TFOA | The Family Office Association

GPU-BASED INFRASTRUCTURE: A NEW FRONTIER FOR FAMILY OFFICE INVESTING

WWW.TFOA.INFO

GPU-Based Infrastructure: A New Frontier for Family Office Investing

Authored by Nikolay Filichkin & Warren Hosseinion and edited by Marc J. Sharpe, M.A., M.Phil., MBA

Abstract

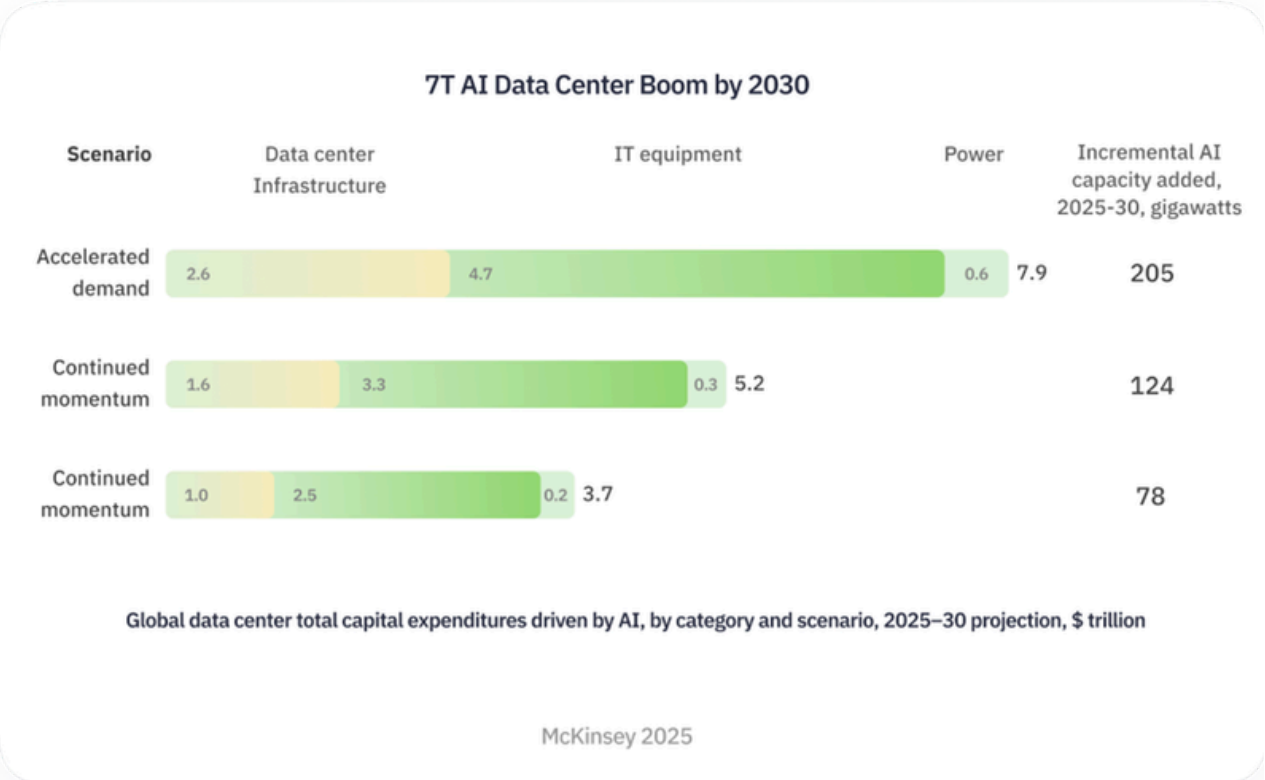
This whitepaper examines GPU (graphics processing unit)-based infrastructure as an emerging asset class for family offices, driven by the AI (artificial intelligence) boom since 2023. It outlines the economics of compute - measured in FLOPS (floating-point operations per second) - highlighting how networked GPUs generate predictable cash flows through training (capital-intensive model building) and inference (recurring real-time applications). Opportunities include structured financing for neocloud operators, yielding net internal rates of return (IRRs) in the mid-teens to mid-twenties over 3–5 years, secured by hard assets with low public market correlation. Key considerations include capital expenditures dominated by hardware, operating expenses led by power and cooling, and depreciation cycles that extend useful life beyond typical accounting periods. Risks span utilization variability and technology obsolescence, yet secondary markets support meaningful residual values. As AI capital expenditures scale significantly, family offices can access diversified yields comparable to mid-market private credit.

Executive Summary

Family offices allocating capital to private credit and real assets are increasingly encountering structured financing opportunities secured by revenue-generating graphics processing units (GPUs). These transactions can offer mid-teens to mid-twenties net IRRs with 3–5-year duration, low correlation to public markets, and hard-asset collateral. This whitepaper maps the underlying economics, risk factors, and structuring considerations in light of the explosive growth of this formerly niche strategy due to the recent boom in Artificial Intelligence (AI) and Large Language Models (LLMs) since 2023.

