

Novel Architecture Exploration

OpenGPT-X: Open Large Language Models

<https://www.opengpt-x.de/>

Chelsea John, Dr. Andreas Herten
Women in HPC SC23 Workshop
Denver, United States.



13.11.2023

Novel Architecture Exploration

- AI models on hardware architectures
 - ❖ Performance?
 - ❖ Energy-efficiency?
 - ❖ Model vs. hardware?
- Define reference benchmarks
- Using two types of models as benchmarks
 - ❖ TensorFlow ResNet-50 Convolutional Neural Network ^[1,2]
 - ❖ PyTorch Megatron-LM^[3] (derived from OpenGPT-X fork)

[1]: https://github.com/HelmholtzAI-FZJ/tf_cnn_benchmarks

[2]: <https://github.com/graphcore/examples>

[3]: <https://github.com/NVIDIA/Megatron-LM>

Tested Hardware



- A100 Node (4x 40 GB GPUs, SXM)^[4]
- H100 Node (4x 80 GB GPUs, PCIe)^[5]



- MI200 Node (4x 128 GB MI250 GPUs)^[5]

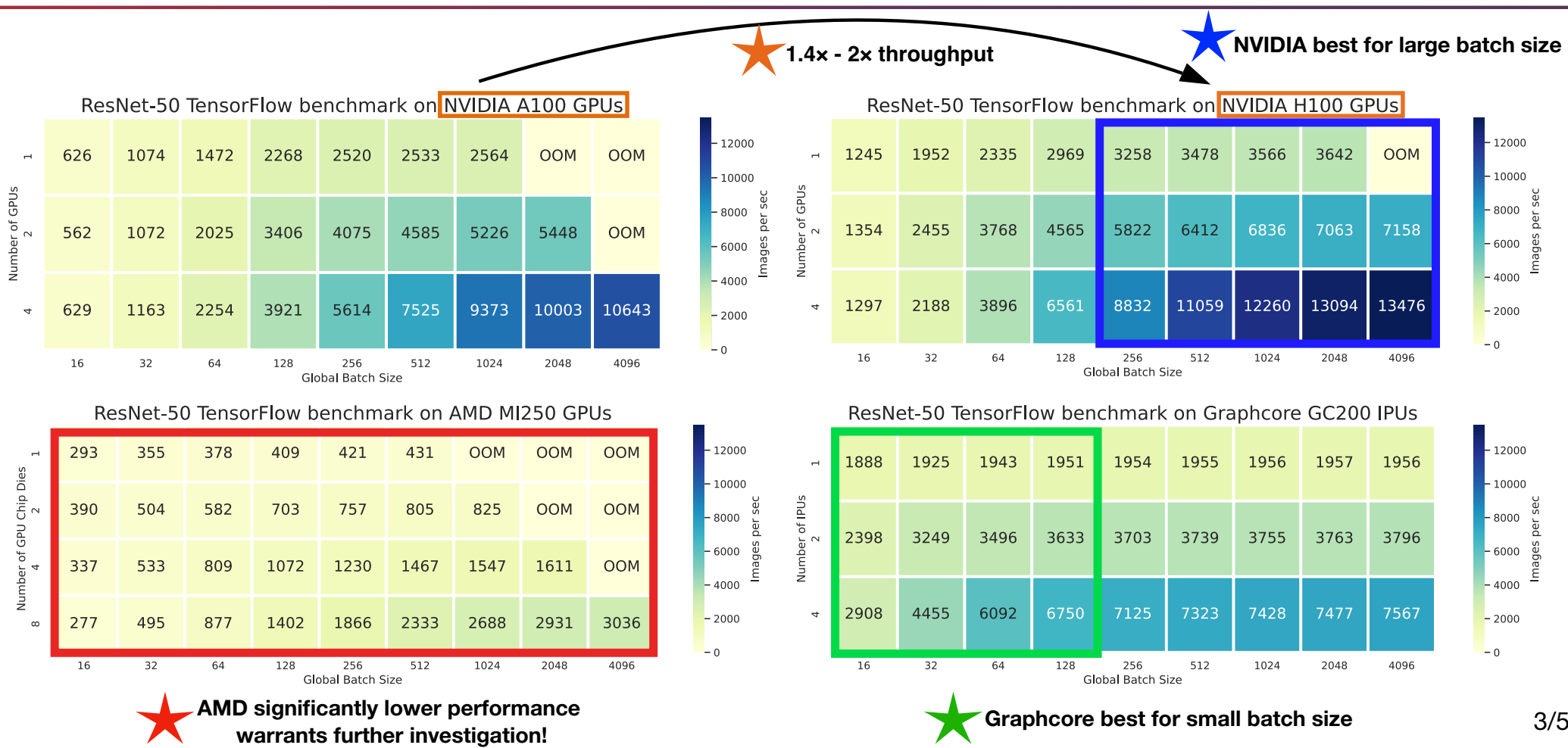


- M2000 POD4 (4x GC200 IPU, ≈ 260 GB)^[5]

