
intel AI
KEY TRENDS IN DEEP LEARNING
FROM EXPERIMENTATION TO DEPLOYMENT
+ FUTURE WORKLOADS
Matthew Beale, Director of Public Sector Business Development
Intel AI Products Group
CREDIT May 2019

LEGAL NOTICES AND 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 complete information visit intel.com/performance.

Intel does not control or audit the design or implementation of third-party benchmark data or websites referenced in this document. Intel encourages all of its customers to visit the referenced websites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

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 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/benchmarks.

The Shift to Inference at Scale

By 2020, the inference to training ratio will be 10:1 or greater

RESEARCH → FIELD OPERATIONS

1 **Scale.** A model goes nowhere if it can't scale everywhere.

2 **Cost.** Rip and replace is not an effective strategy; get more from what you have.

3 **Environment.** Privacy, security, latency demands, & power constraints.

WHAT THIS MEANS: AI hardware ≠ one size fits all

Different workloads, different considerations

Training

- Key metric: time to train

Inference

- Key metric: ???

Different workloads, different considerations

Training

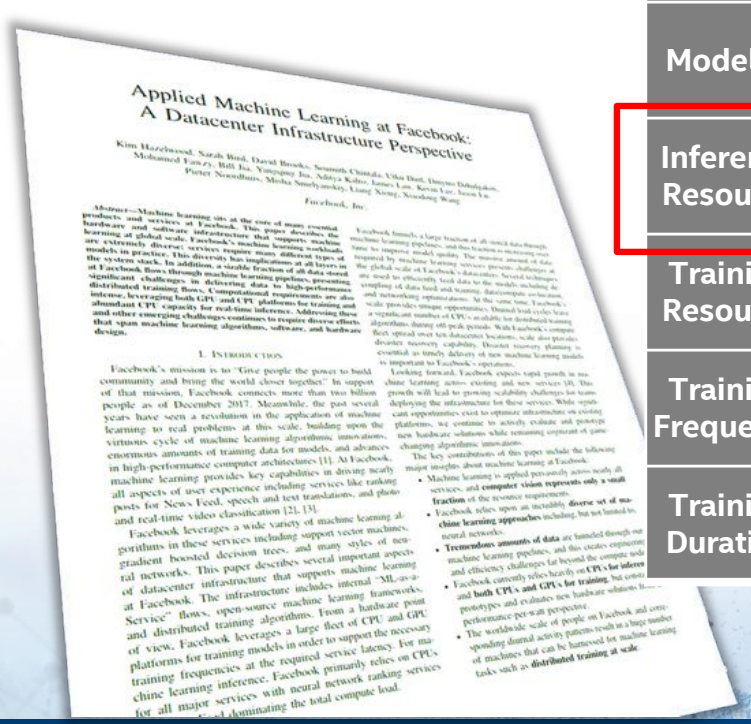
- Key metric: time to train

Inference

- Key metric: ???
  - Throughput
  - Latency
  - SWaP constraints
  - Cost
  - Flexibility to support non-DL workloads
  - Lack of cloud compute access
  - Unified SW stack

Approach #1: Leverage Existing Compute

Large cloud users employ CPU extensively for deep learning


Applied Machine Learning at Facebook:
A Datacenter Infrastructure Perspective

Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmitry Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, James Law, Kevin Lee, Jason Lu, Pieter Noordhuis, Misha Smelyanskiy, Liang Xiong, Xiaodong Wang

Facebook, Inc.

Services	Ranking Algorithm	Photo Tagging	Photo Text Generation	Search	Language Translation	Spam Flagging	Speech
Model(s)	MLP	SVM,CNN	CNN	MLP	RNN	GBDT	RNN
Inference Resource	CPU	CPU	CPU	CPU	CPU	CPU	CPU
Training Resource	CPU	GPU & CPU	GPU	Depends	GPU	CPU	GPU
Training Frequency	Daily	Every N photos	Multi-Monthly	Hourly	Weekly	Sub-Daily	Weekly
Training Duration	Many Hours	Few Seconds	Many Hours	Few Hours	Days	Few Hours	Many Hours

Source Paper: research.fb.com/wpcontent/uploads/2017/12/hpca-2018-facebook.pdf

Intel® AI Customer

IMAGE RECOGNITION

THE NEED

Help radiologists quickly and accurately **read more scans**.

THE CHALLENGE

Cost-effectively deploy deep-learning inference on scanning **machines already in the field**.

THE SOLUTION

Use new software to optimize Intel® Xeon® Scalable processors and **get more out of CPU-based infrastructure**.

THE RESULT

188x acceleration on bone-age model; **37x** on lung model; **value added to 45,000 existing servers**.

