

# Virtualizing Mission-Critical Applications

A VMTurbo White Paper



# TABLE OF CONTENTS

- 1 SUMMARY ..... 3
- 2 CHALLENGES OF VIRTUALIZING MISSION-CRITICAL APPLICATIONS ..... 4
  - 2.1 The Challenges of Silos Inefficiencies ..... 4
  - 2.2 The Challenges of Controlling Interference..... 6
  - 2.4 The Challenges of Managing SPP ..... 9
- 3 BEST PRACTICES .....11
  - 3.1 Smart Consolidation of Critical and Non-Critical Workloads.....11
  - 3.2 Better Placement.....12
  - 3.3 Adaptive, Real Time SPP Management.....13
- 4 IMPLEMENTING AN EFFECTIVE SOLUTION ..... 14
- 5 CONCLUSIONS..... 15
- 6 REFERENCES ..... 15
- 7 ABOUT VMTURBO..... 16

---

## 1 SUMMARY

---

The virtualization of mission-critical applications such as Microsoft Exchange, Oracle SQL server and SAP offers substantial benefits to any organization that can successfully manage its associated complexity. It can eliminate the capacity waste and costs of remaining over-provisioned applications silos, while greatly increasing the efficiency of IT processes-- from operations management and disaster recovery to development, deployment and upgrades of applications. However, to rip these benefits IT organizations need to master the intricacies of virtualizing mission critical applications.

This whitepaper explores the challenges and best practices for virtualizing mission-critical applications. We examine key issues and identify major practical barriers that organizations face in this task. We conclude with advice for implementing an effective solution in your organization.

Traditional IT architectures dedicate hosts to applications and overprovision them to handle peak workloads. These over-provisioned silos can assure applications performance. But the excess capacity required for peak workloads is idle most of the time, leading to substantial waste. Virtualization technology is often first adopted in an effort to capture some of this wasted capacity. Non-critical applications are migrated and consolidated onto a virtualization platform, where hypervisors eliminate silo boundaries to provide more efficient resource sharing. Hypervisors increase resource utilization by exploiting statistical fluctuations in applications workloads; when a resource is not needed by some applications, is used by others. Of course, if the aggregate resource demand exceeds its capacity, applications will experience mutual interferences. These interferences, if lasting, can build into congestion or sustained bottleneck. Non-critical applications can often tolerate some interference. In practice, however, as organizations increase utilization targets, they soon discover the challenges of managing the complex tradeoffs between increased utilization and interferences.

Mission critical applications can be particularly sensitive to interferences in resource access. They thus require special care in managing their resources: First, performance must be assured with the ability to detect, resolve and view performance problems in real time. Second, the existing environment must be continually improved with an ability to automatically execute workload placement and support multiple quality of service classes. Finally, one must plan for the future by analyzing demand and hardware change scenarios to obtain an optimal execution plan.

VMTurbo has developed significant tools and operational experience in the management of virtualized workloads. Here we introduce three best practices to support efficient virtualization of mission critical workloads:

***Maintain a consolidation-balance between performance-sensitive and insensitive workloads***

***Use improved workload placement algorithms***

***Provide adaptive control to optimize resource use and avoid interference***

## 2 CHALLENGES OF VIRTUALIZING MISSION-CRITICAL APPLICATIONS

The virtualization of mission-critical applications such as Microsoft Exchange, Oracle SQL Server and SAP is often viewed as a risky endeavor. The prospects of business downtime, data loss and security breach are perceived to be too great to entrust critical applications to mere virtual machines. But today some organizations successfully confront the challenge of virtualizing mission-critical applications. As a result they reap many benefits that greatly increase their agility and competitiveness in the marketplace.

Consider a few of these benefits

- Unified infrastructures for all apps
- Improved utilization and efficiency
- Streamlined operations management
- Efficient disaster recovery
- Upgrade and deployment ease
- Reduced power and space usage
- System mobility
- Simplified development, testing and training

Let's take a look at the major challenges and resulting transformation when organizations embrace virtualization of mission-critical applications.

### *Benefits of Mission-Critical App Virtualization*

- *Unified infrastructures for all apps*
- *Improved utilization and efficiency*
- *Streamlined operations management*
- *Efficient disaster recovery*
- *Upgrade and deployment ease*
- *Reduced power and space usage*
- *System mobility*
- *Simplified development, testing and training*

### 2.1 The Challenges of Silos Inefficiencies

IT architecture has traditionally been based on silos wherein specific applications get dedicated to specific physical computing resources. The physical resources commonly have over-engineered capacity to accommodate applications' peak demands. This Dedicated Peak Provisioning – called here DPP -- compartmentalizes applications into separately deployed and managed silos. Virtualization replaces these DPP silos with common infrastructures to share resources among applications.

