Addressing sustainability in the financial services industry

Business white paper

HPE GreenLake

How new server technologies and Direct Liquid Cooling can help address sustainability, lower carbon emissions, and achieve net zero emission goals

Introduction

Financial institutions (FIs) face unprecedented challenges. With increasing consumer expectations and megatrends such as mobile banking, alternative lenders, and cloud computing, banks are under intense pressure. They need to innovate faster, offer new services, and deploy new applications to serve their customers better. Despite these imperatives, data center growth rates have become unsustainable. The need for technology innovation has run up against equally pressing requirements to meet sustainability goals.

These challenges demand an **all-of-the-above** approach. This paper will discuss two important aspects of the sustainability puzzle—improving server efficiency and data center power usage effectiveness (PUE).[1] The paper will also explain how data center operators can improve throughput, density, and efficiency using HPE Apollo Gen10 Plus systems powered by the latest AMD EPYC™ processors. Finally, it will outline the benefits of Direct Liquid Cooling (DLC) technology and why it is a potential game changer for reducing power and cooling requirements and addressing sustainability.

A challenging application landscape

For financial firms, high-performance computing (HPC) has long been an important competitive differentiator. Clusters of high-performance systems are used for multiple applications, including options pricing, scenario modeling, portfolio evaluation, backtesting, and quantifying market and credit risk. HPC is essential to helping firms understand their exposures and optimizing capital reserves. Banks that can model and accurately price risk faster than competitors enjoy a decisive time-to-market advantage.

Following the financial crisis of 2008, banks faced a tsunami of new regulations around modeling liquidity risk, counterparty exposures, and stress testing. Most investment banks and fund managers operate large clusters to support risk management environments. HPC is also critical in high-frequency trading (HFT), where banks leverage GPU and FPGA-based systems for order execution, arbitrage, and trend trading strategies. Banks compete in these and other areas based on their HPC infrastructure's effectiveness.

New technologies are transforming banking as if the IT landscape were not already challenging enough. These include mobile commerce, the rise of Big Data, cloud-based delivery models, and the increased adoption of artificial intelligence (AI) across all facets of banking. The impact of AI is transformative. Fueled by massive growth in data and increasingly sophisticated predictive models, AI is emerging as a critical differentiator. It is used in customer service applications, fraud analytics, process automation, and recommendation engines that provide the **next-best** actions. AI is not only transforming analytics in areas such as operations and risk management—it is helping create new customer experiences that improve loyalty, satisfaction, and retention. More than ever, HPC is now essential to competitiveness.

[1] Data center PUE is a ratio that describes how efficiently a computer data center uses energy

New strains on data center capacity

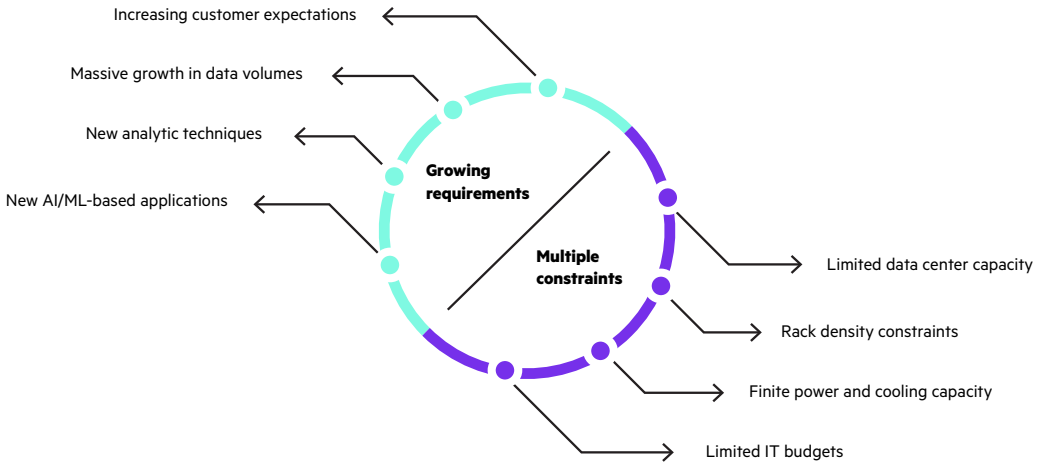

Figure 1. Growing requirements are running up against limited resources and other constraints

Most FIs have made considerable strides in embracing cloud delivery models and improving data center operations. By embracing virtualization, VDI, and containerization, they have reached high levels of efficiency—often running at utilization levels comparable to cloud hyperscalers. Despite these advances, data center growth rates have become unsustainable.

