White Paper

# Mission-Critical Infrastructure for the Data-Driven Enterprise

Sponsored by: Hewlett Packard Enterprise and Intel Corp.

Peter Rutten Ashish Nadkarni
June 2020

## EXECUTIVE SUMMARY

Data is at the core of the modern enterprise, regardless of size or industry or customer type. But data is no longer a homogeneous concept. It is generated in many different ways and managed with a widening set of data management technologies and infrastructure solutions. This has led to an increasingly diversified approach for data management software and data management infrastructure.

In this heterogeneous environment, the mission-critical real-time online transaction processing (OLTP) and online analytical processing (OLAP) platforms that are designed for hosting both core relational database management systems (RDBMS) and modern data management systems stand out. These platforms combine unique characteristics around transactional and analytical performance, reliability, availability, and security. The move to the cloud off these platforms has therefore been slower, and IDC has even seen a return to private cloud on premises, referred to as "repatriation," for transactional and analytics applications.

From a data perspective, a move to a future-ready real-time enterprise includes the convergence of business-centric transaction processing and data-centric analytics systems. To speed up analytics, businesses are deploying in-memory and memory-centric databases, infusing data analytics platforms with high-performance technologies, and using a highly available and secure conduit for data movement between the various application tiers.

The HPE Superdome Flex family of servers is designed to host critical enterprise workloads such as conventional Oracle and Microsoft SQL Server and in-memory databases such as SAP HANA, Oracle Database In-Memory, and Microsoft SQL Server with in-memory capabilities. The servers are also utilized for high-performance computing (HPC) and artificial intelligence (AI) workloads, which, given high data interdependency, commonly run on a single system or cluster node. In addition, Superdome Flex servers are well suited for Unix-to-Linux migrations as these applications benefit from the near-linear compute, memory, and I/O scalability; extreme availability; and simplified management capabilities.

IDC believes that the Superdome Flex family sets a high standard for mission-critical servers for data-driven enterprises and is worthy of consideration by firms embarking on a journey to modernize their applications and infrastructure and, crucially, to unlock the value of their data in a timely manner.

## SITUATION OVERVIEW

Data is at the core of the modern enterprise, regardless of size or industry or customer type. An organization may be midsize and perform mostly business-to-business transactions, or the organization may have a global presence and execute millions of consumer transactions a day. In all cases, the way that businesses interact with their customers, suppliers, ecosystem partners and, indeed, the entire external world, happens increasingly through digital means, supported with ever-larger volumes of data that inform those relationships.

Data is no longer a homogeneous concept. It is generated in many different ways, is managed with a widening set of new and existing technologies and, if anything, has led to a more heterogeneous way of using data. Several new data management types have emerged — from Hadoop to Spark to NoSQL to Kafka to graph databases. Likewise, new infrastructure solutions have been developed to optimally process this variety of data. All this has resulted in a "purpose built" approach to data management and data infrastructure, rather than the homogeneous approach from years past. What this means is that IT is increasingly inclined to match a specific type of data that it wishes to process with the optimal data management technology and, in equal measure, with the best data management infrastructure.

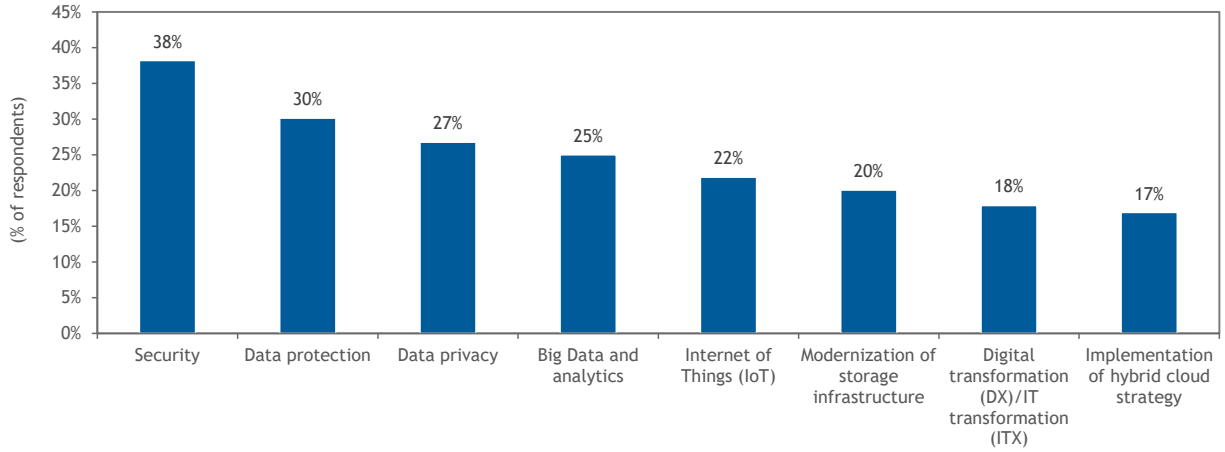
IDC has seen infrastructure vendors — processor manufacturers and storage and server OEMs — embrace this "purpose built" approach for every workload and its specific data and data management requirements. Processors are designed for AI inferencing, for example, or for scale-up platforms. Servers, with varying processor types, I/O options, coprocessors, and storage solutions, are aimed at AI training, for example, or virtualization, or data-intensive workloads.

