MORE INFORMATION:

INFO@DIGITALPOWEROPTIMIZATION.COM

HPC and AI Training:
What It Is and What It Means for Green Power Companies

We attempt to cut through the hype by building a ground-up case for why the rise of HPC and AI computing is relevant to owners of power assets, as well as data center developers and "Bitcoin miners." In short, we believe this technology will spur a tremendous amount of power-intensive data center development in the 2020s, but it lacks the flexibility of proof-of-work (Bitcoin) computing.

INTRODUCTION:

This memo addresses the underlying infrastructure of a technology that has demonstrated 1) the power to fundamentally change the economy, 2) technological breakthroughs that few believed were likely at this stage, and 3) the ability to attract nearly limitless amounts of capital for businesses that show promise. We are, of course, not talking about blockchain, but AI. There are numerous parallels between the hype cycles for blockchain and AI, and the amount of capital raised in the recent blockchain cycle brings into focus the financial incentive for every technology business to suddenly emphasize AI. This attention brings about daily media coverage, predictions of a transformation of the economy, and reinforces the need for businesses to talk about the recently-hyped technology.

One thing the hype cycle does not encourage is an understanding of the underlying technology, which takes time to build. We at DPO have put together this short memo to help our clients better understand how new high-performance computing ("HPC") technologies (including AI computing) are relevant to their businesses.

KEY TAKEAWAYS:

Based on our research, we make the following key points:

- HPC uses new chip technology to perform certain computing functions exponentially faster than was previously possible.
- This new chip technology, and the software it has enabled (such as AI), is expected to drive unprecedented computing demand.
- HPC data centers have many of the same requirements as a traditional cloud computing data center.
  - One major exception is that some HPC facilities do not need to be located in a data center hub.
- HPC data centers are power-intensive, but not nearly as power-intensive as proof-of-work (Bitcoin) data centers.
  - Capital cost per MW is especially high for HPC data centers, making them uneconomic to curtail.
  - Energy (and its cost) is a key consideration of HPC siting, but not nearly as critical as it is for proof-of-work.
- Utilities will care about HPC because it is a large and growing load on their systems, sometimes located near generation.
  - However, unlike proof-of-work, HPC is not a curtailable load that can be switched on and off to balance the grid.
- Thoughtful data center developers and some proof-of-work computing companies can benefit by building HPC rack space quickly. Power companies will benefit by serving these loads, especially if they are placed in locations with excess generation.

Executive Summary:

We begin with an explanation of what HPC is and how it differs from traditional data center computing. Advances in graphics processing unit ("GPU") technology and generative AI software are key drivers of the recent growth in this sector. We can expect demand for this compute to keep growing over the coming years, as the amount of HPC infrastructure available is currently well short of the expected AI-driven demand for HPC. McKinsey expects data center power use to grow ~10% per year in the coming decade (more than doubling and reaching 35 GW by 2030), and we believe this estimate may prove conservative.

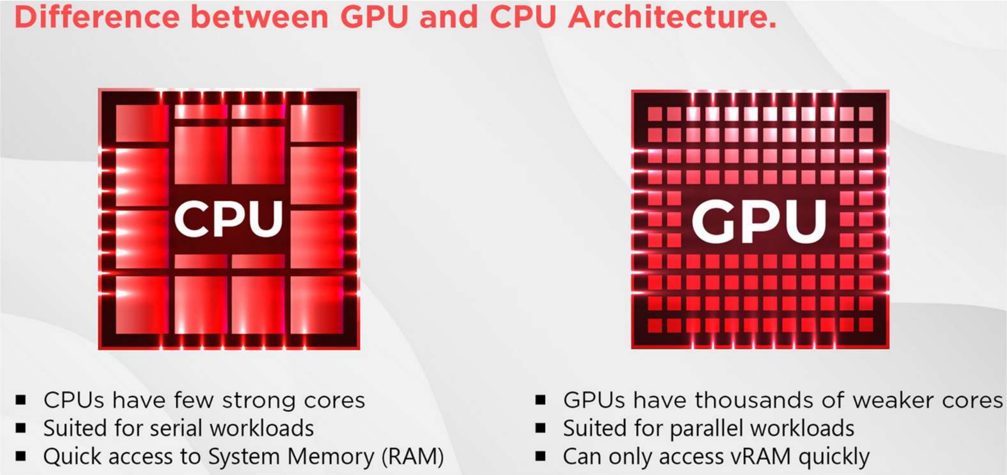
HPC data centers look and feel largely like traditional data centers, but with greater power density and different types of GPU-enabled servers in place. Customers of HPC data centers will not tolerate having their compute functions curtailed voluntarily, and it does not make sense to do so from an economic standpoint given the high value of compute. A high-end HPC deployment is able to sell compute for an equivalent of ~$5,000 per MWh—it would not make sense to curtail that function in order to bring overall power costs down by ~$20 / MWh. This leads us to conclude that HPC is generally unlike proof-of-work or Bitcoin ("PoW") computing in terms of strategic placement by power companies—this is not a load that can be used to help optimize power assets by absorbing power when it is not valuable and releasing power when its external value is higher.

While HPC is not the curtailable load our clients would like it to be, it remains relevant to power companies and "bitcoin miners." It appears that some in the PoW space may be well-suited to become developers of HPC sites, even if they are operated quite differently from a PoW data center. This will all be relevant to power companies for two reasons: 1) HPC data centers can sometimes be placed adjacent to rural solar and wind assets, providing a physical offtake for power, and 2) simply because HPC may be the fastest-growing sector for large loads in the country for the remainder of this decade. We advise those in the power industry to pay attention.

Note that our conclusions and the business implications of HPC are found in the second half of this memo, while the first half works to explain the underlying technology and how it is applied.

High Performance Computing Summary:

HPC refers to the use of powerful computing systems and techniques to solve complex problems or perform computationally demanding tasks. HPC focuses on maximizing processing power, memory capacity, and data throughput to achieve high-performance and fast execution of calculations. HPC computing is exponentially faster than legacy computing. For example, a legacy computing system has a limited number of processing cores and lacks the ability to handle large volumes of genetic data efficiently. It takes *several weeks* to process and analyze a dataset containing thousands of genomes. The analysis involves complex algorithms for comparing and identifying genetic variations, which are crucial for understanding diseases and developing personalized medicine. With the HPC system, the analysis of the same dataset containing thousands of genomes can be completed in a *matter of hours*. The HPC system leverages its vast computational resources to distribute the workload across multiple nodes and process the data in parallel. HPC computing enables researchers to perform large-scale genetic analyses that were previously impractical or unfeasible with legacy systems. It empowers researchers to make significant advancements in understanding complex genetic mechanisms, identifying potential therapeutic targets, and facilitating the development of precision medicine approaches.


Difference between GPU and CPU Architecture.

- CPUs have few strong cores
- Suited for serial workloads
- Quick access to System Memory (RAM)

- GPUs have thousands of weaker cores
- Suited for parallel workloads
- Can only access vRAM quickly

Figure 1 - CPUs are great for executing tasks one by one extremely quickly. GPUs are better at executing thousands of less demanding tasks simultaneously. Source: cgdirector.com. ***Please see the Appendix for detail on differences between HPC and legacy data centers.***

It is critical to note that HPC is not simply a rack full of GPUs. A rack will include multiple cutting-edge servers with high-power CPUs, storage and other infrastructure which work together with the GPUs to maximize compute power. For high-end HPC, the server and its other components can cost nearly as much as the GPUs.

