

A Measurement Based Dynamic Policy for Switched Processing Systems

Ying-Chao Hung

Graduate Institute of Statistics
National Central University
Jhongli 32049, Taiwan
hungy@stat.ncu.edu.tw

George Michailidis

Department of Statistics
The University of Michigan
Ann Arbor, MI 48109-1107
gmichail@umich.edu

Abstract- Switched Processing Systems (SPS) represent a canonical model for many areas of applications of communication, computer and manufacturing systems. They are characterized by flexible, interdependent service capabilities and multiple classes of job traffic flows. Recently, increased attention has been paid to the issue of improving quality of service (QoS) performance in terms of delays and backlogs of the associated scheduling policies, rather than simply maximizing the system's throughput. In this study, we investigate a measurement based dynamic service allocation policy that significantly improves performance with respect to delay metrics. The proposed policy solves a linear program at selected points in time that are in turn determined by a monitoring strategy that detects 'significant' changes in the intensities of the input processes. The proposed strategy is illustrated on a small SPS subject to different types of input traffic.

I. INTRODUCTION

There has been a lot of interest over the last decade for the analysis and performance assessment of Switched Processing Systems (SPS). This interest stems from the fact that such systems serve as canonical models in many areas of application, including high performance computing, wireless networking, call centers and flexible manufacturing. They are primarily characterized by the flexible, interdependent nature of their service capabilities, which for many service allocation policies induces complex queueing dynamics. A basic objective, studied fairly extensively over the years, has been the investigation of throughput maximizing scheduling policies [1, 2, 3, 4, 5, 8, 14, 15, 16]. More recently, the focus has shifted to quality of service (QoS) issues and the performance of various scheduling policies with respect to performance metrics, such as delay. Analytical results for service policies for SPS are in general hard to come by, due to the inherent complexity of the underlying queueing dynamics. Some notable exceptions can be found in the work of Ross and Bambos [11, 12], and in our previous work on the subject [6]. Specifically, a randomized algorithm was introduced in [11, 12] and its QoS performance compared with a class of projective cone policies, originally in-

troduced in [1]. A weighted variant of cone policies was studied in [6], where the weights were calculated by estimating online the intensity of the incoming traffic. By measuring incoming traffic and adjusting the weights correspondingly, the proposed throughput maximizing policy becomes more responsive to QoS aspects.

In this paper, a measurement based dynamic scheduling policy of the system's resources is proposed and its performance investigated. The policy relies on measuring the traffic intensity and identifying changes over time. For the latter purpose, we present a methodology that models the traffic intensities as an exponentially weighted moving average (EWMA) process, which achieves the following objectives: (i) prediction of the intensity for the next time period under consideration and (ii) monitoring for significant shifts of the intensity over time. Thus, the policy is adaptive to traffic fluctuations that occur in practice.

The remainder of the paper is organized as follows. In Section II, the SPS is introduced along with a measurement based control policy. In Section III, the EWMA methodology is introduced and implementation issues discussed. The performance assessment of the proposed policy is illustrated in Section IV. Finally, some concluding remarks are drawn in Section V.

II. SPS: DESCRIPTION AND A DYNAMIC CONTROL POLICY

Consider a multiclass queueing system comprised of Q parallel, infinite capacity, first-in-first-out (FIFO) queues, with each queue corresponding to a different class of job traffic. The class q jobs arrive according to a general process \mathcal{A}_q , $q = 1, \dots, Q$. Suppose the j^{th} job of class q arrives to the system at time $t_j^q \in \mathbb{R}_+$ and carries the service requirement δ_j^q . We model these random quantities as elements of a random marked point process (RMPP)

$$\mathcal{I}_q = (t_j^q, \delta_j^q), \quad j \in \mathbb{Z}_+, \quad (1)$$

defined on some probability space (Ω, \mathcal{F}, P) . Define the *stochas-*

tic traffic intensity of queue q at time t by

$$\lambda_q(t) = \lim_{\Delta t \rightarrow 0} \left[\frac{1}{\Delta t} \sum_{j \in \mathbb{Z}_+} \delta_j^q \mathbf{1}_{\{t_j^q \in (t, t + \Delta t)\}} \right] \quad (2)$$

and let $\Lambda(t) = (\lambda_1(t), \dots, \lambda_Q(t))$. It is further assumed that $\lambda_q(t)$ satisfies

$$0 < \mathbb{E} \int_0^T \lambda_q(t) dt < \infty \quad (3)$$

for any finite planning horizon $[0, T]$, $q = 1, \dots, Q$.

