Inter-Datacenter Bulk Transfers: Trends and Challenges

Long Luo, Hongfang Yu, Klaus-Tycho Foerster, Max Noormohammadpour, and Stefan Schmid

***Abstract*—Many modern Internet services and applications operate on top of geographically distributed datacenters, which replicate large volumes of business data and other content, to improve performance and reliability. This leads to high volumes of continuous network traffic between datacenters, constituting the bulk of traffic exchanged over the Wide-Area Networks (WANs) that connect datacenters. Operators heavily rely on the efficient and timely delivery of such traffic, as it is key for the performance of distributed datacenter applications, and bulk inter-datacenter traffic also plays an important role for network capacity planning.**

In this paper, we discuss the unique and salient features of bulk across-datacenter transfers and extensively review recent technologies and existing research solutions proposed in the literature for optimization of such traffic. Moreover, we discuss several challenges and interesting directions that call for substantial future research efforts.

I. Introduction

Datacenter networks are growing explosively in both size and numbers with the increasing popularity of online services on many fronts (e.g., health, business, streaming, social networking)—they have become a critical infrastructure of today's digital society. The total number of hyperscale datacenters that are capable to scale out on-demand has increased to over 500 in 2019, which has been tripled since 2013 [1]. It takes only two years to build more than 100 hyperscale datacenters, and the rate is accelerating. To offer cloud services with improved performance, providers typically build datacenters in distant regions close to their global customers, see Fig 1. For example as of 2019, Amazon, Microsoft, and Google operate datacenters across dozens of geographical regions and more than one hundred cities.

With these geographically dispersed datacenters, cloud providers can easily support thousands of services such as web and streaming. Operating services globally, in turn, increases the inter-DC WAN pressure in that services continuously produce an increasing amount of data transfers that typically range from tens of terabytes to petabytes among their working datacenters for increasingly stringent availability, performance, and consistency [2]–[5]. For example, Facebook is experiencing an exponential growth of the traffic carried by their inter-DC networks over the years, this trend is expected to continue in the future [6]. A relatively small portion (e.g., 5%-15% [4]) of inter-DC traffic is user-facing (e.g., interactive user search queries) which happens between customers and datacenters. User-facing traffic is highly sensitive to delay and hence always prioritized over application to application traffic and delivered with appropriate reserved bandwidth. On the other hand, the bulk of traffic carried by inter-DC networks is fast-growing internal traffic, which is application to application (e.g., search index synchronization and database replication). Internal traffic can tolerate additional delay and is hence delivered over the leftover bandwidth of high-priority user-facing traffic.

In order to support massive volumes of traffic, inter-DC network operators, such as Amazon and Facebook, need to invest in expensive high capacity WAN backbones that connect their datacenters globally. Moreover, many cloud providers invest large sums of capital in leasing bandwidth from Internet Service Providers (ISPs) to transfer data across their datacenters. In general, major cloud providers spend in the order of hundreds of millions of dollars per year (amortized) on maintaining such connectivity. Therefore, efficient utilization of network bandwidth is critical to maximize cost savings and optimize performance given existing capacity at any given time. Motivated by these trends, the networking community has recently done many efforts on the design of new traffic engineering solutions for inter-DC traffic, especially to deliver bulk transfers efficiently. To this end, their solutions pursue research goals of improving resource utilization [2], [3], [7], meeting transfer deadlines [4], [8], [9], minimizing transfer completion times [10], reducing transmission costs [11], etc.

Our short tutorial paper provides an introduction to current trends and open problems in this area. We discuss communication patterns, performance metrics, and objectives of inter-datacenter bulk transfers. Moreover, we review the transfer techniques used to optimize the large inter-DC transfers in existing solutions and point selected open problems in the literature that will be of interest for future work.

II. Current situation of inter-DC transfers

We first briefly review the inter-DC bulk transfers studied by current works, considering three dimensions: communication patterns, performance metrics, and operator objectives.

A. Communication Patterns

Inter-DC bulk transfers can be broadly classified into the following communication patterns, with respect to the number of destinations on a data transfer, see Fig. 1. One is the *Point to Point (P2P)* transfers that deliver data from the source datacenter to a single destination datacenter. For example, to

Long Luo is with the University of Electronic Science and Technology of China; Hongfang Yu is with the University of Electronic Science and Technology of China and Peng Cheng Laboratory; Stefan Schmid and Klaus-T. Foerster are at the Faculty of Computer Science, University of Vienna; Max Noormohammadpour is with Facebook. Corresponding author: Hongfang Yu.

