Juku.it

Down to earth report: A hyper-converged infrastructure for the rest of us from HP.

Enrico Signoretti

November 2013

Table of contents

Executive summary

End users of all kinds are continuously looking at solutions aimed to reducing the complexity and to drive down the total cost of ownership (TCO) of their IT infrastructures. Especially in the SMB space, where resources are usually limited and skills aren't often developed at best, there is a lot of attention to this issue.

Nowadays most infrastructures are virtualized, and they are based on x86 commodity servers with a standard ethernet networking. The highest costs, both in terms of TCA (Total Cost of Acquisition) and TCO (Total Cost of Ownership), usually come from data storage. In fact, it's easy to find that about 50% of the cost of a virtualized infrastructure relies on an ordinary SAN/NAS infrastructure.

This paper will show a solution, completely designed on industry standard components and HP software, that allows the realization of a complete scale-out virtualized infrastructure without needing traditional shared storage. In a few words: a hyper-converged infrastructure. The converged solution that we are going to describe has all the characteristics that you could expect to address: TCO, ease of use, scalability, efficiency, it's also open and with a low price. In fact, the big advantage comes from the virtualization of the internal disks on the same x86 servers that run the hypervisor, presenting a virtual SAN to the VMs.

Another big advantage is that the whole stack comes from a single vendor, enabling an easier way to get access to an end-to-end support service.

The scale-out model presented in the following pages could have a great impact to the user and, potentially, for the reseller of the solution. In fact, each node (a brick) has a precise amount of resources (CPU, RAM, disk space/IOPS) and a well-known acquisition cost too. Adding a precise amount of VMs to the infrastructure is only a matter of buying one or more bricks and connecting them to the existing infrastructure.

The primary objectives of this paper are to present a reference design, its potential scalability and a standard benchmark score showing its positioning.

Introduction

What is a scale-out infrastructure?

Scalability is the capability of an IT system to be enlarged to sustain a growing amount of work that needs to be managed.

There are two types of scalability: vertical (scale-up) and horizontal (scale-out). The main difference is that scale-up systems are monolithic while the others are composed of small nodes connected together. In practice when you talk about vertical scalability, you talk about the ability of the system to add more resources in the same box (e.g.: adding CPUs, RAM, disks, etc. in the same computer). On the other hand, the expansion of a scale-out system/infrastructure occurs by adding more nodes, with each node adding its own resources to the cluster.

Both approaches have their advantages and tradeoffs but, in the past few years, technology has made many steps further and scale-out systems are now chosen for a wider range of applications. Potential connection latencies between nodes and management complexity are no longer a big problem while it's relatively easy to deploy huge computing systems at reasonable prices.

Scale-out is also becoming more popular in the storage industry to solve performance and space problems when the numbers are huge: BigData Hadoop/HDFS clusters are the most visible examples.

Why is scale-out important to SMB?

SMB doesn't have the same huge needs described above but the concept of scalability is also important. A scale-out infrastructure allows the SMB

enterprise to think only about current needs and invest as less as possible at the beginning while avoiding the risk of forklift upgrades in the future.

Another big advantage for SMB is that a scale-out infrastructure has a more predictable cost and power when expanded: in fact every new node added to the infrastructure brings a precise amount of resources and the ability to carry out a known quantity of work.

Why it is important for you

HP has all the components (servers, networking and storage software) to build a real hyper-converged scale-out system. The unquestionable enabler of this solution is the software component that allows you to realize a true software-defined storage system on top of commodity hardware, and all produced by a primary vendor.

Moreover, this “software-defined” approach allows a lower TCO and faster ROI, if compared to traditional solutions with shared storage: the end user has all the benefits of a full-featured next generation storage system without its “hardware” rigidity.

Last but not least, the end user could expand the cluster at very low and predictable prices (the cost of a server, its local disks and the VSA license) or swap the server with a faster model while maintaining the same licenses.

