Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

A note on large-buffer asymptotics for generalized processor sharing with Gaussian inputs

K.G. Dębicki, M.R.H. Mandjes

REPORT PNA-R0703 FEBRUARY 2007

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

A note on large-buffer asymptotics for generalized processor sharing with Gaussian inputs

ABSTRACT
In a previous study logarithmic large-buffer asymptotics were derived for a two-class generalized processor sharing system with Gaussian inputs, for three of the four possible scenarios. In this note we show how the large-buffer asymptotics for the remaining fourth regime follow from a recently derived result for tandem systems. We also provide a heuristic interpretation of the result.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: Queueing, generalized processor sharing, large deviations

A note on large-buffer asymptotics for generalized processor sharing with Gaussian inputs

Krzysztof Dębicki* and Michel Mandjes†

February 20, 2007

Abstract

In a previous study [3] logarithmic large-buffer asymptotics were derived for a two-class generalized processor sharing system with Gaussian inputs, for three of the four possible scenarios. In this note we show how the large-buffer asymptotics for the remaining fourth regime follow from a recently derived result for tandem systems. We also provide a heuristic interpretation of the result.

1 Introduction and model

In [3] a two-class Generalized Processor Sharing (GPS) queueing system is studied. In two-class GPS traffic of class i ($i = 1, 2$) is guaranteed a service rate $\phi_i c$ (with $\phi_1+\phi_2 = 1$), and the service rate not used by one class is made available to the other class. For further background on GPS' system mechanics, see for instance [2]. Following [3], input of class i ($i = 1, 2$) is modeled as a Gaussian process with stationary increments; with $A_i(s, t)$ denoting the traffic of class i arriving in $[s, t)$, the mean rate is μ_i (so that $\mathbb{E}A_i(s, s+\delta) = \mu_i\delta$) and the variance curve is $v_i(\cdot)$ (so that $\mathbb{V}\mathrm{ar}A_i(s, s+\delta) = v_i(\delta)$). The stability condition $\mu_1 + \mu_2 < c$ applies.

The authors of [3] concentrate, without loss of generality, on the steady-state distribution of the workload in queue 1, say Q_1. They distinguish four scenarios:

(S1) $\mu_2 > \phi_2 c$;

(S2) $\mu_2 < \phi_2 c$, and $v_2(t) = o(v_1(t))$ as $t \to \infty$;

(S3) $\mu_1 < \phi_1 c$, $\mu_2 < \phi_2 c$, and $v_1(t) = o(v_2(t))$ as $t \to \infty$;

(S4) $\mu_1 > \phi_1 c$, and $v_1(t) = o(v_2(t))$ as $t \to \infty$.

For the first three scenarios, the authors of [3] find, under mild conditions on the variance curves, the logarithmic large-buffer asymptotics; more precisely, constants r and κ are identified such that

$$\lim_{u\to\infty} \frac{1}{u^r} \log \mathbb{P}(Q_1 > u) = -\kappa.$$

The counterpart of this result for scenario (S4), however, was not given.

*K. Dębicki (email: Krzysztof.Debicki@math.uni.wroc.pl) is with Instytut Matematyczny, University of Wrocław, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland. KD was supported by KBN Grant No 1 P03A 031 28 (2005-2007).

†M. Mandjes (email: mmandjes@science.uva.nl) is with Korteweg-de Vries Institute, University of Amsterdam, Amsterdam, the Netherlands. MM is also affiliated to CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands, and EURANDOM, Eindhoven, the Netherlands.

In this note we shall argue that a recent result on tandem queues [4], which relies heavily on the infinite-intersections machinery developed in [5], can be directly applied to also solve this regime in the important special case that source 2 is a fractional Brownian motion (fBm). We also provide an insightful heuristic interpretation of the result.

