

SCALING OUT FUTURE-READY AI INFRASTRUCTURE

The Complete Guide to the AMD
Pensando™ Vulcano 800 AI NIC

THE AMD PENSANDO™ VULCANO 800 AI NIC: PURPOSE BUILT FOR HIGH-THROUGHPUT, LOW-LATENCY AI NETWORKING

THE EVOLUTION OF AI WORKLOADS

Across training, distributed inference, and agentic AI, workloads are rapidly scaling beyond the limits of traditional systems.

AI infrastructure can no longer be designed around compute alone – the network is now a primary determinant of what the system can deliver.

MEET THE NEXT GEN AMD AI NIC™

- > ONLY **FULLY PROGRAMMABLE** SOFTWARE AND HARDWARE AI NIC
- > SELECT YOUR **TRANSPORT OF CHOICE** (MRC, UEC-READY RDMA, CUSTOM)
- > FLEXIBLE **ARCHITECTURE BUILDS**

KEY BENEFITS

UP TO 2.4 TBPS¹

On 3 NICs per GPU

UP TO 800 GBPS

Connectivity per NIC

50% MORE

Scale out bandwidth¹

UP TO 13% BETTER

AI job completion time²

UP TO 33% LOWER

Network switching cost³

PERFORMANT, PROGRAMMABLE, AND OPEN SOLUTIONS FOR AI WORKLOADS



FASTER JOB COMPLETION

Unprecedented networking bandwidth accelerates job completion across scale-out clusters.



ENHANCED CLUSTER UPTIME

Improved reliability, availability, and serviceability (RAS) to help reduce interruptions in large-scale AI workloads.



OPTIMIZED CAPEX SPENDING

Optimized network switching costs by scaling AI clusters on fabrics that need fewer transceivers, cables, and optics.



OPERATIONAL EXCELLENCE

AI infrastructure maintains production-readiness with automated deployment, minimized downtime, and cluster-wide visibility.



PROGRAMMABILITY

A fully P4 programmable hardware and software stack enables choice of transport protocol, including MRC, UEC-ready RDMA, and intelligent custom transports.

