# New Architectures for Seismic Data Processing

Date: July, 2020

# 3 Key Oil & Gas Activities – Exploration, Production & Prediction

## Exploration - Seismic Data Processing

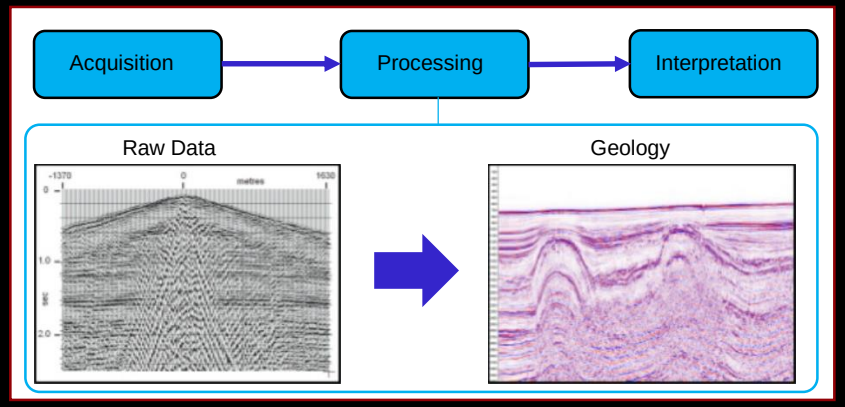
- Where and How Much is the Oil & Gas?
  - ✓ Build accurate HD earth Subsurface Models
  - ✓ Interpret the Models Automatically

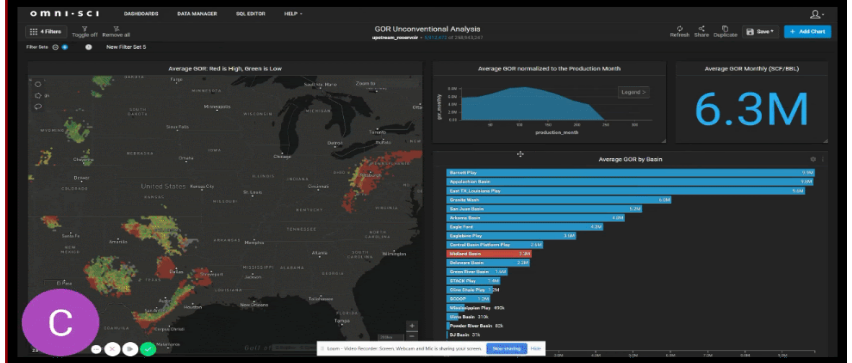
## Production - Drill Operations

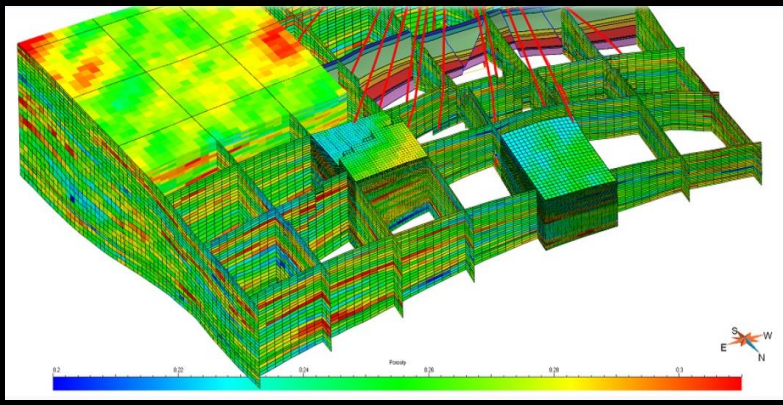
- How are the wells going in Real-Time?
  - ✓ Well Operation & equipment status Monitoring
  - ✓ Predictive Maintenance to avoid operation disruptions

## Prediction - Reservoir Simulation

- How Much Oil & Gas Left & How would the reservoir be changing?



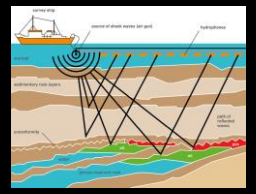
https://www.omnisci.com/blog/the-next-wave-of-oil-and-gas-analytics-reservoir-behavior



FUTUREWEI Technologies

# Monetize the New "Oil" with FWI & Deep Learning

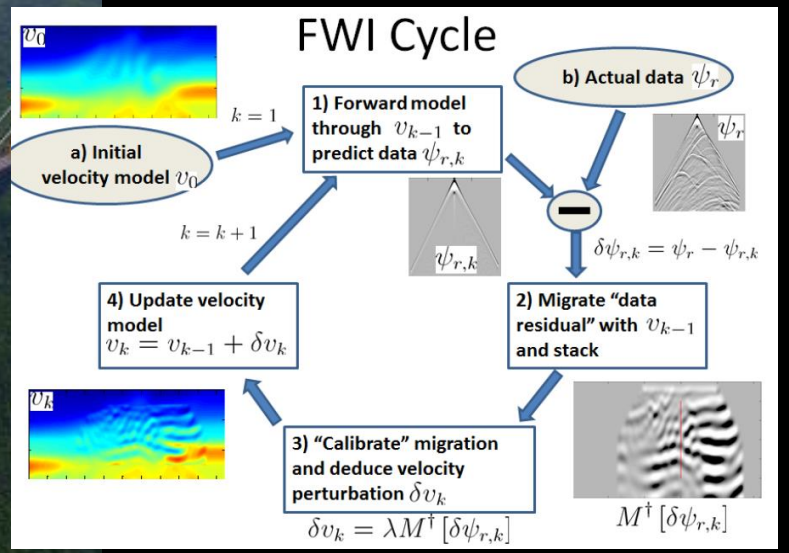
## Market Opportunities



$9.28B Seismic Survey Market by 2022, $11.8B by 2025
(ResearchandMarkets, 2017)

$122.6B Global Oil & Gas Analytics Market by 2025
(Zion Market Research 2019)

$22B Oil&Gas Analytics Market by 2025
(Brandessence Research, 2020)

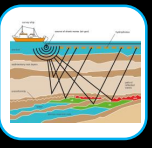



- ✓ FWI (Full-Waveform Inversion) to Seismic Data Processing is DL(Deep Learning) to AI
- ✓ FWI algorithms are not new, but recent progress in advanced & affordable computing power makes 3D/4D FWI imaging & Processing feasible
- ✓ FWI is a process that grows **10X** in compute with every frequency Doubling (DUG, 2019)
- ✓ **DL** is being investigated to reduce FWI computing requirements

## Big Data about Oil & Gas Exploration

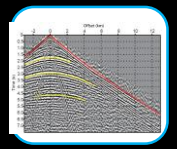
80,000 Sensors 15PB / Life Per on-shore well

>1.3M Active Oil & Gas Wells in USA
(EIA, 2019)



200+TB Data in a 24000 km² survey
(DUG, 2018)

$5M-$8M / onshore
$100M-200M / Offshore Oil Well(USEIA, 2016)



32b FP & FWI SEG-Y In Seismic Data Processing

5000-35000ft Well Depth

Source: FWI Cycle from "A Perspective on Full-Waveform Inversion" by Gary F. Margrave, Kris Innanen, and Matt Yedlin, 2012



HUAWEI

# 3-Step of Seismic Data Processing

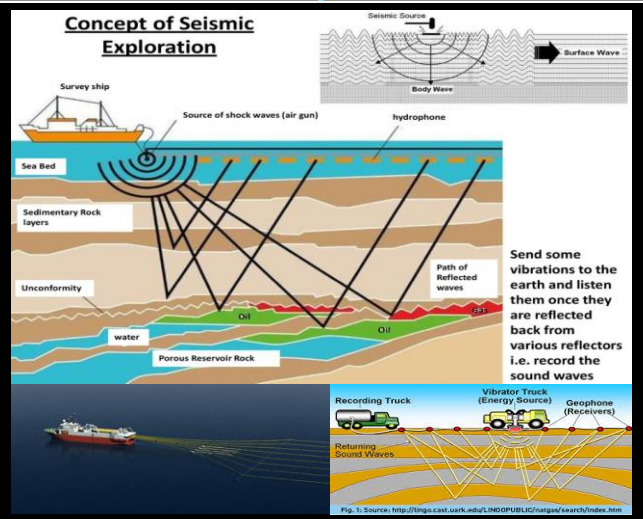
- ✓ High-Frequency FWI is the Game Changer in Seismic Data Processing, the same as DL to AI
- ✓ FWI is an extreme computation-intensive process growing 10X in Compute with Every frequency Doubling
- ✓ Advanced Data-Driven Deep-Learning could play key roles in Seismic Data Processing and Interpretation

✓ 2D/3D/4D Seismic Wave Data in SEG-Y
✓ 100s TB data per Survey

✓ High-Frequency FWI is the Game Changer!
✓ FWI is a process that grows 10X in compute with Every frequency Doubling
✓ 32b Single-Precision Floating Point Operations
✓ Deep-Learning may play a crucial role in FWI

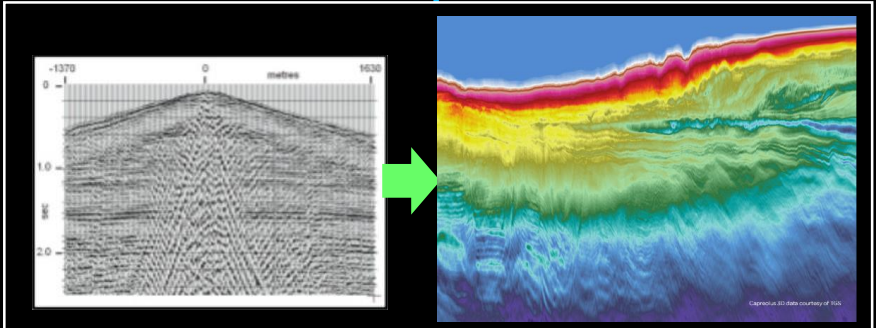
✓ Human, Legacy Computer Vision or Deep-Learning
✓ Advanced Deep-Learning could play a key role for Automated Seismic Data Interpretation

Acquisition → Processing → Interpretation → Reservoir Simulation



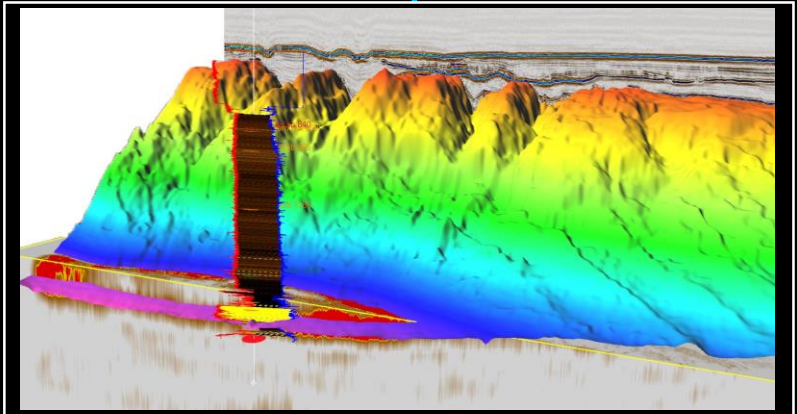

✓ To Collect Seismic Wave Data

Source: https://slideplayer.com/slide/13787934/
https://fluidpowerjournal.com/repairing-hydraulic-valves/



✓ Remove Noises and Transform the Raw SEG-Y Seismic Wave data into Layered HD Velocity Models
✓ Digital Signal Processing Algorithms, Physics-Driven FWI, Data-Driven FWI
✓ High-Frequency FWI is the Game Changer

