Parallels Remote Application Server on Microsoft Azure

Scalability and Cost Analysis

Parallels International GmbH
Vordergasse 59
8200 Schaffhausen
Switzerland
Tel: + 41 52 672 20 30
www.parallels.com

Contents

CHAPTER 1

Introduction to Parallels RAS and Microsoft Azure

Deploying Parallels Remote Application Server (RAS) on a Microsoft Azure cloud provides a flexible and comprehensive desktop and application delivery solution that lets you monitor and manage your entire infrastructure.

RAS on Azure is fast to deploy, robust, scalable, and easy to manage, all for only about $6.10 per user per month for a task worker and $7.37 per user per month for a knowledge worker as identified in the findings of this document.

Parallels RAS 17.1 extends its VDI capabilities on Microsoft Azure by allowing organizations to provision, scale and manage required users' workloads on demand directly on Microsoft Azure in a full cloud or hybrid environments rendering faster deployments, management simplification and cost reductions.

This document presents an analysis of the scalability and cost of Parallels RAS deployed on Microsoft Azure.

Parallels Remote Application Server and Microsoft Azure

Parallels RAS is a comprehensive virtual application and VDI solution that allows your employees to use and access applications, desktops and data from any device. Seamless and easy to deploy, configure, and maintain, Parallels RAS supports both Microsoft RDS and major hypervisors.

Microsoft Azure is a reliable and flexible cloud platform that allows applications to be quickly deployed across Microsoft-managed datacenters. For all Virtual Machines (VMs) that have two or more instances deployed in the same Availability Set, Azure offers monthly service level agreements (SLAs) of 99.95% to meet strict requirements for continuously available services.

Microsoft Azure makes it possible to spin up new virtual machines in minutes and adjust usage as infrastructure requirements change. VMs based on various Azure instance types can support all of the infrastructure required from RAS infrastructure servers to workload VMs.

When deployed on Microsoft Azure, Parallels RAS gives businesses the required flexibility to deliver their Windows applications, desktops or VDIs supporting full cloud or hybrid deployments by extending and integrating with on-premises infrastructure. Organizations can scale their infrastructure on demand by automatically monitoring and adjusting resources to maintain high performance for virtual applications and desktops. Auto-provisioning and auto-scaling of workloads can ensure the right balance between resources' availability and associated compute, networking and storage costs when running on Microsoft Azure.

Scalability and Cost of using Parallels RAS on Microsoft Azure

As Microsoft Azure provides various instance types or VMs to host Parallels RAS infrastructure and users' sessions, it is important to choose the right VMs for the right workload for your business and users. To help understand better the scalability of different Azure VMs, Parallels engineers conducted a series of performance tests on different Azure VM types targeted for General Purpose (**Dsv3**-Series) and Compute optimized (**Fsv2**-Series) VMs to host RAS sessions using different users' workload, based on a Task Worker and Knowledge worker user profile.

In order to understand the associated cost, price-performance comparisons were also carried out and presented.

Login VSI 4.1.39.6 was used in testing to generate user connections to RAS Terminal Servers, simulating typical user workloads running on Azure instances. Microsoft Azure instance types vary according to infrastructure resources provided and relative cost per hour. Pricing for Azure virtual machines varies by region and includes Windows licensing (see https://azure.microsoft.com/en-us/pricing/).

CHAPTER 2

Scalability

In this section, scalability results of Parallels RAS along with users' workloads on different Microsoft Azure instances is presented.

Parallels RAS workloads were evaluated on Azure Dsv3 and Fsv2 series VM instance types. The Dsv3-series instances are a newer version of DSv2-series instances. The Dsv3 instance type is built on the Intel® Xeon® 8171M 2.1GHz (Skylake), Intel® E5-2673 v4 2.3 GHz (Broadwell) or the 2.4 GHz (Haswell) processors which feature Hyper-Threading technology providing a better value proposition for general purpose workloads. The Fsv2-seriesis are based on the Intel® Xeon ® Platinum 8168 with single-core turbo frequency up to 3.7GHz while also feature Hyper-Threading technology. The Fsv2-series have a high CPU-to-memory ratio and at a lower cost per hour list price, making such VMs ideal for this test.

The infrastructure VMs needed to deploy Parallels RAS such as Publishing Agent, Secure Client Gateway and other Microsoft services such as Active Directory and DNS servers, were deployed on D2s v3 instances for testing. The table below shows the configuration and hourly cost for a D2s v3 instance type (based on Central U.S. pricing at the time of this writing).

Instance	Virtual cores	RAM (GB)	Storage (GB)	Storage type	Price per hour
D2s v3	2	8	16	4 data. 1 local HDD	$0.21

Configurations for Scalability Testing

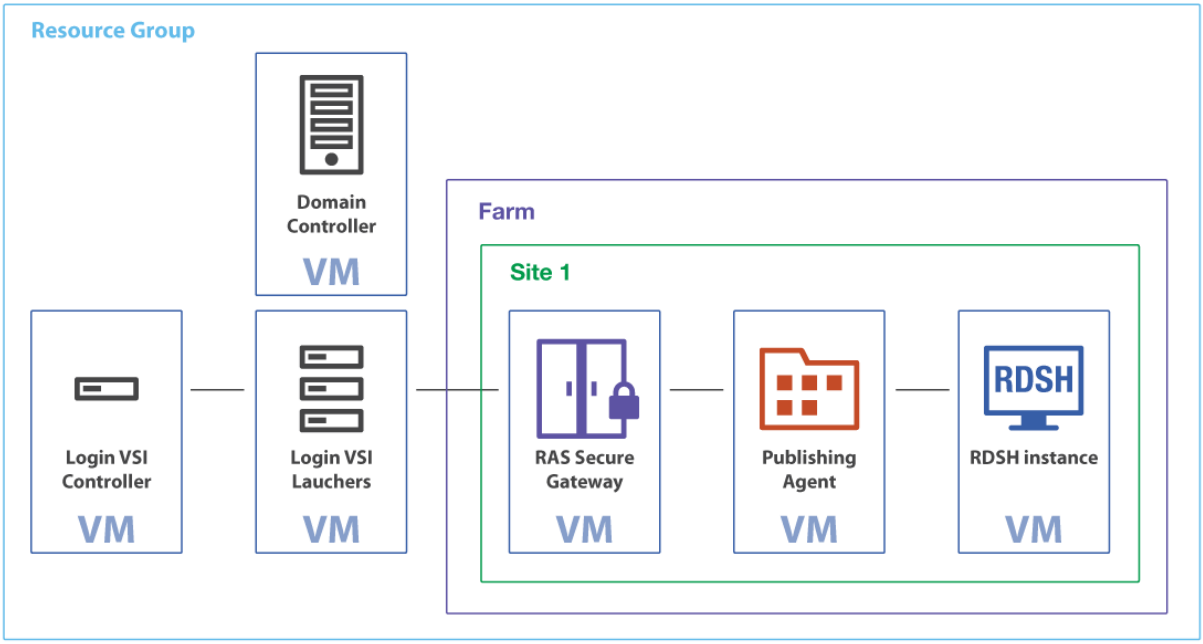
For Parallels RAS 17.1 scalability testing, the infrastructure VMs were configured with Microsoft Windows Server 2016 on Azure instances as follows:

