Use of BIER Entropy for Data Center CLOS Networks

draft-xie-mboned-bier-entropy-staged-dc-clos-00

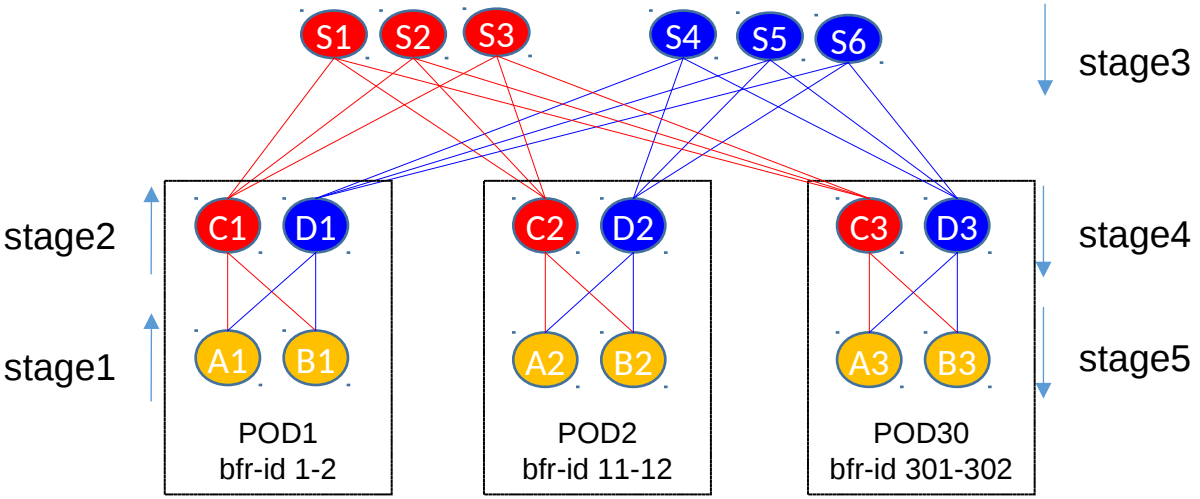
Jingrong Xie, Mike McBride, Gang Yan
@Huawei

Xiaohu Xu
@Alibaba Inc

Problem Statement

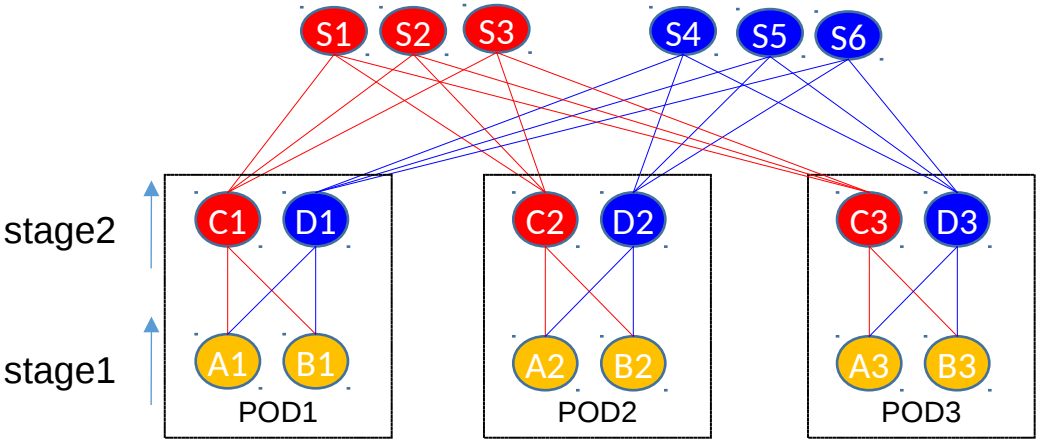
- Describes how BIER Entropy can be useful in DC CLOS networks.
- Large, long lived elephant flows may affect performance of smaller short lived flows and reduce efficiency of per flow load sharing.
- Due to ECMP hash function inefficiencies it is possible to have frequent flow collisions. More flows get placed on one path over the others.
- Isolating faults in the network with multiple parallel paths and ECMP based routing is non trivial due to lack of determinism.
- When BIER is deployed in a multi-tenant data center network for efficient delivery of BUM, an operator may want a deterministic path.
- A deterministic path for a multicast path in the DC, with multiple staged equal cost paths, is comparable to a traffic-engineering path defined in ietf-mpls-spring-entropy-label with multiple hop equal cost paths.
- A deterministic path can be found by part of the 20 bit entropy field. Bit 0 to bit 2 of entropy label can represent a value of 0 to 7, and can be used to select a deterministic path from 8 equal cost paths.

Problem Statement



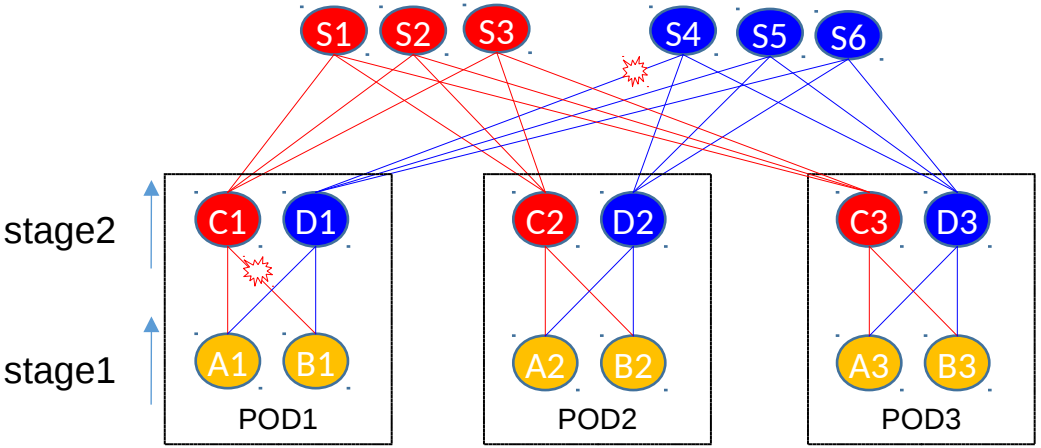
- DC-CLOS network: 3 layers, 5 stages, Northwards Stages (stage 1 & 2) have rich ECMP.
- Problem 1: Steering for elephant flows (A1->C1->S1-->A2)
- Problem 2: Path Division for Tenant flows to different SIs (A1->C1->S1-->A2, A1->D1->S4-->A3)

Solution by using BIER Entropy



- Stage 1: use bit[0] of the 20-bit Entropy to represent RED and BLUE path.
- Stage 2: use bit[1~2] of the 20-bit Entropy to represent 3 paths to each RED/BLUE cluster.
- Similar to the [ietf-mpls-spring-entropy-label] for multi-stage ECMPs along a path by breaking the 20-bit BIER Entropy.

Local convergence and global optimization



- For a flow from A1, originally using Entropy[0]=0, and Entropy[2~1]=00, then
  - upwards path(s): A1->C1->S1
  - downwards path(s): C1->B1(intra-POD), S1->C2->A2/B2(inter-POD).
- When Link between C1 and B1 fail, then A1 can do local convergence
  - upwards path(s): A1->C1->S1, A1->D1. //the BIFT-0 on A1 can converge for BFER<B1> locally.
  - downwards path(s): D1->B1(intra-POD), S1->C2->A2/B2(inter-POD).
- A1 can also do a global optimization by using Entropy[0]=1 and Entropy[2~1]=01 or 02.

Forwarding Procedure

- The use of BIER entropy label to select a path between some equal cost paths is a local configuration matter.
- This draft defines a method to use part of the 20-bit entropy label in each router, and this needs a data-plane to do some bit operation function.
- It is expected to be easier than hashing function.

Thank you !