The New Infrastructure Frontier

Artificial intelligence is the defining technological force of this decade, driving unprecedented demand for one fundamental input: compute. Compute refers to the essential hardware (GPUs, TPUs, CPUs, memory) and the processing power needed for training models and running AI tasks, measured primarily in FLOPS (Floating-Point Operations Per Second), indicating trillions or quadrillions of math calculations per second, showing how much raw power is available for complex AI workloads. It signifies the scale of resources for tasks like data processing, model training, and inference, determining model complexity and capability, with higher FLOPS equating to more powerful AI. Therefore, compute is the unseen energy behind everything from chatbots to autonomous vehicles. And the scale of the compute buildout is already visible in global capital expenditure forecasts, which project trillions of dollars flowing into AI-related data center infrastructure over the coming years.



At the center of this transformation lies the graphics processing unit, or GPU. Once designed for rendering images and games, GPUs now function as the core hardware in machine learning infrastructure. When networked together in servers and racks across data centers, GPUs form the computational fabric that makes artificial intelligence possible.

Yet unlike other essential resources such as energy or real estate, compute has no mature financial infrastructure built around it. Investors can own equity in the companies that use compute or build data centers to host it, but there is still no standardized way to invest directly in the utilization of compute itself. This paper explores how GPUs are used, how their economics work, and why an emerging ecosystem of operators and financial innovators are turning compute into a transparent, yield-bearing asset class.

How GPUs Are Used Today

Every digital interaction today, from searching the web to generating an image with artificial intelligence, relies on compute. GPUs are uniquely efficient at generating compute because they can perform thousands of calculations simultaneously. In AI, that capability enables neural networks to learn, recognize patterns, and make predictions. In essence, GPUs transform electricity and data into ‘intelligence’.

When deployed at scale inside data centers, thousands of GPUs are linked together into clusters that operate as compute farms. Think of them as the digital equivalent of power plants. Just as a turbine converts fuel into energy, these GPU clusters convert electricity into computational power, which AI developers rent by the hour to train or run their AI models.

The efficiency of that conversion – i.e., how much the GPUs are utilized -- determines both the productivity and profitability of a specific GPU cluster. Utilization rate is typically measured on a 0–100% scale, indicating how much of the available compute power is actively in use. A cluster running at 90% utilization is generating revenue nearly around the clock, while one sitting at 50% is leaving half its capacity idle. The closer operators can keep their systems to full utilization, the more consistent and predictable their cash flow becomes.

For investors, this means compute capacity is not a speculative asset but a productive one. When managed by credible operators and leased under transparent contracts, GPU infrastructure produces predictable, yield-bearing cash flows similar to energy infrastructure. The more efficiently the GPUs are deployed and maintained, the more consistent the returns.

Training vs. Inference

AI workloads generally fall into two categories: training and inference. Understanding their distinctions is critical to evaluating risk, return, and duration.

- Training uses massive GPU clusters to build foundational models, similar to constructing an oil refinery. These workloads are capital-intensive and cyclical but yield high utilization for defined periods.

- Inference happens after the model is built. It refers to the process of using that trained model to make predictions, generate content, or answer questions in real time. When someone types a question into ChatGPT, they're triggering inference. These workloads are continuous, lower intensity, and generate recurring revenue similar to power purchase agreements or SaaS contracts.

For investors, both are relevant. Training offers high short-term yield opportunities tied to specific projects, while inference provides more stable, long-term income streams.

