SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

Training Deep Learning Models in the Cloud

Live Webcast

March 15, 2023

10:00 am PT / 1:00 pm ET

Today's Presenters

Erin Farr
Vice Chair SNIA Cloud Storage Technologies Initiative
IBM

Milind Pandit
Habana Labs, an Intel Company

Seetharami Seelam
IBM Research

 © SNIA. All Rights Reserved.

SNIA - By the Numbers

Industry Leading Organizations

180

Active Contributing Members

2,500

IT End Users & Storage Pros Worldwide

50,000

 © SNIA. All Rights Reserved.

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

What We Do

Educate vendors and users on cloud storage, data services and orchestration

Support & promote business models and architectures: OpenStack, Software Defined Storage, Kubernetes, Object Storage

Understand Hyperscaler requirements Incorporate them into standards and programs

Collaborate with other industry associations

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.

 © SNIA. All Rights Reserved.

Today's Topics

- Industry Trends in AI
- Examples of AI Adoption and Benefits
- Considerations for Adopting AI Technologies
  - Scale Out Infrastructure
  - Unified Platform for Training and Inference
  - Middleware Stack and Tools

 © SNIA. All Rights Reserved.

Industry Trends

AI model compute doubling every 10 months

 © SNIA. All Rights Reserved.

WSJ | OPINION

Subscribe | Sign In

Home World U.S. Politics Economy Business Tech Markets Opinion Books & Arts Real Estate Life & Work Style Sports

Search

OPINION | COMMENTARY Follow

ChatGPT Heralds an Intellectual Revolution

Generative artificial intelligence presents a philosophical and practical challenge on a scale not experienced since the start of the Enlightenment.

By Henry Kissinger, Eric Schmidt and Daniel Huttenlocher
Feb. 24, 2023 2:17 pm ET

Bloomberg the Company & Its Products | Bloomberg Terminal Demo Request | Bloomberg Anywhere Remote Login | Bloomberg Customer Support

Bloomberg

US Edition Sign In Subscribe

Live Now Markets Economics Industries Technology Politics Wealth Pursuits Opinion Businessweek Equality Green CityLab Crypto More

Technology

Wall Street Banks Are Cracking Down on AI-Powered ChatGPT

- Banks currently restricting chatbot's use, people familiar say
- ChatGPT has sparked intense interest across industries

By Gabriela Mello, William Shaw and Hannah Levitt
February 24, 2023 at 4:36 AM PST *Updated on February 24, 2023 at 7:10 AM PST*

MIT Technology Review

Featured Topics Newsletters Events Podcasts Sign in Subscribe

ARTIFICIAL INTELLIGENCE

The original startup behind Stable Diffusion has launched a generative AI for video

Runway's new model, called Gen-1, can change the visual style of existing videos and movies.

By Will Douglas Heaven February 6, 2023

SIGN IN / UP

The Register

AI + ML

America: AI artwork is not authored by humans, so can't be protected by copyright

23

Comic *sans* Midjourney images would be eligible for IP protections

Katyanna Quach Fri 24 Feb 2023 01:45 UTC

Images generated by Midjourney and other AI text-to-image tools are not protected by US copyright law since they "are not the product of human authorship", according to the nation's Copyright Office.

Last September, the USCO approved an application for *Zarya Of The Dawn*, a graphic novel created by Kristina Kashtanova, registering the work as intellectual property protected by copyright. The decision reinvigorated the legal debate over whether digital artwork created by the latest AI text-to-image tools could be copyrighted or not.

 © SNIA. All Rights Reserved.

Enterprises Will Increasingly Rely on Deep Learning Workloads

2021 – 2026 projections indicate

- Data center accelerator market **CAGR of 36.7%**
- Deep learning driving the growth
- 1/3 of servers shipped in 2026 will run DL training or inference
- DL to account for majority of cloud workloads
- **Training applications** to be the majority of the server apps by 2026

https://www.businesswire.com/news/home/20210819005361/en/Global-Data-Center-Accelerator-Market-Forecast-to-2026-Artificial-Intelligence-to-Drive-the-Growth-of-Cloud-Data-Center-Market---ResearchAndMarkets.com

 © SNIA. All Rights Reserved.

Explosive Demand for Deep Learning Training

- More applications for AI
- More complex models
- Many iterations:

74% of IDC respondents indicate running 5-10 iterations of training

50% of respondents rebuild models weekly or more often;

26% rebuild daily or hourly

Source: IDC Semiannual Artificial Intelligence Tracker (2020H1, published Jan 2021)

 © SNIA. All Rights Reserved.

Deep Learning Acceleration

Study cited in State of AI 2022 report

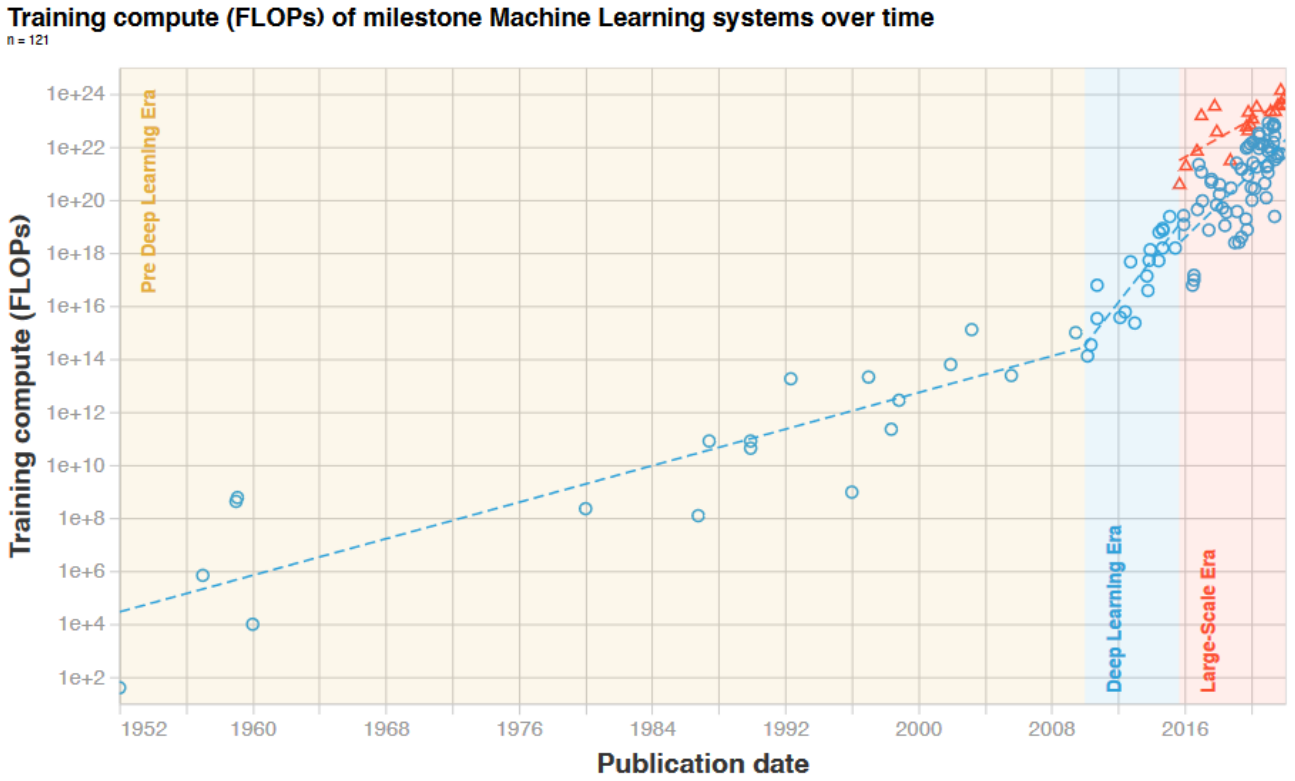


Figure 1: Trends in $n = 121$ milestone ML models between 1952 and 2022. We distinguish three eras. Notice the change of slope circa 2010, matching the advent of Deep Learning; and the emergence of a new large-scale trend in late 2015.

Study by Epoch, University of Aberdeen, Center for the Governance of AI, University of St. Andrews, MIT, Eberhard Karls Universitat Tubingen, Universidad Complutense

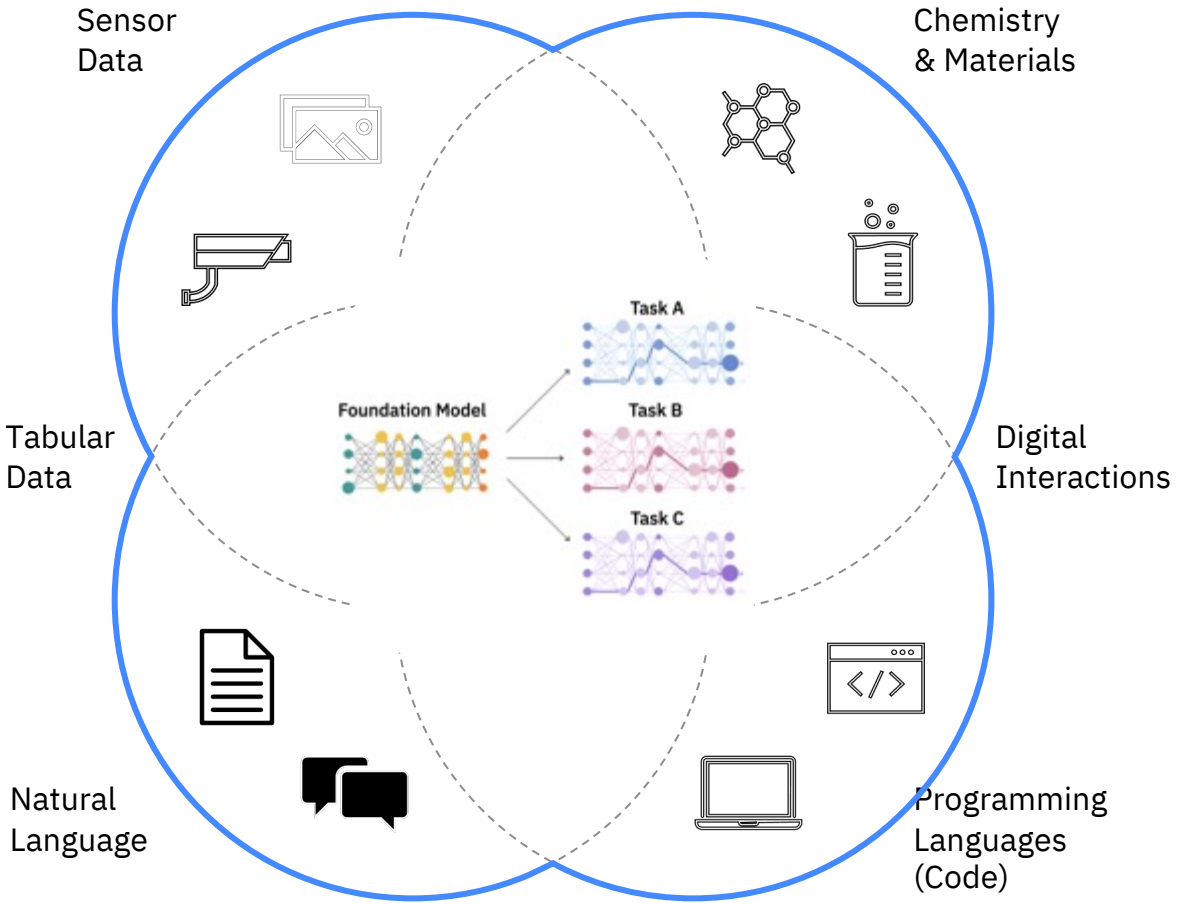
 © SNIA. All Rights Reserved.

