CLOUDPHYSICS

Global IT Data Lake Report; Q4 2016

Executive Summary

CloudPhysics' SaaS delivery model enables the creation of the largest IT Data Lake in the industry. In this report, we offer insights about global trends and changes that are occurring in the world of data center IT. With several key findings, the overarching theme of the report focuses on an abundance of idle and wasted compute and storage resources, suggesting a lack of optimization in data centers globally.

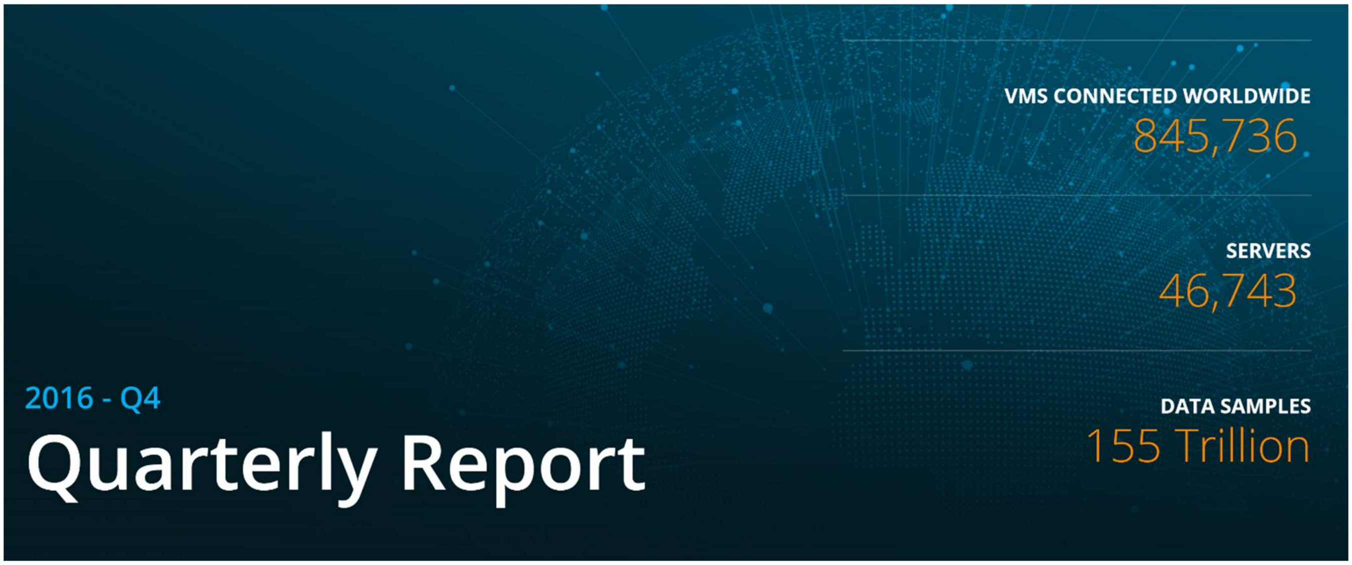
The report is organized into four sections:

- Cloud Migration - How to Accurately Estimate Your Costs
- Storage I/O - Do I Really Need Flash? How Should I Use It?
- What does the use of CPU & Memory Resources look like today?
- An Overview of the CloudPhysics Global IT Data Lake

Select key findings contained in this report include:

- The cumulative financial impact of rightsizing during migration would net to $54.7B annually – more than the GDP of 114 nations
- More than 50 percent of the world's allocated virtual central processing units (vCPUs) could be reclaimed
- 90 percent of the virtual machines (VMs) from CloudPhysics' global dataset average less than 15 percent utilization of the vCPU allocated to them
- 91 percent of VMs globally average 10 input/output per second (IOPS) or fewer.
- 80 percent of an organization's IOPS is used to access only 20 percent of their storage.
- 35 percent of organizations in the CloudPhysics global dataset have meaningful degrees of I/O contentions daily, with 5 percent of their VMs experiencing contentions every day.

What is the IT Data Lake, and How Does CloudPhysics Create It?



The CloudPhysics platform is delivered to customers and partners through a Software as a Service (SaaS) form factor. Following a simple, five-minute vApp installation in the customer data center, machine metadata is transmitted to the CloudPhysics cloud-based data platform, from anywhere in the world.

This approach offers several advantages to our user community, including ease of deployment, ease of management, Internet-wide accessibility, and the ability to view the entire organization through a single pane of glass, even if the organization is globally distributed.

In addition, the CloudPhysics SaaS form factor enables us to aggregate data across our customers' data centers — after first anonymizing it, of course — and run global queries across the IT industry's largest Configuration Management Database (CMDB), surfacing insights about global trends and changes that are occurring in the world of data center IT.

In this Report — the first of many from CloudPhysics — we've asked the Global IT Data Lake a series of questions, all revolving around the query, "has virtualization lived up to its promise?" We then analyzed the answers and documented them for you to enjoy!

Cloud Migration - How to Accurately Estimate Your Costs

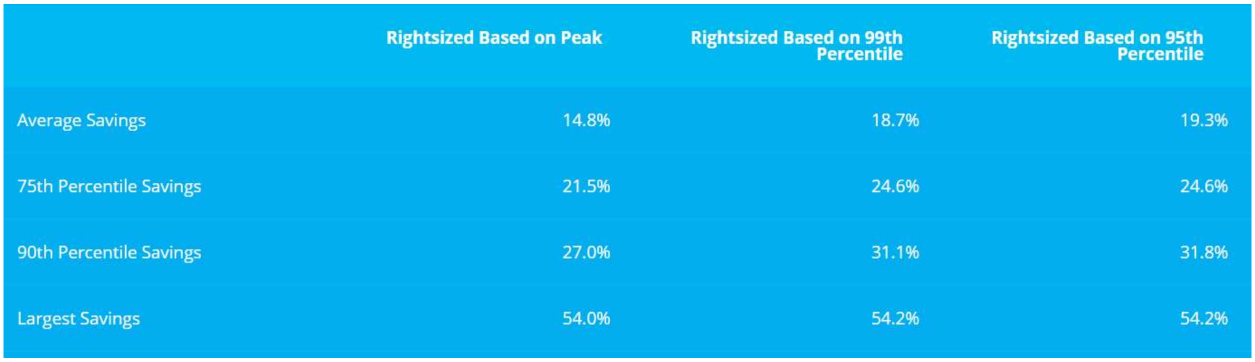
As you will see in our "CPU & Memory Resources" section, most of the world's VMs are provisioned with more resources than they need to service their workloads — sometimes extensively so.

While compute over-provisioning can certainly result in waste in the private cloud data center — waste that can be reclaimed through rightsizing — there is another, more insidious way in which this over-provisioning can result in bad decision-making. This happens when organizations evaluate public clouds for possible workload migration.

On-premises, an over-provisioned VM may be hidden. This is because many system administrators manage resources at the cluster level, rather than focusing intently on the configuration of each and every VM. If an administrator is managing physical resources at the cluster level, then that administrator will continue to load VMs into the cluster until the overall cluster resources begin to be exhausted. Since the focus is on resource utilization, rather than resource configuration, the fact that VMs are over-provisioned can go unnoticed while the workload is on-premises.

However, this is not the case when estimating the cost of running a currently on-premises VM in a public cloud. If an administrator uses the configuration of a VM as a basis for cloud costing — and that VM's configuration is inflated due to over-provisioning — then the estimated costs of running in the cloud will be inflated; sometimes massively so.

How much? We asked the Global IT Data Lake.