Comparison

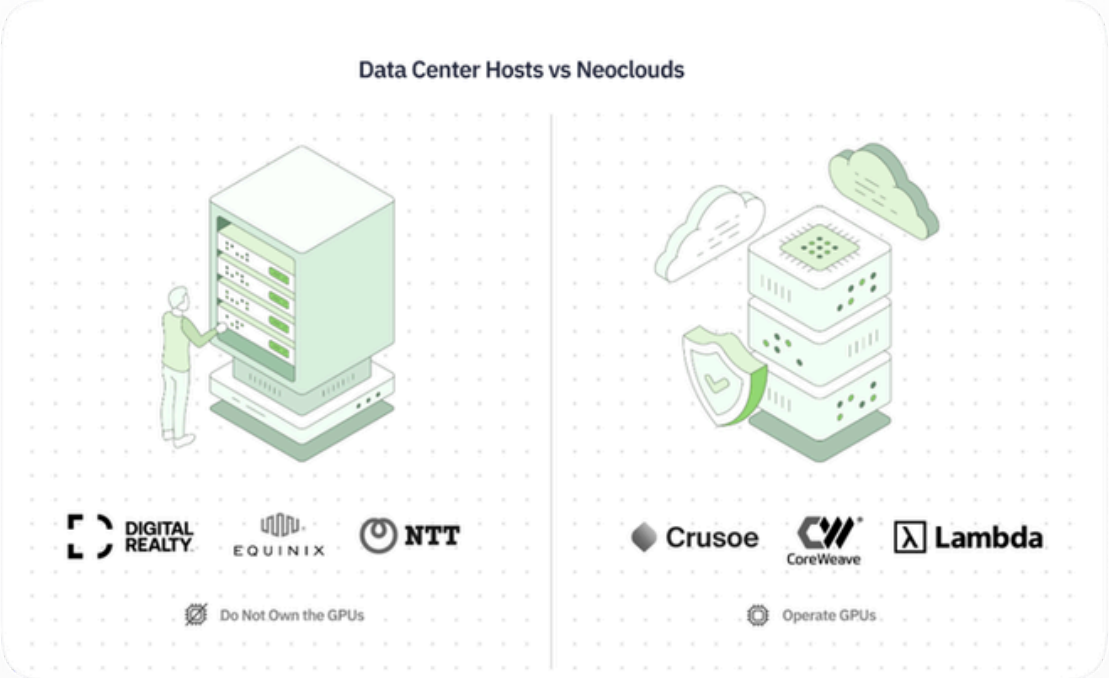
	Training	Inference
Purpose	Building and improving AI Models	Running deployed AI models in real-time
Duration	Long sessions (weeks/months)	Continuous Shorter Tasks
Hardware Intensity	Highest	Moderate
Revenue Stability	Project-based	Subscription/ usage-based

Two Economies of Compute

[2] See Appendix A for step-by-step calculations.

[3] Holland & Knight. (2023). Independent Sponsors: Market Trends and Industry Insights.

Data Center Hosts vs. Neocloud Operators



To understand how GPUs are monetized, it's important to distinguish between data center hosts and neocloud operators.

- Data Center Hosts lease physical space, cooling, and power to tenants. They are essentially real estate and utility businesses and do not own the GPUs themselves.

- Neocloud Operators represent a new class of companies that both own and operate GPUs, offering compute access directly to AI developers on demand.

These neoclouds are effectively the independent power producers of the AI age: capital-intensive, infrastructure-driven, and often seeking external financing to scale GPU clusters quickly. A potentially attractive GPU investment opportunity sits with small to medium-sized neocloud operators as these entities are too small to self-finance like hyperscalers, yet large enough to maintain stable utilization contracts with enterprise clients.

