[2] C. J. Chase and P. J. Ramadge, "On real-time scheduling policies for flexible manufacturing systems," *IEEE Trans. Automat. Control*, vol. AC-37, pp. 491–496, April 1992.

[3] S. H. Lu and P. R. Kumar, "Distributed scheduling based on due dates and buffer priorities," *IEEE Transactions on Automatic Control*, vol. 36, pp. 1406–1416, December 1991.

[4] T. I. Seidman, ""First Come First Serve" is Unstable!," tech. rep., University of Maryland Baltimore County, 1993.

[5] M. Bramson, "Instability of FIFO queueing networks," tech. rep., Mathematics Department, University of Wisconsin, Madison, WI, 1993.

[6] F. P. Kelly, *Reversibility and Stochastic Networks*. New York, NY: John Wiley and Sons, 1979.

[7] A. N. Rybko and A. L. Stolyar, "On the ergodicity of stochastic processes describing open queueing networks," *Problemy Peredachi Informatsii*, vol. 28, pp. 2–26, 1991.

[8] P. R. Kumar and S. P. Meyn, "Stability of queueing networks and scheduling policies," tech. rep., C. S. L., University of Illinois, 1993.

[9] P. R. Kumar and S. Meyn, "Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies," tech. rep., C. S. L., University of Illinois, 1993.

[10] R. L. Cruz, "A calculus for network delay, part I: Network elements in isolation," *IEEE Transactions on Information Theory*, vol. 37, pp. 114–131, January 1991.

[11] R. L. Cruz, "A calculus for network delay, part II: Network analysis," *IEEE Transactions on Information Theory*, vol. 37, pp. 132–141, January 1991.

[12] G. I. Winograd, "The first come first served scheduling policy for deterministic systems," Master's thesis, University of University, Urbana, IL, November 1993.

[13] J. G. Dai, "On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid models," tech. rep., Georgia Institute of Technology, 1993.

Theorem: Performance of FCFS under regulators. *Consider the $M \times N$ re-entrant line shown in Figure 8. Consider the control by N regulators, each of service time $1/\lambda$, placed before the buffers $b_{11}, \ldots, b_{1N}$. Then, for* all *arrival rates λ in the capacity region*

$$\lambda \sum_{j=1}^{N} \tau_{ij} \leq 1 \text{ for } i = 1, 2, \ldots, M,$$

the number of parts in the system is uniformly upper bounded under FCFS by

$$\sigma + (N-1)(M+1) + \frac{NM(M+1)}{2}.$$

What is striking is that the upper bound does not diverge to infinity as $\rho_i \nearrow 1$.

10 Concluding Remarks

By obtaining a sharp bound on a tandem segment of a system, we have shown that one can establish the stability of some topologies under FCFS. We have also shown how to obtain performance bounds by using traffic burstiness regulators. The performance bounds may remain bounded even under heavy traffic.

It would be valuable to extend these results to more classes of systems. It could well be that the stability of systems under other statistical assumptions may well be related to their stability under the burstiness assumptions considered here. Some evidence for this is the recent fluid limit approach to stability of Dai [13].

References

[1] P. R. Kumar and T. I. Seidman, "Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems," *IEEE Trans. Automat. Control*, vol. AC-35, pp. 289–298, March 1990.

burst parameter can ever have is 1, unless the traffic stream is empty).

Now we apply Theorem 1 of Section 3, and deduce that the burst parameter $\widehat{\sigma}_J$ which, together with rate parameter λ, constrains the traffic leaving buffer b_{MJ}, is

$$\widehat{\sigma}_J \leq \delta_{M+1,J} + \lambda \sum_{j=1}^{N} \sum_{i=1}^{M} \delta_{ij} \tau_{ij} \tag{17}$$

subject to the constraints,

$$\sum_{i=i_1}^{i_2} \delta_{ij} \leq 1 \text{ for } 1 \leq i_1 \leq i_2 \leq M+1 \text{ and } 1 \leq j \leq N.$$

For the above bound and the rest of the result we assume only that the capacity constraint

$$\rho_i := \lambda \sum_{j=1}^{N} \tau_{ij} \leq 1 \tag{18}$$

is satisfied, with equality allowed. We see that since $\delta_{ij} \leq 1$, $\lambda \sum_{j=1}^{N} \delta_{ij} \tau_{ij} \leq 1$. Hence we have

$$\widehat{\sigma}_J \leq M+1.$$

Since the traffic leaving b_{Mj} enters a regulator with service time $\frac{1}{\lambda}$, we see that the buffer level of the parts in each regulator is bounded by $(M+1)$. Hence the total number of parts in $(N-1)$ of the regulators is bounded by $(N-1)(M+1)$, while in the regulator preceding b_{11} it is bounded by σ, giving a total bound over *all* regulators of $(N-1)(M+1)+\sigma$.

At the buffer b_{iJ}, a similar argument shows that the bound on the number of parts is $i+1$. Hence an upper bound on the total number of parts in the system is

$$\sigma + (N-1)(M+1) + \frac{NM(M+1)}{2}.$$

To jointly illustrate the power of our bounds, that good control is possible, and that the bounds on delay so obtained remain bounded even when nominal loads are *equal* to 1, we examine the following $M \times N$ re-entrant line shown in Figure 8.

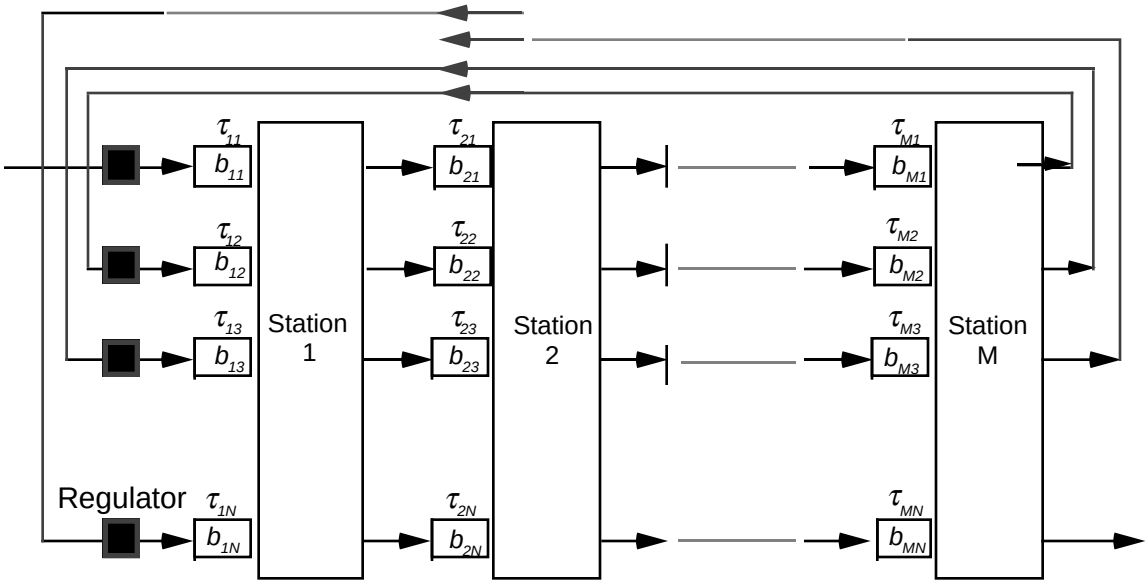

Figure 8: An $M \times N$ re-entrant line with N regulators. The regulators are depicted by small rectangles which precede the buffers of station 1.