We a real world example<sup>1</sup> of a Microsoft Exchange Server to illustrate the challenges of virtualizing mission critical application. Figure 1 shows the IO capacity utilization, by the Exchange workload, through daily operations. The blue curve represents average utilization while the red curve represents utilization peaks. We observe that peak utilization reaches 100% for a very short duration around 10pm and then for a more sustained duration between 3am-4am, with lingering effects until 5am. The workloads peaks are primarily associated with nighttime administrative tasks such as data backup. Microsoft best-practice guidelines for Exchange, recommend provisioning sufficient resource capacity to assure that peak workloads of mission-critical exchange Roles (sub-applications) do not exceed certain

targets utilization levels (e.g., 50%-70% for mailbox processing Roles). This DPP strategy permits non-critical administrative Roles to share the underlying resources during night hours when critical workloads are minimal.

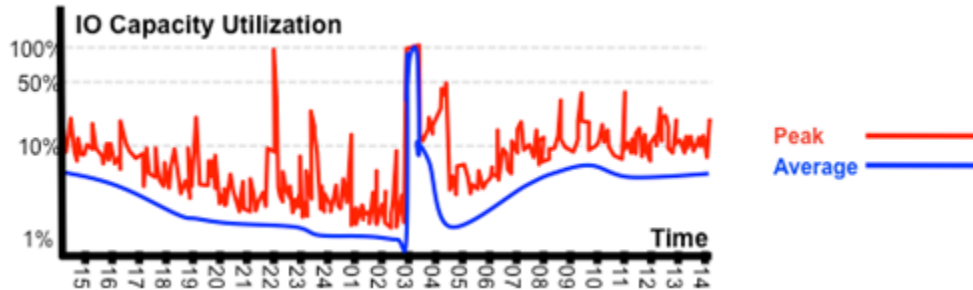


FIGURE 1: IO WORKLOAD OF EXCHANGE SERVER

The administrative Roles are less sensitive to resource availability and can thus tolerate the congestion and delays that come with 100% utilization. Indeed, as observed in Figure 2, the traffic peak starting at 3am creates a massive sustained congestion, reflected by the peak and average utilization rising to 100% until 4am and lingering on until cleared at 5am. Were administrators concerned with performance of these administrative Roles, they could allocate additional capacity to keep peak utilization lower.

We see that the DPP management strategy can help assure the performance of critical Roles. But take a look at Figure 2. During most of the day, the average utilization ranges only 3%-8%! The low utilization “passively” assures that performance-sensitive Roles (e.g., email flow) will have sufficient resources to meet their stringent performance goals. The price of this performance assurance is a waste of some 92%-97% of IO capacity during the majority of each day. Even if one considers peak utilization and not just averages, the peaks utilize 8%-15% of the capacity most of the day, leaving 85%-92% idle.

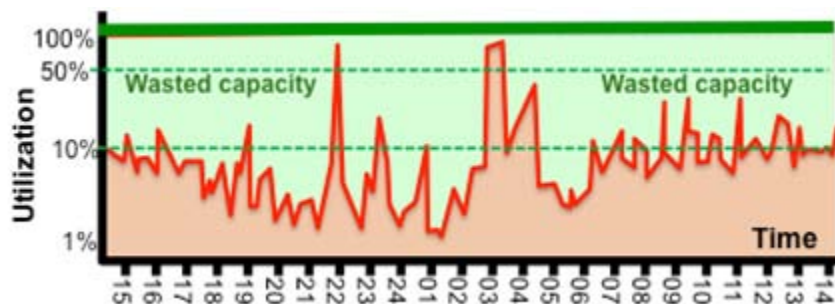


FIGURE 2: 85%-92% OF IO CAPACITY IS WASTED MOST OF THE DAY

As enterprises expand their virtualization infrastructures silos, such as the one used in the Exchange example, become isolated pockets of waste. Capacity waste is but one factor of inefficiency. No less significant are the waste factors associated with different operations management processes for each silo – e.g., disaster recovery to deployment and upgrades. It is little wonder that IT organizations seek to eliminate these remaining silos and their inefficiencies by virtualizing mission critical applications.

