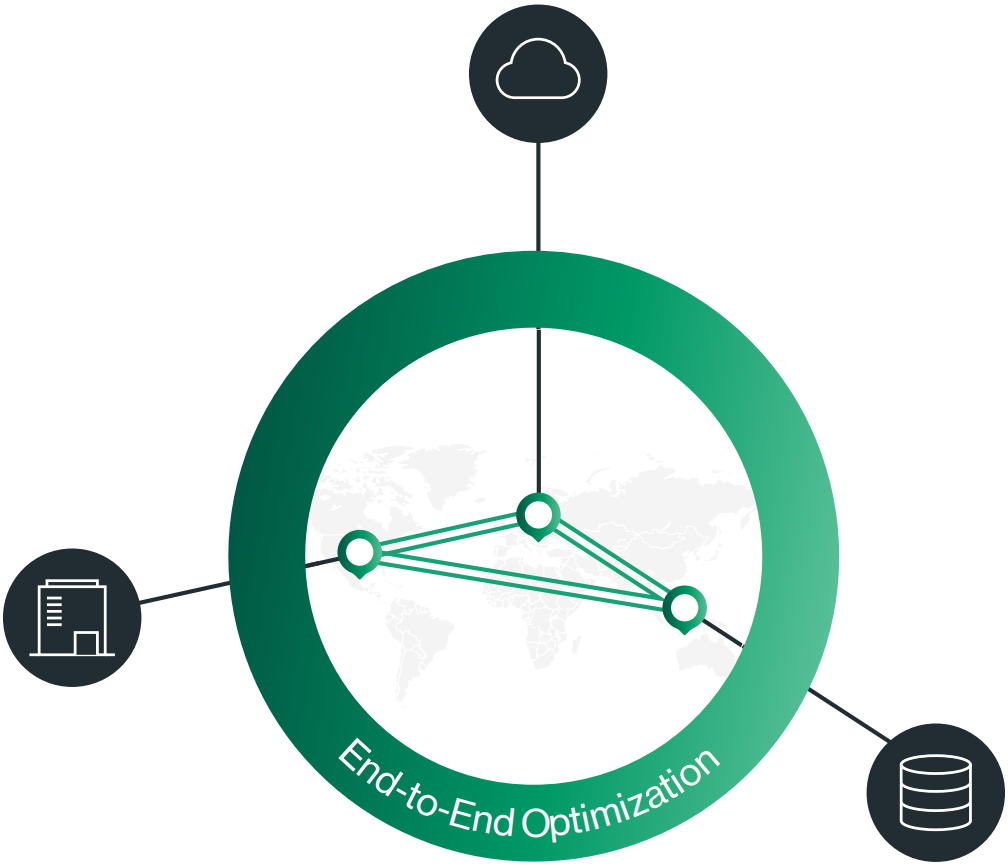
CATO
NETWORKS

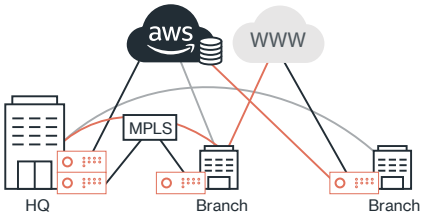
Cato Networks Optimized WAN and Cloud Connectivity



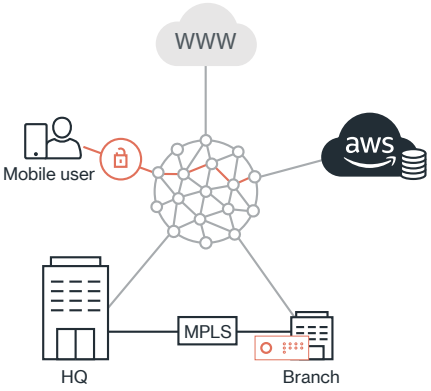
Contents

Overview

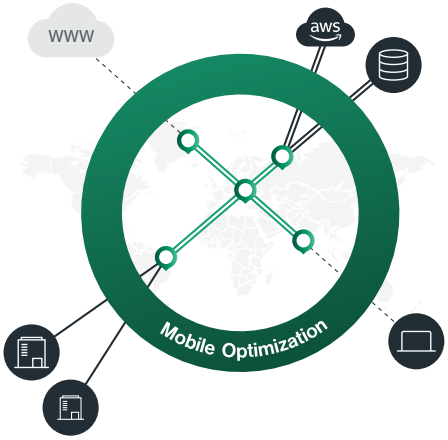
For years, we grappled with the effects of connecting our offices with telco managed MPLS services. Our budgets were taxed with the high-costs of MPLS capacity, forcing us to connect offices with just enough bandwidth. Providing Internet access directly from branch offices would have complicated security management, so we backhauled Internet traffic to secured, centralized Internet portals- and sacrificed the performance of Internet and cloud applications.



