

Addressing the Hidden Costs of AI

Live Webinar

February 27, 2024

10:00 am PT / 1:00 pm ET



Today's Presenters



Erik Smith
Distinguished Engineer
Dell Technologies



Rob Davis
VP of Storage
Networking
NVIDIA



David McIntyre
SNIA Board of Directors
Samsung



Mishali Naik
Principal Engineer and
Architect
Intel



Justin Potuznik
Engineering Technologist
Dell Technologies

The SNIA Community



200
Corporations,
universities, startups,
and individuals



2,500
Active
contributing
members



50,000
Worldwide
IT end users and
professionals

Ethernet, Fibre Channel, InfiniBand®

iSCSI, NVMe-oF™, NFS, SMB

Virtualized, HCI, Software-defined Storage

Storage Protocols (block, file, object)

Securing Data

Technologies We Cover

SNIA Legal Notice

- The material contained in this presentation is copyrighted by SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
 - Any slide or slides used must be reproduced in their entirety without modification
 - SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

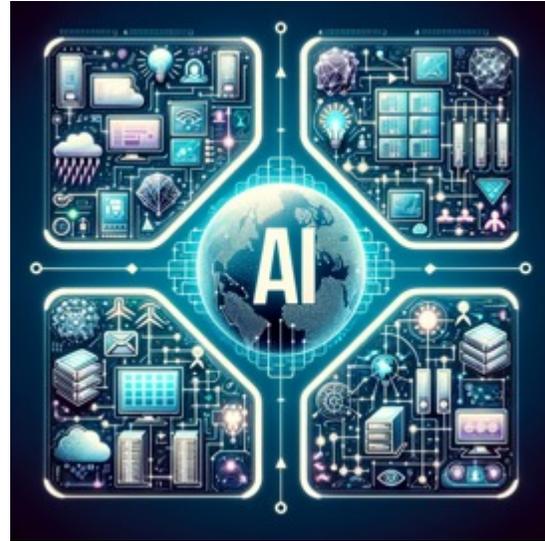
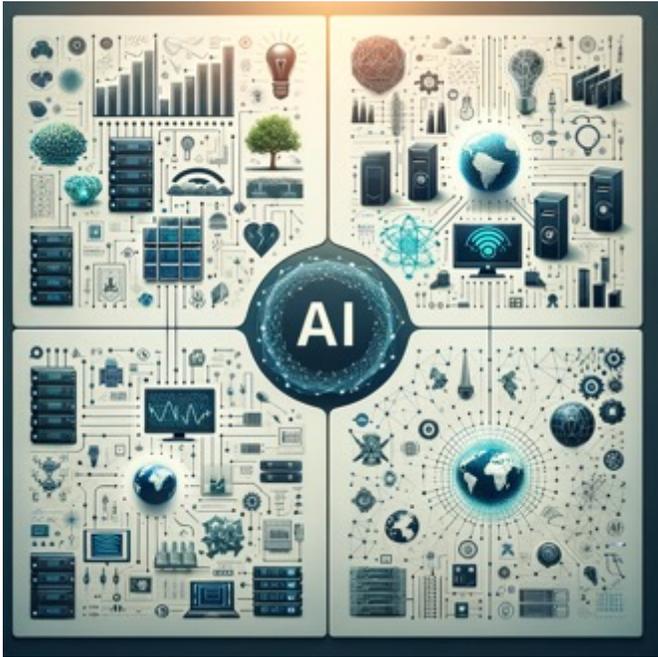
NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Today's Agenda – Hidden Costs of AI

- Prompt Engineering – Erik
- AI's Rapid Evolution – Mishali
- AI Infrastructure – Justin
- Sustainability – David
- Fabric Requirements – Rob
- Q&A



Introduction



Images Generated by ChatGPT 4 (DALL-E 3)

Introduction

ChatGPT 4 ▾



You

Create a image



ChatGPT

Sure, I can create an image for you. Could you please provide a detailed description of what you'd like the image to depict?

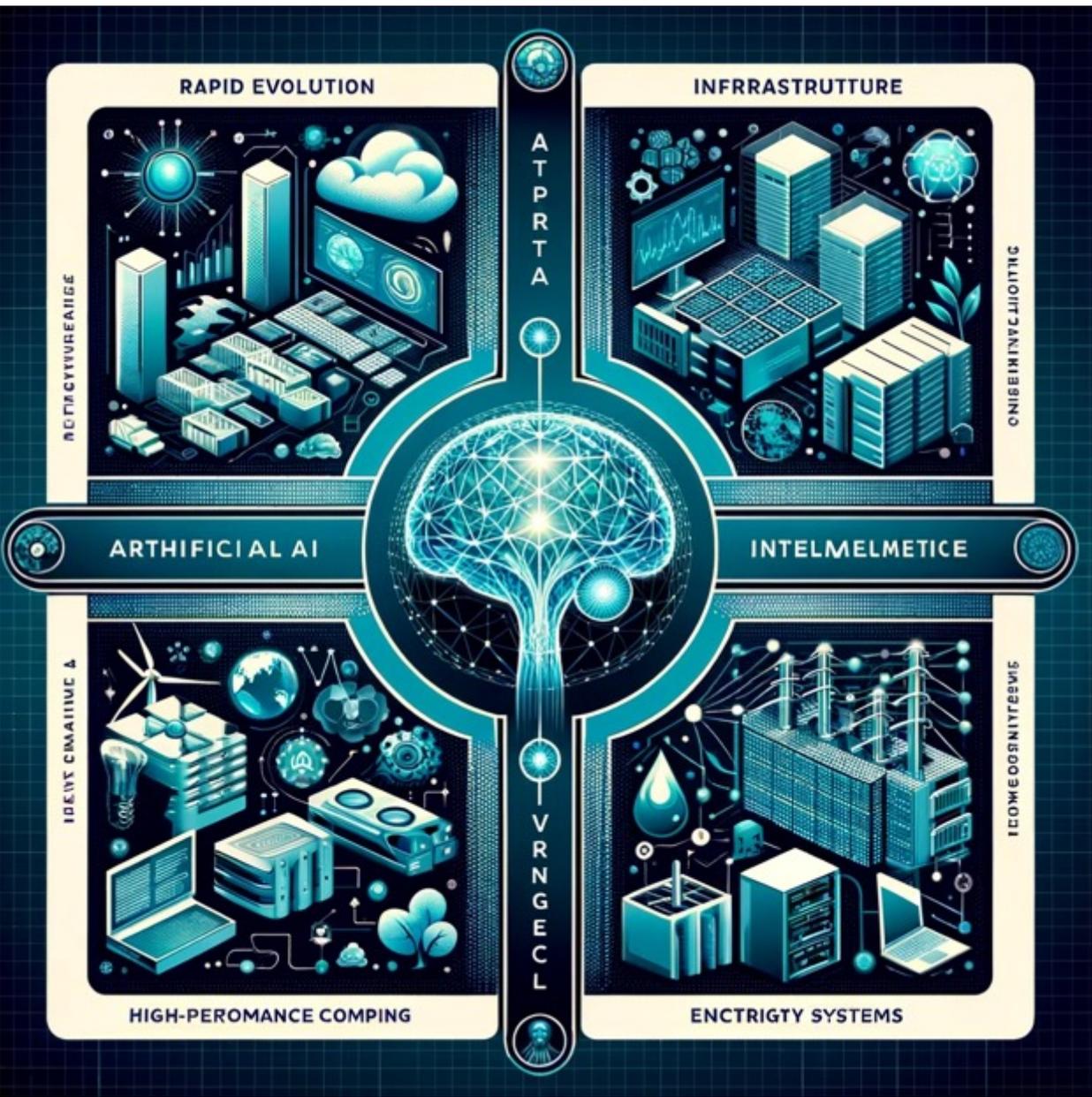
Introduction



You

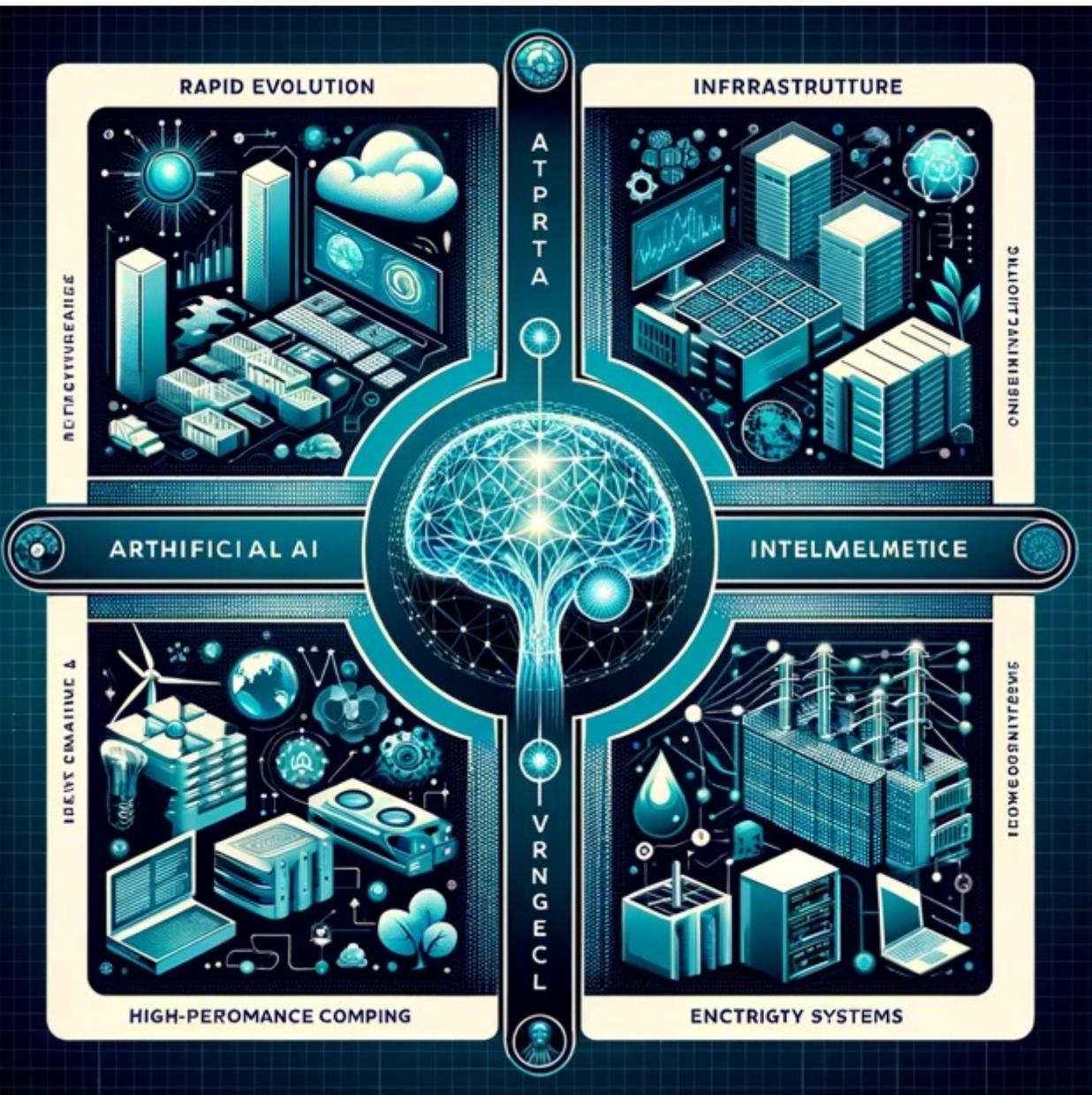
I would like an image that can be embedded into a PowerPoint presentation. The image should consist of four quadrants with each quadrant representing four different aspects of Artificial Intelligence. The four aspects are:

1. The rapid evolution of AI
2. The Infrastructure required to support AI model training
3. The Environmental impact of AI especially with regards to use of electricity
4. The Network required to connect AI infrastructure together



ChatGPT

Here is the image designed for your PowerPoint presentation. It's divided into four quadrants, each visually representing a different aspect of Artificial Intelligence: the rapid evolution of AI, the infrastructure required for AI model training, the environmental impact of AI, and the network necessary to connect AI infrastructure. This design should suit a professional presentation context well.



