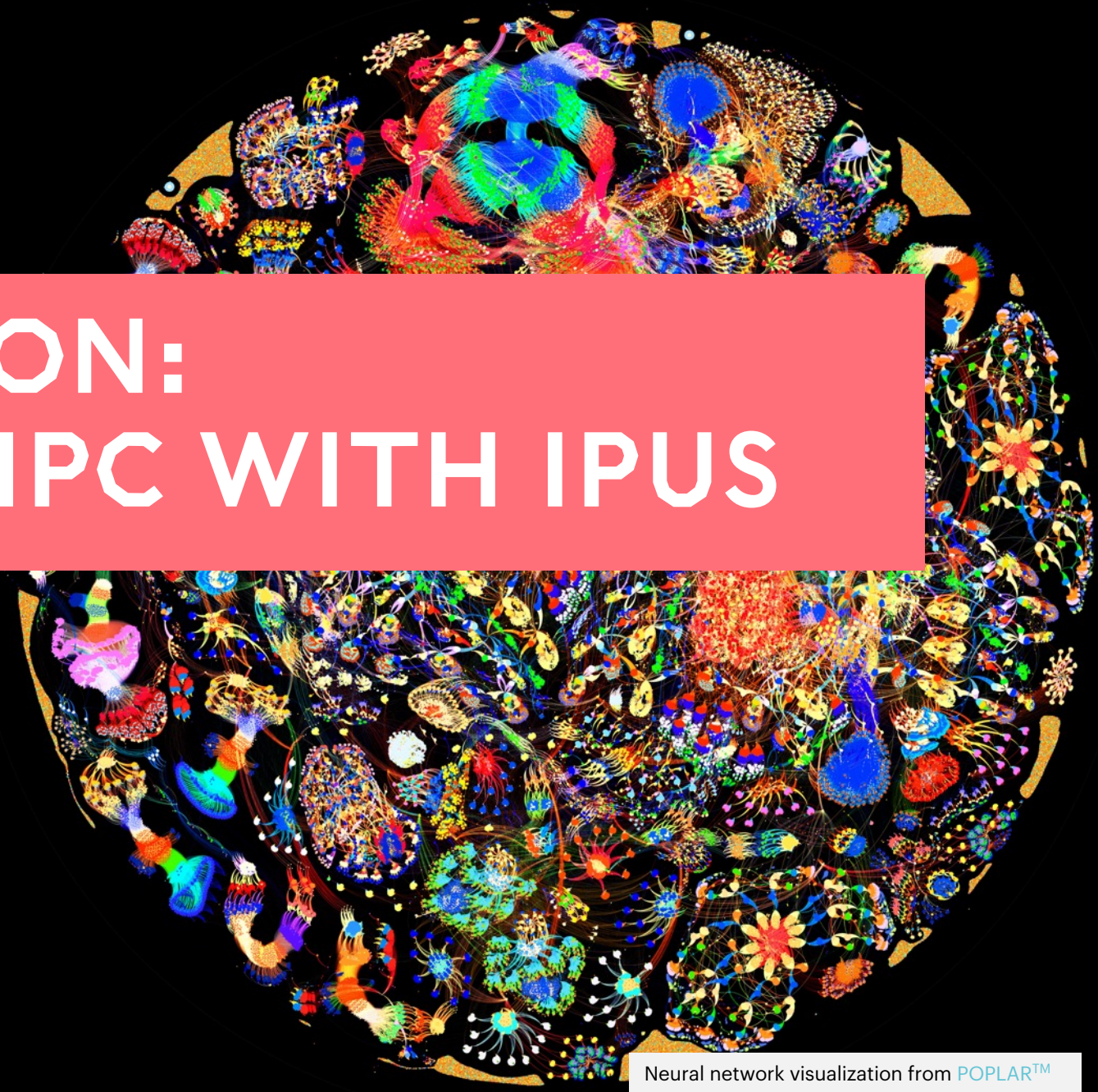
AI FOR SIMULATION: ACCELERATING HPC WITH IPUS



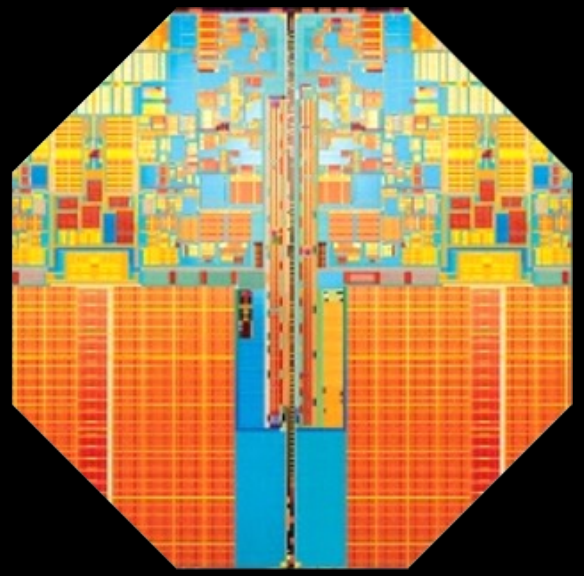
GRAPHCORE

Dr Alex Titterton
Solutions Architect

Neural network visualization from POPLAR™

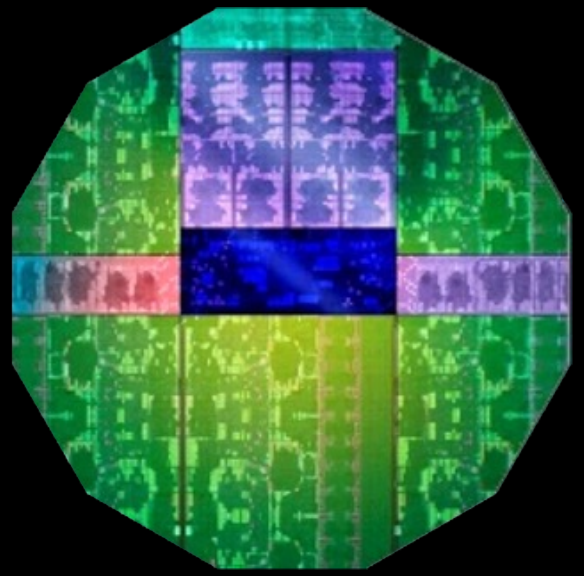
INTELLIGENCE PROCESSING UNIT

DESIGNED FOR AI



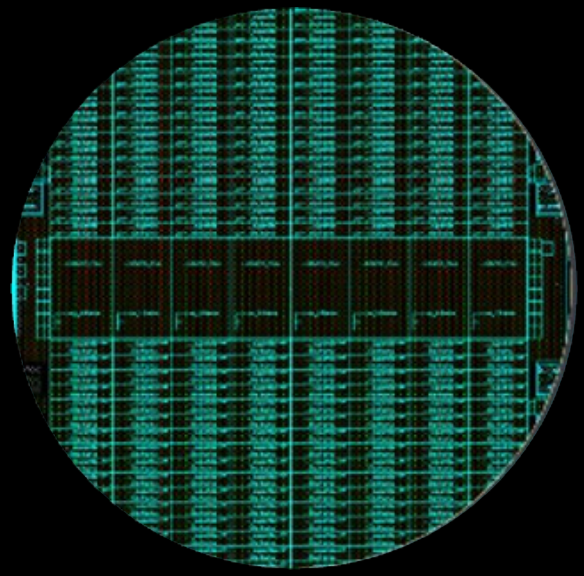
CPU

Scalar



GPU/TPU

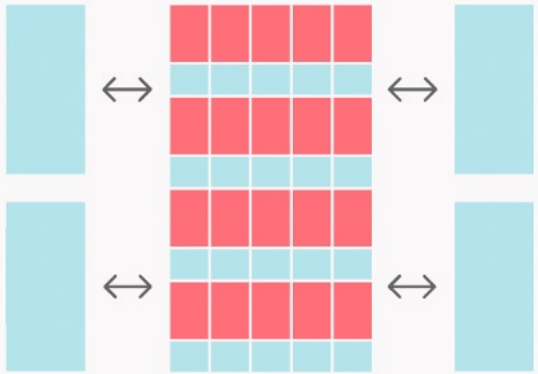
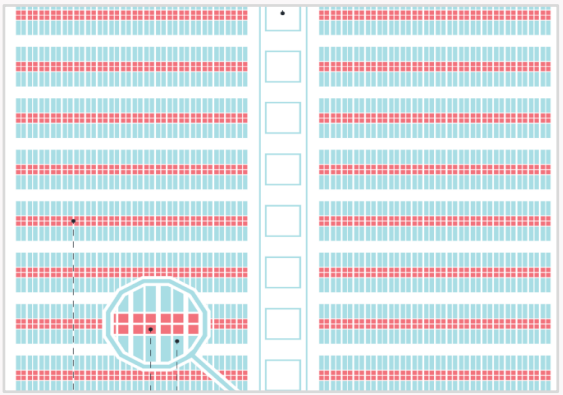
Vector



