HP ProLiant Cluster with Oracle Database 10*g* and Oracle Real Application Clusters

High-end business intelligence platform shows breakthrough 1-TB performance

ORACLE

hp invent

Executive summary

Oracle® and HP recently delivered world-record 1-TB TPC-H benchmark results:

- Query performance of 35,141 QphH at 1,000 GB
- Price to performance ratio of $60/QphH at 1,000 GB

Business intelligence solutions depend on highly scalable data warehouses. This benchmark demonstrates the leadership performance of the HP ProLiant/Linux, Oracle Database 10*g*, and Oracle Real Application Clusters (RAC) architecture as a platform for this workload. The result described in this paper is directly applicable to commercial deployments. Specifically:

- A 12-node cluster of HP ProLiant DL585 Servers, using multiple AMD Opteron x86 processors, delivered performance comparable to large symmetric multiprocessing (SMP) systems.
- The Linux operating system (Red Hat Enterprise Linux AS 3) handled the throughput and processing demands required to achieve the benchmark result.
- The Oracle Database 10*g* with RAC database delivers consistent, high performance query execution in clustered server environments.

This result builds on an earlier 3-TB, eight-node result to demonstrate the commitment of HP and Oracle to this architecture. The benchmark proactively supports current and potential customers that are considering Linux for their data warehouse. HP and Oracle are well-positioned to lead this segment of the data warehouse market.

Business requirements: high-end business intelligence platform

Benchmark results are most usable when they reflect workloads comparable to, or somewhat larger than, the customer application. Thus, this result is best used to assess configurations with raw data requirements similar to, or smaller than, the test system. Benchmark results are only one of several considerations on which customers should base their data warehouse solutions. The suitability for a given customer's data warehouse depends on the anticipated data warehouse workload, current IT environment, and related factors. Customers who are experienced and comfortable with Linux and scale-out clustering should feel confident in this architecture at the 1-TB size range. Customer use cases include midsized retail warehouses and smaller telecom call data record warehouses. The reported configuration costs roughly $2 million in server, storage, and software, with one fact table of six billion rows.

Scalability

Data warehouse solutions consistently grow larger. However, the uncertainty of business environments makes it difficult for managers to predict their future data warehouse requirements. Nonetheless, they want a flexible platform that can scale quickly and easily on demand. The supporting architecture must support large-scale growth without “forklift” upgrades. Business requirements drive growth in two areas:

- Expansion of the core data warehouse as the business grows and new information feeds are added. Successful warehouse deployments consistently expand in terms of raw data, tables, users, and query complexity.
- Deployment and expansion of supporting data marts to support specific businesses. This expansion creates more data, tables, and users.

Performance

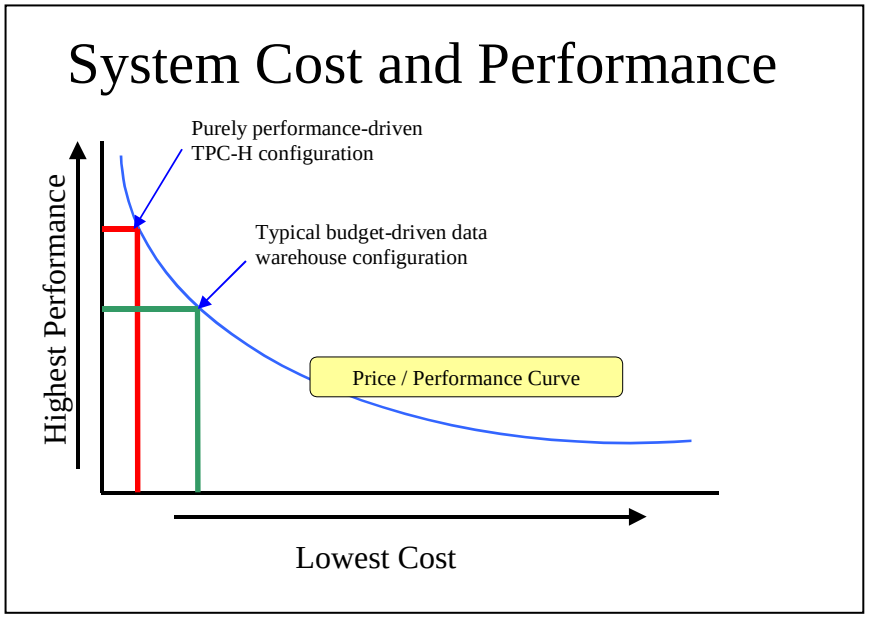
Data warehouse performance requirements can change based on query and load requirements. A large data warehouse typically supports multiple applications, each with its own queries and access patterns. The resulting mixed workload is a demanding one. The data warehouse solution must handle heavy throughput (many concurrent queries) and power (complexity of the queries) requirements. It must also support the seamless addition of processors, memory, I/O bandwidth, and storage capacity to tune for throughput and query performance. The database management system (DBMS) must be easily configurable, and even self-administering, to make best use of the available storage and processing resources

Cost

Customers demand a reasonable return on investment (ROI) from their data warehouses and a low total cost of ownership (TCO). Business managers prefer to add IT computing and storage capacity as needed and to leverage skills and training across multiple applications. The architecture must be flexible enough to allow strong performance from a competitively priced and optimized configuration built from standards-based components.

