

the
GORILLA
GUIDE[®] to...



Creating Business Value with Generative AI

Key Considerations When Advancing GenAI in Your Business

ED TITTEL

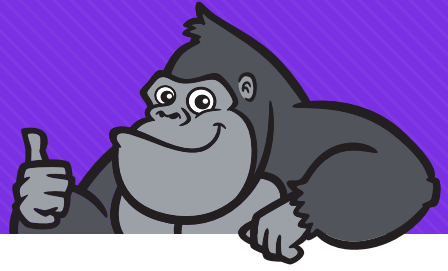


Hewlett Packard
Enterprise

intel[®]

POWERED BY  **ActualTech**
MEDIA

the
**GORILLA
GUIDE[®]** to...



Creating Business Value with Generative AI

Key Considerations When
Advancing GenAI in Your Business

By Ed Tittel

POWERED BY  **ActualTech**
MEDIA

Copyright © 2024 by Future US LLC
Full 7th Floor
130 West 42nd Street
New York, NY 10036

All rights reserved. This book or any portion thereof may not be reproduced or used in any manner whatsoever without the express written permission of the publisher except for the use of brief quotations in a book review. Printed in the United States of America.

www.actualtechmedia.com

PUBLISHER'S ACKNOWLEDGEMENTS

DIRECTOR OF CONTENT DELIVERY

Wendy Hernandez

GRAPHIC DESIGNER

Olivia Thomson

HEAD OF SMARTSTUDIO

Katie Mohr

WITH SPECIAL CONTRIBUTIONS FROM HPE

Robert Checketts

SR. MANAGER, COMPUTE PRODUCT
MARKETING, HPE

Andy DeBernardis, Worldwide
Product Marketing

AI SOLUTIONS, HPE COMPUTE

Aaron Lamond, Worldwide
Product Marketing

AI SOLUTIONS, HPE COMPUTE

ABOUT THE AUTHOR

Ed Tittel is a 30-plus year veteran of the IT industry who writes regularly about cloud computing, networking, security, and Windows topics. Perhaps best known as the creator of the *Exam Cram* series of certification prep books in the late 1990s, Ed writes and blogs regularly for GoCertify.com, TechTarget, ComputerWorld, TekkiGurus, and other sites. For more information about Ed, including a resume and list of publications, please visit EdTittel.com.

ENTERING THE JUNGLE

- Introduction: Generative AI’s Incredible Opportunity and Challenge** **6**

- Chapter 1: Initial Considerations for Business Use of GenAI** **8**
 - Making GenAI Work for Your Business 10
 - Common Use Cases for GenAI 11
 - Why HPE and Intel Excel at GenAI Workloads 13

- Chapter 2: Getting Value from GenAI** **15**
 - Harness GenAI and Make It Your Own 17
 - Are You Ready to Benefit from GenAI? 18
 - Other Key HPE ProLiant Gen11 Factors: Business-Critical for Today’s Enterprise Customers 20
 - Why HPE and Intel Can Help Your Advance Your Initial GenAI Moves 22
 - How HPE ProLiant GenAI Solutions Meet Challenges 23

- Chapter 3: Build Your First GenAI App with HPE ProLiant Gen11 Intel Xeon Servers** **25**
 - Choosing a First GenAI Application 26
 - Corresponding HPE ProLiant Gen11/12 Servers and GPU Options 28
 - HPE Services Does AI for Customers 31

CALLOUTS USED IN THIS BOOK



SCHOOL HOUSE

In this callout, you'll gain insight into topics that may be outside the main subject but are still important.



FOOD FOR THOUGHT

This is a special place where you can learn a bit more about ancillary topics presented in the book.



BRIGHT IDEA

When we have a great thought, we express them through a series of grunts in the Bright Idea section.



DEEP DIVE

Takes you into the deep, dark depths of a particular topic.



EXECUTIVE CORNER

Discusses items of strategic interest to business leaders.



DEFINITION

Defines a word, phrase, or concept.



GPS

We'll help you navigate your knowledge to the right place.



KNOWLEDGE CHECK

Tests your knowledge of what you've read.



WATCH OUT!

Make sure you read this so you don't make a critical error!



PAY ATTENTION

We want to make sure you see this!



TIP

A helpful piece of advice based on what you've read.

INTRODUCTION

Generative AI's Incredible Opportunity and Challenge

Welcome to The Gorilla Guide To...[®] Creating Business Value with Generative AI.

