SCALEWAY NEW H100

PCIe GPU Instance

October, 6 2023

CLOUD MERCATO
CLOUD TRANSPARENCY PLATFORM

ABOUT SCALEWAY

Founded in 1999, Scaleway, helps developers and businesses to build, deploy and scale applications to any infrastructure.

Located in Paris, Amsterdam and Warsaw, Scaleway's complete cloud ecosystem is used by 25,000+ businesses, including European startups, who choose Scaleway for its multi-AZ redundancy, smooth developer experience, data centers that run only on renewable energy, and native tools for managing multi-cloud architectures.

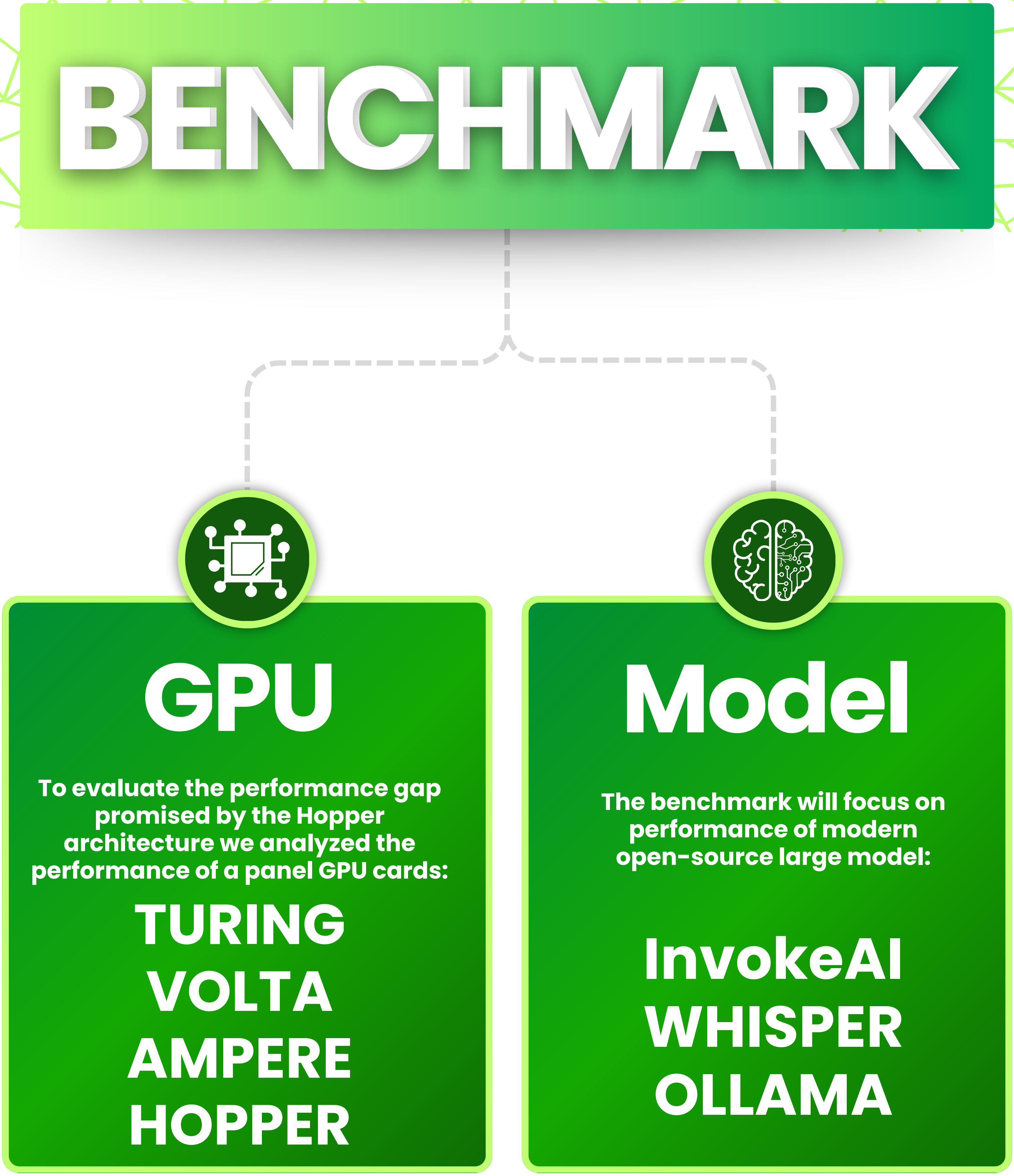
With fully managed offerings for bare metal, containerization and serverless architectures, Scaleway brings choice to the world of cloud computing, offering customers the ability to choose where their customer's data resides, to choose what architecture works best for their business, and to choose a more responsible way to scale.

