CONTROL FORMULATIONS IN THE NONDEGENERATE SLOWDOWN DIFFUSION REGIME

Rami Atar

Technion, Israel

atar@ee.technion.ac.il

with

Itai Gurvich

Northwestern, USA

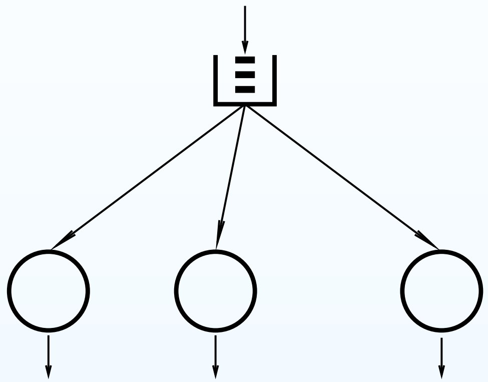
Nir Solomon

Technion, Israel

I. Introduction

Heavy traffic diffusion regimes

Consider a queue with multiple servers



Parametrize by letting

$$\lambda_n \approx n, \qquad N_n \approx n^\alpha, \qquad \mu_n \approx n^{1-\alpha},$$

where $0 \le \alpha \le 1$, so that $\lambda_n \approx N_n \mu_n$.

Obtain:

$\boxed{\alpha = 0}$ ———— $\boxed{\alpha = 1}$

Conventional — Halfin-Whitt

Define slowdown=sojourn time / service time

Slowdown is degenerate at both endpoints

When is the slowdown nondegenerate?

Consider $\alpha = 1/2$.

$$\lambda_n \approx n, \qquad N_n \approx n^{1/2}, \qquad \mu_n \approx n^{1/2}$$

Clearly the service time $\approx n^{-1/2}$

Obtain

DELAY $\quad \sim \quad$ SERVICE TIME

Earlier work (the case of M/M/N):

- * Whitt (Oper. Res., 2003): Convergence of queuelength and delay processes to a RBM ($\alpha = 1/2$)
- * Mandelbaum and Shaikhet (Mandelbaum's EURANDOM lecture notes, 2003): independently, a similar result, ($\alpha = 1/2$); observe that the delay and the time in service are of the same order
- * Gurvich (M.Sc. Thesis, 2004): Convergence of queuelength/ delay processes to a RBM for $\alpha \in [\frac{1}{2}, 1)$.

The above works regard this as a part of the Efficiency Driven regime (the diffusion being RBM, the probability of delay being close to 1)

Our point of view

* The joint law of delay and time in service is interesting

* $\alpha = 1/2$ is the only case where the limit is a nondegenerate pair of processes

* The limiting joint law (and in particular the limiting sojourn time law) is distinct from that under the other two diffusion regimes

