OCP SUMMIT

March 20-21
2018
San Jose, CA

OPEN. FOR BUSINESS.

INTEL DEEP LEARNING PLATFORM

Jordan Plawner, Sr. Director, Product Management
Artificial Intelligence Product Group, Intel

AI Adoption in the Enterprise is Just Beginning

The AI Opportunity: Democratize AI

Dedicated to assist with or take over specific tasks

NARROW AI *DRIVEN BY INDUSTRY*

STRONG AI *DRIVEN BY IT*

PLANNING
Smart R&D and forecasting

PRODUCTION
Optimize production and maintenance

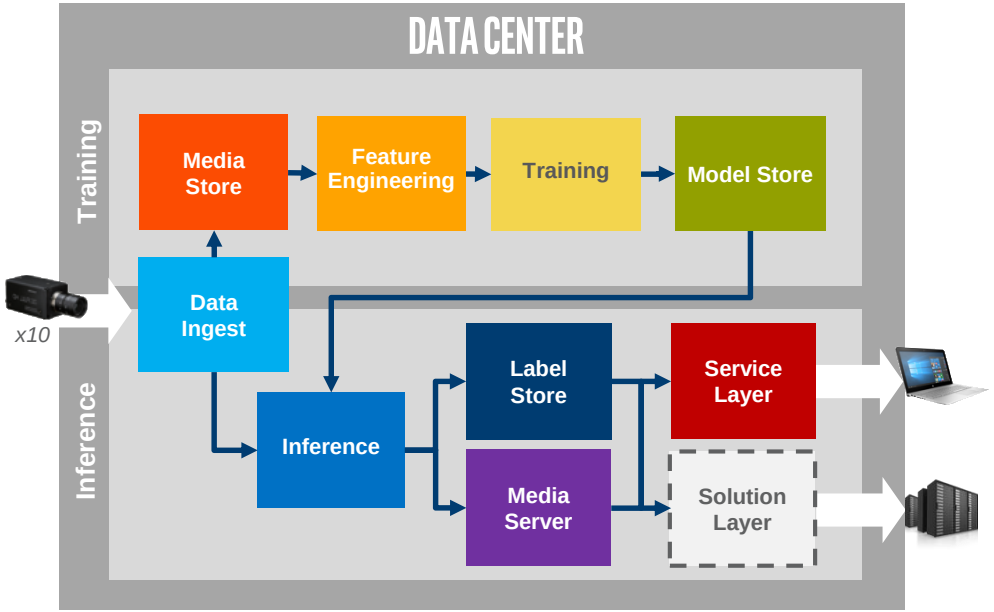
PROMOTION
Targeted sales and marketing

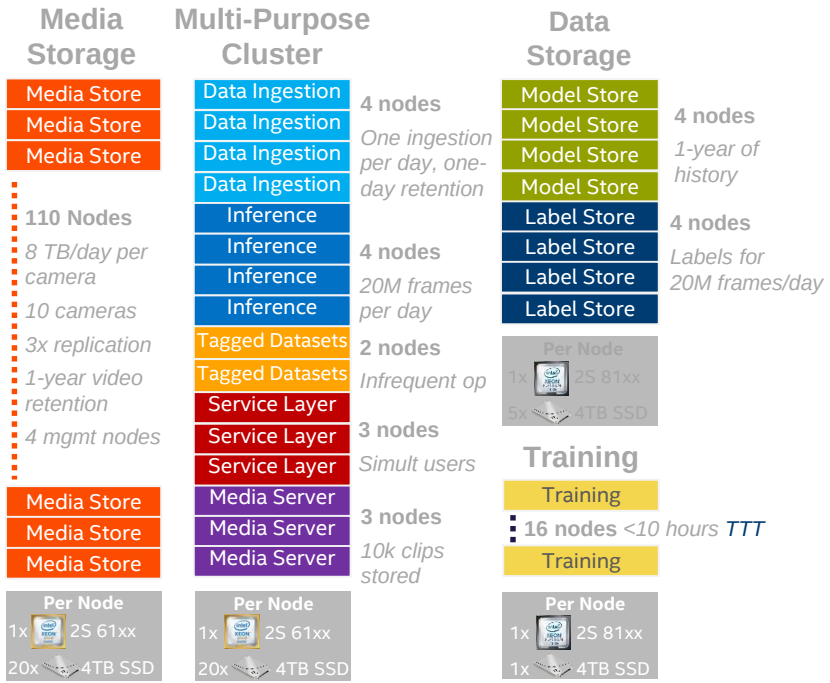
PROVIDE
Enhanced user experiences

IT TECHNICAL SERVICES
Image classification, language translation, facial recognition, speech-2-text, speech generation, time series…