- Five Infrastructure VMs on an D2sv3 instance for Login VSI environment:
    - 1x VM as Login VSI controller and profile server
    - 4x VMs as Login VSI launchers
- Two Infrastructure VMs on a D2sv3 instance for Parallels RAS environment:
    - 1x VM as Parallels RAS Publishing Agent
    - 1x VM as Parallels RAS Secure Client Gateway
- One Infrastructure VM on a D2sv3 instance for Microsoft related services:
    - 1x VM as Active Directory Domain Controller and DNS server

- Various workload VMs where used to test the Parallels RAS user session workloads. Each VM was configured as follows
    - Parallels RAS RD Session Host Agent 17.1
    - Microsoft Office 2016
    - Latest Windows updates available at the time of testing
    - Azure managed standard HDD storage
    - Local users' profiles
    - Out of the box Windows settings were used with no specific Windows optimizations carried out on the VMs

  Creating a virtual machine on Azure also creates an Azure Resource Group container. All virtual machines above resided in the same Resource Group and are siloed on the same virtual network.

The below high level logical diagram shows the architecture and components used for the Parallels RAS scalability testing on Microsoft Azure.



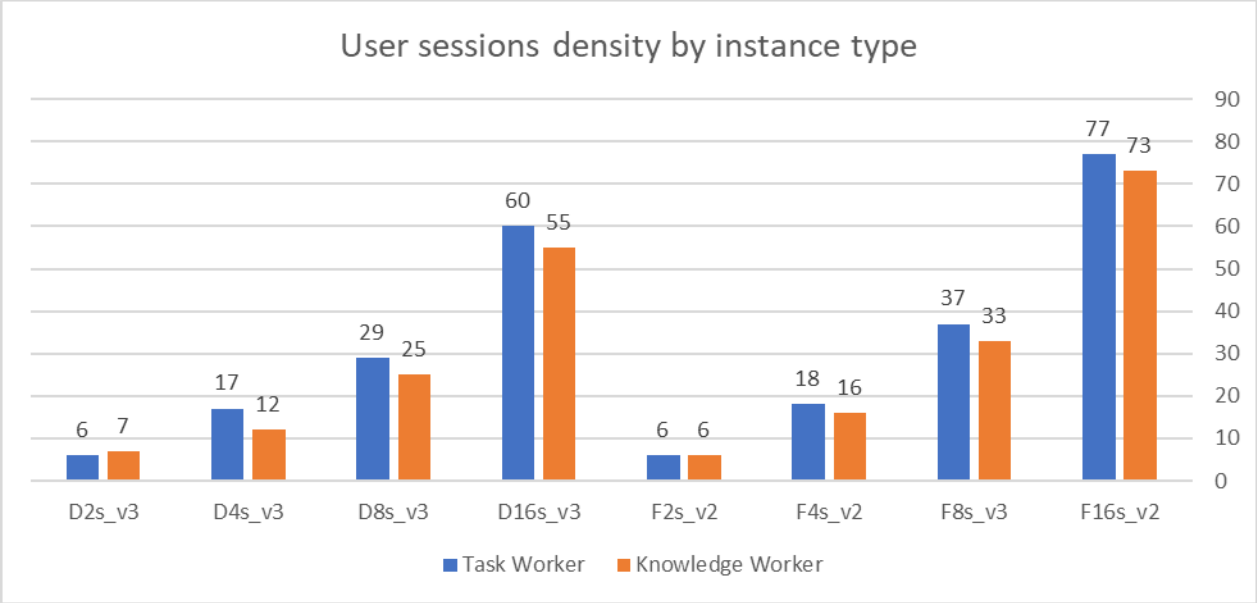
Users connect through Parallels RAS clients for Windows, macOS, iOS, Android, Linux, Chrome OS, or even via Web using HTML5 technology to access their applications and desktops. Login VSI clients simulate such user connections to the Parallels RAS environment. As with the standard Parallels RAS architecture, Publishing Agents distribute the connections and set up service connection between end-users and the RD Session Hosts hosting such applications or desktops.

Findings

The following graphs show side-by-side comparisons of the maximum number of RAS user sessions supported by D-Series and F-Series VM instance types based on a similarly spec'd single server testing.

