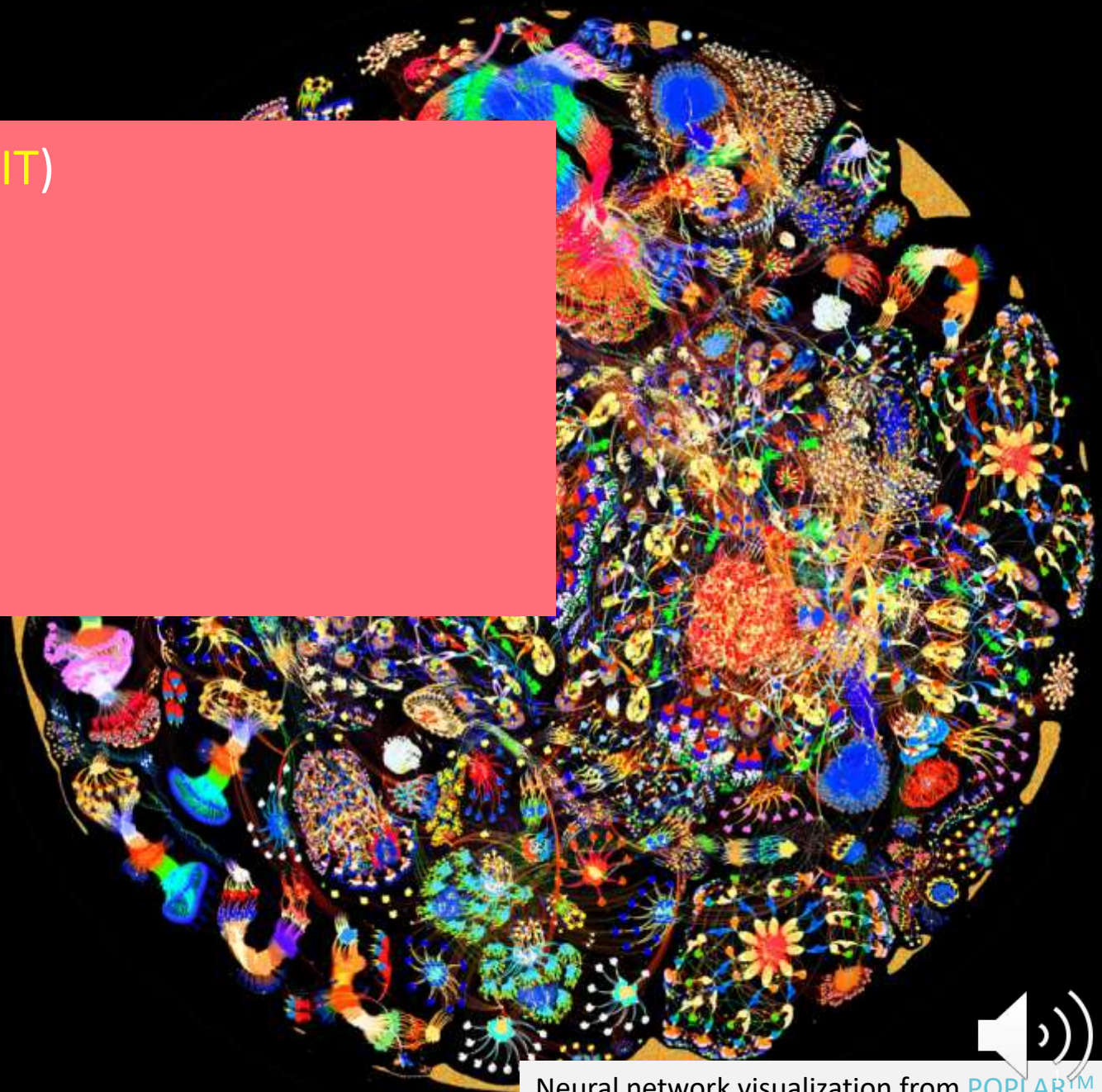
GRAPHCORE **IPU** (INTELLIGNCE PROCESSING UNIT)
IPU-POD: NEW BREAKTHROUGHS IN AI AT SCALE

AI専用　革新プロセッサIPU

中野　守 mamorun@graphcore.ai
Jan 20 2022
Graphcore Japan KK

GRAPHCORE
www.graphcore.ai



Neural network visualization from POPLAR™

GRAPHCORE

2018 July 第一世代 IPU
Begins to ship the
PCI card.

IPU GC200

IPU M2000

IPU POD16

2020 12月～ 第2世代IPU

IPU POD16 POD64/128/256

EXA POD

グラフコア・ジャパン株式会社
東京都千代田区
内幸町1-1-1
日比谷帝国ホテルタワー15F

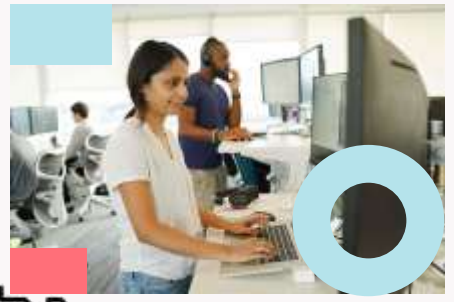
CEO Nigel Toon CTO Simon Knowles

2013 Q4
GC project began in stealth mode
to design the 1st Gen Chip in Bristol.

2016 June Graphcore is officially founded as a company
2021 Jul over 620 employee, HQ Bristol, United Kingdom

2016 Oct
Series A Funding
2020 Dec Series E
Total $730M IN FUNDING

GRAPHCORE OFFICES



OSLO

GDANSK

Paris

Munich

BRISTOL

Cambridge

London

PALO ALTO

Seattle

Austin

New York

BEIJING

Shanghai

Shenzhen

Seoul

Tokyo

India Pune

HSINCHU

Singapore

Sales

R&D

Operations

>620 PEOPLE

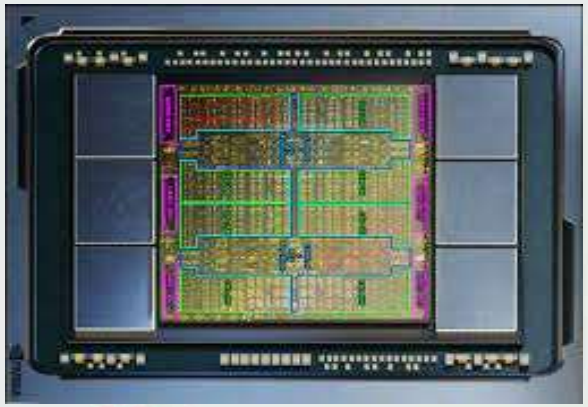
GRAPHCORE

AGENDA　AI専用　革新プロセッサIPU

1. AI専用プロセッサ開発の背景
2. IPUおよびIPU-POD概要
3. 開発環境、PODPLAR SDKその他
4. USECASES, LARGE MODEL、EVAL/POC

＜プログラム＞
13:00-13:05　挨拶
　　佐藤三久（PCCC会長／HPC-OSS部会長／理研）
13:05-13:20　開催趣旨説明
　　滝沢寛之（東北大学）
13:20-13:45「GPUとDPUによるアクセラレーテッドコンピューティング」
　　佐々木 邦暢(エヌビディア合同会社)
13:45-14:10「AMD Instinct GPU and ROCm technical overview」
　　Greg, Oakes (AMD Sr. Solution Architect)
14:10-14:35「AI革新プロセッサIPU」
　　中野　守（Graphcore Japan KK）
14:35-14:45　(休憩)
14:45-15:10「インテルXPU戦略」
　　大内山 浩、髙藤良史(インテル株式会社）
15:10-15:35「ハイブリッドアクセラレータ　SX-Aurora TSUBASA」
　　政岡靖久（日本電気株式会社）
15:35-16:00「SambaNova Systemsのご紹介」
　　鯨岡俊則（SambaNova）
16:00-16:10　(休憩)
16:10-17:30　パネルディスカッション
　　「アクセラレータの王者と挑戦者たち」（仮題）
　　モデレータ：滝沢寛之（東北大学）

MACHINE LEARNING REQUIRES COMPUTE INNOVATION

	CPU	>	GPU	>
	General purpose		Graphics Processing Unit Graphics, part of HPC and ML/DL	
Parallelism	Suitable for scalar processes		SIMD architecture. Suitable for large blocks of dense contiguous data	
				
	CPU Scalar Many core+Multi Die Designed for office apps Evolved for web servers		GPU Vector GPU+HBM2e Designed for graphics Evolved for linear algebra	
Memory				
Access ration, MemB				

Processor Memory

Others

FPGA
TPU
IPU
:

AI REQUIRES COMPUTE INNOVATION BY IPU

A NEW APPROACH IS REQUIRED
AIドメインスペシフィックな革新的アーキテクチャ

- AI本格的な応用段階、ML/DLの精度・機能・性能向上ニーズが急増
- 大規模モデル　パラメータ数は10年間で30万倍以上増加
- 革新的アルゴリズムの台頭
- TCO, 導入コスト,電力コストを大幅に低減

➔ 今後10年間で1000倍以上の性能向上が必要

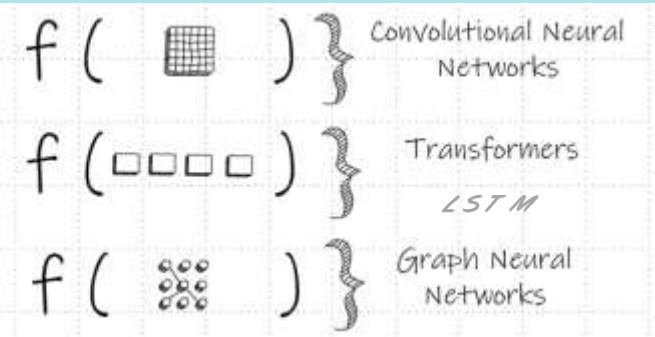
Massive parallelism

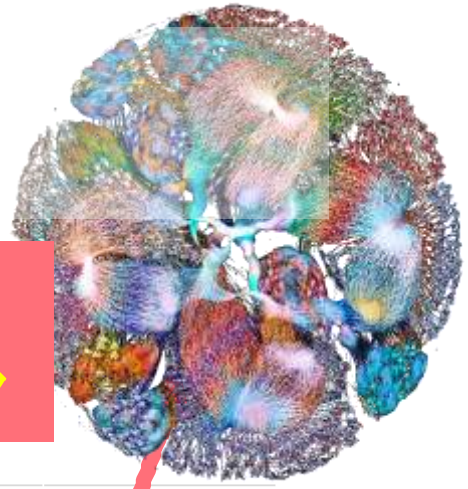
Sparsity in data structures

Low precision compute

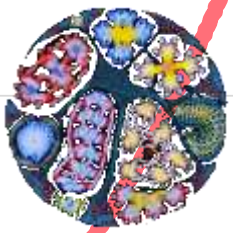
Model parameter re-use

Static graph structure





And more …



GPT3: 175Bn
June 2020
OpenAI が開発

GPT-3 Generative Pretrained Transformer
1750億個のパラメータを使用した文章生成言語モデル. 通常の文章以外にもソースコードやデザインですらGPT-3」は理解し、生成する事が可能といわれている。

BERT-Large: 330M
Oct 2018

ResNet50: 25M
Jan 2016
畳み込みニューラルネットワーク（CNN）
2016

Parameters

2019

Time

日本経済新聞　2021年(令和3年)10月20日(水曜日)

米IT、半導体開発に本腰

アップル・グーグル、新商品に搭載

自社設計が競争力左右

世界のIT（情報技術）大手による半導体の開発競争が本格化している。米アップルは18日に発表したノートパソコンに自前で設計した半導体を採用する。グーグルも28日発売の新型スマートフォンに自社開発の半導体を搭載する。部品である半導体が製品そのものの競争力を左右するようになり、自動車や通信でも自前の開発能力を備える動きが広がる。

アップルは新型パソコン「MacBook Pro」向けに2種類の半導体を自社で設計開発した。2020年に発表した「M1」に続く独自半導体だ。処理性能は最大1・7倍で、動画にCG（コンピューターグラフィックス）を重ね合わせるような操作が容易になる。最先端の回路線幅5ナノ（ナノは10億分の1）の半導体製造プロセスで生産する。

グーグルも日米などで発売する「Pixel（ピクセル）6」に自社開発の半導体「グーグル・テンソル」を採用し、画像処理や翻訳などの機能を高める。例えば、撮影した画像から不要な物体を消去できる「消しゴムマジック」と呼ぶ機能を搭載する。

自動車や通信も

グーグルはこれまで米クアルコムの半導体を使ってきた。ハードウエアを担当するリック・オステロー上級副社長は今夏、日本経済新聞などの取材で「人工知能（AI）で新たな技術革新が起きており、汎用的な市販品では不十分と感じるようになってきた」と半導体の独自開発に乗り出した理由を説明している。

これまで半導体の設計、開発機能を担ってきたのは、米インテルのような垂直統合型の半導体企業や、クアルコム、米エヌビディアといった半導体設計に特化した企業だった。設計開発に専門知識が必要な半導体は外部企業から調達する部品という性格が強く、IT大手が自前で開発や生産能力を持つ必要はなかった。

しかし、AIや高速通信が普及するなか、製品の頭脳である半導体に求められる性能は急速に高まっている。自社の用途に特化した専用半導体を作れば競合と差異化もできる。新たな使用環境や顧客ニーズに合わせた半導体をどれだけ早く開発し、製品に搭載できるかが競争能力を左右するようになりつつある。