Tapping capacity in public clouds is an option. However, there are few incentives for operators that are already highly efficient. Given data gravity considerations, security needs, regulatory requirements, and the proprietary nature of their software, FIs often pursue a mixed strategy, operating their infrastructure and operating in the cloud or using hybrid approaches where it makes sense.

Space is also at a premium. The cost of building new data center space is high, so increasing server density is often a better solution than leasing or building additional data center capacity. Financial firms are increasingly bumping up against hard limits related to floor space, performance, latency, and power and cooling capacities.

Sustainability emerges as a key concern

A key concern for all businesses is the unsustainable levels of carbon dioxide (CO_2) emissions and greenhouse gasses (GHG) driving climate change. CO_2 emissions are forecasted to increase to 54.08 billion metric tons in 2050, compared to 35.3 billion metric tons in 2018.[2] Unchecked, this growth rate is forecasted to lead to catastrophic increases in average global temperatures of over 4°C by 2100.[3] To limit global temperature rise to between 1.5°C and 2.0°C, the Paris Agreement was adopted in 2015.[4] According to the International Energy Agency, data centers and the data transmission networks that connect them account for over 2.1% of global electricity demand, and HPC is a significant contributor.[5]

Most large enterprises have made public commitments around sustainability. These include pledges to accelerate a low-carbon energy transition and achieve net zero emissions by specific dates. Given their sizeable impact, curbing data-center-related emissions looms large in most corporate sustainability plans.

[2] "Forecast of carbon dioxide emissions worldwide from 2018 to 2050," Global CO_2 emissions 2018–2050, Statista, 2020

[3] "Global warming of 1.5°C," An IPCC Special Report on the impacts of global warming

[4] The Paris Agreement, United Nations Climate Change

[5] "Data Centres and Data Transmission Networks," International Energy Agency, 2021: 1% of global electricity use is attributable to data centers and another 1.1%–1.4% is attributable to data transmission networks

Today, many FIs limit a typical data center rack to approximately 10 kW of power with high-density configurations in HPC going from 20 kW to 40 kW.[6] For a large facility with 500 racks and 10,000 servers, simply powering the servers alone requires 43,800 MWh costing $4.6M annually, enough to power over 4,000 homes.[7] At 0.85 lb of CO_2 emissions per kWh, this is equivalent to over 13,000 tons of CO_2 gases released annually.[8]

Cooling—a foreseeable challenge

As high as the figures mentioned are, these power requirements do not include cooling. Based on separate estimates from Hewlett Packard Enterprise, the power required to cool a 10,000-server facility for one year is roughly 20,660 MWh.[9] This costs an additional $2.17M annually, translating into roughly 8,800 tons of CO_2 emissions.

Unfortunately, even as individual servers and components become more power-efficient, the need for performance in financial services is driving power and cooling requirements even higher. These trends are illustrated in Figure 2. For top-bin CPUs, per socket power requirements are moving from 270 watts to 350 watts–380 watts. Next-gen CPUs may require 550 watts per socket. GPUs, increasingly critical for AI-powered applications, are following a similar trend. Internal HPE estimates indicate that per-socket power consumption for GPUs is expected to reach 700W.[10] Silicon designs are increasingly going 3D with components layered on one another. This creates new thermal challenges and requires that case temperatures be cooled to even lower levels to avoid overheating.[11]

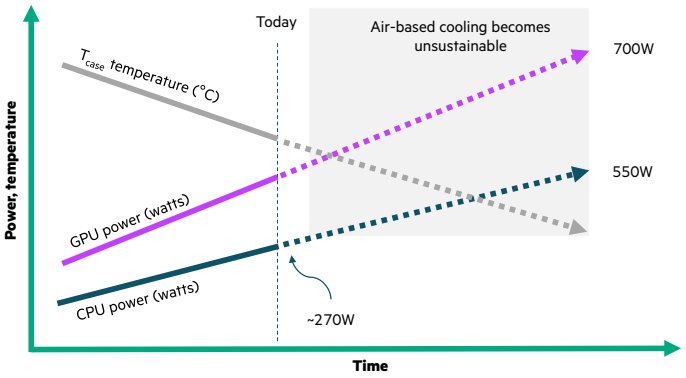

Figure 2. Trends in CPU and GPU power consumption vs. cooling capacity

Also, impacting these trends is increasing data center density. While air cooling is technically possible for some high-end parts, this would require using enormous heat sinks—reducing server density and driving up costs. Given these trends, air-based cooling will likely no longer be an option in dense data center environments—especially those running the latest GPU accelerators.

