Stratusphere UX SpotCheck for Azure Virtual Desktops

Last Updated: 03/04/2025

SpotCheck Methodology Guide

Contents

Document Purpose: To define metrics and thresholds for a SpotCheck as it relates to User Experience in an Azure Virtual Desktops environment utilizing Liquidware Stratusphere UX.

This document is designed to bring together recommendations from many experts in the industry about the metrics that need to be monitored and the thresholds that are deemed acceptable as it relates to User Experience. This document does not make recommendations on changes needed due to the many industry, usage, costing, and application variables that are in play.

What is a "SpotCheck"?

A SpotCheck is a point in time health check that focuses on key User Experience metrics with known acceptable performance levels. The review of data from multiple dates and times is critical before making recommendations or changes to the environment. These thresholds represented below are taken during a one hour level of granularity unless otherwise specified in the description and are key areas that affect User Experience. Normal/High Usage dates and times should be examined based on the industry and user requirements.

Critical Notes:

A. Know your company!

- Know your industry/company/department work habits, loads and applications are critical for data interpretation and threshold evaluation.
    1. Example: Moderate/High storage latency may be acceptable during shift changes with large number of users logging in and out, but this is not acceptable during normal work hours as this impedes productivity.
    2. Example: Law firms and healthcare organizations generally need/want sub ten second login times whereas most other industries are satisfied with under 30 second.

B. Know your data!

- There are many monitoring and diagnostic solutions out on the market. Each of the solutions collect data differently and have different levels of granularity. All of these solutions render/report the data in differently and with unique granular roll ups that can drastically change the data and perspective for the user. For this reason, the metric values represented in this document are only for Liquidware Stratusphere UX and may not apply well to other products.
    1. Example: Depending on the view, you could be looking at averages, peaks or peak averages. Did the data come from the Broker, Hypervisor, "In-Guest", "In Band" or "Out of Band"? How much impact did the "In-Guest" Agent put on the OS? How much impact and time lag is on the "Out of Band" Agent, Broker and Hypervisor?

C. Good Blogs

- SpotCheck Methodology
- Grey Matter is Required – Automated Solutions don't work
- Monitoring vs. Diagnostics

D. Liquidware Community Site and Other Important Links

- **Liquidware Community** - Slack
- **Other Important Links** - SE Field Articles, Product Documentation, etc…

Information Needed for Analysis, Conclusions and Recommendations:

A. Multiple SpotCheck Dates
- MM/DD/YYYY (Monday), MM/DD/YYYY (Wednesday), MM/DD/YYYY (Friday)

B. Multiple Time Frames Examined on each date
- (Time frames for review are based on business requirements)
  9-10AM, 10-11AM, 2-3PM, 4-5PM

The system(s) should be examined on multiple dates and times for the following information based on the max values shown below. **Please do not make a change based on a single data point.**

Critical Sections for Review:

A. Machines/OS Criteria:
- Machine Last Boot – Critical Question – How long has the machine been running?
  1. See Appendix R for more details.
- Login Delay (Industry Average is under 30 Seconds – This is a company preference)
  1. See Appendix L for more details.
- Application Load Time (Industry Average is under 3 Seconds – Company Preference)
- CPU Utilization (Max 80%) – Higher than 50% generally is bad over 60 Minutes
  1. This generally denotes stuck or run-away process(es) on the machine.
- CPU Queue (Should not be more than 1 per vCPU assigned to VM)
  https://technet.microsoft.com/en-us/library/Cc940375.aspx
- Memory Usage (Should be less than 80%)
- Best Practice is to reduce Windows Paging
- Page File Usage (Should be as close to zero as possible)
  1. Windows paging cannot be stopped.
  2. Do not turn off the paging file in windows. Set to minimum and maximum size of the page file.
  3. Do not use "System Managed" – Set the page file start size to ¼ the memory.
  4. Windows Paging causes CPU and Disk Overhead and should be reduced whenever possible. To reduce paging, allocate more memory to the virtual machine.
  5. Soft Page Faults occur in memory and Hard Page Faults occur to the disk.
- Disk Queue (Should be ZERO for 99% of Users)
  1. Disk Queue shows that the OS is waiting on disk reads/writes.
  2. This can be caused by antivirus holding up the IO or latency of the disk sub system.
- Graphics Intensity will be noted has high when over 100 for more than 1/3 of users.
  1. This must be examined to see if graphics off load processor would help
  2. See Appendix G for more details.
- Applications Non-Responsive – (1 per Day/Per Machine/App is OK)
  1. Any more than this requires investigating the apps and services used by the application.