SD-WAN fixes some problems but not others. The erraticness of the Internet, particularly in global networks, makes eliminating MPLS impossible with SD-WAN appliances alone. Appliance sprawl continues to be a problem with SD-WAN appliances requiring additional security appliances at branch offices. Connecting cloud resource is, at best, difficult to configure and time-consuming. Mobile users are completely unserved by traditional SD-WAN. Outsourcing SD-WAN to a telco only masks the problem, only now you pay a lot more to manage that same complexity, while losing the visibility, control, and agility that is so essential to a digital business.



Cato Cloud is a global, secure SD-WAN service whose architecture individually optimizes traffic flows at the last-mile and the middle-mile. As such, network optimizations perform better, allowing Cato to achieve dramatic improvements in throughput (see “Multi-segment Optimization” below). In addition, Cato uniquely extends its benefits beyond physical locations to cloud infrastructure, cloud applications and mobile users.

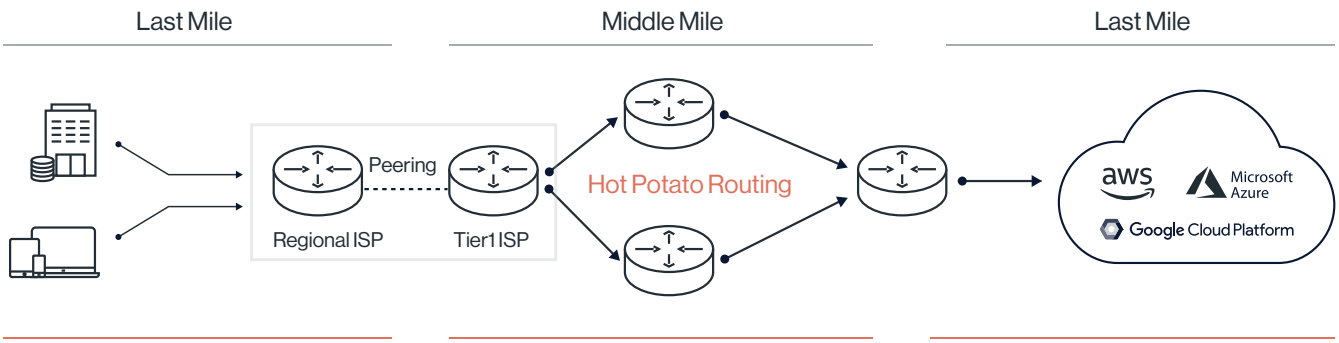


Last- and Middle-mile Optimization Challenges

As applications operate across long distances, throughput is primarily determined by latency and packet loss — not bandwidth. (See "The Impact of Latency and Loss on Throughput" to better understand why that's the case).

Improving throughput across the WAN is a matter of managing latency and loss. For MPLS, a provider assumes that responsibility and engineers its service accordingly. The public Internet is different. The tight control of MPLS gives way to a "free for all" where each network segment is individually managed, and delivers its own latency and loss characteristics.

Broadly speaking, we group these different network segments into the "last mile" and the "middle mile." "Last miles" are between the edge sites and their local ISP networks. The "middle mile" connects the two last miles. Traffic moves between these segments by providers agreeing to free, mutual traffic exchanges (peering agreements) or by one provider paying for access to the other's network (transit agreements).



Internet connections span two last miles and a middle-mile

The last mile connecting the customer premises with the local ISP's network is relatively short, minimizing the impact of latency. On the other hand, packet loss is more prevalent, caused by congestion as customer networks contend for lastmile capacity. Within developing Internet regions in particular, poor physical infrastructure also contribute to increased packet loss rates. And while it's not performance per se, availability is an issue in the last-mile as the lack of redundancy leaves enterprise networks susceptible to outages from cable cuts, router misconfigurations, and other issues.

Within the middle mile, packet loss continues to be an issue, particularly at congested peering points, but it's latency that's most pronounced. Part of this has to do with the long distances as the middle-mile stretches between the last miles. Latency is also exaggerated by today's routing practices. Providers route based on economics not application requirements or optimum performance. The result: the twisted, meandering routes all too familiar to Internet engineers.

Traditional SD-WAN appliances, if they optimize WAN performance at all, treat all segments the same. Packet loss correction and latency mitigation techniques cannot be adapted to the last or middle miles. As such, they carry over the kind of thinking indicative of the old, MPLS world where the network was one and bandwidth was limited.