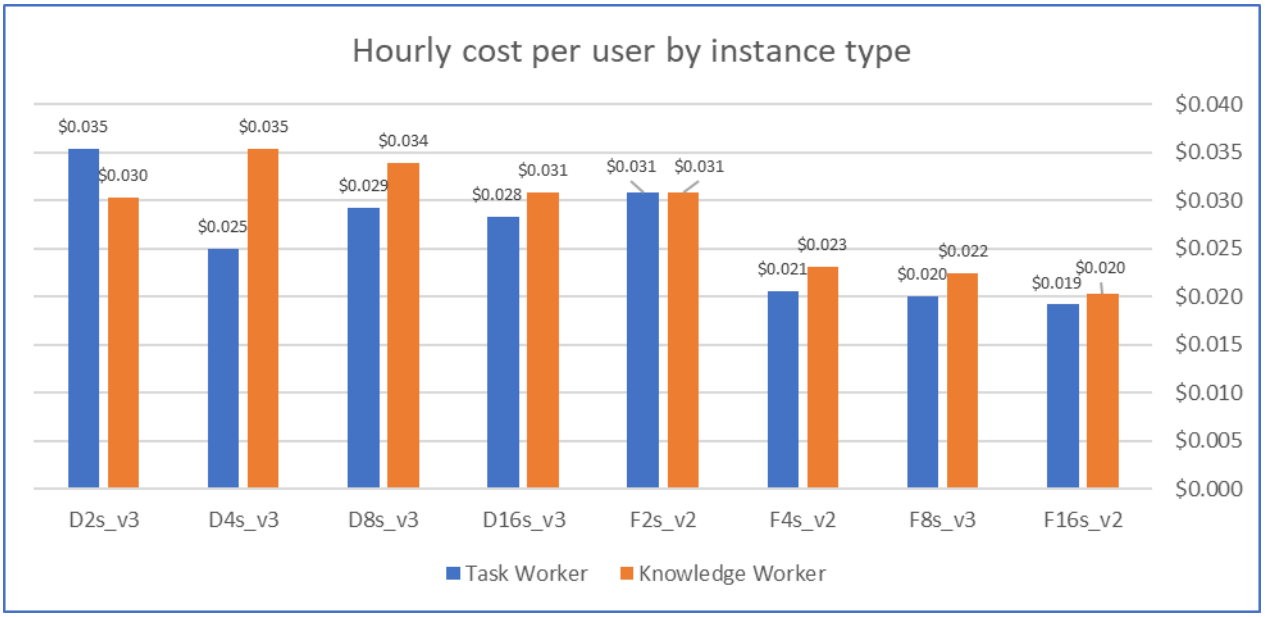
It can be noted that Dsv3-series D2s_V3 (2vCPUs, 8 GB RAM) provided same user density value when compared with the entry level of Fsv2-series which is F2s_v2 (2vCPUs, 4 GB RAM) for Task Worker workload while slightly higher density for Knowledge Worker workload which could be contributed to faster CPUs and more memory resources available.

However, as instance types grow, it was found that the highest density for both Task and Knowledge Worker workloads is achieved on the Fsv2-series instance types with F16s_v2 instance (16 vCPUs, 32 GB RAM) hosting the highest density amongst the tested instances that of 77 and 73 user sessions based on Task and Knowledge Worker workload respectively.



The following graph compares the cost efficiency of each Azure instance type by RD Session Host densities attained in similarly spec'd single server testing based on user.

Pricing for Azure instances depends on region, instance type, and resources provided. Costs shown are based on Central U.S. pricing for standard VM instances, and include Microsoft Windows licensing.



For both Task and Knowledge Worker workload, F16s_v2 instance type shows the lowest cost per user per hour at $0.019 and $0.020 respectively. This is followed by F8s_v3 instance which shows a cost per user per hour of $0.020 for Task Worker and $0.022 for Knowledge worker.

Results clearly show that a user session density can be achieved from the faster processors and higher memory available in the Fsv2-series instances tested. However, if users run applications that are particularly memory-intensive, there may be benefits in using more memory-intensive instances available in the Dsv3-series, though higher costs may be incurred.

Testing methodology

In the scalability testing, Login VSI 4.1.39.6 was used to run a user load on the Parallels RAS session based desktops. 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.

It is important to note that while scalability testing is a key factor in understanding how the platform and the 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 their Parallels RAS environment on Azure should conduct their own Login VSI scale testing using custom workload scripts.

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 very severe demand on the environment and represents users that do not access the system very heavily.
- Knowledge Worker Workload – includes segments with Microsoft 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 more severe

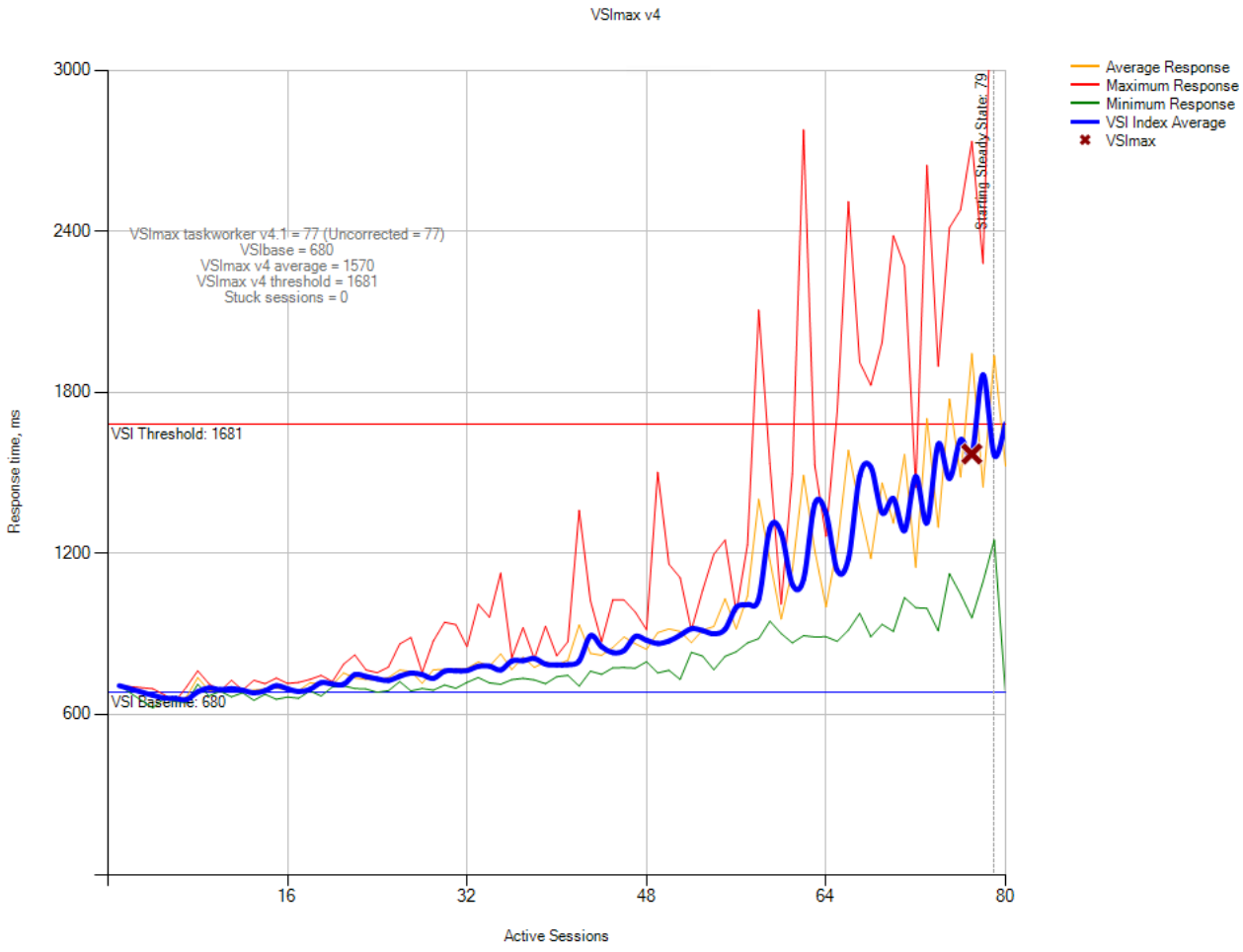
Azure Dsv3-Series and Fsv2-Series VM instance types were tested. To get a baseline showing the densities possible on each instance type in a series, Login VSI client launchers were configured to go through the Parallels Secure Client Gateway server configured in Gateway SSL mode. Performance was measured at user logon and virtual desktop acquisition (ramp-up), user workload execution (steady state), and user logoff. For consistent measurements showing when components were cached, each workload ran for 45 minutes before Login VSI performance was captured. VSI tests were repeated 3 times on each VM instance to get an average number of users that successfully ran the test.

