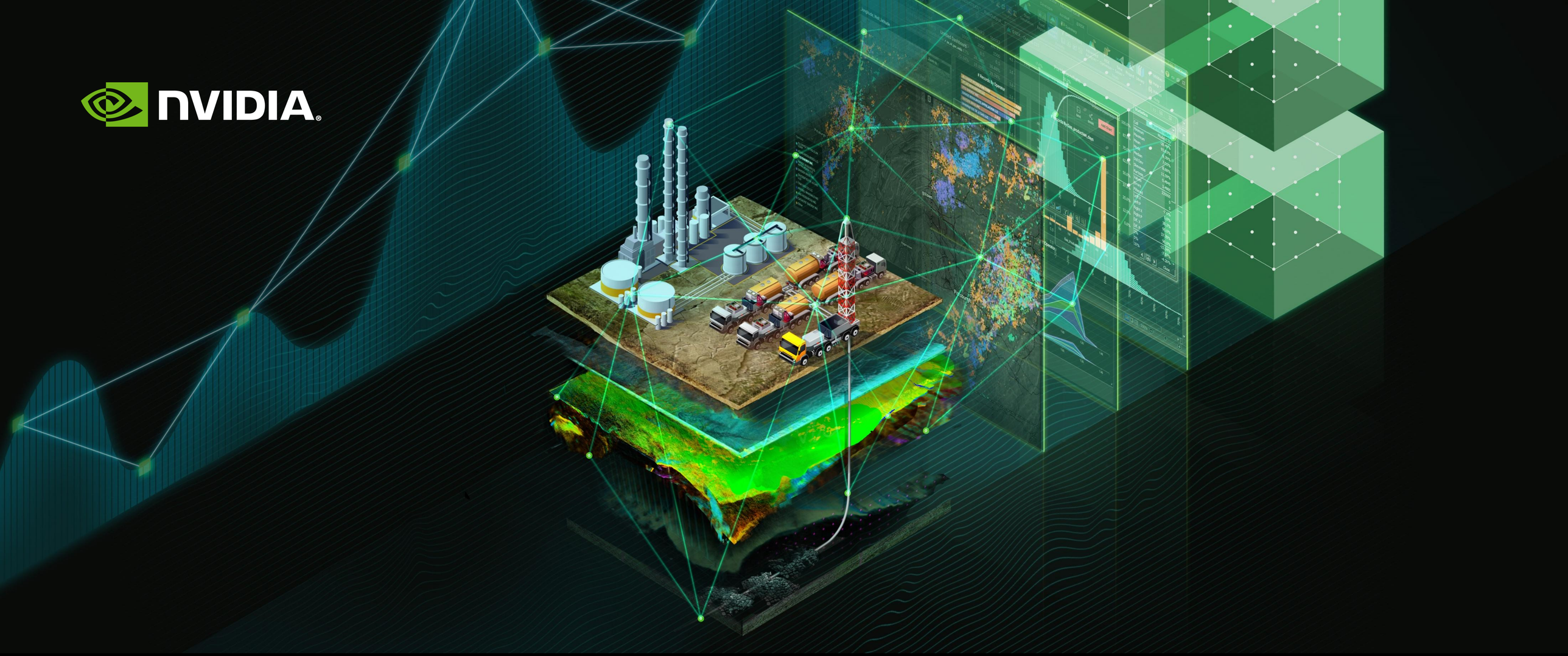


Seismic Imaging at Scale: The Future of HPC and AI – A NVIDIA Perspective

João Paulo Navarro – Sr. Solutions Architect

jpnavarro@nvidia.com

What Is NVIDIA Energy Samples

GPU Accelerated Building Blocks For Expediting Exploration Algorithms,
Maximizing Performance Lowering Operational Costs For Energy Customers

Customer/ISV Applications

Quickly evaluate and integrate GPU accelerated techniques into applications in the cloud and on-prem

Energy SDK

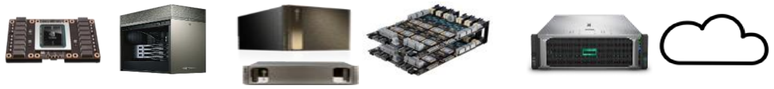
Energy sped-of-light CUDA-first sample algorithms for on-prem and cloud such as **FWI, RTM**, Kirchhoff, ResSim, imaging, etc.

Compression and upscaling techniques for handling and pipelining subsurface, surface, edge, and far edge data

Physics ML & AI models to support digital transformation and energy transition

HPC SDK **RAPIDS** **PhysicsNeMo**

Integrate the best of HPC SDK, RAPIDS, PhysicsNeMo, and GPU Direct Storage (GDS)



Leverage the GPU architectural features and tested on various system configs for on-prem / cloud deployment

Seismic Imaging

Use Cases

- Full Waveform Inversion (FWI)
- Reverse Time Migration (RTM)
- Kirchhoff Depth Migration
- Surface-related multiple elimination (SRME)

Challenges

Massive workloads that require significant memory for processing and interpretation

Solution

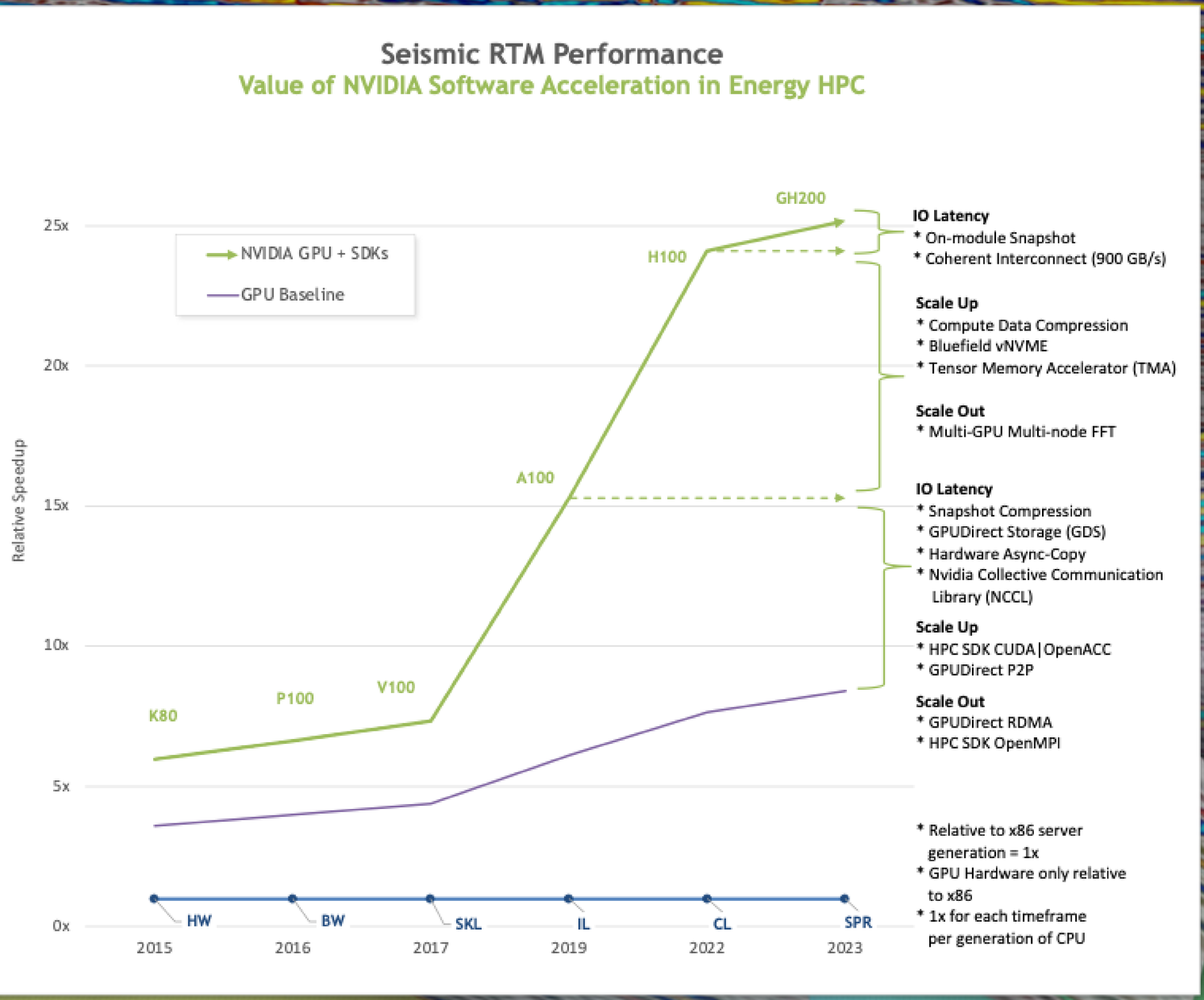
- Internal code development built to be optimized on the GPU
- Solutions provided under NDA to partners, continued development and optimization with their development teams

