quark lepton

WHITEPAPER / August 15, 2021 / QL544018

Comparing IBM Power Systems and Oracle Exadata Database Machine for Transaction Processing

Market Situation

Server and storage architectures continue to evolve to better address the growing size, complexity, and diversity of data workloads. While analytics, artificial intelligence (AI), and multicloud computing topics tend to dominate today's headlines, transaction processing infrastructures continue to be the driving force serving core business operations. At the same time, in-memory software advances and processor innovations have prompted a new generation of computational paradigms and associated platform architectures.

Transactional workloads drive fundamental requirements for speed, availability, concurrency, and recoverability. Organizations depend on immediate and accurate execution of business processes for a range of applications, such as e-commerce and enterprise resource planning (ERP) as well as retail point-of-sale, consumer banking, and reservation systems. Transactional systems continue to evolve to meet the growing demands of new applications and technologies.

This paper examines the fundamental technology differences between two distinct server platforms in support of online transaction processing (OLTP) workloads: the POWER9-based IBM Power Systems platform, running the AIX operating system and Oracle Database; and the Oracle Exadata Database Machine, an engineered system that essentially consists of a networked Intel x86-based compute and storage server bundle running Oracle Linux and Oracle Database.

Comparisons include estimated three-year costs for use of IBM Power Systems with Oracle Database and Oracle Exadata Database Machine as platforms for facilitating enterprise OLTP applications. Cost estimates draw upon the experiences of organizations that have deployed the Oracle Database in support of enterprise OLTP applications on these platforms.

A broad survey of 135 organizations representing a variety of industries was used to construct 10 installation profiles for manufacturing, healthcare, distribution, agribusiness, engineering and construction, media, retail, and IT services companies. Companies ranged from 100 to 18,000 employees with revenues up to $11 billion.

Costs for use of IBM Power Systems compared to Oracle Exadata Database Machine as the underlying infrastructure for running Oracle Database averaged 37 percent less (Figure 1). Hardware employed for comparisons includes Exadata X8M-2 and Exadata X8M-8; and IBM Power Systems S914, S922, S924, E950, and/or E980 servers with IBM FlashSystem storage arrays.

Both Exadata and Power Systems use Oracle Database 19c as the foundation for applications such as Oracle E-Business Suite, PeopleSoft, and/or financial solutions. Principal cost differences between the two platforms are driven by factors such as server architecture and software licensing policies.

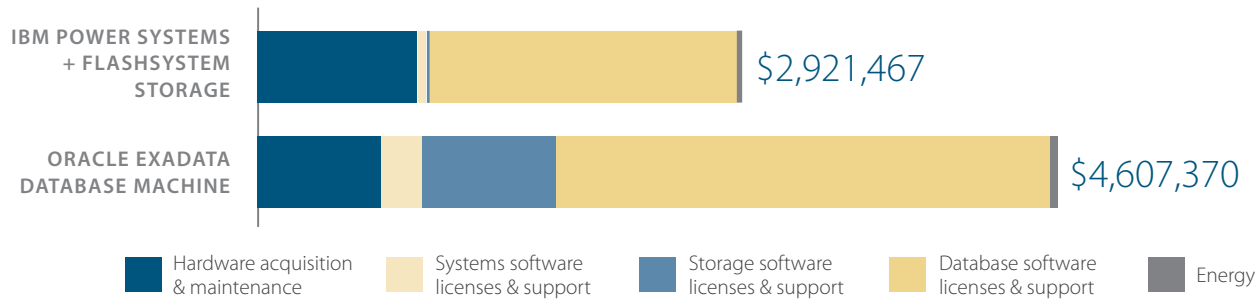
The POWER9 processor-based Power Systems platform, designed for flexibility and scalability, includes a family of standard server models optimized for traditional enterprise workloads such as database, application serving, and e-commerce. All workloads can, in fact, be run on a single system through the utilization of logical partitions. Organizations benefit from robust reliability, availability, serviceability (RAS), and security features offered by Power Systems with PowerVM and AIX in addition to highly granular workload management and innovative autonomic features.

The Oracle Exadata Machine is an engineered system designed to run Oracle databases only and does not support other workloads. Employing a split-tier architecture, Exadata integrates scale-out database and storage servers with a Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) internal network that connects all servers and storage. Specialized database functionality resides with each server tier and all server and storage resources share the single network. Exadata supports all types of database workloads including OLTP, data warehousing (DW), and in-memory analytics. Application workloads must be run on separate systems such as the Oracle Private Cloud Appliance or third-party servers such as IBM Power Systems or x86 servers.

IBM Power Systems Technology

IBM introduced the world's first multicore processor in 2001. The POWER4 dual-core processor, comprising more than 170 million transistors, was a breakthrough in architecture and semiconductor engineering and allowed two processors to work together at a very high bandwidth with large on-chip memories and high-speed buses and I/O channels. The industry response at the time— "this is

