaptiva™ Access Server (1/2)

Front-end server to orchestrate quantum resources and enable HPC and quantum hybridization



It enables the integration of any quantum processing unit (QPU) and emulator into the high-performance computing (HPC) infrastructure.

- Real scheduling of QPUs with SLURM
- Scale-out numerical simulation (MPI + GPU)
- Used in several HPC-QC pilots:
  - HPC-QS by EuroHPC
  - HQI in France
  - Qsolid in Germany

Thank You

Marco Briscolini, PhD

marco.briscolini@gmail.com

Cell: 3357693820

03/19/2025