White Paper

Best Practices in Operationalizing ZDX

By Amit Patel
Team Lead, ZDX Deployments

Contents

Contents

1. Rollout strategy

The first step in rolling out Zscaler Digital Experience (ZDX) is to ensure the firewall and process allowlist requirements are met so that any adverse impact on the end user experience is prevented or identified at an early stage of deployment and tackled.

Although the ZDX client service does not make any forwarding decisions or change configurations on the machine, it is recommended to roll out the entitlement in a staggered fashion. Another reason to start with pilot users is to address any environment-specific issues with a smaller rollout rather than finding the problem at a larger scale.

1.1 Zscaler Client Connector version

Since ZDX runs as a module within ZCC, we have a one-to-many compatibility relationship. Ensuring support for the latest ZDX version is key to take advantage of new features and bug fixes.

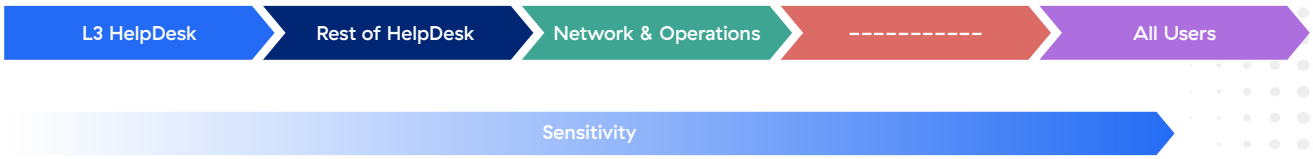
1.2 Pilot rollout

The rollout should begin with the users that are least resistant to change and self-sufficient in debugging basic device issues such as Tier 3 helpdesk.

1.3 Successive staggered rollout

After a week of a pilot group of users running the service without disruptions, creating a staggered rollout plan is next in deploying the entitlement across the entire user base.

It is recommended to start by rolling out across an entire team, such as the helpdesk, before continuing with other teams. The number of rollout phases depends on the organization's size and appetite for change. The recommendation is to onboard up to 5,000 users at a time.



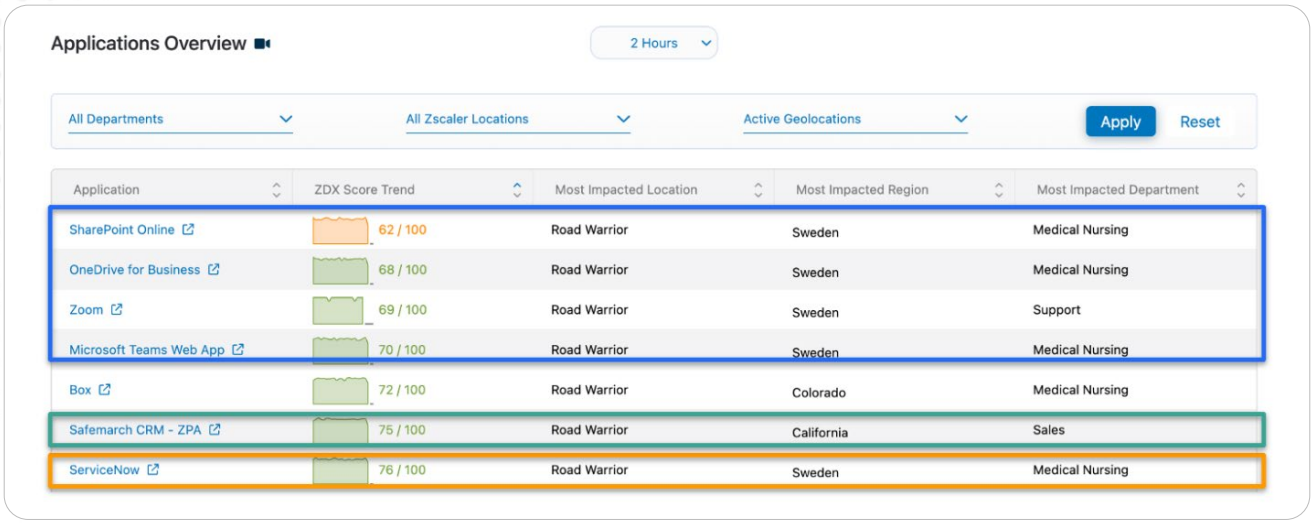
Rollout 101:

1: Satisfy pre-requisites

2: Start small, pilot rollout

3: Progressive rollout across the population

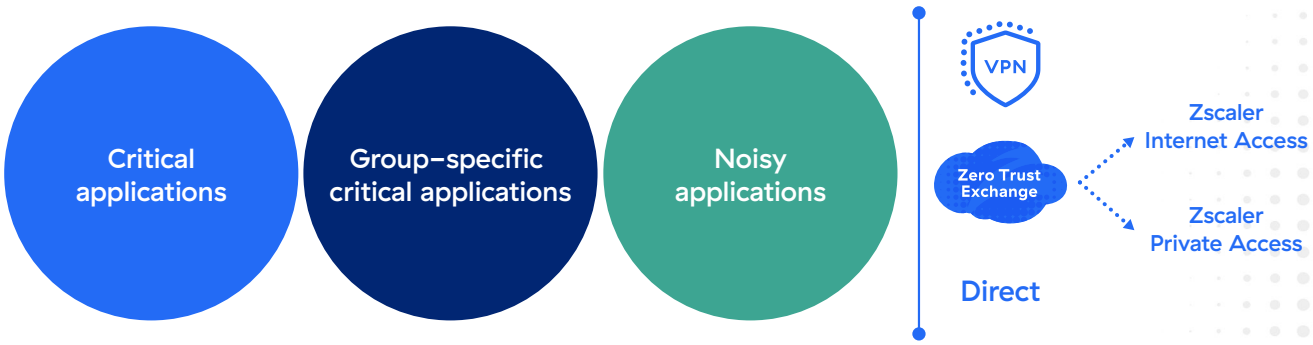
2. Application strategy



To ensure ZDX does not impact the end user experience that we are monitoring in the first place, the number of probes a user can run is limited to 30. This makes identifying which applications to monitor crucial.

2.1 Usage-based

It is recommended to have applications selected based on the three below categories:



2.1.1 Critical applications

These applications are critical from an employee/user productivity perspective and need to be monitored across the board regularly. Common examples include the Microsoft 365 suite, Zoom, etc.

2.1.2 Group-specific critical applications

The group-specific category definition is creating groups of applications segmented based on a team's priority. For example, the PagerDuty application would be critical for the Operations team, but the Salesforce application wouldn't. Rather, the Sales team would prioritize the Salesforce application.

2.1.3 Noisy applications

To derive maximum value from ZDX, monitoring unstable applications and generating the top ticket volumes is essential.

The noisy application category is the only dynamic category, as the application state varies from stable to unstable at a given time. This means it's important to revisit this list periodically.

During quarterly reviews, previously noisy applications can be disabled, and others onboarded. This also leaves the opportunity to run a Deep Trace on disabled noisy applications should a user complain about performance.

2.2 Delivery-based

At any given point in time, an organization relies on a multitude of applications. However, within the context of End User Experience Monitoring, it is not possible to measure performance of them all.

However, a select few delivery paths are utilized. With this in mind, if a good balance between the delivery paths and usage-based app categories is achieved, existing data can be leveraged to baseline performance based on forwarding used. For example, if your organization routes Outlook Online through Zscaler Internet Access (ZIA), that can be used to baseline performance for all apps funneled through ZIA.

2.2.1 Zscaler Internet Access (ZIA)

- Data-driven SaaS applications are generally funneled through the ZIA Cloud Infrastructure.

2.2.2 Zscaler Private Access (ZPA)

- Private applications that an organization manages and are not publicly accessible are primary drivers for ZPA.
- Source IP compliance requirements can also drive apps to be routed through ZPA.

2.2.3. VPNs

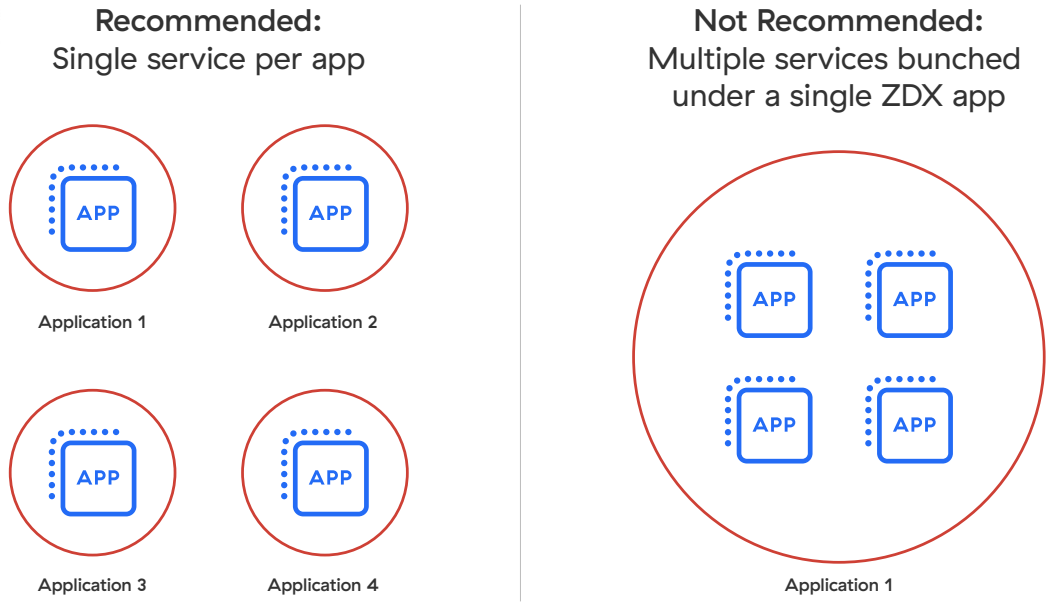
- Organizations yet to adopt full zero trust will have private applications riding the legacy VPN tunnels.

