Advantages of a Dynamic Infrastructure: A Closer Look at Private Cloud TCO

Scott A Bain
Innes Read
John J Thomas
Fehmina Merchant
IBM SWG Competitive Project Office

April, 2009

Table of Contents

Executive Summary

Many companies are finding their need for greater business agility being frustrated by an increasingly costly and rigid IT infrastructure. The culprits are many. Maintenance of the current environment accounts for over 70% of the IT budget, leaving less than 30% available for new projects. Annual operational costs (power, cooling, and management) of distributed systems and networking exceed their acquisition cost by 2-3X and continue to climb. Utilization rates of these commodity servers hover around 5% on average, leading to excess capacity going to waste. Time to provision new servers can be as long as six months, hampering lines-of-business efforts to quickly respond to competitive threats or new opportunities. As a result, LOB units are beginning to go outside the datacenter to public cloud providers like Amazon in hopes of lowering their costs and improving their responsiveness. To avoid disintermediation, IT needs to re-invent the datacenter by moving towards a more dynamic infrastructure. One that takes out cost through the use of virtualization and consolidation to improve utilization levels with a commensurate reduction in power consumption. One that embraces a private cloud model that dynamically provisions IT services in minutes/hours rather than months (and at lower cost) via self-service portals.

This paper examines the Total Cost of Ownership (TCO) for a dynamic infrastructure built around private cloud services and compares it to public cloud alternatives as well as conventional one-application-per-distributed server models. The results show that private cloud implementations built around larger virtualized x86 and System z servers can be up to 80% less expensive than public cloud options over a five year period and almost 90% less than a distributed stand-alone server approach.

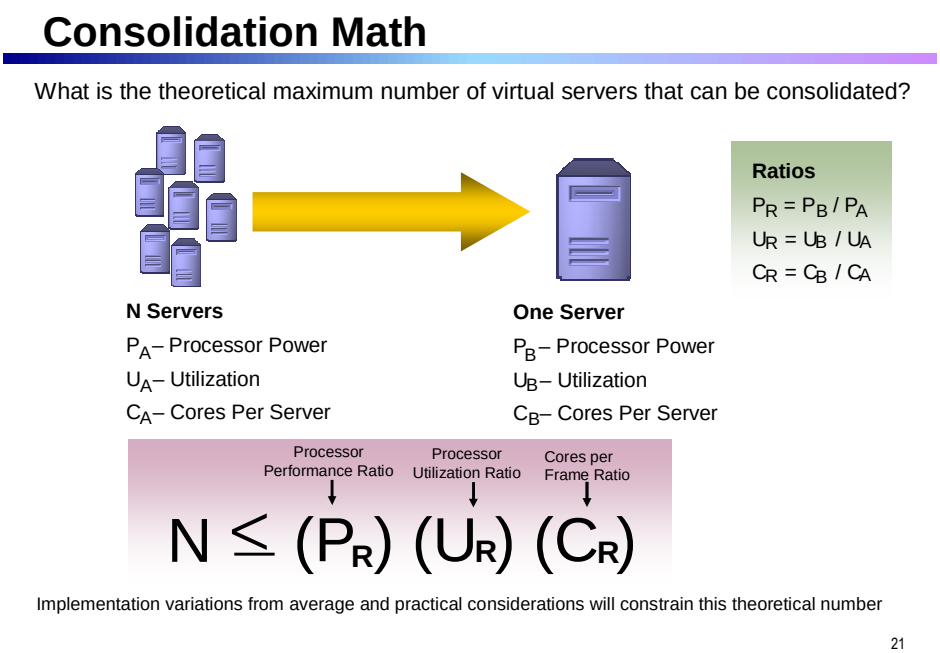
Take Cost Out Through Virtualization and Consolidation

A recent IBM internal study of its nearly 4000 distributed servers showed annual operational costs attributed to each server to be over $34,000, with almost 90% due to software maintenance and systems administration. It stands to reason that reducing the number of physical servers to fewer, larger, more capable machines can serve to greatly reduce these costs. Indeed, the virtues of virtualization and consolidation to accomplish this have been well-publicized. What has proven to be more elusive, however, is the quantification of these benefits. How many workloads can actually be consolidated onto a given platform while maintaining acceptable service level agreements? Which platform gives you the greatest economy of scale, producing the lowest cost per virtual machine image/workload?