IPU

Graph

THE INTELLIGENCE PROCESSING UNIT (IPU)
WHAT MAKES IT DIFFERENT?

	CPU	GPU	IPU
Parallelism	Designed for scalar processing	SIMD/SIMT architecture. Designed for large blocks of dense contiguous data	Massively parallel MIMD architecture. High performance/efficiency for future ML trends
			
Memory Bandwidth	Off-chip memory	Model and Data spread across off-chip and small on-chip cache and shared memory (2TB/s for A100 HBM)	Main Model & Data in tightly coupled large locally distributed SRAM (~65 TB/s for Bow IPU)

INTRODUCING THE BOW IPU
WORLD'S FIRST 3D WAFER-ON-WAFER PROCESSOR

3D silicon wafer stacked processor

350 TeraFLOPS AI compute

Optimized silicon power delivery

0.9 GigaByte In-Processor-Memory @ 65TB/s

1,472 independent processor cores

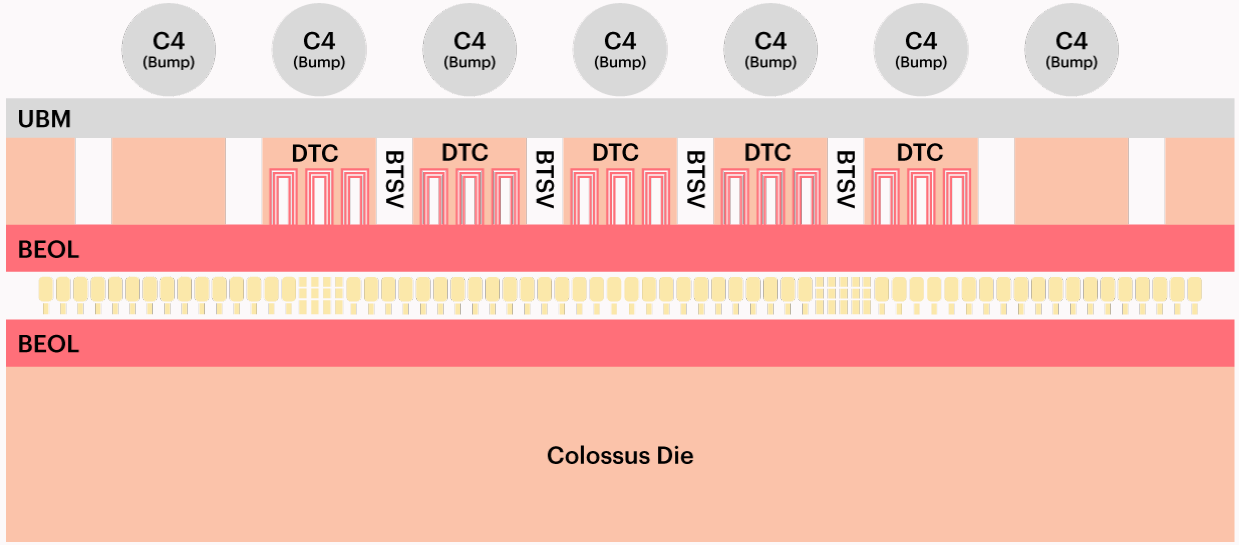
8,832 independent parallel programs

10x IPU-Links™ delivering 320GB/s

GRAPHCORE

BOW IPU: 3D WAFER-ON-WAFER PROCESSOR

GRAPHCORE + tsmc



Advanced silicon wafer stacking technology co-developed between Graphcore and TSMC

World's first commercial deployment using TSMC SoIC-WoW™ technology in Bow IPU

Enabling technology for closely coupled power delivery die to maximize application performance

GRAPHCORE SOFTWARE

NLP/TRANSFORMERS
IMAGE CLASSIFICATION/CNNS
OBJECT DETECTION
LARGE MODELS
MLPERF
CONDITIONAL SPARSITY
GNNS

ML APPLICATIONS

TUTORIALS
CODE EXAMPLES
DOCUMENTATION
VIDEOS
NATIVE IPU CODERS PROGRAM
APPS PORTFOLIO

DEVELOPER ECOSYSTEM

FRONTENDS
FRAMEWORKS
FW BACKENDS
POPLAR®
DRIVERS

JUPYTER NOTEBOOKS
INFERENCE DEPLOYMENT TOOLKIT
ONNX
HALO
Keras
PaddlePaddle
XLA
POPART+
POPDIST
PARTITIONER
POPIR
POPIT
POPLIBS
GCL
POPLAR
GRAPH ENGINE
GRAPH COMPILER
GC DEVICE ACCESS LAYER
IPUOF DRIVER
PCIe DRIVER