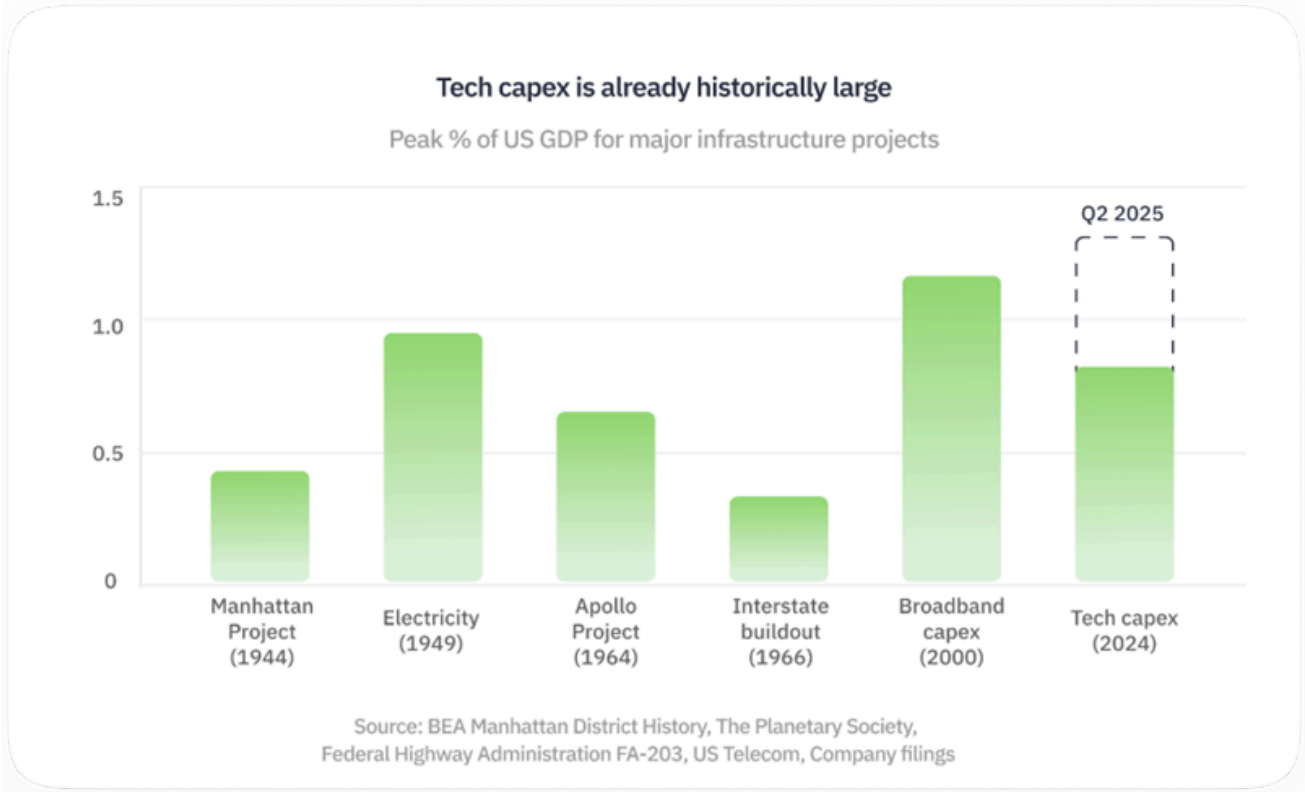
Size and Structure of Operators

Tier	Example Entities	Typical Characteristics
Tier 1 - Hyperscalers	Microsoft, Amazon, Google	Vertically integrated, own global infrastructure, limited external access
Tier 2 – Neocloud Operators	CoreWeave, Lambda, Crusoe, Volt	High growth, capital-hungry, GPU-focused, multi-tenant
Tier 3 – Specialized Operators	Regional AI hosts, vertical cloud startups	Targeted workloads (e.g., medical imaging, simulation), rely on leasing or financing

For investors evaluating compute exposure, Tier 2 and Tier 3 operators present a combination of return potential and risk diversification. These companies have strong, sustained demand for GPU capacity from AI developers but lack the low-cost financing available to large hyperscalers. That funding gap creates opportunity: family office investors who provide capital for GPUs can capture higher returns, secured by tangible assets and underpinned by contracted revenue from active GPU usage.

What Drives Capex and Why It's Changing

We are in the midst of one of the largest capital buildouts in modern history. Technology investment, particularly in AI infrastructure, has reached a scale comparable to the most significant public works projects of the past century. AI-related Capex, including GPUs and data centers, already rivals the peak infrastructure spending related to electricity, broadband, and the interstate highway system.



The key drivers of capital expenditure for GPU infrastructure varies widely depending on hardware generation, power density, and facility readiness. Most of the total cost comes from the GPUs assembled into servers, while the remainder covers networking, cooling, and power delivery systems needed to keep them operating efficiently. In some cases, where there are specific software requirements to service a customer, the software can be included in capex as well.

- Hardware Costs: High-performance GPU servers remain the largest expense. Pricing depends on supply availability and lead time, with newer models commanding a premium until supply catches up.

- Generation Cycles: New GPUs release roughly every 18–24 months, increasing compute density and efficiency—but requiring periodic reinvestment.

- Facility Readiness: Retrofitting existing centers vs. building greenfield sites changes capex profiles dramatically.

- Interest Rates & Financing Costs: As with any capital-intensive industry, macro conditions impact expansion pace.

A notable trend is the rise in asset-light financing. Instead of operators purchasing GPUs outright, structured leases enable rapid scaling without balance-sheet strain, converting capex into manageable operating costs.

The Cost of Keeping Compute Online

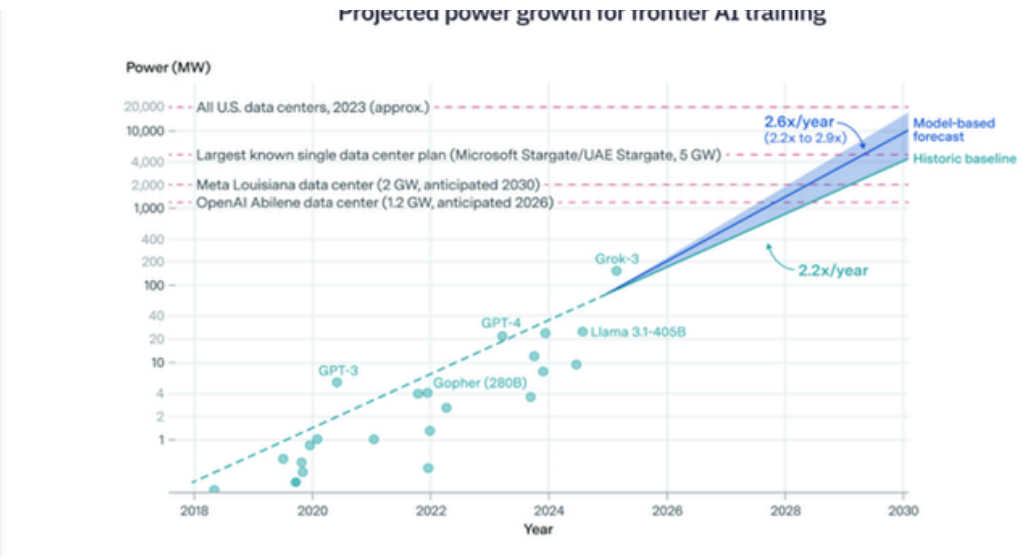
Operating expenses (OpEx) represent the ongoing cash costs required to keep GPU infrastructure running. For most operators, power is the dominant expense, as modern GPU racks can draw 20–40 kilowatts each which is comparable to the load of an entire suburban home. These expenses define the baseline cost structure of a GPU cluster and directly influence the utilization required for sustainable profitability. A typical cash OpEx profile includes:

- Power & Cooling (25–40%) - Electricity and cooling are the largest operating costs. Total expense depends on local energy pricing, power contracts, and the facility's Power Usage Effectiveness (PUE), which measures how efficiently electricity is converted into usable compute.

- Maintenance & Repairs (5–10%) - This includes hardware servicing, component failures, and ongoing software updates. Because GPU clusters operate continuously, routine maintenance is essential to sustaining uptime and predictable performance.

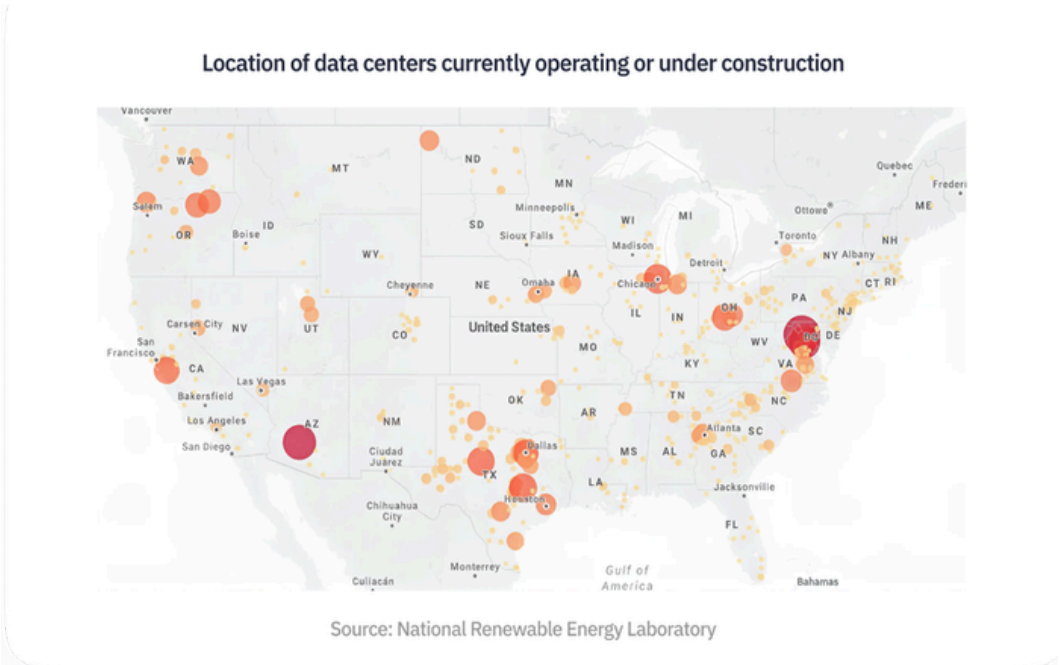
- SG&A + Network Operations (10–15%) - Staffing, monitoring, connectivity, and bandwidth contribute to the operational overhead required to maintain reliable, secure, and responsive compute capacity.

- Other Operating Costs (4–8%) - Insurance, colocation fees, cross-connects, spare parts inventory, and site-specific operational needs. The range varies based on facility type and deployment model.



Location remains one of the most significant determinants of OpEx. Many operators co-locate near regions with low-cost or renewable energy sources, such as hydroelectric, nuclear, geothermal, or flare-gas recovery sites, to reduce electricity costs and improve sustainability profiles. Others deploy containerized or modular micro–data centers to access cheaper power and accelerate deployment timelines without relying on traditional data center infrastructure.

In practice, data-center capacity in the United States is not spread evenly. It concentrates in a handful of metro areas and corridors where power availability, land costs, and network connectivity are most favorable. The map below illustrates this clustering: large numbers of facilities are grouped around a few key hubs rather than being evenly distributed across the country. That concentration reflects deliberate siting decisions designed to manage OpEx and ensure long-term access to reliable power and fiber.



How Compute Generates Cash Flow

Revenue in GPU operations stems from leasing compute hours to customers such as AI startups, research institutions, or enterprise clients. These contracts can take several forms:

1. Fixed-Term Contracts: Multi-month or multi-year agreements that provide stable, predictable cash flows.

2. On-Demand: Pay-as-you-go pricing, like traditional cloud services.

3. Spot Market: Cost-effective option for AI companies offering unused capacity for short periods of time. These can be interrupted with short notice (typically a few minutes) to prioritize on-demand & contracted terms.

