

# Chopin: Combining Distributed and Centralized Schedulers for Self-Adjusting Datacenter Networks

Neta Rozen-Schiff  

School of computer science and engineering, Hebrew University, Israel

Klaus-Tycho Foerster  

Computer Science Department, Technical University of Dortmund, Germany

Stefan Schmid  

TU Berlin, Germany

Faculty of Computer Science, University of Vienna, Austria

David Hay  

School of computer science and engineering, Hebrew University, Israel

---

## Abstract

The performance of distributed and data-centric applications often critically depends on the interconnecting network. Emerging reconfigurable datacenter networks (RDCNs) are a particularly innovative approach to improve datacenter throughput. Relying on a dynamic optical topology which can be adjusted towards the workload in a demand-aware manner, RDCNs allow to exploit temporal and spatial locality in the communication pattern, and to provide topological shortcuts for frequently communicating racks. The key challenge, however, concerns how to realize demand-awareness in RDCNs in a scalable fashion.

This paper presents and evaluates *Chopin*, a hybrid scheduler for self-adjusting networks that provides demand-awareness at low overhead, by combining centralized and distributed approaches. *Chopin* allocates optical circuits to elephant flows, through its slower centralized scheduler, utilizing global information. *Chopin*'s distributed scheduler is orders of magnitude faster and can swiftly react to changes in the traffic and adjust the optical circuits accordingly, by using only local information and running at each rack separately.

**2012 ACM Subject Classification** Networks → Programmable networks; Networks → Data center networks

**Keywords and phrases** reconfigurable optical networks, centralized scheduler, distributed scheduler

**Digital Object Identifier** 10.4230/LIPIcs.OPODIS.2022.4

**Supplementary Material** *Software (Code)*: [https://bitbucket.org/NetaRS/sched\\_analytics](https://bitbucket.org/NetaRS/sched_analytics)

**Funding** Research supported by the European Research Council (ERC), grant agreement No. 864228 (AdjustNet), Horizon 2020, 2020-2025, a grant from Fraunhofer SIT, and the Israeli Innovation Authority through the Peta-Cloud consortium.

## 1 Introduction

Data-centric and distributed applications, including batch processing, streaming, scale-out databases, or distributed machine learning, generate a significant amount of network traffic and their performance critically depends on the throughput of the underlying network [31, 85].

To improve datacenter throughput, researchers and industry, e.g., Google [64], have recently started exploring innovative new datacenter designs that rely on dynamic and *demand-aware* topologies: topologies that self-adjust toward the workload they currently serve. The motivation behind self-adjusting datacenter topologies is twofold.

*First*, empirical studies reveal that datacenter traffic patterns feature much structure [6, 10, 31, 66], i.e., are sparse, skewed, and bursty, which introduces optimization opportunities.



© N. Rozen-Schiff, K.-T. Foerster, S. Schmid, and D. Hay;  
licensed under Creative Commons License CC-BY 4.0

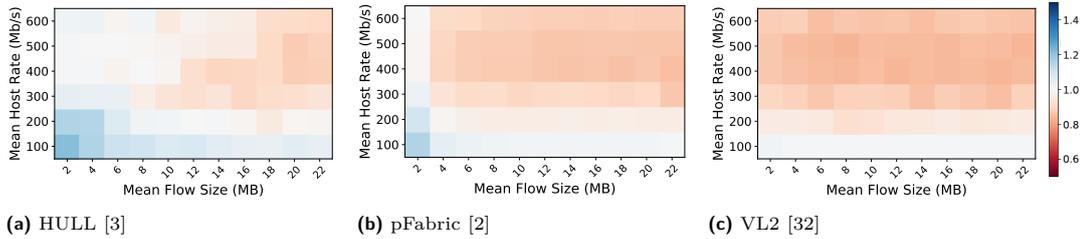
26th International Conference on Principles of Distributed Systems (OPODIS 2022).

Editors: Eshcar Hillel, Roberto Palmieri, and Etienne Rivière; Article No. 4; pp. 4:1–4:22



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



**Figure 1** Comparison between centralized and distributed schedulers under different traffic patterns (each generated by scaling well-known realistic flow size distributions, assuming Poisson flow arrival times under different rates), when the number of ToR switches is 80, and the optical connectivity of each ToR switch is 4. The color represents the ratio between the optical throughput of the distributed scheduler and the centralized scheduler. Blue cells mark settings where the distributed scheduler outperforms the centralized one; red cells mark the opposite. We refer to §6.1 for topology details.

For example, a small number of flows typically carry the majority of traffic (these are called *elephant flows*), while the remainder consists of a large number of flows that carry very little traffic (*mice flows*). Therefore, in a demand-aware network, the elephant flows should be routed through the optical circuits for offloading the electrical bottleneck, which in turn, reduces the latency of the mice flows, and improves the overall throughput.

*Second*, emerging optical technologies and optical circuit switches enable the required very fast reconfigurations [8, 21, 22, 33]. Over the last years, several interesting hybrid optical datacenter networks were suggested and evaluated [83], augmenting an oversubscribed network with inter-rack optical links [7, 14, 22, 24, 25, 31, 53, 69, 70, 77, 79], see [27] for a survey. The number of optical routes from/to each Top-of-Rack (ToR) switch, which we call the *ToR switch optical degree*, is a single-digit number, typically at most 4 [21, 71, 82].

**Challenge: Scalability.** While the vision of self-adjusting networks is intriguing and early solutions show promising results, the main challenge faced by such demand-aware networks concerns the scalability of the control plane. Unlike demand-oblivious networks (i.e., static networks like Clos [15], Slim Fly [12], and Xpander [76] or dynamic networks like RotorNet [59], Opera [58] and Sirius [8]), demand-aware networks require the collection and evaluation of traffic patterns. In particular, performing all topology scheduling decisions *centrally* (i.e., a centralized scheduler) may introduce a bottleneck and can result in slow reaction times. A fully distributed decision making (i.e., a distributed scheduler) on the other hand, may be suboptimal as it is based on incomplete information.

In order to show this tradeoff, we analyzed the *optical throughput ratio*. The optical throughput ratio is defined as the ratio between the throughput routed through the optical circuit and the total datacenter throughput. It is a cornerstone measure as it reflects the utilization of the optical circuit, and therefore reduces the bottleneck over the electrical network. Fig. 1 compares the optical throughput ratio of the distributed-only scheduler and the centralized-only scheduler under different traffic patterns (each traffic pattern follows a distribution measured in a real datacenter, where we have parametrized the mean flow size and each host rate). It demonstrates the tension between the two approaches: there is no “clear winner” and which one is better depends on the traffic pattern. The traffic pattern however is often not known when the datacenter is built and changes over time. For example, consider a datacenter serving a pFabric traffic pattern, with typical mean flow size of approximately 1.7 MB [51], and where each host sending rate is approximately 100 Mbps. In this case, the optical circuit throughput ratio in the distributed scheduler is by 13% higher

compared to the centralized scheduler, as can be seen in the blue cells in Figure 1b. However, for the same datacenter, with the same traffic pattern (pFabric), and the same mean flow size distribution, once the host’s sending rate grows beyond 300 Mbps, the centralized scheduler achieves a higher throughput ratio compared to the distributed one (see the relevant red cell).

Motivated by this insight, and by the desire to provide an efficient control plane for self-adjusting networks, we propose to combine both approaches to achieve the best of both worlds: fast reaction times of distributed decision-making and network utilization benefits of centralized optimization.

**Introducing Chopin.** We present *Chopin*<sup>1</sup>, a novel scheduler for reconfigurable datacenter networks that fully exploits the benefits of self-adjusting networks by relying on an efficient control plane. Specifically, *Chopin* provides demand-awareness at low overhead, by combining centralized and distributed approaches. At the heart of Chopin’s approach lies the idea that a relatively complex algorithm (e.g., Maximum Weight Matching, MWM) should be computed centrally, based on complete information.

However, since such an algorithm cannot be computed fast [11, 16, 48] (e.g., MWM may take around 20 ms for 80 ToR switches), we additionally allow distributed *updates* to the centralized optical circuit allocation, based on a *threshold*. The threshold specifies the flow weight changes from which a distributed scheduler can update the centralized scheduler allocation. For example, if there is a large drop in demand in an allocated optical circuit (e.g., when an elephant flow ends), the distributed scheduler may tear it down and try to establish another circuit. Hence, due to the volatility of many flows, we want a distributed constant-round algorithm (ideally just two rounds) and hence forgo more complex distributed algorithms [9] or dynamic centralized algorithms [13]; the indirection via a centralized controller comes with overheads and delays which render this approach problematic to handle continuous update streams.

**Our Contributions.** In summary, we make the following contributions:

1. We identify and analyze the difference in throughput performance of centralized and distributed schedulers for reconfigurable datacenter networks, for various scenarios and different flow size distributions.
2. We design a hybrid scheduler, *Chopin*, which combines centralized and distributed decision-making based on thresholds. To this end, we present and analyze both a centralized and a local online scheduler, exploring the trade-off between accuracy and running time. *Chopin* relies on commodity devices available today, and required Chopin nodes which can simply be added to existing ToR switches by directing one of the switch ports to them. Moreover, information collection and dissemination of the centralized algorithm can be realized in the control plane using Software-Defined Networks (SDNs).
3. We report on Chopin’s effectiveness through extensive simulations for different settings, showing that Chopin improves upon centralized and distributed approaches. We achieve throughput improvements of up to 20% against centralized and up to 23% against distributed schedulers, *always outperforming both*.

## 2 Optical Background and Related Work

Chopin is motivated by trade-offs between centralized and distributed scheduling, which arise in matching algorithms. We first motivate why matching algorithms are central to Chopin’s setting and then discuss centralized and distributed schedulers in this context.

---

<sup>1</sup> Stands for: Controller for Hybrid OPtical electrical Networks.

