The Impact of Data Center and Server Relocation on Application Performance

A Proven Strategy for Project Success

A Shunra Software Best Practices White Paper By Amichai Lesser
Executive Summary

According to analysts, 88% of enterprises are currently in the early phases of either consolidating servers or executing a major data center relocation. These consolidations and relocations bring measurable benefits. For example, they reduce Information Technology (IT) costs, improve business continuity and information security, optimize service management, and help comply with federal and industry regulations.

With so many benefits, why haven’t more companies completed such initiatives? According to Gartner, the primary factors inhibiting these beneficial projects include internal politics and pressures, often based on business owners’ concerns that their critical applications and services will be adversely affected by such moves.

Such concerns are legitimate. Consolidations and relocations can impact service levels in many ways. Users that were previously local to the servers supporting their business applications, for example, may become remote users. Few applications are “future-proofed” against these types of network changes.

The impact of this physical displacement can be dramatic. It can include:

- A 3x to 100x increase in application response time due to the multiplying effect of network latency and impairments on application performance
- A severe reduction in the productivity of server to server infrastructure processes such as copy jobs, replication and backup.
- A reduction in application scalability for supported users due to the queuing effect, that longer transaction times have on server resource allocation
- Downtime and business disruptions due to unexpected application behavior during the interim stages of relocation or consolidation

In fact, 20-40% of applications will fail to meet service level objectives during or after a data center relocation, some with severe issues that render the application unusable. The remediation of these service level issues will typically require either infrastructure enhancements or the re-coding of applications – both of which mean added cost and delayed time-to-benefit.

IT organizations that wish to prepare for such a move must fully understand in advance how the multiplying effect of network latency and impairments on application performance impacts critical business processes for end-users. Also, it is prudent to understand critical IT tasks such as backup and replication – and how to remediate any potential performance issues before the first server moves.

Fortunately, the success of relocation/consolidation projects can be protected through best practices planning and management. This white paper offers a proven strategy for predicting and resolving the application performance issues that arise when data centers are moved or consolidated. By following this strategy, IT organizations can instill confidence in their business units that service level commitments will be fulfilled and prevent relocation/consolidation projects from turning into high-stakes gambles. This strategy aims to identify problematic applications in advance and provide focused analysis to support the remediation of expected performance problems.

Five-Step Strategy for Success

At the heart of this strategy is the combined use of network emulation technology with application performance analysis that supports the following best practices steps:

**Step 1** Identify applications that will experience performance degradation under the post data center move conditions.

**Step 2** Perform deep transaction analysis on under performing applications and identify performance bottlenecks.

**Step 3** Remediate performance problems and validate solutions in a proof of concept lab.

**Step 4** Assess the latency impact on interdependent back-end servers and establish an appropriate move plan.

**Step 5** Manage user expectations and get business buy-in through hands-on acceptance and performance reports.

Why Are Data Center Moves So Tricky?

Businesses are increasingly dependent on applications and other IT services. Those applications and services, in turn, are heavily dependent on the underlying transport infrastructure that supports them. The relationship between infrastructure and application performance is a subtle and complex one.

Nowhere is the impact of the network on application delivery (and the associated risk) more apparent than when a business goes through a data center relocation.

While there are many types of data center and server relocation projects, they can generally be divided into two categories:
1. **Consolidation of multiple, small local data centers into larger regional data centers**

2. **The physical move of a large central data center to an entirely new location**

In some cases, aspects of these two categories may be combined – such as when data centers are both consolidated and relocated. The principles set forth in this paper are applicable to all such moves. Changing the data center’s location directly impacts the way applications are delivered over the network. Users who were previously local to front-end applications or back-end infrastructure such as policy directories, domain servers and other desktop services find that they have become remote users. Because they access those servers from a distance, they can potentially experience much longer transaction response times in their business applications as well as for common tasks such as file copies, desktop boot up and login.

On top of that, servers that were previously able to support many local users completing quick transactions are called upon to support an increasing number of users processing slower transactions due to added network latency. This has a significant, adverse impact on the server’s scalability as more and more transactions queue up. In addition, data center consolidations are often accompanied by a transition to web-based applications that extend application access to users across the enterprise, at home and on the road. Thus, at the same time latency and scalability issues are introduced, a significant number of low-bandwidth/high-latency users are also added to an already busy system.

Data center moves also present a unique challenge because they expose users to change. With the deployment of new applications, users have no point of reference for which to compare application performance. But with data center and server relocations/consolidations, users immediately notice even small performance degradations. So any shortcomings in planning and management will result in immediate complaints from the business, as well as customers.