https://dug.com/dug-geo/full-waveform-inversion-fwi/



✓ To Determine where, how deep and How Much the Oil & Gas for Where to Drill

https://www.geolink-s2.com/expertise/3d-seismic-interpretation-training

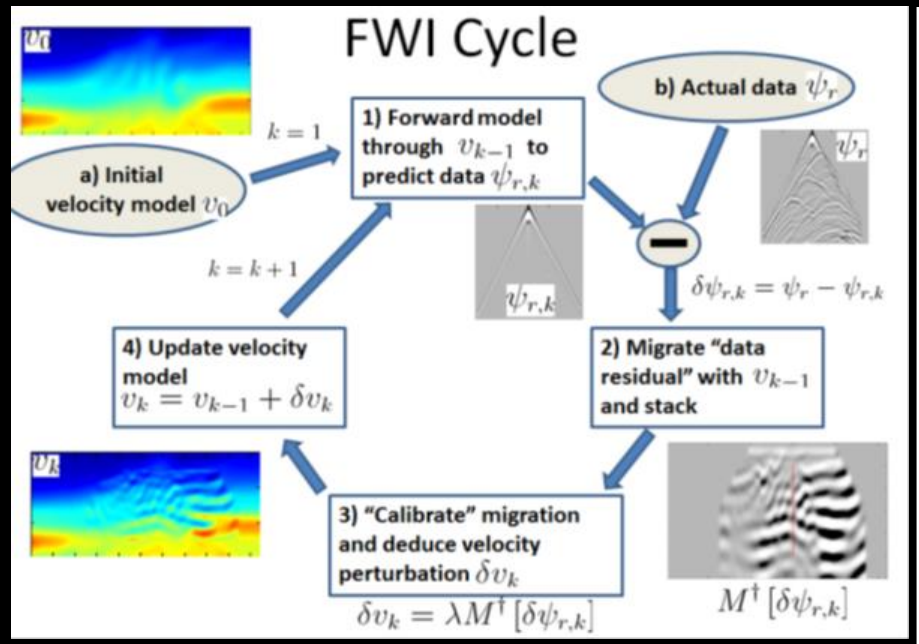


HUAWEI

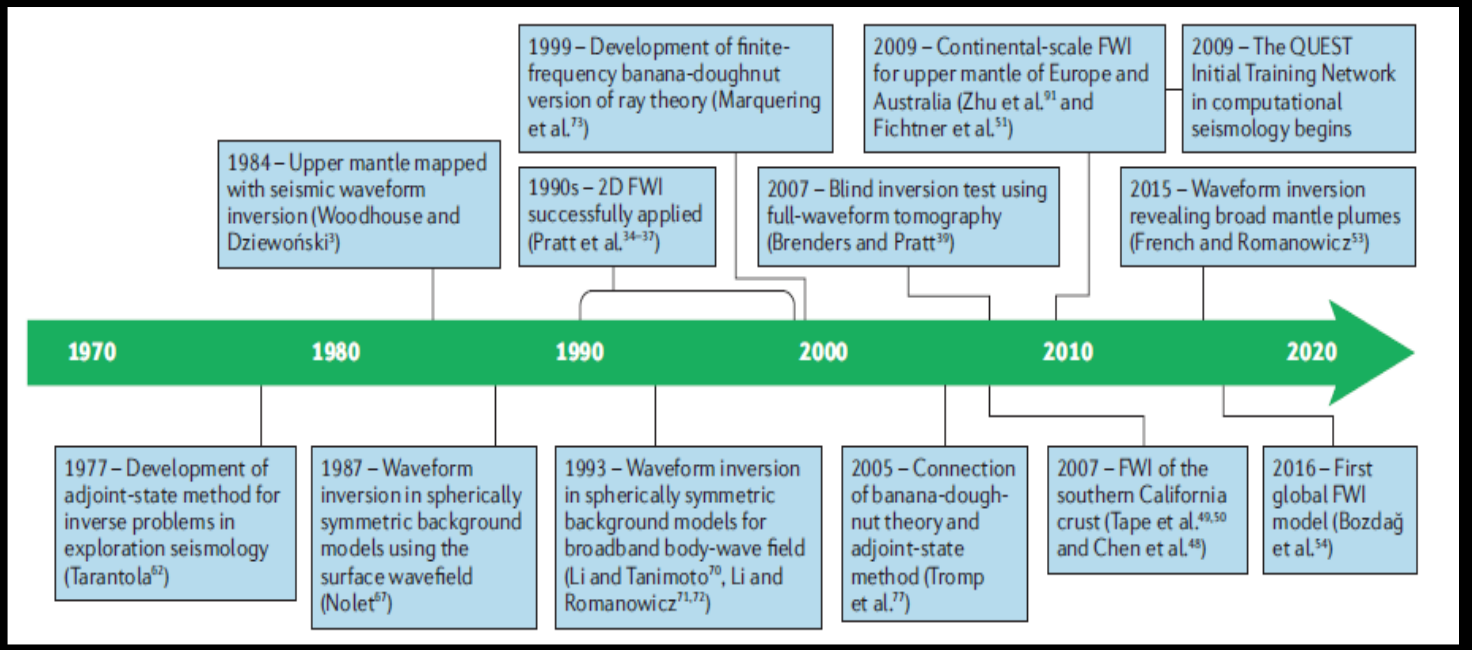
# FWI: The Game Changer for Seismic Data Processing

✓ FWI Algorithms are not New, but recent Computing Power Advancement Making it a Reality, Just as DL to AI

- An extremely complex process of developing a high-quality, three-dimensional (3D/4D) model of the Earth's subsurface
  - ① Perform a seismic survey and record the reflected sound waves from hundreds of thousands of sound impulses made on the Earth's surface
  - ② Process the collected survey data to remove various types of coherent and random noise, leaving just the primary signal of interest.
  - ③ Locate the reflectors in the data spatially into a 3D volume using an earth model, producing images of the reflectors in the subsurface
  - ④ From this data set, update the earth model and perform the imaging again and again to obtain a very accurate earth model and correct placement of the reflecting geological layers beneath the Earth's surface




Source: FWI Cycle from "A Perspective on Full-Waveform Inversion" by Gary F. Margrave, Kris Innanen, and Matt Yedlin, 2012



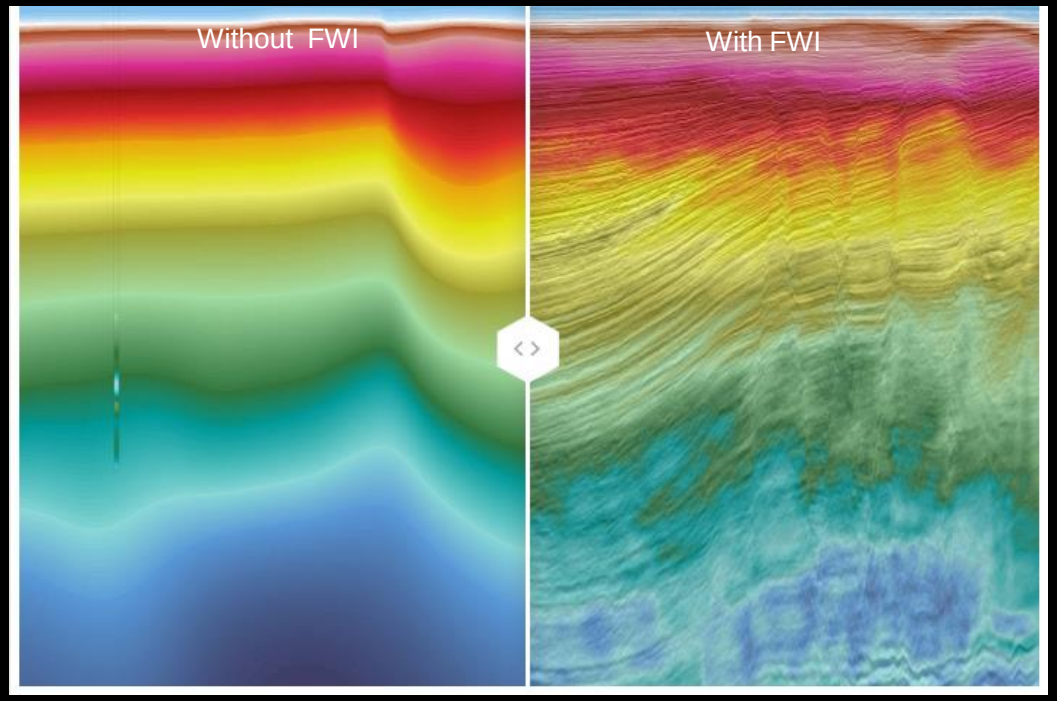

Source: "Seismic wavefield imaging of Earth's interior across scales", by Jeroen Tromp, Nature Reviews Earth & Environment volume, 2020



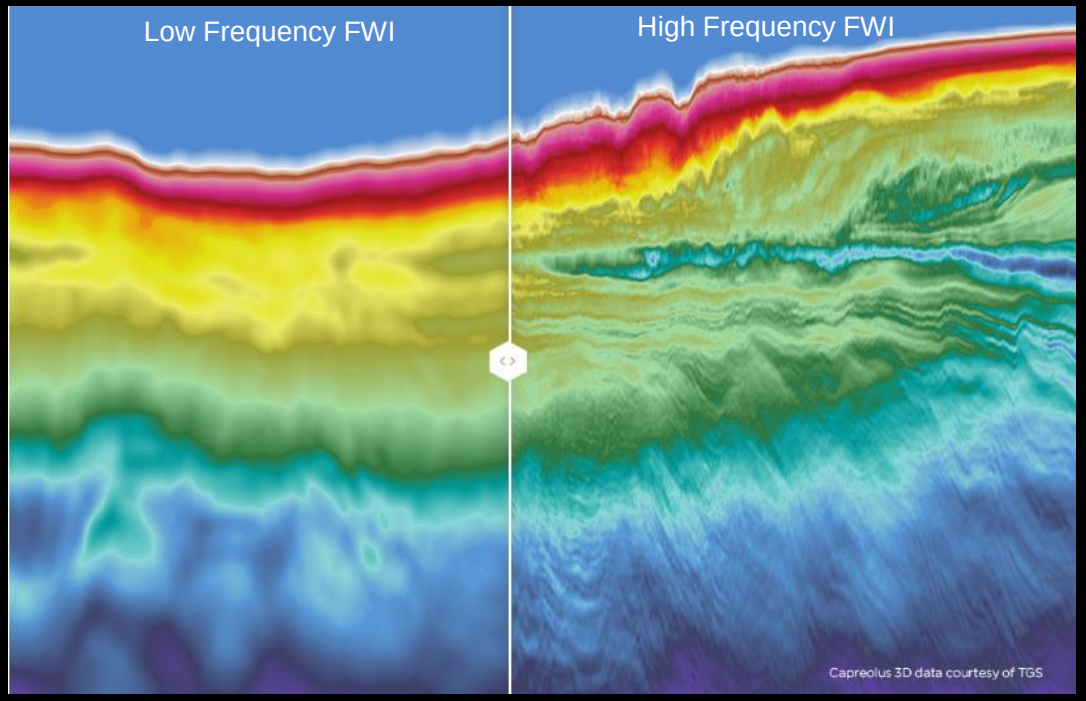
HUAWEI

# High-Frequency FWI is the Game Changer!

- The Goal of FWI: use the entire wavefield, including refractions and reflections, primaries and multiples, to generate a refined, high resolution Earth model.
- λ/2: The highest subsurface resolution provided by FWI, and the speeds of sound wave travelling through underground subsurfaces vary from ~300m/s to 6000+m/s with subsurface media. The higher the FWI frequency, the sharper the structure details:
  - For 100Hz and 3500m/s, the resolution is 3500 / 200 = 17.5m – the subsurface with thickness of larger than 17.5m could be identified
  - The FWI technology has been mainly used for up to 10-12Hz (resolution of 175m-145m @ 3500m/s) due to Computing power constraint
  - DUG states: With 100 Hz FWI, it is OK to get out of the model basically everything that was recorded in that wavefield




Before and after FWI. Smooth starting velocity model prior to FWI (left) and after FWI, co-rendered with the seismic data (right). (Data courtesy of Shell NZ).




Conventional low frequency FWI model and high frequency FWI model. Note the increasingly sharp stratigraphic and structural details in the high frequency model. (Capreolus 3D data courtesy of TGS)

Source: https://dug.com/dug-geo/full-waveform-inversion-fwi/

FUTUREWEI Technologies

# FWI: Physics-Driven vs. Data-Driven

Prevailing

**Physics-Driven Full Wave Inversion**

- ✓Inferring the layered velocity models by solving the Computation-intensive equations with the large collected survey data
- ✓Seismic waves are mechanical perturbations that travel in the medium at a speed governed by the acoustic/ elastic impedance of the medium in which they are traveling as the following acoustic-wave Equation

$$\left[\frac{1}{K(\mathbf{r})}\frac{\partial^2}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho(\mathbf{r})}\nabla\right)\right]p(\mathbf{r},t) = s(\mathbf{r},t),$$

Where ρ(r)is the density at spatial location r, K(r) is the bulk modulus, s(r; t) is the source term, p(r; t) is the pressure wavefield, and t represents time.

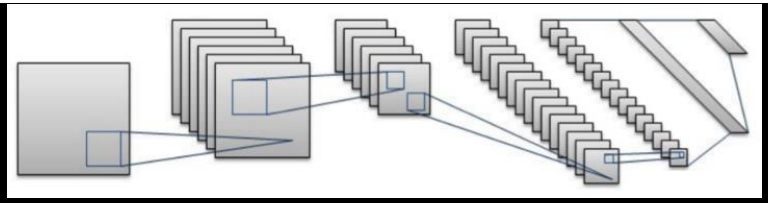
Leverage the FP32 High Computing Performance provided by SIMD Vectors or Tensor Cores and Massively Distributed Parallel Processing HPC System

Emerging/Exploring

**Data-Driven Full Wave Inversion (from data to model)**

- ✓Inferring the layered velocity models by using Deep-Learning Neural Network Model & known Seismic model data, but without solving Computation-Intensive Equations
- ✓Learning the Velocity Models from massive known datasets

$$\text{Seismic Measurements} \xrightarrow{f^{-1}} \text{Velocity Models.}$$



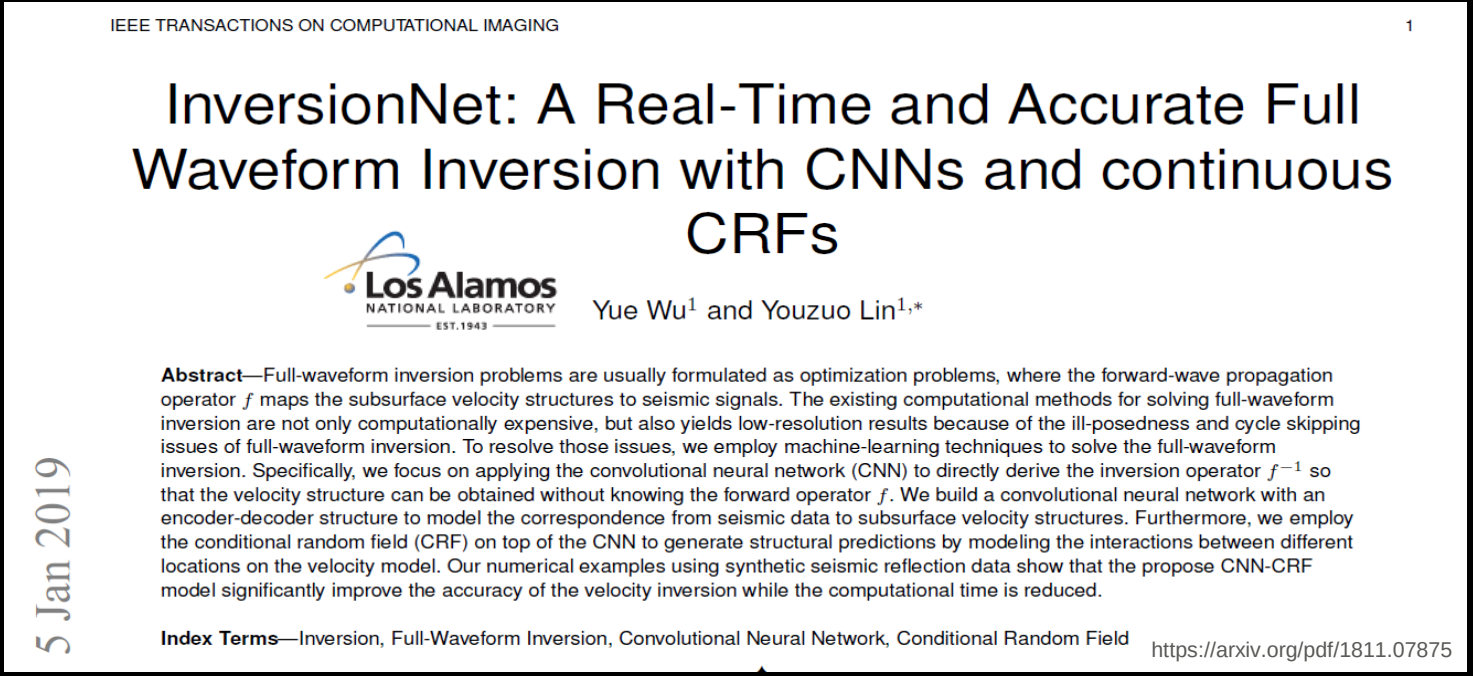
Leverage the Massive Tensor Core 16b FP/8b INT FMA Performance, typically 10X of FP32, if OK

https://arxiv.org/pdf/1811.07875

FUTUREWEI Technologies

# Deep-Learning for Seismic Data Processing

- ✓ Yue Wu & Youzuo Lin from Los Alamos National Lab (LANL) proposed the InversionNet by using Data-Driven Deep Learning technologies to overcome the drawbacks of "Expensive computation and Low-resolution results due to the ill-posedness and cycle skipping issues of Physics-Driven Full-Waveform Inversion(FWI) Methods"
- ✓ InversionNet uses convolutional neural network (CNN) to directly derive the inversion operator $f^{-1}$ so that the velocity structure can be obtained without knowing the forward operator f, not only yields accurate inversion results but also significantly improves the computational efficiency
- ✓ Microsoft has been developing the Open Source DeepSeismic to interpret the Seismic data processing results/images by leveraging state-of-the-art segmentation algorithms (UNet, SEResNET, HRNet) for seismic interpretation on a GPU server



IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING 1

## InversionNet: A Real-Time and Accurate Full Waveform Inversion with CNNs and continuous CRFs

Los Alamos NATIONAL LABORATORY EST.1943

Yue Wu[1] and Youzuo Lin[1,*]

**Abstract**—Full-waveform inversion problems are usually formulated as optimization problems, where the forward-wave propagation operator $f$ maps the subsurface velocity structures to seismic signals. The existing computational methods for solving full-waveform inversion are not only computationally expensive, but also yields low-resolution results because of the ill-posedness and cycle skipping issues of full-waveform inversion. To resolve those issues, we employ machine-learning techniques to solve the full-waveform inversion. Specifically, we focus on applying the convolutional neural network (CNN) to directly derive the inversion operator $f^{-1}$ so that the velocity structure can be obtained without knowing the forward operator $f$. We build a convolutional neural network with an encoder-decoder structure to model the correspondence from seismic data to subsurface velocity structures. Furthermore, we employ the conditional random field (CRF) on top of the CNN to generate structural predictions by modeling the interactions between different locations on the velocity model. Our numerical examples using synthetic seismic reflection data show that the propose CNN-CRF model significantly improve the accuracy of the velocity inversion while the computational time is reduced.

