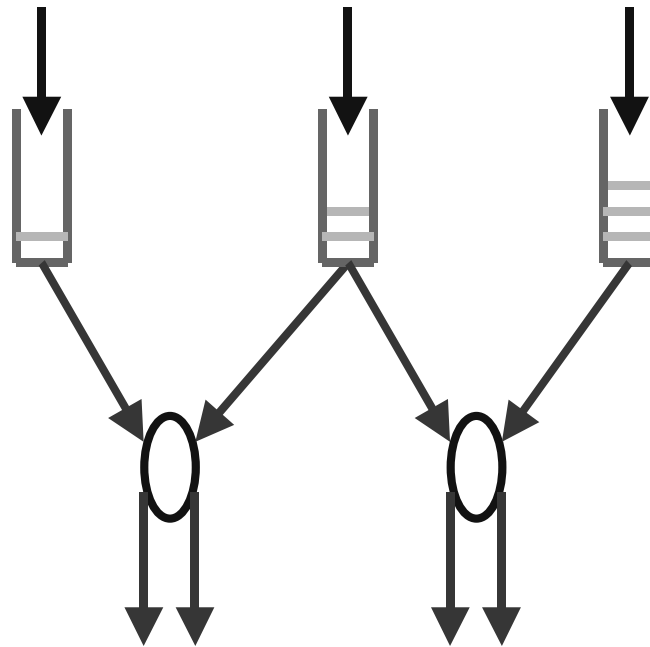


Control of Stochastic Processing Networks: Some Theory and Examples

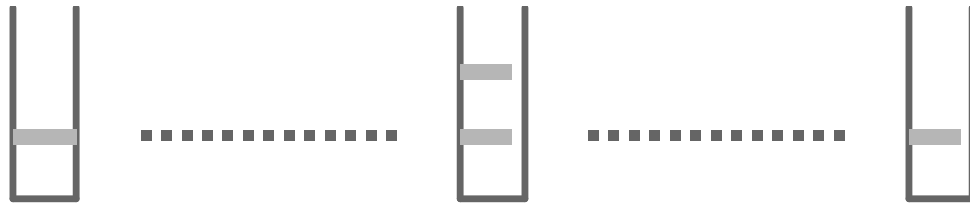


Ruth J. Williams

University of California, San Diego

<http://math.ucsd.edu/~williams>

Stochastic Processing Networks



**I buffers
(classes)**

J activities

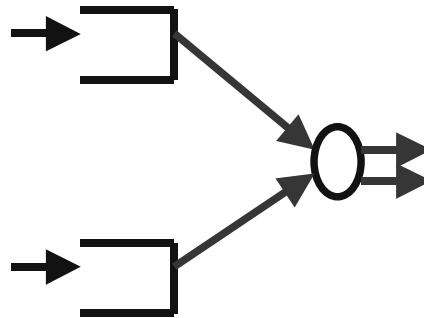
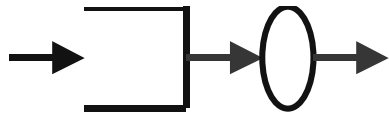


**K servers
(resources)**

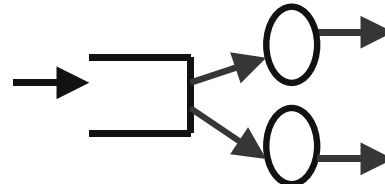
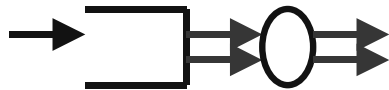
An activity consumes from certain classes,
produces for certain (possibly different) classes,
and uses certain servers.

Stochastic Processing Networks

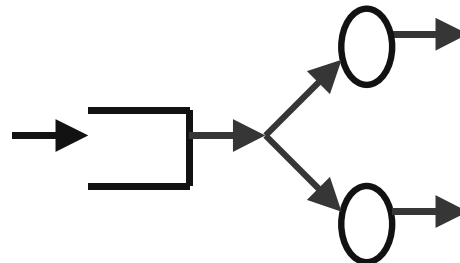
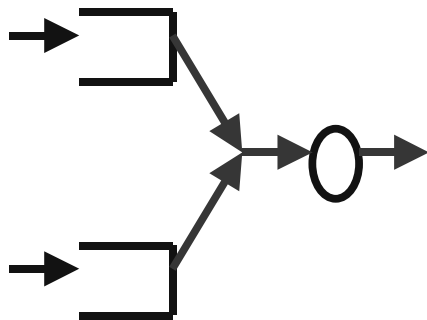
SPN Activities are Very General



Queueing network



*Flexible servers,
alternate routing*



Simultaneous actions

PERSPECTIVE

MQN

SPN

HL

Sufficient conditions for
stability and diffusion
approximations

e.g., parallel server system,
packet switch

Non-
HL

e.g., LIFO, Processor Sharing
(single station,
PS: network stability)

e.g., Internet congestion
control / bandwidth sharing
model

BROWNIAN MODEL APPROACH TO DYNAMIC CONTROL OF SPNs

1. Formulate stochastic network model & associated control problem
2. Define heavy traffic (alternate routing)
3. Formal diffusion approximation: BCP (Brownian control problem)
4. Reduce to EWF (equivalent workload formulation)
5. Solve the BCP (or EWF)
6. Interpret the solution of the BCP
7. Analyze the performance of this policy & prove asymptotic optimality

BROWNIAN MODEL APPROACH TO DYNAMIC CONTROL OF SPNs

1. Formulate stochastic network model & associated control problem
 2. Define heavy traffic (alternate routing)
 3. Formal diffusion approximation: BCP (Brownian control problem)
 4. Reduce to EWF (equivalent workload formulation)
 5. Solve the BCP (or EWF)
 6. Interpret the solution of the BCP
 7. Analyze the performance of this policy & prove asymptotic optimality
- OVERALL APPROACH (ASSUMING HL SERVICE): Harrison '88, Laws '92, Kelly-Laws '93, Harrison-Van Mieghem '97, Harrison '00
 - (More general notion of heavy traffic: Harrison '03: A broader view of Brownian networks, Harrison-W '05, '06: workload reduction; NOT discussed here)
 - HL: head-of-the-line, but not necessarily non-idling

