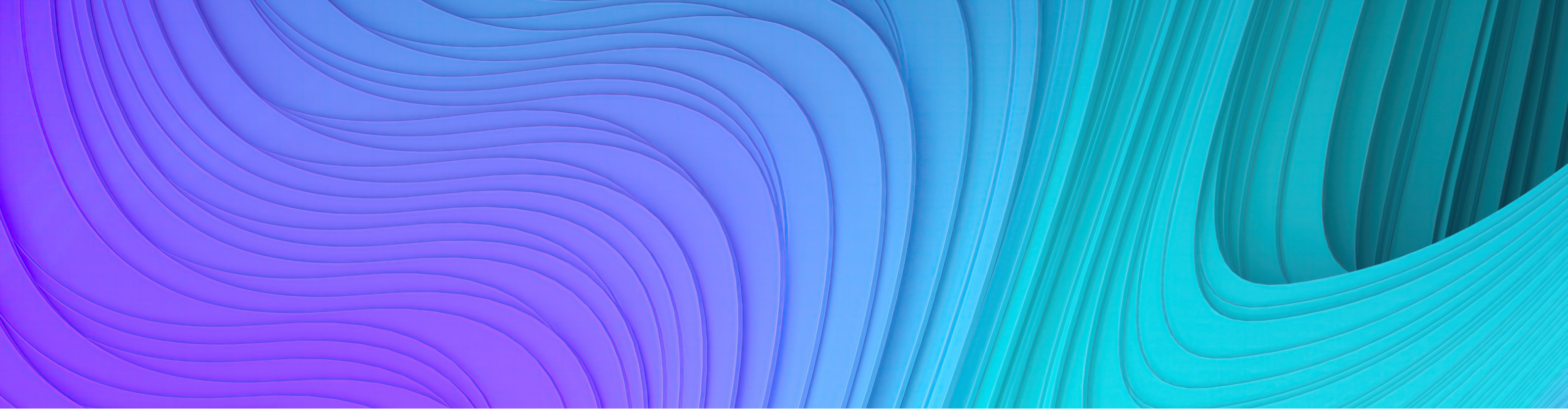
Technical brief

Scaling AI models — when supercomputing speed is a must

Unleashing the power of generative AI

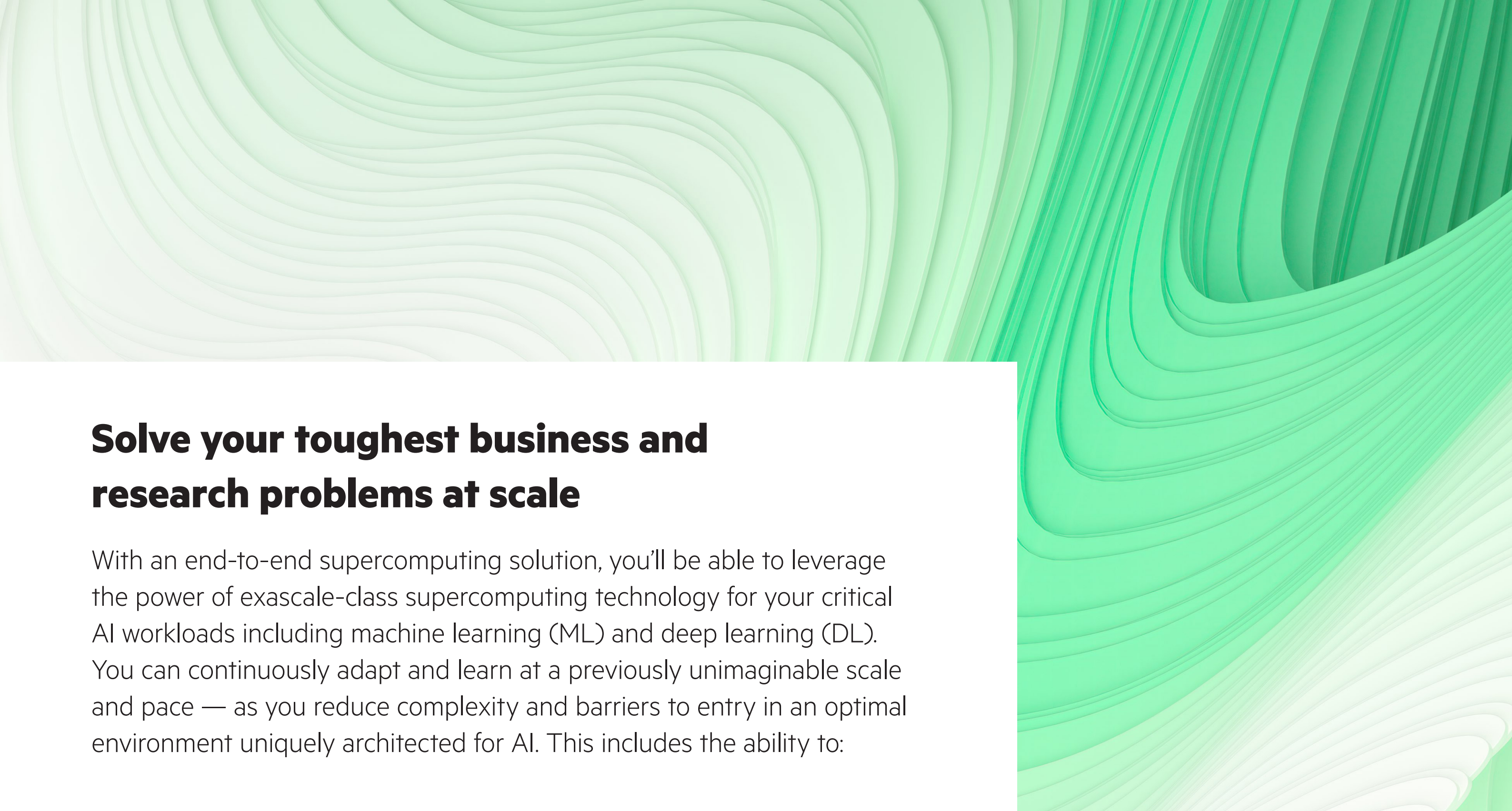
HPE GreenLake



Enterprises, research institutes, and government agencies around the world are training and tuning artificial intelligence (AI) models to drive innovation and unlock breakthroughs. The range of use cases is staggering. But whether you're using AI to solve previously intractable business problems, develop lifesaving medical advances, or build a national initiative to facilitate commerce, these projects have two things in common. The data sets are vast. And speed is an absolute priority.

What's more, for AI workloads such as large language models (LLMs) and deep learning recommendation model (DLRM) training, you need intense computational power that runs 24x7 for long periods of time.

All are good reasons why organizations are turning to supercomputing to accelerate AI model training and outcomes. Supercomputing can deliver the massive AI performance density you need for your biggest workloads and applications. New end-to-end supercomputing solutions that combine hardware, software, and services help you get started quickly and reduce operational friction, as you more efficiently scale, process vast amounts of data, and perform complex calculations at unparalleled speed.



Solve your toughest business and research problems at scale

With an end-to-end supercomputing solution, you'll be able to leverage the power of exascale-class supercomputing technology for your critical AI workloads including machine learning (ML) and deep learning (DL). You can continuously adapt and learn at a previously unimaginable scale and pace — as you reduce complexity and barriers to entry in an optimal environment uniquely architected for AI. This includes the ability to:

Access massive GPU performance density with hybrid-by-design, AI-native architecture

Gain the unprecedented scale and performance required for big AI workloads, such as LLM training and tuning and DLRM training.

Deploy an integrated software stack

You can build and maintain AI applications at scale, as you tap into software tools to build AI applications, customize prebuilt models, and develop and modify code. You can also simplify system management and firmware tasks.

Extend the range of data sets available for modeling

Look back over a longer time window to drive better insights and decisions. Quickly scale up as needed to ingest new data sets.

