I. The Proof of Throughput Optimality of QueueFlower

In this section, we will demonstrate the throughput optimality of the main theorem. We justify QueueFlower's theoretical performance and prove that it is throughput optimal. In other words, QueueFlower can support the maximum flow arrival rate such that every flow has a finite latency. The queue lengths under QueueFlower evolve as Markov chains, and we study their dynamics via Lyapunov drift analysis. First, we define the maximum throughput region in the system.

***Definition** 1:* Let $\mathcal{X} := \{x : \sum_{n\in\mathcal{N}} x_n = X\}$ be the set of all possible resource allocation and the service rate function be $d_n(x_n)$. The throughput region is defined as the set of arrival rates such that

$$\Lambda := \{\lambda^f, f \in \mathcal{F} \mid \sum_{f\in\mathcal{F}_n} \lambda^f \leq d_n(x_n),\ x \in \mathcal{X}\}.$$

Intuitively, the throughput region Λ characterizes the maximum flow arrival rates that can be supportable in the system, i.e., "flows arrival rates $\leq$ service rates". We can present our main results of QueueFlower in the following Theorem.

***Theorem** 1:* Assume the arrival rates are within the maximum throughput region Λ, there exists queue weights (w_1, w_2, w_3) such that the total queue length $\mathbb{E}[\sum_{f\in\mathcal{F}}\sum_{n\in\mathcal{R}^f} q_n^f(t)]$ is bounded under QueueFlower.

Theorem 1 demonstrates that QueueFlower guarantees queue stability (the first-order property) in the microservice system, thereby achieving optimal throughput [1]. Note we need proper weights such that the virtual queues $\{\hat{q}_n(t)\}_n$ are good estimators of the real queues $\{q_n(t)\}_n$ such that QueueFlower does the right queue balancing. Though these weights are hyperparameters, our experiments found them insensitive as the metrics are proportional to the real queues in the system.

To prove Theorem 1, we leverage the Lyapunov drift method, which is a classical technique for bounding a stochastic process [2]. We first introduce the core concepts in our analysis, called *section*, which groups services with nearly identical queue lengths and are part of the same flow, and *section* weight, which is a measurement of the unserved requests of the *section*.

***Definition** 2:* We let $q^f_{\max}(t) \triangleq \max_{i\in f} q_i^f(t)$ where $q_i^f(t)$ is the queue length of node i belongs to flow f at time t. A *section* $S^f(t)$ of flow f at time t is a set of consecutive nodes that can be represented as

$$\left(n_1^{(f,s1)}, n_2^{(f,s1)}, ..., n_k^{(f,s1)}, n_1^{(f,s2)}, n_2^{(f,s2)}, ..., n_s^{(f,s2)}\right)$$

where $q^f_{n_j^{(f,s2)}}(t) = q^f_{\max}(t), \forall j \in [1,s]$, $(1-\delta)q^f_{\max}(t) < q^f_{n_j^{(f,s1)}}(t) < q^f_{\max}(t)$, $\forall j \in [1,k]$ given $\delta \in (0,1)$, with s and k representing the number of services of these two types, respectively. The set of *sections* for flow f at period t is denoted as $\mathcal{S}^f(t)$.

***Definition** 3:* The weight of a *section* at time t is defined as

$$\omega_{S^f}(q^f(t)) \triangleq \frac{\sum_{n\in S^f(t)} q_n^f(t)}{|S^f(t)|},$$

where $|S^f(t)|$ represents the number of nodes in *section* $S^f(t)$.

We define the Lyapunov function based on the concept of *section*, which quantifies the queue length within the system at the *section* level:

$$V(q^f(t)) \triangleq \max_{f\in\mathcal{F}} \frac{1}{\lambda^f(t)} \max_{S^f(t)\in\mathcal{S}^f(t)} \omega_{S^f(t)}(q^f(t)),$$

where $\omega_{S^f(t)}(q^f(t)) \triangleq \frac{\sum_{n\in S^f(t)} q_n^f(t)}{|S^f(t)|}$ is the weight of $S^f(t)$. Intuitively, the drift function represents the maximum average response time across all flows within the *sections*. Under the assumption that there exist weights such that $\hat{q}_n(t) = q_n(t)$, we can generally follow the techniques [1] to establish a negative Lyapunov drift under QueueFlower. Therefore, Theorem 1 is proved by the Forster-Lyapunov theorem [2].

