TT TEXTBOOK TESTBANKS

Test Bank for AI Systems Performance Engineering 1st Fregly

TEST BANK SAMPLE PREVIEW

PUBLISHER
OReilly

COPYRIGHT
2025

RESOURCE TYPE
Test Bank

ISBN
9798341627789

Test Bank for AI Systems Performance Engineering 1st Edition by Fregly

ISBN: 9798341627789

AI Systems Performance Engineering

ISBN 9798341627789 | Chapter 1: Introduction and AI System Overview

Total Questions: 150

Multiple Choice

Q1.
Which company was described as the small startup in China that trained a frontier large language model under export-related hardware constraints?

A) Anthropic
B) DeepSeek.AI ✓ Correct
C) Meta AI
D) Mistral AI

Rationale: The chapter identifies DeepSeek.AI as the Chinese startup that trained a frontier model despite hardware restrictions.

Q2.
Which export-compliant NVIDIA GPU was specifically mentioned as being used by DeepSeek because top-tier GPUs were unavailable?

A) A100
B) B200
C) H800 ✓ Correct
D) L40S

Rationale: The text states that DeepSeek used locally available export-compliant alternatives including the NVIDIA H800 GPU.

Q3.
What was the name of DeepSeek's reasoning model that achieved performance near leading frontier models?

A) DeepSeek-V2
B) DeepSeek-V3
C) DeepSeek-R1 ✓ Correct
D) DeepSeek-XL

Rationale: DeepSeek-R1 is identified as the reasoning model with near-frontier performance.

Q4.
According to the chapter, what resource did DeepSeek's engineers explicitly treat as scarce and optimize aggressively?

A) Disk capacity
B) Communication bandwidth ✓ Correct
C) CPU clock speed
D) Power supply redundancy

Rationale: The chapter emphasizes that DeepSeek treated communication bandwidth as a scarce resource and optimized every byte over the wire.

Q5.
Which optimization technique was specifically cited as one of the methods DeepSeek used to improve performance on constrained GPUs?

A) Federated averaging
B) Model distillation ✓ Correct
C) Rule-based pruning
D) Symbolic execution

Rationale: Model distillation is explicitly named as one of the advanced optimization techniques used by DeepSeek.

Q6.
Approximately how much did the chapter report as the estimated training cost of OpenAI's GPT-4?

A) About $6 million
B) About $25 million
C) About $100 million ✓ Correct
D) About $191 million

Rationale: The chapter reports GPT-4's training cost as approximately $100 million.

Q7.
Approximately how much did the chapter estimate for training Google's Gemini Ultra?

A) About $60 million
B) About $100 million
C) About $150 million
D) About $191 million ✓ Correct

Rationale: Gemini Ultra is cited as having an estimated training cost of about $191 million.

Q8.
DeepSeek claimed that DeepSeek-R1 was trained for less than what amount in compute cost?

A) $1 million
B) $6 million ✓ Correct
C) $25 million
D) $50 million

Rationale: The chapter states that DeepSeek claimed DeepSeek-R1 was trained for less than $6 million in compute.

Q9.
Which event was reported to have briefly shocked US financial markets after DeepSeek's efficiency announcement?

A) AMD stock rose 17% in a day
B) NVIDIA stock dropped about 17% in a day ✓ Correct
C) The NASDAQ halted trading in AI stocks
D) TSMC announced a production freeze

Rationale: The chapter notes that NVIDIA's stock dropped roughly 17% in a single day after the news.

Q10.
The chapter defines AI systems performance engineering as being about more than speed; it is also about making what both possible and affordable?

A) Research and regulation
B) Automation and compliance
C) The previously impossible ✓ Correct
D) Open-source licensing

Rationale: The text states that AI systems performance engineering is about making the previously impossible both possible and affordable.

Q11.
Which description best matches the role of the AI systems performance engineer as presented in the chapter?

A) A specialist focused only on model architecture design
B) A specialist focused on optimizing AI models and the systems they run on ✓ Correct
C) A manager responsible primarily for AI product marketing
D) A compliance officer for AI governance documentation

Rationale: The chapter defines the role as optimizing both AI models and the underlying systems they run on.

Q12.
Which combination of outcomes is explicitly listed as a goal for AI training and inference pipelines?

A) Fast, cost-efficient, performant, reliable, and highly available ✓ Correct
B) Interpretable, regulated, multilingual, and autonomous
C) Portable, decentralized, and blockchain-enabled
D) Low-code, no-code, and vendor-neutral

Rationale: The chapter states that performance engineers ensure pipelines are fast, cost-efficient, performant, reliable, and highly available.

Q13.
Which trio does the chapter emphasize must be codesigned together in AI systems performance work?

A) Security, privacy, and ethics
B) Training, validation, and testing
C) Hardware, software, and algorithms ✓ Correct
D) Finance, operations, and marketing

Rationale: The text repeatedly stresses codesigning hardware, software, and algorithms.

Q14.
Which of the following is listed as a key responsibility of an AI systems performance engineer?

A) Drafting reimbursement policy
B) Benchmarking and profiling ✓ Correct
C) Managing hospital credentialing
D) Conducting qualitative focus groups

Rationale: Benchmarking and profiling are explicitly listed among the engineer's core responsibilities.

Q15.

Benchmarking and profiling in the chapter primarily involve measuring which kinds of metrics?

A) Patient satisfaction and staffing ratios

B) Latency, throughput, and memory usage ✓ Correct

C) Market share and brand recognition

D) Compiler syntax and code style

Rationale: The chapter defines benchmarking and profiling as measuring latency, throughput, memory usage, and related performance metrics.

Q16.

Which tool set was specifically recommended to iteratively identify bottlenecks at different levels of the stack?

A) Wireshark, Grafana, and Tableau

B) NVIDIA Nsight Systems, NVIDIA Nsight Compute, and the PyTorch profiler ✓ Correct

C) TensorBoard, Jenkins, and Git

D) Prometheus, Kibana, and Airflow

Rationale: The chapter explicitly recommends using Nsight Systems, Nsight Compute, and the PyTorch profiler together.

Q17.

Why does the chapter recommend automated performance tests early in the development cycle?

A) To reduce licensing fees

B) To catch performance regressions early ✓ Correct

C) To replace all manual debugging

D) To eliminate the need for benchmarks

Rationale: Automated performance tests are recommended to detect regressions or reductions in performance early.

Q18.

In the chapter, debugging and optimizing performance issues requires tracing problems to what?

A) The most recent software vendor

B) Their root cause ✓ Correct

C) The largest memory device

D) A single benchmark score

Rationale: The chapter emphasizes tracing performance issues to their root cause.

Q19.

Which NVIDIA hardware feature was mentioned as optimized for modern LLMs using the transformer architecture?

A) BlueField DPU

B) Transformer Engine ✓ Correct

C) Omniverse RTX

D) Jetson Orin

Rationale: The chapter notes that the latest NVIDIA Transformer Engine is hardware optimized for modern LLMs.

Q20.
What software change does the chapter suggest when a Python data preprocessing step is bottlenecking a training pipeline?

A) Move the code to JavaScript
B) Replace the model with a smaller dense network
C) Reimplement the code in C++ or use NVIDIA cuPyNumeric ✓ Correct
D) Disable profiling to reduce overhead

Rationale: The text specifically suggests reimplementing the bottlenecking Python code in C++ or using NVIDIA cuPyNumeric.

Q21.
What is NVIDIA cuPyNumeric described as in the chapter?

A) A GPU driver update tool
B) A drop-in NumPy replacement that can distribute array operations across CPUs and GPUs ✓ Correct
C) A filesystem benchmark utility
D) A container orchestration engine

Rationale: The chapter describes cuPyNumeric as a drop-in NumPy replacement that distributes array workloads across CPUs and GPUs.

Q22.
When scaling from small research workloads to ultrascale production, what is the stated goal?

A) To reduce the number of nodes to one
B) To avoid using distributed communication libraries
C) To scale with minimal overhead and loss of efficiency ✓ Correct
D) To keep models dense rather than sparse

Rationale: The chapter states that scaling should occur with minimal overhead and minimal loss of efficiency.

