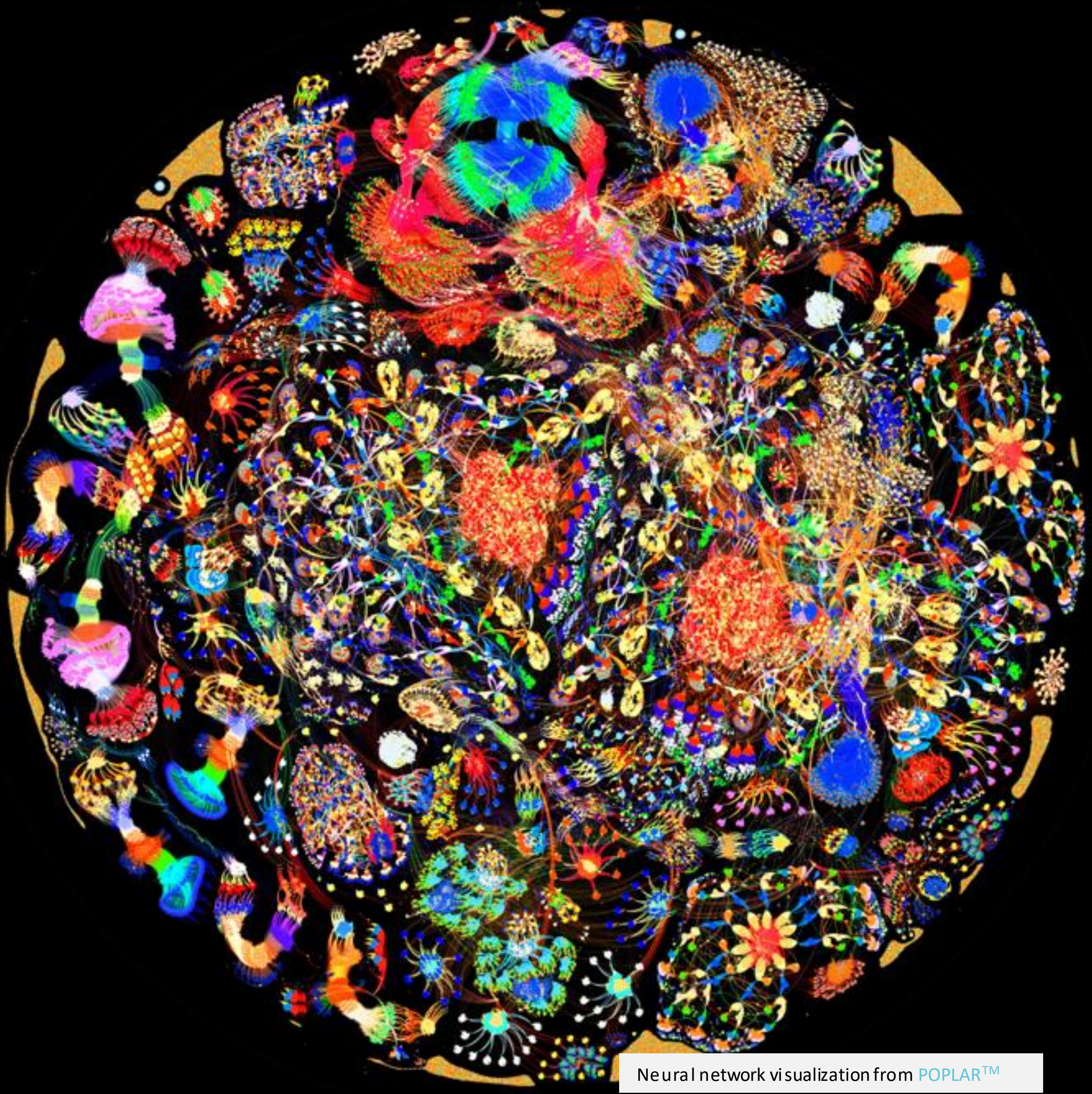
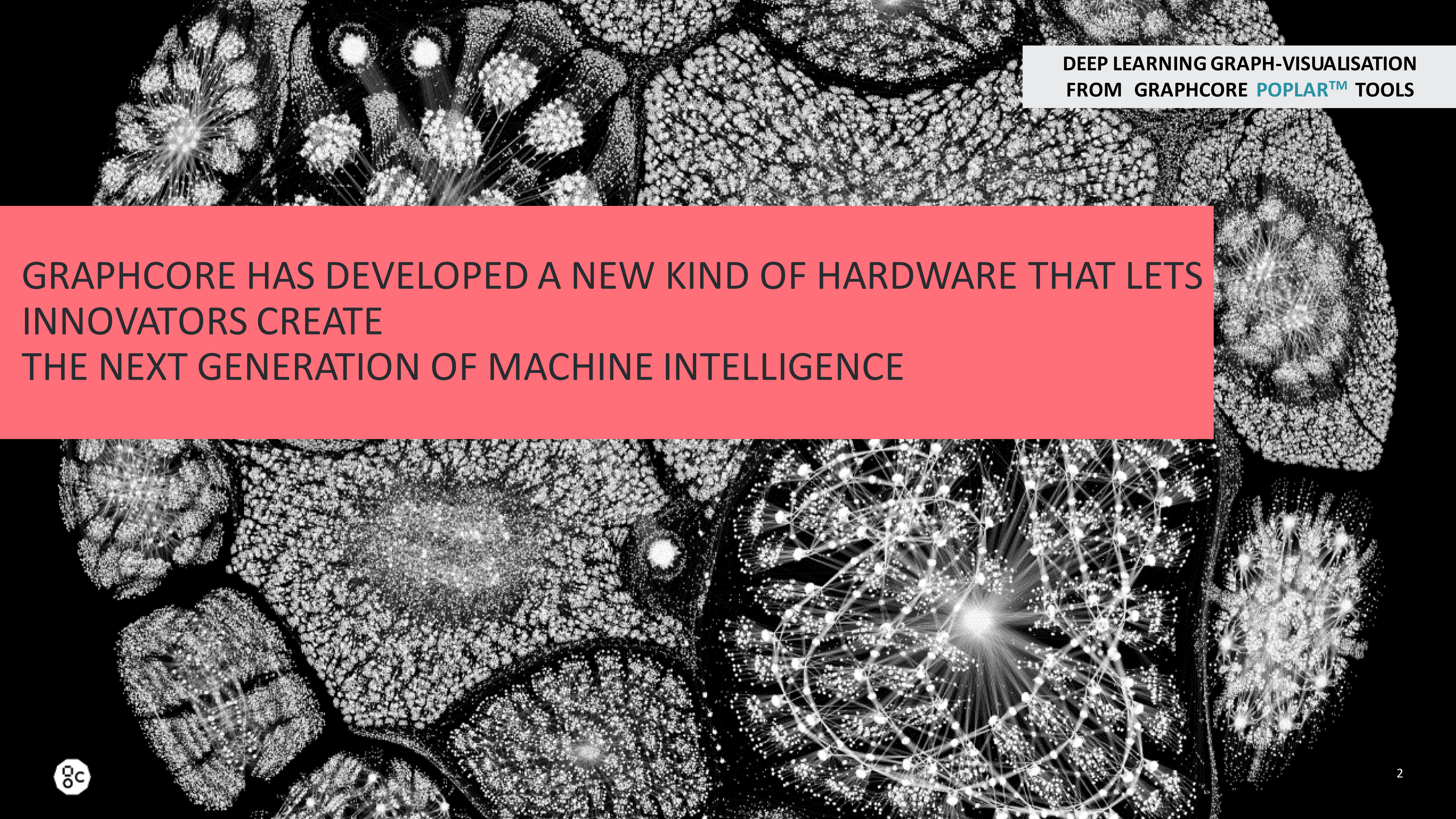
GRAPHCORE

Innovation in Machine Intelligence



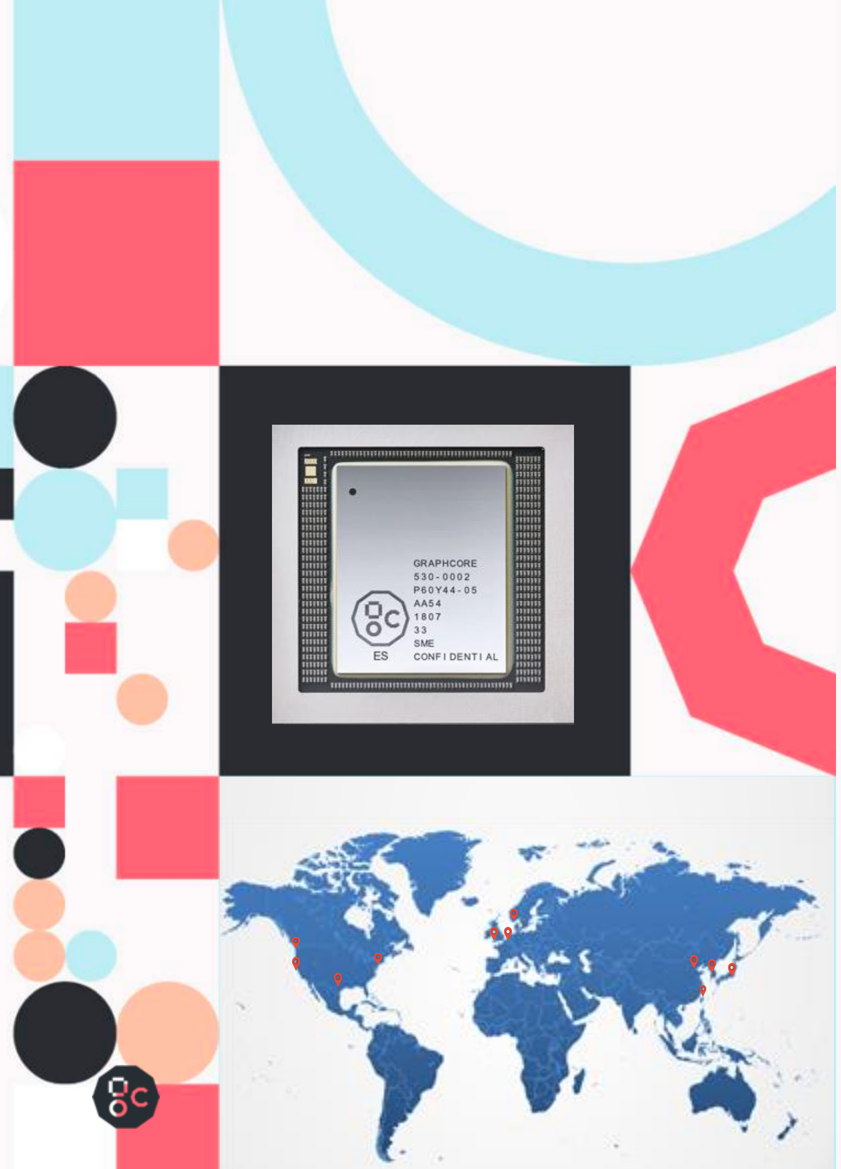
Neural network visualization from POPLAR™



DEEP LEARNING GRAPH-VISUALISATION FROM GRAPHCORE POPLAR™ TOOLS

GRAPHCORE HAS DEVELOPED A NEW KIND OF HARDWARE THAT LETS INNOVATORS CREATE THE NEXT GENERATION OF MACHINE INTELLIGENCE

GRAPHCORE ENABLING MACHINE INTELLIGENCE



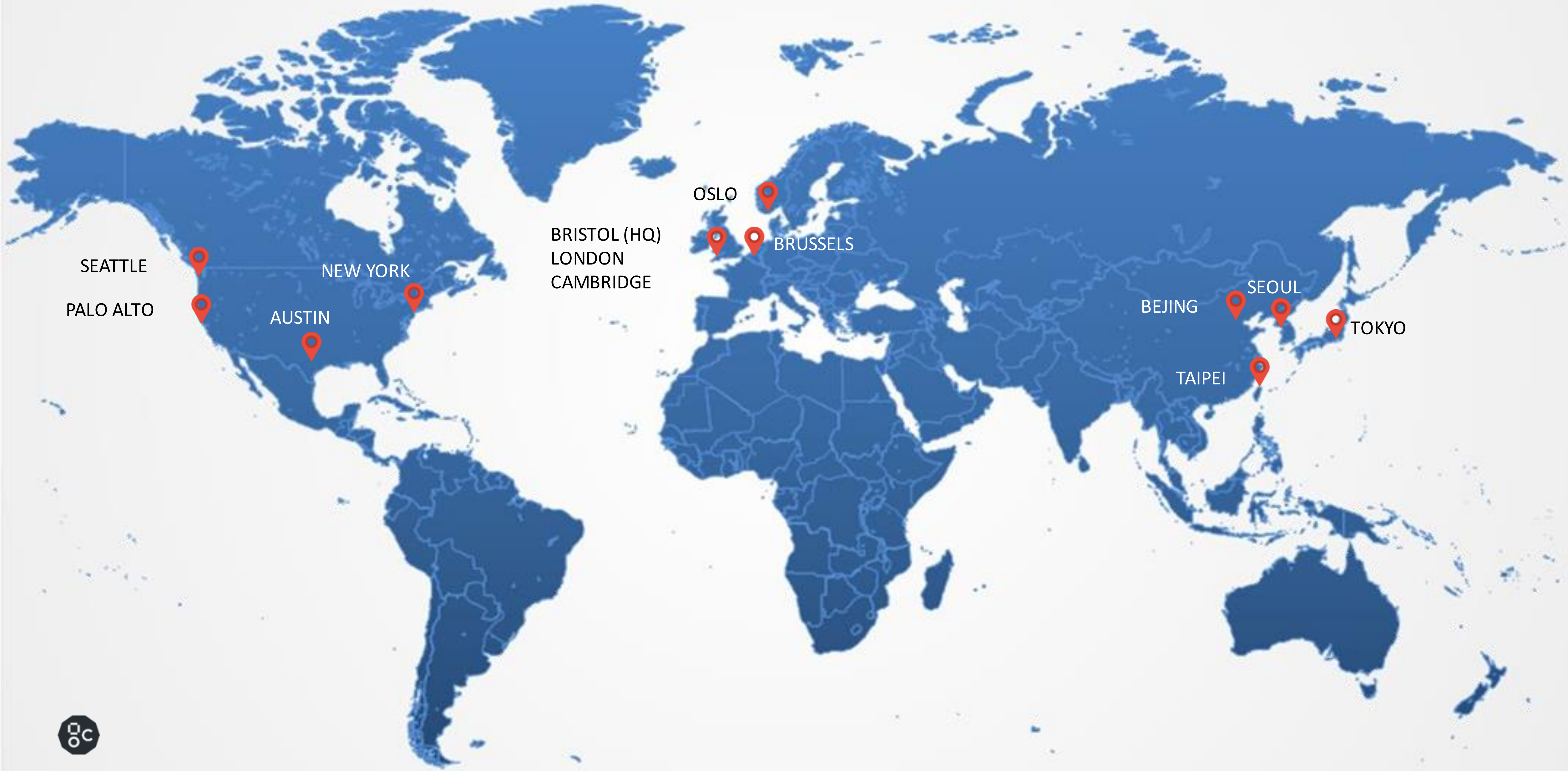
- Founded in 2016
- Technology: Intelligence Processor Unit (IPU)
- Team: approaching 400 globally
- Offices: UK, US, China, Norway
- Raised >$320M

