The Data Crisis is Unfolding – Are We Ready?

ALPHAWAVE SEMI

We live in an era of unprecedented technology adoption. Industries, from technology to financial services to healthcare, are seeing technological advances at a blistering rate. The meteoric rise in interest in generative AI, the Internet-of-Things (IoT), autonomous vehicles, and a litany of other next-gen technologies are moving our society toward a brave new world of efficiency.

That being said, these advancements come at a considerable cost. Never before have we been so beholden to the widespread sharing and processing of data to navigate our everyday lives in ways we don't even realize. The navigational systems for mass transit and shipping logistics, medical diagnoses and treatments, and even the air conditioning we use in a rapidly warming world have become much more advanced but are all controlled by data that has to be moved and processed at speeds we've never seen before.

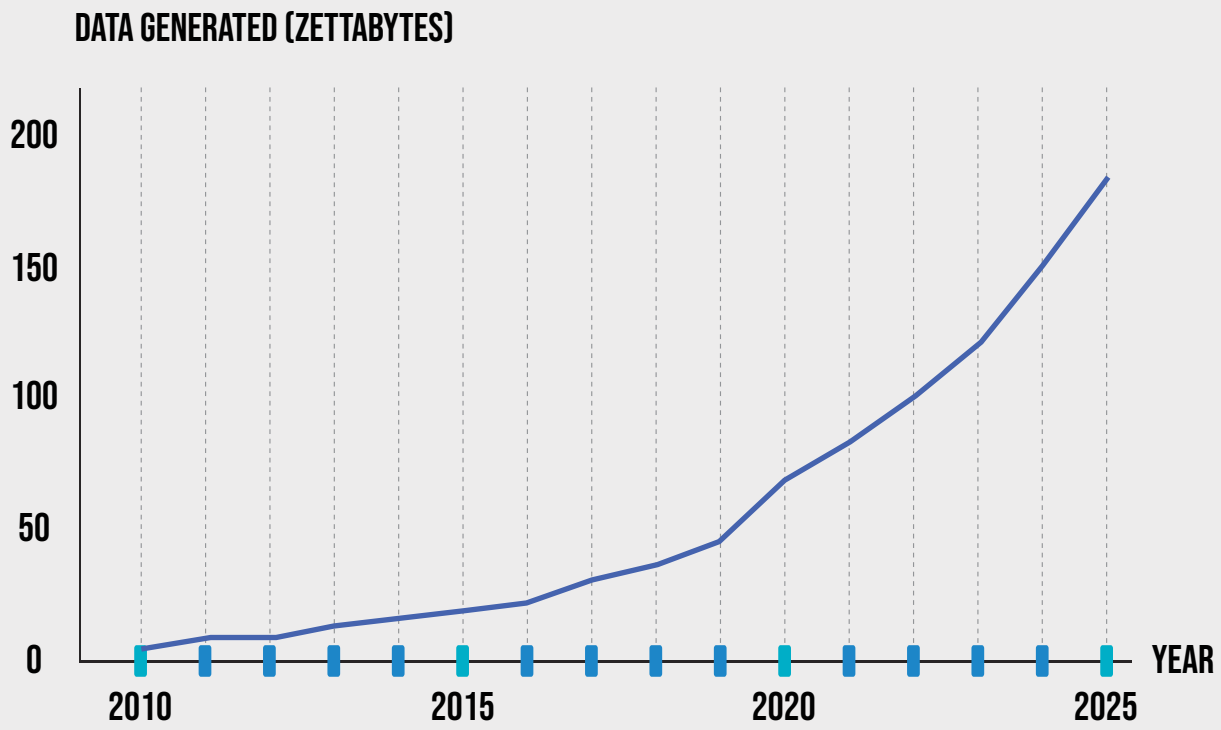
> Our reliance on advanced technology leaves us vulnerable to critical failures if our ability to process data is disrupted, even for a moment. Such disruptions can have severe consequences, threatening our ability to function as a society. We must be mindful of these risks and take steps to protect ourselves.
>
> —Tony Chan Carusone

The usage of global data is increasing every year at an exponential rate. However, any form of bottleneck can hinder the growth of data-dependent technologies, preventing them from taking root or reaching more people to benefit from in a global society. This poses a moral and humanistic dilemma that urgently needs to be addressed.