It consists of M stations, each visited N times "cyclically," before parts leave the system. As shown in Figure 8, preceding the buffers at station 1 are N regulators, depicted by small rectangles. Each regulator is a "server," introduced for purposes of control. In particular, the regulator preceding buffer b_{ij} is a server which takes $1/\lambda$ time units to process a part, where λ is the arrival rate of parts to the system.

The effect of introducing such a regulator is that the traffic input to buffer b_{1j} is constrained by rate parameter λ and burst parameter 1. (We note that the smallest value any

We note that one only requires $\rho_1 := \lambda(\tau_{11} + \tau_{12}) \leq 1$ and $\rho_2 := \lambda(\tau_{21} + \tau_{22}) \leq 1$ for this bound. Suppose now that $\lambda\tau_{12} < 1$ and $\lambda\tau_{22} < 1$. Then we obtain

$$\widehat{\sigma}_{12}(t) \leq \frac{\sigma + \lambda\sigma(\tau_{11} - \tau_{21})^+}{1 - \lambda \max(\tau_{12}, \tau_{22})} \text{ for all } t \geq 0.$$

This establishes the stability of the system.

Theorem: Stability of 2×2 Re-Entrant Lines under FCFS. *Consider the 2×2 re-entrant line shown in Figure 7. Let $\rho_i := \lambda(\tau_{i1} + \tau_{i2})$ be the nominal load of station i. We assume (only) that*

$$\rho_i \leq 1, \lambda\tau_{12} < 1 \text{ and } \lambda\tau_{22} < 1.$$

Then the system is stable under the FCFS scheduling policy. In particular together with the rate parameter λ, all traffic streams have a finite burst parameter. Also, the buffer levels are bounded, and all parts experience a bounded delay.

9 Controlling Systems by Regulating Burstiness

In many systems, it is unknown whether the given system is stable under the FCFS policy. Even it is stable, the performance as measured by the buffer levels attained, or equivalently the delay experienced by parts, may be unacceptable. For both these reasons, one wishes to control a system to render it stable and operate it so that all parts are guaranteed to have a small delay.

One such control scheme is to use "regulators," studied in [10,11]. This is also reminiscent of a "leaky bucket" scheme.

hold, then the system is stable under the FCFS scheduling policy. In particular the traffic streams into and out of every buffer b_{ij} are constrained by a finite burstiness parameter σ_{ij} and rate parameter λ_j. Also, the buffer levels are bounded for all time, and the delays experienced by parts are also bounded.

8 Stability of 2×2 Re-Entrant Lines Under FCFS

Consider now the 2×2 re-entrant line shown in Figure 7.

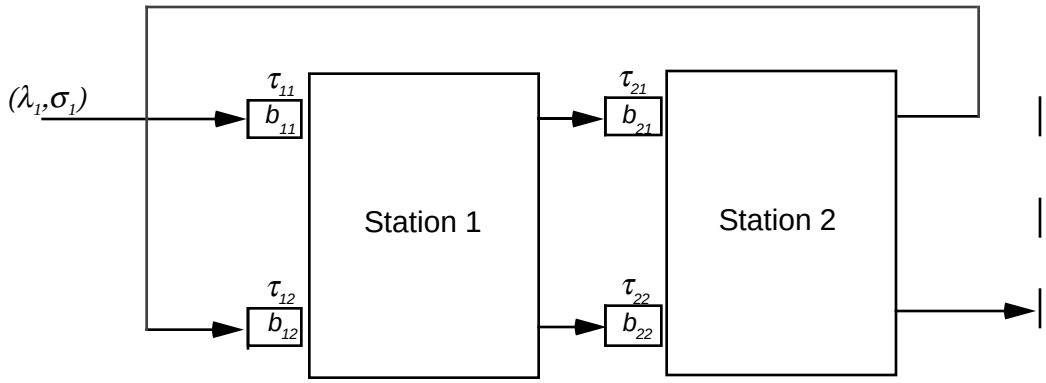

Figure 7: A 2×2 re-entrant line.

Let $\widehat{\sigma}_{12}(t)$ be the smallest burst parameter, which, together with the rate parameter λ, constrains the input to buffer b_{12} over the time interval $[0, t)$.

We will now apply the bound for the two station system of Example 3, treating the input to b_{12} as an exogenous stream, constrained by burst parameter $\widehat{\sigma}_{12}(t)$ and rate parameter λ, over the time interval $[0, t)$. The output from buffer b_{21} then has burst parameter $\widehat{\sigma}_{12}(t)$ (since it is the input to b_{12}) over the same time interval, which satisfies,

$$\widehat{\sigma}_{12}(t) \leq \sigma + \lambda\sigma(\tau_{11} - \tau_{21})^+ + \lambda\widehat{\sigma}_{12}(t)\max(\tau_{12}, \tau_{22}).$$

Following the same line of analysis presented in [10], we wish to show that $(I-A)^{-1} = I + A + A^2 + \cdots$. For this it suffices to show that every eigenvalue γ of A satisfies $|\gamma| < 1$. To see this consider an eigenvalue γ of A with eigenvector $x := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Then since $Ax = \gamma x$, we have

$$\Lambda^{(1)} T^{(2)T} x_2 = \gamma x_1 \text{ and } \Lambda^{(2)} T^{(1)T} x_1 = \gamma x_2.$$

Hence $\Lambda^{(1)} T^{(2)T} \Lambda^{(2)} T^{(1)T} x_1 = \gamma^2 x_1$, and so γ^2 is an eigenvalue of $\Lambda^{(1)} T^{(2)T} \Lambda^{(2)} T^{(1)T}$. However this latter matrix has rank 1, and its only nonzero eigenvalue is $T^{(2)T} \Lambda^{(2)} T^{(1)T} \Lambda^{(1)}$. Hence

$$|\gamma^2| \leq T^{(2)T} \Lambda^{(2)} T^{(1)T} \Lambda^{(1)} = (\sum_{i=n+1}^{R} \lambda_i \tau_{1i})(\sum_{i=1}^{n} \lambda_i \tau_{2i}) < 1$$

due to the capacity conditions

$$\sum_{i=1}^{R} \lambda_i \tau_{1i} < 1 \text{ and } \sum_{i=1}^{R} \lambda_i \tau_{2i} < 1,$$

which we will assume to hold. Hence, noting that A is a nonnegative matrix, we obtain

$$S(t) \leq (I + A + A^2 + \cdots) C$$

for all $t \geq 0$, proving the boundedness of the burstiness parameters of all traffic entering any buffer, with the rate parameter equal to the appropriate λ_j. Hence we have proved the following theorem.

