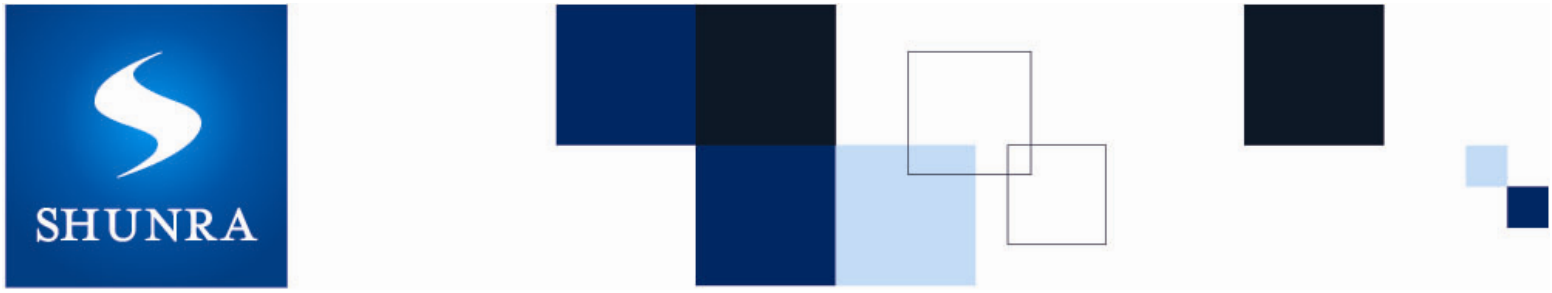
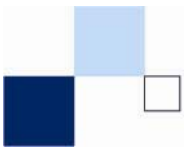


The Impact of Data Center and Server Relocation on Networked Application Performance

A Seven-Step Strategy for Project Success

A Shunra Software Best Practices White Paper



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: 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 started and completed such initiatives? According to Gartner, the primary factors inhibiting these potentially 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 who were previously local to the servers supporting their business applications, for example, may become remote users. And users who were already remote may wind up even further away from these servers. Few applications are "future-proofed" against these types of network changes.

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

- A 10X to 100X increase in application response time due to network latency and impairments
- A reduction in application scalability for supported users due to increased transaction time
- 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 or server consolidation**.** And 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 initially ignore the network-related impact of relocation/consolidation moves on application performance do so at their own peril. It's not enough to simply focus on systems issues such as the right-sizing of new servers or virtualization of storage resources. IT organizations must fully understand *in advance* how the relocation of servers on the network will impact application performance for end-users – and how to overcome any potential performance issues *before* the move or consolidation is performed.

Fortunately, the success of relocation/consolidation projects can be protected through best practices planning and management. This white paper offers a proven seven-step strategy for predicting and resolving the network-related application performance issues that arise when data centers or servers are moved or consolidated. By following this strategy, IT organizations can fulfill their service level commitments and prevent relocation/consolidation projects from turning into high-stakes gambles.

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 technology 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 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

