Power and Performance Analysis of AI Models on NVIDIA Grace Hopper Superchips

Alvajoy Asante, College of Computing and Informatics
Tyler Allen, College of Computing and Informatics

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Introduction

- **Large Language Models (LLM) -** powerful tools in natural language processing, leveraging vast amounts of internet text for inference tasks
- **High performance computing (HPC)** - Also known as supercomputers use of powerful of computers and specialized techniques to tackle complex tasks.
- **Nvidia Grace Hopper Superchip 200 (GH2) -** Referring the new Grace Hopper Superchip 200. (Specs: 900W Power consumption, 96GB of memory, $40K)
- **Motivation** - With ever growing powerful chips there is a need to Improve the AI model power consumption on HPC such as the GH2.

Objectives

- Identify power and memory consumption with model training on GH2.
- Analysis correlation between a LLM size and power consumption.
- Explore solutions to improve LLM models on GH2.
- Collect benchmarks of the CPU and GPU during model training of LLM.

Research Question:

Q1. Will model size affect the overall power consumptions on GH2?

Q2. How can we effectively train an LLM with lower power consumption on GH2?

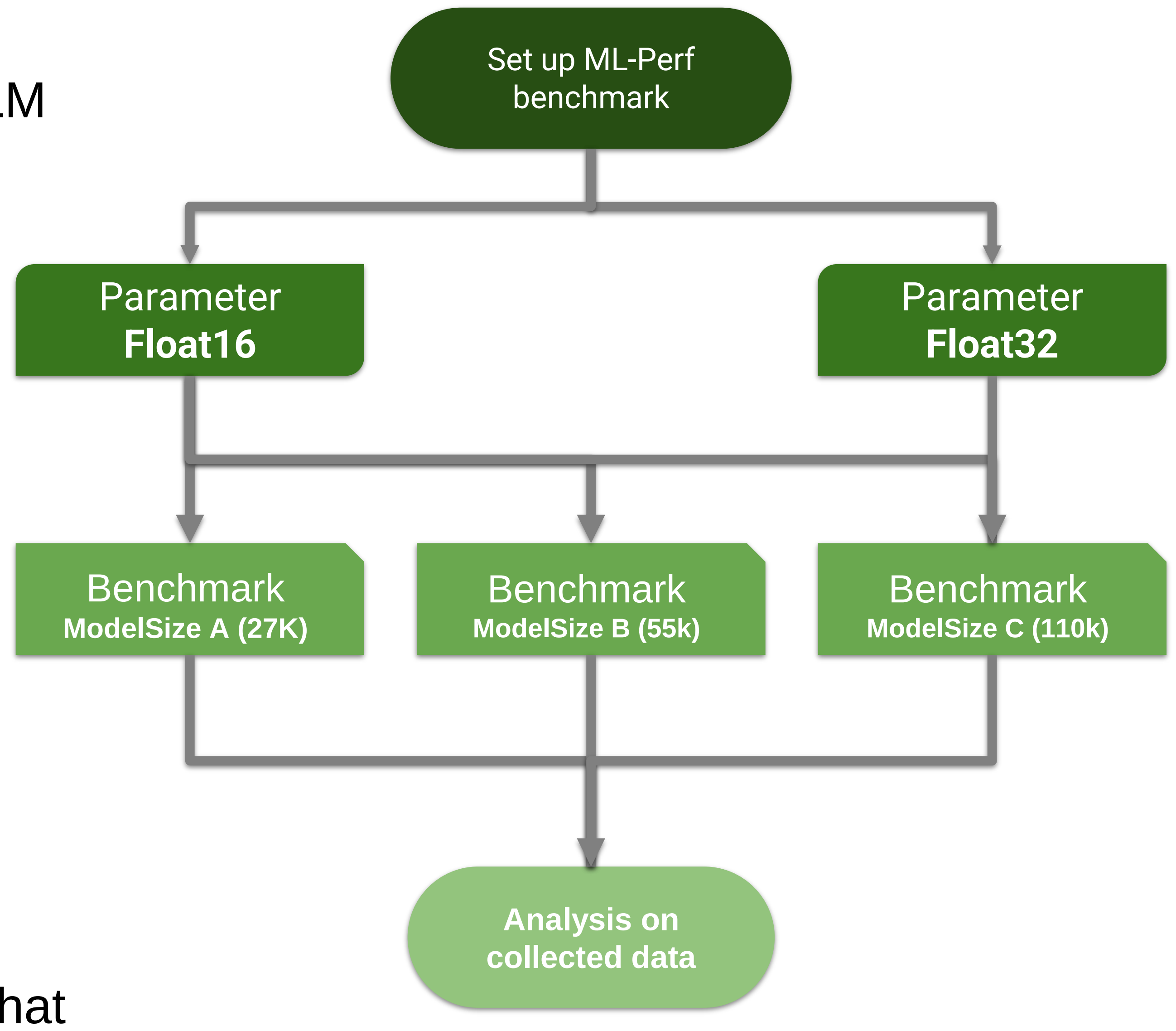
Methodology

Utilized Nvidia ML-Perf, configured the training benchmarks for each selected LLM to run GH2.