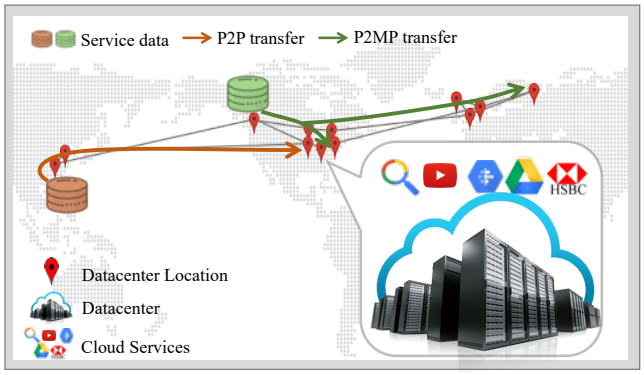

Fig. 1: Illustration of an inter-datacenter WAN and bulk transfers, as in Google's B4 [3].

TABLE I: Representative transfer scheduling objectives

Objective	Description
Fairness	Resources should be fairly allocated among transfers to achieve max-min fairness [2], [3] or completion fairness [5].
Maximizing Utilization	Fully using available network bandwidth is necessary to improve network throughput [2], [3].
Minimizing Transfer Completion Times	Short completion times improve the overall quality of services and network efficiency [10].
Maximize Satisfied Demand	For deadline-constrained transfers of equal value, maximizing the number of deadline-satisfied transfers maximizes the satisfied demand and the network utility [4], [8].
Minimizing Transmission Cost	Large data transmissions over the inter-DC networks of service providers are of substantial cost. A fundamental objective is hence to reduce or even minimize the transmission charges incurred by the inter-DC bulk transfers [11].

protect a datacenter against failures or natural disasters such as snowstorms and floods, cloud backup services may back up data onto a geographically distant datacenter. Quite a few efforts are centered around P2P transfers such as Google B4 [3], Microsoft TEMPUS [5], and Amoeba [4]. Another one is the *Point to Multipoint (P2MP)* transfers which replicate data from the source datacenter to multiple destination datacenters. Geo-replication is usually the main datacenter service that generates transfers with such communication patterns [10]. For example, search engines use geo-replication to periodically (e.g., every 24 hours) synchronize search index updates across their datacenters, video streaming services disseminate high-definition video content to multiple locations close to their regional customers [4], and financial institutions back up their daily transaction records to datacenter sites [12].

P2MP transfers is forming a large portion of workloads over inter-DC WANs (e.g., up to 91% for some large-scale online service providers [13]). Hence, many of the current optimizations on the basis of P2P transmission are becoming less efficient. In this past two years, we have seen an increasing number of efforts (e.g., [10] and [9]) devoted to optimizing P2MP transmission. This trend is likely to continue: geo-replication traffic will continuously grow with the rapid expansion of datacenters and the proliferation of global cloud services, which demands more effort dedicated to P2MP transfers to improve efficiency.

B. Performance Metrics

Inter-DC bulk transfers typically have a size that ranges from tens of terabytes to petabytes, and a key performance requirement of them is timely delivery [5]. Delayed completion will significantly degrade service quality and even violate SLAs between users and service providers. The performance metrics considered by a majority of prior work on inter-DC large transfers can be roughly classified into two types: 1) to meet *deadlines* and 2) to minimize transfer *completion times*.

Deadlines specify the time period that the data delivery needs to be finished, which is imposed by many cloud services [4]. Such deadlines may represent, e.g., different consumer SLAs, or the importance of transfers for businesses and organizations. Most prior works (e.g., [4], [8], [9]) focus on deadlines, where a transfer may be delayed in favor of another one with a closer deadline. Recent work (e.g., [4]) also considers transfers associated with different types of deadlines, i.e., hard and soft deadlines. In the case of hard deadlines, a late transfer is useless to applications and so guaranteeing completion prior to the deadline is necessary to improve the transfer efficiency. Transfers with soft deadlines, on the other hand, can be delayed to some extent given a decreasing utility over time, but still need to finish before some later deadline.

Transfer completion time refers to the total time needed to finish a transfer request from when it arrives at the network. As delivering data as soon as possible is critical to the performance of many global-scale cloud services, completion time is the primary metric to be considered when applications do not specify a deadline. Most work on deadline-unconstrained bulk transfers focuses on the average completion time [10], [12], while a handful of them (e.g., Owan [12]) also consider the total time to complete a given collection of transfers, i.e., the *makespan*.

Performance metrics of bulk transfers can be used to form constraints on data delivery and various objective functions, as outlined in the following section, depending on what is important to service providers and applications.

C. Objectives of Scheduling Inter-DC Bulk Transfers

Table I summarizes the common objectives of prior traffic engineering solutions for inter-DC bulk transfers. Fairness is one of the main objectives of the solutions [2], [3], [5] considered by large technological companies, such as Microsoft and Google. In particular, they focus on max-min (throughput) fairness [2], [3] and completion fairness that maximizes the minimal deadline-meeting fraction when it is not possible to meet all deadlines for transfers with soft deadlines [5].