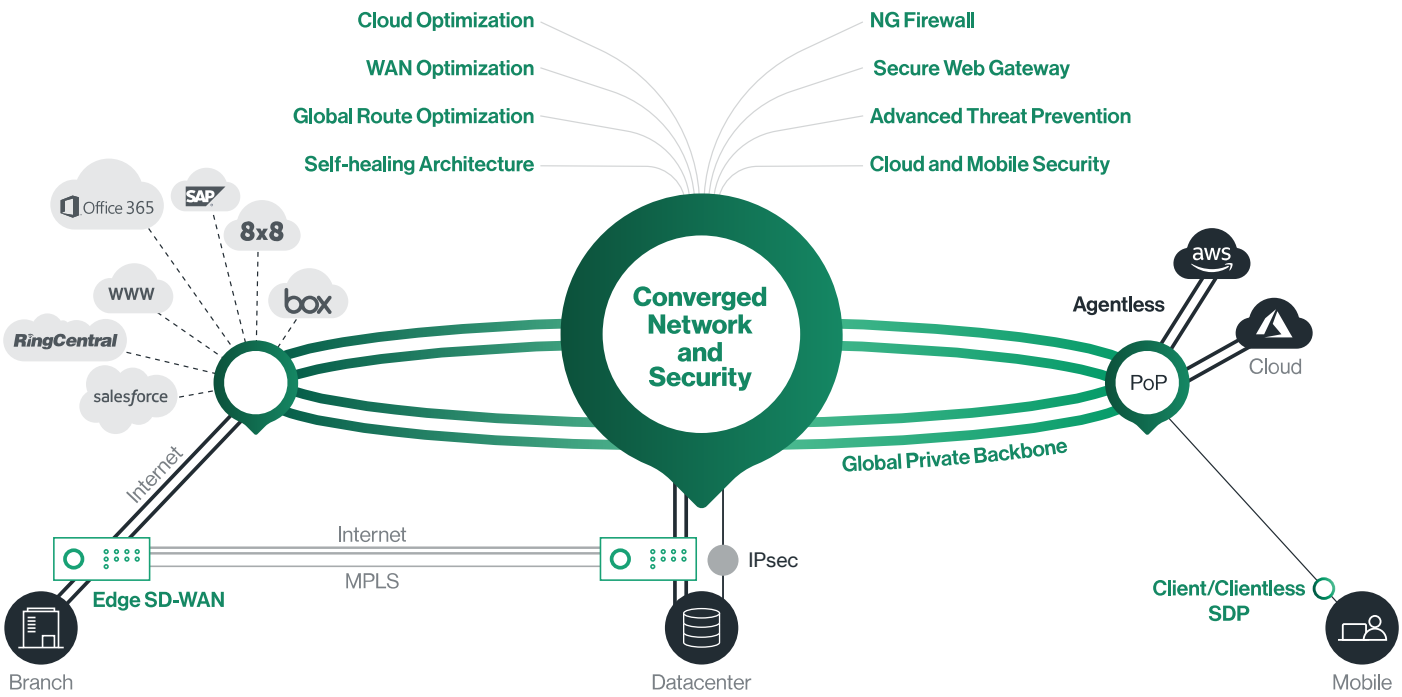
Cato Cloud

Cato Cloud is a secure, global managed SD-WAN service with optimizations designed for today's enterprise connectivity needs. It treats individual segments uniquely, applying optimization techniques according to the specific characteristics of the last and middle miles.

Cato Cloud connects locations using Cato Sockets and the Cato Cloud Network:

- Cato Edge SD-WAN ("Cato Socket"): A WAN edge appliance that connects physical locations to Cato Cloud, and between edge devices, over any last mile transport (Internet, MPLS, 4G/LTE). The Cato Socket provides application-based policy-based routing and packet loss mitigation, driven by quality of service policies and provider link performance, packet loss, and jitter.
- Cato Global Private Backbone: A global, geographically distributed, SLA-backed network of PoPs, interconnected by multiple tier-1 carriers. The backbone's cloud-native software provides global routing optimization, self-healing capabilities, WAN optimization for maximum end-to-end throughput, and full encryption.

All offices, datacenters, and cloud resources connect to the closest PoP by establishing secure tunnels from a Cato Socket. Mobile users connect by running Cato Client on their mobile devices. Network optimizations applied at the edge vary based on the capabilities of these endpoints.



