Jisto White Paper

Recapturing Stranded Server Resources

Bringing Increased Efficiency to Server Utilization

This white paper examines the phenomenon of underutilized physical and virtual servers and the impact of this on the companies that own and maintain them. It then explains how companies can make greater use of their underutilized server resources by using the Jisto Elastic Workload Manager.

Table of Contents

Executive Summary

Many companies invest heavily in server infrastructure. These investments may range from dozens of servers running on site or hundreds or thousands running in a cloud to 100,000 or more spread across multiple data centers around the world.

Given the large cost of ownership associated with this infrastructure, you would think that these companies are taking full advantage of their server infrastructure capacity. Unfortunately, this is not the case. Average server utilization is less than 20%.

What's more, this underutilization occurs whether these servers are running in on-premises data centers, private clouds, or public clouds. Significantly, this problem affects all infrastructure users.

Why such a low utilization rate? The primary reason is that almost every company today has critical applications that require 24/7 availability. A server or Virtual Machine (VM) running one of these applications may experience a peak utilization rate of 95% or higher for as little as a quarter hour each day. Even though these peaks may last no more than a few minutes, they are still placed in dedicated, if underutilized, servers or VMs to preserve the high Quality of Service (QoS) for those applications at all times. Data centers are clearly employing a risk averse strategy to prevent service disruption and assure resource availability.

While this arrangement guarantees that application demands are met at all times, the costs of server underutilization are non-trivial. Companies must invest in additional servers where lower priority applications can be run without jeopardizing the performance of the higher priority critical applications. The lower total ROI on servers reflects this low utilization rate.

There are, however, strategies and tools that make it possible for companies to improve the ROI of their servers. By layering lower priority applications on top of servers dedicated to critical applications, without compromising the latter's performance, companies can greatly increase the utilization of their infrastructure. This much denser server infrastructure will run many more applications, and reduce or eliminate the need for additional servers.

This approach is employed by the Jisto Elastic Workload Manager. It allows companies to improve utilization in an intelligently managed server infrastructure, without service disruption or changes to that infrastructure. It does this by automatically managing otherwise unused server resources—any combination of on-premises data centers, private clouds, or public clouds—while preserving the QoS for high-priority applications that are already running on these servers. In this way, the Jisto Elastic Workload Manager can improve server utilization by 200–300% providing an enormous computing capacity to run additional applications without increasing hardware costs.

Background

Any company that uses or manages a server infrastructure has certain objectives, such as:

- Guaranteeing that critical applications have adequate computing resources in periods of high or peak demand.
- Maximizing the ROI of infrastructure resources by minimizing costs or maximizing the revenue per server.
- Adopting new technologies and keeping software and other resources updated.
- Providing users with secure and reliable environments.

As often happens, an attempt to reach one of these goals may be at odds with efforts to reach another. For example, how can you assure that critical applications always have enough resources while at the same time minimizing server costs? How can you adopt new technologies and upgrade applications while simultaneously providing users with secure and reliable environments?

Conflict 1—Maximizing ROI and Meeting Peak Demand

The conflict between maximizing ROI and assuring that an application always has sufficient resources is independent of the type of data center and cloud strategies practiced, or whether physical servers or VMs are used. The following sections describe why and how this same conflict affects:

- Physical server infrastructures, whether legacy or new.
- Virtualized server infrastructures, whether legacy or new.
- Cloud resources, whether persistent or bursty.

Physical Server Infrastructures

It might be helpful to examine how companies currently attempt to achieve these objectives in order to understand why they are often in conflict. For a company managing a data center made up of physical servers, it may be important to guarantee that critical (e.g., customer-facing) applications be available and running at all times.

At the server level, this means that the data center must allocate enough resources for the successful processing of the applications at peak demands, which may last only a few minutes per day. Achieving this objective, however, results in underutilized server capacity, thereby reducing the ROI of the infrastructure for the rest of the time. In Figure 1, for example, each physical server has an Operating System (OS), and various applications running below full server capacity.

