

The Citrix Administrator's Guide to Understanding and Troubleshooting Citrix ICA/HDX

Written By: Marius Sandbu, Citrix CTP
Goliath Technical Support Team
Updates By: George Spiers, Citrix CTP

Table of Contents

Introduction	3
Defining ICA/HDX	3
A. The History of ICA.....	3
B. How ICA Works.....	4
C. HDX.....	7
D. ICA and Citrix Gateway	8
Adaptive Transport	11
HDX/ICA and its Dependencies	11
A. Networking.....	11
B. Server Performance	13
Measuring ICA Performance	14
A. ICA Latency.....	14
B. ICA RTT	15
C. Frames Per Second	15
Troubleshooting ICA Session Performance.....	16
A. Overview.....	16
B. Impact on User Experience.....	17
C. Troubleshooting	19
ICA Channels.....	22
A. Troubleshooting with ICA Channels.....	24
B. Correlating ICA Channel Usage to ICA Performance Metrics for Troubleshooting	25
Summary	27

Introduction

The goal of this guide is to provide a comprehensive understanding of ICA/HDX, the components that it is built on, how it works, how it can be impacted by network conditions, and how to troubleshoot it. Armed with this information, we hope that you will be better equipped to troubleshoot and resolve what we have come to know as the ubiquitous “Citrix is slow” complaint from end users.

This document was co-authored by the Citrix support team from Goliath Technologies and Marius Sandbu, CTP. Goliath products are used by some of the largest and most sophisticated Citrix deployments worldwide. And, in an environment where support coming from Citrix is not always helpful when solving complex end user experience issues, our support team steps in to assist customers directly. This document is a result of the deep experience and knowledge gained through thousands of hours spent with our customers in the pursuit of solving their Citrix issues.

Defining ICA/HDX

A. The History of ICA

Before we can start digging deep into the technical specifics of ICA/HDX, we must first understand what it is. The Independent Computing Architecture, or ICA is a proprietary protocol developed by Citrix over 20 years ago. The purpose of the technology was to allow for the delivery of applications and desktop computing environments independent of the end user’s computing platform. In other words, to create a client/server computing experience like how Unix applications were delivered during the mainframe era. The goal was to have a way to consume computing resources from any device or platform independent of platform or transport protocol.

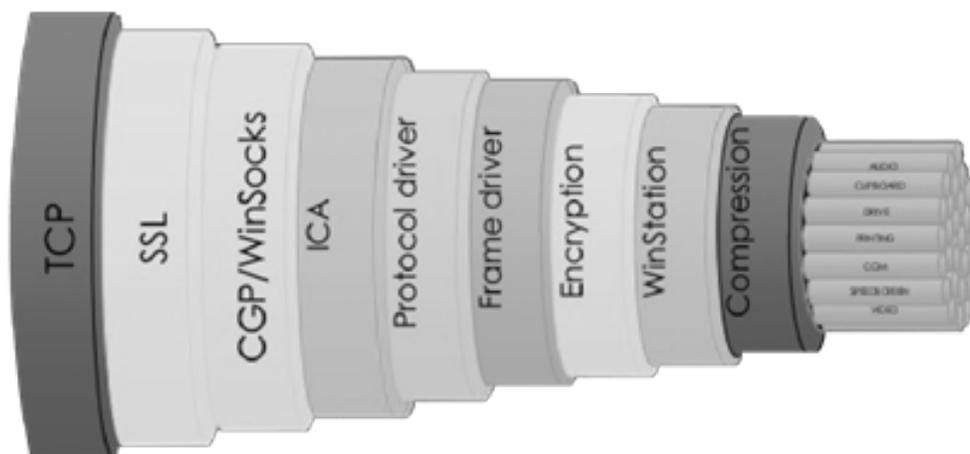
Initially developed in 1992, the protocol was licensed with Microsoft Windows NT as WinFrame. The multi-win engine at its core allowed multiple users to leverage applications from a single Windows operating system simultaneously. In 1997 this technology was licensed to Microsoft and became known as Remote Desktop Protocol (RDP).

Further development from Citrix following the introduction of Windows Terminal Services enhanced the capabilities of ICA, separating it from the basic capabilities offered by Microsoft RDP. The enhancements to ICA allowed Citrix to support application publishing, low bandwidth requirements, encryption and session reliability, among others. This version of ICA is what forms the foundation of the protocol that Virtual Apps and Desktops are delivered from today.

