The scalability and economics of delivering Citrix Virtual App and Desktop services from Microsoft Azure

Given business dynamics—seasonal peaks, mergers, acquisitions, and changing business priorities—deploying Citrix Virtual App and Desktop workloads on a Microsoft Azure cloud can give your IT organization the strategic advantage of agile desktop and application delivery.

**Microsoft Azure Cloud Services**

Microsoft Azure is a reliable and flexible cloud platform that allows applications to be quickly deployed across Microsoft-managed datacenters. Azure offers monthly service level agreements (SLAs) of 99.95% to meet strict requirements for continuously available services. A comprehensive list of Microsoft Azure Service Level Agreements may be found at [Microsoft Azure: Service Level Agreements](#).

By provisioning Citrix Virtual Apps desktops and application workloads on Azure Cloud Services, businesses can avoid expenses for internal infrastructure and rely instead on Microsoft to supply the necessary compute, networking, and storage resources for user workloads.

**Citrix Virtual App and Desktop Service**

The Citrix Virtual App and Desktop Service secures the delivery of Windows, Linux, Web or SaaS applications and desktops to any device, empowering today’s modern digital workspace. Citrix Virtual Apps and Desktops provides advanced management and scalability, a rich multimedia experience over any network and self-service applications with universal device support across a full range of mobile endpoints—including laptops, tablets, smartphones, PCs, and Macs.

With session and application virtualization technologies it is easy for customers to manage centralized applications and apply the optimal combination of local and hosted delivery models to match user requirements. Both Citrix hosted desktop sessions and virtual desktops can be provisioned on Azure.

Exclusively available as a hybrid cloud solution, the Citrix Virtual Apps and Desktop Service allows you to choose the workload deployment option that best aligns with your enterprise cloud strategy. When deployed on Microsoft Azure cloud, Citrix Virtual Apps and Desktops gives IT departments the flexibility of delivering infrastructure services for Windows applications and desktops with the elastic scale of public cloud while extending and integrating current investments from on-premise environments.

The Citrix Virtual App and Desktop Service was used in this series of tests for the control and management of the workloads. The numbers herein focus specifically on the scalability and performance of a single Azure virtual machine instance running Citrix’s Server OS Virtual Delivery Agent (VDA) and not Citrix Cloud services.
Scalability testing of Citrix Virtual Apps on Microsoft Azure

The combination of Citrix Cloud and Microsoft Azure makes it possible to spin up new Citrix virtual resources with greater agility and elasticity, adjusting usage as requirements change. Virtual Machines on Azure support all of the control and workload components required for a Citrix Virtual App and Desktop service deployment.

The goal for this document was to analyze the scalability and economics of Citrix virtual app sessions deployed on Microsoft Azure general purpose DS_v3 and compute optimized FS_v2 series instances. As a part of this exercise, performance and price-performance comparisons are used to evaluate the various VM instance types. Login VSI 4.1.32.1 was used in testing to generate user connections to the Server OS VDA workers, simulating typical user workloads running on Azure instances. Microsoft Azure instance types vary according to infrastructure resources provided and relative cost per hour. Pricing and availability for Azure virtual machines varies by region and includes Windows licensing (see Microsoft Azure Documentation: Products available by region and Windows Virtual Machine Pricing).

The Dsv3-series instances are a newer version of Dsv2-series instances. The Dsv3 instance type is built on the Intel® E5-2673 v4 2.3 GHz (Broadwell) and the 2.4 GHz (Haswell) processors which introduce Hyper-Threading technology. These newer instances provide the same performance at approximately 28% lower cost than the similarly configured Dsv2-series instances using physical cores. The Fsv2-series based on the Intel® Xeon® Platinum 8168 2.7GHz processor (SkyLake) with single-core turbo frequency up to 3.7GHz. The Fsv2-series has lower memory per core and provide a lower cost per hour while increasing the raw compute power available.