There's no doubt that artificial intelligence (AI) is taking the technology world by storm. Indeed, AI [software spending](#) is expected to double from \$33 billion in 2022 to \$64 billion by 2025, with AI-related [overall investment](#) expected to jump from \$156 billion in 2022 to as much as \$275 billion by 2025. Of that amount, spending on AI-related infrastructure is jumping from over half of that total, to an amazing \$300-plus billion by 2031 (a CAGR of nearly 30%). What the fuss is all about, of course, is that AI is already delivering on its promised impacts on business, which include faster and better innovation, increased flexibility and insights, and a shorter time-to-value or time-to-market when turning ideas into cash flow. This has spawned a market efflorescence that [The Economist](#) says has not been seen since the dot-com era.

But AI comes with a unique and formidable set of challenges. On the one hand, organizations need to identify applications or services that can benefit from the insights and content that AI can generate on their behalf. This requires careful observation of what's already

available and what's most urgently needed, a profound understanding of the underlying data and an appreciation of where value can be added, productivity gained, and time-to-market accelerated. On the other hand, incorporating and integrating Generative AI (GenAI) means adopting one or more technology stacks, identifying and preparing data for use in training and refining AI models, and standing up applications to take advantage of the power and potency of GenAI in useful and effective ways. Those are big tasks for organizations to tackle and solve so they, too, can put GenAI to work—preferably sooner, rather than later.

In this Gorilla Guide, you should find the information you need to better understand the values and costs of your own adoption, creation, and piloting of GenAI technologies. This begins with an assessment of impacts and common use cases for GenAI, along with an explanation of how HPE and Intel are uniquely suited to handle such workloads. Next comes an exploration of the value and benefits of deploying GenAI, and how HPE and Intel solutions help speed adoption and time-to-value. Finally, there's a discussion that explains the various HPE and Intel options available at the edge and in the data center, with services and consultation to match and how HPE ProLiant Gen11 servers powered by Intel 5th Gen Xeon processors can help you make those things happen without disrupting or delaying opportunities to improve and grow your business. Chapter 1 starts you down this trail with an explanation of how GenAI can impact your business from the edge to the data center, thanks to a nuanced understanding of what it can do, and how to make it work for you.

CHAPTER 1

Initial Considerations for Business Use of GenAI

Leaving aside the breathless excitement and universal enthusiasm surrounding AI, it's important to understand that running GenAI applications represent a different kind and class of computing workload than many other typical IT tasks. The characteristics of GenAI workloads are tough and demanding, with a special emphasis on compute. Indeed, GenAI workloads require significant processing power to train models and generate content. Large models such as Llama3 or GPT-4o run best on specialized hardware designed to offer them the best possible speed and efficiency. Other aspects of GenAI workloads include the following:

- **Memory consumptive:** AI workloads demand substantial memory where local, fast stores for model parameters and intermediate values eat large amounts of RAM. Certain transformation-based models (GPT-4 or GPT-4o) may use hundreds of millions to billions of values: this requires substantial memory, storage, and bandwidth at great scale to accommodate heavy access to huge datasets.

- **Data dependencies:** AI models, especially those for generative AI (aka GenAI: see “Generative AI (GenAI) Defined” on p. 4) rely on reams of high-resolution data, with many processing steps devoted to data preparation to drive model training.
- **Latency limits:** Certain AI workloads, such as chatbots or voice assistants, impose strict latency requirements for acceptable user experiences. Optimizing AI models to speed response is vital, and may require careful workload placement to minimize latency.
- **Overcoming complexity:** AI workloads (especially GenAI) can be complex, and may require specialized software, hardware, and expertise for best results. This applies especially to applications like computer vision and natural language processing (NLP).

Thus, there are stringent server requirements at work for the brave new work of AI workloads. A flexible, capable architecture that works well for data- and compute-intensive tasks, even for the most demanding workloads, is what enables GenAI to deliver real value. As you’ll learn in the following sections, the [Intel 5th Gen Xeon powered](#) HPE ProLiant Gen11 server family covers all those bases, and meets those foregoing needs, with dispatch and careful design—and does so at the edge, in the cloud, and in the data center.

Many businesses (and business owners) hang onto legacy systems longer than they should. In fact, keeping outdated technology in service can fall short in providing necessary services and innovation, and reduce productivity and business value. That’s how the benefits of replacing outmoded technology more than offset the risks involved in retaining legacy systems.

Making GenAI Work for Your Business

GenAI provides methods to create—that is, to generate—new high-quality, high-level content in response to user requests (also called “prompts” or “queries”). Indeed, GenAI facilitates to produce usable text, images, audio, video, animation and simulations already exist, and keep getting better. Businesses can deploy this kind of content on demand for a variety of purposes from training, to customer service, to sales support, to automate routine IT and other departmental tasks, and more. Such uses can boost productivity, and free up human resources for high-value tasks (creating new offerings, opportunities, markets, and more).

