Parallels RAS Scalability Testing with Login VSI - 250 Users - Task Worker

250 Users – Task Worker workload

White Paper | Parallels Remote Application Server | 2021

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

© 2022 Parallels International GmbH. All rights reserved. Parallels and the Parallels logo are trademarks or registered trademarks of Parallels International GmbH in Canada, the U.S., and/or elsewhere.

Apple, Safari, iPad, iPhone, Mac, macOS, iPadOS are trademarks of Apple Inc. Google, Chrome, Chrome OS, and Chromebook are trademarks of Google LLC.

All other company, product and service names, logos, brands and any registered or unregistered trademarks mentioned are used for identification purposes only and remain the exclusive property of their respective owners. Use of any brands, names, logos or any other information, imagery or materials pertaining to a third party does not imply endorsement. We disclaim any proprietary interest in such third-party information, imagery, materials, marks and names of others. For all notices and information about patents please visit https://www.parallels.com/about/legal/

Contents

Introduction **4**

Scalability **5**

Testing the scalability of Parallels RAS 5

Configurations for scalability testing 5

Testing process 6

Findings 7

Conclusion **10**

Index **12**

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, Azure Virtual Desktop and major hypervisors.

This document presents an analysis of the scalability testing of Parallels RAS 18.2 using Login VSI for around 250 user sessions with Task Worker workload on RD Session Hosts.

CHAPTER 2

Scalability

In This Chapter

Testing the scalability of Parallels RAS 5
Configurations for scalability testing 5
Testing process..... 6
Findings 7

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 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 Connection Brokers 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, two physical servers were used consisting of the following hardware specifications each:

CPU	2x Intel Xeon E5-2680v4, 2.40 GHz
RAM	128 GB DDR4 1600 MHz
Storage	1 x 1 TB SSD
Network	1 x Gigabit Ethernet NIC

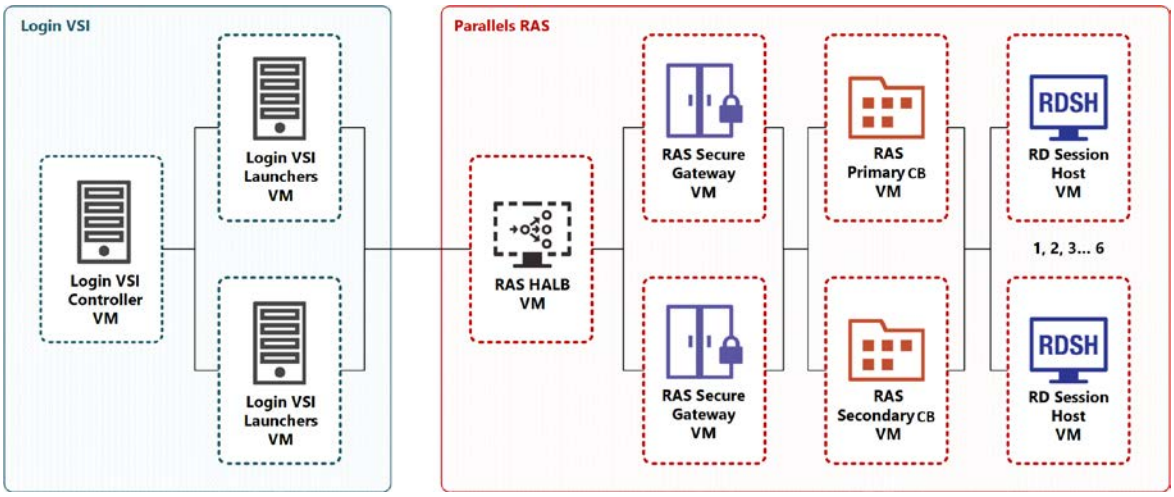
Parallels RAS was deployed on VMware ESXi 7.0.2 on Microsoft Windows Server 2019 in the following configuration*:

Parallels RAS component	Total VMs	vCPU in each VM	RAM in each VM
RAS Connection Broker (CB)	2	2	4 GB
RAS Secure Gateway	2	2	4 GB

High Availability Load Balancing (HALB)	1	1	2 GB
RD Session Host	6	6	24 GB

*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 18.2 using Parallels Client for Windows 18.2 (64-bit). 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 Task Worker workload was selected for this testing. The workload includes segments with Microsoft Office 2016 (Excel, Outlook, Word), Internet Explorer and Adobe Acrobat. While being diverse and not focused on one or two applications, the Task Worker workload does not place a very severe demand on the environment and represents users that do not overload the system with heavy tasks.

