WWTA LOAD-BALANCING FOR PARALLEL-SERVER SYSTEMS WITH HETEROGENEOUS SERVERS AND MULTI-SCALE HEAVY TRAFFIC LIMITS FOR GENERALIZED JACKSON NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yaosheng Xu

August 2022

WWTA LOAD-BALANCING FOR PARALLEL-SERVER SYSTEMS WITH HETEROGENEOUS SERVERS AND MULTI-SCALE HEAVY TRAFFIC LIMITS FOR GENERALIZED JACKSON NETWORKS

Yaosheng Xu, Ph.D.

Cornell University 2022

This thesis explores the BAR approach applied on various of stochastic processing networks. The weighted-workload-task-allocation (WWTA) load-balancing policy is known to be throughput optimal for parallel-server systems. The first part of the thesis concerns the steady-state performance approximation of WWTA policy in heavy traffic. Instead of proving a stochastic process limit followed by a limit interchange – a method that dominates the literature, our method works directly with pre-limit BAR that characterizes the stationary distribution of each pre-limit system. Under a complete-resource-pooling condition, we prove that WWTA achieves a "strong form" of state-space collapse in heavy traffic and that each scaled workload converges in distribution to an exponential random variable, whose parameter is explicitly given by system primitives. Various steady-state performance measures are shown to be approximated from this exponential random variable.

In the second part, we prove that under a multi-scale heavy traffic condition, the stationary distribution of the multi-scaled queue length process in any generalized Jackson network has a product-form limit. Each component in the product-form has an exponential distribution, corresponding to the Brownian approximation of a single station queue. The "single station" can be constructed precisely and has a good intuitive interpretation.

BIOGRAPHICAL SKETCH

Yaosheng Xu grew up in Shandong Province, China. After graduating from Weihai No.2 High School, she finished her undergraduate study in Shandong University majoring in Mathematics. There she continued her graduate study as a master student majoring in Probability and Mathematical Statistics. In 2017, she started her Ph.D in Statistics in the Department of Statistics and Data Science at Cornell University. There she is advised by Professor Jim Dai, working on the stochastic processing network.

I dedicate this thesis to my parents and my husband.

ACKNOWLEDGEMENTS

Five-year journey has been unique experience of study that I had ever gone through. It is full of uncertainty but full of amazing exploration.

I first would like to thank my advisor, Professor Jim Dai, for inducing me to the advanced methodology on exploring the queue theory. I am very grateful for his consistent support and guidance along with my study. I am deeply impressed by his passion for pursuing research, which inspires me a lot on the creative research insights and persistency on performing hard work. Those attitudes enlighten me on scientific research, also benefit me on various of aspects throughout my life.

I further would like to thank my committee members, Professor Martin T. Wells and Professor Yang Ning. Their responsive feedback and insightful advice support me along the way. I also would like to thank my ad hoc committee member, Professor Peter W. Glynn, from Stanford University. I am grateful to work with him on such awesome project idea with his insightful guidance.

I would also like to thank my family who are always there to back me up. My parents understand my thoughts and decisions, and always stand with me. My grandparents, aunts and uncles raised me up with my parents when I was young, and they always support me in many ways. I gain so much love that no matter what I encountered when I am away from home, they are the warmest harbour for the deep inside of my heart.

I would like to especially thank my husband, Xin Bing, for supporting me through my entire Ph.D study. I may not able to be what I am today without his sympathy and companion.

I would like to thank my friends, Siyi Deng, Huijie Feng, Yuxuan Zhao, Xiaoyi Zhu, and all the other friends from Cornell, for being there with me.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

WWTA LOAD-BALANCING FOR PARALLEL-SERVER SYSTEMS WITH HETEROGENEOUS SERVERS

1.1 Introduction

Parallel-server systems are a special class of stochastic processing networks. For each class, jobs only need one-time service before leaving the system. We use "N-model" as illustration and there are two architectures regarding where the jobs are queued. In this paper, we consider the Architecture 1 (Figure 1.1): upon arrival with rate λ, the system manager immediately addresses the new job and decides which buffer the job should be routed to. The decision rule that the system manager applies to select the buffer is called a routing policy. Each server has multiple classes of jobs waiting to be processed. The decision rule for the servers to choose jobs to process is called scheduling policy. A given server processing a given class is called an activity, associated with a buffer in the system with different mean service rates μ.

The performance of the parallel-server systems under different policies in heavy traffic has been studied intensively for the last 20 years, see, for example, [20], [19],[2]. They defined the heavy traffic system based on a *static allocation problem*, which was a linear programming problem that minimizes the utility of the busiest server. The parallel-server system in those papers consider the Architecture 2 (Figure 1.2), which is different from ours in terms of where the jobs are queued. Under their scenarios, jobs queue near the arrival, only scheduling policy is needed when the server is going to process the next one. However, the structure that servers working in parallel on different class jobs allows us to

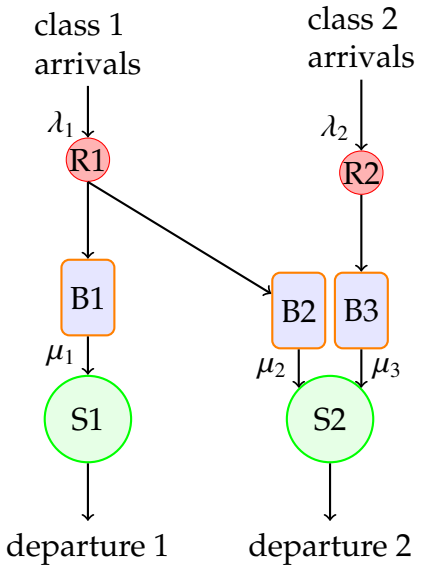

Figure 1.1: Architecture 1

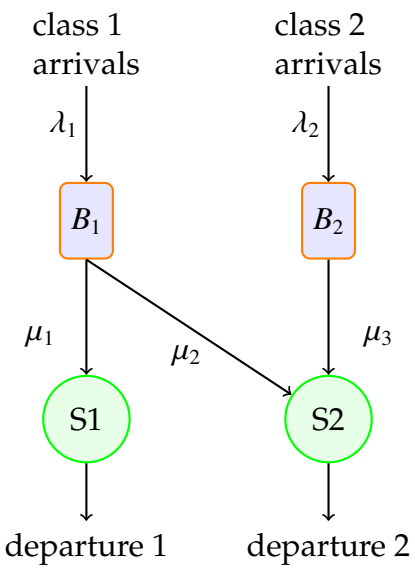

Figure 1.2: Architecture 2

formulate the heavy-traffic analysis based on the same optimal linear programming solution. The optimal solution with each element associated to one activity, indicated what proportion of time each server should allocate to its different job classes when each server is 100% busy. Based on the optimal solution, the activity corresponding to a non-zero proportion of optimal solution was called the *basic activity*, and the activity with zero proportion was called the *non-basic activity*. Assuming the unique optimal solution of static allocation problem, they found a unique way to achieve heavy traffic. They further showed that if the system was capable to balance the workload among the servers (i.e. servers communicate through basic activities), then the equivalent workload formulation among the servers became one-dimensional. This phenomenon was called *complete resource pooling*. In this formulation, state-space collapse to one dimension was crucial to establish the heavy traffic analysis using diffusion limits.

Threshold policy ([1] and [2]) is a scheduling policy designed based on the parallel-server system in which jobs wait near arrival and will be allocated un-

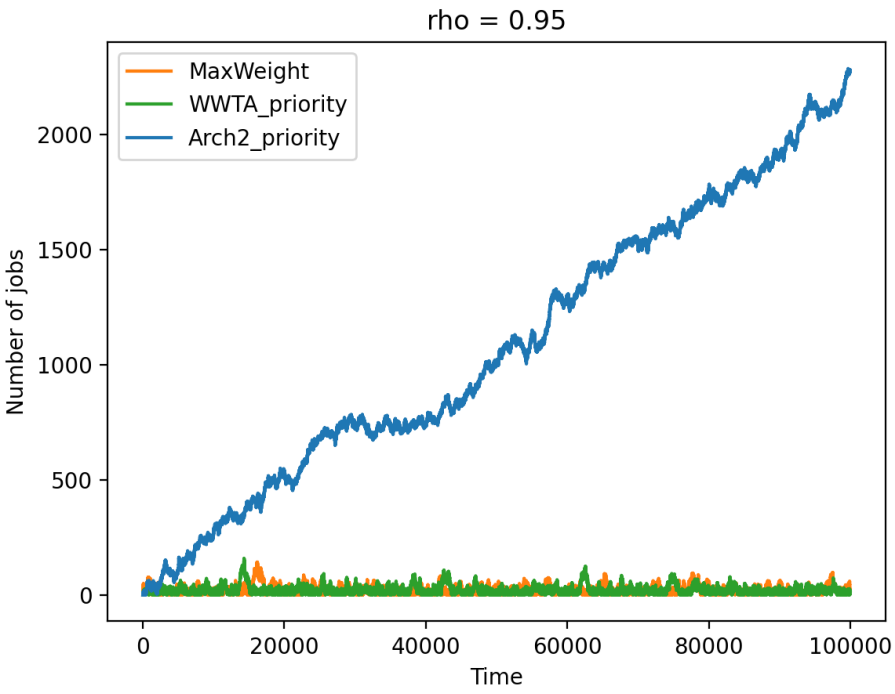

Figure 1.3: Priority scheduling under architecture 2 is not stable

til it will be served immediately. Based on the assumption of heavy traffic and complete resource pooling introduced above, they proved the Threshold policy was asymptotically optimal in the heavy traffic limit. However, the Threshold policy requires the knowledge of basic activities which meant the static allocation problem need to be solved before implementing the policy. Besides, the structure and primitives of the parallel-server system needs to be fully explored to build up a tree from the bottom to the root, aiming to set priority to different activities. Furthermore, the threshold level is a hyper-parameter which needs to be carefully chosen for each concrete issue. The WWTA (Weighed Workload Task Allocation) policy, which we will discuss in this paper, doesn't require the knowledge of optimal solution of static allocation problem. The reason we introduced it was for the performance analysis purpose. WWTA policy only needs the service rate for activities and dynamic queue lengths in the system. In fact,

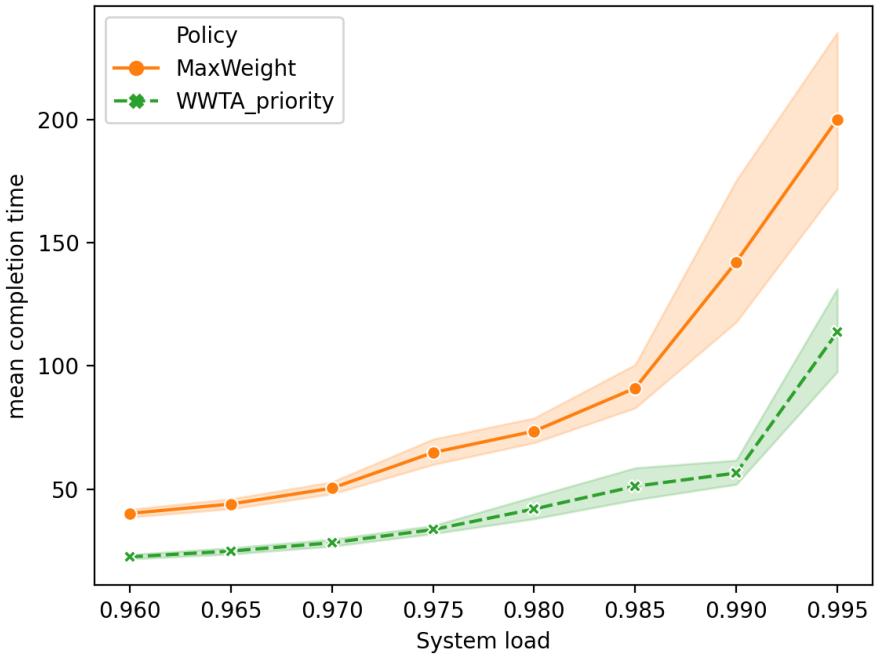

Figure 1.4: WWTA is better than Maxweight - achieving shorter mean completion time

we will show that the WWTA policy make the system work efficiently, magically coordinates the servers to behave as if under the guidance of the optimal solution in the heavy traffic, while without knowing it beforehand.

MaxWeight ([29]) is also a scheduling policy in which jobs wait near arrival until it will be served immediately. It is usually considered in the generalized switch model, which includes a discrete time version of the parallel-server model as a special case. Jobs arriving to the system wait near the arrival side, and MaxWeight policy is applied when the servers are ready to process the next one. According to the weighted queue length, server picks the class i job, if the corresponding weighted queue length is the maximum. Stolyar ([29]) further proved the state-space collapse, and the workload process converges to a Reflected Brownian Motion.

[13] exploited the state-space collapse to derive the moment bounds. More specifically, they used Lyapunov-drift based approach to get the upper and lower bound for the expected steady-state queue lengths, showing that the bounds were tight in heavy traffic limit. As an illustration, they got the moment bounds for the single-class-multi-server systems with join-the-shortest-queue (JSQ) routing policy and multi-class systems with MaxWeight scheduling policy. Based on the drift method, [23] used the Transform method as a generalization to discuss the heavy traffic performance. They used the moment generating function to show that the stationary distribution of scaled queue lengths was exponential. Unlike [13] and [23], in our discussion, we don't need to assume the boundedness of arrival and service process to make sure that the moment generating function exists. Instead, we utilize the Laplace transform since it always exists for a random variable.

In the setting with 3-level data locality, [32] established the throughput optimality of the WWTA policy, and proved heavy traffic delay optimality given the condition of locality and prioritized scheduling. [30] proposed Map Task Scheduling policy that are composed of the JSQ with MaxWeight policy. Comparing with it, WWTA policy is shown to be superior to JSQ-MaxWeight under some traffic scenarios([32]). In our discussion under the heavy traffic parallel-server system, we consider more general system structure and argue by state-space collapse in a workload version, where all moments of the maximum difference among workloads of servers are proved to be bounded in heavy traffic limit. Another generality of this paper is the discussion of scheduling policies. Besides prioritized scheduling used by [32], we have found out any non-idling scheduling policies are capable to work as good companion with WWTA.

Due to the difference of classes processed by each server, an accompanying scheduling policy need to be specified for WWTA policy to be considered simultaneously. But later we will show any non-idling scheduling policies are eligible, which will not put extra constraints, but giving more flexibility on the implementation of WWTA policy. The freedom of the choice for scheduling policy is entitled by good routing (WWTA) policy. As a "N-model" example shows in Figure 1.3, MaxWeight scheduling policy under Architecture 2 is stable, WWTA with priority scheduling policy under Architecture 1 is also stable. Here we give the priority to the class which has faster processing time. However, assigning the same priority for servers under Architecture 2 without routing policy makes the system even unstable.

Since any non-idling scheduling policy is able to accompany WWTA policy, by carefully choosing the scheduling, some performance can be improved. For example, in Figure 1.4 under same parameter setting, when we give priority to the jobs with faster service rate, more shorter-service-time jobs will be processed and discharged from the system under WWTA. As a result, the average completion time is shorter under WWTA than MaxWeight. Therefore, total number of jobs in the system under WWTA with prioritized scheduling is more likely to be less than MaxWeight, that is, even servers work for the same time length under two policies, WWTA picking shorter-service-time jobs means finishing larger number of jobs first. With the help of routing policy, the scheduling side can be optimized to achieve better performance.

As introduced in the beginning, [20] , [2] constructed static allocation problem(linear program) to define the heavy traffic of the parallel-server system. Under the assumption that the static allocation problem had unique solution in

the heavy traffic, they also constructed the dual linear program. They showed that one dimensional workload, unique optimal dual solution, and all servers communicating are equivalent conditions for Complete Resource Pooling. As [20] mentioned, it might be unnecessary to assume the uniqueness of optimal solution of static allocation problem, while multiple optima lead to complicated situations. In our discussion, we consider a more general case, including the case in which the primal solution is not unique. Under the assumption that primal optimal solution might not be unique, we find out there still exists a unique optimal dual solution under some necessary conditions. The uniqueness of dual solution is crucial for our analysis. The sum of tail probability that each server is idle, weighted by optimal dual solution, is proven to be precise w.r.t the load of the system. Besides, the dual solution adjusts the WWTA routing criterion with respect to different classes in a standardized way, which facilitate the discussion of state-space collapse. What's more, the limit of scaled sum of workload, also weighted by optimal dual solution, is proved to be exponential random variable, with parameters also depending on optimal dual solution. Therefore, depending on the uniqueness of dual solution, the assumption for the definition of heavy traffic based on primal solution can be relaxed to be non-unique.

According to the interpretation of static allocation problem, zero-value(non-basic) elements in the optimal primal solution mean the servers should not spare their time working on those activities in the heavy traffic. In fact, the WWTA policy can automatically distinguish the non-basic activities in the heavy-traffic load. As introduced later, WWTA with proper scheduling policies only require the knowledge of service rates and current queue lengths. It is easy to find out that expected proportion of time, P_{ik}, that each server k spare for class i, behaving under WWTA with any non-idling scheduling policy $\mathcal{P}$, actually con-

verges to the optimal solution of static allocation problem in the heavy traffic limit. In other words, we can theoretically calculate the optimal solution to achieve heavy traffic by solving the static allocation problem, however, provided a system where there is a way to achieve heavy traffic, WWTA with any non-idling scheduling policy will automatically choose that optimal way to work efficiently, without knowing the optimal solution beforehand. The robustness of WWTA policy is very appealing.

The rest of the paper has the following structure. In section 2, we formally introduce the static allocation problem, assuming that there exists optimal solution to achieve heavy traffic, and the Complete Resource Pooling condition holds. In section 3, we introduce WWTA policy and some properties as preparation for the heavy traffic analysis. The sum of tail probabilities that each server is idle or working on non-basic activities, weighted by optimal dual solution, is equal to the gap between the current system load and the critical load. In section 4, we state our main results based on WWTA policy. First, we show the state-space collapse in a workload version. In other words, all the moments of the maximum difference among the weighted workload for each server is bounded in the heavy traffic limit. Next, we prove that, the scaled sum of workload for servers, weighted by optimal dual solution, converges to a one-dimensional exponential distribution. Finally, the state-space collapse help us move one step further to get the marginal distribution of each workload. In section 5, we illustrate our results using an example "W" models. Simulation is implemented to justify our results.

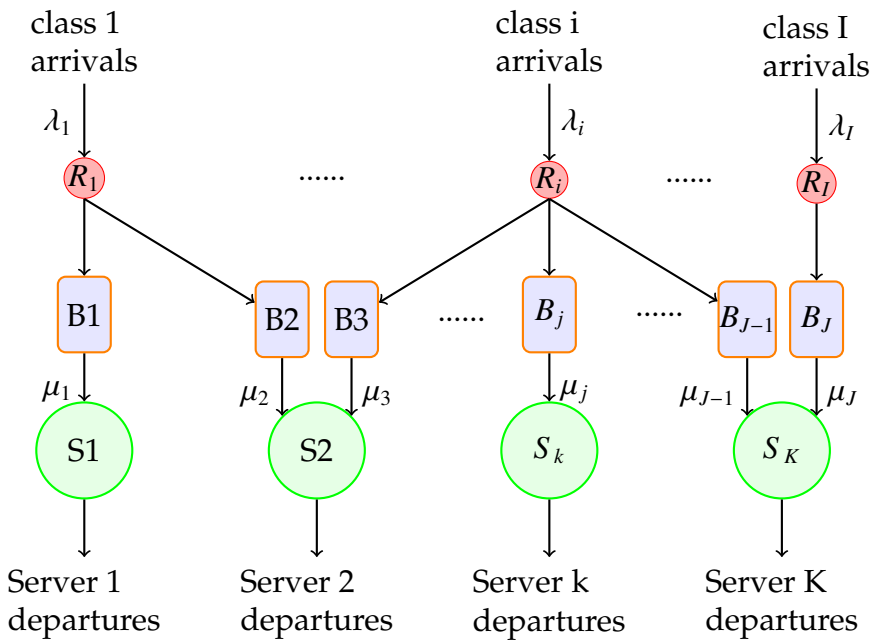

Figure 1.5: General parallel-server system

1.2 Parallel-server systems and policies

The N-model discussed in the introduction is among the simplest parallel parallel-server systems. This paper deals with the general parallel-server systems as illustrated in Figure 1.5. In the general system with K servers and I classes of arrivals, each arrival process is assumed to be Poisson. For class $i = 1, \ldots, I$, let λ_i be the arrival rate for class i arrivals. Each class i job can be processed by any one of the servers in the set $K(i) \subset \{1, 2, \ldots, K\}$. After being processed by a selected server, the job leaves the system. We call it a type $j = (i, k)$ activity when server k processes a class i job. We assume the processing times for type (i, k) activities are independent identically distributed exponential random variables with mean m_{ik}. Define $\mu_{ik} = 1/m_{ik}$ to be the service rate for type $j = (i, k)$ activities. We denote $I(k)$ the set of classes that server k can process. The total number of activities is denoted by J, which is at most $I \times K$.

Each class i arrival is immediately routed to one of the servers in $K(i)$, say, server $k \in K(i)$. If server k is free, the job enters processing immediately. Otherwise, it waits in buffer $j = (i,k)$ until the server is ready to process the job following a service policy to be specified below. Server k maintains multiple buffers, one for each type $j = (i,k)$ with $(i,k) \in I(k)$.

Weighted workload task allocation (WWTA) routing policy. We now formally introduce weighted workload task allocation (WWTA) routing policy. It is a load-balancing routing policy first proposed by [32]. Unlike JSQ routing policy, which simply compares the queue lengths among servers in its routing decisions, the WWTA policy compares workloads among servers in its decisions. In literature, the workload of a server at time t is defined to be the virtual waiting time at time t, which is the waiting time of a fictitious job arriving at the server at time t. In this paper, we assume the mean processing times are observable, but not the actual processing times. Thus, the virtual waiting times are non-observable quantities for a system manager. For our purpose, we define workload for server k to be

$$W_k(z) = \sum_{i \in I(k)} m_{ik} z_{ik},$$

where $z = (z_{ik})$ is the vector of job counts in the system. Here, component z_{ik} is the number of jobs in buffer $j = (i,k)$, including possibly the one in service. We assume at each time t, the jobcount vector z is observable. The WWTA policy routes an arriving job from class i to any server k which satisfies

$$argmin_{k \in K(i)} m_{ik} W_k(z).$$

Ties are broken randomly if there are multiple servers achieving the minimum.

Service policies. Since there might be several buffers associated with a server, we further need to specify a service or scheduling policy that dictates, for each server, from which buffer to choose a job to process next. It is known that under the WWTA routing policy, any non-idling service policy is throughput optimal; see, for example, Section 11.8 of [11]. By non-idling we mean, each server must be busy processing jobs whenever there are jobs waiting at its buffers. In the following, we introduce two types of scheduling policies that will be the focus of this study.

The first scheduling policy is the head-of-line proportional processor sharing (HLPPS) that was studied by [4]. Under HLPPS scheduling policy, all nonempty buffers receive service simultaneously. For class $i \in I(k)$, the proportion of utilization that the server k allocates to the class at any time is

$$P_{ik}(z) = \frac{z_{ik}}{\sum_{i \in I(k)} z_{ik}}, \quad k = 1, ...K \tag{1.1}$$

when the jobcount vector is $z = (z_{ik})$. Here and later, we adopt the convention that $0/0 = 0$. Thus, when $\sum_{i \in I(k)} z_{ik} = 0$, server k idles. Therefore, the dynamic service rate for buffer $j = (i, k)$ is $\mu_{ik} P_{ik}(z)$. The implementation of the HLPPS policy does not require the knowledge of system parameters, such as arrival rates and service rates. It also does not depend on the routing policies, but only on the proportion of the queue sizes at each of the servers.

The allocation in (1.1) can be generalized to

$$P_{ik}(z) = \frac{c_{ik} z_{ik}}{\sum_{i \in I(k)} c_{ik} z_{ik}}, \quad i = 1, ...I, k = 1, ...K, \tag{1.2}$$

where $c = (c_{ik}) > 0$ is a given vector of positive numbers. We call the scheduling policy using allocation (1.2) a *generalized HLPPS* policy with weight $c = (c_{ik})$.

The second type of scheduling policies is the static buffer priority (SBP) poli-

cies. Each SBP policy corresponds to a ranking among buffers. Given a ranking, we use $(i', k) \prec (i, k)$ denotes that buffer (i', k) has a (preemptive) higher priority than buffer (i, k). Formally, we define the SBP scheduling policy for server k by specifying its allocation

$$P_{ik}(z) = \mathbb{1}\Big(\sum_{(i',k)\prec(i,k)} z_{i'k} = 0, z_{ik} > 0 \Big).$$

By ranking buffers according to the shortest meaning processing time first, the corresponding SBP scheduling policy is shown to have superior performance in our simulation studies.

For future usage, for each type of scheduling policy, we define $\bar{P}_k(z)$

$$\bar{P}_k(z) = 1 - \sum_{i\in I(k)} P_{ik}(z) = \mathbb{1}\left(\sum_{i\in I(k)} z_{ik} = 0 \right).$$

It represents the proportion of unused capacity of server k. Under a non-idling policy, $\sum_{i\in I(k)} z_{ik} > 0$ implies that $\bar{P}_k(z) = 0$.

Therefore, throughout the proof, we can use a general notation $\mathcal{P}$ to represent those eligible scheduling policies, with corresponding individual plan $P_{ik}(z)$, indicating how many efforts server k need to make on different classes. Later on, we will keep using this general notation $\mathcal{P}$, since the majority of results doesn't rely on the specific scheduling policies, and it is easy to discuss the system with any one of the proper scheduling policies by replacing $P_{ik}(z)$ with specific expression whenever needed.

1.3 Assumptions and main results

In this section, we will formally introduce two critical assumptions and main results. One assumption is the heavy traffic and the other is the complete resource

pooling. They are standard in literature. These two assumptions are formulated through solutions to a linear program (LP), which was first introduced in [20]. For that, it is useful to adopt the compact notational system in [20]. Central in that system is the concept of activities. In the setting of a parallel server system introduced in Section 1.2, an activity j corresponds to a buffer (i,k) for a certain job class i and a certain server k. We assume J is the total number of activities.

Define a $I \times J$ *constituency* matrix C and a $K \times J$ *resource-consumption* matrix A as follows.

$$C_{ij} = \begin{cases} 1, & \text{if activity } j \text{ processes class } i; \\ 0, & \text{otherwise.} \end{cases}$$

$$A_{kj} = \begin{cases} 1, & \text{if server } k \text{ performs activity } j; \\ 0, & \text{otherwise.} \end{cases}$$

Given these two matrices, each activity $j = 1, \ldots, J$ is uniquely associated with a class i and a server k, allowing us to write $j = (i,k)$. We assume J activities are ordered from 1 to J. We denote the mean service rates of J activities by

$$\mu = (\mu_1, \ldots, \mu_J)^T.$$

Define *output* matrix

$$R = C\,\mathrm{diag}(\mu),$$

where R_{ij} is the job departure rate from buffer (i,k) when $j = (i,k)$ and server k devotes all its effort on the buffer.

We consider the following *static allocation problem*:

$$\begin{aligned} \min \quad & \rho \\ \text{s.t.} \quad & Rx = \lambda \\ & Ax \leq \rho e \\ & x, \rho \geq 0 \end{aligned} \tag{1.3}$$

where $\lambda = (\lambda_1, ...\lambda_I)^T$, $x = (x_1, ...x_J)^T$, $e = (1, ...1)^T \in \mathbb{R}^K$. The vector x can be regarded as a processing plan, with each element x_j interpreted as the long-run proportion of time that activity j is processed by its server and ρ interpreted as the long-run utilization of the busiest server. As [20] considered, the minimization problem (1.3) aims for a relatively even processing plan of allocation among servers. Following [20], we define the following notion of *balanced heavy traffic*: even under the most efficient processing plan, *all* servers are 100% utilized. Formally, we state the following assumption.

Assumption 1 (Heavy Traffic)**.** *The parallel server system is assumed to be in (balanced)* heavy traffic*, namely, the static allocation problem* (1.3) *has an optimal solution* (x^*, ρ^*) *that satisfies*

$$\rho^* = 1 \quad \textit{and} \quad Ax^* = e. \tag{1.4}$$

Remark 1. *We do not assume linear program (1.3) has a unique solution. The uniqueness is assumed in [20], [2], and [19]. Analysis in these papers utilized the uniqueness property critically. Non-uniqueness of the LP solutions means that there might exist multiple ways to allocate the time for servers, such that most efficiently, all servers can be 100% busy.*

Following [20], we define the basic activities. Note that the possible non-uniqueness of LP solution (x^*, ρ^*), therefore, we define activity $j = (i, k)$ is called

a *basic activity associated with some* x^* if $x^*_j > 0$. Otherwise, it is a called *non-basic activities associated with* x^*. Similarly, we consider the communicating servers using following definition:

Definition 1. *Servers k and k' are said to communicate directly, if there exists some x^*, such that both $j = (i,k)$ and $j' = (i,k')$, for some class i, are basic activities associated with x^*. We further call that such server k and server k' are neighbors. Server k and k' are said to communicate, if there exist servers linking them that communicate directly to each other. In other words, if server k and k' communicate, then server k is the neighbor of the neighbors of server k'.*

Then we introduce our second assumption as follows:

Assumption 2 (Complete Resource Pooling (CRP))**.** *There exist optimal solution(s) (x^*, ρ^*) that satisfies (1.4) and all servers communicate.*

Remark 2. *Assumption 2 contains Assumption 1. Under Assumption 2, when we search for basic activities to communicate all the servers, we are allowed to utilize different optimal solutions x^*, when we find the next neighbor of the current server.*

We end this section by stating two lemmas that related to the dual of LP (1.3). The dual LP to the static allocation problem (1.3) is defined as follows:

$$\begin{aligned} \max \quad & v\lambda \\ s.t. \quad & vR \le uA \\ & ue \le 1 \\ & u \ge 0 \end{aligned} \tag{1.5}$$

where $v = (v_1, ...v_I)$, $u = (u_1, ...u_K)$.

Lemma 1. *Under Assumption 1, the dual LP (1.5) has an optimal solution (v^*, u^*), satisfying*

(i) $\sum_{i=1}^{I} \lambda_i v_i^* = 1$

(ii) $\sum_{k=1}^{K} u_k^* = 1,\ u_k^* \geq 0$

(iii) *If* $x_{ik}^* > 0$, *then* $\mu_{ik} v_i^* = u_k^*$

Proof. Under Assumption 1, by strong duality, the dual LP also has optimal solution (v^*, u^*), and the duality gap is zero. Therefore

$$\sum_{i=1}^{I} \lambda_i v_i^* = \rho^* = 1.$$

For the second constrain in dual LP, complementary slackness gives

$$\sum_{k=1}^{K} u_k^* = 1$$

Furthermore, complementary slackness also gives

$$(Ax^*)_k = 1 \quad or \quad u_k^* = 0, \quad k = 1, ...K$$

$$x_j^* = 0 (x_{ik}^* = 0) \quad or \quad (vR)_j = (uA)_j, \quad j = 1, ...J$$

where $(vR)_j = \mu_{ik} v_i^*,\ (uA)_j = u_k^*$. □

Lemma 2. *If Assumptions 1 & 2 holds, then*

(i) *The optimal dual LP solution* (v^*, u^*) *is unique.*

(ii) $u_k^* > 0, v_i^* > 0, \quad \forall k = 1, ...K, i = 1, ...I$

Proof. Under Assumptions 1 & 2, if the primal solution of static allocation problem is further assumed to be unique, this case has been fully discussed by [20]. Lemma 2(*i*) is one of the equivalent statements of the *complete resource pooling*, and Lemma 2 (*ii*) is a corollary under their setting. Our proving steps, however, provide an analytical way to obtain optimal dual solution, which doesn't depend on the uniqueness of primal LP solution.

