

OPEN.



FOR
BUSINESS.



OCP
SUMMIT

OPEN.



FOR
BUSINESS.



OCP
SUMMIT

OPEN.



FOR
BUSINESS.



OCP
SUMMIT

OPEN.



FOR
BUSINESS.



OCP
SUMMIT



WelKom!

Mark Roenigk, Chairman
Open Compute Project

OPEN. FOR BUSINESS





SETTING EXPECTATIONS

Over the next two days, you will:

- Learn from organizations that have adopted Open Source and specifically OCP, even though they were skeptical at first
- See the results of a study we initiated with IHSMarkit, a leading analyst firm, on the current strategies firms are using to deal with data center energy efficiency in the European region, including barriers and best practices, and get to see a case study from the London Stock Exchange
- Experience a Virtual Central Office (VCO) demo
- Review an analysis of Data Center Architectures Supporting OCP by Schneider Electric
- See an Edge Data Center Solution
- Review use cases from the likes of ING Bank, KPN, Yahoo Japan and Nerim
- And much, much more!

OPEN. FOR BUSINESS





GET INVOLVED!

Take advantage of the next 2 days:

1. Tap into the incredible experience in the room in all facets of Open Compute
2. Make the Technical Break Out sessions a priority!

Executive Track: 8 sessions Networking: 8 sessions Telco: 8 sessions

Data Center: 7 sessions Hardware Management & Firmware: 7 sessions

Server & Storage: 7 sessions Rack & Power: 6 sessions

Advanced Cooling: 5 sessions

3. Contribute ideas, products, whitepapers, specifications, case studies

4. Visit with our Open Source partners: OpenPower, OpenCAPI, Linux Foundation Networking and OpenStack

OPEN. FOR BUSINESS





OPEN COMPUTE GLOBAL UPDATE

U.S. Summit is being rebranded as our annual **OCP Global Summit**

- 3,500 attendees, a 17% increase YOY
- OCP is truly global with attendees from 34 countries from around the world
- 81 Sponsors & exhibitors

The **2019 OCP Global Summit** will be **March 14 & 15 in San Jose, CA, USA**. Planning is underway, exhibit locations & sponsorships are more than 65% sold out at this time.

OCP Asia Roadshow is scheduled for 2019 with Engineering Workshops & Technology Days. We are planning now, so if you have an interest in submitting presentation ideas, sponsoring or hosting, please contact the OCP event staff.

We look forward to seeing you here next year!

OPEN. FOR BUSINESS





CURRENT OPEN COMPUTE PROJECTS

This is the core of what we do ... thank you to the community members for your engagement & innovation!



Data Center



High Performance Compute



Hardware Management



Networking



Rack & Power



Server



Storage



Telco

5 Projects in incubation: Open System Firmware, Security, Advanced Cooling Solutions, Archival Storage and Open Edge

OPEN. FOR BUSINESS





INCREASING ADOPTION OF OPEN COMPUTE

This event is the result of increased adoption & global OCP momentum

In March we announced that OCP had reached non-board adoption of **\$1.2 billion**, and is forecast to reach **\$6 billion by 2021**

That study, conducted by independent analyst firm IHSMarkit, broke that forecast down by industry vertical & geography

IHSMarkit specifically identified the need for OCP to have a formal presence in Europe in 2018 based on increasing adoption rates in the region

OPEN. FOR BUSINESS





COMMUNITY UPDATE

Look to the Fall of 2019 for an Asia roadshow of Engineering Workshops & Tech Tracks, we are planning now so if you have an interest in submitting for a talk, or sponsoring please contact the OCP Foundation staff during this event!

Membership continues to broaden to other industries

Contributions are up, there are over 100 products on the OCP Marketplace, with an additional 100 products in the pipeline

Contribution portal is live with over 150 active specifications and growing

OPEN. FOR BUSINESS





UPCOMING REGIONAL OPEN TECHNOLOGY EVENTS

OpenPOWER Summit Europe

Amsterdam RAI (same venue as OCP!)

October 3-4, Amsterdam

TIP Summit (Telecom Infra Project)

London

October 16-17

Use code **OCP30** 30% off a TIP Summit pass.

OPEN. FOR BUSINESS



OPEN.



FOR
BUSINESS.



OCP
SUMMIT



INTRODUCTION TO MOORE'S LAW ON OPEN SOURCE STEROIDS

OR WHY THOSE LITTLE GREEN ARROWS ARE GOING VIRAL

John Laban, Reset Catalyst
Open Compute Project

OPEN. FOR BUSINESS





John Laban

Reset Catalyst

OCP Foundation

john.laban@opencompute.org

[@rumperedis](#)

 +44 7710 124487



OPEN. FOR BUSINESS





A Proven Productivity Technique
POMODORO



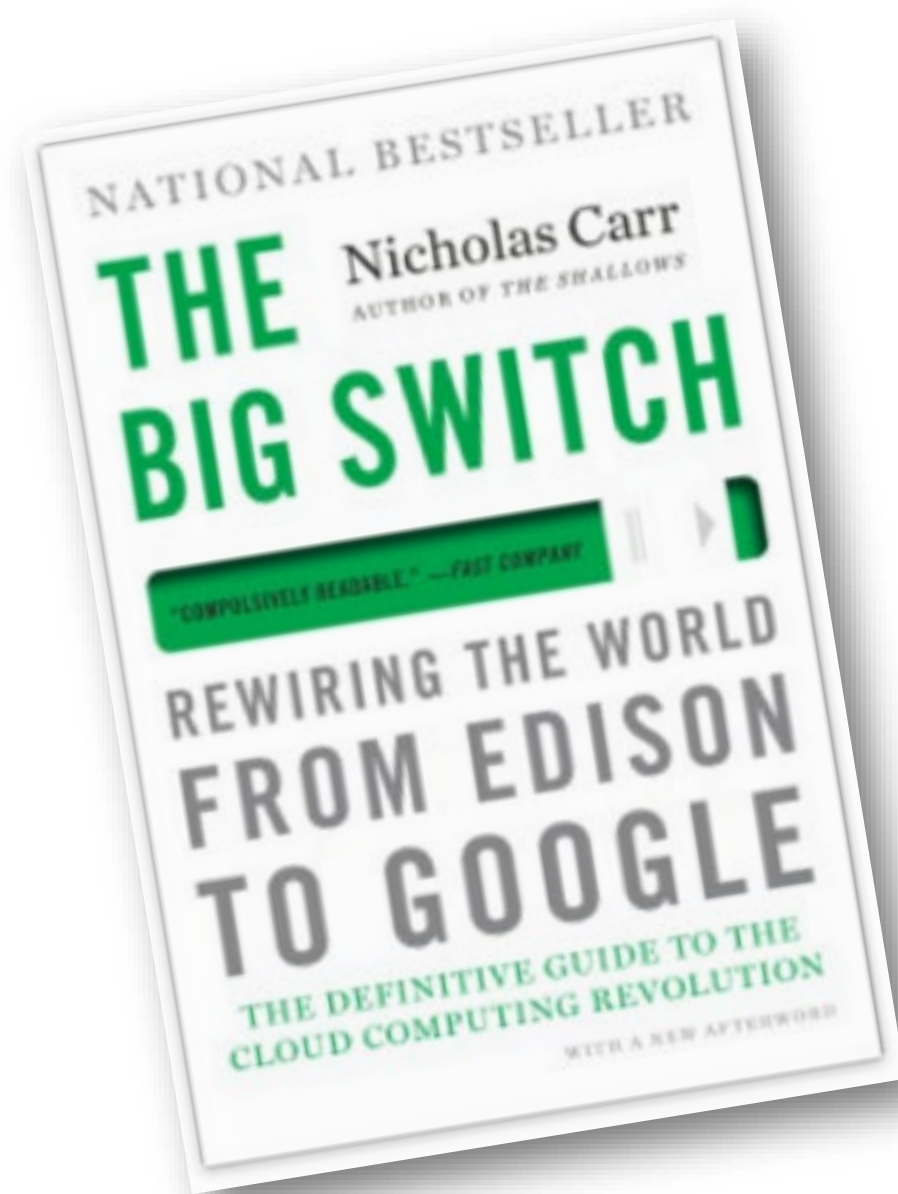
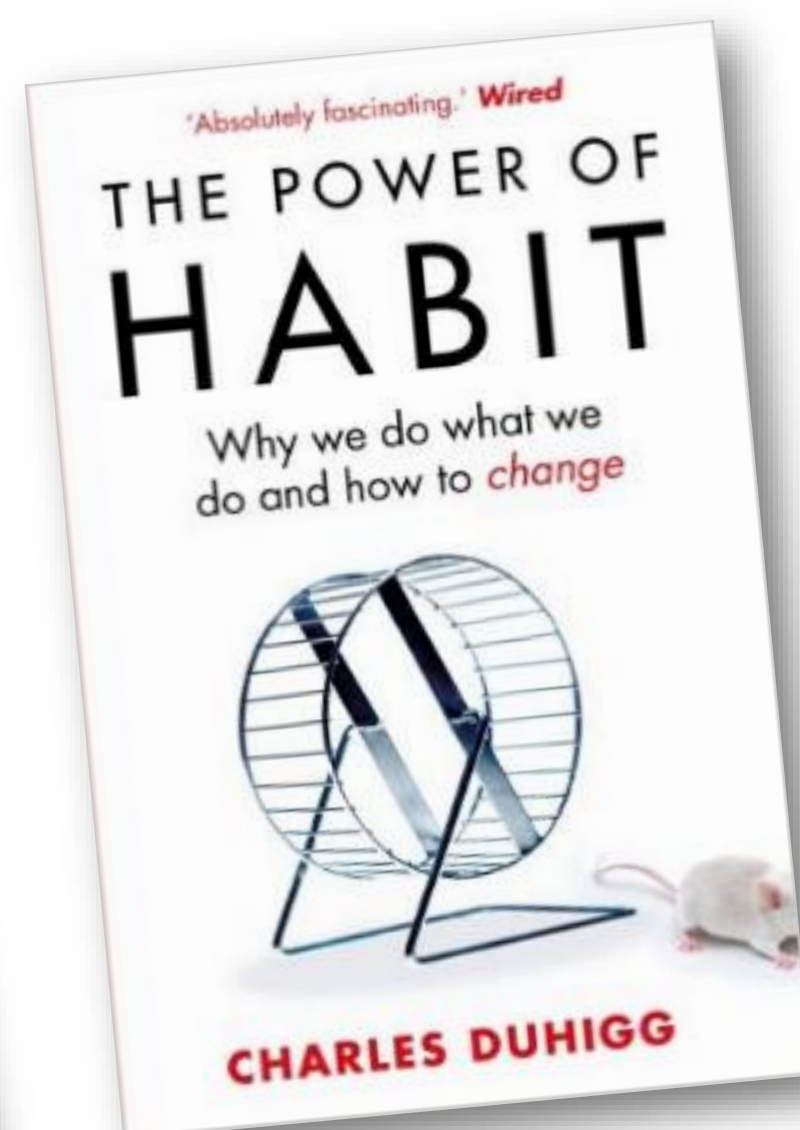
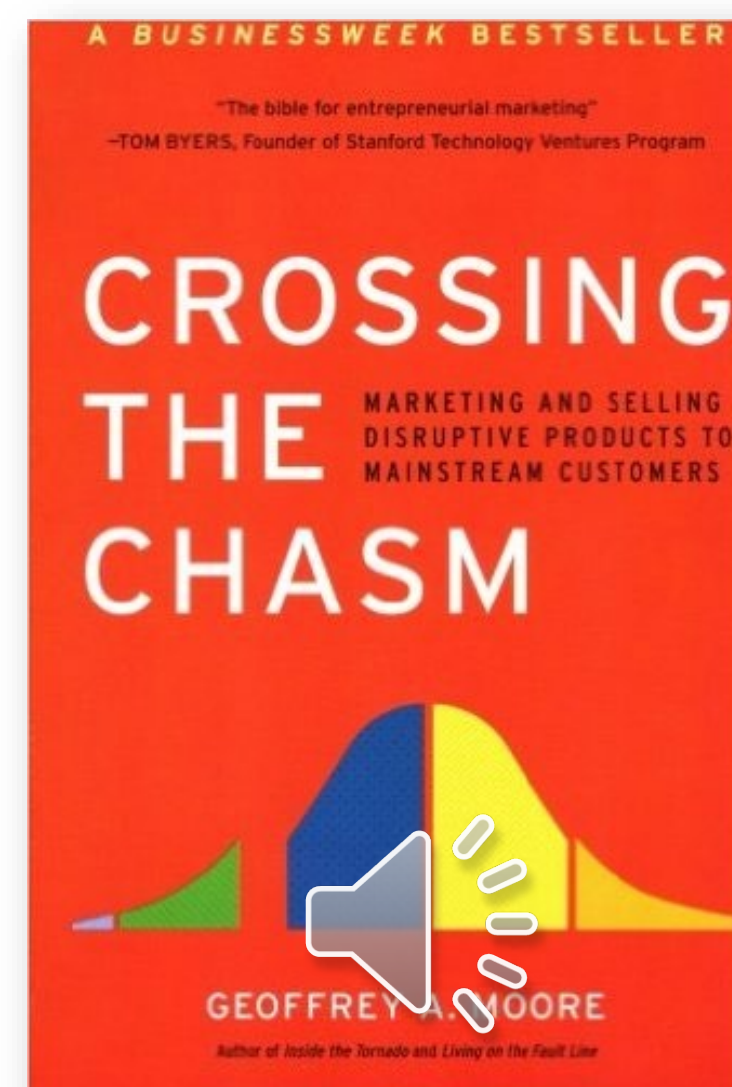
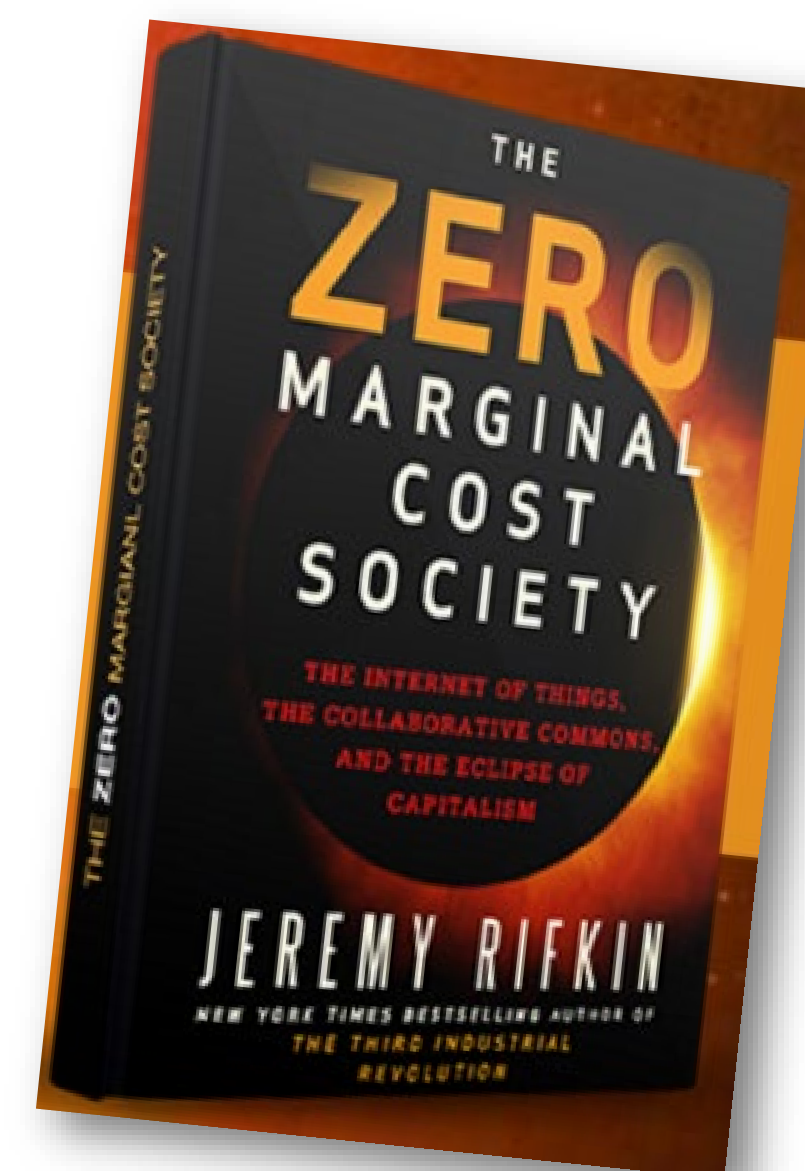
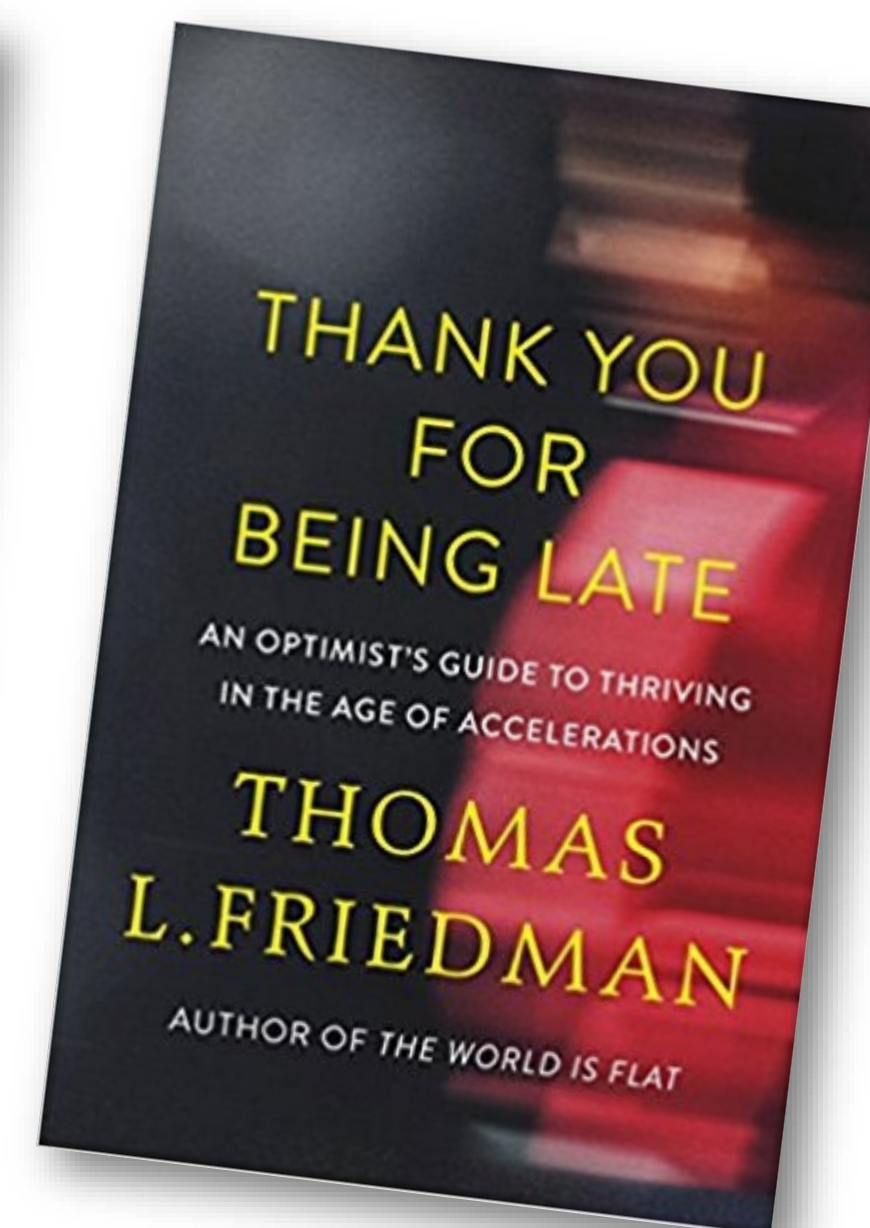
OCP version for open source hackers

