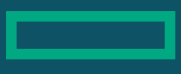
Hewlett Packard Enterprise | Red Hat | NVIDIA

Brochure

ACCELERATE YOUR AI

HPE, Red Hat, and NVIDIA accelerated AI solution

Delivering unparalleled productivity, performance, and flexibility for artificial intelligence (AI)/machine learning (ML) solutions with HPE ProLiant DL360 Gen10 Server and HPE ProLiant DL380 Gen10 Server. Along with, container orchestration with Red Hat® OpenShift Container Platform and AI acceleration by NVIDIA® T4 GPUs and NVIDIA NGC AI/ML containers.



Check if the document is available in the language of your choice.

Key benefits

- Combines the industry's most secure server,[1] the HPE ProLiant DL380, with NVIDIA's industry-leading GPUs and Red Hat's open source OpenShift Container Platform to provide a factory-integrated, workload-optimized AI accelerated solution that is quick to ramp, cost-effective, and simplifies implementation.
- Offers the performance and manageability of a GPU-enabled environment built on HPE ProLiant DL380 Servers, equipped with the new riser to support a dense GPU configuration of up to seven single-wide NVIDIA T4 GPUs or two double-wide NVIDIA V100 GPUs for AI workload acceleration.
- Provides enhanced cyberattack protection through proactive and intrinsic security from silicon-rooted server firmware to vital applications and data.
- Increases AI adoption and removes the technology knowledge roadblocks to speed delivery of accelerated AI by handling the heavy lifting (expertise, time, and compute resources) with pre-trained NVIDIA GPU Cloud models and workflows with best-in-class accuracy and performance.

THE CHALLENGES OF ADOPTING AI/ML INITIATIVES

In the age of digital transformation, AI has the potential to transform every business—in the same way (and possibly more) as the internet has transformed the way we do business. From smarter products and services to better business decisions and optimized business processes, AI has the power to change almost everything. If businesses don't capitalize on this transformative power, they risk being left behind.

AI and related technologies are in production today enhancing existing products and creating new ones. They are optimizing internal and external operations, helping organizations to make better decisions, and freeing up employees to be more creative and to produce higher-value work, among a wide range of other benefits. According to IDC, "By 2024, AI will be integral to every part of the business, resulting in 25% of the overall spend on AI solutions as 'outcomes as a service' that drives innovation at scale and superior business value."[2]

Implementing an AI strategy creates some key challenges:

- Difficult and expensive for customers to set up a development environment for AI initiatives and scale rapidly to production
- Limited expertise designing complex hardware platform for specific workload and use cases
- Lack of experience with Red Hat OpenShift Container Platform for managing AI containers environment
- Limited pool of data scientists in the industry who have experience to develop AI container and libraries from scratch that drive business outcomes based on analytics
- Security risk for intellectual property or mission-critical data