REFERENCES

■ OVERALL APPROACH:

Harrison '88, Laws '92, Kelly-Laws '93, Harrison-Van Mieghem '97, Harrison '00

■ USED TO GENERATE "GOOD" POLICIES:

Harrison-Wein '89, '90, Wein '90, '92, Laws '90, '92, Laws-Louth '90, Kelly -Laws '93,

■ ANALYSIS OF WORKLOAD AND BCP (or EWF):

Bramson-W '03 workload, Budhiraja-Atar '06 HJB, Kumar-Muthuraman '04 numerical

■ SYSTEMATIC INTERPRETATION OF SOLNS (by discretization):

Harrison '96 BIGSTEP, Teh '99, Maglaras '00, Kushner-Dupuis '01

■ FEW PROOFS OF ASYMPTOTIC OPTIMALITY:

Criss-cross: Martins-Shreve-Soner '96, Kushner-Martins '96, Budhiraja-Ghosh '05

Closed two station network: S. Kumar '00

Parallel server: Harrison '98 (Harrison-Lopez '99), Kushner-Chen '99, Bell-W '01, '05, Stolyar '04, Mandelbaum-Stolyar '04, Ata-Kumar '05 (feedback)

■ RELATED WORK: Meyn '01-'04, Dai '04, Shah-Wischik '06, Kang-W '06, Kelly-W '04,.7

STOCHASTIC PROCESSING NETWORK MODEL

First Order Data (average rates)

External arrival rate vector $\alpha \geq 0$ ($\alpha \neq 0$)

α_i = average rate of exogenous arrivals to class i

Input-output matrix R

R_{ij} = average amount of class i material consumed per unit of activity j

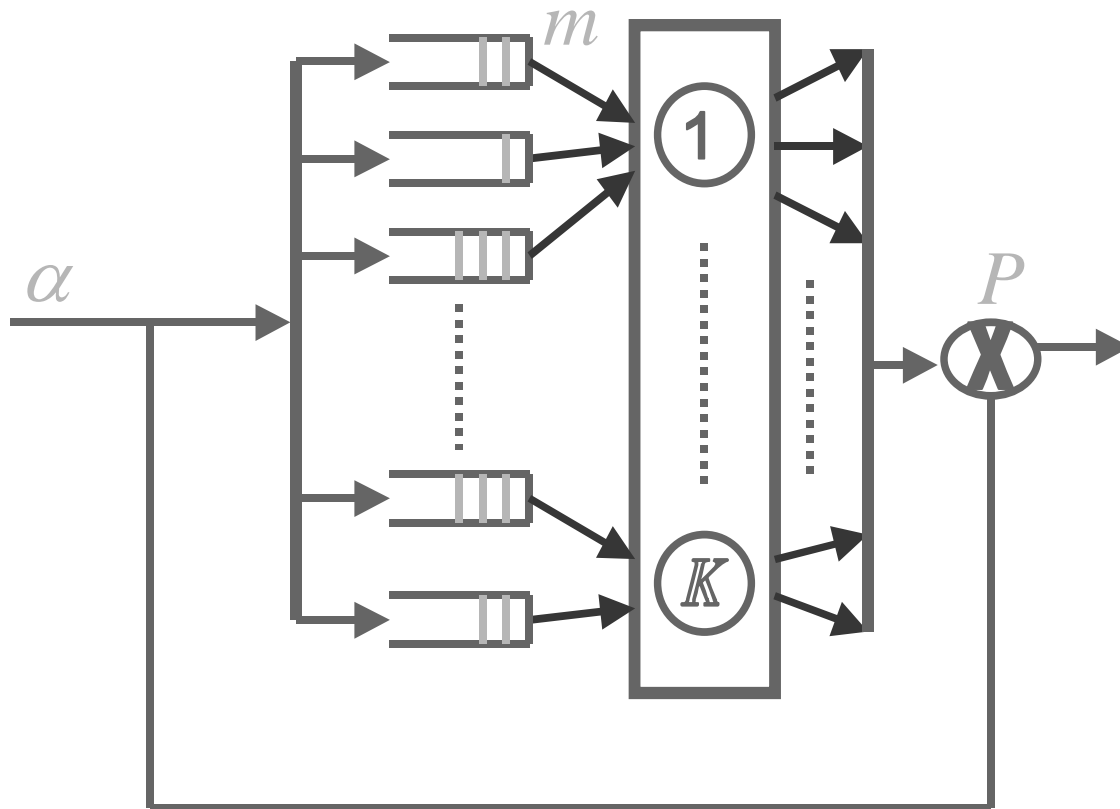
(If $R_{ij} < 0$, then $|R_{ij}|$ is the amount of class i material produced per unit of activity j)

