Calculating the Return on Investment for Clustered Caching and Data Grid Solutions

An Oracle White Paper
Updated May 2007

ORACLE

Calculating the Return on Investment for Clustered Caching and Data Grid Solutions

This document is intended to demonstrate ROI in Oracle Coherence deployments for the purpose of budget allocation and also to demonstrate that ROI can be achieved in the current budget cycle. This information can help you ascertain whether Oracle Coherence holds the same dramatic potential for your organization, by giving you a framework for your own cost justification and ROI analysis.

INTRODUCTION

At a time of increasingly strict financial controls, improvements to operational systems must have a quantifiable return on investment (ROI). These enhancements must either be explicitly budgeted or must pay for themselves in the current budget cycle to avoid a visible budget impact. This document is intended to demonstrate ROI in Oracle Coherence deployments for the purpose of budget allocation and also to demonstrate that ROI can be achieved in the current budget cycle. This information can help you ascertain whether Oracle Coherence holds the same dramatic potential for your organization, by giving you a framework for your own cost justification and ROI analysis.

Oracle worked with several customers to create a common list of ROI factors. These include considerations such as the value of improved service-level performance, full availability with 100 percent reliable data, and improved resource utilization with unlimited scalability. This document includes some of these customers' statistics as examples that may help guide your own analysis.

For each of the targeted categories, you will find a series of pertinent questions. Although the answers will be unique for each customer, we have included what we consider conservative estimates of potential savings impacts, based on customer feedback, where appropriate.

TODAY'S UNIQUE CHALLENGES

We are an information society. For most organizations, whether they are in finance, insurance, communications, or government or are any other type of institution, competitive advantage is determined by how intelligently, efficiently, and cost-effectively they manage and exploit their information systems.

Today's users expect consistent high performance from their applications, no matter how massive or erratic the throughput demand. They demand continuous availability of information with 100 percent reliability, even in the event of hardware or software outages. And the cost of scaling must be predictable and economical, even with explosive user and data growth. But budget constraints often force compromises in high application performance, with disastrous consequences. And data bottlenecks can limit peak performance and frustrate users. One of the

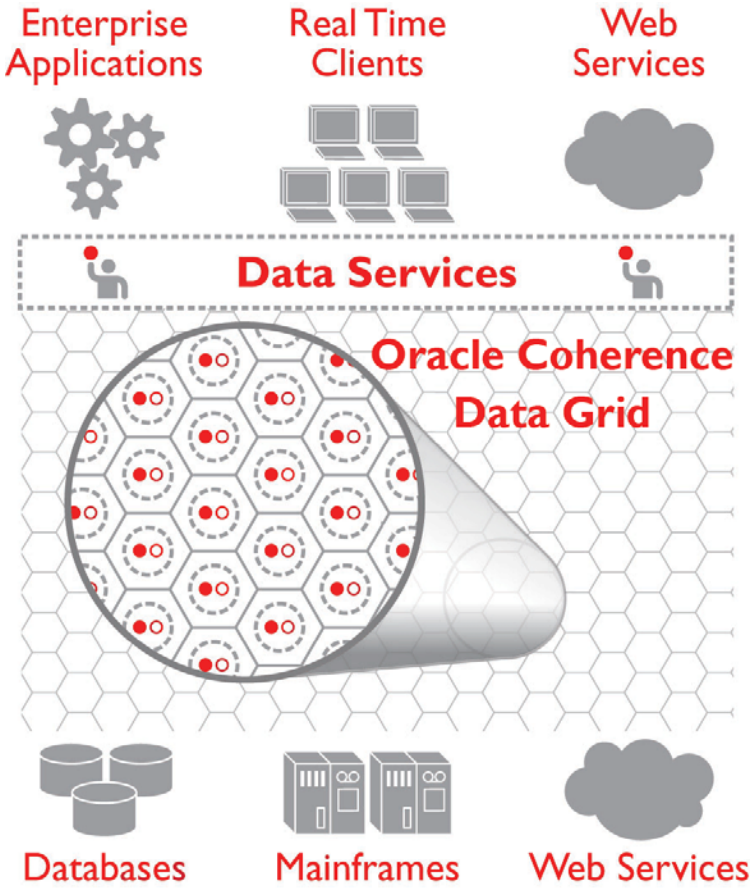
most common bottlenecks limiting peak performance is the rate at which the application can be fed the data for completing its processes. Applications—and, by extension, end users—spend most of their time waiting for data.