Appendix:

Appendix A: RemoteFX - Remote Display Protocol

- Image Quality:
  1. This is a good "Base" metric to gage/monitor the user connection quality. Adjusting compression, RDP efficient multimedia streaming for video playback, redirect video encoding and encoded video quality can greatly affect the image quality for the user. It is recommended to tune these items on a per pool basis as all settings may not be required for all users.
  2. Image quality directly affects the number of frames per second sent from the VM to the end client.
- Session Latency:
  1. General Max Observations:
     - New York to California – 30-50 Milliseconds
     - USA to India – 150-200 Milliseconds
     - Interoffice same city – 10 Milliseconds
     - Interoffice same building – 5 Milliseconds
     - Huge Latency Numbers in Stratusphere shows users dropping on and off the network. (Huge Denotes 800+ ms)
- Protocol: (Good and Bad… This is Just info for you)
  1. RemoteFX is 100% TCP with UDP support coming in the future.
- Packet Loss:
  1. Packet Loss with TCP can cause users poor experiences: mouse lag, screen artifacts, slow screen redrawing, typing latency, etc.
- General Recommendations:
  1. QOS (Quality of Services) should be implemented on all routers.
     - TCP should be right under Voice Over IP and Video.
  2. There are many options for RemoteFX Tuning. All tuning for Azure Virtual Desktops right now involves tuning MTU, fragmentation, and large send offload. Test all scenarios and consult the best practice guides from Microsoft and tune/monitor users for best user experience based on the environment
  3. Good Article on tuning of the Azure Virtual Desktop protocol - TCP/IP performance tuning for Azure VMs

Appendix D: Disk Queue/Disk Response/Disk Latency

1. Definition of Terms:
   - I/O - Throughput aka number of megabytes per second (MBps)
   - IOPs - Input and output operations per second
   - Disk Queue - Number of outstanding transactions waiting for processing on disk
     Note: Affected by disk response time and number of filter drivers (typically many) that are between the disk and file system.
     Example: Click on an Excel spreadsheet. Antivirus will scan the document then let Excel read the data. When Excel saves the document, Antivirus scans again before being written.
   - Disk Response/Latency - Read/Write time from the operating system file system to the underlying disk. Applications typically are highly random IO and IOPs.
2. Virtual Machine Thresholds:
   - Disk Response/Latency should generally be less than 1-2 millisecond
   - Disk Queue is preferred to be 0.02 or less over a one hour timeframe
     Note: Cloud machine IO/IOP limits vary by hyperscaler and tier with bursting based on current load.

Appendix G: Graphics Intensity

3. Graphics rendering is a large part of the user experience. Depending on the application it can use MS GDI, DirectX, OpenGL, CUDA, etc., or many other graphics interface drivers/protocols.
4. There is always a misconception that since there are no extremely graphic intensive applications that GPUs (Graphics Processing Units) are not needed. This is not true; Windows and normal Microsoft Office applications have a lot of graphics requirements. All desktops/laptops built in the last 10 years have a GPU. These processors are used by the OS and applications to offload drawing of boxes, circles and other complex shapes from the main CPU and rendering them on the monitor.
5. Non-vGPU enabled Virtual Machines:
   - Turn off hardware acceleration for all applications. Even though you don't have a vGPU instance in the in guest hypervisor tools still has a driver that looks like a GPU to the OS and applications.
6. Applications that have the option to disable "Hardware Graphics Acceleration" should be done unless you have a GPU installed in the host. Most modern application have a GPO that can turn this off. Note that this is generally a per user-based GPO. Microsoft Office, Google Chrome and Firefox all have GPO settings to turn off hardware acceleration.
   - Note: These simple application changes can result in a 10% CPU reduction on your host operating system. Your results will vary based on the OS and applications You can monitor this with Stratusphere.
   - This can result in large cost savings in Azure.
7. GPU enabled Virtual Machines - (GPU enabled machine instances)
   - GPUs are expensive and sometimes difficult to determine if you are getting the most out of them. The resources are allocated per machine and most settings are around the framebuffer (aka GPU RAM) allocation. Stratusphere can determine if the machine/app is using the GPU memory that is assigned to it.
   - Example: Allocated 2,048MB of GPU framebuffer but it is observed that only 768MB is being used with a non-consistent burst to 1,024MB will allow for lowering the framebuffer allocation to 1,024MB will allow for more GPU enabled machines.

