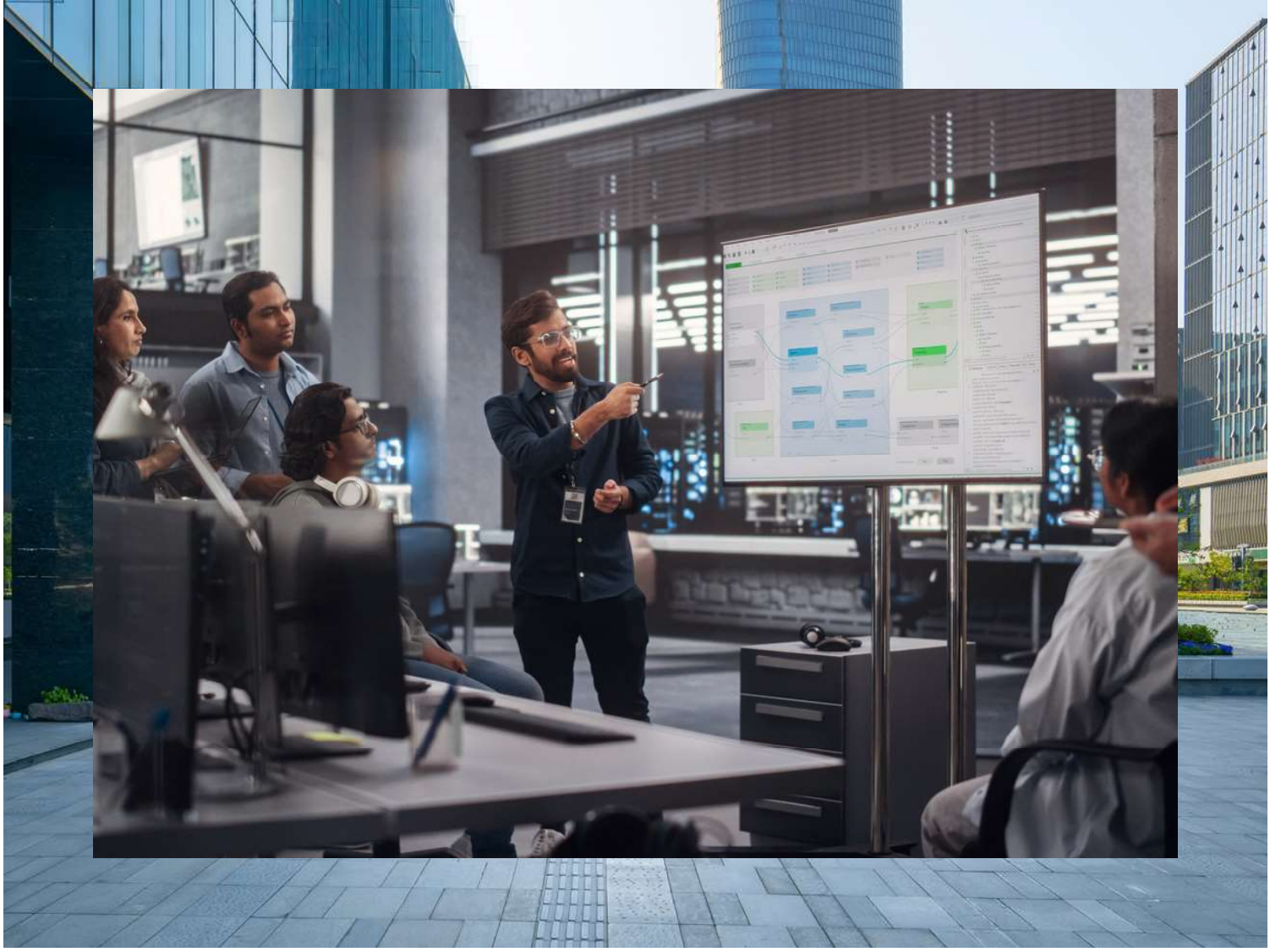


Technical guide

Fueling business transformation with generative AI

Hewlett Packard Enterprise | NVIDIA

A guide for the next era of AI

Artificial intelligence (AI) is profoundly changing how business is done, and while enterprises understand the transformative potential of AI, implementing the rapidly evolving technology is a major challenge. Generative AI (GenAI) and the adoption of large language models (LLMs) and multimodal models mark a significant advancement in AI. The ability of machine systems to generate controlled outputs that are natively human-readable dramatically simplifies their use and accelerates game-changing value for your business.

GenAI is considered the ultimate data-driven hybrid cloud workload. It creates unique and unprecedented demands on IT and business — and introduces new requirements that most enterprises have never encountered before. What worked yesterday won't get you to where you need to be tomorrow.

With GenAI, the pace of change is much faster. Foundational models constantly learn, meaning they must be trained and tuned regularly to remain accurate and up to date. In addition, deploying GenAI into a production is complex. It requires adjusting your operating model and training your team to use it effectively. Given all this, it's no surprise that 73% of enterprises are prioritizing AI overall other digital investments but 89% need help scaling models due to compute scarcity, limited IT expertise, and budget constraints.[1] These stats tell us that being successful is not easy and requires the right approach and partner.

"With the emergence of GenAI, enterprises are quickly realizing that the data and computational demands to effectively run AI models require a fundamentally different approach to technology," said Antonio Neri, president and CEO, Hewlett Packard Enterprise. Harnessing massive amounts of varied, distributed data sets requires a cutting-edge data-first pipeline architecture. That is why HPE is delivering an enterprise-class GenAI solution to help you prepare for the future of AI.

Our open, full-stack AI-native architecture[2] delivers AI solutions that accelerate AI model development and deployment. These solutions bring together HPE's leadership in hybrid cloud, supercomputing, and AI / machine learning (ML) with industry-leading NVIDIA® GPUs, NVIDIA AI Enterprise software, and NVIDIA networking platforms to make your business truly AI-powered. We deliver a curated mix of software and infrastructure specifically to accelerate the AI lifecycle and expand the capabilities of AI for future inferencing requirements. This shift is powering common GenAI use cases like conversational search, business process automation, and content creation.

This guide will help you harness GenAI capabilities that are fueling business transformation. By implementing these strategic principles, you will learn how to reduce the barriers of IT and discover AI-native and hybrid cloud solutions[3] that enable you to effectively train, tune, and deploy AI models from edge to cloud. Let's reinvent your business.

[1] "Among C-suite Leaders, AI is Top Digital Priority in the Path to Operational Resilience, Finds Accenture Study," Accenture, 2023

[2] "Introducing an AI-native architecture for AI-driven transformation," HPE Newsroom, 2023

[3] "Hewlett Packard Enterprise fuels business transformation with new AI-native architecture and hybrid cloud solutions," HPE Newsroom, 2023

Table of contents

Guiding principle 1

Partner with an industry leader

Business challenge

AI is shaping the future of business, but adopting and managing these technologies can be difficult. Enterprises need a partner with deep AI expertise to unlock the power of AI and evolve with new requirements.

GenAI workloads are computationally intensive and require enterprises to efficiently process massive amounts of data. The size of LLMs has grown from below a billion to over a trillion parameters in five years. Being able to handle large models is behind the technical breakthrough of next-generation AI.

Each stage of the AI lifecycle has its own unique demands on data and IT. Training and tuning AI models at scale demands innovative software and powerful supercomputing technologies to improve accuracy and speed results. To realize these benefits, enterprises are on a mission to balance sustainable solutions with exceptionally high performance. They must extend their cloud-native environment to include an AI-native approach if they hope to effectively utilize GenAI.

HPE is a trusted partner for AI adoption. We have proven technology and expertise, built from decades of experience as the global leader in supercomputing, which is essential for training the largest AI models. Our AI strategy is hybrid by design, recognizing that data can live anywhere and that inferencing increasingly must happen at the edge, where business transformation takes place and outcomes are delivered. We are making AI flexible, customizable, and sustainable from the start.

