BLQF: A Prioritized Class of Maximum Weight Matching Algorithms which guarantee 100% throughput

Kevin Ross

June 6, 2002

1 Introduction

In this project I consider the maximum weight matching algorithms developed by McKeown et al [1]. I will show three significant generalizations from the original algorithm:

- A generalization of the algorithm
- A generalization of the arrival traces allowed
- A generalization of the service configuration set available

The results here have significance in theory and in practice. I will demonstrate that the crossbar switch is an example of a larger class of scheduling problems, and the algorithms I propose for these problems will allow users of a switch to prioritize input-output pairs in useful and interesting ways.

I will present these results as follows: First I will define the model and discuss how the switching problem fits into this larger class of scheduling problems. I will briefly outline how stability is defined for these problems. Secondly, I will introduce the algorithm BLQF and state the key results of this project - throughput guarantee for a set of matrices B. I will show results for bernoulli iid traffic and general traffic.

Following the introduction of model and algirithm, I will show the geometric structure of this class of algorithms, and how this leads to intuition in the proof of stability and scalability of the algorithm.

2 Model Definition

Switches can be seen as a special case of queueing systems with the following properties:

- Long term unknown arrival rates to each of a set of queues
- Multiple service configurations possible at any time

Consider a system comprised of indexed queues $q \in \mathcal{Q} = \{1, 2, ..., Q\}$. Arrivals to these queues may be distributed in any way, dependent or independent with the only requirement being the existence of a long-term average arrival rate to each queue. For each queue q, the instantaneous arrival rate at time t is given by $A_q(t)$ and it satisfies the time average condition:

$$\lim_{t \to \infty} \frac{\int_0^t A_q(s)ds}{t} = \rho_q < \infty \tag{1}$$

Note that the only condition restricting arrivals is the integrability of the instantaneous arrival rate. This allows for a very general class of arrival processes. Within this class is continuous arrivals at a constant or variable rate, discrete arrivals of finite but arbitrary size arriving at distinct times and any combination of these two.

In particular if finite jobs arrive to the system in continuous time then for each job the instantaneous arrival process is represented as a $\delta-$function of magnitude the size of its workload. Between two finite times, a finite amount of work may arrive and a finite number of $\delta-$ jumps or job arrivals can occur.

At any time, the system can be in a single service configuration. The indexed set of available configurations is $\mathcal{S} = \{S^m\}_{m=1}^M$. These configurations are defined as Q-vectors, with each component corresponding to the rate at which a nonempty queue will be served when the system is in that configuration. For example, if $Q = 2$, the service configuration $[1, 2]$ refers to serving queue 1 at rate 1 and queue 2 at rate 2. In the case of a 2×2 packet switch the configurations available are $[1, 0, 0, 1]$ and $[0, 1, 1, 0]$ with the standard definitions of the queues. We could incorporate speedup in this model by including configurations such as $[2, 0, 0, 2]$. Since configurations are defined by vectors I use the terms service configuration and service vector interchangeably.

Arriving jobs are buffered immediately in their respective queues, and served $FCFS$ within that queue. The decision concern for design is the scheduling of service configurations, based on the current state of the system and without prior information on the long term arrival rates.

I use vector notation throughout, referring to subscripts only when necessary. The system at time t is fully defined by the lengths of all of its queues at that time, $X(t)$, the arrival rates to each queue $A(t)$ and the current service configuration $S(t)$.

I make one additional assumption on the set of available service configurations.

Property 1 *For the set of feasible service configurations $\mathcal{S}$, if $S^m \in \mathcal{S}$ and q' is any input queue then $S^{m'} \in \mathcal{S}$, where*

$$S_q^{m'} = \begin{cases} 0 & q = q' \\ S_q^m & q \neq q' \end{cases} \tag{2}$$

This is equivalent to saying that it is possible to *turn off* service at an individual queue. If the system can serve a set of vectors it can serve any subset of those vectors as well. This is seen in a crossbar switch, for example, since for any service configuration it is possible to disconnect any individual input without affecting the service provided to the others.

In the case of packet switches, the queues are input-output pairs, and the service configurations are the set of permutatoins with each input and output holding at most one connection.

2.1 Stability Region Definition

On a high level, a system is considered to be stable if it is possible to serve all incoming traffic. Bambos and Walrand [5] talk about rate stability, the property that input rate is equal to output rate for all queues in the system.

Specifically, consider a queueing system as described above. An arrival trace defined by a Q- vector ρ is said to be stable if there is a convex combination of the service vectors which serves each queue at at least its arrival rate. More formally,

Definition 1 *The stability region $\mathcal{R}$ is given by*

$$\mathcal{R} = \{\rho \in \Re_+^Q : \rho \leq \sum_{m=1}^M x^m S^m, \text{for some } x \geq 0, \sum_{m=1}^M x^m \leq 1\} \tag{3}$$

Equivalently, $\mathcal{R}$ is the convex hull of the available service configurations.

The stability region can be defined in terms of inner products and has some interesting consequences. This is discussed in Appendix B.

3 The BLQF Algorithm

In switching problems, the LQF algorithm chooses S^m at each timeslot which maximizes the vector product $\langle S^m, X(t)\rangle$. I will expand such algorithms to include those which maximize $\langle S^m, BX(t)\rangle$, where B is a matrix. I refer to this as the BLQF policy. Formally, the BLQF policy is defined here.

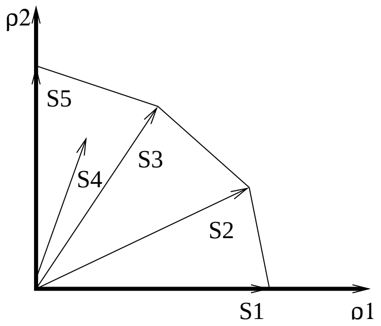

Figure 1: The stability region is defined by the convex hull of the service configurations. S^1, S^2, S^3 and S^5 are extreme configurations, while S^4 is not.

Definition 2 *At time t with workload $X(t)$ in the system, the **BLQF** policy chooses a service configuration S^* which satisfies the equation:*

$$\langle S^*, BX(t)\rangle = \max_{S^m \in \mathcal{S}} \langle S^m, BX(t)\rangle \tag{4}$$

I show that throughput is maximized for all such systems where B satisfies the following properties:

Property 2 *of B:*

a B is positive definite

b B has nonpositive off diagonal elements

c B is symmetric

A direct consequence of the first and second properties is that B has strictly positive diagonal elements.

I show that the combination of these properties is sufficient for stability for any service configuration set and stable arrival process. I show that properties a and b are necessary.

I make one further restriction on the policy.

Definition 3 *A service vector S^m is **minimal** with respect to the workload if there is no q for which $(BX)_q = 0$ and $S^m_q > 0$.*

That is, the if the service vector chosen is minimal it will not choose to serve a queue if serving it will not strictly increase the inner product. Such a policy is possible due to the definition of the service set in property 1.[1]

Given the above model I now state the theorems as the key results.

