Lecture 9: Deterministic Fluid Models and Many-Server Heavy-Traffic Limits

IEOR 4615: Service Engineering

Professor Whitt

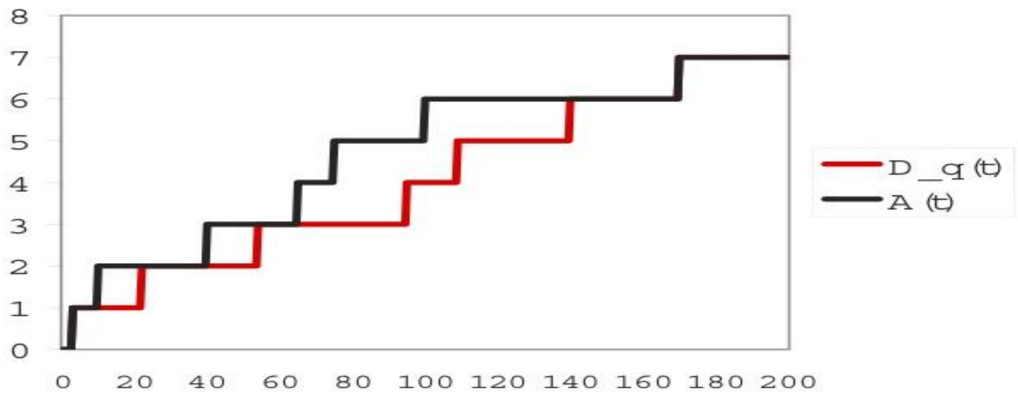
February 19, 2015

Outline

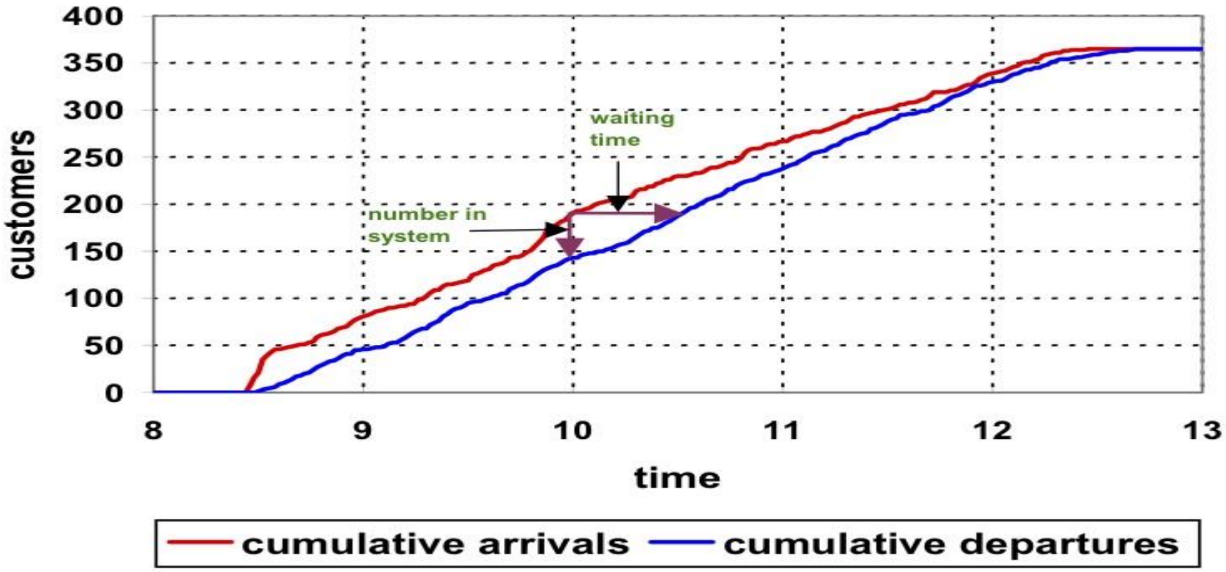
- Deterministic **Fluid Models**
  - Directly From Data: Cumulative Arrivals and Departures
  - Directly from $M_t/M/s_t+M$ BD Model (deterministic view)
- Many-Server Heavy-Traffic **Limits** for Queueing Models
  - Fluid Model Obtained in the Limit as Scale Increases
  - Ultimately, limits explained by LLN and CLT