4. Revenue-Share Models: Operators earn a portion of the client's revenue that depends on compute usage.

Pricing for high-performance GPUs varies widely depending on contract length, supply conditions, and the level of service provided. Rates tend to be higher for dedicated clusters or premium support and lower for shared or short-term access. In practice, an operator's revenue depends on the following factors:

- Services: Bare-metal, or software services on top? The best operators with the most valuable contracts offer more than bare-metal hardware. They offer software services to service their customers and a particular market niche.
- Utilization: The percentage of time GPUs are actively rented or in use.
- Unit Pricing: The effective hourly rate achieved across different clients and workloads.
- Contract Quality: The creditworthiness and reliability of offtake counterparties.

When these factors are managed well, GPU infrastructure behaves like a productive asset—generating recurring, measurable cash flow. Disciplined GPU management and high-quality contracts can make revenue streams from compute infrastructure resemble those of traditional income-generating assets such as energy projects or equipment leasing.

Asset Life Cycles & Depreciation

GPU depreciation behaves very differently from traditional IT hardware. Rather than becoming obsolete when new chips are released, GPUs transition through multiple workload phases and retain meaningful value over long periods. For investors, this creates predictable collateral behavior and durable cash-flow profiles whereby GPUs behave more like income-producing infrastructure than short-lived IT equipment.

- Book Value Depreciation (Accounting) - Under GAAP/IFRS, GPU servers are depreciated over 5–6 years using straight-line depreciation (≈16–20% annually). While this defines how the asset appears on financial statements, it does not reflect real market value or economic life, which typically extends beyond the accounting horizon.

- Economic Depreciation (Real Useful Life) - In practice, GPUs remain revenue-producing for 6–8+ years as workloads cascade and because inference requirements grow more slowly than training, older GPUs continue to serve valuable roles well beyond their book life.
    - Years 0–3: Frontier training and high-performance workloads
    - Years 4–7+: Inference, fine-tuning, research, rendering, batch jobs, enterprise AI

- Market Depreciation (Resale Value Behavior) - Global demand for compute, backward-compatible software (CUDA), and deep secondary markets all support slower, smoother value decline compared to typical IT hardware. Most data-center GPUs retain 45–65% of their value at Year 3 and 25–35% at Year 5.

The Emerging Yield Curve of Compute

While returns vary by structure, GPU-based investments exhibit characteristics familiar to family offices investors who are accustomed to asset-backed credit or equipment leasing.

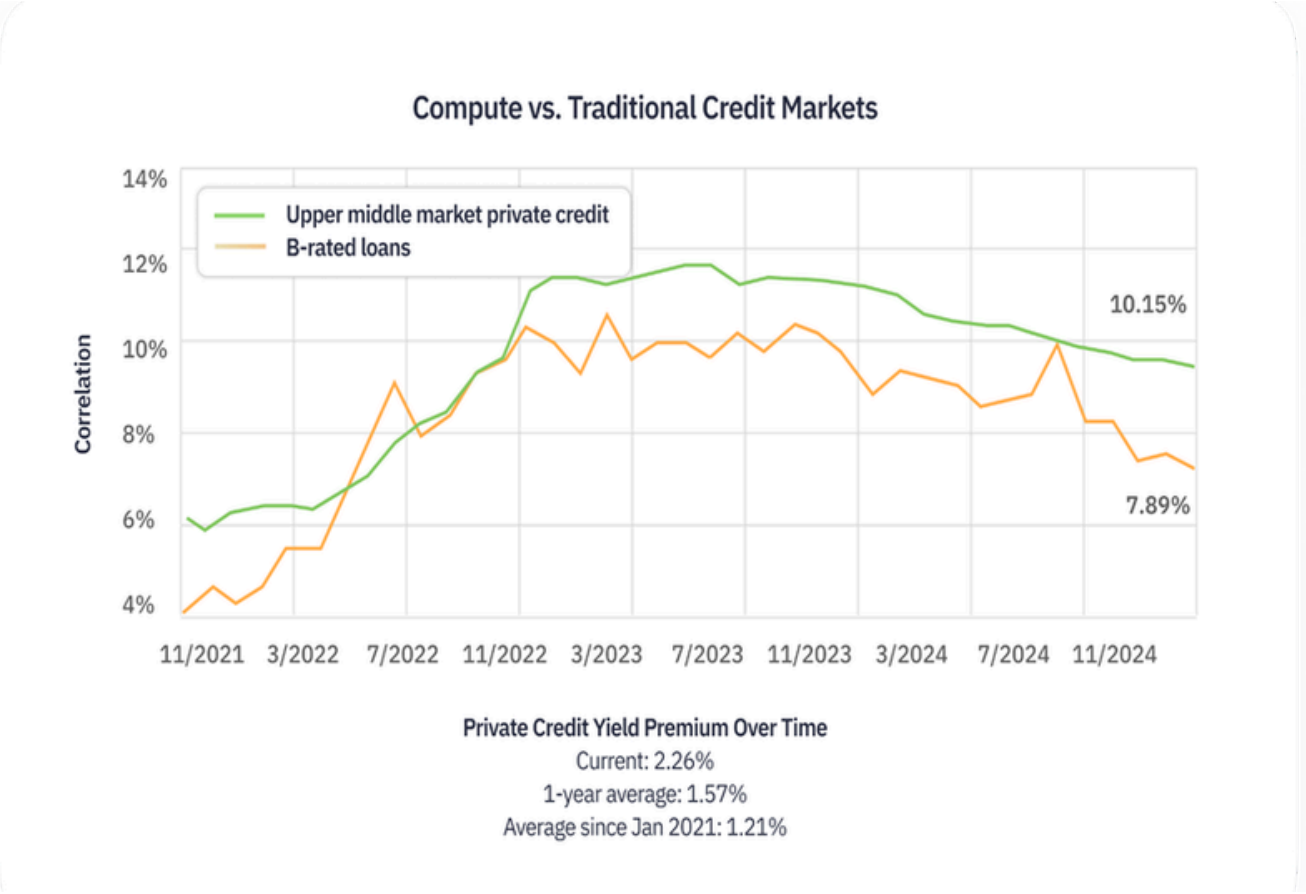
- Gross Yields: 25–60% annually, depending on utilization and cost of capital.

