Real World Nutanix

By James Kilby @Jameskilbynet

About Me

- Senior Cloud Engineer for a Leading UK ISP
- Primary focus on VMware + supporting stack
- Awarded VMware vExpert 2014-2017
- Veeam Vanguard 2017 - Veeam UK usergroup founder
- Working with VMware technology since 2008
- Nutanix Platform Professional

Presented @Scotland year before last on Puppet and automation. Go and see Marks talk

Overview

- So why Nutanix ?
- What is it ?
- The good the bad and the ugly…..

Questions welcome at any point
or come and chat to me after

Road to Nutanix

- vCloud Director Infrastructure refresh project
- Traditional 3 Tier Dell, HP + Cisco…
- Storage HP, Dell, EMC, Nimble, Pure
- HCI - VSAN, Nutanix, Atlantis, Simplivity

Looked at lots more We run SAN both FC and iSCSI by 4 different manufactures
VSAN to expensive - VSPP (Called out by a number of providers) - This has now been changed

7 Clusters Built

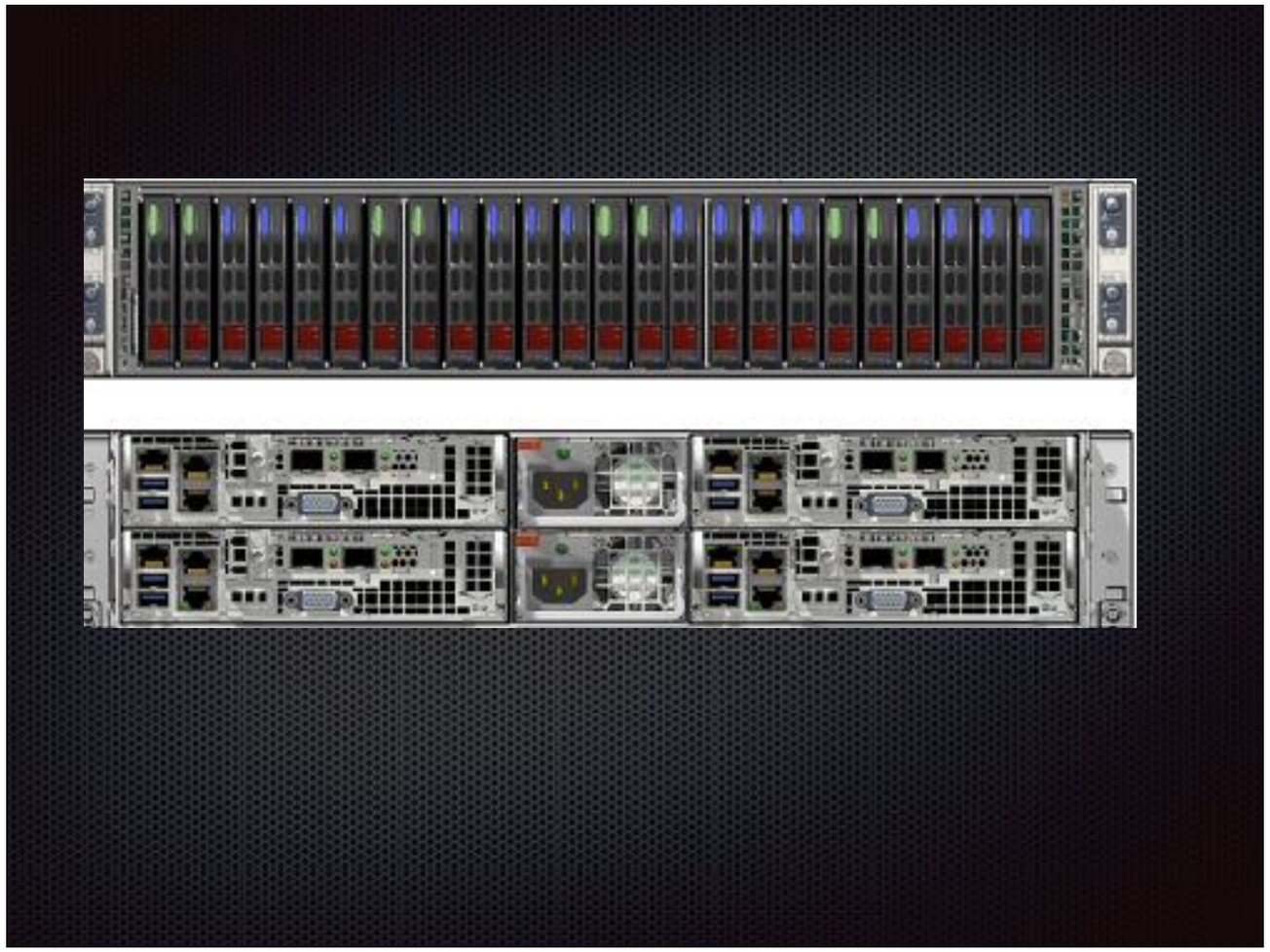
- 2X Management (NX-1065-G4)
- 2X Public Cloud (NX-8035-G4)
- 2X ISP Systems (NX-8035-G4 & G5)
- 1X Test Cluster (NX-3000-G5)

In prod today. Concentrating on scaling out.
very different workloads across the clusters
Slowed down by:
Introduction of NSX
Datacentre consolidation
New WAN!!

So What is Nutanix ??

- Hyper-Converged Infrastructure
- vSphere, Hyper-V, Xen & KVM (Acropolis)
- All built around the CVM(Controller Virtual Machine)
- SuperMicro, Dell, Lenovo or Cisco UCS hardware

No Backplane
All internode transfer through standard 10gb
node=host
block=chassis
Specs are our 8000 series



3000 Series
4 node

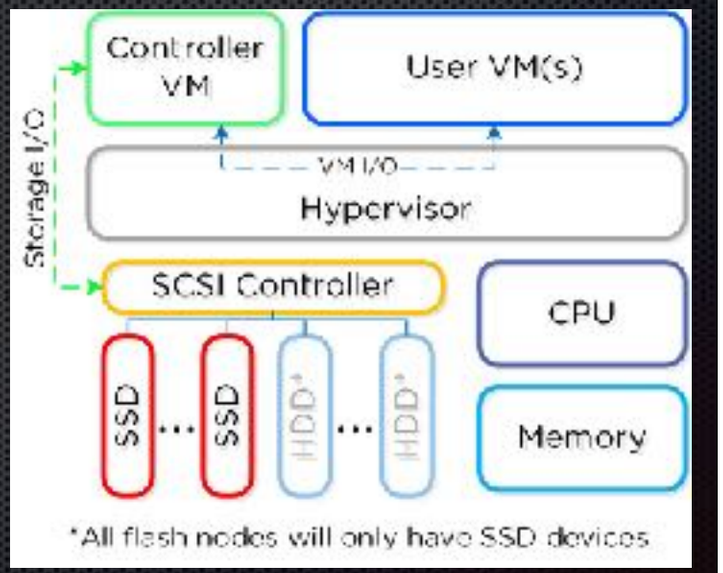
So what is the CVM?

- CentOS based VM - core to Nutanix
- All I/O - compress, dedupe & erasure coding
- Only Nutanix component that uses RAID (mdraid)
- Boot from ISO and SATA DOM

No ASIC’s - All X86
Tiny part of the SSD’s are used for the CVM os hence the raid (in dual sad systems)
All disks/SSD are serving I/O- always.
No hot spares etc

