2020 embedded VISION summit

Acceleration of Deep Learning using Intel® Distribution of OpenVINO™ toolkit: 3D Seismic Case Study

Manas Pathak
Global AI Lead for Oil & Gas
September 2020

intel

Notices and Disclaimers

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

The cost reduction scenarios described are intended to enable you to get a better understanding of how the purchase of a given Intel based product, combined with a number of situation-specific variables, might affect future costs and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs or cost reduction.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

No computer system can be absolutely secure.

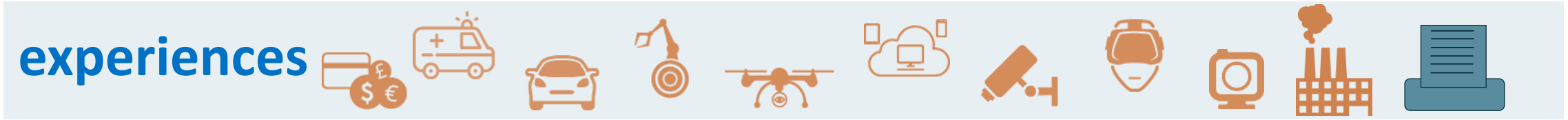
Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

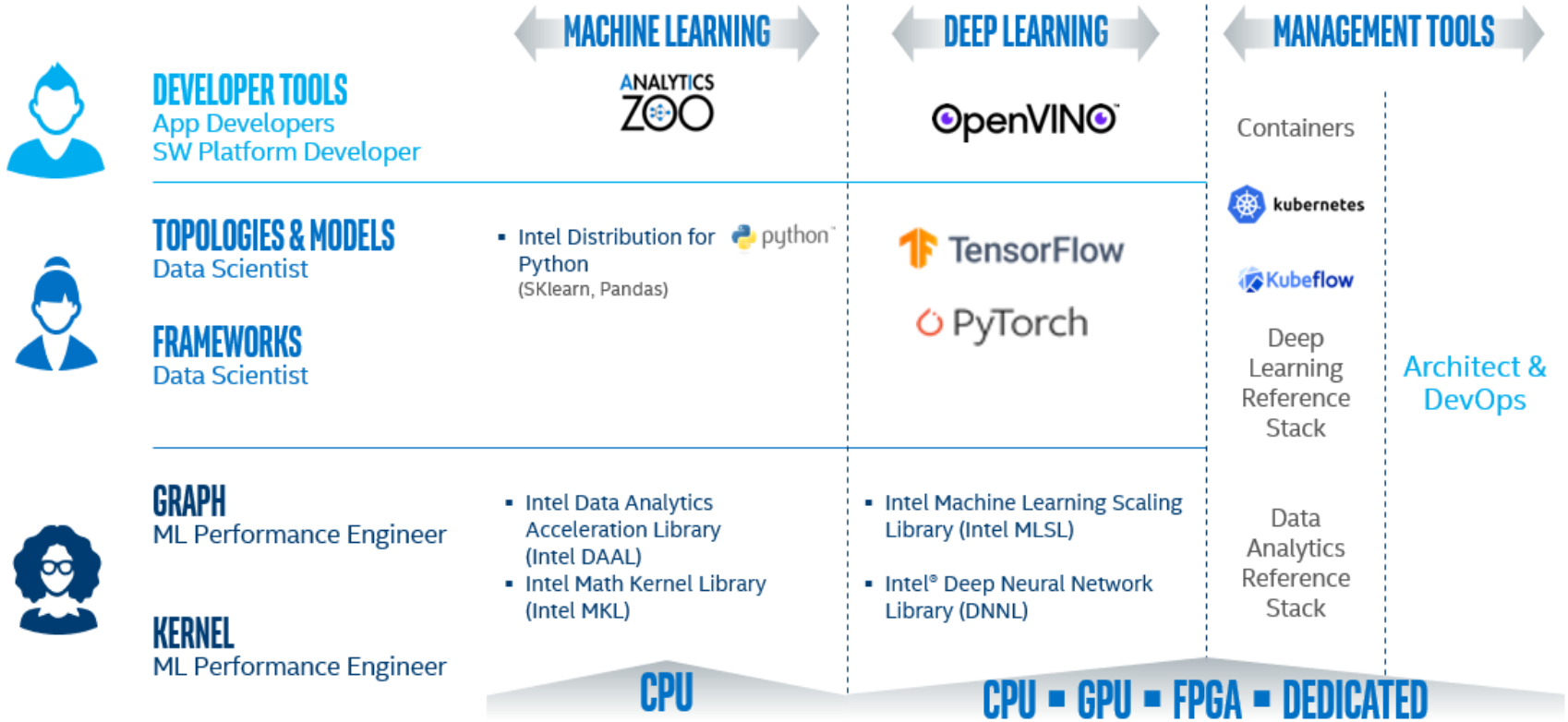
Intel processors of the same SKU may vary in frequency or power as a result of natural variability in the production process.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Combine SW & HW for Full Potential