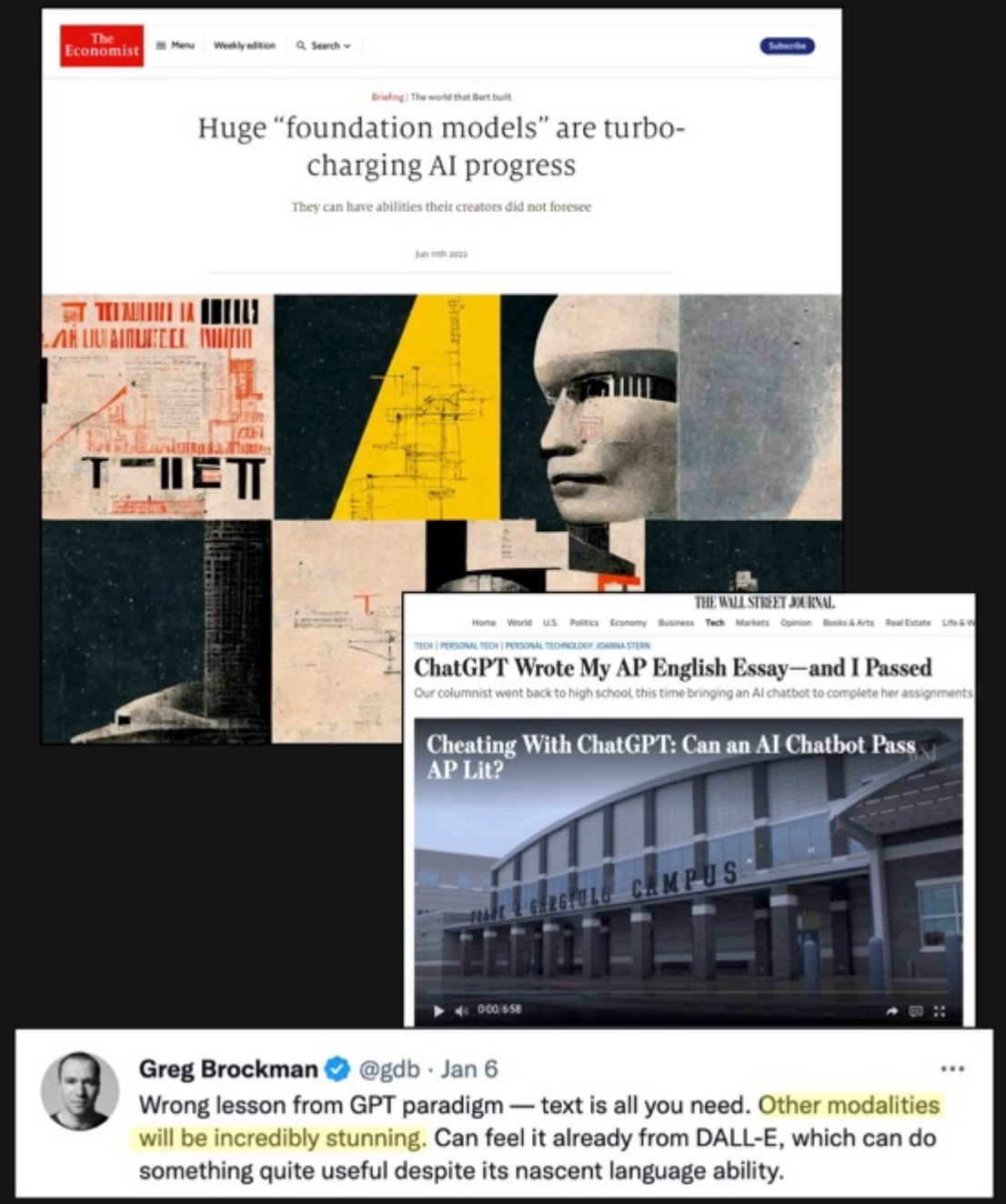
Examples of AI Adoption and Benefits

AI adoption will change the game across many domains

 © SNIA. All Rights Reserved.

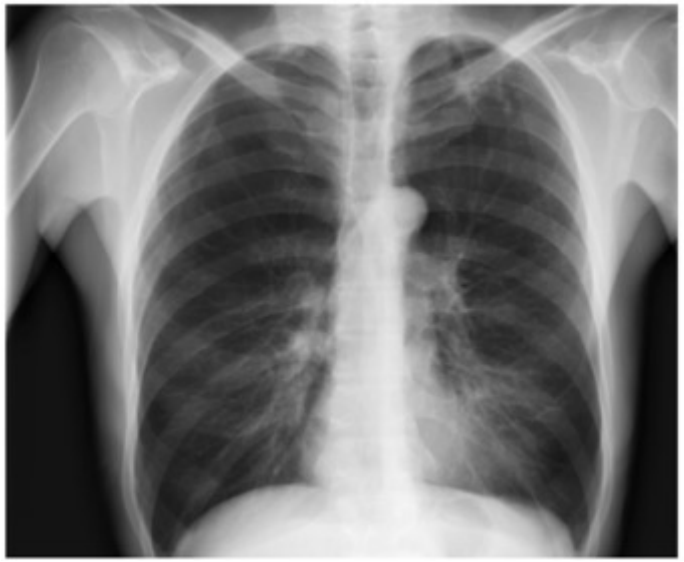
Foundation models are poised to change the game across data modalities and domains

