Queueing Delay Minimization in Overloaded Networks via Rate Control

Xinyu Wu, Dan Wu, Eytan Modiano
Laboratory for Information and Decision Systems, MIT, USA
{xinyuwu1,danwumit,modiano}@mit.edu

***Abstract*—We study the pattern of queueing delay minimization in overloaded networks using link rate control. We show that serving packets with maximum link rates does not minimize queueing delay when networks are on overload. To understand this counter-intuitive observation and identify the optimal solution, we characterize the packet queueing delay explicitly through a fluid queue model, which enables us to characterize the sufficient and necessary condition for a policy to be delay-optimal for single-hop networks. We show that setting the link rates to be proportional to the packet arrival rates minimizes both average and maximum delay among all sources simultaneously. We further prove that a queue-proportional rate control policy, agnostic of packet arrival rates, asymptotically achieves optimal delay performance given any initial state of queue backlog. We also demonstrate that our proposed policies are implementable in distributed manner to reduce communication overhead and computational cost. We evaluate our policies over single-hop networks with different network settings, and demonstrate its superiority compared to the policies that work well in networks not overloaded.**

I. Introduction

Reducing network delay is an old but never-obsolete research problem. Delay-sensitive applications including livestreaming, route directions, remote control, and financial trading, pose increasingly demanding delay requirements on network providers. For enterprises, prompt response to offer service guarantees high revenue and customer loyalty: Google pointed out that advertisement revenues decreases by 20% if web search delay increases from 0.4 seconds to 0.9 seconds, and according to Amazon, an extra 100ms response time decreases the sales by 1% [1]. Network infrastructure providers are dedicated to reducing delay under heavy traffic, for example using smart buffer architectures to absorb network spikes [2], and conducting traffic pacing and shaping [3], [4].

A key component of network delay is the queueing delay, which reflects the waiting time of a packet in network buffers until it can be served. Extensive prior works have been devoted to designing efficient transmission policies to reduce queueing delay, however exact characterization of queueing delay remains a hard problem [5], [6], and optimal scheduling policies have been obtained only for simple structures like parallel queues: The Join-the-shortest-queue policy is proven asymptotically delay-optimal [1], and a power-of-d-choices policy can reduce communication overhead [7].

In this paper, we study queueing delay minimization in *overloaded* networks. Network overload occurs when user demand surpasses network service capacity, under which data packets keep accumulating in network buffers. Overload occurs frequently in datacenter networks and server farms due to higher user demand [8]. Multiple reasons contribute to network overload: demand surge [9], denial-of-service attacks [10], and failure of network components [11], [12]. Network overload can result in detrimental consequences such as throughput reduction [13] and increased latency [14], which impairs quality of service.

We aim to minimize the queueing delay for traffic injected into the networks in finite-time horizon, motivated by the fact that overload is temporary in practice. We observe that many scheduling schemes that achieve good delay performance when network is not overloaded, perform poorly in overload. Consider the 2×1 single-hop network in Fig. 1. Packets arrive to ingress node s_i with rate λ_i $(i = 1, 2)$, and they are transmitted to the egress node d whose service rate is $\mu = 2$, sharing the buffer. The capacity of the two links are 4 and 2 respectively. First, suppose that at most one link can be activated at a time. In this case, maxweight scheduling has been shown to achieve near-optimal delay performance when the network is not overloaded $(\lambda_1 + \lambda_2 < \mu)$, which always activates the link connected to the ingress node with longer queue backlog [15]. However, maxweight fails in delay minimization under for example $(\lambda_1, \lambda_2) = (6, 3)$, as s_1 always has longer queue backlog than s_2, permanently blocking packets in the buffer of s_2. We will also show that even if simultaneous activation is allowed, transmitting packets with maximum rate is in general not delay-optimal under overload. Such counter-intuitive observation motivates us to rigorously design network policies for queueing delay minimization in overloaded networks.

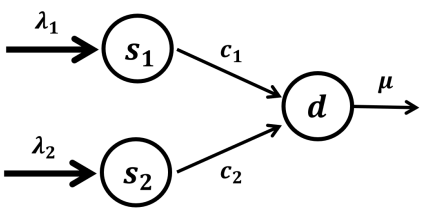

Fig. 1: 2×1 single-hop networks with $(c_1, c_2) = (4, 2)$

One challenge is the technical difficulty of modeling the queueing delay in overloaded networks, as the Little's law [16], the foundation of queueing delay analysis in stationary systems, no longer works in overloaded networks as the long-term expectation of delay is infinite. To overcome

the difficulty, in this paper we formulate the dynamics of traffic through a fluid queue model, where we view packet transmission through a link as a flow. We demonstrate that this continuous fluid model serves as a useful and effective tool to characterize queueing delay in overloaded networks. Meanwhile, it well approximates discrete packet transmission when the time unit is sufficiently small, and thus the results derived through the continuous model can guide the policy design in discrete settings. Based on the fluid model, we propose delay-optimal transmission policies in overloaded networks for traffic that arrive within a bounded time interval. The bounded time corresponds to the temporariness of network overload in practice. As an initial exploration on delay minimization in overloaded networks, we primarily present results in single-hop networks, and leave multi-hop networks as future work.

We summarize the contribution of this work as follows. (i) We formulate the queueing dynamics through a continuous fluid queue model, which can characterize the exact queueing delay of any specific packet arriving at arbitrary time under static flow. (ii) We prove the sufficient and necessary condition of a policy to minimize average queueing delay for any $N \times 1$ single-hop network, and a sufficient condition for any $N \times M$ single-hop network. The conditions demonstrate that a rate-proportional policy is delay-optimal. (iii) We prove that a queue-proportional policy, which only requires real-time queue information, can asymptotically minimize queueing delay under arbitrary initial network state. The queue-based policy is more realistic as packet arrival rates may be time-varying and are generally hard to access [17]. (iv) We show the our proposed policies can be implemented in a distributed manner to reduce communication and computation cost. (v) We validate our results in single-hop network structures under different network settings, compared with backpressure policy [18] and maximum-link-rate policy.

II. Modeling and Problem Formulation

A. Single-Hop Network Structure

We model a single-hop network as a bipartite graph $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V} := \{\mathcal{V}_I, \mathcal{V}_E\}$ denotes the set of ingress nodes by $\mathcal{V}_I$, and the set of egress nodes by $\mathcal{V}_E$, and the set of transmission links between $\mathcal{V}_I$ and $\mathcal{V}_E$ by $\mathcal{E}$. An $N \times M$ single-hop network consists of $|\mathcal{V}_I| = N$ ingress nodes and $|\mathcal{V}_E| = M$ egress nodes. Fig. 2 shows the topology, and real networks that can be modeled by the single-hop structure, including switched networks and server farms. The single-hop structure also serves as the basic structure of datacenter networks [19].

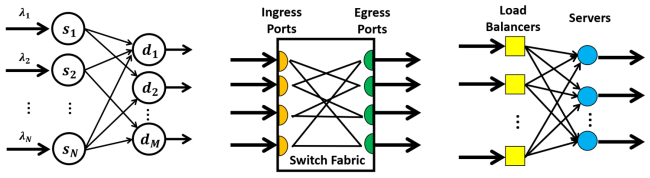

Fig. 2: (a) Single-hop structure; (b) Switch network; (c) Server farm with load balancers as ingress and servers as egress.

Denote the ith ingress node as s_i and the jth egress node as d_j. Each packet arrives at one of the ingress nodes and departs from one of the egress nodes. At each node, packets will be queued in a buffer waiting to be served, and we assume that packets follow the first-come-first-serve service which is common in real network infrastructures [16]. The queue length in node k at time t is denoted by $q_k(t)$.

The packet arrival rate at ingress node s_i, denoted by λ_i, can be interpreted as the average number of packets that arrive to s_i in a time unit. We use $\boldsymbol{\lambda} := \{\lambda_i\}_{i=1}^N$ to denote the packet arrival rate vector. Packets in the buffer of s_i are transmitted to an adjacent egress node d_j through link $(s_i, d_j) \in \mathcal{E}$. The transmission rate on link (s_i, d_j) at time t, denoted by $g_{s_i d_j}(t)$, represents the number of packets transmitted over (s_i, d_j) at time t. Each link (s_i, d_j) is associated with a capacity value $c_{s_i d_j}$, which is the maximum transmission rate, i.e., $0 \le g_{s_i d_j}(t) \le c_{s_i d_j},\ \forall (s_i, d_j) \in \mathcal{E}$. Moreover, $g_{s_i d_j}(t) = 0$ when $q_{s_i}(t) = 0$ for any $(s_i, d_j) \in \mathcal{E}$, which means no packets will be transmitted through (s_i, d_j) when there is no queue backlog in node s_i. We use $\mathbf{g}(t) := \{g_{s_i d_j}(t)\}_{(s_i,d_j)\in\mathcal{E}}$ to denote the transmission rate vector and $\mathbf{c} := \{c_{s_i d_j}\}_{(s_i,d_j)\in\mathcal{E}}$ to denote the capacity vector. Finally, the egress node d_j serves packets in a work-conserving manner: serving with its maximum rate, denoted by μ_j, whenever there exists queue backlog in the buffer. It is clear that work-conserving service at egress nodes is a necessary condition for delay optimality, and thus we can merely study the delay-optimal transmission rate $\mathbf{g}(t)$ between ingress and egress nodes.