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

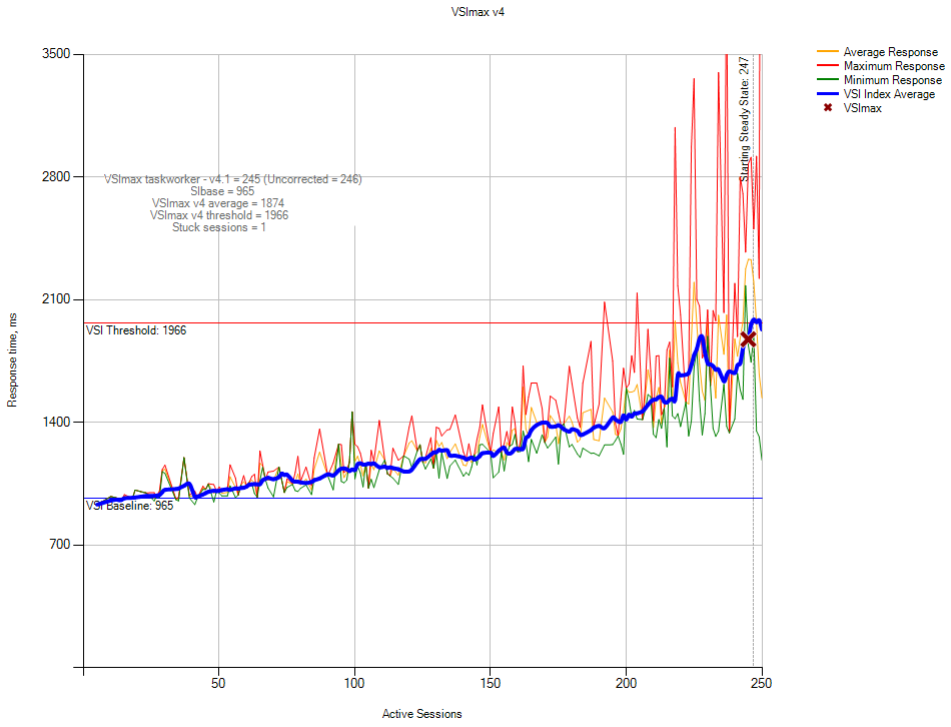
Task Worker Login VSI workload was used to simulate the workload of 250 users on Parallels RAS, the logon phase duration has been configured to take under around an hour and logon rate was set to 1 session per every 10 seconds. Since the goal of this testing was to capture a baseline reflecting the densities possible, Login VSI client launchers were configured to go through Secure Gateway in proxy SSL mode.

Performance metrics were captured during user logon and virtual desktop acquisition, user workload execution (steady state), and user logoff. To achieve consistent measurements that would reflect when components were appropriately cached, each workload ran for 45 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.