- Net Yields: 15–30% after OpEx and depreciation. GPU yields of 15-30% represent a significant premium over the ~10-11% returns usually available in private credit markets.

- Duration: Typical investment horizons of 3–5 years, aligned with GPU upgrade cycles.

- Residual Value: Older GPUs are often redeployed from training to inference workloads, extending useful life and maintaining residual resale value.

Importantly, compute yields are not directly correlated to equity or bond markets. Instead, they follow technology-specific cycles tied to AI adoption and hardware lifecycles. For family offices seeking differentiated yield exposure, this emerging asset class offers diversification anchored in tangible infrastructure. Therefore, a well-structured GPU portfolio should be able to achieve stable returns comparable to mid-market private credit, with the added upside of technology appreciation during high-demand cycles.



Key Risks for Family Offices Investing in GPUs

Investing in GPUs carries risks that family offices should evaluate, including but not limited to operator execution, utilization, and technology cycles. Returns depend on GPUs being deployed with disciplined up-time and consistently high utilization, supported by reliable customers and stable access to power. Variability in workload demand or increases in energy costs can influence margins, which is why underwriting operator capability and site selection remain essential.

Depreciation also plays a role. GPUs retain meaningful utility across multiple workload tiers, but newer generations can shorten the economic life of older hardware if not structured thoughtfully. Clear contract terms, realistic depreciation assumptions, and diversified exposure across operators and regions can help to mitigate this dynamic.

Overall, the risks resemble those found in other private credit and infrastructure opportunities: counterparty quality, operating discipline, depreciation management, and cost control. With careful diligence, compute can function as a transparent, asset-backed exposure to the underlying growth of AI.

Conclusion

Compute is the most fundamental resource of the modern era. It is the raw power that transforms information into intelligence. For family offices and institutional investors, understanding its economics is the first step toward participating in a market expected to surpass $7 trillion in capital expenditure by 2030. As GPUs evolve from technical equipment to structured financial assets, investors who engage early can position themselves at the foundation of the next generational infrastructure cycle. Innovation in financial architecture will define the sustainability of the AI revolution and family offices are ideally positioned as a source of flexible and patient capital to support the next industrial revolution in the intelligence economy.

Bibliography

- McKinsey & Company – "The cost of compute: A $7 trillion race to scale data centers" (2025)
- https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers McKinsey & Company
- McKinsey & Company – "Who's funding the AI data center boom?" (2025)
- https://www.mckinsey.com/featured-insights/themes/whos-funding-the-ai-data-center-boom McKinsey & Company
- NREL – "Data Center Infrastructure in the United States, 2025 (Map)"
- https://research-hub.nrel.gov/en/publications/data-center-infrastructure-in-the-united-states-2025-map Research Hub
- DatacenterDynamics – "NREL launches US data center infrastructure map" (news overview)
- https://www.datacenterdynamics.com/en/news/nrel-launches-us-data-center-infrastructure-map/ DataCenterDynamics
- Investment Research Partners – "Chart of the Week: AI is the New Railroad (Sort of)" (2025)
- https://www.investmentresearchpartners.com/post/chart-of-the-week-8-3-2025 IRP
- Suro Capital – "AI Infrastructure: The Great Mobilization of Our Time" (2025)
- https://surocap.com/wp-content/uploads/2025/09/AI-great-mobilization-suro-7.pdf Surocap
- Goldman Sachs – "Rising power density disrupts AI infrastructure" (2025)
- https://www.goldmansachs.com/insights/articles/rising-power-density-disrupts-ai-infrastructure Goldman Sachs
- ServerLift – "GPUs Power the AI Boom in Modern Data Centers" (2025)
- https://serverlift.com/blog/how-data-center-gpus-have-changed-the-ai-playing-field/ ServerLIFT®
- NVIDIA – "Considerations for Scaling GPU-Ready Data Centers" (technical overview, PDF) https://www.nvidia.com/content/g/pdfs/GPU-Ready-Data-Center-Tech-Overview.pdf NVIDIA