SOLVING KEY CHALLENGES OF TRANSFORMING BUSINESSES WITH ACCELERATED AI

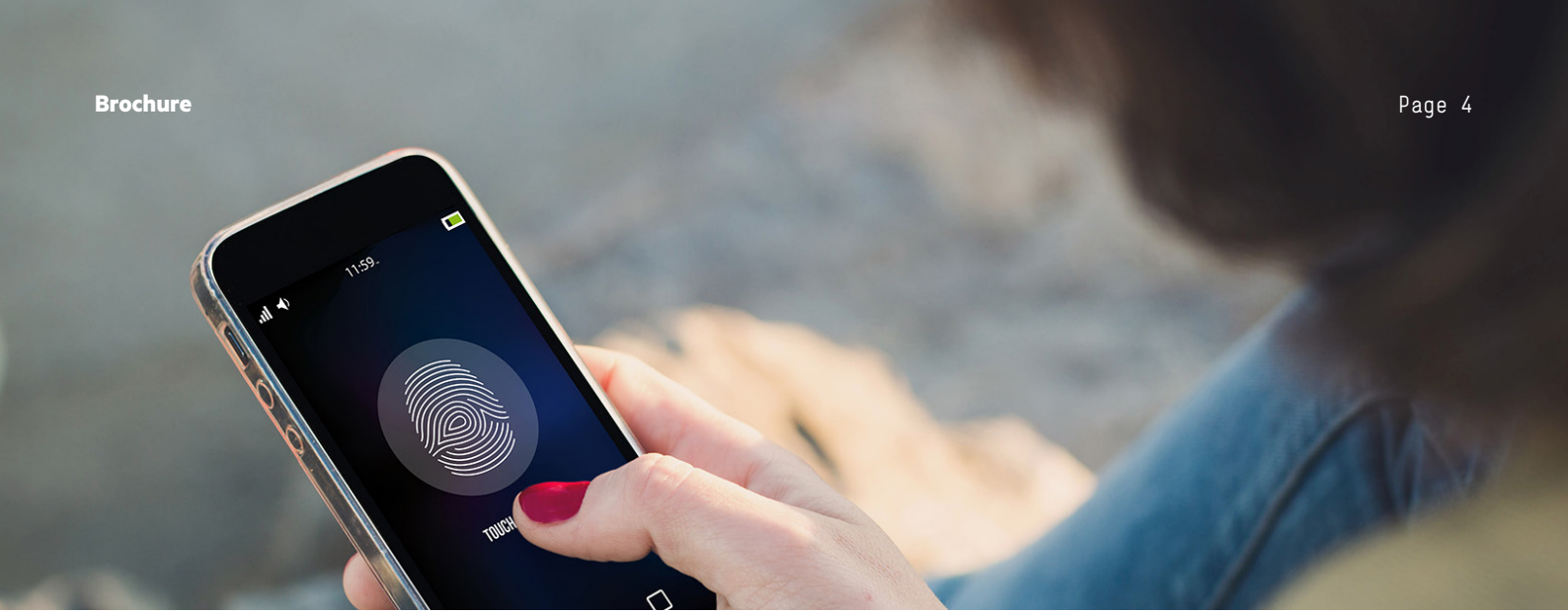
The integrated accelerated AI solution addresses these challenges by offering users the collaboration of the most secure industry-standard Server with the leading GPU vendor, along with leading enterprise hybrid cloud Kubernetes application platform. This offers seamless implementation that is affordable and easy to learn.

The accelerated AI solution is scalable. As your business needs grow, worker nodes (HPE ProLiant DL380 with GPU) can be added to scale. The HPE ProLiant DL380 Server supports a dense GPU configuration of up to two double-wide or four single-wide GPUs for AI workload acceleration. The solution offers a simple acquisition of a complex solution (hardware configuration) removing the requirement of your business to engineer the hardware platform. In addition, the solution leverages the large repository of NVIDIA GPU Cloud containers fine-tuned for specific AI use cases.

Red Hat OpenShift increases AI adoption and removes the knowledge roadblocks to speed delivery of accelerated AI. It includes an enterprise-grade Linux® operating system, container runtime, networking, monitoring, container registry, authentication, and authorization solutions. These components are tested together for unified operations on a complete Kubernetes platform spanning every cloud.

[1] Based on external firm conducting cybersecurity penetration testing of a range of server products from a range of manufacturers, May 2017

[2] Worldwide Artificial Intelligence 2020 Predictions, Document #US45576319, IDC FutureScape, October 2019



IDC AI predictions[3]

- **By 2024, 75%** of enterprises will invest in employee retraining and development, including third-party services, to address new skill needs and ways of working resulting from AI.
- **By 2022, 75%** of enterprises will embed intelligent automation into technology and process development, using AI-based software to discover operational and experiential insights to guide innovation.
- **By 2024**, AI will become the new user interface by redefining user experiences where **over 50%** of user touches will be augmented by computer vision, speech, natural language, and AR/VR.

Hewlett Packard Enterprise addresses these challenges in an effective way with HPE ProLiant DL380 Servers and optionally NVIDIA virtual GPUs (vGPUs). HPE ProLiant DL380 Server is a secure, resilient server that delivers world-class performance and versatility. Its flexible and forward-looking design keeps up with business needs and helps maximize ROI.

HPE ProLiant DL380 Server is an ideal platform for generalized virtualization and high-performance virtualized workloads. It is equipped with industry-leading security and management services that help ensure a secure and easy deployment. With HPE ProLiant DL380 Server's newly designed riser card, up to four single-wide NVIDIA T4 GPUs can be supported in a single server. This provides an option to deploy a range of solutions.

HPE, RED HAT, AND NVIDIA—ACCELERATED AI

HPE ProLiant DL300 Server series, Red Hat OpenShift Container Platform, and NVIDIA T4s and NVIDIA GPU Cloud AI/ML containers collaboration create a factory-integrated, workload-optimized AI solution that is quick to ramp, cost-effective, and easy to understand. It increases AI adoption and removes the knowledge roadblocks to speed delivery of accelerated AI by taking care of the heavy lifting (expertise, time, and compute resources) with pre-trained models and workflows with best-in-class accuracy and performance.