2.2.4 Direct/Zscaler Bypassed

- Real-time applications, such as UCaaS, are usually bypassed from ZIA.
- Users on the production network will usually route DIRECT to private applications.

3. Application arrangement/services

Application layout in ZDX is essential in ensuring dashboards report the accurate status for a particular service.



For instance, let's say SAP is critical for my organization and there are two instances, Europe and Americas, that are independent of each other and either can go down without the other. It's then recommended to set up two applications: SAP Americas and SAP Europe.

Having both services combined under a single SAP application poses the risk of diluting trends for each application and prevents the ability to understand how each instance is performing.

4. Probes

Correct probe setup is at the core of the whole experience monitoring exercise. The below sections specifically describe each factor for the probe configuration.

4.1 Criteria: Run for select users/all

The 30 probe per user limit means utilization of active probes is critical to a platform's health. These approaches depend on the actual probe entitlement of an organization.

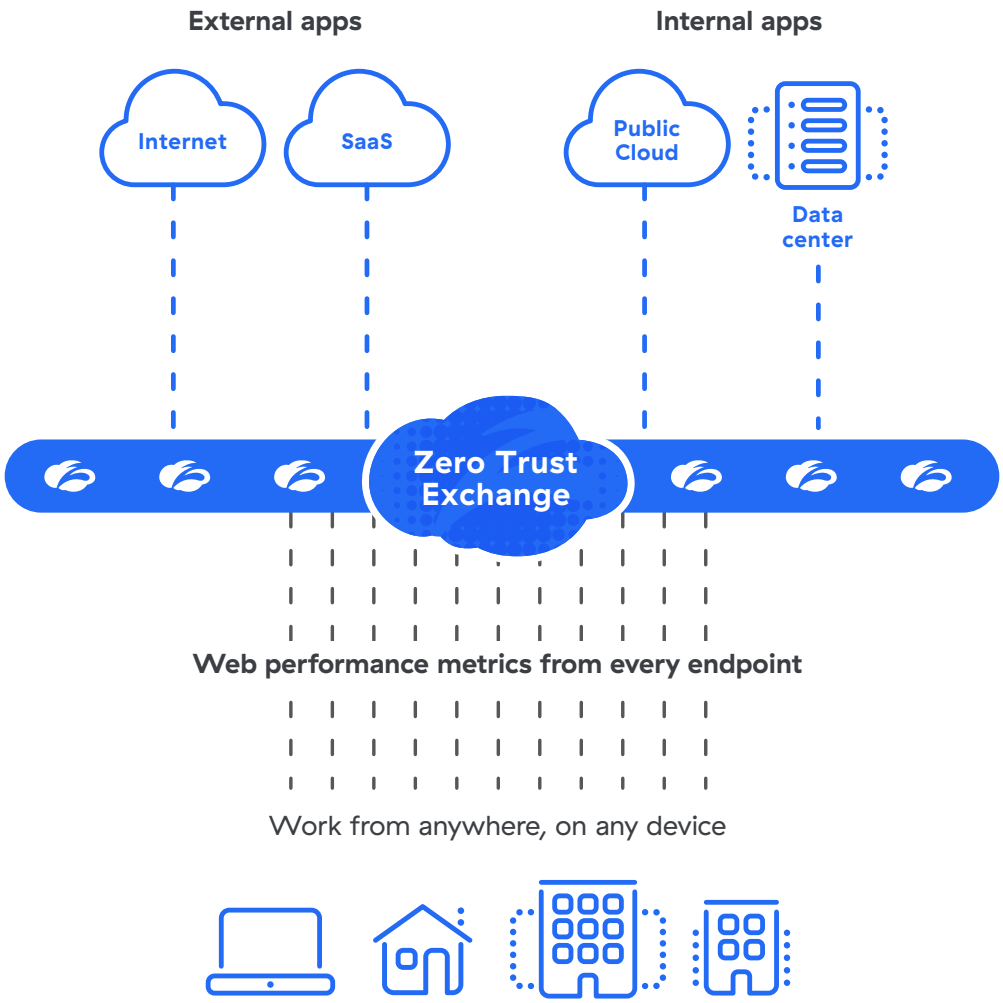
4.1.1 Organizations with 30 probes

Since you can only enable 30 probes on the platform at a time, running each probe from everyone with ZDX enabled provides a more granular understanding of an application's performance. However, if the application is not accessible to everyone, only users with access should be assigned to the probe.

4.1.2 Organizations with more than 30 probes

Here, ensuring the entitled probes can be utilized is vital since the 30 probes per user limit is still enforced, so proper planning is required. The specific group applications in this scenario should only be run for the concerned group, as this helps ensure other users can run probes for maximum applications that are of interest to them.