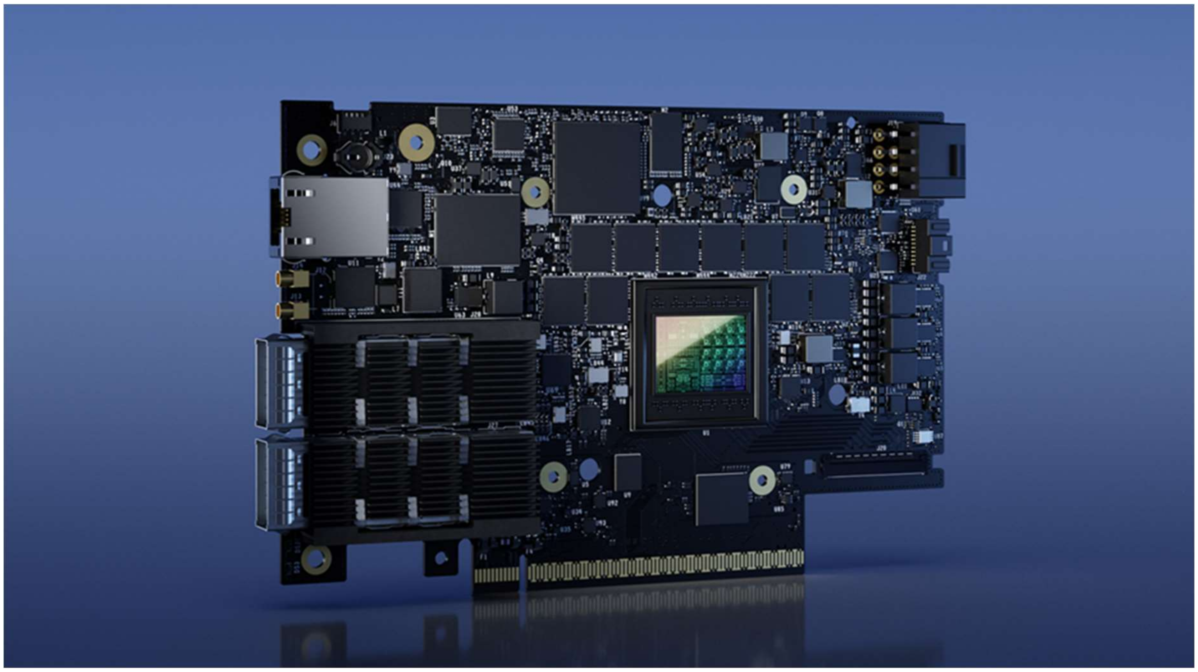
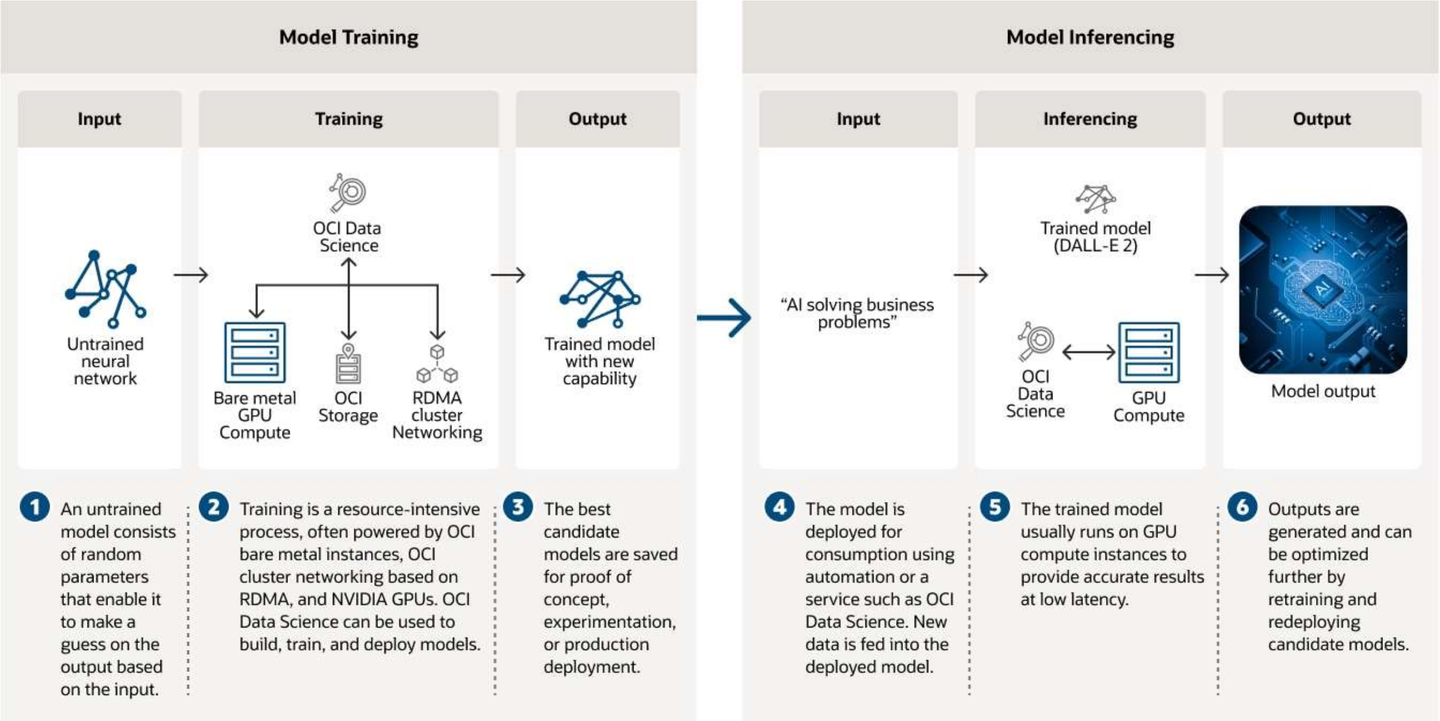


Figure 2 – HPC data center technology trends towards removing bottlenecks between compute, memory, and network by combining them on a single device. NVIDIA BlueField-3 Digital Processing Unit (DPU)

AI machine learning is comprised of two phases, training and inferencing. During the training phase, the machine learning model learns from a large amount of labeled data to recognize patterns and make predictions. It adjusts its internal parameters to optimize its accuracy. Training requires significant computational resources and time. Once the model is trained, it enters the inferencing phase. This is where the model applies what it has learned to new, unseen data to make predictions or draw insights. Inferencing is faster and requires fewer computational resources compared to training. It is also more sensitive to network latency. Inferencing allows the model to make real-time decisions based on its learned knowledge.



1 An untrained model consists of random parameters that enable it to make a guess on the output based on the input.

2 Training is a resource-intensive process, often powered by OCI bare metal instances, OCI cluster networking based on RDMA, and NVIDIA GPUs. OCI Data Science can be used to build, train, and deploy models.

3 The best candidate models are saved for proof of concept, experimentation, or production deployment.

4 The model is deployed for consumption using automation or a service such as OCI Data Science. New data is fed into the deployed model.

5 The trained model usually runs on GPU compute instances to provide accurate results at low latency.

6 Outputs are generated and can be optimized further by retraining and redeploying candidate models.

Figure 3 – The process flow of starting with an untrained model, training it, then deploying it for use in production (model inferencing). Source: Oracle.com

Given the differences in requirements for AI training and AI inferencing, we expect that these functions are likely to migrate to different physical locations.