Theorem: Stability of FCFS for bidirectional systems. *Consider the bidirectional system shown in Figure 6, with arbitrary n and R. If the capacity conditions,*

$$\sum_{i=1}^{R} \lambda_i \tau_{ij} < 1 \text{ for } j = 1, 2$$

$$T^{(1)} := \begin{bmatrix} \tau_{21} \\ \tau_{22} \\ \vdots \\ \tau_{2n} \end{bmatrix}, \qquad T^{(2)} := \begin{bmatrix} \tau_{1n+1} \\ \tau_{1,n+2} \\ \vdots \\ \tau_{1R} \end{bmatrix},$$

$$S^{(1)}(t) := \begin{bmatrix} \widehat{\sigma}_{21}(t) \\ \widehat{\sigma}_{22}(t) \\ \vdots \\ \widehat{\sigma}_{2n}(t) \end{bmatrix}, S^{(2)}(t) := \begin{bmatrix} \widehat{\sigma}_{1,n+1}(t) \\ \widehat{\sigma}_{1,n+2}(t) \\ \vdots \\ \widehat{\sigma}_{1R}(t) \end{bmatrix}, S(t) := \begin{bmatrix} S^{(1)}(t) \\ S^{(2)}(t) \end{bmatrix},$$

$$C := \begin{bmatrix} \sigma_1 + \lambda_1 \sum_{j=2}^{n} \sigma_j \tau_{1j} \\ \sigma_2 + \lambda_2 \sum_{\substack{j=1 \\ j \neq 2}}^{n} \sigma_j \tau_{1j} \\ \quad \vdots \\ \sigma_n + \lambda_n \sum_{j=1}^{n-1} \sigma_j \tau_{1j} \\ \sigma_{n+1} + \lambda_{n+1} \sum_{j=n+2}^{R} \sigma_j \tau_{2j} \\ \sigma_{n+2} + \lambda_{n+2} \sum_{\substack{j=n+1 \\ j \neq n+2}}^{R} \sigma_j \tau_{2j} \\ \quad \vdots \\ \sigma_R + \lambda_R \sum_{j=n+1}^{R-1} \sigma_j \tau_{2j} \end{bmatrix}$$

and

$$A := \begin{bmatrix} 0 & \Lambda^{(1)} T^{(2)T} \\ \Lambda^{(2)} T^{(1)T} & 0 \end{bmatrix}.$$

Then we can rewrite (15, 16) together as the vector inequality,

$$S(t) \leq C + AS(t),$$

i.e., as,

$$(I - A)S(t) \leq C.$$

station 2, and then depart the system. Traffic streams $n+1, n+2, \ldots, R$ enter at station 2, move next to station 1, and then depart the system. Let b_{ij} be the label of the buffer that holds parts from traffic stream j at station i, and let τ_{ij} be the corresponding service time of the parts at that buffer.

Let $\widehat{\sigma}_{ij}(t)$ denote the smallest burstiness parameter, which, together with rate parameter λ_j, constrains the traffic *input* to buffer b_{ij} *over the time interval* $[0,t)$, i.e.,

$$\#\text{ of arrivals to } b_{ij} \text{ in } [a,b) \leq \lambda_j(b-a) + \widehat{\sigma}_{ij}(t) \text{ for all } 0 \leq a \leq b \leq t.$$

Suppose now that $(\lambda_j, \widehat{\sigma}_{1j}(t))$ constrains the input to buffers b_{1j} for $n+1 \leq j \leq R$ over the time interval $[0,t)$. We also know that (λ_j, σ_{1j}) constrains the inputs to buffer b_{1j} for $1 \leq j \leq n$. Now we view station 1 as though it were in isolation, and apply the bounds obtained from Example 1 of the preceding section, to obtain bounds on the burstiness $\widehat{\sigma}_{2J}(t)$ of traffic entering station 2 from buffers b_{1j} with $1 \leq J \leq n$ in the time interval $[0,t)$. This gives the bounds,

$$\widehat{\sigma}_{2J}(t) \leq \sigma_J + \lambda_J \sum_{\substack{j=1 \\ j \neq J}}^{n} \sigma_j \tau_{1j} + \lambda_J \sum_{j=n+1}^{R} \bar{\sigma}_{1j}(t) \tau_{1j} \text{ for } 1 \leq J \leq n. \tag{15}$$

By a symmetric argument, we also have

$$\widehat{\sigma}_{1J}(t) \leq \sigma_J + \lambda_J \sum_{j=1}^{n} \widehat{\sigma}_{2j}(t) \tau_{2j} + \lambda_J \sum_{\substack{j=n+1 \\ j \neq J}}^{N} \sigma_j \tau_{2j} \text{ for } n+1 \leq J \leq R. \tag{16}$$

To rewrite the inequalities (15) and (16) in vector notation, we introduce the following matrices:

$$\Lambda^{(1)} := \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix}, \qquad \Lambda^{(2)} := \begin{bmatrix} \lambda_{n+1} \\ \vdots \\ \lambda_R \end{bmatrix},$$

7 Stability of Bidirectional Manufacturing Systems

We now use the bounds of Theorem 1 to establish the stability of some network topologies under the FCFS policy, for all arrival rates and service rates within the capacity of the system. By "capacity," we mean here that $\rho_i < 1$, for all stations i. However, we will see that in some systems stability also holds for $\rho_i = 1$.

The first network topology for which we establish this strong stabilizing property of FCFS is the class of bidirectional manufacturing systems, an example of which is shown in Figure 6. A special case of this system with $R = 2$ opposing routes was shown to be stable for a fluid traffic model in [7].

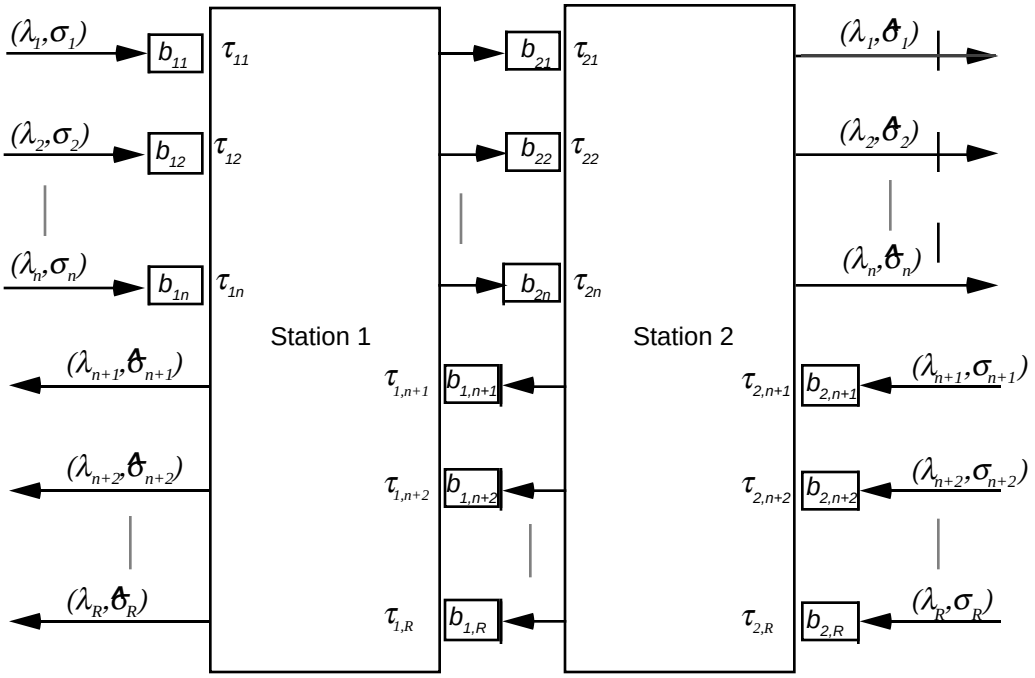

Figure 6: Bidirectional manufacturing system.