You

Please modify the image and ensure only English is used



ChatGPT

The image I created does not contain any text; it's purely visual. Each quadrant visually represents a different aspect of Artificial Intelligence: the rapid evolution of AI, the infrastructure required for AI model training, the environmental impact of AI, and the network necessary to connect AI infrastructure. Since it's already language-neutral, it's suitable for English-speaking audiences or for use in presentations where text is added separately. If you have any specific modifications in mind, please let me know!



 Message ChatGPT... 

[ChatGPT can make mistakes. Consider checking important information.](#)



You

Create an image representing four aspects of Artificial Intelligence:

The rapid evolution of AI

The Infrastructure required to support AI model training

The Environmental impact of AI especially with regards to use of electricity

The Network required to connect AI infrastructure together



ChatGPT

Here is the image, conceptualizing the four aspects of Artificial Intelligence for your presentation. Each quadrant of the image visually represents the rapid evolution of AI, the infrastructure required for AI model training, the environmental impact of AI, and the network necessary to connect AI infrastructure. The design is modern and professional, suitable for academic or professional settings, and contains no textual elements.



You

Can you provide a few examples of a funny way to end my intro?



ChatGPT

Certainly! Here are a few humorous endings that should get a laugh:

3. "And finally, I discovered the most important lesson of all: when you ask an AI for help with a presentation, be prepared for a journey into the unknown. It's a bit like asking a cat to fetch; you're both going to be confused, but only one of you is going to end up with a PowerPoint full of gibberish. So, let this be a lesson to us all: sometimes, the only thing artificial intelligence guarantees is a good laugh."

Today's Agenda – Hidden Costs of AI

- **Prompt Engineering** Arguing with AI – Erik ✓
- AI's Rapid Evolution – Mishali
- AI Infrastructure – Justin
- Sustainability – David
- Fabric Requirements – Rob
- Q&A



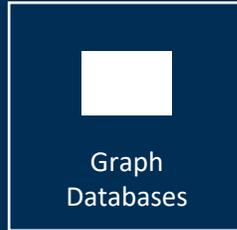


AI's Rapid Evolution

Mishali Naik

AI is Evolving Rapidly

Underlying data technologies....



60%

of data for AI projects will be synthetic by 2024

58%

of CEOs from leading public companies actively investing in AI

More than

50%

of enterprise-managed data will be created & processed outside the data center or cloud

Generative AI is disrupting industries

Growth of model sizes 1T+ parameter models

Growth of smaller model types

AI everywhere

By 2026

50%

of edge deployments will involve AI

Inferencing drives compute costs

Responsible AI

<https://chiefexecutive.net/the-rise-of-the-ai-ceo/>
https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/
Gartner®, Hyperscalers Stretching to the Digital Edge, July 2023. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All right reserved.
By 2026, at least 50% of edge computing deployments will involve machine learning (ML), compared to 5% in 2022. (Building an Edge Computing Strategy, April 2023)

Specialized AI Models

Large Foundational Model

Advantages

- + Incredible all-in-one, out-of-the-box versatility: text, programming, continual natural language conversation and plain summarization
- + Surprisingly, compelling outcomes

Challenges

- Big (>100B parameters), expensive- \$4m+ to train, \$3m per month for inferencing
- Hallucinations; lack of explainability, intellectual property issues
- Frozen in time (sampling)

Domain Specific Models

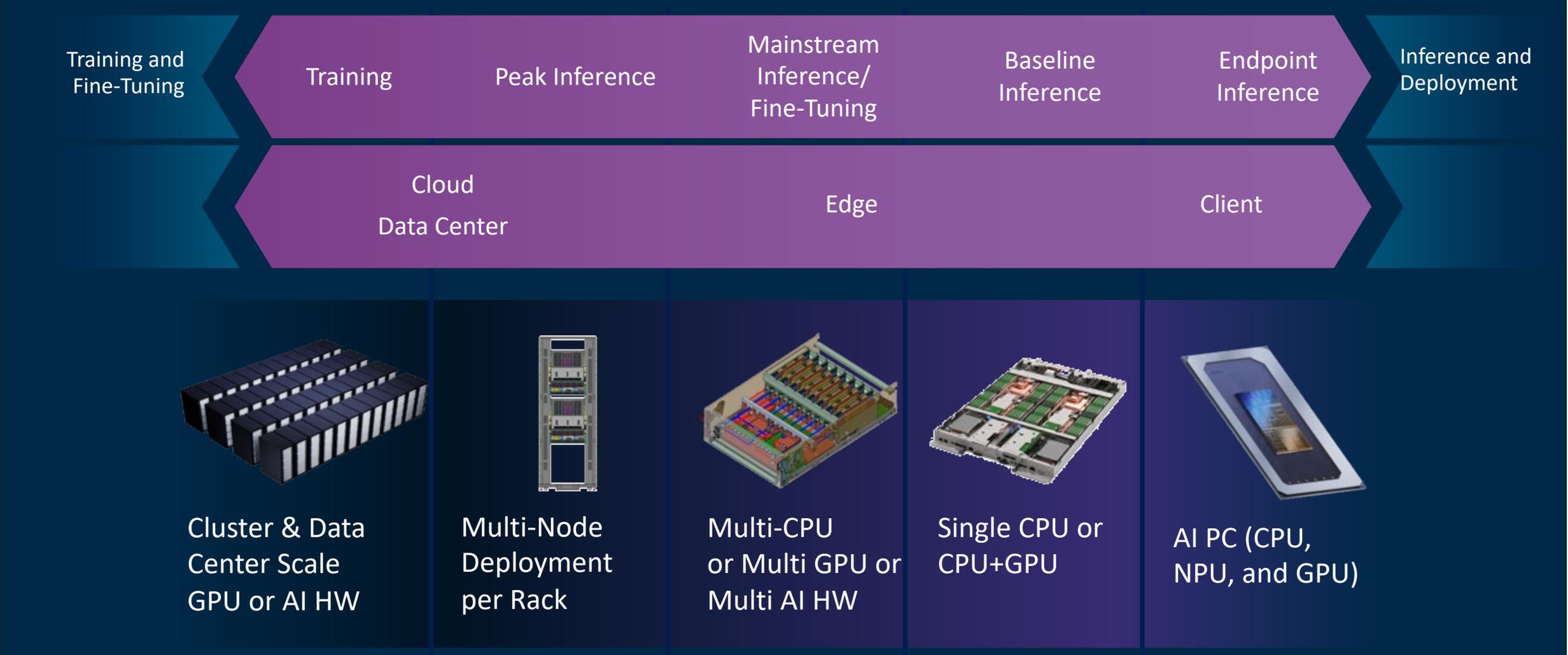
Advantages

- + 10-100x smaller models while maintaining/improving accuracy
- + Economical on general-purpose compute
- + Correctness; Source attribution; Explainability
- + Utilizing private/enterprise data
- + Continuously updated information

Challenges

- Reduced range of tasks
- Requires few-shot fine-tuning and indexing

Scalable Systems for Simple AI Infrastructures





AI Infrastructure

Justin Potuznik

AI Infrastructure- Just Another Workload?

- **Ingredients of AI infrastructure:**

- **Accelerated compute**
 - I.E. GPUs, FPGAs, coprocessors, DPUs, etc.
- **Networks with low latency and high bandwidth, often dedicated**
- **Local and attached storage capable of meeting changing demands**
 - Training & finetuning requires high bandwidth to keep gpu utilization high
 - Inference requires less bandwidth but low latency I/O
- **New software stacks for infra & end users**
- **Different infrastructure configurations needed for optimal performance of Inference, finetuning, and training.**

- **AI Infrastructure is much more resource intensive than many traditional workloads on the entire infrastructure stack**
- **Optimal AI Infrastructure is currently a moving target due to diverging priorities**
- **Three paths are being pursued by different AI models simultaneously:**
 - AI model efficiency in terms of memory use as well as speed and density e.g. how do I do what I already can with less or pack in more users to the same infrastructure?
 - Larger AI models capable of greater quality and higher accuracy
 - Adding new capabilities, currently the hottest models are LLMs that process text to text but text to image or video are coming and will have different requirements on the infrastructure (storage performance requirements will increase as file size increases for img:vid for vid:vid especially)

AI Infrastructure Technology Trends



Large AI Training + Inference adoption

End user adoption of AI is exploding and businesses are looking for ways to use AI in everything



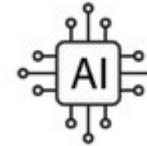
Silicon Diversity & Increasing Power Demands

There are many new options for CPUs, GPUs, and Accelerators than what has historically been available. Power increases across all chips is leading to adoption of Liquid Cooling



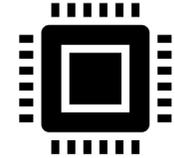
Optimizing LLM Operational Flow

The rapid growth and availability of data coupled with more AI projects has driven a need for optimized data and model operation techniques



Finetuning & RAG

Technical and resource challenges around training models is driving more adoption of open source models + finetuning & techniques such as RAG



