FAR: A Fault-avoidance Routing Method for Data Center Networks with Regular Topology

http://datatracker.ietf.org/doc/draft-sl-rtgwg-far-dcn/

Please send comments to rtgwg@ietf.org
Background

• With rapid development of cloud computing technologies, scale of a data center is growing up quickly.
• Traditional tree-like architectures and routing protocols are not suitable for building large-scale networks.

• Some new network architectures, such as Fat-tree, BCube, are applied to data center networks.

• To maximize benefits of new architectures, some new routing methods are proposed according to the features of Fat-tree, BCube’s topologies.
What is FAR

• FAR is a generic routing method and framework for large-scale data center networks.

• FAR protocol is well designed to fully leverage the regularity in the topology of networks.

• FAR is a high-performance routing method which computes routing tables in a simplistic manner.
Differences between FAR and other routing methods

- OSPF, IS-IS or RIP works in an arbitrary network, but FAR is designed for regular topologies.
  - A regular topology means the distribution of nodes, addressing and connections are well designed, so a node knows the whole topology without learning in a network.

- Other than some routing methods for specific networks such as Fat-tree and BCube, FAR is a generic routing method suitable for any network with a regular topology.
The Principle of FAR

• Network devices, including routers, switches, and servers, are assigned IP addresses according to their location in the network.

• A basic routing table (BRT) is built based on local topology.

• A negative routing table (NRT) is built based on link and device failures in the entire network.

• Look up both BRT and NRT to determine the final route in a routing procedure.

• Final routes = matched routes in BRT - matched routes in NRT.
The Routing Framework of FAR

1. Neighbor/Link Detection
2. Device Learning
3. Link Failure Inferring
4. Link Failure Learning
5. Building the Basic Routing Table
6. Building the Negative Routing Table
7. Querying Routing Table

Link state Learning -> Routing Table Building -> Routing Table Querying
Use Case (Fat-tree Network)
The BRT of aggregation switch 10.1.0.1

- It is easy to build a BRT for a router according to its local topology.
- We take 10.1.0.1 as an example. Its BRT is:

<table>
<thead>
<tr>
<th>Destination/Mask</th>
<th>Next Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1.1.0/255.255.255.0</td>
<td>10.1.1.1</td>
</tr>
<tr>
<td>10.1.2.0/255.255.255.0</td>
<td>10.1.2.1</td>
</tr>
<tr>
<td>10.0.0.0/255.0.0.0</td>
<td>10.0.1.1</td>
</tr>
<tr>
<td>10.0.0.0/255.0.0.0</td>
<td>10.0.1.2</td>
</tr>
</tbody>
</table>
The NRT of aggregation switch 10.1.0.1

- A router’s NRT is determined by locations of link or device failures in the network.

- Suppose the link between 10.0.1.2 and 10.3.0.1 fails, The NRT of 10.1.0.1 is:

<table>
<thead>
<tr>
<th>Destination/Mask</th>
<th>Next Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.3.0.0/255.255.0.0</td>
<td>10.0.1.2</td>
</tr>
</tbody>
</table>
Node 10.1.0.1 forward a packet to node 10.3.1.3

- 1) Calculate candidate hops. 10.1.0.1 looks up its BRT and obtains the following matched entries:

<table>
<thead>
<tr>
<th>Destination/Mask</th>
<th>Next Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.3.0.0/255.255.0.0</td>
<td>10.0.1.1</td>
</tr>
<tr>
<td>10.3.0.0/255.255.0.0</td>
<td>10.0.1.2</td>
</tr>
</tbody>
</table>

So the candidate next hops = \{10.0.1.1; 10.0.1.2\}.
2) Calculate avoiding hops. 10.1.0.1 looks up its NRT and obtains the following matched entries:

<table>
<thead>
<tr>
<th>Destination/Mask</th>
<th>Next Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.3.0.0/255.255.0.0</td>
<td>10.0.1.2</td>
</tr>
</tbody>
</table>

So the avoiding hops = \{10.0.1.2\}

3) Calculate applicable hops.

applicable next hops = \{10.0.1.1; 10.0.1.2\} – \{10.0.1.2\} = \{10.0.1.1\}

4) Finally, forward the packet to the next hop 10.0.1.1.
Advantages of FAR

• FAR is a generic routing method suitable for most data centers with regular topologies.

• FAR is a high-performance routing method which supports very large-scale networks.

• A FAR switch is simple and cheap, so it can lower the constructing and operating cost of a data center.
Drawback of FAR and future work

- FAR in this proposal doesn’t give an universal method to calculate routing tables for various of network topologies. We should design different method for each type of topology.

- Now we are solving the problem above. We have invented a TDL (topology definition language) to describe a regular topology, and based on TDL, we can design an universal method to calculate routing tables for FAR switches.
Requested actions from the WG

• Routing methods based on regular topology have great advantages in large-scale next-generation data centers.

• In the past, no draft has discussed routing problem in regular network topology in Data Centers.

• All we need to do now is to propose the problems in the IETF.

• Requesting IETF Rtg WG to consider adoption of this draft and then standardize the solutions.
Thanks and Q&A!