signal65

February 2026

Azure Maia 200

The Inference Efficiency Frontier

A full-stack inference play across silicon, memory hierarchy, and scale-up networking.

- FP4 TENSOR COMPUTE
- INFERENCE-FIRST ON-CHIP FABRIC
- BALANCED MEMORY HIERARCHY
- INTEGRATED NIC TWO-TIER SCALING
- VARIABLE POWER BALANCING

01 — THE CORE THESIS

Redefining the Inference Efficiency Frontier

AI inference is an efficiency frontier defined by the intersection of accuracy, latency, throughput, cost, and energy. Microsoft positions Maia 200 as part of a portfolio approach — a heterogeneous Azure fleet that matches the best silicon to the customer workload.

Different applications land at different points on this curve.

Interactive workloads tend to be latency sensitive, while batch and throughput-oriented workloads often prioritize cost and sustained token output.

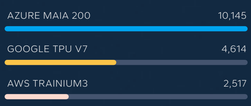
ACCURACY | LATENCY | THROUGHPUT | COST | ENERGY

02 — THE SILICON BET

FP4 Throughput Leadership, Built for Inference Economics

Maia 200 makes a deliberate bet on narrow-precision formats as the path to scaling token generation efficiently. Peak FP4 throughput positions it as a serious competitor to modern AI accelerators in the narrow-precision inference era.

10,145
TFLOPS FP4
Peak dense FP4 throughput



5,072
TFLOPS FP8
Peak FP8 throughput

1,268
TFLOPS BF16
Peak BF16 throughput

Narrow-precision formats, increasingly including FP4, are becoming central to scaling token generation efficiently.

- Narrow-precision tensor compute
- Inference-focused on-chip fabric
- Integrated NIC designed to enable a two-tier scale-up approach

Peak specs are directional indicators; sustained throughput and token economics depend on models, quantization, and scaling efficiency.

03 — THE MEMORY WALL

A Tiered Hierarchy for Effective Bandwidth

Large SRAM and explicit locality controls help keep high-value data on die, reduce repeated HBM accesses, and improve sustained throughput. The practical goal: keep data local when possible, fall back to HBM, and use remote HBM through scale-up paths when needed.

TIER 1 • ON-DIE SRAM

Ultimate Locality

272
MB On-Die SRAM

80
TB/s SRAM Bandwidth

Keep high-value data on die, reduce repeated HBM accesses

TIER 2 • HBM3E

Primary Model Storage

216
GB HBM3E

7
TB/s HBM Bandwidth

Primary working set and model weight storage

BANDWIDTH COMPARISON

WHY IT MATTERS

The Toolbelt Analogy

Tools on your belt are instant. Tools in the truck mean a walk across the jobsite. A trip to the hardware store is even slower. Maia 200's memory hierarchy works the same way: on-die SRAM (80 TB/s) keeps the most critical data at your hip, HBM (7 TB/s) is the truck, and remote memory over scale-up networking (2.8 TB/s) is the store run. A bigger belt means fewer trips.

Inference is frequently memory bound rather than compute bound.

04 — NETWORKING

Reliable Chip-to-Chip Communication at Scale

Maia 200 pursues a scale-up approach based on observed inference patterns, pairing an integrated NIC with commodity Ethernet to reduce cost and power while preserving communication performance across larger clusters.

2.8
TB/s Bidirectional
Scale-up bandwidth per accelerator

6,144
Accelerators
Maximum scale-up domain size

- Integrated NIC reduces cost & power
- Two-tier scale-up on commodity Ethernet
- Predictable collective operations
- Improved scheduling & bin packing

Larger scale-up domains can improve scheduling, bin packing, and utilization across diverse workloads — supporting common parallelism patterns in modern inference with strong tensor-parallel performance.