DATASHEET: June 2025

VMware® Avi™ Load Balancer

One Platform for Load Balancing, Application Security, Container Ingress and Analytics

Key Benefits

- 90% faster provisioning and deployment
- 49% team efficiency improvement with rapid app issue resolution in seconds through app health scores, analytics, security and client insights
- 43% reduction in TCO through on-demand auto scaling
- 27% productivity boost for DevOps with self-service and automation
- A single license for VM, container and bare metal server apps across on-premises and cloud environments

What's Included

A single platform that provides

- L4-L7 load balancing
- Web application firewall (WAF)
- Container ingress
- Global server load balancing (GSLB)
- Real-time application analytics
- Plug-and-play integration with VMware Cloud Foundation (VCF)

Rapid Application Delivery and Simplified Operations

A plug-and-play, fast-to-deploy, easy-to-use load balancing solution is critical for all applications (virtualized or container-based) across on-premises and cloud environments. Legacy hardware load balancers lack the elasticity, flexibility, and simplicity needed to deliver applications quickly, securely and reliably. The rise of containers, AI, automation and self-service mandates a shift to software-defined architectures, overcoming the limitations of the appliance-based approach while delivering enterprise-grade features with a cloud operating model.

Plug and Play with VMware Cloud Foundation

VMware Avi Load Balancer (Avi) is available as a plug-and-play solution seamlessly integrated with VMware Cloud Foundation (VCF), supporting the vision of a unified private cloud experience. Avi automates lifecycle management for virtual machines and containers alike. This integration delivers industry-leading application visibility, accelerating troubleshooting and providing rich telemetry to application owners, DevOps teams, and VCF cloud administrators. Users benefit from real-time application performance insights and advanced anomaly detection to help prevent web application attacks. Avi empowers cloud admins and developers with a consistent, self-service experience, allowing quick and secure load balancer deployment and management without deep expertise, while IT maintains full oversight. Simply stated, Avi helps organizations adopt the cloud operating model across the application lifecycle—build, operate, consume, and protect—offering out-of-the-box simplicity and efficiency for delivering modern applications.

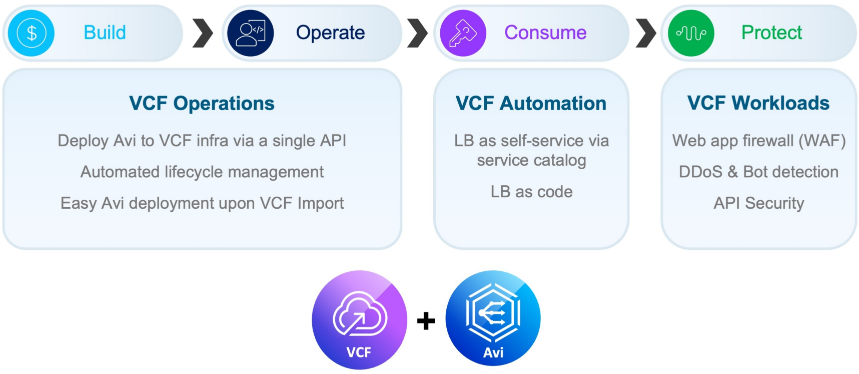
Build
Operate
Consume
Protect
VCF Operations
Deploy Avi to VCF infra via a single API
Automated lifecycle management
Easy Avi deployment upon VCF Import
VCF Automation
LB as self-service via service catalog
LB as code
VCF Workloads
Web app firewall (WAF)
DDoS & Bot detection
API Security
VCF
+
Avi

Figure 1: Avi Integrates with VCF throughout the app lifecycle

Software-Defined Platform Architecture

Avi follows a software-defined architecture that separates the central control plane (Controller) from the distributed data plane (Service Engines). Avi has 100% REST APIs, making it fully automatable and self-service with the CI/CD pipeline for application delivery. Avi auto scales on-demand and auto heals with a resilient fabric, reducing complexities and cost with 10x less devices to manage, configure and upgrade.

Policies and lifecycle management are centralized from the Controller, with full control over licenses across multiple sites via Cloud Console. DR and app resiliency for Avi is a simple license reallocation, without the complexity and cost associated with idle capacities on standby. App latency analytics is the secret sauce for fast troubleshooting and rapid app issue resolution, leading to better end user satisfaction. Avi secures Kubernetes workloads with container ingress / Gateway API and protects applications with web and API security. Security policies are kept current through live threat updates via Cloud Console.

