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

High-end business intelligence platform shows breakthrough 10-TB TPC-H performance

Executive summary 2

Business requirements: High-end business intelligence platform 2
- Scalability 2
- Performance 3
- Cost 3
- Manageability 4
- Reliability and availability 4

TPC-H benchmark 4
- Description 4
- Workload characteristics 5
- Applicability of the benchmark to high-end business intelligence solutions 5

Discussion of the HP/Oracle #1 result 6
- Design principles 6
- Benchmark components and results 6
- Hardware and software configuration: #1 HP/Oracle 10-TB TPC-H benchmark result 6
- Subsystem contributions to the result 8
- Cost breakdown of the #1 result 12

HP/Oracle leadership products for high-end data warehouse environments 12
- Why Integrity Superdome Cluster? 12
- Why HP StorageWorks XP Disk Array storage? 14
- Why Oracle Database 10*g* and Oracle Real Application Clusters? 14

Summary: Findings from the #1 result 16

For more information 18

ORACLE

hp invent

Executive summary

Oracle® and HP recently delivered world-record 10-TB TPC-H benchmark results:

- Query performance of 86,282 QphH at 10,000 GB
- Price to performance ratio of $161 QphH at 10,000 GB

Oracle Database 10*g* and Oracle Real Application Clusters (RAC) ran on a two-node, 64-way, HP Integrity Superdome cluster. This benchmark demonstrates the fitness of the Integrity Superdome/HP-UX and Oracle Database cluster architecture for the largest commercial data warehouses. Specifically:

- Integrity Superdome server clusters scale well in comparison to single-node solutions.
- Integrity Superdome solutions combining scale-up capability with scale-out clustering deliver high-end performance.
- HP-UX 11i v2 handled the rigorous tuning and optimization requirements required to achieve the result.
- HP XP series storage arrays achieved consistent throughput of 1.2 TB/minute.
- Oracle Database 10*g* with RAC database delivers consistent, high-performance query execution in clustered, symmetric multiprocessing (SMP) server environments.
- Customers building extremely large enterprise data warehouses can now consider standards-based clusters.

The reported configuration costs roughly $13 million in server, storage, and software, with one table of 60 billion rows. This result builds on the success of an earlier 10-TB, single-node result. More importantly, this newly announced cluster result demonstrates a superior alternative to proprietary, shared nothing platform architectures. The highly parallel SMP architecture and 64-bit Intel® Itanium® CPUs of the Integrity Superdome, coupled with Oracle RAC clustering, provides an unbeatable high-end solution. The sustained I/O throughput of 1.2 TB/minute proves that the vast majority of businesses and current implementations can scale from a single Superdome to a Superdome cluster. This report guides system designers in evaluating this general architecture for similarly sized production data warehouses.

Business requirements: High-end business intelligence platform

This 10-TB result can be adapted to design large (3 to 50TB raw) data warehouses. The suitability for a given customer's data warehouse depends on several factors, such as anticipated workload, current IT environment, staff capability, and related factors. Customer use cases include large retail warehouses and telecom call-data record warehouses. Commercial data warehouses in this range must address the stringent business requirements detailed in the following sections.

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. Data warehouses of this size often add over 100 GB of new data every day.
- Deployment and expansion of new dependent data marts to support specific businesses.

As a data warehouse demonstrates value, it will include more data, tables, and users. This growth in usage results in higher workloads as the complexity of the queries and data analysis increases.

Performance

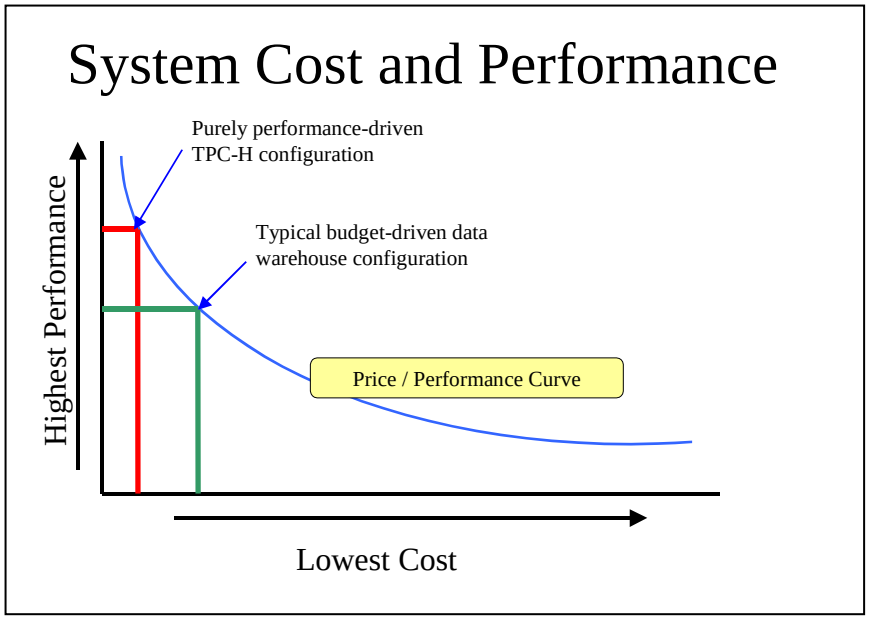
Data warehouse performance requirements can change based on query and load requirements. A large data warehouse typically supports multiple applications, each with its own queries and access patterns. The resulting mixed workload is a demanding one. The data warehouse solution must handle heavy throughput (many concurrent queries) and power (complexity of the queries) requirements. The hardware complex must support the seamless addition of nodes, 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 self-administering to make best use of the available storage and processing resources.