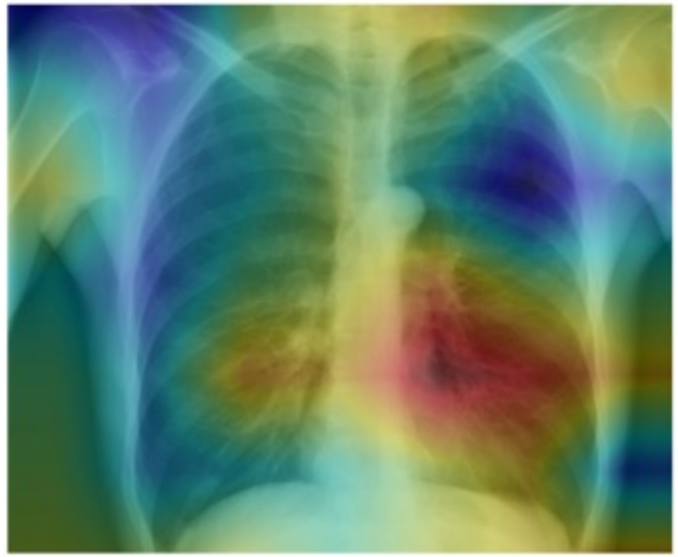




Detection of COVID Pneumonia in Chest X-Rays

leidos

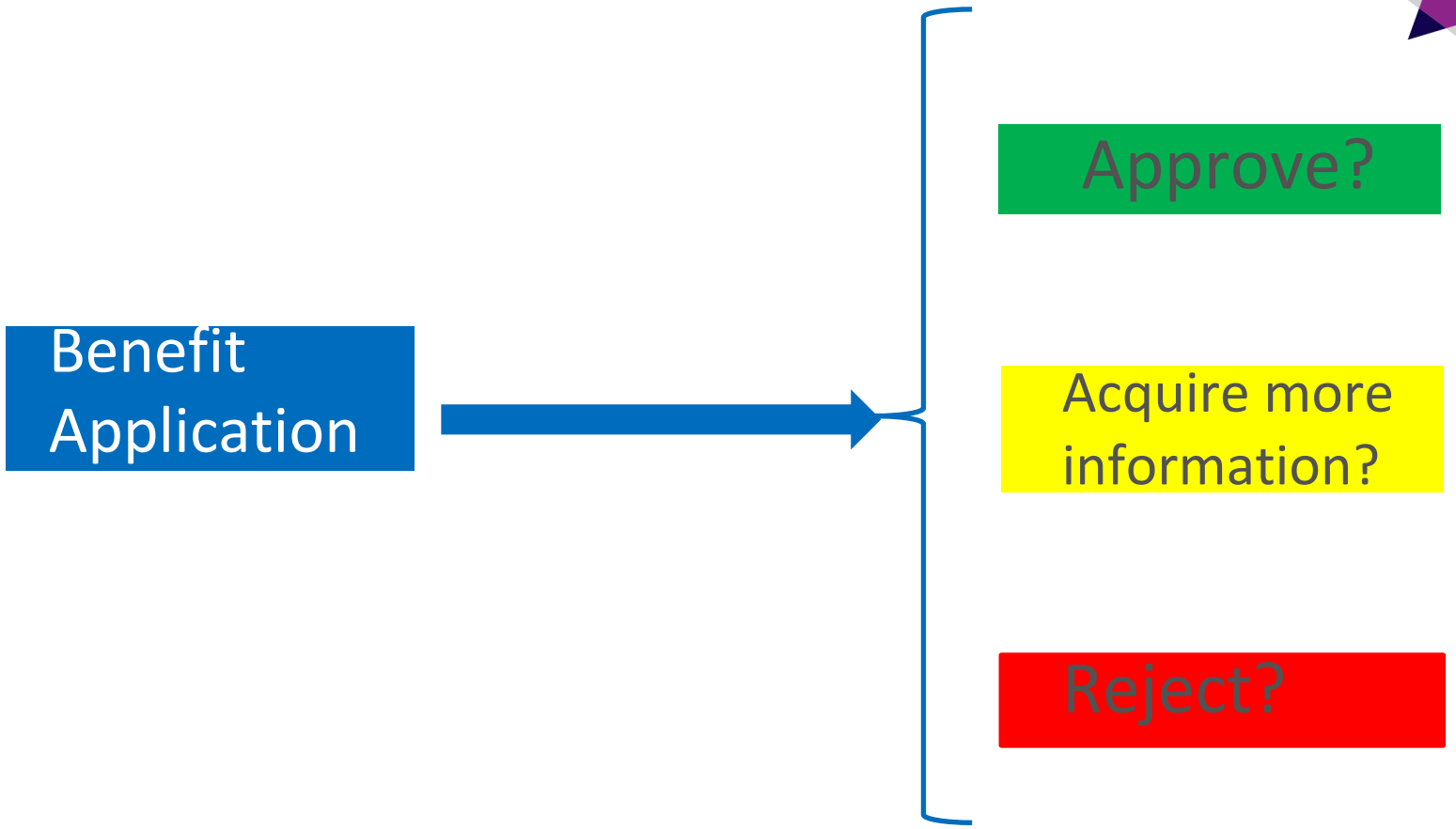




COVID Pneumonia?

 © SNIA. All Rights Reserved.

Model for Dispositioning of Medical Benefit Applications

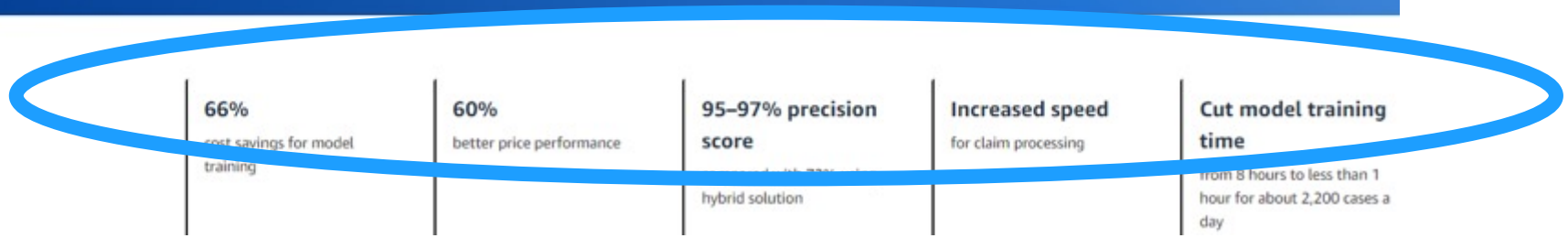


 © SNIA. All Rights Reserved.

Read the Leidos Case: Amazon EC2 DL1 Instances

https://aws.amazon.com/solutions/case-studies/leidos-case-study/

January 2023 AI Forum – Internal NDA



Mobileye

Custom object detection

(2D and 3D) models trained on Gaudi

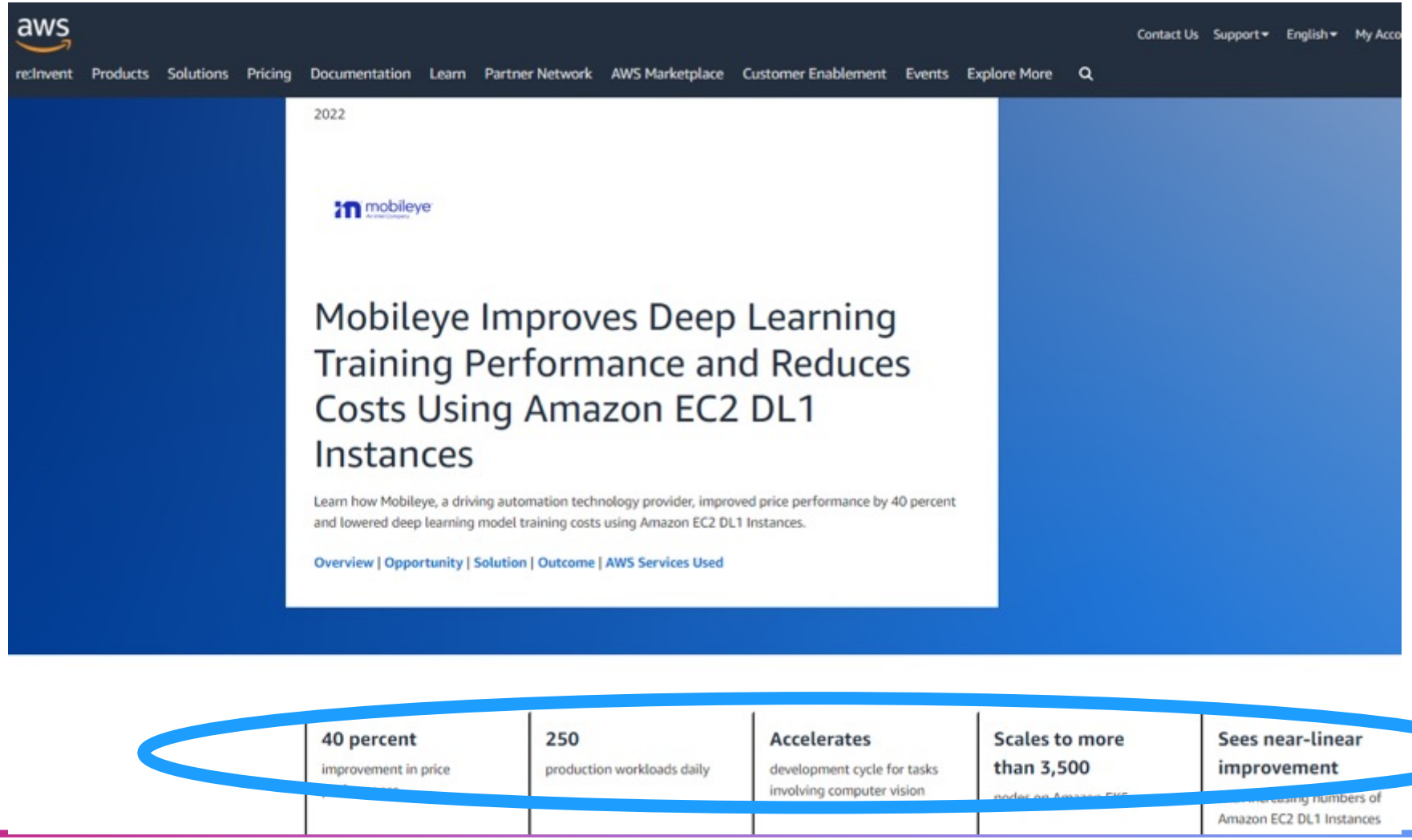
"On our own models the increase in price performance met and even exceeded the published 40% mark."

Chaim Rand, Mobileye

 © SNIA. All Rights Reserved.

Read the Mobileye Case: Amazon EC2 DL1 Instances

https://aws.amazon.com/solutions/case-studies/mobileye-ec2-dl1-case-study/

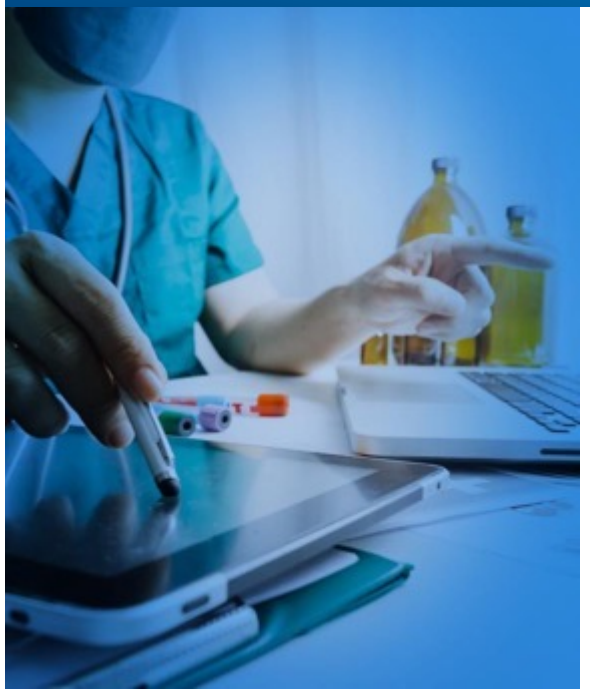


 © SNIA. All Rights Reserved.

