x86 Virtualization: Cure for Server Sprawl?

Data centers are suffering from the effects of server sprawl. The biggest culprit is x86 servers, which multiply like rabbits in the nooks and crannies of IT shops, filling up rack after rack of floor space. Relief may be in sight as virtualization technology, coupled with enterprise class scalable x86 hardware gives customers the ability to radically reduce server counts, reduce management labor, while maintaining or actually increasing service levels. This report examines the benefits provided by x86 server virtualization.

Intel servers have claimed a very large chunk of the overall server market over the past several years, accounting for more than 85% of unit shipments and about half of total revenue. This growth is fueled by rapid technical advances in x86 server technology with the most obvious improvements found in processor performance and system scalability. Modern systems can support up to 32 dual-core Intel processors in a single (SMP) system image – large enough to handle almost any enterprise workload. Reliability, availability and manageability have also improved at a rapid clip – our latest survey of enterprise data center personnel shows **73%** agreeing that *"Properly managed x86 servers are just as available/reliable as UNIX servers*." This result confirms just how far x86 server technology has advanced in the last few years.

Not surprisingly, enterprise customers have embraced 64-bit x86-based servers in a big way – using them both for new workloads and in some cases to supplant aging UNIX and proprietary midrange systems no longer in production. In our recent survey (referenced above), 68% of respondents indicated they were re-hosting UNIX workloads on Intel-based systems and 81% of respondents reported they were using x86-based systems for business critical workloads. As the sales figures above prove, data centers now house huge numbers of x86 servers with more arriving every day…and that's where the problems begin.

Virtualization technology has recently re-emerged as an old solution to the new problem. The benefits to running x86 servers in a virtual environment are many, but IT managers are principally driven by the need to reduce IT labor costs, and the savings are most pronounced when companies choose larger SMP servers to host the virtualized applications. These four- and eight-processor platforms deliver more processing, memory and I/O resource to their hosted applications helping simplify the guesswork associated with new deployments. The more advanced software solutions can even pool the reserve capacity required for peak load conditions across server resources delivering economies of scale that lead to more virtual machines (and productive workloads) per processor license.

Gabriel
CONSULTING GROUP

Sophisticated virtualization technology was, until recently, the exclusive province of the mainframe and high-end UNIX systems. However, feature rich virtualization packages are now available for use with x86 systems. Moreover, these products have been around long enough to earn a track record for stability and maturity. At this point the stage seems set for a rapid proliferation of x86 virtualization technology with the biggest rewards being reaped by those deploying the fewest new systems.

Pain Points & Viscous Circles

While x86 servers have radically changed in the last several years, one thing that hasn't is how customers deploy them. The vast majority of x86 systems host only a single workload and, according to our observations and industry data, average utilization rates for x86 servers hover around 10% or less. These systems are not purposely underutilized; it's just that each system has to be configured to handle the predicted peak load plus some room for growth. Unfortunately, many of today's Internet-exposed software workloads are 'spiky', with high usage peaks of short duration followed by long valleys of low usage. In most data centers, it's easy to find particular servers being pounded sitting alongside others that are simply idling at 1 or 2% utilization.

Compounding this problem is the fact a single new application often brings multiple servers into the data center. New applications generally require a system (or systems) for production along with additional discrete systems for test, development, and perhaps training or failover. All of these systems generally have very low utilization rates over time and are seldom rededicated to new tasks. But, regardless of utilization (or non-utilization), all of these systems have to be managed and maintained. They need to be: connected to the network, individually backed up, patched to ensure they don't fail and cause problems in the infrastructure, monitored for capacity planning purposes, factored into every change in the overall infrastructure, and secured and updated in a timely manner. All labor-intensive operations, but all are necessary to keep the data center stable and secure.

Almost every organization has tens, hundreds, or even thousands of these relatively small servers. The pain arising from this server sprawl is acute, and data centers are experiencing symptoms that weren't on the map five or six years ago - running out of floor space, running out of electrical capacity, with heat output outstripping their ability to provide enough cooling.

While facilities problems might cause some pain, this discomfort pales in comparison to the toll that server sprawl is taking on personnel and budgets. The labor required to manage modern server 'hairballs', as one IT manager calls them, is becoming more and more onerous. Many organizations, despite their best efforts, are finding themselves devoting most or all of their labor efforts to simply keeping the existing infrastructure operating – putting out fires, rather than spending their time performing their 'real jobs', such as capacity planning, infrastructure planning, etc. It becomes a vicious circle, or, perhaps more accurately, a viscous circle; data centers get stuck in the mud of reacting to day-to-day problems and can't make progress towards taking proactive steps that will help them avoid these issues.

This inefficiency drops down to the bottom line as organizations require more and more IT personnel to manage ever increasing number of primarily x86 based servers. While the cost of hardware (and even software) is steadily dropping over time, the cost of IT labor is increasing at close to 10% annually – which means costs will double every 7-8 years. This rate of increase is insupportable over time and will force changes in IT management.

