# Parallels RAS Scalability Testing with Login VSI

1000 Users - Knowledge Worker

White Paper | Parallels Remote Application Server | 2020

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

# Contents

CHAPTER 1

# Introduction

Parallels Remote Application Server (RAS) is a comprehensive virtual application and desktop delivery solution that allows your employees to use applications and data from any device. Seamless and easy to deploy, configure, and maintain, Parallels RAS supports the delivery of applications and desktops via Microsoft RDS, Windows Virtual Desktop and major hypervisors.

This document presents an analysis of the scalability testing of Parallels RAS 17.1 using Login VSI for around 1000 user sessions on RD Session Hosts with Knowledge Worker workload.

CHAPTER 2

# Scalability

## In This Chapter



# Testing the Scalability of Parallels RAS

To validate Parallels RAS configurations, Parallels engineers conducted a series of performance tests. The goal was to analyze the scalability of Parallels RAS sessions running on VMware vSphere virtual machines. As part of this testing, Login VSI was used to generate user connections to RD Session Host servers simulating typical user workloads.

In a typical Parallels RAS deployment, users connect through a Parallels Client application to access remote applications and desktops. Login VSI clients simulate user connections, while RAS Publishing Agents distribute them and set up service connections between end users and RD Session Host servers.

# Configurations for Scalability Testing

For the Parallels RAS scalability testing, a total of 10 Supermicro X9DRi-LN4+ consisting of the following hardware components were used:

| | |
|---|---|
| CPU | 2x Intel Xeon E5-2695v2, 2.6GHz, 20 MB L3, 115W TDP |
| RAM | 128 GB, 16x 8 GB Micron DDR-4-2100 at 1600MHz |
| Storage | Host OS: 1 x 1 TB SATA 7.2k RPM<br>Workload VMs and infrastructure VMs: 2 x 1 TB SSD NO RAID |

## Components

| Component | Software version/build |
|---|---|
| Hypervisor | VMware ESXi 6.7 |
| Hypervisor Management | VMware vSphere 6.7 |
| Network Load Balancing | Parallels HALB Appliance 17.1.0 (build 21783) |
| Domain Services (DNS, AD Functional level) | Microsoft Active Directory Domain Services |
| NTP Server | Microsoft NTP Server |
| File Server for UPD | Microsoft SMB Server |

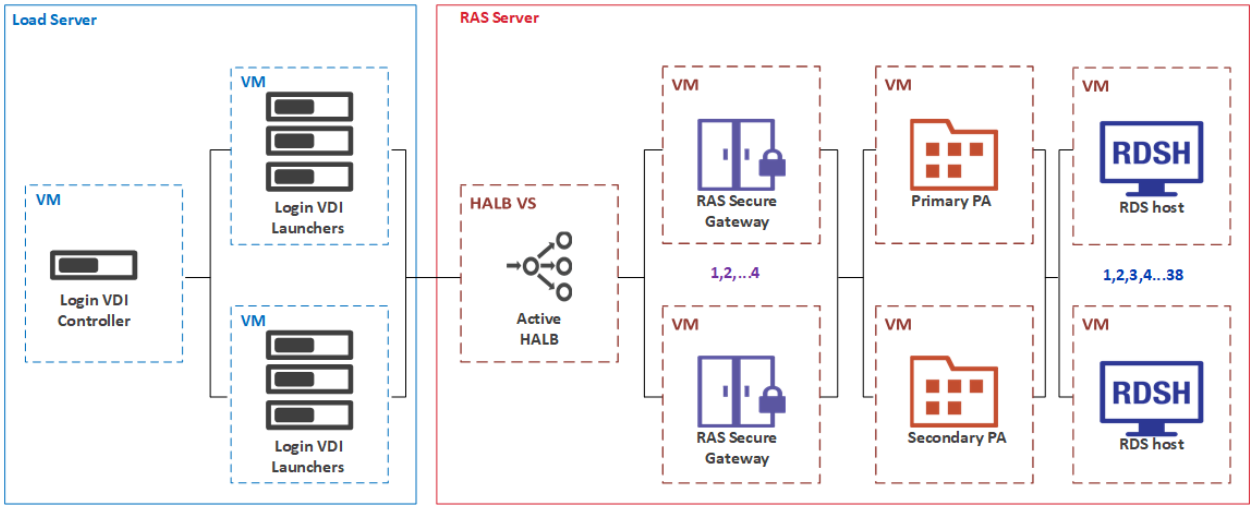
## Parallels RAS VM configuration

| Parallels RAS Component | Total VMs | vCPU in Each VM | RAM in Each VM |
|---|---|---|---|
| RAS Publishing Agent | 2 | 2 | 8 GB |
| RAS Secure Client Gateway | 3 | 2 | 8 GB |
| High Availability Load Balancing | 1* | 1 | 2 GB |
| RD Session Host | 38 | 8 | 20 GB |

* Two or more HALB devices are recommended in product environment for High Availability.

All virtual machines comprising the testing environment are siloed on the same virtual network.

## The Testing Environment Diagram



# Testing Process

In the scalability testing, Login VSI 4.1.40 was used to run a user load on Parallels RAS 17.1 using Parallels Client for Windows 17.1 (x64). 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.

The Knowledge Worker workload was selected for this testing. The Knowledge Worker is designed for 2(v) CPU environments. This is a well-balanced intensive workload that stresses the system smoothly, resulting in higher CPU, RAM and I/O usage.

