White Paper

ScaleN: Elastic Infrastructure

Emerging data center models are based on flexibility and non-disruptive, on-demand scalability. Infrastructure must not only support these capabilities, but be able itself to provide these same benefits and capabilities. F5 ScaleN is a breakthrough in high availability and scalability, offering the robust capabilities required to enable multi-tenant solutions, elastic applications, and infrastructure for any environment.

by Lori MacVittie
Senior Technical Marketing Manager

Contents

Introduction **3**

F5 ScaleN **3**

ScaleN Operational Scaling 4

ScaleN Application Scaling 6

ScaleN On-Demand Scaling 9

Conclusion **11**

Introduction

Elasticity is a relatively new term, introduced with cloud computing, that expands on traditional notions of scalability. Where scalability focuses on expansion or growth of a resource, elasticity also implies the reverse: the ability to contract available resources. Elasticity is considered superior to traditional scalability because it attempts to exactly match resources to demand. Doing so eliminates costly over-provisioning methods used in the past, which ensured the capacity to meet sudden spikes in demand, and improves utilization for a better return on investment.

While elasticity as an operating and business model initially focused on applications, it has become evident that infrastructure services like load balancing and identity and access management must also be elastic. Applications are not islands and cannot properly adapt to demand if the services upon which they are dependent do not also expand and contract to meet demand. As more business environments share critical core services such as identity and access management, a failure to properly scale in the face of overwhelming demand by a single application can have a domino effect across the data center, impacting tens or hundreds of other applications.

75% of all U.S. businesses have experienced interruptions due to:

- Power.
- Hardware.
- Telecommunications.
- Software problems.

A multi-tenant architecture is a requirement for scaling all aspects of business operations. The desire to share infrastructure is noble, but compliance and security concerns may require the same level of isolation on a per-application or business-unit basis as is present in public cloud provider environments. Traditional network infrastructure simply does not support such isolation, and thus presents a challenge for the efficiency- and cost-minded organization.

F5® ScaleN™ technology breaks away from the traditional infrastructure scalability model and introduces a more efficient, elastic, and multi-tenant solution that meets the challenges and demands of modern data center architectures. It further expands the ways in which elasticity can be achieved, offering multiple scalability models to better meet the specific needs of organizations across a wide spectrum of industries.

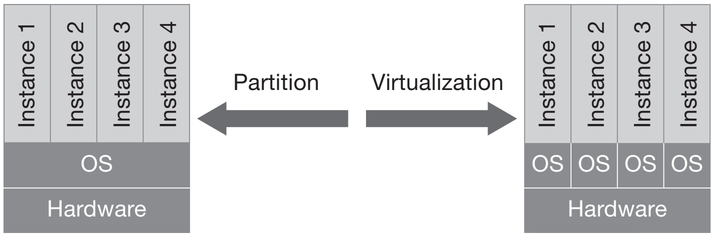
F5 ScaleN

Applications running across networks encounter a wide range of performance, security, and availability challenges. These problems cost organizations an enormous amount in lost productivity, missed opportunities, and damage to reputation. Strategies addressing these challenges have a common, critical factor: scalability. Scalability is an integral component of architectures designed to enable resiliency, improve performance, and optimize resources.

F5 augments its already comprehensive availability solutions—such as trusted N+1 high-availability (HA) architectures—with ScaleN. ScaleN is a unique approach to scalability comprising multiple models designed to meet the diverse requirements of both business and operational stakeholders in elastic and traditional environments. In addition to extending the capabilities of traditional, horizontal-scaling HA architectures, ScaleN adds virtualization and a seamless on-demand scaling option to ensure organizations can meet operational as well as business and architectural requirements.

ScaleN Operational Scaling

ScaleN operational scaling comprises two core concepts: device virtualization and partitioning capabilities.