NVIDIA Solution Stack

- Hardware: A100, H100, H200, GH200
- Software: Energy SDK, HPC SDK

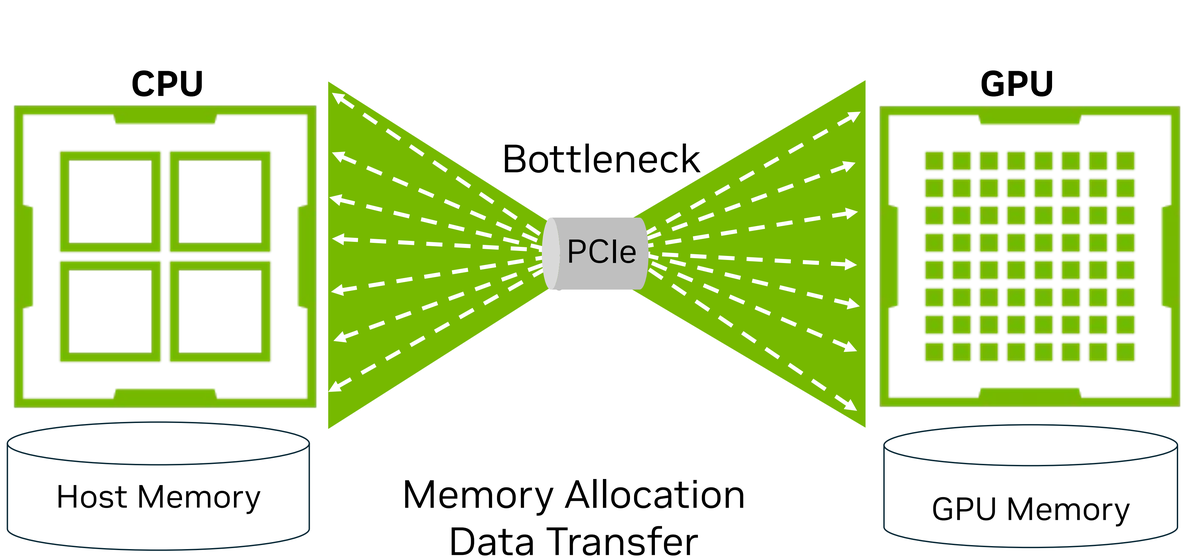
Outcomes

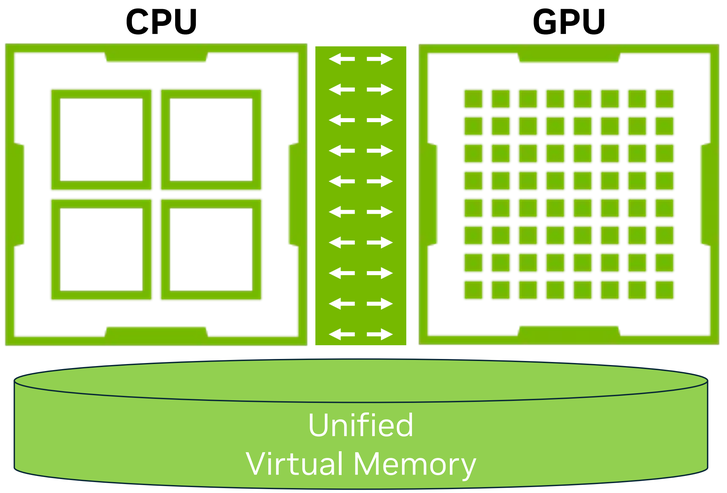
- 9x relative speedup of seismic + GPU
- 25x relative speedup of seismic + GPU + SDK
- Software-defined solution allows for significant acceleration of seismic interpretation
- Key ISVs include SLB, Shearwater, AspenTech (former Emerson)



Challenges with Traditional Accelerated Systems

Bridging the GAP: Unifying CPU and GPU Memory





- ❑ PCIe bottlenecks CPU-GPU communication
- ❑ GPU can't access CPU memory directly
- ❑ Complex Programmability
- ❑ Low CPU Performance per Watt

Legacy PCIe Architecture

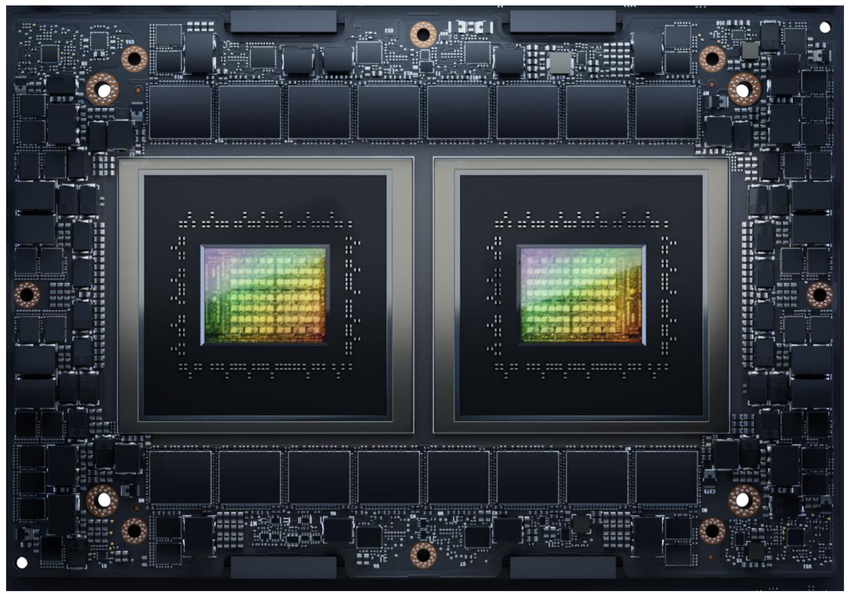
We need a new chip architecture:

- ✓ Accelerate CPU-GPU data transfers
- ✓ Coherent CPU and GPU cache memory
- ✓ Simplify Programmability
- ✓ Deliver more performance per Watt

NVIDIA Grace for Cloud, AI and HPC Infrastructure

Grace CPU Superchip

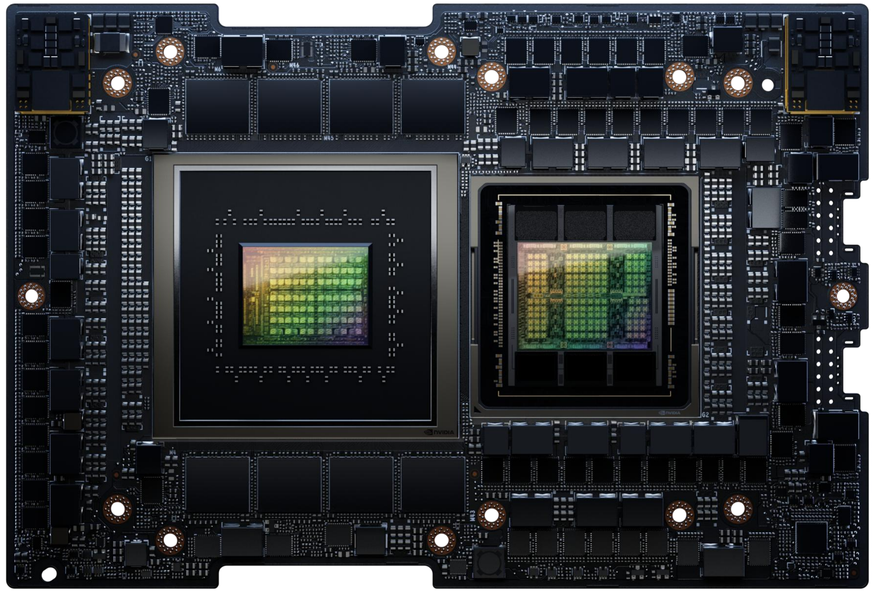
CPU Computing



CPU-based applications where absolute performance, energy efficiency, and data center density matter, such as scientific computing, data analytics, enterprise and hyperscale computing applications