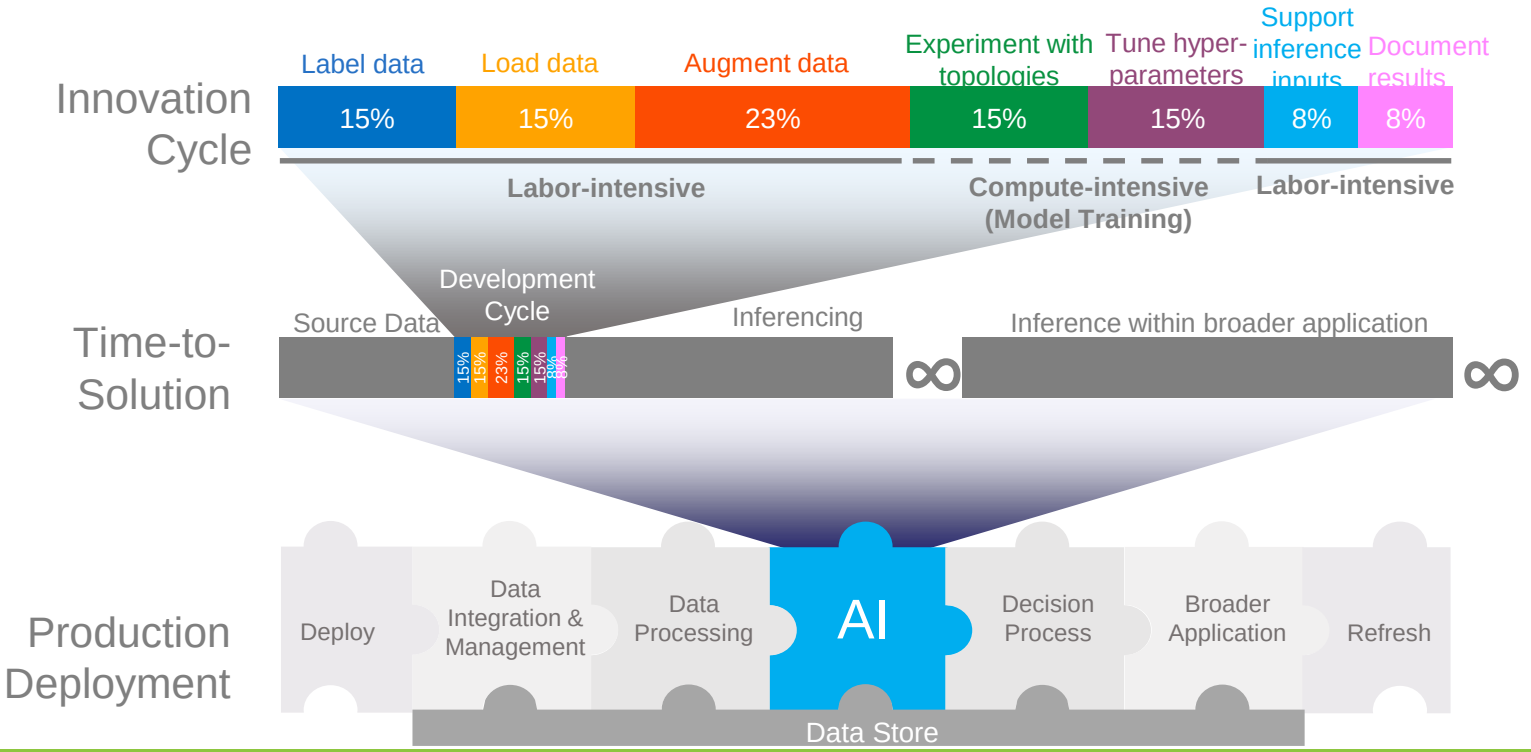
PRODUCTION OPERATIONS
Data pipeline and workflow, logging and monitoring, feature store, life cycle management…

