TECHNICAL WHITE PAPER

Vertical AI Jumpstart Services in a GB10

Bridging the Gap Between AI Ambition and AI Adoption

Business Solutions
1.800.800.0014

Enterprise Solutions
1.800.369.1047

Public Sector Solutions
1.800.800.0019

www.connection.com/helix

CONTENTS

1. The NVIDIA Grace Blackwell GB10 and CNXN Helix Jumpstart Services

Every organization wants AI. Few know where to begin. The gap between AI ambition and AI adoption has become one of the defining challenges of the enterprise technology landscape. **Vertical AI Jumpstart in a GB10**—CNXN Helix's AI jumpstart program with proven vertical use cases and delivery—was built to close that gap.

A New Category of AI Infrastructure

The NVIDIA DGX Spark, powered by the GB10 Grace Blackwell Superchip, represents a fundamentally new category of AI infrastructure. It places a full AI supercomputer—capable of up to 1 petaFLOP of FP4 AI performance—into a compact desktop form factor that runs on a standard wall outlet. There are no racks to install, no cooling systems to engineer, no electrical upgrades to commission. It sits on a desk, plugs into the wall, and delivers the same NVIDIA software stack that powers the world's largest AI data centers.

Available from leading OEMs including Dell, HPE, and Lenovo, the GB10 integrates a 20-core NVIDIA Grace ARM CPU with a Blackwell GPU featuring fifth-generation Tensor Cores—connected through NVLink-C2C for a coherent, unified 128GB memory architecture. Two units can be clustered via QSFP cables and NVIDIA's NCCL (pronounced like "nickel") to deliver 256GB of combined memory and support models up to 405 billion parameters. The device ships with DGX OS (a slightly modified Ubuntu-based to deal with some of the lagging driver support issues from the broader ecosystem), Docker pre-installed, and the full NVIDIA AI software stack including NGC container registry access, CUDA libraries, NIM microservices, and support for leading open models.

This is not a toy or a proof-of-concept device. The GB10 is production-capable for inference, fine-tuning, and RAG workloads at the departmental level. Organizations can run state-of-the-art reasoning models from NVIDIA (Nemotron), Meta (Llama), Google (Gemma), DeepSeek, Qwen, and others—locally, privately, and without cloud dependencies.

The CNXN Helix Difference: Outcome-based, Not Box-based

Hardware alone rarely translates to business outcomes. This is the central insight behind CNXN Helix's approach. While any reseller can ship a GB10 unit, **CNXN Helix delivers a working, industry-specific AI solution with the client's own data—typically in a 2-day onsite engagement**.

Our Vertical AI Jumpstart Services pair with the GB10 hardware to create a complete, outcome-based solution. Approximately 20 hours of professional services cover the full engagement lifecycle: use case selection from a curated menu of industry-specific solutions, planning and data pipeline preparation, device setup and configuration, AI model deployment (including RAG, LoRA fine-tuning, and NeMo Guardrails), end-user training, and project management. The client walks away with a functioning AI assistant—built on their data, running in their environment, solving a real business problem.

This distinction is critical for sellers and clients alike. CNXN Helix is not positioning the GB10 as a device sale. It is positioning it as an AI adoption platform—a right-sized entry point that proves value, builds internal competency, and creates a clear expansion path to data center-scale NVIDIA infrastructure.

2. The Challenge of Getting Started with AI

AI adoption is accelerating across every industry and company size—but the gap between interest and execution remains substantial. Understanding the barriers organizations face is essential to understanding why the GB10 with CNXN Helix approach matters. We like to think of this as the chasm between those leading with AI adoption and experimentation (think frontier model makers and large enterprise organizations) and everyone else.

The Numbers Tell the Story

According to Deloitte's State of AI in the Enterprise 2026 report, based on a survey of over 3,200 senior leaders across 24 countries, improving productivity and efficiency is the leading benefit organizations are achieving from AI—with two-thirds (66%) reporting gains. But revenue growth remains aspirational: 74% of organizations hope to grow revenue through AI, while only 20% are doing so today. The report identifies insufficient worker skills as the single biggest barrier to integrating AI into existing workflows.[1]

56% of CEOs report getting "nothing" from their AI adoption efforts.

