Principal Real Estate

Principal Asset Management℠

AI is transforming the data center

Authors:

SEBASTIAN DOOLEY, CFA
Senior Fund Manager

CASEY MILLER
Managing Director,
Portfolio Manager

PAUL LEWIS
Director,
European Data Centres

BEN WOBSCHALL
Managing Director,
Portfolio Manager

Artificial intelligence (AI), formerly relegated to the pages of scientific journals and science fiction narratives, is now permeating the fabric of our everyday lives—and transforming the data center in the process. Because AI workloads are quite different from traditional workloads, supporting them demands a new approach to data center design and operation. As AI applications progress—for example, from training models to decision-making—its impact on the data center industry will continue to evolve.

Adding to the pressure on data center operators, demand for data centers to support AI is additive to demand for data centers to support traditional workloads. Data center operators will have to support these very different demand profiles—sometimes even within the same facility. This paper explores how.

AT-A-GLANCE

Artificial intelligence has been around since the 1950s but has only recently begun to infuse daily life. Driving this evolution is exponential growth in data, advancements in computing architecture and chip performance, and AI's increasingly broad applicability and accessibility.

The impacts of AI on the data center result from the differences between AI workloads and traditional workloads.

- Compared to traditional workloads, AI workloads have distinct computational demands. They're considerably more power intensive and their power consumption is more highly variable. That drives new power infrastructure and management requirements.
- AI deployments are typically higher density, which drives the need for new approaches to cooling.
- The importance of latency varies depending on the AI application, driving unique facility and size decisions.

AI has been around since the 50s. So what's new?

For decades, artificial intelligence has been—at least according to popular media—poised to take over the world. And yet it's only recently that AI has begun to infuse daily life. (It's still not taking over the world. But it is affecting how data centers are designed and operated.)

As far back as the 1950s, early AI pioneers laid the groundwork for AI as we know it with the development of rudimentary neural networks, inspired by the human brain. In 1997, AI burst into the public eye when IBM's Deep Blue defeated chess grandmaster Garry Kasparov, showcasing the potential of AI in complex decision-making and strategy. In 2011, IBM's Watson demonstrated the power of AI in understanding and generating natural language when it beat Jeopardy's reigning champion Ken Jennings. This evolution has accelerated in recent years. Today, with generative AI we're witnessing the rise of AI that can not only analyze and process information but also create original content.

The increasingly rapid evolution of AI is driven by:

- **The data deluge**—The recent exponential growth in data generation, primarily driven by big data analytics and the Internet of Things (IoT), has provided AI algorithms with the extensive datasets necessary for effective learning and development.
- **Advancements in computing architecture and chip performance**—Processors like GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units) are specifically designed for parallel processing, significantly boosting performance compared to traditional CPUs.
- **AI's increasingly broad applicability and accessibility**—Powerful AI capabilities are becoming accessible to developers of all levels through user-friendly interfaces and cloud-based services. As a result, AI is more accessible and applicable to individuals and enterprises across various sectors.

EXHIBIT 1:
The accelerating evolution of AI

- **1951** Marvin Minsky and Dean Edmonds develop the first artificial neural network
- **1966** Stanford Research Institute develops Shakey, the world's first mobile intelligent robot
- **1981** MIT grad student Danny Hillis designs massively parallel computer consisting of a million processors, each similar to a modern GPU
- **1997** IBM's Deep Blue defeats Garry Kasparov in a historic chess rematch
- **2006** IBM Watson originates with the initial goal of beating a human on the quiz show Jeopardy!
- **2011** Apple releases Siri, a voice-powered personal assistant that can generate responses and take actions in response to voice requests
- **2014** Facebook develops the deep learning facial recognition system DeepFace, which identifies human faces in digital images with near-human accuracy
- **2016** DeepMind's AlphaGo defeats top Go player Lee Sedol
- **2018** OpenAI releases its first Generative Pre-trained Transformer (GPT), paving the way for subsequent large language models (LLMs)
- **2019** Microsoft launches the Turing Natural Language Generation generative language model with 17 billion parameters
- **2020** Open AI releases the GPT-3 LLM consisting of 175 billion parameters to generate humanlike text models
- **2021** OpenAI introduces the Dall-E multimodal AI system that can generate images from text prompts
- **2022** OpenAI releases ChatGPT to provide a chat-based interface to its GPT-3.5 LLM
- **2023** Google releases its AI chatbot, Bard

Source: TechTarget, The history of artificial intelligence: Complete AI timeline, 16 August 2023.

