Revolutionizing Data Centers: A.I.'s Influence on Power Consumption

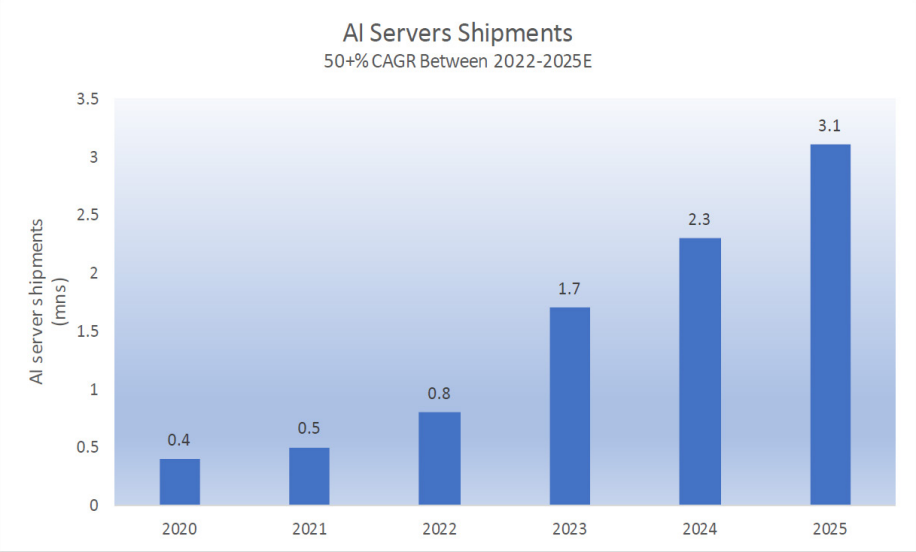
REGIONS ASSET MANAGEMENT

February 2024

AI Wave

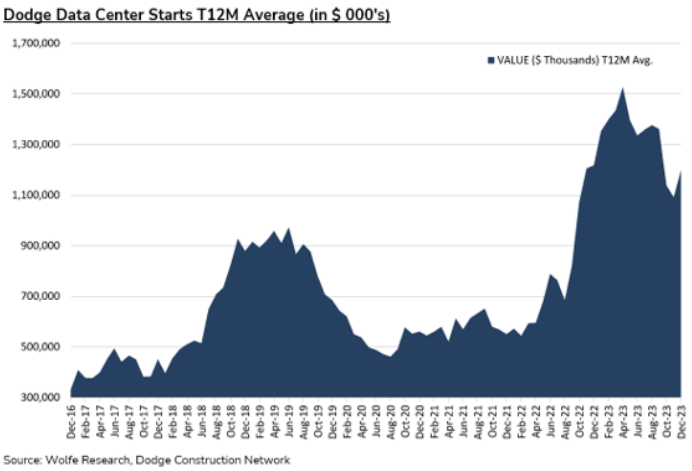
If an investor were to describe the market in 2023, they would likely highlight the dominance of AI and weight loss drugs. I will be focusing on the former and discussing how data centers housing these AI servers will require higher power densities. Additionally, I will explore which companies could benefit from this surge in power usage.

Recently, my colleague Anjali Tagare wrote a white paper titled "Generative Artificial Intelligence: The Next Disruptive Technology." According to her findings, Nvidia (ticker: NVDA) is a market leader for GPU (graphic processing units) semiconductor chips. During 2023, Nvidia rode the AI wave on its way to being the best performing stock in the S&P 500, up over ~230% for the year.[1] Anjali points out that Chat GPT was trained using 10,000 GPUs (NVDA, A100) and is expected to need over 30,000 GPUs going forward.[2] BofA's US semiconductors analyst, Vivek Arya, estimates there will be a rise in AI server shipments, more than doubling in 2023 and then further increasing by 35% in 2024. BofA anticipates AI server shipments to comprise 19% of all server shipments by 2025.[3]



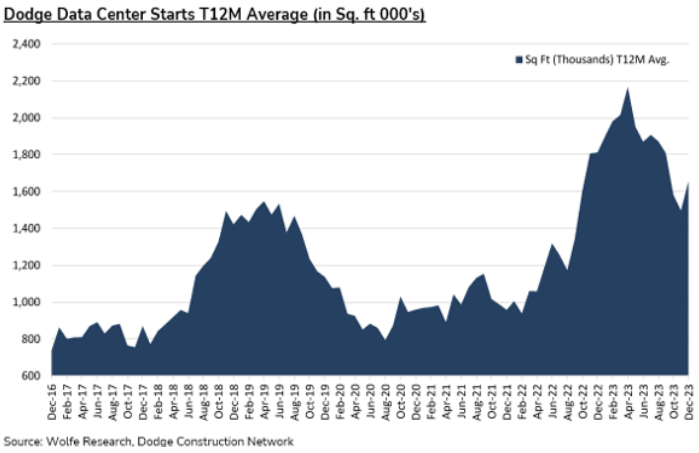
The Rise of Data Centers

The Dodge Construction Network, which provides data and analytics to the commercial construction industry, recently issued its data center starts on a dollar and square footage basis. On a dollar basis, data center starts increased +259% y/y in December and are up 129% on trailing twelve-month average compared to +111% in November. On a square footage basis, data center starts saw an increase of 280% y/y in December and are up +103% on a trailing twelve-month basis compared to +81% in November.[5]