However, virtualizing mission critical applications presents far from trivial challenges, considered by the next section.

## 2.2 The Challenges of Controlling Interference.

Virtualization of Exchange starts with packaging Roles into individual VMs. These VMs may be consolidated with VMs of other applications to exploit the 85%-92% idle IO capacity, we observed in Figure 2. Figure 3 shows the peak (purple) and average (green) utilization curves for the consolidated IO workload. The dashed red and blue curves in the background show the peak and average utilization of the Exchange Role components.

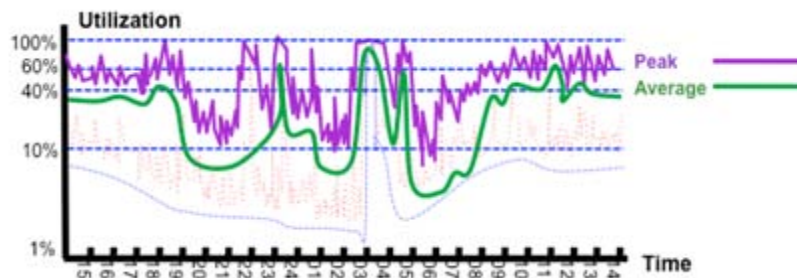


FIGURE 3: HYPOTHETICAL EXCHANGE SERVER WORKLOAD CONSOLIDATION

The workload contributions from additional (non-critical) apps are the differences between the green/purple and blue/red curves. We observe that the additional (non-critical) workload has two very important effects:

1. Significant utilization gains over DPP are realized along with elimination of inefficiencies. The average utilization increases to the 40%-50% range, during the workday, in contrast with the 3%-8% of DPP, eliminating IO capacity waste observed in Figure 2.
2. Interference with the Exchange traffic pushes peak utilization close to 100% for long periods of time. For example, the brief peak of Exchange at 10pm (Figures 2 and 3) is transformed into sustained congestion between 9pm-12pm. During other times interference can create shorter-lived congestion, disrupting mission-critical Exchange Roles.

***When Critical Apps are virtualized resource utilization increases but so does interference.***

In general, non-critical applications, such as print servers and web-development servers, can accommodate large degree of interference without significantly impairing business processes. Virtualizing such applications can realize substantial utilization gains without the steep price of disruptions. In contrast, mission critical apps may be very sensitive to interference. It is therefore necessary to control such potential interference to assure the performance of critical apps.

We use an example to illustrate interference control challenges and best practice solutions. Figure 4(a) shows a virtual-CPU (vCPU) allocation needed by 6 virtual machines (VM) sharing one physical machine (PM). Figure 4(b) outlines the CPU co-scheduling mechanism of ESX. The PM has 8 cores, depicted as circles and colored as the VMs they service. For example, core D is allocated to service the vCPU of VM2, while cores G and H are allocated to service the 2 vCPUs of VM5. The cores colored white (A and B) are available. VMs ready for processing wait in the CPU\_Ready\_Queue until they can get assigned cores to service their vCPU needs.

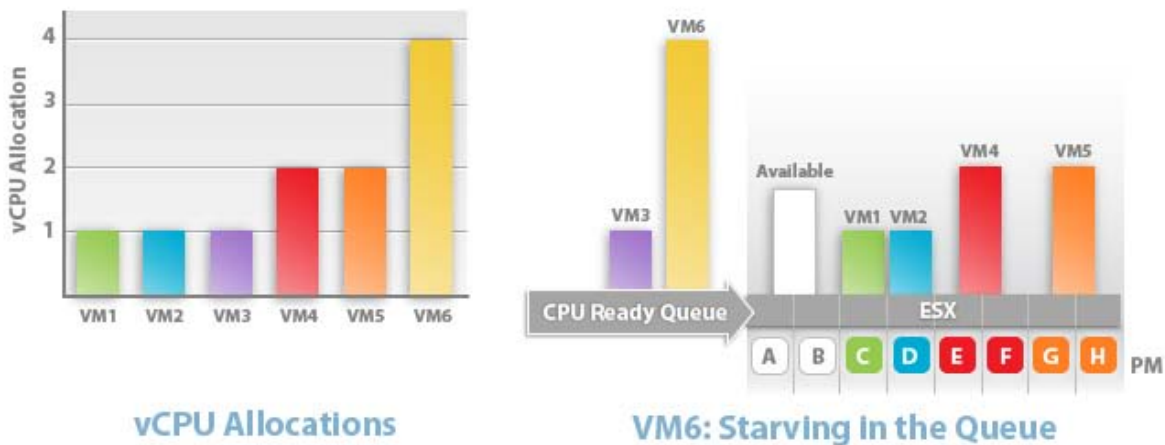


FIGURE 4:

4.A) CPU ALLOCATIONS NEEDED BY VIRTUAL MACHINES, 4.B) MISSION CRITICAL APP STARVED FOR CPU

Let’s assume VM6 runs a mission-critical compute-intensive app that requires a virtual symmetric multiprocessor (vSMP) with 4 vCPUs. It waits in the CPU\_Ready\_Queue for 4 cores to become available for co-scheduling. But the ESX hypervisor scheduler has only two available cores, so it allocates core B to VM3, ahead of VM6. Suppose, next, VM4 terminates its time-slice, and frees cores E and F. The ESX hypervisor will have 3 cores available (A, E and F), yet the available resources are still insufficient for VM6. So VM4 gets assigned its 2 cores, say A and E, ahead of VM6. *VM6 may thus remain starved in the CPU\_Ready\_Queue, for long periods, while other VMs get served ahead of it.* This interference in VM6 access to CPU resources can greatly impair or completely block the mission-critical application.

This interference is caused by two factors:

1. **Peak workloads:** The VMs workloads highly utilize their vCPU allocations. Indeed, if VM1-VM5 are lightly loaded they will free their cores frequently, permitting VM6 to obtain the 4 cores it needs.
2. **Over-commitment of CPU resources:** The aggregate VMs demand, of 11 vCPUs, exceeds the PM capacity of 8 cores. Indeed, suppose the aggregate demand is reduced to 8 vCPUs, by moving VM3 and VM5 to another PM, the needs of all VMs would be met.

Similar Interference effects, involving memory access, are caused by memory over-commitment during peak workloads. For example, suppose the aggregate memory allocations of the 6 VMs, in Figure 4, is 20GB but the PM has only 16GB. Should the aggregate memory workload exceed 16GB at any time, it will be necessary to swap memory pages into/from slower disk storage. Although the mission-critical application will continue to function, such swapping can impair its performance considerably.

Numerous “best practice” guides attempt to resolve such interference by “over-provisioning” virtual machines, much like traditional over-provisioning of physical machines. This strategy, which we call Shared-Peak-Provision (SPP), consists of two best-practices rules:

1. [SPP-I] **Assure peak utilization:** allocate sufficient resources to VMs to handle their peak workloads.
2. [SPP-II] **Avoid over-commitment:** assure that the aggregate allocation of a resource to VMs do not exceed the capacity of the underlying PM.

The SPP rules could easily resolve the co-scheduling starvation scenario of Figure 4. Indeed, according to rule SPP-I, each VM will allocate the vCPUs it needs for its peak demands. The aggregate allocation of the 6 VMs requires 11 cores. To meet SPP-II, the aggregate allocation should not exceed the PM capacity of 8 cores. Therefore, to meet SPP-II one can shift VM3 and VM5 to alternate PM and thus eliminate the possibility of co-scheduling starvation. SPP forms the core of best practice guides for virtualizing critical applications. For example, the VMWare Best Practices Guide (BPG) for virtualizing Exchange 2010<sup>2</sup> recommends:

- **BPG page 8:** “For performance-critical Exchange virtual machines (i.e., production systems), try to ensure the total number of vCPUs assigned to all the virtual machines is equal to or less than the total number of cores on the ESX host machine.”
- **BPG page 26:** “It is recommended that standalone servers with only the mailbox role be designed to not exceed 70% utilization during peak period”
- **BPG page 9:** “Do not over-commit memory on ESX hosts running Exchange workloads.”

Other best practices guides to virtualizing mission-critical applications recommend similar SPP rules.

How does SPP compare with traditional DPP, considered in Section 2.1? DPP dedicates physical resources to service the peak workloads of a mission-critical application. SPP, likewise, seeks to assure that resources are sufficient for peak workloads. Indeed, if each VM reserves the resources for its peak

workloads, SPP is reduced to DPP. However, SPP does not require such over-provisioning through reservations. Applications can flexibly share physical resources during off-peak times.