Citrix Cloud provides a single management plane to deliver unified and reliable access to the apps, desktops, and data that employees need. The test environment leveraged the Citrix Virtual App and Desktop Service for management of the workloads. The remaining infrastructure VMs implemented to support the scalability testing Resource Location — Cloud Connectors, Citrix StoreFront, Active Directory Domain Controllers — were deployed on Standard_D2s_v3 instances. The Azure region used for testing was Azure US West 2.

Topology for the scalability testing

For the scalability testing, the infrastructure VMs were configured with Microsoft Windows Server 2016 on Azure instances as follows:

- **LoginVSI virtual machine on a Standard_D2s_v3 instance containing:**
  - 1x Login VSI controller
  - 4x Login VSI launchers

- **Infrastructure virtual machines on Standard_D2s_v3 instances containing:**
  - 1x StoreFront server
  - 2x Citrix Cloud Connectors
  - 1x Active Directory domain controller, and profile server, and DNS Server

- **Citrix virtual application workloads running on a single Windows Server 2016 instance with the following:**
  - Citrix Server OS VDA 7.17
  - Microsoft Office 2016
  - Microsoft Defender with default settings
  - Latest Windows updates available at time of testing
  - Standard HDD Azure Storage with managed Disks
  - No optimizations to any components: out of the box settings were used across the board

User sessions were simulated using LoginVSI on each instance type in different test runs to analyze the scalability of different Azure instances.
The Delivery Controller, SQL Server, and Studio management console are provided and managed by the Citrix Cloud Virtual App and Desktop service. A StoreFront server was installed on the local network to facilitate testing, along with the Login VSI infrastructure, Active Directory domain controller, and the Server OS VDA. The figure below depicts the test architecture.

Users connect through a StoreFront server to access applications and desktops. (Login VSI clients simulate user connections to the StoreFront server). As in a traditional Citrix architecture, Delivery Controllers distribute the connections and set up service connection between end users and VDAs hosting applications. In a Citrix Cloud based deployment the Delivery Controller functions are distributed between Citrix Cloud and the Resource Location’s Cloud Connectors. All Citrix resources within the Resource Location reference the Cloud Connectors as their Delivery Controllers.

Results summary
The graphs below show side-by-side comparisons of the maximum number of virtual application user sessions supported by the Dsv3-series and the Fsv2-series instance type in single-server scalability testing. The highest densities of 74 and 67 user sessions, for task worker and knowledge worker respectively, were obtained on the F16s_v2 instance type (16 cores, 32 GB RAM) with second place going to the D16s_v3 instance type (16 cores, 64 GB RAM) weighing in at 59 and 56 user sessions for task worker and knowledge worker respectively.
The pricing model for Azure instances varies according to the region, the instance type, and the resources provided. The graph below compares the cost efficiency of each instance type based on VDA user densities achieved in the single server testing. The costs reflect U.S. West 2 Pay-As-You-Go pricing for standard VM instances as of June 2018 and includes the cost of Microsoft Windows licensing.

![Hourly Cost per User by Instance Type](image)

As shown in the graphs, the F16s_v2 and F4s_v2 instance types all show the lowest cost per hour of $0.019 (when rounded) for task worker while the F16s_v2 instance type holds the best hourly cost of $0.0206 for the knowledge worker, followed closely by the F8s_v2 instance type at $0.0230. In the testing, the density results showed a clear benefit from the faster processors available with the Fsv2-series instances. If users run applications that are particularly memory-intensive, the Dsv3-series might be a better choice, though the cost may be slightly higher.

Another item to consider is the impact of the Azure storage type, **Standard or Premium**, on the performance of the instance. With some of the tests, the storage became a bottleneck and some instance types showed extremely high disk response times which in turn lowered the VSImax score. In most cases, the impact was not more than a user or two, but consideration should be given as to whether the extra user density is worth the cost of higher performance storage.

As with any design choice, the utilization of the VMs combined with the organization’s tolerance for resiliency will drive the final decision as to which instance type, size, and storage selection will work best for the organization.