Author: Christopher N. Walker, CFA
Regions Investment Management, Equity Analyst

During its 3Q earnings call, the CEO of Equinix, which is one of the largest data center real estate investment trusts (REIT), referenced a recent study found that 55% of organizations are in pilot or production mode with generative AI. Equinix CEO stated, "We are seeing this (AI) manifest in accelerated interest from both enterprise customers and from emerging service providers looking to service this demand. We see strong similarities between the evolving AI demand and the multi-tiered architectures that have characterized the cloud build-out for the past eight years."[4]



Source: Wolfe Research, Dodge Construction Network

Why this Matters for Data Centers

GPUs require roughly 2x to 3x the power versus traditional CPUs because GPUs are able to process simultaneously thousands of threads of historical data and statistical analysis to generate outputs. Typically, traditional CPUs have an average power density of ~ 10 kilowatts (kW) per rack, whereas AI applications generally use GPUs with an average power density of 20-25kW per rack.[5]

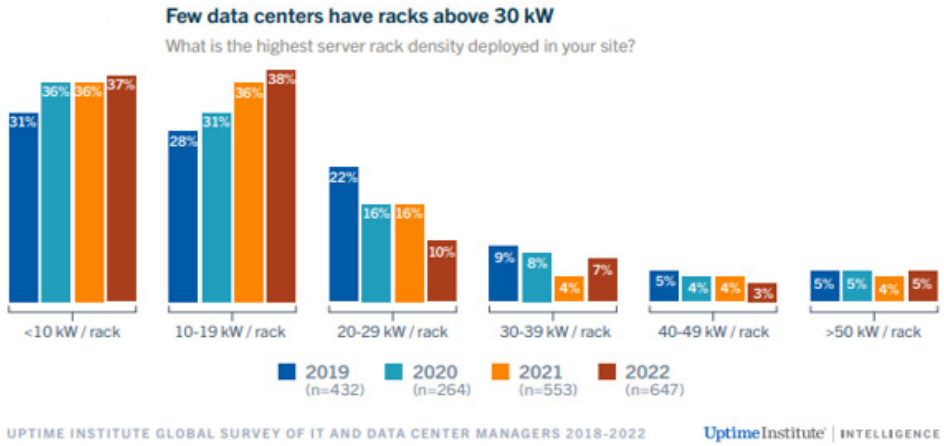
By using higher power density GPUs, data centers risk breaching the limit (25kW per rack) of air-cooled systems. The result of using more GPUs will require data centers that house AI servers to increase the adoption of liquid cooling solutions. Liquid cooling rack densities for AI applications can withstand 30-100 kW per rack. A prime example to better accommodate higher power densities is Meta (Facebook's parent company) pausing the buildout for some its data centers at the end of 2022. Meta announced that the halt in construction for twelve of its data centers was due to a redesign that would allow for more capacity for AI-centric designs. Meta has since restarted two of the sites in Texas and Idaho with each data center being able to support higher GPU capabilities for AI. [6]

Racks are arranged vertically, with all of the various IT equipment sitting on its shelves. The standard rack height is 7 feet and are normally divided into are 48 rack units.[7] The racks house and protect data center equipment such as servers, routers, switches, hubs, and audio/visual components. The data center IT equipment can get very hot, so cooling is required to keep them from overheating and for proper operation. Racks in data centers are either air-cooled or liquid cooled. Traditional data centers that house CPUs are typically filled with 24 servers each with two CPUs on each rack that pull out 150 watts.[8] For traditional CPUs, when considering power storage, networking, and other components, they produce a total of 10kW per rack. BofA provides a breakdown of power usage for traditional CPUs versus GPUs for AI applications, using the Nvidia Hopper 100 GPU chips as an example below.

Comparing Rack Power Density in Traditional vs AI Applications

	AI	Traditional
Rack height(in rack units)	48	48
divided by Server height (in rack units)	4	2
Servers per rack	12	24
x GPUs or CPUs/Server	3.5	2
GPUs or CPUs/Rack	42	48
x Watts/GPU or CPU	700	150
kW/rack for GPUs or CPUs	29	7
kW for storage, networking etc.	3	3
Total kW/rack	32	10

According to the Uptime Institute, there is ~250mn square feet of data center infrastructure globally with an average power density of ~10 kilowatts (kW) per rack and of the 647 data center managers that were surveyed, 85% of data centers operate with maximum rack power densities of less than 30kW per rack.[9]



Reproduced by Permission: Uptime Institute

The traditional air-cooling falls short for AI ready data centers, particularly for large language models that are equipped with a number of NVIDIA equivalent chips. The extensive power consumption and heat rejection linked to these large language models surpasses the capacity for current air-cooling methods. There are a few alternatives to the conventional air-cooling systems:

1. Rear Door Heat Exchanger - a fan less heat exchanger module that installs as the rear door of an equipment rack. This chilled water-based system provides up to 50kW of room-neutral cooling.