Predictably, the BPG<sup>2</sup> recommends avoiding reservations to permit such flexibility:

- **BPG page 8:** *"Setting a CPU Reservation sets a guaranteed CPU allocation for the virtual machine. This practice is generally not recommended because the reserved resources are not available to other virtual machines and flexibility is often required to manage changing workloads."*

A mission critical application may thus use "reservations" to guarantee its performance under average workloads, and use priorities (e.g., by allocating "shares" ) to assure its performance through peak periods. Therefore, unlike DPP, SPP permits applications to assure their performance through peak traffic without dedicating the resources needed.

## 2.4 The Challenges of Managing SPP

Despite their apparent simplicity, the SPP rules are difficult to manage because they involve practices which are quite complicated to handle manually. These complicated practices include

- Mixing production and non-production applications;
- Analyzing temporal behaviors of peaks;
- Tuning resource allocations, reservations and shares to assure the performance of production applications; and
- Adapting these placements as workloads and resources change, which, even for small sites, is too complex for manual handling requiring workload and capacity management tools.

### *Practical Challenges of SPP:*

- *Mixing production and non-production apps*
- *Analyzing time behavior of peaks*
- *Tuning resource allocations to assure critical app performance*
- *Adaptive workload placement and resource changes*

Let's take a closer look at these challenges. To start with, SPP requires constantly monitoring peak workloads and assuring that the SPP rules are valid. Consider again the scenario of Figure 4, adjusted to meet SPP-II, by keeping only VM1, VM2, VM4 and VM6. Suppose VM4 increases its allocation from 2 vCPUs to 4 vCPUs. The aggregate allocation of 10 vCPU violates SPP-II leading to potential starvation. Even a small enterprise with a few scores of VMs may find it difficult to manually manage the SPP rules through similar (minor) allocation changes.

More surprisingly, it has been found that if mismanaged, SPP may waste even more resources than DPP and thus invalidates the overall virtualization effort of mission-critical applications! Figure 5 shows the IO workload of an Exchange Server, with Figure 5(a) depicting the non-virtualized IO workload. Now assume the server is decomposed into 3 virtualization roles (corresponding to Exchange Roles), each packaged into separate VM1, VM2 and VM3. Figure 5(b) depicts the IO workloads of these 3 VMs individually (pink, blue, green) and combined (dotted red).

The non-virtualized Exchange requires sufficient capacity to handle its peak load of 10k IOPS. If the peak utilization desired is, for example, 70%, then DPP will over-provision IO capacity of  $10k/0.7 = 14.3k$  IOPS. How much IO capacity will SPP require for the 3 Exchange Roles VMs? VM1 peaks at 4k IOPS, VM2 at 5k IOPS and VM3 at 7k IOPS. The IO capacity allocations required to support 70% peak utilizations, are respectively 5.8k, 7.2k, and 10k IOPS. The capacity needed to meet SPP-II must exceed the aggregate allocation of 23k IOPS. Therefore, SPP requires a PM with IO capacity of at least 23k IOPS, which is 61% above the 14.3k IOPS over-provisioned by DPP. Therefore, while the virtualized and non-virtualized servers have the same aggregate workloads, the waste factor by SPP far exceeds DPP's by 61%.

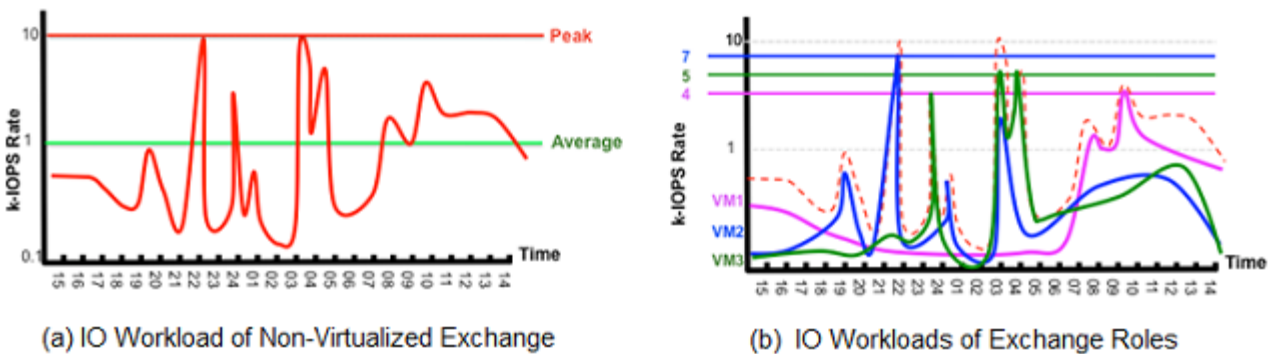


FIGURE 5: IO WORKLOADS OF (a) NON-VIRTUALIZED EXCHANGE, (b) THREE EXCHANGE ROLE VMs.