For many large cloud providers which deploy their own dedicated WANs connecting distant datacenters, most of their inter-DC WAN resources and costs are fixed over relatively short periods of time. As a result, they also focus on maximizing the resource utilization by accommodating as many data transfers as possible [2], [3]. As timely delivery is crucial to the performance of many online services, minimizing the average, median, or tail transfer completion is another goal for many providers [10], [12]. For deadline-sensitive bulk

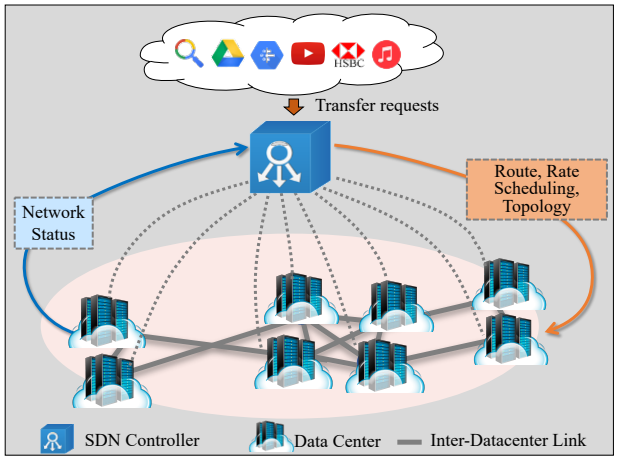

Fig. 2: High-level overview of an SDN-based control system for engineering bulk transfers over inter-DC WANs.

transfers, delivering traffic without paying attention to deadlines could waste precious inter-DC link bandwidth, as late transfers can have much lower utility. In such cases, the goal of minimizing the fraction of missed deadlines or lateness can maximize the value of deadline-sensitive data transfers (satisfied demands) and improve the effective utilization of scarce and expensive WAN resources [4], [8], [9].

On the other hand, cloud service providers which build their private inter-DC WANs often invest huge on building own optics or leasing links from ISPs. One major objective of these providers is to minimize the cost incurred by data transmissions. Hence, many prior works (e.g., TrafficShaper [11]) consider the intensity of inter-DC bulk transfers and aim to minimize the transmission cost of the network bandwidth according to certain usage-based charging models.

III. Techniques for Bulk Transfers

In this section, we first introduce a Software Defined Networking (SDN) based control system that is widely adopted by many solutions for bulk transfer optimization. Then, we will review the major techniques used by prior work in the literature, which rely on bulk transfer traffic characteristics.

A. SDN-based Control Systems

An increasing number of companies and organizations, such as Google [3], Microsoft [2], and Facebook [6], have adopted SDN to operate their inter-DC WANs in the past few years. The reasons are twofold: first, an inter-DC WAN usually has only a few dozens of datacenters, which any SDN-style centralized network control can easily manage without scalability issues. Second, it's the flexibilities and optimization opportunities of SDN, which enhance the network ability to optimize the WAN resources allocation to increase the overall network utility. Fig. 2 provides a high-level overview of SDN-based control systems for inter-DC bulk transfers, where cloud services submit their transfer requests and the providers optimize the delivery (e.g., forwarding routes, transmission rates and scheduling policies) of these requests and even the network topology towards certain objectives, with well-designed transfer allocation approaches and techniques.

In the following, we provide an overview of several selected techniques used in prior research and discuss their respective contributions.

B. Transmission Techniques

Generally, transmission control techniques determine the routing and forwarding of traffic, the rate at which traffic is transmitted, and the delivery scheme of data that needs to traverse the inter-DC WAN. In the case of inter-DC bulk traffic, these techniques largely rely on the characteristics of such traffic and are usually tailored to scheduling objectives concerned by service providers. A *"key characteristics"* [4] are associated deadlines (timely delivery requirement), but at the same time, such transfers are delay-tolerant (i.e., "*can be served more flexibly*" [5]). The combination of both hence enables scheduling techniques, to be also covered in more detail later.

1) Routing and forwarding: Routing schemes decide which forwarding route, including intermediate datacenters and inter-DC links, the traffic takes when traversing the inter-DC WAN from the source towards the destination datacenter(s). With the adoption of SDN, routing decisions are naturally performed in a centralized manner, which allows for (close to) optimal and computationally tractable solutions, as most inter-DC WANs only contain about a few dozen sites [2], [3], [6]. Notwithstanding, multiple controllers can be incorporated as well [14]. The design of centralized routing schemes typically take into account whether the routes change in the process of data delivery and how the data is delivered.

