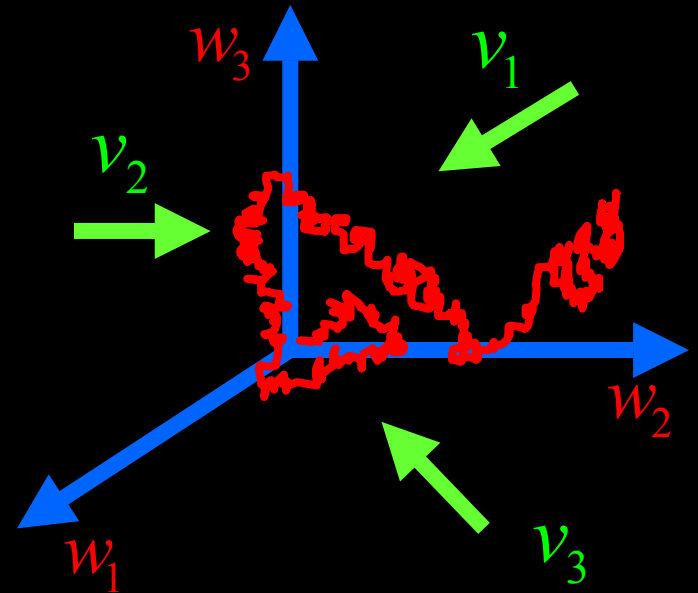
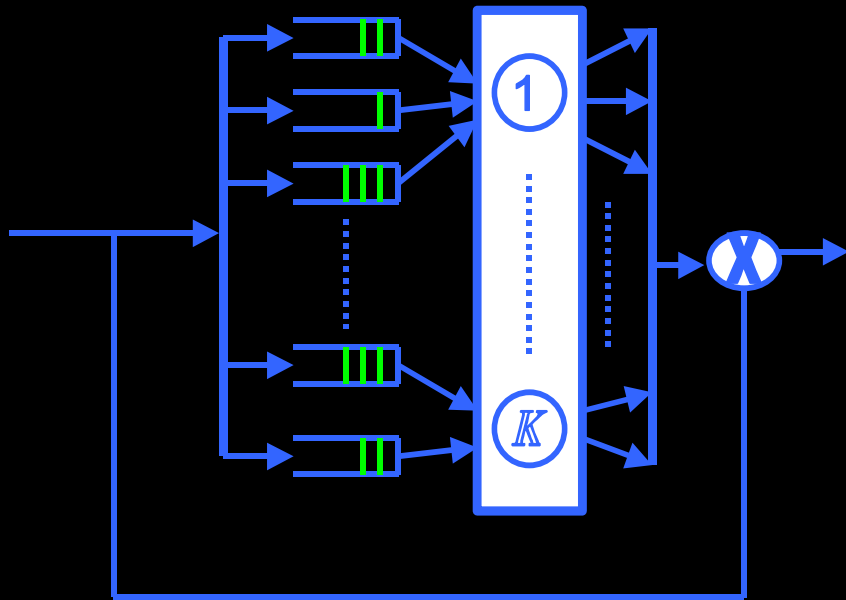


MULTICLASS QUEUEING NETWORKS: PROGRESS AND SURPRISES OF THE PAST 15 YEARS



Ruth J. Williams
University of California, San Diego

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**

Answers

■ STABILITY

- Subcritical fluid models

■ PERFORMANCE ANALYSIS (in heavy traffic)

- Reflecting diffusions and state space collapse via critical fluid models

Scaling

- **Fluid scale:** $\bar{W}^r(t) = W(rt) / r$
- **Diffusion scale:** $\hat{W}^r(t) = W(r^2t) / r = \bar{W}^r(rt)$
- **Performance processes do not require centering for the heavy traffic diffusion approximation. Diffusion scale is obtained by considering large times in fluid scale.**

Assumptions

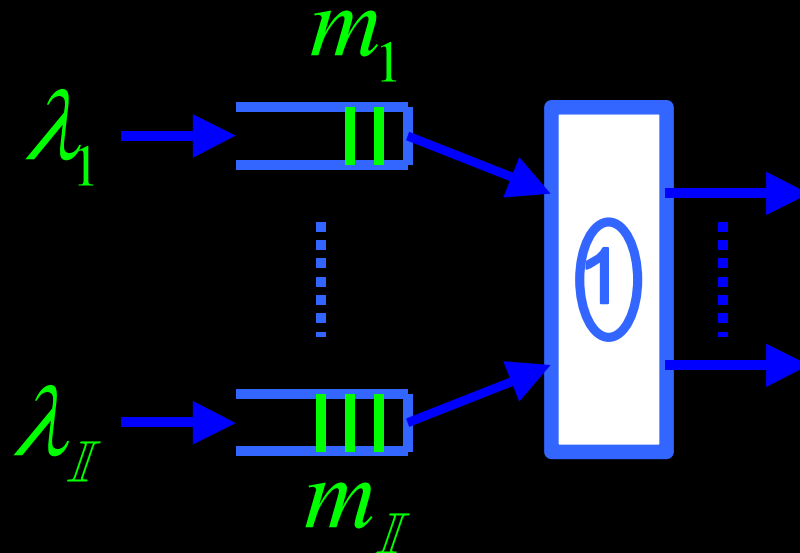
- **HL:** jobs within a buffer are stored in the order in which they arrived and service is always given to the job at the head-of-the-line. Also, the discipline is non-idling.
- Primitive arrival, service and routing processes are assumed to satisfy functional central limit theorems.

Outline

- **SIMPLE EXAMPLE**
- **OPEN MULTICLASS HL NETWORK MODEL**
- **HISTORY UP THROUGH EARLY 1990's**
- **FLUID MODELS AND STABILITY**
- **REFLECTING BROWNIAN MOTIONS**
- **HEAVY TRAFFIC LIMIT THEOREM VIA SSC**
- **FURTHER DEVELOPMENTS**

