AMD AI Networking Direction and Strategy

December 2023

Abstract

AMD is committed to developing Ethernet-based AI networking for GPU nodes, pods, and cluster connections. This involves a fusion of internally-developed products and collaborations with leading Ethernet switch vendors, a strategy that empowers AMD to introduce groundbreaking technology tailored to the distinct challenges posed by AI networks, all while capitalizing on the openness and widespread adoption of Ethernet technology and its ecosystem.

"There is no artificial intelligence without an intelligent network"

At the heart of the AMD AI Network solution lies the AMD P4 programmable NIC. Built upon cutting-edge AMD technologies in high-performance DMA, P4, programmable transport protocols, monitoring, and telemetry, the NIC delivers exceptional performance and ultra-low latency for data transfer across a large number of GPUs.

With its open Ethernet interface, the AMD AI Network solution gives enterprises the freedom to scale up GPU nodes and expand GPU pods/clusters to meet the explosive demand for AI/ML models and parallelism, all without mandating the purchase of proprietary switches and networks. Customers can leverage their decades-long Ethernet expertise to design, create and manage high-performance Ethernet networks that align with their specific requirements.

This remarkable achievement is made possible by continuous improvements in Ethernet technology, including latency reduction, rapid port speed iteration, increased Ethernet silicon capacity, and enhanced intelligence at the edge (DPU, RDMA NIC) and within the network for congestion management, load balancing, and telemetry.

AMD remains committed to open, standards-based AI networking. As a founding member of the Ultra Ethernet Consortium (UEC), AMD has contributed technologies such as congestion management and RDMA transport to the standardization body. As UEC standards evolve, the AMD programmable NIC simplifies compliance with future transport standards. This approach fosters a dynamic environment for AMD, its technology partners, and end users to innovate at an accelerated pace.

Overview

AMD is committed to developing AI-specific high-performance, high-bandwidth, low-latency, programmable transport networks, leveraging standard Ethernet and open ecosystem solutions. These networks are essential for AI / ML workloads, which are very data intensive and require high-speed, low-latency data transfer.

The Need for AI-Specific Networking

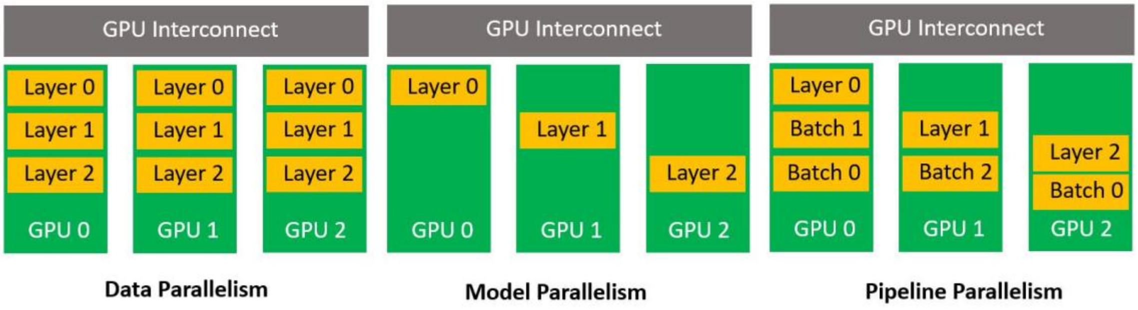
The rapid development of large language and deep learning recommendation models (LLM and DLRM) has led to the creation of increasingly complex and demanding AI networks. These models often have trillions of parameters and require massive amounts of data to train and run. This has put a significant strain on existing networking infrastructure, which is not designed for the bulk synchronous parallel nature of AI workloads.

To address this challenge, a new type of network is needed specifically for AI. This network must be able to support three types of parallelism that are essential for AI training and inference.

1. Data Parallelism
   This involves training or running multiple copies of the same model on different nodes at the same time.

2. Model Parallelism
   If the size of the model is too large, then divide the model into smaller parts and training or running them on different nodes at the same time. Examples include Transformers, GPT-3, and BERT.

3. Pipeline Parallelism
   This involves breaking down the training or inference process into smaller steps and executing them on different nodes at the same time.



In addition to model parallelism, there are different phases of model development:

1. Training
2. Inference

Beyond supporting these types of parallelism and phases, the AI network must also have the following characteristics:

- High bandwidth: AI models require massive amounts of data to train and run. The AI network must be able to provide the high-bandwidth necessary to support this data transfer.

- Low Latency: Network must have low latency, to allow model and data tasks to run smoothly,and minimize the introduction of processing delays.

- Scalability: AI models are becoming increasingly complex and demanding. The AI network must be scalable to support the growing needs of AI workloads.

- Reliability: AI training workloads are often time-consuming and expensive, so it is important to use a networking solution that is reliable and can minimize the risk of disruptions.

AI networks are typically divided into two main types:

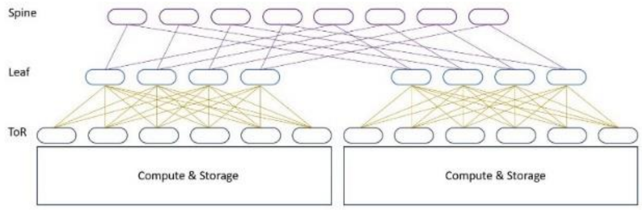
Front-end network: This network connects to the front-end compute and storage servers, which typically handles data ingestion and supports many users and devices that are accessing AI services. They carry two types of traffic [north/south (NS)] traffic to and from the external world (internet or other data centers) and [east/west (EW)] traffic from network endpoints within the same data center. Both have different requirements.

