AI Inference

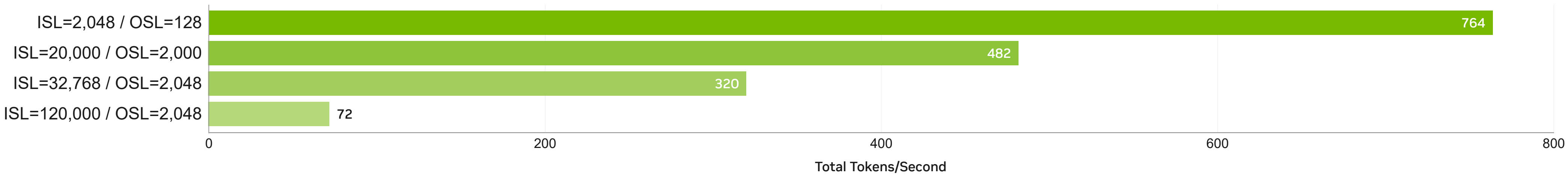
Inference can be deployed in many ways, depending on the use-case. Offline processing of data is best done at larger batch sizes, which can deliver optimal GPU utilization and throughput. However, increasing throughput also tends to increase latency. Generative AI and Large Language Models (LLMs) deployments seek to deliver great experiences by lowering latency. So developers and infrastructure managers need to strike a balance between throughput and latency to deliver great user experiences and best possible throughput while containing deployment costs.

When deploying LLMs at scale, a typical way to balance these concerns is to set a time-to-first token limit, and optimize throughput within that limit. The data presented in the Large Language Model Low Latency section show best throughput at a time limit of one second, which enables great throughput at low latency for most users, all while optimizing compute resource use.

Click here to view other performance data.

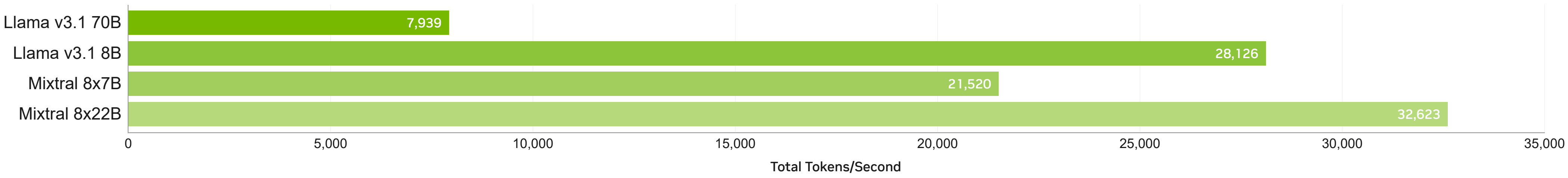
LLM Inference Performance of NVIDIA Data Center Products

Llama v3.1 405B - H200 TRT-LLM High Throughput



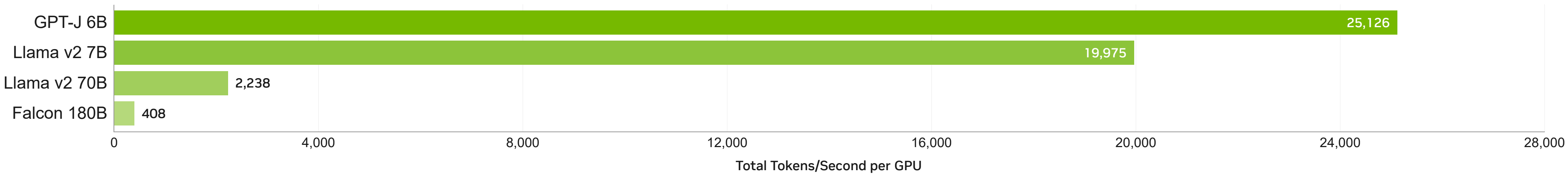
DGX H200 w/ NVIDIA H200 | TensorRT-LLM v0.14a and v0.15.0 | Precision: FP8 | Input Length: 2,048, 20,000, 32,768 and 120,000 | Output Length: 128, 2,000 and 2,048 | Tensor Parallelism: 8 for 20,000/2,000, 32,768/2,048 and 120,000/2,048 and 1 for 2,048/128 ISL/OSL

