The $\Delta_{(i)}/GI/1$ Queue: Fluid and Diffusion Approximations.

Harsha Honnappa, Rahul Jain and Amy R. Ward,
University of Southern California,
Los Angeles, CA 90007.

We consider a model of queueing behavior where demand is *transitory*, and the queues exist only for a short period of time. Consider a finite population of customers who arrive to request a service with known start time. Customers can queue up in anticipation of service, and are served FIFO once service starts. This occurs, for example, when ticket holders arrive at a music venue with unassigned seating - some customers will arrive early (before the concert venue opens), in the hope of securing seats near the front, and other customers will arrive later. Other examples of transitory demand include buy or sell orders for stock submitted to a broker to 'fill-or-kill' during a single day of trading, or passengers arriving at an airport to board a flight.

We assume each customer independently samples an arrival time from a common distribution F that has finite support, with minimum value less than or equal to the service start time. This means that if $X_1, X_2, ..., X_n$ are samples from the distribution F, then the customer arrival times follow the order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$. Customer service times are i.i.d. and follow a general distribution. We call this the $\Delta_{(i)}/GI/1$ queue. Our queueing model is inherently transient; there is no steady-state that arises as time becomes large. Unfortunately, transient performance measures for queueing systems can be very difficult to compute. Therefore, we develop large population approximations for the queue-length and waiting time processes under both fluid and diffusion scaling. We also establish a transient Little's law that links the limiting queue-length and waiting time processes under these two scalings.

The limiting fluid approximations may switch between overloaded, underloaded, and critically loaded periods, as time progresses. This is interesting because this mimics the behavior of the $M_t/M_t/1$ queue in which the arrival and service processes are both time-varying [1]. However, this behavior is extant in our model, even though neither the arrival nor the service process need be time-varying (consider a uniform arrival distribution, for example). The limiting diffusion approximation arises as the directional derivative of the fluid netput process (that is, the difference between the arrival process and the potential service completion process that assumes the server is always busy). This parallels the situation in the $M_t/M_t/1$ queue. However, there is a key difference: the diffusion netput process combines a Brownian Bridge, that arises from the invariance principles related to the Kolmogorov-Smirnov statistics, and a Brownian motion, that arises from the functional central limit theorem for renewal processes.

Finally, the motivation for our arrival process comes from the previous work [2], in which the authors assume that customers choose their arrival time strategically, in order to minimize a cost function that is a weighted sum of the delay and arrival time of the customer. It was shown that, in the fluid limit, the equilibrium mixed strategy for the arrival time is a uniform distribution function. However, for the exact, discrete-event model, it is not possible to analytically solve for the equilibrium arrival strategies. The development of the diffusion approximation for the $\Delta_{(i)}/GI/1$ queue may be viewed as a first step towards identifying the equilibrium arrival times in the refined diffusion limit.

References

[1] A. Mandelbaum and W. Massey, "Strong Approximations For Time-dependent Queues," *Mathematics of Operations Research*, vol. 20, no. 1, 1995.

[2] R. Jain, S. Juneja, and N. Shimkin, "The Concert Queueing Game: To Wait or To be Late," *Discrete Event Dynamic Systems*, vol. 21(1), pp. 103–134, 2011.