

Abstract

The present invention discloses a distributed parallel system and method for multi-AI compute cores interconnected via a high-speed bus. The AI compute cores may be implemented as cascaded AI processor nodes or as multiple cores integrated within a single semiconductor die and interconnected by an internal bus such as AXI. The system addresses the computational bottlenecks of single cores caused by thermal limitations and semiconductor process constraints, thereby achieving near-linear scaling of both memory bandwidth and computational capacity. This architecture enables the deployment of large-scale AI models on edge devices or consumer terminals without requiring high-bandwidth memory (HBM) or server-grade cooling solutions.

The system integrates dedicated hardware circuits with a software compiler to enable direct data communication and synchronization among on-die high-speed memories (e.g., SRAM, Cube Memory) across the cascaded compute cores, thereby reducing off-chip memory access. The compiler employs a scheduling strategy that balances depth-first execution for on-die memory reuse within each AI compute core with inter-core execution timing, ensuring that producer operations complete before consumer operations begin while maintaining efficient local computation.

This approach provides a scalable and cost-effective solution for real-time edge AI computation through multi-core synchronization, optimized on-die memory utilization, and compiler-assisted scheduling.

Fig 1