Cisco HyperFlex and Cisco Workload Optimization Manager

Achieving Agility and Predictable Performance
Contents

IT agility: more important than ever, yet harder to achieve ................................................................. 3
The rise of hyperconvergence and the need for decision automation .................................................. 3
   The solution: Cisco HyperFlex and Workload Optimization Manager .............................................. 4
Next generation hyperconvergence meets decision automation .......................................................... 4
Self-managing workloads require a new approach to automation ....................................................... 5
   Integrated solution architecture .......................................................................................................... 6
   Solution use cases ............................................................................................................................... 6
   Super-cluster optimization .................................................................................................................... 6
   Intelligent independent scaling of compute and storage ..................................................................... 9
   Maximizing elasticity with multicloud control .................................................................................... 10
   Modernize at the pace of your business .............................................................................................. 12
Future directions .................................................................................................................................... 14
Self-managing Kubernetes across Cisco HyperFlex and Multicloud ................................................ 14
Achieving self-managing workloads with Cisco HyperFlex and Workload Optimization Manager ...... 14
Conclusion ............................................................................................................................................... 15
IT agility: more important than ever, yet harder to achieve

Today’s organizations operate in unforgiving times. According to IDC, by 2020 50 percent of the Global 2000 will see the majority of their business depend on their ability to create digitally enhanced products, services, and customer experiences. Those that fail will be competing for smaller and smaller shares of their market. Success relies on an IT organization’s ability to be agile and enable revenue-generating innovation that responds quickly to market needs.

Historically, however, cumbersome infrastructure management and unpredictable workload performance have thwarted these aspirations. As a result, infrastructure and operations teams are stuck in a maintenance-focused model of just “keeping the lights on” rather moving to an innovation model that makes IT a strategic enabler of business demands. In this age of relentless competition and customer-centricity, these IT challenges are more complex and failure to address them more consequential than ever.

The rise of hyperconvergence and the need for decision automation

The deployment of hyperconverged systems has risen out of the need to simplify IT and increase its capacity to adapt quickly. IT leaders want to free their organizations from the yoke of cumbersome infrastructure management and unpredictable performance, simplifying the infrastructure with preinstalled, preconfigured, easy-to-update, and designed-to-work-together solutions. It is therefore no wonder that, in 2017, the market revenue for hyperconverged systems grew 64 percent, a testament to the demand for on-premises simplicity and the speed that it enables. But, as Gartner notes, simplicity is not enough: “. . . the simplicity of software-defined integration will be more than countered by the complexity of an expanding IT cosmos comprising millions to billions of data points and entry/access gateways in a matrix of connections. However, the addition of Artificial Intelligence (AI) and Machine Learning (ML) could balance and reduce this complexity significantly by creating a new paradigm of self-organizing systems.”

The digitization of businesses has prompted a revolution in application architectures and development, which in turn have transformed the way IT must operate. Applications are becoming more distributed and updated more frequently as developers increasingly adopt micro-services and cloud-native technologies. There are more moving parts across multiple clusters, data centers, and clouds, and more frequent changes to the IT environment. These hybrid and multicloud environments are beyond human capacity to manage.

Hyperconverged systems simplify infrastructure deployment, but the fluctuating resource-demands due to recent Google Cloud application trends will continue to challenge IT’s ability to ensure predictable application performance. These challenges require a new approach to infrastructure automation. Cisco HyperFlex™ and Cisco® Workload Optimization Manager together provide customers with the ultimate automation solution: self-managing workloads. With self-managing workloads, day-to-day operations are reliably managed by software, increasing IT productivity and allowing teams to focus on how to bring products, services, and experiences to market faster.

These hybrid and multicloud environments are beyond human capacity to manage.
The solution: Cisco HyperFlex and Workload Optimization Manager

Predictable performance on self-managing systems

The evolution of IT automation has been a process decades in the making. From manual resource management to scripting and, more recently, orchestration—every advancement has been about saving time, and increasing agility and speed—while simultaneously driving performance and efficiency. With each advancement, software has taken increasing control with an IT worker still at the helm. But what if you could free IT teams from the business of managing infrastructure and resource availability? What if, instead, they could spend their time solving business challenges?