**The Impact on Transaction Response Time**

Network-related delays change dramatically during a data center relocation, since geographic distance contributes significantly to the network latency between client and server. However, in the case of data center consolidations, this geographic latency is exacerbated by the large increase in wide area network (WAN) traffic as more users access servers remotely.

It is a dangerous misconception to equate network latency to application latency. The fact that 20 200 milliseconds (0.02 – 0.2 of a second) are added to the network delay doesn’t mean that the application’s latency or response time will only increase by that amount. On the contrary, the relationship between network delay and application response time is far from one-to-one. It depends on...
factors such as:

- The number of messages exchanged between the client and the server for each transaction (often referred to as application “turns” or “ chattiness”).

- The synchronous nature of the application and how blocking are each of the application turns (in other words, how many turns can be processed in parallel). Does the application wait for a certain objects (e.g. css, js, jar) to download before it can continue processing other data?

- The location and distribution of application code (js, jar, asp, php, swf), meta data (ini, XML, etc.) and the actual data (tables, images, objects, XML, etc.) and how these resources are fetched/downloaded in each transaction.

- The configuration of the protocol on the server and client (TCP buffer sizes, authentication methods, time out settings, maximal number of sessions and sockets, etc.)

The following chart, based on a log-in transaction for a CRM application, provides a relatively simplistic example to the relationship between network latency and application latency.

![Figure 3: Transaction response time – local user vs. remote user](image)

Note: The “login page download” arrows represent over 500 network packets broken into 270 application turns

For a local user with less than a millisecond of network delay between the client and the server, the transaction completed in three seconds. When just 50 milliseconds of network delay were introduced (representing a typical cross-country WAN connection), the performance of this transaction did not slow down by only 50 milliseconds. Instead, this same login transaction took a full 30 seconds to complete. This example dramatically illustrates how small changes in network latency result in major problems with application performance.

This deterioration in performance can be clearly understood by analyzing the behavior of the login process. Deep packet transaction analysis shows that the application generates 270 “blocking application turns” to complete this task. As the round-trip time for each application turn increases from 1 millisecond on a LAN to 100 milliseconds across the WAN, the transaction response time increases by 27 seconds as the 100 msec network latency becomes a multiplier of 270 application turns which adds 27 seconds (270 application turns * 100 msec = 27000 msec which are 27 seconds).

Not all applications and business processes were created equal. Some will have different degrees of sensitivity to latency based on the factors described above. That is why any team planning a data center relocation must understand which applications will be most severely impacted by the new latency – allowing IT and engineering to develop appropriate mitigation solutions.

Unfortunately, the network latency impact is usually regarded as a network manager’s responsibility, even though application tuning and architecture changes may be a major contributor in overcoming latency-related performance issues. After all, network managers can’t change the speed of light, or make San Francisco closer to New York. So it is unreasonable to lay the problem entirely on them. In fact, since these types of application performance problems are more a result of application design rather than a network issue, the addition of more bandwidth alone will not have a major impact, and other network enhancements must be examined carefully before investing.

Furthermore, modern application development platforms such as .NET, FLEX, AJAX, SOAP, and J2EE tend to greatly exacerbate the network impact on application performance by enabling application developers to logically link objects across the network without giving any consideration to network characteristics such as distance and latency. This wouldn’t be an issue if all of the objects connected in this manner were local to each other. But they are not. As server moves and consolidations put greater distance between users and resources, application architects, developers and designers must pay more attention than ever to issues of distance and latency – or risk developing applications that simply can’t perform on real-world networks.

Developers and QA engineers need to embrace these best practices to assess and improve the latency tolerances of their applications, so that appropriate performance enhancements can be cost-effectively achieved and maintained for all critical business services.

In a complex project such as a data center move, the challenge of...
assessing the network impact on application performance becomes one of large scale as hundreds of servers hosting thousands of applications need to be tested for the impact of the data center move. Past experience shows that attempts to use only discrete event analyzers or deep packet analyzers to assess the network impact will fail to provide the needed information in a timely manner. A combined approach is needed; one that can quickly identify and isolate problematic transactions and then thoroughly analyze the root cause of (only) those problematic transactions.

**The Impact on Server Scalability and Processing Power**

The addition of network latency caused by server moves can also degrade the scalability and performance of the servers themselves. This often-overlooked phenomenon has an adverse impact on the entire user population – not just remote users. It is almost never caught in the QA process, and it is rarely diagnosed correctly even when it starts causing problems in the production environment.

