NVIDIA-Certified Associate: Generative AI LLMs (NCA-GENL)

Cheat Sheet for NVIDIA GenAI Ecosystem

Author: DolbyUUU (https://github.com/DolbyUUU)

GPU Architecture & Cores

NVIDIA Hopper (H100, H200) / Blackwell (B100, B200, GB200)

These are NVIDIA's current and next-generation flagship AI computing architectures, respectively, providing the core compute power for training and inferencing generative AI models. The Hopper architecture is the current market leader, while the Blackwell architecture, through system-level innovations, offers higher performance and efficiency for future trillion-parameter-scale giant models.

NVIDIA Tensor Cores

These are specialized hardware processing units within the GPU designed specifically for deep learning. Their core function is to dramatically accelerate matrix multiplication and accumulation operations. Since mainstream AI models like Transformers heavily rely on such computations, Tensor Cores are the physical foundation for achieving high-performance AI training and inference.

Transformer Engine

This is an innovative technology that combines hardware and software, built into the Hopper and newer GPU architectures. It can intelligently and dynamically switch between different numerical precisions, such as FP8 and FP16, thereby significantly increasing the operational speed of Transformer models without sacrificing model accuracy.

Compute & Networking Systems

NVIDIA DGX / DGX SuperPOD

NVIDIA DGX is an "all-in-one" AI supercomputer that integrates top-tier GPUs, high-speed networking, and a full suite of AI software, ready to use out of the box. The DGX SuperPOD is a reference architecture for clusters composed of multiple DGX systems, serving as the premier blueprint for enterprises building large-scale LLM training centers.

NVIDIA NVLink & NVSwitch

This is a suite of interconnect technologies for high-speed communication between GPUs, designed to solve the communication bottleneck in multi-GPU collaborative work. NVLink provides direct high-speed channels between GPUs, while NVSwitch acts like an intelligent switch, allowing up to hundreds of GPUs to communicate at full speed as if they were a single, unified processor.

NVIDIA BlueField DPU (Data Processing Unit)

This is a programmable data processor whose primary task is to offload networking, storage, and security infrastructure workloads from the CPU and GPU. In an AI cluster, the DPU can independently handle data communication tasks, allowing valuable CPU and GPU resources to be 100% focused on core AI computations.

NVIDIA Spectrum-X Ethernet Platform

This is an Ethernet networking solution optimized specifically for AI workloads, designed to deliver ultra-high performance comparable to traditional InfiniBand networks. By combining Spectrum switches with BlueField DPUs, it achieves the high bandwidth and low latency required for AI within the open Ethernet ecosystem.

NVIDIA Jetson AGX Orin / Thor

This is NVIDIA's series of embedded computing platforms for edge devices (such as robots and autonomous vehicles), designed to bring data center-level AI capabilities to the endpoint. They have lower power consumption and a smaller form factor, enabling them to run optimized generative AI models directly on the device for real-time intelligent interaction.

Core Platforms & Libraries

NVIDIA CUDA (Compute Unified Device Architecture)

NVIDIA CUDA is the core soul of the entire NVIDIA ecosystem. It is a parallel computing platform and programming model. Through CUDA, developers can use general-purpose programming languages like C++ or Python to directly invoke the powerful parallel processing capabilities of GPUs for various computational tasks, not just graphics rendering.

CUDA-X AI Libraries

CUDA-X AI is the collective name for a series of highly optimized, GPU-accelerated libraries provided by NVIDIA for the AI and data science domains. This collection includes numerous specialized libraries such as cuDNN, NCCL, and cuBLAS, providing developers with pre-optimized software modules needed to build high-performance AI applications.

NVIDIA cuDNN (CUDA Deep Neural Network library)

NVIDIA cuDNN is a GPU library specifically for accelerating fundamental deep learning operations. It provides highly optimized implementations for common neural network operations like convolution, pooling, normalization, and activation functions, serving as the cornerstone for the high performance of mainstream deep learning frameworks like PyTorch and TensorFlow.