Next generation hyperconvergence meets decision automation

The next wave of IT automation is here—where workloads self-manage and IT runs lean and fast. Cisco HyperFlex Systems deliver flexible and scalable, enterprise-class, hyperconverged solutions that simplify infrastructure management. Cisco Workload Optimization Manager automatically and continuously adjusts infrastructure resources to meet dynamic application demand, driving a perpetual state of health in the IT environment. It prevents the resource contention that causes applications to run slowly, and frees IT teams from “firefighting.” Together, Cisco HyperFlex and Workload Optimization Manager give customers the decision automation they need to deliver predictable application performance while enabling IT agility.

- **Cisco HyperFlex Systems** combine software-defined computing in the form of Cisco Unified Computing System™ (Cisco UCS®) servers and software-defined storage with the powerful Cisco HyperFlex HX Data Platform Software. Together, these elements comprise an adaptive infrastructure that seamlessly integrates with your IT environment and provides a unified pool of resources to power applications that meet your changing business needs. When integrated with Cisco Workload Optimization Manager, this infrastructure adapts automatically and intelligently, driving continuous application health.

- **Cisco Workload Optimization Manager** is a decision engine that helps ensure that the applications running on HyperFlex™ get the resources they need when they need them. The platform continuously analyzes workload consumption, costs, and compliance constraints in the environment and automatically allocates resources in real-time. It evaluates the multidimensional resource trade-offs, from the application layer to the infrastructure layer, to deliver workload automation that drives perpetual health across clusters, data centers, and clouds.

Together, Cisco HyperFlex Systems and Workload Optimization Manager deliver self-managing workloads that dynamically adjust to changing resource needs by automating specific actions:

- Workload placement (initial and continuous)
- Workload sizing
- Workload provisioning
- Resizing virtual storage
- Provisioning or suspending virtual storage (for example, data stores)

Cisco Workload Optimization Manager can also make recommendations for:

- Provisioning or suspending HyperFlex compute-only nodes
- Provisioning or suspending HyperFlex node (physical storage)
The infrastructure becomes as elastic as it is performant, adjusted in real time by intelligent software to meet application demand, enabling IT to focus on service delivery to developers, line-of-business end users, and customers.

**Self-managing workloads require a new approach to automation**

While automation is generally accepted as the answer to dealing with complexity and increasing efficiency, what to automate and how to do it has challenged IT. At the heart of this challenge is the workload, whether it applies to the virtual machine, container, or micro-service of today—or the to-be-determined nano-service of tomorrow.

Workloads can only become self-managing if they are automated by a decision engine that meets three critical requirements:

- Has **accurate, comprehensive, real-time data** about the full stack of technologies in the IT environment, from the increasingly interconnected applications to the software-defined infrastructure to the hardware—from the logical to the physical.
- Can **continuously make trade-offs among IT resources** to keep all workloads in a healthy state. Today’s workloads operate in an interdependent environment where a change “here” always has an impact “there.”
- Has the capacity to **make decisions and then automatically take action**, without the need for human intervention. The IT environment will only become more dynamic and distributed. Today it is hybrid and multicloud, but tomorrow it is an interconnected mesh of things, from cell towers to coffee machines.

There is one additional requirement that does depend on IT workers. Workloads can only become self-managing if IT sufficiently trusts the decisions taken by the self-management software to allow it to run automatically. In order for IT to trust those decisions, the impact of each individual decision has to be limited, specific, and low-risk. Each decision needs to deliver an incremental benefit. Cisco WOM calls these incremental benefits "micro-improvements." The cumulative result of a series of "micro-improvements" is what makes Cisco WOM’s approach industry-leading: we focus on keeping the IT environment healthy, instead of fixing problems after the environment is already sick.

**It’s your automation journey**

Achieving self-managing workloads does not happen overnight. Handing the reins over to software is a journey. In order to advance their automation journey, customers must be able to trust the actions that their software takes, and that trust must be earned. With that in mind, Cisco Workload Optimization Manager (Cisco WOM) provides different levels of automation that administrators can set for specific types of actions (for example, placement, scaling, etc.) and for specific parts of the infrastructure (virtual machines, hosts, storage, etc.). These automation levels include:

- **Recommend**: Actions appear but are “view only.”
- **Manual**: Actions can be executed in the user interface by selecting and clicking “Apply.”
- **Automated**: Actions will be executed in real time.

When fully automated, Cisco WOM and HyperFlex deliver self-managing workloads on elastic hyperconverged infrastructure.
**Integrated solution architecture**

Cisco WOM and Cisco HyperFlex are tightly integrated to ensure predictable application performance on elastic hyperconverged infrastructure. Cisco WOM uses APIs to collect data and understand the IT environment. It then analyzes the data to determine the specific actions required to maintain health continuously in the environment.