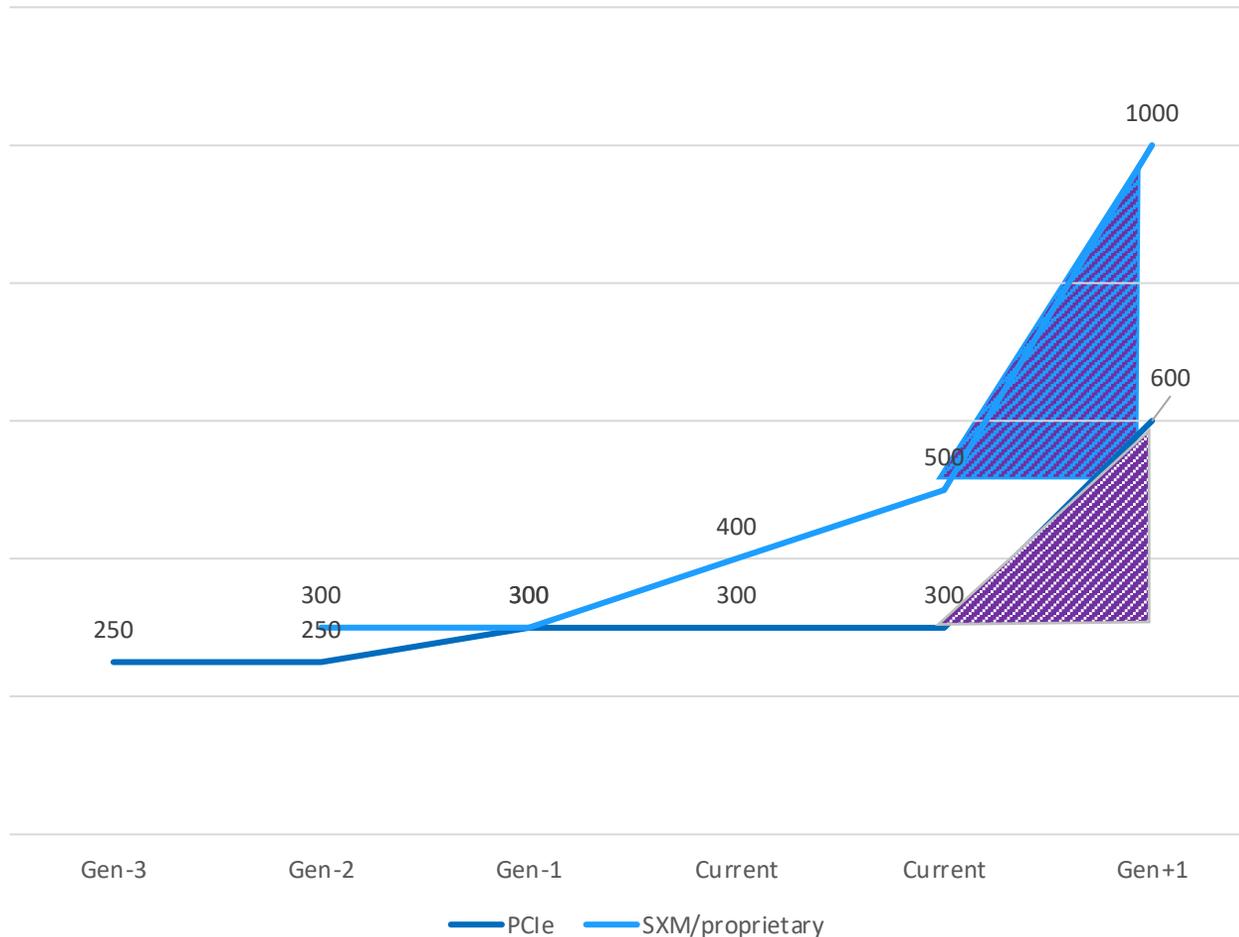
Open, Transparent & Standard

Businesses desire and are adopting Open source software, permissive model licenses, transparent data usage and industry standardization

Rising GPU Power Trends Impact Solution Design

Performance drives up consumption and cooling

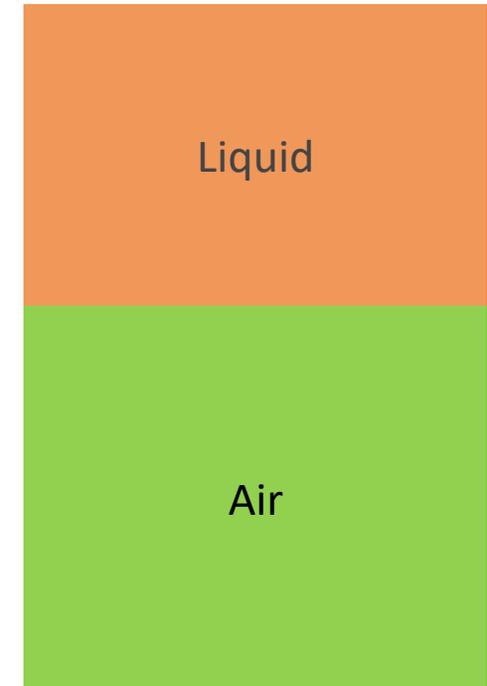
PROJECTED GPU Power Trends



GPU Form factor

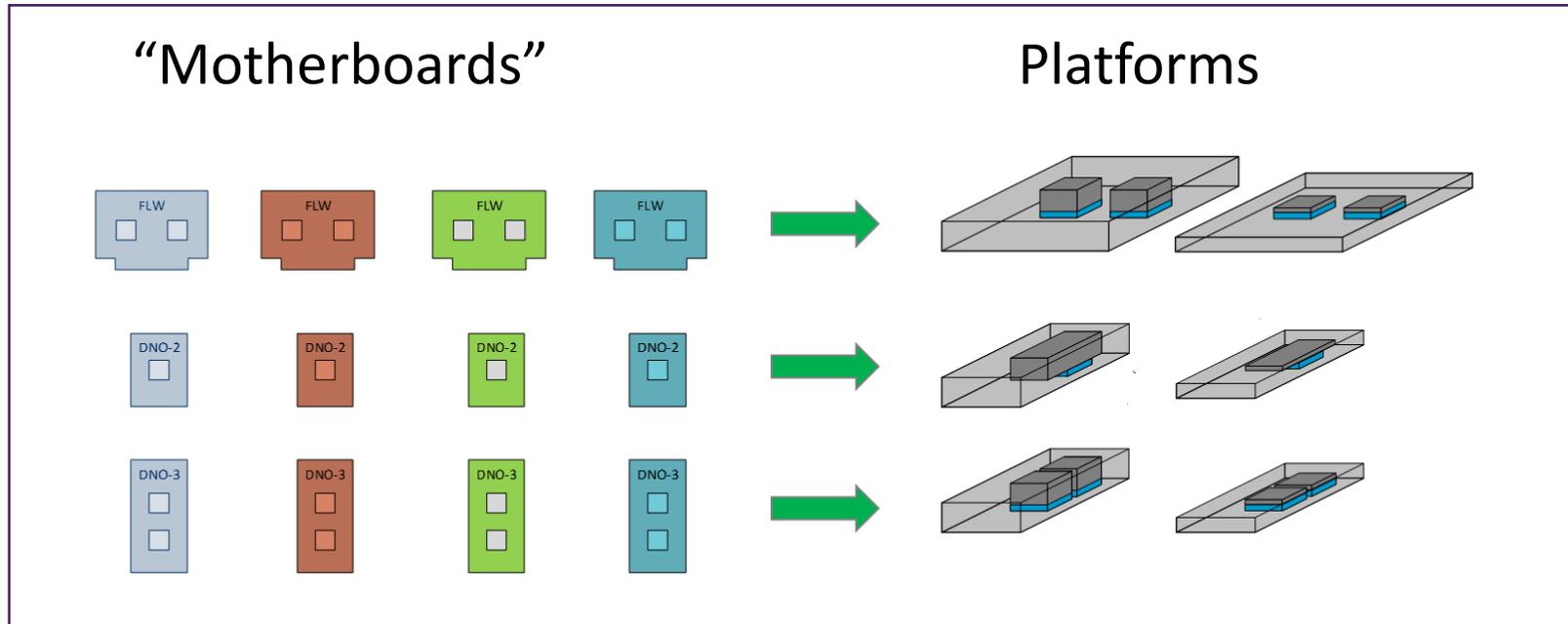


Cooling requirements



OCP's DC-MHS: Silicon Diversity w/Standardization

Objective: Standardize form-factors, connectivity, and subsystems to enable high interchangeability and flexibility with platforms.



Value-props

1. High HW consistency CPU to CPU
2. Improved economy of scale: validation & sourcing
3. Upgrade in place, gen over gen
4. Faster intercept of new technologies

Training is Lengthy & Complicated

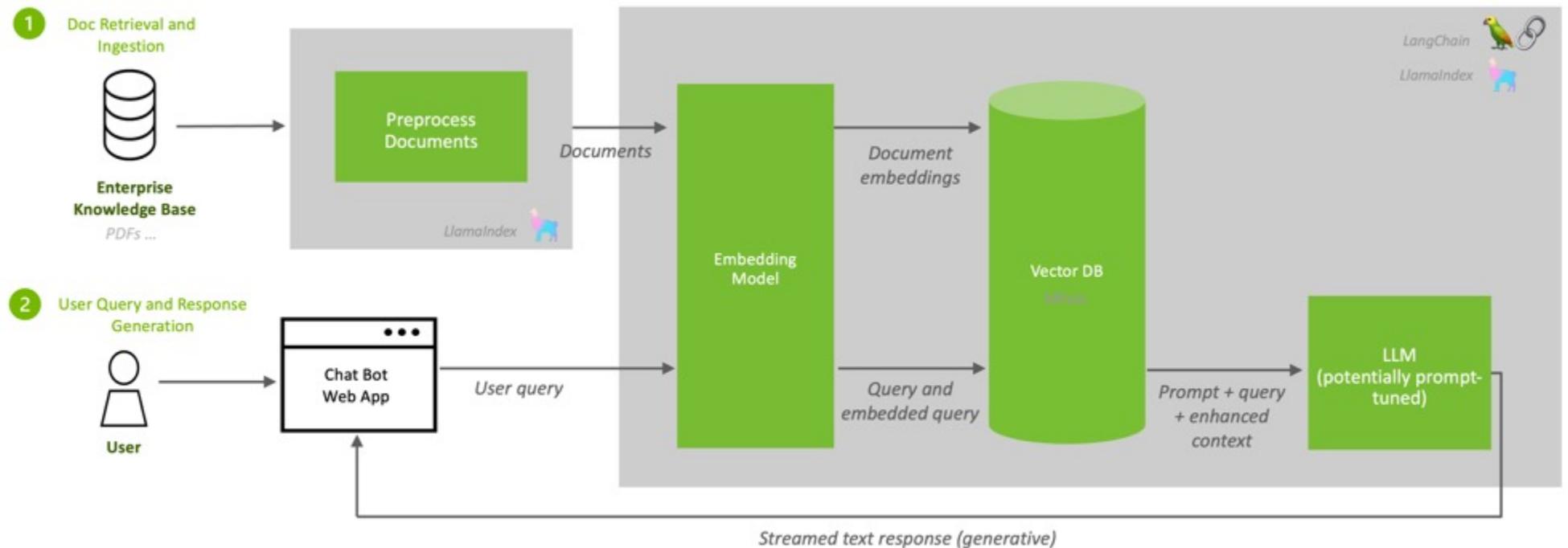
- It takes 128 current generation GPUs (16 servers) 24 days to Train a 70B parameter model
 - This is approximately 100.5 mWh of power usage over the 24 days (174 KW/h sustained.)
- It takes 8 current generation GPUs (1 Server) 5 days to Fine Tune a 70B parameter model
 - This is approximately 1308kWh of power usage over the 5 days (10.9KW/h sustained.)

Finetuning has Many Benefits

- Finetuning allows you to focus on your company data, not general training data + your company data
 - Total infrastructure needed is significantly less
 - Training requires more robust infrastructure across compute, storage, and network
 - Dedicated storage and multiple networks are normal for training clusters
 - Storage use is dramatically less as you only need data you already have
 - Time to value is decreased
 - Staff with AI training capability are rare and hard + costly to acquire

RAG LLMs Provide Quick Time to Value

- Retrieval Augmented Generation (RAG) is an architecture solution to quickly and accurately ingest and serve content with a familiar search box UI



