Architecting Resilient, Scalable SASE-on-SIM for Mobile Operators

The Challenge of delivering a scalable and distributed architecture for mobile subscribers

For Mobile Network Operators (MNOs), Versa SASE on SIM offers a powerful solution to integrate advanced security and networking capabilities directly into the mobile network, enabling seamless, secure access at scale.

Organizations today face the dual challenge of securing increasingly distributed workforces while maintaining high performance and scalability across their networks. Versa SASE on SIM addresses these needs by integrating advanced security and networking capabilities - without the need for client agents or complex device changes.

With Versa SASE on SIM, traffic from SIM-equipped devices is routed through a dedicated SASE enforcement point—the SSE Gateway—where it undergoes comprehensive security checks and policy enforcement before accessing the internet or enterprise resources. However, the challenge is how can subscriber traffic be routed through the SASE enforcement point using a scalable and distributed architecture?

Versa Elastic Services Cluster

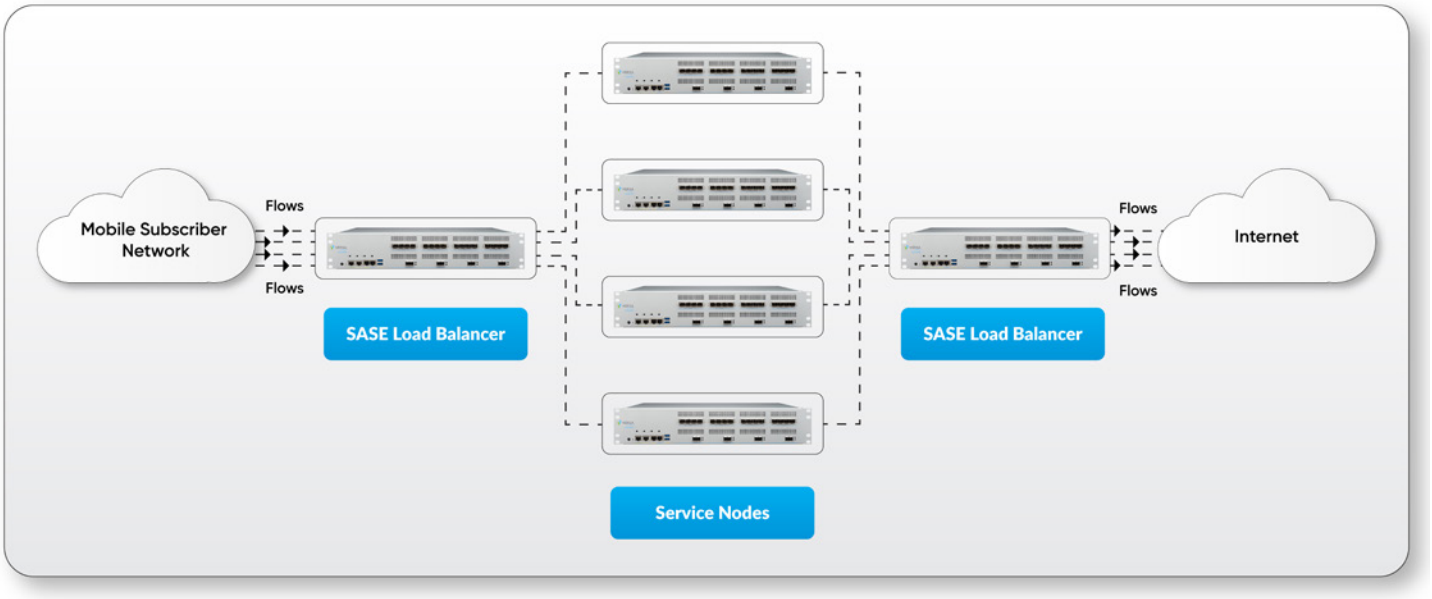

Figure 1: Versa Elastic Services Cluster

Versa's architecture for delivering high-performance security and networking services is built on a distributed, software-defined design that uses disaggregated building blocks rather than proprietary chassis. At the front end, SASE Load Balancers dynamically distributes subscriber sessions across a pool of Service Nodes—ensuring efficient load spreading and flow coherence. These service nodes, built on Versa's VOS-based appliances, perform full L4–L7 functions such as NGFW, Secure SD-WAN, and SSE. The architecture supports flow affinity, linear scaling, and seamless resilience using standard hardware and SDN methods, with optional I/O nodes for additional throughput and control. This elastic, modular approach eliminates the bottlenecks of legacy chassis systems and allows operators to scale services flexibly and independently, without disruption.

At the core of Versa's SASE on SIM architecture is the centralized SASE Load Balancer orchestration layer that intelligently manages and scales a distributed network of SSE Gateways to ensure resilient, efficient, and tenant-aware security enforcement. It dynamically balances subscriber sessions, manages gateway state, and supports multi-tenant environments, empowering MNOs to deliver resilient, efficient, and flexible security services and underpins a distributed, scalable SASE on SIM architecture that ensures optimal performance, high availability, and future-ready adaptability as network demands continue to grow.