NVIDIA NCCL (NVIDIA Collective Communications Library)

NVIDIA NCCL is a communication library designed for multi-GPU and multi-node distributed training, hailed as the "lifeblood" of distributed training. It implements efficient collective communication operations such as All-Reduce, ensuring that GPUs can synchronize data and gradients with extremely high efficiency during large-scale model training.

NVIDIA DALI (Data Loading Library)

NVIDIA DALI is a library for accelerating the data loading and preprocessing pipeline in AI training. By offloading compute-intensive tasks like image decoding and data augmentation from the CPU to the GPU, it effectively resolves the CPU bottleneck in the data preparation stage, thereby improving end-to-end training efficiency.

NVIDIA RAPIDS

NVIDIA RAPIDS is a suite of open-source software libraries designed to migrate the entire data science and analytics workflow—from data preparation to machine learning—to run entirely on the GPU. By leveraging the parallel computing power of the GPU, RAPIDS can accelerate data analysis tasks traditionally handled by the CPU by several to tens of times.

cuDF

cuDF is one of the core components of the RAPIDS suite and can be seen as the GPU-accelerated version of the famous data analysis library, Pandas. It provides an API almost identical to Pandas, allowing data scientists to seamlessly migrate their data processing code to the GPU to achieve significant performance gains.

cuML

cuML is the machine learning library within the RAPIDS suite, positioned as the GPU-accelerated version of the popular machine learning library, Scikit-learn. It implements a variety of classic machine learning algorithms, enabling developers to train traditional models at high speed on the GPU, greatly shortening the time for model iteration and experimentation.

NVIDIA cuBLAS / cuSPARSE

These two libraries are fundamental components in NVIDIA CUDA for accelerating linear algebra operations and are the mathematical computation core for all higher-level AI frameworks. cuBLAS focuses on accelerating dense matrix and vector operations, while cuSPARSE is specialized in optimizing computations for sparse matrices. Together, they form the bedrock of GPU high-performance computing.

LLM/GenAI-Specific Software Stack

NVIDIA NeMo Framework

NVIDIA NeMo is an end-to-end, cloud-native framework designed specifically for building, customizing, and deploying generative AI models. It provides a comprehensive toolchain covering the entire model lifecycle—from data processing and large-scale distributed training to model fine-tuning and efficient inference—greatly simplifying the development process for enterprise-grade GenAI applications.

NVIDIA Megatron-LM

NVIDIA Megatron-LM is a pioneering framework specifically for training Transformer models with enormous parameter counts. By implementing key distributed techniques like tensor parallelism and pipeline parallelism, it successfully overcomes the challenge of a single GPU being unable to accommodate massive models, making it the core engine for training models at the hundred-billion or even trillion-parameter scale.

NVIDIA NeMo Megatron

This is the official integration of the NVIDIA NeMo framework with the Megatron-LM training engine, and it is NVIDIA's current flagship solution for large language model training. It combines the powerful distributed training capabilities of Megatron-LM with the engineering-friendly usability of the NeMo framework, offering developers an LLM development platform that is both high-performance and easy to manage.

NVIDIA FasterTransformer

NVIDIA FasterTransformer is a high-performance C++/CUDA library for Transformer model inference, which achieves minimal inference latency through deep optimization of CUDA kernels. Although it was very successful, many of its core optimization techniques have now been integrated into the more comprehensive TensorRT-LLM library, which is NVIDIA's current primary inference solution.

NVIDIA TensorRT-LLM

NVIDIA TensorRT-LLM is an open-source library designed to accelerate and optimize large language model inference, and it is the premier tool for LLM inference performance optimization today. It integrates cutting-edge techniques such as in-flight batching, paged attention, and INT4/INT8 quantization, which can significantly increase the throughput of LLM services, reduce latency, and decrease memory footprint.