We assume that finely tuned queue-latency weights will aid in approximating the actual queue behavior, and we will analyze based on real queue vectors in the subsequent proof. We first introduce the following definitions of *section* and *section* weight.

Then we can begin the proof of throughput optimality. According to the definition of maximum throughput region, we have for any arrival rate $\lambda^f(t)$, there always exists a real number $\epsilon \in (0, 1/2)$ such that

$$(1+\epsilon)(\sum_{f:f\in\mathcal{F}} \lambda^f(t)) \in \Lambda,$$

where we let $\delta = \epsilon/2 \in (0, 1/4)$. Recall the drift function

$$V(q^f(t)) \triangleq \max_{f\in\mathcal{F}} \frac{1}{\lambda^f(t)} \max_{S^f(t)\in\mathcal{S}^f(t)} \omega_{S^f(t)}(q^f(t)), \quad (1)$$

we will then prove $\frac{D^+}{dt^+}V(q^f(t)) \leq 0$ to show the throughput optimality and system stability.

We first present a lemma from [3].

***Lemma** 1:* If $f(x) = \max_{i\in[M]} f_i(x)$ and $f_i(x)$, $\forall i$, are locally Lipschitz continuous, then we have

$$\frac{D^+}{dx^+} f(x) \leq \max_{i\in\mathcal{K}} \left\{ \frac{D^+}{dx^+} f_i(x) \right\},$$

where $\mathcal{K} \triangleq \{i | f_i(x) = f(x)\}$.

From Lemma 1, we can provide an upper bound for $\frac{D^+}{dt^+}V(q^f(t))$ that

$$\frac{D^+}{dt^+}V(q^f(t)) \leq \max_{\overline{f}\in\overline{\mathcal{K}}} \frac{1}{\lambda^{\overline{f}}(t)} \frac{D^+}{dt^+} \max_{S^{\overline{f}}(t)\in\mathcal{S}^{\overline{f}(t)}} \omega_{S^{\overline{f}}(t)}(q^f(t))$$

where

$$\overline{\mathcal{K}}(t) \triangleq \left\{ \overline{f} \in \mathcal{F} : V(q^f(t)) = \frac{1}{\lambda^{\overline{f}}(t)} \max_{S^{\overline{f}}(t)\in\mathcal{S}^{\overline{f}}(t)} \omega_{S^{\overline{f}}(t)}(q^f(t)) \right\}.$$

We let μ represent the whole capacity of the system within given pods. Furthermore, we denote the capacity of node i

as μ_i and the capacity of node i belonging to flow f as μ_i^f. Subsequently, we introduce the following Lemma, which plays a crucial role in establishing queue stability.

Lemma 2: Assume $\theta(\sum_{f:f\in\mathcal{F}}\lambda^f)\in\Lambda$ for some $\theta>0$. If $q_i^f\neq 0$ and $(i^*,f^*)\in\arg\max_{(i,f)} q_i^f/\lambda^f$, then $\mu_{i^*}^{f^*}\geq\theta\lambda^{f^*}$ under QueueFlower.

This lemma demonstrates that QueueFlower ensures a minimum capacity for service i^* to minimize the maximum value of q_i^f/λ^f, thereby maintaining system stability. Combine this lemma and consider the case $V(q^f(t))>0$, we can get

$$\frac{q_{i^*}^{f^*}(t)}{\lambda^{f^*}(t)}\geq\frac{q_{\max}^{\overline{f}}(t)}{\lambda^{\overline{f}}(t)}\overset{(a)}{\geq}\frac{1}{\lambda^{\overline{f}}(t)}\max_{S^{\overline{f}}(t)\in\mathcal{S}^{\overline{f}}(t)}\omega_{S^{\overline{f}}(t)}(q^f(t))>0,$$

where (a) holds from the definition of *section* weight.

Then we analyze the derivative of $\max_{S^{\overline{f}}(t)\in\mathcal{S}^{\overline{f}}(t)}\omega_{S^{\overline{f}}(t)}(q^f(t))$ to bound the derivative of drift function. Given that $q_i^{\overline{f}}(t)$ is continuous and $q_{\max}^{\overline{f}}(t)>0$, it follows that $q_i^{\overline{f}}(t)$ exhibits the property of preserving signs for any sufficiently small $u>0$. Consequently, we identify two key facts:

(i) The service in $S^{\overline{f}}$ with the maximum queue length at period $t+u$, for a sufficiently small u, also retains the maximum queue length at period t.

