CLASSICAL k-SERVER QUEUES REVISITED: SOME PROBLEMS

D. J. Daley

The Australian National University

(Amplified notes for talk at WORMS'04, Carnegie–Mellon Univ., 19–20 April 2004)

Abstract

The talk surveys both recent and 'classical' studies of k-server queueing systems. There will not be time to do justice to more than a selection of the following topics.

1. EXISTENCE. Kiefer and Wolfowitz' (1955) proof of the existence of a stationary regime for the GI/GI/k FCFS system relies on 'An Essential Lemma'. In reality it hinges on a strong law of large numbers, of which variants are useful in a range of problems including k-fast servers and the tail behaviour of the work-load vector. It depends on ergodicity and not the independence assumptions of GI/GI/k (cf. Loynes' Lemma and Construction).

2. CYCLES. Much work on busy period analysis has relied on either regeneration points or weakened versions of them. The prime underlying result again is one of ergodicity and stationarity, and exploits ideas of point processes, albeit subsequent to Khinchin's (1955) monograph.

3. MOMENTS. The fact that a sub-minimally stable GI/GI/k system has finite mean delay when $E(S^{3/2}) < \infty$ for a generic service-time r.v. S is now better understood. An intermediate result concerns the moment index $\kappa(X) = \sup\{k : E(X^k) < \infty\}$ of independent r.v.s X and Y, with $\kappa(\min(X, Y)) \geq \kappa(X) + \kappa(Y)$. Equality holds if X has a d.f. with a regularly varying tail, but the class is a little larger than this.

4. EXPONENTIAL SYSTEMS. Let two identical ./M/1 systems be fed by the same Poisson arrival process: what is their joint stationary distribution, and correlation of the queue sizes? Alan Brown has conjectured the structure of the d.f. sufficient for use with an expression for the stationary bivariate p.g.f.: is there any 'routine' approach to circumvent the relative intractability of such simple systems?

5. HEAVY TAILS. The same dichotomy between minimally and sub-minimally stable systems as occurs for moment behaviour affects the asymptotic tail-behaviour of the work-load vector when the service-time d.f. is heavy-tailed.

6. SECOND-ORDER PROPERTIES. The stationary waiting-time sequence $\{W_n\}$ in a single-server system has monotonic decreasing correlations (i.e. $\mathrm{corr}(W_0, W_n) \downarrow$ for $0 < n \uparrow$), and its Hurst index is a simple function of the moment index of service times. Does monotonicity hold in GI/GI/k? And what is the Hurst index of $\mathrm{corr}(W_0, W_n)$: is it affected by the system being minimally stable or not?

0. The many-server FCFS queue

c channels

Notation: s servers

$i,\ j, k$ integers

Recall besides the queueing notation: GI/M/1, E_j/D/1, etc, so to use k for the number of servers as in D/G/k reminds us that David George Kendall was responsible for introducing the notation.

Moreover: Realtor's advice to a house-purchaser is to consider three factors: location, location, and location.

Similarly one can contemplate advice to be given to a mathematician: three things to remember: notation, notation, and notation.

1. Kiefer and Wolfowitz (1955) 'Essential Lemma'

Let $\{X_n\}$ be a (doubly infinite) stationary sequence, ergodic and nonnegative, with finite mean $\mu = \mathrm{E}(X_n) < \infty$. Then to $\varepsilon > 0$, there exists $n_\varepsilon(\omega)$ such that $X_1 + \cdots + X_n > n(\mu - \varepsilon)$ for all $n \geq n_\varepsilon(\omega)$.

Define a k-dimensional quasi 'random walk' $\mathbf{Y}_n = (Y_{n1}\ \cdots\ Y_{nk})$ by

$$\mathbf{Y}_{n+1} = \mathbf{Y}_n + X_{n+1}\mathbf{e}_{J_n},$$

where the k-vector $\mathbf{e}_j = (0\ \cdots\ 1\ \cdots\ 0)$ has 1 in the jth place and 0s elsewhere, and

$$Y_{nJ_n} = \text{ smallest component of } \mathbf{Y}_n, \text{ denoted } \min(\mathbf{Y}_n).$$

(If there exists more than one solution $J_n \in \{1, \ldots, k\}$, then J_n is taken to be any one of these solutions uniformly at random.) The following Lemma is proved much as in Kiefer and Wolfowitz (1955, §4) (else, see e.g. Borovkov, 1976, Lemma 25.4).

