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

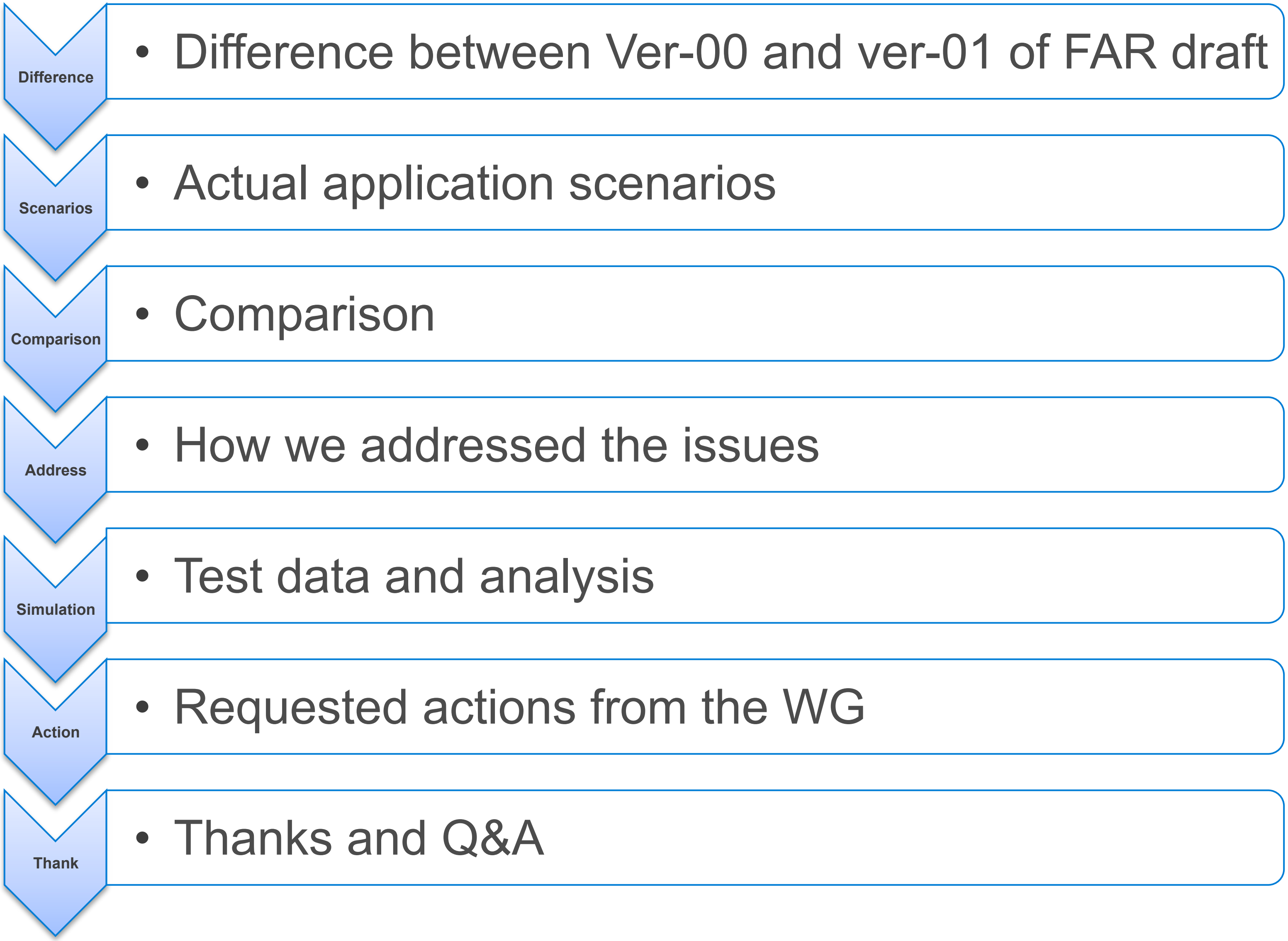
http://datatracker.ietf.org/doc/draft-sl-rtgwg-far-dcn/

Bin Liu, ZTE
Yantao Sun, Jing Cheng, Yichen Zhang
Beijing Jiaotong University

Please send comments to rtgwg@ietf.org

IETF90,Wednesday, 23 July 2014
Fairmont Royal York Hotel
100 Front Street W
Toronto, Ontario, Canada M5J 1E3

Outline



Difference
- Difference between Ver-00 and ver-01 of FAR draft

Scenarios
- Actual application scenarios

Comparison
- Comparison

Address
- How we addressed the issues

Simulation
- Test data and analysis

Action
- Requested actions from the WG

Thank
- Thanks and Q&A

Difference between Ver-00 and ver-01 of FAR draft

- Add description of the actual application scenarios in regular topological with rapid routing protocol(HPC/InfiniBand, FAR, SRP, etc).