Cost

Customers demand a reasonable return on investment from their data warehouses and a low total cost of ownership. Business managers prefer to add IT capacity as needed and 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 off-the-shelf components that can be repurposed as business needs change.

Customer workloads are less predictable than benchmarks. They often require “CPU headroom” for month-end or unanticipated peak loads. Availability and reliability requirements can call for additional, redundant components. These types of considerations add cost. As a result, deployed data warehouses frequently occupy a more conservative 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 must approach 24x7 availability. Downtimes are costly to business and must be minimized, even for system upgrades to handle business and user growth. Production systems should therefore include version migration, high availability, and failover solutions for all key subsystems. Typical features include RAID storage, as well as redundancy and failover in the storage fabric, failover nodes 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 designed under the auspices of the Transaction Processing Performance Council (TPC) Decision Support subcommittee. The database used in this result contains 10 TB of usable, or raw, data. It tests the capability of an implementation to:

- Process and analyze extremely large amounts of data (10 TB in this case)
- Execute analytical queries with a high degree of complexity
- Provide answers to critical business questions

The database allows several types of business analysis which are typical of data warehouses and data marts:

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

Workload characteristics

The 10-TB TPC-H is the largest of the TPC-H family, and requirements are particularly intense.

- The data volumes that are analyzed demand large orders of parallelism in the servers, storage arrays, and software to achieve excellent performance. These components must all be designed with this high level of parallelism in mind.
- Extremely high levels of sustained I/O throughput are required to support the server CPU processing capacity.
- Rapid processing of huge datasets requires that query working sets be efficiently distributed to many high-power processors.
- Results must be shared rapidly across nodes and processors.

The workload mimics real-world deployments by requiring:

- Very large sequential scans of data (multiple terabytes at a time per query)
- Very large aggregations of data (multiple terabytes per query)
- Multitable joins involving TB-sized tables
- Extensive sorting of huge amounts of data

With system pricing of roughly $13 million, the system configuration described in the following sections is larger than almost all data warehouses currently in production, which makes it valuable to organizations planning extremely large-scale or "future-proof" deployments.

Applicability of the benchmark to high-end business intelligence solutions

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

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

The audited results allow analysis and insight into the strengths, weaknesses, and abilities of a specific combination of components. Each result reflects careful tuning and optimizing within the test requirements. An HP or Oracle technical representative can assist with assessing the fit of a specific result to a specific business situation.

Discussion of the HP/Oracle #1 result

Design principles

Production data warehouses 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 fabric to support any-to-any data transfer between nodes and arrays
- 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 many leadership results and will continue to do so. The final choice of configuration for any deployment will vary but will always reflect these design principles.

