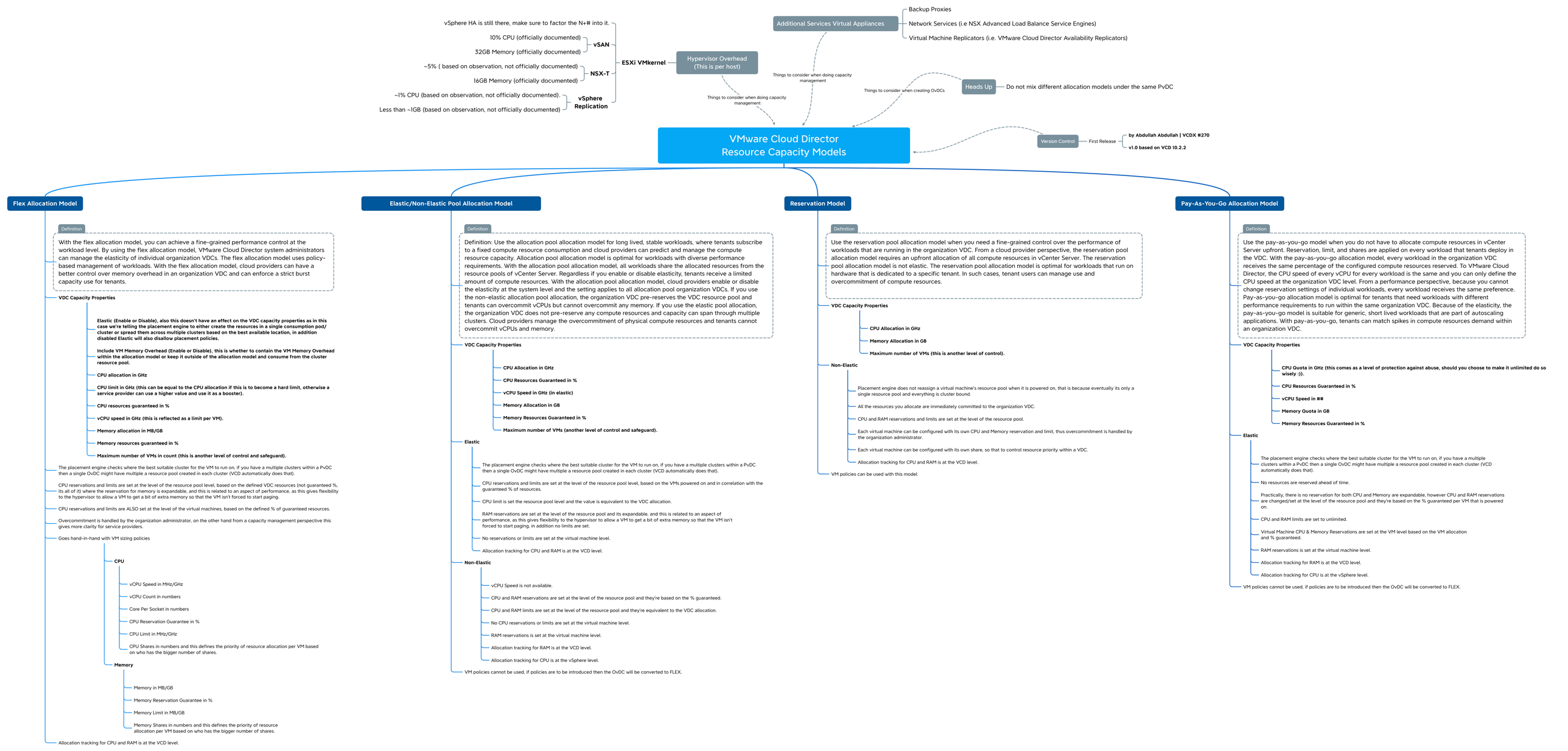
vSphere HA is still there, make sure to factor the N+# into it.
10% CPU (officially documented)
32GB Memory (officially documented)
vSAN
ESXi VMkernel
~5% (based on observation, not officially documented)
16GB Memory (officially documented)
NSX-T
~1% CPU (based on observation, not officially documented).