- Add description of comparison between FAR and OSPF on why OSPF and other conventional routing methods do not work well in a large-scale network with several thousands of routers?

Actual application scenarios --HPC/InfiniBand System with Better Protocol on Regular network topology

- There are many real world scenario where tens of thousands of nodes(or much more nodes) need to be deployed in a flat area, such as infiniband routing and switching system, high-performance computer network, and many IDC networks in China.
- The similar problems have been existed long ago. People have solved the problems through similar solutions, such as the traditional regular topology-based RFC3619 protocol, the routing protocols of IB routing and switching system, and high-performance computer network routing protocol.

HPC system

InfiniBand System

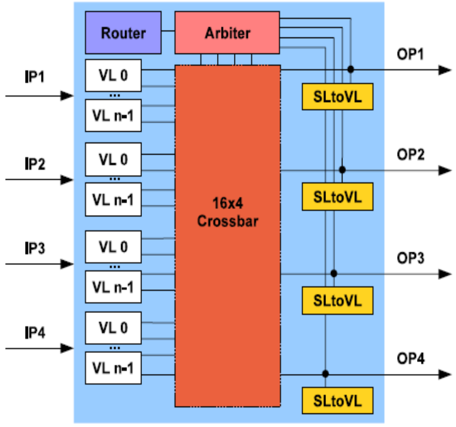


Fig. 4. Assumed InfiniBand switch architecture with 4 ports.

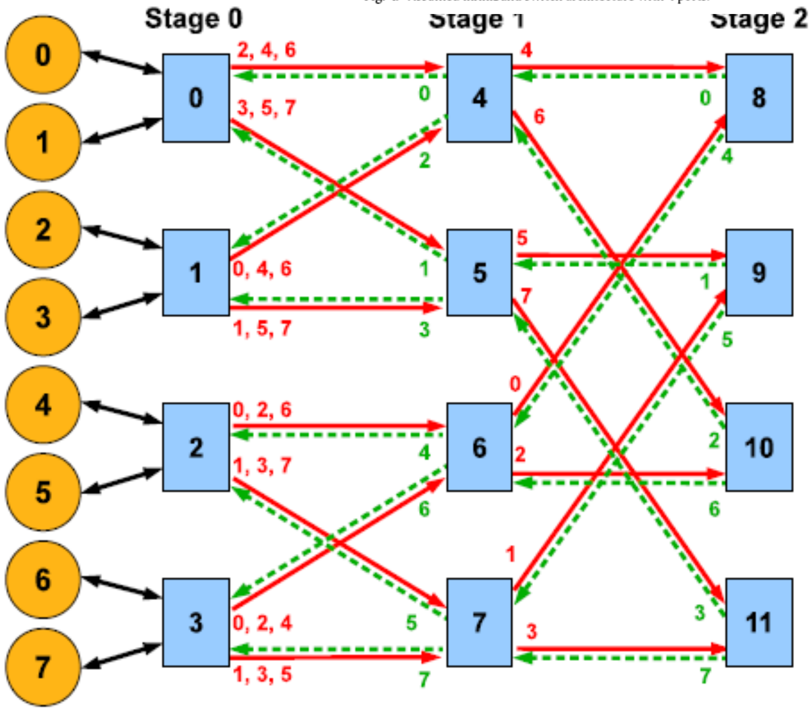


Fig. 3. DESTRO routing algorithm in a 2-ary 3-tree; upward links as solid arrows, downward links as dotted.

