E-Guide

Server hardware for a scalable virtual infrastructure

Expert advice for making your purchase
IT professionals have new options when it comes to buying server hardware, both for equipment and strategy. Should you go with a scale-up or scale-out approach? Expert Scott Lowe explains why the scale-out model may be the best option in the modern data center context.

Buying server hardware for a scalable virtual infrastructure, Part 1
By Scott Lowe

With developments like server virtualization, cloud computing and big data analytics permeating data centers, IT professionals have new options related to buying server hardware. Data centers need equipment that can meet the performance and availability needs of rapidly growing companies and allow for relatively easy scalability as IT needs change in upcoming years.

Organizations may opt to purchase small numbers of new, powerful servers -- in a scale-up strategy -- that allows a few servers to handle large workloads while consuming less energy. Alternatively, they can choose a scale-out approach that uses large numbers of less-powerful commodity machines, which allow for clustering and redundancy, and this architecture may be less expensive up front.

Both of these server hardware strategies have their place, but today’s need for scalable compute power at a moment’s notice has replaced the traditional scale-up models for server hardware architecture with scale-out models. In what follows, we’ll dissect this shift and how it takes shape in the modern data center context.

The pros and cons of scale-up and scale-out strategies
Over the past decade, scale-up architectures took hold as the strategy of choice as IT leaders began to use certain metrics, like server consolidation ratio and number of virtual hosts, to measure IT performance. Based on these metrics and a desire to save on hardware costs, IT shifted toward ongoing server consolidation by using a few powerful servers that could each take on large workloads, thus maximizing the use of costly resources. The
licensing costs for underlying virtualization software were also reduced, and it became easy to simply add resources when needed due to the decoupled nature of many mainstream applications. Organizations still scaled out, but only as scaling up hit practical limits, such as hitting physical resource maximums in a single-host server.

Today, however, as workloads grow and new needs arise, scale-out architectures are re-emerging and affecting how IT buys server hardware. By harnessing raw compute power rather than divvying up that compute power for discrete workloads, scale-out workloads are solving many of today's most critical challenges. Big data analytics, for example, requires the ability to target data sets with major compute power. That compute power can be acquired by deploying many smaller systems tied together to achieve a common goal. This type of system is also well-suited to cloud environments, to which practically unlimited computing power can be brought to bear. An organization can even consider cloud services as an additional platform in an overall scale-out strategy. In general, cloud vendors provide either large scale-up environments or smaller, discrete scale-out environments, depending on the needs of the customer.

Of course, there are downsides to both architectures as well. Scale-up scenarios rarely provide a linear increase in resources, often favoring one or two resources over others. In a generalized virtual environment that is scaled up to use as few host servers as possible, for example, RAM and disk capacity are often exhausted long before processor capacity is reached, leaving "money on the table" when it comes to resources. Scale-up architectures also require a more detailed approach to availability. On the other hand, a scale-out environment may require new ways of thinking about application design and may not accommodate legacy applications.

With that said, it's important to note that the scale-up-versus-scale-out argument is not a mutually exclusive one. It will be increasingly common to see organizations running their scaled-up environments for legacy and operational needs and their scaled-out environment for research or compute-intensive needs.
Choosing server hardware for virtualization

With different methods for application deployment come different hardware platforms on which to operate those applications. In a predominantly scale-up environment, the capability of underlying hardware plays a much more critical role, while a scale-out environment may be able to leverage the commodity hardware that is emerging on the market.

In the past decade, the virtualization race made the x86 server the go-to platform for just about every organization running mission-critical workloads. The x86 server, in many instances, replaced the legacy mainframe, although aspects of traditional mainframes remain in play today.

For example, while many credit VMware with the creation of virtualization, mainframes have used similar technologies for decades for workload separation. Today's growing environments -- both scale-up and scale-out -- have a lot in common with mainframes, as most environments today are tightly integrated hardware components with master scheduling systems that manage resource allocation. However, it's increasingly rare to see organizations making monolithic mainframe purchases today given the plummeting cost of x86 and commodity hardware and the emerging infrastructure options described later in this section.

IT purchasers know what to expect when it comes to buying x86 servers for scale-up virtualization needs. In short, for pure scale-up virtualization, the ability to expand a single host as much as possible is generally the deciding factor. Doing so keeps down the overall costs of virtualization licensing.

In some instances, depending on the size of the environment, companies may consider massively scalable hardware, such as extremely high-end, densely packed servers that include dozens of processor cores, terabytes of RAM and mass storage. Perhaps the biggest challenge in such a scenario is the potential for workload failure in the event that a single hardware device fails.

A couple of emerging infrastructure options exist that are growing in popularity as organizations strive to rein in the complexity that has befallen
many virtual environments. Both revolve around converged infrastructure, but to different degrees.

The first solution is basically a data center in a rack (or set of racks): Companies from across the virtualization spectrum come together on a prebuilt, pretested hardware platform that is supported by a single vendor. The most recognizable of these solutions is probably the Vblock provided by Cisco Systems, EMC and VMware, but other companies have gotten in on the action, such as Dell with its vStart solution. These infrastructure options enable customers to buy “units of infrastructure” that meet current demands without having to worry about whether certain elements will be compatible with others. These solutions are great from a support perspective and can provide organizations with a lot of peace of mind.