On the contrary to similar solutions out there this one has some big advantages:

- based on 100% industry standard hardware;
- compatible with the most common hypervisors (VMware ESXi and Microsoft Hyper-V);
- end-to-end hardware and software solution from a single primary vendor;
- cheaper than many other solutions and perfectly suitable for SMB;
- worldwide and well proven support services;

In the following pages I’m going to discuss all the standard industry components that are ready, off-the-shelf.

Hardware

HP ProLiant DL380p Gen8 Server

HP ProLiant DL380p Gen8 is a 2U, 2-socket datacenter rack server. It's part of the latest family of HP x86 servers and has all the reliability, accessibility, and serviceability characteristics that distinguish the whole line-up. The basic technical specifications of this server are:

- up to 2 Intel Xeon E5-2600v2 series CPUs;
- up to 768GB of RAM;
- 4 1Gb, or 2 10G Ethernet, or 2 x 10G Flexfabric;
- Different internal storage configurations (up to 25 2,5" SAS disks or 12 3,5" disks, SSD support and internal RAID options);
- 6 PCIe 3.0;
- integrated iLO management engine.

We have chosen this server for our report because it has the right characteristics of internal storage (up to 25 2,5" 10K RPM disks) and external connectivity that we need. It also has PCIe slots to accommodate Flash cards or additional NICs.

To maximize the benefit of the internal SAS JBOD we have also decided to add an HP Smart Array Controller model 420 (equipped with 1GB of cache). The controller will help to optimize IO operations by performing basic RAID calculations and caching.

You can find more info about HP ProLiant DL380p Gen8 server on HP website.

HP 2920 Switch Series

For this solution we have chosen to adopt the HP 2920-24G switch, one of which was suggested by HP StoreVirtual best practices documentation.

This is a high-density 1GbE, top-of-rack, cost effective and scalable switch with the ability to support up to four 10Gbit Ethernet ports, as well as two stacking modules.

It also supports jumbo frames (up to 9220 bytes), IPv4, IPv6, VLANs and layer 3 routing, QoS and OpenFlow protocol: all important characteristics to avoid bottlenecks while granting the maximum flexibility in terms of configurations and future expansions. Power supply is removable and easily upgradable. Lifetime support and free software upgrades make this switch particularly attractive to price sensitive users.

You can find more info about HP 2920 switches on HP website.

Why it is important for you

Powerful and well-built affordable hardware is the driver for the success of this kind of deployment, and HP is the world's most important server manufacturer providing support and services in most countries. The importance to deal with reliable primary vendors is a key factor for SMBs who need to be reassured about the presence of local dealers and system integrators even for basic implementations and technical services.

On the other hand, networking is fundamental to grant the performance and reliability needed by server and storage virtualization. HP 2920 switches are certified end-to-end with all the other components of the infrastructure, providing great throughput and next generation features aimed at optimizing the quality of storage and VM traffic. Even more, these particular switches provide up to four 10Gbit Ethernet ports at a very affordable price, enabling the deployment of a simplified and powerful networking backend.

Software

HP StoreVirtual storage VSA

HP StoreVirtual Storage is a family of affordable storage solutions designed to solve SMB needs. The product portfolio shows two different kinds of products: HP StoreVirtual 4000 (hardware products based on HP industry standard x86 servers) and a HP StoreVirtual VSA (Virtual Storage Appliance). software The features of the two products are identical.

HP StoreVirtual 4000 is a complete scale-out iSCSI storage array with all the features that you would expect from a modern storage solution (for example, thin provisioning, snapshots, replication, VMware integration and so on). Due to the nature of the HP StoreVirtual architecture, data on each node of the cluster are striped and replicated on other nodes with a mechanism called Network RAID. This approach allows you to obtain a very highly scalable and resilient environment with off-the-shelf hardware. Adding additional space and performance to the cluster is very easy and new nodes can join old er generation hardware clusters with a few limitations. Automatic restriping enables an immediate use of the new added resources.

The VSA (Virtual Storage Appliance) version of HP StoreVirtual allows to use a VM as a virtual storage controller and it takes advantage of local server disks to provide IOPS and space.

VSA has seen many improvements in the last two years. Recent versions of StoreVirtual's LeftHand operating system (v. 10) were partially rewritten with multithreading capabilities, and now many IO operations are quicker than in the previous versions. VSA has also taken advantage of this new software, in fact the VSA is now configured with 2 virtual CPUs showing huge improvements in terms of performance and scalability.

LeftHand OS version 11, the latest one, also introduced Adaptive Optimization. This is an automatic tiering capability operating at the sub-

volume level that allows mixing of flash memory and traditional spinning hard drives in the same configuration. Adaptive Optimization could dramatically improve storage performance, making the whole cluster more responsive and capable of taking full advantage of all its computational resources.

HP StoreVirtual VSA is licensed on a capacity basis, this means that the end user, using just internal server disk slots, could configure an important amount of storage space on every server host at a relatively low cost. At the moment, there are three different types of capacity licenses, however each one has a capacity limit to consider:

- 4TB: this is the cheapest tier but limited to 3-node installations and does not include the Adaptive Optimization feature.
- 10TB: this is the most common license, this one has no limits in the number of nodes in the cluster and can be useful for the vast majority of the installations;
- 50TB: this is a recently introduced license for space sensitive installations and, also in this case, there are no limits in the number of nodes that can form the cluster.

HP provides upgrade paths for all the licenses: an end user, for example, can buy a 10TB license today and pay the difference to upgrade to a 50TB license in the future, if needed.

Third party hypervisors and benchmarking suite

HP StoreVirtual VSA supports both Microsoft Hyper-V and VMware ESXi Hypervisor.

For our lab tests we opted for VMware because we can use the VMware VMmark suite to run a standard benchmark on the system. We offer the result of this benchmark only for evaluation purposes and it was not submitted to the official VMware reviewing process. On the other hand, this

document reports all the necessary information to reproduce the same results and officially go through the certification process.

We won't use VMmark to show you a sterile number, as that is useless for SMB end users. Each VMmark tile (a tile is a complete set of VMs) reproduces standard application stacks and workloads of various sizes and complexity:

- 1 Mail server: Exchange Server 2007 on Windows 2008 R2 Enterprise Edition,
- 1 Web server for Social Networking application (Olio): SLES11 64 bit,
- 1 Database server for Social Networking application (Olio): MySQL DBs on SLES11 64 bit,
- 3 Web servers for eCommerce application (DVD Store 2): SLES11 64 bit,
- 1 Data base server for eCommerce application (DVD Store 2): SLES11 64 bit,
- 1 Stand-by server: Windows 2003 server without running apps,
- 1 Windows server that is deployed and retired during the test.

While the test is running, some external clients launch a workload generator for each one of the mentioned application stacks. In particular, the Exchange Server machine has a simulated workload of 1000 (heavy profile) users. We know that this is huge for a SMB company, especially because this workload is replicated every time we add a node. On the flip side, Exchange server is a complex application with a DB underneath and it could be assimilated to many heavy loaded applications like, for example, ERPs.

The final goal is to show how many of these machines can run in each node of the cluster to have a real world scenario of what you can expect in your real life environment.

For our benchmarks we used a basic two-node cluster and then we added a third and a fourth node to show the first steps towards larger scale-out configurations.

Due to VMware licensing model, we have performed all the benchmarks with the standard version of ESXi. This version is the most suitable for small companies while granting a quite interesting set of features.

In any case, our goal isn't to give you a VMmark scorecard but to answer two simple and uncomfortable questions. The kind of questions that many small end user often ask their reseller:

- How many virtual machines can run on a node?
- How many nodes do I need to run all my VMs?

Answering these questions is very hard and usually needs a deep knowledge of the infrastructure, the involved workloads and technology. On the contrary, with this kind of pre-tested infrastructure, we are attempting a completely different approach: We tested it with a workload that can match most of the use cases, giving us an average per-node number of VMs that each node can run.

Why it is important for you

The idea behind this white paper is to show the capabilities of an easy to implement LEGO®-like architecture. The enabler of this architecture is the software and HP StoreVirtual VSA perfectly fits in this picture. This software-defined storage solution has all the features of the hardware model but it's cheaper and allows flexible clustered node configurations and transparent hardware upgrades.

Last but not least, availability of HP StoreVirtual VSA on the two most deployed hypervisors of the market grants a great freedom of choice for the end users.

This approach is very suitable for smaller end users where skills to design and deploy an efficient virtualized infrastructure are often absent. Also, resellers can take full advantage of this kind of solution thanks to the tremendous simplification of the sales process.

Architecture

Infrastructure node configuration

For the actual implementation of our infrastructure we have decided to configure medium-sized nodes in terms of CPU, RAM and disks. Our goal is not to push for the fastest possible configuration but, on the contrary, to prove the value of a truly affordable hyper-converged solution for small environments.

At the same time, we have chosen 10GBit Ethernet ports and switches to simplify the configuration, avoiding teaming (link aggregations) and to enable an easy path for future upgrades. 10Gbit/s Ethernet for such a small infrastructure could sound strange but, on the contrary, it's perfect because HP's networking hardware that we have chosen (2920) is very cost effective and also gives the opportunity to make use of 24 more 1Gbit/s Ethernet ports on each switch.

The server node of our scale-out infrastructure will be configured as follows:

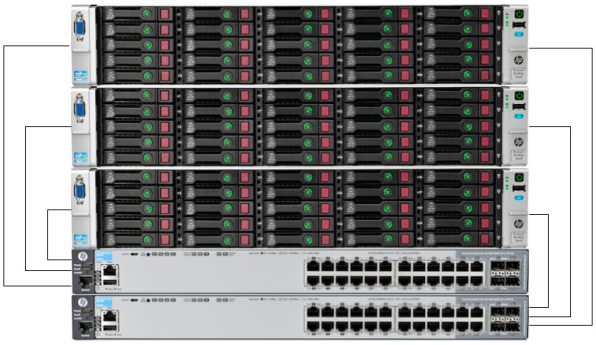
- 2 Intel Xeon E5-2640 (6c/2.5GHz) CPUs;
- 64GB of RAM;
- 2 10G Ethernet ports;
- 18 2,5" 450GB/10K RPM SAS disks;
- HP Smart Array Controller model 420 w/ 1GB of cache.

This configuration, in a basic cluster of three nodes, will give 36 CPU cores, 192GB of RAM and 45 active data disks backed by 3GB of cache (which theoretically means more than 9TB of free space and a good quantity of IOPS, even in the worst scenario): we will be discussing the measured performance of the cluster in the following chapters.

On the networking side, the configuration that we are going to use in our test is based on two HP 2920-24G Switches configured with 4 active 10Gbit ports each. The 10GBit ports are dedicated to connect the nodes of the cluster while the 24 1Gbit ports are free and available to connect other servers and clients.

Cluster and nodes layout

The layout of the cluster is very simple: each node has a network connection to every network switch. The two switches can be used as core switches (in smaller environments) or to provide the adequate number of up-links to an infrastructure already in place.



The node configuration is very simple too and it is particularly focused on the internal storage. Following HP's best practices documentation, HP SmartArray controller can be configured with write-back cache because, in our case, the Network RAID mechanism implemented into the VSAs prevents potential data losses. Internal server disks (18) are configured with 1 spare disk, a small RAID1 volume made of two disks for booting the hypervisor and a RAID50 group made of 15 disks for the VMware Datastores. There is also room for future disk expansions and SSD options.

Why it is important for you

Simplicity and manageability are the most important characteristics of this implementation. Our goal was to keep it as simple as possible and follow all the standard procedures and best practices available on HP websites. We did it to give the end user, and the reseller, a solution that can be easily reproduced without the need of a particularly skilled engineer.

Avoiding complexity while implementing a 100% scale-out infrastructure also reduces management costs, even if we are talking about a few nodes.

