

Redefining AI Infrastructure with Fully Disaggregated, Fully Heterogeneous Architecture

Unified GPU-FPGA-CXL Memory for Scalable AI |
zett.ai



Problem & Opportunity

- VRAM capacity & cluster utilization limit modern AI workloads
- NVLink-only scaling is expensive & inflexible
- Need pooled, tiered memory that scales to 10–100 TB and feels like VRAM

What is ZettEngine?

- Hardware + software platform exposing a unified GPU-FPGA-CXL memory fabric
- ZettFabric SoC: Custom OoO RISC-V with PCIe/CXL 3.0 switch support
- ZettFabric (FPGA): Shipping PCIe/CXL 2.0 bridge extending H100/GB300 memory
- ZettOS + SDK/CLI: CUDA 13 & PyTorch 2.9 integration



ZettEngine – Target Rack (Simplified)

zett.ai Memory Target (Single Rack)



CXL 3.0 switching + PCIe/CXL dataplane; CXL2PCIe Bridge

Software Stack

- ZettOS: pooling, NUMA policies, hot-plug & error isolation
- ZettFabric: Linux driver with CUDA 13.1 / PyTorch 2.9 unified addressing
- ZettCLI / SDK: telemetry, allocation, policies, per-job quotas
- APIs: Python & C/C++; K8s/Slurm ready

FPGA Bridge

- PCIe/CXL 2.0 bridge with x16 Gen5 links
- Extends H100 VRAM via CXL.mem semantics
- < 400 ns round-trip bridge latency
- Works with CUDA & ROCm via ZettBridge SDK
- Developer portal: <https://zettai.us/portal>

Product Line & Pricing

Version	Configuration	Price
ZettEngine	Threadripper PRO 9960X + SMART CXL 512GB + RTX 5090	\$10,000
ZettEngine Pro	Intel Xeon 6787P + SMART CXL 512GB	\$20,000
ZettEngine Pro Max	Xeon 6787P + RTX 5090 + SMART CXL 512GB + ia780i FPGA	\$40,000
ZettFabric (FPGA)	Intel VR3 FPGA dual port CXL2PCIe Bridge	\$30,000

Key Advantages

- Up to 10× lower inference cost vs NVL72
- 95% utilization (eliminate idle compute)
- 5× energy efficiency via CXL pooling
- Unified GPU-CXL-FPGA memory fabric
- Scalable to 100 TB pooled memory with CXL 3.0 switching

Availability & Roadmap

- ZettBridge FPGA (PCIe/CXL 2.0): Available today
- RISC-V SoC with PCIe/CXL 3.0 Switch: 2H 2026 (Q3 sampling)