Now starting with any one of the servers, k, denote $u_k^* = a \geq 0$, and Assumption 2 guarantees that we can find at least one neighbor server for it. Here suppose server k has two neighbors k_1, k_2, as illustration, that communicate directly with server k through some classes i_1, i_2 via basic activities:

$$\begin{array}{ccc} \textit{server } k & \overset{\textit{class } i_1}{\rightarrow} & \textit{server } k_1 \\ \downarrow_{\textit{class } i_2} & & \\ \textit{server } k_2 & & \end{array}$$

In other words, server $k \to k_1$ has basic activities (i_1, k) and (i_1, k_1), $k \to k_2$ has basic activities (i_2, k_2), (i_2, k_2)

By Lemma 1, for any optimal solution x^*, we notice that each $x_{ik}^* > 0$ corresponds to a basic activity with equation $\mu_{ik} v_i^* = u_k^*$. Therefore, we have

$$v_{i_1}^* = \frac{u_k^*}{\mu_{i_1 k}} = \frac{a}{\mu_{i_1 k}} \quad v_{i_2}^* = \frac{u_k^*}{\mu_{i_2 k}} = \frac{a}{\mu_{i_2 k}}$$

$$u_{k_1}^* = v_{i_1}^* \mu_{i_1 k_1} = a\frac{\mu_{i_1 k_1}}{\mu_{i_1 k}} \quad u_{k_2}^* = v_{i_2}^* \mu_{i_2 k_2} = a\frac{\mu_{i_2 k_2}}{\mu_{i_2 k}}$$

which means $u_{k_1}^*$, $u_{k_2}^*$ for server k_1 and k_2 can be expressed as $u_k^* = a$ multiplied by the ratio of some mean service rates.

Similarly, starting with server k_1 and k_2, we can also find other neighbor servers of them, respectively, with the help of basic activities to obtain $u_{k_3}^*, u_{k_4}^* \ldots$, to express them as $u_k^* = a$ multiplied by ratios of some mean service rates. Eventually, by Assumption 2, we will go over all the servers and obtain such expression for each server. As the last step, by Lemma 1: $\sum_{k=1}^{K} u_k^* = 1$, we can solve $u_k^* = a$ uniquely. Then each element in (v, u) can also be solved explicitly.

In the discussion above, we pick one possible i_k in each step to obtain the solution (v^*, u^*). Each choice of i_k might not be unique due to the multiple choices

of basic activities for communication. Therefore, it is likely that there are some unused equations due to unused basic activities. While since Lemma 1 already guarantees the existence of optimal dual solution, this (v^*, u^*) should satisfy the unused equations, otherwise it is not an optimal solution. Therefore, the optimal dual solution is unique.

For proving Lemma 2(ii), it is obvious that $a > 0$, then from the equations utilized above, $u_k^* > 0$, $v_i^* > 0$, $\forall k \in K, \forall i \in I$. □

Lemma 2 is crucial in building up the explicit performance results even under the condition of non-unique optimal solutions of static allocation problem 1.3. Throughout the following discussion, we consider the system for which the Assumptions 1 & 2 hold, i.e. the heavy traffic system that satisfies CRP condition. We have proved in Lemma 2, that optimal dual solution (v^*, u^*) is unique, therefore, from now on, we will always use this unique dual solution, and we omit superscript "*" for simplicity.

In order to discuss the heavy traffic, we construct a sequence of parallel server systems approaching the heavy traffic in the load. That is, each system in the sequence is under the load such that the arrival rates are parameterized by ϵ:

$$\lambda_i^{(\epsilon)} = \lambda_i(1 - \epsilon), 0 < \epsilon < 1$$

with the other settings of the systems being the same. Then the load of the sequence of systems approaches 100%, as ϵ goes to zero. When $0 < \epsilon < 1$, the parallel-server system with the WWTA policy is proven to be throughput optimal([11], [32]), which means the vector of queue length has stationary distribution. Therefore, we can further let $Z^{(\epsilon)}(\infty), 0 < \epsilon < 1$ be the vector of steady-state queue length in the system w.r.t ϵ. Throughout the paper, all of the results

will be discussed based on this steady-state queue length $Z^{(\epsilon)}(\infty)$.

Theorem 3 (Limit distribution for Individual Workload)**.** *Consider a parallel-server system that satisfies Assumptions 1 & 2. Under the WWTA policy and scheduling $\mathcal{P}$, the limit distribution of scaled workload for each server is one-dimensional exponential distribution. Furthermore,*

$$\lim_{\epsilon \downarrow 0} \epsilon \left(W_1\left(Z^{(\epsilon)}(\infty)\right), ..., W_K\left(Z^{(\epsilon)}(\infty)\right) \right) \xrightarrow{d} (u_1, ..., u_K)\, X$$

where X is a random variable that follows exponential distribution:

$$X \sim Exponential\left(\frac{\sum_{k=1}^{K} u_k^2}{\sum_{i=1}^{I} \lambda_i v_i^2} \right)$$

In Theorem 3, the limit distribution only depend on the arrival rates $\lambda_i, i = 1, ..I$ and unique optimal dual solution (v, u). It doesn't depend on the specific scheduling policy $\mathcal{P}$. The proof of Theorem 3 is provided in Section 1.8. The key ingredient for proving Theorem 3 is called *State-space Collapse*, which will be presented in Section 1.6. Other supporting results will be introduced in Section 1.5 and Section 1.7.

1.4 Example and Simulation

In this section we would like give a more complex model to show how to explicit obtain the limit distribution for individual workload. Also we will use this model to compare WWTA policy under different scheduling policies with Maxweight policy. For simplicity, we consider an example whose static allocation problem has unique solution: W model, which is described in Figure 1.6 that has three classes of jobs and two servers. We evaluate the performance in

term of the average completion time for each job starting from arrival to leaving in the system. Its static allocation problem is

$$\begin{aligned} \min \quad & \rho \\ s.t. \quad & x_{11} + x_{21} + x_{31} \le \rho \\ & x_{12} + x_{22} + x_{32} \le \rho \\ & \mu_{11}x_{11} + \mu_{12}x_{12} = \lambda_1 \\ & \mu_{21}x_{21} + \mu_{22}x_{22} = \lambda_2 \\ & \mu_{31}x_{31} + \mu_{32}x_{32} = \lambda_3 \end{aligned}$$

We consider the following setting of parameters in the system: $\mu_{11} = 8$, $\mu_{21} = 2$, $\mu_{31} = 0.25$, $\mu_{12} = 0.25$, $\mu_{22} = 0.5$, $\mu_{32} = 1$, $\lambda_1 = 4$, $\lambda_2 = 1.3$, $\lambda_3 = 0.4$. Then we have unique optimal solution for static allocation problem in W model:

$$x_{11}^* = 0.5, x_{12}^* = 0, x_{21}^* = 0.5, x_{22}^* = 0.6, x_{31}^* = 0, x_{32}^* = 0.4, \rho^* = 1$$

where activities {12} and {31} are non-basic activities by definition. Therefore, the servers communicate in this W model through activities $\{11\}, \{21\}, \{22\}, \{32\}$. The optimal dual solution is also unique according to Lemma 2(i):

$$u_1^* = \frac{4}{5} \quad u_2^* = \frac{1}{5}, \quad v_1^* = \frac{1}{10} \quad v_2^* = \frac{2}{5}, \quad v_3^* = \frac{1}{5},$$

Here we consider the W model under WWTA and HLPPS policy, applying and Theorem 3, we can solve the individual workload as follows:

$$\lim_{\epsilon \downarrow 0} \epsilon W_1(Z^{(\epsilon)}(\infty)) \xrightarrow{d} X_1 \sim Exponential\left(\frac{425}{132}\right)$$
$$\lim_{\epsilon \downarrow 0} \epsilon W_1(Z^{(\epsilon)}(\infty)) \xrightarrow{d} X_2 \sim Exponential\left(\frac{425}{33}\right)$$

In the simulation, we simulated a real processing system under several system loads from 96% to 99.5%. We recorded the time for each job staying in the system which arrived to the system and was served and discharged later. We

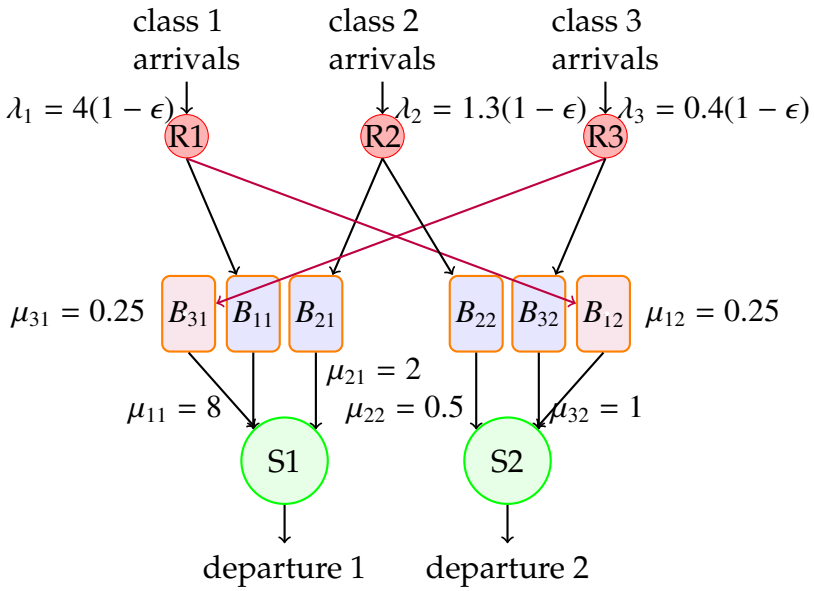

Figure 1.6: W model

called this time as job's "completion time". Under each system load, we run the system under different policies for 30 replicates each of which run for 50000 time units and plotted the average completion time with 95% confidence interval. As we saw in Figure 1.7, WWTA with Prioritized scheduling and WWTA with HLPPS has shorter completion time than the MaxWeight. The 95% confidence intervals almost didn't have overlap, which showed significant differences among the policies in terms of the mean completion time.

1.5 Preliminary Results I

As we introduce the policy assumptions and main results in the previous section, this section will provide the fundamental framework and some general supporting results before utilizing the State-space Collapse which will be introduced in the next section. Recall that in the WWTA Routing Policy, allocation

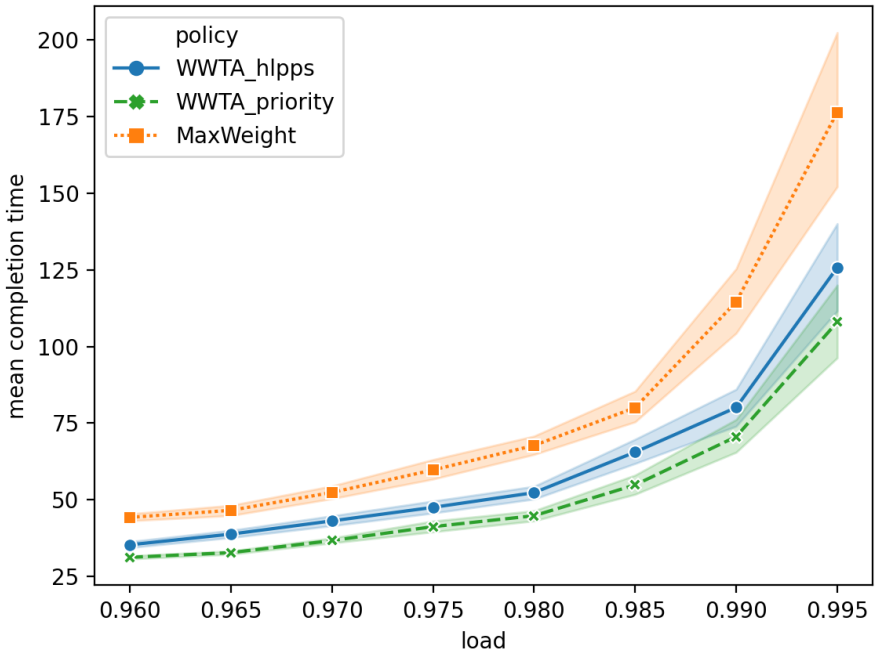

Figure 1.7: Policy comparison under W model

for each class i arrival, depends on the comparing criterion among the eligible servers whoever achieve

$$argmin_{k\in K(i)} m_{ik} W_k(z).$$

For simplicity, we denote $T_{ik}(z) = m_{ik}W_k(z)$. That is, when class i job comes to the system, we route it to server k' if $k' = argmin_{k\in K(i)} T_{ik}(z)$, and route it randomly to one of those minimizers, if there are multiple servers achieve the minimum. Then for each class i, we split the possible routing into the cases with respect to server with index $H^{(i)}(z) \triangleq argmin_{k\in K(i)} T_{ik}(z)$. That is, if we let "$\mathbb{1}$(event)" be the indicator function, then

$$\sum_{k\in K(i)} \mathbb{1}\left(k = H^{(i)}(z)\right) = 1.$$

Additionally, as discussed in the previous section, we use a general notation P_{ik} for any eligible scheduling $\mathcal{P}$, and the proofs can be modified easily for a specific

scheduling policy.

We let $e_{ik} \triangleq e_j = (0, ...1, ...0)$ be the unit vector with only the jth element being nonzero, and we can write the generator of general parallel-server model under the WWTA and scheduling $\mathcal{P}$:

$$\begin{aligned} Gf(z) = &\sum_{i=1}^{I} \lambda_i^{(\epsilon)} \sum_{k \in K(i)} \Big[f(z + e_{ik}) - f(z) \Big] \mathbb{1}\left(k = H^{(i)}(z) \right) \\ &+ \sum_{k=1}^{K} \sum_{i \in I(k)} \mu_{ik} P_{ik}(z) \Big[f(z - e_{ik}) - f(z) \Big] \end{aligned} \tag{1.6}$$

In Lemma 1, we have discussed that if (v, u) is optimal dual solution, then each activity satisfies $u_k \geq \mu_{ik} v_i$, and each basic activity satisfies $u_k = \mu_{ik} v_i$. For simplicity, we denote $u_k - \mu_{ik} v_i \triangleq d_{ik}$, then non-basic activities generally have $d_{ik} \geq 0$. Due to the reasons explained in the following Remark, without loss of generality, throughout the paper, we denote the non-basic activities are those activities with $u_k > \mu_{ik} v_i$, which means we have the following key relation:

$$\mu_{ik} = \frac{u_k - d_{ik}}{v_i}, \quad d_{ik} = \begin{cases} 0 & \text{(i,k) is basic activity} \\ > 0 & \text{(i,k) is non-basic activity} \end{cases} \tag{1.7}$$

Furthermore, we use "b" and "nb" to denote the cases corresponding to basic and non-basic activities whenever needed.

Remark 3. *Note that basic activities might not be uniquely determined due to the potential non-uniqueness of x^*. According to the primal-dual theory, if $u_k > \mu_{ik} v_i$ holds for the activity (i, k), then in any one of the optimal primal solutions, this activity (i, k) should be non-basic with $x^*_{ik} = 0$. In other words, if there exist an optimal primal solution where $x^*_{ik} > 0$, then the activity (i, k) should satisfy $u_k = \mu_{ik} v_i$. These activities with strict inequalities (equivalently, $d_{ik} > 0$) bring difficulties throughout the proof of results, therefore need extra efforts to be carefully handled. Moreover, it is worthy to*

notice that since $u_k > 0$, then $v_i^* > 0$, these activities with $d_{ik} > 0$ have smaller μ_{ik} than it should be utilized as basic activities under any primal solution, resulting in inefficient activities. We allow the existence of such inefficient activities that has $d_{ik} > 0$, and we will further show that they are actually negligible in the heavy traffic. Hence, without loss of generality, throughout the paper, when we discuss the non-basic activities, we mean the non-basic activities with $u_k > \mu_{ik} v_i$.

Lemma 4. *Under Assumptions 1 & 2, with the WWTA and scheduling $\mathcal{P}$,*

(a) Let $f(z) : \mathbb{R}^J \to \mathbb{R}$ be a function. Suppose there exists $n \in \mathbb{N}^+$ such that $|f(z)| \leq C \sum_{k=1}^K W_k^n(z)$ for some $C > 0$ (i.e. $f(z)$ is dominated by a polynomial function of workload). Then the vector of steady-state queue length $Z^{(\epsilon)}(\infty), 0 < \epsilon < 1$ satisfies

$$\mathbb{E}\Big[Gf\Big(Z^{(\epsilon)}(\infty)\Big)\Big] = 0$$

(b) As a consequence of (a), for any $n \in \mathbb{N}^+$, $\exists M_n > 0$, when $0 < \epsilon < 1$,

$$\epsilon \sum_{k=1}^K \mathbb{E}\Big[W_k^n\Big(Z^{(\epsilon)}(\infty)\Big)\Big] \leq M_n$$

The proof of Lemma 4 is provided in the Appendix A.1.1 and relies on the similar discussion as [[5], Lemma 1].

The following Lemma describes the probability that servers encounter idleness or non-basic activities are activated are actually bounded up to order of ϵ. Therefore, as $\epsilon \downarrow 0$, those two events are negligible.

Lemma 5. *Under Assumptions 1 & 2, with the WWTA and scheduling $\mathcal{P}$, the probability that each server in the system w.r.t ϵ is idle is at most $O(\epsilon)$, and more precisely,*

$$\sum_{k=1}^K u_k \mathbb{P}\Big(\sum_{i \in I(k)} Z_{ik}^{(\epsilon)}(\infty) = 0\Big) + \sum_{k=1}^K \sum_{i \in I(k)} d_{ik} \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}(\infty))\right] = \epsilon$$

Proof. We let $f(z) = \sum_{i=1}^{I} \sum_{k\in K(i)} v_i z_{ik}$, where v_i is from optimal dual solution (v, u). Then the generator (1.6) becomes

$$\begin{aligned}
Gf(z) &= \sum_{i=1}^{I} \lambda_i^{(\epsilon)} v_i \sum_{k\in K(i)} \mathbb{1}\Big(k = H^{(i)}(z)\Big) - \sum_{k=1}^{K} \sum_{i\in I(k)} \mu_{ik} v_i P_{ik}(z) \\
&= \sum_{i=1}^{I} \lambda_i^{(\epsilon)} v_i - \sum_{k=1}^{K} \sum_{i\in I(k)} \mu_{ik} v_i P_{ik}(z) \\
&\overset{(a)}{=} 1 - \epsilon - \sum_{k=1}^{K} u_k \sum_{i\in I(k)} P_{ik}(z) + \sum_{k=1}^{K} \sum_{i\in I(k)} d_{ik} P_{ik}(z) \\
&= 1 - \epsilon - \sum_{k=1}^{K} u_k + \sum_{k=1}^{K} u_k \mathbb{1}\Big(\sum_{i\in I(k)} z_{ik} = 0\Big) + \sum_{k=1}^{K} \sum_{i\in I(k)} d_{ik} P_{ik}(z) \\
&\overset{(b)}{=} -\epsilon + \sum_{k=1}^{K} u_k \mathbb{1}\Big(\sum_{i\in I(k)} z_{ik} = 0\Big) + \sum_{k=1}^{K} \sum_{i\in I(k)} d_{ik} P_{ik}(z)
\end{aligned}$$

where (a) and (b) are by lemma 1 with notation (1.7). Then by Lemma 4(a), we have

$$\sum_{k=1}^{K} u_k \mathbb{P}\Big(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)}(\infty) = 0\Big) + \sum_{k=1}^{K} \sum_{i\in I(k)} d_{ik} \mathbb{E}\Big[P_{ik}\Big(Z^{(\epsilon)}(\infty)\Big)\Big] = \epsilon$$

□

Lemma 6. *Under Assumptions 1 & 2, with the WWTA and scheduling $\mathcal{P}$,*

$$\lambda_i^{(\epsilon)} \mathbb{P}\Big(k = H^{(i)}\Big(Z^{(\epsilon)}(\infty)\Big)\Big) = \mu_{ik} \mathbb{E}\Big[P_{ik}\Big(Z^{(\epsilon)}(\infty)\Big)\Big]$$

Furthermore, for non-basic activities (i, k) s.t. $d_{ik} > 0$, we have

$$\mathbb{P}\Big(k = H^{(i)}\Big(Z^{(\epsilon)}(\infty)\Big)\Big) = O(\epsilon) \tag{1.8}$$

Proof. We denote $f(z) = z_{ik}$, then the generator (1.6) becomes

$$Gf(z) = \lambda_i^{(\epsilon)} \mathbb{1}\left(k = H^{(i)}(z)\right) - \mu_{ik} P_{ik}(z)$$

by Lemma 4(a), we have

$$\lambda_i^{(\epsilon)} \mathbb{P}\Big(k = H^{(i)}\Big(Z^{(\epsilon)}(\infty)\Big)\Big) = \mu_{ik} \mathbb{E}\Big[P_{ik}\Big(Z^{(\epsilon)}(\infty)\Big)\Big]$$

By Lemma 5, non-basic activity with $d_{ik} > 0$ has $\mathbb{E}\left[P_{ik}\left(Z^{(\epsilon)}(\infty)\right)\right] \leq \epsilon$, therefore,

$$\mathbb{P}\left(k = H^{(i)}\left(Z^{(\epsilon)}(\infty)\right)\right) = O(\epsilon), \quad (i,k) \text{ is non-basic activity}$$

□

Lemma 7. $\forall x \geq 0$,

$$1 - e^{-x} \leq x$$

1.6 State-space Collapse

Denote

$$T_k(z) = \frac{1}{u_k} W_k(z), k = 1, ...K$$

then we state the result of State-Space Collapse as following, which describes the balance of workload among the servers:

Theorem 8 (State-Space Collapse)**.** *Consider a parallel-server system that satisfies Assumptions 1 & 2. Under the WWTA policy and scheduling $\mathcal{P}$, there exists ϵ_0 s.t. $0 < \epsilon < \epsilon_0$, any nth moments of max difference among the routing criterion are bounded. i.e. there exist constants $M_n \geq 0, n \in \mathbb{N}^+$, such that*

$$\mathbb{E}\left[\max_{k\in\{1,...K\}} T_k\left(Z^{(\epsilon)}(\infty)\right) - \min_{k\in\{1,...K\}} T_k\left(Z^{(\epsilon)}(\infty)\right)\right]^n \leq M_n$$

Remark 4. *Theorem 8 provides the state-space collapse in a workload version. It actually provides a much stronger state-space collapse than what we need in the discussion of limit distribution, which only requires the 1st and 2nd moment boundedness. This is thanks to the utilization of Laplace transform.*

Before proving Theorem 8, we first introduce Lemma 9 below:

Lemma 9.

$$\frac{1}{\sum_{i\in I(k)}^{b}\lambda_i v_i} \le \frac{1}{u_k}, k = 1, ...K$$

Proof. We denote $f(z) = \sum_{i\in I(k)} v_i z_{ik}$, $k = 1, ...K$. Then the generator (1.6) becomes

$$\begin{aligned} Gf(z) &= \sum_{i\in I(k)} \lambda_i^{(\epsilon)} v_i \mathbb{1}\left(k = H^{(i)}(z)\right) - \sum_{i\in I(k)} \mu_{ik} v_i P_{ik}(z) \\ &= \sum_{i\in I(k)} \lambda_i^{(\epsilon)} v_i \mathbb{1}\left(k = H^{(i)}(z)\right) - \sum_{i\in I(k)} (u_k - d_{ik}) P_{ik}(z) \\ &= \sum_{i\in I(k)} \lambda_i^{(\epsilon)} v_i \mathbb{1}\left(k = H^{(i)}(z)\right) - u_k\Big(1 - \bar{P}_k(z)\Big) + \sum_{i\in I(k)} d_{ik} P_{ik}(z) \end{aligned}$$

where d_{ik} comes from with notation (1.7). By Lemma 4(a), we have

$$\begin{aligned} &\sum_{i\in I(k)} \lambda_i v_i \mathbb{P}\Big(k = H^{(i)}\Big(Z^{(\epsilon)}(\infty)\Big)\Big) - \sum_{i\in I(k)} \lambda_i v_i \epsilon \mathbb{P}\Big(k = H^{(i)}\Big(Z^{(\epsilon)}(\infty)\Big)\Big) \\ =& u_k - u_k \mathbb{P}\Big(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)}(\infty) = 0\Big) - \sum_{i\in I(k)} d_{ik} \mathbb{E}\Big[P_{ik}\Big(Z^{(\epsilon)}(\infty)\Big)\Big] \end{aligned}$$

since $0 \le \mathbb{P}\Big(k = H^{(i)}(Z^{(\epsilon)})\Big) \le 1$, with Lemma 5 and Lemma 6, we have

$$\begin{aligned} u_k &= \sum_{i\in I(k)}^{b} \lambda_i v_i \mathbb{P}\Big(k = H^{(i)}\Big(Z^{(\epsilon)}(\infty)\Big)\Big) + \sum_{i\in I(k)}^{nb} \lambda_i v_i \mathbb{P}\Big(k = H^{(i)}\Big(Z^{(\epsilon)}(\infty)\Big)\Big) + O(\epsilon) \\ &\le \sum_{i\in I(k)}^{b} \lambda_i v_i + O(\epsilon) \end{aligned}$$

where as a reminder "b" and "nb" denote basic and non-basic activities, *resp*. Since $u_k > 0, \forall k = 1, ...K$, then let $\epsilon \downarrow 0$,

$$\frac{1}{\sum_{i\in I(k)}^{b}\lambda_i v_i} \le \frac{1}{u_k}$$

□

Proof of Theorem 8. Let

$$f(z) = \left(\max_{k\in\{1,...K\}} T_k(z) - \min_{k\in\{1,...K\}} T_k(z)\right)^{n+1} = R^{n+1}(z)$$

We recall that $T_k(z) = \frac{1}{u_k} \sum_{i \in I(k)} \frac{1}{\mu_{ik}} z_{ik}$ and denote

$$H(z) = argmin_{k \in \{1,...K\}} T_k(z)$$

$$L(z) = argmax_{k \in \{1,...K\}} T_k(z)$$

where if there are multiple servers are minimizers, choose server k, if k is the smallest index of those minimizers; if there are multiple servers are maximizers, we choose server k, if k is the largest index of those maximizers. We first discuss some terms in the generator. Because of the difference of basic and non-basic activities as showed in 1.7, we put "-b" and "-nb" to denote them, respectively, in the following. Then we have

$$\begin{aligned}
&(f(z+e_{ik}) - f(z)) \mathbb{1}\left(k = H^{(i)}(z)\right) \\
=&(f(z+e_{ik}) - f(z)) \mathbb{1}\left(k = H^{(i)}(z), k = H(z)\right) \\
&+ (f(z+e_{ik}) - f(z)) \mathbb{1}\left(k = H^{(i)}(z), k = L(z)\right) \\
=&\left(\left(R(z) - \frac{1}{u_k \mu_{ik}}\right)^{n+1} - R^{n+1}(z)\right) \mathbb{1}\left(ik - b, k = H(z)\right) \\
&+ \left(\left(R(z) - \frac{1}{u_k \mu_{ik}}\right)^{n+1} - R^{n+1}(z)\right) \mathbb{1}\left(ik - nb, k = H(z), k = H^{(i)}(z)\right) \\
&+ \left(\left(R(z) + \frac{1}{u_k \mu_{ik}}\right)^{n+1} - R^{n+1}(z)\right) \mathbb{1}\left(k = L(z), k = H^{(i)}(z), k!bi\right)
\end{aligned} \tag{1.9}$$

where $k!bi$ means server k is the only server for class i that the activity (i,k) is basic. Since all servers communicate, then the number of class i that has only one basic activity is no more than I, i.e. those classes are on the margin of the parallel-server system. Each of these classes only links to the system by one server through basic activities, and don't need another basic activity to link further servers. In reality, some parallel-server systems have fewer such classes or none. Suppose $i_1 ... i_S$ be such S classes, the corresponding basic activities link server $k_1, ...k_S$, *resp*. There are also other terms ruled out in the (1.9), by discussing the corresponding events don't exist. Then by binomial expansion,

with Lemma 1 and notation (1.7), the generator (1.6) becomes

$$
\begin{aligned}
Gf(z) = &-(n+1)\sum_{i=1}^{I}\lambda_i^{(\epsilon)}v_i\sum_{k\in K(i)}^{b}\frac{1}{u_k^2}R^n(z)\mathbb{1}\Big(k=H(z)\Big)\\
&-(n+1)\sum_{i=1}^{I}\lambda_i^{(\epsilon)}v_i\sum_{k\in K(i)}^{nb}\frac{1}{u_k(u_k-d_{ik})}R^n(z)\mathbb{1}\left(k=H(z),k=H^{(i)}(z)\right)\\
&+(n+1)\sum_{s=1}^{S}\lambda_{i_s}^{(\epsilon)}\frac{1}{u_{k_s}\mu_{i_sk_s}}R^n(z)\mathbb{1}\left(k_s=L(z),k_s=H^{(i_s)}(z),k_s!bi_s\right)\\
&+(n+1)\sum_{k=1}^{K}\frac{1}{u_k}R^n(z)\mathbb{1}\Big(k=H(z)\Big)-(n+1)\sum_{k=1}^{K}\frac{1}{u_k}R^n(z)\mathbb{1}\Big(\sum_{i\in I(k)}z_{ik}=0,k=H(z)\Big)\\
&-(n+1)\sum_{k=1}^{K}\frac{1}{u_k}R^n(z)\mathbb{1}\Big(k=L(z)\Big)+(n+1)\sum_{k=1}^{K}\frac{1}{u_k}R^n(z)\mathbb{1}\Big(\sum_{i\in I(k)}z_{ik}=0,k=L(z)\Big)\\
&+LR^{(n)}
\end{aligned}
$$

where we denote the terms whose order is smaller than n as $LR^{(n)}$:

$$
\begin{aligned}
LR^{(n)} \triangleq &\sum_{i=1}^{I}\lambda_i^{(\epsilon)}\sum_{k\in K(i)}^{b}\sum_{\ell=2}^{n+1}\binom{n+1}{\ell}\left(-\frac{1}{u_k\mu_{ik}}\right)^{\ell}R^{n+1-\ell}(z)\mathbb{1}\Big(k=H(z)\Big)\\
&+\sum_{i=1}^{I}\lambda_i^{(\epsilon)}\sum_{k\in K(i)}^{nb}\sum_{\ell=2}^{n+1}\binom{n+1}{\ell}\left(-\frac{1}{u_k\mu_{ik}}\right)^{\ell}R^{n+1-\ell}(z)\mathbb{1}\left(k=H(z),k=H^{(i)}(z)\right)\\
&+\sum_{i=1}^{I}\lambda_i^{(\epsilon)}\sum_{k\in K(i)}\sum_{\ell=2}^{n+1}\binom{n+1}{\ell}\left(\frac{1}{u_k\mu_{ik}}\right)^{\ell}R^{n+1-\ell}(z)\mathbb{1}\left(k=L(z),k=H^{(i)}(z),k!bi\right)\\
&+\sum_{k=1}^{K}\sum_{i\in I(k)}\mu_{ik}\sum_{\ell=2}^{n+1}\binom{n+1}{\ell}\left(\frac{1}{u_k\mu_{ik}}\right)^{\ell}P_{ik}(z)R^{n+1-\ell}(z)\mathbb{1}\Big(k=H(z)\Big)\\
&+\sum_{k=1}^{K}\sum_{i\in I(k)}\mu_{ik}\sum_{\ell=2}^{n+1}\binom{n+1}{\ell}\left(-\frac{1}{u_k\mu_{ik}}\right)^{\ell}P_{ik}(z)R^{n+1-\ell}(z)\mathbb{1}\Big(k=L(z)\Big)\\
=&o\Big(R^n(z)\Big)
\end{aligned}
$$

since

(i)
$$
-\sum_{i=1}^{I}\lambda_i^{(\epsilon)}v_i\sum_{k\in K(i)}^{nb}\frac{1}{u_k(u_k-d_{ik})}R^n(z)\mathbb{1}\left(k=H(z),k=H^{(i)}(z)\right)\le 0
$$

