



A Pragmatic Approach to Server and Data Center Consolidation



A Pragmatic Approach to Server and Data Center Consolidation

“A server consolidation project should be undertaken with specific and clear objectives in mind. Although cost reductions are a frequent goal, there are many equally important reasons to spend time and money consolidating systems. IT organizations should engage in a server consolidation project with three major goals: cost reduction, agility improvement and service-level improvement.”

— Gartner, Inc.,
“Key Issues for Data Center Servers, 2007”,
by Mike Chuba,
March 12, 2007

“Driven by cases of corporations not being able to get additional electric power for data center expansion, corporate IT shops are beginning to consider the energy costs and environmental impact of their computing equipment. Lack of power and lack of expansion space for urban data centers are precursors to lifetime energy costs exceeding the initial acquisition cost of data center equipment. And electricity prices are only heading higher; a federal cap-and-trade system for reducing carbon emissions could raise average electricity prices in the US by 5% to 35% depending on how aggressive a system is eventually adopted.”

— Forrester Research Inc.,
“The Greening of IT”,
by Christopher Mines
and Frank E. Gillett,
April 19, 2007

Executive Overview

In the late 1990s, we saw many organizations decentralizing their IT organization, spurred in part by the availability of inexpensive and powerful x86 servers that could meet the processing needs of departments and branch locations. Beginning in the early 2000s, we saw a reverse trend in which organizations began to move back to a more centralized IT structure in order to obtain cost savings, increased operational efficiencies and control, standardized processes and improved agility.

As a result, most enterprise data centers today are very large and getting larger. It is not uncommon for data centers to grow by hundreds of servers per year, gobbling up precious floor and rack space and creating shortfalls in power and cooling resources. To address the issue of server sprawl, as well as ease system management and lower data center costs, server and data center consolidation has rapidly emerged as the number one priority for IT managers. New technologies like blade servers and virtualization show great promise in helping data centers combat the problem of server sprawl.

The green computing movement is also driving consolidation initiatives. As organizations look to reduce the environmental impact and overall carbon footprint of their IT operations, consolidation is a logical first step. The cost savings and environmental benefits of consolidation and virtualization are closely aligned. By consolidating servers into more energy-efficient virtual machine hosts or blade servers, organizations can retire old, power-hungry hardware and optimize underutilized servers to achieve significant savings in space, power and cooling requirements.

However, successful consolidation initiatives require considerable upfront planning, time and effort, and a thorough understanding of the server workloads that need to be consolidated. A poorly planned or executed consolidation effort can result in complex configurations and increased IT overhead. In some cases, physical server sprawl is merely replaced with virtual machine sprawl. Virtual sprawl occurs when virtual machines are allowed to proliferate in an ad hoc fashion until eventually no one knows how many VMs there are across the enterprise, who owns them or their intended purpose.

Recognizing the challenges involved, data centers are beginning to view server consolidation in a different way. Whereas previously consolidation was seen as a one-off IT project, many data center managers now view consolidation as an ongoing IT strategy with long-term benefits for the organization. New technologies have emerged that support this view by making it easier and faster to monitor, move and consolidate workloads onto the systems where they will run most efficiently – whether physical or virtual. These available technologies are making continuous server consolidation and optimization a reality.

This white paper provides an overview of the factors driving consolidation initiatives and considers best practices for ensuring a successful consolidation initiative and a maximum return on investment. It also provides an introduction to workload profiling and portability technology which has helped thousands of organizations accelerate their consolidation initiatives and optimize their data centers through continuous server consolidation.



The Current State of the Data Center

Today's data center is a heterogeneous mix of different servers, operating systems, applications and data. Large, difficult to manage image archives and the increasing adoption of virtualization infrastructures and high-density blade server technologies further add to the complexity. It is extremely rare for an organization to be standardized on a single hardware platform. The growing size, density and intricacy of enterprise data centers often lead to extremely poor resource utilization, as well as administrative headaches. The outlook is that the data center will only continue to get larger, more heterogeneous and more complex unless IT management philosophies begin to change.

