DELL Technologies

How Dell Makes the AI Factory Real

AI's Unique Needs Demand a Fresh Approach—The AI Factory

FEATURING RESEARCH FROM FORRESTER

The Forrester Wave™: AI Infrastructure Solutions, Q1 2024

It's Dell's strong belief that AI will be the defining technology of the years to come—and we want to help organizations set themselves on the path to success. The AI revolution is here, and it's transforming business at an unprecedented pace. Success in the evolving global economy will depend on how quickly and effectively organizations can harness the potential of AI. To stay relevant, accelerating toward an AI future is no longer a choice; it's imperative.

But the AI transformation is unlike any we've seen before. Its adoption is faster, its scope is broader and it offers unparalleled potential. This means organizations must build a strong AI foundation today—or risk obsolescence.

IN THIS DOCUMENT

What Is an AI Factory

In the age of AI, traditional IT infrastructures and operating models are not enough to handle the rigorous demands of AI. A new approach is required that's tailored to meet AI's needs. This is the AI factory.

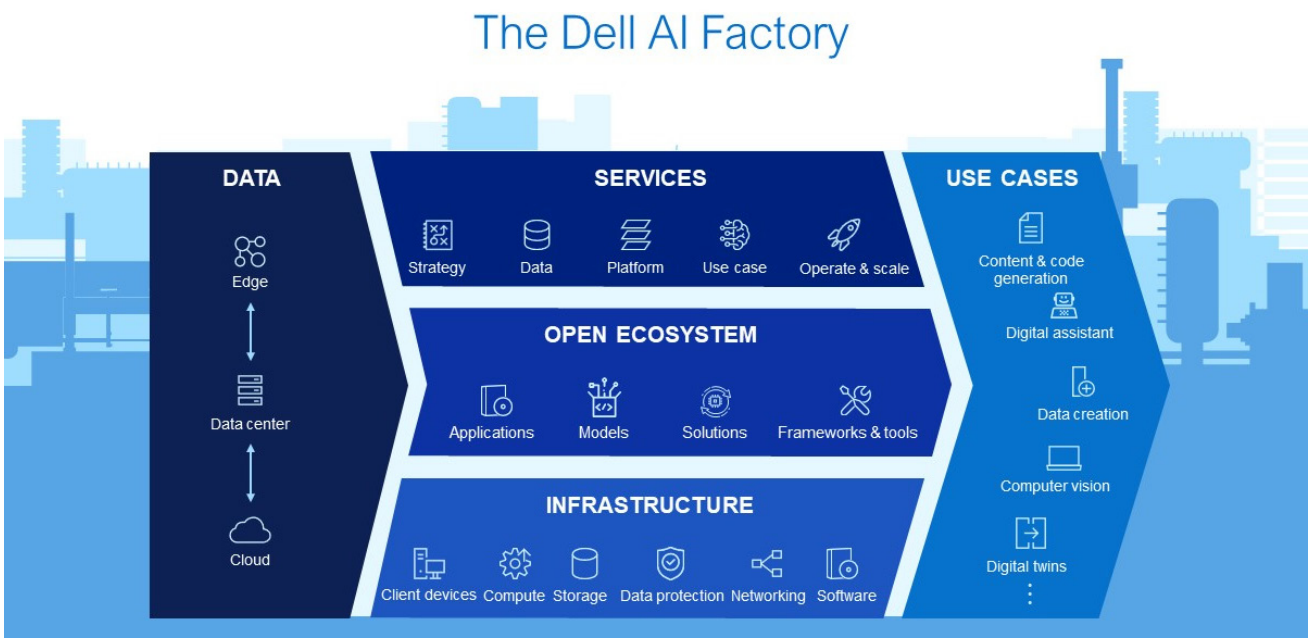
Just like physical factories fueled the industrial revolution, AI factories will drive the AI revolution. But instead of producing physical goods, they create actionable intelligence, fresh content and new insights. Every business will come to need an AI factory, and those most adept at establishing their AI factories to produce fast, repeatable results will have a critical advantage as we move further into the AI era.

What Is Dell's Vision for the AI Factory

We want to help organizations build their own AI factories to create transformative outcomes, consistently, and at scale, so they can gain an edge. And we're doing this through the Dell AI Factory.

The Dell AI Factory is our approach to help accelerate AI innovation in organizations of all sizes. It comprises a portfolio of products, solutions and services tailored for AI workloads and designed for fast, repeatable outcomes. It's versatile and capable of operation across various environments—be it cloud, data centers, workstations, AI PCs or edge locations.

Let's break it down further and look at the Dell AI Factory under the hood.



How the Dell AI Factory Works

As organizations seek to build their own AI factories and take advantage of all AI has to offer, many will want to get to speed as fast as possible. The Dell AI Factory is how we're helping them do exactly that.

