The Performance Impact of Delay Announcements

Taking Account of Customer Response

IEOR 4615, Service Engineering, Professor Whitt

Supplement to Lecture 21, April 21, 2015

Review: The Purpose of Delay Announcements

A service system manager may want to tell each customer an estimate of the delay that customer will experience in order to:

- **Improve customer satisfaction.**
  - **Uncertain Waits** are longer than known finite waits.
- **Improve performance for the customers who are served.**
  - **By inducing some customers to balk or abandon earlier and then retry later** when the system is more lightly loaded.

But what will be the performance impact of the delay announcements?

A Model of Customer Response

- Make **delay announcement** (single-number w) to each new arrival, with number depending upon system state at that time.
- But we need to consider the **customer response**
  - Assume: **balk** (leave immediately) with probability $B(w)$
  - If join, **abandon** before time t with probability $F(t|w)$
- Need to consider **equilibrium:**
  1. Customer response depends on announcement.
  2. Announcement depends upon system state or history (performance).
  3. System performance depends upon customer response.

Example: The All-Exponential Response Model

- **delay announcement** w to each new arrival,

  with number w depending upon system state at that time.
- **customer response**
  - **balk** (leave immediately) with probability $B(w) = 1 - e^{-\beta w}$ for constant $\beta > 0$.
  - If join, **abandon** before time t with probability

$$
\begin{aligned}
F(t|w) &= 1 - e^{-\gamma t}, \quad 0 \leq t \leq w, \\
&= 1 - e^{-\gamma w} e^{-\delta t}, \quad 0 < w \leq t < \infty,
\end{aligned}
$$

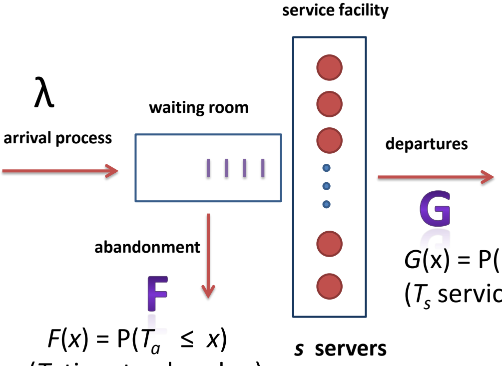
where γ and δ are positive constants.

Problem for Today

- **What to announce?**
  - Delay of Last to Enter Service (**LES**)
  - Fixed Deterministic (**FD**) Announcement, corresponding to equilibrium expected delay
- **How to study performance impact?**
  - deterministic **fluid model**
  - **simulation**
  - iterative numerical algorithm (for FD) for $M/M/s + GI$ using "engineering solution" (approximation) from WW (2005)

Review: The G/GI/s+GI Fluid Model

Approximation for the **G/GI/s + GI** **Stochastic Queueing Model**



$G(x) = P(T_s \le x)$

(T_s service time)

$F(x) = P(T_a \le x)$

(T_a time to abandon)

Many-Server Heavy-Traffic (MSHT) Limit

Increasing Scale

Increasing Scale

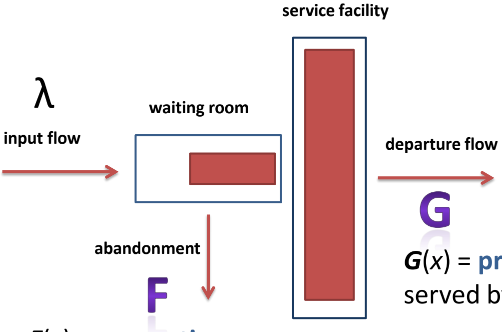
- a sequence of $G/GI/s + GI$ models indexed by n,
- arrival rate **grows**: $\lambda_n/n \to \lambda \quad \text{as} \quad n \to \infty$,

  number of servers **grows**: $s_n/n \to s \quad \text{as} \quad n \to \infty$,
- service-time cdf G and patience cdf F held **fixed** independent of n

  with mean service time 1: $\mu^{-1} \equiv \int_0^\infty x\, dG(x) \equiv 1$.

The G/GI/s+GI Fluid Model

Model data: (λ, s, G, F) and initial conditions.



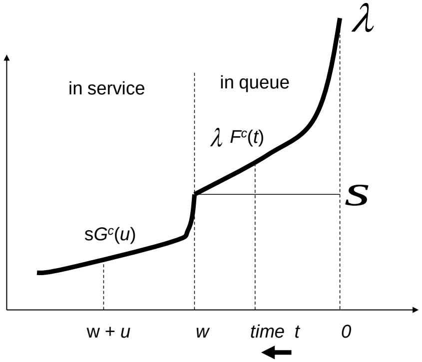
G(*x*) = **proportion** served by time *x*

F(*x*) = **proportion** abandoning by time *x*

capacity S

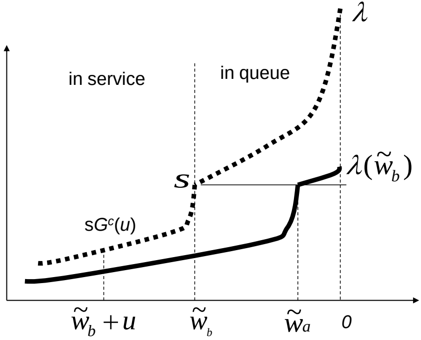
The Overloaded Fluid Model in Steady State

fluid density arriving time t in the past

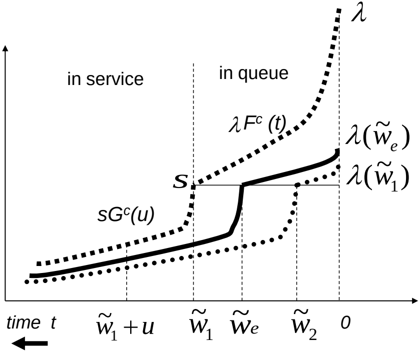


Use Fluid Model to Study Performance Impact

Direct Response to Delay Announcement



Equilibrium Delay



Equilibrium Delay in Overloaded Fluid Model

Recall that : $\lambda > s$ and $\mu = 1$.