SEQUOIA ATOMICO SOFINA Microsoft

BMW i VENTURES DELL BOSCH SAMSUNG Merian GLOBAL INVESTORS

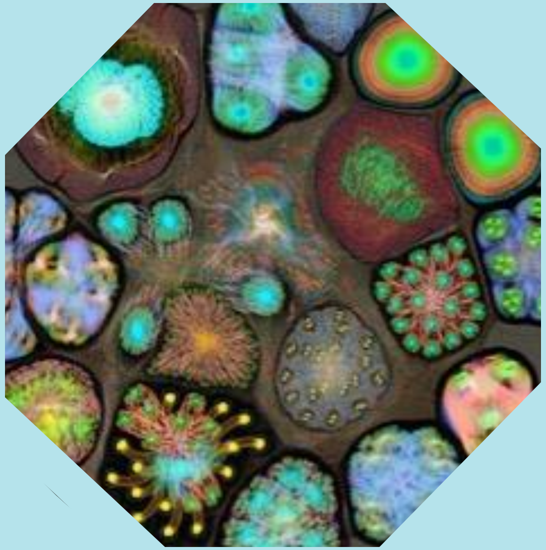
Amadeus Capital Partners pitango VENTURE CAPITAL draperesprit Foundation CAPITAL

GRAPHCORE GLOBAL FOOTPRINT



ABOUT US…

Technology

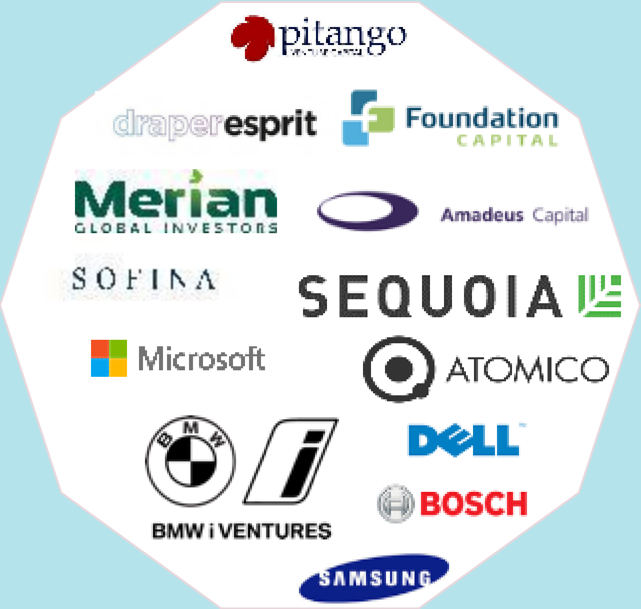


Processors and software solutions designed for AI

Products

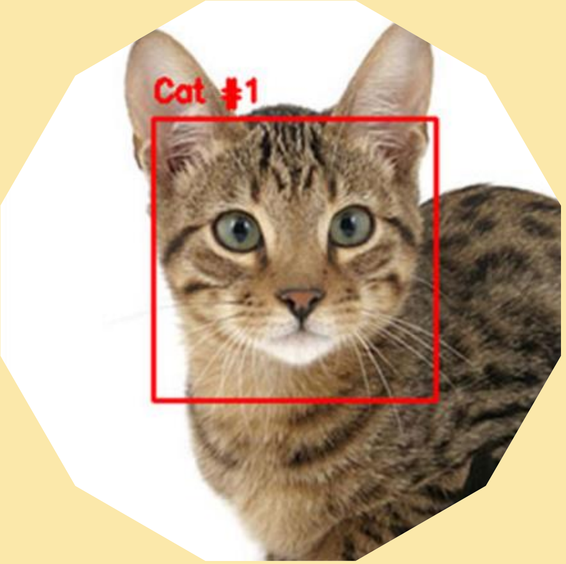
IPU-Processor PCIe Cards and Poplar® software stack

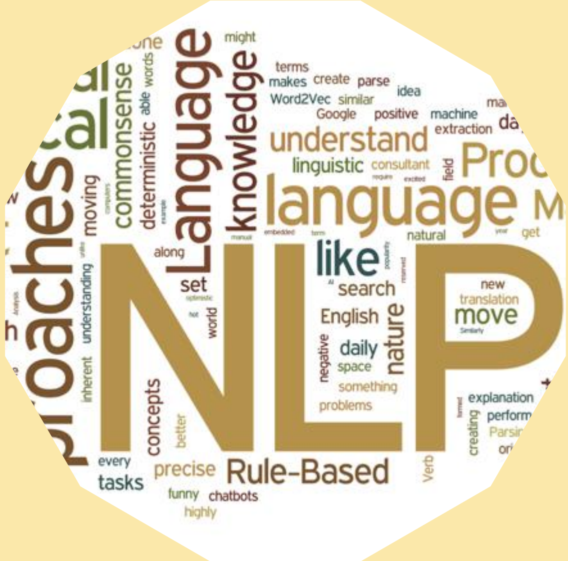
Investors



>$310m in funding

MACHINE INTELLIGENCE EVOLUTION





STEP 1

Simple perception

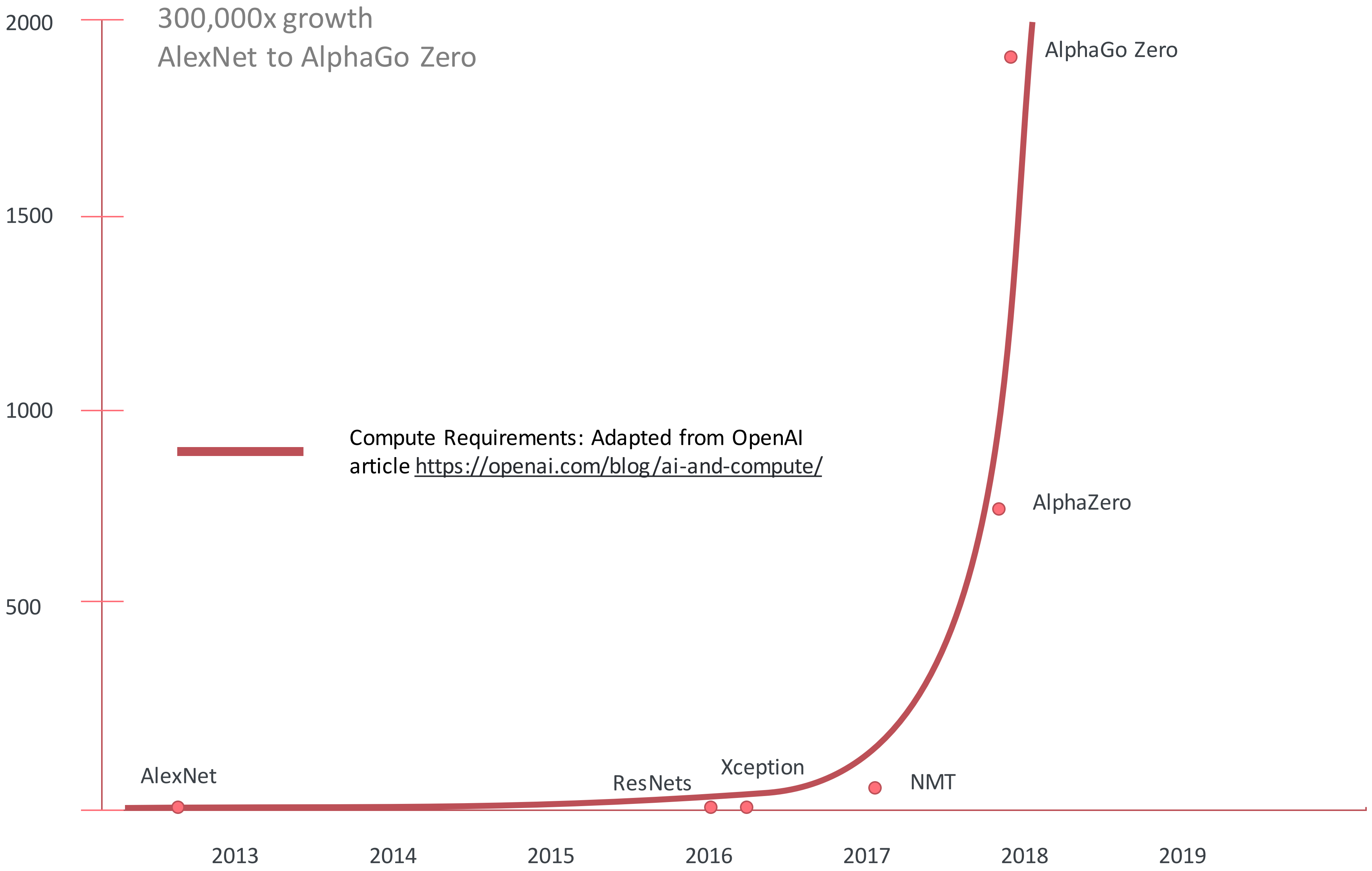
STEP 2

Language understanding

STEP 3

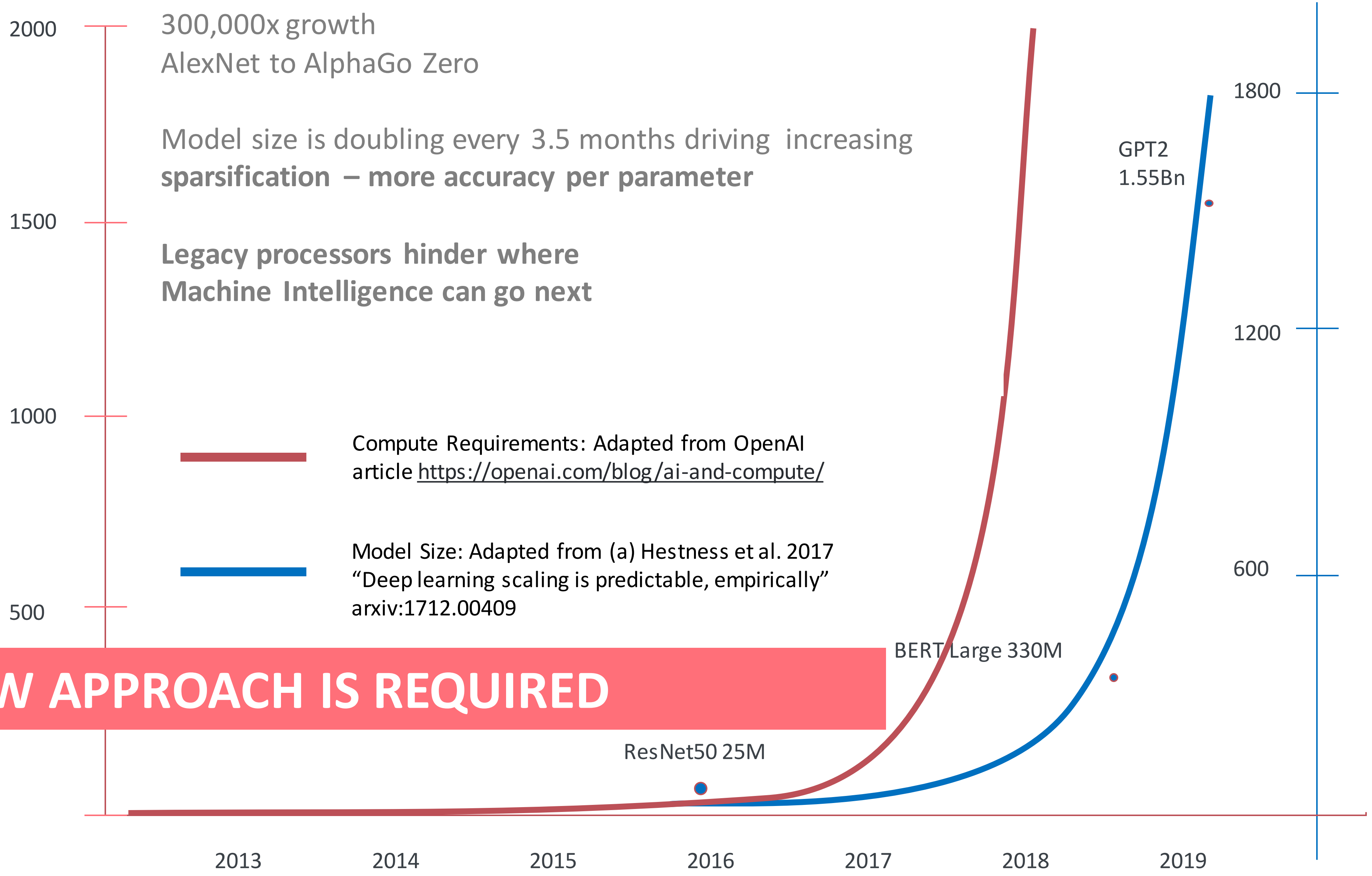
- Advanced perception
- Learning from experience

MACHINE INTELLIGENCE COMPUTE EXPONENTIAL…



Compute Requirements: Adapted from OpenAI article https://openai.com/blog/ai-and-compute/

MACHINE INTELLIGENCE COMPUTE EXPONENTIAL…



300,000x growth
AlexNet to AlphaGo Zero

Model size is doubling every 3.5 months driving increasing **sparsification – more accuracy per parameter**

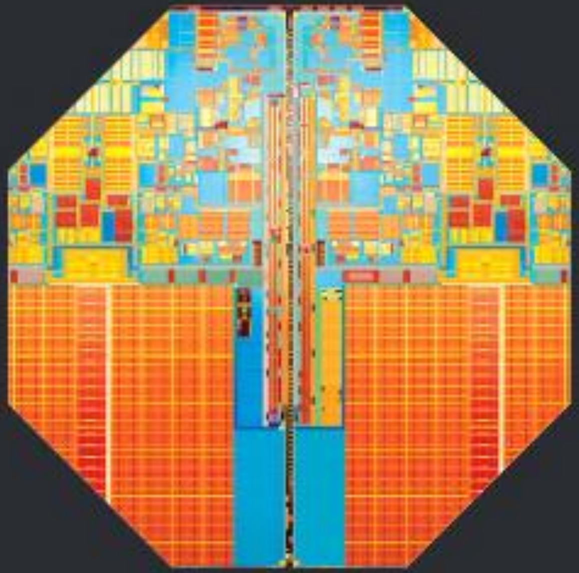
Legacy processors hinder where Machine Intelligence can go next

🟥 Compute Requirements: Adapted from OpenAI article https://openai.com/blog/ai-and-compute/

🟦 Model Size: Adapted from (a) Hestness et al. 2017 "Deep learning scaling is predictable, empirically" arxiv:1712.00409