Vendors aren't completely deaf to the problems above, hardware and software vendors have made some progress towards addressing at least the symptoms of the server sprawl problem. Advances in administration and management automation help reduce the number of hours it takes to provision and manage systems. New form factors, such as blade servers, can significantly reduce the floor space, power and cooling requirements. These innovations are certainly useful, but do not do much to solve the key cause of x86 server sprawl: the x86 server usage model.

x86 Virtualization – Changing the x86 Server Usage Model

The best way to attack the server sprawl problem and to regain control over runaway IT expansion is to implement a virtualized x86 server infrastructure. Virtualization software is giving IT shops around the world the ability to reduce server sprawl, increase efficiency, and reduce total x86 server operating costs. But before we begin a discussion of what virtualization can do for the enterprise, let's define some terms since virtualization can mean different things to different people. Ideally, virtual servers should:

- ✓ Host large numbers of horizontally or vertically scaling applications safely, a failure in a single application does not harm any other application on the system.
- ✓ Provide high overall hardware and O/S availability
- ✓ Reconfigure system resources on the fly – without taking a reboot or workload interruption anywhere in the system. There should never be a need to over-configure a system with a "free pool" or "floater boards" in order to have the resource flexibility necessary to support combined workloads.
- ✓ Manage system resources to apportion capacity to applications according to business need or policy – without operator intervention. SLA requirements can be factored into the management scheme and system configuration (cpu, memory, etc.) is automatically adjusted to meet the SLA.
- ✓ Offer sophisticated chargeback mechanisms that track application and system usage at the user level and allow for costs to be traced back to business units or other constituencies.
- ✓ Hardware sparing to allow failed components (cpus, memory, etc.) to be configured out of the system without workload interruption or reboot.

All of the requirements above, when applied to x86 servers, can be boiled down to a single industry need: the ability to apply a mainframe usage model to x86 systems – while preserving x86 performance and economics. Although current x86 virtualization software and system combinations don't quite satisfy all of the above points, the technology does adequately address the first (and most important) requirement, meaning that these systems can host multiple workloads safely, not allowing a single misbehaving application to harm the entire system. Both VMware, with their ESX Server product, and SWsoft, with their Virtuozzo product, have demonstrated this ability and this ideal is the design point for every other virtualization product.

For their part, x86 servers now have the size and scale needed to easily run multiple applications along with the requisite workload management software. To us, it doesn't make much sense to virtualize small (less than 4-processor) servers – there simply isn't enough capacity to handle the overhead associated with the management software, operating systems and application code. Larger servers (8-processors or more) have the capacity to handle more workloads and the headroom to handle unexpected spikes in demand without missing a beat. Smaller servers may be able to host a few light weight applications, but may not have the capacity to handle more complex business workloads or high peak loads without having to shortchange other system users. Another advantage to larger servers is that they offer more configuration options and generally have more sophisticated availability and manageability features.

x86 Virtualization – A gift that keeps on giving

The benefits from virtualization can be profound. GCG has done extensive research into virtualization and server consolidation dynamics in the UNIX market–still a bit ahead of x86 capabilities. Over the past two years, UNIX customers have realized savings of 30-40% vs. non-consolidated systems. These are pretty good results, but we believe there is even more opportunity for efficiency gains and cost savings in the x86 server world.

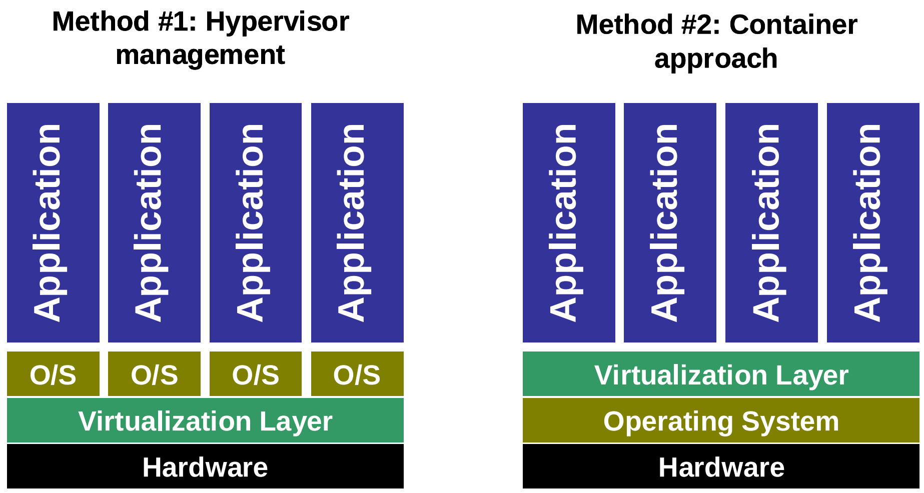
For example, one customer in Australia recently confided to us that they had managed to consolidate the workloads of 380 small test and development servers onto ***three*** larger x86 servers using VMware. Of course, not all of the 380 server images are in use at the same time, each individual environment can be launched as needed using scripts – provisioning a new server image takes just minutes and can be accomplished by non-central IT personnel. Not surprisingly, the company has reaped considerable savings in time and support dollars and is aggressively pursuing other x86 server consolidation opportunities.

