SOLUTION GUIDE

Experiment, prototype and build the future of AI on workstation solutions

Meeting the demands for AI computing resources

Generative AI is bringing profound change across industries, accelerating the adoption of AI-infused technologies at an incredible scale. These new AI-powered workflows offer the promise of new levels of creativity and productivity, improving efficiency across industries.

But they also require significantly more computing power than before. The models used for generative AI are very large, taking weeks or months to train and test on clusters of servers. The results are highly complex models, capable of understanding language, voice, and audio—or creating content such as articles, images, and music, and much more.

95% of executives believe generative AI will compel their organization to modernize its technology architecture[1]. The rush to rapidly add AI computing power to data centers and increase the availability of accelerated cloud instances is straining the availability of hardware. This makes it difficult to meet the continually growing demand.

Key challenges

- **Hardware:** Demand for accelerated AI hardware for data centers and cloud service providers (CSPs) is exceedingly difficult at maintaining cost as opposed to a fix-cost workstation. Current desktop computing resources may not be suitable for AI-augmented workflows.
- **Training and fine-tuning:** AI model size continues to grow, taking months to train and taxing already oversubscribed data center and cloud instance resources. Off-the-shelf models trained on public data require fine-tuning with domain-specific data to provide content relevant to business purposes.
- **Workflow complexity:** Modern professional workflows require running multiple applications simultaneously to maximize productivity. Adding AI-augmented tools and applications puts additional requirements on current computing solutions.

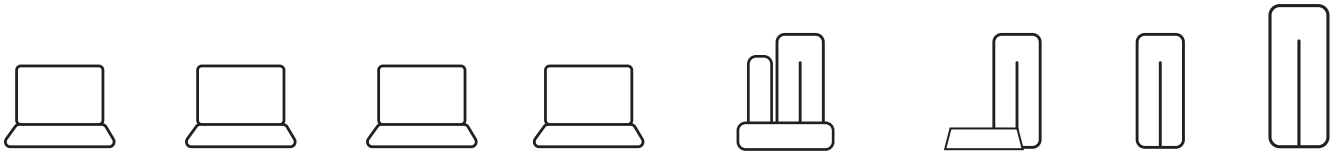
*Please note that several of the products (Z by HP AI Studio GA 2025 waitlist in US/UK, Z by HP Boost GA 2025 in US/UK), and as such, they may not be generally available to all customers at this time. Availability is subject to change, and specific eligibility criteria apply.

We encourage you to reach out to your representative for further details on the products mentioned, including information on how to apply for access to those in the early access phase.

Z AI workstations benefits

- Provide additional AI computing resources to augment data center and cloud instances for development and R&D tasks.
- AI Studio, Z by HP Boost, and a pre-loaded Data Science Software Stack streamline and simplify workflows.
- Large GPU memory configurations enable AI-augmented, multi-application workflows that maximize productivity.
- Enterprise-grade solutions designed to work 24/7/365 and are widely available worldwide.

Recommended Z desktop and mobile workstations for using and developing AI



ZBook Firefly | ZBook Power | ZBook Studio | ZBook Fury | Z2 (Mini, SFF, Tower) | Z4/Z4 Rack | Z6/Z6A | Z8/Z8 Fury

Using AI	• Gen AI/content creation, editing and optimization • Project management and monitoring	• 3D modeling and animation • Data visualization	• Real-time visualizations and simulations • Virtual production, visual effects and composting/finishing • Reality and motion capture
Developing AI	• Exploratory data analysis • Data cleaning and pre-processing • Model development	• End-to-end data processing • Machine learning development • Concept validation	• Advance visualization • Deep learning model training • Computer vision

Ultimate mobility and battery life — Ultimate performance

Z AI workstations powered by NVIDIA RTX

Our comprehensive AI portfolio is two-fold:

- **Differentiated Hardware:** Z AI workstations powered by NVIDIA RTX[1] offer incredible desktop computing performance—perfectly suited for AI training, inference, and data science workflows.
- **Solutions Innovations:** Tools like AI Studio and a pre-loaded Data Science Software Stack streamline workflows and enhance efficiency and productivity.

Together, these robust workstations and solutions are optimized for processing smaller AI models locally with exceptional efficiency, while also providing seamless integration with data center and cloud resources for the development of larger, more complex models. Our most powerful Z desktop AI workstation with up to four NVIDIA RTX 6000 Ada Generation GPUs[2] delivers up to 5,828 AI TOPS for advanced AI. This elegant solution brings data-center-levels of AI computing power to the desktop with ease.