The Knowledge Worker workload uses the following applications:

- Internet Explorer 11 web browser
- Microsoft Outlook 2013
- Microsoft Word 2013
- Microsoft Excel 201
- Microsoft PowerPoint 2013
- Adobe PDF Reader DC
- Doro PDF Writer
- 7-Zip
- Windows Photo Viewer
- Freemind/Java

More info about the Login VSI Knowledge Worker workload can be found at the following link: https://support.loginvsi.com/hc/en-us/articles/360001046100-Login-VSI-Workloads-Default-workloads-information

Knowledge Worker Login VSI workload was used to simulate the workload of 1000 users on Parallels RAS, the logon phase duration has been configured to take under around an hour and logon rate was set to 1 session per every 4 seconds.

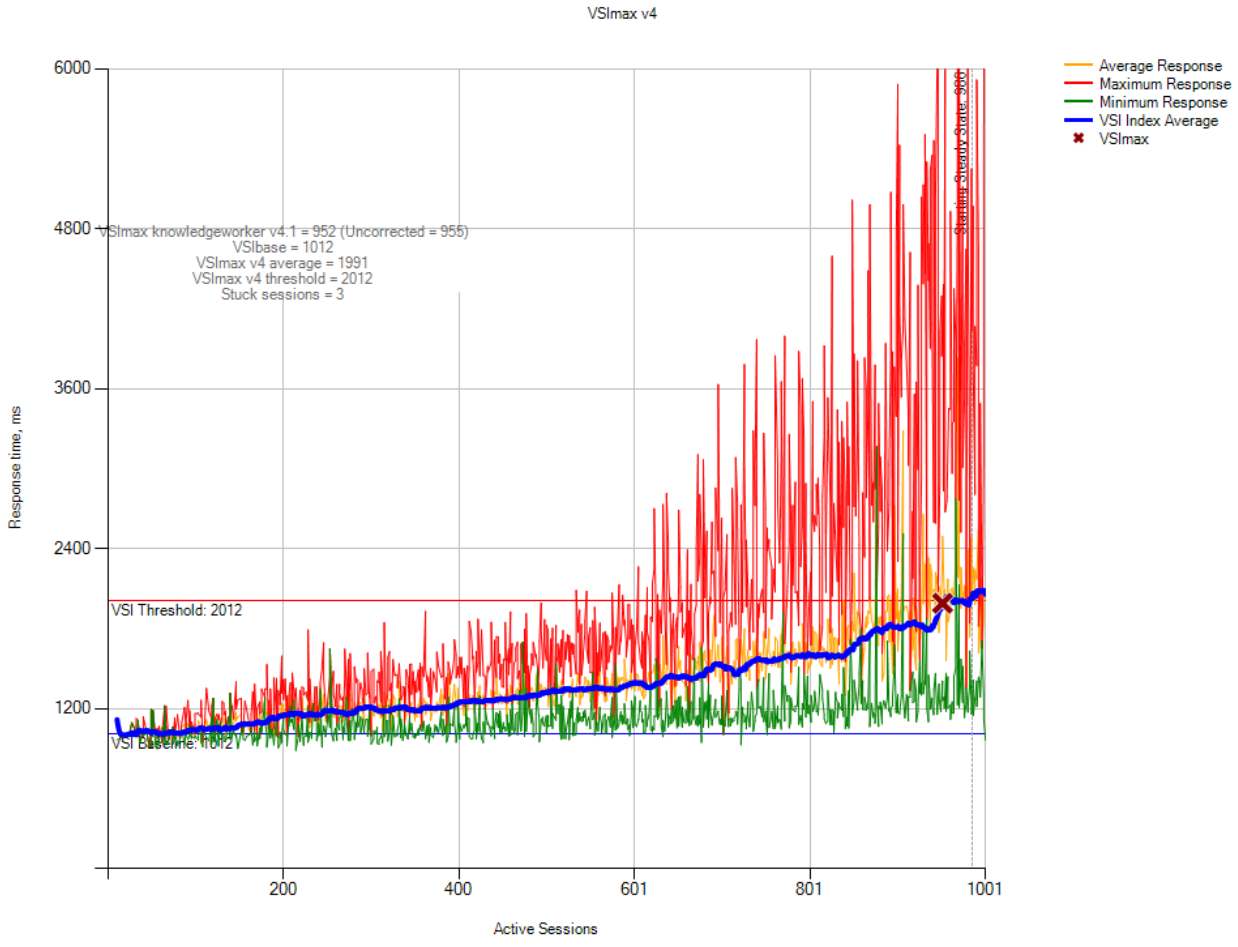
Performance metrics were captured during user logon and virtual desktop acquisition (ramp-up), user workload execution (steady state), and user logoff. To achieve consistent measurements that would reflect when components were appropriately cached, each workload ran for 48 minutes before Login VSI performance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

It is important to note that while scalability testing is a key factor in understanding how a platform and an 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 should conduct their own Login VSI scale testing using custom workload scripts. Additionally, such customers could request Parallels RAS POC/Pilot. Since the goal of this testing was to capture a baseline reflecting the densities possible, Login VSI client launchers were configured to go through Secure Gateway in proxy SSL mode.

# Findings

Following are test results for the Knowledge Worker workload. VSImax v4 (which indicates the maximum user density under a specific workload) is 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.