Lemma (Strong Law of Large Numbers for $\{\mathbf{Y}_n\}$).

$$\frac{\mathbf{Y}_n}{n} \to \frac{\mu}{n}\mathbf{1} \text{ a.s.} \quad (n \to \infty) \qquad \textit{where } \mathbf{1} = (1\,1\ \cdots\ 1).$$

Proof. Introduce

$$U_n = k\max(\mathbf{Y}_n) - |\mathbf{Y}_n| = \sum_{i=1}^{k}\bigl(\max(\mathbf{Y}_n) - Y_{ni}\bigr),$$

where $\max(\cdot)$ denotes the largest component of the k-vector argument and $|\mathbf{Y}_n| = \sum_{i=1}^{k} Y_{ni} = \sum_{r=1}^{n} X_r$, so $|\mathbf{Y}_n|/n \to \mu$ a.s. by ergodicity, and the assertion of the Lemma follows if we can show that $\max(\mathbf{Y}_n)/n \to \mu/k$ a.s.; this property is equivalent to $U_n/n \to 0$ a.s., which in turn is implied by showing that $U_n \leq_d Z$ for some honest r.v. Z.

We have

$$\begin{aligned} U_n &= k\max(\mathbf{Y}_n) - |\mathbf{Y}_n| \\ &= k[\max(\mathbf{Y}_n) - \max(\mathbf{Y}_{n-1})] + k\max(\mathbf{Y}_{n-1}) - |\mathbf{Y}_{n-1}| - X_n \\ &= k[\max(\mathbf{Y}_n) - \max(\mathbf{Y}_{n-1})] + U_{n-1} - X_n\,. \end{aligned}$$

Either $X_n < \max(\mathbf{Y}_{n-1}) - \min(\mathbf{Y}_{n-1})$, and then $\max(\mathbf{Y}_n) = \max(\mathbf{Y}_{n-1})$, in which case $U_n = U_{n-1} - X_n$, or else $X_n \ge \max(\mathbf{Y}_{n-1}) - \min(\mathbf{Y}_{n-1})$ and then $\max(\mathbf{Y}_n) = X_n + \min(\mathbf{Y}_{n-1})$, so that

$$\begin{aligned} k\max(\mathbf{Y}_n) - |\mathbf{Y}_n| &= -X_n + \sum_{i=1}^{k}\big(\max(\mathbf{Y}_n) - Y_{n-1,i}\big) \\ &< -X_n + \sum_{i=1}^{k}\big(\max(\mathbf{Y}_n) - \min(\mathbf{Y}_{n-1})\big) \;=\; (k-1)X_n, \end{aligned}$$

i.e.

$$U_n \le \begin{cases} U_{n-1} - X_n, \\ (k-1)X_n, \end{cases} = \max\big(U_{n-1} - X_n,\, (k-1)X_n\big).$$

By iteration,

$$\begin{aligned} U_n &\le \max\big((k-1)X_n, (k-1)X_{n-1} - X_n, \ldots, (k-1)X_1 - X_2 - \cdots - X_n\big) \\ &=_d \max\big((k-1)X_{-1},\, (k-1)X_{-2} - X_{-1}, \ldots,\, (k-1)X_{-n} - X_{-n+1} - \cdots - X_{-1}\big) \\ &\equiv U'_n. \end{aligned}$$

Ergodicity implies that a.s. we have $(k-1)X_{-n} \ge X_{-1} + \cdots + X_{-n+1}$ only finitely often, and therefore $U'_n \uparrow U'_\infty$ a.s. for some r.v. U'_∞ that is $< \infty$ a.s. We can take $Z = U'_\infty$. $\square$

The Loynes Construction (e.g. Baccelli and Brémaud, 1994, Chapter 2) is an important generalization of this argument.