WHAT IS A H100 PCIE GPU INSTANCE?

HOPPER THE LATEST NVIDIA ARCHITECTURE

Large performance improvements:

- Faster Core/Memory throughput
- Larger L1/L2 cache
- Better power efficiency
- FP8/FP16 acceleration



BENCHMARK

GPU

To evaluate the performance gap promised by the Hopper architecture we analyzed the performance of a panel GPU cards:

TURING
VOLTA
AMPERE
HOPPER

Model

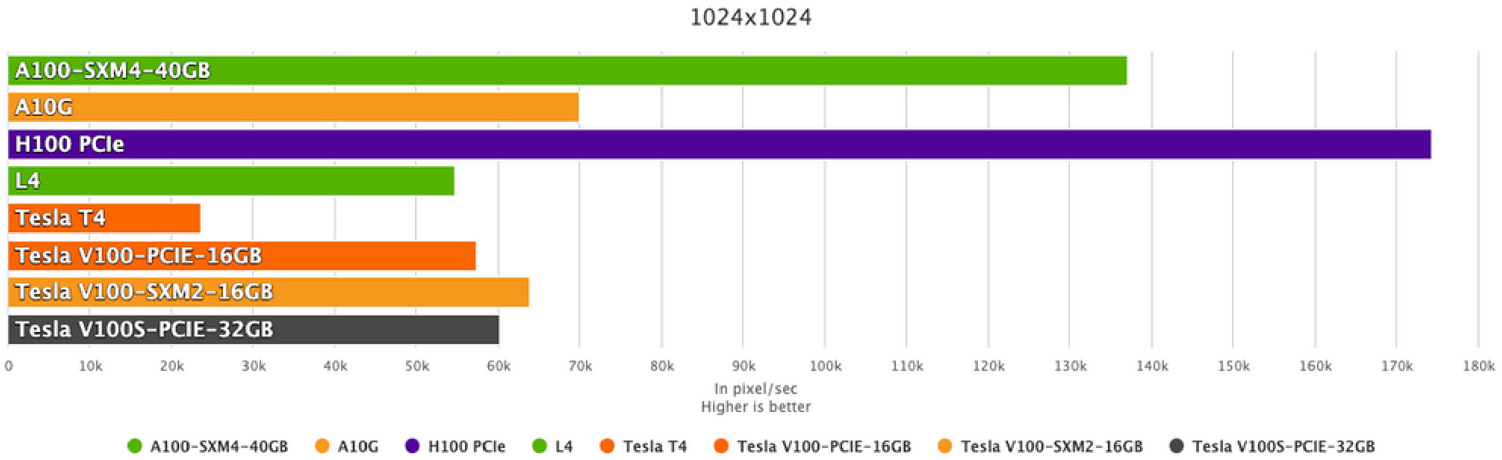
The benchmark will focus on performance of modern open-source large model:

InvokeAI
WHISPER
OLLAMA

InvokeAI

InvokeAI is an image handling and generation framework using Stable Diffusion models.

How fast an image is created?



Observations

The Scaleway H100 PCIe GPU Instance is always among the best in terms of performance.

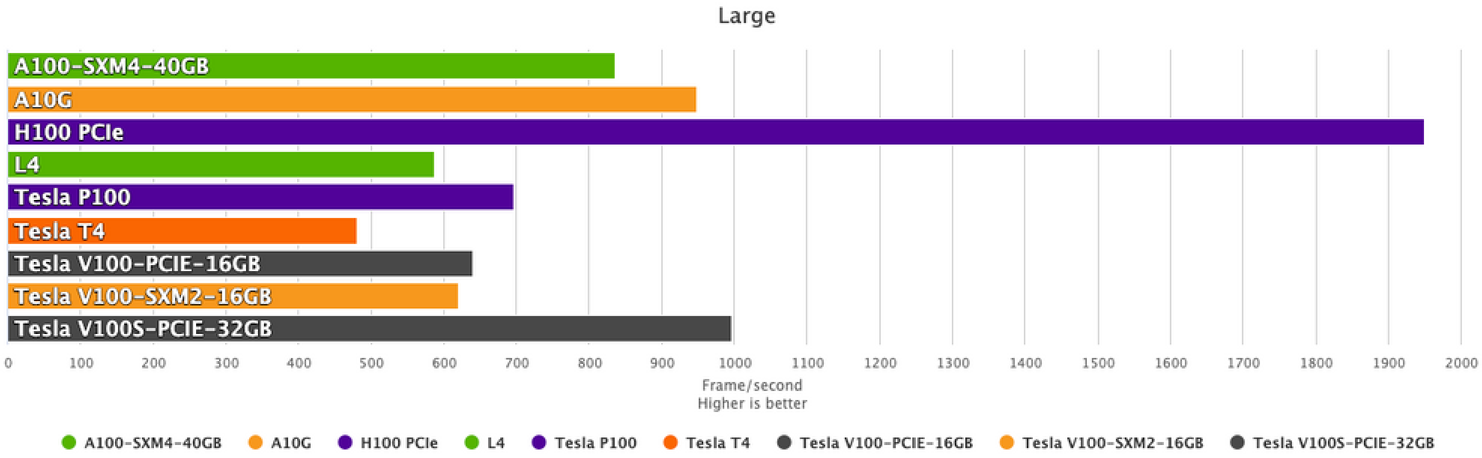
Only a few GPUs can produce large scale image in a decent timing.