Variation in Model Testing

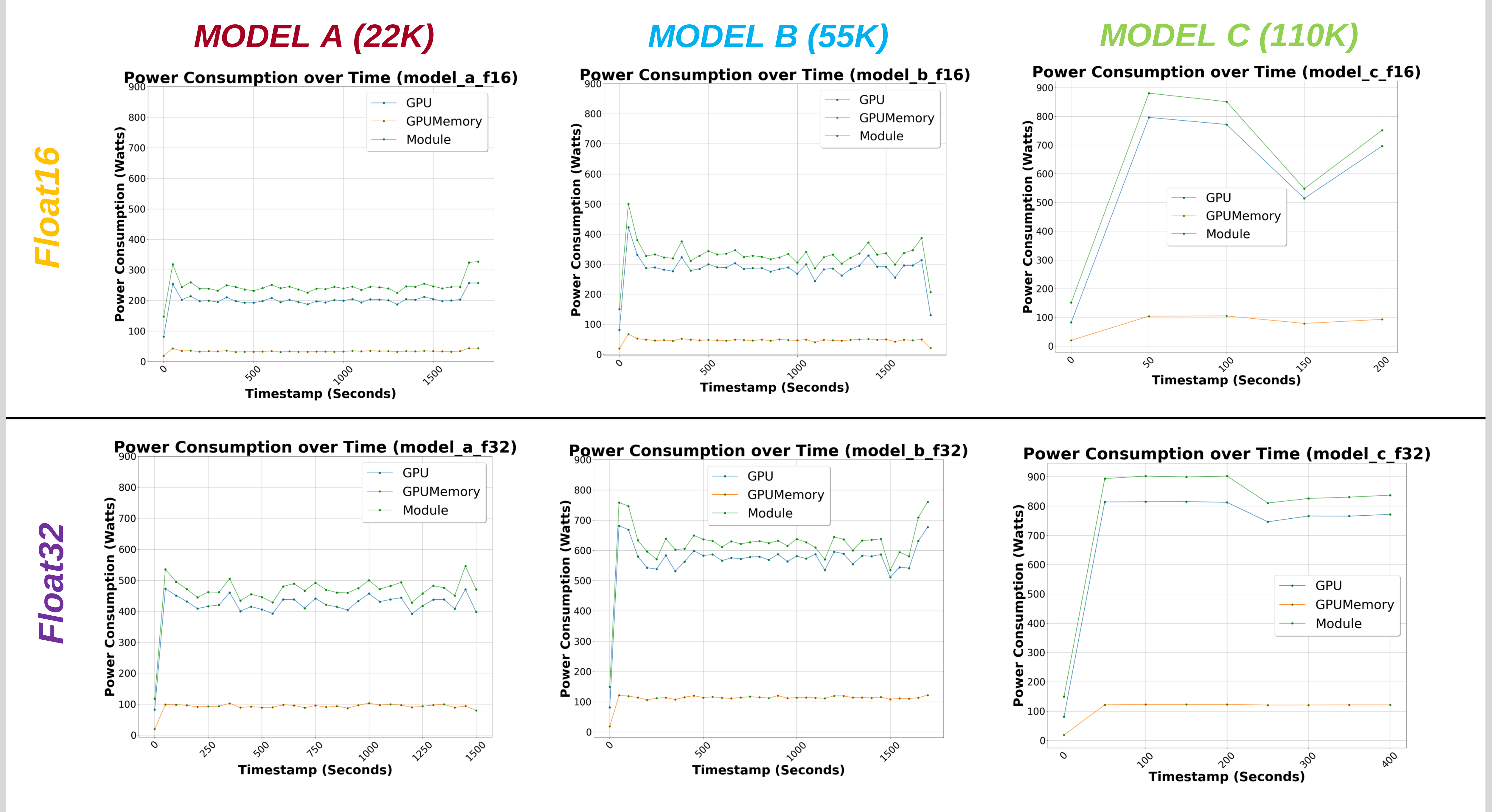
- Tested 3 different model sizes restricted to memory utilization
  - Model memory from 50% - 70%
- Used floating-point precision for the model weights
  - Float16 (F16) vs. Float32 (F32)
- Collected CPU utilization and GPU utilization