Benchmark components and results

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

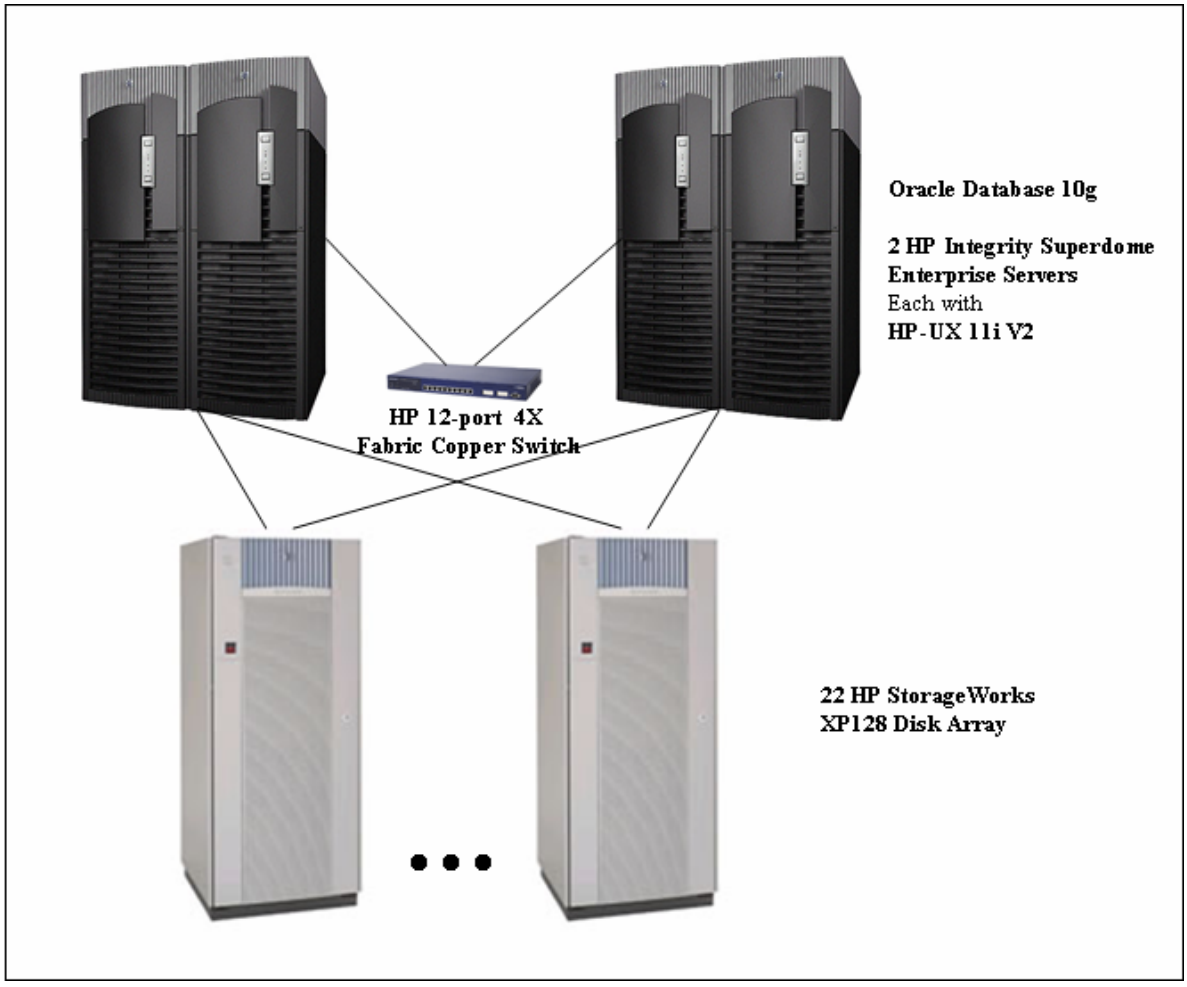
Benchmark metric	Published result	Metric definition
QphH at 10 TB	86,262	The composite queries per hour performance metric for the 10 TB size. It reflects multiple aspects of the capability of the system to process queries. These aspects include the selected database size against which the queries are 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 10 TB	$161	The price:performance metric for the 10 TB size.
Load time (hours)	7:19	The elapsed time measured during database creation, including: • Table creation • Data loading—10 TB • Index creation • Statistics generation

Hardware and software configuration: #1 HP/Oracle 10-TB TPC-H benchmark result

The benchmarked hardware system was a two-node Integrity Superdome cluster connected to a 105-TB HP StorageWorks XP128 Disk Array storage fabric. This configuration produced aggregate measured throughput of 20 GB/s. Performance was achieved by spreading I/O across 22 XP128

Disk Arrays with 1,408 disk drives. Oracle RAC, supported by InfiniBand adapters, provided overall system workload and storage management.

Figure 2. Benchmark configuration



The benchmark configuration shown in the preceding figure is detailed in the following table.