HPE and NVIDIA have introduced a new enterprise computing solution for GenAI. It is an AI tuning and inferencing data center solution that provides the ideal entry point for enterprises of all sizes and is ready-out-of-the-box to jumpstart the AI journey. This breakthrough offering enables you to adopt AI across tuning and inferencing — from our proven expertise in building systems designed for extreme scale and performance to implementing and integrating models into existing environments, so enterprises can use pretrained foundation models with their private data to create production applications (such as AI chatbots).

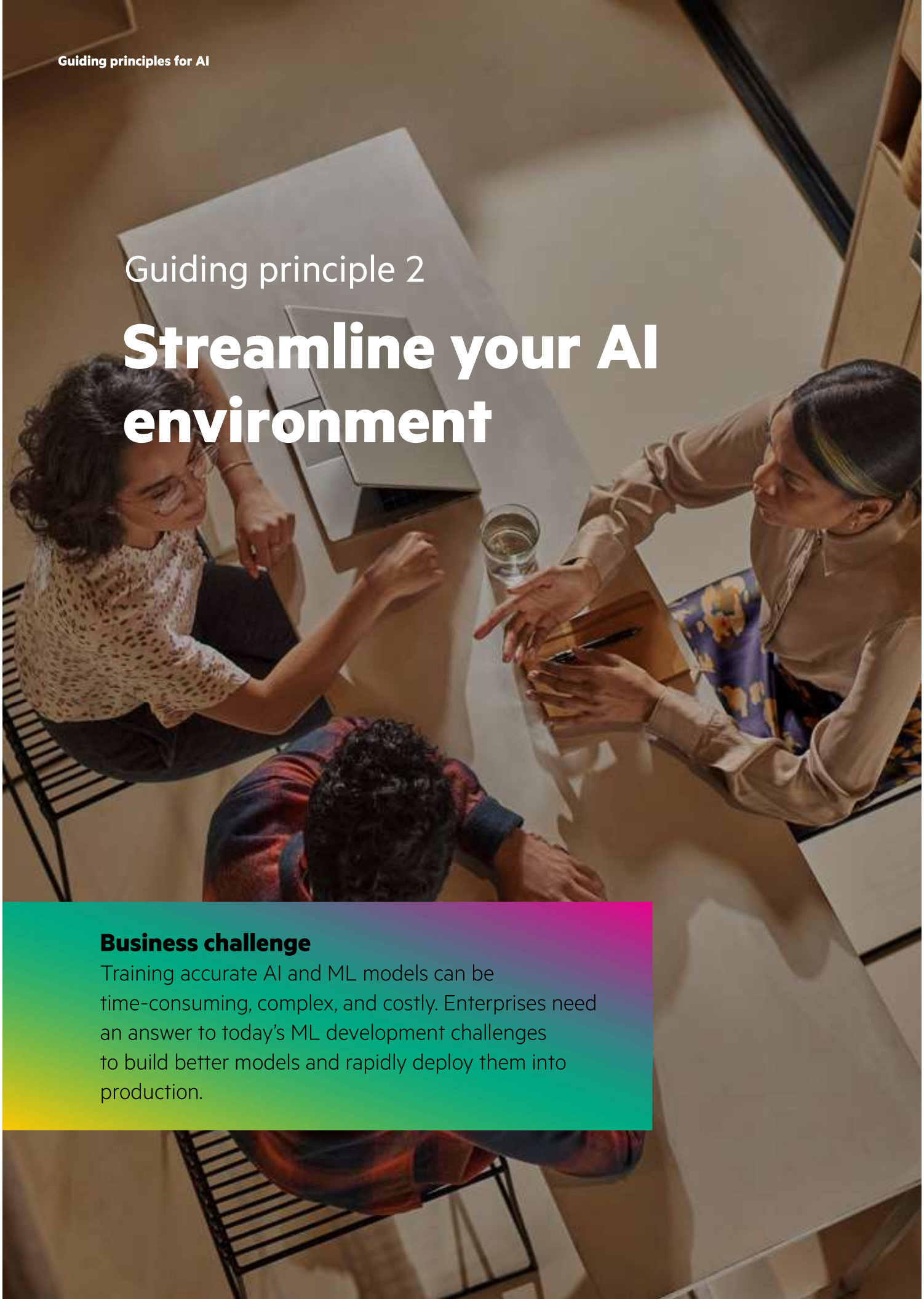
What you can expect from HPE solutions:

- Ready-out-of-the-box offering for enterprises of all sizes
- Pretrained foundation models that can be customized with private data to create production applications
- NVIDIA GPU-enabled applications for faster AI outcomes
- Tuned for edge or data center deployments
- Ultra-scalable architecture for greater inferencing performance
- End-to-end AI Services for GenAI from discovery to deployment and beyond

The HPE AI ecosystem is committed to helping you solve enterprise infrastructure challenges. Our agnostic open approach and trusted partnerships are constantly evolving to power your next transformation.

Expected results

- Simplify AI deployments with integrated hardware, software, and services purpose-built for AI
- Turn your data into intelligence for smarter decisions and faster outcomes
- Connect your data from edge to cloud with the highest level of security
- Future-proof your AI using a holistic approach that prioritizes sustainability from end to end



Guiding principle 2

Streamline your AI environment

Business challenge

Training accurate AI and ML models can be time-consuming, complex, and costly. Enterprises need an answer to today's ML development challenges to build better models and rapidly deploy them into production.

ML engineers and data scientists are on an endless search for solutions that enable them to prioritize innovation and speed up time to value. Their mission is to quickly customize models with enterprise data in a secure way to maximize business value while delivering high performance, flexibility, and ease of use. The most effective solutions include specialized tools and capabilities that enhance AI workflows and help users achieve faster results with accurate data.

If your organization wants to accomplish more with AI, HPE offers a future-focused solution that can help.

HPE Machine Learning Development Environment is designed to uncover hidden insights from your data by helping engineers and data scientists collaborate, build more accurate ML models, and train them faster. The platform integrates an array of software features into an easy-to-use, high-performance solution that allows teams to broaden their AI/ML scale and focus on building better models, rather than managing infrastructure. Combining state-of-the-art AI software with your existing IT makes it faster, simpler, and more cost-effective to develop and deploy AI models into production.