We consider a fluid-queue model to characterize the dynamics in the network: Packets are modeled as continuous flows instead of discrete packet units, which means the queue length can be fractional. We show below that the fluid formulation facilitates characterization of queueing delay. The fluid model is based on the flow conservation law, which states that the net increase of queue length equals to the difference between the number of new arrivals and departures at a node at any time, i.e.,

$$\begin{cases} \dot{q}_{s_i}(t) = \lambda_i - \sum_{d_j:(s_i,d_j)\in\mathcal{E}} g_{s_i d_j}(t),\ \forall i = 1, \dots, N \\ \dot{q}_{d_j}(t) = \sum_{s_i:(s_i,d_j)\in\mathcal{E}} g_{s_i d_j}(t) - g_{d_j}(t),\ \forall j = 1, \dots, M \end{cases} \tag{1}$$

where under the work-conserving mechanism at egress nodes, $g_{d_j}(t) := \mu_j$ if $q_{d_j}(t) > 0$ and 0 otherwise.

Remark: The dynamics (1) offers a simplified framework for flow control analysis compared with discrete queueing model [18]. It is different from the fluid model defined in some prior works which captures the scaled limit of the queue backlog [13], [20], [21], an indicator for queue stability but not suited to study queueing delay.

B. Characterization of Queueing Delay

Based on (1), we can characterize the exact queueing delay of a packet that arrives at any specific time under static flow, where the transmission rate vector $\mathbf{g}$ is time-invariant. A packet goes through an ingress node and an egress node, which follows the two-node model in Fig. 3. Consider the red packet arriving at the node 1 at time t. The queueing delay at

node 1 is $q_1(t)/g_{12}$, as the red packet has to wait for all of the packets ahead of it to be served. The packet departs from node 1 and arrives at node 2 at time $t' := t + q_1(t)/g_{12}$, and thus the queueing delay at node 2 is $q_2(t')/\mu$. Therefore the total queueing delay for this packet is

$$\begin{aligned}&\frac{q_1(t)}{g_{12}} + \frac{q_2(t')}{\mu}\\ &= \frac{q_1(t)}{g_{12}} + \max\left\{\frac{q_2(t) + \frac{q_1(t)}{g_{12}}(g_{12}-\mu)}{\mu}, 0\right\}\\ &= \max\left\{\frac{q_1(t)+q_2(t)}{\mu}, \frac{q_1(t)}{g_{12}}\right\}\end{aligned} \tag{2}$$

where the max term in the second line is to take into account that $q_2(t')$ may reach 0 when $g_{12} < \mu$. It indicates that larger g_{12} such that $g_{12} \geq \frac{q_1(t)}{q_1(t)+q_2(t)}\mu$ guarantees minimum queueing delay for the red packet, while further increasing g_{12} does not make a difference as the delay is bottlenecked by μ.

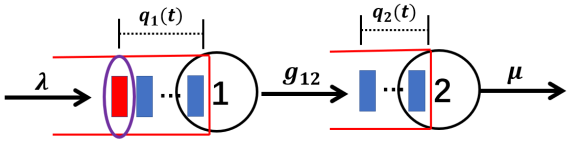

Fig. 3: Queueing delay for a packet in a two-node model.

We now apply the two-node model to a general $N \times M$ single-hop structure. Denote the queueing delay of a packet that arrives at ingress node s_i at time t, and departs at egress node d_j, as $D_{s_i d_j}(t)$. Let $t' = t + \frac{q_{s_i}(t)}{\sum_{d_j:(s_i,d_j)\in\mathcal{E}} g_{s_i d_j}}$, and thus

$$D_{s_i d_j}(t) = \frac{q_{s_i}(t)}{\sum_{d_j:(s_i,d_j)\in\mathcal{E}} g_{s_i d_j}} + \frac{q_{d_j}(t')}{\mu_j} \tag{3}$$

which consists of the queueing delay at s_i and d_j respectively. Equation (3) can be further transformed into a function of $\mathbf{g}$ as in (2), which facilitates our derivation of optimal transmission policies in later sections.

Remark: In this work, we purely shed light on queueing delay for two reasons: (i) The preprocessing and propagation delay are with very small variations, and almost independent of network policies; (ii) In overload the queueing delay dominates compared with preprocessing and propagation delay, as overload causes buffer congestion which leads to severe increase of queue backlog.

C. Problem Formulation

We now formulate the delay minimization problem in a general $N \times M$ single-hop network. We focus on two queueing delay metrics: (i) the average delay $\bar{D}_{\text{avg}}$ (ii) the maximum ingress delay $\bar{D}_{\max}$, with formal definitions introduced later. They measure the delay performance from different angles: $\bar{D}_{\text{avg}}$ reflects the overall delay performance of all the arrived packets, while $\bar{D}_{\max}$ measures the average delay of the packets from different ingress nodes and takes the largest one as the maximum ingress delay. In practice, $\bar{D}_{\text{avg}}$ is relevant to systems whose overall performance is important, for example datacenters and server farms, while $\bar{D}_{\max}$ represents the fairness of processing tasks from different sources, which is important in parallel task completion, coflow analysis, etc.

We focus on both metrics for packets that arrive in some bounded time interval $[t_0, t_0 + T]$. The reason we consider bounded intervals is two-folded: (i) Network overload is a temporary event in practice; (ii) Both metrics approach infinity in unbounded intervals in overloaded networks under any policy.

We now define $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ formally. Denote the average queueing delay of packets arriving at ingress node s_i within $[t_0, t_0 + T]$ as $\bar{D}_i$, which under an $N \times M$ structure is[1]

$$\bar{D}_i = \frac{1}{T}\int_{t_0}^{t_0+T} \sum_{j=1}^{M} \frac{g_{s_i d_j}}{\sum_{k=1}^{M} g_{s_i d_k}} d_{s_i d_j}(t)dt, \ \forall i = 1, \ldots, N, \tag{4}$$

which contains two layers of averaging: (i) averaging over different arrival times t within $[t_0, t_0 + T]$, which is an unweighted integral; (ii) averaging over packets sent to different egress nodes, which is weighted by $g_{s_i d_j}/\sum_{k=1}^{M} g_{s_i d_k}$, i.e., the portion of packets that arrive at s_i at time t and will depart from the network from d_j. Then the two delay metrics to be optimized in this paper can be formulated as:

$$\bar{D}_{\text{avg}} = \sum_{i=1}^{N} \frac{\lambda_i}{\sum_{j=1}^{N}\lambda_j}\bar{D}_i, \tag{5}$$

$$\bar{D}_{\max} = \max_{i=1,\ldots,N} \bar{D}_i. \tag{6}$$

Based on (4), the $\bar{D}_{\text{avg}}$ in (5) introduces an additional layer of averaging, weighted by the ratio $\lambda_i T / \left(\sum_{j=1}^{N}\lambda_j T\right) = \lambda_i / \left(\sum_{j=1}^{N}\lambda_j\right)$ that is the portion of the packets that arrive at ingress node s_i within the time interval $[t_0, t_0 + T]$, while (6) takes the maximum over all $\bar{D}_i$'s, which represents the largest delay among all ingress nodes from s_1 to s_N.

D. Network Overload

We define overload in single-hop networks before introducing our results on minimizing $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$.

Definition 1. *A $N \times M$ single-hop network is overloaded if there is no transmission rate vector $\mathbf{g}$ such that*

$$\begin{cases}\sum_{d_j:(s_i,d_j)\in\mathcal{E}} g_{s_i d_j} \geq \lambda_i, \ \forall j = 1, \ldots, M\\ \sum_{s_i:(s_i,d_j)\in\mathcal{E}} g_{s_i d_j} \leq \mu_j, \ \forall i = 1, \ldots, N\\ g_{s_i d_j} \in [0, c_{s_i d_j}], \ \forall (s_i, d_j) \in \mathcal{E}\end{cases} \tag{7}$$

This definition of overload guarantees that there is no feasible transmission rate vector $\mathbf{g}$ that prevents all ingress and egress nodes from queue overflow. This may occur when demand drastically increases during peak hours, or lack of sufficient transmission resources due to failure, misconfiguration, and resources being occupied by other flows. We purely focus on delay minimization in overloaded networks, since for networks not overloaded, (7) is feasible and thus it is trivial to achieve zero queueing delay in the steady state under the fluid model by finding one of the feasible solutions of (7).

[1]In the following, for $\forall (s_i, d_j) \notin \mathcal{E}$, $g_{s_i d_j}$ equals to 0.

III. Delay Minimization on $N \times 1$ Structure

In this section, we present delay optimality results for $N \times 1$ single-hop structures, where all arrivals share a single egress node. This structure represents a single server that receives requests from multiple sources, or the case that packets arrive from multiple upstream links and share a single port of a downstream switch between two stages in a datacenter [19]. We unveil a counterintuitive result that serving with larger rates may increase delay, and prove that setting transmission rates that follow the proportion of packet arrival rates to different ingress nodes is delay-optimal, which we term as the *rate-proportional* policy. We demonstrate that the rate-proportional policy not only minimizes queueing delay, but also saves transmission resources and thus reduces power consumption compared with serving with maximum rates.

Given that no prior works applied the fluid model to delay minimization in overloaded networks, we first introduce the result with detailed analysis over a 2×1 single-hop network, and then extend the result to the $N \times 1$ structure. To present the intuition behind the rate-proportional policy in this section, we (i) present the main results under unlimited link capacity, and then demonstrate that introducing limited link capacity does not affect the form of the delay-optimal policy; (ii) consider time-invariant transmission rate vector $\mathbf{g}$, which induces static flow and serves as a foundation of real-time queue-based policy design in Section V.

A. 2×1 Overloaded Networks

A 2×1 single-hop network contains 2 ingress nodes and 1 egress node. We present the sufficient and necessary condition for $\mathbf{g} := (g_1, g_2)$[2] to minimize delay metrics $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ in Proposition 1. We derive Proposition 1 under zero initial queues for conciseness and highlighting the intuition behind the rate-proportional pattern, and we clarify how initial queue length affects the result and point out that in practical overload the impact is generally negligible.

