Traffic Modeling for Complex Service Systems

Peter W. Glynn

Stanford University

I-Sim Workshop on Simulation in Complex Service Systems,
Montreal, July 18-20, 2011

Goal of Talk:

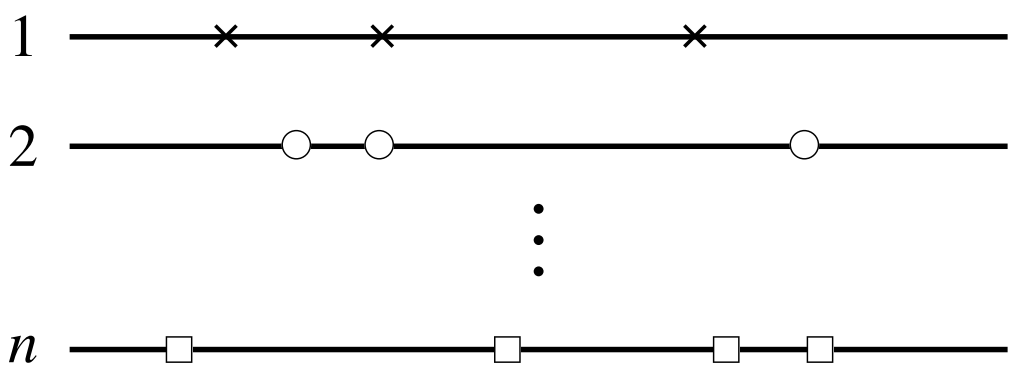
- To provide an overview of some issues that arise in the traffic modeling context
- More questions than answers!

What is traffic modeling?

- A description of the incoming work to a system (usually exogenously defined)
- Service time requirements are very domain-specific
- What about arrivals?

Poisson Arrival Streams:

Key result is the "superposition theorem" (Palm-Khintchine)



$$N_n(\cdot) = \sum_{i=1}^{n} N_{ni}(\cdot)$$

In great generality,

$$N_n(\cdot) \Rightarrow N_\infty(\cdot)$$

where $N_\infty(\cdot)$ is a Poisson process

Question:

- Does a renewal arrival model (i.e. with iid inter-arrival times) ever make sense as a realistic description of arrivals?
- Other than Poisson case...

probably not!

Another General Principle:

In "high volume" traffic settings, traffic often looks Gaussian.

Formulation 1: High volume induced by a single source sending traffic at a fast rate

$$N_n(\cdot) = N(n\cdot)$$

Then,

$$N_n(t) \stackrel{\mathcal{D}}{\approx} \lambda nt + \sqrt{n}\sigma B(t)$$

(Invariance Principle)

Formulation 2: High volume induced by many sources independently generating traffic

$$N_n(\cdot) = \sum_{i=1}^{n} N_{ni}(\cdot)$$

Then,

$$N_n(t) \stackrel{\mathcal{D}}{\approx} \lambda n t + \sqrt{n} Z(t),$$

where $Z(\cdot)$ is a Gaussian process for which

$$\mathrm{cov}(Z(s), Z(t)) = \mathrm{cov}(N_{n1}(s), N_{n1}(t))$$

In either formulation, high volume implies:

- arrival counting process is approximately Gaussian
- standard deviation is of order of $\sqrt{\text{arrival rate}}$
  [Leads to "square root" staffing formula in call center setting]
- conclusions generalize in a natural way to non-stationary context

All well and good... are things really this simple?

Poisson superposition result:

$$N_n(\cdot) \stackrel{\mathcal{D}}{\approx} \text{Poisson Process}$$

High volume arrival processes:

$$N_n(\cdot) \stackrel{\mathcal{D}}{\approx} \lambda nt + \sqrt{n}Z(t)$$

variances don't match! (In Poisson process, variance = mean)

Back to superposition principle:

$$N_{ni}(\cdot) = \widetilde{N}_i(\cdot/n)$$

where $\widetilde{N}_1, \widetilde{N}_2, \ldots$ is a sequence of iid point processes with common intensity λ.

Then:

$$\sup_{A \in \mathcal{F}_t} \left| P(N_n \in \cdot) - P(N_\infty \in \cdot) \right| \to \begin{cases} 0 & \text{if } t \ll n \\ 1 & \text{if } t \gg n \end{cases}$$

- For an order 1 intensity point process, N_n is only Poisson over time scales of order $o(n)$; non-Poisson behavior manifests over longer time scales
- For a high intensity point process, N_n is only Poisson over time scale of order $o(1)$; non-Poisson behavior manifests itself over $O(1)$ time scales and longer
- In statistically testing Poisson assumption at level of inter-arrivals, one will miss the longer time scale non-Poisson behavior

Note that the superposition theorem describes:

> a superposition of many independent sources, each of which is low-intensity and has large "inter point" distances

This does not describe, for example, Internet trafic:

> Flow-level source model:
>
> - flow is initiated according to a low intensity point process
> - once flow is initiated, packets are generated rapidly and regularly

Poisson process fails to describe aggregate packet stream, no matter how fine the time scale.

In high-volume arrival setting (even in presence of time-of-day effects), standard models predict:

- Gaussian count data (over order 1 time scales)
- very low coefficient of variation (standard deviation of order $\sqrt{\text{arrival rate}}$)

Sometimes not reflected in statistical analysis of real data

Conclusion: "Standard models" used to describe high intensity arrival streams are sometimes wrong!