Component type	Product detail (per node or array)	Cluster total
Server components		
Node type	Integrity Superdome	Two nodes
Processors	64 1.5-GHz Itanium 2	128 processors
Memory	256 GB	512 GB memory
Cluster interconnect		One InfiniBand protocol switch
Cluster interconnect (per node)	Two HP PCI-X two-port 4X Fabric Host Channel Adapters (HCA)	Four high performance interconnects
Storage fabric components		
Host bus adapters	44 2-GB dual port PCI	176 adapters

Component type	Product detail (per node or array)	Cluster total
Storage area network	22 HP StorageWorks XP128 Disk Arrays 16 array groups per XP128 Disk Array 64 disks per array 73-GB, 10,000-rpm HDD Ultra320 disks	22 storage arrays 352 disk groups 1,408 drives (1,452 total, including 44 spares) Total Storage: 105 TB
Software		
DBMS	Oracle Database10*g* Oracle Real Application Clusters (RAC)	2 Oracle Database servers managed as a single instance by Oracle RAC

Subsystem contributions to the result

Parallel processing within the cluster

The two-node Integrity Superdome cluster was chosen because it provided a good balance of throughput and processing power to match the 10-TB TPC-H specification. The ability to deliver high levels of SMP is a cornerstone of the Superdome architecture. The combined 128 processors efficiently distributed this high-end workload. In combination with the power of the Itanium 2 processor, they provide rapid processing of large workloads.

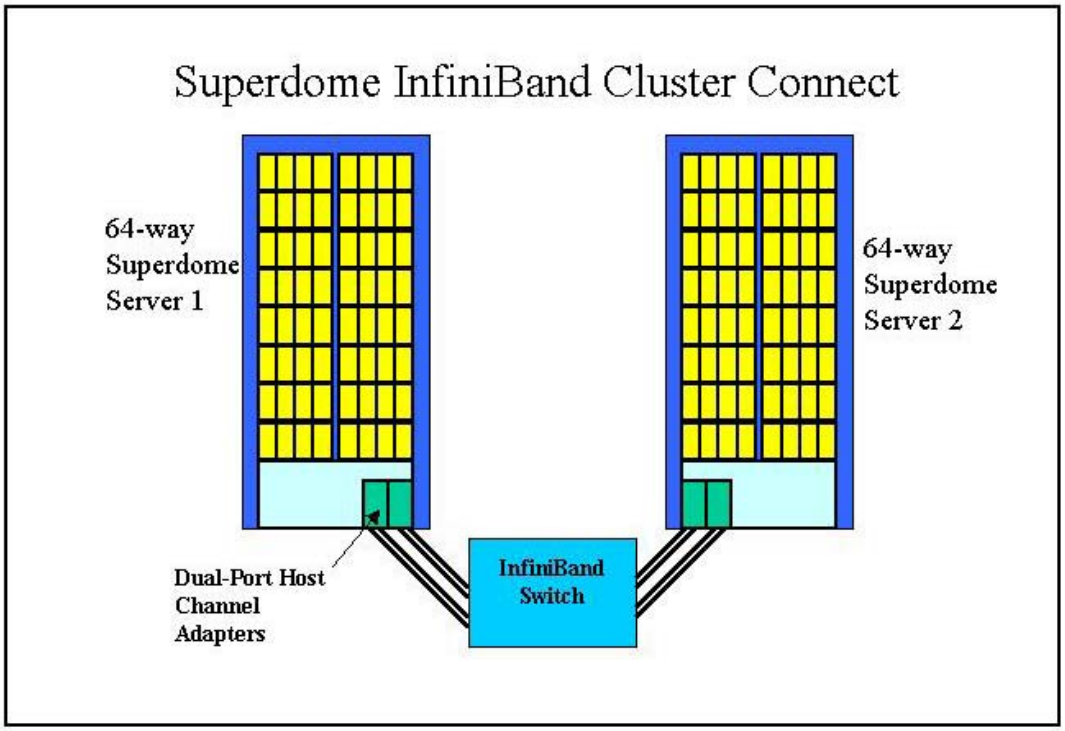
By fully loading each Superdome with 64 CPUs, the configuration maximized the performance of each node. This showcased the price:performance advantages provided by the linear scaling of the Superdome architecture. Partitioning was not used to group CPUs within each Superdome. Besides SMP scaling, the two-node Superdome cluster posted a impressive 91% performance increase over the previously released single-node Superdome result.

The load rate was also a leadership performance result in this 10-TB benchmark. Again, compared to the single Superdome 10-TB result, it proved the increased parallel processing derived from adding the second Superdome to the hardware complex. It also shows that the tight integration of HP platforms with Oracle RAC delivers broad improvement in processing performance.

Because the Oracle System Global Area does not require large amounts of memory in data warehouse environments such as TPC-H, 4 GB per CPU was chosen. In this case, 4 GB per CPU (256 GB per node and 512 GB per cluster) was sufficient to support hash joins and sorts, as well as Parallel Query (PQ) message buffers. This level also optimized performance around the query and update mix. A workload with more updates would perform better with more memory.