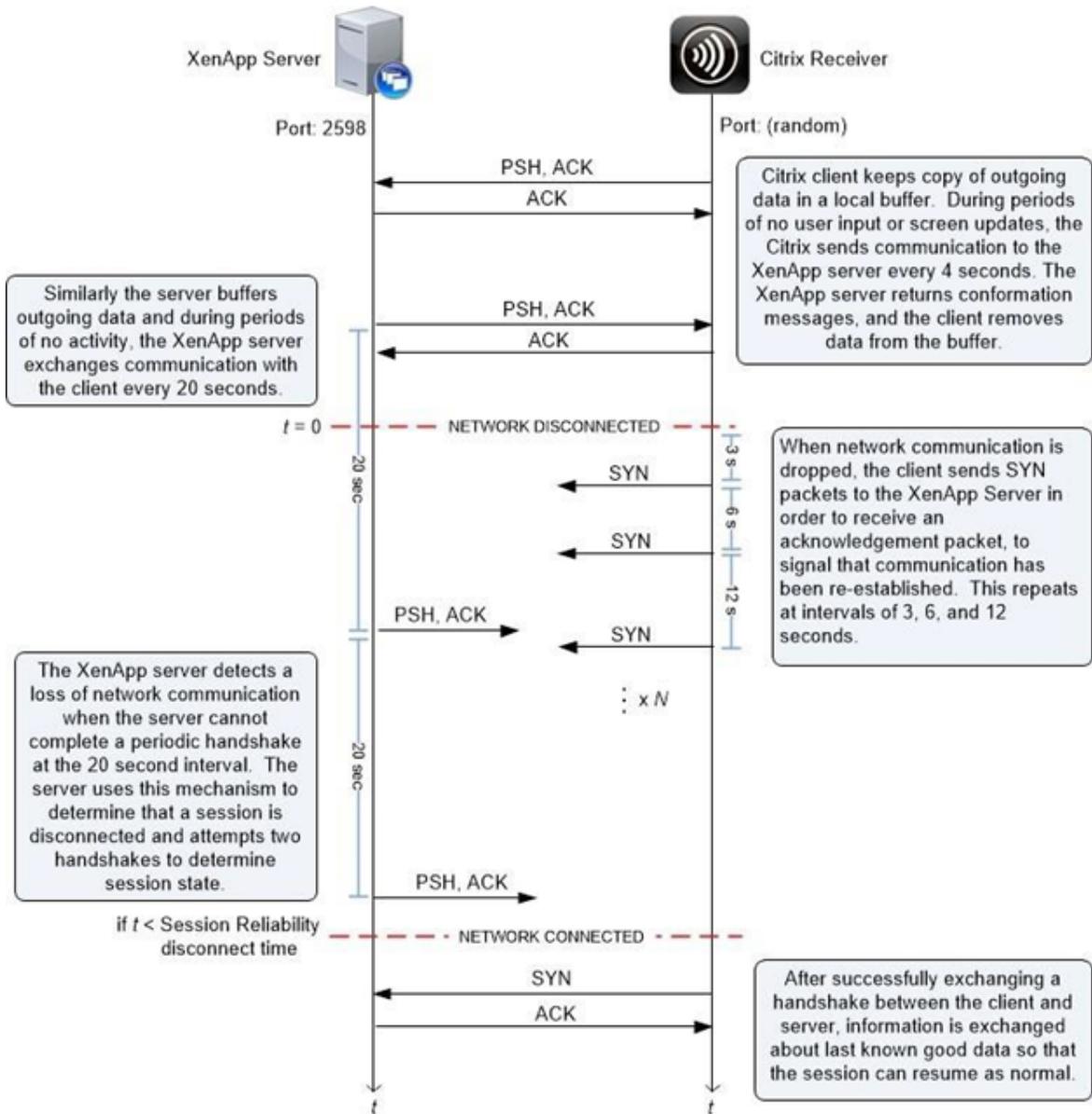
B. How ICA Works

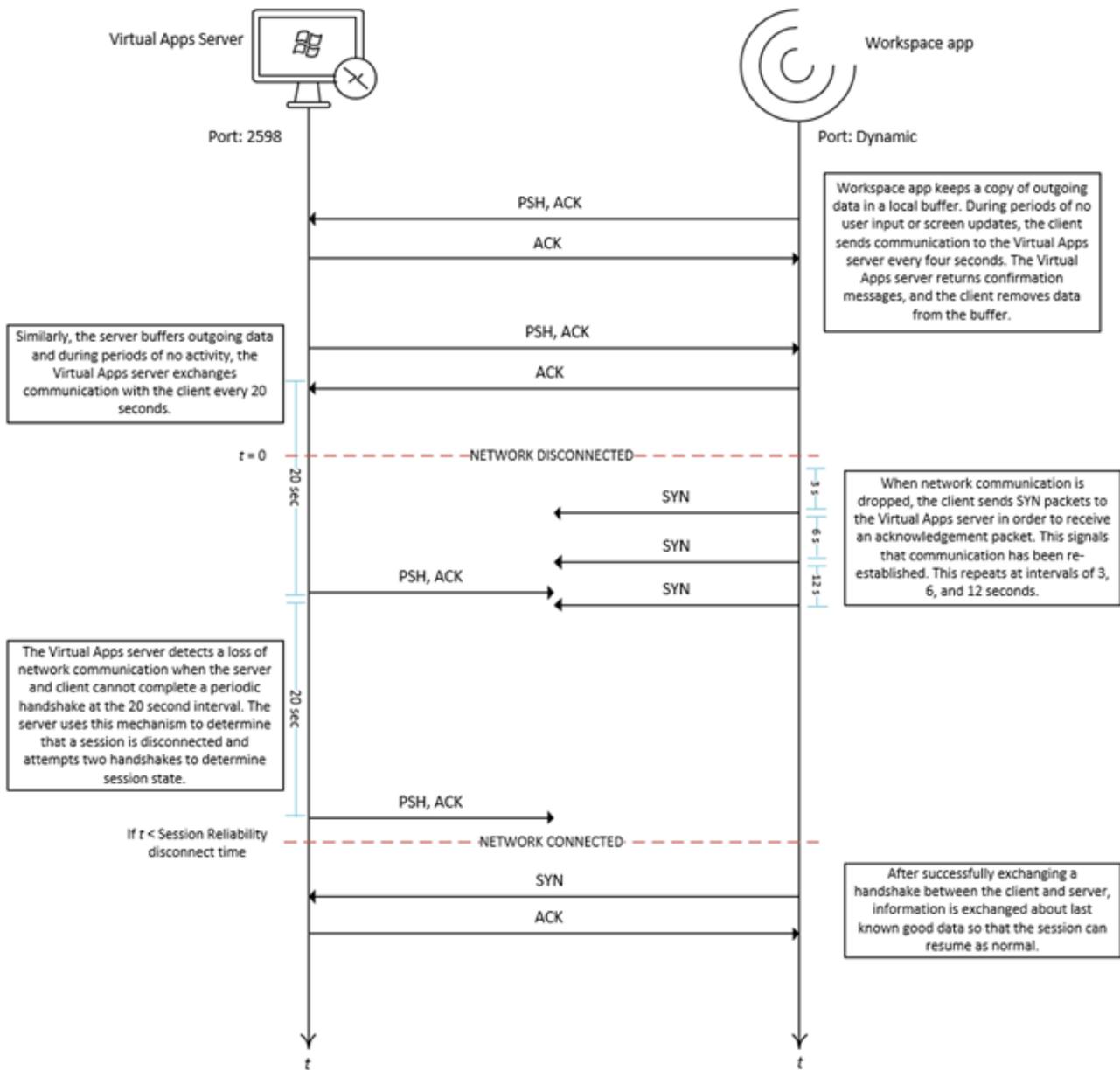
The ICA protocol operates at the presentation layer (layer 6) of the OSI model (we will detail the OSI model later in this document). At this layer, data is prepared to be presented at the application layer (layer 7). It functions by delivering input from the client to the server, as well as providing output such as video and audio from the server back to the user. This presentation layer traffic flows over port 1494 (ICA) and 2598 (session reliability) by creating dynamically allocated TCP or UDP ports for client/server communication. Inside, the protocol virtual channels are used for functions such as printing, typing, video, audio, and USB, among others. Later in this document we will provide more detail around ICA, its functionality and how it can impact end user experience. In addition to leveraging virtual channels for managing traffic, ICA also manages user experience with Session Reliability. This capability allows sessions to remain active on the screen even if the end- user's session connectivity is interrupted. This capability allows for resiliency and consistency even when users are on unreliable network connections or mobile. The specifics on session reliability are shown in the graph below

ICA in Real Life:



Session Reliability On The Wire:





Session reliability is reliant on the Common Gateway Protocol, which is a server -side replay that strips away the CGP layer and then forwards ICA traffic to the ICA listener on 1494. The service buffers traffic if the network link between the client and the Virtual Apps or Desktops server is broken, and presents a frozen screen to the end user.

Once the connection is restored, the buffered ICA data is flushed by the XTE service, and the session continues without any remediation from the end user.

By default, session reliability has a timeout of 180 seconds, after which time the frozen screen will be disconnected if the network connection is not restored. The default value is configurable.

From a bandwidth perspective, the ICA protocol was designed during a time when the most common internet connection was a 56K modem. Needless to say, the protocol is optimized for WAN and high latency network/internet connections. It also supports Quality-of-Service (QoS) and other network features for performance optimization. Over time as applications and user experience have become more graphically intensive, new features have been implemented to help with delivering a quality experience over less than reliable connections. Off-loading of graphics rendering or video display have been implemented to help ease the strain of the ICA session and improve reliability. In the network section of this document, we will discuss in detail how network bandwidth affects ICA/HDX traffic and how it can be optimized to minimize the impact.

C. HDX

HDX has become the modern acronym given to the delivery protocol for Virtual Apps and Desktops. Although many see this as the evolution of ICA, that definition is a bit of a misnomer. HDX in and of itself does not replace ICA. Rather, it is a set of capabilities designed to enhance the user experience for sessions delivered over ICA. There are three key capabilities that comprise the underpinnings for HDX per Citrix. They are intelligent redirection, adaptive compression, and data de-duplication. Each capability works together to:

- ▶ Optimize the user experience and IT control
- ▶ Decrease bandwidth consumption
- ▶ Increase user density per hosting server

Intelligent redirection is an off-loading capability that takes several factors including screen activity, application commands, endpoint device, network, and server capabilities into consideration to determine how and when to best offload processing to the endpoint. This smarter form of redirection also includes devices and peripherals such as webcams, printers, and scanners that all operate at native USB speeds using this capability.

Adaptive compression is proprietary to Citrix and ICA, and it sets the codecs used during sessions. It also intelligently allocates and manages the way the CPU and GPU resources are leveraged.

De-duplication is a form of multicasting and caching that allows for the reduction of duplicate network data traversing the network. Multi-cast is primarily used for multimedia streams to ensure a single delivery of data that may be consumed by more than a single user. Caching de-duplicates data that is commonly accessed such as document files, print jobs, and bitmaps.

HDX are enhanced capabilities that bring a more native feel and user experience to ICA. These capabilities were introduced to allow for graphic and processor-intensive computing to be performed from a virtual computing environment. While HDX technologies accomplish that goal, the foundation remains ICA.

D. ICA and Citrix Gateway

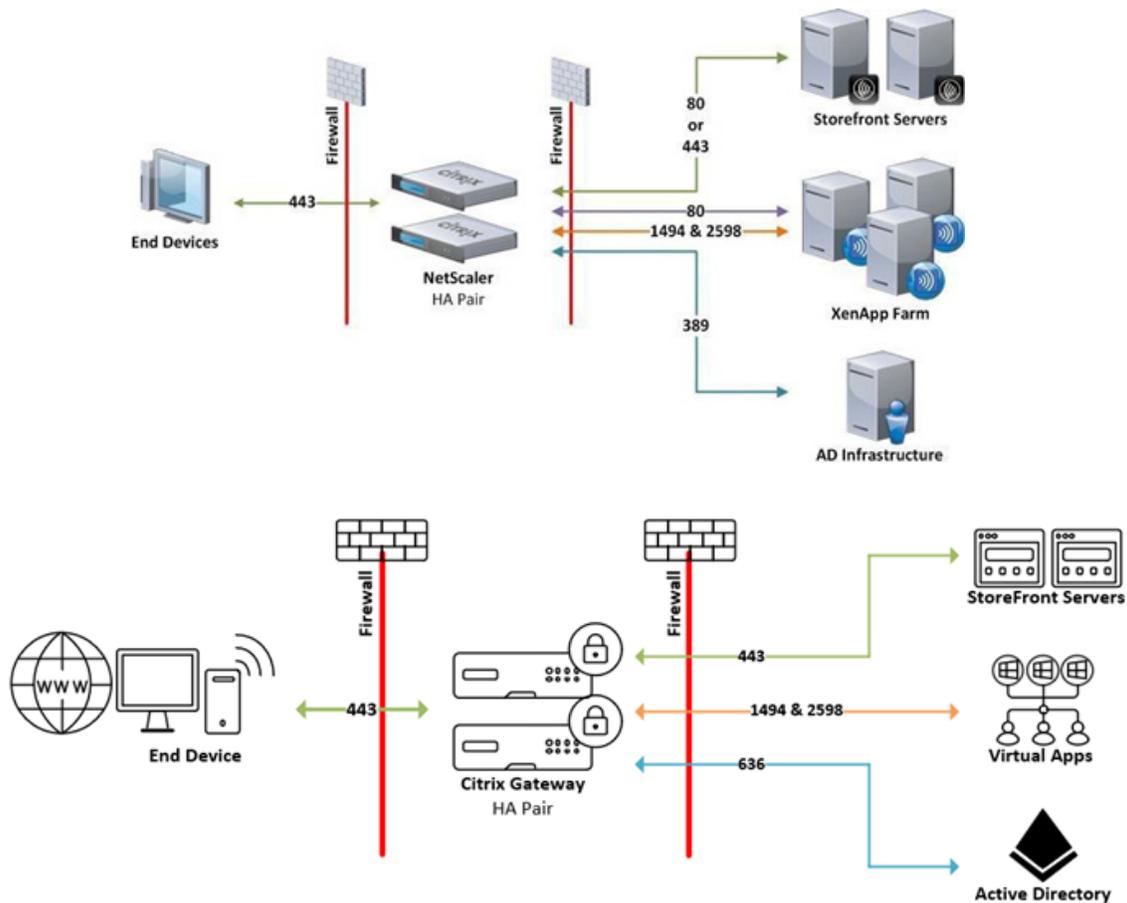
As previously discussed, Citrix ICA was built for low speed, high-latency connections, making it an excellent remote access protocol. The very nature of how it operates ensures a higher quality user experience than a full or split-tunnel VPN. It is also far less expensive than a large-scale optimized WAN solution. However, there are some challenges to deploying an ICA delivered application or desktop to users remotely, natively, as listed below:

Security: The StoreFront server runs on a Windows Server Operating System, and exposing it to the internet could present a risk. StoreFront may be secured with an SSL/TLS certificate, but that is only for authentication and the session would still then need to be encrypted at the Virtual Apps and Desktops level.

