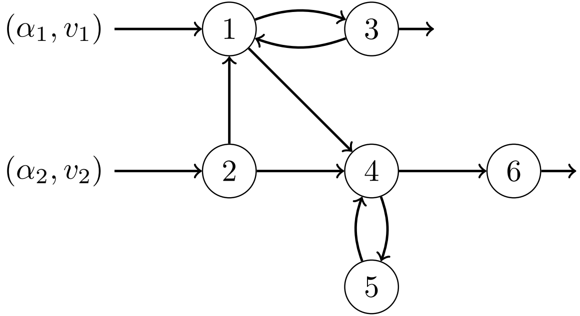
Diffusion Parameters of Flows in Stable Queueing Networks

Yoni Nazarathy
The University of Queensland

Werner Scheinhardt
University of Twente

Presented in the "The Stochastic Brain" session at SPA-2014,
Buenos Aires, July 28, 2014.

A Network of Queues



$$P = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \mu = \begin{bmatrix} 8.25 \\ 8.25 \\ 5 \\ 8.25 \\ 5 \\ 5 \end{bmatrix} \quad \alpha = \begin{bmatrix} 1 \\ 4 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad V^2 = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Network Equations

$$Q_k(t) = Q_k(0) + A_k(t) + \sum_{i=1}^{K} D_{i,k}(t) - \sum_{j=0}^{K} D_{k,j}(t)$$

$$D_{i,j}(t) = \Phi_{i,j}\Big(S_i\big(T_i(t)\big)\Big), \quad i = 1, \ldots, K,\ j = 0, \ldots, K$$

$$\sum_{j=0}^{K} \Phi_{i,j}(\ell) = \ell, \quad i = 1, \ldots, K$$

$$D = \Big[D_{1,1}, \ldots, D_{1,K}, D_{2,1}, \ldots, D_{2,K}, \ldots\ldots\ldots\ldots, D_{K,1}, \ldots, D_{K,K}\Big]'$$

Asymptotic Rates

$$\nu_{i\to j} := \lim_{t\to\infty} \frac{\mathbb{E}[D_{i,j}(t)]}{t}$$

$$\sigma^2_{i\to j} := \lim_{t\to\infty} \frac{\mathrm{Var}\Big(D_{i,j}(t)\Big)}{t}, \quad \sigma_{i_1\to j_1, i_2\to j_2} := \lim_{t\to\infty} \frac{\mathrm{Cov}\Big(D_{i_1,j_1}(t), D_{i_2,j_2}(t)\Big)}{t}$$

Calculating $\nu_{i\to j}$ is easy

Our contribution: Closed formulas for the (co)variance terms with diffusion limits

Fluid and Diffusion Scaling and Limits

For $n = 1, 2, \ldots$ and a function $U(t)$, denote $\bar{U}^n(t) = U(nt)/n$. We say that a fluid limit of U exists if $\lim_{n\to\infty} \bar{U}^n(t) = \bar{U}(t)$ exists uniformly on compact sets (u.o.c) almost surely. Further, when the limit $\bar{U}(t)$ exists, denote,

$$\hat{U}^n(t) = \frac{U(nt) - \bar{U}(nt)}{\sqrt{n}}, \quad n = 1, 2, \ldots.$$

In cases where the above sequence converges weakly on Skorohod J_1 topology to a limiting process, $\hat{U}(t)$, we denote,

$$\hat{U}^n \Rightarrow \hat{U}$$

For discrete time processes replace $U(nt)$ by $U(\lfloor nt \rfloor)$

Assumptions on Primitives

$A_k(\cdot)$, $S_k(\cdot)$ and $\Phi_{k,\cdot}(\cdot)$ independent

FSLLN:

$$\bar{A}_i(t) = \alpha_i t, \quad \bar{S}_i(t) = \mu_i t, \quad \bar{\Phi}_{i,j}(\ell) = p_{i,j}\ell,$$

with $\alpha_i > 0$, $\mu_i > 0$, $p_{i,j} \geq 0$, and $p_{i,0} = (1 - \sum_{j=1}^{N} p_{i,j}) \geq 0$ and $\mathrm{sp}(P) < 1$ so $\nu_{i \to j} := (I - P')^{-1} \alpha \, p_{i,j}$

FCLT:

$\hat{A}_i(t)$ are BM with coefficients $v_i \geq 0$

$$\hat{\Phi}_{k,\cdot}(t) = \left[\hat{\Phi}_{k,1}(t), \ldots, \hat{\Phi}_{k,K}(t)\right]', \quad k = 1, \ldots, K,$$

are BM with cov matrices Γ_k, having entries $p_{k,i}(\delta_{i,j} - p_{k,j})$

UI:

$$\left\{ \frac{(A_i(t) - \alpha_i t)^2}{t}, \, t \geq t_0 \right\} \quad \text{is UI}$$

Stable Scheduling Policy Assumptions

Different policies imply different restrictions on $T(t)$ (single-class, multi-class, preemptive, non-preemptive, etc...)

