CARAML: Systematic Evaluation of AI Workloads on Accelerators

Chelsea John, Stepan Nassyr, Carolin Penke, Andreas Herten

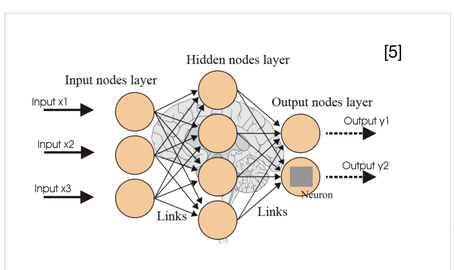
JÜLICH Forschungszentrum

CARAML

- CARAML[1] benchmark suite provides **C**ompact **A**utomated **R**eproducible **A**ssessment of **M**achine **L**earning workloads on novel accelerators
- Automation and reproducibility using **JUBE**[2] benchmarking environment and Apptainer containers
- Performance assessment through curated AI benchmarks in PyTorch and TensorFlow
- Power measurement using **jpwr**[3]

JUBE BENCHMARKING ENVIRONMENT

Energy [4] + **Performance**
- Tokens or images per second
- Tokens or images per Watt hour



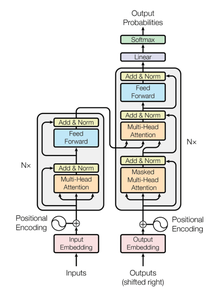
TensorFlow PyTorch

[1]: https://github.com/FZJ-JSC/CARAML
[2]: https://github.com/FZJ-JSC/JUBE
[3]: https://github.com/FZJ-JSC/jpwr
[4]: https://www.vecteezy.com/energyicon
[5]: https://www.analyticsvidhya.com/blog/2016

Benchmarks

Natural Language Processing (NLP)

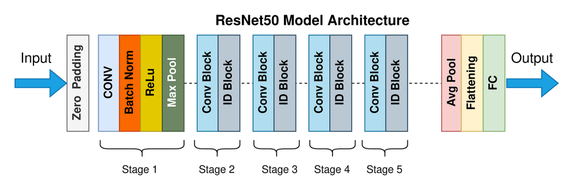
- GPT-based LLM training (OpenGPT-X)
- Curated from forks of Megatron-LM and vendor benchmarks
  - NVIDIA/Megatron-LM
  - Bigcode-project/Megatron-LM
  - Graphcore/examples
  - FZJ-JSC/jubench



GPT Model architecture: Generative Pre-trained Transformer model uses stack of **encoders/decoders** to process data, weighted by **attention** AttentionIsAllYouNeed:https://arxiv.org/abs/1706.03762

Computer Vision (CV)

- ResNet50 model training
- Curated from forks of TensorFlow and vendor benchmarks
  - Tensorflow/benchmarks
  - Graphcore/examples
  - FZJ-JSC/jubench



ResNet50: Residual Networks (ResNet) with 50 Convolutional Neural Network layers used for image classification tasks https://commons.wikimedia.org/wiki/ResNet50

Accelerators

Tested systems as part of JURECA[6], JEDI[7] and WestAI[8]