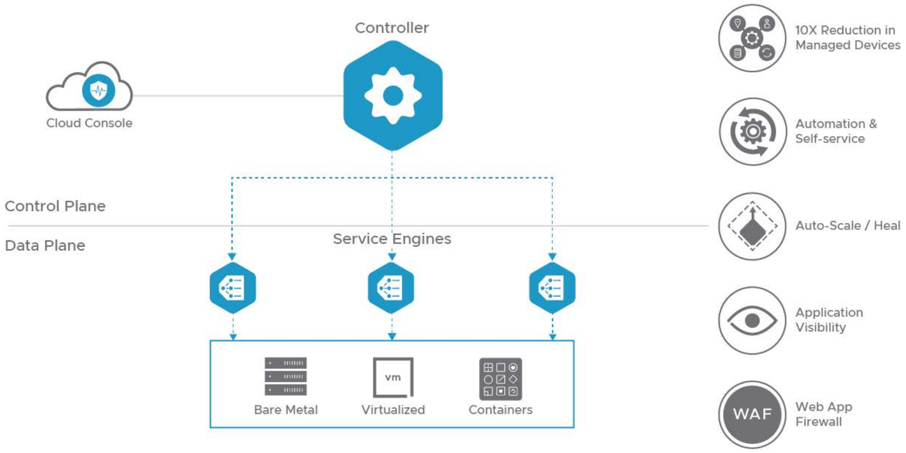

Figure 2: Avi Platform Overview

Local and Global Load Balancing

Avi provides local and global server load balancing in one platform. Controller is the “brain” of the system and acts as a single point of intelligence, management, and control across a distributed fabric of enterprise-grade load balancing, application security, container ingress and analytics. The Controller provides decision automation based on closed-loop telemetries and presents actionable insights based on based on application monitoring, end-to-end timing, searchable traffic logs, security insights, log insights, client insights, and more. The Cloud Console also delivers an always-on, as-a-service consumption model for operational capabilities such as central licensing and security feeds. See Figure 2.

Web App and API Security

Avi features a web application firewall (WAF), bot management and API protection. Customers can enforce security through signatures, positive security model and application learning mode. Avi WAF protects against OWASP Top 10 Threats, updates CRS, supports compliance requirements such as PCI DSS, HIPAA, and GDPR. With an optimized security pipeline, Avi maximizes the efficiency of resource-intensive operations. Cloud Console provides live feeds of new threat updates including IP reputation, bot detection, signatures, and more, and automatically minimize false positives with advanced security analytics, detection, and enforcement modes. With real-time app security insights and analytics provide actionable insights on performance, end-users and security events in a single dashboard with end-to-end visibility. See Figure 3.

Key Features

- Point-and-click simplicity for security policies with central control
- Elastic scale with high performing, load based automatic scale-out architecture
- Granular security insights on traffic flows and rule matches for precise policies
- Automated threat updates through Cloud Services
- Real-time app security insights and analytics
- Protects applications from DDoS attacks and OWASP Top 10 threats

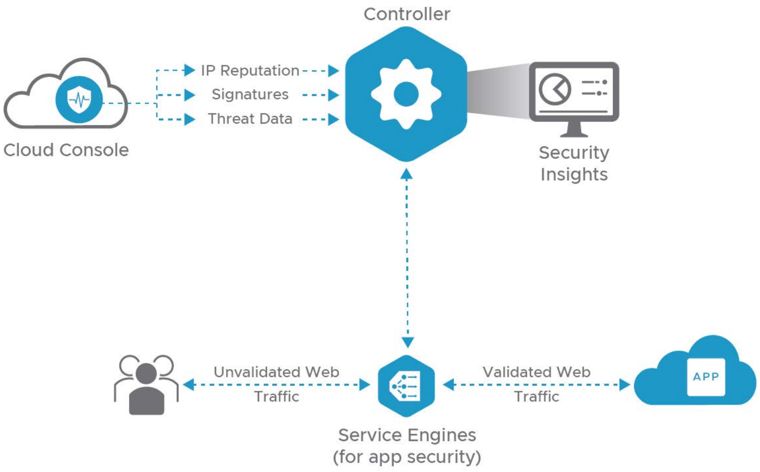

Figure 3: Web App and API Security

Kubernetes Ingress Services

Modern application architectures based on microservices have made appliance-based load balancing solutions obsolete. Containerized applications deployed in Kubernetes clusters need a scalable and enterprise-class solution for load balancing, global and local traffic management, service discovery, monitoring/analytics, and security. However, this should not be done in a disparate way with siloed DIY products to be stitched together all by the platform teams. Enterprises adopting Kubernetes need a cloud-native approach for traffic management and application networking services. Avi introduces the Gateway API, a next-gen Kubernetes ingress with standardized and native deployment to K8s, enhancing automation and future-proofing customers for Kubernetes and serverless workloads. For modern container-based applications, Avi offers a consolidated set of container services including cloud-native, scalable, enterprise-class container ingress traffic management, dynamic service discovery, and security. See Figure 4.