On the storage side, HP StoreVirtual VSA perfectly suits this picture: an easy to use, solid and mature product that was designed from the ground up with this type of architecture in mind and with all the tools needed for a seamless integration with the hypervisor.

Last but not least, redundant 10Gbit Ethernet links allow you to have a very high bandwidth and ease of management at the hypervisor level without needing complex configurations.

Benchmark results

Introduction

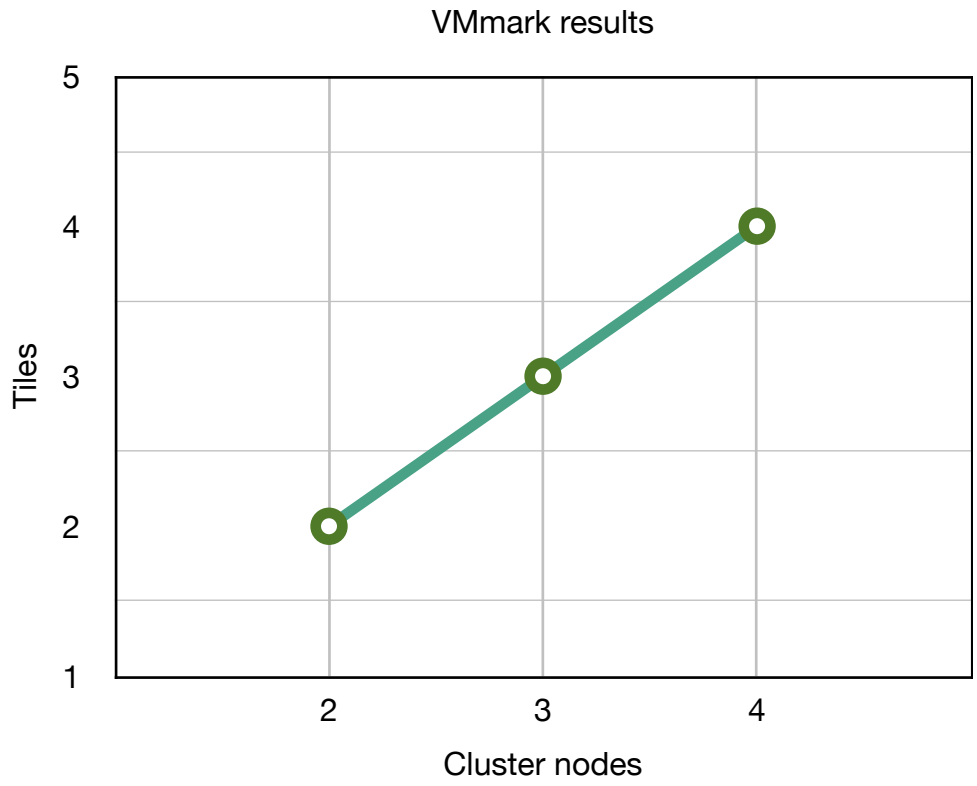
The objective of this paper is to provide the simplest metrics to evaluate this kind of solution. To achieve this goal we ran the VMmark benchmark and we collected all the information for each different configuration of the cluster.

These results show two important things: the number of VMs that the cluster can run simultaneously and its scalability. At the same time, as this is a real life configuration, we never pushed tests and configurations to the limit of the hardware. Our choice was to leave enough room to allow the end user to relocate VMs and have decent performance in case of a node failure.

The only issue that we have found with the VMmark suite is that tiles can't be split on different nodes, a tile can run only on a single node at a time. This means that for our small configuration there are some wasted resources but, as mentioned earlier, this is a real life configuration and we need to have some available resources to sustain the infrastructure in case of a complete fault of a node.

Benchmark results

The following chart shows the graph with the number of virtual machines that this cluster is capable of running, as expressed in "number of tiles". You will immediately discover the linear progression of the cluster. Each node added to the configuration allows it to scale the same number of VMs. CPU and RAM are never a problem: the limit is always found on the disk side. In the case of the 4 node configuration we take full advantage of a higher number of disks and there is some space to run more virtual machines but not enough to run a whole tile.



