Deep Dive into Cisco's Use of QUIC, MASQUE and OS Native Capabilities to Deliver Frictionless Zero Trust Access

Vinny Parla
Principal Architect Cisco Secure CTO Office

BRKSEC-3027

Cisco Webex App

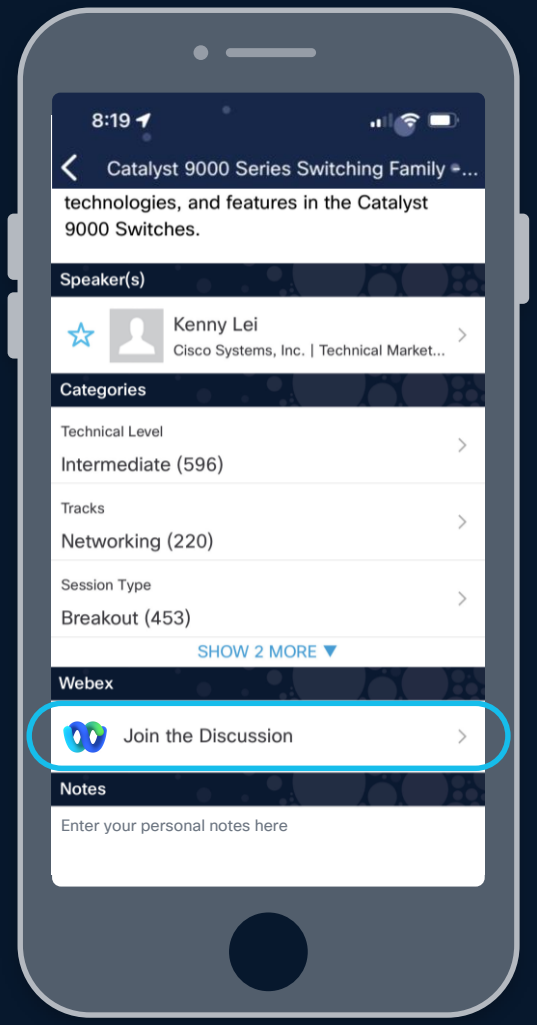
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 13, 2025.



Who is Vinny Parla?

- Father of 2 amazing kids
- Pet hoarder
- Technology Geek

- Principal Architect in the Cisco Secure CTO office
- Developer of many of the remote access technologies you use today
- Creator of the Network Visibility Module (NVM)

- Architect for the new Zero Trust Access ecosystem based on MASQUE & QUIC

Agenda

01 Zero Trust Access Designs

02 MASQUE

03 QUIC

04 Cisco Secure Access (ZTA)

05 New Protocols & Capabilites

06 Conclusion

Hybrid Work. Cloud/SaaS Explosion. Unabated Threats

Need to reduce the attack surface and enforce least privilege access

49%
Employees are remote/hybrid users

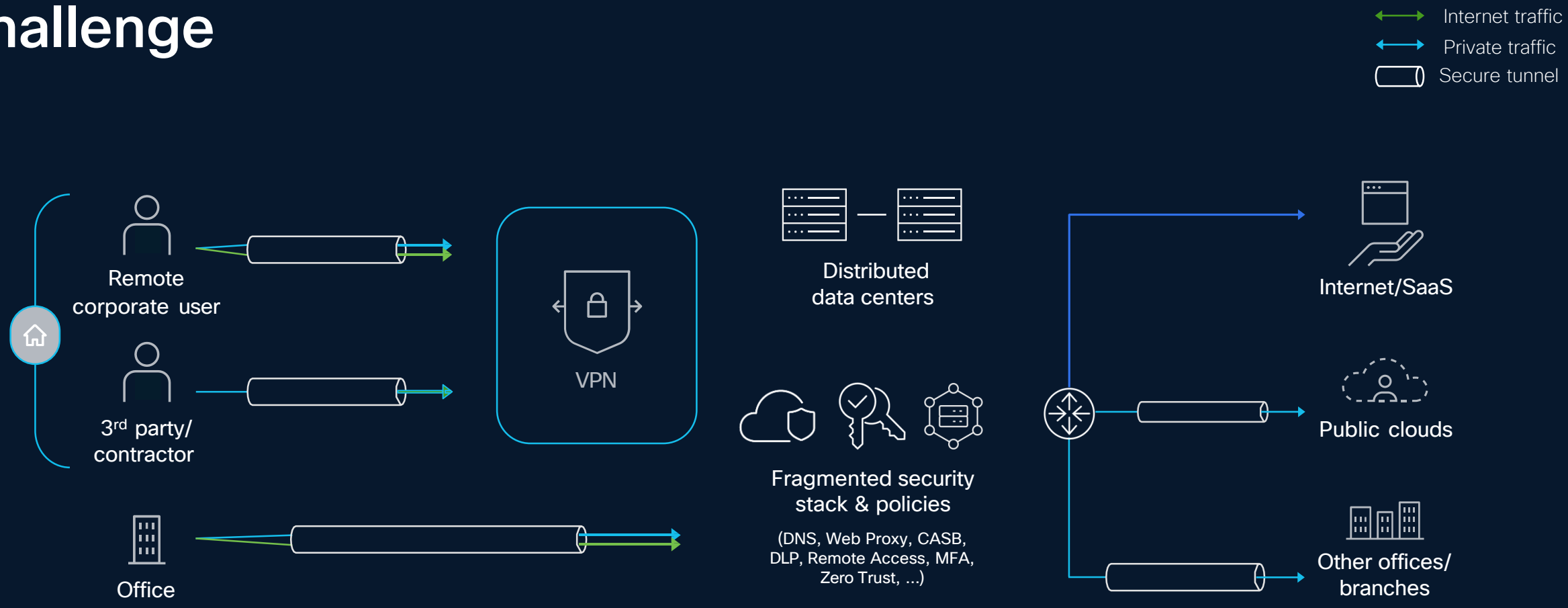
53%
Remote/hybrid workers using DIA

55%
Traffic to/from off-premises, cloud-based facilities

Capacity constraints. User frustration (opting out). Security gaps.

Reference: ESG SSE Survey, June 2023

Challenge

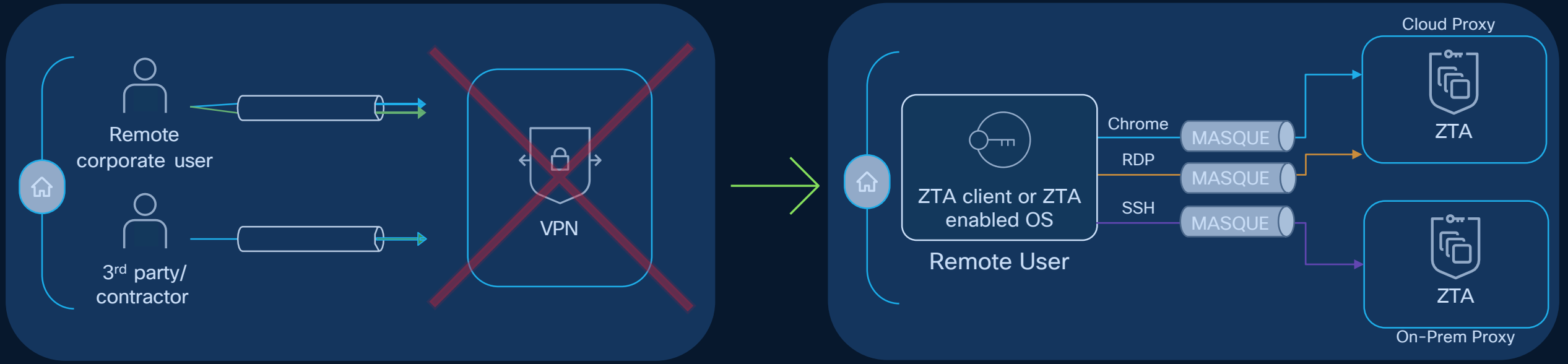


Poor user experience
Lower productivity

Large sets of individual solutions and vendors
Complexity of operations and costs

Gaps in security posture born out of complexity and fragmentation

Evolution of Hybrid Access



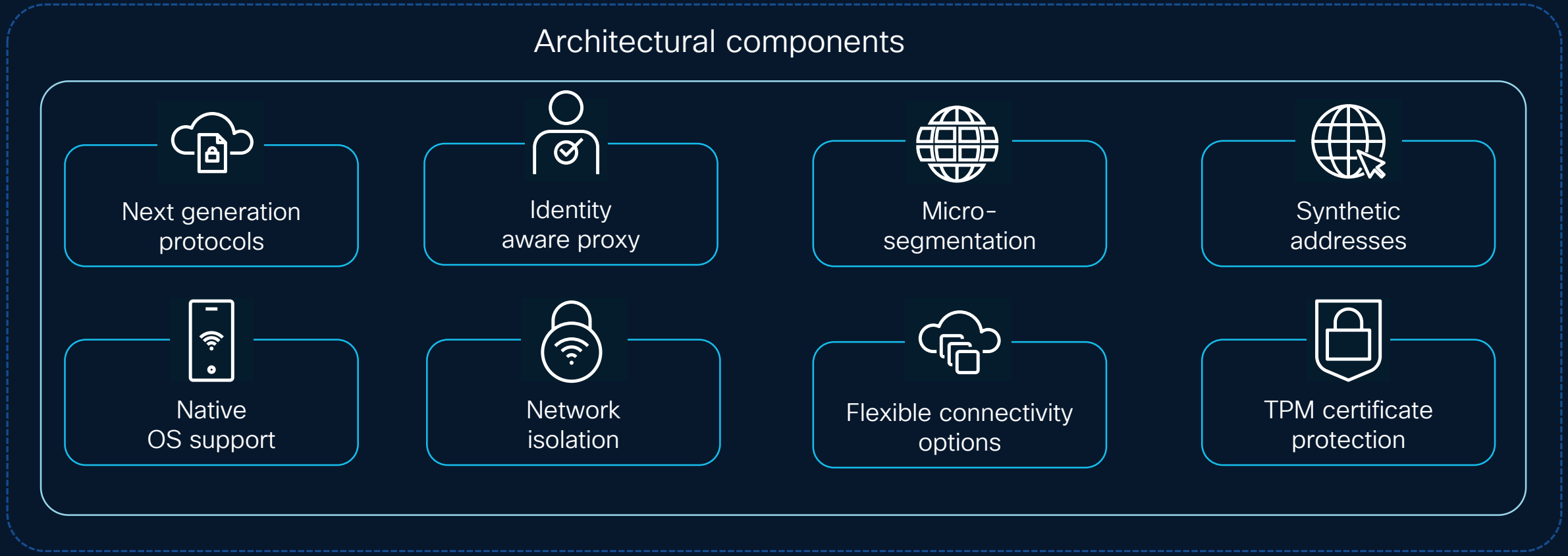
New Protocols Allow for New Approaches

(D)TLS
IPsec

Wireguard
gRPC

MASQUE/
QUIC

Modern Zero Trust Access Design (ZTA)

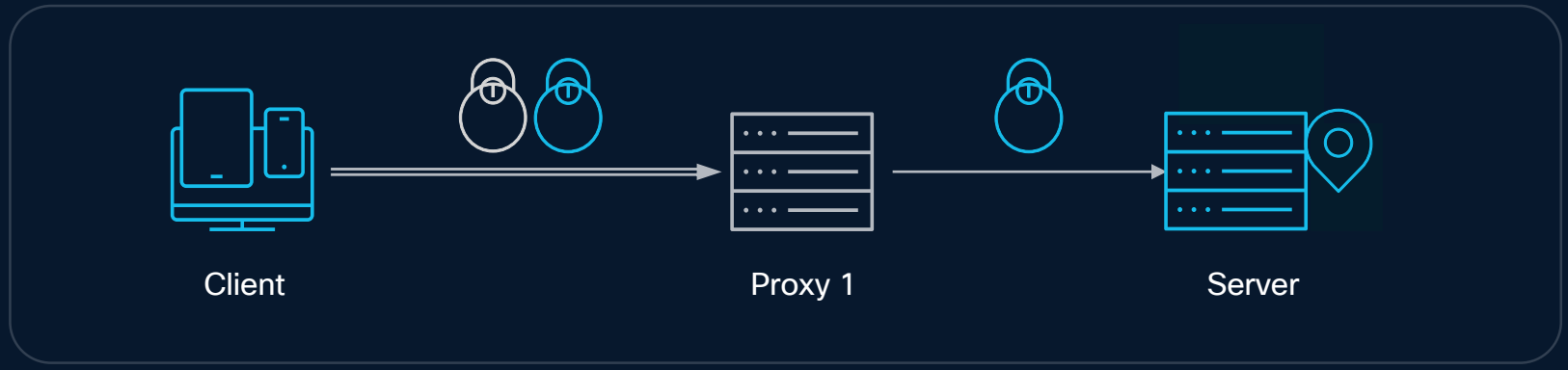


Why MASQUE?

