Proposal of Block storage for Agena's hypervisor stack

Revision history

Revision	Author	Date	Description
0.1	Jyoti Ranjan	05-July-2019	Conceptualized the content.
0.2	Jyoti Ranjan	08-July-2019	Detailed out requirements, goal and author's view.
0.5	Jyoti Ranjan	08-July-2019	Elaborate feature comparison of different storage system.
0.6	Jyoti Ranjan	09-July-2019	Detailed evaluation criteria and drafted point based evaluation system
0.7	Jyoti Ranjan	09-July-2019	Drafted ranking system
0.8	Jyoti Ranjan	10-July-2019	Updated the content with some of feature details with more accuracy
0.9	Jyoti Ranjan	11-July-2019	Re-calibrating the points

Introduction

Overview

Agena's VMaaS platform aims to support virtual instances provisioning as cloud model which is ubiquitous because of its agility, pay as we go and robustness. Every instance needs boot disk and persistent disks (optional) which implicitly translates to requirement of need to support block storage for VMaaS.

Purpose

The document aims to achieve the following: (1) Assimilate block storage requirement, (2) Evaluate choices and (3) Recommend the best fit storage solution. As we know that different hypervisor exposes volume to instances differently even if the mechanism to present the disks to hypervisor host is same. This holds more true for ESXi hypervisor. So, the reader is expected to consider this point in view.

Critical aspects

Considering the variability of many parameters like compute instances, storage system, hypervisor etc, it is important to reflect common understanding of some aspects which is captured below.

What are hypervisor being considered while choosing storage system?

The following hypervisors are being considered:

- ESXi
- HyperV
- KVM

As we know that ESXi overlay the VMFS through the concepts of data store which is unlike KVM and HyperV. So, the extra abstraction does not allow direct presentation of disks presented to host as it is to virtual instance(s) unless and until we use VVOL which is not so popular in user fraternity. As of know the VMaaS is going to support only ESXi. So, the document is focused towards to identifying best fit bock storage system for ESXi. Reader is expected to assume the contents are specific to ESXi unless and until stated otherwise.

What are storage system being considered or evaluated?

As per guideline given, the plan is to use one or combination of the following storage system:

- VMware vSAN
- HPE SimpliVity
- HPE Nimble
- HPE Primera

How do we address scale aspects for storage systems?

The scalability of storage system depends upon two variables:

- Usage of hyper converged system. In this case, horizontal scalability is dependent upon the number of hosts present in the cluster and vertical scalability is dependent upon number of disks used per hosts.
- Usage of external storage system. In this case, there is no horizontal scalability and vertical scalability is limited by the specific models.

Why can not we use same storage system that serves different hypervisors (KVM, HyperV, ESXi) or container?

Technically yes. We can do. We can very well pick any storage array and present volumes to VM or container. But, it will not be wise to use same storage system for all use cases. For e.g. ESXi is very much different than rest of hypervisor as far as consumption of storage is concerned. ESXi overlays VMFS over data store created out of disks presented to host. The disks can be presented from any iSCSI or FC storage devices like HPE Primera, HPE 3PAR, HPE Nimble. But, the overlay of VMFS dilutes the native differentiated storage capabilities unless and until data store is created intelligently. Even if we create different data store of different capabilities intelligently, the manageability becomes an issue especially for the cloud use cases where user pattern can not be predicted with accuracy. The same does not hold true for traditional data center model where user workload pattern could have been predicted with a good degree of accuracy. On the similar line, the expectation of agility is more in case of cloud. So, the usage of fixed ration of storage to compute is not a good choice and hence HPE Primera or HPE Nimble lack this piece especially in the case of ESXi. For KVM and HyperV, it might not be so strong negative.