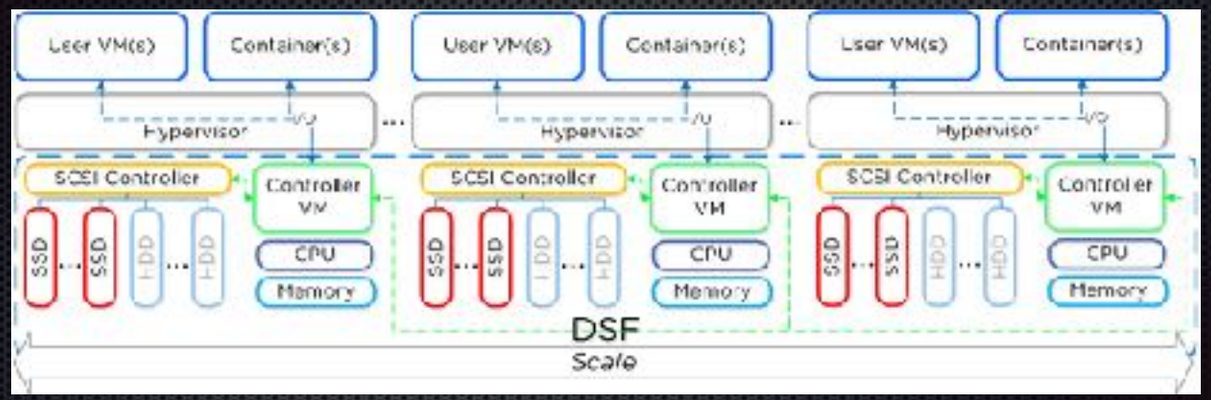
CVM Logical Layout

- One CVM Per Node
- First VM to Boot



Typically 2 SSD. 4 DISK
MGMT 1 ssd 2 disk

Distributed Storage Fabric



DSF is the fabric that containers (different properties) are added onto
these are then presented as NFS back to VMware
Scale out almost limitless

What else does the CVM do ?

- Prism Element (control plane)
- Backup - Nutanix cluster or AWS/Azure
- Replication - Nutanix cluster (async/metro cluster)
- Cluster performance trending, alerting & patching

AWS - Single node cluster M1.XLarge
vCenter, VUM, vROPS, vSphere Replication, vSphere Data Protection, vSphere Web Client, Platform Services Controller (PSC) and the supporting database platform (e.g.: SQL/Oracle/Postgress).

But isn't it hungry ?

- Default 8vCPU and 16GB of RAM

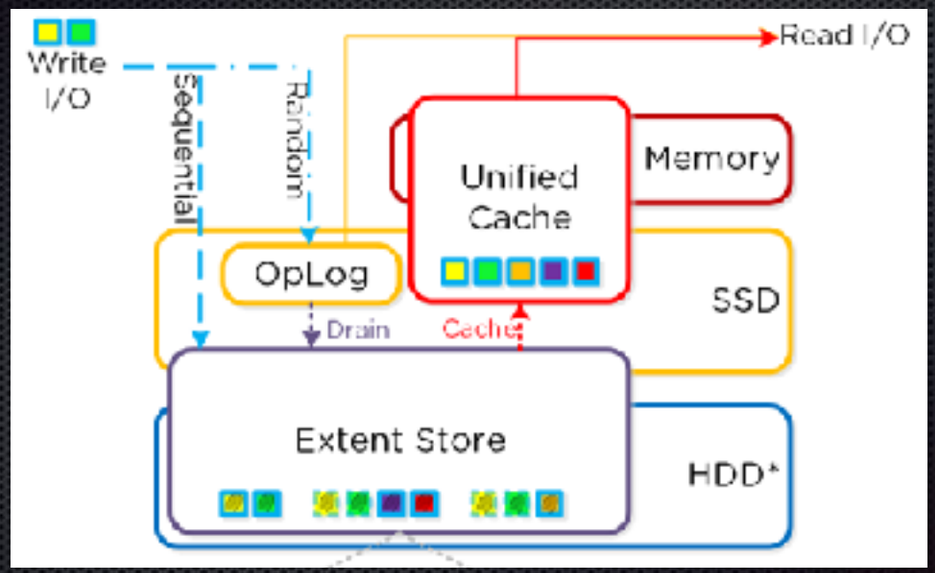
We run our CVM’s with 24-32GB OF RAM. Typically not short of ram in our clusters get extra read cache. 5% of ram.
24GB recommended for dedupe
2x14 CPU with HT. 14-28%
CVM also has reservations

CVM Data Services

- Replication factor + checksumming (every read & write)
- Autopath - Preference to local CVM where possible
- Writes - always local and remote before being acknowledged
- IO - Tiering within CVM

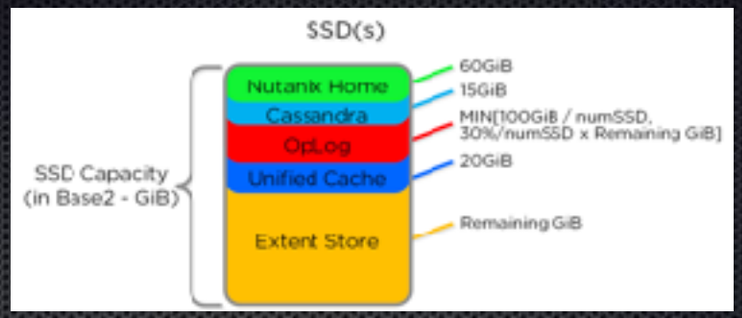
In failed condition I/O served from remote CVM over 10Gb
Not Every I/O is served locally. but most are. (read). verified in our environment
RAM LOCAL SSD REMOTE SSD.   Local or remote disk. Local CVM monitor. (disk queue)