In this context, routing can be *static or dynamic*. Static approaches keep routes fixed after the initial assignment and do not update it throughout the delivery process. Static forwarding paths are convenient for operators, as they avoid the transient routing inconsistency, such as blackholes and forwarding loops [2], which could arise due to dynamic updates. Many prior work [4], [5] precomputes a collection of paths (e.g., k-shortest paths) between every pair of datacenters and uses them as tunnels to deliver bulk traffic among datacenters. To adapt to changes in network condition (e.g., the increase or decrease of traffic intensity, capacity changes, etc.), some research work, such as [9], dynamically recalculates forwarding routes for every transfer request, either periodically (e.g., every several minutes) or when new transfer requests arrive, to improve efficiency and performance. Despite the network having to suffer from transient throughput drops for performing consistent and congestion-free routing updates [2], dynamic routing is still suitable for bulk transfers because such traffic is often long-lived and less sensitive to delay.

Routing of bulk transfers can be *unicasting or multicasting*. Unicasting delivers every copy of data over a separate path every destination, while multicasting can deliver data to multiple destinations simultaneously. It is natural to deliver a P2P transfer via unicasting as leveraged by many existing works [2]–[5]. For P2MP transfers, some works transform the setting into multiple P2P transfers, each with one of the

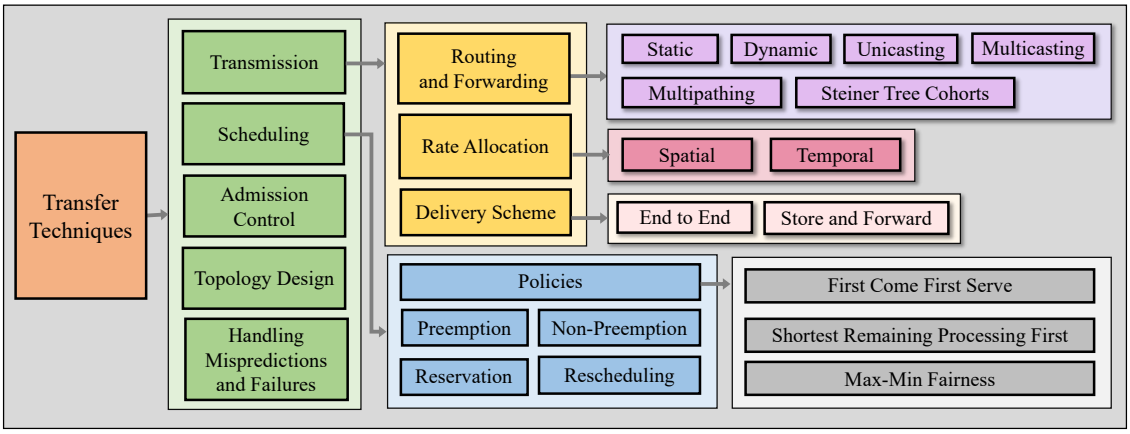

Fig. 3: High-level overview of current techniques on bulk transfers

destinations, and apply unicast routing separately to these P2P transfers [8]. Meanwhile, several works consider multicast networks of SDN capabilities, where they deliver P2MP transfers over Steiner trees originating from the source site, flowing through or ending with the destination sites [9], [10].

Furthermore, in order to fully use scattered network resources, multiple routes, i.e., k-paths (Steiner trees), can also be introduced to deliver a single P2P (P2MP) transfer, referred to as *multipathing (multicast tree cohorts)*.

A popular SDN solution of routing implementation is OpenFlow using which highly sophisticated forwarding patterns can be programmed into network elements. For path-based unicasting, SWAN [2] identifies admissible tunnel paths by labeling, where VLAN IDs are used as labels to determine which tunnel is traversed by every data packet. For tree based multicasting, QuickCast [10] and DaRTree [9] use the group tables in OpenFlow switches to forward a copy of data to multiple outgoing ports.

2) Rate allocation: Allocating appropriate transmission rates on the available routes is also crucial to improve the efficiency of bulk transfers over Inter-DC WANs. Herein, transfer rates are flexible and adaptive to network conditions such as traffic load and topology.

Rate allocation typically exploits flexibility in dimensions of space and time. As network bandwidth differs from location to location, it is common practice to employ multiple routes for a single transfer and distribute traffic among these routes to fully use link bandwidth across different spatial locations. Some rate allocation solutions determine the sending rate on each forwarding route, while some plan the total rate on all admissible routes, as well as a splitting ratio among these routes. Because high-priority interactive traffic typically fluctuates temporally, the leftover bandwidth that can used by bulk transfers accordingly varies over time. Many approaches (e.g., [4], [5]) consider such temporal differences and propose to allocate time-varying transmission rates for bulk transfers. For example, most solutions use a time-slotted system and change flow rates across timeslots. The spatial-temporal dynamics of network resources and the properties (e.g., deadline) of bulk transfers are often combined to enable optimizations.