Platform	GH200 JEDI	GH200 JURECA	H100 JURECA	H100 WestAI	MI200 JURECA	IPU-M2000 JURECA	A100 JURECA
Accelerator	4× NVIDIA GH200-120GB (1× 72 c Grace, 1× H100)	1× NVIDIA GH200-480GB (1× 72 c Grace, 1× H100)	4× NVIDIA H100 GPU (PCIe)	4× NVIDIA H100 GPU (SXM5)	4× AMD MI250 GPU (OAM)	4× Graphcore GC200 IPU	4× NVIDIA A100 GPU (SXM4)
CPU			2× 72 c Intel Xeon Platinum 8452Y	2× 32 c Intel Xeon Platinum 8462Y	2× 48 c AMD EPYC 7443	2× 48 c AMD EPYC 7413	2× 64 c AMD EPYC 7742
CPU–Acc. Connect (intra-node)	NVLink-C2C 900 GB/s		PCIe Gen 5 128 GB/s			PCIe Gen 4 64 GB/s	
Acc.–Acc. Connect (intra-node)[1]	NVLink4 900 GB/s	-	NVLink4[2] 600 GB/s	NVLink4 900 GB/s	Infinity Fabric 500 GB/s	IPU-Link[3] 256 GB/s	NVLink3 600 GB/s
Interconnect internode[4]	4× IB NDR (4×200 Gbit/s)	-	-	2× IB NDR (2×400 Gbit/s)	2× IB HDR (2×200 Gbit/s)	-	2× IB HDR (2×200 Gbit/s)
Memory	4× 120 GB LPDDR5X (CPU), 4× 96 GB HBM3 (GPU)	480 GB LPDDR5X (CPU), 96 GB HBM3 (GPU)	512 GB DDR5-4800 (CPU), 80 GB HBM2e (GPU)	512 GB DDR5-4800 (CPU), 94 GB HBM2e (GPU)	512 GB DDR4-3200 (CPU), 128 GB HBM2e (GPU)	512 GB DDR4-3200 (CPU)	512 GB DDR4-3200 (CPU), 4× 40 GB HBM2e (GPU)
TDP / device	680 W†	700 W†	350 W	700 W	560 W	300 W	400 W
JUBE Tag	JEDI	GH200	H100	WAIH100	MI250	GC200	A100

[1] Bidirectional bandwidths per device.
[2] GPU0 and GPU1 and GPU2 and GPU3 are connected through NVLink bridges, each with 12 NVLink4 connections (each 25 GB/s).
[3] Each IPU in a node is connected to other IPUs in- and out-of-node with 10 *IPU-Links*. Intra-node, an IPU connects to two other IPUs with 2 links, and with one IPU with 4 links. At 32 GB/s bidirectional bandwidth per link, an IPU has hence an accumulated intra-node connection bandwidth of 256 GB/s.
[4] NVIDIA InfiniBand is abbreviated to *IB*.
† The TDP for the GH200 superchips is for the full package, i.e. including the CPU and GPU devices.

[6]: https://apps.fz-juelich.de/jsc/hps/jureca/evaluation-platform-overview.html
[7]: https://apps.fz-juelich.de/jsc/hps/jedi
[8]: https://westai.de/

Benchmark on Graphcore GC200 IPU

NLP Benchmark

- Trained **117M GPT model** (synthetic data) using modified Graphcore benchmark with energy
- Single M2000 POD4 can train 117M GPT model only with pipeline parallelism of 4 and no data parallel
- Performance increases with batch size

Batch Size	Tokens/Time 1/s	Energy/Epoch/IPU Wh	Tokens/Energy 1/Wh
64	64.99	15.68	4.08
128	97.21	18.20	7.03
256	129.96	18.37	13.93
512	155.72	18.56	27.60
1024	172.94	19.07	53.71
2048	183.37	20.05	102.13
4096	188.88	21.88	187.22
8192	191.86	25.47	321.34
16384	193.41	33.00	496.43