We assume **stability:**

(A1) Fluid limits for work allocations: $\bar{T}_k(t) = \frac{\nu_k}{\mu_k} t$

(A2) $\hat{Q}^n \Rightarrow 0$

(A3) Moment growth rates: $\mathbb{E}[\big(Q_k(t)\big)^2] = o(t)$ as $t \to \infty$

In multi-class setting: necessary condition: $\sum_{k \in \mathcal{C}_i} \frac{\nu_k}{\mu_k} < 1$

Assumption (A1) implies,

$$\lim_{n \to \infty} \bar{D}^n_{i,j}(t) := \bar{D}_{i,j}(t) = \bar{\Phi}_{i,j}(\bar{S}_i(\bar{T}_i(t))) = \nu_{i \to j}\, t, \;\; \text{u.o.c.}$$

Theorem

$\hat{D}^n$ converges weakly to Brownian Motion with cov matrix,

$$\Sigma^{(D)} := H \begin{bmatrix} diag(v_k^2) & & & 0 \\ & \nu_1 \Gamma_1 & & \\ & & \ddots & \\ 0 & & & \nu_K \Gamma_K \end{bmatrix} H'$$

where $H := \begin{bmatrix} P_c(I - P')^{-1} & I_{K^2} + P_c(I - P')^{-1} B \end{bmatrix}$, *and*

$$B := \mathbf{1}' \otimes I, \qquad P_c := \begin{bmatrix} P' \, e_{1,1} \\ P' \, e_{2,2} \\ \vdots \\ P' \, e_{K,K} \end{bmatrix}$$

Further $\sigma_{i_1 \to j_1, i_2 \to j_2} = \Sigma^{(D)}_{(i_1-1)K+j_1, \; (i_2-1)K+j_2}$

Proof Idea

$$
\begin{aligned}
0 &= \sum_{j=0}^{K} \hat{\Phi}^n_{i,j}(\ell), \quad \ell = 1, 2, \ldots \\
\hat{Q}^n_k(t) &= \hat{A}^n_k(t) + \sum_{j=1}^{K} \hat{D}^n_{j,k}(t) - \sum_{j=0}^{K} \hat{D}^n_{k,j}(t), \quad t \geq 0
\end{aligned}
$$

Proof Idea (cont.)

$$\tilde{\Phi}^n_{i,j}(t) := \hat{\Phi}^n_{i,j}\Big(\bar{S}^n_i(\bar{T}^n_i(t))\Big), \qquad \tilde{S}^n_k(t) := \hat{S}^n_k(\bar{T}^n_k(t))$$

Lemma: $\hat{D}^n_{i,j}(t) = \tilde{\Phi}^n_{i,j}(t) + p_{i,j}\tilde{S}^n_i(t) + p_{i,j}\mu_i\hat{T}^n_i(t)$

Now manipulate to get:

$$\hat{D}^n(t) = \left[\begin{array}{cc} H & 0_{K\times K} \end{array}\right] \left[\begin{array}{c} \hat{A}^n(t) \\ \tilde{\Phi}^n(t) \\ \hat{S}^n(t) \end{array}\right] - P_c(I-P')^{-1}\hat{Q}^n(t).$$

Now use the FCLT assumptions to get the weak convergence of $\hat{D}^n$

The (co)variance rate convergence requires a bit more work using the UI and moment assumptions

Observation: The asymptotic diffusion processes and parameters do not depend on the service sequences

Reason: Queues are stable

Thought: So how about for a network without delays?

