BitOoda

BitOoda AI Research:
Understanding HPC Hardware for Modern AI Computing

Abstract

This report examines the pivotal role of fast intranode and internode communication in High-Performance Computing (HPC) applications, particularly in AI training and inference. We explore how NVIDIA has capitalized on the post-GPT-3 AI and HPC surge by introducing specialized tools tailored to meet these demands. Additionally, we delve into existing solutions for efficient GPU communication within and between servers, while also highlighting the future potential of NVIDIA's upcoming Blackwell architecture.

Introduction

High-Performance Computing (HPC) hardware serves as the foundation for today's advanced artificial intelligence (AI) and machine learning (ML) applications. As AI models grow more complex, a deep understanding of the underlying hardware becomes crucial. This report explores the essential components of HPC hardware with a focus on AI training and inference tasks, emphasizing the importance of large GPU clusters and fast communication between GPUs for achieving optimal performance. By dissecting these key elements, we aim to offer insights into the current state and future trends of AI computing architectures, supporting more informed decisions in AI infrastructure development and deployment.

Research

Tim Kelly, CEO & Founder
Dhyay Bhatt, Head of AI
Niraj Yagnik, Lead Developer & AI Engineer
David Bellman, Head of Power

Key Takeaways

- **Specialized AI Hardware**: NVIDIA capitalized on the AI boom by building specialized AI chips like the H100 and H200.
- **Efficient GPU Communication**: Technologies like NVLink improve communication between GPUs, enhancing scalability and reducing latency.
- **Future AI Advancements**: NVIDIA's Blackwell architecture will bring even more power, efficiency, and scalability to AI computing.

Why is NVIDIA Dominating the AI Hardware Space

Superior AI Focused Chips

NVIDIA timed the market just right by building GPUs that specialize in AI, along with the right software stack. Below are details on the chips NVIDIA used to capitalize on the AI Hype after GPT 3:

NVIDIA H100 GPU

The NVIDIA H100, based on the Hopper architecture, is engineered specifically for AI workloads.

- **Advanced Tensor Cores:** These accelerate matrix operations essential for training and inference.
- **Transformer Engine:** Optimizes the performance of transformer models used in natural language processing.
- **FP8 Precision:** Introduces 8-bit floating-point precision for faster computations without significant loss in accuracy.
- **High Memory Bandwidth:** Enables rapid data access, crucial for large-scale AI tasks.

NVIDIA H200 GPU

The H200 aims to fix some of the H100's shortcomings. Key improvements include:

- **More Memory:** Think of this as the GPU's brain capacity. The H200 has almost twice as much memory as the H100, allowing it to handle larger and more complex AI tasks.
- **Faster Memory Bandwidth:** The H200 can move data around much quicker than the H100.
- **Better Optimized:** Performs better in the real world, and is more scalable and optimized for AI tasks.

Definitions and Background
Understanding AI Training & Inference

What Is Training:

Training is the process of teaching a machine learning model to make accurate predictions by adjusting its internal parameters based on input data. This involves computationally intensive operations, especially with large datasets and complex models like deep neural networks.

Why Training Requires Large GPU Clusters:

- **Parallel Processing:** Training large models requires significant computational resources. Multiple GPUs can process data in parallel, drastically reducing training time.
- **Memory Capacity:** Complex models often exceed the memory capacity of a single GPU. Distributing the model across multiple GPUs allows for training larger models.
- **Scalability:** Using multiple GPUs enables scaling up computational power to meet the demands of ever-growing datasets and model sizes.

What Is Inference:

Inference is the phase where a trained model is used to make predictions on new, unseen data. It requires the model to process input data and generate outputs, typically with a focus on low latency and high throughput for real-time applications.

Why Inference Requires Large GPU Clusters:

- **High Throughput:** For large-scale inference like real-time recommendations or massive dataset processing, multiple GPUs enable parallel processing of requests, boosting system throughput and response times.
- **Complex Model Execution:** Large models such as LLMs may exceed the memory capacity of a single GPU. Distributing the model across multiple GPUs helps manage memory and ensures efficient inference.
- **Latency Reduction**: Time-critical applications like autonomous driving or financial trading demand ultra-low latency. Multiple GPUs reduce the time to make predictions by sharing computational tasks.

