Product-forms in batch networks: approximation and asymptotics

Peter Harrison[a], Richard A. Hayden[a], William Knottenbelt[a]

[a] *Dept. of Computing, Imperial College London,*
Huxley Building, 180 Queen's Gate, London SW7 2BZ, United Kingdom

Abstract

It is shown that a Markovian queue, with bulk arrivals and departures having any probability mass functions for their batch sizes, has geometrically distributed queue length at equilibrium (when this exists) provided there is an additional special bulk arrival stream, with particular rate and batch size distribution, when the server is idle. It is shown that the time-averaged input rate of the special arrivals tends to zero as the queue becomes saturated, and a heavy-traffic limit for the queue without special arrivals is derived by martingale methods. This is shown to give the same asymptotic queue length probabilities as the geometric model. The product form is then extended to tandem networks of batch queues using the reversed compound agent theorem (RCAT). In order to obtain the product form in this case, it is required that, in addition to special arrival streams, so-called 'partial batches' are discarded immediately from the network when there are not enough customers in the queue to fill an entire departing batch. Somewhat surprisingly it turns out that, in heavy traffic, the product-form network does not always agree with the regulated Brownian motion (RBM) diffusion limit for the standard network without special arrivals and where partial batches are not discarded, but forwarded to the next node. Indeed, we show that the two models agree in heavy traffic if and only if the skew-symmetry condition for the RBM to have a product form is satisfied. When the condition does hold, our theoretical and numerical results thus validate the use of the product-form batch networks as moderate-traffic approximations to the analogous standard queueing network model without special arrivals and where partial batches may be forwarded to the next node instead of being lost. In the case that the condition does not hold, we obtain a new product-form stationary distribution for the associated non-RBM diffusion limit.

1. Introduction

Burstiness in network traffic tends to degrade network performance and for many years traffic-shapers have aimed at "smoothing" unavoidably bursty offered load at devices to increase efficiency. On the other hand, switching off some devices when they are not in use and switching them on again when they are next required is a common method of reducing energy consumption. This switching inherently increases burstiness, not only in power consumption but also in the performance delivered. Stochastic performance analysis of queues and networks of queues with batches has therefore become an important means to address these issues and the energy–performance trade-off in particular.

We define a class of batch queues for which we obtain conditions for a geometric queue length probability distribution to exist at equilibrium. These have regular batch arrivals and batch departures, as well as a special batch arrival stream that is activated only when the queue is empty, and "partial" batch departures with nominal size greater than the current queue length that lead to the empty state, i.e. clear the queue. The geometric distribution, when it exists, allows a product-form to be derived for networks of such queues, called batch networks, which is simply proved using the reversed compound agent theorem (RCAT) of [1] in Section 2.3.

Of course there is the objection that, even if batch networks provide a good representation, it is unlikely that the conditions will be met that lead to a product form, and hence efficient solution. However, direct analytical solutions (solving the underlying Markov chain's global balance equations), or explicit-state bounding approximation schemes such as [2, 3] are generally intractable numerically for high levels of system load, and simulation is expensive and time consuming. This gives at least three important roles for product forms in general:

- To provide exact results when their conditions are met;
- To provide a benchmark against which to assess approximate methods and simulation: a parameterisation of a model would be chosen that does satisfy the conditions for product form and the ensuing exact solution would be compared with the inexact model's output;

- Product forms themselves are often acceptable approximations to the model of interest and may lead to upper and/or lower bounds on the exact solution.

We introduce our geometric batch queue – abbreviated to *GBQ* – in the next section and obtain an equation for its geometric queue-length parameter and an expression for its response time distribution's Laplace-Stieltjes transform (LST). Its reversed process is shown to be another GBQ and the rate equations leading to product-form solution in networks are considered in Section 2.3. Heavy-traffic diffusion limits are derived by martingale methods in Section 3 for standard batch queues (with no special arrivals or departures), which we call *SBQs*, and the corresponding limit for the GBQ is obtained directly from its distribution's parameter. This is found to be the same as the SBQ limit, which may not, at this stage, seem surprising in light of the fact that we also show that the net throughput (i.e. time-averaged rate) of the special arrivals tends to zero as the GBQ's utilisation tends to one.

Tandem networks, considered in Section 4, turn out to be more interesting. Although in a product-form tandem network of two GBQs, the throughputs of the special arrivals at both queues still tend to zero as their utilisations approach one, and the same applies to the discarded, partial batches, the product-form network *does not*, in general, approach the diffusion limit of the standard network with no special arrivals and where partial batches are forwarded to the second queue. A more careful consideration reveals that the special arrivals at the second queue do indeed have no effect on the heavy-traffic limit, as in a single batch queue. However, the special arrivals at the *first* queue do affect the limit at the second queue. Moreover, *forwarding* partial batches to the next queue, rather than discarding them as in the GBQ model, affects the heavy-traffic limit.

Given the zero-throughput results mentioned above, these results are rather surprising. A plausible intuition is to note that the special arrivals at the first queue cause an increase in net arrival rate at the second queue at *all* queue lengths there – in contrast to its own special arrivals. Similarly, discarding partial batches causes a decrease in net arrival rate at all queue lengths. It is shown that when these positive and negative perturbations on the arrival rate to the second queue are equal on average, the heavy-traffic limit for the tandem network of GBQs is the same as that of the SBQs – a multi-dimensional reflected Brownian motion (RBM), as we again prove via martingales. Furthermore, the condition for these time-averages to be equal reduces to the *skew-symmetry* condition known to be required for the RBM components to have a product-form stationary distribution. In addition to this satisfying connection with previous work, the product-form heavy-traffic limit of the GBQs, where the skew-symmetry condition is not satisfied, yields a new product-form limit for the resulting non-RBM diffusion. Numerical results are presented in Section 5 and the paper concludes in Section 6, where we discuss how to apply batch queues to energy-performance optimisation problems.

2. Geometric batch queues

We seek conditions on the batch-size probability distributions and the corresponding instantaneous transition rates that render the equilibrium state probabilities geometric. Product forms in networks of such queues are then easy to identify and write down, using RCAT. It is well known that no such product form exists for queues with only the standard arrival and departure batches described above – unless these are unit-sized with probability one. We therefore introduce additional, "special" batches that can arrive only when the queue is empty. Similarly, partial batches, which truncate a departure batch to the current queue length, leave the queue empty. Our model of batch transitions in a single-server queue is defined as follows, where we assume that the rates are bounded so that the infinite sums exist:

- The state space $\mathcal{S}$ of the queue is the set of non-negative integers;
- *Normal* batch arrivals of size $k \geq 1$ are represented by transitions with constant instantaneous rate $a_k : i \to i + k \;\; (i \geq 0)$, i.e. from states i to $i + k$;
- Additional *special* batch arrivals of size $k \geq 1$ to an empty queue are represented by transitions with constant instantaneous rate $a_{0k} : 0 \to k$;
- Full batch departures of size k are represented by transitions with constant instantaneous rate $d_k : i + k \to i \;\; (i \geq 0)$;
- *Partial* batch departures of size k, leading to an empty queue, are represented by transitions with constant instantaneous rate $d_{k0} = \sum_{j=k+1}^{\infty} d_j : k \to 0$;
- The ordering of individual tasks in the queue is strictly first come first served (FCFS).

Rate generating functions are defined for each batch transition as follows:

$$A(z) = \sum_{k=1}^{\infty} a_k z^k \qquad A_0(z) = \sum_{k=1}^{\infty} a_{0k} z^k \qquad D(z) = \sum_{k=1}^{\infty} d_k z^k$$

We assume that $A(1), A_0(1), D(1) < \infty$, to avoid null mean state holding times (i.e. infinite total instantaneous transition rate out of a state). The functions $A(z), A_0(z), D(z)$ are therefore absolutely convergent and analytic inside the unit disk, which lies inside their circles of convergence. Alternatively, we describe the corresponding processes in terms of the total rate of their underlying Poisson point process along with a batch-size random variable. In particular, the arrival process is Poisson with rate Λ and the batch size α has probability mass function $p^a_\cdot$. Thus, $\Lambda = A(1)$ and $p^a_k = a_k/A(1)$ for $k \geq 1$. Similarly, the supremum of the total normal rate of service completions (or the sum of the rates of *all* service completions in any state, whether normal or special) is denoted by Δ, with batch-size random variable β having mass function $p^d_\cdot$, so that $\Delta = D(1)$ and $p^d_k = d_k/D(1)$ for $k \geq 1$.

The following proposition gives conditions for the length of a special batch queue to have a geometric equilibrium probability distribution, along with its parameter, so that product forms become facilitated in networks by application of RCAT.

Proposition 1. *The batch queue defined above, with $A(1), D(1) < \infty$, has geometrically distributed equilibrium queue-length probabilities with parameter $\rho < 1$, $\pi_n = (1-\rho)\rho^n$ for $n \geq 0$, iff*

$$A_0(z) \quad = \quad \frac{[A(1) + D(1) - D(\rho)]\rho z - A(z)}{1 - \rho z} \tag{1}$$

for $|z| < \min(\rho^{-1}, R)$, where R is the radius of convergence of the series $A(z)$.[1][2] *If $A(\rho^{-1}) < \infty$, then ρ is the unique solution of the equation:*

$$A(\rho^{-1}) + D(\rho) = A(1) + D(1) \tag{2}$$

in the interval $(0, 1)$, whereupon we may write

$$A_0(z) \quad = \quad \frac{A(\rho^{-1})\rho z - A(z)}{1 - \rho z} \tag{3}$$

Proof. At equilibrium, the queue has balance equations, for $i \geq 1$,

$$\Big(A(1) + \sum_{j=1}^{i} d_j + d_{i0}\Big)\pi_i = \sum_{j=1}^{i} a_j \pi_{i-j} + \pi_0 a_{0i} + \sum_{j=1}^{\infty} d_j \pi_{i+j} \tag{4}$$

$$(A(1) + A_0(1))\pi_0 = \sum_{j=1}^{\infty} (d_j + d_{j0})\pi_j \tag{5}$$

Suppose $\pi_i = (1-\rho)\rho^i$. Multiplying equation 4 by z^i and summing from $i = 1$ to ∞ gives (substituting for d_{i0}):

$$A(1)(\Pi(z) - \pi_0) + \sum_{i=1}^{\infty}\sum_{j=1}^{i} d_j \pi_i z^i + \sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty} d_j \pi_i z^i = A(z)\Pi(z) + \pi_0 A_0(z) + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} d_j \pi_{i+j} z^i$$

where $\Pi(z) = \sum_{i=0}^{\infty}(1-\rho)\rho^i z^i = (1-\rho)/(1-\rho z)$ for $|z| < \rho^{-1}$. Dividing by $\Pi(z)$, summing all but one of the series terms and interchanging the order of summation in the remaining one, this becomes:

$$(1-\rho z)A_0(z) = A(1)\rho z - A(z) + D(\rho z) - D(\rho)\rho z + (1-\rho z)\sum_{j=2}^{\infty} d_j \sum_{i=1}^{j-1} \rho^i z^i$$

so that $(1-\rho z)A_0(z) = A(1)\rho z - A(z) + D(\rho z) - D(\rho)\rho z + \sum_{j=2}^{\infty} d_j \big(\rho z - (\rho z)^j\big)$. Equation 1 now follows. The converse is proved by expanding equation 1 in powers of z and comparing coefficients. The trial solution is

[1] In Proposition 2 it is shown that for the reversed process also to have finite total departure rate, $A(\rho^{-1}) < \infty$ so that effectively it is sufficient here that $z \leq 1/\rho$.

[2] In Proposition 2 it is shown that for the reversed process also to have finite total departure rate, $A(\rho^{-1}) < \infty$ so that effectively it is sufficient here that $z \leq 1/\rho$.

therefore valid (it is straightforward to check that equation 1 gives equation 5 at $z = 1$) and the proposition follows by uniqueness of the equilibrium probabilities of an irreducible Markov process.

For the last part of the proposition, note that $A_0(z)$ has radius of convergence greater than R and so is finite at $z = 1/\rho$. Thus, its numerator must vanish there, giving equation 2. Now let $f(x) = A(x^{-1}) + D(x) - A(1) - D(1)$. Then $f(r) > 0$ for some $r < 1$, since $A(y)$ increases in real $y > 0$ (all its coefficients being non negative), and $f(1) = 0$. There is therefore at least one solution to the equation $f(x) = 0$ in the open interval $(r, 1)$ if and only if $\dot{f}(1) > 0$ (whereupon $f(1^-) < 0$), i.e. iff $\dot{D}(1) - \dot{A}(1) > 0$, since $D(z)$ and $A(z^{-1})$ are analytic in the annulus with inner radius r and outer radius 1, and so are continuous in $(r, 1)$. Any such solution is a valid value for the parameter ρ of a geometric probability function that satisfies the queue's Kolmogorov equations and must be unique. $\diamondsuit$

Notice that the derivatives at $z = 0$, $\dot{A}(1)$ and $\dot{D}(1)$ are the task-arrival and task-departure rates respectively, so that $\dot{D}(1) > \dot{A}(1)$ is the expected stability condition for the batch queue. We call a queue satisfying the conditions of Proposition 1 a geometric batch queue, or GBQ, with parameter ρ.