CUDA Graph

CUDA Graph is a core performance optimization feature of the CUDA platform that allows a series of GPU operations to be "recorded" and then "replayed" as a single unit. For LLM inference tasks with a fixed computation flow, this mechanism can eliminate the overhead of the CPU dispatching tasks one by one, thereby significantly reducing latency and improving final performance.

NeMo Guardrails

NeMo Guardrails is an open-source toolkit for establishing programmable "safety guardrails" for applications based on large language models. It helps developers precisely control the model's conversational behavior, ensuring the topical relevance, safety, and factual accuracy of the output, effectively preventing the model from generating inappropriate or off-topic responses.

Inference & Deployment

NVIDIA TensorRT

NVIDIA TensorRT is a software development kit (SDK) for high-performance deep learning inference. Its core function is to deeply optimize already-trained models. Through techniques like layer fusion, precision quantization, and kernel auto-tuning, it creates a runtime engine for the model that runs fastest and most efficiently on a specific GPU, making it a key tool for achieving ultimate inference performance.

NVIDIA Triton Inference Server

NVIDIA Triton Inference Server is an open-source inference serving software designed for deploying AI models at scale and reliably in production environments. It supports models from various frameworks like TensorRT and PyTorch, and provides enterprise-grade features such as dynamic batching and model version management, making it a core component for building standardized, high-throughput AI inference services.

Development, Profiling & Management ToolsNVIDIA Nsight Suite
NVIDIA Nsight is a powerful suite of developer tools designed to help developers debug, analyze, and optimize applications running on NVIDIA GPUs. It includes

several professional tools, such as Nsight Systems and Nsight Compute, for performance profiling at the system-wide and individual kernel levels, respectively.

Nsight Systems

Nsight Systems is the system-level performance analysis tool in the Nsight suite, focusing on capturing and visualizing the interactions between the CPU, GPU, operating system, and network. Its main goal is to help developers identify system-level bottlenecks, such as CPU waiting for GPU or data transfer delays, to enable macro-level optimization.

Nsight Compute

Nsight Compute is the tool in the Nsight suite for in-depth CUDA kernel profiling. It provides developers with extremely detailed feedback and analysis on GPU kernel execution. With this tool, developers can deeply inspect micro-level metrics of a single kernel, such as memory access patterns and compute unit utilization, to perform code-level, ultimate optimization.

NVIDIA System Management Interface (nvidia-smi)

nvidia-smi is a fundamental and widely used command-line tool for real-time monitoring and management of NVIDIA GPU devices in a system. It can quickly query key status information such as GPU utilization, memory usage, power consumption, temperature, and running processes, making it an essential tool for every GPU user and administrator.

NVIDIA DCGM (Data Center GPU Manager)

NVIDIA DCGM can be considered the enterprise-grade and cluster-level version of nvidia-smi. It is a suite of GPU management and monitoring tools designed specifically for data center environments. Compared to nvidia-smi, DCGM offers richer performance metrics, proactive health diagnostics, and integration with cluster management systems, making it suitable for the operation and maintenance of large-scale GPU clusters.

NVIDIA Fabric Manager

NVIDIA Fabric Manager is a software tool specifically for configuring, monitoring, and managing the NVLink high-speed interconnect fabric in multi-GPU systems. It ensures that the NVLink network (Fabric) is correctly initialized at system startup and remains healthy, which is crucial for multi-GPU nodes like DGX to achieve optimal communication performance.

NVIDIA Base Command Manager

NVIDIA Base Command Manager is a software platform for simplifying the lifecycle management of AI infrastructure clusters. It can automate a series of complex operational tasks, from OS deployment and software stack installation/updates on bare-metal servers to the health monitoring and management of the entire cluster.

