# Pretraining a foundation model using MODIS observations of the earth's atmosphere

V. Anantharaj[1], Takuya Kurihana[2], Gabriele Padovani[3], Ankur Kumar[4], Aristeidis Tsaris[1], Udayshankar Nair[4], Sandro Fiore[3] and Ian Foster[2,5]

[1]Oak Ridge National Laboratory, [2]European University of Chicago, [3]University of Trento, [4]University of Alabama Huntsville, [5]Argonne National Laboratory

EGU24-22461

## Opportunities and challenges

The earth and atmospheric sciences research community has an unprecedented opportunity to exploit the vast amount of data available from earth observation (EO) satellites and earth system models (ESM). Smaller and cheaper satellites with reduced operational costs have made a variety of EO data affordable, and technological advances have made the data accessible to a wide range of stakeholders, especially the scientific community (EY, 2023). The NASA ESDS program alone is expected to host 320 PB of data by 2030 (NASA ESDS, 2023). The ascent and application of artificial intelligence foundation models (FM) can be attributed to the availability of large volumes of curated data, accessibility to extensive compute resources and the maturity of deep learning architectures, especially the transformer (Bommasani et al., 2021).

Developing a foundation model involves pretraining a suitable deep learning architecture with large amounts of data, often via self supervised learning (SSL) methods. The pretrained models can then be adapted to downstream tasks via fine tuning, requiring less amount of data than task-specific models. Large language models (LLM) are likely the most common type of foundation encountered by the general public. Vision transformers (ViT) are based on the LLM architecture and adapted for image and image-like data (Dosovitskiy, et. al., 2020), such as EO data and ESM simulation output.

We are in the process of pretraining a Shifted Window Transformer, abbreviated as SwinT-V2 (Liu et al, 2021 and 2022), model for the earth's atmosphere using a select few bands of 1-km Level-1B MODIS radiances and brightness temperatures, from the NASA Terra and Aqua satellites respectively. We are planning to use 200 million image chips of size 128x128 pixels. We are exploring three SwinT-V2 models of sizes 100 million and 600 million and 1.4 billion parameters respectively. The pretrained models will be finetuned for cloud classification and evaluated against AICCA. We will discuss our experiences involving data and computing experiments, and present preliminary results.

## Computational studies and scaling explorations

We have designed two sets of experiments to understand the computational and memory requirements of various SwinT-VZ2. model configurations. All experiments were run on the OLCF Frontier supercomputer, each node with 4X AMD Instinct MI250X GPUs.

| | Model Parameters | | | |
|---|---|---|---|---|
| | 100M | 600M | 1.4B | 3.6B |
| Model | 838.7 MB | 2.39 GB | 5.69 GB | 14.41 GB |
| Batch | 67.15 MB | 67.15 MB | 67.15 MB | 67.15 MB |
| Gradients | 381.47 MB | 2.27 GB | 5.28 GB | 13.51 GB |
| Optimizer | 381.47 MB | 2.27 GB | 5.28 GB | 13.51 GB |
| Total | 1.668 GB | 7.98 GB | 16.3 GB | 41.6 GB |

### Experiment Group 1

Vary sequence length (tokens) for fixed model size (1.4B)

