GENERATIVE AI · TAUGHT WITH VLLM

LLM Inference and Serving

Prefill and decode, the KV cache, continuous batching, quantization, and speculative decoding. You size a GPU for a model and a context length, then defend the number under load.

Advanced | 4 phases | 23 sessions | 35 contact hours

Syllabus

4 phases, 23 sessions. Each phase ends in something you have built.

1 The Machine Underneath

Predict what a model will cost in memory and latency before you rent the GPU.

5 sessions
$108 alone

1.1 One GPU, One Model

vLLM 0.26 serving an OpenAI-compatible endpoint, and the four startup log lines that matter.

FREE PREVIEW

1.2 Prefill and Decode

One request, two workloads: prefill saturates compute, decode starves on memory bandwidth.

1.3 Sizing the KV Cache

Bytes per token per layer, and why 128K of context outweighs the weights you paid for.

1.4 GQA, MLA, Sliding Windows

How latent and grouped-query attention shrink the cache, and what batch size that buys you.

1.5 Where the Time Goes

HBM bandwidth divided by bytes read per token, and why decode leaves the GPU mostly idle.

By the end: **A sizing sheet for one model at 128K context whose predictions land within ten percent of a measured run.**

2 Throughput Without Lying

Get honest throughput out of a single node and report it in a way that survives review.

6 sessions
$108 alone

2.1	**Inside PagedAttention**	Blocks, fragmentation, and the gpu-memory-utilization knob that decides your real concurrency.
2.2	**Continuous Batching**	Requests joining and leaving mid-flight, and reading the waiting queue before you blame the model.
2.3	**Chunked Prefill**	max-num-batched-tokens as the single dial between time to first token and inter-token latency.
2.4	**Prefix Caching, Measured**	Block hashing, hit rate on a shared system prompt, and the workloads where it buys nothing.
2.5	**Load Testing Honestly**	Poisson arrivals, warmup, fixed input and output lengths, and why mean TTFT hides the failure.
2.6	**Latency, Throughput, Goodput**	TTFT, TPOT and ITL, then picking the config that holds p99 instead of the best average.

By the end: **A benchmark report comparing three engine configurations at a fixed p99 time to first token.**

3 Cheaper, Not Worse

Cut cost per million tokens without quietly damaging the model.

6 sessions
$108 alone

3.1 FP8 as the Default
E4M3 weights and KV cache on Hopper via llm-compressor, at close to no measurable accuracy cost.

3.2 FP4 on Blackwell
NVFP4 and MXFP4 block scaling on B200, and the roughly one point of MMLU it can cost you.

3.3 AWQ, GPTQ, and What Died
Where 4-bit integer formats still earn a place, and which quantizers are no longer worth shipping.

3.4 Proving Quantization Safe
One eval before and after, including long context, where a benchmark average hides the damage.

3.5 Speculative Decoding
EAGLE-3, MTP and n-gram drafting, with acceptance rate as the number that decides the win.

3.6 Cost Per Million Tokens
GPU hourly rate over measured output tokens per second, comparing H200 and B200 without hand-waving.

By the end: **A quantized deployment with speculative decoding, an accuracy check, and a measured cost per million output tokens.**

4 Running a Fleet

Serve across many GPUs and many nodes against a latency target you can defend.

6 sessions
$108 alone

4.1	**Tensor and Pipeline Parallel**	Splitting a model that does not fit, and letting NVLink versus PCIe decide which split you use.
4.2	**Expert Parallelism for MoE**	Active versus total parameters, expert parallel with data-parallel attention, and expert imbalance.
4.3	**Disaggregating Prefill and Decode**	Separate pools with NIXL moving KV, and when the transfer costs more than the recompute saved.
4.4	**KV Offload and Tiers**	CPU DRAM and object storage behind HBM via LMCache, and the hit rate that justifies the hop.
4.5	**Routing and Autoscaling**	Prefix-aware routing via the Gateway API Inference Extension, KEDA on queue depth, and cold starts.
4.6	**Observability and Postmortem**	Metrics that predict a stall, and a written postmortem of one deliberately overloaded run.

By the end: **A multi-replica Kubernetes deployment with prefix-aware routing, queue-based autoscaling, and an SLO dashboard.**

Tools you will use

vLLM 0.26 · SGLang 0.5.16 · NVIDIA TensorRT-LLM · llm-compressor / compressed-tensors · NVIDIA TensorRT Model Optimizer (NVFP4) · GuideLLM · LMCache + NIXL 1.3 · NVIDIA Dynamo · Kubernetes Gateway API Inference Extension

What you will build

1. **The Sizing Sheet**
   Memory and throughput predicted, then measured within ten percent
2. **Half the Cost, Same Answers**
   FP8 or NVFP4 plus speculative decoding, with the eval to prove it
3. **The Fleet**
   Multi-node serving, prefix-aware routing, autoscaling against a p99 SLO

Registration

Phase 1: The Machine Underneath	$108
Phase 2: Throughput Without Lying	$108
Phase 3: Cheaper, Not Worse	$108
Phase 4: Running a Fleet	$108
Whole track, all 4 phases	**$360**

Buying every phase separately costs $432, so the whole track saves $72.

No payment is taken online. Registering creates an order and payment instructions are sent with it. Access opens once payment is confirmed.