(ii) Any service i that satisfies $(1-\delta)q_{\max}^{\overline{f}}(t)<q_i^{\overline{f}}(t)<q_{\max}^{\overline{f}}(t)$ still satisfies $(1-\delta)q_{\max}^{\overline{f}}(t+u)<q_i^{\overline{f}}(t+u)<q_{\max}^{\overline{f}}(t+u)$.

Since $\left(n_1^{(\overline{f},s1)},\ldots,n_J^{(\overline{f},s1)},n_1^{(\overline{f},s2)},\ldots,n_K^{(\overline{f},s2)}\right)$ is a *section* with the maximum average queue length at period t, according to the above facts, we can assume $\left(n_1^{(\overline{f},s0)},\ldots,n_I^{(\overline{f},s0)}\right)$ be the set of service that with the queue lengths equal to $(1-\delta)q_{\max}^{f}(t+u)$ and are included in a section at period $t+u$. Then we have a *section* with the maximum queue length at period $t+u$:

$$\left(n_1^{(\overline{f},s0)},\ldots,n_I^{(\overline{f},s0)},n_1^{(\overline{f},s1)},\ldots,n_J^{(\overline{f},s1)},n_1^{(\overline{f},s2)},\ldots,n_K^{(\overline{f},s2)}\right).$$

Then by fact (ii), $\forall i,j,k$, we have

$$q^{\overline{f}}_{n_i^{(\overline{f},s0)}}(t+u)<\min\{q^{\overline{f}}_{n_j^{(\overline{f},s1)}}(t+u),q^{\overline{f}}_{n_k^{(\overline{f},s2)}}(t+u)\},$$

which gives

$$\begin{aligned}&\max_{S^{\overline{f}}(t+u)\in\mathcal{S}^{\overline{f}}(t+u)}\omega_{S^{\overline{f}}(t+u)}(q^f(t+u))\\ \leq&\frac{1}{J+K}\left(\sum_{j=1}^{J}q^{\overline{f}}_{n_j^{(\overline{f},s1)}}(t+u)+\sum_{j=1}^{K}q^{\overline{f}}_{n_j^{(\overline{f},s2)}}(t+u)\right).\end{aligned}\tag{2}$$

In addition, since $q^{\overline{f}}_{n_j^{(\overline{f},s1)}}(t)<q^{\overline{f}}_{n_k^{(\overline{f},s2)}}(t),\forall j,k$, we have

$$\begin{aligned}&\max_{S^{\overline{f}}(t)\in\mathcal{S}^{\overline{f}}(t)}\omega_{S^{\overline{f}}(t)}(q^f(t))\\ \geq&\frac{1}{J+K}\left(\sum_{j=1}^{J}q^{\overline{f}}_{n_j^{(\overline{f},s1)}}(t)+\sum_{j=1}^{K}q^{\overline{f}}_{n_j^{(\overline{f},s2)}}(t)\right).\end{aligned}\tag{3}$$

Combine (2) and (3), we have

$$\begin{aligned}&\frac{D^+}{dt^+}\max_{S^{\overline{f}}(t)\in\mathcal{S}^{\overline{f}}(t)}\omega_{S^{\overline{f}}(t)}(q^f(t))\\ =&\limsup_{u\to 0}\frac{1}{u}(\max_{S^{\overline{f}}(t+u)\in\mathcal{S}^{\overline{f}}(t+u)}\omega_{S^{\overline{f}}(t+u)}(q^f(t+u))\\ &-\max_{S^{\overline{f}}(t)\in\mathcal{S}^{\overline{f}}(t)}\omega_{S^{\overline{f}}(t)}(q^f(t)))\\ \leq&\limsup_{u\to 0}\frac{1}{J+K}(\sum_{j=1}^{J}\frac{1}{u}(q^{\overline{f}}_{n_j^{(\overline{f},s1)}}(t+u)-q^{\overline{f}}_{n_j^{(\overline{f},s1)}}(t))\\ &+\sum_{j=1}^{K}\frac{1}{u}(q^{\overline{f}}_{n_j^{(\overline{f},s2)}}(t+u)-q^{\overline{f}}_{n_j^{(\overline{f},s2)}}(t)))\\ =&\frac{1}{J+K}(\sum_{j=1}^{J}\frac{d}{dt}q^{\overline{f}}_{n_j^{(\overline{f},s1)}}(t)+\sum_{j=1}^{K}\frac{d}{dt}q^{\overline{f}}_{n_j^{(\overline{f},s2)}}(t)).\end{aligned}$$

