Cost-effective capacity provisioning in wide area networks with Shoofly

Rachee Singh, Microsoft
Nikolaj Bjørner, Microsoft
Sharon Shoham, Tel Aviv University
Yawei Yin, Microsoft
John Arnold, Microsoft
Jamie Gaudette, Microsoft

ABSTRACT

In this work we propose Shoofly, a network design tool that minimizes hardware costs of provisioning long-haul capacity by *optically bypassing* network hops where conversion of signals from optical to electrical domain is unnecessary and uneconomical. Shoofly leverages optical signal quality and traffic demand telemetry from a large commercial cloud provider to identify optical bypasses in the cloud WAN that reduce the hardware cost of long-haul capacity by 40%. A key challenge is that optical bypasses cause signals to travel longer distances on fiber before re-generation, potentially reducing link capacities and resilience to optical link failures. Despite these challenges, Shoofly provisions bypass-enabled topologies that can meet 8X the present-day demands using *existing network hardware*. These topologies save 32% of the cost of long-haul capacity while incorporating resilience to aggressive stochastic and deterministic link failure scenarios.[1]

CCS CONCEPTS

• **Networks** → **Wide area networks**; **Physical links**; **Network design and planning algorithms**.

KEYWORDS

Optical Bypass, Backbone Design, Traffic Engineering.

ACM Reference Format:
Rachee Singh, Nikolaj Bjørner, Sharon Shoham, Yawei Yin, John Arnold, and Jamie Gaudette. 2021. Cost-effective capacity provisioning in wide area networks with Shoofly. In *ACM SIGCOMM 2021 Conference (SIGCOMM '21), August 23–27, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3452296.3472895

1 INTRODUCTION

Traffic demands on cloud wide area networks (WANs) are growing rapidly, driven by new workloads like real-time video and cloud gaming. Cloud providers respond to increase in traffic demands by provisioning additional WAN capacity. However, long-haul network capacity is expensive – the median annual cost of 100 Gbps of long-haul capacity is over $100,000 in N. America as per TeleGeography (Figure 1) [32]. Fiber, routers and equipment in the optical line system (OLS) [19] are the key contributors to the cost of capacity in the cloud WAN (Figure 2). Large cloud providers have existing fiber deployments, acquired through purchase or long-term leases. Thus, the marginal cost of provisioning capacity in the cloud is dominated by the additional hardware resources used in the process: *router and optical ports.*

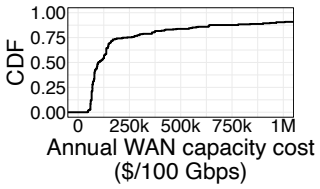


Figure 1: Capacity cost.

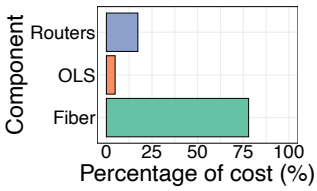


Figure 2: Cost breakdown.

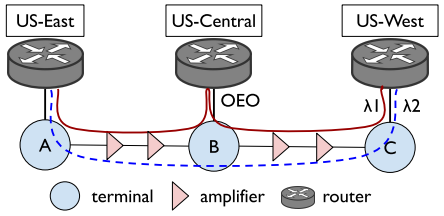


Figure 3: Optical terminals consist of wavelength selective switches (WSS), multiplexers/demultiplexers. They connect light channels to router ports. Transponders plugged into router ports convert the optical signals to electrical signals.

Cloud providers operate *point-to-point* inter-regional networks – where optical signals are converted to electrical signals and back at every geographical region [12]. Thus, inter-regional traffic undergoes optical-to-electrical-to-optical (OEO) conversion at all intermediate regions on the path towards its destination. For example, in Figure 3, wavelength $\lambda 1$ originates at US-East, terminates at US-West but undergoes an OEO conversion at US-central. Once the signals are converted from the optical to electrical domain, centralized traffic engineering systems take control of routing them [18, 20].

This design provides control and flexibility at the network layer, keeping the optical layer uncomplicated and easy to manage. However, the conventional design does not keep in view the nature of traffic demands and the corresponding traffic flow imposed by them. We analyze inter-regional traffic patterns in the backbone of a large commercial cloud provider and find that 60% of traffic traversing through 30% of geographical regions in the WAN is *passing through* – neither originating nor terminating at the region. The pass-through or transit traffic undergoes wasteful OEO conversions at all intermediate regions in point-to-point networks, occupying scarce optical line- and router ports. These ports contribute a majority of the cost of provisioning capacity in cloud networks with existing fiber deployments (Figure 2).

[1] Code and experiments at http://shoofly.network.

SIGCOMM '21, August 23–27, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8383-7/21/08…$15.00
https://doi.org/10.1145/3452296.3472895

In this work, we propose to minimize the hardware cost of provisioning long-haul capacity by removing wasteful OEO conversions in the WAN. We refer to the elimination of an OEO conversion at a region as *optically bypassing* the region *e.g.,* wavelength $\lambda 2$ optically bypasses US-Central in Figure 3. Our analysis shows that 60% of the transit traffic through a region is exchanged between only two neighbors of the region – highlighting the potential of reaping most cost savings with very *few* optical bypasses.

While optically bypassing regions in cloud backbone networks can offer significant reduction in capacity cost, it also introduces new operational challenges. First, wavelengths of light in point-to-point regional networks undergo regeneration at every hop, correcting errors caused by signal attenuation and dispersion during transmission on the fiber. By optically bypassing regions, signals are forced to travel longer fiber paths before regeneration *e.g.,* $\lambda 2$ in Figure 3 travels from US-East to US-West without regeneration whereas $\lambda 1$ undergoes regeneration at US-Central. Longer distances can necessitate lower data rates for error-free transmission *e.g.,* $\lambda 2$ faces more attenuation than $\lambda 1$ since it travels a longer fiber path without regeneration and can only sustain data rates lower than $\lambda 1$'s. Thus, optical bypass can lower the achievable capacity over a channel, in turn hampering the network's ability to meet traffic demands. Second, IP links resulting from signals that optically bypass regions are susceptible to failure if *any constituent fiber link* fails. This expansion of shared risk link groups (SRLGs) can reduce the network's resilience to physical link failures [22].

We tackle these challenges by analyzing the optical signal quality on fiber of a commercial cloud provider and find that the signal quality of 75% of optical channels on fiber is sufficiently high to sustain transmission over longer distances. The remaining optical channels must downgrade their data rates to traverse longer fiber paths without regeneration (§2). We leverage these empirical insights to make the following key contributions towards the design of a bypass-enabled cloud backbone with resilience to link failures:

Optimal optical bypass in cloud WANs. We develop Shoofly, an optical backbone design tool that formulates the problem of identifying optical bypass opportunities in a cloud network with the goal of *minimizing the hardware costs of long-haul capacity*. Shoofly-proposed backbone topology can reduce the hardware cost of long-haul capacity by 40% while continuing to meet up to 8X the present-day demands using *existing hardware and fiber deployments* (§3, §4).

Failure resilience with optical bypass. We extend Shoofly to incorporate the goals of failure resilience while identifying bypass opportunities in the WAN. We show that Shoofly can provision bypass-enabled topologies resilient to both stochastic [5] and deterministic [25] link failures by sacrificing at most 20% of hardware cost savings (§5).

Low logistical burden of enabling optical bypass. We quantify the logistical burden of provisioning the Shoofly proposed bypasses on cloud operators. We show that the bypass-enabled topology can keep up with demand growth over time, up to 8X the present-day demands. Implementing bypasses requires modifications to the physical connections between optical terminals and routers. We show that a small number of such changes are needed since 25% of optical bypasses achieve 80% of the cost savings (§**??**).

The design of optical networks has been well studied by service and content providers [2, 3, 9, 16]. In contrast, cloud WANs with centralized software-defined traffic engineering present a unique opportunity to rethink conventional backbone designs due to the *predictability* [33] of intra-WAN traffic in cloud networks.

Ethics Statement. This work does not raise any ethical issues.

2 QUANTIFYING THE OPPORTUNITY

Cloud providers lease or purchase optical fiber across the world to provision their WAN. This fiber is connected to optical equipment, namely, *optical terminals* and *optical amplifiers* to transmit optical signals over hundreds of kilometers. Signals can originate, terminate or pass through an optical terminal. Signals that terminate undergo a conversion from the optical to electrical domain. The corresponding electrical signals are then de-multiplexed onto ports of a router or switch. Similarly, electrical signals from router ports are converted to optical signals and multiplexed onto wavelengths of light that traverse the fiber. In contrast, signals can *pass through* an optical terminal without originating or terminating at it. These signals are said to *optically bypass* a router (*e.g.,* US-Central is bypassed by wavelength $\lambda 2$ in Figure 3).

2.1 Point-to-point regional backbones

The commercial cloud provider we study has provisioned their wide area backbone in a point-to-point regional topology. In this design, optical signals undergo regeneration at all regional hops. As per conventional wisdom, this design offers two main benefits:

Fine-grained control via Layer-3 traffic engineering. Optical signals are converted to the electrical domain at all regional hops in this design, allowing the Layer-3 traffic engineering algorithms hop-by-hop autonomy to route traffic to suitable next-hops. These algorithms can leverage fine-grained, cross-layer telemetry from the network to make on-the-fly decisions that achieve network-wide goals, *e.g.,* minimum end-to-end latency and maximum throughput.

