Engineering the Future of AI Datacenters with NVIDIA and Enfabrica's Architectural Innovations and Strategic Integration

Preprint · October 2025

DOI: 10.13140/RG.2.2.10436.28804

CITATIONS
0

READS
63

1 author:

Sanjay Basu
Oracle Corporation
9 PUBLICATIONS **8** CITATIONS

SEE PROFILE

Engineering the Future of AI Datacenters with NVIDIA and Enfabrica's Architectural Innovations and Strategic Integration

Author(s)

Dr. Sanjay Basu, Oracle Corporation

October 01, 2025

Executive Summary

NVIDIA's 2025 product evolution represents a paradigm shift in AI datacenter architecture, anchored by the revolutionary Blackwell Ultra and the specialized Rubin CPX architectures. The Blackwell Ultra architecture delivers 1.5x more AI compute FLOPS compared to standard Blackwell GPUs, featuring 5th-generation Tensor Cores with 2x attention-layer acceleration and breakthrough NVFP4 precision capabilities. Complementing this compute evolution, NVIDIA's Rubin CPX GPU is purpose-built to handle million-token coding and generative video applications, designed specifically for the compute-intensive context phase of disaggregated inference.

The strategic significance of NVIDIA's $900 million Enfabrica acquisition cannot be overstated. NVIDIA has shelled out over $900 million to hire Enfabrica CEO Rochan Sankar and other employees at the artificial intelligence hardware startup, and to license the company's technology. This acquisition represents more than talent acquisition—it's a strategic move to vertically integrate revolutionary networking and memory fabric technologies that address fundamental scaling bottlenecks in hyperscale AI infrastructure. Enfabrica says its technology can connect more than 100,000 GPUs together, offering integrated systems around chips so clusters can effectively serve as a single computer.

NVIDIA 2025 Architecture

Blackwell, Rubin, and Beyond

Blackwell Ultra Architecture

Advancing Tensor Computing

The NVIDIA Blackwell Ultra GPU builds on core innovations to accelerate training and AI reasoning, fusing silicon innovations with new levels of system-level integration. The architecture's cornerstone innovation lies in its 5th-generation Tensor Cores, which represent a fundamental advancement in mixed-precision acceleration capabilities.

Tensor Core Evolution and NVFP4 Precision

Blackwell Ultra is composed of two reticle-sized dies connected using NVIDIA High-Bandwidth Interface (NV-HBI), providing 10 TB/s of bandwidth. The 5th-generation Tensor Cores introduce several architectural enhancements:

- **Enhanced Tensor Memory Integration**: The new Tensor Cores are tightly integrated with 256 KB of Tensor Memory (TMEM) per SM, optimized to keep data close to the compute units
- **Dual-Thread-Block MMA**: They support dual-thread-block MMA, where paired SMs cooperate on a single MMA operation, sharing operands and reducing redundant memory traffic
- **NVFP4 Breakthrough**: The introduction of NVIDIA NVFP4, the new 4-bit floating-point format, combines two-level scaling—an FP8 (E4M3) micro-block scale applied to 16-value blocks plus a tensor-level FP32 scale—enabling hardware-accelerated quantization with markedly lower error rates than standard FP4

Performance Scaling and Efficiency Gains

The Blackwell Ultra dense NVFP4 compute capability provides substantial performance uplift over the original Blackwell GPU. While the base architecture delivers 10 petaFLOPS of NVFP4 performance, Ultra pushes that to 15 petaFLOPS—a 1.5x increase compared to Blackwell GPU and 7.5x increase from NVIDIA Hopper H100 and H200 GPUs. This performance boost directly benefits large-scale inference by enabling more concurrent model instances, faster response times, and lower costs per token generated.

Transformer Engine Advancements

The NVIDIA Blackwell Transformer Engine utilizes fine-grain scaling techniques called micro-tensor scaling, to optimize performance and accuracy enabling 4-bit floating point (FP4) AI. This doubles the performance and size of next-generation models that memory can support while maintaining high accuracy.

Rubin CPX Architecture

Specialized Context Processing

Purpose-Built for Massive Context Workloads

The NVIDIA Rubin CPX GPU is a purpose-built solution designed to deliver high-throughput performance for long-context AI workloads, featuring 30 petaFLOPs of NVFP4 compute and 3x attention acceleration compared to NVIDIA GB300 NVL72. Unlike traditional GPU architectures optimized for balanced compute and memory bandwidth, Rubin CPX employs a fundamentally different design philosophy.