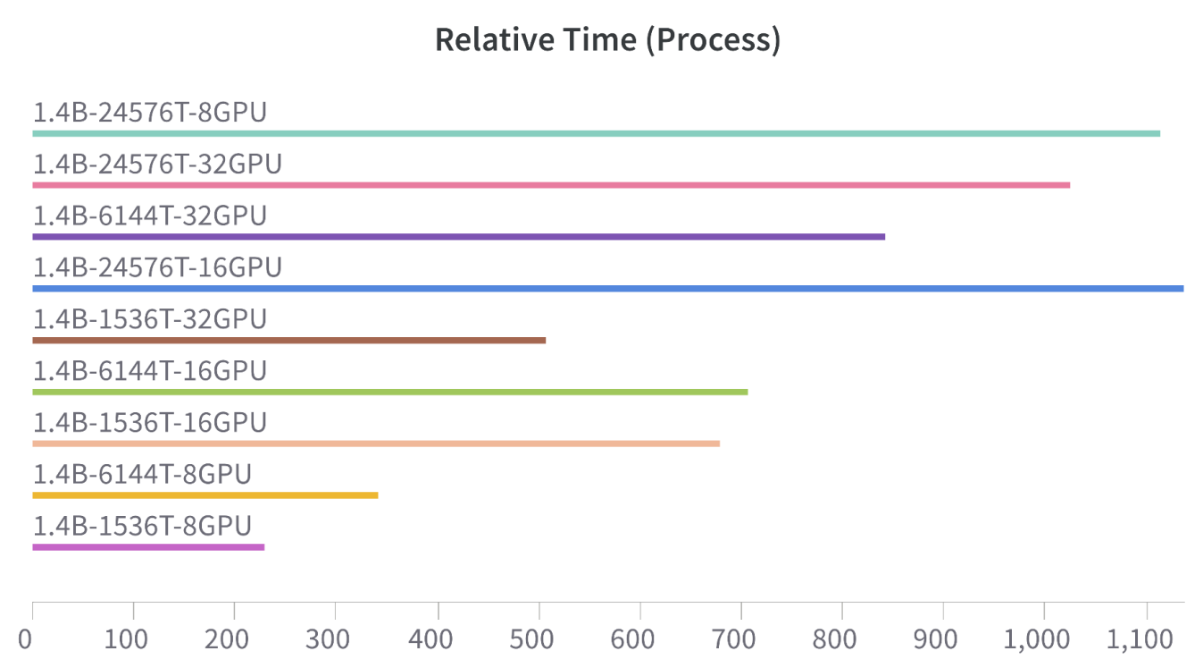
- 9 experiments completed
  - Tile size 128x128; Patch size 8x8: (128x128) / (8x8) * 6 = 1,536
  - Tile size 128x128; Patch size 4x4: (128x128) / (4x4) * 6 = 6,144
  - Tile size 128x128; Patch size 2x2: (128x128) / (2x2) * 6 = 24,576
- Runs using 8, 16, 32 GPUs (1, 2, 4 nodes)
- Number of samples (tiles): 3,000
- Batch size: 32 samples; 4 for 24K tokens
- Max WCT 20 minutes (or 15 epochs)
- Fixed learning rate: 0.0001
- Testing: 1,000 samples (tiles)

