Recast

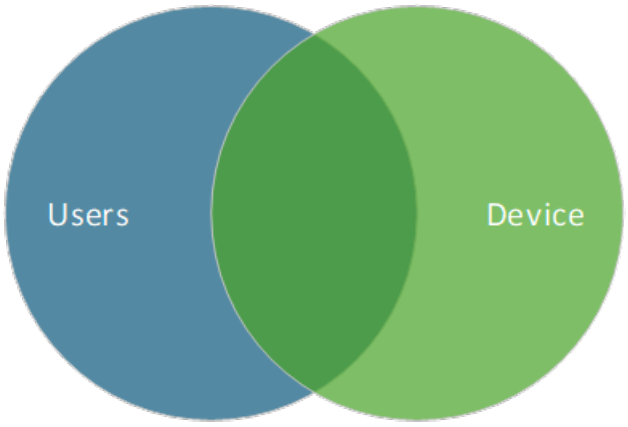
Scaling an Application Workspace System

Last Modified on 2026-09-28

Scaling an Application Workspace System is a hard task. The reasoning for this has to do with a lot of factors. In this article we list the most important factors:

Virtual Workspaces (virtual objects in the server memory)

They are held in the Application Workspace Server memory for every user and device. As soon as we can create a relationship, we are merging them into one Virtual Workspace. Therefore, we cannot calculate based on device or users; if we take the following diagram as a representation of the aforementioned virtual workspace, we need to calculate the complete surface of the diagram.



Content

Application Workspace can serve content to the Agent in a lot of ways. But the server load is more than double if no offloading is done. The offloading can be done by enabling Azure Blob with direct access or by configuring other Content Access Endpoints like an Application Workspace Satellite Server or a DFS replication share with direct access from an Agent. The size and the number of content items also greatly impacts the server's load. In a case where all Application Workspace Servers have local content storage on the VM, the Application Workspace System needs to replicate the content items between servers and that generates a significant increase of the load on servers.

Latency

The Application Workspace System performance can vary greatly based on the connection latency between the Agent and the Application Workspace System. Things like SSL offloading by a load-balancer can benefit an Application Workspace Server.
The Application Workspace Server streams the uploaded content through the memory of the server to the content location. The latency to that content location has a lot of impact on the performance. But the biggest latency impact is between the Application Workspace Server and the MS-SQL servers. If there is high latency or high load between them, Application Workspace will feel slow and unresponsive.

Package

Application Workspace needs to serve packages to devices and users. The number and size of packages, the frequency of the action sets and the filters can impact the load. We recommend you to be careful with the usage of the Refresh agent type of action action inside a package because it can cause looping.

Entitlements

Application Workspace caches entitlements on the Virtual Workspaces. It is a calculation based on the user and device package entitlements. The system load can vary depending on how the entitlement is done (direct, nested, or by a dynamic group). Restricting the entitlement to a single group or context reduces the load compared to creating separate groups for all packages. All the changes in entitlements can determine the need to reload the Virtual Workspace which can cause load on the servers. Events triggered on entitlements can cause a package to run every 30 minutes, which could have a high impact on the system.

Connection

The latest Application Workspace Agent offers significant performance improvements and includes features designed to reduce the load on the Application Workspace System. Old or homemade portal integration can cause high load on the backend as well because it does not honor the caching mechanism managed by the Agent.

Zones

Application Workspace can host multiple zones on the same Application Workspace System. You need to calculate all the individual zones and plan based on the sum of them.

Other

There are probably a lot more factors that could positively or negatively impact the Application Workspace System. The general solution we recommend is to monitor the Application Workspace System and up/down scale accordingly.

Example of a setup

Our advice for most customers is to have the system described below as the starting point and expand it by continuous monitoring and tracking via the Insights functionality.

Start with 2 Application Workspace Servers that can handle 3000 Virtual Workspaces on each server having the following specifications:

- 2 vCPU
- 8GB RAM
- Azure blob storage + direct access
- Load balancer with SSL offloading

We estimate that each Application Workspace Server can handle around 3000 Virtual Workspaces. We calculate Application Workspace Servers with the N+1 for the case in which a server can be placed in maintenance without having impact on the overall system. For this system start with a SQL DTU of 50.

You can scale out or up by adding more servers to or increasing the CPU and memory of the Application Workspace Servers.