IO Tiering

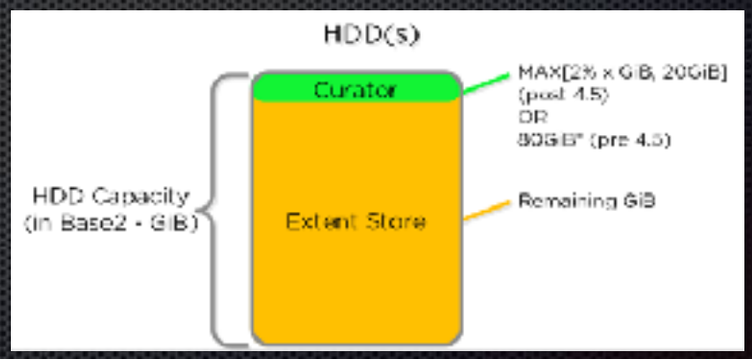


SSD like performance random writes into SSD before draining
Sequential go straight to disk
Unified cache made up of single touch and multi touch pool

SSD

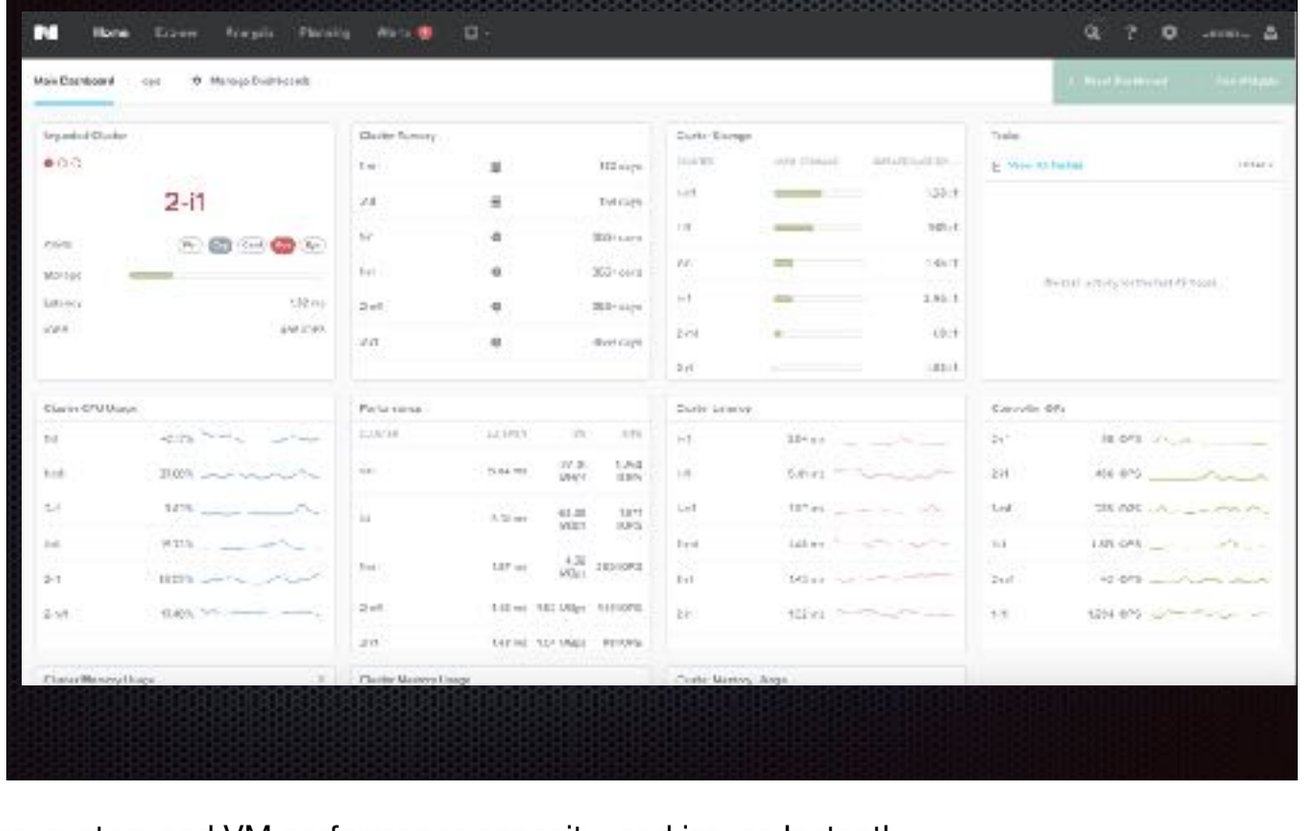


HDD



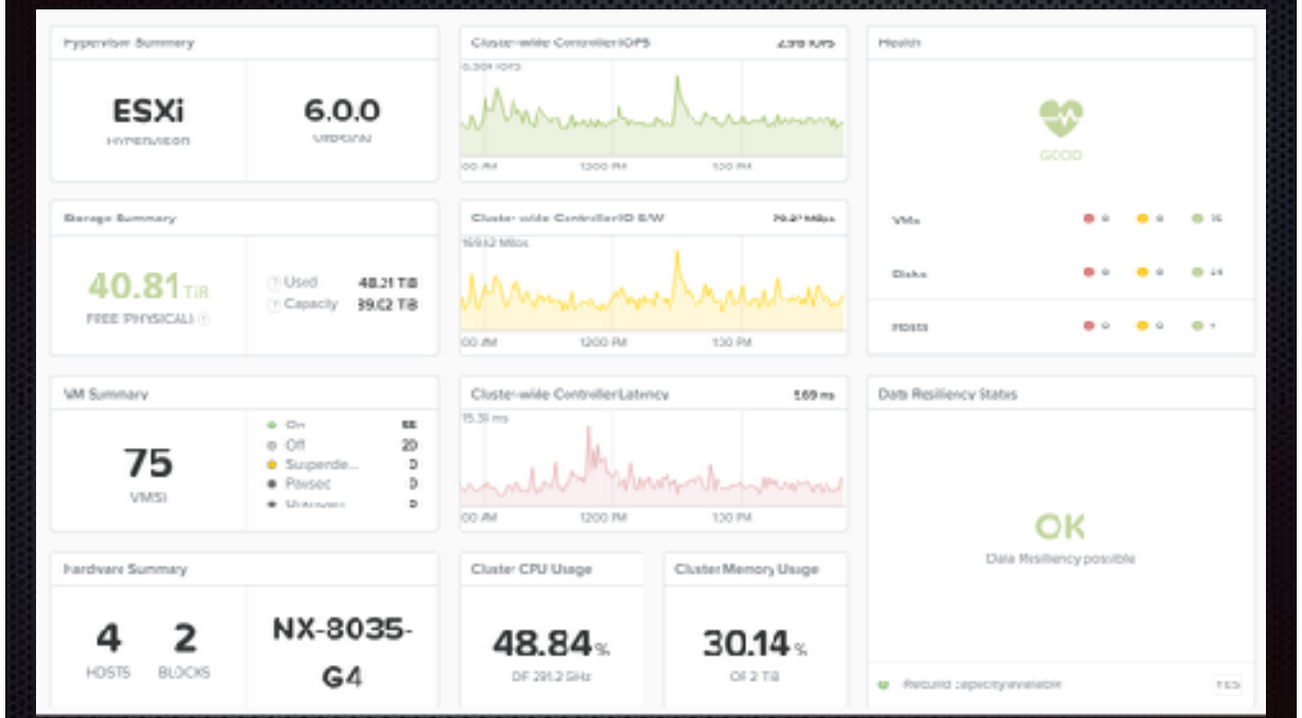
Complex operations going on under the hood to optimise performance and scale