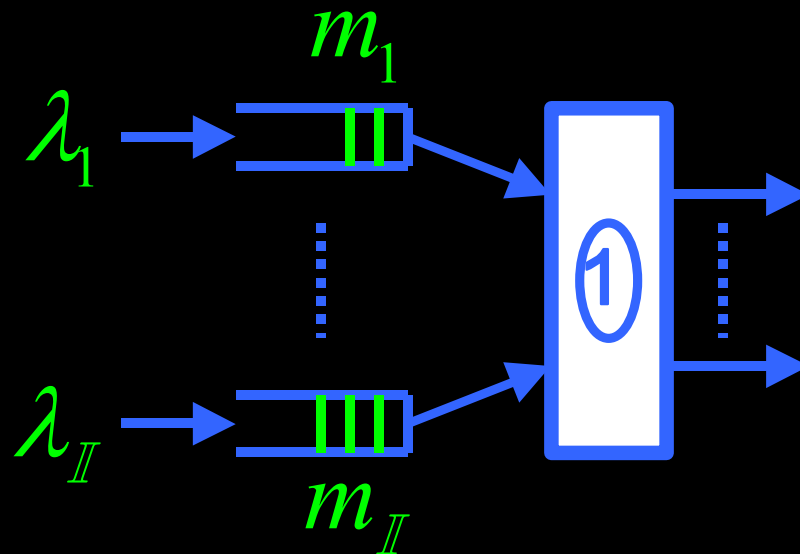
SIMPLE EXAMPLE

Multiclass FIFO Station



- Renewal arrivals to class i at rate λ_i
- i.i.d. service times for class i , mean m_i
- Service discipline: FIFO across all classes

Performance Processes

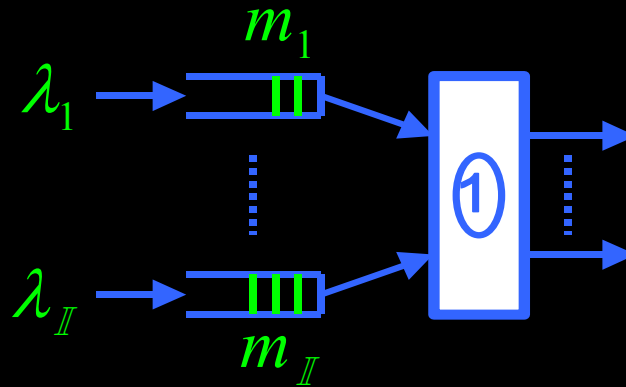


• Queue length for class i : Q_i

• Workload: W

• Idle time: Y

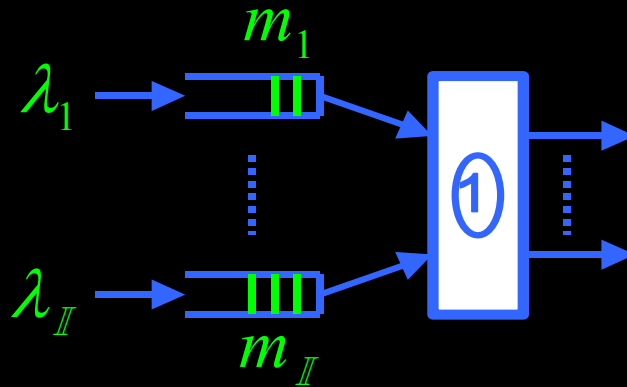
Stability



•Traffic Intensity $\rho_1 = \sum_{i=1}^I \lambda_i m_i$

•Stability iff $\rho_1 < 1$

Stability



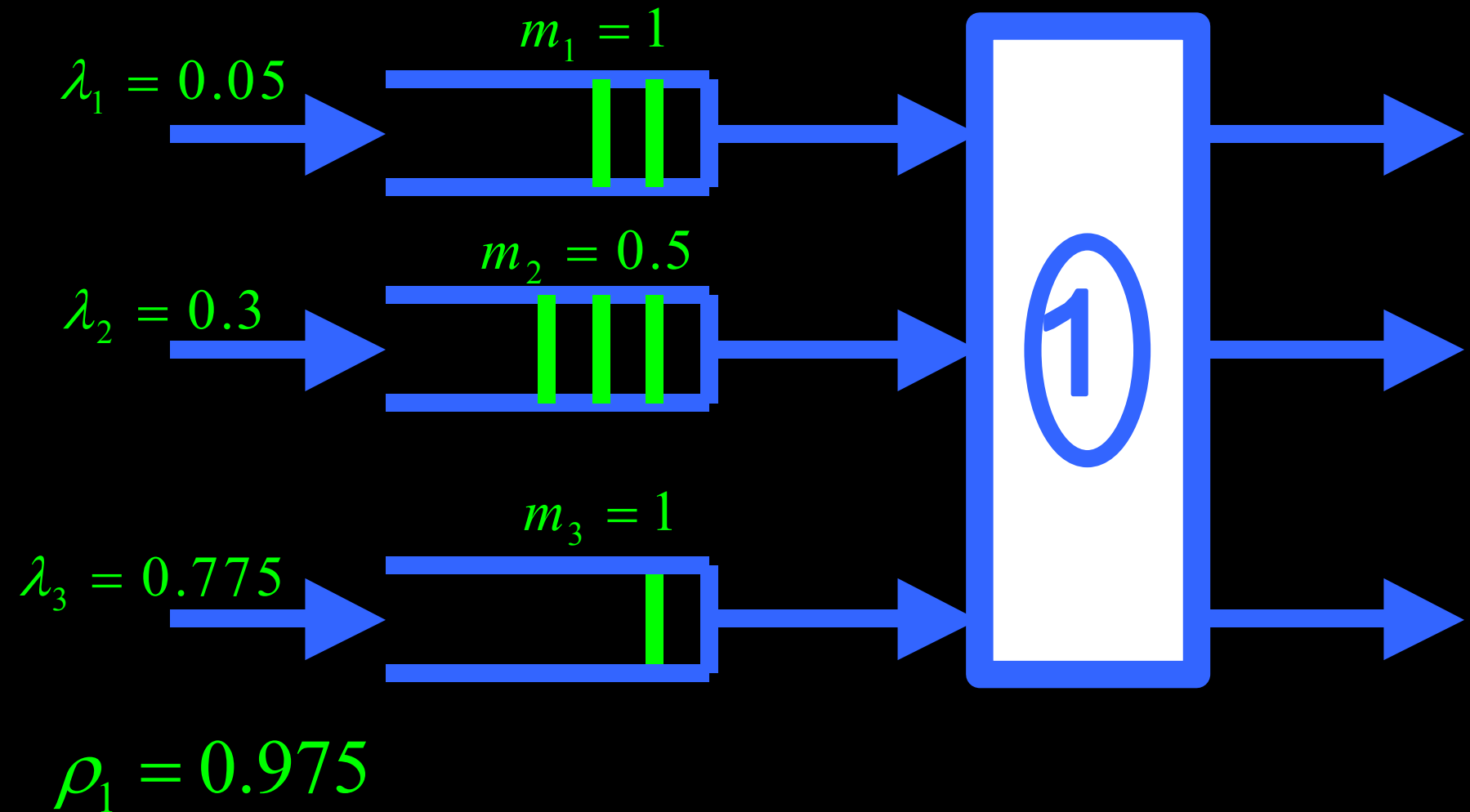
•Traffic Intensity $\rho_1 = \sum_{i=1}^I \lambda_i m_i$