Customer workloads are less predictable than benchmarks. They often require "CPU headroom" for unanticipated peak loads. Availability and reliability requirements can call for additional, redundant components. These types of real-world considerations add cost. As a result, deployed data warehouses frequently occupy a lower position on the price:performance curve, as shown in the following figure.

Figure 1. Relative price:performance of benchmark and production deployments



Manageability

Data warehouses achieve high performance by distributing processing, I/O, and storage across many components. This large component count places an extra burden on the systems and software management systems. These systems must be able to seamlessly extend and reconfigure all components with minimal disruption to on-going operations. Whenever possible, they should automate propagation of administrative tasks across classes of components, such as disks, arrays, or nodes.

Reliability and availability

Data warehouses at this size range provide business-critical services, and availability requirements approach 24x7. Downtimes are costly to business and should be minimal, including system upgrades to handle business and user growth. Production systems should therefore include version migration, high availability, and failover solutions for all key subsystems. Availability solutions include Redundant Arrays of Inexpensive Disks (RAID) storage, as well as redundant adapters, switches, and arrays in the storage fabric, failover nodes and CPUs in the cluster, and high availability server management software. Disaster recovery options should also be available.

TPC-H benchmark

Description

The TPC-H benchmark is an industry-standard data warehouse benchmark approved by the Transaction Processing Performance Council (TPC) Decision Support subcommittee. The database used in this specific benchmarking effort contains 1TB of usable, or raw, data. It tests the capability of an implementation to:

- Process and analyze very large amounts of data
- Execute analytical queries with a high degree of complexity
- Provide answers to critical business questions

The database allows several types of business analysis to be performed, which reflect typical warehouses and marts:

- Pricing and promotions
- Supply and demand management
- Profit and revenue management
- Customer satisfaction
- Market share study
- Shipping management

Workload characteristics

The 1-TB benchmark is the midrange TPC-H benchmark and, at the same time, is significantly larger than most production warehouses. This characteristic makes it valuable to the many organizations planning to upgrade their business intelligence solutions to a “future-proof” architecture.

The 1-TB TPC-H benchmark tests reflect real-world data warehouse workloads:

- Sequential scans of large amounts of data—Up to 6.0 GB/s for some benchmark queries
- Aggregations of large amounts of data—Up to 1 TB for a single query

- Multi-table joins involving TB-sized tables
- Extensive sorting of large sets of data

The workload requirements put these stresses on the system:

- The benchmark calls for a hardware and storage architecture that supports high I/O bandwidth.
- The data volumes demand a high degree of parallelism in the servers, storage arrays, and database software to generate rapid query results.
- Processing large query data sets means that query working sets must be distributed over many processors.

Applicability to high-end business intelligence solutions

The TPC-H benchmark enables buyers to compare solution options in a predefined and audited environment, specifically:

- Database systems
- Operating systems
- Server architectures
- Storage arrays

This comparison allows analysis and insight into the strengths, weaknesses, and abilities of hardware, storage arrays, operating systems, and databases.

Discussion of results

Design principles

Production data warehouse systems share many traits with the test configurations documented with TPC-H results. All successful benchmark configurations include:

- Many CPUs, nodes, or both so that the workload can be efficiently distributed and rapidly completed
- Rapid sharing of results and load balancing across CPUs, nodes, or both
- A high-speed, high-throughput storage area network (SAN) to support any-to-any data transfer between nodes and disks
- Physical striping of data across many disks and volumes to ensure high aggregate sequential read performance
- Partitioning of large tables to reduce the volume of data required to answer queries

HP and Oracle have published a wide range of leadership results across many configurations. The final choice for any deployment will vary but will always reflect these design principles.

Benchmark components and results

The main audited metrics reported for this leadership result are listed in the following table.