Q23.
Which library is identified as the NVIDIA library used for distributed collectives such as all-reduce?

A) cuDNN
B) NCCL ✓ Correct
C) TensorRT
D) UCX-Py

Rationale: NCCL is explicitly named as the library for distributed collectives like all-reduce.

Q24.
How is NCCL pronounced according to the chapter?

A) Nickel ✓ Correct
B) Nackle
C) N-cell
D) N-cycle

Rationale: The chapter explicitly notes that NCCL is pronounced 'nickel'.

Q25.
Which library is described as providing high-throughput, low-latency point-to-point data movement across GPU memory and storage tiers for distributed inference?

A) NIXL ✓ Correct
B) NVTX
C) NVDLA
D) NVML

Rationale: The chapter identifies the NVIDIA Inference Xfer Library, or NIXL, in that role.

Q26.
Which parallelism strategy is particularly relevant when a model is so large that it does not fit on a single GPU?

A) Only batch normalization parallelism
B) Tensor parallelism or pipeline parallelism ✓ Correct
C) Only label parallelism
D) Only storage parallelism

Rationale: The chapter states that tensor parallelism or pipeline parallelism may be needed when the model cannot fit on one GPU.

Q27.
For mixture-of-experts models, which additional form of parallelism is mentioned as advantageous?

A) Spectral parallelism
B) Expert parallelism ✓ Correct
C) Lexical parallelism
D) Checkpoint parallelism

Rationale: The chapter specifically notes that MoE models can take advantage of expert parallelism.

Q28.
Which resource-management technique is given as an example of partitioning GPU resources for better utilization when full GPU power is unnecessary?

A) CUDA graphs
B) MIG ✓ Correct
C) FSDP
D) FlashAttention

Rationale: The chapter cites NVIDIA Multi-Instance GPU, or MIG, as a GPU virtualization technique for partitioning resources.

Q29.
Cross-team collaboration for AI systems performance engineers is described as especially important with which groups?

A) Only legal and procurement teams
B) Only product marketing and sales teams
C) Researchers, data scientists, application developers, and infrastructure teams ✓ Correct
D) Only executive leadership and finance

Rationale: The chapter emphasizes collaboration with researchers, data scientists, developers, and infrastructure teams including networking and storage.

Q30.
What principle does the chapter emphasize regarding evidence in performance engineering?

A) Trust intuition over measurement
B) Measure everything and trust data, not assumptions ✓ Correct
C) Rely on vendor claims when benchmarks are unavailable
D) Prioritize anecdotal speedups over reproducibility

Rationale: The chapter explicitly states that performance engineering should measure everything and trust data rather than assumptions.

Q31.
During DeepSeek's Open-Source Week in February 2025, which project was identified as their optimized attention kernel written in CUDA C++?

A) DeepEP
B) DualPipe
C) FlashMLA ✓ Correct
D) 3FS

Rationale: FlashMLA is described as DeepSeek's optimized attention kernel written in CUDA C++.

Q32.
Which DeepSeek open-source project provides an FP8-optimized matrix multiplication library?

A) DeepGEMM ✓ Correct
B) Fire-Flyer File System
C) EPLB
D) DeepCache

Rationale: DeepGEMM is identified as the FP8-optimized matrix multiplication library.

Q33.
What is DeepEP according to the chapter?

A) A distributed filesystem
B) A highly tuned communication library for MoE models ✓ Correct
C) A GPU virtualization layer
D) A CPU scheduling framework

Rationale: The chapter defines Deep Experts Parallelism, or DeepEP, as a tuned communication library for MoE models.

Q34.
Which DeepSeek project implements a redundant expert strategy that duplicates heavily loaded experts?

A) DualPipe
B) FlashMLA
C) EPLB ✓ Correct
D) 3FS

Rationale: EPLB is described as implementing a redundant expert strategy for heavily loaded experts.

Q35.
What is the main function of DualPipe as described in the chapter?

A) Encrypting model checkpoints
B) Overlapping forward/backward computation and communication to reduce pipeline bubbles ✓ Correct
C) Compressing token vocabularies
D) Virtualizing CPU memory for GPU kernels

Rationale: DualPipe is described as a bidirectional pipeline parallelism algorithm that overlaps computation and communication.

Q36.
Which DeepSeek project is the high-performance distributed filesystem mentioned in the chapter?

A) DeepGEMM
B) 3FS ✓ Correct
C) NIXL
D) TensorRT-LLM

Rationale: 3FS, or Fire-Flyer File System, is identified as DeepSeek's high-performance distributed filesystem.

Q37.
Which benchmark suite is presented as a standard for reproducibly comparing training and inference performance across hardware and software setups?

A) SPECint
B) TPC-C
C) MLPerf ✓ Correct
D) LINPACK

Rationale: The chapter identifies MLPerf as the open benchmark suite for reproducible training and inference comparisons.

Q38.
When citing MLPerf results, what caution does the chapter give about per-GPU numbers?

A) They are the only valid cross-platform metric
B) They should replace end-to-end system metrics
C) They are not the primary metric across platforms and should be treated as component-level indicators ✓ Correct
D) They are useful only for storage benchmarking

Rationale: The chapter cautions that per-GPU results are not the primary cross-platform metric and should be viewed as component-level indicators.

Q39.
Compared with the H100, what two capabilities are reduced in the export-compliant H800 according to the chapter?

A) HBM capacity and tensor core count
B) NVLink interconnect bandwidth and FP64 performance ✓ Correct
C) PCIe support and CPU compatibility
D) CUDA support and driver stability

Rationale: The H800 is described as reducing NVLink bandwidth and FP64 performance relative to the H100.

Q40.
Approximately how much NVLink interconnect bandwidth per GPU does the chapter attribute to the H100?

A) 200 GB/s
B) 400 GB/s
C) 900 GB/s ✓ Correct
D) 1.8 TB/s

Rationale: The chapter states that the H100 provides about 900 GB/s of NVLink bandwidth per GPU.

Q41.
Approximately how much NVLink interconnect bandwidth per GPU does the chapter attribute to the H800?

A) 300 GB/s
B) 400 GB/s ✓ Correct
C) 700 GB/s
D) 900 GB/s

Rationale: The chapter states that the H800 provides about 400 GB/s of NVLink bandwidth per GPU.

Q42.
DeepSeek-V3 is described as what type of model architecture?

A) A recurrent neural network
B) A dense convolutional model
C) A mixture-of-experts language model ✓ Correct
D) A graph neural network

Rationale: The chapter explicitly identifies DeepSeek-V3 as a massive MoE language model.

Q43.
Although DeepSeek-V3 contains about 680 billion total parameters, approximately how many active parameters are used per input token?

A) 9 billion
B) 37 billion ✓ Correct
C) 256 billion
D) 680 billion

Rationale: The chapter states that DeepSeek-V3 uses about 37 billion active parameters per input token.

Q44.
For each token in DeepSeek-V3, how many experts are active according to the chapter's routing description?

A) 1
B) 8
C) 9 ✓ Correct
D) 256

Rationale: The routing scheme uses 1 shared expert plus 8 selected experts, totaling 9 active experts per token.

Q45.
What training strategy is highlighted for DeepSeek-R1's reasoning development?

A) Heavy dependence on human feedback loops
B) A cold-start approach using minimal supervised data and reinforcement learning ✓ Correct
C) Exclusive supervised fine-tuning on labeled chain-of-thought data
D) Purely unsupervised pretraining without reasoning optimization

Rationale: The chapter describes R1 as using a cold-start strategy with minimal supervised data and reinforcement learning.

Q46.
Aspirational 100-trillion-parameter models are compared in the chapter to approximately how many synaptic connections in the human neocortex?

A) 100 billion
B) 1 trillion
C) 10 trillion
D) 100 trillion ✓ Correct

Rationale: The chapter compares 100-trillion-parameter models to the estimated 100 trillion synaptic connections in the human neocortex.

Q47.
About how much GPU memory would be required just to load a dense 100-trillion-parameter model at 16-bit precision, according to the chapter?