The data from other sources paints a similar picture. McKinsey reports that 78% of companies now use AI in at least one function—up from 55% in 2023—but only about 1% describe themselves as "mature" in AI deployment, meaning AI is fully embedded and producing major business outcomes.[2] The OECD's 2025 data shows enterprise AI adoption roughly doubled in two years (from 8.7% to 20.2% of firms globally), yet a steep gap persists between large enterprises (55% in the EU) and small firms (17%).[3]

Perhaps most striking: PwC's 2026 Global CEO Survey found that 56% of CEOs report getting "nothing" from their AI adoption efforts.[4] And MIT's research suggests that 95% of generative AI pilots fail to move beyond the experimental phase.[5] These are not statistics about AI skeptics—they describe organizations that invested in AI and still couldn't cross the gap from pilot to production.

Five Barriers That Keep Organizations Stuck

Perceived Cost and Complexity. When leaders hear "AI infrastructure," they picture massive GPU clusters, multi-million-dollar capital outlays, and specialized data center environments. For small- to mid-sized businesses, education, state and local government, and individual business units within enterprises, the assumption is that AI is simply out of reach. The reality is that inference-focused workloads—the vast majority of enterprise AI applications—can run effectively on a single desktop-class GPU with sufficient memory. The GB10 was designed precisely for this sweet spot.

Lack of In-house AI Expertise. Deloitte's survey confirms that the AI skills gap is the top barrier to integration.[1] Most organizations lack the data scientists, ML engineers, and DevOps talent needed to select models, build data pipelines, configure guardrails, and deploy production-ready AI applications. Even organizations with technical staff often lack experience with the specific NVIDIA ecosystem—CUDA, NIM, NeMo, Nemotron—that powers modern enterprise AI. Without expert guidance, expensive hardware sits idle and projects fail to achieve intended outcomes.

The GB10 offers a fully local AI platform that satisfies the most stringent data sovereignty requirements.

No Clear Starting Point. Organizations struggle to identify a specific, high-value AI use case that justifies investment. Without a defined business problem, a measurable outcome, and a clear deployment path, AI initiatives stall in committee. Research from the UK's Office for National Statistics found that 39% of enterprises that considered AI but cited difficulty identifying relevant use cases as a primary barrier.[6] Canada's business surveys show 78% of non-adopters simply believe AI is "not relevant" to their business[7]—a perception problem, not a technology problem. CNXN Helix can help with this, and it's precisely why we created the Vertical Jumpstart program to be led by business outcomes.

Data Privacy, Compliance, and Sovereignty Concerns. Regulated industries like healthcare (HIPAA, HITECH), education (FERPA), government (CMMC, FedRAMP), and manufacturing (CGMP, ITAR) all face constraints on where their data can reside and how it can be processed. Cloud-based AI services create compliance risk when sensitive data leaves the organization's network. Many organizations are mandating (or at least comfortable starting with) on-premises AI that keeps data within their physical boundaries, but traditional on-premises GPU infrastructure is prohibitively complex. The GB10 offers a fully local AI platform that satisfies the most stringent data sovereignty requirements.

Cloud Cost Uncertainty. Even organizations without strict compliance requirements face legitimate concerns about the economics of cloud AI. GPU-as-a-service pricing is unpredictable, usage-based billing makes budgeting difficult, and latency for real-time inference applications can be unacceptable. A GB10 unit is a one-time purchase with no recurring compute costs—the inference runs locally, unlimited, for as long as the hardware is in service. You can process millions of input and output tokens daily without overages or unexpected bills.

The GB10 directly addresses each of these five barriers: affordable hardware eliminates the cost barrier; CNXN Helix Jumpstart Services eliminates the expertise barrier; our pre-built vertical use cases eliminate the "where to start" confusion; on-premises deployment reduces compliance risk, and a one-time purchase eliminates cloud cost uncertainty.