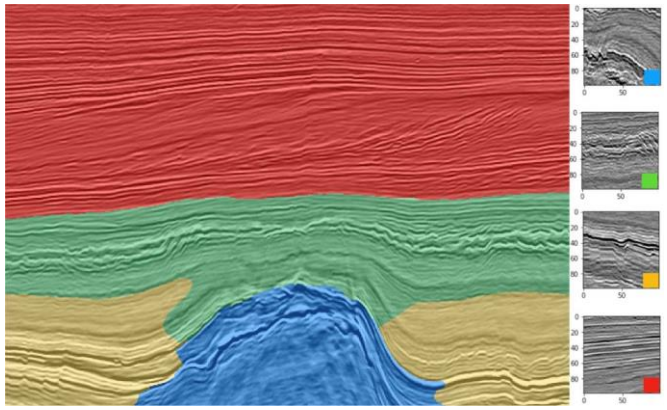




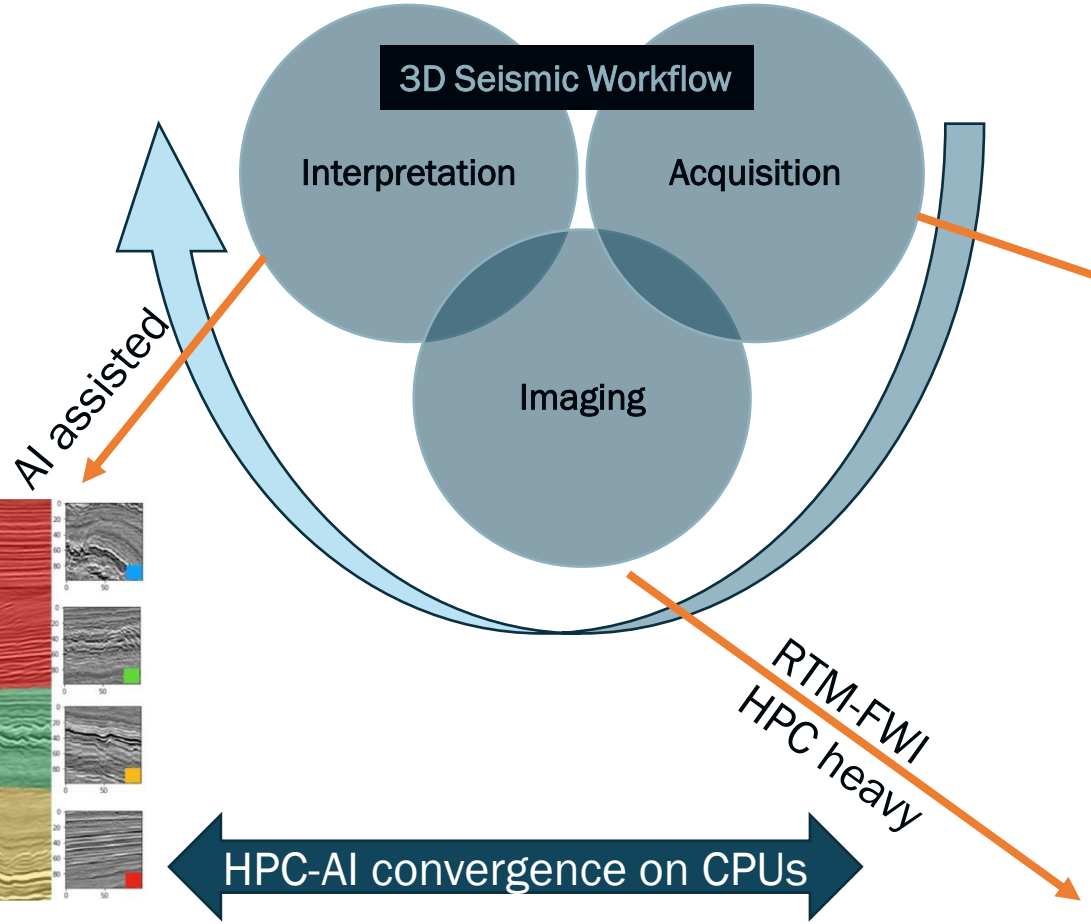
Jnleash Full otential

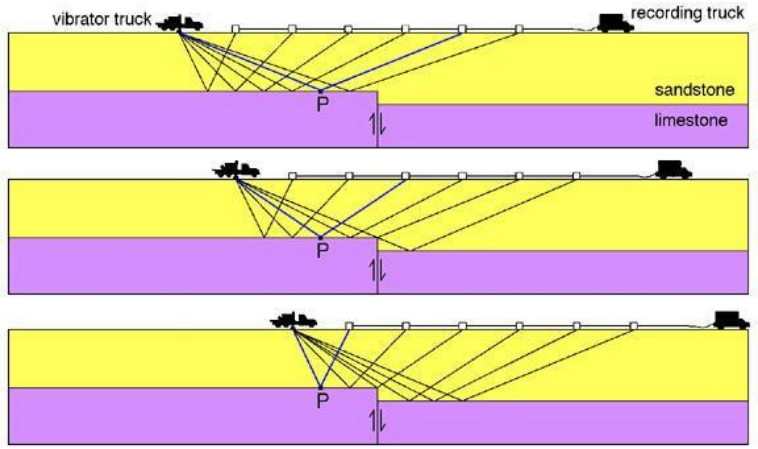
Pain Point in Seismic Analysis — Data Movement

Seismic Analysis is the estimation of shapes and physical properties of Earth's subsurface layers from the returns of sound waves that are propagated through the Earth



https://www.enthought.com/

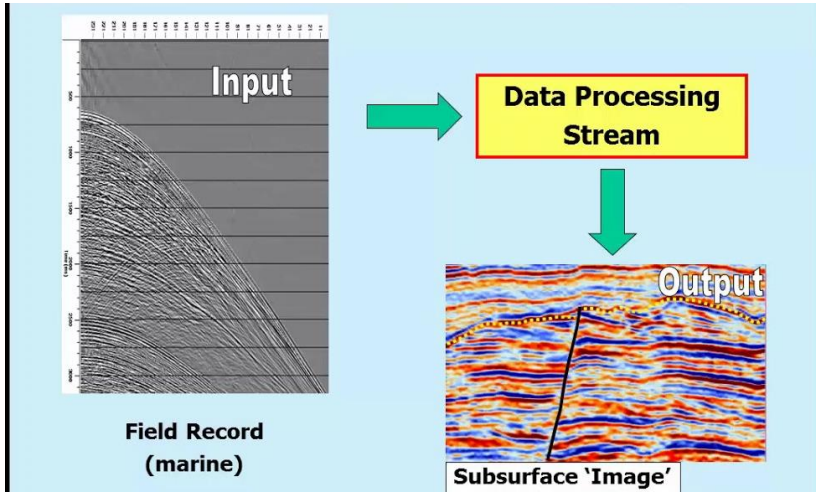




Source: https://www.oilandgaslawyerblog.com/

HPC-AI convergence on CPUs

- Perform Inference where your data is
- Eliminate moving and staging large volumes (in Petabytes) of seismic data
- Eliminate data silos



Source: SelfTraining STO

Accelerate Performance for Deep Learning Inference for 3D Seismic

Objective: General purpose Intel® Xeon® can be used for both HPC and AI workloads in oil and gas – show optimization in AI pipeline

- Full stack solution: Increase performance of DL models for oil & gas datasets on Intel® Xeon® processor using Intel® Distribution of OpenVINO™ toolkit
- A well-cited 3D seismic use-case was used to show this performance boost

Intel® Distribution of OpenVINO™ Toolkit

2020 embedded VISION summit

Tool Suite for High-Performance, Deep Learning Inference

Faster, more accurate real-world results using high-performance, AI and computer vision inference deployed into production across Intel® architecture from edge to cloud