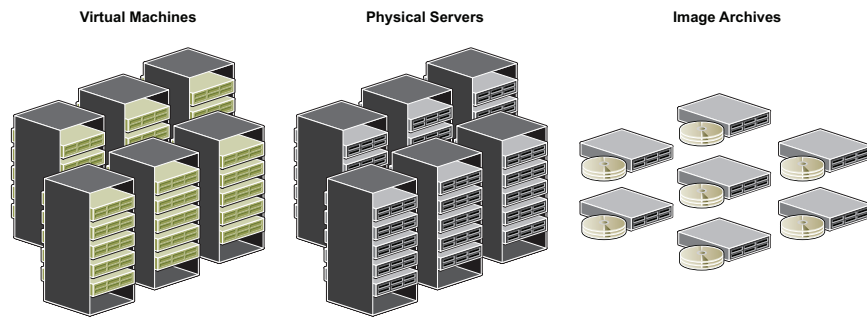


Diagram: The current data center landscape is a mix of virtual machines, physical servers and image archives.

The majority of servers in the typical data center are under-utilized, meaning that they consume more computing, power and cooling resources than can be justified by the workloads running on them. In most organizations, servers handle small or periodic workloads and run at only 10-20% of their capacity. Each server in the mix tends to run a single operating system instance and a single business application. This “one server, one app” model is a major contributor to low CPU utilization and server sprawl.

Under-utilized servers mean that resources are essentially being wasted while the hardware continues to consume costly computing, power and cooling assets – not to mention specialized IT staffing resources and maintenance cycles. Over-utilization of servers is less common, but may occur when workloads grow more rapidly than expected. Depending on the business-critical nature of the workload, an over-utilized server puts business continuity at risk.

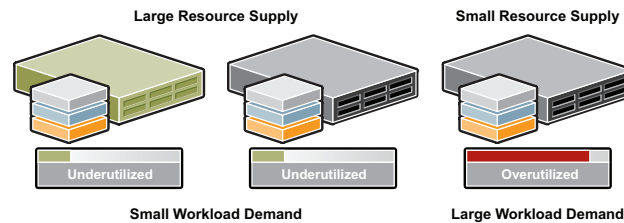


Diagram: Imbalances between workloads and resources in the data center.

Faced with the challenge of optimizing server utilization, combating server sprawl and greening IT, data center managers are increasingly turning to server and data center consolidation. Server consolidation allows organizations to reduce the total number of servers or data center sites required to support the business.



Types of Server Consolidation

There are different types of consolidation scenarios, and successful initiatives typically include a combination of the following:

Physical Consolidation

Physical consolidation involves migrating and/or combining workloads from multiple physical servers onto larger or newer physical hardware configurations such as blade servers. Blade server technology aids physical consolidations by allowing organizations to make the most of data center floor and rack space. Factors driving physical consolidation may include:

- The retirement of legacy or end-of-lease hardware.
- Data center relocations – i.e. moves to regions where power and cooling costs are significantly lower or where local government offers new business incentives.
- Post-merger or acquisition IT consolidations.

Physical consolidation requires the movement of workloads between hardware platforms via physical-to-physical (P2V) workload migration. Consider a multiplatform workload migration solution that supports different hardware configurations and server technologies to accommodate future changes in server infrastructure.

Virtual Consolidation

Virtual consolidation involves migrating workloads from physical servers to virtual hosts running virtualization infrastructures provided by VMware, Microsoft Virtual Server, Virtual Iron or XenSource. Virtualization allows more efficient sharing of physical resources to deliver higher CPU utilization rates. It also reduces the total number of servers needed to run the business, as multiple workloads can be combined and hosted on a single virtual machine host.

Server consolidation through virtualization requires the movement of physical workloads to virtual platforms via physical-to-virtual (P2V) workload migration. This task may be performed over a local network (LAN) or across greater distances using a WAN. In cases where bandwidth or lack of connectivity between sites is an issue, staged workload migrations may be required in which workloads are captured to image archives, redeployed on the virtual hosts at the remote site and then synchronized to capture any changes that occurred during the move.

