Jensen Huang and the Billion-Fold Future of Intelligence

Published October 9, 2025

Christopher Gannatti, CFA
Global Head of Research

Key Takeaways

- NVIDIA's $100 billion partnership with OpenAI signals a paradigm shift, as demand for AI compute infrastructure surges beyond even the boldest forecasts.
- Jensen Huang argues that intelligence compounds, with AI not replacing human ingenuity but multiplying it, fueling explosive growth in jobs, ideas and economic output.
- Far from facing saturation, NVIDIA remains in persistent scramble mode, with accelerated computing replacing CPUs across industries and no sign of a capacity glut in sight.

Ray Kurzweil once predicted that the 21st century would compress **20,000 years of progress into a single century**.[1] Brad Gerstner, **founder, chairman and CEO** of Altimeter Capital, and NVIDIA CEO Jensen Huang cited this in a recent BG2 podcast conversation[2] to emphasize just how profoundly difficult it is for most people to grasp the velocity of technological change. Human intuition is poor at compounding systems—and even worse at exponential ones that accelerate with scale. What looks like incremental progress in a single year can, when stacked, represent civilizational leaps that would previously have taken millennia. Huang and Gerstner's point in bringing this into the discussion was not abstract futurism; it was a reminder that the rate of AI improvement is already outpacing traditional forecasting frameworks, leaving institutions perpetually underestimating the profound shift in capability.

The Expanding Frontier of Intelligence

That exponential framing connects directly to the debate over jobs and automation. Skeptical narratives assume a **finite pool of ideas**: if machines take over tasks, humans have less left to do. Huang flips that assumption in this discussion. The reality, he argues, is that intelligence itself is generative, not zero-sum. Each new intelligent system—whether human or artificial—creates *more* possibilities, not fewer. Just as the steam engine and the microprocessor opened vast new industries, AI creates categories of work and problem-solving that did not previously exist. To think otherwise is to assume that we have reached the end of imagination, a concept that Huang flatly rejects.

Inside NVIDIA, he uses this principle as operational proof. Every engineer, chip designer and software developer at the company now works with AI models as copilots—what he described as **"100% coverage with AI."** Instead of reducing headcount, this integration has expanded it. The company is hiring more

people because augmented productivity opens the door to pursuing *more ideas*. AI is not reducing the need for human ingenuity; it is multiplying the number of projects, prototypes and explorations NVIDIA can undertake. The workforce grows because the frontier expands faster than human bandwidth alone can manage.

The larger insight is that **intelligence compounds**. Surround a group with more intelligence—be it in the form of talented colleagues or AI copilots—and their collective imagination broadens. Huang's conviction is that this dynamic will scale far beyond NVIDIA's labs. If the global economy is 55%–65% driven by human intelligence, a figure they cited as a rough estimate in the discussion, then multiplying the productivity of that intelligence via AI doesn't shrink opportunity; it unlocks trillions of dollars of additional output. Intelligence is not a scarce resource to be divided, but a catalytic one that expands the horizon of what is possible. In a century of 20,000 years of progress, that is the only mindset that makes sense.

The $100 Billion Investment in OpenAI

The centerpiece announcement in the BG2 conversation was NVIDIA's expanded partnership with OpenAI. Over the coming years, OpenAI will bring **10 gigawatts of AI compute capacity online**, roughly equivalent to the power output of 10 nuclear reactors. For NVIDIA, that translates into staggering demand. As Huang noted, if all of those data centers ran on NVIDIA systems, it could represent as much as **$400 billion in revenue** flowing through the company over time. Alongside this, NVIDIA committed to investing up to **$100 billion** directly into OpenAI, both as a technology partner and as a shareholder in what Huang described, in his informed opinion, as a near-certain "next multi-trillion-dollar hyperscaler."

What makes this arrangement unique is its dual nature: **customer and investee**. OpenAI is one of NVIDIA's largest customers, already consuming enormous amounts of graphics processing unit (GPU) capacity through Microsoft Azure, Oracle Cloud and CoreWeave. But now there is a shift. Instead of outsourcing entirely, OpenAI is now **building its own hyperscale AI factories**, in partnership with NVIDIA at every layer—chips, software, systems and end-to-end infrastructure. For Huang, this is a strategic inflection point: it cements NVIDIA not as a chip vendor but as a long-term infrastructure partner in the creation of one of the most consequential technology companies of the century.