**Optical Model: Why Matchings?** From a theoretical viewpoint, we consider the problem of how to augment a static network with (optical) edges in order to improve the total network performance. The reason why this augmentation comes in the form of matchings lies in the underlying hardware, namely optical circuit switches, we refer to Hall et al. [36, §3] for a technological overview. In the simplest case, a set of nodes is connected to the optical circuit switch’s ports by an optical cable each, and the switch “matches” these ports by e.g. adjusting mirrors to steer the light signals s.t. that pairs of ports (and hereby, pairs of nodes) are hence connected by optical circuits. Nodes could also be connected multiple times to the optical switch, or multiple optical switches could be used, giving rise to, e.g.,  $b$ -matchings [26, 39].<sup>2</sup> Conceptionally, other hardware could be used to the same effect (e.g., beamformed wireless connections [37] or free-space optics [7]), but on a graph-theoretic level, they form circuits between pairs of nodes, and as thus, matchings. We refer here to the survey by Foerster and Schmid [27] for a further introduction to the enablers, algorithms, and complexity of reconfigurable datacenter networks. We moreover refer to the article by Zerwas et al. [84] on how system delays can be accounted for for scheduling algorithms.

**Centralized schedulers** operate under the assumption of near-perfect utilization visibility and traffic demands, collected at a centralized location [18], often leveraging SDN. We refer to a recent survey and the references therein [74]. Herein the restriction to large and long-lived flows enables centralized schedulers [24, 79] to also cope with control loop delays. However, these schedulers still suffer from traffic stability assumptions [23].<sup>3</sup> Traffic matrix schedulers [22, 54, 77, 78] on the other hand, adjust packet transmissions to coincide with scheduled circuit reconfiguration, with full knowledge of when bandwidth will be available to particular destinations. However, for the duration of the matching schedule, new flows are not accounted for and might need to wait for the next iteration. In contrast, Chopin’s design ensures rapid reactions to local traffic changes and new flow insertions, due to its additional distributed scheduler part.

**Distributed schedulers.** In practice, the large number of scheduling decisions and status reports can overwhelm centralized schedulers, and in turn lead to long latencies before scheduling decisions are made [18]. ProjecToR [31] initiated a broader interest in distributed scheduling, by proposing a stable-matching algorithm that optimizes for low latency, utilizing high fan-out single hop free-space optics [30]. Via aging of requests, they obtain a constant-factor latency approximation for their online scheduling algorithm [19]. RotorNet [59], Opera [58], and Sirius [8] employ a different approach and use lower fan-out circuits, where the topologies are created in a demand-oblivious manner. RotorNet rotates through matchings independent of the current traffic, that provide eventual connectivity, where traffic is either scheduled to be routed along single hops, or along two hops, via buffering and a proposal and accept mechanism. Sirius follows similar ideas, either transmitting directly or via schemes reminiscent of Valiant’s method. Opera extends RotorNet by also always maintaining an expander graph, motivated by static topologies [46, 76]. Although Opera’s reconfiguration scheduling is deterministic, the precomputation of the topology layouts is in its current form still randomized. Notwithstanding, ProjecToR, RotorNet, Sirius, and Opera can all rapidly deploy traffic along reconfigurable connections, by omitting a centralized control plane. However, it is not clear how to realize the above three distributed systems with

<sup>2</sup> There is also some work that considers multicast by splitting the outgoing light signals [17, 57, 72].

<sup>3</sup> Orthogonal to matching algorithms, Xia et al. [80] investigate how to migrate between Clos and random graph topologies. However, they require specialized 4/6-port converter switches and also rely on a centralized control loop, estimating an update delay “*on the order of seconds*” [80].

off-the-shelf hardware, such as a common optical circuit switch, and hence their application scenario is not as general as with Chopin. Notwithstanding, Decentralized scheduling is also used in several other systems, including SplayNet [68], Cerberus [35], or CacheNet [34].

Lastly, while there is profound research on matching algorithms in the distributed computing community [73], distributed algorithms for maximal matchings in graphs with large degree  $\Delta$  (as for optical circuit switches) are relatively slow [9]. While approximation [55] and dynamic [56] algorithms are considerably faster, here the constraints of the optical datacenter networking and the distributed computing community are quite different and hence the communities (yet) don't overlap much in their research applications: ideally, for optical circuit switching, small-constant round algorithms of low computational complexity are desired, whereas in the distributed computing community, the local algorithms can be more complex, with a focus on asymptotic runtime optimization. As thus, Chopin utilizes a low complexity threshold based distributed algorithm, using just two rounds of communication, which falls in line with the requirements of hybrid datacenters.

### 3 Chopin's Design

In a nutshell, *Chopin's* topology scheduler aims to provide demand-awareness efficiently by combining centrally optimized decision making with fast distributed reactions. The idea is hence analogous to the nervous system of animals, which is typically divided into a slower central nervous system and a faster peripheral nervous system [75].

Specifically, *Chopin's* scheduler uses two different control mechanisms, each carried out in a different location in the datacenter, providing different latency and response times. The **centralized scheduler** is reminiscent of an SDN controller and allows *Chopin* to adapt to global changes (such as traffic rates). This optimization uses traffic measurements across the network and has a (relatively) long response time. Moreover, it may receive additional information (e.g., from applications that have specific repetitive patterns) to make even better decisions. Fig. 2 presents the connectivity between the SDN controller to each of the ToR switches and Chopin's nodes. The **distributed scheduler** is embedded within the ToR switches and is based only on local measurements. It reacts quickly to local changes in traffic and may tear down connections if they become unmatched and establish new connections for new "hot" ToR switch pairs. The tear down and connection establishment are made by updates sent from the ToR switch to its Chopin node, see Fig. 2.

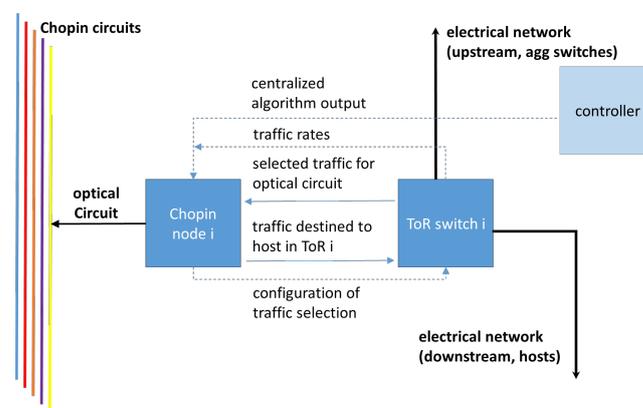


Figure 2 Chopin's design.

The centralized scheduler and the distributed scheduler are discussed in details in Section 4 and Section 5 respectively.

Moreover, by combining these two schedulers, we can strike an optimized trade-off and realize both fast reactions and global and long-term network optimizations, accounting for demand uncertainty. In particular, unlike many existing solutions, which consider only one scheduler, Chopin is flexible and performs better than both.

### 3.1 The Hybrid<sup>4</sup> Topology

Chopin can be used together with any fast switching circuit technology (as in [22, 23, 62]), and implemented within the existing datacenter hardware. We distinguish between two entities in the ToR switches: the electric switch itself (for brevity, we will simply refer to this switch as the ToR switch), and the Chopin node which resides in the switch, serving as the entry point to the optical network. This modular Chopin structure enables us to support existing ToR switches, by directing one of its upstream ports to the Chopin nodes. When clear from the context, we use the terms, ToR switch and Chopin node, interchangeably.

The optical network can be any non-blocking topology, where the only constraint on establishing a circuit between two ToR switches is the availability of a transceiver in the corresponding Chopin node (namely, its optical degree).

Specifically, we assume each Chopin node has an optical degree of  $k$  and optical circuits are symmetric. This implies, that at any given time, a Chopin node can send and receive data *from at most  $k$  Chopin nodes*. For any given time  $t$ , we denote by  $dest_i(t)$  the set of Chopin nodes connected to the Chopin node  $i$ . We observe that as circuits are symmetric, if  $j \in dest_i(t)$  then it also holds that  $i \in dest_j(t)$ .

### 3.2 Problem Formulation

At the heart of Chopin lies the desire to improve network performance and throughput by avoiding scheduling bottlenecks. As Chopin is deployed between ToR switches, the scheduler is oblivious to intra-rack traffic or delays.

We first need to introduce some preliminaries. Let  $n$  be the number of ToR switches in the network and assume that time is slotted, where in each time-slot the distributed scheduler can be invoked (e.g., the length of each time-slot is 1 ms). Let  $X_{i,j}(t)$  be the total amount of traffic sent from rack  $i$  to rack  $j$  at time-slot  $t$ . Now let  $Y_{i,j}(t)$  be an indicator variable to describe whether a pair of ToR switches is connected through a Chopin circuit at time interval  $t$ :  $Y_{i,j}(t) = 1$  if and only if  $j \in dest_i(t)$  (and 0 otherwise). If  $Y_{i,j}(t) = 1$  then  $Y_{j,i}(t) = 1$  as connections through Chopin are symmetric.

Let  $C_t \subseteq S \times S$  be a symmetric relation with all ToR switch pairs that are connected through a Chopin circuit at time interval  $t$  (i.e.,  $(i, j) \in C(t)$  if and only if  $j \in dest_i(t)$ ).

We aim to maximize optical circuit throughput, a standard objective in such topologies [7, 14, 22, 24, 54, 59, 78, 79], namely  $\sum_t \sum_i \sum_j X_{i,j}(t) \cdot Y_{i,j}(t)$ . This relieves the electrically switched network part and reduces the overall latency. This is done by updating the set  $C_t$ , based on local and centralized decisions.

Note that, as optical circuit capacities are typically very high, we assume that the capacity of an optical circuit is always larger than the total amount of traffic sent between two racks (namely  $X_{i,j}(t)$ ). In case this does not hold, and the two racks are connected through an

---

<sup>4</sup> We note that the term *hybrid* can have a different meaning in some networking contexts, e.g., indicating a combination of the LOCAL model with the node-capacitated clique model [5].

optical link, one can send traffic through the optical circuit up to its capacity while the remaining traffic is sent through the electrically switched network.

### 3.3 Schedulers and Definitions

First, a *centralized* scheduler has a global view of the network and, in some cases, even auxiliary information given by the network administrator. This, on one hand, enables the scheduler to perform more informed decisions. But on the other hand, when using a centralized scheduler, it can take much longer to gather, compute, and spread the information across the datacenter. In our model, we assume the centralized scheduler works every  $T$  time-slots (which we call the *centralized scheduler epoch*) and uses slightly outdated traffic information: at time  $t$ , only the measurements  $\{X_{i,j}(t') | t' < t - \Delta, \text{ for every } i, j\}$  can be used, where  $\Delta$  is the *centralized algorithm delay*: the time it takes it to gather all information and make decisions. For example, if the optical degree is 1 (i.e.  $k = 1$ ), the centralized scheduler may use algorithms such as maximum weight matching to optimize the throughput that goes through the optical circuits.

As  $T$  becomes larger, centralized scheduler decisions can deteriorate, as the input on which decisions are based is outdated toward the end of the epoch. Thus, we additionally consider a distributed scheduler that is more fine-grained and runs every time-slot, benefiting from a reduced computation time and avoiding the delays involved in the centralized scheduler; it changes the pairs of connected switches based on local information only and by exchanging messages between ToR switches in two rounds. Specifically, the distributed scheduler of node  $i$  at time-slot  $t$  may use traffic measurements on its node until the computation starts:

$$\{X_{i,j}(t') | j \neq i, t' < t - \delta\} \cup \{X_{j,i}(t') | j \neq i, t' < t - \delta\},$$

where  $\delta < \Delta$  is the distributed scheduler delay. In addition, the distributed scheduler is aware of the information sent to it by other nodes throughout the rounds of computation. Importantly, for each pair  $(i, j)$  that was optically connected through Chopin at time interval  $t - 1$  (namely,  $Y_{i,j}(t - 1) = 1$ ), the distributed scheduler at node  $i$  knows what information was used to establish this connection (e.g., what is the rate reported to the centralized scheduler upon its last invocation) and decides whether the information is stale or not. Table 1 summarizes Chopin schedulers' notations.

Notation	Meaning
$n$	The number of ToR switches.
$dest_i(t)$	The set of racks optically connected to rack $i$ at time-slot $t$
$X_{i,j}(t)$	The total amount of traffic sent from rack $i$ to rack $j$ at time slot $t$
$Y_{i,j}(t)$	Indicator variable. $Y_{i,j}(t) = 1$ iff $j \in dest_i(t)$
$C_t$	The set of rack pairs optically connected at time slot $t$
$K$	Optical degree, the number of available circuits per Chopin node.
$\Delta$	Centralized scheduler delay
$\delta$	Distributed scheduler delay
$\alpha$	Chopin threshold for keeping centralized decisions
$A$	Centralized scheduler aggregation interval
$a$	Distributed scheduler aggregation interval
$T$	Centralized scheduler epoch

■ **Table 1** Chopin's schedulers' notations

## 4 Chopin's Centralized Scheduler

The centralized scheduler is implemented on top of a centralized SDN controller, which is (logically) connected to each of the ToR switches and the Chopin nodes. Upon a request

from the centralized scheduler, the controller collects traffic measurements across the network (namely, counters at ToR switches, current status of Chopin nodes). Based on these measurements, it computes the next optical circuit allocation.

Recall that the *delay*  $\Delta$  is the time it takes to send all the information to the controller, run the centralized algorithm, and send the decisions back to the nodes. The centralized algorithm works in epochs of length  $T$ , where decisions arrive at the nodes at the beginning of each epoch. Assume an epoch starts at time  $t$ . Then, these decisions will be used by nodes until time  $t + T$  (or until altered locally by the distributed scheduler). Furthermore, these decisions are based on information gathered in the interval  $[0, t - \Delta]$ . However, if traffic changes quickly (DC traffic is often bursty [6]), this information may be outdated quickly. This motivates us to define an *aggregation interval*  $A$  for the centralized algorithm, considering only the interval  $[t - (\Delta + A), t - \Delta]$ , see Fig. 3.

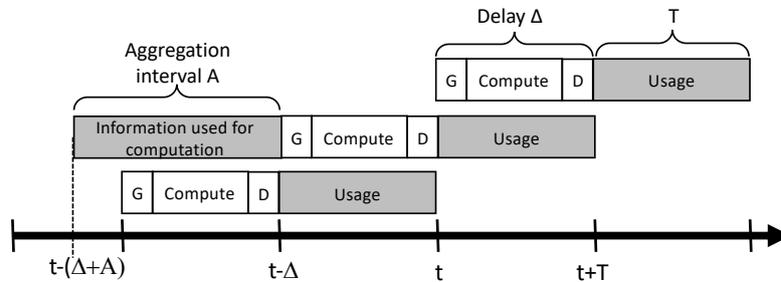
Notice that the delay  $\Delta$  is an important factor for the performance of Chopin. The delay describes the response time of the central scheduler and consists of several steps: contacting tens to hundreds of nodes [20, 44], (2) receiving thousands of flow entry statistics, estimating optical circuit utilization, contacting all nodes again, and updating all rules with new parameters if needed.

Considering common SDN controllers’ capability to handle a few thousands of messages per second [81], we estimate the delay to be in the order of hundreds of milliseconds in most configurations [49] [40]. Furthermore, the computation time of our algorithms can be in the order of tens of milliseconds for hundreds of ToR switches (e.g., when running maximum weight matching–like algorithms, as reported in [24]). Due to these delays, fast changes in the network (occurring within a few milliseconds [66]), may not be detected by the centralized scheduler in a timely manner. Also, the reconfiguration time (approximately 11  $\mu$ sec [31] [63]) is likely negligible compared to a centralized reconfiguration cycle. These observations motivate usage of another scheduling layer, to adapt to traffic in an online manner.

Our high-level goal is to maximize the overall throughput over the optical network. First recall that allocations are constrained by the optical ToR switch degree  $k$ : each ToR switch can be optically connected to at most  $k$  other ToR switches, Accordingly, our centralized algorithm essentially needs to solve a weighted b-matching problem, with  $b = k$ . Specifically, we consider an undirected graph whose nodes are the ToR switches and the weight of each edge  $(i, j)$  is the total traffic between  $i$  and  $j$  in the relevant interval:

$$w_{ij} = \sum_{t'=t-(\Delta+A)}^{t-\Delta} X_{i,j}(t') + X_{j,i}(t').$$

While b-matching algorithms are strongly polynomial [4], their running time can still be prohibitively high in practice [28, 50, 61]. This can lead to high delays  $\Delta$  and in turn, to a



■ **Figure 3** Centralized scheduler timing parameters, where “G” stands for the gathering period and “D” for the disseminate period. Similar parameters are used by the *distributed scheduler*, with delay of  $\delta$ .

significantly reduced overall performance of the system. Thus, we propose to approximate the problem: we compute a maximum weight matching (using Edmond’s MWM algorithm [29]), subtract the weights of the matching’s edges from the graph, and run maximum weight matching again with the new, smaller weights. As in [70], this process is repeated  $k$  times, resulting in  $k$  matchings. Hence each node is connected to at most  $k$  other nodes, as required. We refer to Khan et al. [47] for a further discussion on efficiently approximating b-matchings.

We further reduce computation time by considering only the top- $m$  live flows per ToR switch (instead of all possible pairs between the nodes). Due to the sparse nature of datacenter traffic matrices, even small values of  $m$  provide a highly accurate approximation: there is almost no performance degradation compared to a full-fledged MWM (§5). Note that the top- $m$  flows per switch can be efficiently calculated in each switch since there are only  $n$  possible flows and maintaining  $n$  counters at line rate is supported by switches.

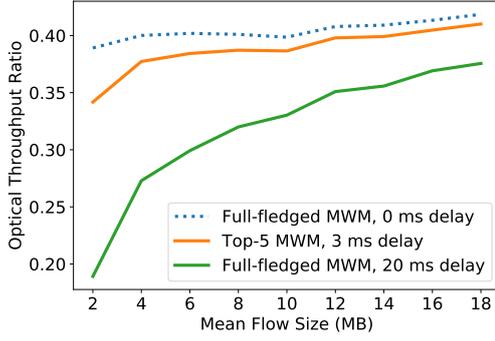
Moreover, focusing only on a constant number of top flows per node enables Chopin to scale with an increasing number of nodes. It also decreases both the MWM computation and the network reconfiguration times, allowing more frequent centralized scheduler reconfigurations. As the frequency of centralized scheduler invocations significantly affects the performance, by considering only  $m$  flow, we can improve the scheduler’s performance. For example, when  $m = 5$  and the number of ToR switches is 80, the time it takes to compute MWM based on top-5 live flows per ToR is 3 ms, while full-fledged MWM takes at least 20 ms. Fig. 4 compares the performance of both algorithms under the pFabric traffic pattern we have described (similar results hold for other traffic patterns as well) and shows that having more frequent reconfigurations is more significant than having slightly better matchings. Our centralized scheduler, based on top-5 live flows with reconfiguration every 3 ms, achieves almost the same results as an idealized online optimal algorithm, that computes full-fledged MWM every 1 ms. Finally, we observe that the optical throughput ratio improves as the mean flow size increases (and the gap between the algorithms shrinks), since longer flows imply that flow information is still relevant even after a long time when computations are infrequent.

In order to explain the throughput differences in Fig. 4, we analyzed the number of reconfigurations in each scenario. We consider the average number of reconfigured pairs in each run, out of the total number of pairs ( $n/2$ ). Fig. 5 presents the reconfiguration average ratio per 1 ms, as a function of the mean flow size. As expected, as the mean flow size increases (and the flows are longer), the number of reconfigurations decreases. Moreover, it shows that the number of reconfigurations is decreasing rapidly when the epoch time is 20 ms. This can be attributed to the fact that short lived connections have less impact on the 20ms long measurements and are less likely to be matched.

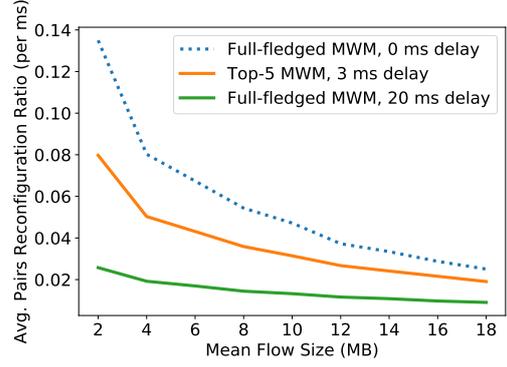
## 5 Chopin’s Distributed Scheduler

The distributed scheduler is a distributed control algorithm, embedded inside each Chopin node. Each ToR switch is connected to a single Chopin node and sends flows to the latter (e.g., by connecting one of its ports to the Chopin node). The traffic that is sent through this port is configured either by our centralized algorithm (as described in Section 4) or by the distributed scheduler that runs on the Chopin node.

The Chopin node is responsible of sending traffic destined for the ToR switch from one of the optical circuits. Each ToR switch in turn is connected to a single Chopin node. We refer to the illustration in Fig. 2 for an overview. Supplemental pseudo-code of Chopin’s distributed algorithm appears in Algorithm 1 in the Appendix.



■ **Figure 4** Comparison between a centralized scheduler which operates every 3 ms and computes the MWM of the top-5 live flows, to a centralized scheduler which operates every 20 ms and computes full-fledged MWM. For brevity only pFabric results are shown.



■ **Figure 5** Average reconfiguration ratio per 1 ms of three scenarios: centralized scheduler which computes MWM over top-5 live flows every 3 ms, a centralized scheduler which operates every 20 ms and computes full-pledged MWM. For brevity only results for the pFabric traffic pattern are shown.

At the beginning of each centralized scheduler epoch, every Chopin node keeps track of the traffic rate according to which its circuit was selected. Namely, for an epoch that starts at time  $t$ , if a circuit was established between Node  $i$  and Node  $j$ , both Node  $i$  and Node  $j$  compute and store:

$$R_{i,j}(t) = \frac{1}{A} \sum_{t'=t-(\Delta+A)}^{t-\Delta} X_{i,j}(t') + X_{j,i}(t').$$

The nodes use these rates to determine if traffic demands stay steady during the epoch. Specifically, we define a *threshold*  $\alpha \geq 0$ , and compute the rate in each time-slot  $\hat{t} \in [t, t+T]$  based only on local information available at the nodes:

$$r_{i,j}(\hat{t}) = \frac{1}{a} \sum_{t'=\hat{t}-(\delta+a)}^{\hat{t}-\delta} X_{i,j}(t') + X_{j,i}(t'),$$

where  $\delta$  is the distributed scheduler delay, and  $a$  is its aggregation interval. Only if  $r_{i,j}(\hat{t}) > \alpha \cdot R_{i,j}(t)$ , then the circuit is marked as matched and the algorithm keeps it connected through this epoch. Otherwise, it strives to replace it with a better connection, as described next. We observe that for  $\alpha = 0$ , all existing connections are kept matched (namely, Chopin just runs the centralized scheduler, and may improve only when its computed b-matching changes). As the threshold increases, it enables the distributed scheduler to tear down almost every centrally-computed connection and to create new ones, based on the current ToR switch traffic. The distributed scheduler itself tries to establish as many circuits as possible to increase the overall traffic through the optical circuit connections. In a nutshell, each Chopin Node  $i$  sends requests to a predetermined number of other nodes needs (this number is denoted by variable `max_reqs`), for which it observes the most bi-directional traffic. These nodes, denoted by `req_nodes`, do not include those kept matched to Node  $i$ ; moreover, `max_reqs`  $> k$  to allow utilizing all the circuits connected to Node  $i$ . After a Chopin node  $i$  sends its requests, it waits to receive requests from other nodes. We distinguish between:

1. Request from a node  $j$  that is in the `req_nodes` set: This means that both nodes  $i$  and  $j$  consider the traffic between them in their top `max_reqs` links. This makes Node  $j$  a candidate for a match with Node  $i$ .

2. Request from a node  $j$  that is not in the `req_nodes` set: This means that while Node  $j$  considers Node  $i$  in its top `max_reqs` links, Node  $i$  has `max_reqs` other links with larger traffic. This request should be denied.

We wait until all requests are received at Node  $i$ : This is indicated by a time-out event, that can be set, for example to half the aggregation interval  $a$  (requests are timestamped, so requests that arrive after the time-out will simply be ignored.). After all requests are received, there will be at most `max_reqs` candidates for matching. However, the number of free circuits (the optical degree minus the number of matched circuit) may be smaller. Therefore, we choose only the top ones so as not to exceed the number of available links. We thus send a `grant` message to all of them and `deny` messages to others.

In the last phase of our algorithm, each node waits until all its requests are either granted or denied. It then connects with all nodes that (i) *it has granted*, and (ii) *a grant message was received from them*. It disconnects all other links, except those made by the centralized algorithm and above the threshold. Note that rate measurements used in an epoch are performed in parallel with the decision making of the previous epoch.

## 6 Evaluation

Chopin aims to maximize the circuit throughput (online), without compromising the datacenter latency, by combining centralized and distributed schedulers. Therefore, in our evaluation, we focus on each of these schedulers' parameters as well as on their contribution to the overall DC performance. The performance is evaluated on several parameters, including the centralized scheduler epoch  $T$ , aggregation intervals  $A$  (for the centralized scheduler) and  $a$  (for the distributed scheduler), as well as the corresponding delays  $\Delta$  and  $\delta$ .

### 6.1 Methodology

**Topology.** We have analyzed Chopin's performance through synthetic simulations, for which we generate traffic according to known datacenter traffic patterns [10, 38, 66]. We used *NetworkX* for topology creation, as well as for matching computations. Our simulation code is available at [67].

Specifically, we have considered real-world datacenter topologies (3-tier) with 8 and 16 aggregation switches, 80 racks and 160 racks, respectively, where each rack contains 10 hosts (i.e., up to 1,600 hosts in the network). In addition to the electrical network, we assume a non-blocking optical circuit switch, which connects to each Chopin node  $k$  times at 10 Gbps, we vary the value of  $k$  between 1, 2, and 4.

Real datacenter's data plane parameters were used. The link capacities are 1 Gbps between servers and ToR switches, 10 Gbps between ToRs and aggregation-level switches as in [10], and 40 Gbps between the aggregation-level switches and cores, as in [44]. The reconfiguration time is approximately  $11\mu\text{sec}$  [31, 63], as discussed in Section 4. As the host's traffic contains hundreds of Mbps on average at all times [38], we analyzed average host demand levels of 200 Mbps.

Chopin's evaluation focuses on increasing the optical throughput. Optimizing Chopin's optical throughput adds some approximation to it, in three aspects: (i) partial maximal weight matching computation, (ii) higher optical degree of Chopin nodes, and (iii) approximated maximal weight matching for higher optical degrees, as discussed next.

Adding several optical routes per Chopin node improves its optical throughput by using higher connectivity between Chopin nodes. This can be achieved, e.g., by a wavelength-

## 4:12 Chopin: Combining Distributed and Centralized Schedulers

Distribution	HULL	pFabric	VL2
Mean	Low (100 KB)	Medium (1.7MB)	High (12 MB)
Variance	Low	Medium	High
Centralized perf.	0.42	0.7	0.77
Distributed perf.	0.49	0.75	0.77
Chopin	0.5	0.76	0.78

■ **Table 2** Throughput ratio for optical degree 4

selective switch (WSS) module at each ToR switch, which is a customized  $1 \times 4$ -port Nistica full-fledged 100 WSS module (as suggested in [21]). This implementation enables each Chopin node to connect other Chopin nodes by up to 4 optical links.

This becomes less attractive for a larger number of channels (namely, greater than 4) because of the additional noise (e.g., the multiplexer enables additive noise funneling from each of the sources into the reconfigurable optical add/drop multiplexer ROADM network) [71]. Furthermore, recent studies show that using  $1 \times 8$  ports increases the system costs by a factor of 10 compared to  $1 \times 4$  ports [82]. Thus, for cost-effective systems, where several optical switches are recommended, we analyze Chopin’s performance where each Chopin node has up to 4 connections to other Chopin nodes (i.e., “optical degree  $k = 4$ ”).

**Traffic patterns.** We generate the traffic flow based on previous studies of traffic characteristics of datacenter networks [10, 45, 65]. Flows are TCP [1] with Poisson flow arrival times [38], whose size distribution follows one of three well-known flow size distributions: (i) HULL [3]; (ii) pFabric [2, 51, 60]; and (iii) VL2 [32].

The distribution of flow arrival time to the ToR switches is modeled as a Poisson process, where the servers use the network heavily, constantly transmitting and receiving several hundreds of Mbps data on average all the time [38]. Such a traffic pattern matches the common inflow rate in today’s datacenters serving a variety of applications, such as video and job-task managers [41, 52].

The dispersion pattern in the simulation was based on the observation that traffic is either rack-distributed or destined for one  $\approx 1\%$ – $10\%$  of the hosts, spread across most of the source’s cluster (tens of racks) [45, 66]. The inter-rack demand per host was set to approximately 150 Mbps, based on [38].

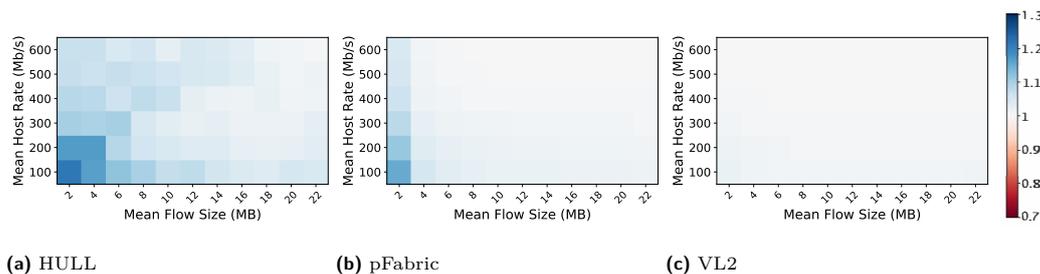
Chopin’s optical circuits throughput is analyzed w.r.t.:

- Different flow traffic distributions (HULL, VL2 and pFabric).
- Different scheduler policies: all-distributed/-centralized, and in-between (varying threshold  $\alpha$  levels).