4.2 Criteria: DDOS concerns



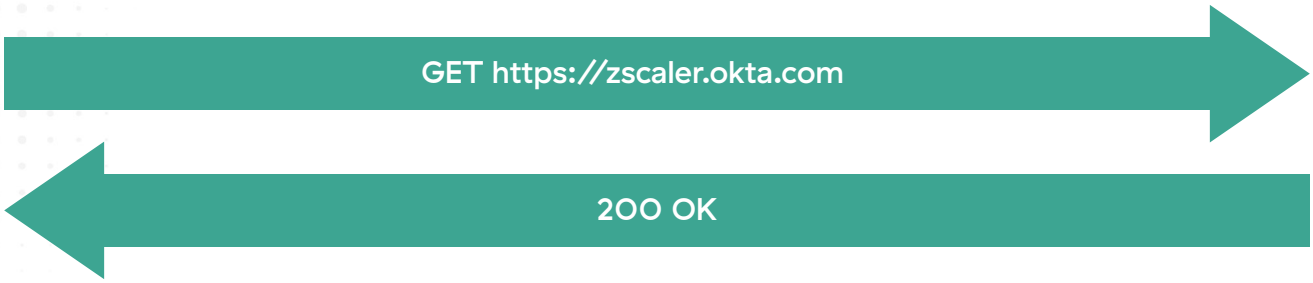
With ZDX, whenever a probe flows through Zscaler Clouds, the service edges perform smart caching to ensure the end service does not get overwhelmed and, at the same time, gives users accurate results back.

When the application is accessed directly by the client, all requests will reach the end service; hence, it's essential to scale testing accordingly. Given that a privately-hosted application might not have the same resources as a SaaS application, starting small with a group/department of geographically dispersed users is recommended. The idea is to keep the total number of users small enough for the service to sustain requests. A Deep Trace can be run on demand if an issue arises for a user who is not running the application probing.

4.3 Web probe redirect/availability best practices

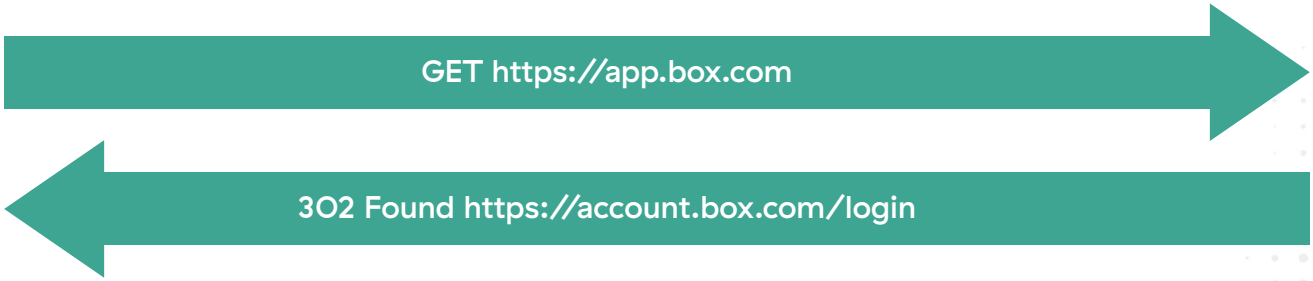
Setting the redirect and expected response codes is critical to ensure the metrics represent the desired performance. The configuration should be based on how the application responds to an unauthenticated HTTP request.

4.3.1 Application that responds with a 200 OK



Here, the success must be measured against the 200 and redirects followed.

4.3.2 Application that redirects to another service such as login with a 302 Found



In this scenario, measuring availability based on a 200 OK coming from a different service than the original renders the metrics skewed. Hence, availability should be measured based on the ability of the original application to redirect.

4.4 Cloud Path best practices

4.4.1 Protocol

Since the internet is "best-effort," and devices on the internet do not consistently respond to a specific protocol for network measurements, it's recommended to use the "Adaptive" protocol for the Cloud Path probe.

In custom routing situations, such as a particular port/protocol being bypassed, it's vital to use the specific protocol and port to ensure correct Cloud Path probe metrics.

4.4.2 Packet Count

Packet Count is the number of traces each cloud path attempts, this plays a pivotal role on the number of connection requests a destination, perimeter firewalls, and intermediate devices receive. Setting this up to an optimum level is critical for efficient probing.

Recommendation is to use a forwarding-based approach here, use 5 for SaaS ZIA//Direct destinations and 3 for Private/ZPA destinations.