IT大手が半導体に参入できるのは、開発と製造を水平分業するサプライチェーン（供給網）の存在が大きい。00年代に入り設計開発に特化したファブレスメーカーが台頭する一方、巨額の設備投資が必要な製造分野では台湾積体電路製造（TSMC）などの生産受託会社（ファンドリー）が製造技術を磨いてきた。

さらにファウンドリーは半導体の設計機能も内製するようになった。半導体に知見の少ない企業も、独自に先端半導体を開発、搭載できるようになった。こうした自社設計の動きはアップルやグーグルにとどまらない。

自動車ではEV（電気自動車）大手の米テスラが19年以降、車の頭脳部を担う半導体を米エヌビディアから自社設計に切り替えている。トヨタ自動車もデンソーとの共同出資会社ミライズテクノの半導体技術者を巡る獲得競争も過熱しており、企業規模で格差が開く可能性は否定できない。

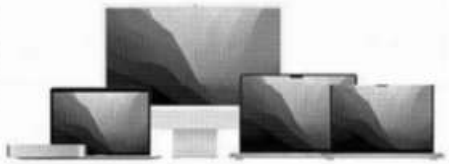
東京大学の黒田忠広教授は「データや演算量の増加をうけ、半導体には高いエネルギー効率が求められている。十分な性能を引き出すには用途に合わせた設計が欠かせない」と指摘する。

コストの重い設計工程を効率化するため、東大はパナソニックやミライズテクノロジーズなどと共同プロジェクトを立ち上げた。省エネ半導体の開発効率を従来の10倍に高めるのが目標だ。

日本の半導体産業は1990年代から衰退が続

自前で半導体を開発する動きが広がる

IT大手	アップル（米）	iPhoneやMacBookPro向けに自前で半導体を開発。5ナノメートルの技術を採用
	グーグル（米）	28日に発売するスマホ「ピクセル6」に独自半導体を採用。機械学習に特化した独自プロセッサーも開発
自動車	テスラ（米）	19年から車載半導体の自社開発品に切り替え
	トヨタ自動車（日）	集積チップやパワー半導体の開発を目的に20年にデンソーと新会社を設立
通信	シスコシステムズ（米）	データセンター向けの機器などで自社開発の半導体を搭載
	ノキア（フィンランド）	通信事業者向けのルーターに独自のプロセッサーを搭載



アップルはMacシリーズで自社半導体の搭載を拡大する

GRAPHCORE IPU（Intelligence Processor Unit）

AI Domain Specific Processor

IPU GC200 (Processor)

TSMC 7nm　823mm2　59.4B Transistors
Many-cores　1472 tiles（Core, In-Proc-Mem）8832 threads
In-proc Mem　Total 900MB/IPU　47.5TB/s Mem BW
Peak Perf (fp16)　250TFlops AI-Float
(fp32)　62TFLOPS（fp32）

M2000

第2世代プロセッサ GC200を4基
1Uの筐体に搭載したAIアクセラレータ

筐体をIPU-Linkで相互接続することで、性能をスケールアップ
IPU環境を容易に使うための POPLAR SDK を提供
PyTorch や TensorFlow 等の主要なフレームワークをサポート

IPU-POD Family

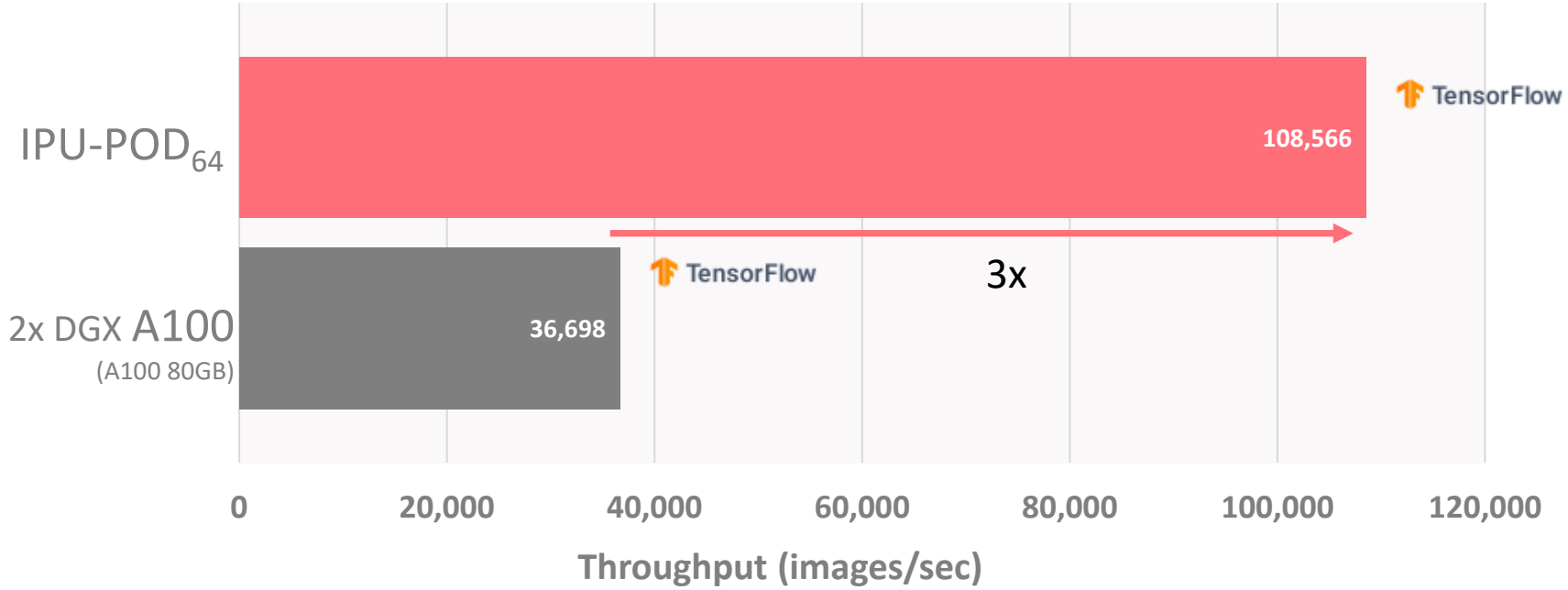
IPU-POD256
IPU-POD16
IPU-POD64

- **1筐体で 1ペタFlops AI-Float(fp16)**

CV　RESNET50 TRAINING

3x Higher Throughput | 3.5x Perf/$ Advantage

同レベルのシステム価格の二つの
システム、IPUとGPUの性能を比較



NOTES: (updated 20th Dec 2021)
ResNet-50 v1.5 Training Highest Throughput | ImageNet2012 Dataset
IPU-POD64 (16x IPU-M2000) | FP 16.16 | SDK 2.4.0 | TensorFlow | https://www.graphcore.ai/performance-results
Best results for DGX A100 640GB (A100-SXM4-80GB) using TensorFlow | Mixed Precision | using 0.90x scaling from 1=>2 DGX A100
DGX A100 results published by NVIDIA (https://developer.nvidia.com/deep-learning-performance-training-inference) on 1 Dec 2021

SUMMARY

MODEL NAME:
RESNET-50 V1.5

ML DOMAIN:
COMPUTER VISION

MODEL DESCRIPTION:
RESNET IS A LEGACY IMAGE CLASSIFICATION MODEL, OFTEN USED AS A PROXY FOR INITIAL PERFORMANCE EVALUATION

DEPLOYMENT:
TRAINING

FRAMEWORK:
TENSORFLOW 1
[TF2 ALSO IN MODEL GARDEN]
[PYTORCH ALSO IN MODEL GARDEN]

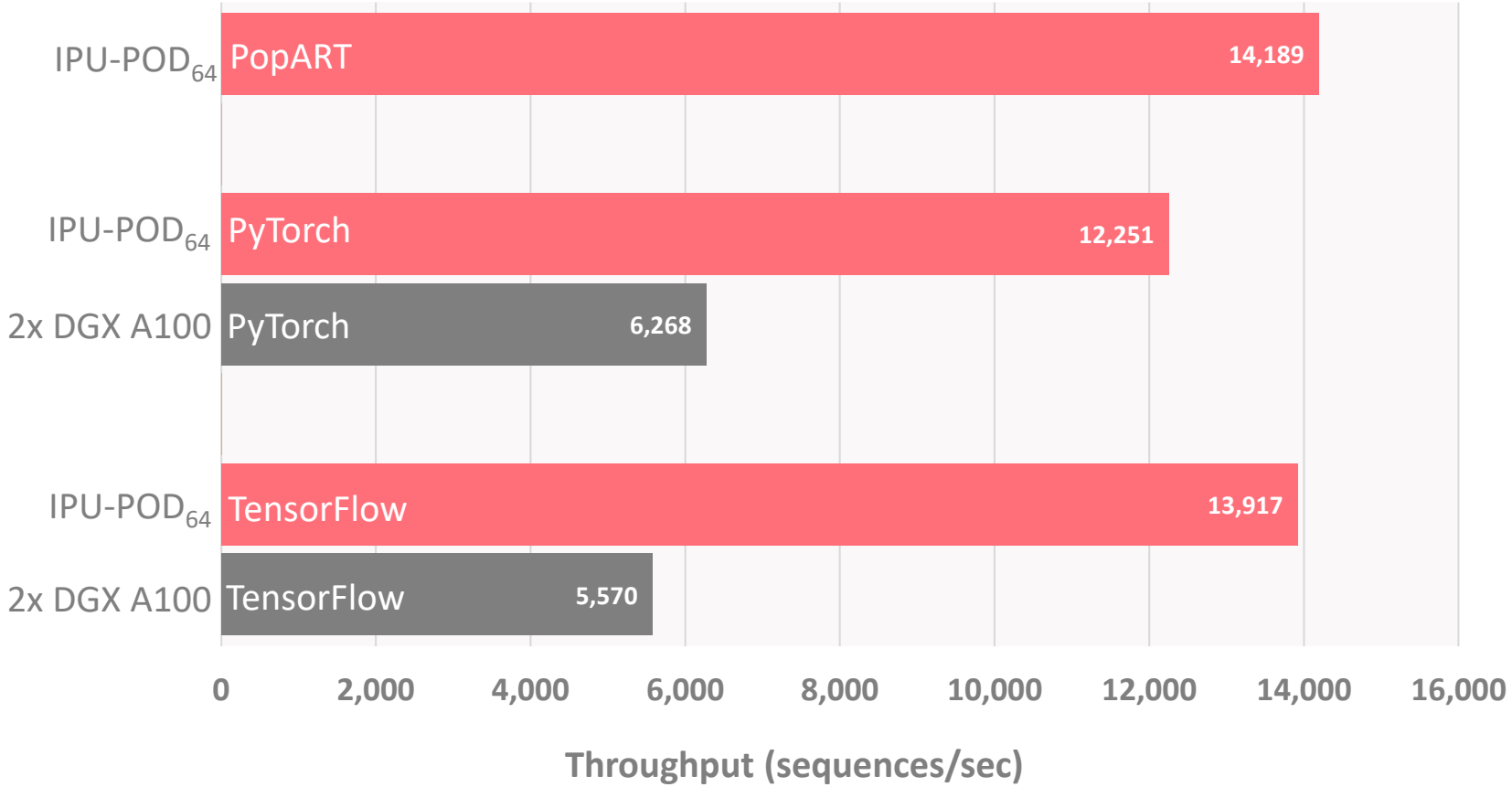
IPU ADVANTAGE:
IPU-POD64 ACHIEVES 3X HIGHER THROUGHPUT VS TWO DGX A100, DELIVERING >3.5X PERF/$ ADVANTAGE

Pricing Assumptions	MSRP
IPU-POD64 (4-host)	~$500k
2x DGX A100 640GB	~$600k

NLP BERT-LARGE PRE-TRAINING

2.3x Higher Throughput | 2x Perf/$ Advantage

同レベルのシステム価格の二つの
システム、IPUとGPUの性能を比較



NOTES: (updated 20th Dec 2021) | headline speedup based on TensorFlow vs TensorFlow throughput comparison
BERT-Large Phase 1 Pre-Training Throughput (SL128)
IPU-POD64 (16x IPU-M2000) | SDK 2.4.0 | https://www.graphcore.ai/performance-results
DGX A100 320GB (A100-SXM4-40GB) | Mixed Precision | Using 97% perf scaling based on published BERT results
DGX A100 PyTorch https://developer.nvidia.com/deep-learning-performance-training-inference 1st Dec 2021
DGX A100 TensorFlow - https://ngc.nvidia.com/catalog/resources/nvidia:bert_for_tensorflow/performance

SUMMARY

MODEL NAME:
BERT-LARGE

ML DOMAIN:
NLP

MODEL DESCRIPTION:
BERT IS AN UNSUPERVISED TRANSFORMER-BASED LANGUAGE REPRESENTATION MODEL. IT IS ONE OF THE MOST WIDELY USED NLP MODELS

DEPLOYMENT:
PRE-TRAINING PHASE 1 (SL128)

