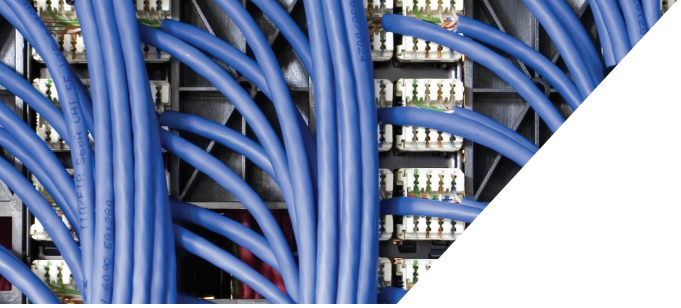


May 2024

The Rural Revolution is Coming as Generative AI Drives Hyperscale Data Centers

Key Points:

- Generative artificial intelligence has taken the data center market by storm and operators are scrambling to stand up new locations to meet its surging demand.
- New chipset technologies from Nvidia are enabling data centers to process unthinkable amounts of data and complex applications that promise to be the backbone of the next industrial revolution.
- Data center demand is wreaking havoc on the energy complex. AI data centers could consume as much as 20% to 25% of U.S. power requirements by 2030, up from 4% today.
- The imbalance in energy markets' supply and demand – combined with generative AI's energy-intensive data processing (training versus inference) – is forcing data center operators into secondary and rural markets.
- These trends present new business opportunities for rural broadband operators that could be parlayed into helping close the digital divide. Data centers also mean new opportunities for rural electric distributors and generation and transmission companies, albeit with greater risk and capital requirements.

Jeff Johnston
Lead Economist,
Digital Infrastructure

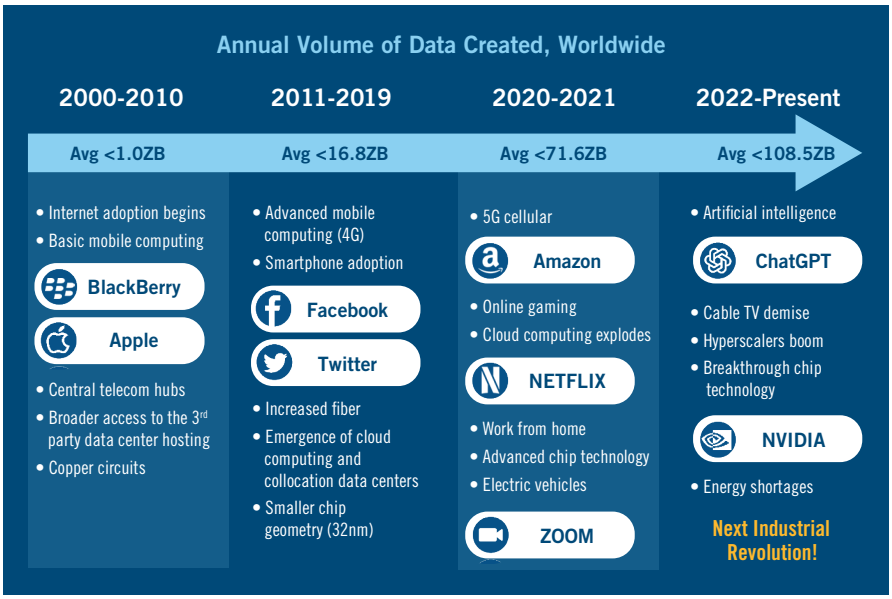
Inside…

Introduction

Generative AI is set to be a game-changer for the hyperscale data center landscape, prompting companies to build in new areas and benefiting rural America in the process. Data center companies have historically built hubs in urban areas for access to power and water, deeply integrated fiber networks, and proximity to Fortune 500 companies and government agencies. However, the processing dynamics of generative AI and the enormous power requirements of AI models are turning that model on its head. Data center operators recognize that moving gigabits of data is easier than moving electrons and are locating where land is available and the path to meet their power needs is clear. Initially they are targeting secondary markets in areas of Ohio, Mississippi, Utah and other states. Over time, we expect new data centers will be built outside of secondary markets and deeper into rural America.

In this report we look at how the digital infrastructure landscape and data creation has evolved, the profound impact generative AI is having on the energy complex, and the opportunities these trends represent for rural broadband operators and rural electric cooperatives.

EXHIBIT 1: Evolution of the Digital Infrastructure



Source: CoBank and Statista

How did we get here?