H200 TRT-LLM High Throughput



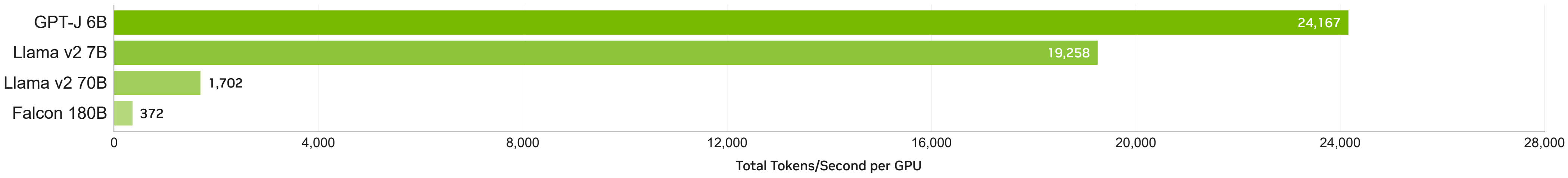
DGX H200 w/ NVIDIA H200 | TensorRT-LLM v0.14.0: Mixtral 8x22B and TensorRT-LLM v0.13.0: Llama v3.1 70B, Llama v3.1 8B, Mixtral 8x7B | Precision: FP8 | Input Length: 128 | Output Length: 2048 | Tensor Parallelism: Llama v3.1 70B and Mixtral 8x7B = 2, Llama v3.1 8B = 1, Mixtral 8x22B = 8

H200 TRT-LLM High Throughput Under 1 Second 1st Token Latency



DGX H200 w/ NVIDIA H200 | TensorRT-LLM 0.9.0 | Precision: FP8 | Batch Size: GPT-J 6B = 512, Llama v2 7B = 512, Llama v2 70B and Falcon 180B = 64 | Input Length = 128 | Output Length: Llama v2 70B and Falcon 180B = 2,048, all other models are 128 | Tensor Parallelism: GPT-J 6B, Llama v2 7B and Llama v2 70B = 1, Falcon 180B = 8 | Highest measured throughput with less than 1 second 1st token latency

H100 TRT-LLM High Throughput Under 1 Second 1st Token Latency



DGX H100 w/ H100-SXM5-80GB | TensorRT-LLM 0.9.0 | Precision: FP8 | Batch Size: GPT-J 6B and Llama v2 7B = 512, Llama v2 70B and Falcon 180B = 64 | Input Length = 128 | Output Length = 128 | Tensor Parallelism: GPT-J 6B, Llama v2 7B and Llama v2 70B = 1, Falcon 180B = 4 | Highest measured throughput with less than 1 second 1st token latency

H200 Inference Performance - High Throughput

Model	PP	TP	Input Length	Output Length	Throughput	GPU	Server	Precision	Framework	GPU Version
Llama v3.1 405B	1	8	128	128	3,953 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 405B	1	8	128	2048	5,974 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200

Model	PP	TP	Input Length	Output Length	Throughput	GPU	Server	Precision	Framework	GPU Version
Llama v3.1 405B	1	8	128	4096	4,947 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 405B	8	1	2048	128	764 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.14a	NVIDIA H200
Llama v3.1 405B	1	8	5000	500	679 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 405B	1	8	500	2000	5,066 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 405B	1	8	1000	1000	3,481 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 405B	1	8	2048	2048	2,927 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 405B	1	8	20000	2000	482 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.14.0	NVIDIA H200
Llama v3.1 70B	1	1	128	128	3,924 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.13.0	NVIDIA H200
Llama v3.1 70B	1	2	128	2048	7,939 total tokens/sec	2x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 70B	1	2	128	4096	6,297 total tokens/sec	2x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 70B	1	1	2048	128	460 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.13.0	NVIDIA H200
Llama v3.1 70B	1	1	5000	500	560 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 70B	1	2	500	2000	6,683 total tokens/sec	2x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 70B	1	1	1000	1000	2,704 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 70B	1	2	2048	2048	3,835 total tokens/sec	2x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 70B	1	2	20000	2000	633 total tokens/sec	2x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 8B	1	1	128	128	28,126 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.13.0	NVIDIA H200
Llama v3.1 8B	1	1	128	2048	24,158 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 8B	1	1	128	4096	16,460 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 8B	1	1	2048	128	3,661 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 8B	1	1	5000	500	3,836 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 8B	1	1	500	2000	20,345 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 8B	1	1	1000	1000	16,801 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Llama v3.1 8B	1	1	2048	2048	11,073 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.13.0	NVIDIA H200
Llama v3.1 8B	1	1	20000	2000	1,741 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x7B	1	1	128	128	16,796 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x7B	1	1	128	2048	14,830 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x7B	1	2	128	4096	21,520 total tokens/sec	2x H200	DGX H200	FP8	TensorRT-LLM 0.14.0	NVIDIA H200
Mixtral 8x7B	1	1	2048	128	1,995 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x7B	1	1	5000	500	2,295 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x7B	1	1	500	2000	11,983 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x7B	1	1	1000	1000	10,254 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x7B	1	2	2048	2048	14,018 total tokens/sec	2x H200	DGX H200	FP8	TensorRT-LLM 0.13.0	NVIDIA H200
Mixtral 8x7B	1	2	20000	2000	2,227 total tokens/sec	2x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x22B	1	8	128	128	25,179 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.14.0	NVIDIA H200
Mixtral 8x22B	1	8	128	2048	32,623 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x22B	1	8	128	4096	25,531 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x22B	1	8	2048	128	3,095 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x22B	1	8	5000	500	4,209 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x22B	1	8	500	2000	27,396 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x22B	1	8	1000	1000	20,097 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.15.0	NVIDIA H200
Mixtral 8x22B	1	8	2048	2048	13,796 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.14.0	NVIDIA H200
Mixtral 8x22B	1	8	20000	2000	2,897 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.14.0	NVIDIA H200