Beyond simple load balancing, the component provides proactive gateway state management, continuously monitoring the operational health and workload of each SSE Gateway. This capability allows for automated failover and session redistribution, minimizing downtime and ensuring high availability.

The following sections provide more details on the SASE Load Balancer.

Load-Aware Load Balancing

As SASE on SIM subscriber sessions transition from the MNO to the SASE Load Balancer, the system ensures optimal performance and resilience by intelligently distributing subscriber sessions across all available SSE Gateways. This load-aware load balancing is a cornerstone of the solution's scalability, enabling the SASE on SIM service to dynamically expand or contract its capacity in response to demand.

The architecture treats SSE Gateways as a unified, distributed farm for SASE enforcement. For each new subscriber, it evaluates multiple load-based metrics—including bandwidth, memory, and CPU utilization—to determine the least loaded SSE Gateway at any given moment. This ensures that every subscriber is always routed to the most suitable gateway, maximizing efficiency and subscriber experience.

Crucially, it is not limited to co-located SSE Gateways. It performs load-aware load balancing across both co-located SSE Gateways within the same Point of Presence (POP) and geographically dispersed SSE Gateways in remote POPs. This capability is instrumental in delivering a truly distributed architecture, which not only enhances scalability but also provides robust geographic resilience—critical for supporting modern enterprise operations.

Tenant Awareness

It is designed with multi-tenant awareness, providing flexibility and scalability for MNOs. This architecture allows MNOs to offer a spectrum of security and service tiers—each represented as a tenant—across a widely distributed pool of SSE Gateways.

For example, an MNO might offer three distinct protection levels to its customers, each hosted as a shared tenant serving multiple organizations. At launch, all SSE Gateways could be provisioned to support all three protection levels. The SASE Load Balancer dynamically identifies which SSE Gateways are associated with each tenant/protection level, enabling intelligent session load balancing and efficient resource utilization.

As demand changes—such as increased adoption of Level 2 Protection following a marketing campaign—the MNO can either reallocate existing SSE Gateways or deploy additional ones specifically for that tenant/protection level. The SASE Load Balancer automatically updates its awareness of tenant-to-gateway mappings, ensuring seamless scaling and optimal performance across a distributed infrastructure. This capability allows the SASE on SIM platform to adapt as SSE Gateways are added or removed, supporting rapid growth and dynamic workload shifts.

The architecture also supports dedicated tenant models, where specific customers are assigned exclusive SSE Gateways—regardless of geographic location. It ensures subscriber sessions are always routed to the appropriate SSE Gateway, even if it is situated remotely. This approach delivers several advantages:

- Dedicated enforcement: Customers with stringent security or compliance requirements can occupy exclusive SSE Gateways.
- Shared enforcement: Other customers can benefit from shared enforcement points within their own tenant, maximizing resource efficiency.
- Flexible Scaling: As customer utilization evolves, the SASE Load Balancer automatically incorporates new SSE Gateways into the farm, simplifying expansion and ensuring the platform remains highly responsive to changing demands.

Highly Resilient Architecture

Versa's SASE on SIM architecture is built to deliver robust resilience, seamless scalability, and intelligent traffic enforcement across a globally distributed network. Through coordinated orchestration layers, including dynamic session management and load-aware routing, the architecture ensures high availability within and across Points of Presence (POPs) while supporting tenant-aware policy enforcement at scale.

Intra-POP Resilience and Scalability

Within each POP, the architecture is engineered for high availability and fault tolerance:

- Resilient SASE Load Balancer Nodes: The nodes cluster is composed of redundant nodes, ensuring protection against individual device failures.
- Dual-Homed SSE Gateways: Each SSE Gateway connects to multiple nodes, providing both control and data plane resilience. This design ensures that even if a node within the cluster fails, there is no impact on the overall capacity or availability of the SASE on SIM service.
- Carrier-Grade Protocols: Resilience is further strengthened by leveraging industry-standard protocols such as BGP and BFD, which enable rapid detection and recovery from failures.

Inter-POP Scalability and Geographic Distribution

To support a highly distributed and scalable network:

- Geographically Diverse Deployments: MNOs can deploy multiple SASE Load Balancers across multiple regions, creating a geographically distributed infrastructure

- **Dynamic Traffic Rerouting:** In the event of a POP failure, subscriber traffic is automatically rerouted to an alternate POP, ensuring continuous service for subscribers.
- **Shared Session State:** Even during network reconvergence, the ecosystem of SASE Load Balancing nodes maintains shared mappings of subscriber and SSE Gateway information. This ensures that traffic arriving at any node is correctly forwarded to the SSE Gateway originally handling the session, supporting stateful inspection and delivering an uninterrupted subscriber experience.