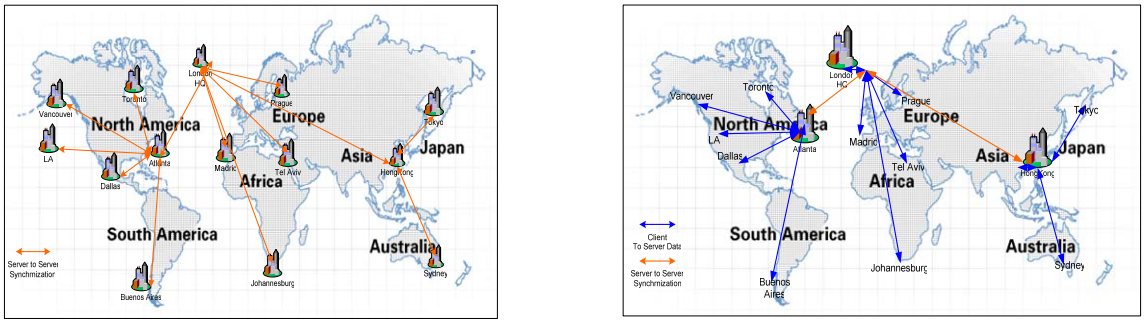


Figure 1: A typical regional consolidation

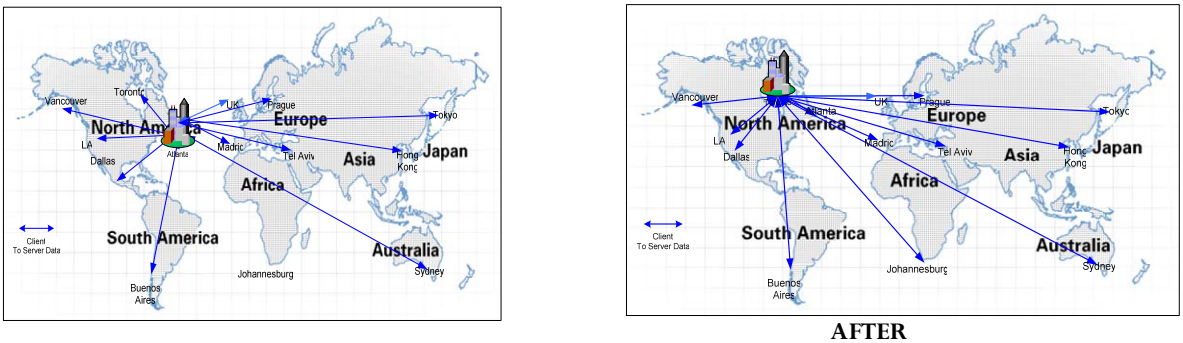


Figure 2: A typical relocation

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 broadly 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 like policy directories find that they have become remote users. Because they access those servers from a distance, they can potentially experience much longer transaction response times.

Similarly, 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.

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 with 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.

The Impact on Transaction Response Time

Network-related delays change dramatically during a data center relocation, since geographic distance contributes significantly to the 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 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 that have to be exchanged between the client and the server for each transaction (often referred to as application "turns")
- The specific transport protocol carrying the application (e.g. connection-oriented or connection-less)
- The configuration of the protocol on the server and client (TCP buffer sizes, 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 description of the relationship between network latency and application latency.

For a local user with a 1 millisecond delay between client and server, the transaction took three seconds. When just 50 milliseconds of network delay was introduced (representing a typical cross-country WAN connection), the performance of this transaction did not slow down by 50 milliseconds. Instead, this same log-in 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.