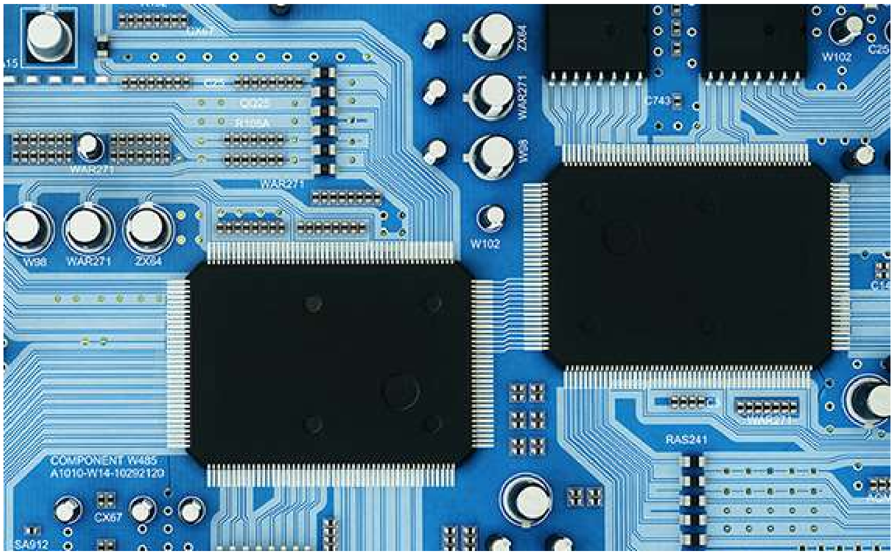
	Rightsized Based on Peak	Rightsized Based on 99th Percentile	Rightsized Based on 95th Percentile
Average Savings	14.8%	18.7%	19.3%
75th Percentile Savings	21.5%	24.6%	24.6%
90th Percentile Savings	27.0%	31.1%	31.8%
Largest Savings	54.0%	54.2%	54.2%

Table A: Spectrum of Potential Savings When Rightsizing before Migrating to Public Cloud

See Table A. The figures in the table represent the potential per-organization savings, if they use data to identify opportunities to rightsize their VMs prior to migrating them to a public cloud. In other words, if you look at the table — if we rightsized based on peak utilization of vCPU — an average customer would shave 14.8% from public cloud cost estimates, while a customer in the 90th percentile would save 27%. In our Global IT Data Lake, we found numerous examples of customers who could shave more than half off their public cloud deployment costs if they rightsized during migration.

FUN FACT: based on various data points within the Global IT Data Lake, and an estimate of 90,000,000 workloads globally, the cumulative financial impact of rightsizing during migration would net to $54.7BN annually — more than the GDP of 114 nations in the world!

Storage I/O - Do I Really Need Flash? How Should I Use It?



All-flash storage has been broadly hailed as a cure-all for Enterprise workload performance woes. It has been taken for granted that Enterprise applications will gain enough benefits from aggressive solid-state disk deployments to justify the additional costs and migrations. Key questions remain, however. How broadly should all-flash storage be deployed? What types of workloads would benefit enough to justify the cost?

We asked the CloudPhysics Global IT Data Lake some questions on this topic: read on for the answers!

VM I/O Load – You are using less I/O than you think!

To begin with, we asked the Global IT Data Lake a simple question: "How much I/O activity do the world's VMs actually generate?"

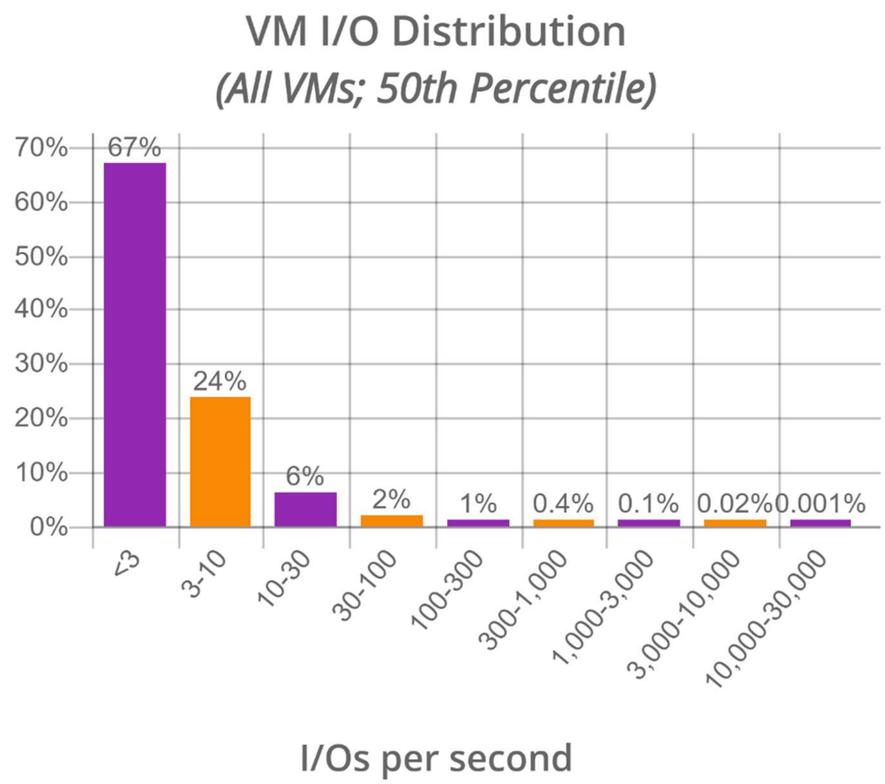


Chart A: VM I/O Distribution; All VMs in the Global IT Data Lake; Median Values over a 7-Day Period

See Chart A. This chart depicts the median I/O load for VMs in the Global IT Data Lake over a period of seven days. (Note that we aggregated vDisks for VMs with multiple virtual disks; note too that this chart might be better depicted using logarithmic scale. However, we kept it using absolute scale for comparison to Chart B.) This chart tells us that a staggering 91% of VMs globally average 10 I/O per second or fewer across a seven day period. That is very low.

Granted, viewing the median I/O characteristics may be of limited value for capacity planning; for that, we would want to look at I/O statistics at peak or near-peak. However, it is still instructive to note that the vast majority of global VMs generate very low I/O loads on average. There is clearly excess capacity available throughout the day.

Now, let’s look at a value closer to peak I/O load, and see what the Global IT Data Lake tells us. While we’re at it, let’s also filter to focus on larger, “beefier” VMs. The theory here is that large VMs, examined at peak load, should demonstrate much larger I/O characteristics than all VMs on average.