Data creation and network technologies have come a long way over the last two decades *(Exhibit 1).* From the early days of the internet 20 years ago when copper circuits and basic 3G wireless coverage were all that we had, worldwide data creation was well below 1 zettabyte (ZB) per year. But as smartphones gained traction and new chipset technologies enabled devices and data centers to store greater amounts of data and to run more complex applications, data creation grew to an average of 16.85 ZBs per year. Then COVID hit and we saw two years' worth of digitization in two months, resulting in "hockey stick" type growth in data generation. And just when the industry thought growth would normalize, generative AI came along and blindsided us with unthinkable growth in new technologies and data center demand.

The generative AI hype is red hot and by most accounts it is justified based on the billions of dollars in capital being deployed by Amazon, Microsoft, Google and Meta. Generative AI is not a new vertical category like Google glasses or the metaverse. Instead, it will be deployed horizontally across devices and transcend industries delivering operational efficiencies and new products and services. It is for these reasons that we believe the technology has significant staying power.

It's important to recognize that hyperscale data centers are not being built on spec – this is not a "build it and they will come" scenario. Instead, data center operators are still trying to keep up with the post-COVID digitization trend, and still have a long way to go before achieving a supply/demand balance for generative AI.

Generative AI and data centers

AI has been around since the 1950s in a program called the logic theorist, a system designed to mimic the problem-solving skills of a human. Since then, we have seen several iterations of AI ranging from Apple's Siri to the way digital ads target users online. The key difference with generative AI and why it has the potential to be a transformative technology is how it teaches itself and how this improves the utility of the technology. Before generative AI, humans taught computers through software coding. With generative AI, computers teach computers which is exponentially accelerating software development and data creation.

At the core of generative AI is the semiconductor company Nvidia. The company's breakthrough graphical processing unit (GPU) chipset technology is enabling large language models (LLMs) to run deep learning algorithms that can recognize, summarize, translate, predict, and generate content using enormously large data sets. All of this is underpinning the unprecedented demand for data centers, storage and compute resources.

U.S. data center absorption rates (amount of new data center capacity used in each year) is surging *(Exhibit 2).* For example, absorption rates increased by ~2750 MWs from the first half of 2023 to the second half, an increase over a 200%. The current rate of new data center construction portends absorption rates that will accelerate in the coming quarters. Vacancy rates are falling fast, reaching 2.4% in the fourth quarter of 2023, a 33% year-over-year decrease that is putting upward

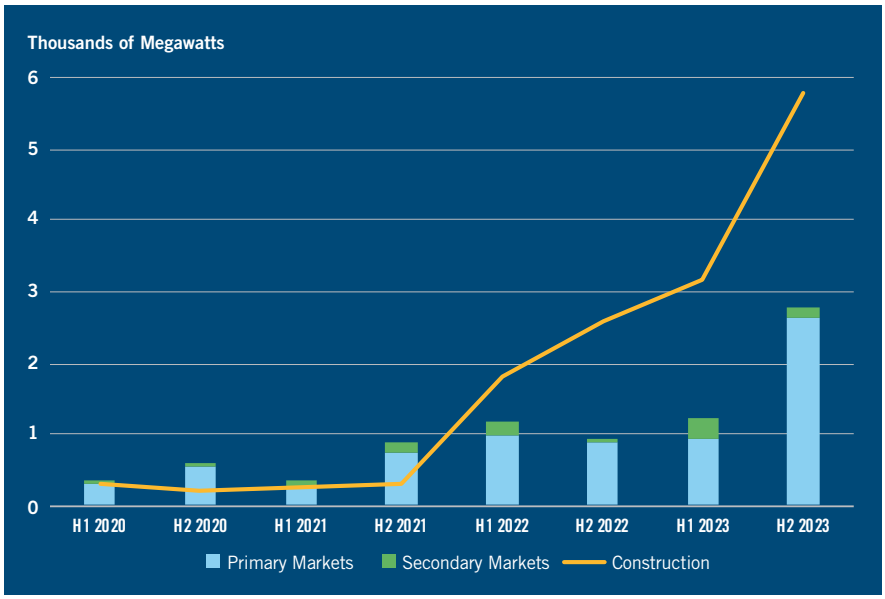
pressure on rental rates. In the first half of 2023, average rental rates in primary markets were up 7.2% over 2022. And in 2022 rates were up 14.5% over 2021. Prior to that, rental rates were declining year-over-year since 2015.

Data center capex is set to explode. Counterpoint Research forecasts total global spend to exceed $400 billion combined in 2024 and 2025, with hyperscale capex representing over 64% of the spend.

HELP WANTED: Electrons!