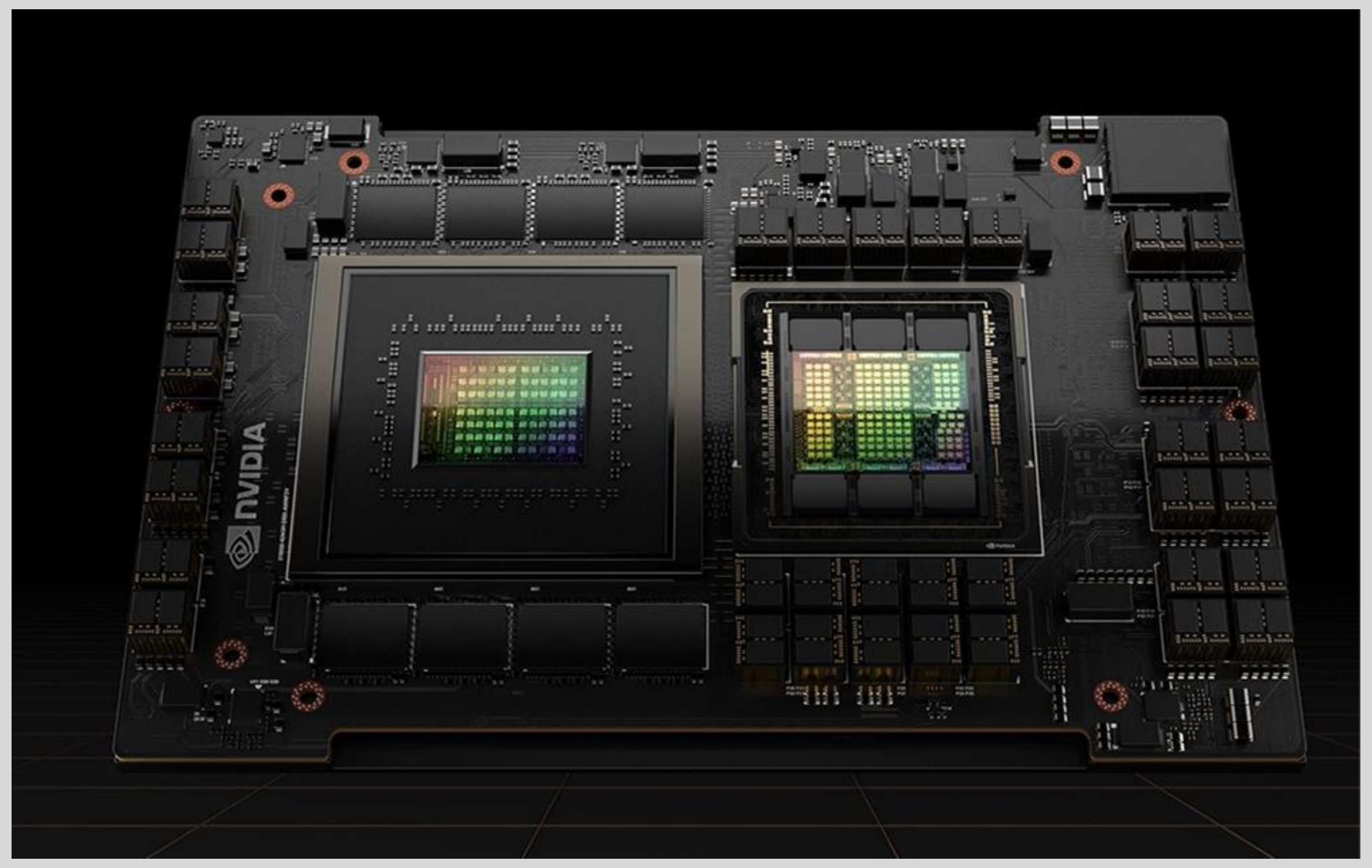
Test was repeated many time to ensure that the model runs smoothly.


Set up ML-Perf benchmark
Parameter Float16
Parameter Float32
Benchmark ModelSize A (27K)
Benchmark ModelSize B (55k)
Benchmark ModelSize C (110k)
Analysis on collected data

Power and Training process for Benchmark and Analysis

Results





Nvidia Grace Hopper superchip (H200)

Conclusions

- Correlation observed between model size and power consumption on GH2.
- Optimizing model architectures and deployment strategies is vital for desired performance with minimal power usage.
- Memory power consumption for Foalt16 is signifyingly lower than Float32.

Future Plans

- Improve energy efficiency in LLM computations on platforms like GH2.
- Explore and test alternative models beyond LLM.

References