POPLAR® SDK

GRAPH ANALYZER
SYSTEM ANALYZER
DEBUGGER
DEVELOPMENT ENVIRONMENT

POPVISION TOOLS

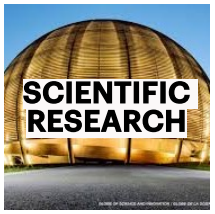
V-IPU

SYSTEM MONITORING
PROMETHEUS
GRAFANA

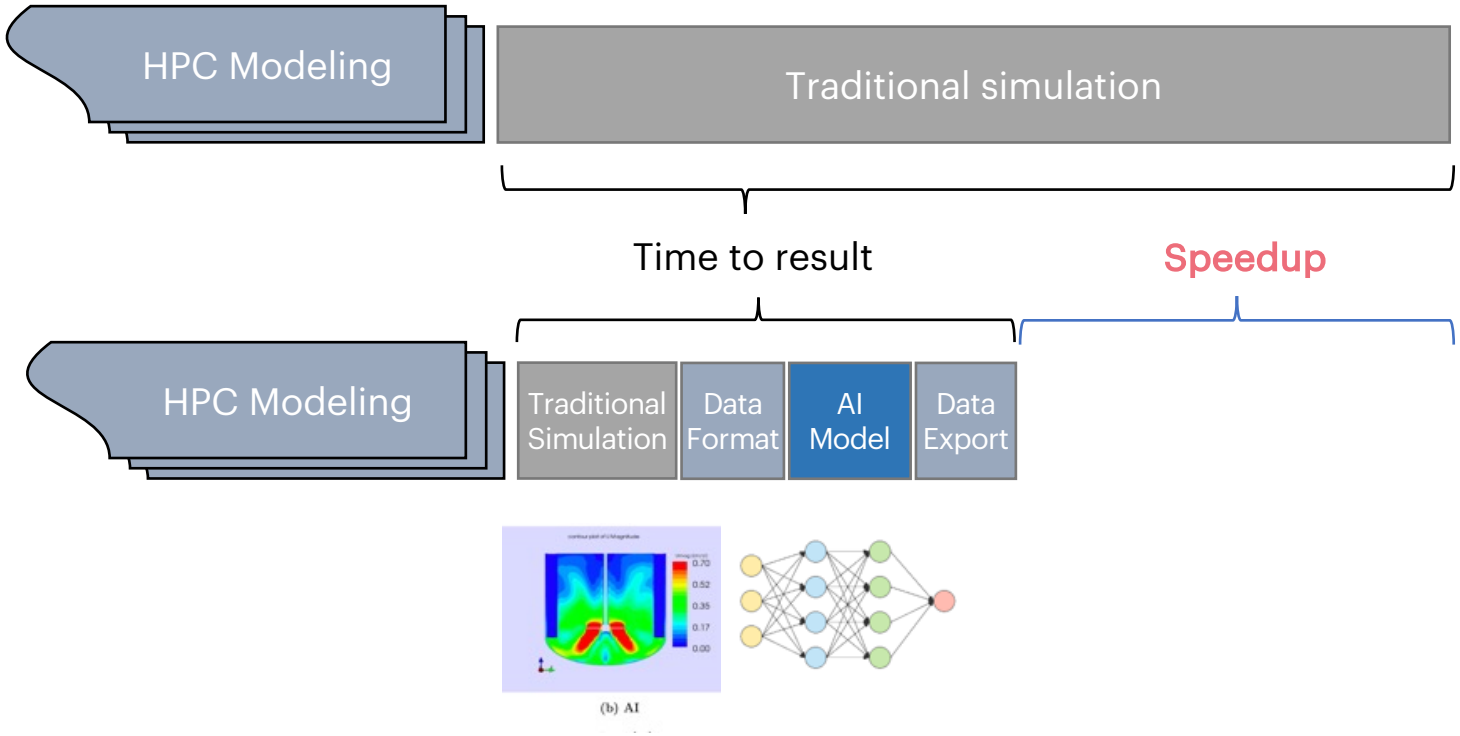
JOB DEPLOYMENT
K8S
SLURM

SYSTEM SOFTWARE

AI FOR SIMULATION

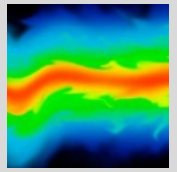


Using AI to accelerate HPC Scientific Applications

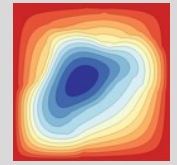


Relevant Application Areas

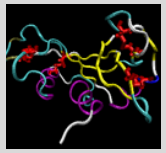
High Energy Physics



Computational Fluid Dynamics

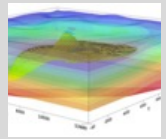


Partial Differential Equations



Protein Folding

Weather Forecasting



Oil & Gas Exploration Simulation

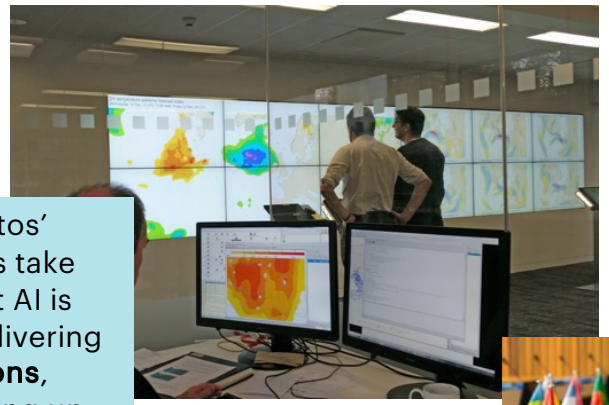
Accelerating HPC with AI: Dramatic Weather Forecasting Accelerations with IPU

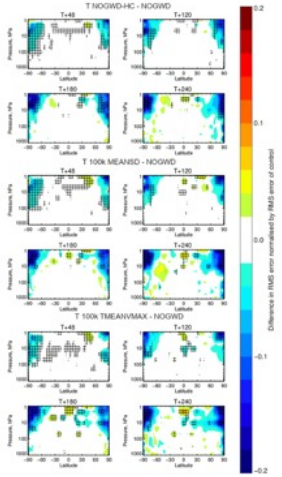
European Centre for Medium-range Weather Forecasting

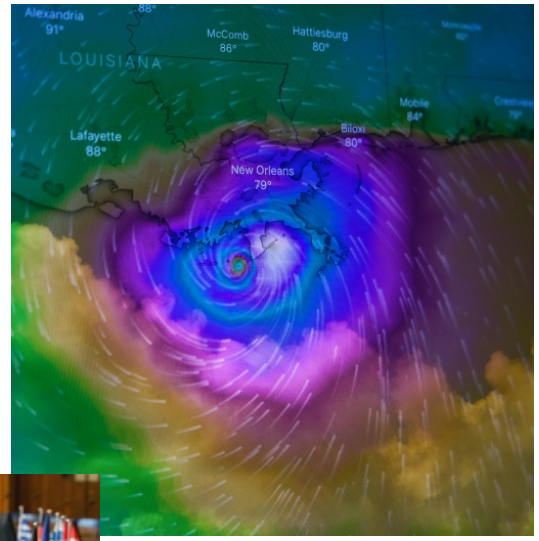
Graphcore IPU trained an ECMWF weather forecasting model **5x faster than a leading GPU (and potentially up to 50x faster than CPUs)**

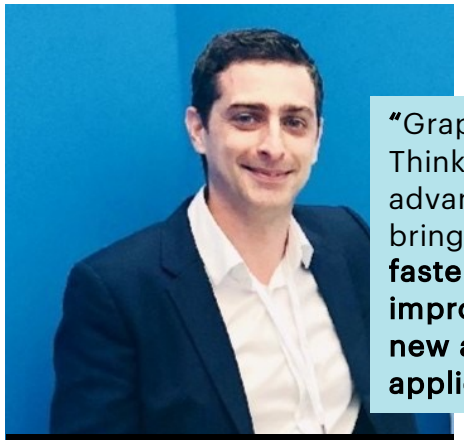
Atos

ECMWF









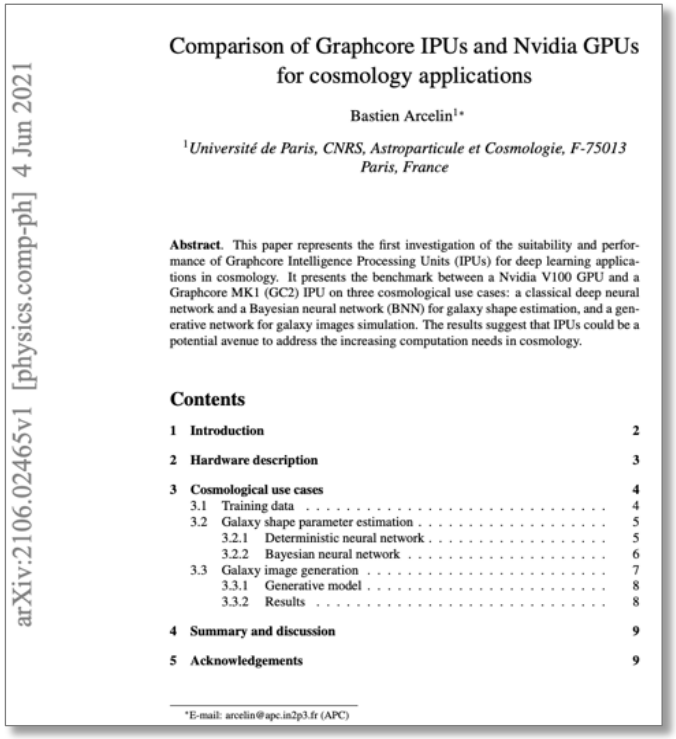
Cedric Bourrasset
Head of HPC AI at Atos

"Graphcore plays a central role in Atos' Think AI solution, helping customers take advantage of the many benefits that AI is bringing to HPC – whether that's delivering **faster and more accurate simulations**, **improving cost efficiency**, or opening up **new areas of research and commercial applications**."

https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002477
https://www.graphcore.ai/posts/climate-change-foreseeing-the-unexpected-with-graphcore-ipus

IPU ACCELERATION FOR COSMOLOGY APPLICATIONS

Université de Paris



arXiv:2106.02465v1 [physics.comp-ph] 4 Jun 2021

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."