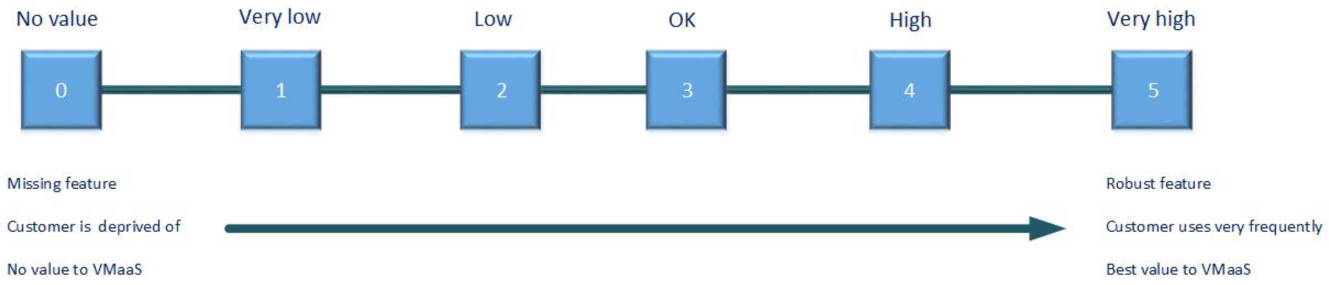
Requirements, Goal and Author's view

ID	Requirement (s)	Goal	Author's view
1	Ability to support multiple supervisory	Leverage same storage system for other hypervisor (e.g. KVM, Hyperv) as and when we support it.	Yes, I agree that it is good to have a same storage system serving block storage requirement for all hyper-visors (e.g. ESXi, KVM, HyperV) so that leverage of design, experience and usability can be extended. However, ESXi falls in different bucket because of it's overlay strategy of creating VMFS over native volume presented to ESXi host. The overlay of VMFS results in dilution of storage capabilities provided by storage system if data stores are not designed properly. So, providing a differentiated storage offering requires some careful and thoughtful design. VVOL has been proposed way of connecting volume to instances directly but it has not been so popular though adopted by some storage vendors. Usability of VVOL in community is very limited and not widely accepted even after its release of multiple years. In a nutshell, the strategy adopted for ESXi does not fit to other hypervisors (KVM and HyperV) as it is.
2	Performance requirement	Ability to support differentiate storage offering.	Because of the the reason of overlaying VMFS over native volume presented to ESXi host, offering a differentiated storage requires careful and thoughtful planning of ESXi data store. In a nutshell, the strategy adopted for ESXi does not fit to other hypervisors as it is.
3	iSCSI support	Instances should be able to connect to volume using iSCSI protocol.	Every volume is presented to ESXi instance as SCSI (not iSCSI). But this is not the case for instances running on KVM or HyperV. As far as presenting native volume to host machine (i.e. ESXi or KVM, or HyperV) is concerned, it does not hold deep value for block storage offering to instances. Yes, it can be critical criteria if we want to boot host using iSCSI volumes but I do not think that requirement is of that here. In a nutshell, the focus of proposal is on presenting volumes to instances. So, iSCSI support in case of ESXi instances is not a true requirement.
4	Support of enterprise feature	Need supportability of features like replication, de-duplication, compression etc.	Yes, the proposal evaluates the advanced enterprise feature. As mentioned above, the ability to pass those features to user requires careful and thoughtful datastore designing.

5	Ability to Scale	Ability to scale horizontally, vertically or segmented.	There are 3 types of scaling: Horizontal, Segmented and Vertical scaling. As we know that hypervisor cluster is more towards horizontal and segmented scaling even if storage system offers one or any combination of horizontal, segmented and vertical scaling. So, the proposal keeps this aspect in mind and clarifies very explicitly where we need to apply which type of scaling based on storage system. In a nutshell, we can not see storage scaling separate from hypervisor cluster in most (if not all) cases.

