Threshold Policies for Single-Resource Reservation Systems

E. G. Coffman Jr. Predrag Jelenković
Columbia University, New York

1 Introduction

Requests for a resource arrive at rate λ, each request specifying a future time interval, called a *reservation interval*, to be booked for its use of the resource. The *advance notices* (delays before reservation intervals are to begin) are independent and drawn from a distribution $A(z)$. The durations of reservation intervals are sampled from the distribution $B(z)$ and are independent of each other and the advance notices. We let A and B denote random variables with the distributions $A(z)$ and $B(z)$ (the functional notation will always allow one to distinguish between our two uses of the symbols A and B).

The following greedy reservation policy was analyzed in [3]: A request is immediately accepted (booked) if and only if the resource will be available throughout its reservation interval, i.e., the resource has not already been reserved for a time period overlapping the requested reservation interval. In [3], the authors compute an efficiency measure, called the *reservation probability*, which is the fraction of time the resource is in use.

This paper studies the reservation probability for a more general greedy policy of threshold type that is defined by two parameters s and τ. If a request has an advance notice less than s or a duration exceeding τ, then the threshold policy makes an attempt to book it under the greedy rule; otherwise, it is rejected even if it could have been accommodated. Our main result is an expression for the asymptotic reservation probability as $s \to \infty$ and the advance-notice distribution becomes progressively more spread out.

The above result relates asymptotics of reservation policies to asymptotics of interval packing policies, a connection first studied in [3]. In the interval packing problem [1], intervals arrive randomly in $\mathbf{R}_+^2$ according to a Poisson process in the two dimensions representing arrival times t and the left endpoints of the arriving intervals. Interval lengths are i.i.d., and since we will map them to reservation intervals, we let their distribution also be denoted by $B(z)$. The intensity is 1, i.e., an average of one interval arrives per unit time per unit distance. For a given $x > 0$, an arriving interval is packed (or accepted) in the 'containing' interval $[0, x]$ under the greedy algorithm if and only if it is a subinterval of $[0, x]$ and it does not overlap an interval already accepted. The problem is to find, or at least estimate, the function $K(t, x)$, which is the expected total length of the intervals accepted by the greedy policy during $[0, t]$, assuming that none has yet been accepted by time 0 ($[0, x]$ is initially empty).

Estimates of $K(t, x)$ were obtained in [3] from its Laplace transform $\mathcal{K}(t, u)$; these results are special cases of the corresponding results for the *threshold* packing policy with parameters s, τ. The threshold packing policy extends greedy interval packing much as we extended the greedy reservation policy: An interval is processed by the greedy packing algorithm if its length is at least τ or if it arrives no sooner than s; otherwise, it is rejected. The next section exhibits the Laplace transform of $H_\tau(s, t, x)$, the expected total length of the intervals accepted during $[0, t]$, $t \geq s$, by the threshold packing policy with parameters s, τ. Note that threshold packing reduces to simple greedy packing if $\tau = 0$ or if $s = 0$. The formulas in the next section will verify that $K(t, x) = H_0(t, t, x)$.

As noted in [3], there are many potential applications covered by models like ours. However, relatively new applications in existing and proposed communication systems, e.g., teleconferencing and video-on-demand systems, have given a fresh impetus to research on reservation systems. Previous work in the communications field is quite recent and focuses more on engineering problems than mathematical foundations; past research has dealt with the implementation issues of incorporating distributed advance-notice reservation protocols in current networks, and with the algorithmic issues concerned with well utilized resources in reservation systems (see [3, 4, 5] for many references). For the analysis of mathematical models different from our own, see the work of Virtamo [5] and Greenberg, Srikant, and Whitt [4].

2 Threshold interval packing

Let $b_\tau := \mathbf{E}B \cdot 1(B > \tau)$, $b_\tau^{(2)} := \mathbf{E}B^2 \cdot 1(B > \tau)$, and $p = p_\tau := \mathbf{P}(B > \tau)$. Denote by $L_\tau(s, t, x)$ the total length of the intervals packed at time $t \geq s$ in $[0, x]$, and let $H_\tau(s, t, x) := \mathbf{E}L_\tau(s, t, x)$.

We compute the rate of change of $H_\tau(s, t, x)$ with respect to s for $x \geq d$ by expressing $H_\tau(s + \Delta s, t, x)$ in terms of $H_\tau(s, t, x)$ and the events occurring in the time interval $[0, \Delta s]$. On doing so, rearranging the recurrence and taking the limit $\Delta \to 0$, we obtain the integro-differential equation

$$\frac{\partial H_\tau(s,t,x)}{\partial s} = -(px - b_\tau)H_\tau(s,t,x) + b_\tau x - b_\tau^{(2)} + 2\int_\tau^d dB(z)\int_0^{x-z} H_\tau(s,t,y)dy. \quad (1)$$