A) 18.2 TB
B) 182 TB ✓ Correct
C) 1.82 PB
D) 18.2 PB

Rationale: The chapter estimates roughly 182 TB of GPU memory to load such a model at 16-bit precision.

Q48.
In NVIDIA's GB200 or GB300 NVL72 naming, what does 'NVL' stand for?

A) Nonvolatile logic
B) Neural vector lattice
C) NVLink ✓ Correct
D) Network virtualization layer

Rationale: The chapter explicitly notes that NVL in NVL72 stands for NVLink.

Q49.
What does the chapter mean by 'mechanical sympathy' in the AI context?

A) Avoiding hardware-specific optimizations to preserve portability
B) Writing software and algorithms with deep awareness of hardware capabilities and constraints ✓ Correct
C) Using only mechanical cooling for GPU clusters
D) Favoring CPU execution over GPU execution

Rationale: Mechanical sympathy is defined as writing software that is deeply aware of the hardware it runs on and codesigning algorithms accordingly.

Q50.
Which metric does the chapter present as the ultimate measure of useful throughput, emphasizing productive work rather than raw FLOPS or utilization?

A) Occupancy
B) Goodput ✓ Correct
C) Clock rate
D) Parameter count

Rationale: The chapter identifies goodput as the key metric for useful end-to-end training or inference throughput.

Q51.
A startup must train a large language model using export-compliant GPUs with reduced interconnect bandwidth. Which action best reflects the chapter's recommended response to this constraint?

A) Delay development until unrestricted top-tier GPUs become available
B) Increase model size immediately to compensate for weaker hardware
C) Prioritize communication-efficient software and algorithmic optimizations ✓ Correct
D) Disable profiling tools to reduce engineering overhead

Rationale: The chapter emphasizes that constrained hardware can still achieve strong results through communication-aware software and algorithmic optimization.

Q52.
An AI systems performance engineer notices that GPUs are frequently idle during distributed training because gradient synchronization is slow. What is the most appropriate first optimization direction?

A) Overlap communication with computation during synchronization ✓ Correct
B) Replace all GPUs with CPUs to simplify scheduling
C) Increase supervised fine-tuning data volume
D) Reduce benchmark frequency to shorten runs

Rationale: The chapter repeatedly identifies communication/computation overlap as a key strategy for reducing synchronization-related idle time.

Q53.
A team reports 98% GPU utilization, yet model training completes far slower than planned because of restarts and data stalls. Which metric should the performance engineer emphasize to leadership?

A) Clock speed
B) Goodput ✓ Correct
C) Parameter count
D) Floating-point precision

Rationale: Goodput captures useful end-to-end work completed and discounts delays, stalls, and failures that raw utilization can hide.

Q54.
During root-cause analysis, a training pipeline is slowed by a Python preprocessing stage that blocks downstream GPU work. Which intervention is most aligned with the chapter's guidance?

A) Increase batch size without changing preprocessing
B) Move preprocessing to a slower storage tier
C) Reimplement the bottlenecked step in C++ or use cuPyNumeric ✓ Correct
D) Add more human-labeled data to the pipeline

Rationale: The chapter specifically cites rewriting Python preprocessing in C++ or using cuPyNumeric to remove such bottlenecks.

Q55.
A model no longer fits on a single GPU in production. Which response best matches the chapter's operational recommendations?

A) Use tensor or pipeline parallelism to distribute the model ✓ Correct
B) Reduce all profiling to save time
C) Convert the job to CPU-only execution
D) Avoid distributed execution because it increases complexity

Rationale: The chapter recommends tensor and pipeline parallelism when a model is too large for one GPU.

Q56.
A team wants to compare two AI systems fairly after tuning custom kernels. Which benchmark practice is most appropriate?

A) Use anecdotal reports from developers who felt one system was faster
B) Use reproducible, standardized benchmarks such as MLPerf ✓ Correct
C) Compare only marketing specifications from vendors
D) Rely only on per-GPU peak FLOPS values

Rationale: The chapter advocates reproducible benchmarking and cites MLPerf as a standard for fair comparison.

Q57.
An organization has many small inference jobs that do not require a full GPU each. Which resource-management approach is most suitable?

A) Disable GPU partitioning to preserve exclusivity
B) Use GPU virtualization such as MIG to improve utilization ✓ Correct
C) Move all jobs to a single CPU socket
D) Force every workload to use maximum tensor parallelism

Rationale: The chapter notes that MIG can partition GPU resources to improve utilization when full GPU power is unnecessary.

Q58.

A performance engineer proposes a new CUDA driver version to improve efficiency, but the change affects multiple services. What is the best professional approach?

A) Deploy it unilaterally because performance takes priority

B) Coordinate with DevOps, infrastructure, and support teams ✓ Correct

C) Wait until the next hardware generation instead of changing software

D) Apply the change only after removing all monitoring

Rationale: The chapter stresses cross-team collaboration for changes such as driver and CUDA updates.

Q59.

A researcher claims a tuning change improved training because the system 'felt faster.' According to the chapter, what is the best response from the performance engineer?

A) Accept the claim because expert intuition is usually sufficient

B) Document the feeling but avoid further testing

C) Require reproducible measurements before concluding improvement ✓ Correct

D) Replace the workload with a smaller benchmark permanently

Rationale: The chapter explicitly warns against anecdotal 'vibe' optimizations and advocates rigorous, reproducible measurement.

Q60.

A company wants to learn from DeepSeek's open releases to improve its own MoE stack. Which open-source component would be most directly relevant for expert communication optimization?

A) 3FS

B) DeepEP ✓ Correct

C) FlashMLA

D) EPLB

Rationale: DeepEP is described as DeepSeek's highly tuned communication library for mixture-of-experts models.

Q61.

A cluster administrator compares two training platforms only by per-GPU numbers and declares one universally superior. Which caution from the chapter is most relevant?

A) Per-GPU results should be ignored in all cases

B) System-level end-to-end throughput is the preferred basis for comparison ✓ Correct

C) Only CPU memory determines training performance

D) Inference metrics are always interchangeable with training metrics

Rationale: The chapter notes that MLPerf cautions against relying primarily on per-GPU figures and recommends system-level throughput.

Q62.
A team needs to keep GPUs fully fed during large training runs and reduce delays from storage access. Which operational focus best aligns with the chapter?

A) Increase model parameters without changing I/O
B) Optimize storage I/O and data delivery to GPUs ✓ Correct
C) Disable asynchronous transfers
D) Reduce CPU core availability to simplify scheduling

Rationale: Efficient resource management in the chapter includes storage I/O optimization and ensuring GPUs receive data at full throttle.

Q63.
A hospital research center uses an 8-GPU node that processes 100,000 tokens in 10 seconds. If the theoretical peak is 12,000 tokens per second across the node, what is the goodput efficiency?

A) 30.0%
B) 60.0%
C) 83.3% ✓ Correct
D) 120.0%

Rationale: The achieved throughput is 10,000 tokens/s, and 10,000 divided by 12,000 equals 83.3%.

Q64.
A performance engineer is asked why small efficiency gains matter financially in large AI clusters. Which explanation best matches the chapter?

A) Minor gains matter only in academic prototypes
B) At scale, small percentage improvements can save millions of dollars ✓ Correct
C) Efficiency gains mainly improve model aesthetics rather than cost
D) Savings occur only when switching from GPUs to CPUs

Rationale: The chapter emphasizes that even incremental efficiency improvements can yield major cost savings at scale.

Q65.
An MoE model is proposed as an alternative to a dense model for a very large deployment. Which expected benefit is most consistent with the chapter?

A) Every parameter is activated for each token, maximizing accuracy
B) Inference latency may be lower because fewer parameters are active per token ✓ Correct
C) Communication overhead is eliminated entirely
D) The model no longer requires memory optimization

Rationale: The chapter explains that MoE models activate only subsets of parameters per token, often reducing inference latency versus dense equivalents.

Q66.

A team is designing a sparse model for scale. Which description best captures how MoE models reduce effective computation?