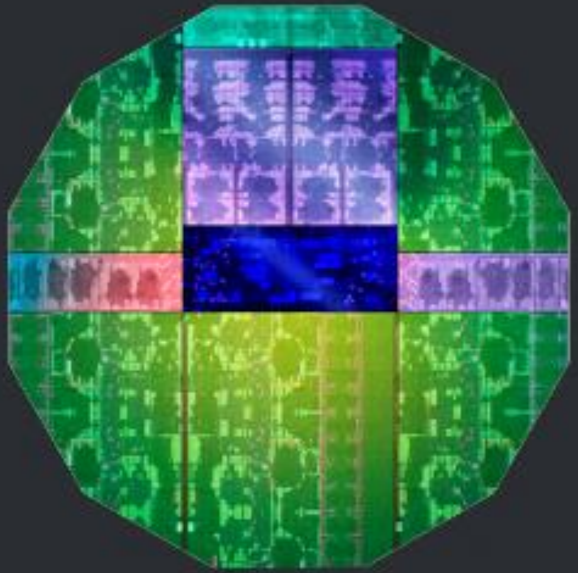
A NEW APPROACH IS REQUIRED

LEGACY PROCESSOR ARCHITECTURES HAVE BEEN REPURPOSED FOR ML



CPU

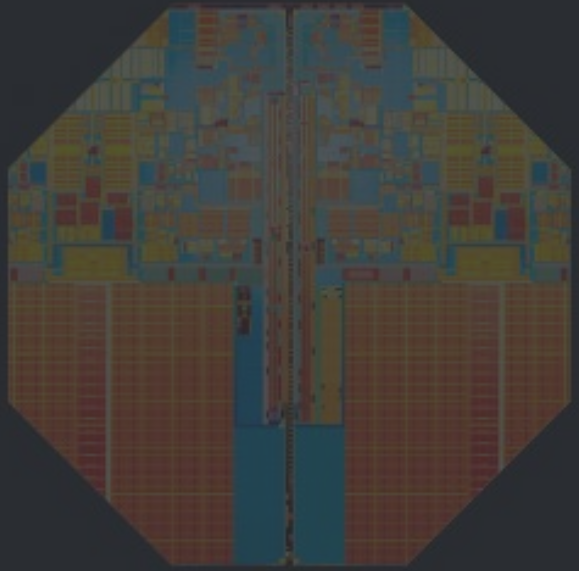
Apps and Web/
Scalar



GPU

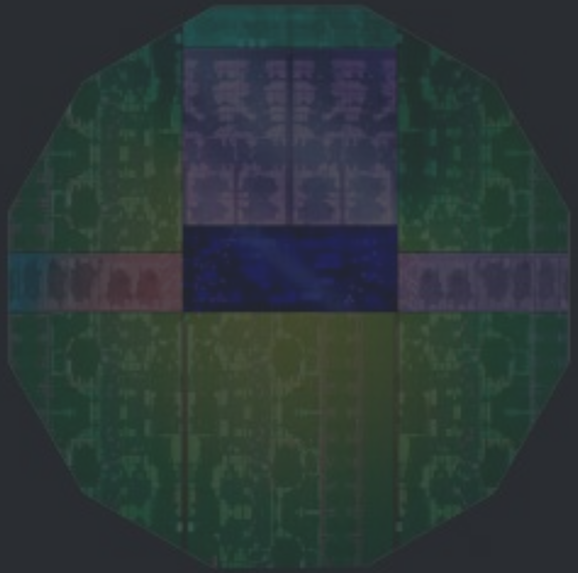
Graphics and HPC/
Vector

A NEW PROCESSOR IS REQUIRED FOR THE FUTURE



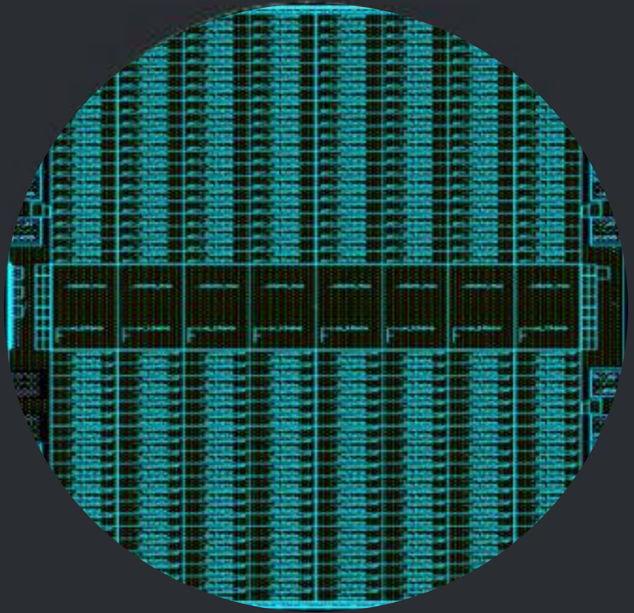
CPU

Apps and Web/
Scalar



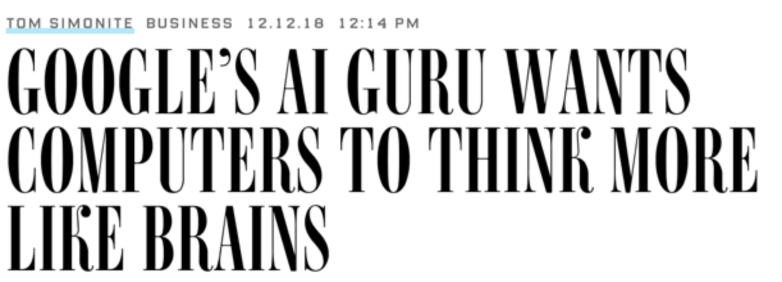
GPU

Graphics and HPC/
Vector



IPU

Artificial Intelligence/
Graph



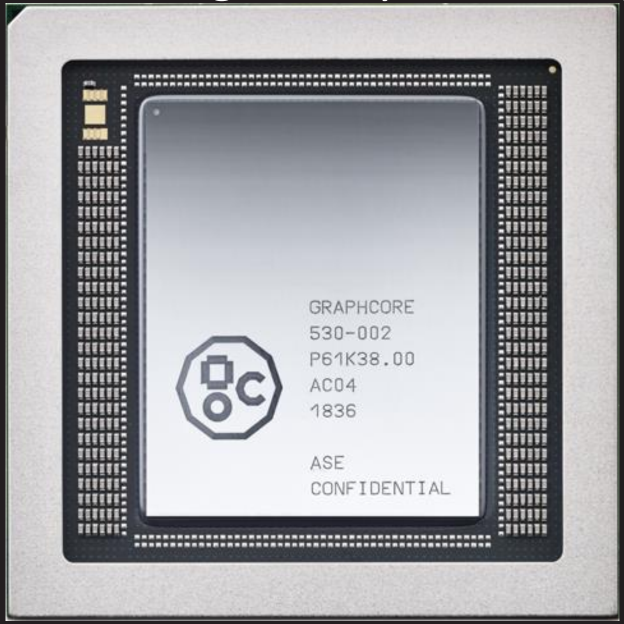
TOM SIMONITE BUSINESS 12.12.18 12:14 PM

GOOGLE'S AI GURU WANTS COMPUTERS TO THINK MORE LIKE BRAINS

WIRED

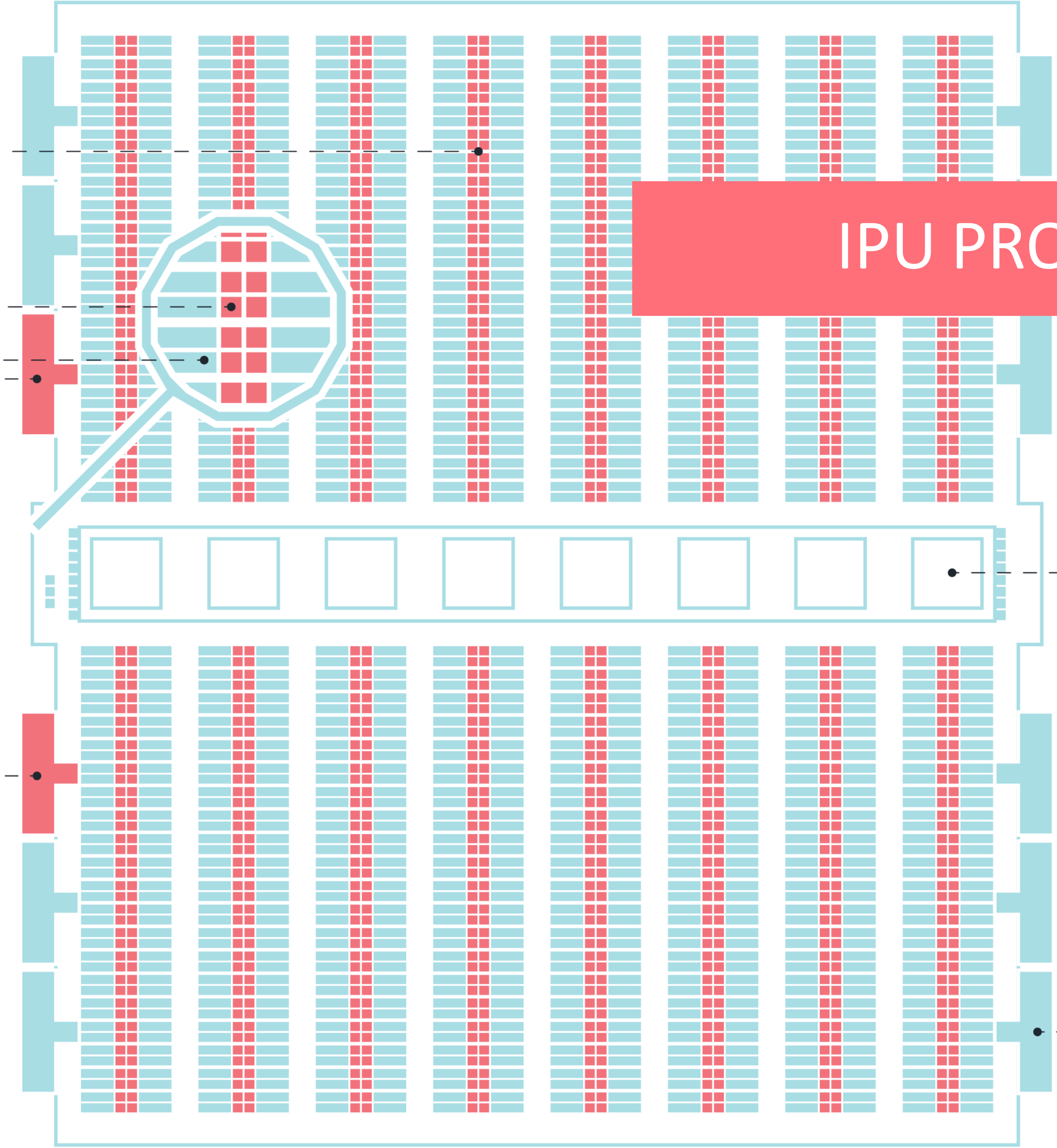
Wired – "How might we build machine learning systems that function more like a brain? "

Geoff Hinton – "I think we need to move towards a different type of computer. Fortunately I have one here…" Hinton reaches into his wallet and pulls out a large, shiny silicon chip:



an IPU processor from Graphcore

IPU PROCESSOR



IPU-Tiles™

1216 IPU-Tiles™ each with an independent IPU-Core™ and tightly coupled In-Processor-Memory™

IPU-Core™

1216 IPU-Cores™ with 7296 programs executing in parallel

In-Processor-Memory™

300MB In-Processor-Memory™
45TB/s memory bandwidth
Whole model held on-chip

PCIe

PCI Gen4 x16
64 GB/s bidirectional bandwidth to host

IPU-Exchange™

8 TB/s all to all IPU-Exchange™
Non-blocking, any communication pattern

IPU-Links™

80 IPU-Links, 320GB/s chip to chip bandwidth

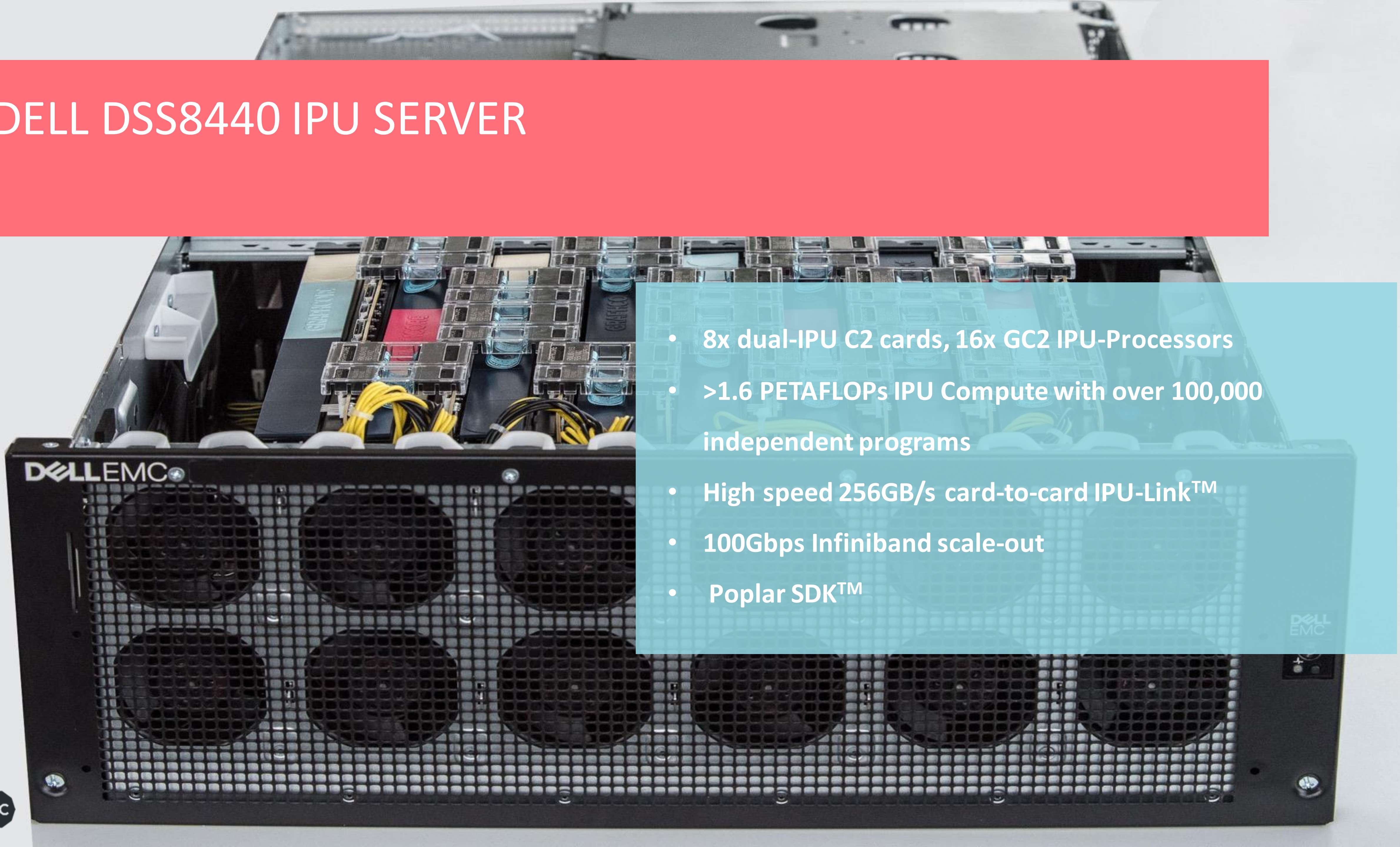
C2 IPU PROCESSOR CARD

2 – COLOSSUS **GC2** IPU PROCESSORS

CARD-TO-CARD **IPU-LINKS™** *(2.5TBps)*

200 TERA-FLOP MIXED PRECISION IPU COMPUTE @ 315W

DELL DSS8440 IPU SERVER

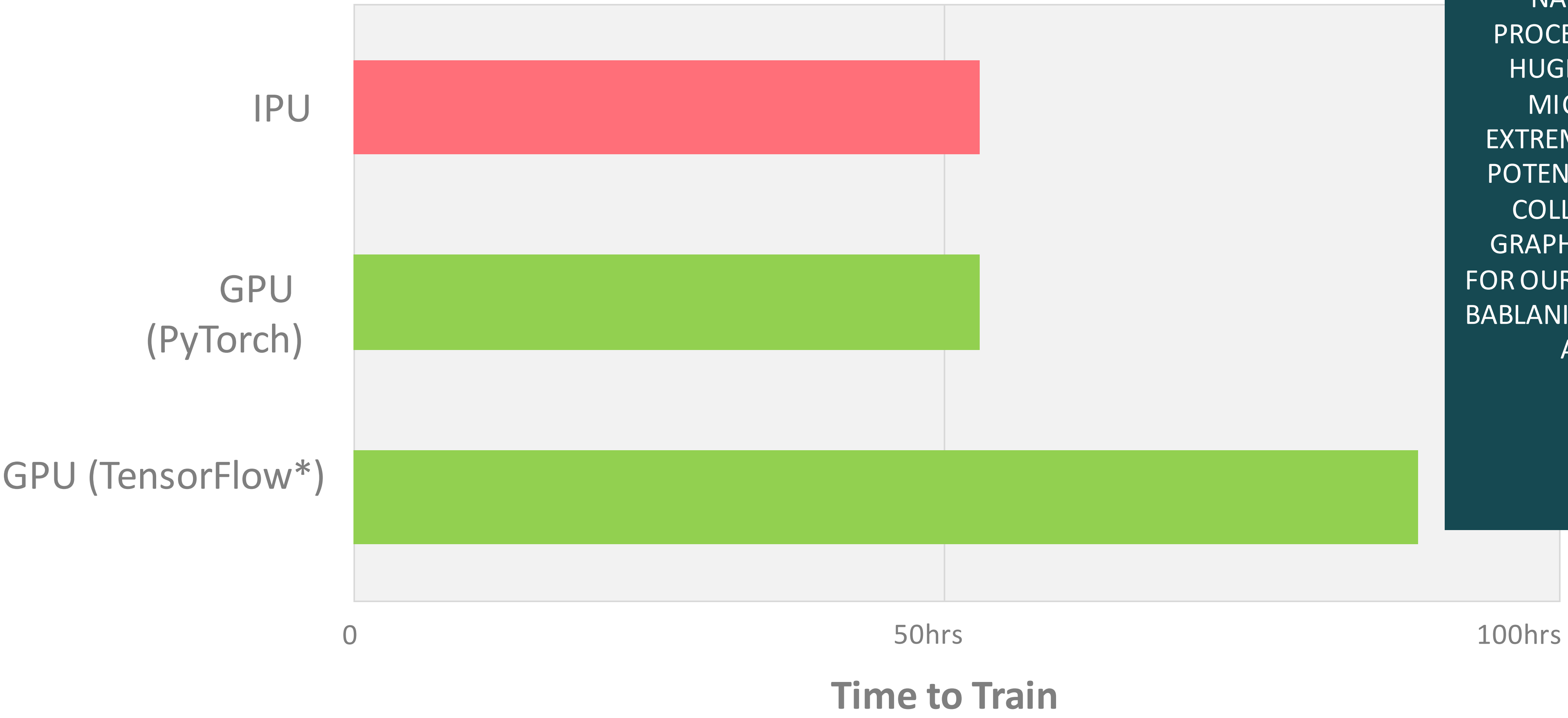


- 8x dual-IPU C2 cards, 16x GC2 IPU-Processors
- >1.6 PETAFLOPs IPU Compute with over 100,000 independent programs
- High speed 256GB/s card-to-card IPU-Link™
- 100Gbps Infiniband scale-out
- Poplar SDK™

IPU ACHIEVES STATE OF THE ART PERFORMANCE ON TODAYS LEADING EDGE MODELS…

BERT-BASE : TRAINING

State of the art time to train: 56 hours on IPU @ 20% lower power

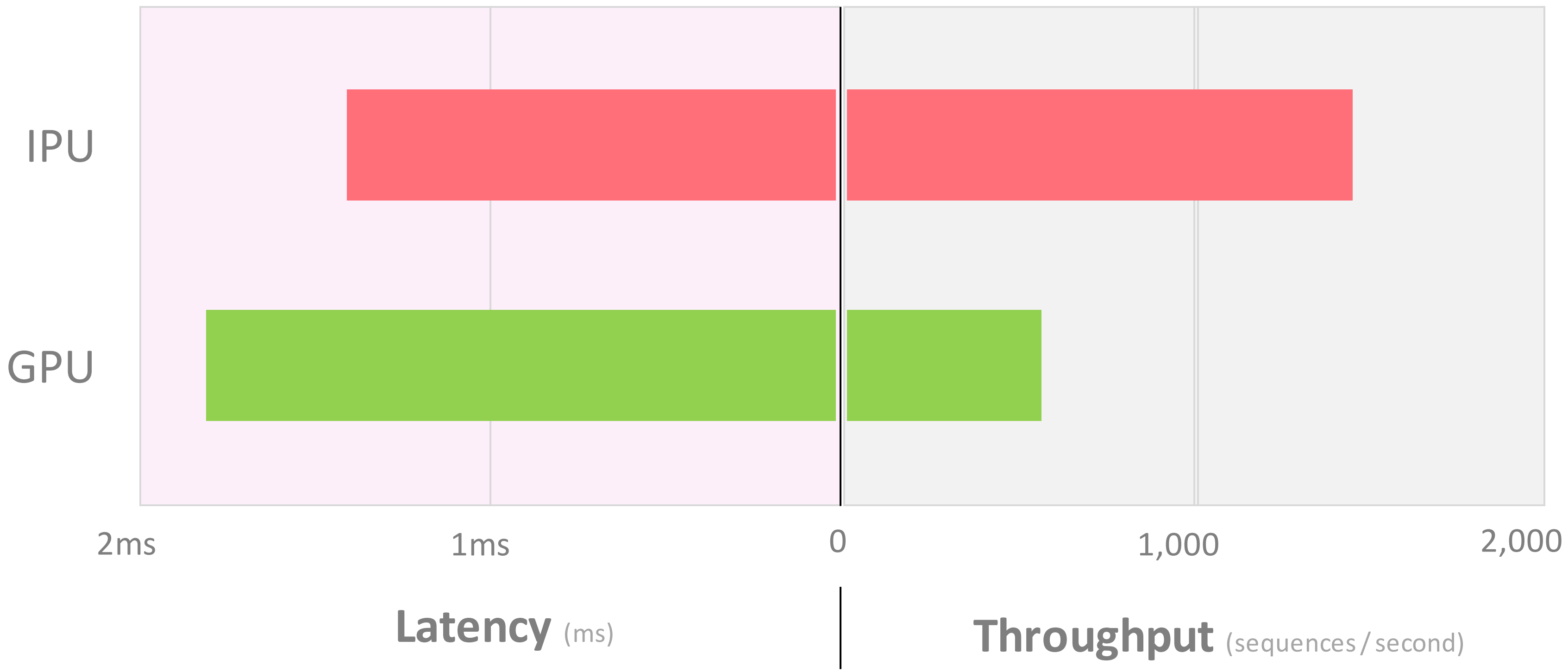


“NATURAL LANGUAGE PROCESSING MODELS ARE HUGELY IMPORTANT TO MICROSOFT. WE ARE EXTREMELY EXCITED BY THE POTENTIAL THAT THIS NEW COLLABORATION WITH GRAPHCORE WILL DELIVER FOR OUR CUSTOMERS,” GIRISH BABLANI, VP AZURE COMPUTE AT MICROSOFT

NOTES:
BERT-Base | Wikipedia dataset + SQuAD 1.1 (EM)
IPU: DSS8440, 7x Graphcore C2 – customer implementation using Poplar
GPU: 8x Leading GPU system using PyTorch and TensorFlow (*estimated)

BERT-BASE : INFERENCE

3x higher throughput at 30% lower latency

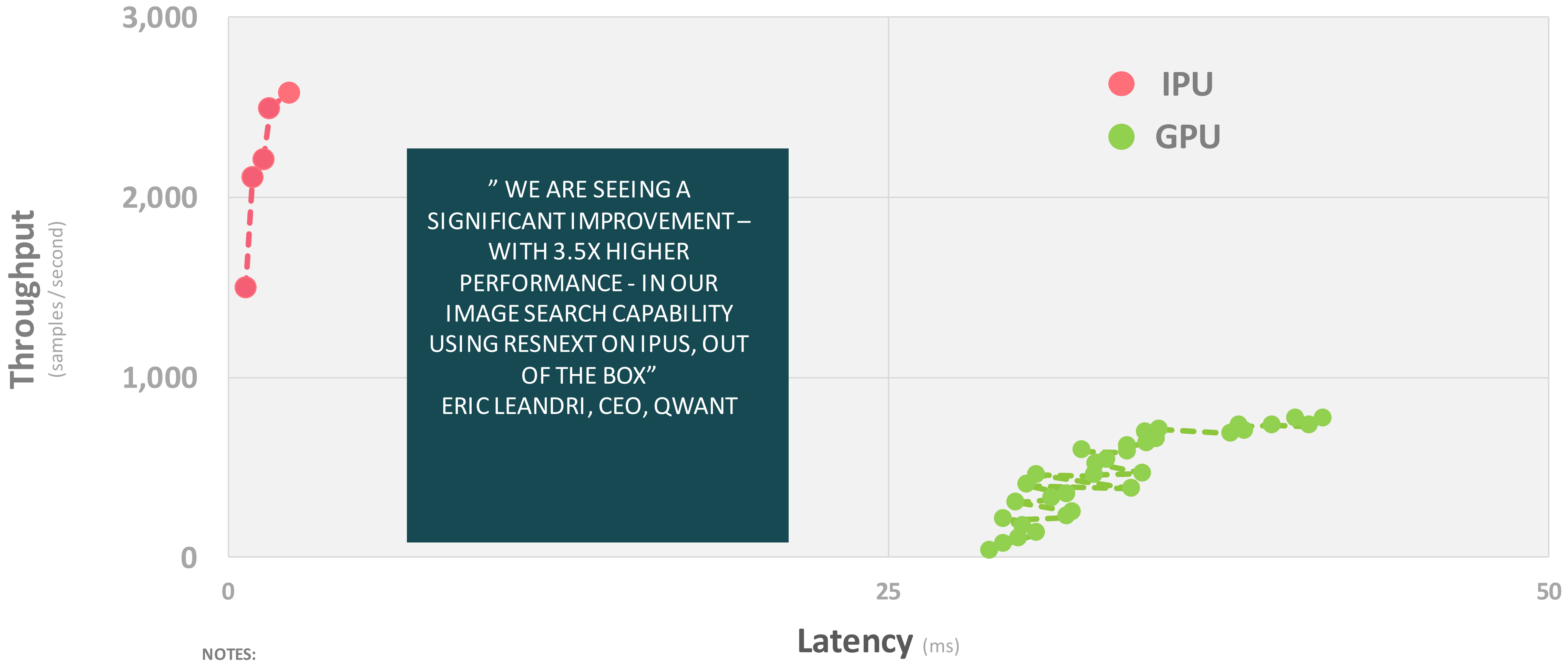


NOTES:
Graphcore results on one C2 Card using two IPUs, on SQuAD v1.1 data,
Graphcore C2 customer implementation using Poplar @ 300W TDP
NVIDIA results for 1xV100 with TensorRT 6.0 using SQuAD v1.1 data, published 6 November 2019
https://developer.nvidia.com/deep-learning-performance-training-inference.

RESNEXT-101 : INFERENCE

Lowest Latency Comparison: 43x higher throughput | 40x lower latency

Highest Throughput Comparison: 3.4x higher throughput | 18x lower latency



" WE ARE SEEING A SIGNIFICANT IMPROVEMENT – WITH 3.5X HIGHER PERFORMANCE - IN OUR IMAGE SEARCH CAPABILITY USING RESNEXT ON IPUS, OUT OF THE BOX"
ERIC LEANDRI, CEO, QWANT

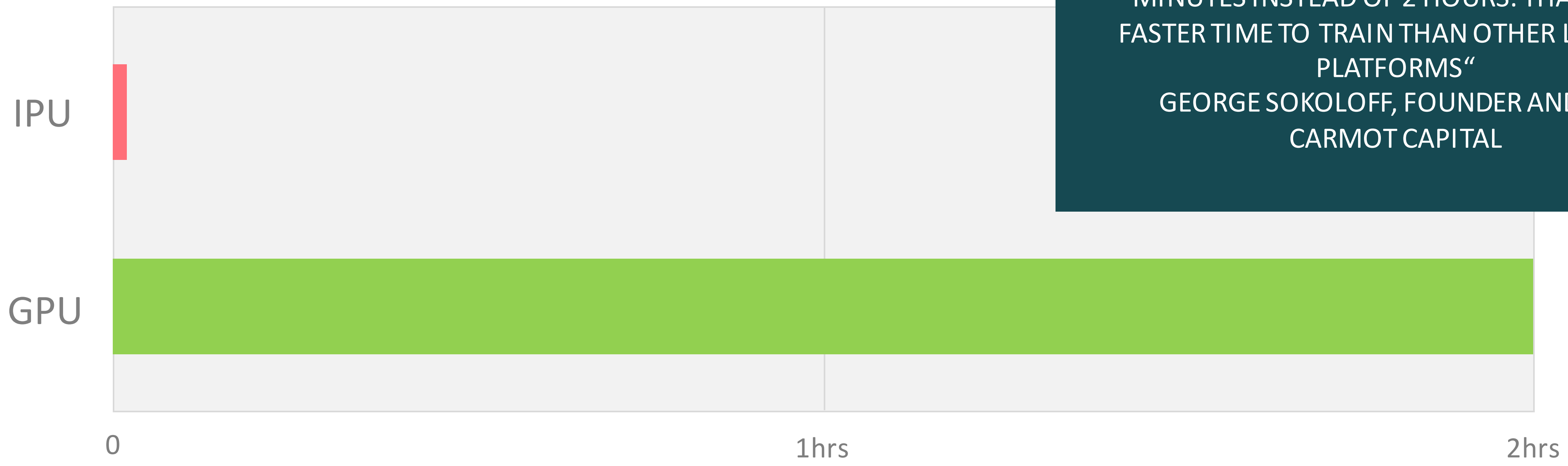
NOTES:
ResNext-101_32x4d | Real data (COCO)
IPU: Graphcore C2 (SDK 1.0.49) using ONNX/PopART (Batch Size 2-12) @ 300W TDP
GPU using Pytorch FP16 (Batch Size 1-32) @ 300W TDP