Evaluation criteria

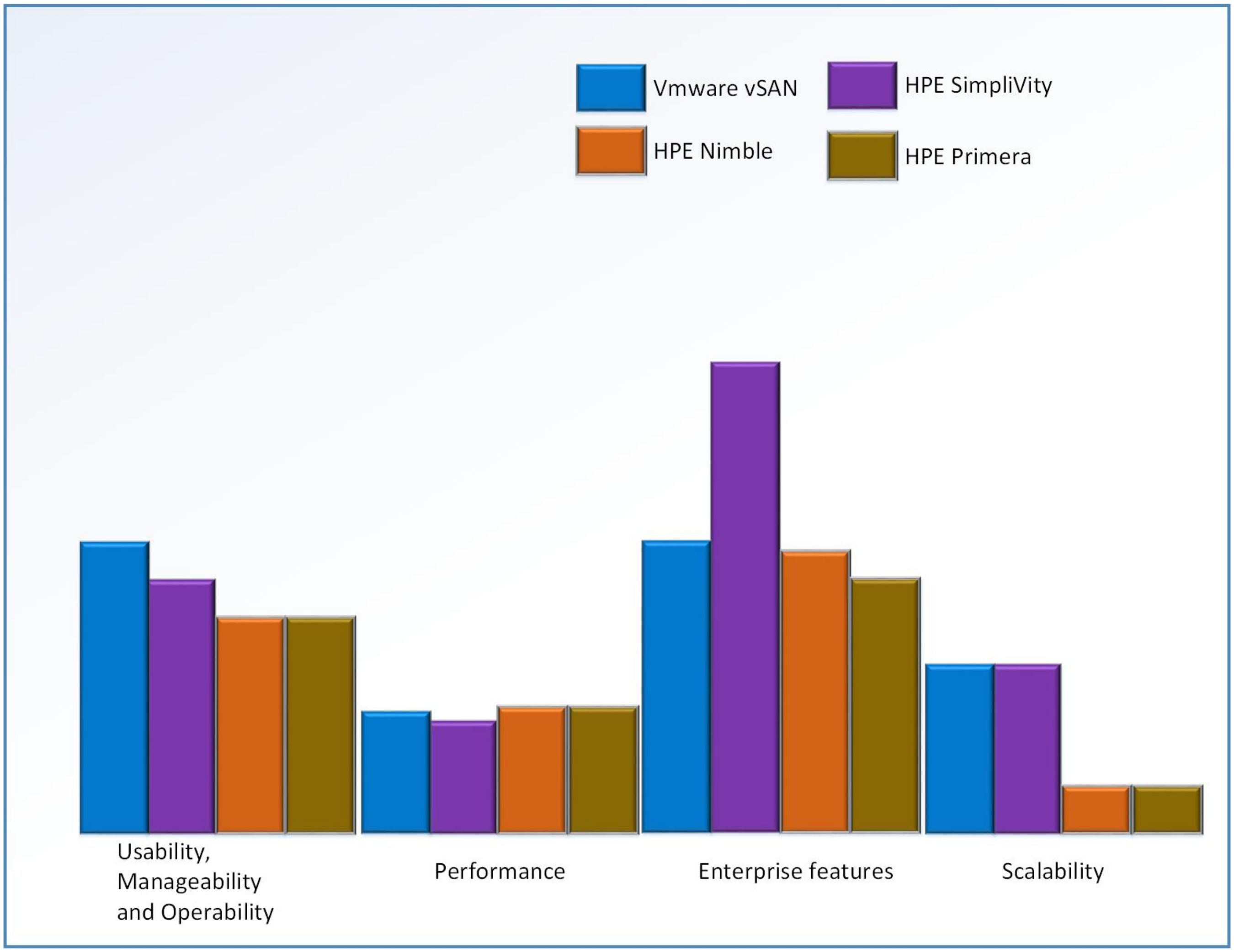
As the considered storage system seems to resemble in features, there is need to drive a objective evaluation system to determine the perfect fit for VMaaS. In this section, the methodology assigns a point in a scale of 1 (low) to 3 (high).