As generative AI tools become integral to professional workflows, the need for powerful GPUs grows. Running large language models (LLMs) like chatbots and code copilots locally on workstations further amplifies these demands. Z AI workstations powered by NVIDIA RTX[2] are designed to handle these intensive workloads.

Key workloads

Z AI workstations for data science

Z AI workstations, equipped with up to four NVIDIA RTX GPUs and substantial GPU memory, are ideal for data science tasks. These powerful workstations reduce latency, facilitating real-time data preprocessing, exploration, visualization, and evaluation of features and models. This saves time and valuable data center and dedicated cloud compute resources. Additionally, our Data Science Stack Manager[3] preloads and automatically updates your workstation with popular open-source software and development tools, allowing your teams to get up and running quickly. It integrates seamlessly with the Z by HP Data Science Software Stack, available on Ubuntu or Windows, and includes open-source software like TensorFlow and PyTorch, as well as developer tools like Git and Visual Studio Code. Furthermore, it supports NVIDIA CUDA®-X libraries, including RAPIDS, a comprehensive open-source collection of GPU-accelerated data science and AI libraries that align with popular open-source data tools. Specifically, RAPIDS cuDF significantly boosts Pandas' performance by up to 110X (compared to CPU-only systems) without any code modifications.

Z AI workstations for using and developing AI

For businesses or individuals starting with AI or using smaller models, Z AI workstations powered by NVIDIA RTX[2] offer a robust and cost-effective solution for AI R&D. These workstations serve as an ideal workspace for developing, evaluating and testing models, augmenting data center servers or cloud resources. In an environment where many competitors are cloud-centric, focusing on on-prem solutions addresses the needs of organizations that prefer or require data processing and AI model development to occur within their own infrastructure for reasons like enhanced security, regulatory compliance, or data sensitivity. Developer teams can fine-tune models as needed and experiment with various datasets and data sizes to optimize results, all while conserving data center or cloud computing resources.

Z AI workstations for inference

Applications with AI-enabled features—Adobe® Photoshop's® Neural Filters, DaVinci Resolve's face tracking, NVIDIA Broadcast's noise and room echo removal, and image denoising in major rendering software—have been available for several years. Generative AI is bringing new levels of capabilities and efficiency to professionals, demanding more computing power and GPU memory. Professionals often work with high-resolution content and use multiple applications simultaneously, which increases these requirements.

As generative AI tools become integral to professional workflows, the need for powerful GPUs grows. Running large language models (LLMs) like chatbots and code copilots locally on workstations further amplifies these demands. Z AI workstations powered by NVIDIA RTX[2] are designed to handle these intensive workloads. The NVIDIA RTX 6000 Ada Generation GPU[2], with 48GB of GPU memory, provides the raw AI computing power and memory necessary to manage high-resolution generative AI content, iterate designs, and integrate with other applications seamlessly, without compromising performance or content fidelity.

Recommended Z desktop and mobile workstations for using AI-enabled applications

ZBook Firefly	ZBook Power | ZBook Studio	ZBook Fury	Z2 (Mini, SFF, Tower)	Z4/Z4 Rack | Z6/Z6A | Z8/Z8 Fury
Up to RTX A500[2]	Up to RTX 3000 Ada[2]	Up to RTX 5000 Ada[2]	Up to RTX 4000 SFF Ada, RTX 4000 Ada, RTX 5000 Ada[2]	Up to 1-4 RTX 6000 Ada[2]

Ultimate mobility and battery life → Ultimate performance

Z by HP AI Studio

US and UK only

Z AI workstations with AI Studio[4] and NVIDIA® NGC™ provide a full-stack solution for AI development. They offer GPU-accelerated models and SDKs to build and deploy solutions across various AI workflows, addressing challenges around trust, security, and scalability.

- Streamline model search and access
- Test and develop directly within AI Studio
- Collaborate remotely with AI Studio

AI Studio's powerful analytics tools provide an overview of available compute resources, helping teams plan and perform operations efficiently. AI Studio ensures consistency in AI development workflows by allowing teams to easily identify the right options, avoiding confusion from different versions of data sets or developer utilities. Its microservices offer optimized model performance with enterprise-grade security, support, and stability.

