Parallels Remote Application Server

5000 User Load and Performance Analysis

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

Contents

CHAPTER 1

Version History

Revision	Description	Date
1.0	Initial release	March 2020

CHAPTER 2

Introduction

Parallels® Remote Application Server (RAS) is an application delivery and VDI solution that allows users to securely access virtual workspaces from anywhere, on any device, anytime. Our cloud-ready software empowers organizations to embrace digital transformation by centralizing management of the IT infrastructure, streamlining multi-cloud deployment, reinforcing data security and improving process automation.

Key benefits of Parallels RAS for your business

- **Superior user experience on any device, anywhere**

  Increase productivity by allowing employees to access their workspaces from any device wherever they are, in the office or on the go. Work on applications and data from any operating system (OS) with platform-specific Parallels Clients for Windows, Linux, Mac®, iOS, Android, Chromebook™ and HTML5 clientless web access. Employees can access their digital workspaces 24/7, switching between devices and locations.

- **Enhanced data security**

  Protecting data is crucial to organizations — any loss can result in huge costs. Parallels RAS reinforces security by centralizing and managing data access. The risk of data loss and malicious activity is reduced by policies that limit access based on user, group permissions, locations, and devices. Parallels RAS supports FIPS 140-2 encryption and multifactor authentication (MFA).

- **IT agility and business readiness**

  The flexible and scalable architecture of Parallels RAS enables organizations to address business demands in real-time and quickly adapt to continuous workplace changes and increase productivity. Auto-scale your IT infrastructure, mix and match different technologies, such as Windows Server OSs, hypervisors, and hyperconverged solutions, and centrally manage multi-cloud deployments.

- **Easy to deploy, configure, and maintain**

  Parallels RAS streamlines the deployment and maintenance of IT infrastructures. Available out-of-the-box, it is easy to set up and manage, reducing IT workloads. A unified and intuitive management console, configuration wizards and a customizable set of tools are used to effortlessly deliver applications, desktops, and data to any device.

- **Reduce total cost of ownership (TCO)**

  As an all-in-one solution, Parallels RAS saves resources, reduces the hardware footprint and lowers overhead costs. A single licensing model incorporates all of the product's comprehensive features, and the learning curve for admins is faster due to a minimal amount of free training needed.

Organizations can embrace digital transformation and grow their business with Parallels RAS by enabling workforce mobility leveraging a secure remote application delivery and VDI solutions. For further information, please visit https://parallels.com/ras.

In This Chapter

Purpose of Document

This document is to be used as guidance to prepare your IT organization for the commissioning of Parallels RAS environment for 5000 concurrent users. A dedicated testing platform has been created to highlight design, architecture, deployment and large-scale validation of Parallels RAS 17, focusing on office workers using published applications and desktops delivered from RD Session Hosts based on Microsoft Windows Server 2016.

Objectives

This undertaking consisted of designing, building, and testing a Parallels RAS environment with the intent to highlight Parallels RAS core components scalability capabilities such as Parallels Publishing Agents (PA) and Secure Client Gateways (SCG) along with showcasing the simple modular design required to handle 5000 concurrent user sessions.

For the Parallels RAS core components scalability testing, the following goals were set:

- A Parallels RAS environment that can handle 5000 concurrent user sessions with a buffer 20% in case of unplanned influx of users.
- All users are to be internal to the environment.
- Uninterrupted access for all users during logon, workload and logoff phases.
- Validate that the Parallels RAS components remain stable with no unexpected behavior during heavy load.
- Confirm that Parallels RAS infrastructure servers are not to exceed an average of 80% CPU utilization at peak load with adequate memory.

In addition, user experience performance-based testing using Login VSI was also carried out to provide more information regarding the scalability of an RD Session Host. In this case the goal set was to help identify the resource utilization requirement for a specific user workload and provide results weather it is decided to scale-up or scale-out. This is meant to help the IT organization to design the environment based on their business requirements.

Document Overview

This document provides a lab environment overview which was created specifically with the intention to load test Parallels RAS components such as RAS Publishing Agents (PAs) and RAS Secure Clients Gateways (SCGs) and validate against a 5000-user workload. Test methodology and configurations are presented and results are highlighted in the **Findings** (p. 30). Additional performance tests were carried out on RD Session Hosts and findings are presented based on scale-up and scale-out approaches.

Target Audience

This document is intended for IT architects and implementation engineers, along with other IT professionals with relevant experience in virtualization technologies, who are interested in the scalability and performance results of a large-scale Parallels RAS deployment. Other stakeholders may also use this document for better understanding on the requirements of Application and Desktop delivery.

The reader should have a strong understanding of Parallels RAS and IT infrastructure such as servers, storage, and network resources. For step by step configuration instructions or further details regarding Parallels RAS, please refer to Parallels RAS documentation at https://www.parallels.com/eu/products/ras/resources.

Executive Summary

This document confirms a successful validation of Parallels RAS setup in a large-scale environment hosting over 5000 users. For this undertaking, a dedicated lab environment was commissioned to run both load and performance tests using different user simulation tools. The tools used for such a simulation included the Parallels in-house load simulation tool, which was used to load test Parallels RAS infrastructure components, such as RAS Publishing Agents and RAS Secure Client Gateways, and Login VSI, which was used to do performance-based testing on differently sized RD Session Hosts. In both cases the Office Worker workload type profile was used running various applications in-session to simulate a common office worker workday. Parallels RAS Performance Monitor and Login VSI reports were used to capture the data which was then presented in the document herein.

With the load equally distributed on three identical Publishing Agents, one primary and two secondary Publishing Agents, each with 4 vCPUs and 8 GB of RAM, results indicate that both CPU and RAM resources available were well adequate for a 5000-user deployment with room to scale. During logon phase the average CPU utilization was at an average of 23% and a maximum of 47%, while memory consumption was consistent at around 50% utilization. During workload phase CPU utilization was noted as steady at an average of 35% and a peak of 60% while memory consumed remained consistent at around 50% utilization.

Load to ten Secure Client Gateways, each with 4 vCPU and 8 GB of RAM, was also equally distributed using Microsoft NLB. This means that each Secure Client Gateway hosted 500 tunneled user sessions with SSL offloading. Results indicate that both CPU and RAM resources provided were well adequate, even if memory assigned would have been much lower, for a 5000-user deployment with room to scale. During logon phase the average CPU utilization was at an average of 14% and a maximum of 38%, while memory consumed was consistent at around only 12% utilization. The highest maximum transfer rate was noted at 10.32 MB/s which translates to 21.1 KB/s per user during logon phase. During workload phase, CPU utilization was noted as steady at an average of 28% and a peak of 55% while memory consumed remained consistently low at around only 12% utilization. The maximum transfer rate was noted at 11.64 MB/s or 23.8KB/s per user during workload phase.

As a reference to aid organizations with sizing RD Session Hosts, two basic strategies were considered, scaling up or scaling out:

- For the scale up approach, an RD Session Host with 12 vCPUs and 64 GB RAM was used with the intention to host around 75 user sessions. Using Login VSI, it was noted that with the mentioned specifications, VSIMax was reached when running 77 user sessions on the same host. It was noted that as user load increased towards the maximum, CPU reaches its maximum while memory consumption increased up to 70% utilization during peak load. The average outbound bandwidth during steady state was noted at approximately 3.83 MB/s or 51.0 KB/s per user each generating an average of 6 IOPS.
- For the scale out approach, an RD Session Host with 4 vCPUs and 14 GB RAM was used with the intention to host around 25 user sessions. Using Login VSI, it was noted that with the mentioned specifications, VSIMax was reached when running 25 user sessions on the same host. It was noted that as user load increased towards the maximum, CPU reaches its maximum while memory consumption increased up to 90% utilization during peak load. The average outbound bandwidth during steady state was noted at approximately 1.4 MB/s or 57.3 KB/s per user each generating an average of 11.76 IOPS.

Abbreviations

The following abbreviations are used in this document.

AD	Active Directory
DHCP	Dynamic Host Configuration Protocol
DNS	Domain Name System
HA	High Availability
HALB	Parallels High Availability Load Balancer
IOPS	Input/Output Operations Per Second
NLB	Network Load Balancer
NTP	Network Time Protocol
PA	Parallels Publishing Agent
RAID	Redundant Array of Independent Disks
RAS	Parallels Remote Application Server
RDSH or RD Session Host	Remote Desktop Session Host
SAN	Storage Area Network
SCG	Parallels RAS Security Client Gateway
SSD	Solid State Disk
VDI	Virtual Desktop Infrastructure
VM	Virtual Machine
VLAN	Virtual Local Area Network
vSAN	Virtual Storage Area Network

Disclaimer

This project was deployed in a dedicated testing environment. It is probable that environment related variables not considered in this document may exist and will affect results. Thus, it is highly recommended to use this document for guidance purposes only and to conduct your own dedicated proof-of-concept (POC) for more accurate results before implementing this solution in your production environment.

CHAPTER 3

Environment Overview

This section provides information on the hardware and software used along with configurations carried out to build the Parallels RAS lab environment.

In This Chapter

Hardware

The following table highlights the hardware used:

Node	Model	Specs	Qty	Usage/roles
1 - 3	SuperMicro SuperChassis 815TQ-600CB	2x Intel Xeon E5-2680 v2 Cores: 2x 10x 2.80 GHz (Dual 10 Core) 128 GB RAM 3x Western Digital Blue 1 TB SSD	3	Parallels RAS infrastructure components (PAs and GWs) and Microsoft Services
4 - 54	HP ProLiant DL360p Gen8	2x Intel Xeon E5-2670 / Cores: 2x 8x 2.60 GHz (Dual Eight Core) 128 GB RAM 3x Western Digital Blue 1 TB SSD	50	Parallels RAS RD Session Hosts
55	Generic standalone host	Intel Xeon E5-2620 v2 RAM: 32 GB RAM SSD: 10* Western Digital Blue 1 TB SSD HW RAID 10	1	File share for UPDs*
N/A	Juniper EX3400	24 10/100/1000BASE-T ports. Four front-panel dual-mode (GbE/10GbE) small form-factor pluggable transceiver (SFP/SFP+) uplink ports and two 40GbE quad SFP+ (QSFP+) ports are also available for connecting the switches to upstream devices.	10	Network resources

Note*: As an alternative to UPDs, Microsoft FSLogix Profile Containers may have been used. For further details refer to https://docs.microsoft.com/en-us/fslogix/overview.

