OpenGPT-X: Novel Architecture Exploration

1st Chelsea Maria John, 2nd Andreas Herten
Jülich Supercomputing Center (JSC)
Forschungszentrum Jülich
Jülich, Germany
{c.john,a.herten}@fz-juelich.de

***Abstract*—The OpenGPT-X project is a German initiative with ten collaborators to build, train, and deploy a multilingual open-source language model. Models trained within the project will be used for pilot cases by industry partners and commercialized through the Gaia-X Federation. Due to the substantial memory and compute resources required for efficiently training large language models, high-performance computing systems such as JUWELS Booster[1] are essential. This paper presents the results of the exploration of novel hardware architecture conducted within the scope of the project.**

I. Introduction

In recent years, the field of Natural Language Processing (NLP) has witnessed success with Large Language Models (LLMs). The most-recent LLMs, like OpenAI's ChatGPT [1] and GPT-4 [2], have attracted enormous attention and shown how LLMs can assist humans in executing tasks efficiently.

The OpenGPT-X project[2] is a German initiative to build and train large-scale Artificial Intelligence (AI) language models for innovative language applications. The project is a collaborative effort of ten partners from industry and academia (see Figure 1). The models developed within the project will be made compatible with the Gaia-X[3] infrastructure, enabling their use in federated European applications.

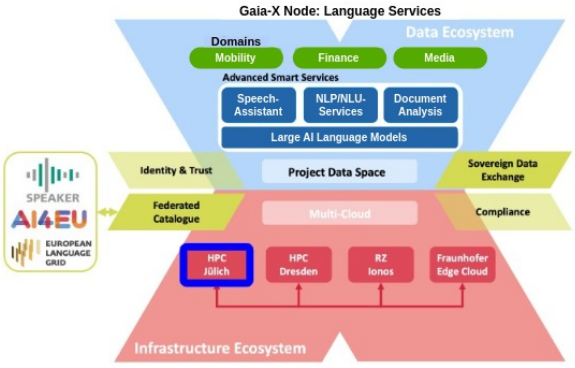


Figure 1. Infrastructure of the OpenGPT-X project. Jülich Supercomputing Centre, JSC, (blue box) provides the JUWELS Cluster and Booster HPC modules for training models.

The key efforts in OpenGPT-X revolve around training multilingual (w.r.t European languages), open-sourced language models. In order to train LLMs, ablation studies on various aspects ranging from data to inference are conducted. This paper delves into initial results regarding the exploration of novel architectures, done under the scope of the OpenGPT-X project.

II. Novel Architecture Exploration

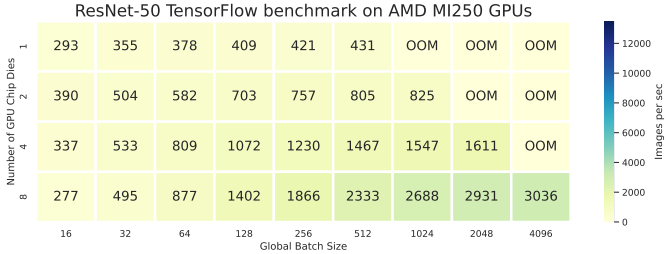
To evaluate future HPC systems for their suitability in NLP, it is crucial to explore a wide range of novel hardware architectures for LLM training. As an initial step in assessing hardware capacities, two benchmarks were evaluated on the resources available in the JURECA-DC[4] supercomputer, especially the JURECA Evaluation Platform[5]. These resources include NVIDIA A100 GPUs (40 GB, SXM), NVIDIA H100 GPUs (80 GB, PCIe), AMD MI250 (64 GB) GPUs, and Graphcore GC200 IPUs (IPU-M2000 POD-4, $\approx$ 260 GB). The first benchmark utilizes TensorFlow ResNet-50 CNNs, offering insights into the overall Machine Learning capacity of the hardware. The second benchmark, derived from the OpenGPT-X fork of Megatron-LM[6], provides us with insights into LLM training.

The Helmholtz AI FZJ fork of TensorFlow ResNet-50[7] was used on the NVIDIA and AMD GPUs. In the case of Graphcore IPUs, a device-optimized version[8] by the vendor was used, since the general TensorFlow setup is not compatible with the IPU architecture.

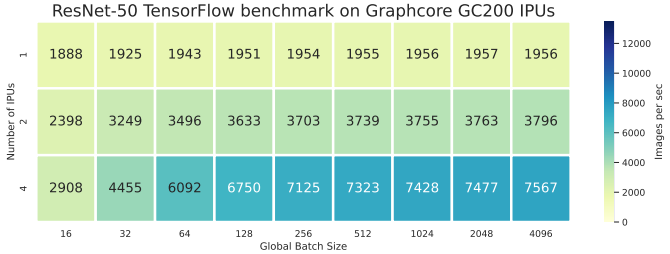
Figure 2 shows heat maps of the training throughput (in images per second) for global batch size plotted against number of devices in a single node. The throughput scales with the global batch size and number of devices. The ResNet-50 model fits into a single device for all the tested hardware, which implies the degree of data parallelism is the same as the number of devices used. The results suggest that Graphcore performs best for small batch sizes, and NVIDIA for large batch sizes. AMD's significantly lower performance warrants further investigation. NVIDIA H100 GPUs, the latest GPU generation, shows $\approx 1.4 - 2\times$ performance compared to NVIDIA A100 GPUs.