It offers enhanced cyberattack protection through proactive and intrinsic security from silicon-rooted server firmware to vital applications and data.

SOLUTION OVERVIEW

The solution includes four HPE ProLiant DL360 Servers and two HPE ProLiant DL380 Servers each equipped with four highly flexible NVIDIA Tesla T4 GPUs. The software includes Red Hat OpenShift Container Platform 4.1 for container orchestration and NVIDIA GPU Cloud containers to provide AI, inference, and analytics.

[3] Worldwide Artificial Intelligence 2020 Predictions, Document #US45576319, IDC FutureScape, October 2019

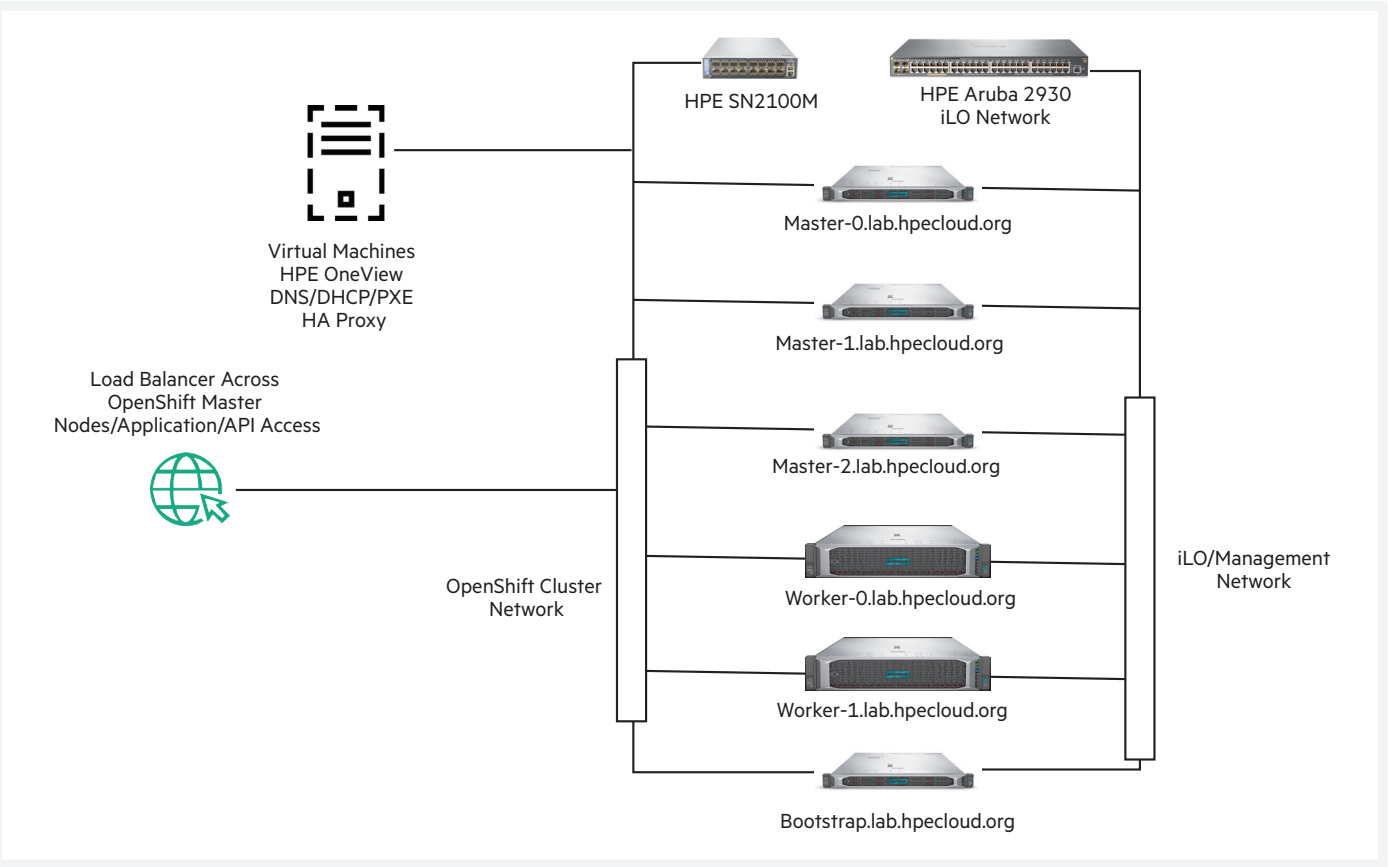


FIGURE 1. Accelerated AI Kit Solution

SOLUTION COMPONENTS

Hardware

The configuration deployed for this solution is described in detail in this section. Table 1 lists the various hardware components in the solution.