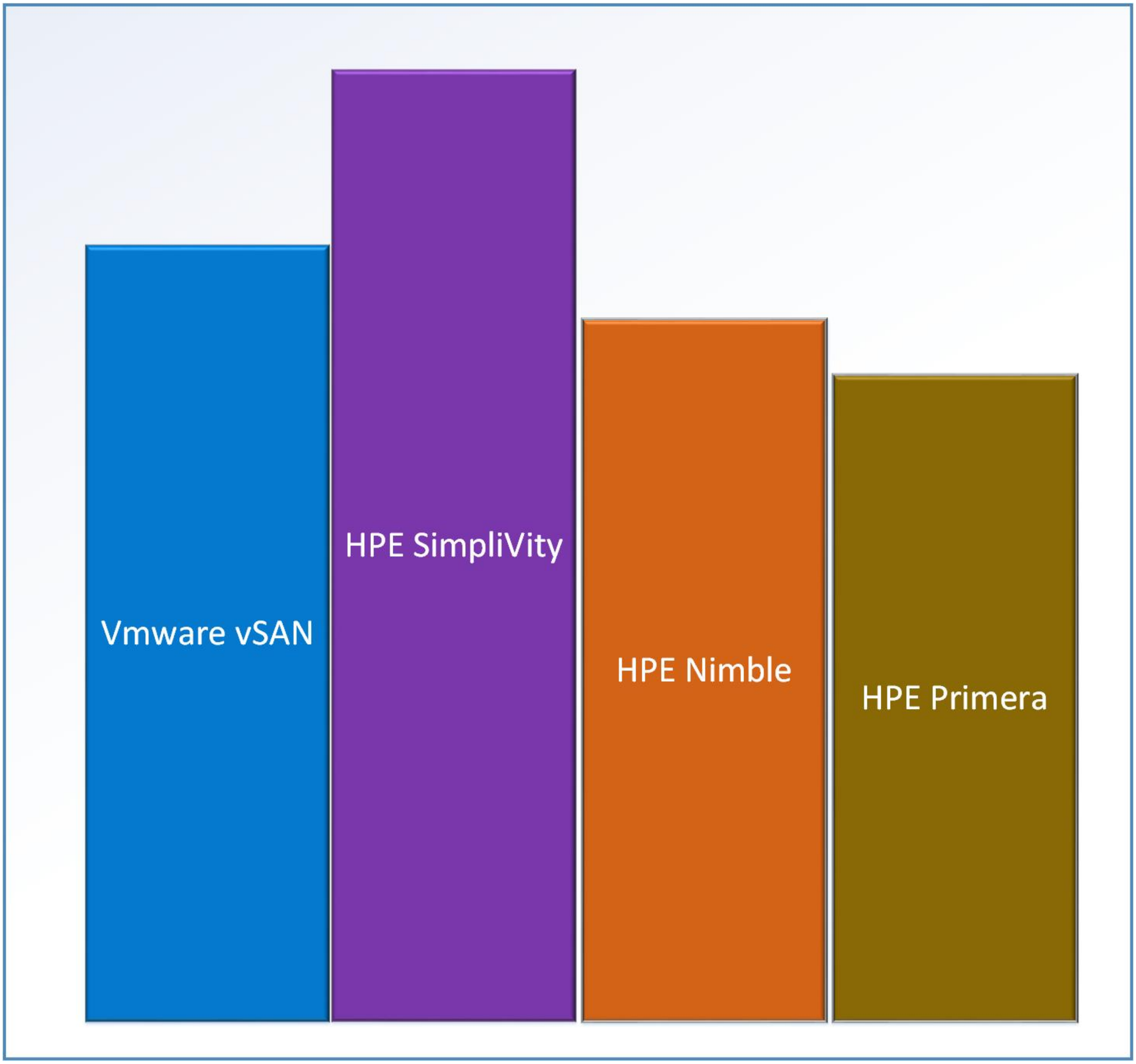


Area	Criteria	vSAN	HPE SimpliVity	HPE Nimble	HPE Primera
Usability, manageability and operability	Leverage as converged or hyper-converged solution	5	5	3	3
	Easy to manage	4	5	5	5
	Initial cost (low cost higher points)	5	5	2	2
	Incremental horizontal scaling cost	5	4	2	2
	Maximum RAW capacity	5	4	5	4
	Extracting value add from HPE hardware with ability to maneuver ratio of storage to compute	5	4	5	5
	Sub-total	**29**	**27**	**22**	**21**
Performance	IOPS	3	4	5	5
	Ability to support differentiated storage based on feature sets (no usage of VVOL)	3	3	4	4
	Ability to support IOPS throttling	5	5	5	5
	Sub-total	**13**	**12**	**14**	**14**
Enterprise feature	De-duplication	4	5	5	5
	Compression	4	5	5	5
	Fault tolerance	5	5	5	5
	Advanced VM based backup	3	5	0	0
	Advanced VM based snapshot	3	5	0	0
	Advanced VM based clone	0	5	0	0
	Advanced VM based DR	3	5	1	1
	Efficiency in data optimization	3	5	5	5
	Availability	4	4	5	5
	Public cloud compatibility	0	0	5	0
	Sub-total	**29**	**44**	**31**	**26**
Scalability	Easy to tweak compute to storage ratio in horizontally (a very critical criteria in case of cloud)	5	4	1	1
	Granular horizontal scalability of standalone storage system	5	5	0	0

	Granular vertical scalability of standalone storage system	3	3	5	5
	Granular segmented scalability of standalone storage system	5	5	1	1
	Sub-total	**18**	**17**	**7**	**7**
	Total	**87**	**105**	**76**	**71**

The above result is represented pictorially below.





Block storage feature(s) comparison

Feature	vSAN	HPE SimpliVity	HPE Nimble	HPE Primera
Hardware requirement	• Array of ESXi hosts with disks to form a vSAN cluster	• Array of OmniStack hosts with disks to form a cluster	• Independent storage controller with bunch of rack based disks.	• Independent storage controller with bunch of rack based disks