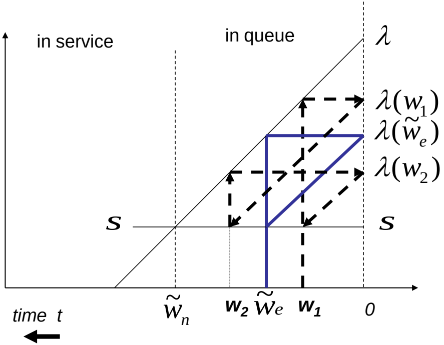
- **Without announcements:** $\lambda F^c(w) = s$
  - M/GI/s+M Model: $\lambda e^{-\theta w} = s$ and $w = \frac{\log(\rho)}{\theta}$, $\rho = \lambda/s$
- **With announcements:** $\lambda B^c(w) F^c(w|w) = s$
  - all-exponential customer response model:

$$\lambda e^{-\beta w_{eq}} e^{-\gamma w_{eq}} = s \quad \text{and} \quad w_{eq} = \frac{\log(\rho)}{\beta + \gamma}$$

- It is reasonable to expect that $\beta + \gamma > \theta$, so that announcements reduce the delay of served customers.

Impact of Delay Announcement in Fluid Model

Cycling Around the Equilibrium Delay



Conclusions of Study, Armony et al. 2009

- Under general conditions, there exists a unique equilibrium delay in the fluid model.
- Direct iteration $w_{n+1} = d(w_n)$ as shown above can lead to oscillations.
- Damped iterations produce convergence: $w_{n+1} = pd(w_n) + (1-p)w_n$
- LES delay and fluid equilibrium delay both work well in simulations

  But LES (1) lower variance and (2) more robust, does not depend on the model.

Experiments: Numerical Comparisons

- Consider overloaded **M/M/s+M** queueing model.
- Consider **all-exponential customer response** model.
- Queueing Model Parameters: $\lambda = 140$, $\mu = 1.0$, $s = 100$, $\theta = 1.0$
- Customer Response Parameters: $\beta = 1.0$, $\gamma = 0.5$, $\delta = 0.5$ and 4.0
- Performance with and without announcements
- Algorithms compared to simulation

Performance With and Without Announcements

Fluid and algorithm for $M/M/100 + M$ with fixed announcement: $\rho = 1.4$

perform. measure	no exact	announce fluid	with $\delta = 0.5$	announce $\delta = 4.0$	$w_{eq} = 0.224$ fluid
P(Balk)	0.000	0.000	0.201	0.201	0.201
P(abandon)	0.286	0.286	0.086	0.087	0.085
$P(B \cup A)$	0.286	0.286	0.287	0.288	0.286
$E[Q]$	40.0	40.0	24.3	17.3	23.7
$E[W\|S]$	0.332	0.336	0.225	0.157	0.224
$SD[W\|S]$	0.100	0.000	0.134	0.088	0.000

Simulation Results With Announcements

Algorithm (INA) and simulation for $M/M/100+M$ with LES and fixed.

perform. measure	$\delta = 0.5$ INA	sim(fixed)	LES	$\delta = 4.0$ INA	sim(fixed)	LES
P(Balk)	0.201	0.201	0.199	0.149	0.144	0.153
P(abandon)	0.086	0.087	0.086	0.137	0.143	0.132
$P(B \cup A)$	0.287	0.288	0.285	0.286	0.287	0.285
$E[Q]$	24.3	24.3	24.2	18.8	18.5	19.4
$E[W\|S]$	0.225	0.225	0.226	0.162	0.155	0.169
$SD[W\|S]$	0.134	0.133	0.091	0.066	0.066	0.072
$E[\|W - W_a\|\|S]$		0.108	0.055		0.052	0.039

SUMMARY

1. We have studied the performance impact of delay announcements.
2. We introduced a model of customer response.
3. The long-run performance is an equilibrium.
4. Fluid, INA and simulation reveal performance of LES.
5. LES and fixed announcements have nearly the same mean.
6. LES is more accurate, i.e., $E[|W - W_a||S]$ is lower.

References

- M. Armony, N. Shimkin and WW. **The Impact of Delay Announcements in Many-Server Queues with Abandonment**. *Operations Research* 57 (2009) 66–81.
- WW. **Improving Service By Informing Customers About Anticipated Delays**. *Management Science* 45 (1999) 192–207.
- WW. **Engineering Solution of a Basic Call-Center Model**. *Management Science* 51 (2005) 221–235.
- WW. **Fluid Models for Multiserver Queues with Abandonments**. *Operations Research* 54 (2006) 37–54. (See lecture 10.)

More References on Fluid Models

- Y. Liu and WW. **The $G_t/GI/s_t + GI$ Many-Server Fluid Queue.** *Queueing Systems* 71 (2012) 405–444.
- Y. Liu and WW. **A Network of Time-Varying Many-Server Fluid Queues with Customer Abandonment.** *Oper. Res.* 59 (2011) 835–846.
- Y. Liu and WW. **Algorithms for Time-Varying Networks of Many-Server Fluid Queues.** *Informs J. on Computing* 26 (2014) 59–73.
- Y. Liu and WW. **A Many-Server Fluid Limit for the $G_t/GI/s_t + GI$ Queueing Model Experiencing Periods of Overloading.** *Operations Research Letters* 40 (2012) 307–312.