Less than ~1GB (based on observation, not officially documented)
vSphere Replication
Hypervisor Overhead (This is per host)
Things to consider when doing capacity management
Additional Services Virtual Appliances
Backup Proxies
Network Services (i.e NSX Advanced Load Balance Service Engines)
Virtual Machine Replicators (i.e. VMware Cloud Director Availability Replicators)
Things to consider when doing capacity management
Things to consider when creating OvDCs
Heads Up
Do not mix different allocation models under the same PvDC
VMware Cloud Director
Resource Capacity Models
Version Control
First Release
by Abdullah Abdullah | VCDX #270
v1.0 based on VCD 10.2.2
Flex Allocation Model
Definition
With the flex allocation model, you can achieve a fine-grained performance control at the workload level. By using the flex allocation model, VMware Cloud Director system administrators can manage the elasticity of individual organization VDCs. The flex allocation model uses policy-based management of workloads. With the flex allocation model, cloud providers can have a better control over memory overhead in an organization VDC and can enforce a strict burst capacity use for tenants.
VDC Capacity Properties
Elastic (Enable or Disable), also this doesn't have an effect on the VDC capacity properties as in this case we're telling the placement engine to either create the resources in a single consumption pod/ cluster or spread them across multiple clusters based on the best available location, in addition disabled Elastic will also disallow placement policies.
Include VM Memory Overhead (Enable or Disable), this is whether to contain the VM Memory Overhead within the allocation model or keep it outside of the allocation model and consume from the cluster resource pool.
CPU allocation in GHz
CPU limit in GHz (this can be equal to the CPU allocation if this is to become a hard limit, otherwise a service provider can use a higher value and use it as a booster).
CPU resources guaranteed in %
vCPU speed in GHz (this is reflected as a limit per VM).
Memory allocation in MB/GB
Memory resources guaranteed in %
Maximum number of VMs in count (this is another level of control and safeguard).
The placement engine checks where the best suitable cluster for the VM to run on, if you have a multiple clusters within a PvDC then a single OvDC might have multiple a resource pool created in each cluster (VCD automatically does that).
CPU reservations and limits are set at the level of the resource pool level, based on the defined VDC resources (not guaranteed %, its all of it) where the reservation for memory is expandable, and this is related to an aspect of performance, as this gives flexibility to the hypervisor to allow a VM to get a bit of extra memory so that the VM isn't forced to start paging.
CPU reservations and limits are ALSO set at the level of the virtual machines, based on the defined % of guaranteed resources.
Overcommitment is handled by the organization administrator, on the other hand from a capacity management perspective this gives more clarity for service providers.
Goes hand-in-hand with VM sizing policies
CPU
vCPU Speed in MHz/GHz
vCPU Count in numbers
Core Per Socket in numbers
CPU Reservation Guarantee in %
CPU Limit in MHz/GHz
CPU Shares in numbers and this defines the priority of resource allocation per VM based on who has the bigger number of shares.
Memory
Memory in MB/GB
Memory Reservation Guarantee in %
Memory Limit in MB/GB
Memory Shares in numbers and this defines the priority of resource allocation per VM based on who has the bigger number of shares.
Allocation tracking for CPU and RAM is at the VCD level.
Elastic/Non-Elastic Pool Allocation Model
Definition
Definition: Use the allocation pool allocation model for long lived, stable workloads, where tenants subscribe to a fixed compute resource consumption and cloud providers can predict and manage the compute resource capacity. Allocation pool allocation model is optimal for workloads with diverse performance requirements. With the allocation pool allocation model, all workloads share the allocated resources from the resource pools of vCenter Server. Regardless if you enable or disable elasticity, tenants receive a limited amount of compute resources. With the allocation pool allocation model, cloud providers enable or disable the elasticity at the system level and the setting applies to all allocation pool organization VDCs. If you use the non-elastic allocation pool allocation, the organization VDC pre-reserves the VDC resource pool and tenants can overcommit vCPUs but cannot overcommit any memory. If you use the elastic pool allocation, the organization VDC does not pre-reserve any compute resources and capacity can span through multiple clusters. Cloud providers manage the overcommitment of physical compute resources and tenants cannot overcommit vCPUs and memory.
VDC Capacity Properties
CPU Allocation in GHz
CPU Resources Guaranteed in %
vCPU Speed in GHz (in elastic)
Memory Allocation in GB
Memory Resources Guaranteed in %
Maximum number of VMs (another level of control and safeguard).