GH200 Grace Hopper Superchip

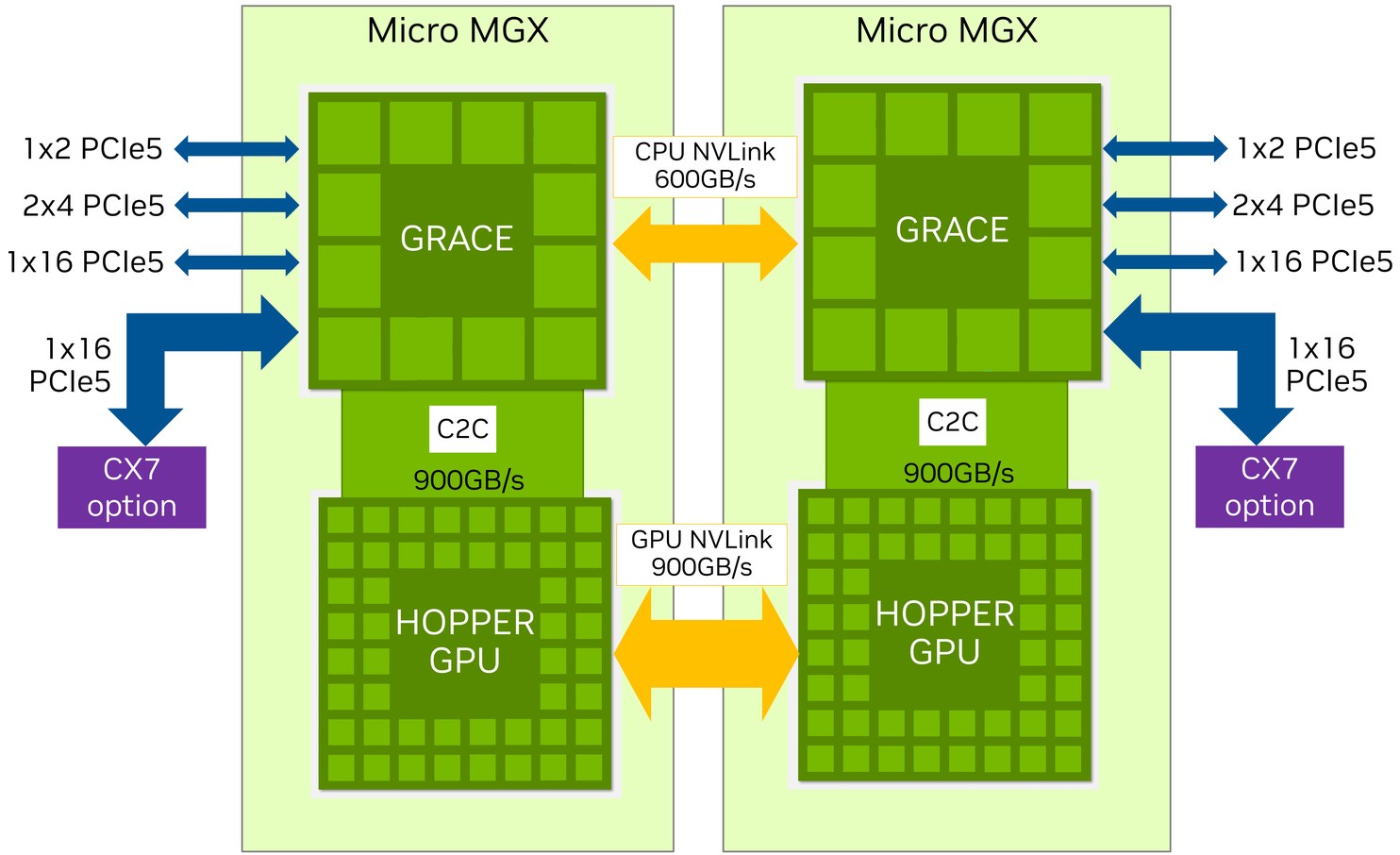
Large Scale AI & HPC



Accelerated applications where CPU performance and system memory size and bandwidth are critical; tightly coupled CPU & GPU for flagship AI & HPC. Most versatile compute platform for scale out.

GH200 NVL2 Node Architecture

Single Large GPU for Era of Accelerated Computing and Gen AI



- Single node, single OS
- Unified address space across 2 CPU + 2 GPU
- Each GPU is cache coherent to the dual CPU
- Option for up to 4x CX7
- 8 PF of AI performance
- 10 TB/s of GPU memory bandwidth and up to 1.2TB total fast memory
- Best performance for compute and memory intensive workloads
  - Mainstream LLM
  - Data Processing
  - RAG

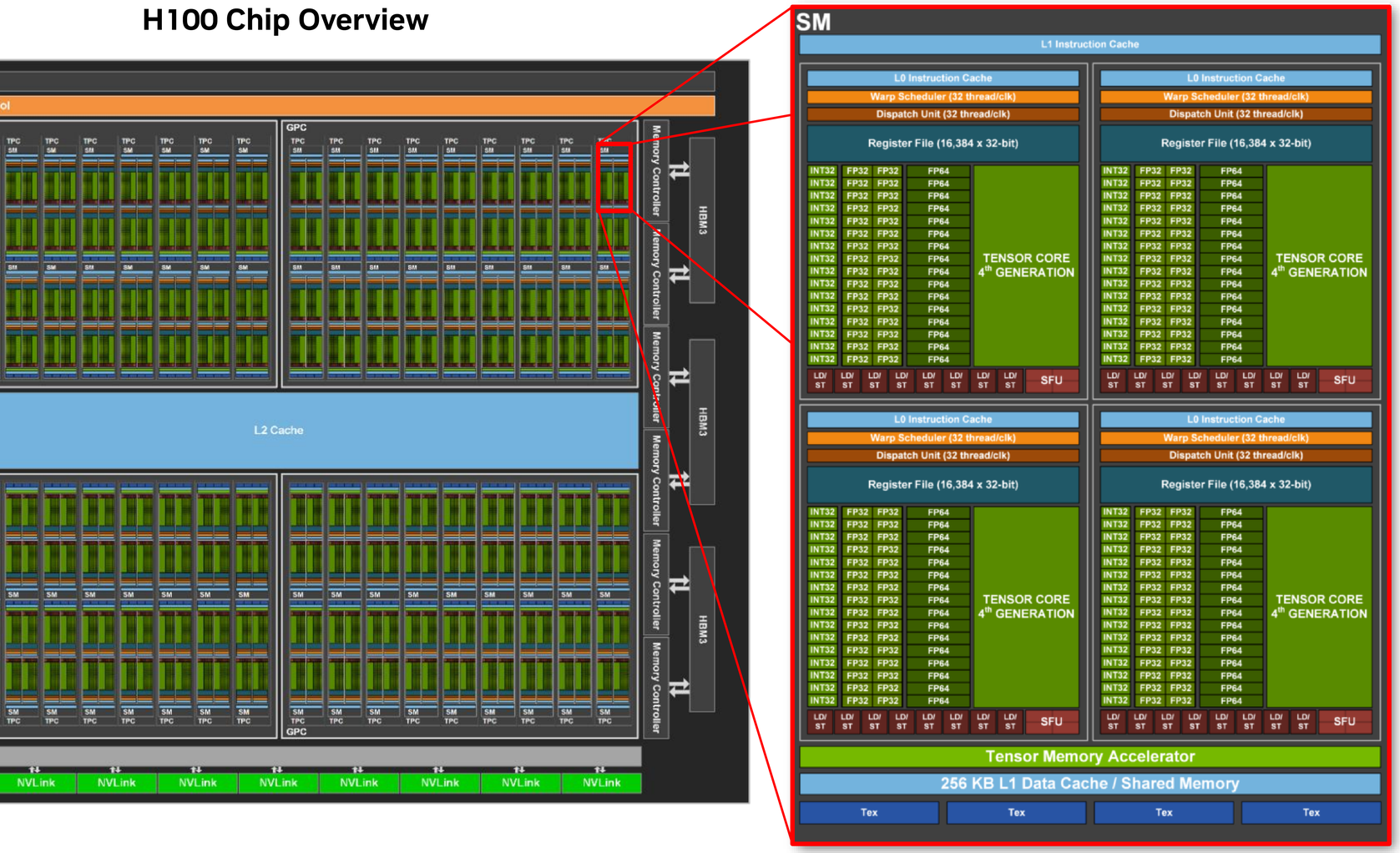
Note: All bandwidth numbers include both directions

How to Go Further With the Same Chip?

GPU: A Processor Made of Processors

Multiple Layers of Specialized Cores Handle Diverse Data Types

H100 Chip Overview



	NVIDIA H100 SXM5[1]
Peak FP64[1]	30 TFLOPS
Peak FP64 Tensor Core[1]	60 TFLOPS
Peak FP32[1]	60 TFLOPS
Peak FP16[1]	120 TFLOPS
Peak BF16[1]	120 TFLOPS
Peak TF32 Tensor Core[1]	500 TFLOPS | 1000 TFLOPS[2]
Peak FP16 Tensor Core[1]	1000 TFLOPS | 2000 TFLOPS[2]
Peak BF16 Tensor Core[1]	1000 TFLOPS | 2000 TFLOPS[2]
Peak FP8 Tensor Core[1]	2000 TFLOPS | 4000 TFLOPS[2]
Peak INT8 Tensor Core[1]	2000 TOPS | 4000 TOPS[2]

Table 1. NVIDIA H100 Tensor Core GPU preliminary performance specs

Mixed-Precision Wave Propagation

Physics vs. Limitations of FP16

Discretization of the Elastic Wave Equation

$5 \times 10^{-7}\ \mathrm{s.m^3.kg^{-1}}$ (subnormal value for FP16)

$$v_x = v_x + \frac{\Delta t}{\rho}(D_x\sigma_{xx} + D_y\sigma_{xy} + D_z\sigma_{xz})$$

$$\sigma_{xx} = \sigma_{xx} + \Delta t(\lambda + 2\mu)D_x v_x + \Delta t\ \lambda\left(D_y v_y + D_z v_z\right) + \Delta t\ s_{xx}$$

24×10^6 Pa.s (overflow in FP16)

- $\mathrm{v_p} = 3500\ \mathrm{m.s^{-1}}$
- $\mathrm{v_s} = 2000\ \mathrm{m.s^{-1}}$
- $\rho = 2000\ \mathrm{kg.m^{-3}}$
- $\Delta \mathrm{t} = 0.001\ \mathrm{s}$

