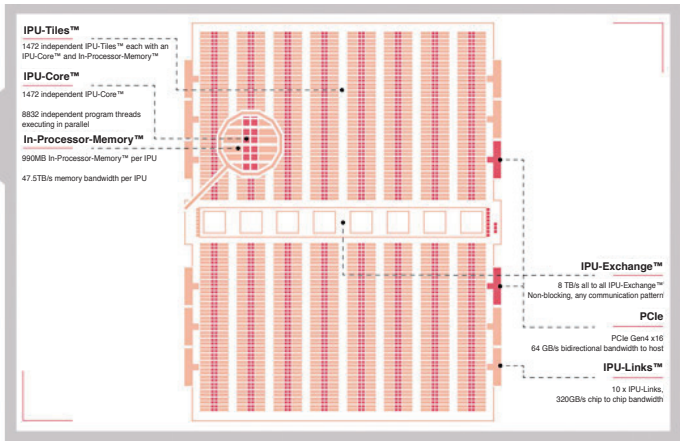
機械学習ワークロードで最大の性能を発揮する大規模並列プロセッサ

SCSK

GRAPHCORE(グラフコア)の第2世代Colossus™ MK2 IPUプロセッサ GC200は、機械学習ワークロードで最大の性能を発揮するよう設計された新しいタイプの大規模並列プロセッサです。GC200は最も複雑なプロセッサですが、GRAPHCOREが合わせて提供するPoplarソフトウェアとIPU-M2000プラットフォームにより、利用者は容易にAIのブレークスルーを実現することができます。

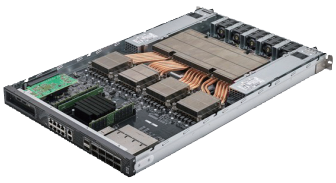
第2世代Colossus™ MK2
IPUプロセッサ GC200

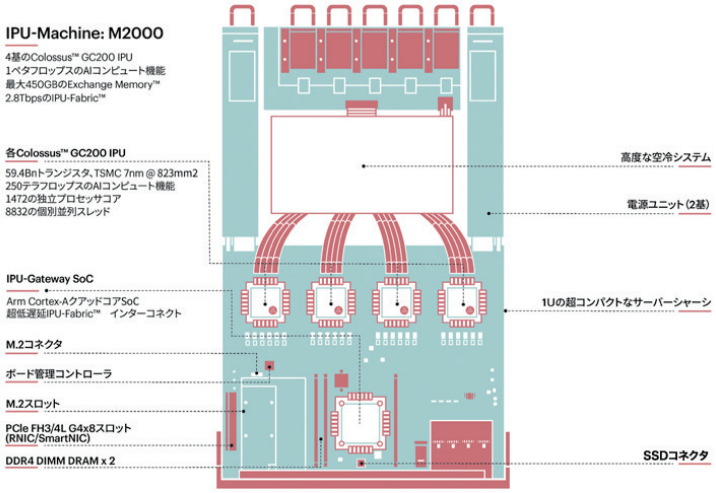


■1472のプロセッサコアと8832の並列スレッドで実現するMIMDアーキテクチャ
■900GBのIn-Processor Memory(SRAM)は47.5TB/sの超高速な帯域を提供