Use Cases: Powering Business Outcomes

The power of a factory lies in what it produces, and the AI factory produces business outcomes powered by an organization's most impactful AI use cases. Dell Technologies assists customers in identifying and prioritizing the AI use cases that drive the greatest business impact and simplifies the deployment and scaling of those AI use cases and applications with our validated solutions.

Data: Fueling the AI Factory

Data is the raw material that powers the AI factory. But the success of any AI initiative depends on the quality of data used. An organization's data often resides in multiple locations, frequently on-premises and at the edge. The Dell AI Factory brings AI as close as possible to where data resides to minimize latency, lower costs and maintain data security by keeping sensitive information within a controlled environment. It also provides a way to prepare this data for use by the AI factory, ensuring that customers are working with quality and accurate data, with easy access and built-in data governance.

Infrastructure: The AI Factory Foundation

Infrastructure—comprising client devices, servers, storage, data protection and networking—forms the foundation of every AI factory, but this foundation must be flexible because the future of AI is rapidly evolving. That's why the Dell AI Factory can support diverse AI requirements with the world's broadest AI solutions portfolio from desktop

to data center to cloud,[1] allowing organizations to right-size their AI investments and gain the flexibility to run AI anywhere. We provide advanced technology across all infrastructure areas, along with the benefits of an end-to-end portfolio for easier deployment, operation, and maintenance of AI factory infrastructure.

Open Ecosystem: Ensuring Seamless Integration

Just as a traditional factory relies on a diverse supply chain, the AI factory depends on a varied and dynamic To learn more about the Dell AI organizations chart their course through this dynamic environment by identifying suitable ecosystem partners and working with them to provide solutions that simplify AI deployment and operation for our customers. With a history of fostering open ecosystems, Dell brings a commitment to inclusivity and access to innovation, offering customers a broader range of technology solutions and more flexibility.

Services: Help When You Need It

An effective AI factory needs a skilled team to succeed, but AI-ready skills are in short supply and the ecosystem is diverse. That's why professional services is a critical layer of the Dell AI Factory, helping customers align their business and technology needs. From strategy formulation to implementation, and management to scaling, Dell provides expert assistance at any stage of an AI journey based on extensive global experience and comprehensive knowledge in deploying AI at scale.

The Dell AI Factory: Helping You Build Your AI Future

The Dell AI Factory is more than just a set of services or products; it's our vision to help organizations harness the full potential of AI. With our broad AI portfolio and leading infrastructure, an open ecosystem and services offerings, we are helping organizations around the world build their own AI factories to transform and innovate.

To learn more about the Dell AI Factory, explore how we're helping organizations bring AI to their data.

[1] Based on Dell analysis, March 2024. Dell Technologies offers hardware solutions engineered to support AI workloads from Workstations PCs (mobile and fixed) to Servers for High-performance Computing, Data Storage, Cloud Native Software-Defined Infrastructure, Networking Switches, Data Protection, HCI and Services (CLM-009277)

WAVE REPORT

The Forrester Wave™: AI Infrastructure Solutions, Q1 2024

The 12 Providers That Matter Most And How They Stack Up

March 18, 2024

By Mike Gualtieri with Sudha Maheshwari, Sarah Morana, Jen Barton

FORRESTER®

Summary

In our 19-criterion evaluation of AI infrastructure providers, we identified the most significant ones and researched, analyzed, and scored them. This report shows how each provider measures up and helps enterprise technology professionals select the right one for their needs.

Additional resources are available in the online version of this report.

Workload Multiplicity And Cost Optimization Matter Most

AI is banging at your door. Don't answer unless you're prepared to feed its voracious appetite for specialized compute, storage, and network. That's AI infrastructure. And it's designed to satiate AI's need for three core AI workloads: data preparation, model training, and model inferencing. Whether it's training custom models for competitive edge or leveraging open-source generative AI models, enterprise technology leaders must invest in AI infrastructure wisely. It's not just a choice between cloud or on-premises. It's about aligning AI infrastructure investments with your overall infrastructure strategy to optimize cost, balanced with internal demand.

As a result of these trends, AI infrastructure solutions customers should look for providers that:

- **Maximize the performance of core AI workloads.** The three core AI workloads are data prep, training, and inferencing. Each has starkly different infrastructure requirements for throughput, latency, fault-tolerance, and cost. Even within the three workloads, there are different requirements. For predictive AI, a data prep workload needs query access and transformations on mostly structured data, while for generative AI it needs to process globs of unstructured data. Deep learning for computer vision or large language models absolutely needs GPUs (or other chip architectures designed for AI), but a predictive model might not benefit from GPUs. AI infrastructure solutions in this evaluation cover all workloads; however, it makes sense for enterprises to choose multiple vendors based on specific needs. An enterprise might choose an on-premises solution for data management and training but choose a hyperscaler for inferencing, and vice versa. Enterprise technology leaders must inventory the AI workloads currently in use and anticipate future AI workloads.
- **Offer a management layer to optimize cost and tame complexity.** AI infrastructure comes with management software to help operations professionals monitor the system, control access, allocate usage, and provision/deprovision infrastructure to optimize costs. This management software is different from AI/ML platforms, which are designed for AI teams to build AI applications rather than to manage AI infrastructure. Some vendors in this evaluation offer both AI infrastructure and AI/ML platforms and thus there may be some overlap and advantages to tight integration between the two. Enterprise buyers must understand how a vendor's AI infrastructure management layer can be incorporated with its existing infrastructure management tools, policies, and ITOps

practices. If an enterprise has already standardized on a vendor's non-AI infrastructure, then using that vendor's AI infrastructure can be attractive from a management point of view.

- **Match your enterprise's commitment to AI.** There are vectors of commitment for each vendor that form its sweet spot. One of the most touted AI superlatives is how many billions of parameters there are in the model of the week. However, your enterprise may never train a model that large. Instead, you may call a very large model from a cloud service or one downloaded from HuggingFace and installed on your own infrastructure. Likewise, you must consider if you have significant use cases (or even just one) for AI on the edge, in which case a cloud-only solution is prohibitive. Your company may do biochemical research and need an AI infrastructure that is tightly integrated or part of a high-performance computing (HPC) environment where massive simulations are also required. If technology leaders take a step back, they will see a bigger context for AI and can make decisions about AI infrastructure based on the totality of the enterprise's AI strategy. For some enterprises, this may mean investing in more than one AI infrastructure vendor.