Interestingly, what stands out in this new purpose-built, heterogeneous data processing environment are the systems that foreshadowed this diversity and that have for many years been built for a specific data processing purpose, namely mission-critical OLAP and OLTP systems that house relational database management systems as well as the latest data management platforms.

IDC sees these mission-critical enterprise platforms not just withstand the winds of change but in many ways fan them on. The platforms have been standardized, modernized, opened up, and cloudified, but they have also maintained their unique characteristics: unmatched transaction processing and analytical performance combined with the highest security, reliability, and data protection. Figure 1 shows the top projects that are driving IT infrastructure spending, with security and data protection ranked highest.

## FIGURE 1

### Top Projects Driving IT Infrastructure Spending



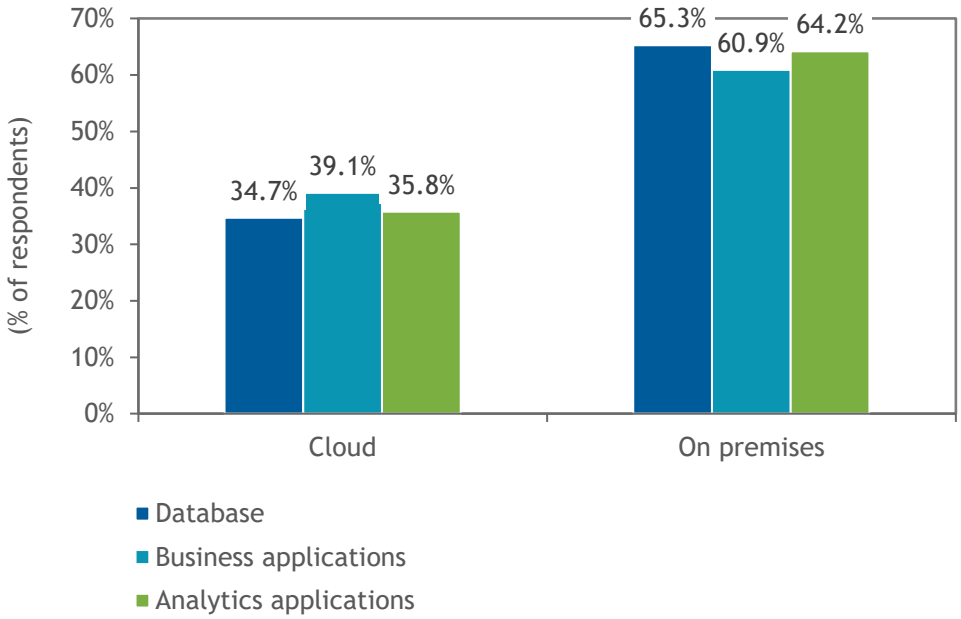

Source: IDC, 2020

The move to the cloud off these platforms has been slow for this very reason, to the point where cloud service providers (CSPs) have begun mimicking mission-critical platforms on their clouds in an attempt to differentiate themselves and attract new enterprise customers. And while the move to cloud certainly continues, IDC is, at the same time, seeing a return to on premises on private cloud, referred to as "repatriation," for transactional and analytics applications.

In a survey of businesses that are running SAP applications on various databases, 65.3% said they run the database on premises, 60.9% said they run the business applications on premises, and 64.2% said they run their analytical applications on premises (see Figure 2).

