Coflow scheduling in input-queued switches: Optimal delay scaling and algorithms

Coflow Scheduling in Input-Queued Switches: Optimal Delay Scaling and Algorithms

Qingkai Liang and Eytan Modiano
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology, Cambridge, MA

***Abstract*—A coflow is a collection of parallel flows belonging to the same job. It has the all-or-nothing property: a coflow is not complete until the completion of all its constituent flows. In this paper, we focus on optimizing *coflow-level delay*, i.e., the time to complete all the flows in a coflow, in the context of an $N \times N$ input-queued switch. In particular, we develop a throughput-optimal scheduling policy that achieves the best scaling of coflow-level delay as $N \to \infty$. We first derive lower bounds on the coflow-level delay that can be achieved by any scheduling policy. It is observed that these lower bounds critically depend on the variability of flow sizes. Then we analyze the coflow-level performance of some existing coflow-agnostic scheduling policies and show that none of them achieves provably optimal performance with respect to coflow-level delay. Finally, we propose the Coflow-Aware Batching (CAB) policy which achieves the optimal scaling of coflow-level delay under some mild assumptions.**

I. Introduction

Modern cluster computing frameworks, such as MapReduce [1] and Spark [2], have been widely used in large-scale data processing and analytics. Despite the differences among these frameworks, they share a common feature: the computation is divided into multiple stages and a collection of parallel data flows need to be transferred between groups of machines in successive computation stages. Often the next computation stage cannot start until the completion of all of these flow transfers. For example, during the shuffle phase in MapReduce, any reducer node cannot start the next reduce phase until it receives intermediate results from all of the mapper nodes. As a result, the response time of the entire computing job critically depends on the completion time of these intermediate flows. In some applications, these intermediate flow transfers can account for more than 50% of job completion time [3].

The recently proposed *coflow* abstraction [4] represents such a collection of parallel data flows between two successive computation stages of a job, which exposes application-level requirements to the network. It builds upon the *all-or-nothing* property observed in many applications [6]: a coflow is not complete until the completion of all its constituent flows. As a result, one of the most important metrics in this context is *coflow-level delay* (also referred to as *coflow completion time* in some literature [5]–[7]), i.e., the time to complete all of the flows in a coflow. To improve the overall response time of a job, it is crucial to schedule flow transfers in a way that the coflow-level delay can be reduced. Unfortunately, researchers have largely overlooked such application-level requirements and there has been little work on coflow-level delay optimization.

In this paper, we study coflow-level delay in the context of an $N \times N$ input-queued switch with stochastic coflow arrivals. In each slot, a random number of coflows, each of them consisting of multiple parallel flows, arrive to the input-queued switch where each input/output port can process at most one packet per slot. Such an input-queued switch model is a simple yet practical abstraction for data centers with full bisection bandwidth, where N represents the number of servers. Due to the large scale of modern data centers, we are motivated to study the scaling of coflow-level delay as $N \to \infty$. In particular, we are interested in the optimal scaling of coflow-level delay, i.e., the scaling under an "optimal" scheduling policy[1]. As far as we know, this is the first paper to present coflow-level delay analysis in a large-scale stochastic system.

The contributions of this paper are summarized as follows.

- We derive lower bounds on the expected coflow-level delay that can be achieved by any scheduling policy in an $N \times N$ input-queued switch. These lower bounds critically depend on the variability of flow sizes. In particular, it is shown that if flow sizes are light-tailed, no scheduling policy can achieve an average coflow-level delay better than $O(\log N)$.
- We analyze the coflow-level performance of several coflow-agnostic scheduling policies, where coflow-level information is not leveraged. It is shown that none of these scheduling policies achieves a *provably optimal* scaling of coflow-level delay. For example, the expected coflow-level delay achieved by randomized scheduling is $O(N \log N)$ if coflow sizes are light-tailed, far above the $O(\log N)$ lower bound.
- We show that $O(\log N)$ is the optimal scaling of average coflow-level delay when flow sizes are light-tailed and coflow arrivals are Poisson. This optimal scaling is achievable with our Coflow-Aware Batching (CAB) policy.

The organization of this paper is as follows. We first review related work in Section II and introduce several mathematical

This work was supported by NSF Grants CNS-1116209, CNS-1617091, and by DARPA I2O and Raytheon BBN Technologies under Contract No. HROO 1 l-l 5-C-0097.

[1]An "optimal" policy should be throughput-optimal, i.e., stabilize the system whenever the load $\rho < 1$, and should achieve the minimum coflow-level delay among all throughput-optimal policies.

tools in Section III. The system model is introduced in Section IV. In Section V, we demonstrate fundamental lower bounds on the expected coflow-level delay that can be achieved by any scheduling policy. In Section VI, we analyze the coflow-level performance of some coflow-agnostic scheduling policies. In Section VII, we propose the Coflow-Aware Batching (CAB) policy and show that it achieves the optimal coflow-level delay scaling under some conditions. Finally, simulation results and conclusions are given in Sections VIII and IX, respectively.

II. Related Work

We start with a brief literature review on coflow-level optimization and delay scaling in input-queued switches.

Coflow-level Optimization. The notion of coflows was first proposed by Chowdhy and Stoica [4] to convey job-specific requirements such as minimizing coflow-level delay or meeting some job completion deadline. Unfortunately, coflow-level optimization is often computationally intractable. For example, it was shown in [5] that minimizing the average coflow-level delay is NP-hard. As a result, many heuristic scheduling principles were developed to improve coflow-level delay. In [8], a decentralized coflow scheduling framework was proposed to give priority to coflows according to a variation of the FIFO principle, which performs well for light-tailed flow sizes. In [5], the Varys scheme improves the performance of [8] by leveraging more sophisticated heuristics such as "smallest-bottleneck-first" and "smallest-total-size-first", where global information about coflows is required. The D-CAS scheme in [11] exploits a similar "shortest-remaining-time-first" principle for coflow scheduling. The Aalo framework [6] generalizes the classic least-attained service (LAS) discipline [9] to coflow scheduling; such a scheme does not require prior knowledge about coflows. Zhong *et al.* [12] develop an approximation algorithm to minimize the average coflow-level delay in data centers. Additionally, Chen *et al.* [7] jointly consider coflow routing and scheduling in data centers. Despite these efforts towards coflow-level optimization, most prior works do not provide any analytical performance guarantee, and there is a lack of fundamental understanding of coflow-level scheduling, especially in the context of large-scale stochastic systems.

Optimal Delay Scaling in Input-Queued Switches. The optimal (*packet-level*) delay scaling in input-queued switches (i.e., the delay scaling under an optimal scheduling policy) has been an important area of research for more than a decade. The randomized scheduling policy [24] (based on Birkhoff-Von Neumann decomposition) achieves an average packet delay of $O(N)$. The well-known Max-Weight Matching (MWM) [26] policy and various approximate MWM algorithms [14], [27] are shown to have an average packet delay no greater than $O(N)$, although it is conjectured that this bound is not tight for a wide range of traffic patterns [27]. Recently, Maguluri *et al.* [15], [16] show that MWM can achieve the optimal $O(1)$ packet-level delay in the *heavy-traffic regime*. Neely *et al.* [21] propose a batching scheme that achieves an average packet delay of $O(\log N)$; this is the best known result for packet-level delay scaling as $N \to \infty$ *under general traffic conditions*. Zhong *et al.* [13] consider the joint scaling of queue length as $N \to \infty$ and $\rho \to 1$. They propose a policy that gives an upper bound of $O(1/(1-\rho)+N^2)$; this joint scaling is shown to be "optimal" in the *heavy-traffic regime* where $\rho = 1 - O(\frac{1}{N^2})$. However, to the best of our knowledge, the optimal scaling of packet-level delay under a general traffic condition is still an open problem in input-queued switches. By comparison, the optimal scaling of *coflow-level delay*, which is an upper bound for packet-level delay, has not been studied before. In this paper, we make the first attempt in deriving the optimal coflow-level delay scaling as $N \to \infty$.

III. Preliminaries

In this section, we briefly introduce some common notations and useful mathematical tools that facilitates our subsequent analysis.

A. Notation

Define $[N] = \{1, \cdots, N\}$. We reserve bold letters for matrices. For example, $\mathbf{X} = (X_{ij})_{N\times N}$ denotes an $N \times N$ matrix with the (i,j)-th element being X_{ij}. For simplicity of notation, we may drop the subscript "$N \times N$" if the context is clear.

We use the traditional asymptotic notations. Let f and g be two functions defined on some subset of real numbers. Then $f(n) = O(g(n))$ if there exists some positive constant C and a real number n_0 such that $|f(n)| \leq C|g(n)|$ for all $n \geq n_0$; similarly, $f(n) = \Omega(g(n))$ if there exists some positive constant C and a real number n_0 such that $f(n) \geq Cg(n)$ for all $n \geq n_0$; finally, $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.

B. Mathematical Tools

(1). Stochastic Dominance. We consider the first-order stochastic dominance whose definition is as follows.

Definition 1 (Stochastic Dominance [22])**.** *Consider two random variables X_1 and X_2. Then X_1 stochastically dominates X_2 if $\mathbb{P}[X_1 \geq x] \geq \mathbb{P}[X_2 \geq x]$ for all $x \in \mathbb{R}$.*

Intuitively, stochastic dominance defines the "inequality relationship" between two random variables in the probabilistic sense. If X_1 stochastically dominates X_2, then X_1 has an equal or higher probability of taking on a large values than X_2. Two useful properties of stochastic dominance are as follows [22].

(P1) Consider two non-negative random variables X_1 and X_2. If X_1 stochastically dominates X_2, then $\mathbb{E}[X_1^n] \geq \mathbb{E}[X_2^n]$ for all $n \in \mathbb{Z}^+$.

(P2) Suppose $\{X_1, \cdots, X_N\}$ is a set of independent random variables and $\{Y_1, \cdots, Y_N\}$ is another set of independent random variables. If X_i stochastically dominates Y_i for all $i \in [N]$, then $\max_i X_i$ also stochastically dominates $\max_i Y_i$.

The above properties will be helpful when establishing inequalities among expectations (and higher moments) of random variables.

(2). Associated Random Variables. The association of random variables is a stronger notion of positive correlation. The formal definition is as follows.

Definition 2 (Associated Random Variables [32])**.** *A collection of random variables $X_1, \cdots, X_n$ are said to be associated if $Cov(f(\mathbf{X}), g(\mathbf{X})) \geq 0$ for all pairs of non-decreasing functions f and g.*

Intuitively, if a collection of random variables are associated, they are usually positively correlated (at least independent). In other words, if X_1 takes on a large value, then $X_2, \cdots, X_n$ are also very likely to take on large values. The followings are some useful properties of associated random variables (see Appendix A of [29]).

(P1) Independent random variables are associated (trivial case).

(P2) Non-decreasing functions of associated random variables are also associated.

(P3) If two sets of associated random variables are independent of one another, then their union is a set of associated random variables.

(P4) If $X_1, \cdots, X_n$ are associated, then $\max_i X_i$ is stochastically dominated by $\max_i X_i'$ where X_i''s are *independent* random variables identically distributed as X_i's.

A particularly important property is (P4) which relates the maximum of a set of (possibly dependent) random variables to the maximum of a set of independent random variables.

IV. System Model

A. Network Model

We consider an $N \times N$ input-queued switch with N input ports and N output ports. The system operates in slotted time, and the slot length is normalized to one unit of time. In each slot, each input can transfer at most one packet and each output can receive at most one packet (this is referred to as "crossbar constraints"). Such an input-queued switch model is simple yet very useful in modeling many practical networked systems. For example, data centers with full bisection bandwidth can be abstracted out as a giant input-queued switch interconnecting different machines. Note that each input port may have packets destined for different output ports, which can be represented as *Virtual Output Queues* (VOQ). There are a total of N^2 virtual output queues, indexed by (i, j) for $i, j \in [N]$, where $[N] \triangleq \{1, \cdots, N\}$ and queue (i, j) holds packets from input i to output j. Figure 1 shows a 2×2 input-queued switch with four virtual output queues.

The schedule of packet transmissions in slot t can be represented by an $N \times N$ matrix $\mathbf{S}(t) = (S_{ij}(t))$ where $S_{ij}(t) = 1$ if the connection between input i and output j is activated. A feasible schedule $\mathbf{S}(t)$ is one that satisfies the crossbar constraints, i.e., $\mathbf{S}(t)$ must be a binary matrix where there is at most one "1" in each row and each column.