90X

It is estimated that as much as 120 zettabytes[1] of data will be created, captured, copied, and consumed in 2023 alone. This number is expected to grow to over 50% to 180 zettabytes of data in 2025. To put this in perspective, in 2010, only two zettabytes of data were generated and used globally. This means that we will witness a **90X increase in just 15 years.**

ZETTABYTES OF DATA/ANNUM FROM 2010-2025 (SOURCE: STATISTA)



[1] Source: https://connect2nonstop.com/monetizing-data-why-companies-are-looking-inside-to-drive-business-success/#:~:text=Data%20volumes%20are%20soaring%20globally,as%20181%20zettabytes%20by%202025.

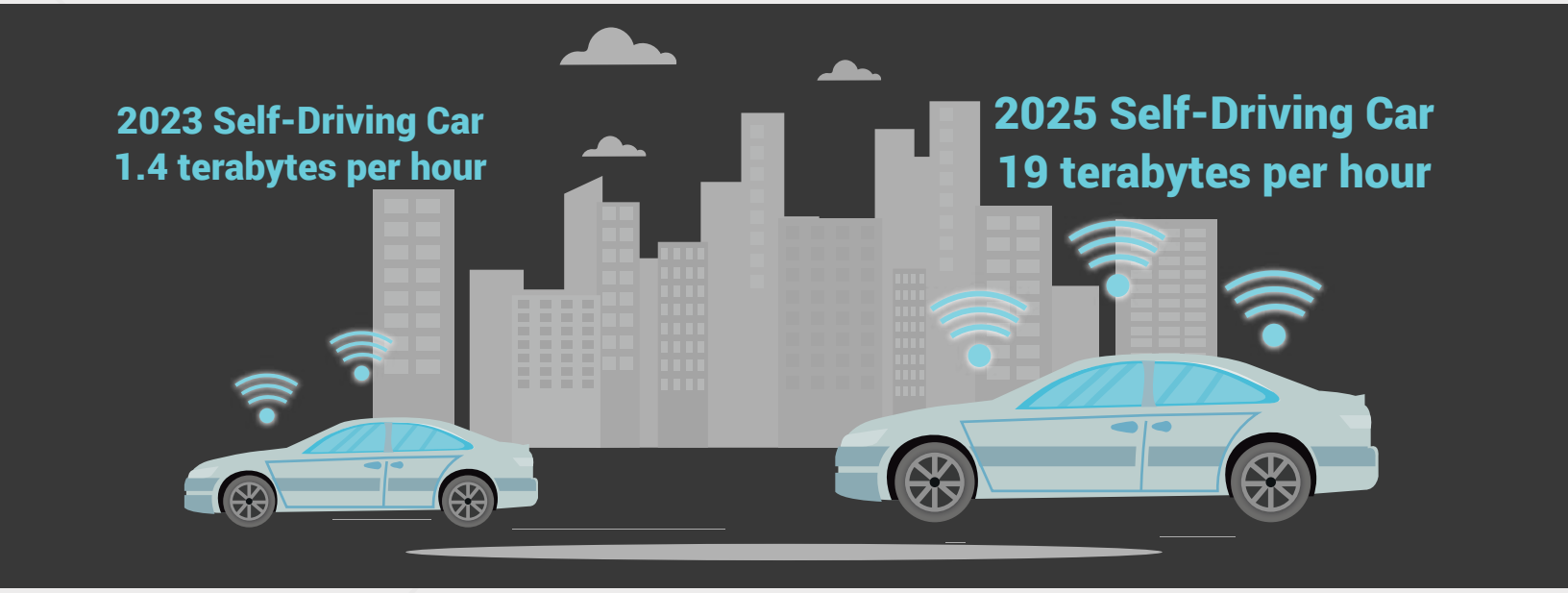
These numbers are absolutely staggering and hard to conceptualize. Take the 120 zettabytes from 2023 – where ~0.33 zettabytes are created every day – here are other ways to conceptualize how much data that really is:

- 120 Billion 1TB iPhone 15 Pro Maxes
- 240 Billion PlayStation 4 console storage chips
- More than 900 Billion 128G Chromebooks

With all this data moving around the world and being processed, a sobering reality comes quickly into focus if the global data infrastructure is not aggressively updated to meet the ever-growing demand for data, the results will be dire, if not catastrophic.

To illustrate this, here is an example using autonomous vehicles.

Right now, an autonomous vehicle, such as a Cruise or Waymo taxi, needs, on average, a minimum of **1.4 terabytes (TB) of data per hour**[2] to conduct itself. This number is expected to grow as autonomous vehicles become more advanced, to roughly **19 terabytes per hour** by 2025. When multiplied by the estimated 840,000 autonomous vehicles that will be hitting the streets in 2030[3] , autonomous vehicles alone will be moving more than **1.6 million terabytes of data an hour between the car and a data center**.



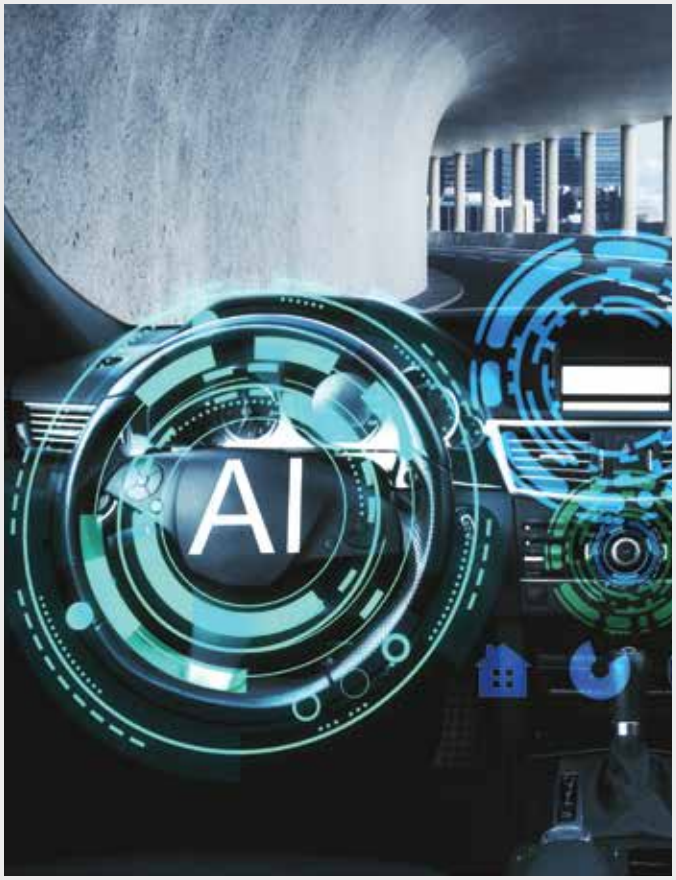
[2] Source: Siemens: https://blogs.sw.siemens.com/polarion/the-data-deluge-what-do-we-do-with-the-data-generated-by-avs/
[3] Source: Guidehouse Insights: https://guidehouseinsights.com/news-and-views/deployments-of-highly-automated-vehicles-are-unlikely-before-2025

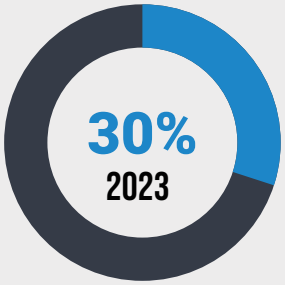
That data needs to be transferred and processed constantly while the vehicles are running to prevent breakdowns, improve resilience in inclement weather, and avoid life-threatening accidents. A malfunction in this instance could have fatal consequences, especially in the context of hundreds of thousands of autonomous vehicles on the road.

If the image of autonomous vehicles potentially careening off the road en masse doesn't conjure alarm, maybe we should imagine something more universally experienced and potentially dreaded: a trip to the doctor's office.

Today, the medical industry already depends on a astonishing amount of data. As of 2021, roughly 88% of US-based doctors[4] have adopted some kind of digital health record for their patients.

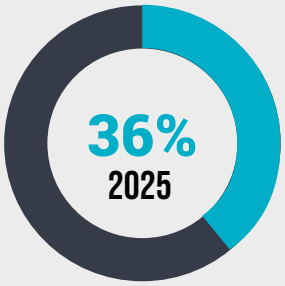
As they continue to become more advanced, AI applications can be deployed across all areas of health-care worldwide – from automating mundane administrative work to reading radiology scans to provide diagnoses to formulating care plans for patients with chronic illnesses. The opportunities for AI to make the healthcare sector more efficient are endless and are to be celebrated.





in 2023, approximately **30% of the global data sphere** is taken up by healthcare data.

HEALTH



With this number expected to increase to **36% in 2025**[5].

[4] Source: HealthIT.gov: https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption
[5] Source: RRC: https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion

While these data-intensive applications can help save time, money, and lives, our current reliance on data to treat patients poses a chilling realization: if any medical application were to experience a widespread inability to process the gargantuan amounts of data efficiently, it could lead to a mass misdiagnosis event and potentially cost lives. Suddenly, the technology meant to serve as a vital force for our collective welfare could be responsible, instead, for needless pain and suffering.

These are just two examples of how the promise of AI and other data-intensive applications can be life-changing, but only if we can secure the infrastructure responsible for transferring this data around the world efficiently, with low latency and low power. To understand the challenges we face in terms of these applications, such as generative AI, it is essential first to understand how they work.

A Brief History and Background of Generative AI

OpenAI's groundbreaking AI-powered chatbot ChatGPT was introduced to the public in November 2022. It's been heralded as the fastest-growing consumer technology in the world, having reached **100 million active monthly users**[6] just two months after its launch.

These users could be students trying to cram in research for a term paper or workers hoping to offload some of their monotonous professional tasks. When hundreds of thousands of queries are asked of generative AI daily, it equals a startling amount of data, fueling an endlessly growing need for power.

In April, a group of researchers from the University of Michigan published a paper discussing a new optimized training method for ML algorithms. They revealed that it took a surprising 1,287 megawatt hours to train the GPT-3 large language model (LLM). To put this into perspective, the amount of electricity used to train this model is enough to **power an average US home for 120 years**. It's worth noting that this energy consumption only accounts for the training process and does not include the energy consumption of the system when it is live, processing millions of daily queries.[7]

[6] Source: Reuters https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
[7] Source: Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training: https://www.usenix.org/system/files/nsdi23-you.pdf

GPT-4, the current version of ChatGPT, is on an entirely new level of computational power, requiring nearly 10 times more computational power to train the LLM due to its vast size and scope. We will elaborate on this in the subsequent discussion. However, the study conducted by the University of Michigan fails to recognize the numerous potential applications of generative AI in the future. Many enterprise companies, such as Disney, The Home Depot, Coca-Cola, and Siemens, have already announced their plans to incorporate AI in order to automate several aspects of their work.

According to Goldman Sachs, about 300 million jobs are at risk of being automated by AI in some capacity, which is a cause for concern for job seekers, particularly those in junior positions, as their roles are more vulnerable to automation, according to job marketplace Hired. As the use of AI continues to grow, we must consider the impact of its massive data transfer on global data infrastructure. Moreover, we must address the issue of increased power requirements for AI applications and data centers worldwide.

We see examples of data infrastructure collapsing under stress due to a data bottleneck, even today. A notable example is the online stock-trading app Robinhood, which almost crashed during the GameStop "meme stock" incident in 2021. There were so many trades happening simultaneously that the company had to halt trading on its platform to ensure that it could support them.

Before we delve into the power concerns of generative AI, it is essential to understand how it operates and its relationship with data centers.

How Next-Gen Technologies Like Generative AI Work and Their Relationship to the Data Center

As discussed, generative AI is hugely taxing on our data infrastructure. It cannot be emphasized enough that the current global data infrastructure is ready for the growth and expansion of generative AI applications. Several factors hinder the adoption of generative AI technologies, but we will focus on two main ones: interconnectivity and power consumption.

Interconnectivity is the spine of the data center. Current data center racks can only support so much computational power, just as our personal computers can only do so many tasks simultaneously. To overcome this, data centers are designed to share the workloads of a query or task over multiple racks to maximize efficiency, which is where interconnect comes in. This technology enables all of the data within data centers to move around freely and easily.

> *Interconnect is the spine of the data center.*
>
> *— Tony Pialis*

This is especially critical in generative AI applications, as the bandwidth required to reference data stored in memory to generate the answer to a query is too much for just one rack to take, but the data has to come back at the same time for the application to answer the query quickly.

To illustrate this, let's imagine that a generative AI query is a puzzle, and each data center rack is a piece of that puzzle. When you ask ChatGPT a question, you've just dumped the puzzle pieces from the box onto the table. As you're putting it together, the separate racks work to generate the answer to your query, but the puzzle isn't a puzzle until all the pieces are put together, and the piece edges are like the interconnect. We need to have the best puzzle piece edges to ensure that the puzzle can be put together quickly, or else we'll have a poorly fitting puzzle.

Current interconnect speeds, or the puzzle, don't allow for the data to be moved around and referenced at the same time quickly enough without causing lag and using unsustainable amounts of power. This brings us to our next primary concern: the power consumption of AI and the data center.

We've alluded to the amount of power AI requires to run, and looking at data released in October 2023, we're only starting to fathom just how much power it will need.

But as more applications come online, what will become increasingly more apparent is the power consumption that inference, or how people interact with AI applications, uses. Let's take Google, for instance, which has started introducing its chatbot Bard into Google search, eventually revolutionizing how we think of search. As of February 2023[8], using an AI-powered search is estimated to cost 10X that of a standard Google search using keywords. If all Google searches were powered by generative AI, models predict that Google's AI function will use the same amount of power as the entire country of Ireland.

BARD + CHATGPT POWER USAGE = IRELAND

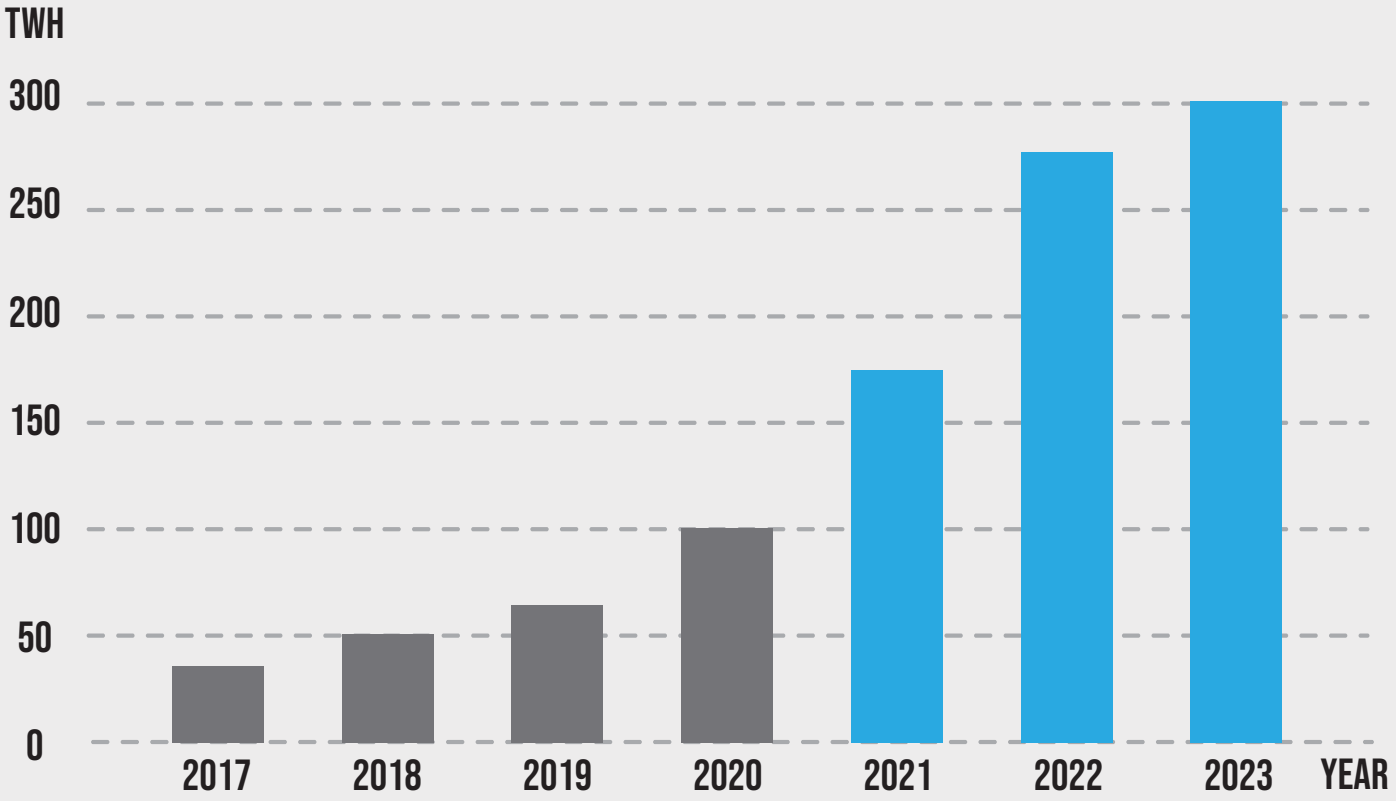
When looking at data centers, the International Energy Agency[9] estimates that in 2022, data centers were responsible for 1-1.3% of the global final electricity demand (excluding cryptocurrency mining power needs). Even though companies and countries are implementing energy efficiency programs, data centers continue to see 20-40% annual growth in energy consumption due to the large data centers that are being put online regularly.

[8] Source: https://www.cell.com/joule/pdf/S2542-4351(23)00365-3.pdf
[9] Source: IEA: https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

The power used by Amazon, Microsoft, Google, and Meta, some of the largest owners and operators of data centers, has more than doubled from 2017 to 2021, using roughly 72 TWh combined in 2021. This equates to enough electricity to power roughly **6.72 million US households** in 2021.

amazon Microsoft Google Meta

POWER USAGE BY MAJOR TECH COMPANIES BETWEEN 2017 AND 2023



Connectivity – Our Way Forward to Solving the Data Crisis

As outlined above, connectivity, or interconnect, is crucial in ensuring that our data infrastructure can rise to the occasion to support generative AI applications and our everyday data needs and to solve the data crisis we've outlined here. AI isn't practical for everything, and its "killer use app" right now is increasing productivity, which means that our data infrastructure will have to remain stable for "normal" workloads such as streaming videos or video calls while simultaneously accommodating the data-intensive workloads that AI necessitates.

Hyperscalers are already acting on this need for faster connectivity to support their generative AI applications. In August 2023, Amazon[10] announced it was moving to custom chip designs for its generative AI applications, citing the need for faster and more accessible chips to continue scaling the business. But to secure the entire data infrastructure, more companies need to invest in faster, more efficient chips that offer orders of magnitude of gains in compute capabilities and, equally importantly, interconnect speeds to ensure the success of AI.

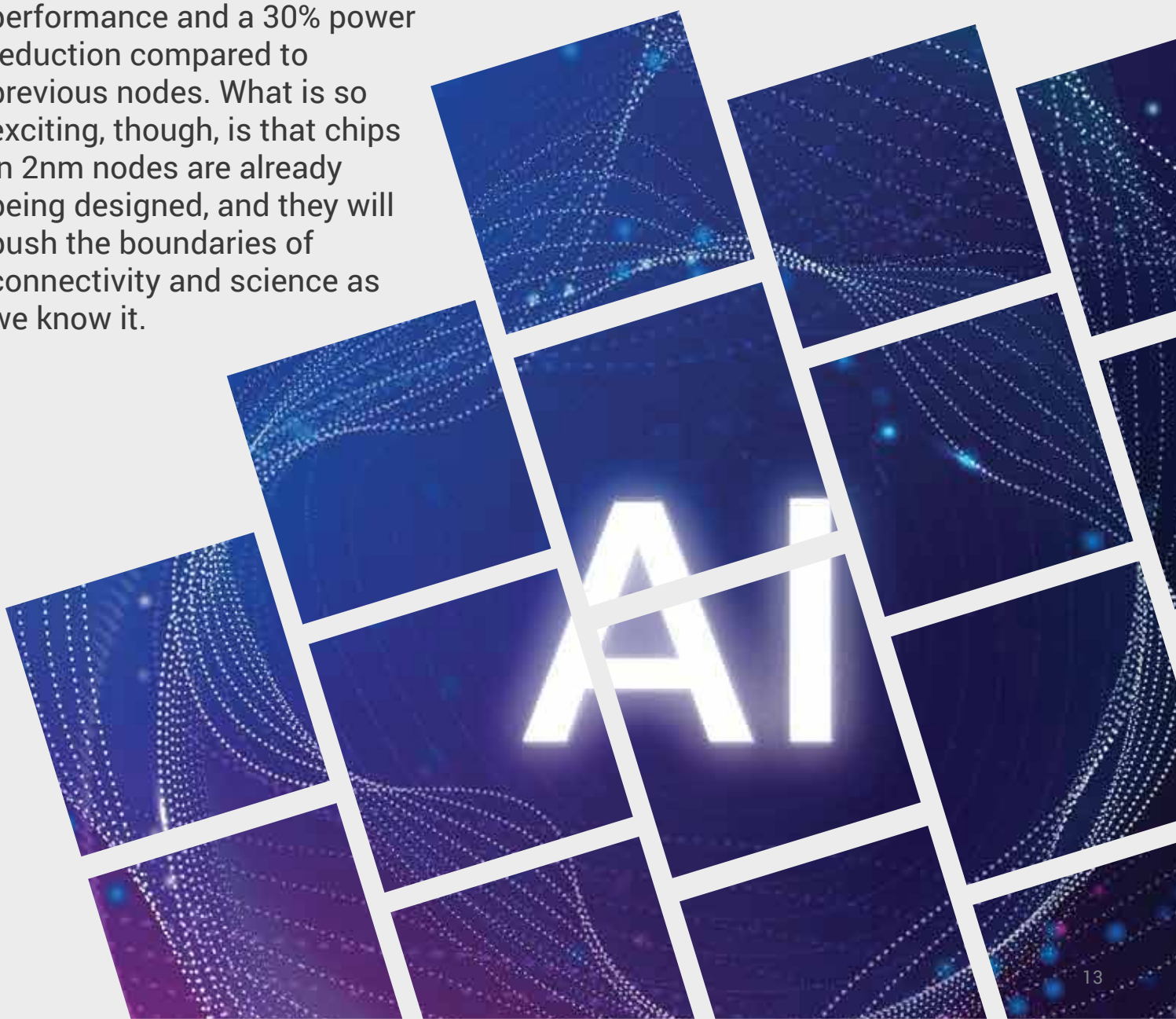
The crisis we face is unlike anything we've seen before, thus, legacy technologies that use monolithic chip structures are not the solution. For the data infrastructure and data centers to survive the mounting computational pressure that we are going to see, utilizing chiplets and custom silicon solutions will be the only way that they can survive to achieve scale, performance and time-to-market.

