

RESOURCE PROVISIONING REPORT

What we learned after studying resource utilization of
148,233 production machines

EXECUTIVE SUMMARY

The throw-more-hardware-at-the-problem mentality has caused an epidemic of over-provisioned servers in today's IT data centers around the globe.

One of the biggest benefits of virtualization is the ability to distribute available physical resources across many workloads. This cuts down on server waste by more fully utilizing the physical server resources and by provisioning virtual machines with the exact amount of CPU and memory that it needs.

Sizing a VM with just the right amount of resources is difficult. It's hard to predict weekly and seasonal spikes, user and application behavior changes over time, and application vendors often recommend more resources than needed.

Over-provision a server with resources it'll never use -- you're wasting your budget. Provision too little and your performance will suffer -- users will complain of slow virtual desktop logon, business applications will run sluggish and some systems may even grind to a halt because of resource contention.

Given the consequences, it's not a surprise that IT organizations tend to use over-provisioning as a de-risking strategy. The question we set

out to answer is whether that is true in real life and if so, how severe is the over-provisioning situation?

For this research we studied over 148,000 production workloads across 943 organizations worldwide, inspected the amount of CPU and memory allocated to each server, and compared that to the amount actually used. We expected to see resource over-provisioning but we were surprised by the magnitude.

These findings are meant to provide confidence for IT organizations to right-size their server resources and provide a better overall IT service. The research plainly illustrates the need for better focus on monitoring resource consumption with advanced tools. Not only are there direct financial benefits of correct provisioning, end user applications, operating system, and server applications will perform better.

Not only are the economics more favorable, effective IT infrastructure monitoring is critical to all aspects of an IT operation.

Rotem Agmon

Rotem is a hopeless geek, addicted to space operas and home improvement shows. When not slacking, he is a cloud and virtualization architect and has presented at several international VMware conferences.

Bassam Khan

Bassam is responsible for product-related communication, strategic product and company direction and various marketing functions.

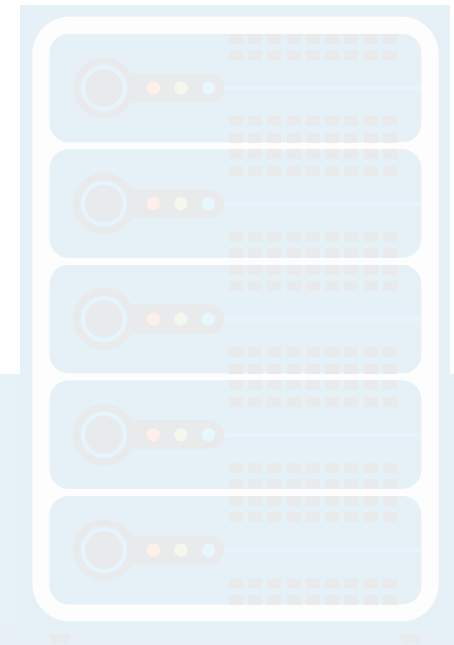


TABLE OF CONTENTS

INTRODUCTION Page 04

REPORT METHODOLOGY Page 10

FINDINGS
CPU PROVISIONING Page 16

FINDINGS
MEMORY PROVISIONING Page 19

FINDINGS
PROVISIONING BY OS Page 22

FINANCIAL
IMPLICATIONS Page 25

DISCUSSIONS Page 27

CONTACT US Page 30

INTRODUCTION


BALANCE OF RESOURCE USAGE AND AVAILABILITY

ControlUp provides a cloud/on-premises hybrid IT management solution to organizations across all industries. The cloud component allows our customers to receive community-based metrics on numerous aspects of IT operations and see where they compare to other organizations in their own industry. Amongst the hundreds of metrics related to resource allocation and consumption in virtualized data centers, this report focuses on the two most important ones; processor and memory. Here is a good description of how physical resources are shared across VMs ⁽¹⁾.

Predicting a server's RAM requirement is complex. For example, a database application will use a CPU's cycles and release it instantaneously, but not so with RAM - it may consume all of the RAM it's given to optimize performance. With end-user computing, user behavior is hard to predict; the more Windows spun up, the more RAM is locked up. When memory resources are not sufficient for the workload, the effect is immediate and obvious, not just to IT but often to the entire company. Impact ranges from slow performance to service outages.

CPU and RAM are expensive. Unlike home PCs, servers require more CPUs and more cores, and a robust socket architecture. Similarly, data center RAM is designed to reduce failures. A server's RAM includes industrial-grade features, such as built-in error correction code, efficient electrical load on the memory controller and better heat management. IT invests in fault-tolerance because service outages and downtime are even more costly. Over-allocating just four DDR4 4x4GB DIMM means a waste of \$1,000⁽²⁾ of IT budget that could have been spent elsewhere. Similar cost structure applies to over-provisioning CPUs⁽³⁾.

While the cost impact of over-provisioning is obvious, more dramatic is the performance impact; over-provisioning either CPU or RAM will hurt a VM's performance, as well as its neighboring VMs.



KEY TAKEAWAY:
In addition to the money-wasting aspect of unused resources, over-provisioning may lead to degraded performance.

INTRODUCTION

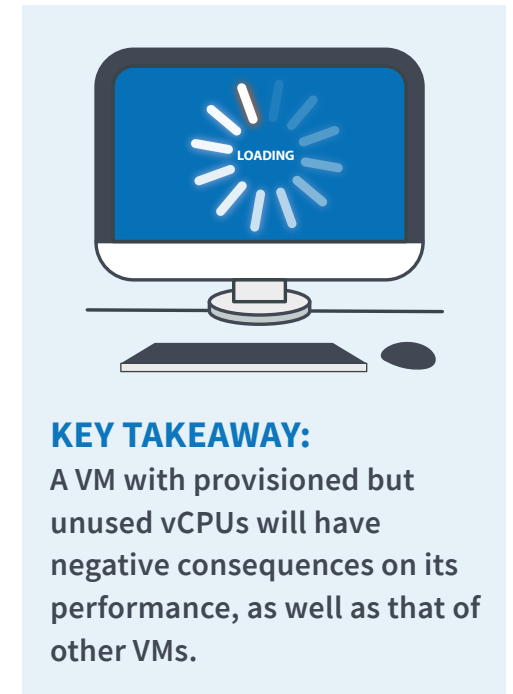
PERFORMANCE IMPACT OF vCPU OVER-PROVISIONING

It all comes down to the application running in the VM. If that application was written with a multi-threaded structure with parallelization support to leverage a multi-core CPU, allocating many vCPUs is advantageous. For all other applications, more vCPUs does not equate to better performance.

The problem arises from CPU scheduling challenges. The virtualization host (e.g., vSphere, Xen, KVM, Hyper-V and other) accepts and forwards processing requests made by the VM to the physical CPU. If the physical CPU is busy, the server will queue new requests until CPU resource frees up. If we allocate 16 vCPUs for one VM, the virtualization host must wait until 16 CPUs are available before accepting workload from that VM. Not only does that VM wait until 16 CPUs free up, other VMs with low vCPU requirements must also wait. During that time the scheduler will not send the physical CPU requests, even if they're idle. This phenomenon is also known as "co-stop".

Requesting more vCPU than the application can utilize is analogous to reserving a large table at a restaurant. If you request a table for 8 at a busy restaurant when you only have 4 people in your group, not only will your party have to wait a longer time for the larger table, your unused seats will also increase wait times for other restaurant patrons. Here is a somewhat dated but solid explanation of CPU scheduling⁽⁴⁾.

In a data center this symptom is indicated by a high CPU Ready time and a low figure for CPU utilization, along with poor overall application performance. New scheduling techniques lessen the impact of over-provisioning vCPUs, but only for the smaller vCPUs; the high-vCPU VM performance still suffer. Proper vCPU allocation will indeed improve a single VMs performance, as well as all other VMs on that physical server.

