



Resource Optimization

Virtual Desktop Service

Toby vanRoojen
February 17, 2021

This PDF was generated from https://docs.netapp.com/us-en/virtual-desktop-service/Management.Cost_Optimization.workload_schedule.html on September 12, 2021. Always check docs.netapp.com for the latest.

Table of Contents


- Resource Optimization 1
 - Workload scheduling 1
 - Wake on demand 1
 - Live Scaling 3
 - VM resource scaling 4
 - Other active resources 7

Resource Optimization

Workload scheduling

Workload Scheduling is a feature that can schedule the time window in which the environment is active.

Workload scheduling can be set to "Always On", "Always Off" or "Scheduled". When set to "Scheduled" the on and off times can be set as granularly as a different time window for each day of the week.

 5.4 Preview

Edit Workload Schedule

Status

Scheduled ▼

Scheduling Options

☐ Run at assigned time interval everyday

☐ Run at assigned time interval on specified days

☒ Run at variable time interval and days

Days

☐ Sun ☒ Mon ☐ Tue ☒ Wed ☒ Thu ☒ Fri ☐ Sat

Current Schedule

4 Day(s) Scheduled.

Cancel Update Schedule

When scheduled to be off, either via "Always Off" or "Scheduled", all tenant virtual machines will shut down. Platform servers (such as CWMGR1) will remain active to facilitate functionality such as wake on demand.

Workload Schedule works in conjunction with other resource optimization features including Live Scaling and Wake on Demand.

Wake on demand

Wake on Demand (WoD) is patent-pending technology that can wake the appropriate VM resources for an end user in order to facilitate unattended access 24/7, even when resources are scheduled to be inactive.

WoD for Remote Desktop Services

In RDS, the VDS Windows Client has built-in Wake on Demand integration and can wake the appropriate resources without any additional end-user actions. They simply need to initiate their normal login and the client will notify them of a short delay which the VM(s) are activated. This client (and thus this automate wake on demand functionality) is only available when connecting from a Windows device to an RDS environment.

Similar Functionality is built into the VDS Web client for RDS deployments. The VDS Web Client is found at: <https://login.cloudworkspace.com>

Wake on Demand functionality is not built into the Microsoft RD client (for Windows or any other platform) nor any other 3rd party RD clients.

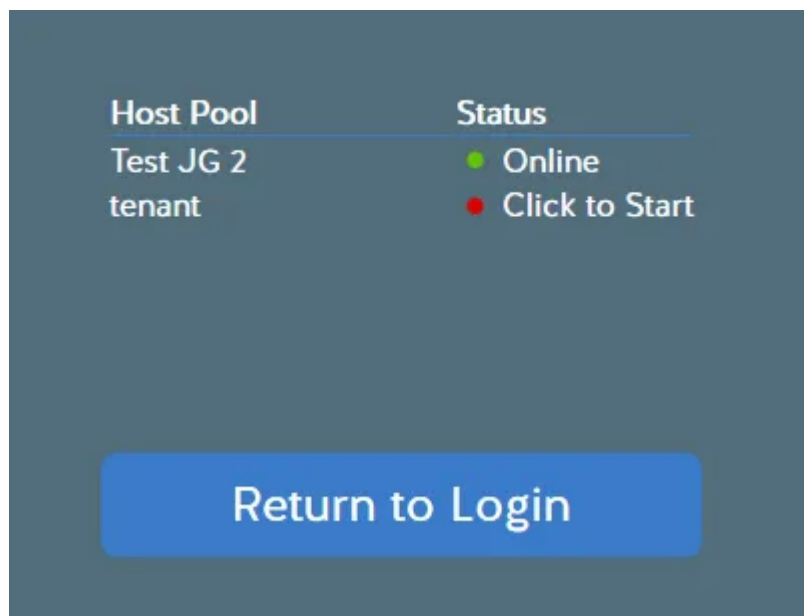
Wake on demand for Azure Virtual Desktop

In AVD, the only clients that can be used to connect are Microsoft provided and thus do not contain the Wake on Demand functionality.

VDS does include a self-service Wake on Demand function for AVD via the VDS Web Client. The web client can be used to wake the appropriate resources, then the connection can be initiated via the standard AVD client.