“DIY AI” in Medicine

Habana and Hugging Face at Northwestern

Northwestern Medicine®

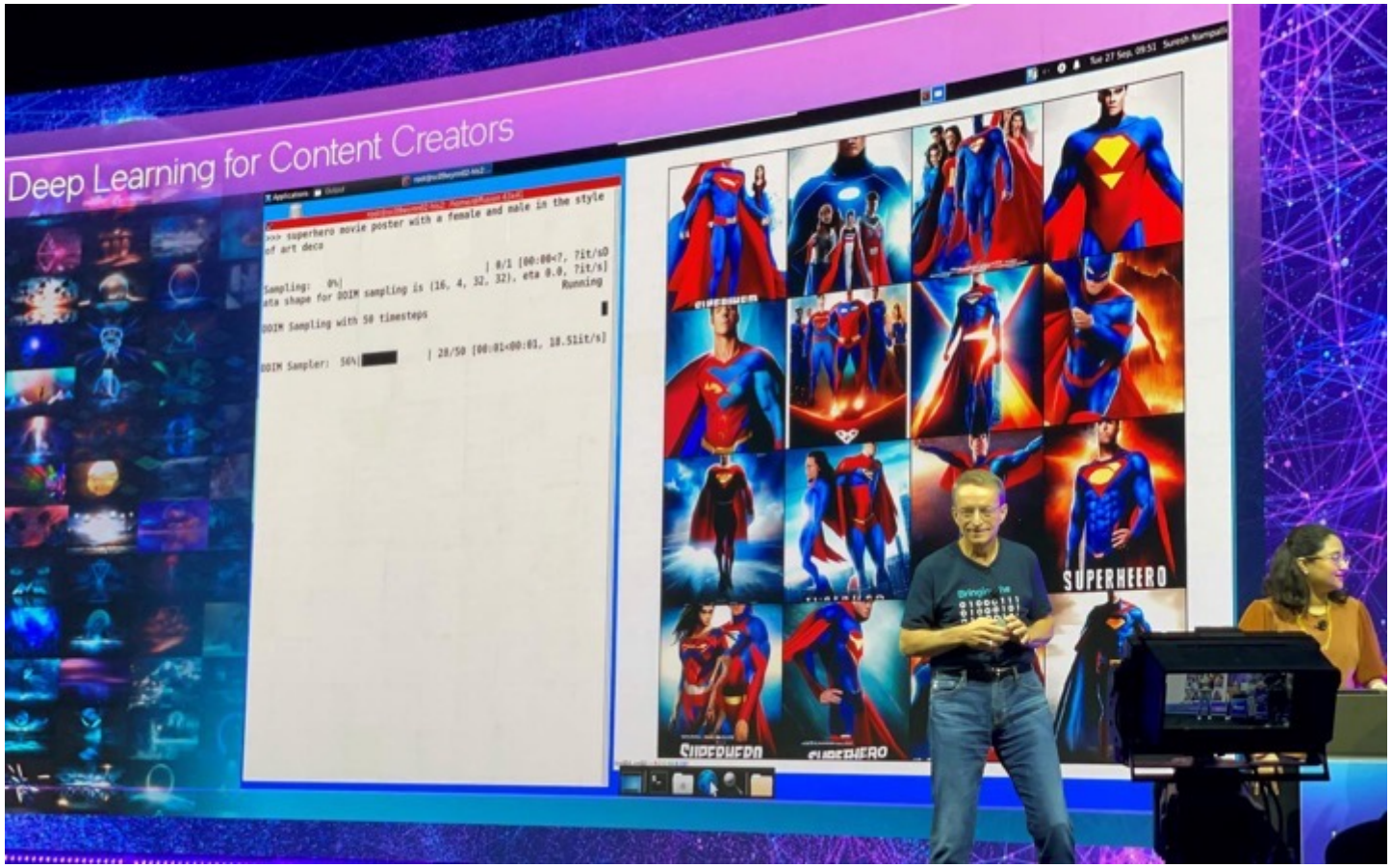


“By partnering with Hugging Face, Habana is democratizing AI. They provide documentation and examples so that a medical expert with minimal data science expertise can produce a useful clinical model in under fifty lines of code. The partnership enables scalability, reduces friction and lowers costs to train large language models. Habana and Hugging Face are providing ‘DIY AI.’”

Mozziyar Etemadi, MD, PhD, Assistant Professor of Anesthesiology and McCormick School of Engineering, Northwestern Medicine

 © SNIA. All Rights Reserved.

Stable Diffusion for Image Generation



Stable Diffusion Model based on https://github.com/pesser/stable-diffusion

Check out Keynote featuring stable diffusion demo at Intel Innovation in Sep'22

 © SNIA. All Rights Reserved.

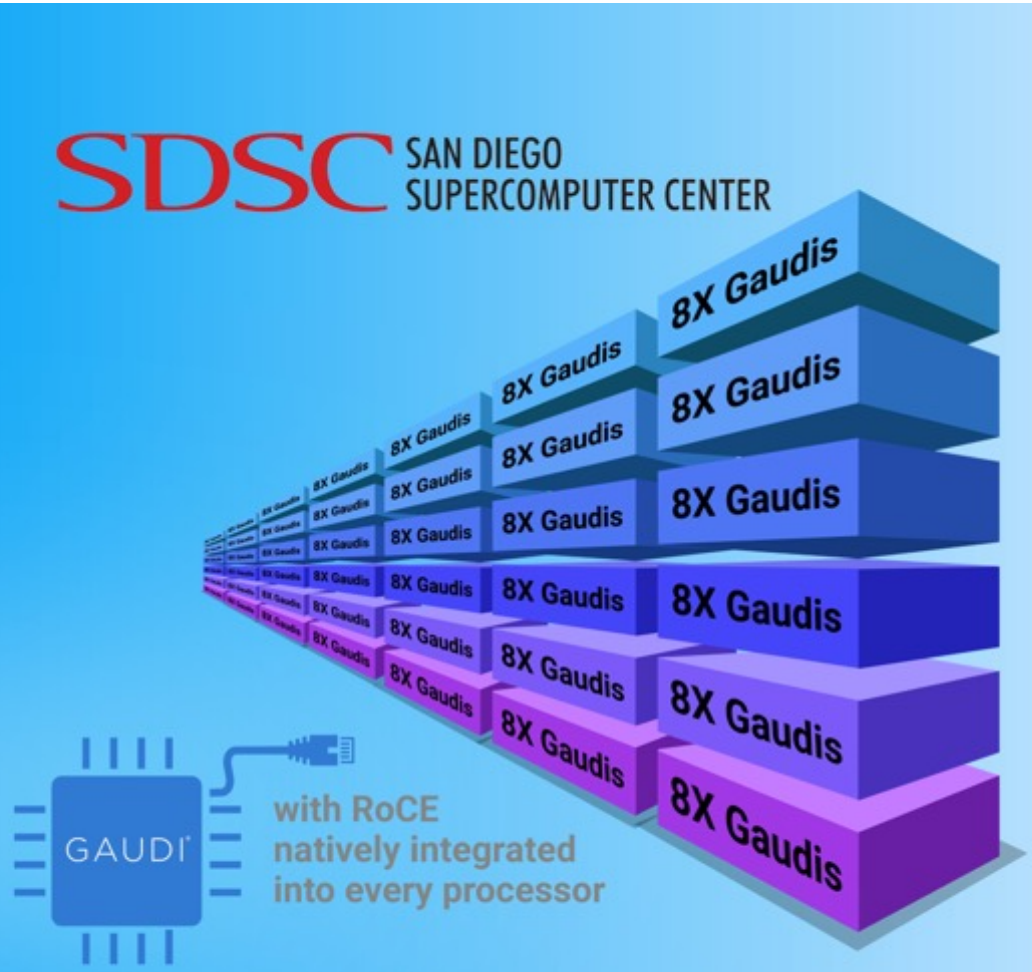
Considerations for Adoption of AI Technologies

Scale Out Infrastructure: High Performance and Cloud Flexibility

 © SNIA. All Rights Reserved.

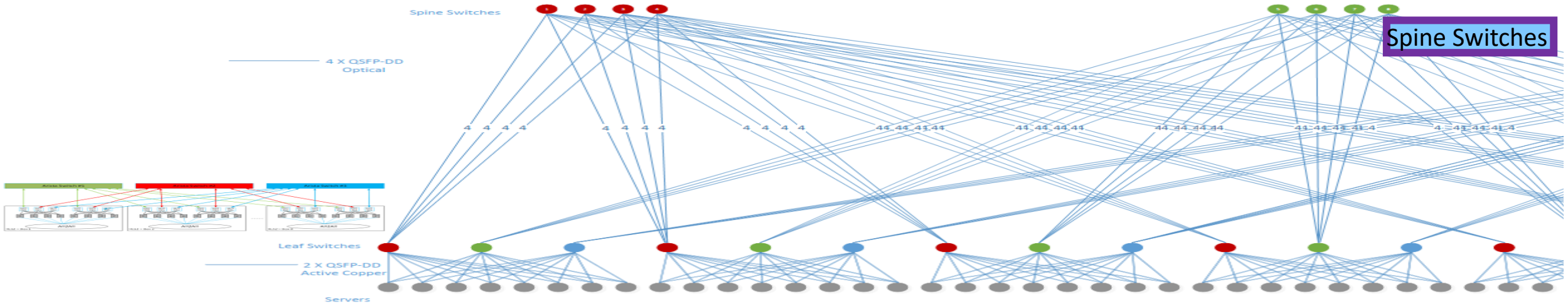
SDSC Voyager Cluster

AI Training Server



 © SNIA. All Rights Reserved.

Massive and Flexible Scale-up and Scale-out



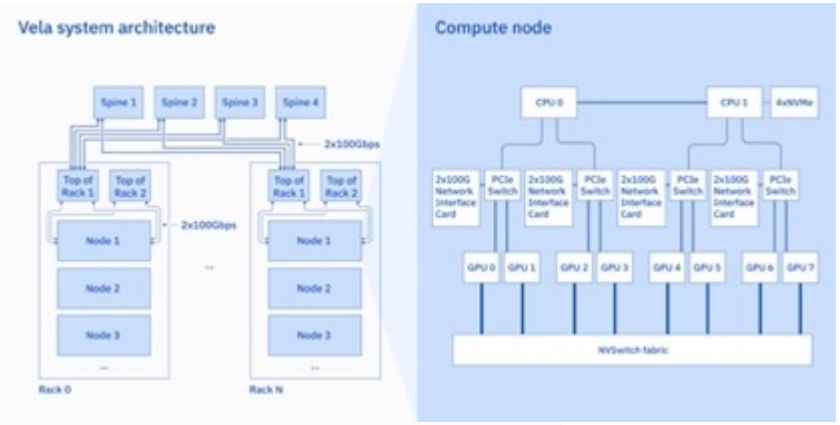
Spine Switches

Leaf Switches

Gaudi2 Servers

 © SNIA. All Rights Reserved.