"In our PACS systems alone, we have probably 45,000 servers already present and deployed across 800 or 900 different healthcare systems in the United States; so my ability to leverage those processors more effectively is of enormous value.

The patient gets a better outcome, and that leads to **much more efficient healthcare, at a much lower cost.**"

- John Huffman
Chief Scientific Officer, Data Science and AI, Philips

AI for Good and Satellite Imagery

Intel AI is working with a major NGO to develop algorithms for on-demand foundational mapping for disaster response

- Deploy trained models on deep-learning optimized CPU cloud instances for country-scale inference

Approach #2: Leverage Accelerators

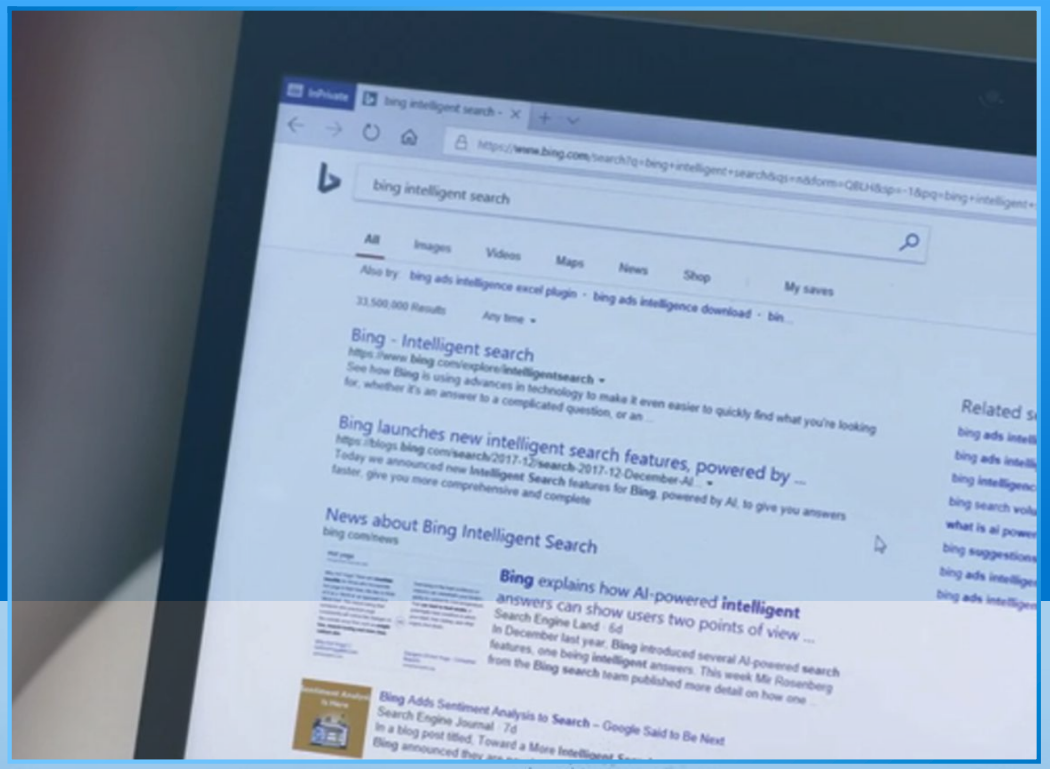
Optimize for:

- Throughput
- Latency
- Power efficiency
- Specific workloads
- …

GPU

Microsoft

Bing



INTEL® ARRIA® AND STRATIX® FPGA

Automated Search with Bing

Detecting poachers with AI at the edge

WHERE DO WE GO FROM HERE?

WHERE HAVE WE SEEN BREAKTHROUGHS, AND WHAT DROVE THEM?

Advances in NLP

- Google and BERT
- OpenAI and GPT-2

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

https://openai.com/blog/better-language-models/#sample2

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)	The incident occurred on the downtown train line, which runs from Covington and Ashland stations. In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief. “The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.” The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials. The Nuclear Regulatory Commission did not immediately release any information. According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

