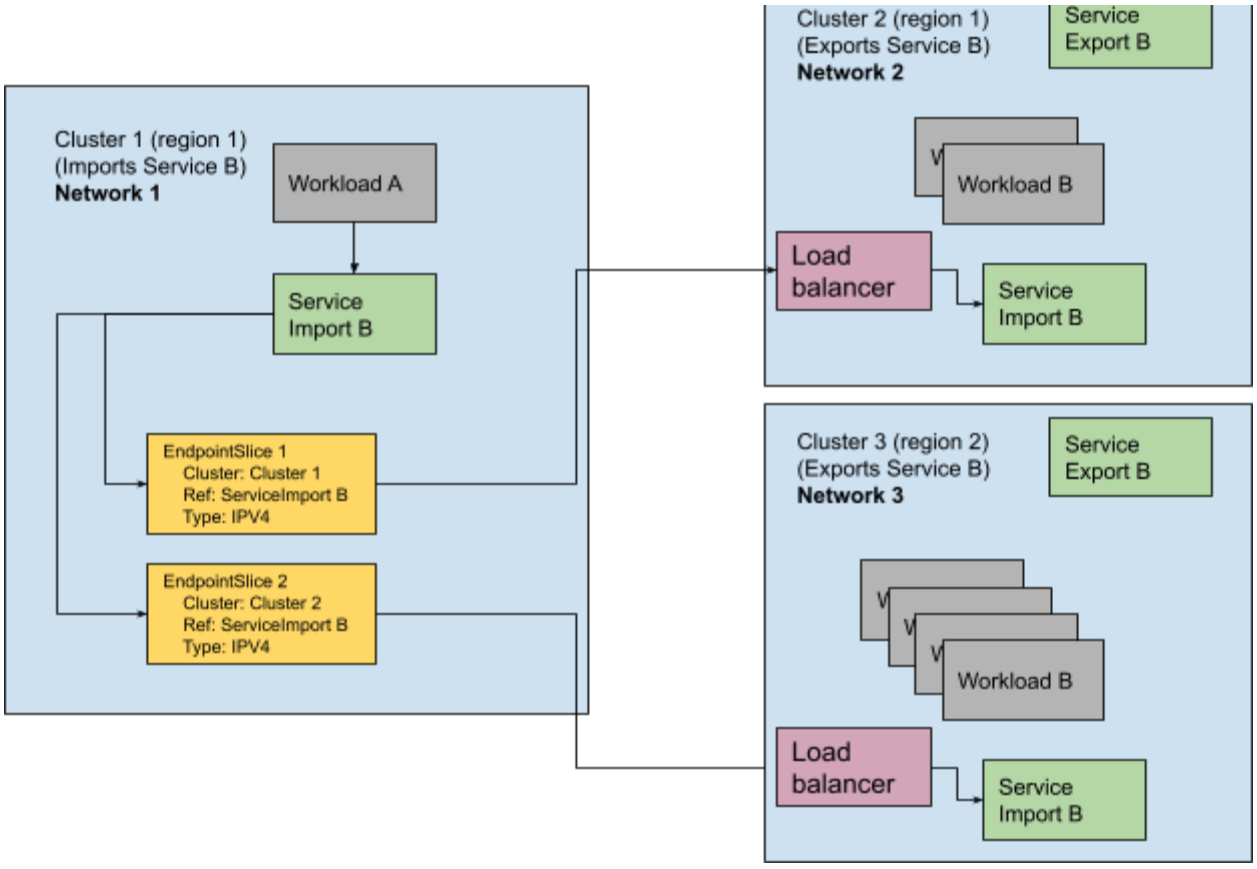
MCS - Multi-network support - Design ideas

MCS for multi-network (non-flat network)

A non-flat network in the context of MCS is one where Pod IPs on one K8s cluster are ***not*** directly reachable from Pods on other K8s clusters participating in forming a multi-cluster setup. In this case the connectivity is established via a Gateway/Load Balancer that acts as an ingress point into the cluster.

Below is the diagram that shows how the EndpointSlices generated for MCS can look like in case of a non-flat network.

EndpointSlices for an MCS Service (non- flat network)



The endpoint slices generated for a workload in a non-flat network would not necessarily represent the actual endpoints (Pod IPs) but rather only that of a Gateway/load balancer that is used to reach the Pods in that cluster. This results in load balancing challenges from the consumer standpoint especially when the endpoints behind EndpointSlices are not identical in number for a workload on two or more clusters.

Proposal

Add an attribute called numberOfActualEndpoints on EndpointSlices that can act as a hint for kube-proxy for better load balancing.

Considerations/Open questions

1. EndpointSlice updates can be very expensive and can cause scalability issues.

    a. The numberOfEndpoints is directly related to the endpoints so the attribute won't change unless the endpoints change at which point the update can be pooled with the changes on EndpointSlice?
2. There are topology hints on each endpoint in an EndpointSlice. How would this work in the context of MCS?

Network identification for services in MCS

Presence of ClusterProperty (network.k8s.io, similar to ClusterId) in a cluster. If this ClusterProperty isn't present on a cluster then its assumed to be a flat network (default)

Questions and Concerns raised

1. `network` is overloaded and can represent Pod network, Overlay network, Ingress network etc.
2. Since its generic, this is going to be used outside of MCS and needs definition and semantics around how this can be used

Alternate proposal

1. Use something specific to MCS, call it **mcs-network.k8s.io** to represent the Pod network for a cluster.
2. ?

References

MCS Spec
ClusterID Spec
MCS DNS Spec