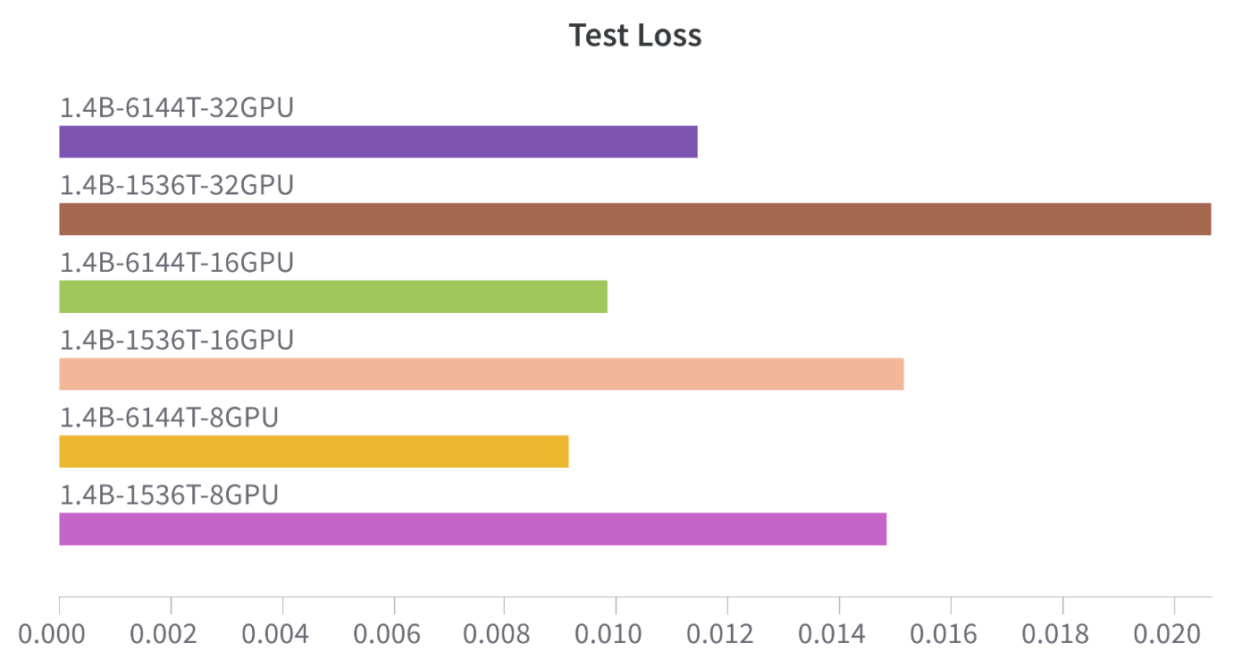
### Experiment Group 2

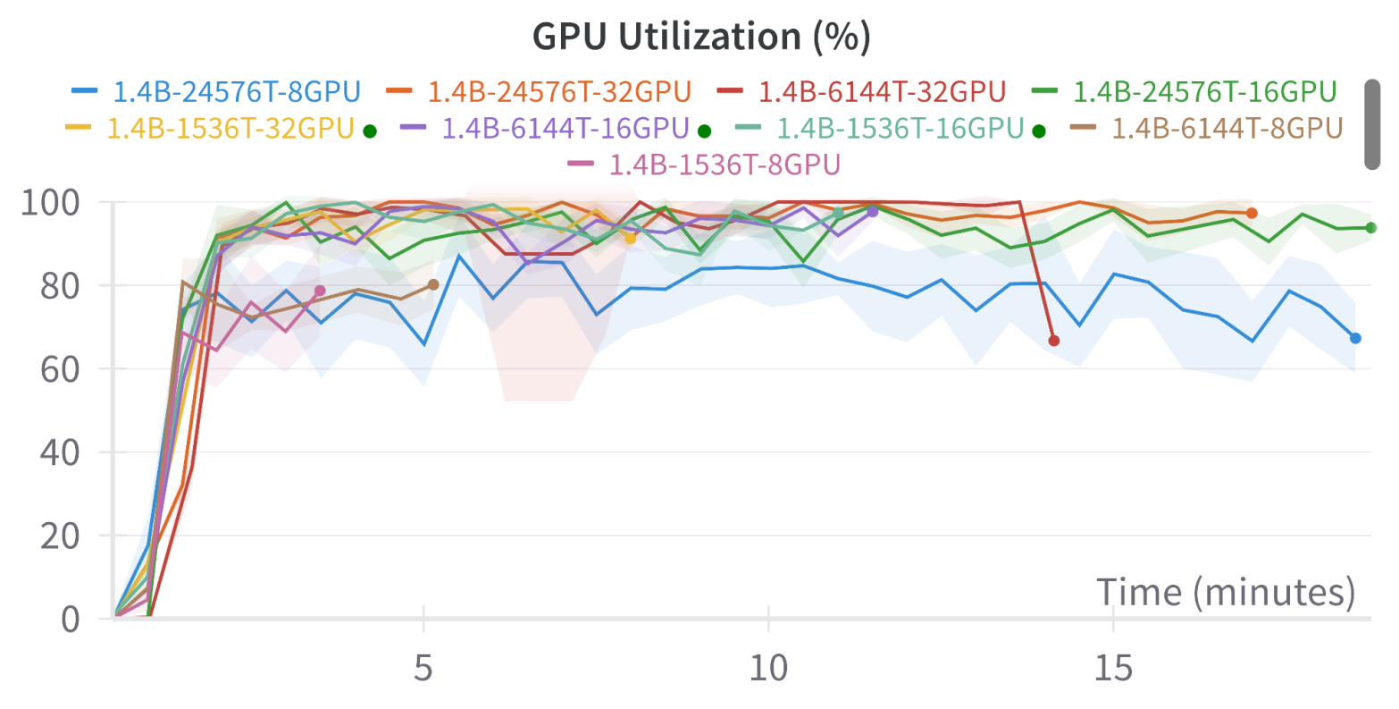
Vary model sizes and number of GPUS for fixed data samples