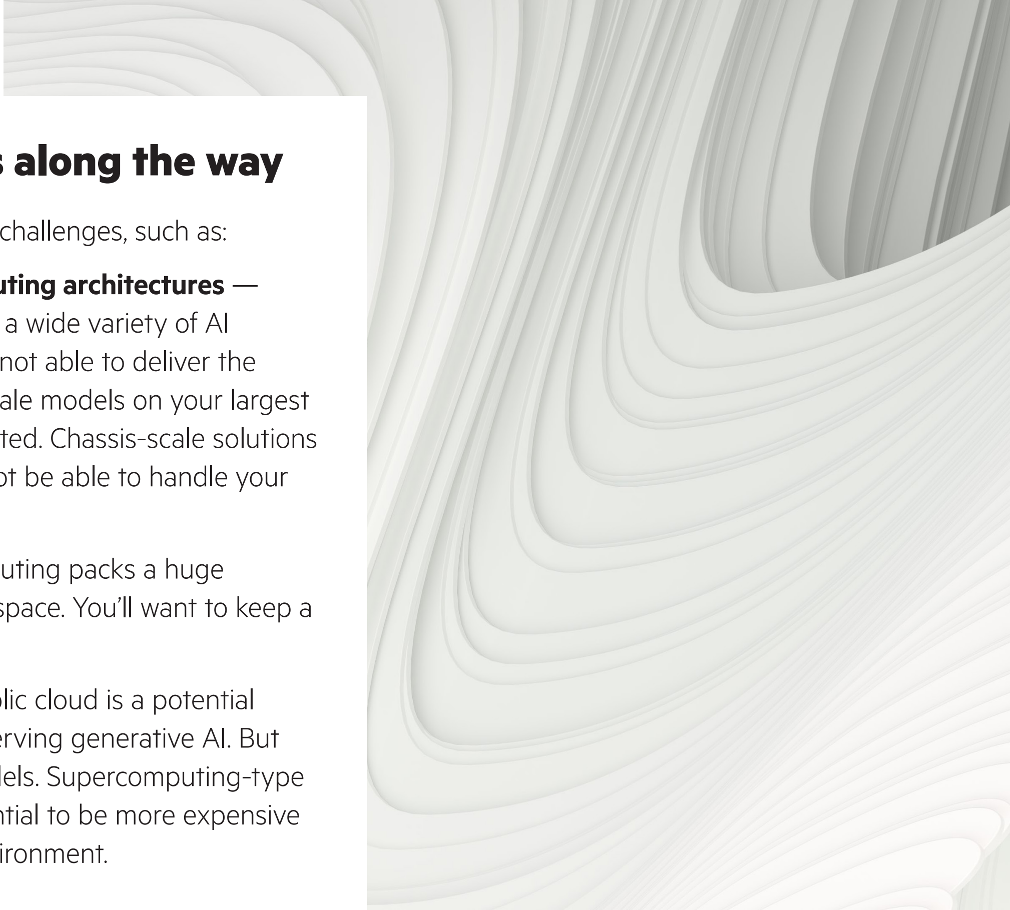
Leverage enterprise-grade lifecycle services

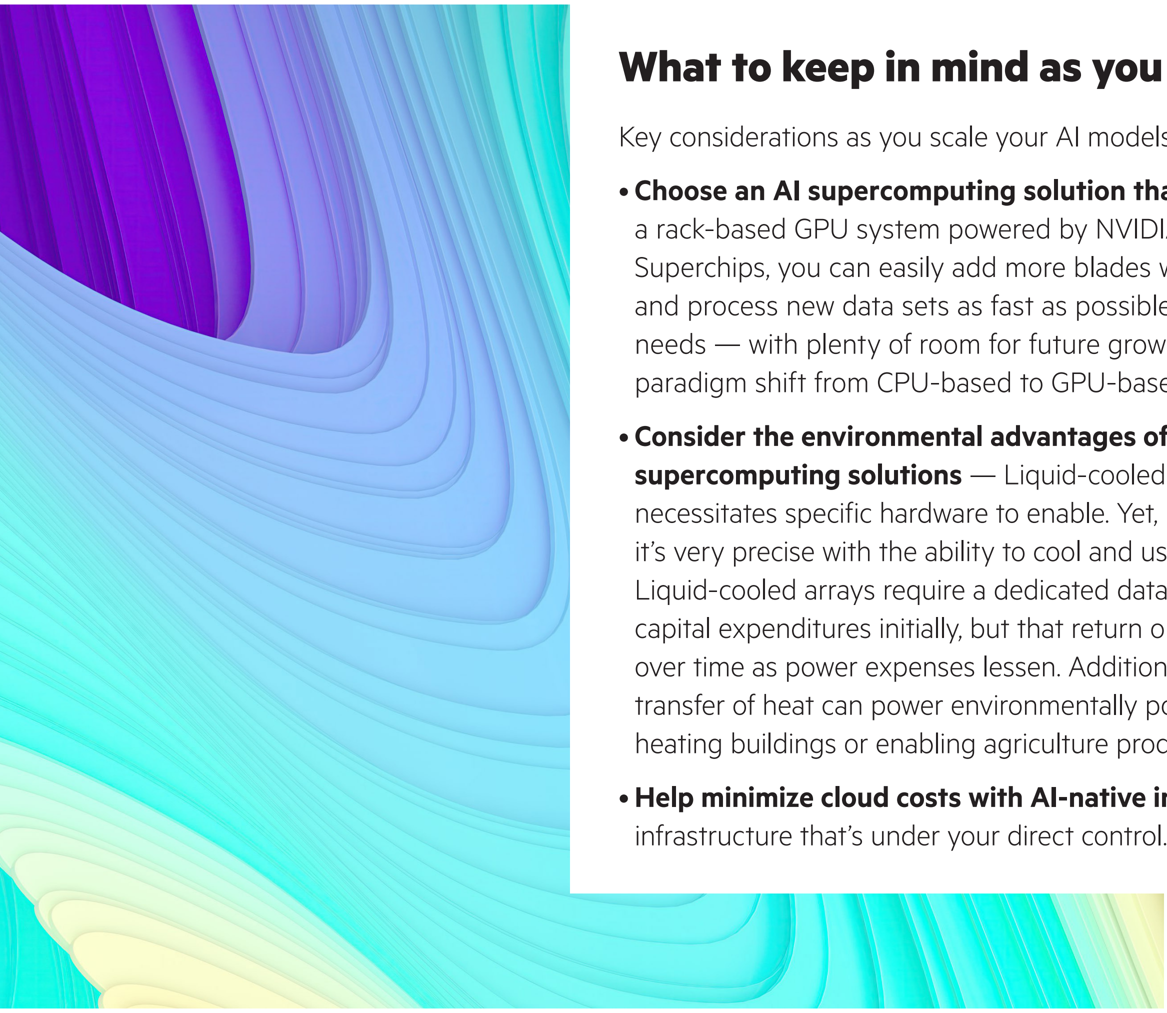
These services can help you bypass complexities and delays with selecting the right package of hardware and software. Additionally, when you benefit from a turnkey, AI-native supercomputing implementation experience, your data scientists can innovate from day one while you maintain the value of your investment with a worry-free operational support experience.