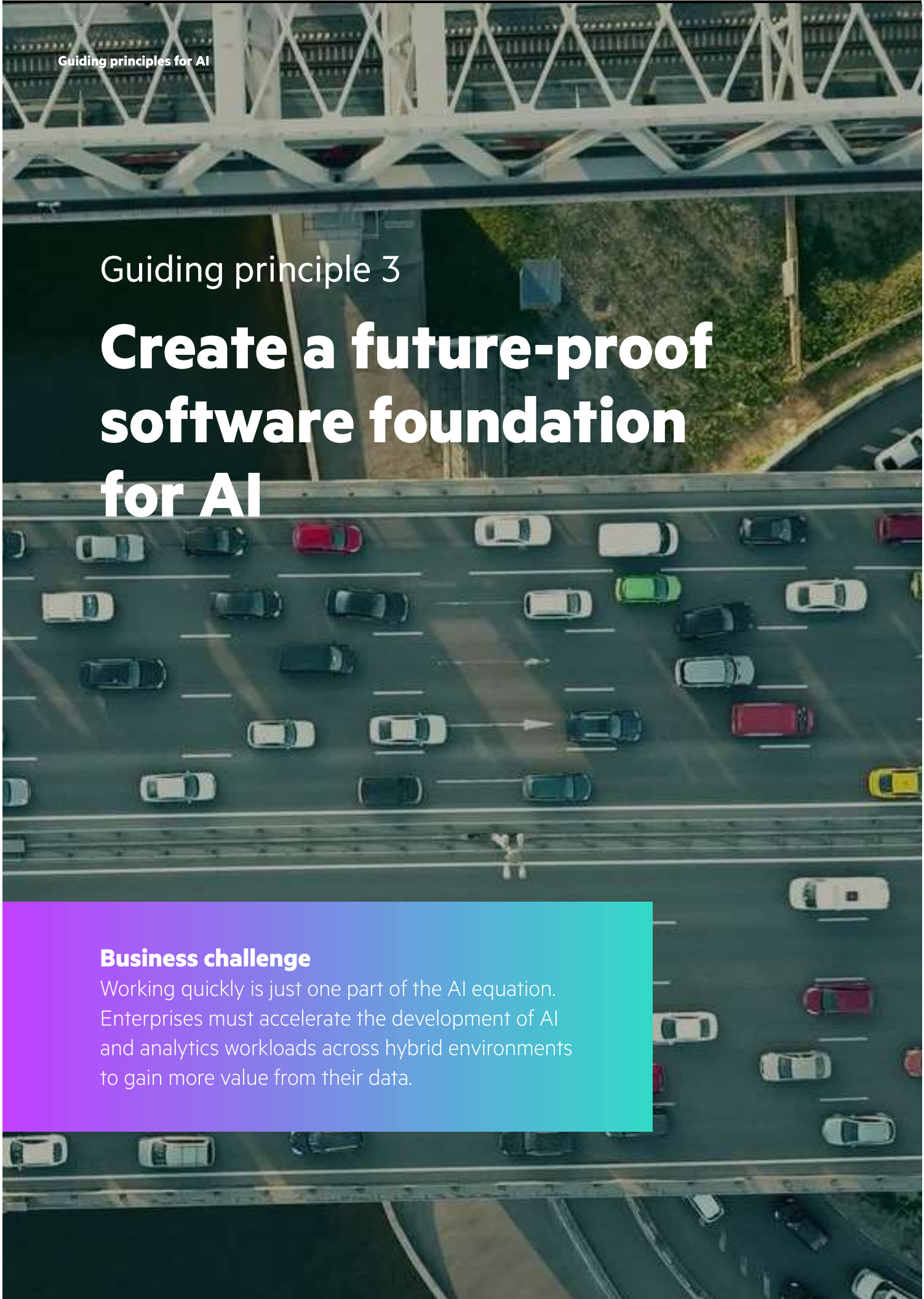
HPE Machine Learning Development Environment offers new GenAI studio features for rapidly prototyping and testing models. You can easily and securely integrate GenAI capabilities with a solution that is preconfigured, fully installed, and ready out-of-the-box. Seamlessly scale multiple nodes and hundreds of GPUs using priority scheduling for the most demanding AI workloads — without rewriting any model code. Teams can share AI infrastructure resources and custom models to boost individual productivity and collaboration. Starting a distributed training job is as simple as changing a configuration setting, so teams can perform distributed training and automatically find high-quality models with state-of-the-art hyperparameter tuning.

HPE Machine Learning Development Environment is now available as a managed service with a complete, adaptable cloud experience for AI/ML model training. This flexible managed service includes quick infrastructure setup delivered on an existing cloud account, which allows you to begin operating within a day. You can also reduce cloud costs with spot instances.

HPE offers the tools and expertise you need to fully optimize your environment. We provide infrastructure code for scaling training as well as support for provisioning machines, networking, data loading, fault tolerance, and much more.

Expected results

- Reduce the complexity and operational overhead of model training with seamless setup and management
- Increase AI adoption with future-proof infrastructure that relieves management, staffing, and processing burdens
- Accelerate time to value by making it easier for IT to set up, manage, secure, and share AI compute clusters
- Implement GenAI initiatives in days with flexible managed services that support the entire AI/ML journey



Guiding principle 3

Create a future-proof software foundation for AI

Business challenge

Working quickly is just one part of the AI equation. Enterprises must accelerate the development of AI and analytics workloads across hybrid environments to gain more value from their data.

Enterprises need a foundation for AI that can adapt and change with new requirements. HPE Software that is purpose-built for AI workloads is strategically engineered to unlock data and insights faster. Forward-looking enterprises are leveraging the latest software offerings — including software as a service (SaaS) — to develop and deploy data, analytics, and AI across hybrid environments.

HPE Ezmeral Software enables you to build and run AI applications on-premises and in the cloud. Integrated with HPE Machine Learning Development Environment, HPE Ezmeral creates a future-proof SaaS foundation for AI that can adapt and scale with the needs of your business. New enhancements are available to further simplify and accelerate model training and tuning with an end-to-end platform that seamlessly operates across hybrid multicloud environments.

HPE Ezmeral Unified Analytics Software is a comprehensive new offering for data engineering, data science, and data analytics delivered through SaaS. A consistent SaaS experience provides access to secure, enterprise-grade versions of the most popular open-source frameworks that help you develop and deploy data, analytics, and AI. While competitive offerings often require you to incur significant costs to move data to the cloud for analysis, HPE Ezmeral is hybrid by design and deployable across edge, colocation, on-premises, and public cloud. This deployment approach helps minimize data ingress and egress costs by enabling you to run analysis when and where it matters most.

The software suite also offers new GPU-aware capabilities to streamline deployment and accelerate data preparation for AI workloads, wherever the data lives. You can optimize NVIDIA GPU allocations across workloads and users to shorten time to value across the entire AI lifecycle with the power of a hybrid data lakehouse.

Other key advantages:

- **Help maximize productivity:** Seamless access to data and extensible open-source environments help eliminate bottlenecks, giving developers consistent self-service access to the evergreen tools they need to deliver across hybrid environments
- **Innovate everywhere:** Deploy AI in advanced analytics where your enterprise requires them to break down barriers and meet regulatory, compliance, and performance needs
- **Operate with confidence:** SaaS delivery provides predictable, transparent economics and precise controls to optimize costs and utilization across environments

HPE Ezmeral Unified Analytics Software enables you to build data pipelines as well as develop, deploy, and monitor models and visualize data with evergreen tools including Apache Airflow, Apache Spark, Apache Superset, Feast, Kubeflow, MLflow, Presto SQL and Ray. New connectors and data sources include Snowflake, MySQL, Delta Lake, Teradata, and Oracle®. With the open and extensible software, teams can easily add applications, locations, and sources as needs evolve.

Purpose-built software from HPE and NVIDIA are simplifying the path to operationalize and expand AI.