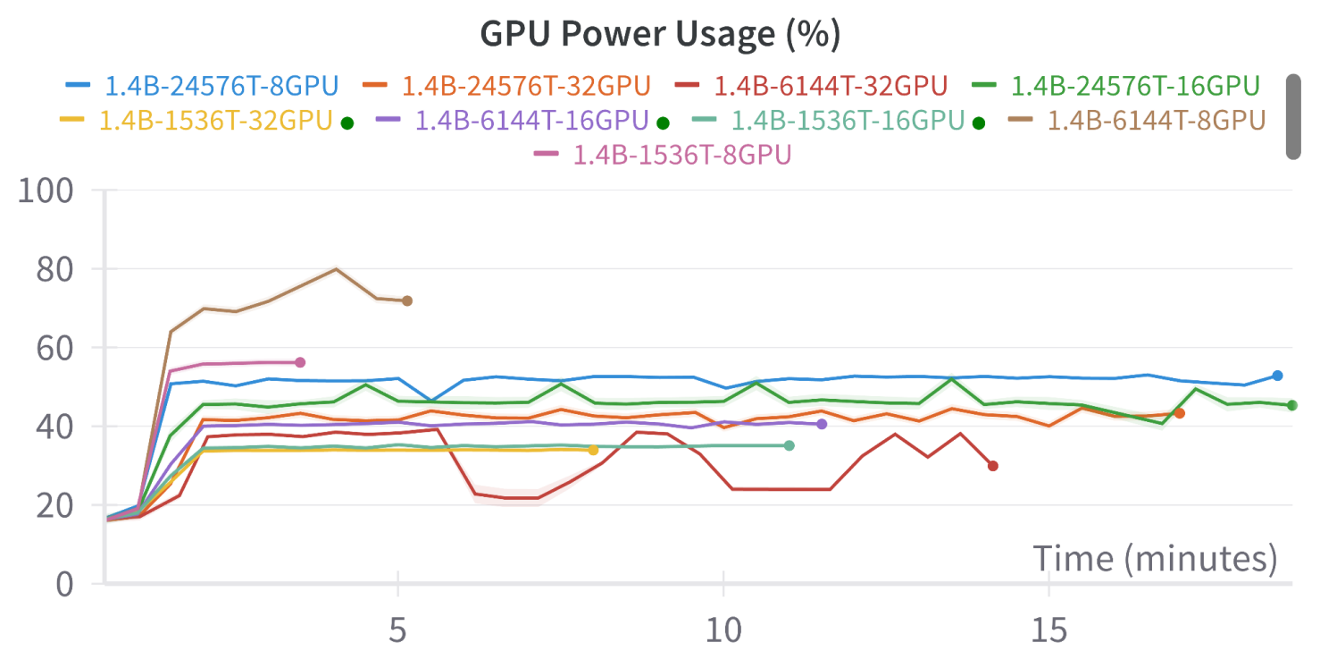
- 9 experiments
  - Model sized 100M, 600M and 1.4B
  - Runs using 8, 16, 32 GPUs (1, 2, 4 nodes)
- Number of samples (tiles): 2M
- Batch size: 64 samples
- Learning rate: 0.0001
- Training loss: Cross entropy

