vSAN by way of RADIO

Alexander Garthwaite
Senior Staff Engineer, VMware
April 4, 2023
Agenda

My background
RADIO and how R&D benefits
vSAN overview and challenges
Additional resources
My Background at VMware

Eight years in the Virtual Machine Monitor group

- Primarily focused on memory management between the vmm and both hosted and bare-metal (ESX) kernels
- Page-sharing, ballooning, swapping, checkpointing, handling large pages, sampling behavior...

“The evolution of an x86 virtual machine monitor”, Agesen, Garthwaite, Sheldon, Subrahmanyam, SIGOPS OSR Vol 44 Issue 4, December 2010

“Optimizing VM Checkpointing for Restore Performance in VMware ESXi”, Zhang, Denniston, Baskakov, Garthwaite, USENIX ATC 2013

”Tesseract: reconciling guest I/O and hypervisor swapping in a VM”, Arya, Baskakov, Garthwaite, VEE 2014

Four years in the Storage and Availability groups: native disk snapshots, core vSAN

Rejoined the Monitor Group three years ago to rearchitect the Virtual Machine Monitor
RADIO
The Research and Development Innovation Offsite

Annual internal R&D conference
• San Francisco, 2.5 days
• Roughly 3000 engineers from around the globe
  – More than half of R&D is now outside of Palo Alto, CA
  – 30% of R&D
• Research program, BoFs, Breakout sessions, keynotes, and the Expo

Research chair in 2018 and 2019
• The core of RADIO
• 470-560 submissions each year, 16% selected (25-30 as talks, 70-90 as posters in Expo)
• Broad sense of what VMware is doing

Origin of many ideas that make it into our products

Lesson: continual renewal due changes in the problems, scale and underlying technologies
Kubernetes and ESXi: better together

Project Pacific
- Released in VSphere 7.0
- Part of VMware Tanzu built around Kubernetes
- Broader impact

Started with a project presented at RADIO 2018
Kubernetes
Background

K8S API Master

Node
kubelet
Pod
Pod
Pod

Node
kubelet
Pod
Pod

Node
kubelet
Pod
Pod
vSphere
Background

- vCenter
- ESXi
- hostd
- VM
- VM
- VM
- VM
- VM
Kubernetes

Pods

- Strongly isolated Containers
  - Networking
  - Volumes
  - Processes
  - Interprocess Communication
vSphere

VMs do it better

VMs are strongly isolated.

- Unique OS Instance!
- Per-VM network stack.
- Per-VM process lists.
- Per-VM volumes.
vSphere
The best of both worlds
New Challenges

We have a number of notions of clusters

- VSphere clusters
- High availability clusters
- vSAN clusters
- Kubernetes clusters
- ...

Challenges:

- Distributed Resource Scheduler (DRS)
- EMM (maintenance mode)
- Health checks

Opportunities

- Learn from the new
- Shift from action-oriented to desired-state-oriented operations
**vSAN**

Virtual SAN

Distributed, Scalable Object Storage

Started with a project presented at RADIO 2012

“vSAN: Modern Distributed Storage”, Fink, Knauft, Zhang, SIGOPS OSR, September 2017

“Hybrid Cloud Storage: Bridging the Gap between Compute Clusters and Cloud Storage”, Gupta, Spillane, Wang, Austry, Feredonny, Karamanolis, SIGOPS OSR, Sept. 2017

“Essential Virtual SAN (vSAN)”, 2nd ed., Hogan, Epping, VMware Press

“Storage Design and Implementation in vSphere 6”, 2nd ed., Khalil, VMware Press

“VMware vSphere 6.7 Clustering Deep Dive”, Denneman, Epping, Hagoort, VMware Press
Modern Object-Based Storage for vSphere

vSAN objects and components

The vSAN datastore is an **object** store

Each object made up of one or more **components**

Data (components) is distributed across cluster based on VM storage policy

 Assign **storage policy** to many VMs, single VM, or single VMDK
Object-Based Storage – A Better Way for Data Protection

Setting failures to tolerate (FTT) to 1 with RAID-1 mirroring

Alternative FTM to RAID-5/6 erasure coding

Data **mirrored** to another host

Witness needed to determine quorum

Requires **fewer hosts** but not as space efficient as RAID-5/6

Additional hosts needed to support greater than FTT 1 or maintenance operations
Object-Based Storage – A Better Way for Data Protection

Setting failures to tolerate (FTT) to 1 with RAID-5 erasure coding

Alternative to RAID-1 Mirroring

Data with parity striped across hosts

For erasure coding, **FTT 1 implies RAID-5**, FTT 2 implies RAID-6

Guaranteed **space reduction**
- 30% savings with RAID-5
- 50% savings with RAID-6

Additional hosts needed to support greater than FTT 1 or maintenance operations
Provide Awareness of Topology and Rack Designs

vSAN custom fault domains

Create **explicit** fault domains to increase availability

Protect against **rack failure**, etc.