3) Delivery schemes: The storage capability of intermediate datacenters may also be exploited to assist bulk transfers, which results in two delivery schemes: *End to End (E2E)* and *Store and Forward (SnF)*. E2E delivery does not rely on storage capability of middle datacenters but requires concurrent availability of nodes and links on transmission routes. Many approaches [2]–[5], [8]–[10] adopt this scheme to avoid the extra storage cost at intermediate datacenters and highly complex network algorithms. With this delivery scheme, the providers only determine routes and flow rates per given route for each transfer. To this end, some solution uses a time expansion graph. We refer to Fig. 4a, where a time expansion graph is built for a P2P transfer from DC1 to DC4 across two timeslots of t_1 and t_2. This transfer uses spatially different path routes, marked using the orange lines in Fig. 4a.

In contrast, SnF delivery allows intermediate datacenters to temporarily store data before delivering it to the next one. Some solutions (e.g., [7]) apply SnF when some inter-DC links are temporally congested, forwarding data at a later time when these links are less congested. With SnF, the time expansion graph of the P2P transfer in Fig. 4a accordingly adapts to the introduction of storage: adding admissible links between the snapshots of the same datacenter at different times, as the blue dotted lines in Fig. 4b. As a result, this request from DC1 to DC4 can traverse a path with temporal storage at datacenter DC3, as in the green dotted line in Fig. 4b. Compared to E2E, SnF introduces storage as an additional dimension for optimization, but storing data at intermediate datacenters incurs additional costs. As the use of data storage, routing, and rate allocation must be carefully coordinated, SnF transfer allocations can also lead to more complex optimization formulations.

C. Scheduling Techniques

As many transfer requests typically coexist in a network, various scheduling techniques can be used to improve the overall performance by arranging how these transfers utilize the network resources. We briefly discuss the topics of policies, preemption, reservation, and rescheduling next.

1) Policies: Scheduling policies is to determine the processing order of coexisting transfer requests. Most solutions

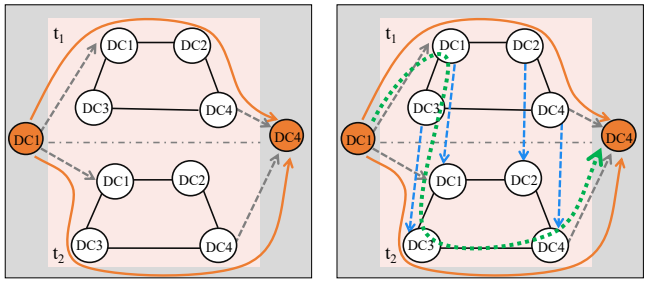

Fig. 4: Time expansion graph models of (a) End-to-End (E2E) and (b) Store-and-Forward (SnF) delivery schemes.

(e.g., [2], [4]) consider transfers with the same weight (priority) and schedule them according to a First Come First Served (FCFS) policy. FCFS processes transfer requests in the order of arrival, which is simple to implement. There are also a few attempts to use other scheduling policies, such as Shortest Remaining Processing Time (SRPT) and Max-Min Fairness (MMF), to improve transfer efficiency (i.e., minimizing completion time [10]). SRPT schedules transfers in an order of remaining completion time, which is optimal for minimizing mean completion times for P2P transfers over a single link. With P2MP transfers, in particular scenarios, MMF can beat SRPT in minimizing mean completion times by a large factor when many multicast trees share a link. In general, it is also possible to apply multiple scheduling policies in a hierarchical fashion to obtain sophisticated scheduling objectives. For example, one can classify transfers into two priorities of low and high and apply SRPT and FCFS across these two classes and within each class, respectively.

2) Preemption: Transfer requests can also be scheduled with support for *preemption* or without it, according to whether their resource allocations can be revoked. Specifically, non-preemptive scheduling does not allow interrupting or revoking transfer requests that are being delivered in the network, where the resources allocated to a transfer can only be released and reclaimed after completion. Some solutions, such as TEMPUS [5] and DaRTree [9], adopt non-preemptive scheduling for bulk transfers with deadlines. In contrast, preemption may temporarily interrupt an ongoing transfer without coordination, which usually occurs when competing for resources with higher priority traffic (e.g., interactive traffic). It can also be specifically designed for scheduling transfers with equal priority to improve efficiency. For example, to reduce the average transfer completion time, preemptive scheduling can be used to interrupt a large slow transfer to schedule more small but fast ones. Compared to the non-preemptive that only schedules new transfers, preemptive scheduling incurs additional complexity and cost, as it needs to maintain the status of existing and new transfers as a reference for strategies.