Back-end network: This network connects AI nodes to each other for distributed computing. Back-end network need to be high performance and low latency to support high-speed communication requirements of AI workloads. There are machines typically called GPU *nodes* (containing one or more CPUs and GPUs). GPU nodes have multiple network interfaces on both networks. Today, GPU nodes have one RDMA NIC per GPU. Collection of nodes form *pods* (group of GPUs that are connected to the same node and can be used for parallel processing). A collection of interconnected nodes that work together to provide accelerated computing power for specific computational tasks form a *cluster*.

Ethernet-Based AI Networking

Ethernet technology is based on an extensive track record of innovation in security, reliability, and performance, and continues to be the networking option of choice for cloud computing and its associated needs. Today, this is leading companies such as Broadcom to call Ethernet "the de facto AI network." Ethernet switch port speeds have increased dramatically, with 400 Gb/s and 800 Gb/s switches now available with ultra-low latency and full support for PAUSE and ECN marking. This makes it ideal to address both the bandwidth and latency requirements of large AI clusters. In addition, increase in silicon capacity has made it possible for Ethernet switches to support 25.6 Tbps, 51.2 Tbps and even 102.4 Tbps; such growth is necessary to handle the large amount of traffic generated by AI clusters.

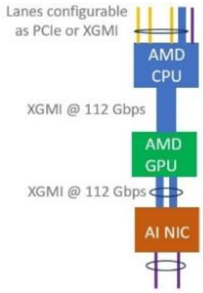
Clos networks are a type of non-blocking network that can provide high bandwidth and low latency between any two nodes in the network. Modern Ethernet networks based on IP routing leverage Clos network design to create large and scalable AI clusters.



One of the key advantages of using Ethernet for AI is that it is a standards-based technology. This means that Ethernet switches from different vendors can interoperate with each other, which makes it easier to deploy and manage a large-scale AI network. However, due to the bulk synchronous parallel nature of AI workloads in both the training and inferencing phases, modern IP networks need to evolve to deal with the problems of ECMP load balancing collisions, congestion management for lossless RDMA, and deadlock/livelock. There are many proposals on how to solve these issues, starting with RDMA over IP (ROCEv2), flow control (802.3 PAUSE), and congestion control (DCQCN).

AMD AI Networking Solution

An AI / ML networking strategy must consider multiple factors, including types of workloads connected on the network, volume and type of data that will be transferred on the network, performance requirements, network scalability, and of course, budget available for the network. The programmable NIC will be built with plug-and-play configurability in mind, and to help improve the overall performance of AI / ML networks by providing high throughput, low latency, high connection scalability and reduced CPU overhead. It will support programmable congestion control, dynamic load balancing, inline security, dynamic connection and many more features.



AMD Infinity Fabric™ (IF) is a high-speed intra-host interconnect that can be used to connect multiple AMD CPUs and GPUs. For scale-up, AMD Infinity Fabric will use a global memory pool for inter-GPU communication. This gives massive bandwidth with smaller domains to run model parallelism traffic. For scale-out, The AMD NIC will support multiple modes that can be used to connect AMD Infinity Fabric nodes and clusters together over Ethernet to build large domains, which can improve the performance of AI / ML training networks that requires both large data and pipeline parallelism with AMD assets.

The AMD ROCm™ Collective Communication Library (RCCL) uses a reliable multicast protocol called Group Multicast Library (GMLC) to ensure that all nodes in the collective communication group receive the same data. GMLC is a tree-based protocol that uses various techniques to help ensure reliability such as error detection, correction, and retransmission. In-network computation, performing AllReduce operations on the NIC, offloads the computation from the CPU, which can improve performance and reduce latency. The AMD Pensando Software-in-Silicon Development Kit (SSDK), in conjunction with RCCL innovation, helps NIC atomic operations and performs the AllReduce operation in-place on the data buffer. The combination of these techniques allows AI deployments to achieve significant improvements for AllReduce operations in AI training.

The AMD programmable AI NIC will support a standards-based congestion control implementation and allow interoperability with other vendors to allow tuning of adaptive primitives based on user needs. The AMD NIC will have built-in capabilities to support real-time telemetry export.

Summary

The AMD AI networking strategy is centered on building a powerful and flexible infrastructure solution, featuring an AI NIC, which offers several benefits for AI / ML workloads and AI training networks. AI-specific networking can improve throughput, provide low latency, and reduce CPU overhead. The AMD NIC plays a pivotal role by acting as the gateway to a programmable transport network for AI / ML workloads. A programmable NIC provides a scale-out architecture solution option, gives users the ability to interoperate with existing Ethernet solutions, and offers high performance, scalability, compatibility, security, and openness, making it ideal for implementing AI / ML networks.

together we advance_networking

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

THIS INFORMATION IS PROVIDED 'AS IS." AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

amd.com/pensando

PSB23001