
Lunch and learn
VMware HA and DRS
Presented by : Joseph Griffiths
@Gortees
contact@jgriffiths.org

HA and DRS

- HA – High availability – technology that provides restart of virtual machines after a compute failure

- DRS – Distributed Resource Scheduler – technology that balances workload across a virtualization cluster

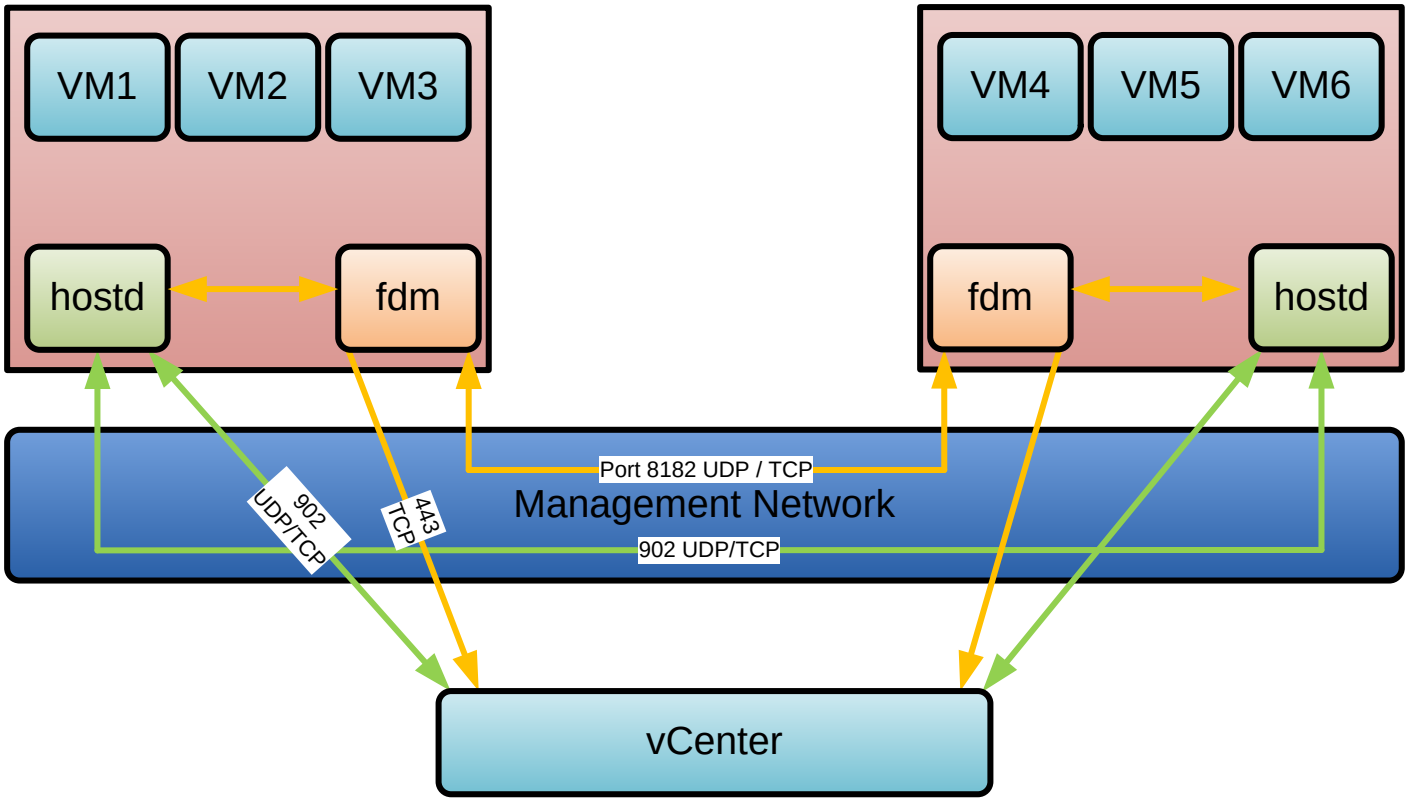
Fault Domain Manager (FDM)

- Handles communication between hosts including – resource information, virtual machine states, HA properties, heartbeats, virtual machine placement, logging
- Uses port 8182 UDP/TCP between ESXi hosts
- Logged to /var/log/fdm.log
- Talks directly to vCenter and hostd process on ESXi

Required Components for HA

- hostd talks to vCenter and handles many functions around virtual machines including power on