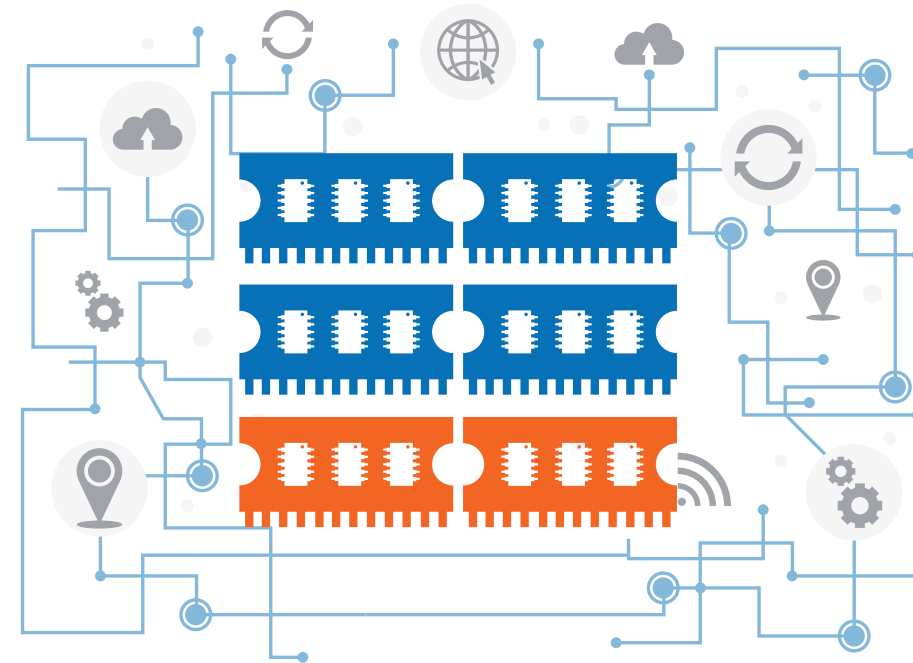


INTRODUCTION

PERFORMANCE IMPACT OF RAM OVER-PROVISIONING

By design, each instance of a “guest OS” will try to optimize its performance by holding on to as much memory as it sees available, which in virtualization means all the memory that is allocated to that VM.

The goal in optimizing performance across all VMs is to match how the amount of memory a VM is actually using (active memory) to what it’s been given. Allocate less memory than is used by its active memory, memory contentions arise and performance for that VM diminishes. Allocate more memory than is used by active memory, there will be less room for other VMs in that server, while overcommitting memory would cause swapping to disk which would hurt the performance of all VMs.



KEY TAKEAWAY:

Over-provisioning RAM results in the VM locking up more memory than is needed, and leads to inefficient operations and suboptimal overall performance.

INTRODUCTION

DEFINITION OF OVER-PROVISIONING

Memory Over-Provisioning

For this test we measure memory usage in each guest VM, and identify the average usage and the peak. By comparing those metrics with what's allocated, we are able to determine whether that specific virtual machine was over-allocated.

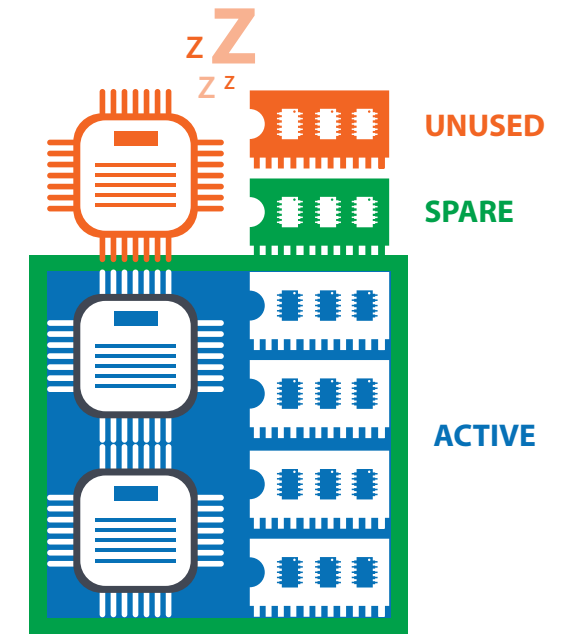
We define "ideal" memory provisioning as what's consumed, plus 1GB for VMs running Windows client operating systems, and 2GB for VMs running server applications. We categorize any memory allocated above that as over-provisioned.

CPU Over-Provisioning

We measure usage across all vCPUs in each guest VM, and calculate the average and peak. Based on those metrics, we are able to determine if vCPUs were over-allocated for that particular virtual machine.

Note, it is not uncommon for a VM's CPU to spike to 100% utilization, and providing more vCPUs will not necessarily lower that spike. In order to collect the true sustained CPU consumption, we discard the top 5% of CPU spikes for each instance. This is a common practice for measuring workload⁽⁵⁾.

To determine the "ideal" amount of CPU allocation, we add 15% to that VM's 95th percentile of CPU utilization. Finally, to determine whether and how much a server has been over-provisioned, we subtract the ideal CPU allocation from 100%.



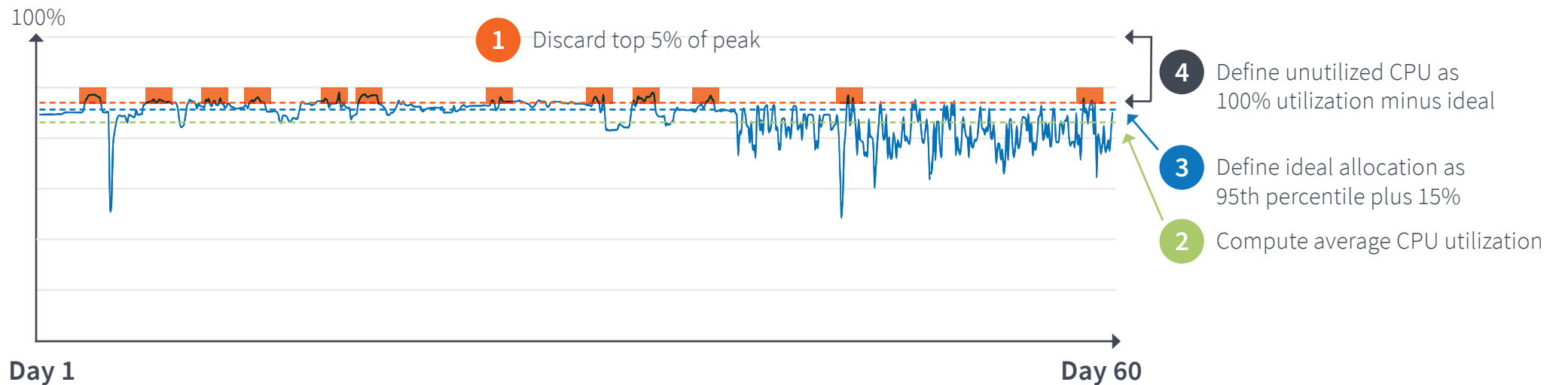
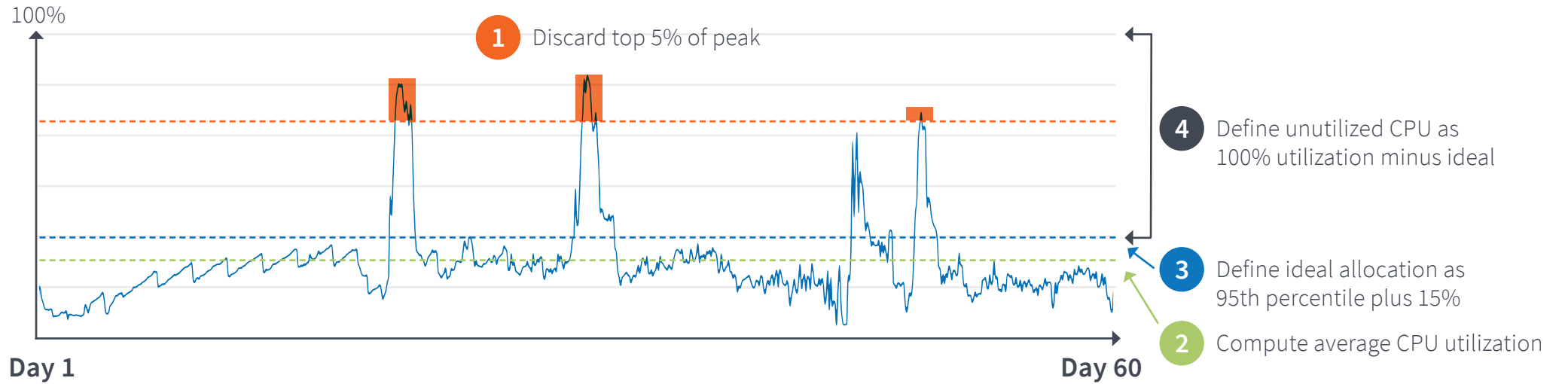
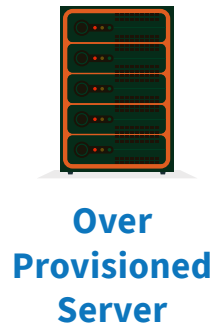
KEY TAKEAWAYS:

Ideal CPU = Actual CPU Consumed + 15% Spare

Ideal RAM = Actual RAM Consumed + 1GB Spare for desktop VMs (2GB for server VMs)

INTRODUCTION

CPU UTILIZATION



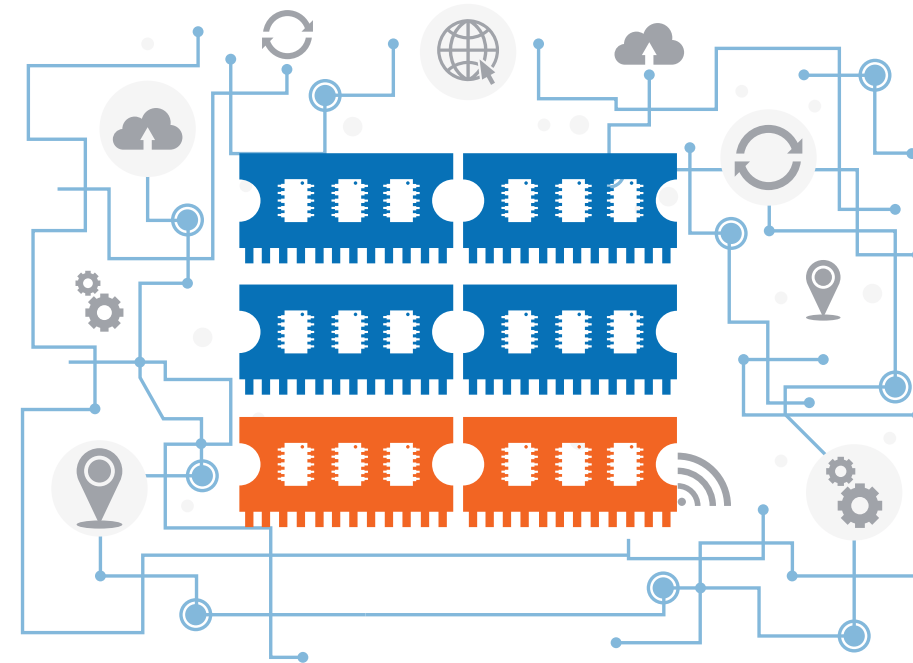
INTRODUCTION

DEFINITION OF MEMORY OVER-PROVISIONING

RAM Over-Provisioning

Similar to CPU, we measure memory usage in each guest VM, and identify low, average and peak. By comparing those metrics with what's allocated, we are able to determine whether there was over-allocation for that particular virtual machine.

We define "ideal" memory provisioning as what's consumed, plus 1GB for VMs running Windows, and 2GB for VMs running server applications. We define memory allocated more than that as over-provisioned.



KEY TAKEAWAYS:

Any allocated vCPU that is not used is considered to be over-provisioned. Unused RAM over 1GB for desktop VMs and 2 GB for server VMs, is considered to be over-provisioned.

REPORT METHODOLOGY

REPORT METHODOLOGY

DATASET OVERVIEW

ControlUp is an IT monitoring and management solution deployed in hundreds of organizations worldwide. For the purpose of historical analysis, benchmarking and troubleshooting, utilization is recorded along with other performance metrics in a global big data warehouse. The accumulated data permits us, with the consent of our

customers, to publish anonymized statistics and research findings based on large representative samples. To improve confidence in the data, we filtered out workloads for which we did not have sufficient volume of data, and ones with time lapses or partially collected data.

After filtering, this report is based on:



Workload utilization from 943 organizations



Data from mostly the US and EMEA, where our customer base is located



60 days of data



Total of 148,233 instances



KEY TAKEAWAY:

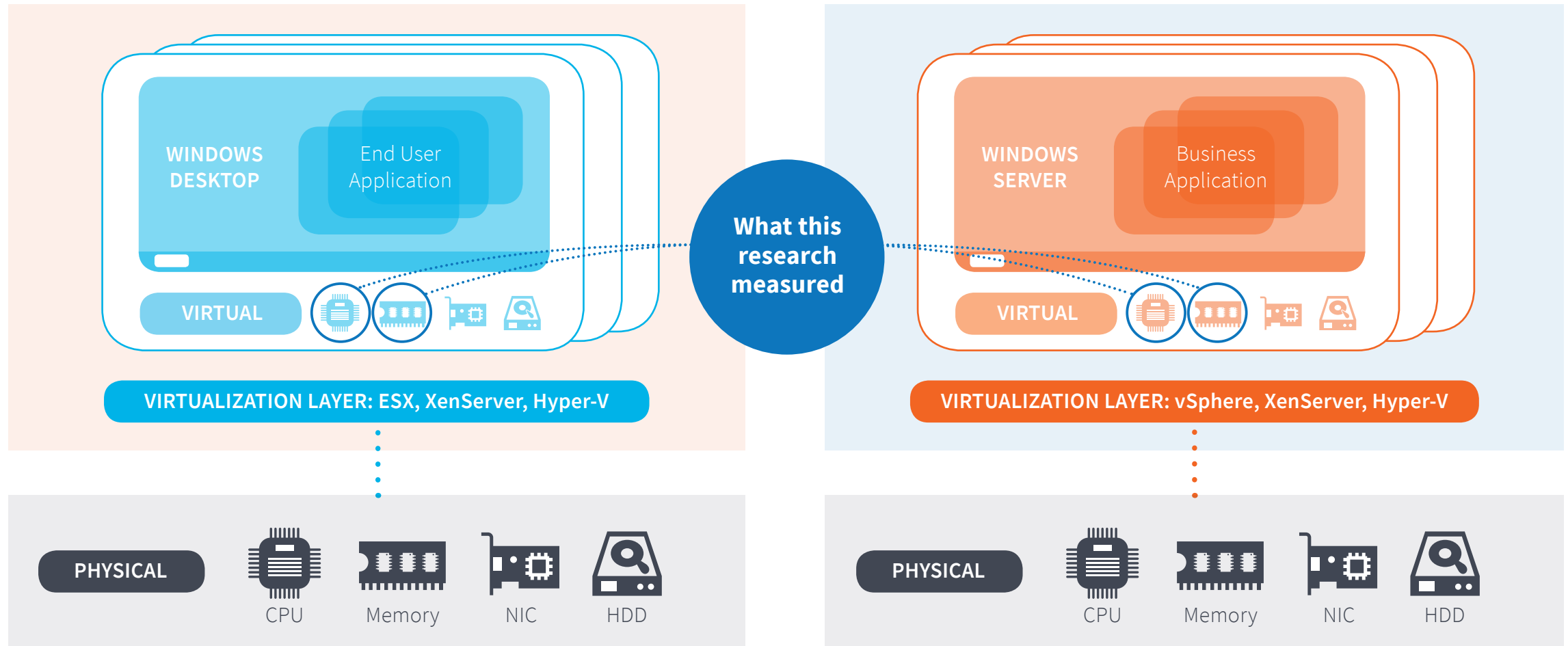
The sample presented in this report is statistically significant, and spans multiple industries, geographical regions and organization sizes.