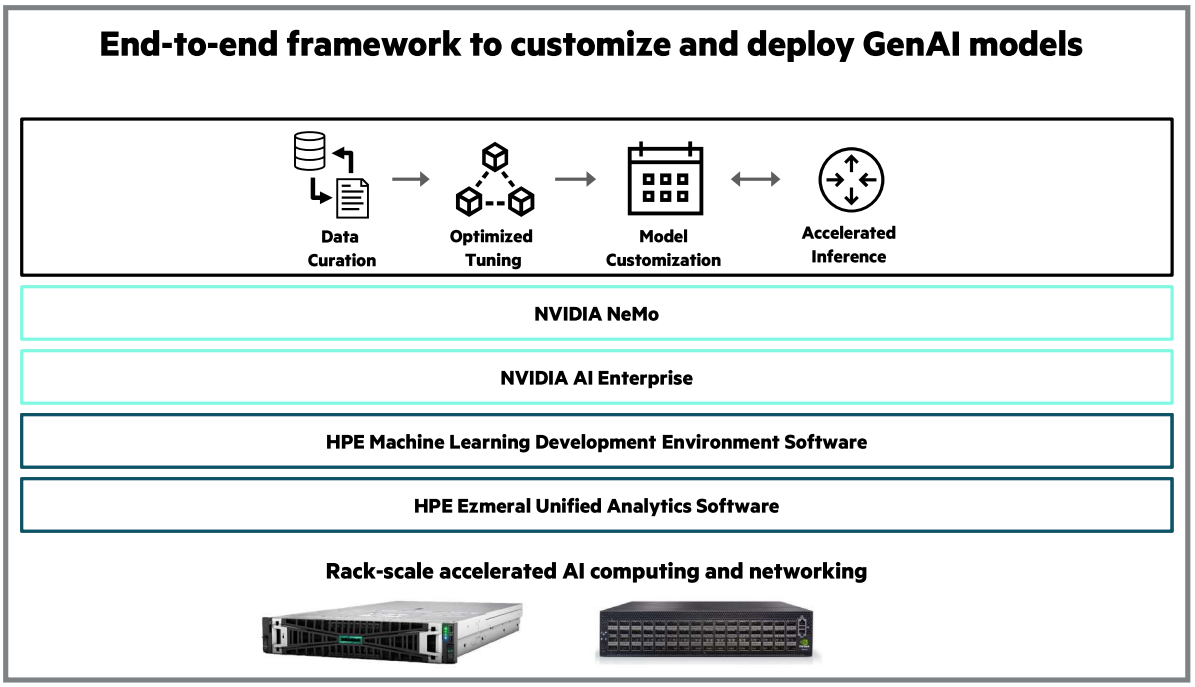
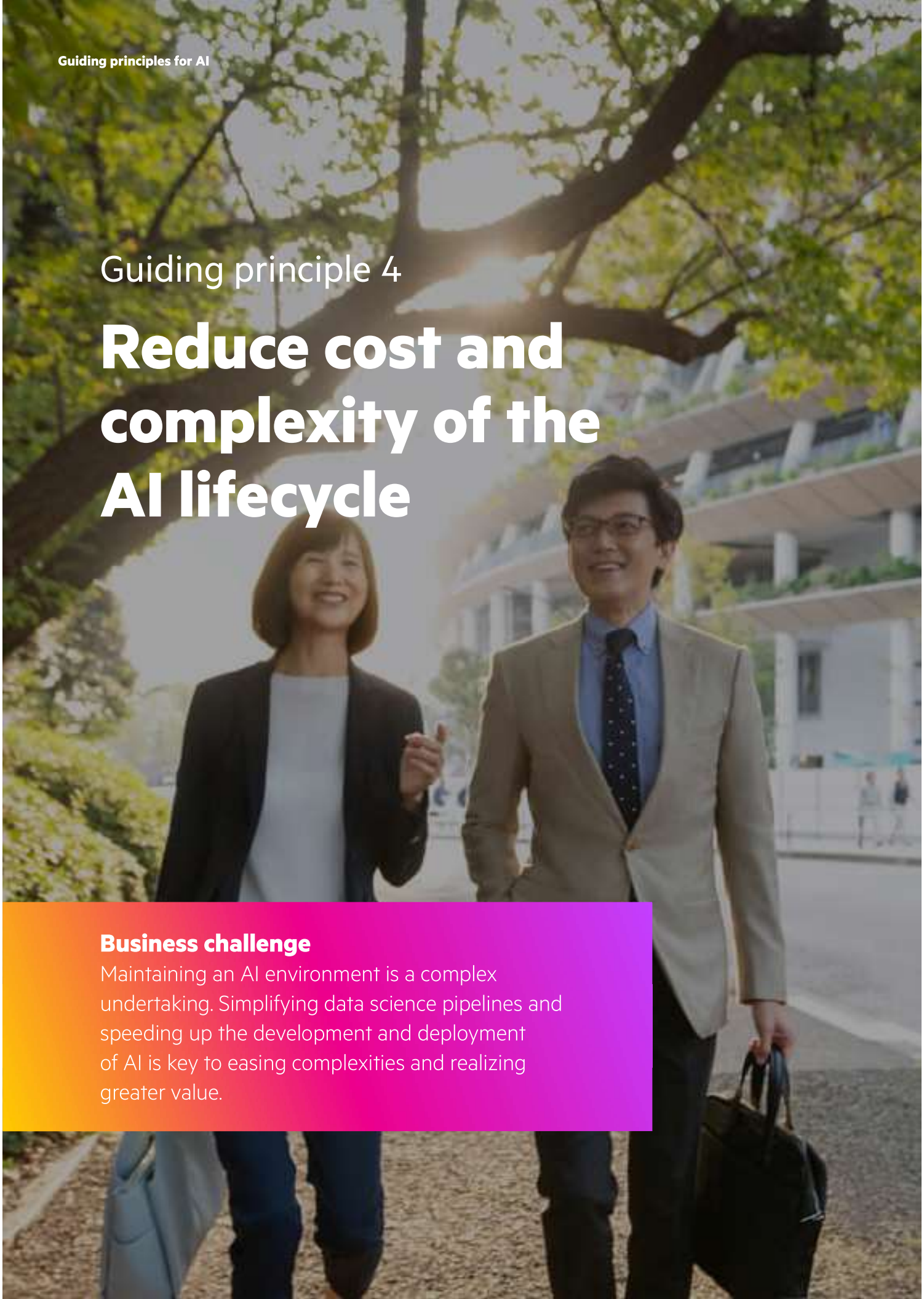


Figure 1. End-to-end framework to customize and deploy GenAI models

Expected results

- Simplify data access, analysis, and governance across on-premises and cloud and on-premises deployments
- Unlock new innovation with a curated GenAI environment, fine-tuned with enterprise data
- Improve productivity with a fully curated solution for open-source analytics tools
- Optimize cost and utilization across environments with a SaaS approach that provides predictable economics and precise controls



Guiding principle 4

Reduce cost and complexity of the AI lifecycle

Business challenge

Maintaining an AI environment is a complex undertaking. Simplifying data science pipelines and speeding up the development and deployment of AI is key to easing complexities and realizing greater value.

As GenAI models grow in size and scope, the complexity of the AI lifecycle is also on the rise. Enterprises running their business on AI require fast, stable, and secure environments that are highly available and easy to use. Investing in enterprise-grade software is a critical strategy for overcoming common roadblocks to AI and accelerating time to value.

With NVIDIA AI, enterprises get the complete platform needed to adopt GenAI, drive differentiation, fuel growth, and shape the future of AI-driven solutions. NVIDIA AI Enterprise is an end-to-end, cloud-native software platform featuring more than 100 frameworks, pretrained models, and development tools. This operating system for enterprise AI accelerates data science pipelines while streamlining the development and deployment of production-grade AI applications — such as GenAI, computer vision, and speech AI. When combined with HPE systems that are NVIDIA-certified, NVIDIA AI Enterprise helps ensure you get the right level of performance, scalability, and enterprise-grade support for any workload.

With NVIDIA AI Enterprise, you can explore NVIDIA AI Foundation Models from your browser, including state-of-the-art pretrained GenAI models and domain-specific frameworks. NVIDIA AI Foundation models include community models that have been further optimized for performance as well as NVIDIA Nemotron models built by NVIDIA with responsibly sourced data. The interface offers a user-friendly experience to explore GenAI models without any setup. Once you select a model, you can customize it and take it to production.

Other benefits include:

- **Enterprise-grade software and support:** Keep AI projects on track with regular security updates, API stability, workload and infrastructure manageability, and enterprise support
- **Optimized for accelerated AI:** Improve productivity and lower total cost of ownership (TCO) with a software development platform optimized for accelerated infrastructure
- **Certified to run everywhere:** Take advantage of workload portability across hybrid and multicloud environments
- **Custom model development:** Build and deploy custom GenAI models faster with reference applications and NVIDIA NeMo™

NVIDIA NeMo is a cloud-native framework that offers enterprises an easy, cost-effective, and fast way to adopt GenAI and is included with NVIDIA AI Enterprise. It includes training and inferencing frameworks, guard railing toolkits, data curation tools, and best-in-class pretrained models. NVIDIA NeMo is a complete solution for the LLM pipeline. You can rapidly build, customize, and deploy LLMs at scale with capabilities to curate data, train large-scale models up to trillions of parameters, and deploy them for inference. These cutting-edge training techniques enable you to build GenAI models anywhere while also reducing time to solution and increasing your return on investment (ROI).