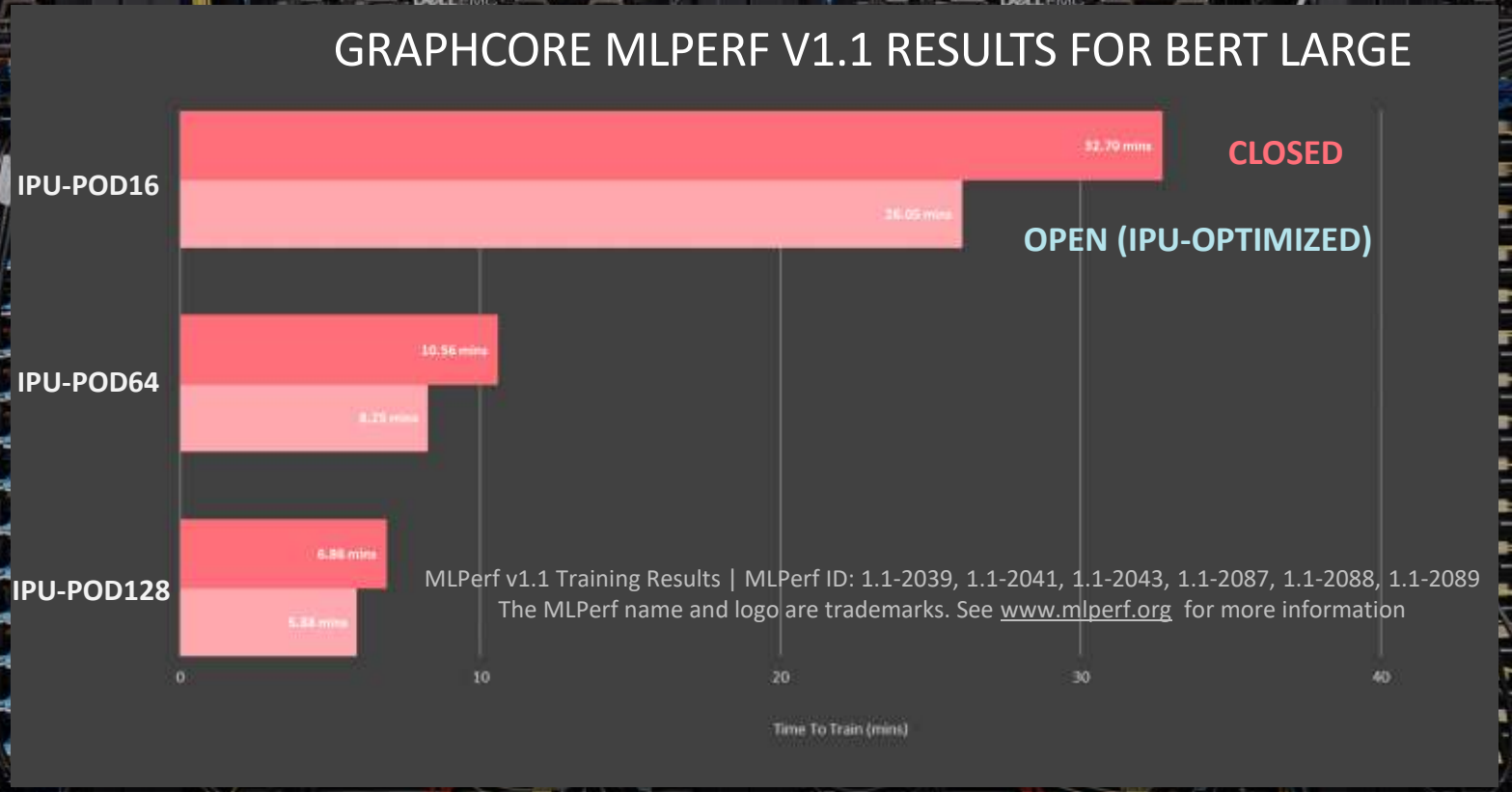
FRAMEWORK:
TENSORFLOW 1
PYTORCH
POPART

IPU ADVANTAGE:
IPU-POD64 ~2.3X HIGHER THROUGHPUT THAN 2X DGX A100 THROUGHPUT, DELIVERING ~2X PERF/$ ADVANTAGE

Pricing Assumptions	MSRP
IPU-POD64 (1-host)	~$440k
2x DGX A100 320GB	~400k

GRAPHCORE

MLPerf



MLPERF 1.1 RESULTS

DECEMBER 2021

既存モデルをIPUアーキテクチャに最適化を施した事例
IPU: トランスフォーマモデルの最適化
2x – 4x 学習時間の高速化 | 3x モデルサイズの縮小

Graphcore Research Grの研究論文

Group Transformer
IPUアーキテクチャを活かしグループ化された畳み込みモジュールを追加
- **3x** パラメータ数の縮小
- **2x** 学習時間の高速化
- 結果の精度向上

https://arxiv.org/abs/2107.02027

Transformer Packing
データ パディングの除去（IPUではなくGPUで必要） 非負の最小二乗法によるヒストグラムのパッキングが 計算効率を大幅に向上。学習時間を2倍に短縮

https://arxiv.org/abs/2106.05822

arXiv:2107.02027v1 [cs.CL] 29 Jun 2021

Packing: Towards 2x NLP BERT Acceleration

Matej Kosec* Graphcore Inc. Palo Alto matej@graphcore.ai

Sheng Fu Graphcore Inc. Palo Alto

Mario Michael Krell* Graphcore Inc. Palo Alto

Abstract

We find that at sequence length 512 padding tokens represent in excess of 50% of the Wikipedia dataset used for pretraining BERT (Bidirectional Encoder Representations from Transformers). Therefore by removing all padding we achieve a 2x speed-up in terms of sequences/sec. To exploit this characteristic of the dataset, we develop and contrast two deterministic packing algorithms. Both algorithms rely on the assumption that sequences are interchangeable and therefore packing can be performed on the histogram of sequence lengths, rather than per sample. This transformation of the problem leads to algorithms which are fast and have linear complexity in dataset size. The shortest-pack-first histogram-packing (SPFHP) algorithm determines the packing order for the Wikipedia dataset of over 16M sequences in 0.02 seconds. The non-negative least-squares histogram-packing (NNLSHP) algorithm converges in 28.4 seconds but produces solutions which are more depth efficient, managing to get near optimal packing by combining a maximum of 3 sequences in one sample. Using the dataset with multiple sequences per sample requires additional masking in the attention layer and a modification of the MLM loss function. We demonstrate that both of these changes are straightforward to implement and have relatively little impact on the achievable performance gain on modern hardware. Finally, we pretrain BERT Large using the packed dataset, demonstrating no loss of convergence and the desired 2x speed-up.

1 Introduction

Since its introduction in 2019, BERT [3] has been the backbone driving the most exciting advances in Natural Language Processing (NLP). Pre-training BERT from scratch requires substantial computational resources which may be out of reach for researchers and industry professionals. To some extent this has been addressed by the public release of pre-trained models of different sizes and depths [20]. Available sizes range from tiny (2 layers with hidden size 128) to large (24 layers with hidden size 1024) [6, 3]. The introduction of ALBERT [14] further improved the accessibility of larger models. However, the dependence on pre-trained models limits the ability of researchers to explore new backbone architectures. Furthermore, it limits the extent to which practitioners in industry can leverage internal datasets and adapt the model to their particular needs. Hence, any approach that speeds up the pre-training process is desirable from an economical as well as environmental perspective.

In this paper, we present some methods to enable researchers to accelerate the pre-training of BERT by as much as 2x. The de-facto pre-training dataset Wikipedia, as well as many other NLP datasets, show a positively skewed distribution of sequence lengths. We show that padding tokens (wasted compute) represent 50% of all tokens of the Wikipedia pre-training dataset at sequence length 512. Overall, the sample lengths range between 5 tokens up to 512 (see Figure 1). Samples of length 512 represent only 23.3% of the dataset.

*These authors contributed equally to the paper.

Preprint. Under review.

GroupBERT: Enhanced Transformer Architecture with Efficient Grouped Structures

Ivan Chelombiev* Graphcore Research ivanc@graphcore.ai

Daniel Justus* Graphcore Research daniel@graphcore.ai

Douglas Orr* Graphcore Research douglas@graphcore.ai

Anastasia Dietrich Graphcore Research anastasia@graphcore.ai

Frithjof Gressmann Graphcore Research frithjof@graphcore.ai

Alexandros Koliousis† New College of the Humanities alexandros.koliousis@nchlondon.ac.uk

Carlo Luschi Graphcore Research carlo@graphcore.ai

Abstract

Attention based language models have become a critical component in state-of-the-art natural language processing systems. However, these models have significant computational requirements, due to long training times, dense operations and large parameter count. In this work we demonstrate a set of modifications to the structure of a Transformer layer, producing a more efficient architecture. First, we add a convolutional module to complement the self-attention module, decoupling the learning of local and global interactions. Secondly, we rely on grouped transformations to reduce the computational cost of dense feed-forward layers and convolutions, while preserving the expressivity of the model. We apply the resulting architecture to language representation learning and demonstrate its superior performance compared to BERT models of different scales. We further highlight its improved efficiency, both in terms of floating-point operations (FLOPs) and time-to-train.

1 Introduction

Deep neural networks have emerged as the leading solution to enabling end-to-end language processing (Hochreiter & Schmidhuber, 1997; Sutskever et al., 2014; Chung et al., 2014). Recently, the Transformer model based on the self-attention mechanism (Vaswani et al., 2017) has become the most promising architecture for language applications (Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020). Attention based models are also increasingly showing promising results for established applications in domains different from natural language processing (Dosovitskiy et al., 2020).

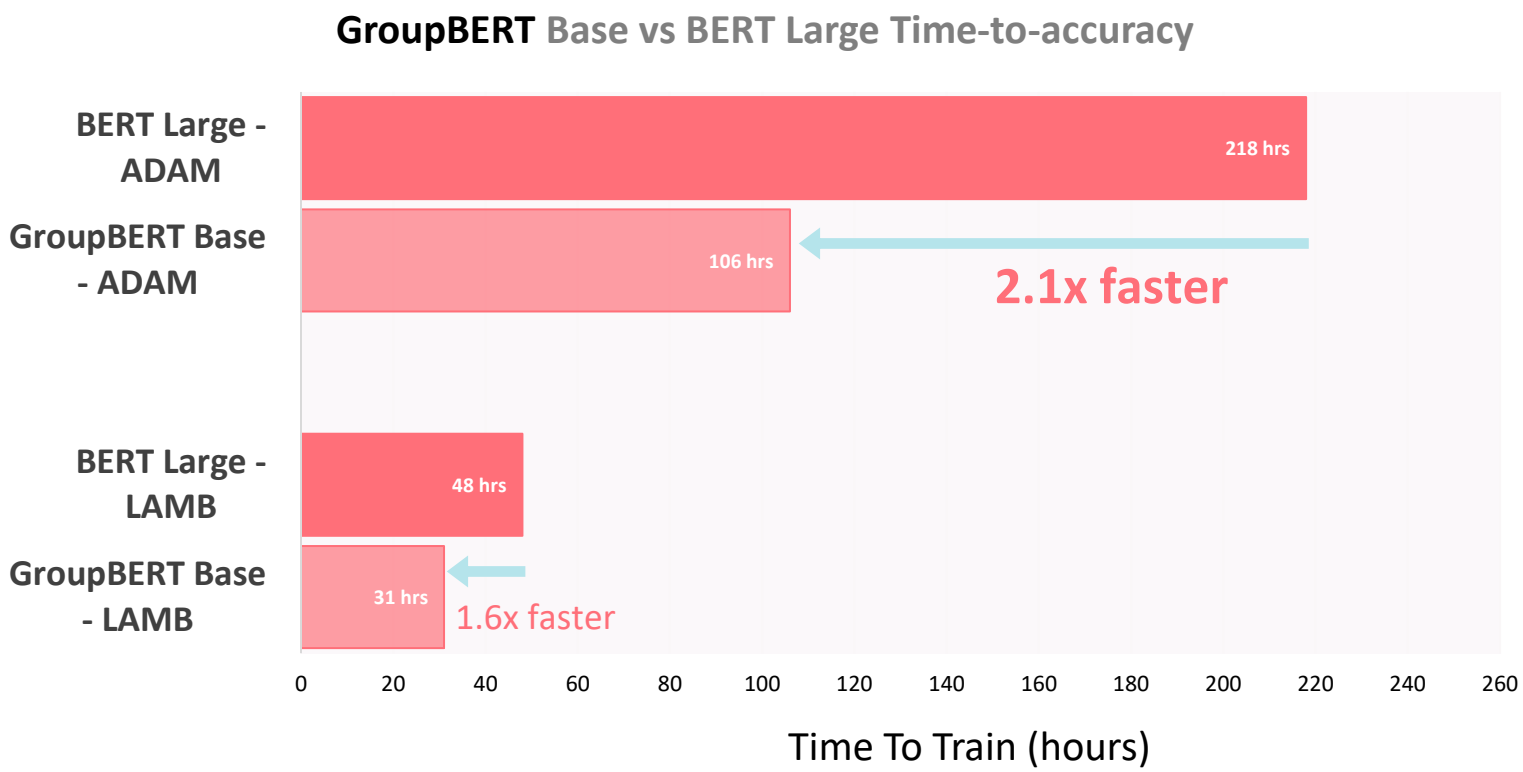
Complementary to the Transformer's improved ability to model long-range dependencies in sequences is its superior potential to scale to larger sizes (Kaplan et al., 2020) and its suitability for execution on existing accelerators. This makes these models favoured over traditional recurrent language models. Given the increased computational demand of these models, there is a growing and pressing interest to develop more efficient architectures (Strubell et al., 2019). Some previous proposals were able to

*Equal contribution.
†Work done while at Graphcore.