Theorem 1 *For admissible iid bernoulli arrivals to a timeslotted crossbar packet switch, the BLQF algorithm is stable for any diagonal matrix B with strictly positive elements.*

Theorem 2 *For any admissible arrival processes , the policy of choosing a minimal $S^m \in \mathcal{S}$ which maximizes the inner product $\langle S, BX(t)\rangle$ at each time t is rate stable for all B satisfying parts a,b and c of property 2.*

Theorem 3 *For a general set of service configuration vectors, property a of property 2 is* necessary *for rate stability.*

Theorem 4 *For a general set of service configuration vectors, property b of property 2 is* necessary *for rate stability.*

Theorem 1 is proved in appendix A, theorem 2 is proved in appendices B and C, and theorems 3 and 4 are proved in appendix D.

[1] I note at this point that the choice I make of having a minimal service vector is for the sake of intuition and proof, not a necessary condition for stability. If a non-minimal service vector is chosen, then the service applied is at least that of the corresponding minimal policy. Thus, while I assume the minimal policy is used, all the results also follow for a non-minimal policy.

4 Geometric interpretation of policies

The nature of these policies can best be seen by considering them geometrically. The matrix B transforms the workload vector in a way that preserves three key properties: First, it continues to give increasing priority to queues whose workload is increasing; second, the priority of these queues decreases as other queues increase in size; finally, the diagonal dominance of the matrix ensures that if one queue increases relative to the other queues, the maximum possible service is applied to that queue.

For each service vector S^m, there is a set of workload vectors for which that service is chosen. The set is a cone because inner product ordering is preserved under scalar multiplication. $\langle S^1, BX\rangle \leq \langle S^2, BX\rangle \Leftrightarrow \langle S^1, B\alpha X\rangle \leq \langle S^2, B\alpha X\rangle$ for any positive scalar α. I denote the cones:

$$C^m = \{X \in \Re_+^Q : \langle S^m, BX\rangle = \max_{m' \in M} \langle S^{m'}, BX\rangle\} \tag{5}$$

for each m. This is the set of workload vectors with inner product maximized by S^m. Note that the workload spaced can be divided into up to M cones, one for each service vector. Cones may share a common boundary. For a particular workload vector I define the set of service vectors which could be chosen by the BLQF policy to be

$$\mu(X) = \{m \in 1, 2, ..., M : \langle S^m, BX\rangle = \max_{m' \in M} \langle S^{m'}, BX\rangle\} \tag{6}$$

This is the index set of service vectors with maximal inner product, and is often a single service vector. For the case with multiple such cones, I can describe the workload-centric cone

$$C(X) = \cup_{m \in \mu(X)} C^m \tag{7}$$

Since the simplest switch is a 2 by 2 switch, it is represented by four queues. This is not easy to illustrate, so instead I provide an example of a two-queue system. This provides the necessary intuition to recognize the geometric nature of the policies.

4.1 Example: A two-queue system

For illustrative purposes I will demonstrate the properties of these policies first through a system with two queues. Assume there are three service configurations, $S^1 = [4, 0], S^2 = [3, 2], S^3 = [0, 3]$. The feasibility region of arrival rates is given by the convex hull of these service configurations, as shown in figure 2. Next to the stability region, if the identity matrix is used for B, the cones are also illustrated.

I consider what happens to the cones if B is different from I. First I consider having only diagonal elements in B. Let B be the matrix [2 0;0 1]. Then the service cones are shifted, giving the cones seen in the left of figure 3.

Next I consider the effect of off-diagonal elements. I consider the matrix $B = [1, -0.5; -0.5, 1]$. The right-hand plot of figure 3 illustrates this case. The off-diagonal elements tend to lean the system to more often serve less queues. This has the effect of making the inner cone smaller and the outer cones cover a larger fraction of the area.

4.2 Scalability of Policies

The key criticism of LQF policies in switching is that they are too slow for practical implementation. Randomized implementations of these policies are commonly proposed to avoid such problems. [4]

The geometric structure of these policies leads to key intuition in this respect. If there is some bound on the amount of instantaneous arrivals to the system, then this is equivalent to a bound on the size of jump in the workload space. Recall that the cones are linear. Thus, as workload increases, the distance between cones increases, and a bound on the workload jump equates to a restriction that the workload will only jump

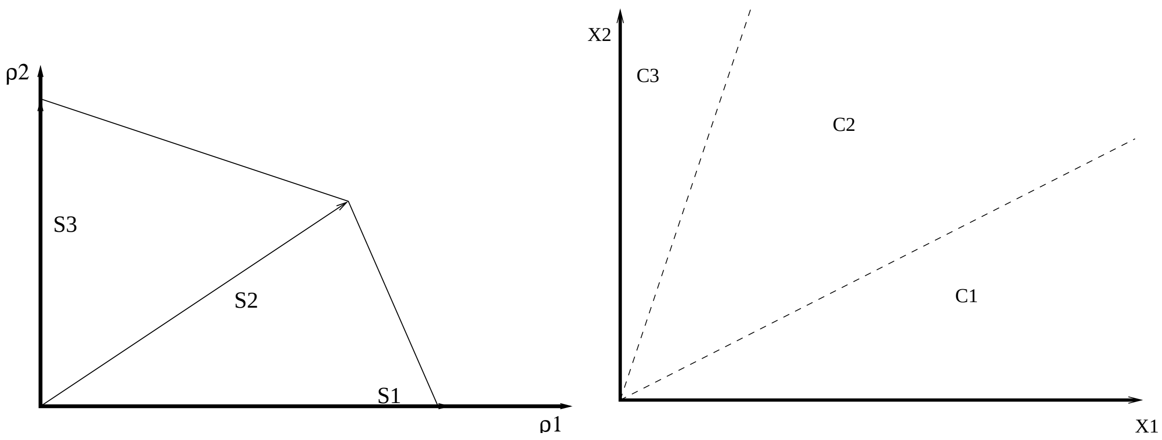

Figure 2: The stability region for our 2-queue example, and the associated cone structure of the workload space with identity B. For example, if the workload at time t is within the cone C^1 then service configuration 1 will be used.

to one of its neighboring cones. In other words, it is not necessary to check all M service configurations and corresponding inner products, only to check the cones which neighbor the current service cone.

All that matters for a the choice of service configuration is which cone the workload lies inside. For implementation, therefore, the cones could be simply considered as colors, and the workload at each time needs only to check which color cone it is inside. If it knows where it was previously, it need only check the cones which neighbor the previous cone. The system can store a list of neighbors for each service configuration, and as the workload varies the service will switch as soon as a new cone is entered. This scalablity is illustrated in figure 4.2.

5 Quality of Service Comparisons

As can be seen in the previous section, the matrix B in the above algorithms has the effect of weighting the queues, giving them different priorities. The strictly diagonal matrix gives a simple prioritizing. The queues assigned larger weights will be served with greater priority.

Geometrically, with the workload space $\Re_+^Q$ divided into cones by the available service configurations, the cones are expanded and contracted by adjusting the weight assigned to a particular queue.