Deploying Deep Learning





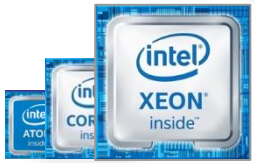
Deploying Deep Learning Time-to-Solution



Intel AI Portfolio

If it computes and is connected it will do AI

GENERAL AI



Mainstream AI

+

Flexible Acceleration

DEEP LEARNING

TRAINING

DATA CENTER/ WORKSTATION

Mainstream Training

+

Intensive Training

INFERENCE

DATA CENTER/ WORKSTATION

Mainstream Inference

+

To be announced

Intensive Inference

Real-time Inference

GATEWAY/EDGE

Mainstream Inference

+

Higher Inference Throughput

Vision 1-20W

Speech/Audio 1-100+mW

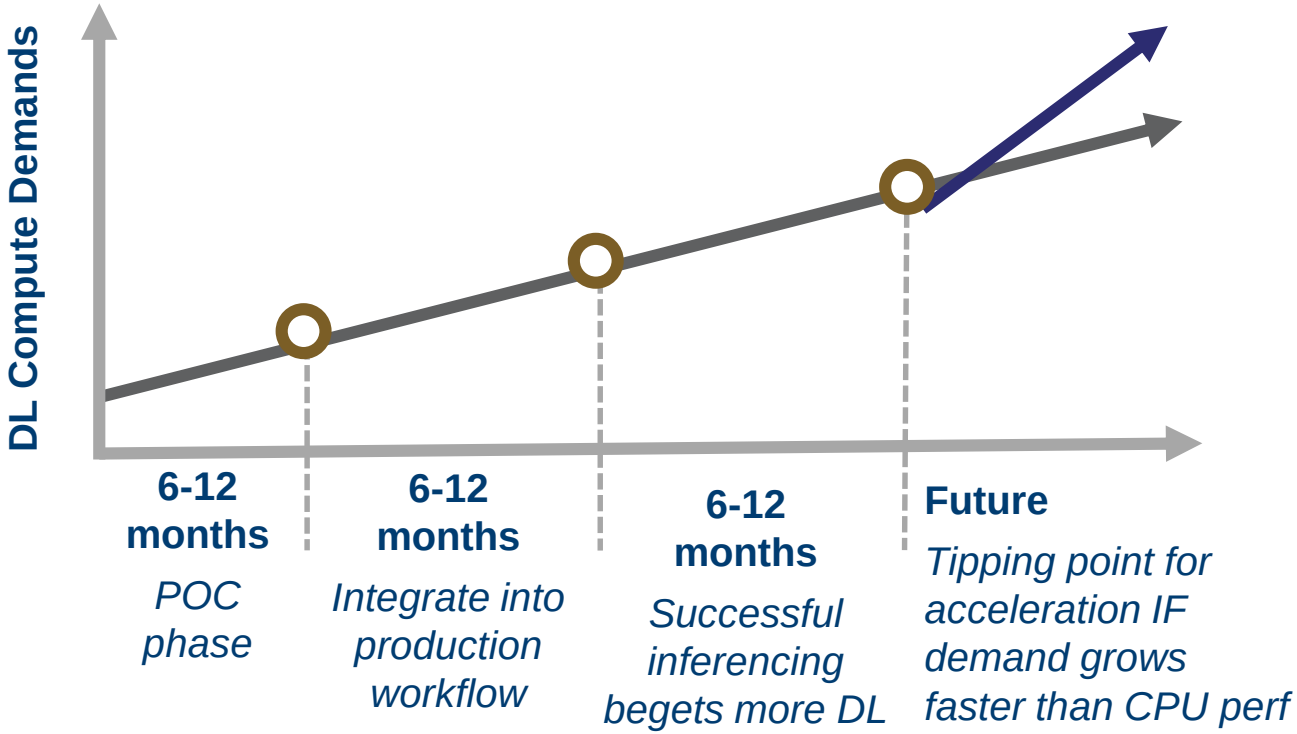
Autonomous driving

Custom Inference

OPEN. FOR BUSINESS.

