Anycast (re)routing of multi-period traffic in dimensioning resilient backbone networks for multi-site data centers

Ting Wang[1], **Brigitte Jaumard**[1], **Chris Develder**[2]
[1] CSE, Concordia University, Montréal (QC) H3G 1M8 Canada
[2] INTEC – IBCN, Ghent University – iMinds, Ghent, Belgium

Abstract

We consider the problem of dimensioning resilient backbone networks for cloud-like scenarios where demand is to be served at one among several candidate data centers (DCs), and where that demand varies over time, which we assume to be slotted. We thus consider multi-period traffic, for which we need to find routes to both a primary DC and a backup DC (in case the primary, or the network connection towards it, fails) and account also for synchronization traffic (following its own routes) between the two chosen DCs. We propose a path formulation and adopt a column generation approach: the (restricted) master problem (RMP) selects "configurations" to use for each demand in each of the time periods, while pricing problems (PPs) construct new, potentially cost-reducing configurations for a given demand. Our model allows for several PPs to be solved in parallel, and we demonstrate the time savings achieved by doing so. We compare several anycast (re)routing strategies, where we allow traffic that spans multiple periods to either (i) not be rerouted in different periods, (ii) only change the backup DC and routes, or (iii) freely change both primary and backup DC choices and routes towards them.

Keywords: Network Virtualization; End-to-End Resilience; Cloud Computing; Anycast Resilience; Multi-period Virtual Network Mapping.

1. Introduction

With the proliferation of cloud computing, today's businesses and end users increasingly rely on data centers (DCs) that serve their applications and content. This evolution has been enabled by the underlying optical network technology [1], as it supports high bandwidths and low latencies. Because of the latter, the exact location of the data center does not matter much to the user it serves. Thus, to serve requests, service providers can adopt anycast routing: they may choose what particular DC to use among several that are distributed over multiple of their operating sites. Furthermore, this anycast principle can also be used for resiliency: if a failure affects either the DC or the (optical) network connection towards it, backup can be provided at an alternate location. Previous work has studied the potential benefits (in terms of server and network resource capacity) of exploiting anycast in network dimensioning problems with static traffic (e.g., [2]).

The current work explicitly considers time-varying traffic (more specifically, we assume that routing and DC assignment can be revised at discrete points in time, i.e., for multiple consecutive periods among which the volume of service requests may vary). We build on our latest work on this topic [3], and contribute with (i) a new column generation model that is path-based (rather than link-based), (ii) an exploration of the effect of parallel execution of multiple so-called pricing problems (PPs), and (iii) more extensive experiments that also consider variations in the choice of DC locations.

2. Problem statement and mathematical model

2.1. Problem statement

We consider a network dimensioning problem spanning multiple time periods, where we are *given*

- The network topology, in terms of network nodes (e.g., OXCs) and links interconnecting them, as well as the subset of nodes that host data centers (DCs), and
- The service requests, with a given source node and the number of unit demands (where a unit represents an amount of bidirectional bandwidth to provide from the source to a DC to be chosen), that are started in each of the consecutive discrete time periods. Each service request also has its specific duration (aka holding time) of 1 or more consecutive periods.

We then need to *find* for each request, and each time period it lasts, the routes from its source to a selected DC, *such that* the total amount of required network resources (i.e., the bandwidth crossing each link, summed over all links) is minimized and each request remains operational under given failure scenarios. For the latter, we will require requests to be protected against single link or single DC failures.

2.2. VNO-resilience

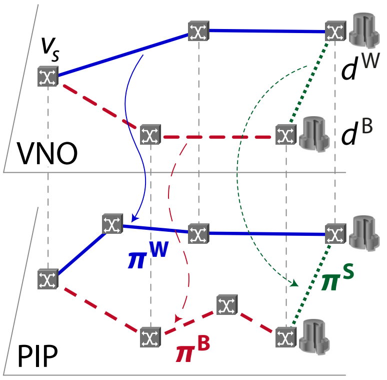


Fig. 1: The VNO-resilience scheme.

We will adopt a VNO-resilience scheme to achieve that requirement. The idea thereof, as previously described in, e.g., [3], is sketched in Fig. 1: it provides 1:1 protection routing in the VNet for network failures, where the working and protection paths of a service have to be physically link/node disjoint. The working path (π^{W}) routes the services from their source node (v_S) towards the primary DC (d^{W}), the protection path (π^{B}) towards the backup DC (d^{B}), while π^{W} and π^{B} are disjoint in their physical layer mapping. In addition, a synchronization path (π^{S}) is established in order to handle migration and failure routing requirements when a DC failure occurs: services then need to be rerouted from the primary d^{W} to backup d^{B}. Further, we assume that there is an automatic switch-back to the original network path and DC once a fault is repaired, and therefore we allow reusing the same network/DC capacity to protect against other failures: backup capacity is shared. Under the assumptions that (A1) the backup DC has a different location than the primary DC, (A2) π^{W} and π^{B} are link disjoint and, (A3) π^{W} and π^{S} are link disjoint, protection is guaranteed against any single link failure and any single DC failure.

2.3. Mathematical model

We here describe only the general ideas of our column generation model, the actual formulas can be found in, e.g., [4] (which is available upon request from the authors). Following the general column generation approach, the overall optimization problem is split into a (restricted) master problem (RMP), and a pricing problem (PP). The RMP's task is to find the optimal combination out of a set of candidate "configurations", C, while the PP's task is to find/create new configuration(s) that, when added to the pool C for the RMP leads to a solution with a better objective value. More concretely, a "configuration" in our model is associated with a given source node v_{S}, and comprises a set of three paths as sketched in Fig. 1: one primary path π^{W} originating at v_{S} towards a primary data center d^{W}, one backup path π^{B} originating at v_{S} towards a primary data center d^{B}, and one synchronization path π^{S} between the primary and the backup data center.

Compared to [3], our new model improves scalability by (i) aggregating all unit demands from a single source node and with the same holding time into a single request, and (ii) check for reconfigurations (which we want to minimize, as a secondary objective next to the total amount of network resources) based on paths rather than links. For details, we refer to [4, Section 4.4].

3. Solution strategy: Serial vs parallel

For our given column generation model, we have various options to decide what PP(s) to solve and find new configuration(s) to add before solving the RMP (with an extended configuration set) again. One strategy is to add one new configuration at a time (for a selected source node, e.g., in round robin fashion). Alternatively, we can solve multiple PPs in parallel (e.g., one for each different source node). We will explore the benefits of such a parallel strategy, sketched in Fig. 2.

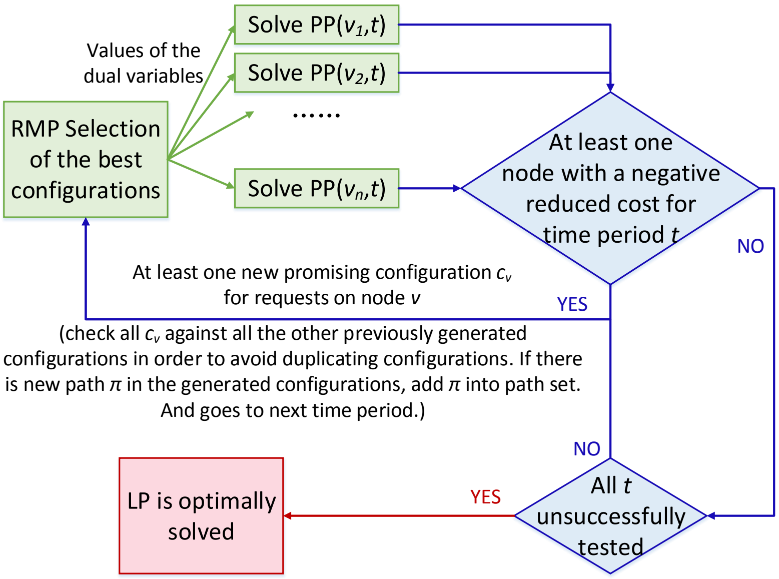

Solve PP(v_1,t)
Solve PP(v_2,t)
……
Solve PP(v_n,t)
Values of the dual variables
RMP Selection of the best configurations
At least one node with a negative reduced cost for time period t
NO
YES
At least one new promising configuration c_v for requests on node v
(check all c_v against all the other previously generated configurations in order to avoid duplicating configurations. If there is new path π in the generated configurations, add π into path set. And goes to next time period.)
NO
All t unsuccessfully tested
YES
LP is optimally solved

Fig. 2: The parallel column generation solution scheme.

4. Case study

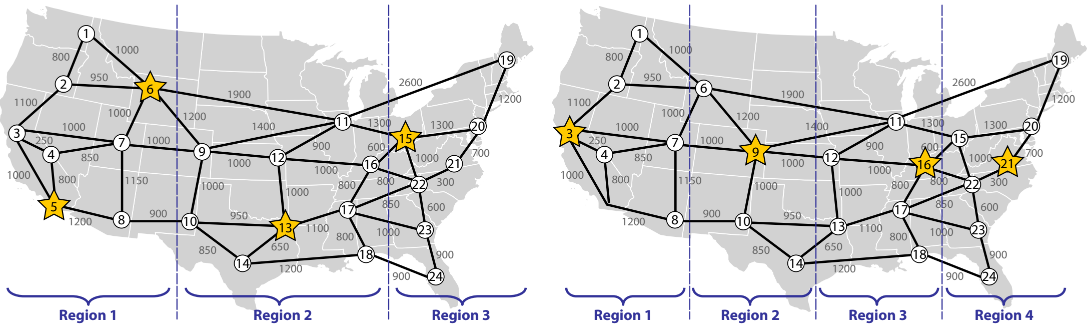
4.1. Experiment setup

We consider a 24-node US network topology, with 43 undirected links, as illustrated in Fig. 3. Given that we adopt anycast routing, traffic is specified in terms of it source only. The amount of traffic generated varies in terms of time-of-day (i.e., period), as well as per region (which are assumed to have distinct, artificial time zones). The split of the total traffic volume across the various regions is as follows: (i) for the 3-region case, 33.3% originates from Region 1, 37.5% from Region 2, and the remaining 29.2% from Region 3, (ii) for the 4-region case, 29.2% originates from Region 1, 16.6% from Region 2, 25% from Region 3, and 29.2% from Region 4. For a given Region, the traffic varies during the day, with 48% of the Region's traffic in 8 am–4 pm, 38% in 4 pm–12 am, and the remaining 14% from 12 am–8 am. The volume of traffic, generated in one of the 3 time periods and a given Region, is further divided in a portion that just lasts that single period and the other portion that will continue into the next period. We consider three patterns, with respectively 20%, 50% or 80% of two-period traffic.

4.2. Resource savings

The first and foremost research question we wanted to address is: What bandwidth savings can we achieve by allowing rerouting (part of) the traffic from one period to the next? Qualitatively, we indeed do expect to see potential (overall) bandwidth savings if we are more flexible, i.e., if we go from Scenario I (no rerouting allowed), over Scenario II (allow changes in backup and/or synchronization paths), to Scenario III (also allow to change the primary working path from one period to the next). Intuitively, we expect to see potentially more substantial savings if the volume of the traffic that we are allowed to reroute increases (e.g., going from Pattern #1, with just 20% of two-period traffic, to Pattern #2 with 80% two-period traffic). The experiments investigate such savings quantitatively, for the two network topologies we defined above.

Figure 4 shows the relative difference in bandwidth requirements for the various rerouting scenarios. From these numerical results, we learn that, compared to the baseline Scenario I, the total bandwidth cost is reduced with on average 5.1% (resp. 6.4%) for Scenario II (resp. Scenario III) with traffic Pattern #1, and by 6.9% (resp. 8.2%) with Pattern #2 (where the average is taken over all traffic volumes). This



(a) 3-Region topology

(b) 4-Region topology

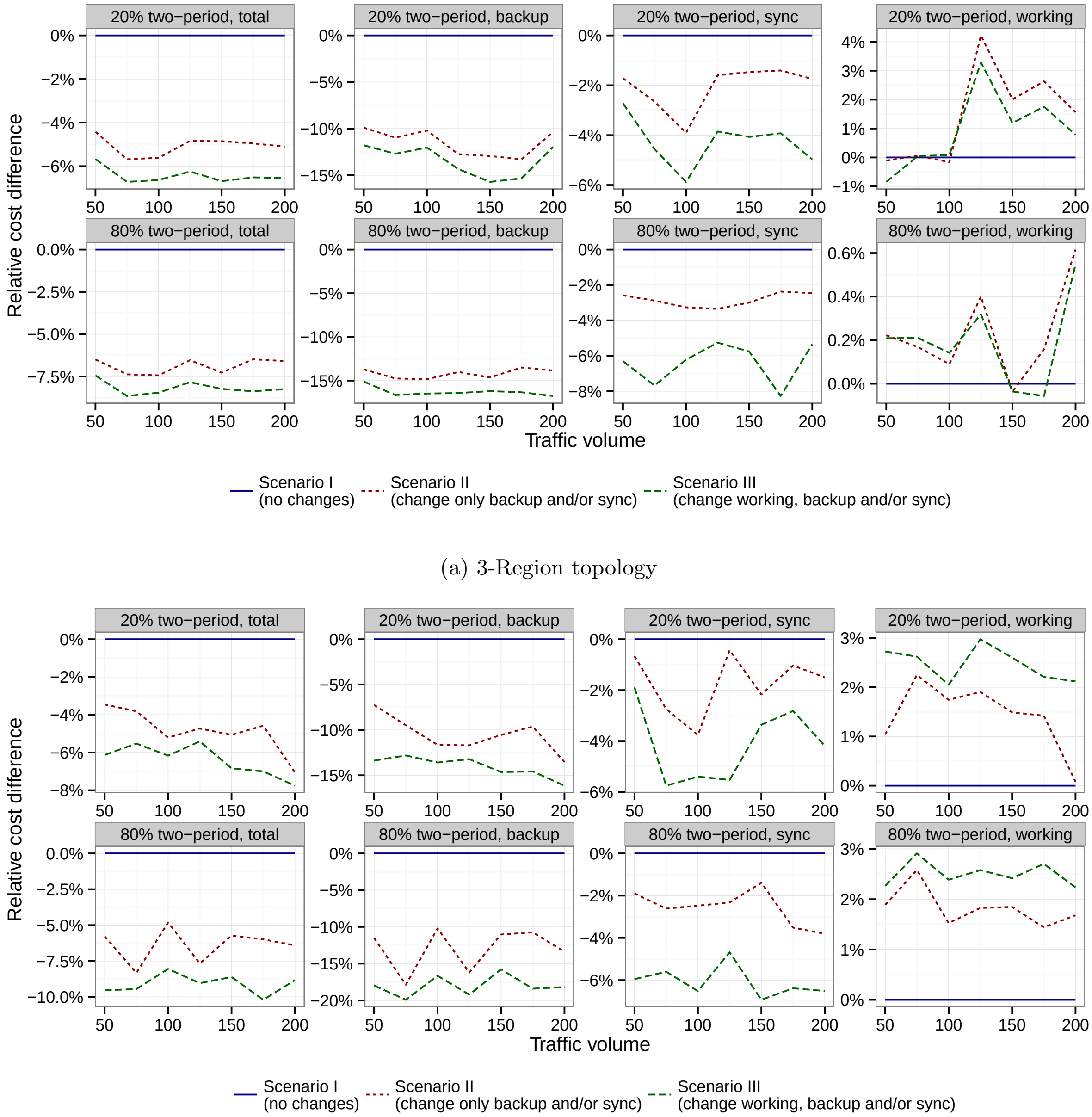
Fig. 3: US network topology with the assumed regions and data center locations indicated with a star.

net saving mainly stems from a reduction of bandwidth for the backup paths, due to increased sharing: we note an average reduction of the backup bandwidth cost of 11.5% (resp. 13.4%) for Pattern #1 (with 20% two-period traffic) and 14.2% (resp. 16.3%) for Pattern 2 (with 80% two-period traffic), when only changing backup/sync paths, i.e., Scenario II (resp. Scenario III, where also the working route can change). We found that these savings do not require all two-period traffic requests to change their routing when going from one period to the next, but only about half of them. Further, these preliminary results suggest that the cost reduction (in terms of bandwidth requirements) can be achieved by only changing the backup/synchronization paths when we consider multiple time periods together (Scenario II): The additional advantage of allowing also the working path to be changed (i.e., the extra benefit of Scenario III compared to Scenario II) is much smaller.

The net savings for this particular case study are modest, but non-negligible for this 3-Region case study. When we move to the more extreme 4-Region case (see Fig. 4b) the savings are a bit higher: we now find savings from rerouting that amount to up to around 10% of the total bandwidth requirements of the no-rerouting Scenario I. Note that these maximal savings are obtained in the 50% two-period traffic scenario. If the portion of two-period traffic increases further (e.g., the 80% two-period traffic case), savings go down. The intuition behind this observed behavior is that savings are realized by wisely choosing backup paths to increase sharing: the amount of traffic that we can share with more freely with (i.e., the next period's newly generated one-period traffic) goes down when going from 50% to 80% of two-period traffic, and so do the savings.

4.3. Benefit of parallellization

To assess the performance impact of our parallel algorithm (as sketched in Fig. 2), we compared it with a straightforward serial strategy. Results are plotted in Fig. 5, where we plot the running time in function of the number of executed rounds. We define a "round" as attempting to find a new configuration for each source node (by solving the corresponding pricing problems, PPs). In the *Parallel* scheme, we re-solve the RMP only after adding all new configurations found by the PPs (that are executed in parallel): one round comprises 1 restricted master problem (RMP) and for each source node one PP. In the *Serial* case, we solve one PP at a time, and re-solve the RMP each time we found a new configuration: one round thus comprises multiple RMPs (1 for each source node where the PP found a new configuration). Since we have in the order of 20 source nodes in our topology, we find that the time per individual round lies about a factor 20 higher for the Serial strategy compared to Parallel. In terms of the number of rounds required by the column generation algorithm to finish (i.e., none of the PPs finds a new, better configuration), the Parallel strategy takes more iterations: in our case study the number of rounds lies about 20% higher for Parallel compared to Serial (thus, the graphs for Serial stop at a lower number of rounds in Fig. 5). Still, the benefit in absolute wallclock time (i.e., the time a user needs to wait before finding the final solution) is substantial: the cumulative time over all rounds until completion lies about 18 times lower for Parallel.



(a) 3-Region topology

(b) 4-Region topology

Fig. 4: The bandwidth requirements for time-varying traffic, for different traffic patterns and topologies. Traffic volume is expressed in number of unit requests.

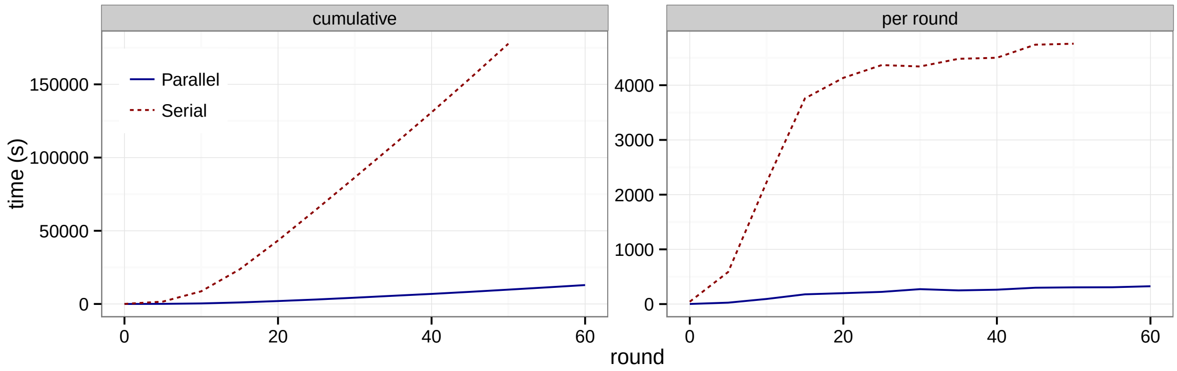


Fig. 5: Running times over subsequent rounds of the column generation algorithm, cumulative up to the current round (left) as well as per individual round (right).

5. Conclusion

We presented a new column generation model to solve a multi-period traffic dimensioning problem for resilient backbone networks for multi-site data centers. From the results presented above, the observations pertaining to bandwidth savings can be summarized as follows:

- The reduction of bandwidth requirements mainly stems from savings in bandwidth for *backup* paths (because of increased sharing with the requests starting in the 2nd period of two-period requests).
- A part of that backup capacity savings is lost due to longer *working*: these longer working paths are chosen to allow more sharing of backup capacity.
- When the working paths are allowed to be rerouted (Scenario III), the bandwidth saving is better than the situation that only backup paths can be rerouted (Scenario II). But the difference is not significant.
- The bandwidth savings from rerouting for the 3-Region topology amounts to less than 10% (in Scenario III, and slightly less for Scenario II).

Furthermore, we also showed that by adopting a parallel pricing problem (PP) solving strategy, we can realize a substantial amount of (wall clock) time. That overall saving is mainly caused by the reduction of the number of times that the (restricted) master problem (RMP) is solved: each iteration, multiple configurations (from the multiple PP solutions) are added, and moreover these PPs are solved in parallel.

ACKNOWLEDGMENT

B. Jaumard has been supported by a Concordia University Research Chair (Tier I) and by an NSERC (Natural Sciences and Engineering Research Council of Canada) grant.

REFERENCES

[1] C. Develder, M. De Leenheer, B. Dhoedt, M. Pickavet, D. Colle, F. De Turck, and P. Demeester, “Optical networks for grid and cloud computing applications,” *Proc. IEEE*, vol. 100, no. 5, pp. 1149–1167, May 2012.

[2] C. Develder, J. Buysse, B. Dhoedt, and B. Jaumard, “Joint dimensioning of server and network infrastructure for resilient optical grids/clouds,” *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1591–1606, Oct. 2014.

[3] T. Wang, B. Jaumard, and C. Develder, “A scalable model for multi-period virtual network mapping for resilient multi-site data centers,” in *Proc. 17th Int. Conf. Transparent Optical Netw. (ICTON 2015)*, Budapest, Hungary, 5–9 Jul. 2015, pp. 1–8.

[4] T. Wang, “Time-varying resilient virtual networking mapping for multi-location cloud data centers,” Master’s thesis, Concordia University, Montreal, Canada, 2016.