**Testing methodology**

In the scalability testing, Login VSI 4.1.32.1 was used to run a user load on hosted shared desktops using the Server OS VDA. Login VSI helps to gauge the maximum number of users that a desktop environment can support. Login VSI categorizes workloads as Task Worker, Knowledge Worker, Power Worker, and Office Worker. In this testing, only the Task and Knowledge Worker profiles were used.

It is important to note that while scalability testing is a key factor in understanding how a platform and overall solution perform, it should not be inferred as an exact measurement for real world production workloads. Customers looking to better assess how applications will perform in a Citrix and Azure solution should conduct their own Login VSI testing using custom workload scripts. All test results here reflect application execution using default Citrix policies and unoptimized default settings for Windows Server 2016, Office 2016, and Windows Defender. Both performance and density can be improved by leveraging optimization tools such as the **Citrix Optimizer**, but the authors of this paper wanted to provide conservative numbers which can be replicated consistently with little to no specialized knowledge of this type of computing workload.
Task Worker and Knowledge Worker workloads were selected for the testing and had the following characteristics:

- **Task Worker Workload** – includes segments with Microsoft Office 2016 Outlook, Excel, and Internet Explorer, Adobe Acrobat and PDF Writer. The Task Worker workload does not place a high demand on the environment and represents users that do not access the system heavily.

- **Knowledge Worker Workload** – includes segments with Microsoft Office 2016 Outlook, Word, PowerPoint, and Excel; Adobe Acrobat, FreeMind, PhotoViewer, Doro PDF Writer and includes viewing of several 360p movies. The Knowledge Worker workload places a higher demand on the environment, including more use of the available memory, and represents users that access the system more heavily.

Azure Dsv3-series and Fsv2-series VM instance types were tested. Since the goal was to capture a baseline reflecting the user densities for each instance type, the Login VSI client launchers were configured to go through the StoreFront server. Performance metrics were captured during user logon and virtual desktop acquisition (ramp-up), user workload execution (steady state), and user logoff. The full Login VSI test loop (48 minutes) was allowed to complete to achieve consistent measurements while the Login VSI performance metrics were recorded. Tests were repeated five times on each VM instance to get an average number of users that successfully ran the test. Numbers reported consist of the five-run average rounded down to the nearest whole number.

The Login VSI workloads calculate the VSImax session count by observing when the user response time has diminished significantly below the expected threshold which was derived from the baseline value taken with only a single user on the system. Historically speaking, the CPU resource has generally been the bottleneck that prevents a system from reaching a higher VSImax session count. However, in the case of the Fsv2-series instances running on faster SkyLake CPUs, the CPU response time was no longer the key constrained resource; instead, the constraint moved to the available memory.

This dynamic had the unique effect on several Fsv2-series tests such that the VSImax was not able to be calculated in the traditional sense by overtaxing the CPU with additional sessions then working backward to find the optimal user count. The additional sessions consumed enough memory such that the Citrix Virtual App session host had to stop processing all sessions and the test prematurely terminated. For the Fsv2-series instance types, where the test was unable to complete in the traditional manner, the VSImax was obtained instead by identifying the highest number of user sessions where the VSImax was not reached but the test would complete successfully. This was the case only for the Knowledge Worker test runs, which consumed more memory than the Task Worker runs on the F2s_v2, F8s_v2, and F16s_v2 instance types.

One other notable point is that on D2s_v3 and D8s_v3 instance types, the standard disk storage was not able to keep up with the write requests, which reduced the VSImax value. If the workload being considered does generate high disk read/write activity, the recommendation is to consider the use of the Premium SSD storage.
**Detailed instance performance results**

Based on this testing effort and current instance pricing at the time of publication, the F16s_v2 instance type is the most economical, for both light workloads (which resemble the LoginVSI Task Worker workload) and for more memory-intensive workloads (which resemble the LoginVSI Knowledge Worker workload).

Detailed test results for the recommended instance type in both scenarios are provided below and should be used as a starting point to determine which instance type provides the best combination of density and resiliency for your organization. Microsoft Azure provided the flexibility to modify the instance sizing on-demand. Citrix recommends customers conduct their own scalability testing to determine the ideal instance type for their organization and workloads.

