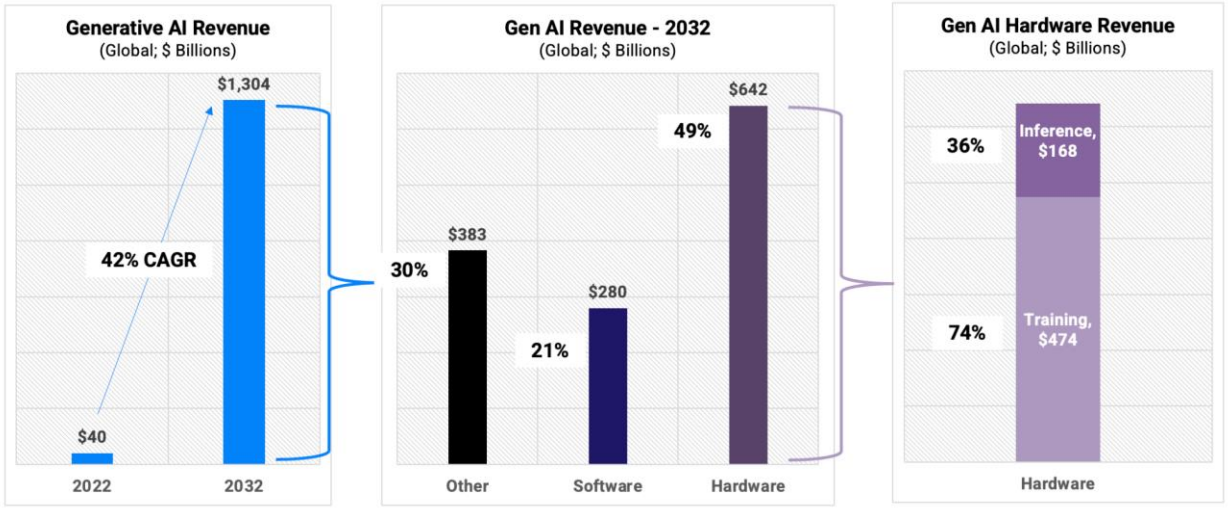
The AI Data Center Dilemma: Skyrocketing Growth vs. Sustainability – Can We Survive the Clash?

Source: Gemini AI

In the relentless race of technological advancement, Artificial Intelligence (AI) is not just a player; it's the game-changer. AI is revolutionizing industries, altering our world in ways once deemed impossible. But there's a ticking time bomb at the heart of this revolution – the data center industry. As the number of AI-driven applications increases, the insatiable thirst for computational power is driving data centers to the brink of their capabilities. This surge in demand comes with a steep cost: a significant increase in energy consumption and associated sustainability concerns. Yet, in a twist of fate, AI itself might be the savior we desperately need to tackle these looming threats. Will we harness AI's power in time to avert disaster?

The Opportunity: Generative AI Is Driving Data Centers to Surge

We are on the brink of a transformative era in artificial intelligence. Until now, machines have never been able to exhibit behavior indistinguishable from humans. However, new generative AI models are not only capable of engaging in sophisticated conversations with users but also creating seemingly original content. The implications for enterprises across various industries are massive, prompting companies to swiftly implement generative AI initiatives. The global Generative AI Market is expected to reach $1,304 billion by 2032, with a staggering compound annual growth rate (CAGR) of 42% from 2022 to 2032[1].



Source: Bloomberg June 2023; JLA Analysis

Generative AI leverages Large Language Models (LLMs) that require vast amounts of training data. For example, Google's Minerva model boasts 540 billion parameters. Training a 1-trillion-parameter model on Nvidia A100 GPUs can cost approximately $308 million over three months[2]. This immense computational demand highlights the necessity for efficient and powerful data center infrastructures to support such intensive training processes. The relentless advancement of AI is driving an unprecedented demand for data centers.

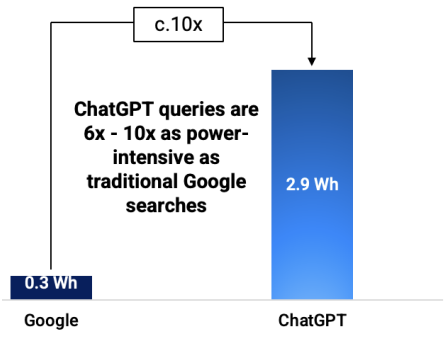
By 2030, generative AI is estimated to drive demand for 15 million incremental servers, significantly increasing the demand for data centers[3].