Proposition 1. *Given a 2×1 single-hop network. For $\forall T > 0$, the set of $\mathbf{g}$'s that minimize $\bar{D}_{avg}$ and $\bar{D}_{max}$ of the packets that arrive within $[t_0, t_0 + T]$ where $\mathbf{q}(t_0) = \mathbf{0}$ is*

$$\left\{(g_1 + g_2 \geq \mu) \cap \left(\frac{\lambda_1}{g_1} = \frac{\lambda_2}{g_2}\right)\right\} \cup \{g_1 \geq \lambda_1, g_2 \geq \lambda_2\} \quad (8)$$

under which $\bar{D}_{avg} = \bar{D}_{max} = \frac{T}{2\mu}\max\{\lambda_1 + \lambda_2 - \mu, 0\}$.

Proof. The main idea of the proof is that we divide the feasible flow region of $\mathbf{g} = (g_1, g_2)$, which is $[0, \infty) \times [0, \infty)$ in a network with unlimited capacity, into 4 regions:

$$\begin{cases} \mathcal{R}_1 := \{\mathbf{g} \mid g_1 \in [0, \lambda_1], g_2 \in [0, \lambda_2]\}, \\ \mathcal{R}_2 := \{\mathbf{g} \mid g_1 \in [\lambda_1, \infty), g_2 \in [\lambda_2, \infty)\}, \\ \mathcal{R}_3 := \{\mathbf{g} \mid g_1 \in [\lambda_1, \infty), g_2 \in [0, \lambda_2]\}, \\ \mathcal{R}_4 := \{\mathbf{g} \mid g_1 \in [0, \lambda_1], g_2 \in [\lambda_2, \infty)\}, \end{cases} \quad (9)$$

and we identify the optimal $\mathbf{g}$'s restricted in each of these regions, denoted as $\mathbf{g}^*_{(1)}, \mathbf{g}^*_{(2)}, \mathbf{g}^*_{(3)}, \mathbf{g}^*_{(4)}$ respectively. We show that each $\mathbf{g}^*_{(i)}$ leads to the same average queueing delay $\bar{D}_{\text{avg}} = \frac{T}{2\mu}\max\{(\lambda_1 + \lambda_2 - \mu), 0\}$ and the same maximum ingress delay $\bar{D}_{\max} = \frac{T}{2\mu}\max\{(\lambda_1 + \lambda_2 - \mu), 0\}$.

Before analyzing the above four cases in order, we first define $D_i(t)$ as the total queueing delay of a packet injected into s_i at time t. Then according to (3), for $i = 1, 2$,

$$\begin{aligned} D_i(t) &= \frac{q_{s_i}(t)}{g_i} + \max\left\{\frac{q_d(t) + \frac{q_{s_i}(t)}{g_i}(g_1 + g_2 - \mu)}{\mu}, 0\right\} \\ &= \begin{cases} \frac{1}{\mu}\left(q_d(t) + \frac{q_{s_i}(t)}{g_i}(g_1 + g_2)\right), & g_1 + g_2 \geq \mu \\ \frac{q_{s_i}(t)}{g_i}, & g_1 + g_2 < \mu \end{cases} \end{aligned}$$

due to $\mathbf{q}(t_0) = \mathbf{0}$ which guarantees that when $g_1 + g_2 < \mu$, $q_d(t)$ will keep zero and thus the only queueing delay is at the ingress nodes. The average delay for packets arrived to ingress node s_i within $[t_0, t_0 + T]$ is $\bar{D}_i = \frac{1}{T}\int_{t_0}^{t_0+T} D_i(t)dt,\ i = 1, 2$ which connects the service rate $\mathbf{g}$ with the two metrics $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ given by (5) and (6).

Case 1: $\mathcal{R}_1 := \{\mathbf{g} \mid g_1 \in [0, \lambda_1], g_2 \in [0, \lambda_2]\}$

In $\mathcal{R}_1$, we first consider the case when $g_1 + g_2 \geq \mu$.

$$\begin{aligned} \bar{D}_i &:= \frac{1}{T}\int_{t_0}^{t_0+T} D_i(t)dt = \frac{1}{T}\int_{t_0}^{t_0+T} \frac{q_d(t) + \frac{q_{s_i}(t)}{g_i}(g_1 + g_2)}{\mu}dt \\ &= \frac{1}{T\mu}\int_{t_0}^{t_0+T} (t - t_0)\max\{g_1 + g_2 - \mu, 0\} \\ &\qquad + \frac{g_1 + g_2}{g_i}(t - t_0)\max\{\lambda_i - g_i, 0\}dt \\ &= \frac{T}{2\mu}\left(\lambda_1\frac{g_1 + g_2}{g_i} - \mu\right),\ i = 1, 2 \end{aligned}$$

Then according to (5) and (6),

$$\begin{aligned} \bar{D}_{\text{avg}} &= \frac{\lambda_1}{\lambda_1 + \lambda_2}\bar{D}_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2}\bar{D}_2 \\ &= \frac{T}{2\mu}\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\left(\lambda_1\frac{g_1 + g_2}{g_1} - \mu\right) + \frac{\lambda_2}{\lambda_1 + \lambda_2}\left(\lambda_2\frac{g_1 + g_2}{g_2} - \mu\right)\right) \end{aligned}$$

and

$$\bar{D}_{\max} = \max\{\bar{D}_1, \bar{D}_2\} = \frac{T}{2\mu}\left\{\left(\lambda_1\frac{g_1 + g_2}{g_1} - \mu\right), \left(\lambda_2\frac{g_1 + g_2}{g_2} - \mu\right)\right\}$$

For $\bar{D}_{\text{avg}}$, we can obtain by Cauchy-Schwartz inequality that the optimal solutions are all $\mathbf{g}$ that satisfy $g_1 + g_2 \geq \mu,\ \frac{g_1}{g_2} = \frac{\lambda_1}{\lambda_2}$ under which the average delay is $\bar{D}_{\text{avg}} = \frac{T}{2\mu}(\lambda_1 + \lambda_2 - \mu)$. For $\bar{D}_{\max}$, we can obtain that the set of $\mathbf{g}$'s that satisfy (8) achieves the minimum $\bar{D}_{\max} = \frac{T}{2\mu}(\lambda_1 + \lambda_2 - \mu)$.

We then consider the case when $g_1 + g_2 \leq \mu$. In this case there will be no queue backlog in the egress node, and thus

$$\begin{aligned} \bar{D}_i &= \frac{1}{T}\int_{t_0}^{t_0+T} \frac{q_{s_i}(t)}{g_i}dt = \frac{\max\{\lambda_i - g_i, 0\}}{g_i T}\int_{t_0}^{t_0+T} (t - t_0)dt \\ &= \frac{T}{2}\frac{\lambda_i - g_i}{g_i},\ i = 1, 2. \end{aligned}$$

Therefore

$$\begin{cases} \bar{D}_{\text{avg}} = \frac{T}{2(\lambda_1 + \lambda_2)}\left(\frac{\lambda_1^2}{g_1} + \frac{\lambda_2^2}{g_2} - \lambda_1 - \lambda_2\right) \\ \bar{D}_{\max} = \frac{T}{2}\max\left\{\frac{\lambda_1 - g_1}{g_1}, \frac{\lambda_2 - g_2}{g_2}\right\} \end{cases}$$

Then under $g_1 + g_2 \leq \mu$, the optimal metric values are $\bar{D}_{\text{avg}} = \bar{D}_{\max} = \frac{T}{2\mu}(\lambda_1 + \lambda_2 - \mu)$, achieved only at $g_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2}\mu,\ g_2 = \frac{\lambda_2}{\lambda_1 + \lambda_2}\mu$, which is on the boundary $g_1 + g_2 = \mu$.

[2] In $N \times 1$ networks, we simplify the notation $g_{s_i d}$ to g_i, $c_{s_i d}$ to c_i.

Case 2: $\mathcal{R}_2 := \{\mathbf{g} \mid g_1 \in [\lambda_1, \infty), g_2 \in [\lambda_2, \infty)\}$

When $g_1 \geq \lambda_1$, $d_{s_1}(t) = \frac{q_{s_1}(t)}{g_1} + \frac{q_d\left(t + \frac{q_{s_1}(t)}{g_1}\right)}{\mu} = \frac{q_d(t)}{\mu}$ as $q_{s_1}(t)$ keeps 0 since $q_{s_1}(t_0) = 0$, which means the queueing delay only occurs at node d. Similar for $d_{s_2}(t)$. Since $\lambda_1 + \lambda_2 > \mu$, packets will accumulate at node d and at time t, and $q_d(t) = (\lambda_1 + \lambda_2 - \mu)t$. Thus

$$d_{s_1}(t) = d_{s_2}(t) = \frac{q_d(t)}{\mu} = \frac{\lambda_1 + \lambda_2 - \mu}{\mu} t.$$

and $D_{s_1} = \frac{1}{T}\int_{t_0}^{t_0+T} d_{s_1}(t)dt = \frac{T}{2\mu}(\lambda_1 + \lambda_2 - \mu) = D_{s_2}$, and hence $\bar{D}_{\text{avg}} = \bar{D}_{\max} = \frac{T}{2\mu}(\lambda_1 + \lambda_2 - \mu)$ for $\forall \mathbf{g} \in \mathcal{R}_2$.

Case 3: $\mathcal{R}_3 := \{\mathbf{g} \mid g_1 \in [\lambda_1, \infty), g_2 \in [0, \lambda_2]\}$

Based on the derivation in case 1 and 2 respectively, we have $D_{s_1} = \frac{T}{2\mu}(\lambda_1 + g_2 - \mu)$ as packets arrived at s_1 only suffer from delay at d, and $D_{s_2} = \frac{T}{2\mu}\left(\frac{\lambda_1 + g_2}{g_2}\lambda_2 - \mu\right)$ where packets arrived at s_2 suffer from delay at s_2 and d. Therefore we can verify easily that any optimal $\mathbf{g} \in \mathcal{R}_3$ that achieves minimum $\bar{D}_{\text{avg}} = \bar{D}_{\max} = \frac{T}{2\mu}(\lambda_1 + \lambda_2 - \mu)$ satisfies $g_2 = \lambda_2$.