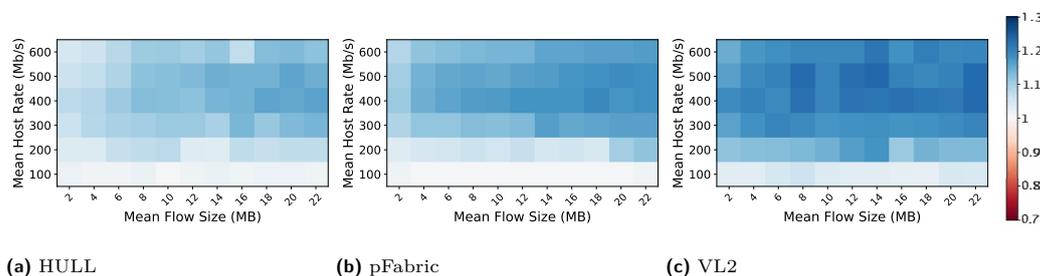
We found that each flow size distribution has special properties, with respect to flow length and flow size variance. These properties have a significant influence on the performance of both the centralized and the distributed scheduler:

- The *HULL* flow distribution is a Pareto distribution where almost all flows are mice<sup>5</sup> ( $< 10\text{KB}$ ). Moreover, flow variance is low. Therefore, the centralized scheduler throughput is low, as there are not many elephant flows, and the differences between the flow carried by the optical links is small compared to the others.
- The *VL2* distribution creates many elephant flows, with high variance. Therefore, the centralized scheduler can optimize the traffic and the distributed scheduler can make decisions which improves the throughput through the optical circuits.

<sup>5</sup> We define mice and elephant flows based on the distinction made by [10].



**Figure 6** Comparison between *Chopin* and the *centralized* scheduler under different traffic patterns (generated by scaling realistic flow size distributions and Poisson flow arrival times under different rates), when the optical ToR switch’s connectivity is 4. The color represents the ratio between the optical throughput of *Chopin* and the centralized scheduler. As the blue cells become darker, *Chopin* more strongly outperforms the centralized scheduler.



**Figure 7** Comparison between *Chopin* and the *distributed* scheduler under the same settings as in Fig. 6. The color represents the ratio between the optical throughput of *Chopin* and the distributed scheduler. As the blue cells become darker, *Chopin* more strongly outperforms the distributed scheduler.

- The *pFabric* distribution includes some elephant flows (but medium mean). With medium variance the distributed scheduler operates as for VL2, but centralized is slightly less effective, due to shorter flows.

The properties of these traffic patterns and their impact on the scheduler performance are described in Table 2, for an optical degree of 4. Notice that Hull, as a Pareto distribution with  $\alpha = 1.05$ , mean=100KB [3] has unbounded variance. *pFabric* has mean value of approximately 1.7 MB [51] and variance of 3.9MB, and VL2 [32] has mean value of 12 MB with variance of 85MB.

## 6.2 Scheduler Implementation

The *Chopin* scheduler consists of centralized scheduler and distributed scheduler. The centralized, described in Section 4, aims to find a Maximum Weight Matching (MWM) solution.

However, due to its complexity, especially as the optical degree ( $k$ ) increases, the centralized scheduler suffers from large running times. In order to reduce delays, two approximations were introduced. First, MWM with degree  $k$  is computed as an iterative Edmond’s MWM algorithm. Second, the centralized scheduler considers only the top- $m$  live flows per ToR switch (instead of all possible pairs between the  $n$  nodes). We found top-5 MWM running time to be within 1% of the MWM over all pairs, since MWM complexity (which is the core of our b-matching solution) scales linearly with the number of to-be-matched edges.

Moreover, as each node reports only its top-5 nodes to the controller, the report can be sent by a single 200 bit packet. Considering 100 switches reporting to a controller with 1Gbps network card, and control plane latency of 0.05 ms, all reports can be sent within 0.07 ms. The reconfiguration commands (for at most 4 links per ToR switch) will have similar latency.

Lastly, we consider the actual update time of the switch internal configuration after the reconfiguration message arrives. However, it is considered as negligible, assuming an optimized implementation with time complexity dominated by TCAM update time which is approximately in the 0.025 ms range [43]. Therefore, the total reconfiguration latency based on top-5 nodes can be bounded by 3 ms.

### 6.3 Scheduler Evaluation Benchmarks

We evaluate Chopin with respect to the following centralized and distributed schedulers.

**Centralized schedulers** are designed for long term datacenter flows. The realistic centralized scheduler was analyzed through different values of the centralized scheduler epoch  $T$ , and with delay  $\Delta$  equals to  $T$ . Namely, in the third epoch, the scheduler uses matching results based on data collected in the first epoch, i.e., data from two epochs ago, recall Fig. 3 (characterized both *centralized scheduler* and *Chopin centralized scheduler*). Similarly to Veisllari et al. [77], we consider an *optimal scheduler*, which runs MWM, *with access to future traffic knowledge*. For each 1 ms interval it uses the optimal matching computed as MWM of that interval. Therefore, it is an upper bound for datacenter performance.

We also consider an *online optimal scheduler*, which has no knowledge of the future but it does not suffer from any delay. For each 1 ms interval it uses an allocation computed as the MWM of the previous interval.

**Distributed schedulers** are designed for bursts and short datacenter flows. According to Roy et al. [66], 90% of the time, 50% of the heavy flows change within 1 ms. Therefore, the distributed scheduler should operate repeatedly in high frequency. *Chopin's distributed scheduler* is set to operate every 1 ms, which is the length of a time-slot in our model. Furthermore, as in the centralized scheduler, both aggregation interval and delay ( $a$  and  $\delta$  respectively) are set to equal the time between two invocations (namely, 1 ms). In addition, the performance of a *distributed scheduler* (unrelated to a centralized scheduler) with the same properties was also analyzed. Moreover, as discussed in §5, a major factor of the Chopin distributed scheduler is the threshold  $\alpha$ , the level under which the centralized allocation can be changed by the distributed scheduler.

As the threshold decreases, Chopin's performance is closer to a centralized scheduler. Similarly, as the threshold increases, Chopin's performance is closer to being distributed. Therefore, we evaluate Chopin for different threshold levels, between 0.1 to 1.3, to capture Chopin's performance scheduling between distributed and centralized scheduling.

### 6.4 Centralized-Distributed Trade-off

How can we find an optimal tradeoff between the centralized scheduler, which provides accurate solutions but relies on outdated information, and the distributed scheduler which relies on more recent information but provides approximate solutions (due to locality)?

This trade-off was analyzed in two related ways: (i) optimal threshold, and (ii) optimal reconfiguration number. The threshold  $\alpha$  is the parameter which enables the distributed scheduler to change the centralized matching, and therefore, to adapt the traffic changes in small time intervals. For example, a circuit allocation between ToR pair with high throughput

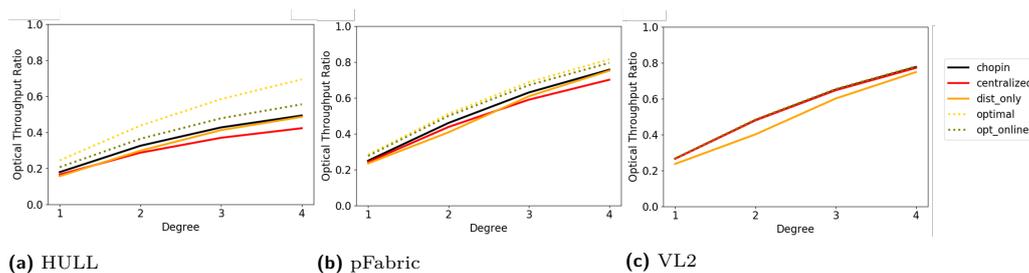


Figure 8 Throughput through the optical circuits, for different optical degrees and flow size distributions.

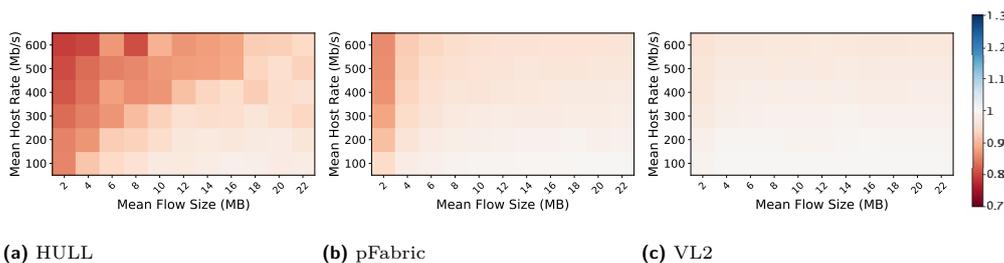


Figure 9 Comparison between *Chopin* and the *online optimal* scheduler under different traffic patterns (each generated by scaling well-known realistic flow size distributions and assuming Poisson flow arrival times under different rates), when the optical ToR switch’s connectivity is 4. As the red cells become darker, the online optimal scheduler performs better than Chopin.

on previous intervals, should be torn down if the flow rate reduces drastically. We found that Chopin’s optimal threshold  $\alpha$  is between  $\approx 0.4 - 0.7$ , depending on the traffic pattern. For the HULL traffic pattern, it achieves higher performance with  $\alpha = 0.4$ , while for DCTCP and VL2 traffic, the optimal threshold is approximately 0.7. Moreover, across this range, the performance across all the traffic patterns were the highest, with low deviation.