3) Reservation: Reserving bandwidth for transfer requests is an important technique to offer predictable and guaranteed performance services. For example, the works in [4], [8] guarantee completion before deadlines for accepted bulk transfers by reserving bandwidth into future timeslots. The reserved bandwidth of a transfer cannot be preempted by other transfers, but may vary over time, as long as the promised service quality (e.g., deadline) is ensured. As a result, a data transfer may be set to temporarily stop delivering data when no bandwidth reserved for it at that time—in contrast to non-preemptive scheduling, which does not allow transmissions to be interrupted before completion.

4) Rescheduling: The bulk transfer problems are often in an online setting where user requests are submitted to the system over time. The optimal allocation decision that was made earlier may become outdated and sub-optimal, due to new unforeseen transfer requests arriving at a later time. Rescheduling existing transfers together with new ones can often improve performance, but at the cost of rapidly increasing computational complexity, especially for the system that needs to plan resource allocations over a long-term time to guarantee transfer deadlines. To reduce such additional computational complexity, selective rescheduling can be performed to only include a few existing transfer requests in a joint rescheduling [4].

D. Other Techniques

1) Admission control: It is necessary to apply admission control to selectively admit a subset of transfer requests submitted to the network out of consideration of guaranteeing transfer deadlines or maximizing total profit. For example, in bulk transfers with a hard deadline, missing their deadline makes the data lose its utility. As a result, most deadline-aware solutions, e.g., [4], [8], [9], only accept transfers that can finish before deadline and analogously reject the transfers that will miss their deadlines. To this end, these solutions use non-preemptive scheduling and reserve the bandwidth to deliver admitted requests.

2) Topology control: With the emergence of reconfigurable network technologies, the topology of inter-DC WANs can be quickly reconfigured at the physical layer. A few solutions [9], [12] optimize the topology together with the transfer allocation, where they use cross-layer optimization algorithms to jointly design the optical topology and the rate allocation for bulk transfers. Although performance could increase noteworthily, cross-layer joint optimization inevitably increases problem complexity and computational cost.

3) Handling mispredictions and failures: Any transmission scheme has to deal with dynamic network conditions, especially mispredictions and failures. As bulk transfers can only use the leftover bandwidth of high-priority interactive traffic, the accurate estimation of interactive traffic is beneficial to efficient transfer allocations. Prior works [2], [5] show that interactive traffic is predictable in periodic patterns within a short time period (e.g., 5 minutes). However, misprediction is nonetheless inevitable. For example, Amoeba [4] observes that predication accuracy decreases with forecast length. Accordingly, to handle such misprediction, Amoeba sets aside different headrooms for different time slots proportional to temporal distance. In case of failure events, existing solutions (i.e., [4]) recalculate resource allocation decisions, but time still matters for bulk transfers.

IV. Open problems

In this section, we point out seven open problems that we consider of special interest for future research. We categorize these problems into optimizing inter-DC bulk transfers for better traffic prediction and measurement, improved scheduling policies and algorithms, and advanced network techniques.

We note that the challenges for intra-datacenter transfers are fundamentally and conceptually different, as *"70% of flows send less than 10 KB and last less than 10 seconds"* [15], and are moreover highly sensitive to delays. These characteristics are in stark contrast to the bulk and replication-heavy transfers between datacenters, which allow for some delay and last for minutes to days, and hence intra-datacenter traffic management relies heavily on e.g. oblivious route management.

1) *Handling uncertainty in traffic demand.* The dynamics of interactive traffic and bulk transfer traffic across datacenters need to be thoroughly considered to develop more efficient solutions. Although high-priority interactive traffic can be largely predicted over short time frames [2], [4], misprediction is inevitable and can lead to inaccurate leftover bandwidth estimates that may degrade the service quality of bulk transfers. Future research focuses on more accurate predictions of interactive traffic that in turn can be leveraged for more reliable solutions for bulk transfers. To facilitate accurate traffic prediction and WAN optimization, designing effective and fine-grained network telemetry systems can be helpful to gain further insights into the current network state (e.g., detecting changes in traffic patterns). In production datacenters, transfer requests typically arrive at the system in an online fashion and disclose their information after arrivals. Hence research on bulk traffic prediction would also be beneficial to designing advanced online solutions. Another open problem is to develop information-agnostic approaches that can efficiently allocate transfers without (nearly) complete knowledge of large transfer traffic or interactive traffic.

2) *Study of various scheduling policies.* Networking researchers have proposed many seminal but also new policies for scheduling intra-DC traffic. From a theoretical perspective, these policies can also be applicable to schedule inter-DC bulk transfers. However, it is not well understood if datacenter scheduling policies also work for such traffic well. So far, there is only one attempt ([10]) that investigates several scheduling policies and focuses on the average completion time of P2MP transfers. A broader and better understanding on how existing datacenter scheduling policies impact bulk transfers with different communication patterns, performance metrics, and objectives is an open problem. Compared to datacenter networks with a single bottleneck, an inter-DC network may have multiple bottlenecks. Any scheduling policy for bulk transfers should thus consider the more complex environments and unique transfer characteristics.