A VSImax v4 density of 1000 users running the Task Worker workload was demonstrated. In our tests, VSImax was reached with 952 sessions. This means that there were already 952 concurrent sessions before any UX degradation was observed based on the current servers specifications.



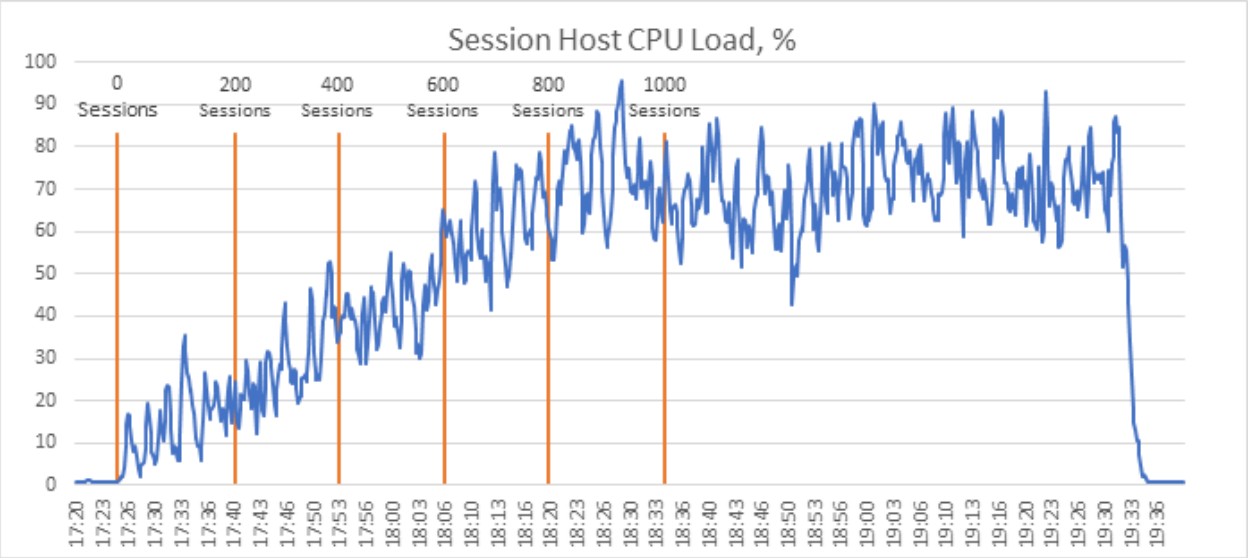
**Note:** Three stuck sessions were detected during the Login VSI test. Ideally there must be no stuck sessions but since the number is low and such sessions are subtracted from the VSImax score, it was decided to keep this results as this is an average number across several test runs.

# RD Sessions Hosts

The following test results for CPU, memory consumption, disk I/O response times and network load are helpful in evaluating performance under the test workload. Each chart below shows data collected from a single average RD Session Host server. Since there were 38 RD Session Host servers and 1000 simulated users, a single RD Session Host server accommodated around 26 users.

In the following two charts, as user load increases, the CPU and memory usage peak where the number of users approaches VSImax v4.

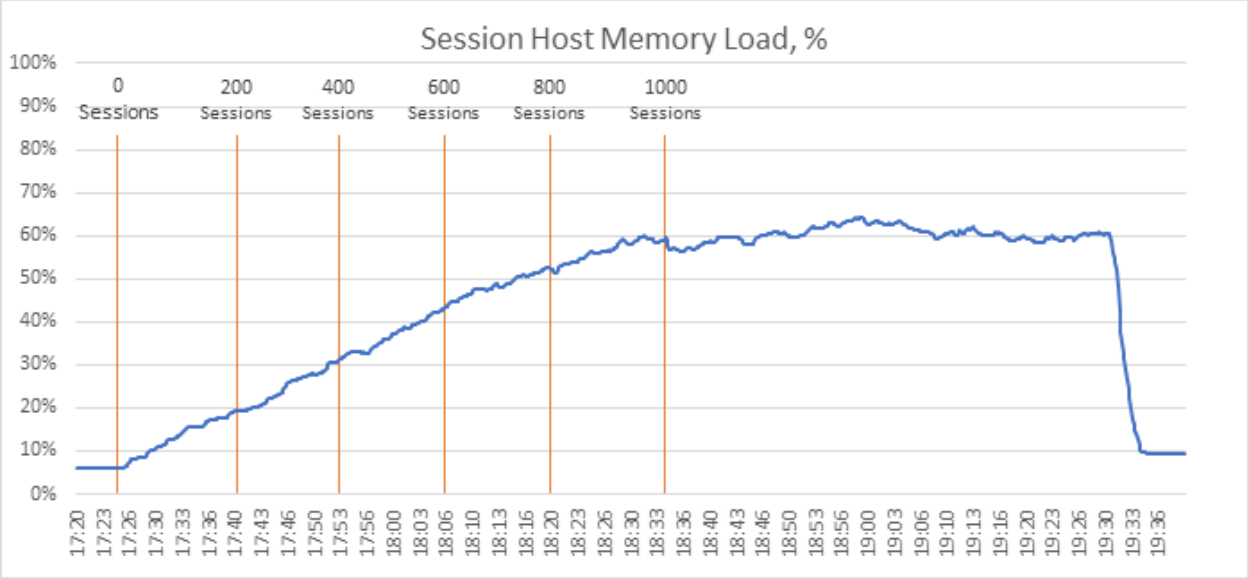
The below graph shows the CPU % utilization of an average RD Session Host during the test.




