How Parallels RAS Autoscaling works with Azure Virtual Desktop

- Parallels Remote Application Server 18.3

This article describes how Parallels RAS auto-scale mechanism works with Azure Virtual Desktop (AVD). When creating new or adding an existing host pool, Parallels RAS allows to configure the following auto-scale settings:

Available settings

Setting name	Explanation
Min number of hosts to be added to the host pool from Template:	This is the number of hosts available always in the host pool. E.g. When host pool is empty, a host will be added to host pool.
Max number of hosts to be added to the host pool from Template:	This is the number of maximum hosts a host pool is allowed to have. At no time the number of hosts in a host pool should exceed this number.
Add new hosts from	This is the workload % of the host pool required to trigger a request (refer to workload

Setting name	Explanation
template when workload threshold is above (%):	calculation explained below).
Number of hosts to be added to the host pool per request:	This is the number of hosts to be added per request, if the number of hosts to be requested is going to exceed the maximum number of allowed hosts, the auto-scaling engine only adds the required hosts.
Drain and remove hosts from host pool when workload is below (%) and remains below this level for:	This is the workload % of the host pool required to trigger un-assignment (removal of host from host pool). If remains below is not set to immediate, then it will only trigger the un-assignment if the workload % of host pool is still below after the remains below time has elapsed. Please note that if in between the first un-assignment trigger and the second check after the remains below time has elapsed the workload % meets the request criteria, the un-assignment request is discarded.

.

Workload calculation

The host pool workload is calculated based on the sessions (active/disconnected). The maximum sessions value configured in the host pool settings is taken into consideration as well. Sessions currently running on the hosts can be viewed from the Site Info page or from the session management page in the Farm category section.

totalSessions = total running sessions (active/disconnected) from all hosts in host pool

maxHostPoolSessions = the limit number of sessions on host multiplied by the number of Hosts with agent state as OK

Hostpool Workload % = (*totalSessions* * 100) / *maxHostPoolSessions*

How the auto-scale action is triggered

The checks for AVD requests and un-assignment are triggered by one of the following scenarios:

- By applying the settings (two minutes after it is processed on the primary Publishing Agent).
- By the session counters (if they changed only) received from each AVD host guest agent.
- Every 30 minutes.

Note: Only the primary Publishing Agent of each site will trigger/process the mentioned checks. The check is not triggered immediately when the apply notification is received because there might be agents that need to be redistributed in a multiple Publishing Agents environment.

Example

Given environment

Host pool type:	Pooled
Provision type:	Template
Load balancer	Breadth
Min number of hosts to be added to the host pool from Template:	1
Limit number of sessions per host	50
Max number of hosts to be added to the host pool from Template:	5
Add new hosts from template when workload threshold is above (%):	60
Number of hosts to be added to the host pool per request:	1
Drain and remove hosts from host pool when workload is below (%)	20
and remain below this level for	Immediate

Scenario

- We have 30 running user sessions on an AVD host **Host1** which equals to 60% of the workload we configured.
- There is another user connected and the workload now is above 60%.
- Parallels RAS will trigger the creation/adding of the additional host **Host2** immediately after the 31st session established.
- The new host provisioning will take few minutes depending on the configuration.
- **Host2** provisioned and operational. 15 users connected to it. On **Host1** at this time we have 40 sessions, in total 55 sessions on two hosts.
- 6 more users connected to **Host2** (21 sessions on **Host2** in total). **Host1** still serves 40 sessions. The workload on both hosts per configuration is 61% now, this triggers the creation of the 3rd Host.
- **Host3** is fully operational and serves 7 sessions. There are also 30 sessions on **Host2** and 40 sessions on **Host1,** the overall Hostpool workload is ~51%
- After the working day users started logging off and there are only 25 sessions running (~16%).
- Since we set the "Immediate" in the Hostpool properties, **Host3** will be immediately set to drain mode first since it has the least number of users sessions running.
- **Host2** will switch to the drain mode during the next check if the total workload at this time is less than 20%.
- Once all users logged off, these hosts will be unassigned from the Hostpool. **Host1** will continue running normally due to the setting "***Min number of hosts to be added to the host pool from Template".***
- In case new users connects at this time and increase the workload back to 60+%, this will trigger the provisioning of the new hosts. (**Host2** and **Host3** once in drain mode, will not accept new user sessions)