3) *Optimizing the network for a mix of transfer patterns, constraints, and objectives.* Datacenter networks typically host a large collection of different cloud services that generate a mixture of bulk transfers among datacenters with different communication patterns, performance metrics and objectives. A provider may need to handle both P2P and P2MP transfers, transfers desiring timely completion and and also a mixture of objectives such as maximizing the number of deadline-meeting requests and minimizing transmission cost. Such a variety of data transfer workloads introduces more complexity, and efficiently optimizing them is an open problem.

4) *Theoretical and mathematical analysis of inter-datacenter networks.* Most current works on bulk transfer optimization problems rely on greedy and heuristic algorithms. At the same time, very few of them investigate algorithms with theoretical guarantees and bounds. Theoretically analyzing the complexity and optimality scenarios of existing transfer algorithms is similarly not well-understood, except for simplified network models with single bottlenecks. However, multiple bottlenecks often occur in inter-DC WANs in practice. Another open problem is to design new algorithms with theoretical bounds for bulk transfer optimizations, especially in networks with multiple bottlenecks.

5) *Leveraging advanced hardware features to further optimize inter-datacenter networks.* Emerging technological innovations in reconfigurable network technologies offer the networking community new dimensions to optimize the explosively growing datacenter workloads. However, only few approaches utilize this technology to improve the performance of inter-DC bulk transfers, e.g,. by reducing the completion time of P2P transfers [12] and maximizing the number of deadline-meeting P2MP transfers [9]. Further exploiting the benefits of reconfigurable network technology to improve performance metrics and scheduling goals is an open issue.

6) *Designing a common framework to capture and simplify inter-datacenter traffic optimization.* Datacenter workloads are growing explosively [6] and this trend will continue in next-generation networks, driven by increasing internet penetration and emerging new applications such as Augmented and Virtual Reality and pervasive artificial intelligence. Next-generation networks will not only need to handle additional across-datacenter workloads, but also new complicated transfer problems with different communication patterns, performance metrics, and objectives. The conventional practice to handle new data transfer problems usually involves addressing new optimizations, which requires significant manual effort and expertise to express, non-trivial computation, and carefully crafted heuristics. In the future, it is necessary and an open problem to develop a general framework that can simplify the manual efforts in optimizing transfer allocation tasks.

7) *Learning-based inter-datacenter traffic optimization.* Given the rapid development of machine learning and its successful application to many complex network problems, we believe machine learning can also be applied to address the increasingly complex bulk transfer problems. For example, tailored machine learning algorithms can intelligently choose the right routing and scheduling policies and allocate cost-efficient transmission rate for mixed transfer workloads. However, corresponding research has been sparse and bridging this gap will be of high interest for future deployments. Moreover, to conduct such research, routing datasets from actual inter-DC networks need to be shared with researchers by operators.

V. Conclusions

The study and optimization of bulk data transfers over inter-DC WANs is a relatively new research area with many technical challenges that still need to be addressed. Inter-DC bulk transfers can lead to different and unique communication patterns, requiring various performance metrics and introducing novel optimization objectives. We presented a short review on existing transfer techniques for developing efficient transfer allocation solutions, focusing on transmission (e.g., routing, rate allocation and delivery scheme), scheduling (e.g., policies), and other techniques such as admission control and topology control. We concluded by discussing several open research questions and future challenges that need further investigation from the networking research community.

Acknowledgment. This work was partially supported by the National Key Research and Development Program of China under Grant 2019YFB1802803; the PCL Future Greater-Bay Area Network Facilities for Large-scale Experiments and Applications under Grant PCL2018KP001; the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme (AdjustNet: Self-Adjusting Networks) under Agreement 864228.

References

[1] "Hyperscale Data Center Count Passed the 500 Milestone in Q3," accessed on October 31, 2019. [Online]. Available: https://www.srgresearch.com/articles/hyperscale-data-center-count-passed-500-milestone-q3

[2] C.-Y. Hong, S. Kandula, R. Mahajan *et al.*, "Achieving High Utilization with Software-Driven WAN," in *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4. ACM, 2013, pp. 15–26.

[3] S. Jain, A. Kumar, S. Mandal *et al.*, "B4: Experience with a Globally Deployed Software Defined WAN," in *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4. ACM, 2013, pp. 3–14.

[4] H. Zhang, K. Chen, W. Bai *et al.*, "Guaranteeing Deadlines for Inter-Datacenter Transfers," *IEEE/ACM Transactions on Networking (TON)*, vol. 25, no. 1, pp. 579–595, 2017.