Data Type	Exponent	Mantissa	Min (normal)	Min (subnormal)	Max
FP16	5	10	$\mathbf{6.104 \times 10^{-5}}$	5.961×10^{-8}	$\mathbf{65,504}$
FP32	8	23	1.176×10^{-38}	1.401×10^{-45}	3.403×10^{38}

Rescaling the Wave Equation

Compute normalized quantities

$$\mathrm{S} = \max(s_{ij})$$

Maximum of source signal

$$\mathrm{C} = \max(C_{ij})$$

Maximum of stiffness coefficient

Scale wavefields, sources, and material parameters

$$\sigma'_{ij} = \frac{1}{\Delta t\,\mathrm{S}}\sigma_{ij}$$

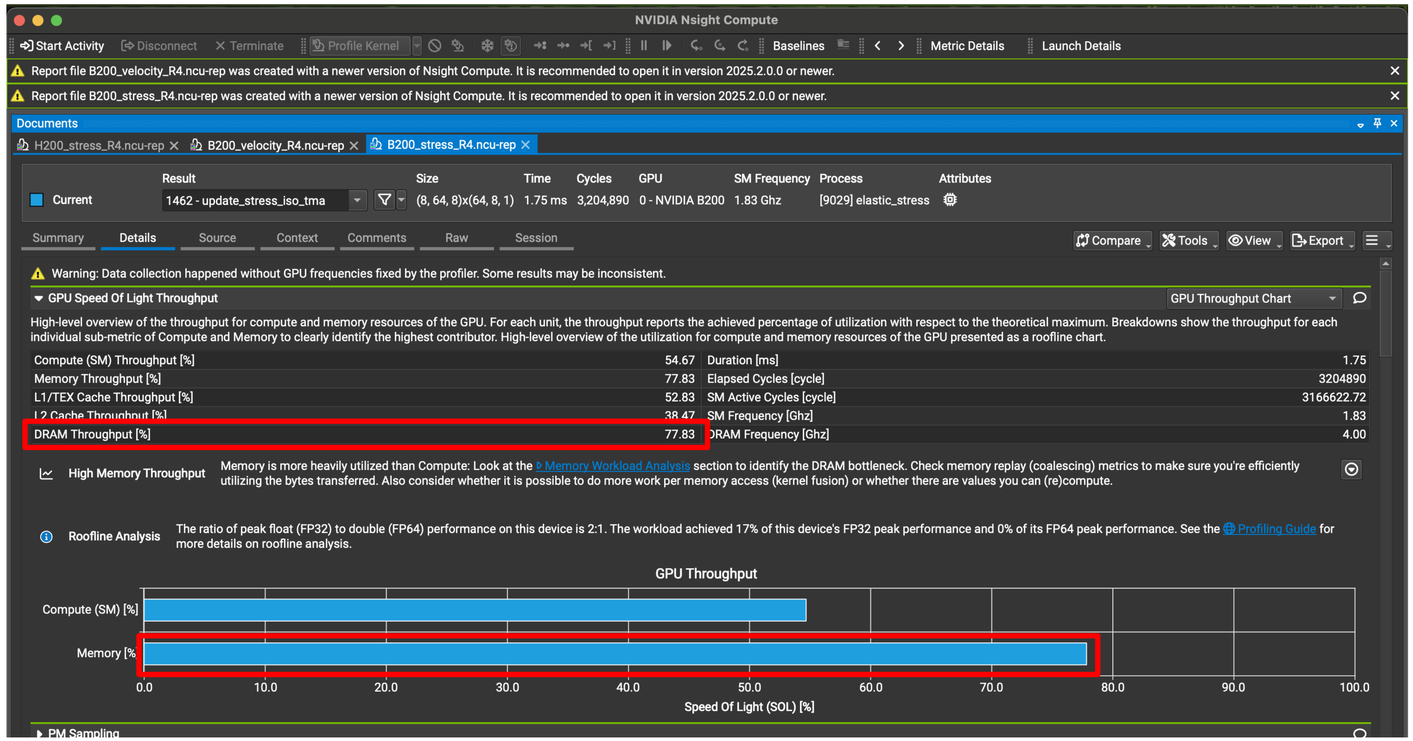
$$s'_{ij} = \frac{1}{\Delta t\,\mathrm{S}}s_{ij}$$

$$C'_{ij} = \frac{1}{\Delta t\,\mathrm{C}}C_{ij}$$

$$v'_{ij} = \frac{\mathrm{C}}{\mathrm{S}}v_{ij}$$

$$\rho'_{ij} = \frac{1}{\Delta t\,/\mathrm{C}}\rho$$

Motivation and Strategy (Memory Bound)



⚠ Elastic kernels are **Bandwidth Bound!**

GPU global memory / L2 cache

Wavefields (FP16)

Elastic models*

FP16 ↓ ↑ FP16

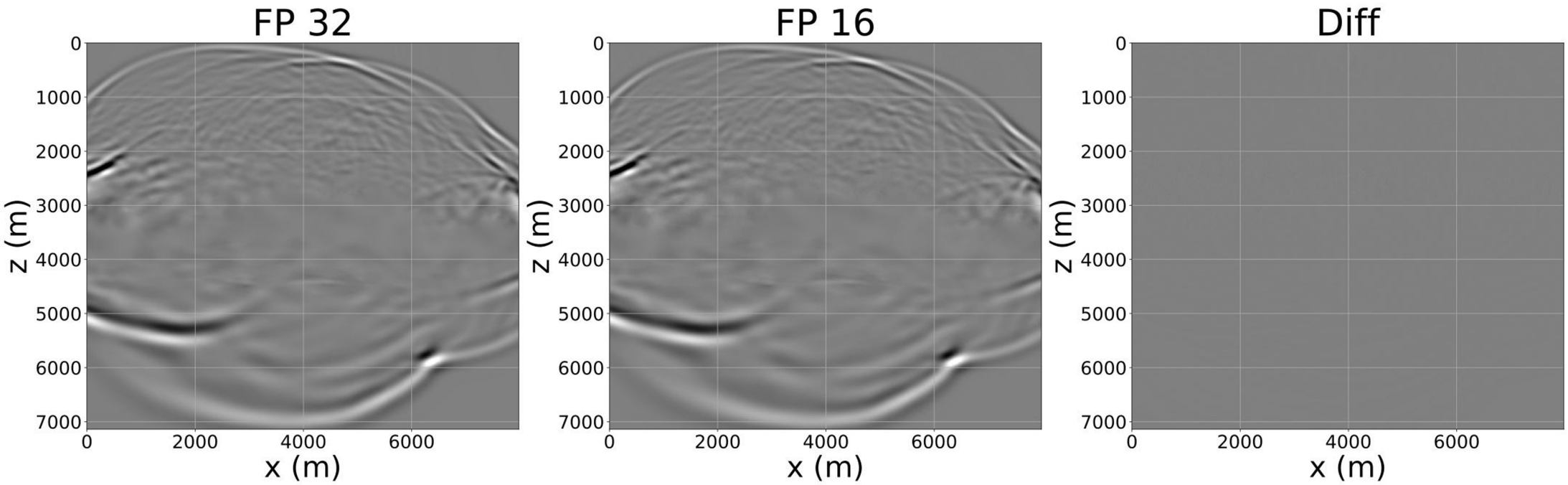
- Load wavefields and elastic models*
- Cast wavefields and elastic models* to FP32
- Update wavefield using FP32 arithmetic
- Store results in FP16