The Graphcore IPU has a unique memory architecture with SRAM distributed into an organized set of small independent

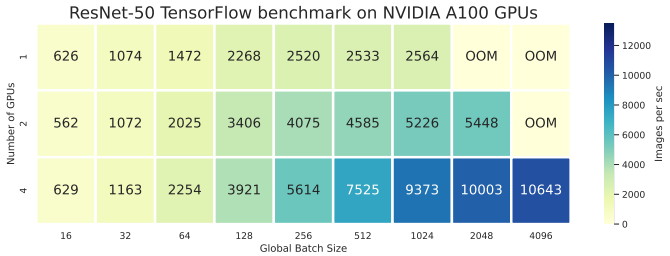
[1]https://doi.org/10.17815/jlsrf-7-183
[2]https://opengpt-x.de/en/
[3]https://gaia-x.eu/
[4]https://doi.org/10.17815/jlsrf-7-182
[5]https://apps.fz-juelich.de/jsc/hps/jureca/evaluation-platform-overview.html
[6]https://github.com/NVIDIA/Megatron-LM.git
[7]https://github.com/HelmholtzAI-FZJ/tf_cnn_benchmarks.git
[8]https://github.com/graphcore/examples.git



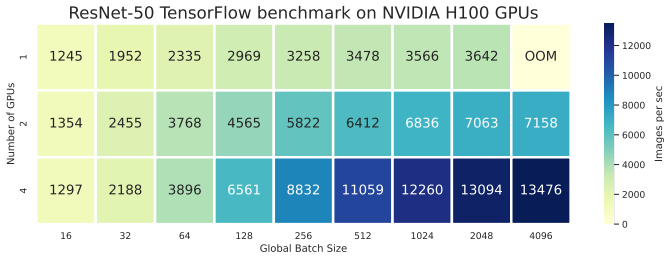
(a) AMD MI250 GPU



(b) Graphcore GC200 IPU



(c) NVIDIA A100 GPU



(d) NVIDIA H100 GPU

Figure 2. Heatmaps: GlobalBatchSize vs. #Devices. Throughput (images per sec) scales with global batch size and number of devices for ResNet-50 model.

memory units, contributing to increased in-processor memory. A set of attached DRAM chips (streaming memory) transfers data to the in-processor-memory via explicit copies within the software. Small batch sizes fit into in-processor memory, which explains the faster throughput, when compared to large batch sizes that would need more communication with the streaming memory. Additionally, the IPU is not a SIMD (Single Instruction Multiple Data) but a MIMD (Multiple Instruction Multiple Data Stream) architecture.

Performance analysis for language model training was also done on a single node (4 GPUs) of the NVIDIA A100 and H100 devices available in JURECA DC and the JU-

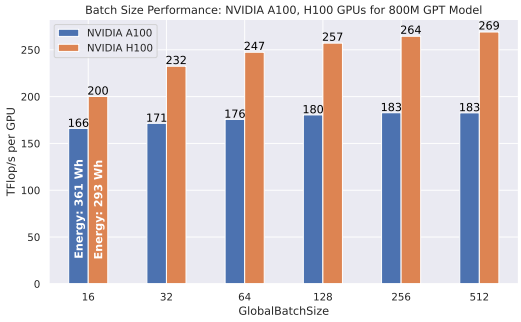


Figure 3. Comparisons of NVIDIA A100 vs. H100 performance (TFlop/s), using a 800M GPT model. Energy usage superimposed for first column pair.

RECA Evaluation Platform. For this, an 800 million parameter GPT model was benchmarked using the OpenGPT-X fork of Megatron-LM. The model fits into a single device and was trained with a data parallelism of 4. Figure 3 shows a bar graph plotting compute throughput in TFlop/s measured per device for different global batch sizes. The NVIDIA H100 improves performance by 1.5 × over the A100, aligning well with the expectations for this latest generation of GPU platform.

Furthermore, to study energy consumption, the 800M GPT model was trained on German data for 1 h on a single node of the NVIDIA A100 and H100 with a data parallelism of 4 and global batch size of 16. The total energy consumed by each device in a node is calculated using power values logged with `nvidia-smi`. A100 GPUs consume an average of 361 Wh and H100 293 Wh, which is an 18.6 % decrease.

III. Conclusion

The ResNet-50 benchmark results shed light on the hardware's performance characteristics. Notably, Graphcore IPUs exhibited superior performance for small batch sizes, whereas NVIDIA GPUs excelled with larger batch sizes. AMD's lower performance raised the need for further investigation. Furthermore, the LLM benchmark revealed that the NVIDIA H100 GPUs, outperformed the NVIDIA A100 GPUs by approximately 1.5 × while consuming less energy.

The results offer valuable insights into the hardware's capabilities, facilitating better-informed choices in hardware selection and optimization within the dynamic field of NLP.

Acknowledgment

OpenGPT-X is funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK) of Germany for the period 2022-2024. We gratefully acknowledge computing time on the JURECA-DC supercomputer and JURECA Evaluation Platform at Forschungszentrum Jülich.

References

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022.

[2] OpenAI, "Gpt-4 technical report," 2023.