The off-diagonal elements have a more subtle effect on the algorithm. Rather than expanding a cone around a specific queue, they shift the boundaries between cones. For example, if a matrix B has a negative element B_{qr}, the algorithm gives *less* importance to queue q if queue r is nonempty. That is, queue q could be considered important except when queue r is large. This could be because of customer perception or complimentary service. For example, one customer may be less dissastisfied with long waiting times if they can see that another customer has similar or greater waiting time. In another case, queues may be divisible into different classes. It may be more important that *some* queue within a class is served, but give little benefit if more than one is served. This allows for a whole class of entangled quality of service based algorithms since the prioritizing of queues relates directly to the quality of service that will be given to the queues.

There is a nice system dynamic at work here. The basic principle is simple: the priority of a particular queue is increasing in its own size and decreasing in the size of every other queue. Service attention will shift toward a queue as its own workload grows, and away as the workload of other queues grows.

The properties of the algorithm are most clearly seen when the variance of workload arrivals is high. In the simulations presented here, a timeslotted system with bernoulli iid arrivals was used.

The simulations demonstrate:

- For a diagonal matrix, a high diagonal element corresponds to a high priority on that queue, and

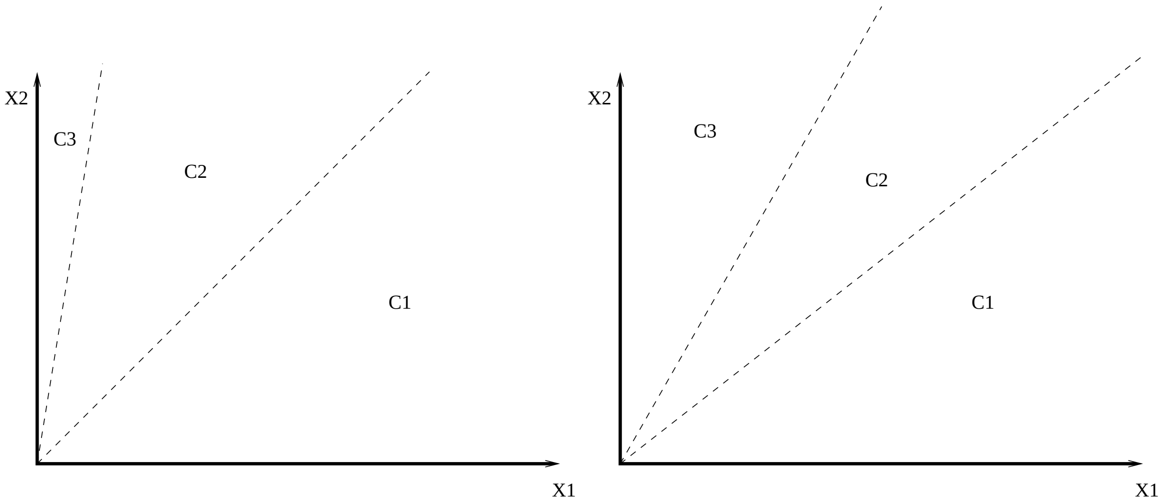

Figure 3: The workload cones for our 2-queue example with diagonal matrix weights and then with off-diagonal weights. The diagonal matrix shifts all of the configuration cone boundaries toward X_2. This corresponds to a greater portion of service being assigned to X_1. The non-diagonal matrix sees the inner cone shrink. This corresponds to the system pulling workload lengths closer together. If the first cone is significantly longer than the 2nd cone, the first will be served.

consequently a lower average queue length

- For a symmetric non-diagonal matrix, the off diagonal elements have the effect of *coupling* the workload in two queues. This results in the third queue having a higher priority, but the other two being closely correlated.

I ran the simulations on the same arrival trace for 100,000 timeslots. For each of three runs, I adjusted the matrix B in the algorithm, and recorded the buffer requirements for each queue. The simulation was on a 4×4 crossbar switch with no speedup. The arrival rate was 99% of the maximum throughput with uniform arrival rates.

Figure 5 shows the trace of queue sizes for the switch. I provide a trace of the first queue from input-output pair (1,1), input-output pair (1,2) and the average of the remaining 14 queues. With identity B the algorithm is identical to the LQF algorithm. The first plot shows the trace for this. When the first component of B is doubled the (1,1) queue is given higher priority. This is shown in the second plot. The third plot shows the effect of adding negative off-diagonal elements. The 1st and 2nd queues are given lower priority but their sizes become coupled.

Figure 5 shows the average buffer size for each of the three runs. This shows more concretely the overall affect of varying the parameters within the matrix B.

6 Conclusions

I have presented the packet-switching problem as a special case of a much broader class of scheduling problems. I have shown the stability of the BLQF algorithm for this whole class. This proof has much weaker assumptions than were required for the Lyapunov-based proof we saw earlier in class.

This has a wide number of applications. I have focussed here on the crossbar swithc but note that the model is very general. The application of this research to multicast switches is of great interest.

I have demonstrated the effect of changing the parameter matrix B to influence quality of service. While no quantatative results can be shown with such general arrival trace restrictions, qualitative analysis shows that these parameters allow a great deal of control to be given to the user.

This research is ongoing, and will appear in part as a technical report currently being filed in Netlab, Stanford University.

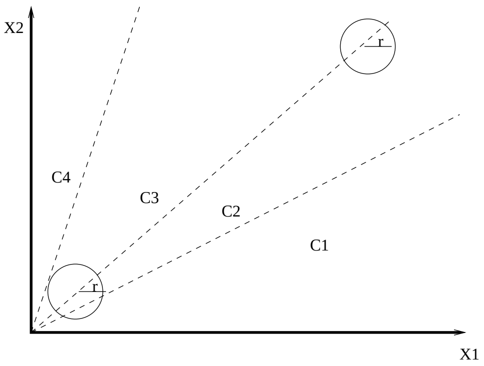

Figure 4: As the size of the workload increases, a fixed jump limit restricts the cone movement to nearest neighbors