3. The Hidden Costs and Complexity of AI Data Center Deployments

For organizations prepared to invest heavily in AI, the traditional data center path introduces challenges that are consistently underestimated. The scale of the global AI infrastructure buildout provides important context for understanding why a desktop-first approach makes strategic sense when you're just starting out.

The Trillion-dollar AI Data Center Boom

The numbers are staggering. McKinsey projects that companies will need to invest $5.2 trillion into AI data center infrastructure by 2030 to meet worldwide demand.[8] Gartner estimates $582 billion will be invested in AI infrastructure in 2026 alone—a 19% increase over 2025.[9] In 2025, tech giants collectively spent roughly $580 billion turning empty fields and abandoned factories into GPU clusters, with Microsoft dedicating $80 billion and Amazon allocating $86 billion for AI infrastructure.

These are the budgets of the world's largest technology companies. For a mid-market enterprise, a healthcare system, a school district, a state agency, or a small manufacturing company, the economics of AI data center deployment look very different—and significantly more daunting.

The Real Cost of an Enterprise AI Deployment

Industry analysts estimate that a 100-megawatt hyperscale data center facility costs $900 million to $1.5 billion to construct, based on an industry average of $9–15 million per megawatt. But most enterprises aren't building hyperscale facilities. They're trying to add AI capability to existing infrastructure within their data center or data closets—and the per-unit costs are equally eye-opening.

GPU Server Hardware. A single enterprise AI server node (such as an NVIDIA DGX or OEM equivalent) costs $200,000–$500,000+. Multi-node clusters with InfiniBand networking easily reach $1–5 million before software licensing. The average cost per AI-capable rack is expected to reach $3.9 million in 2025.

Power Infrastructure. AI servers demand far more power than traditional IT infrastructure. A single AI training workload requires approximately 30 megawatts of continuous power. GPU rack densities have climbed from 40 kW per rack to 130 kW—and are projected to reach 250 kW by 2030. Most enterprise server rooms were built for 5–15 kW per rack. Upgrading to AI-grade power often requires 3-phase 480V electrical feeds, new transformers, upgraded switchgear, and dedicated UPS/PDU infrastructure.

73% of new AI facilities now deploy direct-to-chip or immersion cooling systems.

Cooling Systems. Traditional air cooling is becoming obsolete for AI workloads. AI servers generate up to 1.5 kilowatts of heat per chip—far exceeding conventional cooling capacity. Liquid cooling is roughly 3,000 times more efficient than air cooling for AI hardware, and 73% of new AI facilities now deploy direct-to-chip or immersion cooling systems. Retrofitting an existing server room for liquid cooling is a major construction project.

Specialized Networking. High-speed GPU-to-GPU fabrics (InfiniBand, NVLink Switch systems) require specialized switches, cabling, and network engineering not found in conventional enterprise IT environments. This networking layer alone can cost hundreds of thousands of dollars and requires dedicated expertise to design and maintain.

Facility Upgrades. Beyond the IT stack, AI data center deployments often require structural upgrades: reinforced flooring to support high-density rack weights, enlarged cooling plant capacity, fire suppression modifications, and increased generator capacity for backup power.

Lead Times. Enterprise GPU server lead times range from 12–26+ weeks. Combined with facility preparation, permitting, and construction, an AI data center deployment can take 6–18 months from decision to first workload. In some regions, utilities have paused new power connections entirely—Northern Virginia, the largest data center market in the world, experienced utility connection freezes in 2025–2026 due to grid strain.

Ongoing Operations. Data center AI deployments require dedicated teams for thermal management, power monitoring, firmware updates, driver maintenance, and cluster orchestration—adding significant recurring operational expense on top of the capital investment.

The Compounding Effect

When you combine these costs, the barrier to entry for on-premises AI becomes clear—and daunting for most just starting out on their AI journey. An organization that wants to deploy a modest AI capability—say, a fine-tuned large language model for internal knowledge retrieval—might face a total cost of $500,000 to $2 million for hardware, infrastructure upgrades, and the first year of operations. For a small manufacturer, a county government, or a community college, this is budget they simply don't have.