Capacity consumption matrix $A \geq 0$

A_{kj} = average amount of server k 's capacity consumed per unit of activity j

EXAMPLES

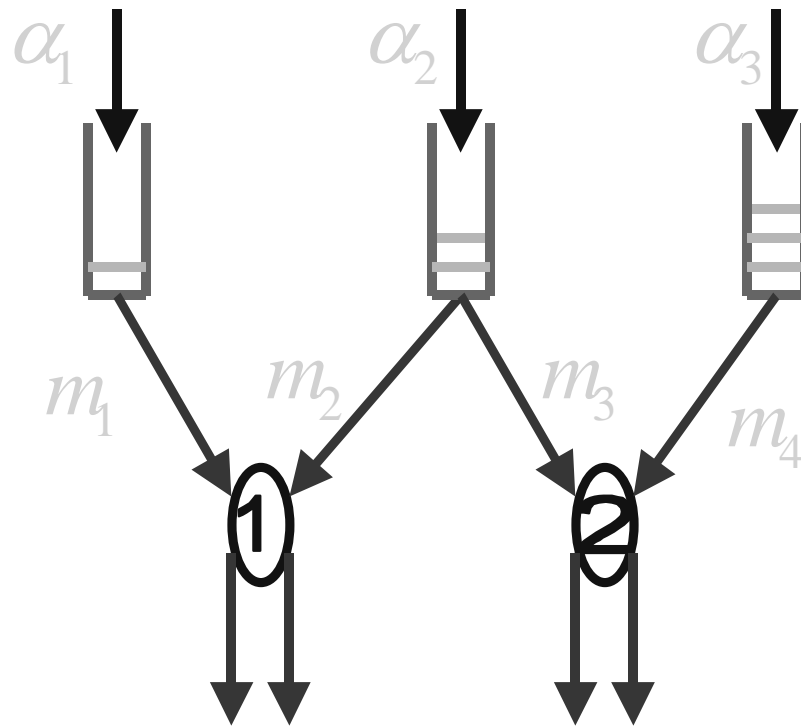
Open Multiclass Queueing Network



$$M = \text{diag}(m) \quad R = (I - P')M^{-1}$$

$A_{kj} = 1$ iff buffer (activity) j is served at k

Parallel Server System



$$R = \begin{bmatrix} m_1^{-1} & 0 & 0 & 0 \\ 0 & m_2^{-1} & m_3^{-1} & 0 \\ 0 & 0 & 0 & m_4^{-1} \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

SPN with Control of Allocations to Activities

(Harrison '00-implicitly HL)

Queue length process

$$Q(t) = Q(0) + E(t) - \sum_{j=1}^J F^j(T_j(t))$$

Idle time process

$$I(t) = \mathbb{1}t - AT(t)$$

Control

$T_j(t)$ = total number of units of activity j used up to time t

Relationship to first order data

$$(E(r\bullet)/r, F(r\bullet)/r) \Rightarrow (\alpha(\bullet), R(\bullet)) \text{ as } r \rightarrow \infty,$$

where $\alpha(t) = \alpha t$, $R(t) = Rt$, $t \geq 0$.

FLUID MODEL AND HEAVY TRAFFIC

Fluid Model

$$\bar{Q}(t) = \bar{Q}(0) + \alpha t - R\bar{T}(t)$$

$$\bar{I}(t) = 1t - A\bar{T}(t)$$

where

$$\bar{Q}_i(t) \geq 0 \text{ for all } i$$

$\bar{I}_k(\bullet)$ is continuous and non-decreasing

$$\bar{I}_k(0) = 0$$

$\bar{T}_j(\bullet)$ is non-decreasing

$$\bar{T}_j(0) = 0$$

Heavy Traffic

Heavy traffic (HT): The following two conditions hold:

- (i) there is a unique fluid control \bar{T}^* under which the fluid model is balanced ($\bar{Q}(t) = \bar{Q}(0)$ for all t),
- (ii) under \bar{T}^* the fluid model incurs no idleness ($\bar{I}(\cdot) = 0$)

Heavy Traffic

Heavy traffic (HT): The following two conditions hold:

(i) there is a unique fluid control \bar{T}^* under which the fluid model is balanced ($\bar{Q}(t) = \bar{Q}(0)$ for all t),

(ii) under \bar{T}^* the fluid model incurs no idleness ($\bar{I}(\cdot) = 0$)

Then $x = \bar{T}^*(1)$, $\rho = 1$ is the unique soln of the linear program:

$$\min \rho \quad \text{s.t.} \quad Rx = \alpha, \quad Ax \leq \rho \mathbf{1}, \quad x \geq 0 \quad (\text{LP})$$

Heavy Traffic

Heavy traffic (HT): The following two conditions hold:

(i) there is a unique fluid control \bar{T}^* under which the fluid model is balanced ($\bar{Q}(t) = \bar{Q}(0)$ for all t),

(ii) under \bar{T}^* the fluid model incurs no idleness ($\bar{I}(\cdot) = 0$)

Then $x = \bar{T}^*(1)$, $\rho = 1$ is the unique soln of the linear program:

$$\min \rho \quad \text{s.t.} \quad Rx = \alpha, \quad Ax \leq \rho \mathbf{1}, \quad x \geq 0 \quad (\text{LP})$$

Equivalent notion of heavy traffic (Harrison '00):

There is a unique optimal solution (ρ^*, x^*) of the linear program and this unique solution satisfies $Ax^* = \mathbf{1}$, $\rho^* = 1$

Heavy Traffic

Heavy traffic (HT): The following two conditions hold:

(i) there is a unique fluid control \bar{T}^* under which the fluid model is balanced ($\bar{Q}(t) = \bar{Q}(0)$ for all t),

(ii) under \bar{T}^* the fluid model incurs no idleness ($\bar{I}(\cdot) = 0$)

Then $x = \bar{T}^*(1)$, $\rho = 1$ is the unique soln of the linear program:

$$\min \rho \quad \text{s.t.} \quad Rx = \alpha, \quad Ax \leq \rho \mathbf{1}, \quad x \geq 0 \quad (\text{LP})$$