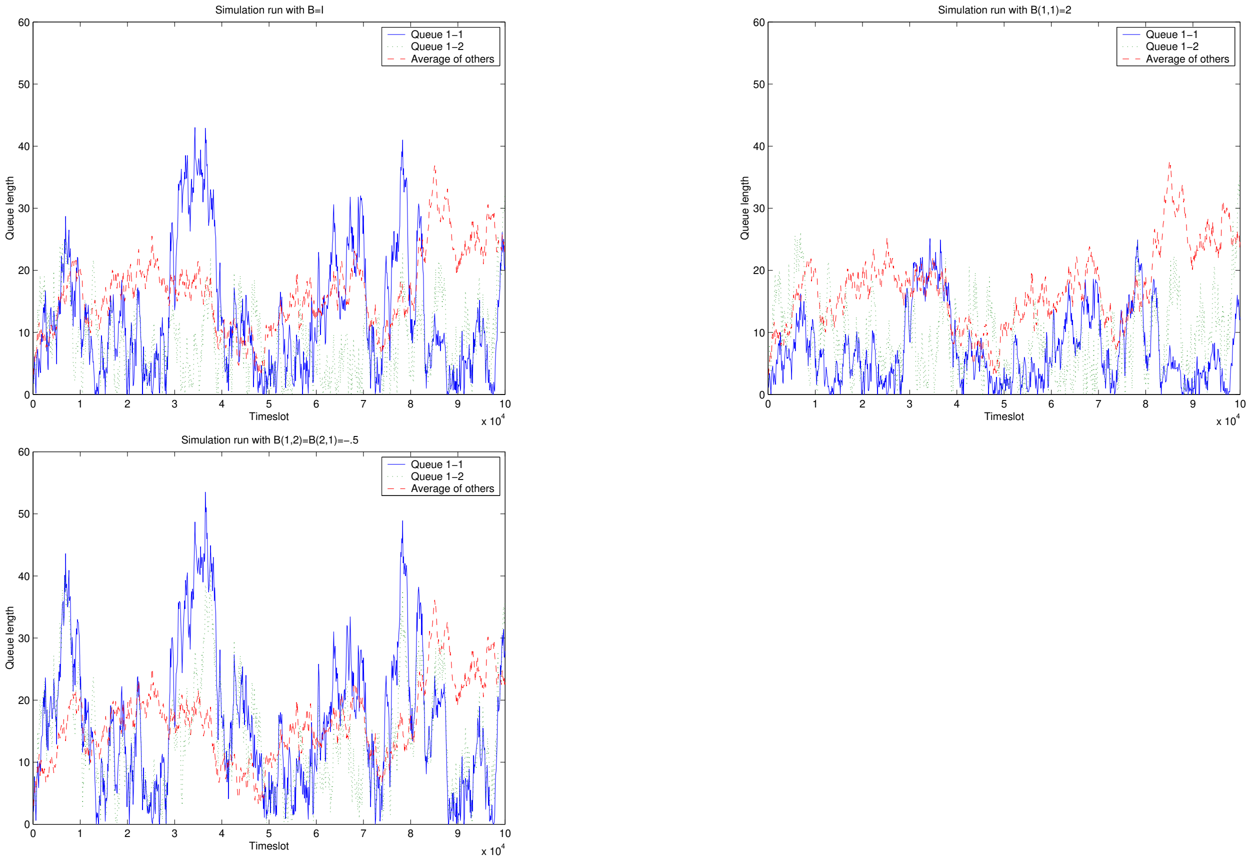

Figure 5: The trace for a 4×4 switch is given with three runs of the simulation. Each run was done on matlab using the same input trace and 100,000 timeslots. The first plot shows the standard LQF algorithm. The 2nd plot gives double priority to the first queue. The third plot gives lower priority to queues 1 and 2 and entangles their traces. The first queue refers to the input-output pair (1,1), the second to i-o pair (1,2).

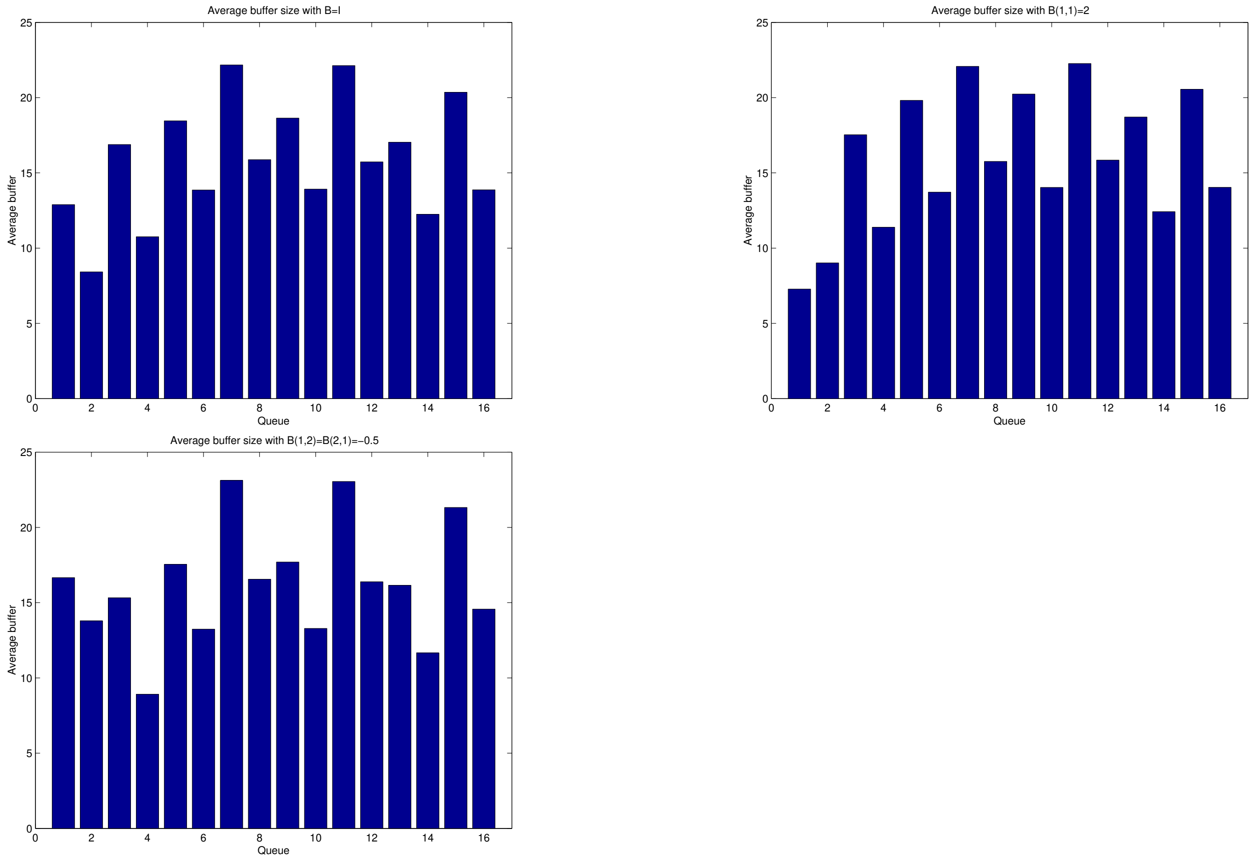

Figure 6: The average buffer size for each queue. This shows the average buffer for the same input traces as shown previously. It can be seen that the 2nd run reduces the average backlog in queue 1, and the third increases the average backlog for both queue 1 and queue 2.

A Proof of stability for positive diagonal B with bernoulli iid arrivals

This proof will follow a similar line to that in McKeown's paper [1]. I use a quadratic Lyapunov function $V(L(n))$ and show that $E[(V(L(n+1)) - V(L(n))|L(n) \ \leq -\epsilon\|L(n)\| + k$, where $k, \epsilon > 0$. According to the argument of Kumar and Meyn [6], it follows that the switch is stable. I will here use the Lyapunov function $V(L(n)) = \langle L(n), BL(n)\rangle$. Since B is positive definite, this defines a norm and the stability of this implies the stability of $L(n)$.

