NUTANIX | intel

Solution Brief

Cost-Effective Options to Advance Your GenAI Success

Nutanix on Intel® Xeon®-Based Platforms with AMX provide AI-ready infrastructure without the GPU price tag, enabling government organizations to successfully deploy small to medium GenAI inference workloads.

As strategic partners, Nutanix and Intel® offer a pragmatic path forward for government organizations by enabling AI workloads to run efficiently on existing CPU infrastructure with Intel Advanced Matrix Extensions (AMX). This solution reduces the need for expensive GPU investments and delivers practical benefits such as improved performance, flexibility, and security while reducing energy consumption, space and total cost of ownership.

Powered by Intel's fifth-generation Xeon processors, Nutanix delivers a single, software-defined platform that simplifies and modernizes IT infrastructure so agencies can run applications (including GPTs and LLMs), data and services anywhere.

Nutanix: An AI-Ready IT Foundation

Nutanix natively integrates compute, storage, networking, and virtualization into a unified and secure IT foundation that can be deployed across data centers, edge locations, and public and managed clouds. Agencies have choice - using any application, hardware, hyperscaler and hyperscaler, in any combination.

Agencies gain unified management and control over VMs, containers and Kubernetes clusters, simplifying monitoring and management of the entire application stack. With a single consumption-based license, agencies can consolidate file, object and block storage while automating database administration and lifecycle management.

Nutanix enhances security through preconfigured STIGs[2] for secure system hardening and automated compliance self-healing. Native security features include app-centric microsegmentation, role based access controls, storage snapshots, FIPS validated data encryption, and S3 compatible WORM storage, with security policies that maintain consistency across environments.

Intel: Industry-leading innovation

Intel delivers comprehensive, secure and high-performance computing solutions, enabling digital transformation and AI capabilities across mission-critical environments.

Optimized for security, performance and total cost of ownership, Intel processors provide government agencies with a reliable technological foundation for modernizing their IT ecosystems and advancing their digital strategy while allowing them to explore AI and GenAI workloads.

Strategic Partners

- Nutanix is a leader in 2024 Gartner Magic Quadrant for Distributed Hybrid Infrastructure, delivering an AI-ready foundation.[1]
- Intel® Xeon® processors are optimized to deliver performance across the greatest range of workloads, while offering improved efficiency and lower power consumption with every generation.

Advance AI without GPU investment

Agencies can deploy AI apps and workloads,like GenAI without a GPU. Intel AMX is a new, built-in CPU accelerator integrated in every Xeon core starting from fourth generation, enabling existing infrastructure to effectively run AI for small and medium LLMs.

Out of the box, performance is validated for over 300 popular AI models. AMX supports inference for GenAI as well as other predictive and interpretive AI workloads, in addition to classic machine and deep learning, without requiring GPUs.

Intel AMX is designed to improve performance for AI workloads including natural language processing, recommendation systems, and image recognition as well as inference activities on smaller to medium, industry- or domain-specific LLMs and models trained on private data.

Common use cases for government include chatbots, document processing, summarization and analytics, basic video analytics, and resource optimization.

Consider a Simple Way to Advance your AI Success

Nutanix GPT-in-a-Box is how agencies can operationalize AI quickly. This full-stack AI platform exposes the full capabilities of AI-accelerated Intel Xeon CPUs as well as an opinionated AI stack by working with AI software ecosystem partners such as Kubeflow, PyTorch and TouchServe.

In addition to out-of-the-box support for Llama2, MosaicML, and FalconGPT, the platform provides integrated access to the Hugging Face library of models, making it easy for agencies to deploy a wide range of GenAI models securely, on-premises, at the edge and in the public cloud.

The Nutanix Federal Innovation Lab has the latest AI-ready Intel Xeon fifth-gen processors with AMX and offers virtual and hands-on experience through proof of concept testing, allowing agencies to evaluate solutions in a controlled environment before deployment. Contact your Nutanix or Intel account manager to engage. You can also get hands-on experience with deploying a GenAI chat app on a Nutanix test drive.

Visit www.nutanix.com/intel and www.intel.com/nutanix

Nutanix is listed on the Department of Defense Information Network Approved Products List, having passed rigorous interoperability and security vetting.

Nutanix was named a visionary in 2024 Gartner Magic Quadrant for File and Object Storage Platforms.

Intel's fifth-gen Intel Xeon processor family offers unprecedented workload acceleration, while achieving an average of 36% higher performance per watt across a diverse range of workloads.[3]

Legal attributions:

1 https://www.nutanix.com/go/gartner-mq-for-distributed-hybrid-infrastructure

2 STIG information https://aplits.disa.mil/

3 Performance + Efficiency for Accelerated AI Workloads and Beyond. Nov. 7, 2024 https://www.nutanix.com/blog/performance-efficiency-for-ai-workloads-and-beyond

NUTANIX

info@nutanix.com | www.nutanix.com | @nutanix

NutanixAndIntelAI-SolutionBrief-FY25Q3-v1 02212025