A) They quantize all weights to zero during inference

B) They route each token through only a subset of experts ✓ Correct

C) They force every GPU to hold the full model replica

D) They eliminate the need for training data

Rationale: MoE efficiency comes from activating only selected experts for each input token rather than the entire model.

Q67.

A researcher asks why a 100-trillion-parameter model cannot simply be loaded onto one modern GPU. Which answer is most accurate from the chapter?

A) The model exceeds the memory capacity of a single GPU by orders of magnitude ✓ Correct

B) The model can fit if profiling tools are disabled

C) The limitation is only due to CPU clock speed

D) The issue is unrelated to memory and only concerns software licensing

Rationale: The chapter estimates about 182 TB would be needed at 16-bit precision, far beyond a single GPU's memory.

Q68.

A systems engineer is planning deployment on an NVIDIA GB200 NVL72 rack. Which statement best reflects the chapter's description of this platform?

A) It contains 72 CPUs and 36 GPUs

B) It integrates 36 Grace CPUs and 72 Blackwell GPUs in one NVLink domain ✓ Correct

C) It is a CPU-only rack optimized for preprocessing

D) It provides uniform memory latency across all attached memory

Rationale: The chapter describes the NVL72 as 36 Grace CPUs paired with 72 Blackwell GPUs in a shared NVLink domain.

Q69.

A developer assumes CUDA Unified Memory makes the NVL72 behave like one uniform memory device. What correction should the performance engineer provide?

A) Unified Memory removes all latency differences across devices

B) Remote memory access over NVLink remains nonuniform in latency and bandwidth ✓ Correct

C) Unified Memory works only for CPU memory, not GPU memory

D) Remote memory is always faster than local HBM

Rationale: The chapter states that pages can migrate or be remotely accessed, but remote memory has distinct performance characteristics.

Q70.
An engineer wants to reduce page-fault stalls when using managed allocations across Grace and Blackwell memory. Which practice is recommended?

A) Rely exclusively on automatic migration with no hints
B) Use explicit prefetching and memory advice APIs ✓ Correct
C) Disable NVLink to simplify memory behavior
D) Store all activations on network storage

Rationale: The chapter recommends cudaMemPrefetchAsync and cudaMemAdvise to reduce stalls with managed allocations.

Q71.
A performance engineer is explaining mechanical sympathy to a cross-functional team. Which description is most accurate?

A) Using software without regard to hardware details to maximize portability
B) Writing hardware-aware software and codesigning algorithms with system capabilities ✓ Correct
C) Buying the newest hardware to avoid software tuning
D) Prioritizing model size over memory behavior

Rationale: Mechanical sympathy in the chapter means deeply understanding hardware and codesigning software and algorithms accordingly.

Q72.
A transformer workload is bottlenecked by attention memory movement on long sequences. Which algorithmic change from the chapter is most likely to help first?

A) FlashAttention ✓ Correct
B) Round-robin CPU scheduling
C) Disabling Tensor Cores
D) Increasing FP64 usage

Rationale: FlashAttention is presented as a hardware-aware attention algorithm that reduces memory traffic and speeds long-sequence workloads.

Q73.
A team is evaluating why FlashAttention became widely adopted so quickly. Which rationale best matches the chapter?

A) It increased memory movement to simplify debugging
B) It reduced attention runtime and memory footprint substantially ✓ Correct
C) It replaced all distributed communication libraries
D) It required no hardware awareness to implement

Rationale: The chapter notes that FlashAttention often produced 2×–4× speedups and lower memory use, making it rapidly attractive.

Q74.
A constrained H800-based deployment seeks an attention kernel tuned to NVIDIA memory hierarchy and Tensor Cores. Which chapter example is most directly relevant?

A) MLA ✓ Correct
B) MIG
C) MLPerf
D) FSDP

Rationale: DeepSeek's MLA is described as a hardware-aware attention algorithm optimized for NVIDIA GPUs, including constrained H800 systems.

Q75.
During architecture planning, a team asks why NVIDIA added Transformer Engine support for reduced precision. Which answer best reflects the chapter's hardware-software codesign theme?

A) Because reduced precision increases software complexity without performance benefit
B) Because transformer workloads motivated hardware features that accelerate their dominant operations ✓ Correct
C) Because FP64 remains the primary format for modern LLM training
D) Because Tensor Cores are mainly intended for storage workloads

Rationale: The chapter explains that transformer and quantization trends drove NVIDIA to add specialized hardware such as Transformer Engine and reduced-precision Tensor Cores.

Q76.
A distributed training job appears fully booked in the scheduler, yet only 30% of theoretical sample throughput is realized because of synchronization and data delays. How should this be characterized?

A) As high goodput because utilization is near 100%
B) As low goodput because useful work is far below theoretical capability ✓ Correct
C) As a memory-capacity problem only
D) As evidence that profiling is unnecessary

Rationale: Goodput reflects useful completed work, so 30% realized throughput indicates poor end-to-end efficiency despite apparent utilization.

Q77.
A performance engineer identifies data-loading stalls as the primary cause of poor goodput. Which intervention is most aligned with the chapter?

A) Introduce caching or asynchronous prefetch ✓ Correct
B) Increase network contention to stress the cluster
C) Reduce all benchmark visibility
D) Switch all jobs to dense models

Rationale: The chapter specifically recommends caching or async prefetch when GPUs are waiting on data loading.

Q78.
A training team wants to improve cluster efficiency by 20%. According to the chapter, why is hiring an AI systems performance engineer often justified?

A) Because the role mainly reduces documentation needs
B) Because efficiency gains can generate returns that exceed the engineer's cost ✓ Correct
C) Because performance engineers eliminate the need for researchers
D) Because only hardware vendors can improve efficiency materially

Rationale: The chapter argues that performance engineers often deliver ROI well above cost through substantial efficiency gains.

Q79.
A manager asks what an AI systems performance engineer must understand to improve goodput effectively. Which answer best matches the chapter?

A) Only model architecture, not infrastructure
B) Only GPU clocks and temperatures
C) Interactions among hardware, software, and algorithms ✓ Correct
D) Only benchmark naming conventions

Rationale: The chapter emphasizes that goodput depends on deep understanding across hardware, software, and algorithmic layers.

Q80.
A team is deciding whether to benchmark, profile, or both while diagnosing inference latency. What approach is most consistent with the chapter?

A) Use only benchmarking because profiling is too detailed
B) Use only profiling because benchmarks are unnecessary
C) Use iterative benchmarking and profiling across different stack layers ✓ Correct
D) Avoid both until the system reaches production scale

Rationale: The chapter recommends iterative use of tools such as Nsight Systems, Nsight Compute, and the PyTorch profiler to identify bottlenecks and track progress.

Q81.
A production team wants to catch performance regressions before release. Which operational practice is recommended in the chapter?

A) Run manual checks only after deployment
B) Set up automated performance tests early in development ✓ Correct
C) Benchmark only once per hardware generation
D) Disable profiling in continuous integration environments

Rationale: The chapter explicitly recommends automated performance tests to catch regressions early in the development cycle.

Q82.
A distributed inference service needs low-latency point-to-point movement of data across GPU memory and storage tiers. Which library from the chapter is most relevant?

A) NIXL ✓ Correct
B) NCCL
C) cuPyNumeric
D) MLPerf

Rationale: NIXL is described as providing high-throughput, low-latency point-to-point data movement for distributed inference.

Q83.
A training engineer needs optimized collective communication for all-reduce across many GPUs. Which tool is the most appropriate according to the chapter?

A) NCCL ✓ Correct
B) PyTorch profiler
C) 3FS
D) MIG

Rationale: The chapter identifies NCCL as the key library for distributed collectives such as all-reduce.

Q84.
A large MoE workload suffers because some experts are overloaded while others are underused. Which DeepSeek open-source project most directly addresses this issue?

A) DualPipe
B) FlashMLA
C) EPLB ✓ Correct
D) 3FS

Rationale: EPLB is described as an expert parallelism load balancer that duplicates heavily loaded experts to handle excess demand.

Q85.
A pipeline-parallel training run has excessive bubbles that reduce throughput. Which DeepSeek contribution is most directly intended to mitigate this problem?