Specific host hardware requirement	• None	• PCIe Accelerator card per host	• None	• None
Network requirement	• 1 dedicated NIC for storage data path. • All other networks for other purposes like vMotion, vCenter management etc are not assumed here.	• 3 NIC to separate management, federation and storage traffic	• Separate traffic for control and data path.	• Separate traffic for control and data path.
Disk storage requirement	• One SSD disk per disk group for a given host. Remaining disks can be non-SSD disks.	• All SSD disks	• Pre-configured model. Number of disks can be changed to vary vertical scale limit.	• Pre-configured model. Number of disks can be changed to vary vertical scale limit.
Models	• Hybrid vSAN • All flash vSAN	• All flash	• Adaptive flash storage array • All flash storage array	• Hybrid flash array (CS 6xx series) • All flash array (AS 6xx series)
Is hyper-converged storage system?	Yes	Yes	• No. But SKU for converged system in the name of ProStack is there (often termed as dHCI).	• No • Needs to be used as independent storage system
Hyper-visor supportability	ESXi	• ESX • HyprerV • KVM (talk is in town but not possible in near future)	• Hyper-visor agnostic. Presents disk to hypervisor disk directly.	• Hyper-visor agnostic. Presents disk to hypervisor disk directly.
iSCSI supportability to instances	• No • Volumes gets presented as SCSI disks	• No • Volumes get presented as SCSI disk	• No, in case of ESXi. • Yes in case of KVM and HyperV	• No, in case of ESXi. • Yes in case of KVM and HyperV
iSCSI supportability to hypervisor host	Not applicable	Not applicable	Yes	Yes
Maximum raw storage capacity	3360 TB / cluster (have not considered advanced way to use 64 nodes)	750 TB / cluster	1100 TB	737 TB
Maximum usable storage capacity	1650 TB / cluster	375 TB / cluster	Depends	Depends
Maximum effective storage capacity (varies with environment)	TBD	1200 TB / cluster	4000 TB	TBD
Fault tolerance	Multiple RAID types	RAID 10 mirroring	• RAID 6 • RAID 6 + Hot sparing • RAID-3P	• Only RAID 6
Availability	TBD	TBD	99.9999%	100%
Thin provisioning	Yes	Yes (always)	Yes	Yes
De-duplication	• Yes • Only available for all flash disk vSAN	Yes	• Yes • Need to check whether it is available with Adaptive flash storage array or not	• Yes • Need to check whether it is available with Hybrid flash storage array or not
Compression	• Yes • Only available for all flash disk vSAN	Yes	• Yes • Need to check whether it is available with Adaptive flash storage array or not	• Yes • Need to check whether it is available with Hybrid flash storage array or not

Granular control on IOPS for instances to leverage pay based on the quality	Yes (using storage policy)	Yes (using storage policy)	Yes. Two ways: Yes (using storage policy) and VVOL	Yes (using storage policy). Not sure about VVOL supportability.
Scalabiltiy	Supports horizontal, vertical and segmented scalability.	Supports horizontal, vertical and segmented scalability.	Supports only vertically scalability.	Supports only vertically scalabliity.
Ability to work in federated mode	Yes	Yes		
Backup of volume presented to instance	Native VMware backup functionality	Yes	Native VMware snapshot functionality	Native VMware snapshot functionality
Native application consistency backup	TBD	Yes	No	No
Snapshot of volume presented to instance	Native VMware snapshot functionality	Yes	Native VMware snapshot functionality	Native VMware snapshot functionality
Clone of volume presented to instance	No	Yes	No	No
Backup of disk volume presented to hypervisor host	Not applicable	Not applicable	Yes	Yes
Snapshot of disk volume presented to hypervisor host	Not applicable	Not applicable	Yes	Yes
Clone of disk volume presented to hypervisor host	Not applicable	Not applicable	Yes (zero-copy)	Yes
Ability to support differentiated storage based on volume types	Yes	No	Requires intelligent, thoughtful and careful design of data store	Requires intelligent, thoughtful and careful design of data store
Disaster recovery	Yes	Yes	Need to apply data center architecture and is not native to cloud eco-system because of lack to operate at VM level	Need to apply data center architecture and is not native to cloud eco-system because of lack to operate at VM level
Manageability	Moderate	Easy	Easy	Easy
Maintainability / Operation cost	Very good	Very good	Excellent	Excellent
Telemetry	Good	Good	Excellent (because of InfoSight)	Excellent (because of InfoSight)
OpenStack compatibility	Yes	Yes	Yes	Yes
Public cloud compatibility	No	No	Yes	No

Recommendation

The recommendation is to use HPE SimpliVity as block storage system for Agena's VMaaS solution. It wins by a significant margin because of its rich data feature set which is VM centric instead of volume centric, a very critical aspect to be considered when offering VM vending in cloud way. If we ignore that edge of HPE SimpliVity then VMware vSAN fits best bill of using HPE Compute and storage in tandem because of its best fitment with Apollo 4200.