IPU DELIVERS MASSIVE PERFORMANCE ADVANTAGE ON DIFFICULT MACHINE LEARNING PROBLEMS

MCMC PROBABILISTIC MODEL : TRAINING

Customer implementation

26x higher throughput



Time to Result

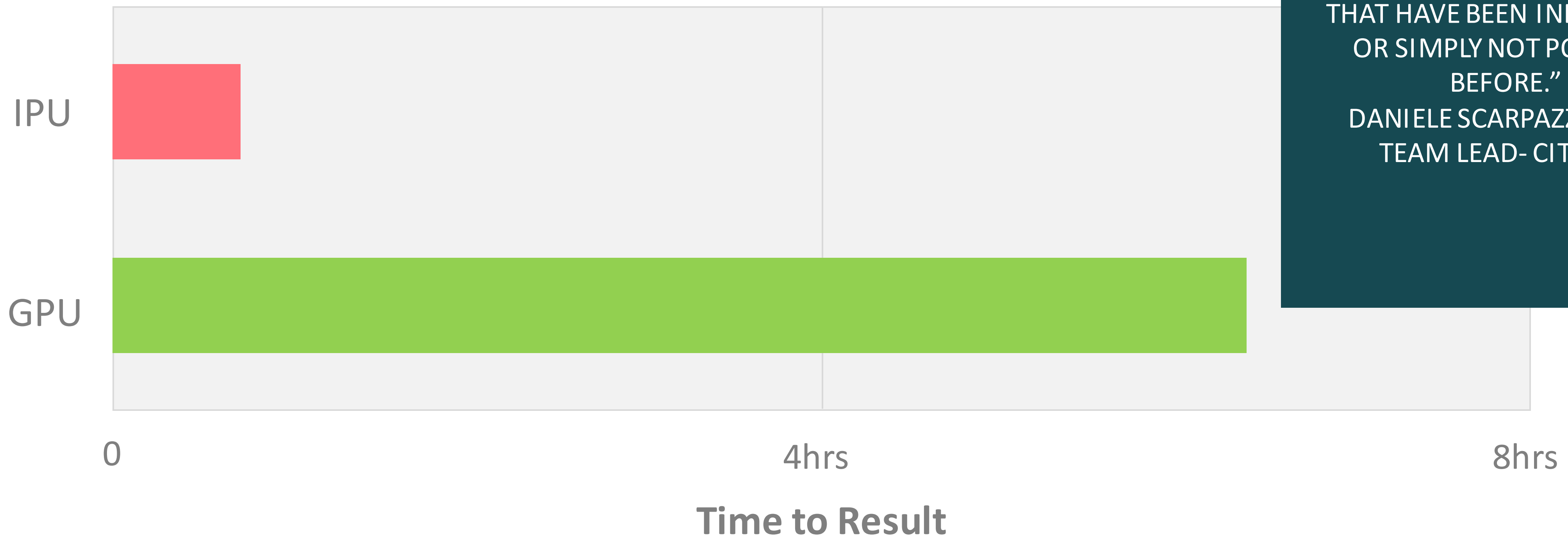
"WE WERE ABLE TO TRAIN ONE OF OUR PROPRIETARY PROBABILISTIC MODELS IN 4.5 MINUTES INSTEAD OF 2 HOURS. THAT'S 26X FASTER TIME TO TRAIN THAN OTHER LEADING PLATFORMS"
GEORGE SOKOLOFF, FOUNDER AND CIO
CARMOT CAPITAL

NOTES:
Graphcore customer Markov Chain Monte Carlo Probability model (summary data shared with customer's permission)
IPU: Graphcore GC2 @ 150W TDP
GPU @ 300W TDP

MCMC PROBABILISTIC MODEL : TRAINING

TensorFlow probability model example

8x faster time to train



> “THE GRAPCORE IPU IS ALREADY ENABLING US TO EXPLORE NEW TECHNIQUES THAT HAVE BEEN INEFFICIENT OR SIMPLY NOT POSSIBLE BEFORE.”
> DANIELE SCARPAZZA, R&D TEAM LEAD- CITADEL

NOTES:

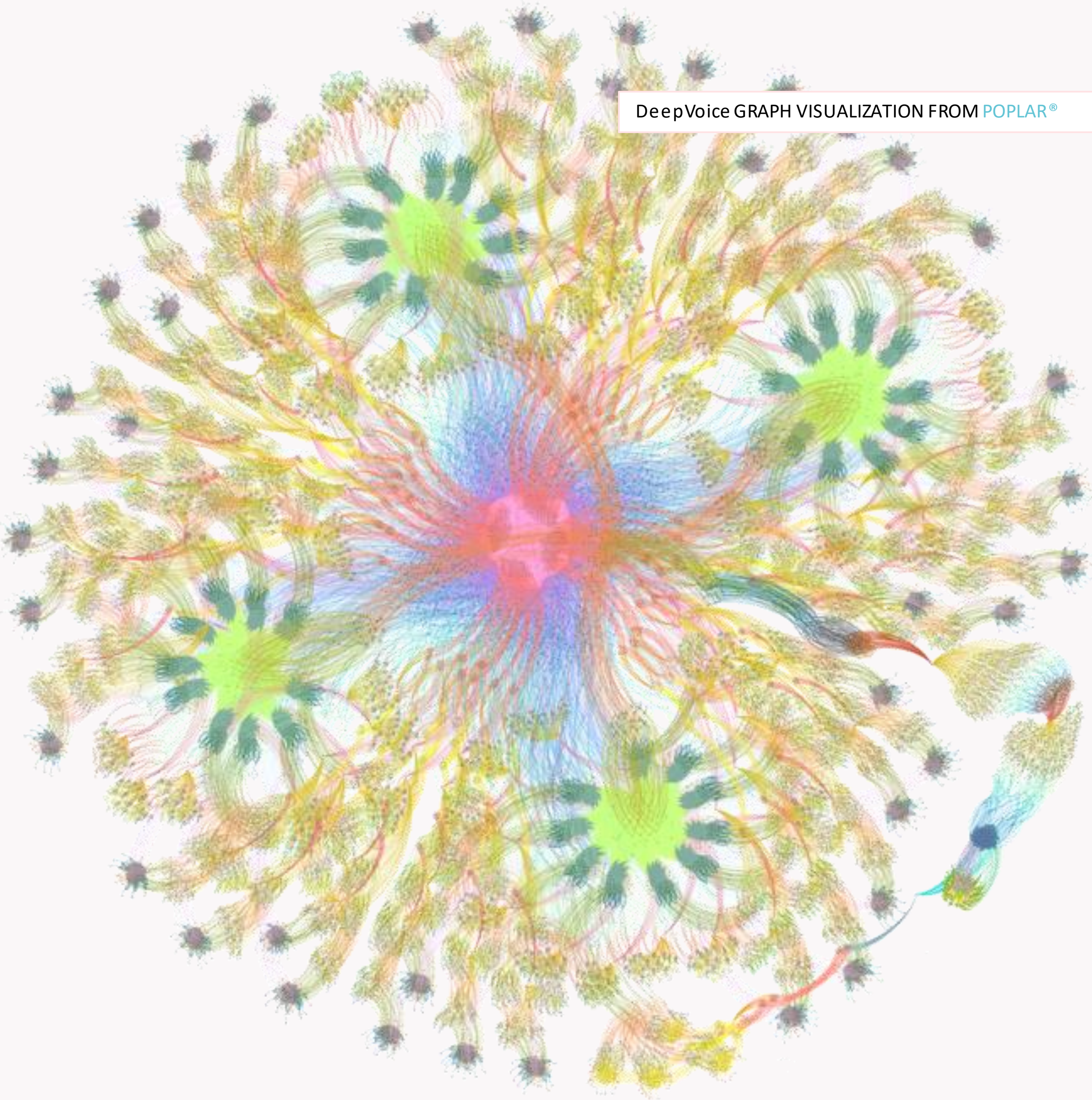
Markov Chain Monte Carlo – Probabilistic model example with TensorFlow Probability, a neural network with 3 fully-connected layers

IPU: Graphcore GC2 @ 150W TDP

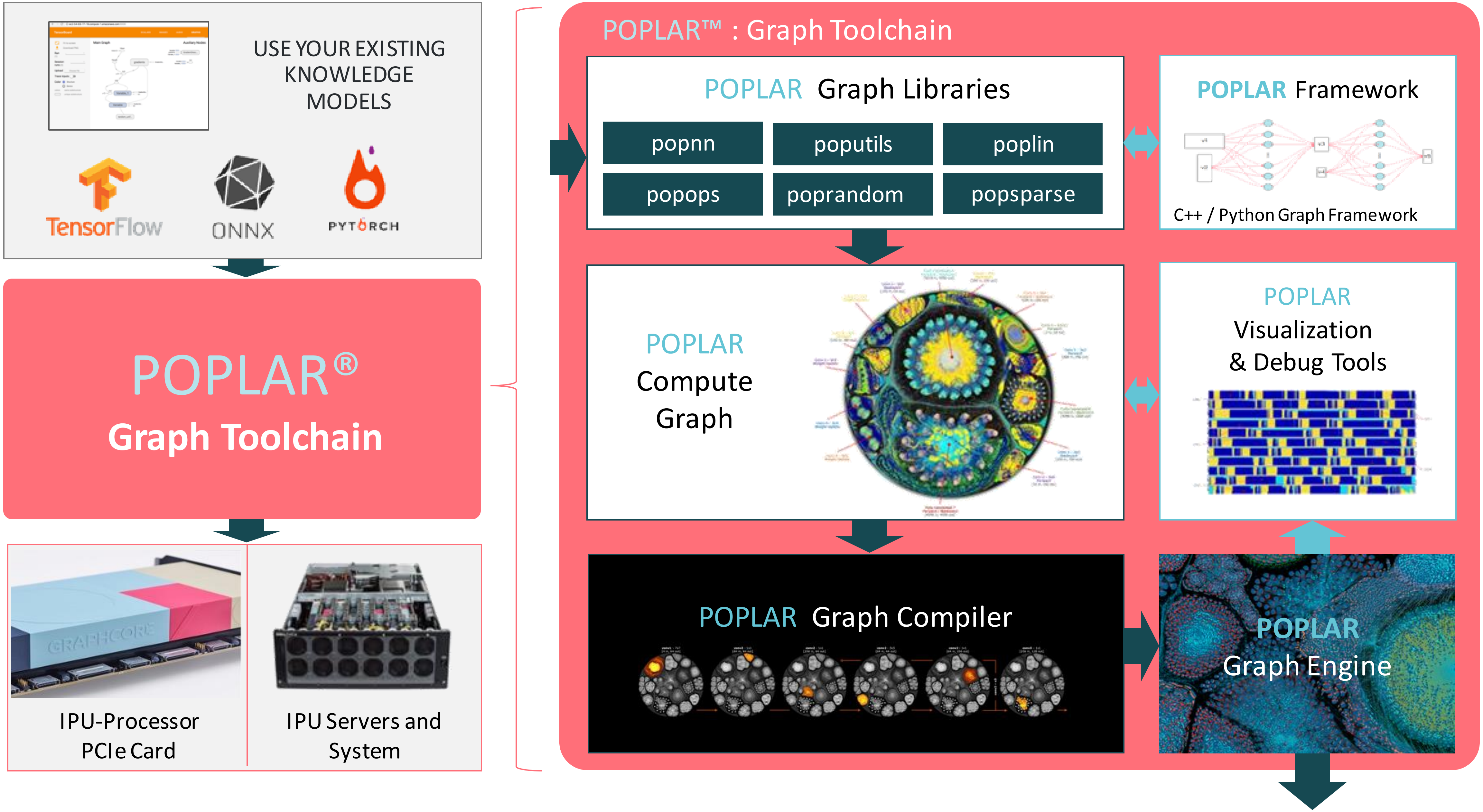
GPU @ 300W TDP

POPLAR®

expands the ML Framework output to a full compute graph.



DeepVoice GRAPH VISUALIZATION FROM POPLAR®



USE YOUR EXISTING KNOWLEDGE MODELS

TensorFlow

ONNX

PYTORCH

POPLAR®

Graph Toolchain

IPU-Processor PCIe Card

IPU Servers and System

POPLAR™ : Graph Toolchain

POPLAR Graph Libraries

popnn	poputils	poplin
popops	poprandom	popsparse

POPLAR Framework

C++ / Python Graph Framework

POPLAR Compute Graph

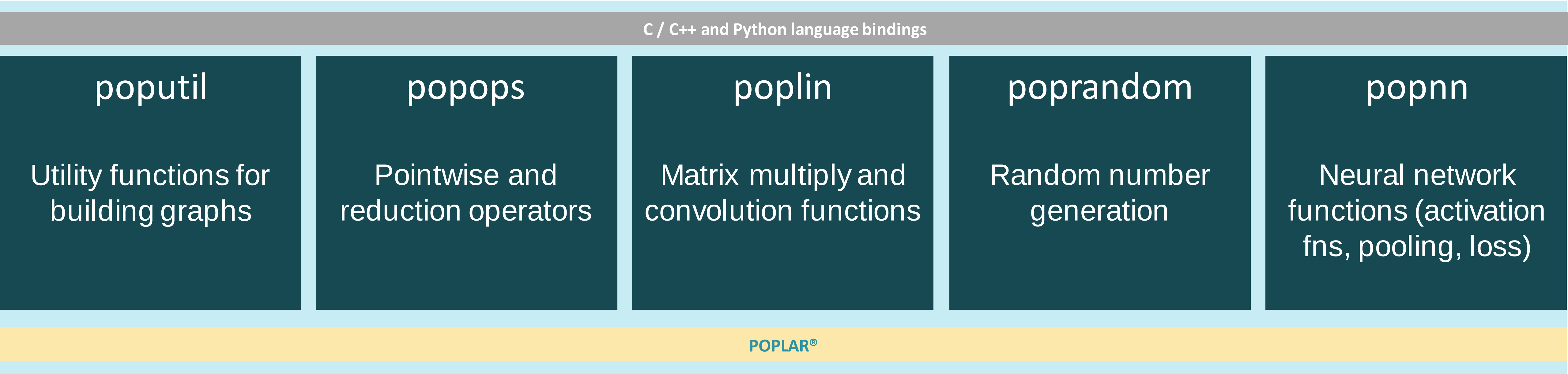
POPLAR Visualization & Debug Tools

POPLAR Graph Compiler

POPLAR Graph Engine

POPLIBS^{TM}

Highly optimized ***open source*** libraries partition work and data efficiently across IPU devices



