

OpenAI Jalapeño: The First Proprietary AI Chip and Its Strategic Impact

MANAGEMENT SUMMARY & CORE TAKEAWAY

OpenAI has launched Jalapeño, its first proprietary ASIC chip for AI inference – developed in just nine months with Broadcom and TSMC. The 'Intelligence Processor' aims to reduce inference costs per token by 50 percent compared to current Nvidia GPUs and is already running GPT-5.3-Codex-Spark internally. Deployment begins late 2026, with full ramp-up by mid-2028. For companies relying on AI APIs, this means potentially lower operating costs and the start of a full-stack strategy that consolidates OpenAI's control over the entire AI infrastructure.

CONTEXT & EXECUTIVE RELEVANCE

In June 2026, OpenAI unveiled its first proprietary chip: Jalapeño, an ASIC (Application-Specific Integrated Circuit) optimised exclusively for large language model (LLM) inference. Unlike general-purpose GPUs from Nvidia, Jalapeño is a specialised 'Intelligence Processor' developed in partnership with Broadcom (silicon design) and Celestica (board and rack integration). TSMC handles manufacturing.

The chip marks a strategic inflection point for OpenAI: moving away from dependence on external hardware providers toward a full-stack platform with control over models, software, and now the underlying hardware. This approach follows the precedent set by Apple (M-series chips), Google (TPU), Amazon (Trainium/Inferentia), and Meta (MTIA) – companies that leverage vertical integration to reduce costs and maximise performance.

CORE ACTION ITEMS

- 01 Engineering samples are already running GPT-5.3-Codex-Spark in internal tests
- 02 Performance per watt is 'substantially better than current state-of-the-art,' according to OpenAI
- 03 Likely eight HBM stacks (High Bandwidth Memory) to minimise latency in inference workloads
- 04 Deployment starts late 2026, full ramp-up by mid-2028

SWISS COMPLIANCE & ARCHITECTURE GUIDELINES

- ☑ Full compliance with revised Swiss Data Protection Act (revDSG).
- ☑ Zero vendor lock-in on closed proprietary models.
- ☑ Seamless integration into Abacus, Bexio and M365 ecosystems.

Non-Binding Process Analysis

Let us review in 20 minutes where AI agents create immediate measurable efficiency in your workflows.

ki-outsourcing.ch