In fact, there is no greater infrastructure challenge for many mission-critical applications than solving the data bottleneck. It prevents predictable and cost-effective scalability, and it directly affects the performance and throughput of every data-intensive application. Solving the data bottleneck offers amazing payback opportunities. Time is money for most mission-critical applications.

THE ORACLE COHERENCE SOLUTION

Oracle Coherence helps organizations meet extreme data access demands while using commodity hardware and unique clustered data management and data grid technologies.

Oracle Coherence helps organizations meet extreme data access demands while using commodity hardware and unique clustered data management and data grid technologies. It provides distributed data management and caching services on top of a reliable and highly scalable peer-to-peer clustering protocol. Frequently used data is maintained safely and securely in memory, delivering instantaneous response and eliminating costly database accesses.



Oracle Coherence data grid solution

With no single point of failure and a sophisticated failover system that distributes its clustered data management services to all available servers, Oracle Coherence

maintains availability and data integrity even when a server becomes inoperative or is disconnected from the network.

Managing data at the application tier rather than at the database tier makes it possible to achieve virtually unlimited and nearly linear scalability simply through addition of servers to the application tier. This avoids the exponential price increases associated with adding database capacity.

Managing data at the application tier rather than at the database tier makes it possible to achieve virtually unlimited and nearly linear scalability simply through addition of servers to the application tier. This avoids the exponential price increases associated with adding database capacity.

SERVICE-LEVEL PERFORMANCE IMPROVEMENTS

Because Oracle Coherence manages data safely and securely in memory, subsecond response time becomes a reality for transaction-based applications, regardless of load.

Similarly, for long-running data-intensive batch applications, such as risk management in the financial sector, the impact of Oracle Coherence can be dramatic. Several customers have reported batch-job times reduced from several hours each day to just a few minutes.

Here are some results Oracle Coherence users reported in this category.

Service-Level Performance	
GEICO	• GEICO immediately improved page-turn time by one whole second, dropping well below its service-level thresholds.
Allegient Systems	• While transaction volume grew by 27 percent, service level to end users improved by 53 percent. Average time for processing rate changes for law firms has been reduced by 50 percent, enabling the company to process 30 percent more invoices per day with no additional hardware or staff. • Productivity improvements have enabled Allegient to delay and reduce staff growth without negatively affecting customer responsiveness, in spite of the 50 percent annual business-volume increase.
Betfair.com	• Betfair.com is providing consistent instantaneous response to all of its users, even through peak surges exceeding 1,000 simultaneous transactions.
TheServerSide.com	• Response time has improved 100 percent, with consistent subsecond response to users, regardless of the load on the system.

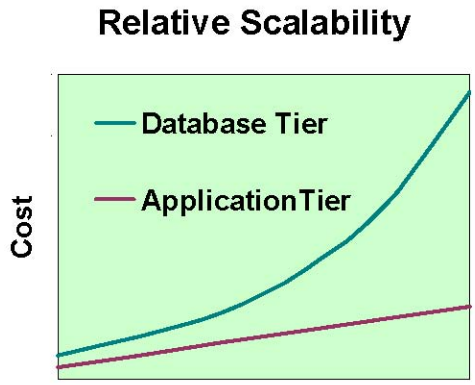
Dealer.com	• In one part of the application that was heavily dependent on the database, a three-hour job was reduced to two minutes. • The use of the XML bean feature has resulted in a cache size reduction of 50 percent and an application performance increase of 30 percent. • Dealer.com makes extensive use of the Oracle Coherence distributed query facility in applications such as QuoteFactory, which enables potential car buyers to rapidly home in on the vehicle of their choice. The search performance is more than twice what Dealer.com measured with a database, driving significantly more leads to Dealer.com customers. • Vehicle searches were improved from 10-second waits to literally instant responses, and back-office processes that used to take hours now take a few minutes. • While talking to prospective dealers on the phone, Dealer.com's sales representatives are able to tailor a demo site to the specifications for each dealer. As a result, sales cycles have shortened, close ratios have increased, and sales productivity has increased by more than 500 percent. • Even though the rate of adding new customers has increased 500 percent, the production team responsible for implementing new customer applications has not increased. • For the internal customer relationship management (CRM) application, the entire account history is maintained in cache. Retrieving and updating account information has been reduced from minutes to less than a second. Customer support call time has been reduced by a factor of 2.
INCERNO	• INCERNO has processed 5,000 transactions per second with no discernable deterioration in performance or reliability.
Large financial institution	• The time to run a risk-management application daily was reduced from nine hours to 15 minutes.