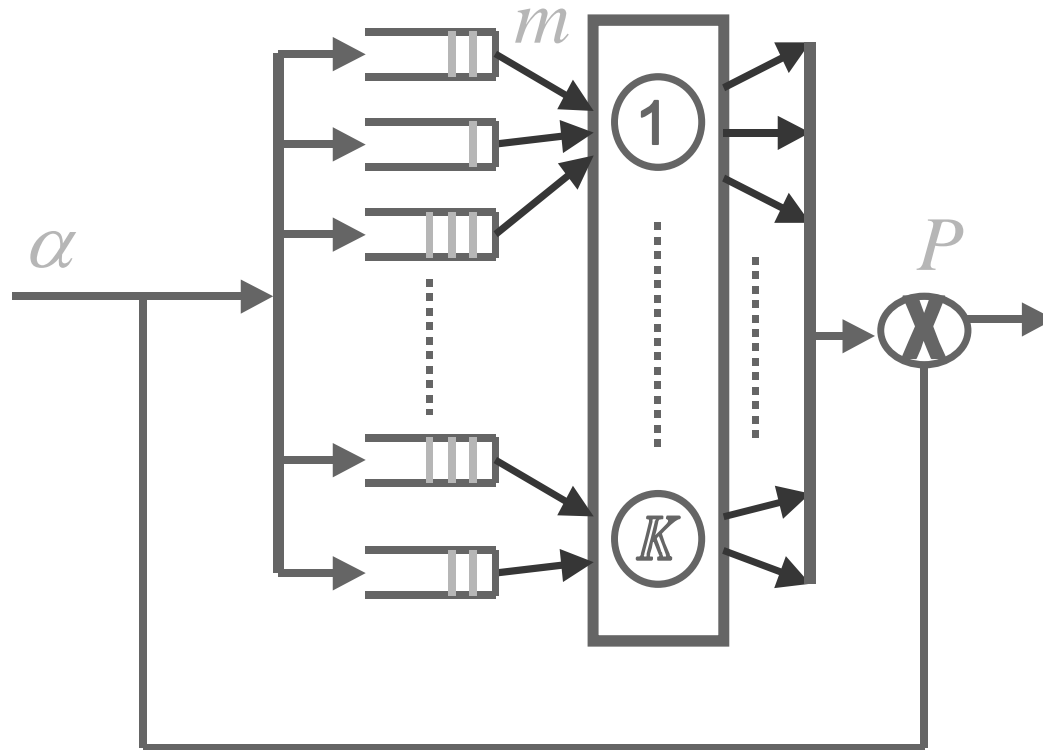
Equivalent notion of heavy traffic (Harrison '00):

There is a unique optimal solution (ρ^*, x^*) of the linear program and this unique solution satisfies $Ax^* = \mathbf{1}$, $\rho^* = 1$

Basic activities: j such that $x_j^* > 0$ \mathbb{B} = number of basic activities

EXAMPLES

Open Multiclass HL Queueing Network



$$M = \text{diag}(m)$$

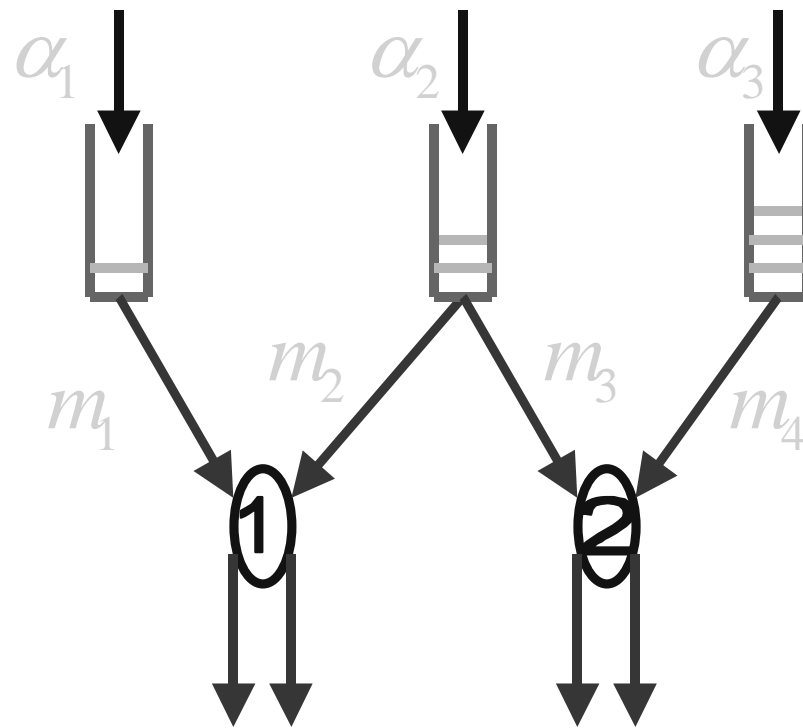
$$R = (I - P')M^{-1}$$

$$\lambda = \alpha + P' \lambda$$

$$\rho = AM \lambda$$

Heavy traffic: $\rho = 1$

Parallel Server System



$$\alpha_1 = 0.2, \alpha_2 = 1.2, \alpha_3 = 0.35 \quad m_1 = 0.5, m_2 = 1, m_3 = 1, m_4 = 2$$

$$x_1^* = 0.1, x_2^* = 0.9, x_3^* = 0.3, x_4^* = 0.7$$

BROWNIAN CONTROL PROBLEM

SPN Control Problem

(with rescaling in heavy traffic)

■ Assume HT holds henceforth

■ Define centered, rescaled processes

$$\hat{E}^r(t) = (E(r^2t) - \alpha r^2t) / r$$

$$\hat{F}^r(t) = (F(r^2t) - Rr^2t) / r$$

$$\hat{Q}^r(t) = Q(r^2t) / r$$

$$\hat{I}^r(t) = I(r^2t) / r$$

$$\hat{Y}^r(t) = (x^* r^2t - T(r^2t)) / r$$

$$\bar{\bar{T}}^r(t) = T(r^2t) / r^2$$

Queue length process

$$\hat{Q}^r(t) = \hat{Q}^r(0) + \hat{E}^r(t) - \sum_{j=1}^J \hat{F}^{r,j}(\bar{\bar{T}}_j^r(t)) + (\alpha - Rx^*)rt + R\hat{Y}^r(t)$$

Idle time process

$$\hat{I}^r(t) = A\hat{Y}^r(t)$$

Cost functional

$$J^r(T^r) = \mathbb{E} \left[\int_0^\infty e^{-\gamma t} h \bullet \hat{Q}^r(t) dt \right]$$