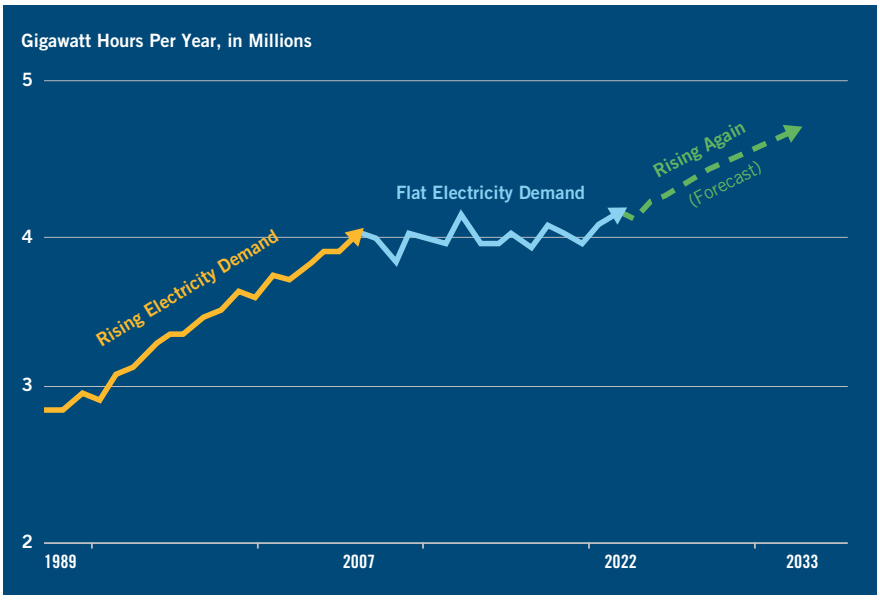
Nvidia GPU-powered generative AI applications require significantly more computing resources than legacy AI applications running central processing unit (CPU) chips. In January, the International Energy Agency said a request to ChatGPT requires 2.9 watt-hours of electricity – that is nearly 10 times as much as the average Google search. Couple that with the explosive growth in AI data centers and you end up with the current supply/demand imbalance in the energy markets. In fact, data center-fueled demand for electricity is growing so fast that it is forcing some states to dramatically revise their demand forecasts in real time. Georgia increased the amount of new electricity it expects it will need to bring online by 2040 by 38% approximately one year after setting its initial forecast. And the CEO of chip-design company Arm said that without greater efficiency, AI data centers could consume as much as 20% to 25% of U.S. power requirements by 2030, up from ~4% today.

EXHIBIT 2: U.S. Data Center Absorption and Construction



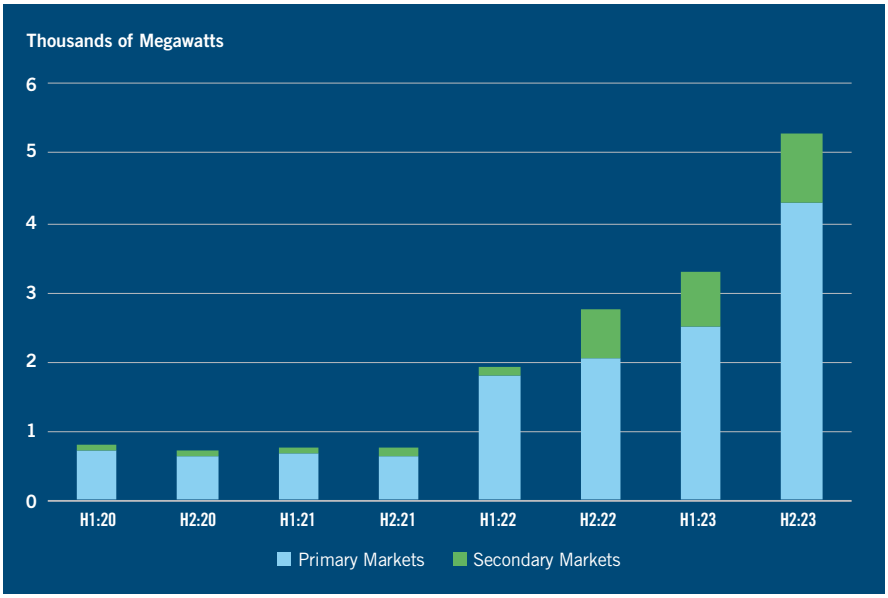
Source: JLL Research

EXHIBIT 3: U.S. Electricity Demand



Source: Data via the North American Electric Reliability Corporation, chart published in *The New York Times*, March 14, 2024. The data reflects annual net energy for load for the United States only, but select years include small portions of Mexico and Canada.

EXHIBIT 4: Data Center Construction



Source: JLL Research

Approximately 1,000 MWs were under construction in secondary markets during the second half of 2023, up from 750 MWs in the first half of 2023 and the second half of 2022. This marks a significant increase over previous periods when construction hovered around 100 MWs.