Throughout the proof I will assume that the $MN \times MN$ matrix B satisfies the assumptions given in the statement of the theorem.

First, I refer to Birkhoff's theorem: The doubly stochastic square matrices form a convex set C with the set of extreme points equal to the permutation matrices S. This is proved in [7].

McKeown noted and proved further that the doubly substochastic $M \times N$ nonsquare matrices form a convex set with the set of extreme points equal to the quasi-permutation matrices S. This leads to

Lemma 1 $L'(n)B(\lambda - S^*(n)) \leq 0, \forall L(n), \lambda$, *where* $S^*(n) = \arg\max_{S\in\mathcal{S}}\langle X(n), BS\rangle$, *the service vector selected by the maximum weight matching algorithm to maximize* $\langle X(n), BS\rangle$.

Note that this also follows from the alternative definition of stability given in the next proof. I are using minimal policies.

Lemma 2

$$E[L'(n+1)BL(n+1) - L'(n)BL(n)|L(n)] \leq NM\|B\| \quad \forall\lambda \tag{8}$$

Proof

$$\begin{aligned} L'(n)BL(n) - L'(n)BL(n) &= (L(n) - S(n) + A(n))'B(L(n) - S(n) + A(n)) - L'(n)BL(n) \\ &= L'(n)B(A(n) - S(n)) + (S(n) - A(n))'B(S(n) - A(n)) \\ &= 2L'(n)B(A(n) - S(n)) + (A(n) - S(n))'B(A(n) - S(n)) \end{aligned} \tag{9}$$

The above uses the symmetry of the matrix B. $(A(n) - S(n))B(A(n) - S(n)) \leq \|A(n) - S(n)\| \cdot \|B\| \cdot \|A(n) - S(n)\| \leq \sqrt{NM} \cdot \|B\| \cdot \sqrt{NM}$ since each component of $A(n) - S(n)$ has absolute value at most 1.

Taking expected value:

$$\begin{aligned} E[L'(n+1)BL(n+1) - L'(n)BL(n)|L(n)] &\leq E[2L'(n)B(A(n) - S(n))] + NM\|B\| \\ &= 2L'(n)B(\lambda - S^*(n)) + NM\|B\| \\ &\leq NM\|B\| \end{aligned}$$

using lemma 1.

Lemma 3 $\forall\lambda \leq (1-\beta)\lambda_m, 0 < \beta < 1$ *where* λ_m *is any rate vector such that* $\|\lambda_m\|^2 = \min(N, M)$, *there exists* $\epsilon > 0$ *such that*

$$E[\tilde{L}'(n+1)B\tilde{L}(n+1) - L'(n)BL(n)|L(n)] \leq -\epsilon\|L(n)\| + NM\|B\| \tag{10}$$

Proof From the previous lemma, it is sufficient to find $\epsilon > 0$ s.t. $2L'(n)(\lambda - S^*(n)) \leq -\epsilon\|L(n)\|$.

Let θ be the angle between $L(n)$ and $[B\lambda_m]$. I first show that $\cos\theta > 0$. Note that $L(n)$ and $B\lambda_m$ are both nonnegative vectors, since all components are nonnegative. If $\cos\theta = 0$ then $L(n)$ and $B\lambda_m$ are orthogonal. If $X_{ij}(n) > 0$ for some (i,j) then there have been arrivals to queue q_{ij}. This can only happen if $\lambda_{ij} > 0$, which implies $[B\lambda]_{ij} > 0$ since $B_{ij,ij} > 0$. Therefore $\cos\theta > 0$ unless $L(n) = 0$.

Now I bind θ strictly away from 0. I know that $\|\lambda\|^2 < \sqrt{NM}$. Let $[\lambda B]_{min} = \min\{[B\lambda]_{ij}, 1 \leq i \leq M, 1 \leq j \leq N\}$, and $X_{max}(n) = \max\{X_{ij}(n), 1 \leq i \leq M, 1 \leq j \leq N\}$. Then $\|L(n)\| \leq [NMX^2_{max}(n)]^{1/2} = \sqrt{NM}X_{max}(n)$. Also, $\|B\lambda_m\| \leq \|B\lambda\| \geq \sqrt{NM}[\lambda B]_{min}/(1-\beta)$. Now I have

$$
\begin{aligned}
cos\theta &= \frac{L'(n)B\lambda}{||L(n)||||B\lambda||} \\
&\geq \frac{X_{max}(n)[\lambda B]_{min}}{||L(n)||||B||(NM)^{1/4}} \\
&\geq \frac{X_{max}(n)[\lambda B]_{min}}{\sqrt{NM}X_y max(n)||B||(NM)^{1/4}} \\
&= \frac{[\lambda B]_{min}}{||B||(NM)^{3/4}}
\end{aligned}
$$

Therefore I have:

$$
\begin{aligned}
L'(n)B(\lambda - S^*(n)) &\leq L'(n)B(\lambda_m - S^*(n)) - L'(n)B(\beta\lambda_m) \\
&\leq 0 - \beta||L(n)|| \cdot ||B\lambda_m||cos\theta \\
&\leq -\frac{\beta||L(n)|| \cdot ||B\lambda||[\lambda B]_{min}}{||B||(NM)^{3/4}} \\
&\leq -\frac{\beta\sqrt{NM}[\lambda B]^2_{min}}{||B||(NM)^{3/4}}||L(n)|| \\
&= -\epsilon||(L(n)||
\end{aligned}
$$

where $\epsilon = \frac{\beta[\lambda B]^2_{min}}{||B||(NM)^{1/4}} > 0$.
Now setting $k = NM||B||$ I have proved theorem 1

B Preliminary results for rate stability

In this section I state and prove some preliminary results needed in the proof of stability give in the following section.

First we consider the stability region definition: Recalling that the inner product is the length of a projection onto the unit vector in that direction, this says that there exists a service configuration which gives a greater projection in that direction than ρ.

Thus I work with the definition:

Definition 4 *The stability region $\mathcal{R}$ is defined by*

$$
\mathcal{R} = \{\rho \in \Re^Q_+ : \forall v \in \Re^Q \exists S^m, \langle \rho, v \rangle \leq \langle S^m, v \rangle \tag{11}
$$

Recalling that the inner product is the length of a projection onto the unit vector in that direction, this says that there exists a service configuration which gives a greater projection in that direction than ρ.

The equivalence of definitions 1 and definition 4 can be seen geometrically and it is worth devoting some discussion to this here. First I define formally the set of extreme configurations, and state equivalence relation.