AI workloads are different than traditional workloads

The impact of AI on the data center is driven by the differences between AI workloads and traditional workloads. Compared to traditional workloads, AI workloads have distinct computational demands. They are considerably more power intensive and their power consumption is more highly variable. AI deployments are typically higher density. And for certain types of AI workloads, low latency (between where the data is produced and consumed and where it is processed) may be a secondary consideration compared to capacity or scale.

AI workloads have distinct computational demands

AI involves computationally intense tasks such as machine learning, deep learning, and neural network training. Unlike traditional workloads that may rely more heavily on CPU-based processing, AI workloads demand high levels of processing power, often necessitating the use of specialized hardware like GPUs and TPUs. AI workloads leverage these advanced processors to handle massive datasets and perform complex calculations at high speeds.

Another distinguishing factor of AI workloads is their data storage and access requirements. AI applications often involve large volumes of data, requiring high-capacity storage solutions with fast data retrieval capabilities. This necessitates the deployment of advanced storage technologies, such as all-flash arrays, and high-speed interconnects to facilitate rapid data access and processing.

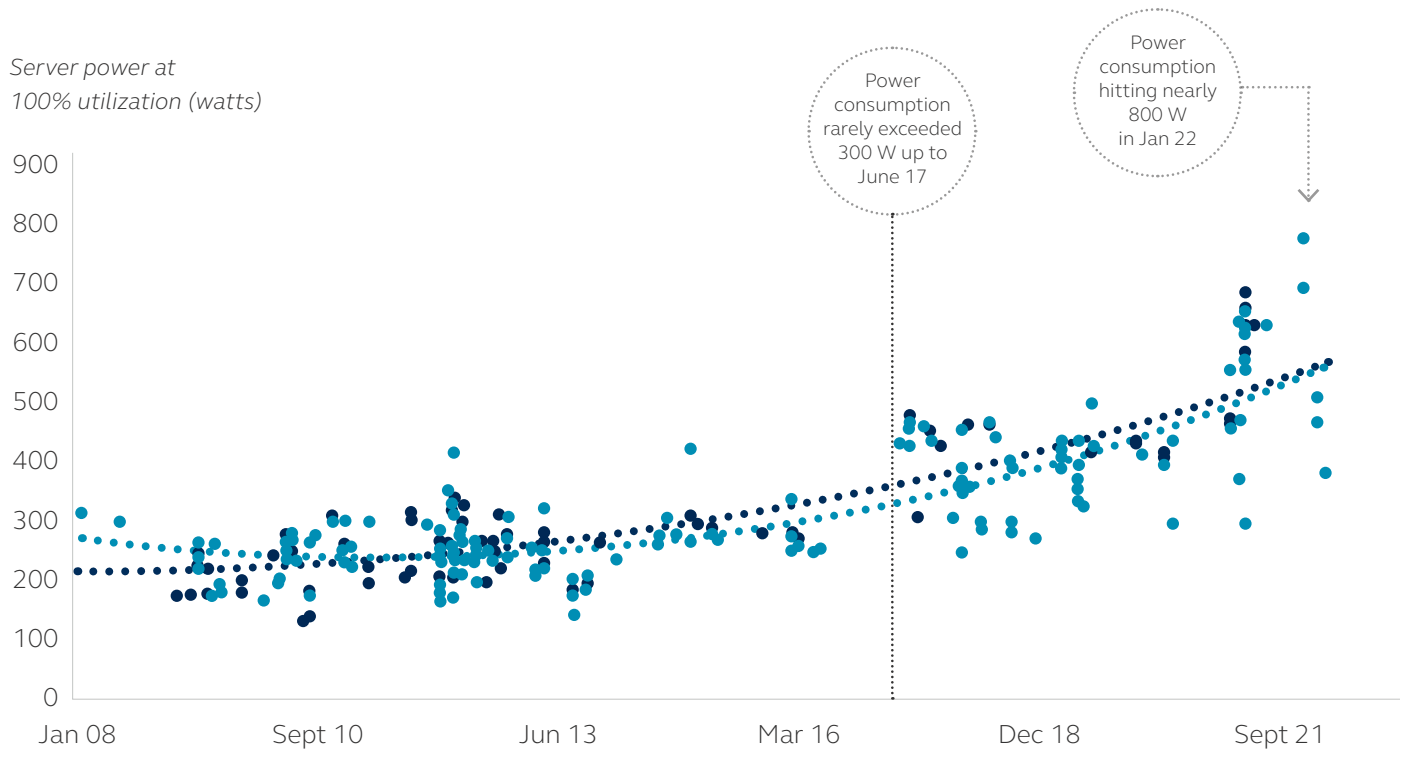
AI workloads are considerably more power intensive than traditional workloads

