Abstract

We present a complete delay analysis for a tandem network of queues with deterministic service and multiple, interfering sources. The model is of a packet-based data collection system consisting of some finite but arbitrary number of stations connected in tandem by a unidirectional, asynchronous net. Packets, or, more generally, tasks, enter the system at every station, are handed from station to station in storeand-forward fashion, and exit at the downstream end; there are no intermediate departures. The stations are provided with infinite storage, and the lines between them operate concurrently and asynchronously. Intermediate sources are Poisson; the source at the head of the network can be somewhat more general than Poisson. Tasks from each source wait at each station downstream of their point of origin. We calculate the joint steady-state moment-generating functions for these waiting times, provided that the line capacities do not increase in the direction of flow; the solution contains as a special case the steady-state moment-generating function for end-to-end delay for each source.
Introduction

The Model

This paper presents a complete delay analysis for the tandem network of queues shown in Figure (1). The links represent the service facilities and the nodes the storage devices; their number is arbitrary. Tasks (the entities receiving service in the network) may enter the network at any node and are handed from node to node in store-and-forward fashion. The node buffers are infinite, and link service is conservative (no server is idle, save when its queue is empty) and in order of arrival. Other service schedules could be accommodated. Our main assumptions are the following:

1. The flow is unidirectional; tasks exit at the downstream end and there are no intermediate departures.
2. Link service times are deterministic (α_j representing the service time, in seconds, at link l_j) and nondecreasing in the direction of flow: α_1 ≤ α_2 ≤ ⋯.
3. The sources S_j are independent; S_j generates Poisson traffic at mean rate λ_j arrivals per second.
4. The system is stable: ρ_j < 1 for all J, where ρ_j is the probability that l_j is busy.

The traffic on the link l_0 is described in Section (1). The main feature of the model is the presence of multiple interfering sources in a network environment that is not of the Jackson type. Our main results include the joint steady-state mgf (moment-generating function) for the waiting-times at l_J, ..., l_{J+N} of tasks from S_J, and the joint steady-state mgf for the waiting-times at l_J, ..., l_{J+N} of tasks originating upstream of S_J. Explicit formulas are offered as well for the corresponding mean waiting-times. We assume for convenience that the S_J are simple Poisson, pointing where appropriate to the straightforward adjustments needed to accommodate batch arrivals. The analysis here extends and refines the solution of the two-fold network described in [7], and constitutes the most complete solution available for the tandem network with multiple sources and deterministic links.

The model has a variety of applications. The application of specific interest to us is to the analysis of packet delay in a store-and-forward, packet-based, data collection system [10]. The links in this case represent the communication channels, which operate concurrently and asynchronously, and the nodes the storage and processing devices. The transmission rate in each channel is constant, the transmission of a packet beginning only when the packet is fully stored in the buffer, and packet lengths are fixed; the link service times are accordingly non-random.

The assumption that the link service times do not decrease in the direction of flow is crucial to the analysis. It entails that link idle times are exponentially distributed and that link busy periods are nested, in the sense that tasks belonging to the same l_J busy period belong to the same l_{J+1} busy period as well. Where the ordering of the link service times is arbitrary, the problem, aside from partial results in [7], is still unsolved.

Related Literature

The two-fold variant of the present problem, where the network consists of two deterministic links in tandem and the sources at each node are Poisson, is solved in Kaplan [7]. This paper refines and extends the techniques introduced there; it describes in a unifying way the structural features of the problem, shows that the multivariate queueing process (one component for each link) enjoys a key regenerating property, and shows how that regenerating property can be exploited so as to complete the solution of the n-fold network.

Of prior results on tandem queues of the type considered here, the most important, in the context of our own work, have to do with so-called “reduction principles” for networks of deterministic servers. There are essentially two cases. In the first case (Friedman [6], Avi Itzhak [1], Rubin [16]) the links are in tandem, and...
the traffic, which is arbitrary, is offered to the first node only, so that the flow interference which constitutes one of the main features of the present problem is absent; the theorem is that end-to-end waiting-times in the network are unchanged when the network is replaced by its slowest link. In the second case (Ziegler and Schilling [19]), the network consists of a single merge node with multiple input links and a single output link, all with identical, constant service times; the reduction principle here is that neither the busy periods on the output line, nor (when the system is ergodic) the overall mean waiting-times in the network, are changed when the input links are removed, provided that the averaging in the calculation of the waiting-times is over sources as well as sample paths. The reduction applies recursively to unidirectional networks where service times are non-decreasing in the direction of flow. This result has a role in the analysis to follow, but seems not to yield waiting-time statistics for individual sources. Morrison [14] contains closely related material for discrete-time queues.

Other papers on single-source tandem queues include Labetoulle et al. [12], where the link service times are deterministic and the buffers finite, and Boxma [2] and Calo [3], where the link service times are random. There is in addition a small literature on multi-source tandem queues which operate synchronously and in discrete-time. This includes Morrison [15], which computes the generating functions for the queue lengths in the case $N = 2$; Konheim et al. [11], where the interfering flows have different priorities and the waiting-time analysis is shown to reduce to that of a single-server queue; and Meister [13], where the setup is similar to that of [11].

Outline of the paper

Section (1) establishes the basic structural properties of the model. The main results are a transformation which leaves the departure process invariant (the effect is to refine a related result in [19], and to remove the discrete-time restriction in [14]), and identification of a set of points in space and time at which the multivariate stochastic process representing the queueing processes at each link can be said, roughly, to regenerate. The results are extended to input processes more general than compound Poisson.

Section (2) derives the steady state mgf for the waiting-times incurred at $l_J$ by tasks arriving exogenously from $S_J$, and also by tasks arriving endogenously from $l_{J-1}$; the trick, borrowed from [7], is to discern in the waiting-time process at each link an embedded M/G/1 queue. Section (2) presents as well the mgf for the number of tasks in an $l_J$ busy period.

Section (3) presents our main result: the joint steady-state mgf for $S_J$ (or $l_{J-1}$) waiting-time at $l_J, \ldots, l_{J+N}$. The calculation is based on the regeneration property described in Section (1), and on the steady-state joint mgf for waiting-time and rank (the location of a task in its busy period) in a batch-arrival M/G/1 queue; the latter is computed in the Appendix.

Section (4) offers simple formulas for the mean waiting-times, and comparison with a widely-used approximation. Section (5) concludes.