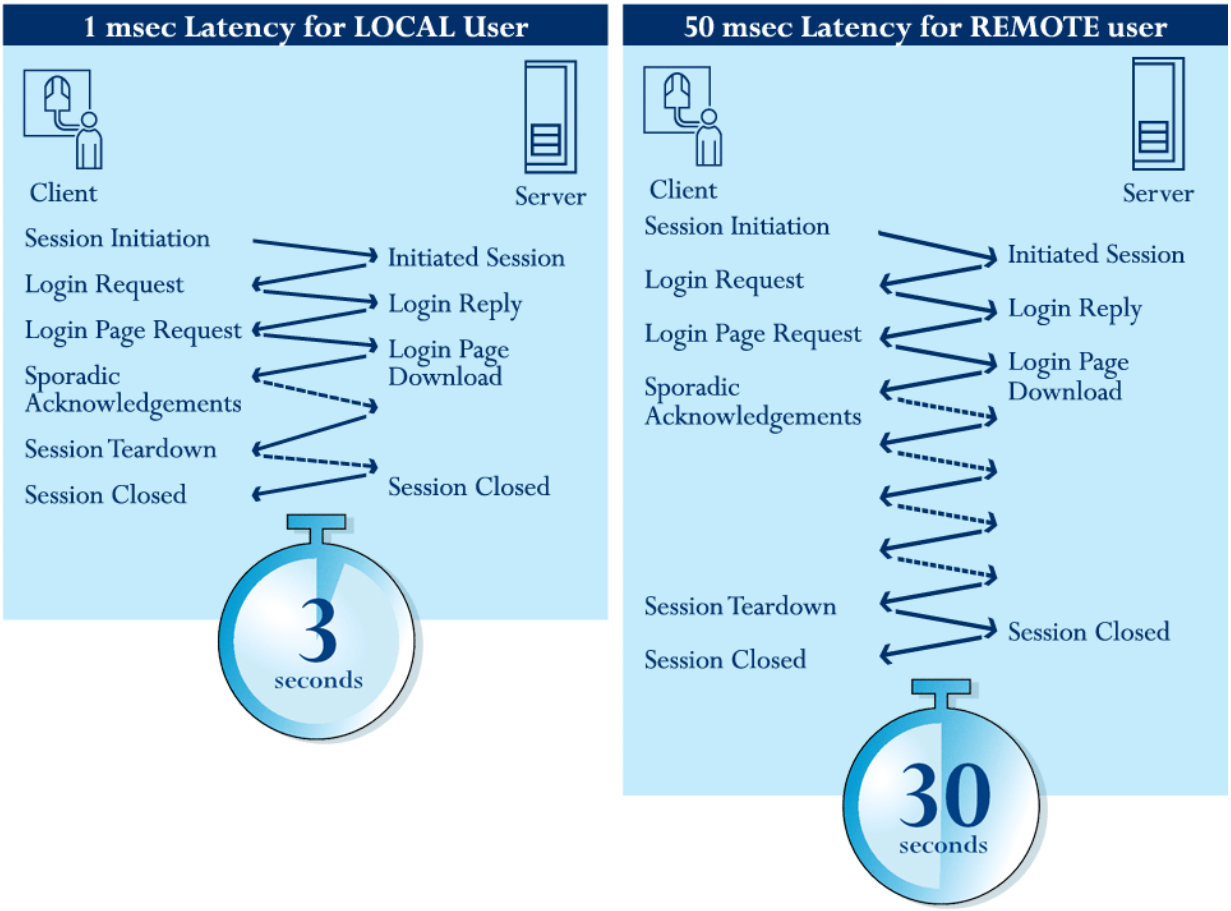


Figure 3: Transaction response time – local user vs. remote user
Note: The "login page download" arrows represent over 500 network packets broken into 180 application turns

This deterioration in performance can be clearly understood by analyzing the behavior of the log-in process. The application generates 180 application turns to complete this task. As the round-trip time for each application turn increases from 2 milliseconds on a LAN to 100 milliseconds across the WAN, the transaction response time increases 1000 percent.

Not all applications and business processes are 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 engineers to develop appropriate mitigation solutions.

Unfortunately, network impact is usually regarded as a network manager's problem, even though application architecture may be a major contributor in overcoming latency-related performance issues. After all, network managers can't change the speed of light, or make Tokyo closer to New York. So it doesn't make sense to lay the problem entirely on them. In fact, if application performance problems are more a result of the application design rather than network, 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, SOAP, and J2EE tend to greatly exacerbate this problem 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 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.

Any planned move therefore provides a great opportunity to re-examine "network-blind" development processes and introduce network-readiness best practices into the software development life cycle. Developers and QA engineers need to embrace these best practices to assess and improve the latency tolerances of their applications, so that appropriate performance can be cost-efficiently achieved and maintained for all critical business services.

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? The answer is rather simple. A server allocates a variety of resources to each concurrent client session. Local clients complete these transactions quickly because their application turns experience minimal network-related delay. Remote transactions, on the other hand, take much longer to complete because each application turn itself takes so much longer.

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. concurrent multi-threaded 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) thread performs the transaction including I/O and network calls to the database, 6) close OS thread and thread resources, 7) close SSL session, and 8) close TCP socket.