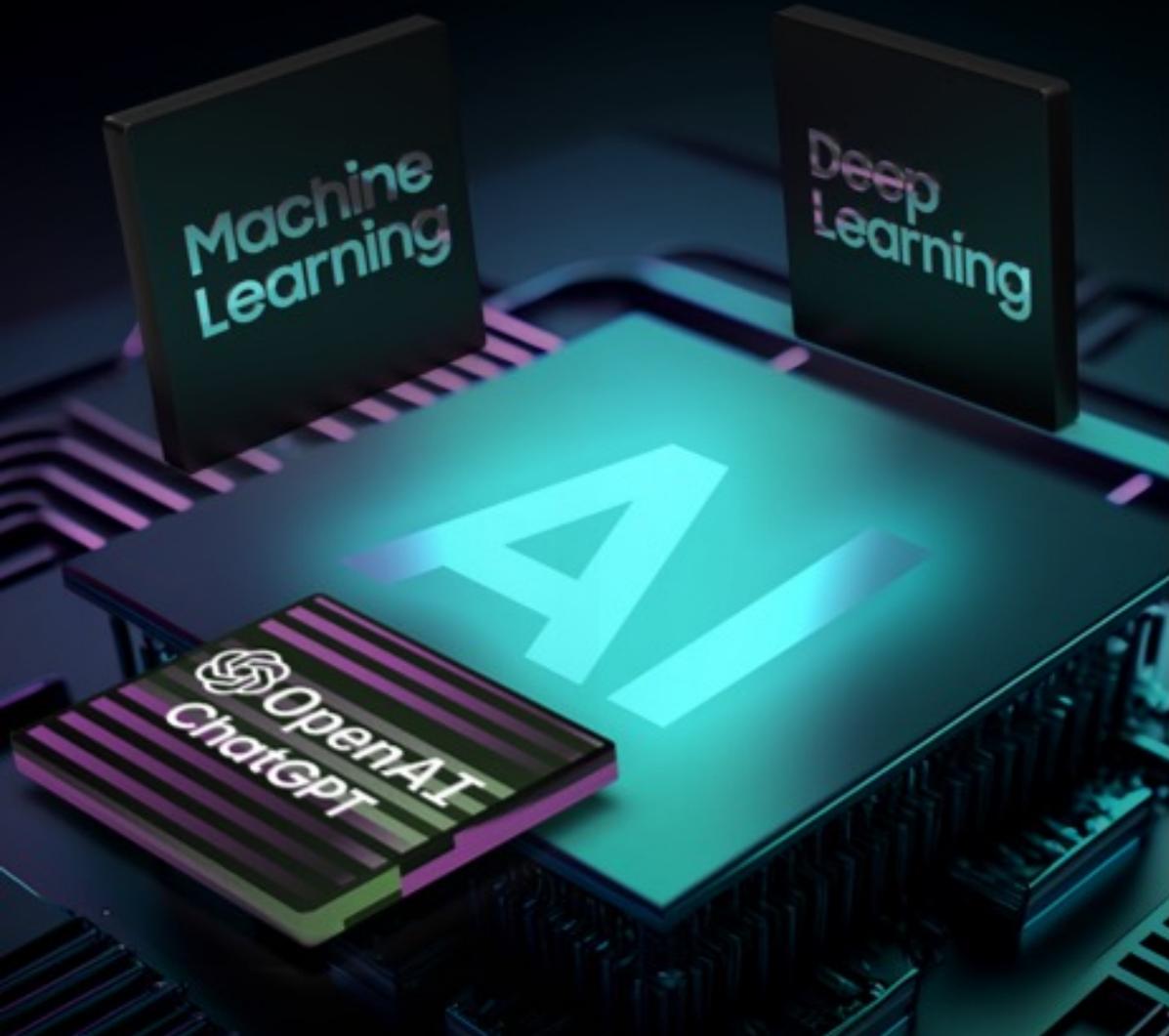


Power Efficiencies with Components

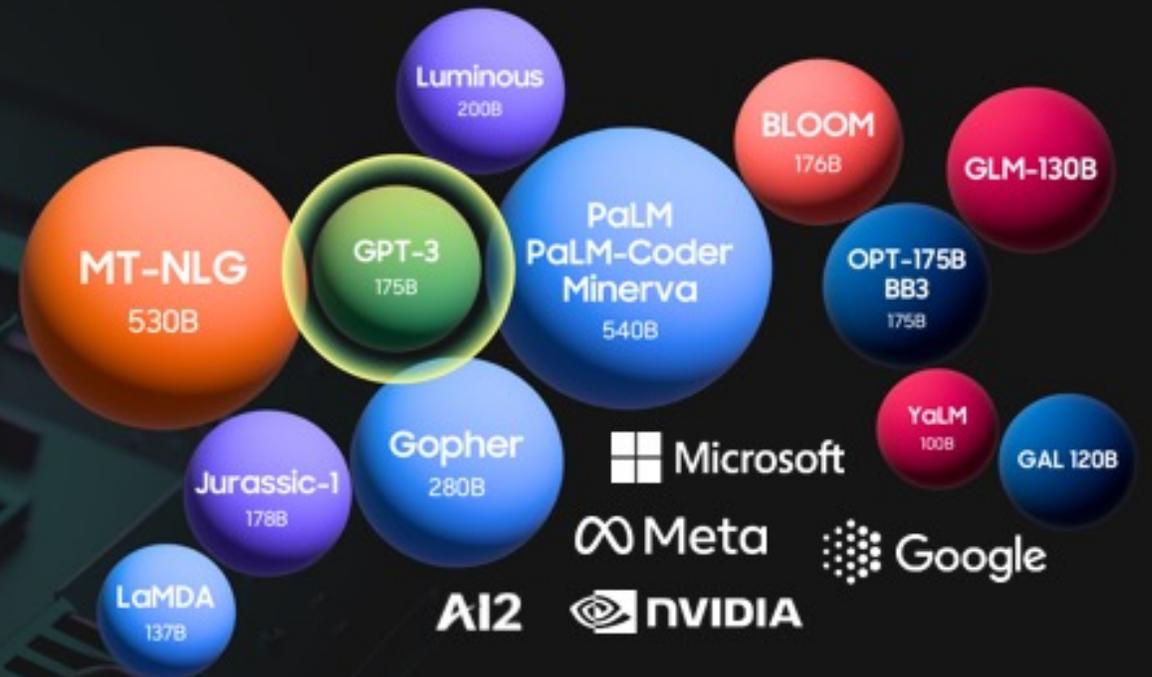
David McIntyre

In the Era of AI & ML

Swift increase in demand for capacity and performance



Language Model Sizes to Dec 2022



*Number of parameters/ Source: <https://lifearchitct.ai/models>

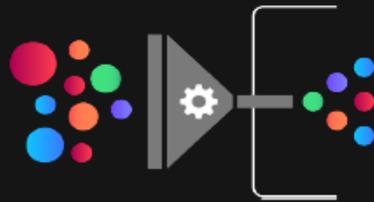
Data-Centric Computing Benefits

Power-optimized scalable processing for large data

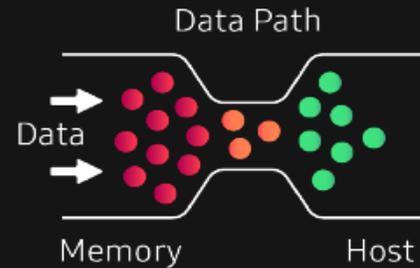
➤ *System to Component Performance with Power Efficiency in Mind*



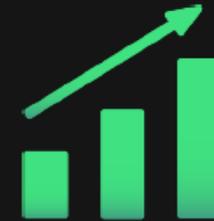
**Low Power
Computing**



**Data
Reduction**



**High Effective
Bandwidth**



**Scalable
Computing**

➤ *Component-level power consumption also scales*

Low-Power Products (Flash Memory)

Next-generation SSD Development Priorities

- System architecture design with power consumption in mind
- Alignment of SSD technology application requirements
- Optimization of component design and selection

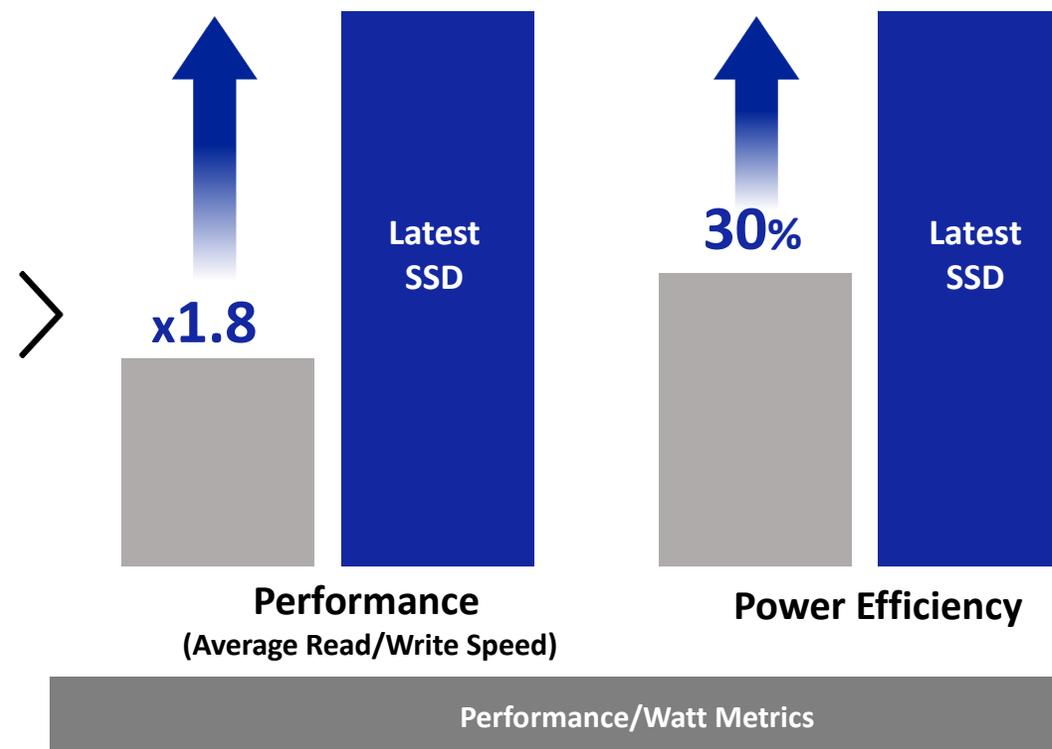


- Advanced process Controller
- Convert DDR4 → LPDDR4
- High Density NAND (512Gb → 1Tb)



- Automatically run low power modes
- Software optimization

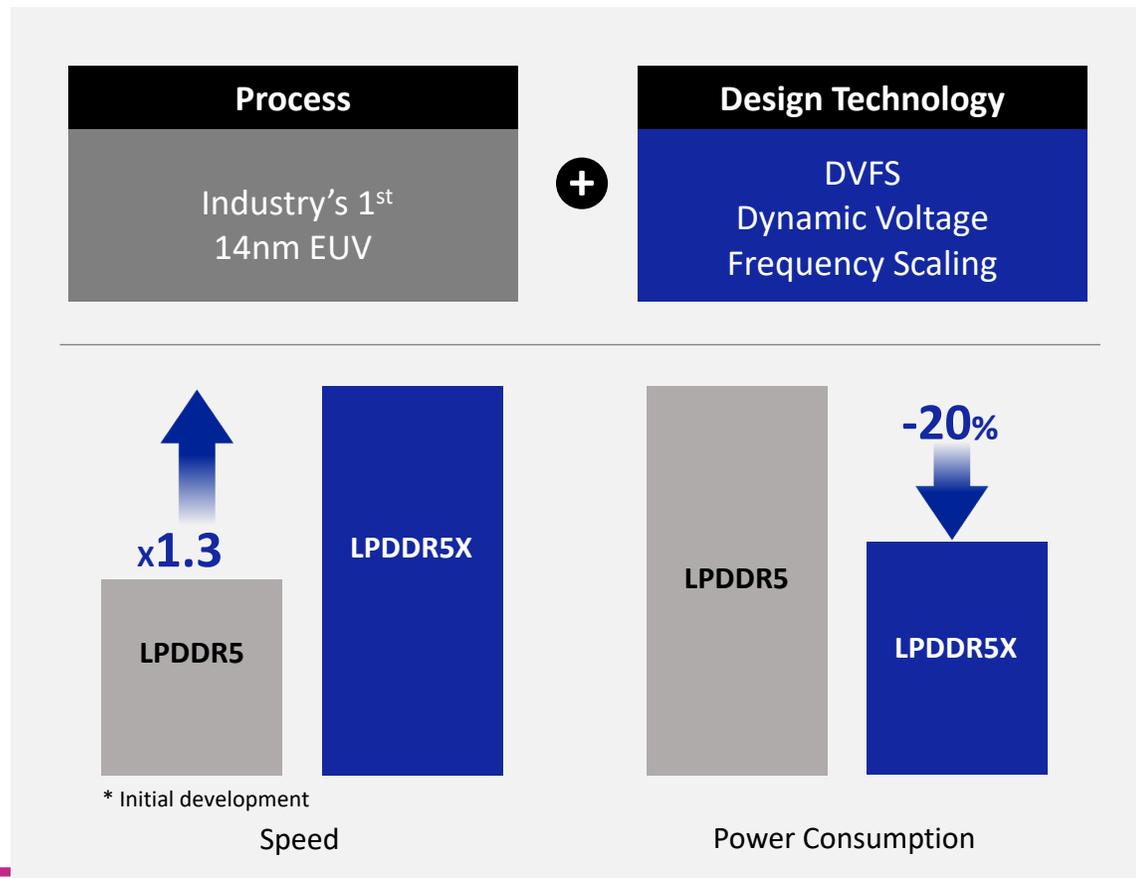
Balancing Performance with Power Efficiency