TABLE 1. Solution components

Component	Description	Quantity
HPE ProLiant DL360 Gen10 Server	OpenShift Master and Bootstrap/compute node	4
HPE ProLiant DL380 Gen10 Server	OpenShift worker nodes	2
HPE StoreFabric SN2100M Switch	Network switch	1
Aruba 2930	Network switch	1
NVIDIA Tesla T4 PCIe V100 GPU accelerator	GPU modules	8

HPE PROLIANT DL380 SERVERS

HPE ProLiant DL380 Server delivers the latest in security, performance, and expandability, backed by a comprehensive warranty. Standardize on the industry's most-trusted compute platform.[4] The HPE ProLiant DL380 Server is designed to reduce costs and complexity, featuring the first- and second-generation Intel® Xeon® Processor Scalable Family with up to a 60%[5] performance gain and 27% increase in cores,[6] plus the HPE 2933 MT/s DDR4 SmartMemory supporting 3.0 TB. It supports 12 Gb/s SAS and up to 20 NVMe drive plus a broad range of compute options. HPE Persistent Memory offers unprecedented levels of performance for databases and analytic workloads. Run everything from the most basic to mission-critical applications and deploy with confidence.

The HPE ProLiant DL380 Servers used in this reference architecture provide a robust platform to run containerized applications. The two HPE ProLiant DL380 Servers in this solution are deployed as OpenShift compute nodes and configured with four NVIDIA Tesla T4 GPU Accelerators.

Table 2 lists the hardware components installed in the HPE ProLiant DL380 Servers.

TABLE 2. HPE ProLiant DL380 Server configuration

Component	Description
Processor	2 x Intel® Xeon® Gold 6126 (2.6 GHz/12-core/120W)
Memory	12 x HPE 16GB 2Rx8 PC4-2666V-R Smart Kit
Network	HPE IB FDR/EN 40/50Gb 547FLR 2QSFP Adptr
Array controller	HPE Smart Array E208i-a SR Gen10 Ctrlr
Disks	HPE 960GB SATA MU SFF SC DS SSD
GPUs	4X NVIDIA Tesla T4 PCIe accelerators

HPE PROLIANT DL360 SERVERS

The HPE ProLiant DL360 Server delivers security, agility, and flexibility without compromise. It supports the Intel Xeon Scalable processor with up to a 60% performance gain and 27% increase in cores, along with 2933 MT/s HPE DDR4 SmartMemory supporting up to 3.0 TB with an increase in performance of up to 82%.[7]

With the added performance that HPE Persistent Memory,[8] NVDIMMs,[9] and 10 NVMe bring, the HPE ProLiant DL360 means business. Deploy, update, monitor, and maintain with ease by automating essential server lifecycle management tasks with HPE OneView and HPE Integrated Lights Out 5 (iLO 5). Deploy this 2P secure platform for diverse workloads in space-constrained environments.

HPE STOREFABRIC SN2100M SWITCH

HPE SN2100M Ethernet switches are ideal for modern server and storage networks. Supporting port speeds of 1, 10, 25, 40, 50, and 100GbE, delivering predictable performance and zero-packet loss at line rate across each port and packet size. Enhanced for storage combined with efficient design, it provides enterprise-level performance with attractive economics and outstanding ROI.

Networks built on HPE SN2100M are fast, reliable, and scalable while also being affordable and easy to manage. It supports primary and secondary storage, providing consistently fair, fast, and low-latency connectivity even under heavy workloads or a mix of different port speeds. This makes them ideal for storage, hyperconverged, financial services, and media and entertainment deployments.

[4] Based on external firm conducting cybersecurity penetration testing of a range of server products from a range of manufacturers, May 2017

[5] HPE measurements conducted in April 2019. Up to 60% performance increase of Intel Xeon Platinum vs. previous generation E5-2600 v4 average gains of STREAM, LINPACK, SPEC CPU 2006, and SPEC CPU 2017 metrics on HPE servers comparing 2-socket Intel Xeon Platinum 8280 to E5-2699 v4 family processors. Any difference in system hardware or software design or configuration may affect actual performance.