Kernel

** Elastic models may be stored in FP32 or compressed and stored in lower precision*

Experiment: Heterogeneous VTI

Results and Validation

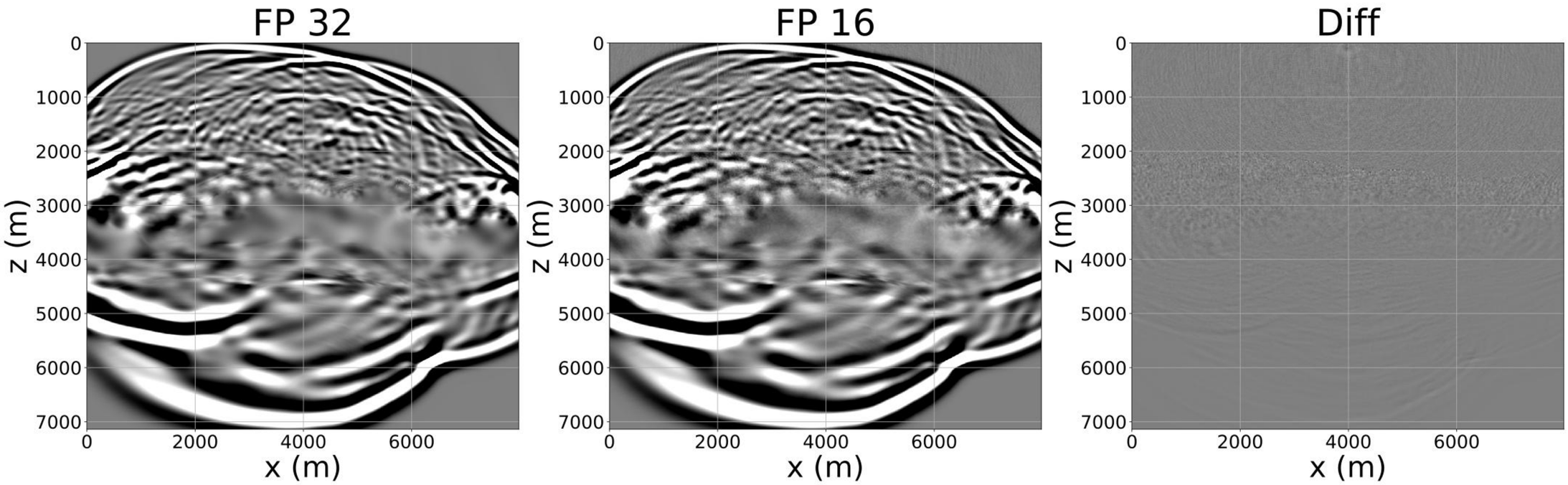


Pressure snapshots

Clip=10%

Experiment: Heterogeneous VTI

Results and Validation



Pressure snapshots

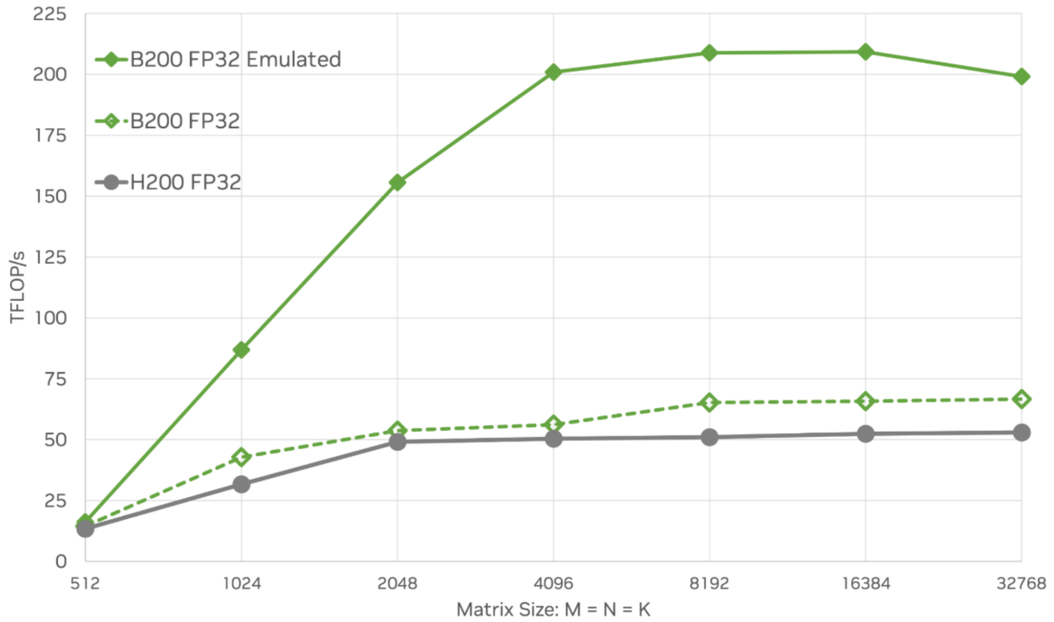
Clip=1%

- We maintained the geophysical accuracy required for modeling and FWI gradients
  - Reducing the memory footprint **by 50%** and unlocking a potential 2x speedup

Half-Precision Seismic Modeling in 3D Anisotropic Media

Floating Point Emulation Methods

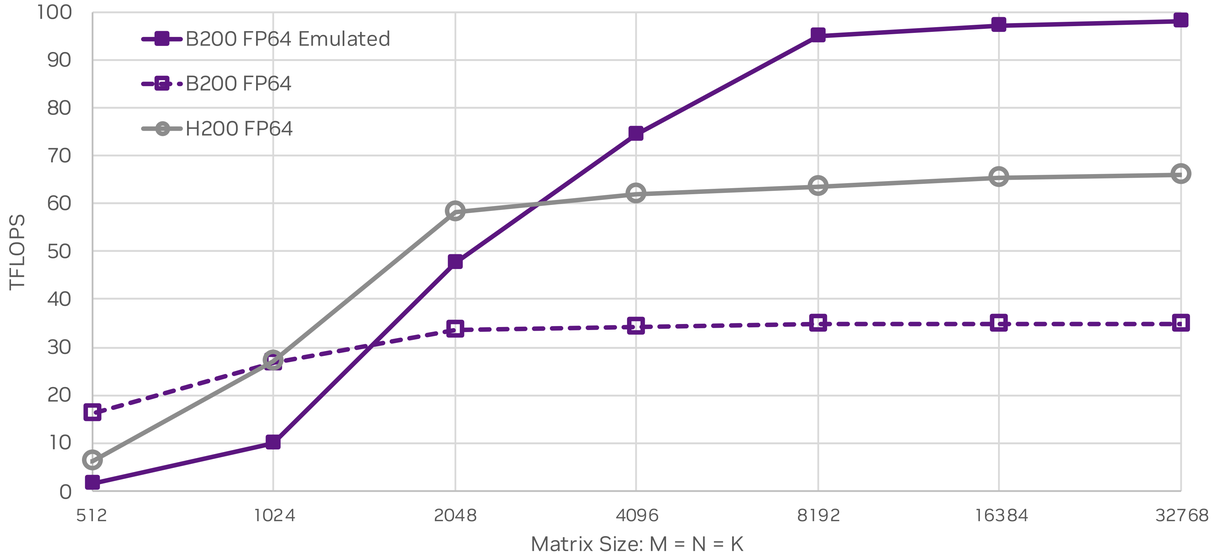
Floating Point Emulation Methods



Performance (TFLOP/s) of emulated FP32 compared to native FP32 on B200 and H200 GPUs

- **FP32 using BF16 Tensor Cores** [1]
  - Released with **CUDA 12.9 in cuBLAS** for Blackwell GPUs
  - Being tested for accuracy and performance impact in:
    - Weather, quantum circuit and condensed matter simulations
    - Dense Linear Algebra (QR, LU)
  - **Uses 9 inner matmuls in BF16 (BF16x9)**

- **FP64 using INT8 Tensor Cores** [2]
  - Just released with **CUDA 13.0.2 in cuBLAS** for Blackwell GPUs
  - Being tested for accuracy and performance impact in:
    - Materials Science, Electronic Structure
    - Molecular Dynamics, Computational Chemistry
    - HPL, Dense Linear Algebra (QR, LU)
  - **Uses variable number of inner matmuls in INT8**



FP64 Emulation using INT8 tensor cores (s=7 or 56 bits)

FP32 Blog Post and **FP64 Blog Post**

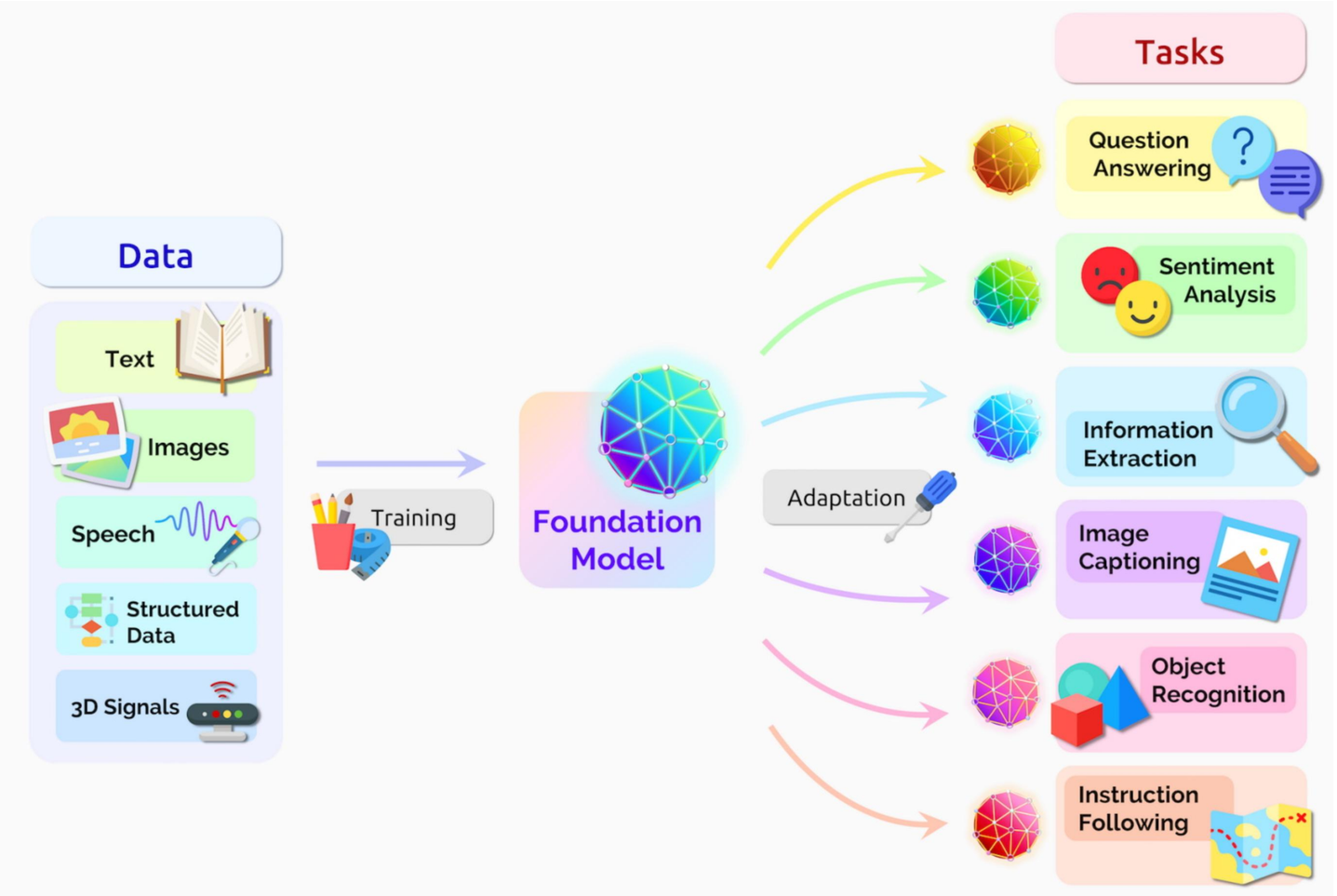
https://arxiv.org/pdf/2203.03341
https://arxiv.org/abs/2306.11975
https://arxiv.org/pdf/1904.06376
https://arxiv.org/abs/2409.13313