The result is a dangerous bifurcation in the market: well-funded enterprises and hyperscalers race ahead with AI, while the vast majority of organizations—the SMBs, public institutions, and individual business units that make up the backbone of the economy—fall further behind.

This is the AI adoption chasm that the GB10 was built to bridge.

The GB10 Alternative: Deploy in Days, Not Months

A GB10 Jumpstart package gets an organization a working AI use case in approcimately two business days.

The GB10 deploys in hours on a standard desk with standard power. It requires no specialized racks, no cooling infrastructure, no electrical upgrades, no facility construction, and no dedicated operations staff. Use it as a rapid starting point to prove AI value—while simultaneously planning or justifying a larger data center investment. The two paths are not mutually exclusive. They're complementary.

A GB10 Jumpstart package—including OEM hardware, peripherals, and full CNXN Helix deployment services—starts at a fraction of the cost of a single enterprise GPU server. The organization gets a working AI use case, set up with their own data, operating in their own environment, in approximately two business days. Compare that to the 6–18-month timeline and multi-million-dollar investment required for a data center deployment, and the strategic value of the GB10 as a starting point becomes obvious.

4. Inside the NVIDIA Grace Blackwell GB10

Understanding the hardware is important for stakeholders evaluating the GB10 as an AI platform. The GB10 is not a consumer device repurposed for AI—it is a purpose-built AI supercomputer miniaturized into a desktop form factor.

The Grace Blackwell Superchip Architecture

The GB10 includes an NVIDIA Blackwell GPU with fifth-generation Tensor Cores and RT Cores, delivering up to 1 petaFLOP of FP4 sparse performance across 6,144 CUDA cores with 24MB of L2 cache. The CPU is an NVIDIA Grace 20-core ARM processor with 10 Cortex-X925 performance cores and 10 Cortex-A725 efficiency cores. The system features 128GB of coherent unified LPDDR5x memory shared seamlessly between CPU and GPU via NVLink-C2C, with approximately 600GB/s of aggregate GPU bandwidth. Storage is up to 4TB NVMe SSD.

Networking includes dual NVIDIA ConnectX-7 QSFP56 ports providing 200GbE aggregate bandwidth, plus Realtek 10GbE and WiFi 7. Two units can be linked via QSFP DAC cables for a 256nGB combined memory cluster supporting models up to 405 billion parameters. The entire system runs on DGX OS (Ubuntu-based)—the same operating system distribution as enterprise

DGX platforms—and draws power from a standard wall outlet with a SoC TDP of approximately 140 watts.

The most impressive part of this device: it fits into the palm of your hand —measuring roughly 6” x 6” x 2”.

What Makes the Grace Blackwell Architecture Unique

The GB10 architecture’s defining advantage is its coherent unified memory model. Unlike discrete GPU systems—where data must be copied between system RAM and VRAM through a PCIe bottleneck—the Grace Blackwell Superchip provides a single memory address space accessible to both CPU and GPU simultaneously via the NVLink-C2C interconnect. This eliminates the data transfer overhead that limits discrete GPU systems and enables the GB10 to load and run large language models far larger than what typical discrete GPUs with comparable VRAM can handle.

The GB10 can load and run large language models far larger than what typical discrete GPUs with comparable VRAM can handle.

To illustrate: an NVIDIA RTX 5090—a $2,000+ consumer GPU—has 32GB of VRAM and can only run models that fit within that memory. The GB10, with 128GB of unified memory, can run models 4x larger. While the RTX 5090 may be faster for small models due to higher memory bandwidth via GDDR7, the GB10 excels at running the large, enterprise-grade models—70B, 120B, even 200B parameters—that organizations actually need for complex reasoning, domain-specific knowledge retrieval, and production inference workloads.

In a 2-node cluster configuration (256GB combined), the GB10 can run models up to 405 billion parameters—including the full Llama 3.1 405B and comparable frontier models. This capability is extraordinary for a desktop device and positions the GB10 as a legitimate development, testing, and departmental production platform.