[6] HPE measurements conducted in April 2019. Up to 27% core increase of Intel Xeon Platinum versus previous generation comparing 2-socket Intel Xeon Platinum 8280 (28 cores) to E5-2699 v4 (22 cores). Calculation 28 cores/22 cores = 1.27 (27%).

[7] HPE measurements conducted in April 2019. Percentage compares Gen10 vs Gen9; Gen10 = 12 channels x 2933 data rate x 8 bytes = 281 GB/sec; Gen9 = 8 channels x 2400 x 8 bytes = 154 GB/Sec; 281/154 = 1.82 or Gen10 is 82% greater bandwidth. Any difference in system hardware or software design or configuration may affect actual performance.

[8] Supported by the second-generation Intel Xeon Scalable processors

[9] Supported by the first-generation Intel Xeon Scalable processors

ARUBA 2930F SWITCH SERIES

Aruba 2930F Switch Series is designed for customers creating smart digital workplaces that are optimized for mobile users with an integrated wired and wireless approach. These convenient Layer 3 network switches include built-in uplinks and PoE power and are simple to deploy and manage with advanced security and network management tools such as Aruba ClearPass Policy Manager, Aruba AirWave, and cloud-based Aruba Central.

A powerful Aruba ProVision ASIC delivers performance, robust feature support, and value with programmability for the latest applications. Stacking with virtual switching framework (VSF) provides simplicity and scalability. The 2930F supports built-in 1GbE or 10GbE uplinks, PoE+, access OSPF routing, dynamic segmentation, robust QoS, RIP routing, and IPv6 with no software licensing required.

Aruba 2930F Switch Series provides a convenient and cost-effective access switch solution that can be quickly set up with zero-touch provisioning. The robust Layer 3 feature set includes a limited lifetime warranty.

NVIDIA TESLA T4 PCIE GPU ACCELERATOR

The NVIDIA T4 GPU accelerates diverse cloud workloads, including high-performance computing, deep learning training and inference, machine learning, data analytics, and graphics. Based on the new NVIDIA Turing™ architecture and packaged in an energy-efficient 70W, small PCIe form factor, T4 is optimized for mainstream computing environments and features multi-precision Turing Tensor Cores and new RT Cores. Combined with accelerated containerized software stacks from NGC, T4 delivers revolutionary performance at scale.

SOFTWARE

Red Hat OpenShift Container Platform
OpenShift Container Platform is Red Hat's enterprise-grade Kubernetes based distribution that provides enterprises the ability to build, deploy, and manage container-based applications. Red Hat OpenShift Container Platform provides enterprises with a full-featured Kubernetes based environment that includes automated operations, cluster services, developer services, and application services to build platform-as-a-service (PaaS) and containers-as-a-service (CaaS) on-premises hybrid cloud solution. Red Hat OpenShift Container Platform provides integrated logging and metrics, authentication and scheduling, high availability, automated over the air updates, and an integrated application container registry.

Red Hat Enterprise Linux CoreOS
OpenShift Container Platform uses Red Hat Enterprise Linux CoreOS (RHCOS), a new container-oriented operating system that combines some of the best features and functions of the CoreOS and Red Hat Atomic Host operating systems. RHCOS is specifically designed for running containerized applications from OpenShift Container Platform and works with new tools to provide fast installation, operator-based management, and simplified upgrades.

RHCOS includes:

- Ignition, a first-boot system configuration for initially bringing up and configuring OpenShift Container Platform nodes.
- CRI-O, a Kubernetes native container runtime implementation that integrates closely with the operating system to deliver an efficient and optimized Kubernetes experience.
- Kubelet, the primary node agent for Kubernetes that is responsible for launching and monitoring containers.

In OpenShift Container Platform 4.1, you must use RHCOS for all control plane machines, but you can use Red Hat Enterprise Linux (RHEL) as the operating system for compute (worker) machines. If you choose to use RHEL workers, you must perform more system maintenance than if you use RHCOS for all of the cluster machines.

HPE ProLiant Servers
HPE ProLiant is an intelligent foundation for hybrid cloud with a fresh, flexible, and software-defined approach, delivering unmatched workload optimization, security, and automation, all available as a service.

Workload optimized

The foundational intelligence of HPE ProLiant transforms IT with insights that help optimize workload performance, placement, and efficiency, delivering better outcomes faster.

360-degree security

Already the world's most secure industry-standard server, HPE ProLiant, provides an enhanced holistic, 360-degree view to security that begins in the manufacturing supply chain and concludes with a safeguarded, end-of-life decommissioning.