•Stability iff $\rho_1 < 1$

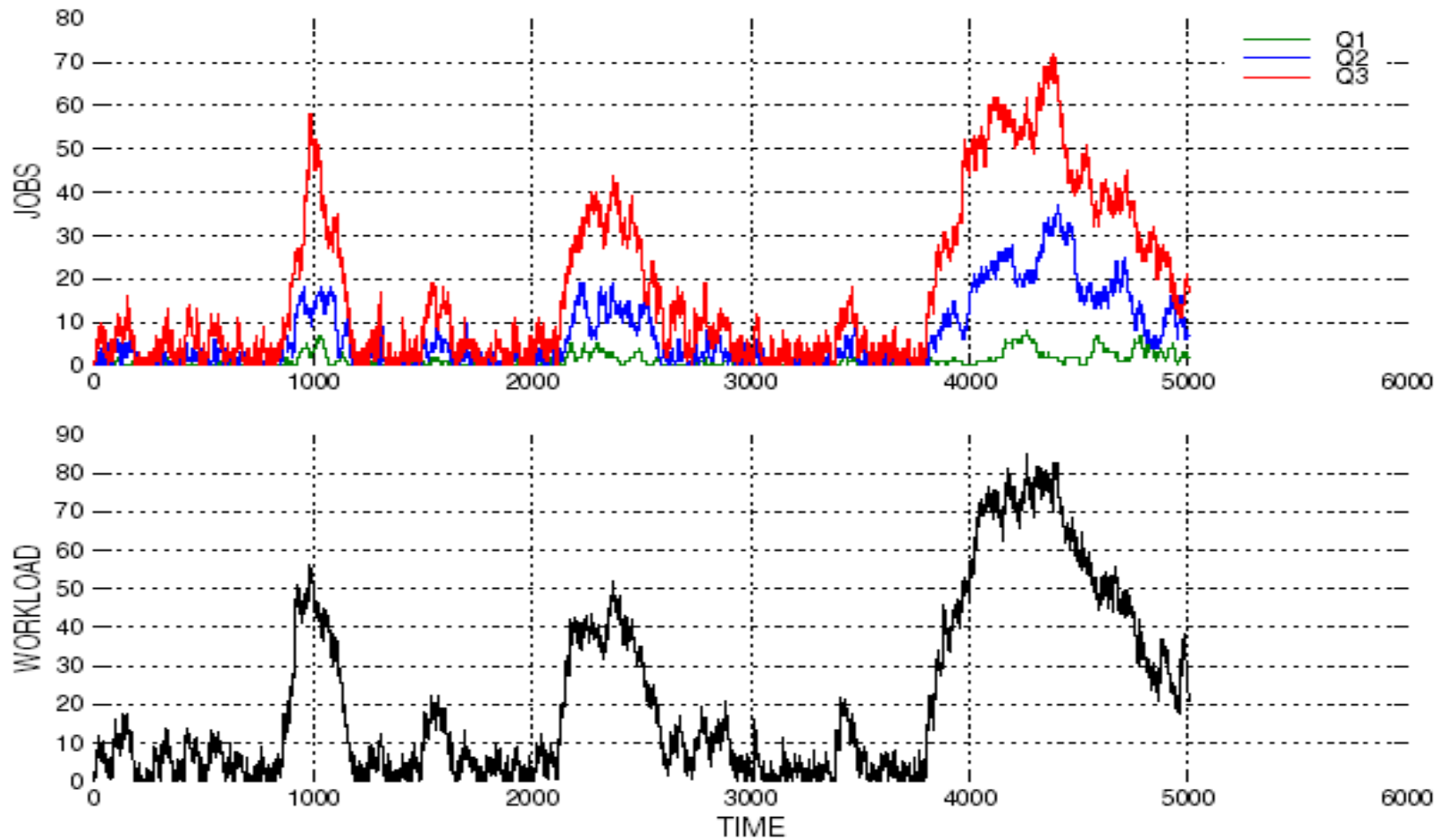
•Heavy traffic $\rho_1 \approx 1$

Simulation of a Multiclass FIFO queue

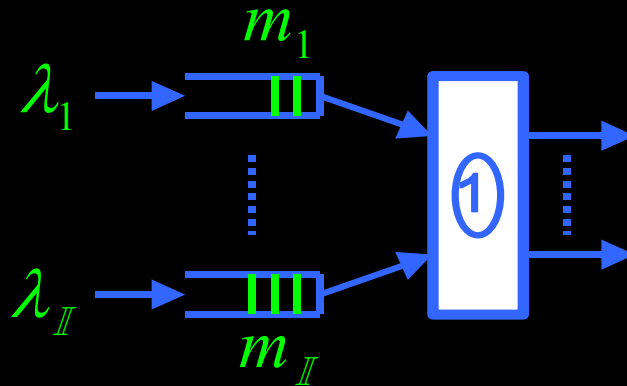
(Poisson arrivals, exponential service times)



Simulation of a Multiclass FIFO queue



Stability



•Traffic Intensity $\rho_1 = \sum_{i=1}^I \lambda_i m_i$

•Stability iff $\rho_1 < 1$

•Heavy traffic $\rho_1 \approx 1$ (assume $\rho_1 = 1$ for simplicity)

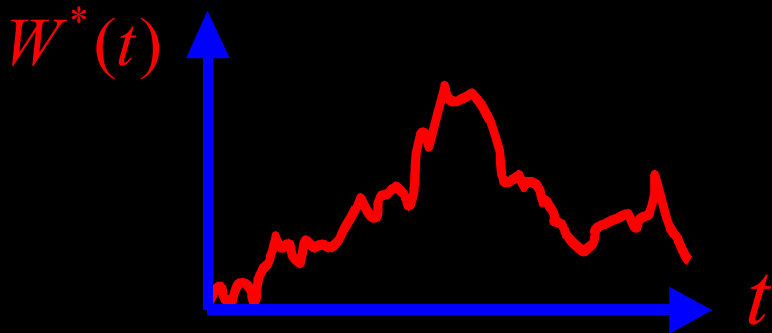
Heavy Traffic Diffusion Approximation

$$\hat{W}^r(t) = W(r^2t)/r, \quad \hat{Y}^r(t) = Y(r^2t)/r,$$

$$\hat{Q}_i^r(t) = Q_i(r^2t)/r, \quad i = 1, \dots, I$$

Theorem (Whitt '71) $(\hat{W}^r, \hat{Y}^r, \hat{Q}^r) \Rightarrow (W^*, Y^*, Q^*)$

where W^* is a one-dimensional reflecting Brownian motion with local time Y^* and $Q^* = \lambda W^*$