NVIDIA NeMo is offered as an optimized, full-stack solution designed to accelerate enterprises with support, security, and API stability as part of NVIDIA AI Enterprise.

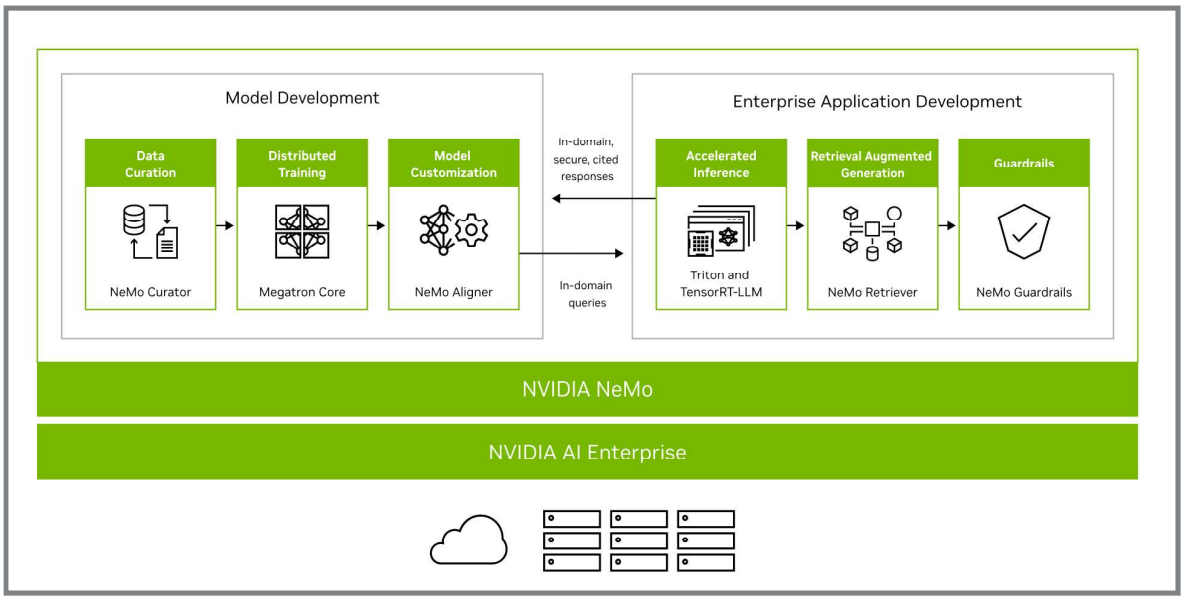
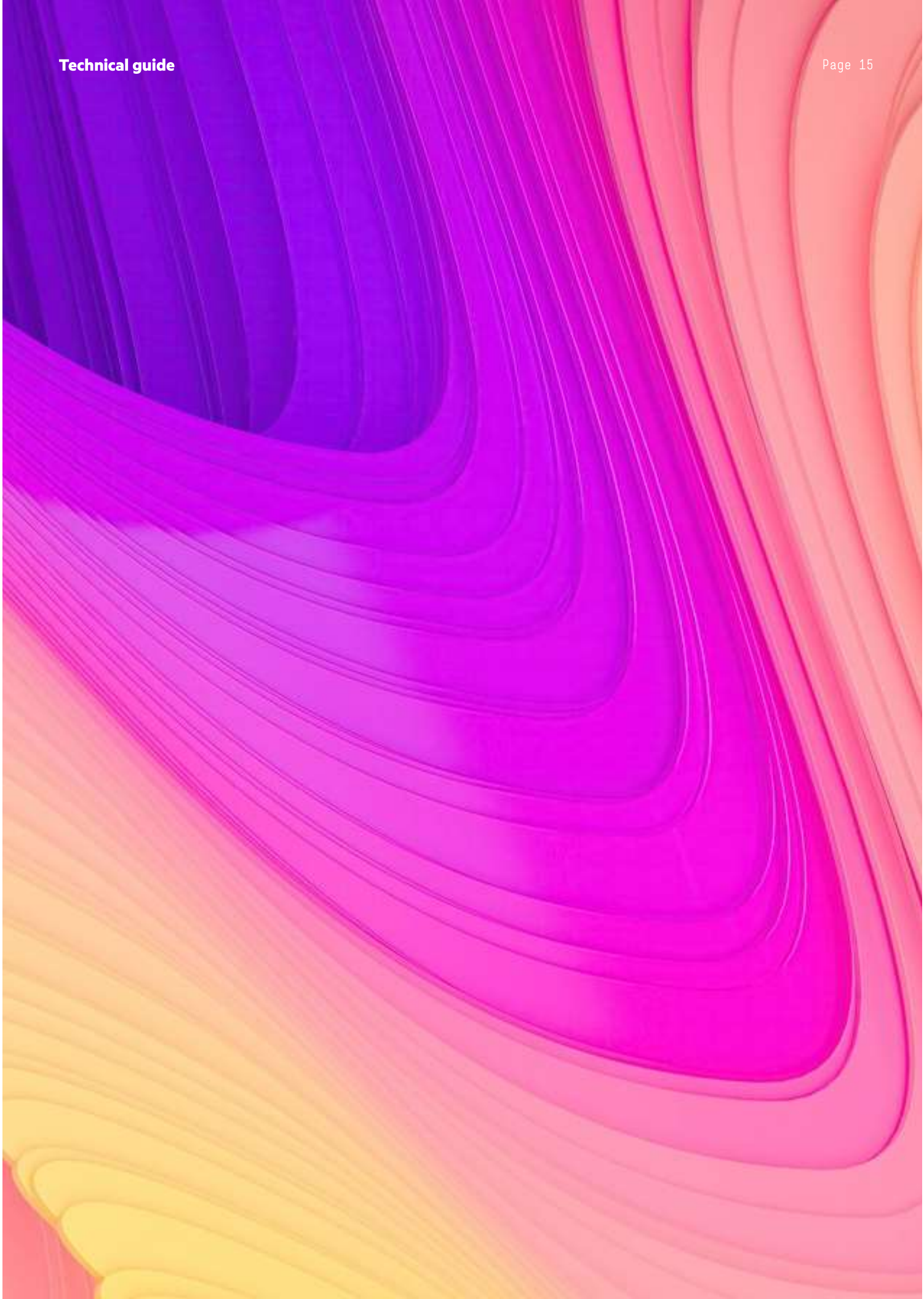
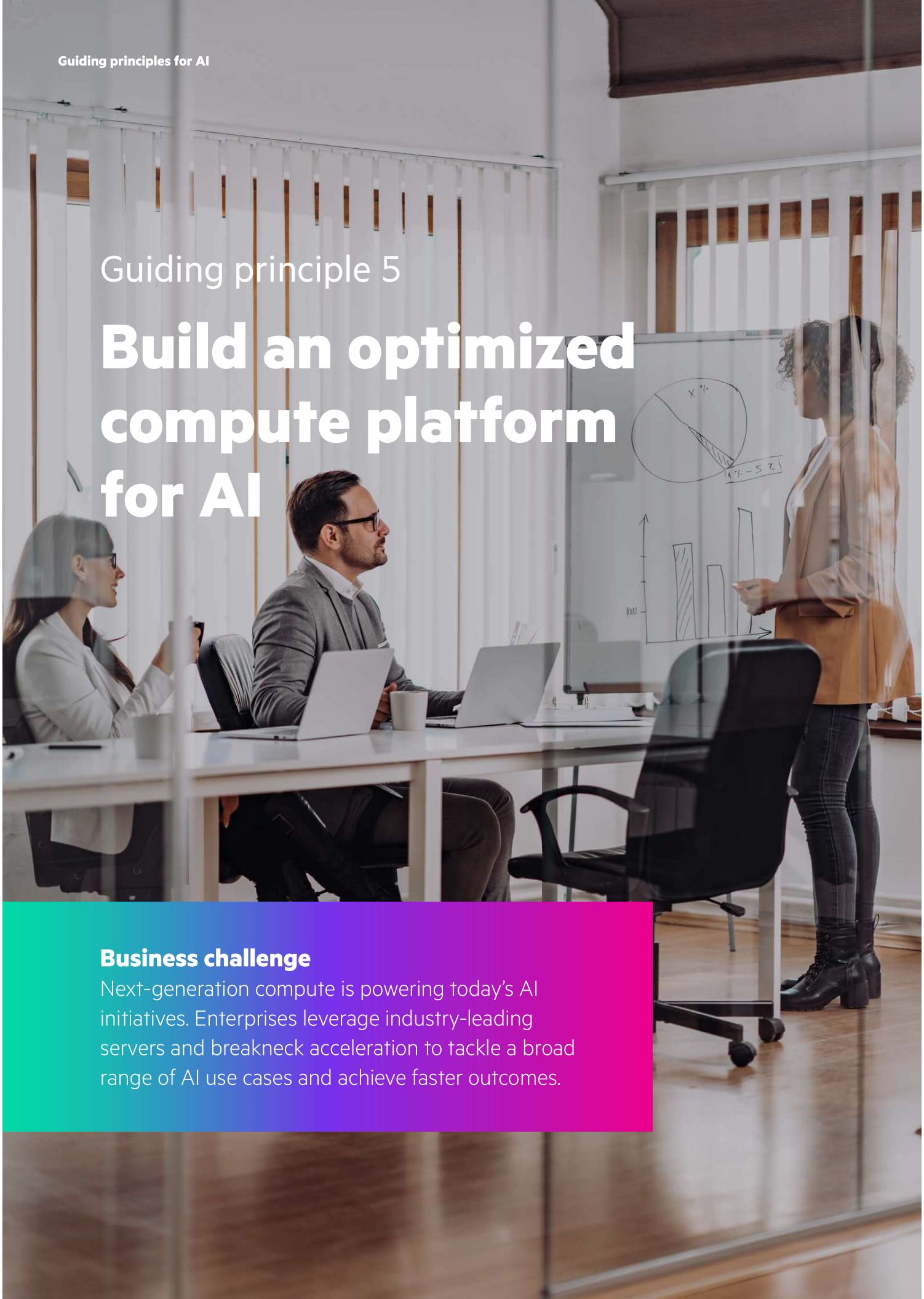


