

AI Networking Solution: Chelsio T7 DPU and S7/T6 SmartNICs

Introduction

As Artificial Intelligence (AI) continues to revolutionize industries, the demand for high-performance and low-latency networking solutions has increased significantly. Chelsio Communications is leading this transformation with its advanced T7 DPU and S7/T6 SmartNIC products. These solutions are designed to accelerate AI workloads, optimize data center operations, and enable high-bandwidth, low-latency communication using 400Gb Ethernet with RDMA, supporting scalable AI training, inference, and HPC environments.

Why 400Gb RDMA for AI Infrastructure

Modern AI training clusters depend on fast, efficient data exchange between GPUs and storage systems. As model sizes grow, traditional networking approaches create bottlenecks in distributed training environments.

The Chelsio T7 platform addresses these challenges with native 400Gb Ethernet and full RDMA support (iWARP and RoCEv2), enabling:

- High-bandwidth GPU-to-GPU and GPU-to-storage communication
- Ultra-low latency for synchronized distributed training
- Reduced CPU overhead through kernel bypass and zero-copy data transfer
- Scalable, standards-based Ethernet fabrics as an alternative to proprietary AI interconnects

Ethernet at 400Gb with RDMA is emerging as a scalable alternative to proprietary AI interconnects for next-generation AI infrastructure.

Features and Benefits of Chelsio T7 DPU and S7/T6 SmartNICs

The Chelsio T7 DPU and S7/T6 SmartNICs have the following features and benefits:

High-Performance Networking

- **Ethernet Connectivity:** T7 DPU supports quad-port 1/10/25/50/100Gb, dual-port 40/100/200Gb, or single-port 400Gb Ethernet. The S7 SmartNIC supports quad-port 1/10/25/50/100Gb or dual-port 40/100/200Gb Ethernet, while the T6 SmartNIC supports dual-port 1/10/25/40/50/100Gb Ethernet. These configurations offer unparalleled data transfer speeds and flexibility.
- **400Gb RDMA Acceleration:** Native iWARP and RoCEv2 support at up to 400Gb Ethernet enables ultra-low latency, high-bandwidth communication for AI training, HPC, and disaggregated storage environments.
- **Full Protocol Offload:** Offloads TCP/IP, iSCSI, NVMe-oF/TCP, and RDMA, reducing host CPU utilization and enabling efficient GPU and storage resource usage.

Scalability and Flexibility

- **PCIe and Memory Options:** The T7 DPU features a PCIe Gen5 interface for high-speed data transfer, integrated PCIe switch and bridge capabilities, and card memory and memory-free operation support. The S7 SmartNIC supports memory-free operation using host memory, offering cost-effective scalability. The T6 SmartNIC supports a unified wire approach for LAN, SAN, and cluster applications, simplifying network infrastructure and reducing costs.

Enhanced Security and Reliability

- **Advanced Encryption:** Both the T7 DPU and S7/T6 SmartNICs provide robust security features, including AES 128/256, SHA1/SHA2 offload, TLS/SSL, and IPsec support, ensuring secure data transmission.
- **Data Integrity and Protection:** The T7 DPU includes integrated features such as deduplication, compression, RAID, and erasure coding support, ensuring high data integrity and reliability.

Use Cases in AI Training and Inference

Use Case 1: Accelerating the AI Training

Training the AI models involves processing massive datasets, requiring significant computational resources, and efficient data handling. The T7 DPU and S7/T6 SmartNICs significantly enhance AI training environments with the following features:

- **High-Bandwidth RDMA Data Movement:** 400Gb RDMA enables efficient GPU-to-GPU and GPU-to-storage communication, eliminating network bottlenecks in distributed training environments.
- **CPU Bypass and Zero-Copy:** RDMA and full protocol offload eliminate unnecessary data copies and CPU intervention, accelerating training cycles and improving system efficiency.
- **Scalability:** The compatibility with T7, S7, and T6 silicon software enables seamless scalability, supporting the growth of data and model complexity without requiring significant infrastructure changes.

Use Case 2: Enhancing AI Inference

AI inference requires processing real-time data to provide immediate responses, making low latency and high reliability critical. The T7 DPU and S7/T6 SmartNICs offer the following benefits for an AI inference:

- **Ultra-Low Latency Networking:** RDMA-enabled communication ensures deterministic low-latency performance for real-time inference and microservices-based AI deployments.
- **Data Security and Integrity:** Robust security features protect sensitive data during AI inference, including TLS/IPsec offloading, and data-at-rest encryption.
- **Cost-Effective Scalability:** The memory-free operation reduces overall cost and allows more inference nodes to be deployed without significantly increasing expenses. This scalability supports real-time analytics across various applications.

Conclusion

The Chelsio T7 DPU and S7/T6 SmartNICs provide a scalable, high-performance Ethernet-based AI interconnect, combining 400Gb RDMA, full protocol offload, and programmable acceleration. This enables organizations to build efficient, standards-based AI infrastructure for training and inference while avoiding the complexity and cost of proprietary networking solutions.

Join Us

Explore how the Chelsio AI networking solutions can transform your data center infrastructure. Contact us today to learn more about our T7 DPU and S7/T6 SmartNIC products and schedule a consultation with our experts.

Related Links

[T7 Product Brief](#)

[S7 Product Brief](#)