Findings

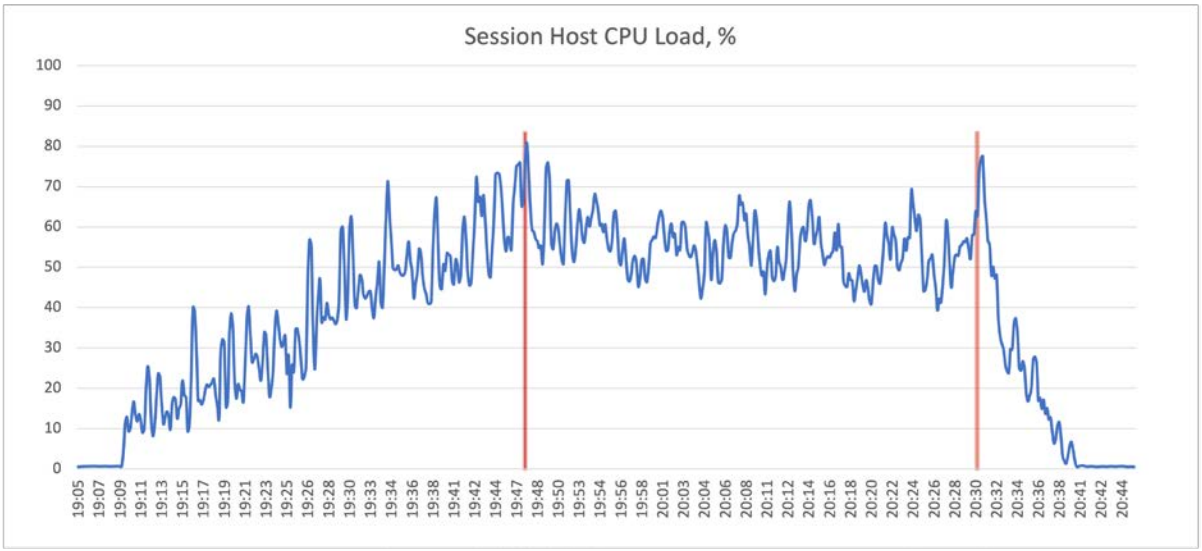
Following are test results for the Login VSI Task Worker workload. VSImax v4 (which indicates the maximum user density under a specific workload) is determined from the VSI Baseline and VSI Threshold metrics. VSI Baseline represents a pre-test Login VSI baseline response time measurement that is determined before the normal Login VSI sessions are sampled. A VSImax v4 density of 250 users running the Task Worker workload was demonstrated. In our tests, VSImax was reached with 245 sessions. This means that there were already 245 concurrent sessions before any UX degradation was observed based on the current servers' specifications.



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 under the Test Worker workload. Since there were 6 RD Session Host servers and 245 simulated users, a single RD Session Host server accommodated around 41 users.

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

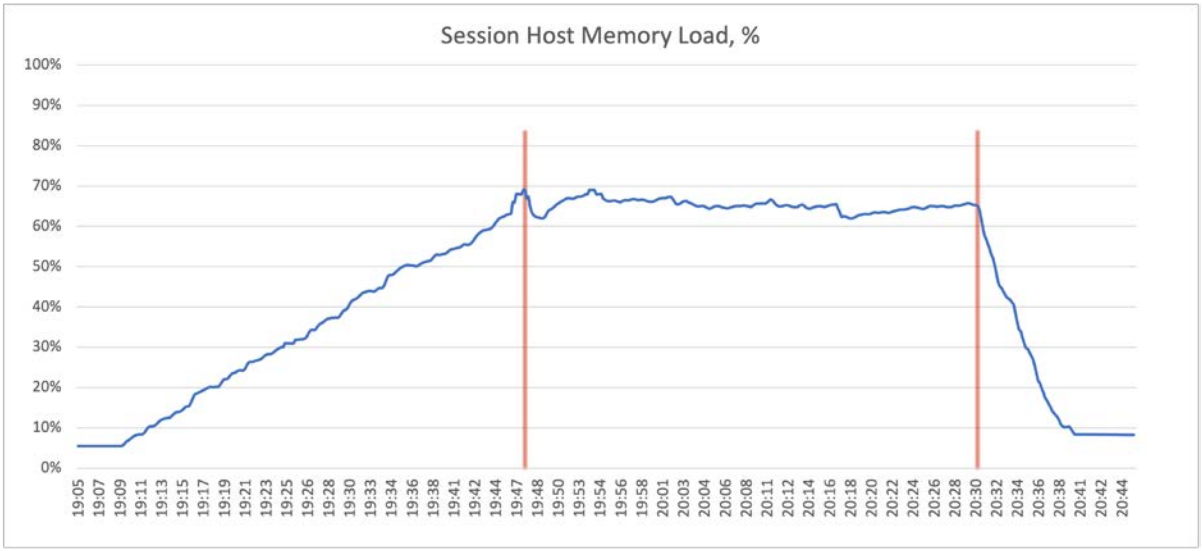
The below graph shows the CPU % utilization of an average RD Session Host under the Test Worker workload during the test.