This inefficiency in applying SPP results from two factors. First, the sum of the peak workloads used by SPP is greater than the peak of the aggregate workload. SPP cannot ignore the possibility of concurrent peaks, even if these are unlikely to occur. It thus over-provisions capacity for such worst case.

Second, and more importantly, consolidating multiple mission-critical applications can be troublesome and thus requires such extreme over-provisioning. Suppose, instead, that the mission-critical app of VM1 is consolidated with two performance-insensitive apps. One could allocate much lower capacity to VM2 and VM3, while reserving sufficient capacity for VM1 to assure its priority. This would permit one to meet the SPP rules with much less capacity, assure the performance of the mission-critical app and permit the performance-insensitive apps to utilize the capacity left by the mission-critical app.

An even more intriguing possibility for efficient virtualization of mission-critical applications is to consider the temporal behaviors of workloads in managing the SPP rules. Most applications generate their peak workloads during specific times of the day. For example, the Exchange mailbox server, using

VM1 in Figure 5(b), has its peaks between 9am-11am, while the Roles of VM2 and VM3 require marginal IO capacity, peaking at very different times.

The SPP rules do not specify the periods over which one should apply them. The inefficiency of SPP relative to DPP arises when one tries to apply SPP over a daily period. Suppose, instead, that one applies SPP for the workday 7am-6pm only. The peak workload of VM1 remains 4k IOPS as before. However, the peak workloads of VM2 and VM3 are under 1k IOPS. Therefore, during the workday, SPP would require an aggregate IO capacity of 8.6k IOPS [ $8.6=(4+1+1)/0.7$ ] which is certainly more reasonable than 23k IOPS. Therefore, in applying SPP one should consider not only what the peaks are, but also the times at which they occur.

### 3 BEST PRACTICES

VMTurbo working with its customers has developed significant operational experience in the management of virtualized workloads similar to those presented above. In this section we introduce three best practices that we have developed to improve the fundamental SPP inefficiencies of mission critical workload consolidation. These best practices are

1. **Maintain a consolidation-balance between performance sensitive and insensitive workloads.**
2. **Use improved workload placement for optimal packing by accounting for all constraints, exploiting flexibility of performance-insensitive workloads and dynamics of workload peaks & troughs.**
3. **Use adaptive workload control to exploit dynamics to reduce waste of static policies while eliminating any emerging interference scenarios.**

Let's consider each in turn and relate it to the challenges of virtualization that we have encountered.

#### 3.1 Smart Consolidation of Critical and Non-Critical Workloads

To avoid over-provisioning waste, when using SPP, we must balance different types of workloads in an intelligent manner.

Consider the example in Figure 5. An inefficient SPP management practice requires 24k-IOPS for a workload that averages only 1k-IOPS. In order to avoid the waste, one may consolidate additional workload, of non-critical apps, to utilize larger share of the IOPS capacity. For example, consolidate spam-filter and email archive applications, with the critical performance-sensitive Exchange Roles. The performance-insensitive applications would increase capacity utilization during non-peak times. When

*Best Practice #1  
Maintain a consolidation-  
balance between  
performance sensitive and  
insensitive workloads*

critical applications require IO resources during a peak time, the performance-insensitive applications are pre-empted to release these resources to the critical application until the peak has passed.

### 3.2 Better Placement

SPP sets constraints on the allocations and placements of VMs. In particular, the aggregate allocation of resources by VMs placed at a PM, must not exceed its capacity (SPP-II). An optimized placement must thus pack the VM allocations into a minimal number of physical machines subject to a variety of constraints. This optimization problem belongs to a class of complex algorithmic problems known as bin-packing.

Manual placements often pursue a “first fit” strategy resulting in highly inefficient solutions, as compared with optimized algorithmic placements. To illustrate the opportunity, consider the “first fit” mapping of VMs to PMs shown in Figure 6. Blue, green and pink bars show CPU, memory, and IO requirements of each VM. The challenge is to place the 10 workloads into as few PMs as possible. Figure 6a shows results for a simplistic first fit algorithm. Starting from left to right, each VM is sequentially placed into the next available PM. Once capacity is exceeded in the given PM, the next PM is provisioned. A total of 5 VMs are required to place the VM workloads. Consider the placement of Figure 6b using an optimized workload placement algorithm. All constraints for the ten VMs are taken into account. Flexibility of performance-insensitive workloads is exploited, in addition to the dynamics of the workload peaks & troughs. The resulting optimal placement saves 40% of PM capacity by fitting the workload into only 3 VMs.