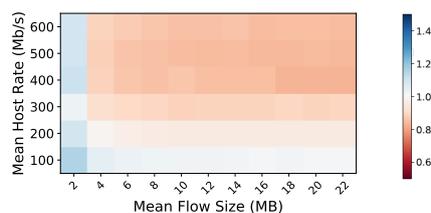
## 6.5 On the Benefit of Hybrid Scheduling

To analyze Chopin’s (the hybrid scheduler) improvement over distributed and centralized schedulers, we consider the optical throughput ratio. Fig. 6 and Fig. 7 describe Chopin’s improvement ratio for each of the flow patterns, compared to centralized and distributed schedulers (respectively), when considering a centralized compute epoch of 3 ms, see §6.2.

The results show that Chopin outperforms the centralized scheduler for *every* traffic pattern. We found that Chopin’s optical throughput ratio is higher than the centralized scheduler optical throughput ratio by up to 20% in the HULL distribution, 15% for pFabric pattern, and 2% for datacenters with a VL2 flow size distribution (as shown in Fig. 6a). Moreover, Chopin also achieves a higher throughput ratio compared to the distributed scheduler across all traffic patterns. Specifically, Chopin increases the optical throughput ratio of the distributed scheduler by up to 16% in the HULL distribution, 20% in the pFabric and 23% in VL2 traffic (see Fig. 7b). Therefore, Chopin outperforms both centralized and distributed schedulers.

## 6.6 Optical Degree Improvement

Next, we examine the improvement as a function of Chopin nodes’ optical degree. Therefore, we have focused on the optimal threshold for each of the flow patterns, where the centralized



■ **Figure 10** Comparison between centralized and distributed schedulers, as in Fig. 1, but with 160 ToR switches. For brevity we only present pFabric results.

scheduler epoch is 3 ms, as discussed in Section 4.

Fig. 8 presents the ratio between the throughput through optical circuits to the overall throughput (electrical and optical networks combined). This *optical throughput ratio* changes with the optical degree and flow patterns, as shown in Fig. 8. In each flow pattern, all the schedulers were considered. It is shown that as the degree increases, the throughput among all the schedulers improved, and that Chopin’s throughput is higher than both the centralized and the distributed schedulers. Moreover, as the number of elephant flows increases (as in VL2), Chopin’s throughput is getting closer to the optimal. It is consistent with Chopin’s aim to carry elephant flows over optical circuits. Therefore, flow patterns with high number of elephant flows benefit more from using Chopin.

## 6.7 Chopin VS Online Optimal Scheduler

We compared between Chopin performance, and online optimal scheduler performance (where centralized updates are being sent to Chopin nodes every 1 ms instead of 3 ms respectively). Fig. 9 shows that even if Chopin’s centralized scheduler updates were sent every 1 ms (such as in the online optimal scheduler), there is no significant improvement for the VL2 flow pattern (see Fig. 9c). In other words, Chopin is closer to the optimal scheduler as the flows become larger, because as the mean flow is longer, the changes over small time intervals (such as 1 ms) become minor. Therefore, in these cases, the added value of high frequent scheduling updates, even with the “future” information (as in the optimal scheduler), decreases. Chopin can benefit from higher frequent centralized updates mostly in HULL distribution, (where the flows are usually shorter), by approximately 20%.

## 6.8 Sensitivity Analysis

We further analyzed our results by considering different sizes of networks, e.g. with 100 ToR switches and for 160 ToR switches, where there are 10 hosts per ToR switch, and for different rates per host (100–600 Mbps). Across all the networks that were examined, for each of the traffic patterns, we observe that on certain conditions the distributed scheduler outperforms the centralized scheduler and vice versa, with respect to higher optical throughput. For instance, see the heatmap for a network with 160 ToR switches, each with an optical degree of 4 connections, under the pFabric traffic pattern in Fig. 10. It shows that under the pFabric distribution, when the mean flow is larger than 5 MB, the centralized scheduler achieves higher performance compared to the distributed one. This phenomenon was also observed for the 80 ToR network, as in Fig. 1b.

Moreover, Chopin’s performance can scale. We demonstrate its effectiveness over a concrete network topology (specified in Section 6.1), but faster links with higher demand

will create the same bottleneck and resolve with Chopin in the same way.

## 7 Conclusion

Chopin aims to combine the benefits of centralized scheduling with distributed scheduling, to provide high throughput and fast reaction. While centralized and distributed scheduling has also been combined in all-static non-hybrid networks, e.g., Facebook’s Express Backbone [42], hybrid networks with optical circuits pose structurally different challenges. In particular, we find that distributed decisions benefit from being closer in time to the measurements they are based on, which is more critical than the rate of decisions.

We believe that our work opens several interesting avenues for future research. In particular, while we achieve significant performance gains, our approach is more complex than the state-of-the-art and it would be useful to simplify it further. Furthermore, our distributed schedulers use the same threshold for all nodes as a homogeneous strategy. While this succinct representation is sufficient for the settings described in this paper, it can be interesting to explore heterogeneity, e.g., to increase the threshold on very congested racks. Finally, the trade-off between an elephant flow’s duration and the time before it starts to route through optical circuits can be considered for future optimization.

---

## References

- 1 Mohammad Alizadeh. Empirical traffic generator. Cisco DC Repositories, 2015.
- 2 Mohammad Alizadeh, Albert Greenberg, David A. Maltz, Jitendra Padhye, Parveen Patel, Balaji Prabhakar, Sudipta Sengupta, and Murari Sridharan. Data center tcp (dctcp). In *ACM SIGCOMM*, 2010.
- 3 Mohammad Alizadeh, Abdul Kabbani, Tom Edsall, Balaji Prabhakar, Amin Vahdat, and Masato Yasuda. Less is more: Trading a little bandwidth for ultra-low latency in the data center. In *NSDI*. USENIX Association, 2012.
- 4 Richard P. Anstee. A polynomial algorithm for b-matchings: An alternative approach. *Inf. Process. Lett.*, 24(3):153–157, 1987.
- 5 John Augustine, Kristian Hinnenthal, Fabian Kuhn, Christian Scheideler, and Philipp Schneider. Shortest paths in a hybrid network model. In *SODA*, pages 1280–1299. SIAM, 2020.
- 6 Chen Avin, Manya Ghobadi, Chen Griner, and Stefan Schmid. On the complexity of traffic traces and implications. In *Proc. ACM SIGMETRICS*, 2020.
- 7 Navid Hamed Azimi, Zafar Ayyub Qazi, Himanshu Gupta, Vyas Sekar, Samir R. Das, Jon P. Longtin, Himanshu Shah, and Ashish Tanwer. Firefly: a reconfigurable wireless data center fabric using free-space optics. In *SIGCOMM*. ACM, 2014.
- 8 Hitesh Ballani, Paolo Costa, Raphael Behrendt, Daniel Cletheroe, István Haller, Krzysztof Jozwik, Fotini Karinou, Sophie Lange, Kai Shi, Benn Thomsen, and Hugh Williams. Sirius: A flat datacenter network with nanosecond optical switching. In *SIGCOMM*, pages 782–797. ACM, 2020.
- 9 Alkida Balliu, Sebastian Brandt, Juho Hirvonen, Dennis Olivetti, Mikaël Rabie, and Jukka Suomela. Lower bounds for maximal matchings and maximal independent sets. *J. ACM*, 68(5):39:1–39:30, 2021.
- 10 T. Benson, A. Akella, and D.A. Maltz. Network traffic characteristics of data centers in the wild. In *ACM IMC*, pages 267–280, 2010.
- 11 André Berger, James Gross, Tobias Harks, and Simon Tenbusch. Constrained resource assignments: Fast algorithms and applications in wireless networks. *Management Science*, 62, 11 2015.