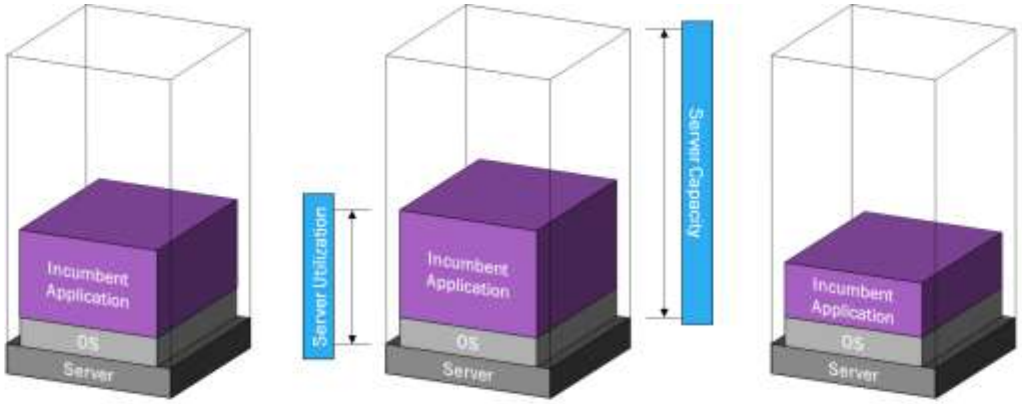

Figure 1—Server capacity and server utilization on physical servers

If a data center over-provisions most of its server resources in this way, then the only way to increase the number of applications that it supports is to increase the number of servers. This translates to greater fixed costs for hardware and physical space, and greater variable costs for items such as maintenance and electricity.

There are, of course, different strategies that data centers have used to increase server utilization. They can, for example, run two or more applications on the same server at the same time, thereby lowering server resource use per application. With an approach like this, however, adequate system resources may not be available at peak application demand. Inevitable resource conflicts, such as memory saturation, a situation when two applications simultaneously experience a memory demand spike, could create cascading application and server crashes.

Static schedulers such as IBM Platform LSF/Symphony and Univa Grid Engine have been used to deploy applications to servers more efficiently, but have several drawbacks. These tools typically schedule one application at a time, either per server or core or statically-provisioned “slot.” These types of schedulers require the server or core or slot allocation to be completely idle before they can schedule the next application.

Many schedulers also require a non-trivial amount of time (i.e., 30–60 seconds) to make the scheduling decision, leaving multiple servers idle while waiting for the next application. While these schedulers define which applications should run and when, they unfortunately reserve a huge amount of server capacity for peak load events that occur very rarely. This means that there is a tremendous amount of processing power that still lies idle most of the time.

Virtualized Server Infrastructures

The challenges that physical servers face as described above also apply to virtualized data centers that run hypervisors such as VMware vSphere, Microsoft Hyper-V, and Red Hat KVM on their servers. Since several VMs are provisioned on a single physical server, and since a VM does not adjust its resources in response to changing application resource demands, similar overprovisioning occurs within a VM. This results, of course, in the underutilization of the underlying physical server.

Figure 2 illustrates this underutilization and excess capacity in a virtualized server environment. Each physical server runs a hypervisor, VMs with an OS, and critical Incumbent Applications running in the VMs.