AI tasks involve processing large volumes of data and running complex models, which necessitates continuous operation of high-performance computing systems at or near full capacity for extended periods—leading to substantial power consumption. Generative AI is particularly power intensive as large language models like those underpinning OpenAI's ChatGPT can be up to 100 times larger than other AI models.

Specialized processors such as GPUs and TPUs are designed to handle parallel processing at a large scale, which is essential for the rapid processing of vast datasets typical in AI tasks. But high power usage is not just a function of the sheer volume of computations; it also stems from the nature of AI algorithms. Many AI processes involve iterative and complex calculations that require these processors to work at peak performance repeatedly over the course of the workload.

Using GPUs and TPUs for computationally intense workloads consumes more power compared to standard CPUs used in traditional workloads, and these processors themselves are increasingly powerful. For example, the maximum power consumption of NVIDIA's latest GPU is 160% higher than that of the company's previous generation chips.

EXHIBIT 2: Rising server power consumption



The data shows the sustained maximum power consumption of two-socket servers when running the SPECpower_ssj2008, which simulates a Java-based business logic. Results for 1U and 2U form factors. Data as at June 27, 2022.

Source: Uptime Institute, Too hot to handle? Operators to struggle with new chips, 15 March 2023

AI workloads' power consumption is more variable than traditional workloads

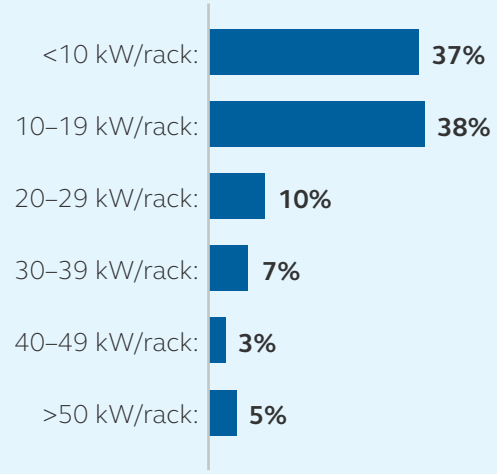
Power requirements are highly variable because AI workloads can vary significantly in their computational intensity over time. For instance, a deep learning model may require intense computational power during its training phase but much less during inference or testing phases. This leads to fluctuating power demands, with sharp increases during periods of high computational activity. For example, an AI model may run in excess of 100% of design utilization during training and only 30% of design utilization during inference or testing.

AI deployments are higher density than traditional workloads

AI workloads often require higher density than traditional data center workloads due to the nature of the computational and storage resources they demand. Specialized hardware is physically dense and consumes more power per unit area than traditional CPUs. The physical arrangement of GPUs in a pod, designed to act as a single computer, increases density. Furthermore, processing vast amounts of data necessitates a substantial amount of high-speed storage and memory, which contributes to the overall density.

The integration of this high-performance computing equipment in a relatively compact space leads to increased rack densities in data centers. While average server deployments in existing data centers remain approximately 10 kW per rack, recent deployments catering to AI are reaching five or even ten times that level. Over the next three to five years as AI continues to advance and become more prevalent and the next generation of technology infrastructure is deployed, high density will be the norm.

