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.

MLPerf Inference
Large Language Model
Inference
Triton Inference Server
Cloud Inference

MLPerf Inference v4.1 Performance Benchmarks

Offline Scenario, Closed Division

Network	Throughput	GPU	Server	GPU Version	Target Accuracy	Dataset
Llama2 70B	11,264 tokens/sec	1x B200	NVIDIA B200	NVIDIA B200-SXM-180GB	rouge1=44.4312, rouge2=22.0352, rougeL=28.6162	OpenOrca
	34,864 tokens/sec	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB-CTS	rouge1=44.4312, rouge2=22.0352, rougeL=28.6162	OpenOrca
	24,525 tokens/sec	8x H100	NVIDIA DGX H100	NVIDIA H100-SXM-80GB	rouge1=44.4312, rouge2=22.0352, rougeL=28.6162	OpenOrca
	4,068 tokens/sec	1x GH200	NVIDIA GH200 NVL2 Platform	NVIDIA GH200 Grace Hopper Superchip 144GB	rouge1=44.4312, rouge2=22.0352, rougeL=28.6162	OpenOrca
Mixtral 8x7B	59,335 tokens/sec	8x H200	GIGABYTE G593-SD1	NVIDIA H200-SXM-141GB	rouge1=45.4911, rouge2=23.2829, rougeL=30.3615, (gsm8k)Accuracy=73.78, (mbxp)Accuracy=60.12)	OpenOrca, GSM8K, MBXP
	52,818 tokens/sec	8x H100	SMC H100	NVIDIA H100-SXM-80GB	rouge1=45.4911, rouge2=23.2829, rougeL=30.3615, (gsm8k)Accuracy=73.78, (mbxp)Accuracy=60.12)	OpenOrca, GSM8K, MBXP
	8,021 tokens/sec	1x GH200	NVIDIA GH200 NVL2 Platform	NVIDIA GH200 Grace Hopper Superchip 144GB	rouge1=45.4911, rouge2=23.2829, rougeL=30.3615, (gsm8k)Accuracy=73.78, (mbxp)Accuracy=60.12)	OpenOrca, GSM8K, MBXP
Stable Diffusion XL	18 samples/sec	8x H200	Dell PowerEdge XE9680	NVIDIA H200-SXM-141GB	FID range: [23.01085758, 23.95007626] and CLIP range: [31.68631873, 31.81331801]	Subset of coco-2014 val
	16 samples/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	FID range: [23.01085758, 23.95007626] and CLIP range: [31.68631873, 31.81331801]	Subset of coco-2014 val
	2.3 samples/sec	1x GH200	NVIDIA GH200 NVL2 Platform	NVIDIA GH200 Grace Hopper Superchip 144GB	FID range: [23.01085758, 23.95007626] and CLIP range: [31.68631873, 31.81331801]	Subset of coco-2014 val
ResNet-50	768,235 samples/sec	8x H200	Dell PowerEdge XE9680	NVIDIA H200-SXM-141GB	76.46% Top1	ImageNet (224x224)
	710,521 samples/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	76.46% Top1	ImageNet (224x224)