**Best Practice #2**  
*Use improved workload placement for optimal packing by accounting for all constraints, exploiting flexibility of performance-insensitive workloads and dynamics of workload peaks & troughs*

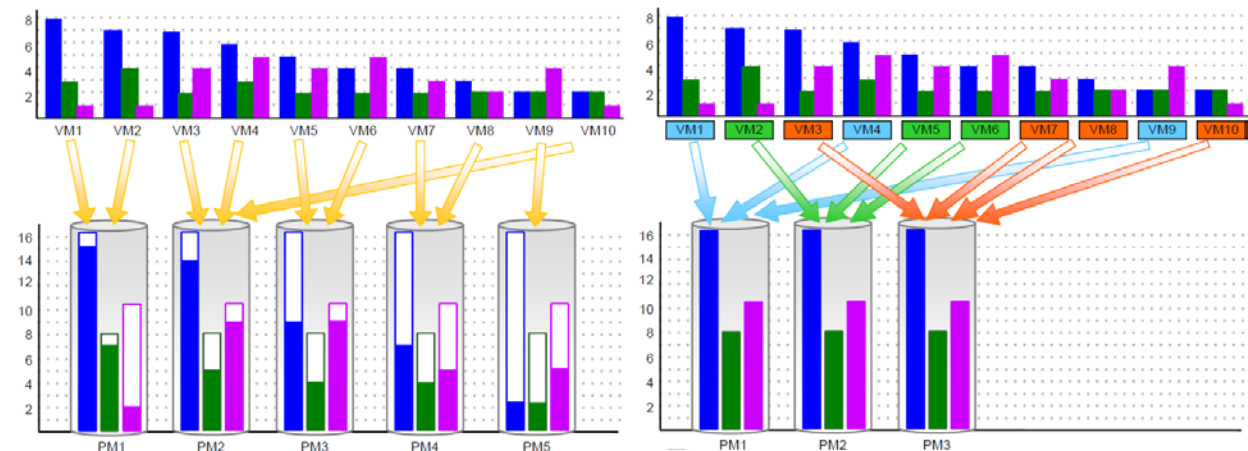


FIGURE 6: WORKLOAD PLACEMENT (A) FIRST FIT ALGORITHM, (B) OPTIMAL PLACEMENT ALGORITHM.

### 3.3 Adaptive, Real Time SPP Management

Workload demands and interferences change over time. A static SPP management can result in significant inefficiencies by ignoring these dynamics. For example, the Exchange administrative Roles (VM2, VM3), depicted in Figure 5, are active through the night. A static SPP management may keep their reservations and SPP constraints during the day, even when these are irrelevant. In fact, A static SPP management will require 16k-IOPS to meet SPP II. In contrast, adaptive SPP management will use only 4k-IOPS during 7am-6pm, permitting 12k-IOPS (75% of the capacity) to be used by additional VMs.

A static SPP management, furthermore, has poor ability to handle emerging problems such as interference with critical workloads resulting from unexpected fluctuations. In contrast, adaptive SPP management reduces waste by exploiting the workload dynamics and can efficiently apply SPP rules to avoid interference resulting from short term fluctuations.

As administrators move toward implementation of an effective solution using the three best practices for virtualizing mission-critical applications presented here, they may feel overwhelmed. Effective SPP management is a tall order, requiring administrators to

- Master voluminous details of hypervisor and applications internals
- Manage interference and waste problems manually
- Manage resource allocations and move applications as workloads change
- Maintain tight-coordination between virtualization and application administrators

In the next section we provide guidance toward implementing an effective solution for your own organization.

**Best Practice #3**  
*Use adaptive workload control to exploit dynamics to reduce waste of static policies while eliminating emerging interference scenarios.*

## 4 IMPLEMENTING AN EFFECTIVE SOLUTION

The implementation of effective SPP management practice is a challenge for many organizations. Let's take a look at some of the major challenges they face and the success criteria involved in implementing an effective solution.