How can network latency affect server performance? Well it all starts with the fact that network latency causes transactions to perform slower. As transactions take longer to complete, server resources are allocated for longer periods, queuing up these resources until some are exhausted and become a bottleneck, since a server allocates a variety of resources to each concurrent client session. Local clients (for example, prior to a data center move) complete these transactions quickly because their application turns experience minimal network-related delay. Remote transactions (for example, after a data center move), on the other hand, take much longer to complete because each application turn itself takes so much longer and as we mentioned before, application turns are multiplied by the added network delay.

It is important to understand that servers allocate resources at the beginning of a process, lock them for the duration of the process, and then only free them when the process is completed. Thus, when remote users communicate with a server, its resources are allocated and kept busy for a longer period of time. This prevents the server from releasing those resources for use by other clients – severely limiting its performance and ability to scale.

For example, consider a web server with a back-end database behind it. Client access to the system may be secured by the SSL protocol, and each client request may be handled by a dedicated operating system thread (i.e. a concurrent multi-threaded server architecture). In such a scenario, the sequence of tasks on the server for each client session would be as follows:

1) TCP socket handshake, 2) SSL handshake, 3) Open OS thread, 4) open separate TCP and SSL sockets for that thread, 5) The thread performs the transaction including I/O and network calls to the database, 6) close OS thread and thread resources, 7) close SSL session, 8) close TCP socket.

**Figure 4. Server Resource Allocation Timeline – comparing a remote user (top) with a local user (bottom)**

As shown, local users complete their transactions in a few seconds, allowing server resources (sockets, threads, database I/Os, etc.) to be quickly allocated, released and made available for the next client. With remote users, these resources remain occupied for a significantly longer period of time. In this particular example, a remote user’s transaction consumes resources that could have been reused nine times on local transactions. In other words, the introduction of additional remote users radically shifts the server’s processing power from delivering actual application services to being “on hold”.

It is critical to understand this phenomenon in order to appropriately remedy the performance problems that inevitably result when remote access grows as a result of a server move. More specifically, when a server reaches a scalability limit with local users, performance and scalability can usually be improved with a CPU upgrade or by adding a server to the cluster. However, when a server reaches a scalability limit with remote users, it typically makes more sense to upgrade RAM, tweak the process scheduler, increase the concurrent thread limit, and off-load SSL processing.

Because of this relatively unknown phenomena, it might be useful to analyze the server’s performance under load in combination with the network latency and provide metrics that the business can understand.

One such metric is the “latency scalability factor”. This factor can be regarded as the difference in server scalability when serving local users vs. remote users.
For example, a server that can scale to support 1000 local users is then stress tested with a ramp up of remote users over a 50 msec latency link and found to only scale to support 700 users. This application is said to have a 0.7 latency scalability factor (at 50 msec latency).

Unfortunately, conventional QA and performance practices typically focus on back-end scalability, with little or no attention given to the network latencies between users and those servers. That is why IT organizations rarely address this phenomenon prior to a data center move. However, that approach can be detrimental to the success of this project as this scalability issue can manifest itself as a dramatic increase in transactions in flight (processed or queued transactions) beyond the server’s capacity. This can cause transactions to be dropped from the queue and even (in extreme cases) require the reboot of servers to release queued resources every so often. Also, since this type of problem only manifests when a certain combination of local and remote users access the system, it tends to be intermittent and difficult to diagnose in production, making these issues particularly aggravating for system administrators and network managers.

The server capacity planning team has a crucial role to play in preventing this phenomenon from affecting enterprise applications as a result of a data center relocation. In their testing, the team must consider that servers will have to process requests from a mix of local and remote users, and they must ensure that sufficient and appropriate server resources are put in place to support both local users (where the quick creation and release of resources is key) and remote users (where resources are “kept open” for a longer period of time).

The question “How many concurrent users can the server support?” should actually be rephrased as “Can the server support the combination of local and remote users we anticipate?”

As we will see, this second question requires the use of a test bed capable of simulating local and remote users across the projected enterprise environment.

Maintaining Business Continuity During Data Center Moves

One of the more challenging aspects in a data center move is maintaining business continuity during the migration. Ideally, an enterprise might pack all of its servers in one weekend, load them on moving trucks, unpack in the new location, and be up and running by Monday. The reality is quite different. Enterprise data centers typically consist of hundreds or thousands of servers and other devices. It can take weeks, months, or even years - depending on the project - to complete a move to a new location. Thus, during this interim stage, some systems will operate from the original data center while others will operate from the new data center.

This transitional period forces IT managers to ask some critical questions:

- What happens to servers with inter-dependencies on back-end systems when they are no longer located together in the same data center?
- Which servers must be moved with other servers (move groups)?
- When should infrastructure servers such as network drives, Active Directory, domain servers and other desktop infrastructure be moved?
- Which servers will need to be replicated for the duration of the move?
- How will backup processes storage and business continuity processes (such as Robocopy, RSynch, VMotion, Dataguard, etc.) be affected during and after the migration?