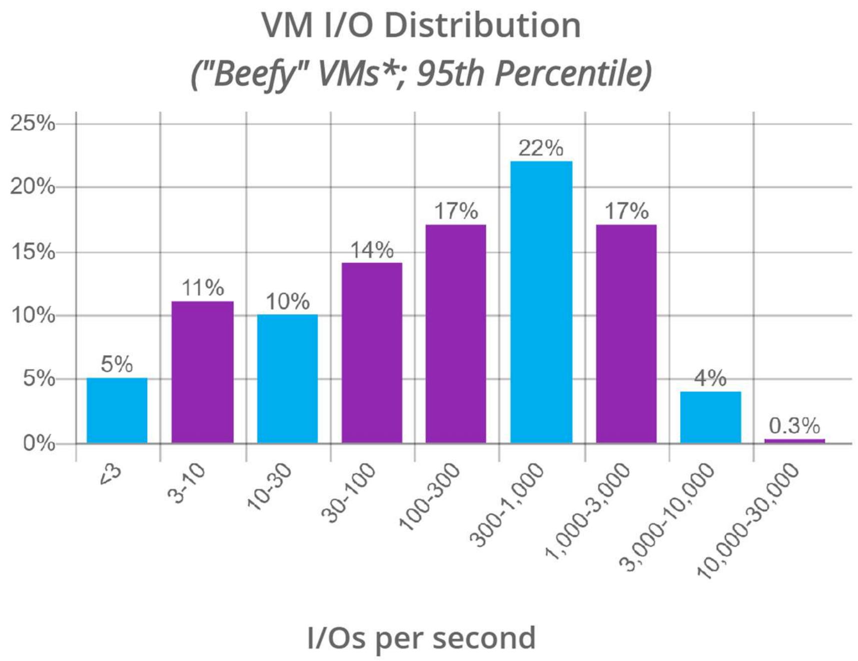


Chart B: VM I/O Distribution; “Beefy” VMs; 95th Percentile Values over a 7-Day Period

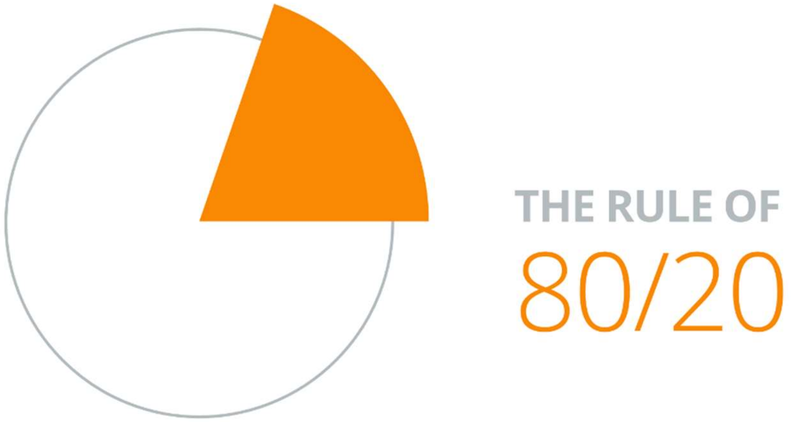
** At least 4vCPU, 8GBs RAM, and 1TB disk*

See Chart B. This chart depicts the 95th percentile I/O load for only “beefy” VMs in the Global IT Data Lake over a period of seven days. (“Beefy” VMs are provisioned with at least 4vCPU, 8GBs RAM, and 1TB disk.) As anticipated, these data show us a greater distribution of I/O per second, with 60% seeing > 100 I/O per second at close to peak load. However, 100 I/O per second is not a lot, and only 22% of beefy VMs experienced over 1000 I/O per second at the 95th Percentile.

Granted, all flash arrays improve storage read response, irrespective of I/O load, but it stands to reason that VMs which average a very low I/O load would see small benefit from all-flash storage.

The takeaway here is that you should not assume that your VMs have I/O characteristics that require additional assistance; your VMs are probably not working as hard as you think! It is vital to use fine-grained data to make your decisions about where to deploy all-flash in the data center to derive the most benefit at the lowest cost.

All Workloads Are Not Created Equal – 80% of your IOPS are used to access only 20% of your footprint.



Now we know that VMs generate a lower I/O load than might be anticipated. We also have a hint that the I/O loads are skewed throughout the data center, with some VMs generating more than others. Let's dig a little deeper at how VMs consume I/Os across an organization. Unsurprisingly, the Global IT Data Lake suggests that I/O distribution is not uniform, and VMs consume more than others. Those workloads with the highest IOPS will likely benefit most from caching.

But how skewed is the work at most enterprises?

It turns out that average organizations have a great deal of workload skew; in fact, 80% of IOPS is used to access only 20% of storage footprint in the average CloudPhysics organization.

It gets even more concentrated than that! In the average organization, 90% of IOPS is used by only 45% of storage footprint. This means that over half of your storage accounts for less than 10% of your IOPS!

This workload skew supports the idea of all-flash storage, but it underscores the need to deploy it judiciously in a data-driven process, using tools that can analyze workloads at the finest time granularity.

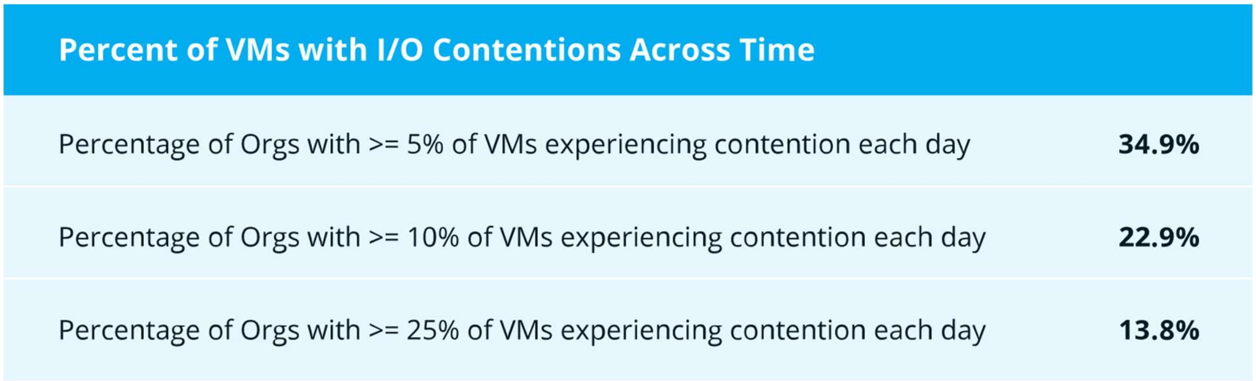
Bad Neighbors – I/O Contentions can invisibly slow workload performance.

So far in this report, we've examined the I/O distribution among VMs in the Global IT Data Lake. We then looked at the way VM I/O generation is skewed within Enterprises, with 80% of IOPs coming from 20% of the storage. However, given the importance of storage I/O capacity to app performance (more than 90% of performance problems are related to storage contention), we need to dig deeper.

Virtualization workloads are constantly moving and changing, which makes I/O-based performance issues difficult to track and pin down. Despite being armed with the management and optimization tools available today, it can take even the most experienced data center administrator hours or days to determine when and how I/O contention has impacted app performance.

Good news! CloudPhysics' data scientists have spent countless hours analyzing storage contention, providing unprecedented visibility into disk I/O workloads, and automating the discovery of "culprit" and "victim" VMs. This allows us to show our users which VMs are impacted by disk I/O issues in a single click, and which are causing the problems.

It also allows us to look across the globe to determine just how prevalent I/O contention is between the world's virtual machines.



Percent of VMs with I/O Contentions Across Time	
Percentage of Orgs with >= 5% of VMs experiencing contention each day	**34.9%**
Percentage of Orgs with >= 10% of VMs experiencing contention each day	**22.9%**
Percentage of Orgs with >= 25% of VMs experiencing contention each day	**13.8%**

Table B: VM I/O Contentions Rate among Organizations

It turns out that far more organizations experience I/O contention events than we anticipated (especially given that the Global IT Data Lake shows that VMs average fewer I/O per second than we expected). Nearly 35% of organizations in the CloudPhysics Global IT Data Lake have some degree of contention, with 5% of their VMs experiencing contentions every day. Some organizations experience a great deal of I/O contention; nearly 14% of CloudPhysics organizations had over one-quarter of their VMs experiencing contentions each day!

Clearly, many VMs experience I/O starvation periods every day. This discovery also supports the idea of all-flash storage, but it further underscores the need to utilize tools that are able to identify where its deployment will most benefit the enterprise.