Vela – Cloud AI System

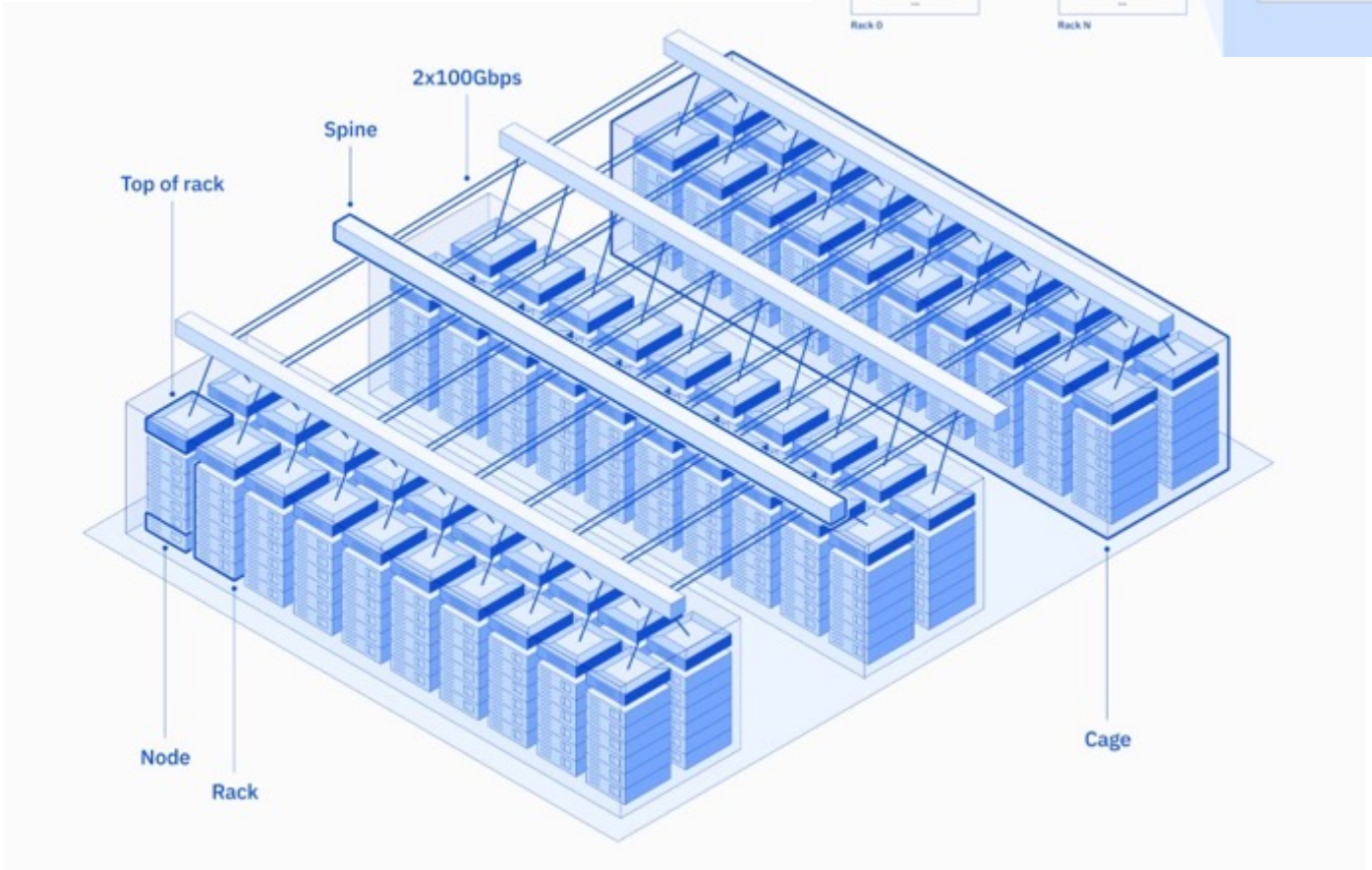


- System specifications

– Nodes: NVIDIA HGX™ A100 (8 GPUs with 80GB/GPU)
– GPUs interconnected with NVIDIA® NVLink® and NVIDIA® NVSwitch™
– Cascade Lake CPUs, 1.5TB of DRAM,
– Four 3.2TB NVMe drives
– Redundant connections between nodes, TORs and spine
– 2 x 100G NICs from each node – NCCL benchmarks show we drive close to line rate

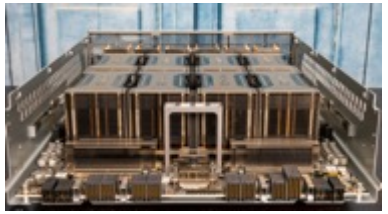
– Configure resources through software (APIs)
– Broad ecosystem of available cloud service
– Leverage data sets on Cloud Object Store
– Collaborate, leveraging IBM Cloud VPC
– Standard, flexible, scalable infrastructure design (vs traditional HPC)
– Near bare metal performance (within 5%, single node)

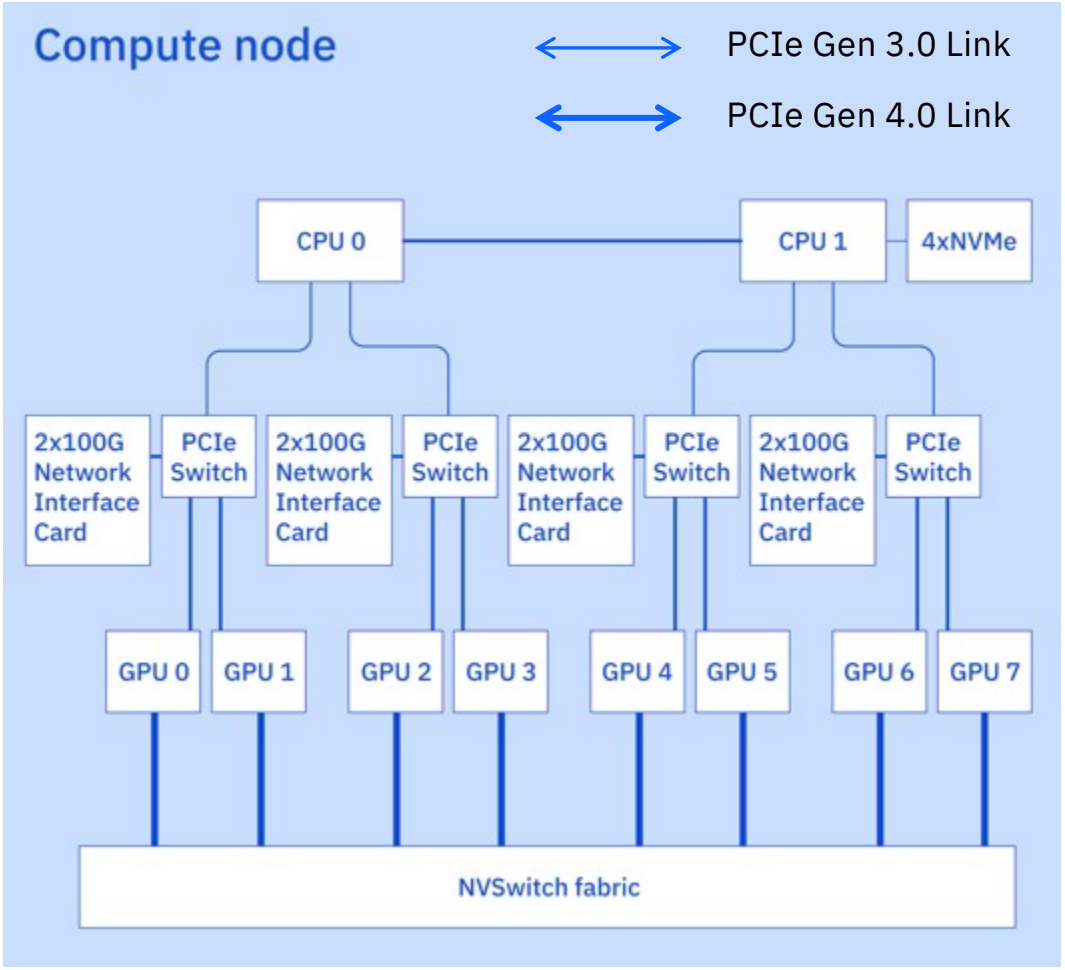
- https://research.ibm.com/blog/AI-supercomputer-Vela-GPU-cluster



Node Architecture

- Compute nodes have complex system architecture
- A layer of virtualization to abstract this complexity is critical to drive usability





Node Virtualization Software

- Workloads
- Application Runtime
- Virtual Machine + Guest OS
- Host OS + Hypervisor
- Bare-Metal Node
- Networking

NVIDIA CUDA

NVIDIA/nccl

Optimized primitives for collective multi-GPU communication

NVIDIA/apt-packaging-fabric-manager

NVIDIA

NVIDIA

Guest

libvirt VIRTUALIZATION API

QEMU

MLNX-OFED for Linux

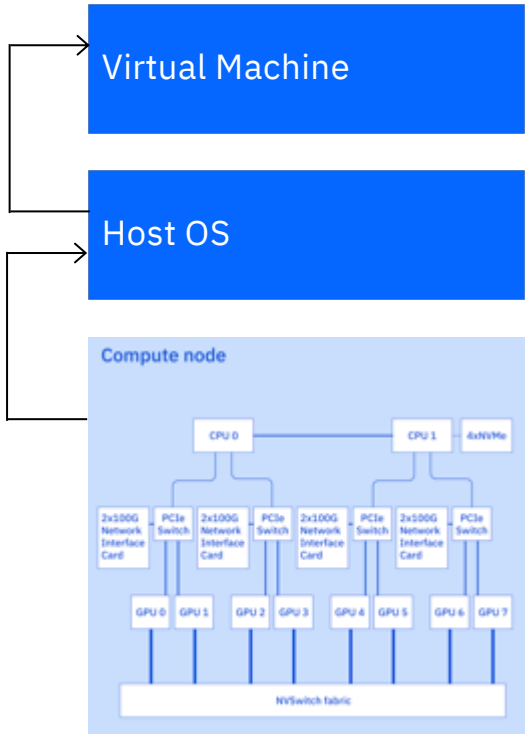
UBUNTU 20.04 FOCAL FOSSA

Linux KVM

Host

 © SNIA. All Rights Reserved.