Last-mile Optimizations

Cato provides the following network optimizations for the last mile between the Cato Sockets and the nearest Cato PoP, typically 25-30 ms away:

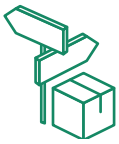
Last-mile Optimizations

Packet Loss Mitigation

Breaking the connection into segments, reduces the time to detect and recover lost packets. Where connections are too unstable Cato duplicates packets across active-active connections for all or some applications. The receiving end accepts the first packet, ignoring duplicate ones. Duplicating packets improves last-mile resiliency and can be crucial in locations where the line quality is low or when packet loss impacts the application user experience, such as when using Voice-over-IP (VoIP).

Active-Active

Cato's SD-WAN connects and manages multiple Internet links, routing traffic on both links in parallel. Using active-active, customers can aggregate capacity for production use instead of having idle backup links. The Cato Sockets and PoPs constantly monitor last-mile link performance and place traffic on the link with the least packet loss.



Brownout Mitigation

In case packet loss jumps, Cato automatically detects the change and switches traffic to the alternate link. When packet loss rates improve to meet predefined thresholds, traffic is automatically returned to primary links. "Flapping," where traffic constantly bounces between links, is prevented through configurable interval settings.

Latency Mitigation and Throughput Maximization

TCP Proxy with Advanced Congestion Control

Each Cato PoP acts as TCP proxy server, reducing latency. The proxy server "tricks" the TCP clients and servers into "thinking" their destinations are closer than they really are, allowing them to set larger TCP windows. In addition, Cato implemented an advanced version of TCP congestion control, allowing endpoints connected to the Cato Cloud to send and receive more data and better utilize available bandwidth. This increases the total throughput and reduces the time needed to remediate errors.

Dynamic PoP Selection

The Cato Sockets and the Cato Clients connect to the nearest available PoP. To ensure the best performance in the last mile, PoP selection is based on the least latency and packet loss as measured during the connection. While connected, the Cato Socket or Client continuously look for better alternatives, using real-time information updates about other available PoPs. When a better alternative is available for a predefined period of time, the client switches to the alternate PoP.

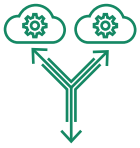
Application Quality of Service (QoS)

Application Priority

The administrator can prioritize business-critical applications, such as voice or video conferencing, or cloud services, such as Office 365, over non-critical applications. Application priority guarantees designated applications access to available bandwidth, serving other applications on a best-effort basis.

Dynamic Path Selection and Policy-Based Routing (PBR)

Cato classifies and dynamically allocates traffic in real-time to the appropriate link based on predefined application policies and real-time link quality metrics. The business requirements and prioritization contained in the policy define the application's service level. For loss-sensitive applications, such as voice and video, Cato will choose the path with the least packet loss.



Bandwidth Management

Bandwidth usage can be controlled for specific applications. Bandwidth Throttling Management allows an administrator to define rules that specify the maximum bandwidth available to an application. YouTube, for example, can be limited to a specific bandwidth or a percentage of link bandwidth.

Identity-aware Routing

Cato correlates Microsoft Active Directory (AD) data and real-time AD login events to associate a unique identity with every packet flow. Once defined, networking and security policies can be specified based on user identity and group affiliation, as well as more conventional means, such as application and IP addressing information. Abstracting policy creation from the network and application architecture makes the network simpler to configure and manage.

Middle-mile Optimizations

The middle mile is inherently "longer" than the last-mile, making latency a much larger factor. The following optimizations minimize the effects of latency across the middle-mile:

SLA-backed Global Private Backbone

The Cato Cloud is a global, geographically distributed, SLA-backed network of PoPs built from tier-1 IP transit services. With IP transit, providers pay to access other networks, avoiding the erraticness that comes from provider peering. IP transit services are backed by "five 9s" availability and .1% packet loss guarantees. As such, the Cato-provided middle mile has predictable and consistent latency and packet loss, unlike the public Internet.

Dynamic Path Selection

Cato constantly measures latency and packet loss of the providers connecting the PoPs. Traffic is placed on the best path available and routed across that provider's network end-to-end.

Optimal Global Routing

Cato's global PoPs are connected in a full-mesh topology. The Cato software calculates multiple routes for each packet to identify the shortest path across the mesh. Direct routing to the destination is often the right choice, but in some cases traversing an intermediary PoP (or two) is the better route.

