

AMD INSTINCT™ CODER 355

FRONTIER-CLASS AGENTIC CODING, ON OWNED HARDWARE, AT A FRACTION OF THE TOKEN BILL

EXECUTIVE SUMMARY

One of the most significant uses of AI today is writing code. Most of this use is done today with hosted, proprietary frontier models billed by the token. That bill is small in a pilot and steep in production, because agentic coding multiplies token usage. Unlike traditional AI coding, where a prompt typically generates a single model response, agentic coding can trigger dozens or hundreds of automated model calls as the agent reads, drafts, tests, and retries. The more your developers lean on the tools, the faster the meter spins, and the harder it gets to forecast or govern.

AMD Instinct™ Coder 355, a joint effort from AMD, Spectro Cloud, and Dell Technologies, runs most coding requests on GLM-5.3 on hardware you own, and calls out to a hosted frontier model only for the small share that needs one. AMD supplies the compute foundation with AMD EPYC™ CPUs and AMD Instinct™ MI355X GPUs. Spectro Cloud supplies the operating layer with PaletteAI, managing the Kubernetes-based AI platform across deployment, operations, updates, and lifecycle management. Inference Launchpad's smart router keeps everyday inference on the server while metering and governing every token. Dell Technologies delivers it as a tested, validated, and bundled solution on a single Dell PowerEdge XE9785 server, which can be deployed in a mainstream data center.

The result is the frontier experience developers expect, with the cost control and data residency the enterprise needs, built on open software and open-weight models. The same platform also carries a team from code through test and deployment, so the whole workflow runs in one place.

YOUR COSTS MOUNT AT THE TOKEN METER

Every AI coding assistant call is a metered transaction. Even a single developer running an agent that cycles through the model hundreds of times a day runs up real costs. Multiply that across a few hundred developers, and the usage curve outpaces any budget set for it. In application development, cost is now the most cited limiting factor. Most organizations cannot see their spending clearly enough to control it.

About 44% of organizations now name token and compute cost as the single biggest factor limiting further AI use in software engineering.[1]

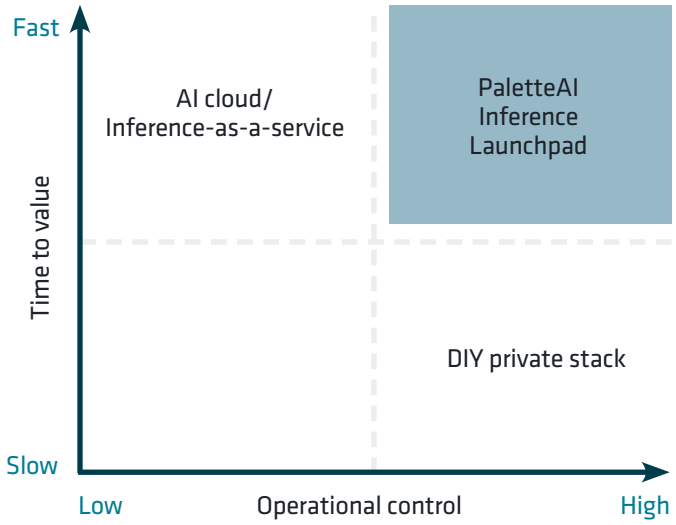
The value of AI-assisted coding is too high to give up, so teams look for control instead. Today, they face two unsatisfying options. Building an on-premises AI service yourself gives maximum control but is slow to stand up and heavy to operate. Renting from a hyperscaler or neocloud means an organization can be up and running in days, but every request still costs, the stack isn't its own, and code and data leave the enterprise perimeter, raising security and sovereignty questions. Neither lets an organization maintain the frontier developer experience while regaining control of costs.



THE OPTIMAL PATH

The Instinct Coder solution is the optimal path: validated, optimized, and pre-configured for rapid time-to-value and fully private for tight control of cost, data, and operations. It ships as a single server that can be stood up in days rather than the months a do-it-yourself build takes, and it does not require a specialist AI infrastructure team to run: the platform handles the GPU enablement, drivers, and day-2 operations that would otherwise need scarce in-house expertise.

From that point forward, everyday inference runs on the customer's own GPUs, inside its own perimeter, at a predictable cost. The open models on the server handle the everyday requests; a hosted frontier model is no longer the default that quietly drains the budget. For teams under strict data rules, routing policy can keep sensitive requests on the server, so the model, the code, and the data stay under the organization's control.



WHAT AMD BRINGS: THE COMPUTE FOUNDATION

AMD EPYC™ CPUs and AMD Instinct™ MI355X GPUs provide the compute foundation for running large open-weight models on-premises, delivering the responsiveness and model quality developers expect from hosted services on infrastructure the enterprise owns. Each AMD Instinct MI355X GPU carries 288GB of HBM3E memory. Within the complete software stack delivered by Spectro Cloud, Spectro Cloud and the GLM-5.3 model provide the AI coding assistance. Developer tools connect directly to this ready-to-use, optimized service, eliminating the need to build and tune one from the ground up.