Software Compatibility: Desktop to Data Center

Critically, **what works on the GB10 will work on NVIDIA data center equivalents**. The Grace Blackwell architecture spans from desktop (GB10) to server (GB200/GB300) to full AI factory scale (DGX SuperPOD). The same CUDA libraries, NIM microservices, NeMo frameworks, NGC containers, and AI models that run on a GB10 can be deployed to enterprise NVIDIA infrastructure without re-engineering—making the GB10 both a practical production tool and a strategic evaluation platform.

This is not an incremental feature—it is the strategic foundation of the entire GB10 value proposition. An organization that prototypes and validates an AI use case on a GB10 can scale that exact solution to data center infrastructure when ready. The GB10 is the on-ramp to the full NVIDIA enterprise AI stack and what makes it so appeal to help our customers jumpstart their AI journeys.

5. CNXN Helix Vertical AI Jumpstart Services

Hardware without deployment expertise is a recipe for shelfware. This is a lesson the technology industry has learned repeatedly—from the early days of ERP implementations to the cloud migration era. AI is no different. In many ways, the risk is greater because the technology is newer, more complex, and the talent pool shallower.

CNXN Helix's Vertical AI Jumpstart Services exist to ensure that every GB10 deployment translates directly into a working, measurable business outcome. The services are not optional add-ons—they are the core differentiator that separates CNXN Helix from every other reseller shipping NVIDIA hardware.

The Services Engagement Model

Each Jumpstart engagement follows a structured delivery process designed for repeatability, quality, and speed. The typical engagement spans approximately 20 hours of professional services, including a two-day onsite deployment.

- **Phase 1: Use Case Selection and Scoping**—CNXN Helix consultants work with client stakeholders to identify the right AI use case from a curated set of proven, industry-specific solutions. Rather than asking the client to define their own AI use case—a common failure point—CNXN Helix presents a curated menu of pre-built, industry-specific use cases, each demonstrated through short videos showing the solution running on GB10 hardware with real data.
- **Phase 2: Planning and Data Preparation**—CNXN Helix engineers collaborate with client IT and data teams to prepare source data, ensuring the AI solution is tailored to the organization's unique context and knowledge.
- **Phase 3: Onsite Deployment (Two Days)**—CNXN Helix engineers deploy and configure the GB10 hardware, integrate client data, and bring the AI solution to production readiness with end-to-end validation.
- **Phase 4: Training and Handoff**—The client receives hands-on training, architecture documentation, and guidance to independently operate, maintain, and expand the solution.

CNXN Helix Jumpstart Services span six strategic verticals, plus cross-industry solutions.

Industries and Use Cases Served

CNXN Helix Jumpstart Services span six strategic verticals, plus cross-industry solutions, each with a growing library of pre-built, deployable use cases:

- **Education (K–12)**—AI tools supporting student services, regulatory compliance, multilingual communication, and staff productivity
- **Higher Education**—Solutions for student support, research environments, and campus-wide AI enablement
- **State and Local Government**—AI assistants for public-facing services, policy navigation, compliance, and multilingual access
- **Healthcare**—Compliance, clinical documentation, and workflow automation—all designed for strict data sovereignty with no cloud dependencies
- **Manufacturing**—Regulatory compliance, quality control, maintenance support, and engineering documentation AI
- **Retail**—Employee onboarding, associate support, and omnichannel data integration for in-store AI experiences
- **Cross-industry**—Cybersecurity compliance, code review, IT service management, and corporate knowledge management

Why Services Matter

The case for bundling services with hardware is not just intuitive. The AI skills gap—identified as the number one barrier globally—means that most organizations cannot self-deploy effectively, regardless of the quality of the hardware.

CNXN Helix's Jumpstart model directly addresses this: expert engineers handle the technical complexity, the client's team provides domain knowledge and data access, and the result is a working AI solution that demonstrates tangible value within days. This "first AI win" is often the catalyst that unlocks executive sponsorship, departmental expansion, and eventually the data center-scale AI investment that benefits the entire organization.

