NEBIUS

NVIDIA Preferred Partner

NVIDIA® GPUs: Performance overview

H100, A100, V100, L4 and L40

Author’s foreword

Thank you for downloading this overview.
I wrote it on a basis of research papers, NVIDIA’s documentation, public benchmarks and opinions of leading ML industry experts. We will overview NVIDIA’s current line of GPUs, offering insights into the new deep learning capabilities of H100.

Find me on LinkedIn for any additional questions on this paper.

Igor Ofitserov,
Senior Technical Project Manager at Nebius AI

Role of Transformers in new GPUs

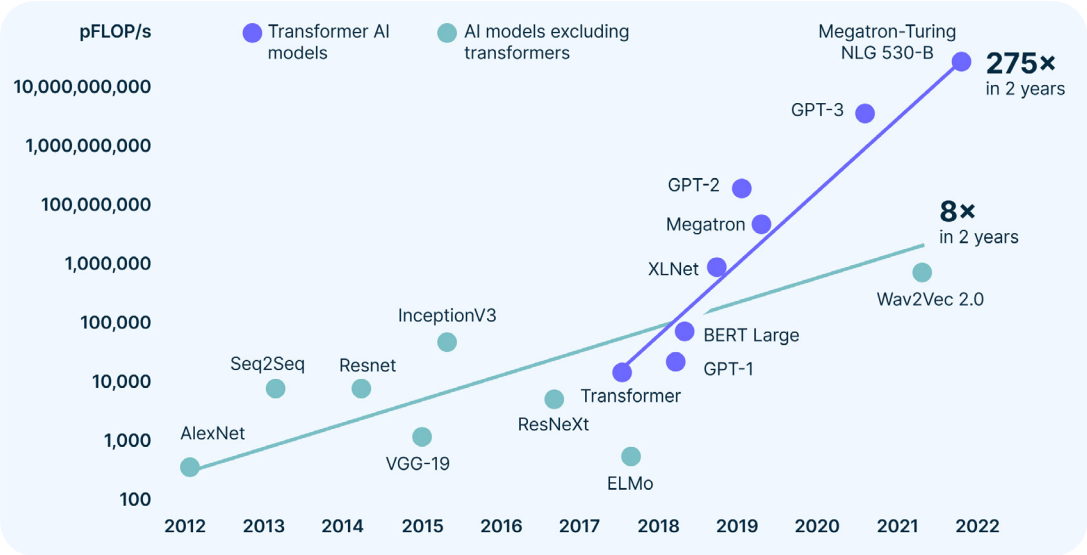
Before we start comparing GPUs, I would like to highlight the trend that Transformers from the LLM domain are becoming the primary tool for most AI tasks across all domains.

Supporting this trend, GPU manufacturers have also optimized their products for Transformers. That's why the new generation of GPUs gives boost in efficiency specifically for Transformers.

Let's see why.

If you're new in the field, I would recommend to watch Andrej Karpathy's seminar "Introduction to Transformers" on Stanford Online YouTube channel

So, why Transformer models are trending?



We have to understand the trend of Transformers to realize the role of FP8 support in the new NVIDIA GPU line (H100, L4, L40)

Transformers demonstrate superior quality compared to other models with pre-training on a large of unlabelled datasets, followed by fine-tuning on a small set of labelled high-quality or specific data. While this training method existed before, Transformers have proven to be more effective.

Transformers parallelize significantly better than other models, as they don't have as many loops and sequential operations. They also utilize Tensor cores better because most algorithms inside are reduced to matrix multiplication.

The possibilities of scaling and improving the quality of models by increasing datasets have not yet been exhausted. In addition, generation of synthetic data is now actively researched.

The role of numbers' precision

FP64

The **double-precision** binary floating-point format is used for scientific computations / HPC with rather strong precision requirements. Typically, **not used in deep learning**.

FP32

The **single-precision** binary floating-point format was the workhorse of deep learning for a long time. Weights, activations and other values in neural networks have long been represented in FP32 by default.