Key Features

Traffic Management & Service Discovery

- Local and global load balancing
- DNS / IPAM / Circuit Breaking
- Health Monitoring
- TLS termination, Cert management / automation
- CI/CD and Blue-Green / Canary deployments

Security & Observability

- WAF
- Authentication
- Allowlist / Denylist
- Rate Limiting
- DOS detection / mitigation
- Application and infra performance metrics
- Transaction tracing & fine-grained logging
- Bot Detection

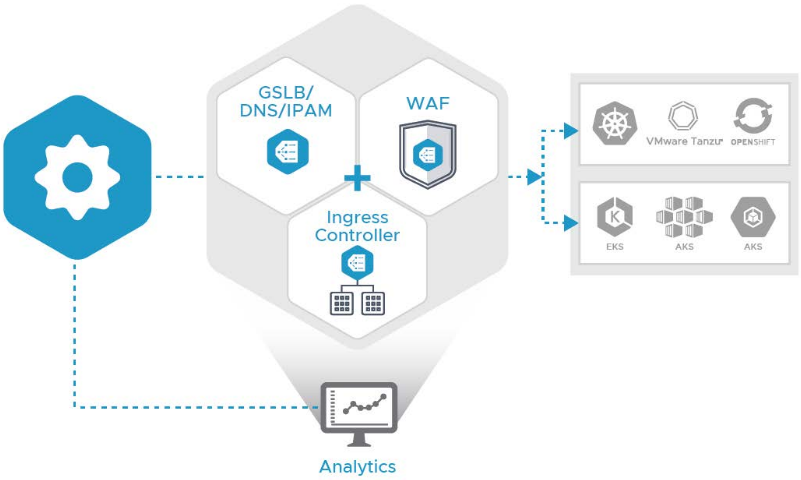

Figure 4: Kubernetes Ingress Services

Real-time Application Insights

Avi accelerates troubleshooting by delivering end-to-end visibility, AI-driven insights, and real-time configuration context through an intuitive, unified dashboard. It displays essential metrics like color coded application health scores, client and server round-trip times and application response times, allowing IT teams to monitor application health across data centers and clouds and quickly identify cross-application issues. Leveraging machine learning, Avi analyzes traffic and application behavior to automatically highlight and categorize critical logs and anomalies, such as HTTP errors or latency spikes, so teams can focus on the most pressing problems and reduce guesswork. Additionally, Avi links performance metrics to configuration changes as they happen, mapping latency and anomalies to specific updates. This enables IT teams to rapidly pinpoint root causes, resolve issues faster, and minimize downtime, ultimately improving operational efficiency and reliability.

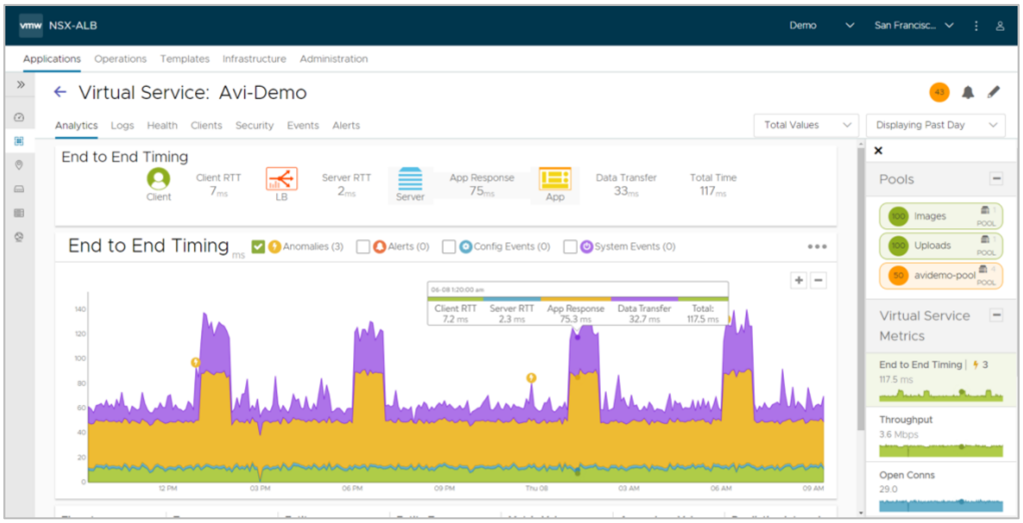

Figure 5: Avi Dashboard

Avi Benefits At a Glance

90%
Faster Provisioning[1]

AGILITY FROM AUTOMATED PROVISIONING AND SELF-SERVICE

- Automated virtual service provisioning with per-app load balancing services
- Application provisioning in seconds
- Full automation with REST APIs to support faster application rollout in Blue/Green and Canary deployments and enable DevOps teams with self-service portals
- Simplified operations with centralized policies

43%
Lower TCO[1]