Defining $\max_r(\mathbf{Y}_n) = r$th largest component of $\mathbf{Y}_n$, we can set

$$\begin{aligned} U_{n[r]} &= \max(\mathbf{Y}_n) + \cdots + \max_{r-1}(\mathbf{Y}_n) + (k+1-r)\max_r(\mathbf{Y}_n) - |\mathbf{Y}_n| \\ &= \sum_{i=r+1}^{k}\big(\max_r - \max_i\big)(\mathbf{Y}_n). \end{aligned}$$

Then a similar argument shows that $\{U_{n[r]}\}$ converges weakly to a stationary r.v. dominated by $U'_{[r]}$ that satisfies

$$U'_{[r]} =_{\mathrm{d}} \max\big(U'_{[r]} - X,\, (k-r)X\big),$$

where X is a generic element of the stationary sequence $\{X_n\}$.

When $\{X_n\}$ are i.i.d. r.v.s with $\mathrm{E}(X^2) < \infty$, we can deduce that

$$0 \le \mathrm{E}(U'_{[r]}) - (k-r)\mathrm{E}(X) \le \mathrm{E}(X) + (k-r+1)\frac{\mathrm{var}(X)}{\mathrm{E}(X)}.$$

The quantity of major interest in GI/GI/k FCFS system is $\min(\mathbf{Y}_n)$.

Problem: Is there an analogue for $\min(\mathbf{Y}_n)$ of $U'_{[r]}$?

The SLLN Lemma is needed in Foss and Kornuchov asymptotics for the heavy-tail behaviour of $\mathbf{W}_n = (W_{n1}\ \cdots\ W_{nk})$.

2. Busy periods and cycles in k-server systems

(This section was skipped in the verbal presentation at WORMS'04)

In a single-server system, a *busy period* is a time interval from the arrival of a customer into an empty system until the first departure thereafter that leaves the system empty. Its k-server analogue is more problematic: the literature in the '80s is dominated by wanting regeneration points (to facilitate Markov chain analysis), e.g. 'all servers idle at an arrival epoch' (see Daley and Servi (1998) for a short review). But such a definition may be what is appropriate in an infinite-server system, unless . . . ! !

The literature uses the term 'full busy period' to denote a time interval from an arrival epoch when the system switches from at least one server idle to all servers busy until the next ensuing departure epoch after which not all servers are busy.

These 'full busy period' starting epochs constitute a subset of the arrival epochs $\{t_n\}$ say, of all customers into the system, where in terms of the Kiefer and Wolfowitz work-load vector $\mathbf{W}_n = (W_{n1} \ \cdots \ W_{nk})$ we can define the indicator r.v.s

$$I_n = I_{\{W_{n1}=0\}},$$
$$J_{nj} = \begin{cases} I_{\{W_{nj}=0<W_{n,j+1}\}} & (j = 1, \dots, k-1), \\ I_{\{W_{nk}=0\}} & (j = k). \end{cases}$$

Clearly, when $\{\mathbf{W}_n\}$ is stationary, so too are $\{I_n\}$ and $\{J_{nj}\}$. From the definitions, $I_n = J_{n1} + \cdots + J_{nk}$. Set $\pi_{0j} = \mathrm{E}(J_{nj})$ and $\pi_0 = \mathrm{E}(I_n)$ so that $\pi_0 = \pi_{01} + \cdots + \pi_{0k}$.

These quantities give

$$\pi_{01} > 0 \qquad \Longleftrightarrow \qquad \mathrm{E}(\text{full busy period}) < \infty$$

because a full busy period $\le$ (full) busy cycle where

$$\text{busy cycle} = t_{n_{i+1}} - t_{n_i}$$

and the t_{n_i} are the arrival epochs where successive busy periods start, i.e. $\{n_i\} = \{n\colon J_{n1} = 1\}$.

(1) In GI/GI/k the t_{n_i} are stopping times for $\{(T_n, S_n, \mathbf{W}_n)\}$.

(2) The events $\{I_n = 1\}$ and $\{J_{nj} = 1\}$ defined above constitute stationary point processes on the integers (Breiman, 1968, §6.10), with

$$\pi_{01} = \mathrm{E}(I_{\{J_n=1\}}) = \frac{1}{\mathrm{E}(n_{i+1} - n_i)},$$

and $\pi_{01} > 0$ if and only if $\mathrm{E}(n_{i+1} - n_i) < \infty$.

(3) Finally, E(full busy cycle length) $= \mathrm{E}(T)\,\mathrm{E}(n_{i+1} - n_i)$ where T is a generic inter-arrival time.