Suppose there are M admissible service modes and for each mode $m \in \{1, \dots, M\}$, the jobs of the queue q receive service at rate μ_{mq} . Therefore, mode m is associated with the non-degenerate service rate vector $R_m = (\mu_{m1}, \mu_{m2}, \dots, \mu_{mQ})$. The $M \times Q$ matrix $R = (\mu_{mq})$ is called the *service rate matrix*. It can be seen that the service modes introduce intricate dependencies amongst the queues. We further introduce a null service mode $R_0 = \vec{0}$, which is used when the system idles; i.e. when no jobs are present in the system. Although, in practice splitting of service modes is not possible, we nevertheless make such an assumption for mathematical convenience. Let $r_m(t)$ denote the level at which service mode m is participating at time t , with $0 \leq r_m(t) \leq 1$ for all $m \in \{0, 1, \dots, M\}$, and $\sum_{m=0}^M r_m(t) \leq 1$ for every point in time. In practice, the quantities $r_m(t)$ capture the proportion of time, service mode m is used over a fixed period of time.

Let the row vector $r(t) = (r_1(t), \dots, r_M(t)) \in \mathbb{R}_+^M$ denote the *control process* of the M admissible service modes and further assume that it is non-anticipative. We also assume that the job service requirements are mutually independent and in addition independent of the arrival processes. The *workload process* under any scheduling policy π can be written as

$$W_q^\pi(t) = W_q^\pi(0) + \sum_{j \in \mathbb{Z}_+} \delta_j^q \mathbf{1}_{\{t_j^q \in (0, t]\}} - \sum_{m=1}^M \mu_{mq} \int_0^t r_m(s) \mathbf{1}_{\{W_q^\pi(s) > 0\}} ds \quad (4)$$

for every $q = 1, 2, \dots, Q$.

We discuss next system stability issues. Define the *long-term traffic intensity* of queue q by

$$\lambda_q = \lim_{t \rightarrow \infty} \left[\frac{1}{t} \int_0^t \lambda_q(s) ds \right] \quad (5)$$

where $q = 1, \dots, Q$. The *stability region* of the system under consideration is given by

$$S = \left\{ \Lambda \in \mathbb{R}_+^Q : \lambda_q < \sum_{m=0}^M \omega_m \mu_{mq} \text{ for all } q \in \{1, \dots, Q\} \right\} \quad (6)$$

where

$$\omega_m = \lim_{t \rightarrow \infty} \left[\frac{1}{t} \int_0^t r_m(s) ds \right] \quad (7)$$

represents the long-run proportion that service mode m is employed at a 100% level, $m = 0, \dots, M$, and $\sum_{m=0}^M \omega_m = 1$. Intuitively, the stability region is comprised of all traffic intensity vectors that the system can accommodate in the long-run (in the sense that all workload processes would remain finite) under some resource allocation policy. It can be shown (proof omitted due to space considerations) that if $\vec{\Lambda} \notin S$, then the workload of at least one queue would explode to infinity. It can also be seen that S is the *convex hull* generated by all service vectors R_m , $m = 1, \dots, M$. To illustrate, the stability region for a 2-queue system with 4 service vectors $R_1 = (3, 0)$, $R_2 = (2, 3)$, $R_3 = (0, 4)$, and $R_4 = (2.5, 1)$ is shown in the left panel of Figure 2. The stability region for another 2-queue system with 3 service vectors $R_1 = (0, 5)$, $R_2 = (2, 4)$, and $R_3 = (3, 3)$ is shown in Figure 1.

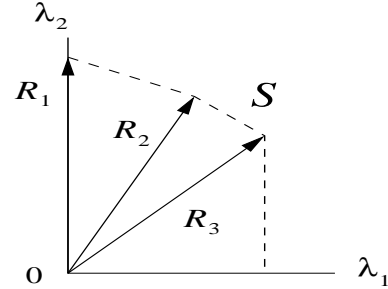


Figure 1: (Right panel): The stability region S for a 2-queue system with $R_1 = (0, 5)$, $R_2 = (2, 4)$ and $R_3 = (3, 3)$.

At certain points in time, the system needs to decide which service mode m should be used. In [1], a dynamic throughput maximizing class of policies was introduced that chooses the mode m that maximizes the inner product $\langle \vec{W}(t), \mu_m \rangle$. A variant of this policy that utilizes a weighted inner product $\langle \vec{W}(t), \mu_m \rangle_{\vec{\alpha}}$ was investigated in our previous work [6]. Careful choice of the weights $\vec{\alpha}$ leads to improved performance from an average delay point of view. We discuss next a measurement based dynamic policy, whose justification is based on a fluid approximation [4].