Feature	Partition	Virtualization
Resource Allocation	Shared/Flexible	Static/Dedicated
Operating System	Single Shared	Multiple Unique

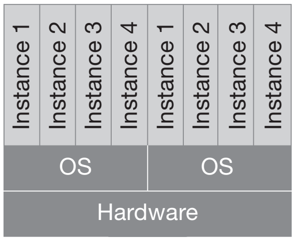
Figure 1: Approaches to infrastructure multi-tenancy include partitioning and virtualization.

There are distinct advantages to each of the two most common approaches to multi-tenancy. By adopting both approaches, a ScaleN-enabled F5 BIG-IP® Application Delivery Controller (ADC) supports a multi-tenant environment, providing network isolation and fine-grained, role-based administrative control over tenant domains. ScaleN operational scaling enables true multi-tenant architectures, inside and out, while maintaining operational separation throughout the system, including management.

Partitioning and F5 vCMP

The most common approach to multi-tenancy is used by Software as a Service (SaaS) providers, in which customers share the same software but are able to personalize its behavior by organization. For network infrastructure, this partitioning must extend into the network and include the ability to isolate routing and networking domains. Additionally, shared infrastructure is often dismissed because different product versions can require sacrificing the capabilities or performance of one application for the benefit of another. ScaleN eliminates the need to balance capabilities, performance, or costs across tenants with ScaleN operational scaling, a unique multi-tenant, virtualized architecture capable of simultaneously supporting a variety of BIG-IP versions and solutions.

This is made possible by the unique F5 Virtual Clustered Multiprocessing® (vCMP) technology, which provides the isolation required to enable per-tenant configuration, policy enforcement, and administration. Each vCMP guest can further be divided using multi-tenant features such as partitions and route domains, providing the means by which IT can support diverse business, application, and departmental requirements without sacrificing predictable performance or the simplified management of a single, consolidated application delivery platform.



ScaleN combines both partitioning and virtualization to support multi-tenancy and role isolation.

To ensure support for the varied compute needs of tenants, vCMP Flexible Allocation allows customers to designate resources such as CPU cores (and blades in VIPRION® chassis systems). Flexible Allocation on chassis-based systems further supports dynamic scaling, a capability unique to F5 that enables automatic resizing of guest clusters to support true elasticity.

> "F5 has the only solution that enables us to manage data and video traffic on a per-subscriber basis to optimize and scale these services for all smartphone users."
>
> —Senior systems engineer at a global mobile services provider

Within each virtual domain, organizations can isolate and secure configuration and policies by leveraging a role-based access system for greater administrative control. Route domains provide isolation of networks such that overlapping subnets and IP addresses do not result in conflicts that can lead to outages or disruption of services.

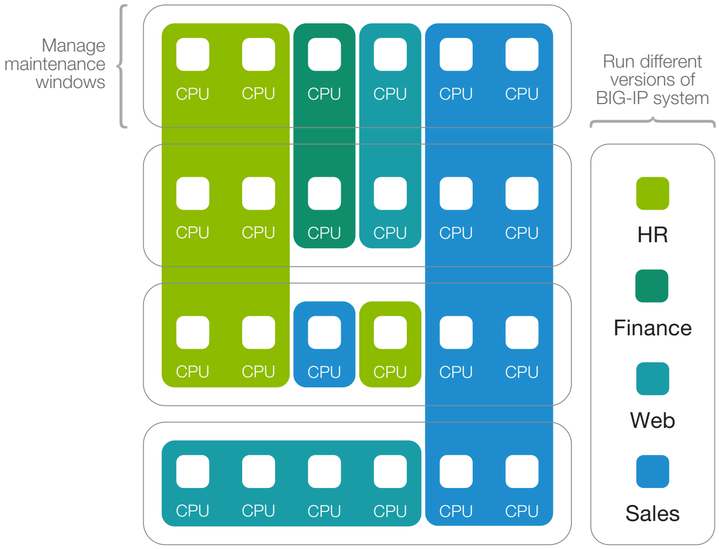

Manage maintenance windows
CPU
CPU
CPU
CPU
CPU
CPU
Run different versions of BIG-IP system
HR
Finance
Web
Sales

Figure 2: ScaleN operational scaling provides robust, multi-tenant support with flexible resource provisioning.