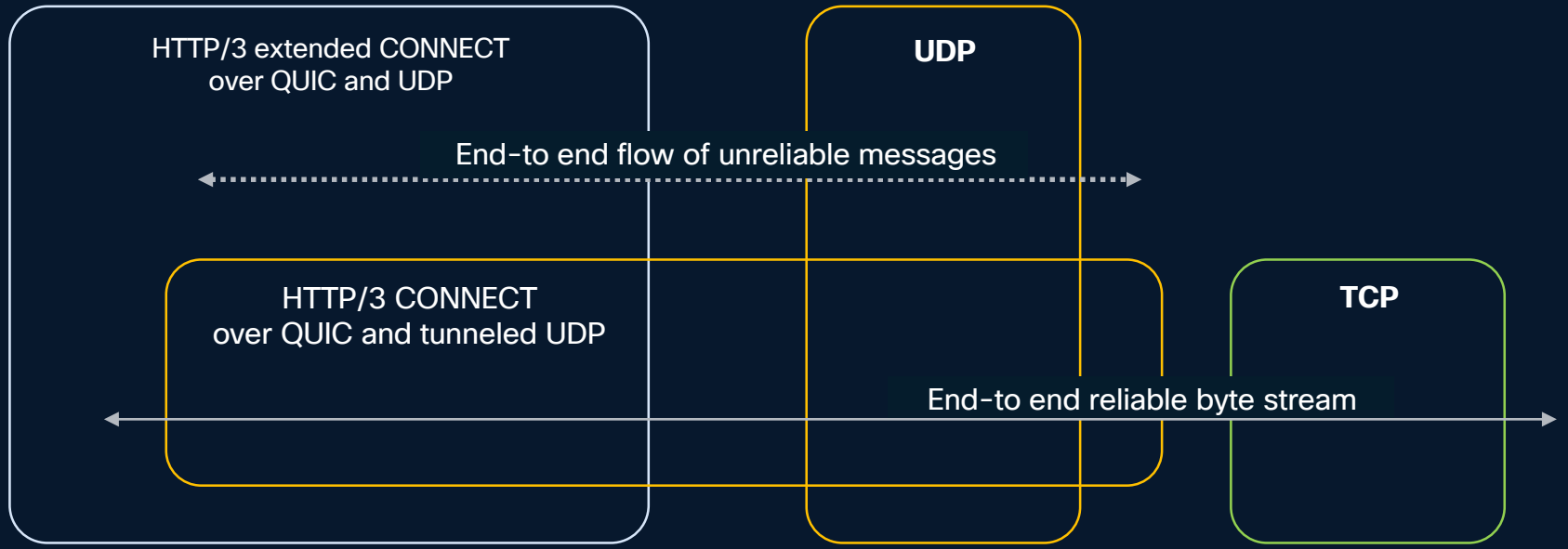
HTTP Formats for Forward Proxying

MASQUE Forward Proxy Protocol

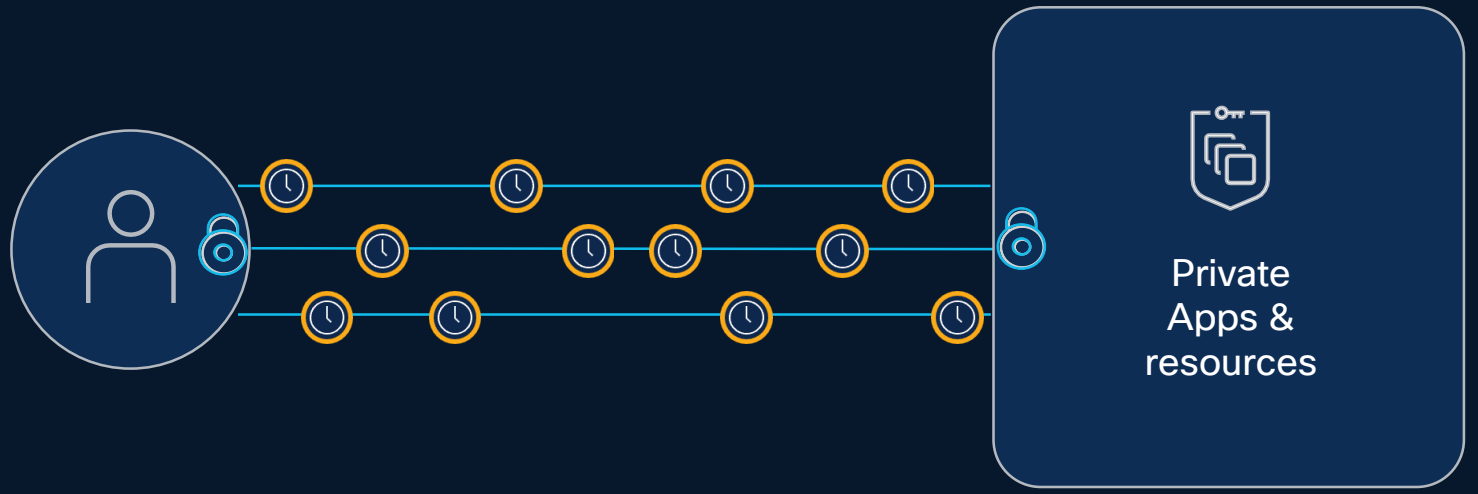
Overview



Protocol view



MASQUE Protocol Micro-segmentation



Value achieved

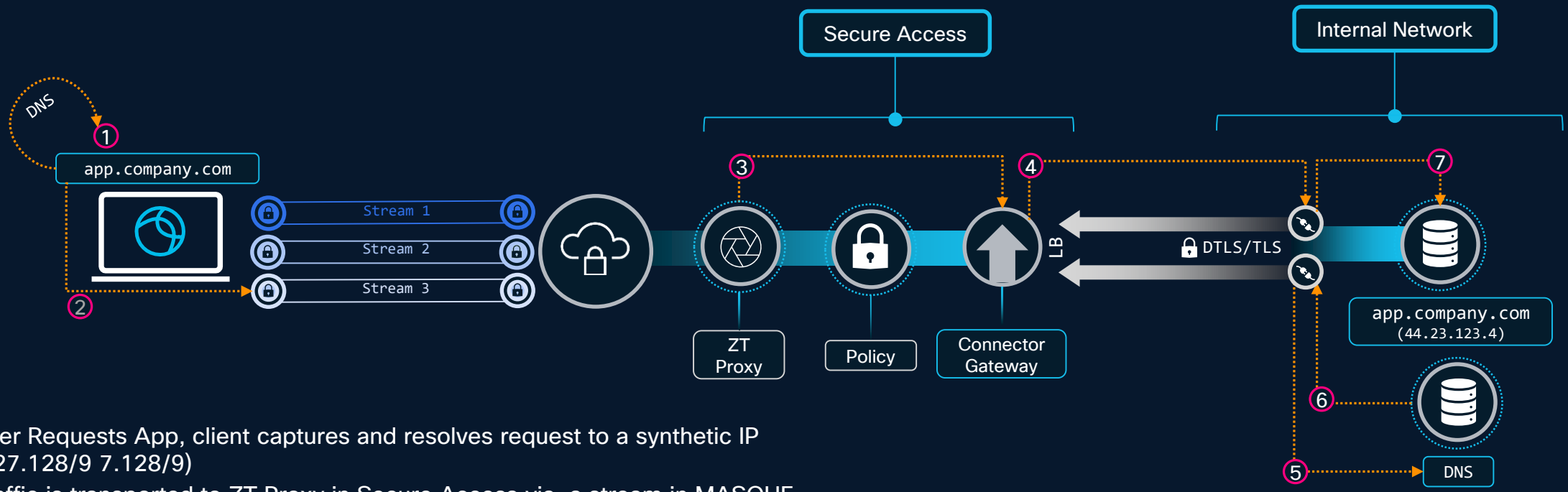
- Unified policy makes creating / managing user based zero trust policies easy
- Trust state is evaluated for every flow
- Least privileged access enforced for every resource
- Users only see the resources policy says they should
- Micro-segmentation directly from inside the OS at a process level

Zero Trust Access benefits of MASQUE

- Multiplexed Application Substrate over QUIC Encryption (HTTP/2 and HTTP/3)
- No direct resource access
- Broad application support
- Proxy any TCP or UDP protocol
- Automatic fallback to TLS if QUIC is blocked
- Process-level multiplexing and segmentation
- Native OS support

Introducing MASQUE – unlocking QUIC Proxying

Seamless and Secure Access for everyone



① User Requests App, client captures and resolves request to a synthetic IP (127.128/9 7.128/9)
② Traffic is transported to ZT Proxy in Secure Access via a stream in MASQUE
③ TCP connection is proxied and built to the mapped resource connector, Policy is enforced on traffic
④ Gateway establishes connection to resource connector
⑤ Resource connector resolves APP IP
⑥ Local DNS responds with App IP
⑦ Resource connector establishes connection to App

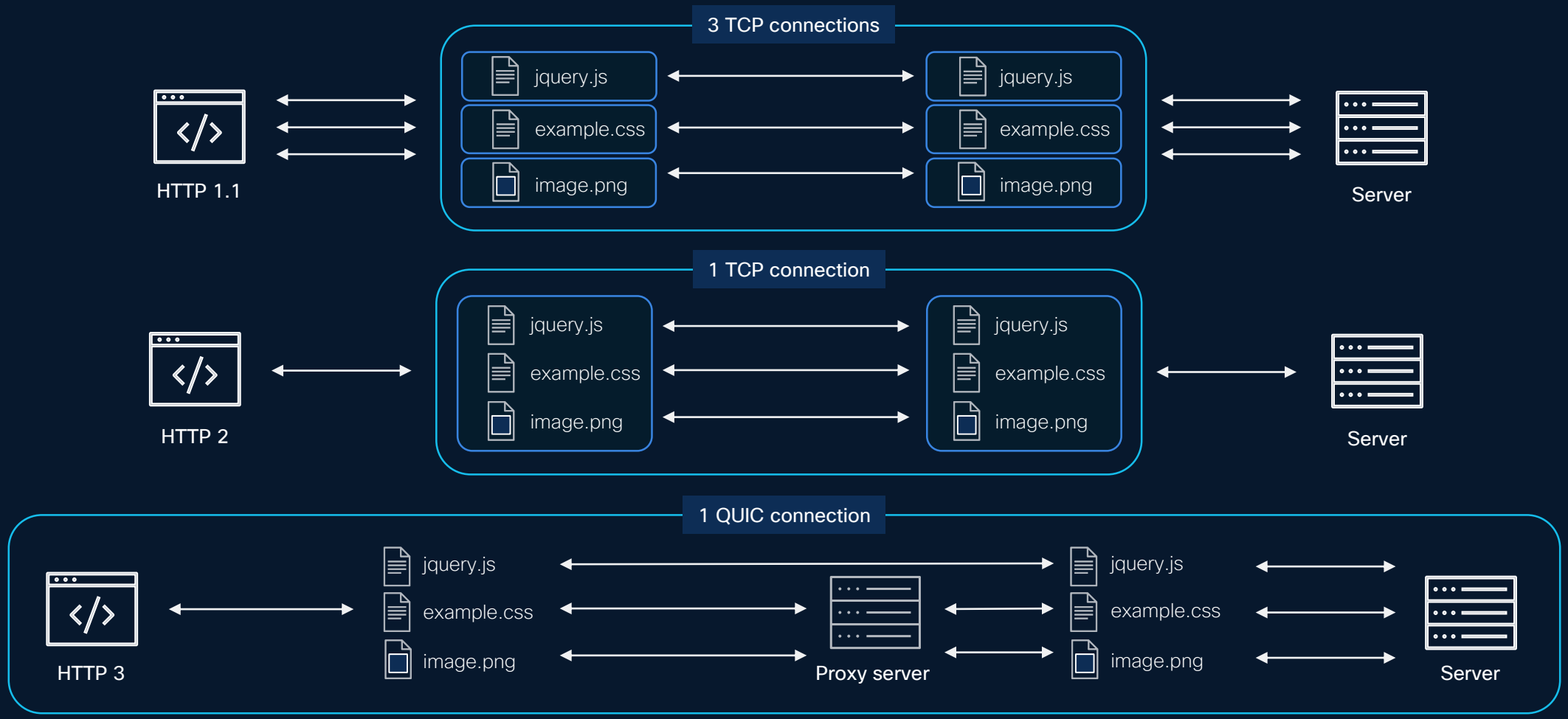
- Seamless In Office Experience
- Full separation between user's network and enterprise network

Why QUIC?