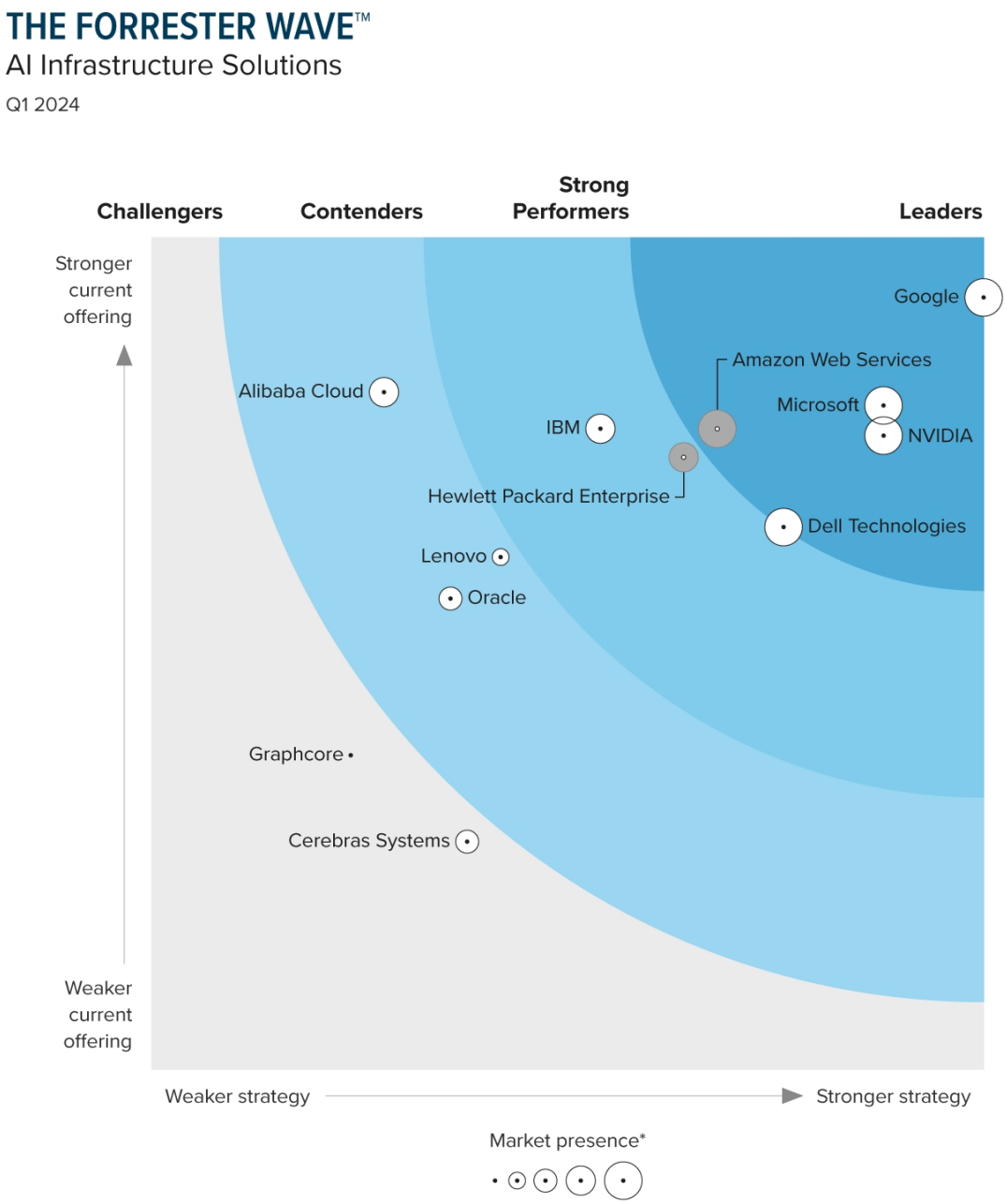
Evaluation Summary

The Forrester Wave™ evaluation highlights Leaders, Strong Performers, Contenders, and Challengers. It's an assessment of the top vendors in the market; it doesn't represent the entire vendor landscape. You'll find more information about this market in our reports on AI infrastructure solutions.

We intend this evaluation to be a starting point only and encourage clients to view product evaluations and adapt criteria weightings using the Excel-based vendor comparison tool (see Figures 1 and 2). Click the link at the beginning of this report on Forrester.com to download the tool.

Figure 1

Forrester Wave™: AI Infrastructure Solutions, Q1 2024



*A gray bubble or open dot indicates a nonparticipating vendor.

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Figure 2

Forrester Wave™: AI Infrastructure Solutions Scorecard, Q1 2024

	Forrester's weighting	Alibaba Cloud	Amazon Web Services*	Cerebras Systems	Dell Technologies	Google	Graphcore	Hewlett Packard Enterprise*	IBM	Lenovo	Microsoft	NVIDIA	Oracle
Current offering	50%	4.07	3.85	1.37	3.26	4.64	1.89	3.68	3.85	3.08	3.99	3.81	2.83
Solution	25%	3.00	3.66	1.00	4.34	4.34	1.00	3.66	4.32	4.32	4.34	5.00	1.66
Workloads	25%	3.66	4.32	1.68	3.68	5.00	2.34	3.66	3.66	3.00	5.00	3.02	3.66
Tools	25%	5.00	4.00	1.00	2.00	5.00	2.00	4.00	4.00	2.00	4.00	3.00	3.00
Deployment	25%	4.60	3.40	1.80	3.00	4.20	2.20	3.40	3.40	3.00	2.60	4.20	3.00
Strategy	50%	1.40	3.40	1.90	3.80	5.00	1.20	3.20	2.70	2.10	4.40	4.40	1.80
Vision	20%	1.00	3.00	3.00	5.00	5.00	1.00	3.00	3.00	1.00	3.00	3.00	1.00
Innovation	25%	1.00	3.00	3.00	3.00	5.00	1.00	5.00	3.00	1.00	5.00	5.00	1.00
Roadmap	25%	1.00	3.00	1.00	3.00	5.00	1.00	1.00	1.00	3.00	5.00	5.00	1.00
Partner ecosystem	10%	1.00	5.00	1.00	5.00	5.00	1.00	3.00	3.00	3.00	5.00	5.00	3.00
Pricing transparency	10%	5.00	5.00	1.00	3.00	5.00	3.00	3.00	3.00	1.00	5.00	3.00	5.00
Supporting services and offerings	10%	1.00	3.00	1.00	5.00	5.00	1.00	5.00	5.00	5.00	3.00	5.00	3.00
Market presence	0%	4.00	5.00	3.00	4.00	5.00	1.00	4.00	4.00	2.00	5.00	5.00	3.00
Revenue	50%	3.00	5.00	3.00	5.00	5.00	1.00	5.00	5.00	3.00	5.00	5.00	3.00
Number of customers	50%	5.00	5.00	1.00	3.00	5.00	1.00	3.00	3.00	1.00	5.00	5.00	3.00

All scores are based on a scale of 0 (weak) to 5 (strong).
*Indicates a nonparticipating vendor

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Vendor Offerings

Forrester evaluated the offerings listed below (see Figure 3).