C / C++ and Python language bindings

poputil

Utility functions for building graphs

popops

Pointwise and reduction operators

poplin

Matrix multiply and convolution functions

poprandom

Random number generation

popnn

Neural network functions (activation fns, pooling, loss)

POPLAR®

GitHub

github.com/graphcore/poplibs

BULK SYNCHRONOUS PARALLEL (BSP)

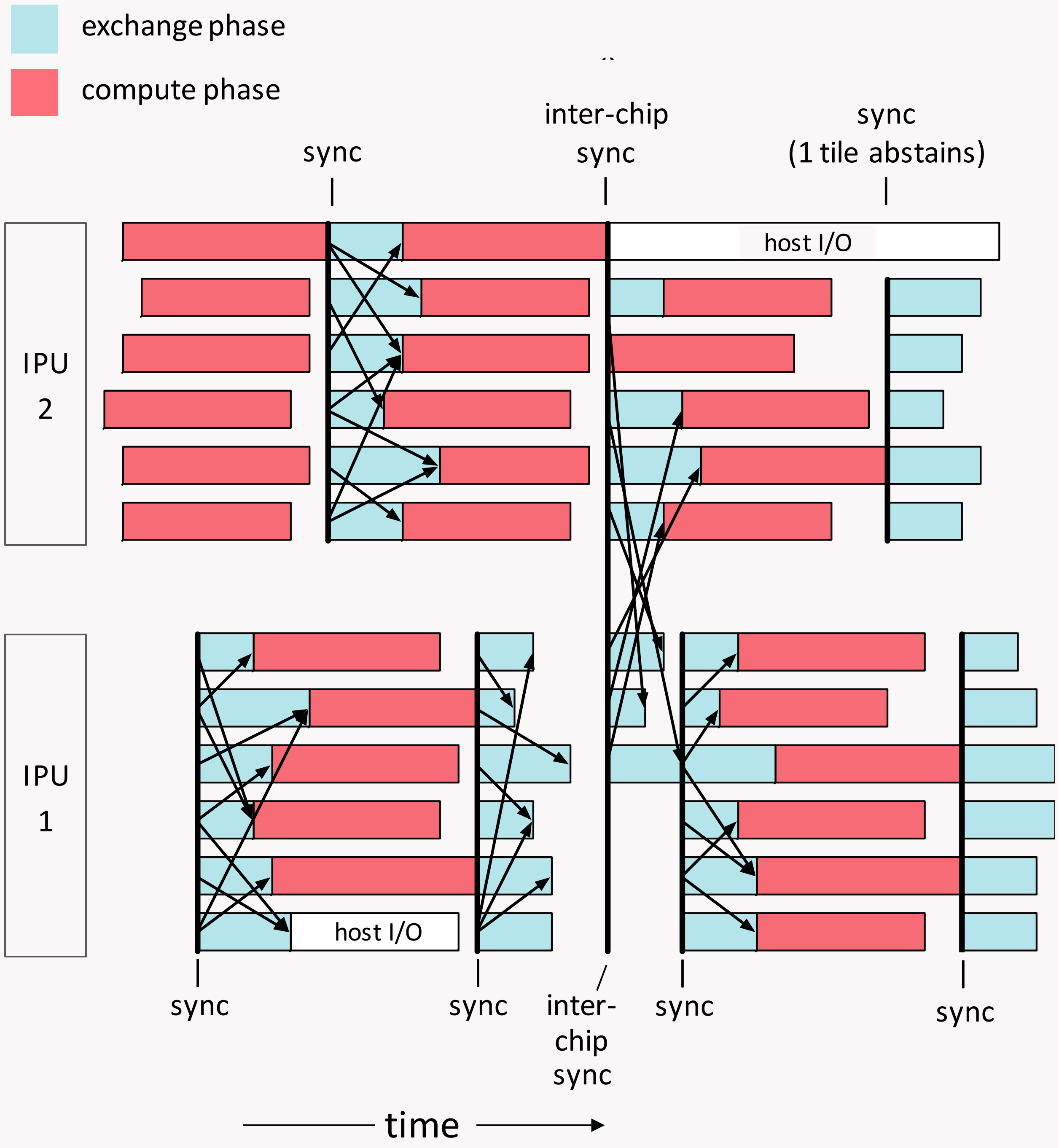
BSP software bridging model – massively parallel computing with no concurrency hazards

3 phases: compute, sync, exchange

Easy to program – no live-locks or dead-locks

Widely-used in parallel computing – Google, FB, …

First use of BSP inside a parallel processor



BULK SYNCHRONOUS PARALLEL (BSP)

Software bridging model for parallel computing

Compute

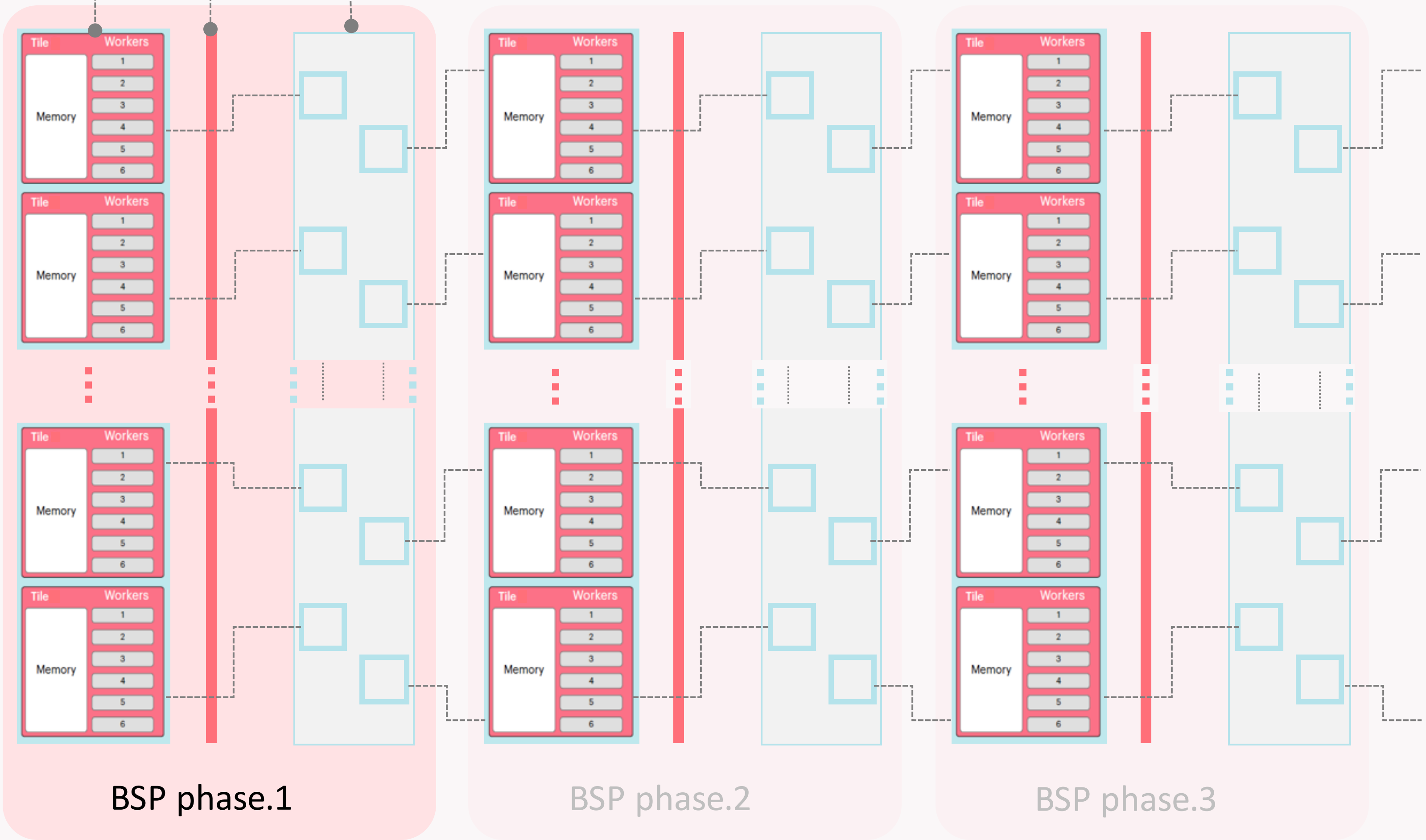
10,000s of compute threads
all operating in parallel
each with all the data that
they need, held locally

BSP Sync

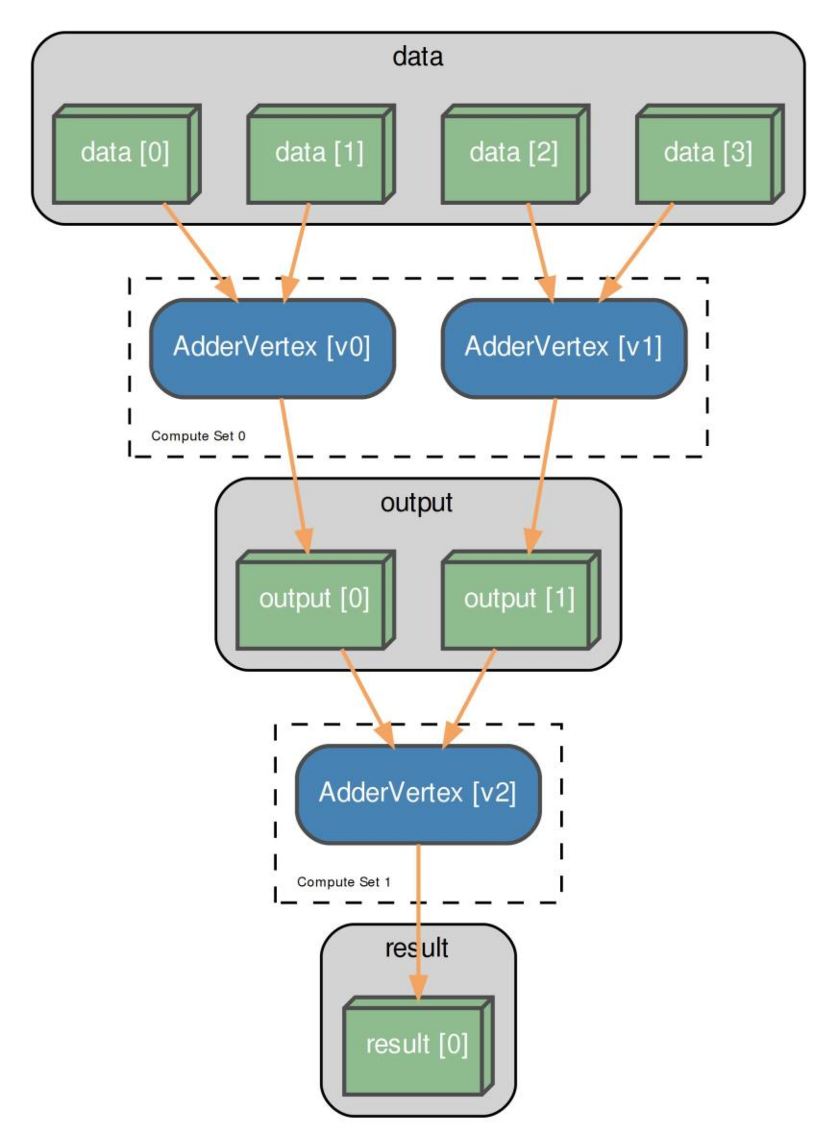
All threads are
synchronized

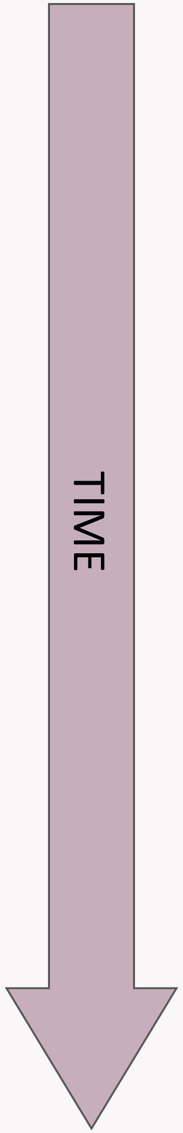
Exchange

Data is exchanged so that every thread
has all the data that it needs for the
next phase of Compute