FIGURE 1: Average Three-year IT Costs by Platform for OLTP Workloads

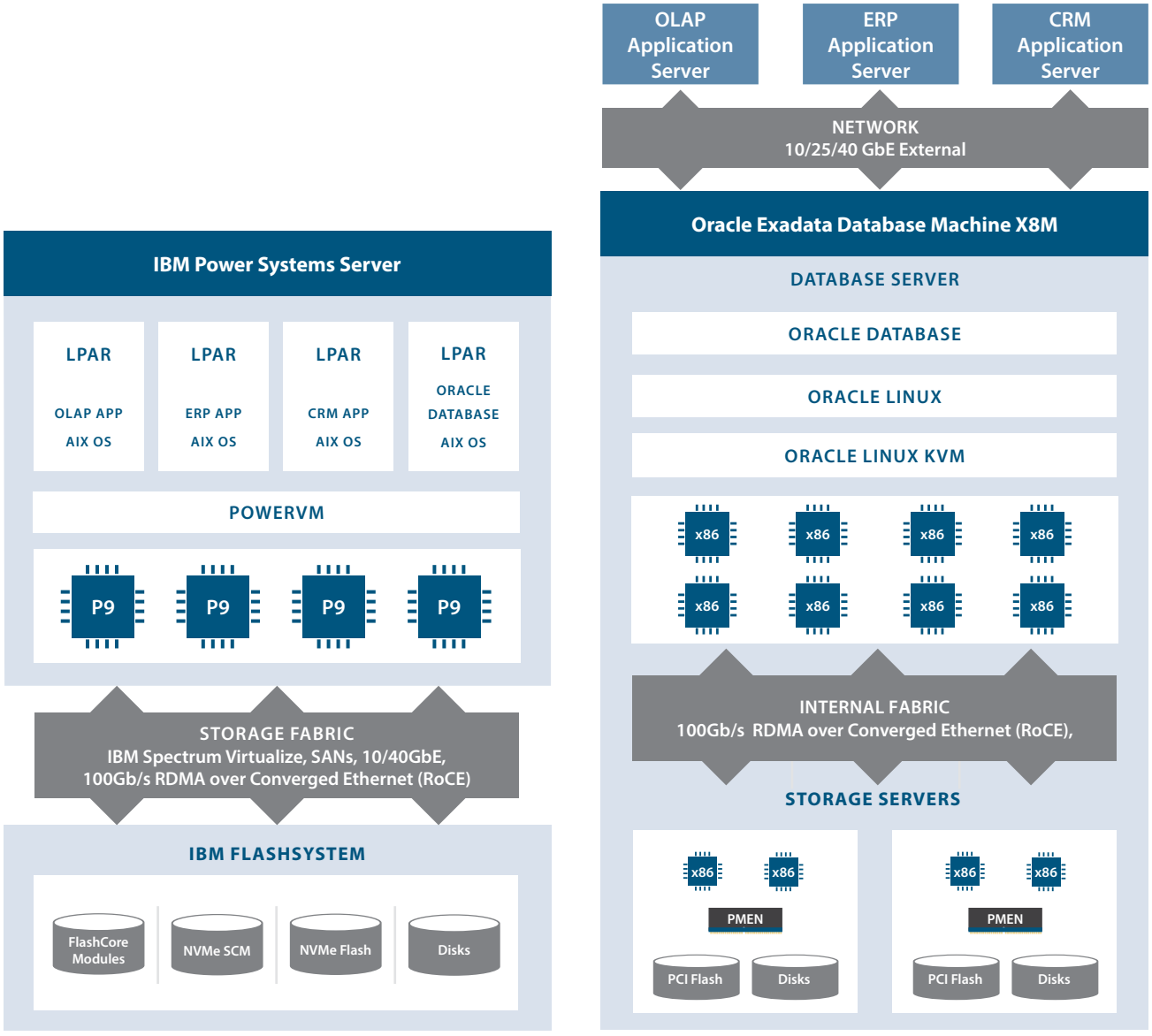


SOURCE: Quark + Lepton (August 2021)

aggressive" and "amazing technology"—led to broad customer acceptance of Power Systems as a leading general-purpose UNIX server.

Power Systems technology has evolved significantly since the introduction of POWER4. With each new generation of technology, IBM expertise in processor design and manufacturing, supercomputer and mainframe design, and operating system and compiler optimization has focused on providing broad support for modern workloads with industry-leading, high-performance computing and data throughput. Power Systems architecture and flexible component-level interconnectivity is illustrated in Figure 2.

FIGURE 2: IBM Power Systems and Oracle Exadata Database Machine Designs



SOURCE: Quark + Lepton (August 2021)

IBM introduced POWER9 processor technology in late 2017 and POWER9-based servers a few months later. The POWER9 processor comprises eight billion transistors and supports up to 12 cores with 96 threads through simultaneous multithreading (SMT8). With eight threads per core, it is an absolute beast for data-intensive workloads. POWER9-based systems include a broad family of servers optimized for different workloads. This paper will focus on those optimized for the PowerVM AIX ecosystem and traditional enterprise data-intensive workloads.

The scale-out servers are designed for traditional data center clusters utilizing single- and dual-socket implementations. The scale-up servers, designed for high-end systems with four or more sockets, support large amounts of memory capacity and throughput.

IBM POWER9-based, scale-out offerings include the Power System S914, S922, and S924 servers. These cloud-enabled servers with integrated PowerVM Enterprise capabilities incorporate strong security and reliability innovations. In July 2020, Power Systems scale-out servers were updated with new Peripheral Component Interconnect Express Gen4 (PCIe4) input/output (I/O) switches and additional NVMe adapter support. These upgrades are designed for high-performance and low latency computing to satisfy the demands of varied complex workloads, such as in-memory databases, cognitive, and cloud-enabled applications.

The Power S914 server supports a one single chip module (SCM) socket and may be configured with either a 4-,6-, or 8-core POWER9 processor module. Core speeds range from 2.3 to 3.8 GHz. Each server has 12 PCIe4 slots and supports up to 1 TB of DDR4 memory. This granularity extends throughout the product line and is shown in some detail in Table 3 on page 12.

The Power S922 server supports up to two SCM processor sockets configured with 8-, 10-, or 11-core POWER9 processors and up to 4 TB of DDR4 memory. The Power S924 server supports two SCM processor sockets which can be configured with 8-, 10-, 11-, or 12-core POWER9 processors and up to 4 TB of DDR4 memory.

IBM's current scale-up offerings for the POWER9-based servers became available in Fall 2018 and include the E950 and E980 servers. These large-scale, cloud-enabled symmetric multiprocessing (SMP) servers with built-in virtualization and flexible capacity are designed to deliver enterprise-class performance, scalability, availability, and security for large environments.

The Power E950 server has two, three, or four SCM processor sockets with 8-, 10-, 11-, or 12-core POWER9 processors, providing up to 48 processor cores per system. The E950 can be configured with up to 16 TB of DDR4 memory and has 10 PCIe4 slots.

The Power E980 system may be comprised of up to four nodes. Each node has four SCM processor sockets, supporting 8, 10, 11, or 12 POWER9 processor cores. Four interconnected nodes create a 16-socket system with up to 192 processor cores, 64 GB of memory, and 32 PCIe4 slots.

IBM also announced two important POWER9-based system implementations in 2018 that are worthy of review here. Together with industry partners, IBM designed and implemented two supercomputer systems that feature a unique architecture combining HPC and AI capabilities. While these systems are

not focused on database workloads, they are illustrative of early POWER9-based system acceptance for high performance computing environments.

In June 2018, the US Department of Energy unveiled the world's fastest supercomputer. The new system, called "Summit," is eight times more powerful than "Titan," which had previously been the fastest supercomputer in the US. Designed by IBM and Nvidia, Summit is powered by IBM POWER9 processors and NVidia's Tensor Core GPUs. At 200 petaflops, the compute cluster consists of about 4,600 nodes, with two CPUs and six GPUs per node, for a total of 9,200 CPUs and 27,648 GPUs.

A few months later in October 2018, IBM announced that "Sierra" was operational at Lawrence Livermore National Lab. Designed by IBM and Nvidia, Sierra is comprised of 4,320 nodes and reaches a performance of 125 petaflops. Each node consists of two IBM POWER9 CPUs and four Nvidia V100 GPUs. Sierra is supported with 154 petabytes of IBM Spectrum Scale storage and delivers 1.54 TB/s of I/O performance.