GPU Communication
The Bottleneck

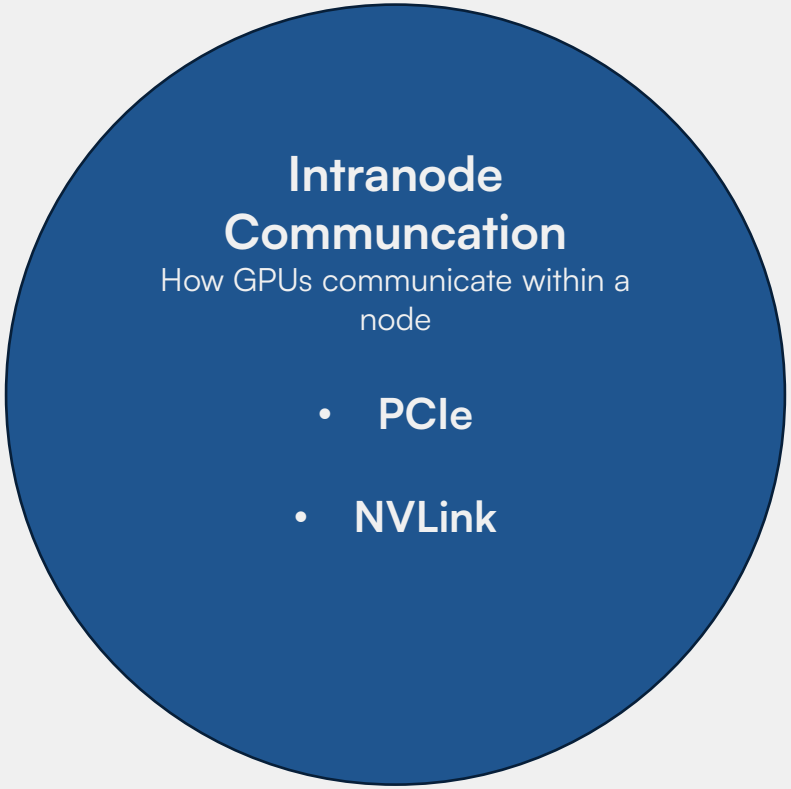
What is a node?

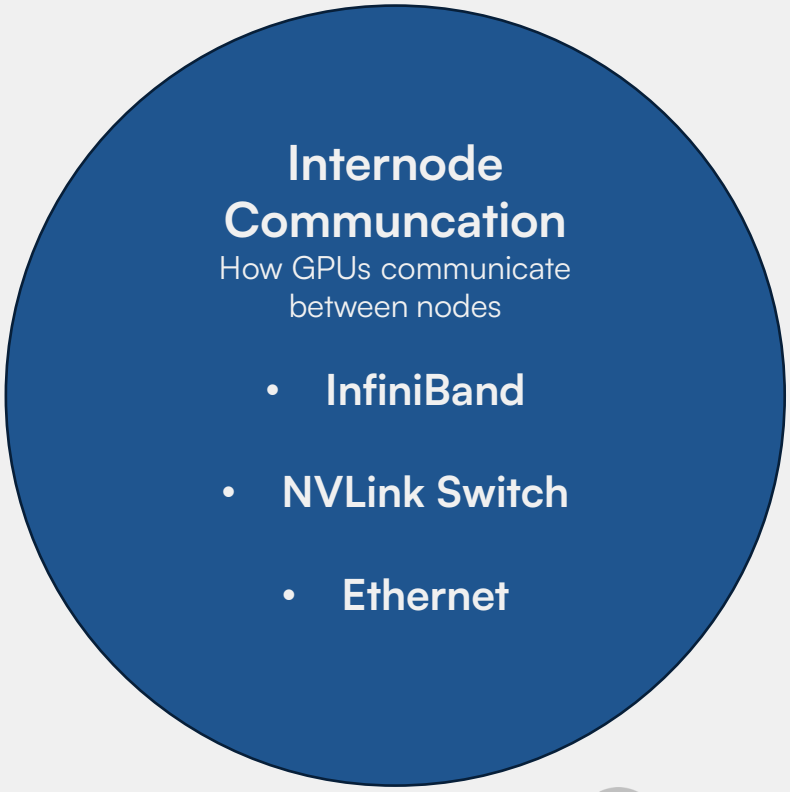
Node: A node is a server typically containing 8 GPUs used for processing tasks in a distributed system. Its significance is in enabling parallel GPU computations, allowing for faster processing.