An InfiniBand switch provided the cluster interconnect, with four Host Channel Adapters (HCAs) per Superdome. InfiniBand provides the communications path for Oracle RAC workload management. InfiniBand-based interconnects are particularly useful for Oracle RAC scalability because they provide a higher bandwidth, lower latency, and better CPU utilization than Gigabit Ethernet. The execution of this high-end benchmark has enabled HP and Oracle to extend and improve the Oracle RAC algorithms to improve performance.

Figure 3. InfiniBand node clustering



HP 4x Fabric cluster interconnect solutions are based on InfiniBand technology and offer 4.6 μs one-way latency and a single stream bandwidth of 760 MB/s and 924 MB/s with aggregated streams, all using industry-standard technologies and off-the-shelf components. While each Superdome could support eight HCAs, four were sufficient to support communications for both the Power and Throughput tests.

Node and CPU failover solutions are provided by a combination of Oracle RAC and HP Instant Capacity.

I/O bandwidth

High-end customer data warehouse workloads are I/O intensive because of the large volumes of data needed to answer typical queries. Large sequential scans and large multitable joins gain the most from large I/O bandwidth. The sustained I/O throughput achieved in this benchmark was approximately 20 GB/s. This metric was calculated from the amount of data that was analyzed by queries in a given result time, which translates into 1.2 TB of data analyzed per minute. The measured 20 GB/s demonstrates the well-balanced performance of all the components (server and software) supporting system I/O.

A significant contributor to the aggregate throughput of the system is the I/O bandwidth of the two Superdomes. Each of the 22 XP128 Disk Arrays had four Fiber Channel connections to each Superdome, for a total of 88 Fibre Channel connects. Each connection could deliver approximately 900 MB/s throughput. The cluster total of 176 connections prevented I/O bottlenecks during the intensive query runs. The Superdome cluster has a combined total of 384 I/O slots, so Fibre Channel connects could be doubled if necessary.

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 are mostly saturated, and there is

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 limitation, data is spread thinly across many small, lightly loaded disks to improve read performance. RAID 1+0 provided redundancy and contributed to the "raw data" multiplier of 10, which was higher than the four to six factor found in many data warehouses. The lower "raw data" factor in commercial deployments also 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.

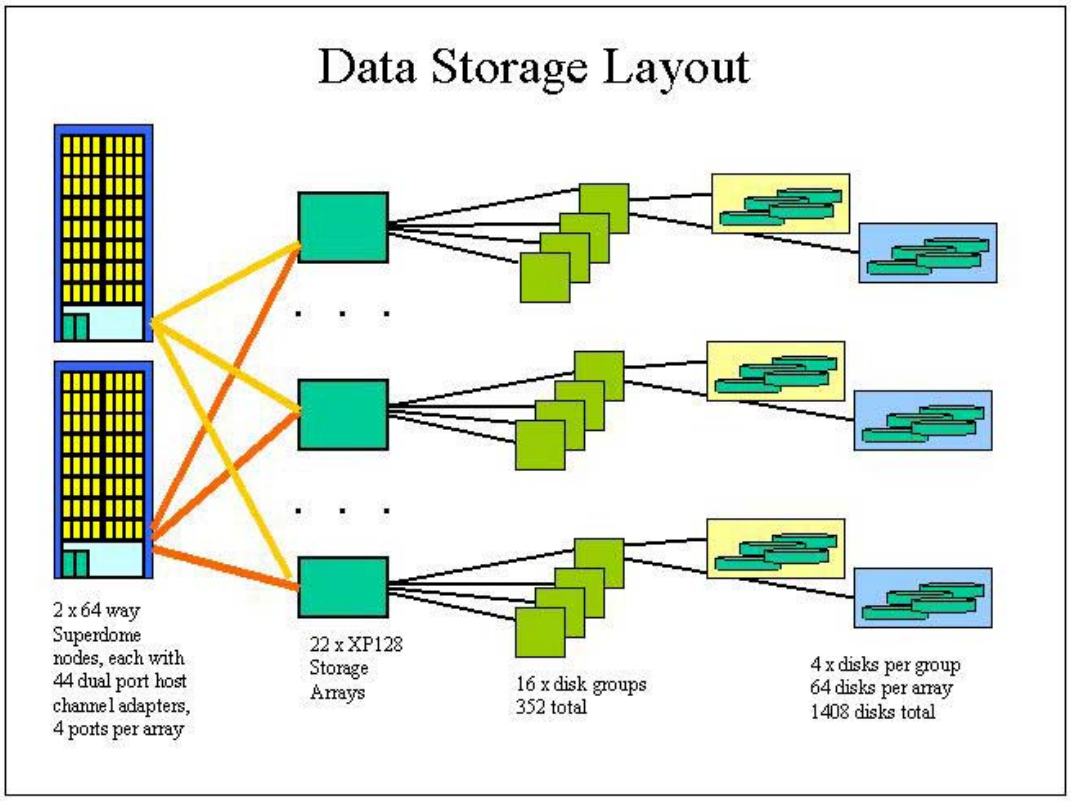
This sustained I/O throughput number is also significant because it indicates that the vast majority of businesses and current implementations can scale comfortably up through a single Superdome and into a cluster of Superdomes without ever having to swap hardware platforms.