One approach to answering these questions and estimating the TCO of a private cloud environment, as well as compare it to other alternatives, is as follows:

1. Determine the expected consolidation ratio for a given workload type (e.g. Windows, Linux, etc.)
2. Estimate the annual cost to operate the virtualized servers (over 3 or 5 years)
3. Compare to stand-alone provisioning or public cloud services

It is possible to predict the theoretical maximum numbers of virtualized servers that can be consolidated using a mathematical formula dubbed “Consolidation Math”:



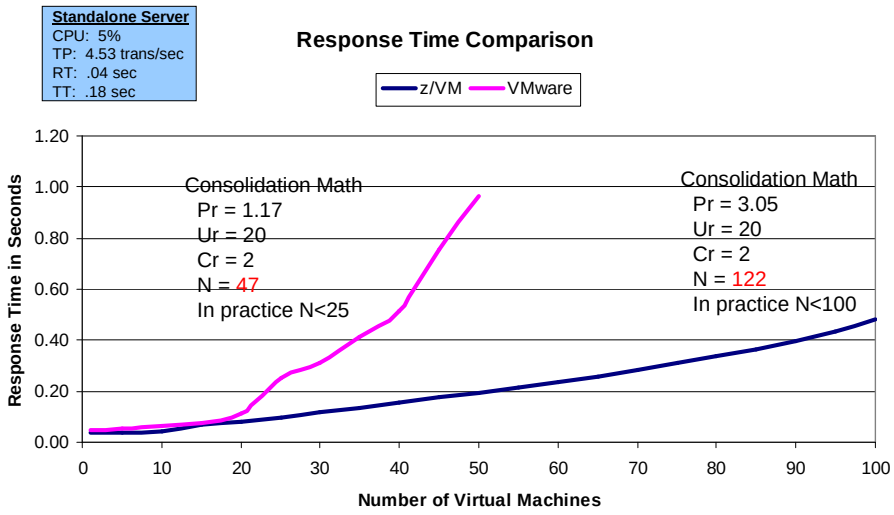
This formula shows that the more virtual servers you can consolidate, the lower your cost. Thus, the goal is to maximize ‘N’ in the consolidation math equation. P_R is the Performance Ratio and represents the impact of moving from an older processor technology to a newer one with a higher clock speed.

This number typically ranges between 1.0 and 3.0. U_R is the Utilization Ratio and speaks to the ability of a given consolidation platform to achieve to achieve higher utilization rates over the older or non-consolidated workload. This ratio is typically between 10-20, with larger values reflective of moving from older x86 systems (~5% average utilization) to newer mainframes (80-100% utilization levels). C_R is the Core Ratio and reflects the advantages of using new, multi-core processors over previous generations with fewer cores.

While the consolidation math formula sets the upper bound, other factors occur in actual productive use that serves to reduce these consolidation ratios. The efficiency of the platform hypervisor, for example, can adversely impact the results. Example metrics include CPU utilization, memory demand and over-commit capabilities, and I/O demand. Other factors include variability in workload demand and the application of Service Level Agreements (SLA).