Figure 3

Evaluated Vendors And Product Information

Vendor	Product evaluated
Alibaba Cloud	PAI (Platform for Artificial Intelligence)
Amazon Web Services	AWS Machine Learning Infrastructure
Cerebras Systems	Cerebras CS-2 Artificial Intelligence System and Wafer-Scale Clusters
Dell Technologies	Dell Generative AI Solutions
Google	Vertex AI Platform, Cloud GPU, Cloud TPU, Dynamic Workload Scheduler, GKE, Dataproc, Dataflow, Deep Learning VM, Deep Learning Container
Graphcore	Bow Pod platforms (latest version of IPU-POD platforms)
Hewlett Packard Enterprise	HPE Machine Learning Development Environment (MLDE), HPE Machine Learning Data Management (MLDM), HPE Machine Learning Development System (MLDS), HPE Cray EX Supercomputers, HPE Cray XD 6500 Supercomputers, HPE Cray XD 2000 Supercomputers, HPE ProLiant servers, HPE Ezmeral Data Fabric
IBM	IBM Z, IBM z16, IBM LinuxONE Emperor 4, IBM LinuxONE Rockhopper 4, IBM Power10, IBM Storage Scale, IBM Storage Scale System, IBM Storage Fusion HCI, IBM Cloud, IBM watxonx.ai, watsonx.data, watsonx.governance, IBM Watson Studio, IBM Cloud Pak for Data Systems, Red Hat OpenShift, IBM Db2 Analytics Accelerator, IBM Db2 AI for z/OS, BM Operations Analytics for z Systems, IBM Cloud Pak for AIOps, IBM OMEGAMON, IBM Instana, IBM Turbonomic
Lenovo	Lenovo ThinkSystem; Lenovo ThinkEdge
Microsoft	Microsoft Azure
NVIDIA	NVIDIA DGX Platform
Oracle	OCI Data Science, OCI Generative AI service, OCI AI/ML Services

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Vendor Profiles

Our analysis uncovered the following strengths and weaknesses of individual vendors.

Leaders

- **Google offers the whole package for AI workloads.** AI continues to be a core capability of Google's many consumer and business services, such as internet search and advertising. So to say Google has a head-start is an understatement. Doing AI efficiently at Google-scale is a feat that few other companies in the world are capable of. Google brings that experience and infrastructure to Google Cloud

AI infrastructure. Google's early and ongoing investments in AI for its other businesses drives its vision of "where the puck is going to be" for enterprise AI. Google's superior roadmap and innovation is to make Google-scale accessible to all customers, whether a bright tiny startup or a large global enterprise, while at the same time abstracting the complexity with easy-to-use tools.

Google has strengths across the board with the highest scores of all the vendors in this evaluation. The only caveat is that customers don't have the option of using the totality of Google AI infrastructure on-premises. Internet-native customers appreciate the efficient elastic scale to run spikey workloads and control costs. Reference customers appreciate the tight integration with Google's AI platform, Vertex AI, to build AI solutions using both custom models and LLMs. Google is a good fit for existing Google Cloud customers and checks all the boxes for existing and new customers that wish to make Google Cloud its strategic public cloud.

- **Microsoft makes supercomputer AI infrastructure easy to use at cloud scale.** Microsoft offers numerous sizes of GPU-optimized virtual machines for direct use. The Azure AI portfolio offers several AI-centric services, such as Azure OpenAI Service and Azure AI Studio, to help customers develop custom AI applications that use Microsoft's underlying AI infrastructure. Microsoft's strategy is to bring AI to every application, every business process, and every employee. Microsoft plans to achieve this through a combination of business and productivity applications and by making Microsoft Azure AI infrastructure attractive for AI developers. Its $13 billion investment in OpenAI adds proof to the pudding. The company's superior innovation and roadmap is driven by infusion of AI into all of the company's business applications, developer tools, and cloud services.

  Microsoft has strengths in architecture, ecosystem, data preparation, model training, inferencing, and development tools. Microsoft can improve by adding AI-specific infrastructure management tools and expanding AI workload capabilities on Azure Arc for on-premises deployments. Reference customers appreciate the breadth of services Microsoft provides for data preparation, application development, cognitive services, and pretrained models such as OpenAI through its partnership. Microsoft AI infrastructure is a good fit for customers that have standardized on Microsoft Azure and need cloud-scale AI infrastructure.

- **NVIDIA sets the pace for AI infrastructure worldwide.** Without NVIDIA's GPUs, modern AI wouldn't be possible. In addition to providing its GPUs to most other vendors in this evaluation, NVIDIA offers AI infrastructure to customers directly via the NVIDIA DGX platform, which runs workloads on cloud with NVIDIA DGX Cloud

and/or on-premises. The company's innovation, roadmap, and vision are clear and have kept it moving at lightspeed compared to other semiconductor manufacturers for AI chips. The AI infrastructure it offers directly to customers isn't meant to aggressively compete with its many AI infrastructure partners, but rather to serve as a template for what a state-of-the-art system should look like.

NVIDIA has strengths in system architecture, partner ecosystem, configuration options, model training, and system management. NVIDIA has always been strong on model training and is improving on inferencing. The company relies on cloud service providers and other partners to accommodate scalable data preparation workloads. Now with NVIDIA DGX Cloud, customers have more workload deployment flexibility for experimentation and/or bursty workloads. Reference customers appreciate NVIDIA systems that are designed by and directly available from NVIDIA on-premises and now also in the cloud. Customers that want a state-of-the-art, on-premises system for model training will find NVIDIA an attractive option.