**Index Terms**—Inversion, Full-Waveform Inversion, Convolutional Neural Network, Conditional Random Field

5 Jan 2019

https://arxiv.org/pdf/1811.07875

DEEPSEISMIC

DEEP LEARNING FOR SEISMIC IMAGING AND INTERPRETATION

https://github.com/microsoft/seismic-deeplearning

https://en.wikipedia.org/wiki/Comparison_of_free_geophysics_software



FUTUREWEI Technologies

# InversionNet: Replacing Physics-Driven FWI with Data-Driven



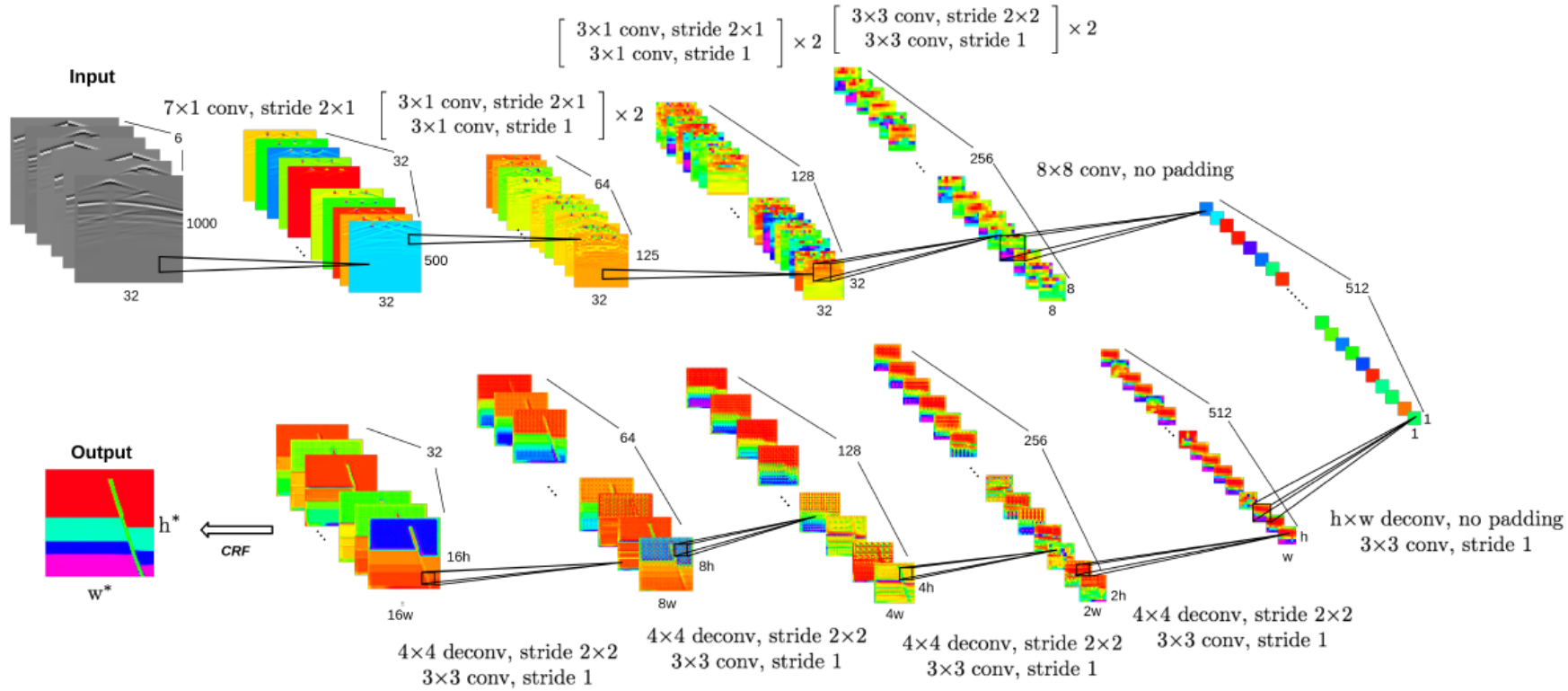



Fig. 1: An illustration of the proposed framework of the data-driven model. The CNN has an encoder-decoder architecture. The encoder (the top pipeline) is primarily built with convolution layers, which extract high-level features from the input seismic data and compress them into a single high-dimensional vector. The decoder (the bottom pipeline) then translates those features into velocity models through a set of deconvolution layers. The specification of each layer is provided in the figure. We build a locally connected CRF on top of the final feature map to generate the final predictions.

https://arxiv.org/pdf/1811.07875

# DownUnder GeoSolutions (DUG) HPC for Seismic Processing

- ✓ A leading global geoscience company focusing on Seismic Data Processing HPC services, located in Perth, Australia, and having HPC data centers in Perth (Australia), London (UK), Kuala Lumpur (Malaysia) and Houston (USA)
- ✓ Key focus: 32-bit SP and A lot of cores per Socket, no need for global MPI communications, MPI only with a few dozens nodes

## DUG Bubba HPC System in Houston

- Max 250 32-bit SP Petaflops w/ 40,000 Nodes of Intel Xeon Phi (each 7250 has 68 cores and 16GB High Bandwidth MCDRAM for 6.1 TeraFLOPS SP)
- High Efficiency Water cool for low PUE of 1.03
- Geophysical layer written in Python
- 4x Xeon Phi nodes in a 2U Box sharing a 50GE CX4 in Socket-Direct/Multi-host Mode, and 200 Nodes on one Mellanox Spectrum ASIC/Leaf Switch: 50 Downlink Ports and up to 14 Uplink Ports
- Each Node consumes about 10MB/s-30MB/s Network Bandwidth

## DUG Insight software stack

- Including modules for loading and managing data, seismic preprocessing, regularization, time and depth imaging, seismic inversion, rock physics, and high frequency FWI
- 32b SP is being used, no use of FP64 & FP16

**Full Waveform Inversion**

Literally the best software stack in the business. Perfect for R&D as well as production. Solutions for loop skipping, anisotropy, absorption, reflections, high contrast models, elastic models, and land. Super high frequency - no problem!

**Seismic Data Processing**

Streamer, land and OBN; 2D, 3D and 4D; State of the art toolkit all written this century - most modern toolkit available.

**Depth Imaging**

From fault constrained, high resolution reflection tomography and FWI, through to amplitude preserving Kirchhoff migrations and least squares RTM.

**Basin-wide Velocity Model Building**

As big as you like, no limit to the number of 2D and 3D seismic surveys. Tying all wells, what more needs to be said?

**Petrophysical Processing and Interpretation**

From operational to regional studies. All remedial and interpretive work undertaken.

**Quantitative Interpretation Services**

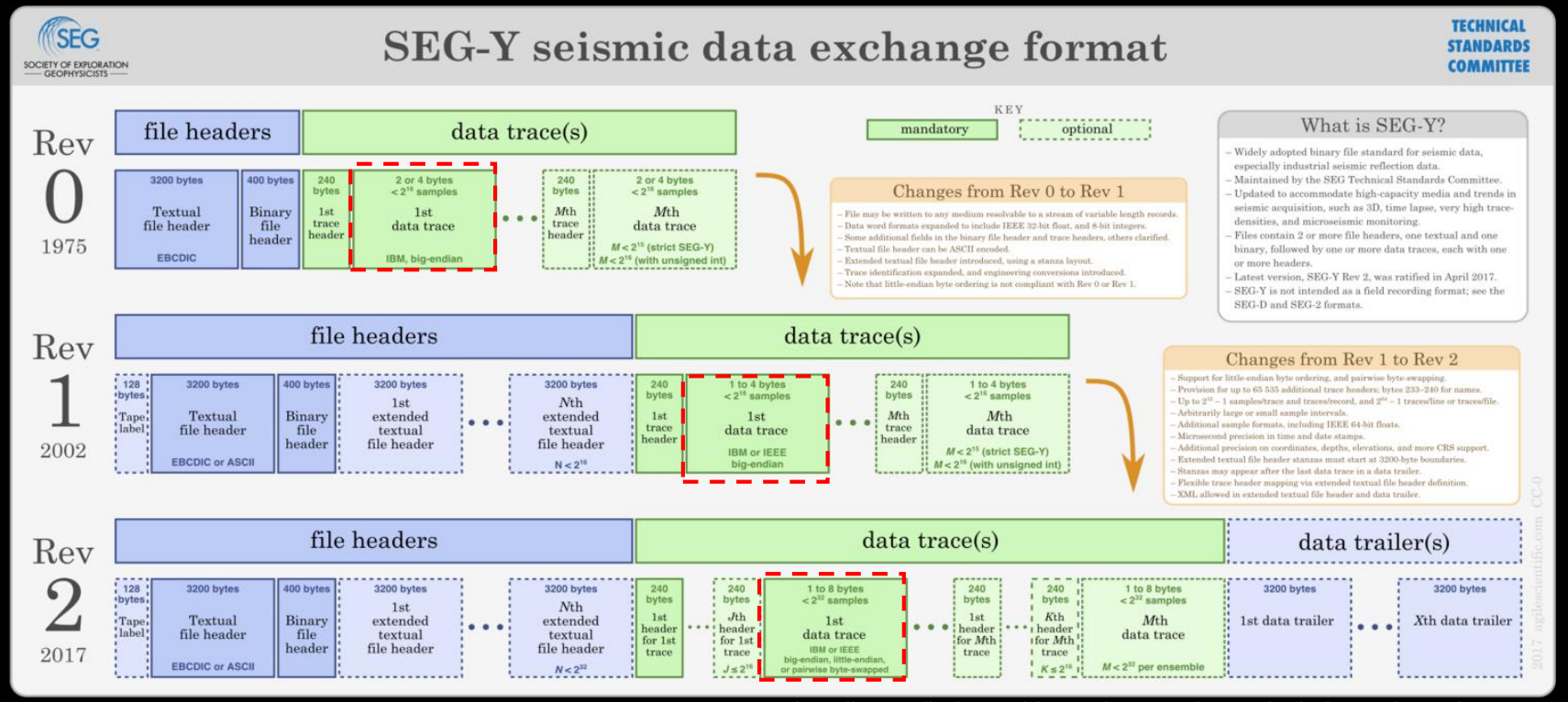
Rock physics, AVA inversion, stochastic inversion, probabilistic lithology and fluid prediction.

https://www.nextplatform.com/2019/05/16/dug-sets-foundation-for-exascale-hpc-utility-with-xeon-phi/

# SEG-Y: Seismic Survey Data Format

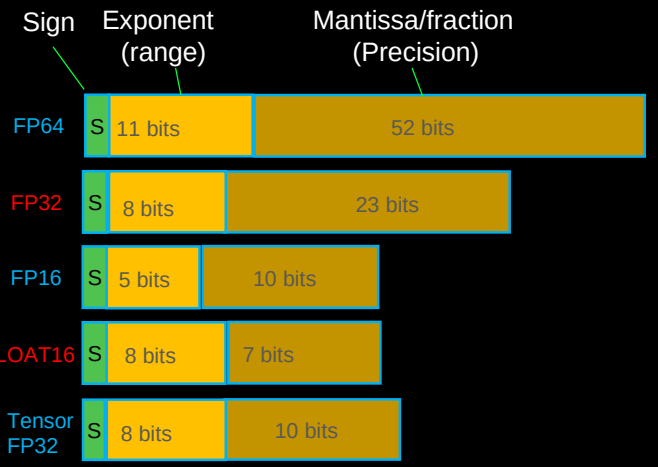
- ✓ A seismic trace represents the response of the elastic wavefield to velocity and density contrasts across interfaces of layers of rock or sediments as energy travels from a source through the subsurface to a receiver or receiver array
- ✓ The Seismic wave data collected by the survey instrument is arranged & saved in the SEG-Y format defined by SEG.
- ✓ As shown, each trace data log is an IEEE 754-1985 32-bit Single-Precision Floating-Point value, while other old format such as IBM 32-bit Single-Precision Floating-Point may be used in archived data.