Authentication: Like security, StoreFront servers have a limited amount of integration with multi-factor authentication providers, unlike Citrix Gateway which can apply many different forms of authentication to end users to ensure that access is prohibited to all but those who are truly governed to use the system remotely.

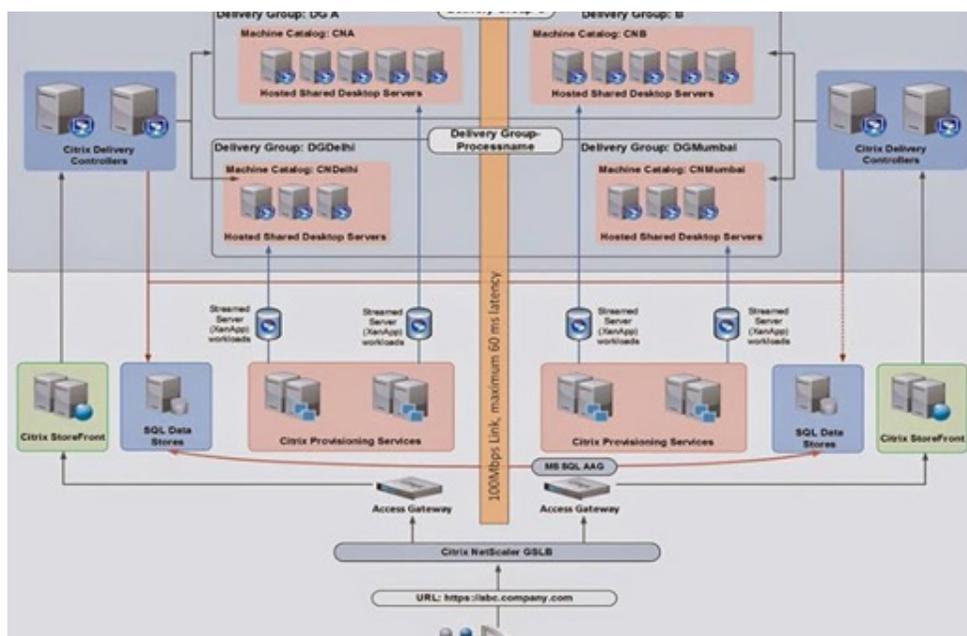
It isn't supported: Citrix simply does not support exposing Virtual Apps and Desktops or the StoreFront server to the public internet.

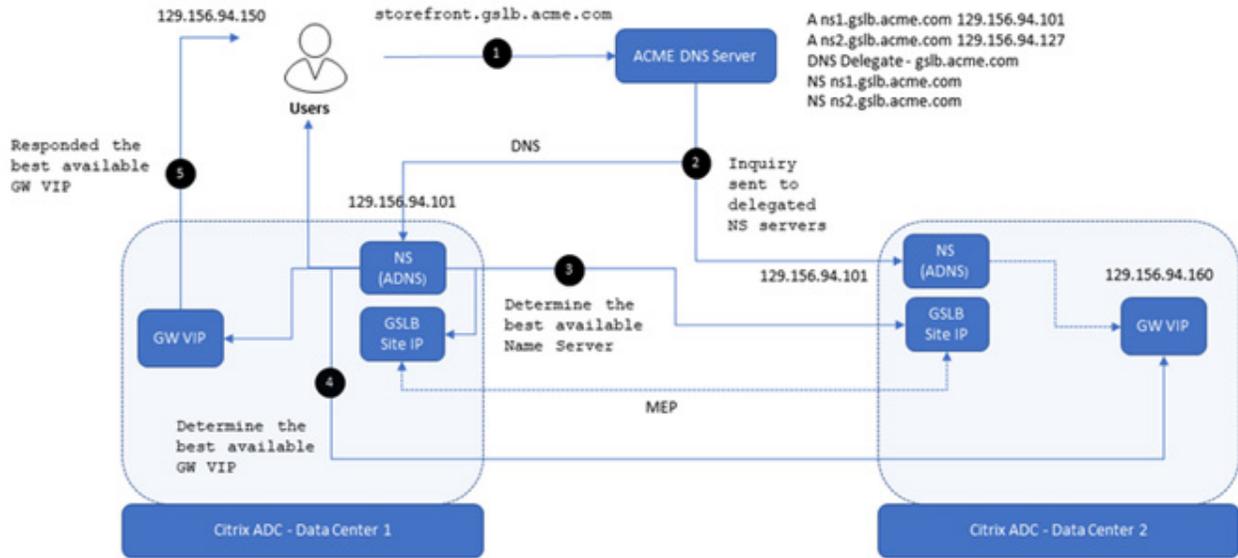
The Citrix Gateway is leveraged to solve the remote access problem. The gateway is a secure FreeBSD appliance that is hardened and designed to be internet facing. All the connections to the internal Citrix Infrastructure



are proxied over an SSL micro VPN tunnel via the Gateway. Therefore, only a single SSL/TLS port needs to be allowed from the public internet to the Citrix Gateway, and then standard ICA and HTTPS ports from the Citrix Gateway allowed into the Citrix environment: The Citrix Gateway accepts incoming connections over SSL/TLS and then passes traffic back to internal services securely over the appropriate port. While deploying the Citrix Gateway does provide a seamless method of remote access, user count and SSL overhead must be considered when sizing the Gateway to ensure that it will not impact overall session performance for the user. Since the Citrix Gateway is the main entry point to the entire presentation layer, there are several architectural decisions that must be considered as to not impact performance and/or availability. The first is high availability (HA). Regardless of if the Gateway deployment is physical (MPX) or virtual (VPX), it is a critical component and should be redundantly configured for hardware failover. All session activity traverses the Gateway and should it fail, user sessions will be disconnected. HA ensures that if an appliance fails for any reason, sessions will continue without interruption.

The second consideration, especially in large and distributed deployments, is Global Server Load Balancing (GSLB). If we have a Virtual Apps and Desktops deployment that spans multiple datacenters and geographies, we want users to access the resources that are most likely to perform optimally for them. If a single Gateway or Gateway HA pair is deployed to North America and users are traversing that North American Gateway to access Virtual Apps resources in Asia, it is very likely there will be additional latency resulting in poor performance. However, if there is a Citrix Gateway in both locations and users access those resources respective to the closest available geography, there is a much higher likelihood that they will have a high-quality user experience. GSLB solves this problem by leveraging DNS to direct traffic to the appropriate location based on several possible variables. GSLB enables a virtual server on the ADC to act as an authoritative DNS (ADNS) resolver for the domain/hostname of the Citrix Gateway infrastructure. Each ADC has the same DNS tables and each vServer is listed as a DNS resolver for the domain. In other words, every ADC in your deployment is an available name server for resolving your Citrix Gateway hostname. When the client queries that ADNS vServer, it returns the IP of the Citrix Gateway based on several policy rules configured in GSLB. These rules could be location based, throughput and performance based, and/or availability based. Regardless of how it is configured, GSLB ensures that a user will access the most available, best-performing Gateway to the presentation layer. The following image details how ADC powered GSLB fits into a Virtual Apps and Desktops deployment running at enterprise scale:





Adaptive Transport

Adaptive transport is a data transport mechanism for Citrix Virtual Apps and Desktops. It is leaner, able to scale, improves application interactivity, and is more interactive on challenging long-haul WAN and internet connections. Adaptive transport maintains high server scalability and efficient use of bandwidth. By using adaptive transport, ICA virtual channels automatically respond to changing network conditions. Adaptive transport uses the Enlightened Data Transport (EDT) to transport ICA traffic between the end user and Virtual Apps or Desktop server. Adaptive transport improves data throughput for all ICA virtual channels including Thinwire display remoting, file transfer (Client Drive Mapping), printing, and multimedia redirection. The same setting is applicable for both LAN and WAN conditions, but is more favorable towards WAN connections where latency is more prominent.