Architectural Differentiation

Each Rubin CPX chip will be a monolithic SoC on a conventional flip chip BGA package. Instead of HBM, Rubin CPX will have 128GB of GDDR7 DRAM. Switching from using HBM to cheaper GDDR7 memory reduces the cost per GB by more than 50%. This design choice reflects a deep understanding of inference workload characteristics:

- **Context Phase Optimization**: Disaggregated inference enables the context and generation phases to be processed independently, allowing for targeted optimization of compute and memory resources
- **Monolithic Die Efficiency**: The monolithic die configuration represents a departure from the dual-GPU packages characteristic of NVIDIA's current Blackwell and Blackwell Ultra architectures
- **Integrated Video Processing**: The processor integrates four NVENC and four NVDEC video encoders directly on-chip, enabling streamlined multimedia workflows without external processing dependencies

Trillion-Parameter Model Support

The Rubin CPX addresses computational bottlenecks in extended-context scenarios where AI models process millions of tokens simultaneously. This capability proves critical for applications including comprehensive software codebase analysis and hour-long video content processing, which can require up to one million tokens.

Grace/Blackwell Server Platforms and Petaflop-Level Performance

System-Level Integration

The NVIDIA GB200 Grace Blackwell Superchip connects two NVIDIA B200 Tensor Core GPUs to the NVIDIA Grace CPU over a 900GB/s ultra-low-power NVLink chip-to-chip interconnect. This integration represents more than component aggregation—it's a holistic platform approach to AI computation.

Fifth-Generation NVLink Advancement

Fifth-Generation NVLink delivers groundbreaking 1.8TB/s bidirectional throughput per GPU, ensuring seamless high-speed communication among up to 576 GPUs for the most complex LLMs. This interconnect capability becomes critical for scaling to petaflop-level node performance across distributed workloads.

Rack-Scale System Architecture

The GB200 is a key component of the NVIDIA GB200 NVL72, a multi-node, liquid-cooled, rack-scale system for the most compute-intensive workloads. It combines 36 Grace Blackwell Superchips, which include 72 Blackwell GPUs and 36 Grace CPUs interconnected by fifth-generation NVLink.

The Bottleneck in AI Datacenters

Networking and Memory Fabric Challenges

Traditional Cluster Interconnect Limitations

Modern AI datacenter architectures face fundamental scaling challenges that traditional networking approaches cannot adequately address. The limitations manifest across multiple dimensions:

Ethernet and InfiniBand Scaling Constraints

Traditional Ethernet and InfiniBand topologies, while robust for conventional datacenter workloads, encounter significant bottlenecks when scaling to the 100,000+ GPU clusters required for foundation model training. These limitations include:

- **Bandwidth Aggregation Challenges**: Traditional NICs provide point-to-point connectivity that doesn't efficiently aggregate bandwidth across multiple network paths
- **Fault Tolerance Limitations**: Single points of failure in large-scale topologies can cascade, affecting entire cluster partitions
- **Memory Access Locality**: Rigid memory-to-compute binding creates resource stranding and limits dynamic allocation

NVLink/NVSwitch Topology Constraints

While NVIDIA's NVLink technology provides exceptional intra-rack connectivity, it faces scaling challenges in hyperscale deployments:

- **Hierarchical Bandwidth Degradation**: Inter-rack communication suffers from significantly reduced bandwidth compared to intra-rack links
- **Limited Fault Isolation**: Switch failures can partition large portions of the compute fabric
- **Static Resource Allocation**: Fixed topologies limit dynamic workload adaptation

The Linear Scaling Imperative

Enfabrica says its technology can connect more than 100,000 GPUs together. It's a solution that could help Nvidia offer integrated systems around its chips so clusters can effectively serve as a single computer. Achieving linear scaling in clusters of this magnitude requires fundamentally different approaches to fabric architecture.

Foundation Model Training Requirements

Modern foundation models with trillion-parameter counts require unprecedented levels of inter-GPU communication. The training process demands:

- **All-to-All Communication Patterns**: Gradient synchronization across tens of thousands of GPUs
- **Dynamic Load Balancing**: Adaptive redistribution of computation based on model complexity
- **Fault-Tolerant Execution**: Continued operation despite individual component failures

Enfabrica's Technology Deep Dive

The ACF-S Millennium SuperNIC

Bridging Compute and Network Domains

Architectural Innovation

The Enfabrica ACF-S "Millennium" is a 3.2Tbps network device on one side and then has 128 PCIe lanes on the other side. That 3.2Tbps is actually 32x 112G lanes providing a lot of flexibility on the networking side. This unique architecture bridges two traditionally separate domains—PCIe/CXL and high-speed networking—in a single silicon solution.

