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

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 : INFERENC

3x higher throughput at 30% lower latency

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 6th November 2019
RESNEXT-101 : INFERNECE

Lowest Latency Comparison: 43x higher throughput | 40x lower latency
Highest Throughput Comparison: 3.4x higher throughput | 18x lower latency

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
GROUP CONVOLUTION KERNELS

up to 77x higher throughput

NOTES:
Results averaged over 10,000 iterations. Filter/kernel size 3x3, field size 7x7, number of filters 512, batch size 32
Repeated for varying group dimensions from standard convolution (512) to full depth-wise (1)
Same code on IPU (Graphcore C2) and GPU using TensorFlow | Forward pass only | Both @ 300W TDP
MCMC PROBABILISTIC MODEL : TRAINING

Customer implementation

26x higher throughput

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

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
MCMC PROBABILISTIC MODEL : TRAINING

VAE model example

4x faster time to train

NOTES:
Scalar control variate used in place of a vector.
IPU: Running on single Graphcore GC2 @ 150W TDP (SDK 1.0.49) | GPU @ 300W TDP
Both using TensorFlow, real data MNIST, Batch size 100 (as in ICML paper).
TIME SERIES ANALYSIS : TRAINING
SALES FORECASTING MODEL | Multi-Layer Perceptron (MLP) + Embedding

>5.5x higher throughput (faster time to train)

NOTES:
Multi-Layer Perceptron (MLP) + Embeddings model for forecasting, Real data
IPU: Graphcore C2 using TensorFlow @ 300W TDP
GPU: using TensorFlow @ 300W TDP
DENSE AUTOENCODER: TRAINING
for content recommendation and ranking

2.2x higher throughput (faster time to train)

Throughput (samples / second)

NOTES:
Deep autoencoder with 6 fully-connected layers and constrained decoder, BS 64 | Content recommendation | Training using open source Netflix 3m data-samples
IPU: Graphcore C2 @ 300W TDP
GPU: @ 300W TDP
REINFORCEMENT LEARNING POLICY: TRAINING

up to 13x higher throughput (faster time to train)

### Throughput (samples / second)

<table>
<thead>
<tr>
<th></th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPU</td>
<td>Batch Size: 4</td>
</tr>
<tr>
<td></td>
<td>0 - 12,500</td>
</tr>
<tr>
<td>IPU</td>
<td>Batch Size: 8</td>
</tr>
<tr>
<td></td>
<td>0 - 25,000</td>
</tr>
<tr>
<td>IPU</td>
<td>Batch Size: 16</td>
</tr>
<tr>
<td></td>
<td>0 - 25,000</td>
</tr>
</tbody>
</table>

#### Batch Sizes

- Batch Size: 4
- Batch Size: 8
- Batch Size: 16

#### Notes:

Reinforcement policy model training | representative of large-scale reinforcement learning systems using LSTM
IPU: Graphcore C2 using TensorFlow @ 300W TDP
GPU: using TensorFlow @ 300W TDP
OUR IPU LETS INNOVATORS CREATE THE NEXT BREAKTHROUGHS IN MACHINE INTELLIGENCE
THANK YOU