HDX/ICA and its Dependencies

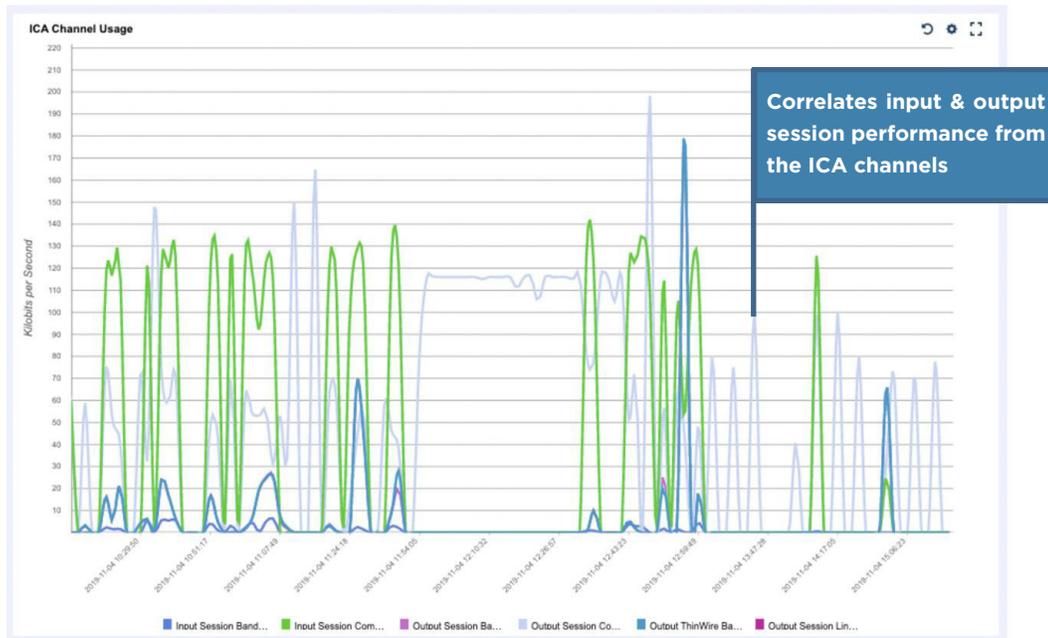
One of the challenges with Citrix, whether with implementation or troubleshooting, has been its dependencies on other parts of the architecture. In fact, the condition of 'ICA latency' alludes to the strong relationship between networking and ICA performance, but is also calculated with consideration to the resource availability of the Citrix Session host. Before we delve into how to troubleshoot ICA latency, we must first understand its performance dependency to the network and session host.

A. Networking

Because ICA/HDX is a remoting protocol and commonly used to deliver users, across geographic distances and various connection mediums, access to resources, an enormous amount of responsibility for a good end-user experience is dependent on the network layer. There are four primary characteristics that affect ICA/HDX behavior, influencing the quality end user experience: Output session line speed, available bandwidth, network latency, and packet loss/TCP retransmits. While many are familiar with these concepts individually, let's explore their relationship specifically to ICA/HDX protocol and end user experience.

Output Session Line Speed:

Output Session Line Speed or Workspace App Connection Speed is specifically defined as the amount of connection speed available to Receiver as a subset of the overall connection speed for the endpoint. The available connection speed to Workspace app must be carefully distinguished between the endpoint's link speed, connection speed, and what is ultimately made available to Workspace app. If a user is on a thin client, for the most part the available link speed and connection speed is being made available in its entirety to Citrix. If a user is on a laptop or workstation, there may be other activities at the OS layer between other applications, OS updates, downloads, and activities that Citrix must contend with to secure connection speed for the user's Citrix Session. Furthermore, if a user's endpoint includes a 1Gbit network, the card speed may be 1 gigabit, but the possible throughput for a connection will be limited by network latency and may be a fraction of that speed. This is especially true in the case of wireless connections, and in either circumstance the Workspace app connection speed will be a fraction of this speed.



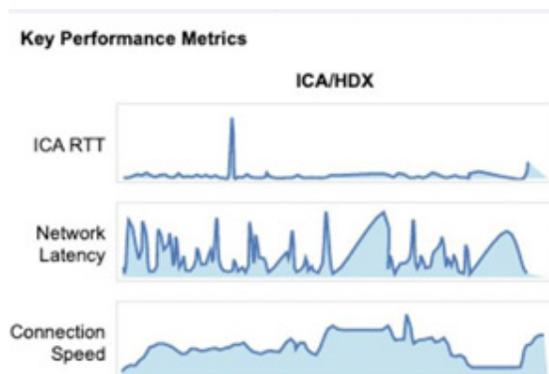
Bandwidth Consumption:

ICA/HDX is highly efficient with bandwidth consumption, using client-side caching and intelligence to only send changing pixels to deliver a high-fidelity experience, even over slow and highly latent connections. User behavior, however, may supersede the protocol's efficiency when multimedia is sent through the channel, or a user prints a large file causing contention with the display channels.

Thus, when troubleshooting, the composition of the bandwidth by breaking down the ICA channels provides good perspective to understand when bandwidth contention is due to user behavior, traffic requirements, or available output session line speed.

Network Latency:

Network latency, as most engineers know, is a measure of time delay from one networked point to another. In the context of ICA/HDX, this must be measured through the channel itself and not through standard network tools. Because the ICA/HDX protocol establishes a microVPN tunnel to the session host, behavior 'inside' the protocol will be different than if it is measured without this consideration. Network teams will use WAN optimization and QoS to optimize ICA packets, or at the endpoint side, it may be deprioritized over VoIP traffic. So, to determine the true network latency of the ICA/HDX packets, it must be measured through the protocol's connection.



Correlating ICA RTT, Network Latency, and Connection Speed can help administrators quickly determine if slow session performance is due to network connection (i.e. home WIFI) or not.

It is worth noting that there are adjustments that can be made on the Citrix Gateway to improve the ICA connection. The default TCP settings on the ADC are configured to be compatible with most network configurations. For ICA traffic, which is “chatty” by nature (e.g., small, frequent packets), the best practice is to define the `nstcp_default_XA_XD_profile` to the Citrix Gateway virtual server, which enables window scaling, changes congestion algorithm, and enables Nagle algorithm. Overall, it will increase the throughput and is better adjusted to handle congestion and packet loss.

Packet Loss / TCP Retransmits:

Obviously, packet loss is not good; when packets get dropped and must be retransmitted, transmissions are delayed, may cause disconnects, and at the very least, especially with a remoting protocol, it will cause slowness and blurriness.

The protocol performs best when there is consistency in connection speed, bandwidth availability, packet loss, and network latency. Consistent variation in any of these will result in a poor end-user experience. This is not to say that the protocol lacks any tolerance for poor connection speed, low bandwidth, or high network latency. As previously mentioned, the protocol was built to handle low speed connections - so how can both be true? The distinction is the variability of performance and consistency. If there is a spike in Network Latency, even to 2000ms, a user may not even notice if it is brief enough, as the protocol will cache the presentation and keystrokes to preserve the experience. However, if the network latency and drop-in connection speed is long enough, the protocol will not be able to mask the change. The presentation will drop to a lower frame rate, causing blurriness but allowing Citrix to refresh fast enough to maintain the connection.

B. Server Performance

The server’s performance impacts the ICA/HDX at two points in the packet’s journey: the processing through the TCP stack and the presentation layer executing the requested action and returning the response data. In fact, a portion of what composes ‘ICA latency’ is the server’s availability to process the requests by the ICA protocol. If a server has its CPU and Memory resources fully disposed, the operating system will not have enough head space to process the ICA requests fast enough, and this in turn will impact the user experience. Consequently, tracking CPU Usage, CPU Queue length, and the CPU Ready state at the host and VM level will help identify if server resources are going to impact ICA performance.

The following is a breakdown of how each element impacts ICA performance:

CPU Usage: Tasks at the user and kernel level, or in terms of the OSI Model the Presentation and Session level, are performed by the CPU. If the OS has CPU resources committed to supporting other processes, there will not be enough resources available to handle the ICA packets.

Memory Usage: Memory usage predominantly supports the application and user environment, and while memory contention would not directly affect ICA handling performance, it will affect the user experience with presenting the delivered application and desktop.

CPU Queue Length: Any delay in CPU processing will result in ICA packet processing delays. CPU Queue length is a proactive indicator of an impending issue with CPU load.

Host CPU Ready: In a virtualized environment, a user’s session host may have low CPU usage, with users experiencing high latency and slowness.

If CPU resource availability is to blame despite the low usage, the contention may actually be occurring at the host level and caused by another VM. The quickest way to determine that is to look at the host CPU Ready and CPU Usage. The ready state is the percentage of time threads spend waiting to execute, like CPU Queue length. CPU Ready over 7- 8% will begin to affect user experience. VM CPU Ready: Similar to host CPU Ready, at the VM level, the CPU Ready state can increase independently of the CPU usage depending on VM level application resource consumption versus a host impacted by another VM's CPU requests.

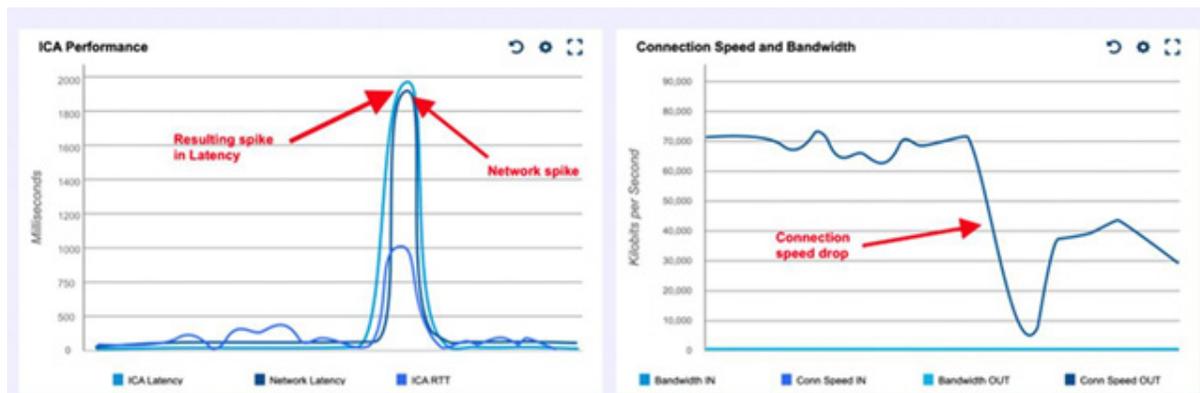
Measuring ICA Performance

ICA Performance and the user's experience is best measured in three key values: ICA Latency, ICA Round Trip Times (ICA RTT), and Frames per second. The metrics are calculated and composed by the ICA protocol based on network latency and server performance. Let's explore how they work.