Note that in this 'description', busy cycles and full busy periods are in one-to-one correspondence.

Armed with the indicator variables J_{nj}, we can as easily define a j-busy period as a sub-interval of the time intervals between epochs $t_{n_{ji}}$ defined by the ordered set of indices

$$\{n_{ji}\} = \{n\colon J_{nj} = 1\}.$$

Problem: Check Kiefer and Wolfowitz and then Loynes (1962) for conditions under which the π_{0j} are positive. Whitt (1972) has condition for $\pi_{0k} > 0$ (i.e. for 'all servers idle at an arrival epoch' to be a positive recurrent regenerative phenomenon). Presumably, conditions for positivity or not of π_{0j} $(j = 1, \ldots, k)$ are expressible in terms of the positivity or not of the quantities $\Pr\{S > T_1 + \cdots + T_j\}$ and $1 - \Pr\{S < T_1 + \cdots + T_j\}$. See also Baccelli and Brémaud (1994, §2.2.4).

3. Moment index and moment order

For a nonnegative r.v. X, define its moment index $\kappa(X)$ by

$$\kappa(X) = \sup\{k\colon \mathrm{E}(X^k)\lambda < \infty\} \;=\; \liminf_{x\to\infty} \frac{-\log \Pr\{X > x\}}{\log x}$$

(the first relation here is the definition; the second equality is fairly easy to prove). Then for independent r.v.s X and Y,

$$\begin{aligned} \kappa(cX) &= \kappa(X) \qquad (0 < c < \infty), \\ \kappa(X^c) &= \kappa(X)/c \qquad (0 < c < \infty), \\ \kappa(X + Y) &= \min\big(\kappa(X),\ \kappa(Y)\big), \\ \kappa(\max(X, Y)) &= \min\big(\kappa(X),\ \kappa(Y)\big), \\ \kappa(\min(X, Y)) &\ge \kappa(X) + \kappa(Y), \end{aligned}$$

with equality above for all Y if and only if

$$\limsup_{x\to\infty} \frac{-\log \Pr\{X > x\}}{\log x} = \kappa(X),$$

i.e. the limit of the ratio defining $\kappa(X)$ exists. Strict inequality is possible (there is an example in Daley, 2001). It is also the case that equality can occur for a r.v. X whose d.f. has a tail that is not regularly varying (Daley and Goldie, 2004). The relation for $\kappa(X + Y)$ holds for arbitrary nonnegative X, Y (just use the c_r inequality), and then because $X \le \max(X, Y) \le X + Y$, the relation for $\kappa(\max(X, Y))$ does not need independence of X, Y either.

Kiefer and Wolfowitz (1956) showed that in GI/GI/1, the stationary waiting time r.v. W satisfies $\kappa(W) = \kappa(S) - 1$, and that in the k-server system GI/GI/k,

$$\kappa(W_{n1} + \cdots + W_{nk}) = \kappa(S) - 1.$$

It follows from work in Scheller-Wolf and Sigman (1997) that in GI/GI/k with

$$\tau \equiv \frac{\mathrm{E}(S)}{\mathrm{E}(T)} < k - 1,$$

$\kappa(W_1) = 2\kappa(S) - 2$ (for suitable d.f.s for S).

Problem: Modify the above for general k and τ.

Problem: Relate the result to the Recurrence relation equation

$$Z_{n+1} = (Z_n + Y_n)_+$$

where $\{Y_n\}$ is ergodic and has $\mathrm{E}(Y_n) < 0$. For suitable such Y_n, this implies the existence of stationary $\{Z_n\}$ satisfying the Recurrence relation equation. 'Suitable' here includes

$$\mathrm{E}(Y_{n+1} \mid \mathcal{F}_n) \le -\delta < 0 \qquad (\text{all } n),$$

which should then give

$$\kappa(Z) = \kappa(Y) - 1$$

'because' for large Z_1 = 'large' Y_1,

$$\begin{aligned}\mathrm{E}(Z^\alpha) &\approx \frac{Z_1^\alpha + (Z_1 + Y_1)^\alpha + \cdots + (Z_1 + Y_1 + \cdots + Y_n)^\alpha}{n} \\ &\approx \frac{Z_1}{|\mathrm{E}(Y_2)|}(\vartheta Z_i^\alpha + \text{ terms of smaller order}).\end{aligned}$$

