Criteria for Next-Generation Hyperconvergence

Solution Brief
March 2016

Highlights

We Define Next-Generation Hyperconvergence

- We outline goals for the next generation of hyperconverged systems so that they overcome the shortcomings of today's offerings.

Requirements

- Interoperability
- Hybrid cloud support
- Automated data optimization
- Broad workload support
- Complete infrastructure convergence
- Policy-based security
- Flexible and granular scaling

Applications dictate infrastructure.

Therefore you must be able to optimize the relationship among computing, networking, and storage resources to support the needs of different applications. Traditional virtualization clusters completely separate computing and storage resources, requiring complex SAN technology and costly enterprise storage systems. Web-scale workloads employ servers with local disk storage using application software that is infrastructure aware and supports resilience with a fail-in-place model.

Existing infrastructure models fail to meet the everyday needs of IT organizations. The cost and complexity of virtualized environments make them less effective than they would otherwise be in supporting business applications. The lack of built-in, application-level resilience of most enterprise applications puts the web-scale model out of reach.

First-Generation Hyperconvergence

Hyperconvergence promised a low-cost, easy way to support a wide range of applications on a scalable, resilient platform with data distributed across the cluster servers' local storage. First-generation hyperconverged products included many compromises that caused them to fall short of the promise. For example, they have:

- **Inefficient scaling:** Most products were based on an appliance model that scales clusters only in fixed ratios of computing and storage resources, not in ratios tuned to meet the unique needs of applications.
- **Insufficient data optimization:** Many products are based on file systems that weren't designed to reduce write response times and increase performance of spinning disks. They typically lack, enterprise-class data services such as data deduplication and compression, fast, space-efficient clones and snapshots, and thin provisioning.
- **Narrow workload support:** First-generation solutions supported a limited range of hypervisors, with no plan to address the broader range of application requirements,

such as the needs of containerized and bare-metal workloads.

- **Performance shortcomings:** The network is essential to cluster and application performance, but it was left as an unspecified, manually configured, do-it-yourself project.
- **New management silos:** New GUIs simplified deployment and operation of cluster nodes, but they added new tools that didn't fit in with existing data center best practices. These tools lacked capabilities such as automated server management and APIs to support programmable infrastructure and integration with higher-level tools. Today's DevOps environments need these capabilities.
- **Security risks:** Hyperconverged environments are dynamic, but they sometimes sacrifice security to quickly move virtual machines from server to server. Network security is difficult to enforce because virtual networks are treated differently than physical ones.

Defining Next-Generation Hyperconvergence

A lack of clarity about what IT organizations really need from hyperconverged infrastructure has made hyperconvergence difficult to define as well. Numerous companies may call their products "hyperconverged," but these offerings all have different features and shortcomings that make them impossible to compare. We propose a definition of next-generation hyperconvergence that addresses these shortcomings (Table 1).

Table 1. Requirements for Next-Generation Hyperconvergence

Characteristic	First Generation	Next Generation
Interoperability	• Creation of new management islands • Isolated data not managed by data center best practices • Isolated infrastructure • No interoperability with other clusters or clouds	• Single point of management • Consistent policy management across computing, networking, and storage resources to reduce security risk • Integration with data center best practices and existing tools • Integration with hybrid cloud and support for public cloud storage • Open API that enables integration with higher-level tools and provides programmability
Hybrid cloud support	• Help in creating private clouds	• Integration with hybrid cloud solutions
Data optimization	• Features, if available, built as add-ons	• Integrated, always-on enterprise storage features • Data lifecycle management
Workload support	• Virtualized workloads only	• Virtualized workloads with broad hypervisor support • Containerized workloads to support lightweight services • Bare-metal workloads running directly on nodes
Infrastructure convergence	• Software-defined storage	• Software-defined computing with composable infrastructure • Software-defined networking • Software-defined storage
Security	• Virtual networks with limited visibility and control	• Automated and policy-based • Isolated application tiers, application instances, and tenants • Microsegmentation to provide enhanced security for east-west traffic within the data center • Incorporation of physical servers and virtual machines with equivalent visibility and control
Scaling	• Rigid, monolithic appliances	• Microscaling of all resources on a highly granular basis

Interoperability

Next-generation hyperconvergence needs to be fully integrated and interoperable with the data centers that IT organizations have in place today and will have in place in the future. It must include these features:

- **Centralized management** must allow hyperconvergence to be deployed across local data centers, corporate campuses, remote data centers, and edge computing environments. Management should be consistent with other data center tools so that hyperconverged infrastructure can connect to other private resources, such as hardware appliances and bare-metal servers, other virtualized services within the data center, and public cloud services beyond the data center, providing hybrid cloud capabilities.
- **Data lifecycle management** must be supported with features that help integrate the cluster's data with the rest of the organization's data. Required features include fast, space-efficient snapshots to support backup operations and asynchronous replication; thin provisioning to make more efficient use of storage; and rapid, space-efficient clones to support today's agile development processes.
- **Hybrid cloud support** must include self-service, management, administration, and chargeback capabilities that are simple and consistent with other data center infrastructure. Next-generation environments should include support for public cloud storage for low-cost data archival, backup, and disaster-recovery operations. It must integrate with hybrid cloud computing platforms to support additional use cases.
- **Policy-based consistency** is necessary to specify and enforce operation best practices for the storage, networking, and computing elements of the physical and virtual infrastructure. Next-generation infrastructure must directly map the application intent with the infrastructure policy required. This approach facilitates continuous service delivery while providing secure isolation between applications and tenants.