From Data to Fluid Models

- To analyze data, we plot ***cumulative*** arrival and departure functions:



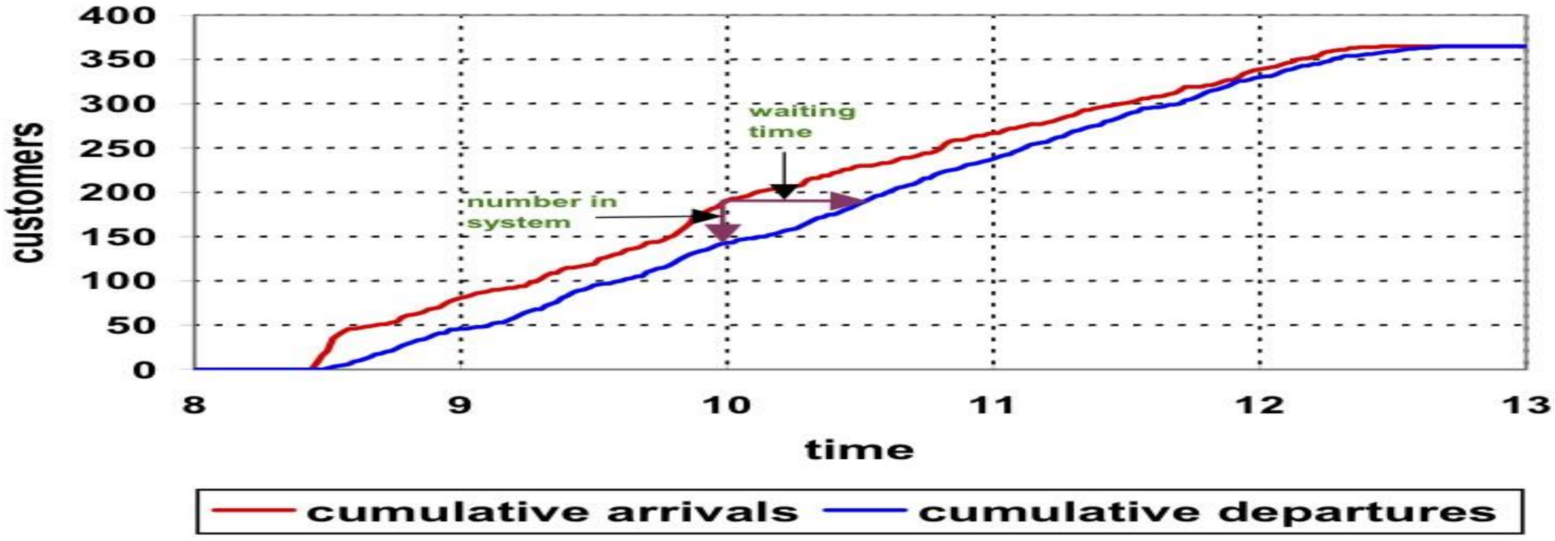
- For large systems (bird's eye view), the functions look smoother.



From Data to Fluid Models

- Directly from event-based (call-by-call) measurements.
- For example, an isolated service-station:
  - A(t) = cumulative # arrivals from time 0 to time t;
  - D(t) = cumulative # departures from system during [0, t];
  - L(t) = A(T) − D(t) = # customers in system at t.

Arrivals and Departures from a Bank Branch Face-to-Face Service



Deterministic Fluid Model

- Describe impact of Predictable Variability
  - Time-Varying Arrival Rate
  - Ignoring Stochastic Variability
- Idealistic Smooth Model
  - Ordinary Differential Equation (ODE)

Phases of Congestion

Hall, textbook:

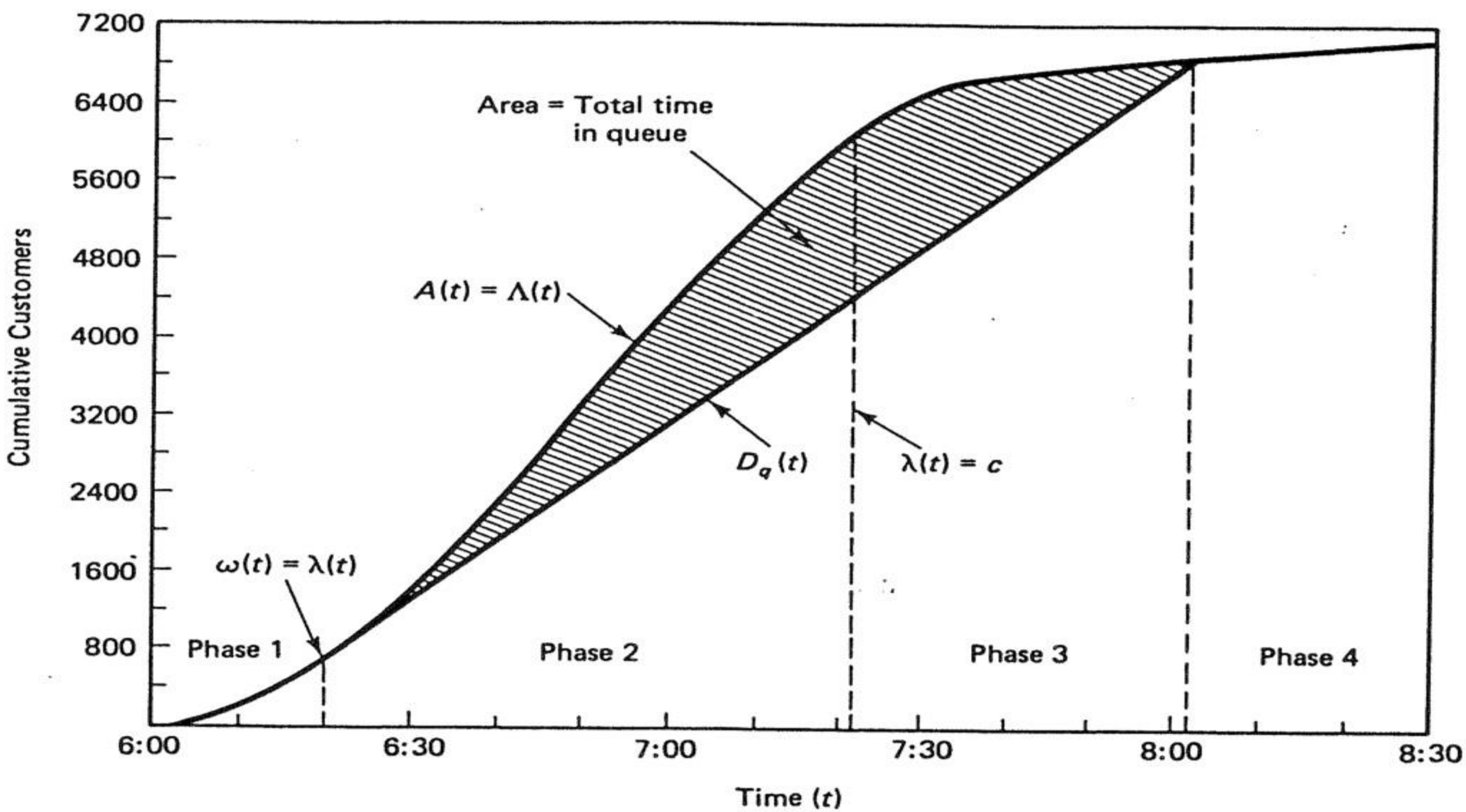
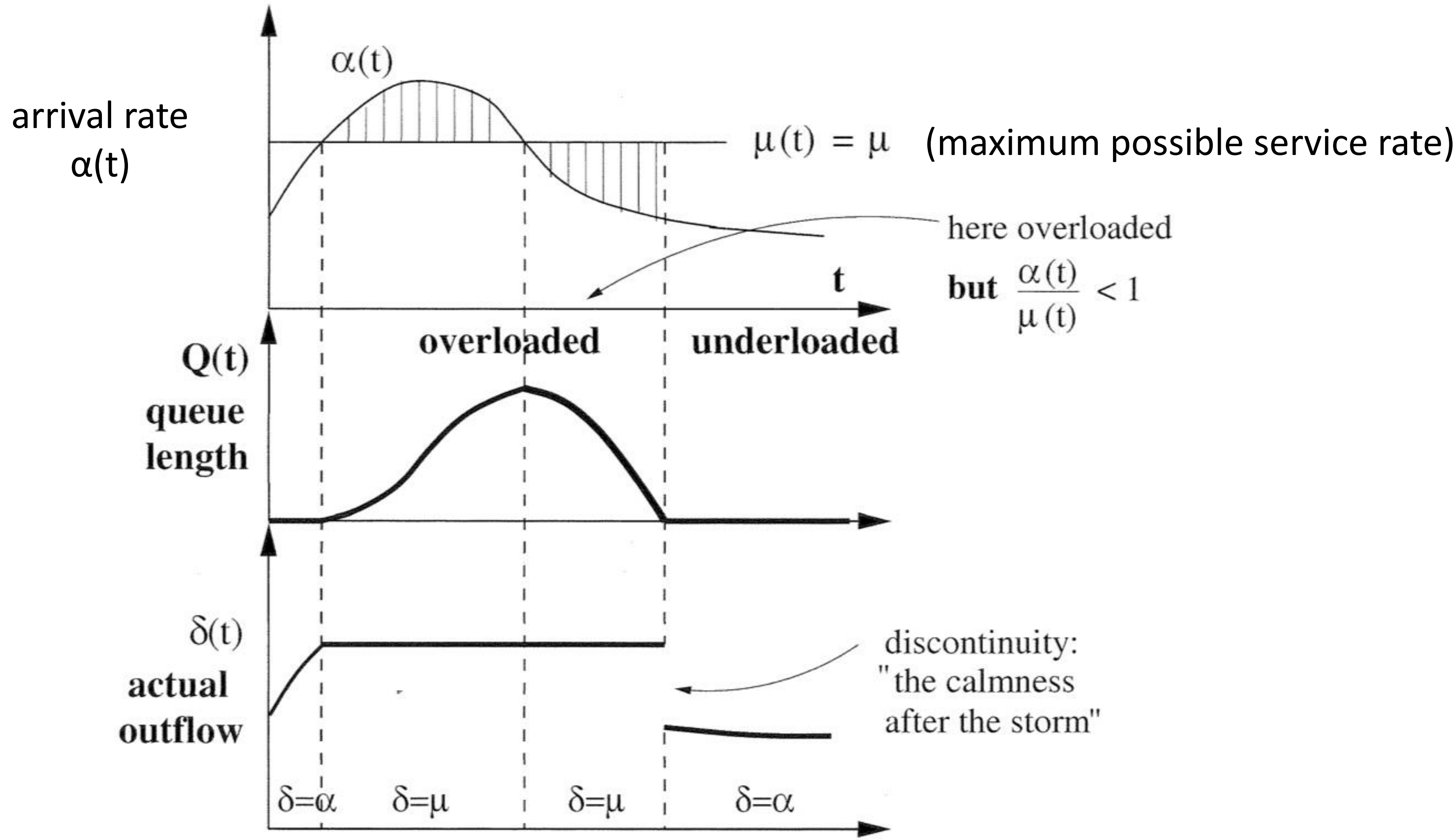


