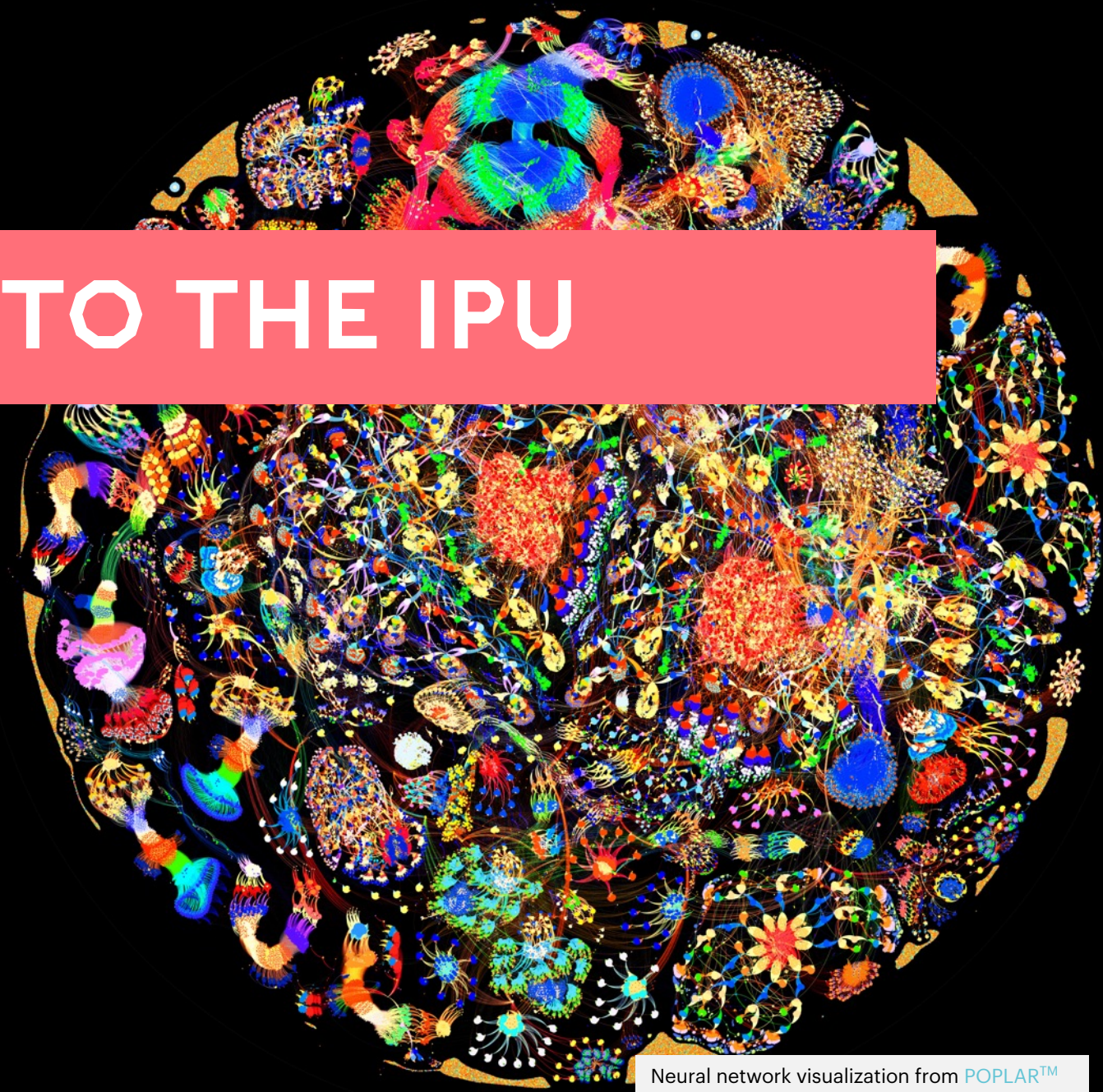
INTRODUCTION TO THE IPU

GRAPHCORE



Neural network visualization from POPLAR™

The Hardware Lottery

Sara Hooker

Google Research, Brain Team
shooker@google.com

Abstract

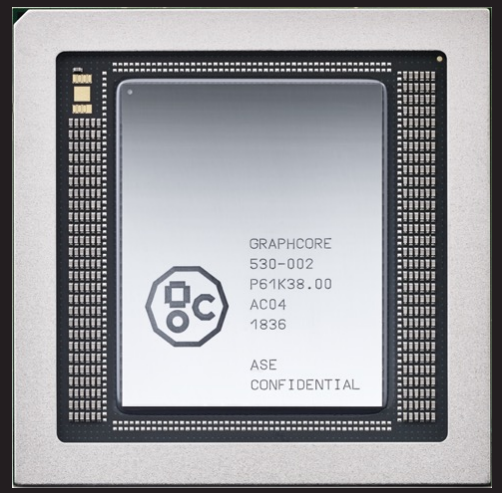
Hardware, systems and algorithms research communities have historically had different incentive structures and fluctuating motivation to engage with each other explicitly. This historical treatment is odd given that hardware and software have frequently determined which research ideas succeed (and fail). This essay introduces the term hardware lottery to describe when a research idea wins because it is suited to the available software and hardware and *not* because the idea is superior to alternative research directions. Examples from early computer science history illustrate how hardware lotteries can delay research progress by casting successful ideas as failures. These lessons are particularly salient given the advent of domain specialized hardware which make it increasingly costly to stray off of the beaten path of research ideas. This essay posits that the gains from progress in computing are likely to become even more uneven, with certain research directions moving into the fast-lane while progress on others is further obstructed.

OUR IPU LETS INNOVATORS CREATE THE NEXT BREAKTHROUGHS IN MACHINE INTELLIGENCE

WIRED

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

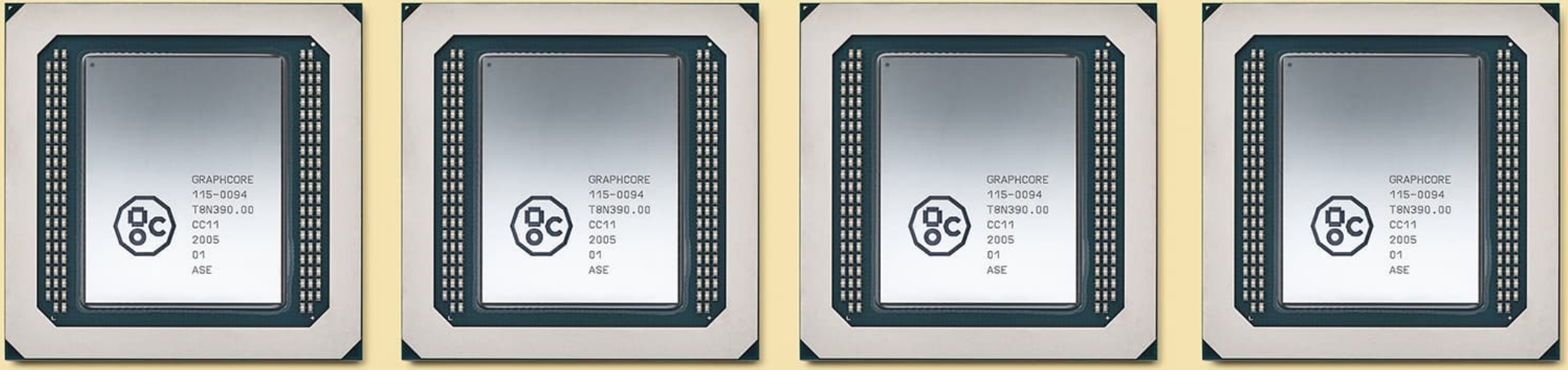
Geoff Hinton – "I think we need to move towards a different type of computer. Fortunately I have one here..." Hinton reaches into his wallet and pulls out a large, shiny silicon chip:



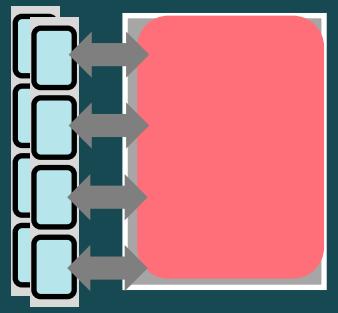
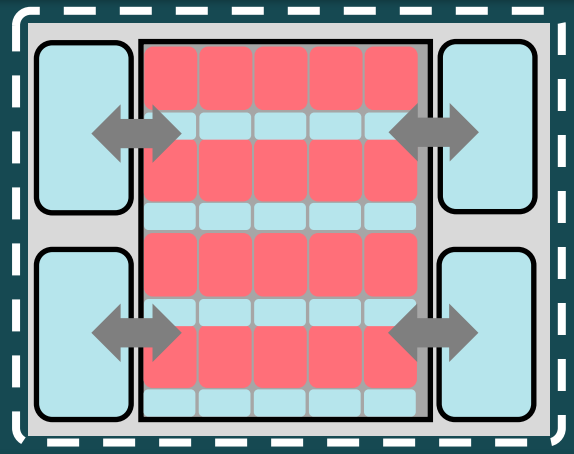
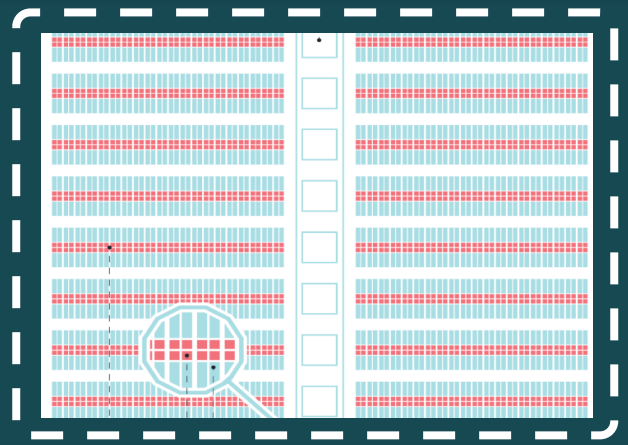
an IPU processor from Graphcore

A DELIBERATELY DIFFERENT TECHNOLOGY

- 2 key differentiators: memory access speed and parallelism
- Orders of magnitude more **parallelism** due to **MIMD** architecture (multiple instruction, multiple data): thousands of independent, programmable cores
- 30x the **memory bandwidth of** GPU due to enormous amount of **memory on the chip**
- Disaggregated approach that allows users to configure **AI compute to CPU** ratio to optimise performance while also maximising power and space available in data centre



PARALLELISM & MEMORY ACCESS

	CPU	GPU	IPU
Parallelism	Suitable for scalar processes	SIMD/SIMT architecture. Suitable for large blocks of dense contiguous data	**Massively parallel MIMD. High performance/efficiency as ML trends to sparsity & small kernels**
			
Memory Access	Off-chip memory	Model and Data spread across off-chip and small on-chip cache and shared mem.	**Model & Data in tightly coupled large locally distributed SRAM**