OPEN. FOR BUSINESS





RECOMMENDED READING TO HELP RESET INTUITITIVE MINDSETS



OPEN. FOR BUSINESS



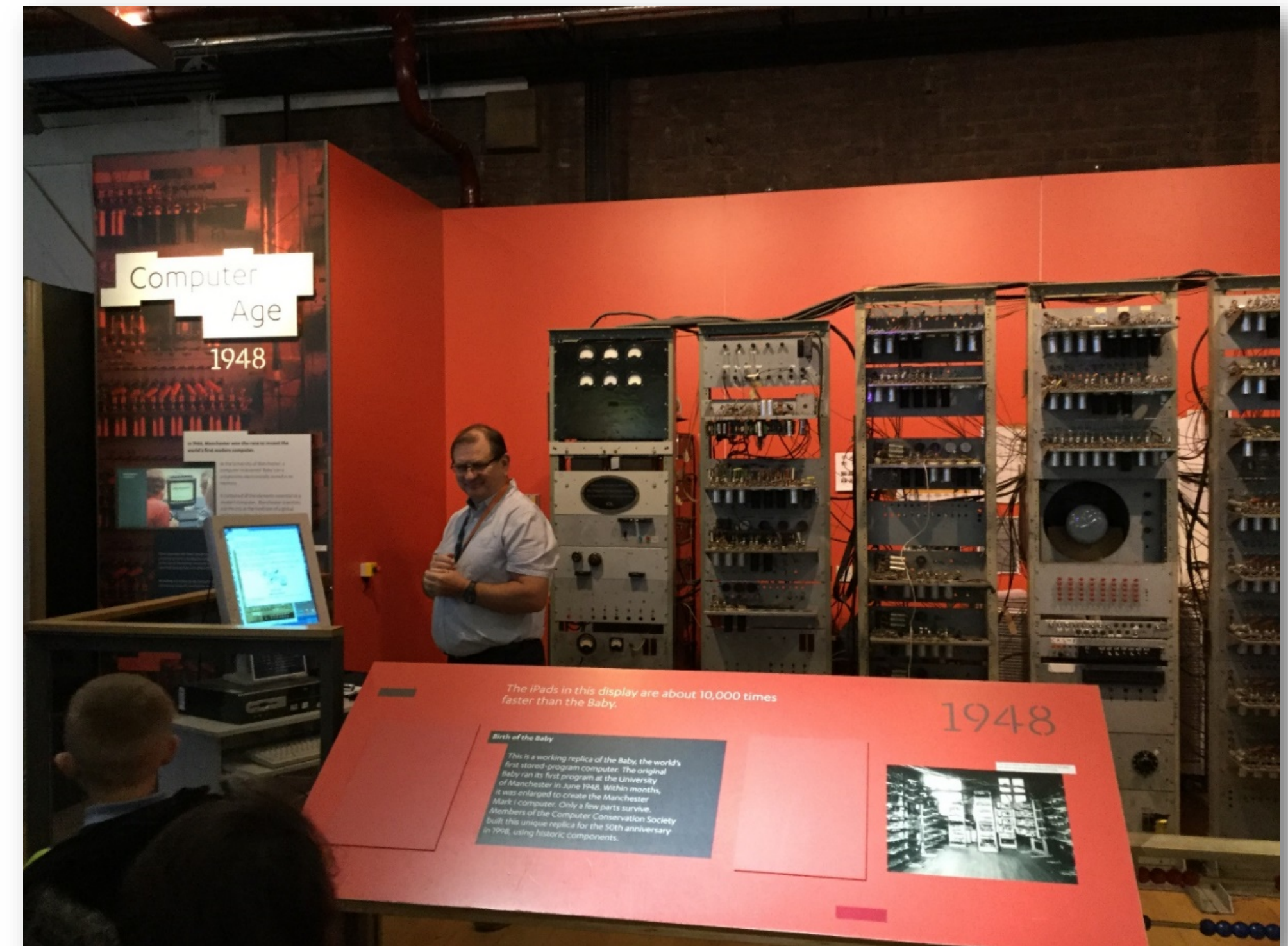


OPENNESS ALWAYS INNOVATES BIGGER IN THE BEGINNING

Manchester University's "The Baby"

The world's first stored programme controlled computer

Celebrates its 70th anniversary in 2018

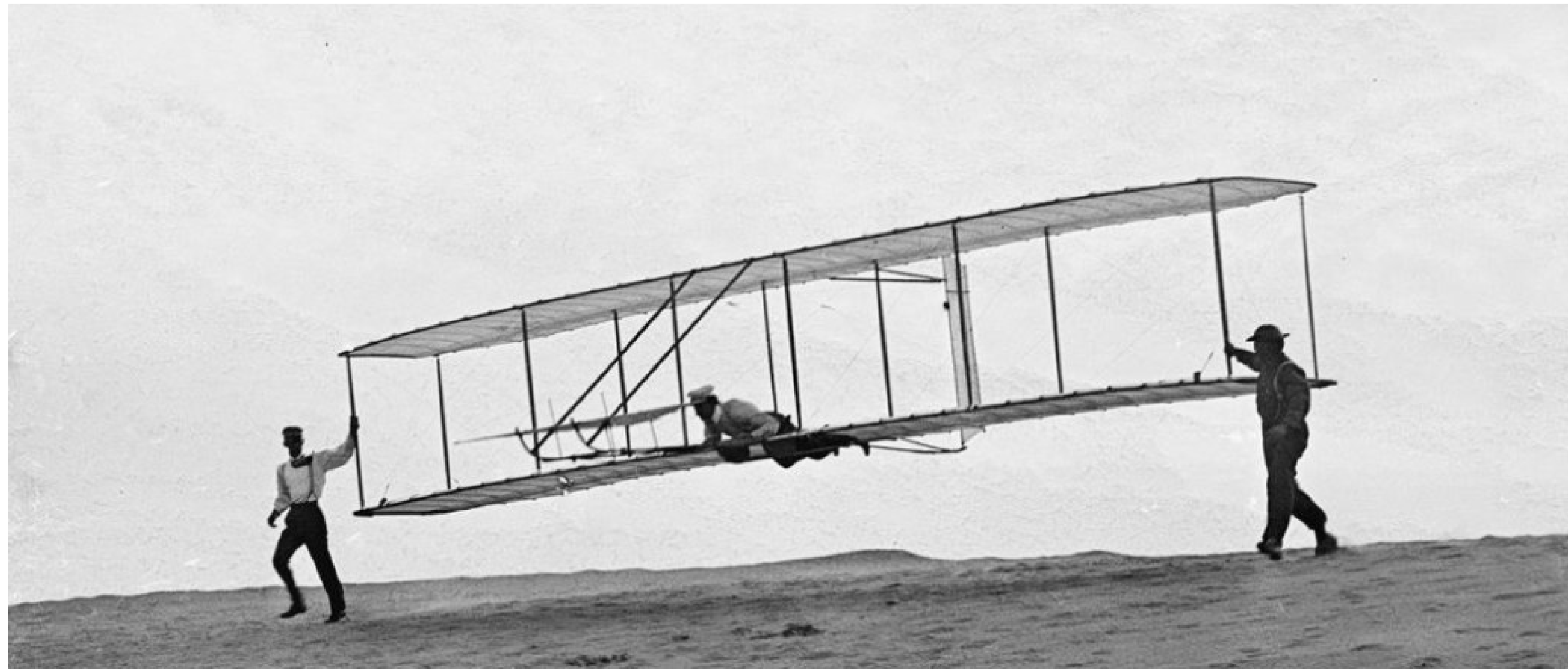


OPEN. FOR BUSINESS





OPENNESS ALWAYS INNOVATES BIGGER IN THE BEGINNING

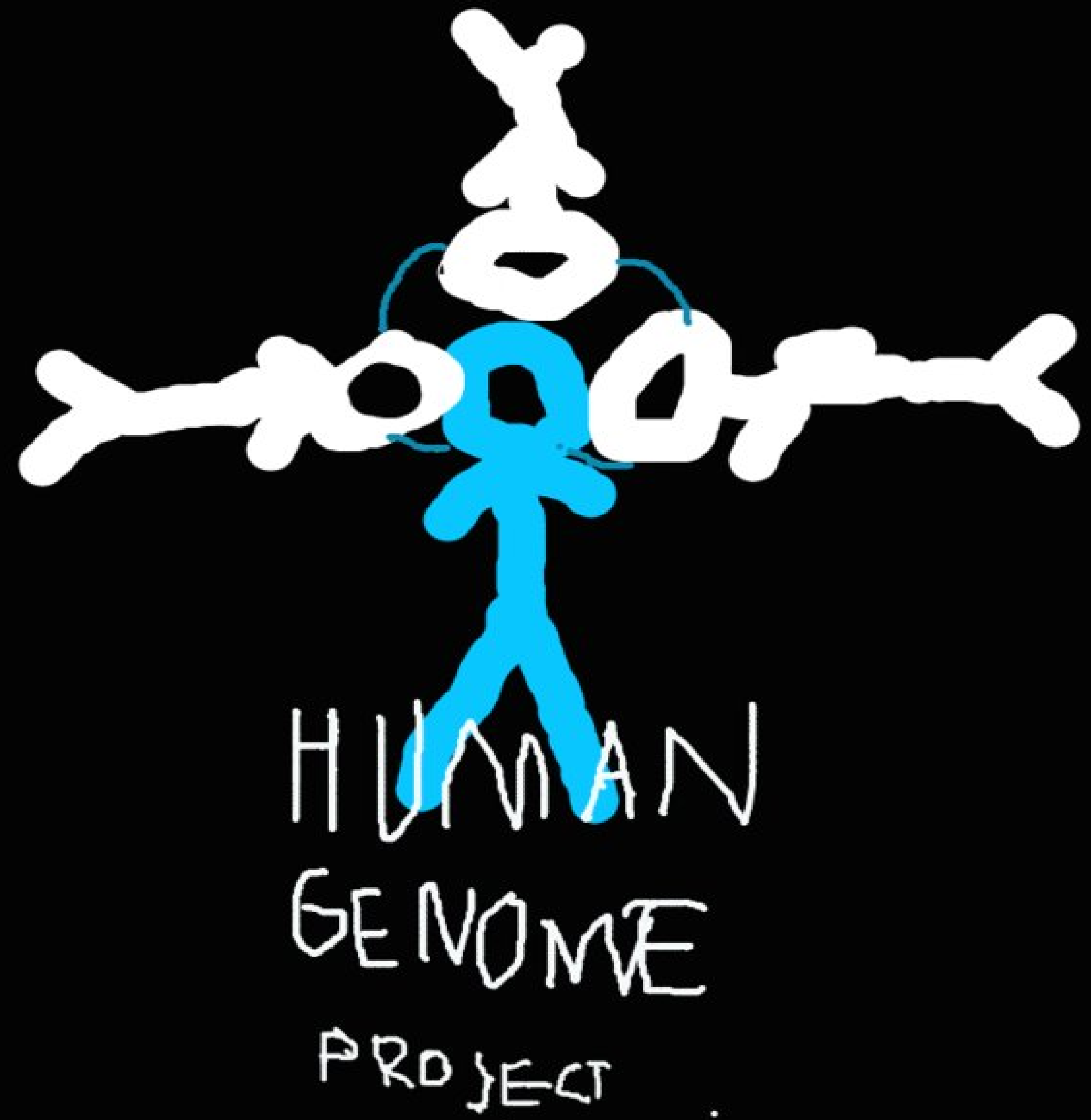


WRIGHT BROTHERS First Flight

OPEN. FOR BUSINESS



SHARING, OPENNESS & COLLABORATION ARE HARD WIRED IN THE HUMAN GENOME





SHARING, OPENNESS & COLLABORATION ARE HARD WIRED IN THE HUMAN GENOME



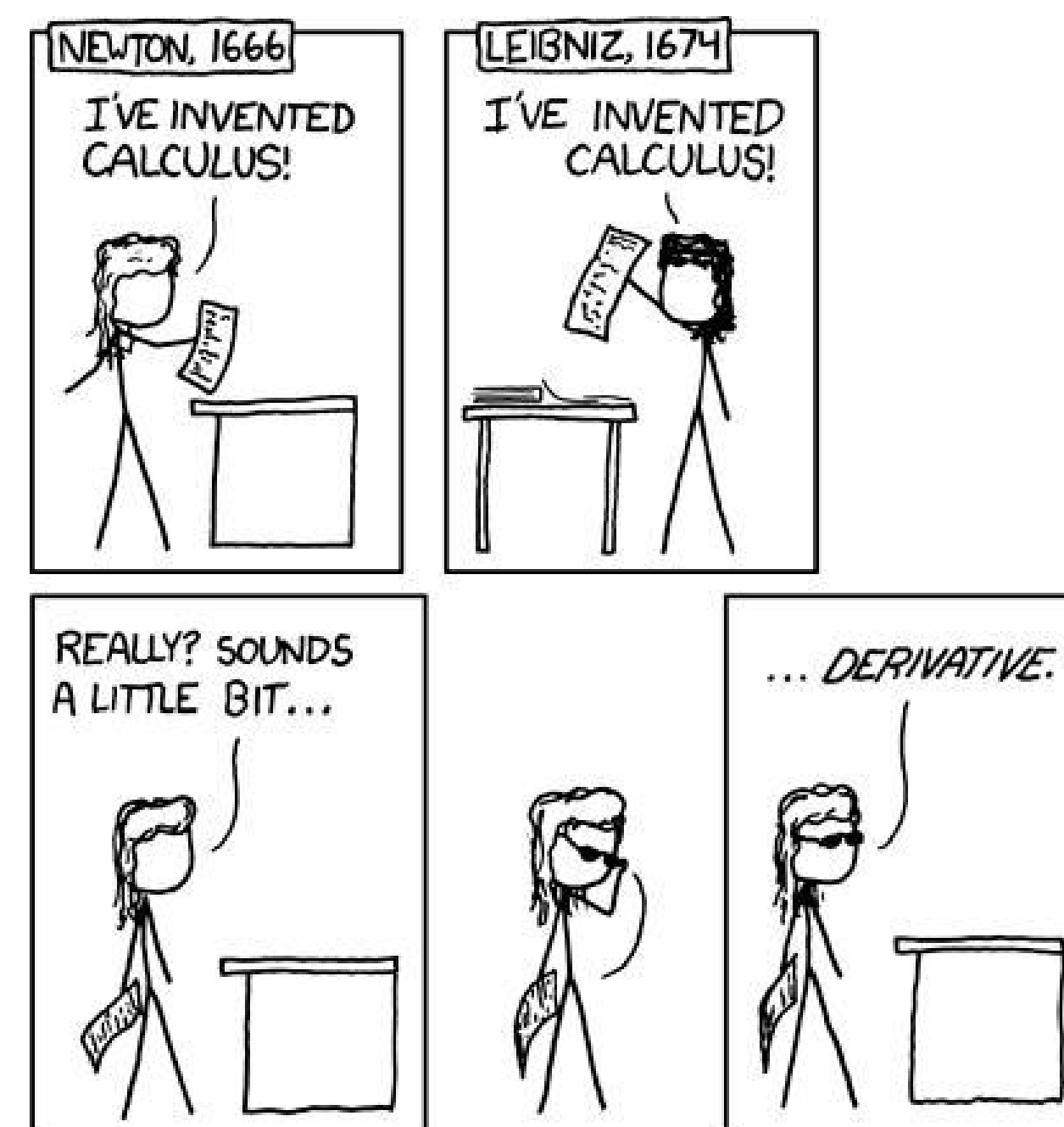
OPEN. FOR BUSINESS



OPEN SCIENCE vs OPEN SOURCE SCIENCE – IT'S ALL ABOUT DISCLOSURE



Newton wrote about Calculus in 1666
He did not publish until 1693



OPEN. FOR BUSINESS



ACADEMIC STUDY OF OPEN INNOVATION

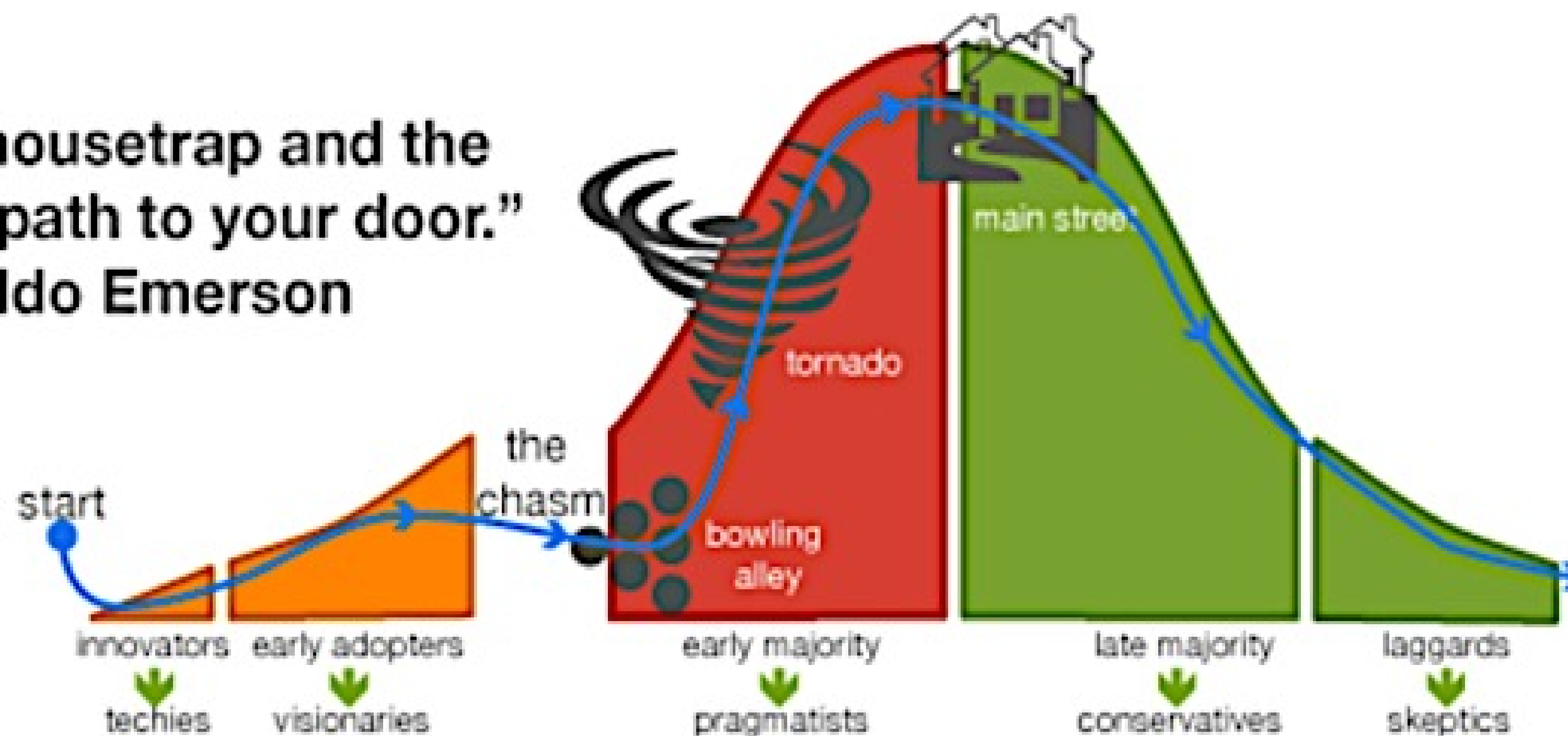


OPEN. FOR BUSINESS



OCP GEAR HITS THE TORNADO PHASE OF ADOPTION

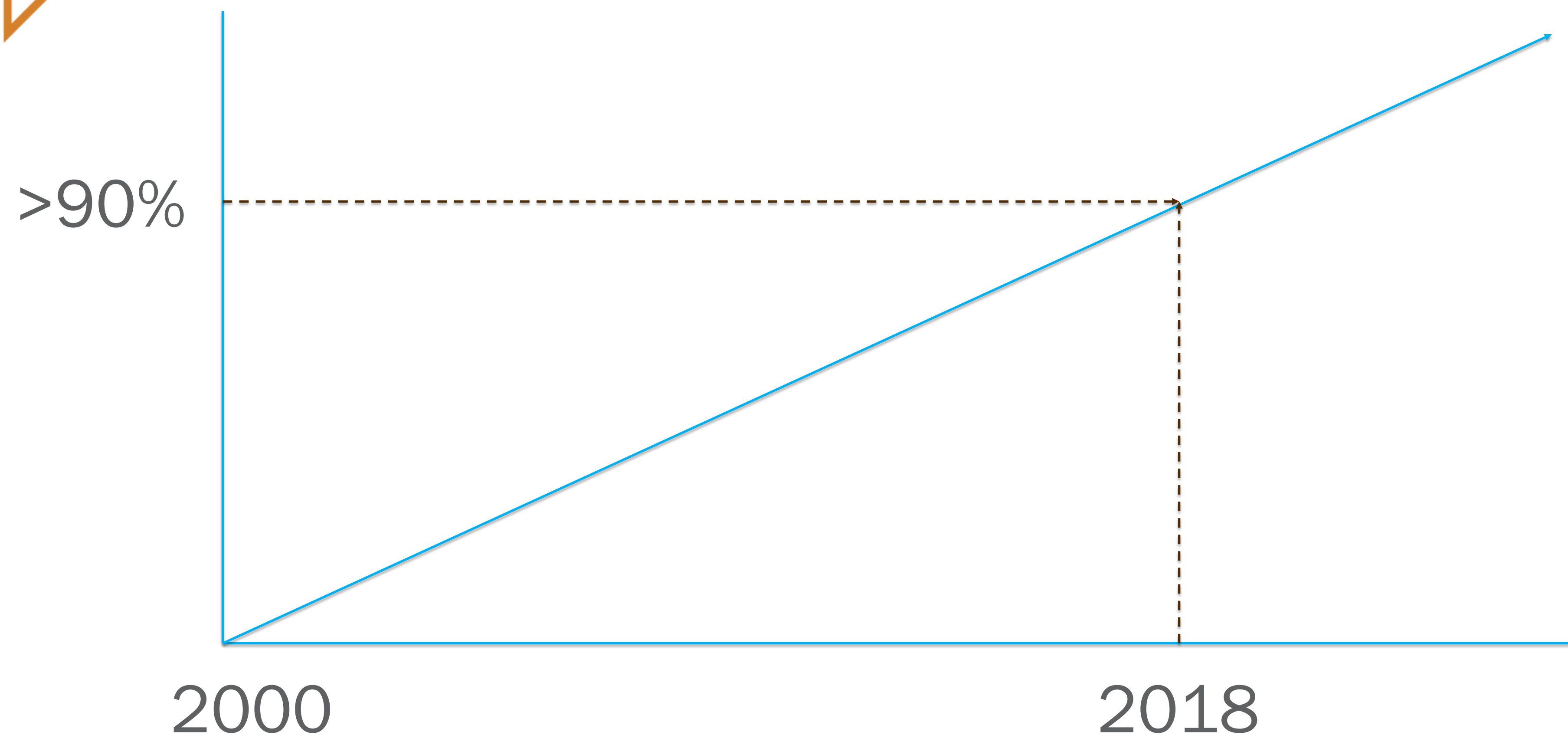
**“Build a better mousetrap and the world will beat a path to your door.”
- Ralph Waldo Emerson**