Close to Bare Metal Performance with Improved Virtualization



Before	TCP		RoCE		RoCE with GDR	
	BM	VM	BM	VM	BM	VM
1 NIC	7.92 →	4.5	8.11 →	1.5	12.12 →	1.5

After	TCP		RoCE		RoCE with GDR	
	BM	VM	BM	VM	BM	VM
1 NIC	7.92 →	8.12	8.11 →	8.09	12.12 →	12.41
2 NIC	15.03	16.41	16.46	16.48	24.71	24.71
4 NIC	15.32	16.48	16.45	16.47	48.13	48.35
8 NIC	15.7	15.75	16.46	19.94	97.01	96.71

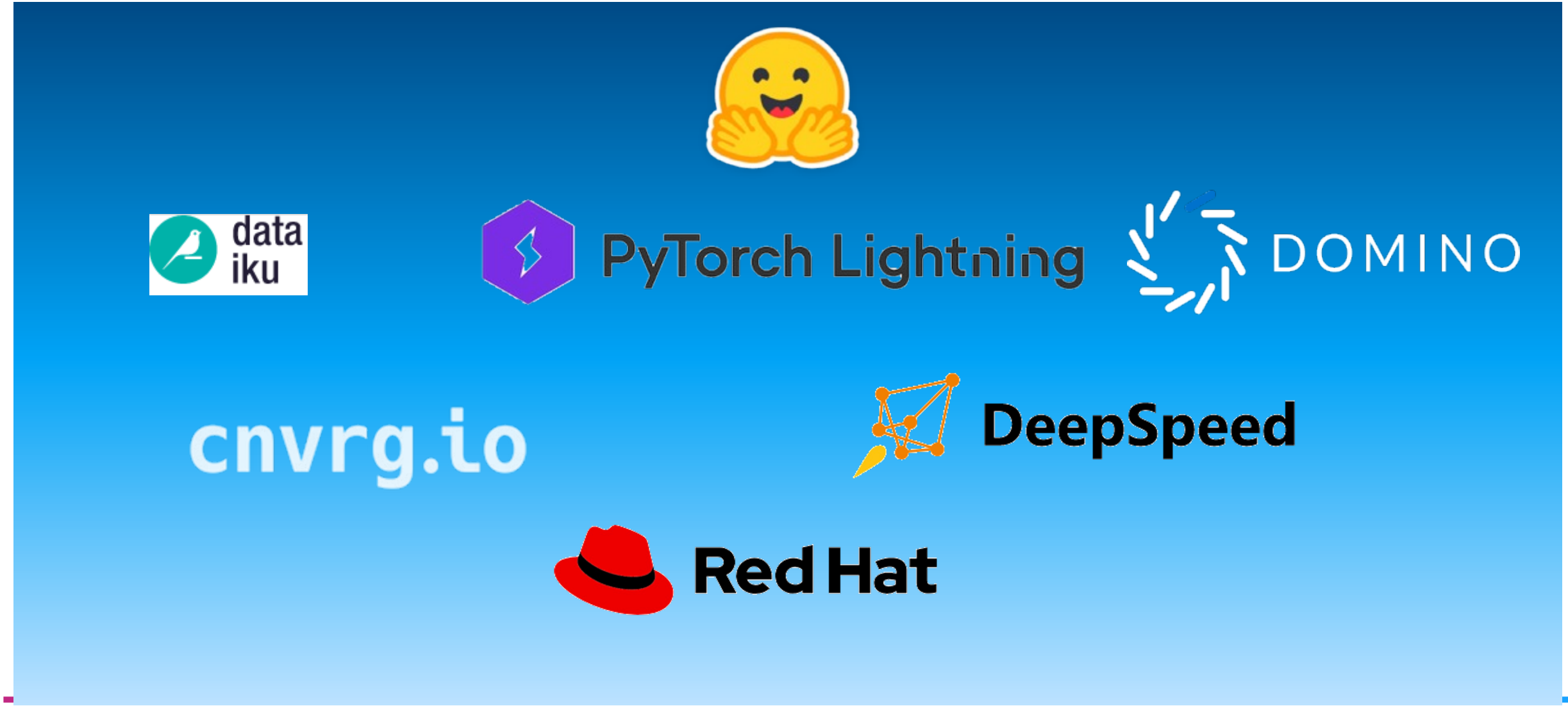
See details in the presentation at GTC 2021

Considerations for Adoption of AI Technologies

Unified Platform for Training and Inference

 © SNIA. All Rights Reserved.

Software Ecosystem



 © SNIA. All Rights Reserved.

A Platform for High-performance, Distributed AI Anywhere

Foundation model stack for Training and validation

RAY

PyTorch

Cloud-native middleware for high end user productivity across model creation life cycle (data preparation, training, and model validation workflows)

Hybrid Cloud Platform

kubernetes

Enhanced Job Management and queueing (MCAD)

Enhanced placement and auto-scaling (InstaScale)

Advanced network configurations (multi-NIC CNI)

...

Scale-out Infrastructure

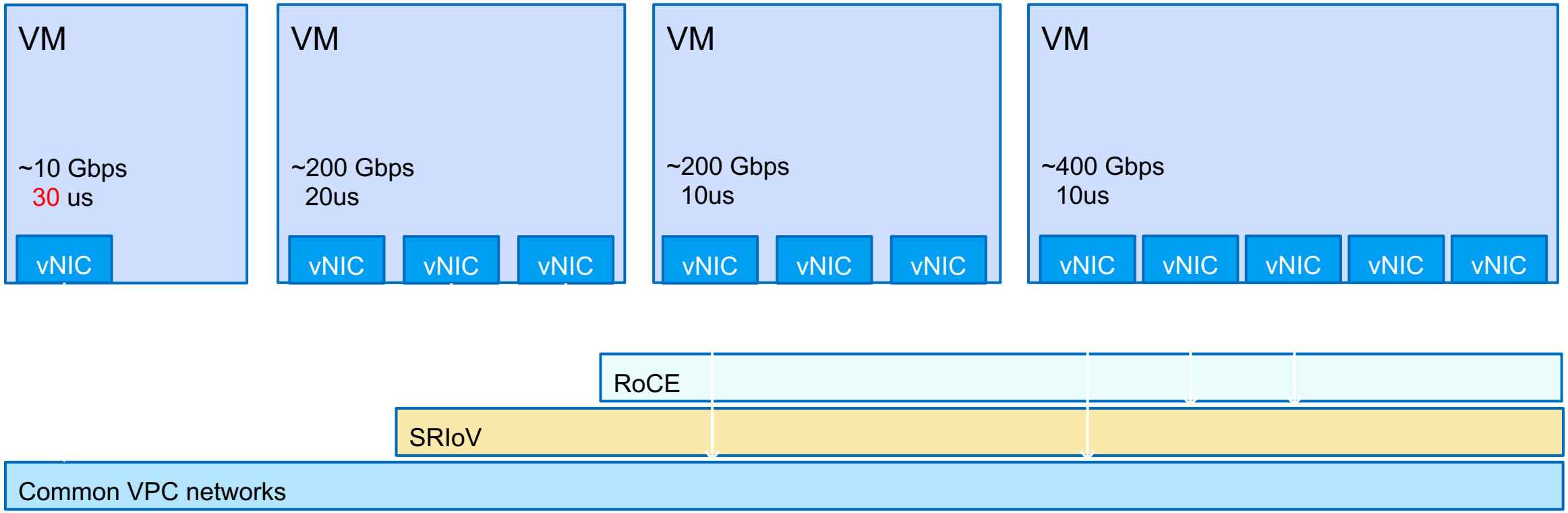
IBM Cloud

On-Prem

Other cloud

Supercomputing-class performance for foundation model training, with cloud-like flexibility and optimized for cost

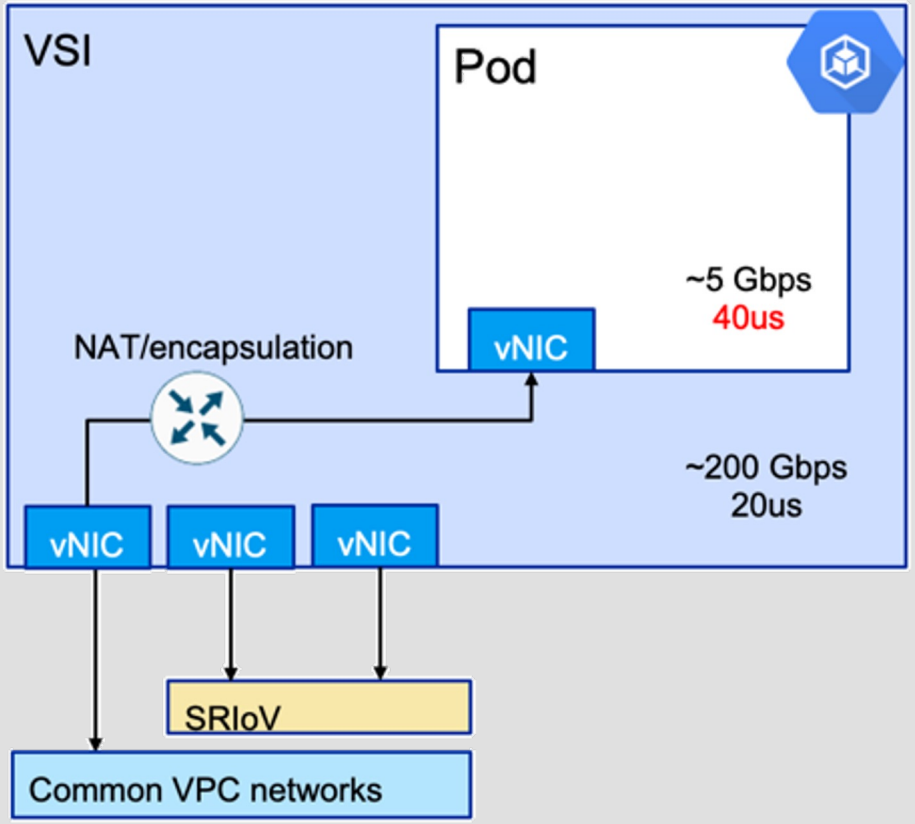
Networking in AI Infrastructure in the Cloud