1. Build

Trained Model

TensorFlow Caffe
KALDI mxnet
ONNX

Open Model Zoo
100+ open sourced and optimized pre-trained models;
80+ supported public models

2. Optimize

Model Optimizer
Converts and optimizes trained model using a supported framework

Read, Load, Infer

IR Data

Intermediate Representation (.xml, .bin)

Post-Training Optimization Tool

Deep Learning Workbench

Deep Learning Streamer

OpenCV | OpenCL™

Code Samples & Demos
(e.g. Benchmark app, Accuracy Checker, Model Downloader)

3. Deploy

Inference Engine
Common API that abstracts low-level programming for each hardware

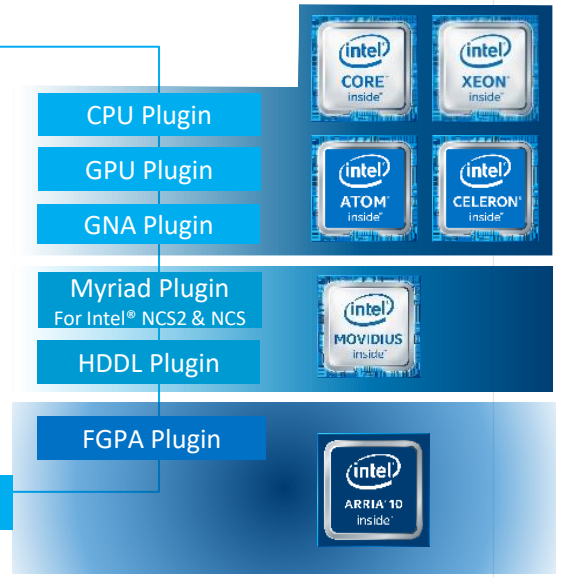
Deployment Manager

CPU Plugin
GPU Plugin
GNA Plugin

Myriad Plugin
For Intel® NCS2 & NCS

HDDL Plugin

FGPA Plugin

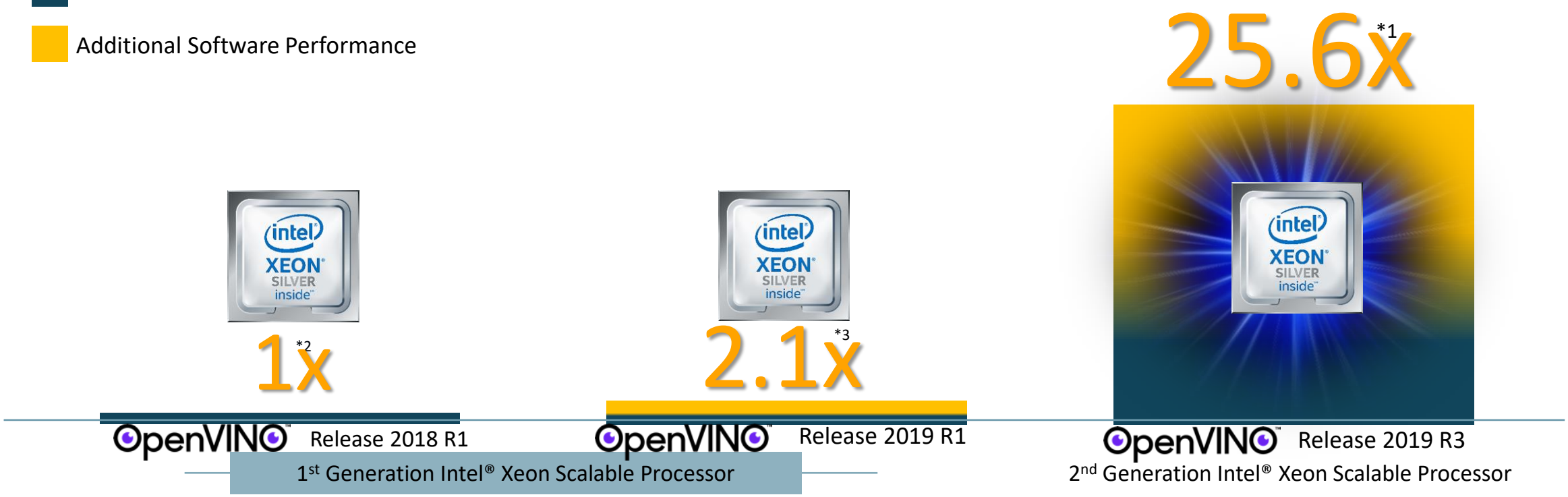


The Compounding Effect of Both Hardware and Software

Improvements Means Exponential Performance

Baseline Performance

Additional Software Performance



For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See backup for configuration details.
Comparison of Frames Per Second utilizing Mobilenet SSD, Batch 1.

The Compounding Effect in Production Deployment

Powered by the Intel® Distribution of OpenVINO™ toolkit

Improvements made by pairing together Intel® architecture-based systems and deep learning acceleration powered by the Intel® Distribution of OpenVINO™ toolkit

Sewer pipe inspection analysis

Inference time was improved with a reduction of up to **80%** using Intel Xeon processors with the OpenVINO toolkit

Solution Brief

Cardiac Examination

Cardiac magnetic resonance imaging (MRI) exams to evaluate heart function, heart chamber volumes and myocardial tissue accelerated by **5.5X**. Learn more

Medical Imaging

Medical imaging accelerated bone age prediction model by **188X** and lung segmentation model by **38X** in inference performance. Learn more

Security Against Social and Digital Attacks

Performance improvements of up to **2.3x faster**, reducing latency by up to **50 percent** for threat detection and remediation to protect businesses against targeted social and digital attacks.