Three primary areas (Figure 7) drive a successful implementation. First of all, because performance is critical to the mission of the organization, it must be assured. Performance assurance involves detection of problems that impact the mission-critical applications; resolving and preventing those problems; and sharing key performance metrics with stakeholders. The associated *success criterion is the ability to detect, resolve, and view performance problems in real time.*

Second, the existing operating environment must be continuously improved. Existing resource capacity must be optimized. Operating expenses must be reduced. More and varied classes of workloads must be supported. The associated *success criteria are ability to optimize in real-time using several metrics, automatically executing workload placement (as opposed to static, manual placement), and finally the support of multiple quality of service classes.*

<b>Assuring Performance</b>	
<ul style="list-style-type: none"> <li>• Do I have problems impacting applications?</li> <li>• How do I resolve &amp; prevent problems?</li> <li>• How do I share key metrics with stakeholders?</li> </ul>	<ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> Automated Real-time problem detection</li> <li><input checked="" type="checkbox"/> Real-time resolutions</li> <li><input checked="" type="checkbox"/> Real-time dashboards and historical reports</li> </ul>
<b>Optimizing the Environment (OpEx)</b>	
<ul style="list-style-type: none"> <li>• How do I maximize resource capacity?</li> <li>• How do I reduce operating expenses?</li> <li>• How can I support more classes of workloads?</li> </ul>	<ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> Real-time optimization based on multiple metrics</li> <li><input checked="" type="checkbox"/> Automatic execution of workload placement</li> <li><input checked="" type="checkbox"/> Support multiple QoS classes</li> </ul>
<b>Planning for the Future (CapEx)</b>	
<ul style="list-style-type: none"> <li>• How do I virtualize more applications?</li> <li>• What is the impact of changes in demand?</li> <li>• What is the impact of hardware changes?</li> </ul>	<ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> Automatic analysis of multiple scenarios</li> <li><input checked="" type="checkbox"/> Wizard to design and evaluate scenarios</li> <li><input checked="" type="checkbox"/> Provides an optimal execution plan</li> </ul>

FIGURE 7: MAJOR IMPLEMENTATION CHALLENGES AND SUCCESS CRITERIA

Third, successful organizations plan for the future. This includes virtualization of more applications being used throughout the organization. The ability to perform “what-if” impact analyses for changing application demands and hardware changes. The associated *success criteria include the ability to analyze multiple scenarios, speedy ways to design and evaluate scenarios, and ability to obtain an optimal execution plan.*

## 5 CONCLUSIONS

---

The virtualization of mission-critical applications such as Microsoft Exchange, Oracle SQL Server and SAP can be successfully managed provided that several best practices are incorporated into the implementation. The initial efforts to virtualize traditional IT “silos” may result in immediate short term productivity gains, improved IT efficiency and application performance. However, significant challenges arise as an organization moves from virtualization of “low hanging fruit” applications towards mission-critical, performance-sensitive applications. Virtualization of these critical applications requires new performance management technology to handle the numerous factors of complexity.

Best practices, Shared-Peak-Provisioning rules, provide a useful start. But, in practice, their manual management is too complex and often results in significant inefficiencies including: over-provisioning waste, reduced workload performance and poor physical machine utilization. Therefore, SPP needs to be augmented by automated intelligent workload management tools incorporating three primary mechanisms. First, the elimination of over-provisioning waste through balanced consolidation. Second, the improvement of workload placement for optimal packaging using constrained workload dynamics. Finally, adaptive control of workloads to exploit their dynamics and eliminate emerging interference between workloads.

## 6 REFERENCES

---

[1] Narayanan, D. et.al, “Everest: Scaling down peak loads through I/O off-loading”, 8th USENIX Symposium on Operating Systems Design and Implementation,  
[http://www.usenix.org/events/osdi08/tech/full\\_papers/narayanan/narayanan.pdf](http://www.usenix.org/events/osdi08/tech/full_papers/narayanan/narayanan.pdf)

[2] [http://www.vmware.com/files/pdf/Exchange\\_2010\\_on\\_VMware\\_-\\_Best\\_Practices\\_Guide.pdf](http://www.vmware.com/files/pdf/Exchange_2010_on_VMware_-_Best_Practices_Guide.pdf)

---

## 7 ABOUT VMTURBO

---

VMTurbo delivers an Intelligent Workload Management solution for Cloud & Virtualized environments. VMTurbo uses an Economic Scheduling engine to dynamically adjust resource allocation to meet business goals. Using VMTurbo our customers ensure that applications get the resources they need to operate reliably, while utilizing infrastructure & human resources in the most efficient way.

VMTurbo is headquartered in New York, with offices in California, Massachusetts, United Kingdom and Israel.