The fluid analogue of the workload process is given by

$$\bar{W}_q(t) = \bar{W}_q(0) + \int_0^t \lambda_q(s) ds - \sum_{m=1}^M \mu_{mq} \int_0^t r_m(s) ds, \quad (8)$$

for every $q = 1, 2, \dots, Q$, $\bar{W}_q(t) \geq 0$ and $t \geq 0$, where $\bar{W}_q(t)$ represents the fluid level of queue q at time t , which is derived by taking the limit of the following scaled process:

$$\bar{W}_q(t) = \lim_{r \rightarrow \infty} \frac{1}{r} W_q(rt). \quad (9)$$

Suppose that the above fluid model is differentiable at time t , then we have that

$$\dot{\bar{W}}_q(t) = \frac{d\bar{W}_q(t)}{dt} = \lambda_q(t) - \sum_{m=1}^M r_m(t)\mu_{mq}. \quad (10)$$

Note that this approximation of the primitive processes eliminates lower order stochastic fluctuations, but is still influenced by the stochastic intensity $\lambda_q(t)$. For the whole system, we can write

$$\frac{d\bar{W}(t)}{dt} = \Lambda(t) - \sum_{m=1}^M r_m(t)R_m = \Lambda(t) - r(t)R. \quad (11)$$

where the operation between vectors is considered component-wise.

Consider next a fixed planning horizon during which all time-dependent quantities remain fixed at their time 0 levels except for the workload process $\bar{W}(t)$; i.e.

$$\Lambda(t) \equiv \Lambda \quad \text{and} \quad r(t) = r \quad \text{or} \quad \vec{0}, \quad (12)$$

where $\Lambda = (\lambda_1, \dots, \lambda_Q)$ and $r = (r_1, \dots, r_M)$. Therefore, we have that $\frac{d\bar{W}(t)}{dt} = \Lambda - rR$ for all $t \in (0, T]$. This implies that the flow balance equations are given by $\Lambda = rR$. For the original time interval $(0, T]$ we obtain the instantaneous flow balance equations $\Lambda(t) = r(t)R$. The latter suggests that an optimal scheduling policy (i.e. the optimal control process r) solves the following LP problem:

$$\begin{aligned} & \underset{r}{\text{minimize}} && \max_q C_q \left(\frac{\lambda_q}{(rR)_q - \lambda_q} \right) \\ & \text{subject to} && \Lambda \leq rR \end{aligned} \quad (\text{LP})$$

where $(rR)_q$ denotes the q th element of vector rR (i.e. the average service rate devoted to queue q under control process r), and $C_q > 0$ is some cost component, $q = 1, \dots, Q$. It is noted that LP-1 can be translated to (approximately) minimizing the maximal value of the average (linear) holding cost among all queues. This is simply true since

$$\frac{\lambda_q}{(rR)_q - \lambda_q} \approx \mathbb{E}[W_q(t)] \quad \text{for all } t \in [0, T] \quad (13)$$

under fairly strong assumptions such as Little's law and all the input processes \mathcal{I}_q are not away from Compound Poisson with exponentially distributed service requirements.

Remark: Under equal holding costs for all queues, the solution to the above LP problem can be related to the LP problem given in [1] that leads to the so-called 'FastEmpty' policy.

It should be noted that for a SPS with stationary input processes (and hence constant traffic intensities rates Λ) the optimal solution to the above posited LP problem performs poorly with respect to various QoS metrics. However, if the traffic intensity vector is estimated from observing the input processes,

then the control policy $r(t)$ would be more responsive to traffic fluctuations, thus achieving better performance. However, if Λ is estimated and the LP has to be solved continuously, that makes this strategy computationally expensive. Further, such a strategy may lead to a large number of service mode switchings, a shortcoming shared by the cone policies studied in [1].

In order to avoid these issues, the traffic intensity is monitored for changes using an exponential moving average process and the system switches to a different set of service modes only when 'significant' changes are detected. Hence, decision times correspond to change points, when the LP needs to be solved and a new service allocation made. It can also be seen that this strategy easily accommodates non-stationary input processes.

IV. ESTIMATION AND SHIFT-MONITORING OF TRAFFIC INTENSITY

In this section we proceed to make the proposed policy operational. Recall that knowledge of input rates Λ is necessary, which is achieved by estimating them over time by tracking the amount of work coming into the system. For this purpose, suppose that time is divided into sequential intervals (Δ_k, Δ_{k+1}) , $k = 0, 1, 2, \dots$, so that at least n job arrivals are observed in each queue during that interval. Experience shows that $n = 30$ is reasonable in practice. Then, the traffic intensity of queue q over the n th time interval can be estimated by averaging all the service requirements; that is,

$$\hat{\lambda}_q(k) = \frac{1}{(\Delta_k - \Delta_{k-1})} \sum_{j=1}^{n_q} \delta_j^q, \quad k = 1, 2, \dots, \quad (14)$$

where n_q denotes the number of jobs that arrived during this time interval. Thus, a sequence of estimated traffic intensities $\{\hat{\Lambda}(1), \hat{\Lambda}(2), \dots\}$ for all queues are obtained, where $\hat{\Lambda}(n) = (\hat{\lambda}_1(n), \dots, \hat{\lambda}_Q(n))$, $n = 1, 2, \dots$