Most cost-effective instance

Based on results achieved, F16s_v2 instance proved to be the most cost-effective for Parallels RAS 17.1. As such, the following sections show user density and performance metrics for the F16s_v2 instance type under the Login VSI Task Worker and Knowledge Worker workloads.

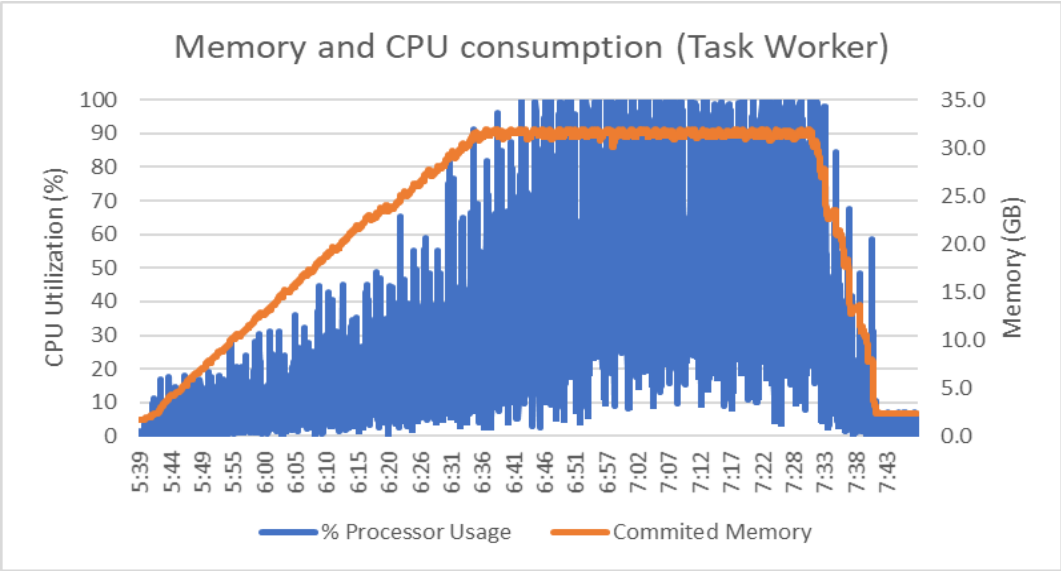
Task Worker workload result

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 was determined from the VSI Baseline and VSI Threshold metrics. 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 shows a VSImax v4 density of 77 users running the Task Worker workload.

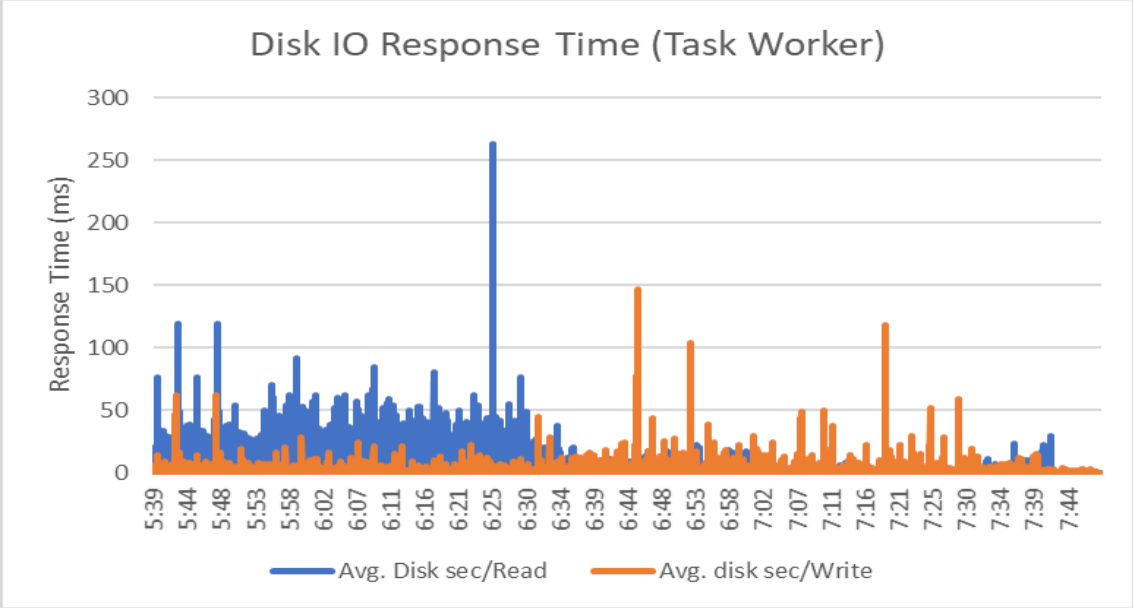


The following test results for CPU and memory consumption and disk I/O response times are helpful in evaluating performance under the Task Worker workload.