The Impact: Average Rack Density will Significantly Increase

Data centers are rapidly transforming to meet the unprecedented demands of artificial intelligence. Today, AI workloads utilize a combination of CPUs and GPUs for both training and inference. However, by 2030, AI accelerators powered by specialized ASIC (Application-Specific Integrated Circuits) chips are expected to dominate most AI workloads in data centers.[4]

To satisfy the immense computational demands of AI, data centers are significantly increasing rack density. By 2027, hyperscale data centers are projected to reach an average rack density of 50 kW per rack, up from 36 kW in 2023[5]. This densification is critical to managing the colossal computational load required for AI workloads during both the training and inference phases. For instance, ChatGPT queries consume approximately 6 to 10 times more power than traditional Google searches[6]. Consequently, AI data centers will drive exponentially higher power demand compared to traditional data centers.

Power Consumption per query/search
(Wh)



Traditional x86 vs. AI Server Power Consumption

Server Rack Type	Cooling	Servers	Storage	Networking	Other	Total Power Consumption
Traditional x86 (5-15 kW)	35%	25%	15%	10%	15%	100%
AI (50-100 kW)	45%	35%	7%	5%	8%	100%

AI servers are a lot more energy-intensive than traditional servers

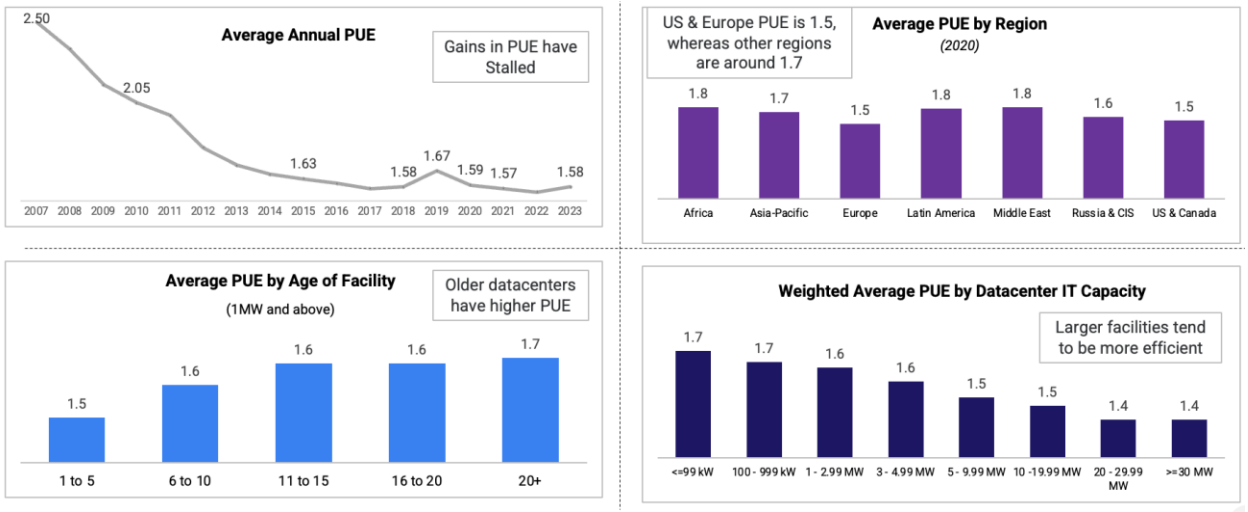
Source: Google, SemiAnalysis, Goldman Sachs, 650 Group; JLA Analysis

The Problem: Escalating Power Demand and Efficiency Stagnation

The adoption of AI is set to drive a staggering 160% increase in data center power demand by the end of the decade. According to Goldman Sachs, data centers currently account for 1%-2% of global power demand, and by 2030, this figure could soar to 3%-4%[7].

The data center power surge coincides with a deceleration in power usage efficiency, significantly compounding the sustainability challenge.