We will refer to it as the Non-Degenerate Slowdown (NDS) regime

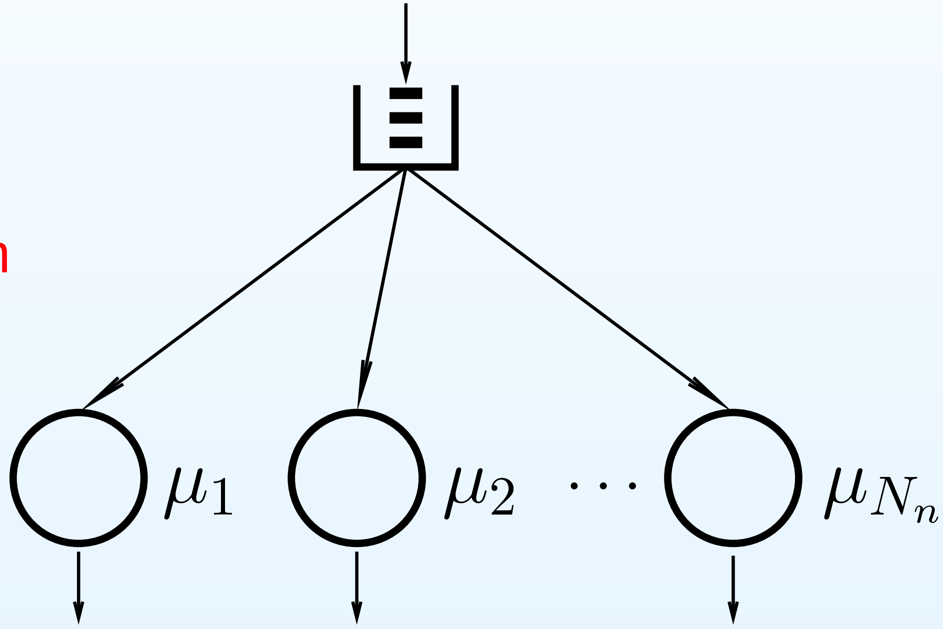
II. Some diffusion limit results

Model

renewal arrivals
each requiring a single
non-interruptible service

routing mechanism

N_n heterogenous
exponential
servers



Assumptions. The NDS Regime ($\alpha = 1/2$)

- Arrivals: $\lambda_n = \lambda n + \hat{\lambda} n^{1/2} + o(n^{1/2})$
- Number of servers $N_n = n^{1/2} + o(n^{1/2})$
- Individual service rates $\mu_{1n}, \mu_{2n}, \ldots, \mu_{N_n\, n}$
- With $\mu_n = \sum_{k=1}^{N_n} \mu_{kn}, \quad n^{-1}\mu_n \to \mu \in (0, \infty)$
  $\hat{\mu}_n = n^{-1/2}(\mu_n - n\mu) \to \hat{\mu} \in (-\infty, \infty)$
- Critical load condition: $\lambda = \mu$

Assumptions (cont.)

- The empirical measure of $\{\hat{\mu}_{kn} := \mu_{kn} n^{-1/2}\}$ converges weakly, namely

$$\frac{1}{N_n}\sum_{k=1}^{N_n} \delta_{\hat{\mu}_{kn}} \to m,$$

for some probability measure m on $\mathbb{R}_+$

Assumptions on the routing policy

- Work conserving
- Nonanticipating

Includes, for example,

- Always route to the slowest available server
- Always route to the fastest available server

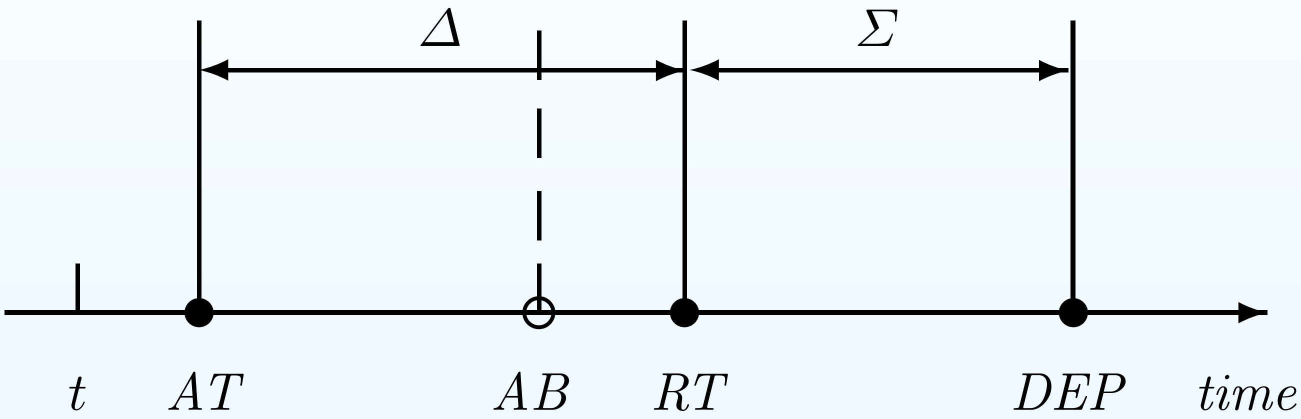
Processes of interest

$\varDelta_n(t)$= delay experienced by the first customer to arrive at or after time t

