From *SIAM News*, Volume 40, Number 1, January/February 2007

The Long and the Short of It: Queueing Theory Goes Dynamic

By Barry A. Cipra

William Massey (right) of Princeton University accepted the 2006 Blackwell–Tapia Prize from Douglas Arnold, director of the Institute for Mathematics and Its Applications. Massey was honored for outstanding achievements in queueing theory, stochastic networks, and modeling of communications systems, and for his contributions to increasing diversity in the mathematical sciences.

Waiting in line—be it at the post office or the grocery store, or on the phone, listening to outsourced call-center music—would seem to be an inherently stationary activity. But to William Massey, a professor of operations research and financial engineering at Princeton University, waiting in line is totally dynamic. In a lecture at the Blackwell–Tapia Conference, held November 3 and 4 at the Institute for Mathematics and Its Applications at the University of Minnesota (see "IMA Hosts 2006 Blackwell–Tapia Conference"). Massey described some of the tools used to study queues as dynamical systems. In part for work described in the talk, Massey received the 2006 Blackwell–Tapia Prize at the conference.

People have been standing in lines since the dawn of bureaucracy, but queueing theory is only about a hundred years old, an offspring-turned-midwife of the telecommunications era. The first explicit analysis is credited to the Danish mathematician Agner Krarup Erlang, who, while working for the Copenhagen Telephone Company, published "Sandsynlighedsregning og Telefonsamtaler" (Theory of Probabilities and Telephone Conversations) in 1909, followed by "Løsning af nogle Problemer fra Sandsynlighedsregning af Betydning for de automatiske Telefoncentraler" (Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges) in 1917, with other papers in between and after. (His principal works, in English translation, are available at http://oldwww.com.dtu.dk/teletraffic/Erlang.html.) Erlang brought the Poisson distribution to telephony and gave an exact solution to the delay problem in the case of the single, overwhelmed switchboard operator.

Multi-server queues should obviously speed things up, provided they're used intelligently. (It would be nice, for example, if poky drivers on the Interstate would stay to the right.) Attendees at this January's Joint Mathematics Meetings in New Orleans got a vertical taste of some modern queueing technology at the new elevator banks in the headquarters hotel: Instead of simply pressing an up or down button to summon a car, riders were required to enter their destination on a keypad; a small screen then told them which of the several alphabetically labeled elevators to board. In principle, such a system will put, say, all six people going to the 13th floor onto one elevator, instead of having six separate cars make the same unlucky stop. (Some, presumably non-SIAM, mathematicians could be seen struggling on their ascent of the learning curve as they pondered the meaning of the keypad and its terse, monosymbolic instructions. Lively onboard discussions were common, focusing on the system's efficiency or perceived lack thereof, and on the algorithms and objective functions it might be using.)

Massey sees queueing theory as "a tandem network of simplifying modeling assumptions." Among the most stark of the assumptions is that the probabilistic nature of the line—the rate at which customers tend to arrive, the demands they make, and the rate at which servers process the workload—is unchanging. If the only variability is stochastic, with invariant statistics, the appropriate thing to analyze is the steady-state behavior of the queue.

If, for example, a single processor gives equal attention to all current customers, who arrive in Poisson fashion at rate r with average demand d (i.e., the probability of a new customer joining the queue in a time interval of duration Δt is $r\Delta t$, and the average job, if given the processor's undivided attention, takes d units of time), the fraction of the time during which n customers will be in the queue is $(1 - p)p^n$, where $p = rd$, provided that $p < 1$, and a customer arriving with a job of size x can expect a "sojourn time" of $x/(1 - p)$. (If $p \geq 1$, the queue just gets longer and longer, so that with the possible exception of a few early birds, no customer's job is ever finished in a reasonable amount of time.)

Real-life lines are not so orderly. In addition to stochastic variations, they are usually subject to *predictable* variability: anticipated changes in a queue's statistics, such as diurnal or seasonal fluctuations or a one-time switch when, say, a new call center opens. Procedures designed to optimize steady-state processing can still be used, of course—you can do anything you want if you don't care how well it works—but they're unlikely to produce optimal outcomes. The proper management of time-varying queues calls for new mathematical tools.

Or maybe creative application of some old ones.

