GPU Comparison Matrix

Key Specifications for Deep Learning Workloads (2024-2025)

GPU Model	Memory	Compute	Bandwidth	Cost	Best Use
A100 80GB **2020** Ampere	**80 GB** HBM2e	**312 TFLOPS** FP16 156 TF (TF32)	**2,039 GB/s**	**$15-20K**	Large model training Production
H100 **2022** Hopper	**80 GB** HBM3	**989 TFLOPS** FP16 3.2× A100	**3,350 GB/s**	**$30-40K**	Cutting-edge training Max performance
H200 **2024** Hopper+	**141 GB** HBM3e	**989 TFLOPS** FP16 Same as H100	**4,800 GB/s**	**$35-45K**	Very large models Memory-intensive
B200 **2025** Blackwell	**192 GB** HBM3e	**2,500 TFLOPS** FP16 2.5× H100	**8,000 GB/s**	**$50-70K**	Next-gen frontier models Massive scale
B300 **2025** Blackwell+	**288 GB** HBM3e	**3,000 TFLOPS** FP16 3× H100	**10,000 GB/s**	**$70-100K**	Ultra-large models Extreme scale
L4 **2023** Ada Lovelace	**24 GB** GDDR6	**242 TFLOPS** FP16	**300 GB/s**	**$3-5K**	Inference Fine-tuning Cost-optimized

Selection Guidelines by Generation

Ampere (2020-2022):

A100: Mature, widely available, proven for production. Best value for established workloads.

Hopper (2022-2024):

H100: 3× faster than A100, current high-end standard. H200: 76% more memory, 43% more bandwidth.

Blackwell (2025+):

B200: 2.5× H100 performance, 192GB memory. B300: 3× H100, 288GB for trillion-parameter models.

Blackwell enables training models 2-3× larger than Hopper generation at same cost.

Cost-Optimized:

L4: Best for inference and fine-tuning. 1/10th cost of A100, sufficient for most serving workloads.