Such queues are used in an "assembly-transfer network" in Chapter 8 of [4], which is appropriate for models of certain manufacturing systems. The interpretation is that batches of size less than or equal to the queue length are "full batches" and the others are "partial batches" (referring to a size equal to the queue length n which is less than the intended batch size k), which are discarded in the assembly line.

2.1. The reversed batch queue

Although to apply RCAT requires the reversed rates of only the active actions – in our case the normal departures – we will need the whole reversed process later when we consider sojourn times in a network of batch queues. Notice that the structure of a batch queue is symmetric: there are normal batch arrivals and departures; together with special batch arrivals and departures, out of and into state 0 only, respectively. Any such queue is specified entirely by its rate generating functions A, A_0, D. Because of the symmetry, the reversed process is also a batch queue with rate generating functions A', A'_0, D', say. We now calculate these and confirm, using Proposition 1, that the reversed queue has the same geometric queue-length probability distribution at equilibrium (assumed to exist, with $\rho < 1$) as the original (forward) queue.

Proposition 2. *The reversed process of a geometric batch queue with parameter ρ and rate generating functions $A(z)$, $A_0(z)$, $D(z)$ is also a geometric batch queue with parameter ρ and rate generating functions: $A'(z) = D(\rho z)$; $A'_0(z) = D_0(\rho z)$; $D'(z) = A(\rho^{-1}z)$. It has finite total outgoing rate in each state provided $A(\rho^{-1}) < \infty$.*

Proof. The rate of a transition in the reversed process, denoted by a prime, is the corresponding forward rate multiplied by the ratio of the (forward transition's) source-state over destination-state equilibrium probability. In the reversed process, the reversed arrival transitions cause decreases in the queue length and so become departures, and similarly, the reversed departure transitions become arrivals. Moreover, the special transitions out of, respectively, into state 0 map into special transitions in the reversed process into, respectively, out of state 0. Using the hypothesis that the equilibrium queue length probabilities are geometric, we have $a'_k = \rho^k d_k$, $a'_{0k} = \rho^k d_{k0}$ and $d'_k = \rho^{-k} a_k$. The rate generating functions of the reversed process then follow as stated. The reversed total departure rate is $D'(1) = A(\rho^{-1}) < \infty$ and $A'(1) = D(\rho) < \infty$. The reversed process therefore satisfies the conditions of Proposition 1 and has equilibrium queue-length probability distribution with the same parameter ρ. $\diamondsuit$

2.2. Example

In the simplest batch queue that requires special arrivals, (normal) arrival batches have size either 1 or 2 and departure batches are all unit. Hence $A(z) = \lambda_1 z + \lambda_2 z^2, D(z) = \mu_1 z$. Equation 2 then yields $\mu_1\rho^3 - (\lambda_1 + \lambda_2 + \mu_1)\rho^2 + \lambda_1\rho + \lambda_2 = 0$ which factorises into $(\rho - 1)P(\rho) = 0$, where $P(x) = (\mu_1 + \mu_2)x^2 - (\lambda_1 + \lambda_2)x - \lambda_2$. Since $\rho = 1$ is invalid, ρ must satisfy $P(x) = 0$. Now, $P(0) < 0$ and $P(1) = \mu_1 - \lambda_1 - 2\lambda_2 > 0$ under the stabilility condition. Therefore, there is a geometric equilibrium probability distribution for the queue length, *viz.* the solution of the quadratic equation $P(x) = 0$ in the interval $(0, 1)$. Equation 3 of Proposition 1 gives the arrivals-in-state-0 rate generating function:

$$A_0(z) = \frac{\rho z(\lambda_1/\rho + \lambda_2/\rho^2) - \lambda_1 z - \lambda_2 z^2}{1 - \rho z} = \lambda_2 z/\rho$$

Thus, only unit-batch special arrivals are required for a geometric solution to exist. In the special case that arrivals are always single, $\lambda_2 = 0$ and we find $A_0(z) = 0$. Indeed, provided that all arrivals are single (unit batches), departure batches of arbitrary size yield a geometric equilibrium queue-length probability distribution (assuming equilibrium exists), without requiring any special arrivals.

2.3. Product-form batch networks

By construction, networks of geometric batch queues – batch networks – may have product forms when their nodes are interconnected such that normal batch departures become the normal batch arrivals at another node. The special departures must leave the network and the special arrivals must also be external; their parameters are determined by the rate equations of RCAT combined with equation 1 or 3. This idea itself is not new and product forms have been obtained for special cases in [4, 5]. More generally, product forms were proved for a similar class of batch networks using quasi-reversibility in [6]. Normal internal departure streams may be split probabilistically to several other nodes by using parallel active departure transitions, just as in conventional queueing networks [7, 8]. Thus, the enabling constraints of RCAT are satisfied in that the passive transitions are normal arrivals that are enabled in every state and, similarly, the active transitions come into every state, these being normal departures. It remains to solve the rate equations which equate the reversed rates of the active transition types, a say, to an associated variable x_a– see [9] for a practical description. Notice that the reversed rates will in general depend on the set of x_a and can be found from Proposition 2.

Here, however, we focus on tandem networks of batch queues. In a tandem of J queues, the reversed rates of the departures from queue j that go to queue $j+1$ w.p. one have generating function (by Proposition 2) $D_j(\rho_j z)$ and so the RCAT rate equations are:

$$D_j(\rho_j/\rho_{j+1}) + D_{j+1}(\rho_{j+1}) = D_j(\rho_j) + D_{j+1}(1) \tag{6}$$

for $j = 1, \ldots, J-1$. In addition, the first node behaves as if in isolation so that

$$A_1(1/\rho_1) + D_1(\rho_1) = A_1(1) + D_1(1) \tag{7}$$

We investigate the asymptotic behaviour of tandem geometric batch queues in terms of both equilibrium state probabilities and response-time probability distribution in subsequent sections.

2.4. Special arrivals in heavy traffic

In heavy traffic, the special arrivals have negligible effect in the sense that their throughput tends to zero as a batch queue's utilisation tends to one. This claim can be proved as follows. First, writing $A_0(z) = -[A(1) + D(1) - D(\rho)] + \frac{A(1)+D(1)-D(\rho)-A(z)}{1-\rho z}$, the throughput of special arrivals (measured in tasks per unit time) is the value of $(1-\rho)\dot{A}_0(z)$ at $z = 1$, i.e.

$$\frac{(1-\rho)[A(1) + D(1) - D(\rho) - A(z)]\rho}{(1-\rho z)^2} - \frac{(1-\rho)\dot{A}(z)}{1-\rho z}\bigg|_{z=1} = \frac{[D(1) - D(\rho)]\rho}{1-\rho} - \dot{A}(1)$$

As $\rho \to 1$, the throughput of the special arrivals therefore tends to $\dot{D}(1) - \dot{A}(1)$. But in the heavy-traffic limit, $\dot{D}(1) = \dot{A}(1)$. Thus we might expect the product form to become, in heavy traffic, an increasingly accurate approximation of the corresponding more realistic model without the special arrivals. We verify this conjecture, in terms of both equilibrium state probabilities and sojourn-time distributions, in the next section and the appendix, respectively, by comparing with the corresponding heavy-traffic limits. In Section 4, however, we find that this result does not extend in the expected way to networks.

3. SBQ heavy-traffic limit

In this section we consider a sequence of *standard* batch queues indexed by $n \geq 1$. Arrivals of batches occur at Poisson rate λ^n to the nth queue and departures at Poisson rate μ^n. We assume that arrival batch sizes are distributed according to the discrete probability distribution p^a_k for $k \geq 1$ and departure batch sizes according to p^d_k for $k \geq 1$. The batch-size distributions are independent of n and can be arbitrary other than requiring the existence of the first four moments, which we write as $\mathbb{E}[\alpha^m] = \sum_{k=1}^{\infty} k^m p^a_k$ and $\mathbb{E}[\beta^m] = \sum_{k=1}^{\infty} k^m p^d_k$, respectively, for $m = 1, \ldots, 4$. We define, correspondingly, the rate generating functions $A^n(z), D^n(z)$ and $A^n_0(z)$, where $a^n_k = \lambda^n p^a_k$, for example.

Define the aggregate task arrival and service rates $\bar{\lambda}^n = \mathbb{E}[\alpha]\lambda^n$ and $\bar{\mu}^n = \mathbb{E}[\beta]\mu^n$, respectively. The heavy-traffic regime we consider assumes that for some fixed $\nu > 0$, as $n \to \infty$, $\bar{\lambda}^n \to \nu$ and $\bar{\mu}^n \to \nu$ (so that the utilization $\bar{\lambda}^n/\bar{\mu}^n \to 1$). We also assume, in the usual way for heavy-traffic limits, that $\sqrt{n}(\bar{\lambda}^n - \bar{\mu}^n) \to \theta$ for some $\theta \in \mathbb{R}$.

3.1. SBQ queue-length process limit

We construct the queue length process here in terms of, for each $n \geq 1$, independent Poisson processes $A^n(t)$ and $D^n(t)$ of rates λ^n and μ^n, respectively, and $\{\alpha_i\}_{i=1}^{\infty}$ and $\{\beta_i\}_{i=1}^{\infty}$ mutually independent and identically distributed sets of positive-integer-valued random variables with distributions p_k^a and p_k^d for $k \geq 1$, respectively, which are also independent of each other. Then, assuming that the initial state of the nth queue is distributed as $Q^n(0)$ (assumed independent of A^n, D^n, $\{\alpha_i\}_{i=1}^{\infty}$ and $\{\beta_i\}_{i=1}^{\infty}$), define:

$$X^n(t) = Q^n(0) + \int_0^t \alpha_{A^n(s)} \, \mathrm{d}A^n(s) - \int_0^t \beta_{D^n(s)} \, \mathrm{d}D^n(s)$$

Then by the memoryless property the queue-length process for the nth queue is equal in distribution to the one-dimensional reflection of X^n: $Q^n(t) = R[X^n(t)] = X^n(t) - \inf_{s\in[0,t]}(X^n(s) \wedge 0)$. The heavy-traffic limit will arise in the usual way through a central-limit-theorem rescaling of the queue-length process in space by $1/\sqrt{n}$ and in time by n. This has the effect of increasing the frequency but reducing the magnitude of jumps in precisely the correct manner to achieve the limiting result. To this end, define $\bar{A}^n(t) = A^n(nt)$, $\bar{D}^n(t) = D^n(nt)$, $\bar{X}^n(t) = (1/\sqrt{n})X^n(nt)$ and $\bar{Q}^n(t) = (1/\sqrt{n})Q^n(nt)$. We now define additionally *compensated* versions of the integral component processes $\int_0^t \alpha_{\bar{A}^n(s)} \, \mathrm{d}\bar{A}^n(s)$ and $\int_0^t \beta_{\bar{D}^n(s)} \, \mathrm{d}\bar{D}^n(s)$ in order to obtain martingales which can then be exploited to prove the limit theorems, adapting the style of methodology in [10]. Informally, the martingales are obtained by subtracting averages to 'balance the processes in expectation':

$$\bar{M}_A^n(t) = \frac{1}{\sqrt{n}} \left(\int_0^t \alpha_{\bar{A}^n(s)} \, \mathrm{d}\bar{A}^n(s) - \bar{\lambda}^n nt \right) \quad \text{and} \quad \bar{M}_D^n(t) = \frac{1}{\sqrt{n}} \left(\int_0^t \beta_{\bar{D}^n(s)} \, \mathrm{d}\bar{D}^n(s) - \bar{\mu}^n nt \right).$$

Formally, we use Lemma 1 in Appendix C to show that $\bar{M}_A^n$ and $\bar{M}_D^n$ are locally square-integrable martingales with respect to the filtration $\mathcal{F}^n(t) = \sigma\{Q^n(0), \bar{A}^n(s), \bar{D}^n(s), \alpha_{\bar{A}^n(s)}, \beta_{\bar{D}^n(s)} : 0 \leq s \leq t\}$

We now write the rescaled queue-length process in terms of the martingales: for $\bar{Q}^n(0) = (1/\sqrt{n})Q^n(0)$, we have $\bar{Q}^n(t) = R[\bar{X}^n(t)]$, where $\bar{X}^n(t) = \bar{Q}^n(0) + \bar{M}_A^n(t) - \bar{M}_D^n(t) + \sqrt{n}t(\bar{\lambda}^n - \bar{\mu}^n)$ Our intention is to apply the martingale functional central limit theorem (FCLT) to obtain limits for the martingale terms and then to use the continuous mapping theorem to generate limits for the queue-length process from these. In order to apply the martingale FCLT, we exploit Lemma 1 again which gives a characterisation of their predictable quadratic variation processes (see e.g. [10] for more of the supporting theory): $\langle \bar{M}_A^n \rangle(t) = \mathbb{E}[\alpha^2]\lambda^n t$ and $\langle \bar{M}_D^n \rangle(t) = \mathbb{E}[\beta^2]\mu^n t$. We have further that the predictable covariations $\langle \bar{M}_A^n, \bar{M}_D^n \rangle(t) = \mathbf{0}$ w.p. 1 since $\bar{A}^n$ and $\bar{D}^n$ are independent and thus share no jumps w.p. 1. We wish to apply the specific version of the martingale FCLT in [11, Theorem 2.1]. To do so, it remains to verify that the jumps of the martingale terms decay asymptotically in the following sense, for any $T \in \mathbb{R}_+$:

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{t\in[0,T]} \|\bar{M}_A^n(t) - \bar{M}_A^n(t-)\|^2 \right] = 0$$

and similarly for $\bar{M}_D^n$. We give the reasoning only for $\bar{M}_A^n$ since it is identical in both cases.