Flexibility to meet new demands. Since wavelengths undergo conversion to electrical signals at every region in a point-to-point WAN, traffic from one region can be IP routed to any other region in the network. This flexibility allows the network to meet new and emerging traffic demands between regions without requiring any changes to the optical backbone. In this design, the optical and IP topology of the network bear close resemblance. Each optical terminal maps to an IP router and fiber connections between neighboring regions underpin the IP links between the corresponding routers (Figure 4). Enabling optical bypass can hamper the ability of certain regions from being the origin or destination of traffic. In an extreme case, a region could be bypassed entirely by all optical wavelengths (*e.g.,* region B in Figure 4).

Despite the flexibility and control offered by point-to-point wide area backbones, we propose to rethink this design in the context of traffic flow patterns in the network. In this section, we demonstrate that the nature of inter-regional traffic flow enables a significant potential to save hardware costs of capacity provisioning in point-to-point backbones.

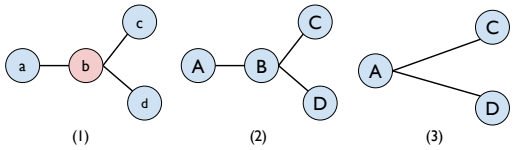


Figure 4: (1) shows the physical topology of a network with 4 optical terminals (a, b, c, d) and the fiber between them. (2) shows the IP layer topology of the same physical network in a point-to-point design – each terminal maps to a router (A, B, C, D). (3) shows the IP layer topology of the physical network in (1) where all signals optically bypass router B.

2.2 Wasteful OEO Conversions

We study the traffic flow between geographical regions of a large commercial cloud provider. The cloud provider has datacenters in approximately 100 geographical regions in the world, connected by a dedicated optical backbone. Demands between regions are routed through a centralized software-defined traffic engineering (TE) controller [18, 20]. At a high level, the TE controller solves a k-shortest path formulation of the multi-commodity flow problem [1]. Demands are denoted by their source and destination regions. For every demand pair, there exist pre-computed k paths across which flow is distributed. We measure hourly demands between regions in the WAN and the corresponding k paths over which they are routed from August 1, 2020 to December 31, 2020.

All traffic transits at least one geographical region. For every demand source and destination pair, we measure the fraction of traffic that traverses WAN paths of different lengths. Each hop on these paths is a geographical region in the WAN. We compute the average length of the WAN path for all demand pairs, weighted by the fraction of demand traffic carried by the paths. Nearly all demands are routed through indirect regional paths – with 75% demands encountering at least 3 intermediate regional hops (Figure 5a).

Majority of traffic routed at any region is *pass-through*. For each region, we measure the three categories of traffic it observes: *origin traffic* originates at the region, *sink traffic* terminates at the region and *transit traffic* passes through the region. We find that all regions observe a high volume of transit traffic. For 30% of geographical regions, the transit traffic volume is over 60%. On average, over 50% of traffic observed at a region is transit traffic (Figure 5b). The high volume of transit traffic through regions contributes to wasteful OEO conversions that occupy routers ports and transceivers. The cost of long-haul capacity can be lowered if the traffic optically bypassed transit regions, staying in the optical domain for longer distances until reaching its destination.

Few regional pairs contribute most of transit traffic. Finally, we measure the neighboring regions that ingress and egress transit traffic through a region in the WAN. We calculate the fraction of transit traffic through a region that is contributed by one ingress-egress neighbor pair. We find that over 50% of the transit traffic through any region is exchanged between two neighbors of the region (Figure 5c). The distribution of transit traffic through regions by ingress-egress regional pairs sorted on transit traffic volume, in Figure 5c shows that optically bypassing the region between few ingress-egress region pairs will achieve most of the cost savings of optical bypass in WANs.

While the flexibility of point-to-point backbones (§2.1) is an important feature, our analysis of an inter-regional cloud WAN over five months shows that the traffic flow between geographical regions of the cloud WAN is suitable for a bypass enabled design. Despite the seasonality of demands, predictable traffic engineering algorithms have imparted *stability* to traffic patterns – few regions are responsible for most of the transit traffic. We propose to leverage this stability to design a cost-efficient optical backbone.

3 OPTICAL BYPASS WITH SHOOFLY

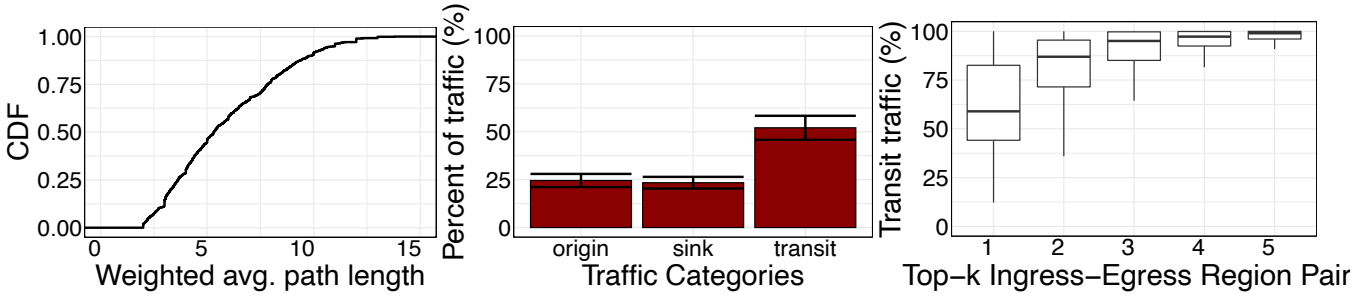
We propose to leverage insights from inter-regional traffic patterns in the WAN to enable optical bypass in the network. Our goal is to reduce the cost of long-haul capacity by preventing wasteful OEO conversions while using existing network hardware and software.

Cost of long-haul capacity. Large cloud providers either purchase fiber or acquire it through long-term leases that span decades, making fiber an infrequent capital investment. Cloud providers also purchase optical (*e.g.,* amplifiers, terminals, transceivers) and electrical equipment (*e.g.,* routers) to provision capacity between different regions of their wide-area backbone. Over time, wavelengths are *lit* on the fiber to meet growing traffic demands. In steady state, the cloud provider has existing fiber deployments. Thus, provisioning additional capacity requires *lighting* a new wavelength or signal on the fiber that originates and terminates at the target source and destination regions. The marginal cost of provisioning the wavelength is determined by the router and optical line ports it uses. The combined cost of a router and optical port can be as high as tens of thousands of dollars. In point-to-point networks, each wavelength uses at least one router and optical line port at every intermediate regional hop in addition to the source and destination regions.

How does bypass save cost? Optical bypass of a region by a wavelength of light saves two router ports and up to two line ports in an industry standard deployment of routers and optical terminals, allowing a significant reduction in the $/Gbps cost of long-haul capacity. However, the ability to bypass a region is conditional on physical constraints on the network topology and signal quality on fiber. We elaborate on these constraints in the next section.

3.1 Physical constraints on optical bypass

Optical bypass of geographical regions forces the signals to travel longer distances on fiber before they can be regenerated. Signals traveling longer distances undergo more attenuation leading to lower optical signal quality. Signal quality, measured through optical signal-to-noise ratio (OSNR), ultimately decides the data rate of the optical signal – higher the OSNR, higher the signal data rate. State-of-the-art optical transponders support three data rates per wavelength of light: 200 Gbps, 150 Gbps and 100 Gbps by modulating the signals in 16-QAM, 8-QAM and QPSK formats respectively. Thus, optical bypass by some wavelengths in the WAN can lower their achievable data rates due to increased transmission distance. In this section we explore the constraint posed by distance on potential optical bypasses in the cloud network. Table 1 shows the relationship between modulation formats, their OSNR thresholds and data rates.



(a) Avg. path length. **(b) Traffic types through regions.** **(c) Transit traffic through regions.**

Figure 5: 5a shows the average regional path length for demands in the WAN weighted by traffic volume. 5b shows the three categories of traffic observed by regions. 5c shows the transit traffic volume per $k = 1, \ldots, 5$ top ingress and egress neighbors of a region. Over 60% of transit traffic through all geographical regions is between one pair of neighbors.

Optical signal quality in the WAN. The OSNR of all wavelengths in the cloud WAN is higher than 15 dB (Figure 6a), highlighting that nearly all wavelengths in the cloud network currently support data rates of 150 Gbps (8-QAM) or higher. 75% of the wavelengths have an OSNR over 20 dB which is higher than the threshold OSNR for the highest data rate of 200 Gbps in the cloud network. Since their OSNR is over 3 dB above the threshold for 16-QAM and 5.5 dB above the threshold for 8-QAM formats, these wavelengths can travel longer distances without regeneration while still maintaining the same data rate. The amount of extra distance that a wavelength can travel without requiring a downgrade in modulation format depends on its current OSNR and modulation format. We discuss this in the next sub-section.

Modulation format	QPSK	8-QAM	16-QAM
Minimum OSNR	10dB	14.5 dB	17 dB
Data rate	100 Gbps	150 Gbps	200 Gbps
Optical reach	5,000 km	2,500 km	800 km

Table 1: OSNR thresholds, data rates and optical reach of modulation formats of signal on fiber.