Appendix I: Important Links

- Liquidware SE Field Articles
- Liquidware Documentation Site - Official Product documentation
- Liquidware Community - Slack, Blogs, Linkedin, X, Facebook, Youtube, etc…

Appendix L: Login Delay

- Time consumed with users logging into a machine is a large part of the user experience. Stratusphere can breakdown the machine boot and login processes. Due to the complexity of active directory and the environments we can only offer a few guiding hints in this document. For a complete login breakdown session please engage Liquidware SE or partner.
- Domain Controller (DC) Discovery Time
    1. DC Discovery happens at boot and login time.
    2. Healthy response times are 300-500 milliseconds.
- Changing of the DC during boot and login shows a potential issue.
    1. DC Discovery Times over 500ms:
        - DC Overloaded – Cannot process request fast enough.
        - Network latency from the machines to the DC.
        - Sites and Services – Machine/User is talking to a DC in another location.
- Long running processes
    1. AD GPOs, Item Level Targeting and Scripts.
        - Need to review these in Stratusphere Login Breakdown.
        - AD Lookups and Local machine WMI Queries are very slow.
        - Mapping a drive/printer to a machine that does not exist, or the user does not have access to can make the login excessively long.
    2. Antivirus Scanning
        - Don't forget that batch files, PowerShell, VB Scripts are all interpreted languages. Meaning that each line in the batch file or script is executed one line at a time. AV systems scan each line then all the previous lines of the script to ensure it is not a virus.
- Domain Overview
    1. Understand which Domain Controllers are processing logins.
    2. How long was the average authentication process on each Domain Controller?
    3. Understand which Domain Controllers have a large number of abnormal events.
- Physical Desktops and Persistent virtual machines need to be treated differently than non-persistent virtual desktops.
    1. Broken and/or Corrupt GPOs.
        - A yearly (at a minimum) review of the GPOs should be performed. Example: IE7 GPOs should not be applied to Windows 10.
        - Conducting GPO reviews can help with login times and with security.
    2. Sites and Services
        - This is one of the top issues found with Stratusphere UX login break down.
        - A machine in New York should not be authenticating to a domain controller in Canada.
        - With the speed needing to provide Work from Home/Work from Anywhere new virtual desktop pools or new VLANs were deployed supporting these initiatives and zoning properly in the correct sites and services for authentication can be something that is easily missed.
- Animated GIF on how to get to Login Breakdown - Login Breakdown
- Animated GIF on how to get to Domain Overview - Domain Overview

Appendix R: Machine Last Reboot Time

- Knowing how long a machine has been running is a critical question. Applications can have memory, graphics and CPU process "Leaks" over time which can/will degrade performance. Machines running longer than one month are also missing critical security/feature patches that put them out of security compliance and at risk.
- Below is a recommendation only of reboot policies based on experience of Liquidware engineers. This is not a Liquidware recommendation as there are no official recommendations from Microsoft.

  Note: The below recommendations also must conform to company business practices and change control policies.

  1. ***Domain Controllers:***
     - Monthly Reboot – Primarily for OS Security Patches
  2. ***Critical infrastructure machines running Windows Server OS:***
     - Monthly Reboot – Primarily for OS Security Patches
  3. ***Single User Virtual Machines (persistent and non-persistent):***
     - Minimum of a Weekly Reboot – Your mileage will vary based on the applications being used by the users. A Daily reboot is ideal to ensure users have the best experience.
     - Minimum of a Monthly Reboot for OS Security Patching.

Appendix V: VoIP – Voice over IP

- Voice over IP Solutions are critical to business meetings and user to user calls. There are many solutions on the market for VoIP and team chat solutions, but they all rely on the network to provide good call quality.
- Most voice over IP solutions and chat systems can sustain a good voice quality up to 200 milliseconds of latency.
  Poor voice quality is introduced when "Jitter" is over 5 Milliseconds.
  Jitter: Is the difference in latency millisecond to millisecond.
- CPU being overloaded can cause latency and this is commonly overlooked. See Machines/OS Criteria section for more information on CPU utilization.
- Why does Jitter Happen:
  1. User network overloaded with other apps downloading/uploading information.

     Note: Many VoIP solutions can offload voice connections from a virtual machine back to the end user device, thereby reducing latency and jitter.