Problem: In the recurrence relation as in S-W&S, there arises the question of $\kappa(\min(X, X_e))$ where the nonnegative r.v.s X and X_e are independent and X_e is an 'equilibrium' version of X, i.e. its has a density function proportional to $\Pr\{X > x\}$ (and X has finite mean). I showed straight after the workshop (it is in Daley and Goldie, 2004) that

$$\kappa(\min(X, X_e)) = \kappa(X) + (\kappa(X) - 1) = 2\kappa(X) - 1.$$

This almost certainly rules out the possibility of having $\kappa(\text{delay})$ taking a value other than $\kappa(S) - 1$, or $\kappa(S) - \frac{1}{2}$ etc.

4. Exponential systems: 2 or more ./M/1

(Not talked about)

5. Second order properties

In stationary GI/GI/1, a stationary waiting-time sequence $\{W_n\}$ has the properties

$$\operatorname{cov}(W_n, W_0) \downarrow 0 \qquad (0 < n \uparrow \infty),$$
$$\sum_{n=0}^{\infty} \operatorname{cov}(W_0, W_n) < \infty \qquad \text{iff} \qquad \mathrm{E}(S^4) < \infty.$$

Further,

$$\operatorname{var}(W_1 + \cdots + W_n) \approx \begin{cases} \text{const.}\, n^{5-\kappa} & (3 < \kappa(S) < 4), \\ \text{const.}\, n & (\kappa(S) \geq 4), \end{cases}$$

so that in the former case here, W_n has Hurst index $\frac{1}{2}(5-\kappa)$ (this is a measure of longe-range dependence).

Problem: In GI/GI/k,

(1) Is $\operatorname{cov}(W_0, W_n)$ monotonic decreasing in n as for $k = 1$?
[Is there an analogue of stochastic monotonicity?]

(2) (a) When is $\sum_{n=0}^{\infty} \operatorname{cov}(W_0, W_n)$ finite?
(b) What is the Hurst index of $\{W_n\}$?
[Presumably answers here may depend on $\tau = \mathrm{E}(S)/\mathrm{E}(T)$ in relation to k. For example, the condition may be the same as for GI/GI/1 when $k-1 < \tau < k$ but some weaker condition suffices when $\tau < k-1$.]

References

Baccelli, F. and Brémaud, P. (1994). *Elements of Queueing Theory*. Springer-Verlag, Berlin.

Borovkov, A.A. (1976). *Stochastic Processes in Queueing Theory*. Springer-Verlag, New York.

Breiman, Leo (1968). *Probability*. Addison-Wesley, Reading MA.

Daley, D.J. (1997). Some results for the mean waiting-time and work-load in GI/GI/k queues. In *Frontiers in Queueing Systems: Models and Applications in Science and Engineering*, (ed. J. H. Dshalalow), CRC Press, Boca Raton FA, pp.35–59.

Daley, D.J. (2001). The moment index of minima. In *Probability, Statistics and Seismology (A Festschrift for David Vere-Jones)*, J. Appl. Probab. **38A**, 33–36.

Daley, D.J. and Goldie, C.M. (2004). The moment index of minima (II). (Manuscript).

Daley, D.J. and Servi, L.D. (1998). Idle and busy periods in stable M/M/k queues. *J. Appl. Probab.* **35**, 950–962.

Loynes, R.M. (1962). The stability of queues with non-independent interarrival and service times. *Proc. Cambridge Philos. Soc.* **58**, 497–520.

Kiefer, J. and Wolfowitz, J. (1955). On the theory of queues with many servers. *Trans. Amer. Math. Soc.* **78**, 1–18.

Kiefer, J. and Wolfowitz, J. (1956). On the characteristics of the general queueing process with applications to random walks. *Ann. Math. Statist.* **27**, 147–161.

Scheller-Wolf, A. and Sigman, K. (1997). Delay moments for FIFO GI/GI/s queues. *Queueing Systems* **25**, 77–95.

Whitt, W. (1972). Embedded renewal processes in GI/G/s queues. *J. Appl. Probab.* **9**. 650–658.