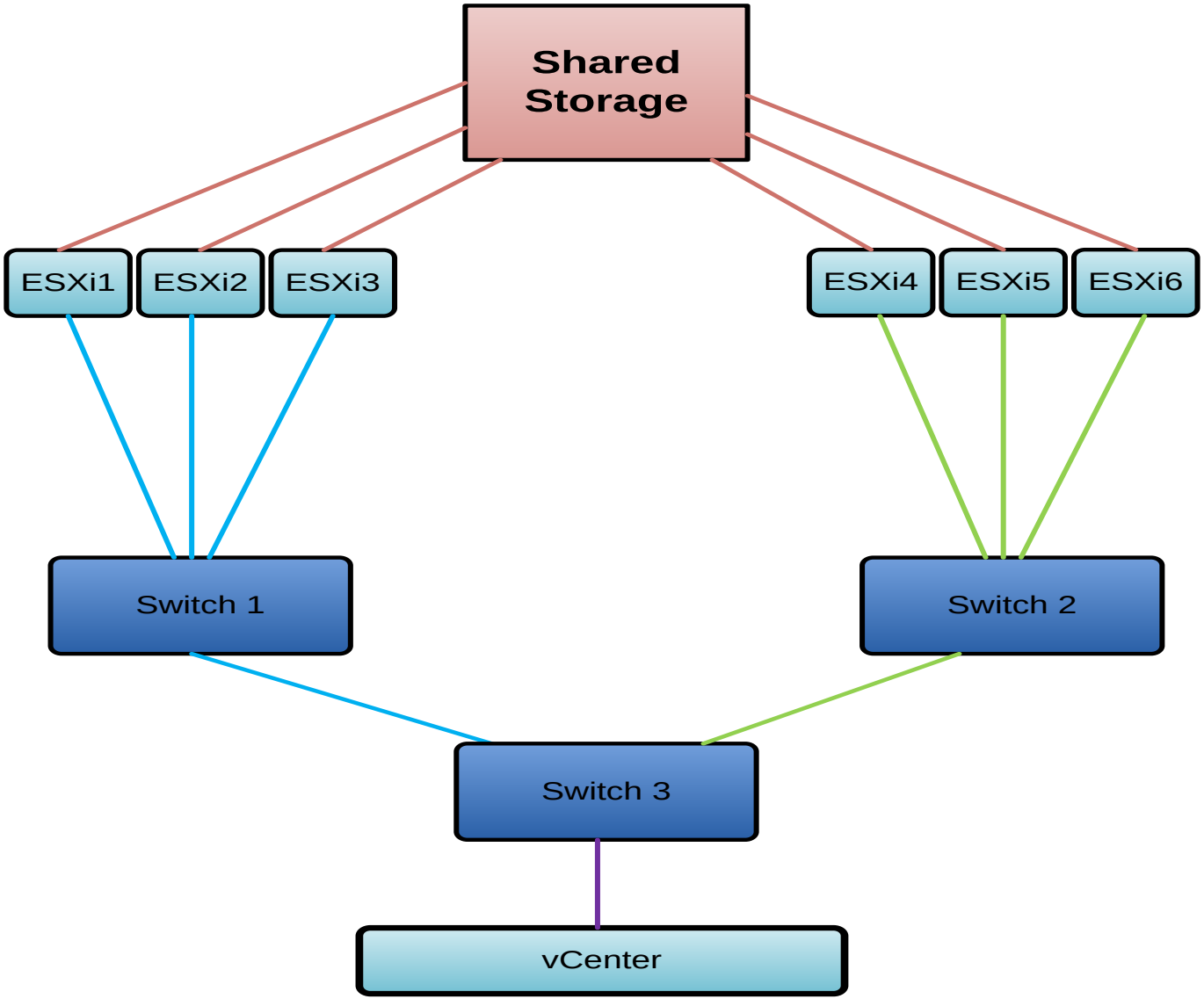
- vCenter is responsible for deploying HA agents (FDM), communication of cluster changes to HA Master, protection of virtual machines


VM1
VM2
VM3
VM4
VM5
VM6
hostd
fdm
fdm
hostd
Port 8182 UDP / TCP
Management Network
902 UDP/TCP
443 TCP
902 UDP/TCP
vCenter