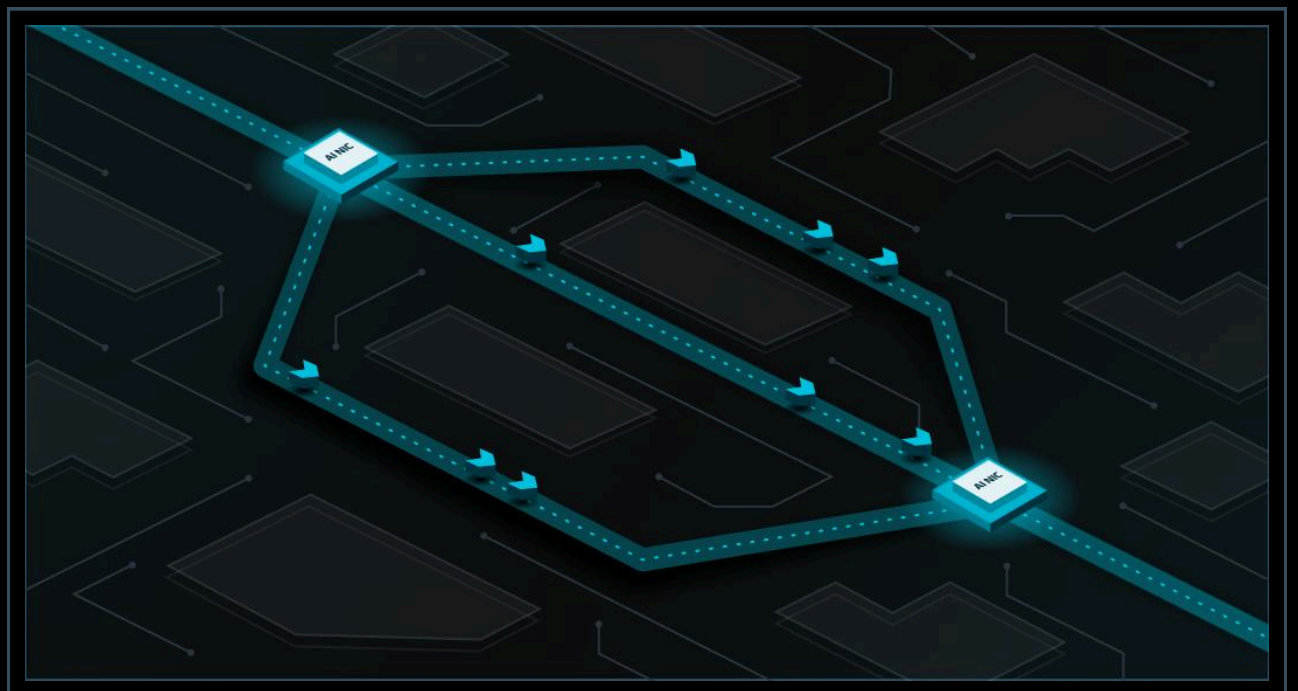
KEY FEATURE: LOAD DISTRIBUTION

Higher bandwidth, smarter distribution:

2.4 Tbps bandwidth combined with programmable software helps deliver up to 13% faster AI job completion times².

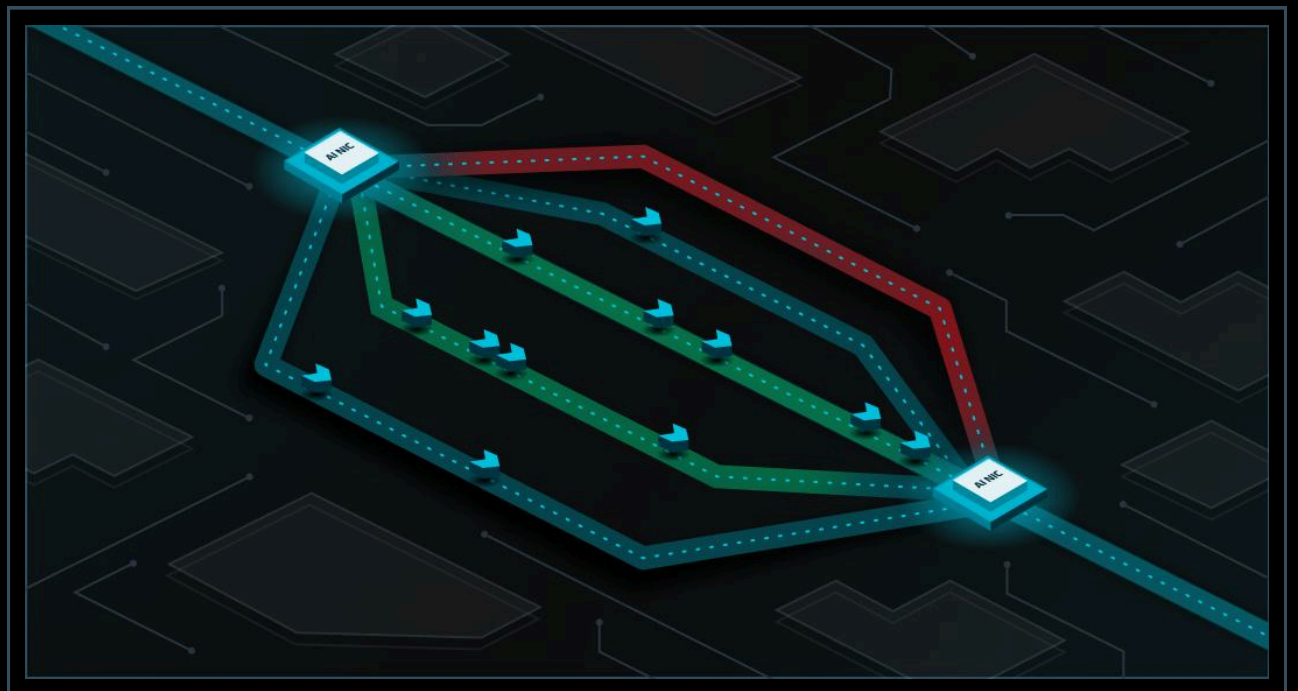
INTELLIGENT PACKET SPRAY

Distributes packets efficiently across multiple paths, maximizing utilization and scalability.



PATH-AWARE CONGESTION CONTROL AND DYNAMIC LOAD BALANCING

Proactively reroutes traffic around congestion, helping prevent performance cliffs and simplifying network performance management.