In the rest of the proof, we omit the time index t for simplicity. Given that the arrival rate of a *section* matches the processing rate of its preceding service, and corresponds to the flow's arrival rate when the initial service of this *section* is the `frontend` service, we consider the following two cases:

(i) The first service $n_1^{(\overline{f},s1)}$ in the *section* isn't the `frontend` service.
We let n_{j_-} denote the previous service of j-th service, then $n_{1_-}^{(\overline{f},s1)}$ represents the previous service of *section* $S^{\overline{f}}$. Then we have

$$\begin{aligned}&\frac{1}{\lambda^{\overline{f}}}\frac{D^+}{dt^+}\max_{S^{\overline{f}}\in\mathcal{S}^{\overline{f}}}\omega_{S^{\overline{f}}}(q^f)\\ \leq&\frac{1}{\lambda^{\overline{f}}(J+K)}(\sum_{j=1}^{J}(\mu^{\overline{f}}_{n_{j_-}^{(\overline{f},s1)}}-\mu^{\overline{f}}_{n_j^{(\overline{f},s1)}})\\ &+\sum_{j=1}^{K}(\mu^{\overline{f}}_{n_{j_-}^{(\overline{f},s2)}}-\mu^{\overline{f}}_{n_j^{(\overline{f},s2)}}))\\ =&\frac{1}{\lambda^{\overline{f}}(J+K)}\left(\mu^{\overline{f}}_{n_{1_-}^{(\overline{f},s1)}}-\mu^{\overline{f}}_{n_s^{(\overline{f},s2)}}\right).\end{aligned}\tag{4}$$

If $q^{\overline{f}}_{n_{1_-}^{(\overline{f},k)}}=0$, then $\mu^{\overline{f}}_{n_{1_-}^{(\overline{f},k)}}=0$ and we have

$$\frac{1}{\lambda^{\overline{f}}}\frac{D^+}{dt^+}\max_{S^{\overline{f}}\in\mathcal{S}^{\overline{f}}}\omega_{S^{\overline{f}}}(q^f)\leq-\frac{\mu^{\overline{f}}_{n_s^{(\overline{f},s2)}}}{\lambda^{\overline{f}}(J+K)}.\tag{5}$$

From Lemma 2, we have

$$\begin{aligned}\frac{q_{\max}^{\overline{f}}}{\lambda^{\overline{f}}}&\geq\frac{1}{\lambda^{\overline{f}}}\max_{S^{\overline{f}}\in\mathcal{S}^{\overline{f}}}\omega_{S^{\overline{f}}}(q^f)\\ &\geq\frac{1}{\lambda^{f^*}}\max_{S^{f^*}\in\mathcal{S}^{f^*}}\omega_{S^{f^*}}(q^f)\\ &\geq\frac{(1-\delta)q_{i^*}^{f^*}}{\lambda^{f^*}}.\end{aligned}\tag{6}$$

Then we can obtain

$$\begin{aligned}\mu^{\overline{f}}_{n_s^{(\overline{f},s2)}} &\overset{(a)}{=} \frac{q^{\overline{f}}_{n_s^{(\overline{f},s2)}}}{q^{f^*}_{i^*}}\mu^{f^*}_{i^*}\\ &= \frac{q^{\overline{f}}_{\max}/\lambda^{\overline{f}}}{q^{f^*}_{i^*}/\lambda^{f^*}}\frac{\lambda^{\overline{f}}}{\lambda^{f^*}}\mu^{f^*}_{i^*}\\ &\overset{(b)}{\geq}(1-\delta)(1+\epsilon)\lambda^{\overline{f}}\\ &\overset{(c)}{\geq}(1+\frac{\epsilon}{4})\lambda^{\overline{f}}, \end{aligned} \tag{7}$$

where (a) follows the resource estimation mechanism of QueueFlower, (b) utilizes equation (6) and Lemma 2, (c) is true since $\delta = \epsilon/2$ and $\epsilon \in (0, 1/2)$.