An "all-of-the-above" approach to curbing emissions

FIs face a dilemma. Business and competitive pressures drive the need for more capacity and denser data centers. But then, they are under enormous pressure to reduce power consumption and curb emissions. Fortunately, there are solutions on the horizon. While modern high-end processors are power hungry, they are also becoming more energy-efficient, offering higher levels of throughput per watt with each generation. GPU-accelerated computing is also playing a role. It is no accident that GPU-based servers dominate the GREEN500 list.[12] By employing high-performance GPUs and faster, more efficient processors, banks can potentially reduce the size of their compute grids and associated power and cooling costs. Technologies such as DLC can also help substantially reduce power requirements.

[6] HPE Apollo with HPE Apollo Direct Liquid Cooling System solution brief

[7] 10 kW x 500 racks x 24 x 365 = 46,800,000 kWh. Average power consumption per home of 10.7 kWh. The calculation assumes an average of $0.105 per kWh

[8] Carbon dioxide produced per kilowatt of U.S. electricity generation

[9] HPE internal estimate, June 2022: based on a cost estimate of $216.77 per year to cool a single server using air cooling. $2,167,700 / $0.105 = 20,664,762 kWh. Total CO2 = 20,664,762 x 0.85 = 17,548,048 lb or 8,774 tons

[10] HPE internal estimate, June 2022

[11] T_{case} refers to the temperature at the interface between a CPU package and its heatsink

[12] GREEN 500 list as of June 2022

Energy efficient servers

Denser, more energy-efficient data centers start with energy-efficient servers. As the market leader in HPC and a pioneer in exascale computing, HPE plays a key role in promoting energy efficiency and sustainability.[13] Innovation in Top500 supercomputing environments trickles down to commercial offerings resulting in more performant and energy-efficient systems. The HPE Apollo Gen10 Plus system is an excellent example.

HPE Apollo 2000 Gen10 Plus systems

The HPE Apollo 2000 Gen10 Plus system is a density-enhanced multiserver with shared power and cooling resources that delivers high levels of efficiency and system scaling. Each 2U chassis supports up to four HPE ProLiant XL225n Gen10 Plus servers, each with up to two industry-leading 2nd or 3rd generation AMD EPYC processors.[14, 15] Depending on customer requirements, these systems provide storage and compute flexibility and can be air cooled or liquid cooled.

HPE Apollo 2000 Gen10 Plus systems are ideal for HPC applications such as risk analytics. They are also highly efficient when used as a general-purpose platform for cloud-based application delivery. Given their dense design and exceptional throughput, they can support virtualized and containerized applications and a wide range of applications from analytics to retail banking.

The HPE Apollo 2000 Gen10 Plus system with HPE ProLiant XL225n Gen10 Plus servers powered by AMD EPYC 7003 series processors has achieved 18 world records in energy efficiency.[16] These multinode servers deliver real space and power savings to data centers of any size. Taking energy efficiency to the maximum, the HPE ProLiant XL225n Gen10 Plus server has the highest result of 17,696 overall ssj_ops/watt for 4-node configurations on the SPECpower_ssj® 2008 benchmark.[17] With the HPE Apollo 2000 Gen10 Plus system, customers can realize up to $15,000 in annual energy cost savings vs. competitive systems.[18]

HPE Apollo 6500 Gen10 Plus systems

For customers running GPU-intensive applications such as GPU-powered risk simulations or AI model training, the HPE Apollo 6500 Gen10 Plus system is an excellent solution. With dense HPE ProLiant XL675d server nodes, this system supports up to 10 double-wide PCIe or 16 single-wide PCIe GPUs per server. In addition to AMD EPYC processors, HPE Apollo 6500 systems support the latest GPUs, including AMD Instinct™ MI100 and AMD Instinct MI210 accelerators with 2nd generation Infinity Fabric™ technology along with other GPU accelerators.[19]