Isolate and protect tenants

A ScaleN-enabled BIG-IP platform is fully multi-tenant aware, enabling both administrative and network isolation on a per-tenant basis. This isolation ensures the security of each tenant by preventing network oversubscription or routing errors from affecting another tenant. Fine-grained administrative control on a per-tenant basis further protects tenants from inadvertent changes to policies or network configuration by other tenants.

Leverage flexible provisioning

ScaleN operational scaling further includes robust, flexible resource provisioning capabilities that allow operators to manage utilization on a per-tenant basis, regardless of how the organization defines a tenant. This capability empowers tenants to take advantage of the unmatched F5 programmable architecture, leveraging the built-in customization available with F5 iRules®, iApps®, and iControl® without raising concerns about the effects on other applications and tenants.

ScaleN Application Scaling

The most common means of addressing the challenge of increasing demand on services is to increase resource capacity by scaling out horizontally. This strategy is also commonly used to combat failure. Employing redundancy ensures failure of a single component does not cause downtime. This strategy has been used successfully for several decades, but it often leads to costly over-provisioning and low utilization rates that impede the return on investment of the entire architecture.

ScaleN application scaling enhances the traditional model by eliminating the need for the dedicated, standby elements, a primary source of the operational overhead. ScaleN application scaling accomplishes this by taking a platform approach, enabling BIG-IP devices to act in concert irrespective of form factor. A group of ScaleN-enabled BIG-IP devices forms a trusted delivery fabric through which applications can be scaled, secured, and delivered reliably and elastically. Because all BIG-IP devices rely upon the same platform, they can scale out via physical or virtual form factors as well as into the cloud.

ScaleN application scaling achieves this multi-directional scaling through two forms of horizontal scale: Application Service Clustering, which focuses on application scalability and high availability, and Device Service Clustering, which is designed to efficiently and seamlessly scale BIG-IP application delivery services.

Consolidate with shared infrastructure

A ScaleN-enabled BIG-IP platform applies a flexible scalability model to applications, eliminating the "all or nothing" approach to application failure associated with traditional models. Previously, critical applications have often required dedicated delivery infrastructure to avoid being affected by any failure of other applications on shared infrastructure. This model was effective, but expensive—especially when coupled with a requirement for a highly available architecture.

Through Application Service Clustering, business stakeholders can confidently take advantage of the lower costs of shared infrastructure while reducing management and maintenance overhead associated with maintaining per-project or per-business unit infrastructure services.

In the past, organizations with multiple business units often deployed multiple, dedicated BIG-IP devices to minimize possible disruptions due to the failure of an application or excessive use of shared BIG-IP resources. Application Service Clustering isolates applications, ensuring that a failure affects only that application and not the entire device. Failover can occur at the application level, rather than at the device, enabling business units and projects to share a single BIG-IP device without fear of disruption from other tenants.

Application Service Clustering

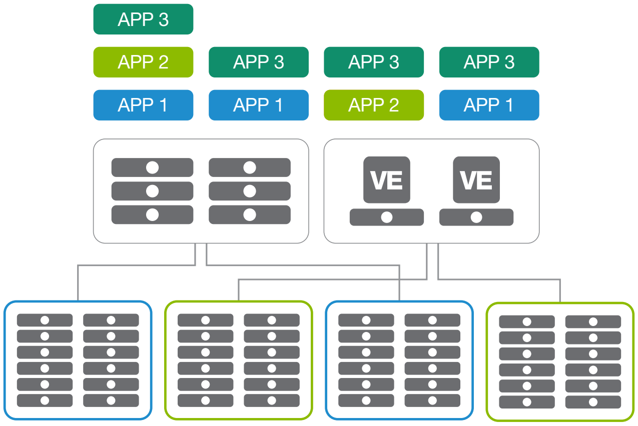


Figure 3: Application Service Clustering ensures fault isolation at the application layer for non-disruptive, lower-cost scalability and high-availability architectures.