KEY FEATURE: RELIABILITY

OUT-OF-ORDER PACKET HANDLING

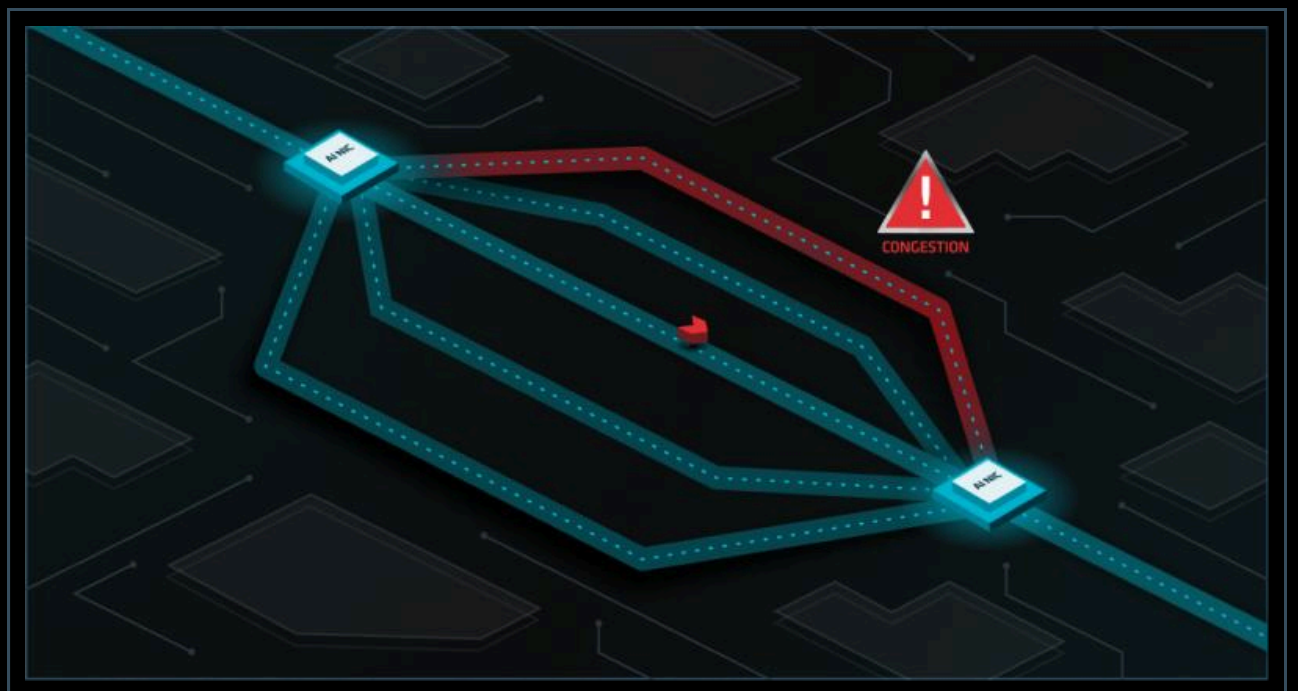
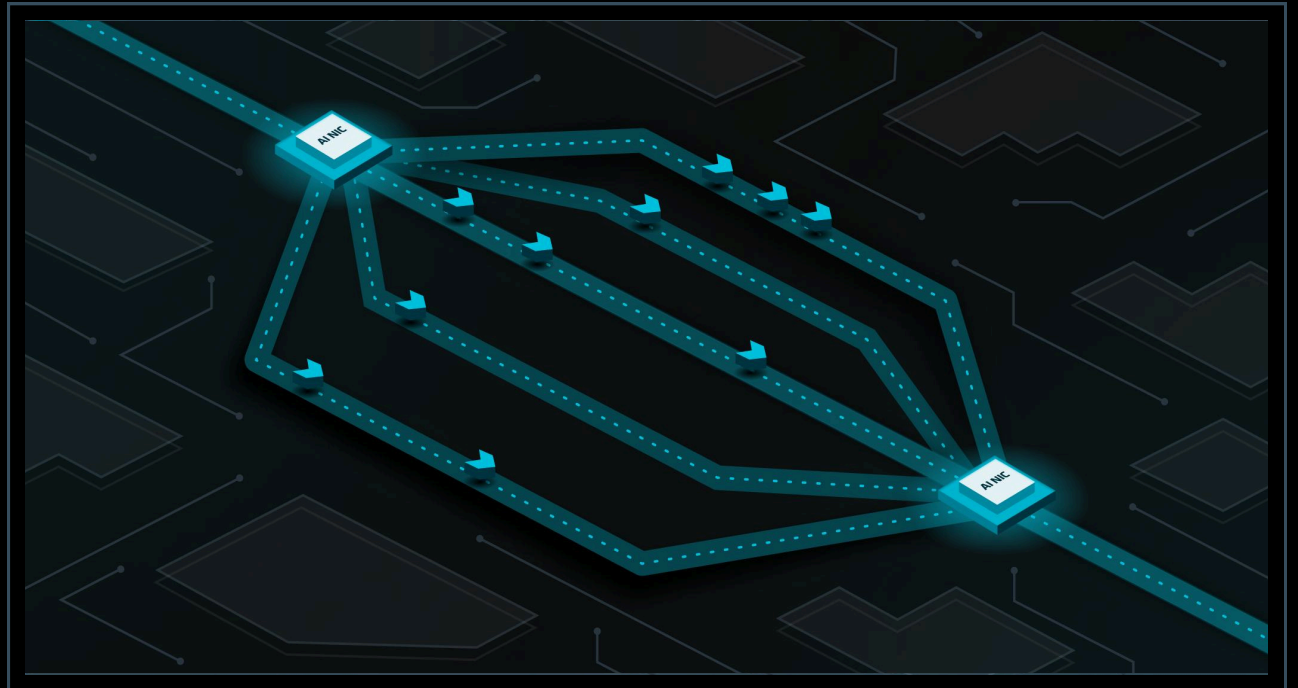
Direct placement into GPU memory without buffering enables fast data throughput.