Oracle Exadata Database Machine Technology

Today's Oracle Exadata Database Machine began as an internal development project for an intelligent storage solution that would improve the performance of data warehouse workloads using Oracle software and an open hardware stack solution. Introduced as a commercial solution in 2008, Exadata's engineered system design has evolved with only limited innovation over the years and is largely reliant on Intel commodity technology improvements. This reliance on commodity high-core count processors limits performance across a spectrum of workloads and increases Oracle software license costs.

With a split-tier architecture, the Oracle Database and Real Application Cluster (RAC) workload is distributed across two separate subsystems, each with its own compute nodes and supporting memory infrastructure. Oracle Database and RAC processing is handled in the Database Server subsystem while the Storage Server subsystem takes on the I/O processing. New with Exadata X8M, internal communication between these subsystems is accomplished using 100Gbps RDMA over Converged Ethernet (RoCE) network fabric. Previous generations of Exadata, up to and including the X8, had InfiniBand network fabric. External connectivity is supported with 10Gbps and 25Gbps Ethernet. The inherent complexity of this split-tier architecture should not be underestimated and is illustrated in Figure 2 on page 3.

The Exadata Database Machine runs Oracle Linux, a full Linux distribution based on Red Hat Enterprise Linux, that has been modified to accommodate specific Oracle software and hardware requirements. Oracle Linux is available free and open source; however, a support contract with Oracle is required for related implementation, performance, and operating issues. Oracle also offers maintenance services for Exadata Database and Storage Server hardware as well as support for Oracle Database and Exadata Storage Server software.

Exadata Storage Servers include a collection of specialized technologies created over the years to assist with I/O processing communication challenges. Specifically, Hybrid Columnar Compression (HCC), Smart Scan, Storage Indexes, Smart Flash Cache, and Write Back Flash Cache are storage server-based

features that together play a critical role in working to minimize traffic over the internal network fabric. These technologies contribute to online analytical processing (OLAP) optimization, but provide only minimal, if any, benefit to OLTP workloads. Exadata's storage subsystem remains essentially a "black box," as only Oracle Database takes advantage of these features. In addition, Exadata Storage Servers are only available to workloads within the Exadata system in which they reside.

Exadata lacks certain advanced storage capabilities, such as hardware compression and encryption, storage replication and FlashCopy, that IT organizations have come to rely on for efficient operations. The system is also missing hardware-level data protection, relying instead on mirroring for availability requirements. Storage capacity requirements can easily grow substantially as organizations implement double or even triple mirroring to meet availability requirements.

Exadata's split-tier Database Server / Storage Server design has been maintained from the X2 through the current X8M generation. The hybridization of its data storage format and processing, supporting a unique I/O read/write hand-off from Database Server compute nodes to Storage Server compute nodes, remains at the heart of the system's design.

This specialized hybridization undermines the practicality of hosting large enterprise database-dependent applications on the Exadata platform directly. Oracle does not support applications on Exadata, and organizations are encouraged to purchase Oracle's Private Cloud Appliance or other Oracle solution at significant additional cost if they wish to deploy new applications. The overall result of this hybrid architecture does not significantly benefit OLTP workload processing performance and may only marginally improve mixed workload performance. Commercial ISV applications will not necessarily benefit from the Exadata specialized technologies without required structural changes within the applications and this is typically not allowed by ISVs.

Exadata X8 generation machines became available in April 2019 and included two models—the 2-socket X8-2 and large-scale 8-socket X8-8. Only six months later, in October 2019, Exadata X8 was essentially replaced by newly-designed Exadata X8M equivalents. Unlike previous generations, X8 models were not upgradeable.

Exadata Database Machine X8M introduced Intel Optane DC Persistent Memory (PMem) and 100 Gbps RoCE Network Fabric. These changes prevent integration with any previous generation Exadata components, which can be disruptive to a typical IT organization.

The Exadata X8M-2 Database Server consists of two 24-core Intel Xeon 8260 2.4 GHz (HT2) processors and 384 GB memory—expandable to 1.5 TB—and provides a total of 48 processor cores. Exadata X8M-8 Database Server includes eight 24-core Intel Xeon 8268 2.9 GHz (HT2) processors and 3 TB of memory—expandable to 6 TB—and provides a total of 192 processor cores. Both the X8M-2 and X8M-8 processors support two threads per core with hyperthreading.

Exadata Storage Servers are available in two primary configurations—High Capacity (HC) or Extreme Flash (EF). HC Storage Servers include four NVMe PCIe3 Flash cards, each with 6.4 TB Exadata Smart Flash Cache and twelve 14 TB 7,200 RPM disks. EF Storage Servers include eight NVMe PCIe3 Flash cards, each with 6.4 TB storage capacity for an all-Flash configuration. Both HC and EF storage servers are

powered by two 16-core Intel Xeon 5218 2.3 GHz processors. Exadata X8M HC and EF Storage Servers include Intel Optane DC PMem modules for increased performance. Each server supports 1.5 TB of PMem which serves as a new memory tier between DRAM and flash.

A third storage option—Exadata X8M Extended (XT) Storage Server—was introduced with Exadata X8. This new option enables customers to extend Exadata data management to long-term or low use data retention requirements. Each XT Storage Server includes twelve 14 TB SAS disks with 168 TB total raw capacity, and has no flash storage. The XT Storage server is powered by one 16-core Intel 5218 2.3 GHz processor.

The minimum configuration for an Exadata system includes two database servers and three storage servers and can be expanded as needed.

Cost Comparison

This detailed cost section presents total cost of ownership estimates derived from survey results and summarizes key differentiators responsible for driving cost variations between these two database platforms. Configurations for Exadata and Power Systems in this report used the latest-generation models and were selected based on estimated comparative workload performance.

Overall results from cost comparisons are summarized in Figure 1 on page 2. Even allowing for aggressive Oracle discounting, costs for use of Power Systems for all solutions averaged 37 percent less than for Exadata systems.

Costs for both platforms include system acquisition and maintenance; systems, storage, and enterprise software licenses and support, including operating systems, Oracle Database 19c, clustering, virtualization, diagnostics, tuning, and other appropriate Exadata options' costs; and energy costs.

Key drivers in the higher costs for using Exadata are software licensing and support (Table 1). Systems software license and support costs average 80 percent less for use of Power Systems compared to

TABLE 1: Average Three-year Costs for Transaction Processing Systems

COST CATEGORY	IBM POWER SYSTEMS + FLASHSYSTEM STORAGE	ORACLE EXADATA DATABASE MACHINE
Hardware acquisition & maintenance	984,047	650,372
Systems software licenses & support	48,365	244,951
Storage software licenses & support	15,322	694,876
Database software licenses & support	1,835,864	2,963,818
Energy	37,869	53,353
THREE-YEAR TOTAL ($)	**2,921,467**	**4,607,370**

SOURCE: Quark + Lepton (August 2021)

Oracle Exadata, enterprise software license and support costs average 38 percent less, and storage software license and support was 98 percent less for Power Systems.