Figure 2. Complete solution for building enterprise-ready LLMs

Expected results

- Fuel business transformation with easy-to-use recipes and tools for GenAI
- Optimize model development with multinode and multi-GPU training and inference
- Ensure a smooth transition from pilot to production
- Help maximize throughput and help minimize LLM training time with advanced customization tools





Guiding principle 5

Build an optimized compute platform for AI

Business challenge

Next-generation compute is powering today's AI initiatives. Enterprises leverage industry-leading servers and breakneck acceleration to tackle a broad range of AI use cases and achieve faster outcomes.

GenAI workloads span a variety of use cases across industries including healthcare and life sciences, financial services, manufacturing, energy, and retail. These workloads are often compute-intensive and data-heavy, requiring next-generation compute platforms that are efficient, scalable, and secure. Many enterprises struggle to meet these massive performance requirements from data center to edge to cloud.

Enterprises can capitalize on the AI revolution by adopting an optimized compute platform for AI.

HPE ProLiant Gen11 Servers are engineered for a hybrid world. With a cloud operating experience, built-in security, and groundbreaking performance to drive your business forward. The portfolio delivers the best of on-premises and cloud computing to resolve today's hybrid cloud infrastructure challenges:

- An intuitive cloud operating experience
- Optimized for diverse AI workloads
- Ultra-scalable architecture
- HPE trusted security by design

HPE ProLiant Gen11 Servers with NVIDIA® L40S GPUs are powering the world's most complex AI, ML, and deep learning workloads. With up to 33% more GPU density[4] in a range of form-factors, you can power a wide range of applications. Delivering breakthrough performance and 48 GB of memory capacity, combined with enterprise-class stability and reliability, the NVIDIA L40S is the ideal accelerator for AI- and GPU-enabled applications. The NVIDIA L40S reduces the time to completion for model training and development and data science data prep workflows by delivering higher throughput and support for a full range of precisions, including FP8.

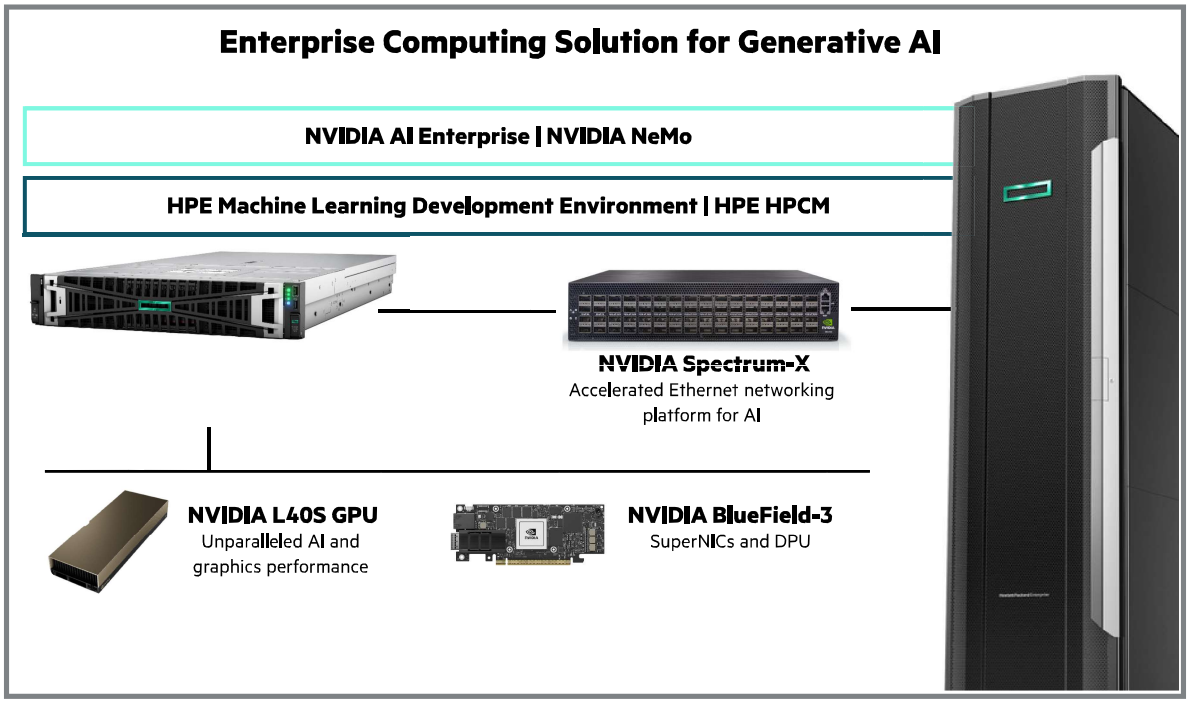


Figure 3. Enterprise computing solution for Generative AI

[4] hpe.com/us/en/servers/proliant-artificial-intelligence.html

The HPE ProLiant Gen11 portfolio takes a fresh approach to GPU support — moving the GPU to the front of the chassis to improve airflow and provide a dedicated power supply to each GPU to improve uptime. This way, HPE ProLiant Gen11 servers can effectively match the requirements for a variety of AI-specific workloads by powering high-wattage GPUs. The front-end cage design enables NVIDIA GPUs to communicate directly with each other. This communication allows the available memory of the GPUs to be combined, which in turn increases the speed of data exchange. You have the flexibility to add or remove GPUs as your AI processing needs change, giving you the ability process AI models with 100+ billion parameters.