Preprint. Under review.

NLP GroupBERT

leverages IPU architecture capabilities
adds grouped convolution module
2x reduction in # parameters
2x faster time to train
improved accuracy of results



NOTES: (updated 20th Dec 2021)

SUMMARY

MODEL NAME:
GROUPBERT

ML DOMAIN:
NLP

MODEL DESCRIPTION:
GROUPBER IS AN OPTIMISED IMPLEMENTATION OF THE BERT MODEL THROUGH USE OF GROUPED TRANSFORMATIONS

DEPLOYMENT:
TRAINING

FRAMEWORK:
TBC

IPU ADVANTAGE:
GROUPBERT DELIVERS 2X REDUCTION IN NUMBER OF PARAMETERS AND 2X FASTER TIMETO TRAIN FOR EQUIVALENT MODEL PERFORMANCE VS STANDARD BERT

https://arxiv.org/abs/2106.05822

MACHINE LEARNING REQUIRES COMPUTE INNOVATION

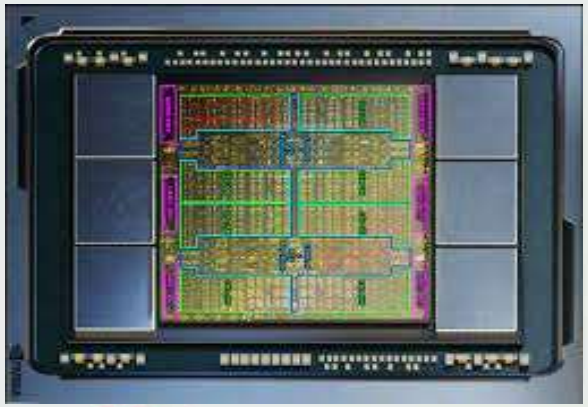
	CPU	>	GPU	>	IPU
	General purpose		Graphics Processing Unit Graphics, part of HPC and ML/DL		Intelligence Processing Unit **ML/DL**
Parallelism	Suitable for scalar processes		SIMD architecture. Suitable for large blocks of dense contiguous data		**Massively parallel MIMD.** **High performance/efficiency as ML trends to sparsity & small kernels**
Memory					
Access ration, MemB					

CPU

Scalar Many core+Multi Die

Designed for office apps

Evolved for web servers

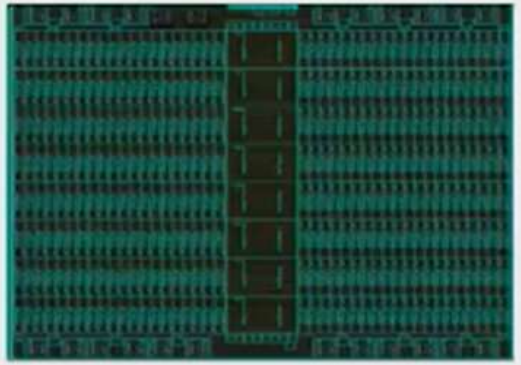


GPU

Vector GPU+HBM2e

Designed for graphics

Evolved for linear algebra



IPU

Graph

Designed for intelligence

Processor Memory

GC200 IPU PROCESSOR

IPU-Tiles™

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

IPU-Core™

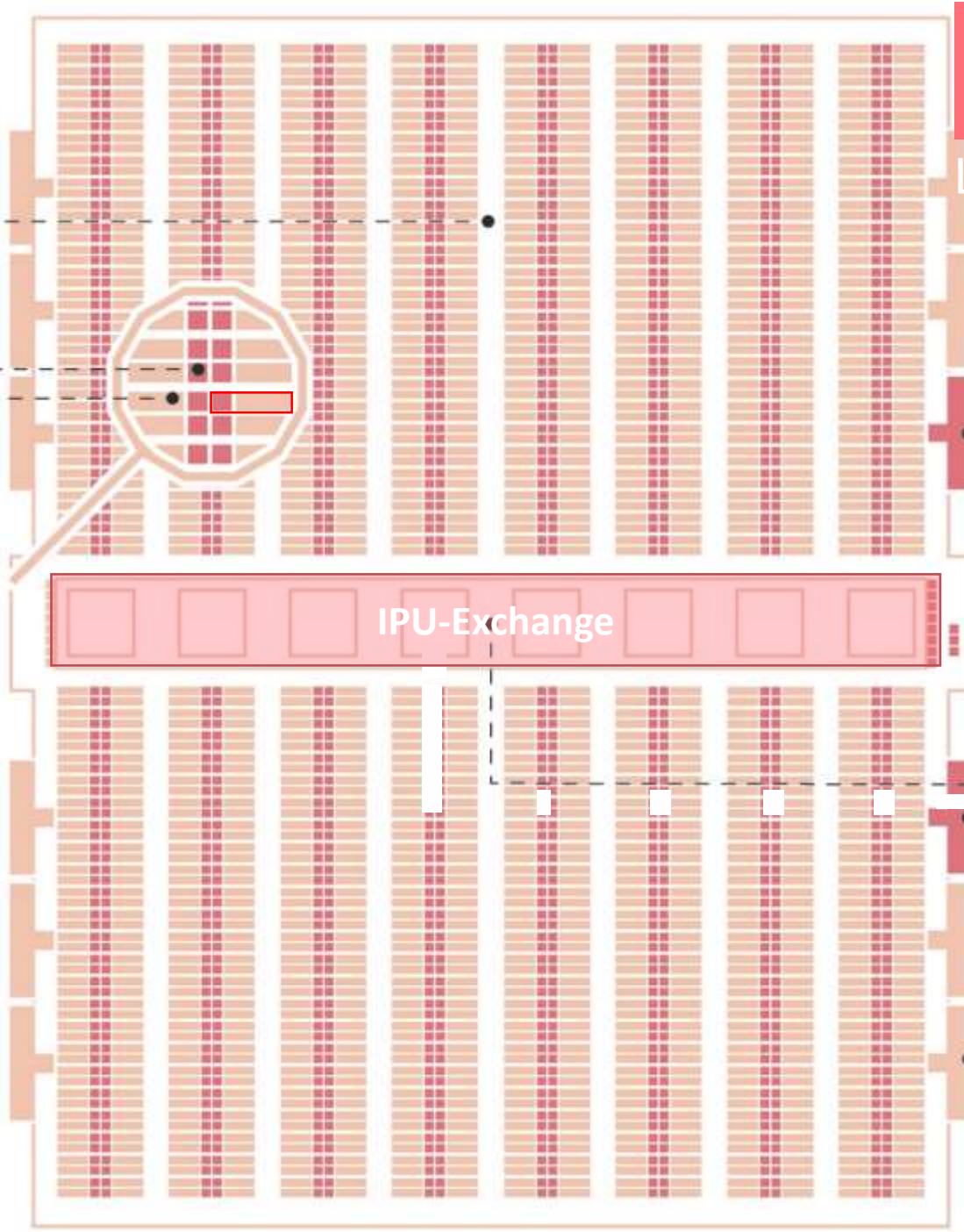
1472 independent IPU-Core™

8832 independent program threads executing in parallel

In-Processor-Memory™

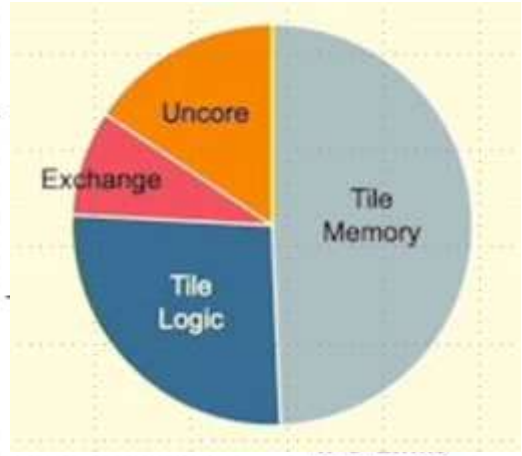
900MB In-Processor-Memory™ per IPU

47.5TB/s memory bandwidth per IPU



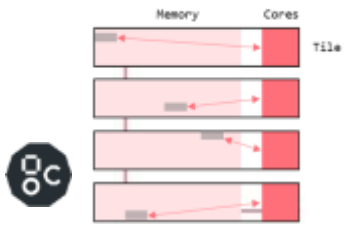
GC200 TSMC 7nm @ 823mm2
250TFlops AI-Float (fp16)
@ 1.325 GHz

- **IPU Tiles** **1472 (Core+Mem)**
  8832 threads
- ***in-Processor Memory***
  900MB per IPU (620KB/tile)
  47.5TB/s Memory BW per IPU
  (33GB/s per tile)
- ***IPU-Exchange***
  8 TB/s All to All
- ***PCIeG4x16*** ***x2***



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

IPU's distributed SRAM



IPU-Machine: M2000

4 x Colossus™ GC200 IPU
1 petaFLOPS AI compute
Up to 526GB Exchange Memory™
2.8Tbps IPU-Fabric™

Each Colossus™ GC200 IPU

59.4Bn transistors, TSMC 7nm @ 823mm2
250 teraFLOPS AI compute
1472 independent processor cores
8832 separate parallel threads

IPU-Gateway SoC

Arm Cortex-A quad-core SoC
Super low latency IPU-Fabric™ interconnect

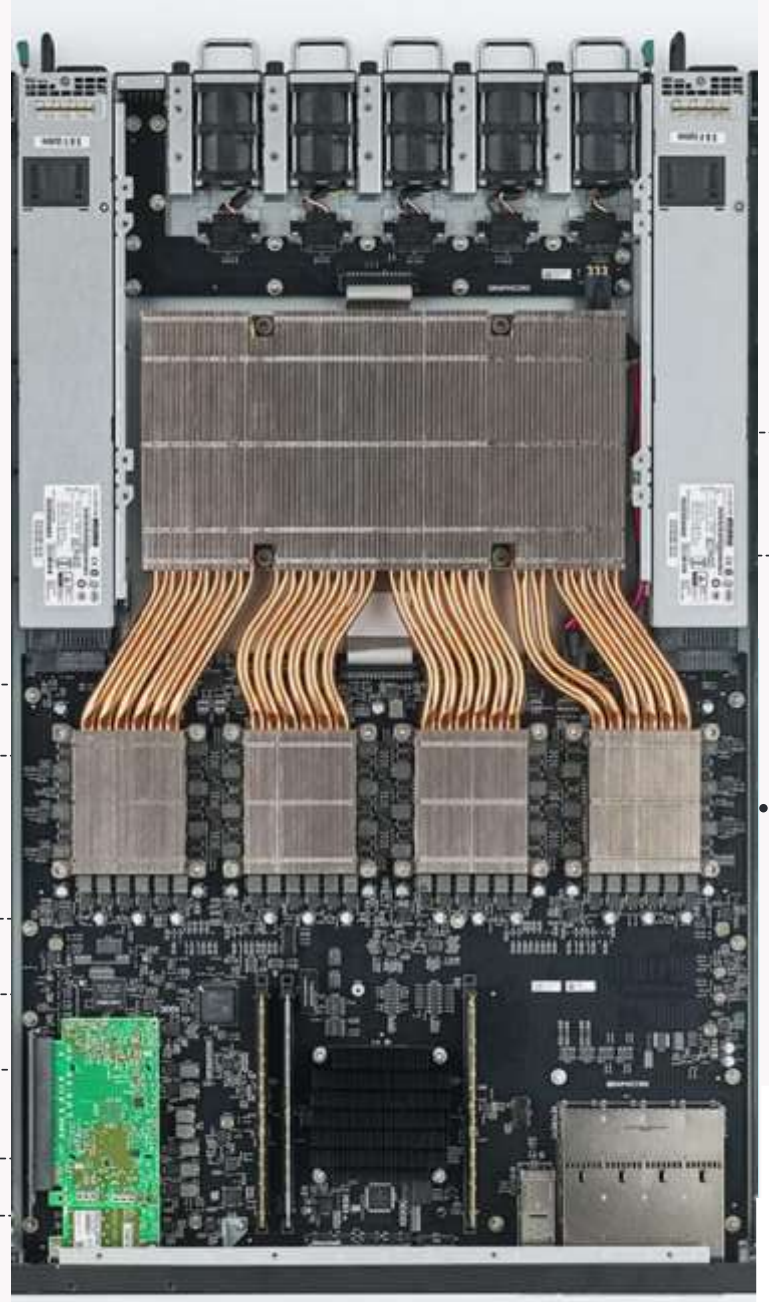
M.2 Connector

Board Management Controller

M.2 Slot

PCIe FH3/4L G4x8 Slot (RNIC/SmartNIC)