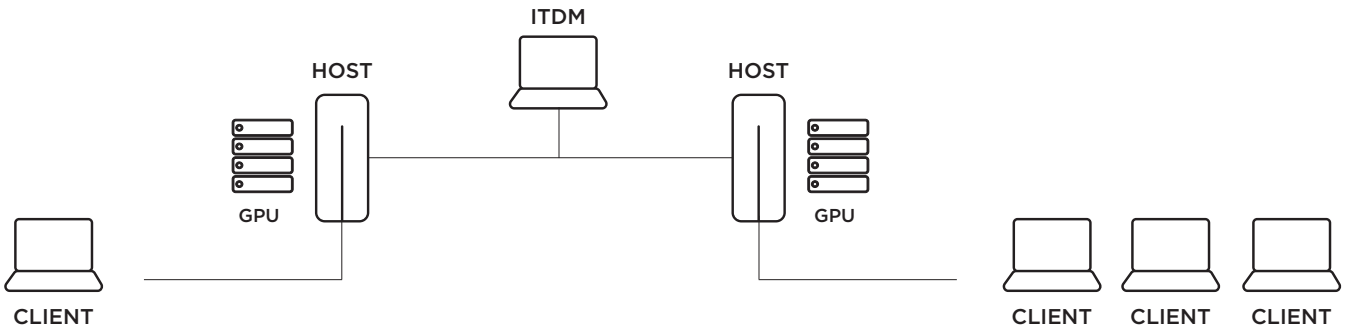
Z by HP Boost

US and UK only

Optimize access to existing and future GPU compute and deliver a performance boost to your Data Scientists and AI developers when they need it. Z by HP Boost facilitates client device access to up to 4 GPUs on a host workstation (must be a Z by HP workstation) via standard network protocol and VPN/encryption. One or more users can access the shared GPUs in a single host through GPU pools. Boost offers many benefits.

- **On-demand performance boost** – choose shared GPUs using a simple drop-down menu when users need them.
- **Accelerate AI innovation** – access GPU power when you need it for faster model development.
- **Efficient cost management** – reduce variable costs associated with cloud compute[6] and make the most of idle GPU compute you own now and in the future.
- **Anytime productivity** – tackle complex AI workflows from nearly anywhere[7] with GPU access through standard network protocols.
- **Securely train AI models** - locally or via VPN/encryption in an individual user session.

On-demand GPU compute, turning your workstation into a sharable resource



Enterprise-ready solutions

Z AI workstations powered by NVIDIA RTX[2] are designed and built for demanding enterprise deployments. Powered by the latest generation of workstation CPUs, Z AI workstations are ready to tackle demanding AI development workflows. The latest generation of Z by HP desktop and mobile AI workstations are available now and ready to ship.

With a full stack of enterprise-level deployment, support, and optimization tools, Z AI workstations easily fit into existing IT infrastructure, providing drop-in solutions for AI training, developing, deploying, and inferencing on the desktop through Z by HP AI Studio[4]. The NVIDIA GPU[2] architecture scales with the use of Z by HP Boost[5] from cloud to data center, desktop, mobile, and embedded device, supporting the same software stack across devices, which enables users to move workloads seamlessly between them.

Ready to get started?

To learn more about the Z AI workstations, visit:
hp.com/z

To learn more about NVIDIA RTX, visit:
nvidia.com/ai-workstations

Contact sales at **hp.com/us-en/workstations/z-support.html**

1. Source: https://www.accenture.com/us-en/blogs/blogs-careers/the-future-making-it-human
2. Products sold separately and must be configured at purchase.
3. Z by HP Data Science Stack Manager requires Windows 10 version 21H2 (Build 19044) and higher or 64-bit Ubuntu 20.04 and is available on select Z workstations.
4. Z by HP AI Studio requires Windows 10 (build 19041 and higher) or Windows 11 OS, Intel Core™ i5 12th gen or higher, AMD Ryzen™ 9 or higher processors, minimum 50 GB of available storage, 16GB of RAM, and Internet access. To enable GPU compute, Z by HP AI Studio requires any NVIDIA® GPU compatible with driver version 528.89 or newer. A minimum of 8GB of VRAM is recommended. Join waitlist for access to Z by HP AI Studio early access program.
5. Z by HP Boost requires client device with Windows 10 or Windows 11 OS and Vulkan Runtime for devices without a GPU, or Ubuntu 22.04 LT or greater, Debian 11 or greater with libatomic1 or libnuma1 packages installed. Host device for sharing GPU(s) must be a Z by HP workstation with NVIDIA professional grade workstation GPU(s) of Pascal architecture of later with an installed driver version 535 or greater, and a CUDA driver version 11.2 through 12.3, Windows 10 or Windows 11 OS, or Ubuntu 22.04 LT or greater or Debian 11 or greater with libatomic1, libnuma1, libvulkan1, libgl1, libgilb2.0-0 packages installed. Network connection required. Join waitlist to be considered for selection for early access.
6. Cost reduction of Z by HP Boost relative to variable costs with cloud compute.
7. Network connection required.