Brownian Control Problem

$$\min_{\tilde{Y}} \mathbb{E} \left[\int_0^\infty e^{-\gamma t} h \cdot \tilde{Q}(t) dt \right]$$

$$\tilde{Q}(t) = \tilde{Q}(0) + \tilde{X}(t) + R\tilde{Y}(t)$$

$$\tilde{I}(t) = A\tilde{Y}(t)$$

where \tilde{X} is a Brownian motion,

$\tilde{Q}_i(t) \geq 0$ for all i

$\tilde{I}_k(\cdot)$ is continuous and non-decreasing, $\tilde{I}_k(0) \geq 0$, for all k

$\tilde{Y}_j(\cdot)$ is non-increasing and $\tilde{Y}_j(0) \leq 0$ for all non-basic j

\tilde{Y} does not anticipate the future of \tilde{X}

Workload

(Harrison and Van Mieghem '97)

Space of reversible queue length displacements:

$$\mathcal{R} \equiv \{ R y : A y = 0, \quad y_N = 0 \}$$

Let \mathcal{L} be the dimension of \mathcal{R}^\perp .

Let Γ be an $\mathcal{L} \times \mathcal{L}$ matrix whose rows are a basis for \mathcal{R}^\perp .

Harrison and Van Mieghem reduce the BCP to an equivalent formulation in which \tilde{Q} is replaced by a "workload" process:

$$\begin{aligned} \tilde{W}(t) &= \Gamma \tilde{Q}(t) \\ &= \tilde{W}(0) + \Gamma \tilde{X}(t) + \Gamma R \tilde{Y}(t) \\ &= \tilde{W}(0) + \Gamma \tilde{X}(t) + G \tilde{I}(t) - H \tilde{Y}_N(t) \end{aligned}$$

Note: Harrison '00 proposes a way to choose Γ via dual LP.

Workload Dimension

(Bramson-W '03)

Theorem Suppose that each column of A contains at most one strictly positive entry. Then, $\mathbb{L} = \mathbb{I} + \mathbb{K} - \mathbb{B}$

Examples:

Multiclass Queueing Networks: $\mathbb{L} = \mathbb{K}$ (total workload)

Parallel server systems: formula applies

(special case of $\mathbb{L}=1$ proved by Harrison-Lopez '99)

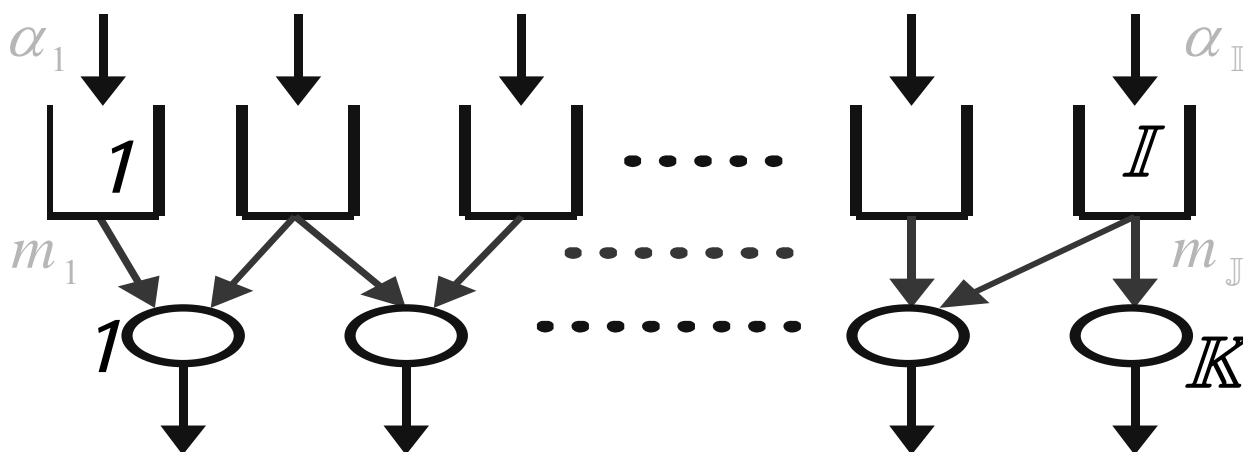
Notes: There is a more general version of the formula without the assumption on A in Bramson-W '03 and also sufficient conditions for Γ , G and H to have all entries non-negative.

BROWNIAN MODEL APPROACH TO DYNAMIC CONTROL OF SPNs

1. Formulate stochastic network model & associated control problem
2. Define heavy traffic (alternate routing)
3. Formal diffusion approximation: BCP (Brownian control problem)
4. Reduce to EWF (equivalent workload formulation)
5. Solve the BCP (or EWF)
6. Interpret the solution of the BCP
7. Analyze the performance of this policy & prove asymptotic optimality

PARALLEL SERVER SYSTEM

COMPLETE RESOURCE POOLING



Theorem (Harrison-Lopez '99) The following are equivalent.

- (i) the workload is one-dimensional,
- (ii) $\mathbb{B} = \mathbb{I} + \mathbb{K} - 1$,
- (iii) there is a unique solution of the dual to (LP),
- (iv) all servers communicate via basic activities.

