vOLT-HA High Availability

Implementation Proposal

High Level Concepts

- Docker in swarm mode is used as the platform
- Each service is independently started and clustered (individual compose files)
  - Facilitates individual scaling to expected load
  - Allows for service specific optimizations
- All services are run in load-balancing clusters.
- Number of servers (or VMs) underlying the cluster is 2N+1; N>0.

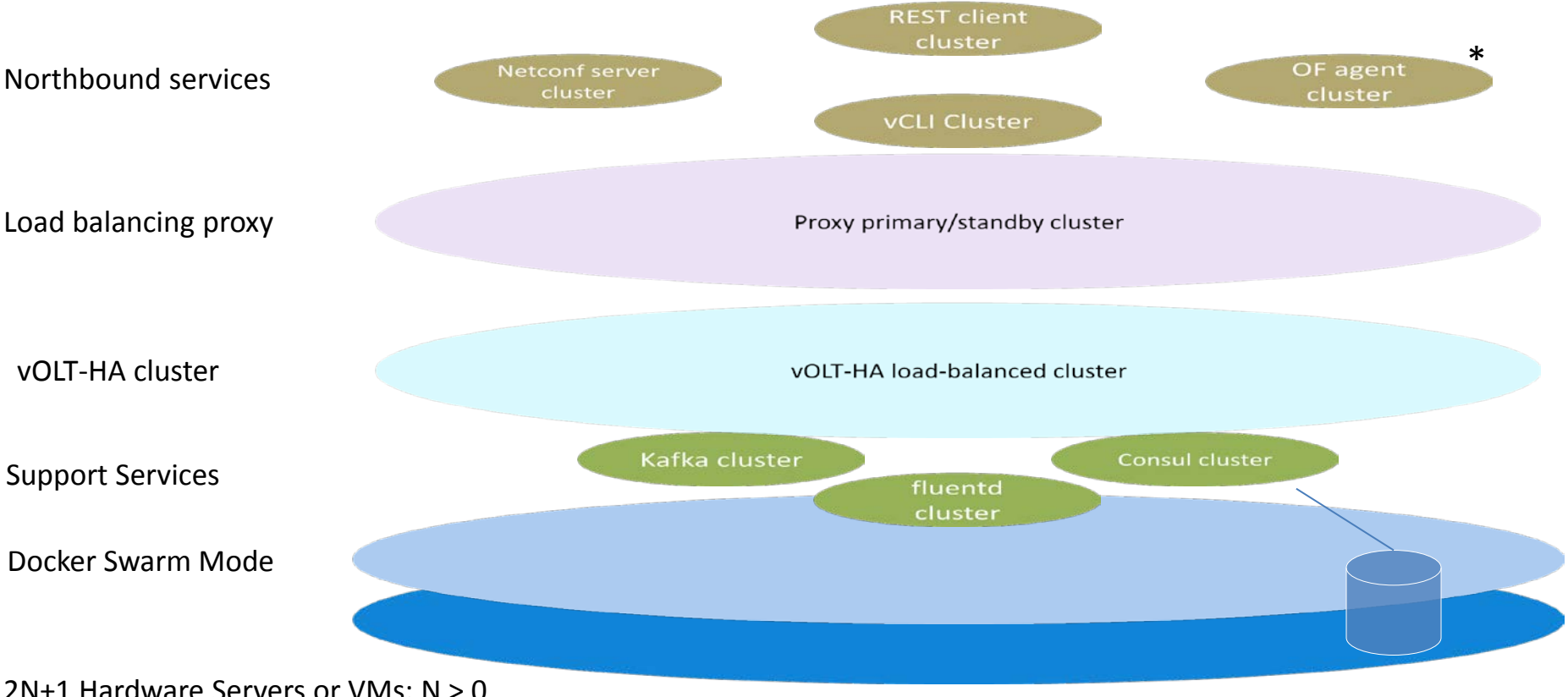
High Level Concepts

- An HTTP2 proxy is used to distribute northbound agent requests.
- All agents are run in primary/backup vIP clusters.
- The proxy load-balances creation (pre-provision) requests across all running vOLT-HA instances.
- The vOLT-HA instance forms a cluster with other running vOLT-HA instances.

High Level Concepts

- The vOLT-HA instance receiving a read / update request will forward to the appropriate instance.
  - This function will be moved to the proxy as a plugin filter in future releases.

High Level Diagram



2N+1 Hardware Servers or VMs; N > 0

Cluster Startup

- Fluentd, Consul, Zookeeper/Kafka
  - All of these can be started simultaneously
- 2N+1 vOLT-HA instances are started after the services are live (N>0)
- Redundant envoy proxy service is started once the vOLT-HA instances are all live.
- Finally, the NBI servers are started once the proxy is up and running.
- All these dependencies are handled by Docker in swarm mode.

Additional Notes

- Docker swarm mode's DNS service eliminates the need for the "registrator" service.
- Docker swarm mode supports an overlay network in a multi-host deployment.
  - Single IP address space across all servers.
- Docker swarm mode supports encrypted communication between containers.
  - Both self signed or using a custom root CA. Simplifies security
- External load balancers / proxies are supported
  - This will allow future disaggregation of the forwarding function in vOLT-HA to the load balancing proxy.

vOLT-HA instance Specifics

- There are 3 vOLT-HA functions to support HA
  - The forwarder
  - The coordinator
  - Persistence
- The forwarder will forward any “non-creation” request to the appropriate instance to handle the request.
  - “Creation” requests are load balanced by the proxy to ensure homogeneous distribution of OLTs across all running vOLT-HA instances.
- The coordinator will perform 2 functions.
  - Mastership election in a cluster formation
  - Assignment of work to new instances (usually after one instance crashes).
- Persistence through consul K/V store
  - Enables the coordinator to provide context to new instances.