There are two stations. Traffic streams labeled $1, 2, \ldots, n$ enter at station 1, move next to

$$\delta_{1j} + \delta_{2j} \leq \sigma_j \quad \text{for} \quad j = 1, \ldots, R, \tag{14}$$

$$\begin{aligned} \delta_{1J} + \delta_{2J} + \delta_{3J} &\leq \sigma_J, \\ \delta_{2J} + \delta_{3J} &\leq \sigma_J. \end{aligned}$$

From Theorem 1, we know that one can choose $\delta_{2J} = \sigma_J$, and attain the maximum. Also, from (13, 14) we see that setting $\delta_{1j} = \sigma_j$ if $\tau_{1j} \geq \tau_{2j}$ and $\delta_{2j} = \sigma_j$ if $\tau_{1j} < \tau_{2j}$ also realizes the maximum.

Hence, the upper bound is,

$$\text{Max} \quad \widehat{\sigma}_J$$

subject to:

$$\begin{aligned} \widehat{\sigma}_J &= \sigma_J + \lambda_J(\delta_{1J}\tau_{1J} + \delta_{2J}\tau_{2J}) + \lambda_J \sum_{j \neq J} \sigma_j \max(\tau_{1j}, \tau_{2j}), \\ \delta_{1J} &\leq \sigma_J, \\ \delta_{2J} &\leq 0, \\ \delta_{1J} + \delta_{2J} &\leq 0. \end{aligned}$$

This yields the bound,

$$\widehat{\sigma}_J = \sigma_J + \lambda_J \sigma_J (\tau_{1J} - \tau_{2J})^+ + \lambda_J \sum_{j \neq J} \sigma_j \max(\tau_{1j}, \tau_{2j}).$$

□

□

Cruz ([10]) has obtained a similar bound in his analysis of the FCFS multiplex network element in isolation.

Example 3: Two stations in tandem.

Consider the two station system shown in Figure 5.

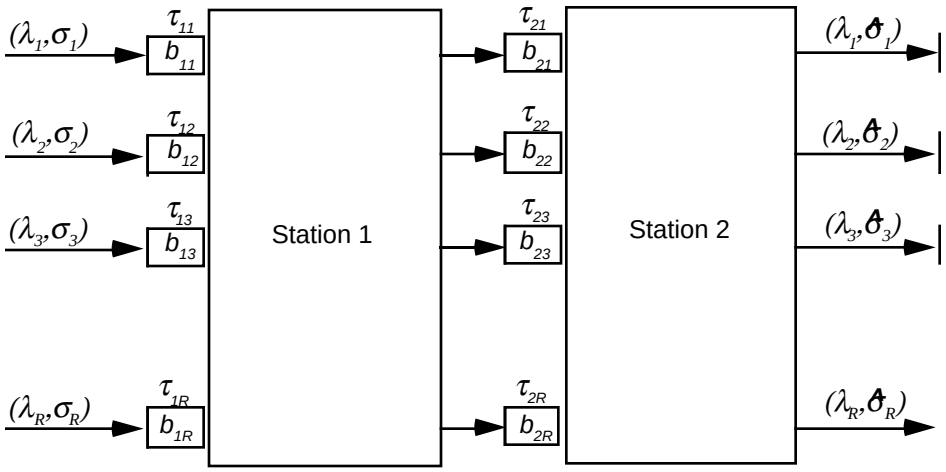

Figure 5: Two stations in tandem.

Then the linear program for $\widehat{\sigma}_J$ is:

$$\text{Max} \quad \widehat{\sigma}_J$$

subject to:

$$\widehat{\sigma}_J = \delta_{3J} + \lambda_J \sum_j (\delta_{1j}\tau_{1j} + \delta_{2j}\tau_{2j}),$$

$$\delta_{1j} \leq \sigma_j \text{ and } \delta_{2j} \leq \sigma_j \quad \text{for} \quad j = 1, \ldots, R, \tag{13}$$

6 Application of the Burstiness Bounds

We now apply these bounds to two specific systems, obtaining explicit expressions for the burst parameter constraints. These specific bounds will be used later to establish the stability of some network topologies under FCFS.

Example 2: A single station.

Consider the system shown in Figure 4, consisting of a single station, i.e., $M = 1$.

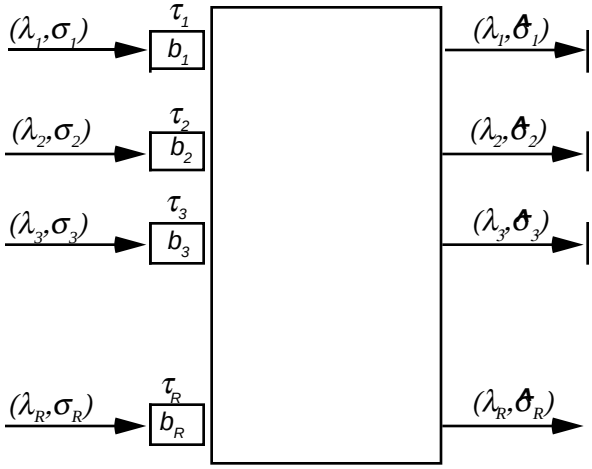

Figure 4: A single station system.

There are R traffic streams, with rate parameter λ_j and burst parameter σ_j for the j-th incoming stream.

We suppose that $\rho := \sum_{j=1}^{R} \lambda_j \tau_j \leq 1$. Then,

$$\hat{\sigma}_j \leq \sigma_J + \lambda_J \sum_{\substack{j=1,\ldots,R \\ j \neq J}} \sigma_j \tau_j.$$

of the behavior of all stations $i < I$ do not affect the chain of busy periods. The required bound can then be obtained by considering only the terms for $I \leq i \leq M$. Two cases need to be considered for i.

Case 1: $k_i < k_{i+1} \leq 0$.

All parts participating in the chain of busy periods at station i are then part of the initial conditions. These parts can be considered to have entered the system at time 0. The length of the time interval necessary to process these parts at station i is simply the sum of their individual service times at station i. Therefore,

$$\begin{array}{rcl} t_{k_{i+1}} - t_{k_i} & = & 0 \\ t'_{k_{i+1}} - t'_{k_i} & = & \sum_{k=k_i}^{k_{i+1}} \tau_{i,r_k} \end{array} \tag{11}$$

where r_k is the route of part π_k.

Case 2: $k_i \leq 0 < k_{i+1}$.

