The etcd solution

Etcd is a distributed key value store used internally by IBM Cloud Private to store state information. It uses a distributed census algorithm called *raft*. The etcd-based VIP manager leverages the distributed key/value store to control which master or proxy node is the instance holding the virtual IP address. The virtual IP address is leased to the leader, so all traffic is routed to that master or proxy node.

The etcd virtual IP manager is implemented as an etcd client that uses a key/value pair. The current master or proxy node holding the virtual IP address acquires a lease to this key/value pair with a TTL of 8 seconds. The other standby master or proxy nodes observe the lease key/value pair.

If the lease expires without being renewed, the standby nodes assume that the first master has failed and attempt to acquire their own lease to the key to be the new master node. The master node that is successful writing the key brings up the virtual IP address. The algorithm uses randomized election timeout to reduce the chance of any racing condition where one or more nodes tries to become the leader of the cluster.

Note: Gratuitous ARP is not used with the etcd virtual IP manager when it fails over. Therefore, any existing client connections to the virtual IP address after it fails over will fail until the client's ARP cache has expired and the MAC address for the new holder of the virtual IP is acquired. However the etcd virtual IP manager avoids the use of multicast as ucarp and keepalived requires.

The ucarp solution

Ucarp is an implementation of the common address redundancy protocol (CARP) ported to Linux. Ucarp allows the master node to "advertise" that it owns a particular IP address using the multicast address 224.0.0.18.

Each node sends out an advertisement message on its network interface that it can have a virtual IP address every few seconds. This is called the *advertise base*. Each master node sends a skew value with that CARP (Common Address Redundancy Protocol) message. This is similar to its priority of holding that IP, which is the *advskew* (advertising skew). Two or more systems both advertising at one second intervals (`advbase=1`), the one with the lower advskew will *win*.

Any ties are broken by the node that has the lower IP address. For high availability, moving one address between several nodes in this manner enables you to survive the outage of a host, but you must remember that *this only enables you to be more available and not more scalable*.

A master node will become master if one of the following conditions occurs:

- No one else advertises for 3 times its own advertisement interval (advbase).
- The `--preempt` option is specified by the user, and it "hears" a master with a longer (advertisement) interval (or the same advbase but a higher advskew).

An existing master node becomes a backup if on of the following conditions occur:

- Another master advertises a shorter interval (or the same advbase, but a lower advskew).
- Another master advertises the same interval, and has a lower IP address.

After failover, ucarp sends a gratuitous ARP message to all of its neighbors so that they can update their ARP caches with the new master's MAC address.