Software

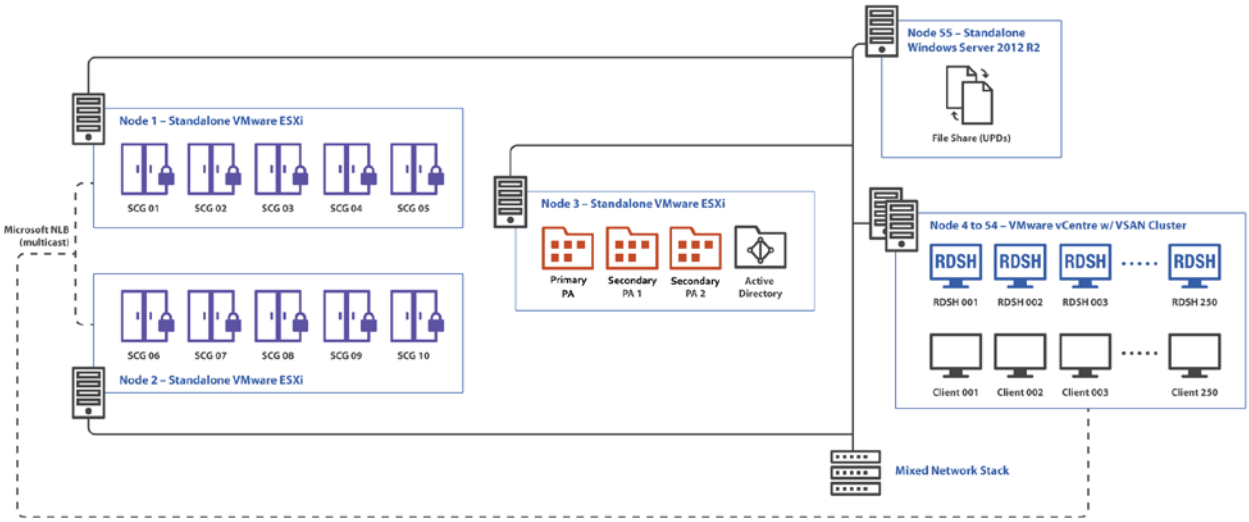
The following table highlights the software used:

Component	Software version/build
Application Delivery Server	Parallels Remote Application Server 17.0 (build 21289)
Application Delivery Client	Parallels Client for Windows (x64) 17.0 build (21282)
Hypervisor	VMware ESXi 6.7
Hypervisor Management	VMware vSphere 6.7
Network Load Balancing	Microsoft NLB Cluster*
Domain Services (DNS, AD Functional level)	Microsoft Active Directory Domain Services
NTP Server	Microsoft NTP Server
File Server for UPD	vCenter Cluster w/ vSAN Cluster Storage
Workload Generator	In house load testing tool was used for Parallels PA and SCG scalability testing. Login VSI (4.1.25) performance testing tool was used for RD Session Hosts scalability testing.
Workload Applications	Google Chrome web browser (inhouse test only) Internet Explorer 11 web browser (Login VSI test only) Microsoft Outlook 2013 & 2016 Microsoft Word 2013 & 2016 Microsoft Excel 2013 & 2016 Microsoft PowerPoint 2013 & 2016 Adobe PDF Reader DC Doro PDF Writer 7-Zip Windows Photo Viewer

Note*: Parallels RAS HALB may have been used instead of the Microsoft NLB cluster. Separate validation is required for this large-scale deployment.

Deployment

The following diagram depicts the testing environment platform setup:



Servers

As shown on the diagram above, node 1 to node 3 are standalone VMware servers hosting the RAS core infrastructure components as VMs including RAS PAs, SCGs and Microsoft services including AD DC, NTP, DNS and DHCP. Node 4 to node 54 are clustered VMware servers which are hosting the workload VMs which in this case are the RD Session Hosts configured with Parallels RAS RD Session Host Agent. Another standalone physical server was used to host the file share for UPDs. These have been illustrated in the table below.

Node	VMs	Roles	OS
1	5	Parallels SCG	Windows Server 2016 running on VMware ESXi 6.7
2	5	Parallels SCG	Windows Server 2016 running on VMware ESXi 6.7
3	5	3 Parallels PAs, 2 DCs (Inc. DNS, DHCP, NTP)	Windows Server 2016 running on VMware ESXi 6.7
4 to 54	250	RDSH	Windows Server 2016 running on VMware ESXi 6.7 (in vCenter)
55	1	FS for UPDs	Windows Server 2012R2

All internal client traffic was being generated by clients on the VMware vCenter cluster. The front-end traffic is then load balanced equally to the ten SCGs using a Microsoft NLB Cluster.

Note: In production environments, to ensure HA, it is recommended that Parallels RAS components are deployed on separate physical nodes considering possible hardware failures.

Storage

Both the HP ProLiant and the Supermicro Serves were equipped with local disks which were used for both local and shared services. For the latter, VMware vSAN Cluster Storage was used.

VMware vSAN is a distributed layer of software that runs natively as a part of the ESXi hypervisor. The local disks of the nodes were used to create a single storage pool shared across all hosts in the vSAN cluster. This was built on 50 nodes, referred to as nodes 4 to 54, each with two SSDs used as part of the vSAN Cluster. One disk was assigned for shared storage and the other disk used for cache.

Network

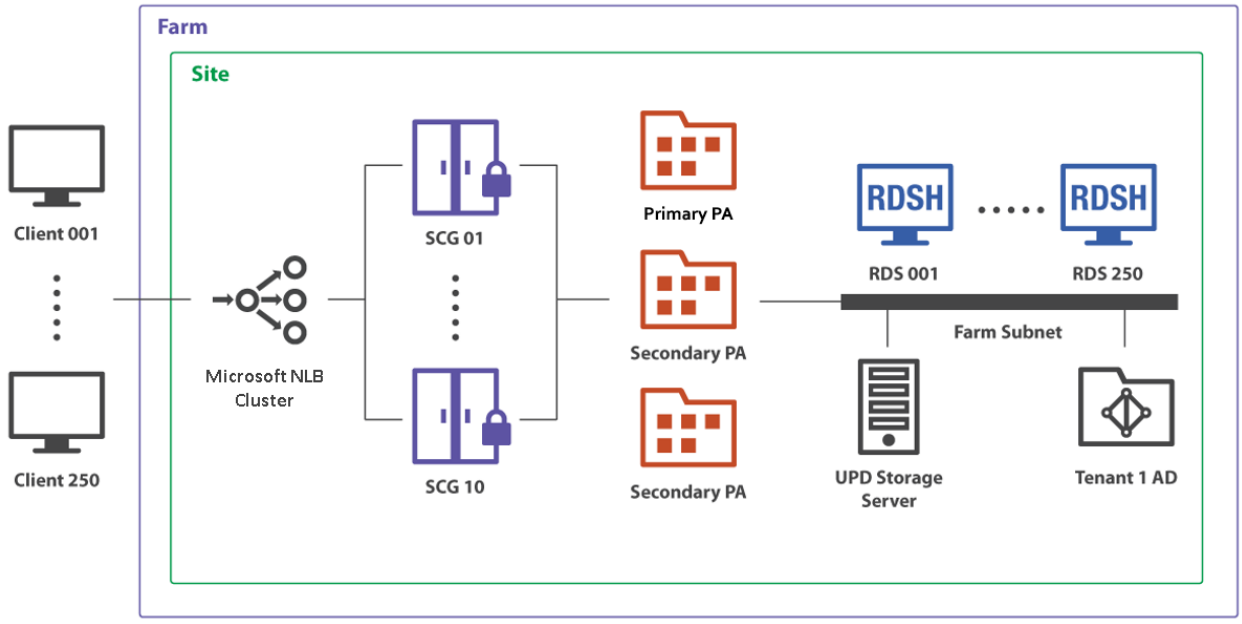
In production environment, it is recommended that network segmentation based on VLANs for different networks such as DMZ, Production, Storage is used.

For the scope of this document, a flat network was used. All physical nodes were connected to a single network stack while all VMs resided in a flat /16 subnet with 255.255.0.0 netmask with 65536 available addresses.

Note: A smaller subnet such as /19 may have also been used.

Parallels RAS Farm

The following logical diagram provides an overview of Parallels RAS Farm deployment:



Virtual Machines Specifications

Parallels RAS Component	No. of VMs	vCPU	RAM (GB)
Publishing Agent	3	4	8
Secure Client Gateway	10	2	8
RDSH Host	250	4	14

Configurations

The below Parallels RAS configuration was carried out:

- Default configuration for RAS PAs and SCGs was used.
- UPDs were configured with default max size of 20 GB for each UPD.
- UPD storage was configured as dynamically allocated*.
- Published resources included both applications and desktops.
- Doro PDF Writer was configured as the default printer via RAS client policies.
- Clients connected to the Parallels RAS environment by pointing to the Microsoft NLB address, which load balanced the traffic equally to different RAS SCGs.
- Client connections were configured in Gateway SSL mode to check scalability under constant utilization of SCG resources.

Note*: Results show that the average user profile was around 600 MB for each user.

For internal users, Direct, Direct SSL or Gateway connection mode may have also been used. Please refer to the Parallels RAS Administrators Guide for more information https://www.parallels.com/eu/products/ras/resources/.

CHAPTER 4

Test Methodology

The primary focus of testing was to validate Parallels RAS core components under heavy load in a large-scale deployment of 5000 concurrent user sessions. In order to do so, the Parallels RAS core components were isolated from RD Session Hosts and Parallels Clients, residing on separate hardware nodes. This was to ensure that Parallels RAS core components' performance is not affected by other VMs. To generate required traffic, a Parallels in-house tool was used simulating users' logon and session workload. Tests carried out were divided into two main phases, logon phase and workload phase, where corresponding metrics were noted during logon and workload playback respectively.

Test metrics were gathered from all RAS PAs and SCGs to ensure that their resource consumption under intended load does not exceed an average of 80% during peak load.

Following this, a separate test using Login VSI was also carried out, this time directly on RD Session Hosts, with the aim to provide performance related metrics with regards to RD Session Host scalability. This was done by looking at different RD Session Hosts hardware specifications and the number of user sessions able to be hosted. Since the goal of this testing was to capture a baseline reflecting the possible user densities, Login VSI client launchers were configured to go through RAS Secure Client Gateway in Gateway SSL mode.

In This Chapter

Test Overview

The following test procedure was used for each test run to ensure consistent results:

1 Preparations

- Before each test, all workload VMs and clients were cleanly started.
- All workload VMs and client launchers were put to Idle for an hour for consistent results.

2 Logon Phase

- All user sessions were launched and stayed active in a one-hour time window.

- 15 user sessions were launched every 10 seconds, thus launching an average of 1 session every 1.5 seconds.

3 Workload Phase

- Total session number (5000) was maintained for 72 hours.
- In-session workload playback for Office Worker was repeated every 3 hours.
- After each workload playback, sessions were logged off and logged back in.