LOWER COST WITH SIMPLIFIED OPERATIONS

- Elastic load balancing and just-right-size capacity without overprovisioning
- Flexible, subscription- based licensing model that eliminates static capacity
- Reduced OpEx through simplified operations of central management
- Consistent application services across any environment without reconfiguration

49%
Improved team efficiencies[1]

RAPID RESOLUTION IN SECONDS

- Near real-time visibility into network transactions to troubleshoot quickly
- Application health score for a quick snapshot of network posture
- End-to-end round-trip times with latencies between each hop
- Real-time logging, recording and replaying traffic and security events

[1] IDC Business Value Study of VMware NSX Advanced Load Balancer: A Study of Enterprises Using Next-Generation Application Delivery

VMWARE INTEGRATIONS		
vCenter	Google Cloud VMware Engine	VCF Automation (Aria or vRealize Automation)
VMware NSX	Oracle Cloud VMware Service	VCF Automation Orchestrator (Aria Automation Orchestrator)
VMware Cloud on AWS	vSphere Kubernetes Service (VKS)	VCF Operations (Aria or vRealize Operations)
VMware Cloud Foundation (VCF)	VMware Horizon	VCF Network Operations (Aria Operations for Networks)
Azure VMware Solution	VCF Automation Multi-tenancy (VMware vCloud Director)	VCF Operations Log management (Aria Operations for Logs)

3rd PARTY INTEGRATIONS	NOTES
OpenStack	Queens, Rocky, Stein, RH OSP, Keystone v3
Bare Metal	RHEL, CentOS, Ubuntu, Oracle Enterprise Linux
Public Cloud	Microsoft Azure, Amazon Web Services (AWS), Google Cloud Platform (GCP), IBM Cloud, Oracle Cloud
Container	Kubernetes, Tanzu, Rancher, OpenShift, Amazon EKS, AKS, GKE

3rd PARTY SUPPORTED PLATFORMS	NOTES
Automation	Ansible, Terraform, Python/Java/Go SDKs, vRO plugin
Analytics / Monitoring	Splunk, Cisco AppDynamics, Graphite, Datadog, Logstash, Elasticsearch, InfluxDB, Syslog, Prometheus, Zabbix
IPAM / DNS	DNS, Azure DNS, Azure DNS Private Zones, AWS Route 53, Infoblox, Custom DNS integration, Custom IPAM Integration

PERFORMANCE - OBSERVED ON A SINGLE SERVICE ENGINE		
	Baremetal Server (24 Core) with xl710 (40 Gbps)	Service Engine running as vCenter VM (6C / 6GB)
Max SSL (EC) Connections	50K per second	12K per second
Max SSL (RSA 2K) Connections	18K per second	4000 per second
Max HTTP Requests	700K per second	185K per second
Max L4 TCP Connections	400K per second	130K per second
Max SSL throughput	38 Gbps	10 Gbps
Max tenants (shared data plane)	Unlimited	Unlimited
Max tenants (isolated data plane)	200	200
Max Service Engines per cluster	200	200

CATEGORY	FEATURE
Enterprise-class load balancing	TLS 1.3 support, SSL termination, default gateway, GSLB, DNS, wildcard VIP and other L4-L7 services
Multi-site, Multi-AZ load balancing	Intelligent traffic routing across multiple sites and across private or public clouds, global server load balancing supported with Canary upgrades of leader and follower sites
Application performance monitoring	Monitor performance and record and replay network events with granular logging
Predictive autoscaling	Application and load balancer scaling based on real-time traffic patterns
Cloud connectors	VMware, SDN controllers, OpenStack, AWS, GCP, Azure, Linux Server Cloud, VMware Cloud on AWS, Google Cloud VMware Engine, Azure VMware Solution (customer-managed)
Distributed application security fabric	Granular app insights from distributed service proxies to secure web apps in real time
Application security	Bot management, Positive security model and learning mode for WAF, CSRF Protection
SSO / client authentication	SAML 2.0 authentication and authorization for back-end HTTP applications
Automation and programmability	REST API based solution for accelerated application delivery, extending automation from networking to developers with self-service portal enabled
Application analytics	Real-time telemetry from a distributed load balancing fabric that delivers millions of data points in real time
Centralized management and upgrade	Policy-based management and ability to selectively upgrade data plane with Flexible Upgrade
Networking protocols support	BGP, RHI and ECMP, BFD, IPv6, VLAN & trunking, VRF awareness, Radius and SIP, SCTP
Consolidated container services	Kubernetes Services including ingress, WAF, GSLB, DNS/IPAM on a scalable platform with support for multi-cluster, multi-site and multi-AZ container clusters, Gateway API for native K8s deployment
Central License and Visibility Platform	Controls all Cloud Services Licensing as well as providing comprehensive Global and Controller dashboarding from a centralized cloud service