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 55.17% with a maximum recorded value at 80.83%. 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.

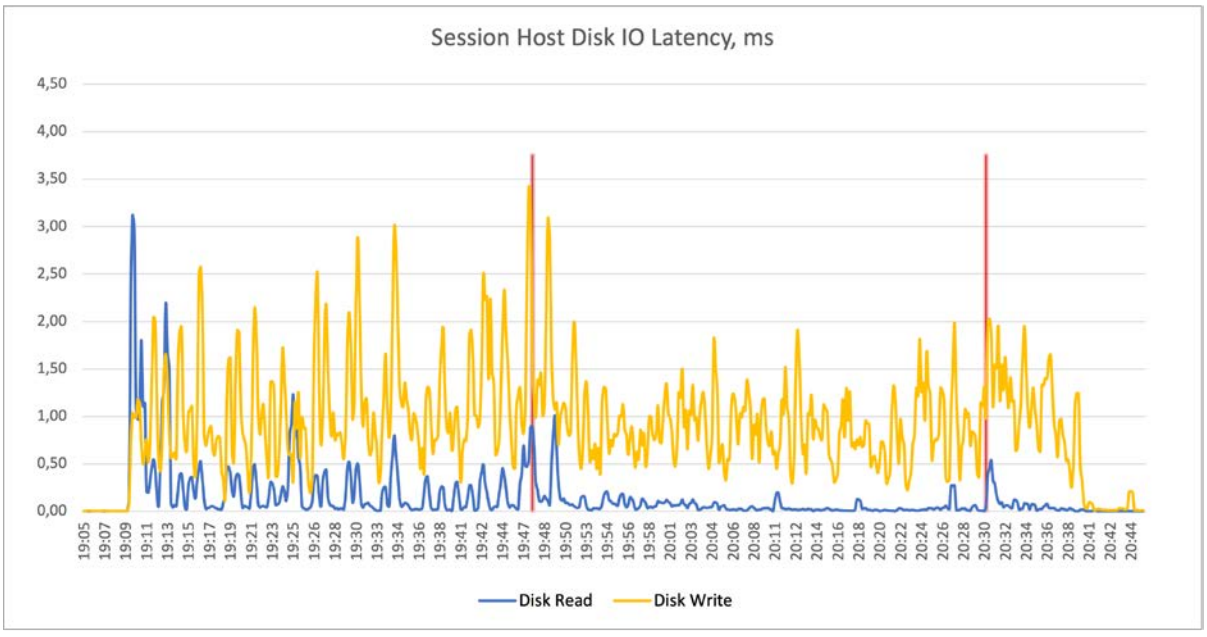
It should be mentioned that Session Host CPU usage during the workload phase is highly dependent on in-session activity. For instance, video playback increases both network and CPU usage, while more “static” applications require less screen updates, less I/O operations and thus less processor time.

The graph below shows the average memory (RAM) consumption of an average RD Session Host during the test. Load has been equally distributed by RAS Connection Broker to all 6 RD Session Hosts participating in the test. RAM consumption grows steadily during logon phase until the workload phase where it stays at an average of 65% (15 GB).