Storage array configuration

The XP128 Disk Arrays provide best-in-class, industry-leading performance for high-end business intelligence (BI) workloads. BI workloads are characterized by large sequential scans (large I/Os and read intensive). Not all storage arrays perform well for both BI and traditional Online Transaction Processing (OLTP) workloads (small I/Os and write intensive). The XP array family is particularly well suited to both types of workloads, given their high sustainable I/O throughput, large number of host ports, large number of controllers, high bandwidth internal crossbar switch, and high scalability.

Optimizing throughput for the benchmark required distributing data across 1,408 disks and enabling 176 channel connects. Relative to the capacity of the storage arrays, the attached storage for each array is small (4.67 TB, 16 four-disk array groups). For this benchmark, the XP128 Disk Array provided the most throughput for the least cost, based on this amount of attached storage. A configuration based on XP1024 Disk Arrays, with more storage per array, can be more cost-effective in commercial deployments. In such cases, higher storage per array is more important than ultimate throughput.

Figure 4.



All data storage disks are configured at RAID 1+0 to ensure adequate data redundancy. Data is distributed in seven logical units (LUNs). Table distribution across the arrays and disks was accomplished through the LUNs. Every table and index had a separate sizing and striping configuration. Describing the layout in detail is beyond the scope of this paper.

Note
Large warehouses and table structures require extensive planning and testing to create high-performance layout and striping configurations. Consult your Oracle or HP consultant for more information.

Database optimization

At runtime, Oracle RAC determines whether to execute a large query by running parallel execution server processes on only one node or on multiple nodes. In general, Oracle tries to use only one node when sufficient resources are available. This practice reduces cross-instance message traffic and synchronization, decreasing throughput loads on the InfiniBand switch.

Oracle Database partitioning powered the record benchmark performance by limiting the amount of data examined and processed and enabling parallel execution. In particular, the benchmark queries benefited extensively from the Partition pruning (PP) and Partition-wise (PW) joins. For instance, the Order and Lineitem tables were range partitioned on the date column to allow for Partition pruning, which gives much of the benefit of a shared nothing architecture because PW joins happen locally on

a node. Oracle 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 query processing, 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. It eliminates static memory allocation parameters. This functionality enables database administrators (DBAs) to leave configuration of the PGA to Oracle Database 10*g* to eliminate time-consuming manual optimization.

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

Cost breakdown of the #1 result

The total cost for the test system is approximately $13 million for hardware, storage, and software. Relative expenditures for each major component are shown in the following table.

Test component	Cost
Server hardware	26% (of system total)
• Processors	• 35% (of server total)
• Memory	• 26% (of server total)
Storage and storage attach	58% (of system total)
Server software	14% (of system total)
• Database software	• 92% (of software total)
• Operating system	• 8% (of software total)

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

Why Integrity Superdome Cluster?

Integrity Superdome Cluster

It has been noted during the analysis of numerous high-end customer data warehouse workloads that the workload can be easily decomposed into subtasks that execute quickly in highly parallel environments. The more processors that are available, the more subtasks can be run in parallel. The Superdome/Oracle Database solution combines the large CPU count of the Superdome cluster with the workload distribution of Oracle RAC to achieve a highly distributed processing cluster.

Very large data warehouses can be expected to grow two or three times in processing, I/O, and storage capacity. The original physical design must transparently accommodate this growth. For instance, current Superdome systems can each handle up to 128 processors. If long-term capacity planning shows requirements exceeding this capacity, multiple Superdome frames can be included in the initial configuration, which makes it easy to add processors and I/O capacity through Oracle RAC as demands increase. The frames can be preloaded with HP Instant Capacity CPUs, which can be called on by Oracle RAC for peak demand or as additional permanent resources.