(ii)

$$
\begin{aligned}
&- \sum_{k=1}^{K} \frac{1}{u_k} R^n(z) \mathbb{1}\Big(\sum_{i \in I(k)} z_{ik} = 0, k = H(z) \Big) \\
&= - \sum_{k=1}^{K} \frac{1}{u_k} \max_{k=1,...k} T_k^n(z) \mathbb{1}\Big(\sum_{i \in I(k)} z_{ik} = 0, k = H(z) \Big) \leq 0
\end{aligned}
$$

(iii)

$$
\sum_{k=1}^{K} \frac{1}{u_k} R(z) \mathbb{1} \left(\sum_{i \in I(k)} z_{ik} = 0, k = L(z) \right) = 0,
$$

then the generator becomes

$$
\begin{aligned}
Gf(z) \leq & -(n+1) \sum_{i=1}^{I} \lambda_i^{(\epsilon)} v_i \sum_{k \in K(i)}^{b} \frac{1}{u_k^2} R^n(z) \mathbb{1}\Big(k = H(z)\Big) \\
& + (n+1) \sum_{s=1}^{S} \lambda_{i_s}^{(\epsilon)} \frac{1}{u_{k_s} \mu_{i_s k_s}} R^n(z) \mathbb{1}\left(k_s = L(z), k_s = H^{(i_s)}(z), k_s ! b i_s\right) \\
& + (n+1) \sum_{k=1}^{K} \frac{1}{u_k} R^n(z) \mathbb{1}\Big(k = H(z)\Big) - (n+1) \sum_{k=1}^{K} \frac{1}{u_k} R^n(z) \mathbb{1}\Big(k = L(z)\Big) \\
& + L R^{(n)}(z)
\end{aligned}
\tag{1.10}
$$

Now we discuss the S terms:

$$
\lambda_{i_s}^{(\epsilon)} \frac{1}{u_{k_s} \mu_{i_s k_s}} R^n(z) \mathbb{1}\left(k_s = L(z), k_s = H^{(i_s)}(z), k_s ! b i_s\right), \quad s = 1, ...S
$$

For sth term: First, for the only basic activity for class i_s in the LP 1.3, we have $\lambda_{i_s} = \mu_{i_s k_s} x^*_{i_s k_s}$. Second, since all servers communicate, the corresponding server k_s must link other classes i through basic activities $x^*_{i k_s} > 0$. This means server k_s can only strictly spare partial efforts for class i_s, i.e. $x^*_{i_s k_s} < 1$. Hence $\lambda_{i_s} < \mu_{i_s k_s}$. Therefore

$$
\frac{\lambda_{i_s}}{u_{k_s} \mu_{i_s k_s}} < \frac{1}{u_{k_s}}
$$

$$
\frac{\lambda_{i_s}}{u_{k_s} \mu_{i_s k_s}} \mathbb{1}\left(k_s = L(z), k_s = H^{(i_s)}(z), k_s ! b i_s\right) < \frac{1}{u_{k_s}} \mathbb{1}\Big(k_s = L(z)\Big), \quad s = 1, ...S
$$

Then there exist $a_{k_s} > 0, s = 1, ...S$

$$
\begin{aligned}
&-\frac{1}{u_{k_s}} R^n(z) \mathbb{1}\Big(k_s = L(z)\Big) + \lambda_{i_s}^{(\epsilon)} \frac{1}{u_{k_s} \mu_{i_s k_s}} R^n(z) \mathbb{1}\left(k_s = L(z), k_s = H^{(i_s)}(z), k_s ! b i_s\right) \\
=&-\frac{1}{a_{k_s}} R^n(z) \mathbb{1}\Big(k_s = L(z)\Big)
\end{aligned}
$$

We further let $a_k \triangleq u_k, \forall k \neq k_1, ...k_S$, then we have $a_k > 0$, for each $k = 1, ...K$. Then (1.10) becomes

$$
\begin{aligned}
Gf(z) \leq &-(n+1) \sum_{k=1}^{K} \left[\frac{\sum_{i \in I(k)}^{b} \lambda_i^{(\epsilon)} v_i}{u_k^2} - \frac{1}{u_k} \right] R^n(z) \mathbb{1}\Big(k = H(z)\Big) \\
&-(n+1) \sum_{k=1}^{K} \frac{1}{a_k} R^n(z) \mathbb{1}\Big(k = L(z)\Big) + L R^{(n)}(z)
\end{aligned}
$$

Now we add some terms and minus some terms, the generator becomes

$$
\begin{aligned}
Gf(z) \leq &-(n+1) \sum_{k=1}^{K} \left[\frac{\sum_{i \in I(k)}^{b} \lambda_i v_i}{u_k^2} - \frac{2}{u_k} + \frac{1}{\sum_{i \in I(k)}^{b} \lambda_i v_i} \right] R^n(z) \mathbb{1}\Big(k = H(z)\Big) \\
&-(n+1) \sum_{k=1}^{K} \frac{1}{a_k} R^n(z) \mathbb{1}\Big(k = L(z)\Big) + L R^{(n)}(z) \\
&+(n+1) \sum_{k=1}^{K} \left[\frac{1}{\sum_{i \in I(k)}^{b} \lambda_i v_i} - \frac{1}{u_k} \right] R^n(z) \mathbb{1}\Big(k = H(z)\Big) \\
&+\epsilon(n+1) \sum_{k=1}^{K} \sum_{i \in I(k)}^{b} \lambda_i v_i \frac{1}{u_k^2} R^n(z) \mathbb{1}\Big(k = H(z)\Big)
\end{aligned}
$$

Since

$$
\begin{aligned}
&\left[\frac{\sum_{i \in I(k)}^{b} \lambda_i v_i}{u_k^2} - \frac{2}{u_k} + \frac{1}{\sum_{i \in I(k)}^{b} \lambda_i v_i} \right] \\
&= \left[\frac{\sqrt{\sum_{i \in I(k)}^{b} \lambda_i v_i}}{u_k} - \frac{1}{\sqrt{\sum_{i \in I(k)}^{b} \lambda_i v_i}} \right]^2 \geq 0
\end{aligned}
$$

then with Lemma 9, the generator becomes:

$$\begin{aligned}
Gf(z) \leq & -(n+1)\sum_{k=1}^{K}\frac{1}{a_k}R^n(z)\mathbb{1}\Big(k=L(z)\Big)+LR^{(n)}(z)\\
& +(n+1)\sum_{k=1}^{K}\left[\frac{1}{\sum_{i\in I(k)}^{b}\lambda_i v_i}-\frac{1}{u_k}\right]R^n(z)\mathbb{1}\Big(k=H(z)\Big)\\
& +\epsilon(n+1)\sum_{k=1}^{K}\sum_{i\in I(k)}^{b}\lambda_i v_i\frac{1}{u_k^2}R^n(z)\mathbb{1}\Big(k=H(z)\Big)\\
\leq & -\frac{n+1}{\bar{a}}\sum_{k=1}^{K}R^n(z)\mathbb{1}\left(k=L(z)\right)+LR^{(n)}(z)\\
& +\epsilon\frac{n+1}{\underline{u}^2}\sum_{i=1}^{I}\lambda_i v_i\sum_{k=1}^{K}R^n(z)\mathbb{1}\Big(k=H(z)\Big)\\
= & -(n+1)\Big(\frac{1}{\bar{a}}-\epsilon\frac{1}{\underline{u}}\Big)R^n(z)+LR^{(n)}(z)
\end{aligned} \tag{1.11}$$

where $\bar{a}=\max_{1,\dots K}a_k, \underline{u}=\min_{1,\dots K}u_k$. In the following, we will apply the Lemma 4(a) and utilize the Induction procedure to conclude our proof. Denote $f(z)=R^n(z), \forall n\in\mathbb{N}^+$, then by Lemma 4(a), we have

$$\mathbb{E}\Big[Gf\Big(Z^{(\epsilon)}(\infty)\Big)\Big]=0.$$

Now taking expectation on both sides of (1.11), we have

$$0\leq -(n+1)\Big(\frac{1}{\bar{a}}-\epsilon\frac{1}{\underline{u}}\Big)\mathbb{E}\Big[R^n\Big(Z^{(\epsilon)}(\infty)\Big)\Big]+\mathbb{E}\Big[LR^{(n)}\Big(Z^{(\epsilon)}(\infty)\Big)\Big]$$

When $n=1$,

$$\begin{aligned}
LR^{(1)}(z)= & \sum_{i=1}^{I}\lambda_i^{(\epsilon)}\sum_{k\in K(i)}^{b}\frac{1}{(u_k\mu_{ik})^2}\mathbb{1}\Big(k=H(z)\Big)+\sum_{i=1}^{I}\lambda_i^{(\epsilon)}\sum_{k\in K(i)}^{nb}\frac{1}{(u_k\mu_{ik})^2}\mathbb{1}\left(k=H(z),k=H^{(i)}(z)\right)\\
& +\sum_{s=1}^{S}\lambda_{i_s}^{(\epsilon)}\frac{1}{(u_{k_s}\mu_{i_sk_s})^2}\mathbb{1}\left(k_s=L(z),k_s=H^{(i_s)}(z),k_s!bi_s\right)\\
& +\sum_{k=1}^{K}\sum_{i\in I(k)}\mu_{ik}P_{ik}(z)\frac{1}{(u_k\mu_{ik})^2}\mathbb{1}\Big(k=H(z)\Big)+\sum_{k=1}^{K}\sum_{i\in I(k)}\mu_{ik}P_{ik}(z)\frac{1}{(u_k\mu_{ik})^2}\mathbb{1}\Big(k=L(z)\Big)\\
\leq & \sum_{i=1}^{I}\lambda_i\sum_{k\in K(i)}\frac{2}{(u_k\mu_{ik})^2}+\sum_{k=1}^{K}\sum_{i\in I(k)}\frac{2}{(u_k^2\mu_{ik})}\triangleq M_0
\end{aligned}$$

There exists ϵ_0 such that $\left(\frac{1}{\bar{a}} - \epsilon\frac{1}{\underline{u}}\right) > 0$ for any ϵ satisfying $0 < \epsilon < \epsilon_0$, we have

$$E\left[R\left(Z^{(\epsilon)}(\infty)\right)\right] \leq \frac{\mathbb{E}\left[LR^{(1)}\left(Z^{(\epsilon)}(\infty)\right)\right]}{2\left(\frac{1}{\bar{a}} - \epsilon\frac{1}{\underline{u}}\right)} \leq \frac{M_0}{2\left(\frac{1}{\bar{a}} - \epsilon\frac{1}{\underline{u}}\right)} \triangleq M_1$$

Then by Induction, suppose $E\left[R^{\ell}\left(Z^{(\epsilon)}(\infty)\right)\right] \leq M_\ell$, $1 \leq \ell \leq n$, we therefore have

$$E\left[R^{n+1}\left(Z^{(\epsilon)}(\infty)\right)\right] \leq \frac{\mathbb{E}\left[LR^{(n)}\left(Z^{(\epsilon)}(\infty)\right)\right]}{(n+1)\left(\frac{1}{\bar{a}} - \epsilon\frac{1}{\underline{u}}\right)} \triangleq M_{n+1}$$

To conclude, by Induction, we have $\forall n \in \mathbb{N}^+$, when $0 < \epsilon < \epsilon_0$,

$$\mathbb{E}\left(\max_{k\in\{1,\dots K\}} T_k(Z^{(\epsilon)}) - \min_{k\in\{1,\dots K\}} T_k(Z^{(\epsilon)})\right)^n \leq M_n$$

□

1.7 Preliminary Results II

In this section, all the results utilize the State-space Collapse Theorem 8 with up to the second order moment boundedness.

Lemma 10. *Under Assumptions 1 & 2, with the WWTA and scheduling $\mathcal{P}$,*

$$\mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'} W_{k'}\left(Z^{(\epsilon)}(\infty)\right)\right)\mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)}(\infty) = 0\right)\right] = O(\epsilon^{1/2})$$

Proof. We omit (∞) in $Z^{(\epsilon)}(\infty)$ for simplicity.

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'} W_{k'}(Z^{(\epsilon)})\right)\mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
=&\mathbb{E}\left[\left|\sum_{k'=1}^{K} u_{k'} \sum_{i\in I(k')} \frac{1}{\mu_{ik'}} Z_{ik'}^{(\epsilon)} - \sum_{k'=1}^{K} u_{k'}^2 \frac{1}{u_k} \sum_{i\in I(k)} \frac{1}{\mu_{ik}} Z_{ik}^{(\epsilon)}\right| \mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
\overset{(a)}{\leq}&\mathbb{E}\left[\left[\max_{k\in\{1,\dots K\}} T_k(Z^{(\epsilon)}) - \min_{k\in\{1,\dots K\}} T_k(Z^{(\epsilon)})\right] \mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
\overset{(b)}{\leq}&\mathbb{E}\left[\left(\max_{k\in\{1,\dots K\}} T_k(Z^{(\epsilon)}) - \min_{k\in\{1,\dots K\}} T_k(Z^{(\epsilon)})\right)^2\right]^{1/2} \mathbb{P}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)^{1/2} \\
\overset{(c)}{\leq}&M_2^{1/2}\mathbb{P}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)^{1/2} \overset{(d)}{=} O(\epsilon^{1/2})
\end{aligned}$$

where (a) is by Lemma 1, (ii) and Lemma 2, (ii), (b) is by Cauchy-Schwarz Inequality, (c) is by Theorem 8, and (d) is by Lemma 5. □

Lemma 11 (Negligibility for non-basic activities)**.** *Under Assumptions 1 & 2, with the WWTA and scheduling $\mathcal{P}$, for any non-basic activity (i,k) s.t. $d_{ik} > 0$, we have*

$$\mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'} W_{k'}\left(Z^{(\epsilon)}(\infty)\right)\right)\mathbb{1}\left(k = H^{(i)}\left(Z^{(\epsilon)}(\infty)\right)\right)\right] = O(\epsilon^{1/2}) \tag{1.12}$$

$$\mathbb{E}\left[P_{ik}\left(Z^{(\epsilon)}(\infty)\right)\left(\sum_{k'=1}^{K} u_{k'} W_{k'}\left(Z^{(\epsilon)}(\infty)\right)\right)\right] = O(\epsilon^{1/2}) \tag{1.13}$$

Besides, the scaled first moment of sum of non-basic activities is also negligible:

$$\epsilon\mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}(\infty)\right)\right] = O(\epsilon^{1/2}) \tag{1.14}$$

The proof of Lemma 11 is provided in Appendix A.2.1.

1.8 Proofs for Theorem 3

In this section, we first introduce the following Theorem 12, and Corollary 13. Then the presentation for this section is planned as following: we introduce two

Lemmas in Sections 1.8.1 and 1.8.2 which provide crucial ingredients in proving Theorem 12, In Section 1.8.3 we prove Theorem 12, then followed by the proof of main result Theorem 3 in Section 1.8.4.

Theorem 12 (Limit distribution)**.** *Consider a parallel-server system that satisfies Assumptions 1 & 2. Under the WWTA policy and scheduling $\mathcal{P}$, for each $\theta \leq 0$,*

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}(\infty)\right)}\right] = \frac{1}{1-\theta\sum_{i=1}^{I}\lambda_i v_i^2}$$

that is, the limit is the Laplace transform of an exponential random variable with mean

$$m = \sum_{i=1}^{I} \lambda_i v_i^2$$

Therefore, the scaled sum of queue length, weighted by optimal dual solution, converges in distribution to an exponential random variable:

$$\epsilon\Big(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}(\infty)\Big) \xrightarrow{d} \tilde{X} \sim Exponential(1/m), \quad as\ \epsilon \downarrow 0$$

Corollary 13 (Workload Version of Limit Distribution)**.** *Under the same conditions of Theorem 12, the scaled sum of workload, weighted by optimal dual solution, converges in distribution to random variable $\tilde{X}$, i.e.*

$$\epsilon\Big(\sum_{k=1}^{K} u_k W_k\Big(Z^{(\epsilon)}(\infty)\Big)\Big) \xrightarrow{d} \tilde{X}, \quad as\ \epsilon \downarrow 0$$

where $\tilde{X} \sim Exponential(1/m)$.

Remark 5. *Corollary 13 is equivalent to the Theorem 12, if all the activities are basic or having $d_{ik} = 0$ as in the (1.7), since by Lemma 1,*

$$\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}(\infty) = \sum_{k=1}^{K} u_k \sum_{i\in I(k)} \frac{1}{\mu_{ik}} Z_{ik}^{(\epsilon)}(\infty) = \sum_{k=1}^{K} u_k W_k(Z^{(\epsilon)}(\infty))$$

However, if there exists non-basic activities which have $d_{ik} > 0$, then

$$\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}(\infty) = \sum_{k=1}^{K}\sum_{i\in I(k)} \frac{u_k - d_{ik}}{\mu_{ik}} Z_{ik}^{(\epsilon)}(\infty) = \sum_{k=1}^{K} u_k W_k\Big(Z^{(\epsilon)}(\infty)\Big) - \sum_{k=1}^{K}\sum_{i\in I(k)} \frac{d_{ik}}{\mu_{ik}} Z_{ik}^{(\epsilon)}(\infty) \tag{1.15}$$

where the additional term due to non-basic activities($d_{ik} > 0$) makes the equivalence nontrivial.

The proof of Corollary 13 can be found in Appendix A.3.1.

1.8.1 Ingredient I for Theorem 12

Lemma 14. *Under the condition of Theorem 12, for each $k = 1, ...K$,*

$$\begin{aligned}
&\sum_{k=1}^{K} u_k \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}(\infty)\right)} \mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)}(\infty) = 0 \right)\right] \\
&+ \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)}(\infty)) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}(\infty)\right)}\right] \\
&= \sum_{k=1}^{K} u_k \mathbb{P}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)}(\infty) = 0 \right) + \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}\left(Z^{(\epsilon)}(\infty)\right)\right] + O(\epsilon^{3/2})
\end{aligned}$$

where d_{ik} comes from with notation (1.7).

Proof. We omit (∞) in $Z^{(\epsilon)}(\infty)$ for simplicity. Denote the residual difference

$$\begin{aligned}
\Delta \triangleq &\sum_{k=1}^{K} u_k \mathbb{P}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0 \right) + \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\right] \\
&- \sum_{k=1}^{K} u_k \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0 \right)\right] \\
&- \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right]
\end{aligned}$$

Then we will prove $\Delta = O(\epsilon^{3/2})$:

$$
\begin{aligned}
\Delta = & \sum_{k=1}^{K} u_k \mathbb{E}\left[\left(1 - e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right) \mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
& + \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(1 - e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right)\right] \\
\overset{(a)}{\leq} & \epsilon|\theta| \sum_{k=1}^{K} u_k \mathbb{E}\left[\left(\sum_{k=1}^{K} u_k \sum_{i\in I(k)} \frac{1}{\mu_{ik}} Z_{ik}^{(\epsilon)}\right) \mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
& - \epsilon|\theta| \sum_{k=1}^{K} u_k \mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)} \frac{d_{ik}}{\mu_{ik}} Z_{ik}^{(\epsilon)}\right) \mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
& + \epsilon|\theta| \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{k=1}^{K} u_k \sum_{i\in I(k)} \frac{1}{\mu_{ik}} Z_{ik}^{(\epsilon)}\right)\right] \\
& - \epsilon|\theta| \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{i=1}^{I}\sum_{k\in K(i)} \frac{d_{ik}}{\mu_{ik}} Z_{ik}^{(\epsilon)}\right)\right] \\
\leq & \epsilon|\theta| \sum_{k=1}^{K} u_k \mathbb{E}\left[\left(\sum_{k=1}^{K} u_k W_k(Z^{(\epsilon)})\right) \mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
& + \epsilon|\theta| \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{k=1}^{K} u_k W_k(Z^{(\epsilon)})\right)\right] \\
\overset{(b)}{\leq} & O(\epsilon^{3/2})
\end{aligned}
$$

where (a) is by Lemma 7, (b) is by Lemma 10 and Lemma 11, where either $d_{ik} = 0$ or $d_{ik} > 0$ for non-basic activities. □

1.8.2 Ingredient II for Theorem 12

Lemma 15. *Under the condition of Theorem 12, for each $\theta \leq 0$ and each $i = 1, ...I$,*

$$
\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\left(\lambda_i v_i - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)}(\infty))\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}(\infty)\right)}\right] = 0 \tag{1.16}
$$

Remark 6. *Equivalently, lemma 15 reflects the flow balance:*

$$
\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\left(\lambda_i - \sum_{k\in K(i)} \mu_{ik} P_{ik}(Z^{(\epsilon)}(\infty))\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}(\infty)\right)}\right] = 0 \tag{1.17}
$$

The equivalence of (1.17) *is straightforward because of* (1.7) *and Lemma* 5 *with* $\theta \leq 0$ *that*

$$d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)}(\infty))e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)}v_i Z_{ik}^{(\epsilon)}(\infty)\right)}\right] \leq d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)}(\infty))\right] = O(\epsilon) \tag{1.18}$$

Proof. Throughout the proof, we omit (∞) in $Z^{(\epsilon)}(\infty)$ for simplicity. Part (1) is trivial case with $\theta = 0$. In part (2) with $\theta < 0$, we will first show the following

$$\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\left(\lambda_i v_i - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)})\right)e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)}v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')}v_{i'}Z_{i'k}^{(\epsilon)}\right)}\right] = 0$$

is true for $t > 0$, then we use Moore-Osgood Theorem([16]) to perform the interchange of limit to prove the case with $t = 0$, which is (1.16) that we intend to prove.

(1) Trivial case: $\theta = 0$. We let $f(z) = \sum_{k\in K(i)} z_{ik}$, $\forall i \in I$, then the generator (1.6) becomes

$$\begin{aligned} Gf(z) =& \lambda_i^{(\epsilon)} \sum_{k\in K(i)} 1\left(k = H^{(i)}(z)\right) - \sum_{k\in K(i)} \mu_{ik}P_{ik}(z) \\ =& \lambda_i^{(\epsilon)} - \sum_{k\in K(i)} \mu_{ik}P_{ik}(z) \end{aligned}$$

by Lemma 4(a), we have

$$\sum_{k\in K(i)} \mu_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\right] = \lambda_i^{(\epsilon)}$$

taking limit and multiply v_i on both sides, by (1.7), we have

$$\lambda_i v_i = \lim_{\epsilon\downarrow 0}\sum_{k\in K(i)}(u_k - d_{ik})\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\right] \overset{(a)}{=} \lim_{\epsilon\downarrow 0}\sum_{k\in K(i)} u_k\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\right]$$

where (a) is by Lemma 5.

(2) When $\theta < 0$, we let $f(z) = e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)}c_i z_{ik}\right)}, \theta < 0, c_i \geq 0$, then with (1.7), the

generator (1.6) becomes

$$\begin{aligned}
Gf(z) &= \sum_{i=1}^{I} \lambda_i^{(\epsilon)} \sum_{k\in K(i)} \left(e^{c_i\epsilon\theta} - 1\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} \mathbb{1}\left(k = H^{(i)}(z)\right) \\
&\quad + \sum_{k=1}^{K} \sum_{i\in I(k)} \mu_{ik} P_{ik}(z) \left(e^{-c_i\epsilon\theta} - 1\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} \\
&= \sum_{i=1}^{I} \lambda_i^{(\epsilon)} \left(e^{c_i\epsilon\theta} - 1\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} \\
&\quad + \sum_{k=1}^{K} \sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i} P_{ik}(z) \left(e^{-c_i\epsilon\theta} - 1\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)}
\end{aligned}$$

By second order Taylor expansion,

$$e^{c_i\epsilon\theta} = 1 + c_i\epsilon\theta + \frac{1}{2}c_i^2\epsilon^2\theta^2 + O(\epsilon^3)$$

then since $0 \le f(z) \le 1$, we have

$$\begin{aligned}
Gf(z) &= \sum_{i=1}^{I} \lambda_i \Big(c_i\epsilon\theta + \frac{1}{2}c_i^2\epsilon^2\theta^2\Big) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} \\
&\quad - \sum_{i=1}^{I} \lambda_i \Big(c_i\epsilon^2\theta + \frac{1}{2}c_i^2\epsilon^3\theta^2\Big) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} \\
&\quad + \sum_{k=1}^{K} \sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i} P_{ik}(z) \Big(- c_i\epsilon\theta + \frac{1}{2}c_i^2\epsilon^2\theta^2\Big) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} + O(\epsilon^3) \\
&= \sum_{i=1}^{I} \lambda_i c_i\epsilon\theta e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} \\
&\quad + \sum_{k=1}^{K} \sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i} (-c_i\epsilon\theta) P_{ik}(z) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} \\
&\quad - \sum_{i=1}^{I} \lambda_i c_i\epsilon^2\theta e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} + \frac{1}{2}\sum_{i=1}^{I} \lambda_i c_i^2\epsilon^2\theta^2 e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} \\
&\quad + \frac{1}{2}\sum_{k=1}^{K} \sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i} (c_i^2\epsilon^2\theta^2) P_{ik}(z) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i z_{ik}\right)} + O(\epsilon^3)
\end{aligned}$$

by Lemma 4(a), we have

$$
\begin{aligned}
&\sum_{i=1}^{I} \lambda_i c_i \epsilon\theta \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i Z_{ik}^{(\epsilon)}\right)}\right] + O(\epsilon^3) \\
&+ \sum_{k=1}^{K}\sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i}(-c_i\epsilon\theta)\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&- \sum_{i=1}^{I} \lambda_i c_i \epsilon^2\theta \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i Z_{ik}^{(\epsilon)}\right)}\right] + \frac{1}{2}\sum_{i=1}^{I} \lambda_i c_i^2 \epsilon^2\theta^2 \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&+ \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i}(c_i^2\epsilon^2\theta^2)\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i Z_{ik}^{(\epsilon)}\right)}\right] = 0
\end{aligned}
\tag{1.19}
$$

Now rewrite (1.19) by only considering the lower order terms w.r.t ϵ:

$$
\begin{aligned}
&\sum_{i=1}^{I} \lambda_i c_i \epsilon\theta \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i Z_{ik}^{(\epsilon)}\right)}\right] + O(\epsilon^2) \\
&+ \sum_{k=1}^{K}\sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i}(-c_i\epsilon\theta)\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} c_i Z_{ik}^{(\epsilon)}\right)}\right] = 0
\end{aligned}
$$

Let $c_{i'} = v_{i'}(1+t), c_i = v_i, i' \neq i = 1, \ldots I, t \geq 0$, and divide $\epsilon\theta$ on both sides:

$$
\begin{aligned}
&\lambda_{i'} v_{i'} t \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)}\right] \\
&+ \sum_{i=1}^{I} \lambda_i v_i \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)}\right] \\
&- \sum_{k\in K(i')} (u_k - d_{ik}) t \mathbb{E}\left[P_{i'k}(Z^{(\epsilon)})e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)}\right] \\
&- \sum_{k=1}^{K} u_k \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)}\right] \\
&+ \sum_{k=1}^{K} u_k \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)} \mathbb{1}\Big(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\Big)\right] \\
&+ \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik} \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)}\right] \\
&= \lambda_{i'} v_{i'} t \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)}\right] \\
&- \sum_{k\in K(i')} u_k t \mathbb{E}\left[P_{i'k}(Z^{(\epsilon)})e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)}\right] + O(\epsilon) \\
&= 0
\end{aligned}
$$

where we use Lemma 1 (i) and (ii) that

$$\sum_{i=1}^{I} \lambda_i v_i = 1, \quad \sum_{k=1}^{K} u_k = 1$$

and Lemma 5 with following:

$$\mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)} \mathbb{1}\Big(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0 \Big) \right] \leq \mathbb{P}\Big(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0 \Big) = O(\epsilon)$$

$$\sum_{k=1}^{K} \sum_{i\in I(k)} d_{ik} \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)} \right]$$

$$\leq \sum_{k=1}^{K} \sum_{i\in I(k)} d_{ik} \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) \right] = O(\epsilon)$$

That is, we have

$$\begin{aligned} &\lambda_{i'} v_{i'} t \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)} \right] \\ &- \sum_{k\in K(i')} u_k t \mathbb{E}\left[P_{i'k}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)} \right] \\ &= O(\epsilon) \end{aligned} \tag{1.20}$$

There are two cases w.r.t $t \geq 0$:

(a) If $t > 0$, divide t on both sides of (1.20), and take $\epsilon \downarrow 0$. For any $t > 0$, we have $\forall i = 1, ...I$,

$$\lim_{\epsilon\downarrow 0} \mathbb{E}\left[\Big(\lambda_i v_i - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)}) \Big) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i')} v_{i'} Z_{i'k}^{(\epsilon)}\right)} \right] = 0 \tag{1.21}$$

(b) When $t = 0$, we need to prove the following lemma:

Lemma 16. *Suppose (1.21) holds for $t > 0$. Then it also holds when t=0.*

We put the proof of Lemma 16 in Appendix A.3.2, where we use Moore-Osgood Theorem([16]) to perform the interchange of limit.