A) DualPipe ✓ Correct
B) DeepGEMM
C) 3FS
D) MIG

Rationale: DualPipe is presented as a bidirectional pipeline parallelism algorithm that overlaps computation and communication to reduce pipeline bubbles.

Q86.
An inference cluster is saturating neither its NVMe SSDs nor RDMA links, causing data starvation. Which DeepSeek open-source project best matches this layer of optimization?

A) FlashMLA
B) DeepEP
C) 3FS ✓ Correct
D) EPLB

Rationale: 3FS is described as a high-performance distributed filesystem intended to saturate NVMe SSD and RDMA bandwidth.

Q87.
A team wants to use a matrix multiplication library optimized for FP8 dense and sparse operations. Which DeepSeek component is the best fit?

A) DeepGEMM ✓ Correct
B) DualPipe
C) NIXL
D) MLA

Rationale: DeepGEMM is described as an FP8-optimized matrix multiplication library for dense and sparse operations.

Q88.
A project lead asks why DeepSeek's open-sourcing strategy mattered beyond publicity. Which answer best reflects the chapter?

A) It reduced the need for benchmarking by others
B) It enabled reproducibility, credibility, and community learning ✓ Correct
C) It guaranteed stock market stability
D) It eliminated hardware constraints for other teams

Rationale: The chapter emphasizes that open sourcing allowed others to reproduce results, validate claims, and learn from the methods.

Q89.
A clinician-researcher new to AI asks why brute-force scaling alone is not a practical path to 100-trillion-parameter models. Which response best matches the chapter?

A) Because current GPUs lack any support for transformer workloads
B) Because compute, memory, communication, energy, and cost requirements become enormous ✓ Correct
C) Because sparse models cannot exceed one trillion parameters
D) Because benchmark suites prohibit models of that size

Rationale: The chapter argues that extreme-scale models are limited by massive compute, memory, communication, power, and financial demands.

Q90.
A team wants lower request-response latency at very large parameter counts. Which architectural direction from the chapter is most likely to help?

A) Use a sparse MoE design with fewer active parameters per token ✓ Correct
B) Force every token through the full dense model
C) Increase FP64 usage for all inference kernels
D) Disable expert routing to simplify computation

Rationale: The chapter notes that MoE models often have lower inference latency because only part of the model is active per token.

Q91.
A system architect proposes using only hardware upgrades and ignoring algorithm redesign. Which chapter principle most directly challenges this plan?

A) Mechanical sympathy and hardware-software codesign ✓ Correct
B) Per-GPU metrics over system metrics
C) CPU exclusivity for all workloads
D) Static scheduling without profiling

Rationale: The chapter consistently argues that hardware, software, and algorithms must be codesigned rather than improved in isolation.

Q92.
A training run on constrained interconnects needs to hide communication delays as much as possible. Which strategy best reflects DeepSeek's approach?

A) Serializing all communication after computation completes
B) Overlapping communication with ongoing computation ✓ Correct
C) Disabling expert routing during training
D) Using only default collectives without kernel changes

Rationale: DeepSeek's approach is described as overlapping communication and computation to mask bandwidth limitations.

Q93.
A manager asks what distinguishes the AI systems performance engineer from a narrowly focused software developer. Which answer best fits the chapter?

A) The role concentrates only on application UI responsiveness
B) The role spans benchmarking, profiling, debugging, optimization, scaling, and resource management across the stack ✓ Correct
C) The role avoids collaboration to preserve objectivity
D) The role is limited to selecting cloud vendors

Rationale: The chapter defines the role as multidisciplinary, spanning hardware, software, algorithms, and operational optimization tasks.

Q94.
A team sees slower multi-GPU scaling on H800 than expected compared with H100. Which hardware difference from the chapter most directly explains this issue?

A) Lower HBM capacity on H800
B) Lower NVLink interconnect bandwidth on H800 ✓ Correct
C) Absence of CUDA support on H800
D) Lack of tensor parallelism on H800

Rationale: The chapter highlights H800's reduced NVLink bandwidth as a key factor limiting inter-GPU scaling.

Q95.
A reviewer asks why DeepSeek-V3 could still be computationally manageable despite its roughly 680 billion total parameters. Which explanation is most accurate?

A) It used no attention mechanism
B) It activated only a subset of experts, about 37 billion active parameters per token ✓ Correct
C) It stored all weights only on CPU memory
D) It required no distributed communication

Rationale: The chapter explains that DeepSeek-V3 is an MoE model with only a fraction of total parameters active per token.

Q96.
A team is deciding whether to prioritize FLOPS, utilization, or useful completed work when evaluating optimization success. Which choice best aligns with the chapter's central metric philosophy?

A) Prioritize useful completed work through goodput ✓ Correct
B) Prioritize utilization only because it is easiest to report
C) Prioritize parameter count because bigger models are always better
D) Prioritize rack power draw over all other measures

Rationale: The chapter presents goodput as the ultimate measure because it reflects useful work done per time, cost, and energy.

Q97.
A research lab wants to improve inference throughput by exploiting the NVL72's single NVLink domain across 72 GPUs. Which software examples from the chapter are specifically mentioned as able to use this efficiently?

A) PyTorch for training and vLLM for inference ✓ Correct
B) Only spreadsheet software and shell scripts
C) TensorRT for storage and NCCL for visualization
D) Only CPU-based orchestration frameworks

Rationale: The chapter explicitly mentions PyTorch and vLLM as frameworks that can exploit the NVL72's single NVLink domain.

Q98.

A performance engineer must explain why cross-team collaboration is essential when improving AI system performance. Which scenario from the chapter best illustrates this need?

A) A single-threaded script requires no external coordination

B) Updating model code and CUDA drivers may require researchers plus infrastructure and DevOps teams ✓ Correct

C) Benchmarking always occurs in isolation from operations

D) Inference tuning never affects storage or networking teams

Rationale: The chapter emphasizes that many performance improvements span researchers, infrastructure, DevOps, and support teams.

Q99.

A team plans a new benchmark report and wants to present the most scientifically sound workflow. Which sequence best matches the chapter's methodology?

A) Assume a bottleneck, optimize once, and publish only the best result

B) Develop a hypothesis, measure reproducibly, adjust, rerun, and share results ✓ Correct

C) Tune by intuition, then estimate gains from utilization alone

D) Skip profiling and compare only vendor claims

Rationale: The chapter advocates a rigorous scientific process of hypothesis, reproducible measurement, iteration, and transparent reporting.

Q100.

An executive asks for the most strategic takeaway from Chapter 1 when planning future AI infrastructure investments. Which answer best captures the chapter's conclusion?

A) Performance engineering is optional once enough hardware is purchased

B) Optimization across hardware, software, and algorithms is necessary to make extreme-scale AI practical and affordable ✓ Correct

C) Only new model architectures matter for future capability gains

D) Financial outcomes are largely unaffected by efficiency breakthroughs

Rationale: The chapter concludes that co-optimization across the stack is essential because brute-force hardware spending alone is insufficient and inefficient at scale.

Q101.

A company claims its 2,048-GPU training cluster runs at 98% GPU utilization, yet model progress per day is far below target because jobs frequently stall on synchronization and restart after faults. Which metric should the performance engineer prioritize to best evaluate true system effectiveness?

A) Peak theoretical FLOPS per GPU

B) Goodput normalized to useful training completed over time ✓ Correct

C) Aggregate installed HBM capacity

D) Average GPU temperature under load

Rationale: Goodput captures useful end-to-end work and discounts stalls, communication overhead, and recovery time that utilization alone hides.

Q102.

An executive proposes solving poor training throughput simply by doubling the number of GPUs in a communication-bound workload. Based on the chapter's central argument, what is the strongest critique of this plan?

A) Additional GPUs always reduce model accuracy in distributed training

B) Scaling hardware without addressing bottlenecks can amplify inefficiency and cost rather than improve useful throughput ✓ Correct

C) Larger clusters cannot use NCCL collectives effectively

D) Inference optimizations are more important than training optimizations in all cases

Rationale: The chapter emphasizes that brute-force scaling is insufficient when communication, software, or algorithmic bottlenecks dominate.