IN-ORDER MESSAGE DELIVERY

Maintains sequencing across multipathing, protecting AI workload integrity.

SELECTIVE RETRANSMISSION

Only retransmits dropped or corrupted packets, helping reduce recovery, latency, and avoid redundant traffic.



LOAD BALANCING ARCHITECTURE

The AMD Pensando™ Vulcano 800 AI NIC supports three distinct load balancing architecture modes, giving operators the flexibility to match network behavior to their infrastructure, workload, and switching environment.

The P4 programmable stack decouples transport behavior from hardware, allowing architecture to evolve alongside emerging standards without requiring a hardware refresh.

	AI NIC-BASED PACKET SPRAY	AI NIC-BASED SOURCE ROUTING	SWITCH-BASED PACKET SPRAY
WHERE SPRAY DECISION IS MADE	AI NIC	AI NIC	Switch
WHERE THE PACKET SPRAY HAPPENS	Switch	AI NIC	Switch
USE CASES	Large fabrics with many equal-cost multi paths	Application communication patterns are well defined	Environments where switching infrastructure natively supports per-packet DLB

AI NIC-BASED PACKET SPRAY

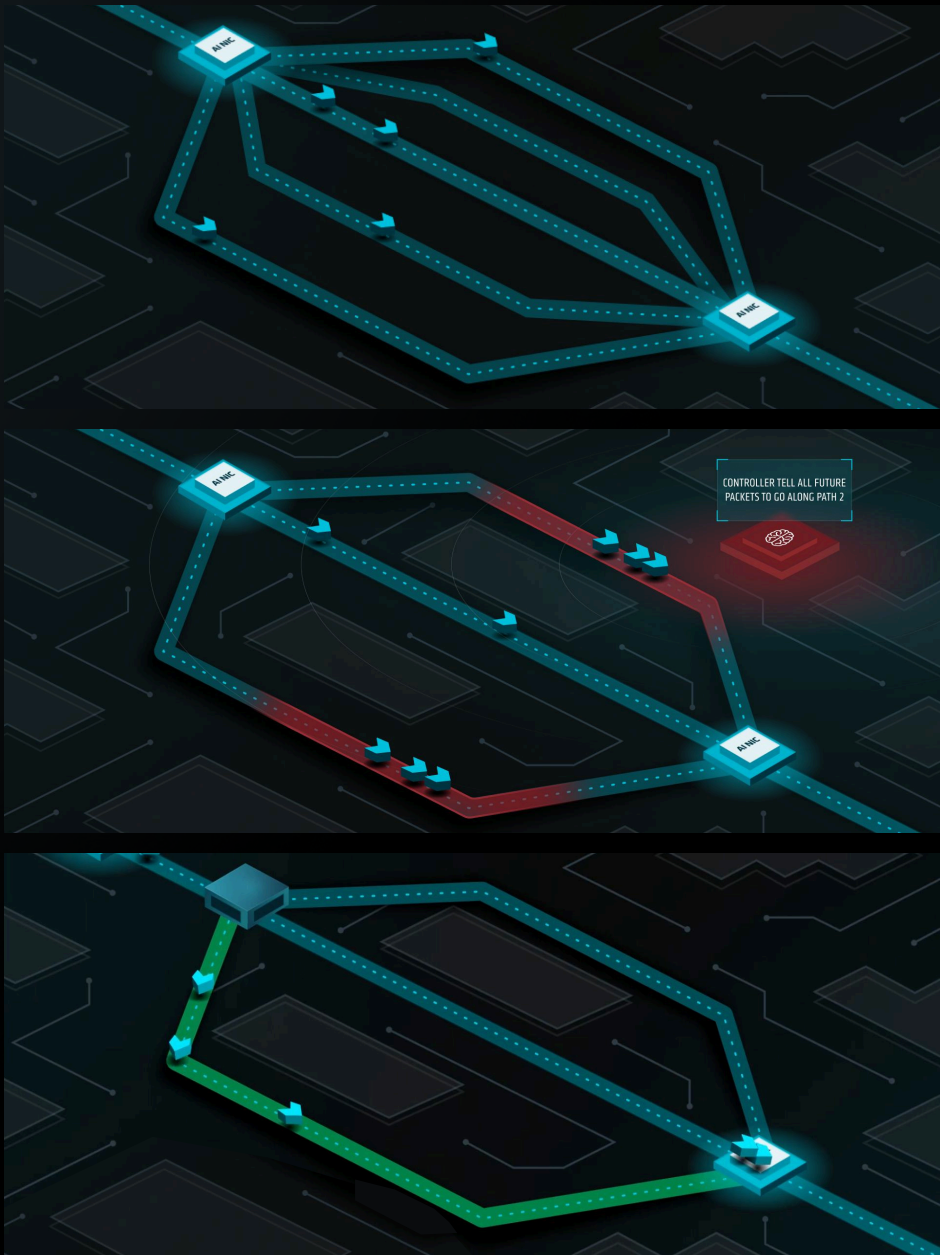
The AI NIC generates entropy values that drive path distribution at the switch, with path-aware congestion control operating independently at the AI NIC. This keeps fabric utilization high without placing any additional configuration requirements on the switching infrastructure.

AI NIC-BASED SOURCE ROUTING