THE WORLD'S MOST COMPLEX PROCESSOR

COMPUTE

COLOSSUS MK2 IPU

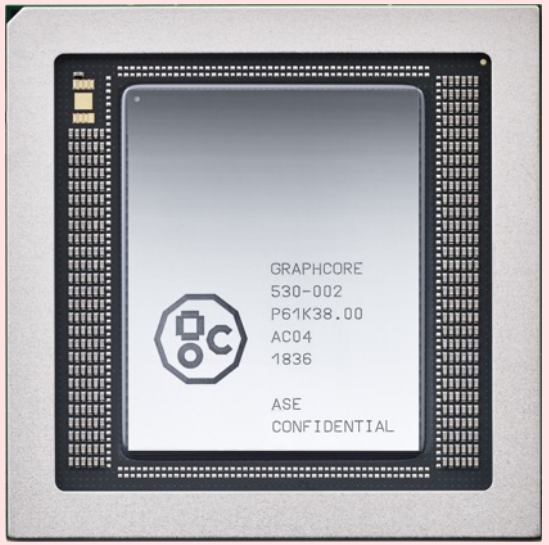
59.4Bn transistors, TSMC 7nm @ 823mm^2

250TFlops AI-Float | 900MB In-Processor-Memory™

1472 independent processor cores

8832 separate parallel threads

>8x step-up in system performance vs Mk1



GC200 IPU

GC200 IPU PROCESSOR

IPU-Tiles™

1472 independent IPU-Tiles™ each with an IPU-Core™ and In-Processor-Memory™

IPU-Core™

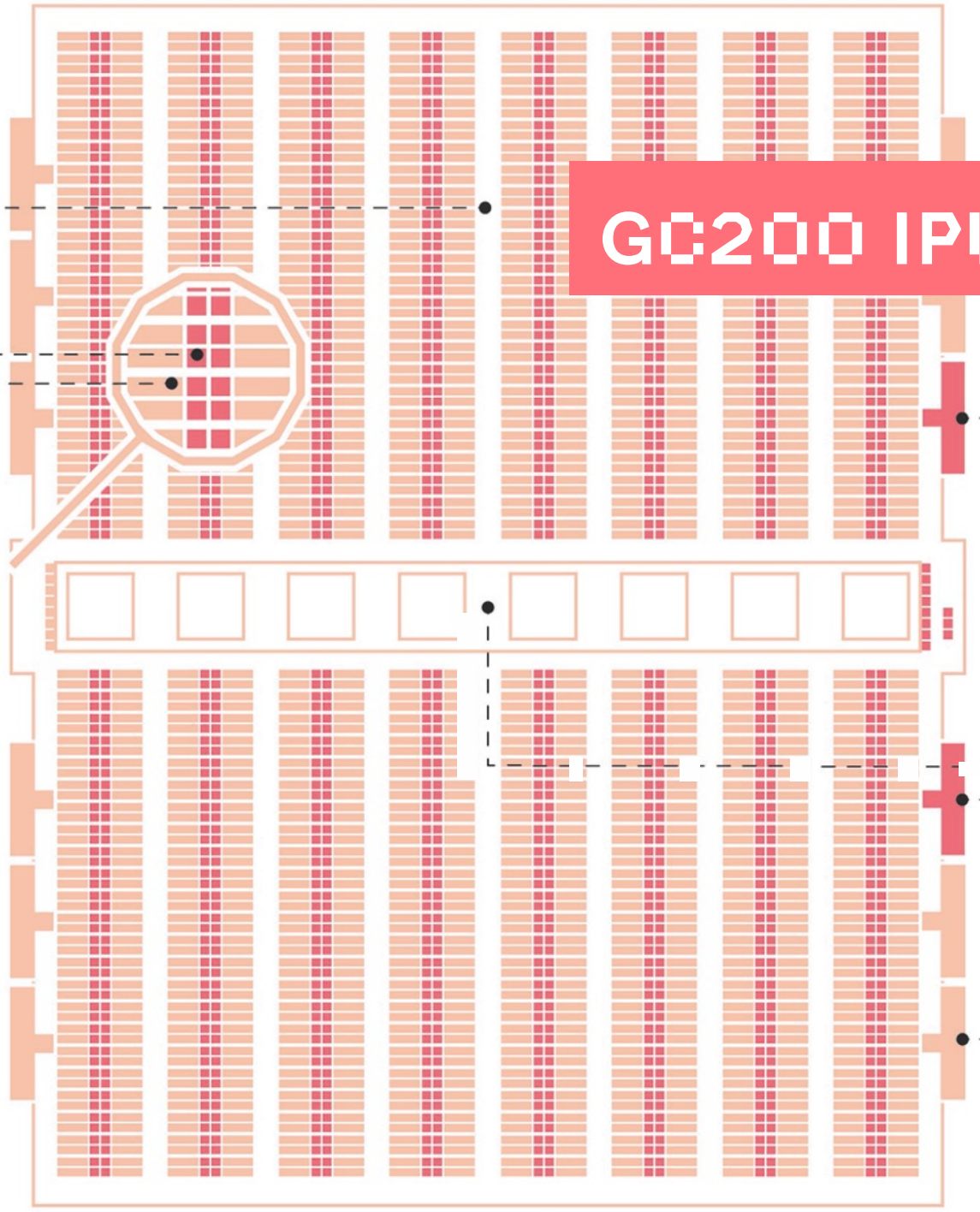
1472 independent IPU-Core™

8832 independent program threads executing in parallel

In-Processor-Memory™

900MB In-Processor-Memory™ per IPU

47.5TB/s memory bandwidth per IPU



IPU-Exchange™

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

PCIe

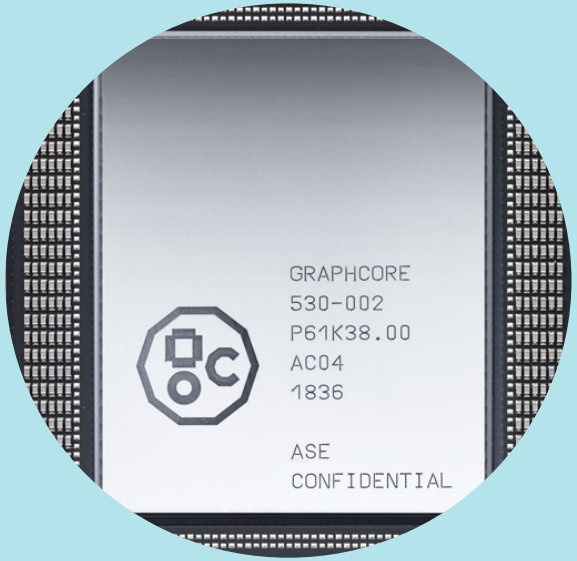
PCI Gen4 x16
64 GB/s bidirectional bandwidth to host

IPU-Links™

10 x IPU-Links,
320GB/s chip to chip bandwidth

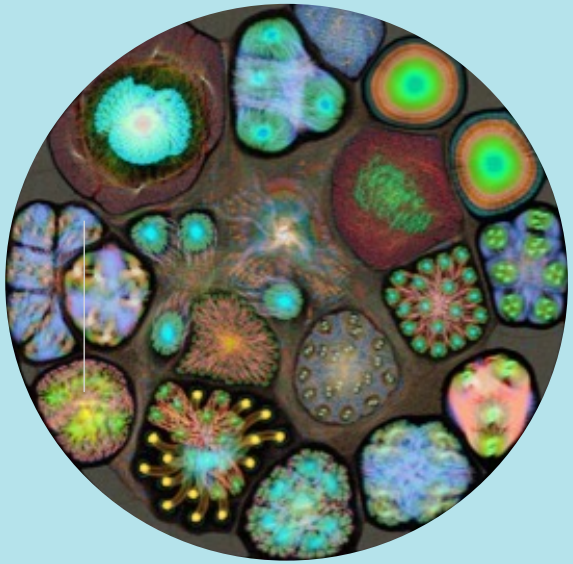
NEXT GEN AI SOLUTIONS

Hardware



IPU processors designed for AI

Software



Poplar® software stack & development tools

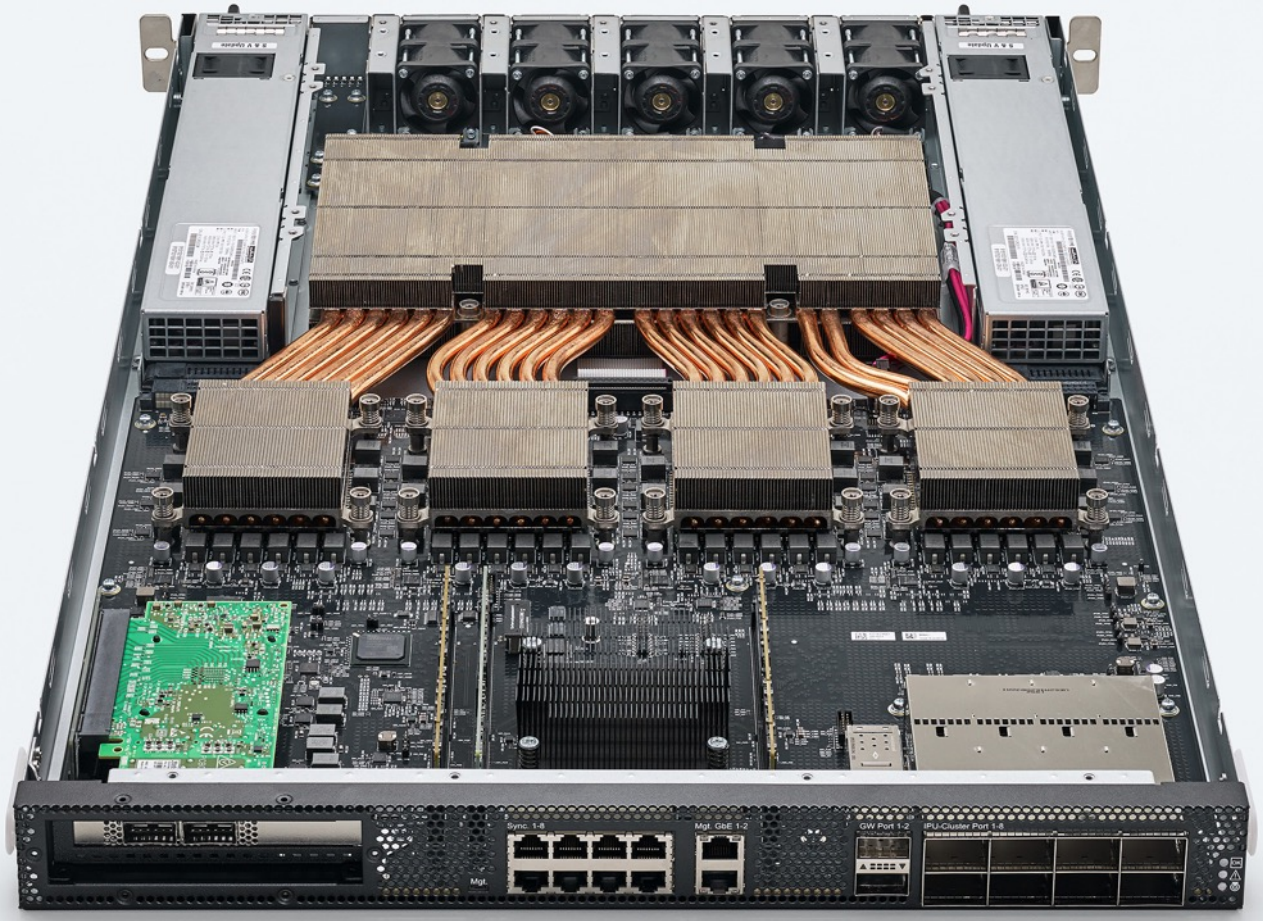
Platforms

M2000 and Server IPU-POD$_{64}$ scale-out

IPU-MACHINE

IPU-M2000

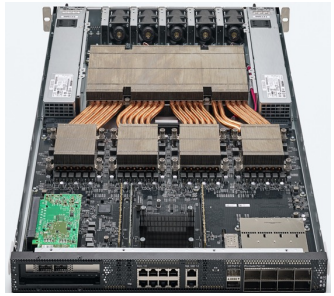
1 PetaFlop IPU compute
2.8Tbps IPU-Fabric™