Training a model involves determining relationships between values in training data sets, and recording or encoding those determinations to establish the model’s parameters and behaviors. Inference

Generative AI (GenAI) Defined

GenAI employs a model that can create text, images, sound, videos, or other data using machine learning to gain its intelligence from training on huge volumes of the kinds of data it’ll ultimately be asked to emit. GenAI models learn patterns, structure, and characteristics from input training data and then create new data with similar elements in response to user queries or prompts. This can be text, images, video, or speech, depending on the training data and intended uses that GenAI models seek to support.



occurs when a model generalizes from such parameters and behaviors to interpret new data it receives for inspection and analysis. The quality of that data, the interpretation and insights that GenAI supplies, also enables GenAI to add business value.

When an AI model is handling input data, it processes those inputs as they arrive in real time. A user query is compared with the information it has established during the training phase for its parameters and behaviors. Responses depend on the tasks involved, where the goal is to calculate and deliver actionable insights or results.

Because inferencing can involve expenditure of significant resources (compute, storage, power, cooling, time, and so forth), optimizing speed and efficiency during inferencing helps ensure the best user experiences and minimizes costs and consumption. This goes straight to the bottom line, and provides improvements in speed, quality, and accuracy that add value and productivity. That explains the relentless push in AI systems to speed up inferencing, reduce costs, and find efficiencies at all levels (hardware, platforms and frameworks, and application software).

Common Use Cases for GenAI

Businesses and enterprises are increasingly using GenAI to build applications, including chatbots, digital assistants, analysis and prediction tools, and developer aids of all kinds. While the possibilities are many and incredibly varied, common use cases for GenAI-based models include the following:

- **Customer service and support:** Because AI agents can listen to voice inputs, understand the underlying needs and wants that drive them, and recommend products or services, or help solve problems, enterprises can put such chatbots to ready and productive use for employees, customers, and partners. The more routine

or tedious work comes off the shoulders of support staff, freeing them up to deal with more valuable, interesting, and useful activities.

- **Employee, user, or customer information navigators:** Generative AI can assist humans in finding information about benefits, insurance, and other complex offerings among which they must (or want to) choose. It can help them find and understand information resources, get better results from systems and tools, manage selection and deployment of programs, and more.
- **Creative outputs and ideas:** Generative AI helps content and other creators formulate new concepts, designs, images or videos, and all kinds of text content, based on textual descriptions (“prompts”) that draw from training data, models, and ongoing user interactions. Creators can be more productive as they can benefit from AI-based help at every step in the creative process from brainstorming, to creation, to refinement, to release and feedback collection.
- **Data analysis and interpretation:** Generative AI excels at assembling and synthesizing data to understand customer or user behavior, profiles and needs to generate content or insights for better marketing, communication, and services delivery. It holds out the possibility of fabulous user or customer experiences based on a real understanding of general and narrowly focused needs, wants and behaviors.
- **Code creation:** Generative AI helps automate coding tasks based on deep understanding of existing examples and models, common workflows across the development lifecycle, and extremely flexible, customized code that fits an organization’s security, governance, and compliance requirements and internalizes business goals and ethics.

The HPE ProLiant Gen11 server family uses the latest 5th Gen Intel Xeon processors. Each of the foregoing use cases draws on advanced AI inferencing, analysis and outputs, actions or recommendations designed to deliver optimal and speedy performance that's also energy-efficient.

Why HPE and Intel Excel at GenAI Workloads

When it comes to GenAI computing, more is not simply better—abundant but affordable resources and capability are essential for providing positive user experiences and manageable TCO. HPE ProLiant Gen11 servers deliver more and faster graphical capabilities, enabling organizations to innovate using the advanced GPU accelerators and the ultra-scalable architecture that GenAI workloads need. By deliberate design, the HPE ProLiant Gen11 server family powered by Intel supports up to one-third higher GPU density, with increased flexibility for GenAI workloads at the edge (chatbots, data analysis for IoT and other local resources, loss prevention based on video surveillance, and other latency-sensitive uses) and in the data center (model creation and training, coding, creative support, and more).

Organizations that deploy HPE ProLiant Gen11 compute also benefit from a more capable resulting IT infrastructure. These servers bring higher efficiency, improved scalability, and better economics to speed and enhance business outcomes while lowering TCO. Higher density also means more workloads run in the same rack-space (server footprint) for improved ROI and better performance for GenAI workloads.

Using HPE ProLiant Gen11 servers with Intel 5th Gen Xeon processors also helps organizations control costs through rightsizing and scaling on-demand. Indeed, GenAI workloads run faster on HPE ProLiant Gen11 clusters with predictable, completely visible costs.

You can scale capacity up or down on-demand—using a next-generation architecture with continuous monitoring—to right-size or even add burst capacity on-site. All this adds up to improved GenAI performance and capability.