Therefore, for each $i = 1, ...I$ and each $\theta \leq 0$, we have

$$\lim_{\epsilon\downarrow 0} \mathbb{E}\left[\Big(\lambda_i v_i - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)}) \Big) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \right] = 0$$

□

1.8.3 Proof of Theorem 12

Proof. Let $f(z) = e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)}, \theta \le 0$, then the generator (1.6) becomes

$$\begin{aligned}
Gf(z) &= \sum_{i=1}^{I} \lambda_i^{(\epsilon)} \sum_{k\in K(i)} \left(e^{v_i\epsilon\theta} - 1\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} \mathbb{1}\left(k = H^{(i)}(z)\right) \\
&\quad + \sum_{k=1}^{K} \sum_{i\in I(k)} \mu_{ik} P_{ik}(z) \left(e^{v_i\epsilon\theta} - 1\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} \\
&= \sum_{i=1}^{I} \lambda_i^{(\epsilon)} \left(e^{v_i\epsilon\theta} - 1\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} \\
&\quad + \sum_{k=1}^{K} \sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i} P_{ik}(z) \left(e^{v_i\epsilon\theta} - 1\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)}
\end{aligned}$$

By second order Taylor expansion,

$$e^{v_i\epsilon\theta} = 1 + v_i\epsilon\theta + \frac{1}{2} v_i^2 \epsilon^2\theta^2 + O(\epsilon^3)$$

then since $0 \le f(z) \le 1$, we have

$$\begin{aligned}
Gf(z) &= \sum_{i=1}^{I} \lambda_i \left(v_i\epsilon\theta + \frac{1}{2} v_i^2\epsilon^2\theta^2\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} \\
&\quad - \sum_{i=1}^{I} \lambda_i \left(v_i\epsilon^2\theta + \frac{1}{2} v_i^2\epsilon^3\theta^2\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} \\
&\quad + \sum_{k=1}^{K} \sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i} P_{ik}(z) \left(v_i\epsilon\theta + \frac{1}{2} v_i^2\epsilon^2\theta^2\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} + O(\epsilon^3) \\
&= \sum_{i=1}^{I} \lambda_i v_i \epsilon\theta e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} \\
&\quad + \sum_{k=1}^{K} \sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i} v_i \epsilon\theta P_{ik}(z) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} \\
&\quad - \sum_{i=1}^{I} \lambda_i v_i \epsilon^2\theta e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} + \frac{1}{2} \sum_{i=1}^{I} \lambda_i v_i^2 \epsilon^2\theta^2 e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} \\
&\quad + \frac{1}{2} \sum_{k=1}^{K} \sum_{i\in I(k)} \frac{u_k - d_{ik}}{v_i} (v_i^2\epsilon^2\theta^2) P_{ik}(z) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\right)} + O(\epsilon^3)
\end{aligned}$$

by Lemma 4(a), we have

$$\begin{aligned}
&\sum_{k=1}^{K} u_k \epsilon\theta \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
&+ \epsilon\theta \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&- \sum_{i=1}^{I} \lambda_i v_i \epsilon^2\theta \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&+ \frac{1}{2}\sum_{i=1}^{I} \lambda_i v_i^2 \epsilon^2\theta^2 \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&+ \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in I(k)} u_k v_i \epsilon^2\theta^2 \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] + O(\epsilon^3) = 0
\end{aligned}$$

By Lemma 5, Lemma 14, and Lemma 1, we divide $\epsilon\theta$ on both sides and have

$$\begin{aligned}
&\sum_{k=1}^{K} u_k \mathbb{P}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right) + \sum_{k=1}^{K}\sum_{i\in I(k)} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\right] + O(\epsilon^2) \\
&- \sum_{i=1}^{I} \lambda_i v_i \epsilon \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] + \frac{1}{2}\sum_{i=1}^{I} \lambda_i v_i^2 \epsilon\theta \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&+ \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in I(k)} u_k v_i \epsilon\theta \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&= \epsilon + O(\epsilon^2) \\
&- \epsilon\mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] + \sum_{i=1}^{I} \lambda_i v_i^2 \epsilon\theta \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&+ \frac{1}{2}\sum_{i=1}^{I} v_i \epsilon\theta \mathbb{E}\left[\left(\sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)}) - \lambda_i v_i\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&= 0
\end{aligned}$$

Dividing ϵ on both sides, taking limit w.r.t ϵ, and plugging in the result of Lemma 15, we have

$$\left(1 - \theta\sum_{i=1}^{I} \lambda_i v_i^2\right) \lim_{\epsilon\downarrow 0} \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] = 1$$

i.e.

$$\lim_{\epsilon\downarrow 0} \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] = \frac{1}{1 - \theta\sum_{i=1}^{I} \lambda_i v_i^2}$$

the limit is the Laplace transform of an exponential random variable with mean

$$m = \sum_{i=1}^{I} \lambda_i v_i^2$$

Let $\tilde{X}$ be a random variable such that $\tilde{X} \sim Exponential(1/m)$. Therefore,

$$\epsilon \left(\sum_{i=1}^{I} \sum_{k \in K(i)} v_i Z_{ik}^{(\epsilon)} \right) \xrightarrow{d} \tilde{X}, \quad as\ \epsilon \downarrow 0$$

□

1.8.4 Proof of Theorem 3

Proof. Denote

$$\tilde{X}^{(\epsilon)} \triangleq \epsilon \Big(\sum_{k=1}^{K} u_k W_k(Z^{(\epsilon)}) \Big) = \epsilon \Big(\sum_{k=1}^{K} u_k^2 T_k(Z^{(\epsilon)}) \Big)$$

then as shown by Corollary 13, and recall that $e = (1, ...1)^T$, we have

$$\tilde{X}^{(\epsilon)} e \xrightarrow{d} \tilde{X} e, \quad as\ \epsilon \downarrow 0 \tag{1.22}$$

Now let

$$Y_{k'}^{(\epsilon)} = \epsilon \sum_{k=1}^{K} \frac{u_k^2}{u_{k'}} W_{k'}(Z^{(\epsilon)}) = \epsilon \sum_{k=1}^{K} u_k^2 T_{k'}(Z^{(\epsilon)}) \quad k' = 1, ...K$$

Denote $Y^{(\epsilon)} = (Y_1^{(\epsilon)}, ..., Y_K^{(\epsilon)})^T$, , then

$$\begin{aligned}
&\left| \tilde{X}^{(\epsilon)} e - Y^{(\epsilon)} \right| \\
=& \sum_{k'=1}^{K} \left| \tilde{X}^{(\epsilon)} - Y_{k'}^{(\epsilon)} \right| \\
=& \sum_{k'=1}^{K} \epsilon \left[\sum_{k=1}^{K} u_k^2 \left| T_k(Z^{(\epsilon)}) - T_{k'}(Z^{(\epsilon)}) \right| \right] \\
\leq& \sum_{k'=1}^{K} \left[\epsilon K \left(\max_{k \in \{1,...K\}} T_k(Z^{(\epsilon)}) - \min_{k \in \{1,...K\}} T_k(Z^{(\epsilon)}) \right) \right] \\
=& \epsilon K^2 \left(\max_{k \in \{1,...K\}} T_k(Z^{(\epsilon)}) - \min_{k \in \{1,...K\}} T_k(Z^{(\epsilon)}) \right)
\end{aligned}$$

since $0 < u_k \leq 1$. By Theorem 8, $\limsup_{\epsilon \downarrow 0} \mathbb{E}\left[\max_{k\in\{1,...K\}} T_k(Z^{(\epsilon)}) - \min_{k\in\{1,...K\}} T_k(Z^{(\epsilon)})\right] \leq M_1$, then we have

$$\left|\tilde{X}^{(\epsilon)} e - Y^{(\epsilon)}\right| \xrightarrow{L^1} 0, \quad as\ \epsilon \downarrow 0$$

By Markov Inequality,

$$\lim_{\epsilon\downarrow 0} \mathbb{P}\left(\left|\tilde{X}^{(\epsilon)} - Y_{k'}^{(\epsilon)}\right| \geq a\right) \leq \lim_{\epsilon\downarrow 0} \frac{\mathbb{E}\left(\left|\tilde{X}^{(\epsilon)} - Y_{k'}^{(\epsilon)}\right|\right)}{a} = 0, \forall a > 0$$

then

$$\left|\tilde{X}^{(\epsilon)} e - Y^{(\epsilon)}\right| \xrightarrow{p} 0, \quad as\ \epsilon \downarrow 0 \tag{1.23}$$

Combining (1.22) with (1.23), by [3] (Theorem 3.1), we have

$$Y^{(\epsilon)} \xrightarrow{d} \tilde{X} e, \quad as\ \epsilon \downarrow 0$$

i.e.

$$\epsilon\left(\sum_{k=1}^{K} u_k^2\right)\left(\frac{1}{u_1} W_1(Z^{(\epsilon)}), ... \frac{1}{u_K} W_K(Z^{(\epsilon)})\right)^T \xrightarrow{d} \tilde{X} e$$

where $\tilde{X} \sim Exponential\left(\frac{1}{\sum_{i=1}^{I} \lambda_i v_i^2}\right)$. Then by scaling property of exponential distribution, we have

$$\epsilon\left(\frac{1}{u_1} W_1(Z^{(\epsilon)}), ... \frac{1}{u_K} W_K(Z^{(\epsilon)})\right)^T \xrightarrow{d}_{\Delta} X$$

where $X \sim Exponential\left(\frac{\sum_{k=1}^{K} u_k^2}{\sum_{i=1}^{I} \lambda_i v_i^2}\right)$. Therefore,

$$\lim_{\epsilon\downarrow 0} \epsilon\left(W_1(Z^{(\epsilon)}), ..., W_K(Z^{(\epsilon)})\right) \xrightarrow{d} (u_1, ..., u_K)\, X$$

□

CHAPTER 2

MULTI-SCALE HEAVY TRAFFIC LIMITS FOR GENERALIZED JACKSON NETWORKS

2.1 Introduction

[24] pioneered the study of a class of queueing networks, known as open Jackson networks. The defining characteristics of a Jackson network are (a) all customers visiting a service station are homogeneous in terms of service time distribution and the routing probabilities, and (b) all interarrival and service time distributions are exponential. For a Jackson network, the queue length process is a continuous time Markov chain (CTMC). Jackson's pioneered contribution is that when the traffic intensity at each station is less than one, the CTMC has a unique product-form stationary distribution, meaning that the steady-state queue lengths at various stations in the queueing network are independent. The product-form result makes the computation of steady-state performance measures scalable with respect to the number of stations in the network.

When interarrival and service times distributions are general, the corresponding network is known as a generalized Jackson network. For a generalized Jackson network, the product-form result no longer holds in general. Creating and justifying approximations of generalized Jackson networks have been an active research area for more than 50 years; see, for example, [31] and Whitt and You (2021), https://doi.org/10.1002/nav.22010. [27] and [25] proved a functional central limit theorem, stating that the scaled queue length process converges to a multi-dimensional reflecting Brownian motion (RBM) under a heavy traffic condition. The class of RBMs was first developed

in [21]. The stationary distribution of the RBM is shown to exist and is characterized by a basic adjoint relationship (BAR) in [17]. However, the stationary distribution is not of product form in general [18]. By developing an elaborate limit-interchange procedure, [14] justified that the stationary distribution of a generalized Jackson network is well approximated by that of the corresponding RBM under the same traffic condition. Numerical algorithms exist to compute the stationary distribution of an RBM in low dimension; see [10, 28]. These algorithms are not scalable with respect to the number of stations. A sequential bottleneck decomposition (SBD) method was proposed in [12].

In this paper, we prove that under a multi-scale heavy traffic condition, the stationary distribution of the multi-scaled queue length process in any generalized Jackson network has a product-form limit. Each component in the product-form has an exponential distribution, corresponding to the Brownian approximation of a single station queue. The "single station" can be constructed precisely and has a good intuitive interpretation, consistent with the bottleneck analysis advanced in [9].

Our proof critically relies on the BAR-approach recently developed in [6] for queueing networks with general interarrival and service time distributions. The BAR-approach provides an alternative to the limit interchange procedure in [14] and [8]. [7] further demonstrates the superiority of employing the BAR-approach for multiclass queueing networks under priority service disciplines. Here, we demonstrates that the BAR-approach is a natural approach to discover and justify the type of results in this paper.

2.2 Generalized Jackson Network

We first define a generalized Jackson network, following closely Section 2.1 of [6] both in terms of terminologies and notations. There is a single class of jobs arriving to the network which has J stations, and will exit the network in finite time with probability one. All the jobs are homogeneous in terms of service time and routing. The service in each station follows the first-come-first-serve (FCFS) discipline. The external interarrival time, service time, and routing decisions are assume to be independent, each of which is assumed to follow an i.i.d sequence of random variables. Let $\mathcal{J} = \{1, \ldots, J\}$ be the set of stations. We consider a discrete time Markov chain(DTMC) on state $\{0\} \cup \mathcal{J}$ with the transition constructed from the routing matrix P. Here we use $\{0\}$ to denote the exit state. This network is open, that is, the routing matrix P on $\mathcal{J}$ is assumed to be transient or equivalently $(I - P)$ being invertible. Upon the service completion at one particular station, jobs either go to another station or exit the network, following the routing matrix P.

The following definitions follow [7] closely. Define $T_{e,j} = \{T_{e,j}(i), i \in \mathbb{N}_+\}$ and $T_{e,j} = \{T_{s,j}(i), i \in \mathbb{N}_+\}$ as two i.i.d. sequences of random variables associated with station $j \in \mathcal{J}$. Assume $T_{e,1}, ..., T_{e,J}, T_{s,1}, ..., T_{s,J}$ are independent. Following [7], we assume $\mathbb{E}[T_{e,j}] = 1$, $\mathbb{E}[T_{s,j}] = 1$ such that $T_{e,j}$ and $T_{s,j}$ are unitized. Denote by $c_{e,j}^2 = \mathrm{Var}\left(T_{e,j}(1)\right)$ and $c_{s,j}^2 = \mathrm{Var}\left(T_{s,j}(1)\right)$. We further assume

$$\mathbb{E}[T_{e,j}]^3 < \infty, \quad \mathbb{E}[T_{s,j}]^3 < \infty.$$

Denote by α_j the external arrival rate, and μ_j the service rate at station j. It follows that $T_{e,j}(i)/\alpha_j$ represents the interarrival time between the $(i-1)$th and ith external arriving jobs at station j and $T_{s,j}(i)/\mu_j$ denotes the the ith job service

time at station j. It is known that $c_{e,j}^2$ and $c_{s,j}^2$ are the squared coefficients of variation for interarrival time and service time, respectively.

Traffic Equation. For the queueing networks, let λ be the unique solution to the traffic equation

$$\lambda = \alpha + P'\lambda, \tag{2.1}$$

where P' is transpose of P. Here the solution λ is referred to as the nominal total arrival rate to station j, including the external arrival and the arrival from other stations. Define the traffic intensity by $\rho = \lambda_j/\mu_j$.

Markov process. For $t \geq 0$ and $j \in \mathcal{J}$, let $Z_j(t)$ be the number of jobs in station j, including possibly one being in the service. Let $R_{e,j}(t)$ be the residual time until the next external arrival to station j. Let $R_{s,j}(t)$ be the residual service time for the job being processed in the station j. If $Z_j(t) = 0$, the residual service time is the service time of next station j job, meaning $R_{s,j}(t) = T_{s,j}(i)$ for an appropriate $i \in \mathbb{N}_+$. We write $Z(t), R_e(t), R_s(t)$ as the vectors of $Z_j(t), R_{e,j}(t)$, and $R_{s,j}(t)$, respectively. Define

$$X(t) = (Z(t), R_e(t), R_s(t)), t \geq 0,$$

then $\{X(t); t \geq 0\}$ is the Markov process with respect to the filtration $\mathbb{F}^X \equiv \left\{\mathcal{F}_t^X; t \geq 0\right\}$ defined on the state space $\mathbb{Z}_+^J \times \mathbb{R}_+^J \times \mathbb{R}_+^J$, where $\mathcal{F}_t^X = \sigma(\{X(u); 0 \leq u \leq t\})$. Dai (1995) proves the following assumption holds under some distributional assumption on interarrival times:

Assumption 3. *For each $r \in (0, 1)$, the Markov process $X^{(r)}(\cdot)$ is positive Harris recurrent and it has a unique stationary distribution.*

2.3 Assumptions and Main Results

We consider a family of generalized Jackson networks indexed by $r \in (0,1)$. We denote by $\mu_j^{(r)}$ the mean service rate at station j. Recall the traffic intensity at station j, one has

$$\mu_j^{(r)} = \lambda_j / \rho_j^{(r)}. \tag{2.2}$$

Assumption 4. ***Multi-scale heavy traffic.*** *We assume as $r \downarrow 0$, we have $\mu_j^{(r)} \to \mu_j$, and the traffic intensity for the jth station is*

$$\rho_j^{(r)} = 1 - r^j, \quad j \in \mathcal{J}. \tag{2.3}$$

That is, without loss of generality, we assume the stations have increasing order of intensities. Heavier intensity means $\rho_j^{(r)} \to 1$ faster as $r \downarrow 0$.

Therefore, for each $r \in (0,1)$, with Assumption 3, we define

$$X^{(r)} = (Z^{(r)}, R_e^{(r)}, R_s^{(r)}) \tag{2.4}$$

as the steady-state random vector.

Assumption 5. *For general interarrival and service time distributions, we assume each properly scaled steady-state queue length satisfies some moment boundedness; that is, there exist $M^{(j)} > 0$ and $j \in \mathcal{J}$, such that*

$$\mathbb{E}\left[(r^j Z_j^{(r)})^{2J}\right] \le M^{(j)}, \quad j \in \mathcal{J} \tag{2.5}$$

for all $r \in (0, r_0)$ with some $r_0 \in (0,1)$.

Remark 7. *Assumption 5 holds under the Poisson external arrival and exponential service time, because $Z_j^{(r)}$ is geometrically distributed with mean $\rho^{(r)}/(1-\rho^{(r)})$. Indeed, we will show in Lemma 23 that any orders of the moment of properly scaled queue length*

are bounded for each station under the exponential distribution condition. We expect (2.5) holds under some general arrival and service distributions, such as phase-type distributions. We leave the exploration of this claim to a future study.

Definition 2. *For station $j \in \mathcal{J}$, we denote by $\{i, i < j\} \subset \mathcal{J}$ and $\{i, i > j\} \subset \mathcal{J}$ the set of stations that are lighter and heavier, respectively, than station j. For each $i \in \mathcal{J}$, let w_{ij} be the probability that starting state i, the DTMC will eventually visit state j, avoiding states in $\{0\} \cup \{i, i > j\}$.*

Lemma 17. *$(w_{1j}, \ldots, w_{J,j})$ satisfies the following equations.*

$$w_{ij} = P_{ij} + \sum_{i'<j} P_{ii'} w_{i'j}, \quad i \in \mathcal{J}. \tag{2.6}$$

In particular,

$$(w_{1j}, \ldots, w_{j-1,j})' = [(I-P)_{\{i<j\}}]^{-1} P(:, j)_{\{i<j\}},$$
$$w_{jj} = P_{jj} + P(j, :)_{\{i<j\}} (I-P)_{\{i<j\}}]^{-1} P(:, j)_{\{i<j\}}.$$

Proof. By the first-step method. □

Theorem 18 (Limit distribution)**.** *For $\theta \in \mathbb{R}_-^J$, where $\mathbb{R}_-^J$ is the J-vector with each element nonpositive,*

$$\lim_{r \downarrow 0} \mathbb{E}\left[\exp\left(\sum_{j \in \mathcal{J}} \theta_j r^j Z_j^{(r)}\right)\right] = \prod_{j \in \mathcal{J}} \frac{1}{1 - d_j \theta_j} \tag{2.7}$$

with $d_j = \frac{\sigma_j^2}{2\beta_j}$, where

$$\begin{aligned}\sigma_j^2 =& \alpha_j c_{e,j}^2 + \sum_{i<j} \alpha_i w_{ij}^2 c_{e,i}^2 + \lambda_j c_{s,j}^2 (1 - w_{jj})^2 + \sum_{i>j} \lambda_i w_{ij}^2 c_{s,i}^2 \\ &+ \sum_{i<j} \alpha_i w_{ij}(1 - w_{ij}) + \sum_{i \geq j} \lambda_i w_{ij}(1 - w_{ij})\end{aligned}$$

and $\beta_j = \lambda_j(1 - w_{jj})$. The limit on the RHS of (2.7) is the product of Laplace transforms of J exponential distributed random variables, with means equal to $d_j, j \in \mathcal{J}$ defined above. Marginally, for each $j \in \mathcal{J}$,

$$r^j Z_j^{(r)} \xrightarrow{d} Y_j,$$

where Y_j is the exponential random variable with mean d_j. That is, the properly scaled steady-state queue length for each station converges in distribution to an exponential random variable.

Corollary 19 (Limit distribution under Exponential case)**.** *Under Poission external arrival and Exponential service time, $d_j = 1$ for each $j \in \mathcal{J}$, that is*

$$\lim_{r \downarrow 0} \mathbb{E}\left[\exp\left(\sum_{j \in \mathcal{J}} \theta_j r^j Z_j^{(r)}\right)\right] = \prod_{j \in \mathcal{J}} \frac{1}{1 - \theta_j}. \tag{2.8}$$

Furthermore, if $\theta_j = \theta$, $\forall\, j \in \mathcal{J}$, then one has

$$\lim_{r \downarrow 0} \mathbb{E}\left[\exp\left(\theta \sum_{j \in \mathcal{J}} r^j Z_j^{(r)}\right)\right] = \left(\frac{1}{1 - \theta}\right)^J, \tag{2.9}$$

which means the summation of properly scaled steady-state queue length for each station converges in distribution to the Erlang$(J, 1)$ distribution.

To prove Theorem 18, we first prove Corollary 19 in Section 2.4, as it sheds light on the proof under general distribution in Section 2.5. Readers can see it is straightforward to extend the results under exponential distribution to more general distributions under Assumption 5. In the following, we provide a heuristic approach to obtain the quantity $d_j, j \in \mathcal{J}$ which is interpretable and easily applicable.

2.3.1 Heuristic Interpretation of $d_j's$

We consider a 3-station generalized Jackson network as illustration and discuss d_2 for station 2 as an example. Any general Jackson network applies same heuristic procedure. The routing matrix is assumed to be

$$P = \begin{pmatrix} 0 & P_{12} & P_{13} \\ P_{21} & 0 & P_{23} \\ P_{31} & P_{32} & 0 \end{pmatrix}$$

Assume $1 - \rho_j^{(r)} = r^j$ for $j = 1, 2, 3$. For station 2, one has

$$Z_2(t) = Z_2(0) + A_2(t) - S_2(B_2(t)),$$

where $S_2 = \{S_2(t), t \geq 0\}$ is the renewal process associated with the station 2 service, and $B_2(t)$ is the cumulative busy time of server 2 in $[0, t]$. Here, $A_2(t)$ is the cumulative number of arrivals (including both external and internal ones) to station 2 in $(0, t]$. It is

$$A_2(t) \approx E_2(t) + \sum_{k=1}^{E_1(t)} \xi_{12}(k) + \sum_{k=1}^{S_3(B_3(t))} \xi_{32}(k) + \sum_{k=1}^{S_2(B_2(t))} \xi_{22}(k) \triangleq \tilde{A}_2(t),$$

where $E_j(t)$ is the external arrival process to station j, $\xi_{j2}(k)$ is the Bernoulli random variable, which is equal to 1 if the kth departure from station j will eventually go to to station 2, without going through station 3. The approximation $\approx$ would be equality if every job goes station 1 experiences no delay at all, which can be achieved if the service time at station 1 is 0.

From Lemma 17,

$$\mathbb{E}(\xi_{12}(i)) = P_{12} \equiv w_{12},$$

$$\mathbb{E}(\xi_{32}(i)) = P_{32} + P_{31}P_{12} \equiv w_{32}$$

$$\mathbb{E}(\xi_{22}(i)) = P_{21}P_{12} \equiv w_{22}.$$

Define

$$\hat{E}_j(t) = E_j(t) - \alpha_j t,$$

$$\hat{S}_j(t) = S_j(t) - \mu_j^{(r)} t.$$

Then,

$$\begin{aligned}\tilde{A}_2(t) =& \hat{E}_2(t) + \alpha_2 t + \sum_{k=1}^{E_1(t)} (\xi_{12}(k) - w_{12}) + w_{12}\hat{E}_1(t) + w_{12}\alpha_1 t + \\ &+ \sum_{k=1}^{S_3(B_3(t))} (\xi_{32}(k) - w_{32}) + w_{32}\hat{S}_3(B_3(t)) + w_{32}\mu_3^{(r)} B_3(t) + \\ &+ \sum_{k=1}^{S_2(B_2(t))} (\xi_{22}(k) - w_{22}) + w_{22}\hat{S}_2(B_2(t)) + w_{22}\mu_2^{(r)} B_2(t).\end{aligned}$$

Therefore,

$$Z_2(t) \approx Z_2(0) + \eta_2(t) + \hat{\beta}_2 t + w_{32}\mu_3 I_3(t) + (1 - w_{22})\mu_2 I_2(t),$$

where

$$\begin{aligned}\eta_2(t) =& \hat{E}_2(t) + \sum_{k=1}^{E_1(t)} (\xi_{12}(k) - w_{12}) + w_{12}\hat{E}_1(t) + \sum_{k=1}^{S_3(B_3(t))} (\xi_{32}(k) - w_{32}) + w_{32}\hat{S}_3(B_3(t)) \\ &+ \sum_{k=1}^{S_2(B_2(t))} (\xi_{22}(k) - w_{22}) - (1 - w_{22})\hat{S}_2(B_2(t))\end{aligned}$$

$$\hat{\beta}_2 = \alpha_2 + w_{12}\alpha_1 + w_{32}\lambda_3 - (1 - w_{22})\mu_2^{(r)} = -r^2(1 - w_{22})\mu_2^{(r)}$$

$I_j(t) = t - B_j(t)$ the cumulative idle time at station j.

Assume $I_3(t) = 0$. Z_2 can be modeled as a $(\hat{\beta}_2, \sigma_2^2)$-RBM (reflected Brownian motion), whose stationary distribution is exponential with mean

$$d_2 = \frac{\sigma_2^2}{2|\hat{\beta}_2|},$$

where σ_2^2 is the asymptotic variance of η_2

$$\begin{aligned}\sigma_2^2 =& \lim_{t\to\infty} \frac{\mathrm{Var}(\eta_2(t))}{t} = \alpha_2 c_{e,2}^2 + \alpha_1 w_{22}(1 - w_{22}) + w_{12}^2 \alpha_1 c_{e,1}^2 \\ &+ \lambda_3 w_{32}(1 - w_{32}) + w_{32}^2 \lambda_3 c_{s,3}^2 + \lambda_2 w_{22}(1 - w_{22}) + (1 - w_{22})^2 \lambda_2 c_{s,2}^2.\end{aligned}$$

Furthermore, since Z_2 is non-scaling queue length, then $r^2 Z_2$ can be modeled as a $(\tilde{\beta}_2, \sigma_2^2)$-RBM, where

$$\tilde{\beta}_2 \triangleq \frac{\hat{\beta}_2}{r^2} = -(1 - w_{22})\mu_2^{(r)},$$

which matches the theoretical results in Theorem 18, where β_2 is the limit of $|\tilde{\beta}_2|$ as $r \downarrow 0$.

2.4 Proof under exponential distribution

Under the exponential distribution assumption on the interarrival and service time, $Z^{(r)}(\cdot) = \left\{Z^{(r)}(t), t \geq 0\right\}$ is the continuous time Markov Chain (CTMC) with discrete state space $\mathbb{Z}_+^J$ for each $r \in (0, 1)$. Define the generator G for the CTMC $Z^{(r)}(\cdot)$ as following

$$Gf(z) = \sum_{i \in \mathcal{J}} \alpha_i \Big[f(z + e^{(i)}) - f(z) \Big] + \sum_{j \in \{0\} \cup \mathcal{J}} \sum_{i \in \mathcal{J}} \mu_i^{(r)} P_{ij} \Big[f(z - e^{(i)} + e^{(j)}) - f(z) \Big] \mathbb{1}(z_i > 0), \quad (2.10)$$

where $e^{(i)}$ is the J-vector with jth component being 1 and 0 for all other components, and $e^{(0)}$ is the J-vector of zero. The following lemma can be directly referred to Lemma 5.1 in [7]:

Lemma 20. *Let $f(z) : \mathbb{Z}_+^J \to \mathbb{R}$ be a bounded function. Then the vector of steady-state queue length $Z^{(r)}, r \in (0, 1)$, satisfies the BAR*

$$\mathbb{E}\Big[Gf\Big(Z^{(r)}\Big) \Big] = 0.$$

The following lemma follows the similar discussion as Lemma 4 of [33] and Lemma 1 of [5], and the proof is imbedded in the proof of Moment boundedness in Lemma 23 (see Appendix B.1, Remark 8) by applying the same polynomial test function $f(z)$.

Lemma 21. *Let $f(z) : \mathbb{Z}_+^J \to \mathbb{R}$ be a function. Suppose there exists $n \in \mathbb{N}_+$ such that $|f(z)| \le (\sum_{j\in\mathcal{J}} c_j z_j)^n$ for some constants $c_j \ge 0, j \in \mathcal{J}$ (i.e. $f(z)$ is dominated by a polynomial function). Then there exists $r_0 \in (0,1)$, such that the vector of steady-state queue length $Z^{(r)}, r \in (0, r_0)$ satisfies*

$$\mathbb{E}\left[Gf\left(Z^{(r)}\right)\right] = 0.$$

Lemma 22 (Idle probability)**.**

$$\mathbb{P}\left(Z_j^{(r)} = 0\right) = 1 - \lambda_j/\mu_j^{(r)} = r^j, j \in \mathcal{J}.$$

Proof of Lemma 22. We provide a simple proof using test function with Lemma 21 directly. An alternative proof can be referred in [6] (Lemma 4.4) under general distribution. For station $k \in \mathcal{J}$, let $f(z) = e^{(k)\prime}(1-P')^{-1}z$, then the generator (2.10) can be written as a matrix form:

$$\begin{aligned}
Gf(z) =& e^{(k)\prime}(1-P')^{-1}\alpha - e^{(k)\prime}(1-P')^{-1}\begin{pmatrix} \mu_1^{(r)}(1-P_{11})\mathbb{1}(z_1>0) - \sum_{i\in\mathcal{J}\setminus\{1\}}\mu_i P_{i1}\mathbb{1}(z_i>0) \\ \vdots \\ \mu_J^{(r)}(1-P_{JJ})\mathbb{1}(z_J>0) - \sum_{i\in\mathcal{J}\setminus\{J\}}\mu_i P_{iJ}\mathbb{1}(z_i>0) \end{pmatrix} \\
=& e^{(k)\prime}(1-P')^{-1}(1-P')\lambda - e^{(k)\prime}(1-P')^{-1}(1-P')\begin{pmatrix} \mu_1^{(r)}\mathbb{1}(z_1>0) \\ \vdots \\ \mu_J^{(r)}\mathbb{1}(z_J>0) \end{pmatrix} \\
=& \lambda_k - \mu_k^{(r)}\mathbb{1}(z_k>0).
\end{aligned}$$

With Lemma 21 and (2.2), taking expectation, one has

$$\mathbb{P}(Z_k^{(r)} = 0) = 1 - \mathbb{P}(Z_k^{(r)} = 0) = 1 - \frac{\lambda_k}{\mu_k^{(r)}} = r^k.$$

□

Lemma 23 (Moment boundedness)**.** *For each station $j \in \mathcal{J}$, $\forall n \in \mathbb{N}^+$, there exists $M_n^{(j)} > 0$ and $r_0 \in (0,1)$ such that when $r \in (0, r_0)$,*

$$\mathbb{E}[(r^j Z_j^{(r)})^n] \le M_n^{(j)}.$$

That is, any order moment of properly scaled steady-state queue length for each station, is bounded.