***? Could Tensor FP32 or BFLOAT16 be used for FWI Seismic Data Processing for Higher Performance?***




The seismic data recorded for one channel

https://www.troika-int.com/about-us/resources/segy-seismic-data-exchange-format




FP32:
Exponent Range: $2^{-126}$ to $2^{+127}$
Data Range: -3.4E-38 to 3.4E+38
Precision: ~0.000001 (6-7 digits)

BFLOAT16:
Exponent Range: $2^{-126}$ to $2^{+127}$
Data Range: -3.4E-38 to 3.4E+38
Precision: ~0.000001 (6-7 digit)

FUTUREWEI Technologies

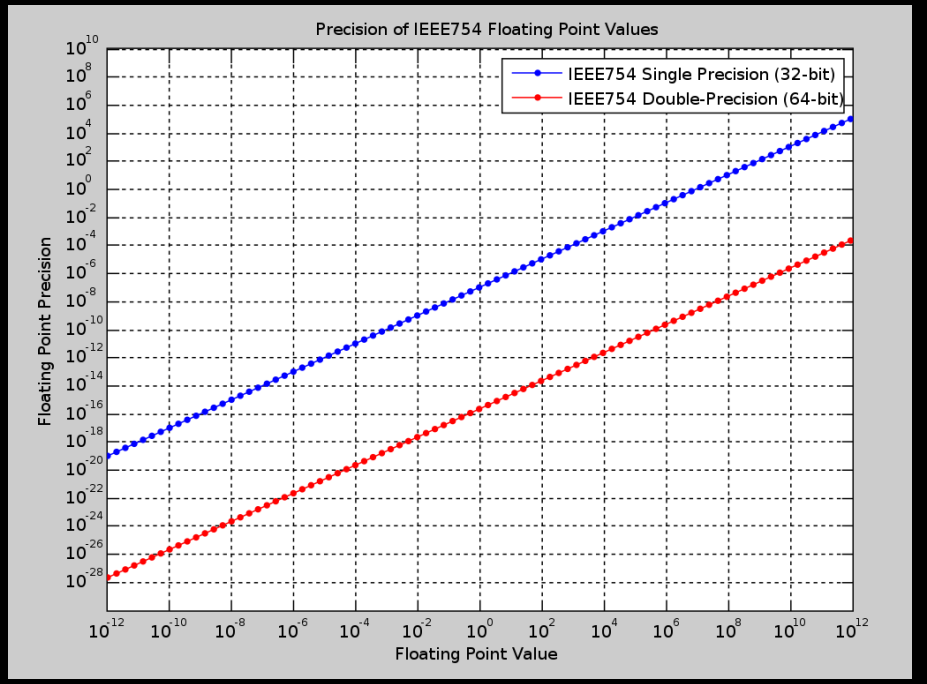
# IEEE Floating-Point Data Format, Ranges and Precision

- ✓ IEEE-754 defined 8 floating-point data formats (5 binary & 3 decimal), but only the FP64 & FP32 are commonly used
- ✓ The number of Mantissa bits determines the accuracy, but the relative accuracy matters. For 32b SP, it is about $10^{-7}$
- ✓ 32b SP has been used for DL model training, but Bfloat16 & Tensor FP32 are being adopted due to reduced memory BW
- ✓ If Bfloat16 or Tensor FP32 could be used for seismic data processing, throughput could be improved significantly (2X)

| Format | 64b Double Precision | 32b Single Precision | 16b Half Precision | Tensor FP32 | Bfloat16 |
|---|---|---|---|---|---|
| Total Bits | 64 | 32 | 16 | 19 | 16 |
| Significand Bits | 52+1 | 23+1 | 10+1 | 8+1 | 8+1 |
| Sign Bits | 1 | 1 | 1 | 1 | 1 |
| Exponent Bits | 11 | 8 | 5 | 8 | 8 |
| Exponent Bias | 1023 ($2^{10}$-1) | 127 ($2^{7}$-1) | 15 ($2^{4}$-1) | 127 | 127 |
| Exponent Max | +1023 | +127 | +15 | +127 | +127 |
| Exponent Min | -1022 | -126 | -14 | -126 | -126 |
| Decimal Exponent Max | 307.95 ($2^{1023}$) | 38.23 ($2^{127}$) | 4.51 ($2^{15}$) | 38.23($2^{127}$) | 38.23 ($2^{127}$) |
| Mantissa Bits | 52 | 23 | 10 | 10 | 7 |
| Relative Accuracy | $2^{-52}/2$ | $2^{-23}/2$ | $2^{-10}/2$ | $2^{-10}/2$ | $2^{-7}/2$ |
| Min Normalized Positive number | $2^{-1022}$ | $2^{-126}$ | $2^{-14}$ | $2^{-126}$ | $2^{-126}$ |
| Max number | $2x2^{1023}$ ($1.8x10^{307}$) | $2x2^{127}$ ($3.4x10^{38}$) | $2x2^{15}$ | $2x2^{127}$ | $2x2^{127}$ |
| Latest Supported | | | GPU, NPU | Nvidia A100 | Intel, Nvidia |
| Typical Applications | Technical Computing HPC | HPC, Gaming, DSP | | DL, AI, HPC | DL, AI |

BF16

Up to 2X Performance to FP32
8X Smaller Multiplier than FP32
2X Smaller Multiplier than FP16




https://en.wikipedia.org/wiki/IEEE_754#:~:text=The%20IEEE%20Standard%20for%20Floating,and%20Electronics%20Engineers%20(IEEE).&text=Many%20hardware%20floating%2Dpoint%20units%20use%20the%20IEEE%20754%20standard.

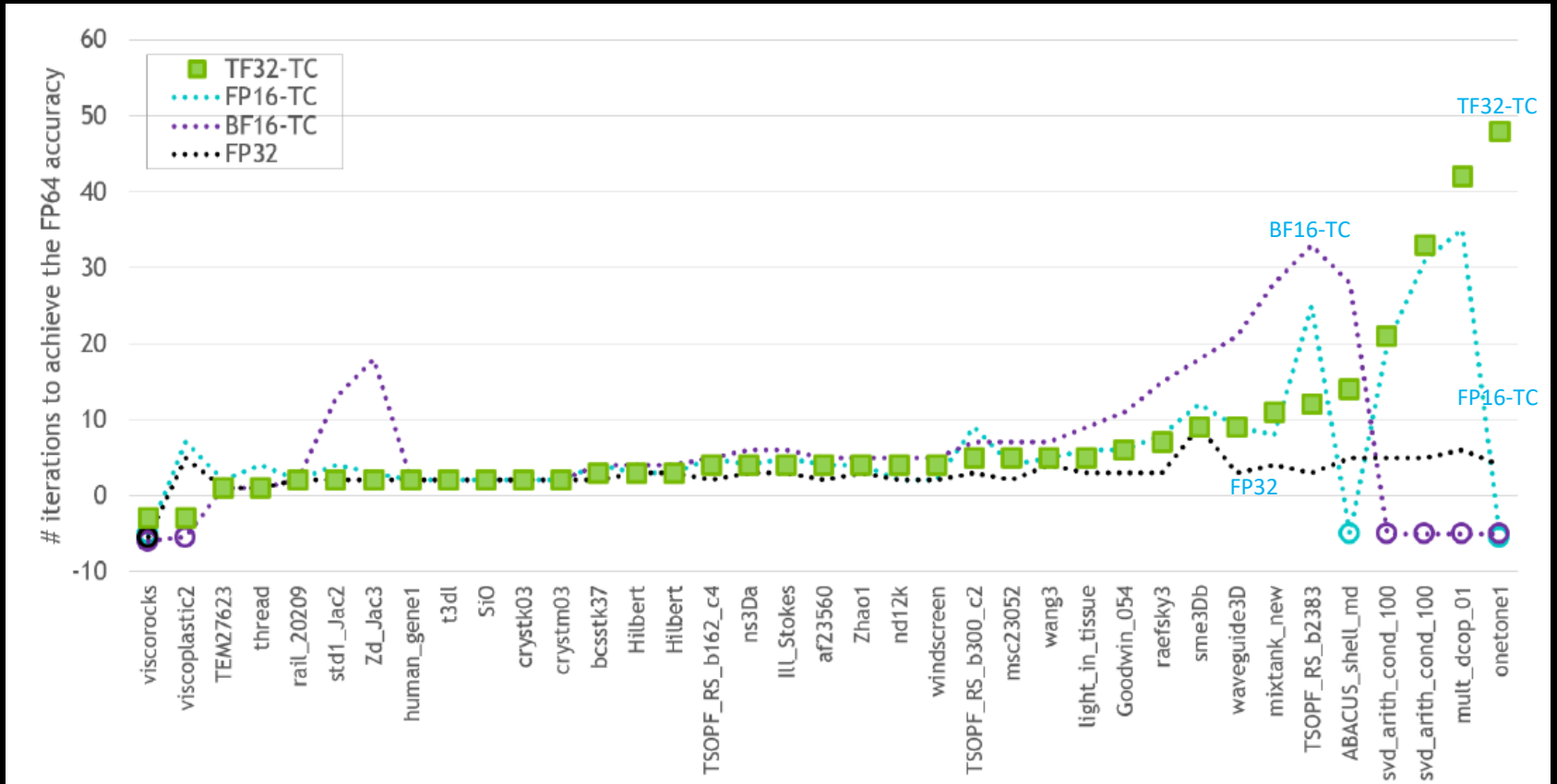
https://blog.demofox.org/2017/11/21/floating-point-precision/#:~:text=A%20float%20has%2023%20bits%20of%20mantissa%2C%20and%202%5E23,of%20precision%2C%20regardless%20of%20exponent.

https://floating-point-gui.de/formats/fp/ https://en.wikipedia.org/wiki/Machine_epsilon

# A100: Iterations of TCAIRS Solver to Converge to FP64 Accuracy

✓ Automated mixed TF32 + TF64 precision should be used to cover all scenarios as some linear solvers not converge if only TF32 used

- Mixed-Precision proven as very effective in DL training
- Nvidia A100 applying the same scheme for Linear Solvers, the Tensor Core Accelerated Iterative Refinement Solver (TCAIRS) in cuSOLVER automates usage of mixed precision for this application with both FP16, BF16 & TF32 for 64b Double-Precision Linear Solvers, and achieving good speed up for TF32.
- Number of iterations taken by the TCAIRS solver compared for 37 different problems to converge to FP64 accuracy for different precisions. The results are sorted by increasing number of iterations for TF32. Negative figures indicate the solver did not converge with reduced precision and fell back to a full FP64 solution.



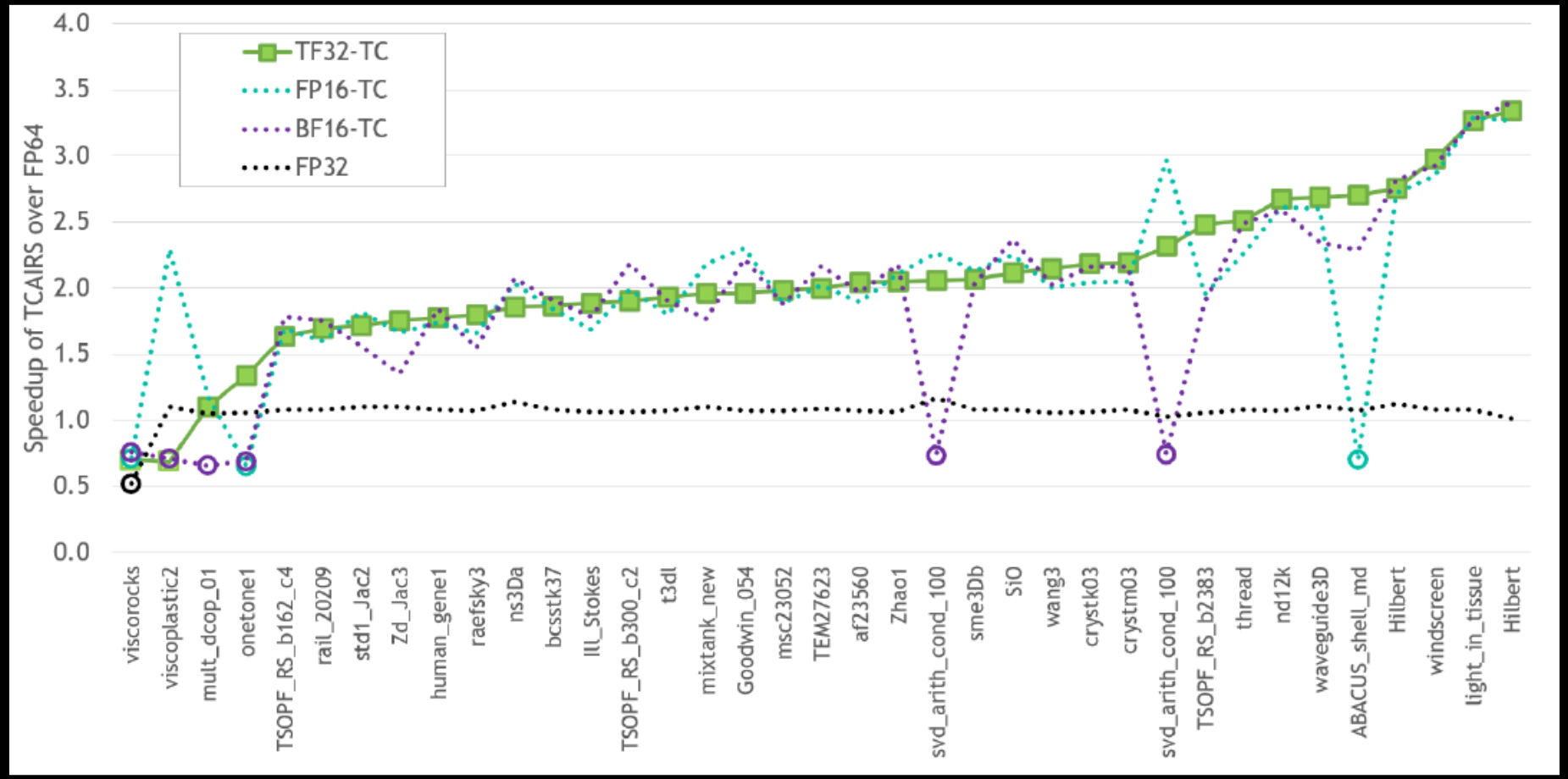

Number of iterations taken by the TCAIRS solver compared for 37 different problems to converge to FP64 accuracy for different precisions. The results are sorted by increasing number of iterations for TF32. Negative figures indicate the solver did not converge with reduced precision and fell back to a full FP64 solution.

Source & Credit: Nvidia, "NVIDIA A100 Tensor Core GPU Architecture White Paper"

# A100: TCAIRS Solver Speedup to Converge to FP64 Accuracy

- ✓ For FP64 solvers, Automated mixed TF32 + TF64 precision delivers more than 2X speedups than FP32+FP64
- ✓ What would be speedups for TF32 over FP32-only Linear Solvers? >> Seismic Data Processing & Digital Signal Processing are FP32 Only, no FP64 used.