To wake VM resources in AVD:

1. Connect to the VDS Web Client at <https://login.cloudworkspace.com>
2. Login with the user AVD credentials
 - A warning message will prompt *"You have Microsoft's AVD services available. Click HERE to view the status and start offline Host Pools."*
3. After clicking "HERE" you'll see a list of available Host Pools along with a link to "Click to Start" link under the status column



4. Click to Start the link and wait 1-5 minutes for the status to change to "Online" and show a green status icon
5. Connect to AVD using your normal process

Live Scaling

Live Scaling works in conjunction with Workload Scheduling by managing the number of online session hosts during the scheduled active time as configured in Workload Scheduling. When scheduled to be offline, Live Scaling won't control session host availability. Live scaling only impacts Shared Users and Shared Servers in RDS and AVD environments, VDI Users and VDI VMs are excluded from these calculations. All other VM types are unaffected.



The AVD *load balancer type* setting interacts with this configuration, so care should be taken in choosing that setting as well. Cost savings are maximized with a depth-first type while end user performance is maximized with a breadth-first type.

Enabling Live Scaling with no options checked, the automation engine will automatically select values for the Number of Extra Powered on Servers, Shared Users Per Server, and Max Shared Users Per Server.

- The *Number of Extra Powered on Servers* defaults to 0, meaning 1 server will run 24/7.
- The *Shared Users Per Server* defaults to the number users in the company divided by the number of servers.
- The *Max Shared Users Per Server* defaults to infinite.

Live Scaling turns the servers on as users log on and turns them off as users log off.

Powering an additional server is automatically triggered once the total active users reaches the number of Shared Users per Server multiplied by the total number of Powered On Servers.

e.g. With 5 Shared Users per Server set (this is the default # we'll use for all examples in this article) and 2 servers running, a 3rd server won't be powered up until server 1 & 2 both have 5 or more active users. Until that 3rd server is available, new connections will be load balanced all available servers. In RDS and AVD Breadth mode, Load balancing sends users to the server with the fewest active users (like water flowing to the lowest point). In AVD Depth mode, Load balancing sends users to servers in a sequential order, incrementing when the Max Shared Users number is reached.

Live Scaling will also turn off servers to save costs. When a server has 0 active users, and another server has available capacity below *Shared Users per Server* the empty server will be powered down.

Powering on the next server can take a few minutes. In certain situations the speed of logins can outpace the availability of new servers. For example, if 15 people login in 5 minutes they'll all land on the first server (or be denied a session) while a 2nd and 3rd power up. There are two strategies that can be used to mitigate overloading a single server in this scenario:

1. Enable *Number of Extra Powered on Servers* so that the additional server(s) will be on and available to accept connections and allow time for the platform to spin up additional servers.
 - a. When activated, the number is added to the calculated need. For example, if set to 1 extra server (and with 6 users connected) two servers would be active because of the users count, plus a 3rd due to the *Extra Powered on Servers* setting.
2. Enable *Max Shared Users Per Server* to place a hard limit on the number of users allowed per server. New

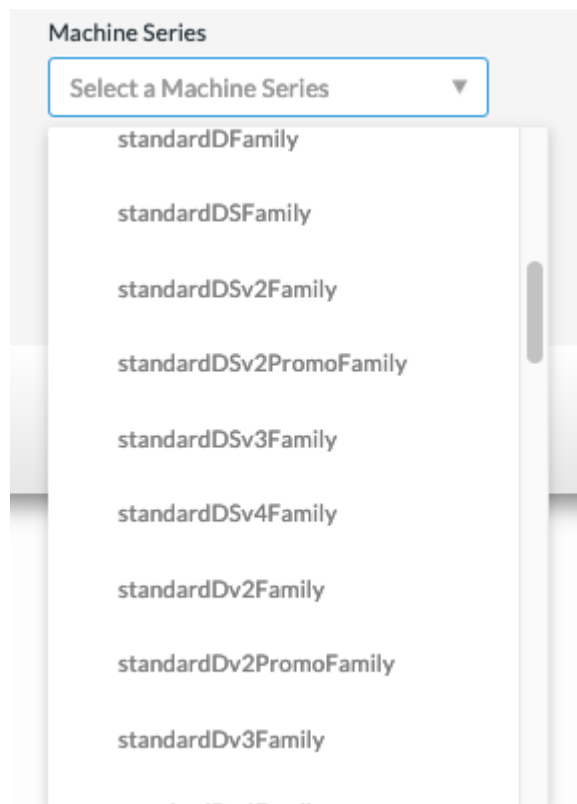
connections that would exceed this limit will be refused, the end user will get an error message and need to try again in a couple minutes once the additional server is available. If set, this number also defines the depth of AVD Shared servers.