OPEN. FOR BUSINESS



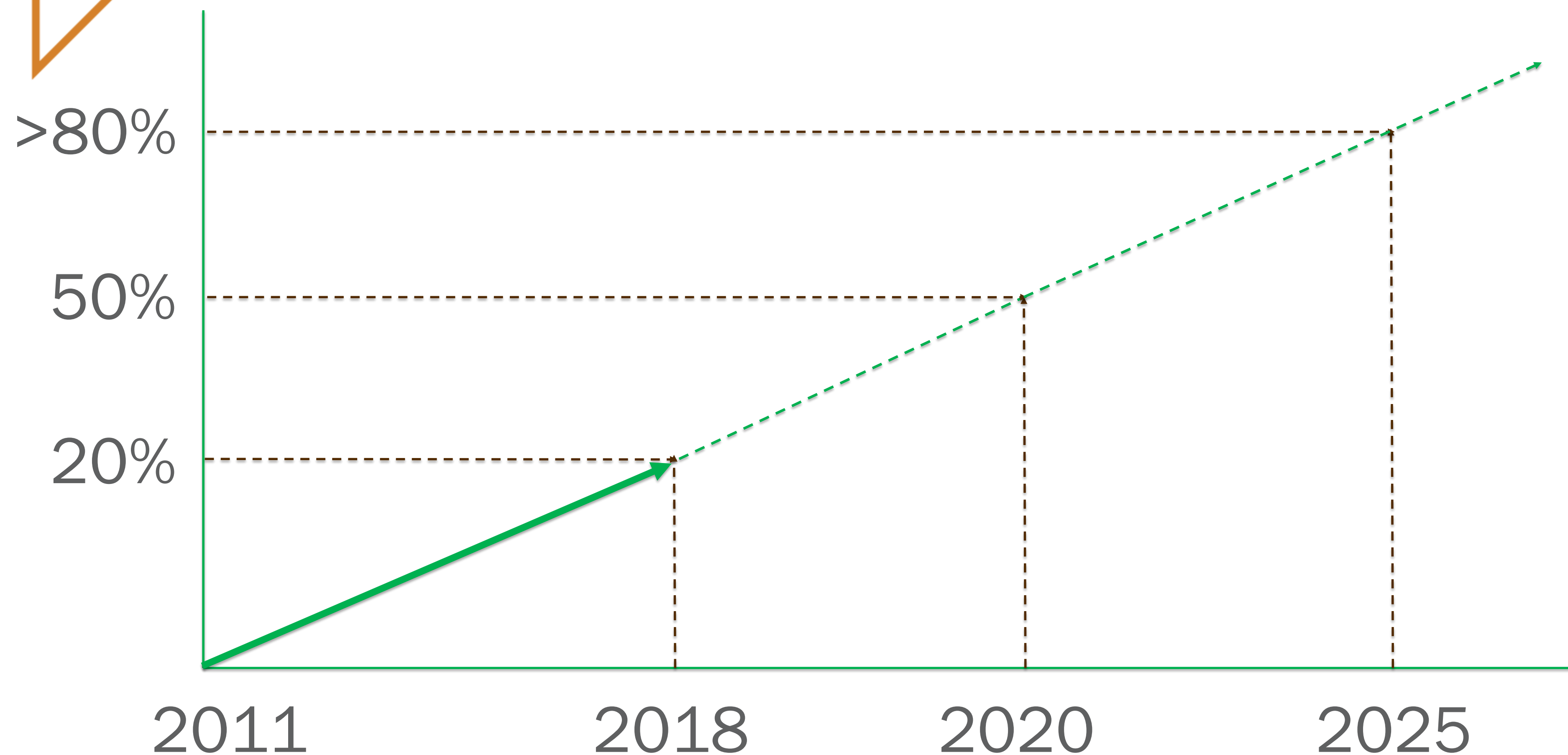
LINUX FOUNDATION & DATA CENTRE OPEN SOURCE SOFTWARE ADOPTION



OPEN. FOR BUSINESS²⁶



OCP FOUNDATION & DATA CENTRE HARDWARE ADOPTION



OPEN. FOR BUSINESS



BUILDING A BETTER MOUSE TRAP



OCP Modular Data Centre

66% reduction in CAPX

*Compared to traditional
20ft containers*

OPEN. FOR BUSINESS





OPEN COLLABORATION

**Better
Products
Faster**

“Yesterday, there was a wall of Tesla patents in the lobby of our Palo Alto headquarters. That is no longer the case. They have been removed, in the spirit of the open source movement, for the advancement of electric vehicle technology.”

Elon Musk, 2014

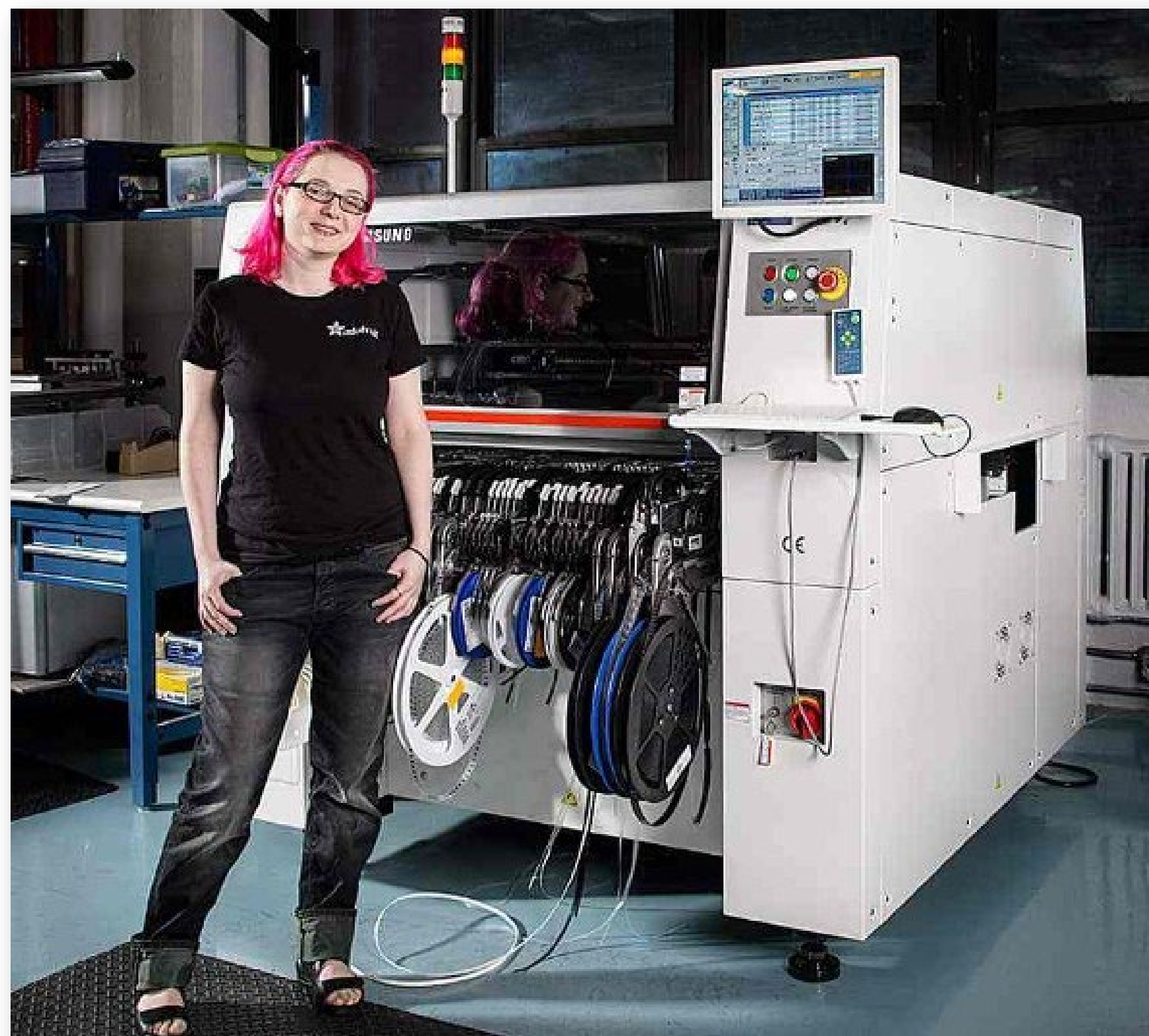
OPEN. FOR BUSINESS





FASTEST GROWING MANUFACTURER IN THE USA

Adafruit grew
700% each
year for the
last 3 years



OPEN. FOR BUSINESS





OPENNESS ALWAYS WINS IN THE END

Linux Dominates Cloud, Embedded & HPC

Market Segment	Linux Dominance
Supercomputer Market Share	99%
IBM's Top Mainframe Customers	90%
Public Cloud Workload	90%
Embedded Systems Market	62%

Every market Linux has entered it eventually dominates

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT NORTH AMERICA

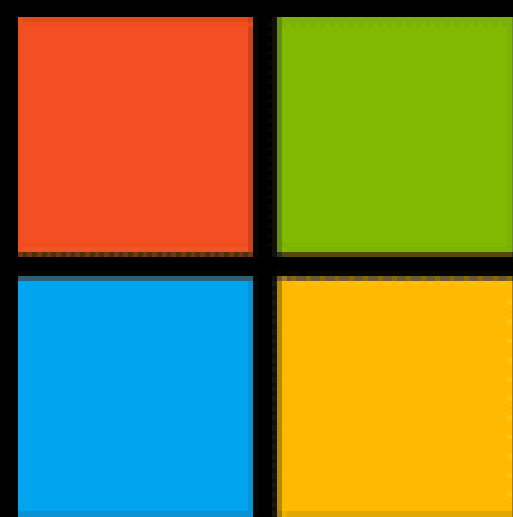
1:55 / 26:55

OPEN. FOR BUSINESS





OPEN SOURCE – A GAME WON STORY



Microsoft



Buys



GitHub

OPEN. FOR BUSINESS



OCP
SUMMIT



OCP SERVERS ARE STREETS AHEAD OF EU ENERGY LEGISLATION

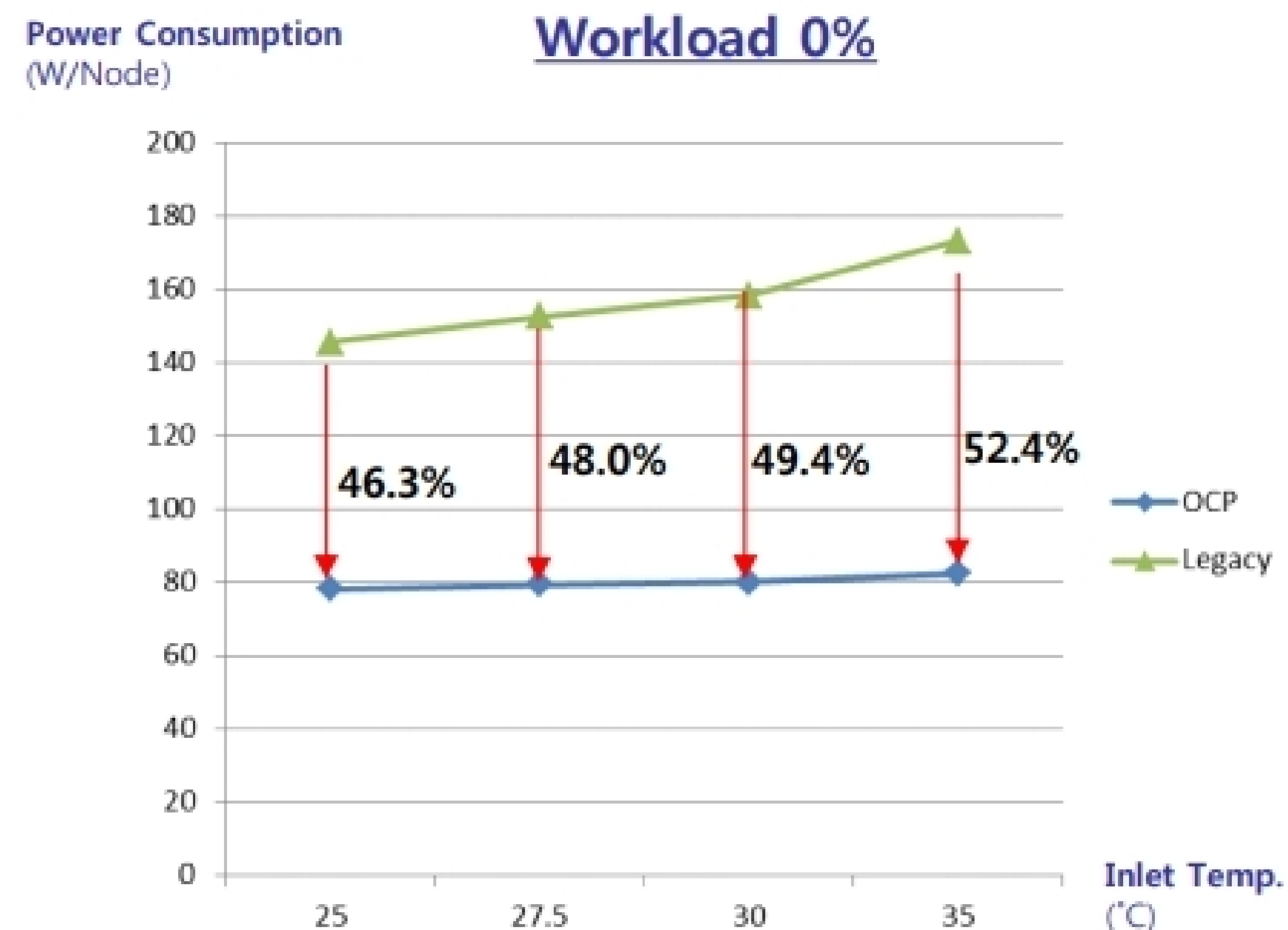
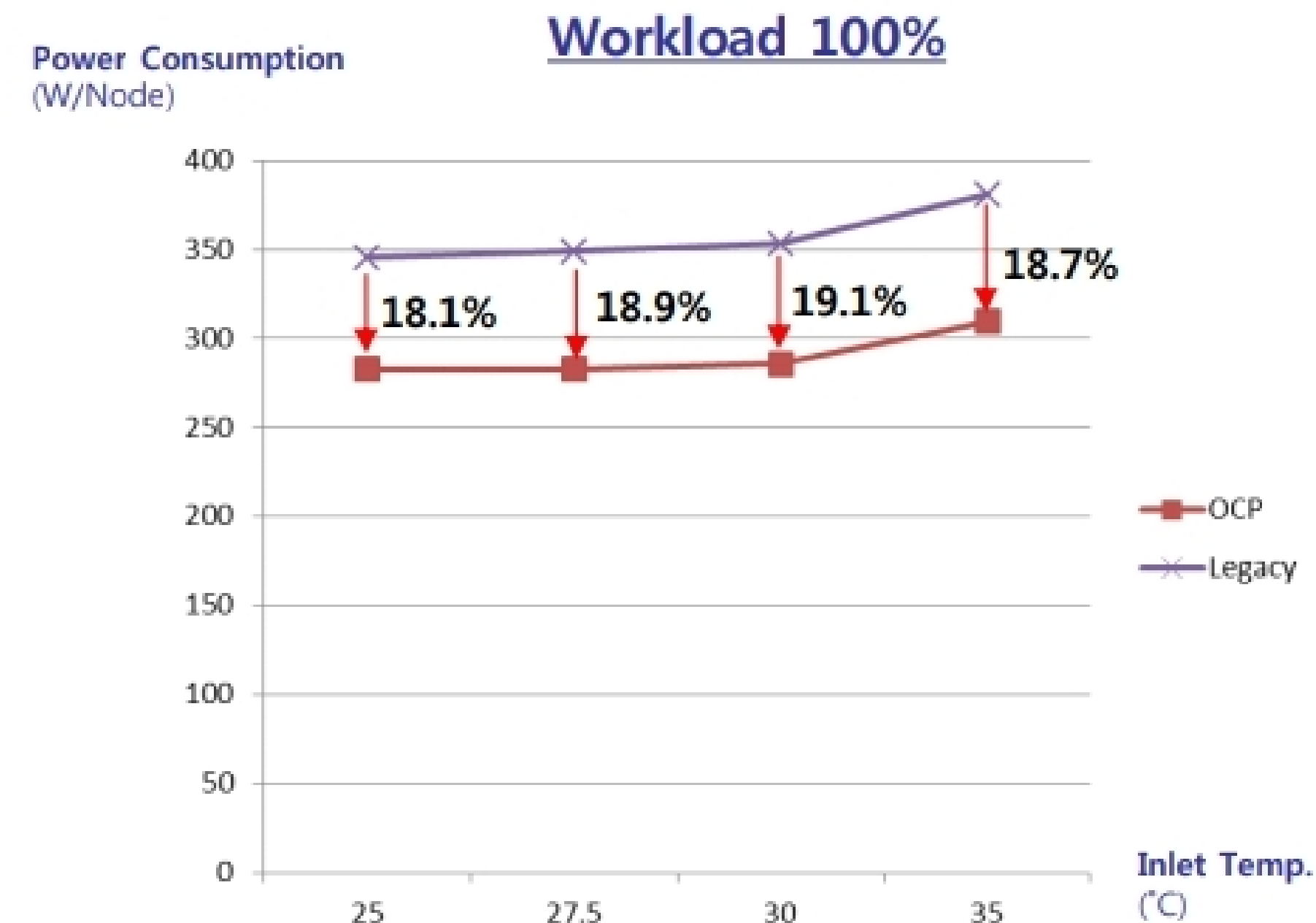


OPEN. FOR BUSINESS

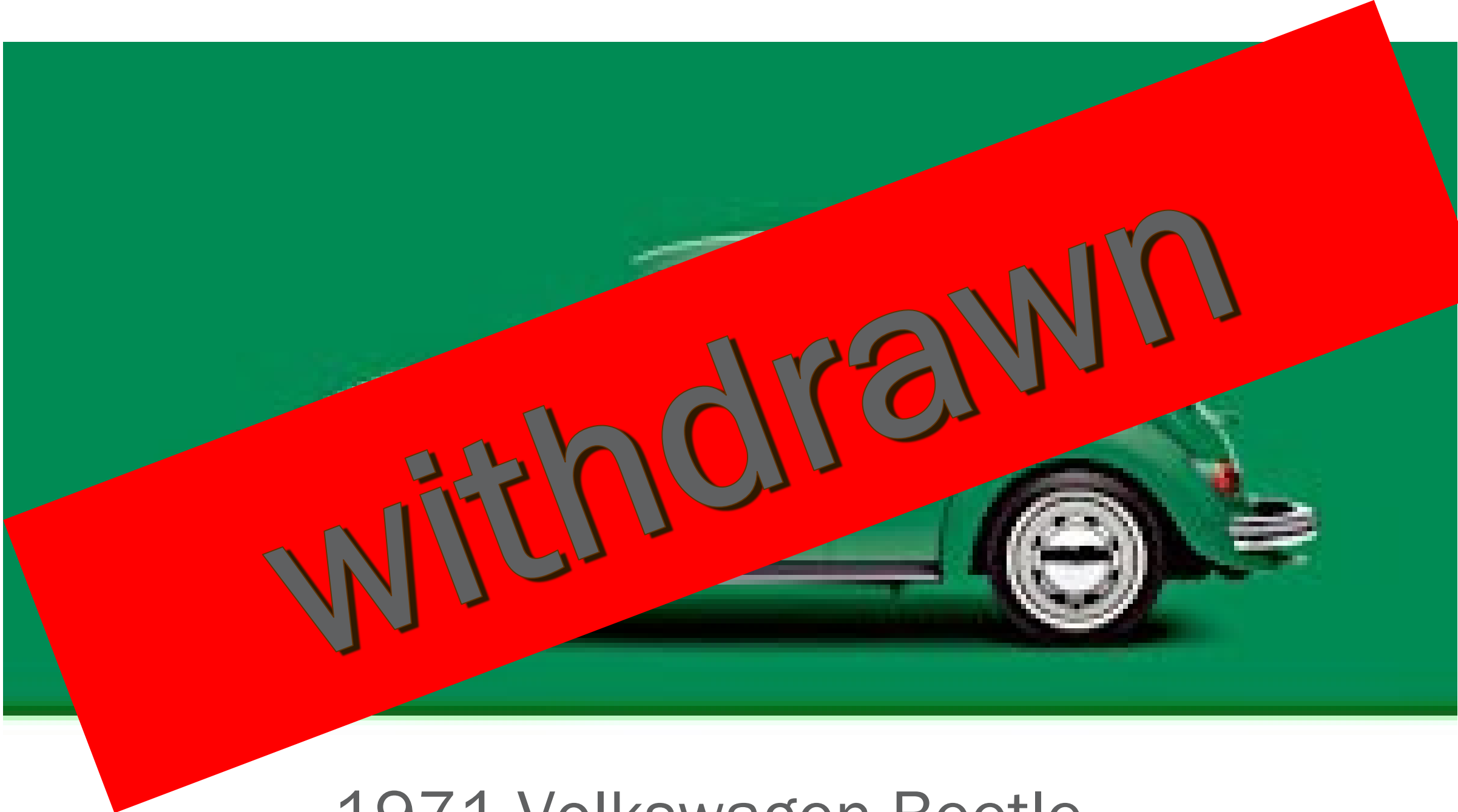
Power Consumption Comparison



- In idle state, OCP server bettered legacy server under all the room temperatures
- In extreme workload, power consumption difference drops slightly



Note : CERN OCP energy study showed a 29% reduction in energy at 80% server utilisation



withdrawn

1971 Volkswagen Beetle

In 2015, in the 50th anniversary year of Moore's Law, Intel engineers did a rough calculation of what would happen had a 1971 Volkswagen Beetle improved at the same rate as microchips did under Moore's law.