Flexible deployment options

The flexibility to leverage any combination of ScaleN-enabled BIG-IP physical and virtual editions positions organizations to better control costs and explore opportunities to expand into cloud environments without sacrificing the benefits of a common, consistent point of control through which security, access, and delivery policies are enforced.

One F5 customer realized the benefits of Amazon Web Services (AWS) without incurring additional overhead or the complexity of multiple delivery systems by leveraging the capabilities of ScaleN-enabled BIG-IP products to scale into the cloud. Using a traditional BIG-IP device and a BIG-IP virtual edition (VE), the customer was able to seamlessly scale into the cloud on demand, realizing the cost benefits of cloud resources without giving up the ability to secure, accelerate, and manage all aspects of delivery. This process was enhanced with iApps, which leverage templates and automation to enable efficient, repeatable deployment of delivery policies. An iApp specifically for AWS ensures successful implementation of a cloud-bursting architecture with minimal effort.

	Private	Public
AWS		✓
Citrix XenServer	✓	✓
Microsoft Hyper-V	✓	✓
KVM	✓	✓
VMware vSphere	✓	✓

Figure 4: ScaleN-enabled BIG-IP products are available in a wide variety of physical and virtual form factors.

Using ScaleN application scaling, idle resources can be eliminated without sacrificing high availability and flexible scalability options. This improves overall utilization and reduces the time required to realize a full return on investment for an application delivery infrastructure.

Improve utilization and operational consistency

Implementing a highly available, elastic delivery infrastructure with ScaleN-enabled BIG-IP devices eliminates the need for idle and costly standby resources. Applications and delivery services can both scale elastically and maintain availability in the event of failure using members of a ScaleN Device Service Cluster. Policies can be systematically shared within the Device Service Cluster, reducing the operational overhead associated with manual configuration and ensuring consistent enforcement of the security and access controls critical to maintaining regulatory compliance and a strong security posture.

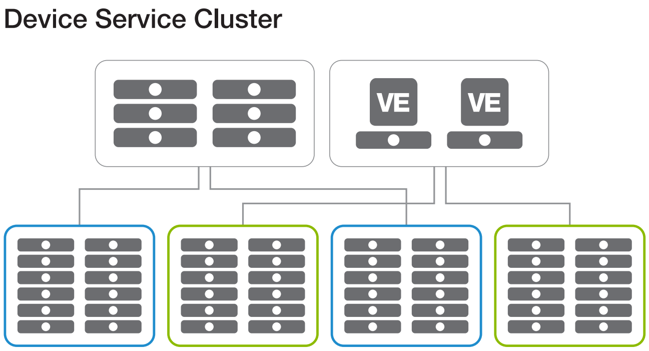


Figure 5: ScaleN scales across applications, not servers, and enables elasticity of BIG-IP services across virtual and physical instances in a Device Service Cluster.

Traditional HA architectures synchronize configurations to redundant and similar components to ensure the ability to fail over rapidly. ScaleN Device Service Clusters enable synchronization and sharing of policies with other ScaleN-enabled BIG-IP devices. This allows customers to quickly expand the capacity of BIG-IP application delivery services using ScaleN-enabled BIG-IP virtual or physical editions, on the premises or off, without requiring days or weeks of pre-positioning and configuration.

ScaleN On-Demand Scaling

Vertical scale—the addition of memory, processors and, in the case of network components, bandwidth—remains a valid method of scaling infrastructure. In most cases it requires new hardware that must then be configured and inserted into the network. This can be—and often is—disruptive, especially when the component in question is in the critical path for delivery of revenue-generating applications.

F5 supports scale-up methods traditionally, of course. Upgrading to a new, more powerful platform is always an option. But F5 also offers a non-disruptive method of scaling up as well, one that requires no configuration changes, no migration, and no maintenance windows.