Intelligent automation

The intelligence built into HPE ProLiant simplifies and automates management tasks, establishing a solid foundation for an open, hybrid cloud platform enabled by composability.

Delivered as a service

HPE provides customers choice in how they acquire and consume IT. Beyond traditional financing and leasing, HPE offers options that free trapped capital, accelerates infrastructure updates, and provides for on-premises pay-per-use consumption.

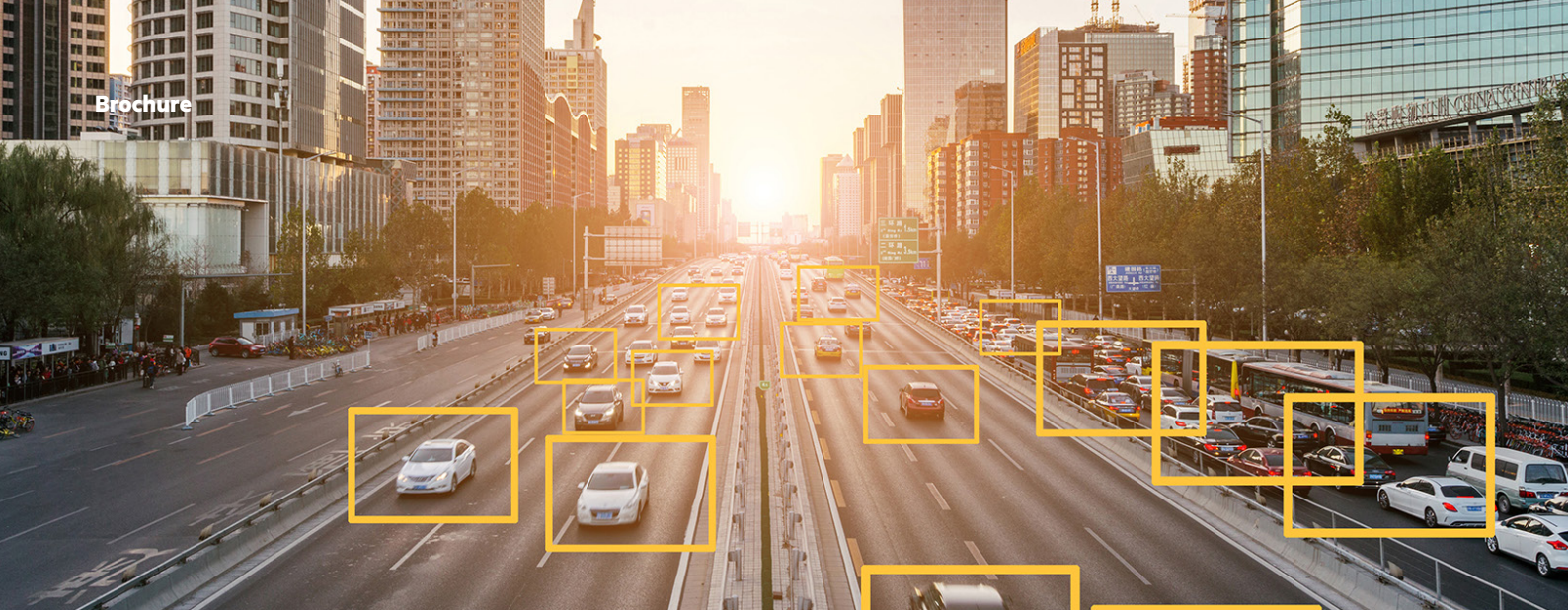


Table 3 lists the versions of Red Hat OpenShift Container Platform and RHCOS used in the creation of this solution. The installer should ensure they have downloaded or have access to this software.

TABLE 3. Software versions

Component	Version
Red Hat CoreOS	4.1
Red Hat OpenShift Container Platform	4.1

SUMMARY

Just as AI and related technologies are enhancing existing products and creating new ones, this solution provides guidance to accelerate your AI initiatives. Information garnered from the massive amounts of disparate data collected from multiple sources throughout an enterprise can be used to set strategic goals and provide a competitive advantage. HPE ProLiant DL Servers powered by NVIDIA Tesla T4 GPUs combined with Red Hat OpenShift Container Platform allow CPU-intensive data analytics tools and applications to be rapidly deployed to end users through a self-service OpenShift registry and catalog.

LEARN MORE AT

hpe.com/servers/dl380

Accelerated AI Reference Architecture

a50000769ENW, January 2020

Hewlett Packard Enterprise