Solution Brief

Operational Improvement

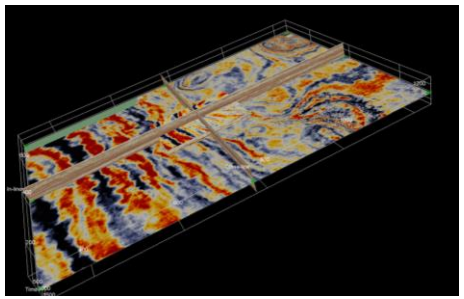
11X increase in performance on Intel® architecture and **19X** with Intel® Vision Accelerator that lead to operational improvements in manufacturing. Learn more

Autonomous Sea Navigation

Autonomous and assisted sea navigation for autonomous ships delivered **4.8X** image throughput compared to unoptimized baseline. Learn more

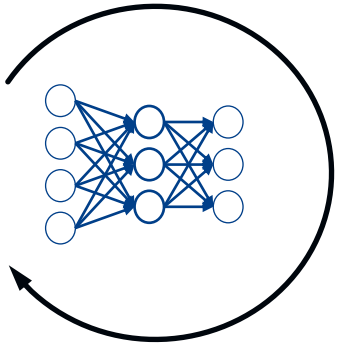
Success Stories ▶ https://intel.com/openvino-success-stories

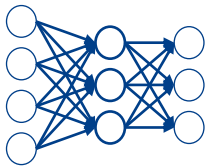
Deep Learning Seismic Workflow



Seismic Data Collection

Data labeling, DL Model Selection & Training

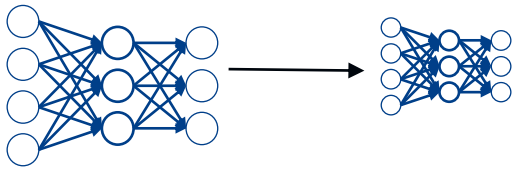




Pre-trained Model

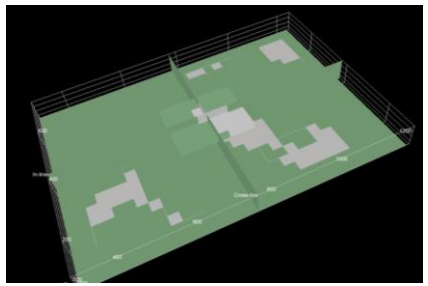
Training

DL Inference (Salt-body identification) via Intel® Distribution of OpenVINO™ toolkit



Model Optimizer (MO)

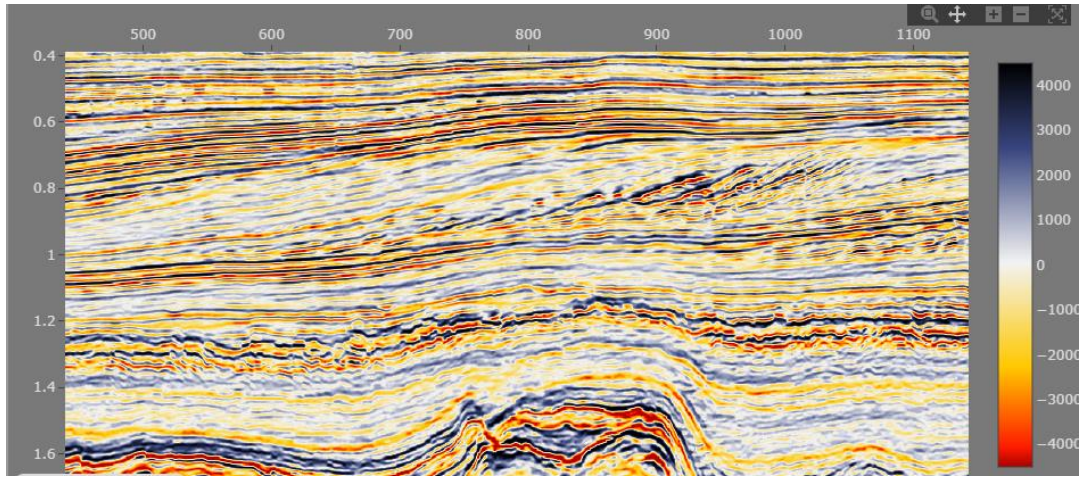
Inference Engine (IE)



Inference/Deployment

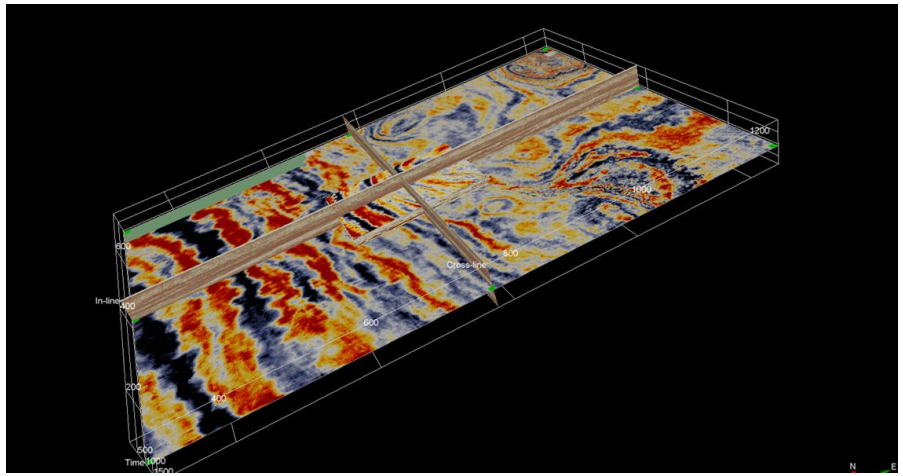
Use-case: Salt Identification in 3D Seismic

- Data used
  - 3D seismic data of F3 Dutch block

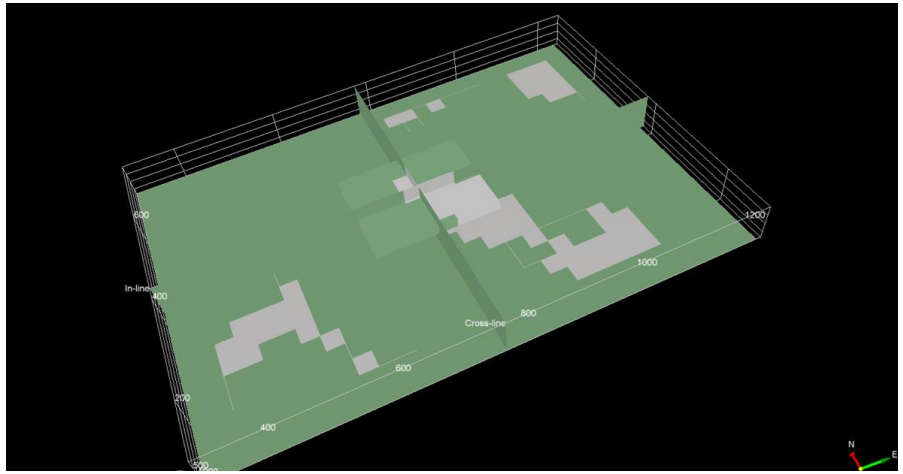


Source: dGB

- Model used
  - CNN for ASI by Waldeland, 2018*



Original F3 3D-seismic section



Salt Prediction

*Convolutional neural networks for automated seismic interpretation, Anders U. Waldeland, Are Charles Jensen, Leiv-J. Gelius, and Anne H. Schistad Solberg, The Leading Edge 2018 37:7, 529-537

Why Salt?