Master and Slaves

- Roles
  - Master
    - Tracks state of virtual machines
    - Takes action on protected virtual machines
    - Tracks state of slaves and reports to vCenter
  - Slave
    - Monitors it’s virtual machines and reports to master
    - Takes actions as directed by master
    - Tracks state of it’s self and reports to master
  - HA Communication between hosts in encrypted TCP on management network

Failure, Isolation and Partitioned



Failure Times

Failure of Slave

Timeline of slave failure

- T0 Failure
- T3s - Master monitors datastore heartbeats for 15 seconds
- T10s - Host listed as unreachable and master pings management
- T15s - if no heartbeat datastores configured host is declared dead
- T18s - if heartbeat datastores are configured host is declared dead

Failure of Master

Timeline of master failure

- T0 - Master failure
- T10s - Master election process
- T25s - New master elected and reads protectedlist
- T35s - New master initiates restarts for all vm's on protectedlist not running

Who is the new master

- Master election happens when
  - HA is enabled or reconfigured
  - Master failed
  - Network partitioned or isolated
  - Disconnected from vCenter
  - Master in standby or Mantenance Mode
- Takes ~15 seconds and is done via UDP
- During election no HA actions will happen

Isolation Times

Isolation detection

- Done with heartbeats between master and slave if a single beat gets through then it is not isolated
- Then will ping isolation address and continue until it works or heartbeats return (by default gateway – by be up to 9 addresses)
- Checks datastore for isolated access

Isolation of slave

Timeline of slave isolation

T0 Isolation of host
T10s Slave enters election state
T25s Slave elects itself as master
T25s Slave pings "isolation address"
T30s Slave declares itself as isolated
T60S Slave triggers isolation response
(Can be adjusted with das.config.fdm.isolationPolicyDelaySec