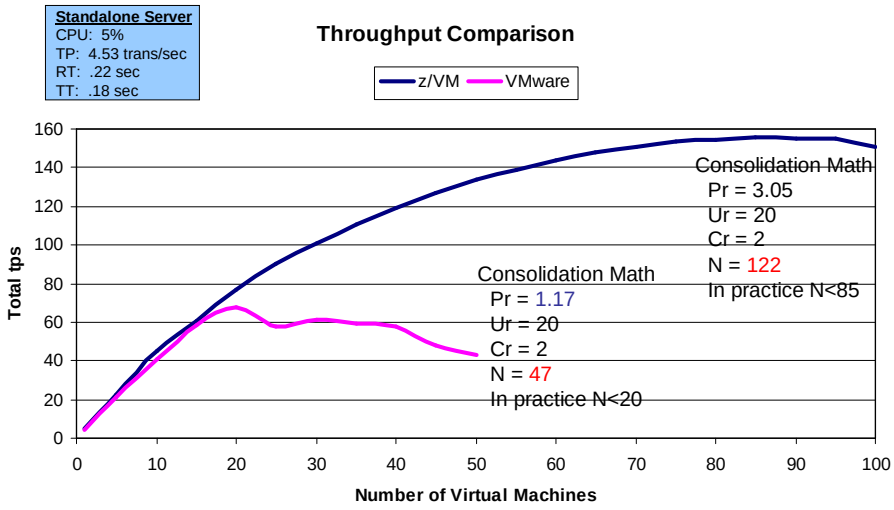
To gauge the impact of these factors in determining actual consolidation ratios, the CPO conducted a series of performance benchmarks. A sample banking application was built using IBM WAS and DB2 running on Linux and run on an older 4-way (single core each) IBM x366 server using 3.66GHz Intel processors and 1GB of memory. Average CPU utilization was 5%, throughput was 4.5 transactions per sec, and average response time was 40 milliseconds. A VM image of that workload was then created and placed on an 8-core IBM x3950 server (four 3.5GHz dual-core processors) with 64GB of total physical memory and running VMware as a hypervisor. Multiple running instances of this VM image were added incrementally to the server until it could no longer handle any additional throughput. CPU utilization, throughput, and response time metrics were captured throughout. This same test was then applied to a single frame IBM z10 EC machine (8 IFL cores @ 4.4 GHz) running z/VM as a hypervisor. The results are shown in the charts below:

Response Time Comparison



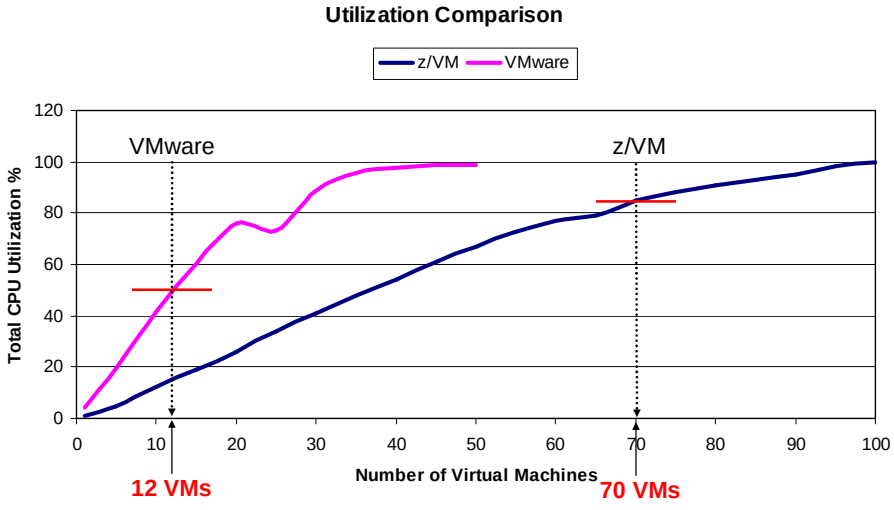
25

Throughput Comparison



26

Service Level Agreement Defines Achievable Consolidation Ratio Versus Utilization



29

For the VMware case, the maximum number of VM images that could be supported with acceptable response time was <25, almost half of the theoretical maximum (47 images). The z/VM case, on the other hand, showed a maximum of <100 whereas the theoretical limit was 122 images – only a 20% degradation. Using maximum throughput as the metric reduced the observed maximums to <20 and

