AMD INSTINCT™ CODER

FRONTIER CLASS AGENTIC CODING AT A FRACTION OF THE COST, ON OWNED HARDWARE

A pre-integrated, pre-validated stack that routes, meters, and governs token usage across local open-weight models and frontier providers.

Intelligent multi-model routing	Meter usage and cost	Govern consumption
• Route by user, team, app, workload, endpoint, or policy. • Fall back to frontier models based on quality, capacity, or policy.	• Track input/output tokens, requests, estimated cost. • Compare local vs. frontier traffic to find savings opportunities.	• Set quotas, rate limits, and time-bound controls. • Audit usage, routing decisions, and policy enforcement.

THE PROBLEM

AI adoption is scaling faster than token cost controls. Coding assistants, agents, and AI-powered workflows have moved from pilots to broad adoption. Token and API costs are rising with usage. Most organizations cannot see where the spend is going, which teams are driving it, or which requests actually need a frontier model.

Innovation cannot slow down. Teams need to keep using AI, with control over cost, model usage, data exposure, and governance.

THE DILEMMA

- **Build it yourself:** This gives maximum control, but is slow to deliver and operationally heavy.
- **Use AI cloud or inference-as-a-service:** This option is fast to start, but consumption costs continue to rise while giving you less control over your stack.
- **Deploy the joint solution:** A fast, private path to control which models to use, where they run, and what executes locally vs. with frontier providers.

$1.2M
current spend

$370K
annual spend with AMD Instinct Coder

$830K
savings

THE AMD INSTINCT™ CODER SOLUTION

Together, AMD, Spectro Cloud, and Supermicro deliver an enterprise-ready AI coding solution with complete control over data, cost, and governance.

Hardware	**Supermicro AS -8126GS-TNMR server** • 2 x AMD EPYC™ 9575F, 64 cores, 3.3GHz CPUs • 8 x AMD Instinct™ MI325X • 3TB (24 x 128GB) DDR5 RDIMM 6400 ECC • 2 x 960GB NVMe PCIe® Gen4 V6 M.2 • 8 x 7.68TB PCIe Gen5 TLC U.2 SSD • 10 x AMD Pensando™ Pollara 400 HHHL PCIe NIC, 400GbE • 6 x 5250W redundant (3+3 configuration) titanium-level high-efficiency power supplies
Software	**Spectro Cloud PaletteAI Inference Launchpad** • Full-Stack AI lifecycle management • Enterprise governance at scale • Intelligent local-first model routing, frontier when justified • Full visibility and control of AI usage and cost
AI models	• AMD Inference Microservices optimized GLM-5.2 • Frontier models from Anthropic, OpenAI, Google (when justified)
Developer tools	Claude Code, OpenAI Codex, Visual Studio Code, Cursor
Support	• Comprehensive full stack software support from Spectro Cloud • 3 yr next business day (NBD) on site from Supermicro for hardware
Scale	Supports up to 50 developers per node (30 concurrent)
Performance	Up to 95% of Frontier model benchmark performance

WHY AMD INSTINCT™ CODER

The joint solution from Spectro Cloud, AMD, and Supermicro is a packaged private AI solution. This brings the best of both worlds as it is faster than DIY but enables full control compared to AI clouds.

- Intelligent multi-model routing that route prompts to the local models by default and frontier models only when needed.
- Token governance to meter, quota, rate-limit, and audit usage before costs get out of control.
- Start on a single Supermicro server and scale out by adding servers as capacity grows.
- Manage locally at first, then switch to central management for fleet-wide consistency.

WHAT EACH PARTNER BRINGS

AMD Instinct GPUs: High-memory-bandwidth accelerators sized for local inference performance and headroom.

Supermicro AS -8126GS-TNMR: Pre-integrated 8×MI325X server platform, ready for team-scale inference workloads.

Spectro Cloud PaletteAI Inference Launchpad: The routing, metering, and governance layer that turns silicon and server into a controlled inference platform.

Talk to your AMD, Spectro Cloud, or Supermicro contact about evaluating your current AI coding costs and scheduling a demonstration of the Instinct Coder Solution.

*All performance and/or cost savings claims are provided by Spectro Cloud and have not been independently verified by AMD. Performance and cost benefits are impacted by a variety of variables. Results herein may not be typical. GD-181a.