WELCOME TO
INTEL AI DEVCON
OPERANT CONDITIONING 1938
SPIKING NEURON 1952
FIRST NEUROSCIENCE DEPARTMENT 1964
TRANSISTOR 1947
FIRST COMPUTER SCIENCE DEPARTMENT 1962
THE TURING MACHINE 1936
INTEL FOUNDED 1968
FIRST 1 BILLION TRANSISTOR PROCESSOR mid 2000s
FIRST COMPUTER SCIENCE DEPARTMENT 1962
DEEP LEARNING PREVALENCE mid 2000s
Deep learning prevalence
THE AI REVOLUTION IS REALLY A COMPUTING EVOLUTION
OUR VISION OF AI
BEST TOOLS TO BUILD TOWARDS

THE BEST POSSIBLE FUTURE
ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION

TOOLS

HARDWARE

COMMUNITY
Enable developers to achieve their AI vision

Tools
Hardware
Community
JASON KNIGHT
HEAD OF SOFTWARE PRODUCTS
INTEL AI
OPEN SOURCE

INTEL® MATH KERNEL LIBRARY FOR DEEP NEURAL NETWORKS (INTEL® MKL-DNN)

HELPS REALIZE THE INCREDIBLE BENEFITS OF DIRECT OPTIMIZATION

PRIMITIVES

MATRIX MULTIPLICATION

BATCH NORM

NORMALIZATION

POOLING

ACTIVATION

CONVOLUTION

Intel/mkl-dnn
OPTIMIZING TENSORFLOW TO SUPERCHARGE AI WORKLOADS

RAJAT MONGA
TENSORFLOW ENGINEERING DIRECTOR
GOOGLE
nGRAPHTM
OPEN SOURCE COMPILER ENABLING FLEXIBILITY TO RUN MODELS ACROSS A VARIETY OF FRAMEWORKS AND HARDWARE

NervanaSystems/ngraph
Other names and brands may be claimed as the property of others.
ENABLING DEEP LEARNING TO TAKE ADVANTAGE OF SCALABLE SPARK AND HADOOP CLUSTERS

BigDL: DISTRIBUTED DEEP LEARNING

intel-analytics/BigDL
USING DATA SCIENCE TO
IMPROVE PATIENT OUTCOMES

DR. RACHAEL CALLCUT
UCSF DIRECTOR OF DATA SCIENCE AND
TRAUMA SURGEON

Other names and brands may be claimed as the property of others
USING DATA SCIENCE TO IMPROVE PATIENT OUTCOMES
VIDEO IS THE ULTIMATE IOT SENSOR

- Emergency Response
- Financial Services
- Machine Vision
- Cities/Transportation
- Autonomous Vehicles
- Responsive Retail
- Manufacturing
- Public Sector
ANNOUNCING: OpenVINO™ SOFTWARE TOOLKIT
VISUAL INFERENCING AND NEURAL NETWORK OPTIMIZATION

DEPLOY COMPUTER VISION AND DEEP LEARNING CAPABILITIES TO THE EDGE

WRITE ONCE + SCALE TO DIVERSE ACCELERATORS

BROAD FRAMEWORK SUPPORT

HIGH PERFORMANCE, HIGH EFFICIENCY FOR THE EDGE

Other names and brands may be claimed as the property of others.

VPU = Vision Processing Unit (Movidius)
ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION

TOOLS

HARDWARE

COMMUNITY
ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION

TOOLS

HARDWARE

COMMUNITY
Applied Machine Learning at Facebook:
A Datacenter Infrastructure Perspective

Facebook’s mission is to ‘give people the power to build community and bring the world closer together.’ In support of that mission, Facebook connects more than two billion people as of December 2015. Meanwhile, the past several years have seen a revolution in the application of machine learning to real-world problems at this scale, resulting upon the virtuous cycle of machine learning algorithm innovation, exorbitant amounts of training data for models, and advances, in high-performance computer architectures. As Facebook’s machine learning provides key capabilities, including nearly all aspects of user experience including serving images, the top 20 news feed posts for News Feed, search, and two translations, and gate, and real-time video classification [1, 14].