REPORT METHODOLOGY

WHAT WE MEASURED

Desktop Workload

Server Workload



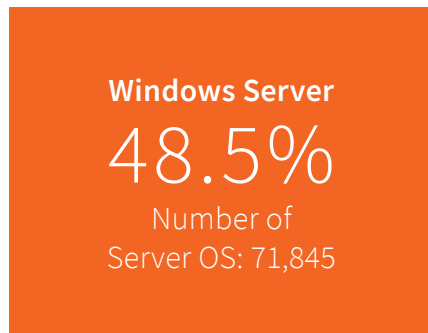
REPORT METHODOLOGY

DATASET'S OS DISTRIBUTION

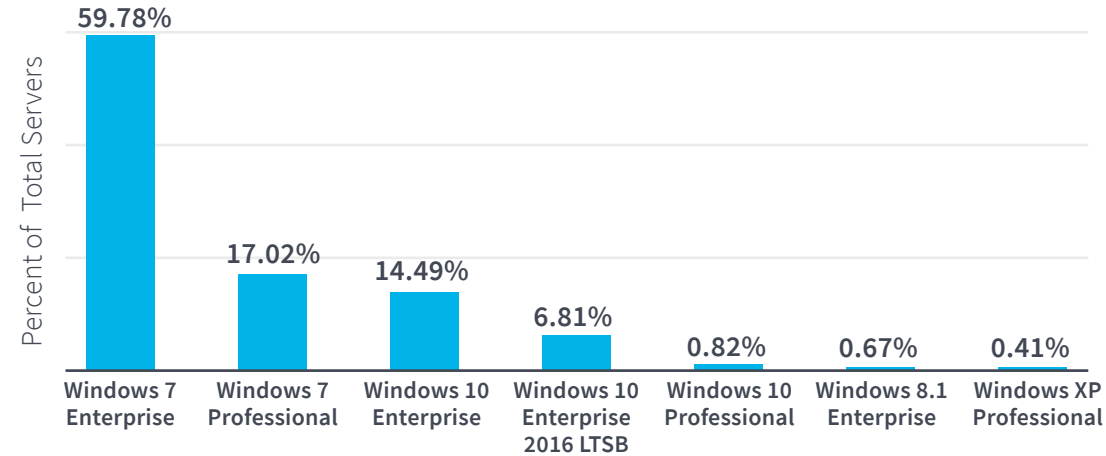
OS Distribution



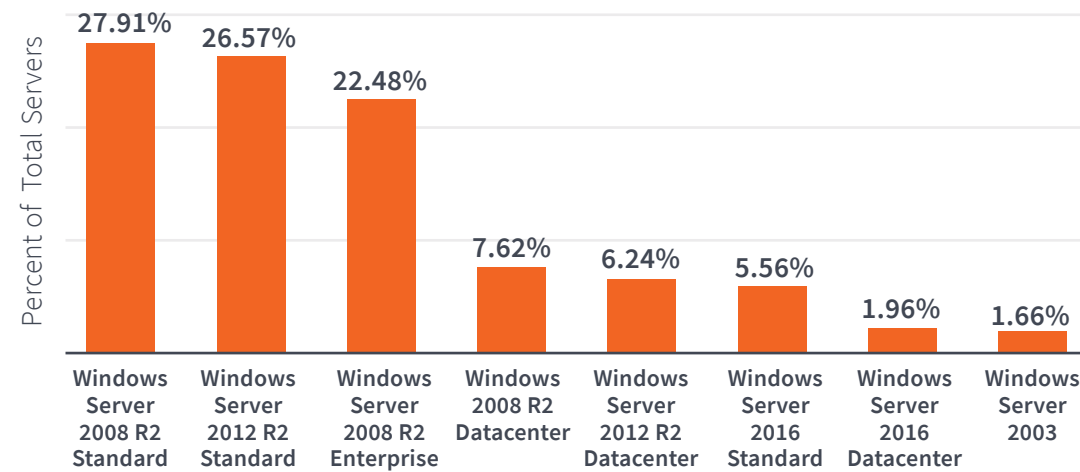
Total number of
instances: 148,233



Desktop OS Distribution



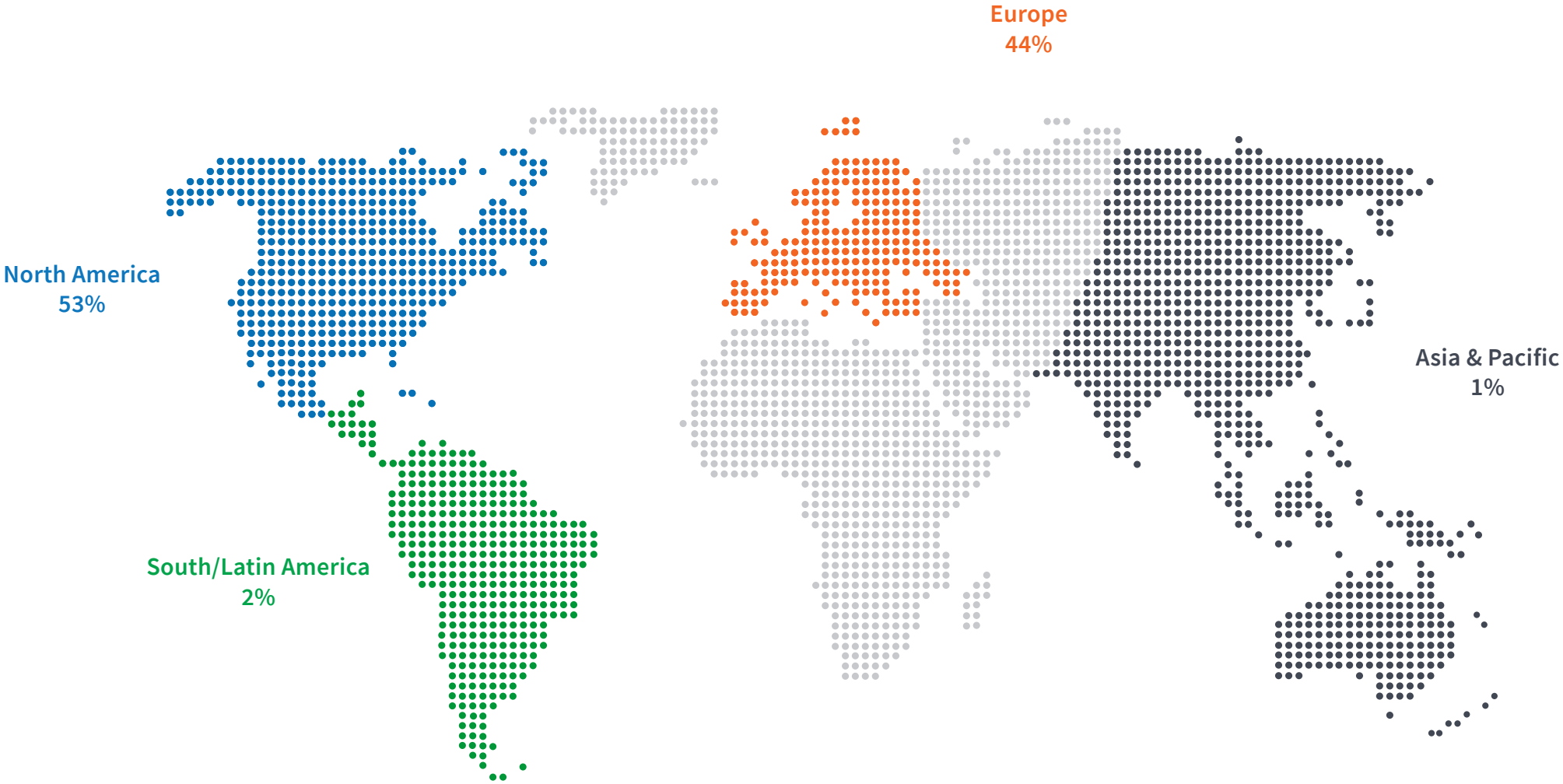
Server OS Distribution



KEY TAKEAWAY:
Both Windows desktop and Windows server use cases are well represented.