Multi-Domain Connectivity

The chip has the networking side, PCIe side, and the NIC pipeline side. To give some sense of scale, today's AI servers use 400GbE NICs and the GPUs have PCIe Gen5 x16 connections. This is enough on either side to handle eight NICs and eight GPUs. The ACF-S provides unprecedented flexibility by enabling:

- **Programmable Protocol Support**: With the middle portion, Enfabrica has programmability to help implement custom protocols
- **High-Radix Multipath Architecture**: One option is to have four 800G ports going to four switches. The other, and more interesting option might be having 100G connections to 32 switches
- **Enhanced Fault Tolerance**: If a switch goes down in a more traditional setup, that might mean a GPU loses communication or 800Gbps of network bandwidth is lost. With ACF-S in a 32x 100G configuration, a failed link means only around 3% of the bandwidth is lost

Zero-Copy Data Movement

The Enfabrica ACF SuperNIC, with its high-radix, high-bandwidth, and concurrent PCIe/Ethernet multipathing and data mover capabilities, can uniquely scale-up and scale-out four to eight latest-generation GPUs per server system. This capability enables efficient data movement without CPU intervention, critical for maintaining low latency in large-scale systems.

Enfabrica's Elastic Memory Fabric (EMFASYS)

Revolutionary Memory Architecture

RDMA-Ethernet Integration with CXL Memory

EMFASYS combines Remote Direct Memory Access (RDMA) Ethernet networking with ComputeExpressLink (CXL) DDR5 memory to provide pooled, rack-scale memory accessible by any GPU server over standard Ethernet ports. This integration represents a fundamental shift from static memory allocation to dynamic, fabric-attached memory resources.

Performance Characteristics

The ACF-S switch enables the striping of memory transactions across a wide number of memory channels and Ethernet ports. It delivers read access times in microseconds and a software-enabled caching hierarchy hides transfer latency within AI inference pipelines. Key performance metrics include:

- **Bandwidth Aggregation**: Powered by Enfabrica's 3.2 Terabits/second (Tbps) Accelerated Compute Fabric SuperNIC (ACF-S) elastically connecting up to 144 CXL memory lanes to 400/800 Gigabit Ethernet (GbE) ports
- **Memory Capacity Scaling**: Shared memory targets of up to 18 TeraBytes (TB) CXL DDR5 DRAM per node, networked using 3.2 Tbps RDMA over Ethernet
- **Latency Optimization**: Uncompromised AI workload performance with read access times in microseconds and software-enabled caching hierarchy that hides transfer latency within inference pipelines

Economic Impact

When deployed at scale with Enfabrica's EMFASYS remote memory software stack, the solution enables up to 50% lower cost per token per user, allowing foundational LLM providers to deliver significant savings in a price/performance tiered model.

Architectural Innovations in Scale-Out and In-Node Connectivity

Elastic Memory Fabric Architecture

Enfabrica CEO Rochan Sankar stated: "AI inference has a memory bandwidth scaling problem and a memory margin stacking problem. As inference gets more agentic versus conversational, more retentive versus forgetful, the current ways of scaling memory access won't hold".

Technical Implementation

Features include high-throughput, zero-copy, direct data placement and steering across a four or eight-GPU server complex, or alternatively across 18-plus channels of CXL-enabled DDR

memory. Its remote memory software stack based on InfiniBand Verbs enables massively parallel, bandwidth-aggregated memory transfers between GPU servers and commodity DRAM.

Performance Advantages

The switch is claimed to outperform flash-based inference storage alternatives with 100x lower latency and unlimited write/erase transactions. This performance differential becomes critical for real-time inference workloads where storage latency directly impacts user experience.

Strategic Integration

What NVIDIA Gains

Vertical Integration of the AI Stack

Complete Platform Control

Nvidia intends to combine Enfabrica's tech stack with its own hardware and software to offer more efficient training clusters for more powerful frontier models. This integration enables NVIDIA to control the entire AI infrastructure stack from silicon to system software, providing several strategic advantages:

- **Optimized Co-Design**: Hardware and software optimization across the entire stack
- **Reduced Vendor Dependencies**: Elimination of third-party networking bottlenecks
- **Enhanced Differentiation**: Unique capabilities unavailable to competitors

DGX/HGX Stack Enhancement

The integration of Enfabrica technology into NVIDIA's DGX and HGX product lines promises to deliver unprecedented system-level capabilities. The combination of ACF-S and EMFASYS might help Nvidia unlock higher GPU utilization rates and lower total cost of ownership — key metrics for hyperscalers and LLM developers operating at the cutting edge of AI.