Why Salts are important:

- Salt-bodies are important subsurface structures with significant implications for hydrocarbon accumulation and sealing in offshore petroleum reservoirs
- If Salts are not recognized prior to drilling, they can lead to number of complications if encountered unexpectedly while drilling the well

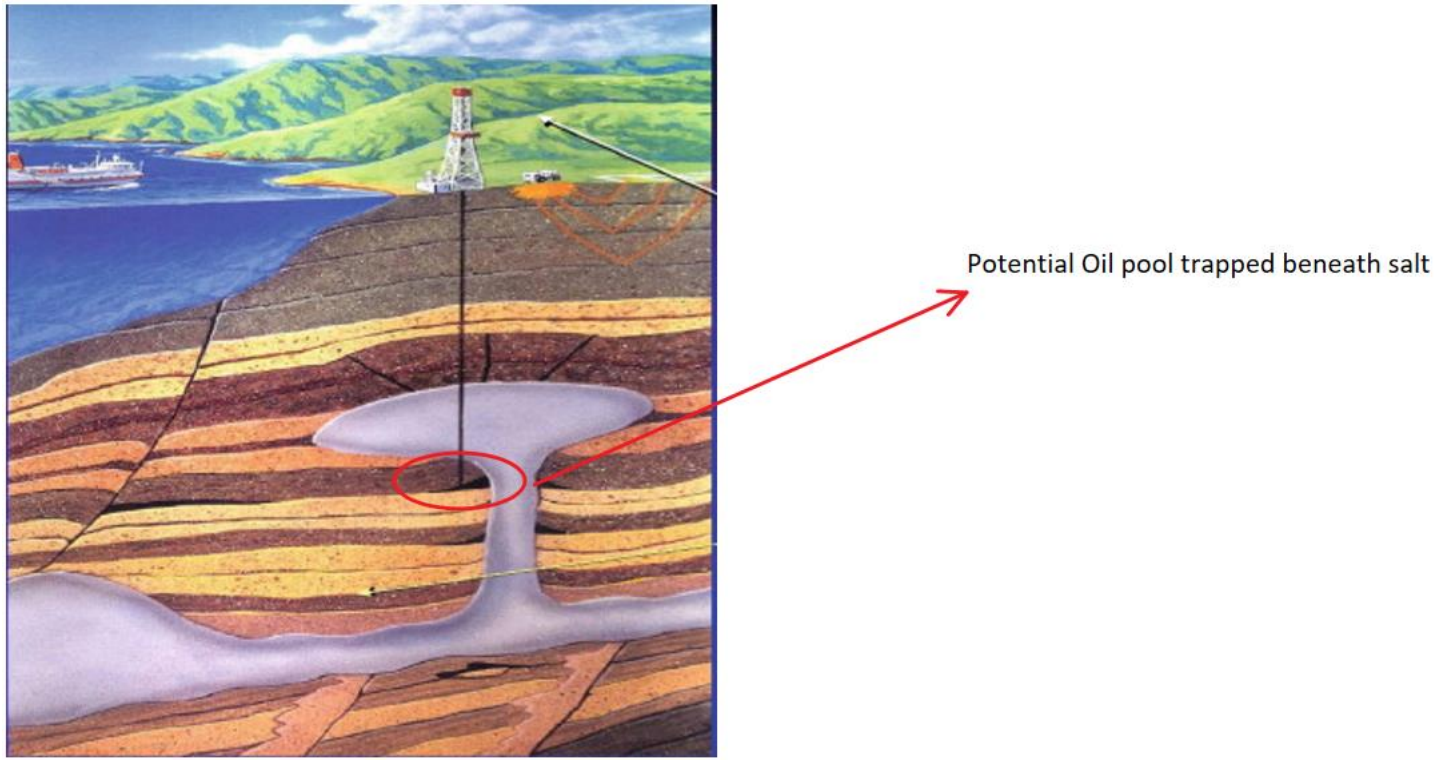
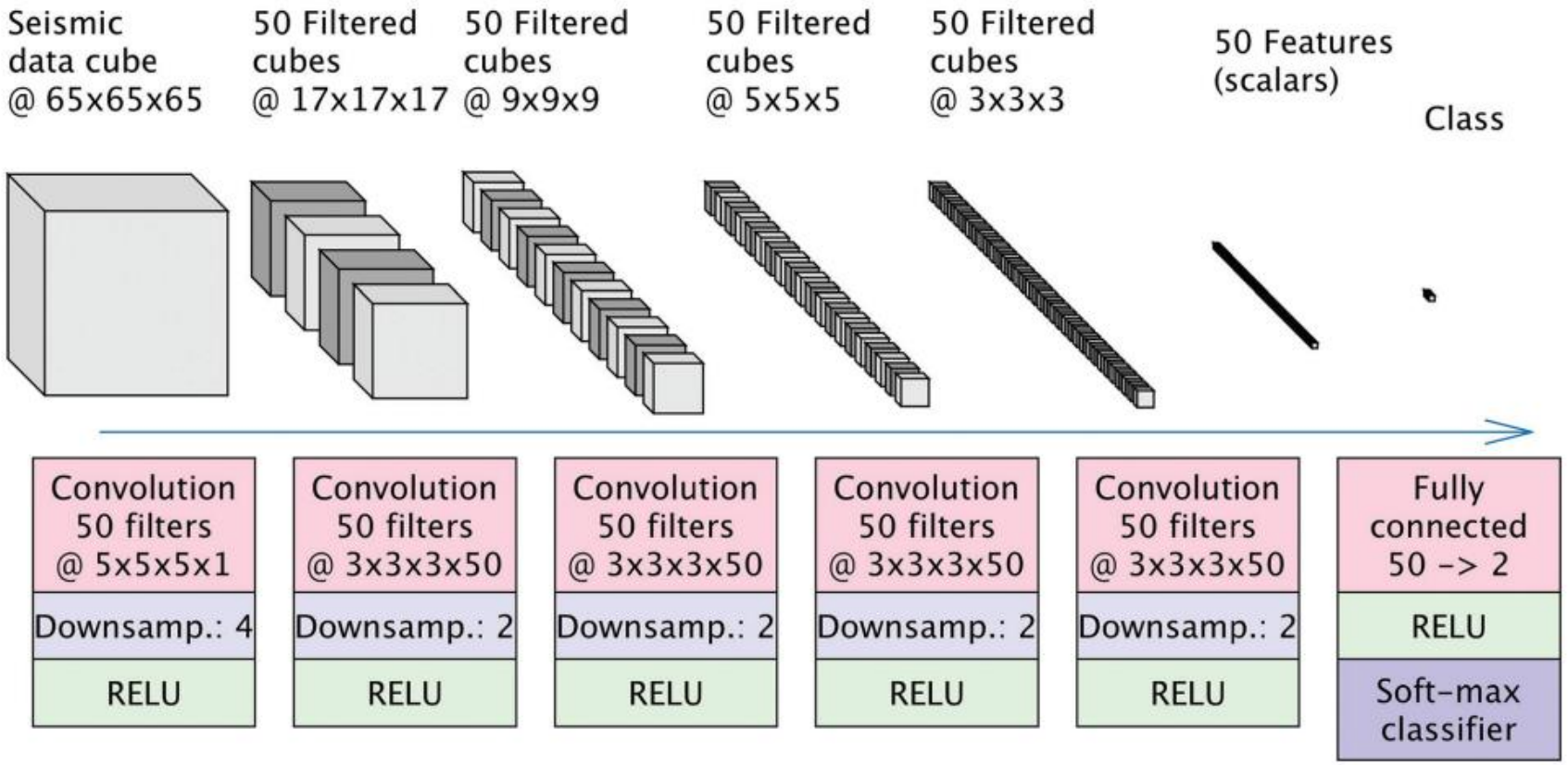


