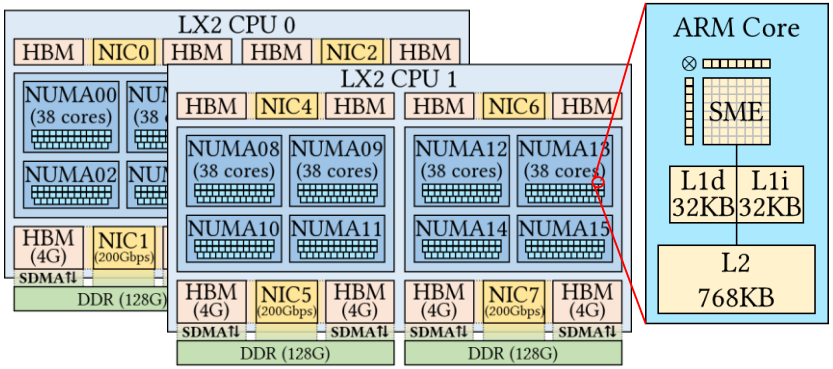
LineShine: Online Acceleration for HPC-AI Converged Scientific Computing

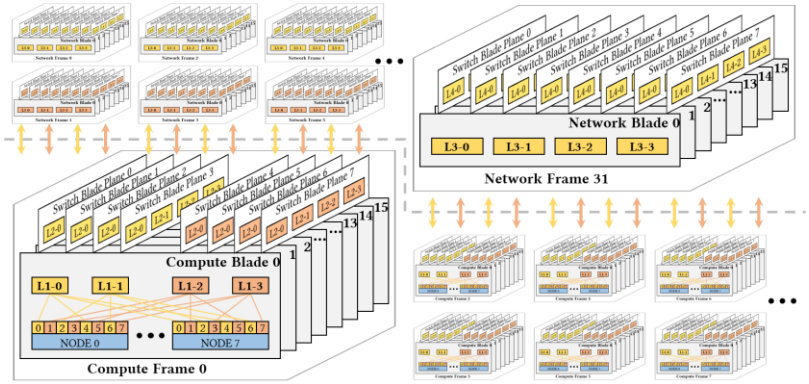
Yutong LU, Guangnan FENG, Haohuan FU

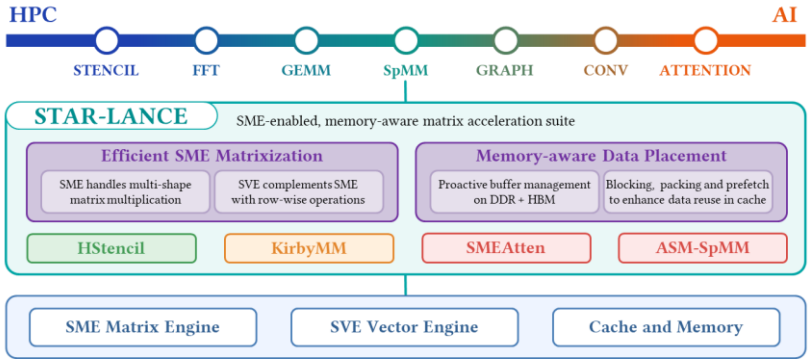
Frontiers of Computer Science, DOI: 10.1007/s11704-026-61591-w

LineShine: New Arch. Supercomputer for HPC-AI

- **High Performance** CPU
  - Online Acceleration, SME + SVE
  - 304 cores, 60.3TFlops@FP64
  - Multi-Precision, FP64/32/16, INT8
  - HBM 4 TB/s, Chiplet Architecture
- **High Scalability** Interconnect
  - Lingqi network, 2 plane × 4 rail Fat-tree
  - Mini Lat. 1.07us, bisection BW 3.5Pbps
  - Scalable to over 100,000 nodes
- **High Capacity** Storage
  - Symmetric Distributed Tiered Storage
  - Globally shared namespace & lifecycle Date
  - management.200PB – 1000PB of capacity
- **High Productivity** Software
  - Architecture-aware system software
  - HPC-AI convergence Framework
    - STAR-Lance, STAR-Flux, and STAR-Helix