4 Data collection

- All user sessions were logged off after the workload phase.
- Test run reports and data were generated.
- All workload VMs and clients were shut down.

Success Criteria

The Parallels RAS core component testing success criteria was as following:

- RAS PAs and SCGs are not overloaded by targeted number of sessions with an average CPU utilization lower than 80% and adequate memory available.
- All sessions are to be launched successfully, reaching the RD Session Hosts as expected.
- All sessions are to be gracefully logged off successfully on the logoff command.
- No service failures detected during deployment, logon, workload, and peak load and logoff phases.

The RD Session Host sizing success criteria was as following:

- Identify resource utilization of a larger RD Session Host to host at least 75 users based on the Office Worker workload profile without user experience degradation.
- Identify resource utilization of a smaller RD Session Host to host at least 25 users based on the Office Worker workload profile without users experiencing degradation.

In-session Workloads

Parallels developed its own in-house load simulation tool which was used to simulate users' workload during this test. Specifically, this tool was designed to simulate in-session user activity based on the Office Worker workload. The simulated user tasks include working with the following applications:

- Google Chrome web browser
- Microsoft Outlook
- Microsoft Word

- Microsoft Excel
- Microsoft PowerPoint
- 7-Zip
- Doro PDF Writer
- Adobe PDF Reader DC
- Windows Photo Viewer

In addition, with respect to the performance test, Login VSI 4.1.32 was used to simulate the Office Worker workload based on 25 and 75 users on RD Session Hosts with different hardware specifications.

The Login VSI Office Worker workload includes working with the following applications:

- Internet Explorer 11 web browser
- Microsoft Outlook
- Microsoft Word
- Microsoft Excel
- Microsoft PowerPoint
- 7-Zip
- Adobe PDF Reader DC
- Windows Photo Viewer

While being diverse and not focused on one or two applications, the Office Worker workload does not place a very severe demand on the environment and represents users that do not overload the system with heavy tasks.

Data Capture

RAS Performance Monitor and Login VSI reports were used to capture windows performance counters from all nodes hosting RAS components.

CHAPTER 5

Findings

This section of the document presents both the load and performance test results achieved during various cycles of a user session, including logon and workload phases. The user experience performance results on RD Session Hosts based on the user's session workload are also presented on differently sized hosts.

In This Chapter

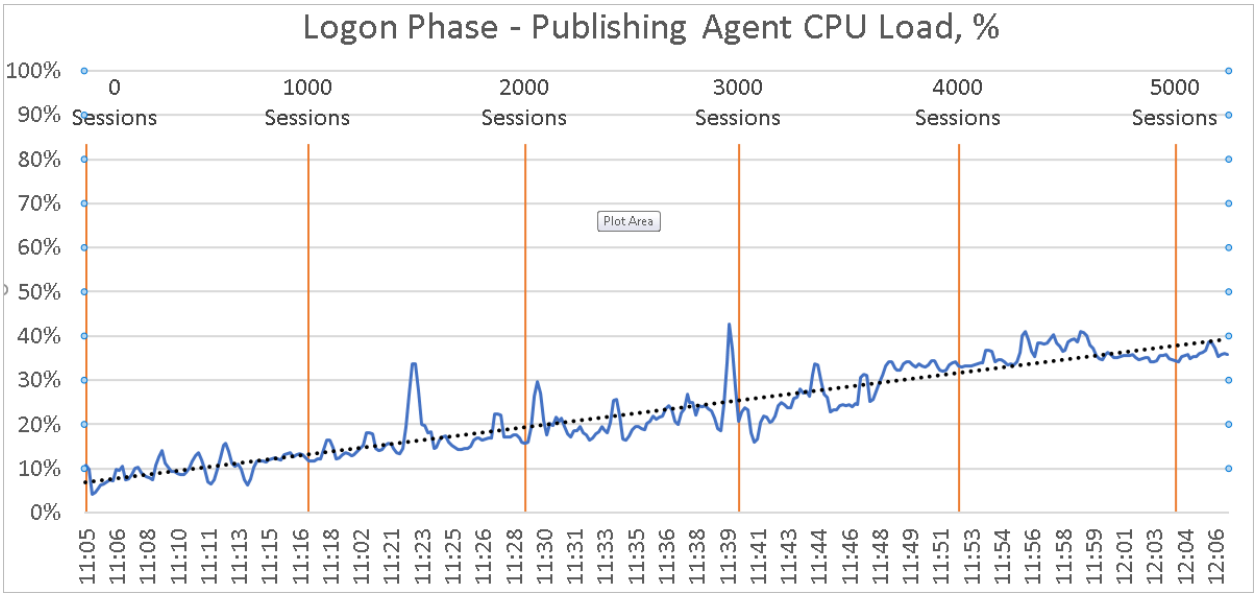
Logon Phase

For the scope of this validation, the logon phase duration has been configured to take under one hour. This means that the user logon rate was set to around 15 sessions in every 10 seconds or 1.5 sessions per second.

During this phase the hardware resources of the RAS core infrastructure components, such as PAs and SCG, have been monitored and presented in the following graphs.

Publishing Agents

The below graph shows the average CPU % utilization of the three PAs in the RAS environment with respect to the number of sessions being launched during the logon phase. Load has been equally distributed to all PAs, that is, one primary and two secondary PAs.

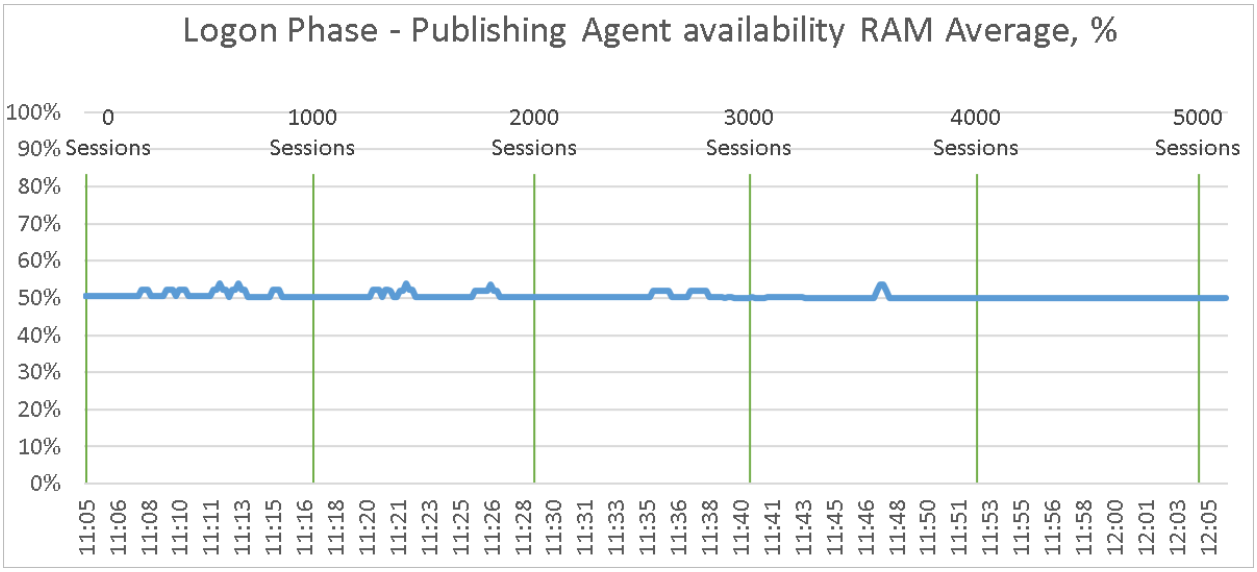


The below table provides a snapshot of the estimate average PA CPU % utilization at different intervals during the Logon stage.

No. of Users	PAs average CPU utilization (%)	Delta
0	7	—
1000	12	5
2000	19	7
3000	24	5
4000	33	9
5000	35	2

It can be noted that the average PAs CPU utilization is on a steady increase in-line with the increase of user sessions as expected. CPU utilization shows that the CPU specifications were well adequate for the load generated with an average of 23% CPU utilization and a maximum of 47% CPU utilization.

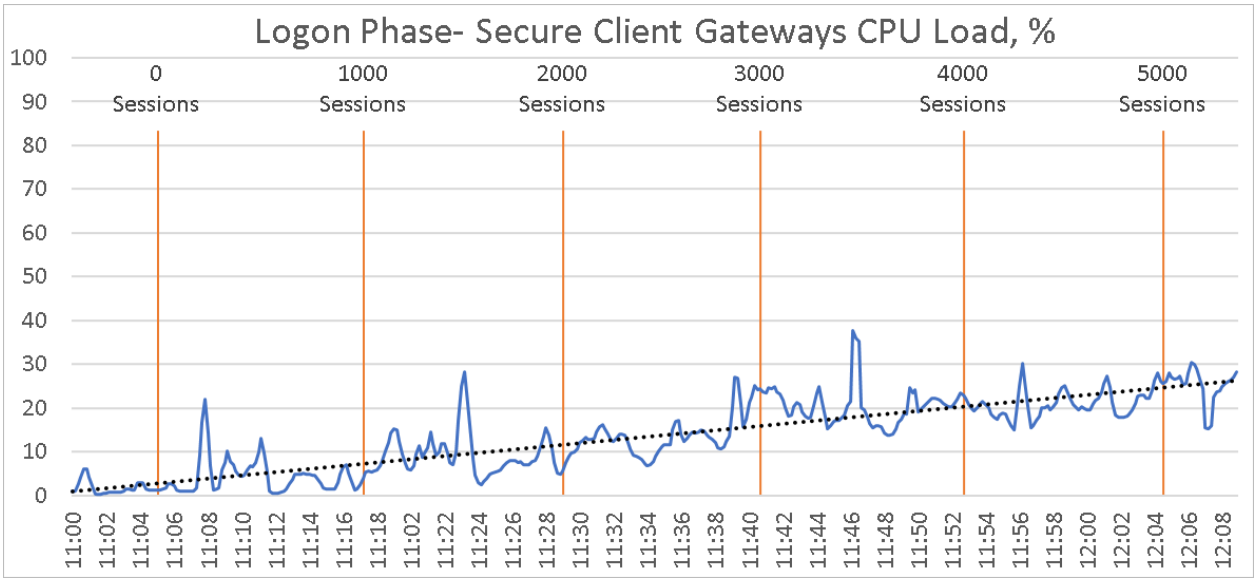
The below graph shows the average memory (RAM) % availability of the three PAs in the RAS environment with respect to the number of sessions being launched during the logon phase. Load has been equally distributed to all PAs, that is, one primary and two secondary PAs.



It can be noted that the average PAs memory (RAM) availability is adequate at a consistent average of 50% available memory even when reaching the 5000 users session mark. This shows us that the number of sessions only have a slight impact on the memory utilization.