REPORT METHODOLOGY

DESCRIBING THE DATA SET - GEOGRAPHY

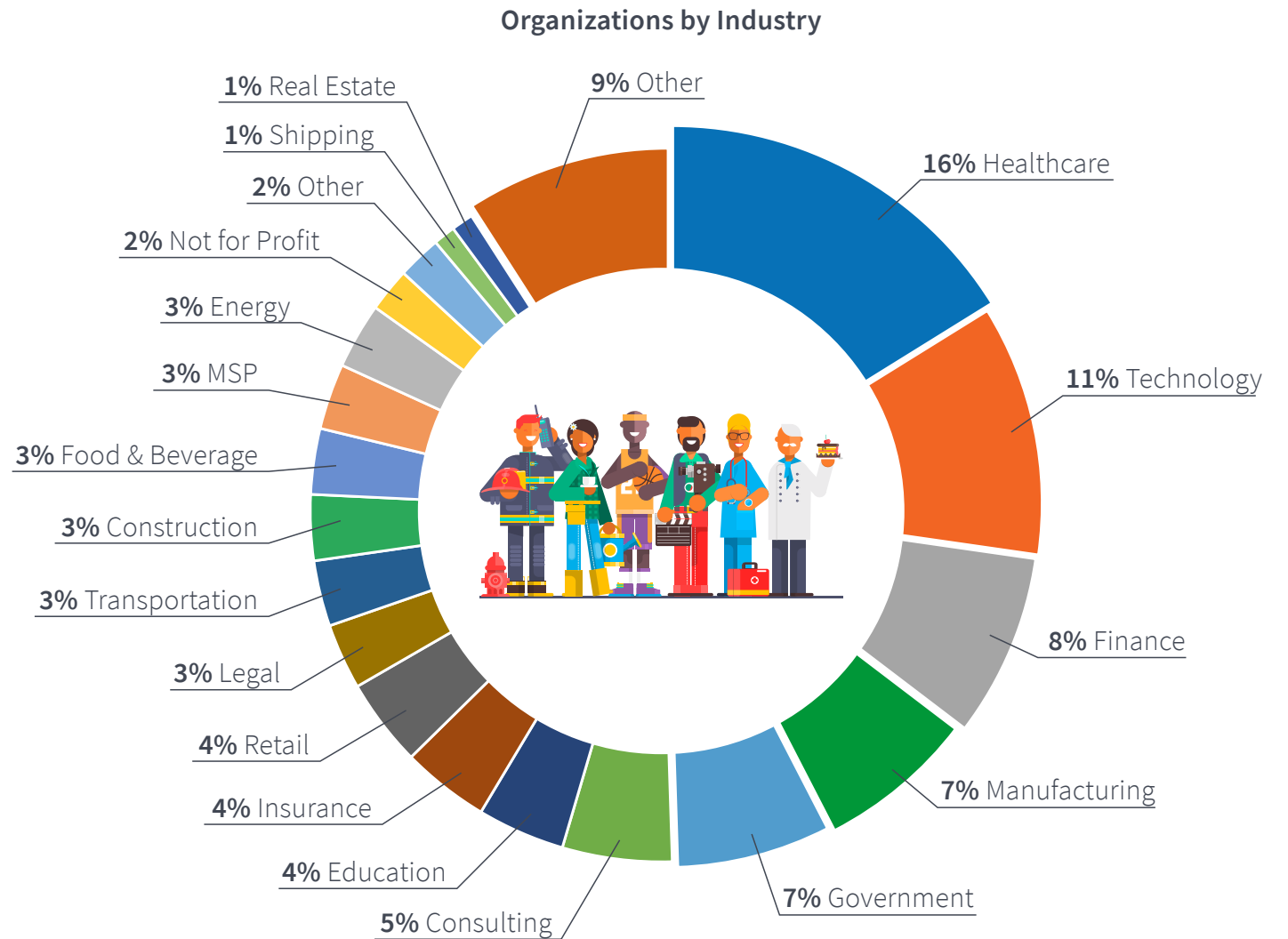


REPORT METHODOLOGY

DESCRIBING THE DATA SET - INDUSTRY

The organizations participating in the dataset were classified by industry, producing the distribution shown here.

The averages obtained in the popular industry categories will be compared to the grand average in the following section of this report.



FINDINGS

CPU PROVISIONING

FINDINGS - CPU PROVISIONING

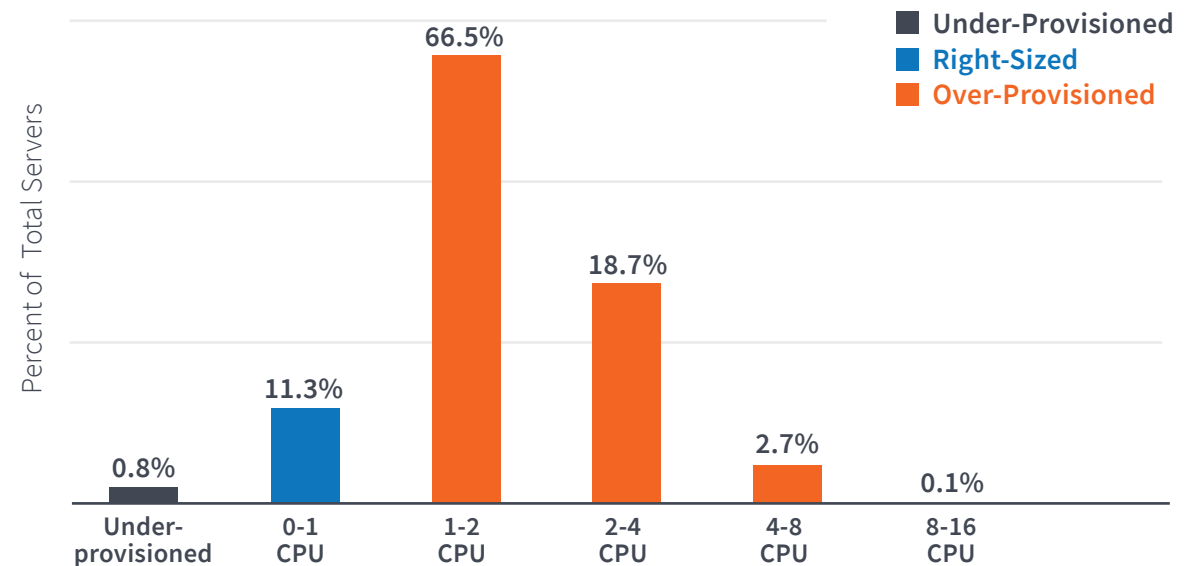
SPARE vCPU, DESKTOP OS

Our data reveals a very high amount of desktop OS vCPU over-provisioning - nearly 9 out of 10 VMs. Across all of those VMs, an average of two virtual CPUs sit mostly unused.

We found almost no under-provisioning. Out of the seventy-six thousand workloads less than 500 instances would have benefited from additional vCPU allocation.

Only 1 in 10 desktop VMs were appropriately allocated with processor resources.

Min 1 vCPU	Max 24 vCPU	Avg 2.5 vCPU
Median 2 vCPU	Stdev 3.9 vCPU	Spare 1.78 vCPU
n: 76,388		



KEY TAKEAWAY:

88% of Desktops OS are over-provisioned with more than 1 spare vCPU.
Average amount of spare vCPUs per desktop OS is 1.8 vCPUs.