[4]: [JURECA-DC](#). [5]: [JURECA Evaluation Platform](#)

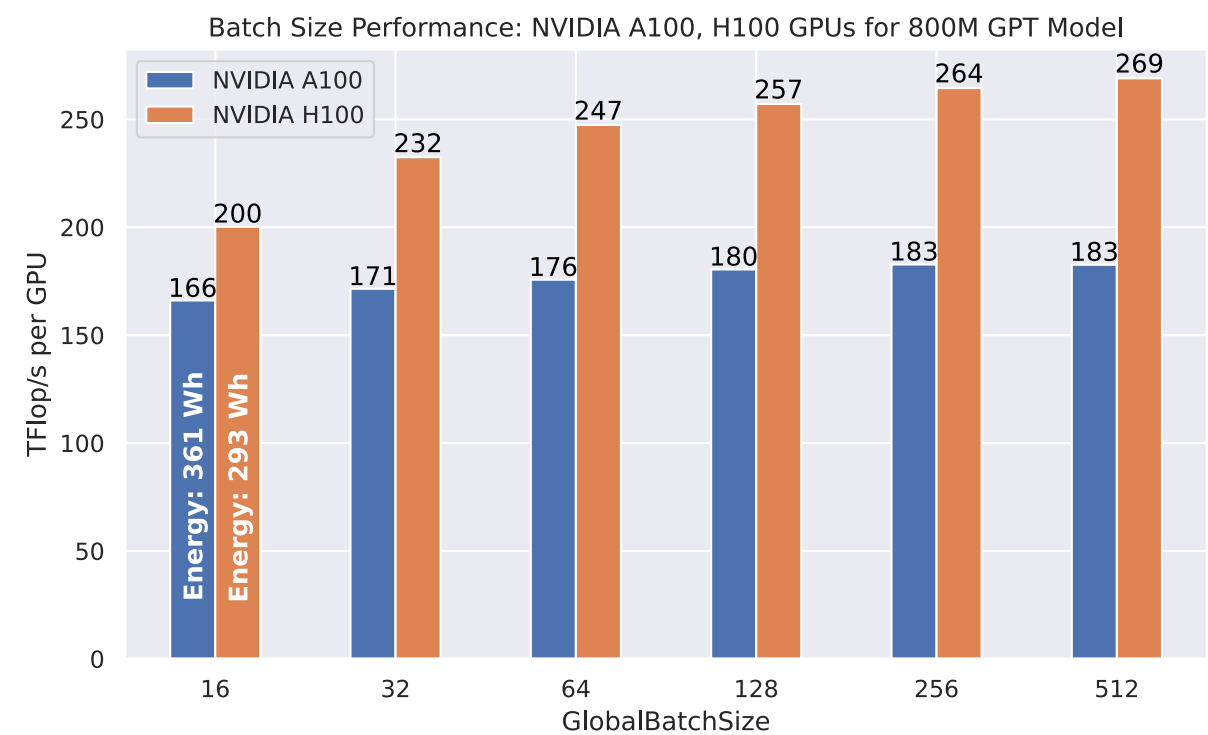
ResNet-50 Benchmark Result

Heat-maps: GlobalBatchSize vs. #Devices showing throughput in images per sec




Megatron-LM Benchmark Result

NVIDIA A100 vs. H100 performance (TFlop/s) against Batch Size



- **800 Million** parameter **GPT Model**
- Trained on single node with **4 GPUs**
- Model replicated 4× (Data Parallel = 4)
- Energy measured using nvidia-smi



**1.5× performance,
19% less energy**

Conclusion

- Reference benchmarks for Computer Vision and NLP
- IPU architecture works best for small batch sizes that fit into in-processor memory
- GPU architecture works best for large batch size and scaling
- NVIDIA H100 GPUs performs 1.5× - 2× than NVIDIA A100 with 19% less energy consumption

Thank you!
c.john@fz-juelich.de