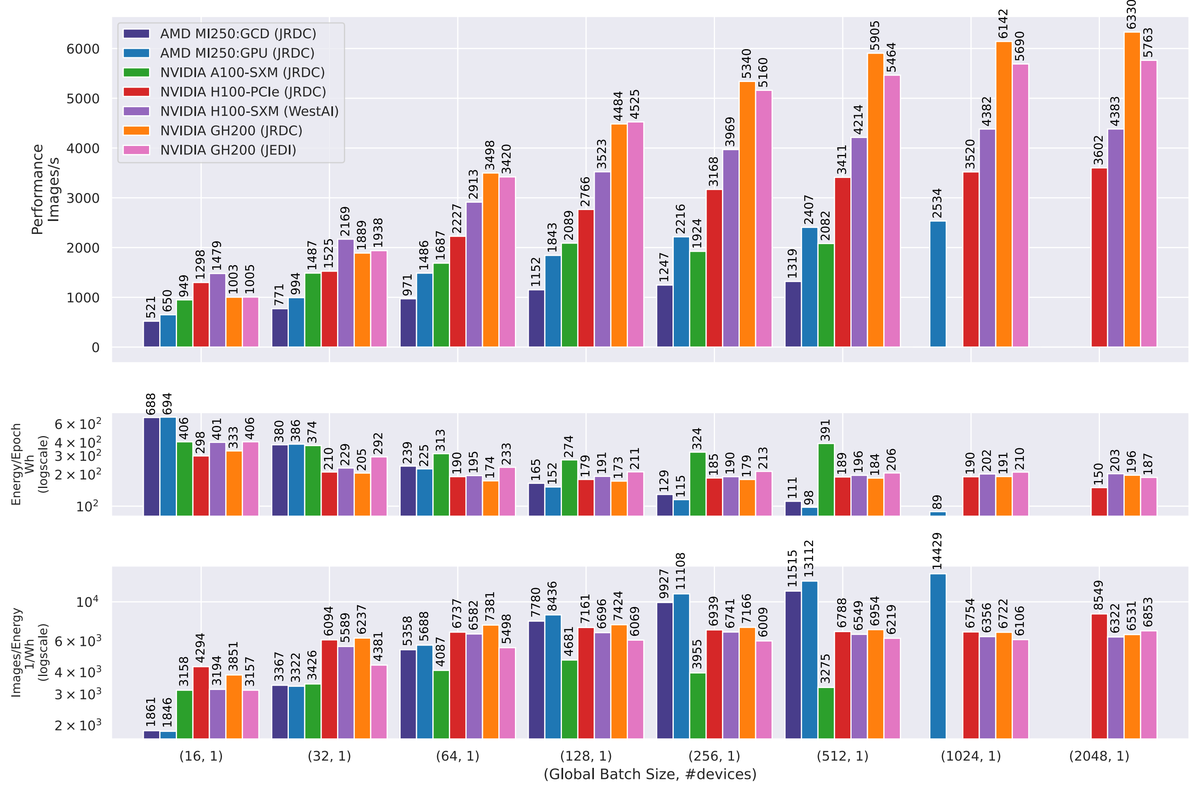
CV Benchmark

- Trained **ResNet50 (**ImageNet data) on 1 IPU using modified Graphcore benchmark with energy
- Performance does not scale with batch size due to multiple sequential calls to DRAM to combat limited SRAM
- Energy efficiency is promising compared to GPUs
- Model graph compilation (≈ 60 min) time is excluded from result

Batch Size	Images/Time 1/s	Energy/Epoch Wh	Images/Energy 1/Wh
16	1827.72	32.09	39925.87
32	1857.90	31.73	40382.19
64	1879.29	31.75	40346.18
128	1888.11	31.67	40452.50
256	1887.23	31.58	40563.65
512	1891.74	31.49	40689.85
1024	1893.07	31.50	40668.79
2048	1889.87	31.53	40636.28
4096	1891.58	31.51	40660.14