Queueing theory has many analogs in fluid dynamics. Massey and colleagues, including graduate student Robert Hampshire, have studied

fluid limits for queues with time-varying rates, using a technique of "uniform acceleration" pioneered by Massey in his 1981 PhD dissertation. In uniform acceleration for a single-server queue, the instantaneous arrival and processing rates λ and μ are both scaled by a factor η, and the scaling effect on the number of customers in the queue is considered in the limit as η tends to infinity. If λ and μ are constants and λ is the smaller of the two rates, there is no scaling effect; the limit is simply the average number of customers in the steady-state limit. But when these rates are variable, and in particular in cases where $\lambda(t)$ is occasionally greater than $\mu(t)$ (which generates periods of overload that can persist even when $\lambda(t)$ reverts to being the smaller of the two), the number of customers does scale with η. Dividing by η gives, by definition, the fluid limit (which is 0 if the queue is underloaded).

Uniform acceleration for a multiple-server queue can be thought of as supply (the number of servers) keeping pace with demand (the arrival rate λ) in a growing economy (η). Alternatively, a single processor that gives equal attention to all current jobs can view uniform acceleration as the quantization of jobs into ever finer pieces as the number of full jobs is proportionally scaled upward. These scaling limits transform the stochastic model into a deterministic process and simplify the analysis of the expected sojourn time and other variables of interest. The fluid limit, of course, only approximates the actual behavior of the original stochastic model, but the asymptotic nature of the results means that these fluid limit models become more accurate as the actual demand and supply rates become larger. Massey and colleagues have found that their fluid limit formulas agree with results from simulations for a variety of examples (see Figure 1).

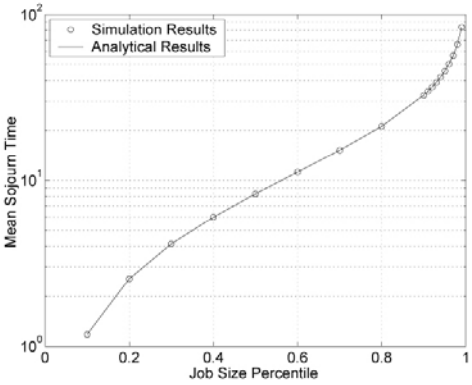

Figure 1. *Go with the flow. A comparison of analytic and simulated results for the mean sojourn time for a queue with sinusoidally fluctuating arrival rate. (From "Fluid and diffusion limits for transient sojourn times of processor sharing queues with time varying rates" by Robert C. Hampshire, Mor Harchol-Balter, and William A. Massey, in* Queueing Systems: Theory and Applications, *June 2006.)*

These fluid limits, Massey points out, also give queueing theory a rich set of analogies with classical mechanics. The length of a queue (i.e., the number of customers in it) at time t is analogous to the generalized coordinate $q(t)$, and the customer flow rate to the velocity $\dot{q}(t)$. The "value rate" of a queue—that is, its rate of net profit (or cost of operation)—is analogous to the Lagrangian $L(q,\dot{q},t)$. (In particular, the distinction between customers lost to blocking or abandonment and those being actively served has a lot in common with the distinction between potential and kinetic energy.) The opportunity cost per customer is like the conjugate momentum $p(t) = dL/d\dot{q}$, and the opportunity cost rate is like the Hamiltonian $H(p,q,t)$. Finally, the principle of least action in classical mechanics carries over to the "Bellman value function," which applies generally to issues in optimal control.

Fluid limit analysis lends itself to applications, as Massey and Hampshire have shown in a study of call-center staffing. In their model a call center consists of a variable number of agents $L(t)$ and a variable number of additional phone lines $K(t)$, which "answer" incoming calls by putting the caller on hold and, stereotypically, playing music. The queueing process is characterized by a time-varying Poisson arrival rate $\lambda(t)$, an exponential service rate μ (assumed to be constant in the model), and exponential abandonment rates β for callers who get fed up listening to music and γ for those who get busy signals (which happens when the number of callers exceeds $K + L$) and don't bother to redial. The staffing problem is to find functions K and L that optimize the profit of the call center given cost functions $c(L)$ and $d(K + L)$ for staffing and provisioning, a per-customer reward for service completion, and per-customer penalties for music and busy-signal abandonments.

Massey and Hampshire have shown that in the fluid limit, the optimal functions $K^*(t)$ and $L^*(t)$ are related to a system of "competing" Euler–Lagrange equations. An amusing quirk of their model is that K^* and L^* are complementary variables, meaning that their product is 0. This implies, perversely but believably, that an efficiently operating (in terms of profit optimality) call center might intentionally schedule times when there are phone lines to accept calls but no one to answer them!

Barry A. Cipra is a mathematician and writer based in Northfield, Minnesota.