In this chart it can be noted that as user load increases towards the maximum, CPU and memory usage peaks.

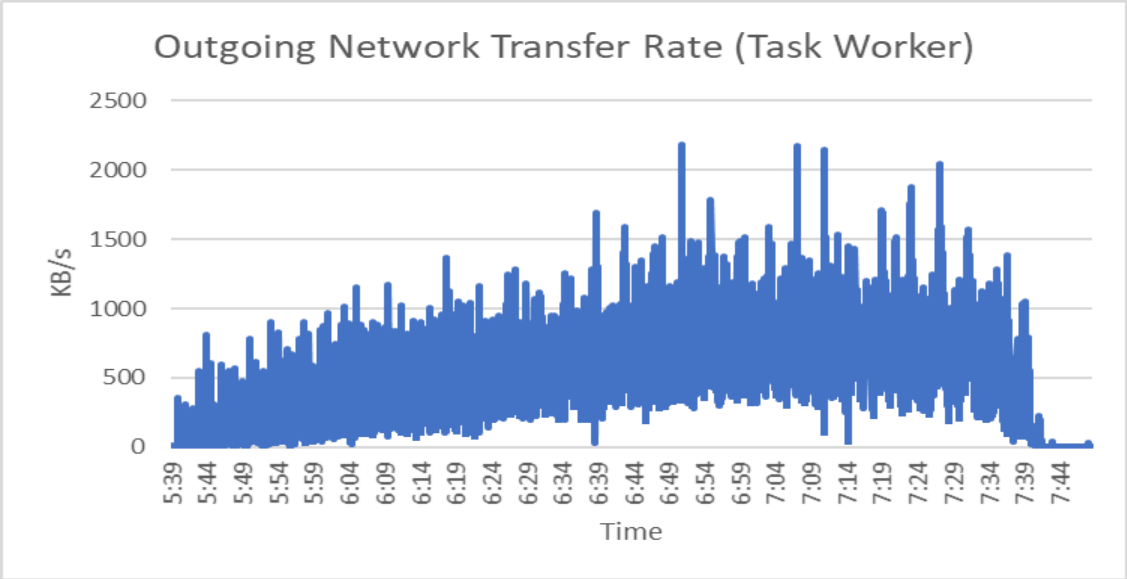


Disk I/O response time metrics for the Task Worker workload are shown below. The write I/O response time averaged around 2.19 milliseconds (ms) while the read I/O response times averaged around 1.14 milliseconds (ms).

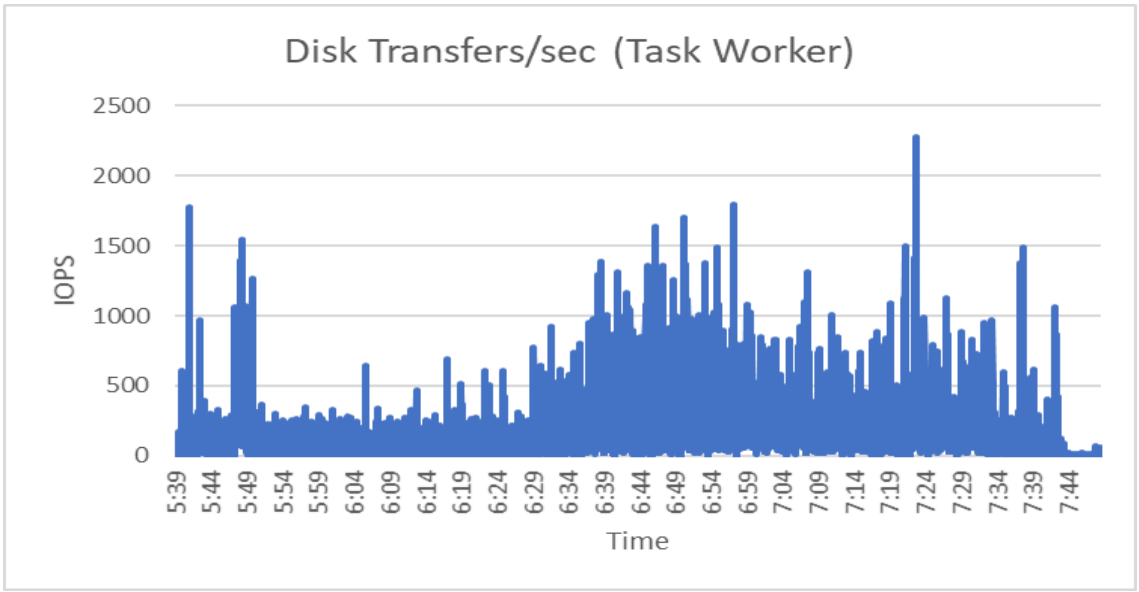


The next two graphs show resource consumption for network transfers and disk performance, both of which affect scalability and cost. The first graph shows networking transfer rates for data going out from Azure data centers. Microsoft charges for outbound data (and inbound data is free).

For the Task Worker workload, the average outbound bandwidth during steady state is approximately 704.9 KB/s for the test workload of 77 users. Therefore the outgoing transfer rate per user is approximately 9.15 KB/s (794.9/77 = 9.15 KB/s). Outgoing network transfers during logoff occur as user profile data is transmitted.