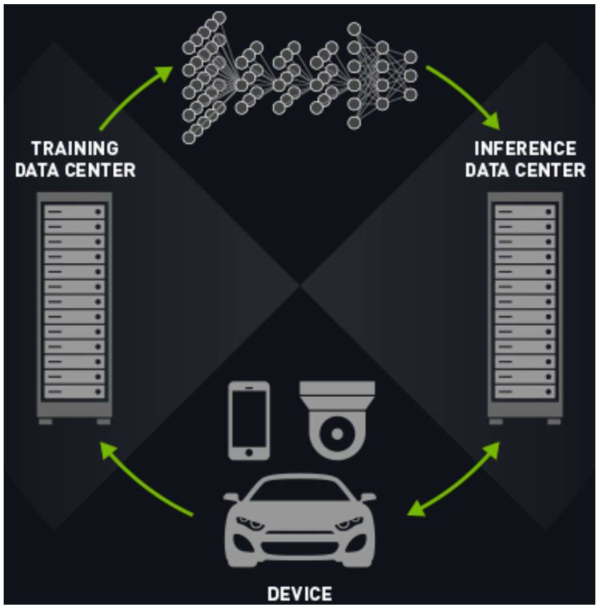


Figure 4 – A dataset is collected from the devices then used to train the model, then the model is deployed for inferencing that the devices will use to operate in real time. Source: NVIDIA.com

HPC Growth Drivers:

HPC's rapid growth is rooted in technological breakthroughs in the chip industry. Initially GPUs were used mostly for graphics processing and video games. In the mid-2000s, machine learning shifted to GPUs due to their parallel processing capabilities. The emergence of deep learning in the 2010s further accelerated the use of GPUs for machine learning. The use case for PoW compute began to emerge during this time as well. The development of specialized GPUs designed specifically for machine learning and AI workloads has further enhanced their performance and efficiency in this space. Critically, this has enabled generative AI models that connect billions of parameters and create software products that few would have expected just two years ago. There are a number of other functions well-suited to GPU compute, which has helped launch an entirely new class of data centers based around these chips. The very high demonstrated value of running this kind of computing, particularly for AI, has created strong and sustainable growth for HPC data centers.

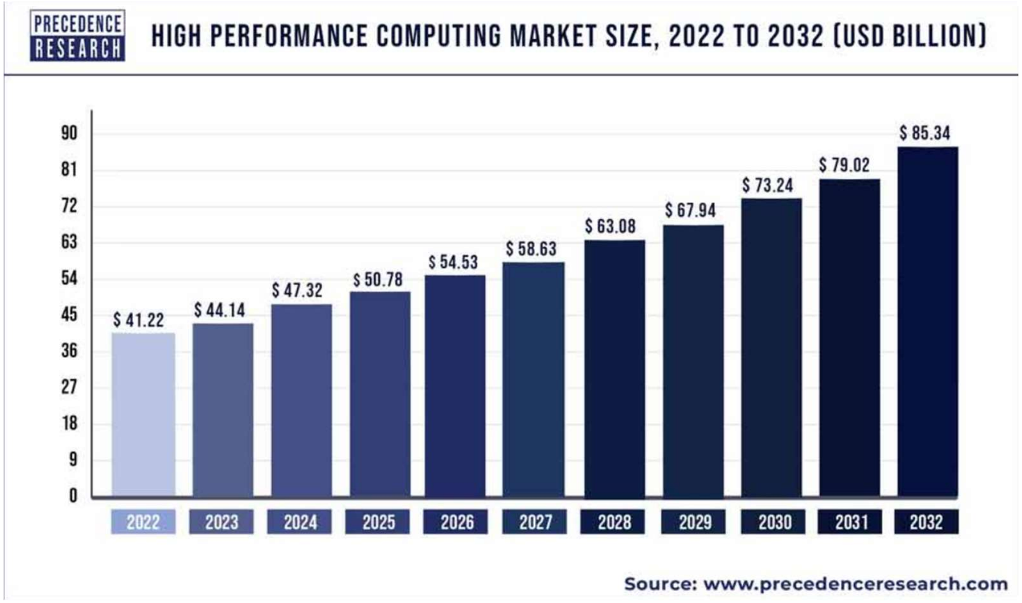


Figure 5 – The HPC Compute market size is projected to double in size by 2032. Source: Precedence Research

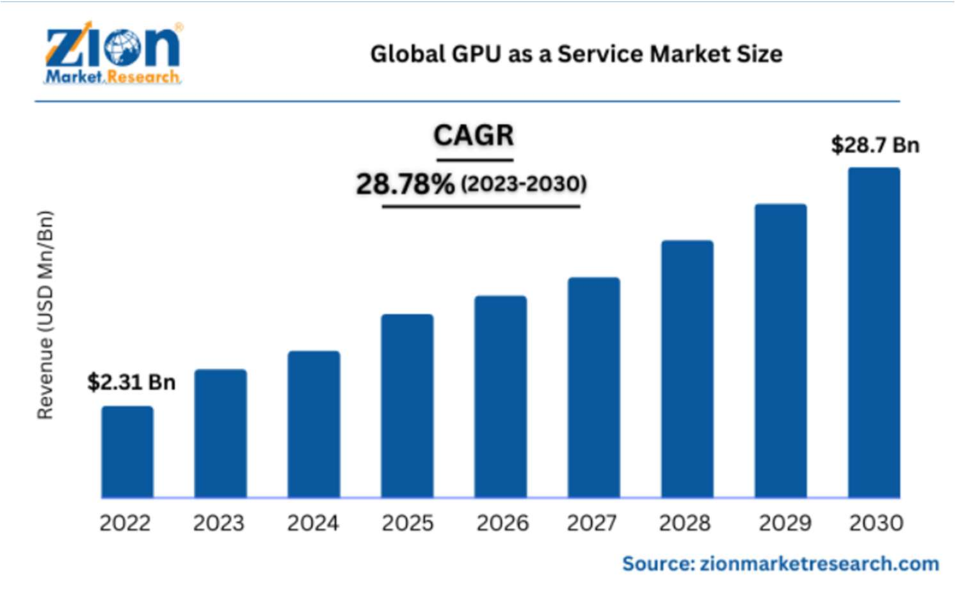


Figure 6 – The Global GPU as a Service market is projected to 10x over the remainder of this decade. Source: Zion Market Research

Below, we list several growth drivers for HPC computing. While AI training will be an impactful contributor, there are many other fast-growing applications that receive less attention:

1. **Artificial intelligence and machine learning:** The rapid growth of artificial intelligence (AI) and machine learning (ML) has increased the demand for HPC. Training and inference processes in AI and ML often require massive computational power, and HPC systems with specialized hardware, such as GPUs or TPUs, are well-suited for these tasks. HPC enables researchers and data scientists to process vast amounts of data and train complex models, driving advancements in AI and ML applications.
2. **Scientific research and simulation:** HPC is crucial for scientific research, enabling researchers to perform complex simulations, modeling, and data analysis. Scientists and researchers in fields such as physics, chemistry, biology, climatology, and engineering utilize HPC to tackle computationally intensive tasks and gain insights into complex phenomena. The demand for HPC in scientific research continues to grow as researchers aim to solve larger and more intricate problems.
3. **Big data analytics:** As the volume, velocity, and variety of data continue to increase, organizations rely on HPC to analyze and extract insights from massive datasets. HPC systems provide the computational resources required to process, store, and analyze large-scale data efficiently. Industries such as finance, healthcare, retail, and manufacturing leverage HPC to gain actionable insights, improve decision-making, and identify patterns or anomalies in extensive datasets.
4. **Engineering and product design:** HPC plays a vital role in engineering and product design by enabling simulations, finite element analysis, computational fluid dynamics, and other computationally intensive tasks. Engineers and designers use HPC to model and test designs, optimize performance, and reduce the time and costs associated with physical prototyping. As engineering processes become more complex, the demand for HPC in this domain continues to grow.
5. **Energy exploration and production:** The energy sector, including oil and gas exploration and renewable energy research, relies on HPC for advanced simulations and analysis. HPC systems are used to model reservoirs, optimize drilling processes, simulate seismic activity, and improve energy extraction techniques. The need for HPC in energy exploration and production arises from the industry's continuous search for more efficient and sustainable energy solutions.
6. **Financial services and risk analysis:** The financial industry utilizes HPC to perform complex risk analysis, portfolio optimization, algorithmic trading, and fraud detection. HPC enables financial institutions to process large amounts of financial data in real-time, improve accuracy in risk assessments, and make data-driven decisions more efficiently. As financial markets become increasingly data-intensive and complex, HPC becomes essential for staying competitive.
7. **The increasing availability of cloud-based HPC services:** Cloud-based HPC makes it easier and more affordable for businesses and organizations to access HPC resources without the large upfront costs to stand up their own infrastructure.
8. **GPU Cloud as a subset of HPC:**
    a. <u>Gaming and media streaming:</u> The gaming industry has experienced significant growth, with an increasing demand for immersive and graphically rich gaming experiences. Cloud GPU services enable game developers and publishers to offload the rendering and processing tasks to powerful GPUs hosted in the cloud, delivering high-quality graphics and reducing the need for users to own expensive gaming hardware. Additionally, cloud GPU services are also utilized for media streaming services, where GPUs help in video encoding, decoding, and transcoding tasks.

    b. <u>Virtual desktop infrastructure (VDI):</u> With the rise of remote work and the need for secure and flexible access to desktop applications, virtual desktop infrastructure has gained popularity. GPUs play a crucial role in delivering a smooth and responsive user experience for graphics-intensive applications. Cloud GPU services enable organizations to provide virtual desktops with GPU acceleration, allowing users to access demanding applications from any device with an internet connection.

    c. <u>Data analytics and visualization:</u> As organizations generate and collect large amounts of data, there is a growing need for efficient data analytics and visualization. GPUs excel at parallel processing, enabling faster data analysis, machine learning model training, and interactive visualizations. Cloud GPU services provide the computational power required for complex data analysis tasks, enabling organizations to derive insights and make data-driven decisions more efficiently.

    d. <u>Autonomous vehicles and robotics:</u> The development and deployment of autonomous vehicles and robotics rely on advanced computational capabilities. GPUs are used for perception tasks, sensor fusion, deep learning algorithms, and real-time decision-making in these applications. Cloud GPU services offer the necessary GPU power and scalability to support the development and training of autonomous systems.

9. **The growth of the Internet of Things (IoT).**
10. **The development of new HPC technologies such as quantum computing and exascale computing.**

HPC Requirements:

The requirements for HPC data centers are closer to those of traditional data centers, with a key difference: they do not need to be located close to hyperscale data centers in major data center markets. This opens up a large number of sites that would not work for large or hyperscale data centers, particularly those sites that are in rural locations near power generation assets. On the other hand, HPC data centers have little to do with the buildouts currently used for PoW computing—these sites can operate with much lower connectivity and are more rudimentary all-around. The below table illustrates the differences between traditional data center requirements, HPC data center requirements, and PoW data center requirements:

Requirement	Proof of Work Data Center	HPC Data Center	Traditional Data Center
Suitable Locations	**Nearly Anywhere with Power**	**Sites with Power and High-Speed Fiber**	**Specific Sites within Major Markets**
Amount of Low-Voltage Power	**High**	**Medium**	**Medium**
Power Availability (% of Hours)	**80%+**	**99%+**	**99.9%+**
Internet Speed Requirement	**Low**	**Fiber**	**Extremely Low Latency**
Skilled Workforce Requirement	**Low**	**Some**	**High**
Power Density within Site	**High**	**Medium**	**Low**
Capital per MW	**Low**	**High**	**Medium**

We note that the requirements for HPC data centers vary based on what kind of clients the data center hopes to attract. It is possible to implement cheaper systems for GPU cloud rendering applications, for example, relative to training Large Language Models (LLMs). The requirements for the internal network interfacing the compute nodes in intensive AI/ML training is specialized and higher cost.

Generally speaking, an HPC data center requires the following:

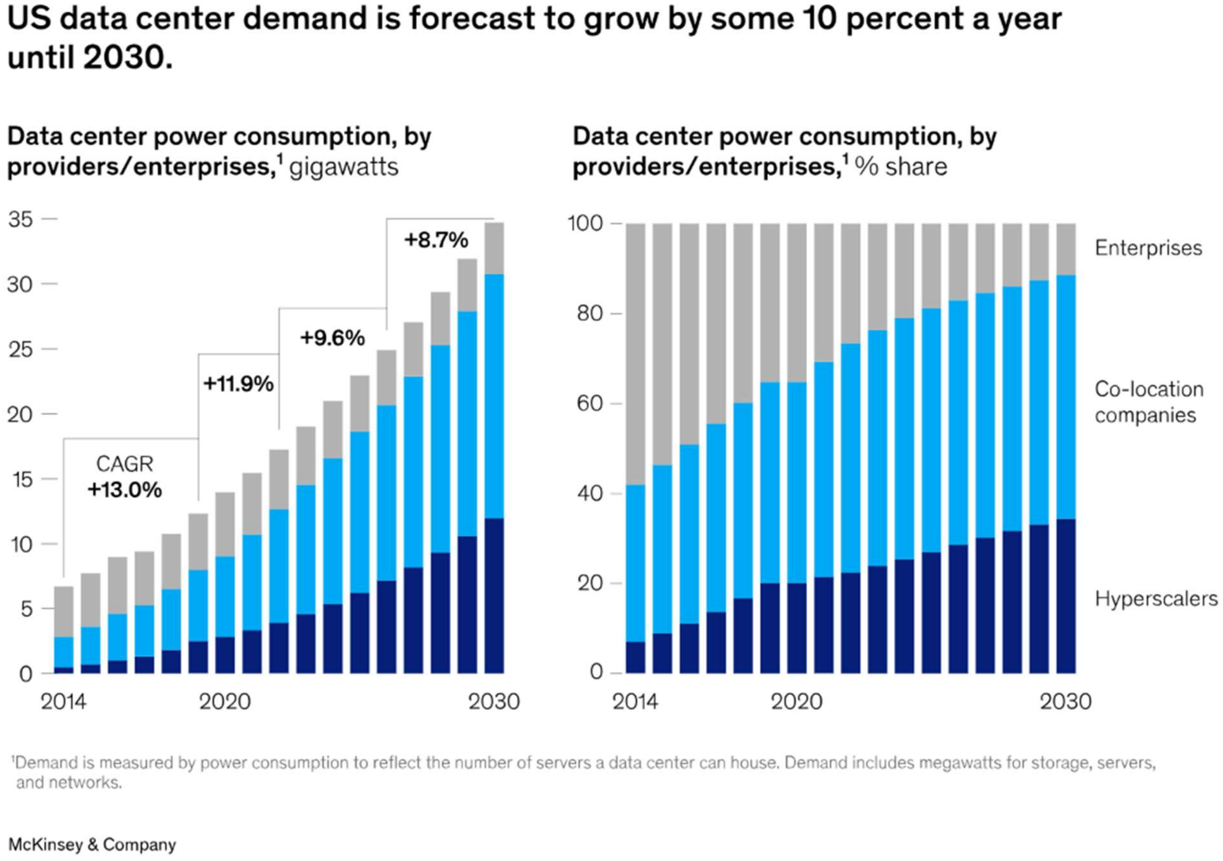
- Environmentally controlled building.
- Robust servers with fast CPUs, ample storage and memory, multiple slots for GPUs, fast network connections.
- Enterprise grade NVIDIA GPUs.
- Robust internal data center network.
- Reasonable distance to a Telco fiber drop. The cost of labor and materials for connecting to a Telco fiber backbone access point increases with distance.
- Redundancy. In order to reach more than 99% uptime, redundancy is a key requirement. There are several areas in which this can be incorporated. It can be taken to the extreme in the highest-end buildouts, which can become very expensive. Below are some examples outside of the main power and liquid cooling infrastructure.
    - Network:

        - Wide Area Network (WAN) connections.
            - Primary fiber connections to different providers with physically diverse cable paths.
            - Backup wireless connections (Microwave, 5G etc.)
        - Local Area Network (LAN) infrastructure within the location.
            - Redundant network switches and routers.
            - Redundant connectivity between compute nodes and network switches.
    - Server/compute design considerations:
        - Multiple storage drives on servers so that if one fails, there is no impact.
        - Using a pool of storage in addition to local storage drives on the individual servers with the GPUs. Think multiple racks with 100s of hard drives that the compute nodes are connected to.
        - Diverse cable paths within the data center.
        - Dual power supplies on devices connected to different PDUs which are connected to different breakers.
        - High Availability. Using virtual compute instances that are not dedicated to specific server hardware and can “float” around the compute infrastructure between different compute nodes.
- Access to water.
- Liquid cooling infrastructure is preferred.
- 480V 3 phase power. Step down to 200-240V.
    - Some newer data centers are using DC Voltage, which reduces amperage. It remains to be seen how common this will become.
- Building generator and UPS for IT equipment.
- A highly skilled technical remote team can support many locations at scale.

There are minimal onsite requirements for day-to-day operations for small-scale data centers, while large buildouts will need a larger team. This includes facility management, security, and ability to act as “smart hands” for remote technical team. There will be a need for on-call support to resolve downtime quickly.

Traditional Data Center Business:

The data center business has traditionally been split between companies who own and operate their own data centers (the majority up until recently) and real estate owners who seek to earn a return by owning this infrastructure. We will generally focus on the third-party ownership model here, as it is more relevant to new HPC developments in the coming years. Per a 2023 McKinsey report, enterprises that own and operate their own data centers will be <20% of the market by 2030 (chart below).



US data center demand is forecast to grow by some 10 percent a year until 2030.

Data center power consumption, by providers/enterprises,[1] gigawatts

Data center power consumption, by providers/enterprises,[1] % share

[1] Demand is measured by power consumption to reflect the number of servers a data center can house. Demand includes megawatts for storage, servers, and networks.

McKinsey & Company

Figure 7: McKinsey & Company: "Investing in the Rising Data Center Economy" January 17, 2023

Data center real estate is generally considered an alternative asset class akin to medical office / lab space or cell phone towers. This is due to 1) the relatively small size of the market versus bigger asset classes such as housing or retail, 2) the very specific infrastructure needs and use case for the asset class, and 3) the very high ratio of CapEx to land value. These factors have limited the amount of developer capital that focuses on this sector relative to traditional asset classes. That is not to say they aren't being developed—total data center power use is set to more than double from 2020 to 2030 (see chart above)—but that they remain a niche sector within the real estate space.

In the case of traditional data centers, they have generally been built in areas with abundant power infrastructure and large data pipelines. This allows for more efficient operations for the tenants. There are also synergies in having data centers physically close to each other because it reduces the latency when they "talk" to each other. These factors have led to the country's main hub existing in Northern Virginia (2.6 GW), with Silicon Valley, Chicago, Dallas and Phoenix comprising the other major hubs. The Northern Virginia hub is 4x the size of any other US market.

Securing land in a data center hub, low-voltage power and an extremely high-speed data connection provides a site where a data center can be built, but the majority of a data center's cost is in the physical infrastructure that is built (not the land). Per Raymond James' 2023 Telecommunications Services Outlook (published 1/18/23), the cost to build a data center is up to $1.8k per square foot. Building high-end office space in Manhattan costs under $1k / SF, and building standard condos in the Midwest costs under $300 / SF. Much of this higher cost is due to the dense electrical infrastructure, networking equipment, and especially redundancy of all the key items. Standard data centers are rated from Tier 1 to Tier 4, but in truth they are all trying to run nearly 100% of the time. Tier 1 runs 99.7% of the time and Tier 4 runs 99.995% of the time, per the Uptime Institute. This all requires redundancy and massive CapEx outlays.

This brings us to the users of traditional data center real estate—businesses who want their compute to happen offsite, but don't want to worry about downtime. This is because the end users do not expect downtime to be part of their experience. For example, if a user wants to access Gmail, they want it to work immediately. This is true for practically all cloud-based software, service, or storage solutions.

While HPC differs in a number of ways from the traditional data center asset class, we believe that HPC data center investments will largely be viewed through the lens of the existing data center paradigm. They may be smaller, more power intensive, and less latency-focused, but they are ultimately still a way to own infrastructure that converts power + data into compute in order to earn a yield on that infrastructure. This is likely to mean that existing players in the data center sector will follow innovators like Coreweave into building this new type of data center—they have the expertise, capital and mandate to do it, and it will allow their businesses to grow.

HPC Economics:

Here we will examine the economic profile of HPC data centers that could reasonably be located in rural areas near power plants ("**Distributed HPC**"). This can include AI training, graphics rendering, or other uses that involve massive computing resources without constant communication. AI inferencing, for example, would more likely be located within major data center markets. For the remainder of the memo, mentions of HPC will refer to Distributed HPC unless noted otherwise. We also use the term "**Powered Shell**." For our purposes, this refers to a data center that includes everything except the racks and servers (land, building, power infrastructure, networking infrastructure, power purchase agreement, high-speed fiber access, etc.). Given servers with GPUs are the most expensive component of a HPC data center, we have chosen to break them out from what we consider the underlying infrastructure.

The economics of a Distributed HPC deployment are distinct from those of a traditional data center and a Bitcoin / PoW data center. While HPC deployments are more power-intensive than a traditional data center, power costs are actually less relevant to an HPC data center simply because the margins are so high.

Even with low $30 / MWh power costs, a Bitcoin hosting facility will spend ~45% of its revenues on power alone. Traditional data centers have higher power prices given their infill locations and their need to draw power during peak periods, which causes their power costs to be ~25% of powered shell revenues.

HPC data centers, on the other hand, spend less than 15% of gross revenues on power costs (in a powered shell), and this can go as low as 1% in a vertically integrated buildout. The value of GPU-based HPC compute is so high in 2023 (if it can be delivered with 99%+ uptime and relatively low latency) that the power cost is simply not relevant to a successful deployment. **For example, a 1 MW high-end HPC buildout equipped with fully integrated NVIDIA A100 GPUs can produce nearly $40 million of revenue per year by selling that compute on a cloud marketplace. The same size PoW buildout would generate approximately $250k of revenue.**

The reason for this extremely high margin appears to be a shortage of high-performance compute resources relative to rapidly-increasing demand. Much like the market for PoW ASICs in H1 2022, these GPUs are made by relatively few companies (with NVIDIA having a monopoly on the very top of the market) and are susceptible to price bubbles. High-end NVIDIA A100 GPUs are currently selling for $13k - $15k apiece, and a 1 MW buildout would require 2,600 of them (total GPU-only cost of over $35mm). Specialized servers to run these GPUs for HPC functions are also in bubble territory, with a current additional cost per GPU of $15k. Lower-end HPC GPUs are still expensive, with RTX 4000 units trading for $700 apiece and a 1 MW buildout requiring 5,300 ($3.7mm in GPUs, roughly in-line with PoW ASIC cost / MW at the peak of the Bitcoin market).

In the context of these kinds of costs and revenues, the difference between $30 and $70 / MWh power is largely irrelevant. A user would never voluntarily curtail a compute function that was worth $2k - $8k / MWh in order to get slightly cheaper power, even if the activity could be toggled on and off seamlessly (which it is not set up for today). That said, we should expect the market to reach an equilibrium closer to that of the traditional data center as the market matures. We can see that even traditional data center infrastructure costs ~$9mm per MW to build and is effectively selling power to hosting clients at $600 / MWh. **Again, power costs are not nearly as relevant as CapEx costs or uptime in this asset class.**

[See table on following page]

Compute Type	Proof of Work	HPC - Low End	HPC - High End	BTC Hosting	HPC Hosting - Low	Traditional DC
Per 1 MW of Capacity Deployed	*M30S++*	*RTX 4000 8GB*	*NVDA A100 80GB*	*Powered Shell*	*Powered Shell*	*Powered Shell*
Power Usage Effectiveness ("PUE")	1.05	1.50	1.55	1.05	1.50	1.55
KW used for ASIC / GPU / CPU	952	667	645	952	667	645
KW draw per ASIC / GPU / CPU	3.20	0.13	0.25	3.20	0.13	N/A
Number of ASIC / GPU / CPU	298	5,333	2,581	298	5,333	N/A
Uptime %	95%	80%	90%	95%	80%	65%
2023 Rev / MWh Used in Compute	$80	$1,792	$7,740	$70	$469	$616
2023 Revenue / MWh of Capacity	$72	$956	$4,494	$63	$250	$258
Assumed Power Cost / MWh Used	$30	$50	$70	$30	$45	$103
Power Cost / MWh of Capacity	$29	$40	$63	$29	$36	$67
Other OpEx / MWh of Capacity	$16	$40	$40	$16	$40	$53
EBITDA / MWh of Capacity	**$28**	**$876**	**$4,391**	**$19**	**$174**	**$138**
Annualized EBITDA per MW ($mm)	**$0.25**	**$7.67**	**$38.47**	**$0.16**	**$1.52**	**$1.21**
EBITDA Margin	*39%*	*92%*	*98%*	*30%*	*70%*	*54%*
Land Cost / MW of Capacity ($mm)	$0.02	$0.10	$1.00	$0.02	$0.10	$3.00
Build Cost / MW of Capacity ($mm)	$0.45	$4.00	$6.00	$0.45	$4.00	$6.00
Server Cost / MW of Capacity ($mm)	$0.43	$9.60	$84.35	$0.00	$0.00	$0.00
Total Cost / MW of Capacity ($mm)	$0.90	$13.70	$91.35	$0.47	$4.10	$9.00
EBITDA Yield on Total Cost	**27.5%**	**56.0%**	**42.1%**	**35.1%**	**37.2%**	**13.5%**
Power Cost % of OpEx	**64%**	**50%**	**61%**	**64%**	**47%**	**56%**
Power Cost % of Revenues	**39%**	**4%**	**1%**	**45%**	**14%**	**26%**

HPC Fit with “Bitcoin Miners”:

We have recently seen many bitcoin mining entities pivot to HPC. While it may seem they are simply following the capital to the next buzzword, there are good underlying reasons for this pivot.

The fit between bitcoin miners and HPC is tangential, and different than it may first appear. Many in the “mining” sector initially viewed HPC as another type of flexible, curtailable load that could operate much like ether mining when it was secured by PoW and run on GPUs. As demonstrated above, this is not the case. Clients demand high uptime and power costs are not terribly relevant to the bottom line. HPC data centers look much more like a traditional data center than they do a PoW deployment.

That said, there are some good reasons to believe certain bitcoin mining entities will succeed in the HPC space. Crucially, Distributed HPC data centers can tolerate higher latency than traditional data centers. This makes it much easier to build them in remote areas outside of traditional data center hubs, which happens to be where bitcoin miners already operate. The table below shows the requirements for a Distributed HPC data center along with how well-suited traditional bitcoin miners and data center developers are to meeting each requirement:

Requirement	Bitcoin Miners	Data Center Developer
Fast access to large amounts of low-voltage power	Yes	Some
Power market expertise / access to cheap power	Yes	Some
Expertise building power infrastructure	Yes	Yes
Buildable land with fiber connectivity	Yes	Some
Expertise building redundant data centers	No	Yes
Experience with creative cooling solutions	Yes	Some
Relationships with server suppliers	No	Yes
Large team of development staff	No	Yes
Access to large amounts of capital	Some	Yes
Experience monetizing compute resources and managing sales of compute	No	Yes
A successful existing strategy in place competing for focus (inertia)	No	Yes
Financial incentive to enter new business line	Yes	Some

A good example of a PoW entity that is poised to succeed in the HPC space is Applied Digital. They have gained control of a site in Jamestown, ND with access not only to very cheap power, but direct fiber connectivity (See map below). They have successfully convinced parties like Super Micro Computer and Hewlett-Packard, as well as the market generally, that they will be building a competitive HPC data center on that site. We would expect many more “miners” to attempt to secure similar sites and use Applied Digital as a model for future capital raises.