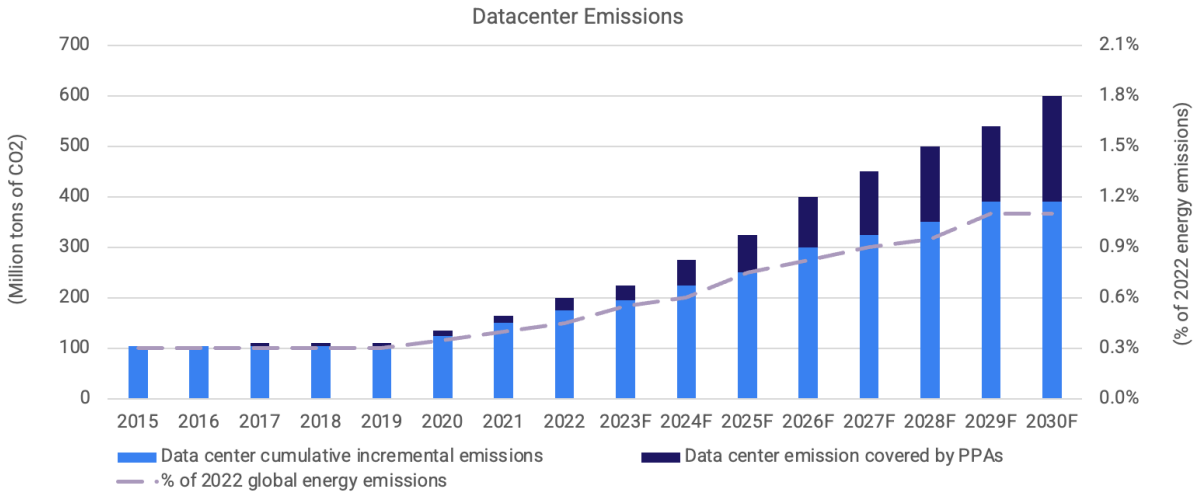
Data centers have long pursued improved power efficiency, measured by Power Usage Effectiveness (PUE). However, since 2020, gains in PUE have stagnated, raising major concerns. This plateau is particularly alarming given the escalating power requirements driven by AI technologies. While regions like the US and Europe have achieved a PUE of around 1.5, other areas lag, averaging closer to 1.7. Older data centers, in particular, struggle with higher PUE, indicating inefficiencies that are increasingly difficult to overcome with aging infrastructure[8].



Source: Uptime Institute, JLA Analysis

Major Push to Make Data Centers More Sustainable

The sharp rise in power demand from data centers will result in a significant increase in CO2 emissions. Projections indicate that emissions will double by 2030 compared to 2023 levels, even accounting for power purchase agreements (PPAs) from technology companies[9]. This underscores the urgency for data centers to reduce energy consumption and enhance sustainability, as their environmental impact becomes increasingly evident.



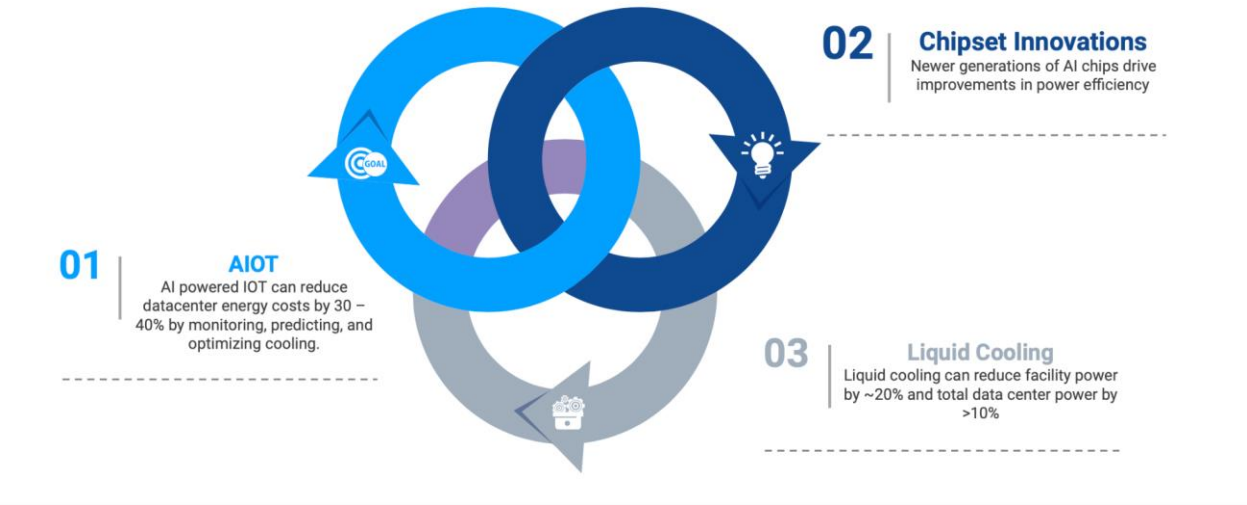
Source: Goldman Sachs Global Investment Research; JLA Analysis

Data center operators are under immense pressure from all stakeholders to slash power consumption and boost operational efficiency. Investors, employees, and tenants are demanding enhanced sustainability, while governments in Singapore, Amsterdam, and Germany are mandating that all new data centers achieve PUE levels of around 1.2-1.3[10].

This relentless pressure has made energy optimization the most formidable challenge and top priority for data center operators. The stakes have never been higher: those who fail to adapt risk falling behind in an industry where efficiency and sustainability are no longer optional but imperative. The race to optimize energy use is not just about cutting costs; it's about leading the future of data-driven innovation in a world that can no longer afford inefficiency.

AI and Technology Provide the Solution

Technological advances have historically played a more significant role in avoiding emissions than even renewable energy generation. AI presents the solution to the problem it presents. Data center operators can leverage three innovative technologies to improve sustainability and enhance efficiency:



Source: NVIDIA, Goldman Sachs; Fierce Electronics; JLA analysis

- **AI-Powered Internet of Things (AIOT)**

AIOT and operational technologies can monitor, predict, and optimize energy usage in data centers, reducing energy costs by 30-40% and meeting sustainability objectives. For instance, Google's DeepMind AI platform has demonstrated a 40% reduction in cooling costs, equivalent to a 15% reduction in PUE, by using neural networks trained on various operating scenarios[11].

- **Innovations in AI Chipsets**

Chipset manufacturers are pushing the boundaries of physics to increase energy efficiency. Newer chipsets not only elevate max power consumption per server but also significantly boost computing speed, leading to meaningful reductions in power intensity[12]. Innovations like Nvidia's DGX series and Sapeon's X330 chip represent leaps in computational performance and power efficiency.

- **Liquid Cooling Technologies**

Data center operators and chip manufacturers are embracing liquid cooling technologies such as direct-to-chip and immersion cooling to reduce energy costs. Liquid cooling involves the use of coolants to absorb and remove heat directly from the components, offering a more efficient thermal management solution than traditional air cooling. Liquid cooling technology can reduce facility power by approximately 20% and total data center power by over 10%[13]. Companies like Nvidia, Meta, Microsoft, and Google are leading the way in developing and deploying these technologies.

Conclusion

AI is both a boon and a bane for data centers. It drives unprecedented demand and offers transformative potential but also exacerbates sustainability challenges. However, AI itself holds the key to overcoming these hurdles, heralding a new era of efficient, powerful, and sustainable data centers. As data center operators navigate this complex landscape, the crucial strategy is to harness AI not only to fuel growth but also to spearhead innovation in energy efficiency and sustainability. Those who master this balance will not just survive but dominate in the future of digital infrastructure.

JLA empowers datacenter operators and sustainability solution providers to seize the immense market opportunities driven by AI. From identifying expansion opportunities to recommending new technologies for operational efficiency, JLA leverages its deep experience and trusted business acumen to craft cutting-edge strategies. For more information on how we can help further your business goals, reach out to us at info@jlaadvisors.io.

Author

Sangit Rawlley

Sr. Advisor & AI Practice Lead

JLA Advisors

John Trobough

Founding Partner

JLA Advisors

JLA Advisors is a boutique consulting firm providing our clients with a comprehensive end-to-end suite of services. We specialize in strategy development, technology

architecture design and execution, and software operational excellence, all with a strong emphasis on innovation.

References

1. Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds
2. Generative AI & The Future of Data Centers: Part I - The Models
3. AI Data Centers' Global Power Surge and the Sustainability Impact
4. Generative AI: The Next S-curve for the Semiconductor Industry?
5. AI and the Green Energy Transition Will Bring New Challenges and Opportunities
6. AI Data Centers' Global Power Surge and the Sustainability Impact
7. AI Data Centers' Global Power Surge and the Sustainability Impact
8. Global PUEs — Are They Going Anywhere?
9. AI Data Centers' Global Power Surge and the Sustainability Impact
10. Bring on regulations for data center sustainability, say Europe and APAC
11. DeepMind AI Reduces Google Data Centre Cooling Bill by 40%
12. AI Data Centers' Global Power Surge and the Sustainability Impact
13. More Than a Third of Enterprise Data Centers Expect to Deploy Liquid Cooling by 2026