**Ensures** other copy/replica does NOT live in the same rack as first copy
Low Cost Local and Remote Protection

Integrated stretched clustering technology

Redundancy **locally** and **across** sites

Upon site failure, vSAN maintains availability with local redundancy in surviving site

No change in stretched cluster configuration steps

**Optimized** site locality logic to minimize I/O traffic across sites
vSAN Components

- **CMMDS**: Cluster monitoring, membership and data svc
- **CLOM**: Cluster-level object manager
- **DOM**: Distributed object manager
- **LSOM**: Local log-structured object manager
- **Mgmt**: Management services
LSOM
Local log-structured object manager

Manages local storage
• Storage organized into disk groups
  – Disk group split into caching and capacity tiers
  – All-flash: caching tier used for writes, destaging
• Services
  – On-disk encryption
  – Compression
  – Deduplication of blocks within disk groups
• Unit is a disk group for fault domains

Critical for overall performance
DOM

Distributed object manager

Three layers
- DOM client: on host where disk object is accessed (e.g., VM is running)
- DOM owner: the host that coordinates access to the disk object
- DOM component manager: on each host with components for the object

All three layers run on each host (symmetry)

Optimizations
- Collocate DOM client and owner
- Read locally
- I/O scheduling (resync vs. data path)

DOM owner splits/combines IOs to/from component managers
- Singletons, mirroring, and erasure-encoding
- Two-phase commit protocol
- Detects CRC/integrity issues in blocks, local repair of components
RDT
Reliable Data Transport

Common network layer abstract
- In-order, reliable
- Datagrams, not a stream
- Implemented via TCP connections

Hides underlying transport
- TCP/IP
- RDMA (RoCE2)

Fail-over with multiple NICs

Pairwise communication between hosts

Checksumming
Data Path for an I/O Through Components

RAID-1 Example

```
Application
GuestOS
DOM client
DOM owner
DOM component
LSOM (disks)
```

VM

```
Namespace (VM directory)
VM swap
vmdk(s)
```

```
DOM component
LSOM (disks)
```

```
RD T
```
CLOM

Cluster-level object manager

Overall coordination of storage in the cluster
- Placement of components
- Handling of evacuation events
  - Maintenance mode: do nothing, ensure availability, evacuate all
- Rebalancing
- Rebuilding absent or degraded components

Fault domains
- FTT0: no redundancy
- FTT1: mirroring
- FTT2: RAID5 (requires 4 nodes)
- FTT3: RAID6 (requires 6 nodes)
- Stretch-clusters: Primary vs. secondary fault domains
CMMDS
Cluster monitoring, membership and data service

Peer-to-peer: leader-election forming cluster
- Roles: leader, backup, agents

Detect network partitions and absent hosts within cluster

Distributed key-value store
- Each host owns its own entries, published to all
- Enables subscription service for all events driven by these entries on any node
- Every entry checksummed
Handling Partitions
Cluster monitoring, membership and data service

- Peer-to-peer: leader-election forming cluster
  - Roles: leader, backup, agents

- Detect network partitions and absent hosts within cluster

- Distributed key-value store
  - Each host owns its own entries, published to all
  - Enables subscription service for all events driven by these entries on any node
  - Every entry checksummed
Handling Partitions

Cluster monitoring, membership and data service

Peer-to-peer: leader-election forming cluster
- Roles: leader, backup, agents

Detect network partitions and absent hosts within cluster
Handling Partitions
Cluster monitoring, membership and data service

Each subpartition elects a leader

VMs/clients and accessible objects
- Use of delta components

When cluster partition resolved, leaders abdicate to a single one
Management Services

- Interface to overall management plane and APIs (vCenter)
- Health checks, performance monitoring, and guards
- Maintenance mode workflows, cross-cluster operations
- What-if services
  - Impact of changes to cluster or per-object policies
HCI Mesh
Compute/storage disaggregation

Enable hosts in one cluster to mount vSAN objects from another
• Allow more flexible placement of compute workloads vs. storage
• Avoid stranding storage

Changes the dynamic:
• Multitenancy
• DOM client/owner split

Presented RADIO 2020
Challenges Going Forward

New technologies and scale
• Scale of hosts, data centers, network capabilities
• Storage technologies: persistent memory, larger-scale devices
• Offloading mechanisms, smartNICs, …
• Rack-scale systems

New Services
• Distributed file systems services
• Automated problem identification and remediation
• More flexible failover with network partitions

New frameworks
• Tanzu and Kubernetes: supporting containerized workloads
Additional Resources

“Remote memory in the age of fast networks”, Aguilera, Amit, Calciu, Deguillard, Ghandi, Subramaynyam, Suresh, Tati, Venkatasubramanian, Wei, SoCC 2017

“Scalable and practical locking with shuffling”, Kashyap, Calciu, Cheng, Min, Kim, SOSP 2019

“Project PBerry: FPGA Acceleration for Remote Memory”, Calciu, Puddu, Kolli, Nowatzyk, Ghandi, HotOS 2019

“Don’t shoot down TLB shootdowns!”, Amit, Tai, Wei, EuroSys 2020

“IOctopus: Outsmarting Nonuniform DMA”, Smollyar, Markuze, Pismanny, Eran, Zellweger, Bolen, Liss, Morrison, Tsafrir, ASPLOS 2020


“Rethinking Software Runtimes for Disaggregated Memory”, Irina Calciu, M. Talha Imran, Ivan Puddu, Sanidhya Kashyap, Hasan Maruf, Onur Mutlu, Aasheesh Kolli ASPLOS | April, 2021

https://research.vmware.com
Thank You