IPU-Machine:M2000

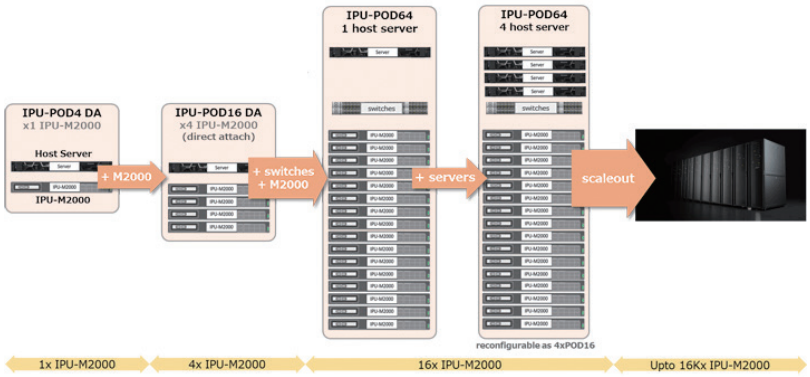
IPU-M2000は、第2世代IPUプロセッサGC200を4基搭載したGRAPHCOREの中核となるハードウェアプラットフォームです。
1Uシャーシに1ペタFLOPSのAIコンピュート機能、3.6GBのIn-Processor Memory(SRAM)、最大450GBのExchange Memory(DRAM)を搭載し、要求の厳しい機械学習ワークロードを処理します。ホストサーバやM2000同士の接続用に100Gbpsの高速・低遅延のインターフェイスを有しています。





スケーラビリティ

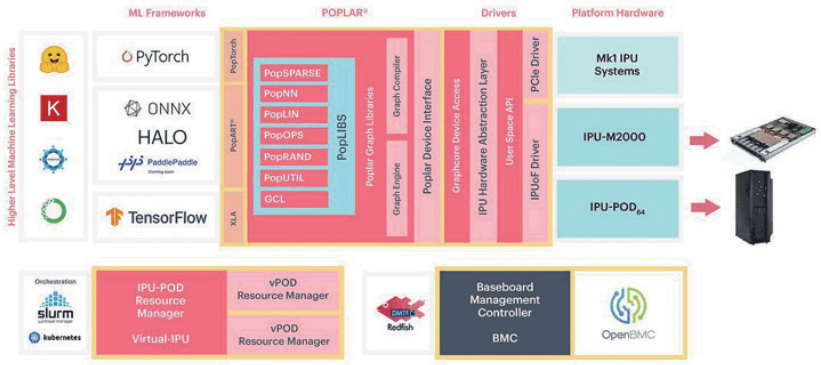
ホストサーバとIPU-M2000の1対1接続を最小構成とし、M2000 4台を相互接続したIPU-POD16、スイッチを介した16台構成のIPU-POD64と、需要に合わせて柔軟にシステムを拡張していくことができ、最大数千台といったスーパーコンピュータの規模まで拡張可能です。



次世代AIプラットフォーム GRAPHCORE IPU

Poplarソフトウェア

GRAPHCOREは、その革新的なIPUプラットフォームと合わせて、Poplarというソフトウェア群を提供します。Poplarがあれば、大規模なIPU環境でも、1台のマシンと同様の容易さで使うことができます。さらに、Poplarがすべてのスケーリングと最適化を行ってくれるので、ユーザはモデルとその結果に集中することができます。複数のユーザが異なるワークロードを同時に実行したい場合は、GRAPHCOREのVirtual-IPUソフトウェアを使い、AI計算を動的に共有することも可能です。



Poplarは、TensorFlow、PyTorch、ONNX、PaddlePaddleなどの標準的な機械学習フレームワーク、およびOpen BMC、Redfish、Dockerコンテナや、SlurmおよびKubernetesなどの業界標準のツールをサポートしているため、実用展開も容易です。

ベンチマーク

自然言語処理 - BERT

IPUは、様々な産業分野や事例において広く使われている各種の自然言語処理モデルにて非常に優れた性能を発揮します。BERT-Largeモデルでは、IPU-POD64は最新のGPUと比べて学習時間を大幅に短縮できます。

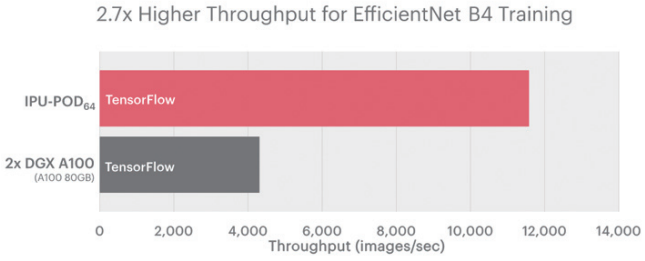
BERT-Large：TTT (time-to-train)