Case 4: $\mathcal{R}_4 := \{\mathbf{g} \mid g_1 \in [0, \lambda_1], g_2 \in [\lambda_2, \infty)\}$ Similar to case 3, where any optimal $\mathbf{g} \in \mathcal{R}_4$ satisfies $g_1 = \lambda_1$. □

We term (8) as the *delay-optimal region* of $\mathbf{g}$. We observe that delay-optimal regions for both $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ in a 2×1 single-hop network are the same. It means that we can simultaneously achieve minimum average delay and optimal delay fairness of packets arrived to different ingress nodes. The delay-optimal region (8) is illustrated as the blue region in Fig. 4(a), consists of a line segment connecting the points $\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\mu, \frac{\lambda_2}{\lambda_1+\lambda_2}\mu\right)$ and (λ_1, λ_2), and the polytope $\{\mathbf{g} \mid g_i \geq \lambda_i, i = 1, 2\}$. $\mathcal{R}_1$ to $\mathcal{R}_4$ in (9) are also depicted.

We have the following insights from (8): (i) Setting g_i no less than λ_i for both $i = 1, 2$ achieves minimum queueing delay, while further increasing g_1 and g_2 does not make a difference. This is because for any $\mathbf{g}$ that $g_i \geq \lambda_i$, the buffers of s_1 and s_2 are empty, thus all queueing delay being at the egress node, and hence the bottleneck of delay is only μ. (ii) In $\mathcal{R}_1$, it is possible to achieve the global optimum delay as serving in $\mathcal{R}_2$ by setting $\mathbf{g}$ such that $g_1/g_2 = \lambda_1/\lambda_2$, i.e., to follow the arrival rate ratio, and $g_1 + g_2 \geq \mu$, i.e., to guarantee maximum throughput. It indicates that using the rate-proportional policy suffices to minimize delay, which requires much fewer resources and consumes much less power compared with serving with higher rates as in $\mathcal{R}_2$. (iii) It is not true that serving with higher rates leads to lower queueing delay. For example serving with transmission rates in $\mathcal{R}_3$ and $\mathcal{R}_4$ is inferior to controlling the transmission rates on the optimal line segment in $\mathcal{R}_1$. The counter-intuition is because packets from s_1 and s_2 share an egress node, where the imbalance between g_1 and g_2 leads to severe delay increase of packets arrived to one of s_1 and s_2.

We now extend Theorem 1 to limited link capacity. Based on (8), we can obtain directly that the delay-optimal region for limited capacity case is simply the intersection of (8) and $\{\mathbf{g} \mid g_1 \leq c_1, g_2 \leq c_2\}$ as limited capacity does not affect the proof. We illustrates the optimal region when $\lambda_i > c_i$, $i = 1, 2$ in Fig. 4(b), where only the solid blue line segment is the optimal region. It is clear that serving both links with maximum rates

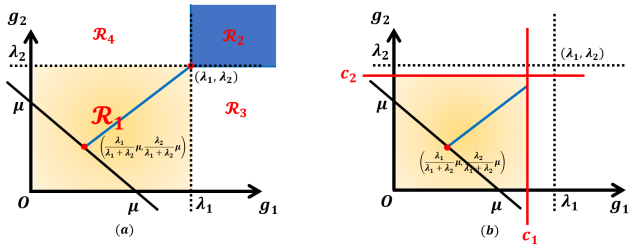

Fig. 4: Delay-optimal region in a 2×1 single-hop network: (a) unlimited capacity; (b) limited capacity ($c_i \leq \lambda_i$, $i = 1, 2$)

is not delay-optimal in general, which suggests that refined control of link rates according to (8) is required for delay minimization.

Finally we discuss the impact of the initial queue length $\mathbf{q}(t_0)$ over the result. Following the proof we can straightforwardly obtain the delay-optimal policy among $\mathcal{R}_1$ to be

$$\frac{g_1}{g_2} = \sqrt{\frac{\lambda_1(\lambda_1 + q_{s_1}(t_0)/T)}{\lambda_2(\lambda_2 + q_{s_2}(t_0)/T)}}, \tag{10}$$

under which any ingress queue length will not reduce to zero. Note that (10) also follows the rate-proportional pattern, and $g_1/g_2 \to \lambda_1/\lambda_2$ when $q_{s_1}(t_0)/q_{s_2}(t_0) \to \lambda_1/\lambda_2$[3] or T is sufficiently large. For $\mathcal{R}_2$, $\mathcal{R}_3$ and $\mathcal{R}_4$, derivation is of higher complexity as we need to analyze if the queue length at the ingress nodes will or will not change from non-zero to zero within $[t_0, t_0 + T]$, which involves at least two cases for each ingress node. In practice, however, (i) the initial queue length $q_{s_1}(t_0)$ and $q_{s_2}(t_0)$ are generally very small before overload occurs[4], and (ii) in practice we generally care about rate control for relatively long T instead of instantaneous overload. Therefore $q_{s_1}(t_0)/T$ is generally small and thus (10) is approximately $g_1/g_2 = \lambda_1/\lambda_2$, matching (8). For the above reasons and proof conciseness, we neglect initial queue length in the results below, and verify empirically in Section VI that initial queue length does not affect the overall performance.

B. $N \times 1$ Single-Hop Networks

Generalizing the idea in 2×1 single-hop networks, the result for $N \times 1$ single-hop networks is stated in Theorem 1, where the rate-proportional policy remains delay-optimal.

Theorem 1. *Given a $N \times 1$ single-hop structure. For $\forall T > 0$, the set of $\mathbf{g} = \{g_i\}_{i=1}^N$ that minimizes $\bar{D}_{avg}$ and $\bar{D}_{max}$ of the packets that arrive within $[t_0, t_0 + T]$ where $\mathbf{q}(t_0) = \mathbf{0}$ is*

$$\left\{\left(\sum_{i=1}^N g_i \geq \mu\right) \cap \left(\frac{\lambda_1}{g_1} = \cdots = \frac{\lambda_N}{g_N}\right)\right\} \cup \{g_i \geq \lambda_i,\ \forall i = 1, \ldots, N\}, \tag{11}$$

under which $\bar{D}_{avg} = \bar{D}_{max} = \frac{T}{2\mu}\max\{\sum_{i=1}^N \lambda_i - \mu, 0\}$.

The proof of Theorem 1 is similar to Proposition 1, the 2×1 case. The delay-optimal region is in the same form:

[3]This fact is related to the queue-based policy design in Section V.

[4]Under fluid model (1), the ideal case of practical queueing systems, when network is not overloaded, the queue length at any node keeps 0.

a line segment connecting two points $\left\{\frac{\lambda_i}{\sum_{j=1}^N \lambda_j}\mu\right\}_{i=1}^N$ and $\{\lambda_i\}_{i=1}^N$, and a polytope, the i-th entry of any point in which is greater than λ_i. This indicates that to achieve minimum delay, if there exists one link with service rate higher than its corresponding packet arrival rate, then all the other links should be as well; if the service rates of all the links are less than the corresponding arrival rate at the ingress node, then the rate-proportional policy, under which the link rates follow the ratio of arrival rates, minimizes delay.

IV. Delay Minimization in $N \times M$ Structure

In this section, we follow the idea in $N \times 1$ networks to show delay-optimal policy in $N \times M$ single-hop networks. The $N \times M$ structure is more complex where the number of links is one order-of-magnitude higher, for which we find it hard to derive an exact form of the delay-optimal region. Nevertheless, we identify a sufficient condition for a policy to achieve globally minimum delay, which is a subset of the delay-optimal region. Any policy that satisfies the sufficient condition follows the *rate-proportional* pattern, combining both $\boldsymbol{\lambda}$-*proportional* pattern as $N \times 1$ case, and $\boldsymbol{\mu}$-*proportional* pattern which means to control link rates so that packet injection rates to egress nodes matches the ratio among $\{\mu_1, \ldots, \mu_M\}$.

We present the results under unlimited capacities, and the extension to limited capacities is similar to $N \times 1$ case, simply with an intersection of capacity constraints. We first introduce the proof details and intuition behind the results using a 2×2 structure, stated in Proposition 2, and extend it to $N \times M$ case.

Proposition 2. *Given a 2×2 single-hop network. For $\forall T > 0$, a sufficient condition for the transmission rate vector $\mathbf{g}$ to minimize $\bar{D}_{avg}$ and $\bar{D}_{max}$ of the packets that arrive within $[t_0, t_0 + T]$ where $\mathbf{q}(t_0) = \mathbf{0}$ is*

$$\begin{cases} \frac{g_{11}+g_{12}}{g_{21}+g_{22}} = \frac{\lambda_1}{\lambda_2}, \ \frac{g_{11}+g_{21}}{g_{12}+g_{22}} = \frac{\mu_1}{\mu_2} \\ g_{11} + g_{21} \geq \mu_1, \ g_{12} + g_{22} \geq \mu_2 \end{cases} \tag{12}$$

under which $\bar{D}_{avg} = \bar{D}_{max} = \frac{T}{2(\mu_1+\mu_2)} \max\{\lambda_1 + \lambda_2 - \mu_1 - \mu_2, 0\}$, equal the global minimum among all feasible $\mathbf{g}$'s: $g_{11}, g_{12}, g_{21}, g_{22} \in [0, \infty)$. Moreover, among $\{\mathbf{g} \mid g_{11} + g_{12} \leq \lambda_1, \ g_{21} + g_{22} \leq \lambda_2\}$, (8) is the sufficient and necessary condition to minimize $\bar{D}_{avg}$ and $\bar{D}_{max}$.