Notation

Where $Y$, $Z$ are random variables, and $\{X(i)\}$ is a sequence of random variables that are identically distributed, we write $Y \approx Z$ to indicate that $Y$, $Z$ have the same distribution, and $X$ (a random variable) to represent the common distribution of the $X(i)$. The notation $X * M$, where $X, M$ are independent random variables and $M$ is integer-valued, stands for the random sum of $M$ iid random variables each distributed like $X$; where $M = 0$, $X * M = 0$. The operation $*$ is associative, but not commutative; the iterated variant $U(k) * [U(k-1) * \cdots * U(1)]$ (note the order of the indices), where all variables except perhaps the first are integer-valued, is abbreviated $\prod_{i=1}^{k} * U(i)$. We adopt the convention that $\sum_{i} \cdot$, $\prod_{i} \cdot$ are zero, unity, and unity respectively when $i > j$. The brackets $[x]$, where $x$ is real, stand for the integer part of $x$. 

3
1. Basic Properties

We focus here on the structural properties of the model. Lemma 1 enumerates certain combinatorial features that are implied by the assumed ordering of the link service times. Theorem (1), unifying and elaborating results due to Ziegler and Schilling [19] and Morrison [14], presents a transformation which leaves invariant the departure process from $l_J$ and simplifies the computation of its statistics. Lemma (1) contains the key to the joint queue analysis presented in Section (3); specialized to Poisson arrivals, it implies that the queueing processes at $l_1, l_2, \cdots, l_J$ regenerate, in a sense to be made precise, at $l_J$ idle instants. Refer to figure (2).

Assume for the moment that the source statistics are arbitrary. By virtue of the conservative, deterministic character of the links, a task finding $l_{J-1}$ idle is released to $l_J$ $\alpha_{J-1}$ seconds later, while each task finding $l_{J-1}$ busy is released to $l_J$ $\alpha_{J-1}$ seconds after its predecessor.

Let $t_{J-1}(k, n)$ denote the instant of release from $l_{J-1}$ of the $k$-th task in the n-th $l_{J-1}$ busy period, and $\eta_{J-1}(n)$ the number of tasks served in the n-th $l_{J-1}$ busy period. Then $t_{J-1}(k, n) = t_{J-1}(1, n) + (k - 1)\alpha_{J-1}$, $t_{J-1}(1, n + 1) > t_{J-1}(\eta_{J-1}(n), n) + \alpha_{J-1}$, and the n-th $l_{J-1}$ busy period starts at $t_{J-1}(1, n) - \alpha_{J-1}$ and ends at $t_{J-1}(\eta_{J-1}(n), n)$.

**DEFINITION:** The interval $(t_{J-1}(1, n), t_{J-1}(\eta_{J-1}(n), n) + \alpha_{J-1})$ will be called the n-th active period of $l_{J-1}$; it is a replica of the n-th busy period of $l_{J-1}$, delayed by $\alpha_{J-1}$ seconds. The intervals between active periods will be said to be inactive. The n-th inactive period is a replica of the n-th idle period, delayed by $\alpha_{J-1}$ seconds.

The assumed ordering of the link service times, $\alpha_{J-1} \leq \alpha_J$, implies the following:

**LEMMA (1):**
(i) $l_J$ is busy during $l_{J-1}$ active periods.
(ii) If $l_J$ is idle at time $t$, then $l_{J-1}$ is idle at $t - \alpha_{J-1}$.
(iii) For $I = 1, 2, \cdots, J - 1$, tasks from $S_{J-1}$ generated after $t - \alpha_{J-1} - \cdots - \alpha_{J-1}$ remain upstream of $l_J$ at $t$; if $l_J$ is idle at $t$, then tasks from $S_{J-1}$ generated prior to $t - \alpha_{J-1} - \cdots - \alpha_{J-1}$ are downstream of $l_J$ at $t$.

Statements (i) and (ii) express a key property of the model, that busy periods do not break in the direction of flow. The meaning of (iii) can be visualised with the help of Figure (3). Curve $C_J(t)$ reflects the transport delay, neglecting waiting-time, in the network. Tasks which enter $l_I$ $(I < J)$ in that region of the $l_I$ time-axis to the right of $C_J(t)$ are upstream of $l_J$ at $t$. From this it follows that the $l_J$ load at $t$ depends only on tasks which enter the network in that region of the time-axis to the left of $C_J(t)$. These are downstream of $l_J$ at $t$ whenever $t$ is an $l_J$ idle instant.

Fix $J$. From Lemma (1), the busy and idle periods of $l_J$ are unchanged if $l_1, \cdots, l_{J-1}$ are replaced by pure delays $\alpha_1, \cdots, \alpha_{J-1}$ respectively (a pure delay of $\alpha$ seconds amounts to a $\star/D/\infty$ queue with service times $\alpha$). Hence:

**THEOREM (1):** The departure instants of tasks from $l_J, l_{J+1}, \cdots$, where $J$ is fixed, are unchanged if $l_1, \cdots, l_{J-1}$ are replaced by pure delays $\alpha_1, \cdots, \alpha_{J-1}$ respectively.

We emphasize that the invariance in the departure process is a sample path property and applies to arbitrary input processes. Note that the order of release of tasks from $l_J$ is scrambled by the transformation of Theorem (1), the scrambling being limited to within an active period of $l_{J-1}$. Note also that if the arrival processes are stationary and independent, then for the purpose of computing the statistics of the departure process from $l_J$, the delays can be removed. Theorem (1) then implies that the network reduces to a single
link with service time $\alpha_J$. This reduction principle was first stated in [19].

Let $\eta_J$ (a random variable) stand for the number of tasks served in an $l_J$ busy period, and $\gamma_J(z) = E z^{\eta_J}$ its mgf. The Lemma and the Theorem acquire particular significance in the case of Poisson arrivals.

COROLLARY (Poisson arrivals):

(i) If $t$ is an $l_J$ idle instant, then the queueing process at $l_J$ prior to $t$ is independent of the arrival process at $l_J$ at or after $t$ and of the arrival process at $l_I$ at or after $t - \alpha_{J-1} - \cdots - \alpha_I$, $I = 1, \ldots, J - 1$.

(ii) The departure process from $l_J$ is statistically identical to that from an M/D/1 queue with rate $\sum_{i=1}^J \lambda_i$ and service time $\alpha_J$; in particular, the inactive periods at $l_J$ are exponentially distributed with mean $\Lambda_J^{-1} = (\sum_{i=1}^J \lambda_i)^{-1}$, and $\eta_J$ is Borel ([9], equation 5.158), with mgf

$$\gamma_J(z) = z \exp\{-\rho_J(1 - \gamma_J(z))\} = \sum_{k=1}^{\infty} \frac{(k\rho_J)^{k-1}}{k!} \exp\{-k\rho_J \} z^k,$$

where $\rho_J = \alpha_J \sum_{i=1}^J \lambda_i$ is the traffic intensity at $l_J$.