Definition 5 *A service configuration is extreme if the stability region defined in definition 1 would be strictly smaller if configuration S^m was removed.*

Theorem 5 *Definitions 1 and 4 describe the same stability region $\mathcal{R}$*

Proof of theorem 5 It is sufficient to show the two following properties:

- If $\rho \in \mathcal{R}$ in definition 1 then $\rho \in \mathcal{R}$ in definition 2.
- If $\rho \notin \mathcal{R}$ in definition 1 then $\rho \notin \mathcal{R}$ in definition 2.

First I prove the first assertion. If ρ is stable, then consider any vector $v \in \Re^Q$. From definition 1 there exists an x with $\rho = \sum_{m=1}^{M} x^m S_m, \sum_{m=1}^{M} x^m = 1, x \geq 0$. Also, from property 1 I know that any component of a service vector can be set to zero.

Consider any vector $v \in \Re^Q$. Let $P(v) = q \in \mathcal{Q} : v_q < 0$. For $q \in P(v)$, I set $S_q^m = 0$ for all m. That is, I replace all S^m configurations with $S_q^{m'}$ as definited by equation 2. This leads to a subset of service configurations (with repeated vectors) where $S_q = 0$ whenever $v_q < 0$.

Now for each $q \in P(v), \rho_q v_q \leq 0 = \sum_{m=1}^{M} x^m \hat{S}_m^q \rho_q$. For all other $q, \rho_q v_q = \sum_{m=1}^{M} x^m S_m^q \rho_q = \sum m = 1^M x^m \hat{S}_m^q \rho_q$ since these components of S^m have not been changed. Then I have for each q in $\mathcal{Q}$, $\rho^q v^q \leq \sum_{m=1}^{M} x^m \hat{S}_m^q = \sum_{m=1}^{M} \hat{x}_m S_m^q$, where $\hat{x}$ is definied by regrouping the S^m vectors. Consequently,

$$\langle \rho, v \rangle \leq \sum_{m=1}^{M} \hat{x}_m \langle S_m, v \rangle \leq \max_{S^m \in \mathcal{S}} \langle S^m, v \rangle \tag{12}$$

which proves the assertion.

Now I prove the second result.

Assume that ρ is not containted in $\mathcal{R}$ from definition 1.

Let p be a solution to $p = \arg\min_{p \in \mathcal{R}} \langle \rho - p, \rho - p \rangle$. Here p refers to the *nearest* feasible arrival vector to the ρ and by our assumption the inner product is strictly positive.

Let $v = \rho - p$. Let $S^* = \max_{S^m \in \mathcal{S}} \langle S^m, v \rangle$. It suffices to show $\langle S^*, v \rangle < \langle \rho, v \rangle$.

Since $p \in \mathcal{R}$, I have $p = \sum_{m=1}^{M} x^m S^m$ for some x. Hence $\langle v, p \rangle = \sum_{m=1}^{M} x^m \langle v, S^m \rangle \leq \langle v, S^* \rangle$.

Therefore

$$\langle S^* - p, v \rangle \geq 0 \tag{13}$$

The proof procedes by contradiction. Assume that the assertion is false. That is, assume

$$\langle S^* - \rho, v \rangle \geq 0 \tag{14}$$

Then if equality holds in equation 13 then $\langle S^* - \rho, v \rangle = \langle p - \rho, v \rangle = -\langle v, v \rangle < 0$ which would bring an immediate contradiction to equation 14. Hence I assume the inequality in 13 is strict.

Now p and S^* are both in $\mathcal{R}$. Since $\mathcal{R}$ is convex, $p_\epsilon = p + \epsilon(S^* - p) \in \mathcal{R}$ for all $\epsilon \in [0, 1]$.

$$\begin{aligned} \langle \rho - p, \rho - p \rangle - \langle \rho - p_\epsilon, \rho - p_\epsilon \rangle &= 2\epsilon \langle S^* - p, \rho - p \rangle - \epsilon^2 \langle S^* - p, S^* - p \rangle \\ &= \epsilon(2\langle S^* - p, v \rangle - \epsilon \langle S^* - p, S^* - p \rangle \end{aligned} \tag{15}$$

Since both terms in equation 15 are strictly positive, a sufficiently small ϵ can be chosen to give $\langle \rho - p, \rho - p \rangle > \langle \rho - p_\epsilon$, which contradicts the definition of p as the nearest feasible point to ρ.

Hence the assertion is proved by contradiction.

Lemma 4 *Let $A(\cdot)$ and $S(\cdot)$ be the arrival and service loads generated by an arbitrary traffic trace which satisfies equation 1 and assume the BLQF policy is used. Then letting $X(t) = A(t) - S(t)$ be the workload in the system at time t, if $\lim_{t \to \infty} \frac{X_q(t)}{t} = 0$ then $\rho_q^{in} = \lim_{t \to \infty} \rho_q^{out}(t)$.*

That is, the long term departure rate is equal to the long term arrival rate to the system.

Proof Since I are dealing with minimal policies, lemma 7 implies that

$$\frac{X(t)}{t} = \frac{\int_0^t A(s) ds}{t} - \frac{\int_0^t S(s) ds}{t} \tag{16}$$

Letting $t \to \infty$ in equeation 16 and rearranging terms I have

$$\lim_{t\to\infty} \frac{S_q(t)}{t} = \rho_q \tag{17}$$

for all $q \in \mathcal{Q}$. Since service is never to idle queues, this establishes that the output rate is equal to the input rate.

Lemma 5 *Consider two increasing, unbounded time sequences* $\{t_n\}_{n=1}^{\infty}$ *and* $\{s_n\}_{n=1}^{\infty}$. *If* $\lim_{n\to\infty} \frac{t_n - s_n}{t_n} = 0$ *or equivalently* $\lim_{n\to\infty} \frac{s_n}{t_n} = 1$, *then*

$$\lim_{n\to\infty} \frac{X(t_n) - X(s_n)}{t_n} = \lim_{n\to\infty} \frac{X(t_n) - X(s_n)}{s_n} = 0 \tag{18}$$

Proof For every policy and for all $m \in 1, 2, ..., M, q \in \mathcal{Q}$, I have $\int_{s_n}^{t_n} I_{\{S(t)=S^m\}} dt \leq t_n - s_n$. Dividing by t_n gives $\lim_{n\to\infty} \frac{\int_{s_n}^{t_n} I_{\{S(t)=S^m\}} dt}{t_n} = 0$. Recalling that $X_q(t_n) - X_q(s_n) = \int_{s_n}^{t_n} A(t)dt - \sum_{m=1}^{M} S_q^m \int_{s_n}^{t_n} I_{\{S(t)=S^m\}} dt$, dividing by t_n and letting $n \to \infty$ I get

$$\lim_{n\to\infty} \frac{X_q(t_n) - X_q(s_n)}{t_n} = \lim_{n\to\infty} \frac{\int_{s_n}^{t_n} A(t)dt}{t_n} - \lim_{n\to\infty} \frac{\sum_{m=1}^{M} S_q^m \int_{s_n}^{t_n} I_{\{S(t)=S^m\}} dt}{t_n} = 0 \tag{19}$$

Moreover, $\lim_{n\to\infty} \frac{X_q(t_n) - X_q(s_n)}{s_n} = \lim_{n\to\infty} \frac{X_q(t_n) - X_q(s_n)}{t_n} . \frac{t_n}{s_n} = 0$.