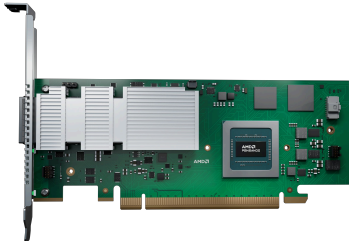
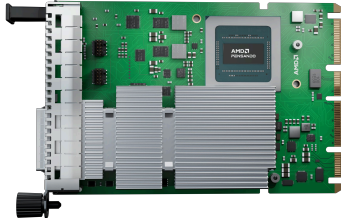
The AI NIC assigns per-packet path identifiers, giving operators deterministic control over exactly how traffic traverses the fabric across independent network planes. When links fail, the AI NIC detects and rebalances traffic automatically, without relying on the switch to respond.

SWITCH-BASED PACKET SPRAY

The switch drives per-packet load balancing across available paths, while the AI NIC manages packet reordering to deliver in-order message completions. The result is that multipathing complexity stays in the network layer, invisible to the application.



TECHNICAL SPECS

				
	PCIe®	OCP	VULCANO AI NIC BOARD - 4	VULCANO AI NIC BOARD - 6
MAX BANDWIDTH PER GPU	800 Gbps		1.6 Tbps	2.4 Tbps
NICS/GPU	NA		2 NICs/GPU	3 NICs/GPU
FORM FACTOR	HHHL PCIe CEM	OCP 3.0 TSFF	Custom	
CONNECTOR	1 x OSFP-800		2 x 1.6T OSFP	3 x 1.6T OSFP
ETHERNET INTERFACE	1 x 800G / 2 x 400G / 8 x 100G		4 x800G / 8x400G / 16x200G	6 x800G / 12x400G / 24x200G
NIC COOLING	Air-Cooled		Liquid-Cooled	
MANAGEMENT	PLDM over MCTP/ USB RBT NC-SI Over MCTP/ SMBus			
BOARDS PER COMPUTE TRAY	1	8/4	2	

THE AMD RACK SCALE SOLUTION

The AMD Helios AI Rack Solution is purpose-built to meet the demands of frontier training and large-scale distributed inference. The AMD Pensando Vulcano 800 AI NIC is the scale-out networking cornerstone of this solution.