Facebook leverages a wide variety of machine learning algorithms in these services, including support vector machines, random forests, boosted decision trees, and many styles of non-parametric decision trees. This paper describes several important aspects of Facebook’s datacenter infrastructure that supports machine learning training at scale. The infrastructure includes internal “ML-as-a-Service” flows, open-source machine learning frameworks, and high-performance computing (HPC) clusters. Facebook leverages a large fleet of CPUs and GPU view, Facebook leverages a large fleet of CPUs and GPUs for training and deploying machine learning models in order to support the necessary training frequencies at the required scale. Facebook primarily relies on GPU computing for machine learning inference.

L. Infrastructure

Facebook’s mission is to ‘give people the power to build community and bring the world closer together.’ In support of that mission, Facebook connects more than two billion people as of December 2015. Meanwhile, the past several years have seen a revolution in the application of machine learning to real-world problems at this scale, resulting upon the virtuous cycle of machine learning algorithm innovation, exorbitant amounts of training data for models, and advances, in high-performance computer architectures. As Facebook’s machine learning provides key capabilities, including nearly all aspects of user experience including serving images, the top 20 news feed posts for News Feed, search, and two translations, and gate, and real-time video classification [1, 14].

Facebook leverages a wide variety of machine learning algorithms in these services, including support vector machines, random forests, boosted decision trees, and many styles of non-parametric decision trees. This paper describes several important aspects of Facebook’s datacenter infrastructure that supports machine learning training at scale. The infrastructure includes internal “ML-as-a-Service” flows, open-source machine learning frameworks, and high-performance computing (HPC) clusters. Facebook leverages a large fleet of CPUs and GPU view, Facebook leverages a large fleet of CPUs and GPUs for training and deploying machine learning models in order to support the necessary training frequencies at the required scale. Facebook primarily relies on GPU computing for machine learning inference.

L. Infrastructure

Facebook’s mission is to ‘give people the power to build community and bring the world closer together.’ In support of that mission, Facebook connects more than two billion people as of December 2015. Meanwhile, the past several years have seen a revolution in the application of machine learning to real-world problems at this scale, resulting upon the virtuous cycle of machine learning algorithm innovation, exorbitant amounts of training data for models, and advances, in high-performance computer architectures. As Facebook’s machine learning provides key capabilities, including nearly all aspects of user experience including serving images, the top 20 news feed posts for News Feed, search, and two translations, and gate, and real-time video classification [1, 14].

Facebook leverages a wide variety of machine learning algorithms in these services, including support vector machines, random forests, boosted decision trees, and many styles of non-parametric decision trees. This paper describes several important aspects of Facebook’s datacenter infrastructure that supports machine learning training at scale. The infrastructure includes internal “ML-as-a-Service” flows, open-source machine learning frameworks, and high-performance computing (HPC) clusters. Facebook leverages a large fleet of CPUs and GPU view, Facebook leverages a large fleet of CPUs and GPUs for training and deploying machine learning models in order to support the necessary training frequencies at the required scale. Facebook primarily relies on GPU computing for machine learning inference.

L. Infrastructure

Facebook’s mission is to ‘give people the power to build community and bring the world closer together.’ In support of that mission, Facebook connects more than two billion people as of December 2015. Meanwhile, the past several years have seen a revolution in the application of machine learning to real-world problems at this scale, resulting upon the virtuous cycle of machine learning algorithm innovation, exorbitant amounts of training data for models, and advances, in high-performance computer architectures. As Facebook’s machine learning provides key capabilities, including nearly all aspects of user experience including serving images, the top 20 news feed posts for News Feed, search, and two translations, and gate, and real-time video classification [1, 14].

Facebook leverages a wide variety of machine learning algorithms in these services, including support vector machines, random forests, boosted decision trees, and many styles of non-parametric decision trees. This paper describes several important aspects of Facebook’s datacenter infrastructure that supports machine learning training at scale. The infrastructure includes internal “ML-as-a-Service” flows, open-source machine learning frameworks, and high-performance computing (HPC) clusters. Facebook leverages a large fleet of CPUs and GPU view, Facebook leverages a large fleet of CPUs and GPUs for training and deploying machine learning models in order to support the necessary training frequencies at the required scale. Facebook primarily relies on GPU computing for machine learning inference.
<table>
<thead>
<tr>
<th>Services</th>
<th>Relative Capacity</th>
<th>Compute</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>News Feed</td>
<td>100X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
<tr>
<td>Pacer</td>
<td>10X</td>
<td>Single-Socket CPU</td>
<td>Low</td>
</tr>
<tr>
<td>Lumos</td>
<td>10X</td>
<td>Single-Socket CPU</td>
<td>Low</td>
</tr>
<tr>
<td>Search</td>
<td>10X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
<tr>
<td>Language Translation</td>
<td>1X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
<tr>
<td>Sigma</td>
<td>1X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>1X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
</tbody>
</table>