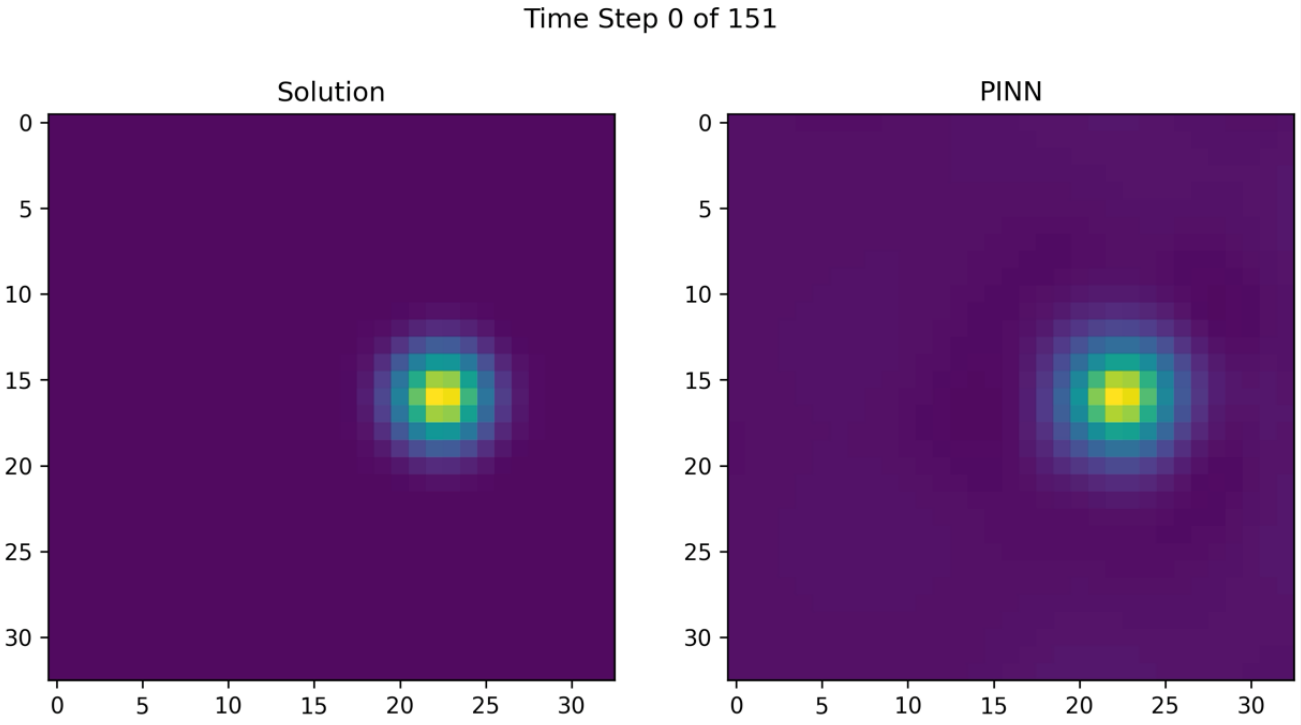
The results show that **IPUs can accelerate various cosmology applications, outperforming GPUs in some cases by as much as 4x faster time to train**"

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

Science and Technology Facilities Council
Hartree Centre

Physics-Informed Neural Networks (PINNs)

UK Atomic Energy Authority



RESULTS ORIGINALLY SHOWN AT SC22, SOLVING A 2D WAVE EQUATION USING A PHYSICS-INFORMED NEURAL NETWORK IMPLEMENTED IN TENSORFLOW 2.

2XIPU FOUND TO BE 11X FASTER THAN 1XA100 GPU, AT SIMILAR MONETARY & ENERGY COST.

WORK DONE IN COLLABORATION WITH STFC HARTREE AND THE UK ATOMIC ENERGY AUTHORITY

Platform	Time to Train / seconds (20k Epochs)	Speedup vs GPU
2 Bow IPUs	41	11x
A100 GPU	530	-

GRAPHCORE

SchNet GNN

Modelling Quantum Interactions in Molecules

Pacific Northwest
NATIONAL LABORATORY

Try on Paperspace

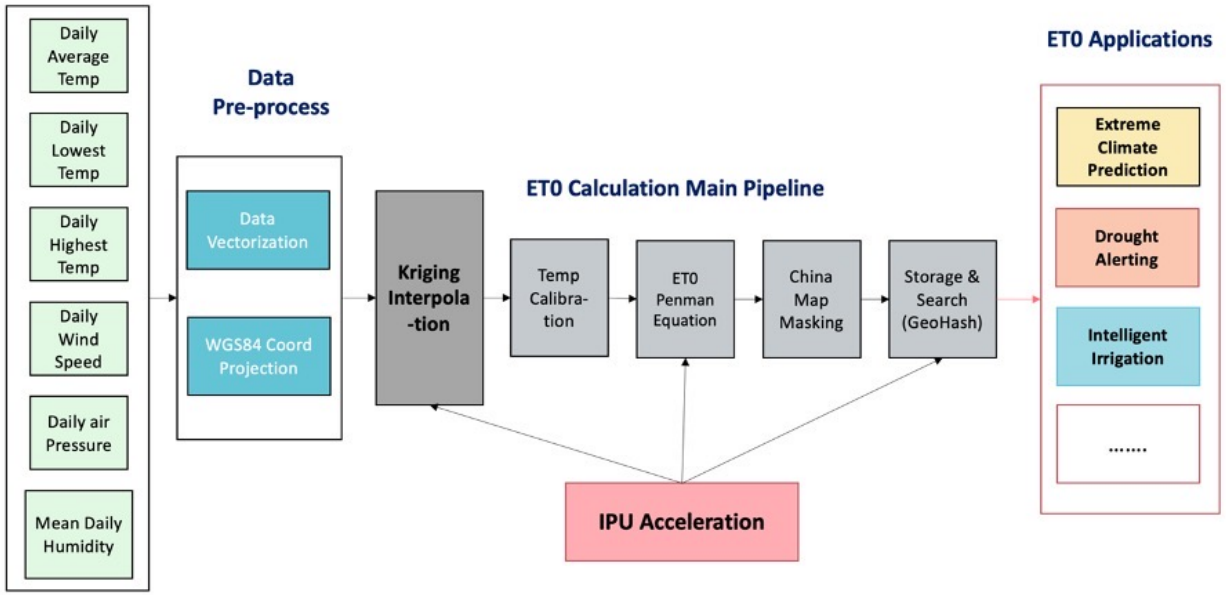
Graphcore engineers successfully trained the **SchNet**[1] model on IPUs on the **500k water clusters** dataset[2], to predict the **potential energy per cluster**.

Preliminary results show a time-to-train of **98 minutes** on 2xIPU-M2000, compared with >**60 hours** on 4xV100 GPUs in PNNL's original paper[2].

[1] JK. T. Schütt1, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. "SchNet – A deep learning architecture for molecules and materials" *J. Chem. Phys.* **148**, 241722 (2018).

[2] Jenna A. Bilbrey, Joseph P. Heindel, Malachi Schram, Pradipta Bandyopadhyay, Sotiris S. Xantheas, and Sutanay Choudhury. "A look inside the black box: Using graph-theoretical descriptors to interpret a Continuous-Filter Convolutional Neural Network (CF-CNN) trained on the global and local minimum energy structures of neutral water clusters" *J. Chem. Phys.* **153**, 024302 (2020).

