Beyond the mega-data center: networking multi-data center regions

Vojislav Dukic[1,2] Ginni Khanna[1] Christos Gkantsidis[1] Thomas Karagiannis[1] Francesca Parmigiani[1] Ankit Singla[2]
Mark Filer[1] Jeffrey L. Cox[1] Anna Ptasznik[1] Nick Harland[1] Winston Saunders[1] Christian Belady[1]
Microsoft Research[1], ETH Zurich[2]

ABSTRACT

The difficulty of building large data centers in dense metro areas is pushing big cloud providers towards a different approach to scaling: multiple smaller data centers within tens of kilometers of each other, comprising a “region”. We show that networking this small number of nearby sites with each other is a surprisingly challenging and multi-faceted problem. We draw out the operational goals and constraints of such networks, and highlight the design trade-offs involved using data from Microsoft Azure’s regions.

Our analysis of the design space shows that network topologies that achieve lower latency and allow greater flexibility in data center placement are, unfortunately, encumbered by their much greater cost and complexity. We thus present and demonstrate a novel optical-circuit-switched architecture, Iris, that lowers these cost and complexity barriers, making a richer topology design space more accessible to operators of regional networks. With Iris, topologies which, in comparison to a simple hub-and-spoke topology can increase the area in which a new DC can be placed by 2-5×, can be implemented at a cost within 1.1× of the simple hub-and-spoke topology, and 7× cheaper than a natural packet-switched network.

CCS CONCEPTS

• **Networks → Network design principles**; **Network design and planning algorithms**; **Data center networks**;

KEYWORDS

data center interconnect, DCI, optical networks, cloud, region, optical switching

ACM Reference Format:
Vojislav Dukic, Ginni Khanna, Christos Gkantsidis, Thomas Karagiannis, Francesca Parmigiani, Ankit Singla, Mark Filer, Jeffrey L. Cox, Anna Ptasznik, Nick Harland, Winston Saunders, Christian Belady. 2020. Beyond the mega-data center: networking multi-data center regions. In *Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication (SIGCOMM ’20), August 10–14, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3387514.3406220

[This work does not raise any ethical issues.]

1 INTRODUCTION

Cloud computing’s growth has forced commensurate scaling of data center (DC) infrastructure. Until recently, such scaling meant building “mega”-DCs with hundreds of thousands of servers across the world, and interconnecting them into a wide-area backbone.

However, a different scaling strategy has quickly become standard industry practice. Instead of serving each broad geographic area from just one or two mega-DCs, in many geographies, large cloud providers have transitioned to using a collection (typically 5-20) of smaller DCs within tens of kilometers of each other, referred to as a “region”. This shift away from mega-DCs is driven by two pressures: (a) the difficulty of siting and provisioning large facilities in or near dense metro areas due to limited resources such as land, power and connectivity; and (b) the desire for fault tolerance in the face of losing one or two large facilities to catastrophes like flooding and earthquakes. These fundamentals have forced all of the largest DC operators, including Amazon [3], Facebook [44], Google [45], and Microsoft [20], to increasingly rely on such regions.

Large volumes of traffic flow between DCs in a region, thus requiring a high-capacity network typically referred to as a regional Data-Center Interconnect (DCI). The growth of the DCI has led to it incurring significant costs for cloud providers, as, for example, seen by the the explosive increase in the total number of 100G ports deployed: there are two orders of magnitude more regional DC-to-DC ports than WAN-facing ports [20]. High capacity notwithstanding, superficially, the design of such DCIs appears trivial:

- The number of DCs to interconnect is small.
- Each DC has a known available capacity.
- DCs are only a few tens of kilometers apart at most.
- DC-to-DC traffic is expected to be relatively stable.

Yet, as we shall show, DCI design is challenging due to several operational, cost, and technological constraints (§4) that are different from those for both intra-DC networks, and DC-WANs used for inter-region connectivity. These constraints lead to complex decisions on both the network’s topology, and how this topology is realized with appropriate switching technology.

Thus, we broadly address the question: *how should DCI networks be designed?* We outline the design space of DCI topologies, ranging from fully centralized ones with all DCs connected to two hubs, to distributed ones that either eschew such hubs entirely, or reduce dependence on them, by building closer or direct connectivity between some subsets of DCs. We show that DCI design involves more nuance than just the clichéd centralized-distributed dichotomy may suggest, fleshing out its complexity by: (a) analyzing data from several of Microsoft Azure’s regions; and (b) performing testbed experiments that demonstrate the physical-layer constraints.

Our analysis shows that distributed topologies provide much lower DC-DC latency than DC-hub-DC connectivity: compared to a centralized topology, latency reduces for at least 60% of DC-DC

SIGCOMM ’20, August 10–14, 2020, Virtual Event, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7955-7/20/08…$15.00
https://doi.org/10.1145/3387514.3406220

paths, and in more than 20% of cases, the latency is >2× lower. This advantage is of high and growing value: customers are increasingly asking for lower latency service level agreements, and latency-sensitive applications like synchronous replication are going mainstream at the region level [13]. While the latency advantage of direct DC-DC connectivity is unsurprising, we also show that distributed topologies increase flexibility in terms of choosing DC sites. In the analyzed regions, the area in which new DCs could be located increases by 2-5× with distributed topologies.

Unfortunately, as they would be implemented today, with electrical packet switching, distributed topologies fare badly compared to centralized approaches across two key metrics: cost and complexity. The centralized approach is much more cost effective, by as much as 7× in the settings we studied, and is significantly easier to manage, requiring a much smaller number of ports.

To lower the cost and complexity barriers in DCI network design, we propose and demonstrate Iris, which uses an all-optical circuit-switched network core. Compared to electrical DCI networks, Iris simplifies network structure, reducing the total number of ports. The resulting reductions in cost and complexity benefit networks on the entire spectrum from fully centralized to fully distributed, but are much larger for larger-scale regions and more distributed network designs. Thus, Iris makes more of the design space practicable, unlocking the latency and siting flexibility advantages of distributed networks while lowering their cost and complexity. Note that Iris substantially reduces, but does not completely ameliorate the complexity of distributed design, which, with *any* architecture, necessitates the management of in-network equipment across multiple sites, instead of just two hubs. But if the pressure for low latency persists, a shift towards distributed designs may be inevitable.

Iris exploits two key observations: (a) DCI cost is dominated by the specialized electrical-optical transceivers needed for covering DCI distances, and (b) regional fiber is abundant and cheap relative to transceiver cost. Iris's design thus makes an extremely favorable cost trade: some additional fiber in exchange for vastly reducing the number of transceivers. To exploit this cost structure, Iris's all-optical approach gives up the finer switching granularity of packet switching in favor of coarser optical switching.

While optical switching is well-studied for both intra-DC and DC-WAN networks, the constraints of regional DCIs present unique challenges and opportunities. Unlike intra-DC optics [17, 18, 22, 23], fast reconfigurability is not necessary as the traffic is slow-changing; the challenges rather stem from the physical layer, which needs to ensure that the budgets of optical devices for power and signal quality are respected across a wide range of distances, and through a varying number of optical switches. On the other hand, while optical DC-WAN networking accounts for even more stringent physical-layer constraints due to the long and diverse distances, the solutions there typically involve optimizing spectral efficiency and switching at the wavelength granularity, *e.g.,* OWAN [28]. For DCI networking, we find that this is more complex than necessary. Instead, Iris, only switches capacity at fiber granularity, thus requiring minimal support from the physical layer. We find that wavelength switching is *more* expensive for DCIs, making Iris the preferable solution in both cost and complexity.

Using the same data and testbed mentioned above, augmented with large-scale simulations, we evaluate the benefits and feasibility of Iris. We find that Iris: (a) can be implemented using off-the-shelf hardware; (b) involves limited reconfiguration that does not hurt application-layer performance; (c) enables the latency and location-flexibility advantages of the distributed approach; (d) allows the distributed approach to be implemented at a cost within 1.1× of a traditional centralized approach, and in fact, *cheaper* than it in more than 98% of the settings examined; and (e) reduces network complexity by reducing the total number of ports, electrical or optical, that need to be managed.

In summary, we make the following contributions:

- We quantitatively flesh out the trade-offs of regional DCI design: compared to a centralized network, a distributed network increases flexibility in placing new DCs (2-5× more area) and cuts latency (by > 2× in 20% of cases), but is costlier (7×) when implemented using packet switching.
- We propose, Iris, an all-optical network architecture that lowers the cost and complexity of DCI design. Iris's benefits are larger for more distributed topologies, thus making their latency and siting flexibility benefits more accessible to operators.
- We show how Iris can be appropriately provisioned to provide non-blocking connectivity along shortest paths, and to meet any specified constraints on resilience to fiber cuts. Our analysis shows that Iris is >2× cheaper than a packet-switched network, even when Iris guarantees capacity under up to two failures, and the packet-switched network provides no guarantees.
- Using a testbed incorporating *all* the optical components used in Iris, we demonstrate that Iris meets its optical layer constraints *without* a complex synchronized, online control plane to manage optical components.
- We show through simulations that Iris's infrequent reconfigurations do not hurt application-layer performance.

2 THE DCI NETWORK DESIGN PROBLEM

A regional DCI connects 5-20 DC sites within tens of kilometers. The problem of network design in this setting requires 3 **inputs**:

DC site locations. Our focus is the network; DC siting is itself an interesting problem, but requires separate treatment as many of the involved factors are non-networking, *e.g.,* the particular buildings available, their cost, connectivity to not just network providers, but power and ground transit infrastructure, etc.

DC capacities. Based on each DC's size and other business factors, we know each DC's network capacity, *i.e.,* how much traffic a DC can maximally send or receive to other DCs in the region. For convenience, we translate the Gbps capacity into a number of fibers, *e.g.,* capacity B Gbps translates to ${}^{B}/_{C \cdot \lambda}$ fibers, where λ is the number of wavelengths per fiber, and C the bandwidth per wavelength in Gbps. In this example, $P = {}^{B}/_{C}$ is the number of electrical ports, *i.e.,* transceivers, required at each DC.

Fiber map. The region's available fiber is known, in terms of fiber ducts between two types of nodes: DCs and "fiber huts", which are intermediate nodes housing switching and other equipment like amplifiers. Where convenient, huts can co-exist with DCs. For our purposes, fiber ducts are unconstrained in the fiber available to lease: each fiber duct contains hundreds of individual fibers, with typically only a fraction of those lit. This is standard industry practice to amortize the cost of constructing a duct.

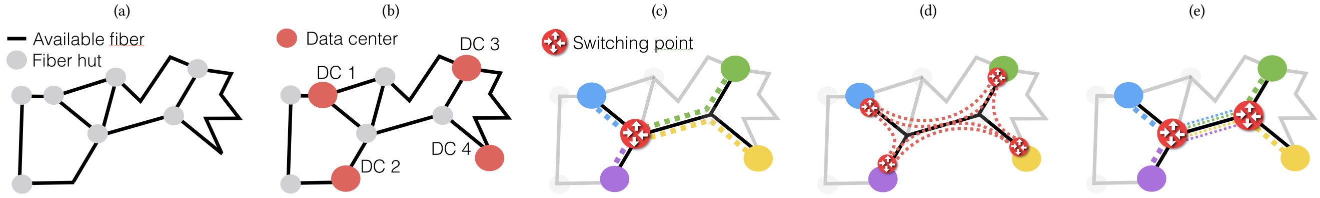

Fig. 1: *DCI design example: (a) The fiber map, which contains all available fiber ducts and huts. (b) The region has* 4 *DCs for which DCI connectivity is to be determined. (c) The centralized approach uses a hub to which all DCs connect; in practice* 2 *hubs are used for resilience, but for clarity only one is shown. (d) An extreme version of the distributed approach, with all pairs of DCs connected directly to each other. (e) A sparser distributed approach, with two pairs of DCs – each pair connects to a hub, and the two hubs connect to each other.*

The above inputs are *outside* the network designer's control. DC sites and capacities are set by operational needs. Expanding the fiber map is possible in some regions, but is typically avoided: it is time-consuming, has a high up-front cost, and is unlikely to improve routes, especially in dense metro areas, that already have plentiful fiber and are space-constrained against further expansion.

A simple example of DCI input specification is shown in Fig. 1. The region's fiber map, including *all* available fiber huts and ducts, is shown in Fig. 1(a), and the 4 DCs the operator has built or plans to build in this region are shown in Fig. 1(b). For this running example, we will assume that all DCs have the same capacity of f fibers each.

Given the DC sites, capacities, and fiber map, we must decide on the following **outputs**:

- Topology: which DC-DC connections are *direct*, *i.e.,* without needing intermediate routing at other DCs or huts? This decision dictates the subset of the fiber map that is used, *i.e.,* which huts and ducts are needed.
- Capacity: what number of fibers are leased in each fiber duct?
- Switching: how is switching (*e.g.,* electrically vs optically) implemented at the DCs and huts?

Loosely, one can think of the topology and capacity decisions as provisioning problems, answers to which depend on the design **goals**: Do we insist on shortest path connectivity, or are longer paths acceptable? Do we provision non-blocking connectivity between all DCs, or is an oversubscribed fabric acceptable? How much failure resilience do we need in terms of fail-over paths?

Switching, on the other hand, is more tied to implementation: What equipment is used at DCs and fiber huts, and how is it interconnected such that it correctly instantiates the topology and capacity decisions? The industry's standard method of switching is to deploy electrical switches. The data travels on each fiber in optical wavelengths, and at given switching points, it leaves the optical domain, such that switches can reroute data as necessary.

However, there is a complex interplay between topology and capacity, and switching: the switching technology can place **constraints** on the topology. For instance, an uninterrupted run of fiber, without amplification or termination at a DC or hut, referred to as a "fiber span", cannot be longer than a particular length.

While we will make the goals and constraints more precise in §3, the above context suffices to examine the design space and trade-offs for DCI networks in terms of two broad approaches.

The centralized approach uses a hub-and-spoke topology: DCs in a region all connect to a centralized hub. In the example in Fig. 1(c), one of the huts is used as a hub, and no other huts are used. There are no direct DC-DC connections, with all connectivity going through the hub. For a non-blocking interconnect, the fiber ducts connected directly at the four DCs will carry f fiber-pairs to connect each DC's full capacity to the hub, where sufficient switching hardware must be provisioned. The remaining central duct carries the $2 \cdot f$ fiber-pairs from the two DCs on the right.

For simplicity, we illustrate and discuss only one hub in our example, but for failure resilience, two hubs are used, and each DC connects to both. The hubs provide a "big switch" abstraction, whereby all DC-pairs are connected in a non-blocking fashion to each other. This approach is presently used in Microsoft Azure [20].

The distributed approach directly connects DCs to each other. An extreme version of this approach would build all pairs of DC-DC connections, *i.e.,* $O(n^2)$ for n DCs, like in Fig. 1(d). In this example, for non-blocking connectivity, $3 \cdot f$ fiber-pairs are needed at the four fiber ducts that originate at the DCs (one fiber-pair each for the other three DCs), with $12 \cdot f$ fiber-pairs on the central duct. We also highlight here the aforementioned interplay with switching: due to technology constraints, it may not be possible to instantiate this design as is, *e.g.,* because some of the DC-pairs that we want to connect directly are too far to be connected over an uninterrupted fiber span, and need amplification at a hut in between.

More generally, one can build a variety of sparser distributed networks, with some DC-DC pairs eschewing direct connectivity in favor of transit through other DCs or huts. An example of this is shown in Fig. 1(e), where two pairs of DCs connect to hubs, with the hubs connecting to each other. In this case, for non-blocking connectivity, f fiber-pairs are needed on the 4 DC-incident fiber ducts, and $2 \cdot f$ fiber-pairs on the central duct. From public resources [3], it appears Amazon AWS broadly uses this approach.

Note: In the above discussion, we highlighted the amount of fiber used primarily to clarify how different connectivity models can be instantiated atop a given fiber map. However, the impact of the design choices is much deeper than just the quantity of fiber used. Different solutions achieve vastly different outcomes in the tradeoff space involving performance, reliability, operational flexibility, and cost, as we discuss next.

2.1 Outcome #1: Latency

An obvious distinction in the centralized and distributed models is the propagation latency they provide between DCs — the distributed approach, provided the right DC-DC links are provisioned, can substantially lower latency by eschewing transit through a hub. Fig. 2 demonstrates this contrast in the Tokyo region.[1] The two

[1] Example regions and fiber maps used throughout the paper use mock-up drawings that resemble but do not represent Microsoft Azure's network maps.

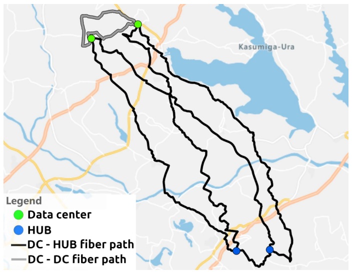

Fig. 2: *DC-hub-DC paths can sometimes be much longer than DC-DC ones.*

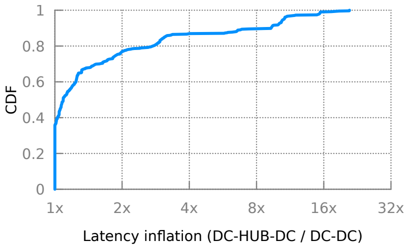

Fig. 3: *Latency inflation of paths via a hub compared to direct ones.*

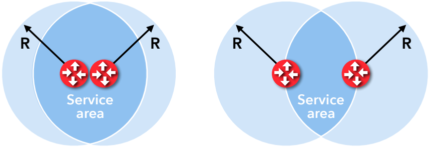

Fig. 4: *Reliability vs. flexibility in the centralized approach. The circles are for intuition; in practice, we must consider real fiber distances.*

hubs are located South of two of the DCs in the region. The DC-hub connections are 53-60 km in terms of fiber distance, resulting in a maximum DC-DC roundtrip latency of 1.2 ms. In contrast, a direct DC-DC connection of 19 km would achieve a 0.2 ms latency, a 6× latency reduction.

Fig. 3 investigates this latency inflation by using Microsoft Azure's DC locations across 22 regions. In some cases, direct DC-DC paths can reduce roundtrip propagation latency by several times, similar to the example in Fig. 2; in more than 20% of cases, the reduction is more than 2×[2]. As not all DCs are connected to one another in these regions, we estimate DC-DC latency using an industry rule of thumb: multiplying the geo-distance by 2× [8, 15].

The astute reader will notice from Fig. 2 that part of the reason the DC-hub-DC paths are much longer is that both hubs are close to each other – if they were more spread out in the region, in many cases, at least one hub-path could be much shorter. Unfortunately, the hub placement is not this flexible, as we discuss next.

2.2 Outcome #2: Siting flexibility

Bounding DC-DC latency requires constraining the locations of DCs and hubs. The maximum latency allowed between any two DCs is typically specified in Regional Service-Level Agreements (SLAs) that implicitly define the maximum DC-DC fiber distance — Azure limits fiber-distance to 120 km for any DC pair [20]. Analyzing data from Microsoft Azure's regions, we show that the resulting siting constraints are much more rigid for the centralized design than the distributed one, making the latter preferable for maximizing deployment flexibility.

For the centralized approach, the 120 km limit restricts each DC-hub connection to at most 60 km of fiber. Thus, once the hubs are placed, a service area for placing DCs is determined as the intersection of their 60 km-radii, as shown in Fig. 4. Comparing the left and right parts of Fig. 4, we see that placing hubs close to each other would maximize the permissible service area (intersection). But this comes at the cost of latency and reliability: (a) if hubs are placed close to each other, DC-hub-DC paths can be longer; and (b) if one hub is lost to a catastrophic event, the other is more likely to be also affected if it is nearby. Thus, in practice, operators using a centralized DCI approach must trade-off latency and reliability if they want greater DC siting flexibility.

In contrast, the distributed approach, by eschewing hubs, simplifies DC siting and alleviates the difficult flexibility-reliability trade-off. We show in Fig. 5 this contrast visually for 4 regions, in the form of permissible area for siting one new DC given existing DCs or hubs. The top and bottom rows of the figure are for the same regions, except in the top row, the hubs are placed nearby (within 4-7 km of each other), while in the bottom row, they are farther apart (20-24 km). For the centralized approach, the service area is smaller when the hubs are closer. The service area for the distributed approach remains the same across the top and bottom rows as it does not use or depend on hubs. In each case, the distributed approach allows much higher flexibility in picking DC sites. This analysis uses real fiber maps and distances, and the same criteria as cloud operation teams follow for DC and hub placement.

Using similar analysis, Fig. 6 shows that the permissible siting area for one new DC (given existing sites) would increase by 2–5× across 33 existing regions with the distributed approach compared to the centralized one. Even though each additional DC that is built constrains future sites in the distributed approach, it is still much more flexible than the centralized one — the number of DCs in the regions used for this analysis ranges from 5–15 existing DCs, with regions with more DCs showing (as expected) smaller, but still sizable (at least 2×), benefits with the distributed approach.

The size of the service area greatly impacts deployment costs and the availability of critical resources like space, especially in busy metro areas. Even a small increase in service area can provide significant flexibility for a provider and reduce capital costs.[3]

2.3 Outcome #3: Implementation ease

The implementation of the centralized approach is simple, effectively breaking up a mega-DC into multiple sites — the uppermost (core) switching tier of what would have otherwise been a mega-DC resides at the hubs, such that connections between this and lower topology tiers are now externalized fiber connections traversing a few tens of kilometers. Operationally, the first step is picking the sites for the hubs and provisioning them anticipating the needed switching capacity. Then over time, the DCs are built such that each DC is within a threshold fiber distance from each hub — as all DC-DC connectivity traverses a hub, this constraint ensures that DC-DC distances (latencies) are bounded per the SLA. The big-switch abstraction further eases management and provisioning; each DC connects all its capacity to the central switching fabric, where a non-blocking network connects it to other DCs. This approach can be easily replicated across regions irrespective of the underlying fiber layout.

A distributed approach requires greater design effort in planning which DC-DC connections are made and at what capacity, such

[2] Inter-connecting DCs within Availability Zones [20] may alleviate some of this latency inflation of centralized topologies similar to semi-distributed topologies as in Fig. 1(e).

[3] Land scarcity has even motivated building vertical DCs [21].

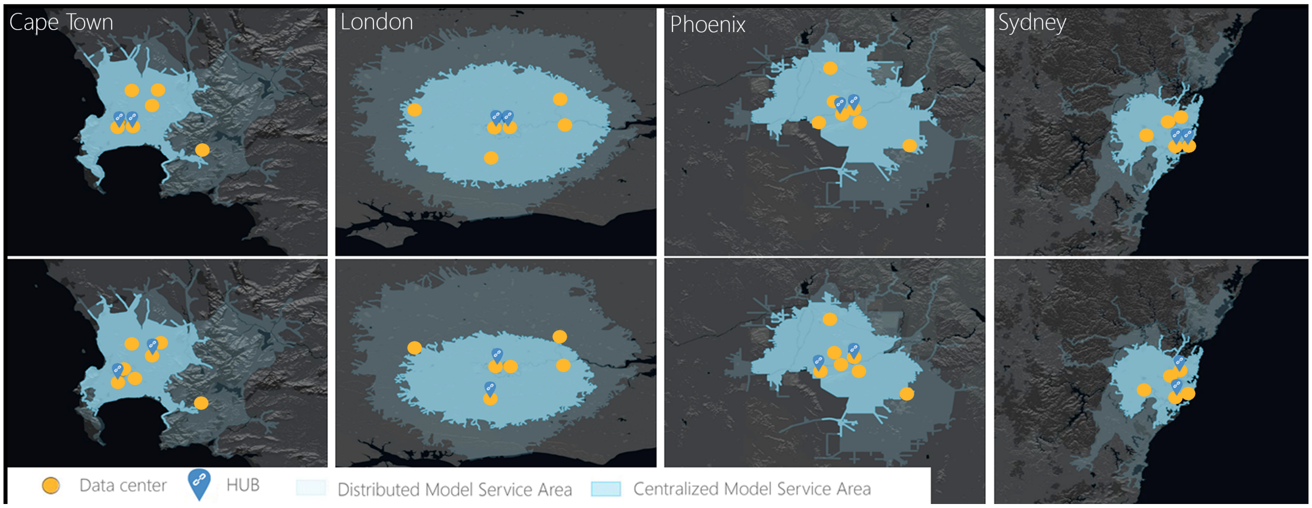
Cape Town
London
Phoenix
Sydney
● Data center
HUB
Distributed Model Service Area
Centralized Model Service Area

Fig. 5: *The distributed approach expands available area for building new DCs. These maps are for hypothetical regions, but with DC and hub placement using real criteria as analyzed by Microsoft Azure's deployment team. The top row shows results with hubs within 4–7 km, and the bottom within 20–24 km. Maximum allowed fiber distance for all DC-DC communication is 120 km for both models. In the distributed model, DCs can be placed in the extended shaded area, which is out of reach in the centralized model.*

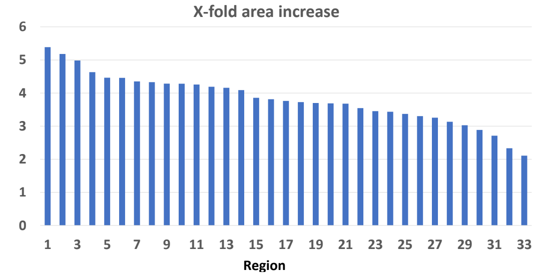

Fig. 6: *Across existing regions (different bars) service area increases by* 2-5× *with a distributed approach compared to a centralized one.*

that appropriate infrastructure can be provisioned at each DC. Operationally, the first DCs can be built in a relatively unconstrained manner, but later DCs must be within a fiber distance threshold of each existing DC. Once it is determined which physical DC-DC links will be built, one must decide on routing such that each DC-DC pair has a path, direct or otherwise, with enough capacity. Given the physical links and routing, DC-DC link capacity can thus be determined, and implemented at the physical fiber layer. For traffic from DC A to C transiting through DC B, the A-B fiber carries both direct A-B traffic and A-C traffic. The A-C traffic is switched using electrical switches installed at B, requiring conversion from the optical domain to electrical, followed by electrical switching, followed by conversion to optics again. Thus, capacity provisioning must account for transit capacity appropriately. Further, small DC facilities are typically severely constrained in terms of available power and space resources and supporting connectivity to multiple other DCs may not be feasible. Thus, care needs to be taken as to which DCs can be inter-connected beyond just fiber capacity.

Thus, for provisioning, the centralized approach is a natural extension of today's Clos networks, while the distributed approach needs additional design effort. Further, expanding a region to add more DCs or capacity at existing DCs also poses different challenges for the two approaches. Centralized DCIs require the hubs to have enough space and power for the *maximum* predicted region scale; accommodating unanticipated growth in a region is thus difficult. The distributed approach requires similar provisioning at multiple (smaller) switching points when a region is expanded.

2.4 Outcome #4: Cost

While we defer a complete cost analysis to §3.3, we can use regional network port counts to coarsely flesh out the design space.

To understand the cost implications of supporting distributed topologies, we look at a simple model of N DCs of capacity P, in terms of physical DCI ports. A DCI port here reflects an electrical switch port of some bandwidth that is dedicated to the DCI network at a particular DC. We further assume that the N DCs are organized in G groups. To simplify, we consider all G groups to be balanced in size, and that all DCs in a group are interconnected using a group-local hub. Further, we assume all-pairs direct connectivity across groups. This simple model allows us to move gradually from centralized towards distributed topologies: $G = N$ represents a fully distributed topology with all DC-DC pairs directly connected, while $G = 1$ represents the centralized topology.

For $G = 1$ and a capacity of P ports per DC, the total number of ports required in the topology is equal to $2 \cdot N \cdot P$, *i.e.*, double the total capacity of all DCs, as $N \cdot P$ ports are required at the hub. For $G > 1$, the number of ports required to connect DCs within a group is ${}^{2 \cdot P \cdot N}/_{G}$. Each group hub needs to support ${}^{P \cdot N}/_{G}$ capacity downstream and ${}^{(G-1) \cdot N}/_{G} \cdot P$ ports upstream to other groups, for a total of $N \cdot P$ ports. This means that the capacity of the hub is essentially independent of the size of the group ${}^{N}/_{G}$; each group hub needs to support the same capacity irrespective of how distributed or centralized the topology is. In total, the topology requires $(G + 1) \cdot N \cdot P$ ports.

This is shown in Fig. 7 using an example region of 16 DCs. The figure further breaks down cost contributions from different hardware components: (a) electrical switch ports, and (b) DCI transceivers, based on realistic cloud-provider prices where a transceiver costs

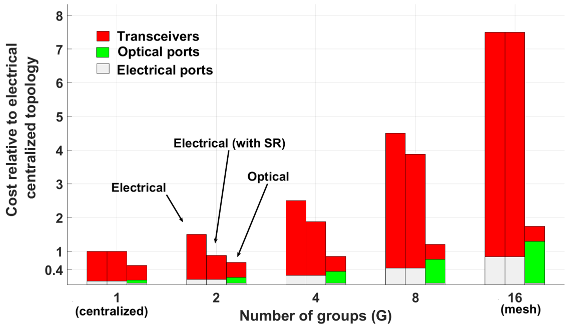

Fig. 7: *Relative port cost breakdown for electrical and optical networks as topologies become more distributed. Total cost is estimated based only on per-port cost. "Electrical with SR" uses cheaper short-reach transceivers for connecting DCs within a group.*

roughly 10× an electrical port. The figure shows that in such a region, the relative cost of supporting a fully meshed distributed topology is roughly 7× the cost of the centralized topology. The semi-distributed topologies are also more expensive than a centralized one, even when we account for group-internal connectivity using cheaper short-reach optical transceivers, which is optimistic, as the required hub-DC distances to be able to use such transceivers (≤2 km) will not always be achievable. The results highlight that the biggest contributor to the cost are the optical transceivers. The third column shows what the cost of an optical DCI network would be, assuming we could replace transceivers with optical reconfigurable ports, the approach we advocate.

2.5 Summary

Our analysis reveals clear pros and cons for each approach: the distributed approach has clear advantages in latency and siting flexibility, but entails greater complexity and cost. Thus, to make the distributed part of the design spectrum more accessible by lowering these cost and complexity barriers, we propose Iris.

3 IRIS GOAL AND CONSTRAINTS

Iris's improvements stem from reducing the large number of transceivers used in electrical DCIs, as well as the total number of switching ports in the network. This results in lowering the bar for realizing distributed topologies by making their cost comparable to centralized ones, and further reducing their complexity in configuration and management by reducing in-network ports by an order of magnitude (§6.1). The core premise of Iris is simple: between its source and destination DCs, traffic never leaves the optical domain. Practically realizing this design philosophy, however, requires addressing a large set of constraints: *operational constraints* that derive from application requirements, and *technology constraints* imposed by the physical characteristics of the optical equipment used.

3.1 Operational constraints

OC1. **Latency SLA** — The maximum roundtrip DC-to-DC latency is bounded by a tight SLA. For existing SLAs, this translates to a maximum DC-DC fiber distance of 120 km (§2.2).

OC2. **Any traffic matrix** — Each DC's aggregate network capacity is known based on its size and other business factors. The DCI should accommodate any traffic demands that are not bounded by DC capacity, as in the hose model [14]. DCI links are typically symmetric, so we do not distinguish between ingress and egress capacities, assuming symmetric demands without loss of generality.

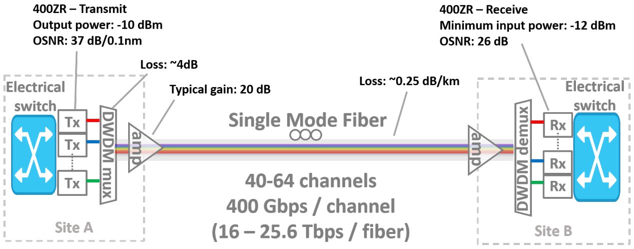

Fig. 8: *Typical DCI optical link, components and 400ZR specifications.*

OC3. **Shortest path** — Traffic between DCs must always use the shortest available physical path. We thus discuss Iris's most complex use case: distributed networks that minimize latency (§2.1). By removing this constraint, simpler designs are easy to build using the same methodology, with appropriate corresponding simplifications.

OC4. **Failure resilience** — Based on reliability goals, an operator specifies a number of fiber cuts that must be tolerated, *i.e.,* for any number of cuts up to the specified tolerance, *OC1-OC3* should continue to hold. A fiber cut here means a *fiber duct destruction*, *i.e.,* all capacity for all fibers traversing the duct is lost. For our description, we use a tolerance of 2 cuts, in line with operational practice, but nothing in our approach depends on the precise value.

3.2 Technology-rooted constraints

Today's electrically-switched DCI networks comprise point-to-point static optical links between any two sites (DC-DC or DC-Hub). Fig. 8 shows a typical example. We consider transceivers that plug directly into DC electrical switches (Tx,Rx in Fig. 8) [20], and in particular, the 400ZR transceivers (400 Gbps, 16 QAM) [39], which have been standardized for DCI and are expected to be deployed soon across most providers. Dense Wavelength Division Multiplexing (DWDM) is used to combine 40 − 64 optical signals at different wavelengths (colors), one per transceiver, covering the C-band.

On the receive side, optical signals need to respect the minimum optical power and optical signal-to-noise ratio (OSNR) thresholds given by transceiver specifications. The received *optical power* is dictated by the sending transceiver's transmit output power minus losses due to optical components in the link, such as the fiber and mux/demux elements. *OSNR* is affected by noise introduced by elements like amplifiers. Fig. 8 includes the details of expected 400ZR OSNR and power values, as well as typical losses for elements on point-to-point links. Any DCI architecture would need to respect these thresholds, which, in turn, lead to the following constraints.

TC1. **Optical link distance** — Optical amplifiers on both side of the link compensate for power losses, and have a typical gain of 20 dB. Thus, assuming a typical fiber loss of 0.25 dB/km [20], the receiving amplifier (Fig. 8) can compensate loss for a maximum DC-DC link distance of 80 km, absent in-line amplification.

TC2. **End-to-end amplifier count** — Additional in-line amplifiers between sites can increase reach (*e.g.,* up to 120 km) and/or allow for extra on-path optical components to enable reconfigurability.

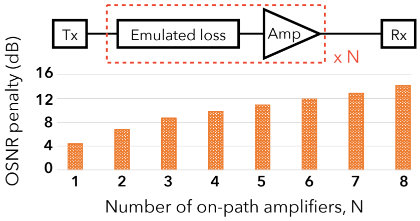

Fig. 9: OSNR *penalty vs. amplifiers. The experimental setup (top) uses attenuators between amplifiers to match the amplifiers' gain.*

Unfortunately, amplifiers add noise, degrading the amplified signal's OSNR [16]. To quantify this, we measure the *OSNR* of transmitted signals at the output of multiple amplifiers in our testbed (Fig. 9). The first amplifier adds an *OSNR* penalty to the unamplified signal equal to the amplifier's specified noise figure (~4.5 dB). Beyond this, each doubling of the number of amplifiers on the line degrades *OSNR* by ~3 dB. The observed penalty agrees with theoretical models that examine the impact of cascaded amplifiers on OSNR [32]. With 400ZR, between sites, we can tolerate up to 11 dB *OSNR* penalty (Fig. 8). Allowing an additional couple of dBs for various transmission impairments and amplifier gain ripples, this translates to an amplifier budget of 9 dB, or a maximum amplifier-count of 3 end-to-end (Fig. 9). Thus, at most one extra in-line amplifier can be added in any reconfigurable physical layer design with maximum distance of 120 km.

TC3. **Power management** — When the optical network is (occasionally) reconfigured, the fiber spans part of a path can change. In turn, some optical amplifiers see their input power change, *e.g.,* if the input fiber span is now shorter, their input signal sees lower loss, and requires less amplification. Absent an adjustment in the amplifier's gain or proper management of the input power, the signal *OSNR* would be degraded. Unfortunately, adjusting the gain of amplifiers region-wide in a synchronized fashion would be severely limiting, as it can take several seconds for optical signals to stabilize [2]. Thus, appropriate management of input power to the amplifiers is mandatory in any architecture where the same amplifier compensates losses across different paths over time.

TC4. **Number of optical reconfiguration elements** — Components that allow optical reconfiguration also cause optical power loss, the degree of which depends on the components used. Reconfiguration can be achieved at two granularities: (a) at the fiber level, with all traffic from one fiber shifted to another, using optical space switches (OSSes) with up to a few hundred ports [9, 40]; and (b) at the wavelength level, shifting individual wavelengths across fibers, using a Wavelength Selective Switch (WSS) with at most a few tens of inputs. Large-scale wavelength-level switching requires combining individual components (de/mux and OSSes) into what is called an Optical Cross-Connect (OXC) [10, 37].

For a maximum distance of 120 km with one extra amplifier (*i.e.,* 40 dB total budget), after accounting for a fiber loss of 0.25 dB/km, we have 10 dB available for optical reconfiguration elements. OXCs and OSSes have typical losses of 9 dB and 1.5 dB, respectively. This translates to at most one OXC or 6 OSSes end-to-end.

3.3 Component costs and operational costs

Besides the above constraints, cost is also crucial in DCI design, especially given that major providers have tens of regions. While our analysis in §6 uses real prices (amortized such that equipment costs and fiber leases can be jointly accounted), we can only disclose coarse component costs in relative terms.

Transceivers are the most crucial cost factor, given their large volume: each electrical port needs one. A DC-DC connection that carries λ wavelengths requires $2 \cdot \lambda$ transceivers. The transceivers used in our analysis are DWDM switch-pluggable transceivers like the 400ZR, or today's 100G equivalent [20] designed to cover DCI distances, *e.g.,* up to 120 km. Prices for such DCI transceivers are not public, with only volume-based prices offered to cloud providers. As a coarse reference point, vendors are estimating such DCI transceivers at roughly \$10/Gbps [7]; this implies an approximate cost of ~\$1,300 per year after accounting for 3-year amortization. Our analysis in §6 uses the true volume-based price charged to cloud providers. Note that while traditional long-haul coherent transceivers designed to cover thousands of kilometers may be used in DCI, their cost is several times the one of custom-designed DCI transceivers [7], and thus are not considered further in our analysis.

Fiber in regional networks is typically inexpensive because already laid out fiber ducts are abundant in metro areas. The caveat is that fiber cannot be arbitrarily added to minimize distances (§2). Fiber-pairs are priced per span, independent of distance, with lease price varying significantly across regions. A ballpark figure is ~\$3,600 per year [1], equivalent to 3× the amortized cost of the above mentioned transceivers. Recall that a single fiber carries data from 40 – 64 transceivers.

OSS ports cost an order of magnitude less than one transceiver, *e.g.,* 100-200 dollars per (unidirectional) port [11].

OXC ports are slightly more expensive than OSS ports, due to the need for de/muxes, but still much cheaper than transceivers.

Amplifiers are equivalent in cost to a few transceivers. However, since each of them amplifies all the wavelengths in a given fiber, their contribution to the cost is not substantial.

Operational costs. While our quantitative analysis only accounts for component costs, we briefly comment on operational costs of two types: (a) network management; and (b) power and equipment space. Precisely appraising management costs is inherently hard, especially for novel, non-operational architectures like Iris. Indeed, we expect that there will be some initial ramp-up cost for developing tooling to manage Iris, but once done, steady-state management cost should be similar to or lower than today's designs across the entire spectrum of centralized to distributed DCI networks, on account of Iris's reduction in the number of ports to be managed. Costs like power and space, on the other hand, are expected to be significantly lower with Iris: most of the optical devices used are passive, requiring orders of magnitude less power than an electrical fabric. In terms of space, optical switches with hundreds of ports are just a few rack-units in size [40], in comparison to the rack-size electrical switches needed at this scale.

3.4 Cost comparison: a motivating example

To motivate Iris's all-optical design strategy, we use a small, toy DCI design example, with a fixed topology implemented both ways, *i.e.,* using either a traditional electrical approach, or Iris's all-optical approach. The topology used is the same semi-distributed one in Fig. 1(e), but is redrawn with labeling in Fig. 10. DC1 and DC2

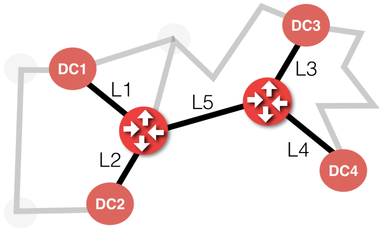

Fig. 10: *An example fiber map with data center placement. Assuming shortest-paths the dark highlighted links are only used.*

connect to one hub and DC3 and DC4 to another. Each of the 4 DCs has a capacity of 160 Tbps. With 400 Gbps for each of 40 wavelengths, this translates to $f = 10$ fiber-pairs.

For the electrical design, L1-L4 each carry 10 fiber-pairs, so each DC's full capacity is connected to its hub. L5 carries 20 fiber-pairs, such that the network is non-blocking. The total number of fiber-pairs is thus $F_E = 60$, and the number of transceivers is $T_E = 2 \cdot F_E \cdot \lambda = 4800$, as each fiber terminates in a transceiver.

With Iris, transceivers are needed only at the DCs, *i.e.,* $T_O = 4 \cdot 10 \cdot \lambda = 1600$ transceivers. However, for optical switching in the network, Iris uses additional fiber and OSS ports. §4 details how this is done, but in this specific example, L1-L4 need 3 additional fiber-pairs, and L5 needs 6 additional fiber-pairs. The total number of fiber-pairs thus increases to $F_O = 78$. Each fiber-pair terminates at OSS ports at both ends, so 312 OSS ports are needed in total.

Using the prices described in §3.3, the electrical design costs 2.7× more than the optical one.[4] This difference is rooted in the fact that transceivers are the overwhelming expense: an OSS port costs an order of magnitude less than a transceiver, and while one fiber's cost is a few times that of a transceiver, the absolute number of fibers needed is nearly two orders of magnitude smaller. Thus, using some extra fiber and OSS ports to reduce the number of transceivers is a very profitable trade.

Iris's advantage is greater for larger regions, and for more distributed topologies. Thus, Iris enables cost-effective networking for larger-scale regions with the favorable characteristics of distributed topologies. Our analysis (§6) using real fiber maps and cloud-provider component costs shows that Iris would be 7× cheaper in the median than an electrical switching implementation.

4 IRIS NETWORK PLANNING

As discussed in §2, planning a regional DCI network entails using the region's fiber map and data center locations and capacities, to decide on the topology, fiber capacity of each connection, and the use of switching to implement the topology and capacity decisions.

We first jointly address topology and capacity, as these derive primarily from operational constraints, and are *largely* the same regardless of switching. For optical switching, meeting the technological constraints sometimes requires revisiting topology and capacity decisions; we discuss such cases separately in §4.3.

4.1 Topology & capacity provisioning

We use a natural graph abstraction: DCs and huts are nodes of graph G, and the available fiber forms edges between them. Fiber edges longer than 80 km can be excluded right away: regardless of electrical / optical switching, longer *point-to-point* connections are not possible (*TC1*). Our task then is to decide which subset of edges are used, and at what capacity. Algorithm 1 achieves this by computing which links lie on shortest paths (*OC1* and *OC3*) in *any* failure scenario (*OC4*) by exhaustively enumerating the latter.

Algorithm 1: Topology & capacity planning.

```
$G_{init} \leftarrow$ fiber map
$\forall$ edge e $\in$ G $_{init}$: capacity$_e \leftarrow 0$
foreach failure scenario do
    G $\leftarrow$ G$_{init}$ \ failed fiber ducts
    SP $\leftarrow$ {shortest paths in G $\forall$ DC pairs}
    foreach edge e $\in$ G do
        sp$_e$ $\leftarrow$ {sp $\in$ SP | sp uses e}
        G$_e$ $\leftarrow$ construct flow graph for e using sp$_e$
        capacity$_e$ $\leftarrow$max(capacity$_e$, max flow of G$_e$)
```

Determining which edges are used is trivial, but assigning their capacities is not. Since each DC-pair uses only its (typically unique) shortest path, one may naively assume that to support the hose traffic model (*OC2*), the capacity of each edge is simply the sum of demands for DC-pairs traversing it, where a DC-pair's demand is the minimum of the two DCs' capacities. However, this leads to needless over-provisioning: *e.g.,* a DC, say *A*, may be part of multiple DC-pairs, say *A-B* and *A-C*, traversing an edge over shortest paths; this naive approach would double-count *A*'s capacity for this edge. A precise solution to capacity provisioning requires a max-flow computation across an appropriately constructed "flow graph". We adapt this from prior work [29], and thus omit the details.

Algorithm 1 yields not only edge capacities, but also which fiber huts are used: if a hut has no edges of non-zero capacity, it is unused. Thus, it fully determines the network's topology and capacity. Note that if shortest paths are unique, as is typically true across real fiber maps, Algorithm 1 yields the unique (and hence optimal) solution for topology and capacity planning: only one set of chosen huts and edges meets the constraint of achieving shortest paths under all failure scenarios (*OC3* and *OC4*). For settings with multiple shortest paths, or when the shortest path constraint is relaxed, this is only a heuristic that still meets all constraints, but does not necessarily provision the minimal infrastructure.

We next discuss three granularities for switching, the last decision needed to fully describe DCI planning, drawing out the reasoning for Iris's choice of optical fiber switching.

4.2 Electrical packet-switched network

Given the topology and capacity provisioning, an electrical packet-switched (EPS) fabric is simple to build: just deploy enough switching capacity at the DCs and huts using standard Clos networking techniques. As noted in §2.4, the key impairment of this approach is its cost: it requires a large number of electrical ports and transceivers, directly proportional to the number of wavelengths per fiber terminated at each fiber hut.

4.3 Optical fiber-switched network

To avoid the explosion of electrical ports, Iris uses an all-optical network core, *i.e.,* data does not leave the optical domain except at

[4] As other costs are much smaller, accounting for only fiber and transceivers arrives at nearly the same number, *i.e.,* $(1300T_E + 3600F_E)/(1300T_O + 3600F_O) = 2.73$.

end-points. As discussed earlier in §2.4, this approach can provide the substantial benefits of a distributed DCI network at cost similar to a centralized one. At each hut, only optical space switches (§3.2) are used to direct all wavelengths carried in a fiber from one port to another, thus reducing port requirements to *one per fiber*. This effectively sets up DC-DC optical circuits through the network. However, this requires deploying appropriate optical equipment at intermediate DCs and huts to address three problems:

- Coarse-grained fiber switching needs more network capacity than computed above in §4.1.
- Since DC-DC data streams travel end-end as optical signals, we must deploy amplification as necessary for the now longer distances (*TC1* and *TC2*).
- We must limit the number of optical switches on each end-to-end path (*TC4*).

The latter two problems are self-evident, based on our earlier description of technology constraints in §3.2, but the first is a significant challenge of fiber switching, and requires some explanation. The need for additional capacity stems from the coarse granularity: while for EPS fabrics, integer number of wavelengths (as we assume DC capacities are specified in) can be flexibly switched, fiber-switching requires rounding to the fiber-level. Consider a DC that has a capacity equivalent to z fibers, and sends x and y to two DCs, such that $x + y = z$, but y comprises only a fraction of one fiber's capacity, such that $\lceil x \rceil = z$. Switching at fiber granularity implies that we now need $z + 1$ capacity from the DC. Worst-case scenarios, which we want to tackle per *OC2*, necessitate that for each DC-pair, one additional fiber is necessary to address this issue, increasing fiber cost by $n \cdot (n - 1)$ fibers for a region with n DCs. Note though, that no additional transceivers are needed: transceivers at the DCs can still be multiplexed across the fibers as necessary. Overall, we find that this is a highly favorable trade-off.

For the second problem, amplification, we use a heuristic to ensure that no umamplified segment exceeds our distance constraint (*TC1*), and each path has at most one amplifier (*TC2*). Our heuristic also tries to greedily reduce the number of amplifiers. The intuition is to examine each failure scenario, identify paths that need amplification, score each potential amplifier location in terms of how many paths it would meet constraints for, add amplifiers as needed to the highest-scoring location, and iterate until the constraints are met. For interested readers, the details are in Appendix A.

For the third problem, limiting each path's switch-count (*TC4*), we use a similar greedy approach. For each path with >6 switching points, we add "cut-through links" that replace one or more switch-points for the path with an uninterrupted fiber between the endpoints of the replaced segment, with adequate capacity for that path. We again attempt to minimize such cut-throughs, by finding ones that resolve constraints for multiple paths.

Put together, the above solutions for capacity provisioning, amplification, and cut-through placement, meet all our constraints. Our heuristics use exhaustive enumeration across failure scenarios, and several iterations by making reassessments after placing each amplifier or cut-through, but still execute within a few minutes for even large region sizes with 20 DCs. Given that this process only executes once for network provisioning, this is sufficiently fast, and as we show later, provides significant cost reduction (§6).

4.4 Wavelength-switched network

While Iris's fiber switched network is many times cheaper than an EPS fabric, one may wonder if the n^2 fiber overhead of coarse-grained fiber switching can be avoided to further reduce expense, using finer-grained *wavelength* switching. Such a design would demultiplex each fiber's wavelengths at the switching points, and switch them into appropriate output combinations, instead of just switching at fiber level. Surprisingly, we find that this design is inferior with the additional components needed for wavelength switching resulting in a pricier design than the n^2 additional fibers for fiber switching (please see details in Appendix B).

While naively switching at the wavelength level is neither feasible nor cost-effective, we also explored a more judicious "hybrid" approach (Appendix B). This approach uses fiber switching for most of the traffic, relying on wavelength switching only to address fractional demands. While indeed it can provide cost savings compared to a fiber-switched-only network in some scenarios, these savings are small (see §6). It also adds substantial complexity, which would deter deployment. We thus conclude that fiber switching is the most viable switching architecture for regional DCI networks.

5 IRIS IMPLEMENTATION

We next discuss Iris's implementation: how different components connect to each other, and how they are managed.

5.1 Putting together Iris components

Fig. 11 shows a full-system view with the details for 2 of the N DCs drawn out, showing the send/receive parts respectively.

Sending from DC1: DC1's internal Clos fabric sends outgoing traffic to its tier-2 (T2 or core) switches. Internal routing to T2 switches can be achieved using standard mechanisms like ECMP and anycast [24], such that traffic for each external destination arrives at the right T2(s) in a load balanced fashion. Each transceiver at each T2 converts this traffic to a wavelength; Fig. 11 shows 3 transceivers / wavelengths for each of the 2 T2s. These wavelengths are mux-ed into fibers (via OSS1), which are then switched towards destination DCs (using OSS2). OSS1's function is allowing any T2-transceiver to be fed into any fiber – thus instead of directly mux-ing wavelengths from T2s, they are first fed into OSS1, whose outputs are then mux-ed. Iris uses tunable transceivers at T2s, such that colors can be assigned to each transceiver to make it trivial to pack them into outgoing fibers. After each fiber is packed, it goes through amplification. OSS2 acts like any other switching point; it switches both: (a) DC1's outgoing capacity of C fibers, plus $N - 1$ fibers to address the "fractional" capacity (§4.3); and (b) any fibers this DC transits for other DC-DC traffic (bottom-left in Fig. 11).

Intermediate switching: Fiber huts and on-path DCs performing intermediate switching use only an OSS and amplifiers (if placed, based on the heuristic in §4.3). As noted earlier, amplifiers can be used by any paths passing through. Implementing this in a configurable manner requires using amplifiers in a "loopback" fashion, whereby their input *and* output are both attached to the OSS, such that an arbitrary fiber can be directed through the OSS to the amplifier, fed back into the OSS post-amplification, and then switched to an arbitrary OSS output. (See hut H1 in Fig. 11.)

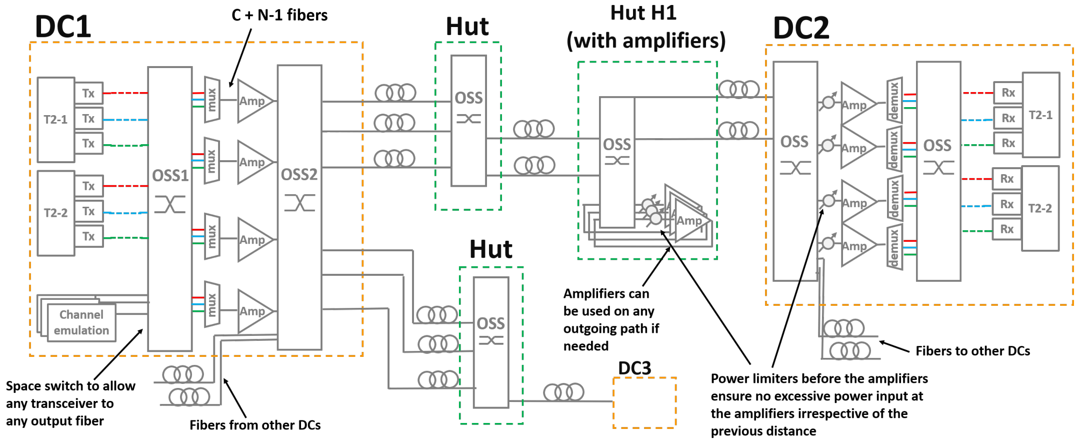
DC1
C + N-1 fibers
Hut
Hut H1
(with amplifiers)
DC2
T2-1
T2-2
Tx
OSS1
OSS2
mux
Amp
Channel emulation
Space switch to allow any transceiver to any output fiber
Fibers from other DCs
OSS
Hut
DC3
Amplifiers can be used on any outgoing path if needed
Power limiters before the amplifiers ensure no excessive power input at the amplifiers irrespective of the previous distance
demux
Rx
Fibers to other DCs

Fig. 11: *Iris puts together available commodity components in a manner that respects all the technology constraints, while still meeting our operational goals.*

Receiving at DC2: The receive side largely mirrors the send. Besides passing the traffic destined to this DC to demux-es and finally transceivers (after amplification), fibers destined to other DCs can be switched towards them by the OSS.

Amplifier power management: The above implementation, if based on appropriate provisioning (§4), suffices to meet all our design constraints except one: amplifier power management (*TC3*, §3.2). We use two methods to ensure that amplifier gains do not need careful management. First, we transmit the full C-band spectrum per fiber, *i.e.,* all wavelengths, even if only some carry data. Doing this using transceivers would be expensive; instead we use amplified spontaneous emission (ASE) noise to fill only the unused spectrum that is then combined through muxes with the “live” channels (“Channel emulation” in Fig. 11). This ensures uniform gain profiles across fiber segments of equal length regardless of their “live” channels. Second, we operate all amplifiers at a fixed gain irrespective of the fiber length that they compensate. To ensure that no excessive power reaches the following component in the physical link (*i.e.,* the next amplifier), we use a power limiter before each optical amplifier to bound its input optical power. These are one-time design decisions rather than continual online management.

5.2 Configuring Iris components

A centralized controller gathers DC-DC traffic demands, and configures the network components appropriately. The small number of sites with only tens of fibers per site, coupled with relatively infrequent reconfigurations, simplify the control problem greatly, especially in comparison to systems using optical reconfiguration in other settings like WAN optimization [28] or data centers [23]. All our key design decisions are further geared towards reducing complexity to make reconfiguration a straightforward process:

- fiber switching based on only simple capacity needs
- basic wavelength management separately in each DC
- no online power management for amplifiers or any other optical component

When the controller decides that a reconfiguration is needed, it first drains traffic from paths that need to be torn down. It then reconfigures the OSSes network-wide to enable new paths. The configuration of transceiver wavelengths, and channel emulation is done independently at each DC.

Reconfiguration time: OSSes are the bottleneck here. While tunable transceivers can switch wavelengths in under 1 ms [22, 42], and unused amplifiers can provide gain in under 2 ms [12, 31], the state of the art for OSSes is ~20 ms [33]. In the future, we expect sub-ms switching for OSSes [25].

Regional IP routing and WAN transit remain the same as today. The higher tier of each DC has full regional route visibility, and a few DCs transit WAN traffic. (Note: WAN traffic is a small fraction of regional traffic.)

6 EVALUATION

Iris, by design, meets the constraints specified in §3. In §2, we further demonstrated its latency and flexibility improvements over the centralized approach in real settings. We thus evaluate three aspects that bridge any potential gap between the system’s abstract design and its practical realization: (a) cost; (b) physical layer feasibility; and (c) impact of circuit switching under fluctuating traffic.

6.1 Cost analysis on real fiber maps

We use 10 real region fiber maps with a randomized placement of $n \in \{5, 10, 15, 20\}$ DCs across each map: the first DC is placed uniformly at random in the service area, and each successive DC is placed randomly (in the more restricted service area given reach from already placed DCs) with probability of a candidate location being inversely proportional to its distance from the nearest already placed DC. In line with typical values [20], we vary DC capacities in terms of number of fibers, $f \in \{8, 16, 32\}$, and λ transceivers per fiber, with $\lambda \in \{40, 64\}$. For each of the 240 combinations of these inputs, we calculate the cost of Iris, and equivalent EPS (§4.2) and hybrid networks (§4.4). We account for all network components with appropriate price amortization (§3.3), and the number of ports per hut can be accommodated with today’s OSSes.

Fig. 12(a) shows the resulting cost comparison in relative terms: in 80% of the examined scenarios, the EPS network is ≥ 5× more expensive than the Iris and hybrid networks. Further the virtually identical costs of the latter two justify Iris’s choice of simpler fiber switching. This analysis includes the transceivers within the DCs, which are fixed across the design space. A sharper contrast between the design choices is revealed when we exclude this fixed cost, and only evaluate in-network components. Iris’s cost is then 10× lower for 80% of the scenarios (the “in-network” line in Fig. 12(a)).

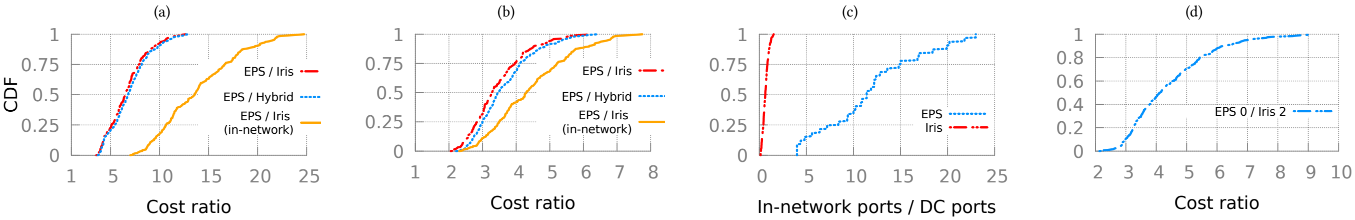

Fig. 12: *Iris is substantially cheaper: (a) Relative cost of Iris, EPS, and hybrid networks across all 240 scenarios. (b) Same as (a) but with DCI transceiver cost assumed (unrealistically optimistically) equal to SR transceivers. (c) EPS uses many more in-network ports, as shown by the ratio of in-network to DC ports across designs. (d) Relative cost of an EPS supporting no failures vs. Iris, which handles up to 2 failures.*

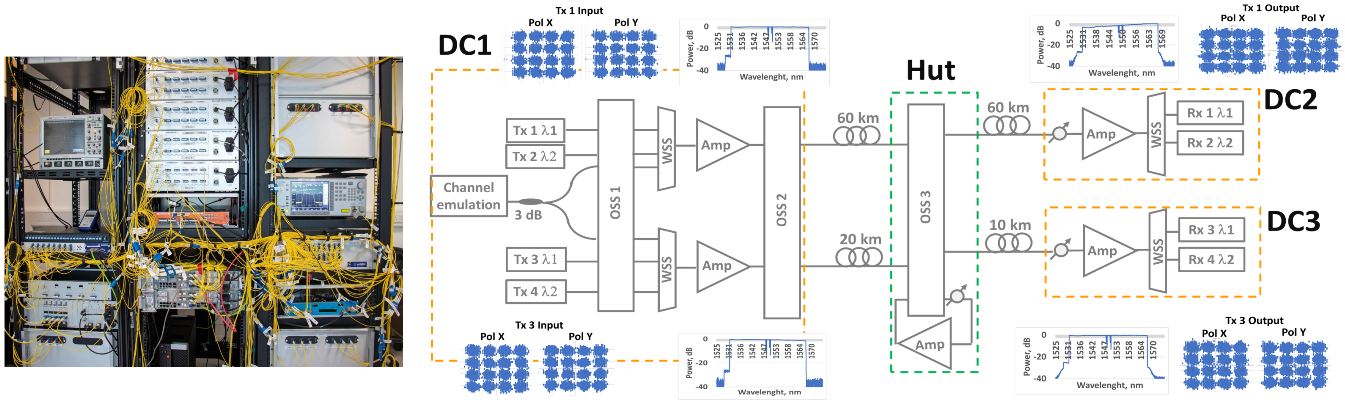

Fig. 13: *(a) A small Iris testbed with all the optical components. (b) Fiber switching experimental set up. Insets: Examples of fully loaded spectra and constellation diagrams following the expected shape of dual polarization 16-QAM signals as measured at different points in the system (details in Appendix C).*

We also emphasize that these cost differences are not ephemeral. The involved components are all commoditized and high-volume, so we believe this analysis to be fair. Nevertheless, to emphasize the disparity between Iris and alternatives, we also examine the potential impact, were DCI transceiver prices to drop (unrealistically) to those of short reach transceivers (presently used for sub-2 km). Fig. 12(b) shows that Iris would still have a substantial cost advantage. The reason is the number of ports (optical or electrical) needed in different systems: as Fig. 12(c) shows, an EPS fabric requires many times more ports in-network than Iris.

Finally, per Fig. 12(d), Iris, which guarantees capacity under up to two failures, is cheaper (by >2× across *all* scenarios) than even an EPS with no guarantees under failures.

6.2 Physical layer feasibility

Fig. 13(a) shows our testbed implementation of Iris, which uses all the components described in §5.1: Multiple fiber spools 5-50 km in length, that allow us to model any regional distance at a granularity of 5 km; Erbium-doped fiber amplifiers from Ciena; OSSes from Polatis, which also provide per-port power-limiting functionality, arranged to model DC OSSes as well as 2 fiber huts; WSSes from Finisar used to mux/demux wavelengths; Channel emulator from BKtel to fill unused spectrum; 4 Acacia tunable transceivers (2xAC400, 2xAC200) that can support varying baud-rates, modulation formats, channel grid spacing, etc. These are not switch-pluggable but controlled via evaluation boards, allowing us fine-grained config to emulate the 400ZR specification.

We have also implemented a software controller (in Python, ≈ 9K LoC) to control the optical devices through a multitude of interfaces (serial port, HTTPS, and NetConf/REST). Our controller implements APIs for all operations of Iris's optical layer, namely channel add/drop, reconfiguration of space switches, checking that the devices are in expected state, etc. Our present testbed evaluation focuses on physical layer feasibility, which then guides our large-scale simulations. Unfortunately, given that our transceivers are controlled through evaluation boards instead of real switches, we cannot run a full control-to-bits evaluation at this time.

Our experimental set-up is shown in Fig. 13(b) and matches the description in §5. We emulate 3 DCs, one sending traffic to the other two, over two distinct paths of 4 fiber spans in total. We switch the two paths at an intermediate hut. Our high-level description below is targeted at most networking readers, with details for optics experts in Appendix C.

We generate 4 optical signals at two different wavelengths together with ASE noise to fill the C-band spectrum in DC1. This traffic is fed into 2 fibers, each carrying the 2 different wavelengths, muxed through the OSS/WSS. The two fiber spans of 20 and 60 km from DC1 terminate at the hut. The following fiber spans from the hut are 60 km (to DC2) and 10 km (DC3). The experiment periodically swaps which spans are connected, producing two combinations A(60-60, 20-10) and B(20-60, 60-10). For both configurations, the shorter distances do not need amplification, while the hut amplifier is used for the two longer ones. Thus, over time, both DC-DC paths interchangeably utilize the hut amplifier. This setup tests each piece of our architecture.

Power management. We measure the full spectrum at uniform power at input/output DCs. Our amplifiers work as desired, not causing any power variations after transmissions of varying lengths with occasional in-line amplification.

BER and reconfiguration. Fig. 14 shows the maximum bit-error rate (BER) before FEC at two of the receivers as we reconfigure every minute. Results collected over multiple day-long runs are

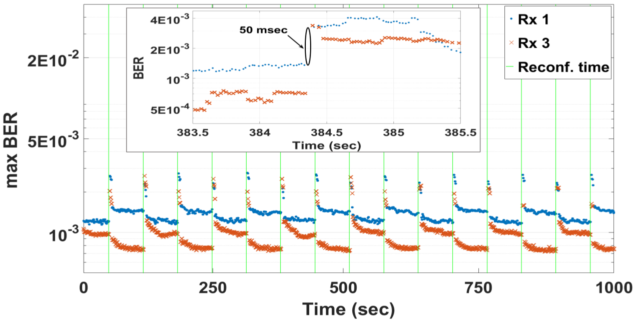

Fig. 14: *BER over time while reconfiguration occurs (inset).*

similar. It takes 50 ms to recover the signal after reconfiguration. The received pre-FEC BERs are well below the soft decision FEC threshold (2x10^{-2}), translating in final BERs below 10^{-15}; this is similar to equivalent *static* optical links. As discussed in §5, no live traffic will be carried by paths during reconfiguration. We performed similar experiments involving reconfiguration across two independent huts with similarly consistent BERs and maximum switching time of 70 ms.

6.3 Impact of circuit transience

Iris uses reconfiguration to respond to failures and (slow) changes in DC-DC traffic. To study how this may impact application performance, we perform region-scale flow-based simulations in a custom simulator. The topologies examined reflect the DC connectivity and scale of the regions analyzed in §6.1. Note that we drain links before reconfiguration, so transport loss is not a concern. Our experiments thus focus on the impact of capacity reduction during reconfiguration where a fiber switch takes 70 ms (§6.2).

Based on experience, we use heavy-tailed traffic between DCs, with a few pairs exchanging most of the traffic; unbounded changes in traffic patterns occur when, *e.g.,* a low-traffic DC-DC pair becomes a high-traffic one. Otherwise, we bound the changes to a maximum % value. We study a broad swath of operating conditions: (a) network utilization in {10%, 40%, 70%}; (b) reconfig frequency of once every 1-30 sec; (c) changes of {1%, 10%, 50%, 100%, unbounded} in DC-DC traffic; and (d) several distributions for flow sizes [4, 41]. Note that these tests include extremes well beyond those we expect to encounter, *e.g.,* DC-DC aggregate traffic changing by up to 50% every 1 sec. Extreme DC-DC traffic volatility at the granularity of seconds would not be expected. Similarly, we chose to examine flow distributions that reflect intra-DC workloads dominated by short flows. This serves as a stress-test for a circuit-switched network — such flows would be the ones most affected by link reconfiguration, as longer flows are throughput and not latency sensitive. Finally, we assume that provisioning is sufficient to handle the traffic before and after the reconfiguration – this is plausible given the predictable nature of DC-DC traffic and substantial capacity over-provisioning.

We compare the Flow Completion Times (FCTs) for Iris to an EPS fabric baseline. Due to space limitations we summarize the main findings here and provide full details in Appendix C. Overall, even at high utilization levels (70%) and large traffic changes (> 50%), the effect is negligible, especially for reconfiguration intervals of 10 sec or above. For shorter intervals, there is a maximum slowdown of 2% across all flows at the 99^{th} percentile with Iris compared to EPS. This is true across all workloads examined.

These results are largely expected: the probability of a short flow ($<50KB$) being affected is small given the reconfiguration interval is much larger than short flow completion times; meanwhile, large flows see only a brief, negligible drop in throughput.

7 RELATED WORK

DCI design is an increasingly important problem for cloud providers. However, to the best of our knowledge, no prior work has exposed the tradeoffs involved, or explored the design space systematically. Nevertheless, we attempt to place our work in a broader context.

Cloud WAN networks, like Iris, interconnect small numbers of sites. However, long-distance WAN links are much more expensive than regional fiber. This results in WAN proposals like OWAN [28], SWAN [26], and B4 [27], optimizing towards maximum utilization of WAN links. Iris's design philosophy is the opposite: exploit the cheap, abundant fiber of metro areas to design a simple and cost-effective network. Further, while wavelength switching, as is often used in metro optical networks [35], would improve spectral efficiency, in DCIs we find this to be unnecessarily complex.

Intra-DC networks using optics are also well-studied. Early efforts in this direction used OSSes [18, 43], while newer work is attempting to tackle frequently changing intra-DC traffic at microsecond scale [5, 6]. Iris, only needs to address slow-changing aggregate DC-DC traffic, but additionally tackle power and signal quality constraints stemming from the longer link distances. These constraints lead to completely different design choices.

Lastly, we note that DCIs are a hot industry topic [38], especially at the lowest layers, *e.g.,* customizing optical components [19, 34], and defining DCI standards [39]. This work fits within current ways of interconnecting these components, like the centralized and distributed models and their implementations discussed in §2.

8 CONCLUSION

Motivated by the growing popularity of multi-data center regions, we study architectures for regional data-center networks, and highlight their trade-offs. We find that while distributed networks offer attractive latency and siting flexibility benefits, their implementation with today's de-facto packet-switching design also engenders greater cost and complexity. To simplify DCI network design and lower the barriers for distributed networks, Iris introduces an all-optical network core. With Iris, data travels between DCs entirely in the optical domain, thus greatly reducing the number of in-network ports. Iris's simple fiber-level circuit switching only requires a minimal control plane, and off-the-shelf optical equipment, as our testbed implementation demonstrates.

ACKNOWLEDGEMENTS

The Iris project has benefited from the work of several interns over the past years to whom the authors are grateful. These include Theofilos Petsios, Marti Sales Llopis, George Mourgias-Alexandris and Shengxiang Zhu. The authors would like to further thank the Optics for the Cloud team at Microsoft Research Cambridge [36] for their continuous support, and in particular Istvan Haller, Kai Shi, Benn Thomsen, Hitesh Ballani, and Paolo Costa, as well as Adrian Power from Azure Networking for his guidance on regional reliability models. Finally, we are thankful to our shepherd Radhika Niranjan Mysore and the anonymous reviewers for their feedback.

REFERENCES

[1] ACG Research. *Migration to 100G campus connectivity.* https://www.inphi.com/pdfs/ACG-100G-Campus-Connectivity-Analysis.pdf.

[2] A. S. Ahsan, C. Browning, M. S. Wang, K. Bergman, D. C. Kilper, and L. P. Barry. "Excursion-free dynamic wavelength switching in amplified optical networks". In: *IEEE/OSA Journal of Optical Communications and Networking* 7.9 (2015), pp. 898–905. DOI: 10.1364/JOCN.7.000898.

[3] Shahbaz Alam, Pawan Agnihotri, and Greg Dumont. *AWS re:Invent. Enterprise fundamentals: design your account and VPC architecture for enterprise operating models.* https://www.slideshare.net/AmazonWebServices/aws-reinvent-2016-enterprise-fundamentals-design-your-account-and-vpc-architecture-for-enterprise-operating-models-ent203.

[4] Mohammad Alizadeh, Shuang Yang, Milad Sharif, Sachin Katti, Nick McKeown, Balaji Prabhakar, and Scott Shenker. "pfabric: Minimal near-optimal datacenter transport". In: *ACM SIGCOMM Computer Communication Review* 43.4 (2013), pp. 435–446.

[5] Hitesh Ballani, Paolo Costa, Istvan Haller, Krzysztof Jozwik, Kai Shi, Benn Thomsen, and Hugh Williams. "Bridging the Last Mile for Optical Switching in Data Centers". In: *Optical Fiber Communication Conference (OFC'18).* OSA, 2018. URL: https://www.microsoft.com/en-us/research/publication/bridging-last-mile-optical-switching-data-centers/.

[6] Hamid Hajabdolali Bazzaz, Malveeka Tewari, Guohui Wang, George Porter, T. S. Eugene Ng, David G. Andersen, Michael Kaminsky, Michael A. Kozuch, and Amin Vahdat. "Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks". In: *Proceedings of the 2nd ACM Symposium on Cloud Computing.* SOCC '11. Cascais, Portugal: Association for Computing Machinery, 2011. ISBN: 9781450309769. DOI: 10.1145/2038916.2038946. URL: https://doi.org/10.1145/2038916.2038946.

[7] Andreas Bechtolsheim. *400G and 800G Ethernet and Optics.* https://pc.nanog.org/static/published/meetings/NANOG75/1954/20190220_Martin_Building_The_400G_v1.pdf.

[8] Ilker Bozkurt, Anthony Aguirre, Balakrishnan Chandrasekaran, P. Godfrey, Gregory Laughlin, Bruce Maggs, and Ankit Singla. "Why Is the Internet so Slow?!" In: Feb. 2017, pp. 173–187.

[9] Calient. *Edge 640 Optical Circuit Switch.* https://www.calient.net/products/edge640-optical-circuit-switch/.

[10] Qixiang Cheng, Meisam Bahadori, Madeleine Glick, and Keren Bergman. "Scalable Space-and-Wavelength Selective Switch Architecture Using Microring Resonators". In: *Conference on Lasers and Electro-Optics.* Optical Society of America, 2019, STh1N.4. DOI: 10.1364/CLEO_SI.2019.STh1N.4. URL: http://www.osapublishing.org/abstract.cfm?URI=CLEO_SI-2019-STh1N.4.

[11] Qixiang Cheng, Sébastien Rumley, Meisam Bahadori, and Keren Bergman. "Photonic switching in high performance datacenters". In: *Opt. Express* 26.12 (2018), pp. 16022–16043. DOI: 10.1364/OE.26.016022. URL: http://www.opticsexpress.org/abstract.cfm?URI=oe-26-12-16022.

[12] Ciena. *Optical Amplifier Modules.* URL: https://media.ciena.com/documents/Optical_Amplifier_Modules_A4_DS.pdf.

[13] James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, JJ Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. "Spanner: Google's Globally-Distributed Database". In: *10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12).* USENIX Association, 2012, pp. 261–264.

[14] N. G. Duffield, Pawan Goyal, Albert Greenberg, Partho Mishra, K. K. Ramakrishnan, and Jacobus E. van der Merive. "A Flexible Model for Resource Management in Virtual Private Networks". In: *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication.* SIGCOMM '99. Cambridge, Massachusetts, USA: Association for Computing Machinery, 1999, pp. 95–108. ISBN: 1581131356. DOI: 10.1145/316188.316209. URL: https://doi.org/10.1145/316188.316209.

[15] Ramakrishnan Durairajan, Paul Barford, Joel Sommers, and Walter Willinger. "InterTubes: A Study of the US Long-haul Fiber-optic Infrastructure". In: *Proceedings of the ACM SIGCOMM* 45 (Aug. 2015), pp. 565–578.

[16] R. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, and B. Goebel. "Capacity Limits of Optical Fiber Networks". In: *Journal of Lightwave Technology* 28.4 (2010), pp. 662–701. ISSN: 1558-2213. DOI: 10.1109/JLT.2009.2039464.

[17] Nathan Farrington, Alex Forencich, George Porter, Pang-Chen Sun, Joseph Ford, Yeshaiahu Fainman, George Papen, and Amin Vahdat. "A Multiport Microsecond Optical Circuit Switch for Data Center Networking". In: *Photonics Technology Letters, IEEE* 25 (Aug. 2013), pp. 1589–1592.

[18] Nathan Farrington, George Porter, Sivasankar Radhakrishnan, Hamid Bazzaz, Vikram Subramanya, Yeshaiahu Fainman, George Papen, and Amin Vahdat. "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers". In: vol. 41. Oct. 2010, pp. 339–350.

[19] M. Filer, S. Searcy, Y. Fu, R. Nagarajan, and S. Tibuleac. "Demonstration and performance analysis of 4 Tb/s DWDM metro-DCI system with 100G PAM4 QSFP28 modules". In: *2017 Optical Fiber Communications Conference and Exhibition (OFC).* 2017.

[20] Mark Filer, Jamie Gaudette, Yawei Yin, Denizcan Billor, Zahra Bakhtiari, and Jeffrey L. Cox. "Low-margin optical networking at cloud scale". In: *J. Opt. Commun. Netw.* 11.10 (2019), pp. C94–C108. DOI: 10.1364/JOCN.11.000C94. URL: http://jocn.osa.org/abstract.cfm?URI=jocn-11-10-C94.

[21] Data Center Frontier. *Vertical Data Centers: 'Watts Per Acre' Guides Construction Economics.* https://datacenterfrontier.com/vertical-data-centers-watts-per-acre-guides-construction-economics/.

[22] Adam Funnel, Kai Shi, Paolo Costa, Philip Watts, Hitesh Ballani, and Benn Thomsen. "Hybrid Wavelength Switched-TDMA High Port Count All-Optical Data Centre Switch". In: *OSA Journal of Lightwave Technology* 35.20 (2017).

[23] Monia Ghobadi, Ratul Mahajan, Amar Phanishayee, Nikhil R. Devanur, Janardhan Kulkarni, Gireeja Ranade, Pierre-Alexandre Blanche, Houman Rastegarfar, Madeleine Glick, and Daniel C. Kilper. "ProjecToR: Agile Reconfigurable Data Center Interconnect". In: *Proceedings of the ACM SIGCOMM 2016 Conference, Florianopolis, Brazil, August 22-26, 2016*. 2016, pp. 216–229.

[24] Albert G. Greenberg, James R. Hamilton, Navendu Jain, Srikanth Kandula, Changhoon Kim, Parantap Lahiri, David A. Maltz, Parveen Patel, and Sudipta Sengupta. "VL2: A Scalable and Flexible Data Center Network". In: *ACM SIGCOMM* 54.3 (2011), pp. 95–104.

[25] Sangyoon Han, Tae Joon Seok, Niels Quack, Byung-Wook Yoo, and Ming Wu. "Monolithic 50x50 MEMS Silicon Photonic Switches with Microsecond Response Time". In: 2014.

[26] Chi-Yao Hong, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Vijay Gill, Mohan Nanduri, and Roger Wattenhofer. "Achieving high utilization with software-driven WAN". In: *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. 2013, pp. 15–26.

[27] Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, and et al. "B4: Experience with a Globally-Deployed Software Defined Wan". In: *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*. SIGCOMM '13. Hong Kong, China: Association for Computing Machinery, 2013, pp. 3–14. ISBN: 9781450320566. DOI: 10.1145/2486001.2486019. URL: https://doi.org/10.1145/2486001.2486019.

[28] Xin Jin, Yiran Li, Da Wei, Siming Li, Jie Gao, Lei Xu, Guangzhi Li, Wei Xu, and Jennifer L. Rexford. "Optimizing bulk transfers with software-defined optical WAN". English (US). In: *SIGCOMM 2016 - Proceedings of the 2016 ACM Conference on Special Interest Group on Data Communication*. SIGCOMM 2016 - Proceedings of the 2016 ACM Conference on Special Interest Group on Data Communication. Association for Computing Machinery, Inc, Aug. 2016, pp. 87–100. DOI: 10.1145/2934872.2934904.

[29] Alpar Juttner, Istvan Szabo, and Aron Szentesi. "On bandwidth efficiency of the hose resource management model in virtual private networks". In: *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428)*. Vol. 1. 2003, 386–395 vol.1. DOI: 10.1109/INFCOM.2003.1208690.

[30] Kazuro Kikuchi. "Fundamentals of Coherent Optical Fiber Communications". In: *Journal of Lightwave Technology* 34.1 (2016), pp. 157–179.

[31] Ko, K.Y. and Demokan, M.S. and Tam, H.Y. "Transient Analysis of Erbium-Doped Fiber Amplifiers". In: *Photonics Technology Letters, IEEE* 6 (Jan. 1995), pp. 1436 –1438. DOI: 10.1109/68.392219.

[32] Thomas L. Koch. *Optical Fiber Telecommunications IIIA (Optics and Photonics)*. Academic Press, 1997.

[33] LIGHTWAVE. *Polatis goes large with 192x192 all-optical switch*. https://www.lightwaveonline.com/optical-tech/transport/article/16664779/polatis-goes-large-with-192x192-alloptical-switch. 2012.

[34] E. Maniloff, S. Gareau, and M. Moyer. "400G and Beyond: Coherent Evolution to High-Capacity Inter Data Center Links". In: *Optical Fiber Communication Conference (OFC) 2019*. Optical Society of America, 2019.

[35] D. M. Marom, P. D. Colbourne, A. D'errico, N. K. Fontaine, Y. Ikuma, R. Proietti, L. Zong, J. M. Rivas-Moscoso, and I. Tomkos. "Survey of photonic switching architectures and technologies in support of spatially and spectrally flexible optical networking [invited]". In: *IEEE/OSA Journal of Optical Communications and Networking* 9.1 (2017), pp. 1–26. ISSN: 1943-0639. DOI: 10.1364/JOCN.9.000001.

[36] Microsoft Research. *Optics for the Cloud Group*. http://opticsforthecloud.com/.

[37] A. Mokhtar and M. Azizoglu. "Adaptive wavelength routing in all-optical networks". In: *IEEE/ACM Transactions on Networking* 6.2 (1998), pp. 197–206. ISSN: 1558-2566. DOI: 10.1109/90.664268.

[38] OFC Conference. *Optical Data Center Interconnect: Hot and Highly Competitive*. https://www.ofcconference.org/en-us/home/about/ofc-blog/2018/february-2018/optical-data-center-interconnect-hot-and-highly/. 2018.

[39] OIF. *400ZR*. https://www.oiforum.com/technicalwork/hot-topics/400zr-2/.

[40] Polatis. *SERIES 7000 - 384X384 PORT SOFTWARE-DEFINED OPTICAL CIRCUIT SWITCH*. URL: https://polatis.com/series-7000-384x384-port-software-controlled-optical-circuit-switch-sdn-enabled.asp.

[41] Arjun Roy, Hongyi Zeng, Jasmeet Bagga, George Porter, and Alex C Snoeren. "Inside the social network's (datacenter) network". In: *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. 2015, pp. 123–137.

[42] John Simsarian, Michael Larson, Henry Garrett, Hong Xu, and Timothy Strand. "Less than 5-ns wavelength switching with an SG-DBR laser". In: *Photonics Technology Letters, IEEE* 18 (Feb. 2006), pp. 565 –567. DOI: 10.1109/LPT.2005.863976.

[43] Guohui Wang, David G. Andersen, Michael Kaminsky, Michael Kozuch, T. S. Eugene Ng, Konstantina Papagiannaki, Madeleine Glick, and Lily B. Mummert. "Your Data Center Is a Router: The Case for Reconfigurable Optical Circuit Switched Paths." In: *HotNets*. Ed. by Lakshminarayanan Subramanian, Will E. Leland, and Ratul Mahajan. ACM SIGCOMM, 2009. URL: http://dblp.uni-trier.de/db/conf/hotnets/hotnets2009.html#WangAKKNPGM09.

[44] Yevgeniy Sverdlik. *Facebook Rethinks In-Region Data Center Interconnection*. https://www.datacenterknowledge.com/networks/facebook-rethinks-region-data-center-interconnection. 2018.

[45] Xiang Zhou and Hong Liu. *Pluggable DWDM: Considerations For Campus and Metro DCI Applications*. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45713.pdf. 2016.

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

A AMPLIFIER AND CUT-THROUGH LINK PLACEMENT

We have to guarantee that transceivers from each data center can reach any other data center in the region. There are two reasons why this could be a potential problem. First, the distance between two sites can be longer than 80 km and the signal requires in-network amplification, and second, there may be too many switching points on the path that reduce the signal power below the power threshold for the receiver.

These problems can be resolved by placing in-network amplifiers (at most one per path) or building "cut-through links" that traverse the switching point without being interrupted (switched), and thus reduce the power loss. Amplifier placement can solve both problems in some situations. Even if the distance is short, but there are many switching points on the path, it may make sense to place amplifiers and increase the signal power instead of building cut-through links that reduce the power loss, because the number of amplifiers needed could be cheaper compared to allocating additional fiber for cut-through links.

Note that there is always *a* configuration that meets all constraints because no links longer that 80 km is allowed in the topology in the first place. This means that a path of 120 km can always be divided into two segments where each of them is not longer than 80 km.

Our goal is to meet all constraints by minimizing the cost. An optimal solution would require exploring every possible combination of failures, amplifier placement and cut-through links. This problem has combinatorial complexity since for each path of h hops, there are 2^h potential cut-through links to be built.

To simplify the process, we place amplifiers using a greedy heuristic described in Algorithm 2. For every failure scenario, we identify all paths that are long and require amplification. For each path, we find all candidate locations where amplifier placement can resolve the power budget constraint. Since one amplifier can amplify only one fiber, the total number of amplifiers needed in a particular location is calculated from the maximum demand on all paths that require amplification at that location, similarly to the maximum capacity calculation in §4.1. We also calculate how many of these long paths suffer from too many hops that reduce the signal power and if amplifier placement at a particular location would resolve that constraint as well. Then, we assign a score to each location based on the total number of constraints resolved versus the total number of amplifiers needed. Finally, we place amplifiers to a location with the maximum score and repeat this process as long as there are paths that require amplification.

Algorithm 2: Algorithm for amplifier placement

```
foreach failure scenario do
    P ← {long paths that require amplification}
    while size(P) > 0 do
        S ← {possible amplifier locations ∀path∈P }
        foreach location ∈ S do
            noa ←# of amplifiers needed at location
            noea ←# of amplifiers already at location
            /* # of amplifiers to be placed      */
            ntbp ←max(0, noa - noea)
            nop ←# of paths resolved by placing amplifiers
              at location
            nhop ←# of paths that resolve the n-hop
              constraint by placing an amplifier at location
            location_score ← $\frac{nop + nhop}{ntbp}$
        mloc ←the location with maximum score
        place amplifiers at mloc
        P ←P −{ paths resolved by mloc }
```

After the amplifiers have been placed, there can still be paths that have too many hops that cause the signal power to drop below the acceptable threshold. Thus, we apply a similar heuristic as for the amplifier placement to place cut-through links and avoid fiber switching at every hop. To do that, we introduce the concept of a segment. If a path does not have an amplification point, the segment is equivalent to the path. However, with an amplification point, a path has two segments, one between the source and the amplifier, and one between the amplifier and the destination. For each segment that has too many hops, we calculate all possible cut-through links that would resolve the power constraint on that segment. Similarly to the previous heuristic, we assign a score to every cut-through candidate based on the number of paths that can utilize the link versus additional fiber needed for that particular link. The cut-through link with the highest score is added to the topology and the process starts again as long as there are segments that have the power budget problem.

The proposed heuristics may not provide an optimal result in terms of cost but they guarantee that all constraints will be met. First, the amplifier placement algorithm assures that there are no long links that require amplification because of distance. Following that, the cut-through placement heuristic guarantees that the distance between source/destination and the amplification point can be bridged with a sufficient power budget.

The cost overhead due to additional amplifiers and cut-through links using the described heuristic is 3% on average (8% in the worst case) compared to the total network cost across all test scenarios.

B WAVELENGTH-SWITCHED NETWORK DESIGNS

Pure wavelength-switching. A design based fully on wavelength switching would demultiplex each fiber's wavelengths at the switching points. This would use an optical cross connect architecture, as noted in §3.2. Wouldn't such a design be *obviously* superior? Surprisingly, the answer is no. While we analyzed this precisely, fleshing out a wavelength-switched design, we only summarize here the reasoning behind this result:

- With at most one OXC switch on path (*TC4*) and only one amplifier per path (*TC2*), it is not feasible to benefit from wavelength switching in many settings.
- Wavelength switching adds complexity, requiring the solution of a graph-coloring problem to avoid wavelength collisions.

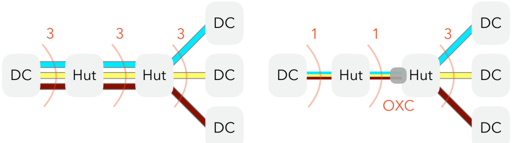

Fig. 15: *(left) Iris requires having one fiber DC pair that causes the total overhead of $O(n^2)$ fibers; (right) Hybrid design reduces that overhead by combining multiple residual links using wavelength switching.*

- Even ignoring the above two issues, and using settings favorable to wavelength switching (*e.g.*, large n=20), at today's prices, the additional components needed for wavelength switching are pricier than the n^2 additional fibers for fiber switching.

Hybrid design. To support any traffic matrix, a fiber switched network requires $O(n^2)$ additional links to carry residual capacity. These links only serve fractional capacity that cannot be accommodated in the base fiber. Intuitively, many of these links could be combined using a finer-grained wavelength switching technology, and thus, reduce the fiber overhead, as shown in the example in Fig. 15. Residual capacity to different destinations can be combined at the source data center, carried in one fiber to a particular fiber hut that is on the shortest path for all combined wavelengths. At the fiber hut, the wavelengths are separated and carried through dedicated fibers to different destinations. The same process applies in the opposite direction as well – residual wavelengths to the same destination can be combined at a fiber hut and carried through one fiber to the common destination.

If we combine x residual fibers into one, we have to guarantee that these x residual fibers combined cannot exceed the capacity of one physical fiber – λ wavelengths.

Observation 1. *Any 2 residual fibers coming from the same source can be combined into one fiber.*

To show this, we have to define the concept of base capacity. The base capacity is the capacity that has to be provided to satisfy operational constraints defined in §3.1, regardless of the technology used for implementation. Iris requires the base capacity plus n^2 residual links. The base capacity links are always fully saturated with λ wavelengths. If the demand to a particular destination is less than λ, the traffic is carried through a residual link.

If two residual fibers carry more than λ wavelength, it means there is at least one fiber provisioned among those in the base capacity. Then, the residual capacity to one destination will be transmitted through one fiber form the base capacity and the other one remaining residual fiber. This result enables a simple optimization that should reduce the n^2 fiber overhead to close to $n^2/2$. However, we show we can potentially save even more.

Observation 2. *Any n residual fibers coming from the same source can be combined into $\lceil n/4 \rceil$ fibers.*

Let us assume there is a data center that can reach n destinations (n residual fibers). Assume that the aggregated traffic demand from this data center to all destinations is D wavelengths. Without loss of generality, we assume that $D \leq \lambda \cdot n$, where n is the total number of destinations (for larger D, the difference would be carried through the base capacity). We want to calculate what is the maximum capacity that *must* be carried through the residual fiber. n residual fibers are shown in Fig. 16. By the definition of base capacity, we know that the base capacity provisions at least $B = D/\lambda$ fibers available, and the rest must be transported through the residual fibers. Since there are n destinations, we will need to provision $R = n - D/\lambda$ residual links atop base capacity.

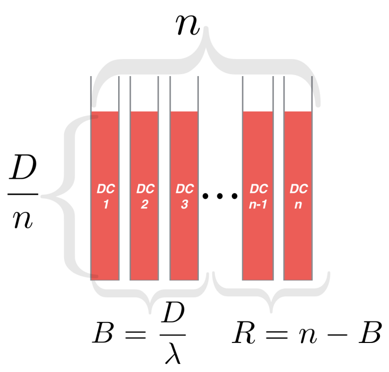

Fig. 16: *Illustration of the worst-case residual capacity allocation. The total capacity that will be carried over residual links is equivalent to $R \cdot D/n$*

We are looking for a traffic matrix that maximizes the capacity carried over these R links. For any traffic matrix, we take the following approach: we sort all demands to n destinations. Largest B fibers will be scheduled using the base capacity, and the remaining part will must go through the R fibers. The total demand in R is maximized if every link carries exactly the same capacity D/n. Thus, the total capacity carried over residual links is $(n - D/\lambda) \cdot D/n$. This function has the maximum for $D = \lambda \cdot n/2$ and the total worst-case capacity on R fibers is $\lambda \cdot n/4$ wavelengths.

This further means that any n residual links coming from the same source will carry at most $\lambda \cdot n/4$ wavelengths. Since each fiber can carry at most λ wavelengths, this means that given residual capacity can be compressed into:

$$\left\lceil \frac{\lambda \cdot n}{4} \cdot \frac{1}{\lambda} \right\rceil = \left\lceil \frac{n}{4} \right\rceil$$

Note that the theorem holds for residual wavelengths that have the same source, as well as for those that have the same destination. This result allows us to merge any 4 residual fibers with the same source/destination.

However, there are two additional challenges that prevent us from minimizing the fiber overhead by a factor of 4:

- Optical devices that are used to pack/unpack wavelengths from the fiber introduce significant signal power loss. Thus, we can afford to have only one wavelength switching device per path.
- Two or more residual fibers can be combined only if they share a subpath from the source / to the destination. For instance, in a distributed network with many direct connections, little fiber can be saved because there are only a few paths that share a subpath.

Note that the devices used for packing and unpacking wavelengths have to be active and dynamic because there are different combinations of residual capacity that these devices have to handle and these combinations change over time, depending on the traffic matrix.

The remaining step in designing a hybrid network is to decide if and where residual fiber links will be aggregated. This problem has similar properties to amplifier placement and cut-through link placement, so we take a similar approach. We compute all possible placements for wavelength switching devices, we give each solution a score based on the potential fiber saving, pick the location with the highest score, place the devices, and repeat this process as long as any fiber saving can be achieved. In our test scenarios, this approach managed to reduce the residual fiber overhead by approximately 50%, which brings some cost reductions, as described in §6, but with the current prices, this is not enough to justify the complexity of managing a network with one more type of devices. However, we envision that this hybrid design could be the first step toward a more sophisticated solution with less fiber overhead.

C EXPERIMENTAL DETAILS

Physical layer experiments. More details of the experiment reported in §6.2 are discussed here for completeness. Four dual polarization (DP) 200 Gbit/s 16 quadrature amplitude modulation (QAM) optical signals are generated by commercially available real-time coherent transceivers, Acacia AC200 and AC400, to produce 2^{31} pseudo random bit sequences. They are spectrally shaped with a root-raised cosine with a 0.2 roll-off factor and with 15% overhead. An amplified spontaneous emission (ASE) source emulates dense wavelength division multiplexed (DWDM) channels ("Channel emulation"), which are then split and multiplexed with the signals in two separate single mode fibers (SMFs) via two *WSS*es to emulate full C-band lines. It is worth pointing out that at the wavelengths of the live signals no ASE was present by the channel emulator since it was properly filtered by the *WSS*. In the experiment the signals wavelengths were tuned within the C-band with similar achieved results. At the receiver side the optical signals under test are demultiplexed and sent to coherent receivers [30] to be converted in the electrical domain. The optical-to-electrical converted signals are fed to the application-specific integrated circuit (ASIC)'s analogue to digital conversion for further processing by the ASIC's digital signal processing, which includes signal recovery, polarization mode dispersion and chromatic dispersion compensation, before SD-FEC decoding. Pre-FEC BER measurements are taken every 10 msec and the received powers are kept within the range of the receiver's optimal performance. Insets of Fig. 13(b) show examples of typical constellation diagrams of the tested signals in our experiments. The constellation diagrams of the tested signals are shown once converted in the electrical domain at different points in the system. They display the signals as a two-dimensional plane diagram in the complex plane at symbol sampling instants. The angle of a point, measured counter-clockwise from the horizontal axis, represents the phase shift of the carrier wave from a reference phase, given by the local oscillator in the coherent receiver. The distance of a point from the plane origin represents a measure of the amplitude or power of the signal. As expected for 16 QAM signals, 16 distinct symbols are visible for both polarizations, x and y. The cloud associated to each symbol is caused by noise. Due to transmission impairments, the constellation diagrams at the end of the system are characterized by a higher degree of noise as compared to the ones before transmission, but are within the range of acceptable received performance, as confirmed by the BER measurements reported in Fig. 14. Insets of Fig. 13(b) further show spectral traces that cover the whole C-band before and after the fiber spans. The traces show how Iris emulates missing channels to fill the unused spectrum while at the same time keeping per-channel power roughly constant so that no per-channel power management is required. These traces are measured in the frequency/wavelength domain using an optical spectrum analyzer.

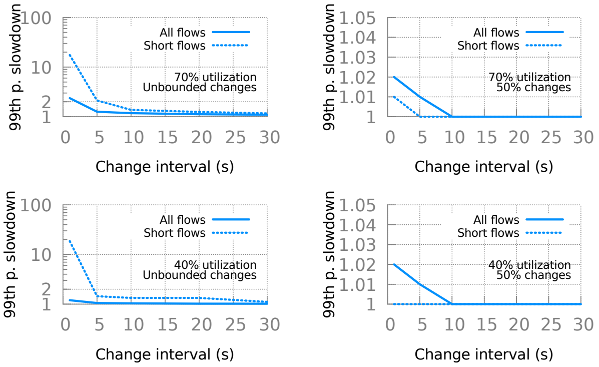

Fig. 17: *Slowdown under reconfiguration (ratio of* 99^{th}*-%ile FCT for Iris vs. EPS). Even at high utilization (70%) and large changes (>50%), slowdown is minimal for reasonable frequencies of reconfiguration.*

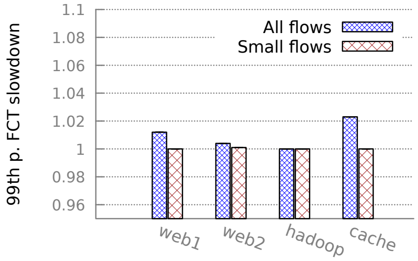

Fig. 18: 99^{th}*-percentile slowdown at 40% util., 50% traffic changes, and reconfig. every 5 sec for various workloads; web1 is from [4] and the rest from [41]. Iris's slowdown is <2% compared to EPS.*

Region-scale simulations. Here, we provide the detailed results and figures from the large flow-based simulations of §6.3.

We compare the Flow Completion Times (FCTs) for Iris to an EPS fabric baseline. Fig. 17 highlights the increase in the 99^{th} percentile of FCT across some of our tested parameters, including the most extreme. As the results show, with the exception of unbounded intensity changes at high utilization, the effect is minimal, especially for reconfiguration intervals of 10 sec or above. Fig. 18 similarly shows that this is the case across all tested flow size distributions.