Secure Client Gateways

The below graph shows the average CPU % utilization of the ten SCGs with respect to the number of sessions being launched during the logon phase. The load has been equally distributed to all SCGs using Microsoft NLB until a peak of 500 sessions per SCG.

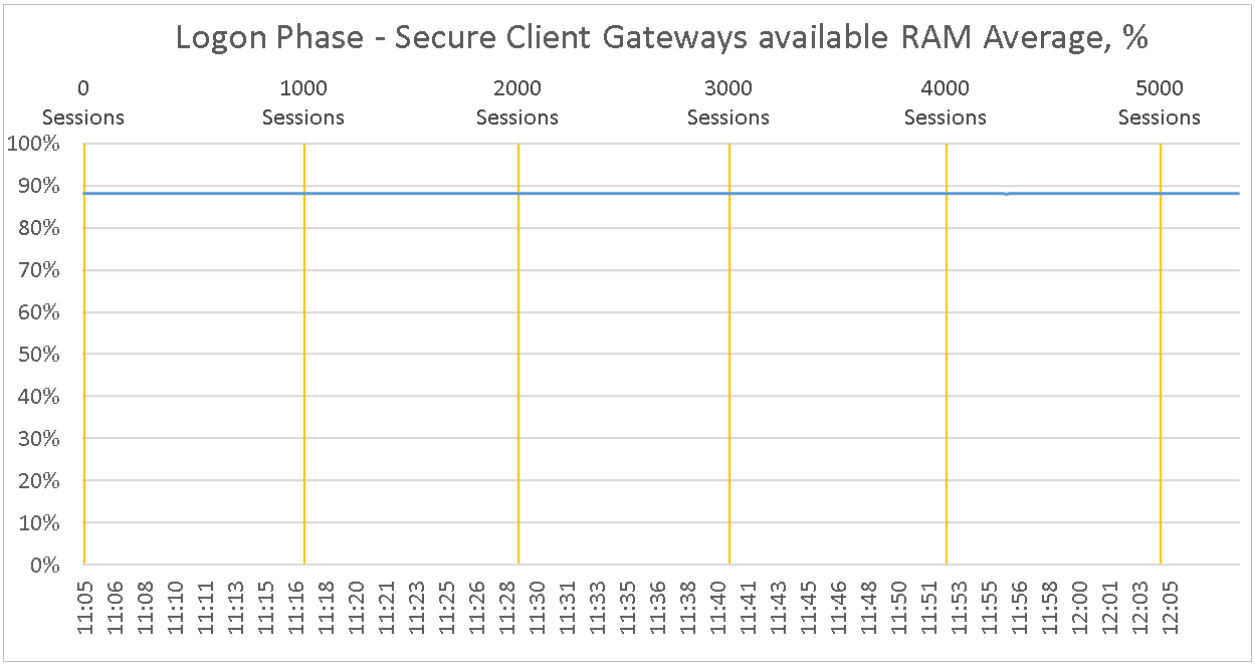


The below table provides a snapshot of the estimate average SCG CPU % utilization at different intervals of the Logon stage.

No. of Users	SCGs average CPU utilization (%)	Delta
0	2	—
1000	6	4
2000	10	4
3000	18	8
4000	21	3
5000	26	5

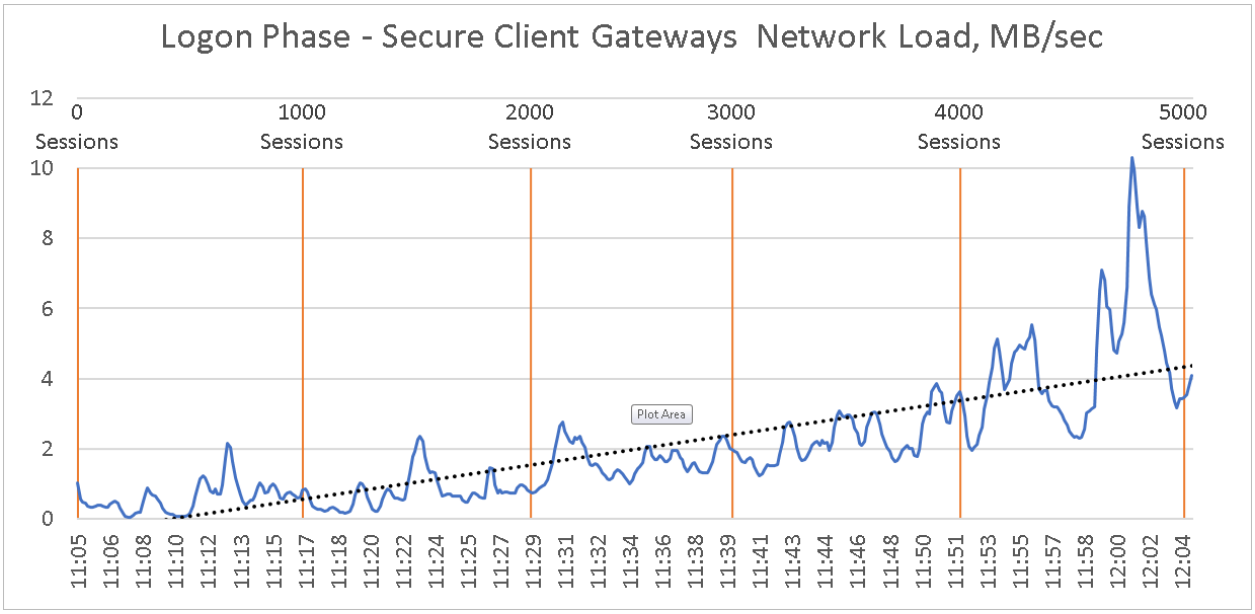
It can be noted that the average SCGs CPU utilization is on a steady increase in line with the increase of user sessions as expected. CPU utilization shows that the CPU specifications were well adequate for the load generated with an average of 14% CPU utilization and a maximum of 38% CPU utilization.

The below graph shows the average memory (RAM) % availability of the ten SCGs with respect to the number of sessions being launched during the logon phase. The load has been equally distributed to all SCGs using Microsoft NLB until a peak of 500 sessions per SCG.



It can be noted that the average SCG memory (RAM) availability is more than adequate at a consistent average of 88% available memory even when reaching the 5000 users session mark. This shows us that the number of sessions only have a slight impact on the memory utilization.

The following graph shows the average network transfer rate through the SCGs during logon phase. As expected, the network utilization increases in line with the number of user session logons. The highest maximum transfer rate was noted at 10.32 MB/s which translates to 21.1 KB/s per user during logon phase.



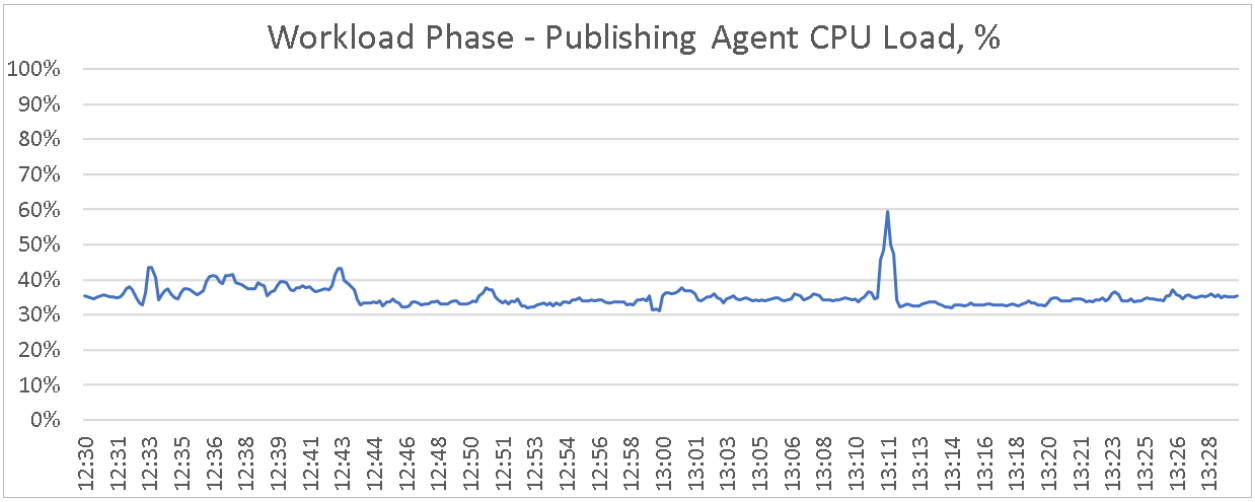
Workload Phase

For the scope of this validation, the workload phase consisted of 5000 user sessions which were maintained for 72 hours. During this phase, each in-session workload has been configured to run for 3 hours. Following this, a graceful session logoff was carried out only to carry out another session logon for another 3 hours of workload playback. This process has been repeated 24 times to reach the 72-hour workload phase testing.

The graphs presented below are a sample of a duration of one hour from the total duration of workload testing.

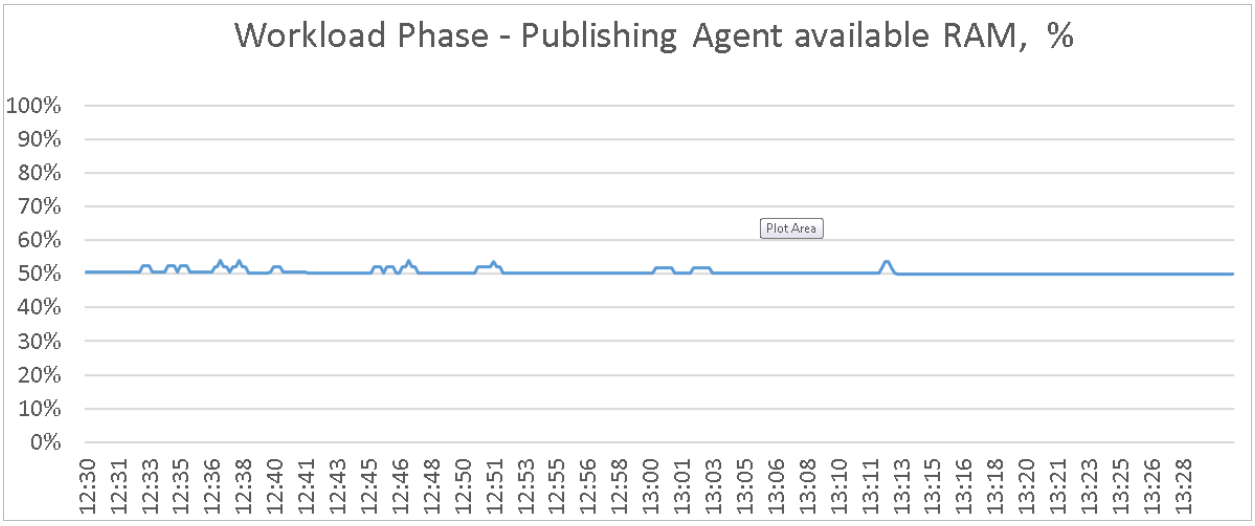
Publishing Agents

The below graph shows the average CPU % utilization of the three PAs in the RAS environment during the workload phase. The load has been equally distributed to all PAs, that is, one primary and two secondary PAs.



