Queueing analysis for priority multiplexing of aggregated VBR on-off regenerative traffic

Michael Shalmon
University of Quebec, INRS-EMT

Abstract-We present the queueing analysis for priority multiplexing of VBR on-off regenerative traffic processes, the on-off periods being i.i.d, the off-periods exponentially distributed and the on-periods general and of variable rate. This traffic model includes as a particular case batch Poisson traffic sources and, under the assumption that for each priority class the aggregated on-rate is not less than the service rate, it is aggregation and input/output invariant enabling the extension of the queueing analysis at an isolated multiplexing station to the queueing analysis of the end to end (source to terminus) delays in traffic concentrating networks.

Index Terms - networks, aggregated traffic, multiplexing, priority, foreground/background, VBR, on-off, regenerative.

1 Introduction

1.1. M-G*V model and input/output invariance: By regenerative on-off we mean that the traffic process regenerates at any instant inside any off-period. The assumed regeneration property implies that the on-off periods are i.i.d and the off-periods exponentially distributed, while the on-periods traffic could be general and of variable rate. We will refer to the regenerative on-off traffic processes described above as M-G*V, where M points to the memoryless duration of the off-periods, and G*V to the general variable rate on-periods. If the on-rate is deterministic, the process will be referred to as M-G*D; if in addition the duration of the on-periods is memoryless the process will be referred to as M-M*D. Note that the output process from a M/D/1 queue with batch arrivals is M-G*D.

Basic M-G*V on-rate assumption: The on-rate is never smaller than the service rate (cell inter-arrival times are not greater than cell service time). The basic assumption is reasonable for aggregated traffic, it implies that M-G*V traffic is input/output invariant (see Lemma 1. below), it applies in particular to batch Poisson traffic for which the analysis is known, and enables a queueing analysis not much more difficult than that for a batch M/D/1 queue. Recall that rate always refers to a local average of the number of cells (constant size packets) per unit time (or inverse of local average of cell inter-arrival times) and that rate modelling is (loosely speaking) equivalent to infinitesimal cells being released continuously and (under the basic assumption) contiguously in continuous time.
We will refer to this model as the fluid model. Alternatively, the traffic process during an on-period could be modelled by batches of discrete size cells arriving in continuous time, the first cell arrival starting the on-period, and such that (under the basic assumption) cell inter-arrival times are not greater than the cell service time; within a batch of cells the cell inter-arrival times are zero. We will refer to this model as the basic model. In a more restrictive modelling, time could be assumed to be slotted, a slot being equal to cell service time. We will refer to this model as the slotted model. As the service time of a cell is practically negligible, the three models are practically equivalent. The analysis of the fluid and slotted models is "neater", as these models sweep delays of the order of cell service time under the rug and do not distinguish between cell preemptive and non-preemptive scheduling. Note that for arbitrary cell traffic the traffic output is on-off, cells inter-departure intervals being equal to cell service time, the output on-period corresponding to the busy period interval delayed by one cell service time.

Assume a single network access multiplexing station and M-G*V exogenous input traffic. For the analysis of the duration of output on-periods and for the analysis of the delays of on-periods with prioritized multiplexing it is enough to specify an M-G*V process by the mean length $\lambda^{-1}$ of a memoryless off-period, and by the joint distribution or the joint m.g.f of the duration of an on-period represented by the random variable $\alpha$ and of the (cumulative work input during an on-period minus the duration of the on-period) represented by the nonnegative random variable $\delta$.

**Lemma 1. M-G*V basic properties**

(i) The stationary probability for a M-G*V process with parameters $\lambda, [\alpha, \delta]$ to be in an on-period equals $\lambda E\alpha/(1 + \lambda E\alpha)$.

(ii) The traffic intensity (mean workload rate) of a M-G*V process with parameters $\lambda, [\alpha, \delta]$ equals $\rho = \lambda(E\alpha + E\delta)/(1 + \lambda E\alpha)$ provided $\rho < 1 \iff \lambda E\delta < 1$.

Assume that the traffic from the i-th source flowing to a network access station is M-G*V with parameters $\lambda_i, [\alpha_i, \delta_i]$; the sources are assumed independent. Define $\lambda = \sum \lambda_i$ and $\rho = \sum \rho_i$, where the summation is over all the sources flowing into the network access station and assume that $\rho < 1 \iff \sum \lambda_i E\delta_i < 1$.

Then:

(iii) The aggregate traffic input at the network access station is M-G*V with parameters $\lambda, \rho$. The traffic output from the network access station is M-G*D with parameters $\lambda, \rho$.

**Proof:** Statements (i),(ii) follow from the law of large numbers and the renewal theorem. Statement (iii) follow from the memoryless duration of the off-periods, from the basic assumption (implying that input on-periods do not break on the output but may coalesce), and from the traffic intensity conservation in stationarity.