Agilor: Using AI for Accurate Climate Modelling

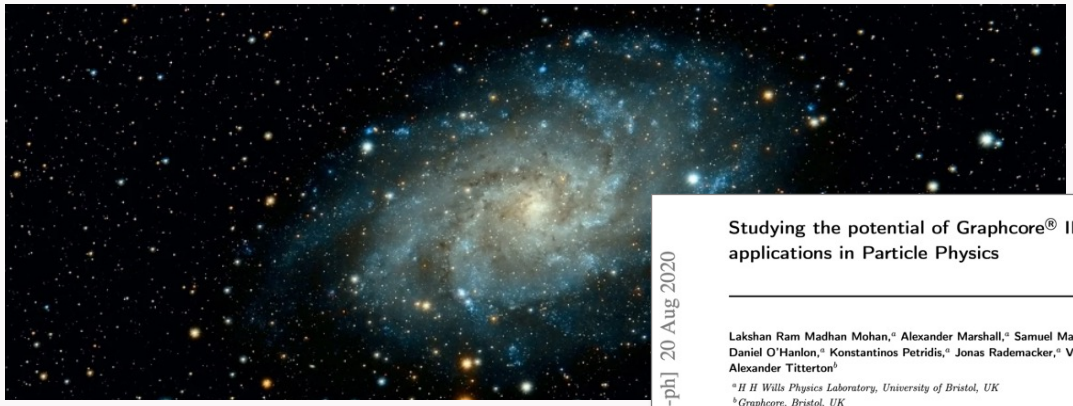


Graphcore engineers have been working with Chinese digital transformation specialists **Agilor**, modelling **evapotranspiration**; the rate at which water moves from surfaces such as plants and soil into the atmosphere.

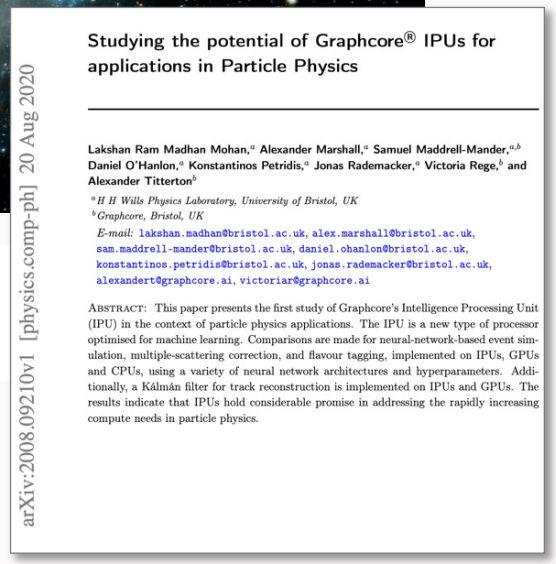
A computationally intensive interpolation technique called Kriging is used in order to give finer-grained approximation results.

Porting PyKrige to the IPU using TensorFlow enabled the interpolation of the entire dataset to be performed in just 21 seconds, compared with 2000 seconds using PyKrige on CPU.

IPUs in Research



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



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

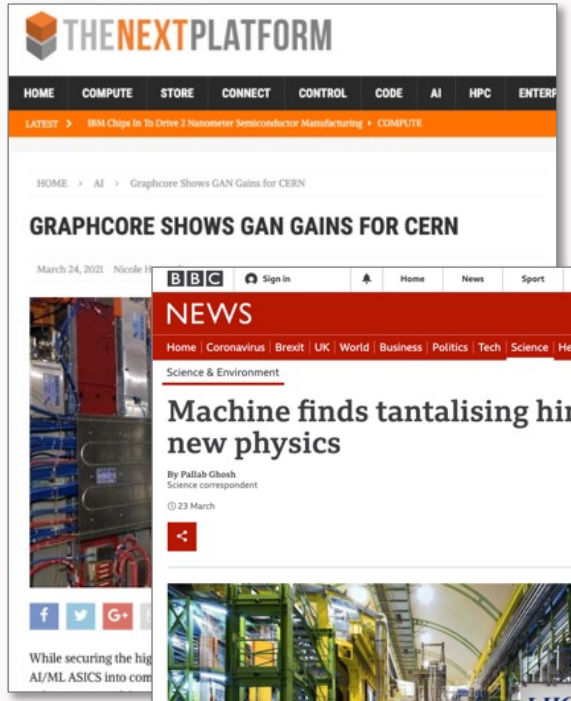
Studying the potential of Graphcore® IPUs for applications in Particle Physics

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

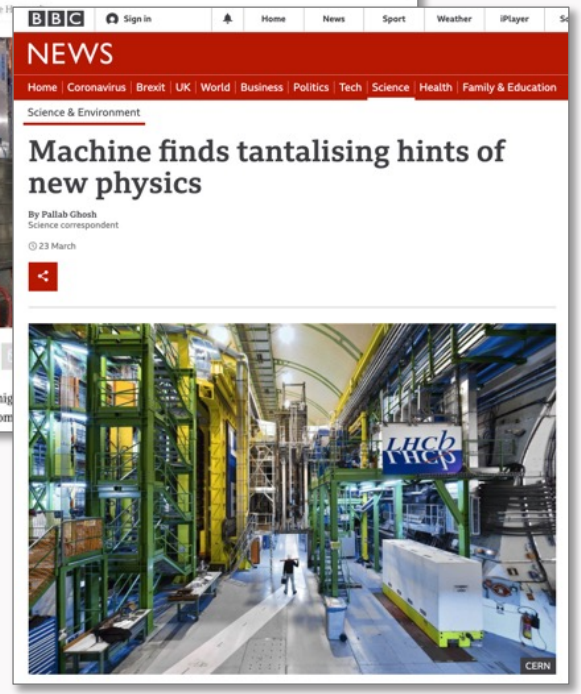
LATEST > IBM Chips In To Drive 2 Nanometer Semiconductor Manufacturing > COMPUTE

HOME > AI > Graphcore Shows GAN Gains for CERN

GRAPHCORE SHOWS GAN GAINS FOR CERN

March 24, 2021

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

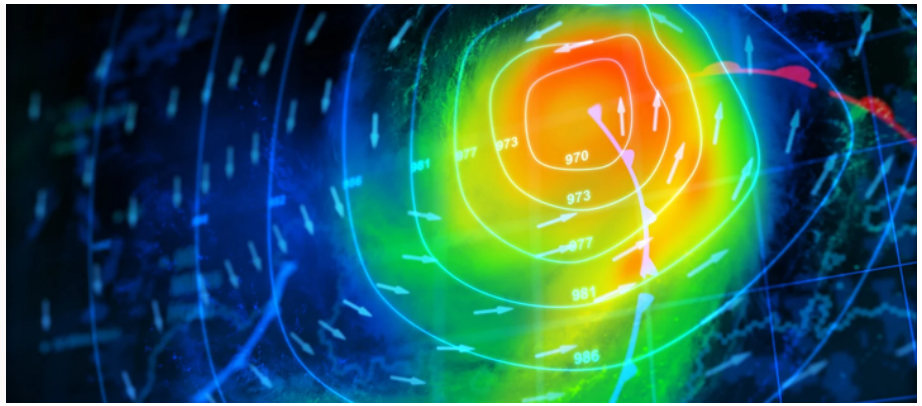
CERN

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

Using AI to accelerate HPC Scientific Applications

Mar 09, 2022 \ AI, HPC

AI FOR SIMULATION: HOW GRAPHCORE IS HELPING TRANSFORM TRADITIONAL HPC



Written By:
Alex Titterton

SHARE:

SUBSCRIBE

Get Updates

For many years High Performance Computing (HPC) techniques have been used to solve the world's most complex scientific problems across a wide range of applications, from modelling Higgs boson decay at the Large Hadron Collider to using Monte-Carlo simulation to predicting whether the weather will improve.

However, due to the immense complexity of the calculations involved in many of these applications, researchers are often waiting a long time for simulation results to arrive. Speeding up these workflows by simply running the same programs on more powerful hardware can be very expensive, with a large cost often giving only a modest improvement in performance.

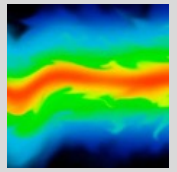
Clearly, a new approach is required to efficiently speed up these workloads, and many researchers are turning to surrogate machine learning models.