Figure is modified from: https://www.sciencedirect.com/topics/earth-and-planetary-sciences/salt-dome

Salt Identification CNN Model*



The full 3D cube is partitioned into smaller mini cubes of 65 × 65 × 65 samples, which are input into the network – to predict the class of the center pixel

*Convolutional neural networks for automated seismic interpretation, Anders U. Waldeland, Are Charles Jensen, Leiv-J. Gelius, and Anne H. Schistad Solberg, The Leading Edge 2018 37:7, 529-537

Performance Result

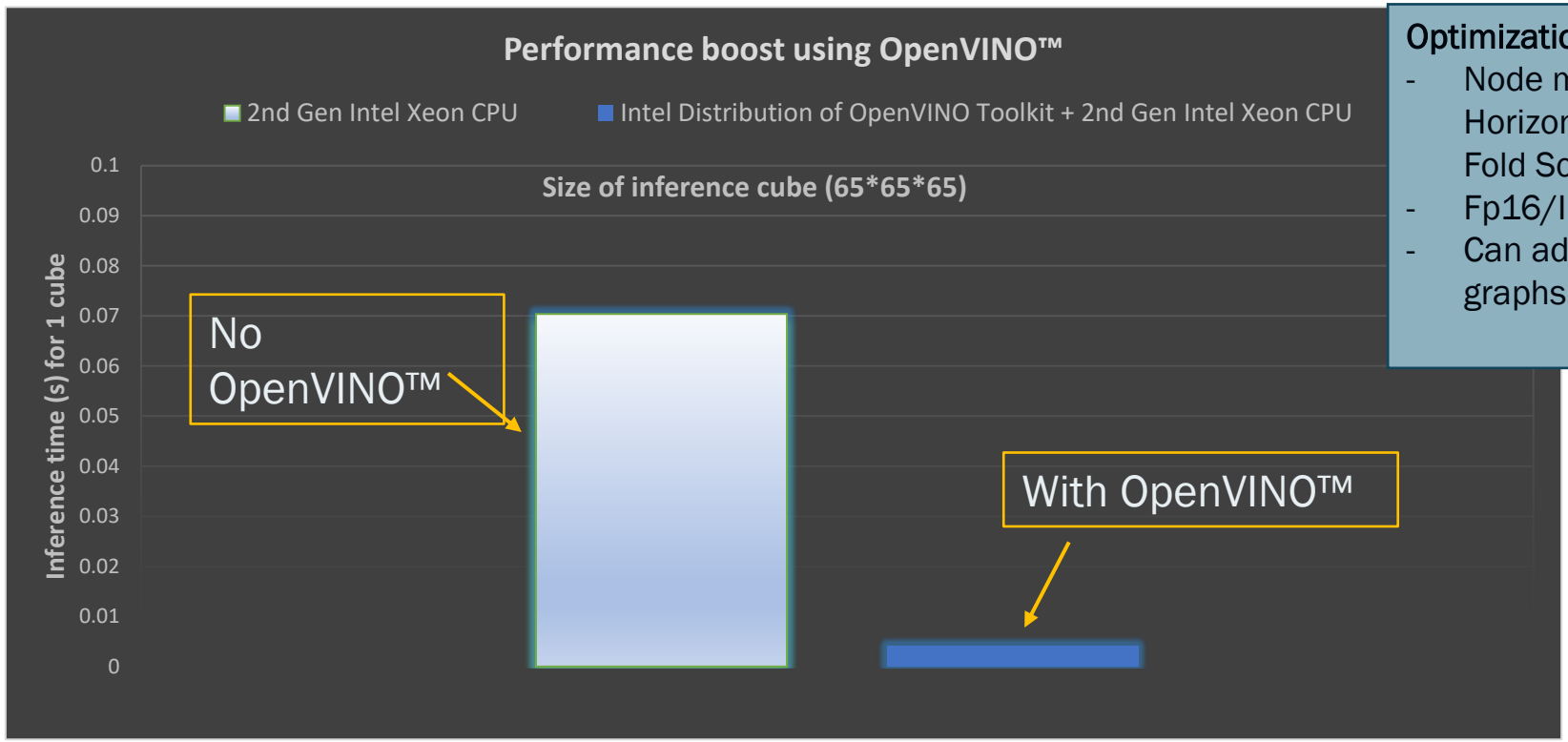
CNN model	System type	OpenVINO™ used (base model)	Inference time* (ms)	Numerical difference in accuracy between OpenVINO™ and base model
5 layers of 3D convolutions (PyTorch model)	Intel Xeon Gold 6252 CPU @ 2.10 GHz (24 cores)	No (PyTorch)	70.20	1.1445E-05
		Yes	4.14	