Recognize the challenges along the way

AI at extreme scale does present unique challenges, such as:

- **Hard-to-scale CPU-based supercomputing architectures** — CPU-based supercomputing can power a wide variety of AI workloads and applications, but they're not able to deliver the accuracy and speed that's needed to scale models on your largest data sets. Their ability to scale up is limited. Chassis-scale solutions may meet your needs today, but may not be able to handle your needs tomorrow.
- **Environmental impacts** — Supercomputing packs a huge amount of compute power into a small space. You'll want to keep a close eye on your energy footprint.
- **Unpredictable cloud costs** — The public cloud is a potential resource for training, fine-tuning, and serving generative AI. But the cloud is not well suited to all AI models. Supercomputing-type workloads that run 24x7 have the potential to be more expensive to run in the cloud than in a private environment.

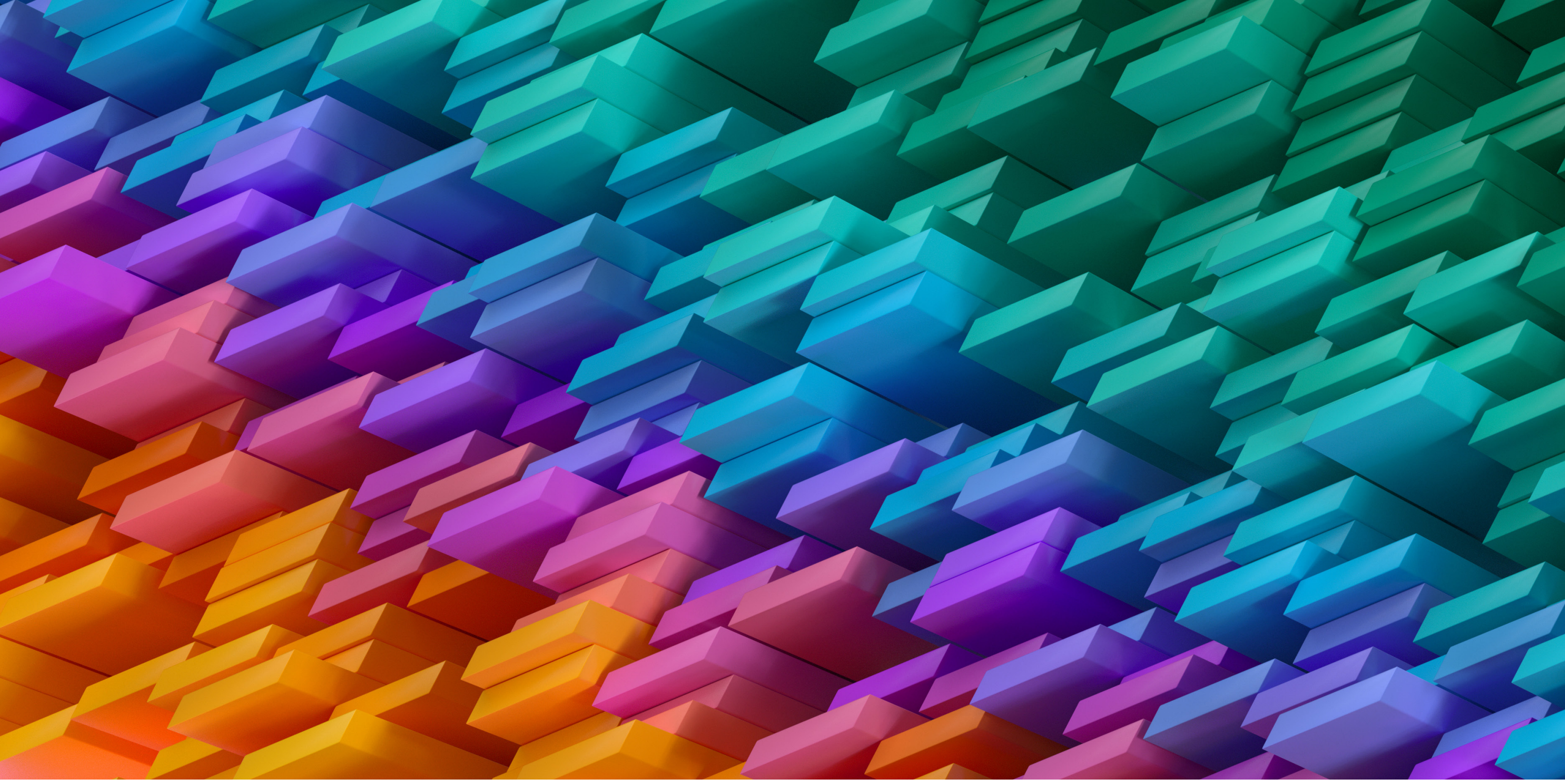




What to keep in mind as you move ahead

Key considerations as you scale your AI models:

- **Choose an AI supercomputing solution that grows with you** — With a rack-based GPU system powered by NVIDIA® GH200 Grace Hopper Superchips, you can easily add more blades when you need them and process new data sets as fast as possible. Meet your immediate needs — with plenty of room for future growth — and get ahead of the paradigm shift from CPU-based to GPU-based supercomputing.
- **Consider the environmental advantages of liquid-cooled supercomputing solutions** — Liquid-cooled supercomputing necessitates specific hardware to enable. Yet, unlike air-cooled systems, it's very precise with the ability to cool and uses less electricity overall. Liquid-cooled arrays require a dedicated data center, which increases capital expenditures initially, but that return on investment is captured over time as power expenses lessen. Additionally, the super-efficient transfer of heat can power environmentally positive use cases, such as heating buildings or enabling agriculture production in greenhouses.