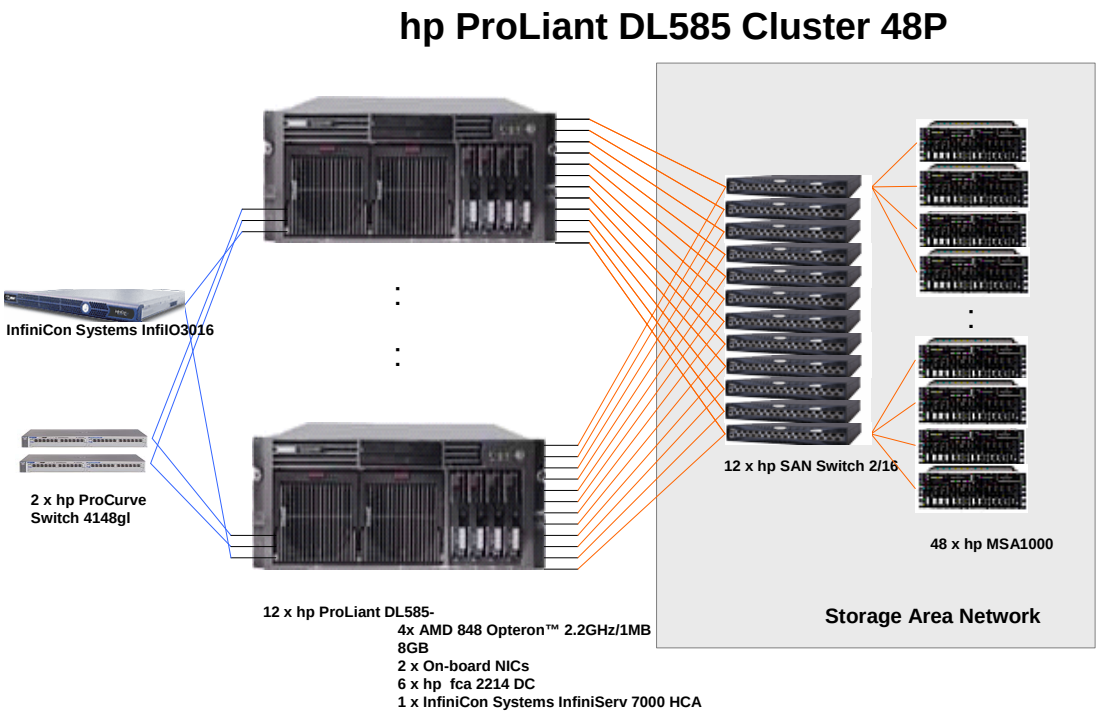
Benchmark metric	Published result	Metric definition
QphH at 1,000GB	35,141	The composite Queries-per-Hour performance metric for the 1 TB size. It reflects multiple aspects of the capability of the system to process queries. These aspects include the selected database size against which the queries are executed, the query processing power when queries are submitted by a single stream, and the query throughput when queries are submitted by multiple concurrent users.
$/QphH at 1,000 GB	$60	The price:performance metric for the 1 TB size
Load time (hours)	3:48[1]	The elapsed time measured during database creation, including: • Table creation • Data loading—1 TB • Index creation • Statistics generation

Configuration: #1 HP/Oracle 1-TB TPC-H benchmark

The extreme performance of the benchmarked hardware system came from a 12-node HP ProLiant Opteron cluster connected to a 14.7-TB HP StorageWorks Modular Smart Array 1000 (MSA1000) storage fabric. This configuration produced aggregate measured throughput of 6 GB/s. Performance was achieved by spreading I/O across 12 SAN switches with 48 arrays and 384 disk drives. Oracle RAC, supported by InfiniBand adapters, provided overall system workload and storage management. The InfiniBand connections provided low-latency, node to node communication so that Oracle RAC could transfer data between nodes without first writing it to disk.

[1] Per the TPC-H specification, database backup time must be included if the database is not RAID protected. The database load time, including backup, was 04:12:23.

Figure 2. Benchmark configuration



The benchmark configuration shown in Figure 2 includes the components listed in the following table.

Component type	Product	Cluster total
Nodes	12 HP ProLiant DL585 Servers	12 nodes
Processors (per node)	Four 2.2-GHz/1-MB AMD Opteron Model 848	48 processors
Memory (per node)	8 GB	96 GB memory
OS disk drives (per node)	Two 36GB 15krpm HDD Ultra 320	24 drives
Network Interface Card (NIC) (per node)	Two on board	24 network connections
Cluster interconnect (per node)	1 Infiniserv 7000	12 high-performance interconnects
Disk controllers (per node)	Six HP StorageWorks FCA 2214DC	72 controllers
SAN	12 HP StorageWorks SAN Switch 2/16 48 MSA1000 408 36-GB, 15,000-rpm HDD Ultra320	12 switches 48 arrays 408 drives (384 for TPC-H data) Total storage: 14,688 GB
DBMS	Oracle Database 10g with RAC, 32-bit version	12 instances managed as a single image by Oracle RAC