AI Perspective

What are Foundation Models

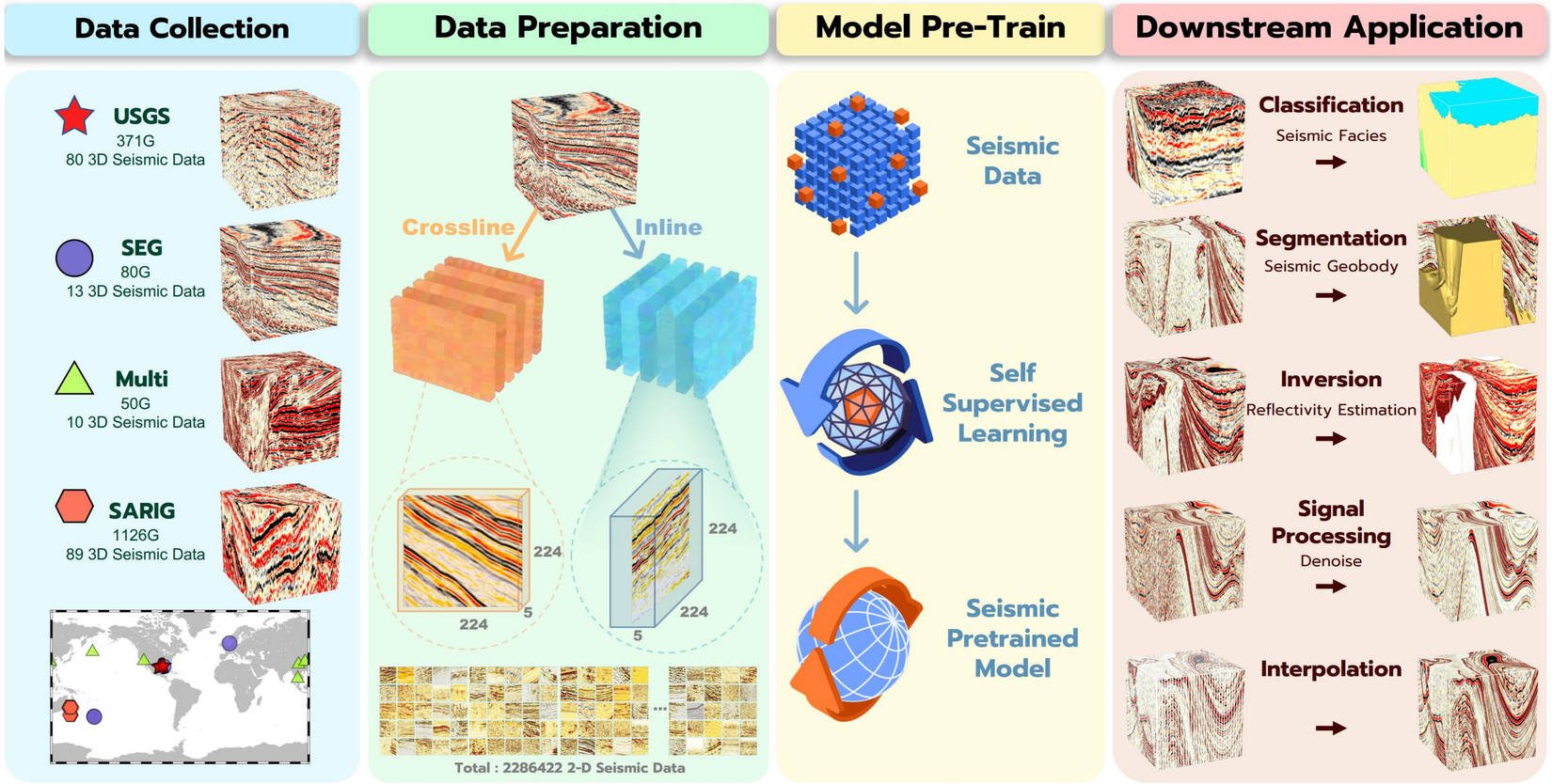
General purpose machine learning models

- Trained on massive unlabeled datasets to handle a wide variety of jobs from translating text to analyzing seismic images.
- General purpose adaptable models that can be fine-tuned for diverse tasks like computer vision, natural language processing with minimal task-specific training.
- Scalable and reusable architectures which leverage learned patterns to perform new tasks efficiently, reducing the need for extensive retraining or custom model development.



Source: What Are Foundation Models? | NVIDIA Blogs

Seismic Foundation Model (SFM)

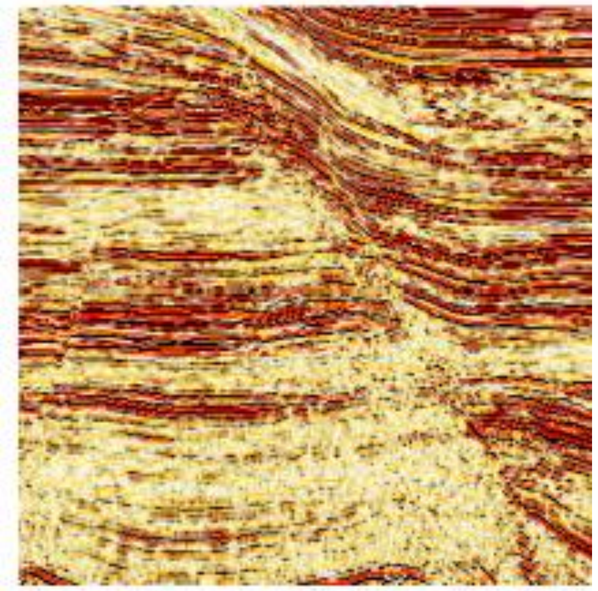


SFM represents the first successful large-scale foundational model trained exclusively within the seismic domain [reference].

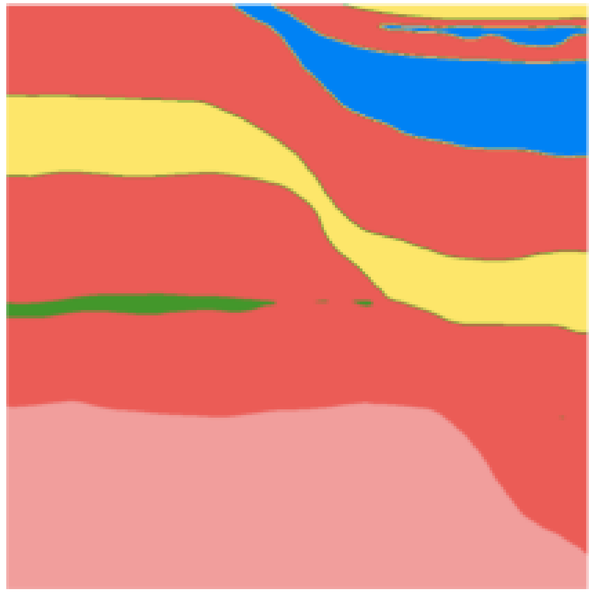
Masked Auto-Encoder fine-tuned on 200 samples

Trained with self-supervised learning on 2.2 million seismic images utilizing NVIDIA TAO