Why it is important for you

The chart presented here gives a clear idea on how many VMs can be spun up on each tested configuration.

Scalability, as predicted at the beginning of this paper, is totally linear and adding more resources has an immediate positive impact on performance.

End users can easily figure out the number of needed cluster nodes to build a new infrastructure by looking at the number of running VMs already in place. Bigger configurations have also the advantage to give a better overall resource utilization, and this is very good to know for future system expansions.

At the time of this publication, the street price of a 3-node cluster like the one used is probably below €25.000 (price are very susceptible to changes and it doesn't consider VSA, hypervisor and Ethernet switches costs). HP is also offering a free 1TB VSA license on the purchase of selected servers that can be considered a good starting point for ROBO and very small installations (more information can be found on HP Website). Even if we add the price of the VSA license (it depends on the amount of disk space required), this is an exceptionally low price. For an entry level solution it could

be easily positioned well under €30.000. In many contexts this price can easily be compared to an entry level configuration of the sole SAN equipment without servers!

Cost savings can also be found by looking at the space occupied in the rack and, especially when compared to traditional FC SANs, in the overall minor complexity of the whole infrastructure.

Bottom line

Words like hyper-converged and software-defined are often abused and they can easily lose their significance but, in this case, I'm pretty confident that we have realized what we were supposed to do: a very entry level software-defined and Hyper-converged solution for the rest of us. Another important key point here is that the whole hardware stack (Server, Storage and Networking) is made with standard off-the-shelf components made by a primary vendor, this has direct consequences on how support services are delivered and automatically leads to a potentially better support experience for the end user.

This is the kind of solution that could help small IT departments to cover all their needs or it can be viewed as a very price sensitive and easy to deploy/manage solution for remote and branch offices in larger organizations.

Furthermore, the benefits of this solution for ROBO is not to be undervalued in many other aspects: for example, data replication capabilities offered by HP StoreVirtual could be used to easily implement DR plans while other products like HP StoreOnce VSA could be the perfect fit to manage remote backups and vault them to a primary site.

With the recently introduced HP StoreVirtual version 11, it's now possible to easily improve performance of this hardware stack by adding a small amount of flash memory on each node while maintaining the price at a reasonable level.

The solutions described in the latter pages fit very well in small and 100% virtualized environments leaving vast possibilities for expansion and improvements if the enterprise grows.

Many software and hardware vendors are working on similar solutions but, at the moment, HP is very well positioned and its offer can start at the very entry level.

Juku

Why Juku

Jukus are Japanese specialized cram schools and our philosophy is the same. Not to replace the traditional information channels, but to help those who make decisions for their IT environments, to inform and discuss the technological side that we know better: IT infrastructure virtualization, cloud computing and storage.

Unlike the past, today those who live in IT should look around themselves: things are changing rapidly and there is the need to stay informed, learn quickly and to support important decisions, but how? Through our support, our ideas, the result of our daily interaction that we have globally on the web and social networking with vendors, analysts, bloggers, journalists and consultants. But our work doesn't stop there, the comparison and the search is global, but the sharing and application of our ideas must be local and that is where our daily experience, with companies rooted in local areas, becomes essential to provide a sincere and helpful vision. That's why we have chosen: "think global, act local" as a payoff for Juku.

Author

Enrico Signoretti, consultant, trusted advisor and passionate blogger (not necessarily in that order). Having immersed into IT environments for over 20 years, his career began with Assembler in the second half of the 80′s before moving on to UNIX platforms (but always with the Mac at heart) until now when he joined the "Cloudland". During these years his job has changed from deep technical roles to management and customer relationship management. In 2012 he founded Juku consulting SRL, a new consultancy and advisory firm highly focused on supporting end users, vendors and third parties in the development of their IT infrastructure strategies. He is constantly keeping an eye on how market evolves and continuously looking for new ideas and innovative solutions. You can find Enrico's social profiles here: http://about.me/esignoretti

4AA5-0132ENW