The Cato Cloud Network Spans 45+ PoPs Worldwide

Active PoPs

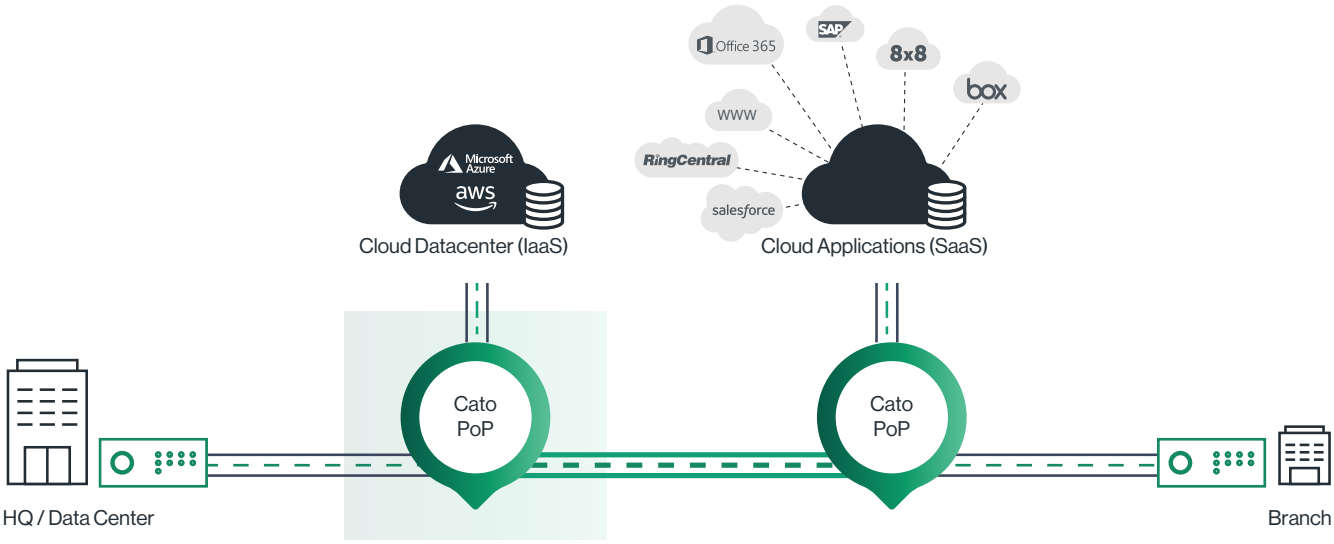
PoPs scheduled to go live over the next two quarters

Cloud Optimizations

Cato natively supports cloud datacenters (IaaS) and cloud applications (SaaS) resources without additional configuration, complexity, or point solutions. Specific optimizations include:

Shared Internet Exchange Points (IXPs)

Like content delivery networks (CDNs), the Cato PoPs collocate in data centers directly connected to the IXP of the leading IaaS providers, such as Amazon AWS, Microsoft Azure, and Google Cloud Platform. Traffic from Cato customer sites and devices is optimized and routed via the shortest and fastest path from the Cato Cloud to the customer's cloud infrastructure provider. As a result, latency to the cloud is comparable to optimized access provided by cloud providers, such as AWS Direct Connect or Azure Express Route, without the additional charge of optimized access offerings.



Cato connects into the same IXPs as IaaS vendors, providing optimized cloud connectivity

Optimized Cloud Provider (IaaS) Access

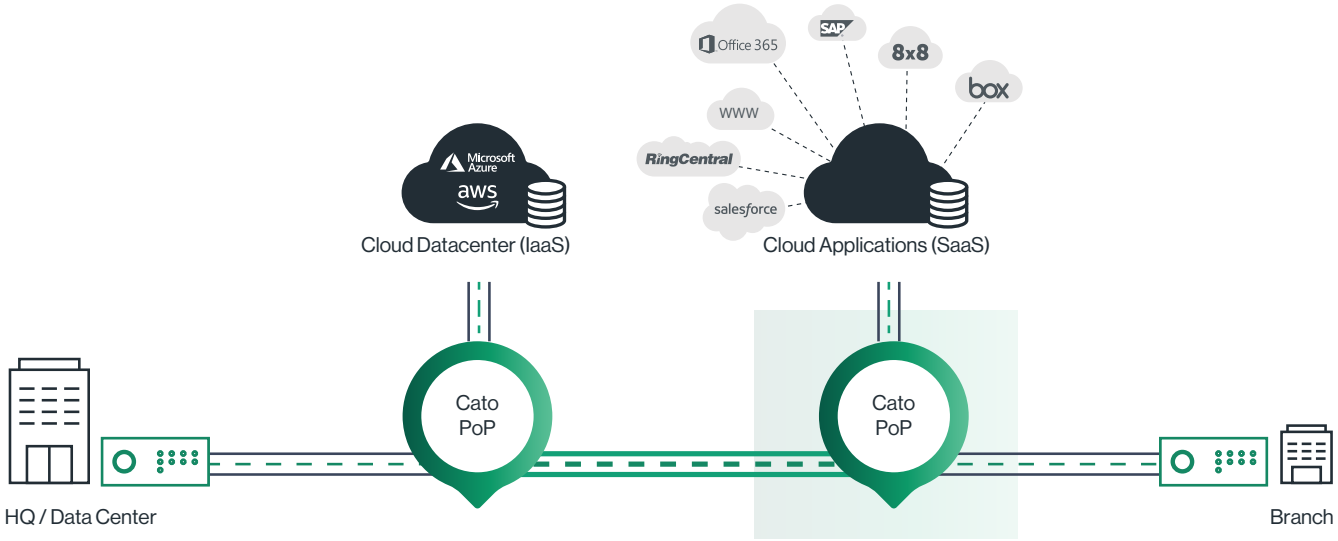
Cato places PoPs on the AWS and Azure infrastructure. By doing so, Cato guarantees that traffic between a customer's AWS virtual private cloud (VPC) or Azure instances routes across the IaaS provider's high performance backbone. Cato is expanding its PoP footprint to run on other providers as well.