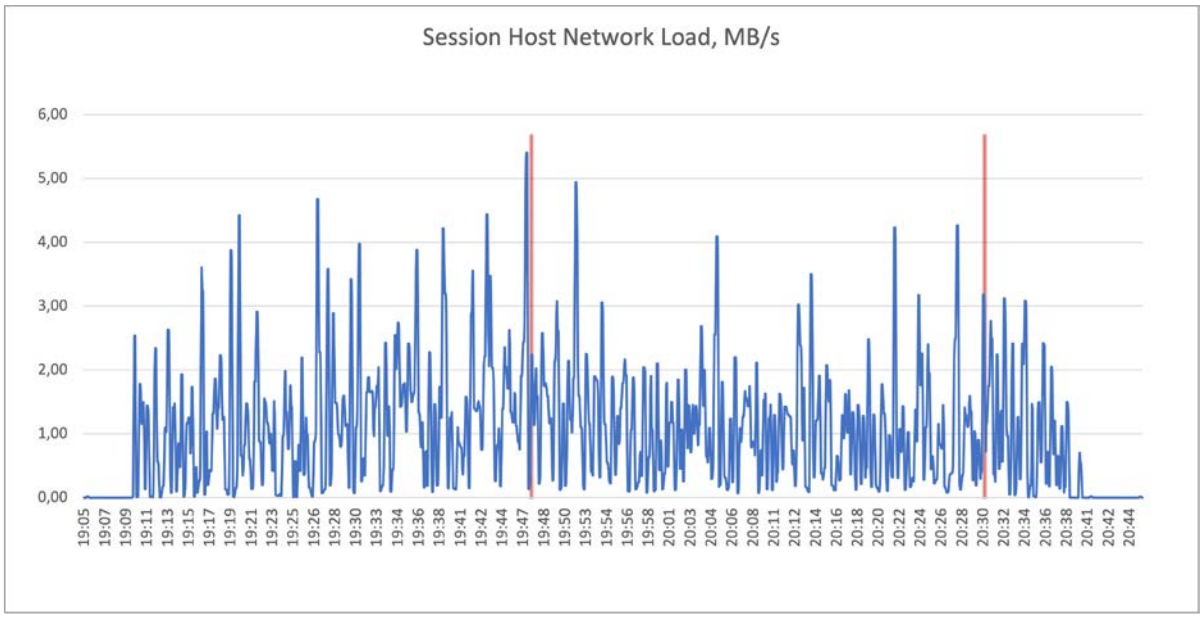


Apart from being dependent on the workload in each individual user session, it can be noted that memory (RAM) consumption on RD Sessions Hosts is also greatly dependent on the number of running sessions as can be seen during the steady phase of the test.

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 0.92 ms and read I/O response times average is 0.07 ms.



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



As follows from the graph, the difference between the logon and stability phases is insignificant, which can be explained by the type of workload.

For the Task Worker workload, the average outbound bandwidth at steady state is about 1.12 MB/s for our test group of 41 users (245 total users divided by 6 RD Session Host servers). Therefore, the outgoing transfer rate per user is about 27.94 KB/s.

It should be mentioned that both Session Host Disk I/O 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.

Conclusion

The Parallels RAS scalability results presented in this document confirm that 245 Login VSI sessions using the Task Worker workload can be successfully launched using the given hardware configuration.

Specifically, a total of six RD Session Host servers with 6 vCPU and 24 GB of RAM each were sufficient to accommodate these sessions with no user-experience degradation.

For the Parallels RAS scalability testing, two physical servers were used consisting of the following hardware specifications each:

CPU	2x Intel Xeon E5-2680v4, 2.40 GHz
RAM	128 GB DDR4 1600 MHz
Storage	1 x 1 TB SSD
Network	1 x Gigabit Ethernet NIC

Parallels RAS 18.2 was deployed on VMware ESXi 7.0.2 on Microsoft Windows Server 2019 as follows:

Parallels RAS component	Total VMs	vCPU in each VM	RAM in each VM
RAS Connection Broker (CB)	2	2	4 GB
RAS Secure Client Gateway	2	2	4 GB
High Availability Load Balancing (HALB)	1	1	2 GB
RD Session Host	6	6	24 GB

The results collected from a single average RD Session Host server under the Test Worker workload during steady phase are:

RD Session Host density	41 users (~20 users/vCPU)
Memory	375 MB per user
Disk average I/O response time	write: 0.92 ms read: 0.07 ms
Network bandwidth	27.94 KB/s per user

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 to 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.

Index

C

Conclusion - 10
Configurations for scalability testing - 5

F

Findings - 7

I

Introduction - 4

S

Scalability - 5

T

Testing process - 6
Testing the scalability of Parallels RAS - 5