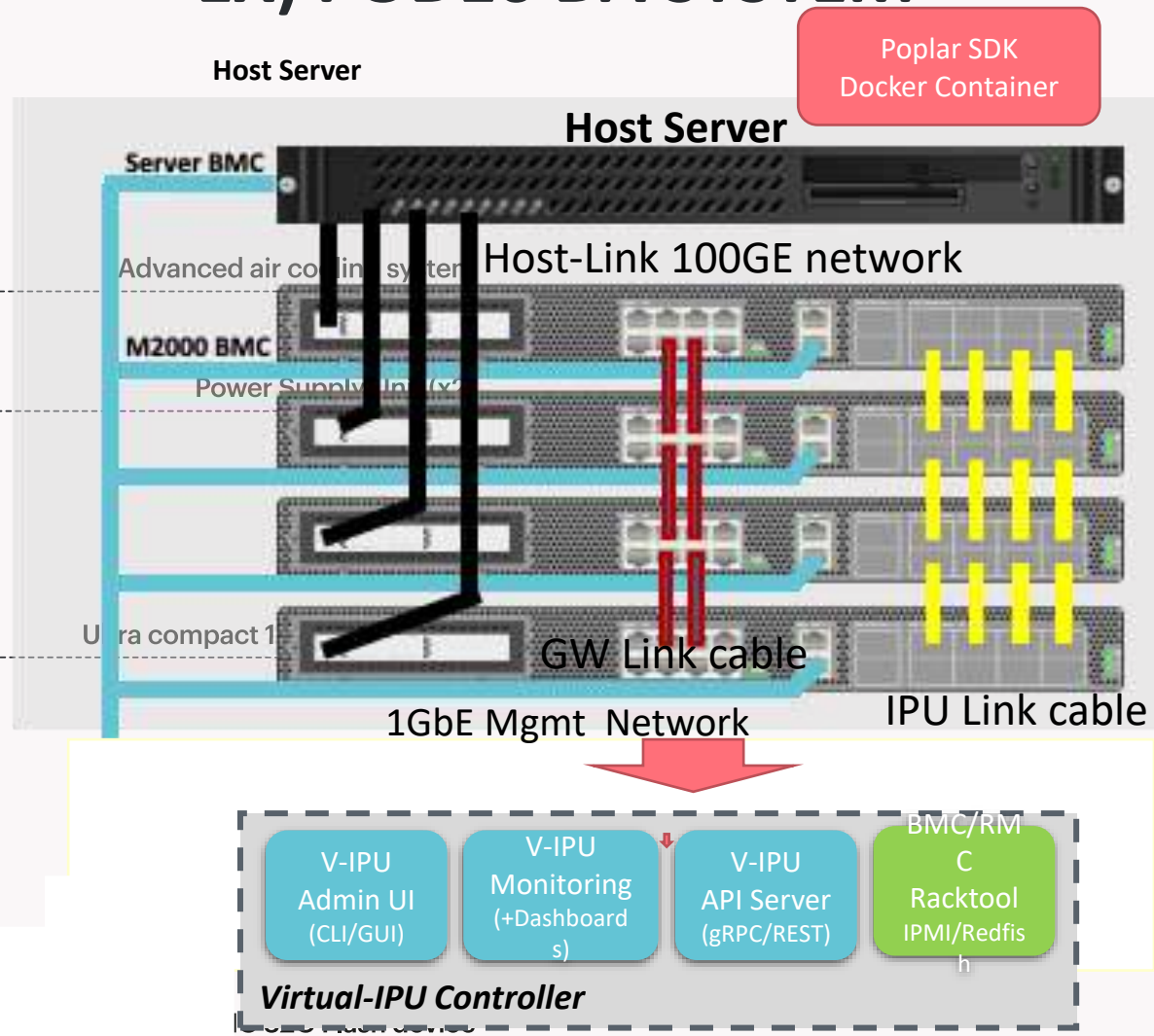
DDR4 DIMM DRAM x 2



For model states streaming

Power
Power consumption 900W-1100W (typical)
Power cut(Max) (1500W)
Input Power 94-256Vac (115-230VAC Normal)

EX) POD16 DA SYSTEM



Network-based disaggregated architecture for host vs. IPU resources

BSP COMPUTE SET EXECUTION

Bulk synchronous parallel

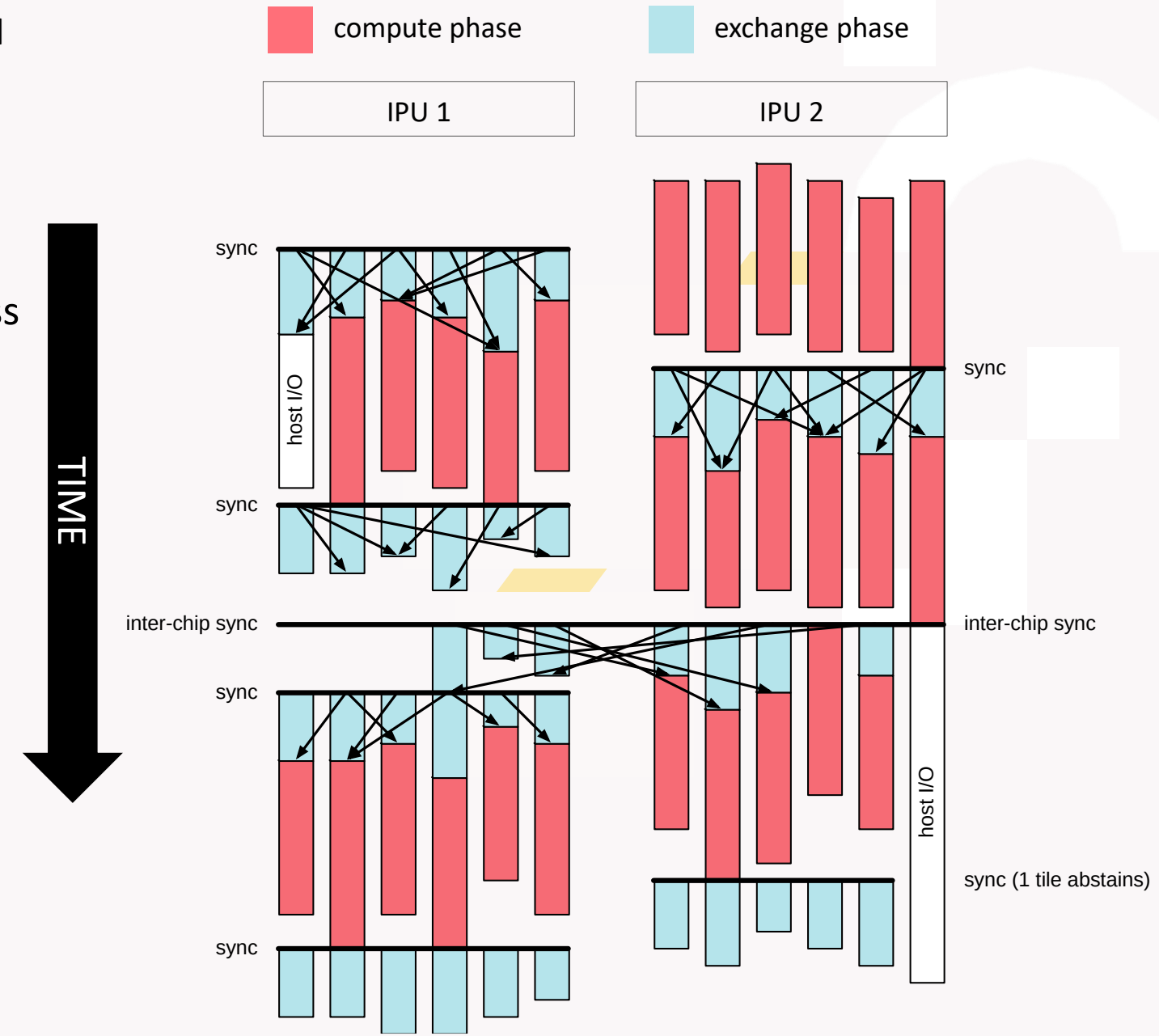
Bulk Synchronous Parallel (BSP)

compute | synchronize | exchange

BSP Compute Set execution works across multiple IPU devices

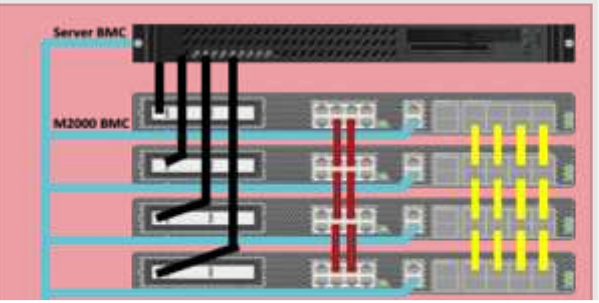
Inter- and Intra-IPU sync phases ensure no live/dead locks or race conditions

Tiles exchanging data with the host can abstain from syncing with other tiles



IPU-POD FLEXIBLE SCALE OUT SYSTEM

Reusable, simple migration between platforms



IPU-POD16 DA（Direct attached）

Eval, POC

IPU-POD4 DA

x1 IPU-M2000

Host Server

Server

IPU-M2000

IPU-M2000

+ IPU-M2000

IPU-POD16 DA

x4 IPU-M2000

(direct attach)

Server

IPU-M2000

IPU-M2000

IPU-M2000

IPU-M2000

Switched version adds
TOR switch & Mgmt switch

+ servers
+ switches
+ IPU-M2000
recable

Disaggregated Server enables configurable Host Server /IPU ratio depending on workload for optimised TCO

IPU-POD64

1 host server

Server

switches

IPU-M2000 ×16

+ servers

IPU-POD64

4 host server

Server ×4

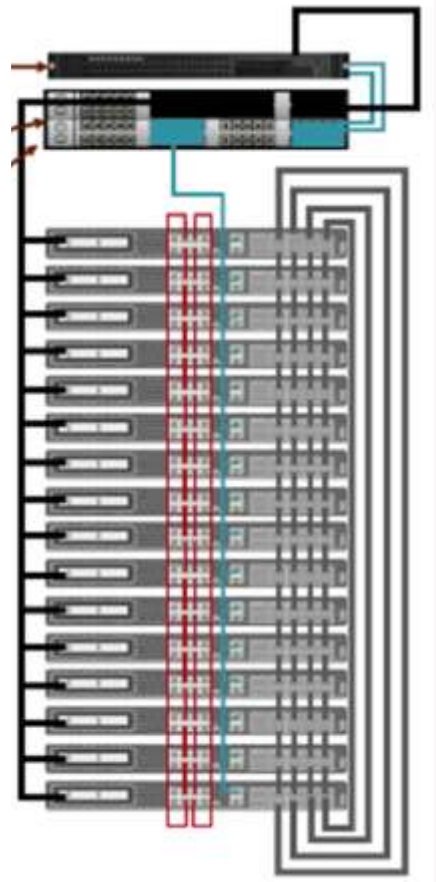
switches

IPU-M2000 ×16

reconfigurable as 4xPOD16

Production

IPU-POD64
(Scale out）



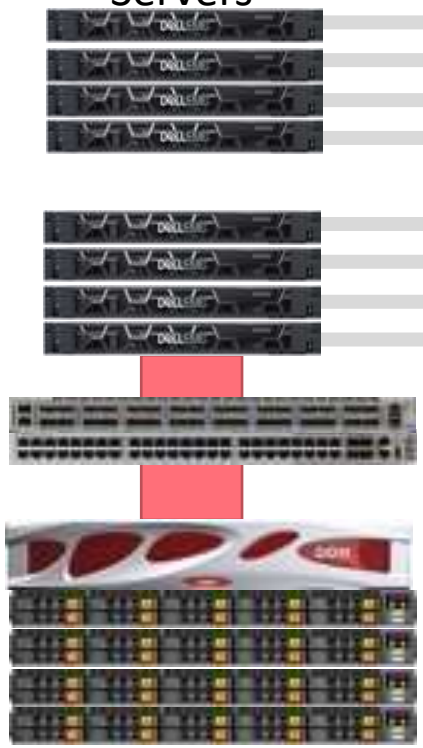
<u>Direct Attached</u>, Pre-configured Model