WHAT SPECTRO CLOUD BRINGS: THE OPERATING LAYER

A rack of GPUs does not serve a developer on its own. Spectro Cloud turns the hardware into a self-service coding platform, built in two layers that sit together.

PaletteAI builds and operates the Kubernetes foundation. It treats the raw Dell PowerEdge server like a cloud, managing everything from the operating system up with no hypervisor to license, and stands up full-featured virtual clusters in minutes rather than the days it takes for a manual bare-metal build. That turns the solution into genuine self-service for developers: they spin up a full virtual cluster to test code or run a microservice, then tear it down again, without filing a ticket or waiting on the platform team. Each virtual cluster carries its own dedicated API server rather than sharing a namespace, allowing multiple developers to work concurrently.

Cluster Profiles capture a known-good configuration, the operating system, the Kubernetes version, the drivers, and the AI stack, so the next environment is a copy rather than a rebuild, and developers write code against a faithful replica of production instead of hitting "it worked on my machine" at deployment. For regulated teams, PaletteAI runs fully on-premises or air-gapped, with built-in local artifact registries and compliance guardrails.

PaletteAI Inference Launchpad is the inference control plane. At its center is a multi-model smart router that evaluates each request for sensitivity, task type, domain, and policy, then directs it to the appropriate model. Everyday coding requests run locally on the GLM-5.3 model using the AMD Instinct MI355X GPUs. When a task or policy requires a hosted proprietary frontier model, the router sends the request through a cloud API. Token metering tracks usage and cost by team, workload, and model, while governance policies enforce quotas and access controls.

Developers keep the tools they already use: Claude Code, Cursor, VS Code, Helix, and Codex, connected in a three-line configuration. Platform teams run it all from a single console that manages a fleet from one server up to a thousand clusters, with role-based access control and routing policies that propagate everywhere at once. Observability rides on standard Grafana and Prometheus dashboards.



WHAT DELL TECHNOLOGIES BRINGS: THE VALIDATED SERVER PLATFORM

Dell Technologies provides the PowerEdge XE9785, a fully configured GPU server that Spectro Cloud has validated as the hardware platform for the solution, backed by Dell ProSupport with next business day onsite service and Dell field deployment services. The Spectro Cloud software stack will be integrated through the delivery partner selected for the final deployment model. Customers can begin with a single server and add capacity as demand grows, while maintaining consistent platform management, upgrades, and drift control across the environment.

THE AMD INSTINCT™ CODER 355

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

Hardware	**Dell PowerEdge XE9785 10U server:** • 2 x AMD EPYC™ 9575F, 64 cores, 3.3GHz CPUs • 8 x AMD Instinct™ MI355X, 288GB HBM3E each • 3TB (24 x 128GB) DDR5 RDIMM 6400MT/s • 2 x 960GB M.2 boot drives on BOSS-N1 (RAID 1) • 8 x 7.68TB NVMe E3.S Gen5 SSD • 2 x Broadcom 57608 400GbE NIC • 12 x 3200W redundant (6+6) Titanium power supplies (3200W at 216 to 240Vac) • iDRAC10
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	• Local model GLM-5.3 • Frontier models Anthropic Claude, OpenAI GPT, or Google Gemini
Developer tools	Claude Code, OpenAI Codex, Visual Studio Code, Cursor
Support	• Comprehensive full stack software support from Spectro Cloud • Dell ProSupport with next business day (NBD) onsite service for 36 months for hardware
Scale	Supports up to 85 developers per node (50 concurrent)
Performance	Up to 98% of Frontier model benchmark performance[2]

$2.04M
current spend

~5-MONTH
payback

75%
annual savings

THE NUMBERS

The solution is built to deliver the responsiveness and model quality developers get from hosted services, at a fraction of the cost. The local experience reaches up to 98% of the benchmark performance of a leading hosted proprietary frontier model on coding tasks[2], so most work never needs to leave the server. A single server can support up to 85 users, 50 concurrently.

From a CFO's perspective, the illustrative model is compelling. For a team of 85 developers whose cloud inference costs roughly $2.04M a year, moving everyday work on-premises with a 10% hosted-frontier fallback can cut the annual cost by up to 75% compared with frontier-only use, on the order of $1.5M a year, providing payback in approximately 5 months.[3]

WHAT YOU GET

Cost

Everyday coding runs locally at a predictable price, while hosted frontier-model usage is reserved for the requests that require it. Token metering provides visibility into usage and costs by team, workload, and model, supporting budgeting, chargeback, and cost governance.