Optimized Public Cloud Application (SaaS) Access

Cato offers a unique capability that optimizes and reduces latency when accessing SaaS applications. Cato customers are assigned specific IP address ranges, which are associated with the Cato PoP closest to the SaaS application datacenter. SaaS traffic sent to the Cato Cloud will route over the Cato backbone, exiting at the PoP nearest to the SaaS application. This is particularly important for applications such as Office 365 where all of a customer's SaaS traffic must reach a specific instance within a geographic location.

Optimized Public Cloud Application (SaaS) Access

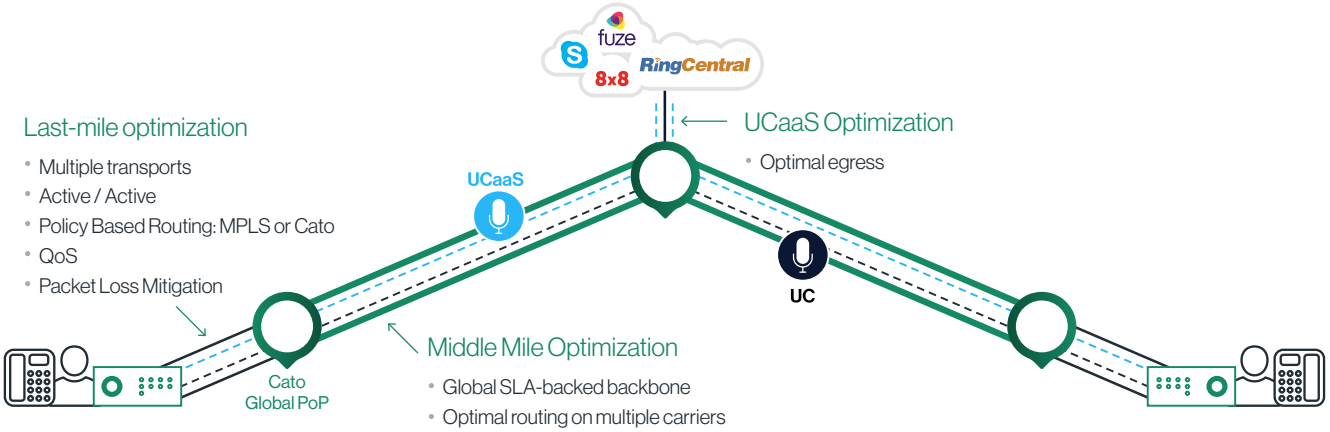
Cato offers a unique capability that optimizes and reduces latency when accessing SaaS applications. Cato customers are assigned specific IP address ranges, which are associated with the Cato PoP closest to the SaaS application datacenter. SaaS traffic sent to the Cato Cloud will route over the Cato backbone, exiting at the PoP nearest to the SaaS application. This is particularly important for applications such as Office 365 where all of a customer's SaaS traffic must reach a specific instance within a geographic location.



Cato optimizes connections to SaaS applications

Optimized UC (Unified Communications) and UCaaS Access

Cato natively supports and optimizes UC and UCaaS traffic. All UCaaS traffic is routed to the optimum UCaaS provider instance anywhere across the globe. UC and UCaaS components connected to Cato Cloud are protected against network attacks without additional dedicated appliances or security services. In addition, Cato's last- and middle-mile optimizations minimize latency and packet loss, and protect against call failures caused by network brownout and blackouts.



Mobile Optimizations

The expansion of the mobile workforce and the use of personal devices to access business data is challenging legacy network and security architectures. Traditional SD-WAN fails to address mobile users; mobile VPN solutions provide no last- or middle-mile optimization. In addition, security controls are limited for mobile VPN solutions. All too often access privileges are very coarse, forcing IT to open access to all network resources. And protecting mobile users still requires additional security tools, such as next-generation firewalls (NGFWs).