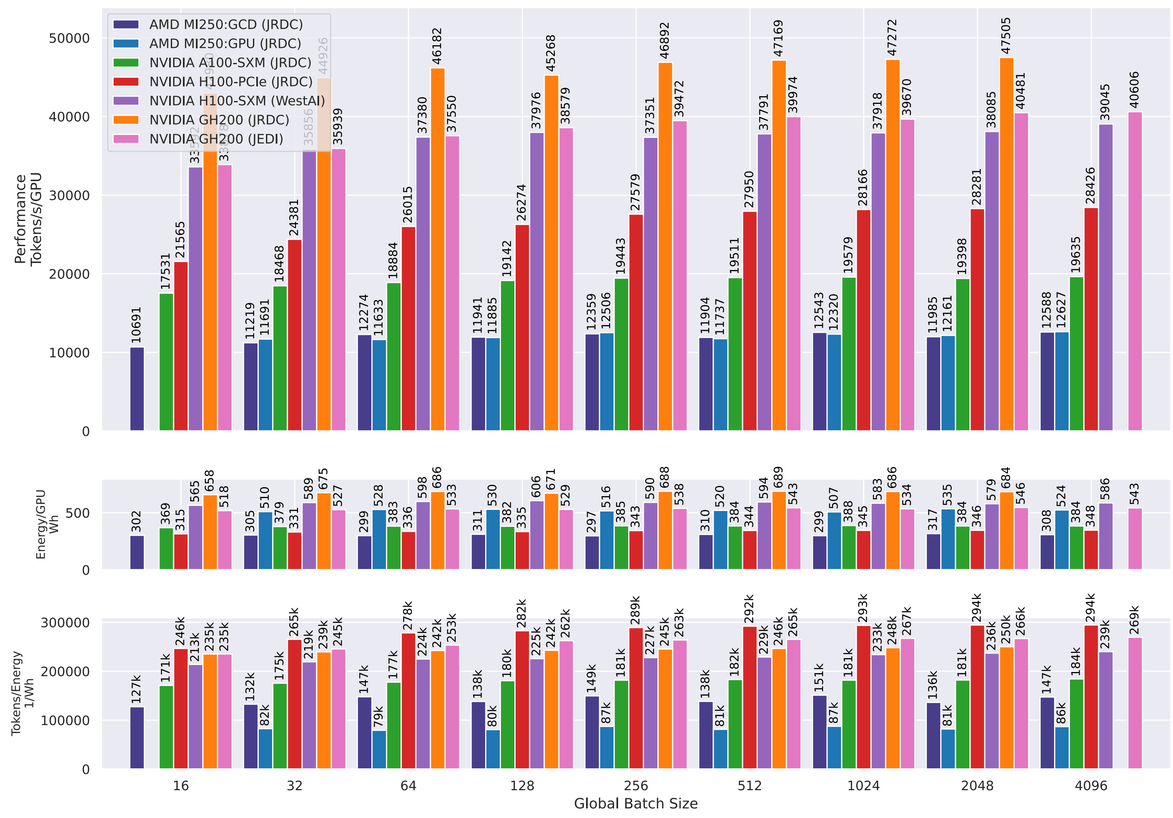
CV Benchmark on GPUs

ResNet-50 TensorFlow Benchmark on 1 Device of Nvidia & AMD Systems
Using ImageNet Data (1 Epoch = 1281167 Samples)



NLP Benchmark on GPUs

Batch Size Training Performance on1 Node of NVIDIA & AMD GPUs Using Megatron-LM
with **800M GPT Model** on OSCAR Data



Conclusions

Performance Trends
- Recent GPU hardware generations show improved performance, with GH200 nodes leading
- GH200 (JEDI) trails GH200 (JRDC) due to data parallelism communication overhead
- H100 (SXM) outperforms H100 (PCIe) due to NVLink's higher bandwidth and SXM GPU form factor
- Distributed training results are detailed in the CARAML paper: https://arxiv.org/pdf/2409.12994

NLP Benchmark
- AMD MI250 with 4 GCDs (2 GPUs) slightly outperforms 8 GCDs (4 GPUs), reflecting data parallelism overhead
- GC200 IPU increases tokens/s with batch size but is less efficient than GPUs due to pipeline bubbles
- H100 (PCIe) leads in energy efficiency due to its PCIe card's power limitations

CV Benchmark
- AMD MI250's throughput is higher using 2 GCDs compared to a single GCD.
- GC200 IPU achieves saturated performance with limited SRAM.
- AMD MI250 is more energy efficient for larger batch sizes, while GH200 and H100 excel with smaller batch sizes

Challenges & Next Steps

- Achieving comparability across accelerators is challenging
- Aligning containers with HPC environments and SLURM schedulers is complex
- Network and system-specific optimizations are essential for better results
- Broaden support for accelerators and add more AI workloads

Acknowledgements

This work was funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through OpenGPT-X (project no. 68GX21007D) and supported by EuroHPC Joint Undertaking under Grant 955513, co-funded by the German Ministry of Education and Research (BMBF, ref. 16HPC029) through MAELSTROM. We also acknowledge the use of JURECA-DC, JURECA-DC Evaluation Platform, WestAI infrastructure, and the JUPITER platform JEDI.

JÜLICH Forschungszentrum

openGPT-X

Bundesministerium für Wirtschaft und Klimaschutz