HA Cluster Plugin

Introduction

At the heart of the ZFS HA Cluster Plug-in is a mature and stable enterprise class high availability product called RSF-1. It was the first commercial HA solution for Sun/Solaris environments and has an 18+ year track record in data centres worldwide providing high-availability assurance for some of the most demanding customers.

In seeking to understand what the Plug-in does and how it does it, it is useful to split out its functions into a conceptual model, with functional modules having discrete capability as follows.

Conceptual Model

RSF-1 Controlling Module: allows a cluster of machines to provide a number of ZFS Pools, and attempts to keep those pools available and accessible by watching for failing machines and starting the pools that were running on those machines elsewhere in the cluster.

Heart beat module: Each machine in the cluster runs a daemon called rsfmon, which exchanges information with the other members of the cluster. This exchange is done by means of 'heartbeats', which both provide an indication that the sending machine is still active, and give information about the states of the pools it is configured to run. These heartbeats are essential to the operation of RSF-1, and so may be duplicated over multiple communications channels. Any combination of TCP/IP networks (public, private & cloud beacons for stretch topologies), shared disc partitions and serial rs232 connections may be used.

Fencing module: This adds a layer of assurance on top of ZFS that prevents split brain. It achieves this by employing a number of techniques and mechanisms to ensure that ZFS pools cannot be imported by more than one machine at a time. There is a sequence of checks that would all have to fail prior to the final failsafe backstop (mhd scsi reservation and failfast) being invoked should a machine attempt to start the import of an already imported pool.

HA Cluster Plugin

Pool attach module: The cluster is expected to provide access to a number of ZFS Pools over a TCP/IP network. To allow the pools to be run on different machines, each one is associated with a different (and sometimes multiple) internet address. RSF-1 handles attaching these addresses to an interface on the machine where the pool is to run. Additionally collections of ZFS pools can be grouped together within one atomically controllable service for start-up and shutdown on the same machine.

Cluster Synchronisation Module: Whilst many applications are able to maintain the meta-data they require to function on shared storage, some like zfs were not originally designed to operate in a clustered environment. RSF-1 therefore takes charge of this synchronisation process; atomically ensuring state changes made on one node in the cluster are propagated throughout. This functionality is essential in pool failover ensuring consistency of COMSTAR targets and providing software ALUA capability.

IP Failover Module: Ultimately each ZFS storage pool is accessed via one or more Virtual IP interfaces. On import of a pool the cluster configures the addresses on the appropriate machine and then issues a gratuitous ARP so that all clients are forced to update their ARP cache and at that stage have instant access to the pool. On manual/graceful failover the cluster software removes those interfaces prior to forcing export of the ZFS storage pool. This module integrates seamlessly with IP multi-pathing and link aggregation on the platform.

Scheduling module: Each pool can be configured so that it may be run on any set of machines in the cluster, from a single machine to all available machines, with a defined order of preference. Pools may be controlled on each machine by marking them as automatic (meaning that they may start on that machine) or manual (they will not be started on that machine, but will continue to run if already running). A running pool may also be manually migrated to other machines in the cluster thus enabling preventative maintenance without downtime.

Import Export Module: When RSF-1 starts or stops a pool, the IP address(es) for that pool is attached or detached from an interface, and a directory of scripts and programmes is used to perform pool control actions. There are a number of these that perform standard operations that dramatically reduces and parallelise pool import times (e.g. use of alternate cache files), whilst there are others that enable non cluster aware ZFS components for use in a storage cluster.

C/C++/Python/Perl/Java API Module: All functions available in RSF-1 via the command line interface are also accessible via an API so that cluster configuration/administration, pool control, status display, etc. are all achievable via a programmable interface.

Licence Module: Basic RSF-1 capability and functionality can be controlled via licence key. The ability to try before you buy, provide cluster capability for a defined period or to force a cluster to operate in a manual only mode (e.g. for distance/metro clusters), are all possible via this mechanism as would any future capacity based approach if that were desirable.

HA Cluster Plugin

Key Points & Benefits

- Volume runs on only one appliance at a time
- Rapid Failover - initiated in seconds uses cachefile removes pool scan for vdevs
- Fencing strategy ensures data integrity - prime concern (no split brain possible)
- Ease of use (2 volume cluster creation < 1 minute)
- Simple but Scalable (up to 16 nodes)
- In-Chassis, Campus Clusters, Stretch Cluster geographic topologies
- Seamless integration with ZFS
- Minimal administration required
- Allows preventative maintenance without downtime
- Parallel restart of pools
- 2 or more head nodes possible
- Multiple Volumes & VIP's per service
- Shared Storage Architecture
- Hot Restart - default (dynamic changes possible to the cluster)
- Active / Active (or Active / Passive if you like!)
- Multiple heartbeats & Multiple Types (Disk, NW, Serial)
- Heartbeat through data disks to identify h/w issues early
- Robust data fencing mechanisms (mhdc, VIP check, heartbeat redundancy, SMITH/STONITH)
- Common Protocols & Hardware Support CIFS, NFS, iSCSI, FC
- Supports SAS, SATA (+certified interposers), FC, SCSI, AoE, iSCSI
- Client volume access via virtual IP(s) and/or COMSTAR targets
- No service "bounce back" after failover
- Agent framework for software component monitoring/action
- Licence control of capability
- Management of COMSTAR LU mappings/views on failover
- Full ALUA support on 2 node clusters
- Platform support for Solaris, Open Solaris / Open Indiana / Illumos, Linux, FreeBSD, Mac OSX
- Intelligent networking monitoring integrated with service failover and ipv6 support