- **Amazon Web Services (AWS) is your one-stop AI shop with a wide range of options.** AWS's AI infrastructure portfolio is extensive. AWS's vision is to offer customers a wide range of options to run AI workloads from preconfigured instances to training services abstracted behind its AI development tool — Amazon SageMaker. Amazon's AI strategic infrastructure portfolio includes expected compute instances/virtual machines based on NVIDIA GPUs, but also instances based on Intel's Gaudi chips. AWS also offers AI infrastructure based on its own chips: AWS Inferentia for inferencing and AWS Trainium for training. Additional services include AWS Neuron SDK to make it easy to use AWS's custom chips, AWS Elastic Inference to optimize cost/performance, and AWS IoT Greengrass for edge inferencing.

  AWS has strengths in solution ecosystem, data preparation services, inferencing, and AI development tools. AWS could strengthen its current offerings by adding more AI workload capabilities for on-premises and edge deployments. Customers appreciate that AWS offers scalable AI infrastructure for the smallest of experiments to the largest production deployments. They also like Amazon SageMaker's integrated tooling. AWS AI infrastructure is ideal for customers that store training data in AWS and standardize on other AWS services that complement AI use cases. AWS declined to participate in the full Forrester Wave evaluation process.

- **Dell offers AI-ready architectures for all, could benefit with more development tools.** Running a growing number of diverse AI workloads can be complicated. Dell Technologies aims to make it easier by publishing meaty reference architectures for numerous AI workload scenarios. Dell offers many options that include its flagship PowerEdge Servers for AI and numerous storage solutions, such as PowerFlex and PowerScale. Dell's superior vision is to offer the quickest, most integrated solution to enterprises for on-premises and colocation deployments. The company can improve its roadmap and innovation with more AI-specific tooling.

  Dell has strengths in architecture (with its reference architectures), configurations, and model training. Dell could improve by designing AI-specific infrastructure management tools and include or partner more tightly with AI development tool providers. Dell has cloud capabilities for some of its AI workload components, but lacks a fully managed AI infrastructure solution. Reference customers appreciate Dell's exceptional level of service in quickly designing custom AI infrastructure that integrates with its existing IT infrastructure. Dell is a good fit for enterprises that wish to deploy on-premises or at a colocation and want an ongoing partnership to smoothly evolve AI infrastructure as demand increases.

Strong Performers

- **HPE powers AI from edge to cloud, but needs sharper messaging.** Hewlett Packard Enterprises (HPE) offers a broad range of hardware configurations from edge to supercomputer that can support AI workloads. Customers can use HPE GreenLake to make it all work as a private cloud. HPE's strategy is to offer a full range of AI infrastructure for on-premises and/or private cloud deployments. HPE also includes resource management software specifically designed for AI workloads. HPE's AI infrastructure is attractive to the most sophisticated AI researchers, but the company must expand its messaging to the growing numbers of enterprise AI developers. To do that, HPE needs to improve its roadmap with AI-specific optimizations that go beyond adding the next generation of GPUs.

  HPE has strengths in configurations, data preparation, management software, and locale (cloud and on-premises). HPE can improve by expanding its solution ecosystem and further optimizing model training. HPE customers appreciate the breadth of deployment options from edge to cloud and how the AI infrastructure integrates seamlessly with other HPE infrastructure. HPE is a good fit for customers who want to run AI workloads in both their own data centers and the

cloud. HPE declined to participate in the full Forrester Wave evaluation process.

- **IBM designs AI infrastructure for mission-critical workloads.** IBM's AI infrastructure is comprised of combinations of IBM Storage, IBM Cloud, IBM Power, and IBM Z. The company's vision is to be trusted AI infrastructure for mission-critical workloads for both cloud and on-premises. Therefore, the company offers AI infrastructure on-premises, in cloud, and in hybrid cloud. IBM Z mainframes still handle the world's most mission-critical, low-latency transactions, so customers with AI applications that make use of that data and low-latency inferencing have no better option than IBM. However, IBM's cloud is up against formidable competition from the three major public clouds (Google, Microsoft, and AWS). In order to compete with them, IBM needs to improve its roadmap by better unifying how customers buy and consume AI infrastructure seamlessly across its ecosystem.

  IBM has strengths in solution ecosystem, configuration, data preparation workloads, locale (cloud and on-premises), and AI development tools. IBM can improve its solution by adding more training workload optimizations and unifying management tools for AI workloads. Reference customers appreciate having the option to run on-premises and/or cloud. They also appreciate the ability to leverage their existing investments in IBM Storage and IBM Z. IBM AI infrastructure is a good fit for existing IBM customers for both on-premises and cloud.