It can be noted that the average PAs CPU utilization is steady at an average of 35% and a peak of 60% CPU utilization. This shows that the CPU specifications were well adequate for the workload of 5000 concurrent sessions.

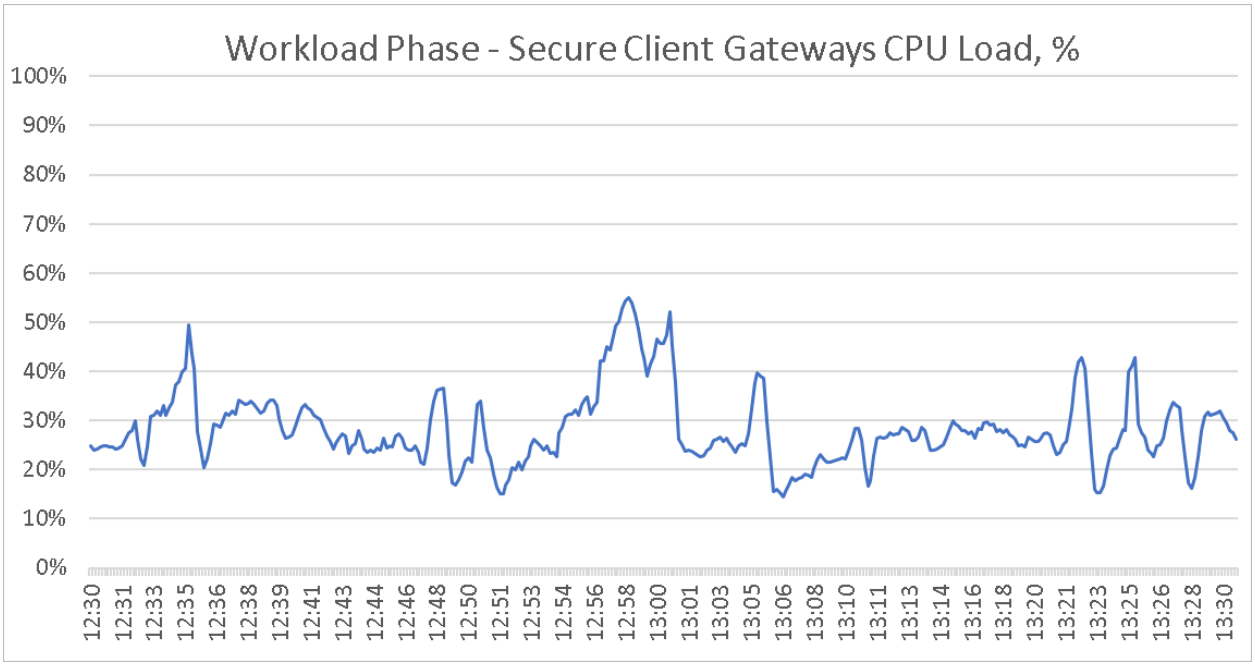
The below graph shows the average memory (RAM) % availability of the three PAs in the RAS environment during the workload phase. The load has been equally distributed to all PAs, that is, one primary and two secondary PAs.



It can be noted that the average PAs memory (RAM) availability is adequate at a consistent average of 50% available memory during the workload phase. This shows us that the number of sessions only have a slight impact on the memory utilization.

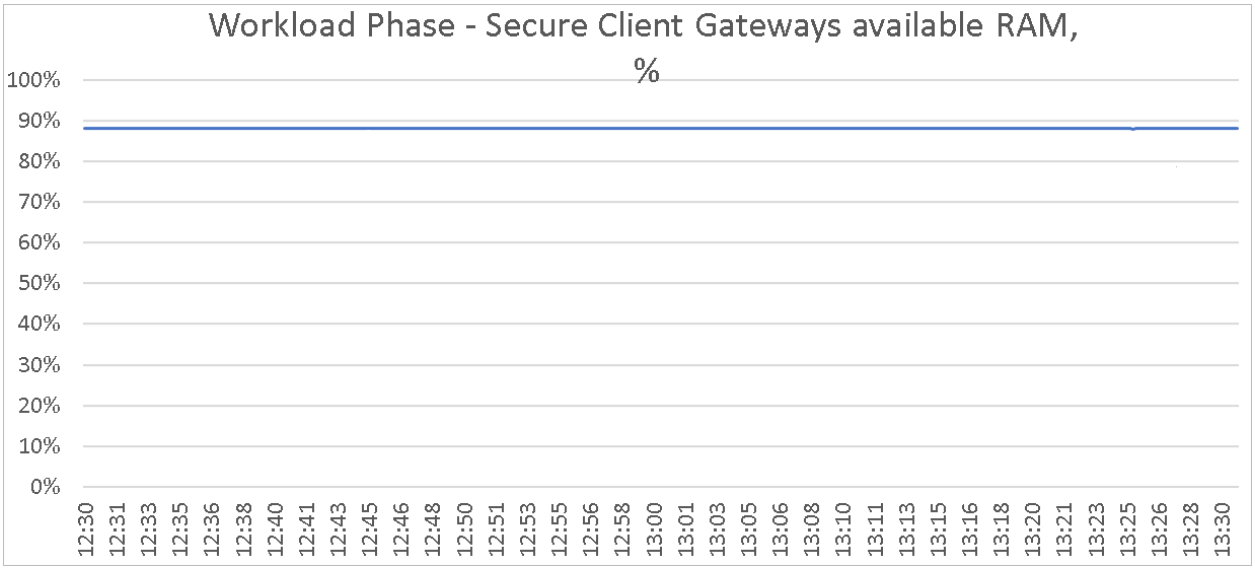
Secure Client Gateways

The below graph shows the average CPU % utilization of the ten SCGs in the RAS environment during the workload phase. The load has been equally distributed to all SCGs using Microsoft NLB.



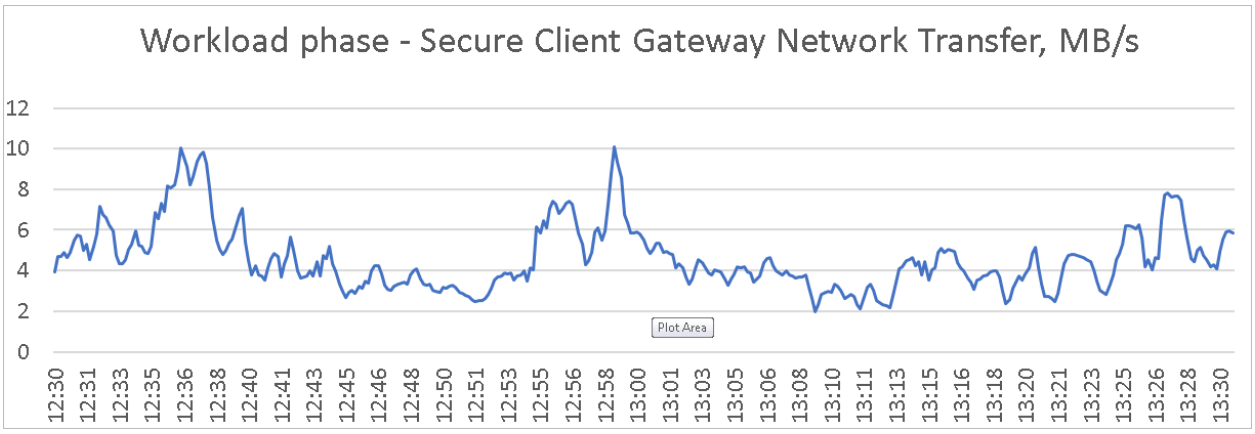
It can be noted that the average SCG CPU utilization varies from a maximum of 55% to a minimum of 15%. The average CPU utilization was noted to be at 28%. As such it can be concluded that the CPU specifications were well adequate for the workload of 5000 concurrent sessions across ten gateways with an average of 500 users per gateway.

The below graph shows the average available memory (RAM) % of the ten SCG in the RAS environment during the workload phase. Load has been equally distributed to all SCGs using Microsoft NLB.



It can be noted that the average SCG memory (RAM) availability is adequate at a consistent average of 88% available memory during the workload phase. This shows us that the number of sessions only have a slight impact on the memory utilization.

The following graph shows the average network transfer rate through the SCGs during workload phase. The average network throughput was recorded at 4.58 MB/s on each SCG for a total of 45.8 MB/s when considering all ten SCGs, which translates to 9.38 KB/s per user. The maximum transfer rate was noted at 11.64 MB/s or 23.8 KB/s per user during workload phase.



However, it should be noted that both SCG CPU and network usage during workload phase is highly dependent on in-session activity. For instance, video playback increases both network and CPU usage, while working with more ‘static’ applications, requiring less screen updates, such as Microsoft Word uses less SCG resources.

CHAPTER 6

RD Session Hosts Sizing Considerations

Following the validation of Parallels RAS core components such as PAs and SCGs in a 5000-user environment as previously highlighted, the focus was shifted on the RD Session Hosts. The test environment was based on 250 RD Session Hosts each with 4 vCPUs and 14 GB RAM hosting 20 users for a total of 5000 users.

However, when it comes to planning and design, sizing the RD Session Hosts is dependent on several factors including hardware specifications, application requirements, users' workload profile, network configuration, system availability and risk tolerance (or mitigation).

It is thus crucial to analyze these requirements and business expectations prior to confirming the sizing of your RD Session Hosts to be used in a production environment.

Nonetheless, there are two basic strategies that can be considered to meet the required performance levels which are scaling up or scaling out. Scaling up refers to using larger hardware for each RD Session Host to serve more users while scaling out is based on using larger number of smaller RD Session Hosts that share the users' workload.

For the scope of this document, Login VSI was used on differently sized RD Session Hosts to provide performance results of 75 user sessions and 25 user sessions as an example of scaling up and scaling out respectively. Environment details and results are shown in the subsections below.