Site Consolidation

Increasingly, we see organizations instituting enterprise-wide mandates to reduce the total number of data center sites and the size of the IT organization required to support them. Data center site consolidations typically involve a combination of physical and virtual consolidation and require both P2P and P2V workload migrations.

Previously, data center site consolidation efforts required the shipping of existing physical servers to new locations or manual rebuilding of new systems from scratch including the installation of operating systems, data, applications and drivers. New hardware is typically rebuilt from an image that is captured on disk and shipped to the new location.

Advanced workload migration tools reduce the manual effort and logistics involved in completing a site consolidation by streaming workloads over the network to both physical and virtual hosts at the consolidated data center site. These solutions also help to mitigate the business risks associated with large-scale data center site consolidations by allowing workloads to be moved to a test environment for thorough testing prior to a production move. Consider a solution that enables workload transfer without taking production servers offline to preserve business continuity during test and production workload migrations.

“Physical consolidation has grown to such an extent that most respondents have completed, or are in the process of, one such project.”

— Gartner, Inc.,
“Poll Confirms Internal Politics Is the Major Problem Area for Server Consolidation Rationalization”,
by John R. Phelps,
February 1, 2007

“It is reasonable to assume that virtualization will improve server use from an average of from 10% to 20% for x86 machines to at least 50% to 60% during the next three to five years. This should indicate a need for fewer servers.”

— Gartner, Inc.,
“Important Power, Cooling and Green IT Concerns”,
by Rakesh Kumar,
January 23, 2007



“Results show that although time to complete a server consolidation project peaks in the seven- to 18-month period, 38% of the projects take more than 18 months.”

— Gartner, Inc.,
“Poll Confirms Internal Politics Is the Major Problem Area for Server Consolidation Rationalization”,
by John R. Phelps,
February 1, 2007

Phases of a Successful Consolidation Initiative

Consolidation initiatives, whether to physical or virtual hosts, require careful planning and are typically completed in phases. This section outlines the different phases involved in assessing, planning and executing a successful consolidation initiative.

Phase 1 – Discover Server Inventory

The first step in planning a server or data center consolidation is to create a thorough inventory of all server assets. This inventory will allow you to make informed decisions about the workloads that may need to be migrated. For very large data centers or distributed enterprises, the discovery of server inventory can be a lengthy and onerous task. Consider investing in a software tool that allows you to remotely discover hardware and software assets, automatically gather detailed data for each server and organize servers by discovered properties. Solutions are available that allow you to complete these tasks without having to install agents or physically touch machines.

For physical-to-physical data center site consolidations or relocations, the asset inventory will tell you what data center equipment will need to be moved or retired. The high cost and logistical issues associated with physically transporting servers between locations typically makes it more economical to purchase entirely new hardware at the new site.

Collecting server inventory is the first step toward developing a detailed workload profile that will allow you to identify ideal candidates for consolidation to physical or virtual hosts.

Phase 2 – Collect Utilization Data to Develop a Workload Profile

To effectively plan a consolidation and properly size and provision the consolidated environment, data center managers must collect utilization data and monitor workloads and their changing resource demands over time.

There are many different kinds of workloads running in the data center, each with their own resource and availability requirements which may change over time. These changes may be cyclical, seasonal or completely random. For instance, financial reporting applications may place heavy demands on server resources at month’s end, or a web server may experience unpredictable traffic spikes.

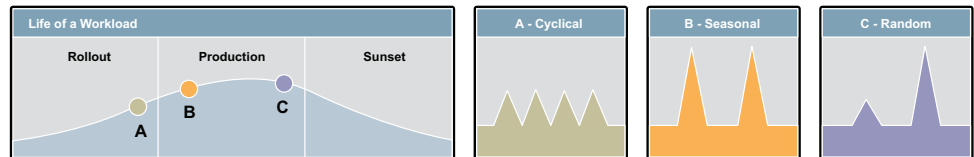


Diagram: Workloads and the demands they place on available server resources change over time.