Speedup of the TCAIRS solver over the baseline FP64 direct solver compared for 37 different problems. Cases where speedups are less than one indicate that the TCAIRS solver did not converge with reduced precision and fell back to a full FP64 solution. The results are sorted by increasing speedup for TF32.




Source & Credit: Nvidia, "NVIDIA A100 Tensor Core GPU Architecture White Paper"



FUTUREWEI Technologies

# New Floating-Point Data Formats for AI/DL

- ✓ Deep Learning Neural Network Model training is an extreme computation-intensive process requiring extreme high memory bandwidth, and High-Bandwidth-Memory (HBM) is integrated on the DL training ASIC for 1TB/s memory BW
- ✓ 32-bit Single-Precision Floating-Point (SPFP) data has been used for DL model training, but its relative accuracy of $2^{-24}$ is not required, thus Bfloat-16(BF16) and Tensor-FP32(TF32) have been proposed and adopted by industry to reduce memory bandwidth requirement and improve/double the DL model training performance. Both easy to convert from FP32.
- ✓ The difference between Bfloat-16 and Tensor-FP32 is the relative accuracy: $2^{-11}$ vs. $2^{-8,}$ but the same range as FP32

## Tensor-Float 32b FP


NVIDIA A100 GPU 2020

TF32 vs. FP32

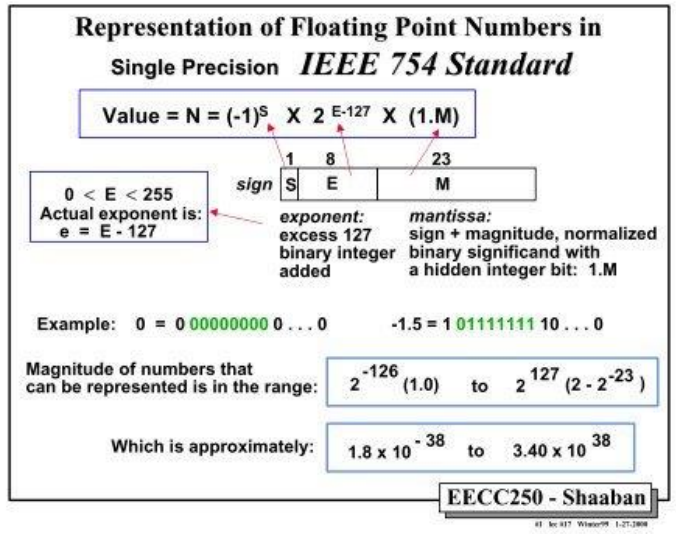
Up to 10X training Performance w/o Sparsity

Up to 20X training Performance w/ Sparsity

Claimed by Nvidia

https://blogs.nvidia.com/blog/2020/05/14/tensorfloat-32-precision-format/

## IEEE754 32b FP




## BFLOAT 16b FP


NVIDIA A100 GPU 2020

intel Cooper Lake CPU 2020

Google TPU2 2018, TPU3 2019

arm ARMv8.6-a, HUAWEI NG

TensorFlow

Up to 2X Performance to FP32

8X Smaller Multiplier than FP32

2X Smaller Multiplier than FP16

https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus



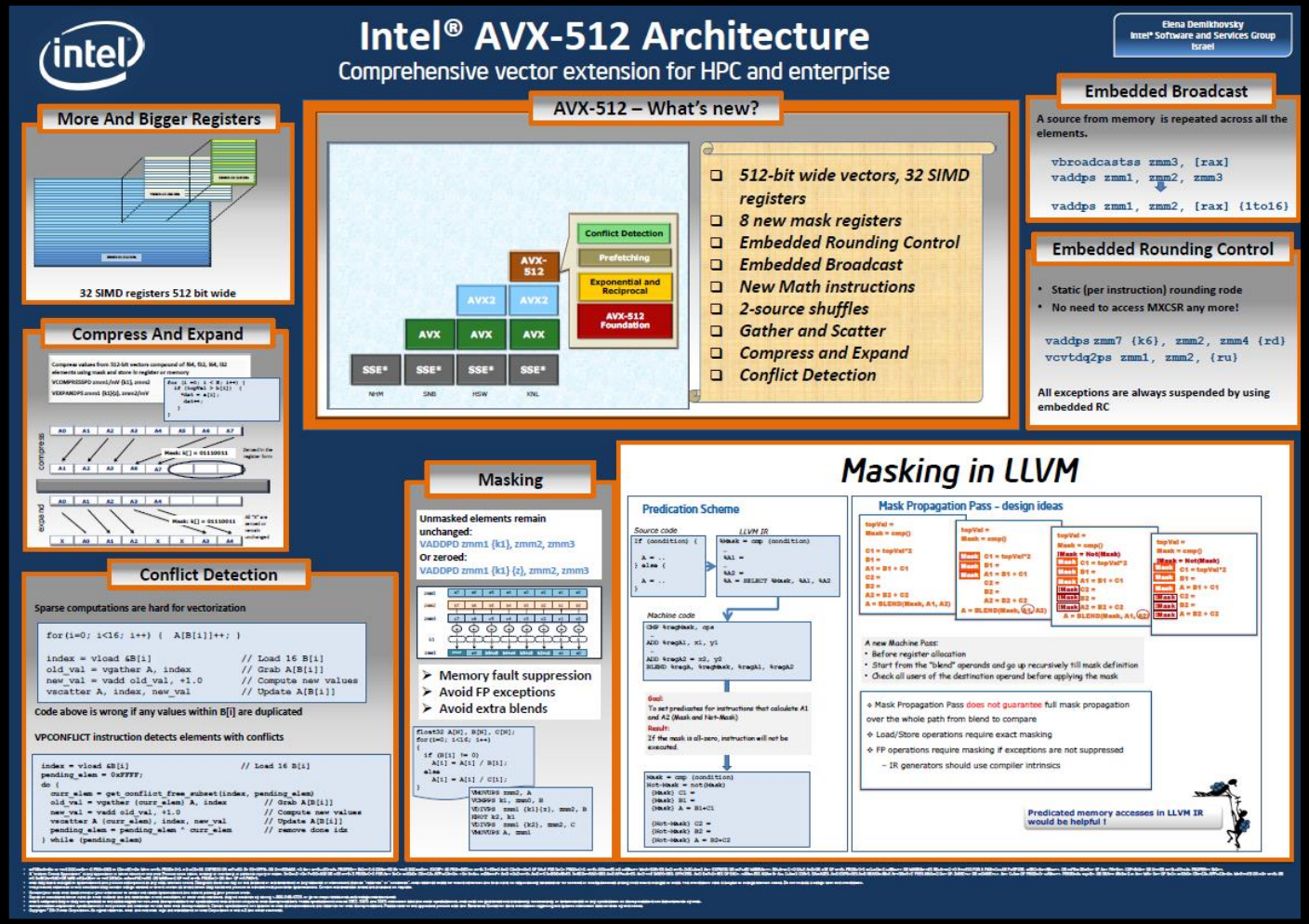
FUTUREWEI Technologies

# Summary

- ✓ Seismic Data Processing is a Complex Computation-Intensive Process with Huge Amount of Data to Compute, but it is also Highly Parallelable in nature due to the Many Seismic Wave Sources and Many Seismic Sensors.
- ✓ Full Waveform Inversion (FWI) is playing a Crucial role in Seismic Data Processing for constructing a HD 2D/3D/4D Subsurface model images. FWI is a process that grows 10x in compute with every frequency doubling. 25 Petaflops required for current ~15Hz means 2.5 Exaflops for ~125Hz to get out of the model basically everything recorded in that wavefield per DUG
- ✓ High Frequency FWI is a game changer that is becoming feasible with the affordable abundant computing power, the same as Deep Learning to the booming modern AI for FWI to seismic data processing.
- ✓ Data-Driven based Deep Learning Neural Network Technologies are being investigated to replace the Physics-Driven based Computational FWI to reduce the amount of compute required for High Frequency FWI, as well as to leverage the DL-optimized Tensor Cores that delivering 10X-100X higher performance than FP32 Matrix-Multiplication Operations
- ✓ Single-Precision Floating-Point Value is used in Seismic Data & Processing, but could less-precise/accurate BFLOAT16 or TF32 Tensor-Float floating data formats be used for more computing and less memory bandwidth?

# Intel Xeon AVX-512 & ARM64 SVE SIMD Vector Unit

- ✓ Some Xeon Processors have 2x AVX-512 SIMD Vector Units for Higher Floating-Point Performance if the Memory subsystem could feed the data fast enough
- ✓ ARM64 Neon SIMD is 128-bit, and SVE is designed for 128-bit to 2048-bit SIMD, up to SoC vendor implementation




- AVX512F – Foundation – most floating-point single/double instructions widened to 512-bit.
- AVX512-DQ – Double-Word & Quad-Word – most 32 and 64-bit integer instructions widened to 512-bit
- AVX512-BW – Byte & Word – most 8-bit and 16-bit integer instructions widened to 512-bit
- AVX512-VL – Vector Length eXtensions – most AVX512 instructions on previous 256-bit and 128-bit SIMD registers
- AVX512-VNNI** (Vector Neural Network Instructions)
- AVX512-VBMI, VBMI2 (Vector Byte Manipulation Instructions)
- AVX512-BITALG (Bit Algorithms)
- AVX512-IFMA (Integer FMA)
- AVX512-VAES (Vector AES) accelerating crypto
- AVX512-GFNI (Galois Field)
- AVX512-GNA (Gaussian Neural Accelerator)

For Skylake 8180 CPU w/ AVX-512 (two units):

- ✓ Clock Frequency = 2.3GHz;
- ✓ 28 cores
- ✓ 32 SP/16 DP FLOPS per AVX-512 Unit
- ✓ 2x AVX-512 Units / core
- ➢ FMA: 2.3G x 28 x 32 x 2 = 4122GFLOPS SP32
- ➢ FMA: 2.3G x 28 x 16 x 2 = 2061GFLOPS DP64

https://www.microway.com/knowledge-center-articles/detailed-specifications-of-the-skylake-sp-intel-xeon-processor-scalable-family-cpus/

# Intel Xeon CPU features for AI & HPC: AVX-512, VNNI, BF16

✓The Peak SP32 Is about 4.122TFLOPS per 28-Core Socket w/ Dual AVX-512 Unit for Skylake-Cascade Lake- Copper Lake

✓The Improvements have been focusing on AI/DL – VNNI & BF16 Support

**Skylake (V5)** — 2017
- AVX-512 F, CD, VL, DQ, BW
- Max 28 cores

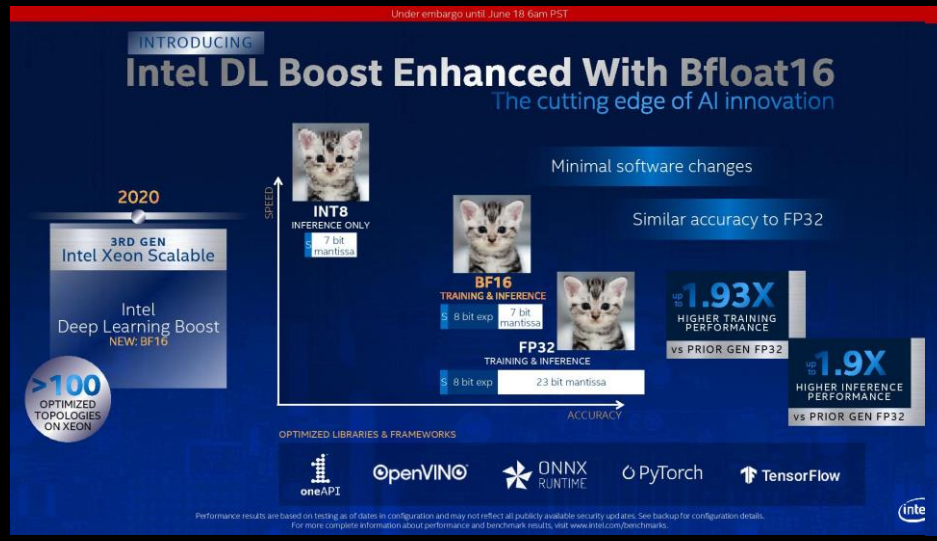
**Cascade Lake (V5)** — 2020
- AVX-512 F, CD, VL, DQ, BW
- VNNI
- Max 28 cores

**Cooper Lake (V6)** — 2020
- AVX-512 F, CD, VL, DQ, BW
- VNNI
- BF16
- Max 28 cores

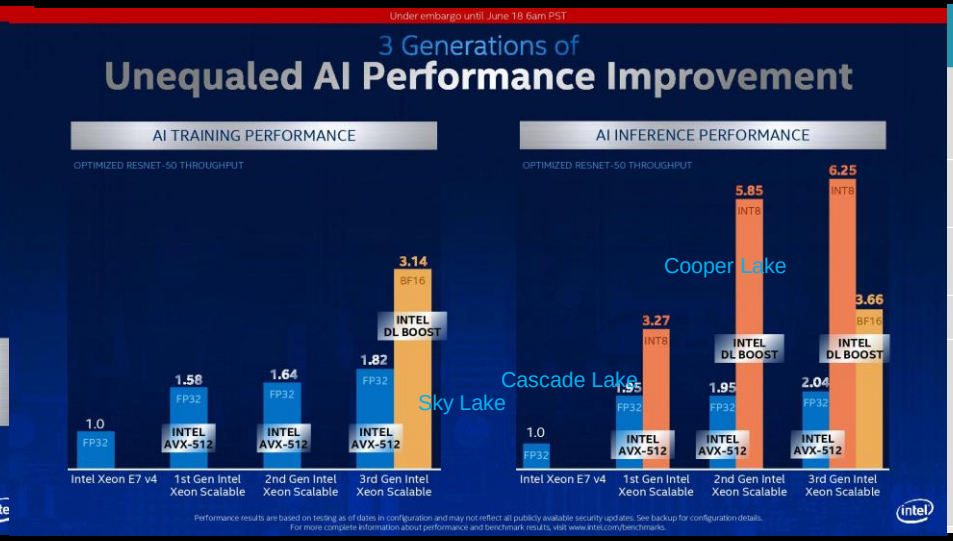
**Ice Lake (V6)** — 2020

- AVX512F – Foundation – most floating-point single/double instructions widened to 512-bit.
- AVX512-DQ – Double-Word & Quad-Word – most 32 and 64-bit integer instructions widened to 512-bit
- AVX512-BW – Byte & Word – most 8-bit and 16-bit integer instructions widened to 512-bit
- AVX512-VL – Vector Length eXtensions – most AVX512 instructions on previous 256-bit and 128-bit SIMD registers
- AVX512-VNNI** (Vector Neural Network Instructions)
- AVX512-VBMI, VBMI2 (Vector Byte Manipulation Instructions)
- AVX512-BITALG (Bit Algorithms)
- AVX512-IFMA (Integer FMA)
- AVX512-VAES (Vector AES) accelerating crypto
- AVX512-GFNI (Galois Field)
- AVX512-GNA (Gaussian Neural Accelerator)