In This Chapter

Test Methodology

Testing was carried out on differently sized VMs promoted to RAS RD Session Hosts running on Microsoft Windows Server 2016. The specifications for each RD Session Host are provided in the following table:

Test	No. of user Sessions on each RDSH	vCPUs	RAM (GB)
Scale up	75	12	64
Scale out	25	4	14

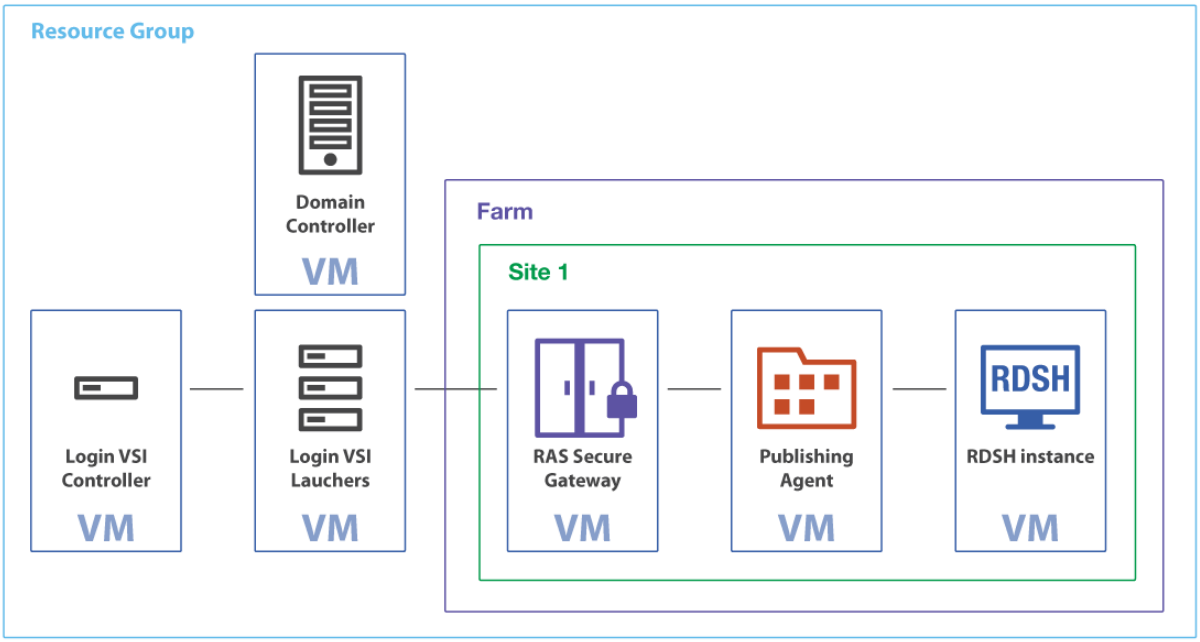
Performance metrics were captured during user logon, user workload execution (steady state), and user logoff. To achieve consistent measurements that would reflect when components were appropriately cached, each workload ran for 48 minutes before Login VSI performance metrics were recorded. Login VSI tests were repeated three times on each workload VM to get an average number of users who successfully ran the test.

Configurations

For Parallels RAS 17 scalability testing, the infrastructure VMs were configured with Microsoft Windows Server 2016 as follows:

- Five Infrastructure VMs for Login VSI environment:
  - 1x VM as Login VSI controller and profile server
  - 4x VMs as Login VSI launchers
- Two Infrastructure VMs for Parallels RAS environment:
  - 1x VM as RAS Publishing Agent
  - 1x VM as RAS Secure Client Gateway
- One Infrastructure VM for Microsoft related services:
  - 1x VM as Active Directory Domain Controller and DNS server Scalability
- Two workload VMs where used to test the Parallels RAS user session workloads. Each VM was configured as follows
  - RAS RD Session Host Agent 17.0 (build 21289)
  - Microsoft Office 2016
  - Latest Windows updates available at the time of testing
  - Local users' profiles
  - Out of the box Windows settings were used with no specific Windows optimizations carried out on the VMs

The below high-level logical diagram shows the architecture and components used for the Parallels RAS scalability testing. Users connect through Parallels 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.



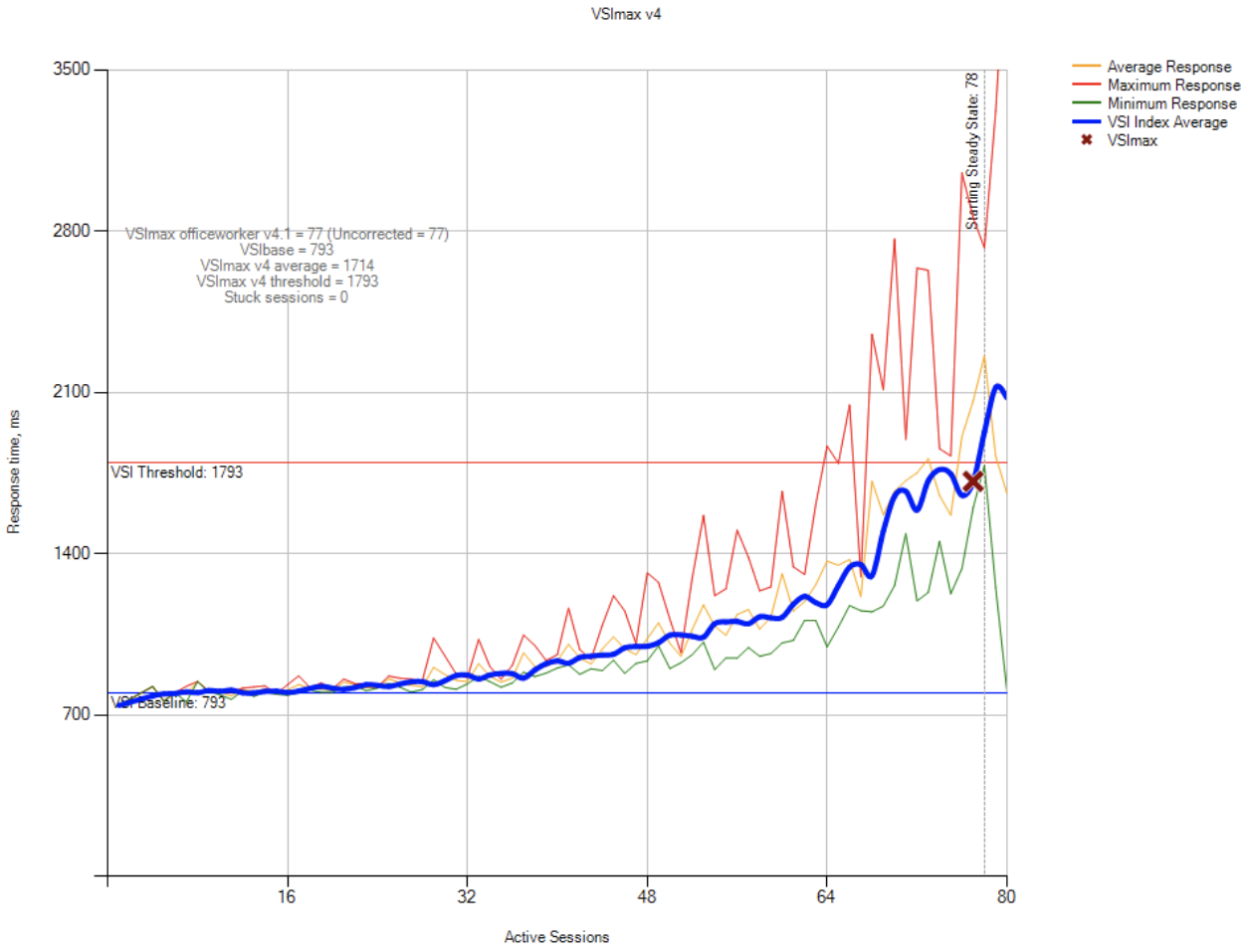
Findings

This section highlights the Login VSI results attained on differently sized RD Session Hosts to simulate scaling up to 75 user sessions on each RD Session Host and scaling out with 25 user sessions on each RD Session Host.

VSImax v4.1, 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.

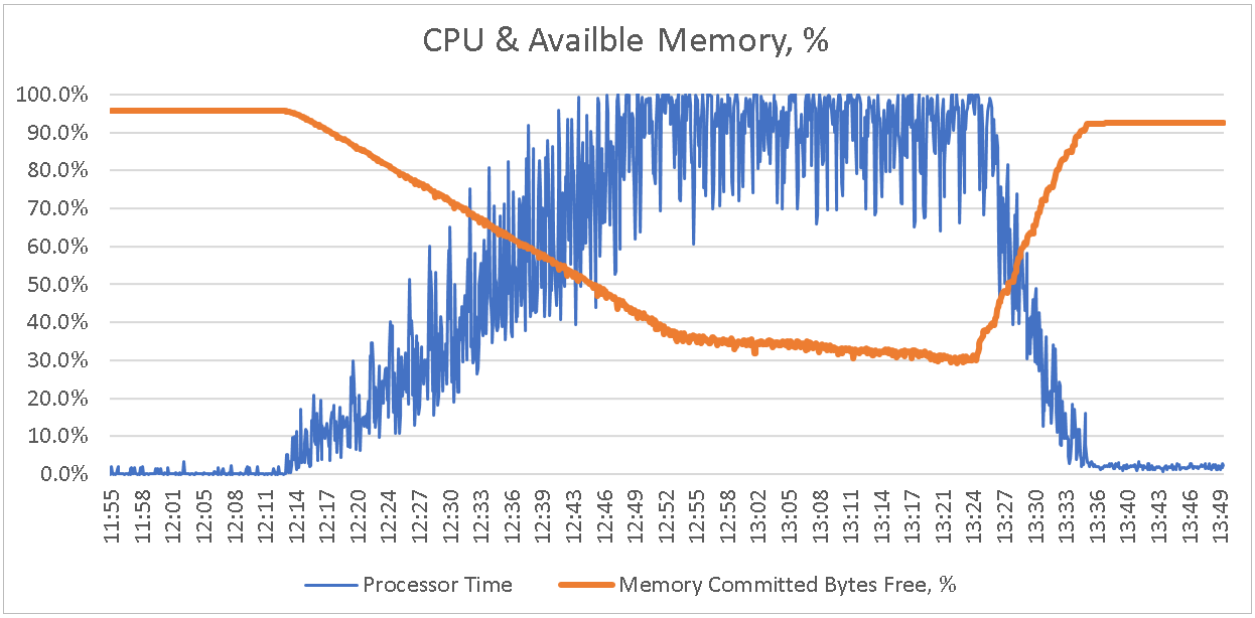
Scale Up

This section describes test results for the Office Worker workload on an RD Session Host designed to handle 75 user sessions. The RD Session Host instance (with 12 vCPUS and 64 GB RAM) shows a VSImax density of 77 users running the Office Worker workload.