## FIGURE 2

### Cloud Versus On-Premises Deployment: Database and Business and Analytics Applications Infrastructure



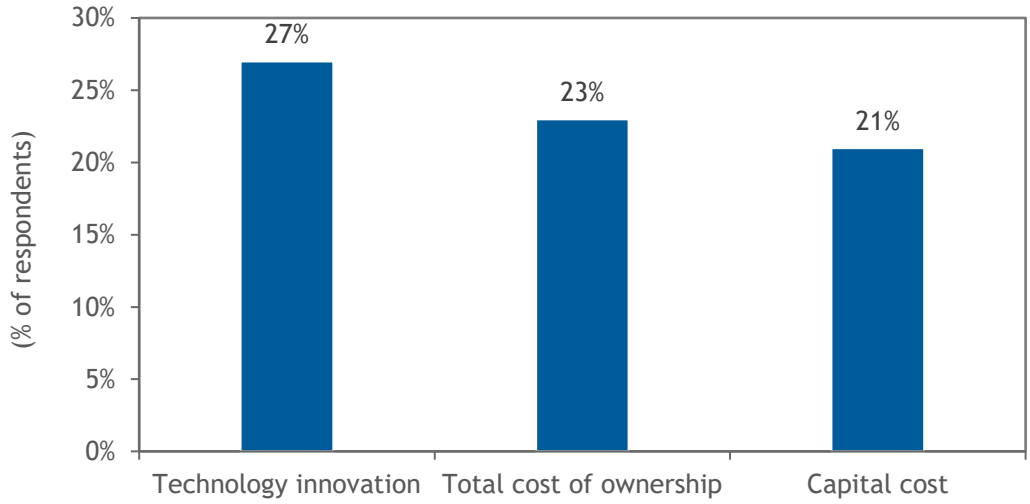

Source: IDC, 2020

Modern analytics environments provide the crucial underpinning for firms transforming themselves into data-driven enterprises. Being data driven enables modern enterprises to compete more effectively in the digital economy via analysis of data coming from core enterprise applications and emerging initiatives such as the Internet of Things, robotics, next-generation security, and next-generation supply chain automation. As firms seek to create and deliver digital offerings and experiences, insights obtained from data are paramount in making key business decisions.

Data is the new basis of competitive advantage. IDC finds that businesses are realizing that accelerating analytics and, ultimately, unlocking the value of data in real time are crucial to their ability to lead in the digital economy. IDC data shows that businesses consider innovation therefore as the most important characteristic of the infrastructure where the data processing and management take place (see Figure 3).

## FIGURE 3

### Top Criteria for Selecting Servers




Source: IDC, 2020

Furthermore:

- As data continues to pervade organizations at all levels at an ever-increasing pace, organizations are challenged to handle the volume, velocity, and veracity of data as leadership strives to derive value from the data and drive business impact in a real-time fashion.
- Generating data intelligence requires the analysis of vast quantities of diverse data, either structured or unstructured and generated by humans or by machines, to uncover patterns and pursue breakthrough ideas.
- AI is the latest stage of analytics with AI training and inferencing becoming an integrated aspect of organizations' analytical capabilities.

Unlocking the intelligence from data in real time requires a modern application and data management environment. The IT infrastructure that hosts these applications and data management platforms serves as a critical foundation layer. The move to a real-time enterprise includes:

- Converging business-centric transaction processing and data-centric analytics systems to increase the quality and timeliness of insight (i.e., systems of record, engagement, and insight)
- Deploying in-memory databases for low-latency response times as part of the application environment
- Infusing data analytics platforms with high-performance technologies to optimize application performance for large data sets
- Using a highly available and secure conduit for data movement between the various application tiers
- Implementing an appropriate data persistence tier that can support the storing, securing, and fast access of rapidly changing data sets
- Preparing for increasingly AI-infused applications and the related data movement and data processing requirements, for example, by introducing hardware acceleration

## The Role of Mission-Critical Platforms for Modern Data-Centric Applications

Firms are increasingly deploying x86-based mission-critical platforms that scale up to optimize their data-driven applications and IT infrastructure transformation. IDC believes that there continues to be a misconception in the market about the merits of scale-up server infrastructure.

### *The Benefits of Scaling Up*

When web infrastructure, collaborative workloads, and application development burst onto the scene some 20 years ago, scaling horizontally became de rigueur. Next, virtualization and cloudification caused the scale-out paradigm to become even more dominant. Along the way, scale-up systems became somewhat misunderstood, even as they were aggressively being modernized.

Business processing, decision support, and analytics have never fared well on horizontally scaled environments. These are demanding workloads that require maximum resources to process multiple terabytes of data. And when these resources — lots of processors that are close together and lots of flat RAM that is globally addressable for in-memory computing — are packaged in a single system, the benefits, compared with scale-out environments, are significant.

The large memory footprints of scale-up systems allow for large and growing databases to be completely held in memory, eliminating the latencies of disk access. Latency is also much reduced because of the use of interconnects that allow for dynamically scaling rather than the complex and extensive networks needed to connect nodes in a scale-out environment. Power consumption and cooling costs are significantly lower, as are software licensing costs.