Note the boundary condition $K(t,x) = H_0(t,t,x)$, the known result for the simple greedy rule. Note also that the effect on (1) of putting $\tau = 0$ is confined to constants depending only on τ. Thus, exactly the same analysis for K in [3] can be applied to H_τ here. We introduce the transforms $\mathcal{H}_\tau(s,t,u) := \int_d^\infty e^{-ux} H_\tau(s,t,x)dx$, $\mathcal{B}_\tau(u) := \int_\tau^d e^{-uz} dB(z)$, and then transform (1) to obtain a pde whose solution is readily found to be

$$\begin{aligned}\mathcal{H}_\tau(s,t,u) &= \frac{1}{u^2}\int_0^s \mathcal{C}_\tau(s-z,t,u+pz)\mathcal{G}_\tau(u,pz)dz \quad (2)\\ &+ \left(\frac{u+ps}{u}\right)^2 \mathcal{K}(t-s,u+ps)\mathcal{G}_\tau(u,ps),\ x \geq d,\end{aligned}$$

where $\mathcal{G}_\tau(u,v) := \exp\left(b_\tau v - 2\int_u^{u+v} \frac{1-\mathcal{B}_\tau(y)}{y}dy\right)$, and where $\mathcal{C}_\tau(s,t,u)$ is the transform of a well-behaved function easily computed by a recursive procedure. (See the expanded version of the paper [2] for details.) As illustrated below, explicit expressions are available for simple distributions $B(.)$.

Estimates of H_τ. Leading-term asymptotics in x are given by the following result, in which

$$\begin{aligned}\alpha_\tau(s,t) &:= \int_0^s \mathcal{C}_\tau(s-z,t,pz)\mathcal{G}_\tau(pz)dz\\ &\quad + \mathcal{K}(t,ps)(ps)^2\mathcal{G}_\tau(ps). \qquad (3)\end{aligned}$$

Theorem 1 *For any* $s > 0$, $t \geq 0$, $H(s,t,x) \sim \alpha_\tau(s,t)x$ *as* $x \to \infty$.

Proof sketch: One first verifies that the functions $\mathcal{C}_\tau$, $\mathcal{G}_\tau$, and $\mathcal{K}$ are such that $\mathcal{H}_\tau$ in (2) satisfies $\mathcal{H}_\tau(s,t,u) \sim \alpha_\tau(s,t)/u^2$ as $u \downarrow 0$.. An application of Karamata's Tauberian theorem and routine manipulations then proves the theorem. ∎

Example 1: Consider the case where interval lengths have only the values 1 or 2, with probabilities $\mathbf{P}(B=2) = 1 - \mathbf{P}(B=1) = q$. Computations give

$$\mathcal{G}(u,v) = \exp\left(2q - 2\int_u^{u+v} \frac{1-qe^{-2y}}{y}dy\right),$$

and

$$\mathcal{C}_\tau = 2qe^{-2u}\left(1 + u\left[\frac{e^{-u}-e^{-2u}}{u} - \frac{e^{-u}-e^{-2u-(1-p)t}}{u+(1-p)t}\right]\right),$$

from which we obtain $\mathcal{K}(t,u)$ and hence an explicit integral formula for $\alpha(s,t)$ in (3) that we can evaluate numerically.

For $t = 7$, $p = 0.4$, we used Mathematica to compute the curve for $\alpha_\tau(s, 10-s)$. The threshold algorithm is strictly better than the greedy policy ($\alpha_\tau(0,10)$) for a large range of values of the threshold parameter s. ∎

In order to obtain higher-order error terms, we use complex analysis and the Cauchy residue theorem to evaluate directly the Laplace-transform inversion formula for $\mathcal{H}_\tau$. This gives the main result needed for the limit law of the next section,

Theorem 2 *For any fixed* $\xi, T > 0$ *with* $\xi > 2T > 0$, *there exists a constant* $\gamma_\tau(t)$, *such that*

$$\sup_{0\leq t_1,t_2\leq T} |H_\tau(t_1,t_2,x) - (\alpha_\tau(t)x + \gamma_\tau(t))| = O(e^{-\xi x}).$$

3 Advanced-notice limit law

Let the advance notice distribution be uniform on $[0,a]$ and for a given δ, $(0 < \delta < 1)$ consider the threshold reservation policy: If the advance notice of a request is in $[\delta a, a]$ and the duration requested is less than a given τ, then the request is rejected. Otherwise, it is processed according to the greedy reservation policy (i.e., it is accepted if does not overlap a time interval already reserved). Let $P_a(\lambda,\tau,\delta)$ denote the reservation probability, i.e., the stationary probability that the resource is in use.

Theorem 3 *As the support of the advance notice distribution tends to infinity* ($a \to \infty$), *we have the asymptotic reservation probability* $P_a(\lambda,\tau,\delta) \sim \alpha_\tau(\delta\lambda,\lambda)$.

The proof follows the approach used in [3] to establish the analogous result for the simple greedy policy.

References

[1] E. G. Coffman, Jr., Leopold Flatto, Predrag Jelenković, and Bjorn Poonen. Packing random intervals on line. *Algorithmica*, 22(1998), 448–476.

[2] E. G. Coffman, Jr. and Predrag Jelenković. Threshold policies for interval packing and reservation systems. Technical Report, Electrical Engineering Dept., Columbia University, New York, NY 10027.

[3] E. G. Coffman, Jr., Predrag Jelenković, and Bjorn Poonen. Reservation Probabilities. *Adv. Perf. Anal.*, 2(1999), 129–158.

[4] A. Greenberg, R. Srikant, and W. Whitt. Resource sharing for book-ahead and instantaneous- request calls. *Proceedings, ITC 15*, pages 539–548, Washington, D.C., June 1997.

[5] J. T. Virtamo. A model of reservation systems. *IEEE Trans. Comm. Sys.*, 40(1992), 109–118.