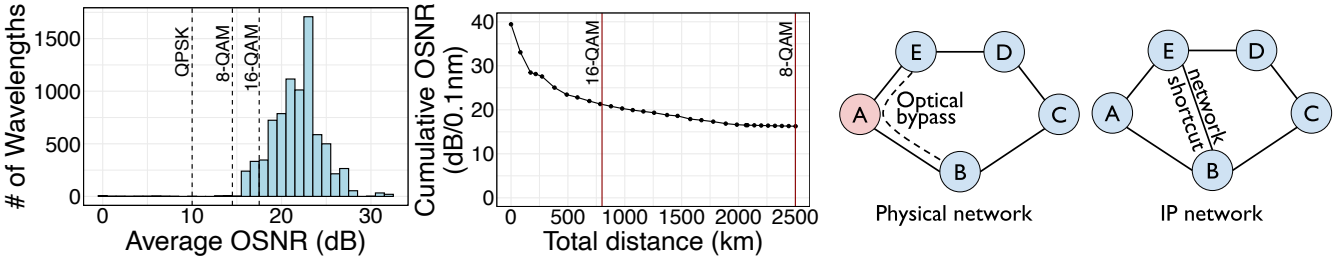
Optical reach. Optical reach is the maximum distance a light signal can travel before it must be re-generated. If the signal is not re-generated within this distance, the OSNR of the signal is too low to merit error-free decoding at the destination. While state-of-the-art transceivers and routers have forward error correction (FEC) abilities to correct transmission errors, if the OSNR is lower than the FEC threshold, error-free transmission is not possible. The reach of a signal depends on characteristics of the physical network, including the launch power of the signal, noise on fiber, gain of amplifiers and optical span lengths. We gather these parameters from the cloud network and simulate the relationship between the OSNR of signals and the distance traversed by them. This OSNR estimation is approximate and is used for network planning. Figure 6b shows the decline in the simulated OSNR as the distances traversed by the signal increases. We observe that after 1,000 km, the OSNR of signals nears the 16-QAM threshold of 17 dB. If the signal is transmitted over this distance threshold, it must be modulated in a lower order format like 8-QAM. Similarly, after traversing more distance, the signal OSNR drops below the 8-QAM threshold.

We subtract a margin of over 200 km from the optical reach estimates derived from Figure 6b to make the reach estimates conservative. Table 1 summarizes the optical reach of a signal as a function of its modulation format – lower order modulation (*e.g.,* QPSK) formats can travel longer distances without re-generation compared to high order modulation formats (*e.g.,* 8-QAM, 16-QAM). Thus, higher order modulation formats enable higher data rates but have lower optical reach. Due to limited optical reach, each signal can bypass a fixed number of regions before a re-generation becomes essential. The number of regions that can be bypassed depends not only on the traffic patterns between the regions but also on the modulation format for the signals. If bypassing a region pushes the transmission of a signal over the optical reach of its modulation, the operator must lower the modulation format and consequently the data rate of the signal. Given the set of demands in the network, lowering the modulation can reduce the network's ability to meet demands. Therefore, selection of regions and wavelengths for optical bypass must navigate the balance between cost saving and the ability of the network to meet demands.

3.2 Bypass as network shortcuts

Our empirical analysis of traffic patterns (§2) shows the potential for optically bypassing regions to save on the hardware cost of capacity in the WAN. However, identifying the set of wavelengths and regions that can be bypassed is hard. The space of potential optical bypasses is constrained by physical factors (*e.g.,* signal quality and optical reach), traffic demands and network tunnels [10] over which they are routed. To effectively enumerate and search the space of potential optical bypasses, we introduce the graph abstraction of *network shortcuts* to represent an optical bypass.

The bypass of one or more regions by a wavelength on fiber introduces a new edge in the corresponding IP network. We refer to this bypass-induced edge as a *network shortcut*. In Figure 6c, the bypass of region A by a wavelength between regions E and B introduces the shortcut EB in the IP network. At the physical layer, the only change is the optical bypass of region A implemented by changing physical connections between the terminal and router at region A. But, higher layers of the networking stack observe a direct connection between nodes E and B as a result of this change. We define one instance of bypass by the corresponding shortcut and its underlying fiber path , *e.g.,* the bypass in Figure 6c is defined by the shortcut EB and $E \rightarrow A \rightarrow B$.



(a) Optical signal quality in WAN. **(b) OSNR vs. fiber distance.** **(c) Bypass as shortcuts.**

Figure 6: 6a shows the histogram of OSNR of wavelengths in the cloud WAN. 6b shows the decline in OSNR as the transmission distance increases. (The graph is truncated at 2,500 km since the decline in OSNR is slow.) 6c shows optical bypass at the physical layer translates to adding a *shortcut* edge in the IP network, *e.g.,* one wavelength from node E bypasses node A and terminates at node B. This change adds a shortcut edge between node A and B.

Feasible network shortcuts. Using the network shortcut abstraction, we enumerate all potential shortcuts in the cloud backbone. To do this, we compute the shortest regional path between every pair of regions in the network. After 3 intermediate regional hops between the source and destination regions, the OSNR is too low and the signal must be regenerated regardless of the other physical constraints. Thus, feasible network shortcuts can have up to 5 hops including the source and destination region. This enumeration yields a list of all potential shortcuts in the network.

Wavelengths as minimum unit of capacity. After enumerating all feasible bypasses in the network as shortcuts, the remaining unknown is the capacity of each shortcut. Our goal is to find the capacity of enumerated shortcuts such that physical constraints and traffic demand constraints are met while the bypass-enabled cost savings are maximized. Similar to the capacity of existing IP links in the network, the capacity of a shortcut is the product of the number of wavelengths that constitute the shortcut and the data rates of their modulation formats. These wavelengths originate at the start of the shortcut, terminate at the end and bypass all intermediate nodes. The modulation format of a shortcut is determined by the OSNR and shortcut length. While the modulation formats of wavelengths in the original cloud topology are known, we determine the modulation formats of wavelengths on shortcuts using the shortcut length (Table 1). If the shortcut length is higher than the optical reach of the wavelengths' modulation format, the shortcut can sustain transmission at a lower modulation format.

3.3 Optimal optical bypass

We propose Shoofly, a tool that formalizes the task of minimizing the hardware cost of long-haul capacity by optically bypassing regions in the WAN. Shoofly leverages the network shortcut abstraction (§3.2) and enumerates all feasible shortcuts, $s \in S$. Since the shortcuts are pre-computed, the per-wavelength data rates (u_s) of shortcuts are also an input to Shoofly. The operator can prune the set of feasible shortcuts to impose policy decisions, *e.g.,* only allow the bypass of one region at a time.

Decision Variables. Shoofly allocates wavelengths to each network shortcut, s. The wavelengths on s are bounded by the number of wavelengths on each edge constituting the shortcut in the original network. The decision of allocating wavelengths w_s to shortcut s implies that w_s light signals between the start and end regions of the shortcut do not regenerate at intermediate regional hops by bypassing them. The capacity of a shortcut is a product of the wavelengths assigned to it by Shoofly and the data rate of the wavelengths' modulation format, u_s. We note that w_s is an integer. The remaining decision variables in the optimization are auxiliary and we define them in the following.

Algorithm 1: Optimal Optical Bypass in WANs

Inputs:

$G\langle V, E\rangle$:	network G, vertices V and edges E
c_e:	capacity of edge e
u_e:	capacity of one wavelength of edge e
D_d:	traffic demand between src_d and dst_d
T_d:	set of tunnels for demand d
s:	network shortcut due to optical bypass
u_s:	capacity of one wavelength of shortcut s

Outputs:

$flow_t$	$\in \mathcal{R}_{\geq 0}$	flow allocated over tunnel t
x_e^t	$\in \mathcal{R}_{\geq 0}$	flow allocation on edge e for tunnel t
w_s	$\in \mathcal{N}$	number of wavelengths on shortcut s
y_s^t	$\in \mathcal{R}_{\geq 0}$	flow allocation on shortcut s for t

Maximize: $\sum_s |s| \cdot w_s$

subject to:

(1)	$D_d \leq \sum_{t \in T_d} flow_t$,	$\forall d \in D$
(2)	$0 \leq x_e^t$,	$\forall t \in T, e \in t$
(3)	$0 \leq y_s^t$,	$\forall t \in T, s \in t$
(4)	$flow_t \leq x_e^t + \sum_{s \ni e} y_s^t$,	$\forall t \in T, e \in t$
(5)	$\sum_{t \ni s} y_s^t \leq u_s \cdot w_s$,	$\forall s$
(6)	$x_e := \sum_{t \ni e} x_e^t$,	$\forall e$
(7)	$x_e + u_e \cdot \sum_{s \ni e} w_s \leq c_e$,	$\forall e$
(8)	$w_s = w_{\overleftarrow{s}}$,	$\forall s$

Objective function. To maximize the cost saving from optical bypass, Shoofly maximizes the number of router and optical ports that are *freed* by allocating wavelengths to shortcuts. The OEO conversion of each wavelength at a regional hop occupies a router port and optical port in both ingress and egress directions. Thus, a shortcut s with w_s wavelengths frees 2 router and optical ports per wavelength at every intermediate hop in the shortcut. The cost

saving from a shortcut s is proportional to $w_s \cdot |s|$ where $|s|$ is the number of hops in the shortcut.