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in[0,T]} \|\bar{M}_A^n(t) - \bar{M}_A^n(t-)\|^2 \right] &= \mathbb{E}\left[\max_{1\leq k\leq A^n(nT)} \{(1/n)\alpha_k^2\} \right] = \sum_{j=1}^{\infty} \mathbb{P}\{A^n(nT) = j\} \mathbb{E}\left[\max_{1\leq k\leq j} \{(1/n)\alpha_k^2\} \right] \\
&\leq \frac{\mathbb{E}[\alpha^2]}{n} + \sqrt{\mathrm{Var}[\alpha^2]} \frac{e^{-\lambda^n nT}}{n} \sum_{j=1}^{\infty} \frac{(\lambda^n nT)^j (j-1)}{j!\sqrt{2j-1}}
\end{aligned}$$

using a simple bound for the expected maximum order statistic of independent and identically distributed random variables [e.g. 12, Section 4.2]. The sum on the right is bounded above by $\sum_{j=0}^{\infty} \sqrt{j} \frac{(\lambda^n nT)^j}{j!} = \exp(\lambda^n nT)\mathbb{E}[\sqrt{X}]$ where X is a Poisson random variable with mean $\lambda^n nT$, which, in turn, is bounded above by $\exp(\lambda^n nT)\sqrt{\lambda^n nT}$ by Jensen's inequality, from which the required limit follows. The bound used above is the reason for requiring the existence of moments of the batch-size distributions up to order four rather than just two. It may be possible to relax this restriction.

The conditions of the martingale functional central limit theorem have now been verified so that, as $n \to \infty$:

$$(\bar{M}_A^n, \bar{M}_D^n) \Rightarrow \left(\sqrt{\nu \frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}} B_A, \sqrt{\nu \frac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]}} B_D \right)$$

where B_A and B_D are independent standard Brownian motions and the convergence is weak on $D_{\mathbb{R}^2}[0,\infty)$

endowed with the uniform topology. Assuming that $\bar{Q}^n(0) \Rightarrow Y(0)$ in $\mathbb{R}$ as $n \to \infty$, the continuous mapping theorem yields $\bar{X}^n \Rightarrow Y$, where Y is defined by:

$$Y(t) = Y(0) + \sqrt{\nu \frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}} B_A(t) - \sqrt{\nu \frac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]}} B_D(t) + \theta t \stackrel{\mathrm{d}}{=} Y(0) + \sqrt{\nu \left(\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} + \frac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]} \right)} B(t) + \theta t$$

where B is another standard Brownian motion. Finally by another application of continuous mapping, we have the following result [e.g. 13].

Proposition 3. *Let $\bar{Q}^n$ be the sequence of rescaled queue-length processes of the batch queues without special arrivals. Then, under the above assumptions and if $\bar{Q}^n(0) \Rightarrow Y(0)$ in $\mathbb{R}$ as $n \to \infty$: $\bar{Q}^n \Rightarrow R[Y]$ in $D[0,\infty)$, where $R[Y]$ is a regulated (or reflected) Brownian motion (RBM) with initial state $Y(0)$, drift θ and variance parameter $\nu \left(\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} + \frac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]} \right)$. For $\theta < 0$, its stationary distribution is exponential with mean $-\frac{\nu}{2\theta} \left(\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} + \frac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]} \right)$.*

3.2. GBQ queue-length process limit

To make the nth queue of the above sequence of SBQs a geometric batch-queue, special arrivals must be added in state 0 with rates given by Proposition 1. Then, the same proposition gives the parameter ρ in equation 2. In heavy traffic this equation must have a double root at $\rho = 1$; there is always one root at 1 and as n increases the valid root must also tend to 1. Thus, defining $f(x) = A(x^{-1}) + D(x) - A(1) - D(1)$, a necessary condition for a heavy-traffic solution is $\dot{f}(1) = 0$. i.e. $-\dot{A}(1) + \dot{D}(1) = 0$ or $\dot{A}(1) = \dot{D}(1)$, as expected. To find the asymptotic form of ρ_n for the nth queue, $f(x)$ is expanded as a Taylor series about $x = 1$ to second order, which gives the correct asymptotic result when the third derivatives of A and D exist (which is implied by finite third moments of the batch-size distributions) by the mean value theorem:

$$f(1+\epsilon) = f(1) + (-\dot{A}^n(1) + \dot{D}^n(1))\epsilon + (\ddot{A}^n(1) + 2\dot{A}^n(1) + \ddot{D}^n(1))\epsilon^2/2 = 0$$

so that $(\theta/\sqrt{n})\epsilon = (\ddot{A}^n(1) + 2\dot{A}^n(1) + \ddot{D}^n(1))\epsilon^2/2$ or

$$\epsilon = \frac{2\theta}{\sqrt{n}(\ddot{A}^n(1) + 2\dot{A}^n(1) + \ddot{D}^n(1))}$$

since $\epsilon \neq 0$. Now writing $\epsilon = -\kappa^n/m$, where $\kappa^n = -2\theta/\left(\ddot{A}^n(1) + 2\dot{A}^n(1) + \ddot{D}^n(1)\right)$ and $m = \sqrt{n}$, $\rho^n = (1 - \kappa^n/m)$ and so the equilibrium probability that the length of the geometric batch queue is greater than $N-1$ is $(1-\kappa^n/m)^{m(N/m)}$. Now, according to the asymptotic regime defined in the previous subsection, the queue length scales by $1/m$. Thus, $\pi_N^n = (1-\kappa^n/m)^{mv} \to e^{-\kappa^n v}$. The heavy-traffic limit is therefore an exponential random variable with mean

$$\lim_{n\to\infty} 1/\kappa^n = \lim_{n\to\infty} -\frac{\ddot{A}^n(1) + 2\dot{A}^n(1) + \ddot{D}^n(1)}{2\theta}$$

Noting that $\dot{A}^n(1) = A^n(1)\mathbb{E}[\alpha]$ and $\dot{D}^n(1) = D^n(1)\mathbb{E}[\beta]$, $\lim_{n\to\infty} \dot{A}^n(1) = \lim_{n\to\infty} \dot{D}^n(1) = \nu$. Similarly, $\ddot{A}^n(1) = A^n(1)(\mathbb{E}[\alpha^2] - \mathbb{E}[\alpha])$ and $\ddot{D}^n(1) = D^n(1)(\mathbb{E}[\beta^2] - \mathbb{E}[\beta])$, so that $\lim_{n\to\infty} 1/\kappa^n = -\frac{\nu}{2\theta} \left(\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} + \frac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]} \right)$, which is the same as the limit for the queue without special arrivals obtained in the previous section. This verifies the conjecture that the special arrivals have no effect in the heavy-traffic limit. The question of how good an approximation the geometric batch queue is to the corresponding queue with no special arrivals in moderate traffic is considered numerically in Section 5.

4. Heavy-traffic limits for a tandem network

We consider now a sequence of tandem networks of J batch queues in series indexed by n. Arrivals of batches occur at Poisson rate λ^n only to the first node and departures at Poisson rate μ_j^n from the jth node for $j \in \{1, \ldots, J\}$. We assume that arrival batch sizes are distributed according to the discrete probability distribution p_k^a for $k \geq 1$ and departure batch sizes according to $p_k^{d,j}$ and $p_k^{d,j}$, respectively, for $k \geq 1$, As before, the batch-size distributions are independent of n and can be arbitrary other than requiring the existence of the first four moments, which we write as $\mathbb{E}[\alpha^m]$ and $\mathbb{E}[\beta_j^m]$, respectively, for $m = 1, \ldots, 4$.

As in the previous section, we define the aggregate task arrival and service rates $\bar{\lambda}^n = \mathbb{E}[\alpha]\lambda^n$, $\bar{\mu}_j^n = \mathbb{E}[\beta_j]\mu_j^n$, respectively. The heavy-traffic regime we consider here assumes that for some fixed $\nu > 0$, we have, as $n \to \infty$, $\bar{\lambda}^n \to \nu$ and $\bar{\mu}_j^n \to \nu$ for all $j \in \{1, \ldots, J\}$; and that $\sqrt{n}(\bar{\lambda}^n - \bar{\mu}_1^n) \to \theta_1$ and $\sqrt{n}(\bar{\mu}_{j-1}^n - \bar{\mu}_j^n) \to \theta_j$ for $j \in \{2, \ldots, J\}$ and some $\theta_j \in \mathbb{R}$.

Analogously to the single node case, we first consider the heavy-traffic limit for the version of the network without special arrivals and where partial batches do not leave the network but are forwarded to the next node as normal batches are. We then consider the same limit in the case of the geometric batch queue.

4.1. The queue-length process limit for a standard tandem network

For each $n \geq 1$ and $j \in \{1, \ldots, J\}$ construct independent Poisson processes $A^n(t)$ and $D_j^n(t)$ of rates λ^n and μ_j^n, respectively and $\{\alpha_i\}_{i=1}^{\infty}$ and $\{\beta_{j,i}\}_{i=1}^{\infty}$ mutually independent and identically distributed positive-integer-valued random variables with distributions p_k^a and $p_k^{d,j}$ for $k \geq 1$, respectively. Then, assuming that the initial state of the jth node is distributed as $Q_j^n(0)$ (assumed independent of A^n, D_j^n, $\{\alpha_i\}_{i=1}^{\infty}$ and $\{\beta_{j,i}\}_{i=1}^{\infty}$), define:

$$X_1^n(t) = Q_1^n(0) + \int_0^t \alpha_{A^n(s)} \, \mathrm{d}A^n(s) - \int_0^t \beta_{1,D_1^n(s)} \, \mathrm{d}D_1^n(s)$$

$$X_j^n(t) = Q_j^n(0) + \int_0^t \beta_{j-1,D_{j-1}^n(s)} \, \mathrm{d}D_{j-1}^n(s) - \int_0^t \beta_{j,D_j^n(s)} \, \mathrm{d}D_j^n(s)$$

for $j \in \{2, \ldots, J\}$. Then the vector-valued queue-length process $\mathbf{Q}^n(t) \in \mathbb{R}_+^J$ can be obtained as the J-dimensional regulated version of $\mathbf{X}^n$, $\mathbf{Q}^n = R[\mathbf{X}^n]$, defined by $Q_1^n(t) = X_1^n(t) + Y_1^n(t)$ and for $j \in \{2, \ldots, J\}$:

$$Q_j^n(t) = X_j^n(t) - Y_{j-1}^n(t) + Y_j^n(t) \tag{8}$$

where $\mathbf{Y}^n$ is the minimal (simultaneously in all coordinates $j \in \{1, \ldots, J\}$) non-decreasing process in $D_{\mathbb{R}^J}[0, \infty)$ such that each $Q_j^n(t) \geq 0$ for all $t \in \mathbb{R}_+$. If we define the *reflection matrix* $\mathbf{R} \in \mathbb{R}^{J \times J}$ by:

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

then we may write alternatively $\mathbf{Q}^n(t) = \mathbf{X}^n(t) + \mathbf{R}\mathbf{Y}^n(t)$. Existence, uniqueness and continuity of the multi-dimensional reflection map are established in many places in the literature [e.g. 14, Section 14.2].

Then as in the last section define $\bar{A}^n(t) = A^n(nt)$, $\bar{D}_j^n(t) = D_j^n(nt)$, $\bar{\mathbf{X}}^n(t) = (1/\sqrt{n})\mathbf{X}^n(nt)$, $\bar{\mathbf{Q}}^n(t) = (1/\sqrt{n})\mathbf{Q}^n(nt)$ and:

$$\bar{M}_A^n(t) = \frac{1}{\sqrt{n}} \left(\int_0^t \alpha_{\bar{A}^n(s)} \, \mathrm{d}\bar{A}^n(s) - \bar{\lambda}^n nt \right), \bar{M}_{j,D}^n(t) = \frac{1}{\sqrt{n}} \left(\int_0^t \beta_{j,\bar{D}_j^n(s)} \, \mathrm{d}\bar{D}_j^n(s) - \bar{\mu}_j^n nt \right)$$

Note that, $\bar{\mathbf{Q}}^n(0) = (1/\sqrt{n})\mathbf{Q}^n(0)$ and $\bar{\mathbf{Q}}^n = R[\bar{\mathbf{X}}^n]$, where $\bar{X}_1^n(t) = \bar{Q}_1^n(0) + \bar{M}_A^n(t) - \bar{M}_{1,D}^n(t) + \sqrt{n}t(\bar{\lambda}^n - \bar{\mu}^n)$ and $\bar{X}_j^n(t) = \bar{Q}_j^n(0) + \bar{M}_{j-1,D}^n(t) - \bar{M}_{j,D}^n(t) + \sqrt{n}t(\bar{\mu}_{j-1}^n - \bar{\mu}_j^n)$ for $j \in \{2, \ldots, J\}$.

As in the previous section, by Lemma 1, all of the $\bar{M}_A^n$ and $\bar{M}_{j,D}^n$ are locally square-integrable martingales with respect to the filtration $\mathcal{F}^n(t) = \sigma\{Q^n(0), \bar{A}^n(s), \bar{D}_j^n(s), \alpha_{\bar{A}^n(s)}, \beta_{j,\bar{D}^n(s)} : 0 \leq s \leq t, j \in \{1, \ldots, J\}\}$ and their predictable quadratic variation processes are $\langle \bar{M}_A^n \rangle(t) = \mathbb{E}[\alpha^2]\lambda^n t$ and $\langle \bar{M}_{j,D}^n \rangle(t) = \mathbb{E}[\beta_j^2]\mu_j^n t$ for $j \in \{1, \ldots, J\}$. As before, we have that the predictable covariations $\langle \bar{M}_A^n, \bar{M}_{j,D}^n \rangle(t) = \mathbf{0}$ w.p. 1 and $\langle \bar{M}_{i,D}^n, \bar{M}_{j,D}^n \rangle(t) = \mathbf{0}$ w.p. 1 when $i \neq j$. Decay of the mean squared jumps of the martingale terms can be proved exactly as in the last section.

Application of the martingale functional central limit theorem then yields, as $n \to \infty$:

$$(\bar{M}_A^n, \bar{M}_{1,D}^n, \ldots, \bar{M}_{J,D}^n) \Rightarrow \left(\sqrt{\nu \frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}} B_A, \sqrt{\nu \frac{\mathbb{E}[\beta_1^2]}{\mathbb{E}[\beta_1]}} B_{1,D}, \ldots, \sqrt{\nu \frac{\mathbb{E}[\beta_J^2]}{\mathbb{E}[\beta_J]}} B_{J,D} \right)$$

where the B_A and $B_{j,D}$ are independent standard Brownian motions and the convergence is weak on

$D_{\mathbb{R}^{J+1}}[0,\infty)$ endowed with the uniform topology. Assuming that $\bar{\mathbf{Q}}^n(0) \Rightarrow \mathbf{Y}(0)$ in $\mathbb{R}^J$ as $n \to \infty$, the continuous mapping theorem yields $\bar{\mathbf{X}}^n \Rightarrow \mathbf{Y}$, where:

$$Y_1(t) = Y_1(0) + \sqrt{\nu \frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}} B_A(t) - \sqrt{\nu \frac{\mathbb{E}[\beta_1^2]}{\mathbb{E}[\beta_1]}} B_{1,D}(t) + \theta_1 t$$

$$Y_j(t) = Y_j(0) + \sqrt{\nu \frac{\mathbb{E}[\beta_{j-1}^2]}{\mathbb{E}[\beta_{j-1}]}} B_{j-1,D}(t) - \sqrt{\nu \frac{\mathbb{E}[\beta_j^2]}{\mathbb{E}[\beta_j]}} B_{j,D}(t) + \theta_j t$$

for $j \in \{2, \ldots, J\}$. So $\mathbf{Y}(t)$ is a J-dimensional correlated Brownian motion with drift vector $\theta = (\theta_1, \ldots, \theta_J)^T$ and covariance matrix:

$$\mathbf{\Gamma} = \nu \begin{pmatrix} \sigma_1^2 & -\frac{\mathbb{E}[\beta_1^2]}{\mathbb{E}[\beta_1]} & 0 & \cdots & 0 & 0 \\ -\frac{\mathbb{E}[\beta_1^2]}{\mathbb{E}[\beta_1]} & \sigma_2^2 & -\frac{\mathbb{E}[\beta_2^2]}{\mathbb{E}[\beta_2]} & \cdots & 0 & 0 \\ 0 & -\frac{\mathbb{E}[\beta_2^2]}{\mathbb{E}[\beta_2]} & \sigma_3^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{J-1}^2 & -\frac{\mathbb{E}[\beta_{J-1}^2]}{\mathbb{E}[\beta_{J-1}]} \\ 0 & 0 & 0 & \cdots & -\frac{\mathbb{E}[\beta_{J-1}^2]}{\mathbb{E}[\beta_{J-1}]} & \sigma_J^2 \end{pmatrix}$$

where $\sigma_1^2 = \frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} + \frac{\mathbb{E}[\beta_1^2]}{\mathbb{E}[\beta_1]}$ and $\sigma_j^2 = \frac{\mathbb{E}[\beta_{j-1}^2]}{\mathbb{E}[\beta_{j-1}]} + \frac{\mathbb{E}[\beta_j^2]}{\mathbb{E}[\beta_j]}$ for $j \in \{2, \ldots, J\}$. Finally by another application of continuous mapping, we have the following result.

Proposition 4. *Let $\bar{\mathbf{Q}}^n$ be the sequence of rescaled queue-length processes of the tandem networks of batch queues without special arrivals and where partial batches are not discarded. Then, under the above assumptions and if $\bar{\mathbf{Q}}^n(0) \Rightarrow \mathbf{Y}(0)$ in $\mathbb{R}^J$ as $n \to \infty$: $\bar{\mathbf{Q}}^n \Rightarrow R[\mathbf{Y}]$ in $D_{\mathbb{R}^J}[0,\infty)$, where R is the J-dimensional regulator map with reflection matrix $\mathbf{R}$ defined above and $R[\mathbf{Y}]$ is an RBM.*

Such an RBM has a unique stationary distribution if and only if [15]:

$$\mathbf{R}^{-1}\theta = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \theta = \begin{pmatrix} \theta_1 \\ \theta_1 + \theta_2 \\ \theta_1 + \theta_2 + \theta_3 \\ \vdots \\ \theta_1 + \ldots + \theta_J \end{pmatrix} < 0 \tag{9}$$

where the inequality is component wise. Furthermore, this stationary distribution is of product form if and only if, additionally the *skew-symmetry* condition holds [15], $2\mathbf{\Gamma} = \mathbf{R}\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{R}^T$, where $\mathbf{\Lambda}$ is the diagonal matrix with diagonal entries Λ_{jj} equal to Γ_{jj}. In our case this equation is:

$$2\mathbf{\Gamma} = \nu \begin{pmatrix} 2\sigma_1^2 & -\sigma_1^2 & 0 & \cdots & 0 & 0 \\ -\sigma_1^2 & 2\sigma_2^2 & -\sigma_2^2 & \cdots & 0 & 0 \\ 0 & -\sigma_2^2 & 2\sigma_3^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2\sigma_{J-1}^2 & \sigma_{J-1}^2 \\ 0 & 0 & 0 & \cdots & -\sigma_{J-1}^2 & 2\sigma_J^2 \end{pmatrix}$$

So in order for this to be satisfied, we require that $\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} = \frac{\mathbb{E}[\beta_1^2]}{\mathbb{E}[\beta_1]}$ and $\frac{\mathbb{E}[\beta_j^2]}{\mathbb{E}[\beta_j]} = \frac{\mathbb{E}[\beta_{j-1}^2]}{\mathbb{E}[\beta_{j-1}]}$ for all $j \in \{2, \ldots, J-1\}$, i.e. that

$$\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} = \frac{\mathbb{E}[\beta_1^2]}{\mathbb{E}[\beta_1]} = \cdots = \frac{\mathbb{E}[\beta_{J-1}^2]}{\mathbb{E}[\beta_{J-1}]} \tag{10}$$

In this case, the marginal distribution at the jth node is exponential with mean [15]: $([-2\mathbf{\Gamma}^{-1}\mathbf{R}^{-1}\theta]_j)^{-1} = \frac{\nu\sigma_j^2}{-2(\theta_1+\ldots+\theta_j)}$.

4.2. The queue-length process limit for a tandem network of GBQs

The product-form solution for a tandem network of geometric batch queues is given by the rate equations of RCAT, as discussed in Section 2.3. Equations 7 and 6 can be solved successively: equation 7 gives ρ_1, then equation 6 applied successively for $j = 1, 2, \ldots, J-1$ gives $\rho_2, \rho_3, \ldots, \rho_J$ respectively. Each instance of j gives an equation in two variables and to find the asymptotic behaviour of ρ_{j+1} we again expand to second order in small quantities (valid when the third moments of the batch sizes are finite). Let $f_j(x,y) = D_j(x/y) + D_{j+1}(y) - D_j(x) - D_{j+1}(1)$. The equation $f_j(x,y) = 0$ is always satisfied by $y = 1$ for all x, so a necessary condition for a heavy-traffic solution (where all queues are saturated) is $\partial f/\partial y|_{y=1} = 0$.

Differentiation of f gives, at the point $x = y = 1$, $\nabla f = (0, -\dot{D}_j(1) + \dot{D}_{j+1}(1))$ and the Hessian of f is

$$H(f) = \begin{bmatrix} 0 & -\dot{D}_j(1) - \ddot{D}_j(1) \\ -\dot{D}_j(1) - \ddot{D}_j(1) & 2\dot{D}_j(1) + \ddot{D}_j(1) + \ddot{D}_{j+1}(1) \end{bmatrix}$$

Therefore, to second order,

$$\begin{aligned} f_j(1+\epsilon_j, 1+\epsilon_{j+1}) &= f_j(1,1) + (\epsilon_j, \epsilon_{j+1}).\nabla f + (\epsilon_j, \epsilon_{j+1}).H(f).(\epsilon_j, \epsilon_{j+1})^T/2 \\ &= \epsilon_{j+1}(\dot{D}_{j+1}(1) - \dot{D}_j(1)) - \epsilon_j\epsilon_{j+1}(\dot{D}_j(1) + \ddot{D}_j(1)) \\ &\quad + \epsilon_{j+1}^2(2\dot{D}_j(1) + \ddot{D}_j(1) + \ddot{D}_{j+1}(1))/2 \\ &= 0 \end{aligned}$$

Thus, since $\epsilon_{j+1} \neq 0$, for $j = 1, \ldots, J-1$, we obtain the linear recurrence:

$$\begin{aligned} \epsilon_{j+1} &= \frac{\dot{D}_j(1) - \dot{D}_{j+1}(1) + \epsilon_j(\dot{D}_j(1) + \ddot{D}_j(1))}{\dot{D}_j(1) + (\ddot{D}_j(1) + \ddot{D}_{j+1}(1))/2} = \frac{\theta_{j+1}/\sqrt{n} + \epsilon_j(\dot{D}_j(1) + \ddot{D}_j(1))}{\dot{D}_j(1) + (\ddot{D}_j(1) + \ddot{D}_{j+1}(1))/2} \qquad (11) \\ \epsilon_1 &= \frac{\theta_1}{\sqrt{n}(\dot{A}^n(1) + (\ddot{A}^n(1) + \ddot{D}_1^n(1))/2)} \qquad \text{(as for the single batch queue)} \end{aligned}$$

Under condition 10, the recurrence 11 simplifies to

$$\epsilon_{j+1} = \epsilon_j + \frac{\theta_{j+1}}{\sqrt{n}(\dot{D}_j(1) + \ddot{D}_j(1))} = \epsilon_j + \frac{2\theta_{j+1}}{\sqrt{n}\nu\sigma^2}$$

for $j = 1, \ldots, J-2$, with

$$\epsilon_1 = \frac{\theta_1}{\sqrt{n}(\dot{A}^n(1) + (\ddot{A}^n(1) + \ddot{D}_1^n(1))/2)} = \frac{2\theta_1}{\sqrt{n}\nu\sigma^2}$$

where $\sigma = \sigma_1 = \ldots = \sigma_{J-1}$. Thus, $\epsilon_j = \frac{2\sum_{k=1}^{j}\theta_k}{\sqrt{n}\nu\sigma^2}$ for $j = 1, \ldots, J-1$ and, finally,

$$\epsilon_J = \frac{\theta_J/\sqrt{n} + \epsilon_{J-1}(\dot{D}_{J-1}(1) + \ddot{D}_{J-1}(1))}{\dot{D}_{J-1}(1) + (\ddot{D}_{J-1}(1) + \ddot{D}_J(1))/2} = \frac{2\theta_J}{\sqrt{n}\nu\sigma_J^2} + \frac{2\sigma^2\sum_{k=1}^{J-1}\theta_k}{(\sqrt{n}\nu\sigma^2)\sigma_J^2} = \frac{2\sum_{k=1}^{J}\theta_k}{\sqrt{n}\nu\sigma_J^2}$$

similarly. The limit is therefore a product form of exponential distributions with means $\frac{-\nu\sigma_j^2}{2\sum_{k=1}^{j}\theta_k}$ at nodes $j = 1, \ldots, J$, in agreement with the heavy-traffic limit of the previous section. However, the product forms still have a limit even when condition 10 does not hold. This suggests that a new heavy-traffic diffusion limit should hold for the GBQ networks in this case, and, furthermore, that they have a product-form stationary

distribution given by the above explicit limit.

The subject of a paper currently in preparation is to characterise the corresponding diffusion limit explicitly. It is fairly clear to see however that it will behave similarly to a standard multi-dimensional Brownian motion while it remains inside the non-negative orthant but, induced by the non-neglible limiting effect of special arrivals and non-forwarded partial batches, will have additional special boundary behaviour when one or more of its components is at the origin. This kind of process is likely related to various forms of *sticky Brownian motion* and other Brownian motion processes [16, 17, 18], and, in light of the results of this paper, we know *in advance* that this new class of diffusion limit will admit a product-form stationary distribution.

4.3. Limiting regimes in a two-node tandem network

In a tandem pair of batch queues, the special arrivals to node 1 cause an increase in net arrival rate to node 2 at *all* queue lengths there, which is not the case for the special arrivals to node 2. We therefore expect that the special arrivals to node 2 have no effect in heavy traffic (as in a single batch queue), but that this may not be the case for the special arrivals to node 1. Similarly, discarding partial batches causes a *decrease* in net arrival rate at all queue lengths of node 2 – the special arrivals to node 1 and discarding of partial batches from node 1 work "against each other" as far as the queue length at node 2 is concerned. The following proposition shows that when these net throughputs are equal, the GBQ limit is the same as the SBQ heavy-traffic limit.

Proposition 5. *In a two-node tandem batch-network of GBQs at equilibrium, in which the batch sizes have finite first two moments, the GBQ limit is the same as the SBQ heavy-traffic limit given by the composition of RBMs in Section 4.1 if and only if the time-averaged arrival rate of the special arrivals at the first node is equal to the time-averaged rate of discarding partial batches from the first node.*

Proof. For a single GBQ with the usual notation, the time-averaged arrival rate of the special arrivals is $(1-\rho)\dot{A}_0(1)$ where $A_0(z) = \frac{A(\rho^{-1})\rho z - A(z)}{1-\rho z}$. Now,

$$\begin{aligned}
\dot{A}_0(1) &= \frac{\rho(A(\rho^{-1})\rho - A(1))}{(1-\rho)^2} + \frac{A(\rho^{-1})\rho - \dot{A}(1)}{1-\rho} = \frac{\rho(A(\rho^{-1}) - A(1)) - (1-\rho)\dot{A}(1)}{(1-\rho)^2} \\
&= \frac{\rho\big(-\dot{A}(1)(\rho-1) + (\ddot{A}(1) + 2\dot{A}(1))(\rho-1)^2/2\big) - (1-\rho)\dot{A}(1)}{(1-\rho)^2} = \frac{\ddot{A}(1)\rho(1-\rho)^2/2 + (1-\rho)^3\dot{A}(1)}{(1-\rho)^2} \\
&= \frac{\ddot{A}(1)}{2}
\end{aligned}$$

to zero order in $1-\rho$ on expanding $A(\rho^{-1})$ in powers of $\rho - 1$. Similarly, the time-averaged rate of discarding partial batches from node 1 is

$$
\begin{aligned}
(1-\rho_1)\sum_{i=1}^{\infty}\rho_1^i i\sum_{j=i+1}^{\infty}d_j &= (1-\rho_1)\sum_{j=2}^{\infty}d_j\rho_1\sum_{i=0}^{j-1}i\rho_1^{i-1}\\
&= \frac{\rho_1}{1-\rho_1}\sum_{j=2}^{\infty}d_j\Big(1-\rho_1^j-(1-\rho_1)j\rho_1^{j-1}\Big) \qquad \text{after some simplification}\\
&= \frac{\rho_1\big(D_1(1)-D_1(\rho_1)-(1-\rho_1)\dot{D}_1(\rho_1)\big)}{1-\rho_1}\\
&= \frac{\rho_1\big(-\dot{D}_1(1)(\rho_1-1)-\ddot{D}_1(1)(\rho_1-1)^2/2-(1-\rho_1)(\dot{D}_1(1)+\ddot{D}_1(1)(\rho_1-1))\big)}{1-\rho_1}\\
&= \ddot{D}_1(1)(1-\rho_1)/2 \qquad \text{to first order in } 1-\rho_1
\end{aligned}
$$

Thus, as $\rho_1 \to \infty$, the time-averaged special arrival and partial batch discard rates are equal iff $\ddot{A}(1) = \ddot{D}(1)$, whereupon the GBQ limit is the same as the SBQ heavy-traffic limit by the result of Section 4.1. ♢

5. Numerical results

5.1. Single node

For the regime defined in terms of the scaling parameter n in Section 3, we chose various probability distributions, $p_\cdot^a$ and $p_\cdot^d$, with finite support for the arrival and departure batch sizes (α and β), respectively. The maximum batch size is 3 for arrivals and 2 for departures and we write $p_i^a = \mathbb{P}(\alpha = i)$ $(i = 1, 2), p_3^a = 1 - p_1^a - p_2^a, p_1^d = \mathbb{P}(\beta = 1), p_2^d = 1 - p_1^d$. The respective batch-arrival and batch-departure rates are $4/\mathbb{E}[\alpha] - 1/\sqrt{n}$ and $4/\mathbb{E}[\beta] + 1/\sqrt{n}$. Thus, in the heavy-traffic limit, the task-arrival and task-departure rates are both 4. The results are summarised as cumulative distribution functions (CDFs) in Figures 1 and 2, which show graphs for two parameterisations of the queue at increasing utilisations $(0.664720, 0.832337, 0.871748)$ and $(0.647551, 0.830431, 0.871148)$. It can be seen that the GBQ approximates well the exact result (with no special arrivals), and also approaches the heavy-traffic limit quite quickly. The CDFs are similar for other cases, for example with geometric batch sizes (not shown).

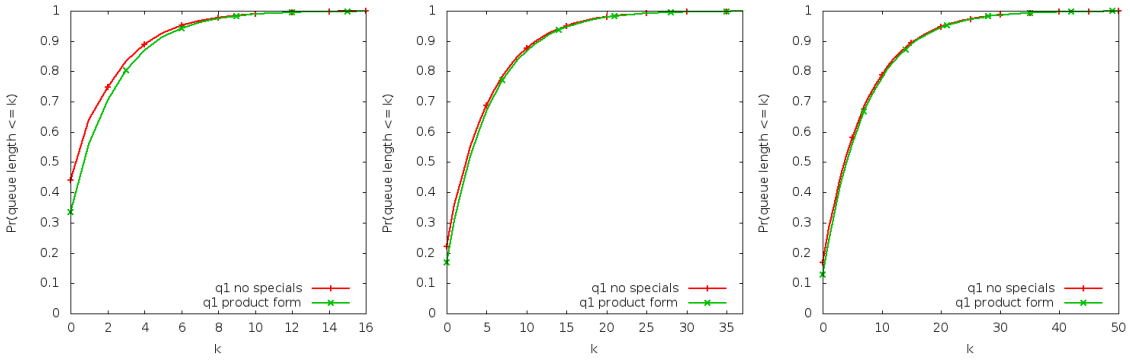

Figure 1: Single-node stationary queue lengths at scaling parameter $n = 1, 5$ and 9 from left to right. The model parameterisation is given by $p_1^a = 0.857143, p_2^a = 0, p_3^a = 0.142857, p_1^d = p_2^d = 0.5$.

5.2. Tandem network

Tandem networks are more interesting because the input point-process to the second queue is no longer Poisson and there is also the choice of discarding partial batches (as assumed in the product-form GBQ

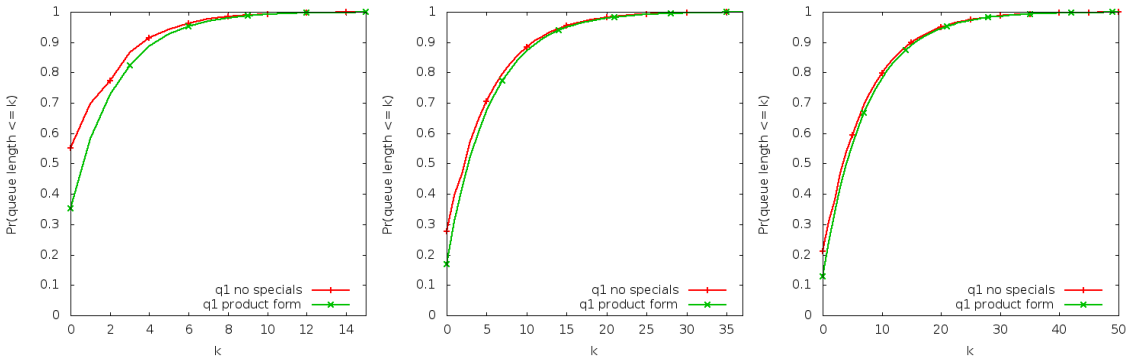

Figure 2: Single-node stationary queue lengths at scaling parameter $n = 1, 5$ and 9 from left to right. The model parameterisation is given by $p_1^a = 0.6, p_2^a = 0, p_3^a = 0.4, p_1^d = p_2^d = 0.5$.

network) or forwarding them to the second queue (as in the SBQ network). We considered a tandem network of two nodes under the regime defined in terms of the scaling parameter n in Section 3.1, where the skew-symmetry condition *does not hold*. We used the same parameterisation for the first queue as in the previous section and fed its output into the second queue – with either discarding or forwarding of partial batches when the queue length at the first queue is 1 – i.e. less than the maximum batch size of 2. The external batch-arrival rate at queue 2 is zero and its batch-departure rate is $4/\mathbb{E}[\beta] + 1/(8\sqrt{n})$. Thus, in the limit, the task-arrival and task-departure rates are all 4 again. The equilibrium behaviour of the tandem pair of GBQs is given by the solution (ρ_1, ρ_2) of equations 6, 7, which are solved numerically for each model parameterisation and scaling factor n.

To examine the accuracy of the GBQ approximation in networks with discarded partial departure-batches, as well as the effect of discarding, we considered a series of parameterisations corresponding to scaling parameter $n = 5, 10, \ldots, 50$ for a tandem pair given by $p_1^a = 1, p_1^d = p_2^d = 0.5$ at the first queue, together with no external normal arrivals and unit batch-size departures at the second queue. The mean queue lengths at the two queues and their standard deviations are listed in Table 1 and show good agreement (first two sets of four columns). However, when partial batches are forwarded instead of being discarded (last four columns), the product-form approximation is very poor, which shows the significance of such forwarding when the skew-symmetry condition does not hold, even at heavy traffic when their influence could be expected to be negligible since then queue 1 would rarely have only one task.

5.2.1. Skew symmetry

We next considered a network where the skew-symmetry condition holds, specifically with parameterisation $p_1^a = 0.857143, p_2^a = 0, p_3^a = 0.142857, p_1^{d,1} = p_2^{d,1} = 0.5, p_1^{d,2} = 1$. CDFs for the length of queue 2 are shown at increasing utilisations $(0.752707, 0.885707, 0.914116)$ in Figure 3, which reveals increasingly good agreement between the SBQ and the GBQ as the traffic intensity increases, as we expect from the equivalence established in Section 4.2. However, in the case with no forwarding, the agreement is less accurate, revealing the potentially non-negligible influence of forwarding in the heavy-traffic limit. The first queue is parameterised exactly as in the previous section and so has CDF as in Figure 1.

	No special arrivals, no forwarding				Product-form				No special arrivals and forwarding			
	Queue 1		Queue 2		Queue 1		Queue 2		Queue 1		Queue 2	
n	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std dev	Mean	Std Dev	Mean	Std Dev
5	4.308	4.782	5.856	6.551	4.308	4.782	6.083	6.564	4.308	4.782	7.857	8.087
10	6.287	6.769	8.674	9.385	6.287	6.769	8.908	9.395	6.287	6.769	11.56	11.60
15	7.805	8.290	10.84	11.56	7.805	8.290	11.07	11.56	7.805	8.290	14.40	14.31
20	9.084	9.571	12.66	13.39	9.084	9.571	12.90	13.39	9.084	9.571	16.80	16.58
25	10.21	10.70	14.27	15.00	10.21	10.70	14.51	15.00	10.21	10.70	18.91	18.59
30	11.23	11.72	15.72	16.45	11.23	11.72	15.97	16.46	11.23	11.72	20.82	20.40
35	12.17	12.66	17.06	17.79	12.17	12.66	17.30	17.80	12.17	12.66	22.57	22.07
40	13.04	13.53	18.31	19.04	13.04	13.53	18.55	19.04	13.04	13.53	24.21	23.62
45	13.86	14.35	19.47	20.21	13.86	14.35	19.72	20.21	13.86	14.35	25.74	25.07
50	14.63	15.12	20.58	21.32	14.63	15.12	20.82	21.32	14.63	15.12	27.19	26.45

Table 1: Means and standard deviations in the tandem pair of GBQs with arrival batch-size probability mass function $p_1^a = 1$ at node 1 and departure batch-size probability mass functions $p_1^{d,1} = p_2^{d,1} = 0.5$ at node 1 and $p_1^{d,2} = 1$ at node 2. n is the scaling factor.

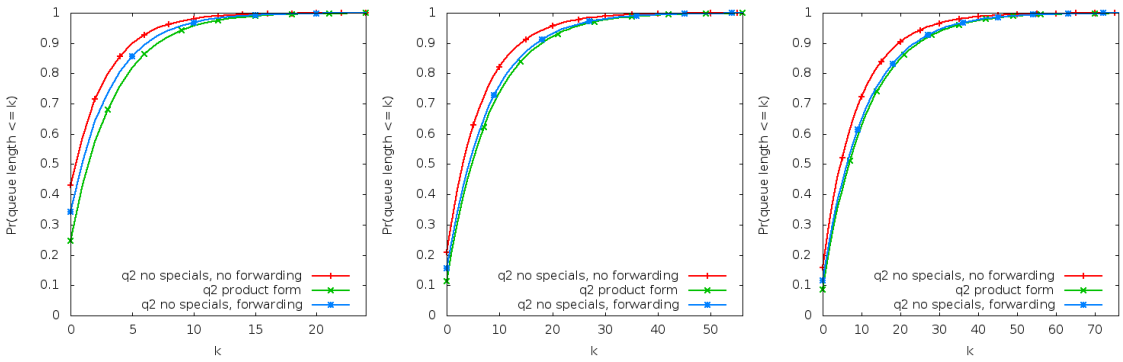

Figure 3: Second node with skew-symmetry condition: stationary queue lengths at scaling parameter $n = 1, 5$ and 9 from left to right. First node unchanged from Figure 1.

The CDFs for a similar tandem structure, in which the skew-symmetry condition does not hold,[3] are shown at increasing utilisations $(0.729600, 0.883224, 0.913342)$ for queue 2 in Figure 4; again the CDFs for queue 1 are the same as in Figure 2. The parameterisation in this case is $p_1^a = 0.6, p_2^a = 0, p_3^a = 0.4, p_1^{d,1} = p_2^{d,1} = 0.5, p_1^{d,2} = 1$ and, as expected, we notice much poorer agreement.

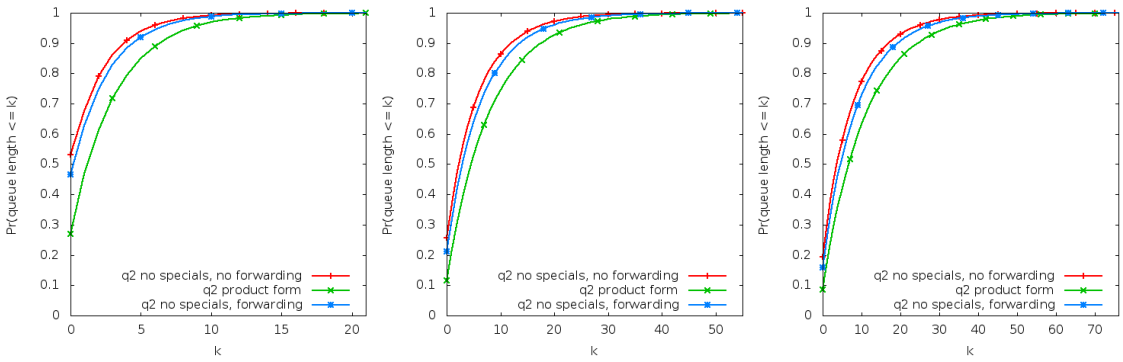

Figure 4: Second node without skew-symmetry condition: stationary queue lengths at scaling parameter $n = 1, 5$ and 9 from left to right. First node unchanged from Figure 2.

[3]To be specific, we take $\mathbb{E}[\alpha^2]/\mathbb{E}[\alpha] = 2\mathbb{E}[\beta_1^2]/\mathbb{E}[\beta_1]$.

5.2.2. Larger batch sizes

Tandem networks of two queues with larger batch sizes exhibit the same qualitative effects. Figure 5 shows the CDFs for queue 2, where the skew-symmetry condition does (left, with better agreement) and does not[4] (middle, worse agreement) hold. The respective parameterisations, both at scaling parameter $n = 430$, are $p_1^a = 0.797101, p_{10}^a = 0.202899, p_1^{d,1} = 0.4, p_8^{d,1} = 0.6, p_1^{d,2} = 1$ (giving utilisation $\rho_2 = 0.99164$) and $p_1^a = 0.94697, p_{10}^a = 0.05303, p_1^{d,1} = 0.4, p_8^{d,1} = 0.6, p_1^{d,2} = 1$ (giving utilisation $\rho_2 = 0.990437$).

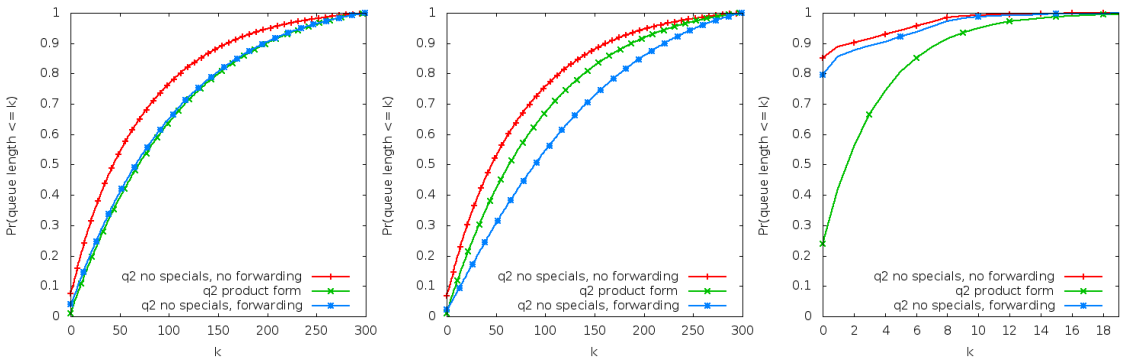

Figure 5: Second node stationary queue lengths with bigger batches: skew-symmetry, non-skew-symmetry, lighter traffic & skew-symmetry.

5.2.3. Lower utilisation

Finally, to see the extent of the error in the approximation at lower traffic intensities, we reduced the scaling parameter to $n = 0.8$, with larger batches at the previous parameterisation for skew symmetry, i.e. $p_1^a = 0.797101, p_{10}^a = 0.202899, p_1^{d,1} = 0.4, p_8^{d,1} = 0.6, p_1^{d,2} = 1$ (giving utilisation $\rho_2 = 0.761357$). This gives the third graph of Figure 5.

6. Conclusion

Batch queueing networks are a key modelling tool for the analysis of bursty traffic which is a crucial feature of many energy-efficient systems since it introduces significant idle periods during which servers can be powered down. However, in general, networks of batch queues do not admit a tractable analysis.

To illustrate a potential role of such models, recall that, in general, burstier traffic gives poorer performance, but facilitates greater energy saving through lower device power levels. We consider briefly one optimisation strategy for batch networks that allows variable batch sizes but fixed mean task, or byte, arrival rate. To aid load balancing, we introduce additional arrivals to idle nodes. These can reasonably be chosen to be larger in intensity for a node with lower utilisation and so the special arrival process is at least a plausible candidate. The average energy saving per unit time, E, is then given by the mean idle time, the reciprocal of the batch arrival rate, proportional to the batch size. The mean response time, R, of a task, on the other hand, can be obtained from the mean node occupancy and Little's result. Such quantities are

[4]Here, we took $2\mathbb{E}[\alpha^2]/\mathbb{E}[\alpha] = \mathbb{E}[\beta_1^2]/\mathbb{E}[\beta_1]$.

routine to obtain in product-form networks, [e.g. 19], so GBQs provide at least a benchmark – a model parameterisation for which exact numerical performance indices can be computed efficiently. The aim is to maximise energy saving and to minimise response time, so a suitable utility function to maximise is the quantity $E^{\alpha}R^{-\beta}$ for suitable positive constants $\alpha, \beta > 0$ chosen from user profile data and the specific objective function. Alternatively, we could maximise E subject to a maximum specified value allowed for R. Often, optimisation with respect to higher moments or quantiles is required, for example maximising average energy saving given that 90% of response times must be less than one second. Such analysis is possible for GBQ models using our results on response-time distributions; see the Appendix.

In any operating system for enterprise-scale storage, there is always a considerable amount of background workload arising from tasks like deduplication, RAID reconstruct and garbage collection. These are highly bursty in nature but can be broken up into subtasks and scheduled by the operating system according to the installation's particular objectives (e.g. low latency, low energy consumption, high reliability etc). It may be worthwhile to schedule such tasks when nodes are idle in such a way that a product form is attained for the relevant network of devices. Any excess background load can be scheduled along with the normal load so as to approximate batch size distributions jointly compatible with the background load to yield a product form. If this can be done to a good degree of accuracy, performance would become predictable, facilitating a range of dynamic, run-time optimisations.

To summarise, we have characterised a class of batch queues and tandem networks of such queues that have a geometric product-form distribution at equilibrium, which we have called *geometric batch queues*. The product form arises due to the addition of "special" arrival streams which are activated only when a node is idle, and the requirement that partial service batches at any node must leave the network. We show that, somewhat surprisingly, these additional arrival streams and the non-forwarding of partial batches is, in general, non-negligible in heavy-traffic. In fact, these extra effects can only be ignored in heavy traffic when a particular *skew-symmetry* condition holds, which matches that required for the limiting regulated Brownian motion diffusion limit to have a product-form stationary distribution.

This has two interesting implications, one practical and one more theoretical. Firstly, it validates the use of the geometric batch network as a moderate-traffic approximation to the standard network without special arrivals and with or without the forwarding of partial batches, *but only* when the skew-symmetry condition holds. Secondly, in the case that the skew-symmetry condition *does not hold*, the heavy-traffic limit of the product-form geometric batch network still exists suggesting the existence of a new diffusion limit with a product-form stationary distribution.

References

[1] P. Harrison, Turning back time in Markovian process algebra, Theoretical Computer Science 290 (3) (2003) 1947–1986. URL `http://aesop.doc.ic.ac.uk/pubs/rcat/`

[2] S. Mahevas, G. Rubino, Bound computation of dependability and performance measures, IEEE Transactions on Computers 50 (5) (2001) 399–413.

[3] P. Buchholz, Bounding stationary results of Tandem networks with MAP input and PH service time distributions, in: Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '06/Performance '06, ACM, New York, NY, USA, 2006, pp. 191–202.
[4] X. Chao, M. Miyazawa, M. Pinedo, Queueing networks: customers, signals and product form solutions, Wiley, 1999.
[5] P. Harrison, Compositional reversed markov processes, with applications to G-networks, Performance Evaluation 57 (2004) 379–408.
URL http://aesop.doc.ic.ac.uk/pubs/product/
[6] M. Miyazawa, P. Taylor, A geometric product-form distribution for a queueing network with non-standard batch arrivals and batch transfer, Adv. Appl. Prob. 29 (1997) 523–544.
[7] J. Jackson, Jobshop-like queueing systems, Management Science 10 (1) (1963) 131–142.
[8] D. Gross, C. M. Harris, Fundamentals of Queueing Theory, Wiley-Sons, 1985.
[9] P. Harrison, C. Llado, R. Puigjaner, A unified approach to modelling the performance of concurrent systems, Simulation Modelling Practice and Theory 17 (2009) 1445–1456.
[10] G. Pang, R. Talreja, W. Whitt, Martingale proofs of many-server heavy-traffic limits for Markovian queues, Probability Surveys 4 (2007) 193–267.
[11] W. Whitt, Proofs of the martingale FCLT, Probability Surveys 4 (2007) 268–302.
[12] H. A. David, H. N. Nagaraja, Order statistics, 3rd Edition, Wiley, 2003.
[13] J. M. Harrison, Brownian Motion and Stochastic Flow Systems, John Wiley & Sons, 1990.
[14] W. Whitt, Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues, Springer, 2002.
[15] J. M. Harrison, R. J. Williams, Brownian models of open queueing networks with homogeneous customer populations, Stochastics 22 (2) (1987) 77–115.
[16] A. V. Skorokhod, Stochastic Equations for Diffusion Processes in a Bounded Region, Theory of Probability & Its Applications 6 (3) (1961) 264–274.
[17] J. M. Harrison, A. J. Lemoine, Sticky Brownian Motion as the Limit of Storage Processes, Journal of Applied Probability 18 (1) (1981) 216–226.
[18] M. Z. Racz, M. Shkolnikov, Multidimensional sticky Brownian motions as limits of exclusion processes, Preprint submitted to arXiv (Feb. 2013).
URL http://arxiv.org/abs/1302.2678
[19] G. Bolch, S. Greiner, H. Meer, K. Trivedi, Queueing Networks and Markov Chains, Modeling and Performance Evaluation with Computer Science Applications, Wiley, 2006.
[20] A. A. Puhalskii, M. I. Reiman, The multiclass GI/PH/N queue in the Halfin-Whitt regime, Advances in Applied Probability 32 (2) (2000) 564–595.
[21] A. Puhalskii, On the Invariance Principle for the First Passage Time, Mathematics of Operations Research 19 (4) (1994) 946–954.

Appendix

Appendix A. Sojourn-time distributions

Appendix A.1. Forward and reversed sojourn times in a single GBQ

Sojourn times in a tandem pair of batch queues are investigated (in Section 4.2) in terms of the reversed sojourn time at the first node and the forward sojourn time at the second node. We therefore now consider both the forward and reversed sojourn times in a single batch queue. The service time random variable S, i.e. time to the next departure of either kind in a non-empty queue of length k, is an exponential random variable with constant parameter $D(1)$, with probability distribution function having LST $S^*(\theta) = \frac{D(1)}{\theta + D(1)}$. Let the random variable R_m denote the remaining sojourn time of a task in position $m+1$ in a batch-queue; i.e. that has m tasks in front of it. Then R_m has probability distribution function $R_m(t)$ with LST $R_m^*(\theta) = \mathbb{E}[e^{-\theta R_m}]$. Define the generating function

$$G_R(x;\theta) = \sum_{m=0}^{\infty} R_m^*(\theta) x^m$$

Then we have the following result (we will often omit the argument θ when it's absence is obvious):

Proposition 6. *In an equilibrium geometric batch queue defined by the generating functions $A(z), D(z)$,*

$$G_R(x;\theta) = \frac{D(1) - D(x)}{(1-x)(\theta + D(1) - D(x))}$$

The corresponding generating function for the LSTs of the remaining sojourn time distributions in the reversed process is:

$$G'_R(x;\theta) = \frac{S'^*(\theta)(A(\rho^{-1}) - A(\rho^{-1}x))}{(1-x)(A(\rho^{-1}) - A(\rho^{-1}x)S'^*(\theta))}$$

where $S'^*(\theta) = A(\rho^{-1})/(\theta + A(\rho^{-1}))$.

Proof. When the task is at the front of the queue, with no task in front of it, $R_0^* = S^*$. Otherwise, for $n > 0$,

$$R_n^*(\theta) = \frac{S^*}{D(1)} \left[\sum_{i=n+1}^{\infty} d_i + \sum_{i=1}^{n} d_i R_{n-i}^*(\theta) \right] \tag{A.1}$$

Thus,

$$\begin{aligned} G_R(x) &= \frac{S^*}{D(1)} \sum_{n=0}^{\infty} \left(x^n \sum_{i=n+1}^{\infty} d_i + \sum_{i=1}^{n} d_i R_{n-i}^* x^n \right) = \frac{S^*}{D(1)} \left(\sum_{i=1}^{\infty} \sum_{n=0}^{i-1} d_i x^n + \sum_{i=1}^{\infty} \sum_{n=0}^{\infty} d_i R_n^* x^{n+i} \right) \\ &= \frac{S^*}{D(1)} \left(\frac{D(1) - D(x)}{1-x} + D(x) G_R(x) \right) \qquad \text{as required.} \end{aligned}$$

The result for the reversed process follows from Proposition 2. ♢

When normal departure batches have geometric distribution, with probability mass function $D(1)(1-\delta)\delta^{k-1}$, straightforward algebra yields

$$G_R(x) = \frac{D(1)}{\theta(1-\delta x) + (1-x)D(1)} \tag{A.2}$$

Appendix A.2. Response times in a tandem pair of GBQs

We consider response times on a path through a tandem network of two nodes, in which tasks that pass between the nodes may arrive at the first one and leave from the second one in either normal or special batches. We consider the *middle state* of the pair of nodes at the instant the tagged task leaves node 1 and enters node 2. This is defined as the ordered pair comprising the numbers of tasks left behind at node 1 and in front of the tagged task at node 2 – i.e. excluding the tagged task in both cases. We consider the reversed sojourn time $\tilde{T}_1(N_1, N_2, M, K)$ at node 1 and the forward sojourn time $T_2(N_1, N_2, M, K)$ at node 2, conditioned on the middle state (N_1, N_2) and numbers of tasks behind and in front of the tagged task in its own batch, (M, K). The conditional response time random variable is then $\tilde{T}_1(N_1, N_2, M, K) + T_2(N_1, N_2, M, K)$ and its probability distribution can be computed by deconditioning with respect to the middle-state and position-in-batch probabilities. In a product-form, tandem batch network, the probability distribution of the middle state (s, s') is precisely the set of equilibrium probabilities $(1-\rho_i)(1-\rho_{i+1})\rho_i^{s_i}\rho_{i+1}^{s_{i+1}}$, where (ρ_1, ρ_2) is the solution of the rate equations 6, 7. This is simply proved by flux arguments since the process is Markovian. The size of an in-transit tagged task's batch at equilibrium is independent of the middle state, and has probability generating function $D_1(\rho_1 z)/D_1(\rho_1)$; this can be shown similarly.

For simplicity, we assume that normal batches remain the same size when passing between nodes, which is also the typical situation. In a two-node batch network, given the probability mass function of the position of a tagged task in a batch at the instant the batch transits from node 1 to node 2, we can find the Laplace-

Stieltjes transform (LST) of the joint probability distribution of the two node-sojourn-times, and hence of the network's response-time distribution, at equilibrium. We define f_{mk} to be the joint probability that, in an arriving batch of size $m+k+1$, there are k tasks in front of and m tasks behind the tagged task. Three commonplace possibilities arise, where the tagged task is: (1) First in batch, whence $f_{mk} = \delta_{k0} d_{1,m+1}/D_1(1)$; (2) Last in batch, whence $f_{mk} = \delta_{m0} d_{1,k+1}/D_1(1)$; or (3) Random position, so that $f_{mk} = d_{1,m+k+1}/\dot{D}_1(1)$, by the standard backwards-forwards recurrence result for renewal periods.

Using the reversed process for the first queue, the forward process for the second queue and the equilibrium property of the middle state, as described above, we immediately have:

Proposition 7. *The tagged task's joint probability distribution of node-sojourn times has LST:*

$$T^*_{12}(\theta_1,\theta_2) = (1-\rho_1)(1-\rho_2)\sum_{m=0}^{\infty}\sum_{k=0}^{\infty} f_{mk} G^m_{\tilde{R}_1}(\rho_1;\theta_1) G^k_{R_2}(\rho_2;\theta_2)$$

where the functions $G^m_{\tilde{R}_1}, G^k_{R_2}$ are defined as in Proposition 6.

Unfortunately, this argument cannot be extended simply to larger networks because tasks propagating through the network must pass from node to node as normal tasks and not be discarded. However, for tandem networks of two nodes, our method of utilizing the "middle state" guarantees that the tagged task does pass between the two nodes and we can obtain numerical results at any traffic intensity when the support of the batch sizes is small (e.g. 2 or 3 different possible batch sizes) or for special cases such as geometric batch size.

Appendix A.2.1. Geometric batch sizes

The difficulty in general is calculating $G^k_R(x)$ but we can do this when the departure batch sizes at the node concerned are geometric; there is no constraint on the arrival batch sizes. Splitting the expression for $G_R(x)$ in equation A.2 into partial fractions and expanding in powers of x yields, for $m \geq 0$, $R^*_m(\theta) = \left(\frac{D(1)}{\theta+D(1)}\right)\left(\frac{\theta\delta+D(1)}{\theta+D(1)}\right)^m$, giving $G^k_R(x;\theta) = \sum_{m=0}^{\infty} x^m R^*_{k+m}(\theta) = \left(\frac{\theta\delta+D(1)}{\theta+D(1)}\right)^k \frac{D(1)}{\theta+D(1)-(\theta\delta+D(1))x}$. We then have:

Proposition 8. *When the arrival batch sizes at node 1 and the departure batch sizes at node 2 are both geometric, with generating functions $A_1(z) = A_1(1)(1-\alpha_1)z/(1-\alpha_1 z)$ and $D_2(z) = D_2(1)(1-\delta_2)z/(1-\delta_2 z)$ respectively, the LST of the joint node-sojourn time probability distribution in the tandem pair is:*

$$\begin{aligned} T^*_{12}(\theta_1,\theta_2) &= (1-\rho_1)(1-\rho_2)\left(\frac{A_1(\rho_1^{-1})}{\theta_1 + A_1(\rho_1^{-1}) - (\theta_1\alpha_1 + A_1(\rho_1^{-1}))\rho_1}\right)\left(\frac{D_2(1)}{\theta_2 + D_2(1) - (\theta_2\delta_2 + D_2(1))\rho_2}\right) \times \\ &\quad F\left(\frac{\theta_1\alpha_1 + A_1(\rho^{-1})}{\theta_1 + A_1(\rho^{-1})}, \frac{\theta_2\delta_2 + D_2(1)}{\theta_2 + D_2(1)}\right) \end{aligned}$$

where $F(z_1,z_2) = \sum_{m=0}^{\infty}\sum_{k=0}^{\infty} f_{mk} z_1^m z_2^k$ is the joint generating function of the probabilities f_{mk}.

Proof. The result follows by applying the above expression for $G^k_R(x;\theta)$ to the forwards process at node 2 and to the reversed process at node 1, and then substituting into Proposition 7. ♢

When the tagged task is first, last or randomly positioned in its batch in the middle state, it is straightforward to show that the generating function F is respectively defined by: (1) $F_f(z_1,z_2) = z_1^{-1} D_1(z_1)/D_1(1)$; (2) $F_\ell(z_1,z_2) = z_2^{-1} D_1(z_2)/D_1(1)$; and (3) $F_r(z_1,z_2) = \frac{D_1(z_2)-D_1(z_1)}{D_1(0)(z_2-z_1)}$.

Appendix B. Sojourn-time limits

We now consider the heavy-traffic limits of sojourn time in a single queue, first without and then with special arrivals, i.e. for the geometric batch queue in the latter case.

Appendix B.1. SBQ response time limit

Let the arrival and departure processes of the queue be $I^n(t) = \int_0^t \alpha_{A^n(s)}\,\mathrm{d}A^n(s)$ and $O^n(t) = I^n(t) - Q^n(t)$. Define the virtual waiting time process: $W^n(t) = \inf\{s \geq 0 : O^n(s) > I^n(t)\} - t$. Note that by the PASTA property $W^n(\infty)$ has the same distribution as the steady-state response time for the first customer in an arriving batch. Write $\bar{W}^n(t) = (1/\sqrt{n})W^n(nt)$ and note that:

$$\bar{W}^n(t) = \sqrt{n}(\inf\{s \geq 0 : (1/n)O^n(ns) > (1/n)I^n(nt)\} - t)$$

Recall from Section 4.1 that jointly:

$$\sqrt{n}((1/n)I^n(nt) - \bar{\lambda}^n t) \Rightarrow \sqrt{\nu\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}}B_A(t)$$

and:

$$\sqrt{n}((1/n)O^n(nt) - \bar{\lambda}^n t) = \sqrt{n}((1/n)I^n(nt) - \bar{\lambda}^n t) - \bar{Q}^n(t) \Rightarrow \sqrt{\nu\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}}B_A(t) - R[Y(t)]$$

Then using Lemma 2 in Appendix C regarding the continuity of the inverse map, we have the following result.