Elastic
The placement engine checks where the best suitable cluster for the VM to run on, if you have a multiple clusters within a PvDC then a single OvDC might have multiple a resource pool created in each cluster (VCD automatically does that).
CPU reservations and limits are set at the level of the resource pool level, based on the VMs powered on and in correlation with the guaranteed % of resources.
CPU limit is set the resource pool level and the value is equivalent to the VDC allocation.
RAM reservations are set at the level of the resource pool and its expandable, and this is related to an aspect of performance, as this gives flexibility to the hypervisor to allow a VM to get a bit of extra memory so that the VM isn't forced to start paging, in addition no limits are set.
No reservations or limits are set at the virtual machine level.
Allocation tracking for CPU and RAM is at the VCD level.
Non-Elastic
vCPU Speed is not available.
CPU and RAM reservations are set at the level of the resource pool and they're based on the % guaranteed.
CPU and RAM limits are set at the level of the resource pool and they're equivalent to the VDC allocation.
No CPU reservations or limits are set at the virtual machine level.
RAM reservations is set at the virtual machine level.
Allocation tracking for RAM is at the VCD level.
Allocation tracking for CPU is at the vSphere level.
VM policies cannot be used, if policies are to be introduced then the OvDC will be converted to FLEX.
Reservation Model
Definition
Use the reservation pool allocation model when you need a fine-grained control over the performance of workloads that are running in the organization VDC. From a cloud provider perspective, the reservation pool allocation model requires an upfront allocation of all compute resources in vCenter Server. The reservation pool allocation model is not elastic. The reservation pool allocation model is optimal for workloads that run on hardware that is dedicated to a specific tenant. In such cases, tenant users can manage use and overcommitment of compute resources.
VDC Capacity Properties
CPU Allocation in GHz
Memory Allocation in GB
Maximum number of VMs (this is another level of control).
Non-Elastic
Placement engine does not reassign a virtual machine's resource pool when it is powered on, that is because eventually its only a single resource pool and everything is cluster bound.
All the resources you allocate are immediately committed to the organization VDC.
CPU and RAM reservations and limits are set at the level of the resource pool.
Each virtual machine can be configured with its own CPU and Memory reservation and limit, thus overcommitment is handled by the organization administrator.
Each virtual machine can be configured with its own share, so that to control resource priority within a VDC.
Allocation tracking for CPU and RAM is at the VCD level.
VM policies can be used with this model.
Pay-As-You-Go Allocation Model
Definition
Use the pay-as-you-go model when you do not have to allocate compute resources in vCenter Server upfront. Reservation, limit, and shares are applied on every workload that tenants deploy in the VDC. With the pay-as-you-go allocation model, every workload in the organization VDC receives the same percentage of the configured compute resources reserved. To VMware Cloud Director, the CPU speed of every vCPU for every workload is the same and you can only define the CPU speed at the organization VDC level. From a performance perspective, because you cannot change reservation settings of individual workloads, every workload receives the same preference. Pay-as-you-go allocation model is optimal for tenants that need workloads with different performance requirements to run within the same organization VDC. Because of the elasticity, the pay-as-you-go model is suitable for generic, short lived workloads that are part of autoscaling applications. With pay-as-you-go, tenants can match spikes in compute resources demand within an organization VDC.
VDC Capacity Properties
CPU Quota in GHz (this comes as a level of protection against abuse, should you choose to make it unlimited do so wisely :)).
CPU Resources Guaranteed in %
vCPU Speed in ##
Memory Quota in GB
Memory Resources Guaranteed in %
Elastic
The placement engine checks where the best suitable cluster for the VM to run on, if you have a multiple clusters within a PvDC then a single OvDC might have multiple a resource pool created in each cluster (VCD automatically does that).
No resources are reserved ahead of time.
Practically, there is no reservation for both CPU and Memory are expandable, however CPU and RAM reservations are changed/set at the level of the resource pool and they're based on the % guaranteed per VM that is powered on.
CPU and RAM limits are set to unlimited.
Virtual Machine CPU & Memory Reservations are set at the VM level based on the VM allocation and % guaranteed.
RAM reservations is set at the virtual machine level.
Allocation tracking for RAM is at the VCD level.
Allocation tracking for CPU is at the vSphere level.
VM policies cannot be used, if policies are to be introduced then the OvDC will be converted to FLEX.