Lemma 6 *Consider an increasing unbounded time sequence* $\{t_n\}_{n=1}^{\infty}$.

$$\lim_{n\to\infty} \frac{X(t_n) - X(t_n)^-}{t_n} = 0 \tag{20}$$

Proof Clearly the result holds when arrivals to the system are continuous. Thus I need to consider the case of instantaneous arrivals which shift the workload by an increasing linear amount. Let t_n be the time of arrival to queue $q \in \mathcal{Q}$, j_n the index of that job and $\alpha_{j_n}^q$ the workload added by the job. Then it is sufficient to show that $\lim_{n\to\infty} \frac{\alpha_{j_n}^q}{t_n} = 0$.

$$\alpha_{j_n}^q = \int_0^{t_n^+} A(t)dt - \int_0^{t_n^-} A(t)dt \tag{21}$$

Dividing equation 21 by t_n and letting $n \to \infty$, I have $\lim_{n\to\infty} \frac{\alpha_{j_n}^q}{t_n} = \rho_q - \rho_q = 0$.

Lemma 7 *If at any time the workload on a queue q is zero, and B satisfies parts a, b and c of property 2 then the minimal BLQF policy sets $S_q = 0$.*

Proof Consider the alternative. Let S^{m*} maximize the inner product $\langle S^m, BX \rangle$. Since the off diagonal elements of B are nonpositive, if $X_{q'} = 0$ then $[BX]_q \leq 0$. If the inequality is strict, then for optimality $S_q^{m*} = 0$. If it is an equality then $S_q^{m*} = 0$ since it is minimal.

C Rate Stability Sufficiency

The proof of this section follows a similar thread to that developed by by Armony and Bambos in [2]. I include details for completeness.

The objective of the proof is to show that $\lim_{t\to\infty} \frac{X_q(t)}{t} = 0$ for each $q \in \mathcal{Q}$ if $\rho \in \mathcal{R}$. This is sufficient for rate stability, as established in lemma 4. Since B is a positive definite matrix, it is sufficient to show that $\lim_{t\to\infty}\langle \frac{X(t)}{t}, B\frac{X(t)}{t}\rangle = 0$.

The proof proceeds by contradiction. Assume that $\limsup_{t\to\infty}\langle \frac{X(t)}{t}, B\frac{X(t)}{t}\rangle > 0$ and let $\{t_a\}_{a=1}^{\infty}$ be an increasing unbounded time sequence on which the supremum limit is obtained and η the corresponding limit of $\frac{X(t)}{t}$. Such a sequence exists by the compactness of the set of possible values for $X(t)$ [2]. I will construct a related unbounded time sequence $\{s_a\}_{a=1}^{\infty}$ and show that it has the property $\lim_{a\to\infty}\langle \frac{X(s_a)}{s_a}, B\frac{X(s_a)}{s_a}\rangle > \lim_{a\to\infty}\langle \frac{X(t_a)}{t_a}, B\frac{X(t_a)}{t_a}\rangle > 0$. This contradicts the definition of the supremum limit.

I establish the contradiction by first finding an increasing, unbounded subsequence $\{t_c\}_{c=1}^{\infty}$ of $\{t_a\}_{a=1}^{\infty}$ and a sequence $\{s_c\}_{c=1}^{\infty}$ which satisfies the following two properties:

Property 3 *1.* $\lim_{c\to\infty} \frac{t_c - s_c}{t_c} = \epsilon \in (0,1)$ *and* $s_c < t_c$ *for each* c.

2. $C(X(t)) \subset C(\eta)$ *for all* $t \in (s_c, t_c]$ *and each* c. *This corresponds to the workload drifting within a single cone throughout the time interval* $(s_c, t_c]$.

Intuitively, the sequence s_c corresponds to the times that the workload vector reenters the cone containing the unbounded sequence t_c. Since the boundary of this cone is becoming further from its center, the increasing distances lead to the contradictory sequence.

Lemma 8 *If the sequences* $\{t_c\}_{c=1}^{\infty}$ *and* $\{s_c\}_{c=1}^{\infty}$ *satisfies the two conditions or property 3 then the supremum limit is not attained on the sequence* $\{t_c\}_{c=1}^{\infty}$.

Because of our assumption of minimal service and lemma 7 implying that no service is applied whenever $X_q = 0$, I know that for all q,

$$X_q(t_c) - X_q(s_c) = \int_{s_c}^{t_c} A_q(t) - \int_{s_c}^{t_c} S_q(t) \tag{22}$$

Multiplying both sides of equation 22 by $[B\eta]_q$ and summing over all components $q \in \mathcal{Q}$,

$$\begin{aligned}\langle X(t_c) - X(s_c), B\eta\rangle &= \langle \int_{s_c}^{t_c} A(t), B\eta\rangle - \langle \int_{s_c}^{t_c} S(t), B\eta\rangle \\ &= \langle \int_{s_c}^{t_c} A(t), B\eta\rangle - \langle S^m, B\eta\rangle(t_c - s_c)\end{aligned} \tag{23}$$

for any $m \in \mu(\eta)$. The time follows because the policy is non-idling and the sequence remains in the same cone, fixing $\langle S^m, B\eta\rangle$ within the time period.

Further, $\lim_{n\to\infty} \frac{\int A(t_n)}{t_n} = \rho$ for any unbounded sequence $\{t_n\}$. Dividing equation 23 by $(t_c - s_c)$ and letting $c \to \infty$ I see that

$$\lim_{c\to\infty}\langle \frac{X(t_c) - X(s_c)}{t_c - s_c}, B\eta\rangle = \langle \rho, B\eta\rangle - \langle S^m, B\eta\rangle = -\gamma(\eta) \le 0 \tag{24}$$

The inequality is due to the stability region definition 4 and letting $v = B\eta$. From the definition of η and the fact that t_c was defined as a thinning of t_a, I have $\lim_{c\to\infty}\frac{X(t_c)}{t_c} = \eta$, hence the first property of s_c gives

[2]For an arrival trace, $X(t) \le A(t) \Rightarrow \frac{X(t)}{t} \le \frac{A(t)}{t} \to \rho$

$$\begin{aligned}
\lim_{c\to\infty} \langle \frac{X(s_c)}{s_c}, B\eta \rangle &= \lim_{c\to\infty} \{ \frac{s_c - t_c}{s_c} \langle \frac{X(t_c) - X(s_c)}{t_c - s_c}, B\eta \rangle + \frac{t_c}{s_c} \langle \frac{X(t_c)}{t_c}, B\eta \rangle \} \\
&= \frac{\langle \eta, \eta \rangle + \epsilon\gamma(\eta)}{1-\epsilon} \\
&> \langle \eta, B\eta \rangle \qquad (25)
\end{aligned}$$

The inequality is due to the facts that $\epsilon \in (0,1)$, and $\gamma(\eta) \geq 0$.