Benefits of the QUIC Protocol

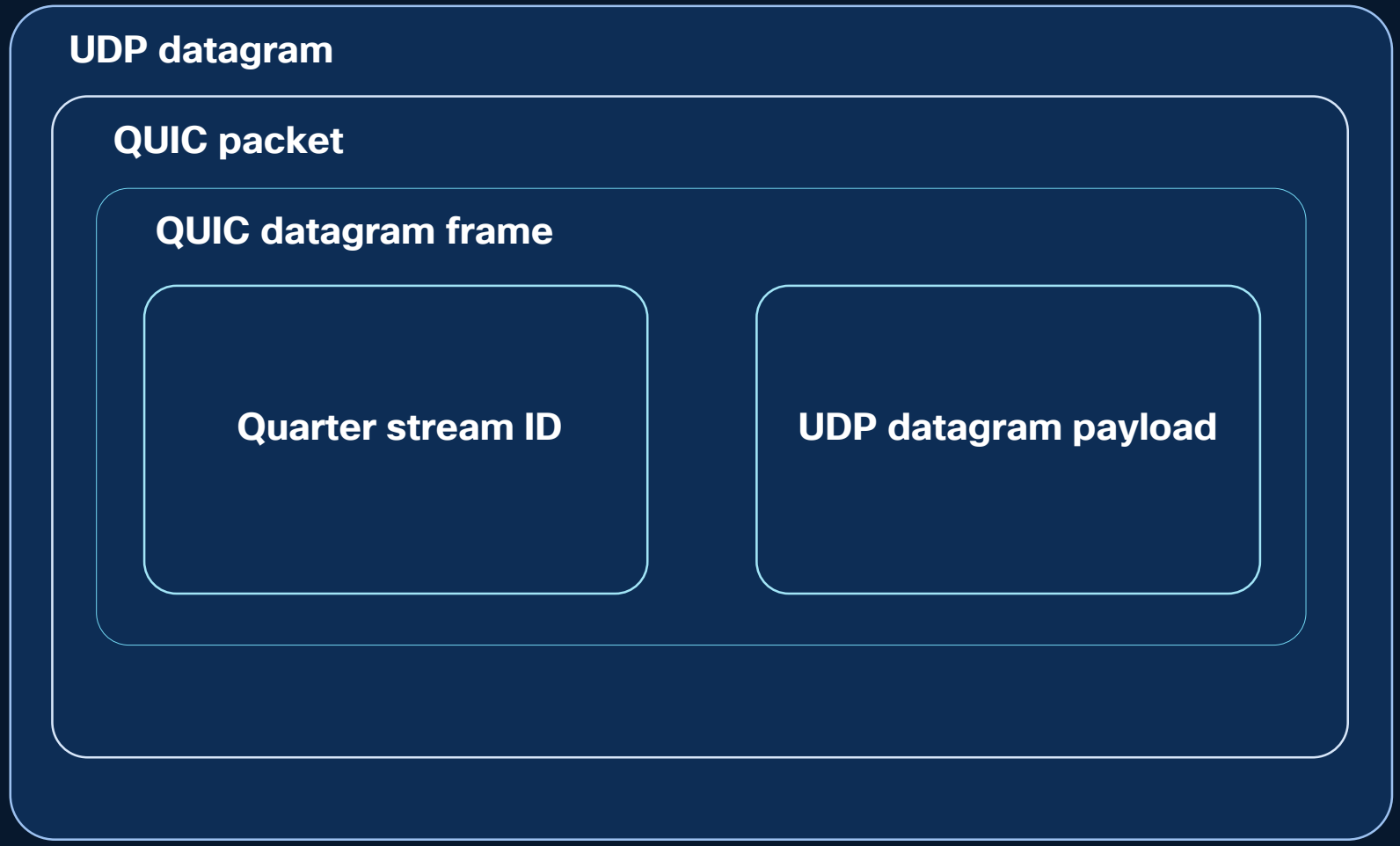
Fast	Better user experience than TCP/TLS for HTTP/2
Secure	Always-encrypted end-to-end security
Evolvable	Not set in stone, new versions deployed quickly
Compatible	Supports all TCP content while avoiding known TCP issues

Connection Evolution from TCP to QUIC



QUIC Packets

- MASQUE can use QUIC for faster performance
- Automatic fallback when QUIC is not possible
- Each connection from an endpoint process is isolated in its own stream and tracked by the Quarter stream ID in QUIC
- Very high segmentation of network traffic



Zero Trust Access benefits of QUIC

- Ability to Change IPs without Renegotiation
- No waiting for Partially Delivered Packets
- Not Vulnerable to TCP-Meltdown
- No Head-of-Line Blocking
- Can Simultaneously use Multiple Interfaces
- Faster Speeds means Faster Access

Vendors embracing QUIC and Next Generation Protocols

Enterprise Relay (MASQUE) Zero Trust Access clients and proxies, QUIC capable firewalls

OS native Enterprise Relay (MASQUE) Zero Trust Access Framework

Private Relay, OS native Enterprise Relay (MASQUE)
Zero Trust Access client, kernel QUIC implementation

SMB over QUIC, Kernel QUIC implementation, edge browser

Chrome browser, Chromium open-source, Envoy & Quiche MASQUE Proxies (50% of traffic is QUIC)

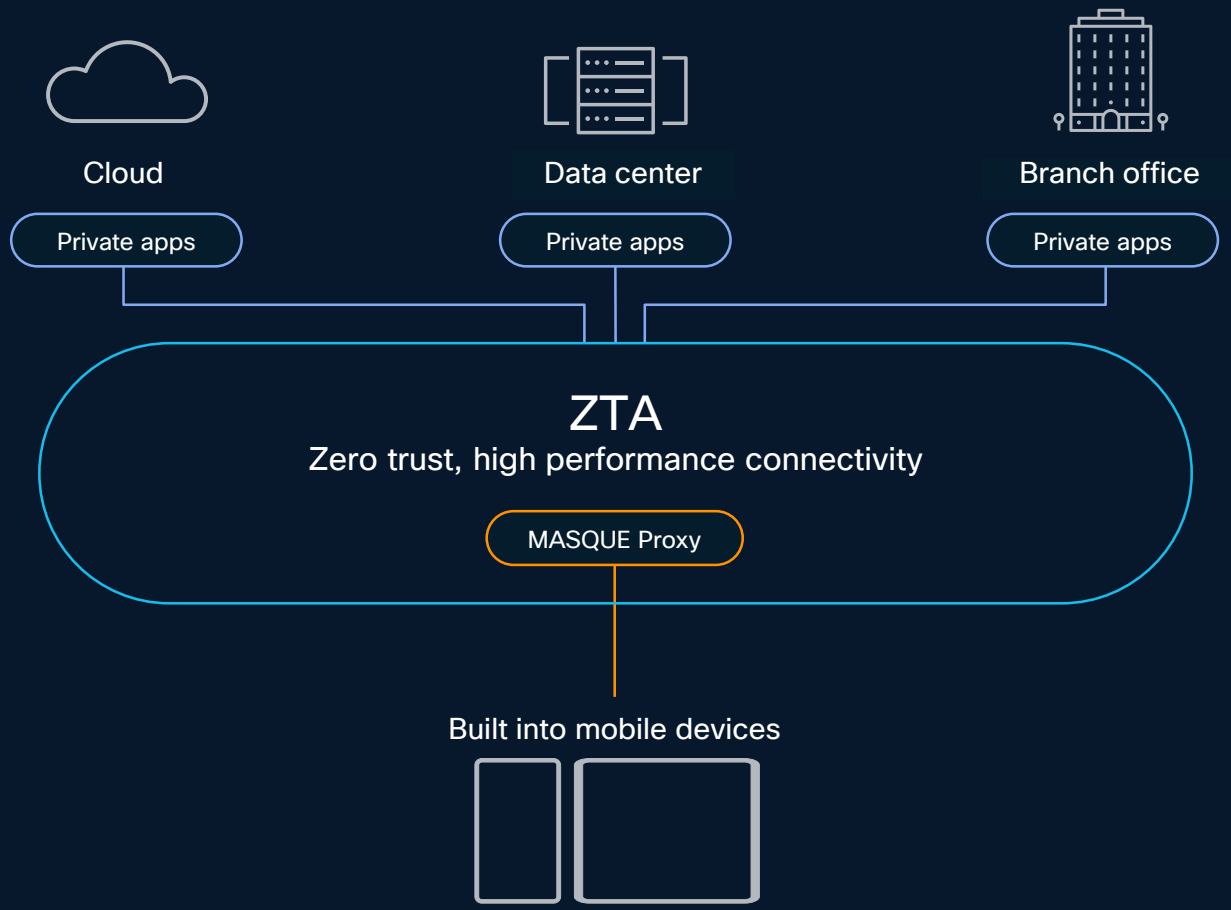
h2o open-source MASQUE proxy used for Private Relay and Enterprise Relay ecosystems worldwide

Quiche open-source MASQUE proxy used for Private Relay worldwide. Zero Trust Access client

Quiche open-source MASQUE proxy used for Private Relay worldwide. Zero Trust Access client

Zero Trust Access Architecture Overview

Operating System Native ZTA



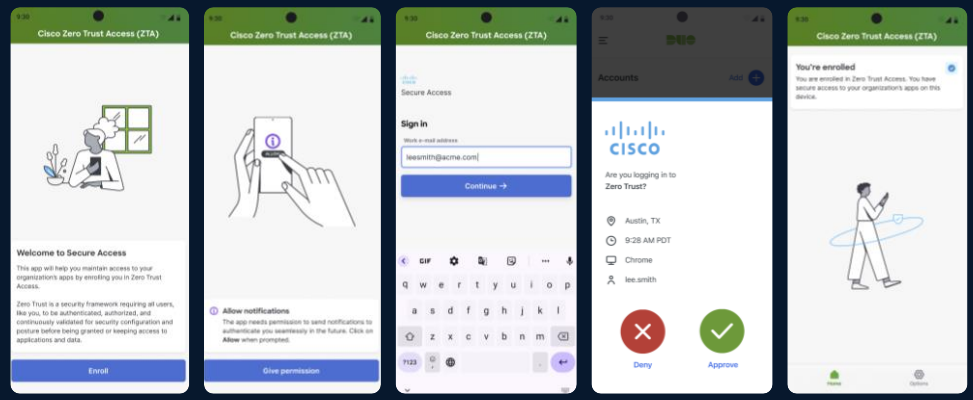
- New OS native ZTA functionality built into the operating system
- Transparent user experience for users—no need to start or wait for VPN
- Delivers low latency and high throughput connectivity by directly intercepting traffic within the application
- Preserves battery life by eliminating the need for device-wide, continuously running VPN connections
- iCloud Private Relay compatible (iOS only)
- Built on industry leading technologies: MASQUE and QUIC
- Supports all applications, ports and protocols—not just web applications

iOS Native Zero Trust

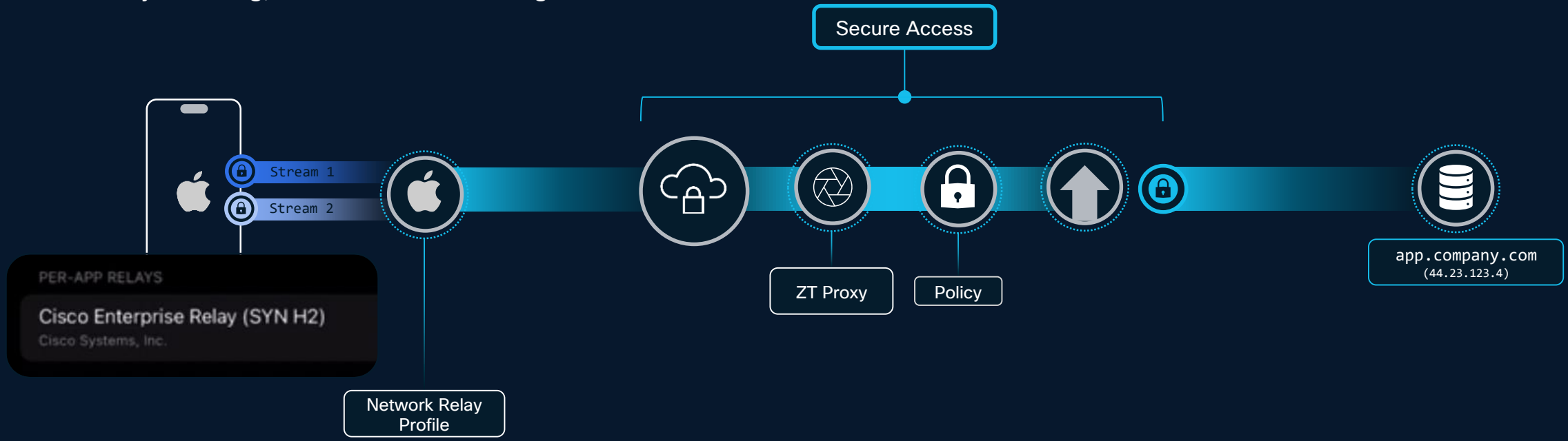
Apple + Cisco Enterprise Relay (iOS 17 +)