Beyond access to power, the greater storage and computing requirements for generative AI applications is also driving data centers deeper into secondary and rural markets.

Generative AI has two core characteristics: training and inference. In training, a model analyzes vast amounts of data to develop a knowledge base, like how a child absorbs information during her formative years. The model then transitions to the inference phase where it uses the knowledge base developed in training and predicts outcomes based on new data. These models operate in a virtuous cycle – during inference, the model gets feedback about what it liked, what worked, what didn't work, etc. and then feeds it back to the training model to further develop the knowledge base, which ultimately improves the outcomes during inference.

From an infrastructure perspective, training can be done in secondary and rural markets because the process is not latency dependent. As training requires a significant amount of processing horsepower and energy, non-core markets are ideal candidates for these data centers. Inference is performed closer to where applications are being used as they tend to have stricter latency requirements and are not as compute intensive as training. As generative AI grows in popularity, the virtuous cycle of training and inference will result in an exponential growth of data that needs to be stored and processed, with secondary and rural markets playing an increasingly larger role in the data center architecture.

All of this is putting the energy complex in unfamiliar territory. Electric cooperatives and generation and transmission companies are being asked to spend capital to support this explosive growth – driven by a small number of companies – and agree to terms they are not used to. Over the last 15 years, energy demand has been relatively flat *(Exhibit 3)* as consumer electronic companies designed more energy efficient products and we used more efficient sources of lighting, among other factors. This predictable environment has been the status quo that the industry has been working in; now, almost out of left field, it is being upended by a massive spike in demand thanks to generative AI. And if all that wasn't enough, a major challenge will exacerbate the current energy supply/demand imbalance: It takes a lot longer to build out new sources of energy than to build a data center.

New markets

As a result of the energy issues and how AI applications are being managed, data center operators are moving into secondary markets and exploring rural areas where there is a clearer path to sources of energy. According to JLL Research, data center construction in secondary markets has been growing over the last six quarters *(Exhibit 4)*.

Rural infrastructure

Supporting hyperscale deployments in rural America will require a significant amount of new infrastructure and capital. For example, it's likely these locations lack the required broadband connectivity. This presents an opportunity for broadband operators to partner with and provide connectivity to a long-life asset (think 25-50 years). And perhaps this investment could be parlayed into offering fiber-to-the-home in underserved/unserved neighboring towns.

For energy companies, the opportunity is much more involved. The opportunity for some is a step function increase in demand, but it requires a lot of capital and comes with increased risk. Some electric distributors are leery of making these kinds of commitments given the challenges they have had with colocation operators. However, the hyperscale model is different and has less risk. None the less, rural energy providers need to explore the risks and opportunity costs of supporting hyperscale data centers.

Conclusion

The AI hype is real, and the technology is set to play a larger role in how we live and work. From a labor market perspective, the country's unemployment rate is still very low with more job openings than people looking for work. Generative AI can deliver operational efficiencies enabling companies to do more with less. All of this suggests that the demand for new data centers, energy and broadband show no signs of slowing down, and rural America is being looked on to serve this need. However, it's not all rainbows and butterflies for rural infrastructure providers as the capital required and risk exposure is pushing some outside of their comfort zone. A careful analysis and a deep understanding of the digital infrastructure ecosystem is imperative for those looking to enter the market. It's not for everyone, but the growth opportunity cannot be denied. ■

Sources Used

"Large Language Models Explained," Nvidia website, accessed April 15, 2024. https://www.nvidia.com/en-us/glossary/large-language-models/

"North America Data Center Trends H1 2023," CBRE website, Sept. 6, 2023. https://www.cbre.com/insights/reports/north-america-data-center-trends-h1-2023

Anton Shilov. "Nvidia's H100 GPUs will consume more power than some countries — each GPU consumes 700W of power, 3.5 million are expected to be sold in the coming year," Tom's Hardware website, Dec. 26, 2023. https://www.tomshardware.com/tech-industry/nvidias-h100-gpus-will-consume-more-power-than-some-countries-each-gpu-consumes-700w-of-power-35-million-are-expected-to-be-sold-in-the-coming-year

Peter Landers. "Artificial Intelligence's 'Insatiable' Energy Needs Not Sustainable, Arm CEO Says," The Wall Street Journal, April 9, 2024. https://www.wsj.com/tech/ai/artificial-intelligences-insatiable-energy-needs-not-sustainable-arm-ceo-says-a11218c9

CoBank's Knowledge Exchange Division welcomes readers' comments and suggestions. Please send them to KEDRESEARCH@cobank.com.