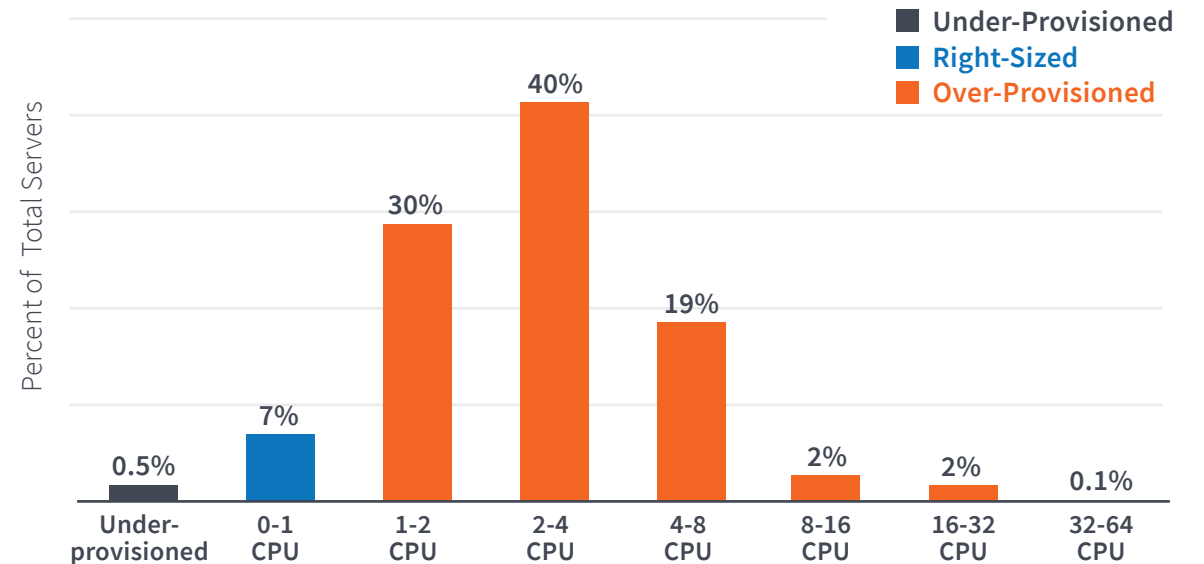
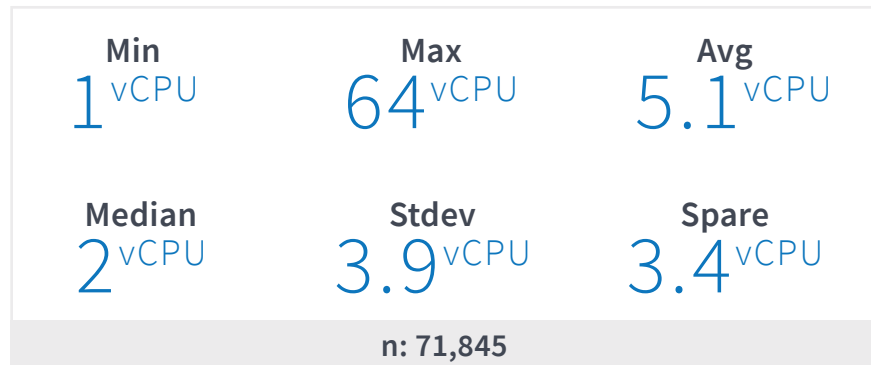
FINDINGS - CPU PROVISIONING

SPARE vCPU, SERVER OS

Our data shows an even higher level of vCPU over-provisioning with server VMs -- 93%. On the average between two and four virtual CPUs sit idle on those VMs.

We also found virtually no under-provisioning. Out of the seventy-one thousand servers, less than 500 were under-provisioned.

Lastly, only 7% server VMs were appropriately allocated with processor resources.



KEY TAKEAWAY:

93% of server OS are over-provisioned with more than 1 spare vCPU.
Average amount of spare vCPUs per server OS is 3.4 vCPUs.

FINDINGS

MEMORY PROVISIONING

FINDINGS - MEMORY PROVISIONING

SPARE RAM, DESKTOP OS

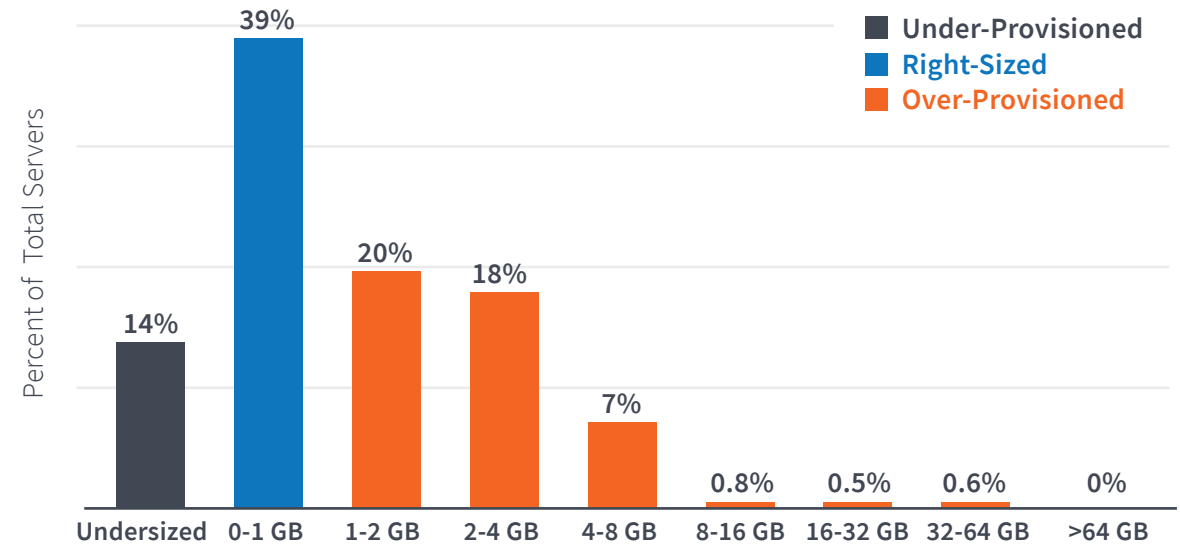
Given the difficulty in accurately predicting RAM and consequence of under-allocating memory, it's not a surprise that we found significant over-provisioning when it comes to memory.

Almost half of seventy-six thousand desktop VMs were given too much RAM, and 39% were properly allocated.

Even more interesting, we found that 14% of desktop OSs need more RAM and are almost certainly experiencing performance degradation.

Min 1 ^{GB}	Max 64 ^{GB}	Avg 5.8 ^{GB}
Median 8 ^{GB}	Stdev 3.2 ^{GB}	Spare 1.93 ^{GB}
n: 76,388		

GB of Over-Provisioned RAM



KEY TAKEAWAY:

47% of desktops OS are over provisioned with more than 1GB of spare RAM.
The average amount of spare RAM per desktop OS is 1.9GB

FINDINGS - MEMORY PROVISIONING

SPARE RAM, SERVER OS

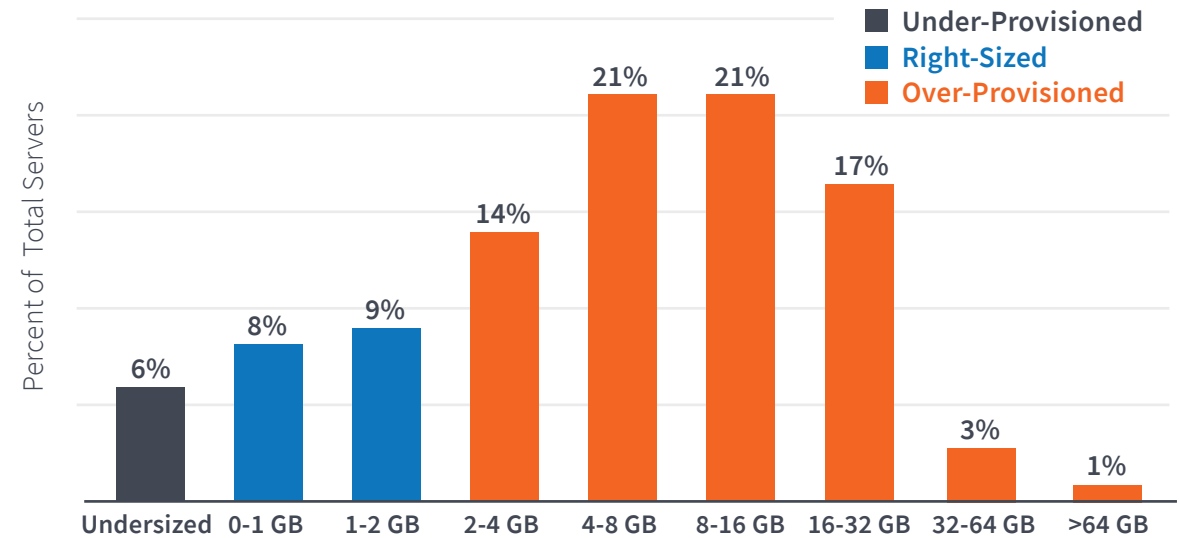
One can make the case that a server VM running out of memory is generally more damaging than when a desktop VM runs out of memory. It's the difference between a cluster of users being impacted vs. an organization's mission critical applications.

