

Silent AI ES

Series R · Series H · Series Z

ON-PREMISES AI APPLIANCE

Designed by FAST LTA in Munich, Made in Germany

Integrated AI Appliance · self-hosted · On-Premises · no cloud dependency · incl. German Hardware SLA

TECHNICAL SPECIFICATIONS

	Series R	Series H	Series Z
PLATFORM · Common to all Series			
Form Factor	4U · 19" rack · 400 mm depth		
CPU	Intel Xeon Gold 6414U · 32 Cores · 2.0 GHz (3.9 GHz boost) · 250 W TDP		
System RAM	128 GB or 512 GB DDR5 ECC		
PCIe Architecture	PCIe 5		
System Storage	8 TB NVMe		
Data Networking	Dual 10GbE (RJ45 / SFP+) or Dual 25 / 100 GbE (QSFP)		
Max. Power Input	2,400 W		
GPU CONFIGURATION · Per Series			
GPU	1–3× NVIDIA RTX Pro 6000	1–2× NVIDIA H200	2× H200 + 1× RTX Pro 6000
GPU Architecture	Blackwell	Hopper	Hopper + Blackwell
GPU VRAM per GPU	96 GB GDDR7 ECC	141 GB HBM3e	141 GB / 96 GB
Max. Combined VRAM	up to 288 GB (3× GPU)	up to 282 GB (NVLink)	378 GB total
Memory Bandwidth	1,597 GB/s per GPU	4,800 GB/s per GPU	4,800 + 1,597 GB/s
FP32 Performance	up to 360 TFLOPS (R3000)	up to 134 TFLOPS (H2000)	up to 254 TFLOPS (Z3000)
Peak AI Perf. (FP4)	up to 12 PFLOPS (R3000)	—	—
SILENT AI ENTERPRISE SEARCH SOFTWARE · Included with every appliance			
Licensing	Unlimited users · unlimited token usage · flat rate per appliance — no per-seat licenses, no API quotas		
AI Platform	Easy-to-use chat experience · conversation history · full-text search · file upload		
Knowledge Base	RAG document integration · private knowledge base management · vector search		
Access Control	Role-based access management (RBAC)		

Subject to product changes. Technical specifications available on request. As of 8 August 2026.