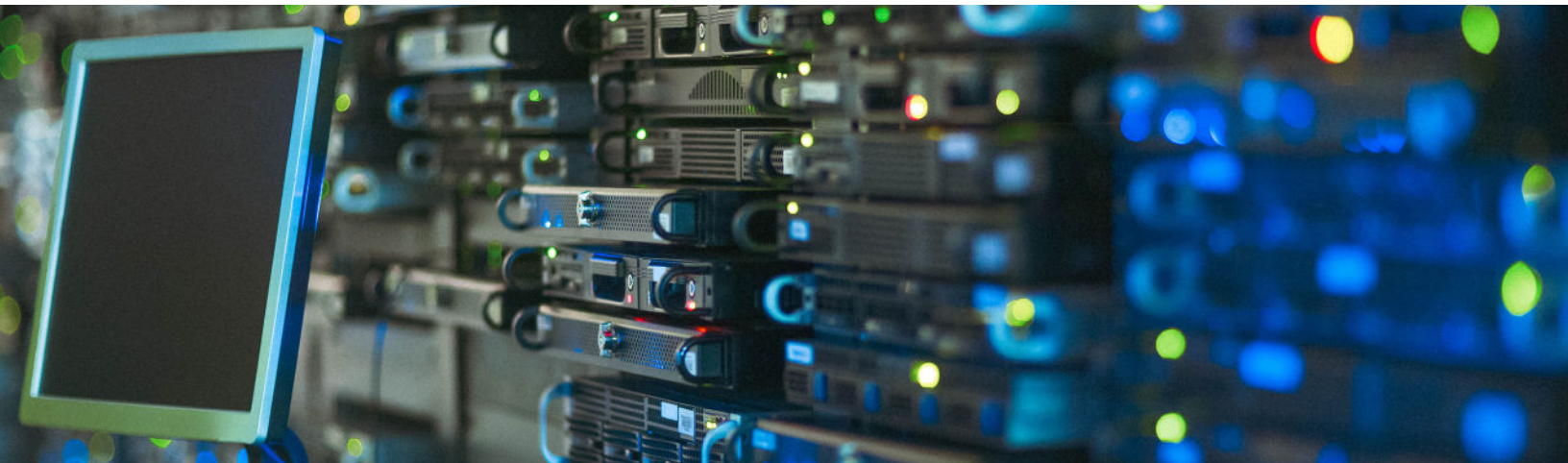
EXHIBIT 3: Rising rack densities



Source: Uptime Institute, Global Data Center Survey 2022, 14 September 2022

For certain types of AI workloads, low latency is critical

In an AI model's training phase, low latency connectivity to end users is typically not as important as other data center deployments. Training involves processing large datasets where the focus is on throughput and accuracy rather than immediate response times. When most AI workloads are training, data center location is not dependent on where the data is generated or consumed. But once an AI model begins inference—real-time data processing and decision-making—low latency becomes critical. Then, the AI deployment will likely have to be located in a data center close to where the data is developed or the model's outputs are utilized by end-users.



Supporting AI workloads demands a new approach to data center design and operation

The differences between AI workloads and traditional workloads drive new requirements for the data center. The fact that AI workloads are generally more power intensive but with wide variability drives new power infrastructure and management requirements. The fact that AI workloads are typically deployed in higher density configurations drives the need for new approaches to cooling. And the fact that some AI workloads are latency sensitive drives facility location and size decisions.

Supporting AI workloads demands new power infrastructure and approaches to power management

The scale of data centers is growing, with many data centers being built today over 50 megawatts. Data center campuses can be in the 100s of megawatts, with some over a gigawatt.