[13] AMD EPYC-based systems have been chosen as the basis of exascale supercomputers. Design wins include Frontier, a collaboration between the U.S. Department of Energy (DOE), ORNL, AMD, and HPE. Frontier is the first supercomputer to break the exascale barrier delivering 1.1 exaflops. AMD EPYC processors will also power El Capitan, a collaboration between U.S. DOE, LLNL, and HPE expected in early 2023

[14] HPE Apollo 2000 Gen10 Plus System

[15] See AMD EPYC™ Processor World Records and accompanying footnotes

[16] HPE Apollo 2000 Gen10 Plus System with HPE ProLiant XL225n Gen10 Plus Servers Achieve 18 World Records in Energy Efficiency

[17] The results referenced are as of March 15, 2021. For details on the four-node HPE Apollo XL225n Gen10 Plus benchmark result, see: spec.org/power_ssj2008/results/res2021q1/power_ssj2008-20210223-01073.html

[18] Internal HPE calculation: Annual energy cost and rack space calculated based on the performance envelope of a 42U rack populated with HPE ProLiant XL225 Gen10 Plus servers running at 100% utilization vs. the energy and rack space required by competitor products to achieve the same performance

[19] HPE Apollo 6500 Gen10 Plus System QuickSpecs

The “EPYC” advantage

Built on 7nm technology, AMD EPYC processors bring together high core counts, large memory capacity, extreme memory bandwidth, large cache sizes, and massive I/O with the right ratios to enable exceptional HPC workload performance.[20] This can translate into faster pricing calculations and more timely analysis for financial workloads, helping deliver a competitive advantage.

AMD EPYC, with its high core counts and memory footprint, is emerging as a leading platform for collapsing many workloads into a single socket while maintaining low, deterministic latencies.

Featuring up to 128 threads per socket and a highly parallel design with eight separate core complex dies (CCDs) and memory channels, AMD EPYC processors are also a favorite processor for cloud-based application delivery models. These processors deliver high throughput on multithreaded workloads enabling more virtual machines (VMs) and threads per physical socket, more tenants per server, and higher data center efficiency.

Financial services customers can also benefit from leadership AMD Instinct GPUs and AMD Xilinx Solarflare™ network adapters. Widely used in FinTech, these FPGA-based 100G NICs deliver exceptional performance and latency—up to 100 million packets per second while consuming less than 64 watts of power.[21]

DLC improves sustainability

For customers with suitably equipped data centers, a new HPE Apollo Gen10 Plus systems option is plug-and-play support for DLC. The DLC option allows customers to increase power density and data center efficiency. HPE server racks connect directly to facility water supplies without secondary plumbing. Options are available for CPU only, CPU plus memory, and GPU cooling.

Figure 3. The DLC option is available in CPU, CPU plus memory, and GPU configurations

The DLC option, pictured in Figure 3, is available for HPE Apollo 2000 and HPE Apollo 6500 Gen10 Plus systems. Liquid cooling is especially helpful in the dense GPU configurations supported by HPE Apollo 6500 Gen10 Plus systems. As illustrated in Figure 4, DLC can have a dramatic impact on improving energy efficiency and helping organizations meet sustainability goals.

[20] AMD EPYC 7003 Series Processors

[21] AMD Xilinx Press Release: Xilinx to Acquire Solarflare

Based on an internal analysis by HPE, the cost of cooling a single server for a year in an air-cooled environment is $216.77.[22] In air-cooled environments, ~18 servers are typically deployed per rack due to power-density limitations. Both cooling costs and data center density can be dramatically improved using DLC. Based on the same analysis, the annual cost of cooling a single server in a DLC environment is just $29.74 per server per year. Also, using DLC, up to four HPE Apollo XL225n Gen10 Plus dual-processor servers can be housed per two rack units, for a total of 80 servers per rack.[23]

In a 10,000-server environment, the impact of DLC on cooling costs and data center density is impressive. Organizations can potentially reduce annual cooling costs and associated CO_2 emissions by 87.3%—enough energy to power ~1,800 homes for a year.[24]