Since $Z_j^{(r)}$ is geometrically distributed, Lemma 23 is trivial. While here we provide a new proof in the Appendix B.1 to shed lights on the future study of proof under the general distribution. Now consider a particular family of test functions given by

$$f(z) = \exp(\langle \theta, z \rangle), z \in \mathbb{R}_+^J, \theta \in \mathbb{R}_-^J,$$

where $\mathbb{R}_+^J$ is the J-vector with each element nonnegative. Then the generator becomes

$$\begin{aligned} Gf(z) = & \sum_{i \in \mathcal{J}} \alpha_i (\exp(\langle \theta, e^{(i)} \rangle) - 1) f(z) \\ & + \sum_{j \in \{0\} \cup \mathcal{J}} \sum_{i \in \mathcal{J}} \mu_i^{(r)} P_{ij} (\exp(\langle \theta, e^{(j)} - e^{(i)} \rangle) - 1) f(z) \mathbb{1}(z_i > 0). \end{aligned} \tag{2.11}$$

Lemma 24. *Let $f_{\theta(r)}(z) = \exp(\langle \theta(r), z \rangle), z \in \mathbb{R}_+^J$, and $\theta(r) \in \mathbb{R}_-^J$ is a function of r, with $\theta(r) \uparrow 0$ as $r \downarrow 0$. Then the following asymptotic version of BAR for the CTMC $Z^{(r)}$ holds:*

$$\begin{aligned} & \theta(r)'(1 - P') \begin{pmatrix} r\mu_1^{(r)} \\ \vdots \\ r^J \mu_J^{(r)} \end{pmatrix} \mathbb{E}\left[f_{\theta(r)}\left(Z^{(r)}\right) \right] - \theta(r)'(1 - P') \begin{pmatrix} \mu_1^{(r)} \mathbb{E}\left[f_{\theta(r)}\left(Z^{(r)}\right) \mathbb{1}(Z_1^{(r)} = 0) \right] \\ \vdots \\ \mu_J^{(r)} \mathbb{E}\left[f_{\theta(r)}\left(Z^{(r)}\right) \mathbb{1}(Z_J^{(r)} = 0) \right] \end{pmatrix} \\ & - \theta(r)'(1 - P') \mathrm{diag}(\mu) \theta(r) \mathbb{E}\left[f_{\theta(r)}\left(Z^{(r)}\right) \right] = o(|\theta(r)|^2), \quad \text{as } r \downarrow 0. \end{aligned} \tag{2.12}$$

Proof of Lemma 24. Note that $\exp(\langle \theta(r), z \rangle) \leq 1$ for any $\theta(r) \in \mathbb{R}_-^J$ and $z \in \mathbb{R}_+^J$, then

with the second order Taylor expansion, the generator 2.11 becomes

$$
\begin{aligned}
&Gf_{\theta(r)}(z) \\
=&\sum_{i\in\mathcal{J}} \alpha_i(\theta(r)'e^{(i)} + \frac{1}{2}\theta(r)'e^{(i)}e^{(i)\prime}\theta(r)))f_{\theta(r)}(z) + o(|\theta(r)|^2) \\
&+ \sum_{j\in\{0\}\cup\mathcal{J}}\sum_{i\in\mathcal{J}} \mu_i^{(r)}P_{ij}(\theta(r)'(e^{(j)}-e^{(i)}) + \frac{1}{2}\theta(r)'(e^{(j)}-e^{(i)})(e^{(j)}-e^{(i)})'\theta(r))f_{\theta(r)}(z)\mathbb{1}(z_i>0) \\
=&\theta(r)'\alpha f_{\theta(r)}(z) - \theta(r)'(1-P')\mu^{(r)}f_{\theta(r)}(z) \\
&+ \theta(r)'(1-P')\begin{pmatrix}\mu_1^{(r)}f_{\theta(r)}(z)\mathbb{1}(z_1=0)\\ \vdots \\ \mu_J^{(r)}f_{\theta(r)}(z)\mathbb{1}(z_J=0)\end{pmatrix} \quad (\triangleq G_1(\theta(r))) \\
&+ \frac{1}{2}\theta(r)'\mathrm{diag}(\alpha)\theta(r)f_{\theta(r)}(z) \\
&+ \frac{1}{2}\sum_{j\in\{0\}\cup\mathcal{J}}\sum_{i\in\mathcal{J}} \mu_i^{(r)}P_{ij}\theta(r)'(e^{(j)}-e^{(i)})(e^{(j)}-e^{(i)})'\theta(r)f_{\theta(r)}(z) \quad (\triangleq G_2(\theta(r))) \\
&+ o(|\theta(r)|^2).
\end{aligned}
\tag{2.13}
$$

With (2.1), (2.2) and (2.3) , the first order terms become:

$$
\begin{aligned}
G_1(\theta(r)) =&\theta(r)'(1-P')(\lambda-\mu^{(r)})f_{\theta(r)}(z) + \theta(r)'(1-P')\begin{pmatrix}\mu_1^{(r)}f_{\theta(r)}(z)\mathbb{1}(z_1=0)\\ \vdots \\ \mu_J^{(r)}f_{\theta(r)}(z)\mathbb{1}(z_J=0)\end{pmatrix} \\
=&\theta(r)'(1-P')\begin{pmatrix}-r\mu_1^{(r)}\\ \vdots \\ -r^J\mu_J^{(r)}\end{pmatrix}f_{\theta(r)}(z) + \theta(r)'(1-P')\begin{pmatrix}\mu_1^{(r)}f_{\theta(r)}(z)\mathbb{1}(z_1=0)\\ \vdots \\ \mu_J^{(r)}f_{\theta(r)}(z)\mathbb{1}(z_J=0)\end{pmatrix},
\end{aligned}
$$

and the second order terms can be written as:

$$
G_2(\theta(r)) = \frac{1}{2}\theta(r)'(\mathrm{diag}(\alpha)+Q)\theta(r)f_{\theta(r)}(z),
$$

where Q is a symmetric matrix with each element defined as following:

$$\begin{cases} Q_{ii} = \mu_i^{(r)} - \sum_{\ell \in \mathcal{J}} P_{\ell i} \mu_\ell^{(r)}, & i \in \mathcal{J} \\ Q_{ij} = Q_{ji} = -\mu_i^{(r)} P_{ij} - \mu_j^{(r)} P_{ji}, & i, j \in \mathcal{J}. \end{cases}$$

Define $Q^{(2)}$ to be the matrix having

$$\begin{cases} Q_{ii}^{(2)} = 0, & i \in \mathcal{J} \\ Q_{ij}^{(2)} = \mu_j^{(r)} P_{ji} - \mu_i^{(r)} P_{ij}, & i, j \in \mathcal{J}. \end{cases}$$

Denote $Q^{(1)} \triangleq \mathrm{diag}(\alpha) + Q - Q^{(2)}$. Then with (2.1), (2.3) and $\alpha_i = \lambda_i - \sum_{j \in \mathcal{J}} P_{ij} \lambda_j$, one can verify that $Q^{(1)}$ has

$$\begin{cases} Q_{ii}^{(1)} = 2(1 - P_{ii}) \mu_i^{(r)} - r\mu^{(r)} + \sum_{j \in \mathcal{J}} r^j P_{ji} \mu_j^{(r)}, & i \in \mathcal{J} \\ Q_{ij}^{(1)} = -2\mu_j^{(r)} P_{ji}, & i, j \in \mathcal{J}. \end{cases}$$

Furthermore, it is easy to check that for any $\gamma \in \mathbb{R}^J$,

$$\gamma' Q^{(2)} \gamma = 0.$$

Therefore, one has

$$(ii) = \frac{1}{2} \theta(r)' Q^{(1)} \theta(r) f_{\theta(r)}(z) = \theta(r)'(1 - P') \mathrm{diag}(\mu^{(r)}) \theta(r) f_{\theta(r)}(z) + o(|\theta(r)|^2),$$

where the second equality follows from $-r\mu^{(r)} + \sum_{j \in \mathcal{J}} r^j P_{ji} \mu_j^{(r)} = O(r)$. With Lemma 20, we take expectation on each term above and apply bounded convergence theorem to conclude the proof.

□

Proof of Corollary 19. We first give the following lemmas:

Lemma 25. $\forall x \geq 0$,

$$1 - e^{-x} \leq x \quad \textit{for } x \geq 0$$

$$|e^x - 1| \leq e^{|x|} |x| \quad \textit{for } x \in \mathbb{R}$$

Lemma 26. *For $\theta \in \mathbb{R}_-^J$, we consider the jth server, $j \in \mathcal{J}$. Denote a J-vector by*

$$\delta^{(r)}(\theta, j) = (\delta_1^{(r)}(\theta, j), ..., \delta_J^{(r)}(\theta, j))',$$

where

$$\delta_k^{(r)}(\theta, j) \triangleq \begin{cases} \sum_{\ell \geq j} w_{k\ell} r^\ell \theta_\ell, & k < j \\ r^k \theta_k + \sum_{\ell > k} w_{k\ell} r^\ell \theta_\ell, & k \geq j, \end{cases} \tag{2.14}$$

and $\{w_{k\ell}, k, \ell \in \mathcal{J}\}$ are defined in Lemma 17. Then for each fixed $\theta \in \mathbb{R}_-^J$

$$\delta^{(r)}(\theta, j)'(1 - P') = \Big(0, ..., 0, (1 - w_{jj}) r^j \theta_j, O(r^j), ..., O(r^j)\Big), \quad \text{as } r \downarrow 0.$$

Proof of Lemma 26. Denote $R = 1 - P'$. By the matrix form in Lemma 17, we have the following equations:

$$\begin{aligned} \sum_{i<j} w_{ij} R_{ik} + R_{jk} &= 0, \quad k < j \\ \sum_{i<j} w_{ij} R_{ij} + R_{jj} &= 1 - w_{jj}. \end{aligned} \tag{2.15}$$

By construction of, $\delta^{(r)}(\theta, j)$ in 2.14, $\delta^{(r)}(\theta, j)'(1 - P')$ is a J-vector with kth element as follows:

$$[\delta^{(r)}(\theta, j)'(1 - P')]_k = \begin{cases} \sum_{\ell \geq j} \Big(\sum_{i<\ell} w_{i\ell} R_{ik} + R_{\ell k} \Big) r^\ell \theta_\ell, & k < j \\ \Big(\sum_{i<j} w_{ij} R_{ij} + R_{jj} \Big) r^j \theta_j + \sum_{\ell \geq j+1} \Big(\sum_{i<\ell} w_{i\ell} R_{ij} + R_{\ell j} \Big) r^\ell \theta_\ell, & k = j \\ \sum_{\ell \geq j} \Big(\sum_{i<\ell} w_{i\ell} R_{ik} + R_{\ell k} \Big) r^\ell \theta_\ell, & k > j. \end{cases}$$

With (2.15), it is easy to check that when $k < j$, for each $\ell \geq j$,

$$\Big(\sum_{i<\ell} w_{i\ell} R_{ik} + R_{\ell k} \Big) r^\ell \theta_\ell = 0.$$

Similarly, when $k = j$, for each $\ell \geq j + 1$

$$\Big(\sum_{i<\ell} w_{i\ell} R_{ij} + R_{\ell j} \Big) r^\ell \theta_\ell = 0,$$

and we have

$$\Big(\sum_{i<j} w_{ij}R_{ij} + R_{jj}\Big)r^j\theta_j = (1 - w_{jj})r^j\theta_j.$$

(2.15) doesn't support the case when $k > j$, while for each $\ell \geq j$, we just need and have

$$\Big(\sum_{i<\ell} w_{i\ell}R_{ik} + R_{\ell k}\Big)r^\ell\theta_\ell = O(r^j).$$

□

Lemma 27. *For $\theta \in \mathbb{R}^J_-$ and $j \in \mathcal{J}$,*

$$\begin{aligned}(1-\theta_j)\mathbb{E}\Big[\exp\Big(\langle(0,...0,r^j\theta_j,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\Big]\\ =\mathbb{E}\Big[\exp\Big(\langle(0,...0,0,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\Big]+o(1),\end{aligned}$$

where $o(1) \to 0$ as $r \downarrow 0$.

If Lemma 27 holds, then one has

$$\begin{aligned}\mathbb{E}\left[\exp\left(\sum_{j\in\mathcal{J}}\theta_j r^j Z_j^{(r)}\right)\right] &= \frac{1}{1-\theta_1}\mathbb{E}\left[\exp\left(\sum_{j\geq 2}\theta_j r^j Z_j^{(r)}\right)\right]+o(1)\\ &= \frac{1}{1-\theta_1}\frac{1}{1-\theta_2}\mathbb{E}\left[\exp\left(\sum_{j\geq 3}\theta_j r^j Z_j^{(r)}\right)\right]+o(1)\\ &= \ldots = \prod_{j\in\mathcal{J}}\frac{1}{1-\theta_j}+o(1)\end{aligned}$$

and taking $r \downarrow 0$, which concludes the proof of Corollary 19. □

The remaining work is to prove Lemma 27.

Proof of Lemma 27. We defined $\tilde{\delta}^{(r)}(\theta, j)$ to be a J-vector as following:

$$\tilde{\delta}^{(r)}_k(\theta,j) \triangleq \begin{cases} w_{kj}r^{j+1/2}\theta_j + \sum_{\ell>j} w_{k\ell}r^\ell\theta_\ell, & k<j\\ r^{j+1/2}\theta_j + \sum_{\ell>j} w_{j\ell}r^\ell\theta_\ell, & k=j\\ r^k\theta_k + \sum_{\ell>k} w_{k\ell}r^\ell\theta_\ell, & k>j.\end{cases} \tag{2.16}$$

Then we list the following lemmas and the proofs will be provided in the end of this section.

Lemma 28. *Suppose $\delta^{(r)}(\theta, j)$ and $\tilde{\delta}^{(r)}(\theta, j)$ are defined by (2.14) and 2.16, respectively. Then we have*

(a)

$$\begin{aligned} & r^j(1-\theta_j)\mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta, j), ..., \delta_J^{(r)}(\theta, j)), Z^{(r)}\rangle\right)\right] \\ & = \mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta, j), ..., \delta_J^{(r)}(\theta, j)), Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)} = 0)\right] + o(r^j). \end{aligned} \tag{2.17}$$

(b)

$$\begin{aligned} & r^j\mathbb{E}\left[\exp\left(\langle(\tilde{\delta}_1^{(r)}(\theta, j), ..., \tilde{\delta}_J^{(r)}(\theta, j)), Z^{(r)}\rangle\right)\right] \\ & = \mathbb{E}\left[\exp\left(\langle(\tilde{\delta}_1^{(r)}(\theta, j), ..., \tilde{\delta}_J^{(r)}(\theta, j)), Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)} = 0)\right] + o(r^j). \end{aligned} \tag{2.18}$$

Lemma 29. *Suppose $\delta^{(r)}(\theta, j)$ and $\tilde{\delta}^{(r)}(\theta, j)$ are defined by (2.14) and 2.16, respectively. Then we have*

(a)

$$\begin{aligned} & r^j(1-\theta_j)\mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta, j), ..., \delta_J^{(r)}(\theta, j)), Z^{(r)}\rangle\right)\right] \\ =& r^j(1-\theta_j)\mathbb{E}\left[\exp\left(\langle(0, ..., 0, r^j\theta_j, r^{j+1}\theta_{j+1}, ..., r^J\theta_J), Z^{(r)}\rangle\right)\right] + o(r^j), \end{aligned} \tag{2.19}$$

$$\begin{aligned} & \mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta, j), ..., \delta_J^{(r)}(\theta, j)), Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)} = 0)\right] \\ =& \mathbb{E}\left[\exp\left(\langle(0, ..., 0, 0, r^{j+1}\theta_{j+1}, ..., r^J\theta_J), Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)} = 0)\right] + o(r^j). \end{aligned} \tag{2.20}$$

(b)

$$\begin{aligned} & r^j\mathbb{E}\left[\exp\left(\langle(\tilde{\delta}_1^{(r)}(\theta, j), ..., \tilde{\delta}_J^{(r)}(\theta, j)), Z^{(r)}\rangle\right)\right] \\ =& r^j\mathbb{E}\left[\exp\left(\langle(0, ..., 0, 0, r^{j+1}\theta_{j+1}, ..., r^J\theta_J), Z^{(r)}\rangle\right)\right] + o(r^j), \end{aligned} \tag{2.21}$$

$$\begin{aligned} & \mathbb{E}\left[\exp\left(\langle(\tilde{\delta}_1^{(r)}(\theta, j), ..., \tilde{\delta}_J^{(r)}(\theta, j)), Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)} = 0)\right] \\ & = \mathbb{E}\left[\exp\left(\langle(0, ..., 0, 0, r^{j+1}\theta_{j+1}, ..., r^J\theta_J), Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)} = 0)\right] + o(r^j). \end{aligned} \tag{2.22}$$

Combining Lemma 28 and 29, one has

$$\begin{aligned} & r^j(1-\theta_j)\mathbb{E}\left[\exp\left(\langle(0,...,0,r^j\theta_j,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\right] \\ &= \mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right]+o(r^j), \end{aligned} \tag{2.23}$$

and

$$\begin{aligned} & r^j\mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\right] \\ &= \mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right]+o(r^j). \end{aligned} \tag{2.24}$$

Note the RHS's of 2.23 and 2.24 are the same. Dividing both sides by r^j, we therefore have

$$\begin{aligned} & (1-\theta_j)\mathbb{E}\left[\exp\left(\langle(0,...,0,r^j\theta_j,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\right] \\ =& \mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\right]+o(1), \end{aligned} \tag{2.25}$$

which concludes the proof. □

Proof of Lemma 28. (a) For the jth server, let $f_{\delta^{(r)}(\theta,j)}(z) = \exp(\langle\delta^{(r)}(\theta,j),z\rangle)$, where

$z \in \mathbb{R}_+^J$. Then with Lemma 26, the asymptotic BAR (2.12) becomes

$$\begin{aligned}
0 =&\Big(0, ..., 0, (1-w_{jj})r^j\theta_j, O(r^j), ..., O(r^j)\Big)\begin{pmatrix} r\mu_1^{(r)} \\ \vdots \\ r^j\mu_J^{(r)} \end{pmatrix}\mathbb{E}\left[f_{\delta^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right] \\
&-\Big(0, ..., 0, (1-w_{jj})r^j\theta_j, O(r^j), ..., O(r^j)\Big)\begin{pmatrix} \mu_1^{(r)}\mathbb{E}\left[f_{\delta^{(r)}(\theta,j)}\left(Z^{(r)}\right)\mathbb{1}(Z_1^{(r)}=0)\right] \\ \vdots \\ \mu_J^{(r)}\mathbb{E}\left[f_{\delta^{(r)}(\theta,j)}\left(Z^{(r)}\right)\mathbb{1}(Z_J^{(r)}=0)\right] \end{pmatrix} \\
&-\Big(0, ..., 0, (1-w_{jj})r^j\theta_j, O(r^j), ..., O(r^j)\Big)\begin{pmatrix} \mu_1^{(r)}\delta_1^{(r)}(\theta,j) \\ \vdots \\ \mu_J^{(r)}\delta_J^{(r)}(\theta,j), \end{pmatrix}\mathbb{E}\left[f_{\delta^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right] + o(r^{2j}) \\
=&r^{2j}\mu_j^{(r)}\theta_j(1-w_{jj})\mathbb{E}\left[f_{\delta^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right] - r^j\mu_j^{(r)}\theta_j(1-w_{jj})\mathbb{E}\left[f_{\delta^{(r)}(\theta,j)}\left(Z^{(r)}\right)\mathbb{1}(Z_j^{(r)}=0)\right] \\
&- r^{2j}\mu_j^{(r)}\theta_j^2(1-w_{jj})\mathbb{E}\left[f_{\delta^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right] + o(r^{2j}),
\end{aligned} \tag{2.26}$$

where we utilize Lemma 22 that for $\theta \in \mathbb{R}_-^J$,

$$\mathbb{E}\left[f_\theta\left(Z^{(r)}\right)\mathbb{1}(Z_j^{(r)}=0)\right] \le \mathbb{E}\left[\mathbb{1}(Z_j^{(r)}=0)\right] = r^j,$$

and the terms with order higher than r^{2j} are put in $o(r^{2j})$. Dividing both sides by $r^j\mu_j^{(r)}\theta_j(1-w_{jj})$, one has

$$r^j\Big(1-\theta_j\Big)\mathbb{E}\left[f_{\delta^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right] = \mathbb{E}\left[f_{\delta^{(r)}(\theta,j)}\left(Z^{(r)}\right)\mathbb{1}(Z_j^{(r)}=0)\right] + o(r^j). \tag{2.27}$$

(b) For the jth server, let $f_{\tilde{\delta}^{(r)}(\theta,j)}(z) = \exp(\langle\tilde{\delta}^{(r)}(\theta,j), z\rangle)$. By similar argument as in (a), one has

$$\tilde{\delta}^{(r)}(\theta,j)'(1-P') = \Big(0, ..., 0, (1-w_{jj})r^{j+1/2}\theta_j, O(r^{j+1/2}), ..., O(r^{j+1/2})\Big).$$

Then the asymptotic BAR (2.12) becomes

$$
\begin{aligned}
& o(r^{2j+1/2}) \\
=& \Big(0, ..., 0, (1-w_{jj}) r^{j+1/2}\theta_j, O(r^{j+1/2}), ..., O(r^{j+1/2})\Big) \begin{pmatrix} r\mu_1^{(r)} \\ \vdots \\ r^J \mu_J^{(r)} \end{pmatrix} \mathbb{E}\left[f_{\tilde{\delta}^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right] \\
& - \Big(0, ..., 0, (1-w_{jj}) r^{j+1/2}\theta_j, O(r^{j+1/2}), ..., O(r^{j+1/2})\Big) \begin{pmatrix} \mu_1^{(r)} \mathbb{E}\left[f_{\tilde{\delta}^{(r)}(\theta,j)}\left(Z^{(r)}\right) \mathbb{1}(Z_1^{(r)}=0)\right] \\ \vdots \\ \mu_J^{(r)} \mathbb{E}\left[f_{\tilde{\delta}^{(r)}(\theta,j)}\left(Z^{(r)}\right) \mathbb{1}(Z_J^{(r)}=0)\right] \end{pmatrix} \\
& - \Big(0, ..., 0, (1-w_{jj}) r^{j+1/2}\theta_j, O(r^{j+1/2}), ..., O(r^{j+1/2})\Big) \begin{pmatrix} \mu_1 \tilde{\delta}_1^{(r)}(\theta, j) \\ \vdots \\ \mu_J \tilde{\delta}_J^{(r)}(\theta, j), \end{pmatrix} \mathbb{E}\left[f_{\tilde{\delta}^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right], \\
=& r^{2j+1/2} \mu_j^{(r)} \theta_j (1-w_{jj}) \mathbb{E}\left[f_{\tilde{\delta}^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right] \\
& - r^{j+1/2} \mu_j^{(r)} \theta_j (1-w_{jj}) \mathbb{E}\left[f_{\tilde{\delta}^{(r)}(\theta,j)}\left(Z^{(r)}\right) \mathbb{1}(Z_j^{(r)}=0)\right] \\
& - r^{2j+1} \mu_j^{(r)} \theta_j^2 (1-w_{jj}) \mathbb{E}\left[f_{\tilde{\delta}^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right].
\end{aligned} \tag{2.28}
$$

Dividing both sides by $r^{j+1/2}\mu_j^{(r)}\theta_j(1-w_{jj})$, (2.28) becomes

$$
r^j \mathbb{E}\left[f_{\tilde{\delta}^{(r)}(\theta,j)}\left(Z^{(r)}\right)\right] = \mathbb{E}\left[f_{\tilde{\delta}^{(r)}(\theta,j)}\left(Z^{(r)}\right) \mathbb{1}(Z_j^{(r)}=0)\right] + o(r^j). \tag{2.29}
$$

□

Proof of Lemma 29. For (a), with Lemma 25, one has

$$
\begin{aligned}
&r^j\mathbb{E}\Big[\exp\Big(\langle(0,...0,r^j\theta_j,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\Big]-\mathbb{E}\Big[\exp\Big(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\Big)\Big]\\
&=r^j\mathbb{E}\Big[\exp\Big(\langle(0,...0,r^j\theta_j,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\cdot\\
&\Big(1-\exp\Big(\langle(\sum_{\ell\geq j}w_{1\ell}r^\ell\theta_\ell,...\sum_{\ell\geq j}w_{j-1,\ell}r^\ell\theta_\ell,\sum_{\ell>j}w_{j\ell}r^\ell\theta_\ell,\sum_{\ell>j+1}w_{j+1,\ell}r^\ell\theta_\ell,...,w_{J-1,J}r^J\theta_J,0),Z^{(r)}\rangle\Big)\Big]\\
&\leq r^j\mathbb{E}\Big[\sum_{\ell\geq j}w_{1\ell}r^{\ell-1}|\theta_\ell|rZ_1^{(r)}+...+\sum_{\ell\geq j}w_{j-1,\ell}r^{\ell-j+1}|\theta_\ell|r^{j-1}Z_{j-1}^{(r)}+\sum_{\ell>j}w_{j\ell}r^{\ell-j}|\theta_\ell|r^jZ_j^{(r)}\\
&\qquad+\sum_{\ell>j+1}w_{j+1,\ell}r^{\ell-j-1}|\theta_\ell|r^{j+1}Z_{j+1}^{(r)}+...,w_{J-1,J}r|\theta_J|r^{J-1}Z_{J-1}^{(r)}+0\Big]=O(r^{j+1})=o(r^j),
\end{aligned}
\tag{2.30}
$$

where the last inequality is by Lemma 23. Similarly, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\langle(0,...0,r^j\theta_j,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\mathbb{1}(Z_j^{(r)}=0)\Big]\\
&\qquad-\mathbb{E}\Big[\exp\Big(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\Big)\mathbb{1}(Z_j^{(r)}=0)\Big]\\
&=\mathbb{E}\Big[\exp\Big(\langle(0,...0,r^j\theta_j,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\cdot\\
&\Big(1-\exp\Big(\langle(\sum_{\ell\geq j}w_{1\ell}r^\ell\theta_\ell,...\sum_{\ell\geq j}w_{j-1,\ell}r^\ell\theta_\ell,\sum_{\ell>j}w_{j\ell}r^\ell\theta_\ell,\sum_{\ell>j+1}w_{j+1,\ell}r^\ell\theta_\ell,...,0),Z^{(r)}\rangle\Big)\mathbb{1}(Z_j^{(r)}=0)\Big]\\
&\leq\mathbb{E}\Big[\Big(\sum_{\ell\geq j}w_{1\ell}r^{\ell-1}|\theta_\ell|rZ_1^{(r)}+...+\sum_{\ell\geq j}w_{j-1,\ell}r^{\ell-j+1}|\theta_\ell|r^{j-1}Z_{j-1}^{(r)}+\sum_{\ell>j}w_{j\ell}r^{\ell-j}|\theta_\ell|r^jZ_j^{(r)}\\
&\qquad+\sum_{\ell>j+1}w_{j+1,\ell}r^{\ell-j-1}|\theta_\ell|r^{j+1}Z_{j+1}^{(r)}+...+0\Big)\mathbb{1}(Z_j^{(r)}=0)\Big]=o(r^j),
\end{aligned}
\tag{2.31}
$$

where in the last equality, we apply Cauchy-Schwarz inequality, Lemma 22 and Lemma 23 to obtain that for $k\in\mathcal{J}$,

$$
r\mathbb{E}[r^kZ_k^{(r)}\mathbb{1}(Z_j^{(r)}=0)]\leq r\mathbb{E}[(r^kZ_k^{(r)})^{2j}]^{1/(2j)}\mathbb{P}(Z_j^{(r)}=0)^{\frac{j-1/2}{j}}=O(r^{j+1/2})=o(r^j).
$$

For (b), using similar discussion as (2.30)and (2.31), one has

$$
\begin{aligned}
&r^j\mathbb{E}\Big[\exp\Big(\langle(0,...0,0,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\Big]-\mathbb{E}\Big[\exp\Big(\langle(\tilde{\delta}_1^{(r)}(\theta,j),...,\tilde{\delta}_J^{(r)}(\theta,j)),Z^{(r)}\rangle\Big)\Big]\\
&\quad=r^j\mathbb{E}\Big[\exp\Big(\langle(0,...0,0,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\cdot\\
&\Big(1-\exp\Big(\langle(w_{1j}r^{j+1/2}\theta_j+\sum_{\ell>j}w_{1\ell}r^\ell\theta_\ell,...,r^{j+1/2}\theta_j+\sum_{\ell>j}w_{j\ell}r^\ell\theta_\ell,\sum_{\ell>j+1}w_{j+1,\ell}r^\ell\theta_\ell,...,,0),Z^{(r)}\rangle\Big)\Big]\\
&\le r^j\mathbb{E}\Big[w_{1j}r^{j-1/2}|\theta_j|rZ_1^{(r)}+\sum_{\ell>j}w_{1\ell}r^{\ell-1}|\theta_\ell|rZ_1^{(r)}+...+r^{j+1/2}|\theta_j|Z_j^{(r)}\\
&\quad+\sum_{\ell>j}w_{j\ell}r^{\ell-j}|\theta_\ell|r^jZ_j^{(r)}+\sum_{\ell>j+1}w_{j+1,\ell}r^{\ell-j-1}|\theta_\ell|r^{j+1}Z_{j+1}^{(r)}+...,w_{J-1,J}r|\theta_J|r^{J-1}Z_{J-1}^{(r)}+0\Big]\\
&\quad=O(r^{j+1})=o(r^j),
\end{aligned}
\tag{2.32}
$$

and

$$
\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\langle(0,...0,0,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\mathbb{1}(Z_j^{(r)}=0)\Big]\\
&\qquad-\mathbb{E}\Big[\exp\Big(\langle(\tilde{\delta}_1^{(r)}(\theta,j),...,\tilde{\delta}_J^{(r)}(\theta,j)),Z^{(r)}\rangle\Big)\mathbb{1}(Z_j^{(r)}=0)\Big]\\
&=\mathbb{E}\Big[\exp\Big(\langle(0,...0,0,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\Big)\cdot\\
&\Big(1-\exp\Big(\langle(w_{1j}r^{j+1/2}\theta_j+\sum_{\ell>j}w_{1\ell}r^\ell\theta_\ell,...,0,\sum_{\ell>j+1}w_{j+1,\ell}r^\ell\theta_\ell,...,0),Z^{(r)}\rangle\Big)\mathbb{1}(Z_j^{(r)}=0)\Big]\\
&\le\mathbb{E}\Big[\Big(w_{1j}r^{j-1/2}|\theta_j|rZ_1^{(r)}+\sum_{\ell>j}w_{1\ell}r^{\ell-1}|\theta_\ell|rZ_1^{(r)}+...+w_{j-1,j}r^{3/2}|\theta_j|r^{j-1}Z_{j-1}^{(r)}\\
&+\sum_{\ell>j}w_{j-1,\ell}r^{\ell-j+1}|\theta_\ell|r^{j-1}Z_{j-1}^{(r)}+\sum_{\ell>j+1}w_{j+1,\ell}r^{\ell-j-1}|\theta_\ell|r^{j+1}Z_{j+1}^{(r)}+...+0\Big)\mathbb{1}(Z_j^{(r)}=0)\Big]=o(r^j).
\end{aligned}
\tag{2.33}
$$

□

2.5 Proof under general distribution

The proof under general distribution will follow closely to [7] in term of parametrization such that one can see clearly the powerfulness of the BAR ap-

proach on dealing with general distribution under various networks. Recall that $X^r, r \in (0,1)$, is the steady-state in the rth queueing network as defined in (2.4). The following lemma is a direct result from [6] (Lemma 4.4).