Prism Central



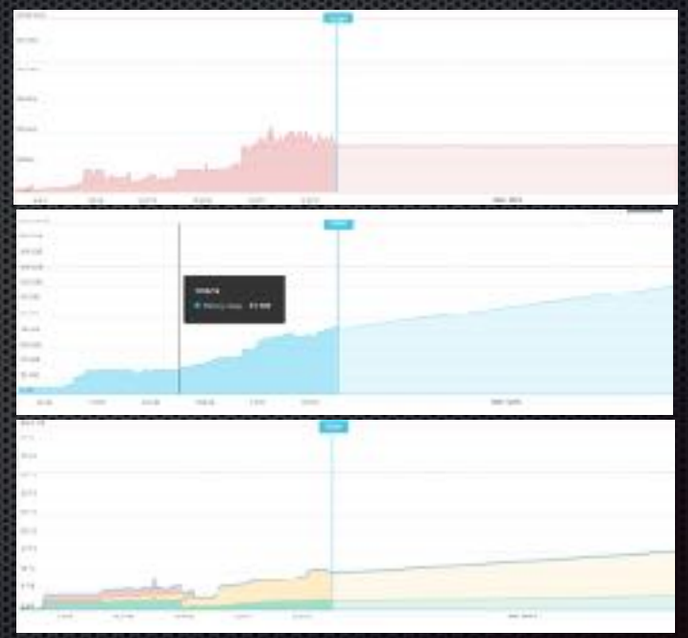
Shiny html5 snappy. Able to see hardware system and VM performance capacity and issues Instantly running 100’s of checks to ensure cluster is healthy and performing as expected

Prism Element



Capacity Management

- CPU
- RAM
- DISK



Constantly calculating the days required before you runout
4 host cluster. dotted line at 75%
As a customer were looking for them to trend at the same rate

Drill Down Metrics

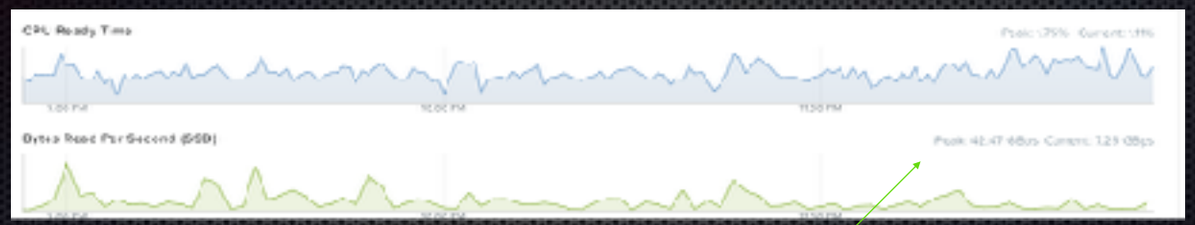
Powerful diagnostics. = OPS team use heavily
IAAS customer. SQL Server 48gb RAM 2TB Previously on Dedicated SSD Tier
last 3hrs at glance can add to custom dash and go back in time

Apollo steady state 1500-2000 with peaks around 8-1000

Drill Down Contd

IAAS Customer _ previosuly all flash tier of trad san
Read by SSD 99.93%
HDD 0.07

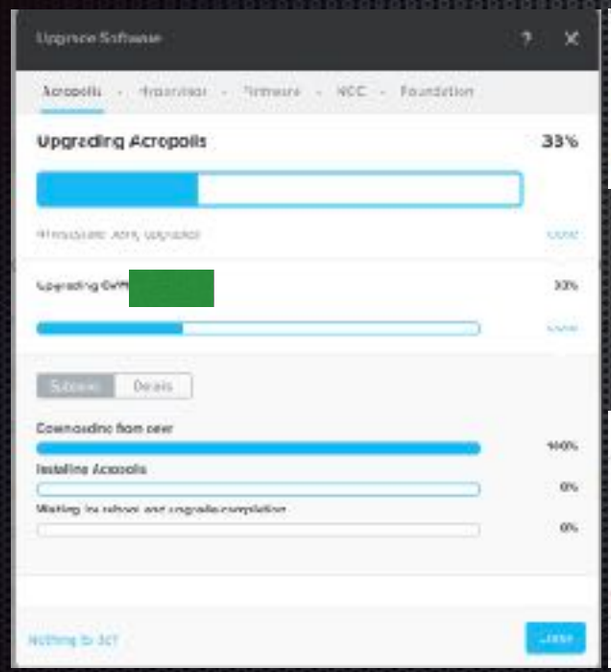
Real World Performance
Log Insight Instance

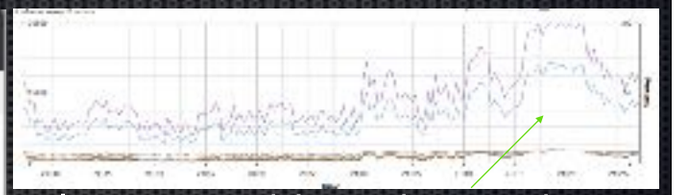


Peak 42GBps

Log insight Mgmt Cluster (much less flash)
Peak 42GBps
3x Node. 500GB SSD + 4TB Disk per node