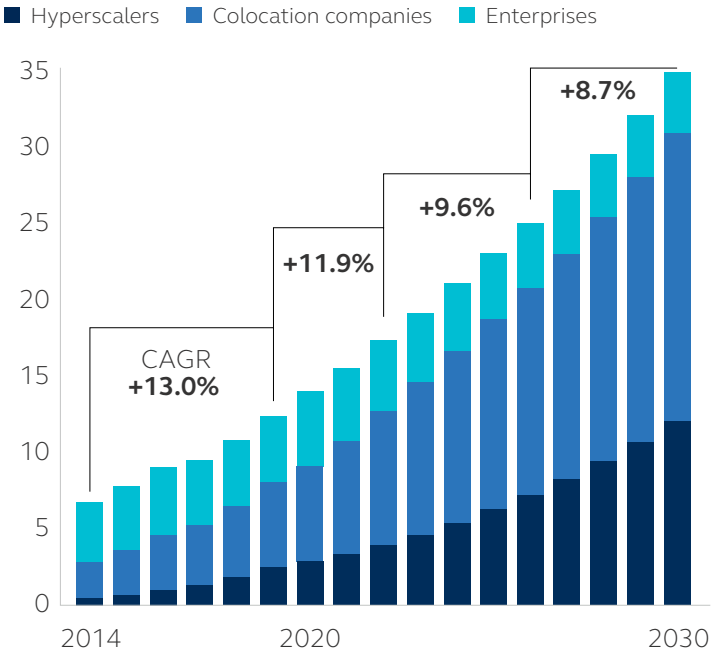
These massive power requirements are making new data center capacity increasingly hard to come by in many major markets, where utility power is increasingly constrained. As CBRE reports, "Sourcing enough power is a top priority of data center operators. Certain secondary markets with robust power supplies stand to attract more data center operators."[1]

Where there is sufficient utility grid capacity, an existing data center can be retrofit to increase power capacity. This involves a systematic upgrade of the power infrastructure—adding transformers, expanding electrical backup systems (uninterruptible power supplies and standby generators), and installing additional switchgear and power distribution units (PDUs).

Retrofitting a data center to enhance its power capacity in the absence of sufficient grid capacity presents a more complex challenge. The data center operator might consider on-site power generation or the development of a microgrid, which can operate independently or in conjunction with the main power grid. Ultimately, the operator may have to work with the local utility to enhance the grid's capacity through distribution infrastructure upgrades or the construction of new power generation facilities.

EXHIBIT 4: Rising data center power consumption

Data center power consumption by providers/enterprises[2], gigawatts



Source: McKinsey & Company, Investing in the rising data center economy, 17 January 2023

[1] CBRE, Global Data Center Trends 2023, 14 July 2023.
[2] 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.

In addition to demanding more power than traditional workloads, AI workloads are also more highly variable. Accommodating power usage spikes may require new power infrastructure as well as new approaches to data center power management.

To effectively support AI workloads, data centers must be capable of handling more variable loads with higher peaks. Accomplishing that might involve upgrading transformers, PDUs, and wiring to handle the increased power draw without compromising safety or efficiency. Intelligent power distribution systems can monitor and dynamically adjust the power supply to different racks or zones based on real-time demand, ensuring that AI workloads receive sufficient power during peak processing times.

Incorporating energy storage solutions, like battery banks, can help in smoothing out the power spikes. During periods of lower demand, these batteries can store energy, which can then be utilized during periods of high demand, thereby reducing the strain on the primary power infrastructure.

Beyond new power infrastructure, new approaches to data center power management are essential to supporting highly variable AI workloads. Power monitoring and management software can predict power usage patterns, monitor real-time consumption, and provide alerts for potential issues. Such predictive analysis helps in optimizing power usage and planning for future capacity needs.

Supporting AI workloads demands new cooling infrastructure and strategies

Because AI workloads are typically deployed in higher density configurations, the cooling requirements for AI workloads are substantially higher than for traditional workloads—driving the need for new cooling infrastructure and strategies.

Traditional forced air cooling, which operates by circulating cool air and expelling warm air, can encounter several challenges at higher densities. The sheer volume of heat generated in a high-density rack can overwhelm the air cooling capacity. Air has a lower specific heat capacity compared to liquid coolant, meaning air can carry less heat per unit volume. This limitation necessitates larger volumes of air to be circulated at a higher velocity.

Moving the heat exchange closer to the servers can increase cooling efficiency and provide supplemental cooling to localized loads, augmenting a more traditional “flooded room” cooling strategy. A more effective solution for high-density racks, liquid cooling systems work by circulating a coolant directly through heat exchangers or cold plates positioned close to heat-generating components. This approach allows for more direct and efficient heat transfer.