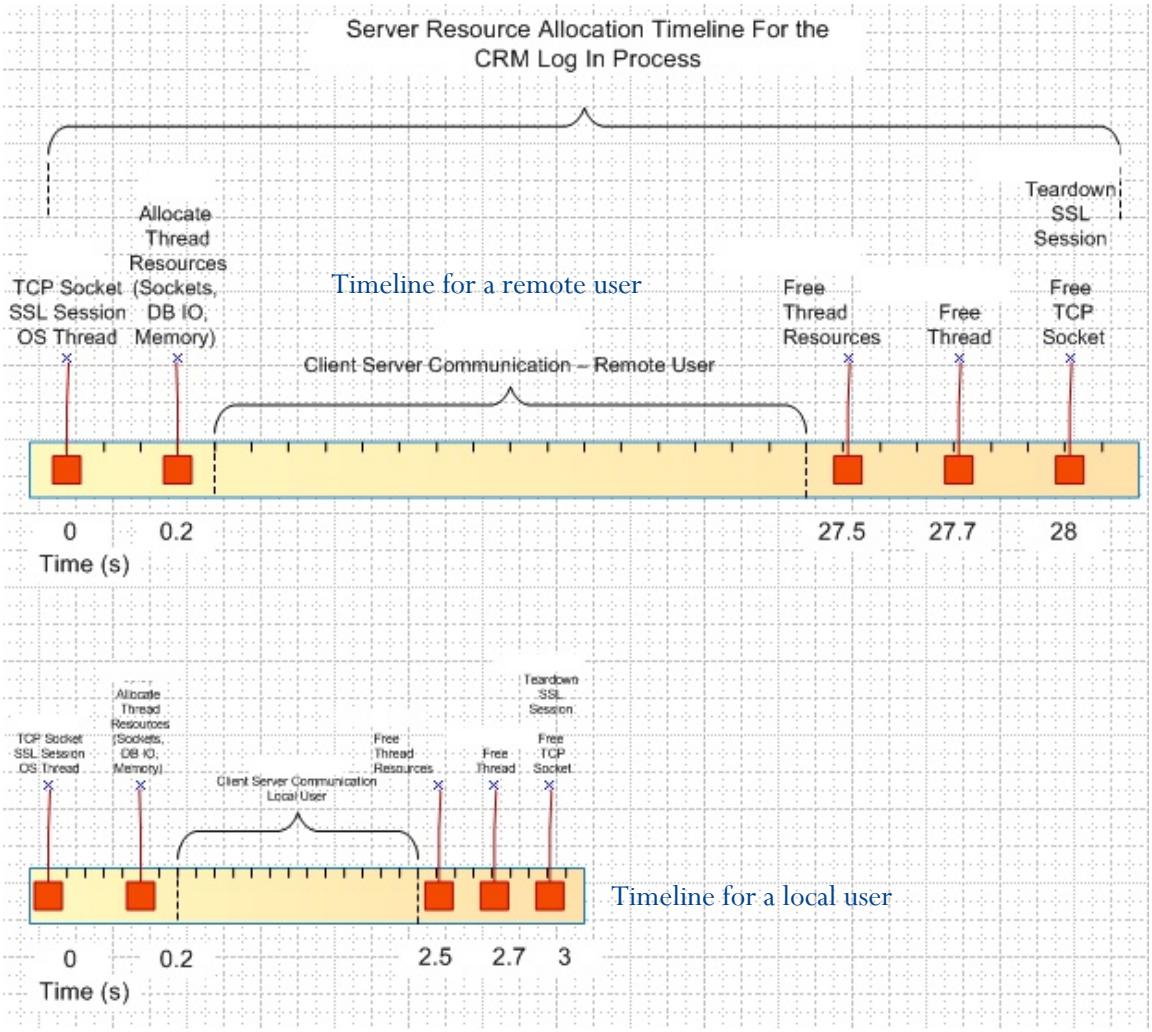


Figure 4. Server Resource Allocation Timeline

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, as *local users* are added to a server, performance and scalability can usually be improved with a CPU upgrade or by adding a server to the cluster. As *remote users* are added to a server, it typically makes more sense to upgrade RAM, tweak the OS scheduler, and/or off-load SSL processing (see Figure 5).