[5] S. Kandula, I. Menache, R. Schwartz, and S. R. Babbula, "Calendaring for Wide Area Networks," in *ACM SIGCOMM computer communication review*, vol. 44, no. 4. ACM, 2014, pp. 515–526.

[6] M. Jimenez and H. Kwok, "Building Express Backbone: Facebook's new long-haul network," accessed on October 29, 2019. [Online]. Available: https://engineering.fb.com/networking-traffic/building-express-backbone-facebook-s-new-long-haul-network/

[7] N. Laoutaris, M. Sirivianos, X. Yang, and P. Rodriguez, "Inter-Datacenter Bulk Transfers with NetStitcher," in *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4. ACM, 2011, pp. 74–85.

[8] L. Luo, Y. Kong, M. Noormohammadpour, Z. Ye, G. Sun, H. Yu, and B. Li, "Deadline-Aware Fast One-to-Many Bulk Transfers over Inter-Datacenter Networks," *IEEE Transactions on Cloud Computing*, 2019.

[9] L. Luo, K.-T. Foerster, S. Schmid, and H. Yu, "DaRTree: Deadline-Aware Multicast Transfers in Reconfigurable Wide-Area Networks," in *IEEE/ACM IWQoS*, 2019.

[10] M. Noormohammadpour, C. S. Raghavendra, S. Kandula, and S. Rao, "QuickCast: Fast and Efficient Inter-Datacenter Transfers using Forwarding Tree Cohorts," in *IEEE INFOCOM*, 2018, pp. 225–233.

[11] W. Li, X. Zhou, K. Li *et al.*, "TrafficShaper: Shaping Inter-Datacenter Traffic to Reduce the Transmission Cost," *IEEE/ACM Transactions on Networking*, vol. 26, no. 3, pp. 1193–1206, 2018.

[12] X. Jin, Y. Li, D. Wei *et al.*, "Optimizing Bulk Transfers with Software-Defined Optical WAN," in *SIGCOMM*. ACM, 2016, pp. 87–100.

[13] Y. Zhang, J. Jiang, K. Xu, X. Nie, M. J. Reed, H. Wang, G. Yao, M. Zhang, and K. Chen, "BDS: A Centralized Near-Optimal Overlay Network for Inter-Datacenter Data Replication," in *EuroSys*, 2018, pp. 1–14.

[14] H. Yu, H. Qi, and K. Li, "WECAN: an Efficient West-East Control Associated Network for Large-Scale SDN Systems," *MONET*, vol. 25, no. 1, pp. 114–124, 2020.

[15] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. 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.

Biographies

Long Luo is currently working toward the PhD degree from the University of Electronic Science and Technology of China (UESTC). She received the BS degree in communication engineering from Xi'an University of Technology in July 2012 and MS degree in communication engineering from the UESTC in July 2015. Her research interests include software-defined networks, inter-datacenter traffic engineering and data-driven networking.

Hongfang Yu is a Professor at University of Electronic Science and Technology of China. She received the BS degree in electrical engineering in 1996 from Xidian University, and the MS and PhD degrees in communication and information engineering in 1999 and 2006, respectively, from the University of Electronic Science and Technology of China. From 2009 to 2010, she was a visiting scholar at the Department of Computer Science and Engineering, University at Buffalo (SUNY). Her research interests include SDN/NFV, data center network, network for AI system and network security.

Klaus-Tycho Foerster is a Postdoctoral Researcher at the Faculty of Computer Science at the University of Vienna, Austria since 2018. He received his Diplomas in Mathematics (2007) & Computer Science (2011) from Braunschweig University of Technology, Germany, and his PhD degree (2016) from ETH Zurich, Switzerland, advised by Roger Wattenhofer. He spent autumn 2016 as a Visiting Researcher at Microsoft Research Redmond with Ratul Mahajan, joining Aalborg University, Denmark as a Postdoctoral Researcher with Stefan Schmid in 2017. His research interests revolve around algorithms and complexity in the areas of networking and distributed computing.

Mohammad Noormohammadpour received the Ph.D. degree from the Ming Hsieh Department of Electrical Engineering, University of Southern California, USA. He is currently a Research Scientist with the Network Planning and Risk team at Facebook, Menlo Park, USA. His research interests include Network Risk Analysis and Capacity Planning, as well as Distributed Systems Optimization and Scaling in general.

Stefan Schmid is a Professor at the Faculty of Computer Science at the University of Vienna, Austria. He received his MSc (2004) and PhD degrees (2008) from ETH Zurich, Switzerland. In 2009, Stefan Schmid was a postdoc at TU Munich and the University of Paderborn, between 2009 and 2015, a senior research scientist at the T-Labs in Berlin, Germany, and from the end of 2015 till early 2018, an Associate Professor at Aalborg University, Denmark. His research interests revolve around fundamental and algorithmic problems in networked and distributed systems.