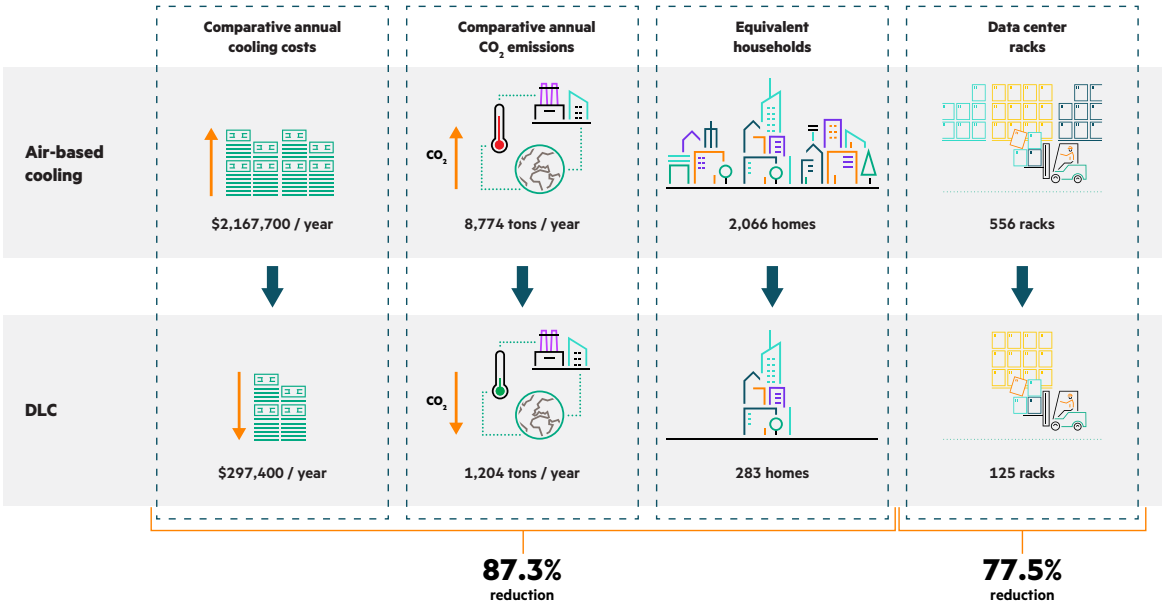

Figure 4. Cooling costs and CO_2 emissions in air-cooled vs. direct-liquid cooled environments

While not every data center is equipped for DLC today, it is relatively easy to add to existing facilities. Installation costs vary depending on the extent of facility modifications required. Payback periods are generally short, given the significant financial savings. Also, power savings recur once a one-time investment in DLC cooling is made.

DLC delivers multiple benefits

As illustrated in Figure 5, the advantages of HPE Apollo Gen10 Plus systems with DLC extend beyond sustainability. HPE DLC solution is fully integrated, installed, and supported by HPE. It is also easy to implement, requiring only a connection to existing on-site water facilities.

With improved cooling, processors can operate at high clock rates for sustained periods enabling higher throughput and predictability for compute-intensive workloads. Also, components such as GPUs are less likely to be taken out of service by workload managers due to high-temperature and power conditions, prolonging server components' life, reducing failures, and improving throughput and availability.

[22] Based on HPE internal estimate, June 2022: Assumes a cost of $0.105 per kWh

[23] Assumes a 48U rack equipped with DLC. See HPE Apollo 2000 Gen10 Plus System QuickSpecs for details

[24] Based on a reduction of annual power requirements for cooling of 20,664,762 kWh – 2,832,381 kWh = 17,832,381 kWh and an average emission of 0.85 lb of CO2 per kWh