Perhaps that is why we see an even more dramatic over-provisioning of server OS machines - 4 times as much as desktop VMs.

All of these sites have ControlUp, they are seeing these statistics, and yet they're not taking action. This may be due to the challenges of working cross-departmentally, which we discuss later.

Min	Max	Avg
1 ^{GB}	1024 ^{GB}	26.2 ^{GB}
Median	Stdev	Spare
8 ^{GB}	14.4 ^{GB}	11.7 ^{GB}
n: 71,845		

GB of Over-Provisioned RAM



KEY TAKEAWAY:

77% of server OS are over provisioned with more than 2GB of spare RAM.
The average amount of spare RAM per server OS is 11.7GB

FINDINGS PROVISIONING BY OS VERSIONS

FINDINGS

PROVISIONING BY DESKTOP OS VERSIONS

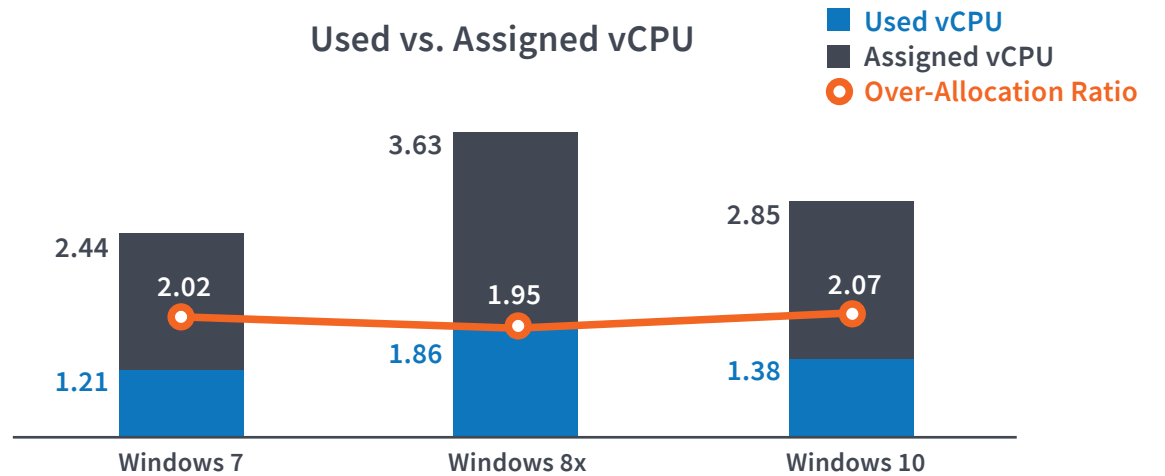
Our research extended to examining resource provisioned and usage levels per OS, and generation of OS. There is an industry perception that newer OSs need more resources for the same workload.

Our study shows that may have been the case over the years, but the pattern was definitely changed with the latest versions of OSs. We found that Windows 10 OS family is more efficient than the previous generation (Windows 8.X) and is practically as efficient as Windows 7.

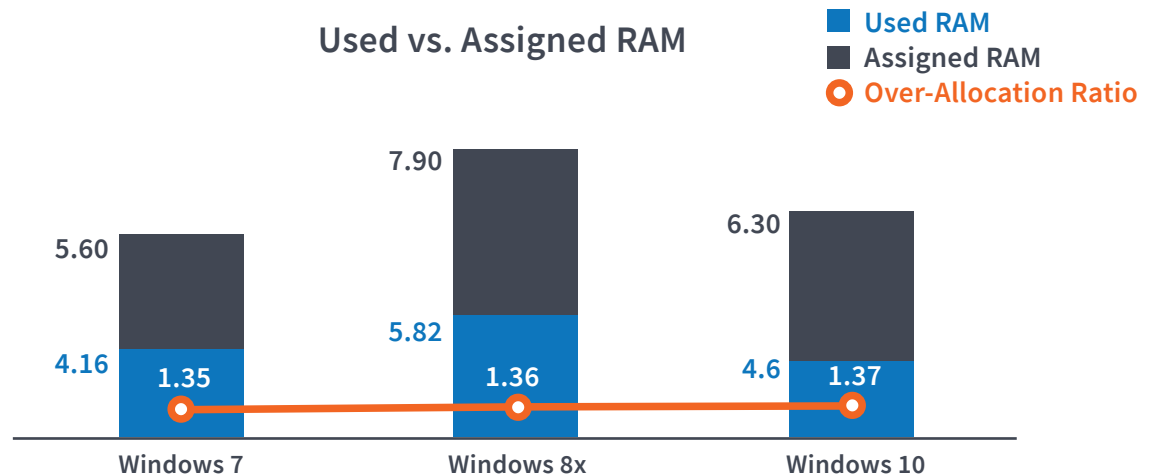
KEY TAKEAWAY:

One should not assume that a newer desktop OS would need, or better utilize, additional resources when compared to older versions. The level of vCPU and RAM over-provisioning by sysadmins was virtually the same across Windows 7, 8, and 10.