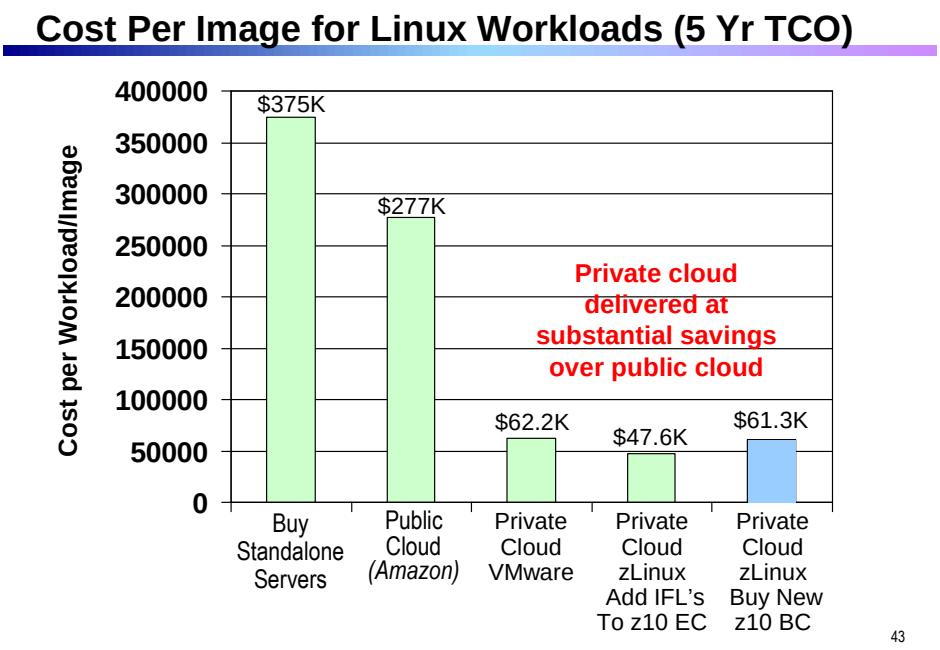
<85 VM images, respectively. Some customers use CPU utilization as their SLA metric. One customer in particular found that limiting x86 servers running VMware to 50% CPU utilization and 85% on an IBM System z platform with z/VM produced acceptable results. Applying these SLA cutoffs to our benchmark data yielded a maximum of 12 images for VMware and 70 images for z/VM.

The next step in the analysis called for estimating the TCO over 5 years for running 100 Linux images using four different platforms to see which one delivered the lowest cost per image or workload:

1. Buy stand-alone x86 servers (running one image/workload on each)
2. Rent Amazon EC2 instances (running one image/workload on each)
3. Buy large x86 servers and provision virtual servers using VMware (private cloud)
4. Upgrade an existing z10 EC machine and provision virtual servers using z/VM (private cloud)

TCO components included hardware, software, maintenance, facilities (power/cooling), and administration and assumed 24x7 operation. Administrative costs were derived from IBM RACE and other internal studies.

The results of this TCO comparison appear below:



As expected, doing "business-as-usual" and buying stand-alone (or rack) servers is the most expensive option ($375,000 per image/workload). Upgrading an existing z10 EC system and provisioning your own virtual servers is the least expensive alternative ($47,600 or 87% less), while doing the same provisioning on larger x86 servers and VMware resulted in a $62,200 charge per image (83% less). What may be surprising to some people is the fact that the public cloud option using Amazon EC2 instances was so expensive ($277,000 per image). This is due to several factors. First, using Amazon EC2 instances

continuously for 24x7 operation results in higher runtime platform charges (almost 3X) than acquiring physical servers and provisioning workloads in-house to achieve higher utilization rates. In addition, customers must purchase software on a per EC2 instance basis rather than being able to take advantage of on-premise multi-core systems that can support multiple images on a given hardware platform. Finally, although individual server management is eliminated with the public cloud, there is still significant labor costs involved in the administration of each running application instance. A breakdown of each of the cost components for each scenario appears below:

Detailed Cost Breakdown for Linux Workloads (5 Yr TCO)

	Buy Another Server	Rent a Virtual Server	Provision Your Own (VMware)	Provision Your Own (z/Linux)
Runtime Platform	100 IBM x3250 with 4 cores each	100 Amazon Extra Large EC2 instances	Five IBM x3950 with 8 cores each	7 IFLs added to existing IBM z10 EC
Hardware Costs ▪ Server ▪ Storage ▪ Networking	$5,000,000	$2,880,000	$1,080,000	$2,030,000
Software Costs ▪ OS (Linux) ▪ Hypervisor (VMware, z/Linux) ▪ App Server (IBM WAS) ▪ Database (IBM DB2)	$21,490,000	$20,840,000	$2,380,000	$1,520,000
Facilities and Admin ▪ Power ▪ Floor space ▪ Maintenance ▪ Systems admin	$11,020,000	$4,020,000 *(admin only)*	$2,760,000	$1,210,000
Total Cost	$37,510,000	$27,740,000	$6,220,000	$4,760,000
Number of Workloads/Images Supported	100	100	100	100
Total Cost per Image	$375,100	$277,400	$62,200	$47,600

40

Perhaps more impressive is the cost per image of a new IBM z10 Business Class (BC) system for those who don't currently utilize IBM System z hardware. At $61,300 per image/workload, this system is slightly less expensive than even x86-based servers running VMware.

All told, customers electing to provision their own virtual servers in a private cloud setting will find it far less expensive than either conventional stand-alone servers or public cloud alternatives.

Service Management for Better Visibility, Control, and Automation

The cost savings associated with a private cloud implementation also extend to the service management arena. With fewer servers to manage in a virtualized and consolidated environment, less software and fewer administrators are required, resulting in lower overall cost.

Aside from basic monitoring of individual IT resources, effective service management requires the following solution components:

- end-to-end monitoring of applications, where parts (WAS, DB2) often run on different servers
- holistic view of the state of business-level services (e.g. order entry process, credit check process, etc.) that often rely on shared infrastructure spread out across multiple physical machines
- a database to store and track changes to hardware, software, and networking configurations for better control
- a service desk for administrators to use to handle service requests and the automated processes required to resolve them

IBM Tivoli provides solutions for each of these requirements: 1) Composite Application Manager (ITCAM) addresses the need for end-to-end application monitoring, 2) Business Service Manager (TBSM) to provide better visibility into business-level services and their status, 3) Change and Configuration Management Database (CCMDB) to capture and manage hardware, software, and networking configurations, and 4) Service Request Manager (TSRM) for a service desk.