B. Coflow Abstraction

A coflow is a collection of parallel data flows belonging to the same job. It has the all-or-nothing property: a coflow is not complete until the completion of all its constituent flows. Coflows are a useful abstraction for many communication patterns in data-intensive computing applications such as MapReduce (see [4] for applications of coflows). Note that the traditional point-to-point communication is a special case of coflows (i.e., a coflow with a single flow).

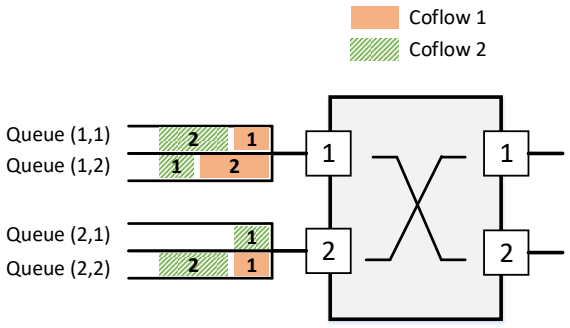

Fig. 1. A 2×2 input-queued switch with two coflow arrivals.

Formally, we represent each coflow by a random traffic matrix $\mathbf{X} = (X_{ij})$ where X_{ij} is the number of packets in this coflow that need to be transmitted from input i to output j. Note that each coflow may contain many small flows from input i to output j, that are aggregated into a single batch X_{ij} for ease of exposition. In the following, X_{ij} will be referred to as the "**batch size**" or "**flow size**" from input i to output j. We assume that all the packets in a coflow are released simultaneously upon the arrival of this coflow. Figure 1 illustrates two coflow arrivals whose traffic matrices are $\mathbf{X}_1 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{X}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Let $\mathbb{E}[X_{ij}] = \beta_{ij}$ and assume that coflows arrive to the system with rate λ. Then the arrivals of packets to queue (i, j) is a *batch arrival process* with rate λ and mean batch size β_{ij}. Also define $\beta \triangleq \max\left(\max_i \sum_j \beta_{ij}, \max_j \sum_i \beta_{ij}\right)$. In our analysis, we also make the following assumptions:

(1) The arrival times and the batch sizes of different coflows are independent.

(2) X_{ij}'s are independent random variables.

(3) If the n-th moment of X_{ij} is finite, we assume $\mathbb{E}[(\sum_j X_{ij})^n] = O(1)$ for all $i \in [N]$ and $\mathbb{E}[(\sum_i X_{ij})^n] = O(1)$ for all $j \in [N]$ as $N \to \infty$.

(4) The sub-critical condition holds: $\rho = \lambda\beta < 1$.

In this paper, we focus on optimizing *coflow-level delay*, i.e., *the time between the arrival of a coflow until all the packets associated with this coflow are transmitted*. In particular, we are interested in the scaling of coflow-level delay in a large-scale system, as $N \to \infty$. Our objective is to find a scheduling policy that achieves the best dependence of coflow-level delay on N while stabilizing the system whenever $\rho < 1$.

V. Lower Bounds on Coflow-level Delay

Before we investigate any specific scheduling policy, it is useful to study the fundamental scaling properties of coflow-

level delay as $N \to \infty$. In this section, we develop lower bounds on coflow-level delay in input-queued switches. These lower bounds serve as the baselines when we evaluate the coflow-level performance of a scheduling policy. We first introduce the notion of *clearance time*.

Definition 3 (Clearance Time)**.** *The clearance time of a coflow* $\mathbf{X} = (X_{ij})$ *is*

$$\tau(\mathbf{X}) = \max\Big(\max_i \sum_j X_{ij}, \max_j \sum_i X_{ij}\Big). \tag{1}$$

Clearly, $\tau(\mathbf{X})$ is the maximum number of packets in a row or a column of $\mathbf{X}$. Since each input/output port can process at most one packet per slot, the minimum time to clear all the packets in $\mathbf{X}$ must be no smaller than $\tau(\mathbf{X})$. In fact, $\mathbf{X}$ can be cleared in exactly $\tau(\mathbf{X})$ slots by using the *optimal clearance algorithm* described in [17]. As a result, $\tau(\mathbf{X})$ is the minimum time needed to transmit all the packets in a coflow $\mathbf{X}$. In the rest of this section, we investigate the scaling of clearance time as $N \to \infty$ and its relationship to coflow-level delay.

Depending on the distributions of X_{ij}'s (the flow sizes), the scaling of clearance time exhibits different behaviors. First, we consider the general case where the flow size distribution is arbitrary (as long as $\mathbb{E}[X_{ij}] < \infty$ for all $i, j \in [N]$).

Lemma 1. *For a coflow* $\mathbf{X} = (X_{ij})$ *with* $\mathbb{E}[X_{ij}] < \infty$*, the expected clearance time is* $\mathbb{E}[\tau(\mathbf{X})] = O(N)$*. Moreover, there exist distributions of* X_{ij}*'s such that* $\mathbb{E}[\tau(\mathbf{X})] = \Omega(N^{\frac{1}{1+\epsilon}})$ *for any* $\epsilon > 0$*.*

Proof. The $O(N)$ upper bound is nearly trivial. It is clear that

$$\mathbb{E}[\tau(\mathbf{X})] \le \mathbb{E}[\sum_{i,j} X_{ij}] \le N\beta = O(N).$$

To prove the lower bound, we find some distributions of X_{ij}'s such that $\mathbb{E}[\tau(\mathbf{X})] = \Omega(N^{\frac{1}{1+\epsilon}})$ for any $\epsilon > 0$. Consider the scenario where $\mathbf{X}$ is a diagonal matrix: $X_{ij} = 0$ with probability 1 for $i \neq j$ and X_{ii} has the power law for all $i \in [N]$, i.e.,

$$\mathbb{P}[X_{ii} \ge k] = \frac{1}{k^{1+\epsilon}},\ k = 1, 2, \cdots,$$

where $\epsilon > 0$. Note that $\mathbb{E}[X_{ii}] = \frac{1+\epsilon}{\epsilon}$ but it can be easily scaled or shifted to have an arbitrary expectation. By simple calculation on the order statistics, we have

$$\begin{aligned}
\mathbb{E}[\tau(\mathbf{X})] &\ge \mathbb{E}[\max_i \sum_j X_{ij}] \\
&= \mathbb{E}[\max_i X_{ii}] \\
&= \sum_{k=1}^{\infty} \Big[1 - (1 - \frac{1}{k^{1+\epsilon}})^N\Big] \\
&\ge \sum_{k=1}^{N_\epsilon} \Big[1 - \Big(1 - \frac{1}{N_\epsilon^{1+\epsilon}}\Big)^N\Big] \\
&\ge N_\epsilon \Big[1 - \Big(1 - \frac{1}{N}\Big)^N\Big],
\end{aligned}$$

where $N_\epsilon = \lfloor N^{\frac{1}{1+\epsilon}} \rfloor$. Since $1 - \Big(1 - \frac{1}{N}\Big)^N \xrightarrow{N\to\infty} 1 - e^{-1}$, we can conclude that $N_\epsilon \Big[1 - \Big(1 - \frac{1}{N}\Big)^N\Big] = \Theta(N_\epsilon) = \Theta(N^{\frac{1}{1+\epsilon}})$ as $N \to \infty$. As a result, there exist (heavy-tailed) distributions of X_{ij}'s such that $\mathbb{E}[\tau(\mathbf{X})] = \Omega(N^{\frac{1}{1+\epsilon}})$ for any $\epsilon > 0$. □

If X_{ij} also has a finite variance, we can obtain a better scaling behavior of clearance time as $N \to \infty$. In this case, we assume $\text{Var}(\sum_j X_{ij}) \le \sigma^2$ for all $i \in [N]$ and $\text{Var}(\sum_i X_{ij}) \le \sigma^2$ for all $j \in [N]$, where σ^2 is a constant independent of N.

Lemma 2. *For a coflow* $\mathbf{X} = (X_{ij})$ *with Var*$(X_{ij}) < \infty$ *for all* $i, j \in [N]$*, the expected clearance time is* $\mathbb{E}[\tau(\mathbf{X})] = O(\sqrt{N})$*. Moreover, there exist distributions of* X_{ij}*'s such that* $\mathbb{E}[\tau(\mathbf{X})] = \Omega(N^{\frac{1}{2+\epsilon}})$ *for any* $\epsilon > 0$*.*

Proof. Devroye [28] shows that if $Y_1, \cdots, Y_n$ are (possibly dependent) random variables with finite means and finite variances, then

$$\mathbb{E}[\max_i Y_i] \le \max_i \mathbb{E}[Y_i] + \sqrt{n} \max_i \sqrt{\text{Var}(Y_i)}.$$

If $\text{Var}(X_{ij}) < \infty$ for all $i, j \in [N]$, it follows that

$$\begin{aligned}
\mathbb{E}[\max_i \sum_j X_{ij}] &\le \max_i \mathbb{E}[\sum_j X_{ij}] + \sqrt{N} \max_i \sqrt{\text{Var}(\sum_j X_{ij})} \\
&\le \beta + \sqrt{N}\sigma.
\end{aligned}$$

Similarly, it can be shown that $\mathbb{E}[\max_j \sum_i X_{ij}] \le \beta + \sqrt{N}\sigma$. As a result, we have

$$\mathbb{E}[\tau(\mathbf{X})] \le \mathbb{E}[\max_i \sum_j X_{ij}] + \mathbb{E}[\max_j \sum_i X_{ij}] \le 2\beta + 2\sqrt{N}\sigma.$$

Consequently, $\mathbb{E}[\tau(\mathbf{X})] = O(\sqrt{N})$. The $\Omega(N^{\frac{1}{2+\epsilon}})$ lower bound can be proved in a similar way to Lemma 1 with the power being $2 + \epsilon$ instead of $1 + \epsilon$ such that the variance is finite. □

Furthermore, if X_{ij}'s have light-tailed[2] distributions, the scaling of clearance time is logarithmic.

Lemma 3. *For a coflow* $\mathbf{X} = (X_{ij})$ *with light-tailed* X_{ij}*'s, the expected clearance time is* $\mathbb{E}[\tau(\mathbf{X})] = O(\log N)$*. Moreover, there exist distributions of* X_{ij}*'s such that this bound is tight.*

Proof. See Appendix A. □

Finally, if X_{ij}'s are deterministic, it is clear that $\mathbb{E}[\tau(\mathbf{X})] \le \beta = \Theta(1)$. Since clearance time is the minimum time to transmit all the packets in a coflow, it is a natural lower bound on coflow-level delay (which is the time between the arrival of a coflow until all the packets associated with this coflow are transmitted). Consequently, the above results essentially impose fundamental limits on the coflow-level delay that can be achieved by any scheduling policy.

[2]In this paper, a light-tailed distribution is the one with a finite Moment Generating Function in the neighborhood of 0. In other words, it has an exponentially decreasing tail.

TABLE I
LOWER BOUNDS ON COFLOW-LEVEL DELAY

Condition	Coflow-level Delay
$\mathbb{E}[X_{ij}] < \infty$	$\Omega(N^{\frac{1}{1+\epsilon}})$ for any $\epsilon > 0$
$\text{Var}(X_{ij}) < \infty$	$\Omega(N^{\frac{1}{2+\epsilon}})$ for any $\epsilon > 0$
X_{ij}'s are light-tailed	$\Omega(\log N)$
X_{ij}'s are deterministic	$\Omega(1)$

Theorem 1. *The expected coflow-level delay achieved by any scheduling policy cannot be better than $O(g(N))$, where*
(1) $g(N) = N^{\frac{1}{1+\epsilon}}$ for any $\epsilon > 0$ if $\mathbb{E}[X_{ij}] < \infty$;
(2) $g(N) = N^{\frac{1}{2+\epsilon}}$ for any $\epsilon > 0$ if Var$(X_{ij}) < \infty$;
(3) $g(N) = \log N$ if X_{ij}'s have light-tailed distributions;
(4) $g(N) = 1$ if X_{ij}'s are deterministic.

The scaling properties of expected coflow-level delay are summarized in Table I. It can be observed that the lower bound on coflow-level delay critically depends on the *variability* of X_{ij}'s: the less X_{ij}'s vary, the smaller lower bound on coflow-level delay we can obtain.

In the rest of this paper, we mainly focus on the case where X_{ij}'s have light-tailed distributions unless otherwise stated. The heavy-tailed case is left for future work.

VI. COFLOW-AGNOSTIC SCHEDULING

To gain further insights into the design of coflow-level scheduling policies, we study the performance of some *coflow-agnostic* scheduling policies where coflow-level information (e.g., which packets/flows belong to the same coflow) is not leveraged. In particular, we study the coflow-level performance of two simple scheduling policies: randomized scheduling and periodic scheduling.