Even better, HPE makes the operating experience intuitive (just like the cloud). You can simply use a single set of controls from edge to cloud via the Intel powered HPE ProLiant Gen11 cloud computing experience. This enables fully digital business operations and transformation, with global visibility and insight through a singled unified console with HPE GreenLake for Compute Ops Management. Organizations can swiftly and painlessly automate tasks for efficient deployment, simplified support and lifecycle management. This helps GenAI workloads integrate seamlessly into existing enterprise operations.

Finally, the HPE ProLiant Gen11 Server family with the latest Intel Xeon processors embodies Trusted Security by Design. It is built to protect your infrastructure, workloads, and data from threats to hardware and third-party software. It does this through a trusted edge-to-cloud security approach that's based on HPE's silicon root of trust, trusted supply chain, zero trust hardware and software technologies. That means reduced security risks and exposures, and better data protection, for your advanced and data-heavy GenAI workloads.

In the next chapter, we switch gears from surveying the business considerations and requirements for GenAI to take a look at how businesses and organizations can extract value from using GenAI.

CHAPTER 2

Getting Value from GenAI

GenAI has been setting the headlines ablaze, as well as firing a multitude of business strategies and projects. Organizations, both public and private, are seeking earnestly to get ahead of the curve when it comes to GenAI. For those not already in the know, GenAI is an area of artificial intelligence (AI) that uses leading-edge machine learning (ML) techniques to generate content or data. What gets generated depends on the specific training data that drives ML. These days, that includes text, images, animation, 3D models, video and more. From a carefully curated and managed collection of training data meant to model some reality sufficiently to make it accurate and convincing, generated content is meant to be useful, engaging, and meaningful.

Compute matters tremendously for making GenAI successful, because the models that drive it are huge and complex. This requires enormous amounts of compute horsepower, when a model may include millions of parameters (or more) and involve billions upon billions of data objects for training up front. Then, when a model is determined to be sufficiently accurate and representative of reality, it ingests even larger volumes of real-world data to generate content in keeping with its carefully trained-up view of reality. That view

keeps changing and getting refined over time as it becomes more accurate and informed by the data it handles and creates, and as the input stream changes to reflect changes in the real world.

GenAI models come with a variety of value-adds they can bring to organizations that deploy them. These include the following and are closely tied to the kind of data used for training and expected by way of output:

- **Automated content creation:** With the right training data to drive it, GenAI can produce text, graphics, videos animations, and more. Be it for summaries, reports, help files, and so forth, AI can handle routine or repetitive tasks that opens an opportunity for people to be more productive and pursue innovation. Likewise, GenAI can produce images, clip art, diagrams and timelines, plus other visual media that saves human time and effort.
- **Improved customer support:** GenAI drives chatbots that can use customer history (individual and in the aggregate) to provide status, answer questions, and speed customers through quality interactions and encounters. Human agents can focus on complex topics, solve problems, and build better customer relationships.
- **Optimize workflows:** GenAI can help with automating data entry, assembling more efficient delivery routes, organizing pick and pack assignments, and tracking task progress and completion. This is how GenAI-driven solutions can reduce manual effort, steer faster completions, and boost productivity.
- **Drive innovation:** GenAI offers insights and recommendations that can do more than optimize and improve existing processes and methods. From the huge volumes of data it ingests, GenAI can identify patterns to suggest new approaches to existing products or

services and cultivate new ideas. These can lead to new products (or new product categories), suggest marketing strategies, and help elaborate new business models.

Judicious application and careful use of GenAI can provide organizations with better ways to do what they already do, and help them find, develop, and deliver innovations and new opportunities.

The HPE ProLiant Gen11 server family has solutions engineered specifically for GenAI, at the edge and in the data center, and in the cloud. It offers breakthrough performance, and ultra-scalable architectures with more graphics capabilities than ever before available. This open-ended architecture lets you grow your business at scale with adherence to industry standards, with the flexibility to incorporate and accommodate new technologies, and innovate wherever your apps and data live.

Harness GenAI and Make It Your Own

There's a big learning curve involved in putting GenAI to work in an enterprise or organization. That's why HPE's approach is to help enterprises to be able to better take advantage of a number of common use cases for which GenAI solutions are ready to jump right into AI inferencing. A pre-configured solution saves developers the time and effort involved by utilizing a pretrained model (e.g., Llama-3, Nemotron-3, Mistral, and other open source models) with ready-to-deploy toolsets, frameworks, and optimized infrastructure with necessary GPU performance. This enables enterprises to skip the training phases which is considered computational-intensive and costly while enabling them to focus on fine-tuning and utilizing RAG with inference to apply domain knowledge for GenAI applications (e.g., edge- or data center-oriented).