$\varSigma_n(t)$= time in service of the same customer

Diffusion scaling:

$$\widehat{\varDelta}_n = n^{1/2}\varDelta_n \qquad \widehat{\varSigma}_n = n^{1/2}\varSigma_n$$



AT = Arrival Time
RT = Routing Time
DEP = Departure Time
AB = Abandonment Time
Δ = Delay
Σ = Service Time

Diffusion-scale limit result

<u>THEOREM:</u> The joint law of $(\widehat{\Delta}_n, \widehat{\Sigma}_n)$ converges to

(RBM, f-White noise)

in finite dimensional distributions.
That is, given j and $0 < t_1 < \cdots < t_j < \infty$, we have

$$(\widehat{\Delta}_n(t_1), \widehat{\Sigma}_n(t_1), \dots, \widehat{\Delta}_n(t_j), \widehat{\Sigma}_n(t_j)) \Rightarrow (\bar{\xi}(t_1), \eta_1, \dots, \bar{\xi}(t_j), \eta_j),$$

were, $\bar{\xi} = \xi/\mu$, ξ is the RBM

$$\xi(t) = \xi_0 + (\widehat{\lambda} - \widehat{\mu})t + \sigma w(t) + l(t),$$

and η_i are independent of ξ, i.i.d., with p.d.f.

$$f(x) = \frac{1}{\mu} \int y^2 e^{-yx} m(dy), \quad x \in [0, \infty).$$

Interpretation of f

* Draw a random variable Y from the distribution

$$\frac{ym(dy)}{\int zm(dz)},$$

* Let η be exponentially distributed with mean Y.

Extension to case with abandonment

Customers abandon the queue while waiting to be served, at fixed rate γ (according to an exponential clock).

The result holds, with

$$\xi(t) = \xi_0 + (\widehat{\lambda} - \widehat{\mu})t - \gamma \int_0^t \xi(s)ds + \sigma w(t) + l(t)$$

Expressions for slowdown (formal)

Without abandonment ($\gamma = 0$) need to assume $\widehat{\lambda} - \widehat{\mu} < 0$, and then

$$\text{slowdown} = 1 + \frac{\sigma^2}{2(\widehat{\mu} - \widehat{\lambda})}$$

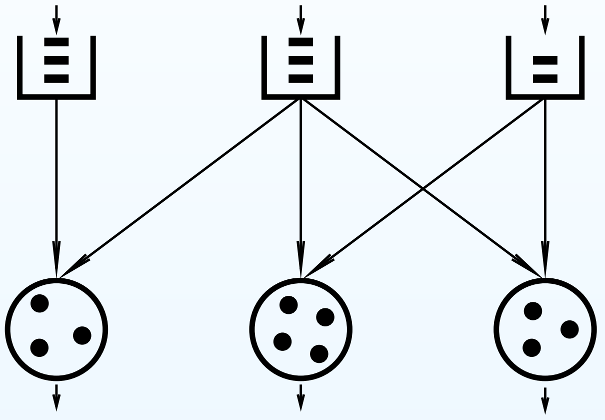
With abandonment ($\gamma > 0$)

$$\text{slowdown} = 1 + \frac{\int_0^\infty x e^{-(x-b)^2/2c^2} dx}{\int_0^\infty e^{-(x-b)^2/2c^2} dx}$$

where $(b, c^2) = (\frac{\widehat{\lambda}-\widehat{\mu}}{\gamma}, \frac{\sigma^2}{2\gamma})$.