Subsystem contributions to the #1 result

Parallel processing within the cluster

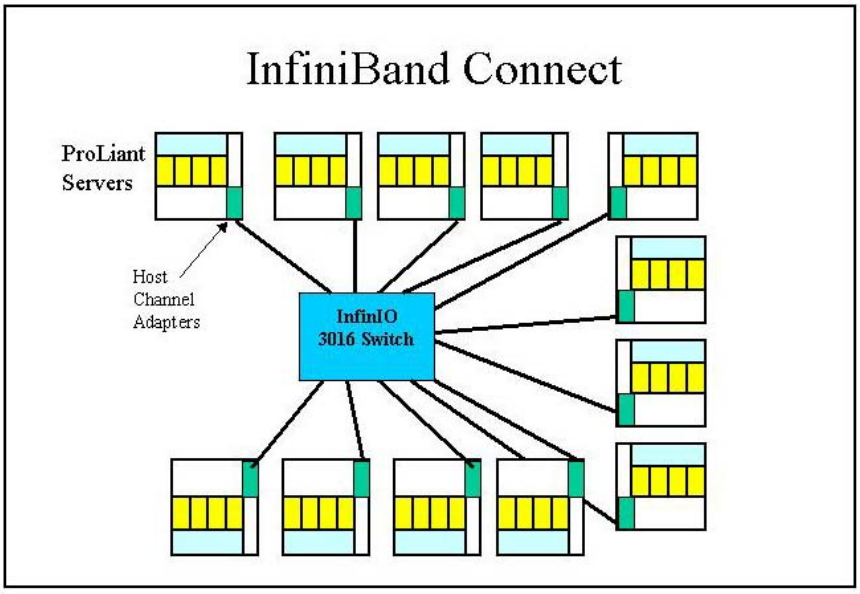
HP ProLiant DL585 Server systems provide excellent scalability when fully configured with four Opteron CPUs. The 12-node configuration was selected for the benchmark because it optimized storage array throughput at the SAN switch. Each 16-port SAN switch was fully utilized, with 12 ports used for node connections and four used for storage array connections. This approach allowed the test system to be CPU-bound, optimizing price:performance with the TPC-H specification.

Each node in the cluster can contain a maximum of 64 GB of PC2100 memory. Because the Oracle Database System Global Area (SGA) does not require large amounts of memory in data warehouse environments such as TPC-H, 8 GB per node was selected. In this case, 8 GB per node (96 GB per cluster) was sufficient to support hash joins and sorts, as well as Parallel Query (PQ) message buffers. Basically, the large size of the data sets makes holding all required data in memory impractical, and minimizing necessary memory reduced the configured price. This level also optimized performance around the query/update mix. A workload with more updates would perform better with more memory.

User Datagram Protocol over Internet Protocol (UDP/IP) over InfiniBand (IB) was used for the high-speed node interconnect.[2] The InfiniBand approach was chosen because it provides higher performance and lower latency than Gigabit Ethernet. It routes messages, data, and other cluster communications traffic, coordinating the access of each node to shared resources. This functionality improves performance within the "shared everything" architecture of Oracle Database 10*g* with RAC. The InfinIO 3016 is one of several roughly equivalent products to implement this approach. Each node is connected through a Host Channel Adapter (HCA) to the InfiniBand switch, as shown in the following figure.

[2] Other approaches, such as User Direct Access Programming Library (uDAPL) uDAP/IP, were not available when this benchmark was in progress.

Figure 3. Node clustering With InfiniBand



Oracle 10*g* with RAC provided node and CPU failover. A production environment would be expected to also employ products such as HP ServiceGuard to manage failover and recovery of the non-Oracle products.

I/O bandwidth

High-end customer data warehouse workloads require large sequential reads. This need stems from the large volumes of data needed to answer typical queries. Large sequential scans and large multi-table joins gain the most from high I/O throughput. For each array, eight disks with throughput of 70 MB/s each were chosen to optimize for the sequential and random access loads of the benchmark.

Node to SAN switch	SAN switch to storage array
Six dual-port adapters per node at 200 MB/s = 2.4 GB/s 12 nodes at 1.2 GB/s = 14.4 GB/s	48 storage arrays at 200 MB/s = 9.6 GB/s

For the MSA1000, its combination of low cost and high throughput was a key factor in the #1 result. The cluster obtained actual performance of 6 GB/s total or about 500 MB/s per node. Benchmark clusters are configured for high throughput to optimize the system for the Power test (single query at a time performance). During the throughput test, the CPUs were mostly saturated and with I/O bandwidth to spare. A typical customer warehouse will have many queries running at the same time, some consuming CPU, others consuming I/O. Typical warehouses often require lower I/O throughput to meet query performance needs.

Data striping is essential to high-performance data warehouse systems. At the same time, the TPC-H specification restricts the use of some indexes and views. To overcome this restriction, data was spread thinly across many small disks to improve read performance. This solution made the “raw data” multiplier 14, much higher than the four to six factor found in many data warehouses. The lower “raw data” factor comes from using indexes and views rather than spindles to increase

performance. Customers often choose a fewer number of larger disks, opting for lower overall cost and higher disk usage over ultimate throughput.

