QED Queues

Avishai Mandelbaum, Technion, ISRAEL

As an introduction for what is to come, I recommend skimming through
"Telephone Call Centers: Tutorial, Review, and Research Prospects",
with Noah Gans and Ger Koole, 2003. It is downloadable from
http://iew3.technion.ac.il/serveng2003/References/CCReview.pdf
(The reference numbers that are used in the sequel are taken from this paper.)

QED Queues arise in large service systems that are both **Q**uality and **E**fficiency-**D**riven (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 the M/M/N (Erlang-C) and M/N/N/N (Erlang-B) models. 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 tutorial is thus to survey ongoing research that addresses some of these challenges. This covers, broadly speaking, empirically-based research (Queueing **Science**), design principles (Service **Engineering**, for example skills-based-routing) and **Management** rules-of-thumb (ex. square-root safety-staffing), all within the **asymptotic** framework of **fluid** and **diffusion** approximations for queueing systems.

More formally, the tutorial will tentatively be divided into the following four parts:

1. Introduction [51, 28, 36, 113]: 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 [74] or M/D/N [88] 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 + z R, 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** [151, 28], 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 [124, 138, 164]. Our base-model here is M/M/N+M, namely M/M/N in which customers actually abandon if they are not served within an exponentially distributed patience-time. Fluid and diffusion analysis of M/M/N+M [56, 63] provides insights to its QED operation, and suggests generalizations to M/M/N+G [20, 117].

3. Time-Varying Dynamics, which captures phenomena such as peak congestion at predictable times-of-day, periodic loads or time-based staffing [89]. The base-model is M_t/M/N_t, which will be generalized to accommodate also abandonment and redials [107, 109]. Here, in order to be useful, fluid and diffusion analysis must preserve the transient behavior.

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 [62]. 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 via special cases ([17, 78], and ongoing work by Armony, Gurvich, Yahalom) that constitute building blocks for the yet intractable general model.

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 levels 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 [74, 63, 88].

The advantages of the QED operational regime were already clear to A. K. Erlang and his co-workers at the Copenhagen Telephone Company. (Erlang described square-root staffing in his 1924 paper "On the rational determination of the number of circuits" [51].) But a rigorous mathematical articulation of the QED regime had to await the wonderful paper by S. Halfin and W. Whitt ("Heavy traffic limits for queues with many exponential servers" [74]), which will be our theoretical starting point.