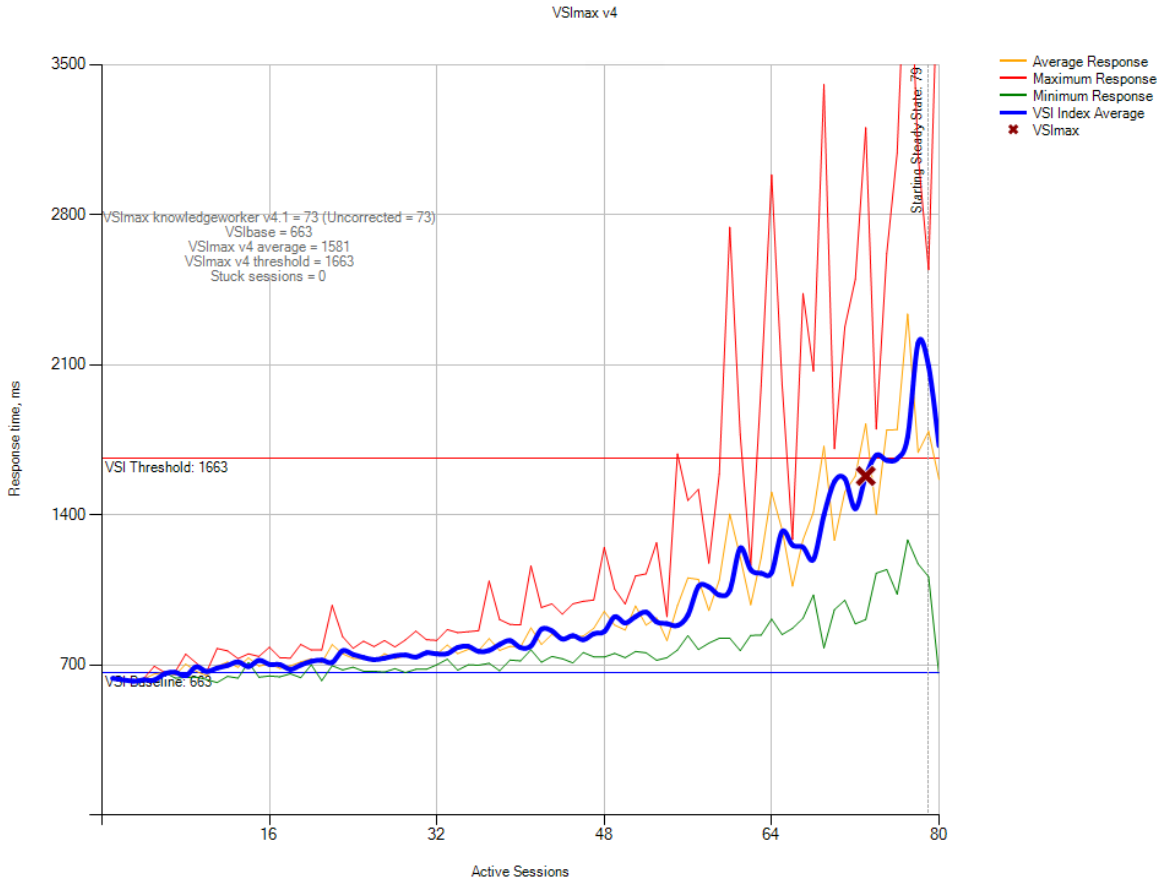


The graph below shows disk transfer metrics. For the Task Worker workload, disk transfers during steady state averaged about 255 IOPS for the test group of 77 users, or about 3.31 IOPS per task worker user. The peak value was 2277 IOPS for 77 users or about 29.6 IOPS per task worker user. Disk transfer activity is also visible during the logoff period as user profile data is recorded.



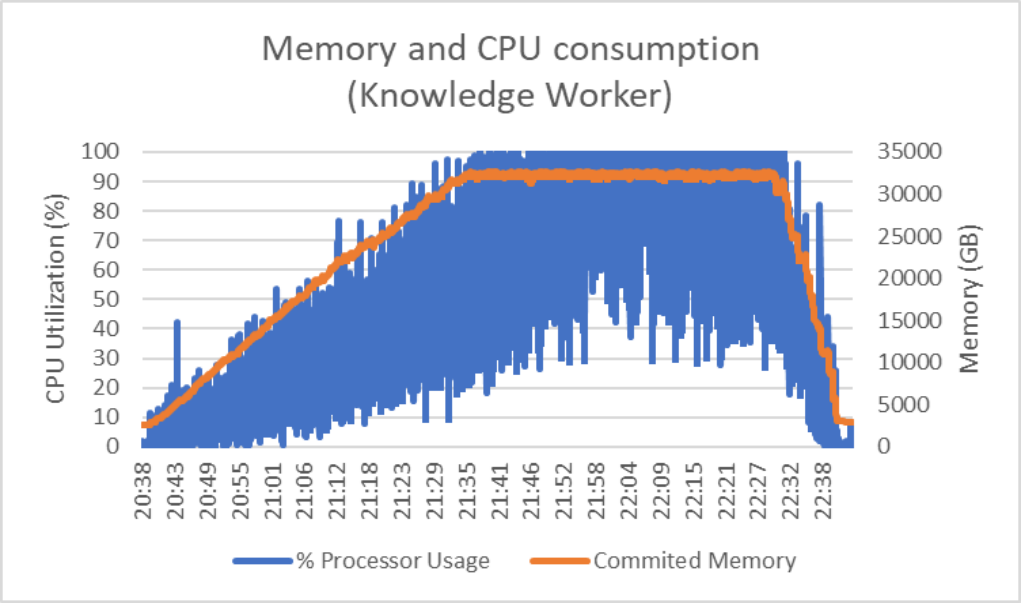
Knowledge Worker workload results

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 73 users running the Knowledge Worker workload.

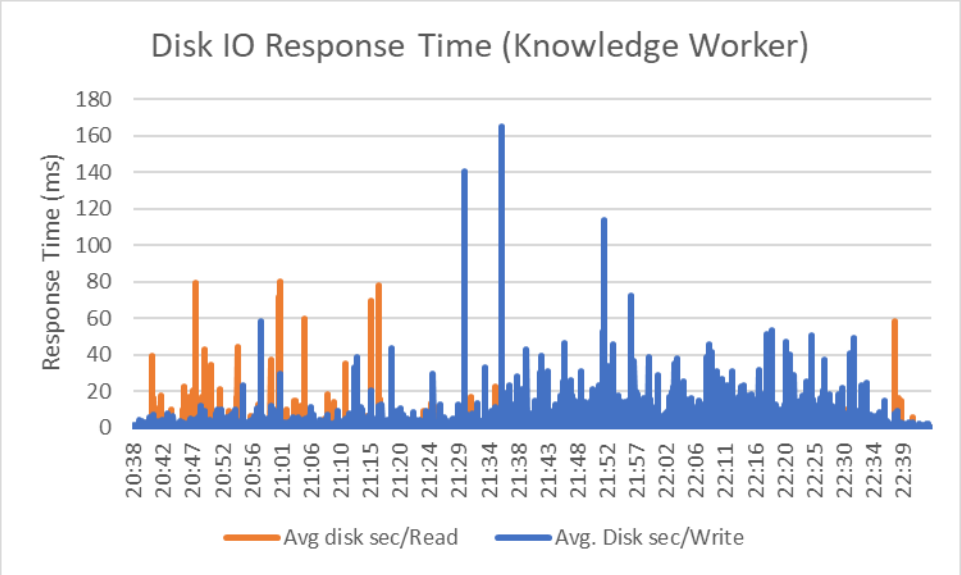


The following test results for CPU and memory consumption and disk I/O response times are helpful in evaluating performance under the Knowledge Worker workload.