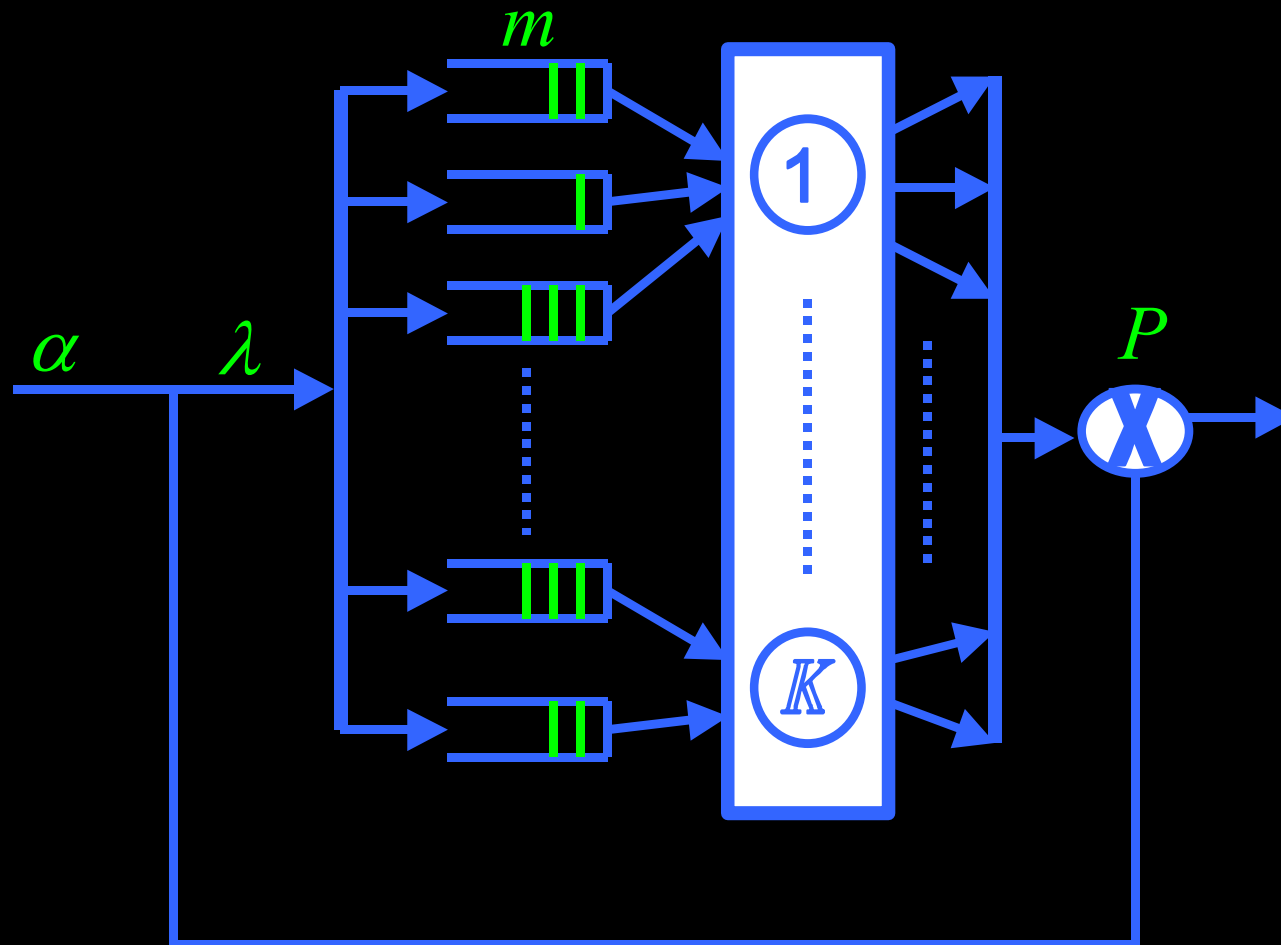
$$W^*(t) = X^*(t) + Y^*(t)$$

$$Y^*(t) = \sup\{-X^*(s) : 0 \leq s \leq t\}$$

X^* = Brownian motion

OPEN MULTICLASS HL NETWORK

Open Multiclass **HL** Queueing Network



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

$$\rho_k = \sum_{i \in k} \lambda_i m_i, \quad k = 1, \dots, K$$

Natural Conjectures

- Stability: Network is stable provided

$$\rho_k < 1 \text{ for each } k = 1, \dots, K$$

- Heavy traffic diffusion approximation:

If $\rho_k \approx 1$, $k = 1, \dots, K$, then $(\hat{W}^r, \hat{Y}^r, \hat{Q}^r) \approx (W^*, Y^*, Q^*)$

where $Q^* = \Delta W^*$ for some $I \times K$ lifting matrix Δ

(that depends on the HL service discipline), and

$W^* = X^* + RY^*$ is a reflecting Brownian motion (RBM) in the K -dimensional orthant.

HISTORY

Affirmative Answers

(Refs. are for diffusion approximations through early 1990s)

■ SINGLE CLASS (FIFO):

- Single station: Iglehart-Whitt ('70)
- Acyclic network: Iglehart-Whitt ('70), Tandem queue: Harrison ('78)
- Network: Reiman ('84), Chen-Mandelbaum ('91)

■ MULTICLASS:

- Single station, priorities: Whitt ('71), Harrison ('73)
- Network, priorities: Johnson ('83, SP), Peterson ('91, feedforward)
- Single station, feedback, round robin & FIFO: Reiman ('88), Dai-Kurtz ('95)

Rely on continuous mapping construction of RBM and do not cover multiclass networks with general feedback.