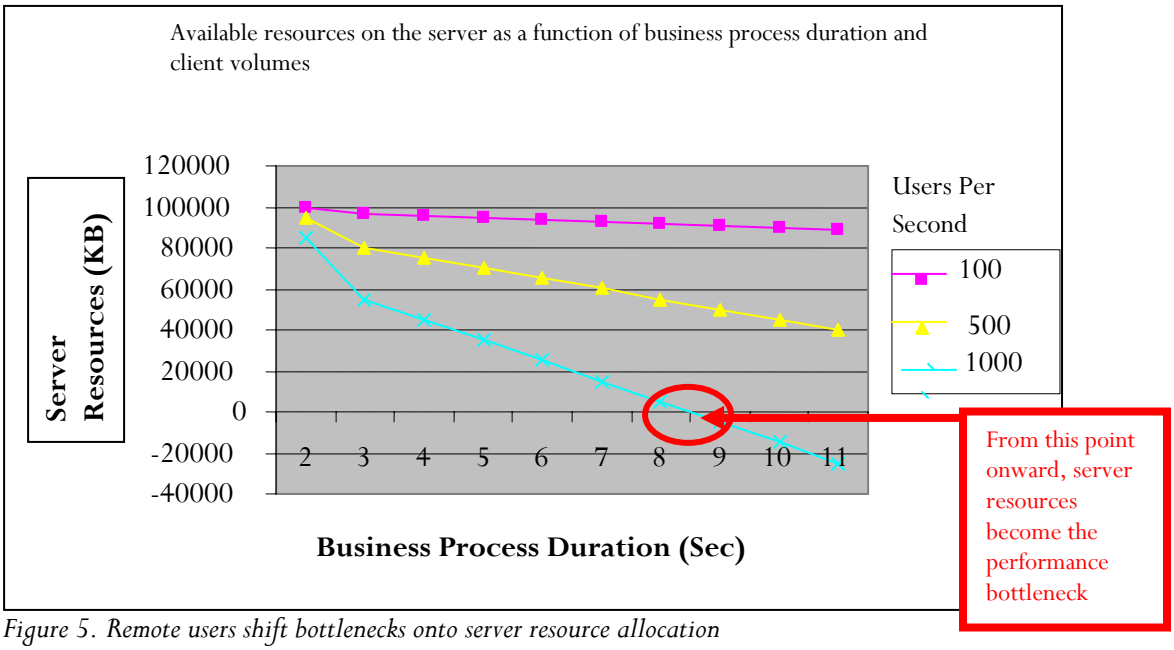


Figure 5. Remote users shift bottlenecks onto server resource allocation

Unfortunately, conventional QA and performance practices typically focus on back-end scalability, with little or no attention given to network latencies which affect users of the system. That is why IT organizations rarely address this phenomenon prior to going live with a data center move. 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 sys admins 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 network environment.

Maintaining Business Continuity During Data Center Moves

One of the most challenging aspects in a data center move is maintaining business continuity. 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?
- When should Active Directory servers be moved?
- Which servers will need to be replicated for the duration of the move?
- How will storage and business continuity architecture be affected during the move?

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 servers are almost never designed to accommodate significant latency between them.

Consider this typical scenario that can greatly impact application performance: 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. Client accesses the credit application server (10 application turns)
2. Credit application server authenticates client on the AD server (50 application turns)
3. Credit application server gets credit data from database server (200 application turns)
4. Client receives 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 different back-end components. Thus the distance between them during the move is likely to have a severe impact on application performance. See Figure 6.