Network	Throughput	GPU	Server	GPU Version	Target Accuracy	Dataset
	95,105 samples/sec	1x GH200	NVIDIA GH200-GraceHopper-Superchip	NVIDIA GH200 Grace Hopper Superchip 96GB	76.46% Top1	ImageNet (224x224)
RetinaNet	15,015 samples/sec	8x H200	ThinkSystem SR685a V3	NVIDIA H200-SXM-141GB	0.3755 mAP	OpenImages (800x800)
	14,538 samples/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	0.3755 mAP	OpenImages (800x800)
	1,923 samples/sec	1x GH200	NVIDIA GH200-GraceHopper-Superchip	NVIDIA GH200 Grace Hopper Superchip 96GB	0.3755 mAP	OpenImages (800x800)
BERT	73,791 samples/sec	8x H200	Dell PowerEdge XE9680	NVIDIA H200-SXM-141GB	90.87% f1	SQuAD v1.1
	72,876 samples/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	90.87% f1	SQuAD v1.1
	9,864 samples/sec	1x GH200	NVIDIA GH200-GraceHopper-Superchip	NVIDIA GH200 Grace Hopper Superchip 96GB	90.87% f1	SQuAD v1.1
GPT-J	20,552 tokens/sec	8x H200	ThinkSystem SR680a V3	NVIDIA H200-SXM-141GB	rouge1=42.9865, rouge2=20.1235, rougeL=29.9881	CNN Dailymail
	19,878 tokens/sec	8x H100	ESC-N8-E11	NVIDIA H100-SXM-80GB	rouge1=42.9865, rouge2=20.1235, rougeL=29.9881	CNN Dailymail
	2,804 tokens/sec	1x GH200	GH200-GraceHopper-Superchip_GH200-96GB_aarch64x1_TRT	NVIDIA GH200 Grace Hopper Superchip 96GB	rouge1=42.9865, rouge2=20.1235, rougeL=29.9881	CNN Dailymail
DLRMv2	639,512 samples/sec	8x H200	GIGABYTE G593-SD1	NVIDIA H200-SXM-141GB	80.31% AUC	Synthetic Multihot Criteo Dataset
	602,108 samples/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	80.31% AUC	Synthetic Multihot Criteo Dataset
	86,731 samples/sec	1x GH200	NVIDIA GH200 NVL2 Platform	NVIDIA GH200 Grace Hopper Superchip 144GB	80.31% AUC	Synthetic Multihot Criteo Dataset
3D-UNET	55 samples/sec	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	0.863 DICE mean	KiTS 2019
	52 samples/sec	8x H100	AS-4125GS-TNHR2-LCC	NVIDIA H100-SXM-80GB	0.863 DICE mean	KiTS 2019
	7 samples/sec	1x GH200	GH200-GraceHopper-Superchip_GH200-96GB_aarch64x1_TRT	NVIDIA GH200 Grace Hopper Superchip 96GB	0.863 DICE mean	KiTS 2019

Server Scenario - Closed Division

Network	Throughput	GPU	Server	GPU Version	Target Accuracy	MLPerf Server Latency Constraints (ms)	Dataset
Llama2 70B	10,756 tokens/sec	1x B200	NVIDIA B200	NVIDIA B200-SXM-180GB	rouge1=44.4312, rouge2=22.0352, rougeL=28.6162	TTFT/TPOT: 2000 ms/200 ms	OpenOrca
	32,790 tokens/sec	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB-CTS	rouge1=44.4312, rouge2=22.0352, rougeL=28.6162	TTFT/TPOT: 2000 ms/200 ms	OpenOrca
	23,700 tokens/sec	8x H100	AS-4125GS-TNHR2-LCC	NVIDIA H100-SXM-80GB	rouge1=44.4312, rouge2=22.0352, rougeL=28.6162	TTFT/TPOT: 2000 ms/200 ms	OpenOrca
	3,884 tokens/sec	1x GH200	NVIDIA GH200 NVL2 Platform	NVIDIA GH200 Grace Hopper Superchip 144GB	rouge1=44.4312, rouge2=22.0352, rougeL=28.6162	TTFT/TPOT: 2000 ms/200 ms	OpenOrca