With up to 2.4 Tbps of bandwidth per GPU, multiplane support and operational resilience, and a fully programmable stack, the next generation AMD AI NIC delivers the intelligence to accelerate job completion and maximize compute utilization at scale.

READY TO SCALE?

[Request a Consultation](#)



FOOTNOTES

1. Result of Vulcano board 3-NICs-per-GPU configuration.
2. MI400-018: Based on AMD Engineering silicon modeling and AMD synthetic benchmark simulation, using the MOE_4p5T hi_sparsity_ea9 benchmark test to project time- to- solution speed up in days using the FP8 training datatype on a simulated LLM system modeled with 8000 AMD Instinct MI455X GPUs and (3) versus (2) AMD Pensando Vulcano NICs.
Results reflect analysis of pure training compute (decode layers only) Configuration evaluated with a global batch size of 4096 and a sequence length of 4K, using SwiGLU activation. Results exclude evaluation and checkpointing overhead. FlashAttention v3 with matmul–softmax overlap is assumed. AllReduce and All2All communication costs are fully accounted for and not hidden via tiled compute–communication overlap. Gradient synchronization and FSDP weight prefetch are evaluated across varying levels of overlap to assess scale-out sensitivity. Results assume ideal and not fully optimized real-world behavior and may vary when actual product(s) are released in market.
3. PEN-022:
AMD comparison and pricing as of May 18, 2026, for network fabric costs to support 32,000 GPUs. Comparison of a Vulcano-based NIC (VULCANO-CUSTOM-2.4T) deployed as part of a Helios rackscale system with a network based on 1.6T Tomahawk 6 switching with 200G SerDes versus using a competitor 800G NIC with 800G Tomahawk 6 switching with 100G SerDes. Both fabrics were fat-tree topologies built on Tomahawk 5 800G switching platforms, with NIC costs considered comparable. The Vulcano-based design is estimated to deliver up to 33% savings in network switching costs by enabling a more cost-effective architecture with fewer switching platforms, more bandwidth per port on the network, and reduced transceiver cables/optics.

TH6-100G Serdes Fat-Tree (Competition):
Switching (TH6C BCM78914 - 128x800G):

• Leaf Units	1,000
• Spine Units	500
• Total Switches	1,500
• TH6 100G Unit Price	\$79,587
• Total Switching Cost	\$60M

Cables/Optics:

• NIC Transceivers (800G-DR8)	64,000 @ \$500
• Leaf/Spine Transceivers (800G-DR8)	192,000 @ \$500
• MPO Cables	256,000 @ \$89

Total Optics/Cables	\$64M
Total Fabric Cost TH6-100G (Switches + Cables/Optics):	\$124M

TH6-200G Serdes Fat-Tree (Vulcano-Custom2.4T / AMD Solution):
Switching (TH6P BCM78910 - 64x1.6T):

• Leaf Units	1,000
• Spine Units	500
• Total Switches	1,500
• TH6 200G Unit Price	\$66,336
• Total Switching Cost	\$50M

Cables/Optics:	
NIC Transceivers (1.6T-DR8)	32,000 @ \$900 (50% fewer vs. competition)
Leaf/Spine Transceivers (1.6T-DR8)	64,000 @ \$900
MPO Cables	128,000 @ \$89 (50% fewer vs. competition)
Total Optics/Cables	\$43M
Total Fabric Cost TH6-200G (Switches + Cables/Optics):	\$93M
Capex Savings (Fabric only):	
Savings \$:	\$30.7M
Savings %:	33.1%

Pricing sources: SemiAnalysis Hyperscaler Networking Model data used with permission; full analysis available via SemiAnalysis subscription. Edgecore switch pricing as of May 17, 2026. Results may vary based on system configuration.