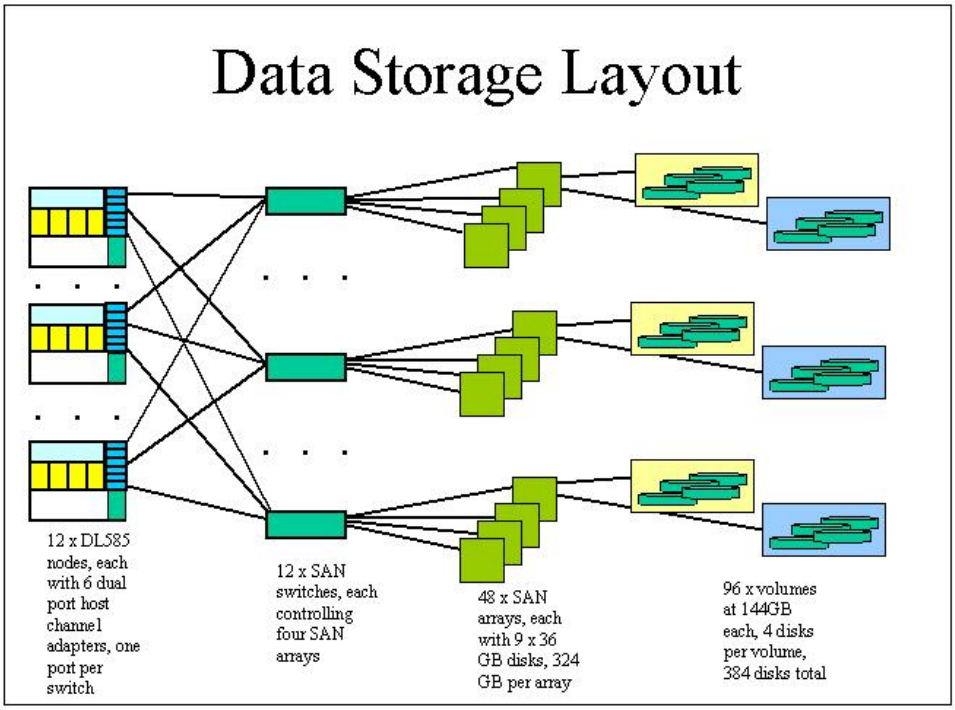
Other options, for instance adding storage bandwidth, such as 4-GB switches or 24- to 32-port SAN switches, enable additional nodes to use the storage fabric. Lowering the number of arrays or disks will reduce aggregate throughput.

Storage array configuration

The HP StorageWorks product set provides strong read performance. It can be configured to support many relatively high-performance yet low-cost Fibre Channel connections between the nodes and the arrays. Storage array performance is 30,000 I/Os per second (IOPS) and 200 MB/s throughput. For data warehouse systems, large sequential reads predominate, so only the throughput number is relevant.

The SAN contains 12 HP SAN Switch 2/16s and 48 MSA1000s. Each HP ProLiant DL585 Server contains six dual-port HP StorageWorks Fibre Channel Adapters (FCA) 2214DC host bus adapters (HBAs). Each port connects to one of 12 HP SAN Switch 2/16s. Each HP SAN Switch 2/16s has four MSA1000s connected to it. Each MSA1000 has two RAID 0 volumes of four 36-GB, 15,000-rpm HDD Ultra320 each for data. Production systems will usually use RAID 1+0 (drive mirroring and striping) or RAID Advanced Data Guarding (ADG) to ensure data availability. RAID 1+0 was not used in the test to reduce costs and improve performance.

Figure 4. Storage optimization for TPC-H performance



Hand-crafted, finely tuned disk and table partitioning are critical to high performance in the benchmark environment. Many optimizations were made, such as partitioning of large tables and distributing tables and temporary files across all 96 volumes. In commercial deployments, a volume manager or Oracle Database 10*g* Automatic Storage Management (ASM) can also give good performance.

The TPC-H specification calls out tables and scaling; table names with row sizes and partitioning data are given in the following table. Total table size for the named tables was 1,000 GB. Table striping with many small drives was done to minimize table read times. The data storage ratio was 14.688, with 1,000 GB used by the DBMS.

Table name	Table size	Row count	Rows per volume	Rows per disk
Region	Less than 1 MB	5	N/A	N/A
Nation	Less than 1 MB	25	N/A	N/A
Supplier	2 GB	10,000,000	104,167	26,042
Customer	26 GB	150,000,000	1,530,612	390,625
Part	30 GB	200,000,000	2,083,333	520,833
Partsupp	110 GB	800,000,000	8,333,333	2,083,333
Orders	150 GB	1,500,000,000	15,625,000	3,906,250
Lineitem	660 GB	5,999,994,267	62,499,940	15,624,985

Consider these areas when developing the physical design for a commercial warehouse:

- Performance was enhanced by using many small disks to improve parallel read performance. Substituting a smaller number of larger disks and reducing the total number of arrays and switches will decrease cost, although it will also reduce peak throughput. A key factor in making these decisions is the maximum throughput required to achieve query performance. Determining this factor will require modeling with real customer data.
- Overall storage requirements could be reduced significantly by using compressed data in the larger tables. This condition will decrease maximum performance, but it can yield significant savings in terms of storage. As mentioned previously, the configuration described maximizes throughput to drive CPU utilization.
- Availability and reliability could be enhanced by adding redundancy in the storage fabric (HBAs, switches, and arrays), as well as storage management software such as HP StorageWorks SecurePath.