Logon phase on this chart has a lot of spikes and they are due to session logons, as usually every new session logon leads to a user profile load and is an expensive operation in general. Later, during the steady test phase, the CPU load stays at around 70.51% with a maximum recorded value at 93.03%. It shows that the RD Session Host capacity is not yet reached and proves that the VM configuration was properly selected according to the test load.

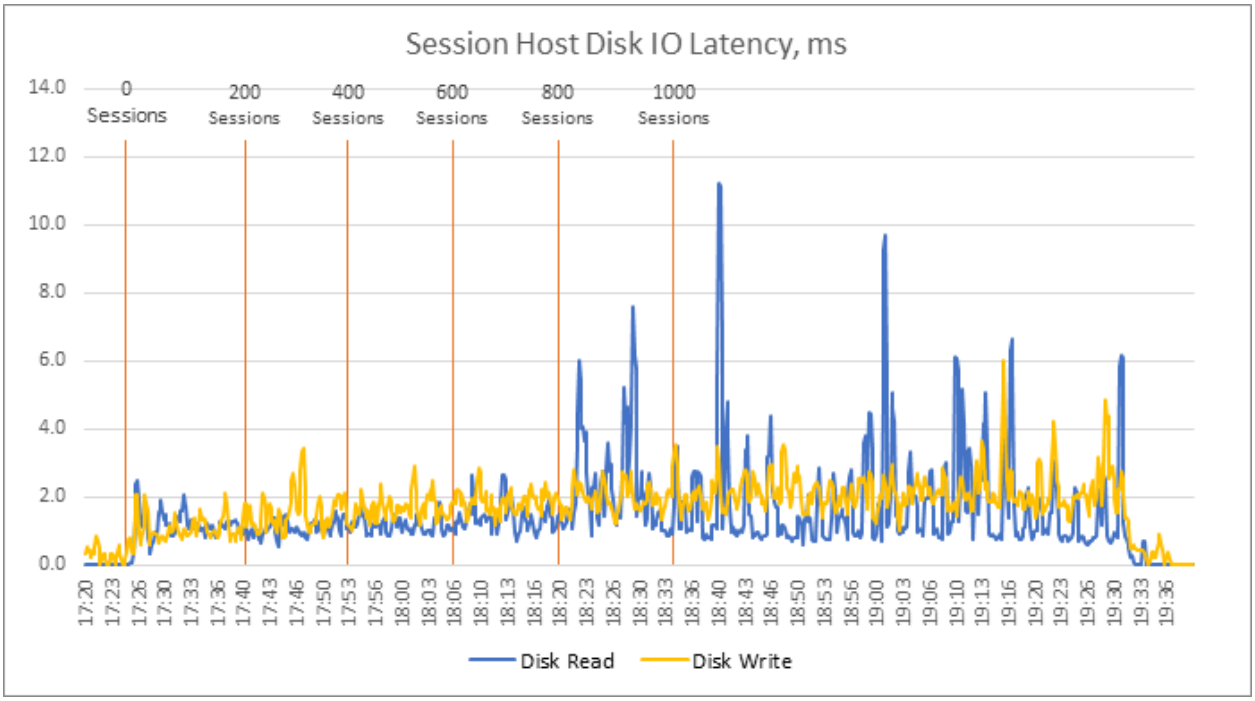
It should be mentioned that Session Host CPU usage during the workload phase is highly dependent on in-session activity. For instance, video playback increases both network and CPU usage, while working with more ‘static’ applications, requiring less screen updates, less I/O operations and thus less processor time.

The below graph shows the average memory (RAM) consumption of an average RD Session Host during the test. Load has been equally distributed by RAS Publishing Agent to all 38 RD Session Hosts participating in the test. RAM consumption growths steadily during logon phase until the workload phase where it stays at an average of 59% (11.8Gb).