Network	Throughput	GPU	Server	GPU Version	Target Accuracy	MLPerf Server Latency Constraints (ms)	Dataset
Mixtral 8x7B	57,177 tokens/sec	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	rouge1=45.4911, rouge2=23.2829, rougeL=30.3615, (gsm8k)Accuracy=73.78, (mbxp)Accuracy=60.12)	TTFT/TPOT: 2000 ms/200 ms	OpenOrca, GSM8K, MBXP
	51,028 tokens/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	rouge1=45.4911, rouge2=23.2829, rougeL=30.3615, (gsm8k)Accuracy=73.78, (mbxp)Accuracy=60.12)	TTFT/TPOT: 2000 ms/200 ms	OpenOrca, GSM8K, MBXP
	7,450 tokens/sec	1x GH200	NVIDIA GH200 NVL2 Platform	NVIDIA GH200 Grace Hopper Superchip 144GB	rouge1=45.4911, rouge2=23.2829, rougeL=30.3615, (gsm8k)Accuracy=73.78, (mbxp)Accuracy=60.12)	TTFT/TPOT: 2000 ms/200 ms	OpenOrca, GSM8K, MBXP
Stable Diffusion XL	17 samples/sec	8x H200	ThinkSystem SR680a V3	NVIDIA H200-SXM-141GB	FID range: [23.01085758, 23.95007626] and CLIP range: [31.68631873, 31.81331801]	20 s	Subset of coco-2014 val
	16 samples/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	FID range: [23.01085758, 23.95007626] and CLIP range: [31.68631873, 31.81331801]	20 s	Subset of coco-2014 val
	2.02 samples/sec	1x GH200	NVIDIA GH200 NVL2 Platform	NVIDIA GH200 Grace Hopper Superchip 144GB	FID range: [23.01085758, 23.95007626] and CLIP range: [31.68631873, 31.81331801]	20 s	Subset of coco-2014 val
ResNet-50	681,328 queries/sec	8x H200	GIGABYTE G593-SD1	NVIDIA H200-SXM-141GB	76.46% Top1	15 ms	ImageNet (224x224)
	634,193 queries/sec	8x H100	SYS-821GE-TNHR	NVIDIA H100-SXM-80GB	76.46% Top1	15 ms	ImageNet (224x224)
	77,012 queries/sec	1x GH200	NVIDIA GH200-GraceHopper-Superchip	NVIDIA GH200 Grace Hopper Superchip 96GB	76.46% Top1	15 ms	ImageNet (224x224)
RetinaNet	14,012 queries/sec	8x H200	GIGABYTE G593-SD1	NVIDIA H200-SXM-141GB	0.3755 mAP	100 ms	OpenImages (800x800)
	13,979 queries/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	0.3755 mAP	100 ms	OpenImages (800x800)
	1,731 queries/sec	1x GH200	GH200-GraceHopper-Superchip_GH200-96GB_aarch64x1_TRT	NVIDIA GH200 Grace Hopper Superchip 96GB	0.3755 mAP	100 ms	OpenImages (800x800)
BERT	58,091 queries/sec	8x H200	Dell PowerEdge XE9680	NVIDIA H200-SXM-141GB	90.87% f1	130 ms	SQuAD v1.1
	58,929 queries/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	90.87% f1	130 ms	SQuAD v1.1
	7,103 queries/sec	1x GH200	GH200-GraceHopper-Superchip_GH200-96GB_aarch64x1_TRT	NVIDIA GH200 Grace Hopper Superchip 96GB	90.87% f1	130 ms	SQuAD v1.1
GPT-J	20,139 queries/sec	8x H200	Dell PowerEdge XE9680	NVIDIA H200-SXM-141GB	rouge1=42.9865, rouge2=20.1235, rougeL=29.9881	20 s	CNN Dailymail
	19,811 queries/sec	8x H100	AS-4125GS-TNHR2-LCC	NVIDIA H100-SXM-80GB	rouge1=42.9865, rouge2=20.1235, rougeL=29.9881	20 s	CNN Dailymail
	2,513 queries/sec	1x GH200	NVIDIA GH200 NVL2 Platform	NVIDIA GH200 Grace Hopper Superchip 144GB	rouge1=42.9865, rouge2=20.1235, rougeL=29.9881	20 s	CNN Dailymail
DLRMv2	585,209 queries/sec	8x H200	GIGABYTE G593-SD1	NVIDIA H200-SXM-141GB	80.31% AUC	60 ms	Synthetic Multihot Criteo Dataset
	556,101 queries/sec	8x H100	SYS-421GE-TNHR2-LCC	NVIDIA H100-SXM-80GB	80.31% AUC	60 ms	Synthetic Multihot Criteo Dataset