*Inference time for one 65*65*65 cube

- Intel® Distribution of OpenVINO™ toolkit provides over 15X boost on 2nd generation Xeons
- Details of benchmarking:
  - Data used: 1 Gb 3D seismic data of F3 Dutch block
  - OpenVino: 2019 R2 release (latest version)
  - Python version: 3.6
  - Py-torch version: 1.1.0
  - Cuda version 9.90
  - Complier : GCC 7.4.0
  - OS: Ubuntu

Configurations and benchmark details can be found on Appendix 1. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Performance results are based on testing as shown in configuration and may not reflect all publically available security updates. No product can be absolutely secure. See configuration disclosure for details. Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Performance Result



Optimization Techniques

- Node merging, Constant folding, Horizontal fusion, Batch Norm fusion, Fold Scale Shift convolutions, Dropout
- Fp16/Int8 quantization
- Can add scaling, normalization, cut sub-graphs

- **DL Boost Technology (VNNI) in 2nd gen**
  - Int8 convolutions – not utilized yet
- **Concern:** Do we want int8 for seismic?
  - Larger tiles boosts accuracy far exceeding the loss from quantization

Configurations and benchmark details can be found on Appendix 1. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Performance results are based on testing as shown in configuration and may not reflect all publically available security updates. No product can be absolutely secure. See configuration disclosure for details. Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Intel R&D: Small V/S Large Tiles — Accuracy is Better With Larger!

Springer for Research & Development

International MICCAI Brainlesion Workshop

BrainLes 2018: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries pp 25-36 | Cite as

Adverse Effects of Image Tiling on Convolutional Neural Networks

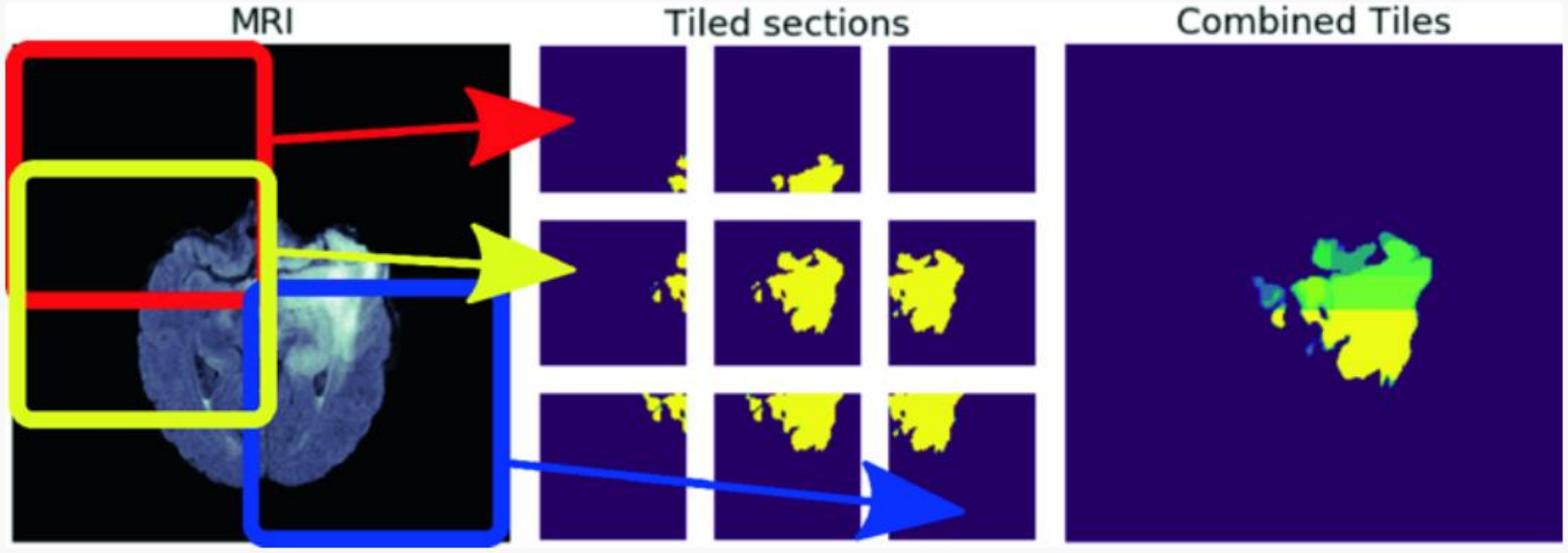
Authors Authors and affiliations

G. Anthony Reina, Ravi Panchumarthy

Conference paper

First Online: 26 January 2019

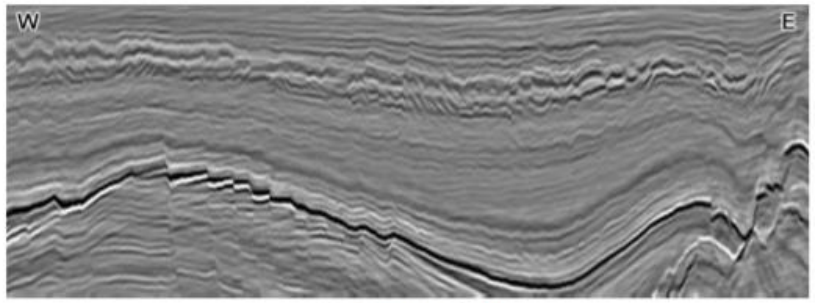
Inf: FULL	Inf: TILED
Inf Mem Req: High	Inf Mem Req: Low
Inf Accuracy: High	Inf Accuracy: Low