Taking your decision-making to the next level, HPE ProLiant Gen11 servers that are NVIDIA-certified offer you a reference design for building scalable units of accelerated computing. The NVIDIA certification helps ensure optimum performance, reliability, and scale for a diverse range of AI workloads.

AI inferencing solutions from HPE are designed to speed your time to value. These systems are certified and tuned for edge or data center deployments targeting computer vision, generative visual AI, and NLP inference. The solutions are built on the new HPE ProLiant Gen11 servers and integrated NVIDIA GPUs, along with key AI frameworks. These HPE ProLiant AI solutions boost inference performance by more than 5x from previous generation systems.[5]

Expected results

- Deploy easy-to-upgrade servers that scale to meet your AI needs from edge to data center to cloud
- Run large, complex models and tackle increasingly challenging AI tasks with high efficiency
- Support parallel processing and reduce training time using multiple GPUs
- Simplify and automate operations across the server lifecycle, no matter where infrastructure lives

[5] hpe.com/us/en/servers/proliant-artificial-intelligence.html

Guiding principle 6

Accelerate and secure AI infrastructure

Business challenge

AI requires exceptional networking speeds delivering high bandwidth and low cut-through latency for cloud-native applications. A flexible, secure configuration is critical to run accelerated workloads at scale.

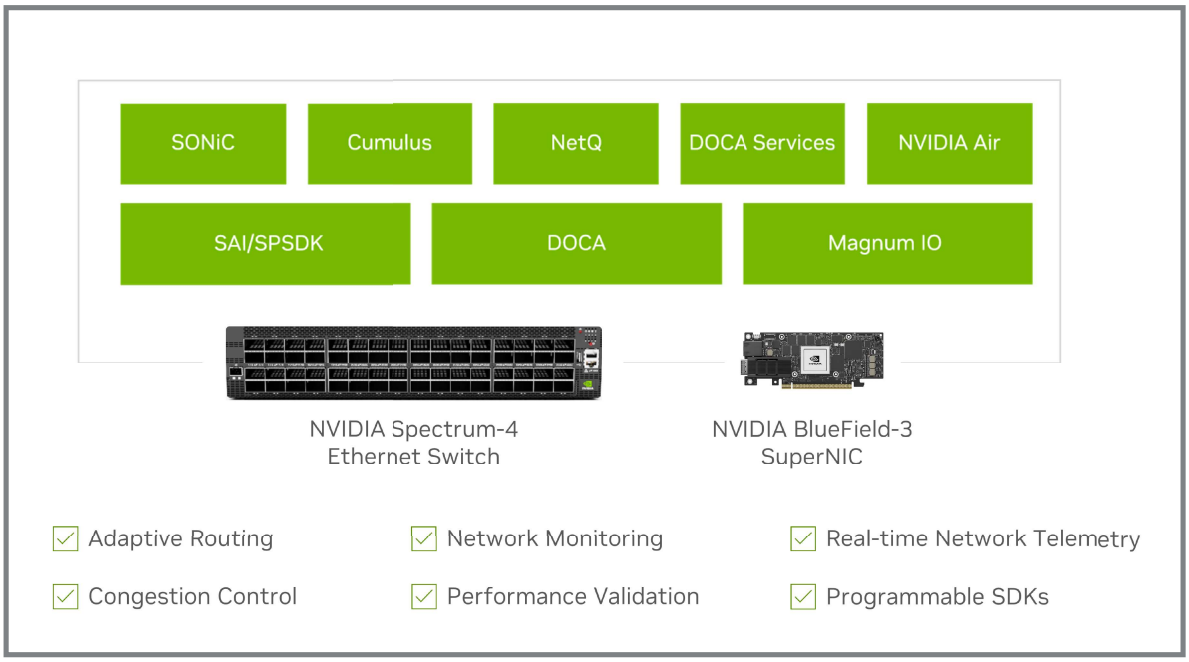
Next-generation compute technologies are powering the world's data centers. Savvy organizations are investing in new innovations to transform how they work with unmatched agility and security. NVIDIA understands the demands of GenAI and offers purpose-built solutions to help you run today's workloads and evolve for what comes next.

NVIDIA Spectrum™-X[6] is the world's first Ethernet networking platform for AI. This revolutionary solution enables you to build multitenant, hyperscale AI clouds with Ethernet. By implementing an accelerated networking platform, you can easily provision and manage powerful AI applications, significantly improve the performance and power efficiency of AI clouds, and gain higher predictability and consistency. NVIDIA Spectrum-X enhances GenAI network performance by 1.6x[7] to ramp up the processing, analysis, and implementation of AI workloads. This enables you to achieve faster time to market for your AI solutions.

Other key benefits include:

- Open network operating system with end-to-end stack optimization
- Deterministic performance and performance isolation
- Nearly perfect bandwidth and extremely low latency at scale
- RoCE network connectivity between GPU servers for peak workload efficiency

Powered by the tight coupling of the NVIDIA Spectrum-4 Ethernet switch[8] and the NVIDIA® BlueField®-3 networking platform,[9] NVIDIA Spectrum-X delivers the highest possible performance for AI, ML, and NLP processing for diverse industry applications. NVIDIA Spectrum-4 switches are designed to accelerate hyperscale GenAI fabrics, delivering port speeds of up to 800 Gb/s. With high performance, consistent low latency, and support for advanced data center networking features, the technology is ideal for improving AI cloud networks and end-to-end data center fabrics.



[6, 7] nvidia.com/en-us/networking/spectrumx/

[8] nvidia.com/en-us/networking/ethernet-switching/

[9] nvidia.com/en-us/networking/products/data-processing-unit/

Additionally, NVIDIA Spectrum-4 offers a complete security suite for all switch layers. Hardware, firmware, and software are authenticated by a built-in root of trust, from the basic input/output system (BIOS) to the network operating system (NOS). Any attempt to run a modified component or image that wasn't originally signed by NVIDIA is automatically blocked, helping ensure the safest network installation within data centers.

NVIDIA BlueField-3 drives unprecedented innovation for modern data centers and supercomputing clusters. Supporting both Ethernet and InfiniBand connectivity, this novel class of network accelerators is ushering in a new era of AI cloud computing offering speeds up to 400 Gb/s.

Drawing on the platform's extensive capabilities, NVIDIA BlueField DPUs and NVIDIA BlueField SuperNICs can transform your traditional computing environments into fast, efficient, and sustainable data centers suitable for any workload at any scale.