Randomized Scheduling. Let $M_1, \cdots, M_{N!}$ be the $N!$ perfect matchings (permutation matrices) associated with the $N \times N$ switch. With the Birkhoff-Von Neumann decomposition, we can find probabilities $\{p_1, p_2, \cdots, p_{N!}\}$ such that the matrix $(\lambda\beta_{ij}) \leq \sum_{k=1}^{N!} p_k M_k$ (where $\beta_{ij} = \mathbb{E}[X_{ij}]$). Such a decomposition is always feasible since $(\lambda\beta_{ij})$ is sub-stochastic by Assumption (4) in Section IV-B. In each slot, the randomized policy uses matching M_k as the schedule, with probability p_k. Under uniform traffic, a simple way to implement the randomized policy is to connect the N input ports with a random permutation of the N output ports. Such a policy is guaranteed to stabilize the network as long as $\rho < 1$ although λ and (β_{ij}) need to be known in advance. The detailed description of this policy can be found in [24] and it can be easily shown that the randomized policy achieves $O(N)$ average *packet-level* delay [21].

Periodic Scheduling. This policy is similar to randomized scheduling except that the scheduling decisions are deterministic. Specifically, for some (sufficiently long) period T, we use matching M_k for exactly $p_k T$ times every T slots. Under uniform traffic, a simple way to implement periodic scheduling is to connect each input port i to output port $[(i+t) \bmod N]+1$ in slot t. This policy also achieves $O(N)$ average packet-level delay whenever $\rho < 1$ [21].

Now we analyze the coflow-level delay achieved by the above two policies. In contrast to the simple analysis of packet-level delay, it is non-trivial to analyze the coflow-level delay achieved by these policies, due to the correlation between packets (e.g., packets belonging to the same coflow arrive simultaneously). For ease of exposition, we assume that traffic is uniform such that $\mathbb{E}[X_{ij}] = \frac{\beta}{N}$ and $\text{Var}(X_{ij}) = \frac{\sigma^2}{N}$ for all $i, j \in [N]$. We also assume that coflow arrivals are Poisson. The analysis can be easily extended to the general case.

Theorem 2. *Suppose X_{ij}'s have light-tailed distributions. The expected coflow-level delay achieved by the randomized or periodic scheduling policy is $O(N \log N)$ whenever $\rho < 1$.*

Proof. See Appendix B. □

Remark 1. The proof to Theorem 2 also shows that the average coflow-level delay achieved by the randomized or periodic policy is $O(\frac{N \log N}{1-\rho})$ as $\rho \to 1$ and $N \to \infty$.

Remark 2. Comparing with the $O(N)$ packet-level delay, we can observe a coflow-level delay "dilation" factor of $O(\log N)$. Intuitively, the delay dilation is due to the additional "assembly delay": packets processed earlier must wait for packets (in the same coflow) that are processed later.

Finally, it is worth mentioning that the randomized or periodic scheduling policy is the simplest throughput-optimal policy in input-queued switches, but it sheds light on the non-triviality of coflow-level analysis (e.g., the correlation between packets) and the potential weakness of coflow-agnostic algorithms (as can be seen from the coflow-level delay dilation). The coflow-level analysis of more sophisticated policies, such as MaxWeight Matching (MWM) scheduling, are very challenging and left for future work. In fact, even the *packet-level* delay of MWM is still an open problem [15], [16].

In conclusion, there has been no throughput-optimal scheduling policy that achieves the provably optimal scaling with respect to coflow-level delay. In the next section, we propose a new coflow-aware scheduling policy that achieves the optimal scaling of coflow-level delay while maintaining throughput optimality and requiring no traffic statistics.

VII. COFLOW-AWARE SCHEDULING

In this section, we develop a coflow-aware scheduling policy that achieves $O(\log N)$ expected coflow-level delay whenever $\rho < 1$ under the assumption that arrivals of coflows are Poisson and flow sizes X_{ij}'s are light-tailed. The policy is called the Coflow-Aware Batching (CAB) scheme. Note that in Section V, we showed that if X_{ij}'s are light-tailed, no scheduling policy can achieve an expected coflow-level delay better than $O(\log N)$. As a result, the CAB policy attains the lower bound, which implies that $O(\log N)$ is optimal scaling of coflow-level delay (under Poisson arrivals and light-tailed flow sizes).

A. Coflow-Aware Batching (CAB) Policy

The basic idea of the CAB policy is to group timeslots into frames of size T slots and clear coflows in batches, where

one batch of coflows correspond to the collection of coflows arriving in the same frame. Coflows that are not cleared during a frame are handled separately in future frames. By properly setting the frame size T, the CAB policy can achieve the desirable $O(\log N)$ average coflow-level delay. Note that Neely *et al.* [21] proposed a similar batching scheme to reduce *packet-level* delay. By comparison, our CAB policy explicitly leverages coflow-level information (e.g., which packets belong to the same coflow) to reduce *coflow-level* delay. More importantly, as mentioned in Section VI, coflow-level delay analysis is fundamentally different from packet-level analysis. The detailed description of the CAB policy is as follows.

Coflow-Aware Batching (CAB) Scheduling Policy

Setup.

- Timeslots are grouped into frames of size T slots.

Notation.

- Denote by $\mathbf{L}(k) = \big(L_{ij}(k)\big)$ the aggregate traffic matrix of all the coflows arriving in the k-th frame, where $L_{ij}(k)$ is the total number of packets from input i to output j that arrive during the k-th frame.

Procedures.

(1) In the $(k+1)$-th frame, we try to clear the coflows that arrived in the k-th frame, i.e., $\mathbf{L}(k)$. Let $\mathbf{B}(k+1)$ be the traffic matrix we choose to clear in the $(k+1)$-th frame (which may be less than $\mathbf{L}(k)$), and denote by τ the clearance time of $\mathbf{L}(k)$. If $\tau \leq T-1$, then $\mathbf{L}(k)$ can be cleared within the first $T-1$ slots in the $(k+1)$-th frame. In this case, we just set $\mathbf{B}(k+1) = \mathbf{L}(k)$. Note that we only use the first $T-1$ slots in a frame while the remaining slot is reserved for clearing "overflow" coflows as discussed below. If $\tau > T-1$, then *overflow* occurs and only a subset of $\mathbf{L}(k)$ can be cleared in the $(k+1)$-th frame. In this case, we sequentially add coflows to $\mathbf{B}(k+1)$ in order of their arrival in the k-th frame until $\mathbf{B}(k+1)$ becomes maximal, i.e., adding any other coflow will make the clearance time of $\mathbf{B}(k+1)$ exceed $T-1$. If a coflow is selected to $\mathbf{B}(k+1)$, it is referred to as a ***conforming coflow*** otherwise it is called a ***non-conforming coflow***.

(2) All the conforming coflows that arrive in the k-th frame are scheduled during the $(k+1)$-th frame by clearing $\mathbf{B}(k+1)$ in minimum time using an optimal clearance algorithm (e.g., see [17]).

(3) All the non-conforming coflows are put into a separate FIFO queue. In the last slot of each frame, this FIFO queue gets served by the switch. Note that non-conforming coflows are served one at a time, and the service time (measured in the number of *frames*) of each non-conforming coflow is its clearance time.

In words, the first $T-1$ slots in a frame are used to serve conforming coflows arriving in the previous frame and the remaining slot is reserved to serve non-conforming coflows in a FIFO manner. Note that conforming coflows (that arrive in the same frame) are cleared together in a batch while non-conforming coflows are served one at a time in the separate FIFO queue. Under the CAB policy, either all the packets in a coflow are conforming or none of them are conforming.

B. Performance of the CAB policy

The following theorem shows that the CAB policy achieves $O(\log N)$ expected coflow-level delay whenever $\rho < 1$ (under Poisson coflow arrivals and light-tailed flow sizes).

Theorem 3 (Average Coflow-level Delay)**.** *Suppose coflows arrive according to a Poisson process and flow sizes are light-tailed. By selecting a proper frame size $T = O(\log N)$, the CAB policy achieves $O(\log N)$ expected coflow-level delay if $\rho < 1$.*

The choice of T will be specified later in Section VII-C. In fact, the CAB policy not only guarantees that the *average* coflow-level delay is $O(\log N)$ but also ensures that the $O(\log N)$ delay is achievable for an arbitrary coflow with high probability.

Corollary 1 (Tail Coflow-level Delay)**.** *By selecting a proper frame size $T = O(\log N)$, the CAB policy achieves $O(\log N)$ delay for an arbitrary coflow with probability $1 - O(\frac{1}{N^2})$ whenever $\rho < 1$.*

In the following, we present a proof for the above results. The proof itself suggests the choice of T.

C. Proof to Theorem 3 and Corollary 1

For simplicity, we assume that traffic is uniform such that $\mathbb{E}[X_{ij}] = \frac{\beta}{N}$ and $\text{Var}(X_{ij}) = \frac{\sigma^2}{N}$ for all $i, j \in [N]$. The analysis can be easily extended to non-uniform traffic.

We discuss the expected coflow-level delay experienced by a conforming and a non-conforming coflow, respectively.

- Conforming coflows are cleared within 2 frames: the frame where they arrive plus the frame where they are cleared. As a result, the coflow-level delay experienced by a conforming coflow is at most $2T$ time slots, i.e.,

$$\text{Delay(conforming)} \leq 2T.$$

- A non-conforming coflow first waits for at most T slots (the frame where it arrives) and then waits in the separate FIFO queue. As a result, the coflow-level delay experienced by a non-conforming coflow is

$$\text{Delay(non-conforming)} \leq T + \text{Delay(FIFO queue)}.$$

Let η be the long-term fraction of non-conforming coflows. Then the average coflow-level delay of an *arbitrary coflow* is

$$\mathbb{E}[W] \leq (1-\eta)2T + \eta[T + \text{Delay(FIFO queue)}]. \quad (2)$$

In the following, we choose $T = O(\log N)$ (the specific value of T will be made clear later). Under such a choice of T, we prove that η is miniscule and Delay(FIFO queue) is not very large. Thus, it can be concluded that $\mathbb{E}[W] = O(\log N)$.

Step 1: Determine the value of T.

We first show that the overflow probability decreases exponentially with the frame size T.

Lemma 4. *Let P_o be the overflow probability in an arbitrary frame. If $\rho < 1$, there exists some constant $\gamma > 0$ such that*

$$P_o \leq 2N \exp(-\gamma T). \tag{3}$$

Proof. Note that an overflow occurs in the k-th frame if the clearance time of $\mathbf{L}(k)$ is greater than $T-1$ slots, i.e., if any of the following $2N$ inequalities is violated.

$$\begin{aligned} \sum_j L_{ij}(k) < T,\ i = 1, \cdots, N, \\ \sum_i L_{ij}(k) < T,\ j = 1, \cdots, N, \end{aligned} \tag{4}$$

where $L_{ij}(k)$ is the number of packets that arrive to queue (i,j) during the k-th frame. Clearly, $\sum_j L_{ij}(k)$ (or $\sum_i L_{ij}(k)$) is the total number of packets that arrive to input i (or output j) during the k-th frame, which corresponds to the number of packet arrivals during T time slots in a *Poisson process with batch arrivals* where the arrival rate is λ and the mean batch size is β.