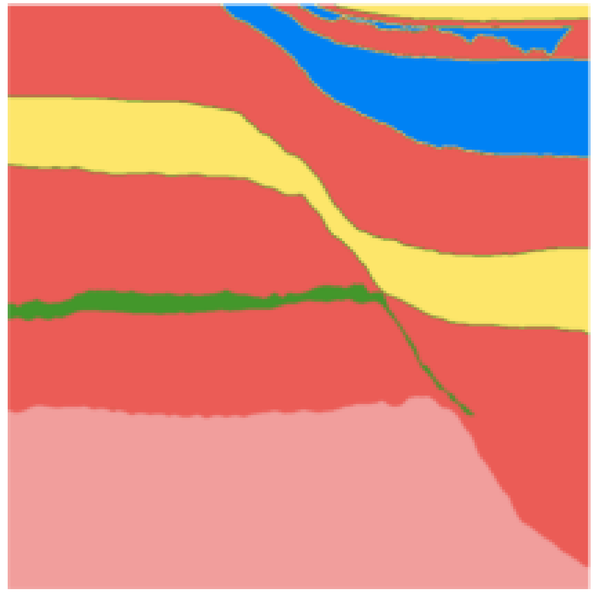
Input Seismic



MAE



Label

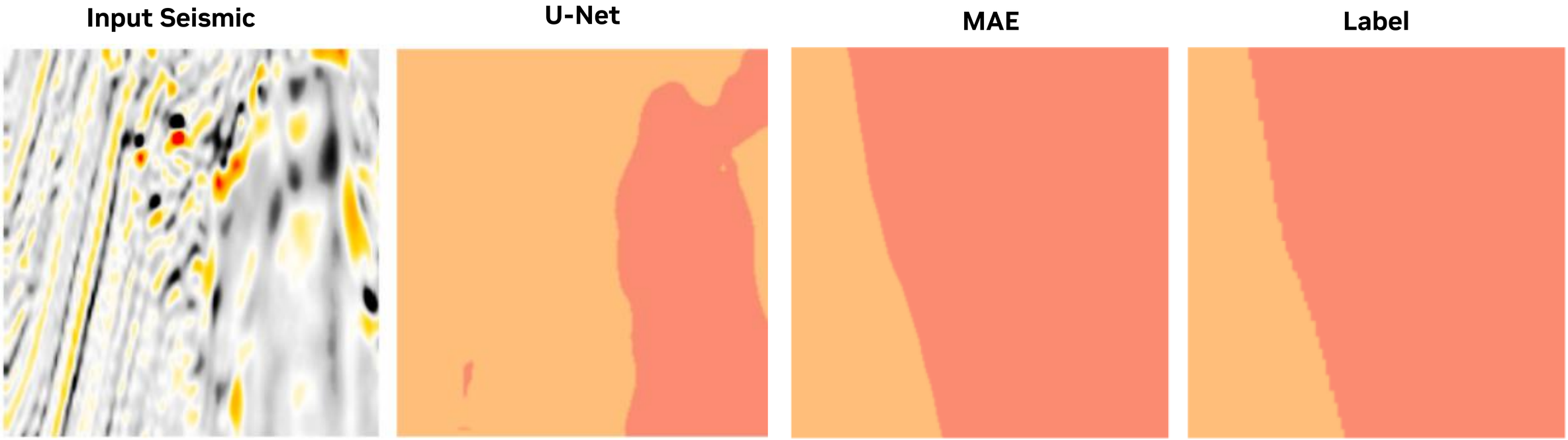


Facies classification

(Results reproduced from Sheng et al. , 2025)
“Seismic Foundation Model (SFM): a new generation deep learning model in geophysics

Masked Auto-Encoder fine-tuned on 3000 samples

Trained with self-supervised learning on 2.2 million seismic images utilizing NVIDIA TAO



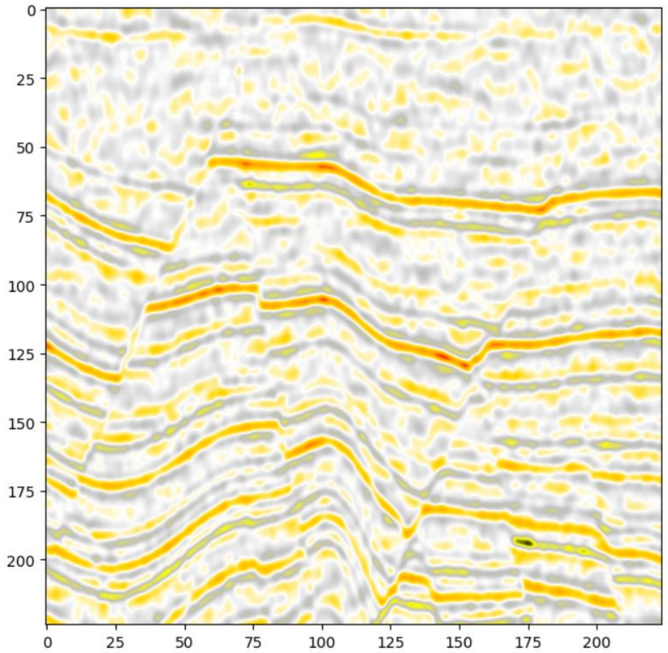
Salt Segmentation

(Results reproduced from Sheng et al. , 2025)
“Seismic Foundation Model (SFM): a new generation deep learning model in geophysics

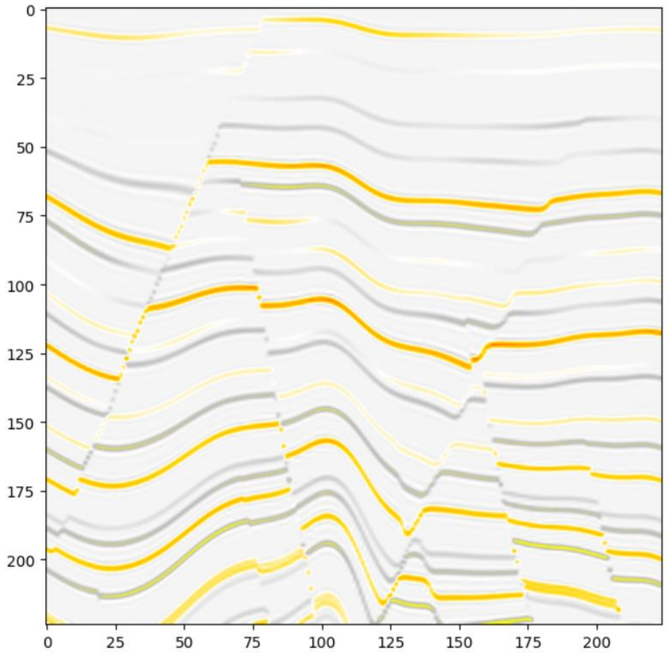
Masked Auto-Encoder fine-tuned on 1000 samples

Trained with self-supervised learning on 2.2 million seismic images utilizing NVIDIA TAO

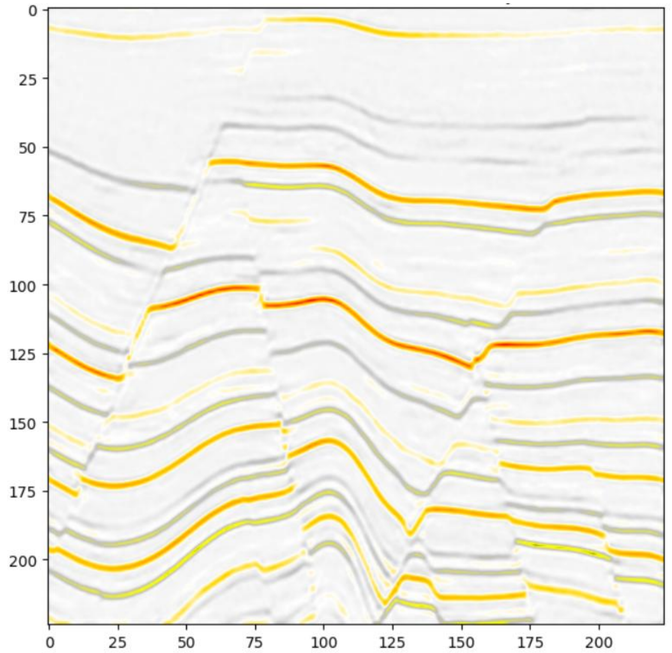
Input Seismic



MAE



Label



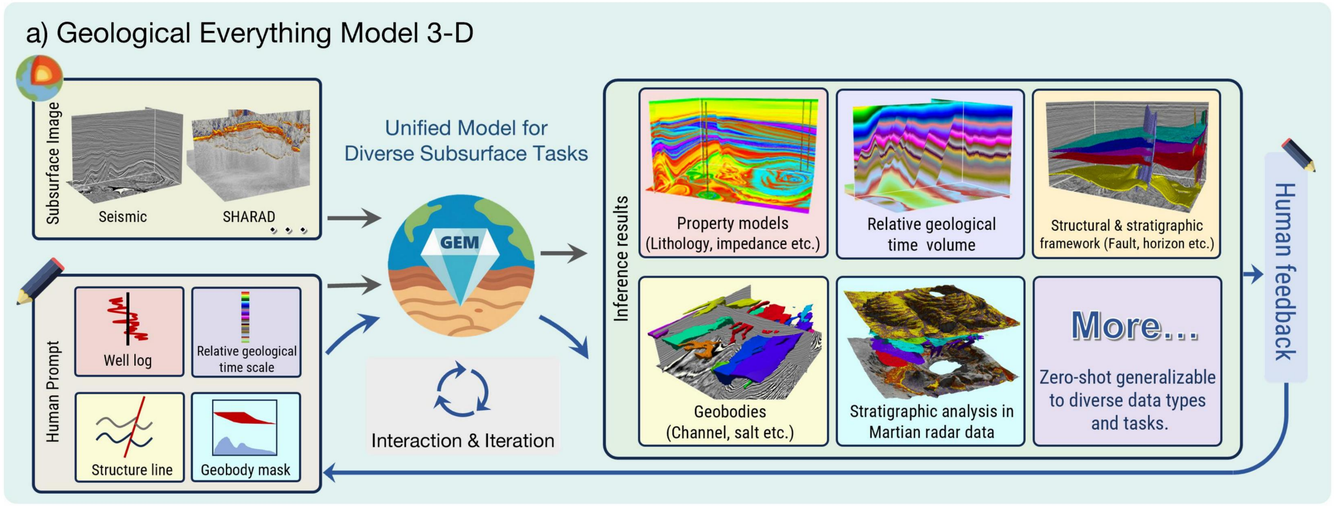
Reflectivity inversion

(Results reproduced from Sheng et al. , 2025)