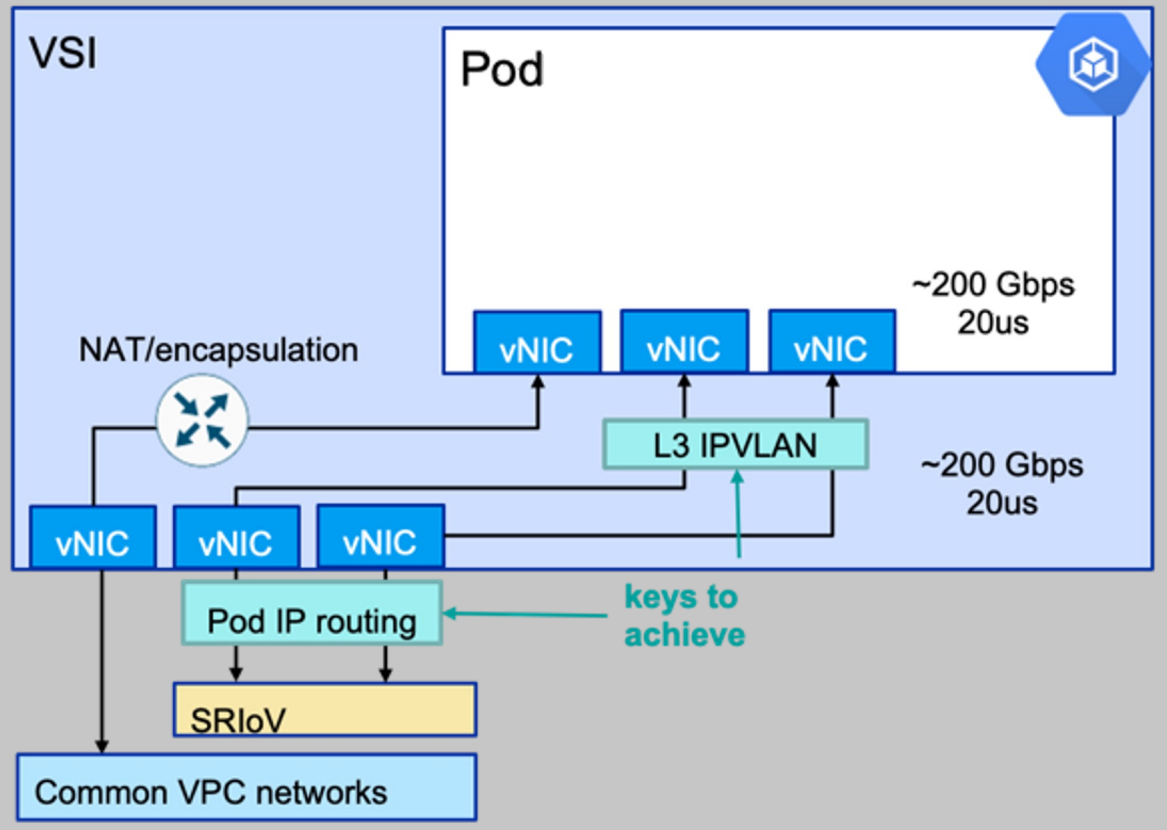
KubeCon Bath Day presentation with details

 © SNIA. All Rights Reserved.

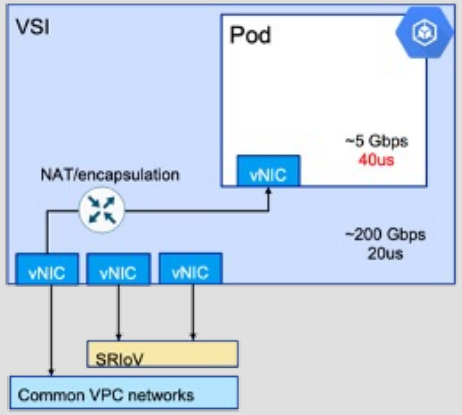
What default networks are



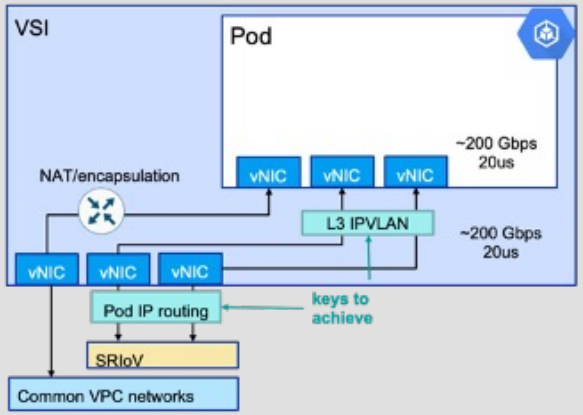
What we need for AI and HPC



What default networks are

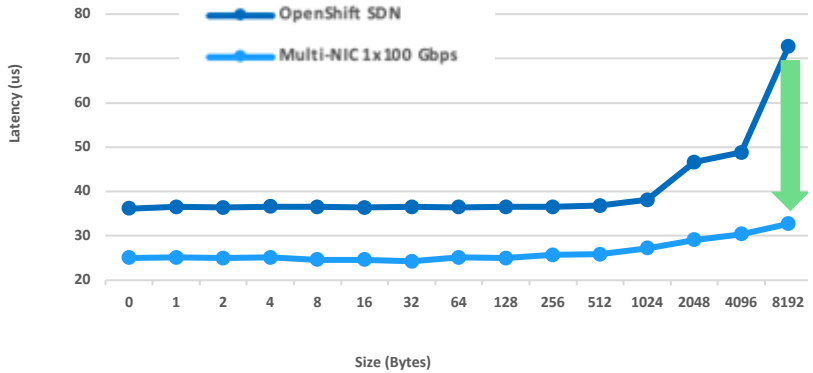


What we need for AI and HPC



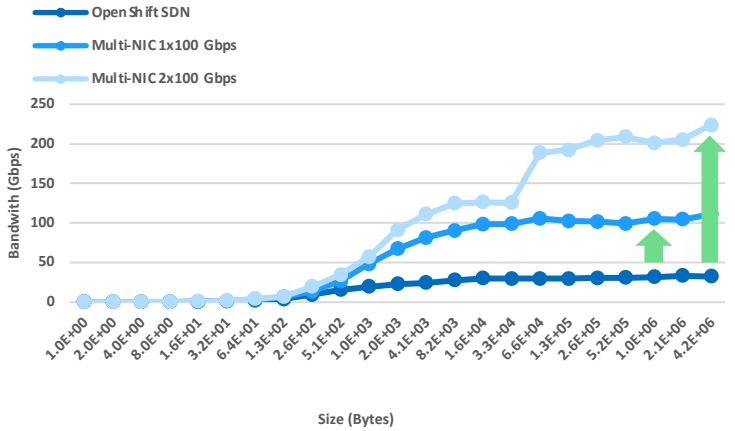
Half Network Latency

OSU MPI Latency Test v5.9



7x Network Bandwidth

OSU MPI Multiple Bandwith v5.9 – 2 Nodes

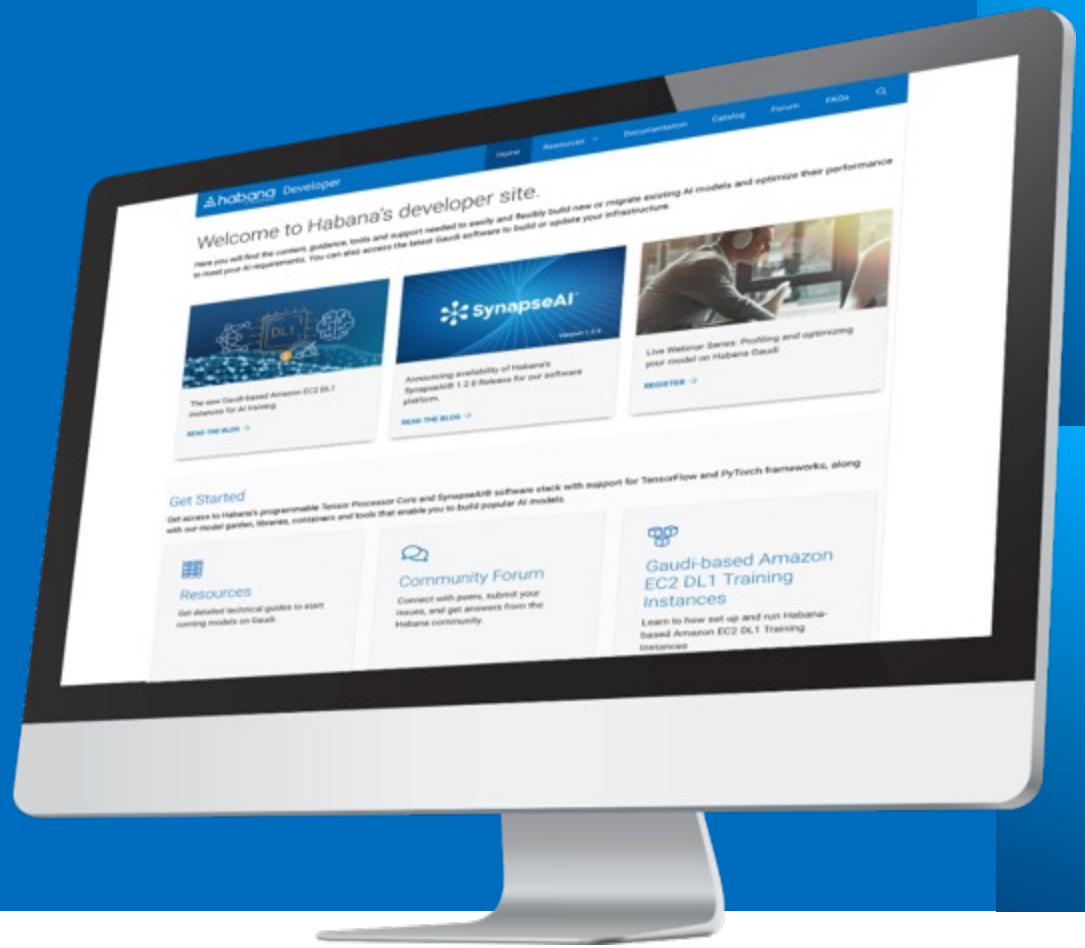


Considerations for Adoption of AI Technologies

Middleware Stack and Tools

 © SNIA. All Rights Reserved.