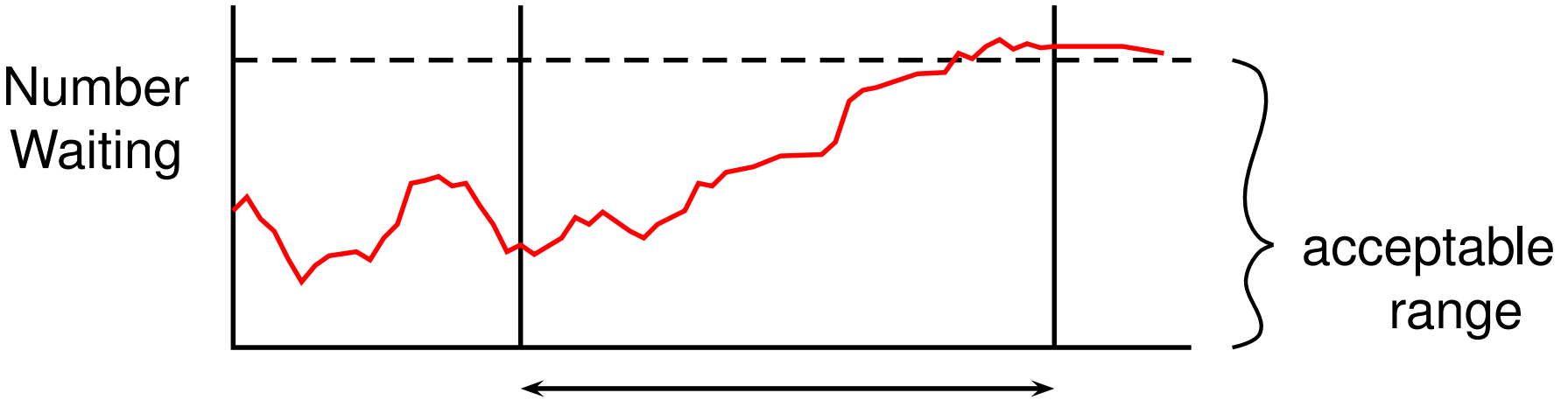
[Puts in question "square root staffing formula"?]

One possible remedy:

Stochastic modulation of arrival rate / volatility / etc
over order 1 time scales (in high volume settings)

Note: we are talking about stochastic modulation not at time scale of individual inter-arrival times, but over much longer scales

Practically speaking, what is $O(1)$ time scale?



The critical time scale is:

> The typical time taken by the system to move from acceptable equilibrium behavior to unacceptable performance region

Remarks:

- In heavily congested systems in "heavy traffic", this generates a time scale of right magnitude

$$O\left(\frac{1}{(1-\rho)^2}\right)$$

- in loss systems involving buffers of order b, this generates a time scale of order b [consistent with large deviations path / statistics]

- Finer scale behavior (e.g. local Poisson structure) is probably less important

Question: What is right class of models that flexibly describe long time scale behavior, look "locally Poisson" at finer time scales, are easy to fit, and are easy to simulate?

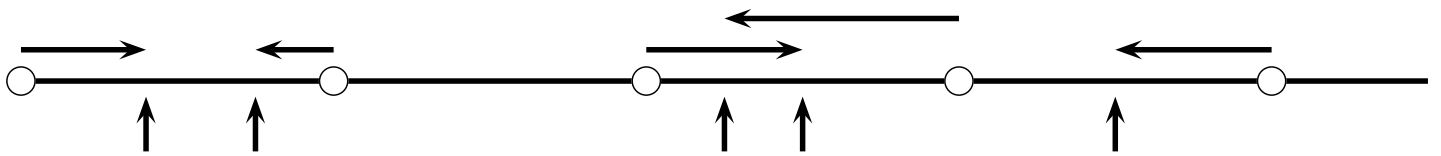
$$N(\cdot) = \widetilde{N}(\Lambda(\cdot))?$$

["doubly stochastic Poisson process"]

An Important Arrival Model for the Simulation Tool-kit

Consider a medical clinic:

- Schedule customers to arrive at 15 minute time intervals
- Customers arrive early / late relative to appointment time



$$N(0,t] = \sum_{-\infty}^{\infty} I(nh + \xi_n + (0,t])$$

ξ: iid

"scheduled traffic"

- This process is highly non-Poisson / non-renewal
- Negatively correlated increments

$$N(0,t] - \lfloor t/h \rfloor \Rightarrow Z_\infty$$

as $t \to \infty$

- Essentially deterministic in "heavy traffic"; but distinctly different in light and moderate traffic

Recommendation:

Should be in "tool-kit" of arrival processes

Question:

How to fit distribution of ξ_n when arrivals are "unlabelled"?

Abandonment Modeling (for Call Centers)

- High volume environment
- In QED setting, many customers are waiting, say $X(t)$.
- Abandonment has an enormous impact on performance
- A variant of superposition principle asserts that cumulative number $N(t)$ of abandonments setisfies

$$dN(t) \stackrel{\mathcal{D}}{\approx} d\text{Poisson}(\gamma X(t))$$

- How to fit γ?

$$\hat{\gamma} = \frac{N(t)}{\int_0^t X(s)ds}$$

What about if waiting times are announced?

Impact on abandonment?

What about if customers who agree to call back later are entered in a lottery to win a small prize?

How to account for human behavior?

Other Issues:

- Use of "coarse grain models" as planning tools for "fine grain simulations"

  e.g. use of fluid and Brownian models as approximations to more complex non-stationary queueing models

- Initialization of real-time simulations

  e.g. simulation model often contains unobserved state variables

- Many other interesting and important questions that arise