- a. Assuming the delta between *Shared Users Per Server* and *Max Shared Users Per Server* is appropriate, the new servers should become available before the max is reached in all but the most extreme situations (unusually large login storms).

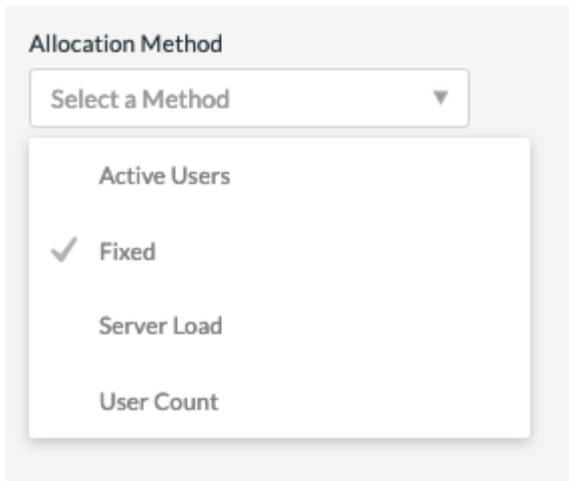
VM resource scaling

VM Resource scaling is a optional feature that can change the size and quantity of session host VMs in an environment.

When activated, VDS will calculate the appropriate size and quantity of session host VMs based on your selected criteria. These options include: Active Users, Named Users, Server Load, and Fixed.



The size of the VMs is contained with the family of VMs selected in the UI which can be changed by dropdown. (e.g. *Standard Dv3 Family* in Azure)



The screenshot shows a dropdown menu titled "Allocation Method". The menu is open, displaying four options: "Active Users", "Fixed" (which is selected and marked with a checkmark), "Server Load", and "User Count". The "Fixed" option is highlighted with a light blue background.

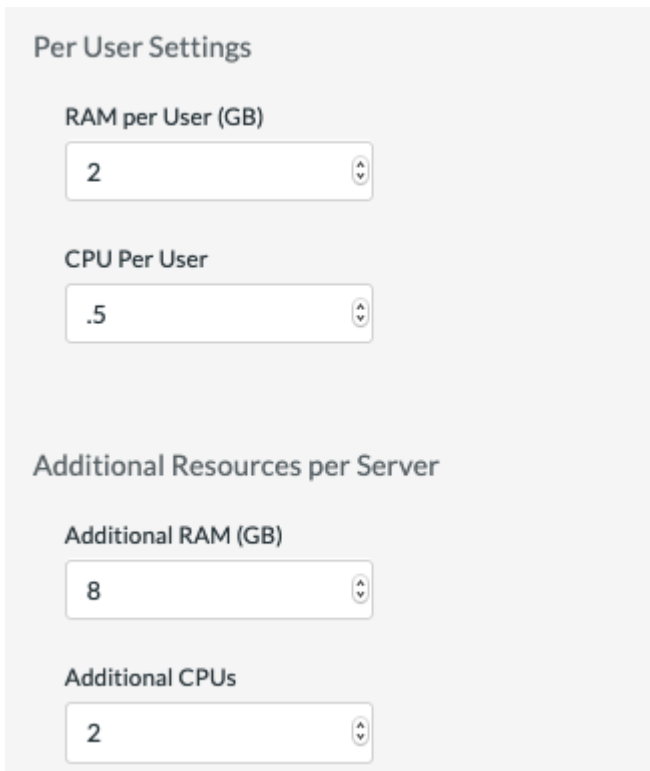
Scaling based on users



The function below behaves the same for either "Active Users" or "User Count". User Count is a simply count of all users activated with a VDS desktop. Active Users is a calculated variable based on the previous 2 weeks of user session data.