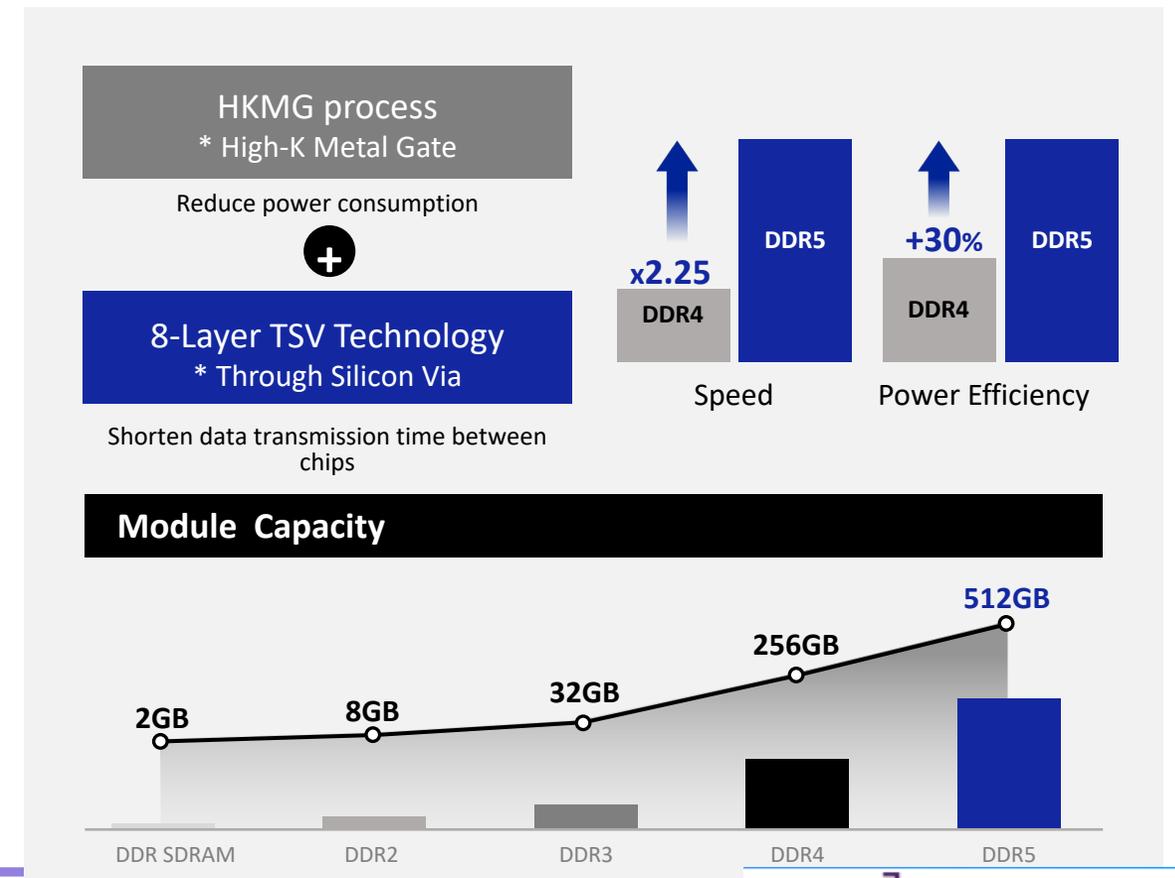
Low-Power Products (DRAM)

- Maximizing energy savings with next generation semiconductor design and manufacturing processes

LPDDR5X



DDR5





Fabric Requirements

Rob Davis

One Problem with Technology is That it Doesn't Get Faster at the Same Time Everywhere



For Example:



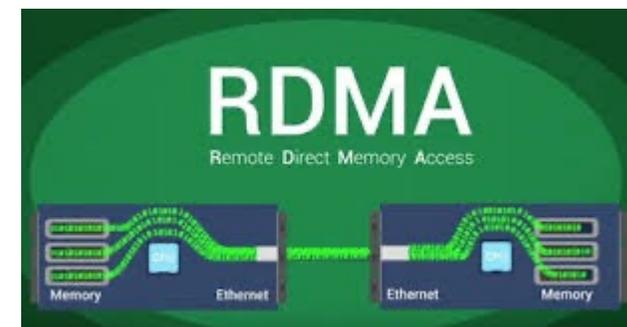
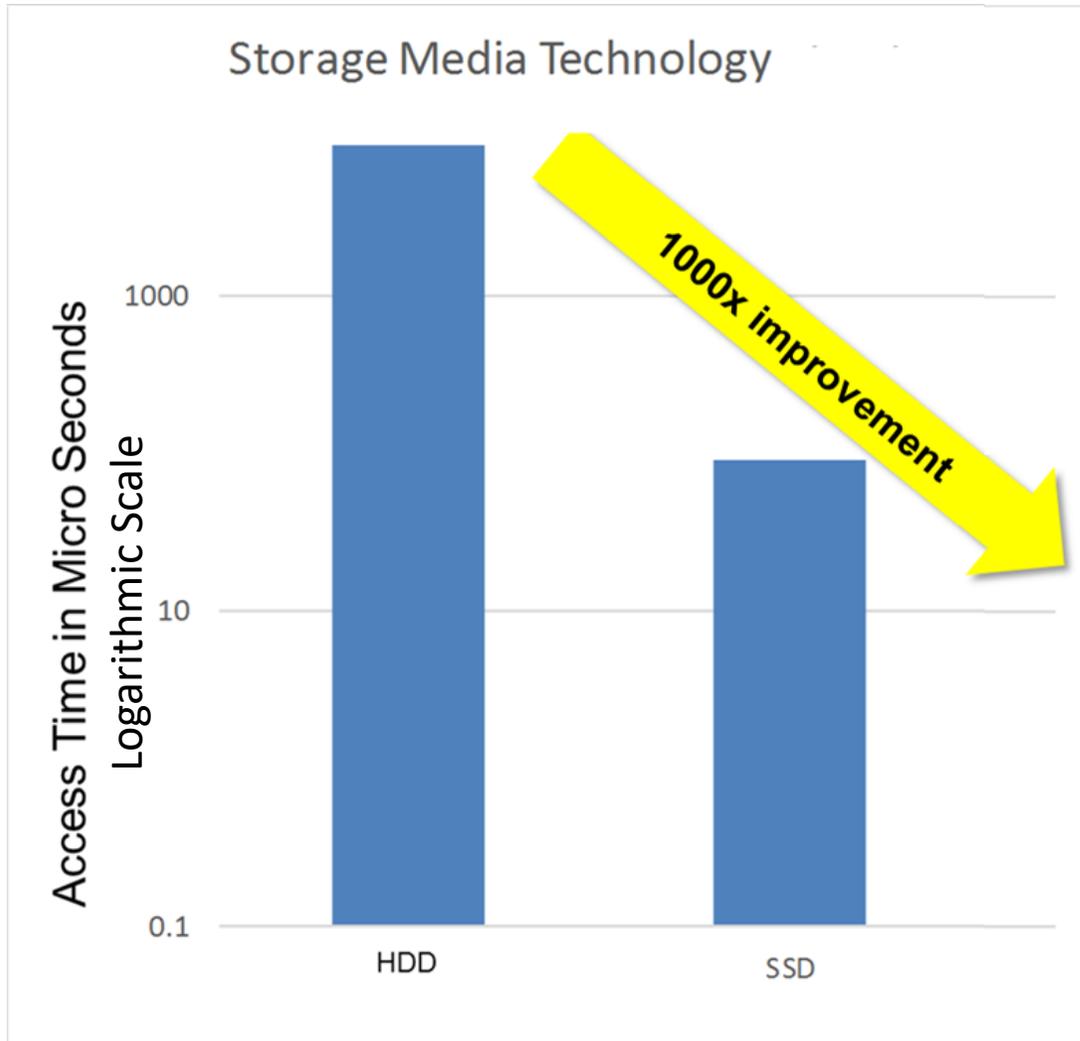
HDD