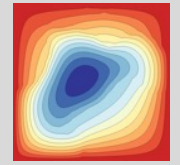
A surrogate model is a machine learning model intended to imitate part of a traditional HPC workflow,

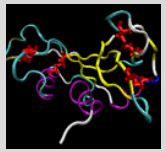
For more information, see our technical blog post:
https://www.graphcore.ai/posts/ai-for-simulation-how-graphcore-is-helping-transform-traditional-hpc

Relevant Application Areas

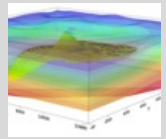

High Energy Physics

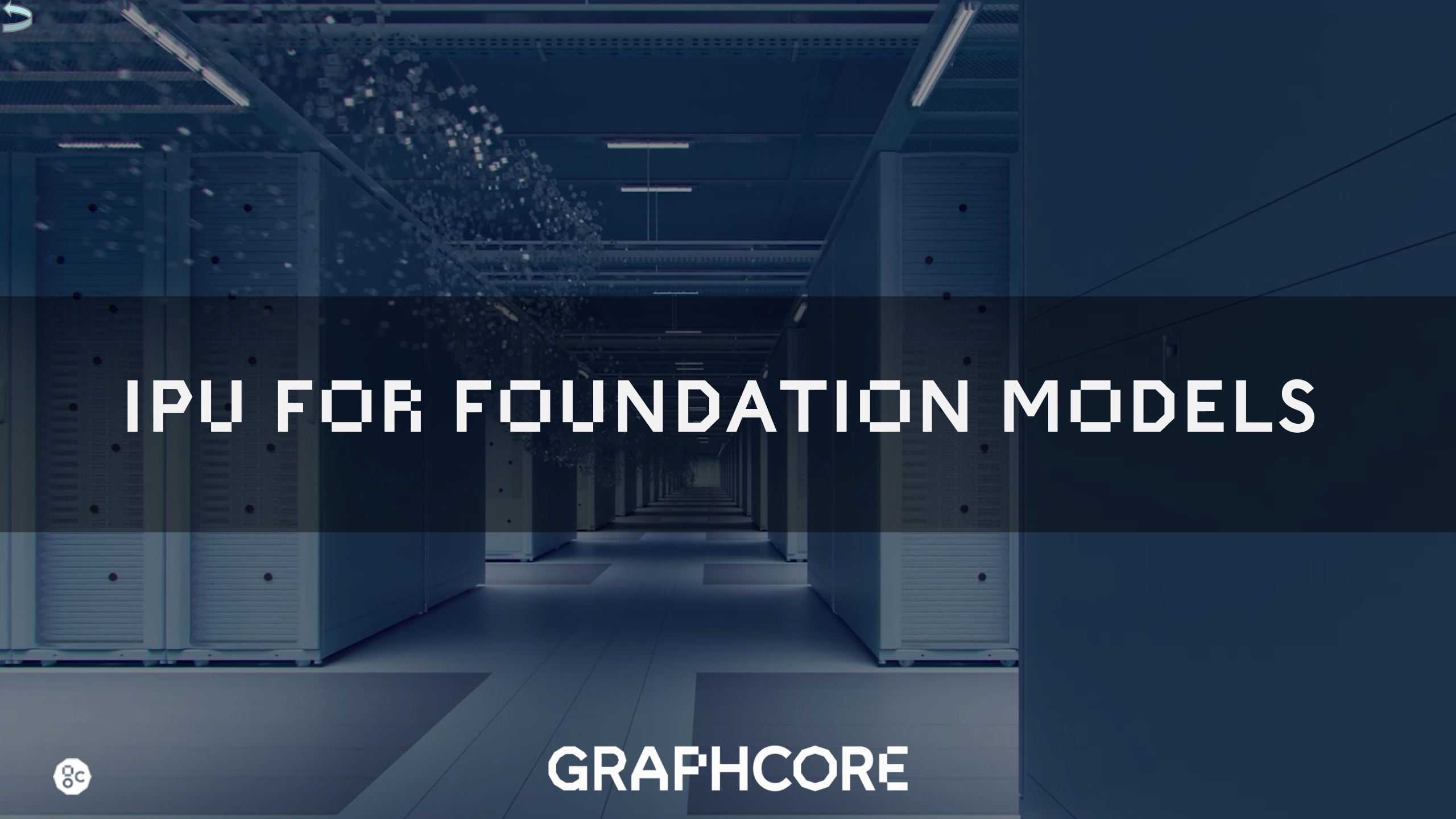

Computational Fluid Dynamics


Partial Differential Equations


Protein Folding

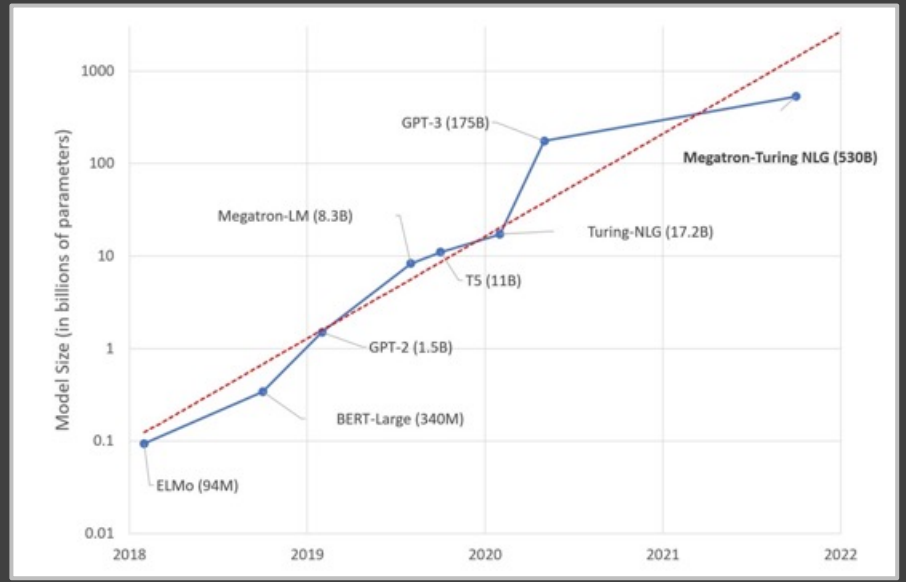

Weather Forecasting


Oil & Gas Exploration Simulation


IPU FOR FOUNDATION MODELS
GRAPHCORE

FOUNDATION MODEL TRENDS

- Models are getting much bigger to deliver ever higher demands on improved accuracy & performance
  - This growth is exponential for dense models
- Multimodal models broaden the learning capability by incorporating different modalities (e.g. linguistic, visual, aural)
  - => larger model demands
- Larger dense models mean more compute, more power, more cost
- Counter to this are economic and societal drivers to reduce energy consumption & cost

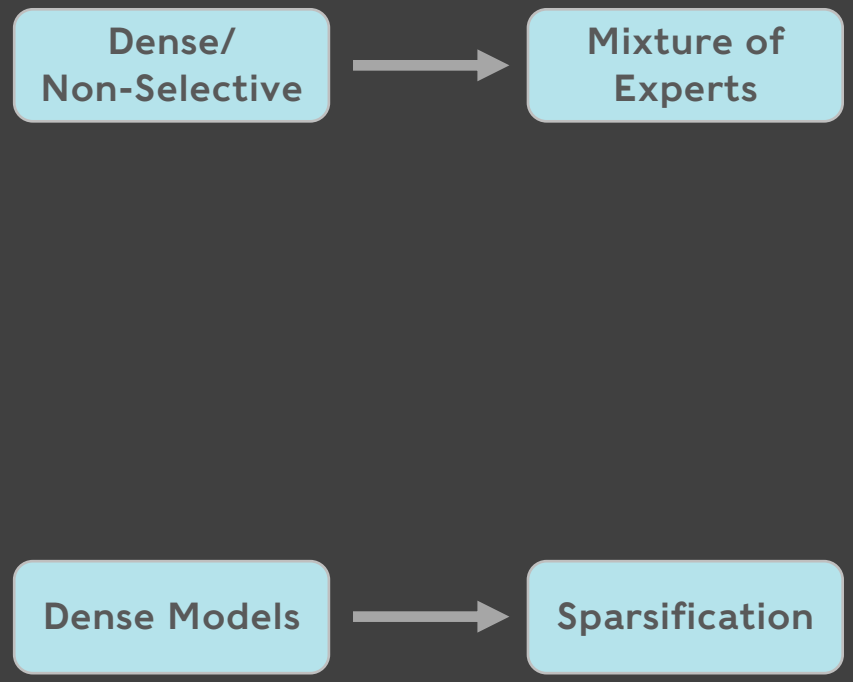


Exponential trend of SOTA NLP models:
Source: Microsoft/NVIDIA https://arxiv.org/abs/2201.11990