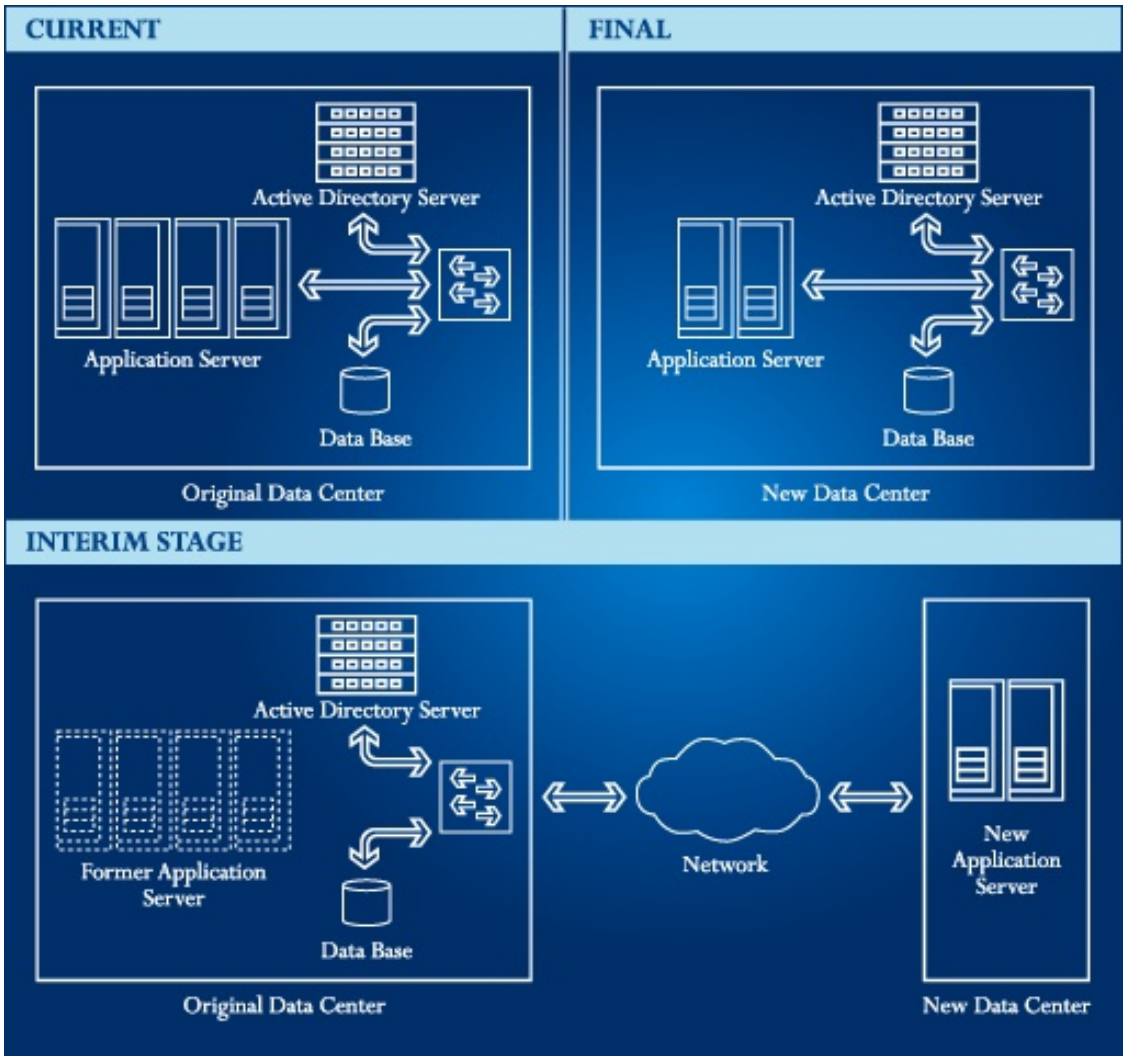


Figure 6. Latencies between servers in transition can disrupt business continuity

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 just hours when the servers were local to each other.

Data center relocation project managers must take these server dependencies into account during planning. Some of these insights may have been gained as a result of Y2K preparations. But this analysis must be updated prior to the move. 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 look up a contact 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 before a single server is relocated.

A Seven-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 has been split into 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 must also follow proven best practices for server consolidation/relocation. These best practices are encapsulated in the following seven-step plan for project success:

1. Build a virtual model of the pre- and post-relocation enterprise environment, as well as all planned transitional phases. All participants in the planning process, including business users, need concrete information about how server moves will impact application performance. This can't be done with paper and pencil. Instead, a complete working model of the environment must be created. This virtual model must simulate all relevant aspects of the projected environment – including the network distances involved, the number and distribution of end-users across the network, and the actual behaviors of the applications involved (i.e. application turns, server-to-server dependencies, etc.).