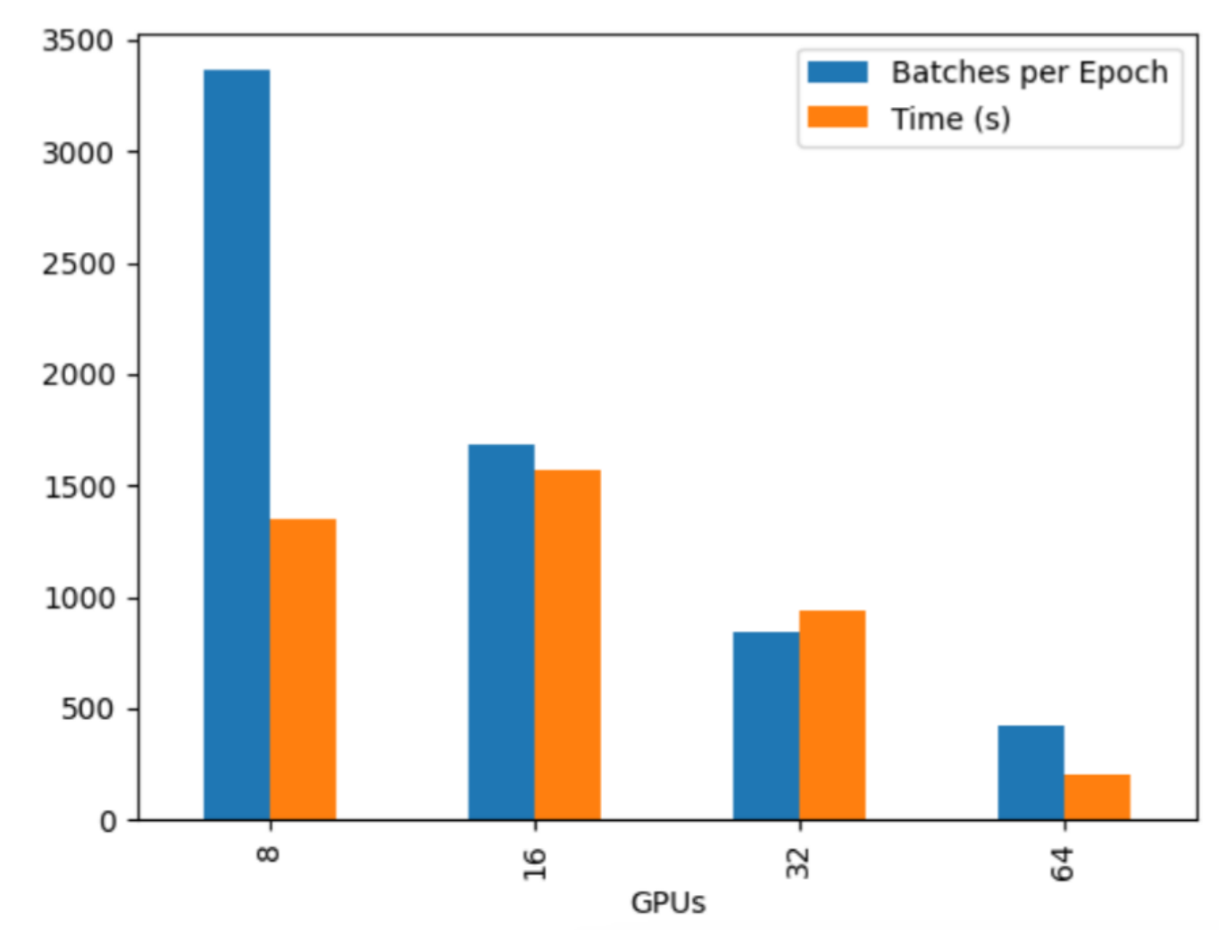
## Preliminary results



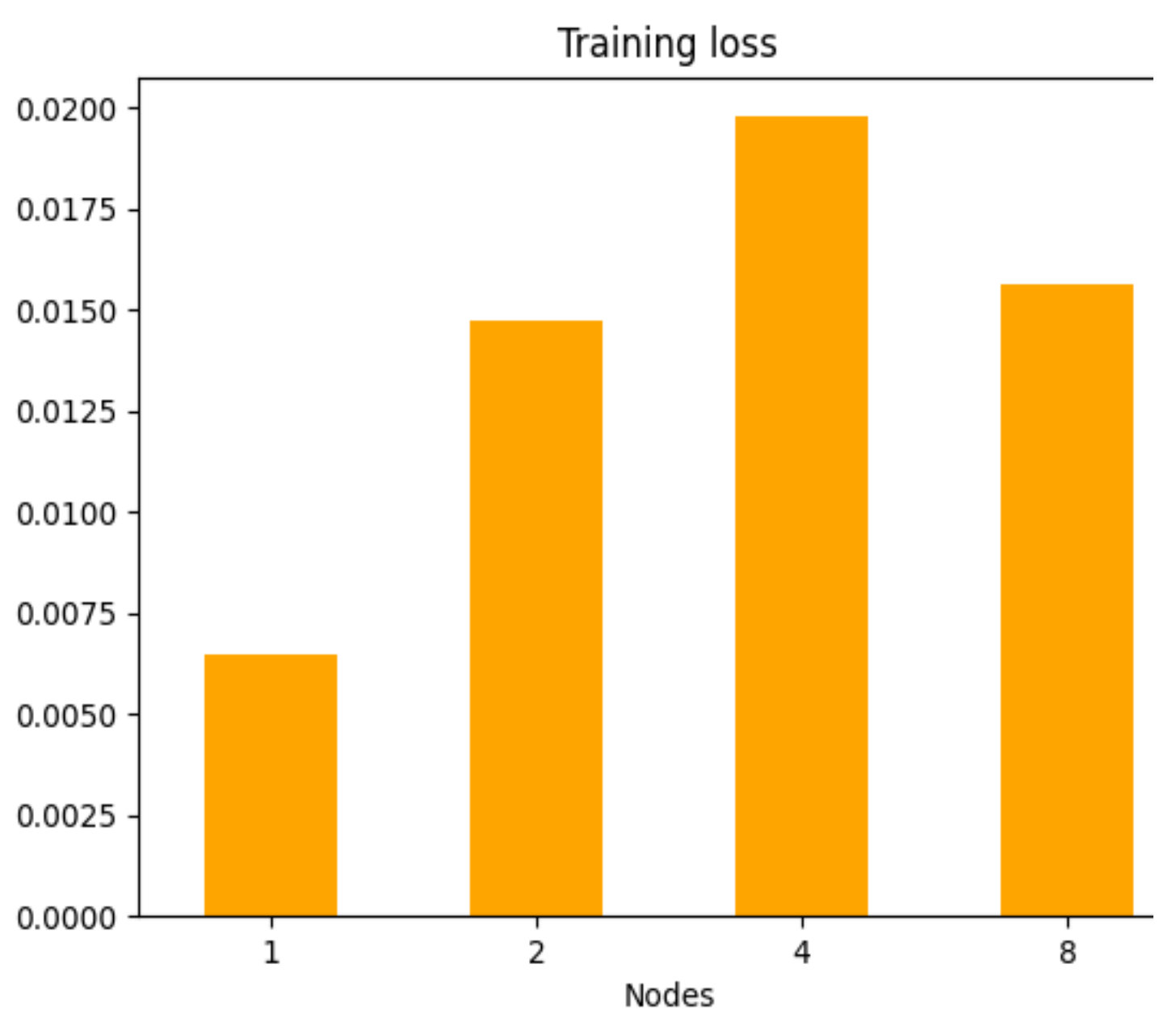


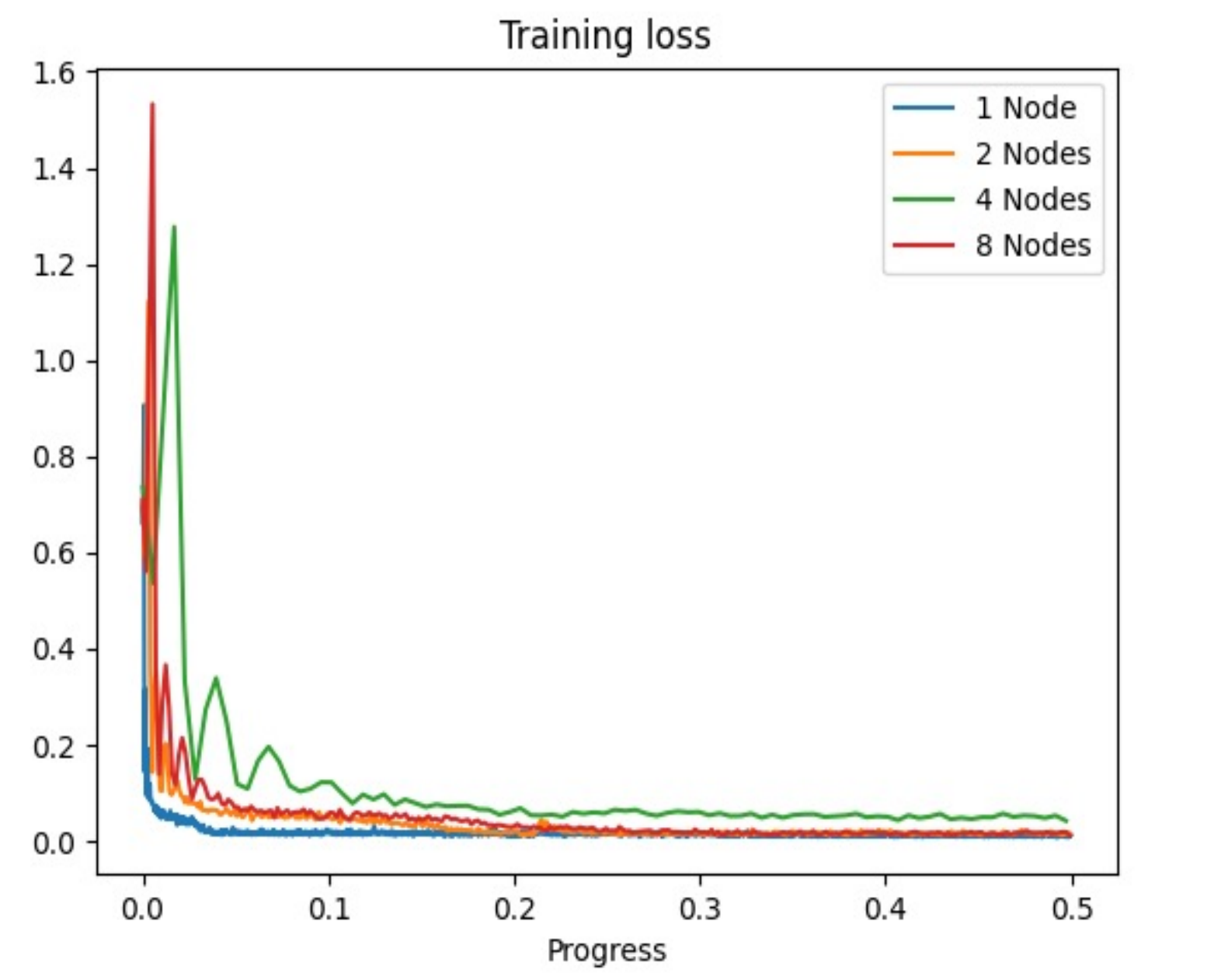






| Number of GPUs | | | | |
|---|---|---|---|---|
| Metric | 8 | 16 | 32 | 64 |
| CPU Usage (%) | 4.03 | 2.4 | 2.29 | 3.09 |
| GPU memory usage (%) | 72.2 | 91.5 | 94.7 | 75.7 |
| Memory usage (%) | 11.55 | 7.75 | 7.75 | 10.21 |
| Emission rate (Kg/s) | 0.26 | 0.006 | 0.0048 | 0.267 |