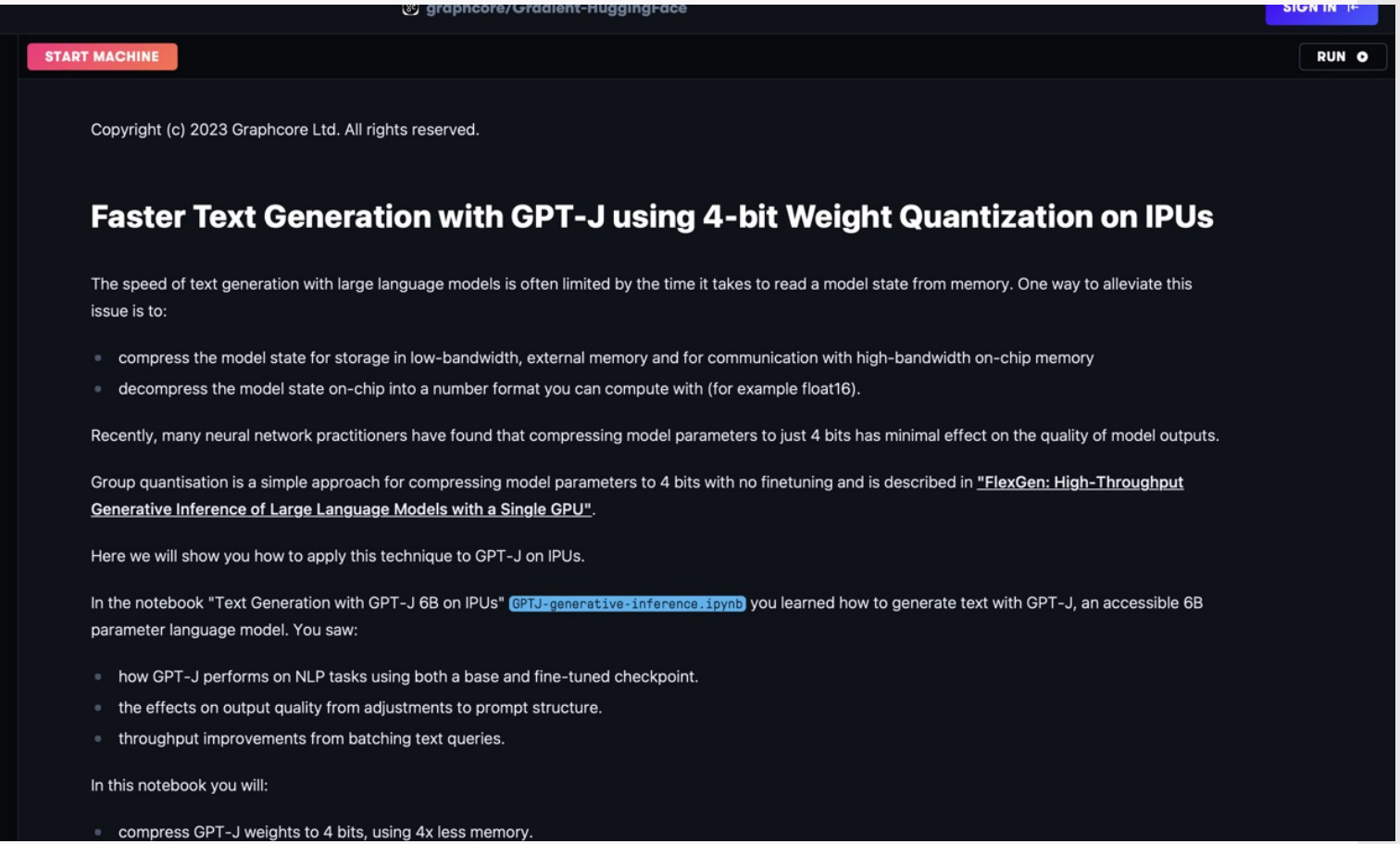
IMPROVING MODEL EFFICIENCY

- Selectivity / Conditional Models
  - Models need to become selective (or conditional), such as Mixture of Experts (MoE) based models
  - Different parts of models are only used when needed
  - This can help reduce compute growth to linear instead of exponential

- Sparsification of models
  - Only incur cost of compute when required
    - ➢ Lower memory requirement
    - ➢ Fewer multiplications
    - ➢ Lower power



QUANTISATION



BLOG:

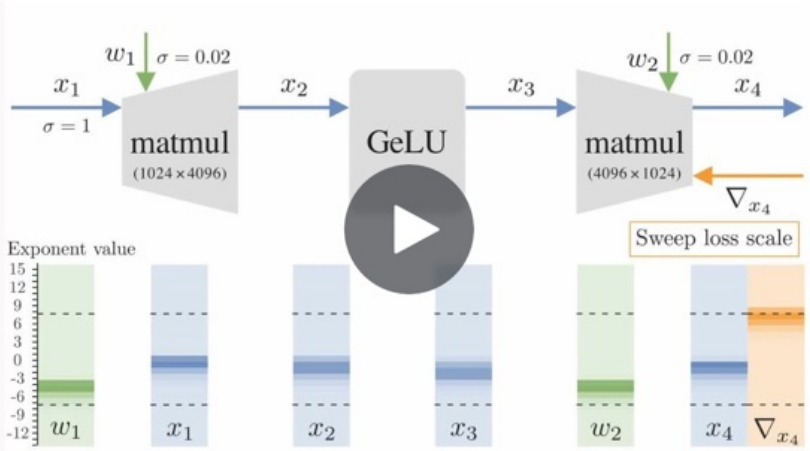
CODE EXAMPLE:

UNIT SCALING FOR TRAINING WITH FP8

Mar 29, 2023

SIMPLE FP16 AND FP8 TRAINING WITH UNIT SCALING

Written By:
Charlie Blake



Popular Topics

- Announcements
- Developer
- Research
- Tutorials
- Case Studies
- Computer Vision
- Natural Language Processing
- Graph Neural Networks
- Generative AI

View all posts →

In recent years the deep learning community has transitioned from the FP32 number format to the FP16 and BFLOAT16 formats. This has led to substantial reductions in memory, bandwidth, and compute requirements-all of which are essential to the trend of increasingly large models.

Now, with the development of FP8-supporting hardware (such as the Graphcore IPU Bow processor used in the C600 PCIe card) further low-precision efficiency savings are possible. However, so far these smaller, low-precision formats have not always been easy to use in practice. With FP8 this may become harder still.

The most significant challenge is that these smaller formats often limit users to a narrower range of representable values. The question thus arises: how do we ensure that our models stick within the range of smaller formats? To address this, Graphcore Research has developed a new method, which we name unit scaling.