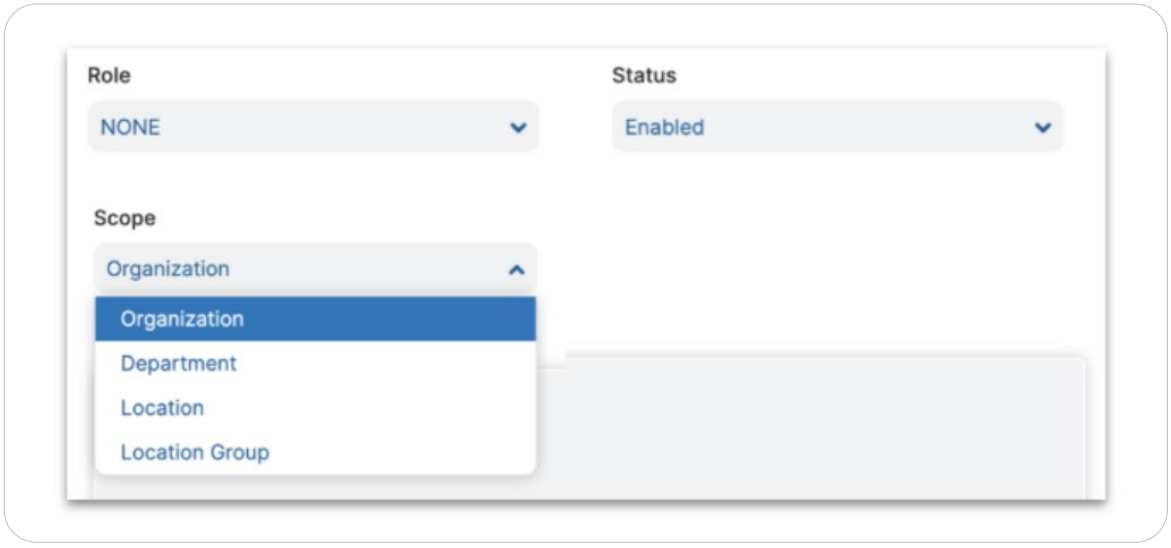
Note: Starting ZDX v3.3.1, packet count is auto-reduced to 3 for all probes when a trusted network is detected to reduce impact. Deep Traces will continue to run with configured values.

5. Admin Access/RBAC

Provisioning sufficient and "just needed" levels of ZDX access to personnel is critical to maintaining platform hygiene. Along with authentication, authorization is a major aspect of Administrator access.

ZDX provides the ability to granularly control the level of access an administrator has in the form of permission-based roles.

These roles can then be assigned to administrators within a predefined scope.



Provisioning access to various different persona can hence be customized, below is an example:

5.1 Platform Owner

Full Read/Write access to the platform. The ZDX Super Admin built-in role should be assigned with organization scope.

5.2 Helpdesk/Servicedesk

RW permission to Deep Trace, no permissions to Administrator and User management. RO permissions to probe configuration and alerts with usernames visible and locations obfuscated.

5.3 Network Operations

RW permission to Deep Trace, no permissions to Administrator and User management. RO permissions to probe configuration, alerts with usernames obfuscated and locations visible.

5.4 T1 Roles

This will be the most restricted role where an administrator is able to simply search for a user and look at performance over the last 2–12 hours depending on the organization's preference.

Here is the recommended RBAC Matrix:

Permissions	Platform Owner	Help Desk / Service Desk	Network Operations	T1 Service Desk
Dashboard Access				
ZDX Dashboard	View Only	View Only	View Only	None
Application Overview	View Only	View Only	View Only	None
Application Dashboard	View Only	View Only	View Only	None
User Overview	View Only	View Only	View Only	None
User Dashboard	View Only	View Only	View Only	None
Device and User Information				
User Name	Visible	Visible	Obfuscated	Visible
Location	Visible	Obfuscated	Visible	Obfuscated
Device Name	Visible	Visible	Obfuscated	Visible
IP Address	Visible	Visible	Visible	Visible
UCaas Monitoring				
Call Quality Configuration	Full	View Only	None	None
Call Quality Meetings	Full	View Only	None	None
Call Quality Applications	Full	View Only	None	None
Analytics				
Static Reports	Full	View Only	View Only	None
Configuration Access	Full	View Only	View Only	None
Administrator Management	Full	None	None	None
User Management	Full	None	None	None

Permissions	Platform Owner	Help Desk / Service Desk	Network Operations	T1 Service Desk
Locations	Full	None	View Only	None
Remote Assistance Management	Full	View Only	View Only	View Only
Deep Tracing	Full	Full	Full	Full
Alerts	Full	View Only	View Only	None
Webhooks	Full	View Only	View Only	None
Zscaler Client Connector Portal	Full	None	None	None
Time Duration	Full	Full	Full	2–12 Hours
Inventory Management	Full	None	None	None

6. Alerting

ZDX collects many data points, but it's not possible to manually sift through them. To address this challenge, you should configure Alerts within ZDX. Alerts drive proactiveness with notifications that act as early alarms before the wider population feels the impact.

The below sections describe best practices with different alert configurations:

6.1 Condition values

Setting alerts on the right conditions is critical to identifying issues before they take shape.

From an end user perspective, performance can be slow, or requests can fail at various stages. Below are the quintessential metrics to be considered:

	Slowness	Failure	Comprehensive
Web Probe	Page Fetch, DNS Time	Availability	ZDX Score, Score Drops
Cloud Path Probe	Latency	Loss	

CPU Utilization, Memory Utilization and Wi-Fi Signal Strength can similarly be configured from the device health perspective.