A. ICA Latency

ICA latency is the time from when a user executes a keystroke or mouse click to when it is processed on the session host. It is inclusive of the network latency and any delay on the session host to process this request. A high ICA Latency should be compared first to the network latency to determine what is causing the delay. If network latency is responsible for the spike, then networking is the culprit; if not, server performance and resource availability is the cause of the delay.

If both ICA latency and network latency spike at the same time it is often indicative of a network bandwidth issue. To confirm it is a network issue, also look at connection speed - if it drops at the same time you can prove to end users and the network team the issue is with the connection (either the speed or allocation of bandwidth to Citrix). An example of this is depicted in the image below.

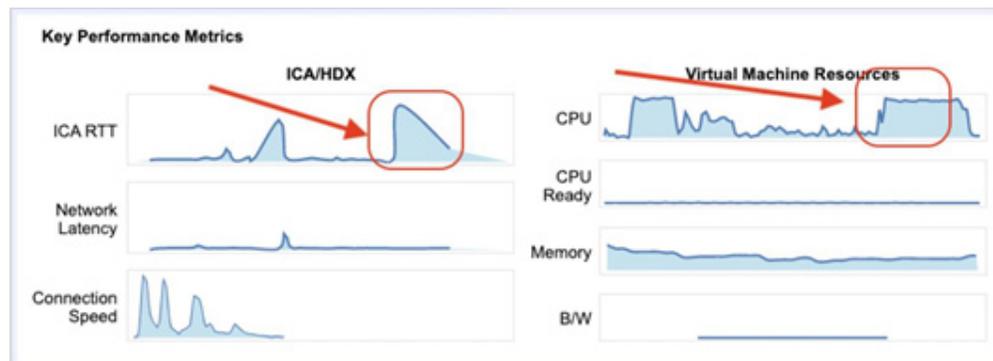


Users will notice slowness when there is a sustained ICA latency over 250ms, and the experience will be significantly impacted when over 400ms for greater than 2-5 minutes. Users that are accessing Citrix-delivered resources from certain regions of the world, especially in Asia, to Europe or North America will often experience ICA latency of at least 180-220ms due to routing and network latency.

B. ICA RTT

ICA RTT is the detected time from when the user hits a key until the response is displayed back at the end point, as calculated by the session experience monitoring service. The ICA RTT should be observed in conjunction with the network latency and ICA latency. The difference between the ICA latency and RTT is the application processing time on the session host. If the ICA latency is the cause of the high RTT, it should be further analyzed using network latency to determine the cause as previously mentioned. If not, then the cause is the application layer and requires troubleshooting of the application and OS performance.

The figure below depicts a scenario where there are spikes in Round Trip Time that correlate with spikes in Virtual Machine Resource utilization. This indicates that resource utilization is actually the driving factor in the high RTT, not network latency.



For those interested in a deeper understanding of how ICA RTT is calculated, it is wholly different from network RTT. In the case of ICA RTT, the client sends special virtual keystrokes when other activity is detected over the ICA channel. The packet is received by the server, processed within the user's session processes subsystems, and a response is sent back to the client. These packets in both directions are encoded with sequence numbers and the client calculates the RTT based on the time difference between starting the ICA RTT check and receiving the response.

C. Frames Per Second

Frames per second is a challenging metric because it can be calculated at several different points. Pure achievable frames per second is calculated at the hardware level, the GPU in the case of customers with NVIDIA vGPU, AMD, or Intel IRIS technologies, however this is hardly useful as the user would never experience this at any point in time. Furthermore, FPS can be calculated at the application level, and while an interesting metrics as it tells us the performance and processing level of the application, it still does not tell us what the user is experiencing. Just like with physical desktops and workstations, the FPS that is ultimately used as a baseline or monitored metric is the FPS presented to the end user.

In a virtual desktop or application delivery environment like with Citrix, this metric is best documented by measuring it at the ICA protocol level. When correlated with ICA RTT, network latency, and connection speed, this metric provides key perspective to any blurriness or pixilation.

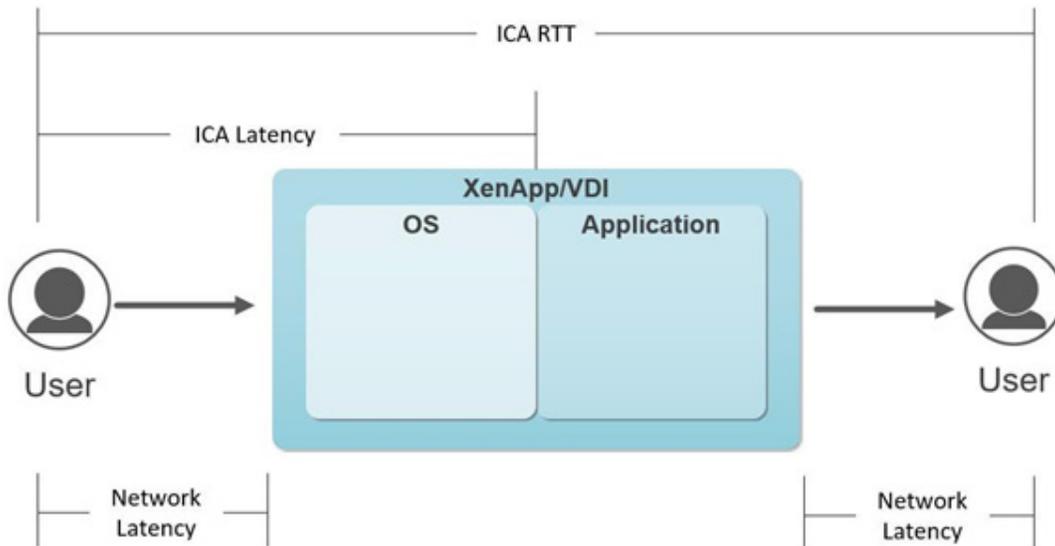
It should be noted that FPS is by default set at 30 on the protocol level. Regardless of how many FPS an application can archive, it will never be more the 30 FPS at the protocol level. This can be adjusted using Citrix policies.

Troubleshooting ICA Session Performance

A. Overview

Before we delve into how the metrics relate to user experience and behavior, we should understand the role they play in a user's connection. The diagram below shows how each metric and its performance relate to each stage of the ICA packet's journey.

- ▶ User executes an action on the end point. Network latency between the end point and session host defines the lag in that stage of the trip.
- ▶ Network latency is combined with CPU processing of the TCP header at the Session host and is measured as ICA latency. Any server-side processing causing delays is calculated as the difference between network latency and ICA latency.
- ▶ OS and Application response to the remote action is processed and the packet is sent back down to the end point.
- ▶ Network lag is the measurable metric for the return journey down to the end point.
- ▶ ICA RTT is the measured metric for network latency from the endpoint to the server + TCP & OS/ Application processing + network latency returning to the end point.



Here is how ICA latency and ICA RTT are broken down into measurable parts. This breakdown allows us to determine the cause of slowness issues and behavior between the session host and the network:

ICA Latency: Network latency should be subtracted from the ICA latency to determine how much delay is coming from the server. It should be noted that network latency may be higher than ICA latency, as it will smooth out latency spikes and improve experience for the end user. The delta between network latency and ICA latency can be attributed to server performance and CPU usage should be checked.

ICA RTT: ICA latency should be subtracted from the ICA RTT to determine how much of the delay is coming from the application layer. The delta between ICA latency and ICA RTT can be attributed to server and application performance. CPU, memory, CPU Ready, and application process performance should be monitored to determine root cause.

B. Impact on User Experience

The impact of ICA latency, ICA RTT, network latency, and connection speed on end user experience generally manifests as blurry screens or pixilation, session slowness, experience of users with highly latent or poor connection speeds, and degraded general user activity.

Blurriness/Pixilation:

Blurriness and pixilation are generally the result of changing line speed, network latency, or FPS maxing out. Citrix includes client-side keyboard and screen caching to overcome the impact of line speed changes and spiking network latency, but when line speed falls to less than 500 kb /sec and network latency is above 400-500ms lag may occur and resolution will drop.

Low Connection Speed & High Latency Connections:

Companies with end users working remotely, traveling, or in Southeast Asia commonly have users with 180-220ms of latency. While Citrix can enable a good end user experience over such high persistent latency, there is not a lot of tolerance on the protocol to overcome spikes in latency and session host resource contention.

Session Slowness:

Session slowness can certainly manifest as blurriness and connection related issues as stated earlier, but root cause may be found in user behavior and the ability of the session host to provide enough CPU, Memory, and storage resources to support it. Screen lag or getting stuck, not to be confused with pixilation, can also find its root cause in resource contention in addition to network latency and connection speed.