Transition from Mellanox InfiniBand to Enfabrica Fabric

Architectural Evolution

While NVIDIA's acquisition of Mellanox brought InfiniBand expertise in-house, the Enfabrica acquisition represents a strategic pivot toward next-generation fabric technologies. The transition addresses several limitations:

- **Ethernet Standardization**: Enfabrica is an active advisory member of the Ultra Ethernet Consortium (UEC) and a contributor to the Ultra Accelerator Link (UALink) Consortium
- **Cost Optimization**: EMFASYS enables cost-effective memory scaling using commodity DRAM rather than expensive HBM
- **Fault Tolerance**: Multi-path architectures provide superior resilience compared to traditional point-to-point links

Impact on Future DGX and Supercomputing Platforms

Blackwell and Rubin Platform Integration

The integration of Enfabrica technology will likely manifest in future DGX platforms as:

- **Native EMFASYS Support**: Built-in elastic memory fabric capabilities
- **Enhanced ACF-S Integration**: Integrated SuperNIC functionality in base platform designs
- **Disaggregated Architecture**: Optimized context/generation workload distribution

Performance and Energy Efficiency Scaling

NVIDIA projects Vera Rubin NVL144 CPX to deliver $5 billion in token revenue per $100 million invested, tying its pitch directly to ROI in long-context inference. This economic model becomes possible through the combined architectural innovations of Rubin CPX and Enfabrica's fabric technology.

Real-World Implications

Practical Deployment Scenarios

Hyperscale Model Training

At first, one might think that having big 800G pipes is far superior. When we get to large-scale clusters, this extra redundancy on the link side can have huge implications to overall cluster reliability. For hyperscale deployments, the multi-path architecture provides:

- **Linear Scaling**: Support for 100,000+ GPU clusters with near-linear performance scaling
- **Improved Reliability**: Graceful degradation under component failures
- **Dynamic Load Balancing**: Adaptive bandwidth allocation based on workload requirements

Low-Latency Generative Inference

Modern models are evolving into agentic systems capable of multi-step reasoning, persistent memory, and long-horizon context. The combination of Rubin CPX for context processing and EMFASYS for memory fabric enables:

- **Million-Token Context Processing**: Efficient handling of extended context windows
- **Real-Time Agent Workloads**: Low-latency response for interactive AI systems
- **Persistent Memory Architecture**: Long-term context retention across sessions

Distributed Workload Optimization

The NVIDIA Vera Rubin NVL144 CPX platform packs 8 exaflops of AI performance and 100TB of fast memory in a single rack. This density enables new deployment models:

- **Rack-Scale Computing**: Single-rack systems with supercomputer-class performance
- **Edge Deployment**: High-performance AI capabilities in distributed locations
- **Hybrid Cloud Architecture**: Seamless workload distribution between on-premises and cloud resources

Operational ROI and Efficiency Gains

Higher GPU Utilization

The EMFASYS solution addresses the critical need for AI clouds to extract the highest possible utilization of GPU and High-Bandwidth-Memory (HBM) resources in the compute rack while scaling to greater user/agent count. Key efficiency improvements include:

- **Reduced Memory Stranding**: Dynamic allocation prevents idle memory resources
- **Load Balancing**: Optimal workload distribution across available compute resources
- **Fault Tolerance**: Continued operation despite individual component failures

Unified Memory Pool Benefits

Enfabrica claims it is no longer necessary to buy more GPUs to get more HBM capacity. The pitch is: use its switch to bulk up DRAM instead and make better use of the GPUs you already have. This architectural shift provides:

- **Capital Efficiency**: Reduced need for expensive HBM-equipped GPUs
- **Operational Flexibility**: Dynamic memory allocation based on workload requirements
- **Cost Optimization**: Use of commodity DRAM for memory-intensive workloads

Decreased Cluster Congestion

Enfabrica contends that ACF-S is more fault tolerant than traditional interconnect systems as it replaces point-to-point GPU connections with a multi-path architecture that reduces congestion, improves distribution of data, and ensures that GPU link failures don't stall compute jobs.

Open-Source and Enterprise Adoption Trends

Standards Participation

Enfabrica, already an actively contributing member of the Ultra Ethernet Consortium (UEC) and a member of its Technical Advisory Committee, is also now a Contributor-level member in the newly-formed Ultra Accelerator Link (UALink) Consortium. This participation ensures compatibility with emerging industry standards.