MAXIMISING ROI ON INNOVATION

M2000 – a building block for next gen data centers

x4

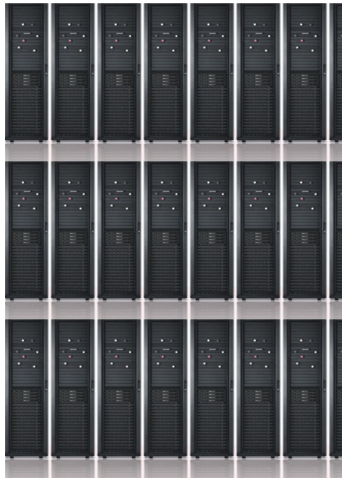


x4

IPU-POD16 Direct Attach

x4

x1024



GC200

IPU-POD$_4$

4 IPUs
1 PetaFlop

IPU-POD$_{16}$

16 IPUs
4 PetaFlop
4x IPU-M2000

IPU-POD$_{64}$

64 IPUs
16 PetaFlop
16x IPU-M2000

IPU-POD$_{64k}$

64k IPUs
16 ExaFlop
<7.4 PB

THE IPU LETS INNOVATORS CREATE THE NEXT GENERATION OF MACHINE INTELLIGENCE

WORKLOAD EXAMPLES GUIDE

Compute intensive benefits from significant amount of compute

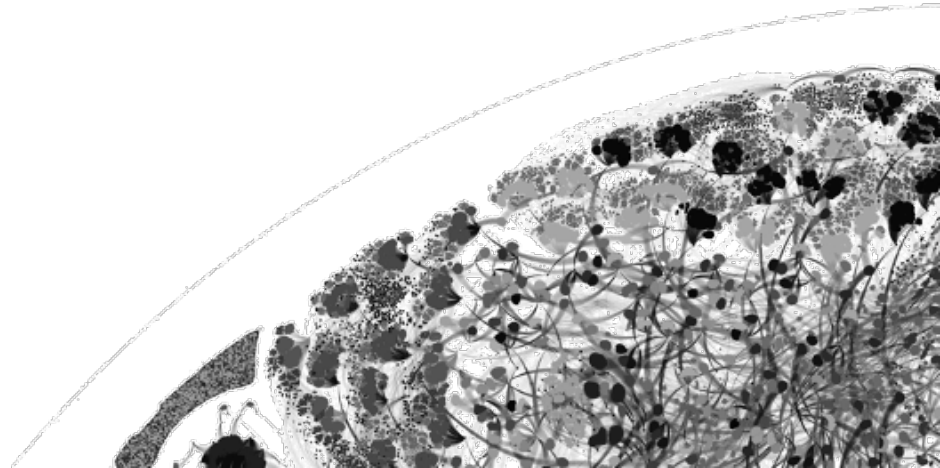
Sparse or fine-grain leverages the unique architecture

Sequential leverages the fast in-processor memory

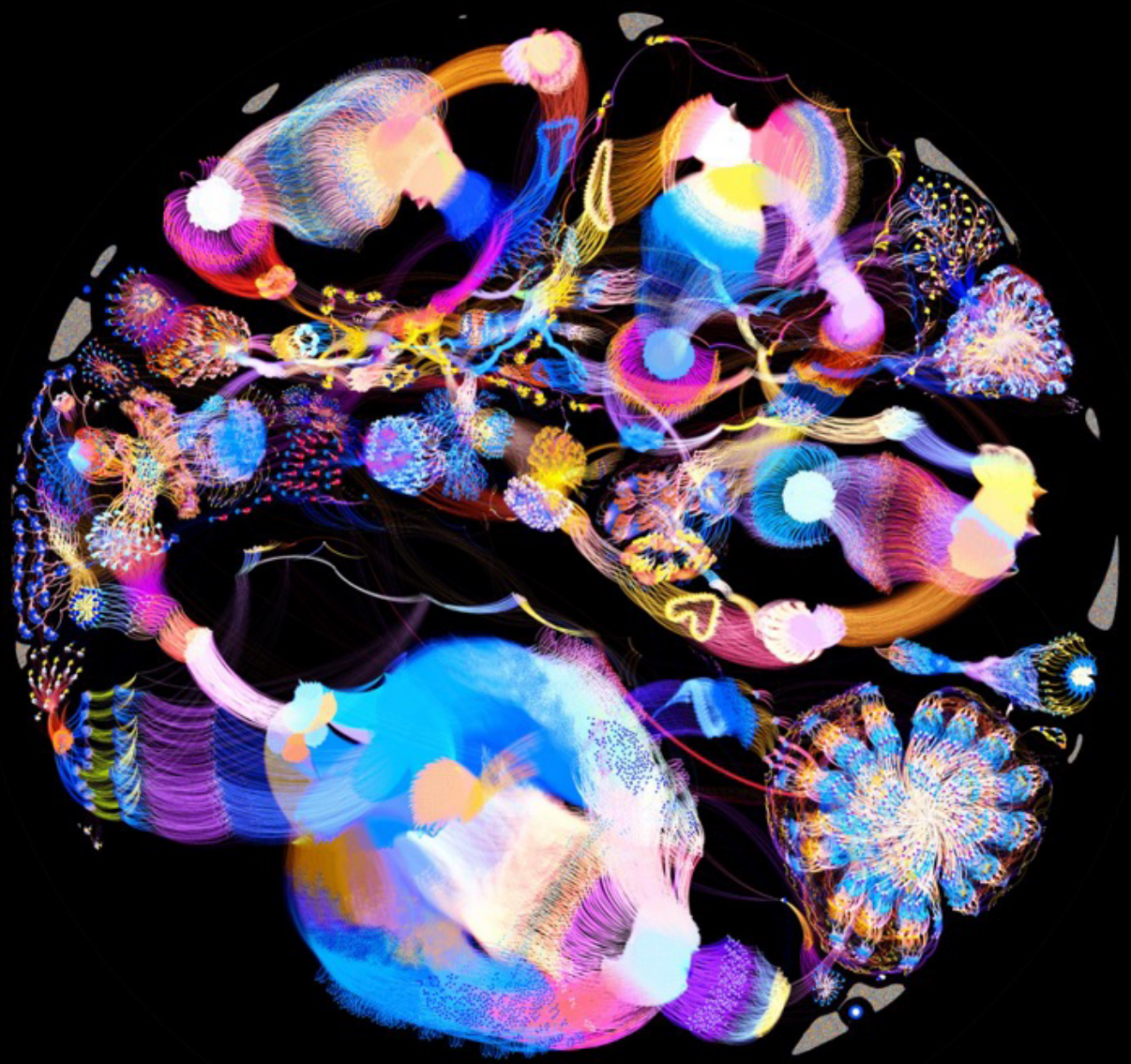
Nontrivial leverages the available parallelism