By substituting inequality (7) into (5), we have

$$\frac{1}{\lambda^{\overline{f}}}\frac{D^+}{dt^+}\max_{S^{\overline{f}}\in\mathcal{S}^{\overline{f}}}\omega_{S^{\overline{f}}}(q^f) \leq -\frac{1+\epsilon/4}{J+K} \leq -\frac{1+\epsilon/4}{N},$$

where the last step follows from the fact that the number of nodes belonging to a section is less than $|N|$, i.e., $J + K \leq |N|$.

If $q^{\overline{f}}_{n_{1_-}^{(\overline{f},s1)}} > 0$, inequality (4) can be written as

$$\begin{aligned}&\frac{1}{\lambda^{\overline{f}}}\frac{D^+}{dt^+}\max_{S^{\overline{f}}\in\mathcal{S}^{\overline{f}}}\omega_{S^{\overline{f}}}(q^f)\\ \leq&\frac{1}{\lambda^{\overline{f}}(J+K)}\left(\frac{q^{\overline{f}}_{n_{1_-}^{(\overline{f},s1)}}}{q_{n_{1-}}}\mu_{n_{1-}} - \frac{q^{\overline{f}}_{n_s^{(\overline{f},s2)}}}{q_{n_s}}\mu_{n_s}\right)\\ =&\frac{1}{\lambda^{\overline{f}}(J+K)}(q^{\overline{f}}_{n_{1_-}^{(\overline{f},s1)}} - q^{\overline{f}}_{n_s^{(\overline{f},s2)}})\frac{\mu_{n_s}}{q_{n_s}}\\ \leq&\frac{-\delta q^{\overline{f}}_{\max}}{\lambda^{\overline{f}}(J+K)}\frac{\mu}{Q}, \end{aligned} \tag{8}$$

where $\mu_{n_s} = \sum_{f:n_s\in f}\mu^f_{n_s}$, $q_{n_s} = \sum_{f:n_s\in f} q^f_{n_s}$, $\mu = \sum_{i\in\mathcal{N}}\mu_i$, $Q = \sum_{i\in\mathcal{N}} q_i$. The equality holds since

$$\frac{\mu_{n_{1_-}}}{q_{n_{1_-}}} = \frac{\mu_{n_s}}{q_{n_s}},$$

and the last inequality holds since

$$\frac{x_i}{y_i} = \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} \text{ if } \frac{x_1}{y_1} = \frac{x_2}{y_2} = \cdots = \frac{x_n}{y_n}.$$

Since (i^*, f^*) maximizes q^f_i/λ^f, we have $q^{f^*}_{i^*}/\lambda^{f^*} \geq q^f_i/\lambda^f$. Thus, we have

$$q_i = \sum_{f:i\in f} q^f_i \leq \frac{q^{f^*}_{i^*}}{\lambda^{f^*}}\sum_{f:i\in f}\lambda^f \leq \mu_i\frac{q^{f^*}_{i^*}}{\lambda^{f^*}}, \tag{9}$$

where the last step follows from the fact that the sum of the capacity of flows that path through node i can't exceed the total capacity of i. Sum the above inequality over N services, we have

$$Q = \sum_{i\in\mathcal{N}} q_i \leq N\frac{q^{f^*}_{i^*}}{\lambda^{f^*}}. \tag{10}$$

Combining (6) and (10), we have

$$\frac{q^{\overline{f}}_{\max}}{\lambda^{\overline{f}}(J+K)}\frac{\mu}{Q} \geq \frac{(1-\delta)\mu}{N(J+K)} \geq \frac{(1-\delta)\mu}{N^2}.$$

Substitute this inequality into (8) and recall the fact that $\delta \in (0, 1/4)$, we have

$$\frac{1}{\lambda^{\overline{f}}}\frac{D^+}{dt^+}\max_{S^{\overline{f}}\in\mathcal{S}^{\overline{f}}}\omega_{S^{\overline{f}}}(q^f) \leq -\frac{\delta(1-\delta)\mu}{N^2},$$

which proves $\frac{D^+}{dt^+}V(q^f(t)) \leq 0$.

(ii) The first service $n_1^{(\overline{f},s1)}$ in the *section* is the `frontend` service.