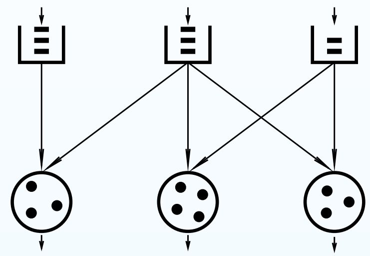
III. Control formulations

Control to minimize sojourn time



- As a diffusion-limited control problem, this set us is meaningful only in the NDS regime

The heavy traffic condition



Following Harrison and Lopez (1999), consider the linear program

Minimize $\rho \in [0,1]$ s.t. $\displaystyle\sum_j \mu_{ij}\xi_{ij} = \lambda_i,\ \forall i, \quad \xi_{ij} \ge 0,\ \forall (i,j), \quad \sum_i \xi_{ij} \le \rho,\ \forall j$

The HT condition: There exists a unique optimal solution (ξ^*, ρ^*), $\rho^* = 1$. Moreover, $\sum_i \xi^*_{ij} = 1$

The complete resource pooling condition

$i \sim j$ — an activity
$\xi^*_{ij} > 0$ — a basic activity

The CRP condition:
* Uniqueness of solutions to a dual program (Harrison and Lopez 1999)
* The graph $\mathcal{G}_b$, of basic activities, is connected (Harrison and Lopez 1999)
* The graph $\mathcal{G}_b$ is a tree (Williams 2000)

Significance:

* High level of cooperation between service stations, so stations work like a single super-server

* Workload is one-dimensional

The diffusion scaling

Denote

$Q_i^n(t)$ = number of class-i customers in the queue at time t

$X_i^n(t)$ = number of class-i customers in the system at time t

$$\hat{Q}_i^n(t) = n^{-1/2} Q^n(t), \qquad i = 1, 2, \ldots, I$$

$$\hat{X}_i^n(t) = n^{-1/2}\Big(X_i^n(t) - \sum_j \xi_{ij}^* N_j^n\Big), \qquad i = 1, 2, \ldots, I$$

The diffusion control problem (Harrison-Lopez 1999)

The DCP consists of r.v.s $X_{0,i}$, BMs W_i, and processes X_i, I_j, Y_{ij}:

$$X_i(t) = X_{0,i} + W_i(t) + \sum_j \mu_{ij} Y_{ij}(t) \geq 0, \qquad t \geq 0, i = 1, 2, \ldots, I,$$

$$I_j := \sum_i Y_{ij} \text{ is non-decreasing and } I_j(0) \geq 0, \qquad j = 1, 2, \ldots, J,$$

$$Y_{ij} \text{ is non-increasing and } Y_{ij} \leq 0, \qquad (i,j) \in \mathcal{E}_{nb}.$$

REM: Y_{ij} are further required in Harrison-Lopez to be adapted; one can drop this requirement (Bell-Williams 2000)

An equivalent DCP

Harrison-Lopez 1999, Mandelbaum-Stolyar 2004

$$X(t) = X_0 + W(t) + Z(t) \in \mathbb{R}^I_+, \qquad t \geq 0,$$

$$\theta' Z \text{ is nondecreasing, and } \theta' Z(0) \geq 0$$

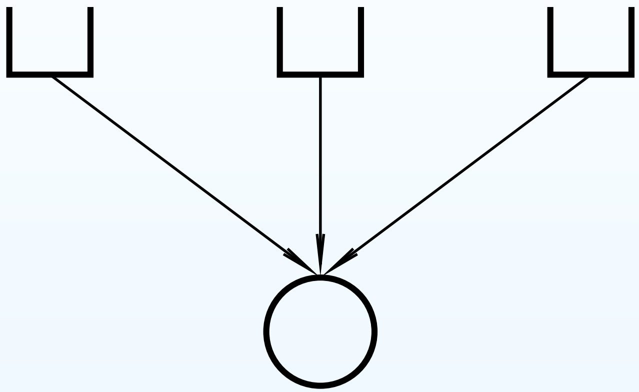
Here, $\theta \in \mathbb{R}^I_+$ is a fixed vector (the workload vector).

<u>THEOREM (with Itai Gurvich):</u> The two diffusion control problems are equivalent.