Retrofitting a data center equipped with forced air cooling to accommodate liquid cooling at the rack level is feasible. For facilities that already have a chilled water loop, the transition from traditional forced air cooling to liquid cooling is easier, as the existing chilled water loop can be extended directly to the racks. This extension typically involves the installation of specialized piping and heat exchangers within or adjacent to the racks to facilitate the direct transfer of heat from the high-density computing equipment to the water.

One of Principal’s assets—a data center commissioned in 2009—provides a great example of the feasibility of retrofitting an existing facility to support more power (and the requisite additional cooling) with relatively modest infrastructure additions. A data hall originally designed to support 2.7 MW of capacity is currently being retrofit to support 12.5 MW, in the same footprint and with the original mechanical equipment.

Supporting AI workloads demands different facility location and size decisions

The fact that some AI workloads are latency sensitive drives facility location and size decisions. In some cases, that will mean smaller data centers in edge locations.

The impact of latency on AI workloads varies depending on the specific application and its requirements. Data center providers catering to AI applications must therefore optimize their infrastructure accordingly, whether it's through the use of edge computing to reduce latency for real-time applications or optimizing network and storage systems for high-throughput, latency-tolerant tasks.

Because AI model training, for example, may not require low latency, such workloads can be located in very large data centers where power is plentiful and inexpensive. But because inference requires low latency, once an AI model is trained it may need to be moved to a data center closer to where the data is produced and consumed. In many cases these data centers will be at the edge—whether that's the city where autonomous vehicles are operating or the availability zone where the enterprise keeps its data. They may also be smaller than most of today's deployments (e.g. 3 MW instead of 50 MW).

There's an added benefit to moving some workloads to smaller data centers. The bigger the data center is, the more expensive it is. The concentration of investment into a campus with hundreds of megawatts or even a gigawatt of capacity can make capitalization more difficult.

AI doesn't make existing data centers obsolete

Many AI deployments will be in facilities dedicated to AI. Deploying AI in a data center outfitted for traditional cloud or enterprise use would typically likely require at least some retrofitting to support the unique network architecture and deployment densities of AI workloads—for example, re-racking the space to widen cold rows or install new network backbone, or the kind of power and cooling infrastructure upgrades mentioned previously.

When a data center has been designed to be future-proof, retrofitting it to support new AI workloads is feasible and may be the most cost-effective option. One of Principal's assets—originally a bank-owned data center commissioned nearly 15 years ago—provides a great example. To support the workloads of the time, the facility was designed for high redundancy and low density. But the data center had the ability to deliver significant amounts of power and cooling so retrofitting it to support modern workloads was feasible. In fact, the retrofit data center was leased to a specialty cloud service provider to deploy an AI/ML strategy that at full deployment will be one of the largest supercomputers in the world. (Learn more in our recent short paper, Data centers: Viable for the long term.)

While running multiple types of workloads in a single facility can make efficiency more difficult to achieve, a data center may be able to efficiently support both traditional workloads and AI workloads. For example, one data hall could have forced air cooling supporting traditional workloads at relatively low density while another data hall in the same building could have liquid immersion cooling to support an extremely high density deployment running AI workloads.

Bottom line

AI has ushered in a new era in data center design and operation. The unique demands of AI workloads, from increased power consumption and cooling requirements to the need for continual learning and updating, require a reimagining of traditional data center architectures. This evolution, while challenging, presents an opportunity for innovation and growth in the data center industry.

As AI continues to advance, it will be imperative for businesses and technology leaders to stay abreast of these changes and adapt their strategies accordingly. The future of data centers lies in their ability to effectively support the dynamic, power-intensive, and ever-evolving workloads.

Risk Considerations

Investing involves risk, including possible loss of Principal. Past Performance does not guarantee future return. All financial investments involve an element of risk. Therefore, the value of the investment and the income from it will vary and the initial investment amount cannot be guaranteed.