Let $Y(T)$ be the total number of packet arrivals during T time slots in the above *batch Poisson process*. It is clear that $Y(T) = \sum_{n=1}^{N(T)} B_n$. Here, $N(T)$ is the number of coflow arrivals during the T time slots, and has a Poisson distribution with rate λT; B_n is the number of packets brought by the n-th coflow to a certain input or output port, which is identically distributed as $\sum_j X_{ij}$ or $\sum_i X_{ij}$ and $\mathbb{E}[B_n] = \beta$. By Wald's equality, we have $\mathbb{E}[Y(T)] = \mathbb{E}[N(T)]\mathbb{E}[B_1] = \rho T < T$ if $\rho < 1$. Suppose the Moment Generating Function (MGF) of B_n is $M_B(s)$, and the MGF of $Y(T)$ is $M_Y(s)$. By the property of a Poisson process with batch arrivals, we have $M_Y(s) = \exp\{\lambda T[M_B(s) - 1]\}$. By the Chernoff bound, we have for any $s > 0$,

$$\begin{aligned} \mathbb{P}[Y(T) \geq T] &\leq M_Y(s) e^{-sT} \\ &= \exp\Big\{\lambda T[M_B(s) - 1] - sT\Big\} \\ &= \exp\big(-f(s)T\big), \end{aligned} \tag{5}$$

where

$$f(s) = \lambda[1 - M_B(s)] + s. \tag{6}$$

It is clear that $f(0) = 0$ and $f'(0) = -\lambda M_B'(0) + 1 = 1 - \lambda\mathbb{E}[B] = 1 - \rho > 0$. As a result, there exists a sufficiently small $\epsilon > 0$ such that $f(\epsilon) > 0$. Let $\gamma \triangleq f(\epsilon) > 0$. By the same argument as in the proof to Lemma 9 (see Appendix A), it can be verified that γ is a constant independent of N under our assumptions on X_{ij}'s. Then we have

$$\mathbb{P}[Y(T) \geq T] \leq \exp(-\gamma T).$$

Applying the union bound, we can conclude that the overflow probability (i.e., at least one of the $2N$ inequalities in (4) is violated) is bounded by

$$P_o \leq 2N \exp(-\gamma T),$$

which completes the proof. □

Remark 1. Lemma 4 implies that if we want to keep the overflow probability below δ, we can choose $T \geq \frac{\log(2N/\delta)}{\gamma}$, where $\gamma > 0$ is some constant independent of N. Since T is an integer, we can choose

$$T = \lceil \frac{\log(2N/\delta)}{\gamma} \rceil. \tag{7}$$

The value of δ will be specified later such that $T = O(\log N)$ and the average coflow-level delay is also $O(\log N)$. The constant γ can also be found in a systematic way (see Section VII-F for details).

Remark 2. Lemma 4 holds only if $\rho < 1$. If $\rho \geq 1$, $\mathbb{E}[Y(T)] = \rho T \geq T$ and the Chernoff bound (5) does not hold.

Step 2: Determine the value of η.

Next, we derive an upper bound for the long-term fraction of non-conforming coflows, i.e.,

$$\eta = \lim_{K \to \infty} \frac{\sum_{k=1}^{K} A_{non}(k)}{\sum_{k=1}^{K} A(k)},$$

where $A(k)$ is the total number of coflows that arrive during the k-th frame and $A_{non}(k)$ is the number of non-conforming coflows in the k-th frame. We begin by identifying a stochastic bound for the number of coflow arrivals in an **overflow** frame.

Lemma 5. *Suppose $N(T)$ is a Poisson random variable with rate λT. Given that an overflow occurs in a frame, the number of coflows arrivals in this frame is stochastically dominated by $N(T) + T/\beta$ when T is sufficiently large.*

Proof. Let $N(T)$ be the total number of coflow arrivals in an arbitrary frame of T time slots. Clearly, $N(T)$ is a Poisson random variable with rate λT. Denote by $\mathbf{X}^{(n)} = (X_{ij}^{(n)})$ the traffic matrix of the n-th coflow, and let $\mathbf{L}$ be the aggregate traffic matrix of these coflows, i.e., $\mathbf{L} = \sum_{n=1}^{N(T)} \mathbf{X}^{(n)}$. It is clear that an overflow occurs if the clearance time of $\mathbf{L}$ is greater than $T-1$, i.e., $\tau(\mathbf{L}) \geq T$. We first find a lower bound on the overflow probability when T is sufficiently large.

Claim 1. *There exists some $T_0 > 0$ such that for any $T > T_0$ the overflow probability*

$$\mathbb{P}\Big[\tau(\mathbf{L}) \geq T\Big] \geq \mathbb{P}\Big[N(T) \geq T/\beta\Big].$$

Proof. First notice that

$$\tau(\mathbf{L}) \geq \max_i \sum_j L_{ij} = \max_i \sum_{n=1}^{N(T)} \sum_j X_{ij}^{(n)}, \tag{8}$$

where the last equality is due to the fact that $L_{ij} = \sum_{n=1}^{N(T)} X_{ij}^{(n)}$. As a result, for any $i \in [N]$

$$\mathbb{P}[\tau(\mathbf{L}) \geq T] \geq \mathbb{P}[\sum_{n=1}^{N(T)} \sum_j X_{ij}^{(n)} \geq T]. \tag{9}$$

By elementary probability calculation, we can derive

$$\mathbb{E}[\sum_{n=1}^{N(T)}\sum_j X_{ij}^{(n)}] = \beta\lambda T = \rho T$$

$$\text{Var}(\sum_{n=1}^{N(T)}\sum_j X_{ij}^{(n)}) = \lambda T(\beta^2+\sigma^2),$$

where $\sigma^2 \geq 0$ is the variance of $\sum_j X_{ij}^{(n)}$. Note that when $\sigma^2 = 0$, then $\sum_j X_{ij}^{(n)} = \beta$ with probability 1 for any $i \in [N]$, which implies that

$$\mathbb{P}[\tau(\mathbf{L}) \geq T] = \mathbb{P}[\beta N(T) \geq T] = \mathbb{P}[N(T) \geq T/\beta].$$

Therefore, we only need to consider the case where $\sigma^2 > 0$.

Note that $\lim_{T\to\infty} N(T) = \infty$. By Central Limit Theorem, we have

$$\lim_{T\to\infty} \mathbb{P}[\sum_{n=1}^{N(T)}\sum_j X_{ij}^{(n)} \geq T] = 1 - \Phi\Big(\frac{T-\rho T}{\sqrt{\lambda T(\beta^2+\sigma^2)}}\Big) \triangleq p.$$

At the same time, since $N(T)$ is a Poisson random variable with mean λT, we can use normal approximation to Poisson distribution:

$$\lim_{T\to\infty} \mathbb{P}[N(T) \geq T/\beta] = 1 - \Phi\Big(\frac{T-\rho T}{\sqrt{\lambda T\beta^2}}\Big) \triangleq p'.$$

As a result, for any $\epsilon > 0$, there exist a sufficiently large T_0 such that for any $T > T_0$

$$\begin{aligned} \Big|\mathbb{P}[\sum_{n=1}^{N(T)}\sum_j X_{ij}^{(n)} \geq T] - p\Big| &\leq \epsilon \\ \Big|\mathbb{P}[N(T) \geq T/\beta] - p'\Big| &\leq \epsilon. \end{aligned} \tag{10}$$

Since $p > p'$ (note that $\sigma^2 > 0$), we can choose $\epsilon = \frac{p-p'}{2} > 0$ and conclude that for any sufficiently large $T > T_0$

$$\begin{aligned} &\mathbb{P}[\tau(\mathbf{L}) \geq T] - \mathbb{P}\Big[N(T) \geq T/\beta\Big] \\ \geq &\mathbb{P}\Big[\sum_{n=1}^{N(T)}\sum_j X_{ij}^{(n)} \geq T\Big] - \mathbb{P}\Big[N(T) \geq T/\beta\Big] \\ \geq &p - \epsilon - (p' + \epsilon) = 0, \end{aligned}$$

where the first inequality is due to (9) and the second inequality is due to (10). □

Next we evaluate the probability that there are at least m coflow arrivals in an overflow frame.

Claim 2. *Given that an overflow occurs in a frame, the probability that there are at least m coflow arrivals in this frame is upper bounded by $\mathbb{P}[N(T) \geq m | N(T) \geq T/\beta]$ when T is sufficiently large, i.e.,*

$$\mathbb{P}\Big[N(T) \geq m \Big| \tau(\mathbf{L}) \geq T\Big] \leq \mathbb{P}\Big[N(T) \geq m \Big| N(T) \geq T/\beta\Big].$$

Proof. If $m \leq T/\beta$, then $\mathbb{P}[N(T) \geq m | N(T) \geq T/\beta] = 1$ and the upper bound naturally holds.

If $m > T/\beta$, then

$$\begin{aligned} \mathbb{P}\Big[N(T) \geq m \Big| N(T) \geq \frac{T}{\beta}\Big] &= \frac{\mathbb{P}[N(T) \geq m, N(T) \geq \frac{T}{\beta}]}{\mathbb{P}[N(T) \geq \frac{T}{\beta}]} \\ &= \frac{\mathbb{P}[N(T) \geq m]}{\mathbb{P}[N(T) \geq \frac{T}{\beta}]} \\ &\geq \frac{\mathbb{P}[N(T) \geq m, \tau(\mathbf{L}) \geq T]}{\mathbb{P}[\tau(\mathbf{L}) \geq T]} \\ &= \mathbb{P}[N(T) \geq m | \tau(\mathbf{L}) \geq T], \end{aligned}$$

where the first equality is due to the rule of conditional probability, the second equality holds because $\mathbb{P}[N(T) \geq m, N(T) \geq \frac{T}{\beta}] = \mathbb{P}[N(T) \geq m]$ when $m > T/\beta$, and the third inequality is due to Claim 1 and the fact that $\mathbb{P}[N(T) \geq m] \geq \mathbb{P}[N(T) \geq m, \tau(\mathbf{L}) \geq T]$. □

Claim 3. *If $N(T)$ is a Poisson random variable, then $\mathbb{P}[N(T) \geq m + r | N(T) \geq r] \leq \mathbb{P}[N(T) \geq m]$.*

This claim was proved in Appendix B of [21].

The above claims imply that when T is sufficiently large

$$\begin{aligned} \mathbb{P}\Big[N(T) \geq m \Big| \tau(\mathbf{L}) \geq T\Big] &\leq \mathbb{P}\Big[N(T) \geq m \Big| N(T) \geq \frac{T}{\beta}\Big] \\ &\leq \mathbb{P}\Big[N(T) \geq m - \frac{T}{\beta}\Big] \\ &= \mathbb{P}\Big[N(T) + \frac{T}{\beta} \geq m\Big], \end{aligned}$$

where the first inequality is due to Claim 2 and the second inequality is due to Claim 3. The above inequalities imply that given an overflow occurs in a frame, the number of coflow arrivals in this frame is stochastically dominated by $N(T)+T/\beta$ when T is sufficiently large. This completes the proof of Lemma 5. □

With Lemma 5, we can find an upper bound for the long-term fraction of non-conforming coflows.

Lemma 6. *If the overflow probability is δ, the long-term fraction of non-conforming coflows among all coflows is upper bounded by $\eta \leq \frac{2\delta}{\rho}$ when the frame size T is sufficiently large.*

Proof. According to Lemma 5, the expected number of non-conforming coflows in an overflow frame is at most $\lambda T + T/\beta$ (note that this is a loose bound since we treat all the coflows in an overflow frame as non-conforming coflows). As a result, the long-term fraction of non-conforming coflows is

$$\begin{aligned} \eta = \lim_{K\to\infty} \frac{\sum_{k=1}^K A_{non}(k)}{\sum_{k=1}^K A(k)} &\leq \lim_{K\to\infty} \frac{\delta K(\lambda T + T/\beta)}{K\lambda T} \\ &= \frac{\delta(\lambda T + T/\beta)}{\lambda T}. \end{aligned}$$

Note that the inequality holds because the number of overflow frames is δK as $K \to \infty$ and the average number of non-conforming coflows in each overflow frame is at most $\lambda T + T/\beta$. Noticing that $T \geq 1$ and $\rho = \lambda\beta < 1$, we have

$$\eta \leq \frac{\delta(\lambda T + T/\beta)}{\lambda T} = \delta(1 + \frac{T}{\rho T}) \leq \delta(\frac{1}{\rho} + \frac{1}{\rho}) = \frac{2\delta}{\rho},$$

which completes the proof. □

Step 3: Determine delay in the separate FIFO queue.

The third step is to find an upper bound for the average delay experienced by non-conforming coflows in the separate FIFO queue. Note that Lemma 4 shows that whenever $\rho < 1$, the overflow probability δ can be made arbitrarily small by setting the frame size T as in (7). In particular, we can choose the frame size to be $T = \lceil \frac{\log(2N/\delta)}{\gamma} \rceil = O(\log N^3) = O(\log N)$ to achieve the overflow probability $\delta = O(\frac{1}{N^2})$. Under such a choice of T, we can prove the following lemma.

Lemma 7. *Under a proper choice of T, the FIFO queue holding non-conforming coflows is stable whenever $\rho < 1$ and the average delay experienced by non-conforming coflows in the FIFO queue is $O(NT^3)$.*

Proof. Note that non-conforming coflows are placed in a discrete-time FIFO queue and are served one at a time. At the end of each frame, a batch of non-conforming coflows arrive to this queue, with probability δ. Since non-conforming coflows arrive to the FIFO queue in batches, the arrivals of non-conforming coflows are dependent. To circumvent this dependence, we notice that the arrivals of different batches of non-conforming coflows are independent: in each frame, there is a batch arrival with probability δ, independent of any other frames. As a result, we overestimate the delay of any *individual* non-conforming coflow by the delay experienced by the *entire batch* of non-conforming coflows (i.e., the time between the arrival of the entire batch and the completion of all the non-conforming coflows in this batch) plus T slots (the size of the frame in which the non-conforming coflow arrive).