Native Experience | Consistent Policy

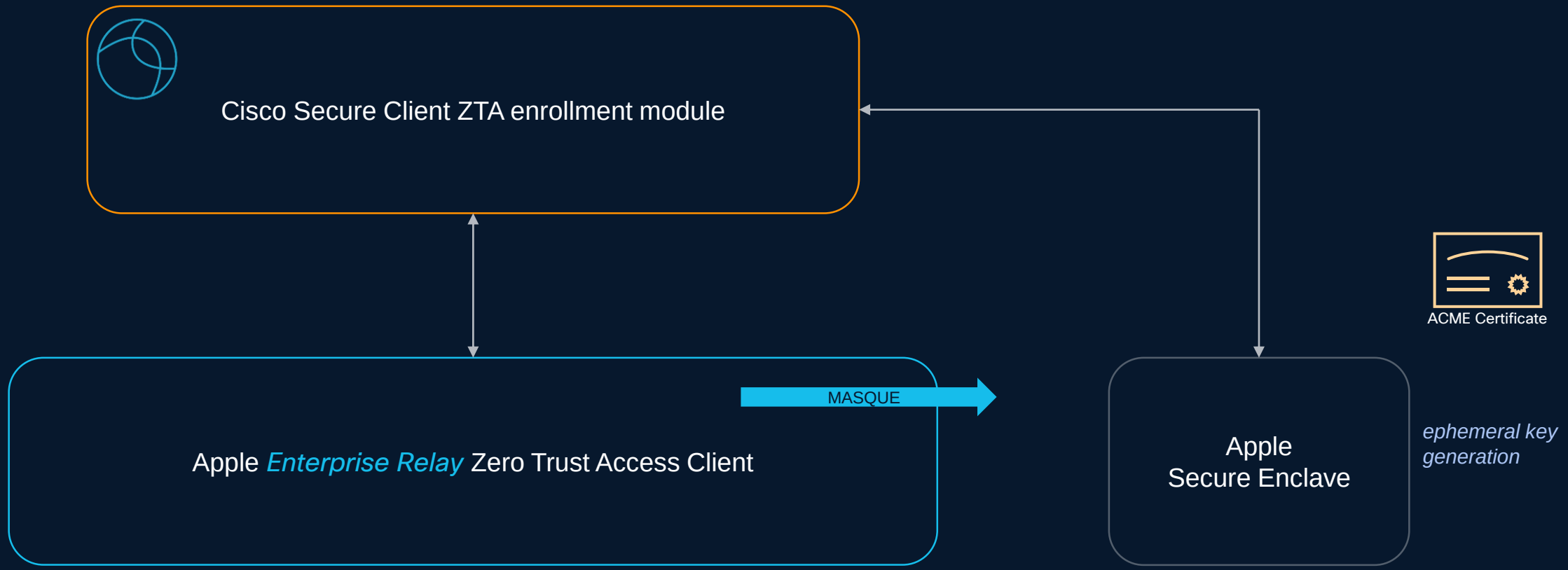
- Co-Developed with Apple
- Native Connectivity without VPN
- Per App visibility & connectivity
- Consistent experience across platforms
- Always running, sustain network changes



Enroll & Get to Work



Apple *Enterprise Relay* Details

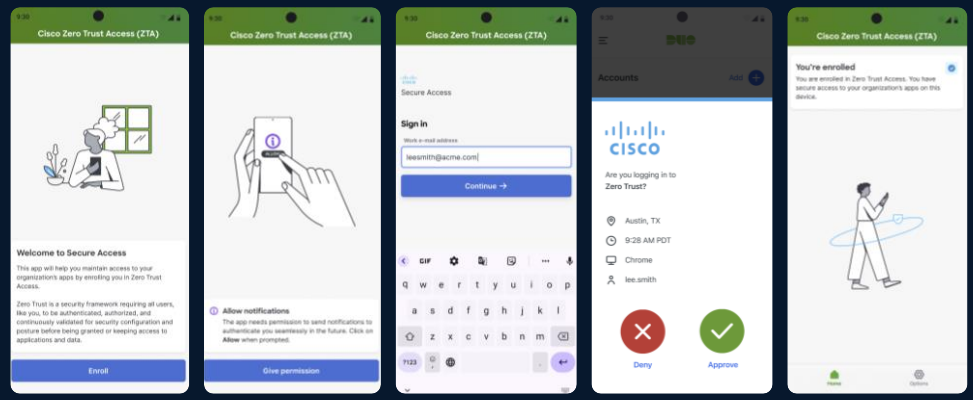


Samsung Android Native Zero Trust

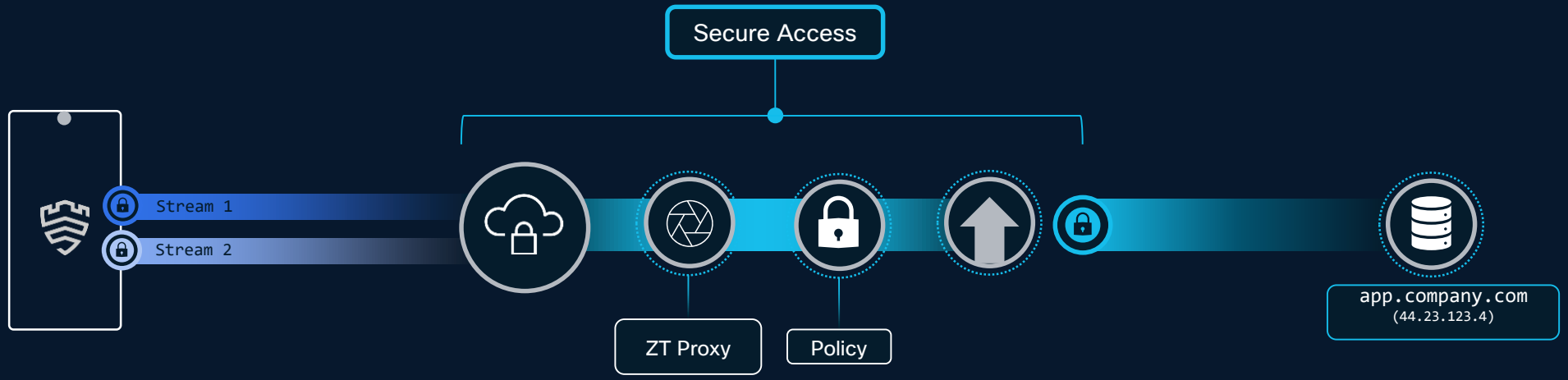
Samsung Knox framework + Cisco (Knox 3.10 +)

Native Experience | Consistent Policy

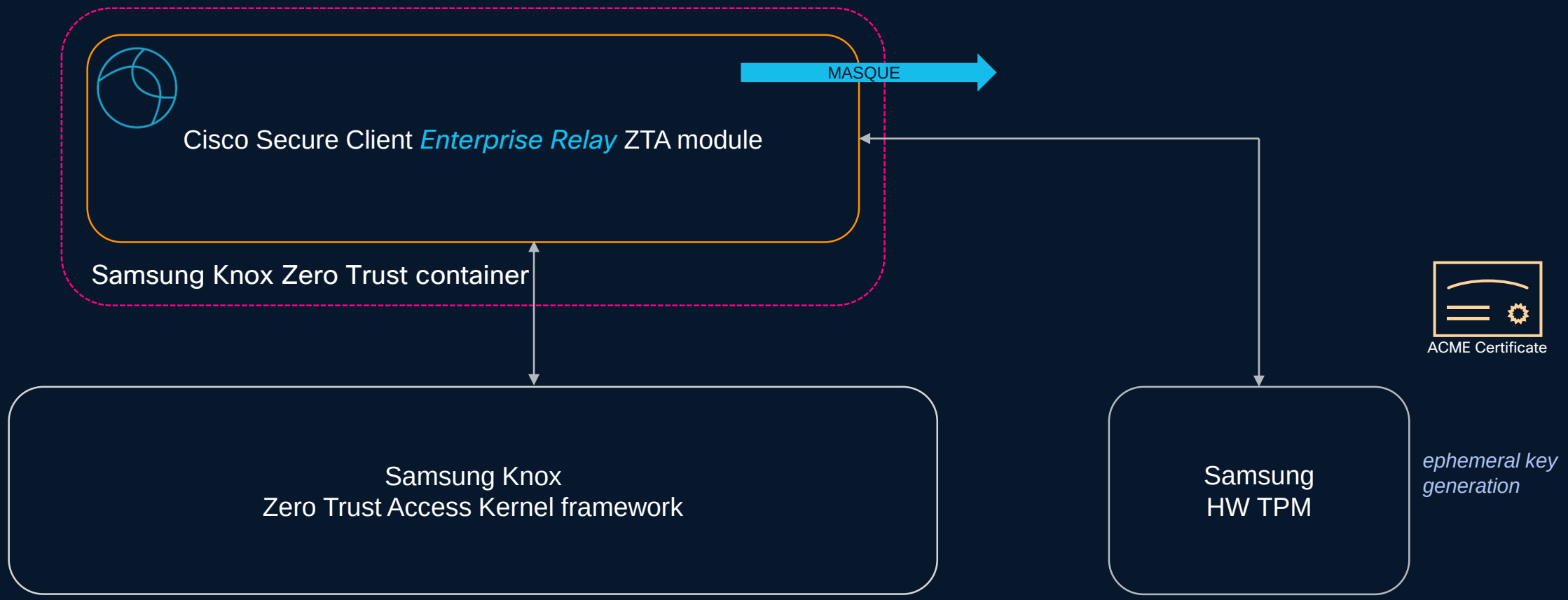
- Co-Developed with Samsung
- Native Connectivity without VPN
- Per App visibility & connectivity
- Consistent experience across platforms
- Always running, sustain network changes