Here, only some of the parts in the busy period at station i are part of the initial conditions, while the rest of the parts arrived later. The length of the busy period at station i is the sum of the service times of the parts present in the initial conditions, and the parts which entered during the time interval $t_{i+1} - t_i$. Therefore,

$$t'_{k_{i+1}} - t'_{k_i} \leq \sum_{k=k_i}^{0} \tau_{i,r_k} + \sum_{j=1}^{M} (\lambda_j (t_{k_{i+1}} - t_{k_i} + \delta_{ij}) \tau_{ij} \tag{12}$$

where r_k is the route number of part π_k.

Using (11) or (12) in place of (7), and condensing all constants into a single constant C, the result is proved. □

route J traffic leaving the system, for rate parameter λ_J, is

$$\overline{\sigma}_J = \widehat{\sigma}_j + C, \tag{10}$$

where C is some constant which depends only on the initial conditions, while $\widehat{\sigma}_j$ is the same as in Theorem 1.

Proof.

For the initial conditions to specify the future behavior, there must be an initial ordering among the parts so that the service stations can implement the FCFS scheduling policy.

Thus the parts may be labeled $\pi_{-K}, \pi_{-K+1}, \ldots, \pi_0$ where K is the total number of parts present in the system at time $t = 0$, and part π_{-k} is guaranteed to exit the system no later than part π_{-k+1}. Also define $t_{-k} \equiv 0$ for $k \geq 0$ (i.e., assume that all these parts were placed in the system at time 0).

Consider a part sequence $\{\pi_{-K} \ldots, \pi_0, \pi_1, \ldots, \pi_F, \ldots, \pi_G, \ldots\}$, and recursively compute k_i, as in Theorem 1. One of two outcomes may occur. It may be that $k_i \geq 1$ for all i. In this case, the computations of Theorem 1 are still valid and the proof goes through as before. This case corresponds to the initial conditions not having any effect on the burstiness achieved between parts π_F and π_G in the exit stream.

Another possibility is that $k_i \leq 0$ for some i. In this case the recursions for k_ℓ may not be defined for all $\ell \leq M$. Specifically, if part π_{k_I} belongs to the initial condition at station I at time $t = 0$, then k_{I-1} is not defined. This means that the busy period at station I is the first busy period in the chain of busy periods starting at time t'_{k_i}. Therefore, the details

since all route 3 service times are 0, thus achieving the upper bound. □

5 Initial Conditions and Bottlenecks

Stations which precede or follow a bottleneck do not contribute much to the burstiness, since the bottleneck stations regulate the traffic. This allows us to refine the bound for some special situations. We refer the reader to Winograd [12] for the proof.

Lemma 2: Bottlenecks. *Consider a tandem system of the form depicted in Figure 1, where there is either a set of consecutive stations indexed by* $I, I+1, \ldots, i$ *such that* $\tau_{i'j} < \tau_{Ij}$ *for all* $I < i' \leq i \leq M$ *and* $1 \leq j \leq R$, *or a set of consecutive stations indexed by* $i, i+1, \ldots, I$ *such that* $\tau_{i'j} < \tau_{Ij}$ *for all* $i \leq i' < I$ *and* $1 \leq j \leq R$. *Then the bound (2) on burstiness can be strengthened by replacing all of the terms in the summation in (3) involving stations* $I+1, I+2, \ldots i$, *or involving stations* $i, i+1, \ldots, I-1$, *by a single constant* C *which does not depend on the* λ_j *and* σ_j *parameters constraining the input to the system.*

Next we turn to the issue of initial conditions. In Theorem 1 we have assumed that the system is initially empty. However non-zero initial conditions are easily incorporated into the burstiness bound, and simply add a constant which depends only on the initial conditions.

Lemma 3: Initial Conditions. *If some number of parts are initially present in the buffers at some of the stations at time* $t = 0$, *then the upper bound* $\bar{\sigma}_j$ *on the burst parameter of*

each station be:

$$\tau_{ij} \quad = \quad \begin{cases} 1 & \text{for } j = 1 \text{ and } i \text{ odd} \\ 1 & \text{for } j = 2 \text{ and } i \text{ even} \\ 0 & \text{otherwise.} \end{cases}$$

We note that even parts with zero service times need service attention, given in FCFS order, before they can move to their next station. An assignment of values to the δ_{ij} parameters which maximizes the bound on $\widehat{\sigma}_2$ is $\delta_{ij} = \sigma$ when $\tau_{ij} = 1$ and $\delta_{ij} = -\sigma$ when $\tau_{ij} = 0$. The optimality of this assignment is evident since $\delta_{ij} \leq \sigma_j$, and since the maximum is over the sum of the product terms $\delta_{ij}\tau_{ij}$. The upper bound on the traffic burstiness exiting from route 2 is therefore $\widehat{\sigma}_i = (M+1)\sigma$.

Now consider the following time sequence of part arrivals. At time $t = 0$, σ parts enter along route 1. Then, an additional part enters along route 1 once per time unit, for the next σ time units. At time $t = \sigma$, σ parts enter along route 2. Then, an additional part enters along route 2 once per time unit, for the next σ time units. This sequence of part arrivals over an interval of length 2σ is repeated at times $t = 2\sigma$, 4σ, 6σ, ... The result of such arrivals is that at time $4m\sigma$ there will be 2σ parts in buffers $b_{31}, b_{51}, b_{71}, \ldots, b_{2m+1,1}$, and 2σ parts in buffers $b_{22}, b_{42}, b_{62}, \ldots b_{2m,2}$. At time $2(M-1)\sigma$, every buffer b_{ij} such that $\tau_{ij} = 1$, will hold 2σ, except for buffer b_{11}. The remaining buffers will be empty.

At this time, let an additional σ parts enter the system along route 1, bringing the route 1 total up to $M\sigma$ parts. If route 1 and route 2 traffic is now suddenly shut off, then it will take $M\sigma$ time units to clear all parts from the system. If $M\sigma + \sigma$ parts enter the system along route 3 during this time interval, then all these parts will leave the system simultaneously

Remark. *In situations where the nominal load ρ_i at station i is strictly less than 1, it is possible to further improve the bound. Briefly, if a coefficient δ_{ij} in the linear program is negative, it implies that the corresponding time interval $[t_{k_i}, t_{k_i+1})$ is at least as large as $-\delta_{ij}$. This can be used to strengthen the inequality in (8).*

4 The Sharpness of the Linear Program Bound

We now show that the above linear program upper bound on traffic burstiness can be tight, even for arbitrarily large systems.

Example 1: A system for which the burstiness bound is attained.

Consider a system with M stations, where M is any odd integer, and let there be $R = 3$ routes through the system. An example of such a system with $M = 3$ is depicted in Figure 3.