**Task Worker workload results**

This section describes test results for the F16s_v2 instance with the Task Worker workload. VSImax v4 (which indicates the maximum user density under a specific workload) is determined from two other metrics, VSI Baseline and VSI Threshold. VSI Baseline represents a pre-test Login VSI baseline response time measurement that is determined before the normal Login VSI sessions are sampled. The F16s_v2 instance demonstrates a VSImax v4 density of 74 users running the Task Worker workload.

The next two graphs depict CPU and memory consumption and disk I/O response times measured during the test.

In the chart below, as user load increases, CPU and memory utilization also increase; notice the memory reaches the max available (32 GB in the F16s_v2 instance type) during the steady state portion.
The write I/O response time averaged around 4.92 milliseconds. Read I/O response times averaged around 1.5 milliseconds.

The following network and disk performance graphs show resource consumption, which can impact scalability as well as cost associated with the solution. The graph below shows networking transfer rates for data going out of Azure data centers. Microsoft charges for
outbound data while inbound data transfers are free. For the Task Worker workload, the average outbound bandwidth during steady state is approximately 705 KBps for the test workload of 76 users. This means the per-user outgoing transfer rate is approximately 9.27 KBps (705/76 = 9.27). Outgoing network transfers during logoff occur as user profile data is transmitted.

Azure charges for disk transfers when standard disks are in use. Disk transfer metrics, which are used in the subsequent cost analysis, are shown below. For the Task Worker workload, disk transfers during steady state averaged about 334 IOPS for the test workload of 76 users, which calculates to an average of about 4.4 IOPS per user. The peak value was 1955 IOPS for 76 users or about 26 IOPS per user. Disk transfer activity is also visible during the logoff period as user profile data is recorded.
Knowledge Worker workload results

Since the F16s_v2 instance also proved to be the most cost-effective for the more intense workload, the following pages show user density and performance metrics for the F16s_v2 instance type running the XenApp 7.17 VDA under the Login VSI Knowledge Worker.

This section shows test results and performance metrics for the F16s_v2 instance under the Knowledge Worker workload. As shown below, the F16S_v2 instance supports a VSImax v4 of 67 users running the Knowledge Worker workload. As noted earlier, the LoginVSI analyzer shows VSImax not reached because this is the highest number of users that could be placed on the F16s_v2 instance and still have a successful completion of the test.

The next two graphs depict CPU and memory consumption and disk I/O response times for the Knowledge Worker workload. These metrics are helpful in assessing performance under the test workload. As the user load increases in the chart below, memory and CPU utilization peaks at the point at which the number of users approaches VSImax v4.
Disk I/O response time metrics for the Knowledge Worker load are shown below. Write I/O response times averaged around 4.2 milliseconds while the read I/O response times averaged to 2.1 milliseconds.

The following graphs show resource consumption metrics for network and disk I/O under the Knowledge Worker workload. The outbound networking transfer rate during steady state averaged around 1818 KBps for the test workload of 67 users. This means the outbound transfer rate per user was approximately 27.13 KB per second ($\frac{1818}{67}=27.13$). During the logoff period, network transfer activity reflects how user profile data is transmitted and captured at logoff.
As shown below, the Knowledge Worker workload imposes an average of 544 IOPS for all 67 users or 8.12 IOPS per user (544/67=8.12). The peak total transfer rate was 1604 IOPS or about 24 IOPS per user.

Economic considerations
This section describes how to estimate the cost of the compute, network, and storage required to deploy Citrix virtual applications on Microsoft Azure. Microsoft Azure deployments are attractive to IT organizations that seek a scalable and agile platform to deploy Citrix infrastructure. Companies pay for only the resources that they use rather than assuming the cost and complexity of developing and maintaining in-house infrastructure. Microsoft provides the required resources from its global network of datacenters, and owns the responsibility and costs associated with infrastructure maintenance, upgrades, monitoring, cooling, power, and service.