As an overestimate, all the coflows that arrive in an overflow frame are treated as non-conforming coflows. Let M be the total number of non-conforming coflows in an overflow frame, and denote by $\tilde{\mathbf{X}}^{(n)} = (\tilde{X}_{ij}^{(n)})$ the traffic of the n-th non-conforming coflow in this overflow frame. Let $\tilde{\mathbf{L}} = \sum_{n=1}^{M} \tilde{\mathbf{X}}^{(n)}$ be the aggregate traffic matrix associated with these non-conforming coflows. It follows that the service time for the entire batch of non-conforming coflows is $U = \tau(\tilde{\mathbf{L}})$ (measured in the number of frames since only one slot per frame is used to serve non-conforming coflows). Clearly, the service times for different batches of non-conforming coflows are independent. As a result, if the entire batch of non-conforming coflows is treated as a "customer", the FIFO queue is a discrete-time GI/GI/1 queue with Bernoulli arrivals (of rate δ per frame) and general service time U. The average waiting time (measured in the number of frames) in such a system can be exactly characterized [34]:

$$\text{Delay(FIFO queue)} = \frac{\delta \mathbb{E}[U^2] - \delta \mathbb{E}[U]}{2(1 - \delta \mathbb{E}[U])} + \mathbb{E}[U]. \tag{11}$$

Now we evaluate $\mathbb{E}[U]$ and $\mathbb{E}[U^2]$. Without loss of generality, we assume $\max_i \sum_j \tilde{L}_{ij} \geq \max_j \sum_i \tilde{L}_{ij}$ so that $U = \tau(\tilde{\mathbf{L}}) = \max_i \sum_j \tilde{L}_{ij}$ (this simplification does not influence the scaling as $N \to \infty$). Note that $\tilde{L}_{ij} = \sum_{n=1}^{M} \tilde{X}_{ij}^{(n)}$. Thus, we have

$$U = \tau(\tilde{\mathbf{L}}) = \max_i \sum_{n=1}^{M} \sum_j \tilde{X}_{ij}^{(n)}.$$

By Lemma 5, M is stochastically dominated by $N(T) + T/\beta$ where $N(T)$ is a Poisson random variable with rate λT. By Lemma 12 (see Appendix C), $\sum_j \tilde{X}_{ij}^{(n)}$ is stochastically dominated by $\sum_j X_{ij} + T$ where (X_{ij}) is the traffic of an arbitrary coflow in an arbitrary frame. As a result, we have

$$\begin{aligned} \mathbb{E}[U] &\leq \sum_i \mathbb{E}[M] \mathbb{E}\Big[\sum_j \tilde{X}_{ij}^{(n)}\Big] \\ &\leq N(\lambda T + T/\beta)(\beta + T) = \Theta(NT^2), \end{aligned} \tag{12}$$

where the second inequality is due to (P1) of stochastic dominance. At the same time, we have

$$\begin{aligned} \text{Var}(U) &\leq \sum_i \text{Var}\Big(\sum_{n=1}^{M} \sum_j \tilde{X}_{ij}^{(n)}\Big) \\ &= \sum_i \left[\mathbb{E}[M] \text{Var}\Big(\sum_j \tilde{X}_{ij}^{(n)}\Big) + \mathbb{E}^2\Big[\sum_j \tilde{X}_{ij}^{(n)}\Big] \text{Var}(M) \right] \\ &\leq \sum_i \left[\mathbb{E}[M] \mathbb{E}\big[(\sum_j \tilde{X}_{ij}^{(n)})^2\big] + \mathbb{E}^2\Big[\sum_j \tilde{X}_{ij}^{(n)}\Big] \mathbb{E}[M^2] \right]. \end{aligned} \tag{13}$$

By (P1) of stochastic dominance, we have

$$\begin{aligned} &\mathbb{E}\Big[(\sum_j \tilde{X}_{ij}^{(n)})^2\Big] \leq \mathbb{E}\Big[(\sum_j X_{ij} + T)^2\Big] = \sigma^2 + (\beta + T)^2, \\ &\mathbb{E}[M^2] \leq \mathbb{E}[(N(T) + T/\beta)^2] = \lambda T + (\lambda T + T/\beta)^2. \end{aligned} \tag{14}$$

Taking (14) into (13), we have

$$\begin{aligned} \text{Var}(U) &\leq N \left\{ (\lambda T + \frac{T}{\beta})\sigma^2 + (\beta + T)^2 \Big[2\lambda T + \frac{T}{\beta} + (\lambda T + \frac{T}{\beta})^2 \Big] \right\} \\ &= \Theta(NT^4), \end{aligned}$$

which implies that

$$\mathbb{E}[U^2] = \text{Var}(U) + \mathbb{E}^2[U] = \Theta(N^2 T^4). \tag{15}$$

Taking (12) and (15) into (11), we have

$$\text{Delay(FIFO queue)} \leq O(T^4) + O(NT^2) = O(NT^2).$$

Note that the above delay is measured in the number of *frames*, which implies that the expected delay (measured in the number of *timeslots*) experienced by a non-conforming coflow in the FIFO queue is $O(NT^3)$.

Finally, it is worth mentioning that the GI/GI/1 queue is stable whenever $\delta \mathbb{E}[U] < 1$. Due to (12), this is true whenever the following condition is satisfied:

$$\delta N(\lambda T + T/\beta)(\beta + T) < 1. \tag{16}$$

Since $\delta = O(\frac{1}{N^2})$ and $T = O(\log N)$, the left-hand side of (16) can be made arbitrarily small (i.e., the queue is stable) for any $N \geq 1$ under a suitably small δ. We will further discuss the exact value of δ in Section VII-F. □

Remark. If $\rho \geq 1$, then Lemma 4 does not hold and the overflow probability δ cannot be made arbitrarily small regardless of the choice of T. In this case, the FIFO queue holding non-conforming coflows is unstable.

Step 4: Putting it all together.

Finally, we can evaluate the average coflow-level delay experience by both conforming and non-conforming coflows. By Lemma 6, the fraction of non-conforming coflows is at most $\eta \leq \frac{2\delta}{\rho}$. If $\rho < 1$, Lemma 4 shows that we can choose the frame size to be $T = \lceil \frac{\log(2N/\delta)}{\gamma} \rceil = O(\log N^3) = O(\log N)$ to achieve the overflow probability $\delta = O(\frac{1}{N^2})$. Under such a choice of T, Lemma 7 shows that Delay(FIFO queue) $= O(NT^3)$. Taking the values of T, η and Delay(FIFO queue) into (2), we can conclude that the average coflow-level delay under the CAB policy is

$$\begin{aligned}\mathbb{E}[W] &\leq (1-\frac{2\delta}{\rho})2T + \frac{2\delta}{\rho}[T + \text{Delay(FIFO queue)}] \\ &\leq 2T + \frac{2\delta}{\rho}\text{Delay(FIFO queue)} \\ &= O(T) + O(\delta N T^3) \\ &= O(T) \\ &= O(\log N).\end{aligned} \tag{17}$$

This completes the proof to Theorem 3.

D. Proof to Corollary 1

The proof to Theorem 3 shows that the coflow-level delay experienced by any conforming coflow is no greater than $2T = O(\log N)$ and the fraction of conforming coflows is more than $1 - \frac{2\delta}{\rho} = 1 - O(\frac{1}{N^2})$. As a result, the CAB policy also ensures that the $O(\log N)$ delay is achievable for an arbitrary coflow with high probability $1 - O(\frac{1}{N^2})$.

E. Heuristic Improvement

In this section, we propose several heuristics that improve the practical performance of the CAB policy up to some constant factor (as compared to N).

Shortest-Clearance-Time-First (SCTF) Rule. This rule simply means that we should first clear the coflow with the smallest clearance time. It is inspired by the optimality of the Shortest-Processing-Time rule in traditional machine scheduling literature. The SCTF rule can be leveraged when we clear conforming coflows. We first order these conforming coflows according to their clearance time. In a certain slot, suppose that queue (i, j) gets scheduled. Instead of transmitting a packet in queue (i, j) according to FIFO, we select to transmit a packet of the coflow with the shortest clearance time that also has a remaining packet in queue (i, j).

Dynamic Frame Sizing. This heuristic was suggested in [21]. In each frame, if all the conforming coflows from the previous frame have been cleared and there are no non-conforming coflows in the system, the system starts a new frame immediately (rather than being idle for the remainder of the frame).

F. Discussions

Robustness to assumptions. Note that the Poisson assumption is not essential in the proof; a similar proof can be constructed for any arrival process such that the number of arrivals during T slots has a light-tailed distribution (referred to as a light-tailed arrival process). On the other hand, if the batch size or the arrival process is not light-tailed, then the overflow probability no longer decreases exponentially with the frame size T and the logarithmic bound does not hold.

Joint scaling as $\rho \to 1$ and $N \to \infty$. It is also interesting to study the joint scaling of the coflow-level delay achieved by the CAB policy as $N \to \infty$ and $\rho \to 1$. Since the majority of coflows experience a delay no greater than $O(T)$, we focus on the scaling of T as $\rho \to 1$. The second-order Taylor series of $f(s)$ (see equation (6)) around $s = 0$ is

$$f(s) = (1-\rho)s - \lambda(\sigma^2 + \beta^2)s^2/2 + O(s^3).$$

When $\rho \to 1$, we can set $s = s^* = \frac{1-\rho}{\lambda(\sigma^2+\beta^2)} \to 0$ so that

$$f(s^*) = \frac{(1-\rho)^2}{2\lambda(\sigma^2+\beta^2)} + O((1-\rho)^3) > 0.$$

Define $\gamma \triangleq f(s^*) = \Theta((1-\rho)^2)$. It follows from (7) that

$$T = \Big\lceil \frac{\log(2N/\delta)}{\gamma} \Big\rceil = O\Big(\frac{\log(2N/\delta)}{(1-\rho)^2}\Big) = O\Big(\frac{\log N}{(1-\rho)^2}\Big)$$

as $\rho \to 1$ and $N \to \infty$. Compared with the $O(\frac{N \log N}{1-\rho})$ scaling under randomized or periodic scheduling, the CAB policy has a much better dependence on N but becomes more sensitive to ρ.

Computational Complexity. The computational complexity of the CAB policy can be analyzed in a similar way to the original batching policy [21]. The computational complexity is $O(N^{1.5} \log N)$ per slot.

Choosing Parameters. In the CAB policy, the frame size is set to be $T = \lceil \frac{\log(2N/\delta)}{\gamma} \rceil$. As a result, we need to choose the parameters δ and γ.

We first fix γ and discuss how to determine the value of δ (the overflow probability). The requirements on δ are:

(1) The Left-Hand Side (LHS) of (16) needs to be made below 1 such that the system is stable.
(2) $\delta = O(\frac{1}{N^2})$ such that the $O(\log N)$ delay is achievable.

Note that when $N\beta \geq 1$ (which is true for relatively large N), the LHS of (16) can be upper bounded by

$$LHS \leq \delta N(\lambda T + T/\beta)(\beta + N\beta T) = \delta NT(\rho+1)(1+NT).$$

Then we can obtain δ by solving the following system of equations:

$$\begin{cases} \delta NT(\rho+1)(1+NT) = \frac{1}{2} \\ T = \lceil \frac{\log(2N/\delta)}{\gamma} \rceil \end{cases} \tag{18}$$

Clearly, the solution to (18) finds $\delta = O(\frac{1}{N^2T^2}) \leq O(\frac{1}{N^2})$ such that the second requirement is met; the first requirement is met due to the fact that $\delta NT(\rho+1)(1+NT)$ is greater than the LHS of (16). Note that the system utilization ρ can be

measured, so equations (18) can solved iteratively for a given γ.

Now we discuss how to estimate the value of γ. To obtain a smaller T, it is desirable to have a larger γ. If we know the coflow arrival rate and the distribution of batch sizes, we can compute γ by maximizing $f(s)$ shown in (6), i.e.,

$$\gamma = \max_{s \geq 0} \lambda[1 - M_B(s)] + s. \tag{19}$$

If the system has no information about λ or batch size distributions, γ can be empirically tuned. We can first pick some (small) arbitrary value of γ and obtain δ and T by solving equations (18). The system then measures the actual overflow probability δ' under such a frame size. If $\delta' > \delta$, the value of γ is reduced by some step size to increase the frame size T; otherwise the value of γ is increased by some step size. The above procedure proceeds until $\delta' \approx \delta$, and a good value of γ is found.