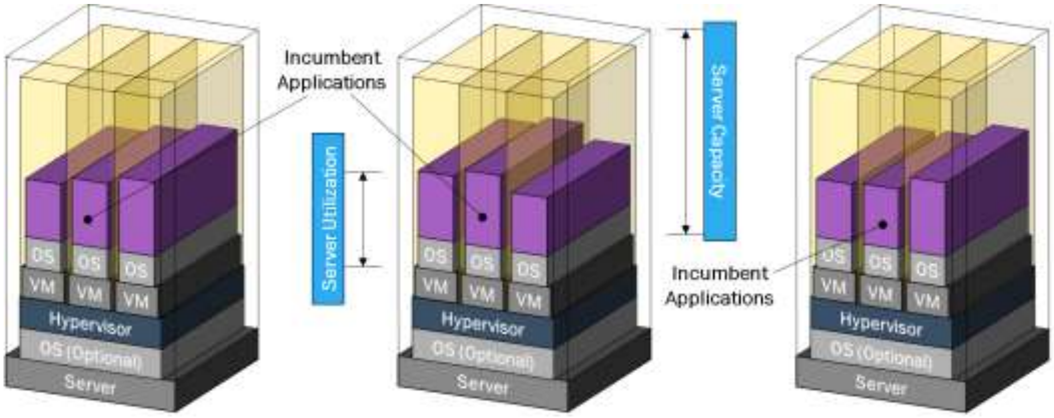

Figure 2—Server capacity and server utilization on virtual machines

In the lifetime of an Incumbent Application running on a VM, there will of course be fluctuations in utilization over time. The average VM/server utilization typically remains well below the VM/server capacity. On average, less than 2 out of 8 cores are used on an 8-core VM.

VM Load Balancers and Server Utilization

Unfortunately, VM load balancers do not significantly improve the total utilization of the virtualized infrastructure. In a perfect VM load balancing strategy, even though utilization may be evenly spread out across the infrastructure, each individual VM will still be underutilized. This translates directly to underutilization of the underlying physical resource and of the infrastructure as a whole.

Cloud Resources

Many companies outsource their server and storage needs to private and public cloud providers such as Amazon Web Services, Microsoft Azure, and Google Cloud Platform. The goals for these companies still include:

- Provisioning enough resources to guarantee resource availability even in unusually-high-demand (peak) load scenarios.

- Making sure not to provision too many resources, to keep costs as low as possible and to maximize their ROI.

For these companies, these goals are in conflict, just as they are for companies with physical and virtualized data center infrastructures. These companies still need to purchase enough servers and server time for peak application demands, creating underutilization for the majority of the time that is outside of these peak scenarios. This means that they over-provision resources, resulting in significant resource underutilization during the average lifespan of these servers.

Server underutilization has a negative impact on the budgets of both the companies that use the cloud resources, as well as the public and private cloud infrastructure providers. Typically, user companies will have used 20% or less of the server resources that they purchase. The cloud provider, on the other hand, will have lost the opportunity to sell the excess server capacity to other companies.

Conflict 2—Assuring Security While Adopting New Technology or Upgrading Software

The goals of providing a secure, dependable software environment and of adopting new technology or upgrading software are in conflict. A new technology or a software upgrade each introduces risks for the successful execution of the application as well as any legacy applications. An upgrade of a server's system stack, for example, creates risks for an existing application.

A strategy that would assure that a new application run successfully in an existing environment and that an existing application be unaffected by new technologies and server upgrades would significantly diminish the conflict between these goals.

Solution

Thankfully, companies have been aware of the challenges facing data center and cloud resource management. There are technologies and tools that can be used to diminish the conflicts between the competing objectives described earlier. These involve the use of:

- *Dynamic schedulers* that automatically select the best resources on which to deploy the applications at any time, on the basis of current utilization of each server.
- *Containers* for efficient application deployment.
- *Elastic workload managers* that automatically reallocate server resources in response to changes in application demand.

Jisto brings together these technologies and enables data center and cloud resources of various kinds to:

- Reduce the number of servers used for application processing.
- Save money spent on hardware, thereby increasing ROI.
- Use existing legacy infrastructure for new applications.
- Upgrade software painlessly.

The following sections explain how the Jisto solution can help different kinds of data centers to meet these goals.

Eliminating Underutilization

In the best of all worlds, it would be possible to run as many applications as possible on any given server at all times, eliminating underutilization and using the full capacity of each server. As we have seen, this seldom occurs in the real world since so many servers are reserved for applications that require a very large part of their server capacity during short periods of high demand. To eliminate the resulting underutilization, the best strategy would be to run non-critical applications on these servers, while assuring that high-priority applications have the resources they need when they need them.

The use of *containers* is one of the techniques that help make this strategy possible. A container is a form of OS-level virtualization that wraps an application with the ports, share drives and other elements that it needs to run. Containers running on the same server share the same underlying OS kernel, giving containers a tiny footprint compared with VMs, where each need a separate OS and kernel. Containers exhibit resource isolation qualities similar to those found in VMs, although they execute much more quickly and are more portable.

While containers offer an effective way to isolate applications from the rest of a server, a way to manage and deploy them is still necessary. Docker is an open source product that provides tools for container creation and management and for the standardization of best container practices as part of an open source ecosystem.

By leveraging Docker and containers in this way, Jisto lets companies:

- Enable additional applications to be intelligently and automatically deployed across servers that are running mission-critical and time-sensitive applications.
- Mitigate potential resource conflicts that occur, ensuring that mission-critical applications in need of resources will obtain them regardless of other applications that are running on the servers.
- Elastically capture and release resources in real time, maximizing server utilization and ROI.
- Preserve the existing software environment—no change is required to the OS, Linux kernel, or any application already running on the server.

The following sections describe how Jisto works with Docker and containers in the following:

- Physical server infrastructures, whether legacy or new.
- Virtualized server infrastructures, whether legacy or new.
- Hybrid clouds, whether persistent or bursty.

Physical Server Infrastructures

Figure 3 shows a physical server environment that runs Jisto.

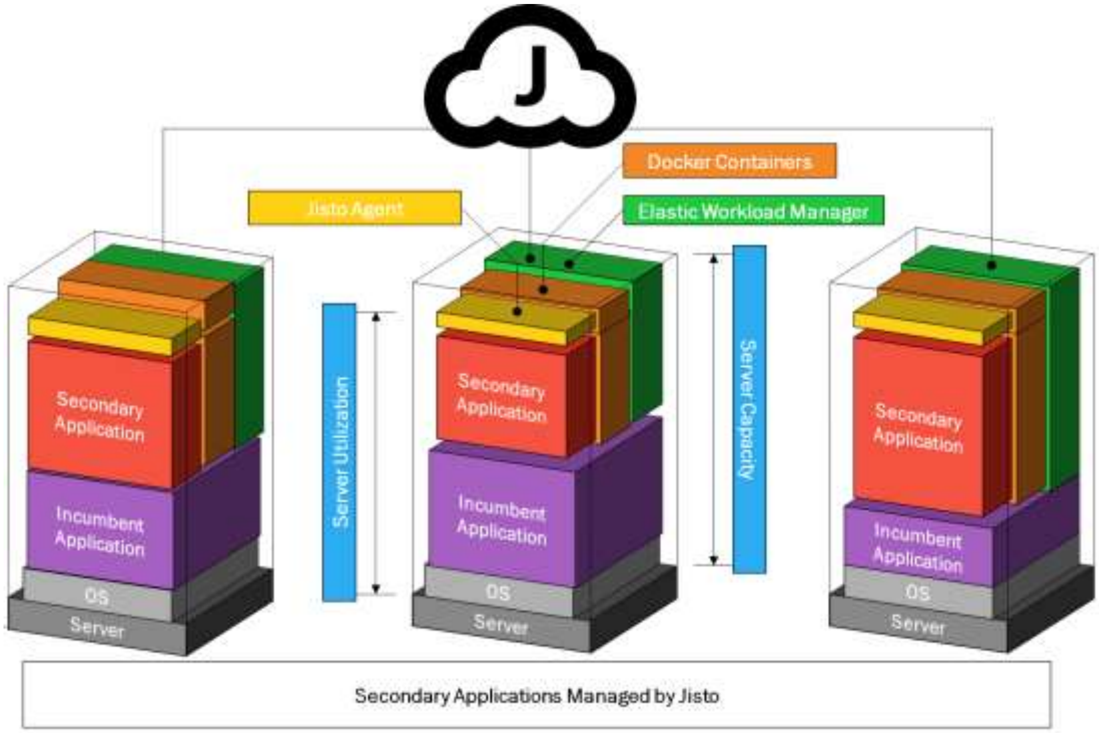

Figure 3—Physical server environment managed by Jisto

In the above diagram, note that each server runs the same OS and the same high-priority applications Incumbent Applications as those shown in Figure 1. The Elastic Workload Manager manages only the lower-priority Secondary Applications by using data from the Jisto Agent to capture resources from or release resources to the high-priority applications in response to fluctuations in their demand. The Incumbent Applications always receive resources when they need them.