Service-Level Questions to Ask

- What are the critical performance objectives (service levels) for your application?
- What happens if your application fails to meet its performance objectives?
    - How often does that occur?
    - Why does that happen?
    - Who else is affected?
- What would it mean to your organization if you could deliver consistent subsecond response to your users, regardless of load or unpredictable surges? Or, stated another way, what does unsatisfactory and/or inconsistent service-level performance cost you today?
    - Staff productivity impact
        - If we could provide consistent subsecond response to this application, how would that affect staff productivity and costs in terms of overtime or staff size and growth?
    - Customer impacts
        - For applications that touch customers, either directly or indirectly, how would it affect customer satisfaction, sales volume, and customer retention if we provided consistent subsecond response time?
    - Penalty avoidance
        - What penalties or charge-backs will we eliminate if we are able to consistently achieve defined service levels?

SCALABILITY AND RESOURCE UTILIZATION IMPROVEMENTS

One of the greatest frustrations that owners of data-intensive, high-growth applications experience is their inability to predict or manage costs. The problem is that the database tier is difficult and expensive to scale. Even with clustered databases, scaling database services usually involves buying larger servers, whose price increases exponentially as the number of CPUs increases. And application servers in a cluster usually have to disable their caches to avoid corrupting the database, which means they put even more load on the database.



Relative cost comparison of scaling the database and application tiers

Although applications can easily flood database servers with queries and transactions, it is easy to scale the application server tier by adding more machines. Applications generally scale almost linearly for business logic and page generation.

It is important to select the caching methodology appropriate for your application environment. Each cache technology has its own performance characteristics, and applying the wrong caching methodology to your applications will not ensure scalability.

Managing frequently used data in memory means that the load on the database can be reduced dramatically. For example, Allegient Systems recently ran a before-and-after test to determine the impact of Oracle Coherence on its database load. Even though the number of concurrent users after the Oracle Coherence test was 27 percent higher than before it, the reduction in database load was amazing:

- Logical reads (data block reads from DBMS memory) declined by 48 percent.
- Physical reads (data block reads from DBMS disks) dropped by 35 percent.
- Costly database sorts were reduced by more than 90 percent.
- The number of statements executed by the DBMS was reduced by almost 60 percent, and the maximum statement execution time was cut in half.

The net effect was that whereas the transaction volume grew by 27 percent, overall database load was reduced by 58.8 percent. At the same time, Allegient Systems measured a 53 percent improvement in the service level it delivered to its end users.

A word of caution: it is important to select the caching methodology appropriate for your application environment. Each cache technology has its own performance characteristics, and applying the wrong caching methodology to your applications means that you are not ensuring scalability. A fully replicated cache, for example, provides the highest data access speed up to a point but does not scale linearly as the cluster grows if there is a high incidence of updates.

Oracle Coherence provides an extensive range of caching services, including replicated caching, distributed (partitioned) caching, and near caching. It also includes a wide variety of unique functions tailored to further enhance scalable performance for specific applications—features such as write-behind caching, HTTP session management, and parallel query capability.