~1.9X Speedup for DL training & Inferencing over FP32




Performance Improvement w/ Both Hardware & Software

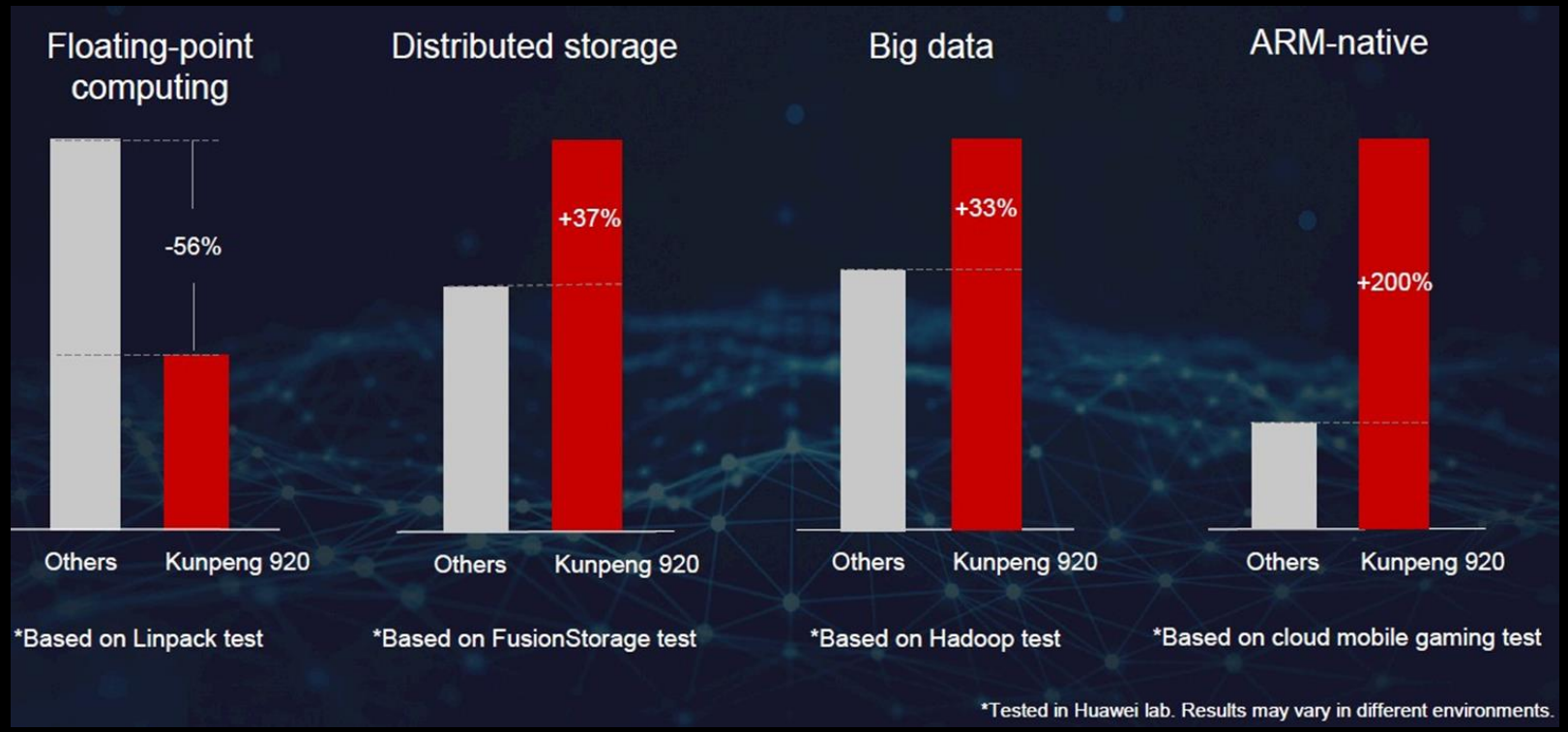



AVX-512 Propagation

| | Xeon | General |
|---|---|---|
| Skylake-SP | AVX512BW<br>AVX512DQ<br>AVX512VL | AVX512F<br>AVX512CD |
| Cannon Lake | AVX512VBMI<br>AVX512IFMA | |
| Cascade Lake-SP | AVX512_VNNI | |
| Cooper Lake | AVX512_BF16 | |
| Ice Lake | AVX512_VNNI<br>AVX512_VBMI2<br>AVX512_BITALG<br>AVX512+VAES<br>AVX512+GFNI<br>AVX512+VPCLMULQDQ<br>(not BF16) | AVX512_VPOPCNTDQ |

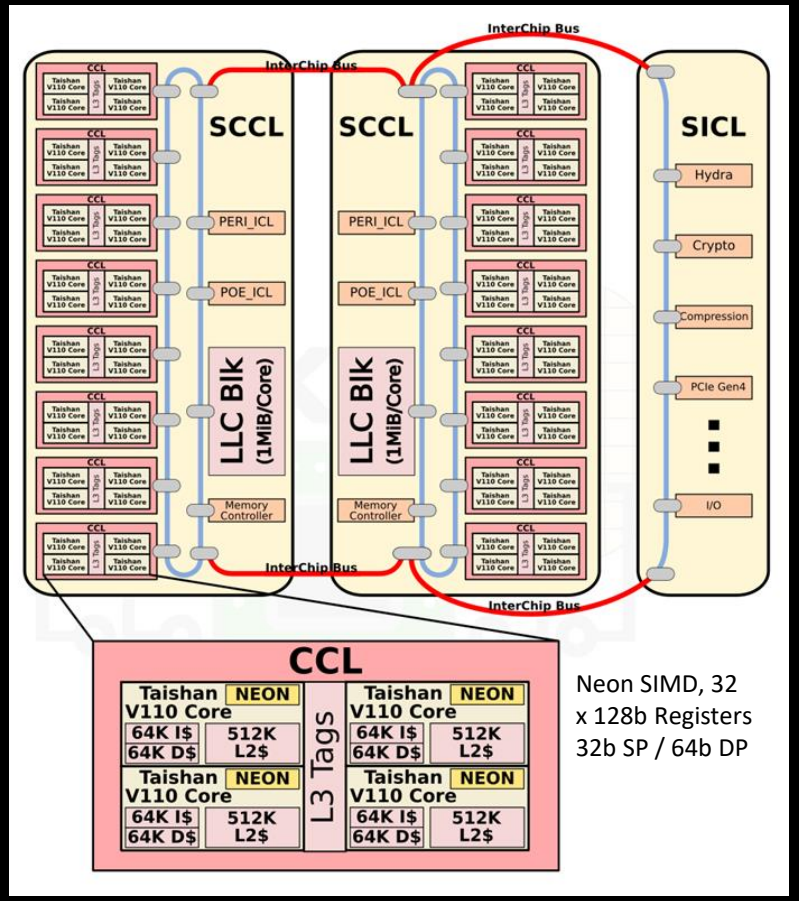
# Huawei KunPeng family ARM64 CPUs

- ✓ KunPeng-920 Provides more 64b cores (64 max) than Xeon (28 max) at similar frequency, and 33% higher Memory bandwidth due to 8-CH vs. 6-CH, and less power
- ✓ In general, KunPeng-920 delivers Lower Floating Point Performance due to its narrower 128-bit Neon SIMD Vector than Xeon AVX-512, equivalent to Middle-Range Xeon Skylake Processors
- ✓ It is expected that our next generation ARM64 will support SIMD to 4 x128 (512 bits) for much higher floating-Point Performance




## KunPeng-920 ARM64 SoC

- Up to 64 cores, 2.6GHz, 7nm, 180W
- 8-Channel DDR4-2933
- PCIe 4.0, 100GE & CCIX
- 128b SIMD, max. 1.331TFLOPS SP, 666GFLOPS DP FMA



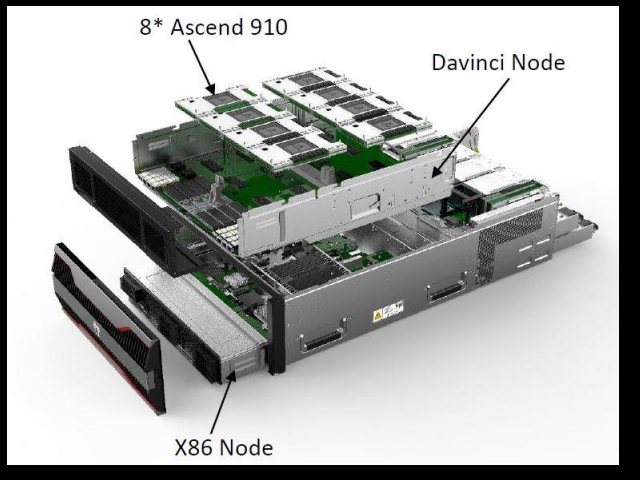

Neon SIMD, 32 x 128b Registers
32b SP / 64b DP

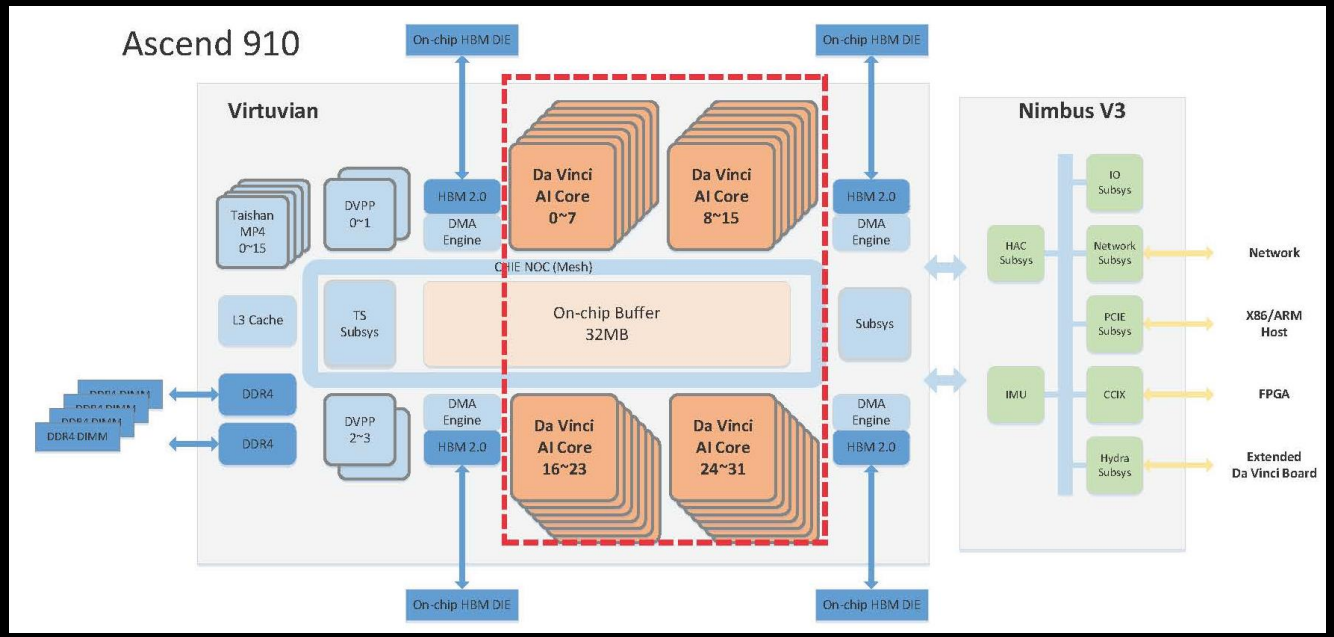


FUTUREWEI Technologies

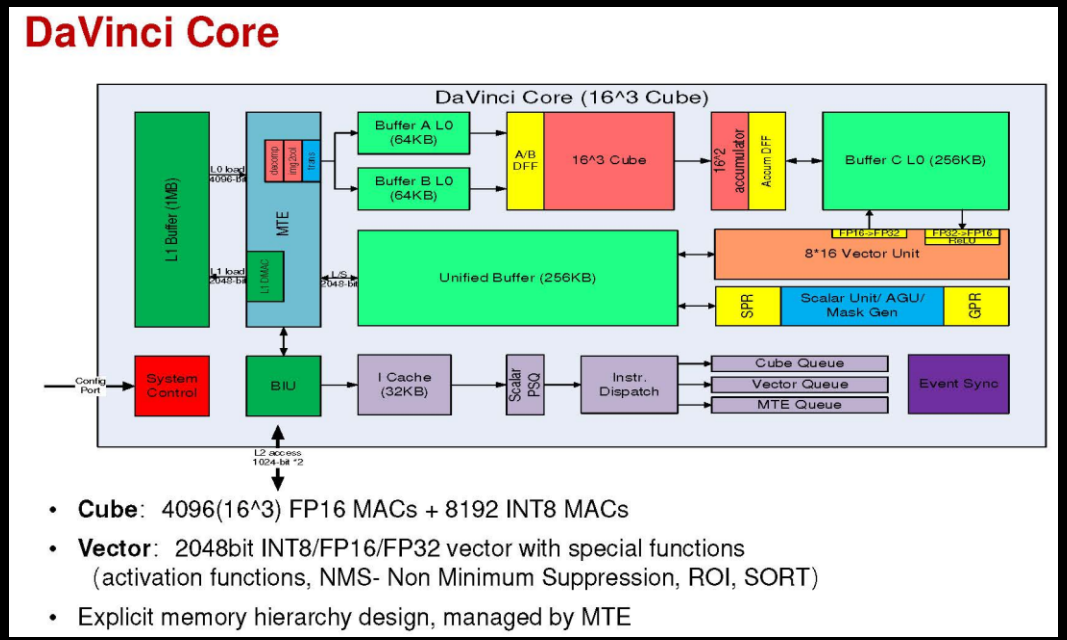
# Huawei Ascend AI Accelerator SoCs Optimized for DL

- 256 TFLOPS FP16 / 512 TOPS INT8 for DL training / inferencing at Cube (max 4096 FMA Operations per Clock Cycle)
- 1.2TB/s & 32GB HBM2E
- FP32 is supported at the 2048-bit Vector Unit, but performance is much lower than the 16x16x16 Cube
- 128-channel H.264/265 hardware Video Decoder



