Parallels RAS HALB Scalability

- Parallels Remote Application Server 18.3
- Parallels Remote Application Server 19.1

Summary

High availability load balancing (HALB) in Parallels RAS is a functionality that load balances RAS Secure Gateways. The load balancer is built into a Parallels HALB appliance, which is a preconfigured virtual machine with the operating system installed and all relevant settings configured.

HALB usage should not exceed 2000 user sessions per HALB appliance.

In the pass-through mode case, the number of user sessions may have a higher value due to the mode does not have a hard limit to the HALB appliance.

Testing environment

VM configuration

Component	N Instances	vCPU	RAM
HALB	1	1	1
Primary PA	1	2	8
Secondary PA	1	2	8
Gateway	5	2	8
RDSH	50	10	20
Client	20	4	10

- Windows Server 2016 for all RAS Core components
- HALB deployed on a separate hardware host
- FSLogix profiles are configured for more stable session results

The testing goal

The goal of the test was to find number of users single HALB instance can handle during 10 min logon storm in production scenario (not causing CPU to cross 90% threshold) and then detect number of users to max out CPU usage during logon storm in given mode.

Totally over 40 HALB trials performed to achieve reproducible results and about 40 trials to detect environment capacity without HALB.

- Production result picked according to 0% failed sessions criteria.
- Accidental session fails start to take place after 2000 sessions and goes from 1% at 2100 sessions to 17-20% at 3500 sessions.
- The amount of failed sessions is relatively small up to 2500 sessions for the environment, which is capable of running more than 2500 sessions without failures (in case of HALB-less deployment where round robin of Gateway sessions is supported via script).

Gateway considerations (include 0% failed session's criteria)

- Max number of simultaneous sessions (single shot): 60
- Max number of sessions in 10 minutes time interval: 700. Each iteration had an increase step 50 sessions. Therefore, to host 2100 sessions 3 gateways used for 3500 sessions – 5 gateways.

HALB pass-through mode: production vs peak

Production

Summary:

N sessions: 2000

Failed: 0%

Max CPU: 70%

Avg CPU: 43%

Max throughput: 17.43 (MB/s)

Avg throughput: 11.82 (MB/s)

CPU utilization:

Network Throughput:

Peak

N sessions: 3500

Failed: ~17%

Max CPU: 100%

Avg CPU: 58%

Max throughput: 30.66 (MB/s)

Avg throughput: 15.27 (MB/s)

Network Throughput:

HALB Offloading mode: production vs peak

Production

N sessions: 2000

Failed: 0%

Max CPU: 87%

Avg CPU: 51%

Max throughput: 30.66 (MB/s)

Avg throughput: 15.27 (MB/s)

Network Throughput:

Peak

N sessions: 3000

Failed: ~33%

Max CPU: 100%

Avg CPU: 60%

Max throughput: 29.27 (MB/s)

Avg throughput: 19.65 (MB/s)

Network Throughput:
