We develop and analyze a queueing model that we call Erlang-R, where the “R” stands for ReEntrant customers. The Erlang-R model accommodates customers that return to service several times during their sojourn within the system. It is most significant in time-varying environments. Indeed, it was motivated by healthcare systems, in which workloads are time-inhomogeneous and patients often go through a discontinuous service process (e.g. in Emergency Wards (EW), physicians are revisited by patients whose service process consists of cycles: examination by a physician, lab tests, treatment by a physician and so forth).

The workforce of a hospital consists of nurses, doctors, laboratory workers and others. Most of these human resources require long and costly training, and jointly contribute as much as 70% to a hospital’s operational budget [6]. Thus, careful management of work-force capacity is naturally called for. There do exist many queueing models for staffing personnel, but they all ignore either time-varying environments or recurrent services or both. For example, Green et al. [1, 2] applies the Erlang-C model with Lag-SIPP for staffing doctors in the EW: while the arrival rate varies during the day, the effect of returning patients is not accounted for.

We model systems as above by an open queueing network with multiple (statistically identical) customers and servers. The customers cycle in the system between two states. The first captures the service process and the second the delay between consecutive services. Figure 1 displays our system graphically. We refer to the service phase as a Needy state, and to the delay phase as a Content state (following Jennings et al. [5]). The main question we address is: how many servers (doctors/nurses) are required (staffing) in order to achieve predetermined service levels.
Staffing transient systems is different from staffing in steady state. Instead of setting performance measures in the long run, one must consider them at every moment in time. We would like to identify staffing procedures so that service performance is stable over time. Specifically, no matter what time of day customers enter the system, they will, for example, always wait on average the same time, and their probability of waiting remains constant. Thus, the staffing algorithm is to attain pre-specified constant service levels and, at the same time, maintain high servers’ utilization. This translates into seeking a QED (Quality and Efficiency Driven) balance, at all times of a transient system.

**Main Results**

We develop expressions for service level measures showing that, in steady state, the system’s behavior is captured by an Erlang-C (M/M/S) model. When considering time-varying environments, however, our system differs significantly from Erlang-C. The difference manifests itself in both amplitude and phase of the offered load which, in turn, is the driver of the system’s dynamics. For example, when the arrival rate is periodic, and the service times are exponentially distributed, the amplitude of an Erlang-R Offered Load (OL) is always smaller than the amplitude of a corresponding Erlang-C OL. On the other hand, the phase of Erlang-R OL sometimes leads and sometimes lags behind the phase of Erlang-C OL. In fact, these differences between the two models are especially pronounced when arrivals vary during the sojourn time of a customer, which is exactly the case in emergency wards. The implication is that Erlang-C will lead to over- or under-staffing at most times. One must thus take into account the discontinuous nature of service, in order to avoid excessive staffing costs or undesirable service levels. This is well demonstrated through several case studies. For
example, Figure 2 presents the probability of waiting when one uses the Erlang-R, Erlang-C and PSA algorithms for staffing in a relatively large system. It clearly shows that, while using Erlang-R stabilizes system’s performance around the pre-specified target (here 0.5), using Erlang-C or PSA does not.

![Figure 2](image_url)

**Figure 2:** $P(W > 0)$ as a function of time, when staffing according to Erlang-R, Erlang-C, and PSA

Based on our theory, we propose a staffing policy that attains pre-specified service levels in the Halfin-Whitt (QED) regime [3]. This policy applies the *Modified Offered Load* (MOL) approximation [4], using the well-known square-root staffing principle. We validate our policy, via simulation, for both large and small systems, and we use an EW simulator to validate its usefulness in realistic scenarios. We find that, in most cases, performance measures such as the probability of timely service, expected waiting, and servers’ utilization are all remarkably stable over time. We show that, despite the fact that our staffing algorithm is based on large-scale approximations, it also stabilizes small systems such as those in hospitals, where the number of ‘servers’ (e.g. doctors) varies between 1 and 10.

Using Taylor-series approximations, we provide means for calculating staffing levels for general arrival rates functions and general service-time distributions. For some special cases, such as when service times are exponentially distributed, numerical methods were developed using fluid models. Lastly, based on diffusion approximations, we developed new MOL approximations for the number of Needy customers and the expected waiting time in the
QED regime, which are also highly accurate.

References