BLOG:

HOW-TO GUIDE:


THE ‘GOOD’ COMPUTER
GRAPHCORE

ROADMAP TO ULTRA-INTELLIGENCE AI

Human brain has around 100 billion neurons

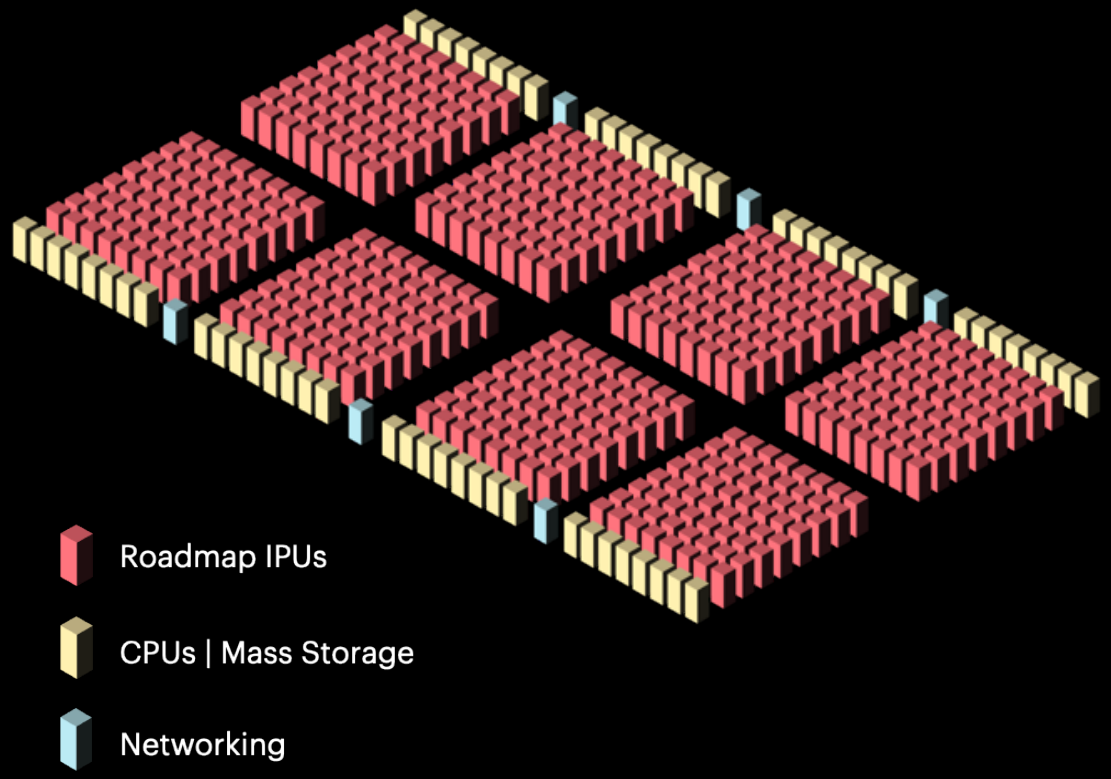
With 100Tn+ synapses, equivalent to parameters in an AI model

Current largest AI models are around 1Tn parameters

Graphcore is developing an Ultra-Intelligence Machine
that will surpass the parametric capacity of the brain

GRAPHCORE

THE 'GOOD' COMPUTER



Over 10 **Exa-Flops** of AI floating point compute from 8,192 roadmap IPUs

3D Wafer-on-Wafer logic stack

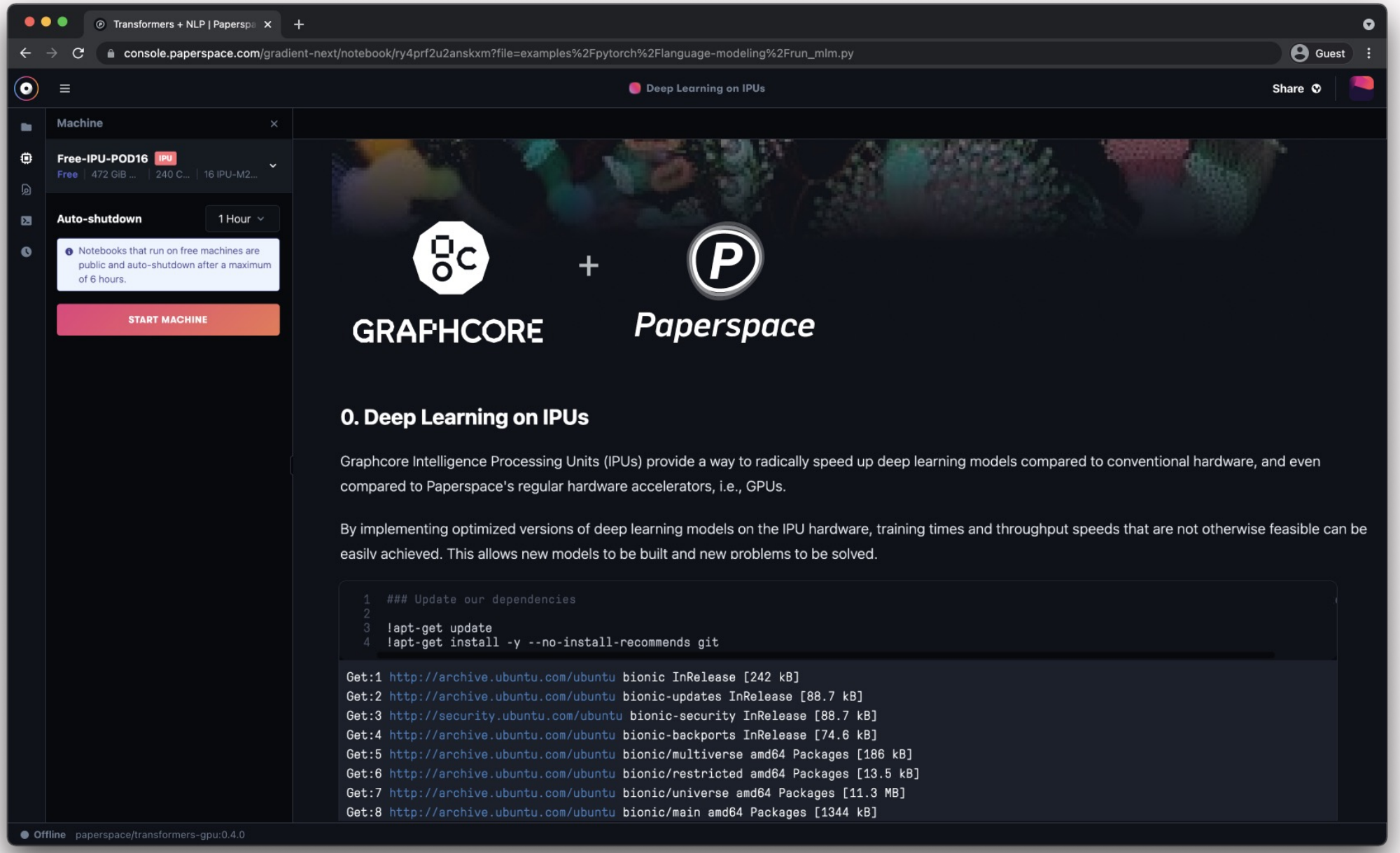
Up to 4 PB of memory with bandwidth of over 10 PB/s

Enabling AI models to be developed with 500 Tn parameters

Fully supported by Poplar® SDK

GRAPHCORE

IPUS IN THE CLOUD



Paperspace

Free 6-hour
IPU Access:

GRAPHCORE ACADEMIC PROGRAMMES

Graphcore's Academic Programme provides tools and technology to educators, researchers and students to enable ground-breaking research and innovation in higher education.

Apply at:
graphcore.ai/academics

MACHINE INTELLIGENCE ACADEMY

The Machine Intelligence Academy is designed to support professors, researchers, and students working in advanced AI and machine learning fields to accelerate their research and develop teaching on IPUs.

Learn more →

ACCELERATED COMPUTING ACADEMY

The Accelerated Computing Academy is aimed at C++ computer scientists looking to build new applications that transcend today's machine learning, ranging from HPC to simulation modelling, and beyond.

Learn more →

THANK YOU

Dr Alexander Titterton, AI Solutions Architect

alexandert@graphcore.ai