Lemma 30.

$$\mathbb{P}\left(Z_j^{(r)} = 0\right) = 1 - \lambda_j/\mu_j^{(r)} = r^j, j \in \mathcal{J}.$$

Following the same argument in [7], we first assume there exists a constant $\kappa > 0$ such that

$$\mathbb{P}\left\{T_{e,j}(1) \le \kappa\right\} = 1, \quad \mathbb{P}\left\{T_{s,j}(1) \le \kappa\right\} = 1, \quad j \in \mathcal{J}. \tag{2.34}$$

The condition (2.34) can be removed by truncating the residual arrival $R_e^{(r)}$ and service time $R_e^{(r)}$ by $1/r$ and making minor modification as discussed in case 2 of Section 7 [7]. We leave this task to a future research.

For $\eta \in \mathbb{R}^J$ and $\xi \in \mathbb{R}^J$, we consider a special class of g that is given by

$$g_\theta(x) = g_\theta(z,u,v) = \exp\left(\langle\theta, z\rangle, \langle\eta, u\rangle, \langle\xi, v\rangle\right), \quad (z,u,v) \in \mathbb{Z}_+^J \times \mathbb{R}_+^J \times \mathbb{R}_+^J.$$

We define

$$\mathcal{A}g_\theta(x) = -\sum_{j\in\mathcal{J}} \frac{\partial g}{\partial u_j}(x) - \sum_{j\in\mathcal{J}} \frac{\partial g}{\partial v_j}(x) 1\left(z_j > 0\right)$$

and $\eta_j\left(\theta_j\right)$ and $\xi_j(\theta)$ via

$$e^{\theta_j}\mathbb{E}e^{\eta_j(\theta_j)T_{e,j}(1)} = 1, \quad j \in \mathcal{J}$$

$$e^{-\theta_j}\left(\sum_{\ell\in\mathcal{J}} e^{\theta_\ell} P_{j\ell} + P_{j0}\right)\mathbb{E}e^{\xi_j(\theta)T_{s,j}(1)} = 1, \quad j \in \mathcal{J}$$

Then following [7] (Lemma 7.1), one has

$$\mathbb{E}\left[\mathcal{A}g_\theta\left(X^r\right)\right] = 0$$

For each $r \in (0,1)$, we further define

$$\eta_j^{(r)}(\theta_j) = \alpha_j\eta_j(\theta_j), \quad \xi_j^{(r)}(\theta) = \mu_j^{(r)}\xi_j(\theta), \tag{2.35}$$

and

$$g_\theta(x) = g_\theta(z,u,v) = \exp\Big(\langle\theta, z\rangle, \sum_{j\in\mathcal{J}} \eta_j^{(r)}(\theta_j)u_j, \sum_{j\in\mathcal{J}} \xi_j^{(r)}(\theta)v_j\Big), \quad (z,u,v) \in \mathbb{Z}_+^J \times \mathbb{R}_+^J \times \mathbb{R}_+^J.$$

Then referring to [7] (Lemma 7.2 and 7.5), we have the following lemmas:

Lemma 31. *If Assumption 3 is satisfied, then the steady-state random variable $X^{(r)}$ follows the following BAR: for each $\theta \in \mathbb{R}_-^J$ and each $r \in (0,1)$,*

$$\sum_{j\in\mathcal{J}} \eta_j^{(r)}\left(\theta_j\right)\mathbb{E}\left[g_\theta\left(X^{(r)}\right)\right] + \sum_{j\in\mathcal{J}} \xi_j^{(r)}(\theta)\,\mathbb{E}\left[g_\theta\left(X^{(r)}\right)\mathbb{1}\left(Z_j^{(r)} > 0\right)\right] = 0. \tag{2.36}$$

Lemma 32. *Set*

$$\bar{\eta}_j\left(\theta_j\right) = -\theta_j, \quad \tilde{\eta}_j\left(\theta_j\right) = -\frac{1}{2}c_{e,j}^2\theta_j^2, \quad j \in \mathcal{J}$$

$$\bar{\xi}_j(\theta) = \Big(\theta_j - \sum_{i\in\mathcal{J}} P_{ji}\theta_i\Big),$$

$$\tilde{\xi}_j(\theta) = -\frac{1}{2}c_{s,j}^2\Big(\theta_j - \sum_{i\in\mathcal{J}} P_{ji}\theta_i\Big)^2 - \frac{1}{2}\sum_{i\in\mathcal{J}} P_{ji}\theta_i^2 + \frac{1}{2}\Big(\sum_{i\in\mathcal{J}} P_{ji}\theta_i\Big)^2, \quad j \in \mathcal{J}.$$

Then as $\theta \to 0$,

$$\eta_j\left(\theta_j\right) = \bar{\eta}_j\left(\theta_j\right) + \tilde{\eta}_j\left(\theta_j\right) + o\left(\theta_j^2\right), \quad j \in \mathcal{J}$$

$$\xi_j(\theta) = \bar{\xi}_j(\theta) + \tilde{\xi}_j(\theta) + o\left(|\theta|^2\right), \quad j \in \mathcal{J}$$

Proof of Theorem 18. Let

$$g_{\delta^{(r)}(\theta,j)}(z) = \exp\Big(\langle\delta^{(r)}(\theta,j), z\rangle, \sum_{j\in\mathcal{J}} \eta_j^{(r)}(\delta_j^{(r)}(\theta,j))u_j, \sum_{j\in\mathcal{J}} \xi_j^{(r)}(\delta^{(r)}(\theta,j))v_j\Big),$$

where $\delta^{(r)}(\theta,j) = (\delta_1^{(r)}(\theta,j), ..., \delta_J^{(r)}(\theta,j))'$ is defined by (2.14). For easy reference to (2.13), we denote

$$\begin{aligned}\hat{G}_1(\delta^{(r)}(\theta,j)) = &\sum_{i\in\mathcal{J}} \alpha_i\bar{\eta}_i^{(r)}(\theta_i)\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right] \\ &+ \sum_{i\in\mathcal{J}} \mu_i^{(r)}\bar{\xi}_i^{(r)}(\theta)\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}\left(Z_j^{(r)} > 0\right)\right], \\ \hat{G}_2(\delta^{(r)}(\theta,j)) = &\sum_{i\in\mathcal{J}} \alpha_i\tilde{\eta}_i^{(r)}(\theta_i)\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right] + \sum_{i\in\mathcal{J}} \lambda_i\tilde{\xi}_i^{(r)}(\theta)\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right].\end{aligned}$$

With (2.35) and Lemma 31, keeping the terms with order r^{2j} or smaller, the BAR 2.36 can be written as

$$\hat{G}_1(\delta^{(r)}(\theta, j)) + \hat{G}_2(\delta^{(r)}(\theta, j)) = o\left(r^{2j}\right).$$

Note that $\hat{G}_1(\theta(r))$ is the same as the first order term of Taylor Expansion under the exponential distribution, where we define $G_1(\theta(r))$ in (2.13), except that function $f_{\theta(r)}(z)$ is replaced by $g_{\theta(r)}(x)$. Then following the proof of Lemma 28, one has

$$\begin{aligned} G_1(\delta^{(r)}(\theta, j)) =& r^{2j}\mu_j^{(r)}\theta_j(1 - w_{jj})\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right] \\ & - r^j\mu_j^{(r)}\theta_j(1 - w_{jj})\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}(Z_j^{(r)} = 0)\right]. \end{aligned}$$

Denote $\tilde{\beta}_j = \mu_j^{(r)}(1 - w_{jj})$, then we have

$$G_1(\delta^{(r)}(\theta, j)) = r^{2j}\tilde{\beta}_j\theta_j\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right] - \tilde{\beta}_j r^j\theta_j\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}(Z_j^{(r)} = 0)\right]. \quad (2.37)$$

Now we consider $G_2(\delta^{(r)}(\theta, j))$. Denote $\tilde{\sigma}_j^2$ such that

$$G_2(\delta^{(r)}(\theta, j)) \triangleq -\frac{1}{2}r^{2j}\theta_j^2\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right]\tilde{\sigma}_j^2,$$

then $\tilde{\sigma}_j^2$ has the following:

$$\begin{aligned} \tilde{\sigma}_j^2 =& \alpha_j c_{e,j}^2 + \sum_{i<j}\alpha_i c_{e,i}^2 w_{ij}^2 + \sum_{i<j}\mu_i^{(r)}c_{s,i}^2(w_{ij} - \sum_{\ell<j}P_{i\ell}w_{\ell j} - P_{ij})^2 \\ & + \sum_{i>j}\mu_i^{(r)}c_{s,i}^2(\sum_{\ell\le j}P_{i\ell}w_{\ell j})^2 + \mu_j^{(r)}c_{s,j}^2(1 - P_{jj} - \sum_{\ell<j}P_{j\ell}w_{\ell,j})^2 \\ & + \sum_{i\in\mathcal{J}}\mu_i^{(r)}\Big[\sum_{\ell<j}P_{i\ell}w_{\ell j}(w_{\ell j} - \sum_{\ell'<j}P_{i\ell'}w_{\ell' j} - P_{ij}) + P_{ij}(1 - \sum_{\ell'<j}P_{i\ell'}w_{\ell' j} - P_{ij})\Big] \\ =& \alpha_j c_{e,j}^2 + \sum_{i<j}\alpha_i c_{e,i}^2 w_{ij}^2 + \sum_{i>j}\mu_i^{(r)}c_{s,i}^2(\sum_{\ell\le j}P_{i\ell}w_{\ell j})^2 + \mu_j^{(r)}c_{s,j}^2(1 - P_{jj} - \sum_{\ell<j}P_{j\ell}w_{\ell,j})^2 \\ & + \sum_{i\in\mathcal{J}}\mu_i^{(r)}\Big[\sum_{\ell<j}P_{i\ell}w_{\ell j}(w_{\ell j} - w_{ij}) + P_{ij}(1 - w_{ij})\Big] \\ =& \alpha_j c_{e,j}^2 + \sum_{i<j}\alpha_i c_{e,i}^2 w_{ij}^2 + \sum_{i>j}\mu_i^{(r)}c_{s,i}^2(\sum_{\ell\le j}P_{i\ell}w_{\ell j})^2 + \mu_j^{(r)}c_{s,j}^2(1 - w_{jj})^2 \\ & + \sum_{i\in\mathcal{J}}\mu_i^{(r)}\Big[\sum_{\ell<j}P_{i\ell}w_{\ell j}(w_{\ell j} - w_{ij}) + P_{ij}(1 - w_{ij})\Big], \end{aligned}$$

where the last term equals

$$
\begin{aligned}
&\sum_{i\in\mathcal{J}}\mu_i^{(r)}\Big[\sum_{\ell<j}P_{i\ell}w_{\ell j}(w_{\ell j}-w_{ij})+P_{ij}(1-w_{ij})\Big]\\
=&\sum_{i\in\mathcal{J}}\mu_i^{(r)}\Big[\sum_{\ell<j}P_{i\ell}w_{\ell j}(w_{\ell j}-1)+\sum_{\ell<j}P_{i\ell}w_{\ell j}(1-w_{ij})+P_{ij}(1-w_{ij})\Big]\\
=&\sum_{i\in\mathcal{J}}\mu_i^{(r)}\Big[\sum_{\ell<j}P_{i\ell}w_{\ell j}(w_{\ell j}-1)+w_{ij}(1-w_{ij})\Big]\\
=&\sum_{i<j}\mu_i^{(r)}\Big[\sum_{\ell<j}P_{i\ell}w_{\ell j}(w_{\ell j}-1)+w_{ij}(1-w_{ij})\Big]\\
&+\sum_{i\geq j}\mu_i^{(r)}\Big[\sum_{\ell<j}P_{i\ell}w_{\ell j}(w_{\ell j}-1)\Big]+\sum_{i\geq j}\mu_i^{(r)}w_{ij}(1-w_{ij})\\
=&\sum_{i<j}w_{ij}(1-w_{ij})\Big[\mu_i^{(r)}-\sum_{\ell\in\mathcal{J}}P_{\ell i}\mu_\ell^{(r)}\Big]+\sum_{i\geq j}\mu_i^{(r)}w_{ij}(1-w_{ij})\\
=&\sum_{i<j}w_{ij}(1-w_{ij})\alpha_i\rho_i^{(r)}+\sum_{i\geq j}\mu_i^{(r)}w_{ij}(1-w_{ij}).
\end{aligned}
$$

Plugging it back into $\tilde{\sigma}_j^2$, one has

$$
\begin{aligned}
\tilde{\sigma}_j^2=&\alpha_j c_{e,j}^2+\sum_{i<j}\alpha_i c_{e,i}^2 w_{ij}^2+\sum_{i>j}\mu_i^{(r)}c_{s,i}^2\Big(\sum_{\ell\leq j}P_{i\ell}w_{\ell j}\Big)^2+\mu_j^{(r)}c_{s,j}^2(1-w_{jj})^2\\
&+\sum_{i<j}w_{ij}(1-w_{ij})\alpha_i\rho_i^{(r)}+\sum_{i\geq j}\mu_i^{(r)}w_{ij}(1-w_{ij}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
o(r^{2j})=&G_1(\delta^{(r)}(\theta,j))+G_2(\delta^{(r)}(\theta,j))\\
=&r^{2j}\tilde{\beta}_j\theta_j\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right]-\tilde{\beta}_j r^j\theta_j\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}(Z_j^{(r)}=0)\right]\\
&-\frac{1}{2}r^{2j}\theta_j^2\tilde{\sigma}_j^2\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right].
\end{aligned}
$$

Divide both sides by $r^j\tilde{\beta}_j\theta_j$, one has

$$
r^j\Big(1-\theta_j\frac{\tilde{\sigma}_j^2}{2\tilde{\beta}_j}\Big)\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right]=\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}(Z_j^{(r)}=0)\right]+o(r^j). \tag{2.38}
$$

Now we let

$$
g_{\tilde{\delta}^{(r)}(\theta,j)}(z)=\exp\Big(\langle\tilde{\delta}^{(r)}(\theta,j),z\rangle,\sum_{j\in\mathcal{J}}\eta_j^{(r)}(\tilde{\delta}_j^{(r)}(\theta,j))u_j,\sum_{j\in\mathcal{J}}\xi_j^{(r)}(\tilde{\delta}^{(r)}(\theta,j))v_j\Big),
$$

where $\tilde{\delta}^{(r)}(\theta, j) = (\tilde{\delta}_1^{(r)}(\theta, j), ..., \tilde{\delta}_J^{(r)}(\theta, j))'$ is defined by (2.16). Similar discussion for (2.38) and Lemma 28 applied on $g_{\tilde{\delta}^{(r)}(\theta,j)}(z)$ gives

$$r^j \mathbb{E}\left[g_{\tilde{\delta}^{(r)}(\theta,j)}\left(X^{(r)}\right)\right] = \mathbb{E}\left[g_{\tilde{\delta}^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}(Z_j^{(r)} = 0)\right] + o(r^j). \tag{2.39}$$

Denote by $\tilde{d}_j \triangleq \frac{\tilde{\sigma}_j^2}{2\tilde{\beta}_j}$. Referring to [7] (Lemma 7.4), we have the following lemma, and the proof will be put in the end of this section.

Lemma 33. *Suppose $\delta^{(r)}(\theta, j)$ and $\tilde{\delta}^{(r)}(\theta, j)$ are defined by (2.14) and 2.16, respectively. Recall $Z^{(r)}$ is the steady-state queue length in $X^{(r)} = (Z^{(r)}, R_e^{(r)}, R_s^{(r)})$. Then we have*

(a)

$$\begin{aligned} &r^j(1-\theta_j\tilde{d}_j)\mathbb{E}\Big[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\Big] \\ &\quad -r^j(1-\theta_j\tilde{d}_j)\mathbb{E}\Big[\exp\left(\langle(0,...0,r^j\theta_j,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\right)\Big] = o(r^j), \end{aligned} \tag{2.40}$$

$$\begin{aligned} &\mathbb{E}\Big[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}(Z_j^{(r)}=0)\Big] \\ &\quad -\mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right] = o(r^j). \end{aligned} \tag{2.41}$$

(b)

$$r^j\mathbb{E}\left[g_{\tilde{\delta}^{(r)}(\theta,j)}\left(X^{(r)}\right)\right] - r^j\mathbb{E}\Big[\exp\left(\langle(0,...0,0,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\right)\Big] = o(r^j), \tag{2.42}$$

$$\begin{aligned} &\mathbb{E}\Big[g_{\tilde{\delta}^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}(Z_j^{(r)}=0)\Big] \\ &\quad -\mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right] = o(r^j). \end{aligned} \tag{2.43}$$

Combining Lemma 33 with (2.38) and (2.39), the steady-state queue length $Z^{(r)}$ in $X^{(r)} = (Z^{(r)}, R_e^{(r)}, R_s^{(r)})$ satisfies

$$\begin{aligned} &r^j(1-\theta_j\tilde{d}_j)\mathbb{E}\left[\exp\left(\langle(0,...,0,r^j\theta_j,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\right] \\ &= \mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right] + o(r^j), \end{aligned} \tag{2.44}$$

and

$$
\begin{aligned}
&r^j\mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\right]\\
&=\mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right]+o(r^j).
\end{aligned}
\tag{2.45}
$$

Note the RHS's of 2.44 and 2.45 are the same. Dividing both sides by r^j, we therefore have

$$
\begin{aligned}
&(1-\theta_j\tilde{d}_j)\mathbb{E}\left[\exp\left(\langle(0,...,0,r^j\theta_j,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\right]\\
=&\mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\right]+o(1).
\end{aligned}
\tag{2.46}
$$

Readers can notices that (2.46) is the same as (2.25) except that here $\tilde{d}_j = \frac{\tilde{\sigma}_j^2}{2\tilde{\beta}_j}$ replacing $\tilde{d}_j = 1$. Denoting by

$$
\begin{aligned}
\beta_j &\triangleq \lim_{r\downarrow 0}\tilde{\beta}_j = \lim_{r\downarrow 0}\mu_j^{(r)}(1-w_{jj}) = \lambda_j(1-w_{jj})\\
\sigma_j^2 &\triangleq \lim_{r\downarrow 0}\tilde{\sigma}_j^2 = \alpha_j c_{e,j}^2 + \sum_{i<j}\alpha_i w_{ij}^2 c_{e,i}^2 + \lambda_j c_{s,j}^2(1-w_{jj})^2 + \sum_{i>j}\lambda_i w_{ij}^2 c_{s,i}^2\\
&\quad + \sum_{i<j}\alpha_i w_{ij}(1-w_{ij}) + \sum_{i\geq j}\lambda_i w_{ij}(1-w_{ij})
\end{aligned}
$$

and $d_j \triangleq \frac{\sigma_j^2}{2\beta_j}$ concludes the proof. □

Proof of Lemma 33. For (a)(2.40), by subtracting and adding a term, we have two summations as follows:

$$
\begin{aligned}
&r^j(1-\theta_j\tilde{d}_j)\left\{\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right]-\mathbb{E}\left[\exp\left(\langle(0,...0,r^j\theta_j,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\right)\right]\right\}\\
=&r^j(1-\theta_j\tilde{d}_j)\left\{\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right]-\mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\right)\right]\right\}\\
&+r^j(1-\theta_j\tilde{d}_j)\left\{\mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\right)\right]\right.\\
&\qquad\left.-\mathbb{E}\left[\exp\left(\langle(0,...0,r^j\theta_j,r^{j+1}\theta_{j+1},...r^J\theta_J),Z^{(r)}\rangle\right)\right]\right\},
\end{aligned}
$$

where the second summation is the same as (2.19) except the constant $\tilde{d}_j$, which has been proved in Lemma 29. Therefore, we only need to consider the first

summation. With Lemma 25, one has

$$
\begin{aligned}
& r^j(1-\theta_j\tilde{d}_j)\left\{\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right]-\mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\right)\right]\right\} \\
=& r^j(1-\theta_j\tilde{d}_j)\mathbb{E}\Big[\exp\left(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\right)\cdot \\
& \left(1-\exp\left\{\langle\eta^{(r)}(\delta_j^{(r)}(\theta,j)),R_e\rangle+\langle\xi^{(r)}(\delta^{(r)}(\theta,j)),R_s^{(r)}\rangle\right\}\right)\Big] \\
\le& r^j(1-\theta_j\tilde{d}_j)\mathbb{E}\Big[\left(1-\exp\left\{\langle\eta^{(r)}(\delta_j^{(r)}(\theta,j)),R_e\rangle+\langle\xi^{(r)}(\delta^{(r)}(\theta,j)),R_s^{(r)}\rangle\right\}\right)\Big] \\
\le& r^j(1-\theta_j\tilde{d}_j)Y_j^{(r)}\exp(Y_j^{(r)}),
\end{aligned}
$$

where we have used

$$
\left|\langle\eta^{(r)}(\delta_j^{(r)}(\theta,j)),R_e\rangle+\langle\xi^{(r)}(\delta^{(r)}(\theta,j)),R_s^{(r)}\rangle\right|\le\left(\left|\eta^{(r)}(\delta_j^{(r)}(\theta,j))\right|+\left|\xi^{(r)}(\delta^{(r)}(\theta,j))\right|\right)C\triangleq Y_j^{(r)}
$$

and $C>0$ is a constant large enough that

$$
\kappa/\lambda_j\le C,\quad \kappa/\mu_j^{(r)}\le C.
$$

From Lemma 32, one has

$$
Y_j^{(r)}\to 0,\quad \text{as } r\downarrow 0, \tag{2.47}
$$

where we use

$$
\left|\eta^{(r)}(\delta_j^{(r)}(\theta,j))\right|=O(r^j),\quad \left|\xi^{(r)}(\delta^{(r)}(\theta,j))\right|=O(r^j)
$$

with the definition of $\delta^{(r)}(\theta,j)$ in (2.14). Therefore,

$$
\begin{aligned}
& r^j\left\{\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\right]-\mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\right)\right]\right\} \\
\le& r^jY_j^{(r)}\exp(Y_j^{(r)})=o(r^j),
\end{aligned}
$$

For (a)(2.41), by subtracting and adding a term, we have two summations as follows:

$$
\begin{aligned}
& \left\{\mathbb{E}\Big[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}(Z_j^{(r)}=0)\Big]-\mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right]\right\} \\
& +\left\{\mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right]\right. \\
& \qquad\left.-\mathbb{E}\left[\exp\left(\langle(0,...,0,0,r^{j+1}\theta_{j+1},...,r^J\theta_J),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right]\right\}
\end{aligned}
$$

where the second summation is the same as (2.20), which has been proved in Lemma 29. Now we only consider the first summation. With Lemma 25, one has

$$
\begin{aligned}
&\left\{\mathbb{E}\left[g_{\delta^{(r)}(\theta,j)}\left(X^{(r)}\right)\mathbb{1}(Z_j^{(r)}=0)\right]-\mathbb{E}\left[\exp\left(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\right)\mathbb{1}(Z_j^{(r)}=0)\right]\right\}\\
=&\mathbb{E}\Big[\exp\left(\langle(\delta_1^{(r)}(\theta,j),...,\delta_J^{(r)}(\theta,j)),Z^{(r)}\rangle\right)\cdot\\
&\left(1-\exp\left\{\langle\eta^{(r)}(\delta_j^{(r)}(\theta,j)),R_e\rangle+\langle\xi^{(r)}(\delta^{(r)}(\theta,j)),R_s^{(r)}\rangle\right\}\right)\mathbb{1}(Z_j^{(r)}=0)\Big]\\
\leq&\mathbb{E}\Big[\left(1-\exp\left\{\langle\eta^{(r)}(\delta_j^{(r)}(\theta,j)),R_e\rangle+\langle\xi^{(r)}(\delta^{(r)}(\theta,j)),R_s^{(r)}\rangle\right\}\right)\mathbb{1}(Z_j^{(r)}=0)\Big]\\
\leq&Y_j^{(r)}\exp(Y_j^{(r)})\mathbb{E}[\mathbb{1}(Z_j^{(r)}=0)]=Y_j^{(r)}\exp(Y_j^{(r)})\mathbb{P}(Z_j^{(r)}=0)=o(r^j),
\end{aligned}
$$

where in the last equality, we use the Lemma 22 and (2.47).

Similarly, for (b) (2.42) and (2.43), we can also subtract and add a term and combine the summation with Lemma 29 (b) (2.21) and (2.22), respectively. Then the proof is concluded.

□

APPENDIX A

APPENDIX OF CHAPTER 1

A.1 Proofs in Section 1.5: Preliminary Results I

A.1.1 Proof of Lemma 4(*a*)

Proof. Assume $f(z) : \mathbb{R}^J \to \mathbb{R}$ satisfy $|f(z)| \leq C \sum_{k=1}^{K} W_k^n(z)$ for some $C > 0$. As discussed in [[5] Lemma 1], a sufficient condition to ensure

$$\mathbb{E}\Big[Gf\Big(Z^{(\epsilon)}(\infty)\Big)\Big] = 0$$

is given by [22] and [15], which requires

$$\mathbb{E}\Big[\Big|G\left(Z^{(\epsilon)}(\infty), Z^{(\epsilon)}(\infty)\right) f\Big(Z^{(\epsilon)}(\infty)\Big)\Big|\Big] < \infty$$

where $G(z, z)$ is the diagonal element of the generator matrix G corresponding to state z. First, we have

$$\begin{aligned}
&\Big|G\left(Z^{(\epsilon)}(\infty), Z^{(\epsilon)}(\infty)\right)\Big| \\
=&\Big| -\left(\sum_{i=1}^{I} \lambda_i^{(\epsilon)} \sum_{k\in K(i)} \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)}(\infty)\right) + \sum_{k=1}^{K} \sum_{i\in I(k)} \mu_{ik} P_{ik}\left(Z^{(\epsilon)}(\infty)\right)\right)\Big| \\
\leq& \sum_{i=1}^{I} \lambda_i + \sum_{k=1}^{K} \sum_{i\in I(k)} \mu_{ik}
\end{aligned}$$