POD4,16,32

1x IPU-M2000 | 4x IPU-M2000 | 16x IPU-M2000 | Upto 16Kx IPU-M2000

IPU-POD256 DATACENTER ARCHITECTURE

Disaggregated Servers



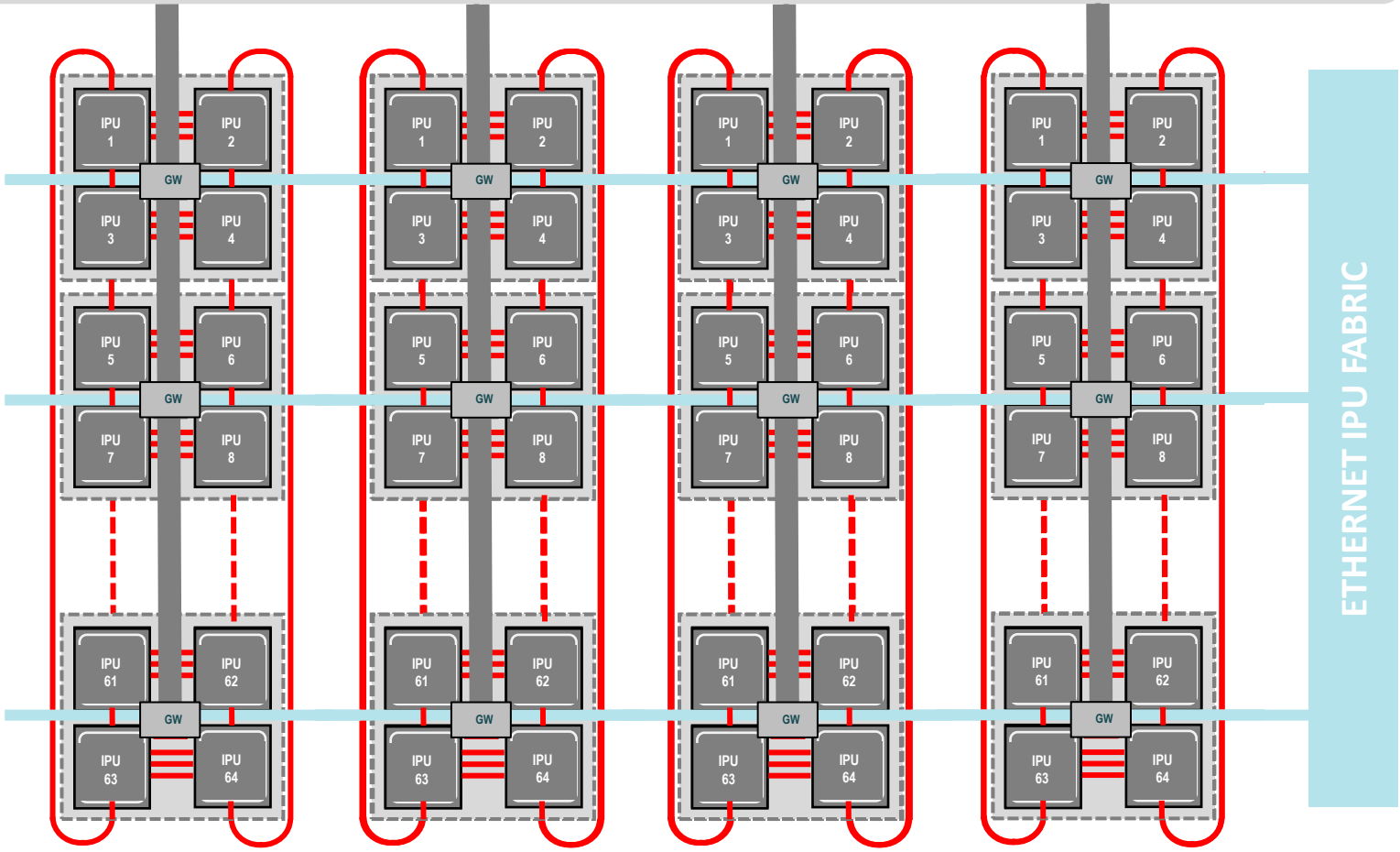
Storage

AI400X

2Uから始められる、
オールNVMeフラッシュ
並列ファイルシステム

Ethernet Switched Fabric

Ethernet Switced Fabric



IPU POD SYSTEM

EXPLORE

IPU-POD$_{16}$

- 4x 1U IPU-M2000
- 1x dual-CPU server
- 4 PetaFLOPS AI compute

BUILD

IPU-POD$_{64}$

- 16x 1U IPU-M2000
- 1x dual-CPU server for BERT
- 4x dual-CPU server for ResNet
- 16 PetaFLOPS AI compute

GROW.1

IPU-POD$_{128}$

- 32x 1U IPU-M2000
- 2x dual-CPU server for BERT
- 8x dual-CPU server for ResNet
- 32 PetaFLOPS AI compute

GROW.2

IPU-POD$_{256}$

- 64x 1U IPU-M2000
- 4x dual-CPU server for BERT
- 16x dual-CPU server for ResNet
- 64 PetaFLOPS AI compute

Large

IPU-POD$_{512,\ 1024}$ EXA-POD

GRAPHCORE

SOFTWARE

For Ease of Use

POPLAR™ COMPUTE GRAPH VISUALISATION

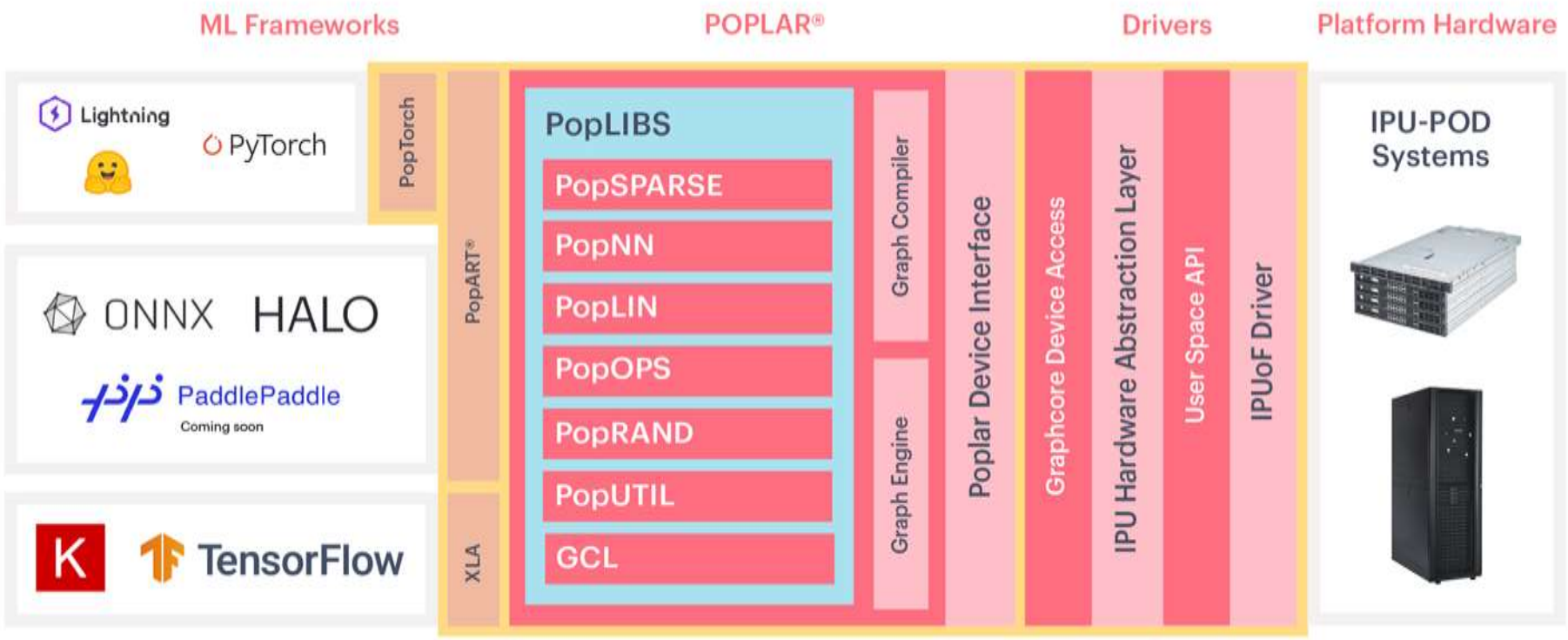
POPLAR SDK

Host Server OS

ubuntu
18.04
20.04 (preview)

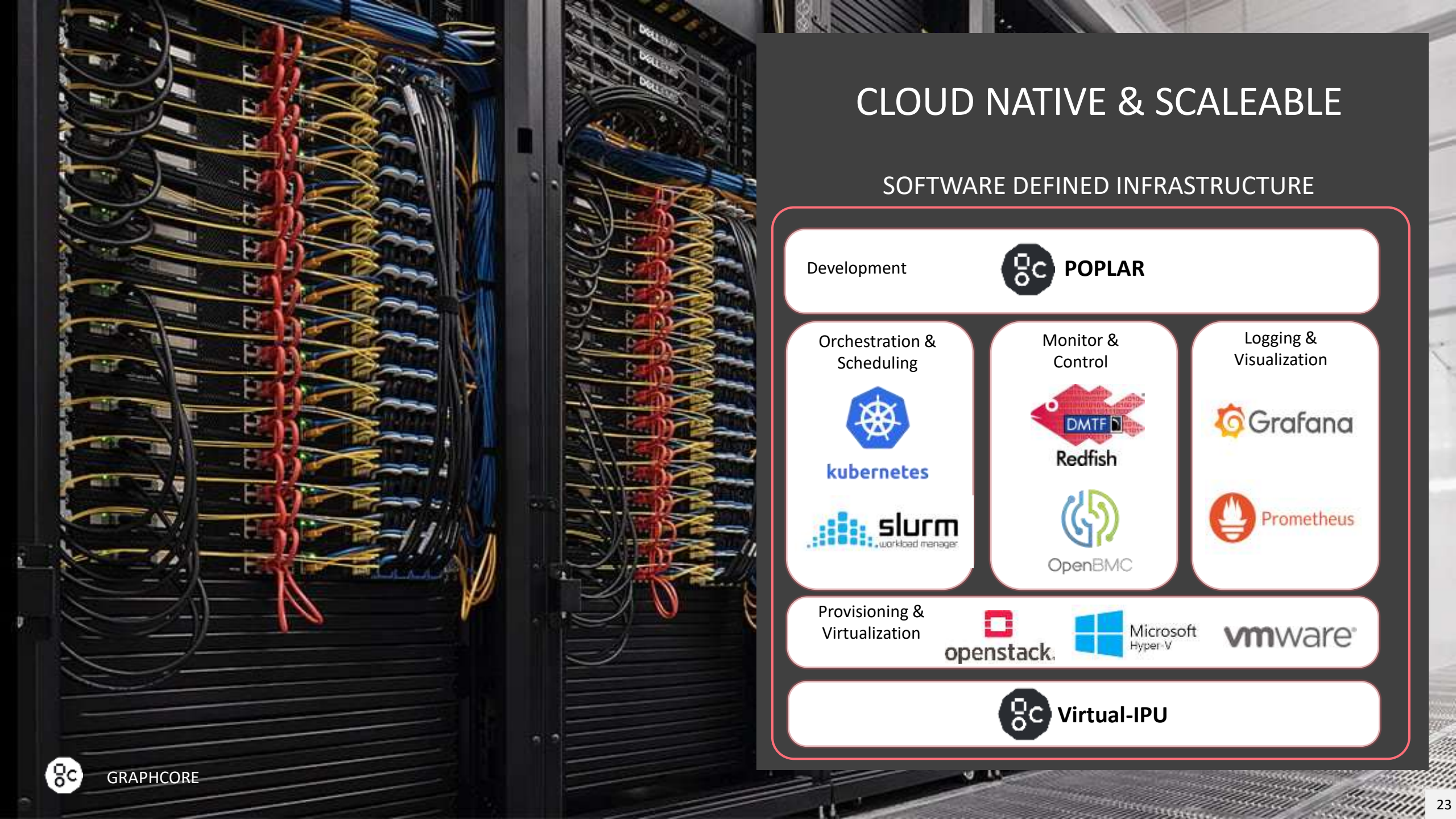
CentOS
7.6
（RHELへの移行）

Debian 10



POPLAR
Visualization
& Debug Tools

Management data can be delivered via Redfish DTMF to management systems such as Grafana.
Poplar software and our IPUs support containerization with Docker.



CLOUD NATIVE & SCALEABLE

SOFTWARE DEFINED INFRASTRUCTURE

Development — **POPLAR**

Orchestration & Scheduling
- kubernetes
- slurm workload manager

Monitor & Control
- DMTF Redfish
- OpenBMC

Logging & Visualization
- Grafana
- Prometheus

Provisioning & Virtualization
- openstack.
- Microsoft Hyper-V
- vmware

Virtual-IPU

DEVELOPER ECOSYSTEM WITH PARTNERS

TensorFlow

Keras

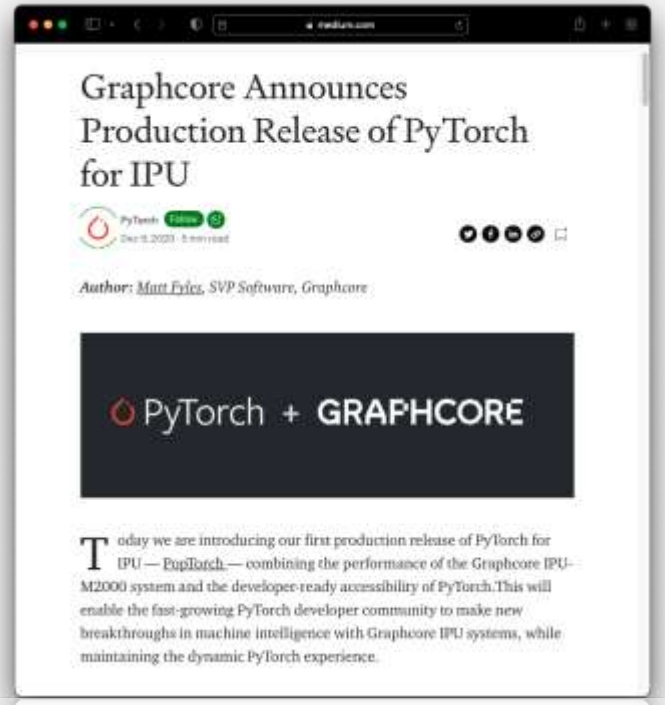
PaddlePaddle

PyTorch

PyTorch Lightning

HALO

jupyterhub

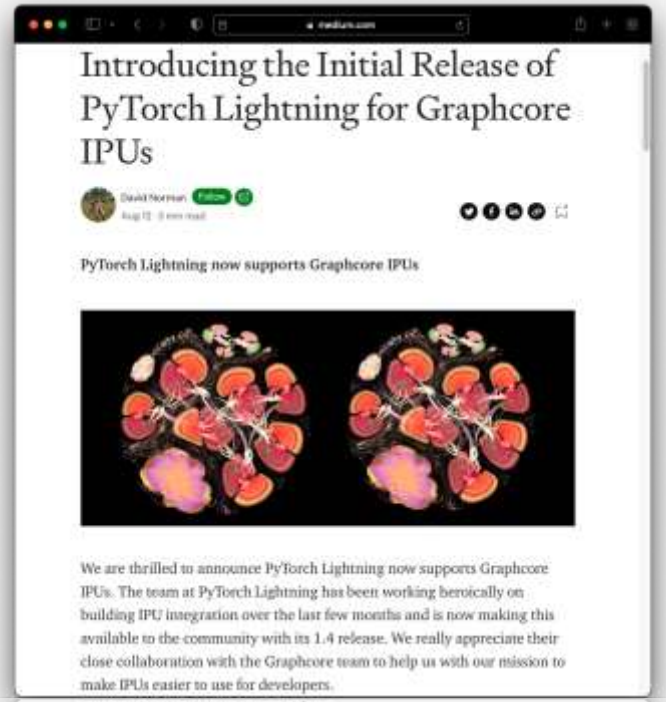


Graphcore Announces Production Release of PyTorch for IPU

Author: Matt Fyles, SVP Software, Graphcore

PyTorch + GRAPHCORE

Today we are introducing our first production release of PyTorch for IPU — PopTorch — combining the performance of the Graphcore IPU-M2000 system and the developer-ready accessibility of PyTorch. This will enable the fast-growing PyTorch developer community to make new breakthroughs in machine intelligence with Graphcore IPU systems, while maintaining the dynamic PyTorch experience.