Nearly 80% of organizations have adopted AI in some capacity, yet fewer than 1% have scaled it to maturity.

One successful GB10 deployment often unlocks additional opportunities across departments. The key is to start focused, deliver value, and expand strategically.

6. Conclusion: Start with GB10 and Helix—Lead with AI

The enterprise AI landscape in 2026 is defined by a paradox: interest is universal, but execution remains rare. Nearly 80% of organizations have adopted AI in some capacity, yet fewer than 1% have scaled it to maturity. The barriers—cost, complexity, expertise, compliance, and the absence of a clear starting point—are well understood. What has been missing is a practical, accessible solution that addresses all of them simultaneously.

The NVIDIA Grace Blackwell GB10, paired with CNXN Helix Vertical AI Jumpstart Services, fills that gap.

What the GB10 + Helix Partnership Delivers

- **Immediate AI Capability.** A production-grade AI supercomputer on your desk, running the same NVIDIA stack used by the world's leading AI organizations—deployed and operational in days, not months.
- **Proven Business Outcomes.** A working, industry-specific AI use case built on your own data, not a generic demo. Clients walk away from the Jumpstart engagement with a solution that solves a real business problem and produces measurable results.
- **On-premises Data Sovereignty.** Your data never leaves your building. No cloud dependencies, no compliance risk, no recurring compute costs. The GB10 satisfies the most stringent requirements for HIPAA, FERPA, CMMC, and other regulatory frameworks.
- **Organizational AI Competency.** Jumpstart Services include training and enablement that builds internal capability. Your team learns the technology, understands the architecture, and gains confidence to operate and extend the solution independently, while gathering buy in across the organization.
- **A Clear Path to Scale.** The GB10 is the entry point to the full NVIDIA ecosystem. Models, code, and workflows developed on the GB10 transition seamlessly to GB300 desktop supercomputers, HCI edge clusters, and full AI factory data center deployments. There is no replatforming, no vendor lock-in, and no wasted investment.
- **A Trusted Technology Partner.** CNXN Helix serves as a single point of accountability for hardware, software, services, and scale—from the first GB10 deployment to enterprise-wide AI transformation.

Start Smart, Scale Confidently

The GB10 is the first step on a well-defined NVIDIA infrastructure scale path. Organizations begin with a 1-node or 2-node GB10 deployment to adopt AI and prove value with CNXN Helix Jumpstart Services. Ultimately, organizations can scale to full AI factory data center deployments using NVIDIA DGX or OEM equivalents for enterprise-wide training and inference as adoption drives a need for more demanding compute infrastructure.

The GB10 investment is additive—it builds toward the organization's long-term AI infrastructure rather than detracting from it.

Ready for Your First AI Win?

Contact your Account Manager today to learn more about our Jumpstart Services.

CNXN HELIX CENTER FOR APPLIED AI AND ROBOTICS | Connection we solve IT

Business Solutions
1.800.800.0014

Enterprise Solutions
1.800.369.1047

Public Sector Solutions
1.800.800.0019

www.connection.com/helix

1 https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html

2 https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

3 https://www.oecd.org/en/about/news/announcements/2026/01/ai-use-by-individuals-surges-across-the-oecd-as-adoption-by-firms-continues-to-expand.html

4 https://www.pwc.com/gx/en/news-room/press-releases/2026/pwc-2026-global-ceo-survey.html

5 https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/

6 https://www.ons.gov.uk/economy/economicoutputandproductivity/productivitymeasures/articles/managementpracticesandtheadoptionoftechnologyandartificialintelligenceinukfirms2023/2025-03-24

7 https://www150.statcan.gc.ca/n1/pub/11-621-m/11-621-m2025011-eng.htm

8 https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers

9 https://www.gartner.com/en/newsroom/press-releases/2025-10-22-gartner-forecasts-worldwide-it-spending-to-grow-9-point-8-percent-in-2026-exceeding-6-trillion-dollars-for-the-first-time