Enroll & Get to Work



Samsung *Enterprise Relay* details

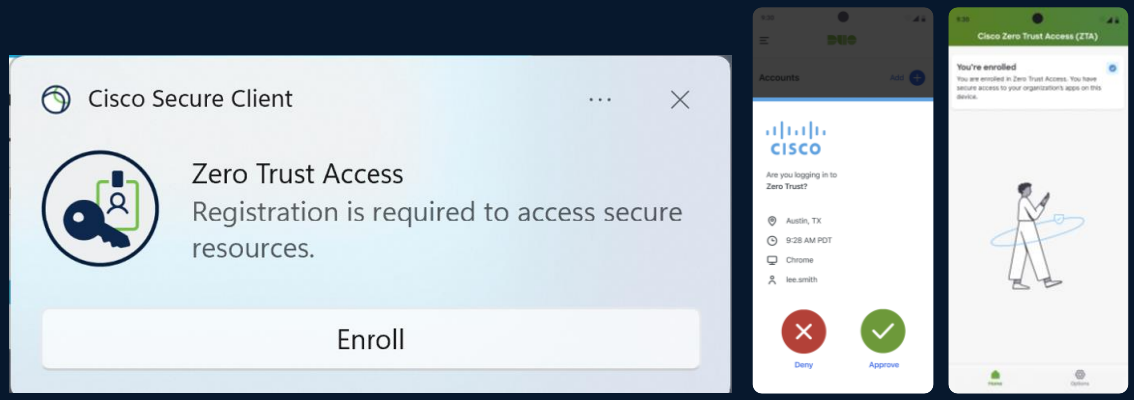


Windows, macOS, Linux Zero Trust

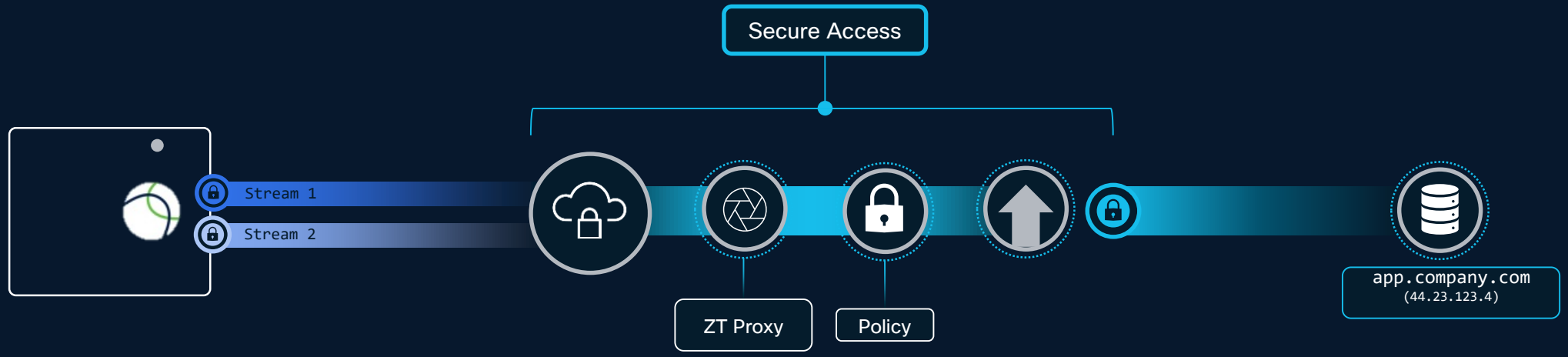
Secure Client 5.1+

Consistent Policy

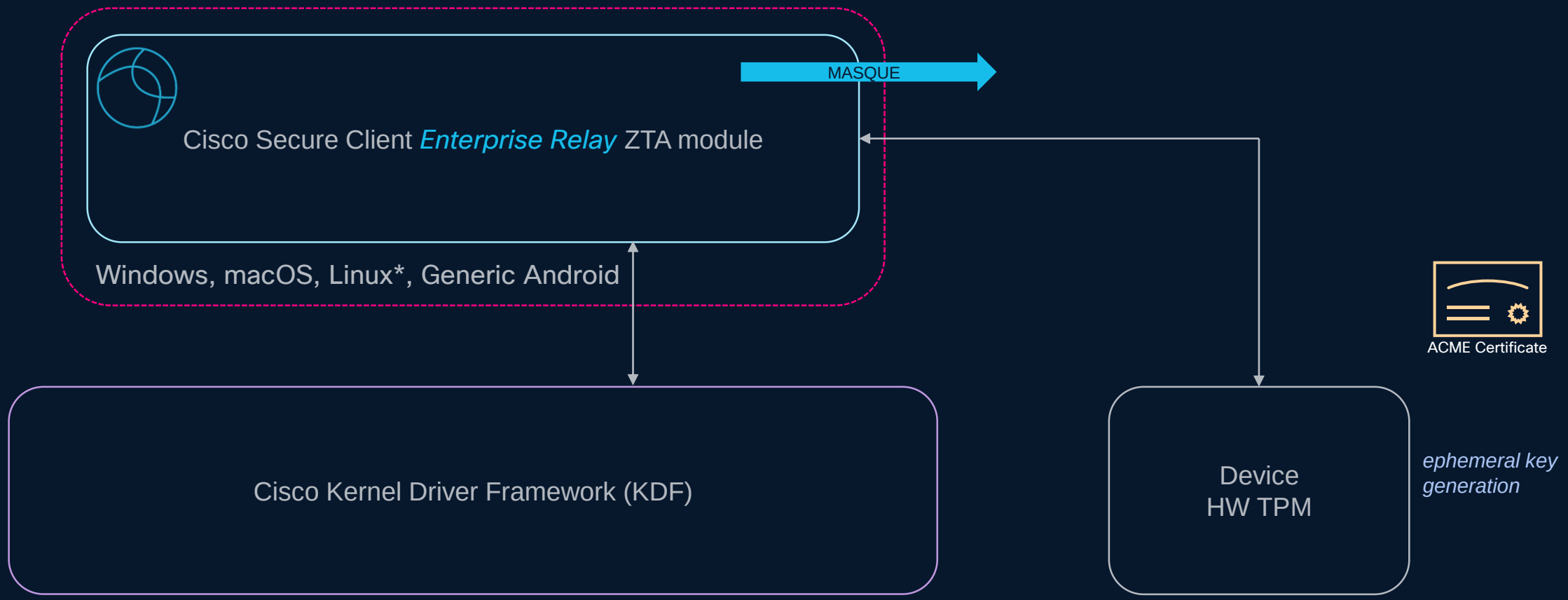
- Developed by Cisco
- Native Connectivity without VPN
- Per App visibility & connectivity
- Consistent experience across platforms
- Always running, sustain network changes



Enroll & Get to Work



Cisco Secure Client *Enterprise Relay* details



TPM & ACME Certificates

What is a TPM?

- TPM is a hardware enclave that protects private keys
- Most TPMs do not store private keys directly, but instead key-wrap them with a TPM bound key
- Binds certificate private key to a specific device preventing key harvesting
- Significantly stronger than simply marking a certificate private key as non-exportable
- On Windows, Andoid and Linux this is known as a TPM 2.0 device
- On macOS and iOS this is called a Secure Enclave
- On Samsung Knox this is called Knox-TPM

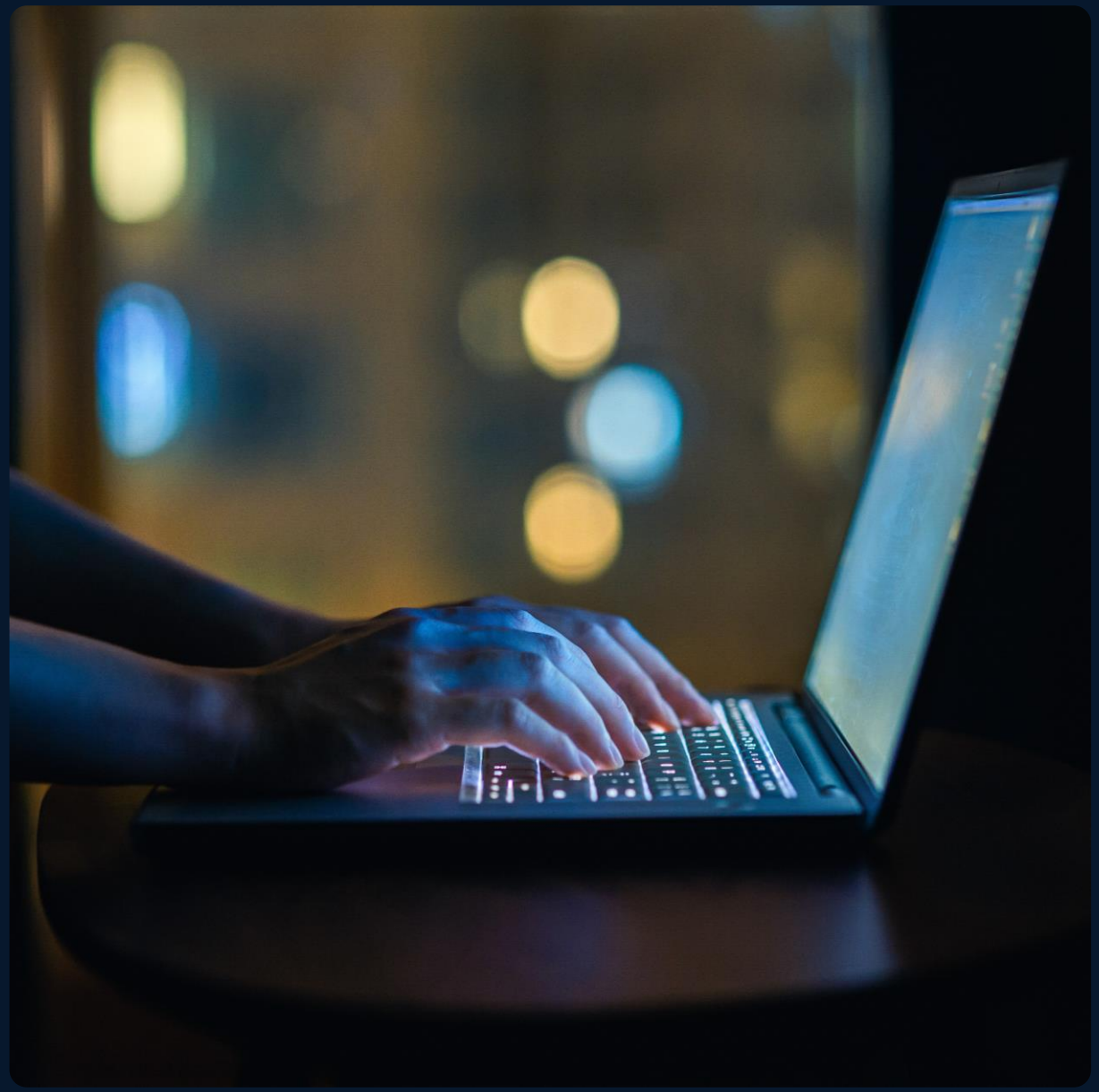
How do ACME Certificates work?

- Users interactively enroll using Biometric Identity via SAML

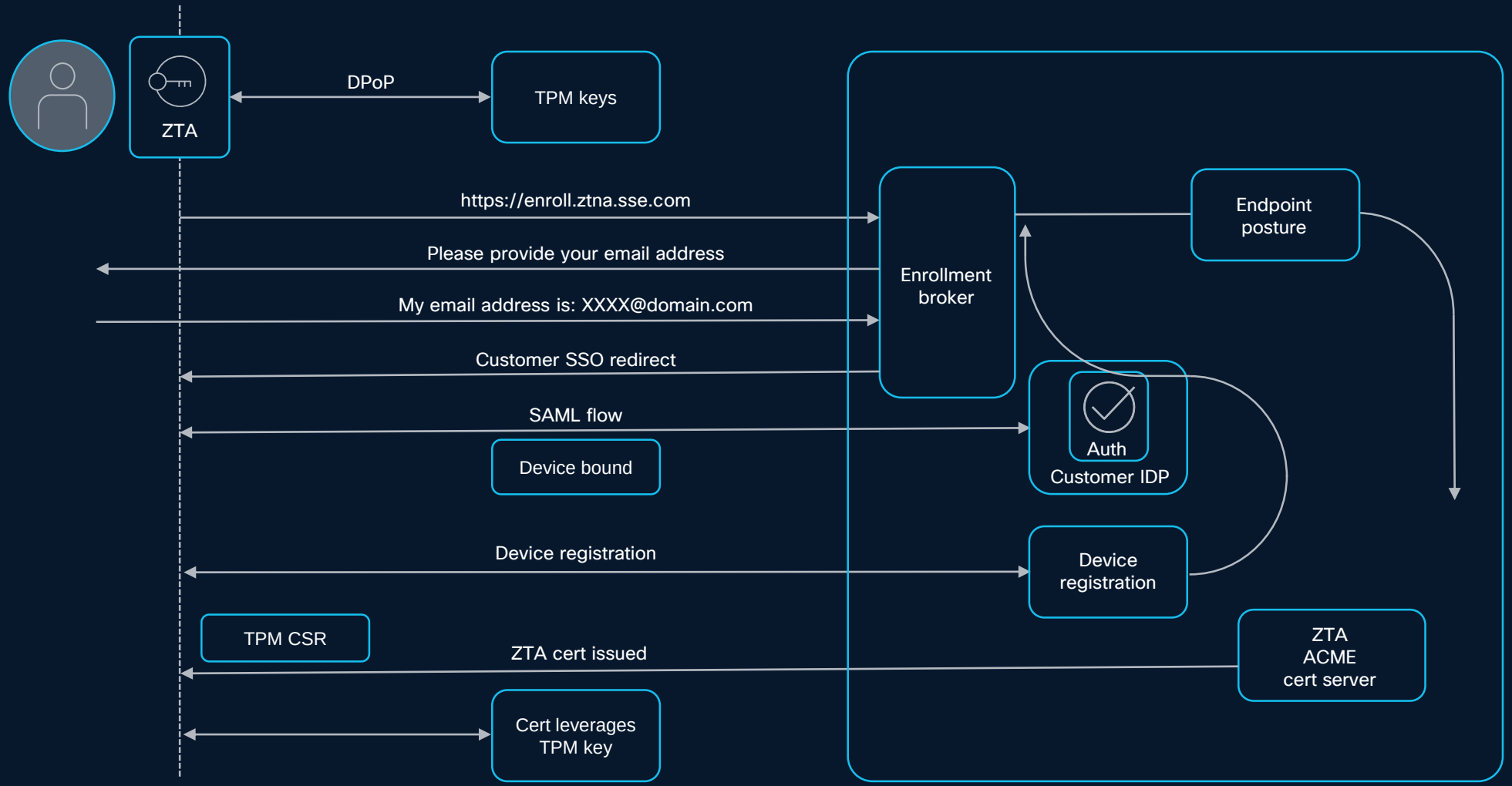
  or

- An Enterprise uses their own PKI Identity Certificate to bootstrap enrollment (Zero Touch Provisioning)

- When interactive enrollment via SAML is used, DPoP binds SAML to the device.
- DPoP is an IETF oAuth standard for Token Binding
- User and Device Identity attributes are bound to the device during enrollment via a Certificate Signing Request.
- The CSR is unique per device using a TPM bound private key.
- ACME Certificates are auto-renewed every 5 weeks using the previously issued certificate for an ACME claim.