Naturally, skeptics raised the question of "circular revenue." If NVIDIA sells billions of dollars of GPUs to OpenAI while simultaneously investing in the company, is this financial engineering rather than economic substance? Huang dismissed the concern outright. The capital that fuels OpenAI's buildout comes from **its own revenue growth, equity fundraising and participation in debt markets at its discretion**—not from NVIDIA's investment being recycled back as purchases. As Huang emphasized, "This is likely going to be the next multi-trillion-dollar hyperscale company, and who doesn't want to be an investor in that?" The demand is real, the economics are grounded in exponential usage growth, and for NVIDIA, the dual role of supplier and investor only deepens its position at the very heart of the AI revolution.

Intel + NVIDIA: A Fusion of Ecosystems

One of the more surprising elements of the conversation was Huang's openness toward **partnering with Intel**, a company that once tried to stamp NVIDIA out of existence. Far from holding a grudge, Huang

framed the collaboration around NVLink Fusion as pragmatic and mutually beneficial. NVIDIA is happy to fuse Intel's enormous enterprise central processing unit (CPU) ecosystem with its own accelerated computing stack, just as it partners with ARM and others. As Huang put it, **"the future is so much greater—it doesn't have to be all us or them, it can be us and them."** That attitude reflects both confidence in NVIDIA's competitive moat and a willingness to expand the factory model by integrating the best components from across the industry, rather than insisting on total exclusivity.

Unlocking the World's Data with Accelerated Compute

That spirit of openness—"us and them" rather than "us versus them"—sets the stage for another underappreciated frontier Huang highlighted: **data processing**. Today, the vast majority of the world's structured and unstructured data workloads, from SQL[3] queries to enterprise databases, still run on traditional CPUs. Huang argued that this is an enormous inefficiency waiting to be addressed. Just as recommender engines migrated from CPUs to GPUs, he sees a similar inevitability for the world's data pipelines: in the future, **AI will be the engine that processes data at scale**. This is not a niche market—structured and unstructured data represent the lion's share of global compute cycles. The discussion stated that moving that domain onto accelerated AI systems could unlock one of the next trillion-dollar transitions in the industry.

30xBetter in One Cycle: The Economics of Staying Ahead

One of the most common investor questions is whether specialized application-specific integrated circuits (ASICs) will one day displace NVIDIA's GPUs as the workhorse of AI. Huang's answer is unequivocal: it's not about a chip; it's about a **system**. Building an AI factory requires co-optimizing models, algorithms, software stacks, networking, memory and the silicon itself. This is what NVIDIA calls **extreme co-design**. In practice, it means that NVIDIA isn't just taping out a chip; it is engineering a coordinated ecosystem where every layer is tuned for performance. ASICs may win in narrow niches like video transcoding and smart NICs,[4] but for the ever-evolving backbone of AI workloads—training, reasoning, multimodality—programmability and co-design matter more than fixed-function efficiency.

The payoff of this system-level approach shows up in the numbers. Between Hopper and Blackwell, NVIDIA delivered a **30ximprovement in performance per watt in just one product cycle**. In a world where power, not chip count, is the limiting factor, that kind of leap is decisive. A hyperscaler with a two-gigawatt budget cannot afford to run older hardware that produces 30 times fewer tokens per unit of energy. Even if a competitor offered chips for free, the opportunity cost of lower performance would wipe out any savings. As Huang explained, customers will always choose the system that maximizes revenue per watt, because that is what ultimately drives returns on the tens of billions spent on land, power and data center shells.

This way of thinking about technology is foreign to most people. In the past, when nations built railroads or strung fiber-optic cables across oceans, the investment was essentially static—an upfront expense followed by decades of incremental use. The trains did not suddenly run 30 times faster in three years. With AI infrastructure, however, that is precisely what happens. The chips at the heart of these systems **depreciate quickly**, but not because they lose function; it's because new generations make them instantly

obsolete by offering orders-of-magnitude better performance. It is a capital cycle more akin to consumer electronics than to traditional infrastructure, but at a scale measured in gigawatts and trillions of dollars.