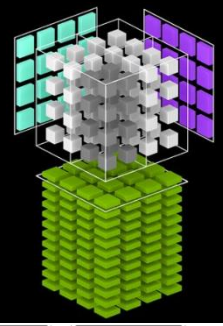

https://www.servethehome.com/huawei-ascend-910-provides-a-nvidia-ai-training-alternative/




- **Cube**: 4096(16^3) FP16 MACs + 8192 INT8 MACs
- **Vector**: 2048bit INT8/FP16/FP32 vector with special functions (activation functions, NMS- Non Minimum Suppression, ROI, SORT)
- Explicit memory hierarchy design, managed by MTE

# Nvidia Ampere A100 GPU: TF32 & BFLOAT16 for Higher Performance

✓ Tensor Cores Matrix-Multiply & Accumulate (FMA), TF32/BF16 Support, and Large Capacity HBM2 are Key to extreme AI/HPC Performance



## Claims by Nvidia:

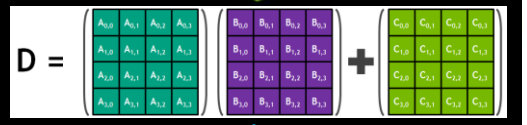
- ✓ 10X Training Performance for TF32 over FP32 w/o Sparsity
- ✓ 20X Training Performance for TF32 over FP32 w/ Sparsity
- ✓ TF32 is applicable & delivers the highest performance to HPC Linear Solvers in a wide range of fields such as earth science, fluid dynamics, healthcare, material science and nuclear energy as well as oil and gas exploration

## High Performance Engines for AI Workloads

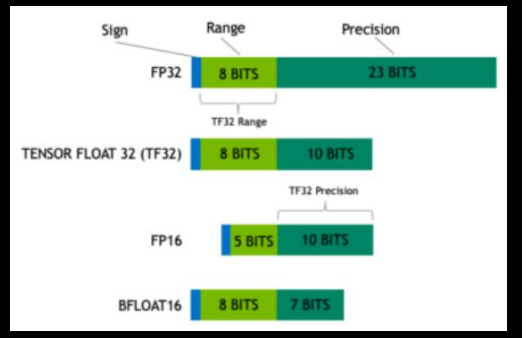
- ✓ BFLOAT16 Tensor Cores – 432 cores for 312/624 teraFLOPS(training)
- ✓ TP32 Tensor Cores – 432 cores for 156/312 teraFLOPS (training)
- ✓ FP32 CUDA Cores – 6912 cores for 19.5 teraFLOPS (training)
- ✓ INT8 Tensor Cores – 432 cores for 624/1248 TOPS (inferencing)
- ✓ INT4 Tensor Cores – 432 cores 1248/2496 TOPS (inferencing)
- ✓ 1.6TB/s 40GB HBM2 Memory

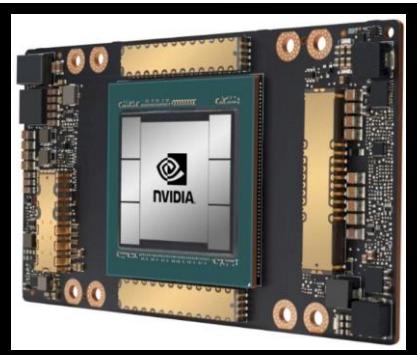
## High Performance Engines for HPC Workloads

- ✓ FP64 CUDA Cores – 3456 cores for 9.7 teraFLOPS
- ✓ FP32 CUDA Cores – 6912 cores for 19.5 teraFLOPS
- ✓ FP64 Tensor Cores – 432 cores for 19.5 teraFLOPS
- ✓ TP32 Tensor Cores – 432 cores for 156/312 teraFLOPS
- ✓ 1.6TB/s 40GB HBM2 Memory




D = A * B + C




Ampere A100 GPU

## CUDA Cores Data Formats Supported

✓FP64 & FP32 & FP16

## Tensor Cores Data Formats Supported

✓FP64 (FP32 as Input not supported)
✓TF32
✓BFLOAT16 & FP16
✓INT8 & INT4

| | Peak Performance |
|---|---|
| Transistor Count | 54 billion |
| Die Size | 826 mm² |
| FP64 CUDA Cores | 3,456 |
| FP32 CUDA Cores | 6,912 |
| Tensor Cores | 432 |
| Streaming Multiprocessors | 108 |
| FP64 | 9.7 teraFLOPS |
| FP64 Tensor Core | 19.5 teraFLOPS |
| FP32 | 19.5 teraFLOPS |
| TF32 Tensor Core | 156 teraFLOPS \| 312 teraFLOPS* |
| BFLOAT16 Tensor Core | 312 teraFLOPS \| 624 teraFLOPS* |
| FP16 Tensor Core | 312 teraFLOPS \| 624 teraFLOPS* |
| INT8 Tensor Core | 624 TOPS \| 1,248 TOPS* |
| INT4 Tensor Core | 1,248 TOPS \| 2,496 TOPS* |
| GPU Memory | 40 GB |
| GPU Memory Bandwidth | 1.6 TB/s |
| Interconnect | NVLink 600 GB/s<br>PCIe Gen4 64 GB/s |
| Multi-Instance GPUs | Various Instance sizes with up to 7MIGs @5GB |
| Form Factor | 4/8 SXM GPUs in HGX A100 |
| Max Power | 400W (SXM) |

*structural sparsity enabled



https://blogs.nvidia.com/blog/2020/05/14/tensorfloat-32-precision-format/

# Nvidia A100 GPU High Performance Architecture

- ✓ Tensor Cores Matrix-Multiply-Accumulate (FMA) and TF32/BF16 Floating Formats are the key for extreme high AI/HPC performance
- ✓ FP32 accuracy as Input/output to Tensor Core not Supported, but FP32 format is Supported, and the accuracy is TF32 or BF16

## CUDA Cores

- Simultaneous FP32 & INT32
- Max 19.5TFLOPS for FP32
- Max 9.7TFLOPS for FP64

FP64:
- 19.5TFLOPS w/ Tensor Cores
- 9.7TFLOPS w/ CUDA Cores

FP32:
- 19.5TFLOPS w/ CUDA Cores
- FP32 not supported at Tensor Cores

TF32: 8-16x of FP32
- 156TFLOPS w/ Tensor Cores
- 312TFLOPS w/ Sparsity

BF16: 2X TF32, 16-32x FP32
- 312TFLOPS w/ Tensor Cores
- 624TFLOPS w/ Sparsity

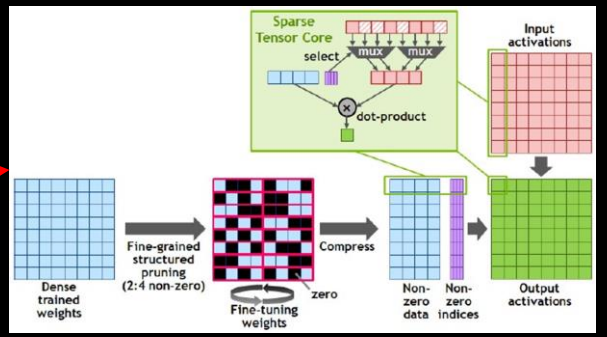
108 SMs / A100 GPU

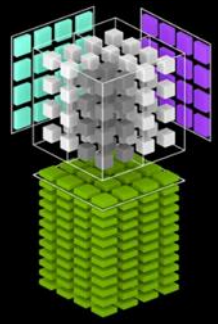

## Each SM: 1024 FP16 or BF16/FP32 OPC

- ✓ 4x SP & 4x Tensor Cores per SM (one Tensor Core/SP)
- ✓ 1024 FP16/FP32 Mixed-Precision FMA Operations per Clock w/ Tensor Cores, 2x Performance w/ Sparsity
- ✓ FP64 Support for 180.56GFLOPS w/ Tensor Cores
- ✓ FP32 Support for 180.56GFLOPS via CUDA Cores

## Each Tensor Core: 256 FP16 or BF16 / FP32 OPC

- ✓ 256 FP16/FP32 Mixed Precision FMA Operations per Clock
- ✓ FP16, BF16, TF32, FP64, INT8, INT4, and INT1 Formats
- ✓ TF32 FMA is 10x faster than FP32 in V100
- ✓ 2x Performance w/ Sparsity
- ✓ FP64 Support for 45.13GFLOPS
- ✓ BF16/FP32 running at the same rate as FP16/FP32
- ✓ FP32 accuracy as input to Tensor Cores Not Supported



Compress Dense Trained Weights



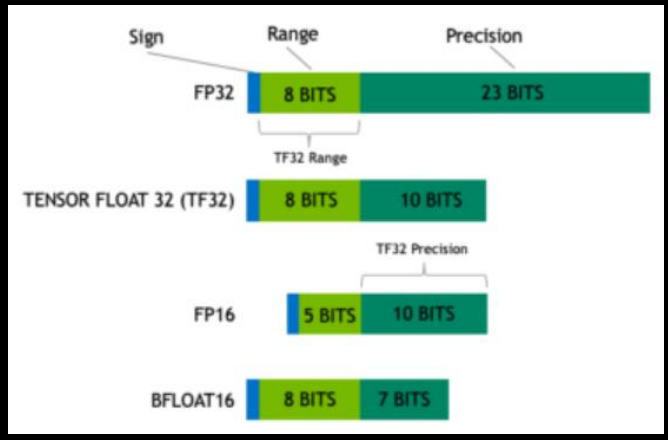
FUTUREWEI Technologies

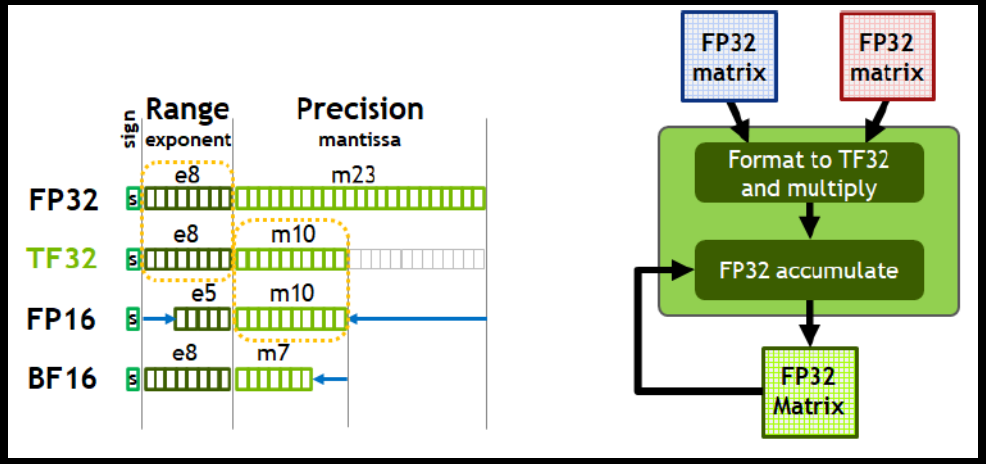
# Nvidia A100 GPU High Performance Architecture

- ✓ Seismic Data Processing is extreme FP32 computation Intensive; and almost all HPC Linear Solvers could be done with TF32 Accuracy
- ✓ FP32 accuracy as input to the Tensor Cores is NOT Supported, but FP32 accuracy as Tensor Core Output is supported
- ✓ If TF32 or BF16 Input accuracy is acceptable to Seismic Data Processing, then High Performance Tensor Cores could be used for Seismic Data Processing; otherwise the lower performance CUDA cores are the choice.

| | V100 | A100 | A100 Sparsity[1] | A100 Speedup | A100 Speedup with Sparsity |
|---|---|---|---|---|---|
| A100 FP16 vs V100 FP16 | 31.4 TFLOPS | 78 TFLOPS | NA | 2.5x | NA |
| A100 FP16 TC vs V100 FP16 TC | 125 TFLOPS | 312 TFLOPS | 624 TFLOPS | 2.5x | 5x |
| A100 BF16 TC vs V100 FP16 TC | 125 TFLOPS | 312 TFLOPS | 624 TFLOPS | 2.5x | 5x |
| A100 FP32 vs V100 FP32 | 15.7 TFLOPS | 19.5 TFLOPS | NA | 1.25x | NA |
| A100 TF32 TC vs V100 FP32 | 15.7 TFLOPS | 156 TFLOPS | 312 TFLOPS | 10x | 20x |
| A100 FP64 vs V100 FP64 | 7.8 TFLOPS | 9.7 TFLOPS | NA | 1.25x | NA |
| A100 FP64 TC vs V100 FP64 | 7.8 TFLOPS | 19.5 TFLOPS | NA | 2.5x | NA |
| A100 INT8 TC vs V100 INT8 | 62 TOPS | 624 TOPS | 1248 TOPS | 10x | 20x |
| A100 INT4 TC | NA | 1248 TOPS | 2496 TOPS | NA | NA |
| A100 Binary TC | NA | 4992 TOPS | NA | NA | NA |