VIPRION hardware provides true linear scalability through modular blades. As blades are added for additional power, they become automatically available to the system, without configuration or changes to the systems.

Both virtual and physical appliances enabled by ScaleN also support on-demand scaling through the capability to license additional capacity. On-demand software licensing enables organizations to right-size application delivery services and support growth without requiring new or higher-capacity hardware. By offering on-demand scalability across chassis, hardware, and software versions of BIG-IP solutions, F5 offers the broadest set of options for scaling application delivery services to match business growth.

Enjoy seamless growth

ScaleN on-demand scaling imbues the chassis-based VIPRION hardware with the ability to non-disruptively scale up through the addition of blades. While most blade-based systems are disruptive to change, vCMP technology allows the system to be expanded in place, seamlessly. As resources are added to the VIPRION device, they are automatically available to the system, without configuration or changes to the system. ScaleN on-demand scaling enables organizations to align capacity growth with business growth without relying on costly idle resources.

> "When we installed VIPRION, we doubled our application delivery capacity overnight. At the same time, we gained the flexibility to increase our capacity again, easily and when needed, by just adding extra blades."
>
> —Thomas Leng, Network Services Manager, Camelot

Benefit from linear scale

Because ScaleN was architected and designed entirely by F5, BIG-IP products have direct access to all available hardware resources. This means that when ScaleN on-demand scaling is employed, the BIG-IP system scales without the penalties associated with general purpose technology or the lifting of artificial limits. When VIPRION hardware running a BIG-IP product is scaled up through the addition of a blade, the entire system gains all of the blade's RAM, CPU, and network resources.

A traditional scale-up approach that merely moves artificial limits imposed on existing resources does not offer additional capacity; it only allows the system to continue to expand its use of those resources. ScaleN enables VIPRION hardware to scale linearly as resources are added to system, providing a better price/performance ratio than that of systems that scale less efficiently.

Optimize resource utilization

vCMP Flexible Allocation allows for the non-disruptive provisioning of CPU and memory resources on-demand. Available on both ScaleN-enabled appliances as well as the VIPRION chassis, Flexible Allocation enables elastic provisioning of resources across blades and CPU cores without requiring a restart, a capability unique to F5.

Flexible Allocation enables administrators to select the number of CPU cores when creating guests. On VIPRION systems, administrators can further designate any assigned slot for guests, a capability that improves mirroring functionality for guests deployed in high-availability pairs.

vCMP is able to automatically adjust resources when they are added or removed from the system. Based on configured maximum and minimum CPU and memory requirements, vCMP dynamically allocates appropriate resources based on system availability.

The ability to dynamically—and more importantly non-disruptively—allocate resources ensures that customers can scale application network services elastically and seamlessly along with application and business demand.

Conclusion

F5 pioneered and continues to innovate in application and infrastructure scalability. By advancing scalability models to a dynamic and more robust model, F5 ScaleN provides the sub-second failover and on-demand scalability required by modern business and IT organizations.

Combining a true multi-tenant infrastructure with the application packaging and isolation of Application Service Clustering, plus the addition of Device Service Clusters, ScaleN offers cloud providers and enterprise customers the flexibility of a modern, elastic application delivery network. In addition, ScaleN delivers this flexibility without sacrificing proven and reliable application delivery capabilities. Because of its integrated, platform approach to delivery services, a BIG-IP device can support layering of its scaling technologies for a scalability strategy that fits both business and IT operational requirements.

ScaleN enables the elastic infrastructure necessary for businesses and cloud providers to realize the agility, efficiency, and cost savings promised by emerging data center technologies.

F5 Networks, Inc. 401 Elliott Avenue West, Seattle, WA 98119 888-882-4447 www.f5.com

F5 Networks, Inc.
Corporate Headquarters
info@f5.com

F5 Networks
Asia-Pacific
apacinfo@f5.com

F5 Networks Ltd.
Europe/Middle-East/Africa
emeainfo@f5.com

F5 Networks
Japan K.K.
f5j-info@f5.com