The previous sections focused on the impact of network latency between clients and servers. Here, the issue becomes network latency that is temporarily introduced between different back-end systems. The impact on application performance when servers are separated across the network can be even more dramatic and unexpected, because those server based systems are almost never designed to accommodate significant latency between them.

Consider this following scenario: A credit application, which authenticates users on an Active Directory (AD) server, accesses a database to validate customer credit scores. The credit application server moves to the new data center location while the AD server and database server remain in the original location (since the latter two servers support other applications and weren’t scheduled to move in this phase). The mechanics of the application work as follows:

1. Business user accesses the credit application server (50 application turns)
2. Credit application server authenticates the user on the AD server (5 application turns)
3. Credit application server gets credit data from database server (200 application turns)
4. Business user receives an answer from the credit application server (5 application turns)

Note that there aren’t many application turns between the client and the front-end credit application server. But there are many application turns between the server and the database. Thus the distance between them during the move is likely to have a severe impact on application performance. See Figure 5.
This issue is rarely given adequate consideration during data center planning. In fact, experience shows that even in relatively well-planned data center relocation projects, this problem is a major cause of delays and service interruptions. For example, one leading pharmaceutical company was completely taken by surprise when a database translation batch process took five days to complete when its servers were separated during a data center move. That same process took a few hours when the servers were local to each other.

Data center relocation project managers must take these server dependencies into account during planning. Special attention, based on validated data, should be given to server move order, server grouping and replication – as well as the location and replication of storage components during the move.

Addressing Service Level Expectations

As stated earlier, Gartner points to political pressure as the main factor inhibiting the initiation of an effective data center relocation. Business managers are justifiably concerned about how the move will impact the performance of their critical applications. This concern intensifies when IT organizations can’t accurately predict how applications will perform during and after the move. Such uncertainty often delays projects and can push IT into over-spending on infrastructure in an effort to minimize the risk that they will fail to meet the expectations of business users.

In fact, managers will often voice concerns about application performance both before and after the move. A business unit manager may ask for dedicated networks, upgraded bandwidth, and/or exceptions to the server move. The problem, of course, is that this may do nothing to remedy potential application performance problems – since those problems often result from the increase in latency from greater physical distances between users and servers, rather than actual bandwidth constraints or other network issues. As a result, budgets are misdirected, plans to move everything are shifted, and the full benefits of the data center relocation or consolidation are not realized. That’s why it is critical to directly address users’ service level expectations up front. It simply doesn’t make sense to set a post-relocation Service Level Objective (SLO) that is identical to what had previously been a local response time. If it takes a local user three seconds to sign in before a move, it is very unlikely that the same process will complete this quickly after the servers are moved across the country. An SLO of seven seconds, for example, may be more reasonable. However, the key is to get business users to accept this new SLO prior to the move as part of the planning process – rather than forcing them to live with the change as a matter of circumstance after the move, and after promises have been made or unrealistic expectations set. To achieve this pre-deployment acceptance, two elements are needed. First, IT must have hard data about post move performance. That is, post-move performance must be predictable. Second, users need a way to see what post-move performance will actually look like. That is, IT needs to simulate post-move performance. This simulation enables IT to set up “acceptance environments” where users can experience post-relocation performance before the move is actually executed.

Predictability and simulation are invaluable to overcoming the concerns of business users, particularly from a psychological perspective. The setting of performance expectations upfront alleviates a primary source of political conflict and reduces the pressure to make unnecessary IT investments.

There is no substitute for allowing business users to see with their own eyes the type of performance they can expect – where they can buy into post-move SLOs or understand the investment needed for certain application remediation, before a single server is relocated.

A Five-Step Strategy for Project Success

From a technical perspective, the problems that may develop in application performance as a result of data center moves and consolidations are not impossible to solve. The real challenge is overcoming the issues that arise from subtle, complex interactions between applications, networks and infrastructure. Traditionally, responsibility for these three areas was split between separate operational “silos.” In truth, such a siloed approach reduces the likelihood that IT organizations will successfully predict and
address performance problems resulting from a data center move. It is therefore essential to take a new collaborative approach that effectively blends the expertise of the application team, systems managers and network architects.