SSE Gateway 'state' Awareness

Versa's SASE on SIM architecture is designed to deliver seamless management of SSE Gateways across a dispersed, resilient infrastructure. The SASE Load Balancer plays a pivotal role in this environment by continuously tracking the operational state of each SSE Gateway and dynamically adjusting gateway farm membership—regardless of their location.

For example, during software upgrades, any SSE Gateway can be placed into maintenance mode. The node is instantly informed of this change and automatically removes the SSE Gateway from the active pool. As new subscriber sessions arrive, they are efficiently distributed among the remaining SSE Gateways, ensuring uninterrupted service and optimal resource utilization. Meanwhile, existing sessions are handled by the SSE Gateway under maintenance until they naturally conclude, ensuring a seamless transition for subscribers.

This approach not only supports a highly distributed and scalable infrastructure, but also ensures that maintenance, upgrades, and scaling operations are managed automatically and with minimal impact. The result is a robust, flexible architecture that adapts to changing demands and delivers a superior subscriber experience—even as the network grows or undergoes routine maintenance.

Cost Effective Port Aggregation

As the number of SSE Gateways within each Point of Presence (POP) grows to support expanding demand, a scalable and cost-effective approach to network connectivity becomes essential. The SASE Load Balancer is central to this architecture, enabling efficient port aggregation for SSE Gateways.

By directly connecting each SSE Gateway to the SASE Load Balancer, the solution delivers a highly scalable network infrastructure. This design not only minimizes costs associated with port utilization but also ensures seamless integration as the SSE Gateway farm scales up or down. It's ability to manage large numbers of SSE Gateways dynamically supports the evolving needs of MNOs, providing robust scalability without compromising on performance or reliability.

Optimising and Conserving SSE Gateway Resources

Not all subscriber traffic requires the advanced security features offered by the SSE Gateway. For example, at the discretion of the enterprise, high-bandwidth content delivered via a Content Delivery Network (CDN) can be routed directly to the internet, bypassing the SSE Gateway entirely.

The architecture enables dynamic, intelligent traffic steering across the platform. When configured, it identifies and routes non-security-sensitive traffic—such as CDN-delivered content—directly to the internet, efficiently managing resource allocation away from the SSE Gateway farm.

This approach not only conserves valuable SSE Gateway resources but also ensures that the platform remains optimized and scalable. By processing only the traffic that requires advanced security, the solution maintains high performance and responsiveness, even as network demands grow.

Key Features for Powering Distributed and Scalable SASE on SIM Deployments

Dynamically distributes subscribers across available SSE Gateways for optimal resource utilization	✓
Seamlessly balances subscribers between SSE Gateways within the same or different Points of Presence (POPs), supporting geographic diversity	✓
Routes subscribers according to real-time SSE Gateway load, ensuring efficient performance and avoiding overload	✓
Enables MNOs to differentiate services (e.g., essential vs. premium) for customers sharing a common enforcement policy	✓
Adapts SASE-on-SIM service allocation based on tenant demand, supporting rapid scaling as needed	✓
Facilitates both shared and dedicated SSE Gateway deployment models to meet diverse customer requirements	✓
Monitors and responds to the operational state of each SSE Gateway, enabling automated management and resilience	✓
Delivers high availability and fault tolerance within each Point of Presence	✓
Ensures resilience and continuity of service across multiple Points of Presence	✓
Directs trusted application traffic directly to the internet, reducing SSE Gateway workload and optimizing resource use	✓

Summary

Through the combination of Sovereign SASE and SASE-on-SIM from Versa, MNOs can deliver scalable security solutions that can adapt to dynamic workloads and evolving threats.

Versa's distributed and scalable SASE on SIM architecture enables mobile network operators to deliver robust, flexible, and resilient security services at scale. It intelligently orchestrates subscriber sessions, dynamically manages SSE Gateway resources, and supports both shared and dedicated enforcement models to meet diverse enterprise needs.

With built-in load-aware traffic distribution, tenant-aware enforcement, and seamless integration across geographically dispersed Points of Presence, the architecture adapts in real time to shifting network demands. It ensures high performance, continuous availability, and stateful session continuity—even during maintenance or unexpected network events.

About Versa

Versa is a trusted leader in Secure SD-WAN, SSE, and SASE, delivering integrated networking and security to federal, defense, and public sector agencies. Our VersaONE platform – validated in the DISA Thunderdome initiative for ZTNA, SWG, NGFW, and CASB – is aligned with DoD Zero Trust architecture and proven across both enterprise and tactical environments. It's built to help you modernize with confidence, reduce risk, and achieve your goals with efficiency and clarity.

Versa Networks, Inc, 2550 Great America Way, Suite 350, Santa Clara, CA 95054
Tel: +1 408.385.7660 | Email: info@versa-networks.com | Website: www.versa-networks.com