A naive approach would be to solve the LP and change the service allocation at every interval k . However, such an approach would be sensitive to small shifts of traffic intensities, thus making it computationally expensive. Therefore, we introduce the Exponentially Weighted Moving Average (EWMA) control chart, whose objectives are: (i) to provide an approximate one-step-ahead predictor of traffic intensities by utilizing all previous observations; and (ii) to monitor 'significant' shifts of traffic intensities in the process. As previously mentioned, the LP would be solved only when such changes are detected.

We discuss first the control chart for a single queue. Consider a sequence of estimated traffic intensities $\{\hat{\lambda}_q(1), \hat{\lambda}_q(2), \dots\}$ for queue q and assume that they are independently and normally distributed as

$$\hat{\lambda}_q(k) \sim N(\lambda_q(k), \sigma_q^2), \quad k = 1, 2, \dots, \quad (15)$$

where λ_q denotes its true mean and σ_q^2 its variance. By estimating λ_q using at least measurements from 30 jobs, the normal assumption holds approximately. Let $\bar{\lambda}_q(n+1)$ denote

the one-step-ahead predictor of $\hat{\lambda}_q(k+1)$, and compute the EWMA statistic by

$$\bar{\lambda}_q(n+1) = \beta \hat{\lambda}_q(n) + (1-\beta)\bar{\lambda}_q(n), \quad (16)$$

where $0 < \beta \leq 1$ represents the weight that the most recent observation carries.

If $\bar{\lambda}_q(0)$ is set to be a target value λ_q , it can be shown that

$$\mathbb{E}\bar{\lambda}_q(n) \equiv \lambda_q \quad \text{and} \quad \text{Var}\bar{\lambda}_q(n) = \sigma_q^2 \left(\frac{\beta}{2-\beta} \right) [1 - (1-\beta)^{2n}]. \quad (17)$$

Based on these results, the control limits of an EWMA chart for $\lambda_q(n)$ can be constructed as follows: the lower and upper control limits (LCL/UCL) of the k -th observation are set as

$$\text{LCL/UCL}(k) = \lambda_q \pm c\sigma_q \sqrt{\left(\frac{\beta}{2-\beta} \right) [1 - (1-\beta)^{2k}]}, \quad (18)$$

with the $+$ corresponding to the UCL and the $-$ to the LCL and $c > 0$ a tuning parameter that captures the sensitivity level that one wishes to achieve. The EWMA chart generates an out-of-control signal at period k if $\bar{\lambda}_q(k) > \text{UCL}(k)$ or $\bar{\lambda}_q(k) < \text{LCL}(k)$. The EWMA control chart for one of the queues of a 3-queue system is shown in Figure 2.

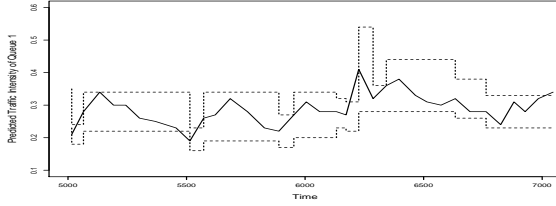


Figure 2: The predicted traffic intensities $\bar{\lambda}_q(k)$ (the solid lines) of a queue monitored by the EWMA chart, with the dash lines representing the UCL and LCL.

Remark: It is assumed that the estimated traffic intensities are independent over time, which may not always be appropriate in practice. If this assumption is violated, then one need to adjust the tuning parameter c in order to avoid too many change detections (positive correlations) or too few (negative correlations).

We now discuss an appropriate extension for the general multiclass SPS. Assume the Q estimated traffic intensities over the k -th time interval follow a multivariate normal distribution with mean Λ and covariance matrix Σ (i.e. $\hat{\Lambda}(n) \sim N(\Lambda(n), \Sigma)$). The Multivariate EWMA statistic (in vector form) is given by

$$\bar{\Lambda}(k+1) = B\hat{\Lambda}(k) + (I-B)\bar{\Lambda}(k) \quad (19)$$

where $B = \text{diag}(\beta_1, \dots, \beta_Q)$, $0 < \beta_q \leq 1$, $q = 1, \dots, Q$, and I is the $Q \times Q$ identity matrix. If there's no specific reason

to weigh past observations differently for the Q input traffic flows, then a single value of β can be chosen for all the queues. Therefore, the asymptotic covariance matrix (as $k \rightarrow \infty$) can be written as $\{\beta/(2-\beta)\}\Sigma$. The MEWMA chart gives an out-of-control signal if

$$T^2(k) = \frac{(2-\beta)}{\beta} \bar{\Lambda}'(k)\Sigma^{-1}\bar{\Lambda}(k) > h, \quad (20)$$

where h is chosen to achieve a specified in-control average run length (ARL) [7].