2. Direct to Chip - places a cooling plate on top of the chip. Only 5% of data centers have any form of liquid cooling according to nVent. Process uses a cold plate that receives the heat from the chip and into the fluid, which is then circulated back to a cooling distribution unit (CDU).

3. Full Immersion - immersion cooling involves fully plunging each component (CPUs, GPUs, memory cards and power supplies) into thermally conductive liquids like mineral oil or dielectric fluids. These liquids can efficiently transfer heat away from the components while protecting them from potential damage due to corrosion or oxidation. Immersion is still seen as being a long way off, and more of a niche application at this point in time. It has cooling densities of over 300kW at a lower cost, although it is not quite at the commercial deployment stage and bears high upfront costs. Bitcoin miners utilize immersion cooling as it cuts down on power costs, however, immersion cooling systems entail high initial costs.

Which Companies Stand to Benefit

According to Schneider's Energy Research Management Center, electrical demand for AI could grow at a 25-33% CAGR over 2023-2028. [10] SC23, which is the international conference for high performance computing, networking, storage, and analysis, was recently held in Denver where attendees of the conference compared the AI craze to dawn of the internet era in the late nineties.[11] Assuming AI is not a bubble, and the hype is real, a few companies may benefit from higher power densities and heat management in an AI world are the "outside the rack companies" like Vertiv (VRT), Eaton (ETN), nVent (NVT), and HVAC players. .

Vertiv has a dominant position in thermal and power management, with 85% of revenues tied to data centers.[12] The company is based just outside Columbus, OH and recently went public in 2018. Emerson Electric (EMR) sold its business unit that was then known as Emerson Network Power to Platinum Equity, who then rebranded the company under the name Vertiv. Its main markets are the US (45% of revenues), China (13%), and other parts of Asia (13%).

Eaton generates approximately 15% of revenue from data centers.[13] The company sells power distribution units (PDUs), uninterruptible power supplies (UPSs), stepdown transformers, power strips, and different rack level cooling offerings. Eaton is based in Dublin, Ireland. Its main markets include the US (60% of revenues), China (3%), and parts of Europe (20%). In their recent 4Q earnings report, Eaton pointed to their data center revenue growing 20% year over year in the fourth quarter, with trailing-twelve-month orders up 30% year over year, and the negotiation pipeline up an impressive 160% year over year.

nVent is another company that appears well positioned in the thermal management space. NVT derives 20% of total revenues from data centers.[14] Their key offerings include cooling distribution units, power management, and racks/server cabinets. nVent is based in a variety of liquid cooling rack products that have a capacity up to 100kW. Based in London, NVT's main markets are the US (63% of revenues) and Europe (20%).

Summary

Bears will argue that the aforementioned companies are relatively well priced in, and 2024 could be a low growth environment. The bulls will say it's just a matter of keeping up with demand as AI workloads become increasingly more widespread. With the highlighted companies all having a vast line of liquid and rack cooling products at their disposal, many also have exposure to the power side of the business that could also profit from this trend. For instance, both Eaton and Vertiv have power management products such as busways, power distribution units (PDUs) and switchgear infrastructure, as well as uninterruptible power sources (UPS), which Caterpillar also engages in. Lastly, as the market evolves towards higher density AI workloads, and immersion gains traction, we anticipate many niche companies that focus exclusively on full immersion could be acquired by more established companies. The consensus is that traditional air to air solutions do not work for GPU cooling. Therefore, companies providing both hybrid air-to-liquid solutions and liquid solutions should capitalize on this theme as opportunities for retrofitting and new buildouts for data centers progress.

References:

1. FactSet Market Data
2. Generative Artificial Intelligence: The Next Disruptive Technology
3. https://rsch.baml.com/report?q=pKbrkwqiO5FWO5cwJNWooA&ps=true&pv=validated
4. Equinix 3Q Earnings Transcript
5. https://rsch.baml.com/report?q=UsuIhWrW-zhDAlkeMqi8SQ
6. Meta 3Q Earnings Transcript
7. https://rsch.baml.com/report?q=UsuIhWrW-zhDAlkeMqi8SQ
8. https://rsch.baml.com/report?q=UsuIhWrW-zhDAlkeMqi8SQ
9. https://uptimeinstitute.com/uptime_assets/6768eca6a75d792c8eeede827d76de0d0380dee6b5ced20fde45787dd3688bfe-2022-data-center-industry-survey-en.pdf
10. Schneider Electric White Paper, The AI Disruption: Challenges and Guidance for Data Center Design https://www.se.com/us/en/download/document/SPD_WP110_EN/
11. https://research.wolferesearch.com/file/38548.pdf?docRef=b708f3d1-824c-4457-b5f5-bbc081e20ca5&ch=2&forceDownload=true
12. FactSet Market Data
13. FactSet Market Data
14. FactSet Market Data