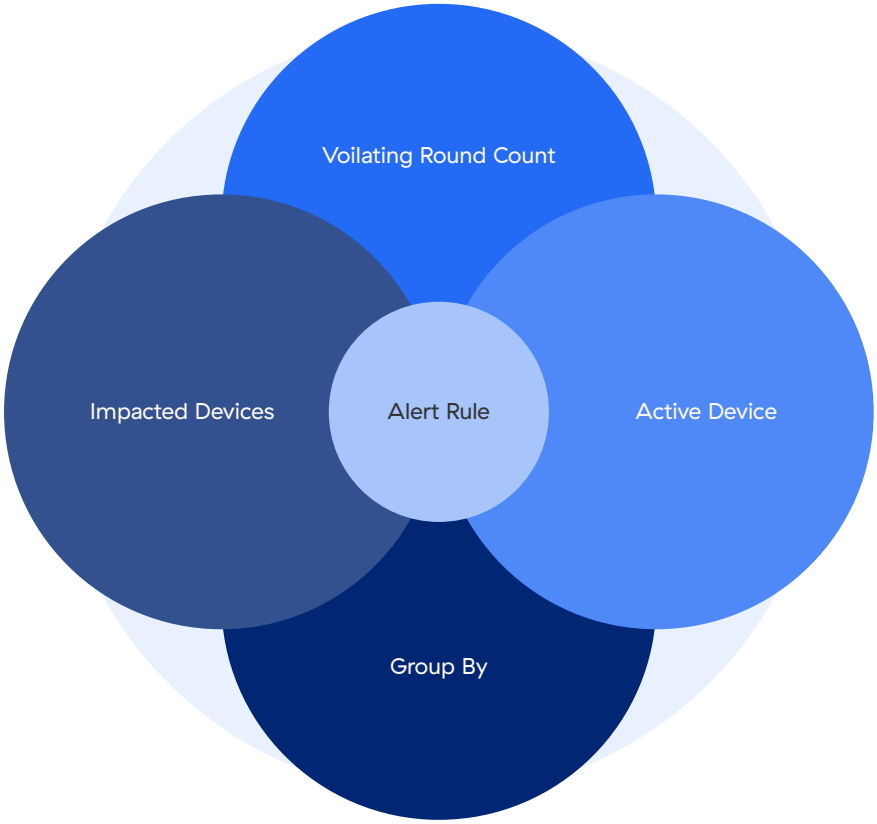
6.2 Nesting

Nesting conditions under an alert rule is an important configuration. The above conditions nested logical 'OR' will cover all the user experience issues.

At the same time, nesting them with an 'AND' won't trigger the alert unless all the conditions are violated. This can cause alerts to stay silent during an actual service degradation and is not recommended.

6.3 Throttles: Violating Count, Devices, and Group By

Setting additional throttles on top of violating conditions helps to get alert rules production ready.



The four available throttling vectors are:

6.3.1 Violating Rounds Count

Setting this to a number >1 is important to prevent “spike” or a “one off” issue alerting, the alerts should rather trigger on sustained violations only, setting this at 2 violating counts is a right balance.

6.3.2 Impacted Devices

This number should be set at the value which translates to a significant number of users seeing bad performance. In order to better scale the alert, it’s recommended to set this to a percentage for violating devices and set a Minimum Device count to the smallest population of the group by entity of interest.

6.3.3 Active Devices

This is very important to set a scalable global alert. This should be set to the total population of the smallest group by entity the organization needs to be alerted upon.

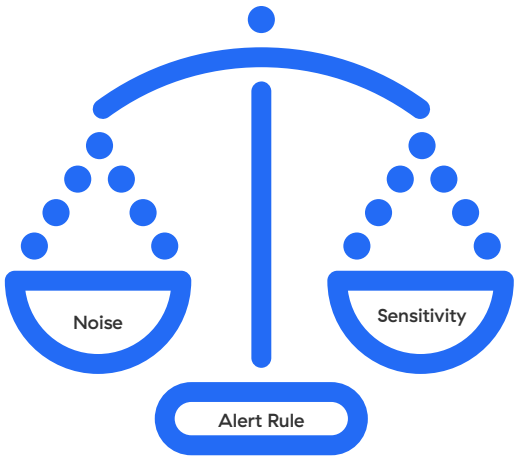
6.3.4 Group By

This value determines the scope for Devices and Violating Count. The preferred value here varies by Organization. An organization with known Zscaler locations available for the majority of users can use the same vs. others can utilize Cities or Regions.