NOTES:
BERT-Large using Wikipedia dataset | end to end pre-training
IPU-POD64 (16x IPU-M2000 Server) using PopART | SDK 1.4.0 | Mixed Precision Ph1 SL=128, Ph2 SL=384.
DGX A100 results calculated using NV published TensorFlow throughput & training scheme | Mixed Precision | Assume linear scaling from 1x to 2x DGX A100
https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/LanguageModeling/BERT#training-accuracy-results

画像認識 - EfficientNet

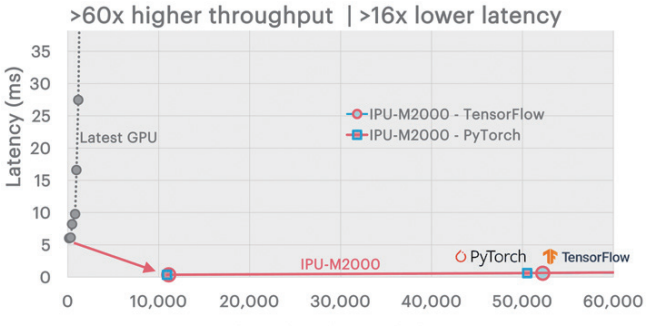
視覚探索エンジンや医療用画像処理といった事例においては、できるだけ小さい遅延で高いスループットを実現することが重要です。EfficientNetのような、より高い精度と効率性を実現した新しいコンピュータビジョンモデルの学習と推論の両方において、IPUは比類のない性能を発揮します。

EfficientNet B4：Training



NOTES: (updated 15th Mar 2021)
EfficientNet-B4 | Real Data (ImageNet)
IPU-POD64 (16x IPU-M2000) | results for TensorFlow | FP 16.32 | SDK 2.0.0 | results for G16-EfficientNet using Group Dim 16
DGX A100 640GB (8x A100-SXM4-80GB) using TensorFlow | Mixed Precision | assume 0.85x scaling from 1=>2 DGX A100
A100 results published by NVIDIA (https://developer.nvidia.com/deep-learning-performance-training-inference) on 2nd Mar 2021

EfficientNet-B0：Inference



NOTES:
EfficientNet-B0 | headline comparisons using lowest latency
1x IPU-M2000 using TensorFlow & PyTorch | FP 16.16 | Synthetic | SDK 1.4.0 | Batch size 4 through 160 | Replicated scaling across IPU-M2000
No GPU results published on NV results website
1x Latest GPU using TensorFlow | Batch Size 1 through 512 | results measured using public Google repo | FP32 support only | Synthetic |
(https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet/).

SCSK株式会社

ITエンジニアリング事業本部　エンタープライズ第一部

豊洲本社　〒135-8110 東京都江東区豊洲3-2-20 豊洲フロント
TEL：03-5859-3011　FAX：03-5859-3085　URL:https://www.scsk.jp

西日本 北浜オフィス　〒541-0041 大阪府大阪市中央区北浜1-8-16（大阪証券取引所ビル）　TEL.06-6223-8800
中 部 オ フ ィ ス　〒460-0003 愛知県名古屋市中区錦2-16-26（SC伏見BLDG.）　TEL.052-209-7500
広 島 オ フ ィ ス　〒730-0012 広島県広島市中区上八丁堀4-1（アーバンビューグランドタワー）　TEL.082-511-6700
九 州 オ フ ィ ス　〒812-0011 福岡県福岡市博多区博多駅前3-30-23（博多管絃ビル）　TEL.092-472-5800

資料請求と価格のお問い合わせは… http://www.clubscs.com/ または hpcteam-staff@ml.scsk.jp へ