The "Fluid View"
or Flow Models of Service Networks

Service Engineering (Science, Management)

December, 2006

1 Predictable Variability in Time-Varying Services

Time-varying demand and time-varying capacity are common-place in service operations. Sometimes, *predictable* variability (eg. peak demand of about 1250 calls on Mondays between 10:00-10:30, on a regular basis) dominates stochastic variability (i.e. random fluctuations around the 1250 demand level). In such cases, it is useful to model the service system as a deterministic *fluid model*, which transportation engineers standardly practice. We shall study such fluid models, which will provide us with our first mathematical model of a service-station.

A common practice in many service operations, notably call centers and hospitals, is to time-vary staffing in response to time-varying demand. We shall be using fluid-models to help determine time-varying staffing levels that adhere to some pre-determined criterion. One such criterion is "minimize costs of staffing plus the cost of poor service-quality", as will be described in our fluid-classes.

Another criterion, which is more subtle, strives for *time-stable* performance in the face of *time-varying* demand. We shall accommodate this criterion in the future (in the context of what will be called "the square-root rule" for staffing). For now, let me just say that the analysis of this criterion helped me also understand a phenomenon that has frustrated me over many years, which I summarize as "The Right Answer for the Wrong Reasons", namely: how come so many call centers enjoy a rather acceptable and often good performance, despite the fact that their managers noticeably lack any "stochastic" understanding (in other words, they are using a "Fluid-View" of their systems).

2 Fluid/Flow Models of Service Networks

We have discussed why it is natural to view a service network as a queueing network. Prevalent models of the latter are *stochastic* (random), in that they acknowledge *uncertainty* as being a central characteristic. It turned out, however, that viewing a queueing network through a "deterministic eye", animating it as a *fluid network*, is often appropriate and useful. For example, the Fluid View often suffices for bottleneck (capacity) analysis (the "Can we do it?" step, which is the first step in analyzing a dynamic stochastic network); for motivating congestion laws (eg. Little's Law, or "Why peak congestion lags behind peak load"); and for devising (first-cut) staffing levels (which are sometime last-cut as well).

Some illuminating "Fluid" quotes:

- "Reducing letter delays in post-offices": "Variation in mail flow are not so much due to random fluctuations about a known mean as they are time-variations in the mean itself ... Major contributor to letter delay within a postoffice is the shape of the input flow rate: about 70% of all letter mail enters a post office within 4-hour period". (From Oliver and Samuel, a classical 1962 OR paper).
- " ... a busy freeway toll plaza may have 8000 arrivals per hour, which would provide a coefficient of variation of just 0.011 for 1 hour. This means that a non-stationary Poisson arrivals pattern can be accurately approximated with a deterministic model". (Hall's textbook, pages 187-8). Note: the statement is based on a Poisson model, in which mean = variance.

There is a rich body of literature on Fluid Models. It originates in many sources, it takes many forms, and it is powerful when used properly. For example, the classical EOQ model takes a fluid view of an inventory system, and physicists have been analyzing macroscopic models for decades. Not surprisingly, however, the first explicit and influential advocate of the Fluid View to queueing systems is a Transportation Engineer (Gordon Newell, mentioned previously). To understand why this view was natural to Newell, just envision an airplane that is landing in an airport of a large city, at night - the view, in rush-hour, of the network of highways that surrounds the airport, as seen from the airplane, is precisely this fluid-view. (The influence of Newel1 is clear in Hall's book.)

Some main advantages of fluid-models, as I perceive them, are:

- They are simple (intuitive) to formulate, fit (empirically) and analyze (elementary). (See the Homework on Empirical Models.)
- They cover a broad spectrum of features, relatively effortlessly.
- Often, they are all that is needed, for example in analyzing capacity, bottlenecks or utilization profiles (as in National Cranberries Cooperative and HW2).
- They provide useful approximations that support both performance analysis and control. (The approximations are formalized as first-order deterministic fluid limits, via Functional (Strong) Laws of Large Numbers.)

Fluid models are intimately related to Empirical Models, which are created *directly* from measurements. As such, they constitute a natural first step in modeling a service network. Indeed, refining a fluid model of a service-station with the outcomes of Work (Time and Motion) Studies (classical Industrial Engineering), captured in terms of say histograms, gives rise to a (stochastic) model of that service station.

3 Some More Details on Fluid Models

The main roles that fluid models play in the world of Service Engineering, as my experience suggests, are as follows: fluid models are interesting and useful in their own right, they provide simple approximations to complicated systems, and they constitute powerful technical tools in the analysis of stochastic systems. Elaborating on these roles:

1. Legitimate models for real systems, with prevalent predictable variability that dominates stochastic variability (verified, for example, by small CV, or by averaging). Examples include (Newell; Hall; Bassamboo, Harrison and Zeevi; EOQ-like models, ...):

- Inventory buildup diagrams (See the Trucks in National Cranberries).
- Mean-value analysis (in Computer Science)
- Transportation engineers often "think fluid" (see Newell's book).
- Airport traffic (planes and people).
- Vandergraft, Hall on staffing.
- Service factories, for example mail-sorting.

2. Useful approximations: first-order deterministic fluid approximations, via Functional (Strong) Laws of Large Numbers (FLLN), to support both performance analysis and control.

- Long-run, detects trends. (See Chen and M.)
- Identify bottlenecks (eg. National Cranberries.)
- Traffic equations, for example in Jackson networks.
- Short-run, captures instantaneous (predictable) variability (Massey, Pats; Bassamboo, Harrison, Zeevi).
- Identify phases in evolution (see Hall, pg. 189-191: starting stagnant, then overloading with queues increasing, and then decreasing, back to stagnant.)

3. Technical Tools for characterizing stability and instability of stochastic networks. (Seminal article by Jim Dai, 1996; currently a very active research venue).

- Lyapounov functions: It is sometimes that case that sample paths of a stochastic system is attracted to Fluid sample-paths. This helps establish stability/instability, weak convergence or asymptotic-control optimality in a stochastic environment, but via a deterministic analysis.
- Mathematical framework for analysis and approximations (reflection), which is amenable to the use of the continuous mapping theorem.

Additional references on the Fluid View are provided in the reading packets within the syllabus.