Automated Data Optimization

To simplify storage deployment, data optimization needs to be automatic, with no tuning or configuration required. Data should be striped across the nodes in a cluster, with automatic placement in tiers to increase performance while reducing cost. Always-on deduplication and compression should reduce the amount of storage needed, helping increase the cost effectiveness of the hyperconverged solution.

Broad Workload Support

Enterprise workloads have varied infrastructure requirements, and next-generation hyperconvergence must support all hypervisors, containerized environments, and bare-metal workloads. The infrastructure must scale up and down quickly and easily to support varying workload demands. First-generation hyperconverged products emulate web-scale infrastructure in which applications are predictable, homogeneous, and less complex than everyday IT workloads. Next-generation hyperconverged solutions need to support a broader spectrum of IT requirements.

Complete Infrastructure Convergence

All resources should be software defined, including computing, networking, and storage resources and even the cluster's software. Computing resources must be composable through software so that applications themselves can create the hardware constructs they need to grow. Software-defined networking capabilities are needed not just to support large and complex clusters, but also to securely isolate different tenants and applications. Next-generation hyperconvergence also requires centralized, zero-touch management of the cluster's entire infrastructure with a unified control plane and APIs that allow access from other tools, including the cluster's applications.

Policy-Based Security

Next-generation hyperconverged environments should have automated, policy-based security. Policies should define the allowed interactions between application tiers, and they should securely isolate different application instances and tenants.

Physical and virtual networks should be able to interconnect easily, with no limitations on connection of physical servers, virtual machines, containers, and physical appliances. They should support microsegmentation for precise isolation and segmentation with advanced services. Networks connecting virtual machines should have the same visibility as physical ones, so that administrators have the same level of control regardless of the implementation model.

Flexible and Granular Scaling

Hyperconverged infrastructure should be characterized by flexible scaling, allowing IT organizations to add resources simply by connecting a new node to the cluster. Resources should be identified automatically, integrated into the cluster, and put into service with single-click simplicity. With this approach, clusters can be scaled quickly and easily so that IT organizations can respond rapidly to changing resource demands.

In next-generation infrastructure, all resources should be capable of granular scaling to allow the infrastructure to be fine-tuned to meet workload needs. Microscaling allows you to add nodes without adding storage. If you need more storage, you should be able to add disk drives to existing nodes. If you need greater storage performance, you should be able to adjust the ratio of caching devices to capacity devices.

Just as cloud computing allows applications to reproduce and scale themselves virtually to accommodate workload needs that they detect, hyperconverged environments should enable applications to scale physically through creation of more physical resources using composable infrastructure.

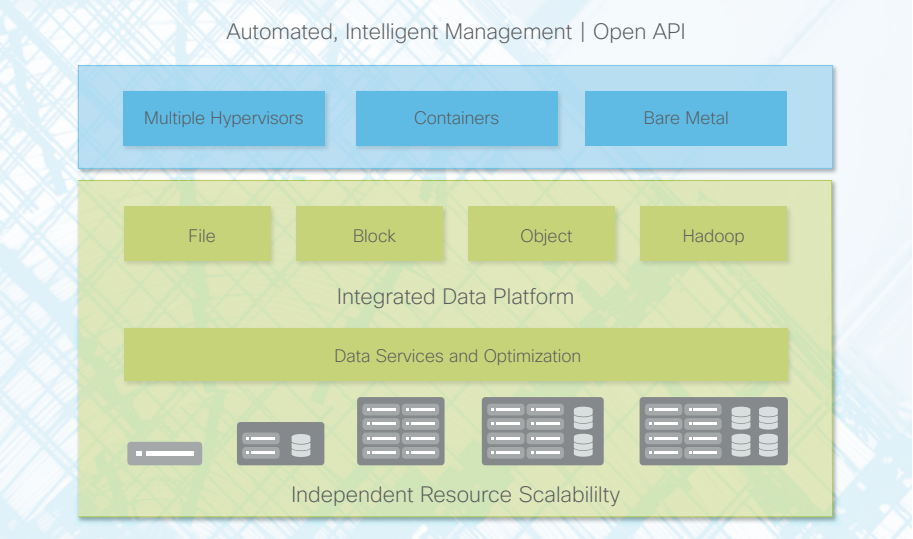

Figure 1. Our Vision for Next-Generation Hyperconvergence

Toward a Vision of Microconvergence

The first generation of hyperconvergence brought together computing and storage in a cluster, simplifying the process of deploying virtualized clusters. We envision a next generation of hyperconvergence that disaggregates all resources so that they can be composed into a cluster that

allows precise control over the balance of computing, networking, storage, and even performance resources. With this vision, we imagine an environment that supports not just virtualized applications, but those that reside in operating system containers and on bare-metal servers, all sharing the solid platform created by the cluster software and that incorporates a high-availability data engine with enterprise-class features (Figure 1).

Our vision of computing has been fabric based and software defined since we introduced Cisco Unified Computing System™ (Cisco UCS®) in 2009. Cisco UCS management enables you to treat infrastructure as code so that you can program hardware as if it were software. Every identity and configuration setting of every device in the system is software defined through Cisco UCS service profiles, and a unified system control plane is made accessible through an open API. If you build your platform beginning with a high-performance, low-latency, unified fabric that is self-aware and self-integrating, building an environment founded on fabric-based, composable infrastructure is a straightforward process. First-generation hyperconvergence moved storage back into servers. Now our fabric-based solution moves the network into the computer. This approach enables precise, microconverged integration of computing, storage, and networking resources for extremely tight coupling of resources to application demands.

For More Information

Visit http://www.cisco.com/go/hyperflex.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices**.

LE-55405-00 03/16