Hard to vectorise but possible to parallelise

Not bottlenecked by host compute or external I/O

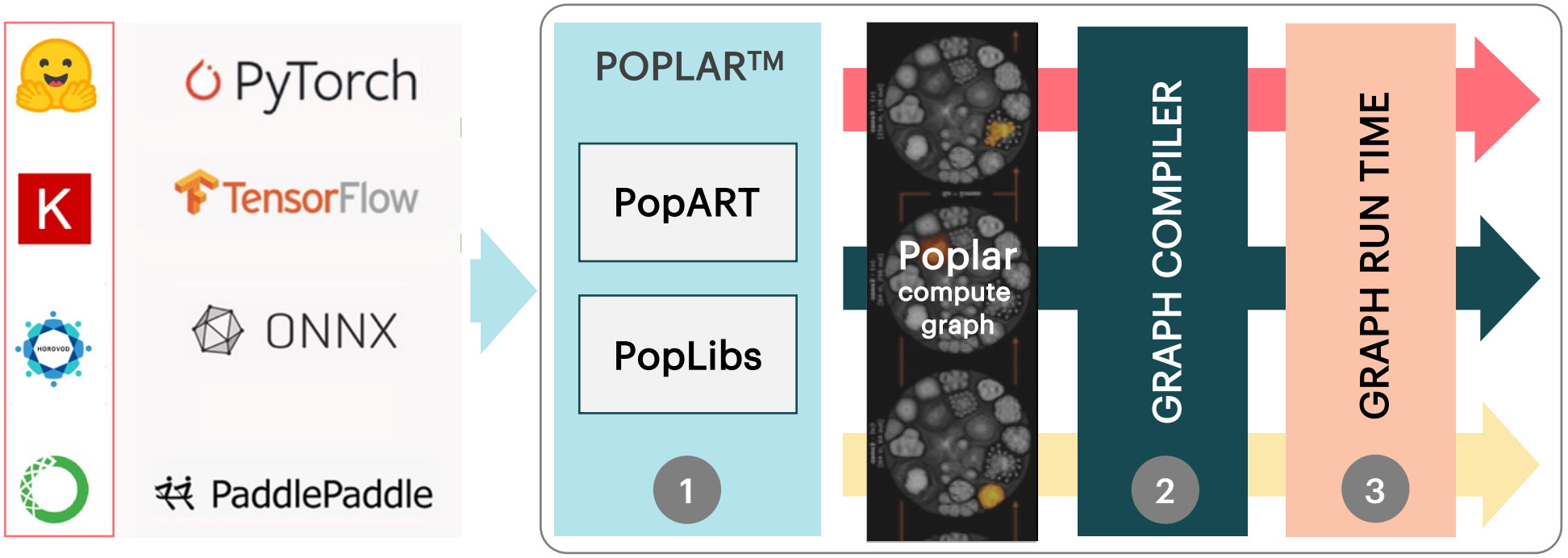


SOFTWARE



POPLAR™ COMPUTE GRAPH VISUALISATION

POPLAR-SDK

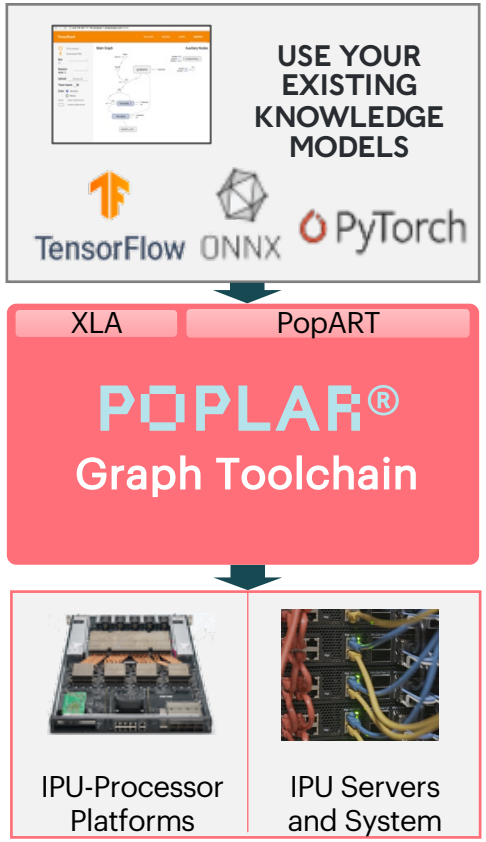


1,000's of IPUs
+ compiled communications

POPLAR EASE OF USE

TensorFlow PYTORCH Azure docker

Keras ONNX kubernetes



Open & Extensible Poplar Libraries

Get access to 50+ optimised functions for common ML models and 750 high performance compute elements. Modify and write custom libraries.

Straightforward Deployment

Pre-built Docker containers with Poplar SDK, Tools and frameworks images to get up and running fast.

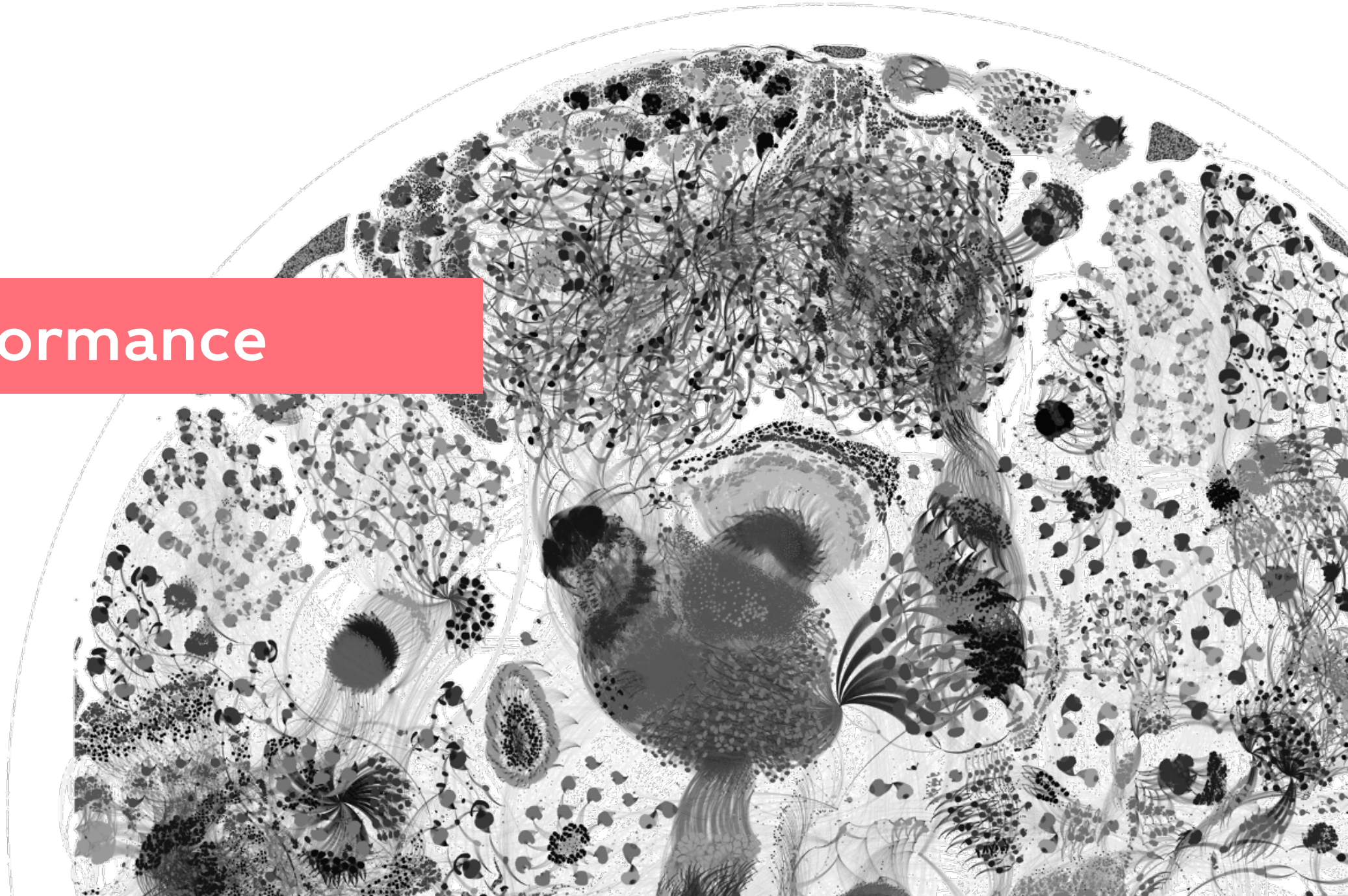
ML Frameworks Support

Support for standard ML frameworks: TensorFlow 1 and 2, ONNX and PyTorch with PaddlePaddle coming soon.

Standard Ecosystem Support

Ready for production with Microsoft Azure deployment, Kubernetes orchestration, Docker containers and Hyper-V virtualization & security.

