Technical Disclosure Commons

Defensive Publications Series

December 08, 2017

# Minimal Restriping For Data Center Expansion

Shizhen Zhao

# MINIMAL RESTRIPING FOR DATA CENTER EXPANSION

## ABSTRACT

A system and a method are proposed to minimize restriping during data center expansion. Minimal restriping technique takes into account the old data center topology, and aims to minimize restriping between the new topology and the old topology.  A low complexity minimal restriping algorithm proposed is used to minimize restriping. The topology solution is also guaranteed to have high network capacity, because it satisfies superblock ports, spineblock ports,and superblock-spineblock level balancedness constraints. The constraints may consider the number of links in original topology and the algorithm may determine the optimum new topology  that minimizes restriping. The disclosed technique may not cause significant bandwidth reduction, and thus greatly shortens the data center expansion time.

## BACKGROUND

A data center is a facility composed of networked computers and storage that is used to organize process, store and disseminate large amounts of data. There are mainly two types of data center expansions. The first type may involve adding a new superblock (SB) when more servers are required by the users for their services. The second type may involve upgrading an existing superblock when the bandwidth utilization of the superblock is very high. Data center expansion is necessary to support the ever-growing user demand. Currently, data center expansion incurs high operational cost and bandwidth reduction, as all the data center links need to be restriped during expansion. If all the links were restriped together, the entire data center bandwidth may be lost. One can reduce the bandwidth impact by performing each expansion in multiple stages but this will take a longer time. Currently, a new data center topology is generated independent of the old topology during each expansion. As a result, the new topology is very different from the old topology, which leads to fabric-wide restriping.

## DESCRIPTION

The disclosure proposes a system and a method for data center expansion. The system and method aims to minimize restriping for data center expansion. Minimal restriping approach takes the old topology into account, and minimizes restriping between the new topology and the old topology. The most straightforward approach for minimal restriping is to introduce an optimization goal. However, this optimization goal introduces significant complexity for the data center topology solver. Thus, a low-complexity algorithm with close-to-optimal restriping ratio is proposed. The disclosed technique significantly reduces the number of links to be restriped during expansion, and thus greatly shortens the data center expansion time.

The low complexity minimal restriping algorithm proposed is used to compute a topology with low restriping ratio. In addition, the new topology is guaranteed to have high network capacity, because it satisfies the constraints that all superblock ports are connected and are evenly distributed among all spineblocks. Additional constraints, e.g., middleblock-spineblock level balancedness constraints and cage level balancedness constraints, can also be incorporated in the low complexity minimal restriping algorithm. The disclosed technique could significantly reduce the network impact during expansion, and thus allow network expansion being done in fewer number of stages.

The concept of minimal restriping for data center expansion by addition of a new superblock is illustrated with an example. The original data center topology for the example as shown in FIG. 1 has three superblocks, three spineblocks (SP) and four PPs (optical circuit switches) initially. In the physical topology, each superblock/spineblock has 4 links, each of which connects to one PP. In contrast, the logical topology has 2 links for block pairs (SB1, SP1), (SB2, SP2), (SB3, SP3), and 1 link for other block pairs. In the data center topology, all the PP-facing superblock ports are connected to some spine ports to provide the maximum possible network bandwidth for each superblock.