Several elements go into a complete Citrix on Azure cost estimate. Typically compute costs associated with the Azure VM instances is the dominant cost factor. These costs include Windows licensing, unless the VMs qualify for Hybrid Use Benefit. Other costs include the cost of outbound network bandwidth and storage.

Citrix licensing costs are also a factor, but since these costs are the same whether deployed on Azure or internal infrastructure, they are not considered in the following cost analysis. The cost calculations presented here determine a monthly cost per user for both Task and Knowledge Worker workloads and reflect West US 2 pricing at the time of this writing.

**Azure compute**

Pay-As-You-Go pricing is calculated on an hourly basis and varies by region according to instance type and the resources that each instance type provides.

For this costing exercise, it is assumed that users work 8-hour shifts and idle and disconnect timers equate to 4 hours. While sessions are active, idle, or disconnected on a Server VDA the compute instance cannot be deallocated. For this reason, the cost estimates given here assume that VMs are allocated and in use for an average period of 12 hours per day. Given this assumption, an F16s_v2 instance has an hourly cost of $1.379, which translates to a monthly cost of about $330.96 at 12 hours per day for 20 days per month. The F16s_v2 can support 74 XenApp Task Workers at a monthly cost of $4.47 per user or 67 XenApp Knowledge Workers at a cost of $4.94 per user.
While the Pay-As-You-Go pricing model is used for the compute calculation in this white paper, additional cost models are available through Microsoft Azure Hybrid Benefit, Reserved VM instances, and Dev/Test pricing. These pricing models can offer substantial savings over the Pay-As-You-Go model. More information on these pricing models and their requirements can be found from Microsoft Azure: [Azure pricing](#).

**Please note:** For VM instances using Pay-As-You-Go pricing, in some cases it may be preferred to have higher numbers of smaller instances. This approach enables individual VDA’s to be more quickly drained of user sessions and deallocated in order to save overall per hour costs. Even though our cost-based recommendation is for the larger F16s_v2 instance size as a single instance, multiple smaller F4s_v2 and F8s_v2 instances may be a better fit within an organization based on their specific VM instance deallocation policies. [Citrix Smart Scale](#) is recommended for dynamic power management and deallocation of Azure VM Pay-As-You-Go Instances based on schedule and load based policies, reducing hourly instance charges to lower total operational expenses. Using Citrix Smart Scale, Azure costs can be optimized by scaling up and scaling down the infrastructure resources to map to the actual usage. For more information about Citrix Smart Scaling, see Citrix Product Documentation: [Smart Scale requirements](#).

**Azure network**

Network costs in a Citrix deployment are primarily associated with the outbound bandwidth of the solution (in GB). Charges vary according to the specific zone (region group) providing services. Pricing is tiered according to the data quantity transferred each month. The monthly cost of most expensive bracket for West US 2 is $0.087 per GB. This number is used for the network calculations in this white paper. For more information on Azure bandwidth pricing please refer to Microsoft Azure: [Bandwidth Pricing Details](#).

As shown in the test results, with the Task Worker workload, an average user consumes outbound network bandwidth at a rate of approximately ~10 Kbps. Additionally, it is assumed when a user session is idle or disconnected (4-hour period) the per user bandwidth utilization is negligible. Therefore, assuming an 8-hour workday, a single XenApp user running a Task Worker workload consumes about ~9 GB of network bandwidth per month—which translates to a cost of $0.78 per month per user. For the Knowledge Worker workload, network utilization is about 28 Kbps per user or ~25 GB for month, which costs about $2.18 per month per user.
**Azure storage**

Azure provides different storage categories and redundancy options. At the time of this whitepaper, Azure storage is divided into Managed and Unmanaged, Standard or Premium, storage. For Standard Unmanaged Disks the storage consumption (in GB) and I/O needs to be considered. For Standard Managed Disks the disk size and I/O needs to be considered. For Premium Managed and Unmanaged Disks, the disk size needs to be considered. Standard HDD Managed Disks were used for this scalability testing. For more information on Azure storage pricing please refer to the Microsoft Azure: [Azure Storage Overview pricing](https://docs.microsoft.com/en-us/azure/storage/about/).