In fact, under any of these conditions, the server-buffer graph with basic activities as edges is a tree. (W '00; Squillante-Xia-Yao-Zhang '00)

Solution of the BCP under complete resource pooling (Harrison-Lopez '99)

- Unique soln of (DP): (y^*, z^*)

- One-dimensional workload process:

$$\tilde{W}(t) = y^* \cdot \tilde{Q}(t) = y^* \cdot \tilde{X}(t) + z^* \cdot \tilde{I}(t) - u^* \cdot \tilde{Y}_N(t)$$

- Holding cost: $h \cdot \tilde{Q}(t) \geq c^* \tilde{W}(t) \quad c^* = \min\{h_i / y_i^* : i = 1, \dots, I\}$

- Minimum workload: $\tilde{W}^*(t) = y^* \cdot \tilde{X}(t) + \tilde{V}^*(t)$

$$\tilde{V}^*(t) = \max\{-y^* \cdot \tilde{X}(s) : 0 \leq s \leq t\}$$

- Optimal queue length and idletime:

$$i^* = \arg \min\{h_i / y_i^* : i = 1, \dots, I\},$$

k^* serves i^* via basic activity

$$\tilde{Q}_{i^*}^*(t) = \tilde{W}^*(t) / y_{i^*}^*, \quad \tilde{Q}_i^*(t) = 0 \text{ for } i \neq i^*$$

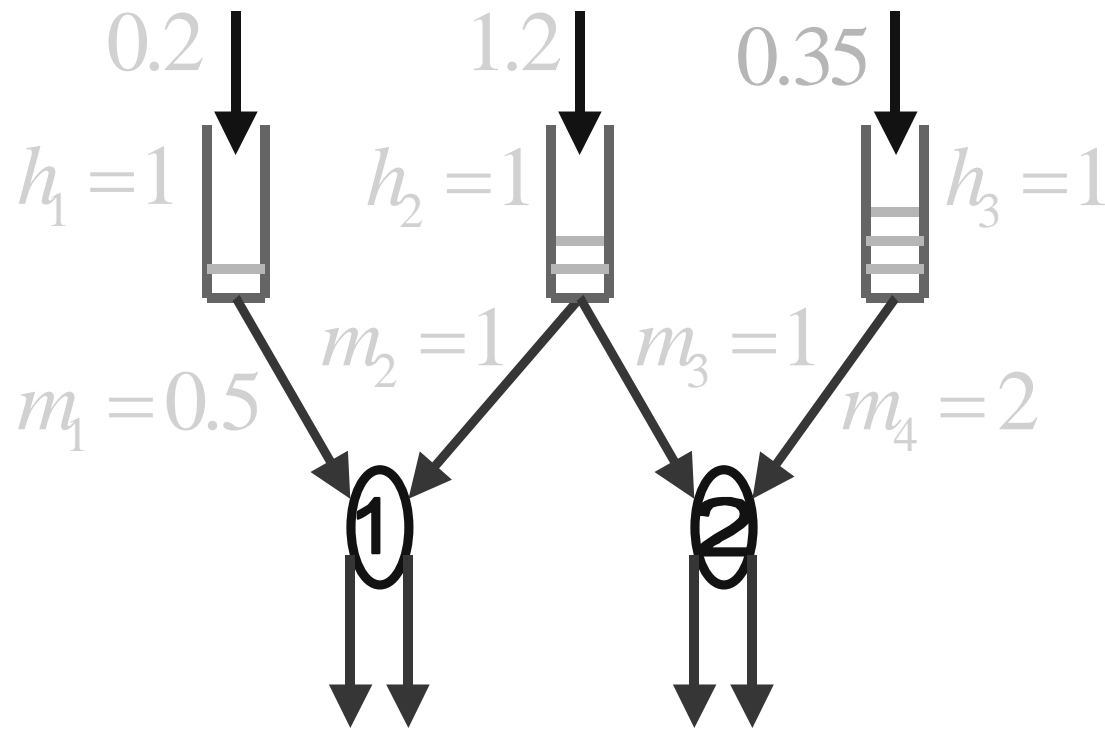
$$\tilde{I}_{k^*}^*(t) = \tilde{V}^*(t) / z_{k^*}^*, \quad \tilde{I}_k^*(t) = 0 \text{ for } k \neq k^*, \quad \tilde{Y}_N^* \equiv 0$$

How can one interpret the solution of the BCP?

Seek a policy that

- (a) keeps the bulk of the work in a buffer i^* with smallest ratio of holding cost to workload contribution,
- (b) incurs idleness only when the system is nearly empty,
- (c) incurs the bulk of the idleness at a server k^* that serves i^* via a basic activity.