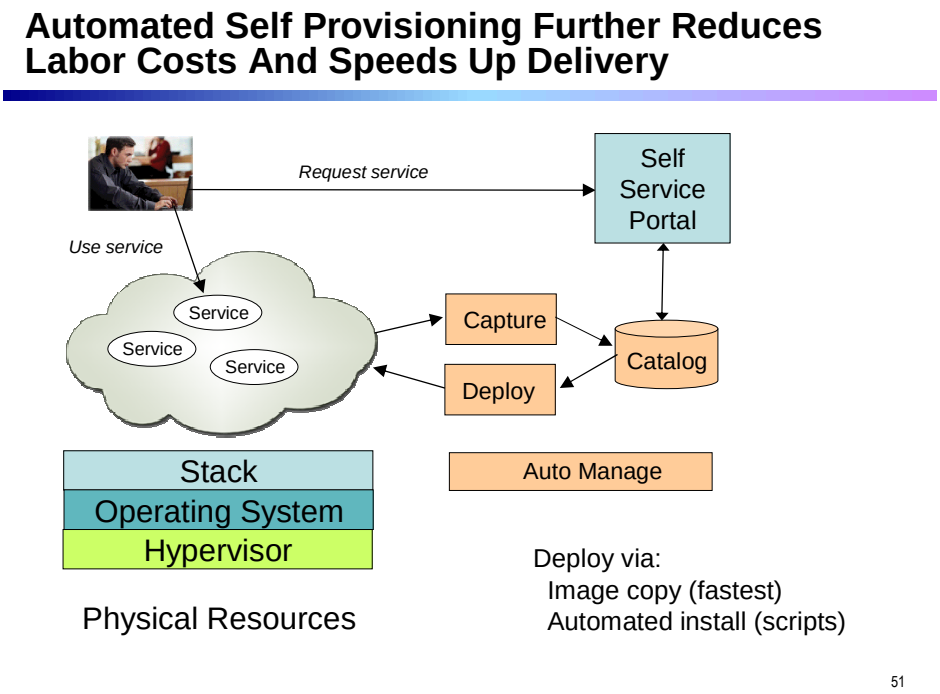
The costs associated with using this software to manage 100 Linux workloads running either on distributed servers or consolidated platforms appears below:

	Distributed IBM x3250	Consolidated IBM x3950	Consolidated z10 EC
Servers Required	100	5	1 *(7 IFLs)*
Administrator Licenses Required	11 *(3 full/ 8 partial)*	3 *(1 full/ 2 partial)*	2 *(1 full/ 1 partial)*
ITCAM Costs *(5 years)*	$2,124,000	$212,400	$89,208
TBSM Costs *(5 years)*	$235,240	$137,320	$126,520
CCMDB Costs *(5 years)*	$271,000	$158,110	$155,934
TSRM Costs *(5 years)*	$114,050	$29,750	$17,350
Total Cost	$2,744,290	$537,580	$389,012

The table shows it is between 80-90% more cost-effective to manage a consolidated environment over a conventional distributed environment.

Request-Driven Provisioning Through Self-Service Portals

While virtualization, consolidation, and service management provide the basic underpinnings of a dynamic infrastructure, it must be accompanied by a self-service portal that enables users to request IT services on demand and have the request fulfilled in minutes/hours versus days/weeks/months.



In this model, services are initially defined/created and stored in a service catalog. Requesters can then browse the catalog to find and select the desired service. After submitting the request, it gets routed for approval and then fulfilled by the underlying infrastructure. The software needed as part of the overall service is typically deployed in one of two ways: image copy (the fastest) or via automated install using scripts. When the service is no longer needed, the affected resources are freed up so that they can be claimed by other subsequent requests. In order for all of this to work seamlessly and transparently to the user, there needs to be automated management software that undergirds each step in the process.

IBM recently introduced Tivoli Service Automation Manager (TSAM) to manage this cloud services lifecycle and deliver request-driven provisioning for a private cloud environment. It leverages TSRM to provide a self-service UI for users to search against the catalog and select the desired service. It also utilizes Tivoli Provisioning Manager (TPM) to provision hardware and software resources according to best practices to satisfy the service request.

Summary

Escalating business requirements will continue to drive companies toward datacenter transformation. This includes pursuing ways to take costs out of their existing infrastructure, such as the use of virtualization and consolidation that reduce the number of physical servers needed and lowers energy consumption. Adding self-service, automated provisioning of IT services on top of this foundation to create a dynamic infrastructure allows IT to respond more quickly to the needs of the business. It also allows them to do so at lower cost than other alternatives (including public clouds) and avoid the threat of disintermediation. While multiple target platforms are available from which to deliver this private cloud environment, customers who have already made investments in IBM's System z platform will find it to be an attractive and cost-effective option.

IBM Corporation
Software Group
Route 100
Somers, NY10589
USA

Produced in the United States

April 2009