Network	Throughput	GPU	Server	GPU Version	Target Accuracy	MLPerf Server Latency Constraints (ms)	Dataset
	81,010 queries/sec	1x GH200	NVIDIA GH200 NVL2 Platform	NVIDIA GH200 Grace Hopper Superchip 144GB	80.31% AUC	60 ms	Synthetic Multihot Criteo Dataset

Power Efficiency Offline Scenario - Closed Division

Network	Throughput	Throughput per Watt	GPU	Server	GPU Version	Dataset
Llama2 70B	25,262 tokens/sec	4 tokens/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	OpenOrca
Mixtral 8x7B	48,988 tokens/sec	8 tokens/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	OpenOrca, GSM8K, MBXP
Stable Diffusion XL	13 samples/sec	0.002 samples/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	Subset of coco-2014 val
ResNet-50	556,234 samples/sec	112 samples/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	ImageNet (224x224)
RetinaNet	10,803 samples/sec	2 samples/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	OpenImages (800x800)
BERT	54,063 samples/sec	10 samples/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	SQuAD v1.1
GPT-J	13,097 samples/sec	3. samples/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	CNN Dailymail
DLRMv2	503,719 samples/sec	84 samples/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	Synthetic Multihot Criteo Dataset
3D-UNET	42 samples/sec	0.009 samples/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	KiTS 2019

Power Efficiency Server Scenario - Closed Division

Network	Throughput	Throughput per Watt	GPU	Server	GPU Version	Dataset
Llama2 70B	23,113 tokens/sec	4 tokens/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	OpenOrca
Mixtral 8x7B	45,497 tokens/sec	7 tokens/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	OpenOrca, GSM8K, MBXP
Stable Diffusion	13 queries/sec	0.002 queries/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	Subset of coco-2014 val
ResNet-50	480,131 queries/sec	96 queries/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	ImageNet (224x224)
RetinaNet	9,603 queries/sec	2 queries/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	OpenImages (800x800)
BERT	41,599 queries/sec	8 queries/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	SQuAD v1.1
GPT-J	11,701 queries/sec	2 queries/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	CNN Dailymail
DLRMv2	420,107 queries/sec	69 queries/sec/watt	8x H200	NVIDIA H200	NVIDIA H200-SXM-141GB	Synthetic Multihot Criteo Dataset

MLPerf™ v4.1 Inference Closed: Llama2 70B 99.9% of FP32, Mixtral 8x7B 99% of FP32 and 99.9% of FP32, Stable Diffusion XL, ResNet-50 v1.5, RetinaNet, RNN-T, BERT 99% of FP32 accuracy target, 3D U-Net 99.9% of FP32 accuracy target, GPT-J 99.9% of FP32 accuracy target, DLRM 99% of FP32 accuracy target: 4.1-0005, 4.1-0021, 4.1-0027, 4.1-0037, 4.1-0038, 4.1-0043, 4.1-0044, 4.1-0046, 4.1-0048, 4.1-0049, 4.1-0053, 4.1-0057, 4.1-0060, 4.1-0063, 4.1-0064, 4.1-0065, 4.1-0074. MLPerf name and logo are trademarks. See https://mlcommons.org/ for more information.
NVIDIA B200 is a preview submission
Llama2 70B Max Sequence Length = 1,024
Mixtral 8x7B Max Sequence Length = 2,048
BERT-Large Max Sequence Length = 384.
For MLPerf™ various scenario data, click here
For MLPerf™ latency constraints, click here

View More Performance Data

AI Pipeline

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