The diagram above is specific to the integration of Cisco WOM and Cisco HyperFlex. As part of this integration, however, Cisco WOM also collects data from the hypervisor to inform its decisions. See Cisco WOM documentation for more information.

**Solution use cases**

There are four key use-cases for the solution incorporating Cisco WOM and Cisco HyperFlex:

1. **Super-cluster optimization**
2. **Intelligent independent scaling of compute and storage**
3. **Maximizing elasticity with multicloud control**
4. **Modernizing at the pace of your business**

The subsequent sections cover each of these use cases in more detail.

**Super-cluster optimization**

Infrastructure silos have traditionally been employed to simplify management. But they also decrease efficiency and increase performance risk, creating situations where workloads can be contending for resources on a heavily loaded cluster while there are resources available on a lightly loaded cluster. Flattening the infrastructure avoids these performance risks, but only when software is ensuring the performance of the increasingly complex and distributed applications that infrastructure supports.

Customers with multiple Cisco HyperFlex clusters can achieve predictable performance while flattening their infrastructure thanks to Cisco WOM. Cisco HyperFlex supports a limited number of nodes in a cluster (64 nodes, as of the 3.0 release), which has traditionally meant that certain workloads are dedicated to certain clusters.
With Cisco WOM, the platform continuously analyzes workload consumption across Cisco HyperFlex and Cisco UCS, as well as gear that is not proprietary to Cisco. It determines when, where, and how to move and/or resize existing workloads to either mitigate resource contention or increase efficiency, maintaining continuous health across clusters.

**Figure 1.** Cisco Workload Optimization Manager continuously analyzes workload consumption across Cisco HyperFlex clusters, executing real-time workload placement, sizing, and capacity to drive continuous health.

There are situations, however, where dedicated infrastructure has nothing to do with simplifying management and everything to with justifiable policies that make sense for the business. These include data sovereignty, business continuity, software licensing, and the like. In such cases, the Cisco WOM platform automatically ingests these policies and ensures that the actions it takes to ensure performance and maximize efficiency always abide by those constraints. Policies can also be defined in the Cisco WOM user interface. But again, with software in control, infrastructure silos are only implemented to meet specific business requirements, not IT’s operational needs.
Figure 2. Cisco Workload Optimization Manager provides a specific action to move the virtual machine “cisco-opsmgr-1.1.3” from host “10.0.210.54” to host “10.0.210.5” to ensure workload performance.
Figure 3. Cisco Workload Optimization Manager provides a specific action to move the virtual machine “turbo61hx” and storage “FNTOP” between clusters—from host “10.0.210.53” to host “m4-esx1.cdnivt.cisco.com”—to ensure compliance with business policies.

With the integration of Cisco WOM and Cisco HyperFlex, workloads dynamically leverage the underlying infrastructure based on application resource needs in real time. Cisco WOM executes workload-placement and sizing actions through VMware vCenter, but the tight integration of vCenter with Cisco HyperFlex (and Cisco UCS) ensures that the workloads are continuously optimizing the use of the underlying compute and storage.

**Intelligent independent scaling of compute and storage**

While the trend in simplifying hardware has taken hold, the reality is that workloads come in all shapes, sizes, and levels of complexity. Hyperconverged solutions let users easily scale by adding additional nodes to the cluster, but this also adds cost in the form of additional software licensing. Cisco HyperFlex is unique because it gives customers the flexibility to scale compute and storage independently of the software stack by seamlessly integrating Cisco UCS compute-only nodes and external storage to maximize cost benefits when customizing to their unique requirements. Cisco WOM further enhances this by scaling these resources based on real-time workload consumption, providing intelligent elasticity in Cisco HyperFlex clusters that preempts performance degradation.
Figure 4. Cisco Workload Optimization Manager provides preemptive actions to provision additional storage or compute, mitigating performance degradation due to lack of resources.