Isolation of Master

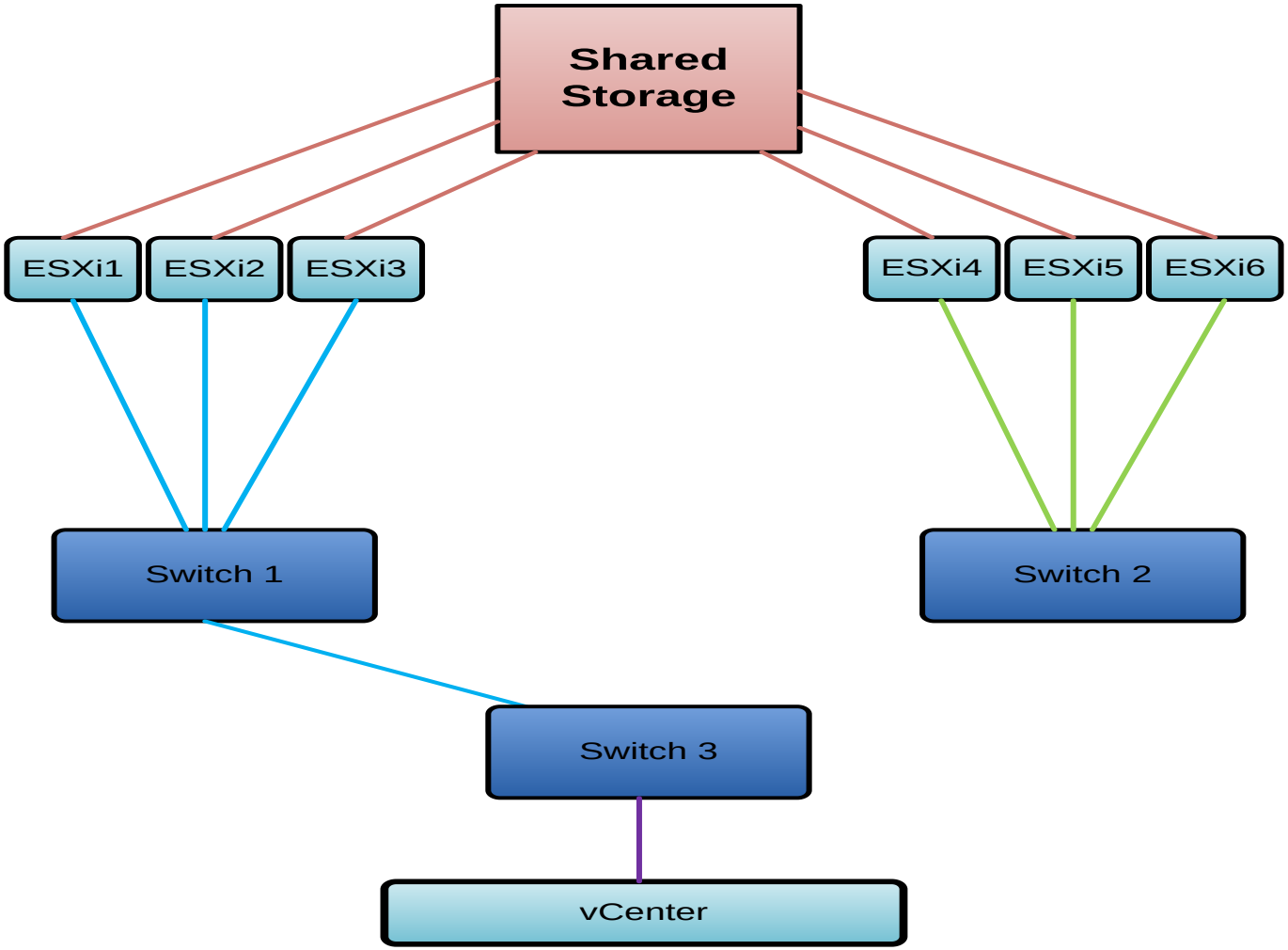
Timeline of master isolation

T0 Isolation of host master
T5s Master declares itself isolated
T35s Master Triggers isolation response

Isolation responses

- Power off - power off virtual machines - pull power cable
- Shut down - guest initiated shutdown timeout can be adjusted by das.isolationShutdownTimeout
- Leave powered on - Leave it on when isolated

What about partitions?



VM Restart

- Restarting virtual machines during failure
- Powered on virtual machines are determined via the poweron file this file is asynchronously read every 30 seconds.
- Machines with a restart priority of disabled are filtered out
- HA does take the following into account
  - CPU and memory Reservation
  - Unreserved capacity of the hosts in the cluster
  - Restart priority of virtual machines relative to other virtual machines to be restarted
  - Virtual machine to host compatability
  - dvPorts required vs available on target host
  - Max vCPU and virtual machines on a host
  - Restart latency - amount of time it takes to initiate virtual machine restarts
  - Checks for required agent virtual machines
- Power on per host is 32 VM's at a time

Is DNS required for HA to operate?

Is vCenter required for HA to Work?

Is networking required for HA to work?

What is the best host isolation response?

Will admission control stop HA from restarting?

Admission control

Function of HA handled by vCenter

- HA will not violate Admission control except in a failure – this is host based not vCenter
- Host failures – based on slot size – highest reservation or 25 Mhz / memory overhead
- Percentage of Cluster resources (100 / number of nodes for n+1)
- Failover hosts

DRS

- Function of vCenter – run as a single thread on vCenter
- Communication between vCenter and vxpa on each ESXi host (443 TCP) is two way
- Host provides vMotion information and power state
- DRS is invoked every 5 minutes, a host is added to the cluster or maintenance mode on a host

Dynamic Entitlement

- Defines a target that represents the ideal amount of resources eligible for use
- Based upon two things
  - Static User defined resource specifications
  - Dynamic element estimated demand and level of contention (Reservations, Shares, Limits)

Resources

DRS uses two metrics for resource scheduling:

- CPU based upon %RUN + %READY (Ready to run state)
- Memory based upon active memory + memory overhead + (idle consumed memory *.25)

Resource Allocation settings

- Reservation - Amount of physical resources guaranteed
- Shares - Relative importance of VM compared to others
- Limit - Upper bound for resources that can be allocated to virtual machine

What can resource allocation settings be applied to?