IV. PERFORMANCE ASSESSMENT

In this Section we examine the performance with respect to delay of the proposed measurement based dynamic policy on three systems: (i) a 2-queue system with compound Poisson input processes, (ii) the same 2-queue system but fed with fractional Brownian traffic (both constant and changing over time) and (iii) a 3-queue system with compound Poisson input processes. We describe the systems and the simulation scenarios next:

A. A 2-queue System with Compound Poisson Input Processes

We consider a 2-queue system with three service modes $R_1 = (0, 4)$, $R_2 = (2, 3)$, and $R_3 = (3, 0)$. Jobs to queue q arrive according to a compound Poisson process \mathcal{A}_q , with mean service requirements are $\mathcal{S}_1 = 1$ and $\mathcal{S}_2 = 10$, and corresponding average interarrival times a $\mathcal{T}_1 = 0.345$ and $\mathcal{T}_2 = 10$, respectively (i.e. hence, the traffic intensities are $\Lambda = (\lambda_1, \lambda_2) = (0.3, 3.7)$). An extensive computer simulation of the system is carried out and various statistics of the delay process recorded. The parameters for constructing the EWMA chart were set to $c = 2$ and $n = 30$. We focus on the system's average delay, as well as the 95th percentile of the delay distribution under the proposed policy, as well as under the dynamic policy that maximizes $\langle \bar{W}(t), \mu_m \rangle_{\bar{\alpha}}$ (MaxProduct policy), for different *but fixed throughout* the simulation choices of weights $\bar{\alpha}$. Note that all the weight vectors $\bar{\alpha} = (\alpha_1, \alpha_2)$ are chosen so that $\alpha_1 + \alpha_2 = 1$, which include all possible directions in \mathbb{R}_+^2 . The resulting average system delays are shown in the left panel of Figure 3, while the the 95th percentile of the delay distribution in the right panel. It can clearly be seen from Figure 3 that by dynamically allocating the system's resources and monitoring changes in traffic intensities, a significantly lower average system delay and 95th percentile of delays are obtained compared to that corresponding to almost any fixed choice for the Max-Product policies. In a large number of cases, the improvement in performance is over 100%, and exceeding 500% for some prespecified weight vectors.

B. A 2-queue System with Fractional Brownian Input Traffic

We consider next the same 2-queue system, fed by the following fractional Brownian type of traffic considered by Norros [10]: $A_q(t) = \lambda_q t + \sigma_q Z_q(t)$, $q = 1, 2$, where $A_q(t)$

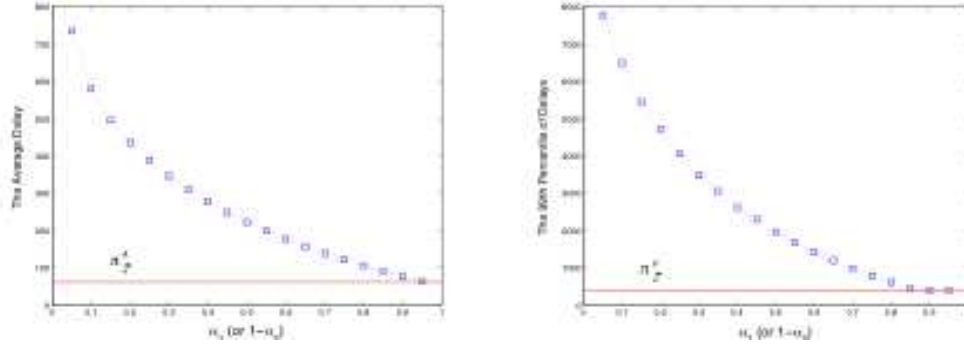


Figure 3: (Left panel): The system’s average delay under compound Poisson inputs for the dynamic measurement based policy and the the MaxProduct policy with all possibly fixed weights $\vec{\alpha}$. (Right panel): The 95th percentile of the system’s delay distribution.