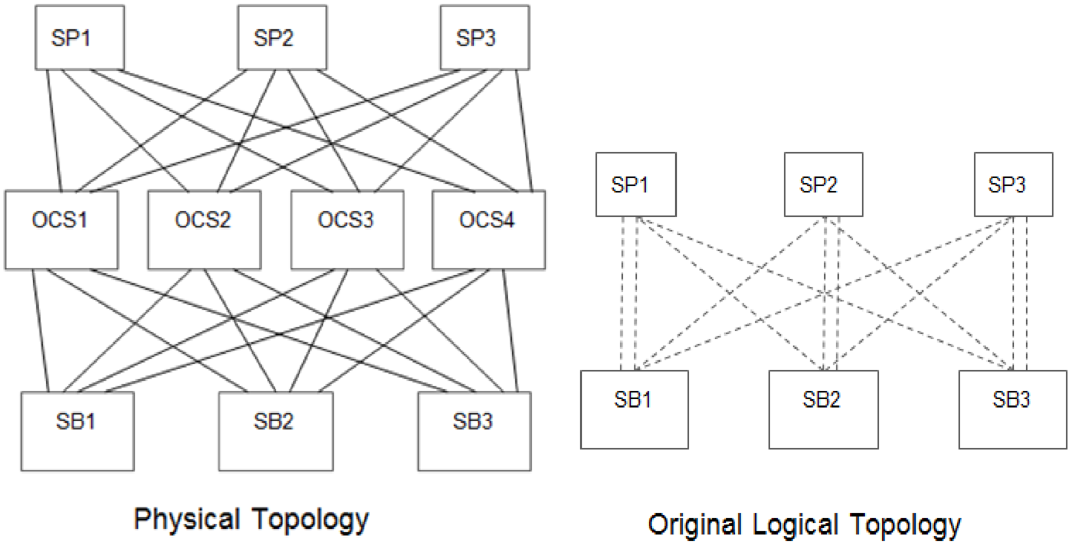
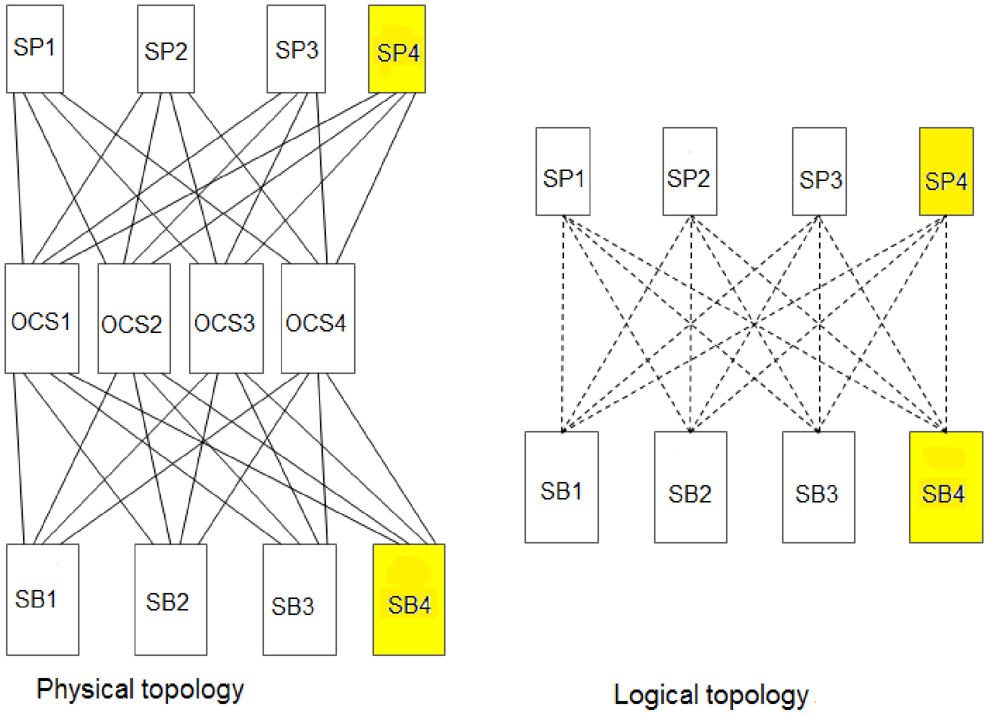


FIG. 1: Original topology

Further, data center is expanded by adding one superblock (SB4) and one spineblock (SP4) to the existing data center. As shown in FIG. 2, both SB4 and SP4 have 4 links, each of the links connects to one PP in the physical topology and the superblock-spineblock pair have exactly one link in logical topology. The PPs has to be configured to interconnect superblock ports and spineblock ports to create a real data center topology. Physical topology is fixed after a superblock and spineblock are added to a data center while logical topology will change based on the changes in PPs configurations.

To ensure high network capacity, all the superblock ports of the post-expansion topology must be connected, and the uplinks of each superblock must be evenly distributed among all the spineblocks. An example post-expansion logical topology is shown in FIG. 2. It is easy to verify that all the superblock ports are connected and the superblock-spineblock-level balancedness constraints are satisfied.



Physical topology

Logical topology

FIG. 2: Post-expansion physical topology and logical topology

The logical topology in Fig. 2 can be reached by minimal restriping. Since the total number of links between SB1-SB3 and SP1-SP3 is reduced from 12 to 9, a minimum of 3 links have to be removed from the original logical topology. This lower bound on the number of links removed is achievable. The PP reconfiguration strategy as shown in FIG. 3 has exactly 3 grey links indicating links removed from the original topology. Black links indicate existing links in the original topology and red links indicate the new topology. This verifies that the final logical topology obtained by minimal restriping is exactly the same logical topology as shown in FIG. 2.

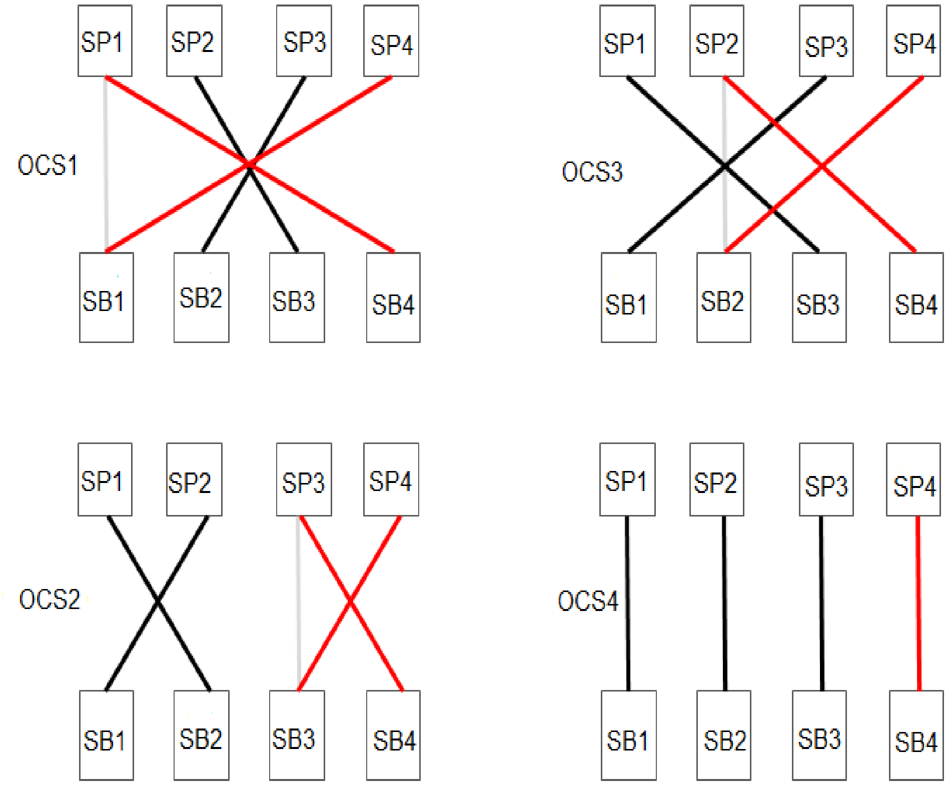


FIG. 3: PP Reconfiguration

An example on minimal restriping is provided. Next is discussed the general case. The minimal restriping problem, is first rigorously formulated and then, an efficient algorithm for minimal restriping is proposed.

The mathematical formulation of minimal restriping is described as follows. Assume that there are M superblocks after expansion, labeled as $L_1$, $L_2$, …, $L_M$, and N spineblocks after expansion, labeled as $R_1$, $R_2$, …, $R_N$ and K PPs, denoted by $O_1$, $O_2$, …, $O_K$. Both superblocks and spineblocks connect to PPs. The number of physical links between the superblock $L_m$ and the

PP $O_k$ are represented by $l_{m,k}$ (m=1~M, k=1~K). The number of physical links between the spineblock $R_n$ and the PP $O_K$ are denoted by $r_{n,k}$ (n=1~N, k=1~K) . The total number of physical links of the superblock $L_m$ and spineblock $R_n$ are represented by $l_m=l_{m,1}+l_{m,2}+...+l_{m,K}$ and $r_n=r_{n,1}+r_{n,2}+...+r_{n,k}$ respectively.

The total number of logical links between the superblock $L_m$ and the spine block $R_n$ through the PP $O_k$ is denoted by $d_{m,n}^{k}$. The data center topology is determined by $d_{m,n}^{k}$ .The following constraints must be satisfied by $d_{m,n}^{k}$ :

All the superblock ports in each PP must be connected:

$\sum_{n=1}^{N} \quad d_{m,n}^{k} = l_{m,k}$----------------------------------------(1)

The Spineblock ports in each PP cannot be overloaded:

$\sum_{n=1}^{N} \quad d_{m,n}^{k} \leq r_{n,k}$------------------------------------------(2)

Superblock-Spineblock level balancedness constraints:

$p_{m,n} \leq \sum_{k=1}^{K} \quad d_{m,n}^{k} \ \leq q_{m,n}$ ---------------------------------(3)

Where $p_{m,n} = \frac{l_m r_n}{(r_1+r_2+\cdots .R_n})$ and $q_{m,n} = p_{m,n} + 1$

Let $b_{m,n}^{k}$ represents number of links in the original topology between the superblock $L_m$ and the spineblock $R_n$ through the PP $O_k$. If the superblock $L_m$ or the spineblock $R_n$ does not exist in the original data center topology, then $b_{m,n}^{k} = 0$. The objective function is given as

$\sum_{m=1}^{M} \quad \sum_{n=1}^{N} \quad \sum_{k=1}^{K} \quad (b_{m,n}^{k} - d_{m,n}^{k})^{+}$ --------------------------------(4)

The objective function (4) calculates the total number of the original links that need to be disconnected in order to migrate to the new data center topology. It has to be minimized to minimize restriping. The factor $d_{m,n}^{k}$ that minimizes the objective function is to be determined. In order to find a solution for $d_{m,n}^{k}$ that minimizes restriping, two methods may be used. The first method involves a high-complexity algorithm that adds the objective function (4) to the integer linear programming problem (1)-(3) and is given as

$$min \sum_{m=1}^{M} \quad \sum_{n=1}^{N} \quad \sum_{k=1}^{K} \quad (b_{m,n}^{k} - d_{m,n}^{k})^{+} \text{ subject to } (1)(2)(3)$$

The second method involves a low complexity algorithm without an objective function for minimal restriping that involves finding $d_{m,n}^{k}$ satisfying the equations (1)-(3) and the following two constraints:

$$b_{m,n}^{k} \ \leq d_{m,n}^{k} \text{ if } \sum_{k=1}^{K} \quad b_{m,n}^{k} < q_{m,n}\text{-------------------------------------(5)}$$

$$b_{m,n}^{k} \ \geq d_{m,n}^{k} \text{ if } \sum_{k=1}^{K} \quad b_{m,n}^{k} > p_{m,n}\text{-----------------------------------(6)}$$

The constraints (5) and (6) are generated based on the Superblock-Spineblock level balancedness constraints (3). The intuition of (5) and (6) is that if the total number of existing links in the original topology is less than the upper bound $q_{m,n}$ in (3), then we should increase the number of existing links $b_{m,n}^{k}$ as we do in (5); otherwise, we should decrease if the number of existing links $b_{m,n}^{k}$ as we do in (6). Even though the low complexity algorithm does not have any objective function, (5) and (6) guarantees low restriping ratio. In addition, (5) and (6) dramatically reduces solver complexity because the searching space is significantly reduced by (5) and (6).

Further, the performance of the low-complexity minimal restriping solver is evaluated. The method for evaluating the performance of the restriping solution is illustrated in FIG. 4. The first step involves setting superblock, spineblock and OCS. Superblock settings are made based on the upgradation required and depending on the total number of superblock ports in each configuration, spineblocks are added. The next step involves enabling superblock-spineblock level constraints, middleblock-spineblock level constraints, and superblock-spineblock-cage level constraints. Although only superblock-spineblock level constraint is mentioned in this application, our low complexity algorithm can easily incorporate additional constraints.

The middleblock-spineblock level constraint indicates that the uplinks of each middleblock must be evenly distributed among all the spineblocks. The superblock-spineblock-cage level balancedness constraint indicates that the total number of links between each superblock and each spineblock should be evenly distributed between different OCS cages. Then, a topology satisfying all the constraints is determined. The superblocks are then upgraded using the minimal restriping algorithm. Finally, the restriping ratio (total restriping/number of links in the pre-expansion topology) is computed and performance is evaluated. The solution that may give lowest restriping ratio is preferred. Practically, when multiple balancedness constraints are present, any balancedness constraint may be used to cut the searching space. This may improve the coverage of the low-complexity minimal restriping solver.

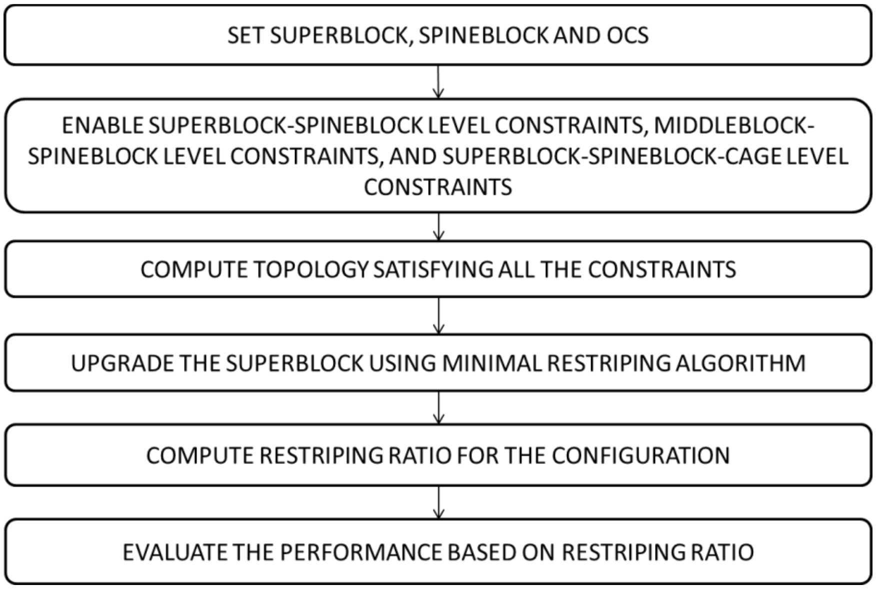

SET SUPERBLOCK, SPINEBLOCK AND OCS
ENABLE SUPERBLOCK-SPINEBLOCK LEVEL CONSTRAINTS, MIDDLEBLOCK-SPINEBLOCK LEVEL CONSTRAINTS, AND SUPERBLOCK-SPINEBLOCK-CAGE LEVEL CONSTRAINTS
COMPUTE TOPOLOGY SATISFYING ALL THE CONSTRAINTS
UPGRADE THE SUPERBLOCK USING MINIMAL RESTRIPING ALGORITHM
COMPUTE RESTRIPING RATIO FOR THE CONFIGURATION
EVALUATE THE PERFORMANCE BASED ON RESTRIPING RATIO

FIG. 4: Flow chart for performance evaluation

The disclosed technique may not cause significant bandwidth reduction, and thus greatly shortens the data center expansion time. The low complexity minimal restriping algorithm does not introduce any objective function, and thus has much shorter solver running time as the integer programming solver only needs to solve one SAT problem. The additional constraints (5) and (6) in the algorithm reduce the range of decision variables, and thus also make it easier to search for a solution.