6.4 Tweaking/soak period

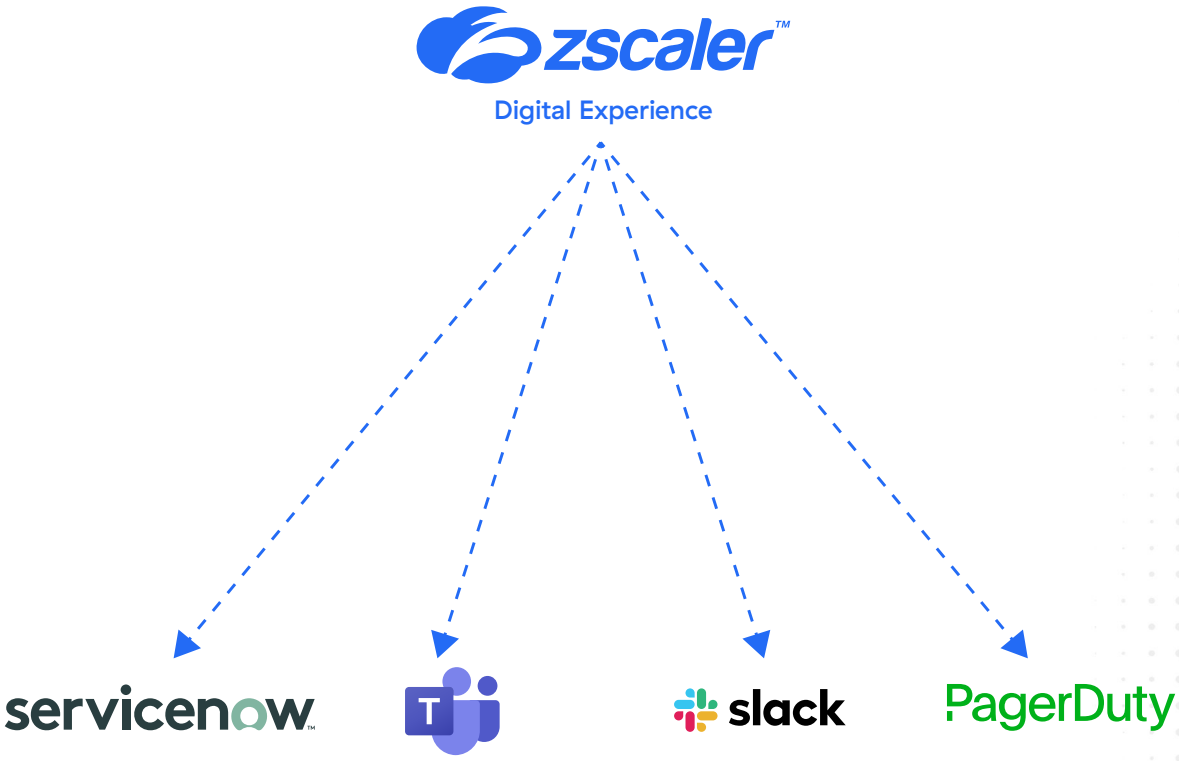
Alerts need constant tuning and tweaking over a soak period to be operationalized. At least a two-week soak period is recommended after ZDX is rolled out to all users.

Getting the right balance between alert noisiness and sensitivity is key to optimal alerting.



6.5 ITSM/IM integrations with webhook

To reap the full benefits of alerting, it's recommended to both configure webhook notifications generating automatic tickets and have a process built to mitigate tickets. This ensures a process is built around alerts and incidents do not go unnoticed for long.



Setting up webhooks to IM tools such as Slack and Microsoft Teams also helps get more eyes on issues before they take bigger shape.

6.6 ITSM Routing

Once the tickets are created, the next step is routing them to the relevant team. Here, an approach needs to be taken depending on the delivery path of the alerted application and the layer of alert rule.

These are our best practice recommendations and proposed handling:

App/Network Alert: ZS forwarded	Network Alert: ZS Bypassed	App/Network Alert: Private App
Route to: Internal Zscaler team	Route to: Internal Zscaler team	Route to: Relevant Zscaler team
Next Steps • If issue in local network/ISP: Forward to internal network team • If issue in ZS network segment: Open a Zscaler support ticket	**Next Steps** • Investigate the problematic network segment/node.	**Next Steps** • Identify if its an Application slowness issue or a network issue and mitigate accordingly

> “ZDX helped us identify the root cause of long-term performance issues and our Helpdesk teams were able to resolve user issues easily.”