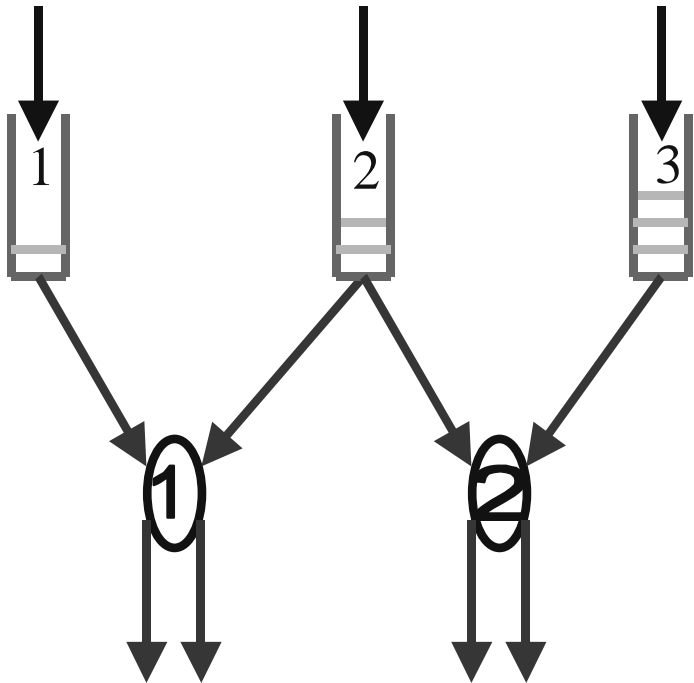
Parallel Server System



$$y^* = (0.25, 0.5, 1), z^* = (0.5, 0.5)$$

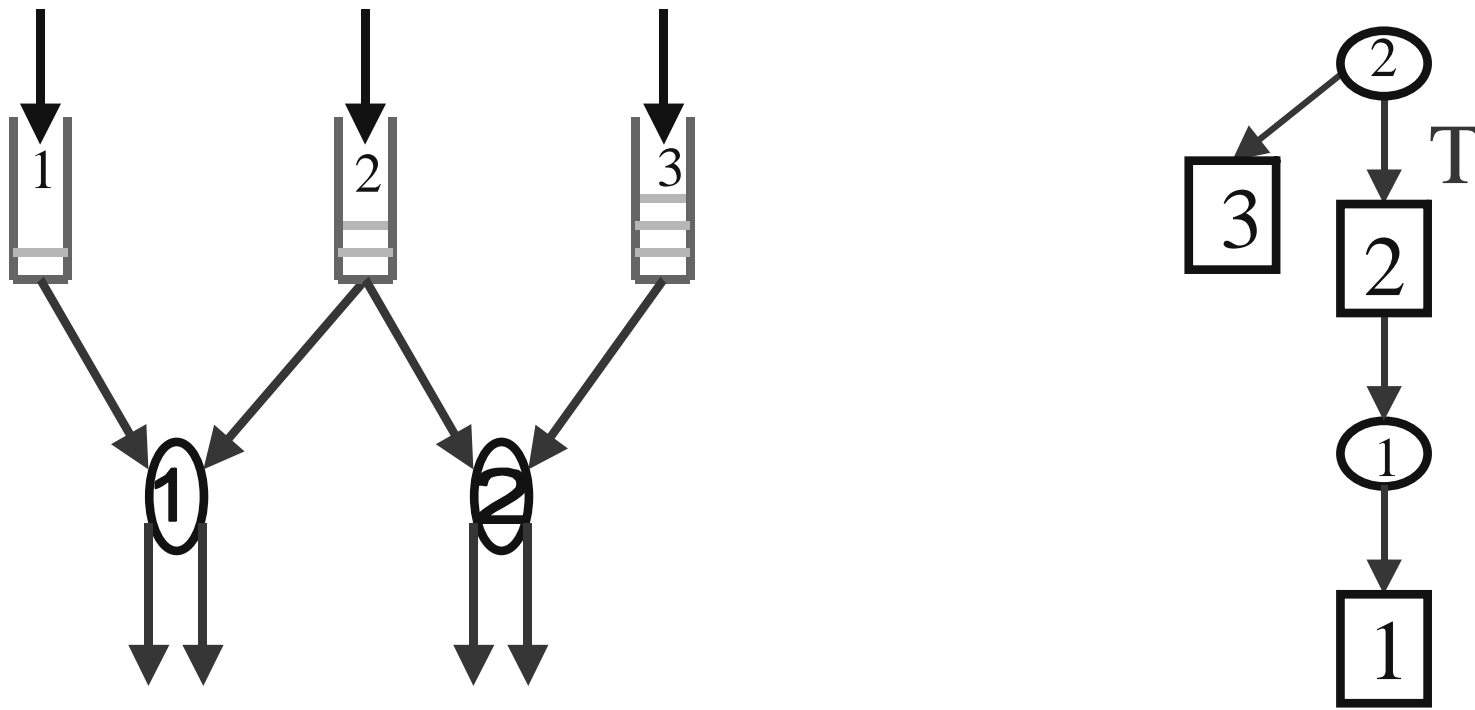
$$i^* = 3, k^* = 2$$

Parallel Server System



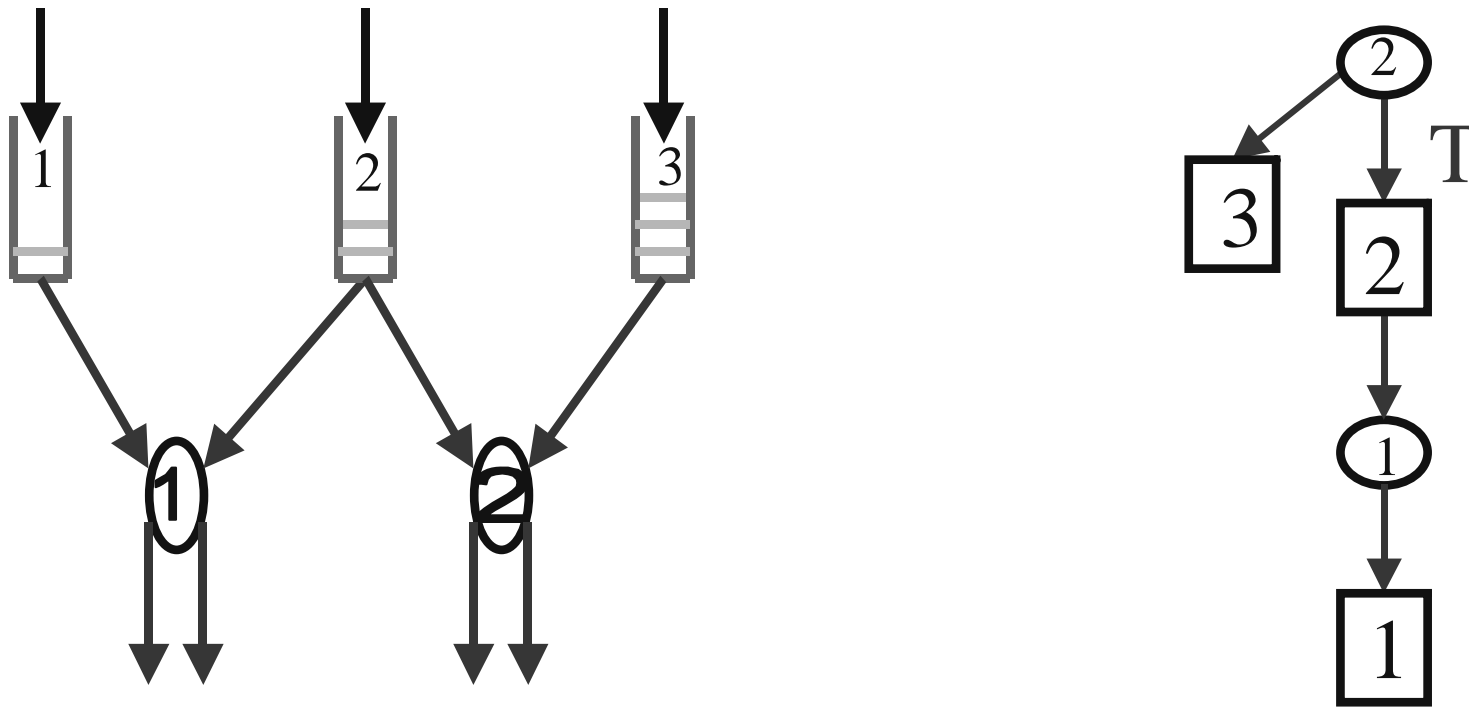
$$i^* = 3, k^* = 2$$

Parallel Server System



$$i^* = 3, k^* = 2$$

Parallel Server System



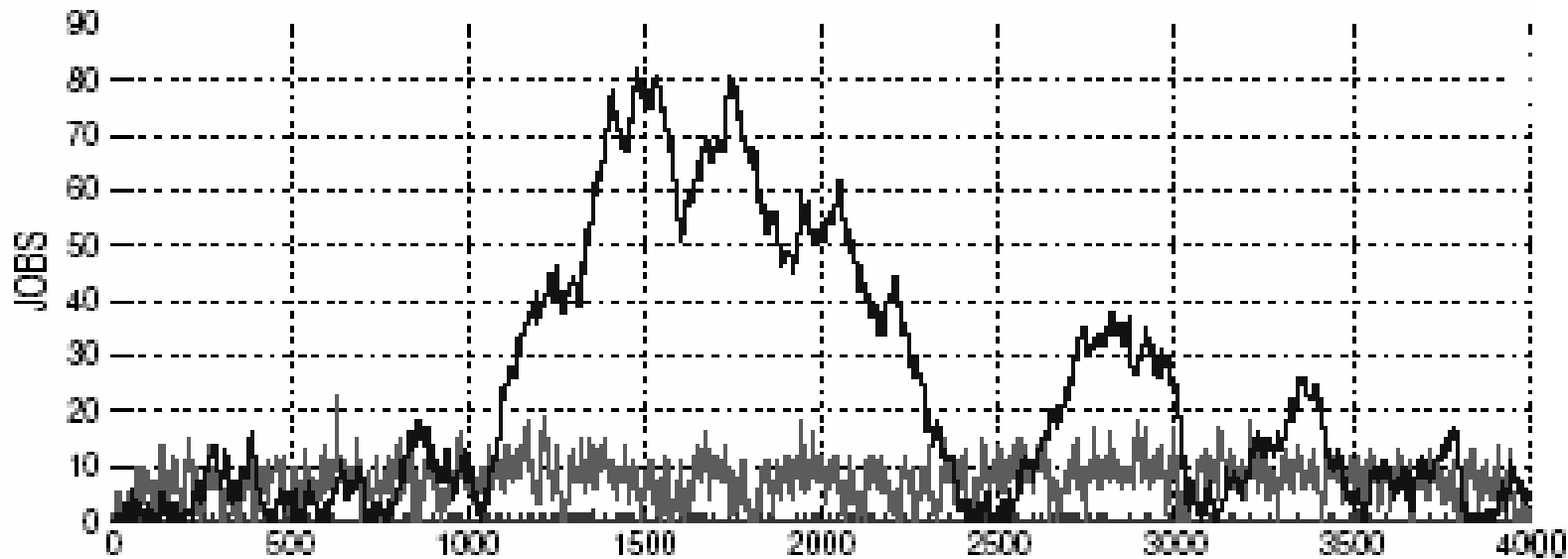
$$i^* = 3, k^* = 2$$

Threshold policy: Server 2 is the root. Buffer 3 has lowest priority. Place threshold on transition activity. Servers give priority to transition activities below them in the tree, except suspend such an activity when the associated buffer is below threshold. Next priority goes to non-transition activities below server. Lowest priority goes to activities *above* server.

Parallel Server System

Simulation with dynamic priority discipline:

server 1 gives priority to buffer 1, server 2 gives priority to buffer 2, except when queue 2 goes below threshold of size 10



Queue lengths for buffer 1 ---, buffer 2 ---, buffer 3 --- versus time

Asymptotic optimality of tree based threshold policy

Assume complete resource pooling, thresholds of order $\log r$, exponential moments and usual heavy traffic conditions.

Theorem (Bell-W '05)