CPU Shares

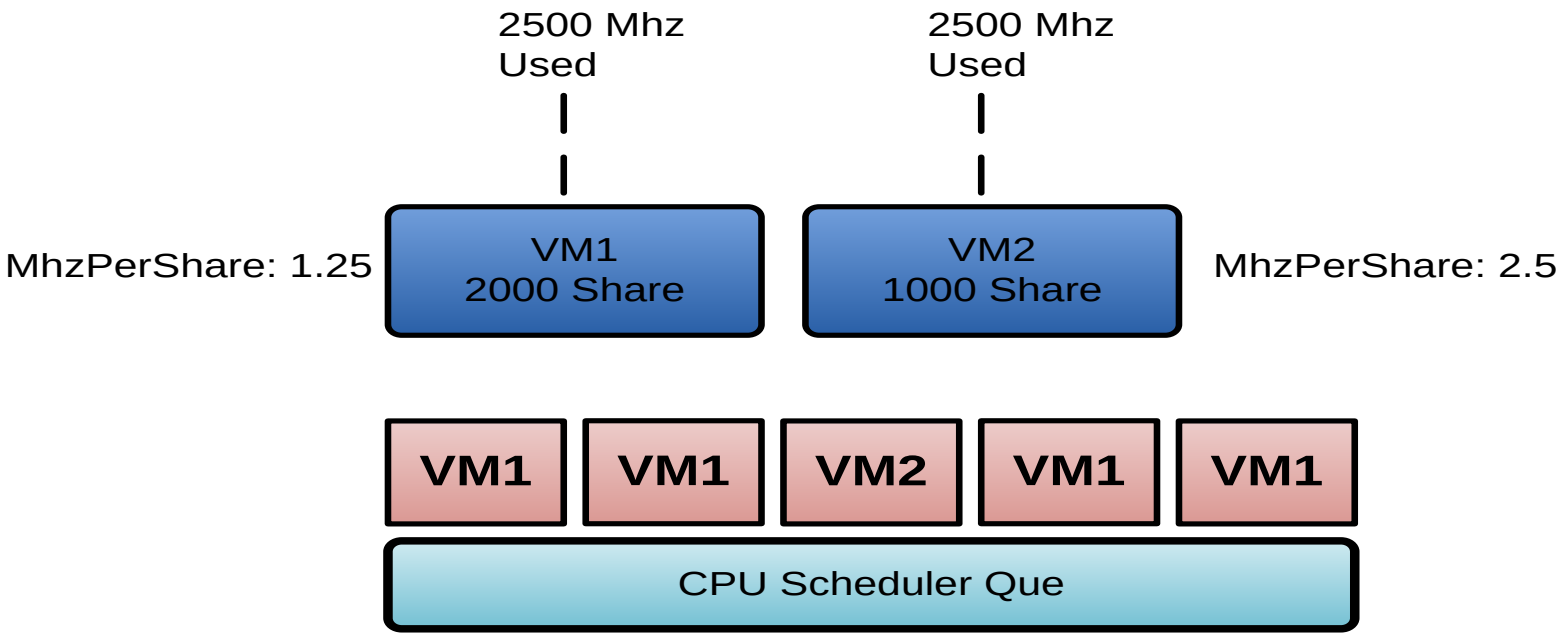
Entitlement of CPU
Ahead – Using more than entitlement
Behind – Not using what is allocated