Database optimization

Oracle RAC determines at runtime whether to execute a large query by running parallel execution server processes on only one node or on multiple nodes. In general, Oracle Database uses only one node when sufficient resources are available to reduce cross-instance message traffic and synchronization.

Oracle Database 10*g* with Partitioning option powered the record benchmark performance by limiting the amount of data examined and processed and enabling parallel execution. Partitioning is essential to achieve scalability; otherwise, performance tends to top off after four nodes for TPC-H. In particular, the benchmark queries benefited extensively from the Partition Pruning and Partition-wise (PW) joins. For instance, the Order and Lineitem tables were range partitioned on the date column to allow for partition pruning. All tables except Nation and Region had 96 horizontal hash partitions: two partitions per CPU to allow for PW joins. This configuration gives much of the benefit of a shared-

nothing architecture because PW joins happen locally on a node. Oracle Database provides the added benefit that any part of the PW join is independent of specific nodes.

The benchmark system also took advantage of the Automatic Process Global Area (PGA) Memory Management feature of Oracle Database 10*g*. In data warehouse systems, a large percentage of the memory used in operations is dynamic and varies from query to query. While this memory comes from PGA, the size of process memory and the number of processes can vary greatly. Automatic PGA optimizes the process memory size used for each query by evaluating the optimal memory required for the query and the amount of memory being used by other queries. This capability enables database administrators (DBAs) leave configuration of the PGA to Oracle Database 10*g* and cuts time-consuming manual memory parameter optimization.

Oracle Database 10*g* features related to partitioning and cluster management are described in detail in the "Why Oracle Database 10g and Oracle RAC?" section.

Cost breakdown of the #1 result

The total price for the test system is approximately $1.8 million for hardware, storage, and software. Relative expenditures for each major component are shown in the following table.

Test component	Cost
Server hardware	25% (total)
• Processors	• 20% of server total (two extra CPUs per node)
• Memory	• 6% of server total (memory only)
Storage and storage attach	37%
Server software	38% (total)
• Database software	• 95% of software total (database software)
• Operating system	• 5% of software total (operating system)

Every customer deployment differs. HP and Oracle technical staff can provide guidance on optimizing for any specific business environment.

HP/Oracle leadership products for high-end data warehouse environments

The components used for this benchmark create a valid and cost-effective foundation for large data warehouse and business intelligence solutions.

Why HP ProLiant Opteron clusters?

Scalability

Each HP ProLiant DL585 Server node contains four processors, with 8 GB per node. Four-way platforms based on Opteron technology provide high system memory bandwidth at lower latency. Also, Opteron technology allows memory and I/O resources to scale with additional processing power, providing increased levels of system resources as computing needs increase. Combining resource scaling with Opteron Hyper Transport technology allows those system resources to be shared effectively across all processing units. This technology benefits multithreaded applications such as Oracle Database 10*g* and provides for high levels of scalability from single- to four-way processing.

HP ProLiant servers are proven high-performance servers. With the inherent high availability features in HP ProLiant servers and the latest Opteron technology, this combination provides a strong, secure roadmap for high-performance cluster computing.

Heavy query loads

Consistent heavy query loads might require additional CPU performance, which can be provided by higher performance processors or upgrading to HP ProLiant servers with additional CPUs. SAN switches with higher port counts allow larger clusters while maintaining the Oracle Database 10*g* with RAC and InfiniBand architecture.

Availability and reliability

Oracle RAC provides complete failover and workload management for all Oracle Database instances in the cluster. However, many production data warehouses include a variety of additional application packages on other nodes in the extended warehouse solution. HP Serviceguard complements the capabilities of Oracle RAC to enable 24x7 availability for all aspects of mission-critical data warehouses.

Why the HP StorageWorks Modular Smart Array 1000?

The MSA1000 provides best-in-class, industry-leading performance for data warehouse workloads, which are characterized by large sequential scans (large I/Os and read intensive). This type of workload differs greatly from traditional OLTP workloads (small I/Os and write intensive). The MSA array family is a good choice for data warehouse projects because:

- Each array can carry fourteen146-GB drives or 2 TB[3] of storage. With 48 arrays, this configuration yields 96 TB of storage, roughly six times that used in the test.
- The arrays can be equipped with backup controllers and switches.
- Disk drives and other key components are hot-swappable to ensure availability. Hot swapping occurs transparently under the control of Oracle RAC.
- HP can enhance overall capability with the parallel database clusters for Oracle.
- The lower cost of the MSA1000, combined with high performance and scalability, offer an exceptional solution.