BF16

The original IEEE FP16 was not designed with deep learning applications in mind, its dynamic range is too narrow. BFLOAT16 solves this, providing dynamic range identical to that of FP32.

BFLOAT16 format now is a trend.

INT8

Usually used for speed up inference via post-training quantization (PTQ). But often it is **inapplicable to Transformers** because it introduces unacceptable accuracy loss. INT8 Quantization aware training (QAT) could be used for Transformer models but it requires additional costs of modifying and re-running the training process.

FP8

In case of inference, the FP8 type delivers the performance of INT8 with the accuracy of FP16 under post-training quantization (PTQ) **for all models, including Transformers**.

Also, It can be used to speed up training after some code changes for mixed-precision support.

GPU performance comparison

H100 vs A100 for Transformer training

H100 is a newer model that builds upon the foundations laid by the A100's Ampere architecture, offering further enhancements in AI processing capabilities and overall performance.

NVIDIA's research has shown that FP8 precision can be used to accelerate specific operations (matrix multiplication and convolutions), without adversely affecting the model quality.

Not coincidentally, the transformer architecture — the core of LLMs and other generative AI models — uses matrix multiplication (matmul) operations extensively.

Parameter	H100 SXM	A100 SXM	Difference
RAM, GB	80	80	0
Bandwidth, TB/s	3.35	2	1.6×
TF32, TFLOP/s	494	156	3.2×
FP16, TFLOP/s	989	312	3.2×
BF16, TFLOP/s	989	312	3.2×
FP8, TFLOP/s	1,979	–	6.3× (vs BF16)

Benchmarks: Training MosaicML GPT Transformer model with H100

Model	Precision	Throughput, tok/sec	TFLOP/s	Speedup over NVIDIA A100 BF16
1B	BF16	43,352	394	2.2×
1B	FP8	53,721	489	2.7×
3B	BF16	22,987	412	2.2×
3B	FP8	29,313	525	2.8×

Source: Mosaic ML

H100 vs A100 for inference

The MLPerf Inference: Datacenter benchmark suite measures how fast systems can process inputs and produce results using a trained model.

MLPerf is an industry standard ML benchmark suite that measures full system performance in real applications

Benchmarks: MLPerf™ Inference

- System Type: Datacenter
- Division/Power: Closed
- Round: v2.1, 1, Single GPU
- Precision: FP32

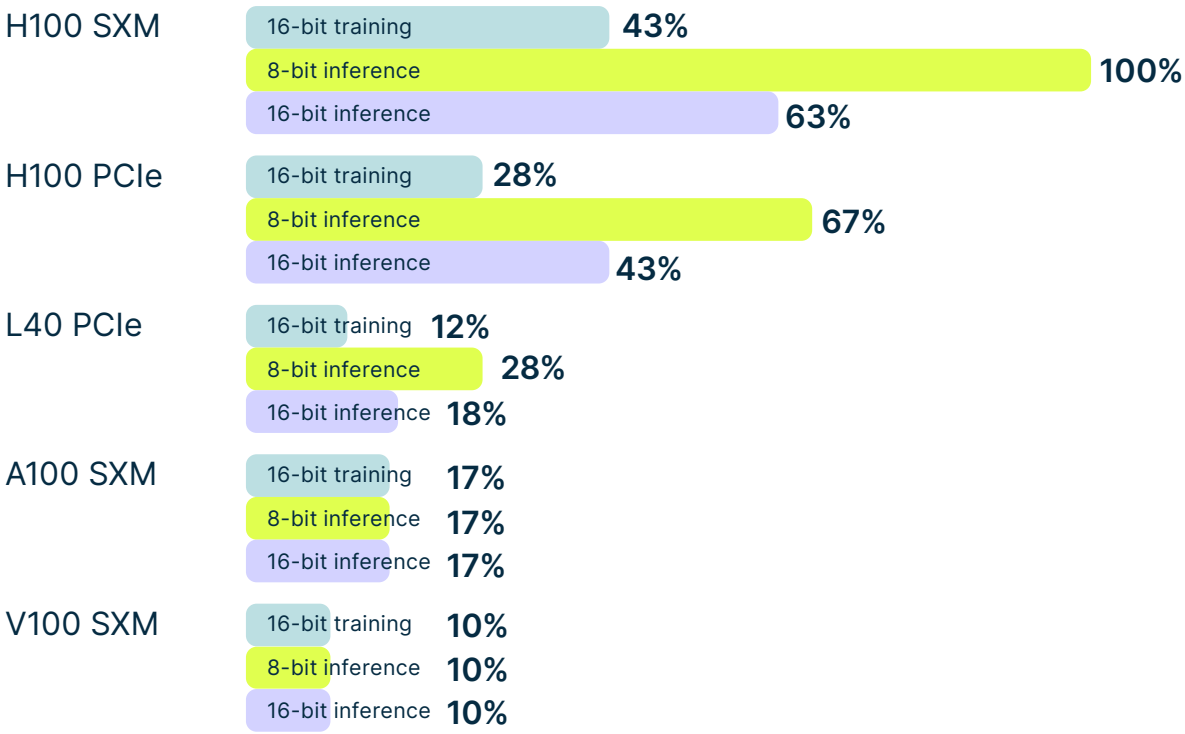
Model	H100 SXM, Inferences/sec	A100 SXM, Inferences/sec	Speedup
DLRM, Recommender	695,298	314,992	2.2×
BERT, Natural Language Processing	8,144	1,757	4.6×
ResNet-50 v1.5, Image Classification	81,292	38,011	2.1×
RNN-T, Speech Recognition	22,885	14,007	1.6×