Safely running more applications on the same resources at the same time with Jisto maximizes the ROI without making changes to underlying hardware or OS.

Virtualized Server Infrastructures

Figure 4 shows an example of a virtualized server environment that leverages Docker containers and Jisto:

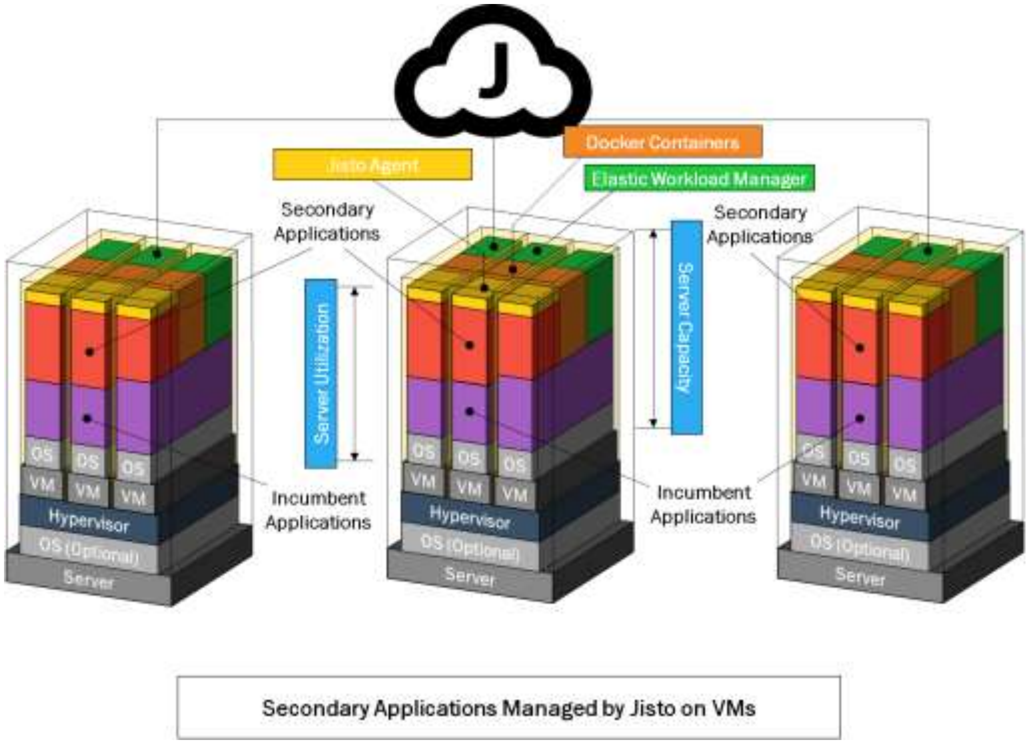

Figure 4—Virtual server environment running Jisto

In the above illustration, each server runs the same hypervisor, VMs, OS and various high-priority Incumbent Applications as shown in Figure 2, with the addition of the Jisto Agent and the Elastic Workload Manager. This manager has the same effect in a virtual server as it does in a physical server, where the Elastic Workload Manager elastically stretches and shrinks the resources available to the lower-priority Secondary Applications, ensuring that there is no resource contention for the high-priority Incumbent Applications. Since more applications are able to run seamlessly in VMs than previously, server utilization increases dramatically, thereby using existing servers more efficiently.