Key Insight. Shoofly's goal is not to design the optical backbone from scratch but to leverage existing software and hardware placement to reduce the amortized cost of long-haul capacity – while continuing to meet existing traffic demands in the network. Therefore, we design Shoofly as an algorithm that computes flow allocations on tunnels in the original regional network topology – similar to traffic engineering algorithms [18, 20]. However, in addition to allocating flows, Shoofly *siphons* as much of the traffic allocations from tunnels to the network shortcuts as possible to increase the number of wavelengths that can participate in bypass. The combination of the siphoned flow (y_s^t) and the flow on existing edges (x_e^t) must meet the traffic demands between regions. Wavelengths on shortcuts must be enough in capacity to carry the siphoned flow on the shortcut. By siphoning flow to the shortcuts, Shoofly facilitates the bypass of network capacity while meeting traffic demands.

Demand Constraints. Each demand d between two regions in the WAN has a demand amount D_d and a set of tunnels T_d associated with it. Tunnels are the pre-computed set of k shortest paths between the demand source and destination regions. The set of tunnels T is the union of tunnels $\bigcup_d T_d$ over all demands d. The sum of flow allocated to all tunnels of a demand, should meet the demand:

$$D_d \leq \sum_{t \in T_d} flow_t \tag{1}$$

Flow conservation constraints. A shortcut s is a path or a sequence of adjacent edges. The shortcut s is said to be *on a tunnel* t if all edges $e \in s$ belong to t. Similarly, $s \in t$ denotes that shortcut s is on the tunnel t. For each tunnel t, edge e, shortcut s on t, we associate non-negative output variables x_e^t and y_s^t, where y_s^t is the flow that passes shortcut s on t, and x_e^t is the flow that passes edge e, outside of all shortcuts.

$$0 \leq x_e^t, \quad \forall t \in T, e \in t \tag{2}$$

$$0 \leq y_s^t, \quad \forall t \in T, s \in t \tag{3}$$

The flow allocated to a tunnel t must be carried either on the edges along the tunnel or shortcuts along it. We prove that Equation (4) ensures conservation of flow as it is siphoned to shortcuts in Appendix A.1.

$$flow_t \leq x_e^t + \sum_{s \ni e} y_s^t, \quad \forall t \in T, e \in t \tag{4}$$

Wavelength constraints. The total flow siphoned off to a shortcut must be bounded by the shortcut's capacity, *i.e.,* the product of the number of wavelengths on the shortcut (w_s) and their corresponding data rates (u_e). For instance, if a shortcut consists of two wavelengths that can support 8-QAM modulation, the total flow siphoned to this shortcut must be bounded by 300 Gbps.

$$\sum_{t \ni s} y_s^t \leq u_s \cdot w_s, \quad \forall s \tag{5}$$

Capacity constraints. The total flow on edges is the sum of allocations across all tunnels.

$$x_e \coloneqq \sum_{t \ni e} x_e^t, \quad \forall e \tag{6}$$

The capacity of edges, c_e, is reduced due to the migration of some wavelengths from edges to shortcuts that contain the edges. The reduced capacity of edges must be sufficient to meet the total flow allocated to the edges. The reduction in capacity is a product of the number of wavelengths bypassing the edge and the modulation format of the edge. For instance, a wavelength that contributed 200 Gbps on an edge (u_e = 200 Gbps) can be assigned to a shortcut containing the edge, thus reducing the edge capacity by 200 Gbps. This wavelength may only contribute 150 Gbps to the shortcut it is becoming a part of since u_s can be lesser than u_e.

$$x_e + u_e \cdot \sum_{s \ni e} w_s \leq c_e, \quad \forall e \tag{7}$$

Bi-directional equality constraints. Links in optical networks can be assumed to be bi-directional. Thus, every shortcut s is also bi-directional and has a reverse $\overleftarrow{s}$. We ensure that shortcuts and their reverse siblings are allocated the same number of wavelengths.

$$w_s = w_{\overleftarrow{s}}, \quad \forall s \tag{8}$$

Alg. 1 summarizes Shoofly's optimization formulation using equations (1)-(8). We will discuss other algorithms that use a subset of the constraints. Thus, we define:

Definition 3.1 (Traffic Allocation Constraints). *Equations (2)-(8) capture all constraints related to traffic allocation in a network with optical bypass. We define the predicate AllocationConstraints as the conjunction of these constraints.*

4 COST SAVINGS WITH SHOOFLY

We implement Shoofly's optimization algorithm using Python 3 bindings of the commercial optimization solver Gurobi [17]. We note that Alg. 1 solves a mixed integer program (MIP). In practice, Gurobi solves the problem efficiently using a linear programming (LP) relaxation and a MIP gap of 0.1%. Solution to the LP relaxation provides an upper bound for the maximization problem of Alg. 1. The MIP gap defines the break condition for the optimization solver *i.e.,* the solver continues to search for a solution using the branch-and-bound strategy until it finds one within 0.1% of the LP optimal. All instances of Alg. 1 we formulate were solved within 10 seconds of runtime, which is acceptable for build planning.

Fiber lengths of shortcuts. We first evaluate Shoofly on the network topology, traffic demand matrix and optical signal quality of a large commercial cloud provider. The cloud provider we analyze has a global footprint with presence in approximately 100 geographical regions. We enumerate all potential shortcuts of 3, 4 and 5 total regional hops in the network. After 5 regional hops, the signal must undergo regeneration and thus shortcuts of more than 5 hops are not feasible. Figure 7 shows the distribution of the lengths of the fiber path in each shortcut. As discussed in Table 1, optical signal quality is too low to sustain transmission even at the lowest possible modulation format of QPSK after traversing 5,000 km on fiber without regeneration. Since the regions in the cloud provider

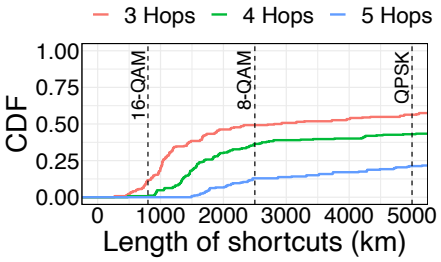


Figure 7: The distribution of fiber lengths of network shortcuts. The dotted lines represent the threshold distance for QPSK, 8-QAM and 16-QAM modulation formats.

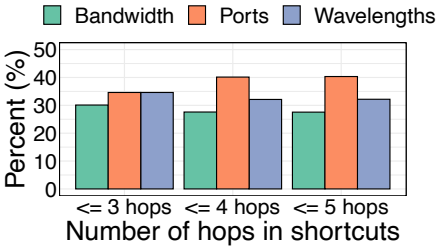


Figure 8: Percentage of bandwidth bypassed, ports saved and wavelengths bypassed by Shoofly.

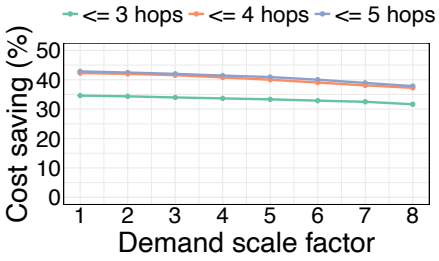


Figure 9: Impact of scaling demands on Shoofly's cost savings. Cost savings reduce by 2% as demand is scaled to 8X.

are geo-distributed globally, the length of network shortcuts can span several thousand kilometers. In fact, over 50% of shortcuts of all hop lengths are longer than 5,000 km, rendering these shortcuts infeasible (Figure 7). Lengths of the remaining feasible shortcuts decide the modulation format that signals on those shortcuts can support. Nearly all 3-hop shortcuts can support 8-QAM or 150 Gbps of data rate per wavelength. Higher hop-count shortcuts can be longer and thus support lower data rates *e.g.,* 100 Gbps.

4.1 Reducing hardware costs of capacity

Shoofly identifies wavelengths in the cloud provider's network that can optically bypass regional hops by allocating capacity to pre-computed feasible shortcuts in the network. In this section, we evaluate the cost savings achieved by Shoofly of various practical topologies.

Impact of shortcut length. We formulate three instances of Alg. 1 – first instance considers shortcuts of 3 hops, second considers shortcuts of 3 and 4 hops and third considers shortcuts of 3, 4, and 5 hops. The shortcut path lengths embody a critical trade-off for Shoofly: longer shortcuts enable higher cost savings by freeing more ports but reduce the data rate of wavelengths on the shortcut. We consider the three different instances of Shoofly based on the maximum permissible shortcut lengths to evaluate this trade-off. We solve the three MIP instances and plot the percentage of total bandwidth allocated to shortcuts, total ports saved by the shortcuts and total wavelengths migrated to shortcuts in Figure 8. We observe that while longer length shortcuts save more ports, the total bandwidth on the shortcuts reduces with hop length. This is a direct consequence of the length vs. data rate trade-off. The number of wavelengths migrated to shortcuts remain similar regardless of the shortcut lengths as they are a function of the traffic matrix which remains the same in all three problem instances. The results of Figure 8 show that **Shoofly can save over 40% of the hardware costs of long-haul capacity** by freeing expensive router and optical line ports at regional hops.

Impact of over-provisioning in networks. Shoofly ensures that existing traffic demands of the network continue to be met in the bypass-enabled topology. However, cloud wide area backbones are often over-provisioned in an attempt to future-proof the network for potential increase in demands. To ensure that Shoofly does not render the bypass-enabled topology incapable of handling increased demands in the future, we evaluate Shoofly on demands that are scaled to 8X the maximum inter-region demands observed in December 2020. Figure 9 shows that there is a very small decline (<2%) in the potential cost savings of bypass as the traffic demand matrix is scaled to 8X the maximum present-day demands. Thus, network operators can provision bypasses proposed by Shoofly using scaled traffic demands to make future-proof bypass decisions without sacrificing on cost savings.

Impact of network topology. Next, we evaluate Shoofly on different network topologies. We have detailed information about the network of the cloud provider we study, referred to as cloud provider WAN or CP-WAN in the figures. Additionally, we evaluate Shoofly using the network topology and demand matrices for prominent production networks, released by previous work [5, 23]. We assume that these networks operate a point-to-point optical backbone. Figure 10a shows the percentage of ports saved by Shoofly for the backbone networks of Abilene, B4, Nextgen, CP-WAN and a custom topology from previous work. We find that Shoofly shows a consistent potential of saving hardware costs in all network topologies, ranging from over 55% cost saving in the Nextgen topology to 15% cost saving in the Custom network topology. Figure 10b shows the fraction of regions in the networks that participate in optical bypass. Over 50% of the regions in the CP-WAN get bypassed by one or more wavelengths, realizing the bypass potential we found in the inter-regional traffic matrices (§2).

4.2 Lower data-rates from optical bypass

One key concern raised by optical bypass is that by forcing signals to travel longer distances, a bypass-enabled topology can reduce the capacity between regions. Our evaluation of Shoofly shows that the bypass-enabled topology can not only meet 8X the present-day traffic demands (§4.1) but also enable 30-40% hardware cost savings (Figure 9).

The reduction in capacity between regions occurs due to a downgrade in signal modulation formats on shortcuts on account of increased transmission distance. Figure 11 shows the modulation formats of the bypasses enabled by Shoofly. We observe that fewer

(a) Percent cost saving by Shoofly. (b) Percentage regions bypassed.

Figure 10: 10a shows that Shoofly can save over 40% of hardware costs of long-haul capacity in CP-WAN. 10b showes that majority of regions in all network topologies get bypassed by one or more wavelengths.

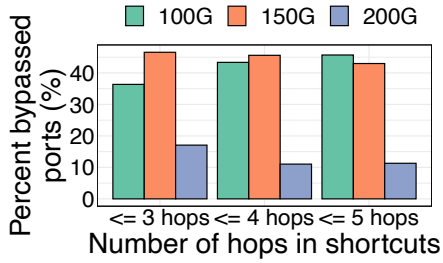


Figure 11: Modulation formats on shortcuts of different lengths. As the number of shortcut hops increases, the fraction of ports in higher modulation formats reduce.

shortcut hops allow Shoofly to keep majority of the wavelengths in higher order modulation formats *e.g.,* of the 3-hop only shortcuts allocated by Shoofly, over 45% can sustain 150 Gbps, 17% can sustain 200 Gbps per wavelength. As Shoofly considers shortcuts with higher hops, it can save more cost (Figure 8) but this higher saving comes at the cost of longer shortcut lengths and consequently lower data rates. Figure 11 shows that fraction of bypasses that can support higher modulation formats reduce as hop lengths increase. When allowed up to 5-hop shortcuts, only 11% of shortcuts can sustain 200 Gbps. We compare the split of *all* links based on modulation formats in the original and bypass-enabled networks in Figure 17 of Appendix A.2.

Network	original	≤ 3-hop	≤ 4-hop	≤ 5-hop
Capacity	160 G	157.5 G	153.5 G	153.5 G

Table 2: Norm. per-link capacity of the original network and networks with 3, 4 and 5 hop shortcuts.

To mitigate the concern of lowered network capacity due to bypass, we compute the normalized per-link capacity of the original network and the bypass-enabled networks in Table 2. The normalized link capacity in the original network is 160 Gbps whereas bypass-enabled networks lower it by 4% in the worst-case.

5 FAILURE RESILIENT OPTICAL BYPASS

Enabling optical bypasses fundamentally changes the impact of individual physical link failures on the IP network's ability to carry traffic. For instance, in point-to-point networks, there is a one-to-one mapping from physical to IP links. Therefore, the failure of a physical link (*e.g.,* fiber cut, amplifier failure) leads to an individual IP link's failure. However, in a bypass-enabled topology one physical link can underpin several IP links and the failure of one physical link can cause *multiple IP links* to fail. Thus, we must revisit failure resilience of the backbone network with optical bypasses.

There is a rich body of work that explores link failure resilience in the *context of traffic engineering* in the WAN. We incorporate their methods of achieving failure resilience to the *capacity provisioning problem* of Shoofly. In doing so, not only do we provision network topologies with link failure resilience baked in, we also show that Shoofly's optimization can be extended to use various reliability objectives. Of these, we design and implement two objectives: resilience to deterministic [25] and stochastic link failures [5]. The work on forward fault correction (FFC) ensures that the cloud TE is resilient to up to k deterministic IP link failures. TeaVaR introduced TE in the presence of *probabilistic link failures* to meet availability guarantees [5]. In this section we incorporate the resilience to *possible* (§5.1) and *probable* link failures (§5.2) in Shoofly.

5.1 K-wise link failures

First, we discuss provisioning bypass-enabled cloud topologies resilient to the *possibility* of k simultaneous physical link failures. This resilience guarantees that even if k physical links were to fail, the resulting network after bypasses can continue to meet traffic demands. This is important since k physical link failures can translate to more than k failures in the bypass-enabled topology. Today, cloud providers provision to be resilient to $k \leq 2$ link failures [25].

We formulate the problem of provisioning bypasses under k simultaneous link failures by building on Alg. 1. In addition to the objective and constraints of Alg. 1, this formulation includes a set of constraints for each link failure scenario i, with set of failing links SRLG_i. For instance, when $k = 1$, each SRLG_i contains each duplex edge. Thus, for each failure scenario i and its set of failing links the following constraints ensure that there is a feasible allocation of flow that meets all demands in the bypass-enabled network:

Definition 5.1 (Shoofly under k-wise failures).
Maximize: $\sum_s |s| \cdot w_s$
subject to:
(1) $D_d \leq \sum_{t \in T_d} flow_t, \forall d \in D$
(2)-(8) $AllocationConstraints(flow, x, y, w)$

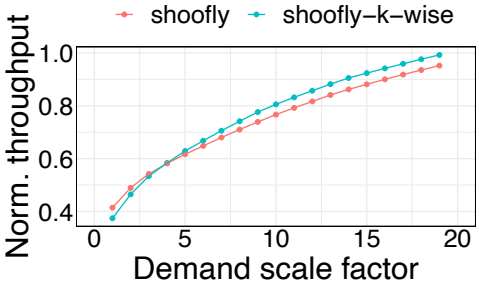


Figure 12: Throughput of failure resililent bypass topology.

for each i

$(w1_i)$ $D_d \leq \sum_{t \in T_d} flow_t^i, \quad \forall d \in D$

$(w2_i)$ $AllocationConstraints(flow^i, x^i, y^i, w)$

$(w3_i)$ $flow_t^i \leq 0, \quad \forall t, e \in t, e \in \text{SRLG}_i$

Provisioning under k-wise failures solves for the objective of maximizing cost savings (as described in Alg. 1) while constraining the problem with the goal of finding feasible flow allocations under *all link failure scenarios*. The failure scenario constraints in Def. 5.1 share the wavelength decision variables, w_s (Alg. 1) but solve for individual flow allocations ($flow_t^i, x^i, y^i$) for each failure scenario i. Thus, the optimization finds wavelength assignments for optical bypasses while ensuring that feasible flow assignments are found for the original network and the network under every failure scenario. The number of constraints of this formulation grow with the number of failure scenarios considered. Recent work has shown ways of translating such optimization formulations to efficiently solvable models [7]. Our current experiments use the less scalable encoding; which is easier to encode and was sufficient for the current evaluation.

Shoofly formulates the k failure resilient bypass provisioning problem using Alg. 1 and Def. 5.1 constraints for all single and double link failures. We compute vanilla (no additional failure resilience), single link failure resilient and double link failure resilient bypass-enabled topologies. We find the cost savings from failure resilient topologies are virtually indistinguishable from the vanilla topologies (Figure 19 in Appendix). Thus, making Shoofly failure resilient *does not reduce hardware cost savings.*

Evaluation. We implement and solve maximum flow traffic engineering on the two bypass-enabled network topologies: the first topology is without failure resilience and the second is resilient to 2 link failures. We solve several instances of the traffic engineering problem by failing 2 randomly selected links in both topologies for each instance. Thus, each TE problem finds traffic allocations on the bypass-enabled topologies in the event of 2 simultaneous link failures. Figure 12 shows the throughput of traffic engineering on the topologies as the demand between nodes is scaled from present-day demands in the cloud network to 20X the present-day demands. As expected, the throughput rises linearly as demand scale increases, until the increase becomes sub-linear due to the network capacity limits. At low demand scales, both networks achieve similar throughput since double link failures do not stress over-provisioned cloud networks. However, at high demand scales, the failure resilient Shoofly topology achieves 5% higher throughput than the vanilla topology.

5.2 Probabilistic link failures

Recent work has proposed a cloud traffic engineering algorithm, TeaVaR, that computes flow allocations that minimize the expected un-met traffic demands, called *loss* or VaR_β, under probabilistic link failure scenarios [5]. The algorithm takes as input the likelihood of link failure scenarios (Q) and target network availability (β) to compute flow allocations. The minimal loss (VaR_β) is guaranteed with probability β and the expectation over all scenarios where the loss is greater than VaR_β is calculated by $CVaR_\beta$ or conditional value at risk. We augment Shoofly with TeaVaR's demand constraints to provision bypasses resilient to probabilistic link failure scenarios. VaR_β and $CVaR_\beta$ are approximated by the outputs α and F_β, in the following optimization problem:

Definition 5.2 (Shoofly with TeaVaR).

Minimize: $F_\beta(\alpha)$

subject to:

(1)	$D_d \leq \sum_{t \in T_d} flow_t,$	$\forall d$
(2)-(8)	$AllocationConstraints(flow, x, y, w)$	
(S)	$\sum_s \lvert s \rvert \cdot w_s \geq S$	$\forall s$
$(t1)$	$F_\beta(\alpha) := \alpha + \frac{1}{1-\beta} \sum_{q \in Q} p_q \cdot s_q$	
$(t2)$	$s_q \geq t_{d,q} - \alpha,$	$\forall d, q$
$(t3)$	$s_q \geq 0,$	$\forall q$
$(t4)$	$t_{d,q} := 1 - \frac{\sum_{t \in T_d} flow'_t z_t(q)}{D_d}$	$\forall d, q$
$(t5)$	$AllocationConstraints(flow', x', y', w)$	

In the formulation, p_q is the probability of the failure scenario q, s_q is the loss in failure scenario q, $z_t(q)$ is 0 iff an SLRG on the tunnel t fails in scenario q. The constraints $(t1) - (t5)$ from Def. 5.2 can be used for fixed values of w for online traffic engineering to minimize the conditional value at risk, $F_\beta(\alpha)$. For optimizing cost savings through bypass, we add constraints (1)-(8) from Alg. 1. Since there are now two competing optimization objectives, maximizing shortcut savings vs. provisioning for stochastic reliability, we introduce constraint (S) for selected lower bounds of savings. By choosing different values of S, Shoofly can find Pareto optical bypasses that meet traffic demands in case of probabilistic failures and save a minimum of S in cost by minimizing the objective $F_\beta(\alpha)$.

Setup. We first we enumerate the likelihoods of link failures in the original network topology by sampling from a Weibull distribution with $k = 0.8$ and $\lambda = 0.0001$ to populate Q, similar to previous work. We use the traffic demands, network topology and enumerated shortcuts of the cloud network (as in §3). Using the simulated failure scenarios in Q, we enumerate availability values (β) for the cloud topology to solve TeaVaR's flow allocation problem. Of the enumerated βs, we choose the one for which TeaVaR can find allocations on the original network with 0 loss and $F_\beta(\alpha) < 0.01$. Equipped with Q and β, we solve for bypass allocations (w_s) using Shoofly and Def. 5.2.

Evaluation. We set S in the savings constraint $\sum_s \lvert s \rvert \cdot w_s \geq S$ to fractions of total savings possible with Shoofly and measure the minimum conditional value at risk (CVaR) calculated in the optimization solution. The total savings possible are found by solving the vanilla Shoofly formulation, without additional failure resilience constraints. Figure 13 shows the relationship between cost savings and CVaR for different maximum shortcut hops. We note that Shoofly with only 3-hop shortcuts can achieve 80% of

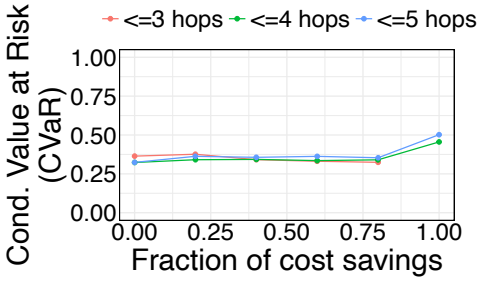


Figure 13: Achievable cost savings vs. CVaR with Shoofly.

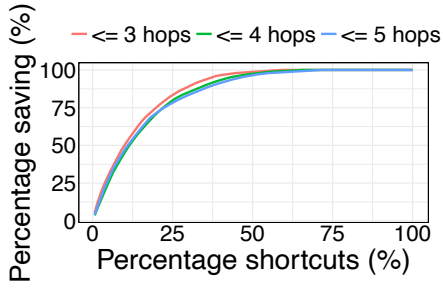


Figure 14: 25% of all bypasses contribute to 80% of all hardware cost savings proposed by Shoofly.

the cost savings possible but when the savings constraint is applied to achieve 100% of the cost savings, a shortcut allocation is not possible. With higher number of shortcut hops, it is possible to achieve the maximum cost savings but the risk of un-met traffic demands increases between 80% and 100% cost saving.

6 OPERATIONAL SAFETY & LOGISTICS

We discuss the implications of a bypass-enabled cloud network topology on the uses it is put to. We focus on the logistical burden of deploying Shoofly's proposed network topology in the cloud WAN. We also discuss the impact of bypass-enabled network on traffic engineering algorithms since these systems and algorithms rely on the network topology to make efficient use of the resources.

6.1 Bypass implementation plan

We devise a *bypass plan* to simplify the logistics of deploying optical bypass in the cloud WAN. We evaluate the contribution of each instance of optical bypass to the hardware cost savings discussed in Section 4. We identify an instance of optical bypass as a triplet of regions $A \rightarrow B \rightarrow C$ where Shoofly proposes that some wavelengths traversing the regions bypass region B en route from A to region C. Each bypass instance represents a unit of logistical overhead faced by the cloud operators to implement a bypass-enabled network. We compute the fraction of total savings enabled by every bypass Shoofly computes. Figure 14 shows the relationship between fraction of shortcuts and the cumulative savings enabled by them. 25% of bypasses contribute to 80% of all hardware cost savings. This shows high return on logistical investment in implementing Shoofly's recommended optical bypasses.

6.2 Traffic engineering with shortcuts

To use Shoofly's bypass-enabled network for TE, the operator must convert the TE tunnels in the original network to new tunnels

Figure 15: Expansion in number of total tunnels after introducing optical bypasses compared to the original network.

in the bypass-enabled topology. This change is needed since (1) edges composing tunnels in the original network may have gone away after bypasses are enabled (2) new edges may have been added between the nodes on the tunnels as shortcuts. Therefore, each original TE tunnel can spawn several tunnels between a pair of nodes. We evaluate the growth in the number of tunnels with the following experiment. For every TE tunnel in the original network, we find the new edges added by bypass and adjust the capacity of existing edges based on Shoofly's output. We calculate all simple paths between the tunnel start and end nodes. We plot the increase in the number of tunnels in the bypass-enabled topology and find that the number of tunnels can increase by a factor of 3 (Figure 15). We suggest that network operators prune the set of new TE tunnels to ensure similar run time of TE algorithms on the new topology.

In place of adding new tunnels, the operator can formulate the TE solver to be shortcut-aware. In this formulation, the standard TE capacity constraints ($\sum_{t \ni e} flow_t \leq c_e$) are replaced by the capacity constraints from Alg. 1, where the output variables w_s are fixed. The remaining output variables capture the allocation of flow on edges and shortcuts. Once the shortcuts capacities have been determined by fixing the number of wavelengths assigned ($\mathbf{w_s}$) to all shortcuts in the network, TE on the resulting network is defined as:

Definition 6.1 (TE with shortcuts).

Maximize: $\sum_{t \in T} flow_t$

subject to:

(TE1) $D_d \geq \sum_{t \in T_d} flow_t, \quad \forall d \in D$

(2)-(8) $AllocationConstraints(flow, x, y, \mathbf{w})$

The inequality on demands is reverse from Alg. 1 since the goal of TE is to maximize throughput using fixed resources.

6.3 Impact of bypass on TE tunnels

Since optical bypasses re-allocate capacity between old and new edges in the network, they can limit communication between some node pairs. Shoofly ensures that demands between all node pairs that communicate *in the present-day cloud network* can be met even if the demand between them increases 8-fold. However, there is no direct traffic demand between some nodes at present and Shoofly can allocate bypasses that prevent the nodes from communication directly. Figure 16 illustrates a case where shortcut allocations can *starve* traffic patterns permissible in the original network.

We can prevent Shoofly from starving traffic patterns that are possible in the original network topology by ensuring that a minimal amount of capacity remains available on edges and tunnels after the allocation of shortcuts. Adding inequalities (9) and (10) to Shoofly's Alg. 1 will achieve this:

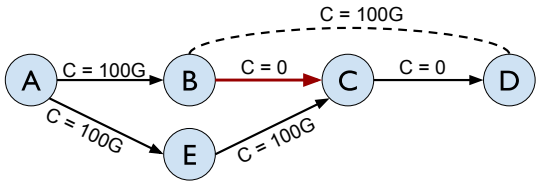


Figure 16: Tunnel AECD is removed when wavelengths of edges BC and CD are allocated to shortcut BD. Nodes B and C can no longer communicate in the bypass-enabled network.

$$x_e \geq lb_e, \quad \forall e \tag{9}$$

$$x_e + \sum_{s \in t, e \in s} u_s \cdot w_s \geq lb_t, \quad \forall t \in T, e \in t \tag{10}$$

7 FREQUENTLY ASKED QUESTIONS

In this section we discuss frequently asked questions about wide-area backbone design in the context of Shoofly.

Why was the network designed in this way? The point-to-point optical backbone was designed to be flexible to meet traffic demands from any pair of source and destination regions. At the time of designing the WAN backbone, flexibility was key since little was known about the traffic that the WAN would carry. Shoofly benefits from observations about the traffic patterns *in hindsight* and proposes optical bypasses to better align the topology with wide-area traffic characteristics.

Can amplifiers expand the optical reach? While amplifiers must be placed at regular intervals on long-haul fiber paths, they cannot be used to expand the optical reach of signals beyond the limits stated in Table 1. Amplifiers boost the signal strength but they also introduce noise in the signal, lowering the SNR.

Does optical bypass need more hardware equipment? Optical bypasses introduced by Shoofly do not need new electrical or optical equipment. In fact, preventing OEO bypasses reduces the hardware required – router ports, optical line port and transceivers.

What happens when new pairs of demands emerge? Shoofly relies on historical information about which regions communicate with each other. These patterns decide the amount of *origin* and *sink* traffic (§2) at WAN regions. However, the emergence of traffic from a new geographical region (*e.g.,* new datacenter deployment) or communication between two existing regions that did not communicate in the past (*e.g.,* due to the launch of a new cloud service) can necessitate re-computing the optical bypasses enabled by Shoofly.

Do the findings apply to other WANs? While our in-depth evaluation (§4) uses the traffic matrices and network topology of one cloud provider, we evaluated Shoofly on several publicly available topologies and demands (§4.1). Shoofly demonstrates the potential of cost savings across different WAN topologies and demands. However, the publicly available demand matrices were limited in details which prevents us from certain evaluations. We believe that our empirical insights about pass-through traffic at WAN regions applies to most large-scale cloud providers.

8 RELATED WORK

In this section we discuss important pieces of work related to Shoofly and set them in the context of our contributions.

Optical network design. Service providers have studied the design of optical networks in depth [3, 4, 6, 9, 16, 27, 28]. We bring two unique perspectives to this rich field of research: first, our analysis leverages optical signal quality and traffic matrices from a production network. Second, our analysis focusses on a large commercial cloud provider. Cloud networks are designed for different workloads than service networks and the demand matrix between regions is a function of the centralized TE algorithms. Internal traffic patterns in SDN controlled cloud networks tend to be stable and predictable [33] helping the design of Shoofly.

Cross-layer network optimizations. Recently, researchers have proposed cross-layer optimizations between IP and physical layers to achieve latency gains for deadline-driven bulk transfers [21]. While related, Shoofly is solving a provisioning problem and not a scheduling one. Similar to the ideas explored by Shoofly, researchers have found that transceivers in data centers can be "stretched" to lower the cost of data center networks [11, 34]. Unlike Shoofly, these works take advantage of high signal quality on data center links to use transceivers for longer distance connections.

Wide area performance monitoring. Researchers have studied optical signal quality in the WAN [12, 14, 15, 30, 31] and found that existing optical signals can be utilized to enable data rates. However, these studies have not utilized the high OSNR to reduce OEO conversion like Shoofly does.

Wide-areaTraffic engineering (TE). Cloud providers have embraced software-defined, centralized TE controllers to assign flow in their WANs to maximize their utilization, guarantee fairness and prevent congestion [18, 20, 24, 29, 31].

Wide-area failure recovery has been extensively studied in the context of network engineering. The original work in [13] considered safe re-allocation of routes without incurring congestion or breaking reachability under network and demand changes. Bringing these ideas to TE, K-wise failure resiliency was developed in [25]. It was first solved using an optimized encoding of enumerating failure scenarios using sorting networks and reformulated using LP dualities in [26]. A general framework based on LP dualities was initiated in [8], and shown practical in [7] based on insights that ensure strong LP dualities. SLA guarantees through TE in probabilistic failure scenarios were explored in [5].

9 CONCLUSION

We analyzed the inter-regional traffic patterns in a cloud WAN and found that 50% of the traffic observed by a region is *passing through* – neither originating or terminating at the region. We propose that such traffic optically bypass regions and stay in the optical domain for as long as is possible, thereby saving hardware costs of long-haul capacity. We propose a tool, Shoofly to find optical bypass opportunities in the WAN such that the hardware cost of long-haul capacity is minimized. We show that despite the physical constraints of limited optical reach of signals on fiber, Shoofly provisions failure-resilient backbones that save 40% of hardware cost using existing network hardware without impacting the network's ability to meet traffic demands.

REFERENCES

[1] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. 1993. *Network Flows: Theory, Algorithms, and Applications*. Prentice hall.

[2] Srivatsan Balasubramanian, Satyajeet Ahuja, Gaya Nagarajan, Andrea Celletti, and Frantisek Foston. 2017. Multilayer planning for facebook scale worldwide network. In *2017 International Conference on Optical Network Design and Modeling (ONDM)*. IEEE, 1–6.

[3] Ajay Kumar Bangla, Alireza Ghaffarkhah, Ben Preskill, Bikash Koley, Christoph Albrecht, Emilie Danna, Joe Jiang, and Xiaoxue Zhao. 2015. Capacity planning for the Google backbone network. (2015).

[4] M. Birk, G. Choudhury, B. Cortez, A. Goddard, N. Padi, A. Raghuram, K. Tse, S. Tse, A. Wallace, and K. Xi. 2016. Evolving to an SDN-enabled isp backbone: key technologies and applications. *IEEE Communications Magazine* (2016). https://doi.org/10.1109/MCOM.2016.7588281

[5] Jeremy Bogle, Nikhil Bhatia, Manya Ghobadi, Ishai Menache, Nikolaj Bjørner, Asaf Valadarsky, and Michael Schapira. 2019. TEAVAR: striking the right utilization-availability balance in WAN traffic engineering. In *Proceedings of the ACM Special Interest Group on Data Communication, SIGCOMM 2019, Beijing, China, August 19-23, 2019*. ACM. https://doi.org/10.1145/3341302.3342069

[6] A. Brzezinski and E. Modiano. 2005. Dynamic reconfiguration and routing algorithms for IP-over-WDM networks with stochastic traffic. *Journal of Lightwave Technology* 23, 10 (2005), 3188–3205. https://doi.org/10.1109/JLT.2005.855691

[7] Yiyang Chang, Chuan Jiang, Ashish Chandra, Sanjay G. Rao, and Mohit Tawarmalani. 2020. Lancet: Better network resilience by designing for pruned failure sets. In *Abstracts of the 2020 SIGMETRICS/Performance Joint International Conference on Measurement and Modeling of Computer Systems, Boston, MA, USA, June 8-12, 2020*, Edmund Yeh, Athina Markopoulou, and Y. C. Tay (Eds.). ACM, 53–54. https://doi.org/10.1145/3393691.3394195

[8] Yiyang Chang, Sanjay G. Rao, and Mohit Tawarmalani. 2017. Robust Validation of Network Designs under Uncertain Demands and Failures. In *14th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2017, Boston, MA, USA, March 27-29, 2017*, Aditya Akella and Jon Howell (Eds.). USENIX Association. https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/chang

[9] Angela L Chiu, Gagan Choudhury, George Clapp, Robert Doverspike, Mark Feuer, Joel W Gannett, Janet Jackel, Gi Tae Kim, John G Klincewicz, Taek Jin Kwon, et al. 2011. Architectures and protocols for capacity efficient, highly dynamic and highly resilient core networks. *IEEE/OSA Journal of Optical Communications and Networking* (2011).

[10] Cisco. (Accessed on 2021-01-20). What is MPLS - Multiprotocol Label Switching. https://www.cisco.com/c/en/us/products/ios-nx-os-software/multiprotocol-label-switching-mpls/index.html. ((Accessed on 2021-01-20)).

[11] Vojislav Dukic, Ginni Khanna, Christos Gkantsidis, Thomas Karagiannis, Francesca Parmigiani, Ankit Singla, Mark Filer, Jeffrey L. Cox, Anna Ptasznik, Nick Harland, Winston Saunders, and Christian Belady. 2020. Beyond the Mega-Data Center: Networking Multi-Data Center Regions. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '20)*. Association for Computing Machinery, New York, NY, USA, 765–781. https://doi.org/10.1145/3387514.3406220

[12] Mark Filer, Jamie Gaudette, Yawei Yin, Denizcan Billor, Zahra Bakhtiari, and Jeffrey L. Cox. 2019. Low-margin optical networking at cloud scale [Invited]. *J. Opt. Commun. Netw.* 11, 10 (Oct 2019), C94–C108. https://doi.org/10.1364/JOCN.11.000C94

[13] Klaus-Tycho Förster, Ratul Mahajan, and Roger Wattenhofer. 2016. Consistent updates in software defined networks: On dependencies, loop freedom, and blackholes. In *2016 IFIP Networking Conference, Networking 2016 and Workshops Vienna, Austria*. IEEE Computer Society. https://doi.org/10.1109/IFIPNetworking.2016.7497232

[14] Monia Ghobadi, Jamie Gaudette, Ratul Mahajan, Amar Phanishayee, Buddy Klinkers, and Daniel Kilper. 2016. Evaluation of Elastic Modulation Gains in Microsoft's Optical Backbone in North America. *Optical Fiber Communication Conference* (2016), M2J.2.

[15] Monia Ghobadi and Ratul Mahajan. 2016. Optical layer failures in a large backbone. In *Proceedings of the 2016 Internet Measurement Conference*. 461–467.

[16] Jennifer Gossels, Gagan Choudhury, and Jennifer Rexford. 2019. Robust network design for IP/optical backbones. *IEEE/OSA Journal of Optical Communications and Networking* 11, 8 (2019), 478–490.

[17] Gurobi. (Accessed on 2019-10-02). GUROBI Optimization. https://www.gurobi.com/. ((Accessed on 2019-10-02)).

[18] Chi-Yao Hong, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Vijay Gill, Mohan Nanduri, and Roger Wattenhofer. 2013. Achieving High Utilization with Software-driven WAN. *SIGCOMM* (2013), 12.

[19] Infinera. (Accessed on 2021-01-10). Optical Line Systems. https://www.infinera.com/optical-line-systems. ((Accessed on 2021-01-10)).

[20] Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, Jon Zolla, Urs Hölzle, Stephen Stuart, and Amin Vahdat. 2013. B4: Experience with a Globally-deployed Software Defined Wan. *SIGCOMM* (2013), 12.

[21] Xin Jin, Yiran Li, Da Wei, Siming Li, Jie Gao, Lei Xu, Guangzhi Li, Wei Xu, and Jennifer Rexford. 2016. Optimizing bulk transfers with software-defined optical WAN. In *Proceedings of the 2016 ACM SIGCOMM Conference*. 87–100.

[22] Juniper Network. (Accessed on 2021-01-10). Shared Risk Link Groups for MPLS. https://www.juniper.net/documentation/en_US/junos/topics/topic-map/srlg-for-mpls.html. ((Accessed on 2021-01-10)).

[23] Praveen Kumar, Yang Yuan, Chris Yu, Nate Foster, Robert Kleinberg, Petr Lapukhov, Chiun Lin Lim, and Robert Soulé. 2018. Semi-oblivious traffic engineering: The road not taken. In *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*.

[24] Wenxin Li, Xiaobo Zhou, Keqiu Li, Heng Qi, and Deke Guo. 2018. Trafficshaper: shaping inter-datacenter traffic to reduce the transmission cost. *IEEE/ACM Transactions on Networking* 26, 3 (2018), 1193–1206.

[25] Hongqiang Harry Liu, Srikanth Kandula, Ratul Mahajan, Ming Zhang, and David Gelernter. 2014. Traffic engineering with forward fault correction. In *ACM SIGCOMM 2014 Conference, SIGCOMM'14, Chicago, IL, USA, August 17-22, 2014*, Fabián E. Bustamante, Y. Charlie Hu, Arvind Krishnamurthy, and Sylvia Ratnasamy (Eds.). ACM, 527–538. https://doi.org/10.1145/2619239.2626314

[26] Hongqiang Harry Liu and Jian Li. (Accessed on 2020-12-19). O(n) improve. ((Accessed on 2020-12-19)).

[27] P. Papanikolaou, K. Christodoulopoulos, and E. Varvarigos. 2017. Joint multilayer survivability techniques for IP-over-elastic-optical- networks. *IEEE/OSA Journal of Optical Communications and Networking* 9, 1 (2017), A85–A98. https://doi.org/10.1364/JOCN.9.000A85

[28] P. Papanikolaou, K. Christodoulopoulos, and E. Varvarigos. 2018. Optimization techniques for incremental planning of multilayer elastic optical networks. *IEEE/OSA Journal of Optical Communications and Networking* 10, 3 (2018), 183–194. https://doi.org/10.1364/JOCN.10.000183

[29] Rachee Singh, Sharad Agarwal, Matt Calder, and Paramvir Bahl. 2021. Cost-effective Cloud Edge Traffic Engineering with Cascara. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. USENIX Association, 201–216. https://www.usenix.org/conference/nsdi21/presentation/singh

[30] Rachee Singh, Monia Ghobadi, Klaus-Tycho Foerster, Mark Filer, and Phillipa Gill. 2017. Run, Walk, Crawl: Towards Dynamic Link Capacities. In *Proceedings of the 16th ACM Workshop on Hot Topics in Networks (HotNets-XVI)*. Association for Computing Machinery, New York, NY, USA, 143–149. https://doi.org/10.1145/3152434.3152451

[31] Rachee Singh, Manya Ghobadi, Klaus-Tycho Foerster, Mark Filer, and Phillipa Gill. 2018. RADWAN: Rate Adaptive Wide Area Network. ACM SIGCOMM. https://www.microsoft.com/en-us/research/publication/radwan-rate-adaptive-wide-area-network/

[32] TeleGeography. (Accessed on 2020-01-20). Wavelengths Pricing Data. https://www2.telegeography.com/wavelengths-pricing-data. ((Accessed on 2020-01-20)).

[33] Asaf Valadarsky, Michael Schapira, Dafna Shahaf, and Aviv Tamar. 2017. Learning to route. In *Proceedings of the 16th ACM workshop on hot topics in networks*. 185–191.

[34] Danyang Zhuo, Monia Ghobadi, Ratul Mahajan, Amar Phanishayee, Xuan Kelvin Zou, Hang Guan, Arvind Krishnamurthy, and Thomas Anderson. 2017. RAIL: A Case for Redundant Arrays of Inexpensive Links in Data Center Networks. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. USENIX Association.

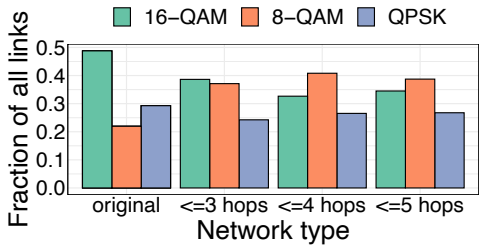


Figure 17: Modulation formats of all links in the original and Shoofly proposed networks.

A APPENDIX

Appendices are supporting material that has not been peer-reviewed.

A.1 Flow conservation in Shoofly

Lemma 1. *Every solution that satisfies AllocationConstraints also has a solution where inequalities (4) are tight and corresponds to a network flow. That is, the flow entering each internal node in a tunnel equals the flow leaving.*

Proof. First note that the inequalities on y_s^t and x_e^t are convex, equations (2) and (3) remain satisfied when decreasing their values. Therefore a solution to the inequalities (4) implies that there is a solution where the inequalities are tight equalities. Let v be an internal vertex on a tunnel t with incoming edge e and departing edge e', then a tight solution to (4) implies that

$$x_e^t + \sum_{s \ni e} y_s^t = x_{e'}^t + \sum_{s \ni e'} y_s^t.$$

Shortcuts that don't terminate in v are included on both sides of the equalities. They cancel out. The equality is preserved for the remaining flows. □

A.2 Overall link modulation formats

We discussed the modulation formats of signals in bypass-enabled topologies in Section 4. In Figure 17 we show the comparison of modulation formats of signals between the original network topology and the bypass-enabled topologies. Nearly 50% of signals in the original network could sustain 16-QAM format but this fraction declines by nearly 10% in the bypass-enabled topology.

A.3 Throughput of TE on bypass-enabled networks.

We discussed the decline in per-wavelength capacity due to bypasses in Section 4. In addition to the average link capacity decline, we simulate the throughput of traffic engineering on both Shoofly provisioned topologies without additional failure resilience and with resilience to 2 simultaneous link failures. First, we show that at as demand is scaled higher, the failure resilient topologies achieve higher network throughput (Figure 18). Second, regardless of the scale of demands, topologies with longer shortcut lengths achieve less throughput, especially at high demand scale factors. At present day demands, the difference between the throughput of 3, 4 and 5 hop shortcut topologies is very similar.

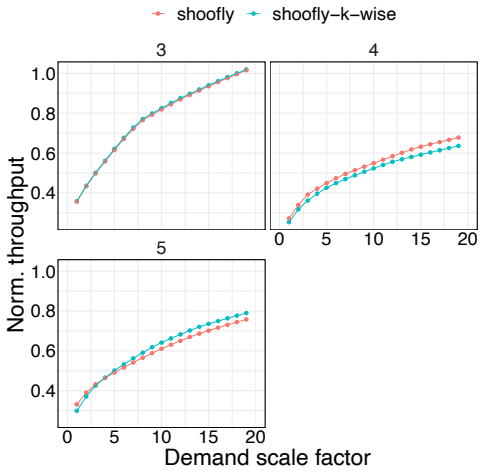


Figure 18: Norm. throughput of traffic engineering on 3, 4, and 5 hop shortcut topologies proposed by Shoofly.

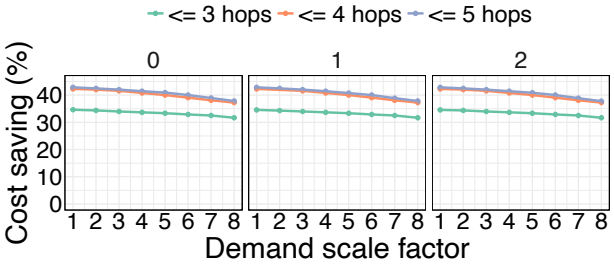


Figure 19: Cost savings by Shoofly in k = 0, 1 and 2 failure resilient scenarios. The cost savings are indistinguishable in all three cases showing that Shoofly can enable k-wise failure resilience without sacrificing cost savings.

A.4 Failure resilience vs. cost saving

In Section 5 we mentioned that k-wise link failure resilience can be incorporated in Shoofly without sacrificing cost savings. We demonstrate this in Figure 19 – cost savings are indistinguishable in case of k = 0, 1 and 2, showing that Shoofly can enable k-wise failure resilience without sacrificing cost savings even at highly scaled demands.