Persistent disks for Azure VMs use “page blobs,” block storage that is optimized for random access. The cost of I/O operations for Standard HDD Managed Disks is $0.0005 per 10,000 transactions to storage, encompassing both read, write, and delete operations. A Task Worker workload imposes about 4.4 IOPS per user on average, which works out to ~3,855,000 transactions per month (assuming an 8-hour workday) at a cost of $0.20 per user. A Knowledge Worker workload requires an average 8.12 IOPS per user. This is about ~7,350,000 IOPS per month per user for an 8-hour day, which means a cost of $0.36 a month for each Knowledge Worker’s storage. Similar to the Azure network calculations, it is assumed that during the period where a user’s session is idle or disconnected (4-hour period) the per user storage utilization is negligible.

Azure maintains storage resources for the Citrix infrastructure even when no users are active, so the analysis of storage consumption charges is based on a 24-hour day.

**Total estimated costs**

The table below shows approximate total costs per user (based on West US 2 pricing) for both Task and Knowledge Worker workloads using the Citrix recommended instance types of F16s_v2. The monthly costs are $10.30 and $12.61 respectively.

<table>
<thead>
<tr>
<th>Component</th>
<th>Instance type</th>
<th>Cost per hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute instance</td>
<td></td>
<td>$6.80</td>
</tr>
<tr>
<td>Network utilization</td>
<td></td>
<td>$0.78</td>
</tr>
<tr>
<td>Storage utilization</td>
<td></td>
<td>$0.20</td>
</tr>
<tr>
<td>Storage capacity</td>
<td></td>
<td>$2.52</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>$10.30</strong></td>
</tr>
<tr>
<td>Task Worker workload</td>
<td></td>
<td>$7.11</td>
</tr>
<tr>
<td>Knowledge Worker workload</td>
<td></td>
<td>$2.18</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>$12.61</strong></td>
</tr>
</tbody>
</table>

Actual costs will vary depending on the region, instance types selected, storage type, instance uptime, and densities achieved with specific user workloads.

**XenApp Infrastructure VM costs on Azure**

In addition to the cost of deploying VMs to support user workloads, a XenApp deployment may include optional* VMs to host infrastructure servers such as StoreFront, or Domain controllers. The table below shows the approximate total cost per hour for each potential XenApp infrastructure VM that may be hosted in the Azure resource location (based on West US 2 pricing).

<table>
<thead>
<tr>
<th>Component</th>
<th>Instance type</th>
<th>Cost per hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>StoreFront (x2) *</td>
<td>DS_v3</td>
<td>$1.88</td>
</tr>
<tr>
<td>Domain controller (x2) *</td>
<td>DS_v3</td>
<td>$1.88</td>
</tr>
<tr>
<td>Citrix Cloud Connectors (x2)</td>
<td>DS_v3</td>
<td>$1.88</td>
</tr>
<tr>
<td><strong>Total for all six:</strong></td>
<td></td>
<td><strong>$1.128</strong></td>
</tr>
</tbody>
</table>

---

*Optional means they can be included as needed.
Conclusion
The Citrix on Azure scalability results presented here should be used only as guidelines in configuring your Azure solution. Before making final sizing and deployment decisions, it is suggested that you run proof-of-concept tests on different Azure instance types using your own workloads. Then use the methodology presented here to approximate your costs.

The Azure instance type that you select to deploy Citrix virtual application workloads is a critical element that determines the user density and scalability, and in turn the cost-per-user for an Azure delivery model. As shown, the different instance types in Azure have advantages for specific workloads, such as high computational requirements or additional memory.

Learn more
For more information about deploying Citrix Virtual Apps workloads on Microsoft Azure Cloud Services, see the Citrix and Microsoft partner web site at http://www.citrix.com/global-partners/microsoft/resources.html.