Contenders

- **Alibaba Cloud offers cloud-scale AI infrastructure, but needs to expand its market.** Alibaba Cloud is a public cloud based in China that offers a full complement of cloud services comparable to other global cloud service providers. Alibaba Cloud's AI infrastructure includes access to a broad range of GPU instances. The company also offers development tools to abstract the complexity of running AI workloads. The company's strategy is to offer scalable AI infrastructure that is cost effective and easy to use. Alibaba must expand beyond the APAC region to be attractive to global enterprises that also do business in other regions. Alibaba's vision, innovation, and roadmap can improve if more informed by its peer hyperscalers.

  Alibaba Cloud has a strong current offering with strengths in data preparation, management tools, development tools, fault-tolerance, and efficiency. Alibaba Cloud can improve by investing in even more optimizations for model training and inferencing and by expanding its solution ecosystem. Alibaba reference customers

appreciate its ability to scale AI infrastructure to handle huge AI workloads and the breadth of other services to deploy AI applications. Alibaba is a good fit for existing Alibaba customers and for new customers that need to run cloud-scale AI workloads in China.

- **Lenovo offers AI infrastructure from pocket to cloud, but could expand AI-specific tools.** Lenovo offers computing platforms from mobile (Motorola) and laptops (ThinkPad) to servers (ThinkSystem), edge (ThinkEdge), supercomputers, and everything in between. The company's vision is to offer AI configurations that satisfy everything from inferencing on mobile devices to the most demanding AI research requiring an AI supercomputer, but it lacks a superlative "why Lenovo?" The company could improve innovation by partnering with AI/ML platform vendors that offer AI full-lifecycle tools that are integrated with Lenovo AI infrastructure.

  Lenovo has strengths in solution ecosystem and configuration options and has on-par scores for model training, data, inferencing, fault-tolerance, and efficiency. Lenovo could improve by offering an AI-specific management tool and further optimizations for AI workloads. Reference customers appreciate the ability to run AI workloads on-premises and at scale. Lenovo is a good fit for customers that want to run AI workloads on mobile, on-premises, at the edge, and/or in private cloud.

- **Oracle offers cost effective AI infrastructure but needs more tooling.** Oracle has emerged as an attractive cloud AI infrastructure provider because it has a mature public cloud, a breadth of complimentary AI services, and the hardware horsepower to back it up. In addition, because of its huge enterprise application business, enterprises already have plenty of training data in the Oracle Cloud. Oracle's strategy is to be a cost-effective alternative to the major cloud service providers. Oracle can improve its strategy by greatly expanding its AI infrastructure vision beyond raw GPU instances to why enterprises should consider a long-term, strategic relationship with Oracle Cloud. Oracle can improve its roadmap with enhancements to its AI-specific development tools.

  Oracle has strengths in data preparation and has on-par scores for model training, inferencing, fault-tolerance, and efficiency. Oracle could improve by adding further optimizations for AI workloads. Reference customers appreciate that Oracle offers a very cost competitive AI infrastructure option compared to other cloud providers. Oracle is a good fit for customers needing cost-effective AI horsepower in the cloud and/or already have data in Oracle's cloud.

Challengers

- **Cerebras Systems aims to revolutionize compute for AI, but hasn't done it yet.** Cerebras is a Silicon Valley AI infrastructure startup founded in 2015 to design semiconductors from the ground up for AI workloads. The company's key innovation is a wafer-scale engine which is purportedly the world's largest chip — the size of a pizza. The idea is that a large chip reduces latencies introduced by interconnects and other optimizations. The company uses this chip to build its CS-2 system offered on-premises, in the cloud, and through custom configurations, including clusters in supercomputer configurations. Existing customers include scientific organizations, life science organizations, and others with heavy AI workloads. The company's strategy is to provide the best price/performance for model training. Since Cerebras' focus is on model training, its roadmap relies on partners for data preparation and inferencing workloads.

  Cerebras has strengths in training workloads, workload efficiency, and locale (on-premises and cloud). Cerebras could improve by expanding the type of training workloads it supports, filling gaps with a more robust solution ecosystem, and generally improving across the board to become more on par with other vendors in this evaluation. Reference customers appreciate Cerebras' training performance. Cerebras is a good fit for customers focused on training performance for models supported by Cerebras and those that already have a solid solution for data preparation and inferencing.

- **Graphcore designs for AI, but hasn't displaced GPUs yet.** Graphcore is a UK-based startup founded in 2016 to design semiconductors specifically for AI workloads. The company calls these the intelligence processing units (IPUs). The company's latest chip, the Bow IPU, uses wafer-on-wafer stacking technology that can deliver 350 teraflops of AI compute for some use cases. Graphcore offers Bow PODs for on-premises systems, Bow IPU processors, and partners with some cloud providers to offer IPUs in the cloud. Graphcore's strategy is to innovate AI chip design to create groundbreaking performance, but the company needs to expand its vision, innovation, and roadmap to compete with larger players. Given the dramatic rise in demand for AI infrastructure, the company can accelerate its go-to-market strategy either through partnerships or by investing more in sales and marketing.