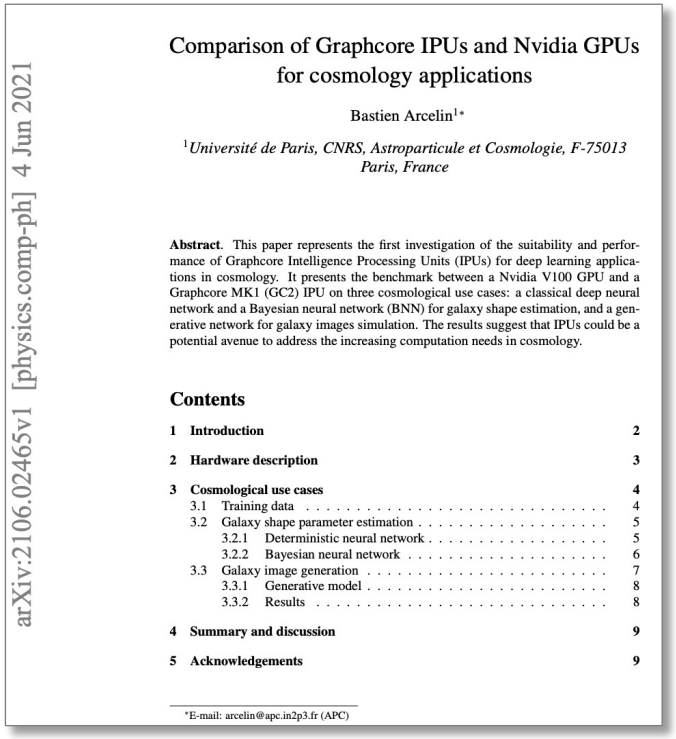
IPU Performance



ACADEMIC & RESEARCH ENGAGEMENTS

IPU ACCELERATION FOR COSMOLOGY APPLICATIONS

Université de Paris



Comparison of Graphcore IPUs and Nvidia GPUs for cosmology applications

Bastien Arcelin[1]*

[1]*Université de Paris, CNRS, Astroparticule et Cosmologie, F-75013 Paris, France*

Abstract. This paper represents the first investigation of the suitability and performance of Graphcore Intelligence Processing Units (IPUs) for deep learning applications in cosmology. It presents the benchmark between a Nvidia V100 GPU and a Graphcore MK1 (GC2) IPU on three cosmological use cases: a classical deep neural network and a Bayesian neural network (BNN) for galaxy shape estimation, and a generative network for galaxy images simulation. The results suggest that IPUs could be a potential avenue to address the increasing computation needs in cosmology.

Contents

*E-mail: arcelin@apc.in2p3.fr (APC)

Jul 09, 2021 \ Research, University

UNIVERSITÉ DE PARIS ACCELERATES COSMOLOGY APPLICATIONS WITH GRAPHCORE IPUS

Written By:

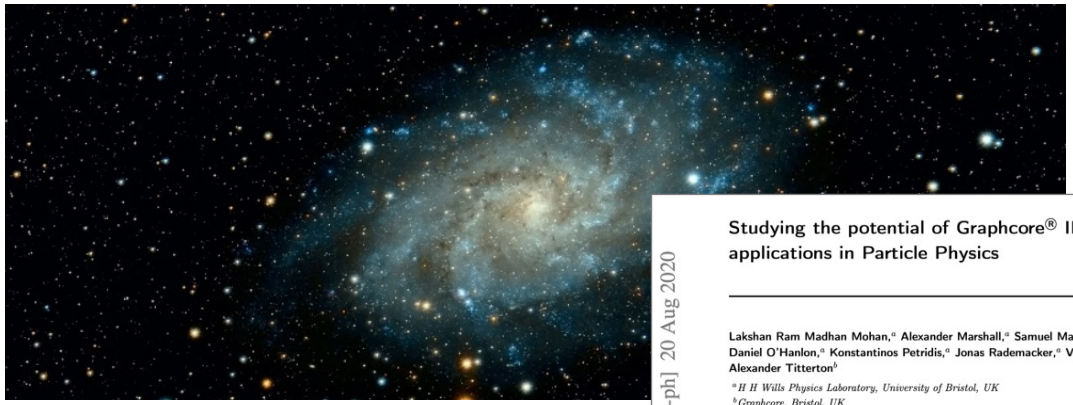
Alex Titterton

This paper represents the first investigation of the suitability and performance of Graphcore Intelligence Processing Units (IPUs) for deep learning applications in cosmology"……"on three cosmological use cases: a classical deep neural network and a Bayesian neural network (BNN) for galaxy shape estimation, and a generative network for galaxy images production."

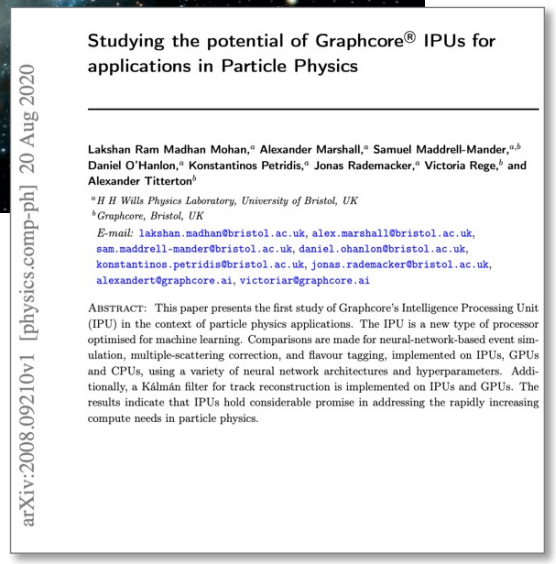
The results show that **IPUs can accelerate various cosmology applications, outperforming GPUs in some cases by as much as 4x faster time to train**"

https://www.graphcore.ai/resources/research-papers

IPUs in Research



UNIVERSITY OF BRISTOL TACKLES CHALLENGES IN PARTICLE PHYSICS WITH GRAPHCORE'S IPU



arXiv:2008.09210v1 [physics.comp-ph] 20 Aug 2020

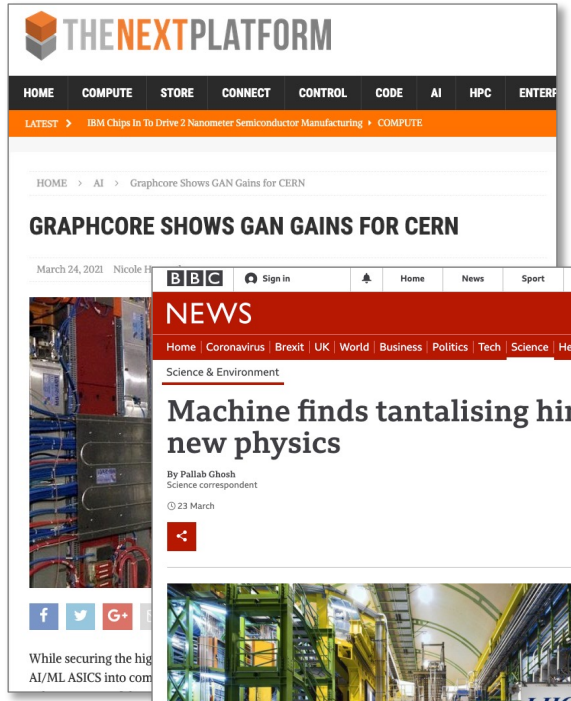
Studying the potential of Graphcore® IPUs for applications in Particle Physics

Lakshan Ram Madhan Mohan,[a] Alexander Marshall,[a] Samuel Maddrell-Mander,[a,b] Daniel O'Hanlon,[a] Konstantinos Petridis,[a] Jonas Rademacker,[a] Victoria Rege,[b] and Alexander Titterton[b]

[a] H H Wills Physics Laboratory, University of Bristol, UK
[b] Graphcore, Bristol, UK

E-mail: lakshan.madhan@bristol.ac.uk, alex.marshall@bristol.ac.uk, sam.maddrell-mander@bristol.ac.uk, daniel.ohanlon@bristol.ac.uk, konstantinos.petridis@bristol.ac.uk, jonas.rademacker@bristol.ac.uk, alexandert@graphcore.ai, victoriar@graphcore.ai

Abstract: This paper presents the first study of Graphcore's Intelligence Processing Unit (IPU) in the context of particle physics applications. The IPU is a new type of processor optimised for machine learning. Comparisons are made for neural-network-based event simulation, multiple-scattering correction, and flavour tagging, implemented on IPUs, GPUs and CPUs, using a variety of neural network architectures and hyperparameters. Additionally, a Kálmán filter for track reconstruction is implemented on IPUs and GPUs. The results indicate that IPUs hold considerable promise in addressing the rapidly increasing compute needs in particle physics.