- **Help minimize cloud costs with AI-native infrastructure** — This is infrastructure that's under your direct control.



Scale your AI models at supercomputing speed with HPE

Unleash the power of generative AI with HPE as you gain the scale and performance you need for computationally intensive generative AI workloads with a preconfigured, purpose-built turnkey solution. HPE's supercomputing solution for Generative AI integrates software tools to help you build AI applications, customize prebuilt models, and develop and modify code. The solution is sustainable by design, with direct liquid cooling that lowers energy consumption. Along with the ability to boost performance for the entire system through an open, Ethernet-based network designed to support exascale-class workloads.

Every organization is at a different stage of using AI to transform. From getting data AI-ready, to quickly getting value from your first AI initiatives — to training, tuning, and inference across the complete AI lifecycle at scale — HPE can help you create your AI advantage.

Select a partner who can help you today, and prepare you for tomorrow

HPE is the first[1] and only company in the world to have built a supercomputer that's broken the exascale barrier.[2] That's how and why we can help you scale your strategic AI models — and drive exceptional results for your customers, shareholders, or citizens at supercomputing speed. With our decades of supercomputing experience and knowledge, we can help you train, fine-tune, and deploy your models with full-stack customer support. This includes premium lifecycle services that remove complexity with access to over 1000 supercomputing specialists available to globally support you.

Learn more

Create your AI advantage with HPE

Unleash the power of generative AI

HPE's supercomputing solution for Generative AI

[1] "Frontier breaks exascale barrier, claims place as world's fastest supercomputer," HPE Blog, 2022

[2] top500.org/lists/top500/2023/11/

Explore **HPE GreenLake**

Chat now (sales)

Hewlett Packard Enterprise

a50010181ENW, Rev. 1