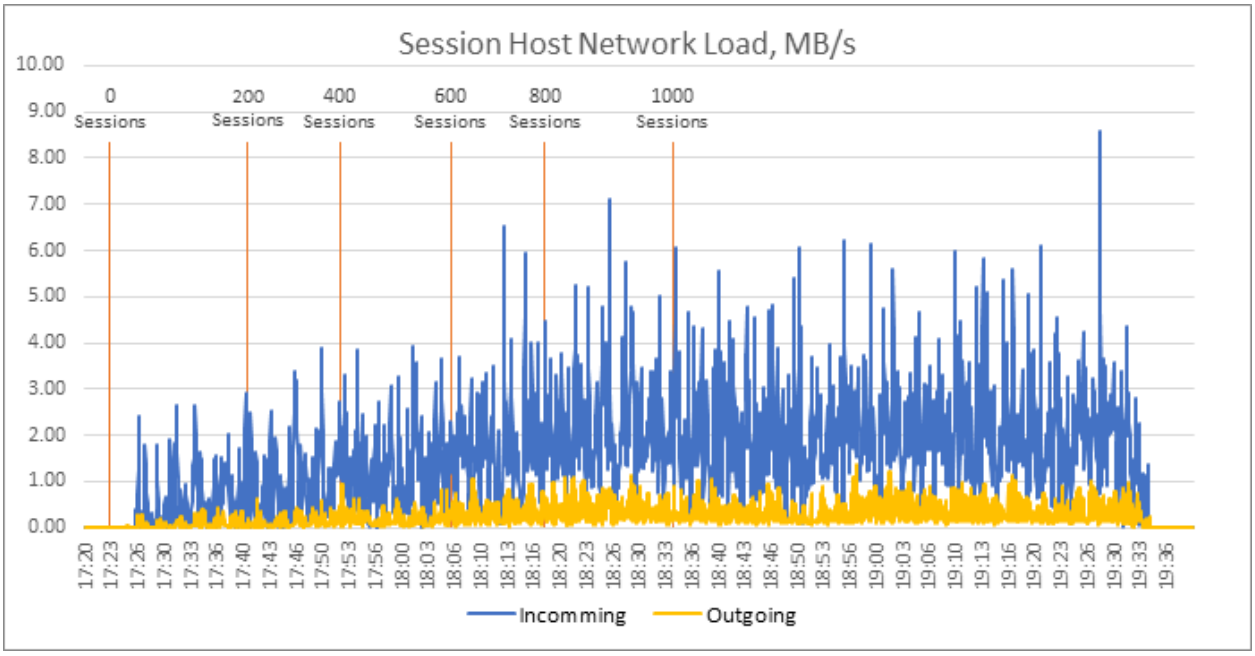



Apart from being dependent on the workload in each individual user session, it can be noted that memory (RAM) consumption on RD Sessions Hosts is also greatly dependent on the number of running sessions as can be seen during the steady phase of the test.

The following chart shows the average disk read and write response time. The average write I/O response time during the workload phase is about 0.646ms and read I/O response times average is 2.01ms.




The following chart shows networking transfer rates for data going in and out of the RD Session Host.



For the Knowledge Worker workload, the average outbound bandwidth at steady state is about 440 kB/s for our test group of 26 users (1000 total users divided by 38 RD Session Host servers). Therefore, the outgoing transfer rate per user is about 16.92 kB/s. Incoming traffic relates mostly to local network operations like profile loads, accessing files on the local network share and as a result we can see higher numbers of the Incoming traffic and the graph fluctuates a lot due to the nature of the session workload.

It should be mentioned that both Session Host Disk I/IO latency and Network load during the workload phase is highly dependent on in-session activity. For instance, video playback from a network share increases both network and disk I/O counters, and these operations generate significant spikes on the corresponding graphs.

CHAPTER 3

# RAS Infrastructure components

## In This Chapter



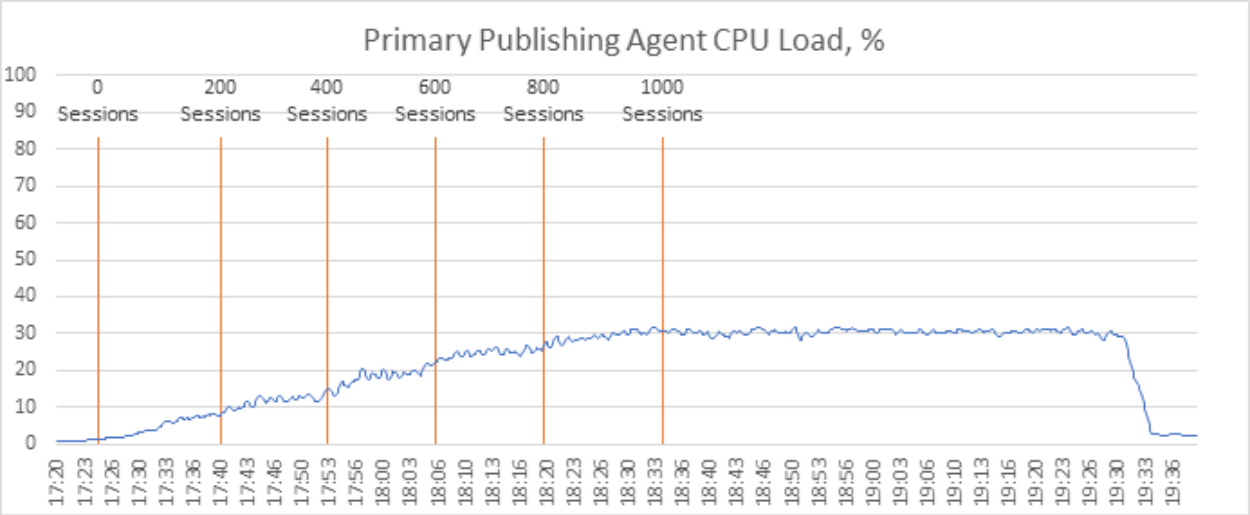
# RAS Publishing Agents

RAS Publishing Agent (PA) provides load balancing of published applications and desktops.