From (4), we can derive

$$\begin{aligned}\frac{1}{\lambda^{\overline{f}}}\frac{D^+}{dt^+}\max_{S^{\overline{f}}\in\mathcal{S}^{\overline{f}}}\omega_{S^{\overline{f}}}(q^f) \leq&\frac{1}{\lambda^{\overline{f}}(J+K)}(\lambda^{\overline{f}} - \mu^{\overline{f}}_{n_s^{(\overline{f},s2)}})\\ \overset{(a)}{\leq}& -\frac{\epsilon}{4(J+K)}\\ \overset{(b)}{\leq}& -\frac{\epsilon}{4N}, \end{aligned} \tag{11}$$

where step (a) follows inequality (7), step (b) follows the fact that $J + K \leq |N|$.

Therefore, we prove that $\frac{D^+}{dt^+}V(q^f(t)) \leq 0$ always holds, which proves the throughput optimality and system stability of QueueFlower.

Proof of Lemma 1.

We prove Lemma 1 by contradiction. We assume

$$\frac{D^+}{dt^+}f(x) > \max_{i\in\mathcal{K}}\left\{\frac{D^+}{dt^+}f_i(x)\right\}. \tag{12}$$

For a sufficient small ρ, there exists a decreasing sequence $\{u_k, k = 1, 2, ...\}$ with $\lim_{k\to\infty} u_k = 0$ such that

$$\frac{f(x+u_k)-f(x)}{u_k} \geq \max_{i\in\mathcal{K}}\left\{\frac{D^+}{dt^+}f_i(x)\right\} + \rho, \forall k = 1, 2, \cdots.$$

Note that $f(x) = f_i(x), \forall i \in \mathcal{K}$. Since there are a finite number of local Lipschitz continuous functions $f_i(x), i = 1, 2, \cdots, K$, there must exist a $j \in \mathcal{K}$ and a decreasing subsequence $\{u_{t_k}, k = 1, 2, \cdots\}$ of $\{u_k, k = 1, 2, \cdots\}$ such that $f_j(x + u_{t_k}) = f(x + u_{t_k}) = \max_{i=1,2,\cdots,K} f_i(x + u_{t_k}), \forall k = 1, 2, \cdots$, which implies that

$$\frac{f_j(x+u_{t_k})-f_j(x)}{u_{t_k}} \geq \max_{i\in\mathcal{K}}\left\{\frac{D^+}{dt^+}f_i(x)\right\} + \rho, \forall k = 1, 2, \cdots.$$

Therefore, we have the contradiction

$$\frac{D^+}{dt^+}f_j(x) \geq \max_{i\in\mathcal{K}}\left\{\frac{D^+}{dt^+}f_i(x)\right\} + \rho. \tag{13}$$

So we have the desired result.

Proof of Lemma 2.

We have $q^{f^*}_{i^*} > 0$. Assume $\mu^{f^*}_{i^*} < \theta\lambda^{f^*}$. Then for any flow within the node $i (f \neq f^*)$, there are two cases:

(i) If $q^f_i = 0$, then $\mu_i = 0$.

(ii) If $q_i^f > 0$, then we have

$$\mu_i^f \overset{(a)}{=} \frac{q_i^f}{q_{i^*}^{f^*}} \mu_{i^*}^{f^*} = \frac{q_i^f/\lambda^f}{q_{i^*}^{f^*}/\lambda^{f^*}} \frac{\lambda^f}{\lambda^{f^*}} \mu_{i^*}^{f^*} \overset{(b)}{\leq} \lambda^f \frac{q_{i^*}^{f^*}}{\lambda^{f^*}} \overset{(c)}{<} \theta\lambda^f,$$

where step (a) follows the resource allocation mechanism of QueueFlower, (b) is true since $(i^*, f^*) \in \arg\max_{(i,f)} \frac{q_i^f}{\lambda^f}$, (c) follows from our assumption.

Combining (i) and (ii), we have $\mu_i^f < \theta\lambda^f$ in any case. Hence we have $\mu < \theta \sum_{\lambda^f \in \Lambda} \lambda^f$, which contradicts the fact that $\sum_{f \in \mathcal{F}_n} \lambda^f \leq d_n(x_n)$, i.e., the service rate vector strictly lies on the boundary of the capacity region Λ. Therefore, we have the desired result.

REFERENCES

[1] B. Li and R. Srikant. Queue-proportional rate allocation with per-link information in multihop wireless networks. *Queueing Systems*, 2016.

[2] R. Srikant and L. Ying. *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*. 2014.

[3] J. M. Borwein and A. S. Lewis. Convex analysis and nonlinear optimization. 2000.