Statement (i) follows from (iii) of Lemma (1) and from the fact that the points $t, t - \alpha_{J-1}, \cdots, t - \alpha_{J-1} - \cdots - \alpha_1$ are regeneration points for the queueing processes in $l_J, \cdots, l_1$ respectively; (ii) follows from Theorem (1). Note that with simple modifications to (1), the results carry over to the case where the arrivals are compound Poisson, the departure process from $l_J$ being equivalent then to that from a bulk-arrival M/D/1 queue. The results in this section, both for arbitrary and for Poisson arrivals, carry over as well to an unidirectional tree network where multiple deterministic links merge at each node and where link service times do not decrease in the direction of flow.

Extensions

The departure process from an M/D/1 queue, consisting of alternating active and inactive periods, exemplifies a class of point processes which we refer to as M-G*D. M signals the memorylessness of the inactive periods, G the fact that the number of points in the iid active periods has a general distribution, and D the fact that the intervals between points within an active period are deterministic. An M-G*D process is specified by the mean length $\Lambda_J^{-1}$ of the inactive periods, the length $\alpha$ of the interval separating points within an active period, and the distribution (represented by an integer-valued random variable $\eta$ with mgf $\gamma(\cdot)$) of the number of points in an active period. The steady state probability that the process is in its active state is $\rho = \Lambda \alpha E\eta/(1 + \Lambda \alpha E\eta)$. For the departure process from an M/D/1 queue, $\gamma(z)$ is Borel with $E\eta = 1/(1 - \rho)$ and $\rho = \Lambda \alpha$. A compound Poisson process corresponds to the particular case $\alpha = 0$.

The M-G*D process has a simple conservation property: a deterministic server with multiple M-G*D inputs has M-G*D departures, provided that each input, when active, produces tasks at intervals no greater than the service time. Part (i) of the Corollary applies when the $S_J$ are sums of M-G*D processes, but part (ii) does not; $\gamma_J(z)$ is not, in general, Borel, except (by Theorem (1)) where the incoming processes at $l_J$, $I = 1, 2, \cdots, J$, are Poisson or Borel M-G*D.

The M-G*D process is a model for a packetized-message source where the delay between successive packet departures within a message is non-trivial. It has been applied as well to the modelling of road traffic (Cowan [5]), and with gradual, rather than instantaneous, inputs, to the analysis of a storage system (Cohen ([4]), Kaspi and Rubinovertich ([8]). The foregoing are concerned generally with a single deterministic link and multiple M-G*D inputs, each of which, when active, presents work to the server at a rate equal to the service rate.

The joint waiting time analysis which follows is carried out for Poisson arrivals at $l_1, l_2, \cdots$ and an additional M-G*D input at $l_1$. Additional related results can be found in Shalmon [17].
2. The Analysis at a Single Link

Assume that $S_1, \ldots, S_J$ are simple Poisson, and that $l_0$ departures are M-G*D with parameters $\gamma_0()$, $\Lambda_0, \alpha_0$, where $\alpha_0 \leq \alpha_1$. Under these conditions, the $l_J$ departure process is also M-G*D, with parameters $\gamma_J()$ (to be computed), $\Lambda_J$, given by

$$\Lambda_J = \Lambda_0 + \sum_{i=1}^{J} \lambda_i,$$

and $\alpha_J$. The probability $\rho_J$ that $l_J$ is busy (equivalently, that the $l_J$ departure process is in its active state) is

$$\rho_J = \frac{\Lambda_J \alpha_J \eta_J}{1 + \Lambda_J \alpha_J \eta_J}.$$

We compute here the steady-state mgf for the waiting-times incurred at $l_J$ by tasks from $S_J$, for the waiting-times incurred at $l_J$ by tasks from $l_{J-1}$, and for the lengths of $l_J$ busy periods. The busy period calculations provide an alternate proof of part (ii) of the Corollary in Section (1), in that the busy period distribution for $l_J$ specializes to the Borel distribution whenever the active periods in $l_0$ are assumed to be Borel. For the waiting-time analysis in this section, we shall need certain results described in [7]; these, adapted to the present setting, are summarized in Theorems (2,a) and (2,b) and in the accompanying discussion. We begin with the waiting-time analysis, assuming throughout that all queues are stable. Refer to figure (2).

The waiting-times

Fix $J$. Let $\Delta r_J(n), \Delta V_J(n)$ denote, respectively, the total number of tasks (both exogenous and endogenous) submitted to $l_J$ during the $n$-th $l_{J-1}$ active period, and the change in the unfinished work at $l_J$ across the same period. By Lemma (1), $\Delta V_J(n)$ is nonnegative and equal to $\Delta r_J(n) \alpha_J - \eta_{J-1}(n) \alpha_{J-1}$. Let $\nu_J$ be a random variable representing the number of $S_J$ tasks generated during an interval $\alpha_{J-1}$ seconds long; $\nu_J$ is Poisson, with mean $\lambda_J \alpha_{J-1}$ and mgf $P_J(z) = E^z^{\nu_J}$ given by

$$P_J(z) = \exp[-\lambda_J \alpha_{J-1}(1 - z)]. \tag{2}$$

The $\Delta r_J(n)$ $(n = 1, 2, \cdots)$ are iid, as are the $\Delta V_J(n)$; they are related to the $\eta_{J-1}(n)$ by

$$\Delta r_J \approx (1 + \nu_J) \ast \eta_{J-1},$$

$$\Delta V_J \approx (\alpha_J - \alpha_{J-1} + \alpha_J \nu_J) \ast \eta_{J-1}. \tag{3}$$

The memorylessness of the inactive periods and the nonnegativity of the change in virtual waiting-time across active periods suggest a compression of the time axis that reveals an embedded M/G/1 queue:

**THEOREM (2,a):** Compress the time axis so that the active periods of $l_{J-1}$ reduce to a point, and insert at the $n$-th compression point a virtual message requiring $\Delta V_J(n)$ seconds of service. Then

(i) The inactive periods of $l_{J-1}$, and the virtual waiting-time at $l_J$ during $l_{J-1}$ inactive periods, are unchanged.

(ii) The compression points are Poisson, with rate $\Lambda_{J-1}$.

(iii) The source $S_J$, consisting of tasks from $S_J$ generated during $l_{J-1}$ inactive periods, is Poisson, with rate $\lambda_J$.

(iv) $l_J$, representing the link $l_J$ in compressed time, is an M/G/1 queue with arrival rate $\Lambda_J = \Lambda_{J-1} + \lambda_J$ and service times distributed as $\theta_J$, where the r.v. $\theta_J$ is $\alpha_J$ with probability $\lambda_J / \Lambda_J$, and $\Delta V_J$ otherwise.