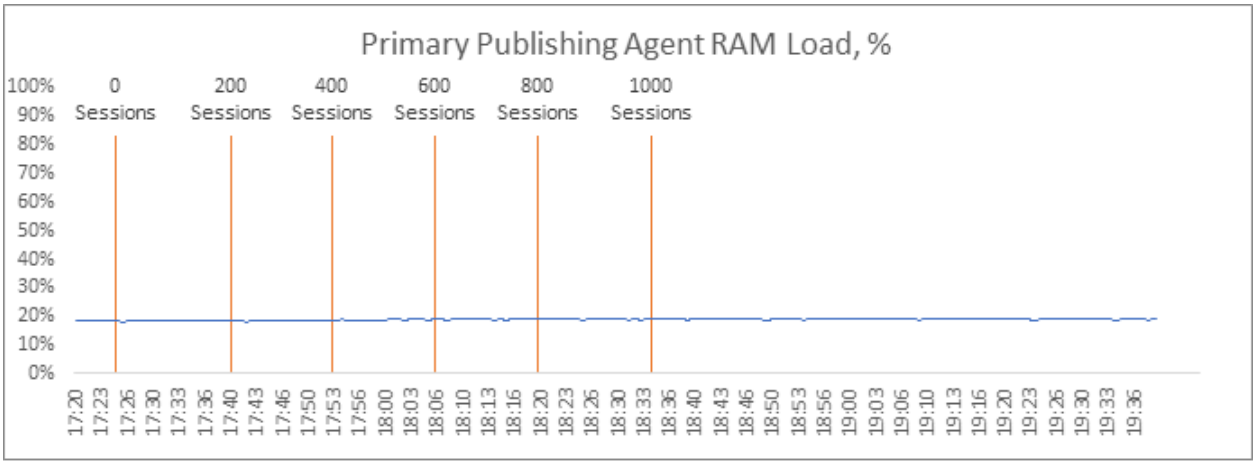
The below graph shows the average CPU % utilization of the two PAs in the RAS environment with respect to number of sessions being launched during the logon phase and later during the test. Load has been equally distributed to all PAs, that is, one Primary and one Secondary PAs.

The below graph shows the average CPU % utilization of the Primary PAs in the RAS environment during the test. Secondary PA CPU % utilization during the test was similar to the Primary PA's.

It can be noted that the average PAs CPU utilization is steady at an average of 30% through the whole test. This shows that the CPU specifications were well adequate for the workload of 1000 concurrent sessions.



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

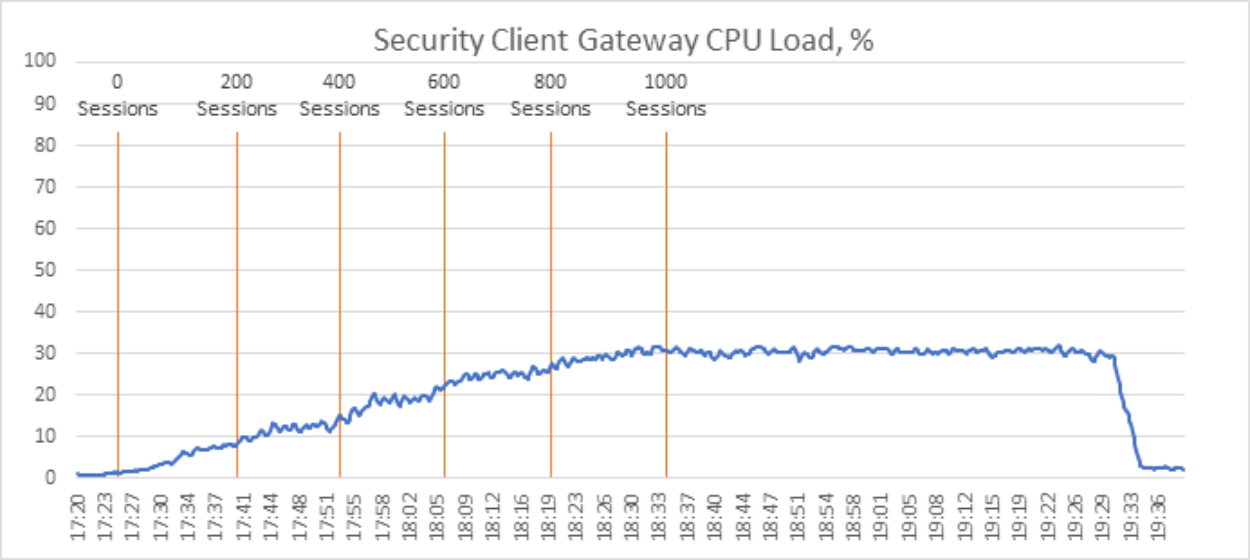



It can be noted that the average PAs memory (RAM) availability is adequate at a consistent average of 19% available memory during almost whole test. This means that lower amount of memory may have been allocated and still be adequate enough for the current number of users and workload.