As an entry-level product line, the MSA1000 provides good sequential read performance when data is striped over many disks and arrays. For additional array throughput and management capability, HP also offers the HP StorageWorks Enterprise Virtual Array (EVA) and HP StorageWorks XP128, XP1024, and XP12000 Disk Array series storage solutions.

Why Oracle Database 10*g* and Oracle RAC?

Oracle Database 10*g* with Oracle RAC delivers proven scalability, performance, availability, and price:performance for high-end business intelligence applications. Oracle Database today supports multiterabyte data warehouses with large user populations, heavy query processing workloads, and complex, dynamic data loading. These features of Oracle Database 10*g* specifically enable scalability, performance, and availability in clustered data warehouse configurations:

- Shared-everything architecture
- Oracle RAC
- Partitioning

Shared-everything architecture

Oracle Database 10*g* with RAC delivers great flexibility for growth because of its shared-everything architecture. In this model, each node has access to all the storage. Customers can add servers, storage, or both separately, based on demand. In the shared-nothing architecture, each node has its

[3] As of benchmark result publication date.

own set of disks. With a shared-nothing cluster, expansion requires server and storage be purchased concurrently and that data be redistributed with each change to servers or attached storage.

Oracle Database 10*g* supports the auto-provisioning capability of grid technology such as the HP Adaptive Enterprise. Customers can add or remove CPUs as needed, even when the application is running. Oracle Database 10*g* recognizes new CPUs instantly and allocates work to them. For example, an enterprise grid could automatically and dynamically add CPUs to a data warehouse with a heavy overnight load process or during periods of high query activity. The grid would then remove them when no longer needed. These examples show how the shared-everything architecture of Oracle Database 10*g* with RAC matches the requirements of 24x7 data warehouse systems.

Oracle RAC for scalability, availability, and manageability

RAC technology of Oracle Database 10*g* enables customers to scale up incrementally to reduce capital expenses by seamlessly adding nodes as the demand increases. Oracle RAC automatically harnesses the processing power of additional nodes as they are brought into the cluster. Oracle RAC eliminates the need for forklift upgrades and makes the capacity upgrade process, easier, faster, and transparent to users.

Oracle Database 10*g* increases application performance and scalability through new features supporting InfiniBand, the next generation high-bandwidth, low-latency switch fabric architecture for cluster interconnects. Cost advantage and manageability are other benefits: an InfiniBand cluster requires fewer server adapters and switch ports, and the environment is easier to manage. The cost advantage improves as cluster and storage requirements increase. Oracle Database 10*g* enables customers to use InfiniBand for cluster interconnects without requiring it for storage and networking infrastructure.

Oracle RAC handles unscheduled outages (for example, instance or node failures) effectively by performing automatic recovery for the failed instance and continuing to provide database service using surviving instances. User data is always accessible if there is at least one available instance running in the cluster.

The linear scalability model of Oracle Database 10*g* with RAC eliminates guesswork regarding the processing ability of the database. If a single node fails in a 12-node cluster, the data warehouse performs at 11/12th of full performance. It also gives the administrator the ability to perform scheduled maintenance on a subset of nodes or components, while continuing to provide service to users. Oracle RAC also can automatically balance new database connection requests among the available instances. It makes decisions based on lowest processing load and fewest connections. Each instance can provide load data to "listeners" at each node and can cross-register with remote listeners, which means each listener is aware of all services, instances, dispatchers, and their current loads, regardless of their location. Each listener can therefore send an incoming client request for a specific service to the least-loaded node, instance, or dispatcher.

One of the major value propositions of Oracle Database 10*g* with RAC is the significant reduction in the management cost of deploying and maintaining an Oracle-based data warehouse. An Oracle RAC environment is a single database accessed by multiple instances. This single system image is preserved across the cluster for all database operations to simplify manageability. Enterprise Manager Grid Control manages grid-wide operations, including management of the entire stack of software, provisioning users, cloning databases, and patch management. DBAs perform configuration, high availability operations, recovery, and monitoring functions just once. Oracle Database then automatically distributes the management functions to the appropriate nodes.

With Oracle RAC, a typical large-scale 24x7 data warehouse needs only two full time DBAs. Oracle Database 10*g* is becoming a self-managing database, with the introduction of automatic performance diagnosis and tuning recommendations. The built-in tools in Oracle Database 10*g*, such as Automatic Workload Repository (AWR), Automatic Database Diagnostics Monitor (ADDM), and SQL Tuning

Advisor, enhance DBA productivity, improve performance of the database, and reduce administrative overhead to keep costs low.

Partitioning for data management and performance

The Oracle Database with Partitioning option allows tables and indexes to be partitioned into smaller, more manageable units, which enables DBAs to pursue a "divide and conquer" approach to data management, either for query optimization or update management. With partitioning, maintenance operations can be focused on particular portions of tables. It allows backing up a single partition of a table, rather than backing up the entire table. A typical use is to support a rolling window load process in a data warehouse, such as loading new data into a table on weekly basis. That table can be range-partitioned so that each partition contains one week of data. The load process is then just adding a new partition. Purging data from a partitioned table means dropping a partition, a cheap and quick data dictionary operation.