Source: ML Commons

Relative Transformer performance of GPUs

Below you can see the raw relative Transformer performance of GPUs. For example, an A100 SXM has about 0.4x performance of an H100 SMX for 16-bit training. In other words, an H100 SXM is 2.5 times faster for 16-bit training compared to an A100 SXM.

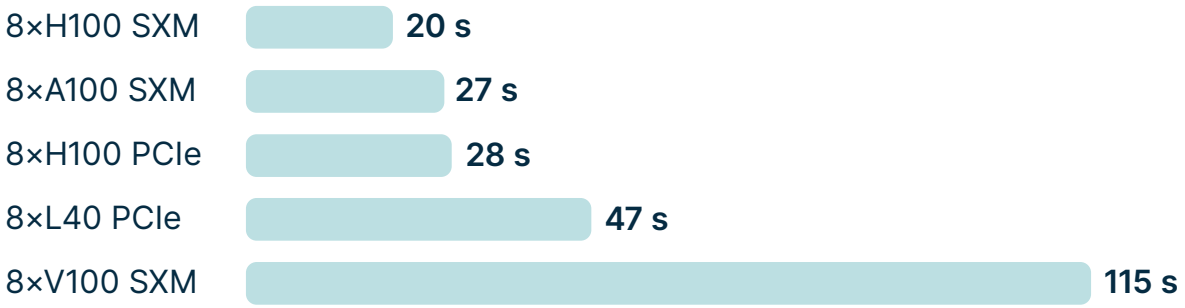
We can also see that there is a gigantic gap in 8-bit performance of H100 GPUs and old cards that are optimized for 16-bit performance.



Relative non-transformer performance of GPUs

Benchmarks: MLPerf™ Resnet training speed

ResNet is a common non-transformer model. This benchmark demonstrates the performance of GPUs in training this type of model.



Sources: Tim Dettmers, ML commons

H100 is 2 to 4 times faster than A100, but A100 can be more cost effective

H100 is always a better choice compared to A100.

The greatest gain of H100 is achieved in the speed of training models with Transformer architecture. The effect becomes more dramatic on mixed-precision FP8 training. 16-bit training of CNN, RNN models can be up to 5% more economical but slower by at least 42%.