+



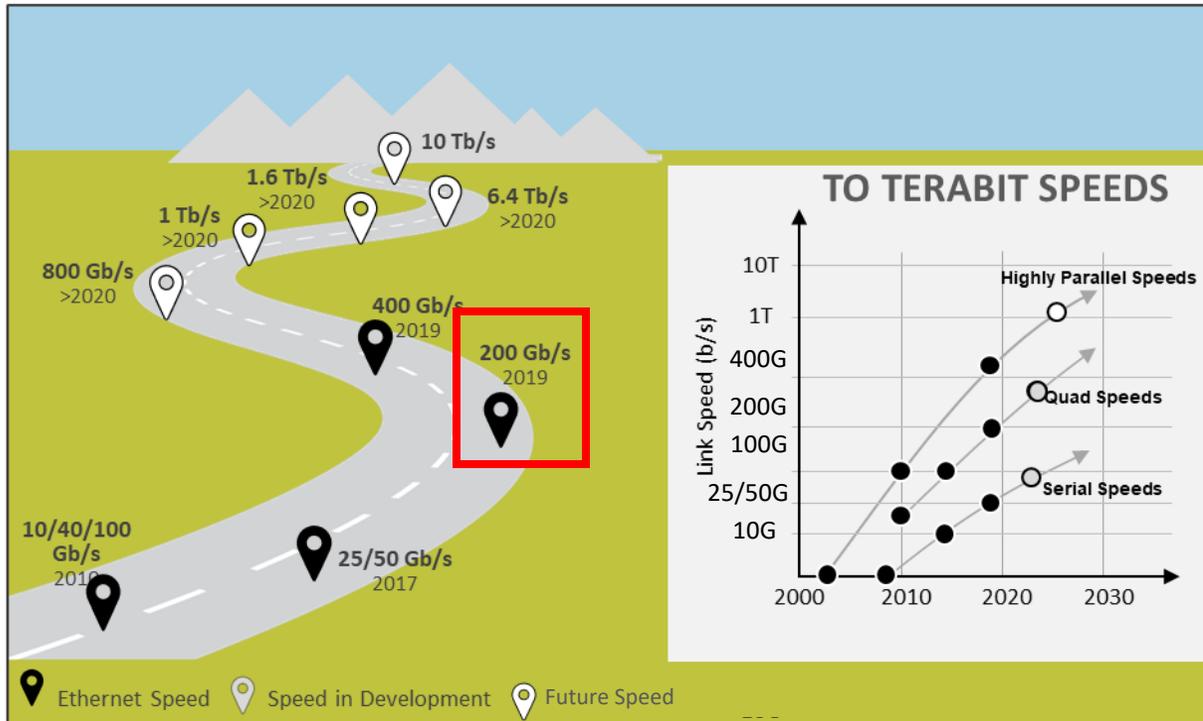
SSD

Faster Storage Needs Faster Protocols

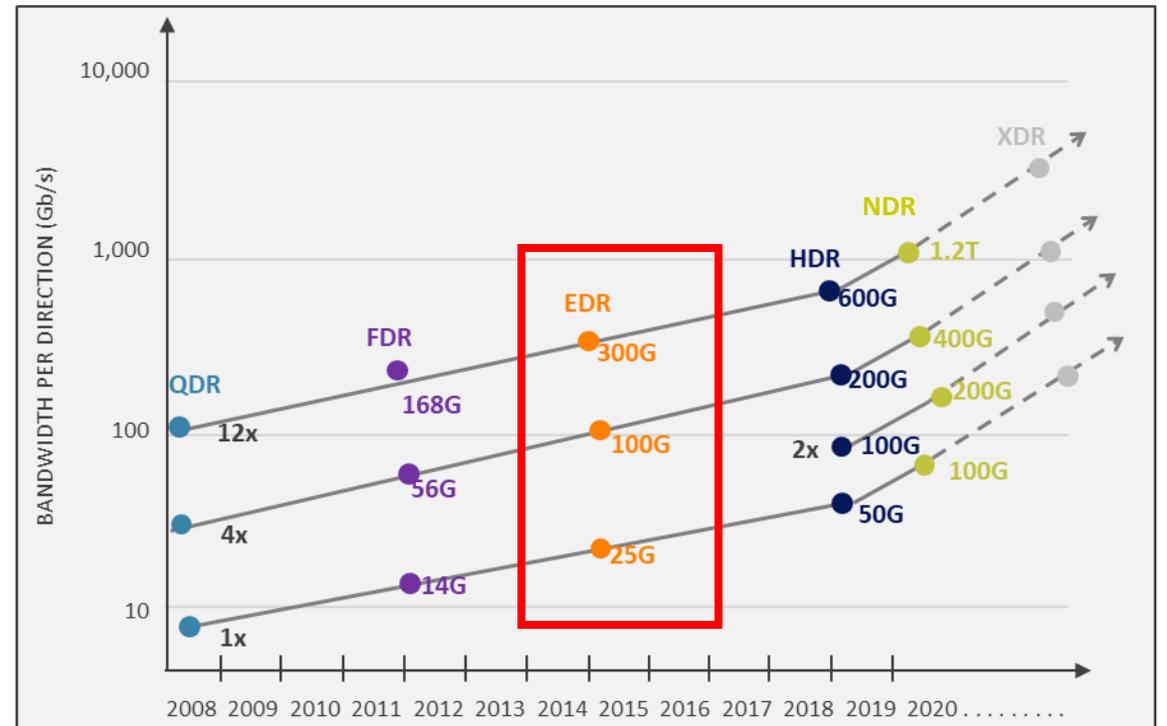


Faster Storage Needs a Faster Network

Ethernet



InfiniBand

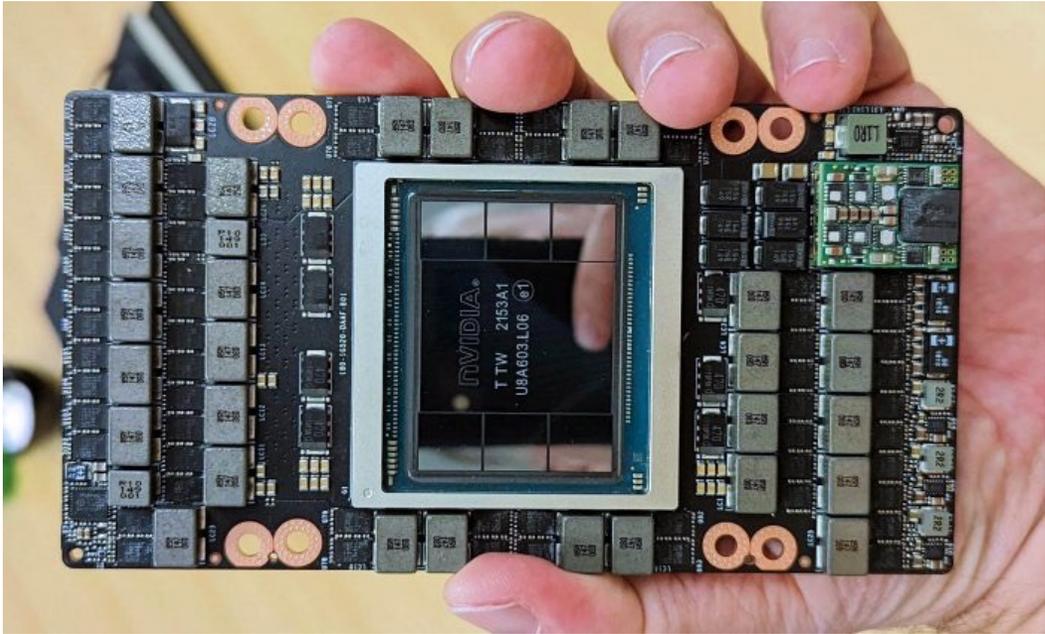


History Seems to be Repeating Itself