One Click Upgrade (almost)

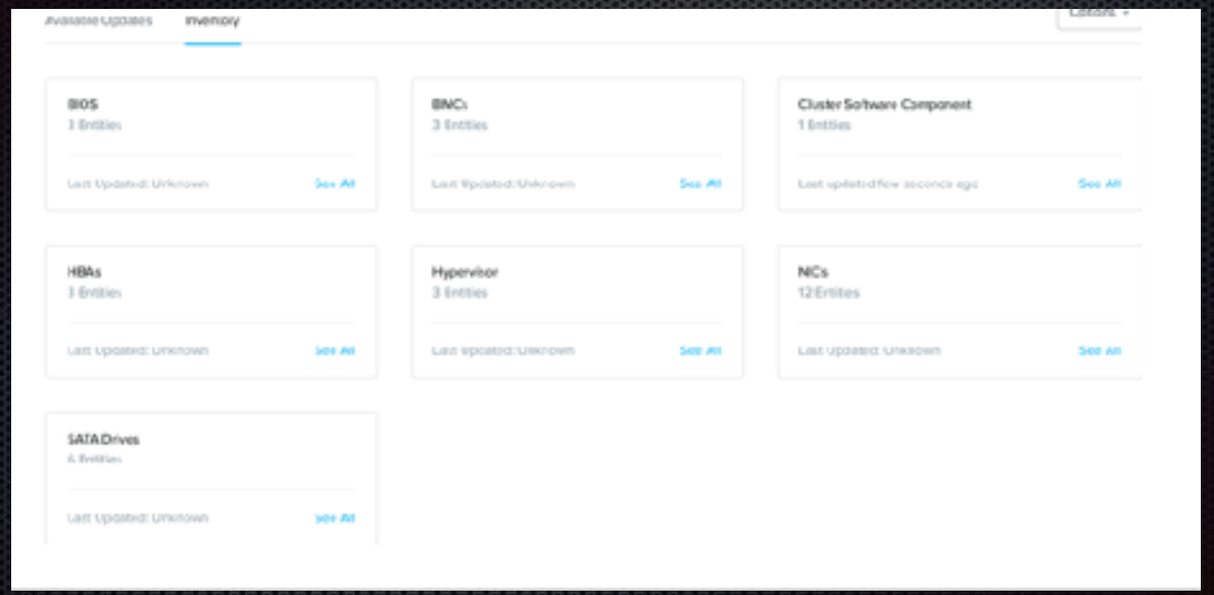




Increased Load on other CVM
Small latency increase

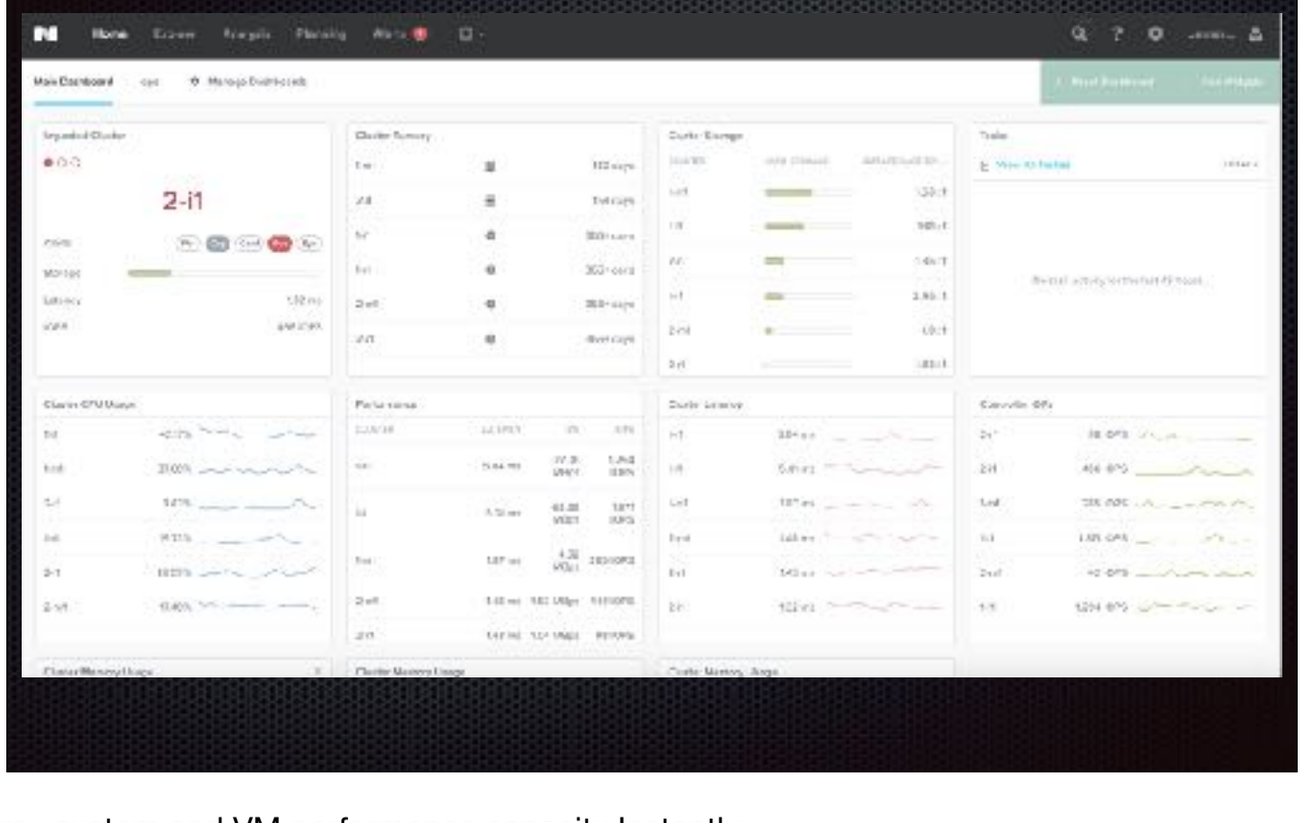
Upgrades of Acropolis (CVM) are done regularly (most of ours are done during the day)
Hypervisor also handled in same way
NCC done whenever new one is published
43% were running 4.6 100 days after release

Lifecycle Management



Everything designed to make your life easier
BIOS BMC HBA. Hypervisor NIC and Disk

Prism Central



Shiny html5 snappy. Able to see hardware system and VM performance capacity Instantly

Failures

- If delayed I/O detected by Stargate disk is offlined
- Hades removes it from the data path and performs health checks
- If unhealthy it tells Zeus to remove from cluster config
- Red disk light turned on and alerts sent
- Data rebalance performed

Because all disks in the cluster are used return to RF happens quickly
Cold-Cold
Hot-Hot. so as not to blow cache
Perfomance consistent during rebuild

Failure Conditions..

- DISK Failure - Non Issue. (Loss of capacity)
- SSD Failure - Reduction of cluster performance
- CVM Failure - I/O distributed to other CVM’s
- Node Failure- Loss of CPU RAM Storage + IOPS

Built appropriately. N-1 means you loose all of the node
Can build RF3

The Really Good

- Support - They have been exceptional
- Expanding clusters - 3 IP’s and your away