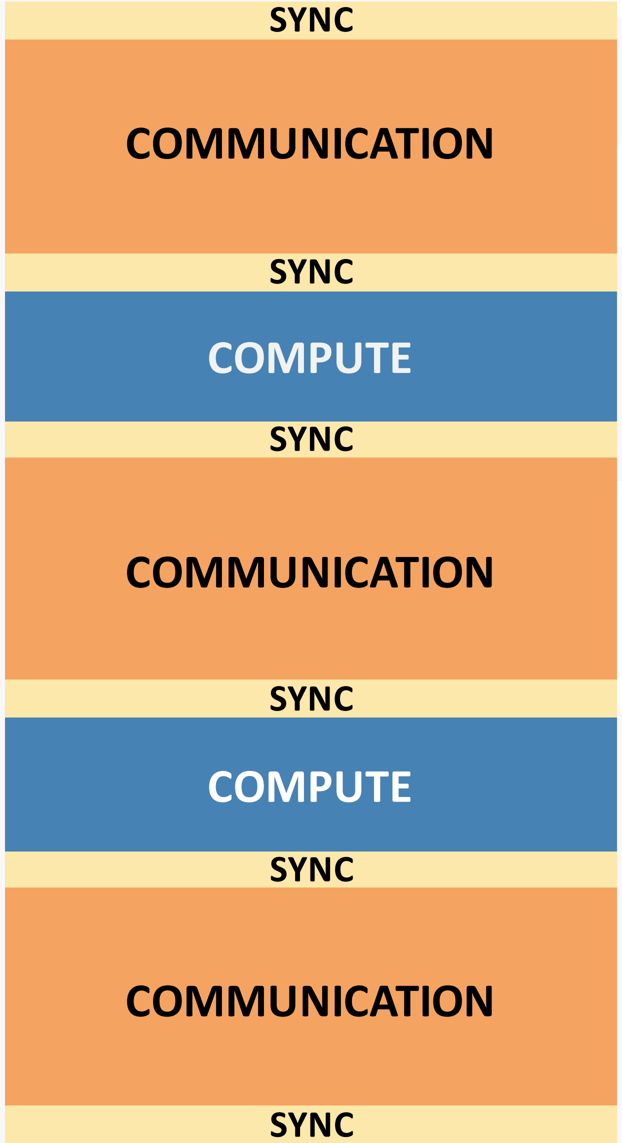


COMPUTATIONAL GRAPH





GRAPH EXECUTION MODEL



SYNC

COMMUNICATION

SYNC

COMPUTE

SYNC

COMMUNICATION

SYNC

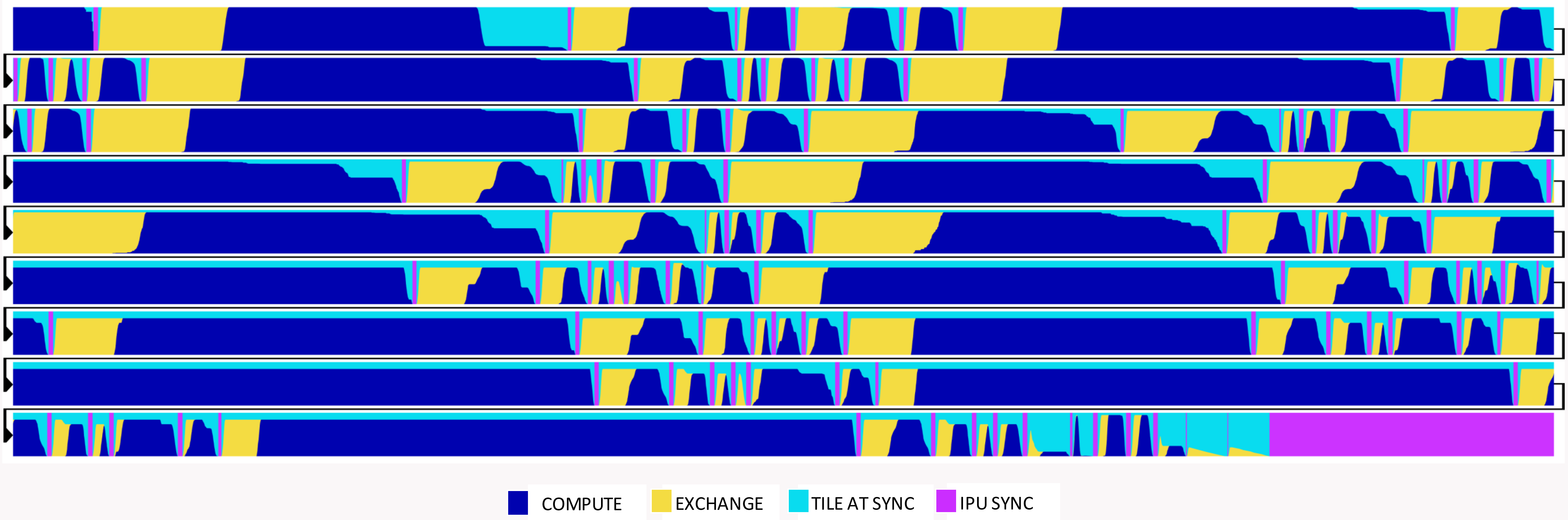
COMPUTE

SYNC

COMMUNICATION

SYNC

IPU BSP EXECUTION TRACE



RESNET-18 INFERENCE BATCH SIZE 1

OPEN-SOURCE GRAPH LIBRARIES

> 50 open-source GRAPH FUNCTIONS

available including (matmul, conv, etc) built from…

> 750 optimized COMPUTE ELEMENTS

such as (ReduceAdd, AddToChannel, Zero, etc)

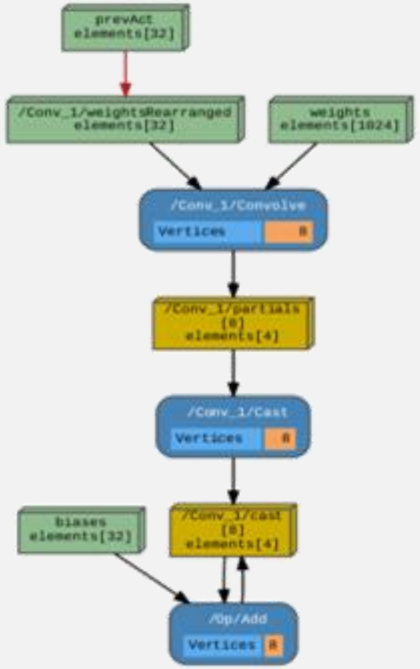
easily create new GRAPH FUNCTIONS

using the library of COMPUTE ELEMENTS

modify and create new COMPUTE ELEMENTS

example GRAPH FUNCTION

32in_32out_Fully_Connected_Layer



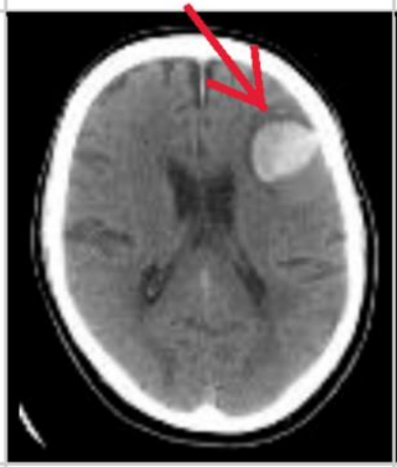
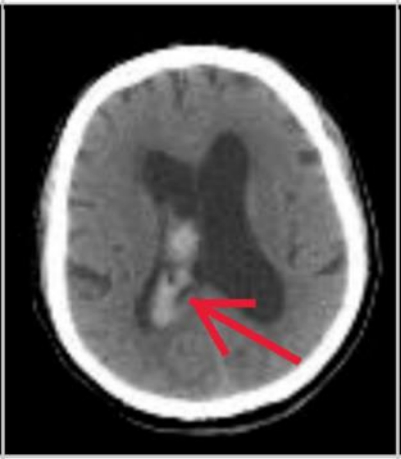
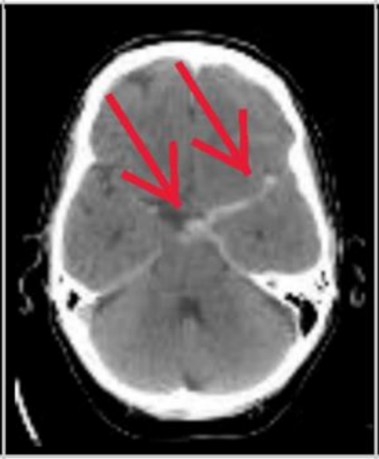
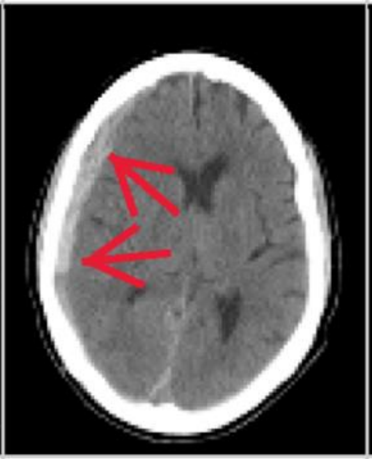
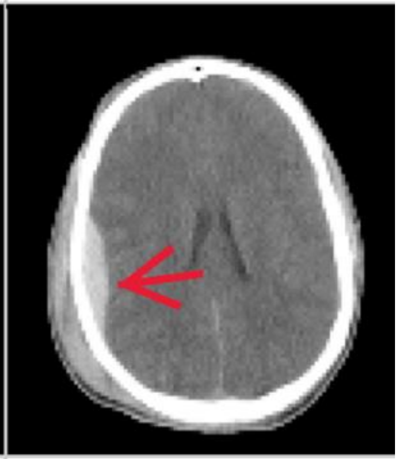
GitHub share library elements and new innovations

IPU-ACCELERATED MEDICAL IMAGING ON MICROSOFT AZURE

Slides & Work Courtesy of:

Microsoft AI & Advanced Architectures Group

INTRACRANIAL HEMORRHAGE

	Intraparenchymal	Intraventricular	Subarachnoid	Subdural	Epidural
Location	Inside of the brain	Inside of the ventricle	Between the arachnoid and the pia mater	Between the Dura and the arachnoid	Between the dura and the skull
Imaging					
Mechanism	High blood pressure, trauma, arteriovenous malformation, tumor, etc	Can be associated with both intraparenchymal and subarachnoid hemorrhages	Rupture of aneurysms or arteriovenous malformations or trauma	Trauma	Trauma or after surgery
Source	Arterial or venous	Arterial or venous	Predominantly arterial	Venous (bridging veins)	Arterial

Courtesy: RSNA

INTRACRANIAL HEMORRHAGE

Trauma: Every case is an emergency; lots of patients, very little time

Extremely Time Critical: Early detection ➔ life-saving implications

Acceleration: Faster inference ➔ timely, precise diagnosis. No patient left untreated.

Deep learning for healthcare – hardware acceleration more relevant than ever!

INFERENCE ON A RESNEXT-50 PRETRAINED MODEL

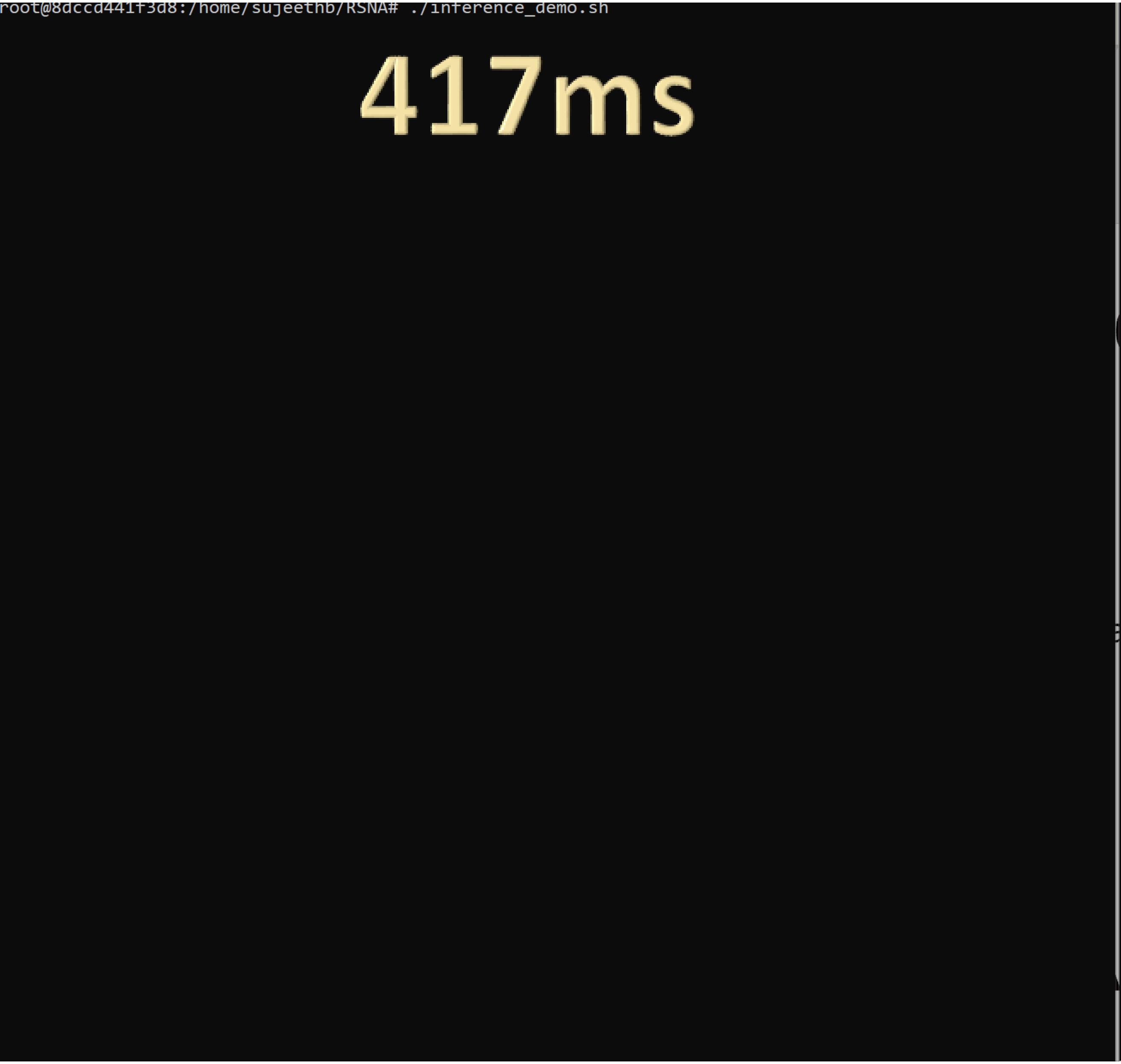
Model: ResNeXt-50 (23M parameters)

Data: 600k randomly selected slices from the ICHD challenge dataset

Data Augmentation: random flip LR & UD, random brightness & contrast, random rotations

Slice-by-slice inference on 3D CT volumes

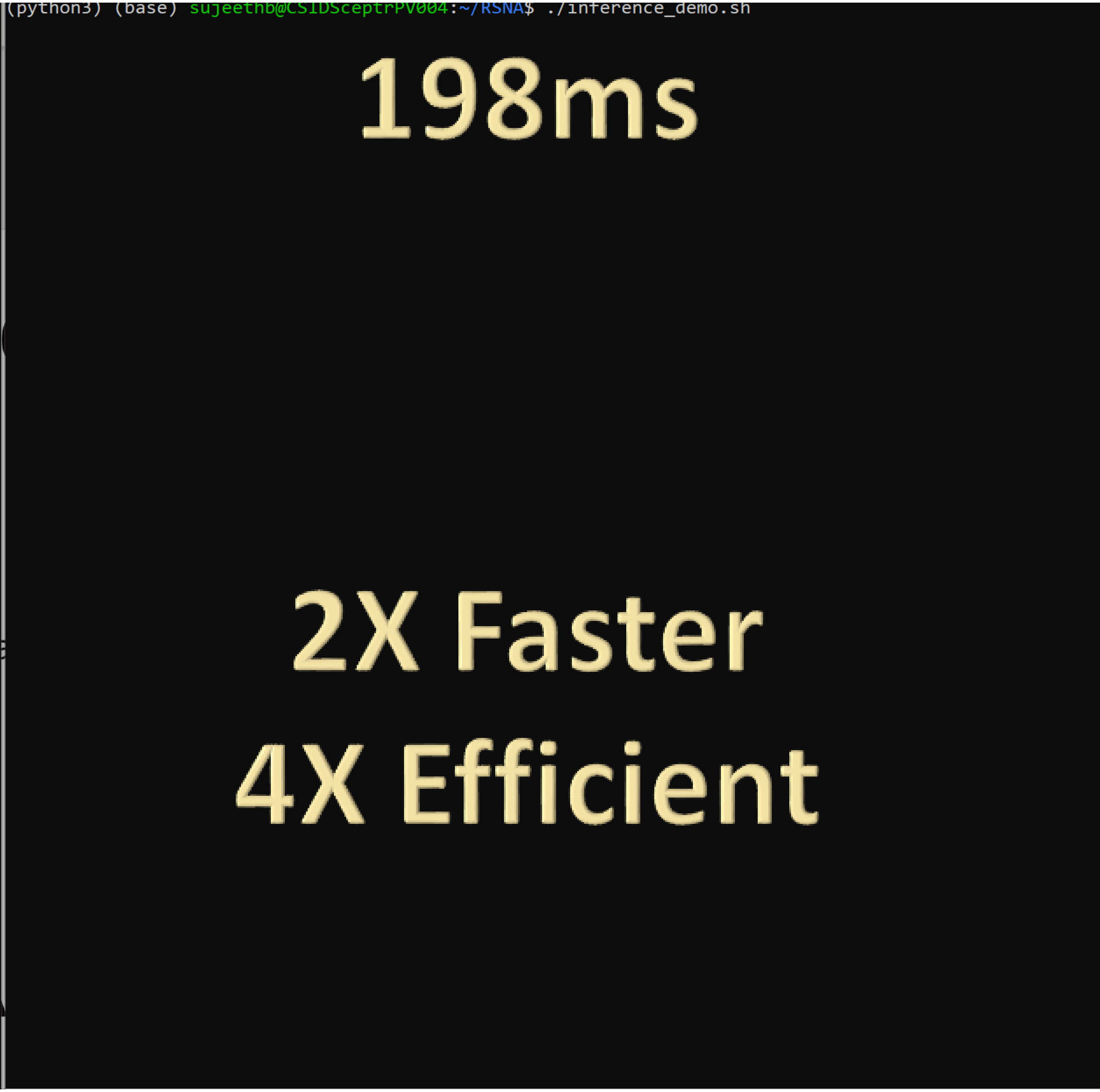
GPU: V100 (300W)



```
root@8dccd441f3d8:/home/sujeethb/RSNA# ./inference_demo.sh
```

