

QED Queues

Avishai Mandelbaum

*William Davidson Faculty of
Industrial Engineering and Management
Technion—Israel Institute of Technology
Technion City, Haifa 32000, Israel*

Abstract

QED queues arise in large service systems that are both **Quality** and **Efficiency-Driven** (QED). Our prime example for such systems are large best-practice telephone call centers: Here, hundreds of agents could cater to thousands of callers per hour, with an average agents' occupancy of 95%, about half of the callers being answered immediately without wait, and the rest delayed for scarcely few seconds.

The design of call center operations, and the management of their performance, has traditionally relied on classical queueing theory, mostly $M/M/N$ (Erlang-C). However, the modern complex call center, and its emerging successor the contact-center (= telephone + IVR + internet + e.mail + chat + ...), are challenging the relevance of this conservative approach. My goal in this review lecture is thus to survey ongoing research that addresses some of these challenges, all within the **asymptotic** framework of **fluid** and **diffusion** approximations for queueing systems. More specifically, the lecture will be tentatively divided to the following four parts:

1. **Introduction:** I shall start with describing call centers and queues in call centers. I then introduce (empirically, numerically and theoretically) queues that, operationally, are Efficiency driven, Quality driven and those that are carefully balanced in the sense that they are QED = Quality AND Efficiency Driven. To explain these regimes of operations concretely, consider the $M/M/N$ or $M/D/N$ queue, with offered load R and a number of servers N that is not small: These queues are efficiency-driven if $N \approx R+x$, quality-driven if $N \approx R+zR$,

and QED if $N \approx R + y\sqrt{R}$; x, y, z are scalars, which must be positive since at least R servers are required to ensure stability. Thus, QED performance is obtained via **square-root safety staffing**, where the safety $y\sqrt{R}$ protects against stochastic variability.

Three characteristics of queues in call centers will be now highlighted and elaborated on :

2. **Human Aspects**, as manifested through callers' impatience and abandonment. Our base-model here is Erlang-**A**, namely $M/M/N$ in which customers actually **A**bandon if not served within an exponentially distributed patience-time. Fluid and diffusion analysis of $M/M/N + M$ provides insights to its QED operation, and suggests generalizations to General patience distributions. Some such generalizations will be described as well.

3. **Time-Varying Dynamics**, which captures phenomena such as peak congestion at predictable times-of-day, periodic loads or time-based staffing. The base-model is $M_t/M/N_t$, which will be generalized to accommodate also abandonment and redials. It will also be shown how to stabilize a time-varying system, via proper hourly staffing, so that its daily performance matches that of a corresponding stationary system.

4. **Heterogeneity of Customers and Servers**, for example VIP and Regular customers that seek technical support for various products, in multi-languages through multiple communication channels. A need hence arises for skills-based routing, which is the online matching of multi-class customers with multi-skilled servers. Some recent progress on this difficult problem will be reported, mainly via special cases.

A common theme in all the models above is that their analysis is carried out asymptotically, as the number of servers increases indefinitely and utilization level approach 100%, so that, in the limit, the fraction of customers that are served immediately without wait is neither 0 (quality-driven) nor 1 (efficiency-driven), but in fact within $(0, 1)$ (QED). The latter $(0, 1)$ -limit turns out equivalent to square root safety staffing, as described above. The advantages of the QED operational regime were already clear to Erlang and his co-workers at the Copenhagen Telephone Company. But a rigorous mathematical articulation of the QED regime had to await the seminal paper by S. Halfin and W. Whitt ("*Heavy traffic limits for queues with many exponential servers*"), which will be our theoretical starting point.

As an introduction for the lecture, I recommend skimming through “*Telephone Call Centers: Tutorial, Review, and Research Prospects*”, with Noah Gans and Ger Koole, 2003. It is downloadable for <http://iew3.technion.ac.il/serveng2004/References/CCReview.pdf>.