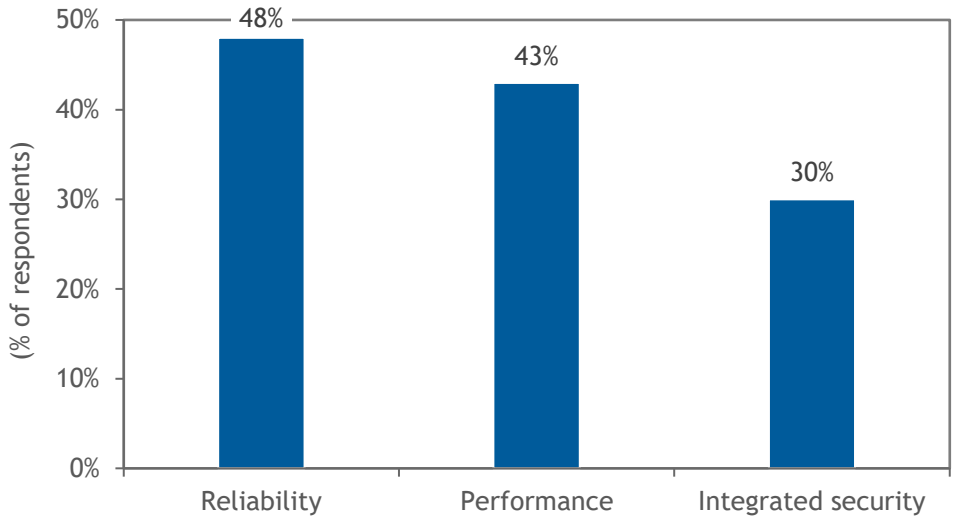
Scale-up systems are also suitable for consolidation projects as they are easier to implement and more efficient to manage and operate than scale-out clusters. They take up a smaller footprint and provide greater reliability and availability (see the Extreme Availability section). In terms of economics, many of today's scale-up systems are nothing like legacy scale-up systems — they leverage the same standardized components (memory, processors, and storage) that scale-out servers are built with, albeit with their own specifications, rather than the proprietary components from the past. The idea that scale-up systems are too costly is simply no longer valid, especially if they are available on a consumption model basis, as many are.

### *Fault Tolerance and Next-Generation Capabilities*

As noted previously, modern scale-up systems leverage standardized hardware, and one trend in scale-up platforms has been standardization on x86. This standardization on x86 for modern scale-up systems has led to another misconception, which is the idea that x86-based platforms cannot be as reliable as the Unix-based systems of the past. This, too, is a myth. In the past five years, IDC has seen scale-up x86-based servers evolve to the highest levels of availability, meeting service levels that are demanded by mission-critical workloads for the data-driven enterprise. They deliver extreme reliability, availability, and serviceability (RAS), even reaching availability level 4 (AL4) in IDC's High Availability framework. Reliability is considered the most important server characteristic (see Figure 4).

## FIGURE 4

### Most Important Server Characteristics




Source: IDC, 2020

IDC classifies servers in four levels of availability, with the highest being availability level 4, or "fault tolerance." At this level, the combination of multiple hardware and software components allows a near-instantaneous failover to alternate hardware/software resources so that business processing continues as before without interruption. In short, modern x86-based scale-up platforms have joined the ranks of the mainframelike fault-tolerant category. This trend is in sync with the increasing desire for zero downtime in today's always-on world.

At the same time, IDC noted that fault-tolerant platforms have been becoming increasingly suitable for the modern datacenter with capabilities such as mobile apps, cloud, APIs, open source software, and next-generation application development.

## A Move to the Cloud — And Back

Cloud adoption continues to accelerate, and now that IaaS and PaaS have matured, businesses are moving some of their critical business applications to IaaS and PaaS environments. While public cloud – IaaS and PaaS – continues to be popular, businesses are also investing more in private cloud (both on premises and off premises) as this allows them to manage a dedicated system for their mission-critical applications and core data that require above-average security and performance while maintaining full control over that system, including its cost.

As a result, businesses are now increasingly managing multicloud environments, often with various cloud service providers. Yet operating these multiple clouds is still difficult, and enterprises often choose to gain more interoperability via a private cloud. According to *Cloud Repatriation Accelerates in a Multicloud World* (IDC #US44185818, July 2018), 80% of businesses interviewed report repatriating some workloads from public cloud environments. Respondents said they expect to move 50% of their public cloud applications to hosted private or on-premises locations. Note that this does not mean that they will consume 50% less public cloud.

IDC does not expect growth in public cloud adoption to slow, but a significant portion of businesses will leverage a private cloud to modernize their large installed base of noncloud applications. Many of these noncloud applications will be their mission-critical applications including core databases, business processing, or apps that execute Big Data and analytics and AI.

## HPE Superdome Flex Family

The HPE Superdome Flex family of servers is made up of the HPE Superdome Flex model, first introduced to the market in 2017, and the HPE Superdome Flex 280 model, introduced at the time of publication. The Superdome Flex family is built upon decades of experience and innovation on scale-up platforms from HPE and Silicon Graphics International (SGI), which HPE acquired in 2016.