Similarly, less GPUs can produce image smaller under 128x128.

WHISPER

Whisper is a speech recognition model allowing speech-to-text from +50 languages.

How fast speech is evaluated to text ?



Observations

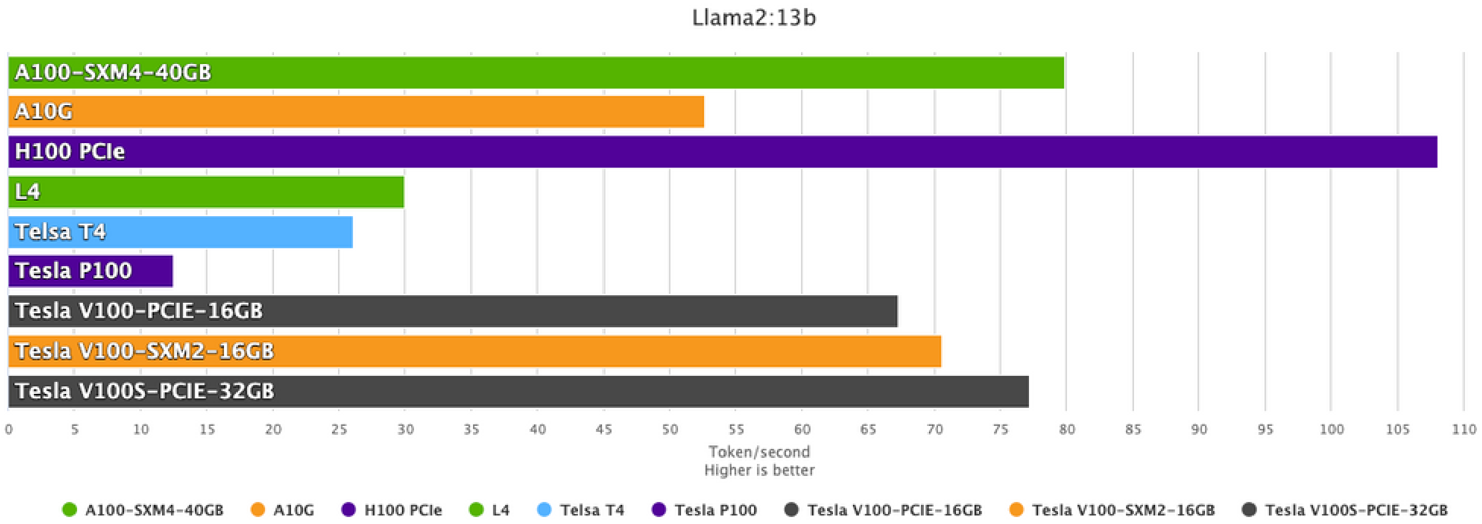
The H100 PCIe GPU Instance always leads the performance ranking.

The Base model on H100 PCIe GPU Instance is faster than the tiny on any other GPU.

OLLAMA

Ollama is an open-source software allowing the usage LLM models such as llama2.

How fast an answer is generated ?



Observations

As the large model require 32GB of VRAM, the panel is very restricted in this category.

H100 PCIe GPU Instance has a performance gap of 30 token/sec.

CONCLUSION

The latest Hopper architecture desmonstrate a large performance gap for large AI models.
Chatbot, Image generation, Speech-to-text, all the modern AI workloads obtain a boost from the PCIe GPU Instance.

For more informations check the full study on Cloud Mercato.