At the virtualization layer—connected via the VMware vSphere APIs—Cisco WOM understands when workloads need more storage or compute and will allocate more or less, based on real-time resource needs. Cisco WOM also understands the resource availability within Cisco HyperFlex clusters through the Cisco HyperFlex API. When existing cluster nodes cannot support the current demand, Cisco WOM recommends provisioning more compute or storage depending on the specific resources needed. Users can then add Cisco UCS compute-only nodes, if more compute alone is required, or another Cisco HyperFlex node, if linear scaling of compute and storage is required. Likewise, when demand subsides, Cisco WOM recommends suspending specific servers or hosts that are no longer required. Customers are able to prevent performance degradation that would have occurred had the necessary infrastructure been unavailable.

Figure 5. Cisco Workload Optimization Manager provides the action to provision a new disk array similar to “CCHX” (a Cisco HyperFlex node) in order to ensure performance.

Maximizing elasticity with multicloud control

Most organizations today are leveraging public cloud resources as part of an overall multicloud strategy. The easy-to-deploy, pay-as-you-go model is ideal for short-lived workloads or as a safeguard against sudden spikes in demand. For example, for online tax services or e-commerce sites, which experience seasonal spikes in demand, the elasticity of public clouds can help at peak times. Throughout the rest of the year, however, these same organizations benefit from their own highly predictable and efficient data centers.
Figure 6. Cisco Workload Optimization Manager drives continuous health in on-premises and public-cloud environments, enabling maximum elasticity across the IT estate.

With Cisco WOM, customers make the most of their investments in Cisco HyperFlex, ensuring performance and achieving elasticity on premises, but also efficiently leveraging public cloud resources when needed. The unified platform controls both on-premises and public-cloud environments. Cisco WOM currently supports Amazon Web Services and Microsoft Azure environments and is adding support for other providers.

Figure 7. Cisco Workload Optimization Manager provides a global view of hybrid infrastructure, in this case Cisco HyperFlex and AWS Cloud, with specific actions for both environments.

Control cloud assets to deliver service levels
- Automatically scales workloads, storage, and databases in real time to ensure performance
- Tracks, reports, and trends compute and storage and database-consumption metrics (CPU, memory, IOPs, latency and DTU) across regions and zones
Minimize public cloud costs

- Continuously optimizes reserved-instance recommendations in concert with resizing automation – to drive the elasticity and cost savings that you want and that you expect from the cloud
- Automatically scales down Azure virtual machines or AWS instances, storage tiers, and database tiers, reducing costs without impacting performance
- Identifies ghost and unattached storage instances
- Suspend or terminates unused instances
- Projects actual cost of workloads by projecting compute, licensing (OS), IP address, and storage costs
- Aggregates monthly bills across services, regions, accounts, specific workloads, and lines of business; tracks in a single view against a defined budget; and projects future monthly bills

Modernize at the pace of your business

Cisco HyperFlex Systems and Cisco WOM seamlessly integrate with the data center you have today so that you can modernize at the pace of your business. In addition to assuring the performance of workloads across your heterogeneous environment, Cisco WOM’s capacity management capabilities allow you to quickly model “what if” scenarios as your business grows and as you retire legacy infrastructure.

Figure 8. Use Cisco Workload Optimization Manager to model infrastructure and workload growth scenarios to determine quickly how much infrastructure you will need and when you will need it.
Cisco HyperFlex

Figure 9. Cisco Workload Optimization Manager analyzes workload growth trends to determine when new hardware will need to be purchased. Note the “Top Clusters” highlighted and their “Time to Exhaustion,” ensuring that you always have the hardware you need to support growth.

Figure 10. These Cisco WOM plan results show that workload demand, in this case 276 virtual machines (see supply chain on left), on 5 legacy hosts can be supported with 1 Cisco HyperFlex host for a one-time investment of $9000. Customers get specific before-and-after views in order to make fully informed decisions as they modernize their infrastructure.
Future directions

Self-managing Kubernetes across Cisco HyperFlex and Multicloud

Gartner predicts that we are on the cusp of a new phase of integrated systems:

Phase 1 is the peak period of blade systems (2005 to 2015); Phase 2 marked the arrival of converged infrastructures and the advent of hyperconverged integrated system (HCIS) for specific use cases (2010 to 2020); Phase 3 represents continuous application and micro-services delivery on HCIS platforms (2016 to 2025).

Cloud-native technologies are central to this third wave as developer teams increasingly adopt micro-services to support digital strategies. Kubernetes has won the battle of the container-orchestration platforms in the decades-long evolution of abstracting away the infrastructure. But, as more organizations deploy Kubernetes at scale, they’re finding that the same resource-availability conundrums exist—the artifacts change, but the challenge is the same: how do you ensure pods and nodes get the resources they need when they need them? (You don’t; intelligent software does it.)