Cluster: A cluster is a collection of GPU-equipped nodes that work together to process large datasets or models. Its significance lies in scaling GPU power, enabling faster training, inference, and handling of computationally intensive tasks across multiple GPUs.

Why do GPUs need to communicate?

- During training, GPUs collaborate to process large datasets and complex models faster by distributing data and computations, requiring communication to synchronize updates and share model parameters.
- During inference, GPUs reduce latency and scale real-time predictions by distributing tasks and ensuring efficient communication, especially for large models and high-demand scenarios.





GPU Communication
Deeper Dive

Intranode

Technology	Description
PCIe	Standard interface connecting GPUs and other components to the motherboard. Enables communication between these components, CPU, and system memory.
NVLink (Connection within a Node)	A high-bandwidth, low-latency interconnect developed by NVIDIA for direct GPU-to-GPU communication. Significantly improves performance in multi-GPU systems compared to PCIe.

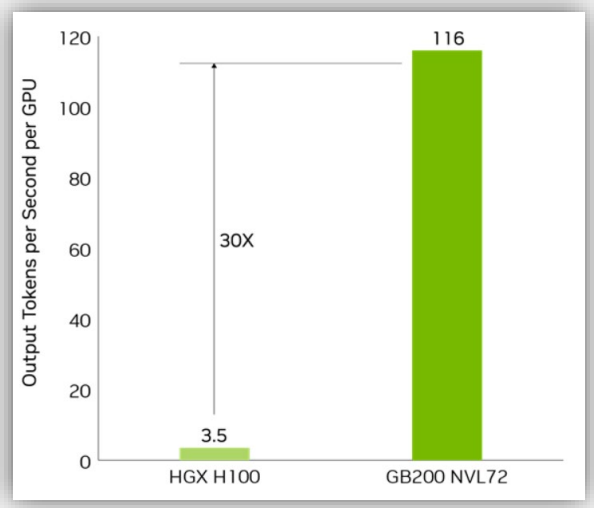
Internode

Technology	Definition
InfiniBand	High-speed, low-latency networking technology for high-performance computing clusters.
Ethernet	Similar to InfiniBand, but generally slower and more common in smaller clusters. More common in smaller clusters or when budget is a concern.
NVSwitch (Extension of NVLink within or between Racks)	Enables all-to-all GPU communication within a server at full NVLink speed. Creates unified memory architecture, allowing any GPU to access any other GPU's memory directly and quickly.

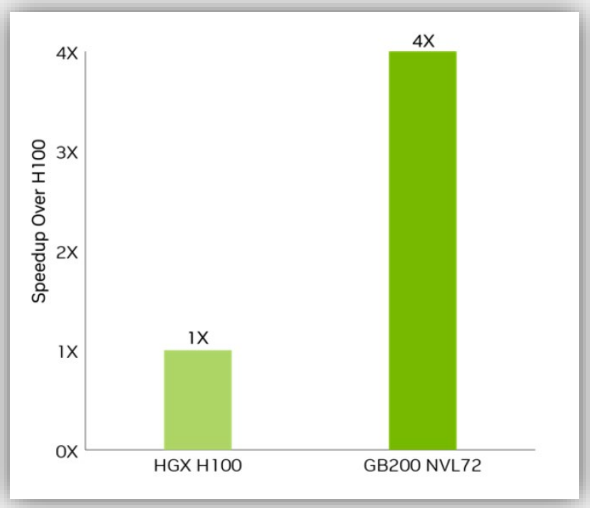
What's Next
NVIDIA Blackwell Architecture

The NVIDIA Blackwell architecture, which succeeds Hopper, is designed to revolutionize AI and high-performance computing. Pricing and availability information is expected soon.