The Zero Service Time Model

$N_{i,j|k}(\ell) \equiv$ The number of times that the ℓ'th customer arriving starting at k traverses $i \to j$

$$\left\{ \left(N_{i,j|k}(\ell),\ i,j \in \{1,\dots,K\}, i \neq j \right),\ \ell = 1,2,\dots \right\},$$

is an i.i.d. sequence (of K^2 dimensional random vectors) with distribution based on the absorbing Markov chain:

$$\tilde{P} = \left[\begin{array}{cc} 1 & \mathbf{0}' \\ \mathbf{1} - P\mathbf{1} & P \end{array} \right]$$

Now define:

$$\breve{D}_{i,j}(t) := \sum_{k=1}^{K} \sum_{\ell=1}^{A_k(t)} N_{i,j|k}(\ell)$$

$$D_{i,j}(t) \leq \breve{D}_{i,j}(t), \qquad \text{a.s.}$$

Denote now,

$$\breve{N}_{i,j}(t) := \breve{D}_{i,j}(t) - D_{i,j}(t)$$

so,

$$\breve{N}_{i,j}(t) \quad =^d \quad \sum_{k=1}^{K} \sum_{\ell=1}^{Q_k(t)} N_{i,j|k}(\ell)$$

Hence $D_{i,j}(t)$ and $\breve{D}_{i,j}(t)$ differ by a "stable" quantity

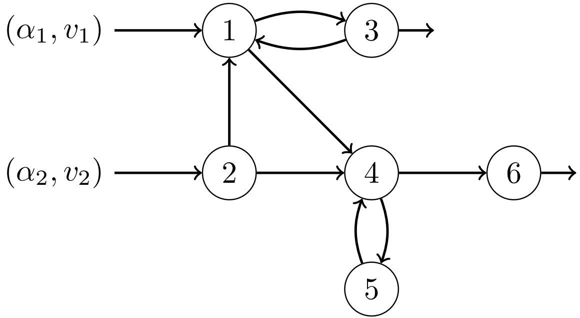
Proposition:

$$\breve{\sigma}_{i_1 \to j_1, i_2 \to j_2} = \sigma_{i_1 \to j_1, i_2 \to j_2}$$

Proposition:

$$\breve{\sigma}_{i_1 \to j_1, i_2 \to j_2} \quad = \quad \sum_{k=1}^{K} \alpha_k \, \mathrm{Cov}(N_{i_1, j_1|k}, N_{i_2, j_2|k}) + \sum_{k=1}^{K} v_k^2 \, \mathbb{E}[N_{i_1, j_1|k}] \mathbb{E}[N_{i_2, j_2|k}]$$

Back to the Example Network



$$P = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \mu = \begin{bmatrix} 8.25 \\ 8.25 \\ 5 \\ 8.25 \\ 5 \\ 5 \end{bmatrix} \quad \alpha = \begin{bmatrix} 1 \\ 4 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad V^2 = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Numerical Example

In this case, $\Sigma^{(D)}$ is 36×36. The diagonals are:

$i \backslash j$	1	2	3	4	5	6
1	0	0	$32/9$	$20/9$	0	0
2	$3/2$	0	0	$3/2$	0	0
3	$31/18$	0	0	0	0	0
4	0	0	0	0	$199/18$	$55/18$
5	0	0	0	$199/18$	0	0
6	0	0	0	0	0	0

$$\rho_{i_1 \to i_2, j_1 \to j_2} := \frac{\sigma_{i_1 \to i_2, j_1 \to j_2}}{\sqrt{\sigma^2_{i_1 \to i_2} \sigma^2_{j_1 \to j_2}}}.$$

e.g. (leaving v_2 free):

$$\rho_{1 \to 3, 2 \to 4} = \frac{v_2^2 - 4}{\sqrt{(v_2^2 + 4)(v_2^2 + 30)}}$$

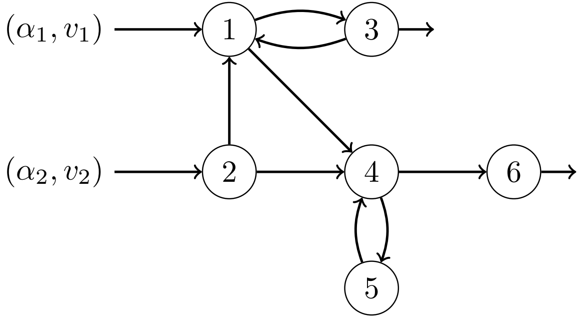
Arrivals to Individual Queues:

$$\Sigma^{(E)} = \begin{bmatrix} 68/9 & 4/3 & 40/9 & 44/9 & 22/9 & 22/9 \\ & 2 & 2/3 & 10/3 & 5/3 & 5/3 \\ & & 32/9 & 10/9 & 5/9 & 5/9 \\ & & & 182/9 & 127/9 & 55/9 \\ & & & & 199/18 & 55/18 \\ & & & & & 55/18 \end{bmatrix}.$$

The diagonal of $\Sigma^{(E)}$ may be useful for network decomposition approximations. Normalizing by ν we get square coefficient of variation:

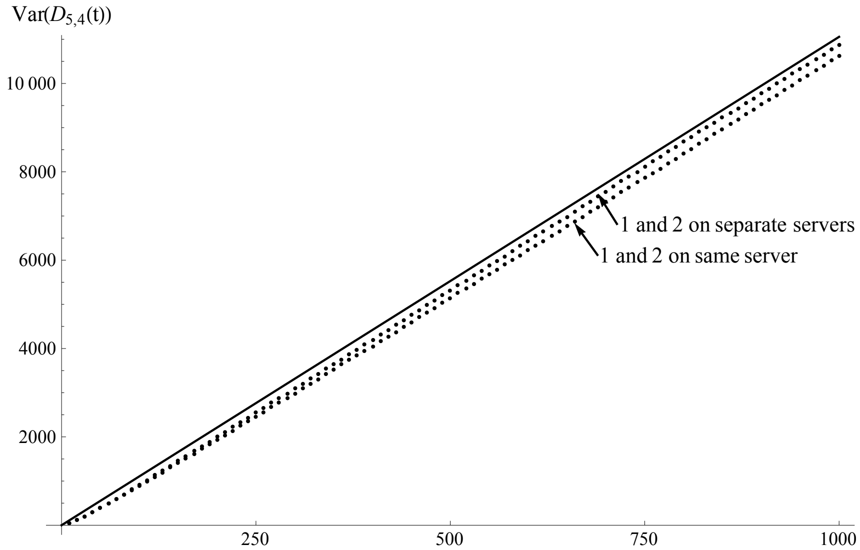
$$c^2 = \begin{bmatrix} 1.89 & 0.5 & 1.78 & 2.53 & 2.76 & 0.76 \end{bmatrix}'$$

Single Class (woman brain) vs. Multi-Class (man brain)

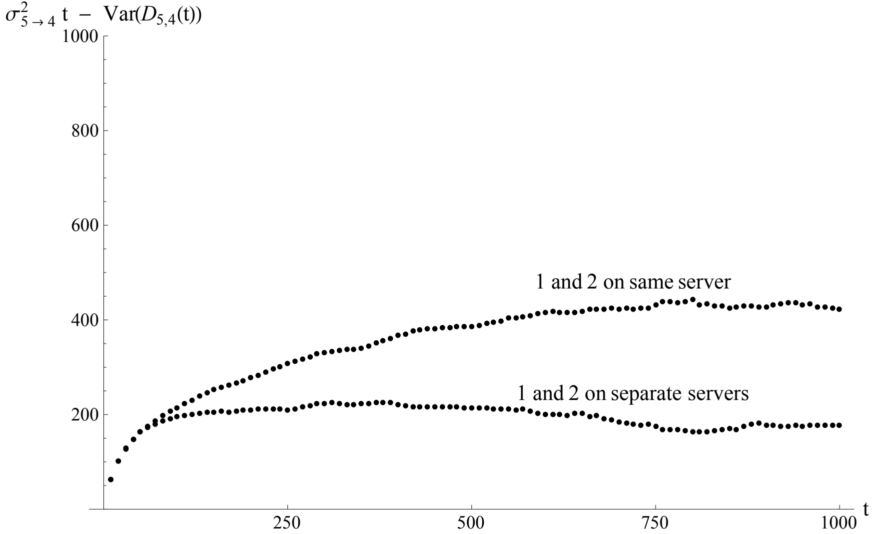


Single-class: Each queue has a dedicated (separate) server (all parts of brain can process in parallel)

Multi-class: Queues 1 and 2 are served by the same server under a non-pre-emptive priority policy giving priority to queue 1 (can't talk while watching football, and priority is obvious). $\Rightarrow$ Load on server of queues 1 and 2 is $\nu_1/\mu_1 + \nu_2/\mu_2 \approx 0.97 < 1$



Var($D_{5,4}$(t))
10 000
8000
6000
4000
2000
250
500
750
1000
t
1 and 2 on separate servers
1 and 2 on same server


$\sigma^2_{5\to4}\,t - \mathrm{Var}(D_{5,4}(t))$
1000
800
600
400
200
250
500
750
1000
t
1 and 2 on same server
1 and 2 on separate servers

Where does the stochastic brain fit in?

My thoughts are quite random on this, so I don't know...

Yoni Nazarathy and Werner Scheinhardt. "*Diffusion parameters of flows in stable queueing networks*" **arXiv preprint arXiv:1311.5610** (2013).