(v) The steady-state distribution of waiting-time in $l_J$, the steady-state distribution of waiting-time in $l_J$ for $S_J$ tasks generated when $l_{J-1}$ is inactive, and the steady-state distribution of waiting-time in $l_J$ for tasks which initiate $l_{J-1}$ busy periods, are all identical.
We use the r.v. $\tilde{W}_j$ to denote the waiting-time distributions described in (v). Recall in what follows that $P_j(\cdot)$ is defined in (2) and that $\gamma_{j-1}(\cdot)$ is the mgf for the number of tasks served in an $l_{j-1}$ busy period; $\gamma_{j-1}(\cdot)$ is given by (1) in the special case that the M-G*D departures from $l_0$ are in fact just Poisson, and otherwise remains to be calculated. The moment-generating functions for $\tilde{\theta}_j$, $\tilde{W}_j$ follow from (3), and from part (iv) of the Theorem augmented by known results for M/G/1 ([9], equation 5.105):

$$E \exp[-s\tilde{\theta}_j] = \frac{\Lambda_{j-1}}{\Lambda_j} \gamma_{j-1}[\exp[-s(\alpha_j - \alpha_{j-1})]P_j(\exp[-s\alpha_j])] + \frac{\lambda_j}{\Lambda_j} \exp(-s\alpha_j),$$

and

$$E \exp[-s\tilde{W}_j] = (1 - \Lambda_j E \tilde{\theta}_j) \frac{s}{s - \Lambda_j (1 - E \exp[-s\tilde{\theta}_j])}.$$

The steady-state $l_J$ waiting-time distributions for tasks excluded in (v) — that is, for exogenous tasks generated during $l_{j-1}$ active periods, and for endogenous tasks which are not first in their $l_{j-1}$ busy periods — can be extrapolated from (5). The key point is that $l_j$ cannot empty during an $l_{j-1}$ active period, so that the unfinished work in $l_j$ some $t$ seconds (say) into an $l_{j-1}$ active period is a simple function of $t$ and of the unfinished work at the start of the active period.

In applying this observation to the calculation of the waiting-time statistics for a packet $P$ arriving at $l_j$ from $l_{j-1}$, it is just $\alpha_{j-1}(r_{j-1} - 1)$, where $r_{j-1}$ is the rank of $P$ in its $l_{j-1}$ busy period; where instead $P$ arrives exogenously from $S_J$ while $l_{j-1}$ is active, $t$ is the corresponding conditional backward recurrence time from the instant of arrival to the start of the active period. In both cases, the unfinished work in $l_j$ at the start of the $l_{j-1}$ active period is distributed like $\tilde{W}_j$ and independent of $t$. With $W$ standing temporarily for $P$’s waiting-time at $l_j$, and $\nu_j(t)$ for the number of $S_J$ tasks generated during the $t$ seconds of the $l_{j-1}$ active period preceding the arrival of $P$ at $l_j$, it follows that

$$W \approx \tilde{W}_j + \alpha_j([t/\alpha_{j-1}] + \nu_j(t)) - t.$$

The distribution of $t$ in (6) is calculated using standard renewal theory [7]; because $S_J$ is Poisson, the calculation for exogenous $P$ is the same as for a virtual arrival at a fixed instant of time [18]. Let $W_j(S_J)$, $W_j(l_{j-1})$ denote the steady-state waiting-times at $l_j$ for exogenous and endogenous $P$ respectively. From (5) and (6), and from the observation that $S_J$ tasks arrive during $l_{j-1}$ active periods with probability $\rho_{j-1}$, we can express the moment-generating functions for $W_j$, $W_j(S_J)$, and $W_j(l_{j-1})$ in terms of the mgf $P_j(\cdot)$, the traffic intensities

$$\rho_j = \frac{\Lambda_j \alpha_j E \eta_j}{(1 + \Lambda_j \alpha_j E \eta_j)},$$

and the mgf $\gamma_{j-1}(\cdot)$, which is to be calculated in the next subsection.

**THEOREM (2,b):**

(i)

$$E \exp[-s\tilde{W}_j] = \frac{(1 - \rho_j)(1 - \rho_{j-1})^{-1}s}{s - \Lambda_j + \lambda_j \exp(-s\alpha_j) + \Lambda_{j-1} \gamma_{j-1}[\exp[-s(\alpha_j - \alpha_{j-1})]P_j(\exp(-s\alpha_j))]}$$

(ii)

$$\frac{E \exp[-sW_j(l_{j-1})]}{E \exp[-s\tilde{W}_j]} = \frac{1 - \gamma_{j-1}[\exp[-s(\alpha_j - \alpha_{j-1})]P_j(\exp(-s\alpha_j))]}{E \eta_{j-1}} \frac{1}{1 - \exp[-s(\alpha_j - \alpha_{j-1})]P_j(\exp(-s\alpha_j))}$$

(iii)

$$E \exp[-sW_j(S_J)] = (1 - \rho_{j-1})E \exp[-s\tilde{W}_j] + \frac{\rho_{j-1}}{\alpha_{j-1}}.$$
\[
\frac{\exp[-s(\alpha_j - \alpha_{j-1})]P_{j-1}(\exp(-s\alpha_j)) - \exp(-s\alpha_j)E\exp[-sW_{j}(l_{j-1})]}{s - \lambda_j[1 - \exp(-s\alpha_j)]}
\]

The busy periods

We calculate here the mgf \( \gamma_j(\cdot) \) for the number of tasks in an \( l_j \) busy period. The idea is that the time compression described in Theorem (2,a) alters the lengths (in seconds) of the busy periods, but not their size, where by the size of a busy period we mean the number of tasks completed; this is so provided that the “virtual” message representing the \( n \)-th compression point in \( \bar{l}_j \) is construed as comprising \( \Delta r_j(n) \) tasks. The service request of the bulk remains \( \Delta V_j(n) \). So the problem here reduces to the analysis of the number of tasks served in the busy periods of \( \bar{l}_j \), a bulk arrival \( M/G/1 \) queue.