By successive thinnings of the components of the workload vector, I can obtain an increasing unbounded subsequence $\{s_d\}_{d=1}^{\infty}$ of $\{s_c\}_{c=1}^{\infty}$ such that

$$\lim_{d\to\infty} \frac{X(s_d)}{s_d} = \psi \qquad (26)$$

From the positive definiteness and symmetry of B, I know that for $\psi \neq \eta, \langle \psi - \eta, B(\psi - \eta) \rangle > 0$. This implies $\langle \psi, B\psi \rangle + \langle \eta, B\eta \rangle > 2\langle \psi, B\eta \rangle$, and hence equations 25 and 26 imply that $\langle \psi, B\psi \rangle > \langle \eta, B\eta \rangle$. Now

$$\lim_{d\to\infty} \langle \frac{X(s_d)}{s_d}, B\frac{X(s_d)}{s_d} \rangle > \langle \eta, B\eta \rangle = \lim\sup_{t\to\infty} \langle \frac{X(t)}{t}, B\frac{X(t)}{t} \rangle > 0 \qquad (27)$$

giving a contradiction to the definition of η. This completes the proof of lemma 8.

It now remains to construct such a sequence $\{s_c\}_{c=1}^{\infty}$. I do so by the intuition referred to at the start of this proof. Formally, I state and prove the following lemma:

Lemma 9 *Suppose* $\lim_{k\to\infty} \frac{X(t_k)}{t_k} = \eta \neq 0$ *for some increasing unbounded seqeuence* $\{t_k\}_{k=1}^{\infty}$ *and nonzero* η*. Let*

$$s_k = \sup\{t < t_k : C(X(t)) \not\subseteq C(\eta)\} \qquad (28)$$

By convention $s_k = 0$ *if the workload has always been within the same cone up to time*[3] t_k*. Then*

$$\lim\inf_{k\to\infty} \frac{t_k - s_k}{t_k} = \epsilon_1 > 0 \text{ for some } \epsilon_1 \qquad (29)$$

Proof Lemma 5 states that if $\lim_{n\to\infty} \frac{t_n - s_n}{t_n} = 0$ then $\lim_{n\to\infty} \frac{X(t_n) - X(s_n)}{s_n} = 0$. Combining this result with $\lim_{n\to\infty} \frac{X(t_n)}{t_n} = \eta$,we have $\lim_{n\to\infty} \frac{X(s_n)}{s_n} = \eta$. Further, to allow for the possibility of large *jumps* in the workload arrivals, I note from lemma 6 that this implies $\lim_{n\to\infty} \frac{X(s_n^-)}{s_n} = \eta$.

Now I are ready to construct the sequence $\{s_c\}_{c=1}^{\infty}$ satisfying the two properties required for the earlier proposition. I rename the sequence defined in equation 28 to be $\{\hat{s_c}\}$ and choose $s_c = \max\{\hat{s_c}, (1-\epsilon_2)t_c\}$, for some ϵ_2. Then I have the properties:

- $\lim_{c\to\infty} \frac{t_c - s_c}{t_c} = \epsilon \in (0,1)$ and $s_c < t_c$ for each c.
- $C(X(t)) \subset C(\eta)$ for all $t \in (s_c, t_c]$ and each c.

This means that $\{s_c\}_{c=1}^{\infty}$ and $\{t_c\}_{c=1}^{\infty}$ satisfy the both aspects of property 3, and lemma 8 completes the proof of rate stability.

[3]This sequence refers to the last time before each time t_k when the workload was in a different workload cone.

D Rate Stability Necessary Conditions

Here I prove that for a general set of service vectors and arrival processes, it is necessary for the matrix B in the BLQF policy to be positive definite and have nonpositive off diagonal elements.

Theorem 3 states that the matrix B must be positive definite.

Proof of theorem 3 Assume that this is not true. That is, there exists a vector y such that $\langle y, By\rangle \leq 0$. I will construct a set of service vectors for which this leads to an unstable policy.

Let $S^1 = y^+, S^2 = -y^-$. That is, S^1 and S^2 are vectors containing the positive and absolute negative elements of y respectively. Each has positive elements only where the other has zeros.

Imagine a deterministic arrival pattern, with $A(t) = S^1$. The policy will choose to serve at rates defined by either S^1 or S^2 based on the greatest inner product value. In this case it must be necessary for it to choose S^1 since otherwise arrivals keep coming to the queues in the support of S^1 and none are being served since the cone policy will always choose S^2. Thus a necessary condition for stability is $\langle S^1, BS^1\rangle > \langle S^2, BS^1\rangle$. Similarly I must have $\langle S^2, BS^1\rangle > \langle S^2, BS^2\rangle$. But by the construction of S^1 and S^2 I have $\langle (S^1 - S^2), B(S^1 - S^2)\rangle \leq 0$, or $\langle S^1, BS^1\rangle + \langle S^2, BS^2\rangle \leq \langle S^1, BS^2\rangle + \langle S^2, BS^1\rangle$, meaning that one of the above conditions must be violated.

Proof of theorem 4 Now I prove that the off diagonal elements of B cannot be positive. A consequence of theorem 3 is that the diagonal elements of B must be strictly positive.

Assume that $B_{ij} > 0$. Let $S_i^1 = B_{jj}, S_j^2 = 0.5 * B_{ij}$ and all other components of S^1 and S^2 be zero.

Now $\langle S^2, BS^2\rangle = 0.25 * B_{jj} * B_{ij}^2 < \langle S^1, BS^2\rangle = 0.5 * B_{jj} * B_{ij}^2$. Thus if arrivals were deterministic to queue j at rate $0.5 * B_{ij}$ then service would consistently serve the wrong queue and the sysem is seen to be unstable.

Again I have a well-defined set of service configurations and arrival trace for which the policy would necessarily lead to queue starvation and hence instability.

References

[1] N.McKeown, V Anantharam, J. Walrand: "Achieving 100% throughput in an Input-Queued Switch", INFOCOM '96, pp. 296-302

[2] M. Armony, N. Bambos "Queueing Dynamics and Maximal Throughput Scheduling in Switched Processing Systems", Technical Report (2001), Stanford University

[3] I. Keslassy, N. McKeown "Analysis of scheduling algorithms that provide 100% throughput in input-queued switches", Allerton 2001

[4] L. Tassiulas, "Linear complexity algorithms for maximum throughput in radio networks and input queued switches", IEEE 1998

[5] Bambos, Nicholas and J. Walrand, "Scheduling and stability aspects of a general class of parallel processing systems". Adv. Appl. Prob., 25 pp176-202

[6] Kumar, P.R., and S.P. Meyn, "Stability of queueing networks and scheduling policies", IEEE Transactions on Automatic Control, Vol. 40 Issue 2, Feb 1995, pp251 - 260

[7] Birkhoff,G., "Tres obervaciones sobre el algebra lineal", Univ. Nac. Tucuman Rev. Ser. A5 (1946), pp. 147-150