Introducing the Initial Release of PyTorch Lightning for Graphcore IPUs

PyTorch Lightning now supports Graphcore IPUs

We are thrilled to announce PyTorch Lightning now supports Graphcore IPUs. The team at PyTorch Lightning has been working heroically on building IPU integration over the last few months and is now making this available to the community with its 1.4 release. We really appreciate their close collaboration with the Graphcore team to help us with our mission to make IPUs easier to use for developers.

HUGGING FACE

vmware

Weights & Biases

docker

SPELL

openstack.

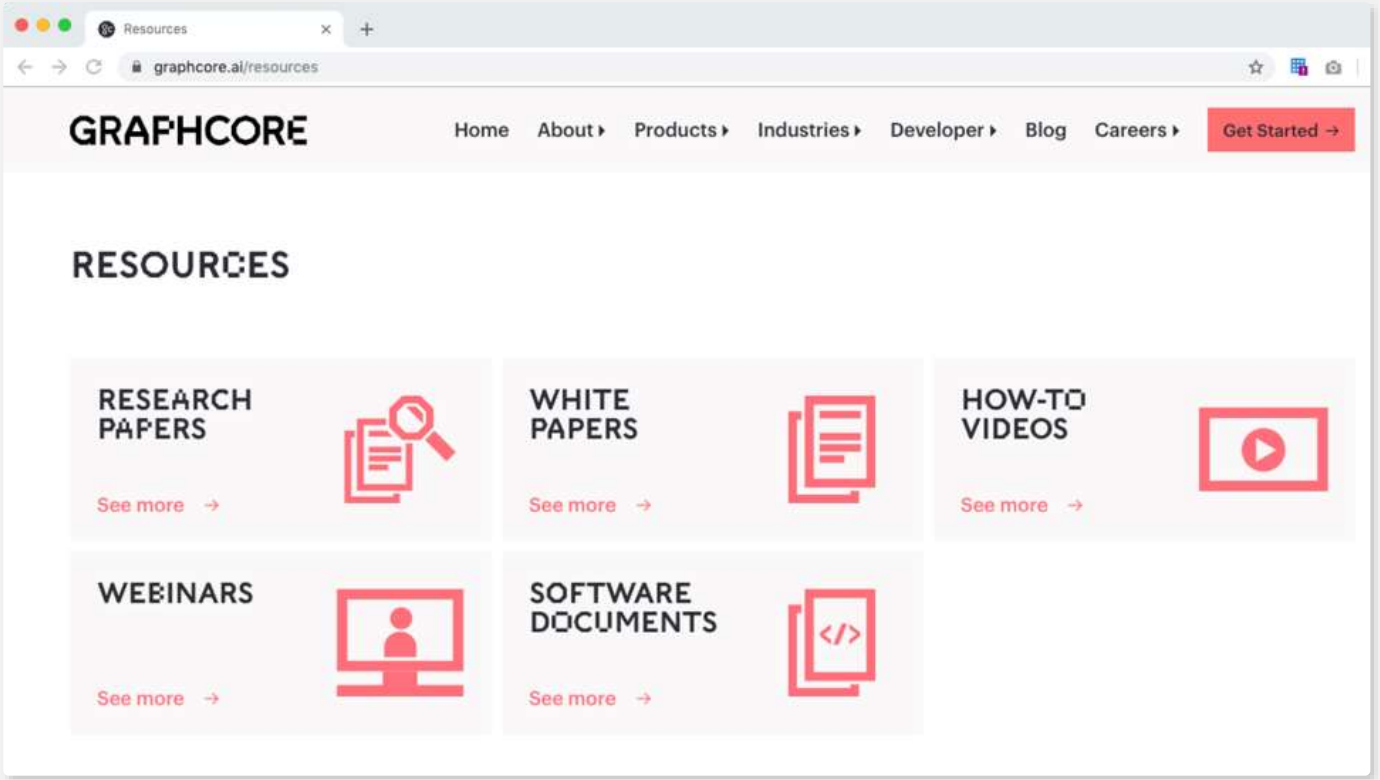
gradient by Paperspace

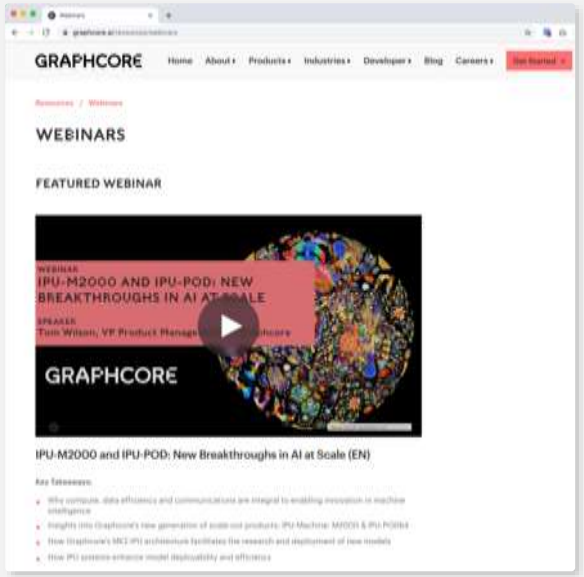
GitHub

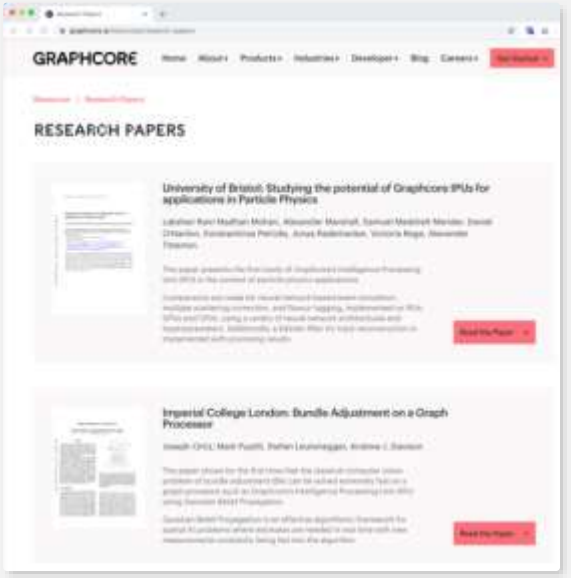
RESOURCES CENTRE

https://www.graphcore.ai/resources

- Central source of research papers, white papers, how-to videos, on-demand webinars and documentation







NEWSFLASH: GRAPHCORE OPEN SOURCES POPLAR GRAPH LIBRARIES (POPLIBS)

HTTPS://WWW.GRAPHCORE.AI/POSTS/GRAPHCORE-OPEN-SOURCES-POPLAR-GRAPH-LIBRARIES-POPLIBS

GitHub

RAPIDLY INCREASING NUMBER OF SERVER PLATFORM PARTNERSHIPS…

STORAGE PARTNERS

GRAPHCORE ELITE PARTNER

ddn

PURE STORAGE

VAST

WEKA

GLOBAL ENGAGEMENT WITH MAJOR HYPERSCALERS AND CSP'S

Cloud Solution Provider

GRAPHCORE SOFTWARE Environment

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

ML APPLICATIONS

Front End

Framework

BACKENDS

TUTORIALS
CODE EXAMPLES
DOCUMENTATION
VIDEOS
NATIVE IPU CODERS PROGRAM
APPS PORTFOLIO

DEVELOPER SUPPORT ECOSYSTEM

POPLAR®

DRIVERS

JUPYTER
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
provisioning software

GRAPHCORE

GRAPHCORE IPUの適用領域

データセンター インターネット業界

・リコメンド
・検索エンジン
・トレンド分析・予測
・スマートアシスタント
・画像検索
・顧客サポート

自動車/電機 製造業

・自動運転
・自動設計
・不良個所の画像認識
・販売支援
・混雑予想
・高度ナビゲーション システム

ヘルスケア ライフサイエンス

・ゲノム研究
・医療画像診断
・疾患診断
・タンパク質の機能解析
・創薬

金融/保険/株式

・株価予測
・リスク分析
・景気予測
・マネロンの検知/防止
・保険料率予測
・顧客の本人確認・認証

研究開発 高等教育

・ニューラルネットによるシミュレーション
・ビッグデータ解析
・実験画像解析
・素粒子物理学
・天文学

CUSTOMER ANNOUNCEMENTS 2021

Academic Program
日本からも1団体参加

Man Group - FinTech

Tractable – Insurance AI as a service

Weather/Climate at ECMWF

Carmot Capital　Fintec

NHN launch　EC/IPU cloud

GRAPHCORE

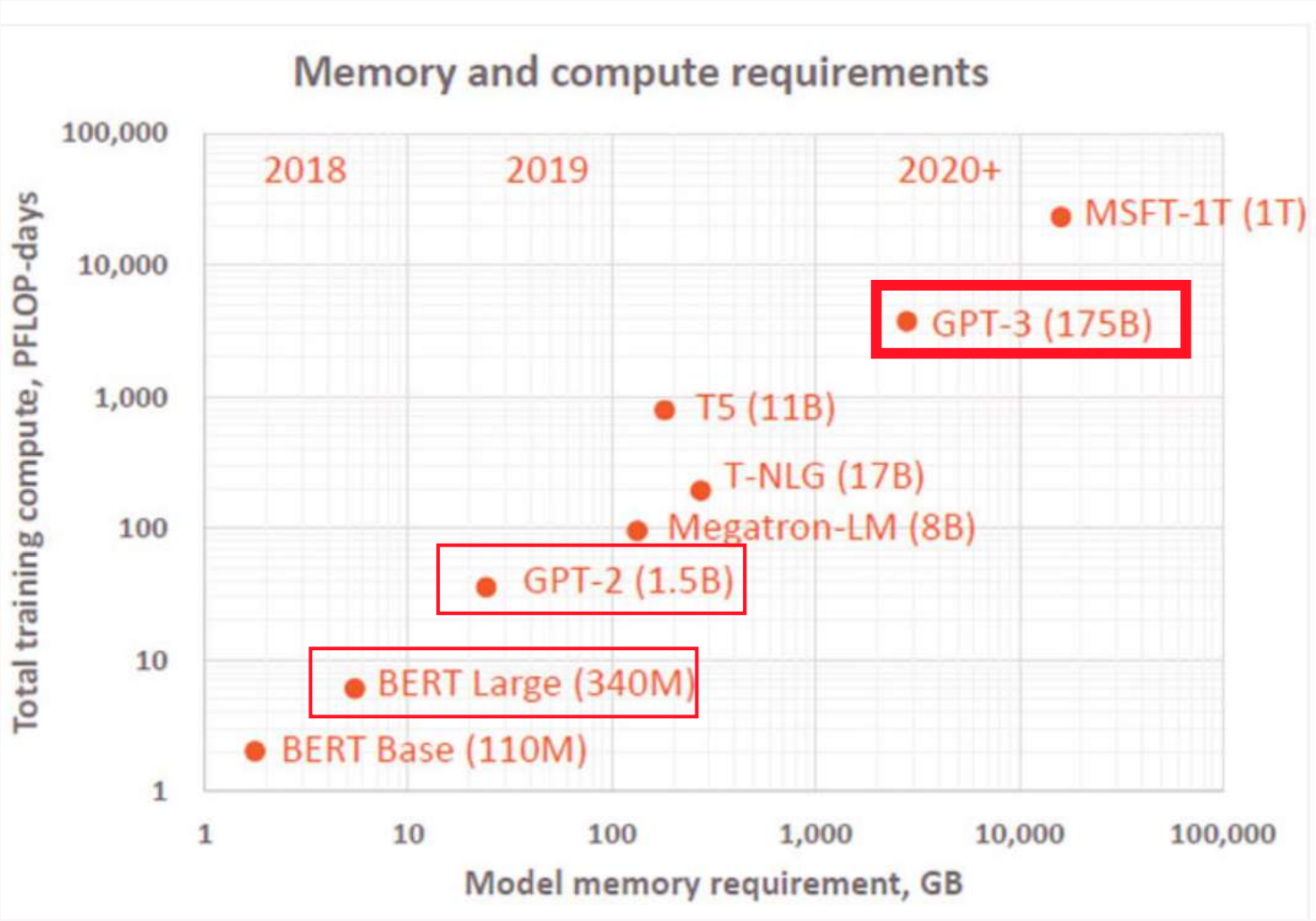
国内導入例　大手製造業、テレコム、金融、国立研究所、国立大学　など

LARGE MODELへの取り組み

- **大規模モデルの規模と計算量は2年間で約1000倍に増大**
- GPT-3は1024GPUを使用し学習に2-４か月かかると言われている。
  また、別の推測では学習にはV100で4096GPUで1か月、ASP利用ではコストが$33M。
  また、別ソースの推測ではA100で2000GPUで1か月、ASPでコスト$50M）