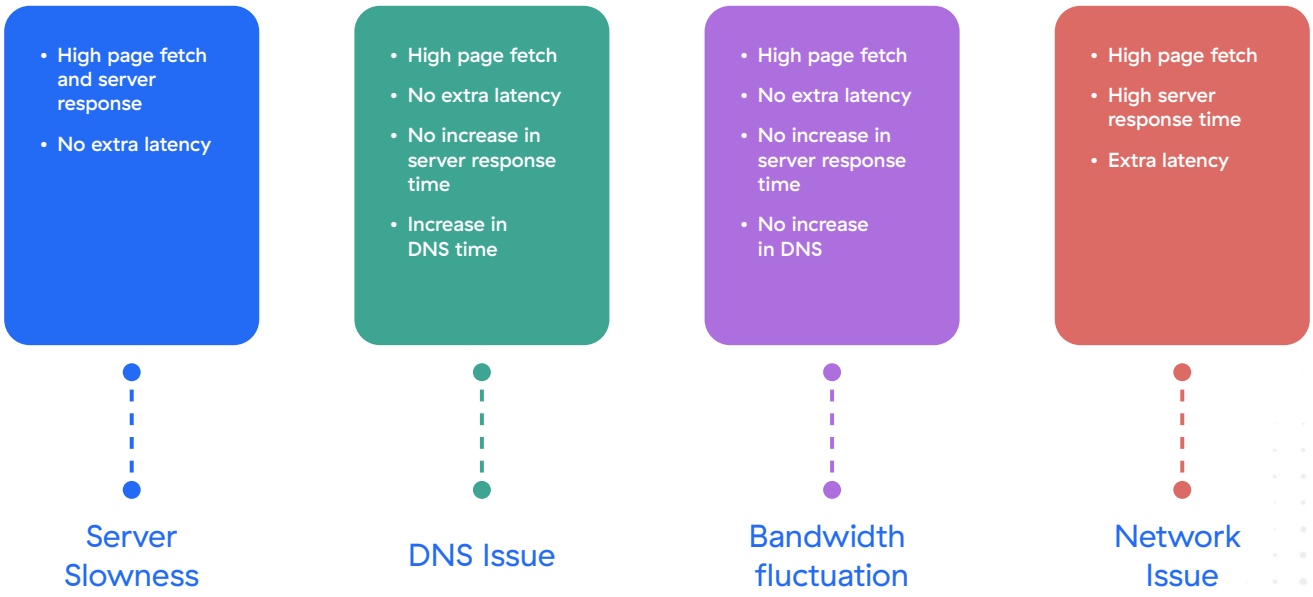
Ballard Mattingly
Principal Cybersecurity Engineer, Liberty Mutual Insurance

7. Data analysis

7.1 Looking at metrics cohesively rather than subjectively

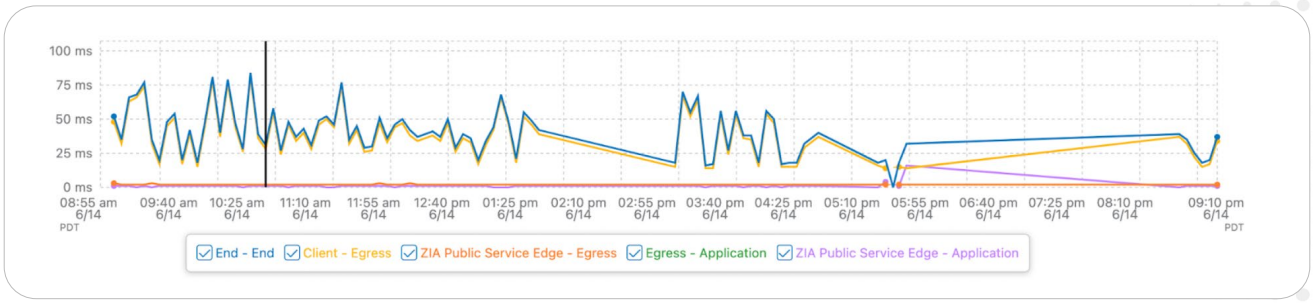
It's important to look at metrics cohesively within the context of other available metrics (same timestamp) such as comparisons of application performance to network and device performance.

Here's a good way of looking at Page Fetch time, Server Response Time, DNS Time and Latency to derive actionable insights from the platform:



7.2 Looking at surrounding data rounds, spike/pattern

When analyzing data, it's important to have historical context on performance to understand where the bottlenecks lie.



On latency peaks, such as those demonstrated in the screenshot above, it's vital to understand if a node is injecting the extra latency each time.

7.3 UCaaS monitoring

When it comes to Call Quality applications, ZDX Score is the most important metric on a user's overall meeting experience.

Looking at network performance, Jitter is an important factor that impacts real time applications in particular.

With ZDX CLI View of Cloud Path, the StdDev column highlights the same, the highlighted hop in below screenshot is adding between 14 to 140 ms of latency driving down the performance.

			Latency (ms)			
Packet Loss	Packets Faile...	Differential	Average	Min	Max	StdDev
-	-	-	-	-	-	-
0%	0/11	3	3	2	7	1.37
0%	0/11	10	48	14	140	41.89
0%	0/11	< 1	13	11	19	1.9
0%	0/11	2	18	12	26	4.12
0%	0/11	< 1	16	11	23	3.52
0%	0/11	< 1	15	11	19	2.71
100%	11/11	-	-	-	-	-
72.73%	8/11	< 1	15	12	18	2.62
0%	0/11	2	17	12	21	2.64
0%	0/11	1	18	13	26	3.38
0%	0/11	< 1	< 1	< 1	< 1	< 1
9.09%	1/11	< 1	< 1	< 1	1	0.49
0%	0/11	1	4	1	25	6.64
0%	0/11	< 1	1	1	3	0.64
100%	11/11	-	-	-	-	-
0%	0/11	< 1	1	1	1	< 1