- 12 Maciej Besta and Torsten Hoefler. Slim fly: A cost effective low-diameter network topology. In *IEEE SC*, pages 348–359, 2014.
- 13 Sayan Bhattacharya, Deeparnab Chakrabarty, and Monika Henzinger. Deterministic dynamic matching in  $O(1)$  update time. *Algorithmica*, 82(4):1057–1080, 2020.
- 14 Li Chen, Kai Chen, Zhonghua Zhu, Minlan Yu, George Porter, Chunming Qiao, and Shan Zhong. Enabling wide-spread communications on optical fabric with megaswitch. In *USENIX NDSI*, 2017.
- 15 Charles Clos. A study of non-blocking switching network. *Bell System Technology Journal*, 32(2):406–424, 1953.
- 16 Shibsankar Das. A modified decomposition algorithm for maximum weight bipartite matching and its experimental evaluation. *Sci. Ann. Comput. Sci.*, 30(1):39–67, 2020.
- 17 Sushovan Das, Afsaneh Rahbar, Xinyu Crystal Wu, Zhuang Wang, Weitao Wang, Ang Chen, and T. S. Eugene Ng. Shufflecast: An optical, data-rate agnostic, and low-power multicast architecture for next-generation compute clusters. *IEEE/ACM Trans. Netw.*, 30(5):1970–1985, 2022.
- 18 Pamela Delgado, Florin Dinu, Anne-Marie Kermarrec, and Willy Zwaenepoel. Hawk: Hybrid datacenter scheduling. In *USENIX ATC*, 2015.
- 19 Nikhil Devanur, Janardhan Kulkarni, Gireeja Ranade, Manya Ghobadi, Ratul Mahajan, and Amar Phanishayee. Stable matching algorithm for an agile reconfigurable data center interconnect. Technical Report 2016-1140, MSR, Jun 2016.
- 20 Fahad Dogar, Thomas Karagiannis, Hitesh Ballani, and Antony Rowstron. Decentralized task-aware scheduling for data center networks. *ACM SIGCOMM CCR*, 44, 08 2014.
- 21 N. Farrington, A. Forencich, G. Porter, P. C. Sun, J. E. Ford, Y. Fainman, G. C. Papen, and A. Vahdat. A multiport microsecond optical circuit switch for data center networking. *IEEE Phot. Techn. L.*, 25(16):1589–92, Aug 2013.
- 22 Nathan Farrington, Alex Forencich, Pang-Chen Sun, Shaya Fainman, Joe Ford, Amin Vahdat, George Porter, and George C. Papen. A 10 us hybrid optical-circuit/electrical-packet network for datacenters. In *OFC/NFOEC*. OSA, 2013.
- 23 Nathan Farrington, George Porter, Yeshaiahu Fainman, George Papen, and Amin Vahdat. Hunting mice with microsecond circuit switches. In *ACM HotNets*, 2012.
- 24 Nathan Farrington, George Porter, Sivasankar Radhakrishnan, Hamid Hajabdolali Bazzaz, Vikram Subramanya, Yeshaiahu Fainman, George Papen, and Amin Vahdat. Helios: a hybrid electrical/optical switch architecture for modular data centers. In *SIGCOMM*. ACM, 2010.
- 25 Thomas Fenz, Klaus-Tycho Foerster, Stefan Schmid, and Anaïs Villedieu. Efficient non-segregated routing for reconfigurable demand-aware networks. *Comput. Commun.*, 164:138–147, 2020.
- 26 Klaus-Tycho Foerster, Maciej Pacut, and Stefan Schmid. On the complexity of non-segregated routing in reconfigurable data center architectures. *Comput. Commun. Rev.*, 49(2):2–8, 2019.
- 27 Klaus-Tycho Foerster and Stefan Schmid. Survey of reconfigurable data center networks: Enablers, algorithms, complexity. *SIGACT News*, 50(2):62–79, 2019.
- 28 Harold N. Gabow. Data structures for weighted matching and extensions to  $b$ -matching and  $f$ -factors. *ACM Trans. Algorithms*, 14(3):39:1–39:80, 2018.
- 29 Zvi Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Comput. Surv.*, 18(1):23–38, March 1986.
- 30 Manya Ghobadi, Ratul Mahajan, Amar Phanishayee, Pierre-Alexandre Blanche, Houman Rastegarfar, Madeleine Glick, and Daniel Kilper. Design of mirror assembly for an agile reconfigurable data center interconnect. Technical Report 2016-1139, MSR, Jun 2016.
- 31 Monia Ghobadi, Ratul Mahajan, Amar Phanishayee, Nikhil Devanur, Janardhan Kulkarni, Gireeja Ranade, Pierre-Alexandre Blanche, Houman Rastegarfar, Madeleine Glick, and Daniel Kilper. Projector: Agile reconfigurable data center interconnect. In *ACM SIGCOMM*, pages 216–229, 2016.

- 32 A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. VL2: A scalable and flexible data center network. *ACM SIGCOMM*, 39(4):51–62, 2009.
- 33 A. Grieco, G. Porter, and Y. Fainman. Integrated space-division multiplexer for application to data center networks. *IEEE J. Sel. Top. Quant. El.*, 22(6), 2016.
- 34 Chen Griner, Stefan Schmid, and Chen Avin. Cachenet: Leveraging the principle of locality in reconfigurable network design. *Computer Networks*, 204:108648, 2022.
- 35 Chen Griner, Johannes Zerwas, Andreas Blenk, Manya Ghobadi, Stefan Schmid, and Chen Avin. Cerberus: The power of choices in datacenter topology design - A throughput perspective. *Proc. ACM Meas. Anal. Comput. Syst.*, 5(3):38:1–38:33, 2021.
- 36 Matthew Nance Hall, Klaus-Tycho Foerster, Stefan Schmid, and Ramakrishnan Durairajan. A survey of reconfigurable optical networks. *Opt. Switch. Netw.*, 41:100621, 2021.
- 37 Daniel Halperin, Srikanth Kandula, Jitendra Padhye, Paramvir Bahl, and David Wetherall. Augmenting data center networks with multi-gigabit wireless links. In *SIGCOMM*, pages 38–49. ACM, 2011.
- 38 Y. Han, J.H. Yoo, and J.W.K. Hong. Poisson shot-noise process based flow-level traffic matrix generation for data center networks. In *IFIP/IEEE IM*, May 2015.
- 39 Kathrin Hanauer, Monika Henzinger, Stefan Schmid, and Jonathan Trummer. Fast and heavy disjoint weighted matchings for demand-aware datacenter topologies. In *INFOCOM*, pages 1649–1658. IEEE, 2022.
- 40 Keqiang He, Junaid Khalid, Aaron Gember-Jacobson, Sourav Das, Chaithan Prakash, Aditya Akella, Li Erran Li, and Marina Thottan. Measuring control plane latency in sdn-enabled switches. In *ACM SIGCOMM*, SOSR '15, pages 25:1–25:6, 2015.
- 41 Netflix help center. Internet connection speed recommendations, 2018. URL: <https://help.netflix.com/en/node/306>.
- 42 Mikel Jimenez and Henry Kwik. Building Express Backbone: Facebook’s new long-haul network, May 2017. URL: <https://engineering.fb.com/data-center-engineering/building-express-backbone-facebook-s-new-long-haul-network/>.
- 43 Mikel Jimenez and Henry Kwik. Ternary Content Addressable Memory (TCAM) Search IP for SDNet - SmartCORE IP Product Guide. Technical report, Xilinx, November 2017. URL: [https://www.xilinx.com/support/documentation/ip\\_documentation/tcam/pg190-tcam.pdf](https://www.xilinx.com/support/documentation/ip_documentation/tcam/pg190-tcam.pdf).
- 44 S. Kandula, J. Padhye, and P. Bahl. Flyways to de-congest data center networks. In *ACM HotNets*, 2009.
- 45 S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken. The nature of data center traffic: Measurements & analysis. In *ACM IMC*, pages 202–208, 2009.
- 46 Simon Kassing, Asaf Valadarsky, Gal Shahaf, Michael Schapira, and Ankit Singla. Beyond fat-trees without antennae, mirrors, and disco-balls. In *SIGCOMM*, pages 281–294. ACM, 2017.
- 47 Arif M. Khan, Alex Pothan, Md. Mostofa Ali Patwary, Nadathur Rajagopalan Satish, Narayanan Sundaram, Fredrik Manne, Mahantesh Halappanavar, and Pradeep Dubey. Efficient approximation algorithms for weighted b-matching. *SIAM J. Sci. Comput.*, 38(5), 2016.
- 48 Viatcheslav Korenwein. The practical power of data reduction for maximum-cardinality matching. Masterthesis, TU Berlin, Januar 2018. Master thesis. URL: <http://ftp.akt.tu-berlin.de/publications/theses/ma-viatcheslav-korenwein.pdf>.
- 49 M. Kuzniar, P. Peresini, and D. Kostic. What you need to know about sdn control and data planes. Technical report, EPFL, 2014.
- 50 Adam N. Letchford, Gerhard Reinelt, and Dirk Oliver Theis. Odd minimum cut sets and b-matchings revisited. *SIAM J. Discret. Math.*, 22(4):1480–1487, 2008.
- 51 Z. Li, W. Bai, K. Chen, D. Han, Y. Zhang, D. Li, and H. Yu. Rate-aware flow scheduling for commodity data center networks. In *IEEE INFOCOM*, pages 1–9, 2017. doi:10.1109/INFOCOM.2017.8057082.

- 52 Xiao Ling, Yi Yuan, Dan Wang, Jiangchuan Liu, and Jiahai Yang. Joint scheduling of mapreduce jobs with servers. *J. Parallel Distrib. Comput.*, 90(C):52–66, April 2016.
- 53 He Liu, Feng Lu, Alex Forencich, Rishi Kapoor, Malveeka Tewari, Geoffrey M. Voelker, George Papen, Alex C. Snoeren, and George Porter. Circuit switching under the radar with reactor. In *USENIX NSDI*, pages 1–15, April 2014.
- 54 He Liu, Feng Lu, Alex Forencich, Rishi Kapoor, Malveeka Tewari, Geoffrey M. Voelker, George Papen, Alex C. Snoeren, and George Porter. Circuit switching under the radar with reactor. In *USENIX NSDI*, pages 1–15, 2014.
- 55 Zvi Lotker, Boaz Patt-Shamir, and Seth Pettie. Improved distributed approximate matching. *J. ACM*, 62(5):38:1–38:17, 2015.
- 56 Zvi Lotker, Boaz Patt-Shamir, and Adi Rosén. Distributed approximate matching. *SIAM J. Comput.*, 39(2):445–460, 2009.
- 57 Long Luo, Klaus-Tycho Foerster, Stefan Schmid, and Hongfang Yu. Optimizing multicast flows in high-bandwidth reconfigurable datacenter networks. *J. Netw. Comput. Appl.*, 203:103399, 2022.
- 58 William M. Mellette, Rajdeep Das, Yibo Guo, Rob McGuinness, Alex C. Snoeren, and George Porter. Expanding across time to deliver bandwidth efficiency and low latency . In *NSDI*. USENIX Association, 2020.
- 59 William M. Mellette, Rob McGuinness, Arjun Roy, Alex Forencich, George Papen, Alex C. Snoeren, and George Porter. Rotornet: A scalable, low-complexity, optical datacenter network. In *SIGCOMM*. ACM, 2017.
- 60 Alizadeh Mohammad, Yang Shuang, Sharif Milad, Katti Sachin, McKeown Nick, Prabhakar Balaji, and Shenker Scott. pfabric: Minimal near-optimal datacenter transport. *ACM SIGCOMM*, 43(4):435–446, 2013.
- 61 Matthias Müller-Hannemann and Alexander Schwartz. Implementing weighted b-matching algorithms: Insights from a computational study. *ACM Journal of Experimental Algorithmics*, 5:8, 2000.
- 62 George Porter, Richard Strong, Nathan Farrington, Alex Forencich, Pang Chen-Sun, Tajana Rosing, Yeshaiahu Fainman, George Papen, and Amin Vahdat. Integrating microsecond circuit switching into the data center. *ACM SIGCOMM*, 43(4):447–458, 2013.
- 63 George Porter, Richard D. Strong, Nathan Farrington, Alex Forencich, Pang-Chen Sun, Tajana Rosing, Yeshaiahu Fainman, George Papen, and Amin Vahdat. Integrating microsecond circuit switching into the data center. In *SIGCOMM*, pages 447–458. ACM, 2013.
- 64 Leon Poutievski, Omid Mashayekhi, Joon Ong, Arjun Singh, Muhammad Mukarram Bin Tariq, Rui Wang, Jianan Zhang, Virginia Beauregard, Patrick Conner, Steve D. Gribble, Rishi Kapoor, Stephen Kratzer, Nanfang Li, Hong Liu, Karthik Nagaraj, Jason Ornstein, Samir Sawhney, Ryohei Urata, Lorenzo Vicisano, Kevin Yasumura, Shidong Zhang, Junlan Zhou, and Amin Vahdat. Jupiter evolving: transforming google’s datacenter network via optical circuit switches and software-defined networking. In *SIGCOMM*, pages 66–85. ACM, 2022.
- 65 Y. Qiao, Z. Hu, and J. Luo. Efficient traffic matrix estimation for data center networks. In *IFIP Networking*, pages 1–9, May 2013.
- 66 Arjun Roy, Hongyi Zeng, Jasmeet Bagga, George Porter, and Alex C. Snoeren. Inside the social network’s (datacenter) network. In *SIGCOMM*. ACM, 2015.
- 67 Neta Rozen-Schiff, David Hay, Stefan Schmid, and Klaus-Tycho Foerster. Chopin implementation code. [https://bitbucket.org/NetaRS/sched\\_analytics/src/master/](https://bitbucket.org/NetaRS/sched_analytics/src/master/), October 2020.
- 68 Stefan Schmid, Chen Avin, Christian Scheideler, Michael Borokhovich, Bernhard Haeupler, and Zvi Lotker. Splaynet: Towards locally self-adjusting networks. *IEEE/ACM Trans. Netw.*, 24(3):1421–1433, 2016.
- 69 Ankit Singla, Chi-Yao Hong, Lucian Popa, and Philip Brighten Godfrey. Jellyfish: Networking data centers, randomly. In *USENIX NSDI*, volume 12, 2012.