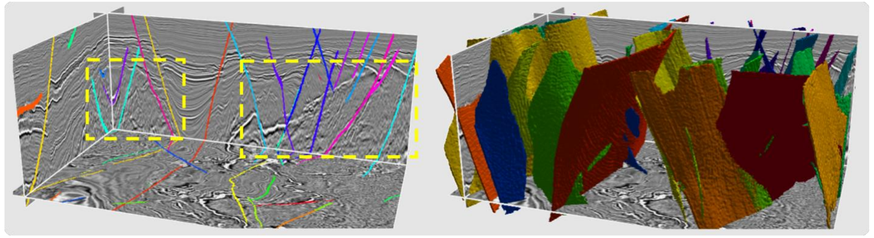
“Seismic Foundation Model (SFM): a new generation deep learning model in geophysics

Geological Everything Model 3D

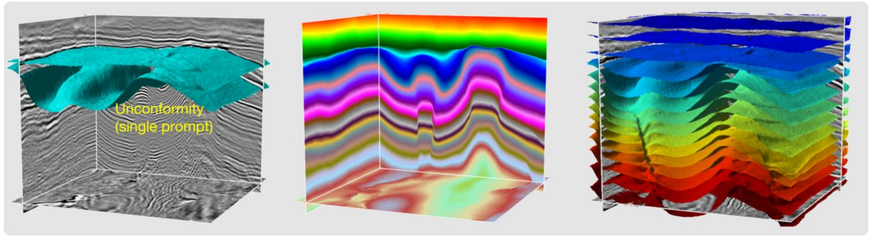
A Promptable Foundation Model for Unified and Zero-Shot Subsurface Understanding



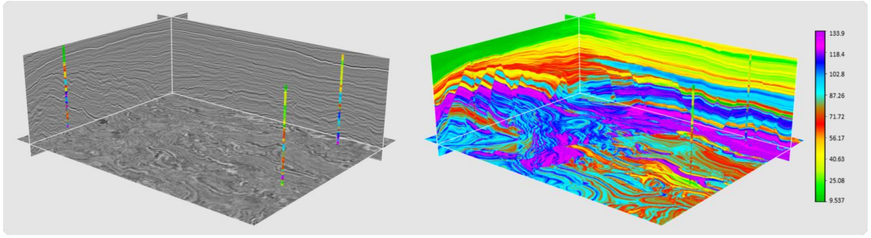
Baiyun Fault



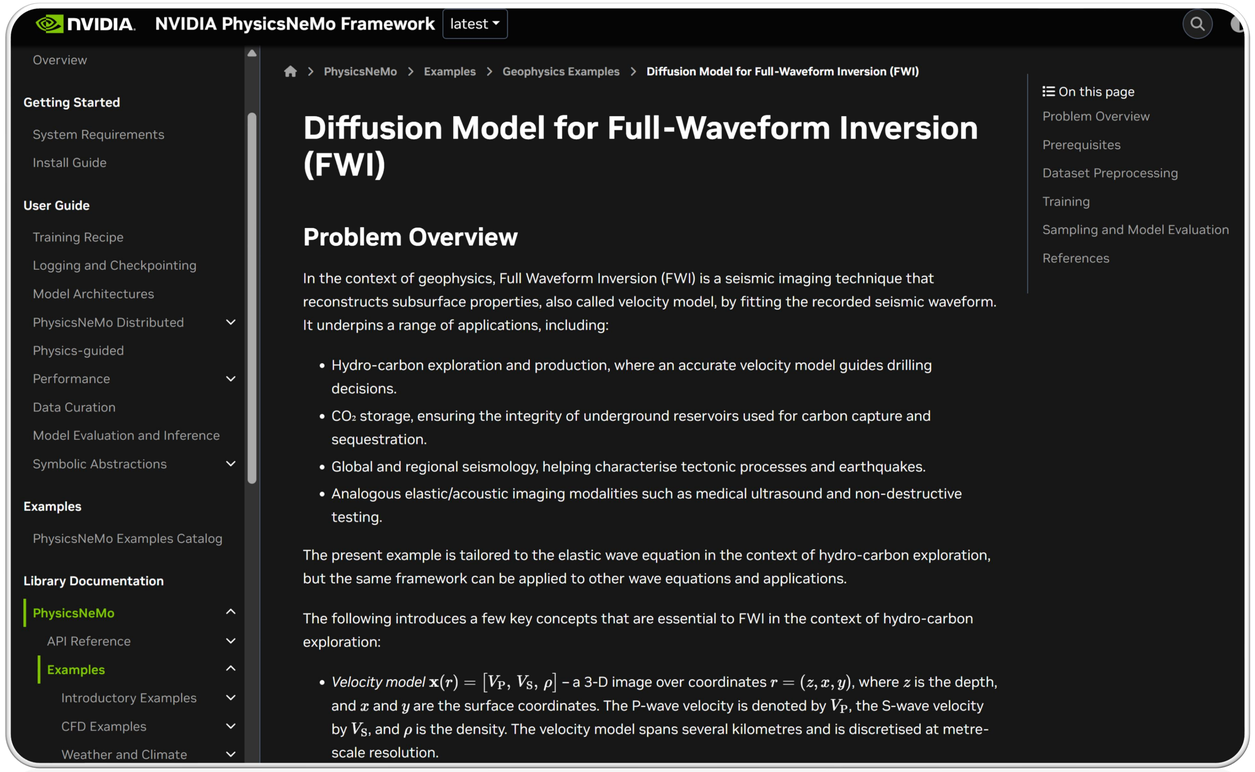
Delft Unconformity and RGT

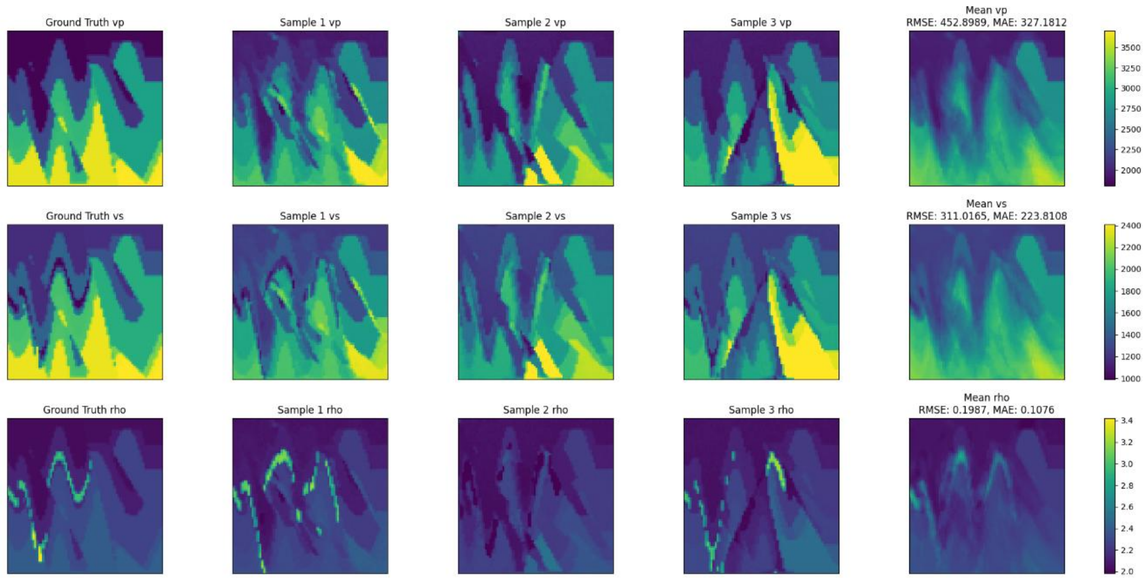


Poseidon GammaRay



Try Diffusion Model for Full-Waveform Inversion Using NVIDIA PhysicsNeMo





https://docs.nvidia.com/physicsnemo/latest/physicsnemo/examples/geophysics/diffusion_fwi/README.html