Ecosystem Development

NVIDIA Rubin CPX will be supported by the complete NVIDIA AI stack — from accelerated infrastructure to enterprise-ready software. The comprehensive software stack includes:

- **CUDA-X Libraries**: Optimized libraries for diverse AI workloads
- **NIM Microservices**: Production-ready AI deployment tools
- **Developer Ecosystem**: Support for 6 million developers and 6,000 CUDA applications

Future Directions and Industry Impact

NVIDIA's Complete AI Stack Ownership Strategy

Vertical Integration Implications

Gutting Enfabrica allows Nvidia to move fast, sidestep messy product overlaps, and reduce regulatory drag. It gets the people and the IP that matter without the burden of integration baggage. This strategy positions NVIDIA to:

- **Control Innovation Pace**: Rapid deployment of integrated solutions
- **Optimize End-to-End Performance**: Hardware-software co-optimization across the stack
- **Differentiate Against Competitors**: Unique capabilities unavailable elsewhere

Competitive Landscape Impact

The acquisition creates significant challenges for competitors like AMD and Intel, who must now compete against an increasingly integrated NVIDIA ecosystem. Perhaps the more interesting wrinkle in this is that the team to do this is not on the AMD or Broadcom side.

Standardization and Open Ecosystem Considerations

Multi-Vendor Compatibility

Despite vertical integration, NVIDIA's commitment to industry standards through Enfabrica's consortium participation suggests a balanced approach to ecosystem development. Key considerations include:

- **Standards Compliance**: Adherence to UEC and UALink specifications
- **Interoperability**: Support for multi-vendor environments
- **Open Source Components**: Continued contribution to open-source AI infrastructure

Regulatory Considerations

The FTC's scrutiny of Microsoft–Inflection and Amazon–Adept shows that Nvidia's play could easily attract attention even if it is not a traditional acquisition. However, the technical benefits may outweigh regulatory concerns given the industry-wide need for advanced AI infrastructure.

Predictions for AI Datacenter Evolution by 2028

Performance Trajectory

Based on current architectural trends and the roadmap implications of Rubin CPX and EMFASYS integration, several predictions emerge:

Hardware Modularity

The idea of building a co-processor around lower-cost memory makes a lot of sense. NVIDIA also has the ability to provide the software stack to make this happen in practice. By 2028, expect:

- **Disaggregated Memory Architecture**: Standard deployment of fabric-attached memory
- **Specialized Compute Units**: Purpose-built processors for specific AI workload phases
- **Adaptive System Configuration**: Dynamic resource allocation based on workload characteristics

Workload Flexibility

Early partners such as Cursor, Runway, and Magic are testing CPX for coding assistants, cinematic content, and agent-driven software engineering with massive context windows. Future capabilities will include:

- **Universal Context Processing**: Support for million-token context windows as standard
- **Real-Time Agent Deployment**: Low-latency AI agents with persistent memory
- **Multi-Modal Integration**: Unified processing of text, video, and code workloads

Economic Models

NVIDIA projects Vera Rubin NVL144 CPX to deliver $5 billion in token revenue per $100 million invested. This economic efficiency will drive widespread adoption and enable new business models around AI infrastructure as a service.

References

[1] NVIDIA Technical Blog - "Inside NVIDIA Blackwell Ultra: The Chip Powering the AI Factory Era" [2] CNBC - "Nvidia just spent over $900 million to hire Enfabrica CEO, license AI startup's technology" [3] ServeTheHome - "Why Enfabrica Has the Coolest Technology" [4] NVIDIA Newsroom - "NVIDIA Unveils Rubin CPX: A New Class of GPU Designed for Massive-Context Inference" [5] NVIDIA Technical Blog - "NVIDIA Rubin CPX Accelerates Inference Performance and Efficiency for 1M+ Token Context Workloads" [6] Network World - "Nvidia reportedly acquires Enfabrica CEO and chip technology license" [7] Blocks and Files - "Enfabrica touts elastic AI memory fabric for GPU workload efficiency" [8] TechPowerUp - "Enfabrica Unveils Industry's First Ethernet-Based AI Memory Fabric System" [9] Unite.AI - "Enfabrica Unveils Ethernet-Based Memory Fabric That Could Redefine AI Inference at Scale" [10] Futurum Group - "NVIDIA Rubin CPX Targets Future of Large-Scale Inference" [11] Business Wire - "Enfabrica Unveils Industry's First Ethernet-Based AI Memory Fabric System"