For that reason, Oracle Coherence provides an extensive range of caching services, including replicated caching, distributed (partitioned) caching, and near caching. It also includes a wide variety of unique functions tailored to further enhance scalable performance for specific applications—features such as write-behind caching, HTTP session management, and parallel query capability. Selecting the appropriate Oracle Coherence caching methodology can provide linear scalability and extreme performance with virtually unlimited data capacity. Here is what our customers tell us:

Scalability / Resource Utilization	
GEICO	• GEICO estimates that it eliminated approximately US$700,000 in capacity expansion costs due to the resource-utilization efficiencies of Oracle Coherence.
Allegient Systems	• While the transaction volume grew by 27 percent, overall database load was reduced by 58.8 percent. Allegient Systems estimates that the more effective use of DBMS resources will save US$300,000 in DBMS hardware, software, and maintenance upgrade fees over the next 12 months. Within a few years, these savings will compound to millions of dollars in cost avoidance of recurring database upgrades to support Allegient Systems' explosive growth.
Betfair.com	• Betfair.com's 1,500 percent improvement in caching effectiveness has significantly reduced both the database load and the application server load. The resulting savings has enabled the company to pursue new capabilities that otherwise could not have been considered because of the cost of necessary incremental hardware. • With Oracle Coherence, Betfair.com is now able to safely and securely grow its cluster to hundreds of servers while still delivering subsecond responsiveness during enormous surges.

Dealer.com	• Dealer.com's Oracle Coherence implementation resulted in an immediate 90 percent reduction in database load. • Before Oracle Coherence was used, each application server was capable of supporting 200 dealer sites. Oracle Coherence doubled the capacity to 400 dealer sites per server. • Dealer.com faced the scalability problem of rapidly diminishing returns as new servers were added. Now that Dealer.com is powered by Oracle Coherence, each additional server adds support for another 400 dealers instead of 200 or fewer, keeping the cost of growth both low and predictable. • The cost of scaling the system to retain performance will remain almost linear, to more than 50,000 dealership customers.
INCERNO	• INCERNO was able to cut hardware costs from an estimated US$300,000 to less than US$60,000.
TheServerSide.com	• At its current growth rate, TheServerSide.com will be managing almost 10 times its current transaction volume within two years. The load on the database server, which was running at 100 percent capacity prior to Oracle Coherence, has been virtually eliminated. • The capacity of its application servers has risen tenfold. This gives TheServerSide.com substantial, immediate incremental capacity, delaying investment in application servers. And when additional capacity is required, each server will satisfy the demand growth 10 times as long.

The preceding results show that users experience a big increase in the effective capacity of their application servers in addition to their database servers. This is due to the reduction in the overhead incurred by the server in executing database reads and writes and reflects dramatic efficiency improvements resulting from the use of the Oracle Coherence distributed cache functionality instead of far-less-efficient clustered server-management alternatives.

Scalability Questions

- What is the peak load on your database today?
- At your expected database load growth rate, when will you reach saturation (100 percent sustained load) of your database server?
- What will be the cost of upgrading, including software, hardware, and maintenance?
- What would be the impact on these costs if you could cut the current load by (__) percent and reduce the rate at which the database load is growing by the same percentage? (We cannot recommend or guarantee a fixed percentage for you to use. It all depends on the nature of your application. However, based on what we have seen, we can say that for most applications, implementing Oracle Coherence would conservatively result in a database load reduction of 50 percent. Many customers have achieved a reduction in database load of greater than 90 percent. TheServerSide.com saw its load go from saturation to virtually nothing.)
- What would be the impact on your costs if the effective capacity of your application servers were increased by (__) percent and the rate at which you had to grow the cluster were reduced by the same percentage? (Again, we are unable to recommend a fixed percentage for you to use. A 20 percent improvement might be a reasonable and conservative estimate, based on our users' feedback.)

AVAILABILITY AND DATA RELIABILITY IMPROVEMENTS

Because Oracle Coherence has no single points of failure, users are assured of 100 percent data integrity, even in the event of a server or database outage.

Because Oracle Coherence has no single points of failure, users are assured of 100 percent data integrity, even in the event of a server or database outage. It automatically and transparently fails over and redistributes its clustered data-management services when a server becomes inoperative or is disconnected from the network. Oracle Coherence also includes network-level fault-tolerance features and transparent soft restart capability to enable servers to self-heal.

To handle increasing load, a user can add extra capacity on demand, simply by expanding the cluster with additional servers, using commodity hardware. When a new server is added or a failed server is restarted, it automatically joins the cluster and Oracle Coherence fails back services to it, transparently redistributing the cluster load.