Cisco WOM controls the full stack of technologies from infrastructure to service. Today, the platform enables Kubernetes to self-manage, automating actions to drive continuous health:

- Automated rescheduling of pods ensures performance
- Intelligent cluster scaling ensures elastic infrastructure
- Full-stack control unites DevOps and infrastructure

With Cisco WOM, the same decision engine that enables workloads to self-manage on Cisco HyperFlex, also enables Kubernetes to self-manage. As customers continue to leverage a mix of on-premises and cloud technologies, the “expanding IT cosmos” requires a universal approach to enabling self-managing systems. The alternatives are point-solutions that cannot understand the very real interdependencies between systems. Cisco is working on providing customers with self-managing workloads on multicloud estates: Kubernetes operating across Cisco HyperFlex, AWS, Azure, Google Cloud, and beyond.

Achieving self-managing workloads with Cisco HyperFlex and Workload Optimization Manager

The integration of Cisco WOM and Cisco HyperFlex is just one example of how the unification of decision automation and infrastructure automation can deliver self-managing workloads, enabling greater elasticity in the infrastructure and agility among teams. Cisco is working to provide customers with self-managing workloads at every layer of the stack, across multiple data centers and clouds—and beyond.
Cisco Workload Optimization Manager

Trusted decision engine that drives continuous health in on-premises and cloud environments with self-managing workloads

<table>
<thead>
<tr>
<th>AppDynamics®</th>
<th>Cisco UCS</th>
<th>Public cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code-level application monitoring ties performance to customer experience. With Cisco Workload Optimization Manager, the infrastructure self-manages to ensure continuous health of applications and with that, great customer experiences.</td>
<td>Unified computing, networking, storage access, and virtualization, enabling industry-leading infrastructure automation.</td>
<td>Elastic capacity for sudden and/or short-lived demand spikes. With Cisco Workload Optimization Manager, cloud elasticity is intelligently leveraged to ensure performance while minimizing costs.</td>
</tr>
</tbody>
</table>

Cisco Tetrion™

Achieve a secure zero-trust model in an application-centric world with pervasive visibility and behavior-based application insight. With Cisco Workload Optimization Manager, workloads are also intelligently localized, based on communication patterns, to reduce network latency.

Cisco CloudCenter™

Deploy and migrate workloads across on-premises and public clouds. With Cisco Workload Optimization Manager, placement accounts for the resource needs of new and existing workloads to ensure performance of all workloads.

Cisco HyperFlex

Next-generation hyperconvergence with independent scaling engineered on the Cisco UCS trusted platform. The HyperFlex Connect management platform provides simplified management of converged resources. With Cisco Workload Optimization Manager, workload performance is ensured across clusters, while HX capacity is intelligently scaled to meet growing compute or storage needs.

Cisco UCS Manager

With Cisco Workload Optimization Manager, workloads self-manage on Cisco UCS servers to ensure performance while maximizing efficiency.

Cisco UCS Director

With Cisco Workload Optimization Manager, the right intent is passed to Cisco UCS Director to orchestrate actions.

Conclusion

Multicloud is the “new” data center, and apps are the new business. In any business, in every industry, applications are the key driver of success. And every data center or cloud is built to do one thing: run apps. However, apps today are not one thing, and they do not “live” in one place. They operate in multiple public, managed, and private cloud environments.

This complexity requires IT organizations to operate with a vastly different and more demanding set of user expectations than even just a few years ago. The digital revolution is driving increased workload volumes that are highly dynamic, eclipsing the ability of IT departments to deliver cost-efficient quality of service that quickly adapts to changing business needs. To operate at scale, intelligent self-managing workloads will be required. IT leaders that maintain the status quo will soon find themselves obsolete.

In this changing landscape, Cisco Data Center’s vision is focused on helping IT teams to maximize application performance, security, and reliability across the multicloud domain by constantly learning, continuously protecting, and instantly adapting to the ever-changing needs of today’s application-driven businesses. This vision includes an open architecture designed to accommodate the broadest range of customer business needs, maximize application performance, mitigate risk, and increase operational agility. To learn more about how Cisco Data Center can help you, visit https://www.cisco.com/go/datacenter.

Learn more about Cisco HyperFlex:

Learn more about Cisco Workload Optimization Manager: