A Multi-Server Fork-Join Network in the Halfin-Whitt Regime

Hongyuan Lu and Guodong Pang

1. Motivation and Model

We consider a fundamental fork-join network with a single class of jobs that will fork into a fixed number of parallel tasks upon their arrival, and then join after service completion. Each parallel task is processed at a multi-server station under the first-come-first-serve (FCFS) and non-idling service discipline, and will join a buffer waiting for synchronization ("unsynchronized queue") associated with the station after service completion. Service times of parallel tasks of each job can be correlated. Tasks are only synchronized if all the parallel tasks of the same job are completed, called "non-exchangeable synchronization" (NES). After synchronization, jobs will leave the system immediately (the synchronization time is irrelevant in our model). Figure 1 depicts such a network model. Unlike classical queueing models, there are two types of delays in this fork-join network: delay for service and delay for synchronization. The objective of this paper is to study the delay for synchronization when each service station is operating in the Halfin-Whitt (Quality-and-Efficiency-Driven, QED) regime. In this regime, the job arrival rate and the number of servers in each service station get large appropriately while fixing service time distributions so that each station is asymptotically critically loaded, achieving both high quality (low delay) and high efficiency (high utilization).

Fork-join networks with NES are used in many applications, including healthcare systems, parallel computing, MapReduce scheduling (e.g., large-scale parallel Web search), disassembly and reassembly systems in manufacturing and so on. In patient flows of hospitals, the treatment and discharge processes are typical examples of fork-join networks with NES: a patient must have all test results ready before a doctor examination and these tests are conducted in different medical units/laboratories and can never be mixed; a patient, after the discharge decision is made, must wait for necessary procedures, pharmacy, transportation, etc., before being physically discharged. In MapReduce scheduling, jobs are processed in two phases: in the map phase, a large-scale data input (e.g., Web processing data) is distributed into individual computation nodes, and each node processes one block of input data, and after the execution of all blocks of the same data input, they will be joined as an output in the reduce phase. In addition, fork-join networks with NES are also natural models in manufacturing and inventory systems, military operations and law reinforcement.

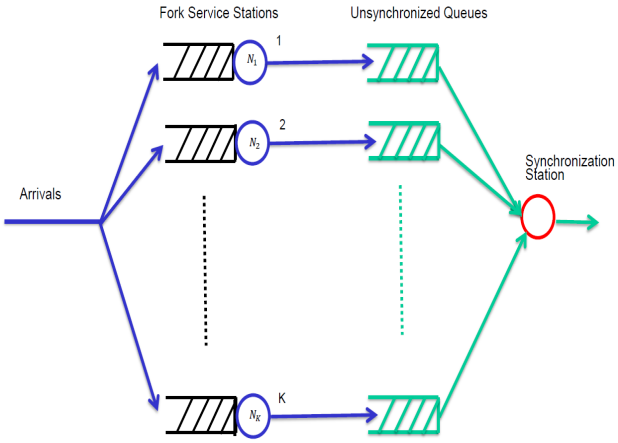


Figure 1: A fundamental fork-join network

The main mathematical challenge in analyzing the multi-server fork-join network with NES is the resequencing of arrival orders after service completion at each service station

due to the randomness of service times and delay for service at all parallel service stations. Dependence of service times for parallel tasks of a job also makes the service completion processes of the parallel tasks dependent, which causes a substantial amount of difficulties in the analysis of the resequencing of the parallel tasks and the synchronization process, as well as the service dynamics at all parallel stations jointly. To the best of our knowledge, our work is the first to study (non-Markovian) multi-server fork-join networks with NES in the Halfin-Whitt regime.

2. Methods and Results

Exact analysis of this model is prohibitively difficult since it is necessary to track the service completion times of all the parallel tasks of each job, which will require an infinite dimensional state space. We develop a new approach to study the resequencing problem in the fork-join networks with NES asymptotically when each station is operating in the Halfin-Whitt regime. Specifically, we establish a relationship between the dynamics of the finite-server fork-join network model and that of the corresponding infinite-server fork-join network model. Thus, the system dynamics (queueing, service, waiting for synchronization, and synchronization) in the multi-server fork-join network model with NES can be represented as functionals of a multiparameter sequential empirical process driven by the service vectors for the parallel tasks as well as the arrival process and the initial quantities.

With this representation, we first show an FLLN (fluid limit) for these processes assuming that the system starts from empty when the arrival rate is allowed to be time dependent. In particular, the fluid limit of the synchronized process is an integral of the minimum of the fluid entering service processes at all stations with respect to the joint service time distribution function. The fluid limits for the unsynchronized queueing processes and the synchronized process capture the impact of the service dependence among all parallel tasks of each job through their joint distribution function. For the fluid limits of service processes, they are reduced to the same limits as in $G/G/N$ queues, depending only on the marginal service time distribution function, and thus are not affected by the dependence structure of parallel service times.

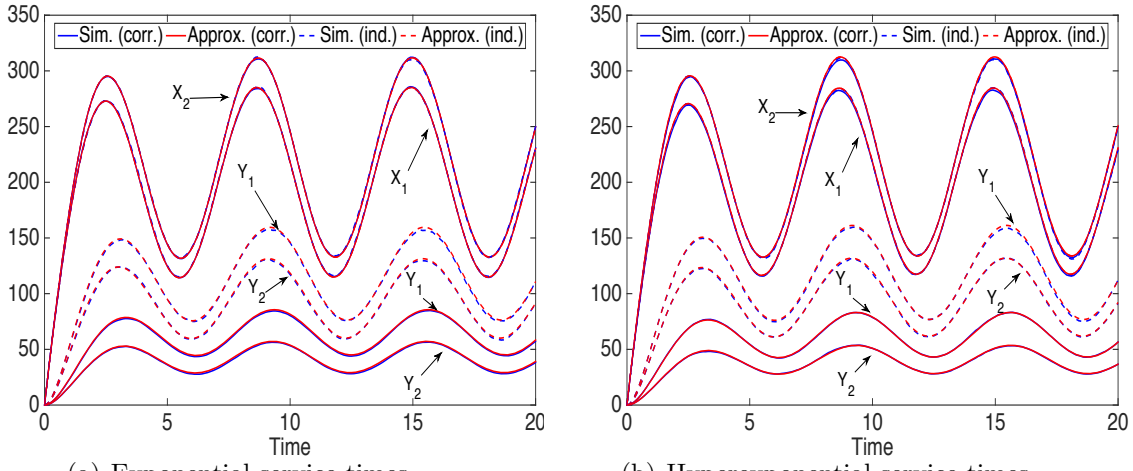
We then prove an FCLT for the aforementioned processes when the arrival rate is constant in the Halfin-Whitt regime and when the number of parallel tasks is equal to two, under some stationarity conditions on the initial quantities. The limits of the diffusion-scaled processes are the unique solution to a set of stochastic integral equations driven by a generalized multiparameter Kiefer process (the limit of the multiparameter sequential empirical processes driven by the service vectors), the arrival limit process and the limiting initial quantities. One important term in the limits of the synchronized process and the unsynchronized queues is an integral of the limit of the diffusion-scaled minimum of "entering service" processes at both stations with respect to the joint service time distribution. Moreover, the limits of service processes for parallel stations are dependent on the joint distribution of service times unlike their fluid limits, and thus are correlated with each other.

Our results show that when all service stations operate in the Halfin-Whitt regime and the arrival rate and the numbers of servers at all stations are of order $O(n)$, the numbers of tasks in the service stations and the numbers of tasks waiting for synchronization are of the same order, $O(n)$. This implies that waiting times for synchronization are $O(1)$,

although waiting times for service are $O(1/\sqrt{n})$. This is an extremely important insight for the management of multi-server fork-join networks with NES in the Halfin-Whitt regime. An intuitive interpretation is that in steady state, for jobs whose tasks are waiting in the associated buffer(s) for synchronization, their other parallel tasks must be already in service with probability one asymptotically. Therefore, in order to minimize the response time - the time duration from the arrival time to synchronization, we conjecture that one must prioritize tasks in each service station dynamically to reduce the waiting time for synchronization to a smaller order.

3. Numerical Examples

We give numerical examples with two parallel stations to show the effectiveness of fluid approximations for the number of tasks X_k in parallel station k (including those in service and in queue) and the size of waiting buffer k for synchronization Y_k, $k = 1, 2$, comparing with simulations. We let the arrival process be Poisson with time-varying rate $200 + 120\sin(t)$, $t \geq 0$. The numbers of servers in stations 1 and 2 are 300 and 340, respectively. In the first numerical example, the service times of the 1st and 2nd tasks are assumed to have a bivariate Marshall-Olkin exponential distribution with means 1 and 10/9, respectively. In the second numerical example, we let the service times of the two parallel tasks have a bivariate Marshall-Olkin hyperexponential distribution with the same means as the first example. In both cases, we consider independent and dependent parallel service times, and the parameters in the Marshall-Olkin exponential and hyperexponential distributions are chosen such that the correlation coefficients are both equal to 0.5. The numerical results are shown in Figure 2, marked with "ind." and "corr.", respectively, for independent and correlated cases. We make two remarks from numerical results. First, the fluid approximations match very well with the simulated results. Second, the positive correlation among parallel service times does not affect the service dynamics (X_k) but reduces the unsynchronized queues (Y_k).



(a) Exponential service times

(b) Hyperexponential service times

Figure 2: Comparison of fluid approximations with simulations.