Figure 6.6 Cumulative diagram illustrating deterministic fluid model. When a queue exists, customers depart at a constant rate. Queues increase when the arrival rate exceeds the service capacity and decrease when the service capacity exceeds the arrival rate.

Four points of view

- Cumulative Arrivals and Departures
- Rates ($\Rightarrow$ Peak Load)
- Queues ($\Rightarrow$ Congestion)
- Outflows ($\Rightarrow$ end of rush-hour)

Phases of Congestion via Rates



- Time lag in congestion after peak arrival rate
- Changing Departure Rate

Mathematical Fluid Models: General Setup

Queueing System as a Tub (Hall, p.188)

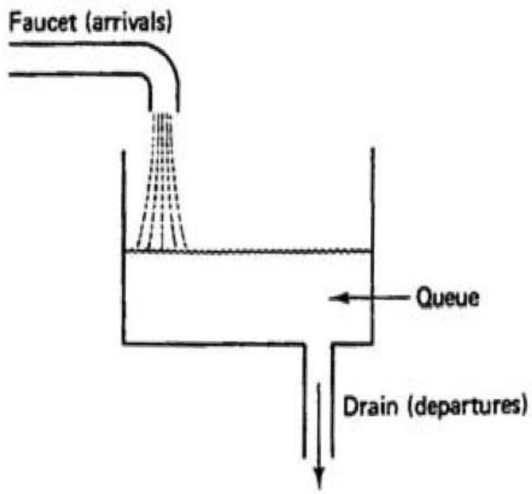


Figure 6.5 In a fluid model, the customers can be viewed as a liquid that accumulates in a tub. Queues increase when the fluid enters the tub faster than it leaves.