Speed

Deploy in days rather than months. Automated Day 2 operations and lifecycle management reduce the effort required for upgrades, maintenance, and platform management, allowing development teams to focus on delivering products.

Control

Sensitive code and user data stay inside the perimeter, and routing follows policy, intent, and sensitivity rather than habit.

WHO IT IS FOR

Three roles usually feel this problem together. The solution is designed to give all three what they need from a single server.

CIO or IT leader: Owns the cost and governance question and wants enterprise AI spend under control without another vendor lock-in.

Platform leader: Delivers cheaper, governed inference across teams without committing to a long build and a heavy Day-2 burden.

Engineering leader: Wants to improve developer productivity with a consistent, uninterrupted AI-assisted coding experience while keeping token costs visible, code context protected, and model quality consistent.



WHERE IT FITS BEST

The solution is especially well suited to environments where code and data must remain under strict organizational control. Banking, health technology, defense, and federal teams can deploy AI coding assistance inside private or air-gapped environments with local security tooling, compliance guardrails, and repository context kept within the enterprise boundary. Inference Launchpad also supports bring-your-own-model deployments, allowing organizations to use Spectro Cloud-certified models validated for AMD infrastructure or their own approved models. The same architecture can support any enterprise seeking greater visibility and control over agentic-coding costs.

NEXT STEPS

Establish a cost baseline.

Use current platform and billing data to understand token consumption, hosted model spend, and which teams or workloads drive it.

Size the local opportunity.

Determine which workloads can run locally based on task complexity, policy requirements, and acceptable model performance, then estimate the potential savings for the proposed server configuration.

Validate the business and technical fit.

Confirm that the expected workload mix, model performance, security requirements, and cost profile support deployment of the solution. Evaluate any available sandbox or assessment environment once the delivery model is finalized.

Plan the production deployment.

Define the initial server configuration, integration approach, governance model, and capacity plan based on validated demand and operational requirements.

Talk to your AMD, Spectro Cloud, or Dell Technologies contact about evaluating your current AI coding costs and scheduling a demonstration of AMD Instinct™ Coder 355.



ABOUT AMD

More than 50 years driving innovation in high-performance computing. AMD Instinct accelerators power AI workloads for leading cloud providers, enterprises, and sovereign AI initiatives worldwide. The AMD commitment to open standards and open ecosystem ensures enterprises can deploy AI with control, flexibility, and long-term vendor independence.

ABOUT SPECTRO CLOUD

Spectro Cloud helps platform teams and cloud providers modernize and manage infrastructure for the AI era without adding more tools or operational complexity. With PaletteAI, enterprises, public sector organizations, neoclouds and sovereign clouds can build, govern, and operate full-stack environments across VMs, Kubernetes, edge, regulated and air-gapped locations, and AI infrastructure.

ABOUT DELL TECHNOLOGIES

Dell Technologies (NYSE: DELL) helps organizations and individuals build their digital future and transform how they work, live and play. The company provides customers with the industry's broadest and most innovative technology and services portfolio for the AI era.

1 Futurum Software Lifecycle Engineering Decision Maker Survey, 2H 2026. Futurum Group. Fielded June 2026.

2 TKS-001 - Based on AMD analysis of a modeled deployment using one server configured with 8 AMD Instinct™ MI355X GPUs. Assumptions include 85 developers spending $2,000 per month each on AI coding/API services ($2.04 million annually); Server list price of $590,000 including software and licensing with a three-year hardware amortization; continued frontier cloud API costs equal to 10% of baseline AI coding/API spend ($204,000 annually); and $50,000 in annual data center operating costs. Under these assumptions, modeled annual costs are approximately $513,500, representing approximately 75% of annual cost savings. The estimated five-month payback period is based on the modeled initial investment, relative to modeled monthly savings. Results are illustrative only. Actual costs, savings, and payback will vary based on workload, utilization, pricing, routing, infrastructure costs, and other factors.

3 On Vals AI SWE-bench Verified leaderboard, updated September 1, 2026, a GLM-5.3 model vs. the Claude Opus 5 model resolved 95.4% and 97.0% of benchmark tasks, respectively, using the Mini-SWE-agent harness. Relative score calculated as 95.4 ÷ 97.0 = 98.4%. Results are specific to the benchmark and evaluation configurations and may not reflect performance on other workloads. For more information see https://www.vals.ai/comparison?m=claude-opus-5,glm-5.3&b=swebench.

©2026 Advanced Micro Devices, Inc. all rights reserved. AMD, the AMD arrow, and combinations thereof, are trademarks of Advanced Micro Devices, Inc. Spectro Cloud, Palette, PaletteAI, and PaletteAI Inference Launchpad are trademarks of Spectro Cloud in the US and other countries. Dell Technologies, Dell, EMC and Dell EMC are trademarks of Dell Inc. or its subsidiaries. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

PID: 265684302