The M-G*D process in its fluid form was first introduced by Kingman and applied to queueing analysis in [7], 1970. The input/output invariance of M-G*D fluid model was observed and the m.g.f of the duration of the output on-period for two i.i.d. M-G*D was first obtained by Rubinovitch, [9], 1973. The multi-
plexer of Poisson arriving G*D on-periods (the limit of an aggregate of very many, very sparse M-G*D independent traffic sources), or in the mainstream terminology, the M/G/∞ process, was analyzed by Cohen, [2], 1974, who obtained a computable formula for the aggregate joint m.g.f of $[\alpha, \delta]$. [9] and [2] were extended in [6]. The relevance of the M-G*D basic model for high speed networks was stated by Shalmon and Kaplan, [11], 1984, who generalize to an arbitrary tandem concentrating network the FCFS analysis of a two station tandem concentrating network with Poisson traffic by Kaplan, [5], 1980. The polling and prioritized link service for an arbitrary tandem concentrating network with Poisson traffic is analyzed by Shalmon, [13], 1987, who points out that the analysis technique is applicable to [M-G*D+Poisson] traffic. A formula for the overall traffic mean delay of M-G*V sources is obtained by Dupuis and Hajek in [3], 1994. The tail distribution of M/G/∞ is analyzed by Jelenkovic and Lazar [4], 1997. For related work on tandem queueing, see W. Scheinhardt and Zwart, [10], 2002. Note that the M-M*D fluid model first analyzed by Kosten, [8], 1974, and further developed by Anick, Mitra and Sondhi, [1], 1982, is not input/output invariant.

1.2. The paper. In section 2. we analyze a foreground-background model for which the foreground M-G*V traffic has NPR (non-preemptive) priority over general background traffic by comparing it to the case where the priority is PR (preemptive-resume); the analysis generalizes that of foreground M-G*D traffic in [13]. In section 3. we present the queueing analysis of the unfinished work and of the M-G*D output on-period duration for a single (or higher priority) M-G*V input at an isolated station: we derive the distributional equation characterizing an output on-period, and following [5], [11], we reduce the unfinished work and delay analysis during the off-periods (also at the start and end of on-periods) to that of an M/G/1 queue. In section 4. we present the analysis of delays for a lower priority M-G*V source, the higher priority traffic being M-G*V, and we derive the distributional equation characterizing the M-G*D output on-period duration for two M-G*V traffic inputs (high and low priority). In section 5. we present the analysis of the unfinished work inside the on-periods in the special case of [M-G*D+ batch Poisson], and point out how to apply it to the analysis of tail distribution in the case of M/G/∞ with on-periods of sub-exponentially distributed duration. Finally, we mention the extension of the queueing analysis at an isolated multiplexing station to the queueing analysis of the end to end (source to terminus) delays in unidirectional traffic concentrating networks; this extension is presented in the companion paper [16].

1.3. Notation, Basic formulas and analysis technique. Where $X, Y, M$ are nonnegative valued random variables and $M$ is integer valued, we write $M_{X,Y,M}(s_1, s_2, z)$ for the m.g.f $E \exp \{-s_1X - s_2Y\} z^M$, $M_X(s)$ for the m.g.f $E \exp \{-sX\}$, and $P_M(z)$ for the p.g.f $E z^M$. For computing moments, one can usually bypass the m.g.f’s. A typical formula in this paper is $[U + \int_{u=0}^{U} d\nu_{\lambda}(u)X(u)]$ where $\nu_{\lambda}(u)$ represents the number of Poisson epochs at rate $\lambda$ during a duration $u$, the $\{X(u)\}$ i.i.d. distributed as
$X$, $U$ being independent of $\{X(u)\}, \{\nu_X(u)\}$. The standard formulas below are companions to the formulas in the paper.

1.3.1 $\text{Exp}\{ -s[U + \int_{u=0}^{U} d\nu_X(u)X(u)]\} = \mathcal{M}_U(s + \lambda(1 - \mathcal{M}_X(s)))$; the mean is $EU(1 + \lambda EX)$; the 2nd moment is $EU\lambda EX^2 + EU^2(1 + \lambda EX)^2$.

Where $U$ represents the duration of a random time interval, we write $[\overline{U}, \overline{U}]$ to represent the [attained, residual] duration of $U$ at a random point. $\mathcal{M}_{\overline{U}}(s) = (1 - \mathcal{M}_U(s))/sEU$ and $\overline{EU}^k = EU^{k+1}/(k+1)EU$.

1.3.2. $\text{Exp}\{ -s[\overline{U} + \int_{u=0}^{\overline{U}} d\nu_X(u)X(u)]\} = [2EU/sEU^2][1 - \mathcal{M}_U(s + \lambda(1 - \mathcal{M}_X(s)))]$; the mean equals $(EU^2/2EU)(1 + \lambda EX)$; the second moment equals $(EU^2/2EU)\lambda EX^2 + (EU^3/3EU)(1 + \lambda EX)^2$.

Throughout the paper, the analysis is based on a decomposition of service duration and of the unfinished work derived via PR (preemptive-resume) priority and having a random tree branching interpretation, Shalmon [12],1985, [14],1988. Basic formulas 1.3.1, 1.3.2 correspond to the PR-Branching decomposition.