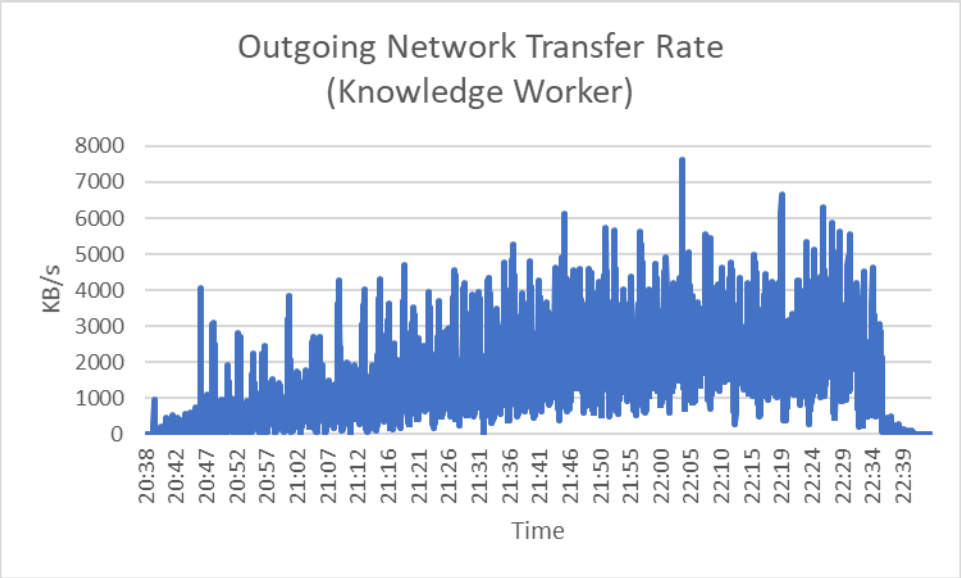
In this chart it can be noted that as user load increases towards the maximum, CPU and memory usage peaks.



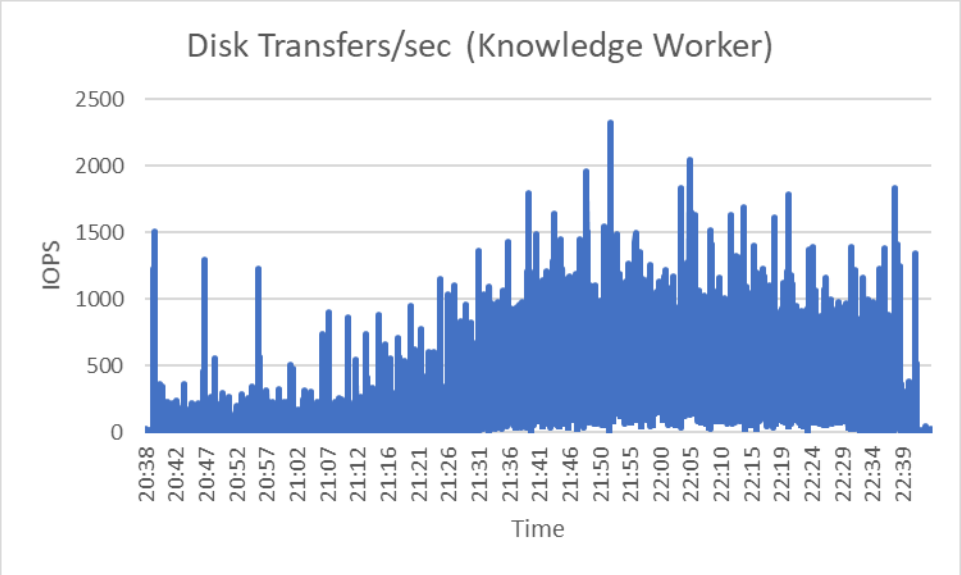
Disk I/O response time metrics for the Knowledge Worker workload are shown below. The write I/O response time averaged around 1.9 milliseconds (ms) while the read I/O response times averaged around 0.8 milliseconds (ms).



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 2132.74 KB/s for the test workload of 73 users. This means the outbound transfer rate per user was approximately 29.22 KB/s (2132.74/73=29.22 KB/s). During the logoff period, network transfer activity reflects how user profile data is transmitted and captured at logoff.



As shown below, for the Knowledge Worker workload, disk transfers during steady state averaged about 410 IOPS for all 73 users or 5.6 IOPs per user (419/73=5.6). The peak value for disk transfer rate was 1925 IOPS, or 26.4 IOPS per user.



CHAPTER 3

Costs

Using Microsoft Azure lets you sidestep the cost and complexity of developing the infrastructure needed for deploying RAS desktops and applications, as well as the considerable ongoing responsibility and costs to maintain that infrastructure. Instead, Microsoft provides all the necessary resources from its own datacenters.

This section describes how to estimate the cost of Infrastructure and workload VMs to deploy Parallels RAS on Microsoft Azure. The cost of Azure instances (which also includes Windows licensing unless Azure Benefits are claimed) is the main consideration in a budget estimate for RAS on Azure. Also to be considered are the cost of network and storage I/O, and storage used. Since Parallels RAS licensing costs are the same whether deployed on Azure or via an internal infrastructure, they aren't considered in our cost analysis. It is also assumed that all Parallels RAS environment is running on Microsoft Azure. Other deployment methods such as Hybrid, leveraging the on-premises infrastructure are expected to reduce further the costs associated with running full workload on Microsoft Azure.