By recording server utilization data over a period of days, weeks or months, you can develop a workload profile that provides a clear picture of server utilization trends and anomalies in CPU, disk, memory and network utilization rates. The workload profile contains the name and inventory of a server (applications and services) as well as the server’s current resource requirements based on real performance data. The profile may also describe the workload in terms of its purpose, departmental owner, level of business criticality, required recovery time and point objectives (RTO and RPO), and so on.



Consider collecting utilization data over a business-significant time period such as a financial quarter-end to ensure that you capture all peaks and valleys in the workload/resource utilization lifecycle.

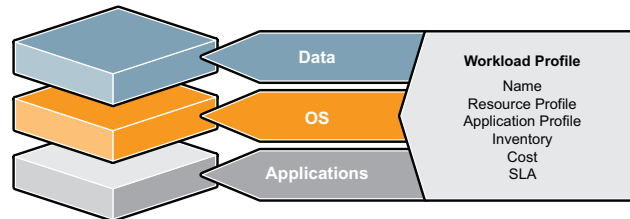


Diagram: The workload profile captures information about the encapsulated data, applications and operating systems residing on a physical or virtual host.

Workload profiling helps organizations accelerate the server consolidation process by automating the capacity planning phase of a consolidation project. Workload profiling also helps data center managers plan the layout and future growth of the consolidated infrastructure and ensure sufficient physical and virtual host capacity in the consolidated environment to accommodate current and future business requirements.

The workload profile follows the workload throughout its entire lifecycle, bringing consistency and predictability to the way organizations manage workloads. Administrators from different departments or sites can utilize the same workload profile to support consistent operational decision making.

Phase 3 – Analyze Workloads

The workload profile, which combines inventory and utilization metrics, provides a much deeper understanding of workloads, enabling data center managers to make more informed and intelligent consolidation decisions. When combined with analysis and forecasting capabilities, the workload profile enables increased visibility into data center operations and allows managers and architects to effectively plan for current and future consolidation initiatives. This awareness of workloads – their lifecycles and how they use resources over time – drives more sophisticated capacity planning. Consider investing in a data center analysis solution with a range of built-in or canned reports and the ability to export reports in a variety of formats (PDF, HTML, text, CSV, MHT, Excel, RTF and so on) to accommodate management and cost/benefit reporting.

When workload profiling data is captured and analyzed, organizations gain greater control over the data center and see marked improvements in the speed and quality of consolidation initiatives.

Phase 4 – Identify Consolidation Candidates

Armed with a detailed profile of workloads in the data center and analysis of real-time utilization data, you can now make informed, intelligent decisions about what workloads to consolidate. Use the server utilization data collected to generate a hardware utilization report that identifies workload and resource mismatches such as under-utilized or over-utilized servers. This report will allow you to identify candidate workloads that can be combined on a single physical host in the consolidated environment.

Phase 5 – Develop a Consolidation Plan

Now that you've discovered your server inventory, monitored utilization patterns over time, and analyzed the workloads running in your data center, it's time to actually build your consolidation plan. Rather than rely on best guesses, consider purchasing a data center planning solution that automates the development of a consolidation plan.

Software solutions are available that provide sophisticated scenario modeling, forecasting and planning capabilities for consolidation initiatives. These planning solutions allow you to more quickly and easily create scenarios for distributing workloads across physical servers and virtual hosts to maximize utilization and minimize resource contention. These solutions should include what-if modeling to determine different combinations of hardware and virtual hosts and proactively account for future growth by using forecasted 1 1 Tflize utilizatvaluativand



Phase 7 – Migrate Workloads

Once the consolidation plan has been developed and tested, you can use the workload migration solution to stream production workloads to any physical or virtual platform.

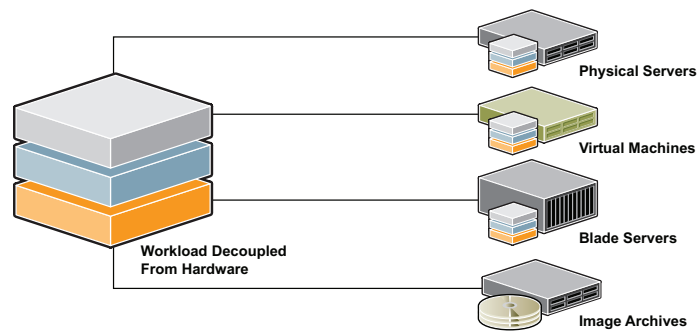


Diagram: Workload portability enables workloads to be migrated across different data center infrastructures.