In addition to working together in a collaborative manner, these IT groups should also follow proven best practices for server consolidation/relocation. These best practices are encapsulated in the following five-step plan for project success:

**Step 1**
**Identify applications that will experience performance degradation under the post data center move conditions.** In this step IT staff and business users go through a “week in a life” simulation of the post data center relocation conditions while using their business application. The simulation of the network conditions to the new data center location is done by the VE Desktop solution that can be deployed seamlessly to all relevant workstations and inject network conditions between the user’s desktop and application servers simulating the post data center move scenario. As users go about their normal business activities they will identify and flag any business process that will exhibit performance issues (at that point the application can be taken out of the simulation, so the user can still remain productive). It is expected that some business units may voice concern from impacting production users for what is considered an IT project. Nonetheless, it is important to communicate to the business units that this critical step will ensure that potential problems are identified early enough so they can be remediated in time for the move. Business units should also realize that it is much easier to stop the simulation (a click of a button) than it is to roll back servers due to unexpected performance problems.

**Step 2**
**Perform deep transaction analysis on under performing applications and identify performance bottlenecks.** In this step Subject Matter Experts (SMEs) and performance engineers take the list of problematic applications and perform deep transaction analysis on those under performing applications. By using the performance intelligence capabilities of the VE solution and isolating performance bottlenecks for each application, IT and engineering can get ahead of the problems and begin remediation efforts.

**Step 3**
**Remediate performance problems and validate solutions in a proof of concept lab.** As developers and IT staff engineer solutions for performance problems they should use the remediation lab to test out the effectiveness of proposed solutions (including code fixes, WAN acceleration, VDI solutions, etc.) and validate performance gains. The remediation lab should include a complete virtual model of the pre- and post-relocation enterprise environment, as well as all planned transitional phases. The need to assess the potential performance gains from hardware based solutions such as WAN accelerators require the use of hardware based network emulation. This can be achieved by using the VE Appliance which combines appliance based WAN emulation and application performance analysis.

**Step 4**
**Assess the latency impact on interdependent back-end servers and establish an appropriate move plan.** In addition to addressing network latencies between clients and servers, the team must fully understand the impact of latencies that may be created between servers during transitional stages of the move. Again, an appliance based network emulation solution is invaluable for accurately measuring the impact of these server-to-server latencies.

**Step 5**
**Manage user expectations and get business buy-in commitments through hands-on acceptance and performance reports.** Users who merely hear that a transaction response time will go from two seconds to five may object out of sheer reflex. When users are allowed to directly experience post-move application performance in advance, on the other hand, they can give their informed consent to the relocation plan.

By following this plan, IT organizations can substantially reduce risk, eliminate unnecessary infrastructure spending, accelerate time-to-benefit, and overcome a wide range of potential political pitfalls. The exclusion of any of these steps, however, greatly increases the likelihood that unforeseen problems will sabotage the project.

To ensure the success of any data center relocation or consolidation initiative, it is thus essential for IT to pool its expertise across multiple technical disciplines and diligently apply simulation/modeling to the planning process.

**Conclusion**

There are many issues that can arise during a data center move. These issues can threaten business operations and undermine the credibility of IT. Many of these issues cannot be predicted by IT organizations that operate in traditional silos, because they’re caused by subtle, complex interactions between networks, servers...
and applications. In fact, after these issues manifest themselves in the post-move production environment, many IT organizations are left struggling to understand and resolve them.

Fortunately, a data center move does not have to be a gamble. By applying proven best practices and appropriate simulation, testing and engineering techniques early in the process, IT organizations can bring predictability to their data center moves. They can effectively simulate each stage of the move to ensure that the business is properly supported throughout the life of the project. Just as important, this disciplined approach to data center and server relocation helps reduce costs, eliminate unexpected delays and optimize user acceptance.

About Shunra
Shunra Software is the market leader in providing WAN emulation solutions that enable organizations to deliver high performing network applications to their global end-users. Shunra has patent-pending technology for predicting and analyzing application performance over a WAN while testing in a pre-deployment lab environment. Using its industry-leading network emulation technology, Shunra provides IT with powerful insights into application performance. Shunra is committed to delivering innovative software and services that provide application performance intelligence, and continues to expand technology into deeper levels of application behavior analysis. Headquartered in Philadelphia, PA, global organizations leverage Shunra to reduce time-to-market for new services, guarantee application deployment success and ensure applications deliver a high quality of experience.

Ask Shunra About Our Data Center and Server Relocation Solutions Today!

Visit www.shunra.com and request to be contacted. Or contact Shunra directly at 1.877.474.8672 or 1.215.564.4046 (worldwide offices listed below)

Call your Local office TODAY to find out more!

For a complete list of our channel partners, please visit our website www.shunra.com

© 2009 Shunra Software Ltd. All rights reserved. Shunra is a registered trademark of Shunra Software.