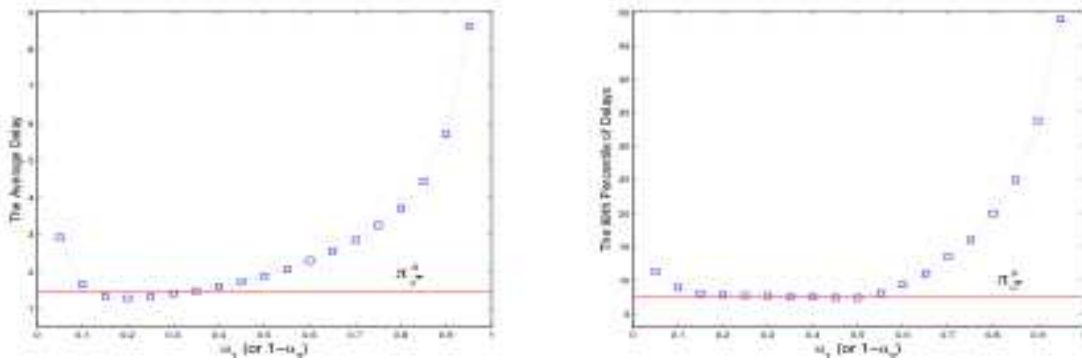


Figure 4: (Left panel): The system’s average delay under FBM inputs for the dynamic measurement based policy and the the MaxProduct policy with all possibly fixed weights $\vec{\alpha}$. (Right panel): The 95th percentile of the system’s delay distribution.

is the total service requirement accumulated in the time interval $(0, t]$, $Z_1(t)$ and $Z_2(t)$ are independent *normalized fractional Brownian motions* (FBM) with the same Hurst parameter H , λ_q is the traffic intensity, and σ_q is the variance of traffic in one time unit. The choice of the parameters were $\Lambda = (\lambda_1, \lambda_2) = (2.66, 1.0)$ and the corresponding variances $(\sigma_1, \sigma_2) = (100, 50)$. Note that these choices of λ_q and σ_q can guarantee that the system is stable and $A_q(t)$ will not go negative, almost surely. In addition, the Hurst parameter is chosen to be $H = 0.7$ for each $Z_q(t)$ so that the two input processes are characterized as *long-range dependent* (so the independence assumption in the EWMA chart is violated). Again, computer simulations were performed to derive the average system delays and the 95th percentiles of delays under policy $\pi_{\vec{\alpha}(\hat{\Lambda})}$ and $\pi_{\vec{\alpha}}$ with all possibly fixed choices of queue weights $\vec{\alpha}$. The results are shown in Figure 4 and to a large extent confirm the previous findings.

We consider next the same fractional Brownian motion input

processes, but with changing input rates for the same 2-queue system. Assume the system starts at time 0 with the initial traffic intensities $\Lambda = (\lambda_1, \lambda_2) = (0.8, 4.0)$ and Λ can change periodically at certain points in time, $t = \{2500, 5000, 7500, \dots\}$. Specifically, define the traffic intensities at time t by $\Lambda(t) = (0.8, 4.0)$ for $t \in [2500i, 2500(i+1))$, $i = 0, 2, 4, \dots$, and $\Lambda(t) = (3.0, 1.8)$ for $t \in [2500i, 2500(i+1))$, $i = 1, 3, 5, \dots$. Note that the long-term traffic intensities for both queues are then $\Lambda = (1.9, 2.9)$, which clearly belong to the stability region. The variance functions for both queues are $\sigma_1(t) = 10$, $t \in [2500i, 2500(i+1))$, $i = 0, 2, 4, \dots$, and $\sigma_2(t) = 50$, $t \in [2500i, 2500(i+1))$, $i = 1, 3, 5, \dots$, respectively. The resulting average system delays and the 95th percentile of job delays under policy $\pi_{\vec{\alpha}(\hat{\Lambda})}$ and $\pi_{\vec{\alpha}}$ with all possibly fixed choices of queue weights $\vec{\alpha}$ are shown in Figure 5. In this case, the measurement based policy outperforms by a wide margin (for most choices of the weights) the MaxProduct policy, both in terms of the average system delay and in terms of the 95th per-

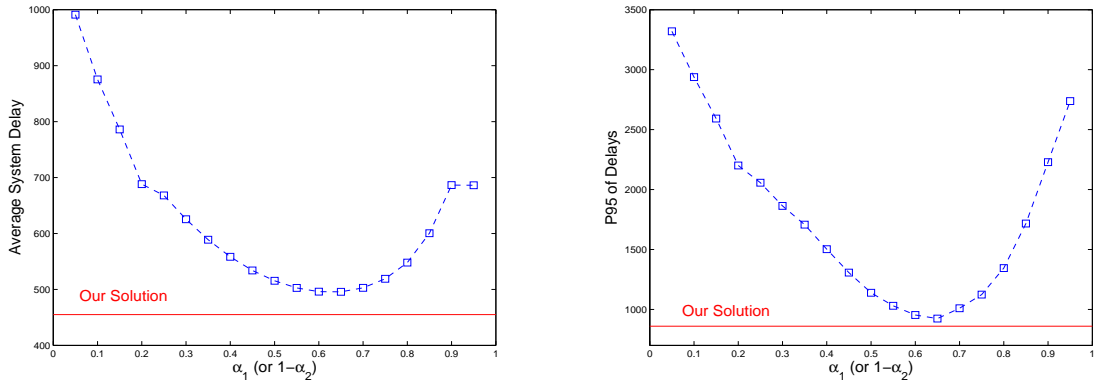


Figure 5: (Left panel): The system’s average delay under FBM inputs for the dynamic measurement based policy and the the MaxProduct policy with all possibly fixed weights $\vec{\alpha}$. (Right panel): The 95th percentile of the system’s delay distribution.