We defer the proof idea to appendix. Proposition 2 demonstrates that a policy which (i) follows $\boldsymbol{\lambda}$-ratio: the ratio of the sum of service rates of the downstream links of s_1 and s_2 is equal to the ratio λ_1/λ_2; (ii) follows $\boldsymbol{\mu}$-ratio: the ratio of the sum of service rates of the upstream links of d_1 and d_2 is equal to the ratio μ_1/μ_2; (iii) guarantees maximum throughput $g_{11}+g_{21} \geq \mu_1, \ g_{12}+g_{22} \geq \mu_2$, can achieve globally minimum delay. Although the result only characterizes a subset of delay optimal $\mathbf{g}$'s, (12) has characterized the minimum total link rates required to minimize delay, as it is the sufficient and necessary condition when $\mathbf{g}$ is limited within $\{\mathbf{g} \mid g_{11} + g_{12} \leq \lambda_1, \ g_{21} + g_{22} \leq \lambda_2\}$. This result shows the way to utilize minimum transmission resource to achieve minimum delay.

We extend the result to $N \times M$ systems stated in Theorem 2, which can be proved in the same way as Proposition 2.

Theorem 2. *Given a $N \times M$ single-hop network. For $\forall T > 0$, a sufficient condition to globally minimize both metrics of the packets that arrive within $[t_0, t_0 + T]$ where $\mathbf{q}(t_0) = \mathbf{0}$ is*

$$\begin{cases} \frac{\sum_{k=1}^M g_{ik}}{\sum_{k=1}^M g_{jk}} = \frac{\lambda_i}{\lambda_j}, \forall i, j = 1, \ldots, N \\ \frac{\sum_{k=1}^N g_{ki}}{\sum_{k=1}^N g_{kj}} = \frac{\mu_i}{\mu_j}, \forall i, j = 1, \ldots, M \\ \sum_{k=1}^N g_{kj} \geq \mu_j, \forall j = 1, \ldots, M \end{cases} \tag{13}$$

under which $\bar{D}_{avg} = \bar{D}_{max} = \frac{T}{2\sum_{j=1}^M \mu_j} \max\{\sum_{i=1}^N \lambda_i - \sum_{j=1}^M \mu_j, 0\}$. Moreover, among $\{\mathbf{g} \mid \sum_{j=1}^M g_{s_i d_j} \leq \lambda_i, \ \forall i = 1, \ldots, N\}$, (13) is the sufficient and necessary condition.

V. Queue-based Policy for Delay Minimization

We have proved the delay-optimal transmission policies. However, they require the complete knowledge of network parameters $(\boldsymbol{\lambda}, \mathbf{c}, \boldsymbol{\mu})$ which in real networks may not be available [17]. In practice, the queue backlog $\mathbf{q}(t)$ is often accessible in real-time. We prove that there exists a queue-based policy, with the idea of being *queue-proportional*, that achieves minimum queueing delay asymptotically. The queue-proportional idea means to set the link rates according to the ratio of queue backlogs in ingress and egress nodes.

A. 2×1 Single-hop Structure

We start from 2×1 single-hop structure to gain intuition of the optimal queue-based policy. The delay-optimality of any $\mathbf{g}$ such that $\frac{g_1}{g_2} = \frac{\lambda_1}{\lambda_2}$ in Section III implies that the following rate control policy where $\mathbf{g}(t)$ satisfies $\frac{g_1(t)}{g_2(t)} = \frac{\dot{q}_{s_1}(t)}{\dot{q}_{s_2}(t)}$ can minimize the average delay $\bar{D}_{\text{avg}}$ and maximum ingress delay $\bar{D}_{\text{max}}$, because $\frac{g_1(t)}{g_2(t)} = \frac{\dot{q}_{s_1}(t)}{\dot{q}_{s_2}(t)} = \frac{\lambda_1 - g_1(t)}{\lambda_2 - g_2(t)} \overset{(*)}{=} \frac{\lambda_1}{\lambda_2}$[5]. The policy $\frac{g_1(t)}{g_2(t)} = \frac{\dot{q}_{s_1}(t)}{\dot{q}_{s_2}(t)}$, although still dependent upon $\boldsymbol{\lambda}$, hints at an idea to propose the following queue-based policy when $q_{s_1}(t), q_{s_2}(t) > 0$,

$$\frac{g_1(\mathbf{q}(t))}{g_2(\mathbf{q}(t))} = \frac{q_{s_1}(t)}{q_{s_2}(t)}, \ g_1(\mathbf{q}(t)) + g_2(\mathbf{q}(t)) \geq \mu, \tag{14}$$

which determines the transmission rates based on the ratio of the queue backlogs at ingress nodes. We show in Proposition 3 that (14) achieves optimal $\bar{D}_{\text{avg}}$ and $\bar{D}_{\text{max}}$ with no initial queue backlogs at the source nodes, and in Proposition 4 that (14) asymptotically converges to the delay-optimal policy (8) starting from arbitrary initial queue backlogs.

Proposition 3. *With $q_{s_1}(t_0) = q_{s_2}(t_0) = 0$, then the policy (14) achieves minimum $\bar{D}_{avg}$ and $\bar{D}_{max}$.*

Proof. Initially, take $\epsilon \to 0$

$$\frac{g_1(\mathbf{q}(t_0+\epsilon))}{g_2(\mathbf{q}(t_0+\epsilon))} = \frac{q_{s_1}(t_0+\epsilon)}{q_{s_2}(t_0+\epsilon)} = \frac{\int_{t_0}^{t_0+\epsilon} \lambda_1 - g_1(\mathbf{q}(s))ds}{\int_{t_0}^{t_0+\epsilon} \lambda_2 - g_2(\mathbf{q}(s))ds}$$
$$= \frac{\int_{t_0}^{t_0+\epsilon} \lambda_1 ds - g_1(\mathbf{q}(t_0+\alpha_1\epsilon))}{\int_{t_0}^{t_0+\epsilon} \lambda_2 ds - g_2(\mathbf{q}(t_0+\alpha_2\epsilon))} \to \frac{\int_{t_0}^{t_0+\epsilon} \lambda_1 ds}{\int_{t_0}^{t_0+\epsilon} \lambda_2 ds} = \frac{\lambda_1}{\lambda_2}$$

[5]The $(*)$ holds since if for some $a, b, c, d \neq 0$ and $a + c, b + d \neq 0$, $a/b = c/d$, then $a/b = c/d = (a + c)/(b + d)$.

where $\alpha_1, \alpha_2 \in [0,1]$. Then in the time interval $[t_0+\epsilon, t_0+2\epsilon]$,

$$\frac{g_1(\mathbf{q}(t_0+2\epsilon))}{g_2(\mathbf{q}(t_0+2\epsilon))} = \frac{q_{s_1}(t_0+2\epsilon)}{q_{s_2}(t_0+2\epsilon)} = \frac{q_{s_1}(t_0+\epsilon)+\epsilon\dot{q}_{s_1}}{q_{s_2}(t_0+\epsilon)+\epsilon\dot{q}_{s_2}}$$
$$= \frac{q_{s_1}(t_0+\epsilon)+\epsilon(\lambda_1 - g_1(\mathbf{q}(t_0+\epsilon)))}{q_{s_2}(t_0+\epsilon)+\epsilon(\lambda_2 - g_2(\mathbf{q}(t_0+\epsilon)))} = \frac{\lambda_1}{\lambda_2}$$

Iteratively, we have

$$\frac{g_1(\mathbf{q}(t))}{g_2(\mathbf{q}(t))} = \frac{q_{s_1}(t)}{q_{s_2}(t)} = \frac{\lambda_1}{\lambda_2},\ \forall t$$

which guarantees minimization of $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ according to Proposition 1. □

Proposition 4. *With arbitrary* $\mathbf{q}(t_0)$, (14) *converges to the state* $\lim_{t\to\infty} \frac{g_1(\mathbf{q}(t))}{g_2(\mathbf{q}(t))} = \frac{\lambda_1}{\lambda_2}$ *which minimizes* $\bar{D}_{avg}$ *and* $\bar{D}_{max}$.

Proof. Under (14), we have when $t \to \infty$,

$$\left|\frac{q_1(t)}{q_2(t)} - \frac{\lambda_1}{\lambda_2}\right| = \left|\frac{q_{s_1}(t_0)+\int_{t_0}^t \lambda_1 - g_1(\mathbf{q}(s))ds}{q_{s_2}(t_0)+\int_{t_0}^t \lambda_2 - g_2(\mathbf{q}(s))ds} - \frac{\lambda_1}{\lambda_2}\right|$$
$$\overset{\text{L'hos}}{\to} \left|\frac{\lambda_1 - g_1(\mathbf{q}(t))}{\lambda_2 - g_2(\mathbf{q}(t))} - \frac{\lambda_1}{\lambda_2}\right|$$

and

$$\frac{g_1(\mathbf{q}(t))}{g_2(\mathbf{q}(t))} = \frac{q_{s_1}(t)}{q_{s_2}(t)} \overset{\text{L'hos}}{\to} \frac{\dot{q}_{s_1}(t)}{\dot{q}_{s_2}(t)} = \frac{\lambda_1 - g_1(\mathbf{q}(t))}{\lambda_2 - g_2(\mathbf{q}(t))}$$

where "L'hos" means using L'hospital's rule. This indicates $\frac{\lambda_1 - g_1(\mathbf{q}(t))}{\lambda_2 - g_2(\mathbf{q}(t))} \to \frac{\lambda_1}{\lambda_2}$. Thus $\left|\frac{q_{s_1}(t)}{q_{s_2}(t)} - \frac{\lambda_1}{\lambda_2}\right| \to 0$ when $t \to \infty$. □

Based on Proposition 4, given arbitrary initial queues, although not necessarily being delay-optimal at any time, the policy (14) keeps driving the queueing dynamics to the state under which delay is minimized globally.

B. Extension to General Single-hop Structures

1) $N \times 1$ Single-hop Structure: The extension of policy (14) to $N \times 1$ structure is straightforward, under which the transmission rates are proportional to the current queue backlogs in their corresponding source nodes, i.e.,

$$\frac{g_{s_i d}(\mathbf{q}(t))}{g_{s_j d}(\mathbf{q}(t))} = \frac{q_{s_i}(t)}{q_{s_j}(t)},\ \forall i \neq j, \quad \sum_{i=1}^{N} g_{s_i d}(\mathbf{q}(t)) \geq \mu \tag{15}$$

2) $N \times M$ Single-hop Structure: We further apply the above methodology to $N \times M$ single-hop network structure, as summarized in Theorem 3.

Theorem 3. *A queue-based policy* $\mathbf{g}(\mathbf{q}(t)),\ \forall t$ *that satisfies*

$$\begin{cases} \frac{\sum_{k=1}^M g_{ik}(\mathbf{q}(t))}{\sum_{k=1}^M g_{jk}(\mathbf{q}(t))} = \frac{q_{s_i}(t)}{q_{s_j}(t)},\ \forall i,j = 1,\ldots,N \\ \frac{\sum_{k=1}^N g_{ki}(\mathbf{q}(t))}{\sum_{k=1}^N g_{kj}(\mathbf{q}(t))} = \frac{q_{d_i}(t)}{q_{d_j}(t)},\ \forall i,j = 1,\ldots,M \\ \sum_{k=1}^N g_{kj}(\mathbf{q}(t)) \geq \mu_j,\ \forall j = 1,\ldots,M \end{cases} \tag{16}$$

achieves (i) optimal D_{avg} *and* D_{max} *as* (13) *with no initial queue backlog; (ii) asymptotically optimal* D_{avg} *and* D_{max} *as* (13) *with arbitrary initial queue backlog.*

The proof idea follows Proposition 3 and 4. The idea behind (16) extends the queue-proportional idea in $N \times 1$ networks: (i) For each pair of ingress nodes s_i and s_j, the policy should guarantee that the total departure rates from s_i and s_j should follow the ratio between the real-time queue backlog in their buffers. (ii) For each pair of egress nodes d_i and d_j, it should guarantee that the total injection rates to d_i and d_j should follow the ratio between the real-time queue backlog in their buffers. (iii) The policy guarantees maximum throughput.

VI. Performance Evaluation

In this section, we validate our proposed queue-based policies over $N \times 1$ and $N \times M$ single-hop networks, in which we simulate discrete packet transmission under different rate control policies. The goal is to demonstrate that our policies derived based on continuous fluid model achieves superior delay performance under packet transmission in practice.

We evaluate three policies: (i) *Max-link-rate* policy under which all links keep activated to serve packets; (ii) *Backpressure* policy that achieves optimal throughput and low latency [18], which serves packets through a link (s_i, d_j) once its upstream node s_j has longer queue backlog than its downstream node d_j; (iii) Our proposed queue-based policies (15) and (16), termed as *Follow-queue-ratio*, and for $N \times M$ networks, we compare our policy combining both $\boldsymbol{\lambda}$-proportional and $\boldsymbol{\mu}$-proportional patterns with the one only considering $\boldsymbol{\lambda}$-proportional pattern, termed as *Follow-queue-ratio-lambda*, to show the necessity of $\boldsymbol{\mu}$-proportional pattern. We do not present the results over maxweight policy, near-optimal in networks not overloaded [15], as $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ can reach infinity in overloaded networks, where an example is provided in Fig. 1.

To verify our results over $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ under different network settings, we evaluate the performance using (i) different $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, which represents different overload levels; (ii) different $\mathbf{c}$, which represents different service capacity. We consider multiple networks instances with randomly sampled values of the above parameters, and measure the empirical cumulative distribution function (CDF) of $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$.

A. $N \times 1$ Structure

For $N \times 1$ structure, we evaluate on a 32×1 single-hop network. We consider 500 different combinations of parameter settings sampled based on the following rules: (i) The arrival rate to each ingress node is uniformly distributed in $[12, 20]$; (ii) The service rate of the egress node is taken to be a random number within $[0.4, 0.6]$ times of the sum of arrival rates at all ingress nodes, which guarantees that the network is overloaded; (iii) Link capacities are uniformly distributed within $[20, 35]$ to represent the case of sufficient capacity and $[5, 15]$ to represent the case of limited capacity[6]. We round any rational number to integer to characterize discrete packet transmission. We consider initial queue length in each node to be a random integer within $[101, 300]$, and we consider the $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ of packets that arrive within first 200 time units during which the network is overloaded.

Fig. 5 and 6 present the results of sufficient capacity case. We observe that our policy and the max-link-rate policy

[6]The sufficiency of link capacity in the experiments for both $N \times 1$ and $N \times M$ networks is defined as whether the sum of downstream link capacity of an ingress node surpasses the packet arrival rate to it in expectation.

achieve much lower $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ than the backpressure, where their medians of $\bar{D}_{\text{avg}}$ are $(200-185)/200 = 7.5\%$ less than that of backpressure, and the medians of $\bar{D}_{\max}$ are around 50% less. The similarity of the curves between our policy and the max-link-rate policy validates our result for sufficient capacity case as Fig. 4, while the little gap arises from the approximation from the fluid model to packet simulation.

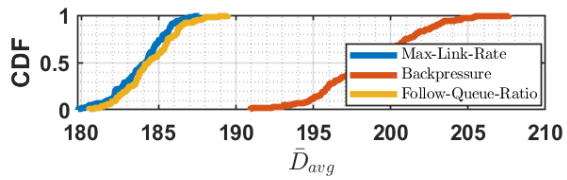

Fig. 5: CDF of $\bar{D}_{\text{avg}}$ in 32×1 network with sufficient capacity

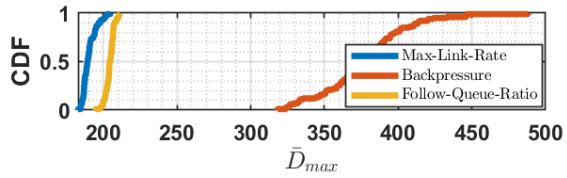

Fig. 6: CDF of $\bar{D}_{\max}$ in 32×1 network with sufficient capacity

Fig. 7 and 8 show the results of limited capacity case, where a major contrast to sufficient capacity case is the big advantage of delay reduction under our proposed policy compared to the max-link-rate policy, which performs poorly. The counter-intuition behind is reflected in Theorem 1 and Fig. 4, where under limited capacity serving with maximum link rates is not delay-optimal.

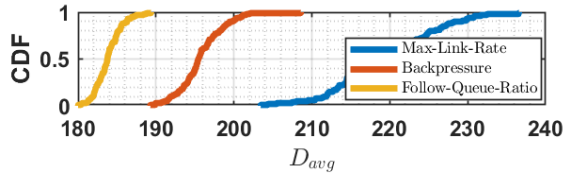

Fig. 7: CDF of $\bar{D}_{\text{avg}}$ in 32×1 network with limited capacity

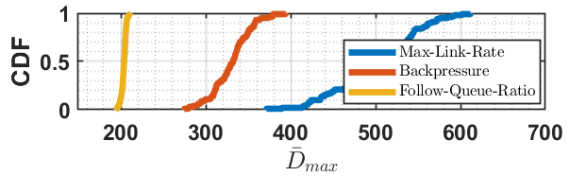

Fig. 8: CDF of $\bar{D}_{\max}$ in 32×1 network with limited capacity

B. $N \times M$ Structure

For $N \times M$ structure, we evaluate on a 32×16 single-hop network. We consider 500 different combinations of parameter settings sampled based on the following rules: (i) The arrival rate to each ingress node is uniformly distributed in $[60, 100]$; (ii) The service rate of each egress node d_j is taken to be the multiplication of a randomly picked normalized weight α_j where $\sum_{j=1}^{M} \alpha_j = 1$ and a random number within $[0.4, 0.6]$ times of the sum of arrival rates at all ingress nodes, which guarantees overload; (iii) The link capacities are randomly picked subject to their sum being equal to 2 times of the sum of arrival rates at all ingress nodes for sufficient capacity case, and 0.8 times for limited capacity case. In this case, we consider the $\bar{D}_{\text{avg}}$ and $\bar{D}_{\max}$ of packets that arrive within first 50 time units, under which the performance gap is already clear among different methods.

We observe that for both sufficient and limited capacity case, the CDF curves are similar. Therefore we only present the results for limited capacity case here in Fig. 9 and 10. We have the following observations: (i) Compared with max-rate-policy and the policy that follows the $\boldsymbol{\lambda}$-proportional pattern only, our proposed policy (16) performs much better in delay reduction, and more stable among all different network settings. This sharp contrast validates the necessity of combination of $\boldsymbol{\lambda}$-proportional and $\boldsymbol{\mu}$-proportional patterns in delay reduction. (ii) Backpressure works well in $N \times M$ networks. The reason is that backpressure try to balance the queue backlogs in ingress and egress nodes, and the effect is amplified compared with $N \times 1$ case as there are $O(M)$-times more links for rate control. (iii) Our proposed follow-queue-ratio policy has greater advantage of balancing the queueing delay of packets from different ingress nodes, as the performance gap of our policy and others is larger for $\bar{D}_{\max}$ than $\bar{D}_{\text{avg}}$.