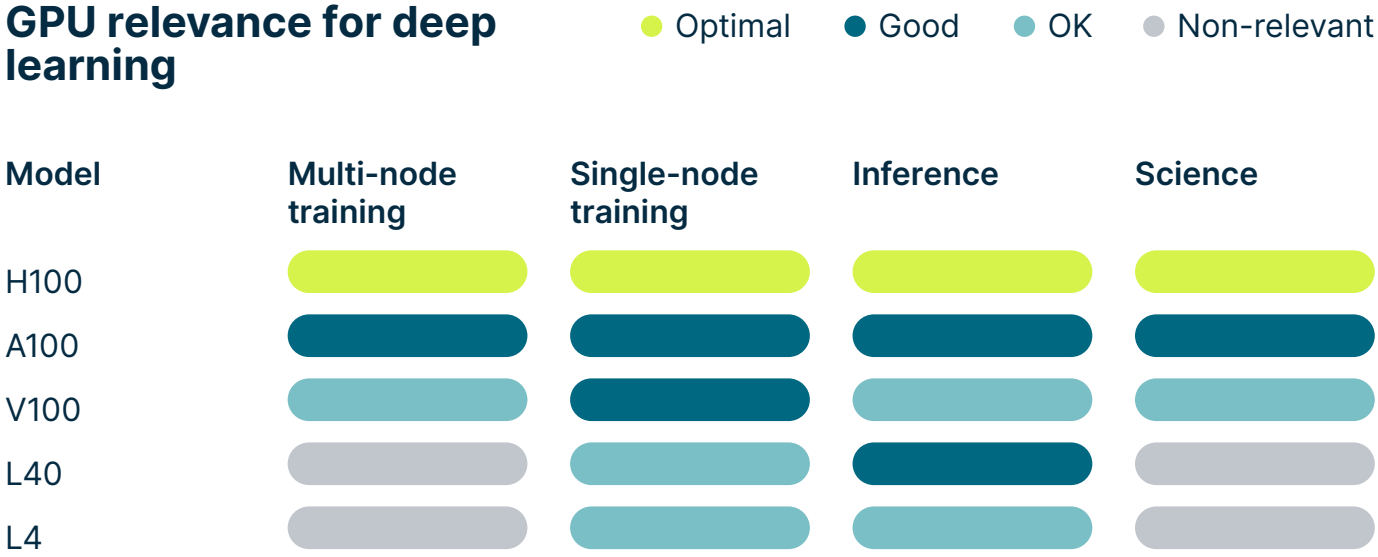
For inference of high memory consumption models with moderate load A100 is more cost effective. H100 should be preferred if speed of processing requests is a priority.

Choosing GPU for deep learning

Most important specifications side by side

	AI training and inference			AI inference			
Parameters	**H100 SXM**	**A100 SXM**	**V100 SXM**	**V100 PCIe**	**H100 PCIe**	**L40 PCIe**	**L4 PCIe**
RAM, GB	80	80	32	32	80	48	24
Memory bandwidth, GB/s	3,350	2,000	900	090	2,000	864	300
Interconnect bandwidth, GB/s	900	600	300	32	128	64	64
FP64, TFLOP/s	67	19	7.8	7	51	–	–
TF32, TFLOP/s	494	156	125	112	378	90	60
BF16, TFLOP/s	989	312	–	–	757	181	121
FP8, TFLOP/s	1,979	–	–	–	1,513	362	242

GPU relevance for deep learning



All of the optimal GPUs are available at Nebius AI, with H100 supporting Infiniband cluster.

Explore Nebius AI

Summary

L4

Entry ticket to the world of GPU computing.
Mostly used for inference models with small workloads.
Choose it if you don't know where to start.

L40

Next level of performance for inference of not-so-large models with intensive workloads. Good choice for generative AI inference.

V100

At the end of its life, but still good for inference. In SXM form factor, it can be used for single-node training, but only with FP32 precision: BF16 is not supported. Comparable to L40 in terms of performance.

A100

Cost effective for single-node training of conventional models and inference of big models with moderate workloads.

H100

Best choice if speed is your top priority. Perfect for bigNLP, LLM, and all models with Transformer architecture

Got questions?

This concludes our overview of different parameters to keep in mind. We hope it helps you in setting up your infrastructure. But since every case is unique, you're welcome to contact us.

We will be glad to discuss your needs and see how we can help.

Thanks for your attention!