Many of our users report continuous uninterrupted operations since installing Oracle Coherence. All report 100 percent data integrity and reliability, even in the event of server or database failure. Some of the stories are quite dramatic. Dealer.com, for example, reports that an unplanned three-hour database outage that occurred during the business day had no impact on any part of its application,

including the transaction-heavy CRM and back-office functions. And no data added or transactions that occurred during the database outage were lost.

Continuous Availability / 100 Percent Reliability Questions

• Do you ever experience unplanned service interruptions, and if so, what are the most significant applications affected?
  o How often does that occur?
  o What are the causes?
• What are the associated costs in terms of
  o Staff productivity loss?
  o Customer sales?
  o Customer satisfaction and retention?
  o Other?
• Do you ever lose data or transactions?
  o What are the associated costs?

BUY-VERSUS-BUILD CONSIDERATIONS

For customers that require only a fraction of the functionality provided by Oracle Coherence, the question of buy versus build must be answered. The decision criteria usually factor in the cost of building; cost of maintenance; and risks associated with new and unproven code in comparison with code that has been tested, refined, and proven by years of experience through hundreds of organizations. Furthermore, whereas home-built code represents a maintenance liability, Oracle is responsible for continually improving and refining Oracle Coherence. Here is some of our customers' feedback:

Buy Versus Build	
Betfair.com	• Oracle Coherence made it unnecessary for Betfair.com to add expensive development staff with experience in clustering and distributed caching. It enabled Betfair.com to focus on its business goals instead of developing distributed cache and HTTP session management software, and it let the company avoid making a long-term commitment to a development project estimated to cost at least US$500,000. Oracle Coherence also eliminated considerable risk, given that an in-house solution would have been new and unproven code.

Jive Software	• Jive Software saved at least US$30,000 in development costs, which is what it would have cost it to produce the minimal services that it would have provided with an internal design. • Jive Software brought a stable, feature-rich product to market at least two months sooner than it would have been able to with a homegrown solution. • The company eliminated the risks associated with designing, developing, and delivering new unproven technology. • It has added important functionality to its product that would not have been possible without Oracle Coherence. • The Oracle Coherence invocation service function has been instrumental in accelerating the ability of Jive Software to implement new functionality in its software products and has enabled it to build unique features such as "dedicated search" into its products that would not have been feasible otherwise.
INCERNO	• INCERNO eliminated several man-weeks of internal software design, development, and maintenance, saving at least US$25,000. The decision to buy rather than build caching capabilities also eliminated the risk of using unproven new software that could put the launch schedule in jeopardy. It also eliminated the ongoing complexities of in-house-software maintenance.

Buy-Versus-Build Questions

- Is there a deadline associated with your project?
    - What are the risk and cost of missing it?
    - Who is affected if you miss it?
- How much experience do you or your staff have in developing Java 2, Enterprise Edition (J2EE) distributed cache applications?
    - What would it cost you to develop the skill and experience necessary to ensure success in this environment?
- How much time would you expect your staff to spend on maintaining the distributed cache environment?
    - What is the cost?
- What is the risk that it will require more time, and what will be the impact?
- How often do you expect to make changes to your topology? What will be the cost of adapting your distributed cache/clustered server applications?

SUMMARY

Once you have quantifiable answers to the preceding questions, you have all the information you need for completing your cost justification and ROI analysis. A completed sample worksheet is included in the appendix. This analysis shows a company's full ROI as a result of more-effective resource utilization and scalability. Because of greatly improved service-level performance, the company also achieved substantial returns from staff efficiency improvements, accelerated customer growth, and reduced customer turnover. This sample is not a precise analysis provided by any user but is a realistic approximation based on stated impacts and costs.

SAMPLE RETURN-ON-INVESTMENT ANALYSIS (US$1,000s)