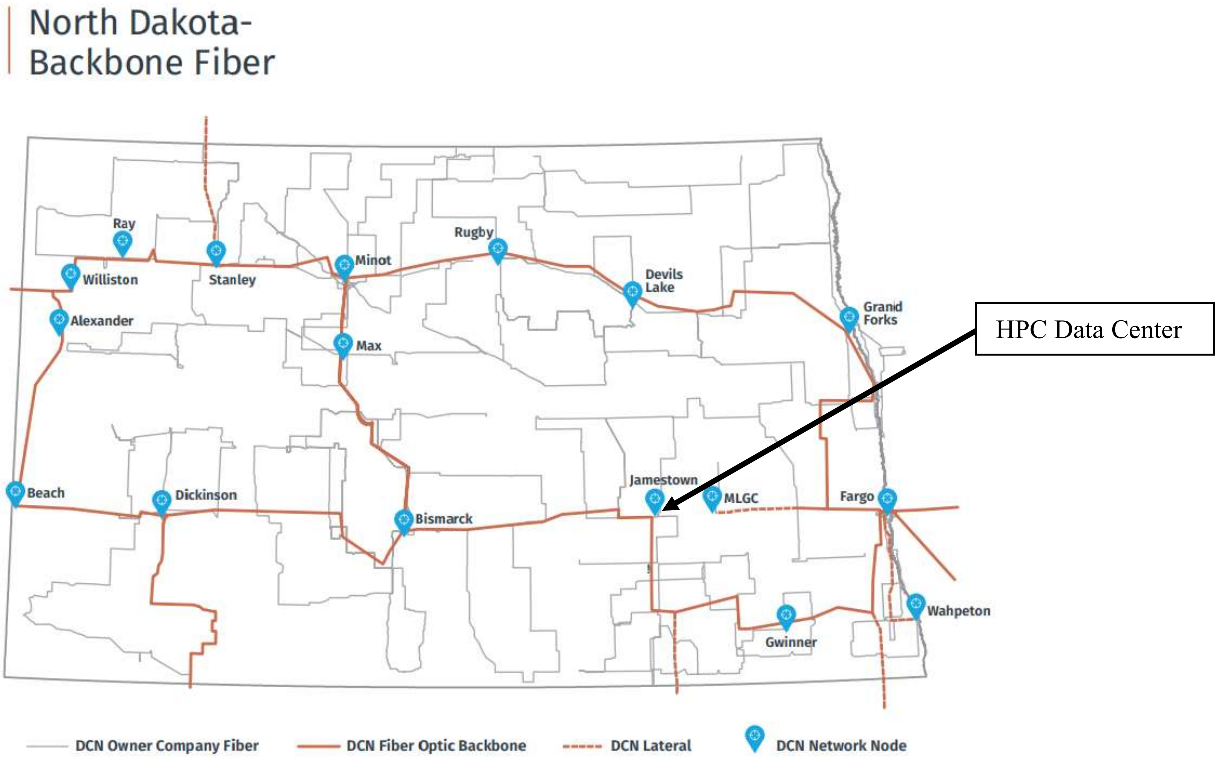


Figure 8: North Dakota has built out a comprehensive fiber network through its more populated areas. Source: Dakotacarrier.com

HPC Relevance to Energy Companies:

In the context of flexible loads such as PoW being a tool to help energy companies operate more efficiently, HPC is irrelevant. However, as shown above, McKinsey expects data center power consumption to more than double (grow by ~20 GW) by 2030. This may prove conservative given the rapid growth in AI computing demand. **Energy companies should view this as a growth driver for their industry and will most likely have strategic outlooks that include directly meeting this demand. Regulated utilities, who compete largely on growing the load profile within their footprint, should make conditions as favorable as possible for new HPC developments to be built within their service areas.**

Beyond the raw power necessary to run Distributed HPC data centers, the location of this power use is unique. While the compute function is not curtailable or nearly as flexible as PoW, the siting of new HPC data centers has much more flexibility than almost any other large load. This means HPC developments can be built in locations that are chronically congested and make use of less valuable power than those loads that are on the other end of the transmission line. In addition, the size can be customized, and HPC data centers today can be competitive at just 1 MW of load.

Power companies, particularly those with renewable assets in congested areas, can benefit from the extremely high value of HPC computing relative to power costs. It will be key for companies to move quickly, as any delays in energizing an HPC facility will be extraordinarily expensive (recall that all-in cost per MW can be >$90mm for high-end HPC data centers). However, those who do help energize facilities quickly will most likely find that HPC data center operators are far less sensitive to power prices than PoW operators and are more likely to be stable customers.

DPO Outlook:

Since February 2020, DPO has helped renewable energy companies evaluate their portfolios and find locations where flexible on-site data centers can create value for the owner of generation assets. This is our core business. We consider HPC to be another tool we can use in executing that strategy.

We have evaluated some sites where a flexible PoW data center does not work long-term because there are higher and better uses for that power, but an HPC data center would drive real value for our client. Other sites work well for both and could host multiple computing functions. We envision wind sites where the first ~10 MW consists of an HPC data center running over 99% of the time (mostly behind the meter but pulling some power from the grid) and there is another 20 MW of PoW computing which only runs when grid prices are suppressed. We will also work with our regulated utility clients on how best to attract new HPC load, large-scale or targeted, in their service areas.

DPO's unique position operating data centers on behalf of power companies gives us insight into how this new technology will impact those in the power generation business. We intend to use that expertise in helping to build out these new computing functions in a way that is most beneficial to our clients.

Authors:

Alex Stoewer, COO

David Carnes, Director of IT

Disclaimer:

All information presented in this memorandum is the product of DPO's research, unless credited otherwise. DPO is presenting this information for informational purposes only and DPO does not guarantee the accuracy of any information presented herein. DPO undertakes no obligation to update this memo, and the information contained herein may be updated, amended or superseded by subsequent documents or disclosures, or may become outdated and inaccurate over time.

Appendix

Key HPC Players:

There are a number of layers to the HPC value chain, with different market participants and very different business models in each layer. At a basic level, HPC is the conversion of power and data into valuable compute functions using infrastructure and bandwidth as the conversion mechanism. Some of that infrastructure is long-term physical infrastructure and some is shorter-lived IT equipment (including GPUs and servers). Key players in certain layers of the value chain are below:

Providing Power and Data:

- Owners of generation and transmission assets such as **Duke, PG&E, and Oncor.** These groups not only generate power but also build transmission and substations that serve new data centers.
- **Various PPA brokers.** These groups work to procure power at the lowest possible price, including any RECs or incorporation of time-matched green energy in order to meet a buyer's carbon footprint goals.
- **Telecom companies such as AT&T, Verizon, or Lumen.** Telecom companies with fiber sell access to very large bandwidths, and also can enter into agreements to expand their service to a site that is not directly on the fiber optic line (at a cost).

Developers / Landlords:

- Traditional data center developers such as **Equinix and Digital Realty.** These two publicly traded entities are by far the largest owners of co-located data center real estate. Their combined EV is over $130 billion. These entities are expert in identifying sites, securing land, securing power, designing cutting-edge deployments that meet the greatest need at the time, and managing sites efficiently on behalf of many clients.
- New HPC developers such as **Coreweave and DCX.** New entrants have appeared with a specific focus on building HPC-style data centers that are different enough from traditional data centers to require a slightly different business plan. Some entities like Coreweave are seeking to vertically integrate ownership of the data center with marketing of discrete units of cloud-based compute.
  - o *DPO's conclusion is that some "bitcoin miners" seek to become something like a data center developer, given their access to attractive sites and power.*
- End users such as **Chase, IBM or Verizon.** Many large entities have historically built and run their own data centers. This has meant that they need an internal team to undertake everything that data center developers do. This was 60% of the market in 2014 but is expected to be <20% of the market in 2030, as this function is largely outsourced to developers and colocation companies.
- EPC and construction firms such as **Schneider Electric.** These entities have expertise in building out data centers, as well as procuring the vast array of equipment necessary to operate one. They are hired by 3rd party owners to perform this task, and earn a large fee without taking material risk.

Makers of HPC IT Infrastructure:

- **NVIDIA.** This is the most critical company in the HPC sector today. Other chip makers are working on powerful GPUs meant for data centers, but at the moment NVIDIA has a near-monopoly on the GPU market. NVIDIA's proprietary software ecosystem (which is very widely used) makes it difficult for a company who is developing in NVIDIAs ecosystem to switch to a competitor without a material switching cost. NVIDIA's head start on GPU technology along with this software system help explain why their sales are growing so dramatically in 2023.
- Server manufacturers such as **Super Micro Computer Inc. and Dell.** These companies make different components (or sometimes all) of the servers that house the GPUs. As described in the sections above, GPUs require CPUs, storage, networking equipment, racks, software, etc. in order to provide the compute needed by clients. Server costs can be as much or more than the GPUs themselves.
- Other chipmakers / manufacturers such as **Intel, AMD and TSMC.** These legacy chipmakers will at a minimum provide many of the chips required in HPC server CPUs and may ultimately take some of NVIDIA's market share in advanced GPUs. It

remains to be seen how competitive they might be, but at a minimum they will be helped by the massively growing demand for compute generally.