The Superdome product line was introduced in 2000 and has since gone through multiple transformations to address emerging market needs. Born on RISC, then transitioned to Itanium, and for many years running the HP-UX operating system (OS), HPE ultimately introduced Superdome X running Linux and Windows. Superdome X addressed the market trend toward standardizing on x86 architectures for mission-critical workloads, delivering a level of reliability previously unseen on standard platforms. With the SGI acquisition in 2016, HPE gained access to SGI's decades-long experience with building some of the most powerful high-performance scale-up systems in the industry, including the SGI UV 300, which HPE started selling as HPE MC990 X.

Superdome Flex was born out of HPE's efforts to combine the best of Superdome X and MC990 X to deliver a modular, standards-based mission-critical system with maximum flexibility, performance, and reliability. According to HPE, the Superdome Flex model has seen strong adoption since its introduction, covering businesses in all geographies and spanning industries from telecommunications to banking, manufacturing, public sector, education, and more. HPE reports that mid to large systems (12–32 sockets) are running in hundreds of production environments, although, notably, the majority of shipped units are four- and eight-socket systems. This reflects both buyer desires to equip for growth and the market requirement for scale-up environments of smaller size. To further address this need, HPE has introduced the HPE Superdome Flex 280 model, starting at a lower two-socket entry point and offering more granular scaling.

### *Noteworthy Characteristics of HPE Superdome Flex*

The HPE Superdome Flex server family features a next-generation multisocket, multicore x86 architecture and is built with HPE's Memory-Driven Computing principles. The new Superdome Flex 280 model features third-generation Intel Xeon Scalable processors, code-named Cooper Lake. The larger Superdome Flex model features second-generation Intel Xeon Scalable processors, code-named Cascade Lake.

#### Unmatched Scale and Flexibility

The Superdome Flex family has a unique modular design that enables firms to start small and scale up as their needs grow. For small to midsize environments, the new Superdome Flex 280 utilizes a 5U chassis and scales seamlessly from two to eight sockets in two-socket increments (up to four sockets per chassis) as a single system. The server is designed to provide 64GB to 24TB of shared memory using DRAM only or in combination with persistent memory. The chassis utilizes a "glueless" architecture and leverages Intel's Ultra Path Interconnect (UPI) links to connect.

As Superdome Flex 280 utilizes third-generation Intel Xeon Scalable processors, there are six available UPI links per CPU, as opposed to three per CPU in second-generation processors. This means higher bandwidth and faster data rates.

For mid to large environments, the HPE Superdome Flex model scales seamlessly from 4 sockets to 32 sockets in four-socket increments and from 768GB to 48TB of shared memory in a single system. The system also features a modular scale-up architecture and a 5U four-socket chassis building block. Within each chassis, however, are two ASICs utilizing eighth-generation HPE technology that connect the chassis to form a high-bandwidth, ultra-low-latency fabric via cabling. This "glued" architecture is unique to HPE and enables Superdome Flex to scale as a single server beyond the eight-socket upper limit of Intel's design. In addition, Superdome Flex can be carved up into nPARs (electrically isolated hard partitions) to physically separate workloads within a common platform.

HPE has developed a diverse product lineup with the HPE Superdome Flex family, leveraging Intel's Gold and Platinum processor variants at various speeds, cache sizes, and core counts. This, in addition to the introduction of the lower entry point Superdome Flex 280, enables businesses to scale and provision to the exact level needed while equipping for growth without having to overprovision.

It's noteworthy that Superdome Flex provides businesses with configuration options that do not require premium-priced high-memory Intel processors. Depending on the amount of memory required per socket, there are two types of Intel Xeon Scalable processor SKUs. The base level allows for up to 1TB per socket, and the L level (only available with platforms featuring second-generation Intel Xeon Scalable processors) allows for up to 4.5TB per socket. If a customer wants a large total memory capacity, Superdome Flex can scale up compute to reach that capacity, thereby avoiding the extra cost associated with the L SKUs. Because other vendors are limited to the eight-socket compute size, they require the pricier SKUs to scale memory. With more processing capacity, Superdome Flex can deliver more memory capacity using lower-priced processors. Spreading memory across more sockets has the additional benefit of increasing the memory bandwidth available to work on the large data sets, rather than restricting access to large amounts of memory behind a fewer number of CPU sockets. Finally, the unique design of Superdome Flex allows the use of Intel Gold SKUs up to 32 sockets while other vendors must use Platinum SKUs at greater than 4 sockets. This allows for a significant number of very useful price/performance combinations.

## Support for Persistent Memory

The Superdome Flex family supports Intel Optane Persistent Memory for HPE. Businesses have the choice, depending on the requirements of their workloads, to run their system either with all DRAM or with a mixture of DRAM and persistent memory. The Superdome Flex 280 model supports the second-generation Intel Optane persistent memory, whereas the Superdome Flex model supports the first-generation of this technology. Intel Optane Persistent Memory for HPE is available on the Superdome Flex family in what Intel calls the "app direct" mode while supporting direct processor load/store access, with speed characteristics that are slower to access than DRAM (especially for writes due to the persistent property supported) but faster than SSDs.