Comparisons are between latest-generation Exadata X8M-2 eighth-, quarter-, half-, and full-rack models and X8M-8 half- and full-rack models with Oracle Linux, Oracle Linux KVM hypervisor, and Exadata EF or HC storage servers; and IBM POWER9-based Power System S914, S922, S924, E950, and E980 server models with the AIX operating system, PowerVM virtualization, and IBM FlashSystem storage.

Comparisons are for production systems only. Although shared production/non-production systems are often deployed in Power Systems environments, many organizations employed separate systems for test, development, quality assurance, other non-production functions, and backup. All calculations except those for energy consumption are based on industry standard discounted prices as reported by users. Energy costs are based on system specifications and national average cost per kilowatt hour (kWh).

Personnel costs for database administrators (DBAs) and application specialists were similar and are not included. There was also no substantive difference in full-time equivalent (FTE) staffing for server and storage administration tasks. However, Exadata DBAs tend to require additional skills that DBAs working with less specialized systems do not require.

Packaging and Deployment

Cost differences also reflect the packaging and deployment options for Oracle Database on Oracle Exadata and Power Systems.

Oracle Exadata X8M-2 eighth-, quarter-, half-, and full-rack models are configured with 48, 96, 192, and 384 Intel database server cores, respectively. Exadata X8M-8 half-and full-rack models both have 384 database cores. One additional database server can be added to the rack configuration for a maximum of 576 cores. CoD can be used to deactivate certain numbers of Exadata database server cores as a means of reducing total cost of ownership (TCO). This, however, can only be done at initial deployment and without further flexibility.

In contrast, Power Systems servers offer more granular configurations of 4, 6, 8, 12, 16, 24, 32, 44, 48, and 192 cores, while individual cores can be activated with CoD as available and as needed. In addition, and as discussed earlier, Power Systems servers are not limited to database workloads and can run application workloads as well.

Lack of CoD for Exadata storage servers, coupled with mirroring recommendations, exacerbates factors increasing licensing costs. Oracle recommends the triple-mirroring high redundancy setting of Oracle Automatic Storage Management (ASM) to ensure optimal high availability. If Data Guard is enabled, ASM normal redundancy (double mirroring) can be used. When using either double or triple mirroring, organizations are required to license the full physical capacity of their Exadata storage. In fact, except for eighth-rack configurations, all storage must be licensed independent of the mirroring configuration.

This requirement significantly drives up costs based on the usable mirrored capacity, although the usable storage capacity is effectively a fraction of the total storage capacity.

Storage software licensing prices for Exadata storage servers are listed as $10,000 per disk and $20,000 per flash drive. A full X8M rack can contain up to 14 storage servers with either eight NVMe flash drives (EF) or 12 disks and four flash drives (HC). Storage licensing costs, before discounts, for a full rack can reach $2.24 to $2.8 million.

Power Systems compatibility with a wide range of storage solutions also affects pricing. Storage hardware and software can be supplied by IBM or other vendors and tailored to customers' needs. An organization's existing storage solutions may also be used. For example, the IBM FlashSystem 5200, when configured as a fully populated 1U enclosure, provides the same capacity as a Exadata Storage Server HC half-rack configuration, and requires no additional software costs.

Higher Exadata X8M core counts, compared to Power Systems, as well as previous generation Exadata systems, result in higher system and enterprise software licensing costs. Power Systems can also support additional operating systems, such as enterprise Linux distributions from Red Hat, SUSE, and Canonical as well as system management and tuning tools from IBM.

Differentiation

The different design strategies implemented by Oracle and IBM in these two Oracle Database server platforms are explored below within the context of performance, virtualization, storage, security, and RAS. These differences are summarized in Table 2.

DATABASE PERFORMANCE

A review of OLAP and OLTP data processing requirements reinforces the differences between Exadata and Power Systems from a database-hosting performance perspective. OLAP workloads typically involve complex queries targeting a large data set and require heavy read I/O and aggregation of data. OLTP workloads are characterized by a large volume of frequent, much smaller I/O requests. Performance for OLAP workloads tends to focus on improving throughput, whereas fast response times are desirable for OLTP workloads.

Recent developments in memory and I/O technologies have further increased the performance of Power Systems. Designed with performance-boosting database, hardware, and storage features, POWER9 offers a 1.5x increase in performance, 2x to 4x more memory per system and 1.8x more memory bandwidth as its POWER8 predecessor. Power Systems are intended for use as a single hosting platform for a broad mix of enterprise software and integrate easily into an existing IT environment.

To address the specific needs of OLTP workloads, POWER9 processors feature a new modular architecture and provide up to 24 cores per socket. This represents a 2x increase in available cores per socket in comparison to POWER8. POWER9-based servers feature 2x the number of available DIMM slots and 2x to 4x more memory capacity per system when compared to its predecessor. The Power S922 and

TABLE 2: Principal Differences Between Oracle Exadata Database Machine and IBM Power Systems with FlashSystem Storage