centile of the job delay distribution.

C. A 3-queue System with Changing Compound Poisson Input Processes

We consider next a 3-queue system with 6 admissible service vectors $R_1 = (0, 0, 6)$, $R_2 = (4, 0, 3)$, $R_3 = (6, 0, 0)$, $R_4 = (4, 5, 0)$, $R_5 = (0, 8, 0)$, and $R_6 = (0, 5, 3)$. Suppose there are four possible input process vectors $\vec{I}_1 = (\mathcal{I}_1^1, \mathcal{I}_1^2, \mathcal{I}_1^3)$, $\vec{I}_2 = (\mathcal{I}_2^1, \mathcal{I}_2^2, \mathcal{I}_2^3)$, $\vec{I}_3 = (\mathcal{I}_3^1, \mathcal{I}_3^2, \mathcal{I}_3^3)$, and $\vec{I}_4 = (\mathcal{I}_4^1, \mathcal{I}_4^2, \mathcal{I}_4^3)$ so that each input process vector is equally likely to happen every 25000 time units. Assume that for each input process vector \vec{I}_i , all the elements \mathcal{I}_i^q are compound Poisson processes, $i = 1, 2, 3, 4$, and $q = 1, 2, 3$. Specifically, we further assume that the mean service requirements are $(\mathcal{S}_1^1, \mathcal{S}_1^2, \mathcal{S}_1^3) = (0.55, 0.55, 2.32)$, $(\mathcal{S}_2^1, \mathcal{S}_2^2, \mathcal{S}_2^3) = (2.32, 0.55, 0.55)$, $(\mathcal{S}_3^1, \mathcal{S}_3^2, \mathcal{S}_3^3) = (0.55, 2.72, 0.55)$, and $(\mathcal{S}_4^1, \mathcal{S}_4^2, \mathcal{S}_4^3) = (1.38, 2.21, 1.22)$, while for simplicity, the mean interarrival times are chosen to be $T_i^q = 1/S_i^q$ for all i, q . Therefore, the long-term traffic intensities for these 4 input combinations are $\Lambda_1 = (0.3, 0.3, 5.4)$, $\Lambda_2 = (5.4, 0.3, 0.3)$, $\Lambda_3 = (0.3, 7.4, 0.3)$, and $\Lambda_4 = (1.9, 4.9, 1.5)$.

Computer simulations are performed to derive the average system delay and the 95th percentile of job delays under the measurement based proposed policy, the MaxProduct policies, as well as the Maximum Weighted Queue Length policy that is a MaxProduct policy with weight vector 1, but employs the queue length process instead of the workload process. Note that for the family of MaxProduct policies with fixed queue weights, it suffices to look at all $(\alpha_1, \alpha_2, \alpha_3)$ combinations satisfying $\alpha_1 + \alpha_2 + \alpha_3 = 1$ (which include all possible directions in \mathbb{R}^3). The resulting contour plots of the average system delays and the 95th percentile of job delays are shown in Figure 6. The minimal average system delay and the minimal 95th percentile of delays are both given around $\vec{\alpha} = (0.55, 0.05, 0.4)$, with corresponding values 33.83 and 121.12, respectively. On the other hand, the proposed policy achieves an average system delay of 21.84 and the 95th percentile of delays 70.16, while

the MWQL policy has an average system delay 28.93 and the 95th percentile of delays 92.36.

Overall, these results indicate that the measurement based policy is competitive for different input processes and is most suitable when the underlying input processes change over time.

VI. CONCLUDING REMARKS

In this paper, a measurement based dynamic service allocation policy was introduced, whose goal is to improve the QoS for switched processing systems. The proposed strategy is intentionally designed to minimize (approximately) the maximal value of the average holding cost among all queues. It employs a statistical technique called Exponentially Weighted Moving Average (EWMA), which provides an approximate one-step-ahead predictor for the traffic intensity and at the same time monitors the shifts of the traffic intensity over time. Further, the proposed policy is throughput maximizing. Simulation results show that it achieves significant improvement in terms of average delay and the high percentile of delays for various types of input traffic.