If $T^{r,*}$ denotes the threshold control in the r th system then for any other sequence of control policies $\{T^r\}$, we have

$$\begin{aligned} \liminf_{r \rightarrow \infty} J^r(T^r) &\geq J^* = \lim_{r \rightarrow \infty} J^r(T^{r,*}) \\ &= \mathbb{E} \left[\int_0^\infty e^{-\gamma t} h \bullet \tilde{Q}^*(t) dt \right] \end{aligned}$$

where

$$J^r(T^r) = \mathbb{E} \left[\int_0^\infty e^{-\gamma t} h \bullet \hat{Q}^r(t) dt \right]$$

Complete Resource Pooling: related work

Parallel server system (linear holding costs):

- Proof of asymptotic optimality of discrete review policy in special two server case: Harrison '98
- Proof that continuous review threshold policy is asymptotically optimal: Bell-W. '05

Parallel server system (strictly convex holding costs):

- Stolyar '04, Mandelbaum-Stolyar '04

General network (with feedback):

- Discrete review policy and proof of asymptotic optimality: Ata-Kumar '05.

Open Problems

■ Control of HL SPNs

- Solve BCP or EWF

(HJB equation: Budhiraja-Atar '06,

Numerical Method: S. Kumar and Muthuraman '04)

- Interpretation of solution of BCP or EWF
- Proofs of asymptotic optimality

■ Performance of HL SPNs

- Theory for limiting diffusions (e.g., for packet switch Kang-W '06, congestion control models Kang-Kelly-Lee-W '06)

■ Control and performance for non-HL SPNs