- 70 Ankit Singla, Atul Singh, and Yan Chen. OSA: An optical switching architecture for data center networks with unprecedented flexibility. In *USENIX NSDI*, 2012.
- 71 T. A. Strasser and J. L. Wagener. Wavelength-selective switches for roadm applications. *IEEE J. Sel. Top. Quant. El.*, 16(5), 2010.
- 72 Xiaoye Steven Sun and T. S. Eugene Ng. When creek meets river: Exploiting high-bandwidth circuit switch in scheduling multicast data. In *ICNP*, pages 1–6. IEEE Computer Society, 2017.
- 73 Jukka Suomela. Survey of local algorithms. *ACM Comput. Surv.*, 45(2):24:1–24:40, 2013.
- 74 Akhilesh S. Thyagaturu, Anu Mercian, Michael P. McGarry, Martin Reisslein, and Wolfgang Kellerer. Software defined optical networks (sdons): A comprehensive survey. *IEEE Commun. Surv. Tutorials*, 18(4):2738–2786, 2016.
- 75 Gerard J Tortora and Bryan H Derrickson. *Principles of anatomy and physiology*. John Wiley & Sons, 2018.
- 76 Asaf Valadarsky, Gal Shahaf, Michael Dinitz, and Michael Schapira. Xpander: Towards optimal-performance datacenters. In *ACM CoNEXT*, 2016.
- 77 R. Veisllari, S. Bjornstad, and N. Stol. Scheduling techniques in an integrated hybrid node with electronic buffers. In *ONDM*, April 2012.
- 78 Shaileshh Bojja Venkatakrishnan, Mohammad Alizadeh, and Pramod Viswanath. Costly circuits, submodular schedules and approximate carathéodory theorems. In *SIGMETRICS*. ACM, 2016.
- 79 Guohui Wang, David G. Andersen, Michael Kaminsky, Konstantina Papagiannaki, T. S. Eugene Ng, Michael Kozuch, and Michael P. Ryan. c-through: part-time optics in data centers. In *SIGCOMM*, pages 327–338. ACM, 2010.
- 80 Yiting Xia, Xiaoye Steven Sun, Simbarashe Dzinamarira, Dingming Wu, Xin Sunny Huang, and T. S. Eugene Ng. A tale of two topologies: Exploring convertible data center network architectures with flat-tree. In *SIGCOMM*, pages 295–308. ACM, 2017.
- 81 Bing Xiong, Kun Yang, Jinyuan Zhao, Wei Li, and Keqin Li. Performance evaluation of openflow-based software-defined networks based on queueing model. *Comput. Netw.*, 102(C):172–185, 2016.
- 82 Haining Yang, Brian Robertson, Peter Wilkinson, and Daping Chu. Low-cost cdc roadm architecture based on stacked wavelength selective switches. *J. Opt. Commun. Netw.*, 9(5):375–384, May 2017.
- 83 Johannes Zerwas, Chen Avin, Stefan Schmid, and Andreas Blenk. Exrec: Experimental framework for reconfigurable networks based on off-the-shelf hardware. In *ANCS*, pages 66–72. ACM, 2021.
- 84 Johannes Zerwas, Wolfgang Kellerer, and Andreas Blenk. What you need to know about optical circuit reconfigurations in datacenter networks. In *ITC*, pages 1–9. IEEE, 2021.
- 85 Danyang Zhuo, Qiao Zhang, Vincent Liu, Arvind Krishnamurthy, and Thomas Anderson. Rack-level congestion control. In *ACM HotNets*, 2016.

## **A** Chopin’s Distributed Scheduler Algorithm

We provide on the next page the pseudo-code of Chopin’s distributed algorithm in Algorithm 1.

---

**Algorithm 1** Chopin Distributed Algorithm Code for Node  $i$ .

$\text{max\_reqs}$  : The number of allowed requests per ToR switch  
 $\text{cur\_nodes}$  : The nodes currently connected with  $i$   
 $\text{centralized\_nodes}$  : The nodes matched to  $i$  by the centralized scheduler in its last invocation  
 $\text{received\_reqs} \leftarrow \emptyset$

Upon the beginning of a distributed scheduler epoch:

```

1: function START:
2:    $\text{matched\_nodes} \leftarrow \emptyset$ 
3:   for  $p \in (\text{cur\_nodes} \cap \text{centralized\_nodes})$  do
4:     if  $r_{i,p} \geq \alpha \cdot R_{i,p}$  then
5:        $\text{matched\_nodes.add}(p)$ 
6:    $\text{req\_nodes} \leftarrow ([n] \setminus \{i\}) \setminus \text{matched\_nodes}$  ▷  $n$  denotes the
number of Chopin nodes in the network
7:    $\text{req\_nodes} \leftarrow \text{GET\_TOP\_NODES}(\text{req\_nodes}, \text{max\_reqs})$ 
▷ Top  $\text{max\_reqs}$  nodes, out of  $\text{req\_nodes}$ , with
the most bi-directional traffic with ToR switch  $i$ .
8:    $\text{grants} \leftarrow \emptyset$ ;  $\text{denies} \leftarrow \emptyset$ 
9:    $\text{SEND\_REQUESTS}(\text{req\_nodes})$  ▷ Send request to all
nodes in  $\text{req\_nodes}$ .

```

Upon receiving a **request** message from  $\text{src\_id}$ :

```

10: function REQUEST_HANDLER( $\text{src\_id}$ ):
11:    $\text{received\_reqs.add}(\text{src\_id})$ 

```

Upon a timeout event (implying the request phase has ended):

```

12: function REQUEST_TIMEOUT_HANDLER:
13:    $\text{nodes} \leftarrow \text{req\_nodes} \cap \text{received\_reqs}$ 
14:    $\text{free\_links} \leftarrow k - |\text{matched\_nodes}|$ 
15:    $\text{granted} \leftarrow \text{GET\_TOP\_NODES}(\text{nodes}, \text{free\_links})$ 
16:    $\text{rejected} \leftarrow \text{received\_reqs} \setminus \text{granted}$ 
17:    $\text{SEND\_DENIES}(\text{rejected})$  ▷ Send deny message to all
nodes in rejected set.
18:    $\text{SEND\_GRANTS}(\text{granted})$  ▷ Send grant message to
all nodes in granted set.
19:    $\text{grant\_sent} \leftarrow \text{true}$ 
20:    $\text{TRY\_EXECUTE\_DECISIONS}()$ 

```

Upon receiving a **grant** message from  $\text{src\_id}$ :

```

21: function GRANT_HANDLER( $\text{src\_id}$ ):
22:    $\text{grants.add}(\text{src\_id})$ 
23:    $\text{TRY\_EXECUTE\_DECISIONS}()$ 

```

Upon receiving a **deny** message from  $\text{src\_id}$ :

```

24: function DENY_HANDLER( $\text{src\_id}$ ):
25:    $\text{denies.add}(\text{src\_id})$ 
26:    $\text{TRY\_EXECUTE\_DECISIONS}()$ 

```

27: **function** TRY\_EXECUTE\_DECISIONS:

```

28:   if  $\text{denies} \cup \text{grants} \neq \text{req\_nodes}$  or not  $\text{grant\_sent}$  then
29:     return ▷ Not all grant/deny were received
30:    $\text{new\_nodes} \leftarrow \text{granted} \cap \text{grants}$ 
31:   for  $p \in (\text{cur\_nodes} \setminus \text{new\_nodes}) \setminus \text{matched\_nodes}$  do
32:      $\text{DISCONNECT}(p)$ 
33:   for  $p \in (\text{new\_nodes} \setminus \text{cur\_nodes} \setminus \text{matched\_nodes})$  do
34:      $\text{CONNECT}(p)$ 
35:    $\text{received\_reqs} \leftarrow \emptyset$ ;  $\text{grant\_sent} \leftarrow \text{false}$ 

```

---