InfiniBand System with Protocol on Regular network topology

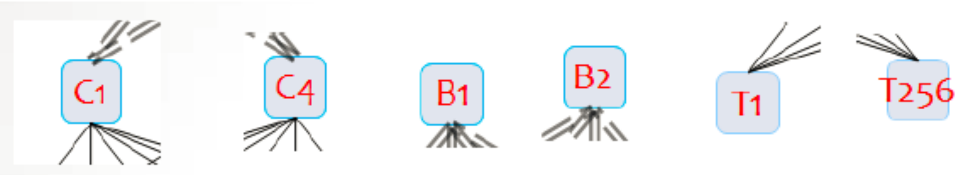
- Scalability of InfiniBand is very high, can support tens of thousands of nodes in a sub network, and each network can have thousands of subnets, each mounting system can have multiple network structure.They have already been used in many large data centers.
- Infiniband network uses the simplified routing protocols, in order to run routing protocol efficiently in the regular network topology. Routing algorithms, based on the InfiniBand, is the routing algorithm based on specific topology, for example, the deterministic e-cube algorithm [11], and its extension under the torus [12, 13] and m-port n-tree routing algorithm (14 to 15), the Up*Down*routing algorithm and its improved algorithm, etc. These algorithms take into account the characteristics of the specific topology structure.

Actual application scenarios – Online Sequoia Routing Protocol on Regular network topology in IDC of Tencent Inc.

- In the large-scale IDC architecture (more than 20,000 switches), routings are simple and controllable.
- reduce the burden of routing protocol (e.g., OSPF) on the control plane.
- convergence time does not increase with the scale growth.
- Rely on regular network topology, split horizon principle to avoid the loop.

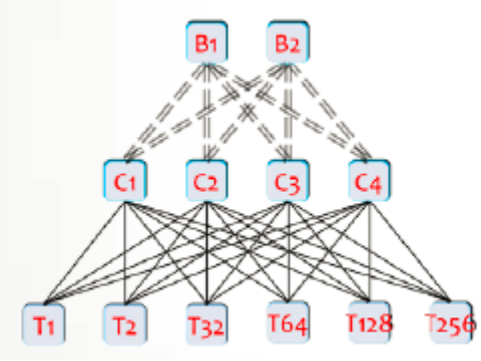
Real data from China's largest IDC laboratory	**OSPF**	**BGP**	**Regular Topology Route Protocol(FAR,etc)**
Convergence time	> 2s	< 1s	< 1s
Scale sensitivity	Performance deterioration with the scale growth	Not sensitive	Not sensitive
Code lines	500000 lines level	300000 lines level	Several thousand lines

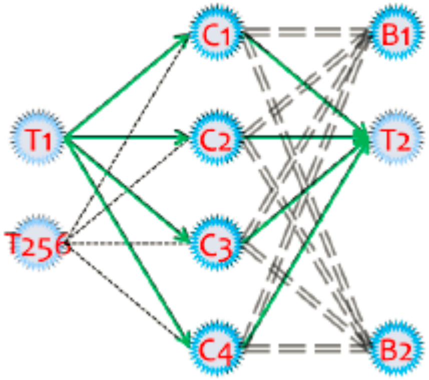
Comparison between FAR and OSPF (1)



1. Generate and globally synchronize LSA

2. The LSAs Are spliced into topology





3. Calculate the shortest paths

As the network scale extends, LSA synchronization and splicing performance deteriorate, paths are difficult to be controlled.

As everyone knows, the OSPF protocol uses multiple databases, more topological exchange information (as seen in the following example in Page9) and complicated algorithm. It requires routers to consume more memory and CPU processing capability. But the processing rate of CPU on the protocol message per second is very limited. When the network expands, CPU will quickly approach its processing limits, and at this time OSPF can not continue to expand the scale of the management. The SPF algorithm itself does not thoroughly solve these problems. On the contrary, FAR… (to be continued…)

Comparison between FAR and OSPF (2)

On the contrary, FAR does not have the convergence time delay and the additional CPU overheads, which SPF requires. Because in the initial stage, FAR already knows the regular information of the whole network topology and does not need to periodically do SPF operation.

Conclusion:

- FAR can manage much larger network than OSPF.
- FAR responds to the network much faster than OSPF.

Comparison between FAR and OSPF (3)

One of the examples of "more topological exchange information":

In the OSPF protocol, LSA floods every 1800 seconds. Especially in the larger network, the occupation of CPU and band bandwidth will soon reach the router's performance bottleneck.

In order to reduce these adverse effects, OSPF introduced the concept of Area, which still has not solved the problem thoroughly). By dividing the OSPF Area into several areas, the routers in the same area do not need to know the topological details outside their area. (In comparison with FAR, after OSPF introducing the concept of Area, the equivalent paths cannot be selected in the whole network scope)

Comparison between FAR and OSPF (4)

OSPF can achieve the following results by Area :

1) Routers only need to maintain the same link state databases as other routers within the same Area, without the necessity of maintaining the same link state database as all routers in the whole OSPF domain.

2) The reduction of the link state databases means dealing with relatively fewer LSA, which reduces the CPU consumption of routers;

3) The large number of LSAs flood only within the same Area.

But, its negative effect is that the smaller number of routers which can be managed in each OSPF area.

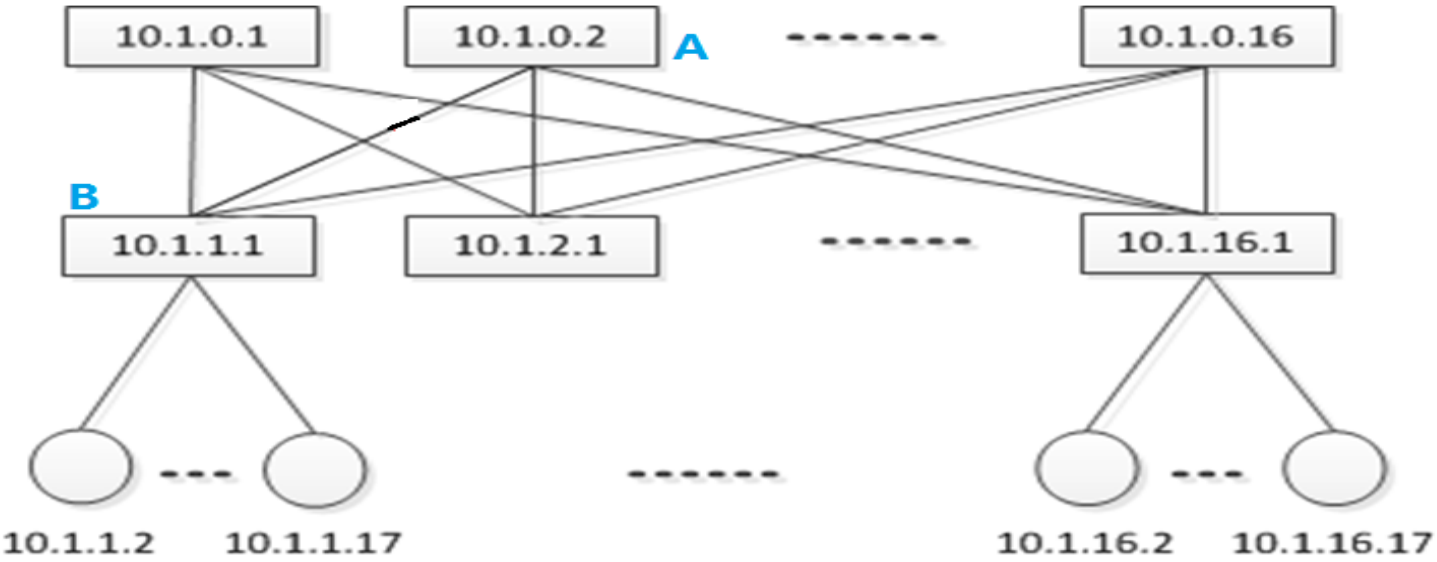
On the contrary, because FAR does not have the above disadvantages, **FAR can also manage large-scale network even without dividing Areas**.

Comparison between FAR and OSPF (5)

The aging time of OSPF is set in order to adapt to routing transformation and protocol message exchange happened frequently in the irregular topology. Its negative effect is:

when the network does not change, the LSA needs to be refreshed every 1800 seconds to reset the aging time. In the regular topology, as the routings are fixed, it does not need the complex protocol message exchange and aging rules to reflect the routing changes, as long as LFA mechanism in the FAR is enough.

Comparison between FAR and OSPF (6)



- **For example**, the router A occurs software failure and cannot send protocol packets, but all physical links are intact. In the OSPF protocol, the aging mechanism is needed to update the A-related routes. However, this processing is unnecessary in the regular topology, because even if the router A occurs software failure and cannot send protocol packets, but the physical link is normal, and there is no any change in the neighbor relationship of A, so there is no need for such a routing update, because the routes have not been changed. The software failure can be discovered by checking port packet statistic and resolved by network management software.