When calculating based on users, the size (and quantity) of the session host VMs is calculated based on the defined RAM and CPU requirements. The administrator can define the GB of RAM, and number of vCPU cores per user along with additional non-variable resources.

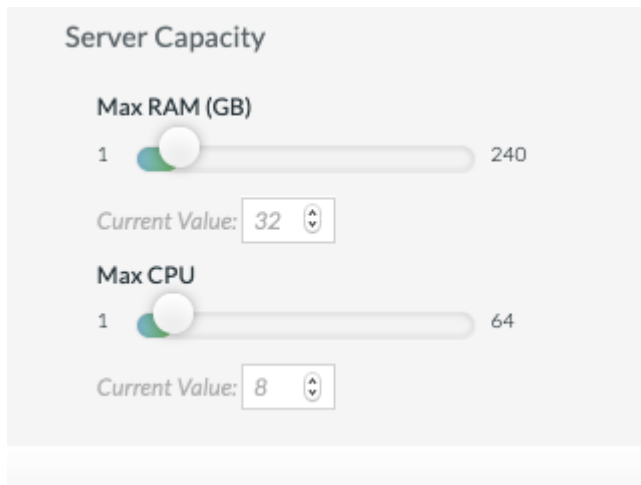
In the screenshot below, each user is allocated 2GB RAM and 1/2 of a vCPU core. Additionally, the server starts with 2 vCPU cores and 8GB RAM.



The screenshot shows two sections of a configuration interface. The first section, "Per User Settings", contains two spinners: "RAM per User (GB)" set to 2 and "CPU Per User" set to .5. The second section, "Additional Resources per Server", contains two more spinners: "Additional RAM (GB)" set to 8 and "Additional CPUs" set to 2. All spinners have up and down arrows for adjustment.

Additionally, the administrator can define the maximum size a VM can reach. When reached, environments will scale horizontally by adding additional VM session hosts.

In the screenshot below, each VM is limited to 32GB Ram and 8vCPU cores.



With all of these variables defined, VDS can calculate the appropriate size and quantity of session host VMs, greatly simplifying the process of maintaining appropriate resource allotment, even as users are added and removed.

Scaling based on server load

When calculating based on server load, the size (and quantity) of session host VMs is calculated based on the average CPU/RAM utilization rates as observed by VDS over the previous 2-week period.

When the maximum threshold is exceeded, VDS will increase the size or increment the quantity to bring average usage back within range.

Like user based scaling, the VM Family and the maximum VM size can be defined.

Manage Resource Pool

Basic Resource Info

Name

Primary Host Pool

Status

☒ Enabled

☐ Disabled

Use Default Deployment Settings

☐ Yes

☒ No

Allocation Method

Server Load

Machine Series

standardDSv2Family

Total Shared Servers

2

Server Load Settings

Peak Hourly Resource Usage

RAM

0%

CPU

0%

Increase Resource Threshold

RAM

70%

CPU

70%

Decrease Resource Threshold

RAM

25%

Server Capacity

Max RAM (GB)

1

240

Current Value: 32

Max CPU

1

64

Current Value: 8

Cancel

Apply to Servers

Other active resources

Workload Scheduling does not control the platform servers such as CWMGR1 as they are needed to trigger the Wake on Demand functionality and facilitate other platform tasks and should run 24/7 for normal environmental operation.

Additional saving can be achieved by deactivating the entire environment but is only recommended for non-production environments. This is a manual action that can be performed in the Deployments section of VDS. Returning the environment to a normal status also requires a manual step on the same page.

bw54deploy.onmicrosoft.com	skk	5.4	Azure	1	Offline	Available	<div>Delete</div> <div>Stop</div> <div></div>
cjdevmherr2.onmicrosoft.com	pht	5.4	Azure	1	Online	Available	<div>Delete</div> <div>Start</div> <div></div>

Copyright Information

Copyright © 2021 NetApp, Inc. All rights reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means-graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system-without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

RESTRICTED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.277-7103 (October 1988) and FAR 52-227-19 (June 1987).

Trademark Information

NETAPP, the NETAPP logo, and the marks listed at <http://www.netapp.com/TM> are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.