The Net Effect – How does all of this impact my latency?

The Global IT Data Lake tells us that all-flash storage — and the I/O improvements that can stem from its use — can certainly deliver benefit if deployed intelligently, with the benefit of data. Ultimately, this all-flash story has a final conclusion, and that conclusion is around latency. If your storage infrastructure is inadequately responsive to the I/O needs of your workloads, your infrastructure will deliver poor read latency, and your application users will experience poor performance. That is a bad thing.

We can, however, ask the Global IT Data Lake to help us measure the read latency of the world's VM workloads, if we layer on some sophisticated analysis.

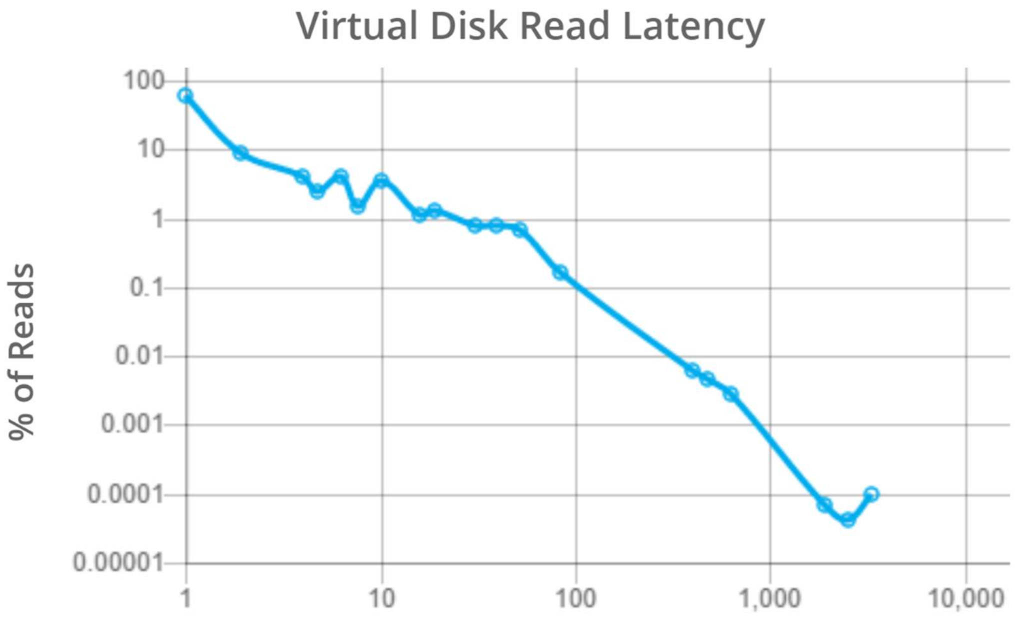


Chart C: Global Read Latency for VMs

See Chart C. Chart C depicts the distribution of read request latency for all VMs in the Global IT Data Lake over a week in October of 2016. The X axis is Read Latency in milliseconds, the Y axis is the percentage of reads that experienced that latency during the week. Note that the scale is logarithmic; otherwise, the line would appear flat after approximately 10 ms of latency.

A couple of observations:

- The vast majority of the world's read requests have latency that would be imperceptible to application users; 92.1% of reads experience latency of 2ms or less, and 96.9% of reads experience sub-10-millisecond latency.

- However, at the extreme, latencies can be very perceptible and meaningful to application users; 0.17% of workloads experienced greater than 100MS of latency.

So this analysis reinforces the Workload Skew analysis. Here again we see that while world's infrastructure as a whole is experiencing low latency and small I/O load, there are VMs on the edges that are providing demonstrably poor performance to users. Again, using a data-driven approach to all-flash deployment can maximize its value at manageable cost.

To ensure that we did not select an anomalous week for our analysis, we compared many billions of latency curves spanning all the weeks in October. This can be seen in the graph below, reflected in consistencies in the curves:

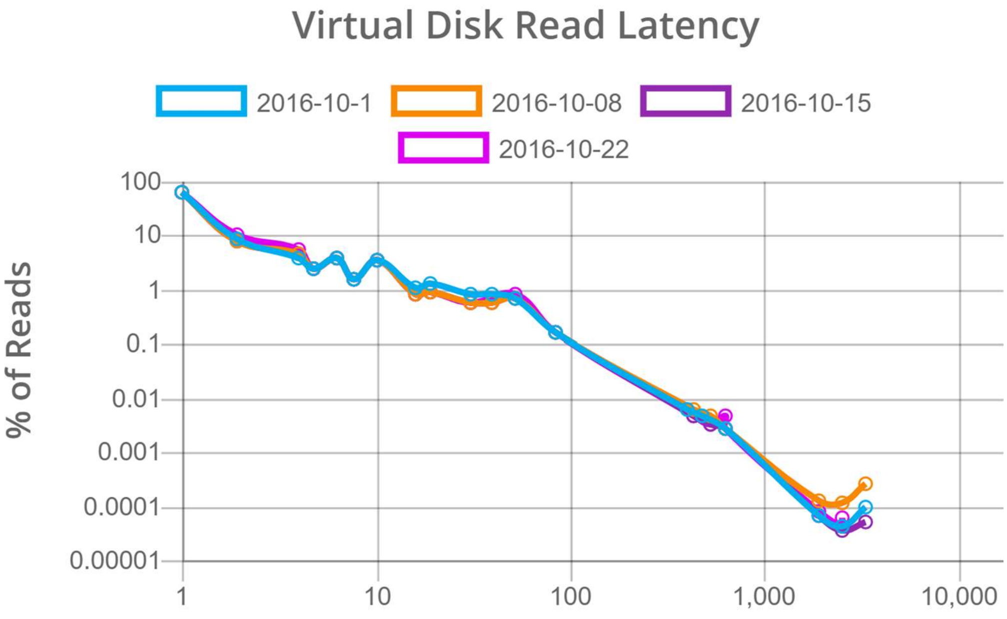
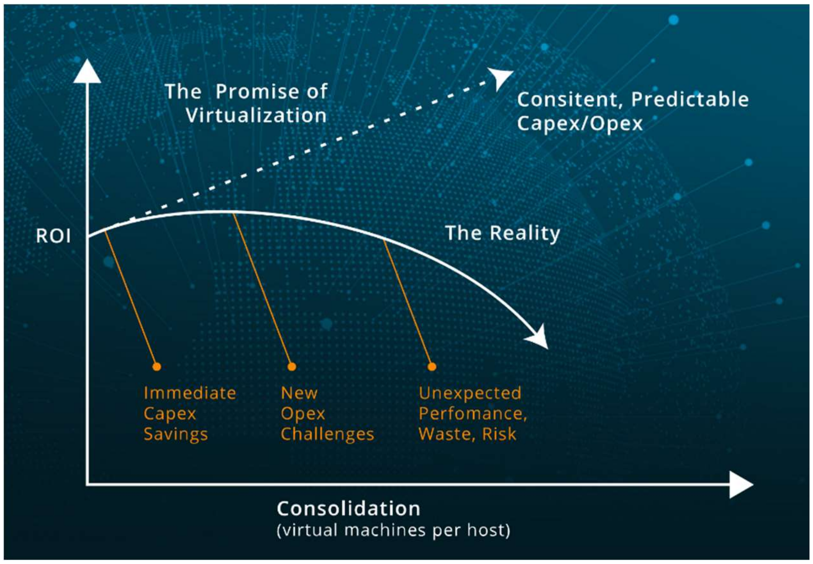


Chart D: Virtual Disk Read Latency Distribution, Multiple Weeks

What does the use of CPU & Memory Resources look like today?