VIII. Simulation Results

In this section, we numerically evaluate the coflow-level performance of the CAB policy.

A. Simulation Setup

In our simulations, coflows arrive to the system according to a Poisson process with rate $\lambda = 0.3$ (per slot). The batch sizes (X_{ij}'s) follow a geometric distribution with mean $\frac{\beta}{N}$ where $\beta = 2.5$ (measured in the number of packets). Hence, the offered load is $\rho = 0.75$. The simulation is run for a sufficiently long time (10^6 slots) such that the steady state is reached. The parameter γ is obtained by solving equation (19) offline; the parameters δ and T are obtained by solving equation (18).

B. Scaling with $N \to \infty$

First, we evaluate the scaling of coflow-level delay as $N \to \infty$. The following schemes are compared:

(1) CAB policy. Note that two heuristics mentioned in Section VII-E are leveraged.

(2) Randomized scheduling as described in Section VI.

(3) Max-Weight Matching (MWM) scheduling: the schedule in slot t is the maximum weight matching of $\mathbf{Q}(t)$ where $\mathbf{Q}(t) = (Q_{ij}(t))$ is the queue length matrix in slot t.

Figure 2 shows the comparison of these schemes with respect to the average coflow-level delay, where the horizontal axis is on a logarithmic scale. As the theoretical bound suggests, the CAB policy achieves the logarithmic scaling as $N \to \infty$ (i.e., a straight line in the figure). By comparison, the average coflow-level delay achieved by the randomized scheme grows much faster with N. Moreover, it can be observed that the CAB policy outperforms the randomized scheme even for very small N (e.g., $N = 40$).

Another interesting observation is that the MWM policy has an exceptional coflow-level performance. It is observed that the MWM policy empirically achieves the optimal logarithmic coflow-level delay scaling as $N \to \infty$. The MWM policy also slightly outperforms the CAB policy by some constant factor. Unfortunately, the coflow-level delay analysis of the MWM policy is very challenging and left for future work.

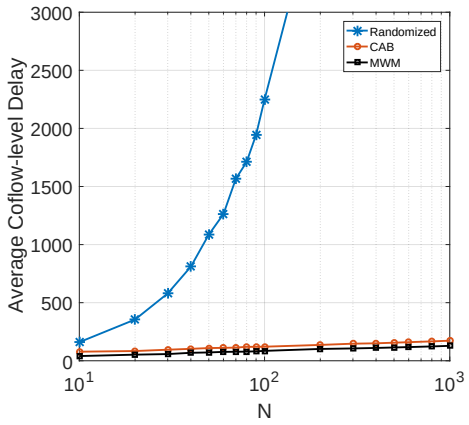

Fig. 2. Average coflow-level delay under different scheduling policies. Note that the horizontal axis is in the log scale.

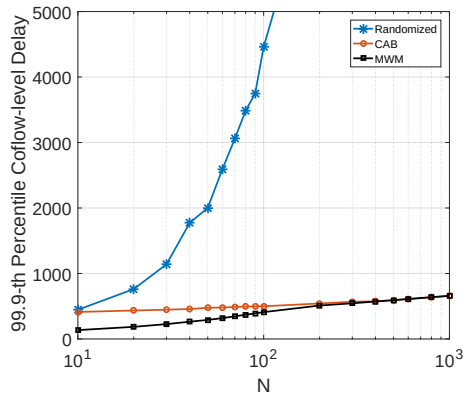

Fig. 3. 99.9-th percentile coflow-level delay under different scheduling policies. Note that the horizontal axis is in the log scale.

In the above, the *average* coflow-level delay is evaluated but in many cases we are also interested in *tail latency*. Note that the CAB policy guarantees that the $O(\log N)$ coflow-level delay is achievable with high probability $1 - O(1/N^2)$ (see Corollary 1), so the coflow-level delay tail under the CAB policy also scales as $O(\log N)$ when N is relatively large. This is illustrated in Figure 3, where the 99.9-percentile coflow-level delay is evaluated. As expected, the CAB policy achieves $O(\log N)$ scaling for the coflow-level delay tail, significantly outperforming the randomized scheme. The MWM algorithm has a similar delay tail as the CAB policy when N is relatively large.

C. Coflow-level Delay Dilation

Next, we compare the coflow-level delay with the packet-level delay under the randomized policy and the CAB policy. In particular, we are interested in the *coflow-level delay dilation factor* which is the ratio between the average coflow-level delay and the average packet-level delay. As is illus-

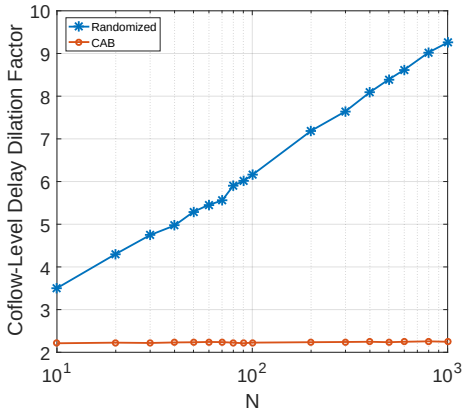

Fig. 4. Coflow-level delay dilation factor (the ratio between average coflow-level delay and average packet-level delay) under different scheduling policies. Note that the horizontal axis is in the log scale.

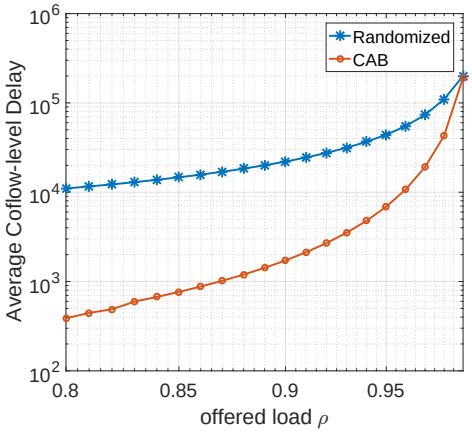

Fig. 5. Scaling of average coflow-level delay as $\rho \to 1$ ($N = 200$). Note that the vertical axis is in the log scale.

trated in Figure 4, the randomized policy has a coflow-level delay dilation factor of $O(\log N)$; this observation empirically validates the tightness of the $O(N \log N)$ bound shown in Theorem 2 (note that the average packet delay achieved by the randomized policy is exactly $\Theta(N)$ under our simulation environment). By comparison, the delay dilation factor for the CAB policy remains at a constant level as $N \to \infty$, which shows the benefits of "coflow-awareness".

D. Scaling with $\rho \to 1$

Finally, we numerically study the sensitivity of the coflow-level performance under different scheduling policies as the offered load $\rho \to 1$. This is shown in Figure 5. Clearly, the CAB policy is more sensitive to the offered load ρ than the randomized policy. In the heavy-traffic regime, the randomized policy even outperforms the CAB policy. Indeed, the average coflow-level delay achieved by the randomized policy grows as $O(\frac{1}{1-\rho})$ as $\rho \to 1$ (see Remark 1 below Theorem 2). By comparison, the CAB policy achieves $O(\frac{1}{(1-\rho)^2})$ average coflow-level delay as $\rho \to 1$ (see Section VII-F). As a result, the price for the better scaling with N is the worse dependence on ρ.

IX. Conclusion and Future Work

In this paper, we investigate the optimal scaling of coflow-level delay in an $N \times N$ input-queued switch as $N \to \infty$. We develop lower bounds on the coflow-level delay that can be achieved by any scheduling policy. In particular, when flow sizes have light-tailed distributions, the lower bound $O(\log N)$ can be attained by the proposed Coflow-Aware Batching (CAB) policy. Thus, the optimal scaling of coflow-level delay is $O(\log N)$ under light-tailed flow sizes.

Future work includes the design of a throughput-optimal scheduling policy that achieves the best scaling of coflow-level delay under a general coflow arrival process and general flow size distributions. Variations of the Maximum Weight Matching (MWM) algorithm (that leverage coflow-level information) may be a promising direction to investigate since our simulation results show that the MWM algorithm has exceptional coflow-level performance. However, the coflow-level performance analysis of the MWM algorithm may be very challenging. Another interesting direction is to consider *correlated* flow sizes and investigate how the correlation influences the scaling properties. Finally, it is also worth studying the case without prior knowledge on coflows such as coflow sizes and release times of flows (currently we assume that all the flows in a coflow are released simultaneously and coflow sizes are known).

References

[1] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. in *OSDI*, 2004.

[2] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, I. Stoica. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. in *USENIX NSDI*, 2012.

[3] M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica. Managing data transfers in computer clusters with orchestra. in *Proceedings of the ACM SIGCOMM*, pp. 98–109, 2011.

[4] M. Chowdhury and I. Stoica. Coflow: A Networking Abstraction for Cluster Applications. in *ACM Hotnets*, 2012.

[5] M. Chowdhury, Y. Zhong, and I. Stoica. Efficient Coflow Scheduling with Varys. in *ACM SIGCOMM*, 2014.

[6] M. Chowdhury and I. Stoica. Efficient Coflow Scheduling Without Prior Knowledge. in *ACM SIGCOMM*, 2015.

[7] Y. Zhao, K. Chen, W. Bai, M. Yu, C. Tian, Y. Geng, Y. Zhang, D. Li, and S. Wang. RAPIER: Integrating Routing and Scheduling for Coflow-aware Data Center Networks. in *IEEE INFOCOM*, 2015.

[8] F. R. Dogar, T. Karagiannis, H. Ballani, and A. Rowstron. Decentralized Task-Aware Scheduling forData Center Networks. in *ACM SIGCOMM*, 2014.

[9] I. A. Rai, G. Urvoy-Keller, and E. W. Biersack. Analysis of LAS scheduling for job size distributions with high variance. in *ACM SIGMETRICS Performance Evaluation Review*, vol. 31, no. 1, pp. 218–228, 2003.

[10] C. Hung, L. Golubchik, M. Yu. Scheduling Jobs Across Geo-distributed Datacenters. in *ACM SoCC*, 2015.

[11] S. Luo *et. al*. Minimizing Average Coflow Completion Time with Decentralized Scheduling. in *IEEE ICC*, 2015.

[12] Z. Qiu, C. Stein, and Y. Zhong. Minimizing the total weighted completion time of coflows in datacenter networks. in *SPAA*, 2015.
[13] D. Shah, N. Walton, and Y. Zhong. Optimal Queue-Size Scaling in Switched Networks. in *ACM SIGMETRICS*, 2012.
[14] D. Shah and M. Kopikare. Delay bounds for the approximate Maximum Weight matching algorithm for input queued switches. in *IEEE Infocom*, 2002.
[15] S. T. Maguluri and and R. Srikant. Heavy-Traffic Behavior of the MaxWeight Algorithm in a Switch with Uniform Traffic. in *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, no. 2, pp. 72-74, 2015.
[16] S. T. Maguluri, S. K. Burle and R. Srikant. Optimal Heavy-Traffic Queue Length Scaling in an Incompletely Saturated Switch. in *ACM SIGMETRICS*, 2016.
[17] T. Inukai. An Efficient SS/TDMA Time Slot Assignment Algorithm. in *Transactions on Communications*, vol. 27, no. 10, pp. 1449-1455, 1979.
[18] B. Eisenberg. On the expectation of the maximum of IID geometric random variables. *Statistics and Probability Letters*, vol. 78, pp. 135-143, 2008.
[19] F. Baccelli, A. M. Makowski and A. Shwartz. The Fork-Join Queue and Related Systems with Synchronization Constraints: Stochastic Ordering and Computable Bounds. in *Advances in Applied Probability*, vol. 21, no. 3, pp. 629-660, 1989.
[20] N. Papadatos. Maximum variance of order statistics. in *Ann. Inst. Statist. Math.*, vol. 47, no. 1, pp. 185-193, 1995.
[21] M. Neely, E. Modiano, and Y. S. Cheng. Logarithmic delay for $n \times n$ packet switches under the cross-bar constraint. in *IEEE/ACM Transactions on Networking*, vol. 15, no. 3, 2007.
[22] E. Wolfstetter. Stochastic Dominance: Theory and Applications. in *Topics in Microeconomics: Industrial Organization, Auctions, and Incentives (Second Edition)*, Cambridge University Press, 2002.
[23] T. L. Lai and H. Robbins. A class of dependent random variables and their maxima. *Z. Wahrscheinlichkeitsch*. vol. 42, pp. 89-111, 1978.
[24] C-S Chang, W-J Chen, and H-Y Huang. Birkhoff-von neumann input buffered crossbar switches. in *Proc. IEEE INFOCOM*, 2000.
[25] R. Srikant and Lei Ying. *Communication Networks:An Optimization, Control, and Stochastic Networks Perspective.* Cambridge University Press, 2014.
[26] N. McKeown, V. Anantharam, and J. Walrand. Achieving 100% throughput in an input-queued switch. in *IEEE INFOCOM*, 1996.
[27] E. Leonardi, M. Mellia, F. Neri, and M. Ajmone Marsan. Bounds on average delays and queue size averages and variances in input-queued cell-based switches. in *IEEE INFOCOM*, 2001.
[28] L. P. Devroye. Inequalities for the completion times of stochastic Pert networks. in *Mathematics of Operations Research*, vol. 4, no. 4, pp. 441-447, 1979.
[29] R. Nelson and A. N. Tantawi. Approximate Analysis of Fork/Join Synchronization in Parallel Queues. in *IEEE Transactions on Computers*, vol. 37, no. 6, 1988.
[30] A. A. Borovkov. Stochastic Processes in Queueing Theory (English translation). Springer-Verlag, New York, 1976.
[31] P. Gao, S. Wittevrongel, and H. Bruneel. Discrete-time multiserver queues with geometric service times. in *Computers & Operations Research*, vol. 31, pp. 81-99, 2004.
[32] J. D. Esary, F. Proschan, and D. W. Walkup. Association of Random Variables, with Applications. in *Ann. Math. Statist.*, Vol. 38, No. 5, pp. 1466-1474, 1967.
[33] Robert G. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2014.
[34] Torben Meisling. Discrete-Time Queuing Theory. in *Operations Research*, Vol. 6, No. 1, pp. 96-105, 1957.