One of the use cases for persistent memory is SAP HANA test and development systems, for which restarts are common and the time lost for data to load can take hours. The main column store (data tables) resides in persistent memory. Because data does not need to be loaded from storage, restart times to bring SAP HANA back can shrink to minutes.

## Unbounded I/O

The Superdome Flex family of products offers ample I/O choice to support a wide variety of workloads. When fully configured, the Superdome Flex 280 model supports up to 32 PCIe 3.0 card slots and the Superdome Flex model supports up to 128 Gen 3 PCIe card slots. These slots can be used for external storage connectivity, hardware accelerators like GPUs (including NVIDIA Tesla and NVIDIA Quadro GPUs), 32Gb Fibre Channel cards, Mellanox InfiniBand, Ethernet cards, NVMe cards, and other peripherals. As HPE does not modify Linux for Superdome Flex, compatibility can be expected with any peripherals running under standard SUSE Linux Enterprise Server (SLES), Red Hat Enterprise Linux (RHEL), and Oracle Linux distributions. Along with the compute capabilities, a highly scalable I/O subsystem enables the deployment of HPC software (which often requires high IOPS and low-latency bandwidth access to storage or accelerator cards).

HPE took great care in architecting the I/O subsystem for the greatest workload benefit. For example, in the Superdome Flex 280, each CPU socket has access to two x8 and two x16 PCIe slots, allowing for powerful, balanced performance. Also, when four full-height, double-width GPU cards are present, each CPU socket also has access to a x16 slot, enabling the fullest possible I/O ingest capability.

## Extreme Availability

Next-generation HPE mission-critical platforms such as Superdome Flex and Superdome Flex 280 are designed to provide "Unix on RISC"-like RAS at a system level, which can be augmented further by using clustering technologies. HPE is a key player in the AL4 market. The Superdome Flex predecessor, Superdome X, is included in the AL4 market, and as the HPE Superdome Flex product family inherits the Superdome X RAS framework, IDC expects it to be classified at this level.

HPE Superdome Flex and HPE Superdome Flex 280 feature many RAS capabilities, including:

- **Firmware First:** This approach ensures error containment at the firmware level, including memory, CPU, or I/O channel errors, before any interruption can occur at the operating system layer. Firmware First covers correctable and uncorrectable errors and gives firmware the ability to collect error data and diagnose faults even when the system processors have limited functionality.
- **Analysis Engine:** This feature reduces human error through predictive fault handling. It monitors resources continuously, predicts hardware faults, and initiates self-repair without operator assistance.
- **Self-healing capabilities:** When faults do occur, Superdome Flex and Superdome Flex 280 provide several mechanisms to avoid unplanned downtime, including disabling failed or failing components during boot and attempting recovery on failed or failing components during runtime.
- **Processor RAS:** Superdome Flex leverages second-generation Intel Xeon Scalable processors, and Superdome Flex 280 leverages third-generation Intel Xeon Scalable processors. These processors include extensive capabilities for detecting, correcting, and reporting hard and soft errors. Because these RAS capabilities require firmware support from the platform, they are often not supported in other industry-standard servers. The Superdome Flex server family implements the full RAS functionality provided in the Xeon Scalable series processors, including corrupt data containment, PCIe live error containment, poison error containment, processor interconnect fault resiliency, and advanced MCA recovery.
- **Memory RAS:** Superdome Flex and Superdome Flex 280 servers use several technologies for enhancing the reliability of memory, including proactive memory scrubbing and Advanced Double Device Data Correction (ADDDC), which HPE enhanced with specific firmware and hardware algorithms to substantially reduce memory outage rates.

- **Platform RAS:** Superdome Flex uses a fabric interconnect scheme featuring adaptive routing capabilities. The system routes traffic down the optimal latency path for performance and provides the ability to route traffic around failing or failed links in the fabric and while the system is running.
- **Application-level RAS:** Superdome Flex and Superdome Flex 280 support Serviceguard for Linux to enable software failover and five-nines availability.

### Multiple Standard Operating Environments

Superdome Flex and Superdome Flex 280 support multiple standard operating environments and virtualization technologies, including SUSE, Red Hat, Windows Server, Oracle Linux, VMware, and KVM. The platform runs on standard, unmodified Linux, which means it supports all the certified stacks from Red Hat and SUSE, including containers and container management software such as Docker and Kubernetes.

### Simplified User Experience

The Superdome Flex family provides a simplified management experience by supporting HPE-specific tools such as HPE OneView, Insight Remote Support, and Proactive Care, as well as the open source Redfish API and OpenStack. Superdome Flex 280 adds an easy-to-use management GUI to further simplify and enhance the user experience.