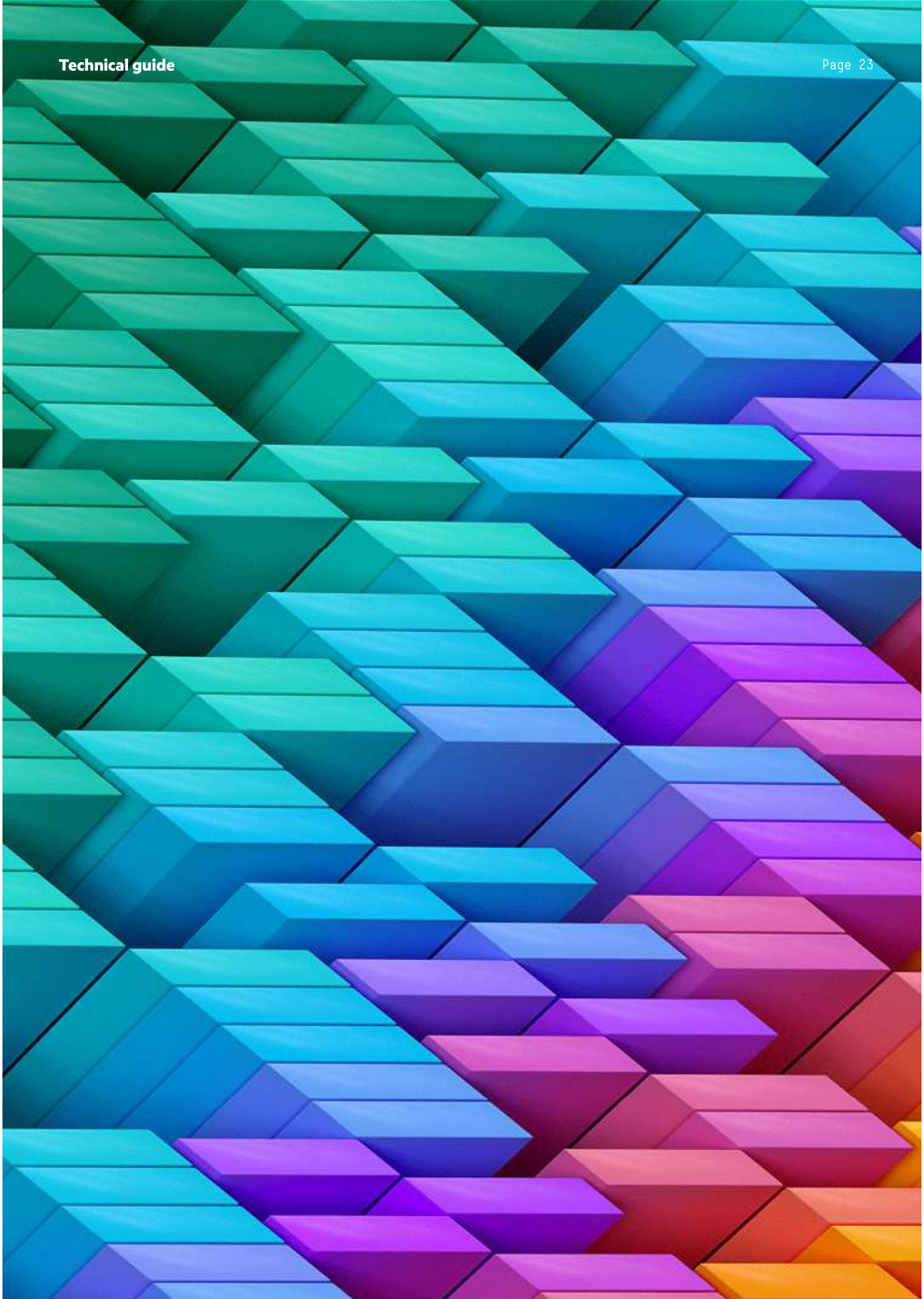
From accelerated AI to hybrid cloud, high-performance computing to 5G wireless networks, BlueField-3 redefines the art of the possible. With powerful computing, high-speed networking, robust cybersecurity, and extensive programmability to deliver software-defined, hardware-accelerated solutions, it creates a secure and agile infrastructure to accommodate massively parallel computing and hyperscale AI workloads in any environment.

Creating a new era of AI cloud computing, the BlueField-3 SuperNIC enables secure, multitenant data center environments with deterministic and isolated performance between jobs and tenants.

You can be among the enterprises worldwide that are investing in NVIDIA BlueField to accomplish more with AI in real time.

Expected results

- Create an optimized AI network with accelerated Ethernet for every data center
- Gain a competitive edge in AI with industry-leading software-defined, hardware-accelerated solutions
- Support cloud-native AI applications with reliable high bandwidth and low latency
- Deliver superior performance, security, and reduced TCO for cloud computing platforms



Guiding principle 7

Develop a plan to accelerate transformation

Business challenge

Expert services can help turn data into game-changing results. HPE Services offers technical, professional, and advisory support to simplify each stage of the AI journey.

From planning to operating and beyond, HPE experts are available to accelerate edge-to-cloud transformation, help maximize IT investments, and optimize how you work. We help you tap into the power of AI, data, and advanced analytics to uncover insights and employ intelligence in revolutionary, ethical, and secure ways. We make it simpler to produce AI-driven automated business outcomes working closely with your data, business, and IT teams.

HPE Services is fueling business transformation with AI and hybrid cloud services. Our extensive list of Advisory and Professional Services for AI and data platforms can advance your journey from AI pilot to production — from edge to cloud. This broad portfolio includes consulting services, workforce training, and deployment solutions to help you overcome the barriers to AI:

- **We build:** Work with HPE experts to develop and deploy your AI solution from end to end
- **We optimize:** Bring your own AI model and HPE optimizes it for you
- **We integrate:** Integrate leading partner solutions into your environment

HPE AI Services guide you through the entire AI lifecycle, from GenAI and LLM discovery to implementation, where you develop the optimum operational models and hybrid cloud data strategies needed to build, deploy, and scale solutions into transformative results. Learn how our AI professionals use your business objectives and desired outcomes to select the GenAI models and technical resources needed for your project. Once you identify a use case and the related data, the next step is to experiment with the implementation, measure the value generated, and bring it to the core of your business. You have access to a broad set of offerings with HPE AI Services to meet your specific requirements. HPE AI experts can even implement the model for you and optimize it for inference, resulting in a ready-to-use solution that can integrate with existing processes or operate as a stand-alone application.

These comprehensive services are supported by HPE Global Centers of Excellence for AI and Data in the United States, Spain, Bulgaria, India, and Tunisia. To learn more, check out HPE consulting services or contact your local representative.

Plan to attend an HPE AI Services — Transformation Workshop, which covers the latest AI trends (such as AI ethics and GenAI domains). You receive a highly interactive full-day experience facilitated by HPE AI advisory consultants. During the workshop, we explore your business needs, use case objectives, and priorities for business, data, and IT stakeholders. We help you align your AI goals and requirements to develop your own initiative road map.

Expected results

- Start your digital transformation in the right direction or give it a boost anywhere along your journey
- Become a data-first organization, enabling unified data access for AI applications at scale
- Learn to bring the cloud experience to workloads everywhere within your hybrid cloud
- Achieve improvement in measurable KPIs from your AI services engagement
- Run large, complex models and tackle increasingly challenging AI tasks with high efficiency
- Support parallel processing and reduce training time using multiple GPUs
- Simplify and automate operations across the server lifecycle, no matter where infrastructure lives

Conclusion

AI is everywhere, disrupting numerous industries and opening unlimited possibilities. AI can turn questions into discovery, insights into action, and imagination into reality.

To help maximize your time and investment, HPE and NVIDIA are committed to defining and optimizing infrastructure for the next era of AI. We bring together market-leading hybrid cloud, supercomputing, and AI capabilities more broadly to the enterprise to drive AI-powered transformation.

The result is our enterprise computing solution for GenAI, which is expertly engineered to enable faster AI outcomes and greater inferencing performance to help maximize the impact of GenAI across your business. Our full-stack AI-native solutions are designed to accelerate time to value. With a curated mix of software and infrastructure, you have the critical tools and capabilities to develop high-quality AI models quickly and securely.

HPE and NVIDIA bring you the advantages of purpose-built AI architecture for today and the future:

- Data-first pipeline to manage public and proprietary data across multigen IT
- AI lifecycle management software to accelerate workflows for training, tuning, and inferencing
- Hybrid by design to run AI anywhere from edge to cloud with data protection
- Supercomputing DNA built into the entire portfolio, sustainable by design, to train the largest models
- Open ecosystem for freedom of choice with no lock-in

Are you ready to use AI to deliver real advantages for your business? We can help you turn your AI strategy into successful projects. Contact HPE to get started.

Learn more at

HPE.com/hpe-enterprise-computing-solution-gen-ai

Chat now (sales)

a00137097ENW, Rev. 1

Hewlett Packard Enterprise