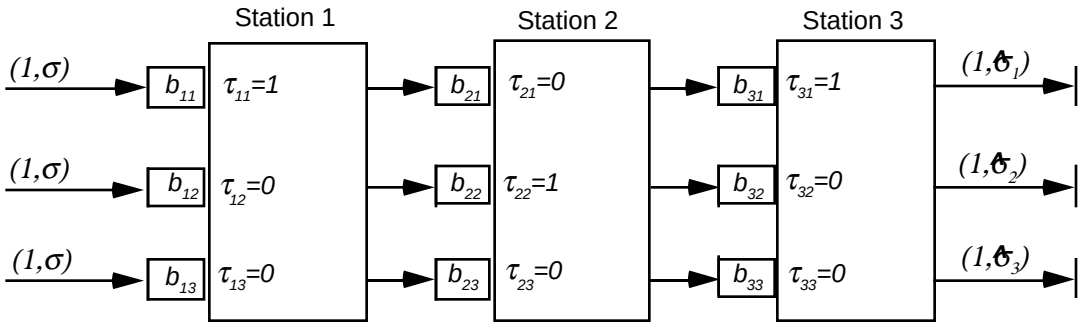

Figure 3: The system which achieves the upper bound on burstiness.

Let the input of parts along each route $j = 1, 2, 3$ be constrained by rate parameter $\lambda_j \equiv 1$, and burst parameter $\sigma_j \equiv \sigma$, for some positive integer σ. Let the service times at

In turn, for $1 \leq i \leq M$,

$$\begin{aligned} t'_{k_{i+1}} - t'_{k_i} &= (\text{Time that } \pi_{k_{i+1}} \text{ left station } i) \\ &\quad -(\text{Time that } \pi_{k_i} \text{ left station } i-1) \\ &= (\text{Time that } \pi_{k_{i+1}} \text{ left station } i) \\ &\quad -(\text{Time that } \pi_{k_i} \textit{ arrived} \text{ at station } i \text{ and initiated a busy period there}) \\ &= \text{Time to serve parts between } \pi_{k_i} \text{ and } \pi_{k_{i+1}} \text{ at station } i \qquad (7) \\ &= \text{Time to serve at station } i, \text{ those parts that arrived} \\ &\quad \text{to the system between } t_{k_i} \text{ and } t_{k_{i+1}} \\ &= \sum_{j=1}^{R} \left[\lambda_j (t_{k_{i+1}} - t_{k_i}) + \delta_{ij}\right] \tau_{ij}. \end{aligned}$$

Hence, from (7), for $1 \leq i \leq M$

$$\begin{aligned} &(t'_{k_{i+1}} - t'_{k_i}) - (t_{k_{i+1}} - t_{k_i}) \\ &\quad = (t_{k_{i+1}} - t_{k_i})[(\sum_{j=1}^{R} \lambda_j \tau_{ij}) - 1] + \sum_{j=1}^{R} \delta_{ij}\tau_{ij} \qquad (8) \\ &\quad \leq \sum_{j=1}^{R} \delta_{ij}\tau_{ij} \qquad (\text{since } \rho_i \leq 1). \end{aligned}$$

Substituting (8) in (6), and then (6) in (5), we obtain

$$\widehat{\sigma}_J \leq \delta_{M+1,J} + \lambda_J \sum_{i=1}^{M} \sum_{j=1}^{R} \delta_{ij}\tau_{ij}.$$

Hence $\widehat{\sigma}_J$ is upper bounded by (2), subject to the constraints (3) and (4).

Now we show that the maximum can be achieved with the choice $\delta_{M+1,J} = \sigma_J$. Note that the terms in (3) which involve the route index J are:

$$\lambda_J[\delta_{1J}\tau_{1J} + \delta_{2J}\tau_{2J} + \ldots + \delta_{MJ}\tau_{MJ}] + \delta_{M+1,J}. \qquad (9)$$

This is to be maximized subject to (4). Now suppose that I is the highest index less than or equal to M, for which $\delta_{IJ} > 0$. Then, if $\delta_{M+1,J} < \sigma_J$, one can transfer a small amount from δ_{IJ} to $\delta_{M+1,J}$. This preserves all the constraints, while increasing the value of (9), since $\lambda_J \tau_{IJ} < 1$. This proves the claim. □

We note that the number of parts that enter the system in $[t_{k_{i_1}}, t_{k_{i_2+1}})$ along route j is then

$$\sum_{i=i_1}^{i_2} \left[\lambda_j \left(t_{k_{i+1}} - t_{k_i}\right) + \delta_{ij}\right] = \lambda_j \left(t_{k_{i_2+1}} - t_{k_{i_1}}\right) + \sum_{i=i_1}^{i_2} \delta_{ij}.$$

However, since traffic entering the system on route j is assumed to be constrained by rate parameter λ_j and burst parameter σ_j, we see that the δ_{ij}'s have to satisfy the inequalities,

$$\sum_{i=i_1}^{i_2} \delta_{ij} \leq \sigma_j \text{ for all } i_1 \leq i_2 \leq M+1.$$

This is true for all $j = 1, 2, \ldots, R$, and so constraint (4) is satisfied by the δ_{ij}'s.

Now let us consider the burstiness of traffic exiting the system along route J, between the departures of π_F and π_G, which were chosen arbitrarily. Let

$$\widehat{\sigma}_J := (\#\text{ of parts along route } J \text{ between } \pi_F \text{ and } \pi_G) - \lambda_J(t'_G - t'_F)$$

be the burstiness achieved, with rate parameter λ_J.

We obtain a bound for $\widehat{\sigma}_J$ as follows,

$$\begin{aligned} \widehat{\sigma}_J &= (\#\text{ of arrivals between } \pi_F \text{ and } \pi_G) - \lambda_J(t'_G - t'_F) \\ &= \lambda_J(t_G - t_F) + \delta_{M+1,J} - \lambda_J(t'_G - t'_F) \\ &= \delta_{M+1,J} + \lambda_J[(t_G - t_F) - (t'_G - t'_F)]. \end{aligned} \tag{5}$$

Now note that,

$$\begin{aligned} (t_G - t_F) - (t'_G - t'_F) &\leq (t_G - t_F) - (t_G - t'_F) \quad (\text{since } t_G \leq t'_G) \\ &= t'_F - t_F \\ &= [t'_{k_1} + (t'_{k_2} - t'_{k_1}) + (t'_{k_3} - t'_{k_2}) + \cdots + (t'_F - t'_{k_M})] \\ &\quad -[t_{k_1} + (t_{k_2} - t_{k_1}) + (t_{k_3} - t_{k_2}) + \cdots + (t_F - t_{k_M})] \\ &= [(t'_{k_2} - t'_{k_1}) + (t'_{k_3} - t'_{k_2}) + \cdots + (t'_F - t'_{k_M})] \\ &\quad -[t_{k_2} - t_{k_1}) + (t_{k_3} - t_{k_2}) + \cdots + (t_F - t_{k_M})] \end{aligned} \tag{6}$$

Let π_{k_M} be the part which initiates the busy period containing part π_F at station M. Recursively define π_{k_n} as the part which initiates the busy period containing $\pi_{k_{n+1}}$ at station n.