MhzPerShare = MhzUsed / Share

Unused entitlements can be used by other machines

Main Scheduler – Select lowest MhzPerShare and execute

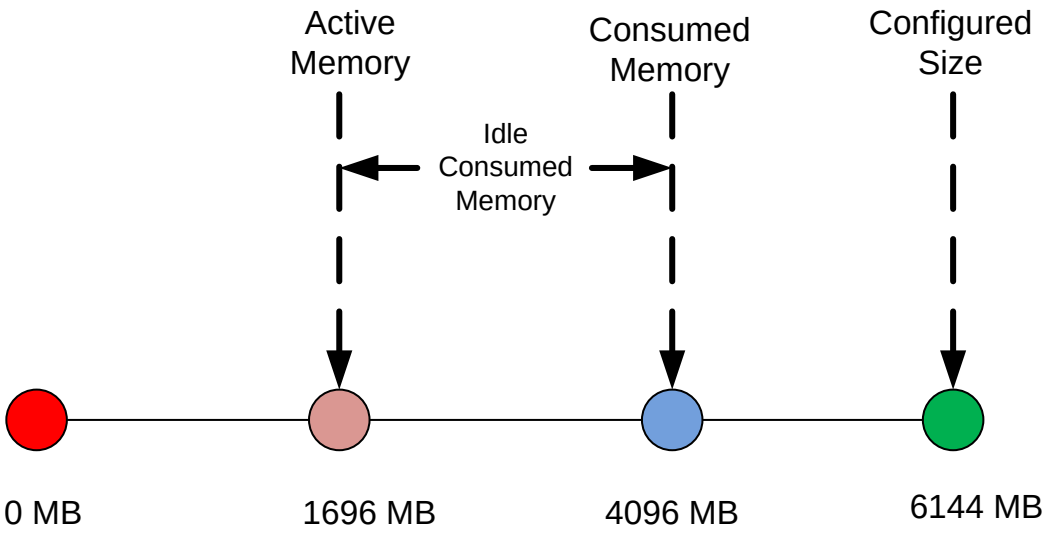
Extra Que – Holding place for ahead



Memory Shares

Memory Demand = Active Memory + 25% idle
Consumed Memory + memory overhead

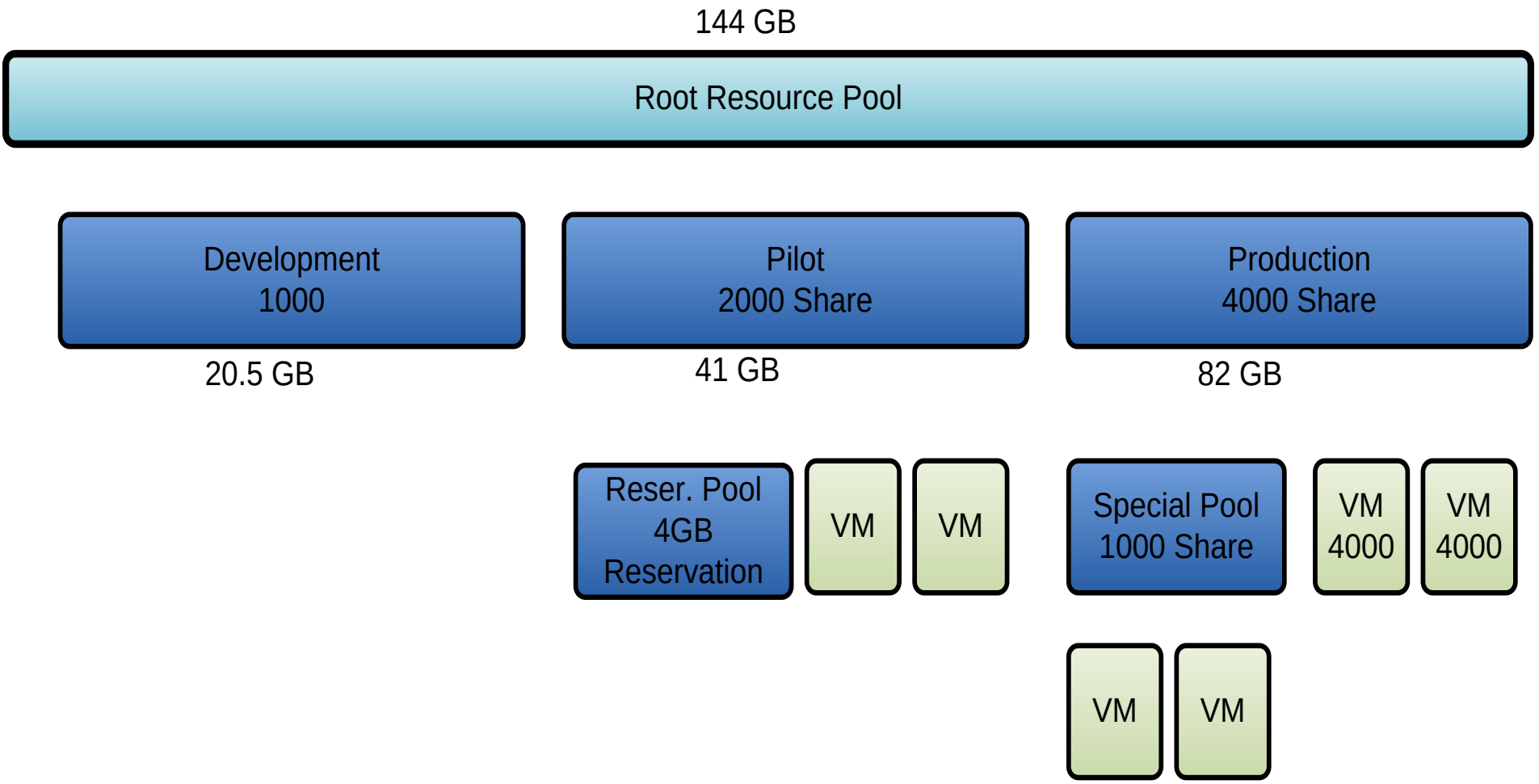
25% idle consumed memory = consumed memory – active memory *.25



If Dynamic entitlement determines contention then memory reclamation techniques are used in order:

- Balloon
- Compression
- Vmdk Swap

Resource Pools



Questions?

Twitter: @Gortees

Email: contact@jgriffiths.org