417ms

IPU: Single Chip (150W)



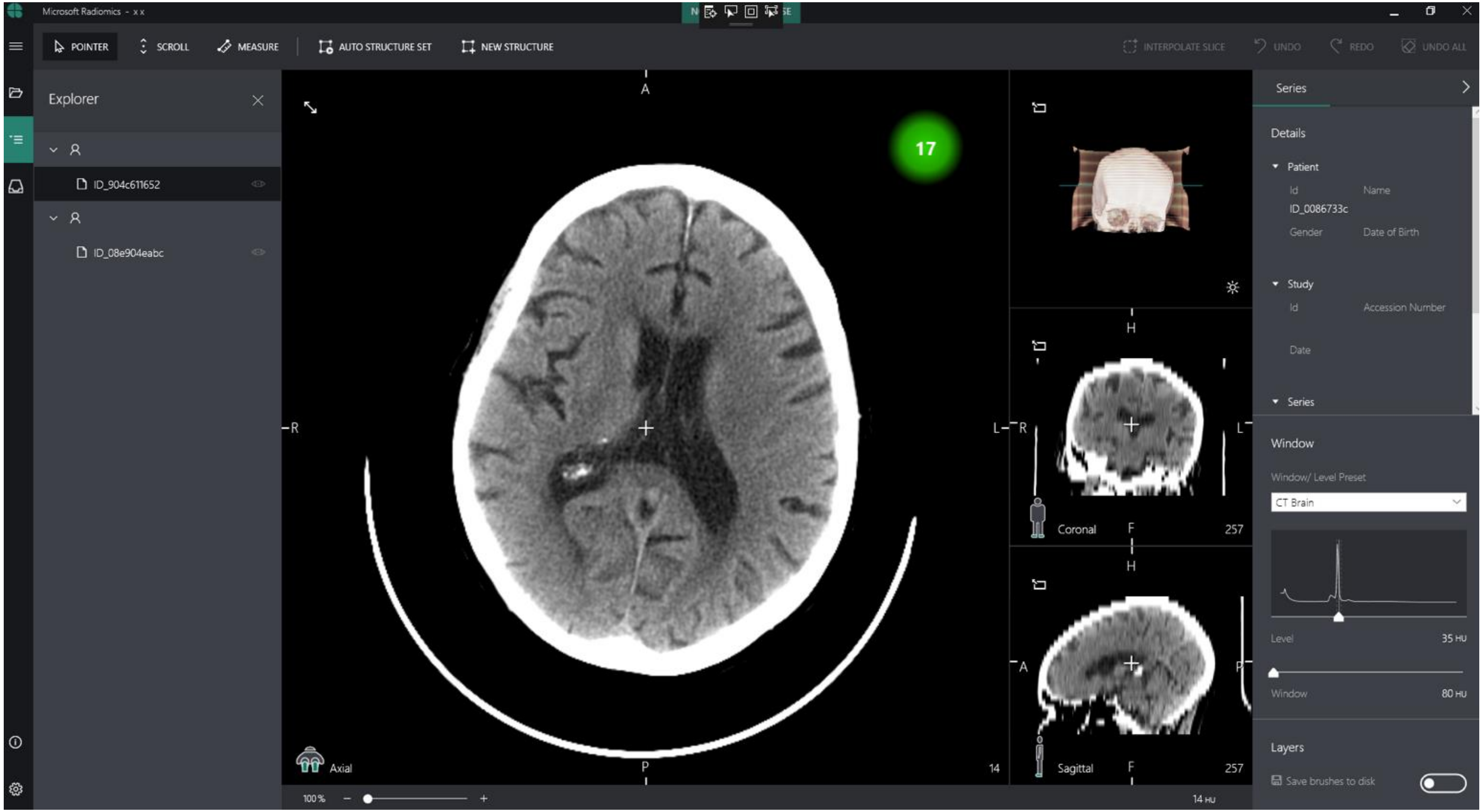
```
(python3) (base) sujeethb@CSIDSceptrPV004:~/RSNA$ ./inference_demo.sh
```

198ms

2X Faster

4X Efficient

INFERENCE RESULTS VISUALIZATION (MICROSOFT INNEREYE)



Id: ID_0086733c

Name

Gender

Date of Birth

Study Id

Accession Number

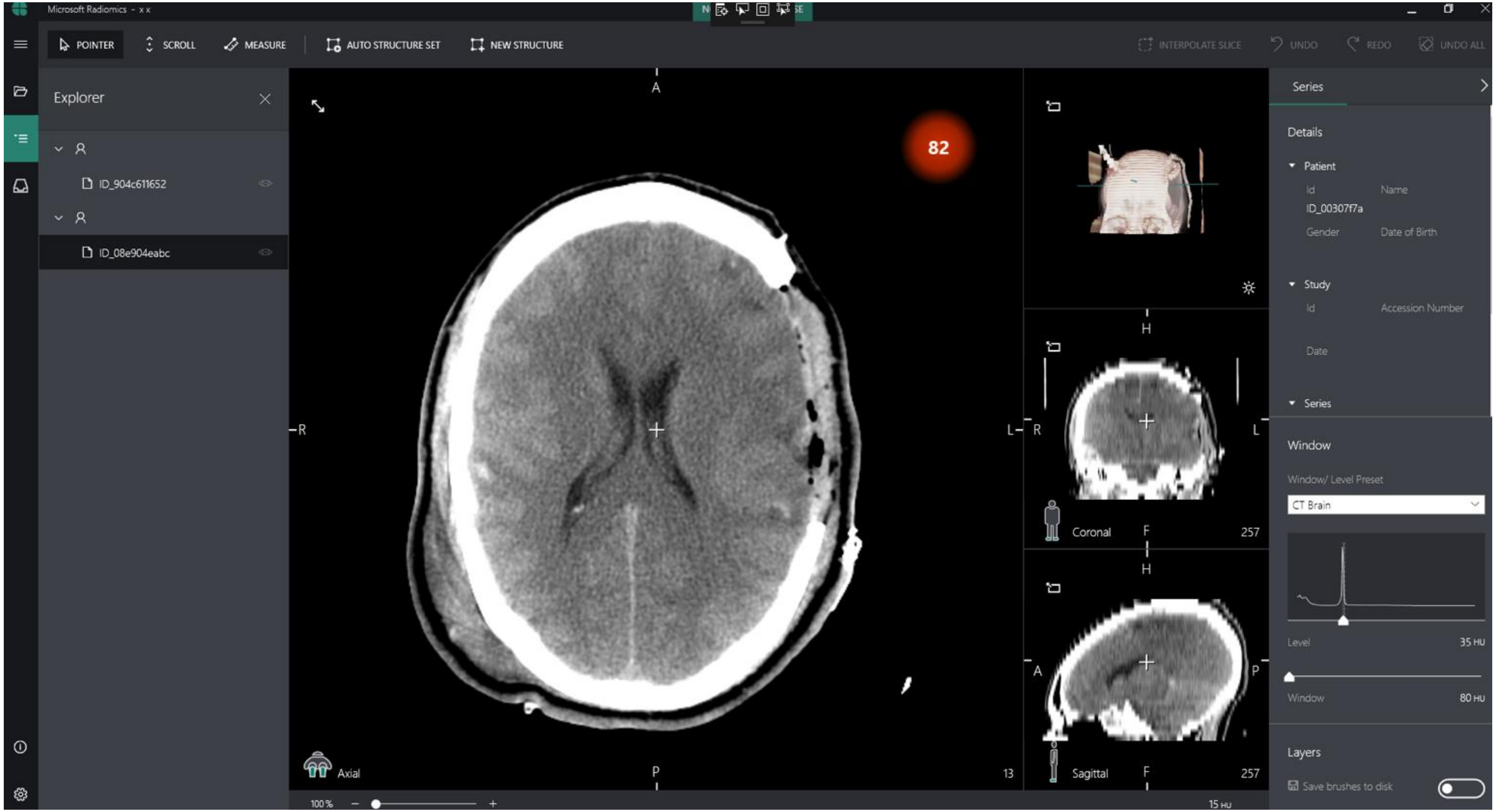
Date

Window/ Level Preset: CT Brain

Level: 35 HU

Window: 80 HU

INFERENCE RESULTS VISUALIZATION (MICROSOFT INNEREYE)



Id: ID_00307f7a
Name
Gender
Date of Birth
Id
Accession Number
Date
Window/ Level Preset: CT Brain
Level: 35 HU
Window: 80 HU

ACCELERATE YOUR RESEARCH WITH STATE OF THE ART PERFORMANCE IPU TECHNOLOGY

Achieving the next big breakthrough in AI is only possible with the right toolkit. The Graphcore IPU Preview on Microsoft Azure allows researchers to run new and complex machine learning models orders of magnitude faster.

Discover what you could achieve with a processor designed specifically for machine intelligence workloads.

GRAPHCORE

Microsoft Azure

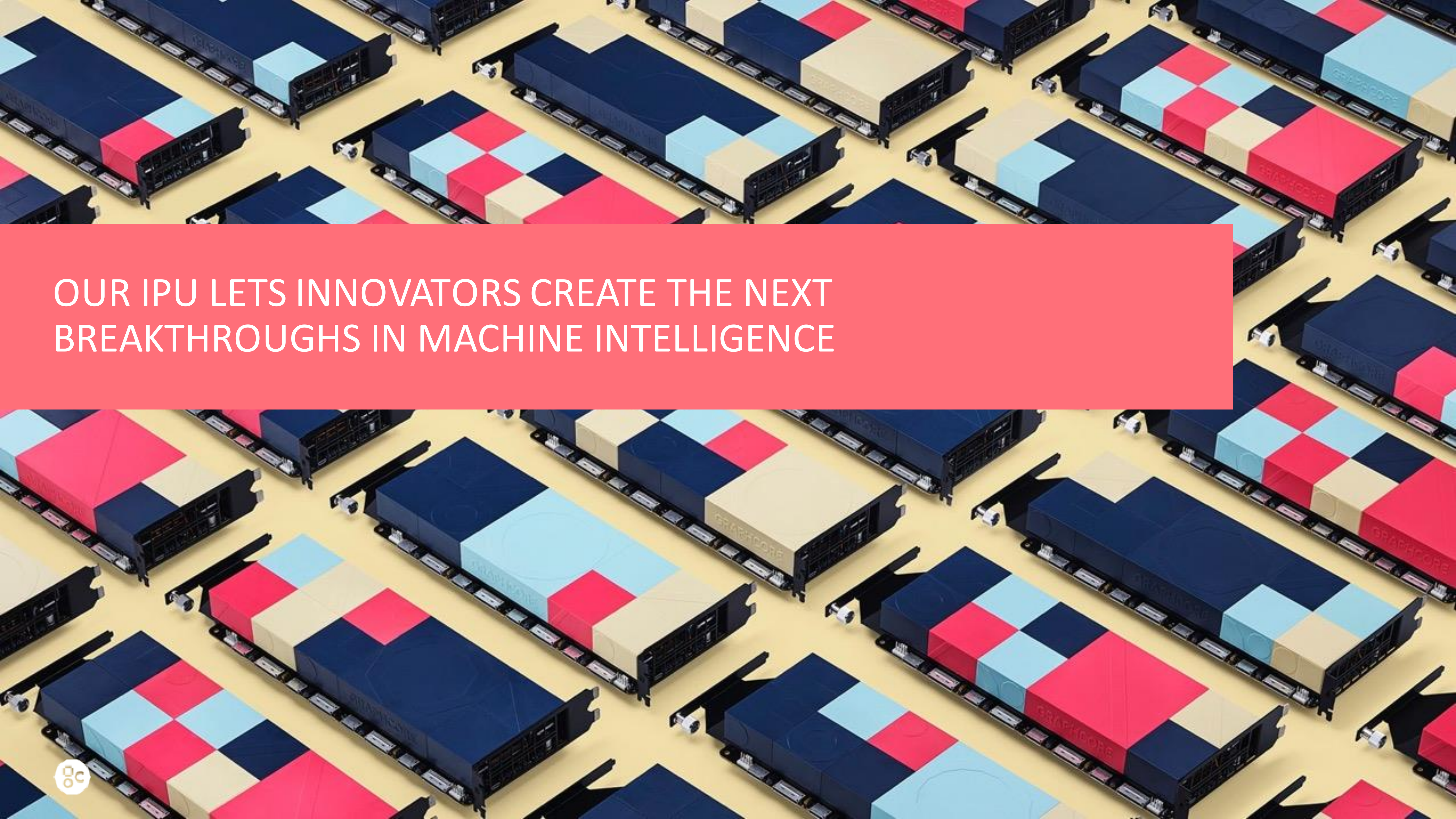
Sign up for IPU preview on Azure

Cirrascale CLOUD SERVICES

Buy now from Cirrascale

DELL EMC

Buy now from Dell:



OUR IPU LETS INNOVATORS CREATE THE NEXT BREAKTHROUGHS IN MACHINE INTELLIGENCE

KEEP IN TOUCH WITH US

BLOG	GRAPHCORE.AI/BLOG
NEWSLETTER	GRAPHCORE.AI/NEWS
TWITTER	@GRAPHCORE.AI
LINKEDIN	LINKEDIN.COM/COMPANY/GRAPHCORE
FACEBOOK	@GRAPHCORE.AI

THANK YOU

Victoria Rege, Director of Strategic Partnerships & Alliances
victoria@graphcore.ai

Alexander Titterton, Product Support Engineer
alexandert@graphcore.ai