	Q1	Q2	Q3	Q4	1st Year
BENEFITS					
Profit from Increased Revenues					
Sales doubled by performance, reliability, and demo	$ 200	$ 200	$ 200	$ 200	$ 800
Recurring revenue increase existing cust	$ 200	$ 200	$ 200	$ 200	$ 800
Added new cust recurring revenue	$ 42	$ 84	$ 126	$ 168	$ 420
Reduced turnover	$ 21	$ 21	$ 21	$ 21	$ 84
Reduced Costs					
Displaced Costs					
Eliminate 1 support rep	$ 25	$ 25	$ 25	$ 25	$ 100
					$ -
Avoided Costs					
Increased productivity defers support rep hires	$ 25	$ 25	$ 50	$ 50	$ 150
Increased productivity defers sales rep hires	$ 25	$ 25	$ 50	$ 50	$ 150
Database upgrade deferred	$ 120				$ 120
Application server effective capacity increase	$ 5			5	$ 10
Total Qtrly Benefits	$ 663	$ 580	$ 672	$ 719	$ 2,634
Cumulative Benefits	$ 663	$ 1,243	$ 1,915	$ 2,634	
COSTS					
Hardware					$ -
Software	$ 50				$ 50
Maintenance					$ -
Services	$ 10				$ 10
Demo application implementation	$ 100				$ 100
Total Qtrly Costs	$ 160	$ -	$ -	$ -	$ 160
Cumulative Costs	$ 160	$ 160	$ 160	$ 160	
NET RETURN					
Qtrly Net Return	**$ 503**	**$ 580**	**$ 672**	**$ 719**	$ 2,474
Cumulative Net Return	**$ 503**	**$ 1,083**	**$ 1,755**	**$ 2,474**	

1st Year Net Return	**$ 2,474**
Break Even Point	**Immediate**
First Year ROI	**1,546%**

Note: The assumptions for this model are explained in the following section. This model is intended only to show how the worksheet should be used and is not intended to suggest reasonable expectations for you. For some users, the elements included here and their associated values would be unreasonable, whereas for others they would be conservative.

VALUE AND COST ESTIMATES: SAMPLE ROI ASSUMPTIONS

Background

This is a fictitious high-growth application service provider (ASP) with IT systems performance and availability problems that are negatively affecting both sales and recurring revenue, causing customer cancellations, and degenerating in-house user efficiencies—particularly in the customer support area.

The database is frequently maxed out. The company estimates that the cost to upgrade to provide the added capacity needed for the immediate future will be $100,000, with a $20,000 increase in maintenance fees. It is finding it necessary to add another application server every four to six months to accommodate new customers.

Increased Revenues

The company believes that if it can deliver performance consistently within defined service-level standards and maintain continuous availability, it will influence revenue significantly in three ways:

- **New sales** – Because its sales model relies heavily on demonstrating how its prospects' personal sites would look, unresponsive or unreliable IT service during sales presentations has a serious impact on sales performance. The company estimates that it could double sales productivity with a reliable high-performing demo site. Quarterly new sales revenues of $200,000 and 20 new customers could be doubled without the addition of staff, eliminating the need to add a new sales rep every two quarters.
- **Recurring revenues** – The current base of 2,000 customers generates an average of $2,000 per customer per quarter in usage fees. That could easily be improved 5 percent with responsive and reliable service, resulting in an additional $200,000 per quarter in recurring revenue. As new customers are added, their usage would be at the higher rate, adding a cumulative $42,000 per quarter.
- **Customer turnover** – Customer turnover is a big problem that could be fixed by elimination of end-user frustrations. The current turnover could be reduced by 10 customers per quarter, saving $21,000 per quarter.

Reduced Costs

Displaced Costs

- **Customer service** – Call activity related to system availability and performance will be eliminated, meaning a staff load reduction of one full-time person ($25,000 per quarter, fully burdened).

Avoided Costs

- **Customer service** – Having full customer records instantly available will reduce time per call by 50 percent, doubling team productivity. This will defer new hires—one in Q1 and another in Q3.
- **Sales** – A high-performance demo and high-quality/high-performance system will double sales productivity, deferring the need to expand staff by one additional hire in Q1 and one in Q3.

- **Systems and database resources** – Database load is reduced by at least 60 percent, eliminating the need for an immediate upgrade that would have cost an additional $100,000, plus an additional $20,000 per year in maintenance costs.

  Currently each application server handles 200 customers. Effective capacity will increase to at least 300 customers per server, deferring the need for an additional server in Q1 and Q4.

Note: These assumptions are intended only to show how the ROI worksheet should be used and are not intended to imply values appropriate for individual usage.

Calculating the Return on Investment for Clustered Caching and Data Grid Solutions
Updated May 2007

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com