1ラック(25KW)のシステムで1日で学習可能にするには？

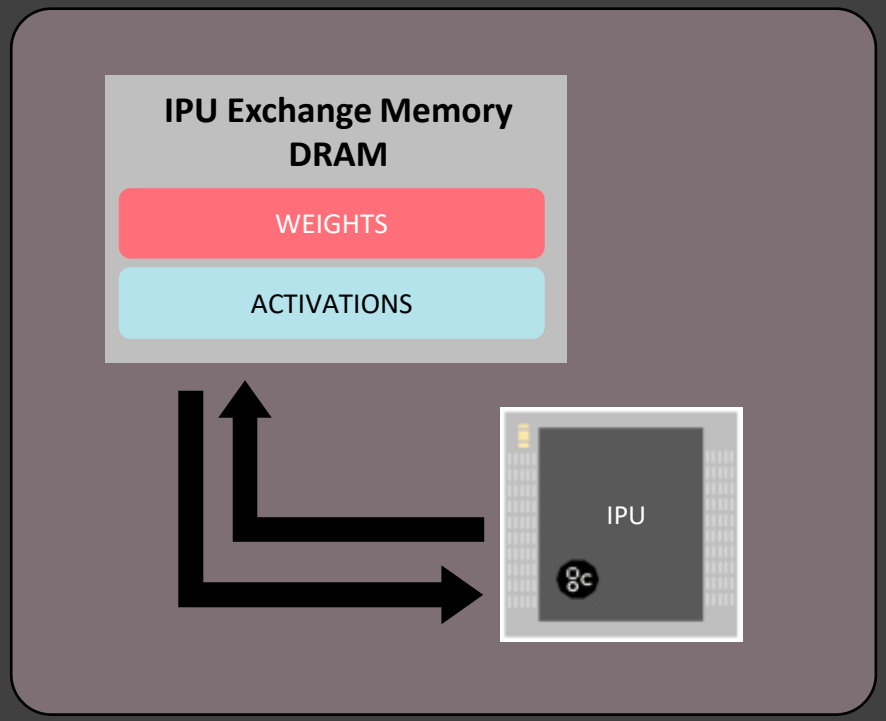
指数関数的に増大



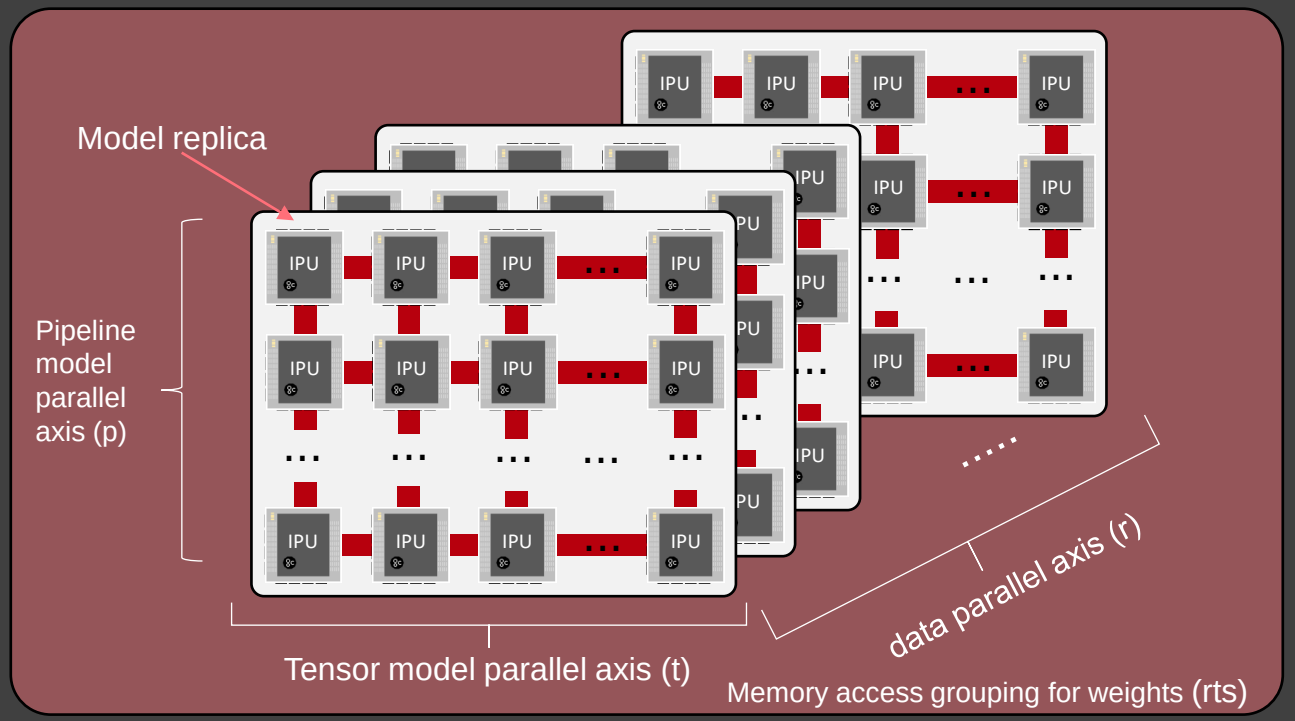
出展　https://news.mynavi.jp/article/20210914-1971929/images/004l.jpg

LARGE MODELへの対応

IPU-ExchangeメモリとIPU-Links™を組み合わせたIPU-PODでの実装

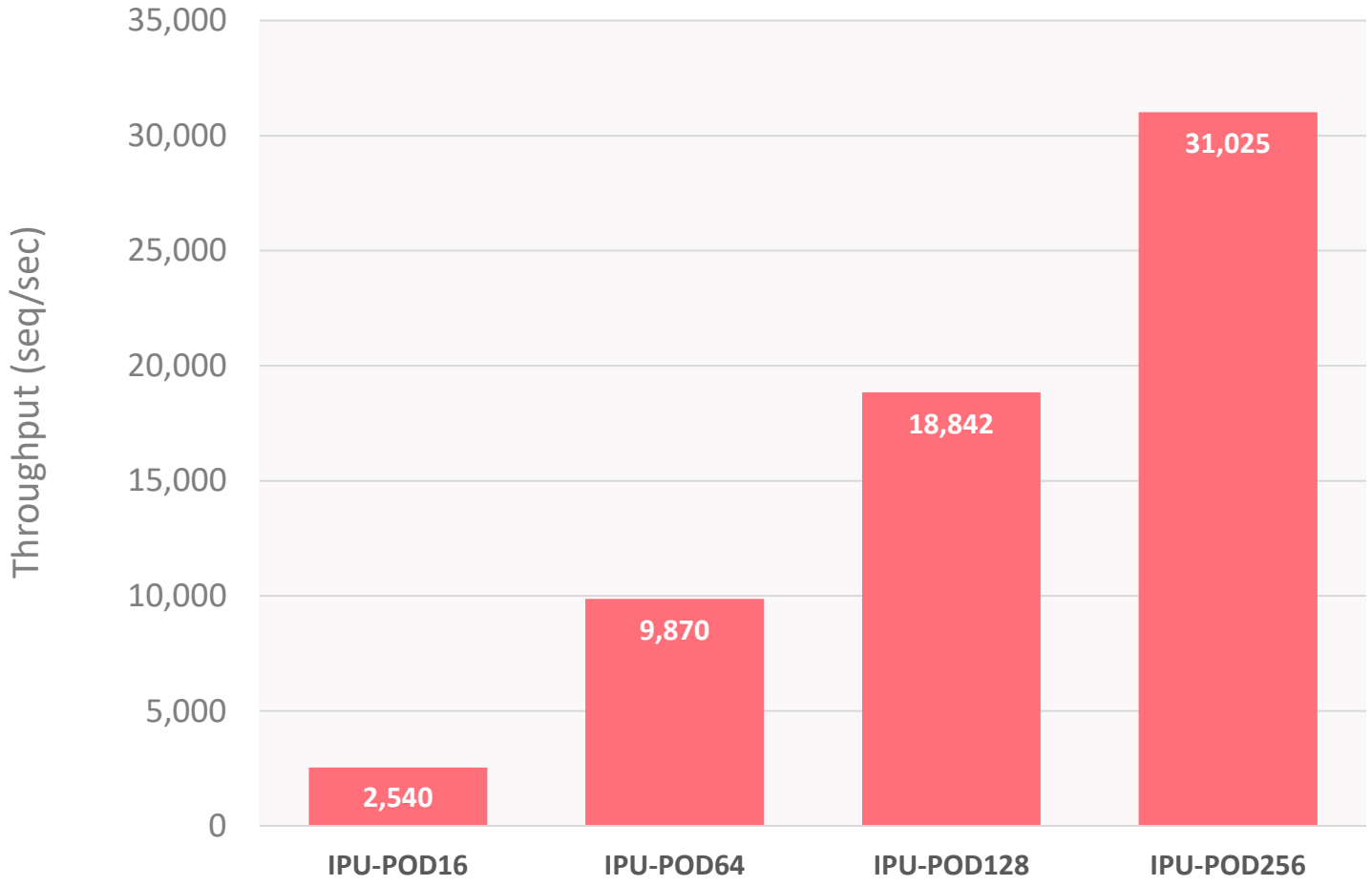


Exchange-MemoryとIn-Processor Memoryの間でウェイトとアクティベーションを“スマート（効率的に）”にスワッピングを行う。



IPU-POD全体でモデルの並列分解 、Poplar®は自動化されたモデル分割をサポートします。

GPT-2 MEDIUM SCALING



NOTES: (updated 20th Dec 2021)
GPT2-Medium Training Throughput | Wikipedia dataset
IPU-POD Platforms | FP 16.16 | SDK 2.3.0 | PyTorch | https://www.graphcore.ai/performance-results

SUMMARY

MODEL NAME:
GPT2-MEDIUM

ML DOMAIN:
NLP

MODEL DESCRIPTION:
GPT IS A GENERATIVE PRE-TRAINING MODE. IT IS UNIDIRECTIONAL TRANSFORMER-BASED MODEL USED FOR PREDICTING LANGUAGE.

DEPLOYMENT:
TRAINING

FRAMEWORK:
PYTORCH

IPU ADVANTAGE:
RESULTS ACROSS MULTIPLE IPU-POD PLATFORMS WITH EXCELLENT SCALING

IPU-POD 性能機能評価

❑ **性能機能評価メニュー**

1. **Graphcloudを利用した評価**
2. **性能評価課題を預かり、弊社にて性能評価、報告**
3. **評価用IPU-PODシステムの貸し出し**
4. **Try＆Buyプログラム**

1. Graphcloudを利用した評価

IPU-POD16 をクラウド上に配置したGraphCloudで2週間無償評価が可能.

- SSH経由
- 無償期間は2週間 (POD64 利用は有償(要相談）)
- 長期利用、IPU-POD64等の利用は有償にて対応可能.
  - 弊社Webもしくは代理店様経由で利用申し込み可能
  - 事前にシステム、プログラミングの速習トーレニングも用意
  - モデル, 入力データ種別を事前にお伝えいただけますと,
    事前準備を対応させて頂きます。

2. 性能評価課題を預かり、弊社にて性能評価、報告

評価目的、実施内容、日程などをお聞きした上で実施可能な場合、
課題を預かり、弊社で移植、最適、試験実施、報告を行います。
評価を実施する時間がない、大きな規模のシステムが必要、
最適化なども含めたサポートが必要な場合にご利用下さい。

3. 評価用IPU-PODシステムの貸し出し

評価目的、実施内容、日程などをお聞きした上で実施可能な場合、
ホストサーバも含めたPOD-4（もしくはPOD8）を貸し出し評価。
セキュリティ高いデータを外部に出せない等の受験科デモ、貸し出し
システムを利用して評価可能になります。
貸し出しに際して、

- 事前にシステム、プログラミングの速習トーレニングも用意
- モデル, 入力データ種別を事前にお伝えいただけますと,
  事前準備等を対応させて頂きます。

4. Try＆Buyプログラム

購入を前提として
評価目的、実施内容、日程などをお聞きした上で実施可能な場合、
ホストサーバも含めたIPU-PODシステムをお貸出しし、評価終了後には
購入いただくプログラムです。
IPU-POD16やPOD64を実際の設置環境に置き、既存システムとの
インテグレーションなども含めて動作検証などを行えます。

ご清聴ありがとう御座いました。

Try IPU　GraphCloud

性能/機能評価(POD16/2Week無償利用）
クラウドサービス(POD16,POD64)
https://www.graphcore.ai/partners

WebinarやVideo,のご紹介
https://www.graphcore.ai/resources

Graphcoreへのお問い合わせ
https://www.graphcore.ai/
Mail　mamorun@graphcore.ai
TEL　03-3507-5600

Graphcore:dellsales@graphcore.ai

IPU-POD$_{64}$

Dell R6525 (3)
IPU-M2000 (16)
PowerSwitch
Rack (1) & PDU (2)