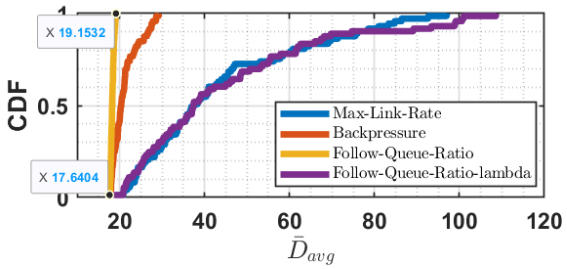

Fig. 9: CDF of $\bar{D}_{\text{avg}}$ in 32×16 network with limited capacity

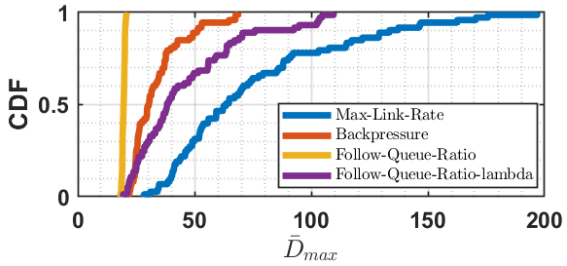

Fig. 10: CDF of $\bar{D}_{\max}$ in 32×16 network with limited capacity

VII. Distributed Policy Implementation

In this section, we propose a distributed implementation of the proposed policies. Consider (16)[7] for a $N \times M$ single-hop network. Solving (16) with variables $\mathbf{g}(\mathbf{q}(t))$

[7](13) is similar in distributed implementation as (16).

gives us a queue-based policy that can be asymptotically delay-optimal agnostic of $(\boldsymbol{\lambda}, \mathbf{c}, \boldsymbol{\mu})$. However, the solving process requires the complete information of $\mathbf{q}(t)$, which needs a centralized controller to collect information and do computation, incurring high communication overhead and computation cost in large-scale networks.

To overcome the scalability problem, we show that finding a solution to (16) can be implemented in a distributed manner. For $N \times 1$ networks, the distributed implementation is straightforward as Algorithm 1. According to (15), the controller at node s_i simply requires current queue backlog and service rate of node s_{i+1}, i.e., $q_{s_{i+1}}(t)$ and $g_{s_{i+1}d}(t)$, to determine its transmission rate $g_{s_i d}$ at $t+1$ as $\frac{q_{s_i}(t)}{q_{s_{i+1}}(t)} g_{s_{i+1}d}(t)$.

Algorithm 1: Distributed Implementation of Delay-Optimal Policy in $N \times 1$ Networks

Input: current queue vector $\mathbf{q}(t)$;
for $i = 1, \ldots, N-1$ **do**
  $g_{s_i d}(t+1) = \frac{q_{s_i}(t)}{q_{s_{i+1}}(t)} g_{s_{i+1}d}(t)$;
$g_{s_N d}(t+1) = \frac{q_{s_N}(t)}{q_{s_1}(t)} g_{s_1 d}(t)$;
Return $\mathbf{g}(t+1)$ as the transmission policy;

For $N \times M$ networks, however, the distributed implementation of (16) is not as straightforward, since the transmission rate $\mathbf{g}(\mathbf{q})$ is involved in multiple equations. Nevertheless, for fully connected $N \times M$ networks, common in datacenter fabrics [19], we can instead target at a subset of delay-optimal $\mathbf{g}$'s which can be elegantly implemented in a distributed way. Consider the set of $\mathbf{g}(\mathbf{q})$'s which satisfy

$$\begin{cases} \frac{\sum_{k=1}^{M} g_{ik}(\mathbf{q}(t))}{\sum_{k=1}^{N} g_{jk}(\mathbf{q}(t))} = \frac{q_{s_i}(t)}{q_{s_j}(t)}, \forall i, j = 1, \ldots, N \\ \frac{g_{1i}(\mathbf{q}(t))}{g_{1j}(\mathbf{q}(t))} = \cdots = \frac{g_{Ni}(\mathbf{q}(t))}{g_{Nj}(\mathbf{q}(t))} = \frac{q_{d_i}(t)}{q_{d_j}(t)}, \forall i, j = 1, \ldots, M \\ \sum_{k=1}^{N} g_{kj}(\mathbf{q}(t)) \geq \mu_j, \forall j = 1, \ldots, M. \end{cases} \tag{17}$$

(17) is a subset of (16), hence the delay optimality of (16) is preserved for any policy that satisfies (17). The new conditions (17) can be implemented in a distributed manner as Algorithm 2. Each ingress node s_i first requests the information of the total departure rate and current queue backlog of s_{i+1} to determines its own total departure rate at the next time step, and then requests real-time queue length from its downstream egress nodes to determine service rates according to the ratio among their queue length. The communication overhead of the distributed controller at each ingress node is $O(M)$, compared with $O(NM)$ using a centralized controller to set the transmission rate at each link. The computation cost is also reduced drastically without solving linear equations in (16).

Algorithm 2: Distributed Implementation of Delay-Optimal Policy in $N \times M$ Networks

Input: current queue vector $\mathbf{q} := \mathbf{q}(t)$;
for $i = 1, \ldots, N$ **do**
  $h_i(t+1) = \frac{q_{s_i}(t)}{q_{s_{i+1}}(t)} \sum_{j=1}^{M} g_{s_i d_j}(t)$;
  $q_{\mathbf{d}}(t+1) = \sum_{j=1}^{M} q_{d_j}(t+1)$;
  for $j = 1, \ldots, M$ **do**
    $g_{s_i d_j}(t+1) = \frac{q_{d_j}(t+1)}{q_{\mathbf{d}}(t+1)} h_i(t+1)$;
Return $\mathbf{g}(t+1)$ as the transmission policy;

VIII. Conclusion and Future Work

We study rate control for queueing delay minimization in overloaded networks. By leveraging the fluid queue model, we prove the delay-optimal rate control policy. We show that the policies that follow the rate-proportional pattern minimize the average delay $\bar{D}_{\text{avg}}$ and the maximum ingress delay $\bar{D}_{\max}$ in single-hop networks, and explain why serving with maximum link rate is not delay-optimal generally. We further extend the rate-proportional result to design a queue-based policy, following the queue-proportional pattern, which can achieve asymptotically minimum delay. We evaluate the performance of our proposed policies under different network settings, and demonstrate its superiority in delay reduction compared with max-link-rate policy and backpressure policy. We finally discuss distributed implementation of our policy in practice. Our ongoing work in future includes extension of the rate-proportional and queue-proportional pattern to some typical multi-hop networks, for example Fat-Tree and Clos structure, where we see some promising preliminary results that their delay-optimality property remains in these structures.

References

[1] Y. Sun, C. E. Koksal, and N. B. Shroff, "On delay-optimal scheduling in queueing systems with replications," *arXiv preprint arXiv:1603.07322*, 2016.

[2] "Broadcom smart-buffer technology in data center switches for cost-effective performance scaling of cloud applications," https://docs.broadcom.com/doc/12358325.

[3] "Achieving data center networking efficiency," https://network.nvidia.com/related-docs/whitepapers/WT-PPR-DC-network-efficiency-WEB.pdf.

[4] G. Kumar, N. Dukkipati, K. Jang, H. Wassel, X. Wu, B. Montazeri, Y. Wang, K. Springborn, C. Alfeld, M. Ryan, D. J. Wetherall, and A. Vahdat, "Swift: Delay is simple and effective for congestion control in the datacenter," 2020. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3387514.3406591

[5] B. Ji, C. Joo, and N. B. Shroff, "Delay-based back-pressure scheduling in multihop wireless networks," *IEEE/ACM Transactions on Networking*, vol. 21, no. 5, pp. 1539–1552, 2012.

[6] M. J. Neely, "Delay-based network utility maximization," *IEEE/ACM Transactions on Networking*, vol. 21, no. 1, pp. 41–54, 2012.

[7] W. Weng and W. Wang, "Achieving zero asymptotic queueing delay for parallel jobs," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 4, no. 3, pp. 1–36, 2020.

[8] C.-p. Li, G. S. Paschos, L. Tassiulas, and E. Modiano, "Dynamic overload balancing in server farms," in *2014 IFIP Networking Conference*. IEEE, 2014, pp. 1–9.

[9] J. David and C. Thomas, "Discriminating flash crowds from ddos attacks using efficient thresholding algorithm," *Journal of Parallel and Distributed Computing*, vol. 152, pp. 79–87, 2021.

[10] Y. Kim, W. C. Lau, M. C. Chuah, and H. J. Chao, "Packetscore: a statistics-based packet filtering scheme against distributed denial-of-service attacks," *IEEE transactions on dependable and secure computing*, vol. 3, no. 2, pp. 141–155, 2006.

[11] "Facebook is back online after a massive outage that also took down instagram, whatsapp, messenger, and oculus," https://www.theverge.com/2021/10/4/22708989/instagram-facebook-outage-messenger-whatsapp-error.

[12] G. Como, K. Savla, D. Acemoglu, M. A. Dahleh, and E. Frazzoli, "Robust distributed routing in dynamical networks—part i: Locally responsive policies and weak resilience," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 317–332, 2012.

[13] D. Shah and D. Wischik, "Fluid models of congestion collapse in overloaded switched networks," *Queueing Systems*, vol. 69, no. 2, p. 121, 2011.

[14] V. Venkataramanan and X. Lin, "On the queue-overflow probability of wireless systems: A new approach combining large deviations with lyapunov functions," *IEEE transactions on information theory*, vol. 59, no. 10, pp. 6367–6392, 2013.

[15] P.-C. Hsieh, I. Hou, X. Liu *et al.*, "Delay-optimal scheduling for queueing systems with switching overhead," *arXiv preprint arXiv:1701.03831*, 2017.

[16] D. Bertsekas and R. Gallager, *Data networks*. Athena Scientific, 2021.

[17] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[18] L. Georgiadis and L. Tassiulas, "Optimal overload response in sensor networks," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2684–2696, 2006.

[19] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano *et al.*, "Jupiter rising: A decade of clos topologies and centralized control in google's datacenter network," *ACM SIGCOMM computer communication review*, vol. 45, no. 4, pp. 183–197, 2015.