  Graphcore is on par with other vendors in this evaluation for model training, inferencing, development tools, efficiency, and locale (on-premises and cloud).

Graphcore can improve by developing more sophisticated management tools and fault-tolerant capabilities and by expanding its solution partner ecosystem to fill capability gaps. Reference customers appreciate the price/performance that Graphcore delivers. Graphcore is a good fit for customers that want to use and/or experiment with new AI chip designs for better price/performance.

Evaluation Overview

We grouped our evaluation criteria into three high-level categories:

- **Current offering.** Each vendor's position on the vertical axis of the Forrester Wave graphic indicates the strength of its current offering. Key criteria for these solutions include solution, workloads, tools, and deployment.
- **Strategy.** Placement on the horizontal axis indicates the strength of the vendors' strategies. We evaluated vision, innovation, roadmap, partner ecosystem, pricing transparency, and supporting services and offerings.
- **Market presence.** Represented by the size of the markers on the graphic, our market presence scores reflect each vendor's revenue and number of customers.

Vendor Inclusion Criteria

Each of the vendors we included in this assessment has:

- **AI infrastructure as identified by Forrester.** Vendors included in this evaluation offer an AI infrastructure solution as defined by Forrester.
- **Comprehensive, differentiated AI infrastructure.** The vendors included in the evaluation offer a solution that is specifically designed for and offered as hardware and/or cloud services to run AI workloads.
- **Market participation.** Each vendor offers a solution that is marketed as AI infrastructure and participates in the market by competing directly with other AI infrastructure vendors.
- **Strong market presence and client interest.** Each vendor has significant interest from our clients in the form of inquiries, advisories, interactions at events, and other conversations.

Supplemental Material

Online Resource

We publish all our Forrester Wave scores and weightings in an Excel file that provides detailed product evaluations and customizable rankings; download this tool by clicking the link at the beginning of this report on Forrester.com. We intend these scores and default weightings to serve only as a starting point and encourage readers to adapt the weightings to fit their individual needs.

The Forrester Wave Methodology

A Forrester Wave is a guide for buyers considering their purchasing options in a technology marketplace. To offer an equitable process for all participants, Forrester follows The Forrester Wave™ Methodology to evaluate participating vendors.

In our review, we conduct primary research to develop a list of vendors to consider for the evaluation. From that initial pool of vendors, we narrow our final list based on the inclusion criteria. We then gather details of product and strategy through a detailed questionnaire, demos/briefings, and customer reference surveys/interviews. We use those inputs, along with the analyst's experience and expertise in the marketplace, to score vendors, using a relative rating system that compares each vendor against the others in the evaluation.

We include the Forrester Wave publishing date (quarter and year clearly in the title of each Forrester Wave report. We evaluated the vendors participating in this Forrester Wave using materials they provided to us by December 20th, 2023 and did not allow additional information after that point. We encourage readers to evaluate how the market and vendor offerings change over time.

In accordance with our vendor review policy, Forrester asks vendors to review our findings prior to publishing to check for accuracy. Vendors marked as nonparticipating vendors in the Forrester Wave graphic met our defined inclusion criteria but declined to participate in or contributed only partially to the evaluation. We score these vendors in accordance with our vendor participation policy and publish their positioning along with those of the participating vendors.

Integrity Policy

We conduct all our research, including Forrester Wave evaluations, in accordance with the integrity policy posted on our website.

FORRESTER®

We help business and technology leaders use customer obsession to accelerate growth.

FORRESTER.COM

Obsessed With Customer Obsession

At Forrester, customer obsession is at the core of everything we do. We're on your side and by your side to help you become more customer obsessed.

Research

Accelerate your impact on the market with a proven path to growth.

- Customer and market dynamics
- Curated tools and frameworks
- Objective advice
- Hands-on guidance

Learn more.

Consulting

Implement modern strategies that align and empower teams.

- In-depth strategic projects
- Webinars, speeches, and workshops
- Custom content

Learn more.

Events

Develop fresh perspectives, draw inspiration from leaders, and network with peers.

- Thought leadership, frameworks, and models
- One-on-ones with peers and analysts
- In-person and virtual experiences

Learn more.

FOLLOW FORRESTER 

Contact Us

Contact Forrester at www.forrester.com/contactus. For information on hard-copy or electronic reprints, please contact your Account Team or reprints@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
Tel: +1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com

ABOUT DELL TECHNOLOGIES

Dell Technologies (NYSE: DELL) helps organizations and individuals build their digital future and transform how they work, live and play. The company provides customers with the industry's broadest and most innovative technology and services portfolio for the data era.