	ORACLE EXADATA DATABASE MACHINE	IBM POWER SYSTEMS & STORAGE ARRAYS
Design	Specialized engineered system designed specifically to run only Oracle Database & is comprised of scale-out database & storage servers, & system software. A change in network fabric in X8M systems has made these systems incompatible with all previous Exadata systems including X8. OLTP, data warehousing, analytics, & mixed workloads may be run on Exadata.	General-purpose server & storage array designs. Servers support high I/O throughput, memory bandwidth, & memory capacity; offer strong core performance; & are designed for horizontal & vertical scaling. May be configured & optimized for simultaneous hosting multiple types of workloads & applications. Processors & systems are designed for data-intensive workloads such as databases (Oracle, SAP HANA, IBM DB2, & others) & analytics.
Configurability	All machines are prebuilt & configured identically. Minimum configurations may be expanded into elastic configurations, adding more database &/or storage servers. With few exceptions, hardware, software, & operating systems may not be modified. CoD is available only for database server cores. After initial installation, active database server cores may only be increased, not decreased. All decisions about sizing must be made prior to installation.	Server & IBM or 3rd-party storage arrays are separately configurable with high levels of granularity. Multiple operating systems & applications are supported on the same server. CoD is available on the E950 & E980. CUoD allows the dynamic activation of additional permanent processor or memory capacity when needed. Elastic CoD (ECoD), temporary activation of processors or memory, may be enabled as needed. Processors & memory may be turned off/on an unlimited number of times. To optimize TCO, Power Private Cloud with Shared Utility Capacity (formerly Power Enterprise Pools 2.0) supports multi-system resource sharing & by-the-minute consumption of on-premises compute resources. Power Private Cloud with Dynamic Capacity extends this support to clients with a private cloud infrastructure on Power Systems.
Availability optimization	Maximum Availability Architecture (MAA) includes standard Intel hardware RAS features, as well as redundant hardware & software for each machine. Oracle software is designed to protect against any single points of failure in hardware. Oracle Database 19c HA features include Flashback, RAC, sharding, & application continuity. Mission critical HA requirements include a second Exadata system as well as activation of multiple database options. Automatic Storage Management (ASM) mirroring capabilities support multiple RAID levels. Triple disk mirroring is recommended.	An integrated system design with advanced hardware RAS & operating system-level availability optimization features minimizes application outages as do redundant power/cooling subsystems & concurrent repair of these. No single points of failure in hardware. Error detection & fault isolation capabilities enable retry & other mechanisms to avoid soft error outages & to allow for use of self-healing features. Live Partition Mobility (LPM) allows planned outages to occur on a server without downtime. Servers & storage arrays support multiple RAID levels, all Oracle Database 19c HA solutions as well as IBM & third-party solutions for HA clustering, failover, & recovery.
Backup & restore procedures	Oracle Recovery Manager (RMAN) software running on database server nodes is used to create backups. Database backups may be directed to Recovery Appliance X8/X8M, Oracle ZFS Storage Appliance, local Exadata storage, Oracle Cloud Infrastructure Object Storage, or other storage device.	RMAN backup & recovery utility is supported & interfaces with IBM storage product Spectrum Protect for the management of physical storage. Storage-level database backup & extended SAN features, including snapshots, remote mirroring, storage tiering, LAN-free backups, & thin provisioning are supported.
Security	Oracle Advanced Security's encryption & redaction policies meet FIPS 140-2 Level 1 security requirements. Multiple Oracle Database security features & options may be activated. Oracle virtualization technology has numerous vulnerabilities reported in the NIST database. Full compliance with all provisions of PCI-DSS, HiPAA, EU GDPR, & others.	Includes the ability to activate the same Oracle Database security measures as in an Exadata environment. Enables KMIP-compliant self encrypting storage. Integrates with IBM PCIe Cryptographic Coprocessors to achieve FIPS 140-2 Level 4 cryptographic security. PowerVM has zero vulnerabilities as reported by the NIST vulnerability database for the analysis period. Data in motion & at rest on server is encrypted. VMs are encrypted & run in secure memory using POWER9's Protected Execution Facility. Storage-level security includes hardware-accelerated AES-XTS 256 encryption for data-at-rest.

SOURCE: Quark + Lepton (August 2021)

S924 support up to 4 TB memory, the E950 16 TB, and the E980 64 TB. This contrasts with the Exadata X8M-2 database server with 1.5 TB and X8M-8 with 6 TB.

For analytics workloads, the increased number of POWER9 cores supported by higher I/O bandwidths provide industry-leading performance levels for OLAP workloads as well. For example, the E950 delivers 230 GB/s memory bandwidth per socket. POWER9 is the first processor built with PCIe4 I/O adapters and infrastructure, delivering twice the I/O bandwidth of PCIe3.

POWER9 processors continue to improve on features first introduced with POWER7 in 2010, enhancing mixed workload management capabilities. Intelligent Threading allows workloads to be executed using from one to eight threads per core. The system can automatically determine which to use for optimum performance, or system administrators may select the number of threads employed. This technology is further enhanced by a new core microarchitecture introduced with POWER9. Intelligent Cache allows systems to dynamically vary cache utilization as workload characteristics change. The system can automatically determine the appropriate level of cache for specific workloads. Continuous performance optimization is provided for both features.

Exadata, in contrast, utilizes a different approach to I/O processing. Each Exadata rack connects compute servers and storage servers over an internal network. This split architecture aims to minimize data transfer through I/O processing done by storage servers using Smart Scan software. Smart Scan allows the majority of OLAP-specific SQL processing to occur in the storage tier, as opposed to the database tier, maximizing efficiency when processing large data sets. This offers only minimal benefit for OLTP-specific SQL processing.

While Oracle has employed unique hardware and software technologies to increase performance in data warehouse and analytical workloads, Exadata's hybrid structure does not necessarily offer specific performance benefits for OLTP workloads. OLTP operations do not involve reading and writing large tables or aggregates. OLTP queries typically involve INSERT, UPDATE, and DELETE transactions for individual records. As a result, storage cores may remain largely dormant when facilitating transaction-intensive database activities.

Exadata X8M-2 racks offer up to 448 cores for SQL offload. For OLTP workloads, however, the compute capabilities these cores provide are dramatically underutilized. The processors in both X8M-2 and X8M-8 database servers support two threads per core. In contrast, IBM's POWER9 cores support eight threads per core, dramatically boosting the number of instructions that can be executed concurrently. In support of multithreading, it should be noted that the sharing of cache among threads can affect system memory performance. POWER9 processors contain larger caches than the processors used in Exadata database servers. A larger cache contributes to better performance, as multithreading can increase the potential for cache misses. There is evidence suggesting that a three times performance decrease may occur when the working set exceeds the L3 cache size. For mixed workloads, a larger cache is beneficial for OLTP tasks.

Exadata X8M specifically targets performance improvements for OLTP workloads by reducing latency and increasing input/output operations per second (IOPS). Exadata X8M allows the database server to

access storage server memory directly over the new 100Gbps RoCE internal network. Exadata X8M also added PMem as essentially another layer of cache memory within the storage server. This non-volatile memory cache together with RoCE are said to reduce latency by 10x and increase IOPS dramatically. These X8M improvements should improve OLTP performance but it is not clear yet to what extent.

Table 3 provides detailed specifications for both IBM Power Systems' POWER9 processors and the Intel Xeon processors used in the latest Oracle Exadata Database Servers.