Hybrid Clouds

Since Jisto supports every type of server environment—on-premises physical servers, on-premises virtual servers, and private and public cloud servers—it can be easily deployed to any combination of these resources as a hybrid cloud. Figure 5 shows how Jisto can maximize server utilization in a hybrid cloud (or environment).

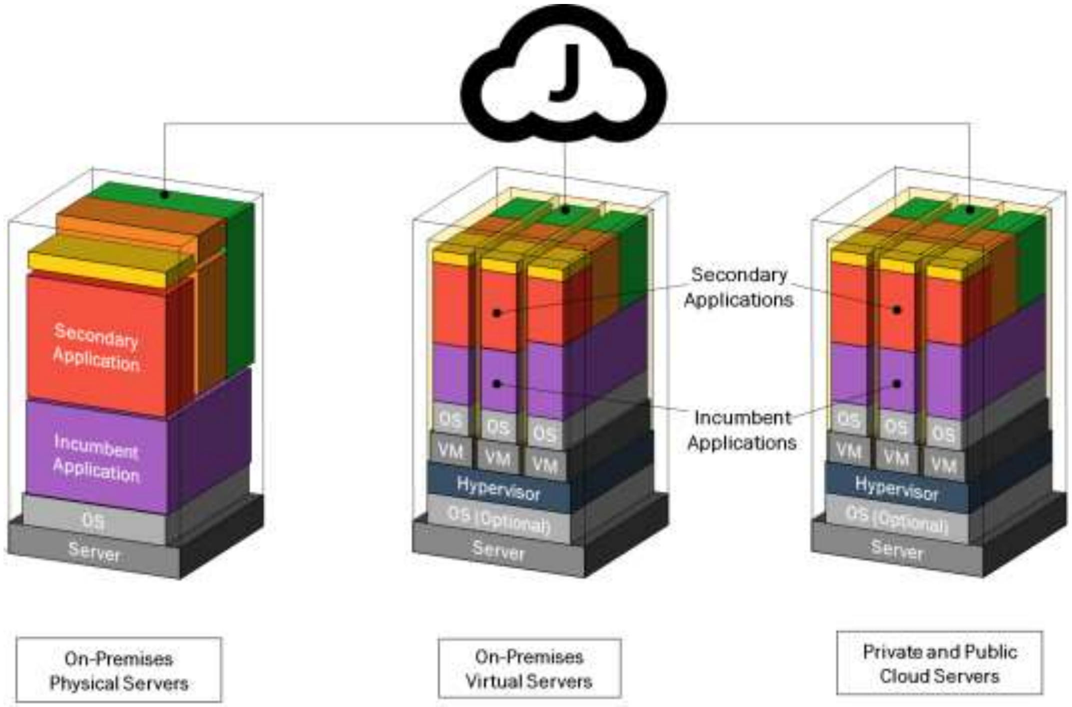

Figure 5—Hybrid cloud environment managed by Jisto

Conclusion

Many companies have invested in server infrastructure that is not optimized for the highest utilization. This results in lower ROI for their infrastructure investments and higher than necessary operating costs.

The best way to counter these problems is to increase server utilization by running non-critical applications inside servers without compromising resource availability for critical applications. Jisto makes this possible by ensuring that critical applications are always able to access the resources they need regardless of other applications running on the server. In addition, the Jisto Elastic Workload Manager assures that multiple applications can be intelligently and automatically deployed across servers that may even already be running other applications.

This strategy enables performance and utilization gains that apply to all types of server infrastructure—on-premises physical servers and VMs, private clouds, and public clouds. Jisto can also be used to deploy and manage applications in a hybrid infrastructure consisting of different infrastructures.

In addition, by using containers, Jisto can deploy applications onto existing infrastructure without requiring other changes. These applications, whether automatically containerized by Jisto or pulled from an existing container repository, are executed in isolation from the underlying OS and its other applications. For this reason, Jisto enables new applications to be run in these server environments without affecting the existing reliability and security.

Interested in Learning More

Please contact us at contact@jisto.com for more information, or visit www.jisto.com to request a demo.