https://openai.com/blog/better-language-models/#sample2

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

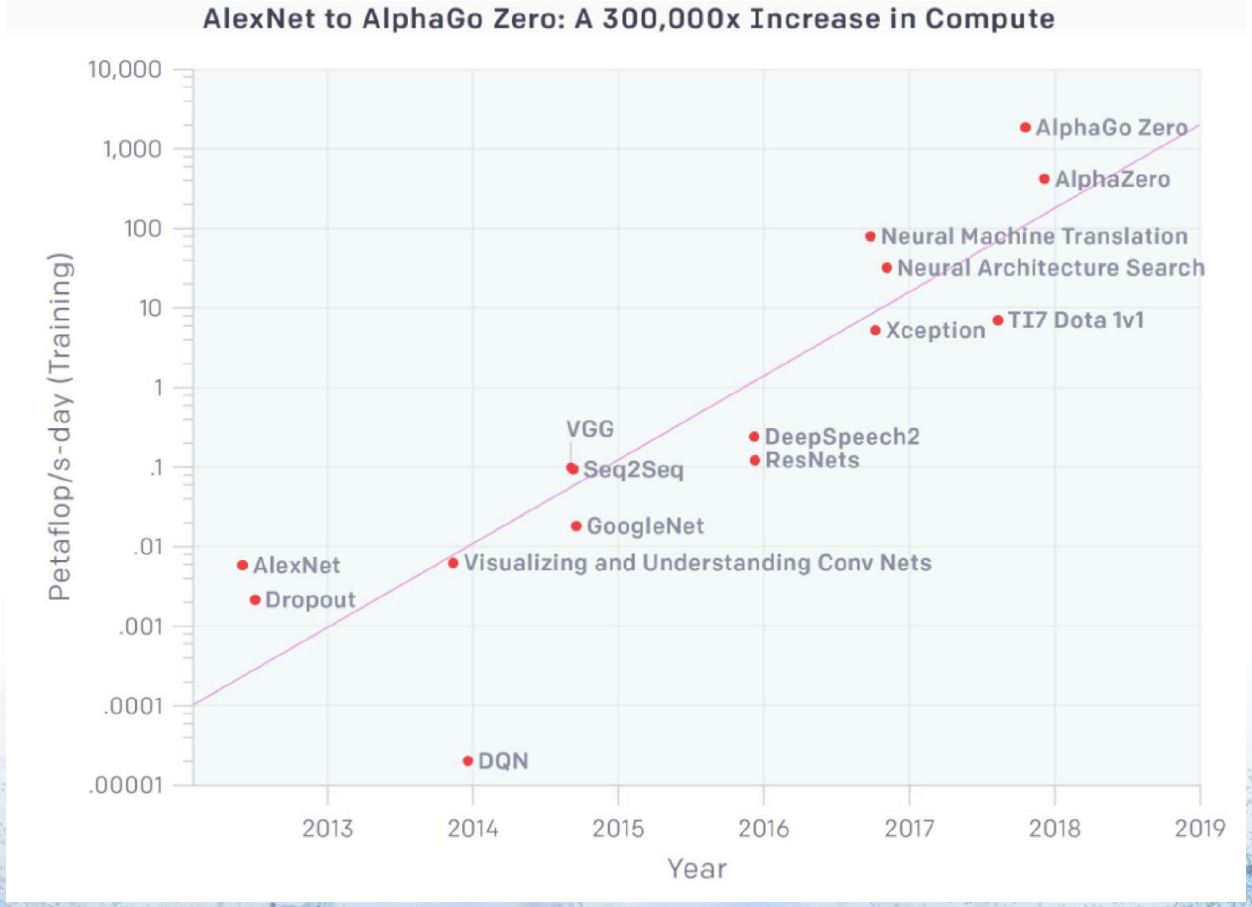
The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Model	# Parameters
Medium-sized LSTM	10M
ELMo	90M
GPT	110M
BERT-Large	320M
GPT-2	1.5B


AlexNet to AlphaGo Zero: A 300,000x Increase in Compute
Petaflop/s-day (Training)
10,000
1,000
100
10
1
.1
.01
.001
.0001
.00001
2013
2014
2015
2016
2017
2018
2019
Year
AlphaGo Zero
AlphaZero
Neural Machine Translation
Neural Architecture Search
Xception
TI7 Dota 1v1
DeepSpeech2
ResNets
VGG
Seq2Seq
GoogleNet
AlexNet
Visualizing and Understanding Conv Nets
Dropout
DQN

Model (Not Just Data) Parallelism

- Data parallelism is the whole model running concurrently (model size limited by single node memory capacity); **model parallelism** is splitting up a model in such a way that each part can be evaluated concurrently, i.e. the order does NOT matter
- Why train large models?
  - Quality improves with model size; understands more about the world
  - Train 1 big model then fine tune it for different things – transfer learning
- Reduces inference latency with computation spread across devices
- **Exploration today:** Language applications (billions of parameters)
- **Challenges:** Very tricky to design the algos

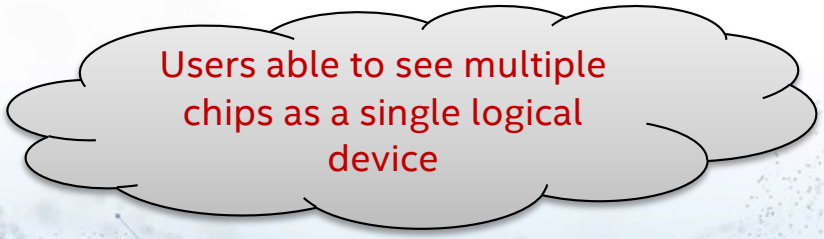
Model And/Or Kernel Parallelism As a Solution