### Superior Security

The HPE Superdome Flex family of servers has been designed with a security strategy to minimize threat exposure through several mechanisms. These include an "air gapped" manageability subsystem that doesn't allow firmware updates from the operating system, as well as limiting the number of common industry interfaces (that add security vulnerabilities) to only those that customers need. In addition, the HPE Superdome Flex 280 model adds Silicon Root of Trust protection implemented directly in HPE-controlled specialized hardware to enable detection of potentially compromised firmware and prevent its execution.

### Consumption Model

Aligned with HPE's strategy to offer its entire portfolio through a range of subscription, pay-per-use, and consumption-driven offerings in the next three years, the HPE Superdome Flex family of servers can be consumed as a service through HPE GreenLake.

## *Target Use Cases and Workloads*

The Superdome Flex family of servers is designed for mission-critical workloads, in-memory databases, data analytics, high-performance computing, and artificial intelligence. The target workloads include SAP HANA, Oracle databases, Microsoft SQL Server, Epic medical record software, and Unix-to-Linux migrations, as well as high-performance computing and AI workloads that benefit from being run holistically within a single server. The Superdome Flex 280 model targets small to midsize environments, whereas the Superdome Flex targets midsize to large environments.

### SAP HANA

SAP has made its in-memory database, SAP HANA, the foundation of the entire environment for combined analytical and transactional processing under SAP S/4HANA. Superdome Flex's abundant memory and modular architecture make it particularly optimized for SAP HANA environments. With Superdome Flex, HPE is offering performance at scale with optimum cost efficiency.

The larger Superdome Flex model scales seamlessly from 4 to 32 sockets in four-socket increments and provides from 1.5TB to 24TB of shared memory for SAP HANA workloads as a single system. Superdome Flex is SAP certified in the complete range of supported configurations for workloads spanning SoH/S/4HANA and BWoH/BW/4HANA.

HPE Superdome Flex 280 has a lower cost-efficient entry point starting at two sockets and is well suited for smaller SAP HANA environments that are not expected to grow beyond eight sockets. SAP certification is pending and is anticipated to be completed in 3Q20.

### Oracle

Oracle continues to innovate on its core data management product. Oracle Database 18c adds new functionality and enhancements to features previously introduced in Oracle Database 12c including multitenant architecture, in-memory column store, and native database sharding.

Oracle 18c can be configured as a scale-up database or scale-out database using clustering via Oracle RAC. By deploying 18c as a scale-up database on Superdome Flex, businesses can increase their database performance per core, resulting in significant TCO savings from lower licensing costs. Scaling up also simplifies deployment, management, and consolidation. In addition, businesses can add in-memory options for real-time workloads. Because of its ample compute resources and memory abundance, businesses can leverage Superdome Flex to run a mix of transactional and analytic workloads on the same Oracle database simultaneously.

### Microsoft SQL Server

Microsoft has brought the enterprise capabilities of SQL Server to Linux as well as Windows and Docker containers. In addition to now-standard features like advanced analytics and machine intelligence, SQL Server 2017 provides exceptional performance and security. Built-in features enable faster transactions with in-memory OLTP and faster analytics with In-Memory Column Store. PolyBase enables easy querying across the SQL Server and data stored in Hadoop. Superdome Flex is ideal for critical SQL Server workloads on bare metal or virtualized server deployments. It is also suitable for database consolidation and migration initiatives, where the target database is SQL Server, and for cases where businesses need very high reliability levels for their critical SQL Server workloads.

### Unix-to-Linux Migration

The Superdome Flex family of servers is ideal for firms that want to standardize on x86-based compute infrastructure but do not want to compromise on performance or RAS. With support for standard operating environments and virtualization technologies, firms get a wide set of options for migrating their mission-critical databases and workloads from Unix systems.

### In-Memory High-Performance Computing

Superdome Flex equips scientific, engineering, and other technical computing environments with the ability to solve complex, data-intensive problems at extreme scalability with "single-system simplicity." These types of problems are often challenging to distribute across multiple nodes in an HPC cluster and benefit from "fat" nodes (more processors and memory) within a cluster. They include CAE, genomics, fraud detection and prevention, and large data visualization. With the introduction of Superdome Flex 280, customers now have even more flexibility sizing these nodes.

### Artificial Intelligence

Complementing HPE's Apollo 6500 for AI workloads, Superdome Flex provides the end-to-end acceleration of AI workloads that a single system can achieve. For example, organizations can outfit Superdome Flex with abundant Ethernet for ingesting data, have CPUs and/or GPUs run AI training or inference on the data sets, keep all the data in memory, and have a number of (unmodified) applications in a workflow pipeline so that they can pipe the data from stage to stage through an in-memory file system.