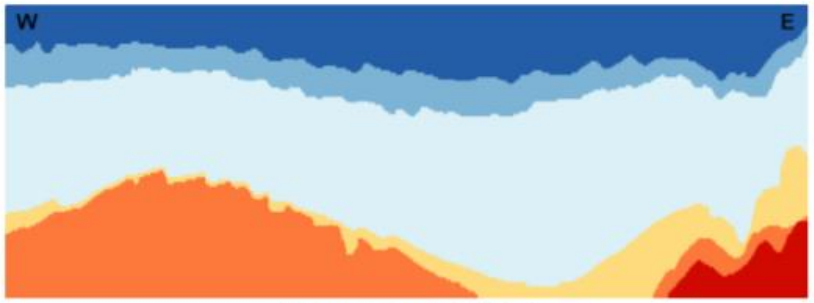
- "Tiling the input to CNN models—while perhaps necessary to overcome the memory limitations in computer hardware—may lead to undesirable and unpredictable errors"

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7020775/

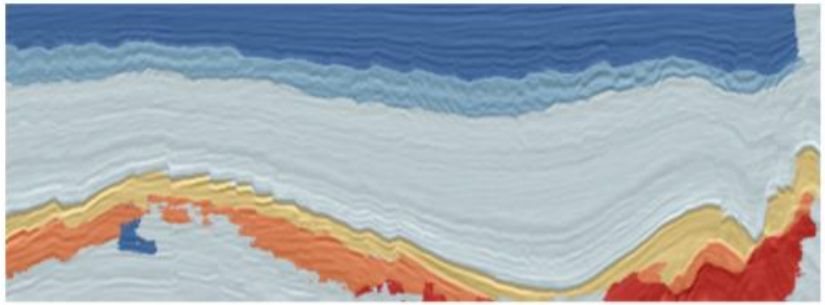
Larger Seismic tile gives better definition for Inference



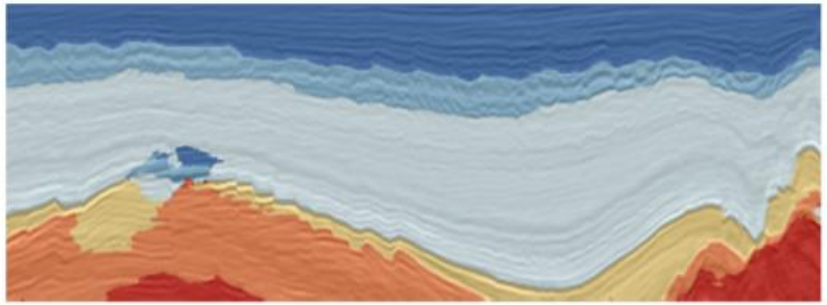
(a) Seismic data



(b) Ground truth labels



(g) Patch-based + aug + skip



(h) Section-based + aug + skip

- Larger tiles boost accuracy
- Spatial context is imp in seismic

Research:

- *Int8 performance on seismic: blog 3 in Resource slide*
- *Int8 on larger tiles! – white paper coming*

https://arxiv.org/pdf/1901.07659.pdf

Model \ Metric	PA	Class Accuracy						MCA	FWIU
		Zechstein	Scruff	Rijnland/Chalk	Lower N. S.	Middle N. S.	Upper N. S.		
Patch-based model	0.788	0.264	0.074	0.499	0.992	0.804	0.754	0.565	0.640
Patch-based + aug.	0.852	0.434	0.221	0.707	0.974	0.884	0.916	0.689	0.743
Patch-based + aug + skip	0.862	0.458	0.286	0.673	0.974	0.912	0.926	0.705	0.757
Section-based model	0.879	0.219	0.539	0.744	0.951	0.872	0.973	0.716	0.789
Section-based + aug.	0.901	0.714	0.423	0.812	0.979	0.940	0.956	0.804	0.844
Section-based + aug + skip	0.905	0.602	0.674	0.772	0.941	0.938	0.974	0.817	0.832

Key Takeaway

The general-purpose 2nd generation Intel® Xeon® Scalable Processor together with Intel® Distribution of OpenVINO™ toolkit provides DL performance that is well suited for Oil and Gas datasets

- ✓ Use OpenVINO™ toolkit to boost inference performance for oil and gas datasets on Intel® Xeon® processor
- ✓ Int 8 quantization with larger tiles on 2nd generation Intel® Xeon® processor gives even better performance boost for seismic data – see *blog 3 in Resources slide*
- ✓ Scale DL on existing CPUs that are used for HPC
- ✓ Use CPUs for other workloads when not doing DL

Resource Slide

2020 embedded VISION summit

Resources for OpenVINO™ toolkit

OpenVINO™ toolkit guide

Download OpenVINO™ toolkit

Resource for Seismic

Seismic model used in this work

Blogs published on OpenVINO™ toolkit + seismic

Blog 1: Accelerating fault detection in 3D Seismic

Blog 2: Accelerating seismic Interpretation using OpenVINO™

Blog 3: Lower Precision Pipeline for Seismic Interpretation

2020 Embedded Vision Summit

Intel Booth

intel

Workshops and demos to visit

General Session Speaker:

- Bill Pearson, VP IOTG, GM Developer Enabling, Intel, Tuesday, September 15, 10:00 a.m. to 10:30 a.m. PDT: **Streamline, Simplify and Solve for the Edge of the Future**

In-depth technical workshops

- Friday, September 18, 9:00 a.m. to 1:30 p.m. PDT: **Using the Intel® Distribution of the OpenVINO™ Toolkit for Deploying Accelerated Deep Learning Applications**
- Monday, September 21, 9:00 a.m. to 1:30 p.m. PDT: **Intel's Edge AI for Retail**
- Wednesday, September 23, 9:00 a.m. to 1:30 p.m. PDT: **Intel's Edge AI for Industrial**

Technical presentations

- Alexander Kozlov, Deep Learning R&D engineer, Intel: **Recent Advances in Post-Training Quantization**
- Dr. Manas Pathak, Global AI lead for oil and gas, Intel: **Acceleration of Deep Learning for 3D Seismic**
- Tara K. Thimmanaik, solutions architect, Intel: **Smarter Manufacturing Achieved with Intel's Deep Learning-Based Machine Vision**
- Gary Brown, Director of AI Marketing, Intel: **Getting Efficient DL Inference Performance: Is It Really All About The TOPS?**
- Rama Karamsetty, Edge AI Marketing Manager, Intel: **Edge Inferencing-- Scaling w/ Vision Accelerator Cards**
- Vaidyanathan Krishnamoorthy, edge inference solutions architect, Intel: **Federated Edge Inferencing**

Dedicated demos and networking space