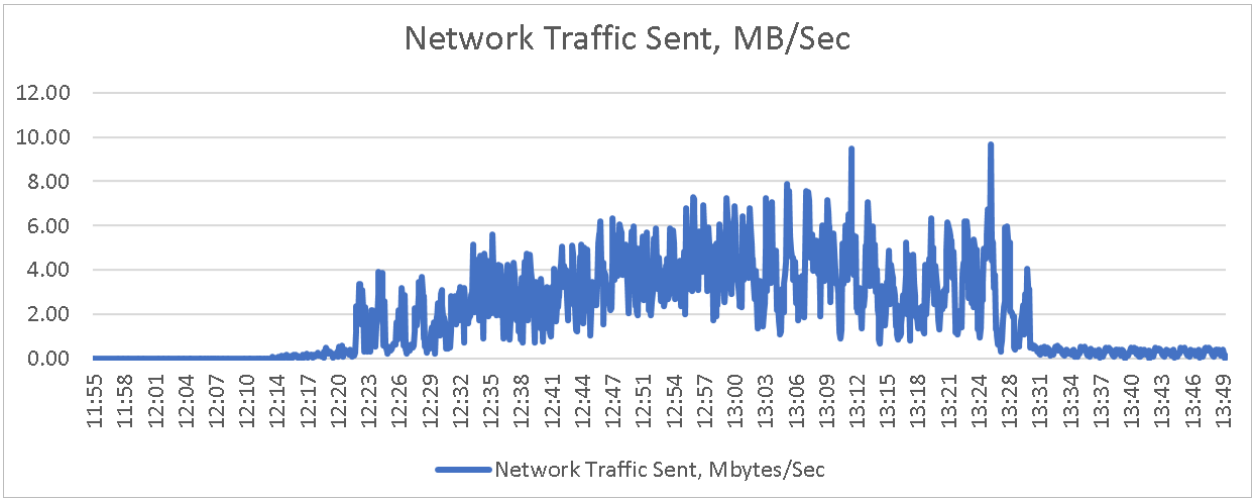
CPU and RAM

The following results show the CPU usage and available memory which are helpful in evaluating performance under the Office Worker workload. In this chart it can be noted that as user load increases towards the maximum, CPU reaches its maximum. As expected, memory consumption is increased, however enough memory is available even at peak usage with around 30% remaining available memory.



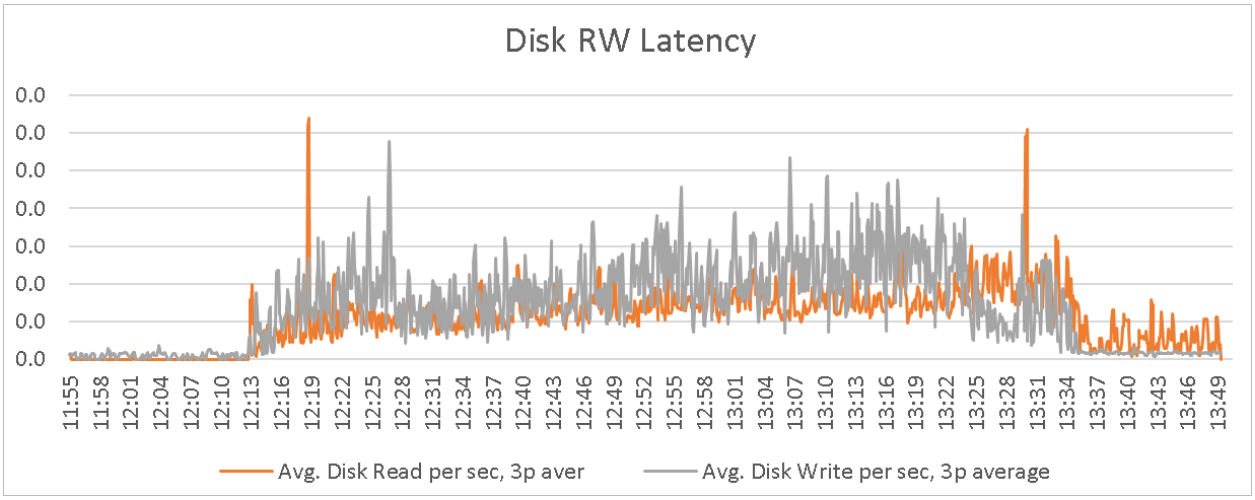
Network

For the Office Worker workload, the average outbound bandwidth during steady state is approximately 3.83 MB/s for the test workload of 77 users. Therefore, the outgoing transfer rate per user is approximately 51.0 KB/s (3.83 MB/s /77 = 51.0 KB/s). Outgoing network transfers during logoff occur as user profile data is transmitted.

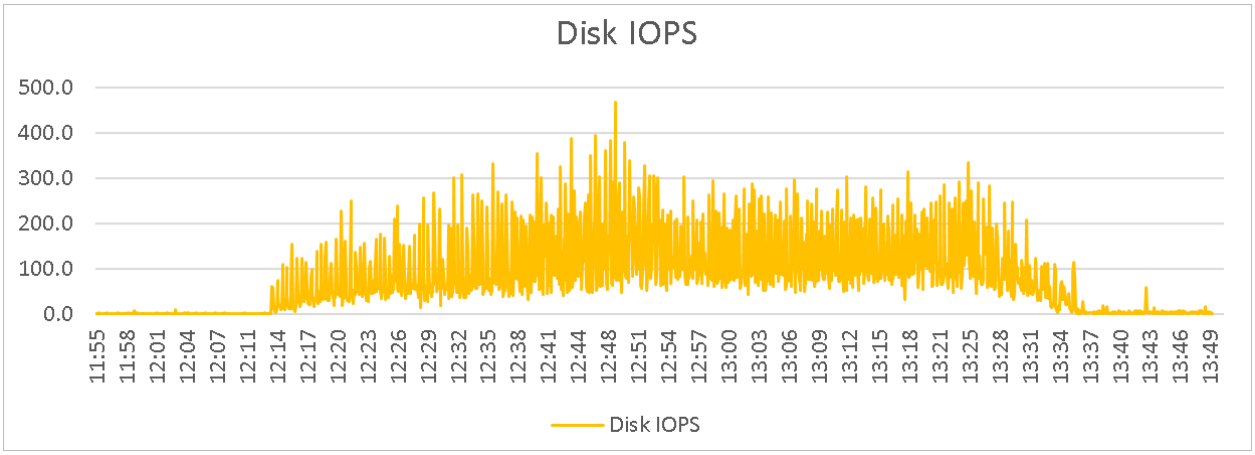


Disk

Disk I/O response time metrics for the Office Worker workload are shown below. The write I/O response time averaged around 2.4 milliseconds (ms) while the read I/O response times averaged around 1.53 milliseconds (ms).

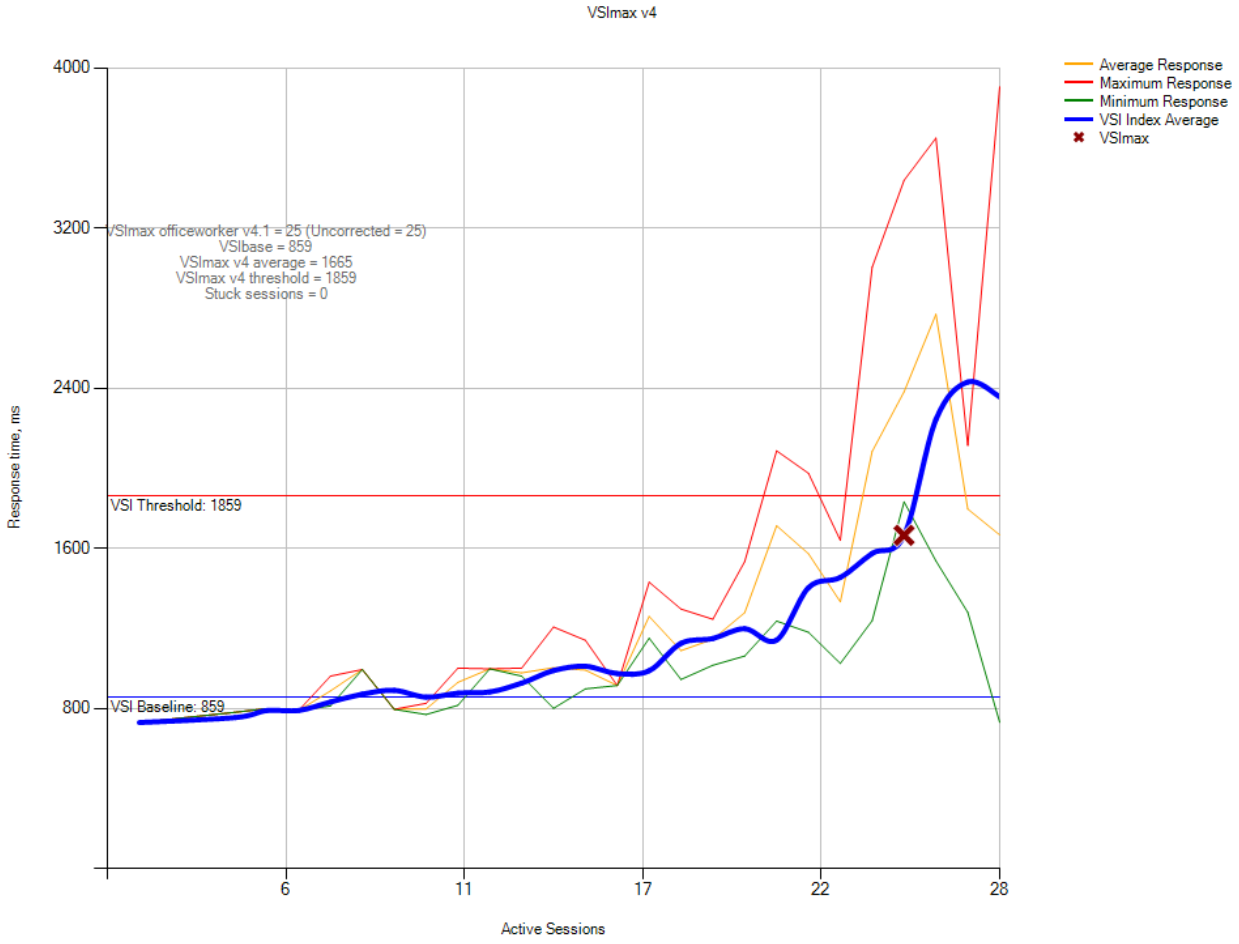


The graph below shows disk transfer metrics. For the Office Worker workload, disk transfers during steady state averaged about 134 IOPS for the test group of 77 users, or about 1.74 IOPS per Office Worker user. The peak value was 468 IOPS for 77 users or about 6.1 IOPS per Office Worker user. Disk transfer activity is also visible during the logoff period as user profile data is recorded.