Oracle continues to place a strong marketing emphasis on another performance feature, Exadata Smart Flash Cache (ESFC). However, performance gains achieved from using ESFC vary depending on workload and configuration and will typically provide only minimal benefit to OLTP workloads. Adding data to ESFC is typically faster than writing to disk, but it is significantly slower than latency associated with writing to memory. Exadata Storage Server, by default, stores only small I/Os in ESFC. Data requested by table scans will not be stored in ESFC unless manually configured to do so. ESFC is also a tertiary cache, used only when requested data cannot be found in the local buffer cache or the RAC

TABLE 3: Specifications for IBM POWER9-based Power Systems and Oracle Exadata Database Machine Database Servers

FEATURES	IBM POWER SYSTEMS				ORACLE EXADATA DATABASE SERVERS	
	E980	E950	S924	S914	X8-8 / X8-8M	X8-2 / X8-2M
Processor	*Power 9040-M9S*	*Power 9040-MR9*	*Power 9009-42G*	*Power 9009-41G*	*Intel Xeon Platinum 8268*	*Intel Xeon Platinum 8260*
Sockets/system*	4 to 16	2, 3, or 4	1 or 2	1	8	2
Cores/socket	8, 10, 11, or 12	8, 10, 11, or 12	8, 10, 11, or 12	4, 6, or 8	24	24
Max cores/system	192	48	24	8	192	48
Max threads/core	8				2	
Max threads/system	1,536	384	192	64	384	96
Frequencies (Base - Max)	3.55-4.0 GHz	3.15-3.8 GHz	3.4-4.0 GHz	2.3-3.8 GHz	2.9-3.9 GHz	2.4-3.9 GHz
Max L-3 cache/socket	120 MB	120 MB	120 MB	80 MB	36 MB	36 MB
Max L-4 cache/socket	128 MB	128 MB	N/A	N/A	N/A	N/A
Max memory bandwidth/socket	230 GB/s	230 GB/s	170 GB/s	170 GB/s	141 GB/s	141 GB/s
System memory (Min - Max)	256 GB-64 TB	128 GB-16 TB	32 GB-4 TB	32 GB-1 TB	6 TB/server	1.5 TB/server
Max PCIe4 slots	32	10	11	8	N/A	N/A
Capacity on Demand	Supported	Supported	N/A	N/A	Supported	Supported

**"System" refers to either one IBM Power System or one Oracle Exadata Database Server*

SOURCE: Quark + Lepton (August 2021), Intel ARK, Oracle Exadata Database Machine Specifications, IBM Power Systems Facts & Features: Enterprise, Scale-out and Accelerated Servers with POWER9 Processor Technology

global cache. Efficient local buffer cache or global cache will reduce any benefits gained by ESFC. Typical workloads on all but exceptionally large enterprise databases do not realize significant improvement in performance from the use of ESFC.

Oracle's primary performance focus on data warehousing and analytics is also evident in the implementation of HCC. This technology utilizes both row and columnar formats for data storage. Columnar compression techniques can result in higher storage savings and improved tables-scan performance. Using HCC, column values for an appropriate set of rows are compressed and stored in compression units. This compression function synergizes with Smart Scan to improve query performance involving large sequential data sets. HCC provides no benefit for OLTP workloads.

Oracle's Advanced Compression, a feature of Oracle Database, can be leveraged by both Power Systems and Exadata.

POWER9-based Power Systems are the clear performance leaders for transactional workloads. Simultaneous multithreading with eight threads per core (SMP8), high-performance cores, large memory capacity and throughput, and integrated PCIe4 I/O enable extreme performance across a breadth of enterprise workloads, including OLTP. Available industry benchmarks for a range of workloads reinforce this leadership position.

CONSOLIDATION AND VIRTUALIZATION

Virtualization and consolidation techniques allow organizations to increase server efficiency and utilization. Multiple systems can be deployed on a single server system—each with its own operating system, system libraries, and supporting programs that are unique to its requirements. This allows for the consolidation of workloads while maintaining isolation of disparate systems with virtual machines (VM). Both Oracle Exadata and IBM Power Systems are promoted as server consolidation solutions, however the strategies and techniques employed are fundamentally different.

Exadata Systems are designed only for database systems and are often used to consolidate multiple database workloads on a single system. Application workloads are not supported on Exadata and must be run on separate servers.

Oracle's virtualization software for Exadata X8M shifted from Oracle VM and Xen hypervisor to Oracle Linux Virtualization Manager with Oracle Linux KVM hypervisor. This is a significant change for existing Exadata customers and one more example of the inability of X8M to integrate with previous models.

Both bare-metal and virtual deployments are supported, however Oracle does not recommend their virtualization technology for heavyweight applications. Soft virtualization technology results in weak isolation of VMs and consolidation limited to similar workloads.

Power Systems with PowerVM allow for the consolidation of both databases and applications on the same platform, including large applications such as Oracle E-business Suite. POWER9-based systems are, in fact, significantly differentiated from Exadata systems in virtualization and workload management. For example, POWER9-based systems running AIX can support higher workload densities, over a greater

number of partitions, and with greater efficiency, than Exadata. PowerVM architecture also supports various operating systems, such as popular enterprise Linux distributions, providing unmatched flexibility to meet the needs of diverse workloads.

PowerVM represents the industry's most sophisticated server virtualization architecture with its unique firmware-based virtualization—allowing for a much higher level of security, overall enhanced performance, and I/O isolation through Virtual I/O Servers (VIOS), a feature not offered by other virtualization technologies. Power Systems with PowerVM support up to 1,000 VMs per server using logical partitions (LPARs) for dividing system resources. In addition, the Live Partition Mobility feature of PowerVM allows for live migration of LPARs from one system to another for organizations that meet the licensing requirements.

To achieve even better workload balance, IBM's Micro-Partitioning, originally a mainframe innovation, allows for the partitioning of processing resources up to a granularity of 1/100th of a core (with a minimum of 1/20th of a core per partition). Hypervisor can dynamically and efficiently move resources as needed.

Power Systems provide far superior virtualization and workload management than Exadata with mature and industry-leading capabilities.

STORAGE

Power Systems support a wide selection of storage solutions, including disk and solid-state arrays or hybrid storage deployments, from IBM as well as other vendors. IBM's FlashSystem Family of storage systems, based on the IBM Spectrum Virtualize (SAN Volume Controller) virtualization software, can be deployed with IBM Real-Time Compression to gain up to 80 percent in storage capacity savings. Thin provisioning can be used with SVC to optimize storage efficiency by flexibly allocating available storage capacity between users and applications. External storage capacity can be dynamically allocated and reallocated between systems. Several different hardware RAID capabilities providing data redundancy are also supported.

The FlashSystem storage systems are NVMe SSD-based solid-state arrays and utilize the Spectrum Visualize storage operating system. These systems support 32 Gbps Fibre Channel (FC) and NVMe over Fabrics (NVMeOF) and enable customers to utilize IBM's proprietary NVMe-based FlashCore Modules (FCMs) as well as commodity NVMe SSDs in the same system. FCMs are available in 4.8 TB, 9.6 TB, 19.2 TB and 38.4 TB capacities.

The FlashSystem 5200 is a compact storage system targeted for the entry enterprise space. A basic system consists of an NVMe dual controller 1U enclosure that can support from three to twelve FCMs, commodity NVMe SSDs or NVMe storage-class memory in the same system. An all-flash system can provide up to 460 TB of raw storage capacity. The 5200 supports FC, NVMeOF, ISCSI, and RoCE RDMA network protocols and allows 4-way clustering to scale to multi-petabyte range.