Notation. We introduce the following notation. $\{Q_1^d(t)\}$ (with $d > \mu_1$) is the steady-state workload process corresponding to queue 1, if it were served in isolation at rate d:

$$Q_1^d(t) := \sup_{s \leq t}\{A_1(s,t) - d(t-s)\};$$

also $Q_1^d := Q_1^d(0)$ (which is, according to Reich's formula, distributed as $\sup_{t\geq 0}\{A_1(-t,0) - dt\}$). Two other 'steady-state random variables' are, with $\mu_1 > \phi_1 c$ and $d < \mu_1$,

$$V^\varepsilon := \sup_{t\geq 0}\{A_2(-t,0) - (c-\mu_1-\varepsilon)t\} - \sup_{s\geq 0}\{A_2(-s,0) - \phi_2 cs\};$$

$$Z^d := \sup_{t\geq 0}\{dt - A_1(-t,0)\}.$$

As follows directly from [6, Lemma 2.4], V^ε can be interpreted as the steady-state workload of a queue fed by arrival process 2, where the first queue is emptied at rate $\phi_2 c$, and the second at rate $c - \mu_1 - \varepsilon$. Likewise, Z^d is distributed as the steady-state workload of a single queue fed at constant rate d, and emptied at a variable rate (that has the same statistical characteristics as the arrival rate of class 1).

2 Auxiliary results

In this section we prove two lemmas. The first relates the steady-state buffer content of queue 1 to the distributions of V^ε and $V^{-\varepsilon}$. The second lemma considers a tandem system with fBm input, and translates the results on the many-sources asymptotics of the downstream queue into the corresponding large-buffer asymptotics.

Lemma 2.1 *For sufficiently small ε, and any $u, x > 0$ and $\delta \in (0,1)$,*

$$\mathbb{P}(V^{-\varepsilon} > u + x)\mathbb{P}(Z^{\mu_1-\varepsilon} \leq x) \;\leq\; \mathbb{P}(Q_1 > u) \;\leq\; \mathbb{P}(V^\varepsilon > (1-\delta)u) + \mathbb{P}(Q_1^{\mu_1+\varepsilon} > \delta u).$$

Proof. Lower bound. Let $B_i^d(-t,0)$ be the amount of service received by class i, if the queue were served in isolation at rate d. Define

$$L^\varepsilon := \sup_{t\geq 0}\{B_2^{\phi_2 c}(-t,0) - (c-\mu_1+\varepsilon)t\}.$$

Following Lemma 3.1 in [1], for all $u, x > 0$ and ε sufficiently small,

$$\mathbb{P}(Q_1 > u) \geq \mathbb{P}(L^\varepsilon > u + x)\mathbb{P}(Z^{\mu_1-\varepsilon} \leq x).$$

As, by definition, $Q_2^{\phi_2 c}(0) = Q_2^{\phi_2 c}(-t) + A_2(-t,0) - B_2^{\phi_2 c}(-t,0)$,

$$\begin{aligned} L^\varepsilon &= \sup_{t\geq 0}\{Q_2^{\phi_2 c}(-t) - Q_2^{\phi_2 c}(0) + A_2(-t,0) - (c-\mu_1+\varepsilon)t\} \\ &\geq \sup_{t\geq 0}\{A_2(-t,0) - (c-\mu_1+\varepsilon)t\} - Q_2^{\phi_2 c}(0) = V^{-\varepsilon}. \end{aligned}$$

Upper bound. Due to Reich's formula, $Q_1 = \sup_{t\geq 0}\{A_1(-t,0) - C_1(-t,0)\}$, where $C_1(-t,0)$ is the capacity available to class 1 in the interval $[-t,0)$. Evidently,

$$Q_1 \leq \sup_{t\geq 0}\{A_1(-t,0) - (\mu_1+\varepsilon)t\} + \sup_{s\geq 0}\{(\mu_1+\varepsilon)s - C_1(-s,0)\}.$$

The first term is interpreted as $Q_1^{\mu_1+\varepsilon}$. The second term corresponds to the workload of class 1 in a system (which we refer to as the 'modified system') in which the amount of traffic put into queue 1 in $[-s,0)$ is $(\mu_1+\varepsilon)s$ rather than $A_1(-s,0)$ (and the input of class 2 is as before). The workload of class 1 is equal to the difference, in the modified system, of the total workload and the workload in queue 2. Evidently, we can write the total workload in the modified system as

$$\sup_{s\geq 0}\{(\mu_1+\varepsilon)s + A_2(-s,0) - cs\}.$$

Notice that in the modified system class 2 does not obtain any unused service capacity from class 1 (as class 1 is generating traffic at a rate higher than its weight: $\mu_1+\varepsilon > \phi_1 c$, due to (S4)). Thus the workload in queue 2 of the modified system is

$$\sup_{s\geq 0}\{A_2(-s,0) - \phi_2 cs\}.$$

We find that $Q_1 \leq V^\varepsilon + Q_1^{\mu_1+\varepsilon}$. This leads immediately to the upper bound. □

In [6, Lemma 2.4] it was shown that

$$\sup_{t\geq 0}\{A(-t,0) - c_2 t\} - \sup_{s\geq 0}\{A(-s,0) - c_1 s\}$$

is distributed as the stationary workload of the downstream queue of a tandem system, fed by the arrival process A, in which queue i is emptied at constant rate c_i (assume $c_1 > c_2$). Furthermore, in case of fractional Brownian motion input (with Hurst parameter H, i.e. with variance curve $v(t) = t^{2H}$), in [4] a constant $\lambda \equiv \lambda(H, c_1, c_2)$ is determined such that

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}\left(\sup_{t\geq 0}\left\{\sum_{i=1}^n A_i(-t,0) - nc_2 t\right\} - \sup_{s\geq 0}\left\{\sum_{i=1}^n A_i(-s,0) - nc_1 s\right\} > n\right) = -\lambda. \tag{1}$$

Interestingly, for $c_1 \geq c_1^\star$, with

$$c_1^\star := \frac{c_2}{H}\left(\sup_{\alpha\in[0,1]} \frac{1+\alpha^{2H} - (1-\alpha)^{2H}}{\alpha}\right),$$

an explicit expression for λ can be given: $\lambda = c_2^{-1}\cdot H/(1-H)$; c_1 does not play a role, and hence the 'shaping effect' of the first queue has no impact on the decay rate of overflow in the second queue.
For $c_1^\star < c_1$, however, the situation is radically different: c_1 *does* play a role in the decay rate λ. In [4] it is explained which path is most probable in this regime, and how the corresponding decay rate can be computed; see in particular Thm. 8 of [4]. Interestingly, in this regime a part of the most probable path is linear. From the construction of this most probable path, as used in the proofs in [4], one sees that $\lambda(H, c_1, c_2)$ is continuous in H, c_1, c_2.

Lemma 2.2 *Let $A(\cdot)$ be a centered fractional Brownian motion with Hurst parameter $H \in (0,1)$ and let $A_1(\cdot), \ldots, A_n(\cdot)$ be i.i.d. copies of $A(\cdot)$. Then,*

$$\lim_{u\to\infty} \frac{1}{u^{2-2H}} \log \mathbb{P}\left(\sup_{t\geq 0}\{A(-t,0) - c_2 t\} - \sup_{s\geq 0}\{A(-s,0) - c_1 s\} > u \right) \tag{2}$$

$$= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\sup_{t\geq 0}\left\{ \sum_{i=1}^{n} A_i(-t,0) - nc_2 t \right\} - \sup_{s\geq 0}\left\{ \sum_{i=1}^{n} A_i(-s,0) - nc_1 s \right\} > n \right) = -\lambda. \tag{3}$$

Proof. First replace u in (2) by $n^{\varrho(H)}$, with $\varrho(H) := 1/(2-2H)$, and rescale time by a factor $n^{\varrho(H)}$. We arrive at

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\sup_{t\geq 0}\left\{ A\left(-n^{\varrho(H)}t, 0\right) - n^{\varrho(H)} c_2 t \right\} - \sup_{s\geq 0}\left\{ A\left(-n^{\varrho(H)}s, 0\right) - n^{\varrho(H)} c_1 s \right\} > n^{\varrho(H)} \right).$$

Because of the self-similarity of fBm, $A(-n^{\varrho(H)}t, 0)$ is distributed as $n^{H\varrho(H)} A(-t,0)$; we obtain after multiplying both sides with $n^{1-\varrho(H)}$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\sup_{t\geq 0}\left\{ \sqrt{n} A(-t,0) - nc_2 t \right\} - \sup_{s\geq 0}\left\{ \sqrt{n} A(-s,0) - nc_1 s \right\} > n \right) = -\lambda,$$

since $\sum_{i=1}^{n} A_i(-t,0)$ is distributed as $\sqrt{n} A(-t,0)$.
In order to complete the proof we formally have to show the existence of limit (2). This, however, straightforwardly follows from the combination of the above part of the proof with the fact that the nominator in (2) is decreasing with u, while the denominator is regularly varying (and hence the three-series theorem can be applied). □

3 Main result

The following theorem is our main result. It requires that A_1 satisfies the conditions:

C1: $v_1(\cdot)$ is $C([0,\infty))$ and ultimately strictly increasing;

C2: $v_1(\cdot)$ is regularly varying at 0 with index $\beta_1 \in (0,2]$, and regularly varying at ∞ with index $\alpha_1 \in (0,2)$.

The above conditions match assumptions C1-C2 of [3]. We note that for scenario (S4) we do not need to assume that C3 of [3] holds.

Theorem 3.1 *Assume that A_1 satisfies C1-C2, A_2 is a fractional Brownian motion with Hurst parameter $H \in (1/2, 1)$ and (S4) holds. Then*

$$\lim_{u\to\infty} \frac{1}{u^{2-2H}} \log \mathbb{P}(Q_1 > u) = -\lambda(H, \phi_2 c, c - \mu_1).$$

Proof. For each sufficiently small $\varepsilon > 0$ and each $x > 0$, $\delta \in (0,1)$, by (1) and Lemma 2.2,

$$\lim_{u\to\infty} \frac{1}{u^{2-2H}} \log \mathbb{P}(V^{-\varepsilon} > u + x) = -\lambda(H, \phi_2 c, c - \mu_1 + \varepsilon).$$

and

$$\lim_{u\to\infty} \frac{1}{u^{2-2H}} \log \mathbb{P}(V^{\varepsilon} > (1-\delta)u) = -(1-\delta)^{2-2H} \lambda(H, \phi_2 c, c - \mu_1 - \varepsilon). \tag{4}$$

Additionally, Lemma 4.2 in [3] (which holds if A_1 satisfies C1-C2) straightforwardly implies that

$$\lim_{u\to\infty} \frac{v_1(u)}{u^2} \log \mathbb{P}(Q_1^{\mu_1+\varepsilon} > \delta u) = -\frac{1}{2}\varepsilon^{\alpha_1}\delta^{2-\alpha_1} \left(\frac{\alpha_1}{2-\alpha_1}\right)^{-\alpha_1} \left(\frac{2}{2-\alpha_1}\right)^{2},$$

which, in view of (S4) and (4), yields for any $\delta \in (0,1)$ and u sufficiently large

$$\log \mathbb{P}(Q_1^{\mu_1+\varepsilon} > \delta u) \ll \log \mathbb{P}(V^{\varepsilon} > (1-\delta)u).$$

Combining this with Lemma 2.1 (and pushing $\varepsilon \downarrow 0$ and $\delta \downarrow 0$), the proof is completed. □

We remark that it is trivial to extend the above theorem from A_2 being (standard) fBm to 'non-standard fBm', i.e., with mean rate μ_2 not necessarily 0, and $v_2(t) = \kappa t^{2H}$ for κ not necessarily 1.

The boundary case where $\mu_1 = \phi_1 c$, and $v_1(t) = o(v_2(t))$ as $t \to \infty$ is not covered by Theorem 3.1 (since $\lambda(H, c_1, c_2)$ is well defined only for $c_1 > c_2 > 0$). We suspect that a different approach is needed to get the logarithmic asymptotic under this scenario. This differs from the case $\mu_1 \geq \phi_1 c$, and $v_2(t) = o(v_1(t))$ as $t \to \infty$, already solved in [3] (Remark 3.2). The asymptotics for scenario $\mu_2 = \phi_2 c$ can be derived by a straightforward combination of Theorems 3.1, 3.2 and 3.3 in [3]. Indeed, providing that appropriate assumptions for A_1, A_2 given in [3] are satisfied and using that $\mathbb{P}(Q_1 > u)$ is nondecreasing as a function of μ_2, we have that both for $v_2(t) = o(v_1(t))$ as $t \to \infty$ and $v_1(t) = o(v_2(t))$ as $t \to \infty$

$$\lim_{u\to\infty} \frac{v_1(u)}{u^2} \log \mathbb{P}(Q_1 > u) = -\frac{1}{2}(\phi_1 c - \mu_1)^{\alpha_1} \left(\frac{\alpha_1}{2-\alpha_1}\right)^{-\alpha_1} \left(\frac{2}{2-\alpha_1}\right)^{2}.$$

4 Discussion and interpretation

It can be checked in [3] that, considering the situation of A_2 being fBm with Hurst parameter H, the logarithmic large-buffer asymptotics of scenarios (S1), (S2), and (S3) do not involve H (class-2 traffic appears in the asymptotics only through its mean rate, or not at all); we find here that in scenario (S4) *H does* appear. The heuristics behind this scenario (and an explanation why its asymptotics resemble those of an associated tandem queue) are the following.