But buying racks at a time isn’t always the best option, particularly for small and medium-sized businesses (SMBs). In fact, smaller organizations may be even more aware of the need to simplify their data center environments, and may have to do so in a more granular way.

This is where a second infrastructure option, hyper-convergences, comes into play. Companies such as Nutanix, Pivot3 and SimpliVity lead in this space. Rather than simply using existing servers and storage, these companies have custom-built units of infrastructure that start at the SMB level and scale to the enterprise level. These individual hardware elements each include compute, RAM and storage resources and often include advanced features in each of these resource categories (e.g., deduplication for storage) intended to maximize their effectiveness. The units of infrastructure are extremely powerful because of their granularity but also because each element can provide a massive amount of resources because of the advanced hardware that’s often included.
Buying server hardware for a scalable virtual infrastructure, part 2
By Scott Lowe

As explained in part 1 of this series, IT admins buying server hardware have complex decisions to make. There's no question that today's virtualized environments need to be scalable, so deciding how to grow an organization's data center operations is extremely important. But the choice can be made with a case-by-case approach, based on different types of workloads. As you design a scalable architecture, here are some additional elements to consider.

Application architecture
In a scale-out environment, there are fewer resources per node compared with a scale-up environment. As such, scale-out environments are generally suitable for workloads that are designed to harness the compute resources across hardware boundaries and leverage these resources as a cohesive whole. Consider a major analytics application, for example. It requires far more resources than could be harnessed in a single system. In a scale-out environment, this application could wrap its arms around all the hardware resources dedicated to it. Your methodology may require new ways of thinking about application design and limits the potential for legacy applications to be moved into scale-out environments. In addition, deploying hypervisor software across a scale-out architecture may be cost-prohibitive from a licensing perspective.

Workload balancing
Workload balancing may be the most important factor in the decision to scale up or scale out. Again, you may want to implement both approaches if needs dictate.

SharePoint, for example, might be considered a legacy tool for some admins, but it provides a good opportunity to explore when to scale up and when to scale out. In an initial SharePoint implementation, you can take either approach. For instance, you may choose the single-server route and simply add resources to that server as needed. Or you may deploy a scale-out
SharePoint environment in which each separate service runs on its own system. Under this scenario, as needs dictate, you would simply add resources to those separate servers or, for additional availability, add servers to support critical roles. In the single-server example, you have the option to add servers in the future, but not in as granular a way.

When it makes sense, I prefer to start with a scale-out environment for such applications, which can be started inside your existing virtualized infrastructure. As needs grow, add resources or move workloads to different hosts or move them to their own physical, scaled-out environment.

**Power-saving options**

In addition to choices surrounding buying server hardware, you should also consider features that can save money. Modern servers are much more power-efficient than older units. With more efficient power supplies and core hardware, today’s servers run cooler and require less power to operate. When coupled with appropriate management software, power efficiency in today’s server environment can be taken to new levels.

Both VMware and Microsoft have power optimization technologies in their respective virtualization management tool that allows the management software to monitor power levels in managed clusters and to shut down hosts when resource requirements are low, and then return hosts to operation as resource needs increase.

With these criteria in mind, how do you go about balancing scale-up and scale-out architectures? For general line-of-business and productivity applications, it’s about balance. Scale out sufficiently to ensure that all mission-critical systems have enough hosts on which to operate and that you’ve accounted for any overhead associated with automated high-availability mechanisms and workload separation. Scale up to ensure that there is enough horsepower to handle these workloads as they’re distributed across the environment.
For scientific, big data or other high-performance computing needs, it's about raw power. Scale out and aggregate the workload using equipment that provides enough processing threads to meet the needs of the application.

A blended architecture that employs both high-end and commodity hardware may be a good choice for organizations looking to scale workloads effectively. The data center of tomorrow will be a combination of scaled-up and scaled-out environments, which will completely depend on workload type.
Free resources for technology professionals
TechTarget publishes targeted technology media that address your need for information and resources for researching products, developing strategy and making cost-effective purchase decisions. Our network of technology-specific Web sites gives you access to industry experts, independent content and analysis and the Web’s largest library of vendor-provided white papers, webcasts, podcasts, videos, virtual trade shows, research reports and more — drawing on the rich R&D resources of technology providers to address market trends, challenges and solutions. Our live events and virtual seminars give you access to vendor neutral, expert commentary and advice on the issues and challenges you face daily. Our social community IT Knowledge Exchange allows you to share real world information in real time with peers and experts.

What makes TechTarget unique?
TechTarget is squarely focused on the enterprise IT space. Our team of editors and network of industry experts provide the richest, most relevant content to IT professionals and management. We leverage the immediacy of the Web, the networking and face-to-face opportunities of events and virtual events, and the ability to interact with peers—all to create compelling and actionable information for enterprise IT professionals across all industries and markets.

Related TechTarget Websites

➢ SearchServerVirtualization
➢ SearchDataCenter
➢ SearchVirtualDesktop
➢ SearchVMware