> *" We are at a critical juncture in the development of data infrastructure, and legacy technologies are no longer sufficient to support the rising demands of artificial intelligence. Chiplets and custom silicon solutions will be how we can achieve the scale, performance, and time-to-market speed required to solve our data crisis. "*
>
> *— Tony Pialis*

[10] Source: CNBC https://www.cnbc.com/2023/08/12/amazon-is-racing-to-catch-up-in-generative-ai-with-custom-aws-chips.html

Chiplets and custom silicon are not new in the semiconductor industry by any means, but they are seeing a surge in demand and importance as data centers and hyperscalers continue to process more data than ever before. Companies like Alphawave Semi and other industry leaders are at the bleeding edge of utilizing these technologies not only to maximize efficiency within the data center but also to lead to reductions in power consumption and latency.

As we move from generic connectivity solutions to the next generation of custom solutions, it is imperative that these solutions continue to adapt to the newest iteration of hardware technologies. Currently, 3nm process nodes are coming online, already offering a 15% improvement in performance and a 30% power reduction compared to previous nodes. What is so exciting, though, is that chips in 2nm nodes are already being designed, and they will push the boundaries of connectivity and science as we know it.



Working with companies to offer flexible, scalable, low-power, and high-performance solutions is a critical success factor in the fight against the mounting data crisis and upgrading our data infrastructure to keep up with generative AI and technologies we have yet to imagine.

The need for constant innovation in connectivity and interoperability may seem at odds with the importance of the widespread adoption of standards and new technologies as industry players move at different speeds – a race of unending progress may appear unnecessary or counterproductive.

But the reality is that connectivity and interoperability are required to stay in front of the rapidly growing need for data and maintain our current hyper-dependence on this data.

The future of technology and progress hangs in the balance. We simply cannot afford to miss this opportunity to work towards securing the future of our data infrastructure before it's too late. Years, if not decades, of progress will be wiped out by our inaction.

ALPHAWAVE SEMI

Alphawave Semi™ is a global leader in high-speed connectivity for the world's technology infrastructure. Our IP and custom silicon solutions meet the needs of worldwide tier-one customers in data centers, compute, networking, AI, 5G, autonomous vehicles, and storage. To find out more about Alphawave Semi™, visit: www.awavesemi.com