Compute Marketplaces:

- Cloud companies such as **Microsoft Azure, Amazon Web Services, IBM Red Hat.** These entities either own their own data centers or lease powered shell space from landlords, use their own equipment, and sell their compute at a retail level. They each have proprietary software to allow users access to their systems. It is not cheap for end users, but it is convenient. While these groups are focused more on traditional data center functionality today, it will be relatively easy for them to compete in the HPC sector as well. As shown by the success of Azure and AWS, they are capturing a massive economic spread by offering their compute power and software while taking care of all of the hardware and logistics themselves. AWS alone generated over $80 billion in revenues and $23 billion in profits in 2022.
- Vertically-integrated new entrants such as **Coreweave and Crusoe.** These companies are looking to build and own vertically integrated HPC data centers while marketing the compute for hours, days or months at a time in an active online marketplace. Coreweave is currently executing this model while Crusoe is in the early stages of doing so.
- HPC-focused cloud compute marketplaces such as **Fluidstack and Cudo Compute.** These new companies seek to take advantage of HPC's unique demand structure by selling hourly, daily, or monthly access to certain specific servers or amounts of compute power. Unlike many traditional compute functions, much of HPC demand is for a specific amount of power for a discrete period of time. By providing an active marketplace for server owners who are willing to sell their compute at a given rate these companies create a visible revenue stream for owners of servers while making costs competitive for end users.

Differences Between HPC and Legacy Data Centers:

HPC data centers differ from traditional/legacy data center computing in the following ways:

1. Focus and Workloads:

- Legacy data center: Typically focuses on providing general-purpose computing resources to support a wide range of enterprise applications, including email servers, databases, web hosting, and business applications. It caters to traditional IT workloads, such as transaction processing, storage, and networking, with an emphasis on reliability, availability, and security.
- HPC data center: Is designed specifically to deliver exceptional computational power for artificial intelligence, 3D graphics rendering, complex scientific simulations, modeling, large-scale data analysis, and other computationally intensive workloads. It is optimized for parallel processing and high-performance computing tasks, utilizing specialized hardware like accelerators (e.g., Graphical Processing Units or GPUs) and interconnects to achieve maximum computational efficiency.

2. Key Differences in Infrastructure:

- Compute Nodes:
  - Legacy data center: The compute nodes (servers) are typically designed for general-purpose computing. They may have multi-core Central Processing Units (CPUs) with moderate processing power and may not include specialized hardware accelerators like GPUs.
  - HPC data center: Features compute nodes optimized for high-performance computing. These nodes often incorporate powerful CPUs with multiple cores, high-speed memory, and, most importantly, specialized hardware accelerators such as GPUs or Tensor Processing Units (TPUs). These accelerators significantly enhance the computational capabilities of the nodes, enabling parallel processing and accelerating tasks like scientific simulations and machine learning algorithms.
- Interconnects:
  - Legacy data center: Typically uses Ethernet as the primary interconnect for networking between compute nodes. Ethernet provides good connectivity but may have higher latency and limited bandwidth compared to specialized interconnects.
  - HPC data center: Employs high-speed interconnect technologies like InfiniBand optimized for low-latency and high-bandwidth communication between compute nodes. These interconnects enable efficient parallel communication and data exchange, critical for scaling and optimizing HPC workloads.
- Storage Systems:

    - Legacy data center: Commonly uses traditional storage systems like RAID arrays or Network Attached Storage (NAS) for data storage. These systems are designed for reliability, data integrity, and data access across various applications.
    - HPC data center: Requires storage systems that can handle large-scale data processing and high-speed Input/Output (I/O). They often utilize parallel file systems to provide high-bandwidth, distributed storage that can support simultaneous access from multiple compute nodes. These file systems are designed for high-throughput and parallel I/O, allowing HPC workloads to efficiently read and write large volumes of data.
- Power and Cooling:
    - Legacy data center: Power and cooling systems focus on maintaining a stable operating environment for general-purpose computing equipment. Redundancy and backup power solutions are typically in place to ensure uptime and data integrity.
    - HPC data center: Requires more substantial power and cooling capabilities due to the high-density computing and the heat generated by specialized hardware like GPUs. They may employ advanced cooling technologies, such as liquid cooling or specialized air cooling solutions, to manage the increased thermal loads. Power distribution and backup systems are designed to handle the power demands of high-performance computing equipment.

3. Scalability and Performance:
    - Legacy data center: Typically focuses on horizontal scalability, allowing organizations to add more servers or storage units to meet increasing demand. They prioritize reliability, stability, and data protection over extreme performance.
    - HPC data center: Emphasize vertical scalability, focusing on maximizing computational power and performance. They employ parallel computing architectures, distributed file systems, and advanced scheduling and workload management systems to efficiently utilize the available computing resources for massive-scale processing.
4. Software and Tools:
    - Legacy data center: Generally supports a wide range of commercial off-the-shelf software applications and operating systems commonly used in enterprise environments. They may also provide virtualization technologies for server consolidation and resource management.
    - HPC data center: Often requires specialized software stacks, libraries, and tools tailored for scientific computing and parallel processing. This includes software frameworks for parallel programming, specialized math libraries, and scientific visualization tools. HPC centers often provide compilers, schedulers, and job management systems designed specifically for HPC workloads.

Different Requirements for Different Applications:

Below, we look at some differences between the requirements for graphics rendering and AI/ML training systems. This illustrates that not all HPC is created equally, and certain deployments will have a very different physical and economic profile. Specifically, we show what we expect will be two of the larger markets for HPC compute, AI training and graphics rendering:

1. Hardware Components:
    - Graphics Rendering: Heavily relies on GPUs for real-time rendering and visualization. High-performance CPUs are also important for managing complex scenes and coordinating rendering tasks.
    - AI/ML Training: Workloads primarily focus on GPUs for parallel processing and accelerating computations. CPUs play a supporting role in managing the training process and coordinating data movement.
2. Memory and Storage:
    - Graphics Rendering: Often requires substantial memory (RAM) to handle large 3D models and textures efficiently. Fast local storage is beneficial for quick access to texture files and temporary rendering data.
    - AI/ML Training: Workloads demand significant memory capacity to handle the size of the datasets and model parameters involved. Fast storage, such as solid-state drives (SSDs), is helpful for efficient data access during training.
3. Networking:

- Graphics Rendering: Local network infrastructure is primarily used for distributing rendering tasks across multiple compute nodes or rendering farm systems.
- AI/ML Training: High-speed network connectivity, such as InfiniBand or Ethernet, is crucial for efficient communication between compute nodes during distributed training or when accessing large-scale datasets.

4. Software and Tools:
   - Graphics Rendering: Involves dedicated software along with specialized render engines for creating and rendering complex scenes.
   - AI/ML Training: Relies on deep learning frameworks such as TensorFlow and PyTorch along with specialized AI libraries and tools. These frameworks provide pre-built functions and algorithms for training and deploying AI models.
5. Data Management:
   - Graphics Rendering: Data management in graphics rendering revolves around scene files, texture assets, and intermediate rendering outputs.
   - AI/ML Training: Often involves managing large-scale training datasets, data preprocessing, and model checkpointing. Data pipelines and distributed storage systems are common components in AI/ML infrastructure.
6. Workload Characteristics:
   - Graphics Rendering: Tasks focus on real-time or near real-time rendering of complex scenes, often requiring interactivity and visualization capabilities.
   - AI/ML Training: Tasks are compute-intensive and involve training models using large datasets. The emphasis is on achieving high accuracy and performance, rather than real-time interactivity.