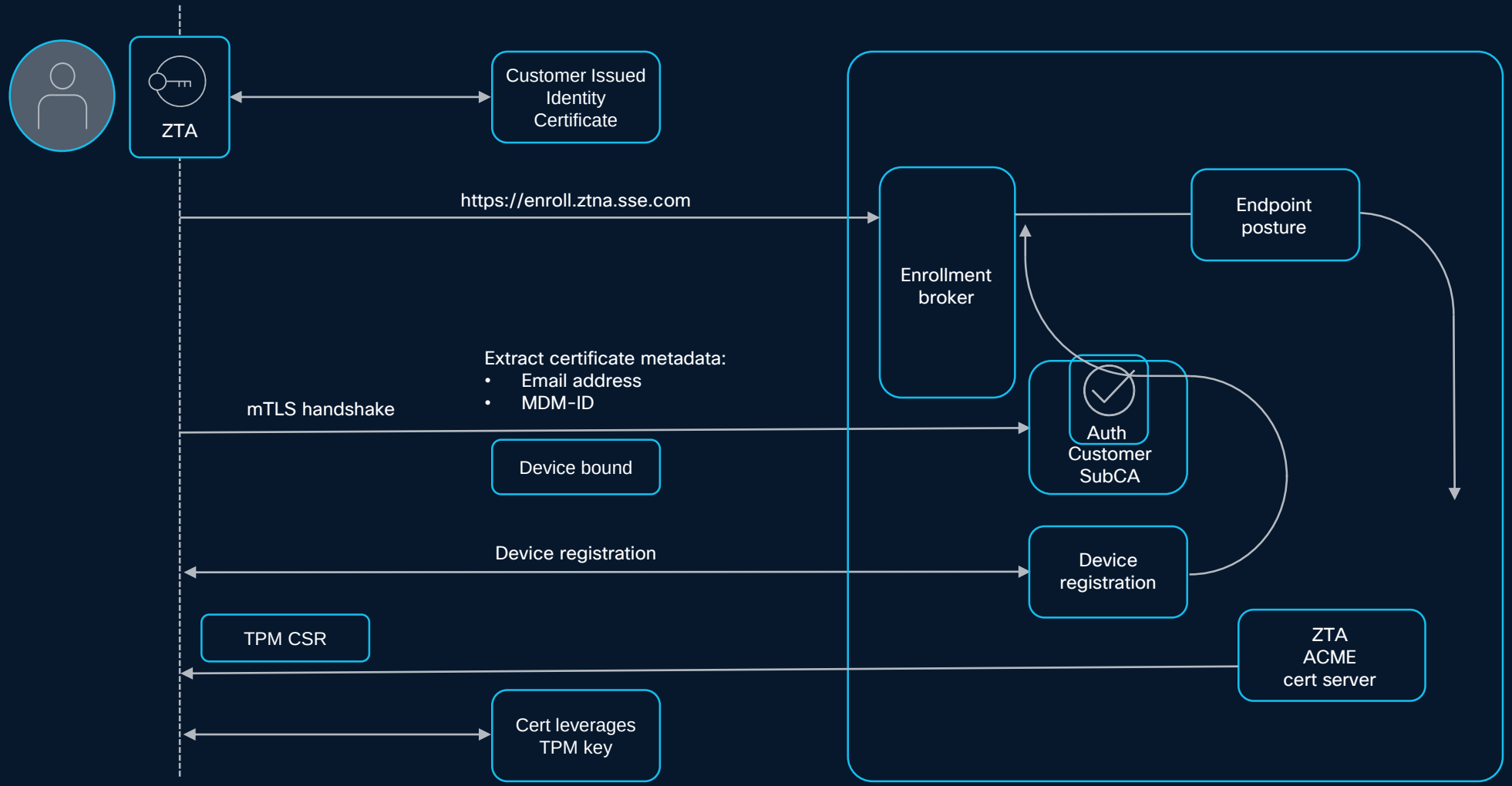


Interactive Enrollment process



- Simple for users
- Advanced level of security

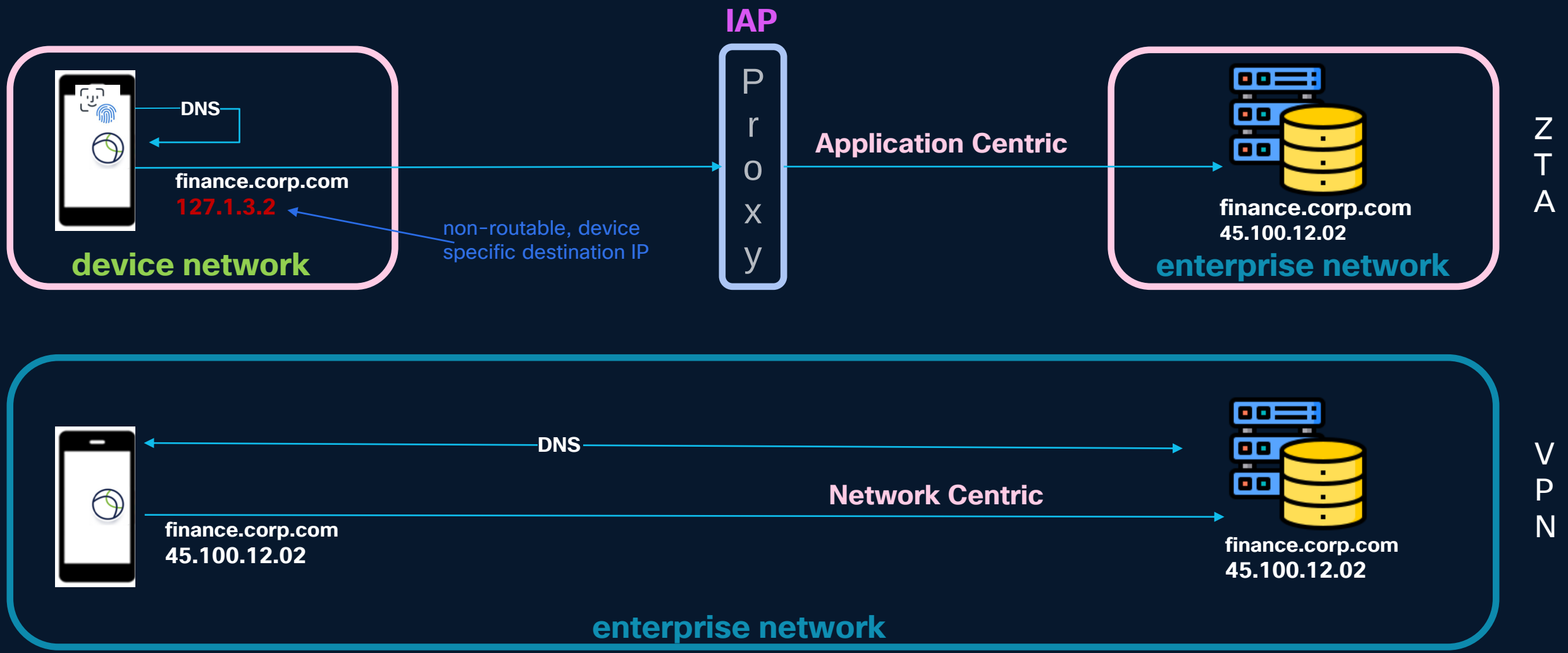
Automated Enrollment process



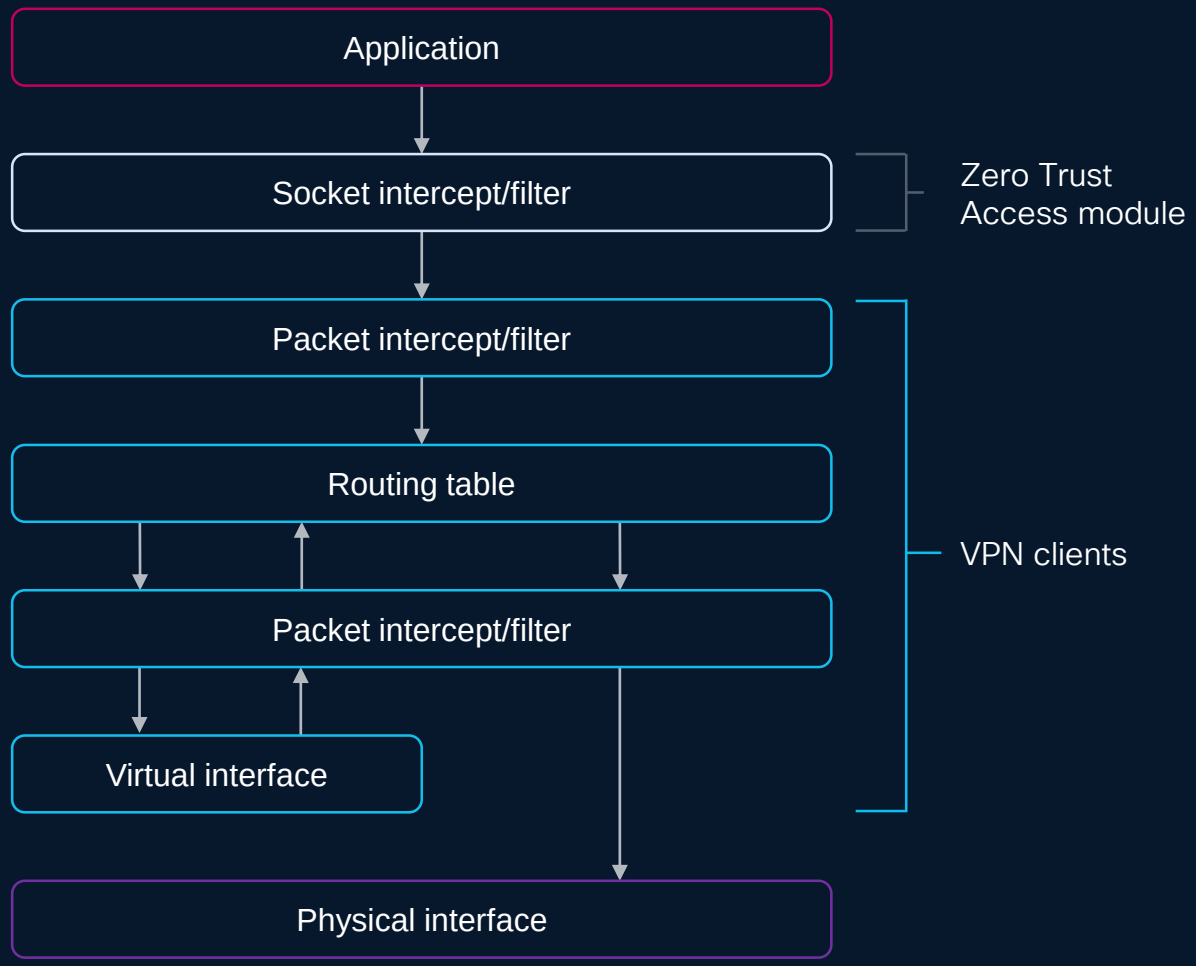
- Simple for users
- Advanced level of security

Packets vs Streams

Zero Trust Access vs Traditional VPN & ZTNA



Zero Trust Access module – socket intercept



Why socket intercept?