Counterexamples

(two-station reentrant lines)

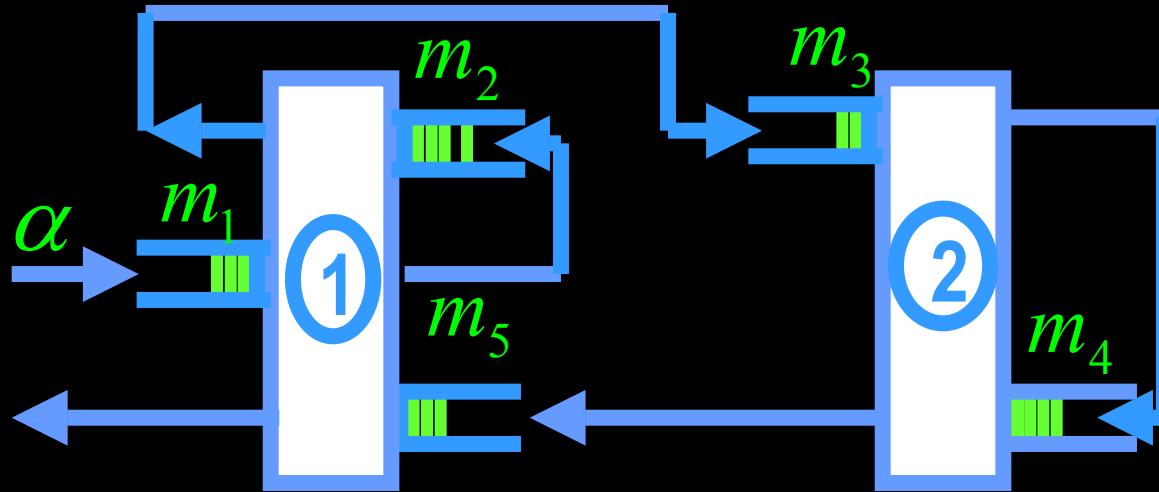
■ STABILITY

- Kumar and Seidman ('90): dynamic policy.
- Lu and Kumar ('91): static priorities, deterministic interarrival and service times.
- Rybko and Stolyar ('92): static priorities, exponential interarrival and service times.
- Seidman ('94): FIFO, deterministic interarrival and service times.
- Bramson ('94): FIFO, exponential interarrival and service times.

■ DIFFUSION APPROXIMATION

- Dai-Wang ('93): FIFO, exponential interarrival and service times.

Dai-Wang '93 FIFO Counterexample

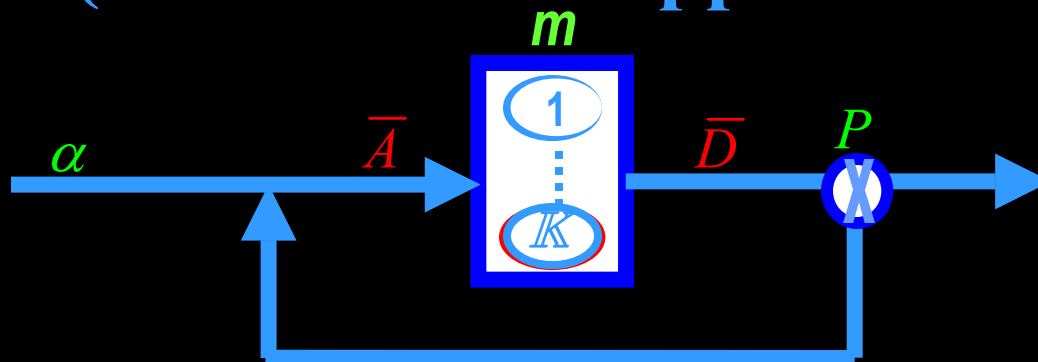


- Poisson arrivals (rate α)
- i.i.d. exponential service times with class means $m = \begin{pmatrix} 1 & 1 & 23 & 4 & 4 \\ 10 & 10 & 27 & 27 & 5 \end{pmatrix}$
- Traffic intensities
 $\rho_1 = \alpha(m_1 + m_2 + m_5) = \alpha$ $\rho_2 = \alpha(m_3 + m_4) = \alpha$
- Proposed workload approximation: $R = \begin{pmatrix} -310 & 16 \\ 27 & -27 \\ 20 & -27 \end{pmatrix}$

STABILITY AND FLUID MODELS

Fluid Model for HL Network

(formal FLLN approximation)

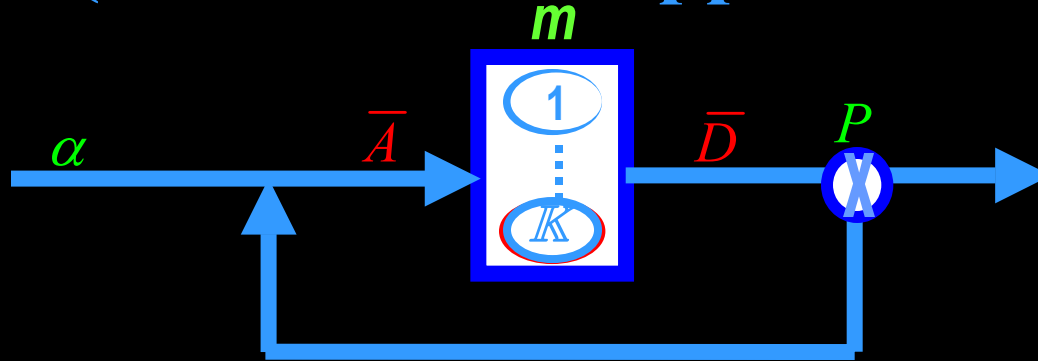


$$\bar{A}(t) = \alpha t + P' \bar{D}(t), \quad M = \text{diag}(m), \quad \bar{D}(t) = M^{-1} \bar{T}(t)$$

$\bar{T}_i(t)$ = total time allocated to class i by t

Fluid Model for HL Network

(formal FLLN approximation)



$$\bar{A}(t) = \alpha t + P' \bar{D}(t), \quad M = \text{diag}(m), \quad \bar{D}(t) = M^{-1} \bar{T}(t)$$

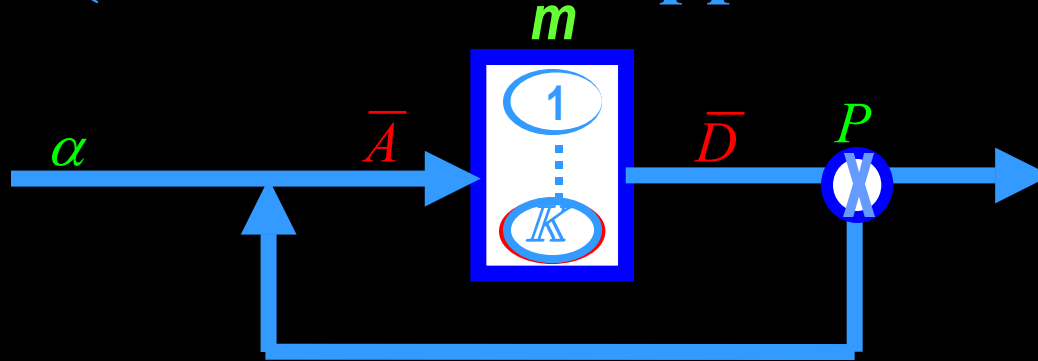
$\bar{T}_i(t)$ = total time allocated to class i by t

$$\bar{W}_k(t) = \bar{W}_k(0) + \sum_{i \in k} m_i \bar{A}_i(t) - t + \bar{Y}_k(t), \quad \bar{Q}_i(t) = \bar{Q}_i(0) + \bar{A}_i(t) - \bar{D}_i(t)$$

$$\bar{Y}_k(t) = t - \sum_{i \in k} \bar{T}_i(t) \text{ is non-decreasing, increases only when } \bar{W}_k = 0$$