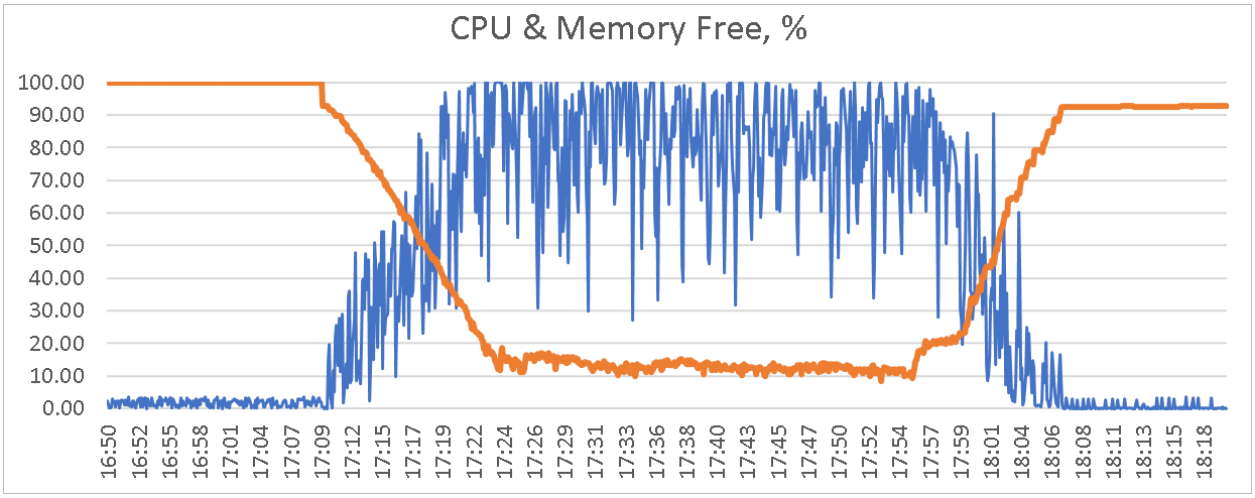
Scale Out

This section describes test results for the Office Worker workload on an RD Session Host designed to handle 25 user sessions. The RD Session Host instance (with 4 vCPUS and 14 GB RAM) shows a VSImax v4 density of 25 users running the Office Worker workload.



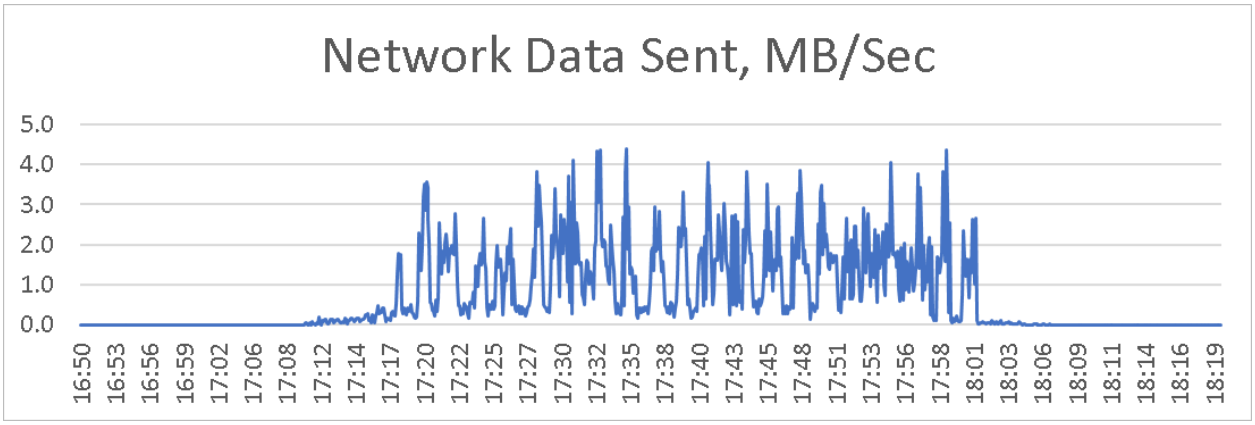
CPU and RAM

The following test results are for the CPU and available memory which are helpful in evaluating performance under the Office Worker workload. In this chart it can be noted that as user load increases towards the maximum, the CPU usage peaks. During peak load, memory is nearly depleted with around 90% memory consumption.



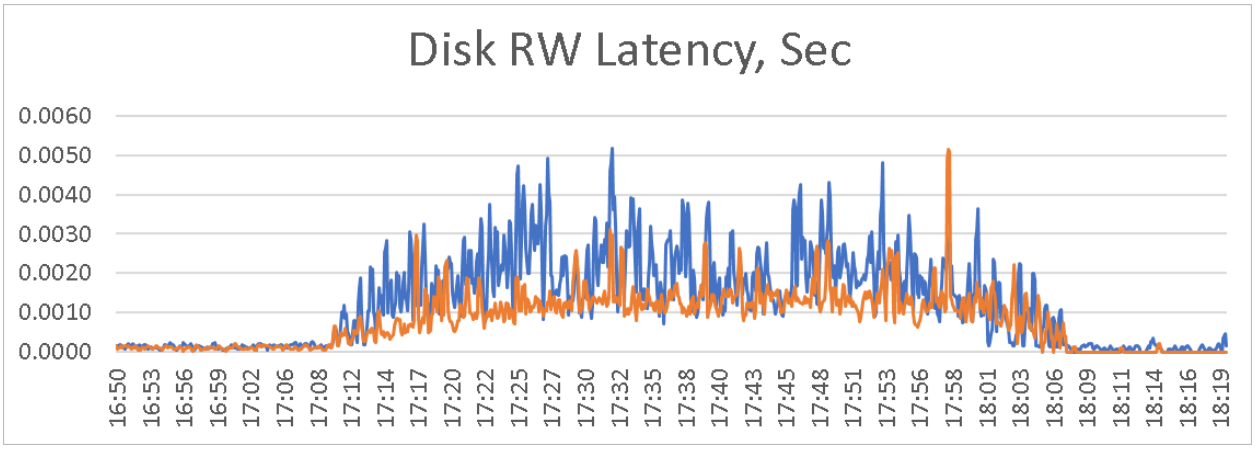
Network

For the Office Worker workload, the average outbound bandwidth during steady state is approximately 1.4 MB/s for the test workload of 25 users. Therefore, the outgoing transfer rate per user is approximately 57.3 KB/s (1.4 MB/s /25 = 57.3 KB/s). Outgoing network transfers during logoff occur as the user profile data is transmitted.

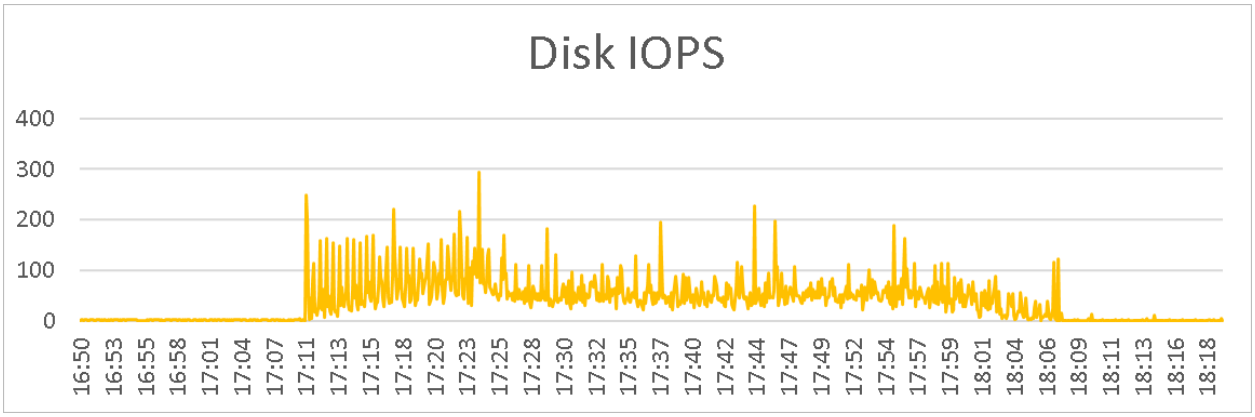


Disk

Disk I/O response time metrics for the Office Worker workload are shown below. The write I/O response time averaged around 2.21 milliseconds (ms) while the read I/O response times averaged around 1.38 milliseconds (ms).



The graph below shows disk transfer metrics. For the Office Worker workload, disk transfers during steady state averaged about 57 IOPS for the test group of 25 users, or about 2.28 IOPS per Office Worker user. The peak value was 294 IOPS for 25 users or about 11.76 IOPS per Office Worker user. Disk transfer activity is also visible during the logoff period as the user profile data is recorded.



CHAPTER 7

Conclusion

This document presented a successful validation of Parallels RAS setup in a large-scale environment hosting over 5000 user sessions utilizing hosted applications and desktops.

Parallels RAS deployment consisted of three Publishing Agents and ten Secure Client Gateways in a single Site in a single Farm. Although for the load test, a total of 250 RD Session Hosts were configured in the Parallels RAS environment, the number of RD Session Hosts may vary depending on the user's workload and applications' resources requirements, among other variables. As such, both scale up and scale out approaches were presented. In case of low to medium workload users, when most of user sessions run the same applications and perform the same tasks, one might consider scaling up and deploying larger but fewer RD Session Hosts that can host a larger amount of sessions on each host. On the other hand, in case of heavier users' sessions workload or in case where a solution requires to lower and distribute the downtime risk among a larger number of hosts, the scale out approach may be considered.

It is important to note that while load and scalability testings are key factors in understanding how a platform and the overall solution performs, the results obtained and presented in this document should not be inferred as an exact measurement for real-world production workloads. It is advised for customers looking to better assess how applications will perform, conduct their own load and scalability testing with their own workload samples. Additionally, Parallels RAS proof of concept (POC) or pilot can be requested to assist in design, deployment, and sizing prior to moving into production.

Parallels RAS, considered a major player in its market, as highlighted in the IDC MarketScape: Worldwide Virtual Client Computing 2019-2020 Vendor Assessment Report (https://vdi.parallels.com/IDC-Report), can provide organizations with an all-in-one application delivery and VDI solution. Providing users with secure access to virtual workspaces from anywhere, on any device, anytime, Parallels RAS empowers organizations to embrace digital transformation by centralizing management of the IT infrastructure, streamlining multi-cloud deployment, reinforcing data security and improving process automation. Its flexible and scalable architecture allows organizations to quickly adapt to business demands while overcoming many common challenges.

For further information about Parallels RAS, features, and benefits, please visit https://parallels.com/ras.