“We needed virtual-to-physical migration for two reasons. V2P provides a means of troubleshooting applications and pinpointing whether issues are caused by the application itself or the virtual layer. This was necessary to get our business units to accept the adoption of virtualization. V2P also gives us a way to easily scale out of the virtual environment should the growth of an application require it. Some of our applications start out with a dozen users and rapidly grow to hundreds of users, requiring them to be moved to a more robust physical server environment.”

– Marco Spoel,
Project Manager,
IT Infrastructure,
Essent

Consider ease of use when evaluating migration solutions (does the solution have an intuitive drag-and-drop interface for migrating workloads from physical servers to virtual machines or blade servers?) and the solution’s ability to perform both local or remote migrations in either a staged or direct mode. The solution should also accommodate multiple concurrent migrations to ensure that workload movement activities can be completed in the most efficient and timely manner.

As discussed above, you should seek a solution with multiplatform support and broader capabilities than simply physical-to-virtual (P2V) workload migrations. Workloads may eventually need to be moved off of virtual infrastructures as their resource requirements grow (virtual-to-physical scale-outs) or may have to be de-virtualized to ensure the integrity of application maintenance and support agreements, many of which have failed to keep pace with virtualization.

The flexibility to move and rebalance workloads in any direction between physical and virtual hosts – physical-to-virtual, virtual-to-physical, physical-to-physical, in and out of imaging formats and so on – ensures optimal data center efficiency. It also enables organizations to better address challenges such as end-of-lease hardware migration and periodic necessities such as the de-virtualization of applications.

Phase 8 – Optimize Workloads

At this point, you’ve successfully planned and executed your consolidation. You’ve consolidated servers and leveraged virtualization and/or blade servers to shrink the physical footprint in the data center, reducing floor and rack-space, as well as power and cooling requirements that contribute to carbon emissions. In the old project-based view of consolidation, you’d be finished – bravo, well done.

In reality, workload optimization is a moving target. As we saw above, workloads and resource utilization change over time, necessitating periodic monitoring and rebalancing of workloads and resources to keep the data center running in an optimal state. Organizations that take a project-based view, rather than a strategic one, and that invest in simple P2V and imaging tools will continually play catch-up, struggling to keep the data center in balance.



“A recent innovation in the x86 environment is the ability to dynamically move running workloads to other servers to provide greater capacity for the workload. Virtual-machine relocation will become a default technology for most large x86 server infrastructures within five years, and it will dramatically change how servers are managed – disaggregating operating-system instances from physical servers. For businesses, the most important benefit will be a much faster response to changing scaling requirements. Virtual-machine relocation will also make capacity planning much easier for users, shifting from an application-specific function to an infrastructure-wide function.”

—Gartner, Inc.,
“Server Capacity on Demand
Spans many Capabilities”,
by John R. Phelps,
February 13, 2007

To find and maintain the balance between resource supply and workload demand and build optimal consolidation plans, data centers need to monitor and measure workloads over time and continuously readjust the balance by moving workloads back and forth between physical and virtual hosts. Data centers also need to size and resize the resources allocated to a given workload to accommodate its changing resource requirements, which may shrink or grow significantly throughout the workload lifecycle. This is where the concept of continuous server consolidation comes in.

Continuous Server Consolidation

After a consolidation project is completed, servers run at more optimal utilization levels. Over time, this point-in-time peak optimization begins to deteriorate. A sudden increase in number of customers may place a greater strain on invoice processing or fulfillment processing. A new service that is brought online may change the resource loads across the server environment, impacting response times.

As business conditions and application resource requirements change, some servers become over-utilized or applications outgrow their virtual hosts, risking vital business processes and service level agreements, while others become under-utilized and require re-consolidation.