[20] J. G. Dai and W. Lin, "Maximum pressure policies in stochastic processing networks," *Operations Research*, vol. 53, no. 2, pp. 197–218, 2005.

[21] M. G. Markakis, E. Modiano, and J. N. Tsitsiklis, "Delay analysis of the max-weight policy under heavy-tailed traffic via fluid approximations," *Mathematics of Operations Research*, vol. 43, no. 2, pp. 460–493, 2018.

APPENDIX

A. Proof of Theorem 2

Proof. We present the proof sketch here. Due to the space limit, we only prove for $\bar{D}_{\text{avg}}$, where $\bar{D}_{\max}$ is similar. For packets arrived at source s_1 at time t and finally transmitted to node d_1, the total queueing delay is

$$\begin{aligned}
D_{s_1d_1}(t) &= \frac{q_{s_1}(t)}{g_{11}+g_{12}} + \frac{q_{d_1}\left(t+\frac{q_{s_1}(t)}{g_{11}+g_{12}}\right)}{\mu_1} \\
&= \frac{q_{s_1}(t)}{g_{11}+g_{12}} + \frac{1}{\mu_1}\max\left\{0, q_{d_1}(t) + \frac{q_{s_1}(t)}{g_{11}+g_{12}}(g_{11}+g_{21}-\mu)\right\} \\
&= \begin{cases} \frac{q_{d_1}(t)}{\mu_1} + \frac{q_{s_1}(t)}{\mu_1}\frac{g_{11}+g_{21}}{g_{11}+g_{12}}, & g_{11}+g_{21} \geq \mu_1 \\ \frac{q_{s_1}(t)}{g_{11}+g_{12}}, & g_{11}+g_{21} < \mu_1 \end{cases}
\end{aligned}$$

Since $\mathbf{q}(t_0) = \mathbf{0}$, then the average delay for packets from s_1 to d_1, denoted as $\bar{D}_{s_1d_1}$, is

$$\begin{aligned}
\bar{D}_{s_1d_1} &:= \frac{1}{T}\int_{t_0}^{t_0+T} D_{s_1d_1}(t)dt \\
&= \begin{cases} \frac{T}{2\mu_1}(g_{11}+g_{21}-\mu_1) \\ \quad + \frac{T}{2\mu_1}\frac{g_{11}+g_{21}}{g_{11}+g_{12}}\max\{\lambda_1-g_{11}-g_{12}, 0\}, & g_{11}+g_{21} \geq \mu_1 \\ \frac{T}{2(g_{11}+g_{21})}\max\left\{\lambda_1-g_{11}-g_{12}, 0\right\}, & g_{11}+g_{21} \leq \mu_1 \end{cases}
\end{aligned}$$

We can verify that among all transmission rate vectors $\mathbf{g}$'s that $g_{11}+g_{21} \leq \mu_1$, the $\mathbf{g}$'s that satisfy $g_{11}+g_{21} = \mu_1$ achieve minimum delay[8]. Therefore the minimum delay achieved under $g_{11}+g_{21} \geq \mu_1$ is exactly the global optimum, under which

$$\bar{D}_{s_1d_1} = \begin{cases} \frac{T}{2\mu_1}\left(\lambda_1\frac{g_{11}+g_{21}}{g_{11}+g_{12}} - \mu_1\right), & g_{11}+g_{12} \leq \lambda_1 \\ \frac{T}{2\mu_1}\left(g_{11}+g_{21}-\mu_1\right), & g_{11}+g_{12} \geq \lambda_1 \end{cases}$$

[8]The intuition is clear that $g_{11}+g_{21} < \mu_1$ does not fully utilize the service capability of node d_1.

Generally, we can obtain that for any $(i,j) \in \{(1,1),(1,2),(2,1),(2,2)\}$,

$$\bar{D}_{s_id_j} = \begin{cases} \frac{T}{2\mu_j}\left(\lambda_i\frac{g_{1j}+g_{2j}}{g_{i1}+g_{i2}} - \mu_j\right), & g_{i1}+g_{i2} \leq \lambda_i \\ \frac{T}{2\mu_j}\left(g_{1j}+g_{2j}-\mu_2\right), & g_{i1}+g_{i2} \geq \lambda_i \end{cases}$$

Therefore, we have

$$\begin{aligned}
\bar{D}_{\text{avg}} = &\frac{\lambda_1}{\lambda_1+\lambda_2}\frac{g_{11}}{g_{11}+g_{12}}\bar{D}_{s_1d_1} + \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{g_{12}}{g_{11}+g_{12}}\bar{D}_{s_1d_2} \\
&+ \frac{\lambda_2}{\lambda_1+\lambda_2}\frac{g_{21}}{g_{21}+g_{22}}\bar{D}_{s_2d_1} + \frac{\lambda_2}{\lambda_1+\lambda_2}\frac{g_{22}}{g_{21}+g_{22}}\bar{D}_{s_2d_2}
\end{aligned}$$

where $\frac{\lambda_i}{\lambda_1+\lambda_2}\frac{g_{ij}}{g_{i1}+g_{i2}}$ denotes the portion of packets from s_i to d_j arrived within $[t_0, t_0+T]$. We again consider the four regions $\{\mathbf{g} \mid g_{11}+g_{12} \lesseqgtr \lambda_1, g_{21}+g_{22} \lesseqgtr \lambda_2\}$ and prove that the optimal solutions constrained in each region are all global optimum. Due to space limit we only show the details of the case $g_{11}+g_{12} \leq \lambda_1,\ g_{21}+g_{22} \leq \lambda_2$. In this case, the average delay of packets arrived in $[t_0, t_0+T]$ among all ingress nodes is

$$\begin{aligned}
\bar{D}_{\text{avg}} &\sim \frac{g_{11}+g_{21}}{\mu_1}\left(\frac{\lambda_1^2}{g_{11}}\left(\frac{g_{11}}{g_{11}+g_{12}}\right)^2 + \frac{\lambda_2^2}{g_{21}}\left(\frac{g_{21}}{g_{21}+g_{22}}\right)^2\right) \\
&\quad + \frac{g_{12}+g_{22}}{\mu_2}\left(\frac{\lambda_1^2}{g_{12}}\left(\frac{g_{12}}{g_{11}+g_{12}}\right)^2 + \frac{\lambda_2^2}{g_{22}}\left(\frac{g_{22}}{g_{21}+g_{22}}\right)^2\right) \\
&= \frac{1}{\mu_1}\left(\lambda_1^2x^2 + \lambda_2^2y^2 + \lambda_1^2x^2\frac{g_{21}}{g_{11}} + \lambda_2^2y^2\frac{g_{11}}{g_{21}}\right) \\
&\quad + \frac{1}{\mu_2}\left(\lambda_1^2(1-x)^2 + \lambda_2^2(1-y)^2 + \lambda_1^2(1-x)^2\frac{g_{22}}{g_{12}} + \lambda_2^2(1-y)^2\frac{g_{12}}{g_{22}}\right) \\
&\overset{\text{(i)}}{\geq} \frac{1}{\mu_1}(\lambda_1x+\lambda_2y)^2 + \frac{1}{\mu_2}(\lambda_1(1-x)+\lambda_2(1-y))^2 \\
&\overset{\text{(ii)}}{\geq} \frac{T}{2(\mu_1+\mu_2)}(\lambda_1+\lambda_2) - \frac{T}{2} = \frac{T}{2(\mu_1+\mu_2)}(\lambda_1+\lambda_2-\mu_1-\mu_2)
\end{aligned}$$

where $\sim$ means removing constant terms, $x := \frac{g_{11}}{g_{11}+g_{12}}$, and $y := \frac{g_{21}}{g_{21}+g_{22}}$. The inequality (i) stems from Cauchy-Schwartz Inequality, which turns into equality when $\frac{g_{11}}{g_{21}} = \frac{\lambda_1 x}{\lambda_2 y}$, $\frac{g_{12}}{g_{22}} = \frac{\lambda_1(1-x)}{\lambda_2(1-y)}$ and equivalently,

$$\frac{g_{11}+g_{12}}{g_{21}+g_{22}} = \frac{\lambda_1}{\lambda_2}. \tag{18}$$

The inequality (ii) holds due to solving

$$\min_{x,y\in[0,1]} \quad \frac{1}{\mu_1}(\lambda_1x+\lambda_2y)^2 + \frac{1}{\mu_2}(\lambda_1(1-x)+\lambda_2(1-y))^2$$

where the optimal (x,y) satisfies

$$\begin{cases} \lambda_1x+\lambda_2y = \frac{\lambda_1+\lambda_2}{\mu_2}\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}\right)^{-1} \\ \lambda_1(1-x)+\lambda_2(1-y) = \frac{\lambda_1+\lambda_2}{\mu_1}\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}\right)^{-1} \end{cases}$$

which, combined with (18), is equivalent to

$$\frac{\lambda_2}{\lambda_1+\lambda_2}\frac{g_{11}+g_{21}}{g_{21}+g_{22}} = \frac{\mu_1}{\mu_1+\mu_2}, \quad \frac{\lambda_2}{\lambda_1+\lambda_2}\frac{g_{12}+g_{22}}{g_{21}+g_{22}} = \frac{\mu_2}{\mu_1+\mu_2}$$

and thus

$$\frac{g_{11}+g_{21}}{g_{12}+g_{22}} = \frac{\mu_1}{\mu_2} \tag{19}$$

suffices to make the inequality (ii) achieve its lower bound. Therefore (18) and (19) with $g_{11}+g_{21} \geq \mu_1,\ g_{12}+g_{22} \geq \mu_2$ give us sufficient and necessary condition for $\mathbf{g}$ to minimize $\bar{D}_{\text{avg}}$ under $g_{11}+g_{12} \leq \lambda_1$ and $g_{21}+g_{22} \leq \lambda_2$. □