By limiting the amount of data to be examined and processed and enabling parallel execution, partitioning improves the data warehouse performance. The performance advantage of partitioning is achieved through:

- Partition pruning—Oracle Database 10*g* optimizes SQL statements to mark the partitions or subpartitions that must be accessed. It eliminates (prunes) unnecessary partitions or subpartitions from access by those SQL statements. In other words, partition pruning is the omitting of unnecessary index and data partitions or subpartitions in a query. For example, if a query only involves March sales data, there is no need to retrieve data for the remaining 11 months. Pruning can dramatically reduce the data volume to improve query performance. Partition pruning works with all of the other Oracle Database performance features. Oracle Database will utilize partition pruning in conjunction with any indexing technique, join technique, or parallel access method.
- PW joins—This feature joins two tables that are partitioned along the join columns. With PW joins, the join operation is broken into smaller joins that are performed sequentially or in parallel, completing the overall join in less time. By taking into account data distribution, PW joins minimize data exchange among parallel slaves during parallel joins.

Summary: findings from the #1 result

The configuration for this record-setting result demonstrates that an HP ProLiant/Linux/Oracle Database 10*g* with RAC solution using MSA1000 storage can be put in production for large data warehouse projects. The solution surpasses the key criteria outlined in this paper.

- **Scalability.** The InfiniBand-enabled cluster demonstrates scaling to 12 nodes. The architecture supports scale-out to many more nodes, including failover nodes. The MSA1000 storage fabric and HP ProLiant cluster size can be extended to support the higher throughput requirements of larger scale-out configurations by using SAN switches with higher port counts. Much larger drives can be added to the arrays to increase total storage by over a factor of 10 from the configured system. In real-world terms, this means that customers can extend the warehouse and add data marts within the same, centrally managed storage environment.
- **Performance.** The cluster contained 48 Opteron CPUs and provided exceptional price:performance. The InfiniBand technology and ability to add I/O throughput means that the overall configuration can easily reach higher levels of performance. Oracle Database 10*g* with RAC and ASM combine to enable expansion of processing and storage components. The MSA1000-based storage fabric enabled throughput of 6 GB/s, with additional throughput available as nodes and faster switches are added to the cluster. Oracle RAC optimizes performance automatically by updating partitioning as new storage is added.

- **Cost.** The standards-based HP ProLiant Opteron servers help to reduce overall solution cost. Nodes can be added inexpensively to improve performance or provide failover redundancy. Oracle RAC can then seamlessly integrate the extra nodes into the cluster. The MSA1000 delivers a cost-effective storage subsystem with a high degree of parallelism, redundancy, and performance.
- **Manageability.** Oracle Database 10*g* with RAC and ASM simplify administrative tasks such as partitioning. Oracle RAC enables all database instances to be managed in parallel, which reduces costs and ensures consistent management across the cluster.
- **Reliability and availability.** HP ProLiant servers and MSA storage offer strong availability and reliability capabilities. These components extend the benchmark configuration to better support mission-critical business intelligence applications.

There are several key design considerations for data warehouses based on this architecture:

- **Servers and memory.** HP ProLiant servers provide industry-leading scalability when fully loaded with four CPUs. Memory requirements are modest unless the workload includes a high percentage of updates. If the overall application requirements include transaction workloads, consider hosting that workload on separate servers configured for that workload.
- **Cluster sizing.** Oracle RAC efficiently allocates query workloads across the cluster using the InfiniBand switch. Adding many nodes without including adequate I/O bandwidth results in low CPU utilization and poor scaling.
- **I/O bandwidth.** In decision support systems, I/O bandwidth is the most common limiter of query processing. Ensure that the storage fabric (switches, arrays, volumes, and disks) is sufficiently distributed to provide adequate read bandwidth.
- **Database optimization.** Oracle RAC manages query processing across nodes and partitions to minimize the data set passed off to each node for processing. Intelligent partitioning of large tables, complemented by PW joins, enables Oracle RAC to make optimal runtime decisions.
- **Reliability and availability.** Modest investments in redundant nodes and storage fabric components can greatly increase the availability and reliability of a large data warehouse. These components can be automatically invoked by Oracle RAC to ensure maximum performance of the DBMS. HP Serviceguard complements Oracle RAC by enabling fine-grained control of the additional hardware, software, and storage components common to large data warehouse environments.

For more information

- Full disclosure report for 1-TB, 12-node, four-way HP ProLiant/Oracle RAC cluster
  http://www.tpc.org/tpch/results/tpch_result_detail.asp?id=104102501
- Full disclosure report for 3-TB, eight-node, four-way HP ProLiant/Oracle RAC cluster
  http://www.tpc.org/tpch/results/tpch_result_detail.asp?id=104030202

5983-1177EN, 02/2005

ORACLE

hp invent