- No DNS on the wire
- Control of DNS and application traffic before VPN clients
- No route table manipulation
- Ability to capture traffic by IP, IP subnet, FQDN and FQDN wildcard
- Interoperability with Cisco and 3rd party VPNs

IP Packet vs Socket Streaming Approach

Packet approach

VPNs and legacy ZTNA use packets

CGNAT is obfuscation not security

Firewalls and NATs rely on timing for IP protocols other than TCP, often 30 seconds or more

Attackers can piggyback on UDP flows to continue them in IP packet systems

Streaming approach

modern ZTA uses socket streams

Socket streaming allows any protocol to be tracked by socket call and terminated at the instance the socket is closed

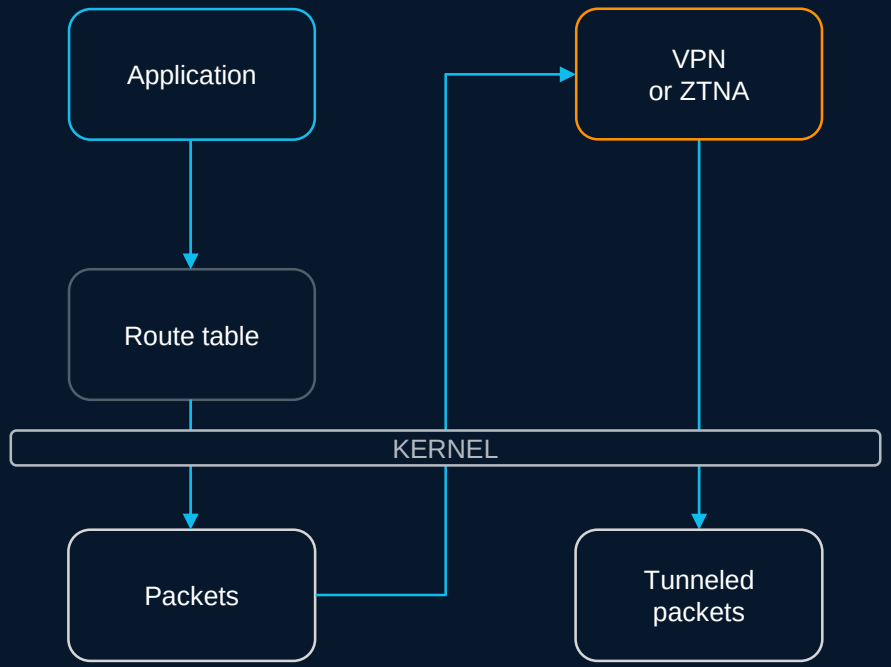
Socket streaming eliminates timers and is deterministic

Flows can’t be continued or hijacked

IP Packet vs Socket Streaming Approach

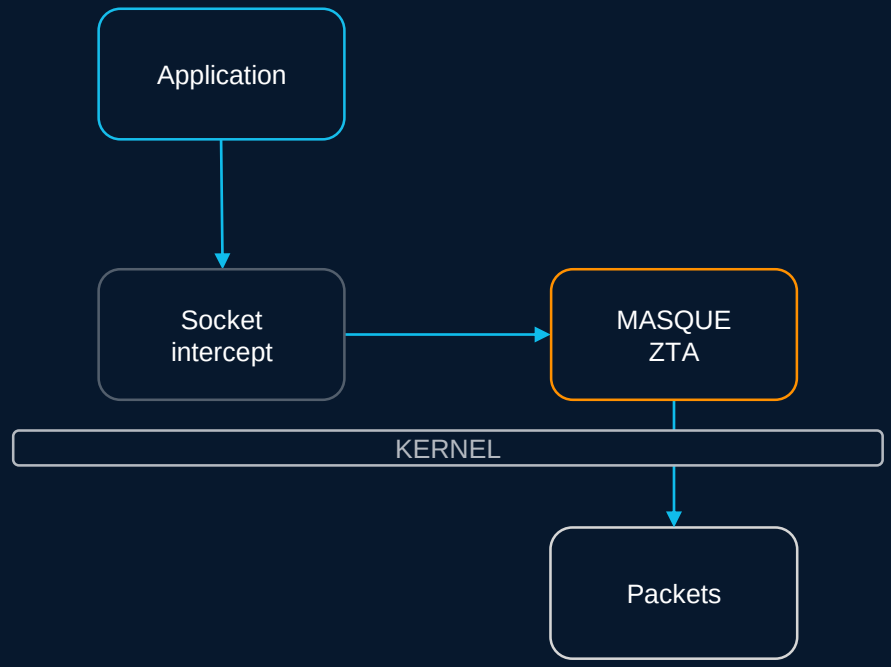
Packet approach

VPNs and legacy ZTNA use packets

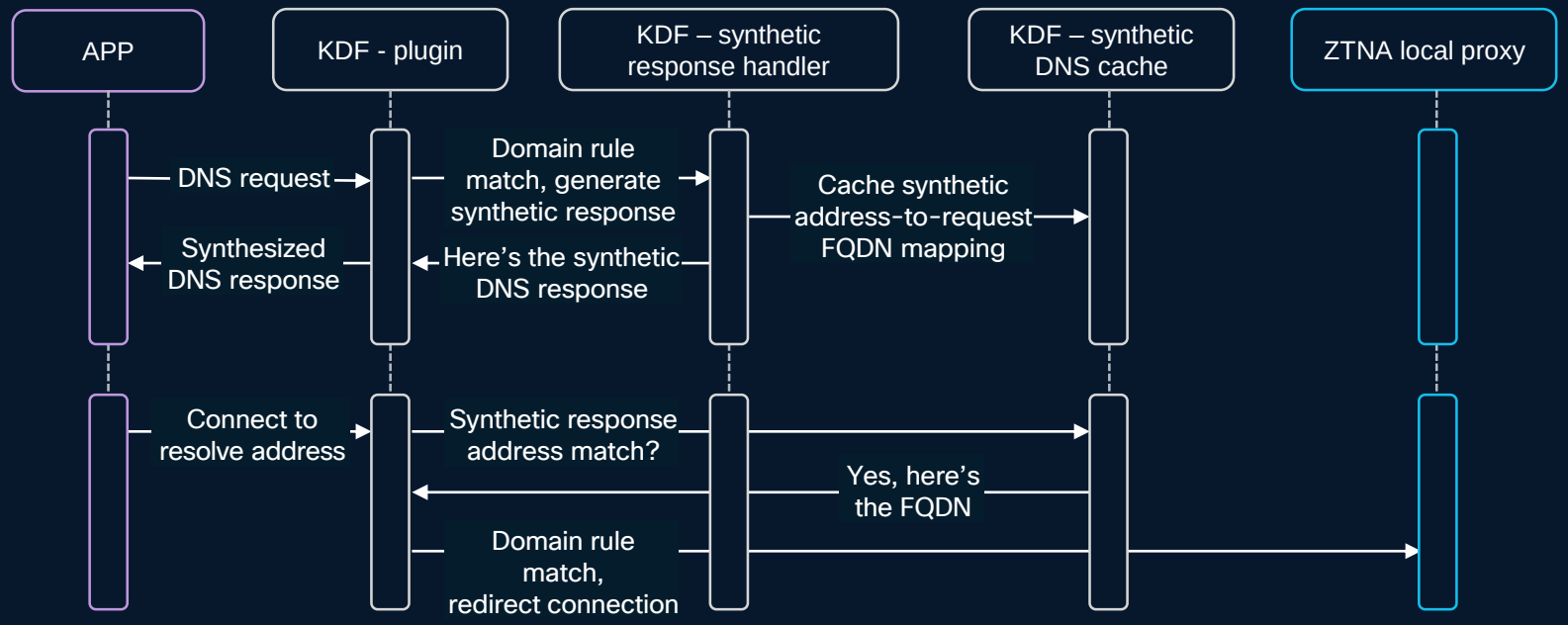


Streaming approach

modern ZTA uses socket streams



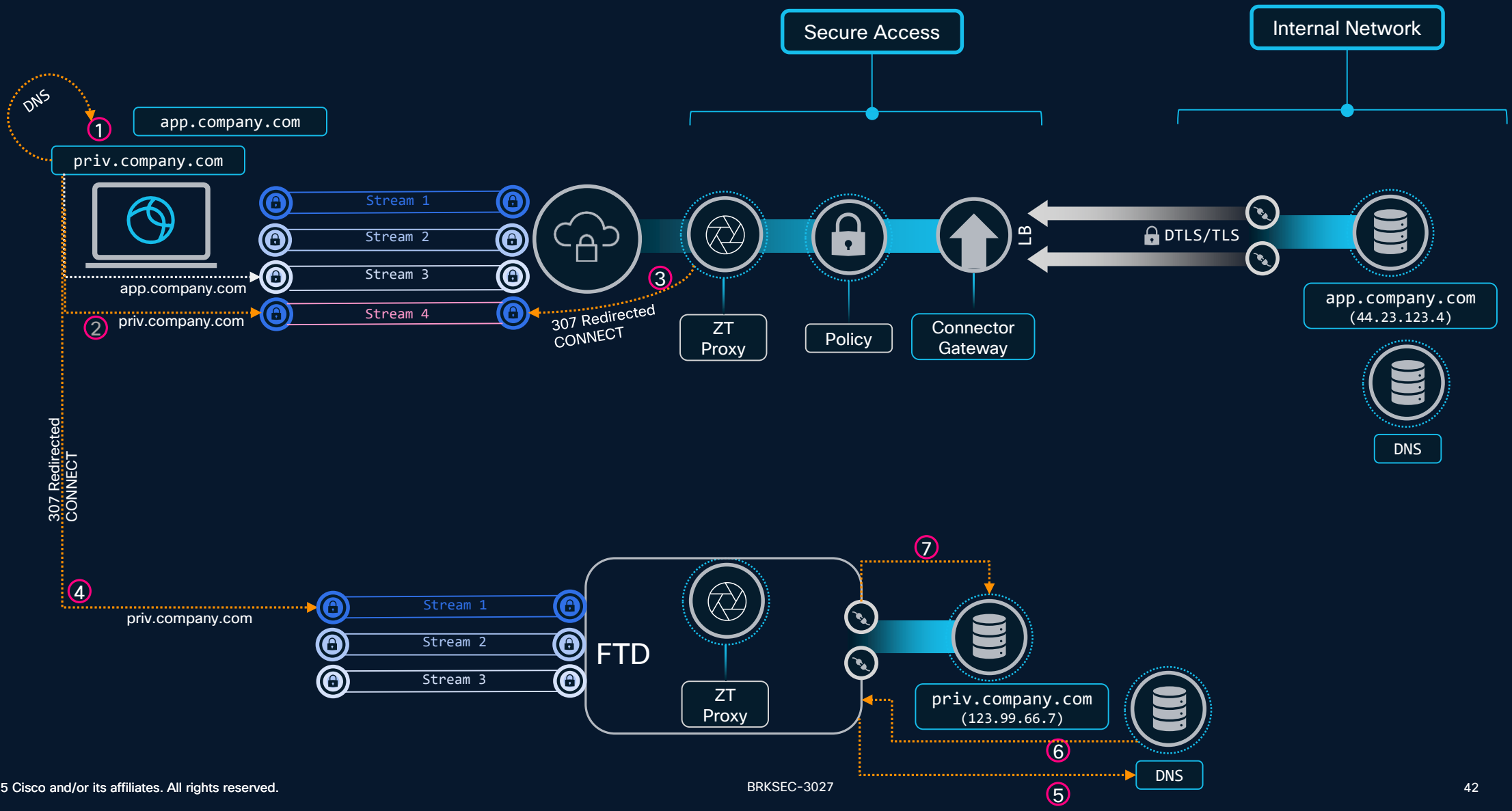
Advantages of Synthetic IPs over CGNAT



- Domain-based flow redirection using a random non-routable loopback synthetic addresses
- No DNS on the wire & no attacker breadcrumbs in the OS
- No flow hijacking or IP piggyback like CGNAT
- Enterprise network IPs never seen at the endpoint like VPN