The Promise of Virtualization
Consitent, Predictable Capex/Opex
ROI
The Reality
Immediate Capex Savings
New Opex Challenges
Unexpected Perfomance, Waste, Risk
Consolidation
(virtual machines per host)

In today's data centers, physical compute and storage infrastructure serves up the resources that are consumed by virtual servers running Enterprise workloads. Some of the most obvious such resources are CPU and memory. But how much of our processing capacity are we using? Do we have enough RAM? Are we over-provisioned with either?

The promise of virtualization was that physical resources would be more saturated by running multiple workloads on physical clusters. Are we realizing that promise?

Once again, the Global IT Data Lake has our answers!

CPU Load: Are You Using Enough of Your Processing Capacity?

As we did with I/O, we'll start our investigation of processor utilization patterns by asking the Global IT Data Lake a simple question: how much of the vCPU we're allocating to our VMs is actually being used?

The answer is, not much at all.

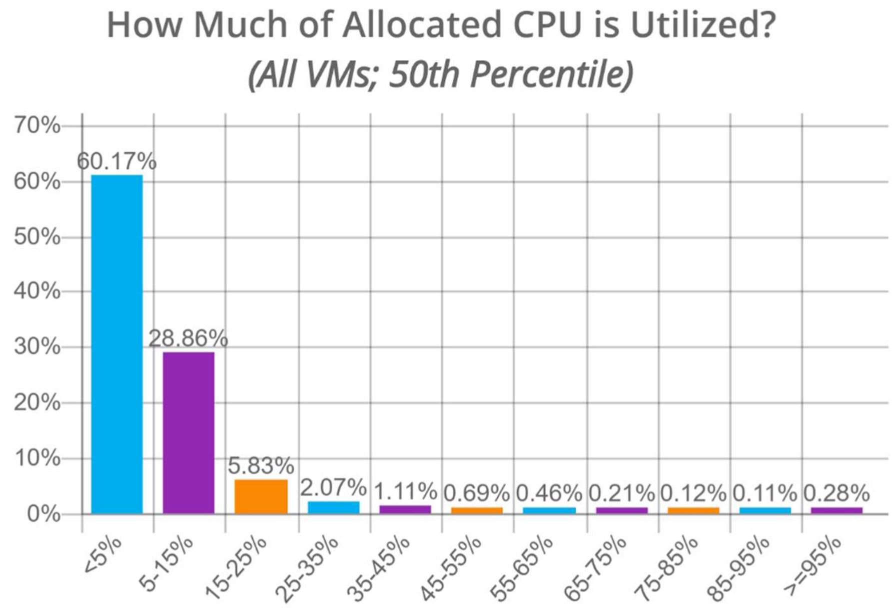


Chart E: vCPU Utilization Distribution; All VMs in the Global IT Data Lake; Median Values Over a 7-Day Period

See Chart E. Nearly 90% of the VMs from our Global IT Data Lake average less than 15% utilization of the vCPU we've allocated to them. Less than 15%! Said differently, only 10% of VMs globally are using a meaningful amount of the vCPU available to them at the median. In fact, barely more than 1% of VMs typically use over 50% of their available vCPU.

Now, that chart is based on median CPU usage for a VM. While understanding the median is instructive for illustrating capacity utilization generally, it is not a very useful metric for capacity planning. Provisioning for median use is not a realistic capacity strategy. A more useful metric for capacity planning would be one closer to expected peak usage. Also, for purposes of gaining insights from a global VM population, it is also interesting to focus on the largest VMs in the Global IT Data Lake; it's reasonable to expect that the largest of our VMs would better saturate their vCPU utilization. So, what do these numbers look like if we look at only "beefy," multi-CPU VMs, at vCPU utilization closer to their peak need?

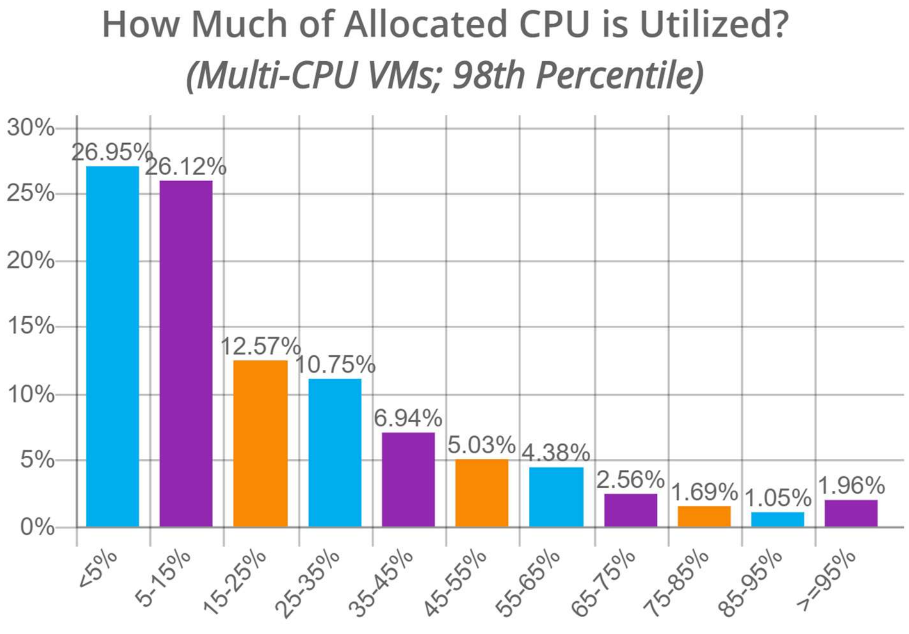


Chart F: vCPU Utilization Distribution; All VMs in the Global IT Data Lake; Median Values over a 7- Day Period

See Chart F. Predictably, this shows a broader vCPU distribution, but still a clear under-utilization of vCPU (remember that this shows 98th percentile usage, which is very close to peak utilization). Still, 11% of VMs are using greater than 55% of their available vCPU.

This is a large number of over-provisioned VMs.

Clearly, there are myriad opportunities to rightsize our VMs! For IT teams considering public cloud or charge-back, this is very important, because pricing in the cloud is typically metered by provisioned resources. More on this point can be found in the “Potential Savings” section.

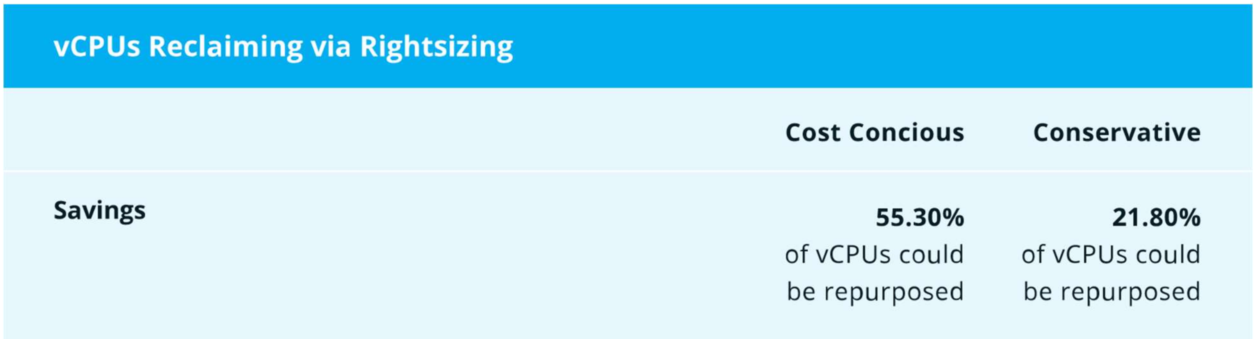
RIGHTSIZING! Could You Be Using Your CPU More Efficiently?