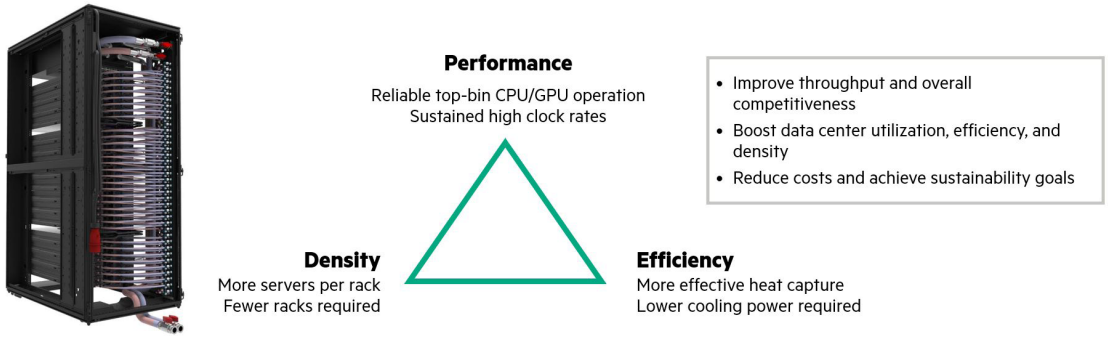

Figure 5. Direct liquid cooling offers benefits beyond sustainability

HPE GreenLake

Beyond more power-efficient servers, there are other paths to a greener, more sustainable data center. Most banks are evolving toward as-a-service delivery models that support virtualized or containerized applications to improve utilization and efficiency. Our transition to a consumption-based, as-a-service company is helping drive sustainable transformation efforts for our customers.

HPE GreenLake edge-to-cloud platform brings cloud experience, self-service, pay per use, and the ability to scale up and down. HPE GreenLake is fully managed by HPE and is deployable in any location, including the network edge, colocation facilities, and data centers. HPE GreenLake is the cloud that comes to you, enabling organizations to free up capital, boost operational and financial flexibility, and free up talent to accelerate important initiatives. With HPE GreenLake, customers can deploy elastic, production-ready cloud services in as little as 14 days.

Customers transitioning to HPE GreenLake from traditional CAPEX models can achieve a greater than 30% reduction in energy costs and total cost of ownership.[25] Multiple financial firms are already realizing the benefits of HPE GreenLake, including YF Life, au Kabucom Securities Co., Ltd., and MKB Bank.[26] According to a recent Forrester study, HPE GreenLake users can receive an ROI of up to 161%.[27]

HPE and AMD—working toward a sustainable future

HPE and AMD have publicly announced corporate sustainability initiatives. Today, AMD EPYC processor-powered systems deliver the industry's highest throughput per watt, holding the top spots in the industry-standard SPECpower_ssj2008 benchmark.[28] Moreover, AMD is on a path to extend this leadership further and is on track to achieve an ambitious goal to deliver a 30x increase in energy efficiency for AMD processors and accelerators powering servers for HPC and AI training from 2020 to 2025. AMD offers a GHG Emissions TCO estimation tool to estimate the potential savings and emission reductions with various AMD EPYC CPUs.[29] By utilizing highly performant AMD CPUs, FIs can run more concurrent simulations per socket and get results faster, significantly boosting throughput per watt and helping to achieve their sustainability goals.

In June 2022, HPE unveiled its annual Living Progress Report for 2021, demonstrating the company's ongoing commitment to creating a more equitable and sustainable world. In this report, HPE announced accelerated targets for net-zero emissions by ten years from 2050 to 2040. By 2030, HPE pledges to reduce emissions within its operations by 70% from 2020 levels, building on the 62% reduction already achieved over the five prior years.

[25] "Hewlett Packard Enterprise Releases 2021 Living Progress Report; Accelerate Net-zero Climate Target by 10 Years," June 2022

[26] The future of financial services—HPE GreenLake for financial services

[27] "The Total Economic Impact™ Of HPE GreenLake," Forrester Consulting, May 2022

[28] See EPYC-028: AMD EPYC provides leading results on SPECpower_ssj2008 benchmark

[29] AMD EPYC Bare Metal and Greenhouse Gas Emissions TCO Estimation Tool

Summary

Financial services firms face intense pressure to deploy new applications to capitalize on market opportunities. They need to keep up with customer expectations and ensure the long-term health and competitiveness of the business. They also face pressures to curb data center growth, reduce power consumption and GHG emissions, and meet aggressive sustainability targets.

HPE offers comprehensive solutions, including systems, services, and an edge-to-cloud platform to help customers meet these challenges. By deploying HPE solutions, financial services firms can:

- Benefit from leading power-efficiency
- Significantly improve density and reduce data center space requirements
- Reduce power and cooling costs and GHG emissions

Learn more at

hpe.com/us/en/solutions/financial-services-industry.html

HPE Apollo with HPE Apollo Direct Liquid Cooling System solution brief

amd.com/en/processors/epyc-7003-series

amd.com/en/corporate-responsibility

Hewlett Packard Enterprise

a50006586ENW