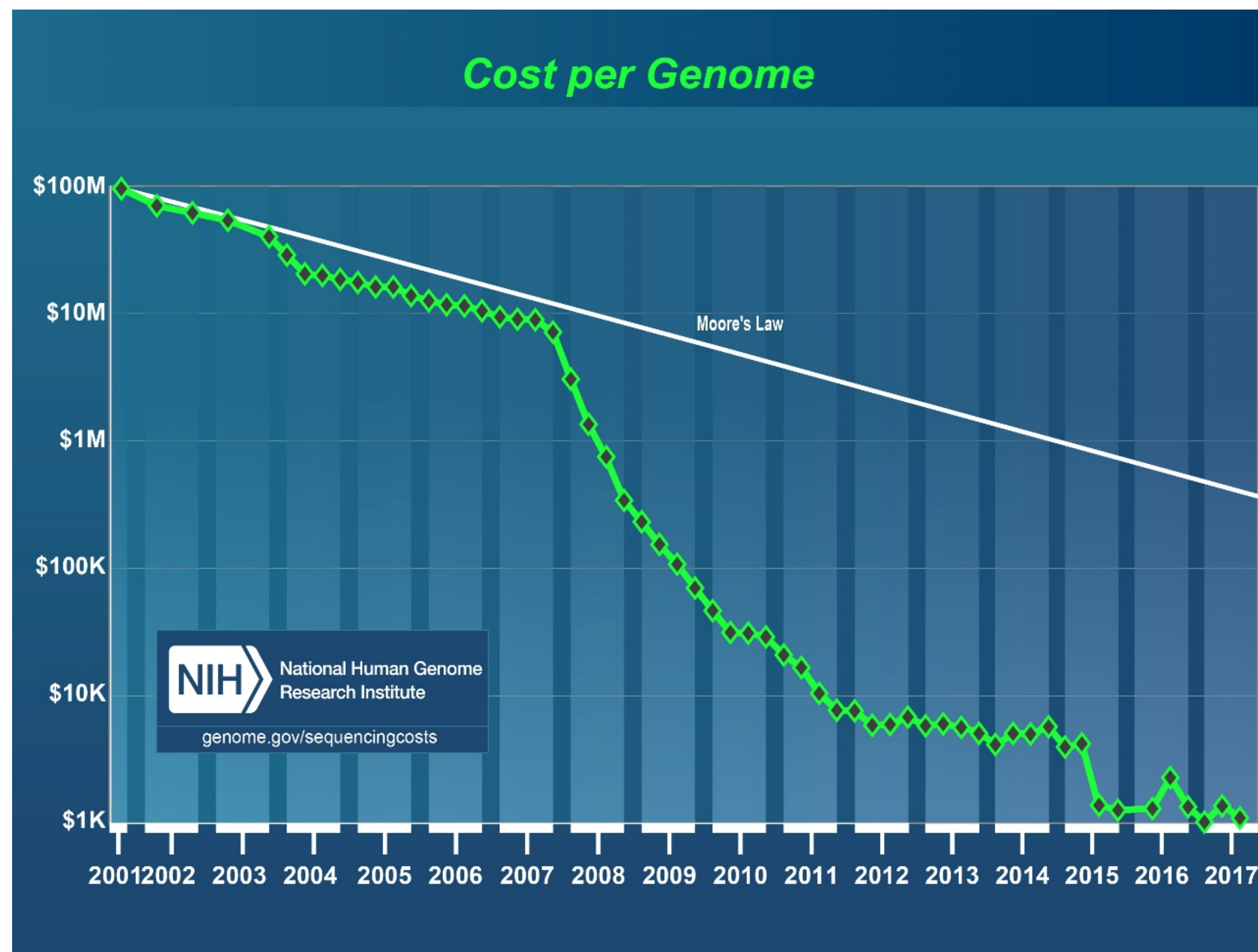
These are the numbers: Today, that Beetle would be able to go about 300,000 miles per hour. It would get two million miles per gallon of gas, and it would cost four cents!

OPEN. FOR BUSINESS





GENOMICS = MOORE'S LAW ON STEROIDS



OPEN. FOR BUSINESS



Genomics Moore's Law on Open Source Steroids.

Three times the rate of improvement compared to microchips.



1971 Volkswagen Beetle

OPEN. FOR BUSINESS

IMAGINE WHAT WOULD HAPPEN IN A WORLD WITHOUT OPEN SOURCE?

“For starters, the Internet and the Web would instantly evaporate. Every Android smartphone, every iPad, iPhone and Mac would go dark. A massive section of our energy infrastructure would cease to function. The global stock markets would go offline for weeks, if not longer. Planes would drop out of the sky. It would be an event on the scale of a world war or a pandemic.”

Steven Johnson



October 1-2
2018
Amsterdam

OPEN. FOR BUSINESS



EUROPEAN SHARING EXPERIMENT



Frenchman

Dutchman

Swede

Brit

OPEN. FOR BUSINESS

OPEN.



FOR
BUSINESS.



OCP
SUMMIT

OPEN.



FOR
BUSINESS.



OCP
SUMMIT

OPEN.



FOR
BUSINESS.



OCP
SUMMIT



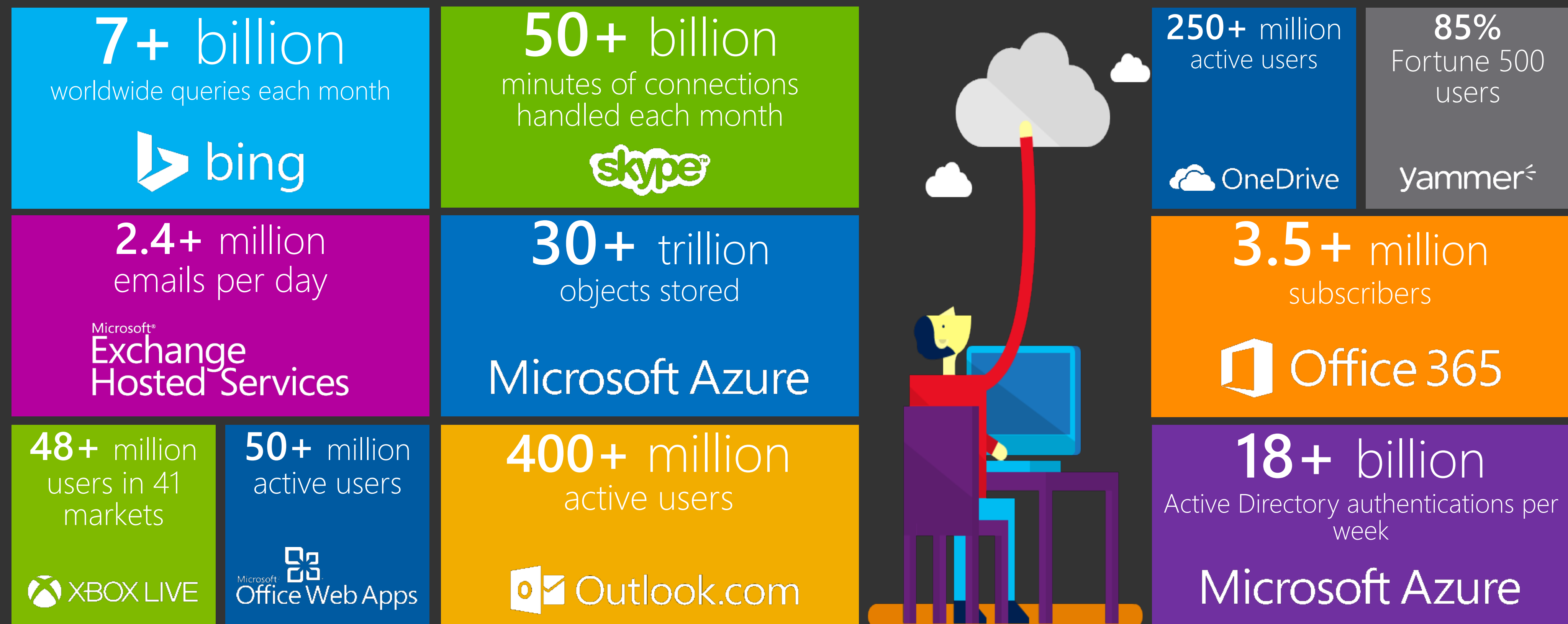
Cloud and AI

Marc Tremblay, Ph.D.
Distinguished Engineer
Microsoft

OPEN. FOR BUSINESS



Azure Cloud



Microsoft Cloud Services

200+ Services · 1+ billion customers · 20+ million businesses · 90+ markets worldwide

3 Millions

Km

intra-datacenter fiber

54

Azure regions

140 countries

100+

datacenters

Millions

of servers

\$B's annual infrastructure investment

Dark fiber network

24 x 7 x 365 support

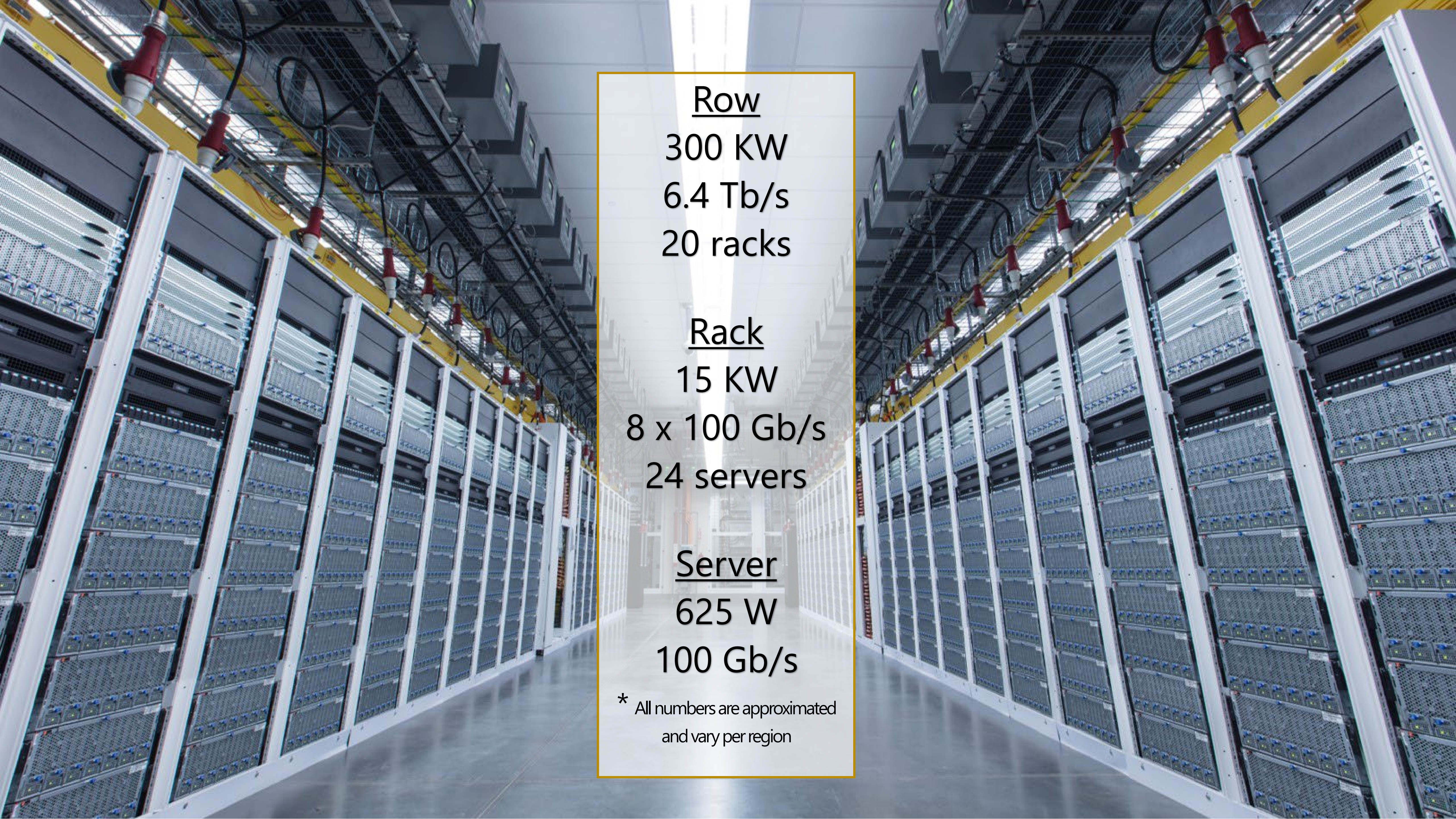
Dublin Datacenter



Looking Outside In - Data Centers

- 5-80 MegaWatts
- Hundreds of TeraBits/s
- Partitioned into ~100 rows





Row

300 KW

6.4 Tb/s

20 racks

Rack

15 KW

8 x 100 Gb/s

24 servers

Server

625 W

100 Gb/s

* All numbers are approximated
and vary per region

Down to the Server

- 2-Socket x86 Server is bread & butter
- Many variations (SKUs) for different workloads
- Rapidly increasing footprint for AI workloads



AI in Azure

AI and Machine Learning

Azure Publicly Available Services

VISION
Computer Vision

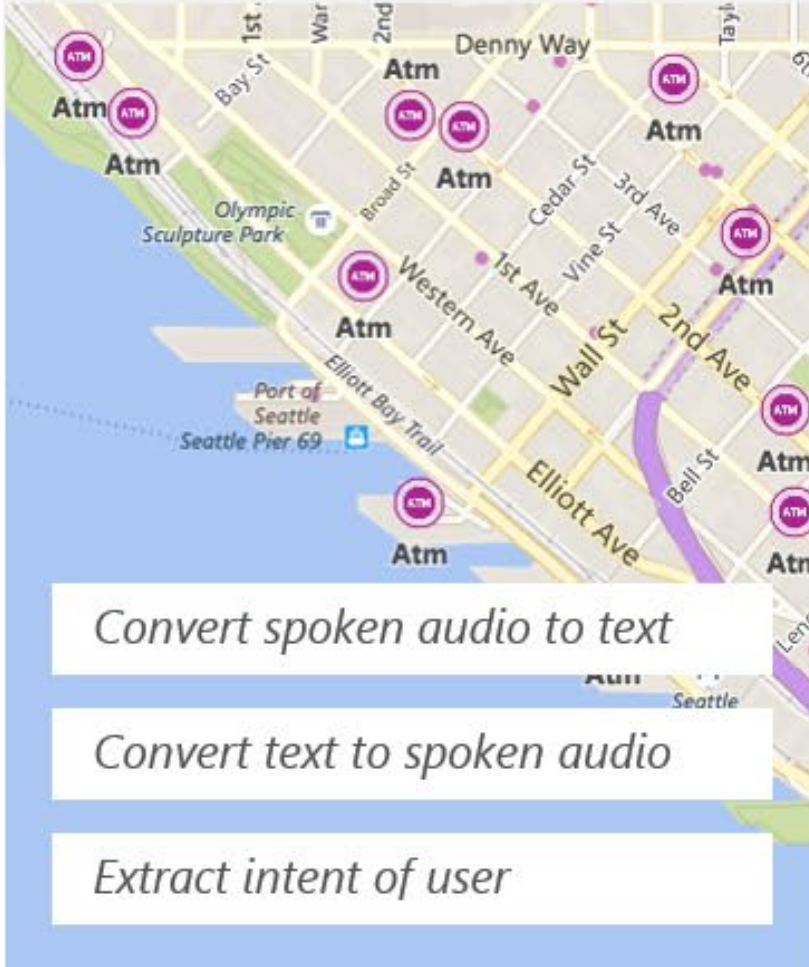
What is in the image?



Category	People; 1 face
Adult / Racy?	False / True
Dominant colors	☐ ☒ ☐
Accent color	☒

SPEECH
Speech-to-Text and Text-to-Speech Service

Give me directions to the closest ATM



Convert spoken audio to text

Convert text to spoken audio

Extract intent of user

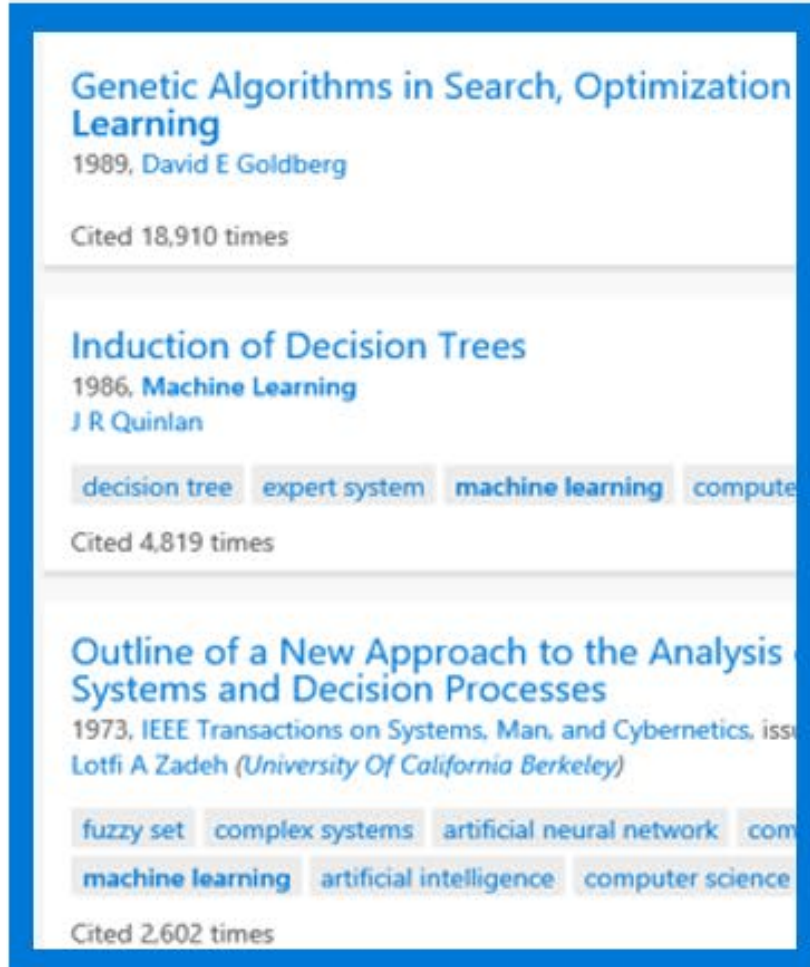
LANGUAGE
Natural Language Processing

Play today's Daily Show



KNOWLEDGE
Knowledge Exploration Service

Top publications in AI?