When we looked at vCPU utilization, we identified a global opportunity to rightsize our VMs, with respect to their over-allocation of vCPU. But how broadly could this rightsizing be deployed? The Global IT Data Lake can tell us this as well.

To examine rightsizing opportunities in a pragmatic way, we ran two scenarios, reflecting two potential risk-tolerant profiles:

- The *cost conscious* scenario metered every VM in the Global IT Data Lake at 98th percentile vCPU utilization, and then rounded up from that 98th percentile to the next-logical vCPU count (i.e., round up to 1 vCPU, or an even number of vCPUs)

- The *conservative* scenario metered every VM in the Global IT Data Lake at actual peak vCPU utilization, and then rounded up to the next-logical vCPU count



vCPUs Reclaiming via Rightsizing

	Cost Concious	**Conservative**
Savings	**55.30%** of vCPUs could be repurposed	**21.80%** of vCPUs could be repurposed

Table C: vCPU Reclamation Possibilities from Rightsizing

The results are reflected in Table C, which shows us that ***more than 50% of the world's allocated vCPU could be reclaimed*** in the cost conscious scenario!! That represents *tremendous* potential savings in the world's data centers, particularly when considering a migration to the Public Cloud, where instance sizes are often fixed at a given resource level.

Even in the conservative case, nearly **one-fourth of the world's vCPUs can be reclaimed**, resetting our understanding of our resource availability, and modifying our cost expectations of different deployment options (Private Cloud vs. Public Cloud, for example).

What about Clusters? How Do They Use CPU and Memory?

VMs are provisioned with more vCPU than they need, but that can be overcome and masked by overloading clusters with over-sized VMs. As long as the clusters do not run out of resources, the impact of over-provisioned VMs can be managed.

Thus, we need to ask the Global IT Data Lake what is happening for resource utilization in the supporting clusters.

Fortunately, the Global IT Data Lake provides insights into a broad swath of the data center infrastructure, capturing data that extends well beyond the virtualization tier itself. This includes a great deal of information about cluster resource utilization.

We decided to plot 99th percentile CPU usage vs. 99th percentile memory usage for a representative sample of reasonably-sized clusters in the Global IT Data Lake.

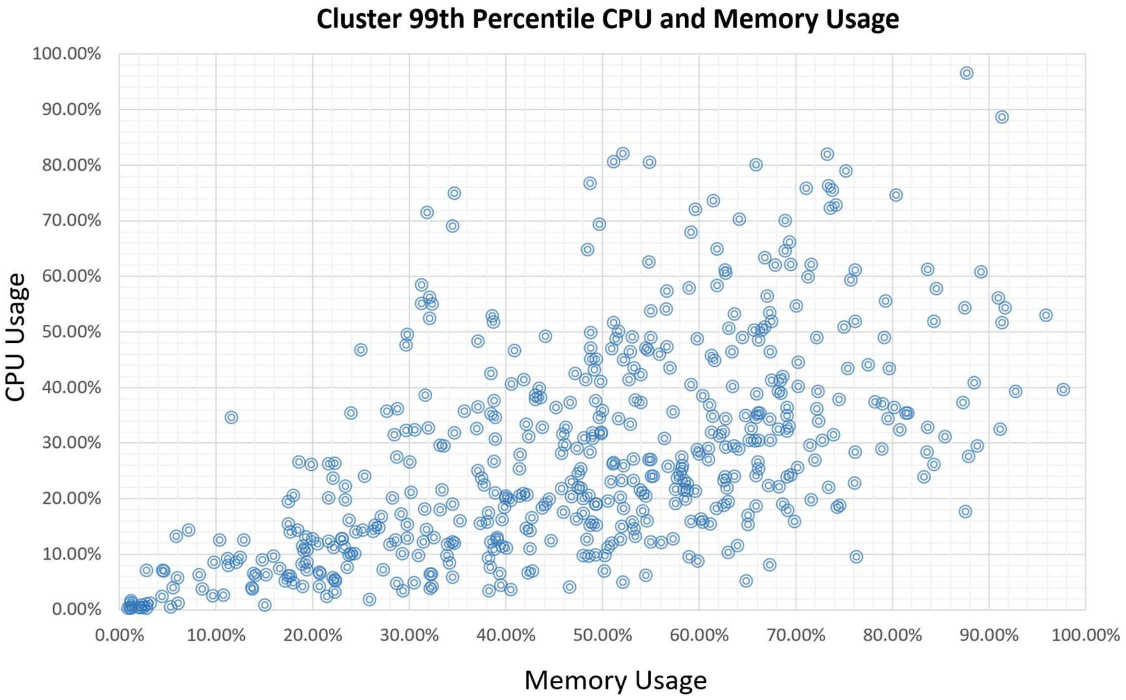


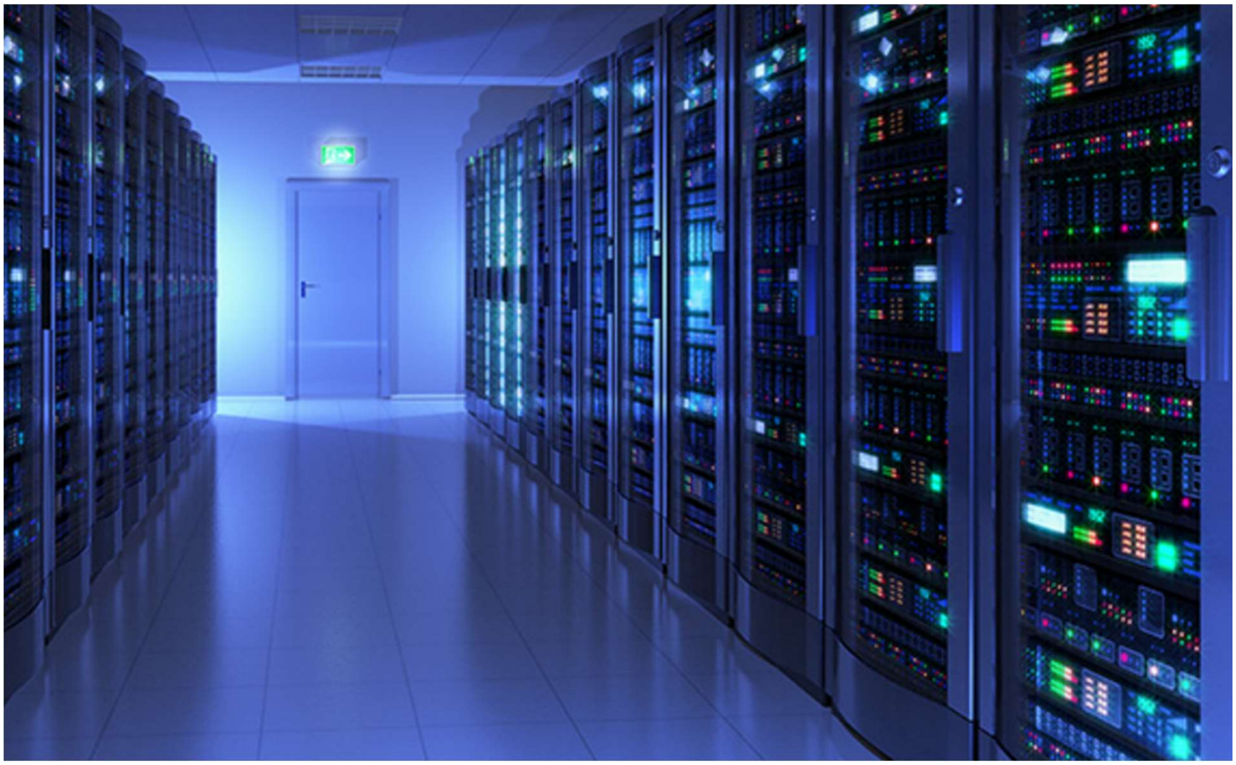
Chart G: Cluster Utilization Analysis – Comparison of 99th percentile CPU utilization vs. 99th percentile Memory utilization

The first thing that jumps off the page is that — once again — there is a lot of unused CPU in our virtualized data centers. Over 85% of the clusters we analyzed used less than 50% of available CPU when metered at the 99th percentile. This means that at very close to peak use, more than half of the CPU was idle for all but 15% of the clusters we analyzed.