The following cost analysis shows a monthly cost per user for both Task and Knowledge Worker workloads, based on current (as of this writing) Central U.S. pricing. For this costing exercised it is assumed that VMs are powered on and operational for 8 hours a day for an average of 20 days per month to exclude weekends for a total of 160 hours per month. Pay as you go model was chosen to calculate costs in this white paper. Other pricing models including Microsoft Azure Hybrid Benefits, Reserved VMs and Dev/Test pricing can provide significant savings on top of the pay-as-you-go model. For updated prices please refer to https://azure.microsoft.com/en-us/pricing/.

Note that costs estimates provided below are for guidance purposes. Actual costs may vary, depending on region, instance infrastructure, and actual densities attained with real-world user workloads.

Cost of Azure Instances

The cost of Azure instances is the main consideration in a budget estimate for Parallels RAS on Azure. Pricing for Azure virtual machines differs by region depending on instance type and resources provided by each instance (see https://azure.microsoft.com/en-us/pricing/details/virtualmachines/windows/).

Use of Azure is priced by the hour. With Azure you can reduce hourly instance costs by shutting down and deallocating virtual machines that aren't in use, but the following cost estimates (except for storage capacity) assume that all VMs are allocated and in use for 8 hours each work day.

With that in mind, a F16s_V2 instance has a monthly cost of $248.32 and can support 77 RAS users under a Task Worker workload and 73 RAS users under a Knowledge Worker workload, for a monthly cost of $3.22 for each Task Worker workload user and $3.40 for each Knowledge Worker workload user.

Cost of Network

Microsoft charges for data going out of Azure datacenters, but not for inbound data. Charges vary depending on the region providing services, and are tiered according to how much data is transferred each month. The monthly cost for the most expensive tier, zone1, is $0.087 per GB. For more details please refer to Azure bandwidth pricing details https://azure.microsoft.com/en-us/pricing/details/bandwidth/.

As found in our tests, the average user running a Task Worker workload uses network bandwidth at about 9.15 KB/s. So for an 8-hour workday, that's about 5 GB of network bandwidth per month which amounts to approximately $0.44 per user per month. For the Knowledge Worker workload, network use is approximately 29.22 KB/s for each user, or about 16 GB for an 8-hour day per month. That's about $1.39 per user per month.

Cost of Storage (I/O)

The cost of I/O (read and write) operations to disk using managed Standard HDD is $0.0005 per 10,000 transactions. For more details please refer to Azure Storage pricing details https://azure.microsoft.com/en-us/pricing/details/managed-disks/.

From results achieved, a user with a Task Worker workload has about 3.31 IOPS, or about 2 million transactions per month, again, assuming an 8-hour workday for 20 days in a month. This brings the storage cost (I/O) of $0.10 per user per month. A Knowledge Worker workload has an average 5.6 IOPS per user, or about 3,225,600 transactions per month. That's a cost of $0.16 per user per month for a Knowledge Worker.

Cost of Storage (Capacity)

Even when no users are active, Azure maintains persistent storage capacity and resources for the RAS infrastructure. Given that, analysis of charges for storage consumption is based on a 24-hour day. Azure provides various storage categories and options for redundancy. Storage pricing is tiered with lower rates for higher levels of consumption.

Total cost estimates

Workload VMs costs

The table below shows approximate total costs per user (based on Central U.S. pricing) for both Task and Knowledge Worker workloads. Based on the F16s_v2 compute instance, the monthly cost per Task Worker workload user is about $6.10. The monthly cost for each user running a Knowledge Worker workload is about $7.37.

Azure Resources	**Task Worker workload per user per month**	**Knowledge Worker workload per user per month**
F16s_v2 compute instance	$3.22	$3.40
Network utilization	$0.44	$1.39
Storage utilization	$0.10	$0.16
Storage capacity	$2.34	$2.42
Total	$6.10	$7.37

RAS Infrastructure VMs costs

In addition to the cost of deploying VMs to support user workloads, a RAS deployment requires VMs to host infrastructure servers. The table below shows the approximate total cost per hour for each RAS infrastructure VM in Azure.

Component	Instance type	Cost per hour
Publishing Agent	D2sv3	$0.211
Secure Client Gateway	D2sv3	$0.211
Domain Controller*	D2sv3	$0.211
Total		$0.633

*If using hybrid or Azure Active Directory (AAD) with Azure Active Directory Domain Services (AADDS) the mentioned Domain Controllers on IAAS VMs would not be required. However additional networking costs may apply as such costs may vary.

Note: Central U.S. pricing based on pay-as-you go. Other pricing models including Microsoft Azure Hybrid Benefits, Reserved VMs and Dev/Test pricing can provide significant savings on top of the pay-as-you-go model.

Note: In production environment it is recommended to consider High Availability (HA) and N+1 redundancy concept.

Note: Microsoft Azure Load Balancer is recommended to be used to Load balance incoming front-end traffic to Parallels Secure Client Gateways when deployed on Microsoft Azure. Please refer to https://kb.parallels.com/en/124539 for further information.

CHAPTER 4

Conclusion

The Parallels RAS on Azure costs and 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 workload requirements.

The Azure instance type that you select to deploy Parallels RAS workloads is the critical element that determines the user density and solution scalability, and in turn the cost-per-user for an Azure delivery model. Different instance types in Azure have advantages for specific workloads.

Appendix

The below tables depict VM sizes used in this document.

General Purpose - DSv2-series

Premium Storage: Supported

Premium Storage caching: Supported

Size	vCPU	Memory: GiB	Temp storage (SSD) GiB	Max data disks
Standard_D2s_v3	2	8	16	4
Standard_D4s_v3	4	16	32	8
Standard_D8s_v3	8	32	64	16
Standard_D16s_v3	16	64	128	32

For further details please see https://docs.microsoft.com/en-us/azure/virtual-machines/windows/sizes-general

Compute Optimized - Fsv2-series

Premium Storage: Supported

Premium Storage caching: Supported

Size	vCPU	Memory: GiB	Temp storage (SSD) GiB	Max data disks
Standard_F2s_v2	2	4	16	4
Standard_F4s_v2	4	8	32	8
Standard_F8s_v2	8	16	64	16
Standard_F16s_v2	16	32	128	32

For further details please see https://docs.microsoft.com/en-us/azure/virtual-machines/windows/sizes-compute

Azure Pricing Calculator

Configure and estimate the costs for Azure products including Virtual Machines from Microsoft Azure Pricing calculator found here https://azure.microsoft.com/en-us/pricing/calculator/.