HPE also provides a broad range of server platform choices that include 5th Gen Intel Xeon processors that range from pay as you go offerings to standalone rack-mounted 1U and 2U configurations suitable for on-premises data centers or edge deployments. Indeed, all these can work together seamlessly through managed services and capacity—including scale-up and scale-out—to support a demand-driven, consumption-based payment model for excess or peak demand capacity. This comes in the context of a single, coherent management process that handles assets and resources alike from edge to data center to cloud, as needed. Then, organizations can take the fullest possible advantage of what HPE and Intel have to offer by way of servers, services, management, control, and flexibility.

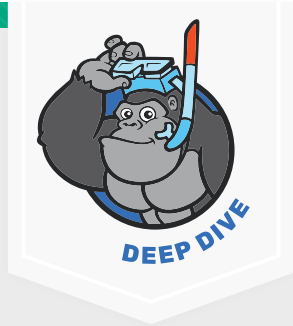
Are You Ready to Benefit from GenAI?

Indeed, organizations must explore and evaluate infrastructure options to get the most out of generative AI deployments. In practice this means identifying and piloting cases at the edge, in the data center, and in the cloud. Some deployments make most sense in the data center for reasons related to security, IP protection, and sensitivity of data, algorithms, or models.

Other situations may require edge deployments to optimize end-user experience by minimizing latency and response times: This is very much the case for computer vision where GenAI uses live security camera feeds as the basis for actions, notifications, or evidence collection. This gives them the flexibility and capacity to accommodate wide swings in uptake and use, or geographical flexibility when latency and customer/user proximity are both important (e.g., in a “follow the sun” pattern that tracks daylight hours around the globe).

AI/GenAI Are Compute Intensive!

Though it might seem that complex models trained up on reams and reams of data is the only way to practice AI, it's just one sub-field in a large, complex undertaking that seeks to use computers to build representations of things so they can be understood, changed around, refined, or extended. But no matter where AI pops up, count on it to be resource-intensive, for compute. That's why you'll find the most demanding GenAI applications running in data center server clusters all the way up to the biggest high-performance supercomputers.



Generative AI is both interesting and challenging because it requires understanding AI workloads and how best to handle them. GenAI in particular, and AI inferencing in general, can require tons of research, trial deployments, and mid-course adjustments to get things “just right.” For organizations on the fast track, it's worth noting that HPE's various GenAI solutions ship ready-to-run on a variety of HPE ProLiant Gen11 servers with 5th Gen Intel Xeon CPUs and appropriate NVIDIA edge- or data center-oriented GPUs. They're also pre-trained for specific use cases, too, so ramp-up and training times are seriously shortened.

In particular, organizations are likely to face certain specific challenges when working toward GenAI deployments:

- 1. Leveraging private data securely:** Intel-powered HPE ProLiant Gen11 takes a zero trust approach from the silicon level all the way up to applications and services under the GreenLake Compute Ops Management

umbrella, backed up with a secure supply chain and a never-trust/always-verify approach to firmware, access controls, and more.

- 2. Optimize infrastructure for maximum AI advantage:** this means providing high performance as and when it's needed, with the ability to manage latency and meet stringent customer experience and SLA requirements.
- 3. Stand up and run infrastructures quickly and securely:** HPE's pay as you go model means management comes as a natural consequence of accessing and using its tools: there's no set-up, no staging, and no added complexity. Hook it up and put it to work!
- 4. Organizations can work from the pre-trained models** to shorten the development cycle with optimized tools, platforms, and so forth. HPE's pre-defined use cases represent real value for organizations, with a shorter, smoother time-to-value backed up by expert consulting services to further shorten that interval.

Other Key HPE ProLiant Gen11 Factors: Business-Critical for Today's Enterprise Customers

Beyond its robust support for GenAI and a set of specific use cases, HPE ProLiant Gen11 servers powered by Intel all share other important characteristics in common. Broadly speaking, these fall under the headings of secure lifecycle management and “as-a-service” offerings based on HPE GreenLake Flex Solutions. Each of these is discussed here:

- HPE ProLiant Gen11 employs **security by design** that extends all the way from a silicon root of trust, with firmware and controller checks at start-up and runtime

to make sure systems remain tamper-free. Even construction and delivery are covered, thanks to the company's secure supply chain, available globally through the [HPE Server Security Optimized Service for HPE ProLiant](#)). In fact, HPE puts zero trust to constant use as well, so that all access requests are constantly checked and verified ("never trust—always verify"). Through HPE GreenLake and its Compute Ops Management (COM) facility, HPE provides security protection for software at the OS, platform, and application layers, with APIs to provide organizations and partner with direct, secure access as well.