Platforms, Applications & Services

NVIDIA AI Enterprise

NVIDIA AI Enterprise is an enterprise-certified and supported end-to-end AI software platform that bundles many key software components from the NVIDIA ecosystem (such as NeMo, TensorRT, Triton, etc.). The platform greatly simplifies the complexity for enterprises to deploy, manage, and scale AI workloads in private, public, or hybrid cloud environments, while providing stability and security guarantees.

NVIDIA NGC Catalog

The NVIDIA NGC Catalog is an official software and model hub, which can be thought of as an "app store" for the AI and HPC fields. From it, developers can easily obtain GPU-optimized software containers, pre-trained models, Helm charts, and industry SDKs, thereby significantly shortening development cycles and rapidly building high-performance applications.

NVIDIA DGX Cloud

NVIDIA DGX Cloud is an AI supercomputing cloud service that allows enterprises to directly rent the top-tier computing power of NVIDIA DGX clusters without bearing the huge cost and complexity of building and maintaining their own data centers. This service is launched in partnership with major cloud service providers and offers convenient, elastic infrastructure for large-scale AI training and inference tasks.

NVIDIA AI Foundation Models

NVIDIA AI Foundation Models are a series of high-quality, pre-trained generative AI models provided by NVIDIA, covering multiple domains such as language and vision. Enterprises can directly obtain these models from NGC for inference tasks or, more commonly, use them as a powerful foundation to quickly build customized models that meet their specific business needs through fine-tuning.

NVIDIA Riva

NVIDIA Riva is a GPU-accelerated software development kit (SDK) specifically for building real-time, multimodal conversational AI applications. It provides a full suite of high-performance APIs covering functions like Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and Text-to-Speech (TTS), enabling developers to easily create fluent and natural voice AI assistants and services.

Ecosystem Partners & Key Open Source Technologies

PyTorch

PyTorch is an open-source deep learning framework led by Meta AI, which has become the de facto standard in the AI research and development field due to its flexibility and ease of use. It is deeply integrated with NVIDIA's CUDA platform, can seamlessly leverage acceleration libraries like cuDNN, and the latest hardware features from NVIDIA (such as FP8 mixed precision) are often first supported in PyTorch.

TensorFlow / JAX

These two frameworks, developed by Google, play key roles in the AI ecosystem: TensorFlow, as a mature end-to-end platform, is widely deployed in enterprise production environments; JAX, with its high performance and functional programming paradigm, is highly favored in cutting-edge research. Both rely on NVIDIA GPUs as the cornerstone of their high-performance computing and depend heavily on CUDA for acceleration.

FlashAttention

FlashAttention is a revolutionary, efficient attention algorithm that optimizes GPU memory read/write patterns by restructuring the computation process, avoiding the instantiation of a huge attention matrix in memory. This technology significantly reduces the memory footprint and greatly increases the computation speed for longsequence Transformer models during training and inference, and has now become a standard in the LLM field.

vLLM

vLLM is a popular open-source library designed specifically for achieving extreme LLM inference throughput. Its core innovation, the PagedAttention algorithm, borrows the concept of virtual memory from operating systems to manage the KV cache, effectively solving the memory fragmentation problem. This leads to higher GPU utilization and more flexible batching, making it one of the leading optimization solutions for deploying LLM services today.

Kubernetes

Kubernetes is the industry-recognized open-source standard for container orchestration. In the AI domain, it is widely used to automate the deployment, scaling, and management of AI workloads on large-scale GPU clusters. NVIDIA, by providing key tools like the GPU Operator, enables Kubernetes to natively recognize and efficiently schedule GPU resources, thus seamlessly extending powerful container management capabilities to AI infrastructure.

InfiniBand

InfiniBand is a computer network communication standard that provides extremely high bandwidth and ultra-low latency. It serves as the central nervous system for building large-scale AI supercomputers like the NVIDIA DGX SuperPOD. It is responsible for high-speed data exchange among hundreds or thousands of GPU nodes, making it a key technology for ensuring that distributed training tasks can be efficiently parallelized and overcome communication bottlenecks.