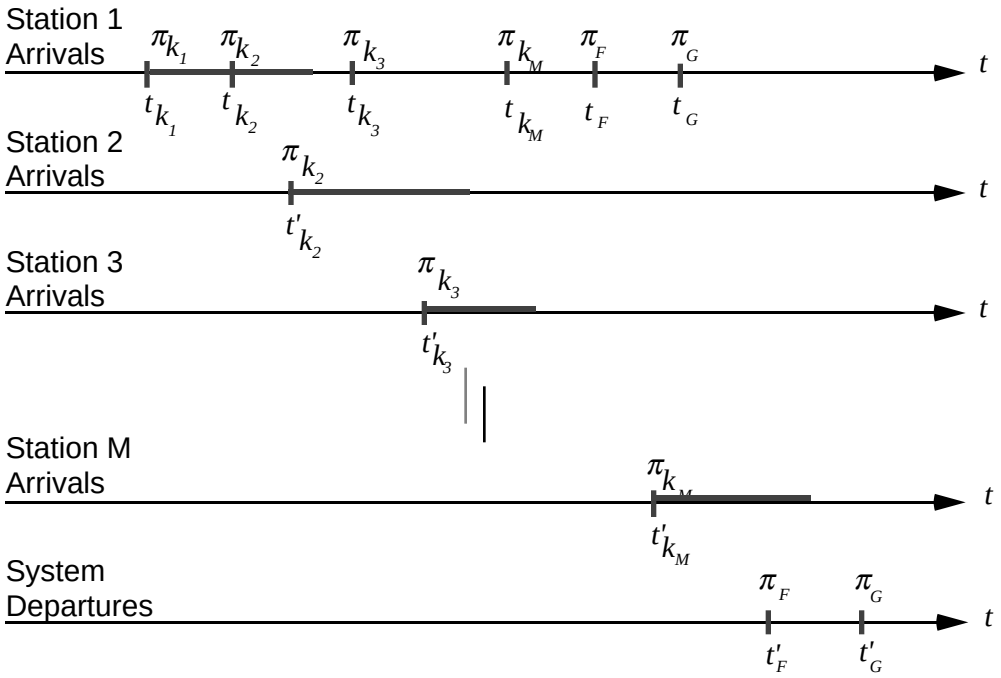

Figure 2: Event times for tandem system. The busy periods are shown in solid black.

Let t'_F and t'_G be the times at which parts π_F and π_G exit the system. Denote by t'_{k_n} the time that part π_{k_n} leaves station $n-1$ and arrives at station n. We take $t'_{k_1} = t_{k_1}$. For simplicity, denote $\pi_{k_{M+1}} := \pi_F, t_{k_{M+1}} := t_F, \pi_{k_{M+2}} := \pi_G$ and $t_{k_{M+2}} := t_G$. Figure 2 shows the sequence of these events.

Let us suppose that the total number of parts which arrive to the system during the time interval $[t_{k_i}, t_{k_{i+1}})$ on route j is $\lambda_j(t_{k_{i+1}} - t_{k_i}) + \delta_{ij}$, where δ_{ij} may be positive or negative.

Theorem 1: Linear Program Bound on Burstiness of Traffic Exiting a Tandem System. *Suppose $\rho_i \leq 1$ for all stations $i = 1, 2, \ldots, R$, and the system is initially empty. Then the route J traffic exiting the system is constrained by rate parameter λ_J and burst parameter $\hat{\sigma}_J$, where $\hat{\sigma}_J$ is the value of the following linear program:*

$$Max \quad \hat{\sigma}_J \tag{2}$$

subject to:

$$\hat{\sigma}_J = \delta_{M+1,J} + \lambda_J \sum_{j=1}^{R} \sum_{i=1}^{M} \delta_{ij} \tau_{ij}, \tag{3}$$

$$\sum_{i=i_1}^{i_2} \delta_{ij} \leq \sigma_j \text{ for all } 1 \leq i_1 \leq i_2 \leq M+1, \text{ and } 1 \leq j \leq R. \tag{4}$$

Also, the maximum above is attained when $\delta_{M+1,J} = \sigma_J$. We note that the variables δ_{ij} have no sign constraints, and that the linear program is bounded.

Proof.

Let $\{\pi_1, \pi_2, \ldots, \pi_F, \ldots, \pi_G, \ldots\}$ be the sequence of parts which arrive to the system in the order shown, and let their arrival times be $\{t_1, t_2, \ldots, t_F, \ldots, t_G, \ldots\}$ and their routes be $\{r_1, r_2, \ldots, r_F, \ldots, r_G, \ldots\}$, respectively. Note that $t_1 \leq t_2 \leq \cdots \leq t_F \leq \cdots \leq t_G$. Without loss of generality, we suppose that $t_1 = 0$. Let us suppose that parts π_F and π_G follow the same route J, i.e., $r_F = r_G = J$.

We note that due to the FCFS scheduling policy, since all parts visit all the stations in the same order, and since there is only one server at each station, no part can ever overtake another part. Hence parts exit each station in the order that they arrive to the system.

3 Linear Programming Delay and Burstiness Bounds for Tandem Systems

It is useful to consider tandem systems of the type shown in Figure 1.

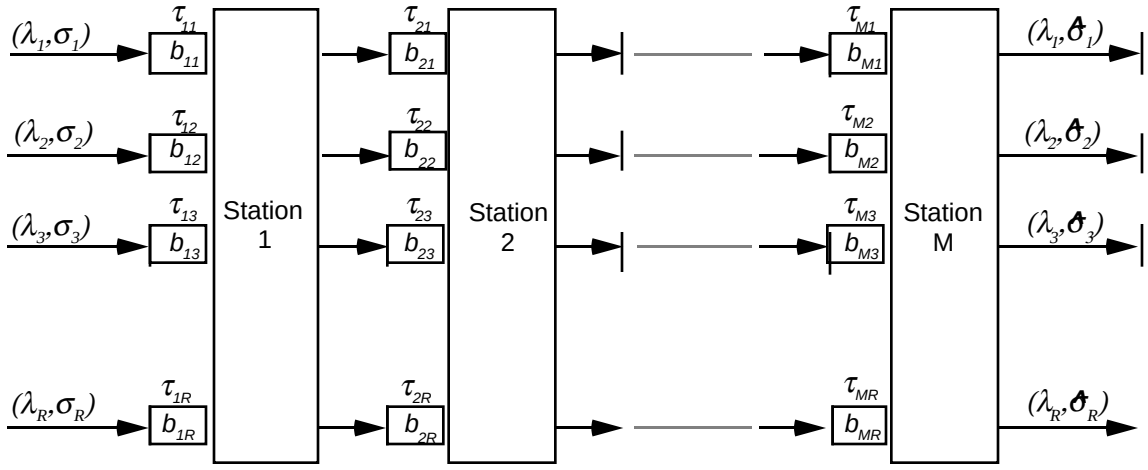

Figure 1: A tandem system.

There are M stations indexed by $1, 2, \ldots, M$. Each of the R routes, indexed by $1, 2, \ldots, R$ visits each station exactly once, in the order $1, 2, \ldots, M$. Let b_{ij} be the buffer holding parts of route type j, at the i-th stage of their route, and let τ_{ij} be the corresponding service time for such parts, which is to be provided by station i. As mentioned earlier, we suppose that traffic entering the system on route j is constrained by rate parameter λ_j and burst parameter σ_j. Note that the nominal load at station i is $\rho_i := \sum_{j=1}^{R} \lambda_j \tau_{ij}$.