Cato Cloud natively supports mobile users with the same optimized routing, security policies and management controls as any other location or resources connected to the Cato Cloud:

Optimized Mobile Access

Cato eliminates the latency from Internet-based connectivity. Mobile users dynamically connect to the closest Cato PoP regardless of location. The PoP uses split-tunneling to securely route Internet traffic directly to the public Internet and WAN traffic across the Cato backbone to datacenters and other company locations. All relevant optimizations performed by the Cato Cloud on traffic from fixed and cloud locations are available for traffic from mobile users.

Granular Access Control

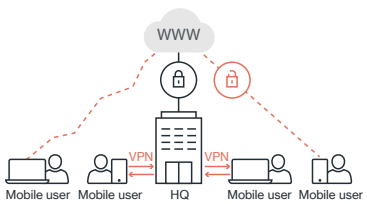
Cato provides fine-grained access control for mobile users. Access can be restricted by applications, Active Directory groups or specific user identity. Organizations can determine the precise resources that can be seen and accessed by the mobile user.

Built-in Advanced Security

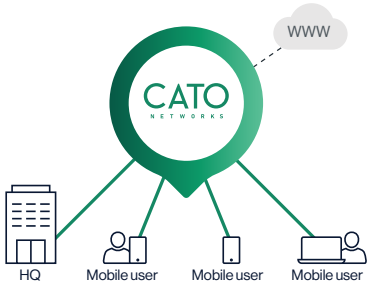
Mobile user traffic is fully protected by Cato's advanced security services that currently include NGFW, secure web gateway (SWG), threat prevention, and managed threat detection and response (MDR). All services reside in all 45+ Cato PoPs, eliminating traffic backhaul and resulting latency incurred when inspecvting traffic in a central location.

Software-Defined Perimeter (SDP) (or "Secure Application Publishing")

Cato reduces risk by restricting access to authorized resources. Mobile users running a Cato Client or through clientless browser access automatically connect to the closest Cato PoP and confirm their identities with multi-factor authentication (optional). Granular policies restrict user to accessing approved applications and resources on premises and in the cloud.



Before Cato: *slow mobile access to datacenter and cloud applications*



With Cato: *secure and optimized mobile access to all applications, globally*

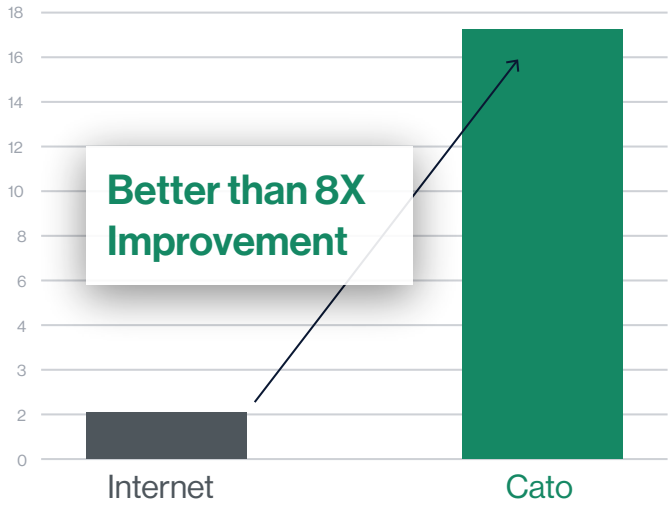
Multi-Segment Optimization

By independently optimizing the last- and middle-miles, Cato Cloud performs more effectively than if the optimizations were only applied at the edge. Within the last-mile, Cato minimizes the likelihood of packet loss and maximizes throughput with our last-mile optimizations. A Cato Socket in a location connects to multiple links (MPLS and multiple ISPs), running them in active/active mode to maximize capacity and availability.

The Cato Socket classifies and dynamically routes traffic based on application type and real-time link quality (packet loss, latency, utilization). By operating at the packet level, the Cato Socket can handle link degradation, "brownouts", not just link failure. To reduce packet loss, Cato uses Packet Loss Mitigation. Bandwidth Management allows customers to prevent applications from consuming too much bandwidth. For maximum availability the Cato Sockets automatically detect the nearest available Cato PoP. In the event a PoP becomes unreachable, the Cato Socket will automatically connect to the closest available PoP.