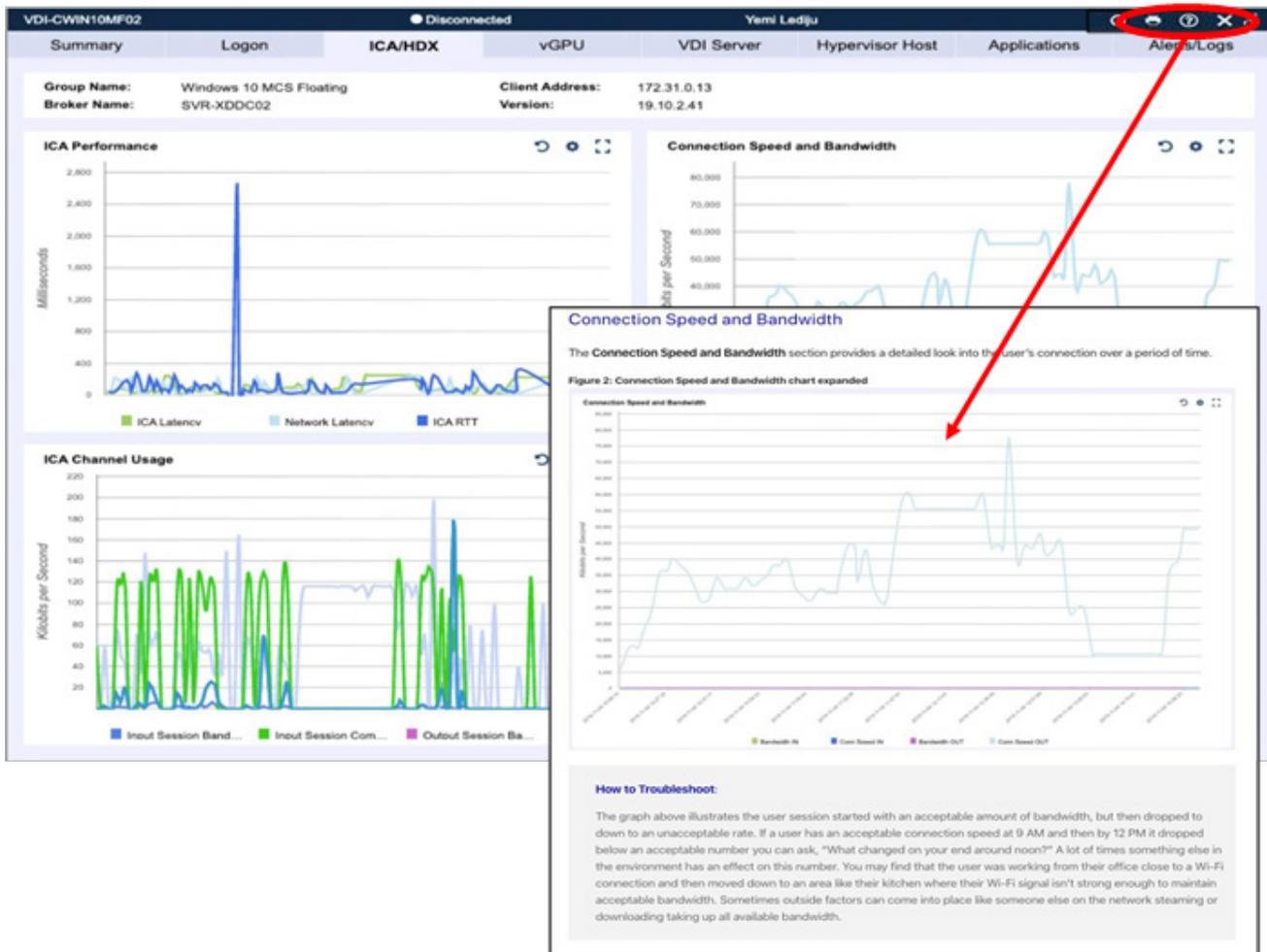
Inactivity:

When a user is not active for any period of time, the ICA channel will stop sending screen updates or passing data. In fact, Citrix will simply record the last data point processed for ICA latency, ICA RTT, and network latency until the session is either reestablished or logged off. When Citrix sessions are minimized or even logged off at the end of a session, there can be a high sustained spike. This behavior is easily recognized when trended as the three metrics are presented as a flatline. When troubleshooting performance problems, data points in these timeframes should be excluded from consideration.

When troubleshooting these use cases, it is often required to not just prove that one of those conditions is occurring, but also that the other conditions are NOT occurring. In scenarios involving slowness there are generally three conditions that can be the root cause and need to be determined:

- ▶ Is the slowness due to a spike in network latency or general networking issues?
- ▶ Is the network fine and the slowness is due to a resource issue?
- ▶ Is the slowness due to user activity that the environment was not designed to support?

In order to troubleshoot, having visibility into the metrics that impact those three conditions are critical. Even more critical is having inline help which not only reminds you of what each metric means, but also advises on how to use the data to effectively troubleshoot (see image below).



Outbound Line Speed:

Beyond the latency issues, did the user's connection speed drop, limiting the available bandwidth to support the user's connection.

ICA Latency & ICA RTT:

Determine the session and ICA connection's ability to support the user activities. This performance should be further broken down between the endpoint to the Citrix Gateway versus the Citrix Gateway to the session host to focus the troubleshooting process.

Network Latency:

Determine if there is network latency above and beyond the expected behavior.

TCP Retransmits:

Packet loss is a primary indicator of issues with user connection, especially when this data is broken down client side versus server side to determine if the problem exists in the last mile.

FPS:

Determine if user activity is contributing to a bottleneck in presentation.

ICA Channel:

Gain context to user behavior and which channels are required to support their activity.

Application Process Performance:

Determine which applications and their usage are contributing to stress on the server and the ICA Protocol.

Virtual Host and VM, CPU and CPU Ready:

Determine if resource consumption and load is impacting the user.

Storage Latency:

Determine if storage performance is impacting user experience.

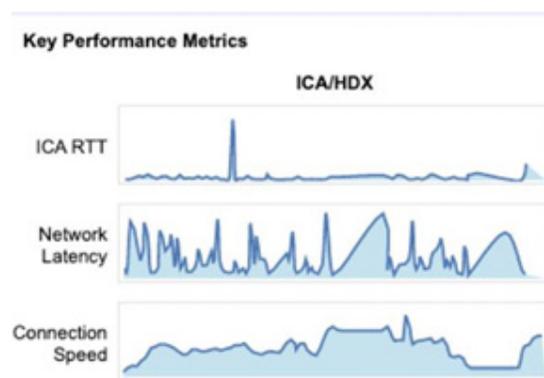
C. Troubleshooting

When troubleshooting, the process is to test each one of the potential root causes from the most basic to the more complex. Here is the process that should be followed:

Identify the users and the sessions where slowness occurred, and, if possible, the time frames when it occurred. Begin with Phase 1.

Phase 1: Determine the culpability of the network in the user experience

- ▶ Review the network latency in their sessions; if there is sustained or persistent latency above 220ms, and this is abnormal, then the problem is with the network. If ICA RTT and ICA latency are being driven entirely by network latency, this is your root cause.

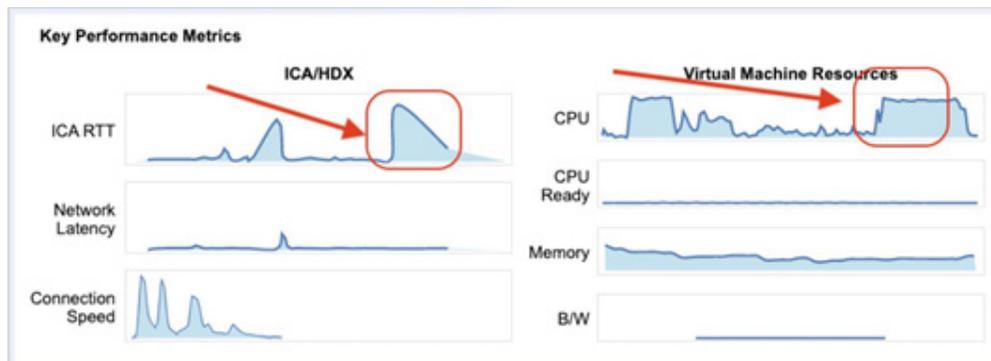


- ▶ Check for Packet Loss and determine if the retransmits are coming from the session host, which would be due to packet loss at the client. This would further point to the networking, and specifically client-side issues as being the root cause.
- ▶ 3. Check the connection speed and ensure that policies for line speed match availability, and that available line speed is sufficient for the bandwidth requirements based on the ICA channel. If outbound line speed is less than 800kbit /sec this is likely the root cause.

Phase 2: Determine if resource contention is root cause

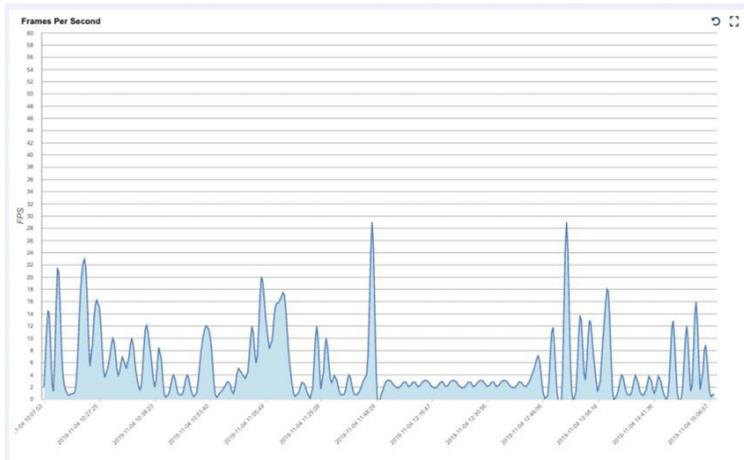
Once network latency and connection speed has been validated, we need to move up the OSI model and review session performance.