Fluid Model for HL Network

(formal FLLN approximation)



$$\bar{A}(t) = \alpha t + P' \bar{D}(t), \quad M = \text{diag}(m), \quad \bar{D}(t) = M^{-1} \bar{T}(t)$$

$\bar{T}_i(t)$ = total time allocated to class i by t

$$\bar{W}_k(t) = \bar{W}_k(0) + \sum_{i \in k} m_i \bar{A}_i(t) - t + \bar{Y}_k(t), \quad \bar{Q}_i(t) = \bar{Q}_i(0) + \bar{A}_i(t) - \bar{D}_i(t)$$

$\bar{Y}_k(t) = t - \sum_{i \in k} \bar{T}_i(t)$ is non-decreasing, increases only when $\bar{W}_k = 0$

+ additional conditions depending on service discipline, e.g., FIFO:

$$\bar{D}_i(t + \bar{W}_k(t)) - \bar{D}_i(t) = \bar{Q}_i(t) \text{ when } i \in k$$

Stability via Fluid Models

Definition: A fluid model is (uniformly) stable if there is $t_0 > 0$ such that $\bar{Q}(t) = 0$ for all $t \geq t_0$ whenever $|\bar{Q}(0)| \leq 1$.

Stability via Fluid Models

Definition: A fluid model is (uniformly) stable if there is $t_0 > 0$ such that $\bar{Q}(t) = 0$ for all $t \geq t_0$ whenever $|\bar{Q}(0)| \leq 1$.

Theorem (Dai '95): Fix an open multiclass HL network and consider an associated fluid model. Under mild conditions*, if the fluid model is stable, then a Markov process describing the multiclass network is positive Harris recurrent.

*include a “spread out” assumption on interarrival and service times

Examples

- Since 1995, many authors have used the fluid model approach to obtain sufficient conditions for stability of open multiclass HL networks (e.g., Bertsimas, Bramson, H. Chen, Dai, Foss, Hasenbein, Meyn, Stolyar, Weiss, ...)
- Bramson '96: FIFO Kelly-type and HLPPS networks (Kelly type: mean service times are station dependent)
 - used subcritical fluid models to establish stability when $\rho_k < 1$ for all k .

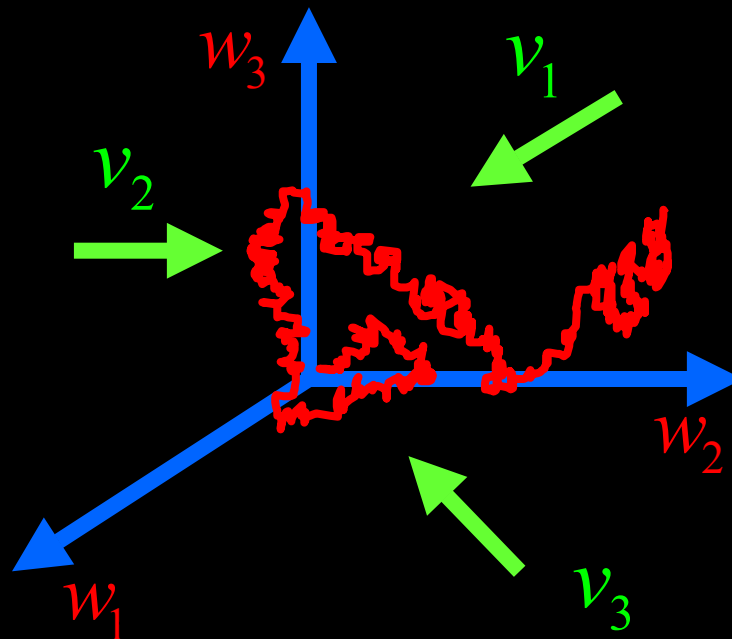
Examples

- Since 1995, many authors have used the fluid model approach to obtain sufficient conditions for stability of open multiclass HL networks (e.g., Bertsimas, Bramson, H. Chen, Dai, Foss, Hasenbein, Meyn, Stolyar, Weiss, ...)
- Bramson '96: FIFO Kelly-type and HLPPS networks (Kelly type: mean service times are station dependent)
 - used subcritical fluid models to establish stability when $\rho_k < 1$ for all k .
 - established asymptotic behavior (as $t \rightarrow \infty$) of critical fluid models ($\rho_k = 1$ for all k) --- uniform convergence to invariant manifold.

**SEMIMARTINGALE REFLECTING
BROWNIAN MOTIONS
(SRBM_s)**

SRBM DATA

- State space: \mathbb{R}_+^K
- Brownian statistics: drift θ , covariance matrix Γ
- Reflection matrix: $R = (v_1, \dots, v_K)$



SRBM DEFINITION (w/starting point x_0)

A continuous \mathbb{K} -dimensional process W such that

(i) $W = X + RY$

(ii) W has paths in $\mathbb{R}_+^{\mathbb{K}}$

(iii) for $k=1, \dots, \mathbb{K}$, $Y_k(0) = 0$, Y_k is continuous,

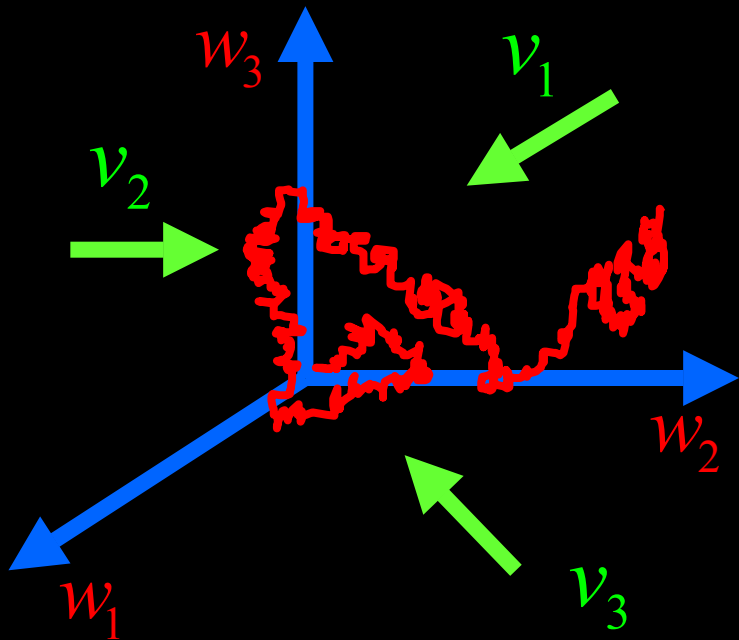
non-decreasing, and it can increase only when $W_k = 0$