We wish to determine the burst parameters $\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_R$, constraining the traffic exiting from the system on routes $1, \ldots, R$, for the rate parameter constraints $\lambda_1, \lambda_2, \ldots, \lambda_R$, respectively.

which n_α parts exit from each buffer b_α. Then,

$$n_\alpha \leq \lambda_{j(\alpha)} T + \sigma_\alpha .$$

From (1), the processing time consumed by these parts is

$$\sum_{\substack{\{\alpha : \text{buffer } b_\alpha \text{ is} \\ \text{served by} \\ \text{station } i\}}} n_\alpha \tau_\alpha \leq \rho_i T + \sum_{\substack{\{\alpha : \text{buffer } b_\alpha \text{ is} \\ \text{served by} \\ \text{station } i\}}} \tau_\alpha \sigma_\alpha .$$

Hence,

$$T \leq \rho_i T + \sum_{\substack{\{\alpha : \text{buffer } b_\alpha \text{ is} \\ \text{served by} \\ \text{station } i\}}} \tau_\alpha \sigma_\alpha .$$

If $\rho_i < 1$ for all stations, then this shows that T, and hence all busy periods, are bounded, proving (i).

For (ii), consider a part that arrives T seconds into a busy period at station i. The work that arrived prior to it in the busy period is bounded above by $\rho_i T + \sum_\alpha \tau_\alpha \sigma_{\alpha'}$, where $\sigma_{\alpha'}$ is the burstiness bound for the buffer preceding b_α, or the external arrivals. However, a work in the amount T is already completed, and so the remaining unfinished work, and hence the queueing delay under FCFS, is bounded by $\sum_\alpha \tau_\alpha \sigma_{\alpha'}$. By Little's Theorem, the number of parts in the system is also bounded. □

Definition. *The* nominal load ρ_i *at station* i *is defined by*

$$\rho_i := \sum_{\substack{\{\alpha : \text{buffer } b_\alpha \text{ is} \\ \text{served by} \\ \text{station } i\}}} \lambda_{j(\alpha)}\tau_\alpha \tag{1}$$

where $j(\alpha)$ *is the route along which buffer* b_α *lies.*

Let us say that the system is *stable* if the number of parts in the system is bounded over time. Clearly

$$\rho_i \leq 1 \text{ for } i = 1, \ldots, M,$$

is necessary for stability.

The following lemma (see [10,11]) shows that if all the flows in the system are appropriately constrained, then the system is stable.

Lemma 1: Bounded Burstiness Implies Stability. *Suppose the flow leaving every buffer* b_α *is constrained by rate parameter* $\lambda_{j(\alpha)}$ *and burst parameter* σ_α.

(i) If $\rho_i < 1$ *for all stations, then all busy periods (i.e., maximal time intervals in which a station is continuously busy) are bounded.*

(ii) If $\rho_i \leq 1$ *for all stations, then the number of parts in the system, and the delays of all parts, are bounded.*

Proof. Consider a subinterval of length T of a busy period, synchronized to begin with the arrival of a part initiating the busy period, and ending with the departure of any part, in

2 The System Model

We consider a queueing network consisting of M stations, indexed by $1, \ldots, M$. There are R types of parts, indexed by $1, \ldots, R$. Each part type has it own deterministic route. To each part type at each stage of its route, we associate a buffer where the corresponding parts are stored until their service is complete. We suppose that service times are deterministic, and allowed to depend on both the part type and stage, i.e., the corresponding buffers. Let τ_α be the service time for parts in buffer b_α. We exclusively consider the First Come First Serve (FCFS) scheduling policy, where each station always selects that part for processing which arrived first to one of its buffers, breaking ties arbitrarily. The service is non-preemptive.

The following notions of rate and burst parameters are introduced in [10,11].

Definition. *Let $f(t)$ denote the number of parts which pass some given point in the network during the time interval $[0, t)$. If*

$$f(t_1) - f(t_0) \leq \lambda(t_1 - t_0) + \sigma \text{ for all } 0 \leq t_0 \leq t_1 < +\infty,$$

then the flow of parts at this point is said to constrained by rate *parameter λ, and* burst *parameter σ.*

We shall suppose that traffic entering the system on route j is constrained only in that it has rate parameter λ_j and burst parameter σ_j.

Few positive results concerning the stability of the FCFS policy appear to be available, except under highly specific assumptions for stochastic networks under which the steady state probability distribution has a "product form;" see [6]. In [7] a fluid model approach is used to establish the stability under FCFS of a certain re-entrant line. In [8,9] it is shown that one can establish stability by finding a feasible solution to a certain linear program. Stability under FCFS of some particular systems has been established by such an approach.

In this paper we conduct an analysis of the FCFS scheduling policy. We adopt the framework and approach of a system model which allows for bursty arrivals, introduced by Cruz [10,11]. First we exhibit an independent result of interest in its own right. Given a tandem segment of a network, we obtain a sharp bound on the burstiness of traffic exiting the segment, in terms of the rate and burstiness of traffic entering the segment. This result allows us to identify some network topologies for which FCFS is stable for all arrival and service rates within the capacity of the system. In particular, we establish the stability under FCFS of bidirectional systems, as well as 2×2 re-entrant lines. Moreover, the burstiness bound also allows us to bound the performance of the FCFS scheduling policy. For an $M \times N$ re-entrant line, we show that by using N traffic smoothing regulators, one preceding each buffer of the first station, one can bound the number of parts, and thus also the delay, over the entire capacity region of the system, i.e., independently of system loading. This is of particular interest since it shows that by this method of control, the delay of parts is bounded in heavy traffic, when all stations are 100% utilized.

The FCFS Service Discipline: Stable Network Topologies, Bounds on Traffic Burstiness and Delay, and Control by Regulators*†

Gil I. Winograd and P. R. Kumar‡

Abstract

We consider the well known First Come First Serve (FCFS) scheduling policy under bursty arrivals. Recently it has been shown that the FCFS policy can be unstable for some non-acyclic network topologies.

We identify some network topologies under which FCFS is stable for all arrival and service rate vectors within the system's capacity. This is done by determining a sharp bound on the burstiness of traffic exiting from a tandem section of the system, in terms of the burstiness of the incoming traffic. This burstiness bound further allows us to provide a bound on the maximum number of parts in the system, and the maximum delay. It also enables us to analyze the performance of some systems controlled by the use of traffic smoothing regulators. The maximum delay can remain bounded even in the heavy traffic limit, when all stations are 100% utilized.

1 Introduction

Queueing networks can be unstable under a certain scheduling policy, even if the nominal load on each server is within the server's capacity; see [1,2,3]. It has been recently shown in [4,5] that this instability can also occur under the well known First Come First Serve (FCFS) scheduling policy.

*Please address all correspondence to the second author at the address below.

†The research reported here has been supported in part by the National Science Foundation under Grant Nos. NSF–ECS–90–25007 and NSF–ECS–92–16487, and by the Joint Services Electronics Program under Contract No. N00014–90–J1270.

‡Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory, University of Illinois, 1308 West Main Street, Urbana, IL 61801.