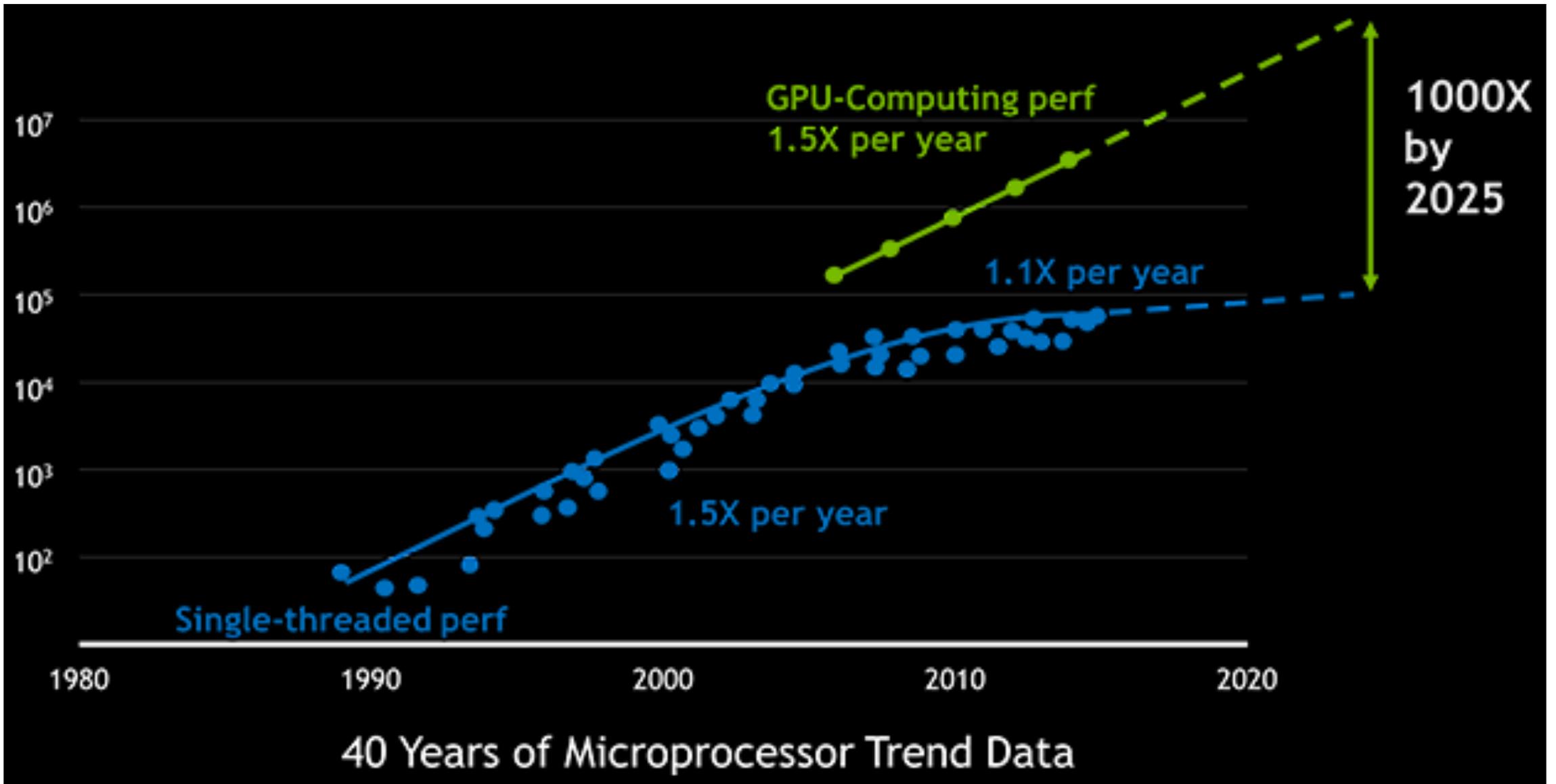


CPU

+



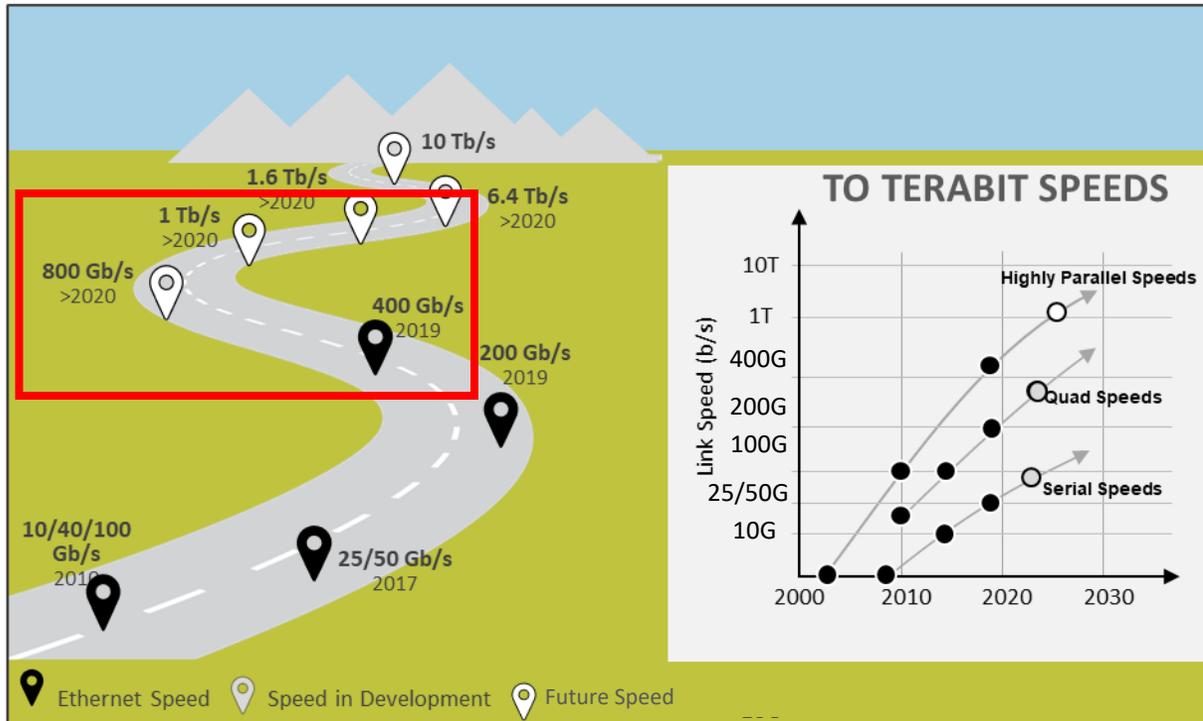
GPU



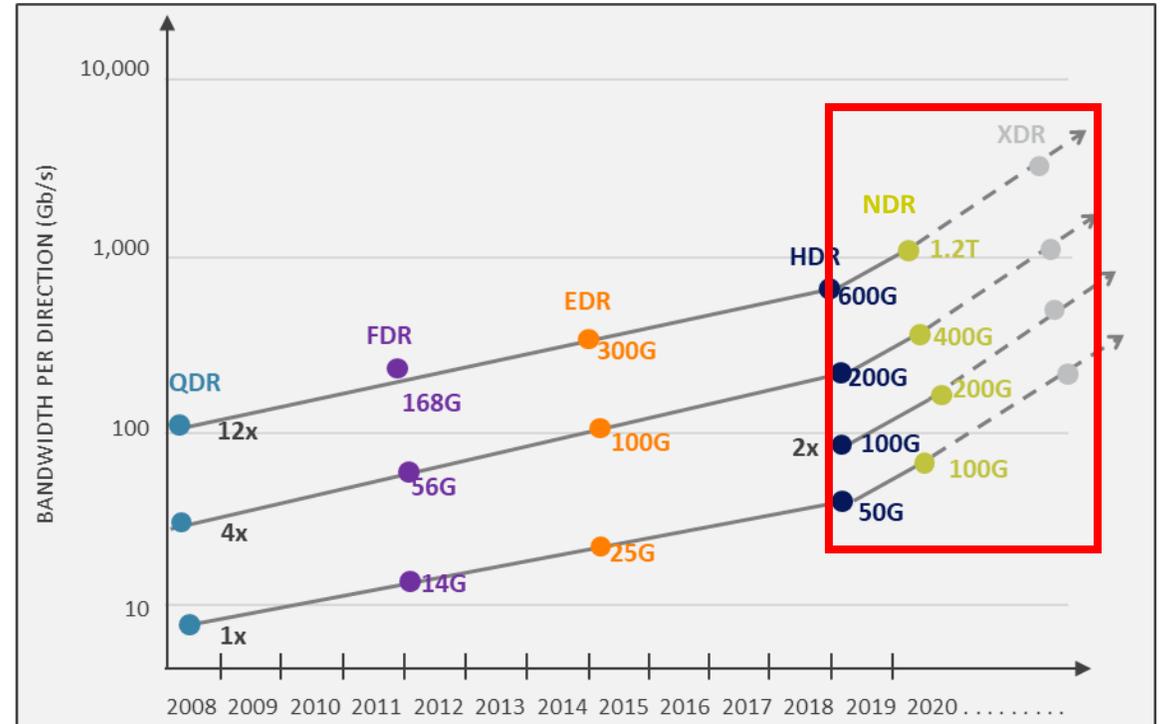
40 Years of Microprocessor Trend Data

Faster Compute Needs an Even Faster Network

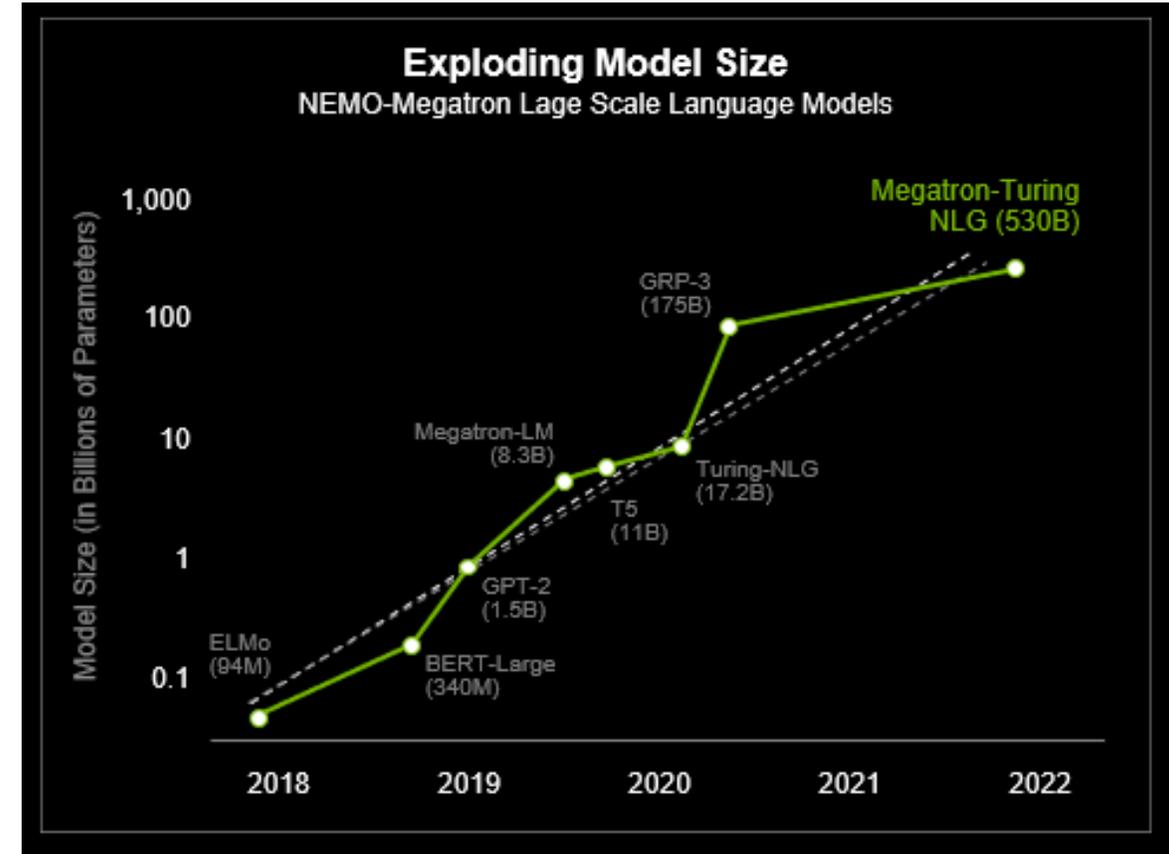
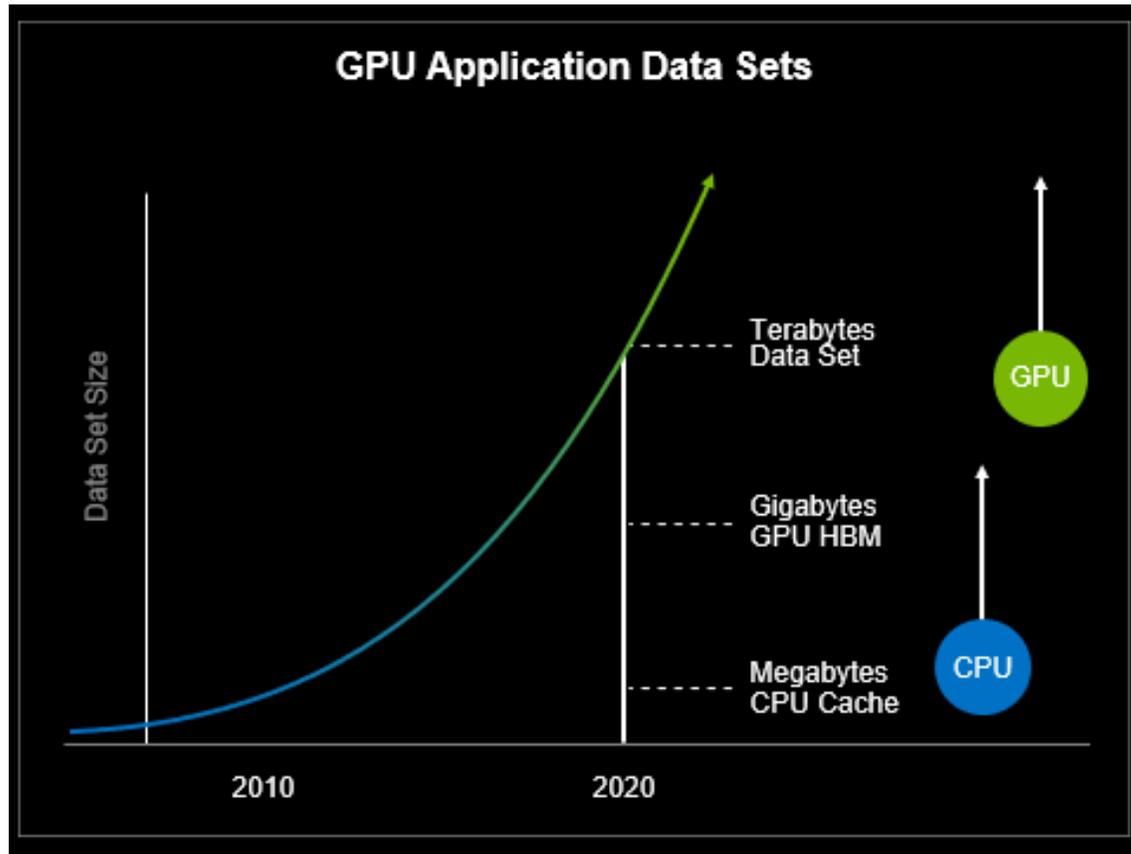
Ethernet