**TABLE III**

RESOURCE REQUIREMENTS OF ONLINE INFERENCE WORKLOADS.
A SCIENTIFIC COLLABORATION BETWEEN INTEL AND NOVARTIS

KUSHAL DATTA, PHD
RESEARCH SCIENTIST
INTEL AI
26X LARGER MULTIPLE OBJECTS

1024 X 1280 X 3

Other names and brands may be claimed as the property of others.
Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors.
Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.
HIGH PERFORMANCE AT SCALE

SCALING OF TIME TO TRAIN

Intel® Omni-Path Architecture, Horovod and TensorFlow®

Speedup compared to baseline 1.0 measured in time to train 1 node

- **1 Node**: Time to Train
- **2 Nodes**: Speedup 6.6X
- **4 Nodes**: Further speedup
- **8 Nodes**: Significant increase

TOTAL MEMORY USED

192GB DDR4 PER INTEL® 2S XEON® 6148 PROCESSOR

- **1 Node**: 64.3GB
- **2 Nodes**: 128.6GB
- **4 Nodes**: 257.2GB
- **8 Nodes**: 514.4GB

MULTISCALE CONVOLUTION NEURAL NETWORK

- Optimized Libraries:
  - Intel® MKL/MKL-DNN, cDNN, DAAL

OPTIMIZED LIBRARIES

INTEL® OMNI-PATH ARCHITECTURE

Other names and brands may be claimed as the property of others.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.
ARTIFICIAL INTELLIGENCE AND INTEL® XEON® PROCESSORS
MACHINE LEARNING IMAGE RENDERING

JAMES JACOBS
CEO OF ZIVA

Other names and brands may be claimed as the property of others
ZIVA VFX Authoring Tools

Shipped as a Maya Plugin

DISTRIBUTED TO SERVER FARM FOR SHOT RENDERS

ZIVA VFX

Intel Paradiso
Intel MKL
Intel Lapack
Intel BLAS

Bones/Muscles
Fascia
Fat and Skin

ZIVA CHARACTERS SIMULATED IN PASSES

ZIVA FEM PHYSICS SOLVER

Intel MKL
Intel Paradiso
Intel BLAS
Intel Lapack

= BIG SHARKS

ZIVA INTEGRATION FRAMEWORK

REAL-TIME TRAINING
CHARACTER TRANSFER AUTOMATION
VOLUMETRIC CAPTURE AUGMENTATION

A.I & M.L TECHNOLOGIES

DISTRIBUTED FUNCTIONALITY AND EMBEDDED PLAYERS

NATIVE NODE GRAPH UI WITH ROBUST API
EMBEDDED PLAYERS IN ANY SOFTWARE
DISTRIBUTED GRAPHS

CLOUD SERVICES

Other names and brands may be claimed as the property of others
FLEXIBLE REAL-TIME INFERENCE

FPGA PRODUCTS
AI For Earth

GPU: 62.4 Images/Seconds
5000 images processed

FPGA: 506.9 Images/Seconds
500,000 images processed
FLEXIBLE REAL-TIME INFERENCEING
FPGA PRODUCTS
• Deploy DNN and Computer Vision at the Edge
• Native FP16 and Fixed Point 8 bit support
• 4 TOPS with 1 TOPS of DNN Compute at 1W
GOALS - HIGH UTILIZATION & MODEL PARALLELISM
Theoretical Reality

Chip X Lake Crest

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance

Chip X GEMM based on DeepBench training data for A(5124, 9124), B(9124,2560) matrix GEMM operations performing DeepSpeech using FP16+ mixed precision at 27.43 TOPs. Source: Lake Crest, Based on Intel measurements on limited distribution SDV, General Matrix-Matrix Multiplication; A(1536, 2048), B(2048, 1536).
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations, and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Chip X GEMM based on DeepBench training data for A(5124, 9124), B(9124, 2560) matrix size GEMM operations performing DeepSpeech using FP16+ mixed precision at 27.43 TOPs.