Given the fact that in (S4) it holds that $v_1(t) = o(v_2(t))$, one would expect that in the most likely way to exceed level u in queue 1, class 1 should transmit roughly at rate μ_1 (which is no rare event; it corresponds to its average behavior). The question that remains is: how much capacity (from the $\phi_2 c$ allocated to queue 2) should class 2 claim? That must be more than μ_2, as otherwise queue 1 does not build up, but how much more? To answer this question, suppose that class 2 generates traffic at rate r. It is 'useless' to choose r larger than $\phi_2 c$, as class 2 can never claim more than $\phi_2 c$. If $r \geq \phi_2 c$, then queue 1 grows at rate $\mu_1 - \phi_1 c$. If r is smaller than $\phi_2 c$, then queue 1 grows at rate $\mu_1 - c + r$. Summarizing, queue 1 builds up essentially linearly, with slope

$$\mu_1 - (c - \min\{r, \phi_2 c\}) = \mu_1 - \max\{c - r, \phi_1 c\}.$$

Put differently, queue 1 grows as the downstream queue in a tandem network whose first queue is fed at rate r and with service capacities $\phi_2 c$ (first queue) and $c - \mu_1$ (second queue); note here that in this situation the output rate of the first queue would be $\min\{r, \phi_2 c\}$. This insight immediately explains why $\mathbb{P}(Q_1 \geq u)$ looks like

$$\mathbb{P}\left(\sup_{t\geq 0}\{A_2(-t,0) - (c-\mu_1)t\} - \sup_{s\geq 0}\{A_2(-s,0) - \phi_2 cs\} \geq u\right),$$

which is interpreted as the buffer content of the downstream queue of a tandem network, as noticed above. It is noted that arguments in the same vein can be found in [8]: the formulae for the decay rates given there show a similar relation between queues operating under GPS and tandem systems.

We believe that Theorem 3.1 can be extended to other than Gaussian classes of input processes. Since Lemma 2.1 does not assume the traffic process to be Gaussian, all what is needed is the logarithmic tail behavior of the downstream node in a tandem system. If the latter is available, the associated decay rate satisfies a continuity property, and A_2 is burstier than A_1 in some sense (see, e.g. [7]), then an analog of Theorem 3.1 should hold. The corresponding scenario for long-tailed traffic flows has been analyzed in [1].

References

[1] S. Borst, O. Boxma, and P. Jelenković (1999). Induced burstiness in generalized processor sharing queues with long-tailed traffic flows. *Proc. 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 316-325.

[2] S. Borst, M. Mandjes, and M. van Uitert (2003). Generalized Processor Sharing queues with heterogeneous traffic classes. *Advances in Applied Probability*, Vol. 35, pp. 806-845.

[3] K. Dębicki and M. van Uitert (2006). Large buffer asymptotics for generalized processor sharing queues with Gaussian inputs. *Queueing Systems,* Vol. 54, pp. 111-120.

[4] M. Mandjes, P. Mannersalo, and I. Norros (2007). Gaussian tandem queues with an application to dimensioning of switch fabrics. *Computer Networks,* Vol. 51, pp. 781-797.

[5] M. Mandjes, P. Mannersalo, I. Norros, and M. van Uitert (2005). Large deviations of infinite intersections of events in Gaussian processes. *Stochastic Processes and their Applications*, Vol. 116, pp. 1269-1293.

[6] M. Mandjes and M. van Uitert (2005). Sample-path large deviations for tandem and priority queues with Gaussian inputs. *Annals of Applied Probability*, Vol. 15, pp. 1193-1226.

[7] D. Wischik and A. Ganesh (2004). The calculus of Hurstiness. To appear in *Queueing Systems.* Available from `http://www.cs.ucl.ac.uk/staff/D.Wischik/Research/hurstiness.pdf`

[8] Z.-L. Zhang (1997). Large deviations and the Generalized Processor Sharing scheduling for a two-queue system. *Queueing Systems*, Vol. 26, pp. 229-245.