TP: Tensor Parallelism
PP: Pipeline Parallelism
For more information on pipeline parallelism, please read Llama v3.1 405B Blog
Output tokens/second on Llama v3.1 405B is inclusive of time to generate the first token (tokens/s = total generated tokens / total latency)

H100 Inference Performance - High Throughput

Model	PP	TP	Input Length	Output Length	Throughput	GPU	Server	Precision	Framework	GPU Version
Llama v3.1 70B	1	2	128	128	6,399 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Llama v3.1 70B	1	2	128	4096	3,581 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Llama v3.1 70B	1	2	2048	128	774 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Llama v3.1 70B	1	2	500	2000	4,776 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Llama v3.1 70B	1	2	1000	1000	4,247 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Llama v3.1 70B	1	4	2048	2048	5,166 total tokens/sec	4x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Llama v3.1 70B	1	4	20000	2000	915 total tokens/sec	4x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Mixtral 8x7B	1	2	128	128	27,156 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Mixtral 8x7B	1	2	128	2048	23,010 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Mixtral 8x7B	1	8	128	4096	47,834 total tokens/sec	8x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Mixtral 8x7B	1	2	2048	128	3,368 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Mixtral 8x7B	1	2	5000	500	3,592 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Mixtral 8x7B	1	2	500	2000	18,186 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.14.0	H100-SXM5-80GB
Mixtral 8x7B	1	2	1000	1000	15,932 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.14.0	H100-SXM5-80GB
Mixtral 8x7B	1	2	2048	2048	10,465 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB
Mixtral 8x7B	1	2	20000	2000	1,739 total tokens/sec	2x H100	DGX H100	FP8	TensorRT-LLM 0.15.0	H100-SXM5-80GB

TP: Tensor Parallelism
PP: Pipeline Parallelism

L40S Inference Performance - High Throughput

Model	PP	TP	Input Length	Output Length	Throughput	GPU	Server	Precision	Framework	GPU Version
Llama v3.1 8B	1	1	128	128	8,983 total tokens/sec	1x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Llama v3.1 8B	1	1	128	2048	5,297 total tokens/sec	1x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Llama v3.1 8B	1	1	128	4096	2,989 total tokens/sec	1x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Llama v3.1 8B	1	1	2048	128	1,056 total tokens/sec	1x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Llama v3.1 8B	1	1	5000	500	972 total tokens/sec	1x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Llama v3.1 8B	1	1	500	2000	4,264 total tokens/sec	1x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Llama v3.1 8B	1	1	1000	1000	4,014 total tokens/sec	1x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Llama v3.1 8B	1	1	2048	2048	2,163 total tokens/sec	1x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Llama v3.1 8B	1	1	20000	2000	326 total tokens/sec	1x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Mixtral 8x7B	4	1	128	128	15,278 total tokens/sec	4x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Mixtral 8x7B	2	2	128	2048	9,087 total tokens/sec	4x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Mixtral 8x7B	1	4	128	4096	5,655 total tokens/sec	4x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Mixtral 8x7B	4	1	2048	128	2,098 total tokens/sec	4x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Mixtral 8x7B	2	2	5000	500	1,558 total tokens/sec	4x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Mixtral 8x7B	2	2	500	2000	7,974 total tokens/sec	4x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Mixtral 8x7B	2	2	1000	1000	6,579 total tokens/sec	4x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S
Mixtral 8x7B	2	2	2048	2048	4,217 total tokens/sec	4x L40S	Supermicro SYS-521GE-TNRT	FP8	TensorRT-LLM 0.15.0	NVIDIA L40S