Kernel Parallelism

- Split a kernel (or single/multi-layers) and distribute across multiple chips, HBMs, TPCs
- Ideal for layers requiring very large compute and/or memory resources even with small batch sizes

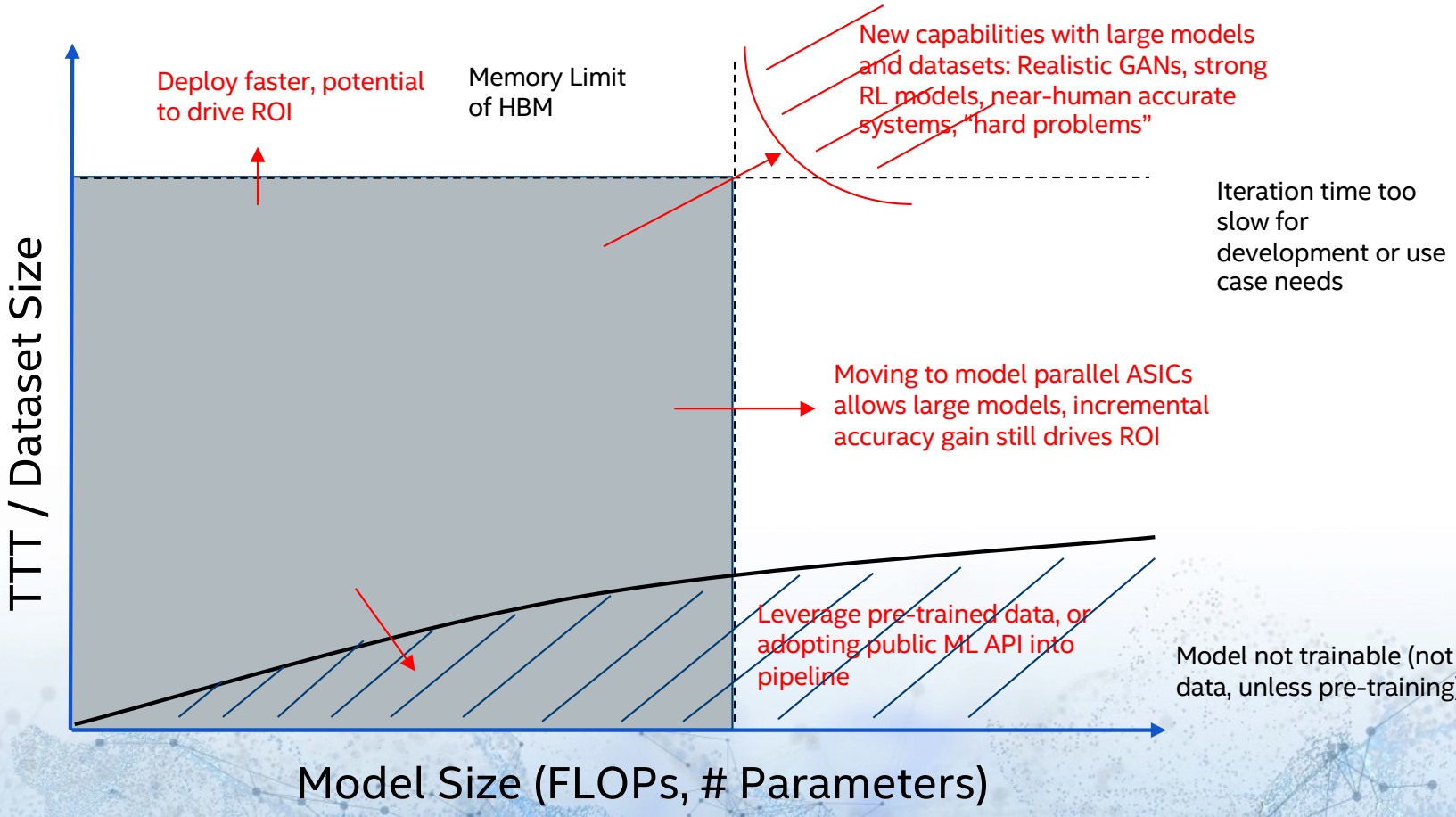
Model Parallelism

- Run multiple different kernels (or layers) across multiple chips/HBMs/TPCs
- Ideal for workloads with several layers that can run in pipelined/concurrent (i.e., for layers without data dependency between them) even with small batch sizes



CPU/GPU -> ASIC -> MULTICHIP

ROI of Moving to New Compute Domains



Matthew Beale
Director of Public Sector Business Development, Intel AI Products Group
matthew.beale@intel.com