FlashSystem 7200 delivers a high-performance, scalable NVMeOF storage solution to meet the most demanding requirements. With support for 16 Gbps or 32 Gbps Fibre Channel with FC-NVMe support,

25Gbps Ethernet with iSCSI, iWARP, and RoCE support, and 10Gbps iSCSI, the FlashSystem 7200 provides broad network flexibility. The FlashSystem 7200 supports four-way clustering enabling a single system with up to 96 NVMe drives and 2944 SAS drives.

The FlashSystem 9200 delivers the highest level of performance, security, and availability for mission-critical enterprise applications. A single 2U system can support up to 24 NVMe FCMs with capacities of 4.8 TB, 9.6 TB, 19.2 TB or 38.4 TB. These FCMs deliver transparent, always-on hardware compression and support FIPS 140-2 Level 1 encryption along with hot-swap capabilities. Storage Class Memory (SCM) NVMe drives can also be utilized to achieve even lower latency for the most demanding workloads. Up to four systems can be clustered and operated as a single system. The FlashSystem 9200 includes dual power supplies and redundant cooling for high availability requirements.

IBM FlashSystem family is a popular flash storage solution for Power Systems that harnesses the high speed and performance of IBM MicroLatency Modules. Using IBM FlashCore, a proprietary, hardware-only controller, the FlashSystem family reduces the overhead generated by software data management produced by other types of flash storage.

For exceptionally I/O-intensive workloads, Power Systems support most new types of all-flash arrays being offered by not only IBM itself, but also a growing number of vendors. Oracle's promotion of its latest X8M Extreme Flash (EF) models highlights the increase in I/O performance as the result of using all flash drives in storage servers. Exadata also offers High Capacity (HC) models featuring hybrid storage. While Exadata flash storage does show an improvement in IOPS rates over HDD devices, it remains comparable to conventional flash storage in access latency.

Although Oracle touts impressive claims of Exadata's I/O throughput and scalability, these only apply to large full-rack systems without mirroring. IOPS rates for smaller Exadata systems, such as a quarter rack, are only a fraction of the rates published for a full rack. Throughput rates published by Oracle also do not consider the effects of mirroring. When using Oracle's recommended ASM high-redundancy mirroring, effective I/O throughput drops to one-third of published rates. Typical workloads will not take advantage of the high throughput Oracle claims Exadata is capable of.

With these performance issues, along with its reliance on mirroring, Exadata requires significantly more storage capacity and storage drives/modules than Power Systems storage. This results in a much higher cost per usable terabyte for Exadata.

SECURITY

Contributing to the resiliency of Power Systems servers are the robust security features enabled throughout the solution stack including processor, operating system, storage, and virtualization.

The AIX operating system can be configured to provide security using an Encrypted File System (EFS) capability, which encrypts and decrypts all files in the file system transparently for non-RAC environments. Once enabled, EFS utilizes user credentials to effect encryption protection for all user-associated files. IBM storage solutions enable data-at-rest encryption through use of embedded data

encryption engines. The drive-level encryption capability protects storage media directly and reduces risk without adding overhead to database servers. Exadata storage offers no similar capability.

PowerVM logical partitions add an additional layer of protection through separation and isolation of system resources in virtualized environments. PowerSC can also be deployed to enable security and compliance automation in virtual environments. Providing centralized security management for virtual environments, PowerSC features real- time file integrity management, trusted logging, compliance automation and reporting, and trusted patch management.

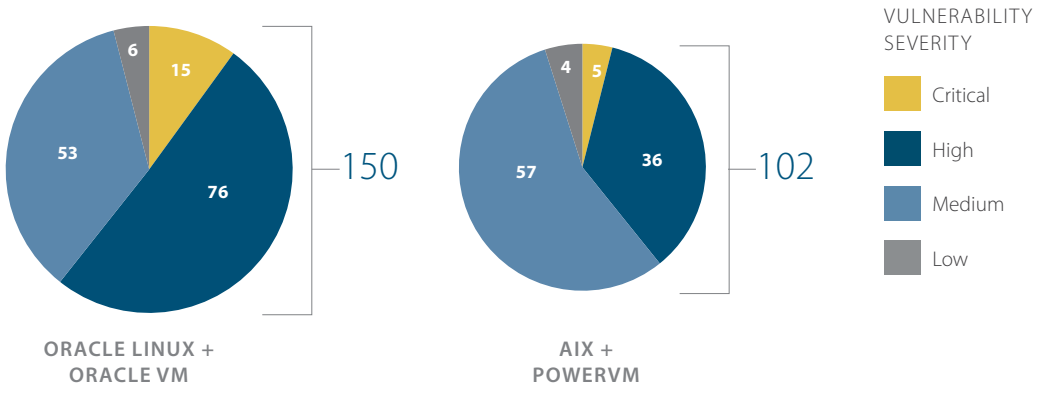
Trusted and Secure Boot is implemented in firmware with POWER9 servers and addressed with AIX 7.2. This allows the verification of firmware and system images via digital signature during boot to ensure authenticity and verify that OS images were not tampered with.

For Exadata, Oracle uses a combination of isolation and access control policies to secure the system. OVM offers software-based virtualization, but software-partitioned environments are less isolated than hardware-based virtualization, offering less protection. Fine grain database isolation can be implemented using Oracle Database Vault, Oracle Virtual Private Database, and Oracle Label Security solutions. Oracle Advanced Security, certified Level 2 through FIPS 140-2, data encryption capabilities and isolation policies can also be deployed. Power Systems implementation is far superior and simpler. Furthermore, Oracle Database Vault and Oracle Label Security may also be implemented in an AIX environment.

From January 2016 through June 2021, IBM AIX received 102 security advisories, as reported by the National Vulnerability Database from the National Institute of Standards and Technology (NIST). PowerVM did not receive a single security advisory during the same period.

In contrast, Oracle Linux and Oracle VM received 124 and 26 vulnerability advisories respectively over the same period. For detailed vulnerability statistics comparing AIX with PowerVM and Oracle Linux with Oracle VM, refer to Figure 3.

FIGURE 3: Comparative Numbers of Vulnerability Notifications: January 2016 – June 2021



SOURCE: Quark + Lepton (August 2021) NIST Computer Security Division, National Vulnerability Database, CVSS Metrics Version 3

RELIABILITY, AVAILABILITY, AND SERVICEABILITY

POWER9-based Power Systems benefit from hardware and microcode-based RAS features derived from earlier models as well as from IBM mainframes. These RAS features include unique capabilities not found in Intel-based designs. Power Systems exploit the greater speed, expanded memory management, and processor allocation capabilities of the POWER9 processor.