InfiniBand throughput requirements are roughly proportional to CPU count. The test system required four channels per 64-way Superdome. A pair of 32-way Superdomes would require two channels each. InfiniBand switches have sufficient ports to accommodate multiple fully loaded Superdomes. The ultimate capacity has not been evaluated.

As discussed earlier, the Superdome capacity for storage channel connections (176 per Superdome) is double that required for the current result, which means I/O throughput capacity has tremendous headroom. System memory requirements are also small relative to capacity, so neither of these is a limiting factor for scaling this configuration far past the published result.

HP-UX 11i v2

The TPC-H benchmark exercises an extremely intense workload. Few commercial applications make use of a system this way 24x7. More commonly, warehouse loads are moderate until the last few days of the month, with dramatically increased reporting loads at the closing of the period. Customers choosing systems from any other vendor, with the possible exception of high-end mainframe solutions, face one of two unsatisfactory options:

- Provision for the lower system utilization experienced most of the month but fail to meet user needs during month-end periods
- Over-provision systems to meet the end of period requirements and pay for unused capacity most of the month

HP-UX 11i, uniquely among UNIX® operating systems offers a third alternative. Using HP-UX 11i virtualization capabilities such as HP Process Resource Manager (PRM), Workload Manager (WLM), and a broad range of Instant Capacity and Pay-Per-Use solutions, applications receive exactly the resources they need. The hardware complex can adjust to the changing requirements automatically and in real time. TPC-H results demonstrate the value of these Adaptive Enterprise components.

Reliability and high availability

In large data warehouse environments, the potential for failure is higher than usual because of the large numbers of I/O components. These components (I/O cards and channels) have been statistically proven to be the highest cause of failures. Additionally, even though they fail infrequently, there are many CPUs.

The Integrity Superdome architecture has many features that have been designed to ensure high levels of reliability. This design strategy prevents system failures caused by traditional areas of component failure.

- Instant Capacity—Automatically replace a failed CPU without a reboot
- Extensive memory error and protection
- Chip spare (or chip kill) to protect against multibit memory errors
- Memory scrubbing
- Page deallocation
- Cache error protection
- Partitioning to isolate applications for availability
- N+1 redundancy of components
- Multipath I/O
- I/O card isolation and Online Addition and Replacement (OLAR) of PCI cards
- Multiple power supplies

Oracle RAC is a highly available clustered database. If any node fails, the remaining nodes immediately recover the work of the failed instance and carry on processing the application workload. However, in a complete enterprise data warehouse, there are many components and

applications not under Oracle RAC management. HP Serviceguard provides complementary high availability cluster infrastructure services solution for these other components. HP Serviceguard has been in production use for many years and has a large installed base of customers.

In addition to being the most robust complement to Oracle RAC, Serviceguard also provides Application Packages—a capability that automatically fails over applications to surviving nodes in the event of node failure. Application Packages can be used in conjunction with Oracle RAC.

Serviceguard has also been extended with data replication solutions to provide disaster-tolerant solutions over geographically separated data centers. This family of solutions includes HP Extended Campus Cluster, Metrocluster, and Continentalclusters.

Why HP StorageWorks XP Disk Array storage?

The Integrity Superdome design reflects extensive data warehouse workload testing. One key result was the integration of a large number of I/O slots and very high throughput capacity. Large sequential scans and large multitable joins will particularly benefit from the large I/O bandwidth of the Superdome cluster.

The XP128 Disk Arrays, as well as other members of the XP storage array family (including the XP1024 Disk Array and the new HP StorageWorks XP12000 Disk Array), provide best-in-class, industry-leading performance for high-end BI workloads. BI workloads are characterized by large sequential scans (large I/Os and read intensive). Not all storage arrays perform well for both BI and traditional OLTP workloads (small I/Os and write intensive). Some arrays are well-suited for OLTP but not as good for BI. The XP array family is particularly well-suited to both types of workloads, given their high sustainable I/O throughput, large number of host ports, large number of controllers, high bandwidth internal crossbar switch, and high scalability (up to 144 TB of usable capacity in the XP12000 Disk Array).

The XP128 Disk Array series also delivers simplified storage management. They give customers a wide range of storage options for large data warehouses. Key benefits of the XP family include extremely flexible configuration, hot replacement, and seamless addition of channel connects.

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 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 the throughput requires modeling with real customer data.
- Overall storage requirements could be reduced significantly by using compressed data in the larger tables. This solution 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.

Why Oracle Database 10*g* and Oracle Real Application Clusters?