- ▶ Start by reviewing the ICA latency values. If the ICA latency values are above 250ms, and it is not being driven by network latency, then the server is struggling to keep up with handling the ICA connections. This is generally expected on App servers hosting multiple user sessions and may be indicative of the fact that the server is struggling to support the concurrent user sessions in either processing the TCP activity or application-level activity. If the issue is with the application activity, there should be correspondingly high ICA RTT, if not reduce the number of users per App server.
- ▶ Review the ICA RTT. ICA RTT is going to include the application's time to respond to the requested action and include any lag at the application layer. When there is high ICA RTT, but low ICA latency, network latency, and enough outbound line speed, then, generally speaking, the architecture to deliver Citrix has the necessary resources to deliver a good experience. However, the resources to deliver the application need to be reviewed and resource contention is a problem.



At this point, in the process of determining the root cause of resource contention, user behavior will be analyzed at the same time and a determination of which is the cause, will be achieved.

- ▶ Check the ICA channel behavior over the course of the session. The goal of reviewing the channel usage is determining what type of traffic and how much is passing through the channel. Specifically, check to see how much bandwidth is being used, as traffic over 500 kb/sec will likely correlate with more resource load. Furthermore, check to see when non- Thinwire (display traffic) based traffic is taking place and if bandwidth usage is contending with connection/outbound line speed.
- ▶ Check the frames per second. Increased Thinwire bandwidth will generally result in increased FPS, but if it doesn't this would result in the experience being constricted. Low FPS and low connection speeds would be indicative of blurriness and pixilation.



A higher FPS number gives a better visual experience but requires more bandwidth and processing on the host machine.

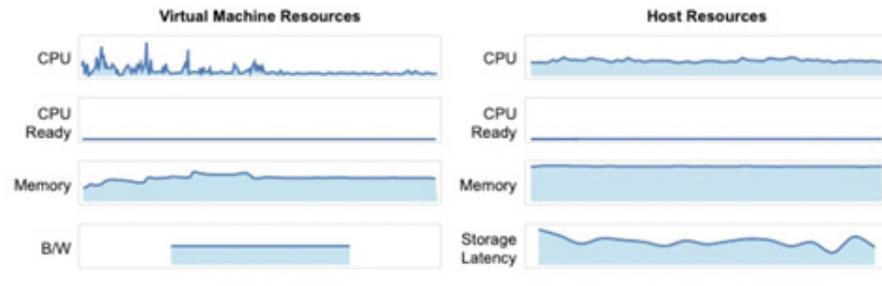
- ▶ Check the Application Performance. Thinwire bandwidth over 500 kb /sec generally results in higher resource utilization if resource consumption is driven by graphics processing. If Thinwire is normal, but ICA RTT times are high, resource contention may be resulting from background process resource consumption. Process CPU, memory, and I/O should be reviewed to determine what is consuming the resources.

At this point one of three determinations can be made:

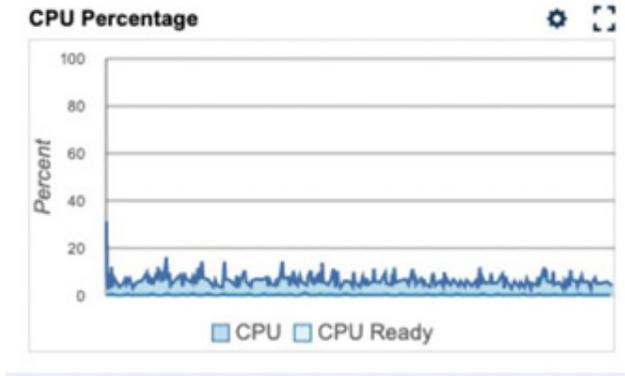
- ▶ If High ICA RTT correlated to high Thinwire bandwidth, then graphics processing was causing the delay, and the process list will advise as to which application was driving the usage. Often, these processes will be using up to or more than a full core of CPU usage.
- ▶ If High ICA RTT, but normal Thinwire bandwidth usage, then either background processes or another VM on the same host is using high CPU or CPU ready resources.
- ▶ Based on the ICA channel usage and the application driving the usage, the culpability of what the user is doing can be assessed to determine if normal or abnormal business operations are contributing to performance problems.

To finish off the troubleshooting process, server and host resource utilization needs to be reviewed:

- ▶ Review VM level CPU and CPU ready usage for the VM the user is on. If it is high, then the user and any other users with similarly high usage are responsible for the performance issues. If the process level usage was low, and/or the VM's CPU and CPU ready usage is low, then the problem is most likely coming from another VM on the same host.



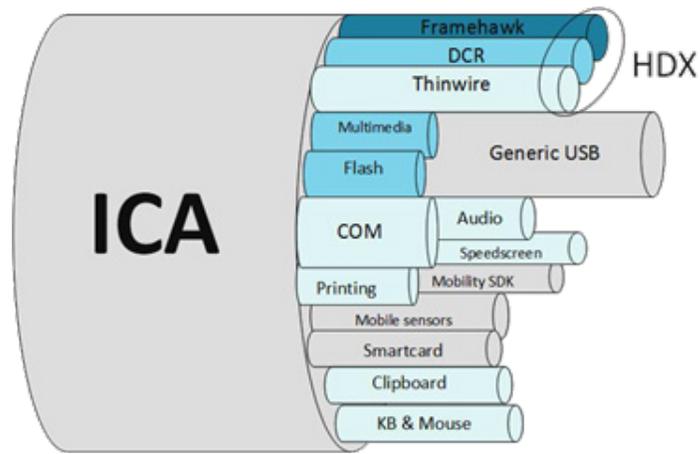
- ▶ Review the Host the VM is on for CPU and CPU Ready Usage. Review performance of other VMs on the same host to determine cause of resource contention.



- ▶ Storage Latency should be checked to ensure its not contributing to performance delays. There have been customers where ICA performance and host resources have been normal, but users experience poor performance. In these circumstances, often the delay has been identified coming from storage latency with persistent spikes up to 7000ms. Sustained storage latency above 150ms should be reviewed.

ICA Channels

To deliver the session to the end point and to retrieve input from the user, Citrix breaks each data type that needs to be remoted into a separate, defined channel. Graphics, audio, disks, ports (com, USB, LPT), printing, smart cards, and even 3rd party custom channels are broken down separately into the remoting protocol. Each virtual channel consists of a client-side virtual driver, as part of Workspace app, that communicates with a server-side application, as part of the VDA. On the client-side each virtual channel corresponds to a driver and DLL, which are provided, often simultaneously, by the WinStation protocol layer. Because multiple channels operate simultaneously, Citrix allows administrators to define policies on channel bandwidth usage and priority to ensure users have a good experience.



From a monitoring and troubleshooting perspective, the ICA channels provide a great mechanism to determine what the users are doing, and specifically how their activity is impacting the ICA traffic. If the greatest impact on the user’s experience is the delivery of the ICA protocol, then the type of traffic that protocol must carry provides a great means of understanding if the user’s behavior – printing, USB redirection, graphic processing, audio, etc. – is impacting ICA performance.

There are over 50 ICA channels including custom channels, but the most common ones that impact user experience and are involved with troubleshooting are the following:

Channel	Purpose	Behavior
Thinwire/Thinwire +	Display protocol. Any changes in pixel presentation, or screen redrawing, is presented through this channel.	Channel usage below 200k /sec is generally word processing and text. 400-800k/sec is generally web page or similar types of traffic. 800-1200 k/sec is graphic rich content, and low-quality video. 1200 k/sec + is high fidelity video traffic.
Audio	Sound and audio delivery. Inbound through microphone or delivery over speakers.	Streaming audio is generally in the 50-60k / sec range, depending on the audio quality set, which is certainly an impact when considered across multiple sessions, rather than a single session, unless the user has a very slow out-bound line speed.
Printing	Print job processing. Print jobs are generally sent in raw format, which can increase the size of the job being sent to the local printer.	Print jobs that begin as PDFs can explode to 8x the size when sent over the protocol. This channel is only engaged when the user sends a print job to a locally mapped printer.
USB Redirection	Redirection of local USB drives.	A heartbeat to ensure the connection to the drive will always occur, but use minimal bandwidth. If a user is reading and opening large files into applications in the session host, this can consume a large amount of bandwidth.
Drive	Redirection of local drives.	A heartbeat to ensure the connection to the drive will always occur, but use minimal bandwidth. If a user is reading and opening large files into applications in the session host, this can consume a large amount of bandwidth.
Clipboard	Redirection of the local clipboard for copy/paste functionality.	The clipboard can contain large amounts of data, which often consume more bandwidth depending on user activity than printing.

A. Troubleshooting with ICA Channels

While the channel breakdown by itself is not indicative of a problem, it does provide important context when used in conjunction with ICA latency, RTT, outbound line speed, and server performance to understand the role of user behavior with their experience. Often, administrators are left to wonder what users are doing in their session, and if a jump in bandwidth or ICA latency is due to printing. By reviewing the ICA channel, the actual activity passing over the wire is known, and thus what the user's applications are requiring the protocol to deliver. Let's look at some scenarios:

Thinwire bandwidth usage:

- ▶ Below 200k /sec is generally word processing and text.
- ▶ 400-800k/sec is generally web page or similar types of traffic
- ▶ 800-1200 k/sec is graphic rich content or low-quality video.
- ▶ 1200 k/sec + is high fidelity video traffic

When looking at the Thinwire utilization, we are looking to determine if there is sustained high bandwidth usage, especially over 1000 kb/sec. This is generally indicative of video traffic, and traffic over 1500-2000 kb /sec is usually HD video. The application responsible for the high bandwidth usage will generally have CPU usage over 30% sustained, and goes up depending on the viewable area the video is being presented in. ICA will only process the changing pixels, usually, in the portion of the screen affected, and so full screen video or large images being viewed will result in large amounts of data and high CPU usage.