Developer Usability



END-USER
EASE OF USE

Build new models

Migrate existing GPU-based models

END-USER
SUPPORT

Developer Site

GitHub

Forums

Trainings & Tutorials

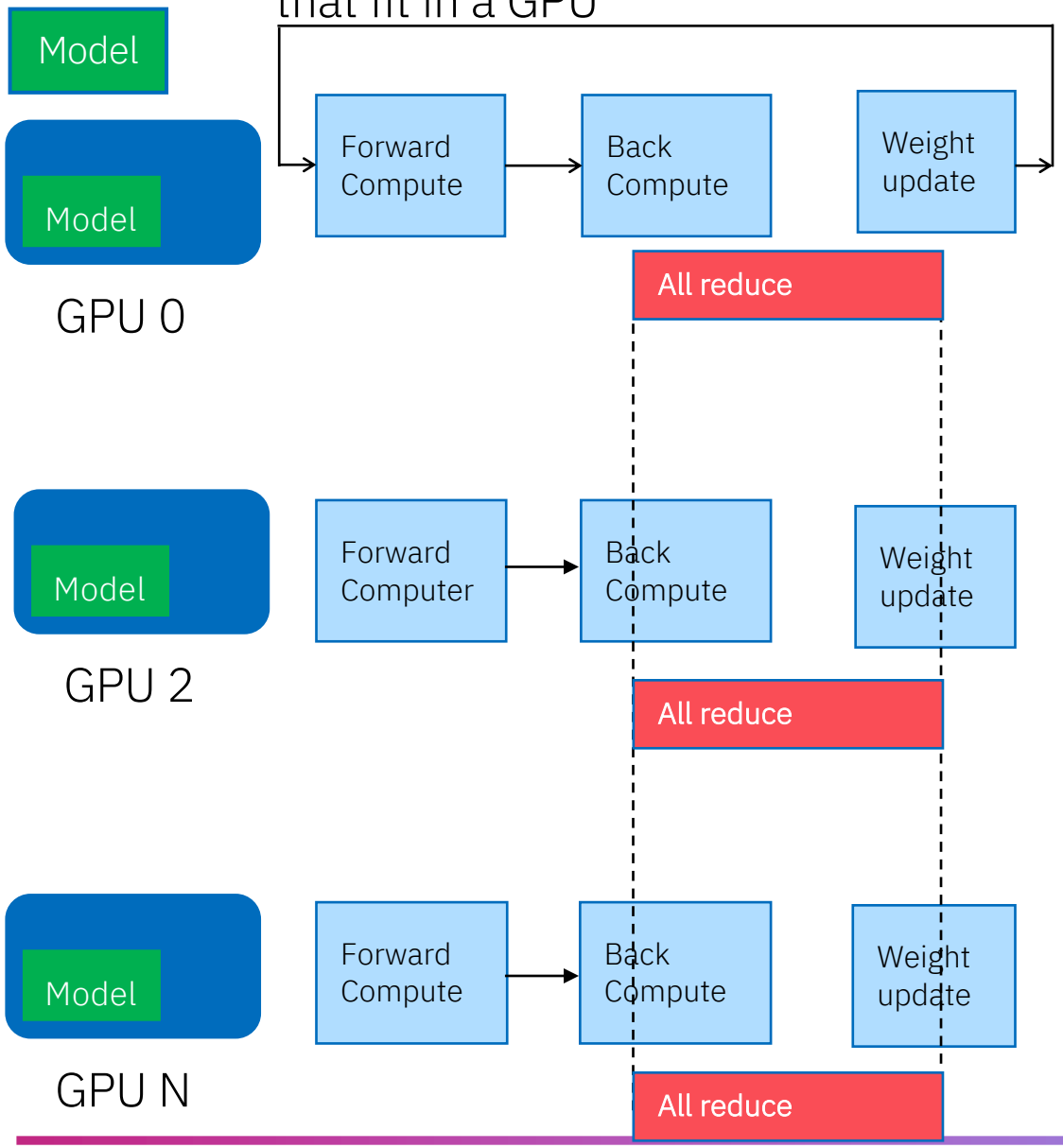
END-USER
MODEL ENABLEMENT

TensorFlow and PyTorch frameworks

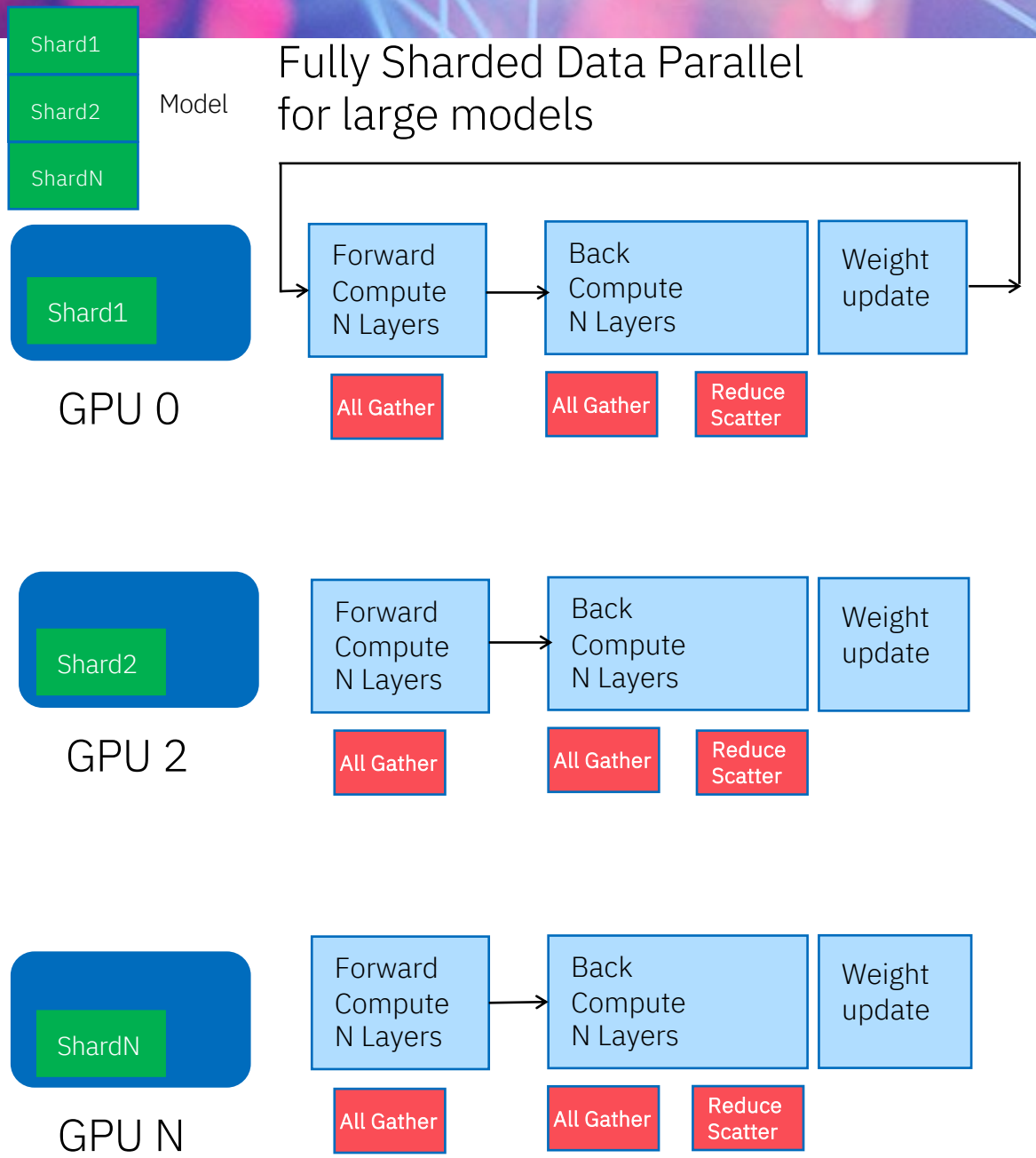
Reference code and scripts for popular models

 © SNIA. All Rights Reserved.

Distributed Data Parallel for Models that fit in a GPU



Fully Sharded Data Parallel for large models



 © SNIA. All Rights Reserved.

Software Stack to Simplify, Automate, and Scale

RAY

Ray Use Cases

Evaluating large language models with Ray in hybrid cloud

Ray Summit 2022

Linsong Chu

- Ray workflows for data pre-processing

Ray Workflows for model validation

GLUE Benchmarking

QA Benchmarking

Sentiment Workloads

PyTorch

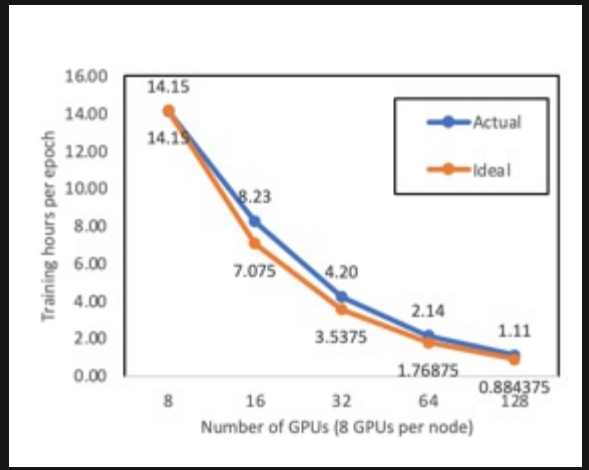
December 15, 2022

Scaling PyTorch FSDP for Training Foundation Models on IBM Cloud

by Linsong Chu, Less Wright, Hamid Shojanazeri, Sophia Wen, Raghu Ganti, Geeta Chauhan

- 80-90+% parallel efficiency has been demonstrated for:
  - Time series model
  - NLP models (BERT, RoRERTa, T5, etc)
  - Cyber Security
  - Code (Project Wisdom)
  - More to come

FSDP T5 Model training (11 B parameters)



Conclusions

- AI workload compute requirements are doubling every 10 months
- AI adoption will accelerate and influence many industries
- Scale out infrastructure that is flexible and cost-effective is critical to capitalize on this new technology
- Users need a unified cloud-native platform for training and inference
- Technologies like Kubernetes, Pytorch, Ray, SynapseAI SDK enable AI developers to quickly leverage AI technologies in their application use cases

 © SNIA. All Rights Reserved.

Thanks for Viewing this Webcast

- Please rate this presentation and provide us with feedback
- This webcast and a copy of the slides are available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webcast will be posted to the SNIA Cloud blog: www.sniacloud.com/
- Follow us on Twitter @SNIACloud

 © SNIA. All Rights Reserved.

Thank You!

 © SNIA. All Rights Reserved.