While the memory was somewhat more saturated, there is still plenty of headroom as well. Over half of the clusters we analyzed used less than 50% of available memory when metered at the 99th percentile; this means that at very close to peak use, more than half of the memory was unused, and therefore available.

Clearly, if data-driven analysis is made available, we can deploy many more VMs on our existing clusters! The moral of this story is to do your own analysis before "throwing hardware at the problem." You may have more resources available than you think!

What is Happening in the Data Center Today? An Overview of the CloudPhysics Global IT Data Lake



We're frequently approached with questions to ask our Global IT Data Lake. Some of those are collected and discussed elsewhere in this report.

In addition, we're frequently asked questions about the Global IT Data Lake itself: what is it, what is in it, how large is it, and so forth. In this section, we drill in on certain characteristics of the Global IT Data Lake to provide a snapshot into today's data centers.

Server Survey – which vendors are supporting the world's virtual infrastructures in Q4 2016?

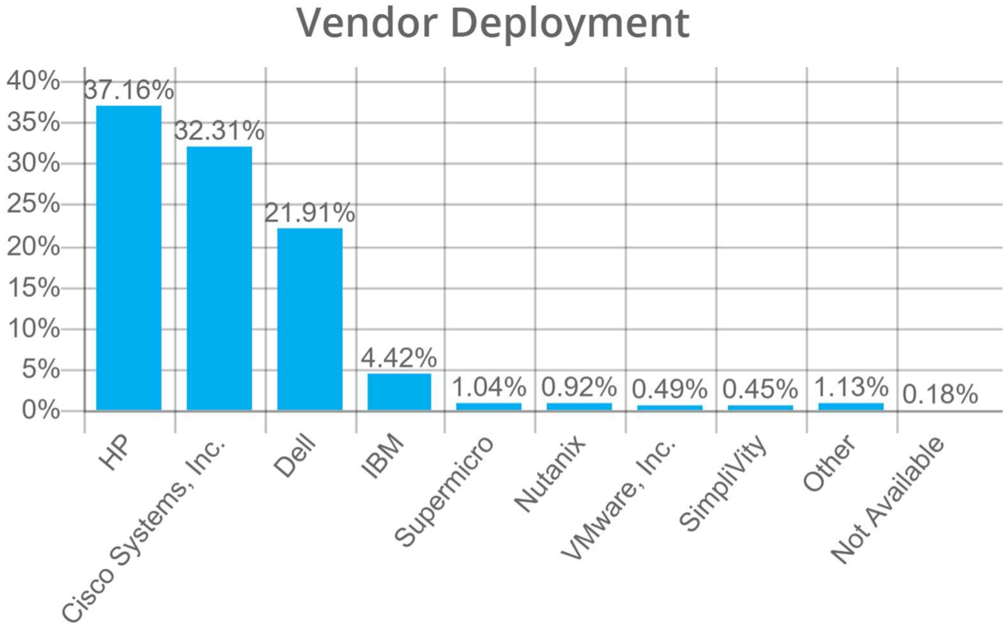


Chart H: Server Vendor Representation

In Chart H, we've depicted the relative representation of server vendors in the Global IT Data Lake. Specifically, the chart depicts the representation of physical servers that run the virtualization infrastructure.

HP leads the way, followed closely by Cisco and Dell, with a "long tail" of other vendors. It is interesting to see the hyper-converged vendors showing up as meaningful contributors to the world's virtual server infrastructure.

To provide some context, we also ran these numbers for Q4 2015. While the big three names were the same as one year ago, their distribution was more skewed; in 2015, HP accounted for 44.4%, Cisco 24.8%, and Dell 19.5%.

Operating System Survey – which OS's are running in virtual servers in Q4 2016?

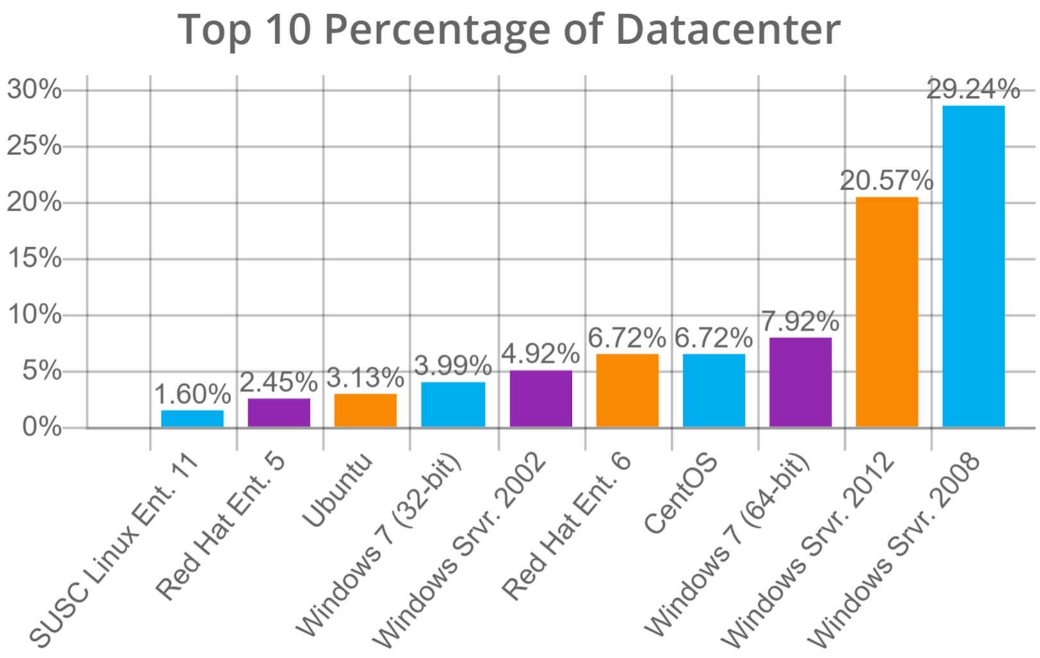


Chart I: Operating System Representation

Chart I depicts the relative representation of Operating Systems running inside the VMs in the Global IT Data Lake. Windows Server 2008 leads the way, with Server 2012 close behind.

A few data points to compare these figures from Q4 2015:

- Windows Server 2012 has increased from 12.5% in Q4 2015, to 20.6% today
- Windows Server 2008 has decreased from 34% in Q4 2015, to 29.2% today
- CentOS (64-bit) has increased from 3.5% to 6.7% today
- RHEL has increased from 4.5% to 6.7% today
- Windows Server 2003 has decreased from 7.5% to less than 5% today

What are the Trends for Windows Server Versions?