Source: Lake Crest: Based on Intel measurements on limited distribution SDV 1 General Matrix-Matrix Multiplication; A(1536, 2048), B(2048, 1536) 2 Two chip vs. single chip GEMM performance; A(6144, 2048), B(2048, 1536) 3 Full Chip MRB CHIP MRB data movement using send/recv. Tensor size = (1, 32), average across 50K iterations.

Chip X Lake Crest

0 1

TOPs

Theoretical

Reality

High Utilization & Model Parallelism

GEMM Operation Utilization\(^1\)

\[
\begin{array}{c|c|c}
 & \text{A(1536, 2048)} & \text{B(2048, 1536)} \\
\hline
\text{Chip X} & 96.4\% & \\
\hline
\text{Lake Crest} & 96.2\% & \\
\end{array}
\]

Multi-Chip Scaling\(^2\)

\[
\begin{array}{c|c|c}
 & \text{A(6144, 2048)} & \text{B(2048, 1536)} \\
\hline
\text{Chip X} & 96.2\% & \\
\hline
\text{Lake Crest} & 96.2\% & \\
\end{array}
\]

Multi-Chip Communication\(^3\)

\[
\begin{array}{c|c|c}
 & \text{OFF CHIP BANDWIDTH} & \text{LATENCY} \\
\hline
\text{Chip X} & \text{<210W} & \text{<790ns} \\
\hline
\text{Lake Crest} & \text{2.4 Tbps} & \\
\end{array}
\]

Power <210W
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operation, and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Chip X GEMM based on DeepBench training data for A(5124, 9124), B(9124,2560) matrix size GEMM operations performing DeepSpeech using FP16+ mixed precision at 27.43 TOPs.

Source: Lake Crest - Based on Intel measurements on limited distribution SDV.

Source: Spring Crest - Intel measurements on simulated product.
FIRST COMMERCIAL NNP
INTEL® NERVANA™
NNP L-1000
in 2019
3-4x training performance
of first generation
Lake Crest product

SPRING CREST
PURPOSE BUILT DESIGN OPTIMIZED ACROSS
MEMORY BANDWIDTH, UTILIZATION, AND POWER
ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION

TOOLS

HARDWARE

COMMUNITY
ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION

TOOLS

HARDWARE

COMMUNITY
PARTNERSHIPS AND OUR DEVELOPER COMMUNITY
AI AND IOT

THOMAS SIEBEL
CHAIRMAN AND CHIEF EXECUTIVE OFFICER
C3 IOT

Other names and brands may be claimed as the property of others
Machine Learning at Amazon: a long heritage

- Personalized recommendations
- Fulfillment automation / inventory management
- Cargo
- Voice driven interactions
- Inventing entirely new customer experiences
Amazon SageMaker

**BUILD**
- Pre-built notebooks for common problems
- Built-in, high performance algorithms

**TRAIN**
- One-click training
- Hyperparameter optimization

**DEPLOY**
- One-click deployment
- Fully managed hosting with auto-scaling
Other names and brands may be claimed as the property of others
GOAL IS TO SELECT A SET OF ML PROBLEMS, EACH DEFINED BY A DATASET AND QUALITY TARGET, THEN MEASURE THE WALL CLOCK TIME TO TRAIN A MODEL FOR EACH PROBLEM.

For more information, see: https://mlperf.org/

Other names and brands may be claimed as the property of others.
HELPING MAKE YOUR AI VISION A REALITY
RISK FACTORS

Today’s presentation contains forward-looking statements. All statements made that are not historical facts are subject to a number of risks and uncertainties, and actual results may differ materially. Please refer to our most recent earnings release, Form 10-Q and 10-K filing available on our website for more information on the risk factors that could cause actual results to differ.
DISCLAIMERS

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to www.intel.com/benchmarks.

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. Notice Revision #20110804 Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at [Intel.com].

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.

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.

Intel, the Intel logo, Xeon, Intel Nervana and Intel Movidius are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

© Intel Corporation.