Used vs. Assigned vCPU



Used vs. Assigned RAM



FINDINGS

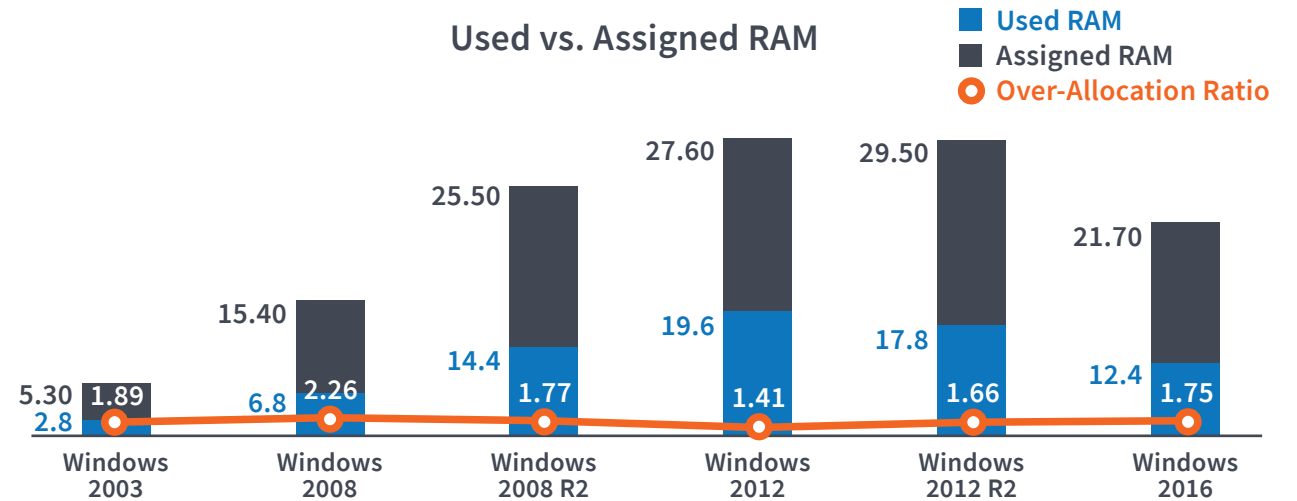
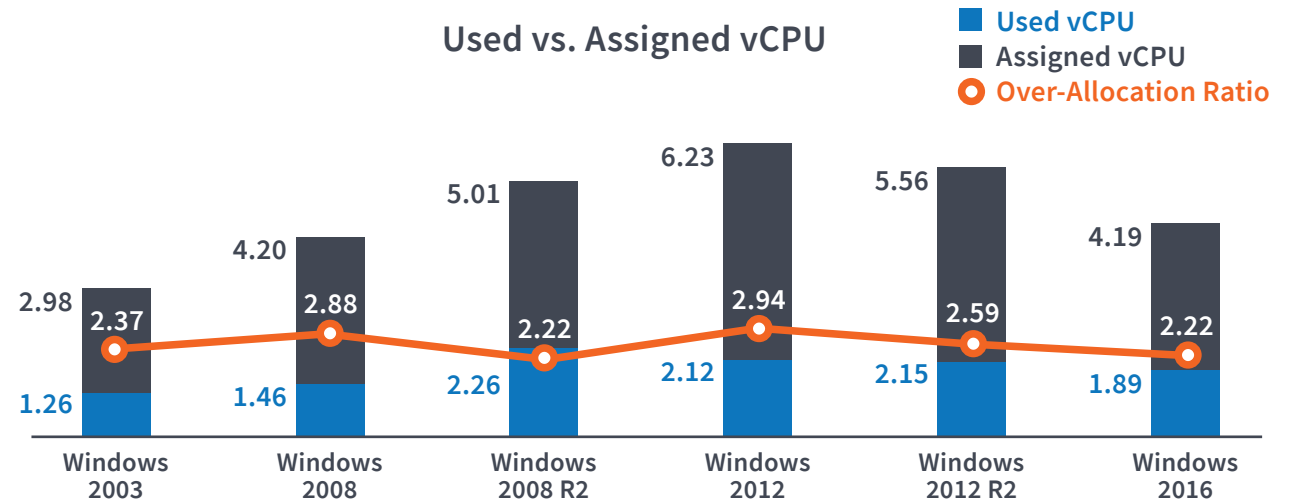
PROVISIONING BY SERVER OS VERSIONS

We found an identical story for server OS versions.

While there was a trend towards each new version of Windows server required more resources than the previous, that trend has definitely reversed over the past two major releases; Windows Server 2012 R2 and Windows 2016.

KEY TAKEAWAY:

Again, we should not assume that a newer server OS would need, or better utilize, additional resources. We also found that levels of vCPU and RAM over-provisioning was virtually the same across all versions of Windows servers.



FINANCIAL IMPLICATIONS


FINANCIAL IMPLICATIONS

FINANCIAL IMPACT OF MEMORY OVER-PROVISIONING

Estimating the financial impact of CPU over-provisioning is somewhat complicated given the continuous sharing of CPUs and cores, but computing cost implications of memory allocation is simple math.

Here we apply our findings' memory over-allocation percentages to the entire dataset we studied. Given the very large size of our dataset, it is safe to extrapolate the same cost implications across the entire industry. Similarly, you can plug in metrics from your monitoring tool into this table and see your organization's situation.

	Servers in this dataset	Avg. RAM Allocated/ Server	Total Spent	% servers over-provisioned	Over-provisioned per server	MSRP Cost ⁽⁶⁾	Total Overspent on RAM
RAM - Desktop OS	76,388	5.8 GB	\$16.3 million	47%	1.93 GB	\$36.80 / GB	\$2.55 million
RAM - Server OS	71,845	26.2 GB	\$69.3 million	77%	11.7 GB	\$36.80 / GB	\$23.82 million



KEY TAKEAWAYS:
The average cost for memory over-allocation per server was \$117.9, costing a total of \$26.4 million across the 148,233 servers. Over 1 out of 3 dollars spent on RAM was wasted.

DISCUSSION

DISCUSSION

THE DIFFICULT NEXT STEPS

Be aware of political and face-saving situations approaching this project. While sysadmins have dominion over their servers, resource provisioning can get messy in certain organizations and situations. For example, it is simpler to right-size resources on an end user computing environment than it is an “it ain’t broke” mission-critical application running at high capacity. The latter often has application owners whose primary goal is stability and performance of his/her application, and they often come to the operations with a budget for high resource levels. They have as much interest in overall data center efficiency as a diner has in a restaurant’s overall efficiency. So what to do?

Step 1 - Involve the application owner on performance-focused monitoring

For application owners, cost is much less of a priority as is performance. Approach the application owner with a performance optimization process where all you’re doing is monitoring to see if there’s room for performance improvement. Sometimes the application owner simply follows their vendors’ recommendation for sizing, which may not be optimal in your environment. You may learn that the resource specifications are for a physical environment, not a virtual one.

The important part of this step is to involve other parties because they will become your co-owners and your champion in this endeavor.

Step 2 - Monitor

To make a preliminary assessment of over-provisioning, you must first decide on a buffer or spare capacity for workloads that would fit the needs of your organization. The answers may vary and will be biased towards your subjective definition of over-provisioning. You may want to assign different answers for different categories’ workloads based on the impact of over-provisioning.

We are now ready to start gathering data using a monitoring tool for assessing a workload’s CPU and memory utilization. Many organizations have cyclical work patterns, so it may make sense to continuously monitor for at least a month in order to obtain a sample that represents usage peaks/spikes.

Lastly, make sure your monitoring tool is built for this purpose. ControlUp’s capabilities for monitoring provisioning have a proven track record in the industry. While there are various measurement tools that can provide you with an over-provisioning report, ControlUp provides an easy and accurate way to obtain accurate in-guest measurements and metrics. An in-guest view is as critical for identifying appropriate provisioning as an x-ray is for an orthopedist.

DISCUSSION

THE DIFFICULT NEXT STEPS (CONT.)

Step 3 - Analyze the data

Is there a processor wait time and low utilization? The best benchmark to measure against other subgroups of servers in your own environment. Historical analysis of granular data over a period of time is key here to ensure you're factoring in usage spikes. You must convince the application owner that right sizing will improve the application's performance.

Signs of sub-optimal vCPU allocation is high in-guest CPU usage, combined with high vCPU allocation and a high CPU Ready time for the VM. CPU Ready time is the time a VM has to wait until its request for its specific number of vCPUs to become available on the physical CPU. In our restaurant analogy, CPU Ready time is the amount of time each restaurant patron waits for their table.

Step 4 - Adjust Provisioning

Start with CPU (vs. RAM) and start with the worst cases first. Look at the highest processor and memory provisioned VMs because that's where you may see the most improvement.

In the case of a new application being rolled out, it may make sense to err on the lower side of vCPU allocation, monitor and measure, and turn up processor allocation until performance gain begins to diminish.

As the final step you may engage the head of the products division responsible for all applications, or even someone higher, like the head of IT.

Step 5 - Institutionalize Right-Sizing

As mentioned earlier, a virtualized data center is a dynamic environment where workload behavior and levels change over time. This research clearly verifies that claim. It is important to establish a procedure for revisiting this issue on a regular cadence, even if the last item on the checklist says "We decided not to review this time".

KEY TAKEAWAY:

Right-sizing a VM's resources is not a trivial undertaking. However, with the right data and the right approach, an organization has much to gain in cost savings and performance improvements.

REFERENCES

1. <https://lelunha.wordpress.com/2012/07/19/140/>
2. <https://pcpartpicker.com/trends/price/memory/>
3. <https://pcpartpicker.com/trends/price/cpu/>
4. <https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/vmware-vsphere-cpu-sched-performance-white-paper.pdf>
5. <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/44279.pdf>
6. <https://pcpartpicker.com/trends/price/memory/>

CONTACT US

We can help you measure, monitor and maximize logon performance, and optimize resource utilization.

It literally takes 15-minutes to install ControlUp in your environment.

or