Therefore, by assumption, it is sufficient to show that $\sum_{k=1}^{K} \mathbb{E}\Big[W_k^n\Big(Z^{(\epsilon)}(\infty)\Big)\Big] < \infty$. Now let

$$V(z) = \frac{1}{n+1} \sum_{k=1}^{K} \frac{1}{u_k^{n-1}} W_k^{n+1}(z),$$

then by binomial expansion, the generator becomes

$$
\begin{aligned}
GV(z) =& \frac{1}{n+1}\sum_{i=1}^{I}\lambda_i^{(\epsilon)}\sum_{k\in K(i)}\frac{1}{u_k^{n-1}}\left(\sum_{\ell=1}^{n+1}\binom{n+1}{\ell}\left(\frac{1}{\mu_{ik}}\right)^{\ell}W_k^{n+1-\ell}(z)\right)\mathbb{1}\left(k=H^{(i)}(z)\right)\\
&+\frac{1}{n+1}\sum_{k=1}^{K}\sum_{i\in I(k)}\frac{\mu_{ik}P_{ik}(z)}{u_k^{n-1}}\left(\sum_{\ell=1}^{n+1}\binom{n+1}{\ell}\left(-\frac{1}{\mu_{ik}}\right)^{\ell}W_k^{n+1-\ell}(z)\right)\\
=&\sum_{i=1}^{I}\lambda_i^{(\epsilon)}\sum_{k\in K(i)}\frac{1}{u_k^{n-1}\mu_{ik}}W_k^n(z)\mathbb{1}\left(k=H^{(i)}(z)\right)\\
&-\sum_{k=1}^{K}\frac{1}{u_k^{n-1}}W_k^n(z)\sum_{i\in I(k)}P_{ik}(z)+o\Big(W_k^n(z)\Big)\\
\overset{(a)}{\leq}&\sum_{i=1}^{I}\lambda_i^{(\epsilon)}\sum_{k\in K(i)}\frac{1}{v_i^{n-1}}\left(\frac{1}{\mu_{ik}}W_k(z)\right)^n\mathbb{1}\left(k=H^{(i)}(z)\right)-\sum_{k=1}^{K}\frac{1}{u_k^{n-1}}W_k^n(z)+o\Big(W_k^n(z)\Big)\\
\overset{(b)}{=}&\sum_{i=1}^{I}\frac{\lambda_i^{(\epsilon)}}{v_i^{n-1}}\min_{k\in K(i)}\left(\frac{1}{\mu_{ik}}W_k(z)\right)^n-\sum_{k=1}^{K}\frac{1}{u_k^{n-1}}W_k^n(z)+o\Big(W_k^n(z)\Big)
\end{aligned}
$$

where (a) is by (1.7), (b) is by the definition of WWTA policy. According to [11] (Lemma 11.2), there exist a set of $\lambda_{ik}^{(\epsilon)}$ such that paraller-server system under WWTA policy satisfies $\lambda_i^{(\epsilon)}=\sum_{k\in K(i)}\lambda_{ik}^{(\epsilon)}$ and $\sum_{i\in I(k)}\lambda_{ik}^{(\epsilon)}m_{ik}<1$. Then the generator

becomes

$$
\begin{aligned}
GV(z) &= \sum_{i=1}^{I} \frac{\sum_{k\in K(i)} \lambda_{ik}^{(\epsilon)}}{v_i^{n-1}} \min_{k\in K(i)} \left(\frac{1}{\mu_{ik}} W_k(z)\right)^n - \sum_{k=1}^{K} \frac{1}{u_k^{n-1}} W_k^n(z) + o\Big(W_k^n(z)\Big) \\
&\le \sum_{i=1}^{I} \frac{1}{v_i^{n-1}} \sum_{k\in K(i)} \lambda_{ik}^{(\epsilon)} \left(\frac{1}{\mu_{ik}} W_k(z)\right)^n - \sum_{k=1}^{K} \frac{1}{u_k^{n-1}} W_k^n(z) + o\Big(W_k^n(z)\Big) \\
&\overset{(c)}{=} \sum_{i=1}^{I} \sum_{k\in K(i)}^{b} \frac{\lambda_{ik}^{(\epsilon)}}{\mu_{ik}} \frac{1}{u_k^{n-1}} W_k^n(z) + \sum_{i=1}^{I} \sum_{k\in K(i)}^{nb} \frac{\lambda_{ik}^{(\epsilon)}}{\mu_{ik}} \frac{1}{(u_k - d_{ik})^{n-1}} W_k^n(z) \\
&\quad - \sum_{k=1}^{K} \frac{1}{u_k^{n-1}} W_k^n(z) + o\Big(W_k^n(z)\Big) \\
&= \sum_{i=1}^{I} \sum_{k\in K(i)} \frac{\lambda_{ik}^{(\epsilon)}}{\mu_{ik}} \frac{1}{u_k^{n-1}} W_k^n(z) - \sum_{k=1}^{K} \frac{1}{u_k^{n-1}} W_k^n(z) + o\Big(W_k^n(z)\Big) \\
&\quad + \sum_{i=1}^{I} \sum_{k\in K(i)}^{nb} \frac{\lambda_{ik}^{(\epsilon)}}{\mu_{ik}} \frac{1}{(u_k - d_{ik})^{n-1}} W_k^n(z) - \sum_{i=1}^{I} \sum_{k\in K(i)}^{nb} \frac{\lambda_{ik}^{(\epsilon)}}{\mu_{ik}} \frac{1}{u_k^{n-1}} W_k^n(z) \\
&= \sum_{i=1}^{I} \frac{1}{u_k^{n-1}} \left(\sum_{k\in K(i)} \frac{\lambda_{ik}^{(\epsilon)}}{\mu_{ik}} - 1 \right) W_k^n(z) + o\Big(W_k^n(z)\Big) \\
&\quad + \sum_{i=1}^{I} \sum_{k\in K(i)}^{nb} \frac{\lambda_{ik}^{(\epsilon)}}{\mu_{ik}} \left(\frac{d_{ik}}{u_k(u_k - d_{ik})} \right)^{n-1} W_k^n(z)
\end{aligned}
$$

where (c) is by considering basic and non-basic activities separately using (1.7) again. Now we discuss the last term above coming from non-basic activities. We will further show that there exist a set of $\lambda_{ik}^{(\epsilon)}$, s.t. $\lambda_{ik}^{(\epsilon)} = 0$ if activity (i,k) is non-basic activity. By our Assumption 1, under the heavy traffic (equivalently, the model is critical as defined in [11]), non-basic activity satisfies $x_{ik}^* = 0$ under any (if not unique) optimal solution of static allocation problem 1.3. This means the parallel-server system can achieve the heavy traffic without working through any non-basic activities. Denote $\lambda_{ik} \triangleq \frac{x_{ik}^*}{m_{ik}}$, then in [11] (Lemma 11.2), λ_{ik} satisfies $\lambda_i = \sum_{k\in K(i)} \lambda_{ik}$ and $\sum_{i\in I(k)} \lambda_{ik} m_{ik} = 1$. Suppose $x^{(\epsilon)}$ satisfying the constraints in the static allocation problem 1.3 with λ_i replaced by $\lambda_i^{(\epsilon)} = \lambda_i(1-\epsilon)$, and $\lambda_{ik}^{(\epsilon)} = \lambda_{ik}(1-\epsilon)$. Then $\lambda_{ik}^{(\epsilon)}$ satisfies all the subcritical condition in [11] (Lemma 11.2), and $\lambda_{ik}^{(\epsilon)} = 0$ for any non-basic activity. Furthermore, $\sum_{i\in I(k)} \lambda_{ik}^{(\epsilon)} m_{ik} = (1-\epsilon) < 1$. Therefore, we

use this set of $\lambda_{ik}^{(\epsilon)}$, and the generator then becomes

$$\begin{aligned} GV(z) &= \sum_{i=1}^{I} \frac{1}{u_k^{n-1}} \left(\sum_{k \in K(i)} \frac{\lambda_{ik}^{(\epsilon)}}{\mu_{ik}} - 1 \right) W_k^n(z) + o\Big(W_k^n(z)\Big) \\ &\leq -\frac{\epsilon}{\bar{u}^{n-1}} \sum_{i=1}^{I} W_k^n(z) + o\Big(W_k^n(z)\Big) \end{aligned}$$

where $\bar{u} = max_{k=1,...K} u_k > 0$. Then there exists some constant $c > 0$, such that for $\sum_{k=1}^{K} W_k^n(z) \geq c_w, k = 1, ...K$,

$$-\frac{\epsilon}{\bar{u}^{n-1}} \sum_{k=1}^{K} W_k^n(z) + o\Big(W_k^n(z)\Big) \leq -c\epsilon \sum_{i=1}^{I} W_k^n(z)$$

then $\exists d > 0$,

$$GV(z) \leq -c\epsilon \sum_{i=1}^{I} W_k^n(z) + d\mathbb{1}\Big(\sum_{i=1}^{I} W_k^n(z) < c_w \Big)$$

invoking [[26], Theorem 4.3], we have

$$\sum_{k=1}^{K} \mathbb{E}\Big[W_k^n\Big(Z^{(\epsilon)}(\infty)\Big)\Big] < \infty$$

□

A.2 Proofs in Section 1.7: Preliminary Results II

A.2.1 Proof of Lemma 11

Proof. 1. In (1.12), for non-basic activity (i, k), we have

$$\begin{aligned} & \mathbb{E}\left[d_{ik} \frac{1}{u_k} W_k(Z^{(\epsilon)}) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \quad (i) \\ \overset{(a)}{<} & \mathbb{E}\left[d_{ik} \frac{1}{u_k - d_{ik}} W_k(Z^{(\epsilon)}) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \quad (ii) \\ \overset{(b)}{\leq} & \mathbb{E}\left[d_{ik} \frac{1}{u_{k'}} W_{k'}(Z^{(\epsilon)}) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \quad (iii) \end{aligned}$$

where (a) is by $d_{ik} > 0$, (b) is because when (i,k) is non-basic activity, there must exist basic activity ik' according to complete resource pooling assumption. When $k = H^{(i)}(Z^{(\epsilon)})$, by WWTA policy, $\frac{1}{\mu_{ik}} W_k(Z^{(\epsilon)}) \leq \frac{1}{\mu_{ik'}} W_{k'}(Z^{(\epsilon)})$. If we divide both sides by v_i, it becomes $\frac{1}{u_k - d_{ik}} W_k(Z^{(\epsilon)}) \leq \frac{1}{u_{k'}} W_{k'}(Z^{(\epsilon)})$. Therefore, $(ii) - (i) \leq (iii) - (i)$ implies

$$
\begin{aligned}
0 <& \mathbb{E}\left[d_{ik} \frac{d_{ik}}{(u_k - d_{ik})u_k} W_k(Z^{(\epsilon)}) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
\leq& \mathbb{E}\left[d_{ik} \left(\frac{1}{u'_k} W_{k'}(Z^{(\epsilon)}) - \frac{1}{u_k} W_k(Z^{(\epsilon)}) \right) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
\leq& \mathbb{E}\left[d_{ik} \left[\max_{k\in\{1,\dots K\}} \frac{1}{u_k} W_k(Z^{(\epsilon)}) - \min_{k\in\{1,\dots K\}} \frac{1}{u_k} W_k(Z^{(\epsilon)}) \right] \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
\overset{(c)}{=}& d_{ik} \mathbb{E}\left[\left(\max_{k\in\{1,\dots K\}} T_k(Z^{(\epsilon)}) - \min_{k\in\{1,\dots K\}} T_k(Z^{(\epsilon)}) \right)^2 \right]^{1/2} \mathbb{P}\left(k = H^{(i)}(Z^{(\epsilon)})\right)^{1/2} \\
\overset{(d)}{=}& O(\epsilon^{1/2})
\end{aligned}
$$

where (c) is by Cauchy-Schwarz inequality, (d) uses Theorem 8 and Lemma 6 ((1.8)). Therefore, we have proved for non-basic activity (i,k),

$$
\mathbb{E}\left[W_k(Z^{(\epsilon)}) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] = O(\epsilon^{1/2}) \tag{A.1}
$$

Furthermore,

$$
\begin{aligned}
& \mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'} W_{k'}(Z^{(\epsilon)}) \right) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
=& \mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'} W_{k'}(Z^{(\epsilon)}) - \sum_{k'=1}^{K} u_{k'} W_k(Z^{(\epsilon)}) \right) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
& + \mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'} W_k(Z^{(\epsilon)}) \right) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
\overset{(e)}{\leq}& \mathbb{E}\left[\left| \sum_{k'=1}^{K} u_{k'} W_{k'}(Z^{(\epsilon)}) - \sum_{k'=1}^{K} u_{k'} W_k(Z^{(\epsilon)}) \right|^2 \right]^{1/2} \mathbb{P}\left(k = H^{(i)}(Z^{(\epsilon)})\right)^{1/2} \\
& + \sum_{k'=1}^{K} u_{k'} \mathbb{E}\left[W_k(Z^{(\epsilon)}) \mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
\overset{(f)}{=}& O(\epsilon^{1/2})
\end{aligned}
$$

where (e) is by Cauchy-Schwarz inequality, (f) uses Theorem 8, Lemma 6 ((1.8)) and A.1.

2. (1.13) and (1.14) will be proved together:

Let $f(z) = (\sum_{i=1}^{I} \sum_{k\in K(i)} v_i z_{ik})^2$, then the generator (1.6) becomes

$$\begin{aligned} Gf(z) = & 2\sum_{i=1}^{I}\sum_{k\in K(i)} \lambda_i^{(\epsilon)}\Big(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\Big)\mathbb{1}\left(k = H^{(i)}(z)\right) + \sum_{i=1}^{I}\sum_{k\in K(i)} \lambda_i^{(\epsilon)} v_i^2 \mathbb{1}\left(k = H^{(i)}(z)\right) \\ & - 2\sum_{k=1}^{K}\sum_{i\in I(k)} \mu_{ik} v_i P_{ik}(z)\Big(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i z_{ik}\Big) + \sum_{k=1}^{K}\sum_{i\in I(k)} \mu_{ik} v_i^2 P_{ik}(z) \end{aligned}$$

By Lemma 4(a), we have

$$\begin{aligned} & \sum_{k=1}^{K}\sum_{i\in I(k)} \mu_{ik} v_i \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\Big)\right] \\ = & \sum_{i=1}^{I}\sum_{k\in K(i)} \lambda_i^{(\epsilon)} v_i \mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] + \frac{1}{2}\sum_{i=1}^{I}\sum_{k\in K(i)} \lambda_i^{(\epsilon)} v_i^2 \mathbb{P}\left(k = H^{(i)}(Z^{(\epsilon)})\right) \\ & + \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in I(k)} \mu_{ik} v_i^2 \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\right] \end{aligned} \tag{A.2}$$

Let $f(z) = \left(\sum_{i=1}^{I} \sum_{k\in K(i)}^{b} v_i z_{ik}\right)^2$, i.e., we consider only the basic activities. Then the generator (1.6) becomes

$$\begin{aligned} Gf(z) = & 2\sum_{i=1}^{I}\sum_{k\in K(i)}^{b} \lambda_i^{(\epsilon)} v_i\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{b} v_i z_{ik}\Big)\mathbb{1}\left(k = H^{(i)}(z)\right) + \sum_{i=1}^{I}\sum_{k\in K(i)}^{b} \lambda_i^{(\epsilon)} v_i^2 \mathbb{1}\left(k = H^{(i)}(z)\right) \\ & - 2\sum_{k=1}^{K}\sum_{i\in I(k)}^{b} \mu_{ik} v_i P_{ik}(z)\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{b} z_{ik}\Big) + \sum_{k=1}^{K}\sum_{i\in I(k)}^{b} \mu_{ik} v_i^2 P_{ik}(z) \end{aligned}$$

By Lemma 4(a), we have

$$\begin{aligned} & \sum_{k=1}^{K}\sum_{i\in I(k)}^{b} \mu_{ik} v_i \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{b} v_i Z_{ik}^{(\epsilon)}\Big)\right] \\ = & \sum_{i=1}^{I}\sum_{k\in K(i)}^{b} \lambda_i^{(\epsilon)} v_i \mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{b} v_i Z_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] + \frac{1}{2}\sum_{i=1}^{I}\sum_{k\in K(i)}^{b} \lambda_i^{(\epsilon)} v_i^2 \mathbb{P}\left(k = H^{(i)}(Z^{(\epsilon)})\right) \\ & + \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in I(k)}^{b} \mu_{ik} v_i^2 \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\right] \end{aligned} \tag{A.3}$$

Then let (A.2) – (A.3), we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{i\in I(k)}^{nb}\mu_{ik}v_i\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}v_iZ_{ik}^{(\epsilon)}\Big)\right]\\
&+\sum_{k=1}^{K}\sum_{i\in I(k)}^{b}\mu_{ik}v_i\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\right]\\
=&\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i^{(\epsilon)}v_i\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k=H^{(i)}(Z^{(\epsilon)})\right)\right]\\
&+\sum_{i=1}^{I}\sum_{k\in K(i)}^{b}\lambda_i^{(\epsilon)}v_i\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k=H^{(i)}(Z^{(\epsilon)})\right)\right]\\
&+\frac{1}{2}\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i^{(\epsilon)}v_i^2\mathbb{P}\left(k=H^{(i)}(Z^{(\epsilon)})\right)+\frac{1}{2}\sum_{k=1}^{K}\sum_{i\in I(k)}^{nb}\mu_{ik}v_i^2\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\right]\\
=&\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i^{(\epsilon)}v_i\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k=H^{(i)}(Z^{(\epsilon)})\right)\right]\\
&+\sum_{i=1}^{I}\sum_{k\in K(i)}^{b}\lambda_i^{(\epsilon)}v_i\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k=H^{(i)}(Z^{(\epsilon)})\right)\right]+O(\epsilon)
\end{aligned}
\tag{A.4}
$$

where the last equality is by Lemma 5 and (1.8).

Let $f(z)=(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iz_{ik})^2$, then the generator (1.6) becomes

$$
\begin{aligned}
Gf(z)=&2\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i^{(\epsilon)}v_i\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iz_{ik}\Big)\mathbb{1}\left(k=H^{(i)}(z)\right)+\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i^{(\epsilon)}v_i^2\mathbb{1}\left(k=H^{(i)}(z)\right)\\
&-2\sum_{k=1}^{K}\sum_{i\in I(k)}^{nb}\mu_{ik}v_iP_{ik}(z)\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iz_{ik}\right)+\sum_{k=1}^{K}\sum_{i\in I(k)}^{nb}\mu_{ik}v_i^2P_{ik}(z)
\end{aligned}
$$

By Lemma 4(a), and Lemma 5 with (1.8), we have

$$
\begin{aligned}
&\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\mu_{ik}v_i\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\right)\right]\\
&-\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_iv_i\mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\right)\mathbb{1}\left(k=H^{(i)}(Z^{(\epsilon)})\right)\right]\\
=&\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i^{(\epsilon)}v_i^2\mathbb{P}\left(k=H^{(i)}(Z^{(\epsilon)})\right)+\sum_{k=1}^{K}\sum_{i\in I(k)}^{nb}\mu_{ik}v_i^2\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\right]\\
=&O(\epsilon)
\end{aligned}
\tag{A.5}
$$

Now, adding (A.4) to (A.5),we have

$$\sum_{k=1}^{K}\sum_{i\in I(k)}^{nb}\mu_{ik}v_i\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}v_iZ_{ik}^{(\epsilon)}\Big)\right]+\sum_{k=1}^{K}\sum_{i\in I(k)}\mu_{ik}v_i\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\right]$$
$$=\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i^{(\epsilon)}v_i\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k=H^{(i)}(Z^{(\epsilon)})\right)\right]$$
$$+\sum_{i=1}^{I}\sum_{k\in K(i)}\lambda_i^{(\epsilon)}v_i\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k=H^{(i)}(Z^{(\epsilon)})\right)\right]+O(\epsilon)$$

then by rearranging some terms,

$$\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i^{(\epsilon)}v_i\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k=H^{(i)}(Z^{(\epsilon)})\right)\right]$$
$$-\sum_{k=1}^{K}\sum_{i\in I(k)}^{nb}\mu_{ik}v_i\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}v_iZ_{ik}^{(\epsilon)}\Big)\right]+O(\epsilon)$$
$$=\sum_{k=1}^{K}\sum_{i\in I(k)}\mu_{ik}v_i\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\right]$$
$$-\sum_{i=1}^{I}\sum_{k\in K(i)}\lambda_i^{(\epsilon)}v_i\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(k=H^{(i)}(Z^{(\epsilon)})\right)\right]$$
$$=\sum_{k=1}^{K}u_k\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\right]-\sum_{k=1}^{K}\sum_{i\in I(k)}d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\right]$$
$$-\sum_{i=1}^{I}\lambda_i^{(\epsilon)}v_i\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\right]-\sum_{k=1}^{K}u_k\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(\sum_{i\in I(k)}Z_{ik}^{(\epsilon)}=0\right)\right]$$
$$=\epsilon\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\right]-\sum_{k=1}^{K}\sum_{i\in I(k)}d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\right]$$
$$-\sum_{k=1}^{K}u_k\mathbb{E}\left[\Big(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}v_iZ_{ik}^{(\epsilon)}\Big)\mathbb{1}\left(\sum_{i\in I(k)}Z_{ik}^{(\epsilon)}=0\right)\right]$$

note that terms with d_{ik} exist only when (i,k) is non-basic activity, then we

have

$$
\begin{aligned}
&\epsilon\mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}\right)\right] \\
&= \sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} \lambda_i^{(\epsilon)} v_i \mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)\mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
&\quad - \sum_{k=1}^{K}\sum_{i\in I(k)}^{nb} \mu_{ik} v_i \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{k'=1}^{K} u_{k'} W_{k'}(Z^{(\epsilon)})\right)\right] \\
&\quad + \sum_{k=1}^{K}\sum_{i\in I(k)}^{nb} \mu_{ik} v_i \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} d_{ik} Z_{ik}^{(\epsilon)}\right)\right] \\
&\quad + \sum_{k=1}^{K}\sum_{i\in I(k)}^{nb} d_{ik} \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}\right)\right] \\
&\quad + \sum_{k=1}^{K} u_k \mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}\right)\mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] + O(\epsilon)
\end{aligned}
\tag{A.6}
$$

Now we discuss each term on the RHS of (A.6). The first term by 1.15 and (1.12) becomes:

$$
\begin{aligned}
&\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} \lambda_i^{(\epsilon)} v_i \mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)\mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
\leq &\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} \lambda_i^{(\epsilon)} v_i \mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'} W_{k'}(Z^{(\epsilon)})\right)\mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] \\
= &O(\epsilon^{1/2})
\end{aligned}
$$

For the third term and the fourth term, by (A.5) and 1.15 ,

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{i\in I(k)}^{nb} \mu_{ik} v_i \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} d_{ik} Z_{ik}^{(\epsilon)}\right)\right] \\
= &\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} \lambda_i v_i \mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}\right)\mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] + O(\epsilon) \\
\leq &\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} \lambda_i v_i \mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'} W_{k'}(Z^{(\epsilon)})\right)\mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] + O(\epsilon) \\
= &O(\epsilon^{1/2})
\end{aligned}
$$

and there exists $C > 0$,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{i\in I(k)}^{nb} d_{ik}\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}\right)\right] \\
\leq & C\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i v_i\mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}\right)\mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] + O(\epsilon) \\
\leq & C\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb}\lambda_i v_i\mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'}W_{k'}(Z^{(\epsilon)})\right)\mathbb{1}\left(k = H^{(i)}(Z^{(\epsilon)})\right)\right] + O(\epsilon) \\
= & O(\epsilon^{1/2})
\end{aligned}$$

where the each last equality is just proved as first term. For the fifth term, by Lemma 10, we have

$$\begin{aligned}
&\sum_{k=1}^{K} u_k\mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}\right)\mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
\leq & \sum_{k=1}^{K} u_k\mathbb{E}\left[\left(\sum_{k'=1}^{K} u_{k'}W_{k'}(Z^{(\epsilon)})\right)\mathbb{1}\left(\sum_{i\in I(k)} Z_{ik}^{(\epsilon)} = 0\right)\right] \\
= & O(\epsilon^{1/2})
\end{aligned}$$

Then (A.6) becomes

$$\epsilon\mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}\right)\right] \leq O(\epsilon^{1/2}) - \sum_{k=1}^{K}\sum_{i\in I(k)}^{nb}\mu_{ik}v_i\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{k'=1}^{K} u_{k'}W_{k'}(Z^{(\epsilon)})\right)\right] \tag{A.7}$$

The LHS of (A.7) is always non-negative, therefore, we have proved the 1.13, that is, for each non-basic activity (i,k),

$$\mathbb{E}\left[P_{ik}(Z^{(\epsilon)})\left(\sum_{k'=1}^{K} u_{k'}W_{k'}(Z^{(\epsilon)})\right)\right] = O(\epsilon^{1/2})$$

therefore, at the same time,

$$\epsilon\mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{k\in K(i)}^{nb} v_i Z_{ik}^{(\epsilon)}\right)\right] = O(\epsilon^{1/2})$$

□

A.3 Proofs in Section 1.8

A.3.1 Proof of Corollary 13

Proof. We omit (∞) in $Z^{(\epsilon)}(\infty)$ for simplicity. Denote

$$\tilde{X}^{(\epsilon)} \triangleq \epsilon\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)$$

Them by Theorem 12,

$$\tilde{X}^{(\epsilon)} \xrightarrow{d} \tilde{X}, \quad as\ \epsilon \downarrow 0 \tag{A.8}$$

Denote

$$U^{(\epsilon)} \triangleq \epsilon\left(\sum_{k=1}^{K} u_k W_k(Z^{(\epsilon)})\right)$$

then by Lemma 11 and (1.15), the difference

$$E\left[U^{(\epsilon)} - \tilde{X}^{(\epsilon)}\right] = \epsilon\mathbb{E}\left[\sum_{i=1}^{I}\sum_{k\in K(i)} \frac{d_{ik}}{\mu_{ik}} Z_{ik}^{(\epsilon)}\right] \to 0, \quad as\ \epsilon \downarrow 0$$

i.e.

$$\epsilon\left|U^{(\epsilon)} - \tilde{X}^{(\epsilon)}\right| \xrightarrow{L^1} 0, \quad as\ \epsilon \downarrow 0$$

which means

$$\epsilon\left|U^{(\epsilon)} - \tilde{X}^{(\epsilon)}\right| \xrightarrow{p} 0, \quad as\ \epsilon \downarrow 0 \tag{A.9}$$

Combining (A.8) with (A.9), we have

$$U^{(\epsilon)} \xrightarrow{d} \tilde{X}, \quad as\ \epsilon \downarrow 0$$

that is,

$$\epsilon\left(\sum_{k=1}^{K} u_k W_k(Z^{(\epsilon)})\right) \xrightarrow{d} \tilde{X}, \quad as\ \epsilon \downarrow 0$$

□

A.3.2 Proof Lemma 16

We omit (∞) in $Z^{(\epsilon)}(\infty)$ for simplicity.

Proof. For $\forall i = 1, ...I$, we have

$$
\begin{aligned}
&\lim_{\epsilon\downarrow 0}\left(\lambda_i v_i \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] - \sum_{k\in K(i)} u_k \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right]\right) \\
&= \lim_{\epsilon\downarrow 0}\mathbb{E}\Bigg[\lim_{t\downarrow 0}\Bigg(\lambda_i v_i e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \\
&\quad - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\Bigg)\Bigg] \\
&\overset{(a)}{=} \lim_{\epsilon\downarrow 0}\lim_{t\downarrow 0}\mathbb{E}\Bigg[\lambda_i v_i e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \\
&\quad - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\Bigg] \\
&\overset{(b)}{=} \lim_{t\downarrow 0}\lim_{\epsilon\downarrow 0}\mathbb{E}\Bigg[\lambda_i v_i e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \\
&\quad - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\Bigg] \\
&\overset{(c)}{=} \lim_{t\downarrow 0} 0 = 0
\end{aligned}
\tag{A.10}
$$

where (a) is by bounded convergence theorem, (b) holds by Moore-Osgood Theorem([16]), (c) is directly by case (a), (1.21). The detailed proofs for $(a), (b), (c)$ are listed as follows:

(a) In this part we will show the interchange of the limit and expectation.

Since $0 < \lambda_i v_i \leq 1$, $0 < u_k \leq 1$, $0 \leq P_{ik}(Z^{(\epsilon)}) \leq 1$ almost surely, then

$$
\mathbb{E}\left[\left(\lambda_i v_i - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)})\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \leq 2
$$

Therefore, by Bounded Convergence Theorem, we have

$$\mathbb{E}\left[\lim_{t\downarrow 0}\left(\lambda_i v_i - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)})\right)e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right]$$
$$= \lim_{t\downarrow 0}\mathbb{E}\left[\left(\lambda_i v_i - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)})\right)e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right]$$

(b) In this part we will show the interchange of the limit w.r.t. t and ϵ.

We first introduce Moore-Osgood Theorem in [16, Theorem 2]

Theorem 34 (Moore-Osgood). *If* $\lim_{x\to p} f(x,y)$ *exists point-wise for each* y *different from* q *and if* $\lim_{y\to q} f(x,y)$ *converges uniformly for* $x \neq p$ *then the double limit and the iterated limits exist and are equal, i.e.*

$$\lim_{(x,y)\to(p,q)} f(x,y) = \lim_{x\to p}\lim_{y\to q} f(x,y) = \lim_{y\to q}\lim_{x\to p} f(x,y)$$

Let $g(\epsilon,t)$ be a function w.r.t ϵ and t:

$$g_i(\epsilon,t) = \lambda_i v_i \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right]$$
$$- \sum_{k\in K(i)} u_k \mathbb{E}\left[P_{ik}(Z^{(\epsilon)})e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right], i = 1,...I$$

Now we discuss the interchange of limit w.r.t ϵ and t. First, by (1.21) , $\lim_{\epsilon\to 0} g_i(\epsilon,t)$ exists point-wise for $t \neq 0$. Next we present the following Lemma 35 to check the second condition of Moore-Osgood Theorem, and the proof is put in the Appendix A.3.3.

Lemma 35. $\lim_{t\to 0} g_i(\epsilon,t)$ *converges uniformly for* $\epsilon \neq 0$, *for* $i = 1,...I$.

Then the conditions for Moore-Osgood Theorem 34 are satisfied, therefore

the limits can be interchanged, i.e.

$$
\begin{aligned}
&\lambda_i v_i \lim_{\epsilon \downarrow 0} \lim_{t \downarrow 0} \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I} \sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t \sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \right] \\
&- \lim_{\epsilon \downarrow 0} \lim_{t \downarrow 0} \sum_{k\in K(i)} u_k \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I} \sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t \sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \right] \\
=&\lambda_i v_i \lim_{t \downarrow 0} \lim_{\epsilon \downarrow 0} \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I} \sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t \sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \right] \\
&- \lim_{t \downarrow 0} \lim_{\epsilon \downarrow 0} \sum_{k\in K(i)} u_k \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I} \sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t \sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \right] \\
=&\lim_{t \downarrow 0} 0 = 0
\end{aligned}
$$

□

A.3.3 Proof of lemma 35

Proof of lemma (35). This is to prove $\forall \eta > 0$, $\exists \delta_i > 0$, when $|t - 0| < \delta_i$, $\forall \epsilon \neq 0$ being sufficiently small, $|g_i(\epsilon, t) - g_i(\epsilon, 0)| < \eta$.

For $i = 1, ...I$,

$$
\begin{aligned}
&|g_i(\epsilon, t) - g_i(\epsilon, 0)| \\
=&\Bigg|\lambda_i v_i \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] - \lambda_i v_i \mathbb{E}\left[e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&- \sum_{k\in K(i)} u_k \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)} + t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right] \\
&+ \sum_{k\in K(i)} u_k \mathbb{E}\left[P_{ik}(Z^{(\epsilon)}) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right]\Bigg| \\
=&\left|\mathbb{E}\left[\left(\lambda_i v_i - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)})\right) e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\left(e^{\epsilon\theta v_i t Z_{ik}^{(\epsilon)}} - 1\right)\right]\right| \\
\leq&\mathbb{E}\left[\left|\left(\lambda_i v_i - \sum_{k\in K(i)} u_k P_{ik}(Z^{(\epsilon)})\right)\right| \left|e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)}\right| \left|\left(e^{\epsilon\theta t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}} - 1\right)\right|\right] \\
\overset{(a)}{\leq}&2\mathbb{E}\left|\left(e^{\epsilon\theta t\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}} - 1\right)\right| \\
\overset{(b)}{\leq}&2|\theta| t\epsilon \sum_{k\in K(i)} v_i \mathbb{E}\left[Z_{ik}^{(\epsilon)}\right] \\
\overset{(c)}{\leq}&2|\theta| v_i t M_0
\end{aligned}
$$

where (a) is by the following: $e^{\epsilon\theta\left(\sum_{i=1}^{I}\sum_{k\in K(i)} v_i Z_{ik}^{(\epsilon)}\right)} \leq 1$, almost surely, since $\theta \leq 0$; $0 \leq P_{ik}(Z^{(\epsilon)}) \leq 1$; and $0 < \lambda_i v_i \leq 1$; $\sum_{k=1}^{K} u_k = 1$. (b) is by Lemma 25; (c) is by Lemma 4(b) with the existence of constant $M_0 > 0$. Hence, we let $\delta_i = \frac{\eta}{3v_i M_0|\theta|}$. When $|t| < \delta_i$,

$$|g_i(\epsilon, t) - g_i(\epsilon, 0)| \leq 2v_i M_0|\theta|\frac{\eta}{3v_i M_0|\theta|} < \eta$$

Therefore, $\lim_{t\downarrow 0} g_i(\epsilon, t)$ converges uniformly for $\epsilon > 0$, $i = 1, ...I$. □

APPENDIX B

APPENDIX OF CHAPTER 2

B.1 Proof of Lemma 23

Remark 8 (Lemma 21)**.** *Following Lemma 4 of [33] and Lemma 1 of [5], a sufficient condition for Lemma 21 is to require that*

$$\mathbb{E}\left[\left|G_{Z^{(r)}}(Z^{(r)}, Z^{(r)}) f(Z^{(r)})\right|\right] < \infty,$$

where $G_z(z, z)$ is the diagonal entry of the generator matrix G_z corresponding to the state z. Since there exist some constant $M > 0$, such that

$$|G_{Z^{(r)}}(Z^{(r)}, Z^{(r)})| \leq \sum_{i\in\mathcal{J}} \left(\alpha_i + \sum_{j\in\mathcal{J}} \mu_j^{(r)} P_{ji}\right) + \sum_{i\in\mathcal{J}} \left(\mu_i^{(r)} + \sum_{j\in\{0\}\cup\mathcal{J}} \mu_i^{(r)} P_{ij}\right) \leq M,$$

and $f(z) \leq (\sum_{j\in\mathcal{J}} c_j z_j)^n$ for some $n \in \mathbb{N}_+$, it suffices to show that $\mathbb{E}[(\sum_{j\in\mathcal{J}} c_j Z_j^{(r)})^n] < \infty$, for $r \in (0, r_0)$. Following that it suffices to show there exist some constants $c, d > 0$ such that for $(\sum_{j\in\mathcal{J}} c_j z_j)^n \geq c_z$ with constant $c_z > 0$,

$$Gf(z) \leq -c(\sum_{j\in\mathcal{J}} c_j z_j)^n + d\mathbb{1}\Big((\sum_{j\in\mathcal{J}} c_j z_j)^n < c_z\Big). \tag{B.1}$$

Proof of Lemma 23. Proof sketch: the proof will start with the lightest station 1 with the first moment. Then the proof will be carried over through Mathematical induction, which including two directions as follows: the first direction is given the low order moment boundedness, to prove the high order moment boundedness. The second direction is given the lighter station moment boundedness, to prove the heavier station moment boundedness.