OCP SUMMIT

Deep Learning Journey



Deep Learning Acceleration Options

	Utilization = TCO	Intermittent training Intermittent or low level of inferencing	Year Round intense training High inferencing compute utilization/node
	Time-to-train	Time-to-train in ~1 day is sufficient with access to additional compute for sale out	Time-to-train in hours is critical for scientist time year round
	Time-to-solution	Fastest by leveraging existing analytics, storage & management infrastructure	Due to high utilization at scale integrating dedicated infrastructure is preferred
	Flexibility& scalability	Resource allocation of shared compute & extend storage pool for unstructured data	No requirement to reallocate compute cycles throughout the year
	Locality	Distributed training and inferencing on edge and premise (eg; factory)	High training and batch inferencing utilization in core datacenter
	Workload	Hybrid of general compute & NN acceleration High definition images exceeding 32GHBM	Mixed precision linear algebra Model layers, weights and data can fit in HBM

Intel® Xeon® Processor Scalable Family

Optimized frameworks and libraries

Up to 113x Gen-to-Gen
DL TRAINING

Up to 138x Gen-to-Gen
DL INFERENCE

FLEXIBLE & GREAT TCO

Scalability across AI & combination workflows

- ❑ Classic ML
- ❑ Deep Learning
- ❑ Reasoning
- ❑ Analytics
- ❑ Emerging AI
- ❑ Many others

DEEP LEARNING PERFORMANCE

Up to **2.4X** performance vs. prior gen via architectural enhancements

- ✓ Up to 28 cores 56 threads / CPU
- ✓ 6 mem channels
- ✓ Mid-level cache up to 1MB per core
- ✓ Intel® AVX-512 with 2 512b FMA units
- ✓ Up to 6.9 FP32 TF & 12 Int8 TOPs

DEEP LEARNING SW OPTIMIZATIONS

Up to 100x performance due to SW optimizations

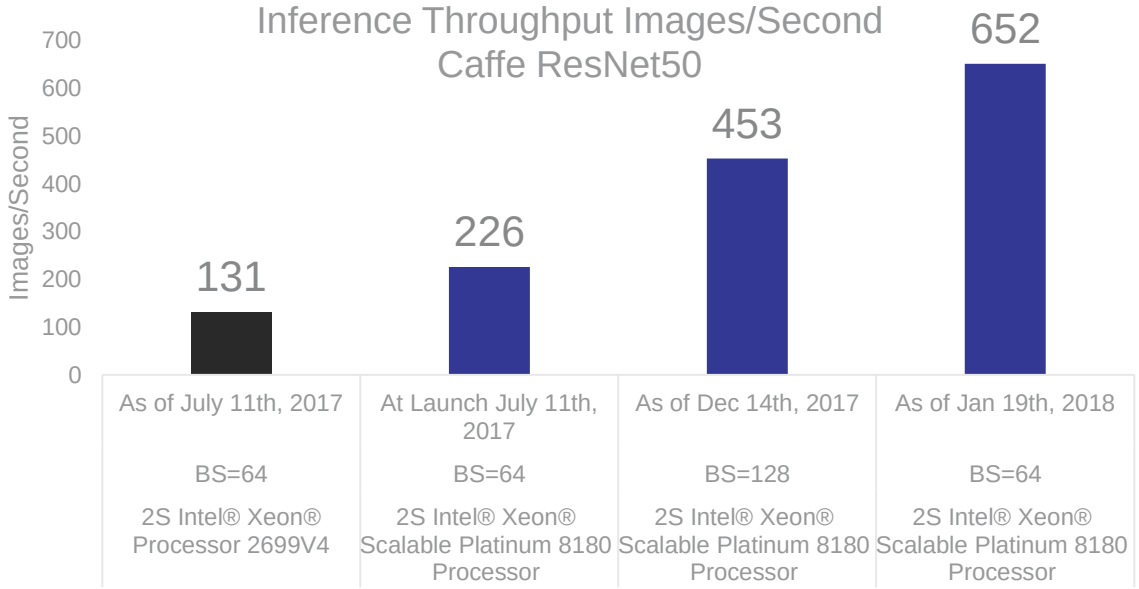
- ➢ Libraries and tools: IA optimized frameworks, Intel MKL-DNN, & DAAL
- ➢ Spark with BigDL, and Python and R with Intel libraries

TensorFlow · Caffe2 · mxnet · BigDL

Xeon SP Performance:

Up to 2.2X improvement with SW optimizations

MEASURED



FUTURE

2019 CPU **+2x**

Move to 8b **+2x**

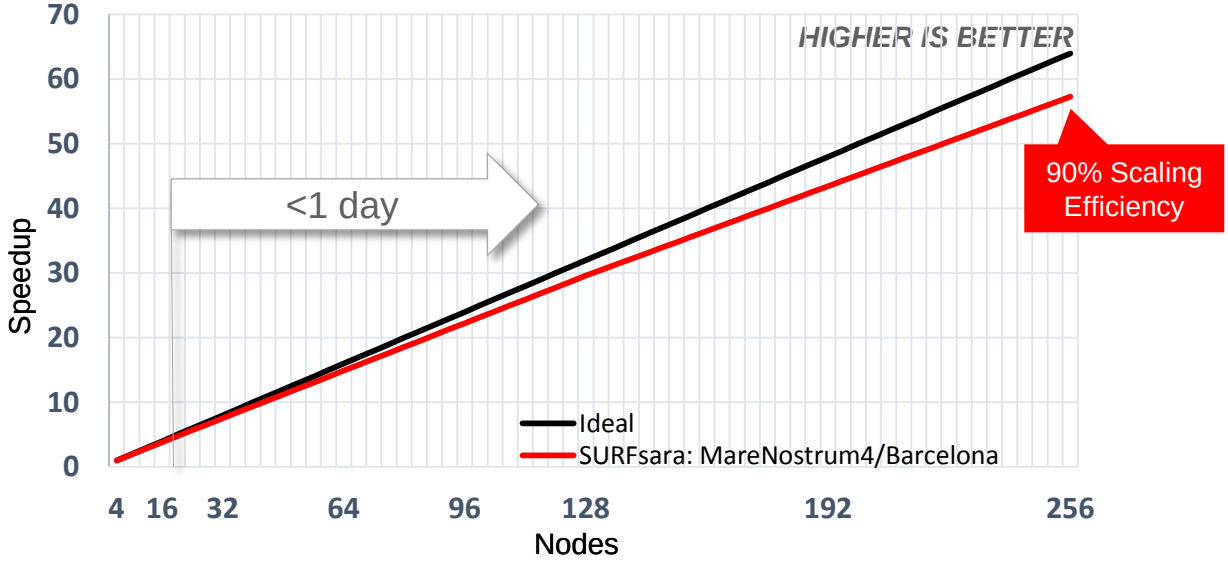
Configuration Details 1, 14, 15

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

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. 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 Source: Intel measured or estimated as of November 2017.

Xeon SP Scaling on CPU

Intel® - SURFsara* Research Collaboration - Multi-Node Intel® Caffe ResNet-50 Scaling Efficiency on 2S Intel® Xeon® Platinum 8160 Processor Cluster



- MareNostrum4 Barcelona Supercomputing Center
- ImageNet-1K
- 256 nodes
- 90% scaling efficiency
- Top-1/Top-5 > 74%/92%
- Batch size of 32 per node
- Global BS=8192
- Throughput: 15170 Images/sec

Time-To-Train: 70 minutes (50 Epochs)

90% scaling efficiency with up to 74% Top-1 accuracy on 256 nodes

Configuration Details 2: Slide 127

Intel® Nervana™ neural network processor (NNP)

PARALLELISM

Massively-parallel compute

Specialized on-die fabrics

Optimized numerics - Flexpoint

SCALABILITY

Large on-die memory

High speed interconnects

Massive inter-chip data transfer

UTILIZATION

Direct SW control for best on-chip memory usage

Managed data-flow paths

ROADMAP

First silicon in 2017

Product roadmap on track to exceed performance goal[1]

¥ Formerly codenamed as the Crest Family

[1]Source: https://newsroom.intel.com/news-releases/intel-ai-day-news-release/?_ga=2.26542141.1088441208.1508441324-198894050.1498491572.

Summary

FINAL THOUGHTS

- Most AI is NOT deep learning
- Most DL is NOT training
- AI value is in convergence

NEXT STEPS

Start your journey with Xeon

- Learn: Focused POCs
- Plan: Develop IT DL technical services and platforms
- Deploy: Integrate with existing IT workflow

Find Out More

LEARN

More information at ai.intel.com

EXPLORE

Use Intel's performance-optimized libraries & frameworks

ENGAGE

Contact your Intel representative for help and POC opportunities

OCP SUMMIT