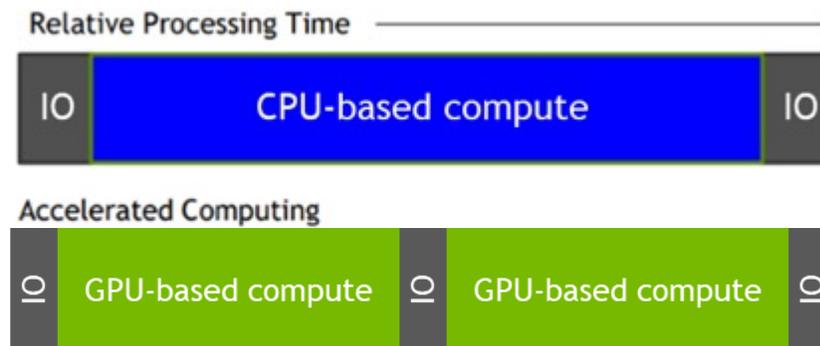
InfiniBand



GPUs Applications Are Hungry for DATA

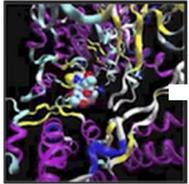


The Goal is to Keep the GPUs Fully Utilized



GPU Optimized Storage Stack

Running AI Storage Workloads on Traditional Ethernet



AI Workload



Significant
Congestion

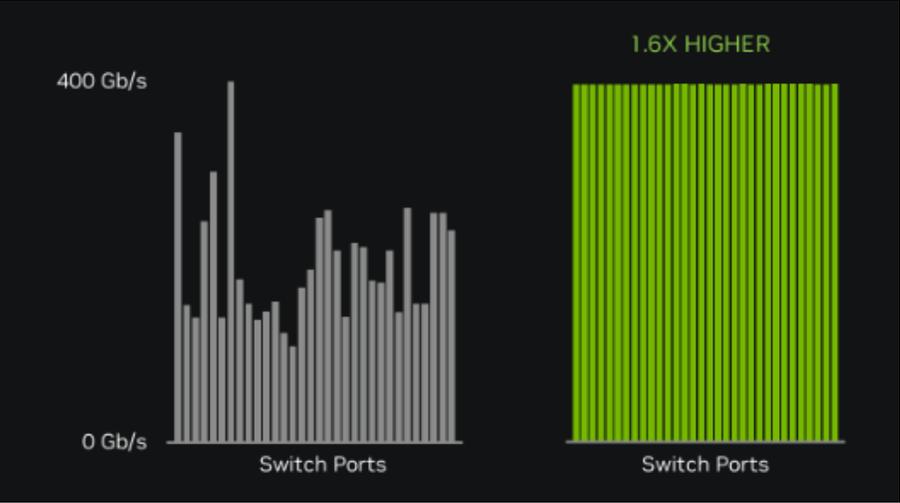


Increased
Latency
Tail Latency

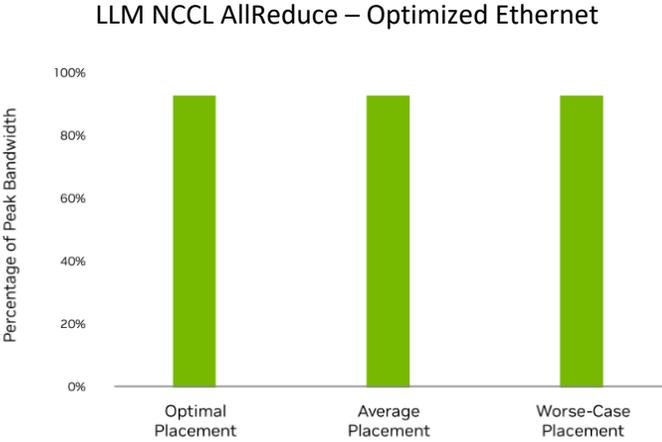
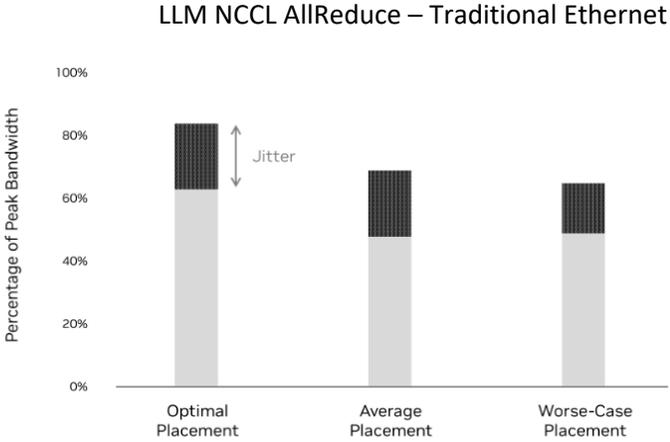


Bandwidth
Unfairness

Results using AI Optimized Ethernet Network



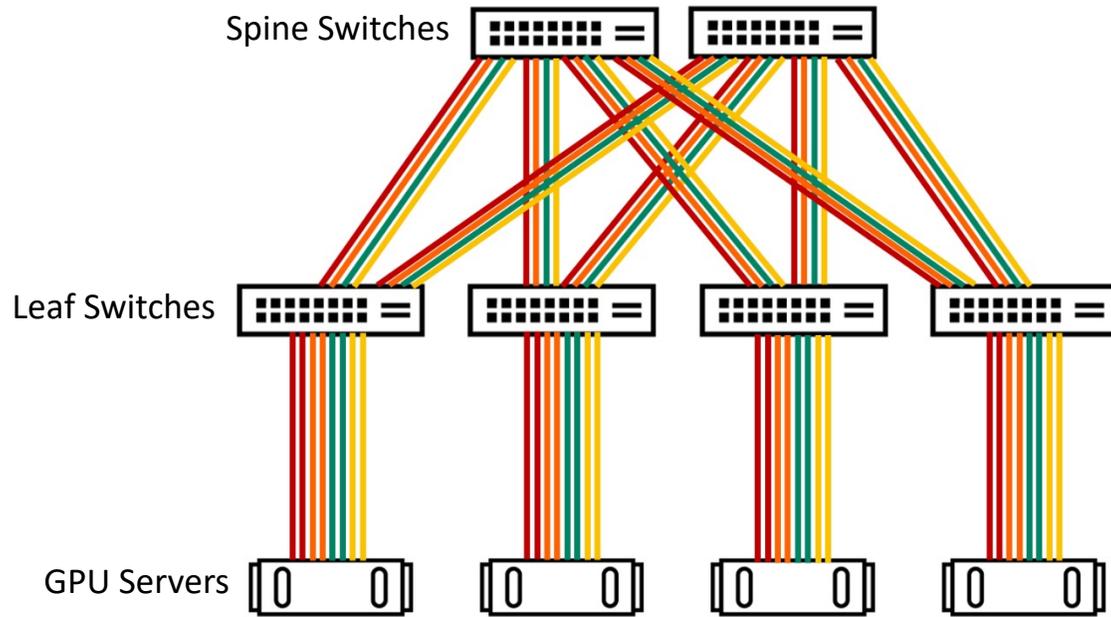
Increased Network Bandwidth Efficiency



Reduce Network Latency Jitter

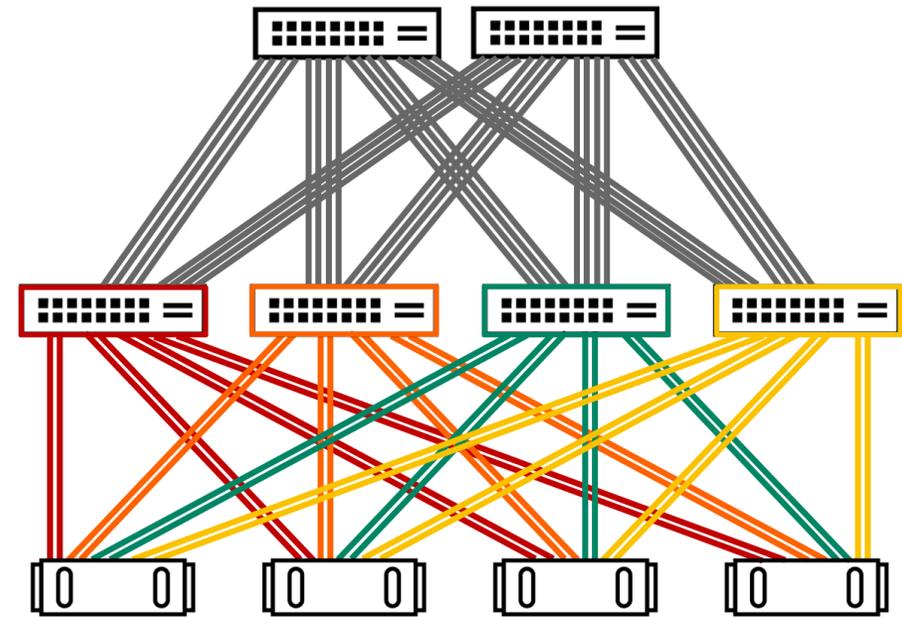
New Topology for AI: Rail-Optimized

Routing must be perfect to ensure all flows use different links.



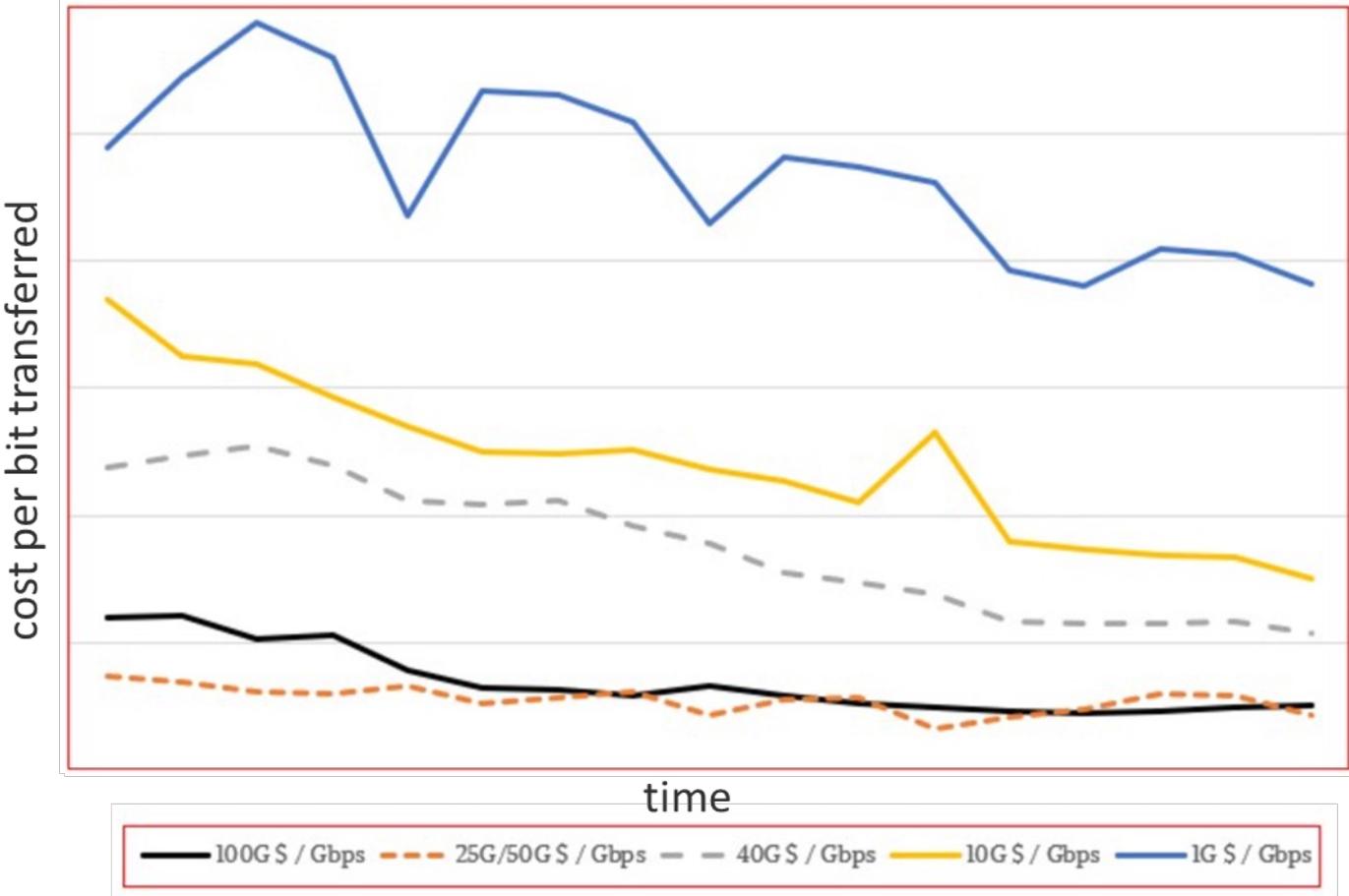
Classic fabric design

All traffic is local to leaf switches. Routing collisions are impossible.



Rail-optimized design

Measuring the Network Cost a Different Way



<https://www.nextplatform.com/2021/08/30/more-than-anything-else-cost-per-bit-drives-datacenter-ethernet/>

Summary

- Using Generative AI requires patience and attention to detail.
- AI is rapidly evolving and there are many use cases. Not all require access to GPUs.
- The infrastructure to train or fine tune a model is highly specialized, resource intensive and continuously improving.
- Optimizing the components used to create infrastructure will reduce power consumption and time to train.
- The network needed to interconnect your AI infrastructure far exceeds anything enterprise environments have seen to date.

Q&A

After this Webinar

- Please rate this webinar and provide us with your feedback
- This webinar and a copy of the slides are available at the SNIA Educational Library <https://www.snia.org/educational-library>
- A Q&A from this webinar, including answers to questions we couldn't get to today, will be posted on our blog at <https://sniansfblog.org/>
- Follow us [@SNIANSF](https://twitter.com/SNIANSF)

Thank You