- NCC - Built in enhanced health checks
- Foundation - Provisioning tool
- Backup - Performance in Veeam increased 5-10x
- Density - 224 pCPU, 2TB of Ram 60TB usable storage
- Prism Search - vm iops <1500

Overall very happy- Some caveats to watch out for backup approx 1-1.2Tb per hour without impacting live all vm’s more than 1500 iops or linux

Things to watch out for

- Make sure you size the CVM correctly
- SVMotion large VM's to Nutanix is Slow….
- Be careful with aggressive DRS settings
- Extreme storage change rate in compressed timescale can cause problems
- Time. Ensure NTP is setup correctly

Easy to change CVM size but try and get it right first time
Never need to worry about Storage DRS (unless changing container)
Change rate need to let Cassandra map reduce full scan
time Casandra ring distributed db. Rolling back time is v hard. forward ok

Whats Next ??

- File Services
- Containers
- All Flash
- Storage Only Nodes
- Acropolis

File services for scale out file services distributed per node. think vdi/citrix profile server etc

Nutanix CE - Try it out

1, 3 or 4 Node AHV Cluster

Feed4ward - It's about the community

- Step up - You might enjoy it.
- Lots of support available
- 10-30 Min slots usually available

Just because you think its trivial. Doesn't mean other people did great talk about doing VMware on a budget
Not out to trip you up

Any Questions ?

LUNCH