Genetic Algorithms in Search, Optimization Learning
1989, David E Goldberg
Cited 18,910 times

Induction of Decision Trees
1986, Machine Learning
J R Quinlan
decision tree expert system machine learning compute
Cited 4,819 times

Outline of a New Approach to the Analysis Systems and Decision Processes
1973, IEEE Transactions on Systems, Man, and Cybernetics, iss
Lotfi A Zadeh (University Of California Berkeley)
fuzzy set complex systems artificial neural network com
machine learning artificial intelligence computer science
Cited 2,602 times

SEARCH
Bing Search

Search for 'cute kittens'

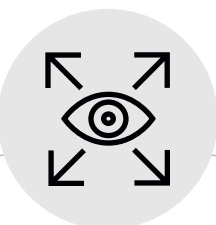


Computers **understanding** the world

Perception

Comprehension

Vision



Speech



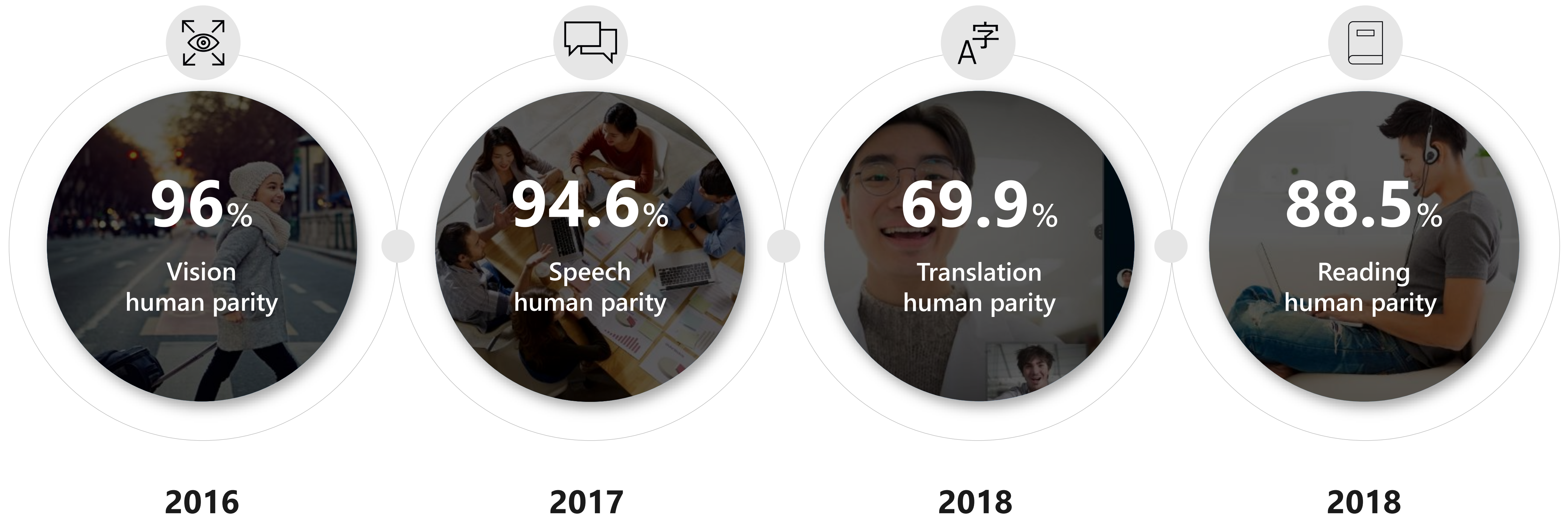
Language



Knowledge



Reaching human parity



Advances that make AI real

Vast amounts of data



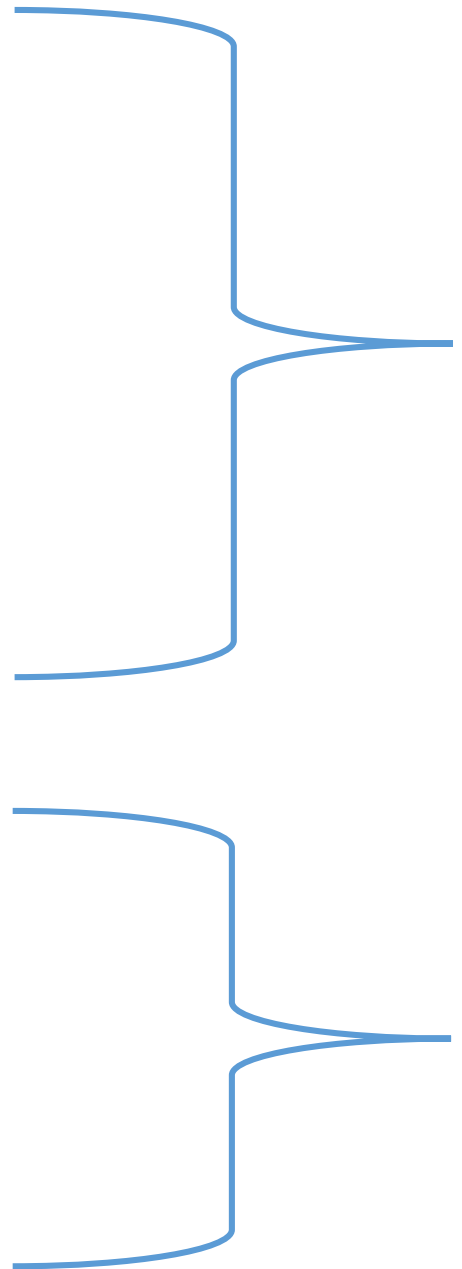
Huge computational power



Microsoft AI Workloads

Large internal workloads

Bing
Office 365
Cortana
Skype translation
Hololens, etc.



TensorFlow, Pytorch, CNTK => ONNX

Thousands of Azure customers

Cognitive services
Training and inference jobs
Etc.

Similar Frameworks

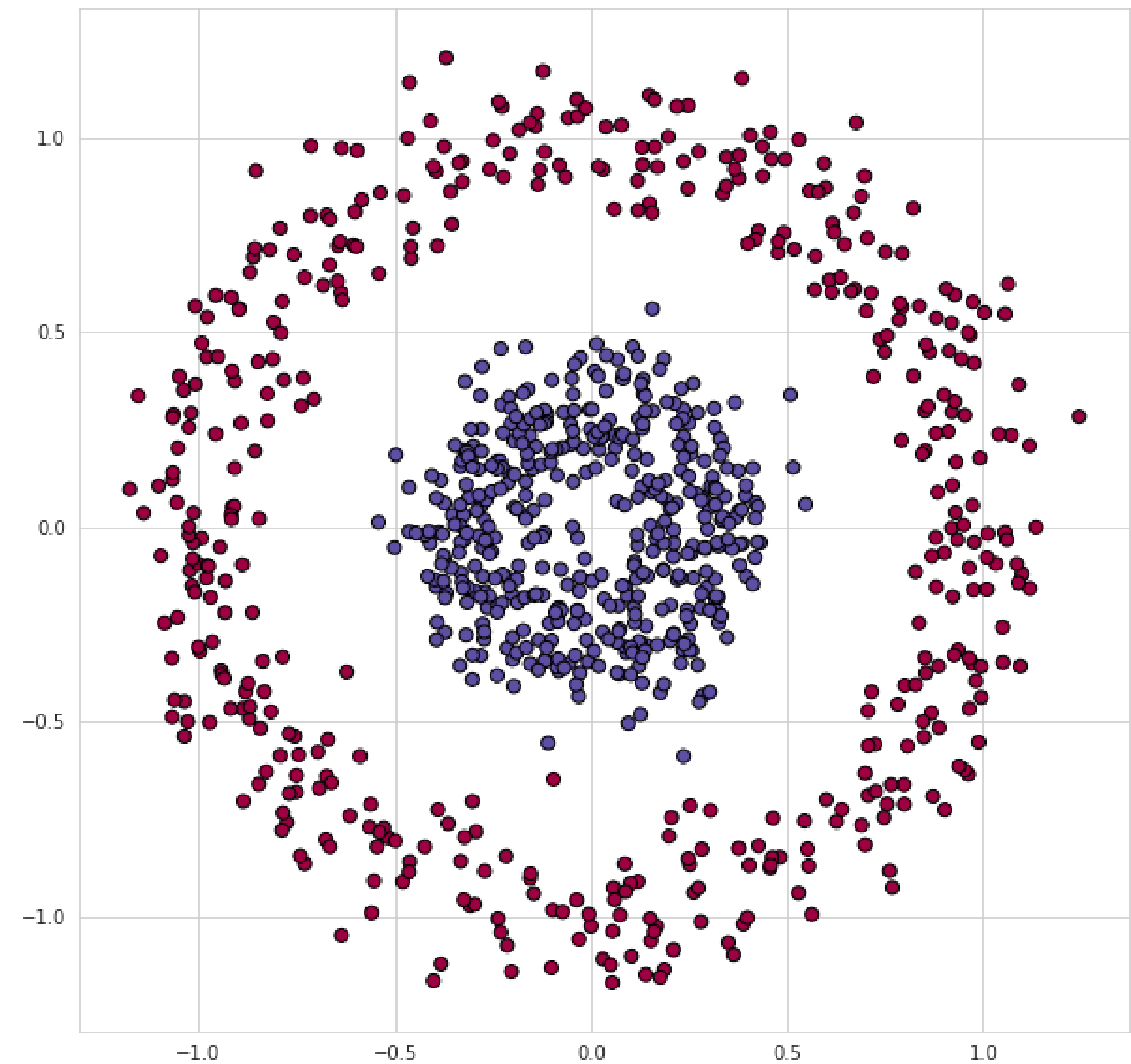
=> Broad class of workloads, frameworks, latency and throughput requirements

More Specifically

- Neural networks evolution:
 - Growing in complexity and size
 - Multiple neural networks per application
- Many RNNs throughout
- Bing architected for Batch Size = 1 for shortest latency
- Training jobs use much larger Batch Sizes (>100)
- Some training jobs take 22 days to run on many GPUs
- Infinite appetite for speed and scale
- Reliability & security paramount

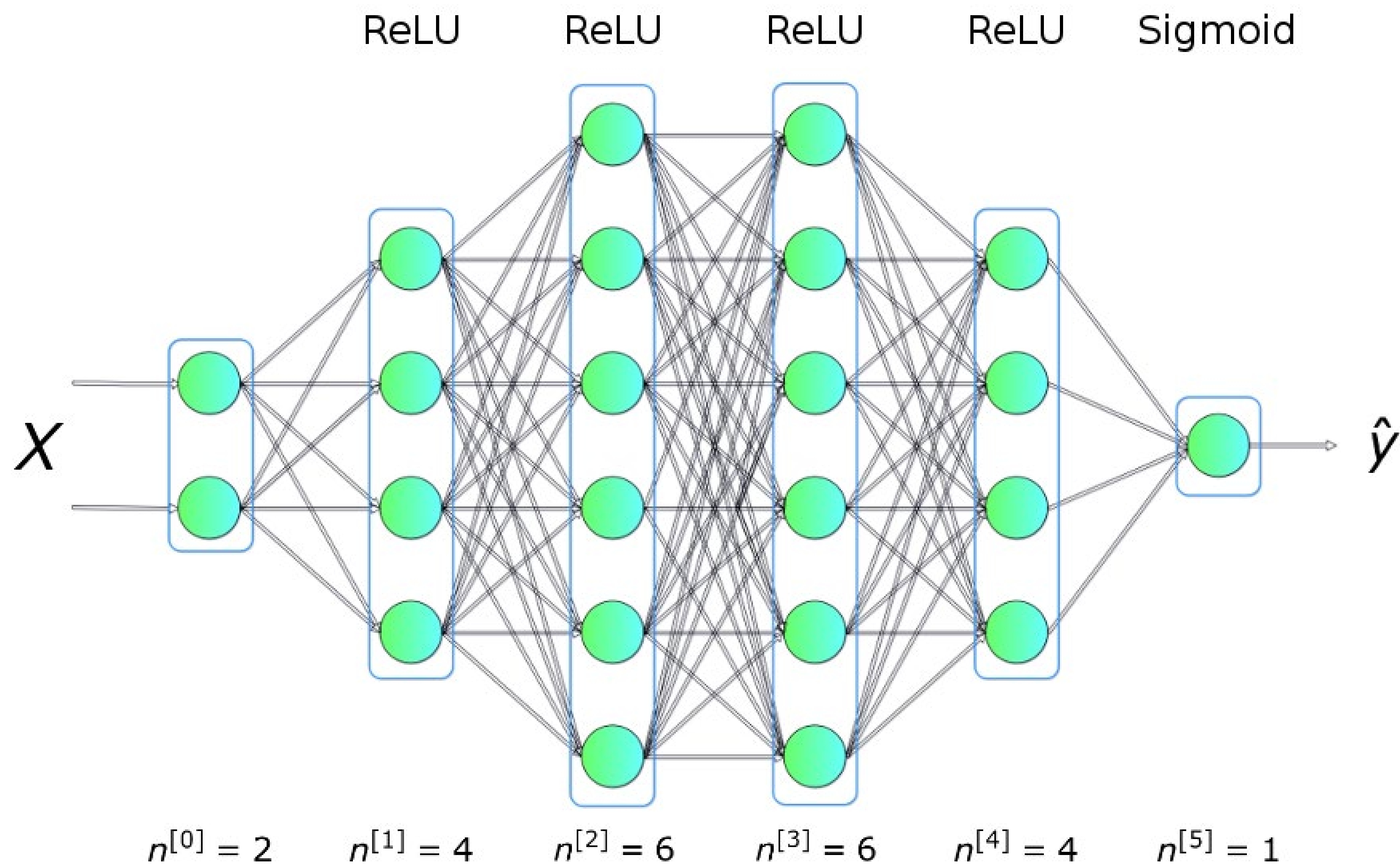
Neural Networks

Classification Problem

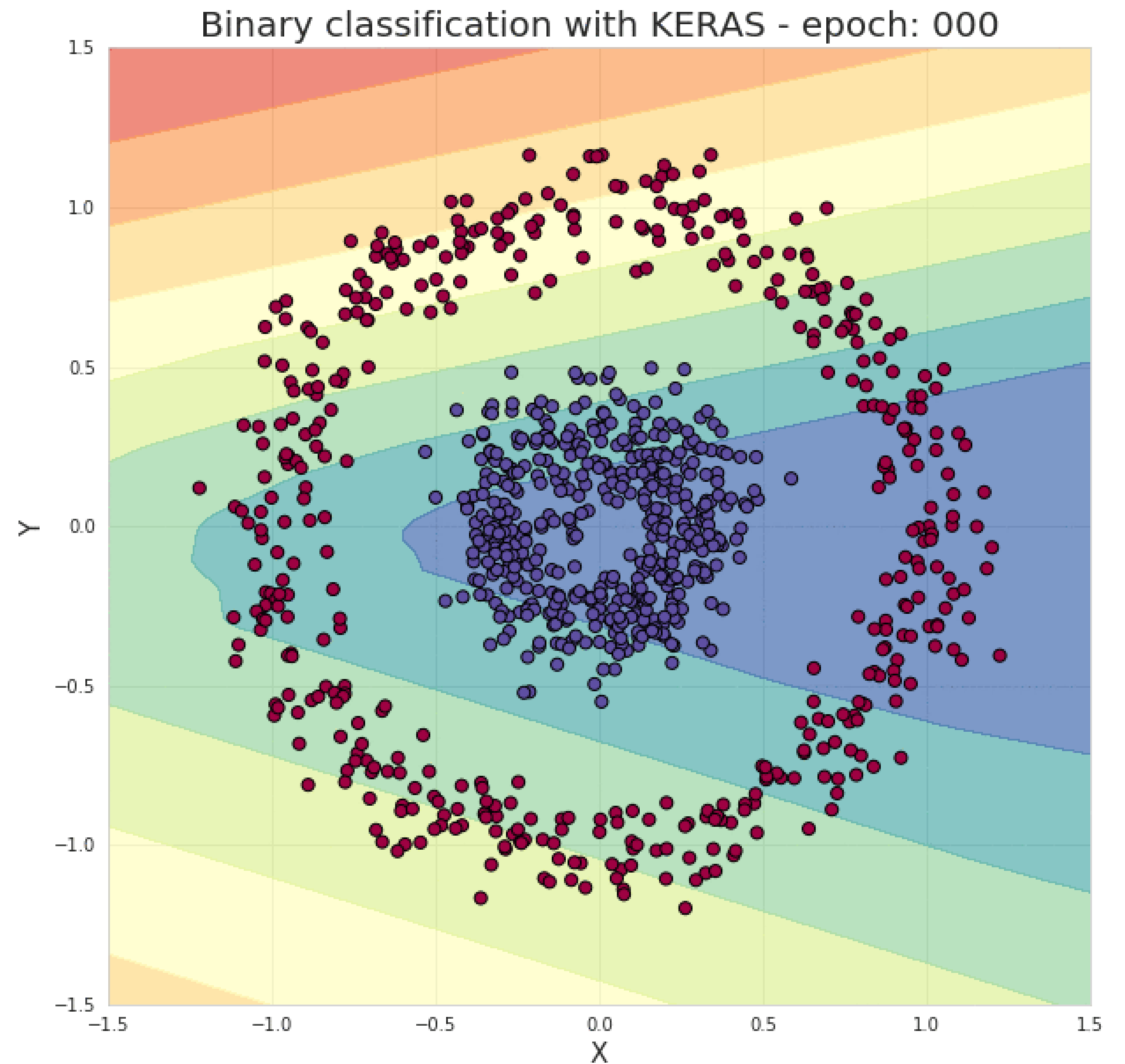


Thanks to: Piotr Skalski

Fully Connected NN

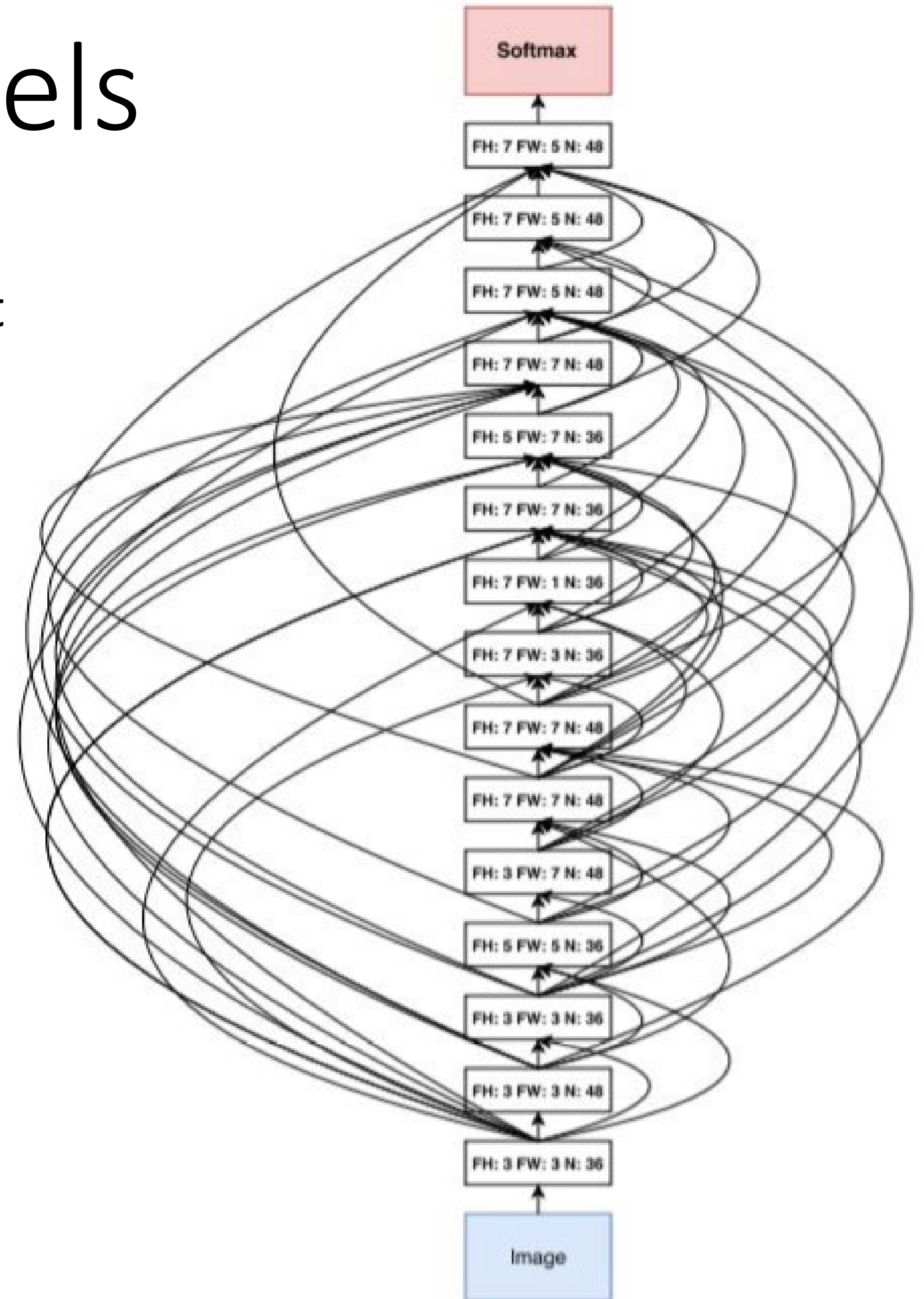


Neural Network Classification Converging



Models Generating Models

- Neural Architecture Search with Reinforcement Learning arxiv.org/abs/1611.01578
- Explorations of model space to find best model through reinforcement learning
- Evolution of Densenets
- But not by humans...