THENEXTPLATFORM

HOME | COMPUTE | STORE | CONNECT | CONTROL | CODE | AI | HPC | ENTERP

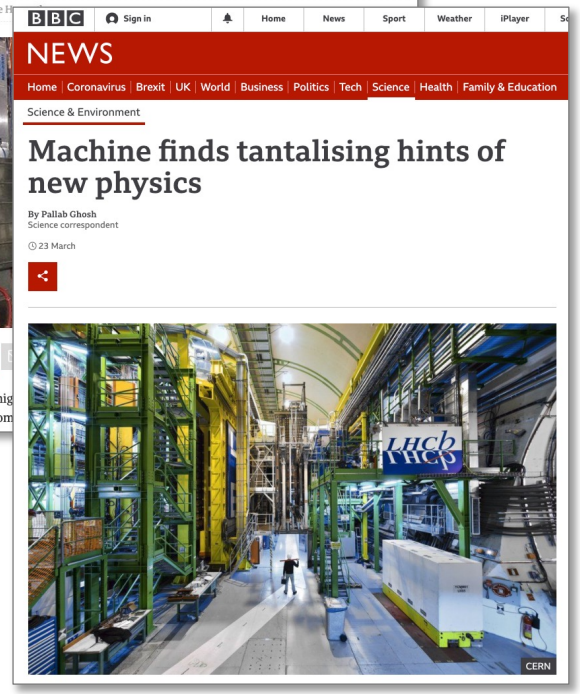
LATEST › IBM Chips In To Drive 2 Nanometer Semiconductor Manufacturing › COMPUTE

HOME › AI › Graphcore Shows GAN Gains for CERN

GRAPHCORE SHOWS GAN GAINS FOR CERN

March 24, 2021 Nicole H

While securing the hig
AI/ML ASICS into com



BBC Sign in | Home | News | Sport | Weather | iPlayer

NEWS

Home | Coronavirus | Brexit | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Science & Environment