Important information

This material covers general information only and does not take account of any investor's investment objectives or financial situation and should not be construed as specific investment advice, a recommendation, or be relied on in any way as a guarantee, promise, forecast or prediction of future events regarding an investment or the markets in general. The opinions and predictions expressed are subject to change without prior notice. The information presented has been derived from sources believed to be accurate; however, we do not independently verify or guarantee its accuracy or validity. Any reference to a specific investment or security does not constitute a recommendation to buy, sell, or hold such investment or security, nor an indication that the investment manager or its affiliates has recommended a specific security for any client account.

Subject to any contrary provisions of applicable law, the investment manager and its affiliates, and their officers, directors, employees, agents, disclaim any express or implied warranty of reliability or accuracy and any responsibility arising in any way (including by reason of negligence) for errors or omissions in the information or data provided. All figures shown in this document are in U.S. dollars unless otherwise noted.

This material may contain 'forward looking' information that is not purely historical in nature. Such information may include, among other things, projections and forecasts. There is no guarantee that any forecasts made will come to pass. Reliance upon information in this material is at the sole discretion of the reader.

This material is not intended for distribution to or use by any person or entity in any jurisdiction or country where such distribution or use would be contrary to local law or regulation.

This document is issued in:

- The United States by Principal Global Investors, LLC, which is regulated by the U.S. Securities and Exchange Commission.
- Europe by Principal Global Investors (Ireland) Limited, 70 Sir John Rogerson's Quay, Dublin 2, D02 R296, Ireland. Principal Global Investors (Ireland) Limited is regulated by the Central Bank of Ireland. Clients that do not directly contract with Principal Global Investors (Europe) Limited ("PGIE") or Principal Global Investors (Ireland) Limited ("PGII") will not benefit from the protections offered by the rules and regulations of the Financial Conduct Authority or the Central Bank of Ireland, including those enacted under MiFID II. Further, where clients do contract with PGIE or PGII, PGIE or PGII may delegate management authority to affiliates that are not authorised and regulated within Europe and in any such case, the client may not benefit from all protections offered by the rules and regulations of the Financial Conduct Authority, or the Central Bank of Ireland. In Europe, this document is directed exclusively at Professional Clients and Eligible Counterparties and should not be relied upon by Retail Clients (all as defined by the MiFID).
- United Kingdom by Principal Global Investors (Europe) Limited, Level 1, 1 Wood Street, London, EC2V 7 JB, registered in England, No. 03819986, which is authorized and regulated by the Financial Conduct Authority ("FCA").
- This document is marketing material and is issued in Switzerland by Principal Global Investors (Switzerland) GmbH.
- United Arab Emirates by Principal Global Investors LLC, a branch registered in the Dubai International Financial Centre and authorized by the Dubai Financial Services Authority as a representative office and is delivered on an individual basis to the recipient and should not be passed on or otherwise distributed by the recipient to any other person or organisation.
- Singapore by Principal Global Investors (Singapore) Limited (ACRA Reg. No. 199603735H), which is regulated by the Monetary Authority of Singapore and is directed exclusively at institutional investors as defined by the Securities and Futures Act 2001. This advertisement or publication has not been reviewed by the Monetary Authority of Singapore.
- Australia by Principal Global Investors (Australia) Limited (ABN 45 102 488 068, AFS Licence No. 225385), which is regulated by the Australian Securities and Investments Commission and is only directed at wholesale clients as defined under Corporations Act 2001.
- Hong Kong SAR (China) by Principal Asset Management Company (Asia) Limited, which is regulated by the Securities and Futures Commission. This document has not been reviewed by the Securities and Futures Commission.
- Other APAC Countries/Jurisdictions, this material is issued for institutional investors only (or professional/sophisticated/qualified investors, as such term may apply in local jurisdictions) and is delivered on an individual basis to the recipient and should not be passed on, used by any person or entity in any jurisdiction or country where such distribution or use would be contrary to local law or regulation.

Principal Global Investors, LLC (PGI) is registered with the U.S. Commodity Futures Trading Commission (CFTC) as a commodity trading advisor (CTA), a commodity pool operator (CPO) and is a member of the National Futures Association (NFA). PGI advises qualified eligible persons (QEPs) under CFTC Regulation 4.7.

Principal Funds are distributed by Principal Funds Distributor, Inc.

MM13872 | 02/2024 | 3344148-122025