Write \( \bar{\mu}_j \) for the number of tasks in a bulk arrival at \( \bar{l}_j \); \( \bar{\mu}_j \) is unity with probability \( \lambda_j/\Lambda_j \), and \( \Delta r_j \) otherwise. Recall from part (iv) of Theorem (2,a) the definition of the corresponding service time \( \bar{\theta}_j \). The joint mgf

\[ Q_j(z, s) = E[z^{\bar{\mu}_j} \exp(-s\bar{\theta}_j)] \]

characterizes the bulk statistics; it is easy to show that

\[ Q_j(z, s) = \frac{\Lambda_{j-1}}{\Lambda_j} \gamma_{j-1}[\exp(-s(\alpha_j - \alpha_{j-1}))P_j[z\exp(-s\alpha_j)]] + \frac{\lambda_j}{\Lambda_j} z \exp(-s\alpha_j). \tag{8} \]

Let \( \gamma_j(\cdot|k, \theta) \) be the conditional mgf for the size of an \( \bar{l}_j \) busy period initiated by a bulk arrival consisting of \( k \) tasks and requiring a total of \( \theta \) seconds of service. Following the line of argument used by Takacs in his analysis of the busy periods in \( M/G/1 \) ([10], pages 111-112), we have

\[
\gamma_j(z|k, \theta) = z^k \sum_{i=0}^{\infty} \frac{[\Lambda_j \theta]^i}{i!} \exp(-\Lambda_j \theta) [\gamma_j(z)]^i
\]

\[ = z^k \exp[-\Lambda_j \theta(1 - \gamma_j(z))]. \tag{9} \]

The following is obtained by averaging (9) with respect to the joint distribution of \((\bar{\mu}_j, \bar{\theta}_j)\), given by (8):

**THEOREM (3):** The mgf for the size of an \( l_j \) busy period is given by

\[
\gamma_j(z) = Q_j[z, \Lambda_j(1 - \gamma_j(z))].
\]

Theorem (3) expresses \( \gamma_j(z) \) in terms of \( \gamma_{j-1}(z) \). It applies generally to a single-server queue with combined M-G*D and Poisson traffic. The simplest way to express \( \gamma_j(\cdot) \) in terms of \( \gamma_0(\cdot) \) is to use the fact, available from Theorem (1) or from the related result in [19], that in calculating \( \gamma_j(\cdot) \) one can assume that the sources \( l_0, S_1, \cdots, S_j \) are connected directly to \( l_j \). Let \( P_{j,0}(\cdot) \) be the mgf for the number of tasks generated by \( S_1, \cdots, S_j \) in an interval of length \( \alpha_0 \); let \( Q_{j,0}(z, s) \) denote the RHS of (9) when the quantities \( \Lambda_{j-1}, \gamma_{j-1}, \alpha_{j-1}, \lambda_j \) are replaced by \( \Lambda_0, \gamma_0, \alpha_0, P_{j,0}, \lambda_1 + \cdots + \lambda_j \) respectively.

**COROLLARY:** \( \gamma_j(z) = Q_{j,0}[z, \Lambda_j(1 - \gamma_j(z))]. \)

In the special case that \( \gamma_0(\cdot) \) is Borel, it can be seen through computation that \( \gamma_j(\cdot) \) is also Borel, as expected from the Corollary in Section (1). The formulas are easily modified to accommodate compound Poisson arrivals.

In the next section, and in the Appendix, we calculate additional attributes of the composition of \( l_j \) busy periods required for the joint queue analysis.
3. The Joint Queue Analysis

The main result in this section is the joint steady-state mgf for $S_J$ waiting-times at $l_1, \ldots, l_{J+N}$; we compute as well the same joint statistic for tasks entering $l_j$ endogenously. The assumptions, as before, are that $S_1, S_2, \ldots, S_N$ are Poisson, and that departures from $l_0$ are M-G*D with parameters $\gamma_0(\cdot), \lambda_0, \alpha_0$.

Without loss of generality, we set $J = 1$. We focus on a particular task $P$ which enters the network, either from $l_0$ or from $S_1$, at $l_1$; P’s waiting-time at $l_k$, and its rank in its $l_k$ busy period, are denoted $W_K$, $r_K$ respectively. The notation $P_K$ designates the task which initiates the $l_{K-1}$ busy period containing $P$; $P_K$ waiting-time at $l_k$ and the rank of $P_K$ in its $l_k$ busy period are represented by $W_K$ and $r_K$ respectively.

In the previous section, we obtained the steady-state statistics of $W_K$ by identifying in $l_K$ the embedded M/G/1 queue $\bar{l}_K$. The following theorem, characterizing the bivariate random process $(\tilde{W}_K, \tilde{r}_K)$, $K = 1, 2, \ldots$, generalizes this result; it is the key to the joint queue analysis. The queue $\bar{l}_K$ plays the same role as before. Recall the batch structure of the virtual arrivals, described in the discussion leading to Theorem (3); the joint mgf for batch size and service time is $Q_K(\cdot, \cdot)$, defined in (9). The generating-function $P_K(\cdot)$ is defined in (2).

**THEOREM (4):** The pairs $(\tilde{W}_K, \tilde{r}_K)$, $K = 1, 2, \ldots$, are independent; the joint mgf $\Phi_K(z, s) = E[z^{\alpha K-1} \exp(-sW_K)]$ is given by

$$
\Phi_K(z, s) = \frac{[(1-\rho_K)/(1-\rho_{K-1})][s - \Lambda_K(1-\gamma_K(z))]}{s - \Lambda_K + \lambda_K z \exp(-s\alpha_K) + \Lambda_{K-1} \gamma_{K-1} \exp[-s(\alpha_K - \alpha_{K-1})]P_K(\exp(-s\alpha_K)).
$$

**Proof:** Fix $K$, and let $J$ run through $\{1, \ldots, K-1\}$. Define $t$ to be the instant at which $\tilde{P}_K$ enters the queue at $l_k$, and $t_J = t - \alpha_{K-1} - \cdots - \alpha_J$. The pair $(\tilde{W}_K, \tilde{r}_K)$ is a function of the arrival process at $l_{K-1}$ prior to $l_{K-1}$, and of $S_K$ arrivals prior to $t$. The following are consequences of Lemma (1): (i) $l_J$ is idle at $t_J$; (ii) all tasks in $l_J$ prior to $t_J$ are downstream of $l_{K-1}$ at $t_{K-1}$; (iii) $\tilde{P}_K, \ldots, \tilde{P}_1, P$ arrive at $l_K$ in the order listed. It ensues that $(\tilde{W}_J, \tilde{r}_J)$ is a function of the arrival process at $l_J$ in $[t_J, \infty)$. By the corollary to Lemma (1), the queuing process at $l_K$ in $(-\infty, t)$, and the arrival processes at $l_J$ in $[t_J, \infty)$ ($J < K$), are independent. The asserted independence of the pairs $(\tilde{W}_K, \tilde{r}_K)$ follows.

The expression for $\Phi_K(z, s)$ is derived by observing that $\tilde{W}_K$, $\tilde{r}_K$ have the same joint distribution as the variables $W$, $r$ representing waiting-time and rank in the batch-arrival queue M/G/1 queue $\bar{l}_K$. The mgf for $(W, r)$ in an arbitrary batch-arrival M/G/1 queue is calculated in the Appendix.