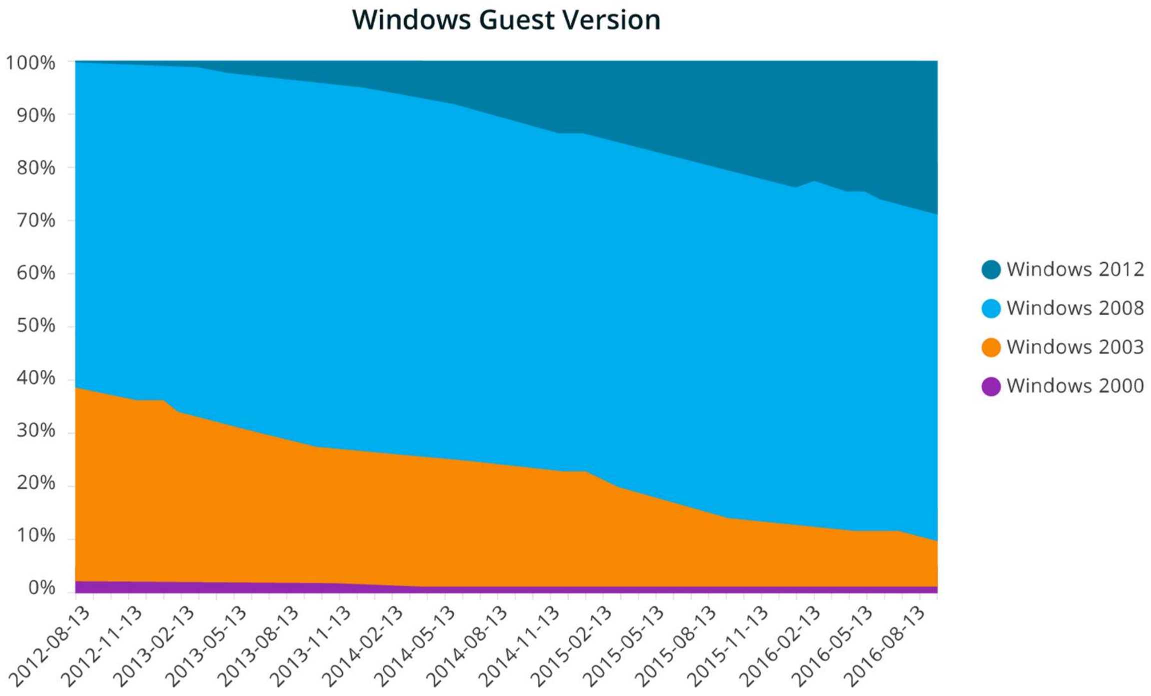


Chart J: Windows Server Distribution across Time

Windows Server operating system version trends were of particular interest to us, for two reasons: first, they are the most broadly-distributed operating system family in the Global IT Data Lake; second, because one version tends to function as direct replacement for an older version. Thus, observing the trends for Windows Server Operating Systems across time can give us a sense of how technologies grow and shrink across time in the world's data centers.

In Chart J, we examined the distribution of Windows Server versions in the Global IT Data Lake going back as far as August of 2012. What we saw was more or less expected: older versions being phased out and replaced by newer versions. What was interesting, however, was the lifecycle of Windows Server 2003.

In August 2012, Windows Server 2003 represented 42% of the Windows Operating Systems in the Global IT Data Lake (2008 held a 56% share). Why is this interesting? Because Windows Server 2003 went off Mainstream Support in July 2010, over 2 years prior.

Extended Support ended five years after Mainstream support, in July 2015; yet at that time, it still held over 13% share among Windows Operating Systems.

Today, 18 months after completely coming off of support, Windows Server 2003 still represents over 8.5% of the Windows Server deployments running in VMs globally.

VMWare Survey – which versions of ESX are most prevalent globally in Q4 2016?

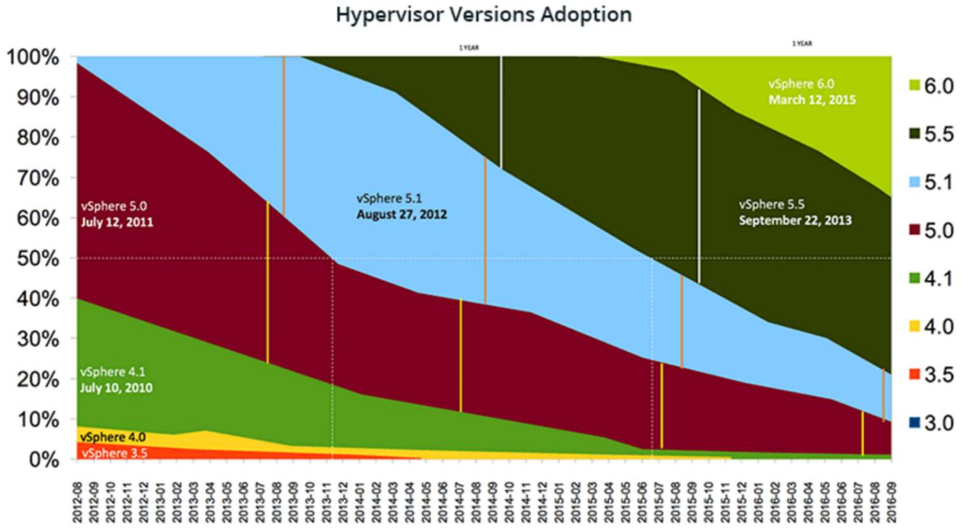


Chart K: Hypervisor Version Distribution

In Chart K, we've depicted the relative representation of hypervisor versions in the Global IT Data Lake.

The distribution today is:

- vSphere 5.5 – 49.2%
- vSphere 6.0 – 31.1%
- vSphere 5.1 – 10.8%
- vSphere 5.0 – 7.5%
- vSphere 4.1 – 1.3%
- vSphere 4.0 – 0.2%

Notably, vSphere 3.1 disappeared from the Global IT Data Lake in February 2014, while vSphere 3.5 didn't disappear until June, 2016, despite VMware having ended general support for it in 2010.

Appendix: Understanding Histograms and Percentiles

Throughout this document, we characterize resource utilization in terms of percentiles: 50th percentile, 95th percentile, and so forth. Some of you may be asking the question, “What is a percentile?”

Wikipedia defines a percentile as “a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value below which 20% of the observations may be found.”

In this report, we’ve used percentiles to describe the utilization of compute and storage resources across time.

For example, we looked at the 95th percentile for VMs I/O generation over a 7-day period. That means, for a given VM in the Global IT Data Lake, we (logically-speaking) “lined up” the I/O per second for each second across 7 days, from lowest value to highest value, then lopped off the top 5% of the values. We then used the next-highest value as a representative of “near-peak” utilization. This is done to remove any infrequent, anomalous, or extreme values from a data set, while still identifying a very high point in its behavior. This means that at the 95th percentile, for example, 72 minutes per day would have higher I/O per second than seen in our graphs (as there are 1440 minutes in a day and 72/1440 = 0.05).

The closer a percentile is to 100%, the closer that reading is to the highest observed utilization.

About CloudPhysics

CloudPhysics provides data-driven insights for smarter IT, delivering unprecedented data center analytics to a broad range of users. CloudPhysics' agile, scalable SaaS solution continuously analyzes customer environments and leverages collective intelligence to yield actionable results that optimize performance, lower costs, reduce risk, and enable smarter business decisions. Headquartered in Santa Clara, CA, CloudPhysics serves thousands of end users worldwide across major industries and supports a robust partner network.

For more information, www.cloudphysics.com