Machine finds tantalising hints of new physics

By Pallab Ghosh
Science correspondent

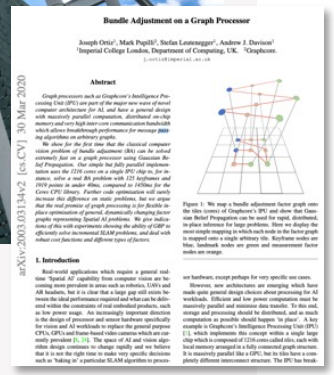
23 March

University of BRISTOL

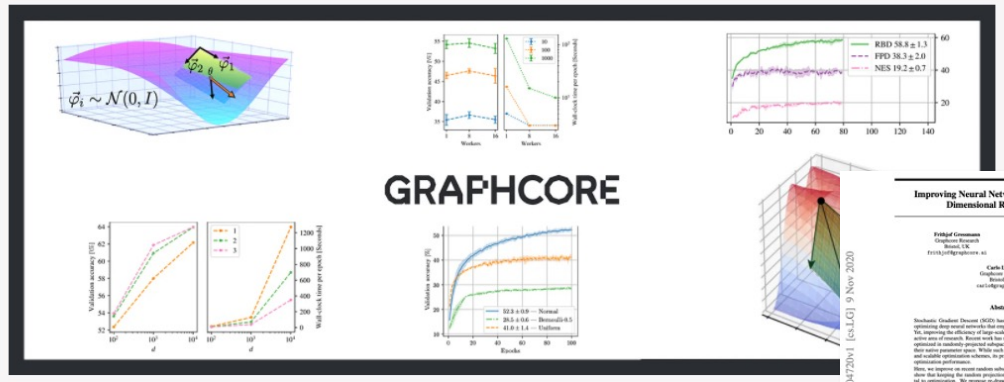
CERN

https://www.graphcore.ai/resources/research-papers

IPUs in Research



IMPERIAL COLLEGE LONDON ACCELERATE CLASSICAL COMPUTER VISION PROBLEM ON IPU



TRAINING NEURAL NETWORKS IN LOW-DIMENSIONAL RANDOM BASES

GRAPHCORE

REVISITING SMALL BATCH TRAINING FOR DEEP NEURAL NETWORKS

University of BRISTOL

UNIVERSITY OF BRISTOL SOLVES SCIENTIFIC PROBLEMS WITH NEW IPU-BASED AI SYSTEMS

https://www.graphcore.ai/resources/research-papers

GRAPHCORE ACADEMIC PROGRAMME

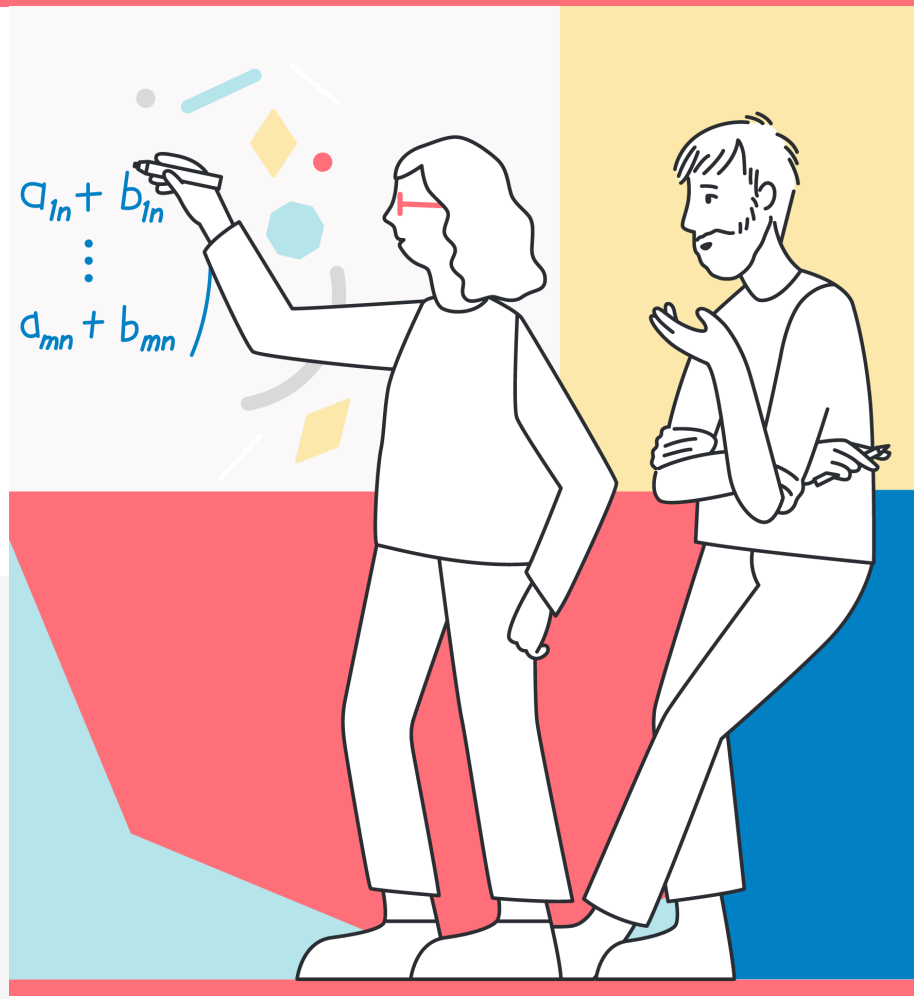
Apply at:
graphcore.ai/academic

Test IPU hardware in the cloud at no-cost

Support letters for grants & funding

Access to Poplar® & PopART® software

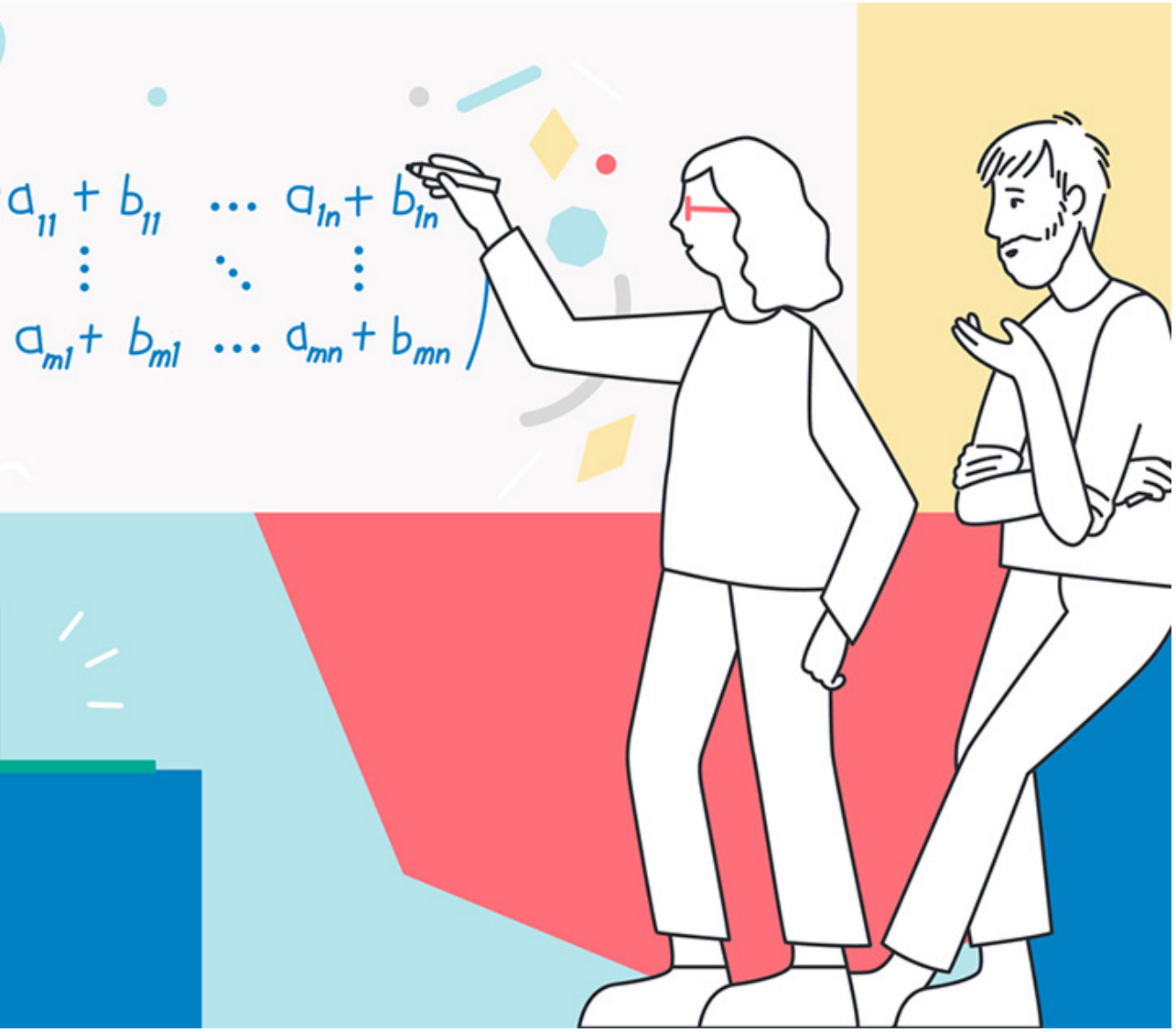
Support from Graphcore Researchers



RESEARCH PRIORITIES:

- Optimisation of Stochastic Learning
- New Efficient Models for Deep Learning & Graph Networks
- Sparse Training
- New Directions for Parallel Training
- Local Parallelism
- Multi-Model Training
- Conditional Sparse Computation

IPU BENEFITS



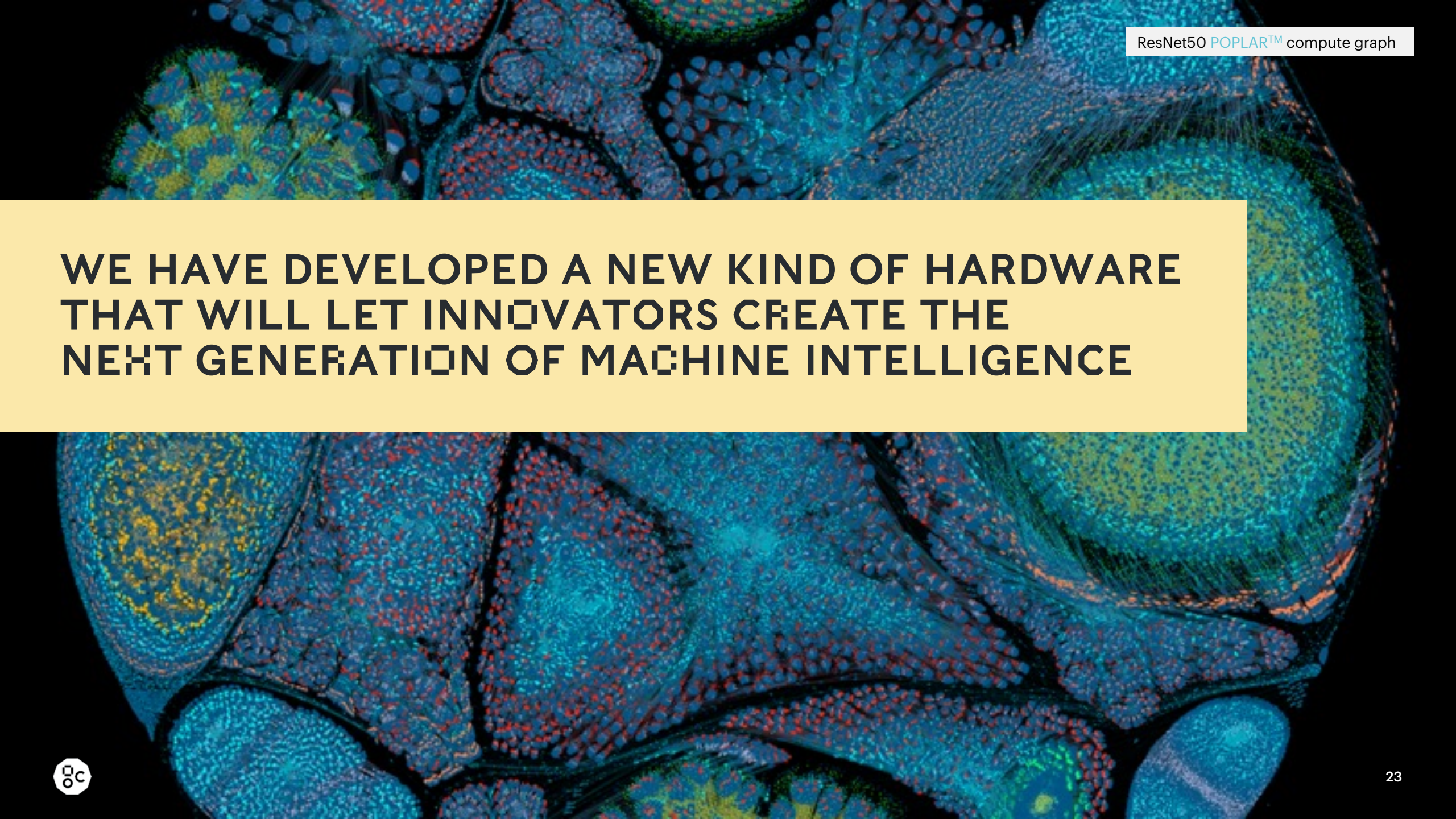
Improving ML model performance

Better performance per dollar

Ability to innovate with new models, previously not feasible on legacy architectures

Easy integration with major ML frameworks like TF and Pytorch

Disaggregated, modular architecture that allows for more flexible scale out solutions



WE HAVE DEVELOPED A NEW KIND OF HARDWARE THAT WILL LET INNOVATORS CREATE THE NEXT GENERATION OF MACHINE INTELLIGENCE

THANK YOU