The rest of this section is devoted to showing how one can extrapolate from Theorem (4) to the joint mgf for $W_1, W_2, \cdots$. The extrapolation is based on (6), which in the present setting can be written as

$$
W_K \approx \tilde{W}_K + (\alpha_K - \alpha_{K-1} + \alpha_K \nu_K) \ast (r_K - 1) - 1,
$$

(10)

where the random variable $\nu_K$, standing for the number of $S_K$ tasks generated in an $\alpha_{K-1}$-second interval following the departure from $l_{K-1}$ of $\tilde{P}_K$, is independent of $W_K$ and of $r_{K-1}$. There is a similar recursion for $r_K$:

$$
r_K - 1 \approx (\tilde{r}_K - 1) + (1 + \nu_K) \ast (r_{K-1} - 1).
$$

(11)

Using (10) and (11), we derive an expression for the quantity $\sum_{1}^{N} s_K W_K$ (the $s_K$ are complex numbers with positive real parts) in terms of the pairs $(\tilde{W}_K, \tilde{r}_K)$, $K = 1, \ldots, N$; the negative of this expression, exponentiated and averaged with the help of Theorem (4), is the generating-function we seek. For convenience in what follows, we allow the index $K$ in references to $\tilde{P}_K, \tilde{r}_K$ (defined above for $K \geq 1$) to take the value zero as well, with $\tilde{P}_0$ meaning $P$ and $\tilde{r}_0 = r_0$. 


Assume first that \( P \) originates at \( l_0 \). The first step is to write \( r_K \) \((K \geq 1)\) in terms of \( r_0, \tilde{r}_1, \ldots, \tilde{r}_K \), starting from (11). To see what the answer should be, note first, applying Lemma (1), that \( \tilde{P}_{K+1}, \ldots, \tilde{P}_1, P \) all belong to the same \( l_K \) busy period, receiving service at \( l_K \) in the order listed. Let \( a_{IJ}, I = 0, \ldots, K \) and \( J \geq I \), be the number of tasks (including \( \tilde{P}_I \) but not \( \tilde{P}_1 \) served at \( l_J \) between \( \tilde{P}_{I+1} \) and \( \tilde{P}_I \) (see Figure (4)). The \( a_{IK} \) \((K \text{ fixed})\) are independent of each other, and \( r_K - 1 \) is their sum:

\[
r_K - 1 = a_{0K} + \cdots + a_{KK}.
\] (12)

The \( a_{IJ} \) \((I \text{ fixed})\) satisfy a simple recursion:

\[
a_{IL} \equiv \tilde{r}_I - 1,
\]

\[
a_{IJ} \approx (1 + \nu_{IJ}) * a_{IJ-I}, \quad J = I + 1, \ldots, K,
\]

where the \( \nu_{IJ} \) \((J \text{ fixed})\) are iid random variables distributed like \( \nu_J \) and independent of \( \tilde{r}_I \); it follows that

\[
a_{IK} \approx \prod_{J=I+1}^{K} [(1 + \nu_{IJ})] * (\tilde{r}_I - 1).
\] (13)

This, applied to (12), yields the desired expression for \( r_K \).

The next step is to write \( \sum s_K W_K \) in terms of \( \sum s_K \tilde{W}_K \) and the \( \tilde{r}_K \). Substitute (12) into (10), multiply both sides by \( s_K \), and sum over \( K \):

\[
\sum_{K=1}^{N} s_K W_K \approx \sum_{K=1}^{N} s_K \tilde{W}_K + \sum_{K=1}^{N} \sum_{I=0}^{K-1} s_K [(\alpha_K - \alpha_{K-1} + \alpha_K \nu_{IK}) * a_{IK-I}].
\]

By interchanging the order of summation on the RHS, applying (13), and letting \( \Gamma_K^N = \Gamma_K^N(s_{K+1}, \ldots, s_N) \) stand for the random variable

\[
\sum_{L=K+1}^{N} s_L (\alpha_L - \alpha_{L-1} + \alpha_L \nu_{KL}) * \prod_{J=K+1}^{L-1} [(1 + \nu_{KJ})],
\]

we get

\[
\sum_{K=1}^{N} s_K W_K \approx \sum_{K=1}^{N} [s_K \tilde{W}_K + \Gamma_K^N * (\tilde{r}_K - 1)] + [\Gamma_0^N * (r_0 - 1)].
\] (14)

The last step is based on the following lemma:

**LEMMA (2):** The variables \( \nu_{IJ} \) \((1 \leq J \leq N \text{ and } I < J)\) are independent of \((\tilde{W}_K, \tilde{r}_K)\), \(1 \leq K \leq N\).

**Proof:** Fix \( J \). The \( \nu_{IJ} \), being attributes of \( S_J \), are independent of the queueing processes upstream of \( l_J \), hence in particular of \((\tilde{W}_K, \tilde{r}_K)\) for \( K < J \). The assertion is extended to \( K \geq J \) by observing that the tasks counted in \( \nu_{IJ} \) follow \( P_K \) at \( l_K \) (see Figure (3)), hence do not affect \( \tilde{P}_K \) queue statistics there.

The lemma implies in particular that the variables \( \Gamma_K^N \) in (14) are independent of the \((\tilde{W}_K, \tilde{r}_K)\). Calculating \( E \exp(- \sum s_K W_K) \) from (14) is thus a straightforward application of Theorem (4). The result is stated below in Theorem (5). Before proceeding to Theorem (5), we indicate how the foregoing argument should be amended so as to apply when \( P \) originates at \( S_1 \), rather than at \( l_0 \).

Suppose, then, that \( P \) belongs to \( S_1 \). There are two cases, according as \( P \) enters the system during an \( l_0 \) active period or not. The two cases occur with probabilities \( \rho_0, 1 - \rho_0 \) respectively, where \( \rho_0 \) is defined in
(7). Where \( P \) arrives when \( l_0 \) is in-active, the only modification needed in the derivations above is removal of all terms containing \( r_0 \) (the desired effect is achieved by setting \( r_0 = 1 \) throughout). The complementary case, where \( P \) arrives, say, \( t \) seconds into an \( l_0 \) active period \((t > 0)\), can be reduced to the one considered above by introducing the variables

\[
\sigma = [t/a_0], \quad \tau = t \mod a_0.
\]

An easy renewal-theoretic argument verifies that \( \sigma, \tau \) are independent of each other, that \( \tau \) is uniformly distributed on \([0, a_0)\), and that \( \Pr\{\sigma = i\} = \Pr\{\eta_0 \geq i + 1\}/E\eta_0 \) \((i \geq 0)\), from which it follows that

\[
Ez^\sigma = \frac{1}{E\eta_0} \frac{1 - \gamma_0(z)}{1 - z}.
\]

The role assigned previously to \( r_0 - 1 \) is here played by \( \sigma \). In particular, equations (10) and (11), while remaining valid for \( K > 1 \), change when \( K = 1 \) to

\[
W_1 \approx \tilde{W}_1 + (\alpha_1 - \alpha_0 + \alpha_1 \nu_1) \sigma + \alpha_1 (1 + \nu_1(\tau)),
\]

\[
r_1 - 1 \approx (\tilde{r}_1 - 1) + (1 + \nu_1) \sigma + (1 + \nu_2(\tau)),
\]

where \( \nu_1(\tau) \) is the number of \( S_1 \) tasks generated during the \( \tau \) seconds preceding the creation of \( P \). The effect in (14), beyond insertion of \( \sigma \) for \( r_0 - 1 \), is to append to the RHS an additional term

\[
s_1(\alpha_1 - \alpha_0 + \alpha_1 \nu_1(\tau)) + \Gamma_N^1 \approx (1 + \nu_1(\tau))
\]

. The independence arguments go through as before.