The Microsoft 2017 Conversational Speech Recognition System

- Several NNs used to reach a 5.1% word error rate
 - CNNs for acoustic modeling
 - CNN-BLSTM for additional acoustic model
 - LSTM for acoustic and language modeling (characters vs. words)
 - LSTM using entire preceeding conversation as history

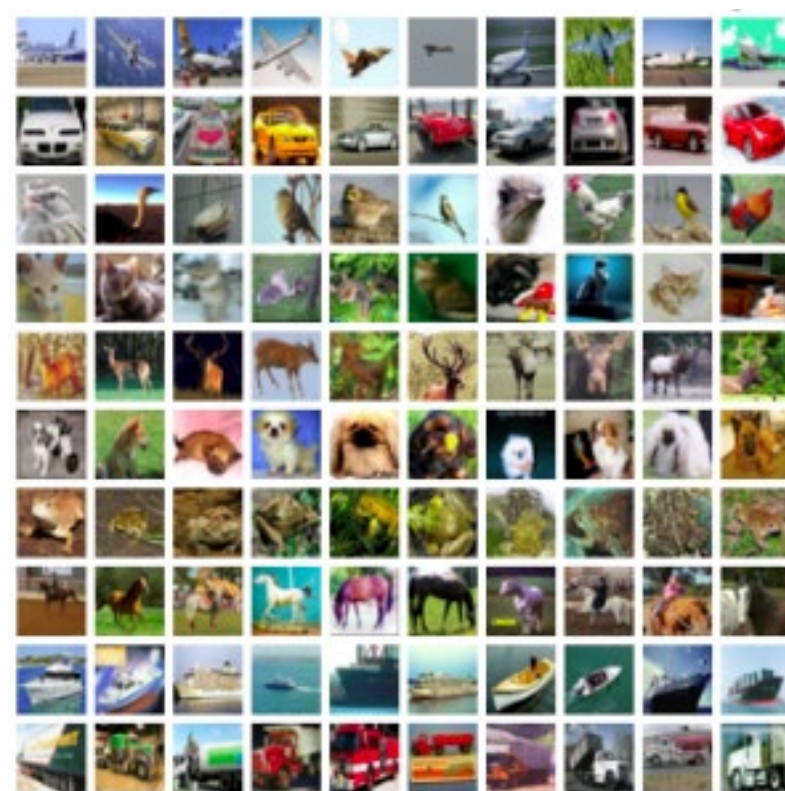
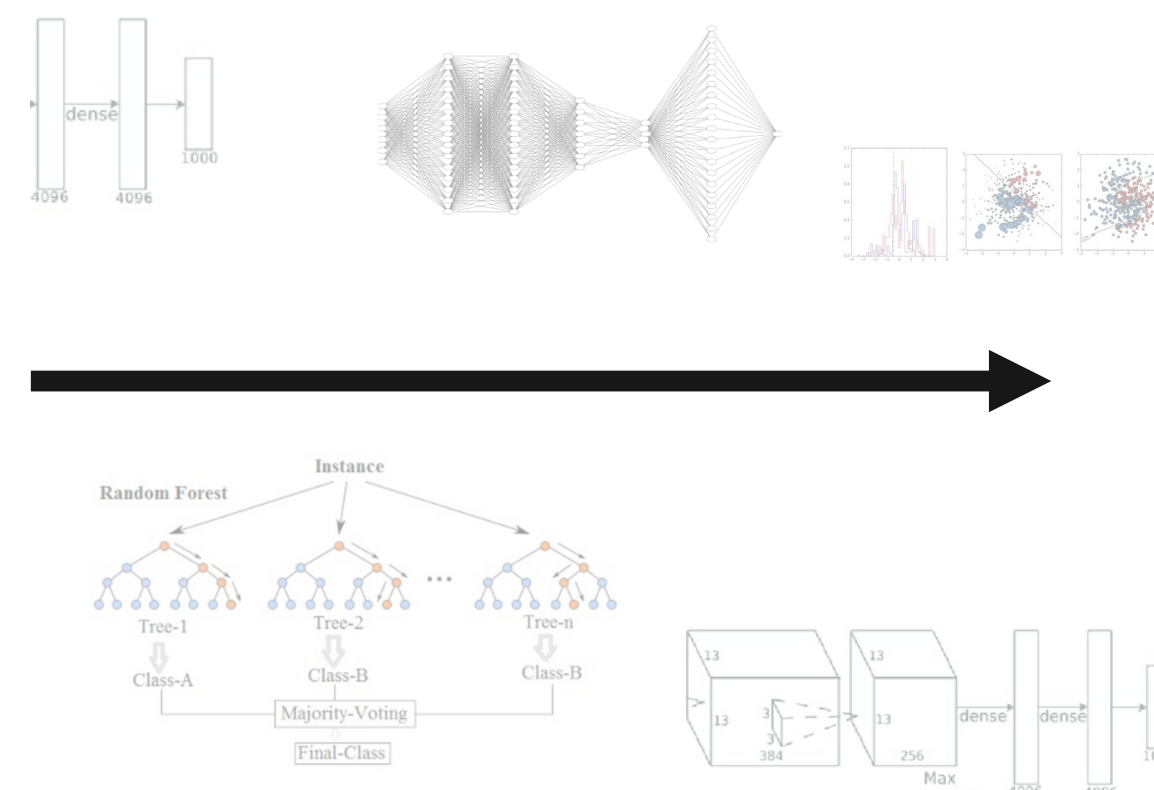
	ResNet	LACE	CNN-BLSTM
Number of parameters	38M	65M	48M
Number of weight layers	49	22	10

- Scoring it all:

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
27k-puhpum	BLSTM	11.3	8.0	9.2	6.3
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.4
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.8
-	Confusion network combination			7.4	5.2
-	+ LSTM rescoring			7.3	5.2
-	+ ngram rescoring			7.2	5.2
-	+ backchannel penalty			7.2	5.1

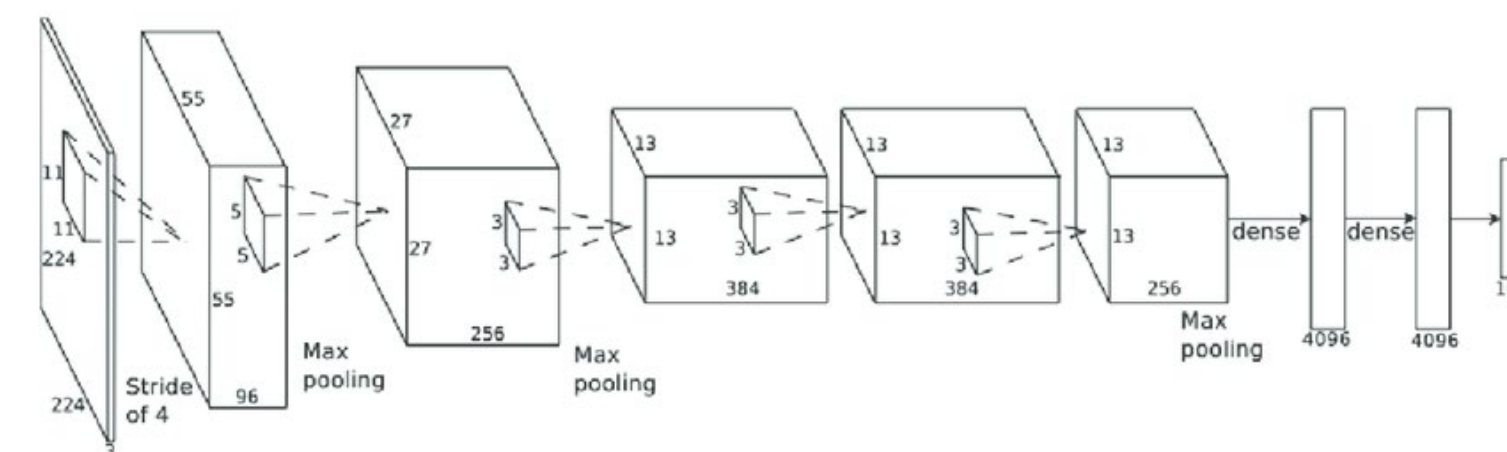
Machine Learning's Weak Spot?

Problem



- How do we automate this for Azure customers?
- Massive compute + AutoML

Solution



Probabilistic Matrix Factorization for Automated Machine Learning

Nicoló Fusi
Microsoft Research
Cambridge (MA), USA
fusi@microsoft.com

Huseyin Melih Elibol
Microsoft Research
Cambridge (MA), USA
v-huelib@microsoft.com

Abstract

The application of machine learning techniques to real-world problems often requires accurate tuning of several key parameters as well as the identification of data pre-processing steps (e.g. whitening images in a DNN) and model ensembling. Automating the identification of the appropriate machine learning workflow (pre-processing + model selection + ensembling), as well as the tuning of all the necessary parameters, has long been one of the goals of the machine learning community. Existing approaches outperform humans in identifying good parameters for a given model and carefully processed data, but haven't fully realized the promise of automating machine learning. We cast this problem as a probabilistic matrix factorization task and use hundreds of different datasets to train a probabilistic latent variable model that embeds different machine learning workflows based

Reconciling Cloud Architecture and AI Requirements

What We Are Looking For

- Faster, more accurate
 - Time per epochs, images/sec, sentences/sec, etc.
- Cheaper hardware and lower power
- Quantized leap over natural time periods
 - Used to be “Render a frame overnight”
 - Now “Train over lunch”
- Easily deployable in our datacenters
 - Doesn’t break our infrastructure
- Robustness
 - For emerging NNs
 - Not yet invented NNs

Inference

- Today:
 - CPUs: Internal workloads e.g. Cortana, Azure ML, Azure IaaS
 - FPGAs: Bing search, Azure Inference Service, etc.
 - GPUs: Bing image search acceleration
- On-line services
 - Strict latency requirements
 - Throughput is where the money is
- Off-line services
 - Throughput is key with reasonable latency

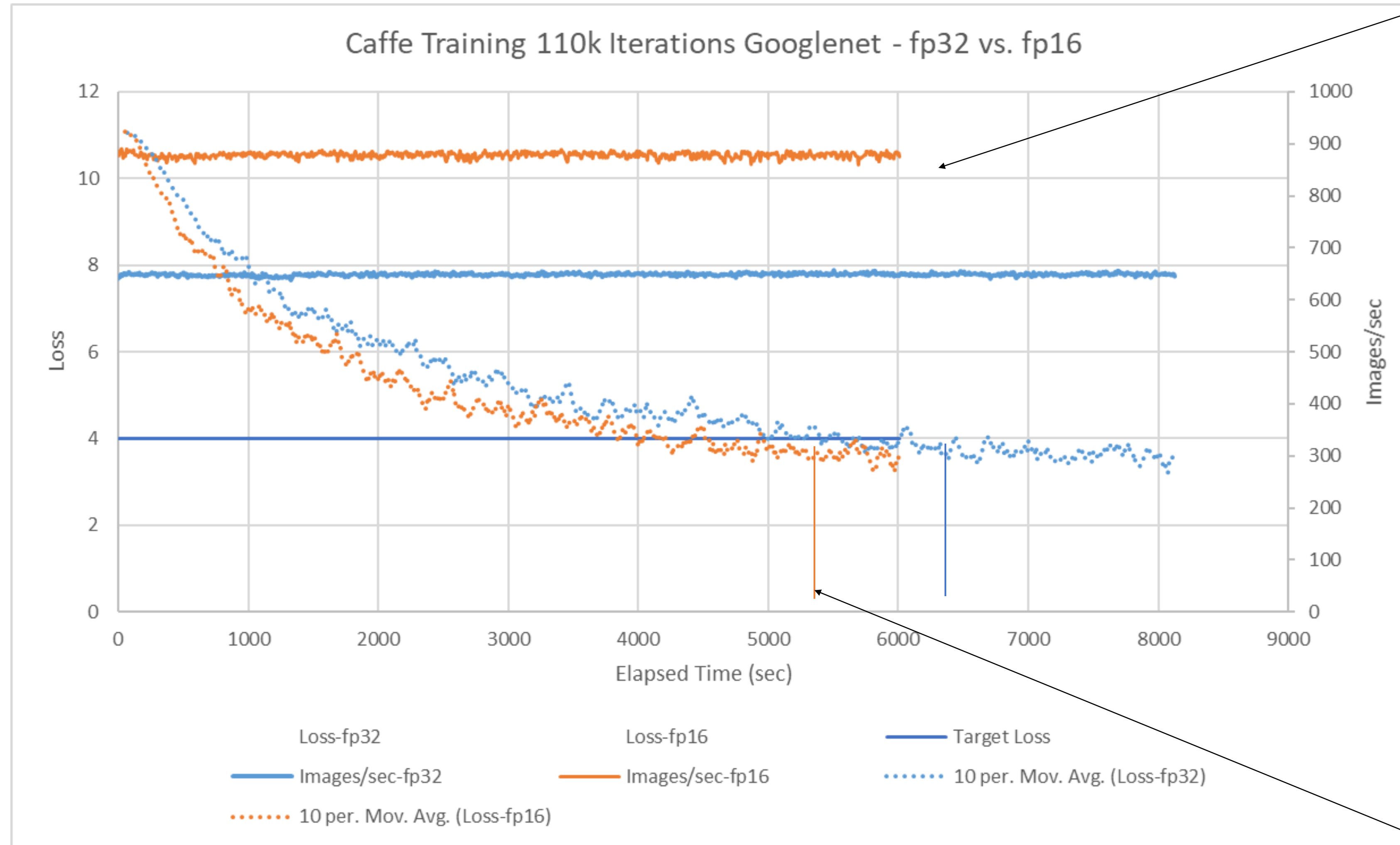
Training

- Today: Nvidia P40, P100, v100, 4-8 cards/system, systems interconnected through Infiniband
- Some of our workloads require 22 days to train, would like to do it “over lunch”
- Data parallelism but clustering often challenging
- Large domains attractive - 100 PetaFLOPs to 1 ExaFLOPS
- 16bit FP mostly working
- Hierarchical bandwidth:
 - Within linear algebra (across processing elements or cores or neurons)
 - Within a mini-batch
 - Across mini-batches

What Matters - Time to Train

FP32 vs. FP16 – 2x V100, BS 96

Fp16 tput 35%
higher



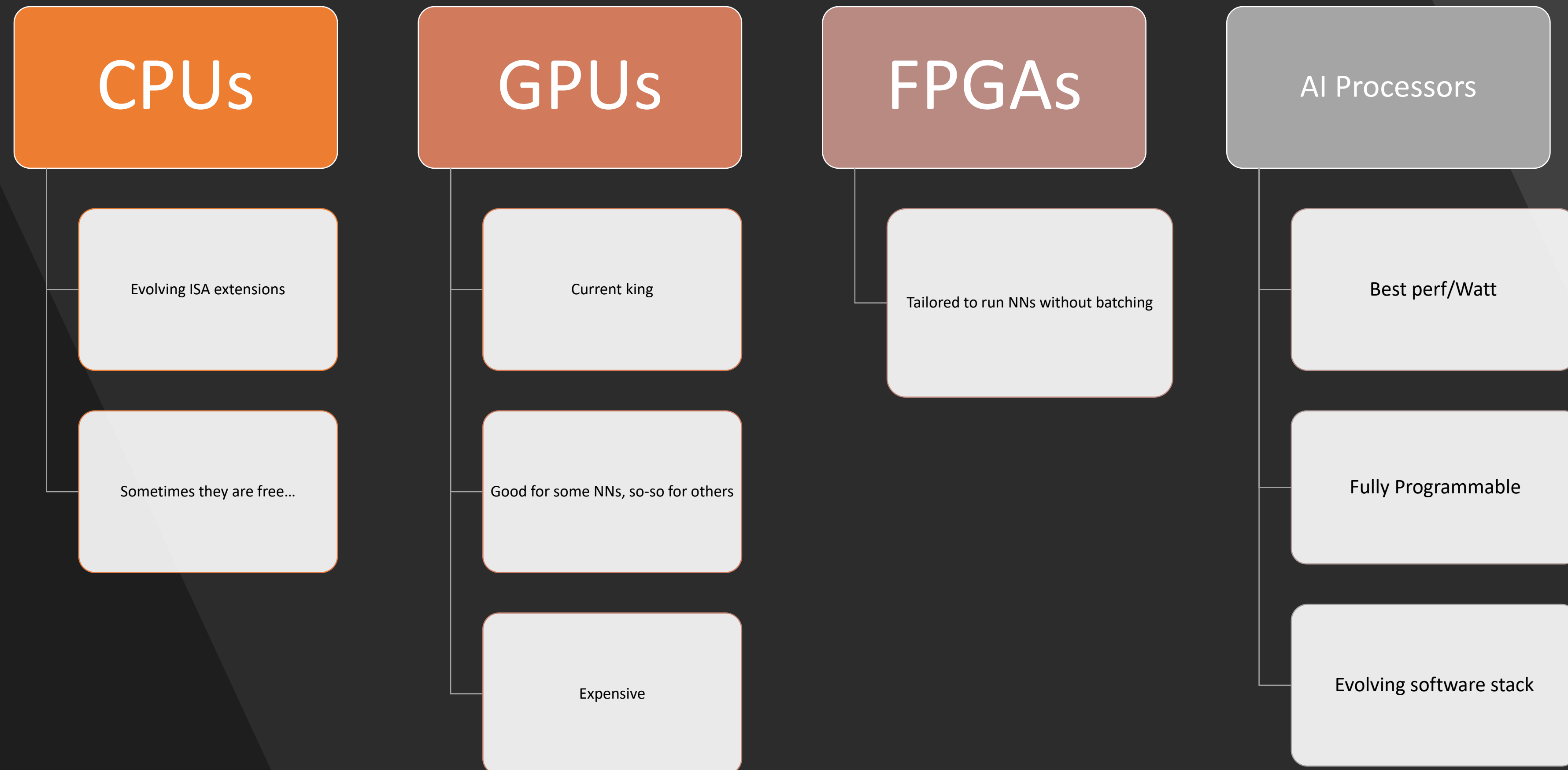
Fp16 reaches target
earlier

Accuracy for inference :

Fp32: top-1 – 50%, top-5 – 76%

FP16: top-1 – 50%, top-5 – 75%

Choices

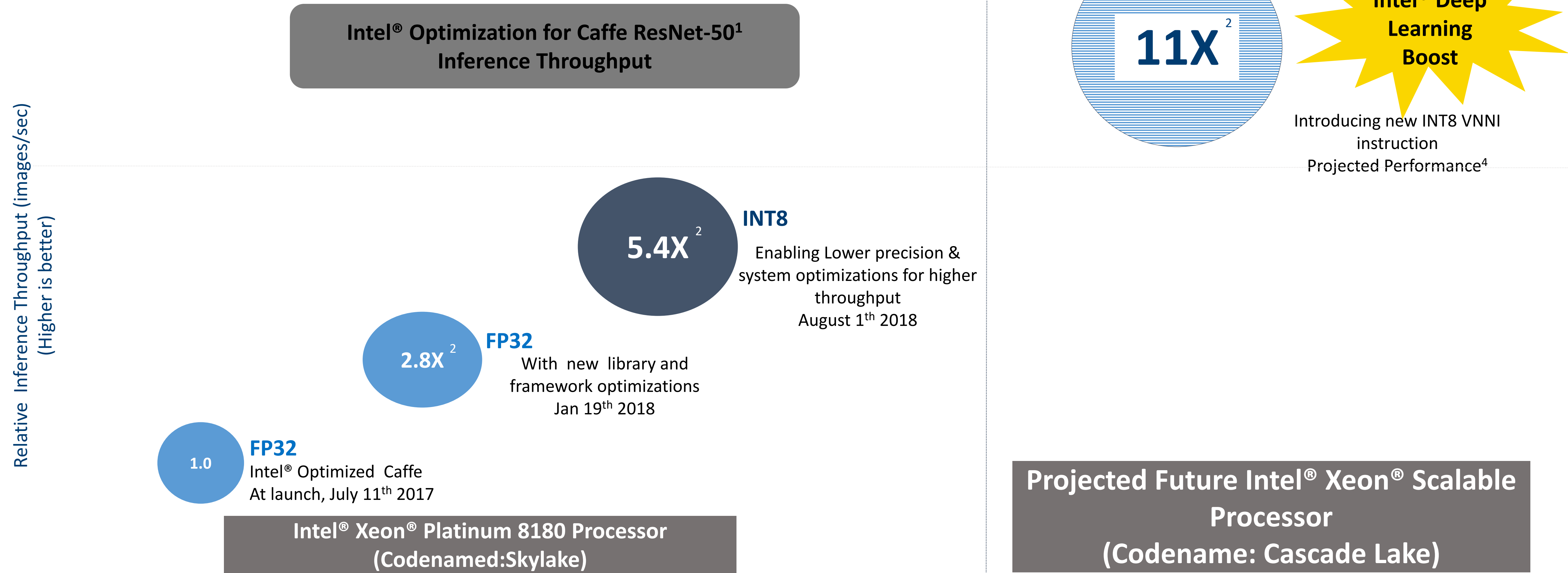


CPU_s

CPUs

- Millions of CPUs in our data centers
- Used extensively for inference
- Perhaps for training by some IaaS customers
- Sometimes “free” (available)
- ISAs evolving rapidly (Intel, AMD, ARM companies)

Continued Innovation Driving Deep Learning Inference Performance On Intel® Xeon® Scalable Processor



¹ Intel® Optimization for Caffe Resnet-50 performance does not necessarily represent other Framework performance.