Within the middle-mile, Cato PoPs choose the best, SLA-backed carrier for every packet. Routing on SLA-backed carriers end-to-end eliminates the packet loss incurred during carrier peering. With the optimum network determined, Cato PoPs act as TCP proxies, dramatically improving end-to-end TCP throughput. As proxies, the PoPs make distant destination appear close to TCP clients and servers, which allows them to set large TCP windows. As a result, clients and servers can pass far more data at once before waiting for acknowledgement. Cato customers report seeing 5x-30x improvement in file download speeds.

Data Throughput: Internet vs. Cato Cloud



While TCP proxying has long been implemented in WAN optimization appliances, the latency between the two edge devices delayed packet loss recovery, reducing throughput. By contrast, since Cato PoPs sit within 25ms of either edge, they can recover rapidly from any last-mile packet loss.

What's more, locating TCP optimizations in the Cato Cloud allows them to naturally extend to any destination including the cloud. Cloud applications are optimized though Cato's ability to define egress points to exit cloud application traffic at the points closest to the customer's application instance. Optimal global routing algorithms can then determine the best path from anywhere in the world to the customer's cloud application instance.

For cloud datacenters, Cato PoPs collocate in the same physical datacenters as leading IaaS services, such as Amazon AWS, and Microsoft Azure, directly connecting to their Internet Exchange Points (IXPs). This means traffic drops right in the cloud's datacenter much like premium connections, such as Direct Connect and Express Route. The combination of TCP optimization and cloud-specific optimizations delivers a superior cloud experience.

To summarize, Cato's unique Multi-Segment Optimization combines edge- and backbone-specific optimizations, allowing Cato to optimize routing and maximize throughput end-to-end to both physical and cloud destinations.

Cato Network Optimization Features: Summary

Packet Loss Mitigation	●	
Link Aggregation and Resiliency		
Active-Active	●	
Brownout Mitigation	●	
Latency Mitigation and Throughput Maximization		
TCP Proxy with Advanced Congestion Control	●	●
Dynamic PoP Selection	●	
SLA-Backed Global Private Backbone		●
Optimal Global Routing		●
Dynamic Path Selection		●
Application Quality of Service (QoS)		
Application Priority	●	
Policy-Based Routing (PBR)	●	
Bandwidth Throttling	●	
Cloud Optimization		
Shared Internet Exchange Points (IXPs)		●
Optimized Cloud Provider (IaaS) Access		
Optimized Public Cloud Application (SaaS) Access	●	
Optimized UC (Unified Communications) and UCaaS Access	●	●
Mobile Optimization		
Optimized Mobile Access	●	●
Granular Access Control	●	
Built-in Advanced Security	●	
Software-Defined Perimeter (SDP)	●	●

Cato Networks is the Cloud-native Carrier

Your business is going digital. It depends on optimized access to applications, data on-premises and in the cloud, and an increasingly mobile global workforce. Enterprise networks of old can't keep pace with the digital business. Stitching together point solutions is difficult and resource intensive; telco services are too expensive and rigid. There has to be a better way.

Cato is the global cloud-native carrier of today. Cato connects all datacenters, branches, mobile users, and cloud resources into a global, optimized, secure, managed SD-WAN service. All WAN and Internet traffic is protected by a comprehensive suite of security services, updated and managed by dedicated security experts. Replacing MPLS and multiple networking and security point solutions with Cato Cloud forms a network so agile and efficient it can meet today's – and tomorrow's – business requirements.

Your business must leap forward to the digital age to stay competitive, and the IT infrastructure can't fall behind. Cato provides the secure and global network that is the new foundation of your digital business.

CONTACT US

Cato.
Network at the Speed of Now.

Cato Cloud

Global Private Backbone

Edge SD-WAN

Security as a Service

Cloud Datacenter Integration

Cloud Application Acceleration

Mobile Access Optimization

Managed Services

Managed Threat Detection and Response (MDR)

Intelligent Last-Mile Management

Hands-Free Management

Appendix

The Impact of Latency and Packet Loss on Network Performance

Most application traffic today rides over TCP. TCP retransmits dropped TCP packets once the sender realizes a packet was lost (due to a timeout). The time it takes for the sender to recover from packet loss doubles the latency, as the sender must wait a round trip time (RTT) before re-sending the packet.

The combination of packet loss and latency is what really degrades total TCP throughput across distance. A 1997 paper by Mathis, Semke, Mahdavi & Ott titled The macroscopic behavior of the TCP congestion avoidance algorithm, documented the impact of packet loss on TCP throughput. Throughput is obviously worse when packets are lost, but the Mathis Algorithm gave us a sense as to how fast it degrades. You can see the impact of just .1% (typical of MPLS services) and 1% (typical of Internet connection) packet loss on throughput in the chart below.

Impact of loss and Latency on Theoretical TCP Throughput