Audio Bandwidth:

While streaming audio will generally not be the cause of an individual's session to perform poorly due to its relatively low bitrate at 48-64kb /sec, it is generally a good indicator to the type of activity on a user's session:

- ▶ If the audio channel is engaged in conjunction with high Thinwire usage, then (generally speaking) video is being played.
- ▶ Audio, like the other channels, is presented both inbound and outbound, and as such inbound traffic is usually indicative of dictation activities, which are sensitive to dips in bandwidth and other activities occurring at the same time.

Please note, when utilizing Citrix Gateway all traffic will be transported over TCP, which has a bad effect on audio based upon the nature of TCP which is a reliable protocol, unless extra configuration has been performed to allow the traffic to use the Enlightened Data Transport (EDT) protocol.

Additionally, this can be adjusted so audio in a remote session is sent via the Real-time Transport Protocol, a Citrix best practice recommendation. Some additional configuration beyond EDT is required to implement audio over RTP.

Printing Bandwidth:

This is often expected to be the culprit for poor performance, and on slower connections it still can be a primary cause of slowness, predominantly because print files are generally sent in RAW format which can balloon the size of data being sent. An 8MB PDF can become an 80MB file. Printing jobs are generally sent infrequently and as such are not a problem, except for a moment in time, and network printers take the stress of printing off the remoting protocol altogether as it does not need to be sent back down to the client.

Clipboard:

Users that perform heavy copy/paste functionality, specifically when copied between ICA sessions, will see high clipboard usage being consumed. This needs to be monitored.

Drive Mapping & USB Redirection:

While bandwidth usage, per se, may not be a problem, usage of the channel is indicative of a user not following best practices for a good experience inside of a session. Problems are prevalent in terms of port mapping during reconnects, especially if files on the redirected drive or device are still open in applications, which can cause crashes or errors.

B. Correlating ICA Channel Usage to ICA Performance Metrics for Troubleshooting

Correlating ICA Channel Usage to ICA Latency, ICA RTT, FPS, and Connection Speed to Troubleshoot and Determine User Behavior

The intelligence that Citrix has added to the protocol has logically become sensitive to user input and behavior, both in terms of user action and inaction. This activity is broken down into the ICA channel which allows us to understand what activity the user is engaging in that needs to be broken down and passed to the user at their end point. Further, this data can be instrumental in troubleshooting as it adds context to what user activity is being supported by the protocol and how it is reacting.

Printing, Clipboard, Drive Redirection:

Commonly administrators may see a spike in bandwidth usage and wonder if it can be simply dismissed as the user printing a large document. By breaking down the channel utilization and correlating it to the overall outbound bandwidth, we can determine what channels are being used at any point in time, and further which channel has priority. Printing is specifically called out as a separate channel and is seldom the cause of a spike or sustained usage. If it is, the channel breakdown will quickly identify this condition. From a troubleshooting perspective, this is useful to add context to a user's activity over the course of their session, and where large files may be redirected for remote users, the impact on ICA RTT.

Audio:

Streaming audio will generally use 48-64kb/sec of audio, including when delivered as part of video. While this usage generally doesn't result in saturating bandwidth, it is indicative of the user's behavior,

and when combined with another channel utilization can result in a poor end user experience. In healthcare, customers using dictation software will need to ensure inbound connection speed is sufficient to support recording bandwidth requirements.

So, when determining if there is a correlation between this type of content being accessed by the end users and poor end user experience, we use ICA channel bandwidth, 3rd party vendor channels, outbound & inbound line speed, and network latency. Here's the process:

- ▶ **ICA Channel Bandwidth: The ICA channel breakdown provides key visibility into the different channels being consumed and their impact on the audio channels. This also provides perspective of what other activity is taking place on the session host that is relying on the bandwidth.**
- ▶ **3rd Party Vendor Channels: 3rd-party vendors, like Nuance, have their own custom channels to ensure priority of dictation traffic and provide additional functionality. Monitoring this traffic may provide key insight into how recording traffic is being impacted by the other channels.**
- ▶ **Check the Inbound/Outbound Connection Speed: Low outbound connection speed may become a problem depending on the other inbound/outbound activities taking place - drive redirection, smart card, attached devices, and high-resolution graphics. Inbound speed, especially if the user is coming over wireless, can impact recording performance upon playback.**
- ▶ **Check the Network Latency: Even the ICA Protocol will use caching and intelligence to smooth out network latency in delivering the session. The nature of the audio channel, especially when dictating, is such that it is still susceptible to network latency spikes.**

High Resolution Content, Web Browsing, and Video:

It's no secret that streaming media can negatively impact end user experience, but the reason has less to do with a given website or video, but rather the condition of rapidly changing pixels on the screen that need to be presented back down to the end point. A presentation that is primarily a slideshow with voiceover that has long pauses between new content being presented would put less stress on the protocol than a blog with rotating ads. Remember ICA channel bandwidth for Thinwire - the display protocol can be used to get a general idea of the type of data being passed through the connection:

- ▶ Below 200k /sec is generally word processing and text.
- ▶ 400-800k/sec is generally web page or similar types of traffic
- ▶ 800-1200 k/sec is graphic rich content or low-quality video.
- ▶ 1200 k/sec + is high fidelity video traffic

So, when determining if there is a correlation between this type of content being accessed by the end user and poor end user experience, we use the ICA channel bandwidth, ICA RTT, FPS, and connection speed to determine the cause of the bottleneck. Here's the process:

- ▶ **ICA Channel Bandwidth:** First and foremost, we need to determine the type of content passing through the channel, and if the bandwidth is even being used by presentation versus printing, clipboard, or drive redirection. Review the Thinwire bandwidth utilization and compare it to the data points above to determine what type of data the user is passing through the connection. Check the connection speed or outbound line speed to determine if there is sufficient connection speed at the end point. If there is not, you will see a correlating spike in ICA latency if it needs to scale up connection speed to accommodate or if there is a bottleneck.
- ▶ Check the FPS to see if the protocol is peaking out in the rate it can send data down to the end point.
- ▶ At the times when we see high bandwidth utilization, check the ICA RTT to see if performance is trending over 400ms persistently or sustained, which will result in the user experiencing slowness.
- ▶ If the session host can handle the traffic, and ICA RTT is under 400ms, CPU, CPU Ready, and memory usage should be reviewed to identify any potential problems.
- ▶ Technologies such as Browser Content Redirection and HDX Optimization for Microsoft Teams offload video and audio processing to the local endpoint, which results in a much better user experience due to the reduction of bandwidth and processing requires within the Citrix session.

Summary

For over 20 years the ICA protocol has evolved from a basic terminal delivery mechanism into a robust, lightweight, and powerful platform for delivering a rich virtual computing user experience. Along the way, Citrix instrumented the protocol such that it has become a primary way to determine why a user may have a poor end user experience. In order to effectively support Citrix deployments, engineers responsible for troubleshooting need to have tools that allow them to follow the path from the user to root cause. These tools must incorporate the key data from each platform necessary to troubleshoot Citrix experience issues: Citrix session, ICA protocol, virtualization, and application performance. For those technologies that have ICA visibility, it's specifically important to ensure key metrics such as ICA latency, ICA RTT, network latency (through the ICA protocol), outbound line speed, ICA channel B/W by channel, and FPS are available with CPU & CPU Ready at the VM and host level to effectively troubleshoot.

Goliath Performance Monitor is a comprehensive troubleshooting, alerting, reporting, and remediation technology that allows engineers to correlate a user's performance between the ICA connection, channel, application, Windows, and hypervisor level to identify root cause. Purpose-built for Citrix delivery infrastructures, Goliath Technologies delivers solutions to Citrix administrators that provide deep and wide visibility, allowing them to understand how ICA performance, network performance, and system performance all impact end user experience.

Goliath Technologies also provides capabilities that go beyond simply monitoring the environment. They are truly proactive and give administrators the capabilities needed to anticipate, troubleshoot, and document all Citrix end user experience issues.

Schedule A Demo Or Download a Free 30-Day Trial of Goliath Performance Monitor

Includes Full Support from Goliath Tech Ops

About the Authors

Marius Sandbu, CTP



Marius Sandbu is a Cloud Architect working for EVERY Cloud Services in Norway. Primarily focusing on Cloud-based offerings and EUC. He also has extensive experience in end user computing solutions such as VMware Horizon, Citrix XenDesktop, Microsoft RDS and Cloud solutions like Microsoft Azure, VMware vCloud Air, Citrix Workspace Cloud, Office365 and Microsoft EMS. Marius is a frequent speaker at Nordic Infrastructure Conference and Citrix User Group, and has been one of the few external speakers on the Citrix NetScaler Masterclass.

Goliath Technical Support Team



Our technical support team leverages their collective decades of Citrix experience and expertise in monitoring and troubleshooting end-user experience issues with the express purpose of assisting the IT professional. In addition to supporting our customers and providing critical feedback and insights to our product management and development teams, they also author helpful technical guides and share tips via our webinar and blog series.

George Spiers



George has been a Citrix Technology Professional (CTP) since 2018 and is also a Citrix Certified Expert in Virtualization (CCE-V), a Citrix Certified Professional in Networking (CCP-N), a Microsoft Certified Solutions Expert and an Epic Certified Administrator. He has experience working with Citrix NetScaler, PVS, XenApp, XenDesktop and XenServer leveraging NVIDIA GRID technology for high-end graphical computing.

Goliath Technologies
(855) 465-4284
techinfo@goliathtechnologies.com
www.goliathtechnologies.com

Version: 20220120
©2022 Goliath Technologies. All Rights Reserved