# RAS Secure Client Gateways

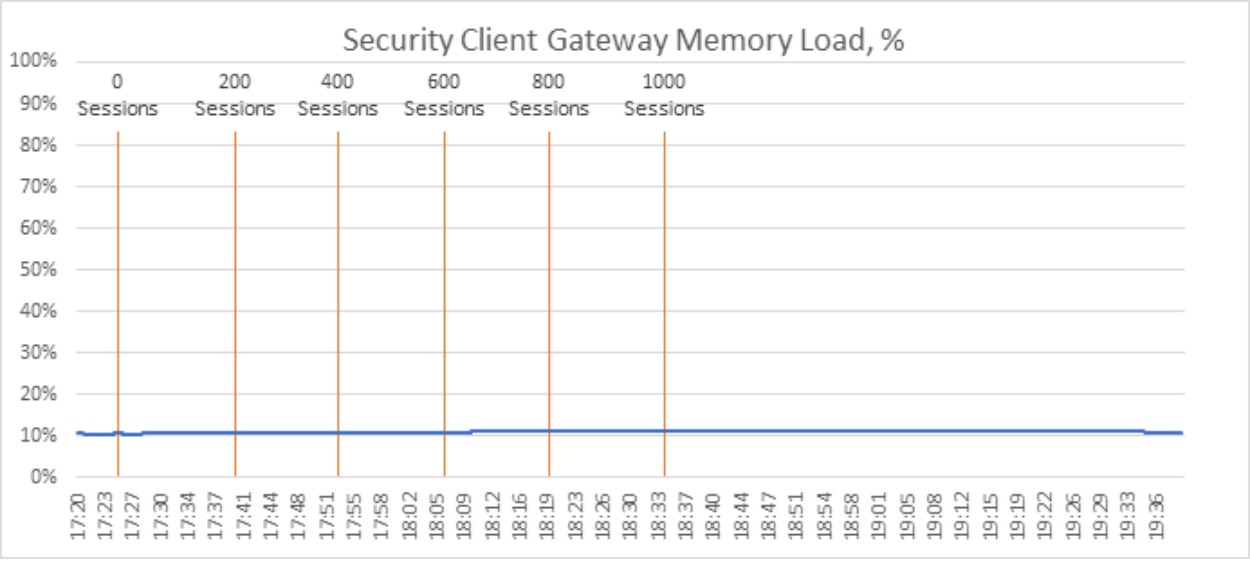
RAS Secure Client Gateway (SCG) tunnels all Parallels RAS data on a single port. It also provides secure connections and is the user connection point to Parallels RAS.

The below graph shows the average CPU % utilization of the 4 SCG with respect to number of sessions being launched during the logon phase. Load has been equally distributed to all SCG using RAS HALB appliance until the peak of 250 sessions per SCG.



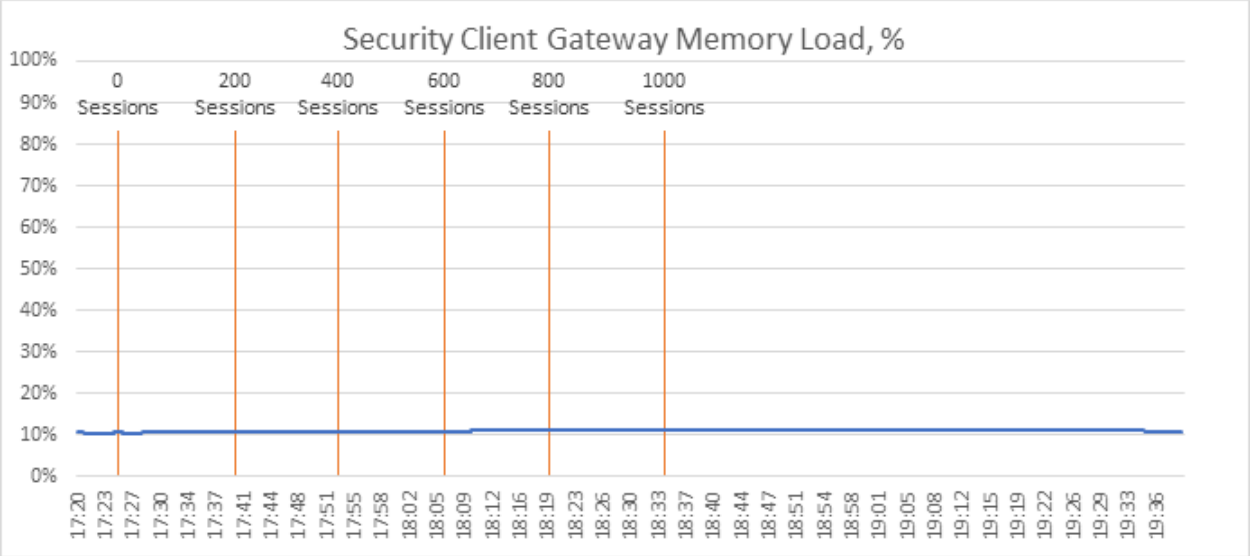

It can be noted that the average SCGs CPU utilization is on a steady increase in line with the increase of user sessions as expected. Also, SCG CPU utilization during the test phase is very stable and stays around 30%. As such it can be concluded that the CPU specifications were well adequate for the workload of 1000 concurrent sessions across 4 gateways with an average of 250 users per gateway.

The below graph shows the memory (RAM) consumption of an average SCG during the test. Load has been equally distributed to all SCGs



It can be noted that the average SCG memory (RAM) consumption is more than adequate at a consistent average of 11% (around 918MB) available memory even when reaching the 1000 user session mark. This means that lower amount of memory may have been allocated and still be adequate enough for the current number of users and workload.

The following graph shows the average network transfer rate through the SCGs. As expected the network utilization increases in line with number of user session logons. During the workload phase (after 1000 session were reached) the average network throughput was recorded at 9.90 MB/s on each SCG for a total of 39.6 MB/s when considering all SCGs, which translates to 39.6KB/s per user. The maximum transfer rate was noted at 12.77 MB/s or 51.08KB/s per user during workload phase.