The close integration of Power Systems, PowerVM, and AIX provides users with the means to manipulate a wider range of RAS variables, including subsystems, threads, processors, cache, main memory and I/O, multiple types of partitioning, multiple threads, and dedicated and pooled processors, and to do so with higher levels of granularity and flexibility than can be managed with the Exadata platform, which is based on commodity Intel servers.

As discussed earlier, IBM FlashSystem storage servers also exhibit strong RAS characteristics. RAID options provide data redundancy, while hot-swappable drives, dual power supplies, dual controllers, and redundant cooling enhance availability and serviceability. Exadata systems, in contrast, suffer single-points-of failure that must be addressed with RAC software and mirroring.

Power Systems may realize significantly higher levels of capacity utilization over time than less well-optimized platforms. Experiences with ERP systems, for example, have shown that, over periods of months to years, Power Systems may execute workload volumes up to 40 percent larger than indicated by point-in-time performance measurements. Such measurements have been used to compare OLTP capacity across competitive systems, including Exadata.

Exadata relies on Oracle RAC to provide high availability (HA) within a single Exadata frame. Power Systems offer multiple solutions with PowerHA, VM Recovery Manager, and Oracle RAC if needed. HA implementation can extend to dual frames in the same room or across multiple physical sites. Exadata relies on Oracle Data Guard for disaster recovery; while Power Systems offer storage replication, Storage Active-Active Clustering, or Data Guard.

Exadata X8M depends on single networking adapters for both internal network and external interconnect. This single-point-of failure represents a significant liability. In contrast, POWER9-based Power Systems provide the ability to use multiple network and SAN cards to provide the required capacity and redundancy.

According ITIC's April 2020 Global Server Hardware, Server OS Reliability Report, IBM Power Systems had only 1.54 minutes of unplanned downtime per server/per year due to hardware flaws, while Oracle x86 with Linux had 37 minutes over the same period. Both platforms met the 99.999 percent reliability/uptime standard. One percent of Power Systems servers and 15 percent of Oracle x86 servers registered over four hours of unplanned annual downtime due to server flaws.

RAS features included in IBM POWER9-based servers running AIX are shown in Table 4.

TABLE 4: Examples of POWER9 Processor-based Systems RAS Technologies and Techniques

RAS FEATURE	DESCRIPTION
First Failure Data Capture (FFDC)	FFDC technology is used to collect information as it occurs about events & conditions that lead to a failure so re-creation of that problem is not necessary. Error detection & fault isolation capabilities maximize the ability to categorize errors by severity & handle faults with the minimum impact possible. Retry & other mechanisms are designed to avoid outages due to soft errors & to allow for use of self-healing features.
Processor Runtime Diagnostics (PRD)	During normal operations, PRD code itself runs in a special hypervisor-partition under the management of the hypervisor, allowing all POWER9-based systems to continue running the PRD code even if the service processor is faulty. PRD handles recoverable errors including certain self-healing features.
Flexible Service Processor (FSP)	Powered separately from the main instruction processing complex, FSP performs serviceability functions including remote power control options, reset & boot features, environmental monitoring, & PowerVM hypervisor & HMC connection surveillance. The service processor can also post a warning & start an orderly system shutdown in certain circumstances.
Cache Error Handling	L2 & L3 caches & cache directories use an ECC code that allows for single-bit correction. When a persistent correctable error occurs in these caches, the system can purge & delete the data in the cache. L1 cache errors can be corrected using data from elsewhere in the cache hierarchy.
Processor Instruction Retry	When a soft error event occurs within the processor core & is detected before a failing instruction is completed, the processor hardware tries the operation again. If successful, the system can recover before an application or server outage occurs.
Predictive Deallocation	If a persistent recoverable error occurs, PowerVM can start a process for deallocating the failing processor dynamically at run time.
Core Checkstop **System Checkstop**	The core checkstop feature may be used on scale-up systems with many partitions when a fault is not corrected by other error-detecting methods. It may restrict an outage to just the partitions running the threads when the uncorrectable fault occurred. In scale-out systems, hypervisor termination or system checkstops would occur when a fault is not corrected by other error detection methods.
PCIe Hub Behavior & Enhanced Error Handling (EEH)	PCIe hub can freeze operations when certain faults occur, & in certain cases can retry & recover from the fault condition, preventing faulty data from being written out through the I/O hub system. EEH for I/O signals device drivers when various PCIe bus-related faults occur. Device drivers may attempt to restart the adapter after such faults depending on the adapter, device driver, & application.
Infrastructure & Concurrent Maintenance	Power supply & fan redundancy allows the system to continue to operate if one element fails.

SOURCE: Quark + Lepton (August 2021); IBM POWER9 Documentation

Conclusions

The commodity hardware, hybridized software, and customized infrastructure of the Oracle Exadata Database Machine have been shown to negatively impact data center TCO costs, integration capabilities, and performance cost-effectiveness when it is used principally to run OLTP workloads. Three-year costs for deployment of Oracle Database averaged 37 percent less on IBM Power Systems than on Oracle Exadata. Restrictions on licensing also increase the costs associated with Exadata.

Architecture and technology differences in database design, virtualization, and storage contribute to cost and performance variances between the two platforms. Although Exadata touts impressive I/O throughput performance, the values are effectively only a fraction of the advertised rates once system redundancy is considered. Storage server processing, although effective for OLAP workloads, does not benefit OLTP workloads.

The wide support IBM Power Systems offer for various operating systems, middleware, applications, and storage solutions provide organizations with great flexibility, while maximizing the compatibility of existing IT investments with future ones. IBM AIX and Power Systems' long history of robust RAS and security features give users the confidence that their investments will deliver protected value, regardless of possible adjustments that may be necessary in optimizing their operational plans.

Index

LIST OF FIGURES

LIST OF TABLES

LIST OF REFERENCES

Intel Xeon Processor Specifications found at https://ark.intel.com/

National Institute of Standards and Technology found at https://nvd.nist.gov/

ITIC, 2020 Global Server Hardware, Server OS Reliability Report found at https://www.ibm.com/downloads/cas/DV0XZV6R

POWER Processor-Based Systems RAS found at https://www.ibm.com/downloads/cas/2RJYYJML

IBM Power Systems Facts & Features: Enterprise, Scale-out and Accelerated Servers with POWER9 Processor Technology found at https://www.ibm.com/downloads/cas/EPNDE9D0

Quark + Lepton™

CORPORATE OFFICE
Boulder, Colorado USA

www.quarkandlepton.com
info@quarkandlepton.com