Another benefit for AI is the system's huge memory capacity within a single OS. The memory footprint of accelerators can be limiting (32GB for the latest GPUs), making it difficult to process, for example, large numbers of very large, high-resolution images at speed. With the Superdome Flex's terabytes of memory, these restrictions are less noticeable. HPE is focusing on genomics, analytics (graph and Big Data), and risk management (Monte Carlo simulations for FSI), leveraging the platform's memory-driven computing.

In addition, Intel has been adding AI capabilities to its Intel Xeon Scalable Processor line. The first generation, code-named Skylake, added performance for AI inferencing without changing the hardware, just through software libraries and framework optimizations, including Caffe2. The second generation, code-named Cascade Lake, introduced a new AVX-512 extension called Vector Neural Network Instructions, which Intel marketed as "DL Boost," to accelerate AI inferencing on the processor. With the third generation, code-named Cooper Lake, Intel adds an extension to the Intel DL Boost Instructions: bfloat16 numerical format support in AVX-512 with embedded acceleration in the CPU. The HPE Superdome Flex 280 leverages the third-generation Intel Xeon Scalable processors and will support this feature.

## CHALLENGES/OPPORTUNITIES

IDC believes that firms are converging their systems of record, engagement, and insight as they advance on their journey to become data-driven enterprises. As a part of this journey, many firms are also standardizing on x86-based infrastructure, even for very data-intensive, mission-critical workloads. The move to the cloud for certain workloads continues unabated, but IDC is also seeing a distinct repatriation of workloads back on premises, notably mission-critical workloads. These workloads run on private clouds as organizations develop a multicloud data-processing approach, and mission-critical infrastructure will play an important role in these private clouds. Indeed, IDC believes that businesses will continue to invest in critical infrastructure platforms that enable them to accelerate their data-driven journey.

IDC believes that a growing appetite for scale-up x86-based platforms such as the Superdome Flex family will continue to command traction among organizations that require:

- Scale-up multisocket design for high-performance scaling
- Security, availability, and reliability for mission-critical deployments
- Flexible and modular design for opex-friendly deployments
- Optimizations for in-memory databases and real-time analytics applications
- Support for an open standards-based and cloud-ready design for deploying hybrid IT
- As-a-service consumption while maintaining on-premises control

The Superdome Flex family is made up of powerful scale-up x86 platforms that combine the best of HPE's mission-critical reliability and SGI's scalable technology. It has been optimized for high-end performance at scale, in-memory databases, and a range of high-availability features throughout the platforms — both hardware and software. It can handle the most demanding workloads quickly and without interruption. Because of its scale-up architecture, Superdome Flex also provides TCO efficiencies that, after a decade of x86 server sprawl and soaring opex in the datacenter, are in high demand.

For HPE, the opportunity lies in providing all the elements of a modern infrastructure environment in which fault tolerance truly matters. It is about positioning Superdome Flex as a family of offerings that:

- Are flexible and powerful enough to handle the massive and growing amounts of data moving through a modern business
- Provide the ability to analyze data from the core to the edge in real time with an optimized in-memory design
- Are modular and cloud ready and the right fit for any business of any size that is pursuing a traditional private cloud or hybrid IT design
- Are well equipped for applications that require an AI inferencing component

Enterprises are also embracing a world in which app developers would want a vibrant open source ecosystem in which they can develop complex, stateful apps that depend on the hardware to maintain their state, sometimes in multiple stages. Stateful apps expect the hardware to not fail, and stateful apps in many industries may have compliance requirements that mean they cannot fail.

Here, HPE should also enlist the developer community and ensure that Superdome Flex stays open and remains developer friendly. HPE should also make sure that its systems are capable of sustaining state without performance downgrades through superior compute, fabric, and storage components.

## CONCLUSION

While traditionally mission-critical systems have represented a smaller part of the server market, they are well poised to grow in new areas as next-generation data analytics, in-memory databases, and AI inferencing and the expansion into the HPC space increase the market demand for this type of platform. With Superdome Flex and now the Superdome Flex 280 model, HPE reaffirms its commitment to delivering high-end x86-based systems for mission-critical workloads running standard operating environments, and with the latest release, HPE demonstrates that it continues to ensure that the platform is ready for emerging workloads such as AI inferencing. Superdome Flex differentiates itself by being a modular and flexible x86-based mission-critical platform. HPE should no longer have to convince prospective customers of the RAS features of Superdome Flex for hybrid IT deployments; the next stage for the mission-critical platforms is to demonstrate how well they can execute AI inferencing, which is becoming an integral part of running data-driven applications. HPE is well-positioned to be the vendor that shows the market how these dynamics can play well together.

## About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

## Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com

---