A Closer Look At Graphcore ML Performance

NEW RESULTS DEMONSTRATE IMPROVED PERFORMANCE AND SCALABILITY

Graphcore, the UK-based AI Unicorn, submitted a raft of new benchmarks to MLCommons in December, which we covered here. Performance improved significantly with the latest software, model optimizations, and scalability. The company published results on Resnet50 and BERT-large models on scaled-out hardware configurations, from the IPU-POD_{16} up to the POD_{256}, demonstrating near-linear scalability for both. The previous submission was for 16-way systems.

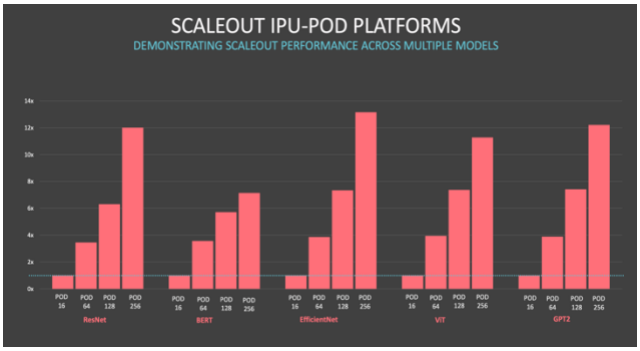
As for competitive comparisons, a 16-way IPU POD showed the roughly equivalent performance to the 8-GPU NVIDIA DGX. Chip-to-chip comparisons are interesting; however, we focus more on the scale-out results since AI training platforms typically deploy at large or even enormous scale.

Graphcore also shared insights beyond the official MLPerf results, which we believe are equally important. Matt Fyles, Software VP at Graphcore, discussed software performance enabled by the latest release, additional benchmarks at scale, an analysis demonstrating CPU-to-Accelerator flexibility, and large model research, which we examine here.

NEW SOFTWARE

Graphcore's latest SDK improves performance and provides scale-out deployment support for up to 256 nodes. As a result, Graphcore customers have realized up to a 50-fold reduction in training time just over the last year. And this is not the final step; we can expect higher levels of scalability going forward. The current MK2 design supports up to 64,000 IPUs once the software can fully enable that scaling level, and of course, a customer wants to buy an entire data center of IPUs.

ADDITIONAL BENCHMARKS AT SCALE



The Graphcore engineering team has gone beyond characterizing the performance of the MLPerf applications, running EfficientNet, ViT vision transformer, and GPT-2. All models scaled nearly linearly, critical for large-scale deployments and a good indicator of the fundamental soundness of the Graphcore hardware and software design. Absolute performance was also impressive, training EfficientNet-B4 on a POD_{16} in only 20.7 hours, compared to an NVIDIA DGX at 70.5 hours.

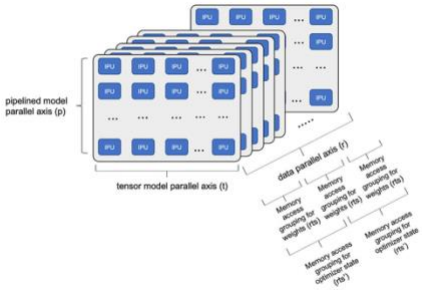
LARGE MODEL RESEARCH

Since Open.ai released access to their 85-billion parameter GPT-3 language model, large AI models have been in the spotlight. Models such as GPT-3 are trained on massive language data sets but have been surprisingly adept at tackling other problems and modalities without retraining. Since

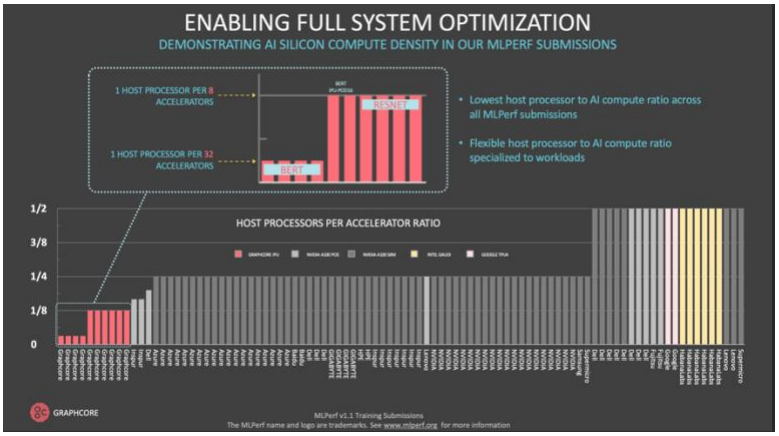
GPT-3, Alibaba, Google, and Microsoft/NVIDIA have announced their massive Large Language Models. We expect to see models exceed ten trillion parameters in 2022.

We anticipated that the large memory footprint (16 TB in an IPU-POD$_{256}$) and fast networking of the IPU-Machine could potentially enable large model training with fewer accelerators, and Graphcore did not disappoint. In a blog posted on December 6, Graphcore's Dave Lacey laid out the path to running brain-scale AI models on IPUs, using the system's optimized data communications, phased execution, and mapping large models across subgraphs.

"With 96 layers in a GPT-3 model, we can still use all the available In-Processor Memory, and all the compute resources efficiently. By expanding to more sub-graphs implemented using Phased execution, we can support even larger models with trillions of parameters. To further reduce training time, we can also scale across multiple, network-connected IPU-POD$_{256}$ systems." -Dave Lacey



Disaggregating CPUs from Accelerators



Graphcore has separated the (expensive) CPUs needed for data preparation and scalar processing in most AI platforms from the accelerators, connected using the company's low-latency protocol running over 100GbE. Some AI models need more or less CPU work than others to run efficiently, and the Graphcore approach provides a dynamic solution, sharing the CPUs across a configurable array of IPU-Machines.

In the case of the MLPerf v1.1 benchmarks, Graphcore only used one host processor per 8 IPUs for Resnet50 and 1 per 32 IPUs for BERT-Large. Other systems that use dedicated custom fabrics like NVIDIA NVLink or PCIe cards cannot do this, incurring large step-functions in incremental cost as you go from 2 CPUs for eight accelerators to 2 CPUs per 4. And these ratios are fixed within the sheet metal of the server, not dynamically allocated as in the IPU-POD architecture. We called this approach a game-changer when it first came out, and we can now imagine (and calculate) the potential TCO savings (acquisition and energy costs) of a giant AI deployment.

Conclusions

When companies have previously announced MLPerf results, the non-benchmark commentary tends to get lost in the head-to-head comparisons, requiring additional parsing to analyze the impacts. In Graphcore's case, the "rest of the story" surpasses the excellent benchmark runs, as it validates the architecture and points to future capabilities, scale, and flexibility that will be available into today's hardware. We look forward to the next chapter in the Graphcore story.

Copyright ©2021 Cambrian AI Research

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR: Karl Freund, Founder, and Principal Analyst at Cambrian-AI Research

INQUIRIES:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

This document, including any supporting materials, is owned by Cambrian-AI Research. This publication may not be reproduced, distributed, or shared in any form without Cambrian-AI Research's prior written permission.

DISCLOSURES

This document was developed with Graphcore funding and support. Although the document may utilize publicly available material from various vendors, including Graphcore, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.

©2021 Cambrian-AI Research. Company and product names are used for informational purposes only and may be trademarks of their respective owners.