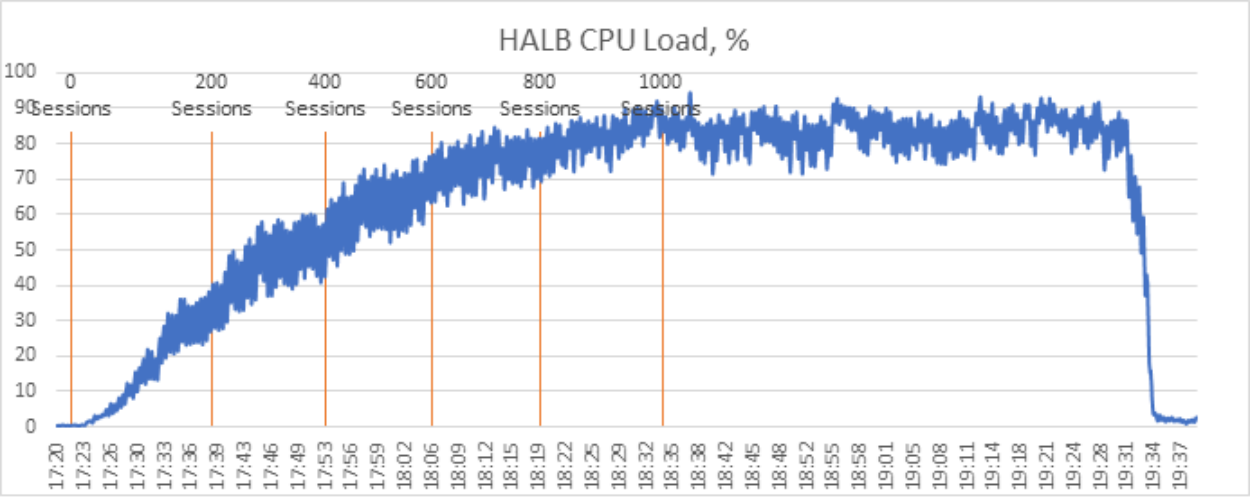



However, it should be noted that both SCG CPU and network usage during workload phase is highly dependent on in-session activity. For instance, video playback increases both network and CPU usage, while working with more ‘static’ applications, requiring less screen updates, such as Microsoft Word uses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

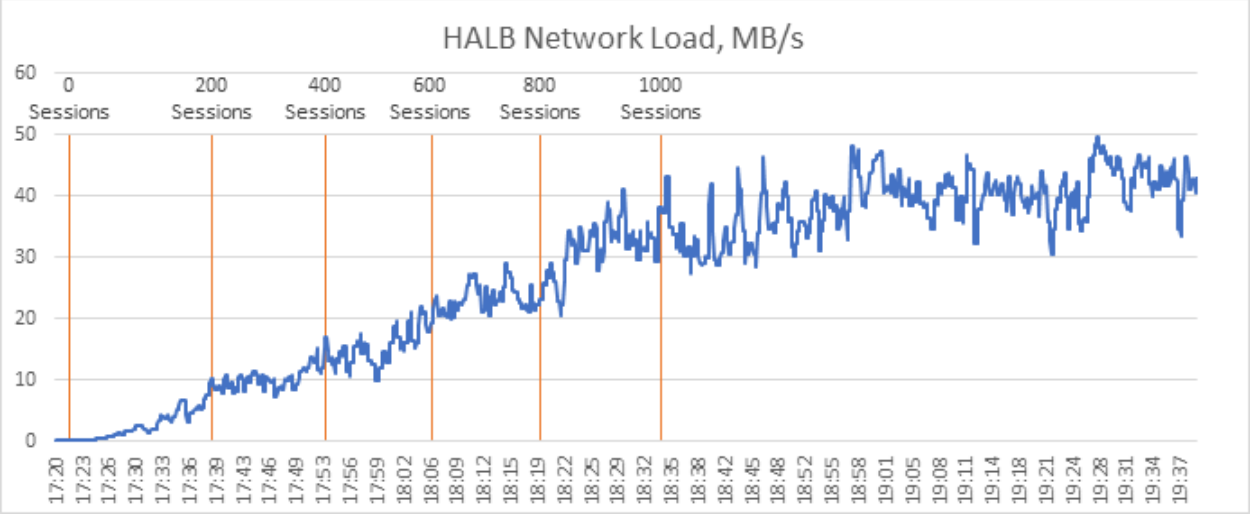
# HALB

High Availability Load Balancing (HALB) is an appliance that provides load balancing for RAS Secure Client Gateways. A Parallels HALB appliance is a preconfigured virtual machine with the operating system installed and all relevant settings configured. All the remote session traffic goes through the HALB.

The below graph shows the average CPU % utilization of the HALB appliance. Average CPU load during the test phase was 76.44% and that means the hardware limit on the HALB was yet to be reached.




On the following graph you can see how network transfer rate through the HALB during the logon phase growths steadily until it reaches the average at 39.16MB/s in the test phase. The maximum transfer rate was noted at 49.74 MB/s or 50.9KB/s per user during workload phase.




For additional HALB scalability information, refer to KB https://kb.parallels.com/125229

CHAPTER 4

# Conclusion

The Parallels RAS scalability results presented in this document confirm that 952 Login VSI sessions using the Knowledge Worker workload can be successfully launched using the given hardware configuration. Specifically, a total of 38 RD Session Host servers with 8 vCPU and 20 GB of RAM each were sufficient to accommodate these sessions with no user-experience degradation.

Parallels RAS was deployed on VMware ESXi 6.7 as follows:

| Parallels RAS Component | Total VMs | vCPU in Each VM | RAM in Each VM |
|---|---|---|---|
| RAS Publishing Agent | 2 | 2 | 8 GB |
| RAS Secure Client Gateway | 4 | 2 | 8 GB |
| High Availability Load Balancing | 1 | 1 | 2 GB |
| RD Session Host | 38 | 8 | 20 GB |

It is important to note that while load and scalability testing 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.

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