- **Lifecycle management** comes to HPE ProLiant Gen11 servers, courtesy of various consoles and management tools. These include the server-at-a-time Integrated Lights-Out (iLO) management facility, but also HP OneView and especially the veritable "single pane of glass" that HPE GreenLake COM provides for servers at the edge and in the data center ... and beyond.
- **Pay as you go offerings** via HPE GreenLake Flex Solutions bring organizations a veritable cornucopia of cloud-based capability. Thus, they can extend their own platforms at the edge and in the data center using consumption-based IT solutions built upon standardized, centrally managed IT modules that create extensible IT infrastructure building blocks in the cloud. Consumption-based pricing, plus predictability and reliability, gives organizations what they need to scale out GenAI solutions as and when they need extra capacity or capability.

Organizations can put Intel-powered HPE ProLiant Gen11 to work in either centralized data centers or distributed edge locations with great results and simple, unified, automated management (through GreenLake COM). This works especially well for cloud-native

workloads, so organization can scale up and scale out, take advantage of infrastructure as code, and achieve agility in fast-moving, fast-changing usage situations.

Why HPE and Intel Can Help Your Advance Your Initial GenAI Moves

HPE ProLiant Gen11 servers with Intel 5th Gen Xeon processors are engineered for peak performance especially for compute-intensive applications like GenAI. That means advanced graphics rendering, data acceleration, large language models, computer vision, and speech and language processing all work well on such servers. They can bring compelling economic returns to organizations, whether GenAI apps run in the data center or at the edge.

HPE's other inferencing solutions include specific elements tailored for different AI applications, with Intel Xeon Scalable processors and NVIDIA edge or data center GPUs to handle such workloads. Available offerings focus on:

- Computer vision AI at the edge, in a compact design for up to 4 edge-oriented GPUs, plus pre-packaged applications for working with cameras in real time for loss-prevention, smart spaces, and safety.
- Natural Language Processing (NLP) provides optimized support for applications that include speech AI and conversational AI, also supported through data center-oriented GPUs, coupled with other optimized HPE ProLiant Gen11 servers with Intel Xeon processors.

Overall, HPE ProLiant Gen11 servers with Intel 5th Gen Xeon processors offer wide-ranging, highly capable and efficient solutions for enterprises seeking to stake out their spot in the arena of GenAI (and similar applications). With enhanced security, reliability, and

industry-leading performance as well, these servers excel at handling demanding, compute-intensive AI workloads. Dig deeper into the economics, sustainability, energy efficiency, and supply chain, and you'll understand how HPE ProLiant Gen11 Intel-powered servers enable organizations to remain agile, manage IT challenges, and focus on long-term innovation and market development.

How HPE ProLiant GenAI Solutions Meet Challenges



By design, HPE ProLiant Gen11 solutions with Intel 5th Gen Xeon CPUs for GenAI are carefully designed. They're built to leverage proprietary data so organizations can obtain unique and lasting value from that data. GenAI solutions will help organizations obtain and use the insights, predictions, or recommendations that get generated.

HPE ProLiant Gen11 solutions, including HPE GreenLake and its COM, lets an organization own and control its entire AI compute stack from end to send. That makes HPE ProLiant Gen11 AI solutions tuned, secured, and certified for edge or data center deployments, where the organization owns and runs everything. Expert consulting and management are available from HPE to assist (or take over) as circumstances dictate. This gives the organization complete control over its own IP.

HPE ProLiant Gen11 servers with Intel 5th Gen Xeon CPUs and NVIDIA GPUs are purpose-built and certified to support an ultra-scalable, ultra-efficient, and ultra-secure platform for AI workloads, across a broad range of use cases. This ensures organizations get the performance they need to keep users happy and provide a worthwhile return on their technology investments.

HPE ProLiant Gen11 offers AI solutions that are tested and optimized by HPE and Intel and their use case partners. Three families of use cases help eliminate the cost and risk typically needed during the training phase, enabling enterprises to fast track computer vision, GenAI, and natural language processing. This means organizations spend less time researching, sifting through alternatives, and building up their GenAI environments. That translates into shorter time-to-value and a quicker payoff on the buy-in.

Finally, HPE ProLiant Gen11 with Intel 5th Gen Xeon CPUs can transform the operating model for IT with HPE GreenLake. This lets organizations access advanced HPE ProLiant servers in a cloud-like, consumption-based environment. In turn, this adds the advantages of on-premises devices to the flexibility, scalability, and accessibility of the cloud. A service-focused approach supports global management at the edge, in the data center, and in the cloud with pay-as-you-go pricing to manage costs and ever-shifting demand. Organizations can use as much or as little technology as they need, and make sure it pays for itself to meet financial targets.

In the next chapter, we'll walk through the steps and considerations involved when evaluating and prioritizing pilot GenAI projects. We'll also explain how HPE and Intel can help with hardware, software, and consulting services to make pilots happen in short and affordable order.