Comparison between FAR and OSPF (7)

- Therefore, in FAR, we can omit many unnecessary processing and the packet exchange. The benefits are fast convergence speed and much larger network scale than other dynamic routing protocol.
- Now there are some successful implementations of simplified routings in the regular topology in the HPC environment.
- **Conclusion**:
  As FAR needs few routing entries and the topology is regular, the database does not need to be updated regularly. Without the need for aging, there is no need for CPU and bandwidth overhead brought by LSA flood every 30 minutes, so the expansion of the network has no obvious effect on the performance of FAR, which is contrary to OSPF.

Comparison between FAR and OSPF (8)

- Comparison of convergence time:
- The settings of OSPF spf_delay and spf_hold_time can affect the change of convergence time. The convergence time of the network with 2480 nodes is about 15-20 seconds(as seen in the following pages); while the FAR does not need to calculate the SFP, so there is no such convergence time.
- These issues **still exist** in rapid convergence technology of OSPF and ISIS (**such as I-SPF**). The convergence speed and network scale constraint each other. FAR does not have the above problems, and the convergence time is almost negligible.

How we addressed the issues

(1)'Data Centers are using existing protocols without problems.'

The similar problems have been existed long ago. People have solved the problems through similar solutions, such as the traditional regular topology-based RFC3619 protocol, the routing protocols of IB routing and switching system, and high-performance computer network routing protocol.

(2) 'Network convergence doesn't follow link state dynamics - Fast reroute exists.'

These issues still exist in rapid convergence technology of OSPF and ISIS (such as I-SPF). The convergence speed and network scale constraint each other. FAR does not have the above problems, and the convergence time is almost negligible.

(3)' First, you need to provide a real world scenario where tens of thousands of nodes need to be deployed in a flat area. Secondly you need to describe why the current IGPs would not be able to do the job or be improved to do it.'

First question , please refer from page4 to page6, Second question, please refer from page7 to page14.

(4)‘seem to ignore the point about fast reroute (FRR).’

Can FRR solve these problems? IP FRR has some limitations. The establishment of IP FRR backup scheme will not affect the original topology and traffic forwarding which are established by protocol, however, we can not get the information of whereabouts and status when the traffic is switched to an alternate next hop. And FRR has the following disadvantages:

1)Additional protocol overhead: For the protection of links, nodes and paths, it is necessary to set up a backup LSP respectively, which causes unnecessary overhead and complex protocol processing; (there is no such protocol overhead for FAR, and because FAR is based on regular topology, path protection&switching process are simple.)

2) Backup LSP failures may exist. As there is no protection mechanism, it cannot fast reroute when it fails; (FatTree network architecture has multiple natural selection.)

3) There is a linear correlation between the **beginning** convergence time and the node number of topologies, while it is not sensitive to the FAR.

4) LDP FRR cannot guarantee that the calculated path is the optimal path, leading to the emergence of new link congestion. But FAT TREE architecture network is a non-blocking network.

How we addressed the issues

(5) ‘First, links have high bandwidth, CPUs are fast and any serious IGP implementation has addressed the bottlenecks you are talking about.’

In the IDC network, though the links between the network device nodes have greatly improved their speeds and the performance of the CPUs has significantly increased, frequently updated data needs to be switched repeatedly between user mode memory and kernel mode memory, in order to maintain the consistency of multiple nodes. As well as the complexity of the TCP/IP protocol parsing and processing, the frequent context switching wastes CPU processing power and the efficient use of memory. Simplifying routing protocol, therefore, to reduce unnecessary protocol message processing, can significantly improve the network performance of each network node, and, more importantly, improve the marginal revenue and the marginal effect. And example can be finded in rtgwg maillist.

Test data and analysis

OSPF
ergence Time - 20 rc

OSPF
rgence Time - 320 r

- These are the test results of two OSPF running, one is a scenario with 20 routers, and another is a scenario with 320 routers. We may draw the conclusion from the test results as follows:

1) With the increase of the number of areas, beginning convergence time will increase (from 25s to 30 s).However, the increase of the scale has no significant impact on the re-convergence time due to subsequent link down problem, because the number of nodes within the area is limited. If the number of nodes within the area increases, LSA synchronization and splicing performance deteriorate, the paths are difficult to be controlled, and the convergence time will increase significantly. This is why the number of nodes in an OSPF domain cannot be too large.

2) Because of a large number of packets transmitted, the network convergence processing, every time the network changes, will have great impact on network performance.

Requested actions from the WG

- 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 RtgWG to consider adoption of this draft and then standardize the solutions.

Thanks and Q&A!