AI INFERENCE SYSTEMS / TSR-CORE + TSR-LINK + TSR-MEM

Inference, scaled Indefinitely

One chip at a time. One connected system.

Proprietary inference compute, flexible memory capacity, and TSR-Link fabric, engineered as one system.



AI demand is scaling. So must the system.

Larger models, longer contexts, and concurrent agents demand more compute, memory capacity, memory bandwidth, and network throughput. Idle capacity raises cost per token.

TSR-Core. TSR-Link. TSR-Mem.

TSR-Core: our proprietary NPU core technology, purpose-built for AI inference.
TSR-Link: a direct fabric designed to connect up to 1,024 chips as one inference machine.
TSR-Mem: decoupled compute and memory capacity, configured to fit each workload.

TOKEN ECONOMICS

8 –10×

modeled cost-per-token advantage

For selected frontier-scale workloads compared with NVIDIA systems.

BUILT FOR DEMANDING WORKLOADS

More models. More users.

Prefill/decode. Long-context analysis. Parallel agents. Private AI. Inference clouds. Connected capacity for demanding models, with cost per token at the core.

BUILT BY SILICON VALLEY VETERANS

Tessora is led by Silicon Valley veterans and chip architects with 20+ years each in GPU, SoCs, and networking, with multiple successful exits.

Contact

Tessora Systems

info@tessorasys.ai
650 215 8585
www.tessorasys.ai

Tessora Systems, Inc.
800 W El Camino Real #180
Mountain View, CA 94040
United States