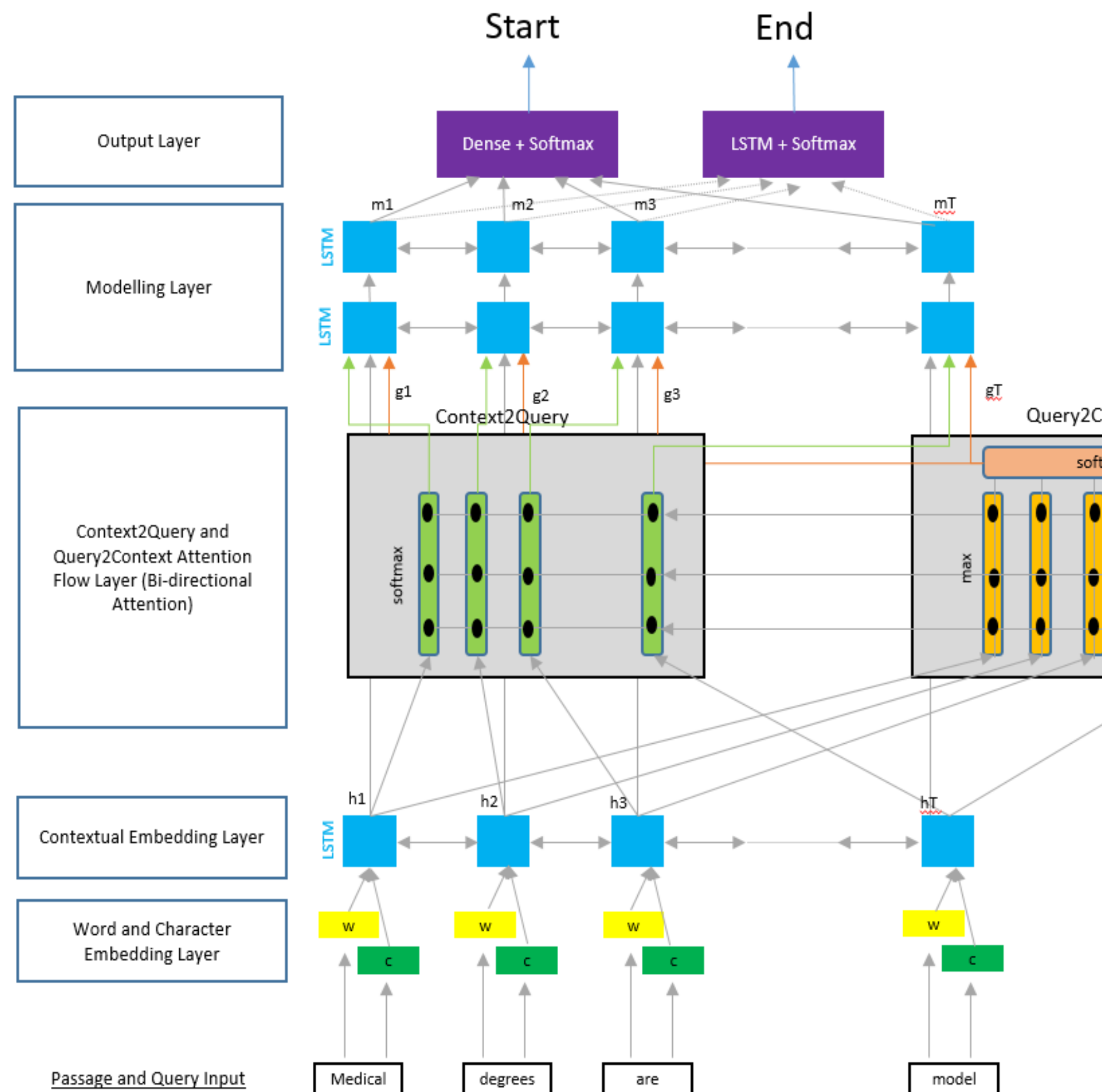
² Based on Intel internal testing: 1X (7/11/2017), 2.8X (1/19/2018) and 5.4X (7/26/2018) performance improvement based on Intel® Optimization for Café Resnet-50 inference throughput performance on Intel® Xeon® Scalable Processor. See Configuration Details 53

³ 11X (7/25/2018) Results have been estimated using internal Intel analysis, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

⁴Inference projections assume 100% socket to socket scaling

Performance results are based on testing as of 7/11/2017(1x), 1/19/2018(2.8x) & 7/26/2018(5.4) and may not reflect all publically available security update. See configuration disclosure for details (config 53). No product can be absolutely secure. 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.

Microsoft's *Deep CPU* Optimization Results



Bing Turing Prototype-3

1. Tensorflow Serving latency: 107ms (**not-shippable**)
2. Target latency: < 10ms

Our Optimization

- Build optimized CPU implementation for TP3
- Fast LSTM and co-attention computation

Same accuracy

Latency: **4.1ms (>20 times speedup)**

Throughput: more than **50 times**

Non-shippable -> **shippable**

GPUs

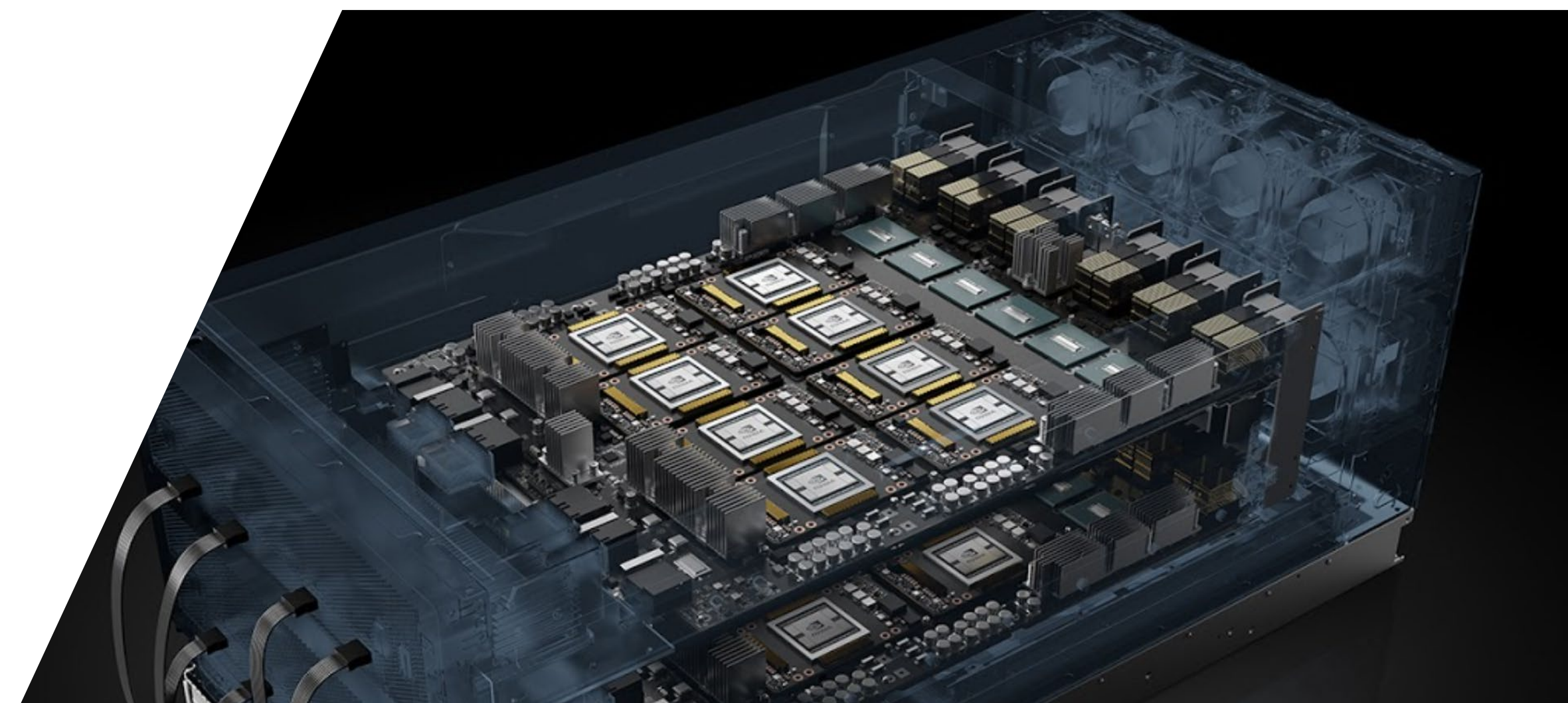
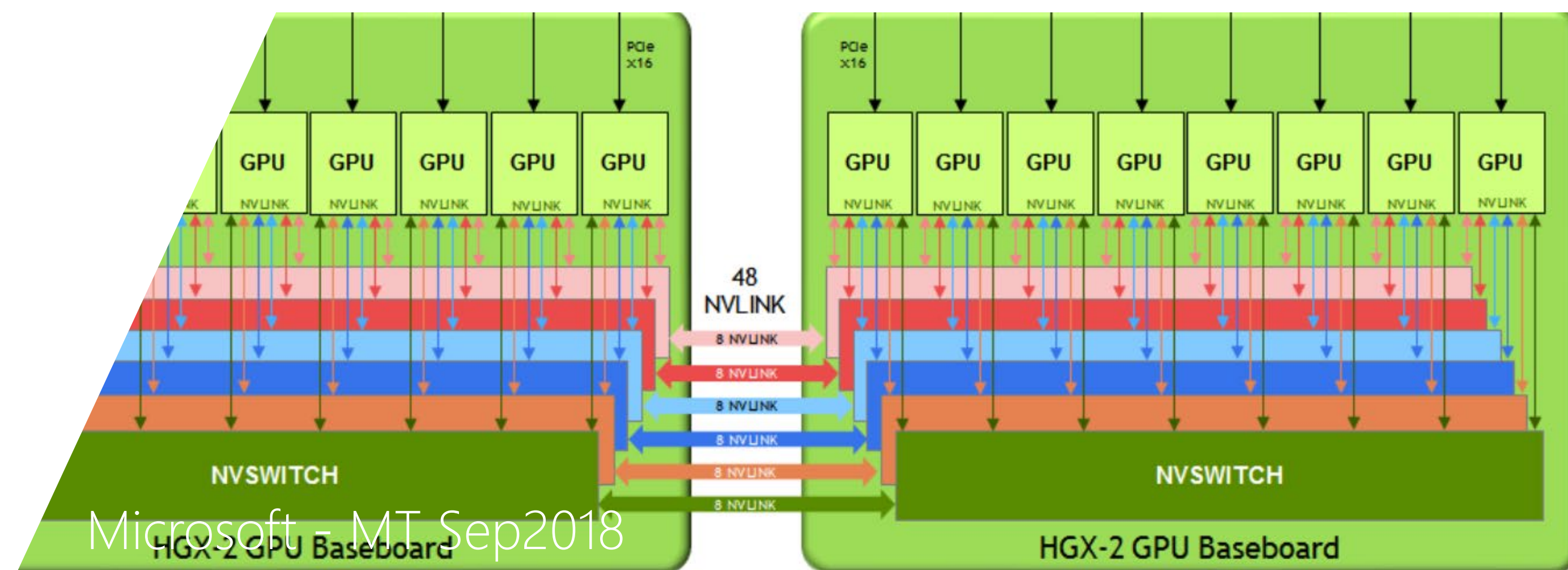
Nvidia Volta 100 – The Current Training King

- Handles graphics, HPC, AI
- 64fp, 32fp, 16fp
- Tensor cores 4x4+4
- 4 HBM2 stacks
- Nvlinks
- Top available training engine
- Pricey
- Low efficiency on many NNs



GPU System – Nvidia DGX2

- 16 chips in a system
- High-end interconnect
- All to all through hot switches
- 1 system per rack in some regions
- \$400k

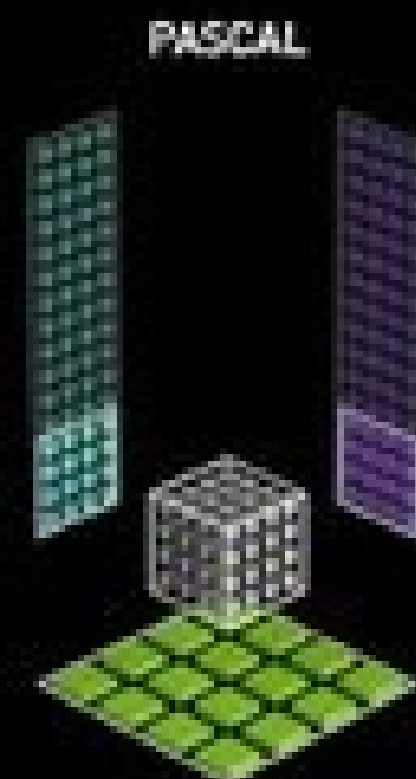
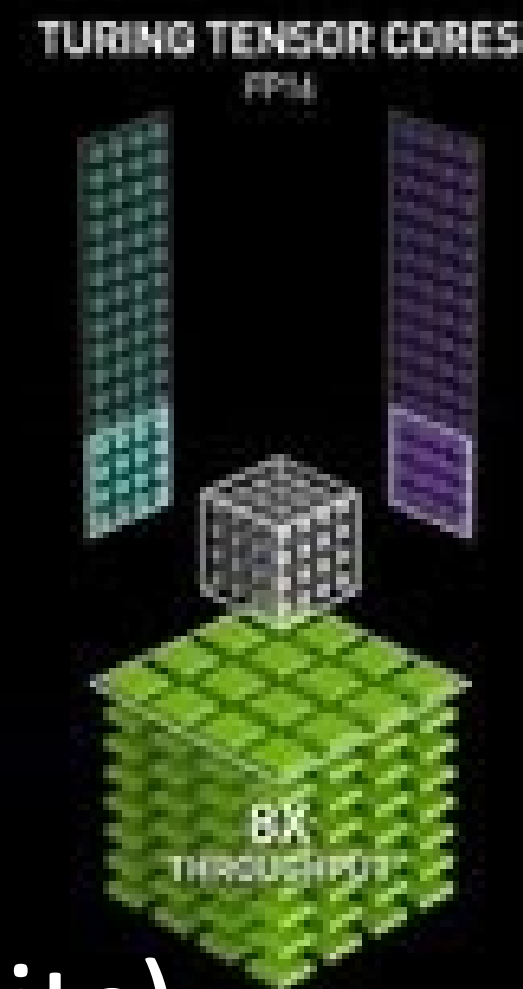
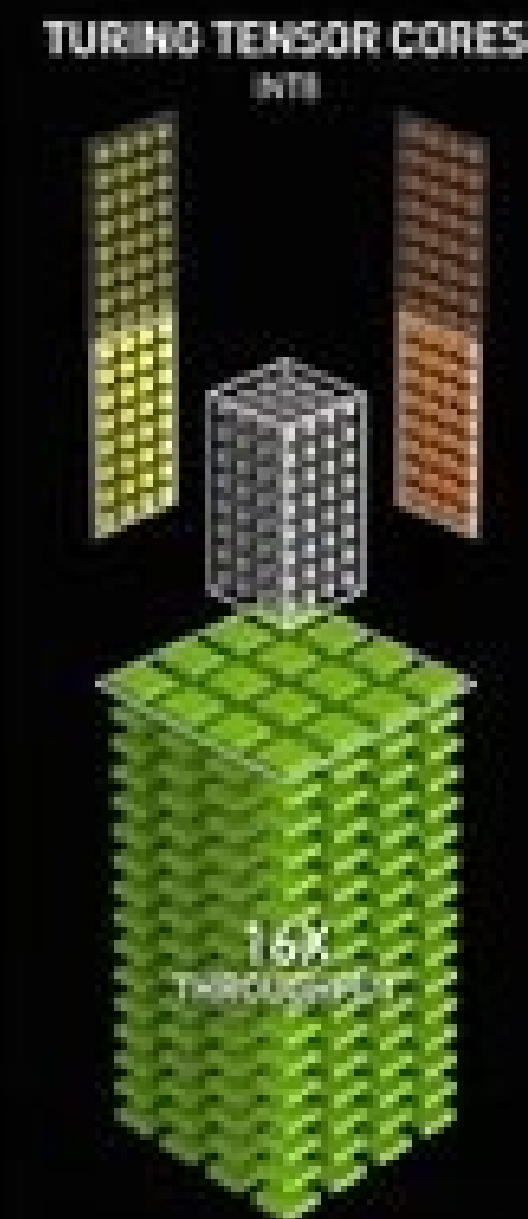
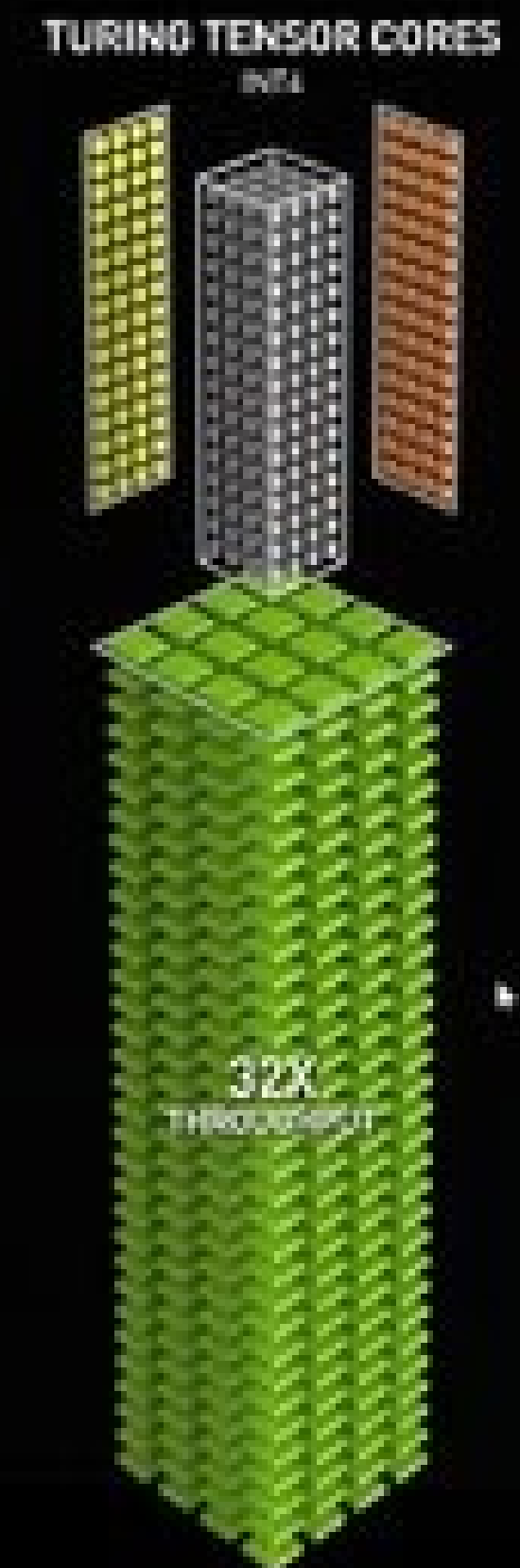


NEW TURING TENSOR CORE

MULTI-PRECISION FOR AI INFERENCE

65 TFLOPS FP16 | 130 TeraOPS INT8 | 260 TeraOPS INT4

- Very interesting
- New Tensor cores supporting lower precision (16, 8, 4 bits)
- Virtualization improvement (SR-IOV)
- Multi-Process Service – Improves throughput for low batch size inference



AI Processors

AI Processors

- Billions of \$\$\$ invested in startups and large companies
 - More than 50 companies
 - Wild wild west
 - Renaissance of processor and computer architecture
- Training and inference
- All using some form of graph-based IR
- Very programmable
- Battle is on for best performance in throughput, latency and cost
- Two major categories emerging
 - Feed the beast through external memory
 - Keep all data on-chip

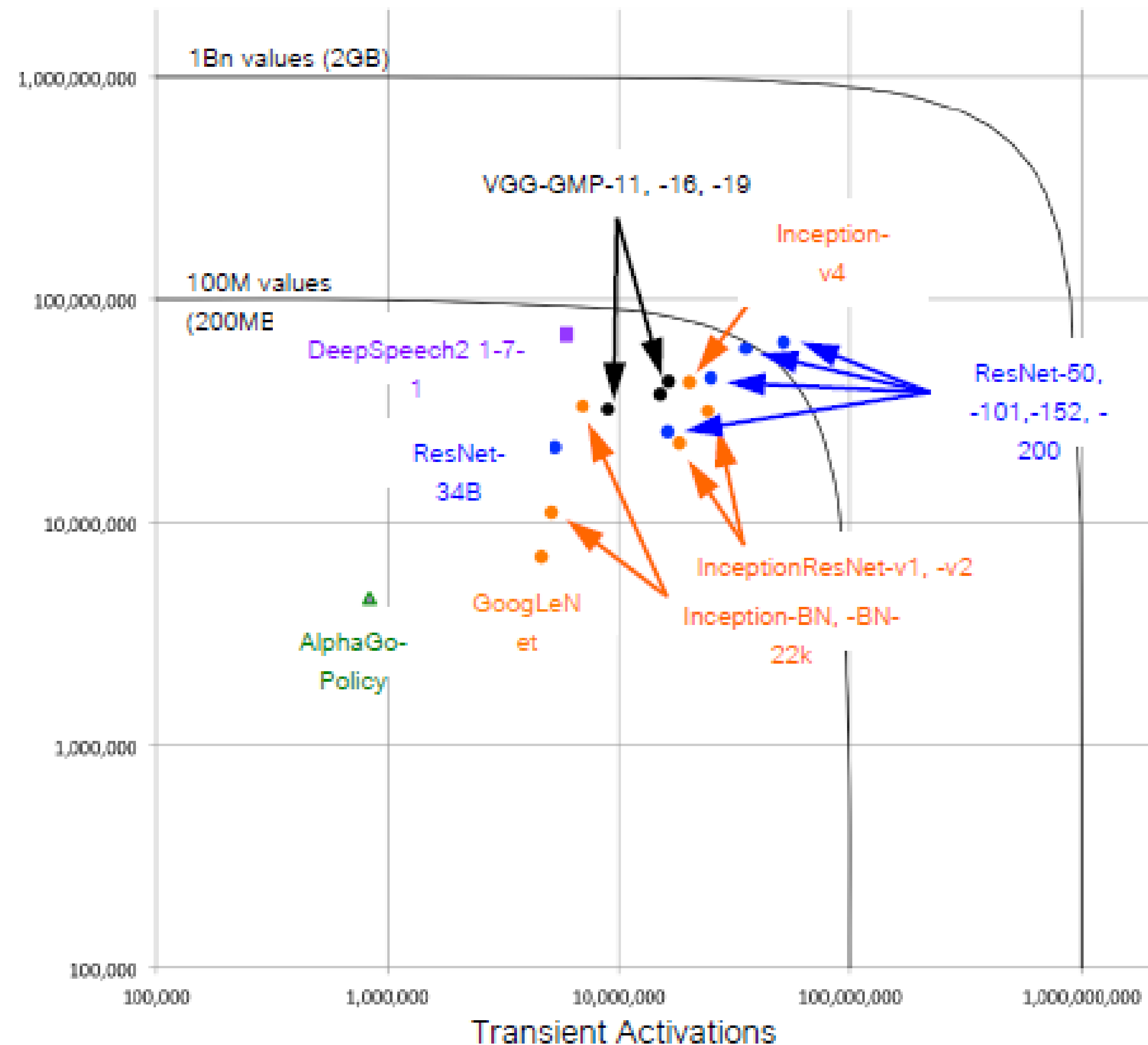
Memory Bandwidth Challenge

- Current state-of-the-art GPU
 - ~100 TFLOPS (16fp)
 - ~1TB/s through 4 HBM2 stacks
 - ⇒ 1 byte per 100 FLOPS
 - Extreme HPC would be 32 bytes per FLOPS (good old Cray Systems)
 - A factor of 3200x
 - Not necessary, but we could use a factor of 10-50x
- Caching, tiling, pinning, tensor operations, smaller data types, all reduce the gap
 - But nowhere close to what is needed
 - Really hurts performance of RNNs and emerging NNs
- New technology on the horizon
 - Faster serdes 56 Gbps, 112 Gbps
 - Advanced packaging
- Heavily used by GPUs, CPUs, and some AI processors