Data center managers are starting to think about server consolidation strategically rather than tactically, moving beyond a “once-and-done” project-based view of server consolidation. The ease with which data centers can now stream workloads in any direction on-demand using workload migration technologies has made continuous server consolidation and ongoing optimization a viable, long-term strategy.

By decoupling software from the hardware layer, and allowing workloads to be streamed onto any platform in any direction, data centers are free to move workloads to where they can run most effectively. Workloads can be virtualized and de-virtualized at will to ensure the ongoing and continuous optimization of resources. Virtual-to-physical (V2P) migration enables V2P scale-outs when workloads need to be moved off of virtual infrastructures as their resource demands grow.

To achieve continuous server consolidation, an anywhere-to-anywhere workload profiling and portability solution must be deployed as a key component of the data center infrastructure. With such a solution in place, data center administrators can regularly monitor, profile and assess workloads to identify resource mismatches and forecast capacity issues so they can be proactively addressed. Administrators may choose to continuously monitor the data center environment, run the workload profiling solution at periodic intervals such as monthly or quarterly, or capture workload profiling data for a few weeks or months prior to any planned data center initiative.

By proactively rebalancing and moving workloads between physical and virtual infrastructures, business service levels can be sustained without interruption or degradation. Technologies that integrate workload profiling, planning and dynamic workload movement effectively automate the process of continuous server consolidation. Data centers gain agility, as well as bottom line cost saving and operational benefits such as lower overhead and total cost of ownership.



Summing Up

Once viewed as a “once-and-done” IT project, server and data center consolidation has become a long-term strategic imperative. Faced with rising energy costs, looming power and cooling shortfalls, and corporate mandates to go green, it is highly likely that data centers will begin to consolidate more and more of their physical infrastructures onto blades and virtual machines.

We are still in the early stages of the adoption of virtualization technology. As virtual technologies mature and comfort levels rise, virtual machines will move from low-risk deployments such as replicating test lab environments into more business-critical production environments. Already, virtualization is being used to support disaster recovery requirements without the need to buy and maintain costly duplicate hardware and software. As the use of virtual machines in production environments grows, and resource utilization rates increase due to consolidation and virtualization, there will be an ongoing need to balance workloads between physical servers and virtual hosts.

Organizations must begin to think differently about the data center. They need to embrace a view of continuous server consolidation where anywhere-to-anywhere workload profiling and migration technology is embedded in the core data center infrastructure to facilitate the free movement of workloads between any physical or virtual host – regardless of virtual infrastructure or hardware configuration. Investing wisely in solutions that support workload profiling and continuous server consolidation will ensure that the data center is equipped to address the current and future requirements of today’s rapidly growing businesses.



About PlateSpin Ltd.

PlateSpin provides the most advanced data center automation solutions designed to optimize the use of server resources across the enterprise. PlateSpin technology liberates software from hardware and streams server workloads over the network between any physical or virtual machine. Global 2000 companies use PlateSpin solutions to lower costs, improve service levels and solve today's most critical data center challenges including server consolidation, disaster recovery and hardware migration.

PlateSpin's patent-pending conversion and optimization solutions transform the enterprise data center by breaking the dependency between hardware infrastructure and server software. Organizations can monitor and manage server workloads to ensure the best fit between server resources and application demands. By enabling the free and flexible interchange of data, applications and operating systems with a simple drag and drop, PlateSpin brings greater flexibility and new efficiencies to the data center.



PlateSpin Ltd.
200 - 340 King Street East
Toronto, Ontario, M5A 1K8
Phone: 416 203 6565
Toll Free: 1 877 528 3774
www.platespin.com

© 2007 PlateSpin Ltd. All rights reserved. PlateSpin and the PlateSpin logo are registered trademarks and PowerConvert, PowerRecon, Server Sync, Workload Portability and PowerSDK are trademarks of PlateSpin Ltd. PlateSpin conversion and optimization technology and related products are patent pending. All other marks and names mentioned herein may be trademarks of their respective companies.