The results in Theorem (5) are expressed in terms of the generating-functions \( E \exp(-\Gamma_K^N) \). These can be computed recursively as follows. Starting from the definition of \( \Gamma_K^N \), note that

\[
\Gamma_K^N = 0,
\]

\[
\Gamma_K^N \approx s_{K+1}(\alpha_{K+1} - \alpha_K + \alpha_K \nu_{K,K+1})
\]

\[
+ \left[ \sum_{L=K+2}^{N} s_L(\alpha_L - \alpha_{L-1} + \alpha_L \nu_{KL}) \right] \prod_{J=K+2}^{L-1} (1 + \nu_{KJ}) \left(1 + \nu_{K,K+1}\right).
\]

The term in square parentheses on the RHS being distributed as \( \Gamma_{K+1}^N \) and independent of \( \nu_{K,K+1} \), it ensues that

\[
E \exp(-\Gamma_K^N) = 1,
\]

\[
E \exp(-\Gamma_K^N) = \exp[-s_{K+1}(\alpha_{K+1} - \alpha_K)]E[\exp(-\Gamma_{K+1}^N)]E[\exp(-s_{K+1} \alpha_{K+1})]E \exp(-\Gamma_{K+1}^N)].
\]

This is the desired recursion.

In the statement of Theorem (5), \( W_K(l_0) \) stands for the waiting-time at \( l_K \) of a task from \( l_0 \), and \( W_K(S_1) \) for the waiting-time at \( l_K \) of a task from \( S_1 \). The quantities \( \gamma_K(\cdot), \rho_K \), and \( \Phi_K(\cdot, \cdot) \) are defined in the corollary to Theorem (3), in equation (7), and in Theorem (4) respectively. The function

\[
F_N(s_1, \ldots, s_N) = \prod_{K=1}^{N} \Phi_K(E \exp[-\Gamma_K^N], s_K)
\]

is what one gets when \( E \exp(-\sum s_K W_K) \) is calculated from (14) assuming \( r_0 = 1 \).

**THEOREM (5):** (i)

\[
F_N(s_1, \ldots, s_N) = \frac{1 - \rho_N}{1 - \rho_0}
\]
\[
\prod_{K=1}^{N} \frac{s_K - \Lambda_K(1 - \gamma_K(E \exp[-\Gamma_K^{N}])))}{s_K - \Lambda_K + \lambda_K \exp(-s_K \alpha_K)E \exp[-\Gamma_K^{N}] + \Lambda_K - 1 \gamma_K(E \exp[-\Gamma_K^{N-1}])}
\]

(ii)
\[
E \exp[-\sum_{K=1}^{N} s_K W_K(l_0)] / F_N(s_1, \ldots, s_N) = \frac{1 - \gamma_0(E \exp[-\Gamma_0^{N}]])}{E \eta_0 1 - E \exp[-\Gamma_0^{N}]}
\]

(iii)
\[
E \exp[-\sum_{K=1}^{N} s_K W_K(S_1)] = (1 - \rho_0)F_N(s_1, \ldots, s_N)
\]
\[
+ \frac{\rho_0}{\alpha_0} E \exp[-\Gamma_1^{N}] E \exp[-\Gamma_0^{N}] - \exp(-s_1 \alpha_1) E \exp[-\Gamma_1^{N}] E \exp[-\sum_{K=1}^{N} s_K W_K(l_0)]
\]

In the next section we provide explicit formulas for the means of the \( W_K \).
4. Mean Waiting-Times

Theorem (1) described a transformation that preserves the departure process from \( l_J \) (albeit permuting the order of service of tasks arriving during \( l_{J-1} \) active periods). There is a corresponding invariance in the waiting-time process which we state as a corollary:

*For ergodic arrival processes, the average end-to-end delay in the network is unchanged by the transformation of Theorem (1), where the averaging is over sources as well as sample paths.*

If the arrival processes are stationary, so that removing the pure delays in the transformed network described in Theorem (1) does not change the mean delay, then the network is equivalent, from the point of view of mean delay, to a single link [19]. In particular, for Poisson arrivals, over-all mean delay (averaged over all sources) is that of an M/D/1 queue:

\[
\sum_{J=1}^{N} \frac{\lambda_J}{\Lambda_N} \sum_{K=J}^{N} E W_K(S_J) = \frac{1}{2} \left( 1 - \frac{\rho_N}{\rho_N} \right) \alpha_N.
\]  

(15)

Individual mean waiting-times \( EW_{J+N}(S_J) \) can be obtained by differentiation from the joint mgf’s already calculated. When all the arrivals are Poisson and \( \alpha_J = \alpha \), we get

\[
\frac{1}{\alpha} EW_J(S_J) = \frac{1}{2} \left( \frac{\rho_J}{1 - \rho_J} - \frac{1}{2} \frac{\rho_{J-1}}{1 - \rho_{J-1}} \rho_{J-1} \right)
\]  

(16)

and

\[
\frac{1}{\alpha} EW_{J+N}(S_J) = \frac{1}{2} \left( \frac{\rho_{J+N}}{1 - \rho_{J+N}} - \frac{1}{2} \frac{\rho_{J+N-1}}{1 - \rho_{J+N-1}} + \lambda_{J+N} \alpha \sum_{K=0}^{N-1} \frac{1}{\alpha} EW_{J+K}(S_J),
\]

where \( N \geq 1 \). In non-recursive form,

\[
\frac{1}{\alpha} \sum_{K=0}^{N} EW_{J+K}(S_J) = \frac{1}{2} \left( \frac{\rho_{J+N}}{1 - \rho_{J+N}} + \frac{1}{2} \sum_{K=0}^{N-1} \left[ \frac{\rho_{J+K}}{1 - \rho_{J+K}}\lambda_{J+K+1} \alpha \right.ight. \\
\left. \left. - \frac{1}{2} \frac{\rho_{J-1}}{1 - \rho_{J-1}} \rho_{J-1} G_J(0, N) \right. \right)
\]  

(17)

where

\[
G_J(K, N) = (1 + \lambda_{J+K+1} \alpha) \cdots (1 + \lambda_{J+N} \alpha) \text{ for } K < N, \text{ and } G_J(K, N) = 1 \text{ for } K = N.
\]

Equation (15) is consistent with (16) and (17). From (16) it follows, as it should, that \( EW_{J+N}(S_J) = 0 \) when \( \lambda_{J+N} = 0 \). From (16) and (17), the mean end-to-end delay for \( S_J \) decreases in \( J \).