Q103.

A team wants to compare two AI platforms using only per-GPU MLPerf numbers because one platform shows better per-device throughput. What is the most defensible evaluation response?

A) Accept the comparison because per-GPU metrics are the primary MLPerf benchmark criterion

B) Reject all benchmark comparisons unless both systems use identical model weights

C) Prefer system-level end-to-end throughput, using per-GPU numbers only as component-level indicators ✓ Correct

D) Use only rack power draw because throughput scales directly with energy consumption

Rationale: The chapter notes that MLPerf cautions against treating per-GPU results as the primary cross-platform metric and recommends system-level throughput.

Q104.

A research group is deciding whether to invest in custom kernels for an export-constrained GPU cluster with weak interconnect bandwidth. Which precedent from the chapter most strongly supports that investment?

A) Dense models always outperform sparse models on reasoning benchmarks

B) DeepSeek used software and algorithmic innovations to approach frontier performance despite H800 hardware limitations ✓ Correct

C) GPU virtualization eliminates all communication bottlenecks in large clusters

D) Blackwell systems remove the need for profiling and benchmarking

Rationale: DeepSeek's example shows that hardware-aware engineering can compensate for constrained hardware and materially improve outcomes.

Q105.

Which scenario best exemplifies mechanical sympathy in AI systems performance engineering?

A) Selecting the largest possible dataset regardless of storage bandwidth

B) Rewriting an attention algorithm to reduce GPU memory traffic and better exploit Tensor Cores ✓ Correct

C) Increasing model parameter count without changing software or hardware strategy

D) Comparing cloud pricing across vendors without profiling workloads

Rationale: Mechanical sympathy means designing software and algorithms with deep awareness of hardware behavior, such as memory hierarchy and compute units.

Q106.

A performance engineer is asked to justify why small efficiency gains matter financially in ultrascale AI systems. Which rationale is most aligned with the chapter?

A) Minor optimizations are valuable mainly because they improve benchmark aesthetics

B) At scale, even modest percentage gains can translate into millions or billions of dollars saved ✓ Correct

C) Small gains matter only for inference, not for training

D) Efficiency gains are secondary because cloud providers absorb most hardware costs

Rationale: The chapter repeatedly stresses that small efficiency improvements compound into major cost savings at large scale.

Q107.

A team evaluating DeepSeek's reported training cost notices uncertainty about what the stated figure includes. What is the most analytically sound conclusion?

A) The claim should be accepted as a complete accounting of total model-development cost

B) The claim is irrelevant because compute cost never affects market perceptions

C) The figure may represent only a subset of costs, so conclusions should distinguish a single run from the full experimentation pipeline ✓ Correct

D) The claim proves that hardware choice no longer matters in AI training

Rationale: The chapter explicitly notes uncertainty about whether the reported amount included only one run or excluded broader development costs.

Q108.

In a postmortem, a team finds that GPUs are idle because a Python preprocessing stage cannot keep up with training. Which response best reflects the chapter's optimization philosophy?

A) Increase batch size until the preprocessing bottleneck disappears

B) Reimplement the bottlenecked preprocessing in C++ or use a distributed array approach such as cuPyNumeric ✓ Correct

C) Disable profiling because the root cause is already known

D) Move the entire pipeline to CPU execution for consistency

Rationale: The chapter gives this exact type of example, recommending lower-level or distributed implementations to remove Python preprocessing bottlenecks.

Q109.

A 680-billion-parameter MoE model activates only a subset of experts per token. Why is this design strategically important for scaling?

A) It guarantees uniform network traffic across all nodes

B) It keeps FLOPS per token from increasing proportionally with total parameter count ✓ Correct

C) It removes the need for any distributed parallelism

D) It ensures all experts are updated on every token

Rationale: Sparse MoE routing allows total parameters to grow without a matching explosion in compute per token.

Q110.

A systems team is planning for future 100-trillion-parameter training using today's methods. Which conclusion from the chapter should most strongly guide strategy?

A) Current brute-force scaling is sufficient if enough racks are purchased

B) Only denser models can achieve that scale efficiently

C) Breakthroughs in hardware-software-algorithm codesign are required because brute force is impractical ✓ Correct

D) Storage throughput is unimportant once model weights fit in memory

Rationale: The chapter argues that 100-trillion-parameter training is not feasible through simple scaling alone and requires new codesigned approaches.

Q111.

A performance engineer must explain why DeepSeek's H800-based cluster faced scaling challenges relative to H100 systems. Which hardware difference is most relevant?

A) The H800 has substantially higher NVLink bandwidth than the H100

B) The H800 reduces interconnect bandwidth, making inter-GPU transfers slower and multi-GPU scaling harder ✓ Correct

C) The H800 lacks GPU memory entirely for model weights

D) The H800 removes support for transformer workloads

Rationale: The chapter highlights the H800's reduced NVLink bandwidth as a key limitation for distributed scalability.

Q112.

A team wants to improve distributed training efficiency on a cluster where gradient synchronization dominates step time. Which strategy is most directly supported by the chapter?

A) Replace all GPUs with CPUs to reduce synchronization overhead

B) Overlap computation with communication and tune collectives such as all-reduce ✓ Correct

C) Disable model parallelism and force single-GPU training

D) Increase only storage capacity to improve synchronization speed

Rationale: The chapter emphasizes optimizing collectives and overlapping communication with computation to reduce synchronization penalties.

Q113.

Which pairing best reflects the chapter's distinction between benchmarking and profiling?

A) Benchmarking identifies legal compliance, while profiling estimates cloud cost

B) Benchmarking measures system performance under workloads, while profiling pinpoints bottlenecks across stack levels ✓ Correct

C) Benchmarking is qualitative, while profiling is purely financial

D) Benchmarking replaces debugging, while profiling replaces optimization

Rationale: Benchmarking quantifies performance outcomes, whereas profiling helps locate root causes of inefficiency.

Q114.

A cluster appears fully occupied, but only 30% of theoretical sample throughput becomes actual training progress because of input stalls and synchronization delays. How should this be interpreted?

A) The cluster has 30% goodput despite high apparent utilization ✓ Correct

B) The cluster has 70% goodput because overhead counts as productive work

C) The cluster is underclocked but otherwise efficient

D) The cluster has optimal performance because utilization is near 100%

Rationale: Goodput measures useful completed work, so realized throughput at 30% of theoretical indicates 30% goodput.

Q115.

A performance engineer must choose the most appropriate explanation for why FlashAttention became widely adopted so quickly. Which is best?

A) It increased model parameter count without additional memory cost

B) It reduced memory movement in attention, producing substantial speedups and lower memory footprint ✓ Correct

C) It removed the need for GPUs in transformer inference

D) It standardized all benchmark methodologies across vendors

Rationale: FlashAttention's hardware-aware tiling reduced memory traffic and significantly accelerated long-sequence attention.

Q116.

A team is considering CUDA Unified Memory across a GB200 NVL72 rack and assumes the entire rack behaves like a single uniform memory pool. What is the best correction?

A) That assumption is correct because NVLink eliminates all memory locality effects

B) Remote pages can be accessed over NVLink, but latency and bandwidth are nonuniform, so explicit prefetch and advice are preferred ✓ Correct

C) Unified Memory is unsupported on Grace Blackwell systems

D) Only CPU memory can be accessed remotely across the rack

Rationale: The chapter states that managed memory can migrate or be remotely accessed, but performance remains nonuniform and should be managed carefully.

Q117.

A manager asks why AI systems performance engineers require cross-team collaboration rather than working independently within model code. Which response is most accurate?

A) Performance improvements often span model code, drivers, networking, storage, and operations, requiring coordinated changes across teams ✓ Correct

B) Collaboration is needed mainly to obtain benchmark screenshots for reports

C) Only infrastructure teams can change model performance meaningfully

D) Researchers should avoid interacting with performance engineers to preserve experimental validity

Rationale: The chapter describes performance engineering as inherently multidisciplinary, often requiring coordinated software, hardware, and operational changes.

Q118.