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

- Shared everything architecture
- Real Application Clusters
- Partitioning

Shared everything architecture

An Oracle data warehouse has great flexibility for growth because of its shared everything architecture. In this model, each node accesses all the storage. Customers can add servers, storage, or both separately, on demand. In the shared nothing architecture, each node has its 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 RAC supports the autoprovisioning capability of "grid" technology such as the HP Adaptive Enterprise. Customers can add or remove CPUs as needed and when the application is running. Oracle RAC 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 RAC matches the requirements of 24x7 data warehouse systems.

Oracle RAC for scalability, availability, and manageability

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

Oracle RAC 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 RAC 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 RAC eliminates guesswork regarding the processing ability of the database. If a single node fails in a 12-node cluster, then the data warehouse performs at $11/12^{th}$ of full performance. 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 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 locations. A listener can send an incoming client request for a specific service to the least-loaded node, instance, or dispatcher.

Another major benefit of Oracle 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. Oracle 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 RAC 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 is becoming a self-managing database, with the introduction of automatic performance diagnosis and tuning recommendations. The built-in tools in Oracle Database, 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 cost low.

Partitioning for data management and performance

The Oracle 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. 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. The 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 by enabling parallel execution, partitioning improves the data warehouse performance. The performance advantage of partitioning is achieved through the following features.

- Partition pruning. Oracle RAC 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, then there is no need to retrieve data for the remaining eleven months. Pruning can dramatically reduce the data volume to improve query performance. Partition pruning works with all other Oracle Database other performance features. Oracle RAC 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 benchmark configuration demonstrates that a Superdome/HP-UX/Oracle Database 10*g* with RAC solution using HP StorageWorks XP Disk Array series storage can be put in production for extremely large data warehouse projects. The solution meets the key criteria outlined at the beginning of this paper.

- **Scalability.** Throughput is the key to enabling performance. It is achieved by balancing CPU capacity and I/O bandwidth. Within the high-end SMP server and XP disk array storage environment, planning ahead for throughput growth will pay dividends in terms of simplifying capacity management and minimizing reconfiguration. For instance, support for long-term growth might be better served by initially deploying multiple lightly loaded Superdomes and XP disk array storage frames than the lowest possible number of fully loaded configurations. This approach allows orderly, evolutionary growth. Building in bandwidth early is relatively inexpensive, especially because HP offers solutions such as Instant Capacity and Pay-Per-Use.
- **Performance.** The cluster contained 128 Integrity CPUs and provided exceptional price:performance relative to traditional high-end, shared nothing solutions. The InfiniBand technology and ability to add I/O throughput means that the overall configuration can easily reach even higher levels of performance. The XP128 Disk Array-based storage fabric enabled throughput of 20 GB/s, with additional throughput available as nodes and processors are added to the cluster. Oracle RAC will optimize performance automatically by updating partitioning as new storage is added to the arrays.

- **Cost.** The Superdome servers help to reduce overall solution cost by providing fewer systems to manage. CPUs can be added inexpensively to improve performance or provide failover redundancy. Oracle RAC can also seamlessly integrate extra nodes into the cluster. The XP128 Disk Array delivers a cost-effective storage subsystem with a high degree of parallelism, redundancy, and performance.
- **Manageability.** Oracle RAC simplifies administrative tasks such as partitioning and query optimization. Because Oracle RAC enables all database instances to be managed in parallel, it reduces costs and ensures consistent management across the cluster.
- **Reliability and availability.** The Integrity Superdome and XP128 Disk Array storage offer strong availability and reliability capabilities. These products extend the benchmark configuration to better support mission-critical business intelligence applications. Serviceguard supplements Oracle RAC to provide a complementary range of high availability services to the rest of the data warehouse hardware and software complex.

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

- **Servers and memory.** Superdome servers provide excellent scalability when fully loaded with 64 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 or partitions configured for that workload. Cluster scaling of 91% demonstrates the exceptionally linear performance gain available from high-end SMP server clusters.
- **Cluster interconnects.** Oracle RAC efficiently allocates query workloads across the cluster using the InfiniBand switch. Adding nodes without adding I/O bandwidth can result 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 (connects and arrays) is sufficiently distributed to provide adequate sequential 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 all hardware, software, and storage components common to large data warehouse environments.

For more information

- Full disclosure report for 10-TB, two-node, 64-way Superdome/Oracle 10*g* RAC cluster
  http://www.tpc.org/results/FDR/tpch/hp_tpch_sd_10TB_RAC1_fdr.pdf
- Full disclosure report for 10-TB, one-node, 64-way Superdome/Oracle 10*g*
  http://www.tpc.org/results/FDR/tpch/hp_tpch_sd_10TB_fdr.pdf

5983-1280EN, 03/2005

ORACLE

hp invent