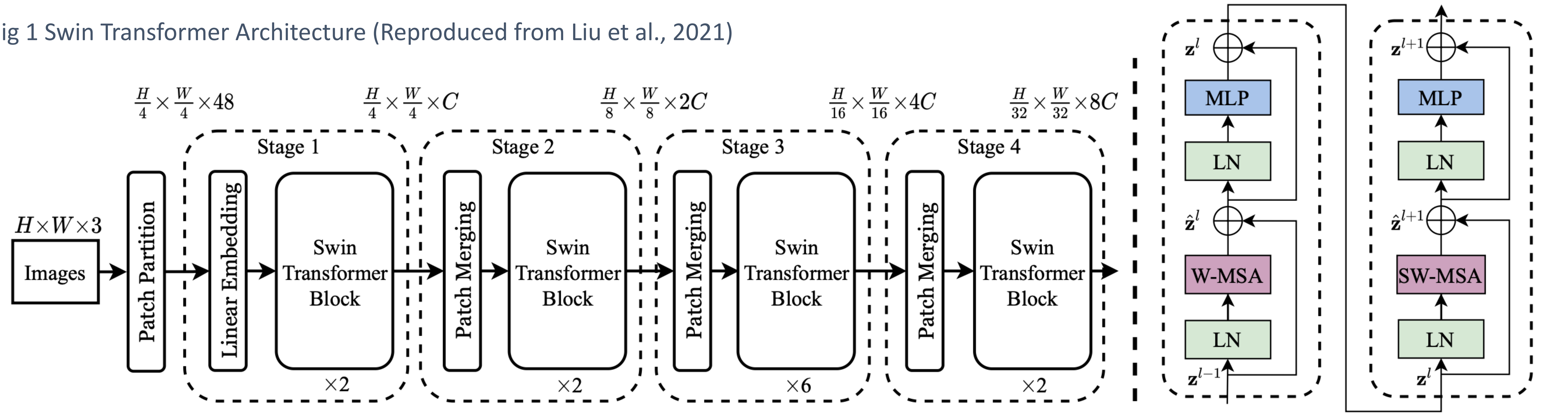




## Early observations and lessons learned

- Longer sequence length could not complete within wallclock budget.
- Training loss converge more or less at the same time (~100 steps) for smaller samples)
- Slightly better testing loss for longer sequence length
- GPU utilization: 1 node / 1.5K Tokens less utilization and finished faster but used more power. More nodes less utilization. Need to optimize usage.
- Training loss: lower for lower number of nodes
- Time vs batches/epoch is reasonable

Fig 1 Swin Transformer Architecture (Reproduced from Liu et al., 2021)



(a) Architecture

(b) Two Successive Swin Transformer Blocks

## References and Acknowledgments

Bommasani, R., et al., 2021: On the opportunities and risks of foundation models. CoRR abs/2108.07258. https://arxiv.org/abs/2108.07258,
Dash, S., et al., 2023: Optimizing Distributed Training on Frontier for Large Language Models. https://arxiv.org/abs/2312.12705
Dosovitskiy, A., et al, 2020.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
Li, Z., et al., 2021: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. https://arxiv.org/pdf/2103.14030.pdf
Li, Z., et al., 2022: Swin Transformer V2: Scaling Up Capacity and Resolution. https://arxiv.org/pdf/2111.09883.pdf
Muennighoff, N., et al., 2023: Scaling Data-Constrained Language Models. https://proceedings.neurips.cc/paper_files/paper/2023/file/9d89448b63ce1e2e8dc7af72c984c196-Paper-Conference.pdf

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. ECMWF also benefited from collaborations funded via ESCAPE-2 (No. 800897), MAESTRO (No. 801101), EuroEXA (No. 754337), and ESiWACE-2 (No. 823988) projects funded by the European Union's Horizon 2020 future and emerging technologies and the research and innovation programmes.