For Huang, this cycle reinforces NVIDIA's moat. By committing to annual release cadences and extreme co-design, NVIDIA ensures that each new generation—Blackwell Rubin Rubin Ultra Feynman—resets the performance curve. Customers are effectively locked into a treadmill of upgrades, not out of brand loyalty, but because the economics of falling behind are intolerable. In the long history of industrial revolutions, from railroads to telecommunications, few technologies have offered compounding leaps of this magnitude. That is why the GPU versus ASIC debate is, in Huang's view, a distraction: the real contest is who can orchestrate **system-wide leaps** at exponential cadence.

Scramble Mode: NVIDIA's Persistent Demand Gap

The most fundamental shift Jensen Huang described is the wholesale transition from **general-purpose computing to accelerated computing**. For decades, CPUs defined the backbone of digital infrastructure. But, as he put it, "general-purpose computing is over." Search, recommender systems and now vast swaths of enterprise workloads are migrating to GPUs because certain things—AI reasoning, multimodal inference and large-scale data processing—**simply cannot be done** without accelerated compute. This is not a theoretical future; it is happening now, and Huang sees it continuing over the coming years until every layer of IT refresh is designed with AI at its core.

That transition is creating persistent demand pressure. Hyperscalers and cloud service providers submit their capacity forecasts to NVIDIA each year, estimating how much GPU infrastructure they will need. But, as Huang admitted, **those forecasts have always been too conservative**. Each cycle, demand outruns expectations, leaving both NVIDIA and its customers scrambling to catch up. "We've been in scramble mode now for a couple of years," he explained, because every forecast turns out to understate real-world usage growth. The exponential curve of adoption makes linear planning models obsolete.

This dynamic also undercuts the argument for an imminent glut. Skeptics worry about overbuilding capacity, but Huang argues that until **all general-purpose workloads have been converted to AI-driven workloads**, the probability of a glut is "extremely low." The replacement cycle is simply too large. From recommender engines to enterprise databases, what used to be CPU territory is steadily migrating to GPU and AI acceleration. Until that migration is complete, the industry will remain in a mode of undersupply, not oversupply.

And, importantly, customers are not overestimating their needs—they are **still underestimating**. Huang pointed out that, so far, "nobody has forecast too high." Each year's estimate has lagged actual usage growth, forcing emergency scale-ups. The compounding effect of AI adoption and reasoning-heavy inference means capacity requirements continue to leapfrog projections. In that context, the idea of a glut becomes less plausible. Instead, the more accurate picture is an industry sprinting to keep up with demand that continually outpaces even the boldest forecasts.

Why Today's Spending Isn't a Bubble

In the short run, technology markets are volatile. Share prices can rise or fall sharply in any given quarter—the fourth quarter of 2025 will be no exception. But over longer horizons, the logic driving today's spending on AI infrastructure is fundamentally different from the late 1990s' dot-com era. Back then, users weren't ready, capital evaporated, infrastructure was built ahead of demand, and many business models were broken. The result was a glut.

Today, the situation is inverted. The demand is real and compounding. OpenAI, Meta, Google, Microsoft and xAI are all **scrambling to keep up**, not speculating in hope. As Huang put it: "Every hyperscaler has realized they dramatically underbuilt. Every forecast we've seen has been too low. We're not building for speculation. We're building for active workloads." This isn't Pets.com IPO[5]-ing on vibes; it's a compute bottleneck with customers underestimating their needs, year after year.

That distinction matters for investors. While stock charts over weeks or months may resemble past bubbles, the underlying economics are very different. NVIDIA's growth is occurring under conditions of shortage, not glut. Infrastructure is being deployed because workloads already exist and are scaling faster than expected. The right way to think about this isn't whether the next quarter will beat consensus, but how quickly the world will complete the shift to accelerated AI computing—and how much of that $50 trillion of human intelligence-driven gross domestic product (GDP) will be augmented by billions of AI co-workers in the years ahead.

[1] Source: R. Kurzweil, (n.d.), "The 21st Century: A Confluence of Accelerating Revolutions," in Writings by Ray Kurzweil.

[2] Unless otherwise noted, the source for everything in this piece is B. Gerstner and B. Gurley (hosts), "NVIDIA: OpenAI, Future of Compute, and the American Dream" [audio podcast episode], BG2Pod, 9/26/25.

[3] SQL refers to structured query language. It is the standard programming language used to manage and manipulate relational databases.

[4] NICs are network interface cards (sometimes called network adapters).

[5] Refers to an initial public offering, when a company enters the public markets with a publicly traded share price.