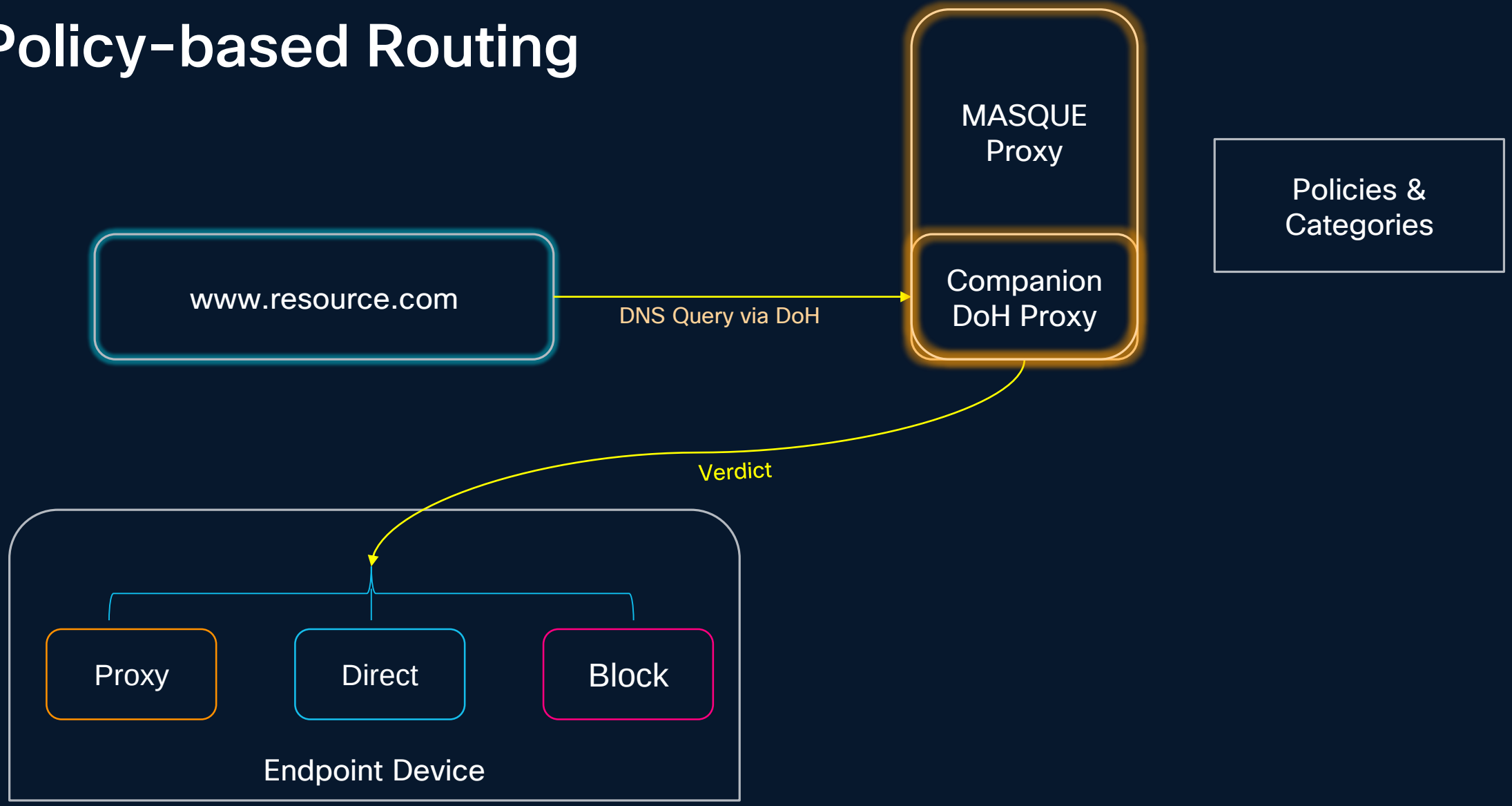
Hybrid ZTA
using MASQUE Redirects

Hybrid ZTA with FTD



Policy Based Routing using DoH via MASQUE

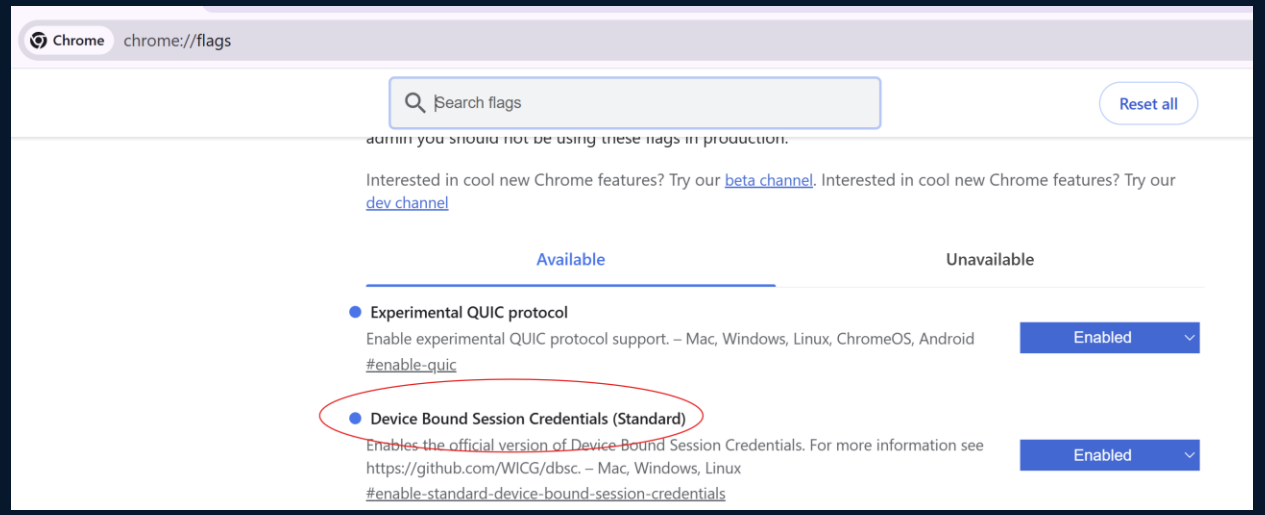
Policy-based Routing



New Protocols In Progress

New Protocols and Capabilities of Interest

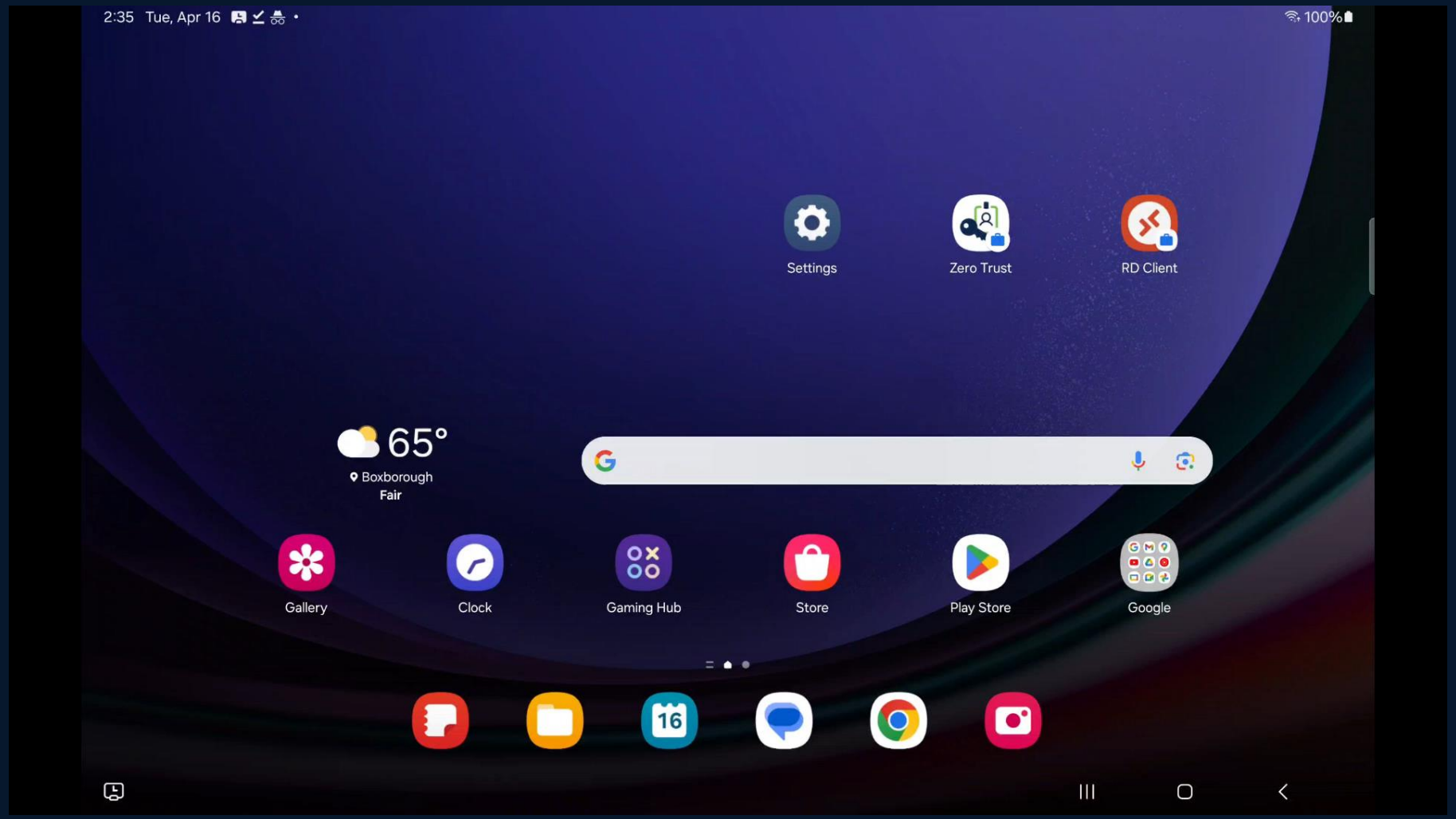
- DBSC & BPoP from Google and Microsoft respectively. Provide for TPM protections of Browser Cookies to prevent harvesting attacks.
  Currently a beta feature in Chrome:



- CONNECT-IP – official IETF standard for sending IP packets via MASQUE. Allows for support of ICMP in the ZTA path
- CONNECT-TCP – IETF standard proposal to mimic CONNECT-UDP syntax for TCP proxying via MASQUE.
- MASQUE PVD – IETF standard proposal to allow configuration to be delivered from a Proxy (e.g. ZTA Proxy node).
- HTTP Redirects for MASQUE – IETF standard proposal to allow 307/8 redirects from a Proxy with the bearer-token passed using the :location redirect header as a parameter. Already being used for Hybrid ZTA.

Demo Time

Samsung ZTA Demo



Apple ZTA Demo

ZTA for the ultimate in secure private access

Least privileged access

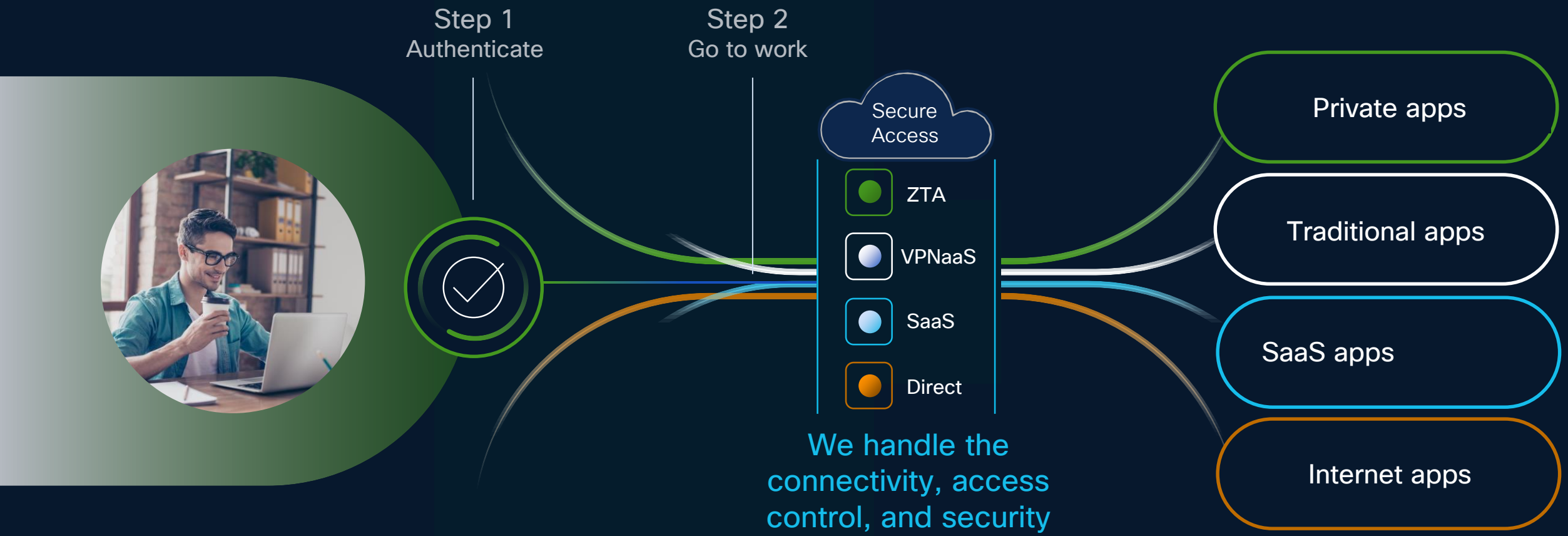
Efficient and secure access to all private apps

- Groundbreaking security
- Incredibly easy to use
- Faster than VPN

- Granular user driven access control protects resources
- Frictionless user experience boosts productivity
- Hardware protected keys to strongly authenticate users
- Hide resources to prevent attacker reconnaissance
- Unified dashboard enables operational simplicity
- Optimized latency/throughput for a superior user experience

ZTA: Next Generation Zero Trust Access

Cisco Secure Access



Enables Hybrid work, securely with no drama, no fuss.

Q & A

Complete your session evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live mobile app.

Continue your education

Visit the Cisco Showcase for related demos

Book your one-on-one Meet the Engineer meeting

Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs

Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Contact me at: *vparla+live25@cisco.com*

Thank you


CISCO