Topics currently under investigation include: (i) the performance of the policy for general non-stationary input processes; and (ii) the computational complexity of obtaining the optimal solution in a SPS comprised of a very large number of queues and service configurations.

References

- [1] M. Armony and N. Bambos, “Queueing Dynamics and Maximal Throughput Scheduling in Switched Processing Systems”, *Queueing Systems: Theory and Applications*, 44(3), pp. 209-252, 2003.
- [2] N. Bambos and G. Michailidis, “Queueing Networks in Random Environments”, *Advances in Applied Probability*, 36, pp. 293-337, 2004.

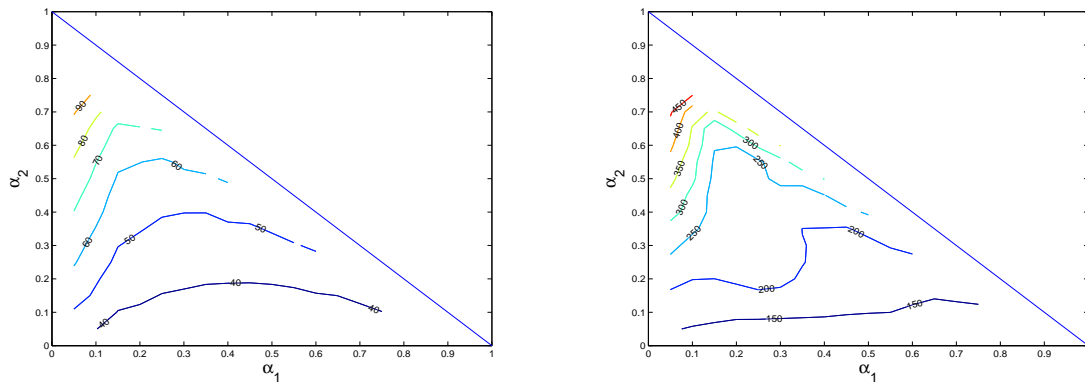


Figure 6: (Left panel): The system's average delay under compound Poisson inputs for the dynamic measurement based policy and the the MaxProduct policy with all possibly fixed weights $\vec{\alpha}$. (Right panel): The 95th percentile of the system's delay distribution.

[3] N. Bambos and G. Michailidis, "Queueing Networks of Random Link Topology: Stationary Dynamics of Maximal Throughput Schedules", *Queueing Systems*, 50, pp. 5-52, 2005.

[4] J.G. Dai and B. Prabhakar, "The Throughput of Data Switches with and without Speedup", *Proceedings of IEEE INFOCOM*, pp. 556-564, 2000.

[5] Y.C. Hung, "Modeling and Analysis of Stochastic Networks with Shared Resources", *Ph.D. thesis*, Department of Statistics, The University of Michigan, 2002.

[6] Y.C. Hung and G. Michailidis. "Improving Quality of Service for Switched Processing Systems", *Proceedings of 11th International Workshop on Computer-Aided Modeling, Analysis and Design of Communication Links and Networks*, 2006

[7] C.A. Lowry and W.H. Woodall, "A Multivariate Exponentially Weighted Moving Average Control Chart", *Technometrics*, 34(1), pp. 46-53, 1992.

[8] G. Michailidis, "Optimal Resource Allocation in a Queueing System with Shared Resources", *Proceedings of the 42nd Conference on Decision and Control*, 2003.

[9] D.C. Montgomery, *Introduction to Statistical Quality Control, 3rd Edition*, John Wiley and Sons, 1996.

[10] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks", *IEEE J. Selected Areas Commun.*, 13(6), pp. 953-962, 1995.

[11] K. Ross and N. Bambos, "Dynamic Quality of Service Control in Packet Switch Scheduling", *Proceedings of IEEE International Conference on Communications*, 2005.

[12] K. Ross and N. Bambos, "Optimizing Quality of Service in Packet Switch Scheduling", *Proceedings of IEEE International Conference on Communications*, 2004.

[13] K. Ross and N. Bambos, "Local Search Scheduling Algorithms for Maximal Throughput in Packet Switches", *Proceedings of IEEE INFOCOM*, 2004.

[14] L. Tassiulas and P.P. Bhattacharya, "Allocation of Interdependent Resources for Maximal Throughput", *Stochastic Models*, 16(1), pp. 27-48, 1999.

[15] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling for maximum throughput in multihop radio networks", *IEEE Transactions on Automatic Control*, 37(12), pp. 1936-1949, 1992.

[16] K.M. Wasserman, G. Michailidis and N. Bambos, "Optimal Processor Allocation to Differentiated Job Flows", *Performance Evaluation*, 63, pp. 1-14, 2006.