(iv) X is a (θ, Γ) BM s.t. $X(0) = x_0$,

$\{X(t) - \theta t, t \geq 0\}$ is a martingale relative to the filtration generated by (W, X, Y)

Necessary Condition for Existence

Defn: R is completely-S iff for each principal submatrix \tilde{R} of R there is $\tilde{y} > 0 : \tilde{R}\tilde{y} > 0$



$$R = (v_1, \dots, v_K)$$

Existence and Uniqueness in Law

Theorem (Reiman-W '88, Taylor-W '93)

There is an SRBM W starting from each point x_0 in \mathbb{R}_+^K iff R is completely-S. In this case, each such SRBM is unique in law and these laws define a continuous strong Markov process.

Analysis of multidimensional SRBMs

■ Sufficient conditions for positive recurrence

Dupuis-W '94, Chen '96, Budhiraja-Dupuis '99, El Kharroubi-Ben Tahar-Yaacoubi '00

■ Stationary distribution

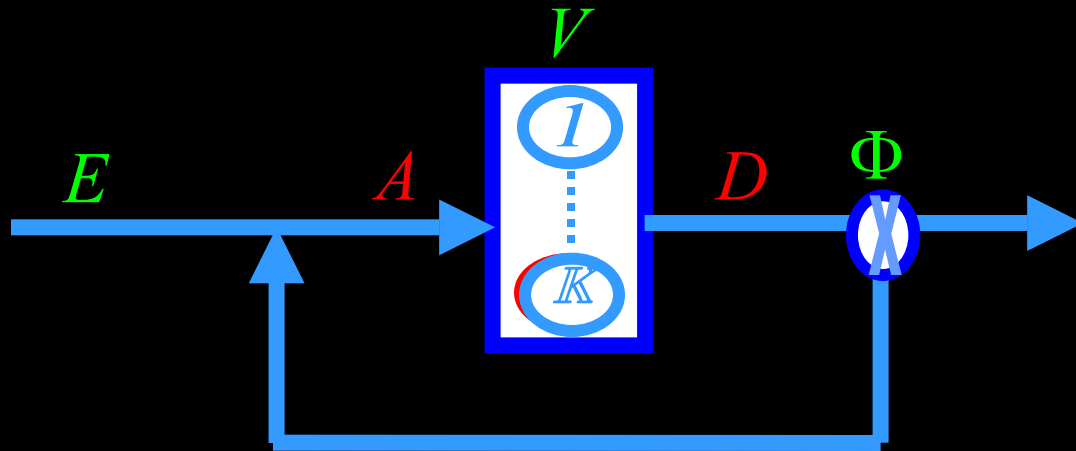
- *Characterization*: Harrison-W '87, Dai-Harrison '92, Dai-Kurtz '98
- *Analytic solutions -two-dimensions*: Foddy '84, Trefethen-W '86
 - *-product form*: Harrison-W '87
- *Numerical methods*: Dai-Harrison '91,'92, Shen-Chen-Dai-Dai '02, Schwerer '01

■ Large deviations

Majewski '98,'00, Dupuis-Ramanan '02

HEAVY TRAFFIC LIMIT THEOREM VIA STATE SPACE COLLAPSE

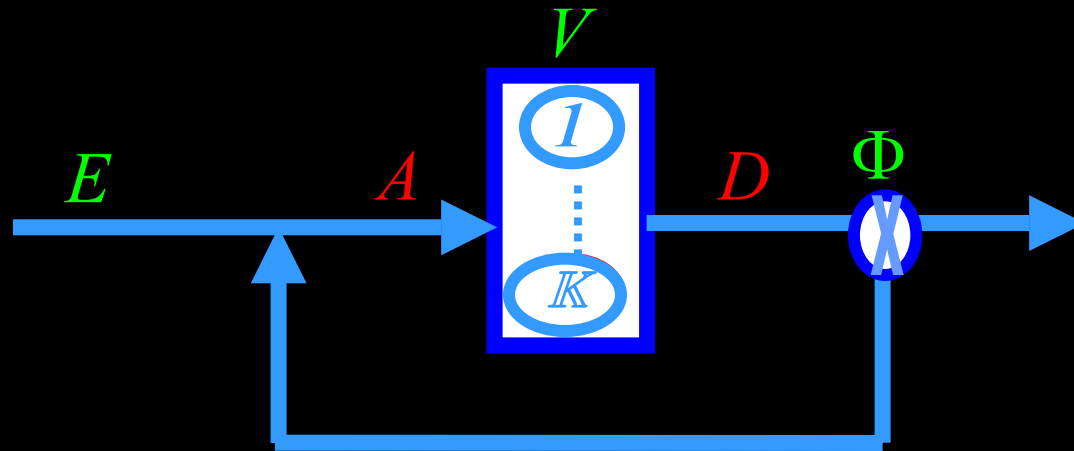
Heavily Loaded Multiclass HL Network Stochastic Model