2. Establish an SLO baseline by measuring application performance before the move. Users' needs and expectations don't exist in a vacuum. Pre-move transaction response times provide essential context for determining reasonable SLOs for after the move.

3. Measure post-move application performance in the virtual environment. The only way to accurately predict the impact of server moves on application performance is to run those applications in a fully simulated post-move environment. This will provide the specific data on potential performance degradations essential for proper planning.

Location	ERP Before DC Relocation	ERP after DC Relocation
NY (15 users)	A (95%)	C (87%)
CA (20 users)	A (97%)	B (92%)
Sao Paolo (5 users)	A (9	
Tokyo (8 users)	A (9	
Brussels (20 users)	B (9	
Sydney (2 users)	B (9	
Application Grade	B (9	

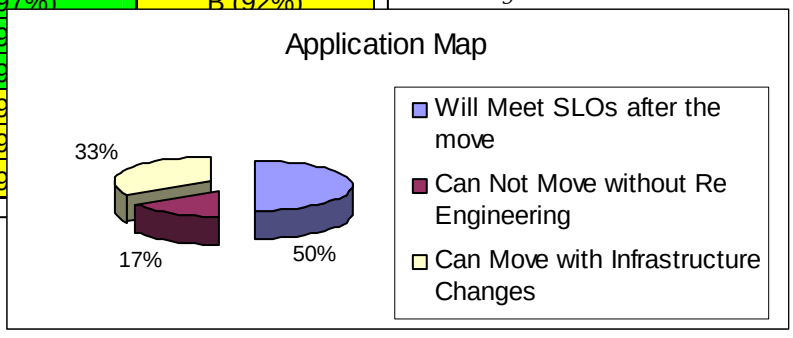


Figure 7. Reports like these are essential for successful planning and execution of data center moves.

4. Identify applications that need special performance tuning. Rather than wasting time, effort and money on beefing up enterprise infrastructure as a whole, it is far more efficient to focus on specific applications or network components that may be particularly problematic.

5. Analyze problems and validate potential fixes for failing applications. Before investing in and deploying a solution, it's important to make sure that it actually works. As noted earlier, IT organizations often add bandwidth or CPU power in the mistaken belief that such infrastructure enhancements will significantly improve application performance. Instead, such assumptions should be tested in the virtual environment before being executed in the production environment.

6. Assess dependencies between back-end servers to 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, a virtual environment is invaluable for accurately measuring the impact of these server-to-server latencies.

7. Manage user expectations and get buy-in commitments through hands-on acceptance. 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, the can give their informed consent to the relocation plan.

By following this simple seven-step 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 virtual modeling 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 provides solutions that empower organizations to address service level and performance concerns up front – before deployment. The Shunra Virtual Enterprise (Shunra VE) solution provides accurate, highly granular insight into how networked applications will function, perform and scale for remote end-users. It creates an exact replica of the production network environment, allowing IT professionals to safely develop, test and experiment with applications and infrastructure before rollout, and effectively plan for growth and change. With solutions tailored for networking and performance professionals, software developers, and quality assurance staff, Shunra VE facilitates collaboration across all IT disciplines – so IT organizations can quickly and more efficiently uncover and resolve problems before they impact the business. This results in more timely, higher quality and cost-efficient IT services, and the ability to "Deliver IT with Confidence".

Solutions for Any Enterprise

More than 1500 customers, including hundreds of *Fortune* 1000 and Global *Forbes* 2000 organizations, from financial institutions to manufacturing companies, retail, energy, media companies, as well as independent hardware and software vendors and telecommunications service providers, have gained measurable returns from Shunra's solutions. Among them are: 3M, Boeing, Cisco, Dow Chemical, EMC, FedEx, General Electric, General Motors, JPMorgan Chase, Kelly Services, Merrill Lynch, Motorola, Nestlé, Pitney Bowes, and Vodafone.

Corporate Information

Shunra's headquarters are located in New York City and Kfar Saba, Israel, with worldwide offices in Singapore, UK, The Netherlands and India. Shunra is also supported through a global network of channel partners.