Proposition 9. *Let $\bar{W}^n$ be the sequence of virtual waiting time processes for the batch queues without special arrivals. Then under the above assumptions as $n \to \infty$, we have:*

$$\bar{W}^n \Rightarrow \frac{\sqrt{\nu\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}}B_A - \left(\sqrt{\nu\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}}B_A - R[Y]\right)}{\nu} = (1/\nu)R[Y]$$

For $\theta < 0$, the stationary distribution of $(1/\nu)R[Y]$ is exponential with mean $-\frac{1}{2\theta}\left(\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} + \frac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]}\right)$.

Appendix B.2. GBQ response time limit

We saw in Proposition 6 that the first task in a batch has exponential response time with mean $1/(D(1) - D(\rho))$ at equilibrium. In heavy traffic, with $\rho = 1 - \kappa^n/\sqrt{n}$, this becomes $\sqrt{n}/(\dot{D}^n(1)\kappa^n)$ – measured in task-units. Rescaling by $\sqrt{n}$ this becomes

$$1/(\dot{D}^n(1)\kappa^n) \to -\nu^{-1}\frac{\nu}{2\theta}\left(\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} + \frac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]}\right)$$

in agreement with the diffusion limit. We conclude this section by showing that the same result holds for all individual tasks in a batch.

We first define $G_R^k(x;\theta) = \sum_{m=0}^{\infty} x^m R^*_{k+m}(\theta)$, so that $G_R^0(x;\theta) \equiv G_R(x;\theta)$, and suppose there are $k \geq 0$ tasks ahead of the task tagged for response time in its arriving batch with probability $p(k)$. Then, by the PASTA property, the LST of the task's response- time distribution is $T^*(\theta) = (1-\rho)\sum_{k=0}^{\infty} p(k)G_R^k(\rho;\theta)$.

Proposition 10. *The response time of any task in a batch arriving at a geometric batch queue in heavy-traffic equilibrium has exponentially distributed response time with mean $-\frac{1}{2\theta}\left(\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} + \frac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]}\right)$, provided batch sizes have finite first and second moments.*

Proof. Using equation A.1, in the notation of Appendix A.1,

$$\begin{aligned}
G_R^k(\rho;\theta) &= \frac{S^*}{D(1)} \sum_{m=0}^{\infty} \left(\sum_{i=1}^{k+m} d_i R^*_{k+m-i} \rho^m + \sum_{i=k+m+1}^{\infty} d_i \rho^m \right) \\
&= \frac{S^*}{D(1)} \left(\sum_{i=1}^{\infty} \sum_{m=i-k\vee 0}^{\infty} d_i R^*_{k+m-i} \rho^m + \sum_{i=k+1}^{\infty} \sum_{m=0}^{i-k-1} d_i \rho^m \right) \\
&= \frac{S^*}{D(1)} \left(\rho^{-k} D(\rho) G_R(\rho;\theta) - \sum_{i=1}^{k-1} \sum_{m'=1}^{k-i} d_i R^*_{k-m'-i} \rho^{-m'} \right) + \frac{S^*}{D(1)} \sum_{i=k+1}^{\infty} d_i \frac{1-\rho^{i-k}}{1-\rho}
\end{aligned}$$

Therefore, by Proposition 6,

$$\begin{aligned}
T^*(\theta) &= \sum_{k=0}^{\infty} p(k) \frac{S^*}{D(1)} \left(\frac{\rho^{-k} D(\rho)(D(1)-D(\rho))}{\theta + D(1) - D(\rho)} - (1-\rho) \sum_{i=1}^{k-1} \sum_{m'=1}^{k-i} d_i R^*_{k-m'-i} \rho^{-m'} + \sum_{i=k+1}^{\infty} d_i (1-\rho^{i-k}) \right) \\
&= -\frac{S^* \epsilon}{D(1)} \left(\frac{D(1)\dot{D}(1)}{\theta - \dot{D}(1)\epsilon} + \sum_{k=0}^{\infty} p(k) \sum_{i=k+1}^{\infty} d_i (i-k) - \sum_{k=0}^{\infty} p(k) \sum_{i=1}^{k-1} \sum_{m'=1}^{k-i} d_i R^*_{k-m'-i} (1-m'\epsilon) \right)
\end{aligned}$$

in task units, to first order in ϵ, with error $o(\epsilon)$ by the second mean value theorem, since the second derivative of D exists by hypothesis. Applying the scaling factor $\sqrt{n}$ to the rates, this becomes

$$T^{n*}(\theta) = \frac{\kappa^n}{\theta + D(1)\sqrt{n}} \left(\frac{D(1)\dot{D}(1)\sqrt{n}}{\theta + \dot{D}(1)\kappa^n} + \sum_{k=0}^{\infty} p(k) \left(\sum_{i=k+1}^{\infty} d_i (i-k) - \sum_{i=1}^{k-1} \sum_{m'=1}^{k-i} d_i R^*_{k-m'-i} (1-m'\epsilon) \right) \right)$$

Now,

$$\left| \sum_{k=0}^{\infty} p(k) \sum_{i=k+1}^{\infty} d_i (i-k) \right| \quad \leq \quad \sum_{k=0}^{\infty} p(k) \sum_{i=0}^{\infty} d_i (i+k) = D(1)(\mathbb{E}[\beta] + \mathbb{E}[\gamma])$$

where γ is the random variable denoting the number of tasks ahead of the tagged task in its arriving batch, with probability mass function $p(\cdot)$, which has finite mean by hypothesis. Finally, since $R^*_{k-m'-i} \leq 1$,

$$\left| \sum_{k=0}^{\infty} p(k) \sum_{i=1}^{k-1} \sum_{m'=1}^{k-i} d_i R^*_{k-m'-i} (1-m'\epsilon) \right| \leq \left| \sum_{k=0}^{\infty} p(k) \sum_{i=1}^{k-1} d_i (k-i)(1-(k-i+1)\epsilon/2) \right| \leq \left| \sum_{k=0}^{\infty} p(k) D(1) k (1-k\epsilon/2) \right|$$

which is bounded since the task's batch size has finite first two moments. Hence, in the limit $n \to \infty$,

$$T^{n*}(\theta) \to \frac{\dot{D}(1)\kappa}{\theta + \dot{D}(1)\kappa} = \frac{\nu\kappa}{\theta + \nu\kappa} \quad \text{where } \kappa = \lim_{n\to\infty} \kappa^n = -\tfrac{2\theta}{\nu} \left(\tfrac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} + \tfrac{\mathbb{E}[\beta^2]}{\mathbb{E}[\beta]} \right)^{-1}.$$

Equivalently, $\sum_{k=0}^{\infty} p(k) G_R^k(\rho;\theta) \to G_R(\rho;\theta)$ in the heavy-traffic limit, which will be useful in Appendix B.3.2.
♢

Appendix B.3. Response time limits in tandem batch-networks

Analogously to the previous section, we first consider the heavy-traffic limit for the tandem network without special arrivals and discarded batches and then that of the geometric batch-queue tandem network.

Appendix B.3.1. Tandem standard batch-queue network

Similarly to the single node case, let the arrival and departure processes of the network be $I^n(t) = \int_0^t \alpha_{A^n(s)}\,\mathrm{d}A^n(s)$ and $O^n(t) = I^n(t) - \sum_{j=1}^J Q_j^n(t)$. Define the virtual waiting time process as before:

$$W^n(t) = \inf\{s \geq 0 \,:\, O^n(s) > I^n(t)\} - t$$

and again by the PASTA property $W^n(\infty)$ has the same distribution as the steady-state response time for the first customer in an arriving batch. Write $\bar{W}^n(t) = (1/\sqrt{n})W^n(nt)$ and note as before that:

$$\bar{W}^n(t) = \sqrt{n}(\inf\{s \geq 0 \,:\, (1/n)O^n(ns) > (1/n)I^n(nt)\} - t)$$

Recall from Section 4.1 that jointly:

$$\sqrt{n}((1/n)I^n(nt) - \bar{\lambda}^n t) \Rightarrow \sqrt{\nu\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}}B_A(t)$$

and

$$\sqrt{n}((1/n)O^n(nt) - \bar{\lambda}^n t) = \sqrt{n}((1/n)I^n(nt) - \bar{\lambda}^n t) - \sum_{j=1}^J \bar{Q}_j^n(t) \Rightarrow \sqrt{\nu\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}}B_A(t) - Z(t)$$

where $Z(t)$ is the sum of the components of the RBM $R[\mathbf{Y}]$. Then using again Lemma 2 in Appendix C, we have the following result.

Proposition 11. *Let $\bar{W}^n$ be the sequence of virtual waiting time processes for the tandem batch queue-network without special arrivals and discarded partial batches. Then under the above assumptions as $n \to \infty$, we have:*

$$\bar{W}^n \Rightarrow \frac{\sqrt{\nu\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}}B_A - \left(\sqrt{\nu\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]}}B_A - Z\right)}{\nu} = (1/\nu)Z$$

Recall from Section 4.1 that the stationary distribution of $R[\mathbf{Y}]$ is product-form exponential if Eq. (9) holds and $\frac{\mathbb{E}[\alpha^2]}{\mathbb{E}[\alpha]} = \frac{\mathbb{E}[\beta_1^2]}{\mathbb{E}[\beta_1]} = \cdots = \frac{\mathbb{E}[\beta_{J-1}^2]}{\mathbb{E}[\beta_{J-1}]}$. Then in this case, we have that the stationary distribution of $(1/\nu)Z$ is a sum of J independent exponential random variables with mean $\frac{\sigma_j^2}{-2(\theta_1+\ldots+\theta_j)}$ for $j = 1, \ldots, J$.

Appendix B.3.2. Tandem geometric batch-queue network

The heavy traffic limit for a tandem pair network of GBQs follows from Proposition 7. We present the result as a corollary to that proposition.

Corollary 1. *In the heavy traffic limit, if f_{mk} has finite first and second moments, the two sojourn times are independent exponential random variables with means $\frac{-\sigma_j^2}{2\sum_{k=1}^j \theta_k}$ for $j = 1, 2$.*

Proof. The proof of Proposition 10 can be used twice, in the reversed process of node 1 and forward process of node 2, these both being geometric batch queues, since in the heavy traffic limit, $\sum_{k=0}^{\infty} p(k)G_R^k(\rho;\theta) \to G_R(\rho;\theta)$, for $R = \tilde{R}_1$ and R_2 respectively. $\diamondsuit$

The corollary extends, by a simple inductive argument, to tandem networks of any finite length provided all nodes are in the heavy traffic regime since then, tasks propagating through the network pass from node to node as normal tasks almost surely. Agreement with the corresponding tandem network of standard batch queues then follows.

Appendix C. Lemmas and proofs

Lemma 1. *Let $A(t)$ be a rate-λ Poisson process defined on some probability space and adapted to some filtration $\mathcal{F}(t)$. Also let $\{\alpha_k\}_{k=1}^{\infty}$ be non-negative, identically distributed and square integrable random variables defined on the same probability space such that for each $t \in \mathbb{R}_+$, the random variables $\mathbf{1}_{\{k \leq A(t)\}}\alpha_k$ for all $k \in \mathbb{Z}_+$ are $\mathcal{F}(t)$-measurable and the random variables $\alpha_{A(t)+1}, \alpha_{A(t)+2}, \ldots$ are independent of $\mathcal{F}(t)$.*

Then the process $M_A(t) = \int_0^t \alpha_{A(s)}\, dA(s) - \mathbb{E}[\alpha]\lambda t$ is a locally square integrable martingale with predictable quadratic variation process $\langle M_A \rangle(t) = \mathbb{E}[\alpha^2]\lambda t$.

Proof. See e.g. Lemma A.1 of [20]. ♢

Lemma 2. *Defined on a common probability space, let the process $X^n(t)$ in $D[0,\infty)$ be unbounded above with $X^n(0) \geq 0$ and let the process $Y^n(t)$ in $D[0,\infty)$ be non-decreasing with $Y^n(0) \geq 0$. Assume that there exist constants $x^n \in \mathbb{R}_{>0}$ and $y^n \in \mathbb{R}_{>0}$ such that $x^n \to x \in \mathbb{R}_{>0}$ and $y^n \to y \in \mathbb{R}_{>0}$ and that there exist continuous processes $U(t)$ with $U(0) = 0$ and $V(t)$ on a common probability space for which it is known that $(c^n(X^n - x^n t), c^n(Y^n - y^n t)) \Rightarrow (U, V)$ in $D_{\mathbb{R}^2}[0,\infty)$ where $c^n \uparrow \infty$. Let $Z^n(t) = \inf\{s \geq 0 : X^n(s) > Y^n(t)\}$. Then $(c^n(X^n - x^n t), c^n(Y^n - y^n t), c^n(Z^n - (y/x)t)) \Rightarrow (U, V, W)$ in $D_{\mathbb{R}^3}[0,\infty)$ where $W(t) = \frac{V(t) - U((y/x)t)}{x}$*

Proof. The lemma is essentially a version of the corollary in [21] adapted to the case of n-dependent centering.

♢