Start system empty

Assume $\rho_k = 1 \forall k$

Heavily Loaded Multiclass HL Network Stochastic Model



Start system empty

Assume $\rho_k = 1 \forall k$

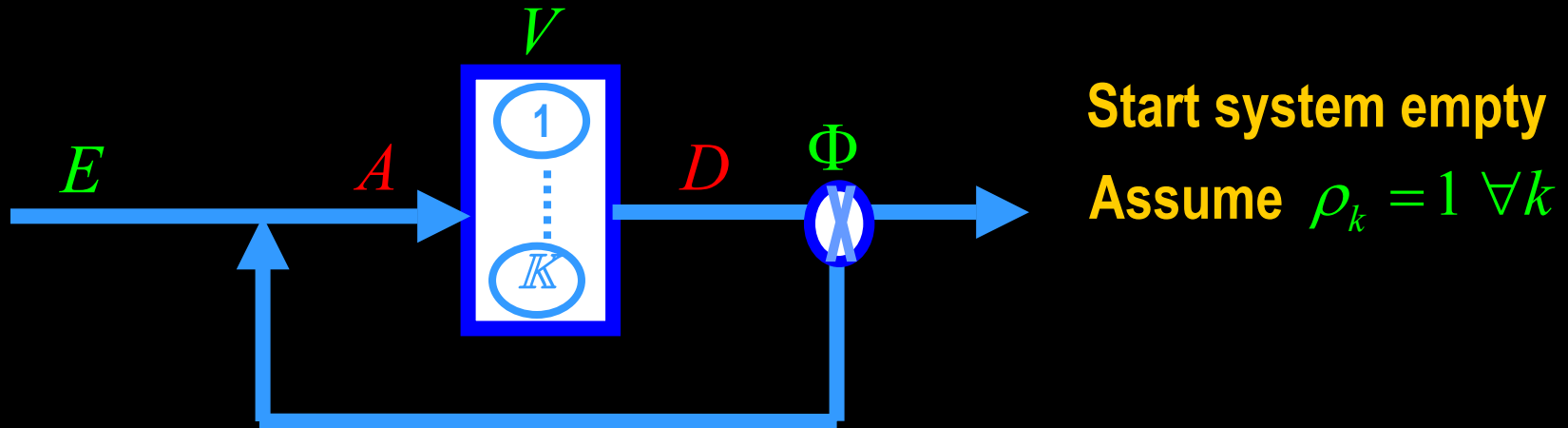
$$A_i(t) = E_i(t) + \Phi_i(D(t))$$

$$D_i(t) = S_i(T_i(t))$$

$$Q_i(t) = A_i(t) - D_i(t)$$

$$W_k(t) = \sum_{i \in k} V_i(A_i(t)) - t + Y_k(t)$$

Heavily Loaded Multiclass HL Network Stochastic Model



$$A_i(t) = E_i(t) + \Phi_i(D(t)) \quad D_i(t) = S_i(T_i(t))$$

$$Q_i(t) = A_i(t) - D_i(t)$$

$$W_k(t) = \sum_{i \in k} V_i(A_i(t)) - t + Y_k(t)$$

+ additional equations depending on service discipline, e.g., FIFO:

$$D_i(t + W_k(t)) - D_i(t) = Q_i(t) \text{ when } i \in k$$

State Space Collapse

Definition: Multiplicative state space collapse (MSSC) holds if there is a $\mathbb{R} \times \mathbb{K}$ matrix Δ such that for each $T \geq 0$:

$$\frac{\|\hat{Q}^r(\cdot) - \Delta \hat{W}^r(\cdot)\|_T}{\max(\|\hat{W}^r(\cdot)\|_T, 1)} \rightarrow 0$$

in probability as $r \rightarrow \infty$.

State Space Collapse

Definition: Multiplicative state space collapse (MSSC) holds if there is a $\mathbb{R} \times \mathbb{K}$ matrix Δ such that for each $T \geq 0$:

$$\frac{\|\hat{Q}^r(\cdot) - \Delta \hat{W}^r(\cdot)\|_T}{\max(\|\hat{W}^r(\cdot)\|_T, 1)} \rightarrow 0$$

in probability as $r \rightarrow \infty$.

Theorem (Bramson '98) “MSSC holds if critical fluid model solutions converge uniformly to the invariant manifold”.
In particular, MSSC holds for FIFO Kelly type and HLPPS networks.

Sufficient Conditions for HT Limit Theorem

Theorem (W '98) Assume standard heavy traffic assumptions and

(i) multiplicative state space collapse,

(ii) the reflection matrix R is completely-S.

Then $(\hat{W}^r, \hat{Y}^r, \hat{Q}^r) \Rightarrow (W^*, Y^*, Q^*)$

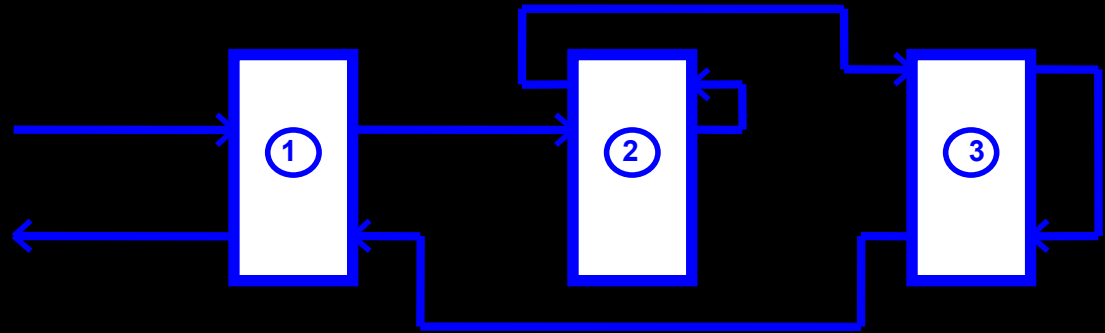
where W^* is an SRBM with pushing process Y^* and

$Q^* = \Delta W^*$.

Examples: FIFO Kelly type and HLPPS networks; FBFS, LBFS reentrant lines; some static priority networks (see e.g., Bramson '98, W '98, Bramson-Dai '01)

Dai-Wang-Wang '92 example

A multiclass FIFO
network of Kelly type



Renewal arrivals (rate α), i.i.d. service times for each class

Assume $m_1 = m_2 = m_3 = m_4 = m_5 = m_6 = m$

Traffic intensities $\rho_1 = \rho_2 = \rho_3 = 2\alpha m$

Reflection matrix for SRBM approximation
to workload process

$$R = \begin{pmatrix} 1 & \frac{2}{5} & -\frac{6}{5} \\ -\frac{2}{3} & 1 & \frac{4}{5} \\ \frac{2}{9} & -\frac{8}{5} & 1 \end{pmatrix} \Lambda$$

No continuous mapping constr. for SRBM

FURTHER DEVELOPMENTS

Some Related Work on Diffusion Approximations for Multiclass Queueing Networks

■ HT limits that are not SRBMs (& have no state space collapse)

- Single station-polling: Coffman-Puhalskii-Reiman '95
- Dynamic HLPS: Ramanan and Reiman '03

■ Non-HL service disciplines

(Markovian state descriptor is typically infinite dimensional)

- LIFO preemptive resume: Single station: Limic '00, '01
- Processor sharing: Single station (Gromoll-Puha-W '01, Puha-W '03, Gromoll '03); network (stability: Bramson '04)
- EDF: Single station (Doytchinov-Lehoczky-Shreve '01), acyclic network (Kruk-Lehoczky-Shreve-Yeung '03), network (stability: Bramson '01)