Which example best demonstrates transparency and reproducibility as described in the chapter?

A) Publishing only final speedup percentages without code or methodology

B) Sharing open-source kernels, communication libraries, and benchmark artifacts so others can reproduce results ✓ Correct

C) Reporting vendor claims without disclosing test conditions

D) Using proprietary internal metrics that cannot be independently verified

Rationale: The chapter praises open sourcing infrastructure optimizations and benchmark methods to enable reproducibility and community learning.

Q119.

A team must compare dense and MoE architectures for a cost-constrained deployment. Which statement is most supported by the chapter?

A) Dense models always deliver lower inference latency than MoE models of similar total size

B) MoE models can scale total parameters while activating fewer parameters per token, often reducing effective compute and latency ✓ Correct

C) MoE models eliminate the need for routing logic and communication tuning

D) Dense models are the only practical path to 100-trillion-parameter systems

Rationale: The chapter explains that sparse MoE routing can keep per-token compute lower and often improve inference efficiency relative to dense equivalents.

Q120.

An engineering lead wants one sentence capturing the business case for the role. Which summary is most faithful to the chapter?

A) The role mainly exists to improve benchmark marketing claims

B) The role converts raw hardware into reliable, scalable, cost-efficient AI throughput by removing bottlenecks across the stack ✓ Correct

C) The role is limited to writing CUDA kernels for training only

D) The role primarily replaces data scientists in model design

Rationale: The chapter defines the role as optimizing training and inference systems for speed, cost, scalability, reliability, and availability.

Q121.

A team wants to improve performance on a workload where communication bandwidth is scarce. Which design principle from DeepSeek's approach is most applicable?

A) Treat communication as effectively free compared with compute

B) Optimize every byte moved and overlap transfers with computation wherever possible ✓ Correct

C) Increase supervised fine-tuning data regardless of system constraints

D) Prefer dense activation of all experts to simplify routing

Rationale: DeepSeek's strategy explicitly treated bandwidth as scarce and focused on minimizing and hiding communication costs.

Q122.
Which recommendation best addresses early detection of performance regressions in an AI development lifecycle?

A) Run manual benchmarks only before production release
B) Set up automated performance tests integrated into the development cycle ✓ Correct
C) Rely on user reports to identify latency slowdowns
D) Use utilization dashboards instead of benchmarks

Rationale: The chapter explicitly recommends automated performance testing to catch regressions early.

Q123.
A model no longer fits on a single GPU. Which response is most aligned with the chapter's scaling guidance?

A) Abandon the model because single-GPU fit is required for efficient training
B) Use tensor parallelism, pipeline parallelism, or other distributed strategies to partition the workload ✓ Correct
C) Convert all model weights to FP64 to simplify memory management
D) Disable communication libraries to reduce overhead

Rationale: The chapter recommends redesigning workloads with distributed parallel strategies when models exceed single-GPU capacity.

Q124.
A systems architect is evaluating the GB200 NVL72 for large-model training. Which feature most directly supports efficient intra-rack multi-GPU communication?

A) An Ethernet-only topology that isolates each GPU memory space
B) A 72-GPU NVLink domain connected by NVSwitch with high aggregate bandwidth ✓ Correct
C) A requirement that all communication route through external storage
D) A CPU-only interconnect fabric between Grace processors

Rationale: The NVL72 is built around an NVLink and NVSwitch fabric that unifies 72 GPUs into a high-bandwidth communication domain.

Q125.
A team is impressed by a system's high theoretical exaFLOPS rating and assumes this alone guarantees fast model training. Which evaluation stance is most appropriate?

A) Agree, because theoretical FLOPS directly equals realized training speed
B) Question the assumption and measure whether hardware, software, and algorithms convert theoretical capacity into goodput ✓ Correct
C) Ignore end-to-end throughput because exaFLOPS is more objective
D) Use rack power only as the deciding metric

Rationale: The chapter distinguishes theoretical capability from realized useful throughput and emphasizes goodput as the more meaningful metric.

Q126.
A performance engineer is selecting a communication library for distributed collectives such as all-reduce during training. Which choice is specifically identified for that role?

A) NCCL ✓ Correct
B) vLLM
C) TensorRT LLM
D) MLPerf

Rationale: The chapter identifies NCCL as the key library for distributed collectives like all-reduce.

Q127.
Which statement best explains why DeepSeek's open-source releases strengthened the credibility of its efficiency claims?

A) They prevented any future hardware improvements from mattering
B) They enabled others to benchmark, reproduce, and inspect the optimizations behind the reported results ✓ Correct
C) They guaranteed all competitors would achieve identical performance
D) They replaced the need for standardized benchmarks such as MLPerf

Rationale: Openly releasing kernels and infrastructure artifacts supports independent reproduction and validation of claims.

Q128.
A cluster engineer must decide whether to use MIG on underutilized GPUs serving small inference jobs. Which rationale from the chapter most supports this?

A) MIG can partition GPU resources to improve overall utilization when full-GPU allocation is unnecessary ✓ Correct
B) MIG increases NVLink bandwidth between nodes
C) MIG is primarily a benchmarking standard for inference
D) MIG eliminates the need for batching or scheduling

Rationale: The chapter cites MIG as a way to partition GPU resources for better utilization when a full GPU is not needed.

Q129.
A team wants to improve model-serving latency by reducing active computation per request while preserving large total model capacity. Which architecture best fits this objective?

A) A dense model activating all parameters for every token
B) A sparse MoE model activating only selected experts per token ✓ Correct
C) A CPU-only recurrent model with no parallelism
D) A full-precision model with increased parameter duplication

Rationale: MoE models preserve large total parameter counts while activating only a subset of experts per token, reducing effective per-request compute.

Q130.

A researcher argues that because a GPU rack offers pooled memory access, memory placement choices are no longer important. Which reply is most accurate?

A) Correct, because pooled access eliminates all locality and page-fault effects

B) Incorrect, because remote memory access remains distinct in latency and bandwidth, so placement and prefetch still matter ✓ Correct

C) Correct, but only for inference and not training

D) Incorrect, because GPU memory cannot be accessed across NVLink at all

Rationale: The chapter warns that pooled access does not make memory uniform; remote access still has different performance characteristics.

Q131.

Which option best captures the chapter's recommended methodology for optimization work?

A) Make intuitive changes until the system feels faster

B) Form hypotheses, measure with reproducible benchmarks, adjust, rerun, and share results ✓ Correct

C) Optimize only the GPU kernels and ignore the rest of the stack

D) Start with hardware replacement before profiling software

Rationale: The chapter advocates a rigorous, scientific, profile-driven optimization process rather than anecdotal tuning.

Q132.

A leader asks why the AI systems performance engineer role commands high salaries. Which explanation is most consistent with the chapter?

A) The role focuses on a narrow coding niche with little business impact

B) The role combines rare cross-stack expertise and directly affects cost, throughput, reliability, and scalability ✓ Correct

C) The role is compensated mainly for cloud-vendor certification requirements

D) The role is valuable only in academic benchmarking competitions

Rationale: The chapter emphasizes the scarcity of cross-disciplinary expertise and the role's direct bottom-line impact.

Q133.

A team needs to move key-value cache data efficiently in distributed inference for disaggregated prefill-decode. Which library is highlighted for high-throughput, low-latency point-to-point movement across memory and storage tiers?

A) NIXL ✓ Correct

B) Nsight Compute

C) PyTorch profiler

D) MIG

Rationale: The chapter identifies NIXL as the library for low-latency, high-throughput data movement in distributed inference contexts.

Q134.
Which finding would most strongly indicate that a training system's apparent hardware saturation is misleading?

A) GPU utilization is high, but effective training time ratio is severely reduced by preemptions and failures ✓ Correct
B) The cluster has more CPU cores than GPUs
C) Per-GPU throughput exceeds a previous hardware generation
D) The model uses reduced precision arithmetic

Rationale: The chapter notes that preemptions, faults, and congestion can drastically reduce useful work even when utilization appears high.

Q135.
A platform architect wants to justify investment in filesystem optimization for AI training. Which chapter insight is most relevant?