The Poisson approximation to the tandem queue consists in calculating mean waiting times at each link as though the net input to that link were Poisson. Equation (17) provides a vehicle for evaluating the Poisson approximation. Since \( \lambda_{J+K+1} \alpha G_J(K + 1, N) < 1 \), it follows from (17) that

\[
\sum_{k=0}^{N} EW_{J+K}(S_J) \leq \frac{\alpha}{2} \sum_{k=0}^{N} \frac{\rho_{J+K}}{1 - \rho_{J+K}}.
\]

The RHS is the Poisson approximation to \( S_J \) end-to-end delay. The Poisson approximation thus overbounds the exact result. The discrepancy between the two can be considerable, tending to be greatest when the flows interfering with \( S_J \) are small, and least when the \( S_J \) flow is small compared to those of \( S_{J+1}, S_{J+2}, \cdots \).
5. Conclusion

We have presented a complete delay analysis for the tandem net with non-increasing link capacities and multiple interfering sources. The analysis applies generally to the case that an M-G*D source encounters Poisson interference at each node in a tandem path. We recount the main steps, letting $P$, $\tilde{W}_J$, and $\tilde{r}_J$ ($J \geq I$) represent, respectively, a task entering the network at node $I$, the waiting-time in $l_J$ of the task which initiates the $l_{J-1}$ busy period containing $P$, and the rank of the initiating task in its $l_J$ busy period:

1. We showed that the pairs $(\tilde{W}_J, \tilde{r}_J)$, $J = I, J = I + 1, \cdots$ are independent; this property, is the key to the calculation of the joint waiting-time statistics across different nodes.

2. We calculated the mgf for the lengths of the busy periods in each link, and also, for each $J \geq I$, the steady-state mgf for the pair $(\tilde{W}_J, \tilde{r}_J)$; this was done through the artifice of a random time compression, the effect of which was to reveal an appropriate embedded M/G/1 queue.

3. From the statistics of the pairs $(\tilde{W}_J, \tilde{r}_J)$, $(J \geq I)$, we extrapolated to the joint mgf for the delays incurred by $P$ at each of $l_I, l_{I+1}, \cdots$.

The analysis accommodates batch arrivals without substantial modification. Waiting-time is related to rank as before. The rank statistics required in the calculation of the delay incurred by a given task $P$ are derived by first conditioning on the position of $P$ in its batch, and then proceeding as in Section (3). The delay incurred by a batch is the delay sustained by the last task in the batch.

Generalization of the analysis to the case that the link capacities are permitted to increase in the direction of flow, or that there are intermediate departures, remains, except for partial results in [7], an open problem.
Appendix

We derive here the joint steady-state mgf for waiting-time \( W \) and rank \( r \) in a batch-arrival M/G/1 queue. The random variables \( W_i, r_i, N_i, U_i \) denote, respectively, the following attributes of the \( i \)-th batch: time in queue incurred by the first task in the batch; the rank in its busy period of the first task in the batch (where in calculating rank tasks, rather than batches, are counted); the number of tasks in the batch; the total batch service time. \( W, r \) represent the steady-state of the Markov process \( \{W_i, r_i\} \). The batch arrival rate is \( \lambda \), and the inter-arrival times are \( \{T_i\} \). The calculations are based on the recursion

\[
W_{i+1} = \max(0, W_i + U_i - T_{i+1})
\]

\[
r_{i+1} = r_i + N_i \text{ when } W_{i+1} > 0, \text{ and } r_{i+1} = 1 \text{ otherwise},
\]

from which it follows that

\[
E[\exp(-sW_{i+1})z^{r_{i+1}}] = zPr\{W_{i+1} = 0\} + E[\exp(-sW_i)z^{r_i}f(s, z; W_i)],
\]

where

\[
f(s, z; w) = E[\exp(-sU_i + sT_{i+1})z^{N_i}|w + U_i - T_{i+1} > 0|Pr\{w + U_i - T_{i+1} > 0\}].
\]

Write \( \phi(s, z), \delta(s, z) \) for the mgf’s of \( (W, r - 1), (U_i, N_i) \) respectively. By first conditioning on the event \( \{U_i = u\} \) in the RHS of the defining expression for \( f \), it is easy to show that

\[
f(s, z; w) = [\lambda/(s - \lambda)](\exp(sw - \lambda w)\delta(\lambda, z) - \delta(s, z)).
\]

This, substituted into (A.1), yields

\[
\phi(s, z) = [A + (s - \lambda)Pr\{W = 0\}]/[s - \lambda + \lambda\delta(s, z)],
\]

where \( A = \lambda\delta(\lambda, z)\phi(\lambda, z) \) is independent of \( s \), but otherwise unknown. To determine \( A \), note first that \( A \) can be expressed in terms of the mgf of \( r - 1 \) simply by setting \( s = 0 \) on both sides of (A.2). So it remains only to calculate the mgf of \( r - 1 \). There are several ways, of which the simplest is the following:

Let \( a \) denote the (steady-state) rank in its busy-period of an arbitrary task, and let \( b \) be the rank of that task in its batch. From standard renewal -theoretic arguments, \( Pr\{a = i\} = Pr\{L \geq i\}/E[L] \), \( Pr\{b = i\} = Pr\{N \geq i\}/E[N] \), where \( L, N \) are, respectively, the length (in tasks) of a busy period and the size (in tasks) of a batch. From \( a = r + b - 1 \), where \( r, b \) are independent, it ensues that

\[
E[z^{r-1}] = E[z^a]/E[z^b] = \{E[N]/E[L]\}\{(E[z^L] - 1)/(E[z^N] - 1)\}.
\]

Now bring it all together, using the fact that \( Pr\{W = 0\} = 1 - \lambda E[U] \) (\( U \) being a typical batch service time) and that \( E[L] = E[N]/(1 - \lambda E[U]) \). The result is

\[
\phi(s, z) = \{1 - \lambda E[U]\}(s - \lambda + \lambda E[z^L])/(s - \lambda + \lambda\delta(s, z)).
\]
References


Figure (1): The Network