Alternative Approach – NN Model on Chip

- Moore's Law gives us the transistors
 - Compelling SRAM cells
 - Large on-chip memory possible
- What can we fit?
- Example - ResNet 50
 - 26M parameters
 - 16M activations in forward pass
 - 42M total
 - 84 MB Using 16-bit FP
 - More due to fragmentation and tem
 - Feasible in 16nm, easy in 7nm
- Bulk of the data streams through
 - Images, sentences, etc.
- Discussed by academia and some AI companies

**Learned
Parameters**



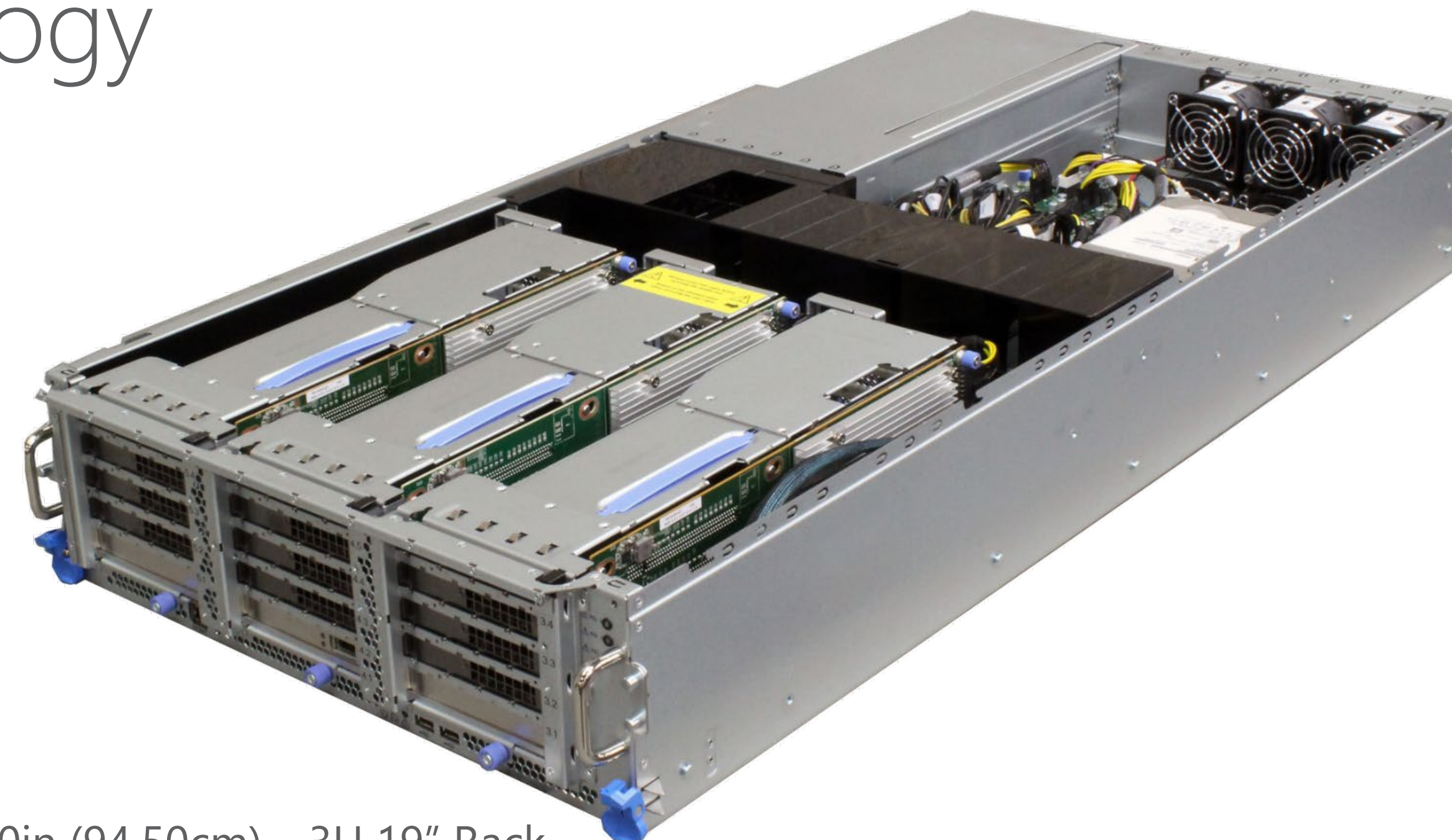
Systems



XPO200 3U PCIe Expansion System featuring NVidia Technology



- **Intel Xeon / NVidia P4 System**
- **Model #:** ZT-XPO200-3UN1810
- **Processor:** 2 x Intel® Xeon® Processor Scalable Family Platinum 8168 (24 core, 2.7 Ghz)
- **Memory** 384GB 2666MHz DDR4 ECC (12 x 32GB RDIMM), expandable to 1536GB (24 total slots)
- **M.2 Storage:** 1 x 960GB M.2 NVMe PCIe SSD (expandable to 4 total M.2 modules on-board)
- **Networking:** 10G Single port SFP+ PCIe 2.0 x8 5GT/s
- **GPU:** 12 x Nvidia P4 GPU Cards
- **Dimensions:** Height: 5.20in (13.20cm) Width: 17.36in (44.10cm) Depth: 37.20in (94.50cm) – 3U 19" Rack
- **Weight:** 90lbs (40.8 Kg)
- **OS Support:** Windows® Server 2016
- **Expansion:**
 - Option 1: 12 (PCIe x16) FHFL Single-Wide slots, 1 (PCIe x16) FHHL slot
 - Option 2: 6 (PCIe x16) FHFL Double-Wide slots, 1 (PCIe x16) FHHL slot
- **Power Supply** 1000W Non-LES (Project Olympus rack with PMDU required)



Designed to deliver a powerful, flexible and cost-effective solution for GPU-intensive scale-out computing.

OPEN. FOR BUSINESS

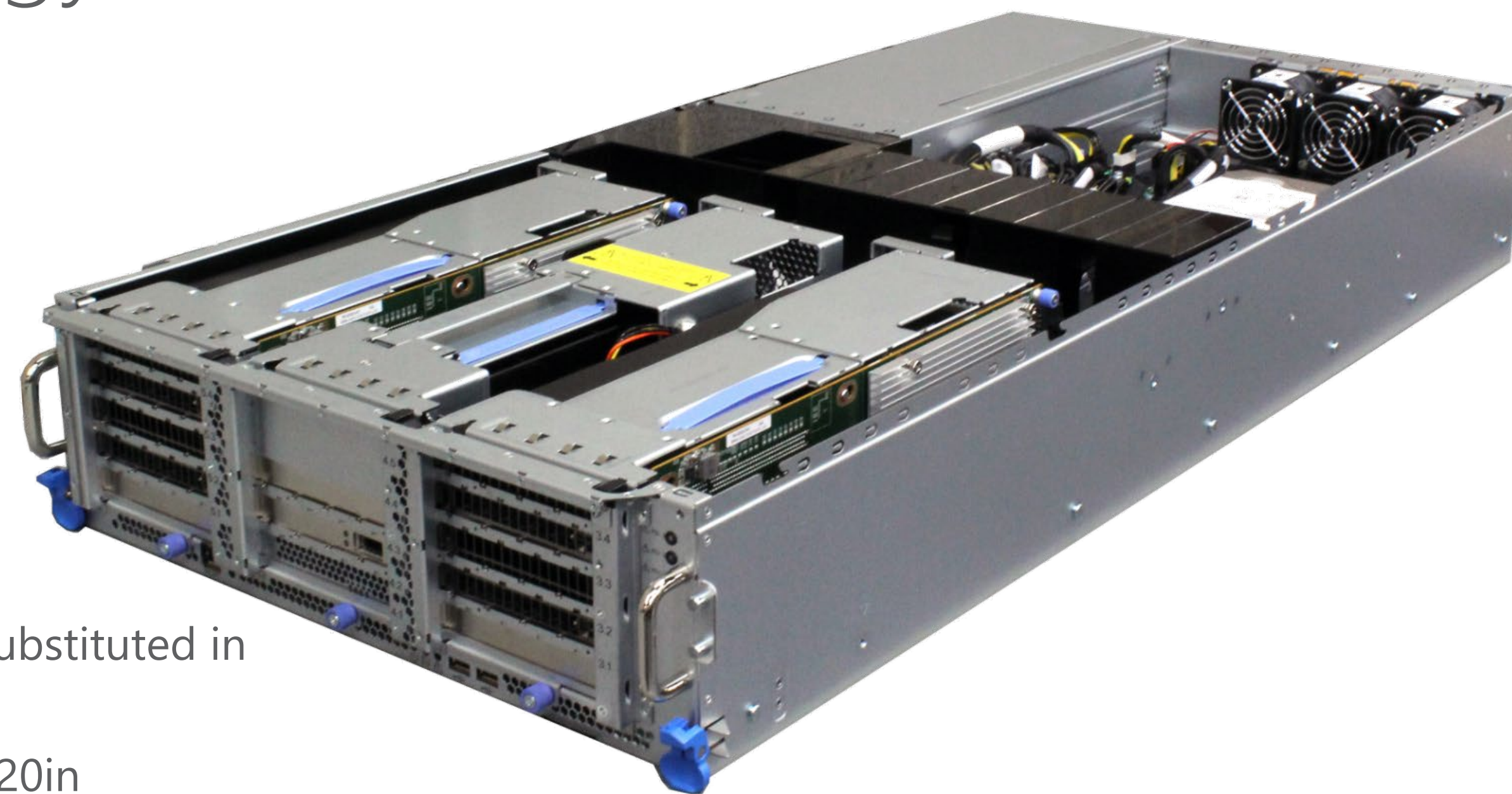




XPO200 3U PCIe Expansion System featuring AMD Technology



- **AMD EPYC / AMD MI25 system**
- **Model #:** ZT-XPO200-3UA1810
- **Processor:** 2 x AMD EPYC™ 7551, 32C, 180W, 2GHz
- **Memory:** 512GB 2666MHz DDR4 ECC (16 x 32GB RDIMM), expandable to 1024GB (32 total slots)
- **M.2 Storage:** 8 x 960GB M.2 NVMe PCIe SSD (4 on-board w/ 4 on an AVA Riser Card)
- **Networking:** 10G Single port SFP+ PCIe 2.0 x8 5GT/s
- **GPU** 4 x AMD MI25 (Vega10) GPU Cards, 16GB Memory (5th card can be substituted in place of M.2 AVA Riser)
- **Dimensions:** Height: 5.20in (13.20cm) Width: 17.36in (44.10cm) Depth: 37.20in (94.50cm) – 3U 19" Rack
- **Weight:** 90lbs (40.8 Kg)
- **OS Support:** Windows® Server 2016
- **Expansion:** 5 (PCIe x16) FHFL Double-Wide slots, 1 (PCIe x16) FHHL slot
- **Power Supply:** 1000W Non-LES (Project Olympus rack with PMDU required)

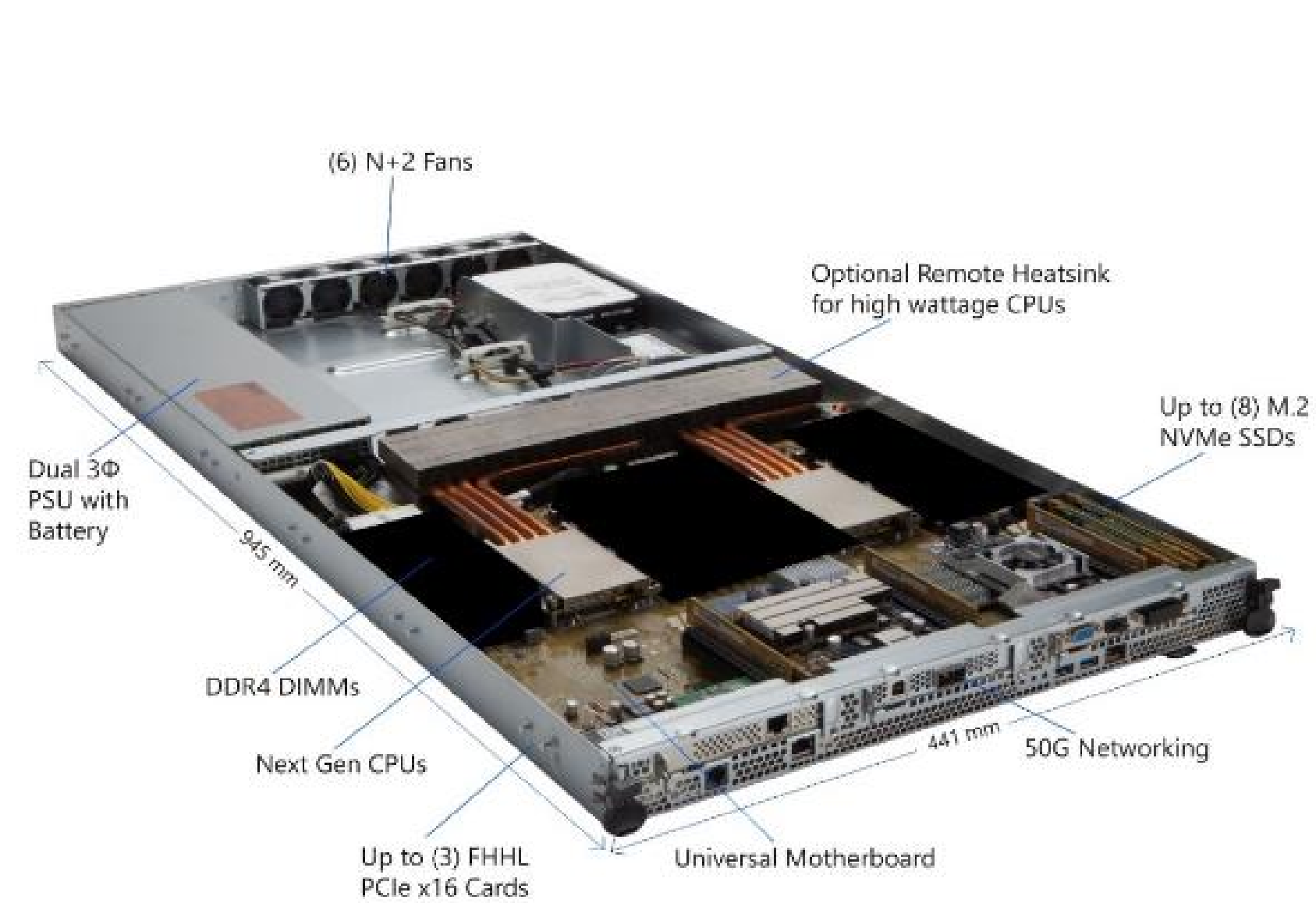
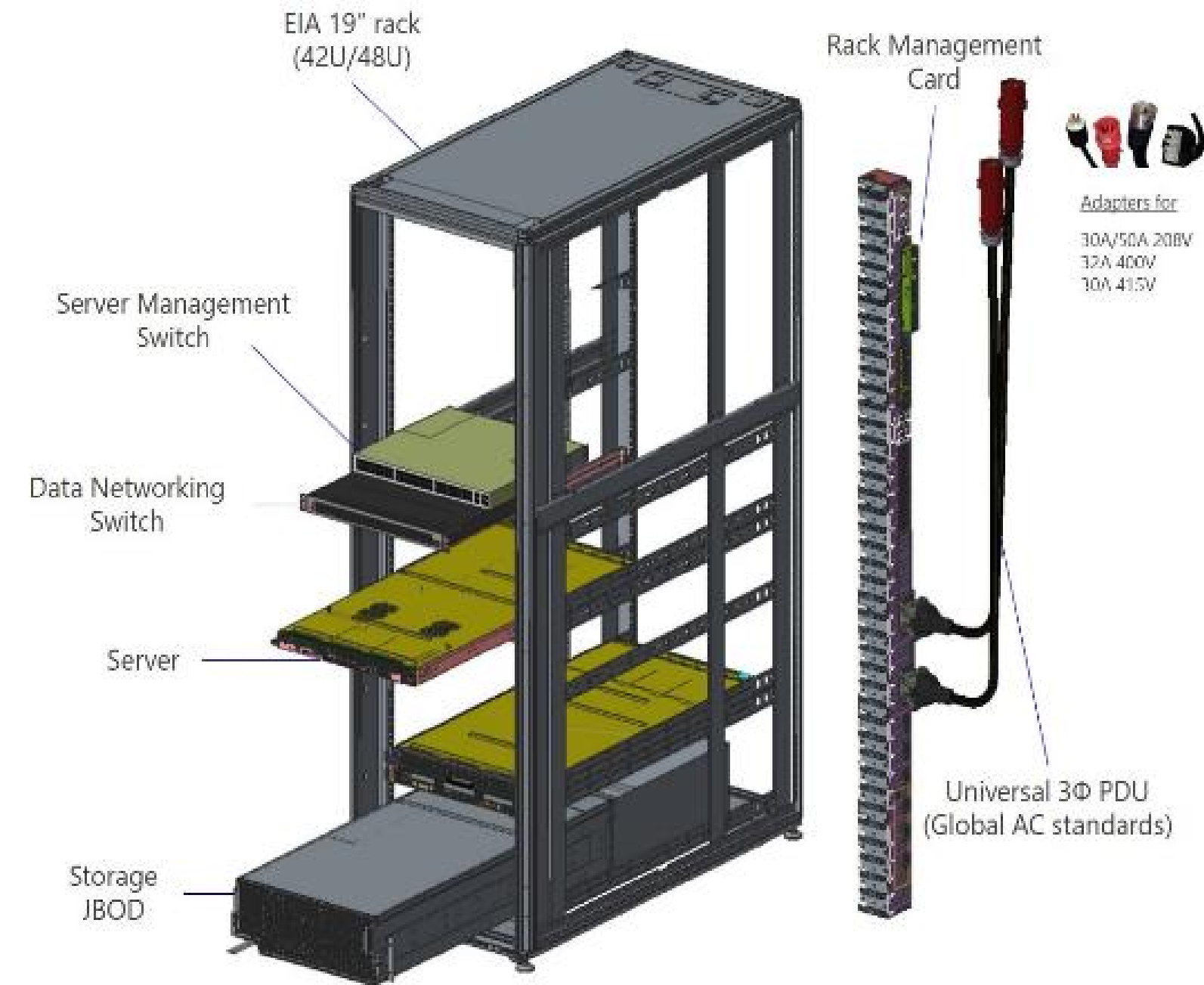


Outstanding flexibility, performance and value for GPU-intensive scale-out computing and Virtual Desktop Infrastructure applications

OPEN. FOR BUSINESS



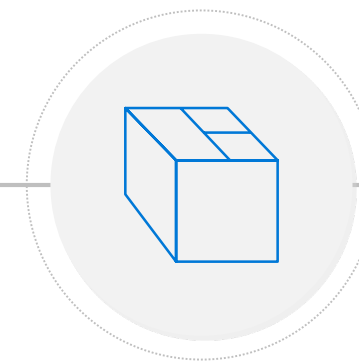
Server & rack design | Project Olympus



Universal
building blocks



High
power efficiency



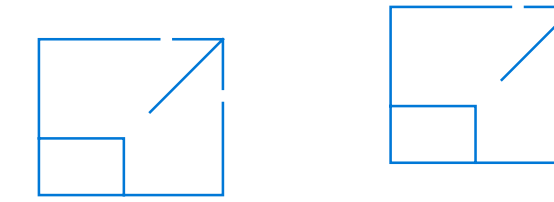
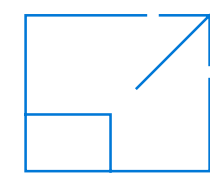
Global
datacenter standards



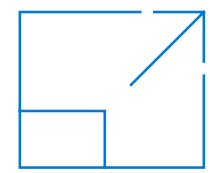
Open
source design

Project Olympus and AI

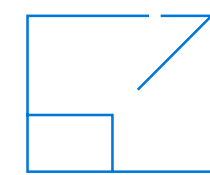
Anticipate AI Needs



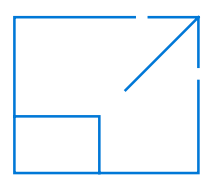
AI data parallelism => processes no longer decoupled!



Inter-chassis bandwidth and latency are key



Higher performance, but higher power density



Project Olympus supports higher density



Conclusions

- Hyperscale datacenters enable **massive scaling**
- Need a factor of 100x in AI computing, **1000x** in a few more years
- Power density, distribution and cooling are limiting factors
- Impact on systems, racks, datacenters
- Impact for on-prem and cloud
- Huge wave coming – better ride it!



OCP
SUMMIT

OPEN.



**FOR
BUSINESS.**