TP: Tensor Parallelism
PP: Pipeline Parallelism

H200 Inference Performance - High Throughput at Low Latency Under 1 Second

Model	Batch Size	TP	Input Length	Output Length	Time to 1st Token	Throughput/GPU	GPU	Server	Precision	Framework	GPU Version
GPT-J 6B	512	1	128	128	0.64 seconds	25,126 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
GPT-J 6B	64	1	128	2048	0.08 seconds	7,719 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
GPT-J 6B	32	1	2048	128	0.68 seconds	2,469 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200

Model	Batch Size	TP	Input Length	Output Length	Time to 1st Token	Throughput/GPU	GPU	Server	Precision	Framework	GPU Version
GPT-J 6B	32	1	2048	2048	0.68 seconds	3,167 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Llama v2 7B	512	1	128	128	0.84 seconds	19,975 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Llama v2 7B	64	1	128	2048	0.11 seconds	7,149 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Llama v2 7B	32	1	2048	128	0.9 seconds	2,101 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Llama v2 7B	32	1	2048	2048	0.9 seconds	3,008 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Llama v2 70B	64	1	128	128	0.92 seconds	2,044 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Llama v2 70B	64	1	128	2048	0.93 seconds	2,238 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Llama v2 70B	4	1	2048	128	0.95 seconds	128 total tokens/sec	1x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Llama v2 70B	16	8	2048	2048	0.97 seconds	173 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Falcon 180B	32	4	128	128	0.36 seconds	365 total tokens/sec	4x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Falcon 180B	64	8	128	2048	0.43 seconds	408 total tokens/sec	8x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Falcon 180B	4	4	2048	128	0.71 seconds	43 total tokens/sec	4x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200
Falcon 180B	4	4	2048	2048	0.71 seconds	53 total tokens/sec	4x H200	DGX H200	FP8	TensorRT-LLM 0.9.0	NVIDIA H200

TP: Tensor Parallelism
Batch size per GPU
Low Latency Target: Highest measured throughput with less than 1 second 1st token latency

H100 Inference Performance - High Throughput at Low Latency Under 1 Second

Model	Batch Size	TP	Input Length	Output Length	Time to 1st Token	Throughput/GPU	GPU	Server	Precision	Framework	GPU Version
GPT-J 6B	512	1	128	128	0.63 seconds	24,167 total tokens/sec	1x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
GPT-J 6B	120	1	128	2048	0.16 seconds	7,351 total tokens/sec	1x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
GPT-J 6B	32	1	2048	128	0.67 seconds	2,257 total tokens/sec	1x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
GPT-J 6B	32	1	2048	2048	0.68 seconds	2,710 total tokens/sec	1x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Llama v2 7B	512	1	128	128	0.83 seconds	19,258 total tokens/sec	1x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Llama v2 7B	120	1	128	2048	0.2 seconds	6,944 total tokens/sec	1x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Llama v2 7B	32	1	2048	128	0.89 seconds	1,904 total tokens/sec	1x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Llama v2 7B	32	1	2048	2048	0.89 seconds	2,484 total tokens/sec	1x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Llama v2 70B	64	1	128	128	0.92 seconds	1,702 total tokens/sec	1x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Llama v2 70B	128	4	128	2048	0.73 seconds	1,494 total tokens/sec	4x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Llama v2 70B	4	8	2048	128	0.74 seconds	105 total tokens/sec	8x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Llama v2 70B	8	4	2048	2048	0.74 seconds	141 total tokens/sec	4x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Falcon 180B	64	4	128	128	0.71 seconds	372 total tokens/sec	4x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Falcon 180B	64	4	128	2048	0.7 seconds	351 total tokens/sec	4x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Falcon 180B	8	8	2048	128	0.87 seconds	45 total tokens/sec	8x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB
Falcon 180B	8	8	2048	2048	0.87 seconds	61 total tokens/sec	8x H100	DGX H100	FP8	TensorRT-LLM 0.9.0	H100-SXM5-80GB

TP: Tensor Parallelism
Batch size per GPU
Low Latency Target: Highest measured throughput with less than 1 second 1st token latency

View More Performance Data

Training to Convergence

Deploying AI in real-world applications requires training networks to convergence at a specified accuracy. This is the best methodology to test whether AI systems are ready to be deployed in the field to deliver meaningful results.

Learn More

AI Pipeline

NVIDIA Riva is an application framework for multimodal conversational AI services that deliver real-performance on GPUs.
Learn More