- **Advanced GPU Design**: The GPU features a unique dual-chip design and specialized engines that boost performance for AI and scientific computing, while improving system maintenance and data security.
- **Fifth-Generation Tensor Cores:** Supercharge AI computing by using new ways to handle numbers (FP4 and FP6), which allow AI systems to work faster and use less power.
- **NVLink 5.0:** Enables faster GPU-to-GPU communication, allowing up to 576 GPUs to work in unison for complex AI workloads.
- **Energy Efficiency:** Reduces cost and energy consumption for LLM inference workloads by up to 25x compared to Hopper3. Emphasizes power efficiency, crucial for data centers running AI workloads continuously.



Projected Improvement in Inference performance
Source: NVIDIA



Projected Improvement in Inference performance
Source: NVIDIA

Disclosures

Purpose

This research is only for the clients of BitOoda. This research is not intended to constitute an offer, solicitation, or invitation for any securities and may not be distributed into jurisdictions where it is unlawful to do so. For additional disclosures and information, please contact a BitOoda representative at info@bitooda.io.

Analyst Certification

Niraj Yagnik, the primary research analyst of this report, hereby certifies that all of the views expressed in this report accurately reflect his personal views, which have not been influenced by considerations of the firm's business or client relationships.

Conflicts of Interest

This research contains the views, opinions, and recommendations of BitOoda. This report is intended for research and educational purposes only. We are not compensated in any way based upon any specific view or recommendation.

General Disclosures

Any information ("Information") provided by BitOoda Holdings, Inc., BitOoda Digital, LLC, BitOoda Technologies, LLC or Ooda Commodities, LLC and its affiliated or related companies (collectively, "BitOoda"), either in this publication or document, in any other communication, or on or through http://www.bitooda.io/, including any information regarding proposed transactions or trading strategies, is for informational purposes only and is provided without charge. BitOoda is not and does not act as a fiduciary or adviser, or in any similar capacity, in providing the Information, and the Information may not be relied upon as investment, financial, legal, tax, regulatory, or any other type of advice. The Information is being distributed as part of BitOoda's sales and marketing efforts as an introducing broker and is incidental to its business as such. BitOoda seeks to earn execution fees when its clients execute transactions using its brokerage services. BitOoda makes no representations or warranties (express or implied) regarding, nor shall it have any responsibility or liability for the accuracy, adequacy, timeliness or completeness of, the Information, and no representation is made or is to be implied that the Information will remain unchanged. BitOoda undertakes no duty to amend, correct, update, or otherwise supplement the Information.

The Information has not been prepared or tailored to address, and may not be suitable or appropriate for the particular financial needs, circumstances or requirements of any person, and it should not be the basis for making any investment or transaction decision. The Information is not a recommendation to engage in any transaction. The digital asset industry is subject to a range of inherent risks, including but not limited to: price volatility, limited liquidity, limited and incomplete information regarding certain instruments, products, or digital assets, and a still emerging and evolving regulatory environment. The past performance of any instruments, products or digital assets addressed in the Information is not a guide to future performance, nor is it a reliable indicator of future results or performance.

All derivatives brokerage is conducted by Ooda Commodities, LLC a member of NFA and subject to NFA's regulatory oversight and examinations. However, you should be aware that NFA does not have regulatory oversight authority over underlying or spot virtual currency products or transactions or virtual currency exchanges, custodians or markets.

BitOoda Technologies, LLC is a member of FINRA.

"BitOoda", "BitOoda Difficulty", "BitOoda Hash", "BitOoda Compute", and the BitOoda logo are trademarks of BitOoda Holdings, Inc.