IV. DCP for sojourn time - an explicit solution

DCP for sojourn time

CASE OF A SINGLE POOL



* Nonlinear cost is of interest

We will consider COST = $\sum_i c_i \mathbb{E}\left[\left(\frac{X_i(t)}{\mu_i} + \Sigma_i\right)^2\right]$

Σ_i -r.v.s representing service time

* Easy to reduce to $\mathbb{E}[C(X(t))]$

Solution of DCP

Denote

$$\rho_i = \frac{\lambda_i}{\mu_i}, \qquad \beta_i = \frac{\rho_i^2}{c_i}, \qquad i = 1, 2, \ldots, I$$

THEOREM (with Nir Solomon): The DCP is solved by bringing $X(t)$ to $X^*(t)$ s.t.

$$\frac{X_i^* + \rho_i}{\mu_i} = \frac{\beta_i}{\sum_k \beta_k} \sum_k \frac{X_k^* + \rho_k}{\mu_k}, \qquad \text{for all } i$$

V. Asymptotics

Asymptotics, the conventional regime

BACK TO THE GENERAL CASE (general number of pools, $J \geq 1$; general cost C)

In conventional heavy traffic:

* Ata-Kumar (2005) - a discretization approach

* Bell-Williams (2001, 2005) - a threshold policy

* Mandelbaum-Stolyar (2004) - a generalized $c\mu$ rule

Asymptotics, the NDS regime

Let $C : \mathbb{R}_+^I \to \mathbb{R}_+$ be a continuous function, increasing wrt usual partial order

$$C^*(a) = \min\{C(q) : q \in \mathbb{R}_+^I,\ \theta' q = a\}$$

Let $q(a)$ be a minimizer. Assumption: q is Lipschitz continuous.

PROPOSED POLICY:

$$\begin{array}{l}
X_1 - - - - - - \qquad\qquad\quad * \\
X_2 - - - - - * - - - - - \\
X_3 - - - - - - - - * - - - - \\
X_4 - - - - - - - - \qquad * \qquad\qquad\qquad \text{LEGEND: } X - - - \qquad q(X) *
\end{array}$$

Priority to overloaded classes

In addition, (i) No use of nonbasic activities, (ii) Work conservation.

Asymptotics, the NDS regime

THEOREM (with Itai Gurvich): Assume C is convex. Fix a finite T.
Then under any policy,

$$\liminf_{n\to\infty} \int_0^T C(\hat{Q}^n(t))dt \geq \int_0^T C^*(Q^*(t))dt,$$

where Q^* is the RBM $\mathit{\Gamma}(\theta' X_0 + \theta' W)$.

Moreover, under the proposed policy,

$$\limsup_{n\to\infty} \int_0^T C(\hat{Q}^n(t))dt = \int_0^T C^*(Q^*(t))dt$$

About the lower bound

The LB does not hold in non-integral form.

* Minimality of the Skorohod map is well-known:

Let $\zeta \in D$. Let $\eta \in D$ be non-decreasing, $\eta(0) \geq 0$. Assume $\zeta(t) + \eta(t) \geq 0$, for all $t \geq 0$. Then

$$\zeta(t) + \eta(t) \geq \Gamma[\zeta](t) \equiv \zeta(t) + \sup_{s \leq t}[\zeta(s)^{-}], \qquad t \geq 0.$$

About the lower bound

The integral LB uses the following perturbation lemma about the Skorohod map:

LEMMA (with Itai Gurvich): Let $T > 0$ and $\varepsilon > 0$, $\varepsilon < T$, be given. Let $\zeta \in D$ and assume $\zeta(0) \geq 0$. Let

$$\alpha = \zeta + \eta + \beta,$$

where $\eta \in D$ is non-decreasing, $\eta(0) \geq 0$, $\beta \in D$ satisfies

$$-\varepsilon^2 \leq \int_0^t \beta(s)ds \leq \varepsilon^2 \qquad t \in [0,T],$$

and $\alpha(t) \geq 0$, $t \in [0,T]$. Then

$$\alpha(t) \geq \Gamma[\zeta](t) + \beta(t) - \text{Osc}(\zeta|_{[0,T]}, \varepsilon) - 3\varepsilon, \qquad t \in [0,T].$$

Thank you!