APPENDIX

A. Proof to Lemma 3

We first introduce a few technical lemmas. The first lemma is regarding the asymptotic bound of order statistics [23].

Lemma 8. *Let $Y_1, \cdots, Y_N$ be i.i.d. $\mathbb{R}^+$-valued random variables whose common CDF $G(y)$ satisfies the following conditions:*

$$G(y) < 1, \quad \textit{for all } y \geq 0,$$

and

$$\lim_{y\to\infty} \frac{1-G(cy)}{1-G(y)} = 0, \quad \textit{for all } c > 1.$$

Under these conditions, the asymptotics

$$\mathbb{E}[\max_i Y_i] = m_N(1+O(1))$$

holds true as $N \to \infty$, where

$$m_N = \inf\Big\{y \geq 0 : 1 - G(y) \leq \frac{1}{N}\Big\}.$$

The second lemma is an application of Lemma 8 to light-tailed random variables.

Lemma 9. *Suppose $Y_1, \cdots, Y_N$ are independent **light-tailed** random variables with $\mathbb{E}[Y_i^n] = O(1)$ as $N \to \infty$ for all $n \in \mathbb{Z}^+$. Then $\mathbb{E}[\max_i Y_i] = O(\log N)$ as $N \to \infty$.*

Proof. Suppose the CDF of Y_i is $G_i(y)$, and let $\mathbb{E}[Y_i] = \mu_i$. Also let $M_{Y_i}(s)$ be the Moment Generating Function of Y_i. Since Y_i is light-tailed, there exists some $s^* > 0$ such that $M_{Y_i}(s) < \infty$ for all $0 \leq s \leq s^*$. By Chernoff bound, for any $0 \leq s \leq s^*$ and $y > \mu_i$

$$1 - G_i(y) \leq M_{Y_i}(s)e^{-sy} \triangleq e^{-yf_i(s)},$$

where

$$f_i(s) = s - \frac{\log M_{Y_i}(s)}{y}.$$

It is clear that $f_i(0) = 0$ and $f_i'(0) = 1 - \frac{M_{Y_i}'(0)}{yM_{Y_i}(0)} = 1 - \frac{\mu_i}{y} > 0$ for any $y > \mu_i$. Due to the continuity of $f_i(s)$ around $s = 0$, there exists a sufficiently small $\epsilon_i > 0$ such that $f_i(\epsilon_i) > 0$. Define $\gamma_i = f_i(\epsilon_i) > 0$. We have

$$1 - G_i(y) \leq e^{-\gamma_i y}, \text{ for all } y > \mu_i.$$

Let $\gamma = \min_i \gamma_i$. Then it follows that for all $i \in [N]$

$$1 - G_i(y) \leq e^{-\gamma y}, \text{ for all } y > \mu_i.$$

Consider a sequence of i.i.d. random variables $Y_1', \cdots, Y_N'$ with shifted exponential distribution $F(y) = 1 - e^{-\gamma(y-\mu)}$ for $y > \mu$, where $\mu = \max_i \mu_i$. It is clear that Y_i is stochastically dominated by Y_i' for all $i \in [N]$. By (P2) of stochastic dominance (see Section III-B), we have $\mathbb{E}[\max_i Y_i] \leq \mathbb{E}[\max_i Y_i']$. Thus, it suffices to show $\mathbb{E}[\max_i Y_i'] = O(\log N)$ as $N \to \infty$. Note that $F(y)$ satisfies the conditions in Lemma 8. Hence, we have $\mathbb{E}[\max_i Y_i'] = m_N'(1 + O(1))$ where

$$m_N' = \inf\Big\{y \geq 0 : 1 - F(y) \leq \frac{1}{N}\Big\} = \frac{\log N}{\gamma} + \mu.$$

Since $\mu = O(1)$, we can conclude that $\mathbb{E}[\max_i Y_i'] = O(\frac{\log N}{\gamma})$ as $N \to \infty$. Now it remains to show that $\gamma = O(1)$ as $N \to \infty$. Taking the Taylor expansion of $f_i(s)$ around $s = 0$, we have

$$f_i(s) = \sum_{n=0}^{\infty} \frac{f_i^{(n)}(0)}{n!} s^n = (1 - \frac{\mu_i}{y})s - \frac{1}{y} \sum_{n=2}^{\infty} \frac{\kappa_n}{n!} s^n, \quad (20)$$

where κ_n is the n-th cumulant[3] of Y_i. Note that κ_n is a degree-n polynomial in the first n moments of Y_i. Since $\mathbb{E}[Y_i^n] = O(1)$ for all $n \in \mathbb{Z}^+$, we have $\kappa_n = O(1)$ as $N \to \infty$. As a result, the second term in (20) (i.e., $\frac{1}{y} \sum_{n=2}^{\infty} \frac{\kappa_n}{n!} s^n$) is also independent of N. Hence, there exists some ϵ_i independent of N such that $\gamma_i = f_i(\epsilon_i) > 0$ and γ_i is independent of N, which implies that $\gamma = \max_i \gamma_i = O(1)$. □

Takeaway. Note that Lemma 9 only shows the scaling of $\mathbb{E}[\max_i Y_i]$ in the case where all the moments of Y_i are constants as compared to N. Sometimes, we are also interested in the case where $\mathbb{E}[Y_i] = O(f(N))$ and $f(N) \to \infty$ as $N \to \infty$. For simplicity, we assume that Y_i's are i.i.d. random variables.

Corollary 2. *Suppose $Y_1, \cdots, Y_N$ are i.i.d. light-tailed random variables with $\mathbb{E}[Y_i] = O(f(N))$ where $f(N) \to \infty$ as $N \to \infty$. If $\mathbb{E}[Y_i^n] = O(f^n(N))$ for all $n \in \mathbb{Z}^+$, then $\mathbb{E}[\max_i Y_i] = O(f(N) \log N)$ as $N \to \infty$.*

The proof to this corollary is omitted for brevity since it is similar to the proof of Lemma 9 except that we explicitly set $\epsilon_i = \frac{1}{f(N)}$.

With Lemma 9, we can easily prove the theorem. It is clear that

$$\mathbb{E}[\tau(\mathbf{X})] \le \mathbb{E}[\max_i \sum_j X_{ij}] + \mathbb{E}[\max_j \sum_i X_{ij}].$$

By our assumption, $\sum_j X_{ij},\ i = 1, \cdots, N$ is a sequence of independent light-tailed random variables with $\mathbb{E}[(\sum_j X_{ij})^n] = O(1)$ as $N \to \infty$ for all $n \in \mathbb{Z}^+$. By Lemma 9, we have $\mathbb{E}[\max_i \sum_j X_{ij}] = O(\log N)$. Similarly, we have $\mathbb{E}[\max_j \sum_i X_{ij}] = O(\log N)$. As a result, we can conclude that $\mathbb{E}[\tau(\mathbf{X})] = O(\log N)$.

To show the tightness, we consider a scenario where $\mathbf{X}$ is a diagonal matrix: $X_{ij} = 0$ with probability 1 for $i \neq j$ and X_{ii} has a geometric distribution with mean β for all $i \in [N]$. In this case, we have $\mathbb{E}[\tau(\mathbf{X})] = \mathbb{E}[\max_i X_{ii}]$. It was shown in [18] that the expectation of the maximum of N i.i.d. geometric random variables is $\Theta(\log N)$ as $N \to \infty$. As a result, $\mathbb{E}[\tau(\mathbf{X})] = \Theta(\log N)$ in this scenario.

B. Proof to Theorem 2

We only prove the result for the periodic scheduling policy; the randomized policy can be analyzed in exactly the same way. We assume the system is initially empty. Coflow arrivals are indexed by $k = 1, 2, \cdots$. Suppose the traffic matrix of coflow k is $\mathbf{X}^{(k)} = (X_{ij}^{(k)})$, and denote by $T^{(k)}$ the inter-arrival time between coflow k and coflow $k+1$. Let $W_{ij}^{(k)}$ be the queuing delay experienced by coflow k in queue (i, j), and denote by $U_{ij}^{(k)}$ the processing time for the batch of coflow k in queue (i, j). Under periodic scheduling, each queue gets served every N time slots. As a result, we have

$$U_{ij}^{(k)} = \Big(Y_{ij}^{(k)} + N(X_{ij}^{(k)} - 1)\Big)^+,$$

where $(x)^+ = \max(0, x)$ and $Y_{ij}^{(k)}$ is the processing time of the first packet of coflow k in queue (i, j). It is clear that $Y_{ij}^{(k)} \le N$ with probability 1 under periodic scheduling. As a result, we have $U_{ij}^{(k)} \le N X_{ij}^{(k)} \triangleq \tilde{U}_{ij}^{(k)}$ with probability 1. Obviously, the delay performance of the original system (with the batch processing time $U_{ij}^{(k)}$) is upper-bounded by the delay performance under a system where the batch processing time is $\tilde{U}_{ij}^{(k)}$. The latter system is referred to as "System 2", and denote by $\tilde{W}_{ij}^{(k)}$ the queuing delay experienced by coflow k in queue (i, j) in System 2. Now we show that $\tilde{W}_{ij}^{(k)}$'s are *associated random variables* (see Section III-B for the definition). First, some simple associated random variables are identified in our context.

Lemma 10. *Define $V_{ij}^{(k)} \triangleq N X_{ij}^{(k)} - T^{(k)}$. For any $k \in \mathbb{Z}^+$, random variables $V_{ij}^{(k)},\ i, j \in [N]$ are associated.*

Proof. Given $T^{(k)} = t$, it is clear that $V_{ij}^{(k)}$'s are independent and thus associated (by (P1) of associated random variables). As a result, by the definition of associated random variables, for any non-decreasing functions f and g

$$\mathrm{Cov}\Big(f(\mathbf{V}^{(k)}), g(\mathbf{V}^{(k)}) \Big| T^{(k)} = t\Big) \ge 0.$$

As a result, it follows that

$$\begin{aligned}
&\mathrm{Cov}\Big(f(\mathbf{V}^{(k)}), g(\mathbf{V}^{(k)})\Big) \\
=&\mathbb{E}\Big[\mathrm{Cov}\Big(f(\mathbf{V}^{(k)}), g(\mathbf{V}^{(k)}) \Big| T^{(k)}\Big)\Big] \\
=&\int_{t=0}^{\infty} \mathrm{Cov}\Big(f(\mathbf{V}^{(k)}), g(\mathbf{V}^{(k)}) \Big| T^{(k)} = t\Big) f_{T^{(k)}}(t) dt \ge 0,
\end{aligned}$$

where $f_{T^{(k)}}(t)$ is the PDF for $T^{(k)}$. Therefore, random variables $V_{ij}^{(k)},\ i, j \in [N]$ are associated, and this conclusion holds for all $k \in \mathbb{Z}^+$. □

With Lemma 10, we can show that random variables $\tilde{W}_{ij}^{(k)}$'s are associated.