1 - Effective TOPS / TFLOPS using the new Sparsity Feature




A100 GPU: The lowest mantissa bits are ignored when read into Tensor Cores



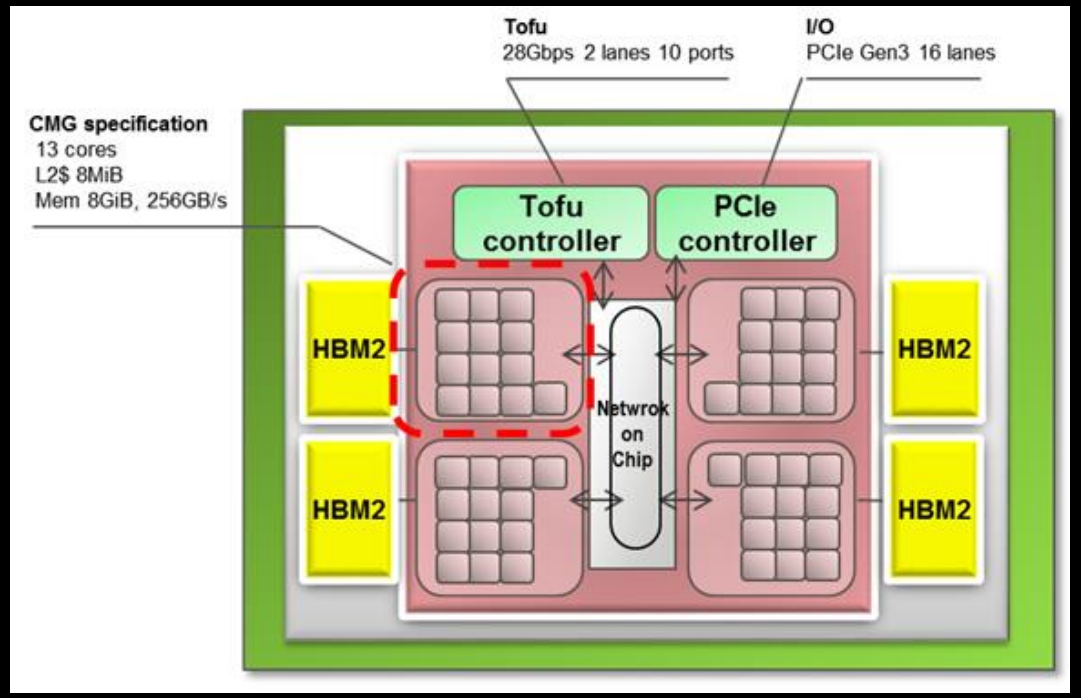
FUTUREWEI Technologies

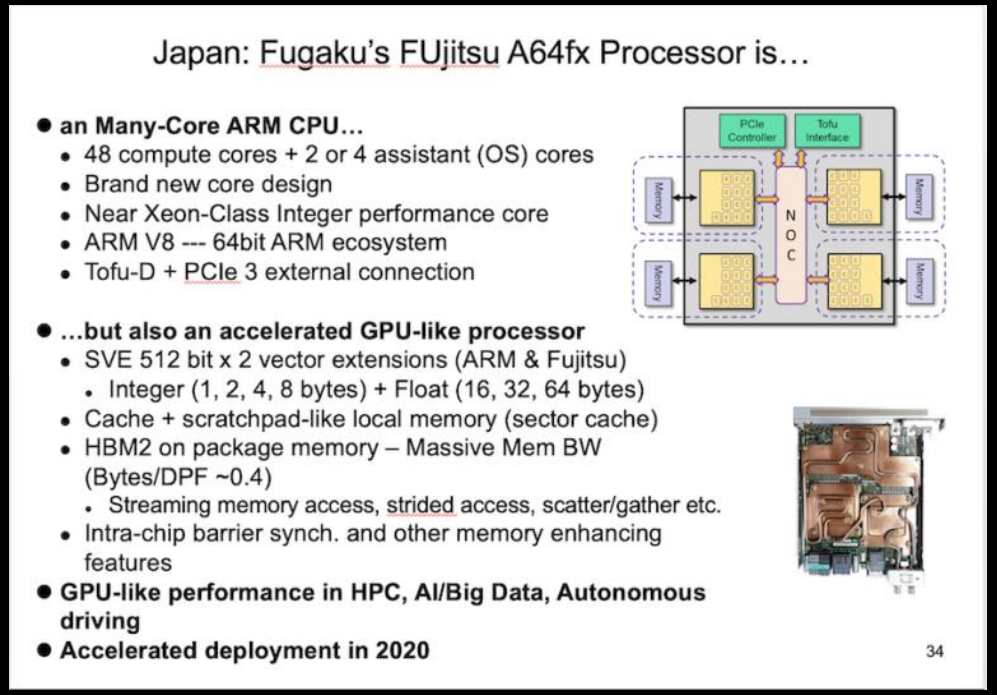
# Fujitsu A64FX CPU for HPC & AI, without external DIMM DRAM

✓ **The ARM64 SoC Optimized for HPC Workloads with SVE accelerators, HBM2 Memory & Tofu Cluster Interconnect**

**Claims by Fujitsu: Performance powered by SVE 512-bit x 2, equivalent to Xeon AVX-512**

- ✓ FP64: 2.7 teraFLOPS (for most HPC)
- ✓ FP32: 5.4 teraFLOPS (AI Training, Seismic Data Processing, due to 48 cores / 96x SVE-512 & HBM2)
- ✓ FP16: 10.8 TeraFLOPS
- ✓ INT8: 21.6 IOPS (AI Inferencing)
- ✓ HBM: HBM2, 32GB, 1024GB/s; No External DRAM DIMM
- ✓ CPU: 48 ARMv8.2-A Cores with 512-bit x 2 SVE Units
- ✓ No BFLOAT16 Floating Point Support



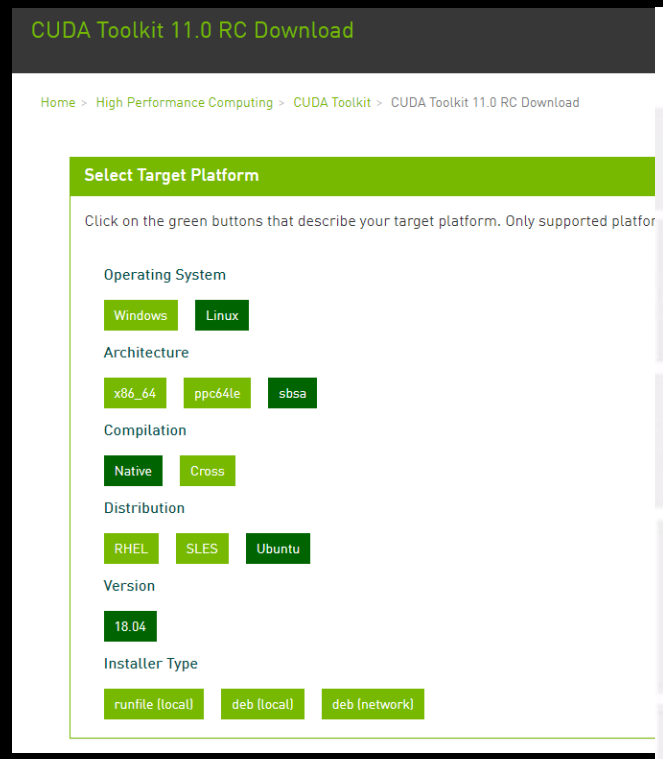



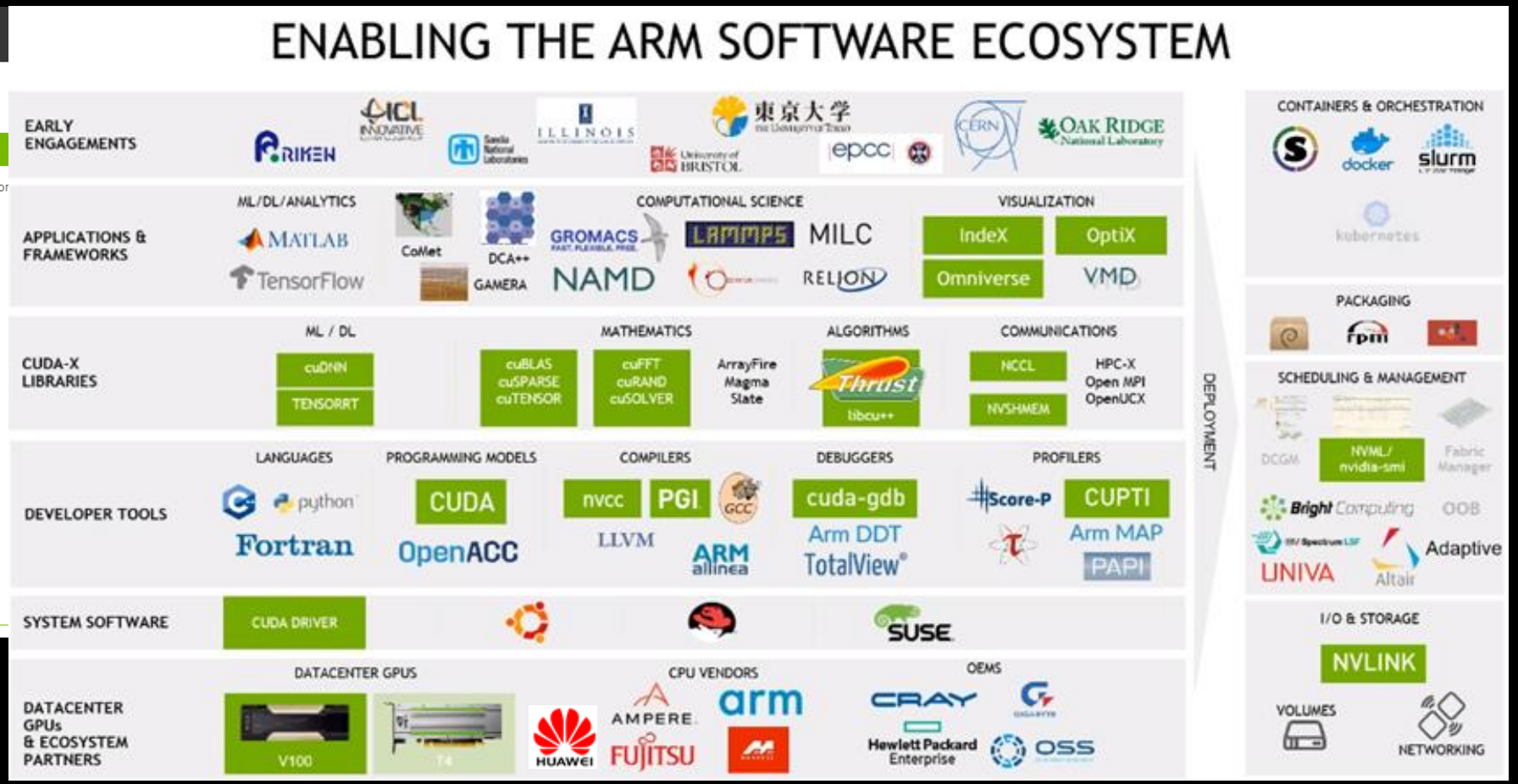



FUTUREWEI Technologies

# Nvidia CUDA for ARM64

- ✓ CUDA 11.0 for ARM64 Platform (including Huawei KunPeng-920) with support for Ampere GPU has been available for download
  - Not Only the CUDA Runtime and Compilers, but also all of the AI & HPC libraries & Tools




https://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html



FUTUREWEI Technologies

# Comparison of some GPU, NPU and CPU Devices

✓ ARM64 + Nvidia GPU could deliver the most cost-effective Highest performance for AI & Seismic Data Processing

| Features | Nvidia A100 | Huawei Ascend 910 | Huawei KunPeng 920 | Intel Cooper Lake | Fujitsu A64FX |
|---|---|---|---|---|---|
| Max FP64 w/ Tensor Cores (TFLOPS) | 19.5 | N/A | N/A | N/A | N/A |
| Max FP32 w/ Tensor Cores (TFLOPS) | N/A | N/A | N/A | N/A | N/A |
| Max TF32 w/ Tensor Cores (TFLOPS) | 156/(312) | N/A | N/A | N/A | N/A |
| Max FP16 w/ Tensor Cores (TFLOPS) | 312/(624) | 256 | N/A | N/A | N/A |
| Max BF16 w/ Tensor Cores (TFLOPS) | 312/(624) | N/A | N/A | N/A | N/A |
| Max FP64 w/ Vectors (TFLOPS) | 9.7 | N/A | 0.666 | ~2 | 2.7 |
| Max FP32 w/ Vectors (TFLOPS) | 19.5 | 4? | 1.332 | ~4 | 5.4 |
| Max FP16 w/ Vectors (TFLOPS) | 78 | 8? | 2.664 | ~8? | 10.8 |
| Max BF16 w/ Vectors (TFLOPS) | 312/(624) | N/A | N/A | ~8? (2x FP32) | N/A |
| On-Chip HBM2 Capacity (GB) | 40 | 32 | N/A | N/A | 32 |
| On-Chip HBM2 Bandwidth (TB/s) | 1.6TB/s | 1.2TB/s | N/A | N/A | 1TB/s |
| DDR4 channels & Speed | N/A | 4x DDR4-2666? | 8x DDR4-2993 | 6x DDR4-3200 | N/A |
| PCIe Speed | Gen 4 | Gen 4 | Gen 4 | Gen 3 | Gen 3 |

# Take-Aways

- ✓ Tensor Cores and TF32 / BF16 Floating Point Data Formats are Key to GPU/NPU's Extreme High DL/AI Neural Network Model training performance, but TF32/BF16's accuracy is much lower than IEEE-754 FP32.
- ✓ Mixed-Precision Operations have been proven very effective in speeding up AI/DL model training with FP16+FP32 or BF16+FP32 or TF32+FP32; and research / evaluations showing very positive (up to 3X speed up) result on leveraging TF32 + FP64 mixed-precision for FP64 Linear Solvers. But for a full solution, automated TF32+FP64 Mixed Precision should be adopted.
- ✓ Per Nvidia analysis, TF32 is applicable & delivers the highest performance to HPC Linear Solvers in a wide range of fields such as earth science, fluid dynamics, healthcare, material science and nuclear energy as well as oil and gas exploration (Seismic Data Processing).
- ✓ The latest generation A100 GPU from Nvidia offers the highest FP64, FP32, TF32, BF16, FP16 & INT8 Peak Performance
- ✓ The latest CUDA software supports both X86 and ARM64 Platform, including Huawei KunPeng 920.
- ✓ For Seismic Data Processing, ARM64+GPU could be a more cost-effective solution.
- ✓ Huawei Atlas family AI Servers are Ready to support both legacy Physics-Driven and DL-based Data-Driven Seismic Data Processing with X86 + GPU or ARM64+GPU, as well as DL-based Data-Driven Seismic Data Processing with ARM64+Ascend or X86+Ascend

# Thank you


https://www.fractracker.org/
https://www.zionmarketresearch.com/report/oil-gas-analytics-market
https://brandessenceresearch.com/energy-and-mining/oil-and-gas-analytics-market
https://www.marketsandmarkets.com/Market-Reports/seismic-survey-market-207832302.html
https://en.wikipedia.org/wiki/Reservoir_simulation
https://petrowiki.org/Reservoir_simulation

# Oil & Gas Open Data sets

https://github.com/microsoft/seismic-deeplearning

✓ Open Source for using Deep Learning for Seismic Imaging and Interpretation

https://wiki.seg.org/wiki/Open_data

✓ List a lot of open geophysical data available for download from the internet, via mail, or through special request. Including but not limited to:

- SEAM Phase 1 Elastic Earth Model Subset – 2D
- SEAM Phase I: Interpretation challenge I – Time
- Elastic 2DEW Classic
- Elastic VSP – 2D Walk-Away
- Well logs
- 2D land seismic data
- 2D marine seismic data
- 3D land seismic data
- 3D marine seismic data
- SEG/DMEC Reference Mineral Exploration Data
- OpenGeoscience at British Geological Survey

https://terranubis.com/osr

✓ TerraNubis is a cloud-based portal for buying, selling and interpreting seismic data sets and interpretations. The portal is developed and maintained by dGB Earth Sciences, the developers of OpendTect seismic interpretation software.

https://seg.org/

✓ Society of Exploration Geophysicists