CHAPTER 3

Build Your First GenAI App with HPE ProLiant Gen11 Intel Xeon Servers

As we've already mentioned, GenAI is reshaping the way enterprises think, innovate, and work across the entire business landscape. Enterprises of all kinds and all sizes are customizing large language models with their own data to support a wide range of GenAI applications from new customer experiences to enhancing employee productivity. This enables them to better handle their unique business needs and situations and empowers AI applications across all aspects of their operations—not just in IT, but in procurement and supply, HR, sales and marketing, manufacturing and logistics, and more.

Building a first GenAI application comes with plenty of interesting challenges. IT operations must learn how to work with hardware and software, specific technology stacks and various frameworks, platforms, and models to support AI operations. HPE ProLiant servers with Intel 5th Gen Xeon processors provide an efficient and high performant foundation that enables enterprises to start small and grow large while maintaining control over data and costs.

Choosing a First GenAI Application

Some organizations start modestly, with one or perhaps two GenAI projects to get the ball rolling. This gives developers, operations, and stakeholders a chance to learn about the technology and what it can do, hopefully through some easy wins to gain interest in and support for further, more significant efforts.

Given the right strategy and approach, research¹ shows that a pilot can launch in as little as a calendar quarter (give or take a month). That strategy can be outlined in the form of a six-step plan, as follows:

- 1. Identify initial use cases:** Look for repetitive, data-driven, or predictable step-oriented tasks like content creation, summarizations and task automation. These are most likely to benefit from applying GenAI to speed completion and boost productivity. Try to start with something already known, where you can work to see measurable results
- 2. Enlist stakeholder support:** Team up with a department or another sponsor to find an AI project to make somebody's job easier. It might be automating a manual task, or combining multiple individual tasks done in sequence into a single task. Work with your partners to agree on what to deliver, then do just that. Get executive sponsorship and buy-in to make sure your project goes forward, fits the overall digital strategy, and is recognized for value.

¹ Gartner Report: [How to Pilot Generative AI](#), July 10, 2023

- 3. Gather and prepare quality data:** Data is what drives AI and determines its relevance and usefulness. Before building any GenAI model, you should collect and groom data to inform the model for proper quality, accuracy, and quantity. Look at the process, document incoming data flows, identify people involved, and learn which systems, applications, and processes touch that data across your organization. This helps illuminate the data lifecycle, and where and how it's stored.
- 4. Find SMEs:** Subject matter experts (SMEs) are the people who regularly work with and best understand the data and operations that you're trying to automate or improve. They're the ones who really know what's going on. Typically, they'll be highly competent and busy. Indeed, only a single data scientist or GenAI expert is plenty for pilots as they work with IT staff/developers and SMEs to define and prioritize use cases, prepare data, build and train simple models, and get ready to pilot their efforts. HPE consulting services can help with staffing if you have no local expert on tap.
- 5. Operationalize a GenAI model:** When the preceding steps are done it's time to deploy into production. When this happens, elements to monitor, manage, govern, and analyze the model's results are needed to make sure it's working and doing what it needs to do. Early on, a full-blown management system may not be necessary (though it will be required once you go past the pilot stage). Instead, get your experts to manage the model from end to end. Make sure it's up-to-date, check that there's no model drift, make sure that responses or results make sense and make a (helpful) difference. This shows an understanding of AI, exposes its value, and makes it has genuine value.

- 6. Select your partner: HPE.** Once you've built a pilot or three and get things into production, you will better understand the systems, workflows, software, and data that models need. This means it's now time to find products and platforms to help meet your GenAI needs for the years ahead (typical planning horizons for GenAI run three to five years). This helps speed model development and training into production. HPE can also provide a unified data foundation that spans hybrid cloud environments from the edge to the data center and into public and private clouds. HPE even offers AI and data transformation services to help companies get started on this journey and find their way to successful outcomes.

Corresponding HPE ProLiant Gen11/12 Servers and GPU Options

Intel and HPE have partnered up to architect specific solutions platforms for typical cases at the edge and in the data center: Each of two possible server packages is described in the sections that follow; both fall under HPE GreenLake for Compute Operations Management (COM) for simplified and automated control of your compute infrastructure in the data center or at the edge.

HPE PROLIANT DL320 GEN11 SERVER FOR AI AT THE EDGE

This 1U single socket server offers the kind of compact, energy efficient yet capable server package that organizations need to position themselves at the network edge for a class of applications where minimizing latency is key. These include loss prevention, site security and surveillance, IoT access, factory floor automation, and more.