About the Authors

Nikolay Filichkin is a business development and partnerships executive with deep experience across enterprise technology and high-growth operating environments. As Chief Business Officer of Compute Labs, Mr. Filichkin leads strategic partnerships and capital deployment initiatives to support the firm's mission to build the capital markets for compute. Prior to joining Compute Labs, Mr. Filichkin held senior roles at Xsolla, where he led cross-border acquisitions, launched new product lines, and built go-to-market functions that supported the company's global expansion. His background spans strategic sales, M&A execution, and operational scale-up across both early-stage and growth-stage organizations. Mr. Filichkin is recognized for fostering durable partnerships and guiding complex initiatives from inception to revenue realization.

Warren Hosseinion is a capital markets and growth professional with experience spanning public-market M&A, investor relations, marketing, and corporate development across AI infrastructure, biotechnology, and emerging technology sectors. As Head of Capital Markets and Growth at Compute Labs, Mr. Hosseinion oversees investor relations, fundraising strategy, and market positioning as the firm develops structured investment products for the compute asset class. Mr. Hosseinion brings previous experience as Vice President of M&A at Voyager Acquisition Corp. (NASDAQ: VACH), where focused on deal sourcing, transaction execution and due diligence. He also brings experience as Chief Operating Officer of Vesicor Therapeutics, helping guide their go-public strategy and daily operations.

Marc J. Sharpe is a global investment executive and board member with a distinguished career spanning family office management, private equity, venture capital, and investment banking. Known for his strategic insight, deep expertise in family office governance, and ability to foster innovation and value creation, Mr. Sharpe has built and led investment platforms that deliver sustainable growth while navigating complex financial and operational challenges. His leadership style emphasizes integrity, continuous improvement, and long-term partnerships that generate significant stakeholder value. Mr. Sharpe is the Founder and Chairman of The Family Office Association, a premier global peer network of single-family offices. Since 2007, he has cultivated a community of senior family office executives and principals representing some of the world's wealthiest families, promoting education, shared-best practices, and co-investment opportunities. Under his leadership, TFOA has become a trusted forum for collaboration, market insight, and proprietary investment deal flow on a global scale. He also teaches an MBA class on "The Entrepreneurial Family Office" as an Adjunct Professor at Rice University and Southern Methodist University. Mr. Sharpe holds an M.A. from Cambridge University, an M.Phil. from Oxford University, and an MBA from Harvard Business School.

Contact: marc@tfoa.me

About TFOA

The Family Office Association is a global peer network that serves as the world's leading single family office community. Our group is for education, networking, selective co-investment, and a resource for single family offices to share ideas, deal flow and best practices. Members are not actively marketing products or services to other members and no contact information or email lists will ever be shared. Since our founding in 2007, TFOA has led the global single family office community by delivering world-class educational content, unique networking opportunities, and exceptional thought leadership to our highly curated network of the world's largest and wealthiest families. If you'd like to access our free library of whitepapers or receive event updates and more timely information please visit: www.tfoa.info

Marc J. Sharpe

Chairman & Founder
The Family Office Association

Disclaimer

All the information provided by The Family Office Association ("TFOA") are for informational and educational purposes only and neither purports nor intends to be, specific trading or investment advice or a recommendation to buy or sell any security, fund, or financial instrument. Information should not be considered as an offer or enticement to buy, sell or trade. You should seek appropriate advice from your broker, or licensed investment advisor, before taking any action. Past performance does not guarantee future results. By subscribing as a member to TFOA, you acknowledge and accept that all trading decisions are your own sole responsibility, and TFOA or anybody associated with TFOA cannot be held responsible for any losses that are incurred as a result. The information contained in this electronic mail message, including attachments, if any, is confidential information. It is intended only for the use of the person(s) named above. Internet emails are not necessarily secure. Each recipient is responsible for carrying out such virus and other checks as it considers appropriate to ensure that the receipt, opening, use or onward transmission of this message and any attachments will not adversely affect its systems or data.