(I) This step will prove moment boundedness by induction for station 1, that is to show $\mathbb{E}[(rZ_1^{(r)})^n] \leq M_n^{(1)}, \forall n \in \mathbb{N}_+, r \in (0, r_1)$ for some $r_1 \in (0, 1)$.

(i) For $n = 1$, we let $f(z) = z_1^2$, then the generator 2.10 becomes

$$\begin{aligned}
Gf(z) =&\alpha_1(2z_1+1) + \mu_1(1-P_{11})(-2z_1+1)\mathbb{1}(z_1>0) \\
&+ \sum_{i\in\mathcal{J}\setminus\{1\}} \mu_i^{(r)} P_{i1}(2z_1+1)\mathbb{1}(z_i>0) \\
=&2e^{(1)\prime}(1-P')\lambda z_1 - 2e^{(1)\prime}(1-P')\begin{pmatrix}\mu_1^{(r)}\mathbb{1}(z_1>0)\\ \vdots \\ \mu_J^{(r)}\mathbb{1}(z_k>0)\end{pmatrix} z_1 \\
&+ \alpha_1 + \mu_1^{(r)}(1-P_{11})\mathbb{1}(z_1>0) + \sum_{i\in\mathcal{J}\setminus\{1\}} \mu_i^{(r)} P_{i1}\mathbb{1}(z_i>0) \\
=&2z_1 e^{(1)\prime}(1-P')\begin{pmatrix}-r\mu_1^{(r)}\\ \vdots \\ -r^J\mu_J^{(r)}\end{pmatrix} + 2z_1 e^{(1)\prime}(1-P')\begin{pmatrix}\mu_1^{(r)}\mathbb{1}(z_1=0)\\ \vdots \\ \mu_J^{(r)}\mathbb{1}(z_k=0)\end{pmatrix} \\
&+ \alpha_1 + \mu_1^{(r)}(1-P_{11})\mathbb{1}(z_1>0) + \sum_{i\in\mathcal{J}\setminus\{1\}} \mu_i^{(r)} P_{i1}\mathbb{1}(z_i>0) \\
\le&2z_1\Big[-r\mu_1^{(r)}(1-P_{11}) + \sum_{i\in\mathcal{J}\setminus\{1\}} r^i\mu_i^{(r)}P_{i1}\Big] - 2z_1\sum_{i\in\mathcal{J}\setminus\{1\}}\mu_i^{(r)}P_{i1}\mathbb{1}(z_i=0) \\
&+ \alpha_1 + \mu_1^{(r)} + \sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{i1} \\
\le&-2rz_1[\mu_1^{(r)}(1-P_{11}) - \sum_{i\in\mathcal{J}\setminus\{1\}} r^{i-1}\mu_i^{(r)}P_{i1}] + \alpha_1 + \mu_1^{(r)} + \sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{i1},
\end{aligned}$$

where there exists $r_1 \in (0,1)$ such that when $r \in (0, r_1)$,

$$\mu_1^{(r)}(1-P_{11}) - \sum_{i\in\mathcal{J}\setminus\{1\}} r^{i-1}\mu_i^{(r)}P_{i1} > 0.$$

Then (B.1) is satisfied, so Lemma 21 holds with $n = 1$. Therefore, taking expectation on both sides gives

$$\mathbb{E}[rZ_1^{(r)}] \le \frac{\alpha_1 + \mu_1^{(r)} + \sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{i1}}{2[\mu_1^{(r)}(1-P_{11}) - \sum_{i\in\mathcal{J}\setminus\{1\}} r^{i-1}\mu_i^{(r)}P_{i1}]} < M_1^{(1)},$$

where $M_1^{(1)} > 0$ is constant which does not depend on r.

(ii) Suppose for $n = \ell$, $\mathbb{E}[(rZ_1^{(r)})^\ell] \le M_\ell^{(1)}$ has been proved for some constant $M_\ell^{(1)} > 0$. Now we will prove the case for $n = \ell + 1$. Let $f(z) = z_1^{\ell+2}$. Using binomial expansion, the generator 2.10 becomes

$$\begin{aligned} Gf(z) =& \alpha_1[(\ell+2)z_1^{\ell+1} + R(z)] + \mu_1^{(r)}(1-P_{11})(-(\ell+2)z_1 + R(z))\mathbb{1}(z_1 > 0) \\ &+ \sum_{i\in\mathcal{J}\setminus\{1\}} \mu_i^{(r)} P_{i1}((\ell+2)z_1 + R(z))(z_i > 0), \end{aligned}$$

where $R(z) \le \sum_{i=0}^{\ell} C_i z_1^i$ includes the terms w.r.t z_1 whose powers are lower than $\ell+1$ and C_i are some constants. Then one has

$$\begin{aligned} Gf(z) =& (\ell+2)z_1^{\ell+1} e^{(1)\prime}(1-P') \begin{pmatrix} -r\mu_1^{(r)} \\ \vdots \\ -r^J \mu_J^{(r)} \end{pmatrix} \\ &+ (\ell+2)z_1^{\ell+1} e^{(1)\prime}(1-P') \begin{pmatrix} \mu_1^{(r)}\mathbb{1}(z_1=0) \\ \vdots \\ \mu_J^{(r)}\mathbb{1}(z_J=0) \end{pmatrix} + R(z) \\ =& -(\ell+2)rz_1^{\ell+1}[\mu_1(1-P_{11}) - \sum_{i\in\mathcal{J}\setminus\{1\}} r^{i-1}\mu_i^{(r)}P_{i1}] \\ &- (\ell+2)z_1^{\ell+1} \sum_{i\in\mathcal{J}\setminus\{1\}} \mu_i^{(r)} P_{i1}\mathbb{1}(z_i=0) + R(z) \\ \le& -(\ell+2)rz_1^{\ell+1}[\mu_1(1-P_{11}) - \sum_{i\in\mathcal{J}\setminus\{1\}} r^{i-1}\mu_i^{(r)}P_{i1}] + R(z). \end{aligned}$$

Since $\mu_1(1-P_{11}) - \sum_{i\in\mathcal{J}\setminus\{1\}} r^{i-1}\mu_i^{(r)}P_{i1} > 0, r \in (0, r_1)$, there exists $c > 0, c_z > 0, d > 0$, such that

$$Gf(z) \le -cz_1^{\ell+1} + d\mathbb{1}(z_1^{\ell+1} < c_z).$$

Then (B.1) is satisfied, so Lemma 21 holds with $n = \ell+1$. Furthermore, by induction hypothesis, there exists constant $M > 0$, s.t.

$$r^{\ell+1}\mathbb{E}[R(Z^{(r)})] = \sum_{i=0}^{\ell} r^{\ell+1-i} C_i \mathbb{E}[(rZ_1^{(r)})^i] \le M. \tag{B.2}$$

Therefore, With Lemma 21, multiply r^ℓ on both sides above, one has

$$\mathbb{E}[(rZ_1^{(r)})^{\ell+1}] \le \frac{r^{\ell+1}\mathbb{E}[R(Z^{(r)})]}{(\ell+2)[\mu_1(1-P_{11}) - \sum_{i\in\mathcal{T}\setminus\{1\}} r^{i-1}\mu_i^{(r)}P_{i1}]} \le M_{\ell+1}^{(1)},$$

where $M_{\ell+1}^{(1)} > 0$ is some constant and the last inequality is by (B.2) and take $r \in (0, r_1)$.

(II) Step (I) has proved the moment boundedness for station $i = 1$. Now using induction again, Step (II) will go through the moment boundedness from station 1 to all heavier stations. Suppose when $i = m$ and there exists $r_m \in (0,1)$, $\mathbb{E}[(r^m Z_m^{(r)})^n] \le M_n^{(m)}, \forall n \in \mathbb{N}_+, r \in (0, \min_{i\le m} r_i)$ has been proved. Now let $i = m + 1$.

(i) For $n = 1$, let

$$f(z) = (t^{(m+1)\prime} z)^2,$$

where $t^{(m+1)\prime} = (w_{1,m+1}, \ldots, w_{m,m+1}, 1, 0, ...0)$ with $\{w_{ij}\}, i, j \in \mathcal{J}$ defined in

Lemma 17. Then the generator 2.10 becomes

$$
\begin{aligned}
Gf(z) =& 2(t^{(m+1)\prime}z)t^{(m+1)\prime}\alpha + \sum_{i\le m+1}\alpha_i t_i^2 \\
&+ 2(t^{(m+1)\prime}z)t^{(m+1)\prime}\sum_{j\in\{0\}\cup\mathcal{J}}\sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{ij}(-e^{(i)}+e^{(j)})\mathbb{1}(z_i>0) \\
&+ \sum_{j\in\{0\}\cup\mathcal{J}}\sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{ij}(t^{(m+1)\prime}(-e^{(i)}+e^{(j)}))^2\mathbb{1}(z_i>0) \\
\le& 2(t^{(m+1)\prime}z)t^{(m+1)\prime}(1-P')\lambda - 2(t^{(m+1)\prime}z)t^{(m+1)\prime}(1-P')\begin{pmatrix}\mu_1^{(r)}\mathbb{1}(z_1>0)\\ \vdots \\ \mu_J^{(r)}\mathbb{1}(z_J>0)\end{pmatrix} \\
&+ \sum_{i\le m+1}\alpha_i t_i^2 + \sum_{j\in\{0\}\cup\mathcal{J}}\sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{ij}(t^{(m+1)\prime}(-e^{(i)}+e^{(j)}))^2 \\
=& 2(t^{(m+1)\prime}z)t^{(m+1)\prime}(1-P')\begin{pmatrix}-r\mu_1^{(r)}\\ \vdots \\ -r^J\mu_J^{(r)}\end{pmatrix} + 2(t^{(m+1)\prime}z)t^{(m+1)\prime}(1-P')\begin{pmatrix}\mu_1^{(r)}\mathbb{1}(z_1=0)\\ \vdots \\ \mu_J^{(r)}\mathbb{1}(z_J=0)\end{pmatrix} \\
&+ \sum_{i\le m+1}\alpha_i t_i^2 + \sum_{j\in\{0\}\cup\mathcal{J}}\sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{ij}(t^{(m+1)\prime}(-e^{(i)}+e^{(j)}))^2.
\end{aligned}
$$

Note the structure of $t^{(m+1)}$ and $(1-P')$, and

$$
t^{(m+1)\prime}(1-P') = (w_{1,m+1},\dots,w_{m,m+1},1,0,\dots 0)\begin{pmatrix}0 & \cdots & 0 & -P_{m+1,1} & \cdots & -P_{J1}\\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 1-P_{m+1,m+1} & \cdots & -P_{J,m+1}\\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0\end{pmatrix},
$$

then one has

$$
\begin{aligned}
Gf(z) =& -2(t^{(m+1)\prime}z)\Bigg\{r^{m+1}\mu_{m+1}^{(r)}[(1-P_{m+1,m+1})-\sum_{i<m+1}P_{m+1,i}w_{i,m+1}] \\
&-\sum_{i>m+1}r^i\mu_i^{(r)}\Big(\sum_{j<m+1}P_{ij}w_{j,m+1}+P_{i,m+1}\Big)\Bigg\} \\
&+2(t^{(m+1)\prime}z)(1-P_{m+1,m+1})\mu_{m+1}^{(r)}\mathbb{1}(z_{m+1}=0) \\
&-2(t^{(m+1)\prime}z)(\sum_{i<m+1}w_{i,m+1}P_{m+1,i})\mu_{m+1}^{(r)}\mathbb{1}(z_{m+1}=0) \\
&-2(t^{(m+1)\prime}z)(\sum_{j>m+1}(\sum_{i<m+1}w_{i,m+1}P_{ji})\mu_j^{(r)}\mathbb{1}(z_j=0)) \\
&+\sum_{i\le m+1}\alpha_i t_i^2+\sum_{j\in\{0\}\cup\mathcal{J}}\sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{ij}(t^{(m+1)\prime}(-e^{(i)}+e^{(j)}))^2 \\
\le& -2(t^{(m+1)\prime}z)\Bigg\{r^{m+1}\mu_{m+1}^{(r)}[(1-P_{m+1,m+1})-\sum_{i<m+1}P_{m+1,i}w_{i,m+1}] \\
&-\sum_{i>m+1}r^i\mu_i^{(r)}\Big(\sum_{j<m+1}P_{ij}w_{j,m+1}+P_{i,m+1}\Big)\Bigg\} \\
&+2(t^{(m+1)\prime}z)(1-P_{m+1,m+1})\mu_{m+1}^{(r)}\mathbb{1}(z_{m+1}=0) \\
&+\sum_{i\le m+1}\alpha_i t_i^2+\sum_{j\in\{0\}\cup\mathcal{J}}\sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{ij}(t^{(m+1)\prime}(-e^{(i)}+e^{(j)}))^2,
\end{aligned}
$$

where the last two terms can be bounded by some constant $C>0$:

$$
\sum_{i\le m+1}\alpha_i t_i^2+\sum_{j\in\{0\}\cup\mathcal{J}}\sum_{i\in\mathcal{J}}\mu_i^{(r)}P_{ij}(t^{(m+1)\prime}(-e^{(i)}+e^{(j)}))^2\le C.
$$

Plugging (2.6), the generator becomes

$$
\begin{aligned}
Gf(z) \le& -2(t^{(m+1)\prime}z)\Bigg\{r^{m+1}\mu_{m+1}^{(r)}(1-w_{m+1,m+1})-\sum_{i>m+1}r^i\mu_i^{(r)}w_{i,m+1}\Bigg\} \\
&+2(t^{(m+1)\prime}z)(1-P_{m+1,m+1})\mu_{m+1}^{(r)}\mathbb{1}(z_{m+1}=0)+C.
\end{aligned}
$$

Now we choose $r_{m+1}\in(0,1)$ such that as $r\in(0,\min_{i\le m+1}r_i)$,

$$
\mu_{m+1}^{(r)}(1-w_{m+1,m+1})-\sum_{i>m+1}r^{i-m-1}\mu_i^{(r)}w_{i,m+1}>0,
$$

then (B.1) is satisfied, so Lemma 21 holds for $(\sum_{j\in\mathcal{J}}c_jz_j)^n=(t^{(m+1)\prime}z)^n$ with $n=1$.

Therefore, one can takes the expectation on both sides and obtain

$$\begin{aligned}
&r^{m+1}\Big(\sum_{i<m+1} w_{i,m+1}\mathbb{E}[Z_i^{(r)}] + \mathbb{E}[Z_{m+1}^{(r)}]\Big) \\
\leq &\frac{(1-P_{m+1,m+1})\mu_{m+1}^{(r)}\mathbb{E}[(t^{(m+1)\prime}Z^{(r)})\mathbb{1}(Z_{m+1}^{(r)}=0)] + \frac{C}{2}}{\mu_{m+1}^{(r)}(1-w_{m+1,m+1}) - \sum_{i>m+1} r^{i-m-1}\mu_i^{(r)}w_{i,m+1}}
\end{aligned}$$

Besides, by Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\Big[(t^{(m+1)\prime}Z^{(r)})\mathbb{1}(Z_{m+1}^{(r)}=0)\Big] = \sum_{i<m+1} w_{i,m+1}\mathbb{E}[Z_i^{(r)}\mathbb{1}(Z_{m+1}^{(r)}=0)] \\
&= \sum_{i<m+1} w_{i,m+1}\mathbb{E}[r^i Z_i^{(r)}\frac{\mathbb{1}(Z_{m+1}^{(r)}=0)}{r^i}] \\
&\leq \sum_{i<m+1} w_{i,m+1}\mathbb{E}[(r^i Z_i^{(r)})^{m+1}]^{\frac{1}{m+1}}\frac{\mathbb{P}(Z_{m+1}^{(r)}=0)^{\frac{m}{m+1}}}{r^i} \\
&\leq \sum_{i<m+1} w_{i,m+1}M_{m+1}^{(i)\frac{1}{m+1}}\frac{r^m}{r^i} = w_{m,m+1}M_{m+1}^{(m)\frac{1}{m+1}} + \sum_{i<m} w_{i,m+1}M_{m+1}^{(i)\frac{1}{m+1}}r^{m-i},
\end{aligned}$$

where the last inequality is by Lemma 22 and induction hypothesis at the beginning of step (II). Therefore, one has

$$\begin{aligned}
&r^{m+1}\mathbb{E}[Z_{m+1}^{(r)}] \leq r^{m+1}\Big(\sum_{i<m+1} w_{i,m+1}\mathbb{E}[Z_i^{(r)}] + \mathbb{E}[Z_{m+1}^{(r)}]\Big) \\
\leq &\frac{(1-P_{m+1,m+1})\mu_{m+1}^{(r)}\Big[w_{m,m+1}M_{m+1}^{(m)\frac{1}{m+1}} + \sum_{i<m} w_{i,m+1}M_{m+1}^{(i)\frac{1}{m+1}}r^{m-i}\Big] + \frac{C}{2}}{\mu_{m+1}^{(r)}(1-w_{m+1,m+1}) - \sum_{i>m+1} r^{i-m-1}\mu_i^{(r)}w_{i,m+1}} \triangleq M_1^{(m+1)}.
\end{aligned}$$

(ii) Suppose for $n=\ell$, $\mathbb{E}[(r^{m+1}Z_{m+1}^{(r)})^\ell] \leq M_\ell^{(m+1)}$ has been proved. Now we will prove the case for $n=\ell+1$. Let

$$f(z) = (t^{(m+1)\prime}z)^{\ell+2},$$

where $t^{(m+1)\prime} = (w_{1,m+1},\ldots,w_{m,m+1},1,0,...0)$ is the same as before. Then

with binomial expansion, the generator becomes

$$\begin{aligned}
Gf(z) =&(\ell+2)(t^{(m+1)\prime}z)^{\ell+1}t^{(m+1)\prime}\alpha + R(z) \\
&- (\ell+2)(t^{(m+1)\prime}z)^{\ell+1}t^{(m+1)\prime}(1-P')\begin{pmatrix}\mu_1^{(r)}\mathbb{1}(z_1>0)\\ \vdots \\ \mu_J^{(r)}\mathbb{1}(z_J>0)\end{pmatrix} \\
=&(\ell+2)(t^{(m+1)\prime}z)^{\ell+1}t^{(m+1)\prime}(1-P')\begin{pmatrix}-r\mu_1^{(r)}\\ \vdots \\ -r^J m u_J^{(r)}\end{pmatrix} \\
&+(\ell+2)(t^{(m+1)\prime}z)^{\ell+1}t^{(m+1)\prime}(1-P')\begin{pmatrix}\mu_1^{(r)}\mathbb{1}(z_1=0)\\ \vdots \\ \mu_J^{(r)}\mathbb{1}(z_J=0)\end{pmatrix} + R(z),
\end{aligned}$$

where $R(z)$ includes the terms with $(t^{(m+1)\prime}z)^i, i < \ell+1$ multiplied by some constants and indicators. That is, there exist some constants, $C_i > 0, i < \ell+1$, s.t.

$$R(z) \leq \sum_{i<\ell+1} C_i (t^{(m+1)\prime}z)^i.$$

Using the structure of $t^{(m+1)}$, $(1-P')$ and $\{w_{ij}\}$ in Lemma 17 again, the

generator becomes

$$
\begin{aligned}
Gf(z) = &-(\ell+2)(t^{(m+1)\prime}z)^{\ell+1}\Big\{r^{m+1}\mu_{m+1}^{(r)}[(1-P_{m+1,m+1})-\sum_{i<m+1}P_{m+1,i}w_{i,m+1}] \\
&-\sum_{i>m+1}r^{i}\mu_{i}^{(r)}\Big(\sum_{j<m+1}P_{ij}w_{j,m+1}+P_{i,m+1}\Big)\Big\}+R(z) \\
&+(\ell+2)(t^{(m+1)\prime}z)^{\ell+1}(1-P_{m+1,m+1})\mu_{m+1}^{(r)}\mathbb{1}(z_{m+1}=0) \\
&-(\ell+2)(t^{(m+1)\prime}z)^{\ell+1}(\sum_{i<m+1}w_{i,m+1}P_{m+1,i})\mu_{m+1}^{(r)}\mathbb{1}(z_{m+1}=0) \\
&-(\ell+2)(t^{(m+1)\prime}z)^{\ell+1}(\sum_{j>m+1}(\sum_{i<m+1}w_{i,m+1}P_{ji})\mu_{j}^{(r)}\mathbb{1}(z_{j}=0)) \\
\leq &-(\ell+2)(t^{(m+1)\prime}z)^{\ell+1}r^{m+1}\Big\{\mu_{m+1}^{(r)}(1-w_{m+1,m+1})-\sum_{i>m+1}r^{i-m-1}\mu_{i}^{(r)}w_{i,m+1}\Big\}+R(z) \\
&+(\ell+2)(t^{(m+1)\prime}z)^{\ell+1}(1-P_{m+1,m+1})\mu_{m+1}^{(r)}\mathbb{1}(z_{m+1}=0).
\end{aligned}
$$

Since $\mu_{m+1}^{(r)}(1-w_{m+1,m+1})-\sum_{i>m+1}r^{i-m-1}\mu_{i}^{(r)}w_{i,m+1}>0, r\in(0,\min_{i\leq m+1}r_i)$, there exists $c>0, c_z>0, d>0$, such that

$$
Gf(z)\leq -c(t^{(m+1)\prime}z)^{\ell+1}+d\mathbb{1}((t^{(m+1)\prime}z)^{\ell+1}<c_z).
$$

Then (B.1) is satisfied, so Lemma 21 holds for $(\sum_{j\in\mathcal{J}}c_jz_j)^n=(t^{(m+1)\prime}z)^n$ with $n=\ell+1$. Furthermore, by induction hypothesis, there exists constant $M>0$, with Minkowskis Inequality repeatedly, one has

$$
\begin{aligned}
&r^{(m+1)\ell}\mathbb{E}[R(Z^{(r)})] \\
\leq&\sum_{i<\ell+1}C_ir^{(m+1)(\ell-i)}\Big(\mathbb{E}[(\sum_{j<m+1}w_{j,m+1}r^{m+1-j}r^jZ_j^{(r)})^i]^{1/i}+\mathbb{E}[(r^{m+1}Z_{m+1}^{(r)})^i]^{1/i}\Big)^i \\
\leq&M.
\end{aligned}
$$

Taking expectation and multiply $r^{(m+1)\ell}$on both sides, with Lemma 21, one has

$$
\begin{aligned}
&r^{(m+1)(\ell+1)}\Big(\sum_{i<m+1}w_{i,m+1}\mathbb{E}[Z_i^{(r)}]+\mathbb{E}[Z_{m+1}^{(r)}]\Big)^{\ell+1} \\
\leq&\frac{M+(\ell+2)(1-P_{m+1,m+1})\mu_{m+1}^{(r)}r^{(m+1)\ell}\mathbb{E}[(t^{(m+1)\prime}Z^{(r)})^{\ell+1}\mathbb{1}(Z_{m+1}^{(r)}=0)]}{(\ell+2)\{\mu_{m+1}^{(r)}(1-w_{m+1,m+1})-\sum_{i>m+1}r^{i-m-1}\mu_i^{(r)}w_{i,m+1}\}}.
\end{aligned}
$$

Besides, by Cauchy-Schwarz inequality, taking $q = \frac{m-\ell}{m+1}$, one has

$$\begin{aligned}
&r^{(m+1)\ell}\mathbb{E}\big[(t^{(m+1)\prime}Z^{(r)})^{\ell+1}\mathbb{1}(Z^{(r)}_{m+1}=0)\big]\\
=&r^{\ell-m}\mathbb{E}\Big[\Big(\sum_{i<m+1} w_{i,m+1}r^m Z^{(r)}_i\Big)^{\ell+1}\mathbb{1}(Z^{(r)}_{m+1}=0)\Big]\\
\le&r^{\ell-m}\mathbb{E}\Big[\Big(\sum_{i<m+1} w_{i,m+1}r^m Z^{(r)}_i\Big)^{m+1}\Big]^{\frac{\ell+1}{m+1}}\mathbb{P}(Z^{(r)}_{m+1}=0)^{\frac{m-\ell}{m+1}}\\
\le&r^{\ell-m}\mathbb{E}\Big[\Big(\sum_{i<m+1} w_{i,m+1}r^m Z^{(r)}_i\Big)^{m+1}\Big]^{\frac{\ell+1}{m+1}}r^{m-\ell} \triangleq M',
\end{aligned}$$

where the last inequality holds under induction hypothesis and Lemma 22. Therefore, for $r \in (0, \min_{i\le m+1} r_i)$, one has

$$\begin{aligned}
\mathbb{E}[r^{m+1}Z^{(r)}_{m+1}]^{\ell+1} \le& r^{(m+1)(\ell+1)}\Big(\sum_{i<m+1} w_{i,m+1}\mathbb{E}[Z^{(r)}_i] + \mathbb{E}[Z^{(r)}_{m+1}]\Big)^{\ell+1}\\
\le&\frac{M + (\ell+2)(1-P_{m+1,m+1})\mu^{(r)}_{m+1} + M'}{(\ell+2)\{\mu^{(r)}_{m+1}(1-w_{m+1,m+1}) - \sum_{i>m+1} r^i\mu^{(r)}_i w_{i,m+1}\}}\\
\triangleq& M^{(m+1)}_{\ell+1} > 0.
\end{aligned}$$

Above all, we choose $r_0 = \min\{r_i, i \in \mathcal{J}\}$, then two directions of induction go through the boundedness for all stations and all the moments of properly scaled steady-state queue length for each station for $r \in (0, r_0)$.

□

BIBLIOGRAPHY

[1] S. L. Bell and R. J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *The Annals of Applied Probability*, 11(3):608–649, 2001.

[2] S. L. Bell and R. J. Williams. Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electron. J. Probab.*, 10(33):1044–1115, 2005.

[3] Patrick Billingsley. *Convergence of Probability Measures, Second Edition*. New York: Wiley, 1999.

[4] M. Bramson. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, 30(1):89–140, 1998.

[5] A. Braverman, J. G. Dai, and J. Feng. Stein's method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. *Stochastic Systems*, 6:301–366, 2016.

[6] Anton Braverman, J.G. Dai, and Masakiyo Miyazawa. Heavy traffic approximation for the stationary distribution of a generalized Jackson network: the BAR approach. *Stochastic Systems*, 7(1):143–196, May 2017.

[7] Anton Braverman, Jim G Dai, and Masakiyo Miyazawa. The bar-approach for multiclass queueing networks with sbp service policies. *arXiv preprint arXiv*, 2022.

[8] Amarjit Budhiraja and Chihoon Lee. Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Math. Oper. Res.*, 34(1):45–56, 2009.

[9] Hong Chen and Avi Mandelbaum. Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.*, 16(2):408–446, 1991.

[10] J. G. Dai and J. Michael Harrison. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *The Annals of Applied Probability*, 2:65–86, 1992.

[11] J. G. Dai and J. Michael Harrison. *Processing Networks: Fluid Models and Stability*. Cambridge University Press, 2020.

[12] J. G. Dai, V. Nguyen, and M. I. Reiman. Sequential bottleneck decomposition: an approximation method for open queueing networks. *Operations Research*, 42:119–136, 1994.

[13] Atilla Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72(3):311–359, 2012.

[14] David Gamarnik and Assaf Zeevi. Validity of heavy traffic steady-state approximation in generalized Jackson networks. *Ann. Appl. Probab.*, 16(1):56–90, 2006.

[15] Peter Glynn and Assaf Zeevi. Bounding stationary expectations of markov processes. *Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz. Selected Papers of the Conference*, 4, 01 2008.

[16] Lawrence M Graves. *The Theory of Functions of Real Variables*. McGRAW-HILL Book Company, 1946.

[17] J. M. Harrison and R. J. Williams. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics*, 22(2):77–115, 1987.

[18] J. Michael Harrison. The diffusion approximation for tandem queues in heavy traffic. *Advances in Applied Probability*, 10:886–905, 1978.

[19] J. Michael Harrison. Correction: Brownian models of open processing networks: canonical representation of workload. *Ann. Appl. Probab.*, 16(3):1703–1732, 08 2006.

[20] J. Michael Harrison and Marcel J. López. Heavy traffic resource pooling in parallel server systems. *Queueing Systems*, 33(4):339–368, 1999.

[21] J. Michael Harrison and Martin I. Reiman. Reflected Brownian motion on an orthant. *Ann. Probab.*, 9(2):302–308, 1981.

[22] S. G. Henderson. Variance reduction via an approximating markov pro cess. *Ph.D. thesis*, 1997.

[23] Daniela Hurtado-Lange and Siva Theja Maguluri. Transform methods for heavy-traffic analysis. *Stochastic Systems*, 2020.

[24] J. R. Jackson. Networks of waiting lines. *Operations Research*, 5:518–521, 1957.

[25] D. P. Johnson. *Diffusion approximations for optimal filtering of jump processes and for queueing networks*. PhD thesis, University of Wisconsin, 1983.

[26] Sean P. Meyn and Richard. L. Tweedie. Stability of Markovian processes III: Foster-Lyapunov criteria for continuous time processes. *Adv. Appl. Probab.*, 25:518–548, 1993.

[27] Martin I. Reiman. Open queueing networks in heavy traffic. *Mathematics of Operations Research*, 9:441–458, 1984.

[28] Xinyang Shen, Hong Chen, J. G. Dai, and Wanyang Dai. The finite element method for computing the stationary distribution of an SRBM in a hypercube with applications to finite buffer queueing networks. *Queueing Systems*, 42:33–62, 2002.

[29] Alexander L. Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability*, 14(1):1–53, 2004.

[30] W. Wang, K. Zhu, L. Ying, J. Tan, and L. Zhang. Map task scheduling in mapreduce with data locality: Throughput and heavy-traffic optimality, 2013.

[31] Ward Whitt. The queueing network analyzer. *Bell System Technical Journal*, 62:2779–2815, 1983.

[32] Q. Xie, A. Yekkehkhany, and Y. Lu. Scheduling with multi-level data locality: Throughput and heavy-traffic optimality. *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9, 2016.

[33] Yaosheng Xu and JG Dai. Heavy traffic performance of wwta load-balance algorithms in parallel-server systems with heterogeneous servers. *arXiv preprint arXiv*, 2022.