A) Filesystems are rarely meaningful bottlenecks once GPUs are provisioned
B) Every layer, including the filesystem, can constrain overall AI system performance and deserves optimization ✓ Correct
C) Storage matters only for inference and not for training
D) Distributed filesystems are incompatible with reproducible benchmarking

Rationale: The chapter uses DeepSeek's 3FS example to stress that even the filesystem layer can materially affect end-to-end performance.

Q136.
A team is debating whether to rely on more human-labeled fine-tuning data or algorithmic innovation for reasoning model development under budget constraints. Which DeepSeek lesson is most relevant?

A) Reasoning quality depends exclusively on larger supervised datasets
B) A cold-start strategy with minimal supervised data and reinforcement learning can reduce cost and time while improving reasoning behavior ✓ Correct
C) Reasoning models cannot benefit from MoE foundations
D) Human feedback loops are unnecessary in any reasoning system

Rationale: The chapter describes DeepSeek-R1 as using minimal supervised data with reinforcement learning to reduce training cost and time.

Q137.
Which choice best explains why a 100-trillion-parameter dense model creates an extraordinary memory challenge even before considering activations or optimizer state?

A) Because 16-bit weights alone would require on the order of hundreds of terabytes of GPU memory ✓ Correct
B) Because dense models cannot be represented in binary formats
C) Because CPU memory cannot be paired with GPU memory in any architecture
D) Because inference requires more weight storage than training

Rationale: The chapter estimates roughly 182 TB just to store 100 trillion parameters at 16-bit precision.

Q138.

A performance engineer must select the most appropriate first action after noticing poor scaling from 8 GPUs to 1,024 GPUs. Which is best?

A) Assume the issue is unavoidable because large clusters always lose efficiency

B) Profile communication, workload balance, and data movement to identify the dominant scaling bottleneck ✓ Correct

C) Increase model depth so compute dominates communication

D) Disable distributed collectives and synchronize manually on CPU

Rationale: The chapter emphasizes profile-driven diagnosis of communication, memory, and load-balance bottlenecks when scaling out.

Q139.

A new engineer asks why hardware, software, and algorithms must be codesigned rather than optimized independently. What is the strongest answer from the chapter?

A) Because improvements in one layer often shift or expose bottlenecks in another, requiring coordinated trade-off management ✓ Correct

B) Because software has no measurable effect once hardware is selected

C) Because algorithms determine cost but not throughput

D) Because codesign is needed only for benchmark submissions

Rationale: The chapter repeatedly stresses that hardware, software, and algorithm choices interact and must be understood together.

Q140.

A team is considering whether to trust anecdotal reports that a configuration change made the model server 'feel faster.' Which response best fits the chapter's guidance?

A) Accept the change if multiple users report subjective improvement

B) Reject any optimization that cannot be explained mathematically

C) Instrument the system and validate the change with reproducible benchmarks and profiling data ✓ Correct

D) Focus on raw device utilization as the only necessary evidence

Rationale: The chapter explicitly warns against 'vibe' optimizations and advocates reproducible measurement.

Q141.

Which statement most accurately characterizes the GB200 or GB300 NVL72 as presented in the chapter?

A) It is a single uniform-memory GPU with no locality concerns

B) It is a rack-scale AI supercomputer integrating 72 GPUs with a high-bandwidth NVLink domain ✓ Correct

C) It is primarily a storage appliance for model checkpoints

D) It is designed only for inference, not training

Rationale: The chapter describes NVL72 as an AI supercomputer in a rack with 72 GPUs connected by NVLink and NVSwitch.

Q142.
A performance engineer wants to explain why modern GPU features such as Transformer Engine and reduced-precision Tensor Cores matter for transformers. Which explanation is best?

A) They mainly improve filesystem throughput during checkpoint writes
B) They accelerate transformer-relevant math and support hardware-aware algorithmic advances such as lower-precision computation ✓ Correct
C) They eliminate the need for attention mechanisms
D) They are useful only for FP64 scientific simulations

Rationale: The chapter links Transformer Engine and reduced-precision Tensor Cores to faster matrix math and transformer-specific acceleration.

Q143.
A distributed MoE training job suffers from overloaded experts while others remain underused. Which DeepSeek-related mechanism most directly addresses this issue?

A) MLPerf per-GPU normalization
B) EPLB redundant expert duplication for heavily loaded experts ✓ Correct
C) MIG partitioning of GPUs into smaller instances
D) CUDA Unified Memory page migration

Rationale: The chapter describes EPLB as duplicating heavily loaded experts to better handle excess demand.

Q144.
A team asks why network and storage engineers should be involved in model performance reviews. Which answer best reflects the chapter?

A) Because AI performance is determined solely by dataset size and model architecture
B) Because bottlenecks in networking and storage can substantially reduce goodput even when GPUs are powerful ✓ Correct
C) Because infrastructure teams are responsible for benchmark publication only
D) Because storage and networking matter only after training is complete

Rationale: The chapter emphasizes that hardware bottlenecks beyond the GPU, including network and storage, can limit useful throughput.

Q145.
Which scenario best illustrates the chapter's distinction between useful throughput and theoretical maximum throughput?

A) A node capable of 12,000 tokens/s peak achieves 10,000 tokens/s actual, yielding about 83.3% efficiency ✓ Correct
B) A node reports 100% utilization and therefore must have 100% goodput
C) A node with more HBM automatically has higher goodput than any smaller-memory node
D) A node with lower rack power necessarily has lower useful throughput

Rationale: The chapter provides this style of example to show how achieved throughput is compared with theoretical peak to estimate efficiency.

Q146.

A systems leader is evaluating where to focus optimization effort for a future-proof skill set. Which emphasis is most aligned with the chapter's roadmap?

A) Only cloud cost negotiation, because hardware details are abstracted away

B) End-to-end understanding from hardware fundamentals through OS, networking, storage, CUDA, distributed training, and inference systems ✓ Correct

C) Exclusive focus on model prompting, because systems tuning is becoming obsolete

D) Only benchmark administration, because profiling tools are vendor-specific

Rationale: The roadmap explicitly spans hardware, system software, CUDA, distributed training, and inference optimization.

Q147.

A team is deciding whether to tune for FLOPS, device utilization, or goodput when the business goal is more completed training per dollar and per joule. Which metric should dominate decision-making?

A) Goodput ✓ Correct

B) Peak clock frequency

C) Installed GPU count

D) Raw FLOPS alone

Rationale: The chapter states that goodput best reflects useful work completed per cost and energy, making it the ultimate success metric.

Q148.

Which conclusion best synthesizes the DeepSeek case with the chapter's broader message about AI economics?

A) Hardware constraints make frontier model development impossible outside the largest labs

B) Efficiency innovations can materially alter training and inference economics and even affect financial markets ✓ Correct

C) Inference efficiency matters less than training cost in all settings

D) Open-source releases reduce the importance of performance engineering expertise

Rationale: The chapter links DeepSeek's efficiency achievements to lower costs, competitive performance, and measurable market reactions.

Q149.

A performance engineer must prioritize one intervention for a training system whose GPUs frequently wait on data from storage. Which choice best increases goodput according to the chapter?

A) Introduce caching or asynchronous prefetch to reduce data-delivery stalls ✓ Correct

B) Increase theoretical FLOPS by overclocking the GPUs only

C) Replace all-reduce with larger checkpoint files

D) Reduce monitoring to free CPU cycles for logging

Rationale: The chapter specifically recommends caching or async prefetch when GPUs are waiting on data loading.

Q150.

A graduate student summarizes the chapter by saying, 'AI systems performance engineering is mainly about making hardware run faster.' Which revision is most accurate?

A) It is mainly about maximizing benchmark visibility for hardware vendors

B) It is about making AI systems fast, scalable, reliable, and affordable by harmonizing hardware, software, and algorithms ✓ Correct

C) It is limited to CUDA kernel tuning for transformer attention

D) It is primarily a procurement strategy for acquiring larger clusters

Rationale: The chapter frames the field as broad cross-stack optimization aimed at useful, reliable, and cost-effective AI performance rather than speed alone.