This isn't an isolated case; we recently heard from SWsoft that enterprises using their Virtuozzo product on the service provider side (leading ISPs and other web services providers) routinely host hundreds of virtual environments on single mid-range x86 servers. In the last several months, they have begun selling their products into corporate accounts and report that these customers are running anywhere from 20 to 60 virtual environments hosting business applications.

The consolidation ratios above (380 to 3, 20 or 60 to 1) are considerably higher than what are generally seen in UNIX consolidation initiatives and leads us to believe that there is huge opportunity for savings in x86 consolidation. The ability to remove such large numbers of underutilized servers from the data center floor, removing the need to power, cool, manage and monitor those systems can provide a provide a huge return on a relatively modest investment.

x86 Virtualization – Two Flavors

There are two distinct approaches to building virtualization management each with their respective advantages and disadvantages. Simply put, the major differences lie in where the virtualization layer of software is located on the system software stack. The pictures below illustrate the differences:



Hypervisor Management: is a virtualization approach where the virtualization software sits directly on the hardware and controls all aspects of resource use by the hosted operating systems. The operating systems cannot see any of the other o/s instances or use any resource that has not been made available to them by the VM software. This is a traditional mainframe virtualization model with the best x86 example being VMware's ESX server product. Hypervisors bring a higher degree of safety to operations as they provide better isolation between o/s instances and thus are more able to ensure that a single application crash will not impact any other o/s instance. The main disadvantage here concerns a significant operational overhead arising from running a large number of o/s instances.

Container Approach: places the virtualization management software on top of (or combined with) a single operating system instance. System resources are allocated to multiple workloads (or virtual environments) by the VM software running under Windows or Linux operating system. This is a hybrid model developed on client/server architectures with the most prevalent example products being Microsoft's Virtual Server and SWsoft's Virtuozzo. The major advantage here is reduced system overhead due to the presence of a single operating system. Another advantage is that this approach may allow more granular dynamic re-allocation of system resources among workloads.

Both of the above methods have mechanisms that automatically adjust resource levels for particular workloads according to business needs without human monitoring or intervention. Even better, these products generally have built-in metering and charge back modules, which will allow IT to track (and charge for) actual system usage.

x86 Virtualization – Getting Started

There are a few key points to keep in mind before approaching any consolidation project, particularly something as new as x86 server consolidation. First, begin with reasonable expectations. While 100% average and peak utilization is possible with mainframes, goals for x86 server utilization should be somewhere around 30-40% average utilization with 50-60% peak utilization. This will ensure that there is plenty of headroom for excessively high peaks and give the VM software enough resource to adequately manage the virtualized partitions or containers. These targets will increase as users gain more experience with x86 virtualization and the technology improves over time.

Another consideration is which applications should (or should not) share space on a virtualized system. Heavy transactional workloads that utilize lots of operating system functionality (best example is databases) and are highly latency sensitive are not great virtualization candidates. Web workloads, file servers, along with test and development tasks, plus light-weight business applications work well on virtualized systems and are the best candidates for consolidation.

Most organizations will find that the biggest hurdles to virtualization/consolidation are internal politics and cultural issues. Any consolidation of workloads is essentially a change in control over of IT assets, which, to whoever is losing control, could be felt as a loss of power. People who feel the process is taking away parts of their kingdom will often do anything they can to derail the process. Of course, consolidation efforts have also been stopped in their tracks by opponents raising valid concerns about SLA adherence, reliability/availability and security. A successful initiative will include a detailed business case that outlines the costs and benefits of the consolidation plus a technical case that covers the nuts and bolts of configurations and project management.

In our experience, the most critical mistake most organizations make is to improperly scope the first virtualization project. We have seen many organizations try to bite off too a big an infrastructure 'chunk' in their first consolidation attempt and get bogged down in details and political warfare. Many of these firms view consolidation or virtualization as a point project, where they do it once and it's fixed forever. However, these are not one-time big bang projects; they are new usage models and need to be viewed as evolutionary changes in IT operations.

With this point in mind, it is far better to start with a small number of systems, running well understood applications that serve a limited (and hopefully friendly) set of users. This makes the project much more manageable, lowers overall risk, and helps the IT department learn through experience. Early successes are the foundation for more ambitious future projects that will have a larger impact on the bottom line.

GCG Recommendations

We are enthusiastic supporters of x86 virtualization and believe the technology is now in place to make multi-application x86 servers a reality. Clients who use virtualization to reduce the number of servers in their data centers will reduce hardware acquisition costs, facilities costs, and, most importantly, personnel costs over time. We will discuss specific recommendations for tackling x86 server consolidation (building the business case, technical case, and product recommendations) in subsequent research reports, so stay tuned....

The information contained in this publication has been obtained from sources believed to be reliable. Gabriel Consulting Group does not warrant the completeness, accuracy, or adequacy of this report and bears no liability for errors, omissions, inadequacies, or interpretations of the information contained herein. Opinions reflect the judgment of Gabriel Consulting Group at the time of publication and are subject to change without notice.

Gabriel
CONSULTING GROUP
phone / 503.372.9389
gcginfo@gabrielconsultinggroup.com
www.gabrielconsultinggroup.com