Key elements of the DL320 architecture include:

- 5th Gen Intel Xeon processor
- Up to 2TB DDR5 RAM
- Up to 10 small form factor SSDs for solid state storage
- Either two double-wide or 4 single-wide GPUs for AI workloads, i.e. up to 4 NVIDIA L4 GPUs
- Choice of networking interfaces (ask your HPE Partner or sales rep for details)

HPE PROLIANT DL380A GEN11 SERVER FOR THE DATA CENTER

This 2U dual socket server offers support for up to two high-end 5th Gen Intel Xeon scalable processors. These rack-mounted servers are ideal for data center deployment and offer an ideal combination of small footprint and energy efficiency for maximum use of cubic volume. They are ideal for resilient, available and capable access to AI, compute and graphics-intensive workloads.

Key elements of the HPE ProLiant DL380a Gen11 architecture include:

- 5th Gen Intel Xeon scalable processors
- Up to 3TB DDR5 RAM
- Up to 8 EDSFF E3.S 1TB SSDs
- Up to 4 double-wide or 8 single-wide GPUs for the most demanding AI workloads, i.e. up to 4 NVIDIA H100 or L40S GPUs
- 4-port GbE OCP3 adapter or 10GbE 2-port adapter (other options available in some markets—ask your HPE Partner or sales rep for details)

HPE PROLIANT COMPUTE DL380A GEN12 FOR THE DATA CENTER

This 4U dual socket server offers support for up to two ultra-efficient Intel Xeon 6 processors, plus up to eight NVIDIA H200 NVL Tensor Core GPUs. These scalable, rack mounted servers aim squarely at large-scale deployment of GenAI-based workloads for scale-up fine-tuning and massive inference workloads. It can also deal with mixed workloads at scales associated with High-Performance Computing, including simulations, weather forecasting, high-end 3D modeling, protein folding and genomics, and more. The overall emphasis is on maximum performance, efficiency and reliability at scale for production data center use.

Key elements of the HPE ProLiant Compute DL380a Gen12 architecture include:

- Up to 2 Intel Xeon 6 processors
- Up to 4TB of DDR5 RAM
- Up to 8 NVIDIA H200 NVL Tensor Core GPUs
- Choice of networking interfaces (ask your HPE Partner or sales rep for details)
- Up to six dedicated and redundant power supplies for GPUs for efficient, reliable performance

At the edge or in the data center, HPE ProLiant Gen11 and HPE ProLiant Compute Gen12 servers with Intel processors offer tremendous value, terrific energy efficiency, and best use of rack space for capability delivered.

HPE Services Does AI for Customers

Don't forget that HPE also offers an [AI Services](#) arm as part of its HPE Services operations. This group of experts can help make the most of AI's potential, which it promises but sometimes fails to deliver. HPE AI Services can find, provide and support the right expertise, data, tools, technology, and partnerships to make sure your AI efforts come out on the winning side. HPE Services personnel can handle the whole lifecycle, from planning, to implementing, and deploying the right AI solutions. They'll bring their data science, ML engineering, and ML operations (ML Ops) experience and knowledge to bear so you can assess and capture business expectations, meet technology and performance requirements, and prepare data for training and deployment safely and securely.

FINDING MORE INFORMATION

That's the end of the trail for this Gorilla Guide To...® Creating Business Value with Generative AI. Hopefully, you've acquired an appreciation for the ways in which GenAI can add value to your organization (and to your bottom line). That need not be the end of the story, though.

HPE is the global [edge-to-cloud company](#) built to transform your business. How? By helping you connect, protect, analyze, and act on all your data and applications wherever they live, from edge to cloud, so you can turn insights into outcomes at the speed required to thrive in today's complex world. Don't forget that Intel plays an important role, too, as it is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, Intel continuously works to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding

intelligence in the cloud, network, edge, and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to Intel.com (see also [HPE ProLiant Gen 11 Servers with Intel](#)).

ABOUT HPE



Hewlett Packard Enterprise

HPE is the global edge-to-cloud company built to transform your business. How? By helping you connect, protect, analyze, and act on all your data and applications wherever they live, from edge to cloud, so you can turn insights into outcomes at the speed required to thrive in today's complex world. <https://www.hpe.com/us/en/about.html#ourpurpose>

ABOUT INTEL



Intel is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to [Intel.com](https://www.intel.com).

ABOUT ACTUALTECH MEDIA



ActualTech Media, a Future B2B company, is a B2B tech marketing company that connects enterprise IT vendors with IT buyers through innovative lead generation programs and compelling custom content services.

ActualTech Media's team speaks to the enterprise IT audience because we've been the enterprise IT audience.

Our leadership team is stacked with former CIOs, IT managers, architects, subject matter experts and marketing professionals that help our clients spend less time explaining what their technology does and more time creating strategies that drive results.

If you're an IT marketer and you'd like your own custom Gorilla Guide® title for your company, please visit actualtechmedia.com.