High Efficient System and Applications

Benchmarks

HPL: 2.1984 EFlops, 80.37% computation efficiency, 52.07GFlops/W, No. 1 in the June 2026 TOP500 rankings

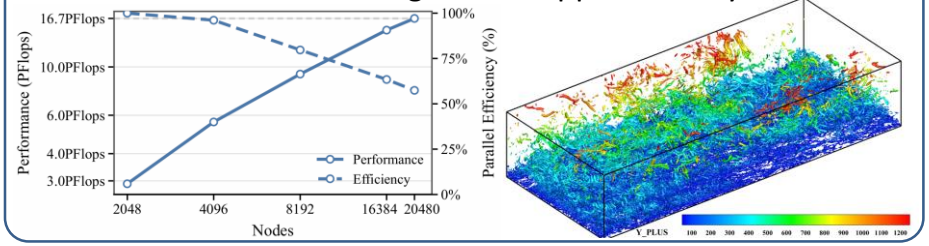
HPCG: 22 PFlops, No. 1 in the June 2026 rankings; HPL-MxP: 7.92 EFlops, 71.10% FP16 computation efficiency

Applications

HPC-AI convergence full-system-scale production applications across many scientific domains

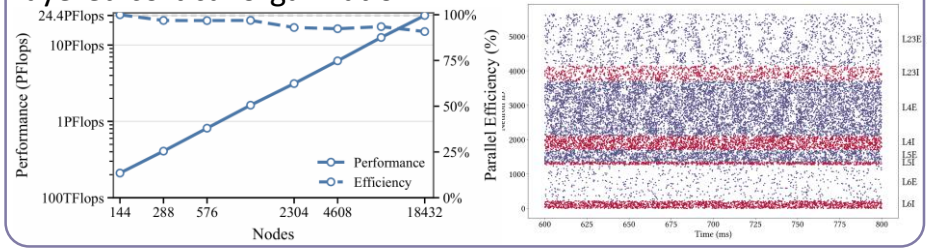
Turbulence Direct Numerical Simulation

PowerLLEL has completed wall-turbulence direct numerical simulation at a friction Reynolds number of approximately 21,000, the highest publicly documented production milestone in this class, and is advancing toward approximately 43,000.



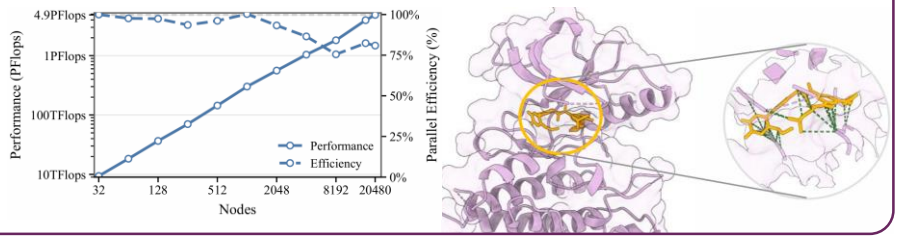
Whole-Brain Spiking Simulation

CerebroSim is the first whole-brain spiking simulator to reach the 100-trillion-synapse scale, simulating 86 billion neurons and 100 trillion synapses with biological constraints and layered cortical organization.



Ultra-Large-Scale Drug Screening

GLARE-Exa combines AI-guided molecular generation with physics-based docking and traverses a synthesizable chemical space of 10 trillion compounds in 22.6 hours, using physics-informed rewards to anchor the AI model.



Generative Compression for Earth Observation

D2AR trains a 6.3-billion-parameter generative Earth-observation reconstruction model at 1.54 EFlops and supports compression ratios of up to 10,000×, enabling observations to be reconstructed on demand from compact representations.