Lemma 11. *For all $k \in \mathbb{Z}^+$, random variables $\tilde{W}_{ij}^{(k)},\ i, j \in [N]$ are associated.*

Proof. We prove by induction on k.

Basis Step. When $k = 1$, random variables $\tilde{W}_{ij}^{(1)},\ i, j \in [N]$ are clearly associated since the system is initially empty and $\tilde{W}_{ij}^{(1)} = 0$ with probability 1 for all $i, j \in [N]$.

Inductive Step. For some $k \ge 1$, assume that random variables $\tilde{W}_{ij}^{(k)},\ i, j \in [N]$ are associated. Now we prove that

[3]The n-th cumulant of Y_i is the n-th order derivative for the logarithm of the MGF of Y_i, evaluated at zero, i.e., $\kappa_n = K^{(n)}(0)$, where $K(s) = \log M_{Y_i}(s)$.

random variables $\tilde{W}_{ij}^{(k+1)},\ i,j \in [N]$ are also associated. It is clear that

$$\begin{aligned}\tilde{W}_{ij}^{(k+1)} &= (\tilde{W}_{ij}^{(k)} + \tilde{U}_{ij}^{(k)} - T^{(k)})^+ \\ &= (\tilde{W}_{ij}^{(k)} + NX_{ij}^{(k)} - T^{(k)})^+.\end{aligned}$$

We identify three sets of associated random variables:

(1) Random variables $NX_{ij}^{(k)} - T^{(k)},\ i,j \in [N]$ are associated due to Lemma 10;
(2) The union of $\{\tilde{W}_{ij}^{(k)},\ i,j \in [N]\}$ and $\{NX_{ij}^{(k)} - T^{(k)},\ i,j \in [N]\}$ is a set of associated random variables since they are two sets of independent associated random variables (by (P3) of associated random variables);
(3) Random variables $\tilde{W}_{ij}^{(k+1)},\ i,j \in [N]$ are associated since they are non-decreasing functions of the set of associated random variables identified in (2) (by (P2) of associated random variables).

Now we have completed the induction. □

Since the above lemma holds for all $k \in \mathbb{Z}^+$, we drop the superscript k for convenience. Define $\tilde{Z}_{ij} = \tilde{W}_{ij}1_{\{X_{ij}\geq 1\}} + NX_{ij}$, i.e., the total time that coflow k needs to wait for until its batch in queue (i,j) is cleared. Note that the $1_{\{X_{ij}\geq 1\}}$ term is due to the fact that if a coflow contains no packets in queue (i,j), it does not need to experience the queuing delay $\tilde{W}_{ij}$. By Lemma 11, $\tilde{W}_{ij}$'s are associated; X_{ij}'s are also associated due to the independence (P1); the union of $\tilde{W}_{ij}$'s and X_{ij}'s is also a set of associated random variables due to the independence of X_{ij}'s and $\tilde{W}_{ij}$'s (P3). Therefore, we can conclude that $\tilde{Z}_{ij}$'s are associated since they are non-decreasing functions of $\tilde{W}_{ij}$'s and X_{ij}'s (P2). Note that $\max_{ij}\tilde{Z}_{ij}$ is the coflow-level delay experienced by a coflow in System 2. By (P4) of associated random variables, we have $\mathbb{E}[\max_{ij}\tilde{Z}_{ij}] \leq \mathbb{E}[\max_{ij}\tilde{Z}'_{ij}]$ where $\tilde{Z}'_{ij}$'s are *independent* random variables identically distributed as $\tilde{Z}_{ij}$'s. In order to evaluate the value of $\mathbb{E}[\max_{ij}\tilde{Z}'_{ij}]$, we identify some important properties of the distribution of $\tilde{Z}_{ij}$.

First, we claim that $\tilde{Z}_{ij}$'s are light-tailed. Indeed, since X_{ij} is light-tailed, the service time $\tilde{U}_{ij} = NX_{ij}$ is also light-tailed. According to [30] (Theorem 11, p. 129), if the service time is light-tailed, the tail of the queuing time in an M/G/1 queue decreases exponentially. Hence, $\tilde{W}_{ij}$'s are light-tailed, which implies that $\tilde{Z}_{ij}$'s also have light-tailed distributions.

Next, we evaluate the value of $\mathbb{E}[\tilde{Z}_{ij}]$ and higher moments of $\tilde{Z}_{ij}$. It is clear that queue (i,j) is a slotted M/G/1 queue with arrival rate λ and service time $\tilde{U}_{ij} = NX_{ij}$. Note that $\mathbb{E}[\tilde{U}_{ij}] = \beta$ and $\mathbb{E}[\tilde{U}_{ij}^2] = (N\sigma^2 + \beta^2)$. According to the Pollaczek-Khinchin formula for slotted M/G/1 queues, we have

$$\begin{aligned}\mathbb{E}[\tilde{W}_{ij}] &= \frac{\lambda\mathbb{E}[\tilde{U}_{ij}^2]}{2(1-\lambda\mathbb{E}[\tilde{U}_{ij}])} + \frac{1}{2} \\ &= \frac{\lambda(N\sigma^2+\beta^2)}{2(1-\rho)} + \frac{1}{2}.\end{aligned}$$

As a result, we have

$$\begin{aligned}\mathbb{E}[\tilde{Z}_{ij}] &= \mathbb{P}(X_{ij}\geq 1)\mathbb{E}[\tilde{W}_{ij}] + \mathbb{E}[\tilde{U}_{ij}] \\ &\leq \frac{\beta}{N}\Big[\frac{\lambda(N\sigma^2+\beta^2)}{2(1-\rho)} + \frac{1}{2}\Big] + \beta \\ &= O(1),\end{aligned}$$

where the inequality utilizes the Markov inequality, i.e., $\mathbb{P}(X_{ij} \geq 1) \leq \mathbb{E}[X_{ij}] = \frac{\beta}{N}$. Furthermore, noting that $\mathbb{E}[\tilde{U}_{ij}^n] = N^n\mathbb{E}[X_{ij}^n] = O(N^{n-1})$ for all $n \in \mathbb{Z}^+$ (by our assumptions on X_{ij}'s), we can similarly obtain $\mathbb{E}[\tilde{Z}_{ij}^n] = O(N^{n-1})$ according to the moment bounds in [30].

Recall that $\tilde{Z}'_{ij}$'s are independent random variables identically distributed as $\tilde{Z}_{ij}$'s. Define $\tilde{Z}'_i \triangleq \max_j \tilde{Z}'_{ij}$ for all $i \in [N]$. It follows from the above analysis that $\tilde{Z}'_i,\ i = 1,\cdots,N$ are a sequence of i.i.d. light-tailed random variables with $\mathbb{E}[\tilde{Z}'_i] \leq \sum_j \mathbb{E}[\tilde{Z}'_{ij}] = O(N)$ and $\mathbb{E}[(\tilde{Z}'_i)^n] \leq O\big(\sum_j \mathbb{E}[(\tilde{Z}'_{ij})^n]\big) = O(N^n)$ for all $n \in \mathbb{Z}^+$. By Corollary 2 (see Appendix A), we have $\mathbb{E}[\max_i \tilde{Z}'_i] = O(N\log N)$ as $N \to \infty$. Then it follows that

$$\mathbb{E}[\max_{i,j}\tilde{Z}_{ij}] \leq \mathbb{E}[\max_{i,j}\tilde{Z}'_{ij}] = \mathbb{E}[\max_i \tilde{Z}'_i] = O(N\log N).$$

Therefore, the average coflow-level delay achieved by the periodic scheduling policy is $O(N\log N)$.

C. Flow size distribution in an overflow frame

In this appendix, we introduce a lemma about the flow size distribution of a coflow that arrives in an **overflow** frame.

Lemma 12. *Let $(\tilde{X}_{ij})$ be the traffic of an arbitrary coflow in an **overflow** frame. Then $\sum_j \tilde{X}_{ij}$ is stochastically dominated by $\sum_j X_{ij} + T$ for all $i \in [N]$, where (X_{ij}) is the traffic MATRIX of an arbitrary coflow in an **arbitrary** frame.*

Proof. Let $\mathbf{L}$ be the total traffic that arrives in an arbitrary frame, and denote by $\mathbf{X} = (X_{ij})$ the traffic associated with an arbitrary coflow in this frame. Clearly, an overflow occurs in this frame if and only if $\tau(\mathbf{L}) \geq T$. As a result, the flow size distribution for an arbitrary coflow in this overflow frame is

$$\mathbb{P}[X_{ij} \geq m | \tau(\mathbf{L}) \geq T] \triangleq \mathbb{P}[\tilde{X}_{ij} \geq m],$$

where (X_{ij}) is the traffic of an arbitrary coflow in an arbitrary frame.

In order to show that $\sum_j \tilde{X}_{ij}$ is stochastically dominated by $\sum_j X_{ij} + T$ for all $i \in [N]$, it suffices to show that for all $i \in [N]$

$$\mathbb{P}[\tilde{X}_{ij} \geq m] \leq \mathbb{P}[\sum_j X_{ij} + T \geq m].$$

To show this inequality, we first prove that for all $i \in [N]$

$$\mathbb{P}[\sum_j X_{ij} \geq m | \tau(\mathbf{L}) \geq T] \leq \mathbb{P}[\sum_j X_{ij} \geq m | \sum_j X_{ij} \geq T]. \tag{21}$$

If $m < T$, we note that the right-hand side of (21) equals to 1, so inequality (21) naturally holds. If $m \geq T$, we notice that

$$\sum_j X_{ij} \geq m \Rightarrow \tau(\mathbf{L}) \geq \sum_j X_{ij} \geq m \geq T,$$

which implies that

$$\mathbb{P}[\sum_j X_{ij} \geq m, \tau(\mathbf{L}) \geq T] = \mathbb{P}[\sum_j X_{ij} \geq m]. \tag{22}$$

Similarly, we can show that when $m \geq T$

$$\mathbb{P}[\sum_j X_{ij} \geq m, \sum_j X_{ij} \geq T] = \mathbb{P}[\sum_j X_{ij} \geq m]. \tag{23}$$

Comparing (22) with (23), we have when $m \geq T$

$$\mathbb{P}[\sum_j X_{ij} \geq m, \tau(\mathbf{L}) \geq T] = \mathbb{P}[\sum_j X_{ij} \geq m, \sum_j X_{ij} \geq T]. \tag{24}$$

Now we rewrite $\mathbb{P}[\sum_j X_{ij} \geq m | \tau(\mathbf{L}) \geq T]$ according to the rule of conditional probability:

$$\begin{aligned}
\mathbb{P}[\sum_j X_{ij} \geq m | \tau(\mathbf{L}) \geq T] &= \frac{\mathbb{P}[\sum_j X_{ij} \geq m, \tau(\mathbf{L}) \geq T]}{\mathbb{P}[\tau(\mathbf{L}) \geq T]} \\
&\leq \frac{\mathbb{P}[\sum_j X_{ij} \geq m, \sum_j X_{ij} \geq T]}{\mathbb{P}[\sum_j X_{ij} \geq T]} \\
&= \mathbb{P}[\sum_j X_{ij} \geq m | \sum_j X_{ij} \geq T],
\end{aligned}$$

where the inequality is due to (24) and the fact that $\mathbb{P}[\tau(\mathbf{L}) \geq T] \geq \mathbb{P}[\sum_j X_{ij} \geq T]$. This completes the proof to (21).

Since $\sum_j X_{ij}$ is light-tailed, we know from [33] that as $T \to \infty$,

$$\mathbb{P}[\sum_j X_{ij} \geq m | \sum_j X_{ij} \geq T] \leq \mathbb{P}[\sum_j X_{ij} \geq m - T]. \tag{25}$$

Taking (25) into (21), we obtain

$$\mathbb{P}[\sum_j X_{ij} \geq m | \tau(\mathbf{L}) \geq T] \leq \mathbb{P}[\sum_j X_{ij} \geq m - T]. \tag{26}$$

Note that the left-hand side of (26) equals to $\mathbb{P}[\sum_j \tilde{X}_{ij} \geq m]$ by our definition of $\tilde{X}_{ij}$, and the right-hand side of (26) can be rewritten as $\mathbb{P}[\sum_j X_{ij} + T \geq m]$. As a result, we have for all $i \in [N]$

$$\mathbb{P}[\sum_j \tilde{X}_{ij} \geq m] \leq \mathbb{P}[\sum_j X_{ij} + T \geq m],$$

i.e., $\sum_j \tilde{X}_{ij}$ is stochastically dominated by $\sum_j X_{ij} + T$. $\square$