- $A(t)$ – cumulative arrivals function.
- $D(t)$ – cumulative departures function.
- $\lambda(t) = \dot{A}(t)$ – arrival rate (dot = **derivative** *d/dt*)
- $\delta(t) = \dot{D}(t)$ – processing (departure) rate.
- $c(t)$ – maximal potential processing rate.
- $q(t)$ – total amount in the system at time t.

Mathematical Fluid Model

Differential equation:

- $\lambda(t)$ – arrival rate at time $t \in [0,T]$.
- $c(t)$ – maximal potential processing rate.
- $\delta(t)$ – effective processing (departure) rate.
- $q(t)$ – total amount in the system at time t.

Then $q(t)$ is a solution of

$$\dot{q}(t) = \lambda(t) - \delta(t);\ \ q(0) = q_0, t \in [0,T].$$

Mathematical Fluid Model: Multi-Server Queue

- $s(t)$ statistically-identical servers, each with service rate μ.
- $c(t) = \mu s(t)$: maximal potential processing rate.
- $\delta(t) = \mu \cdot \min(s(t), q(t))$: processing rate.

$$\dot{q}(t) = \lambda(t) - \mu \cdot \min(s(t), q(t)); \;\; q(0) = q_0, t \in [0, T].$$

i.e., $$q(t) = q(0) + \int_0^t \lambda(u) du - \int_0^t \mu \cdot \min(s(u), q(u)) du.$$

How to actually solve? Discrete-time approximation:

Start with $t_0 = 0,\ q(t_0) = q_0.$ Then, for $t_n = t_{n-1} + \Delta t$:

$$q(t_n) = q(t_{n-1}) + \lambda(t_{n-1}) \cdot \Delta t - \mu \cdot \min(s(t_{n-1}), q(t_{n-1})) \cdot \Delta t.$$

Mathematical Fluid Model: Multi-Server Queue with Abandonment

- θ – Abandonment rate of customers in queue
- Processing rate:

$$\delta(t) = \mu \cdot \min(s(t), q(t)) + \theta \cdot [q(t) - s(t)]^+$$

- The fluid model:

$$\dot{q}(t) = \lambda(t) - \mu \cdot \min(s(t), q(t)) - \theta \cdot [q(t) - s(t)]^+;$$
$$q(0) = q_0, t \in [0, T].$$

- Deterministic View of ***$M_t/M/s_t+M$*** BD Model
  - Parameters: **$\lambda(t)$, μ, $s(t)$, θ**

Many-Server Heavy-Traffic Limit

Sequence of $M_t/M/s_t+M$ Models Indexed by n

Let $n \to \infty$.

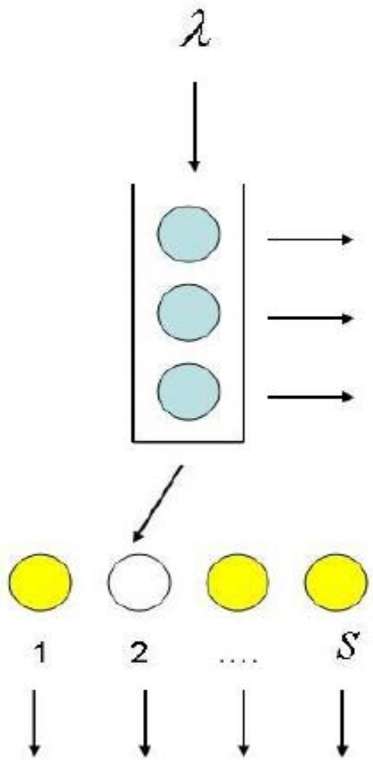
Parameters:

- $\lambda_n(t) = n\lambda(t)$ arrival rate at time t **[heavy-traffic]**
- $s_n(t) = ns(t)$ number of servers at time t **[large scale]**
- $\mu_n(t) = \mu$ individual service rate (constant)
- $\theta_n(t) = \theta$ individual abandonment rate (constant)

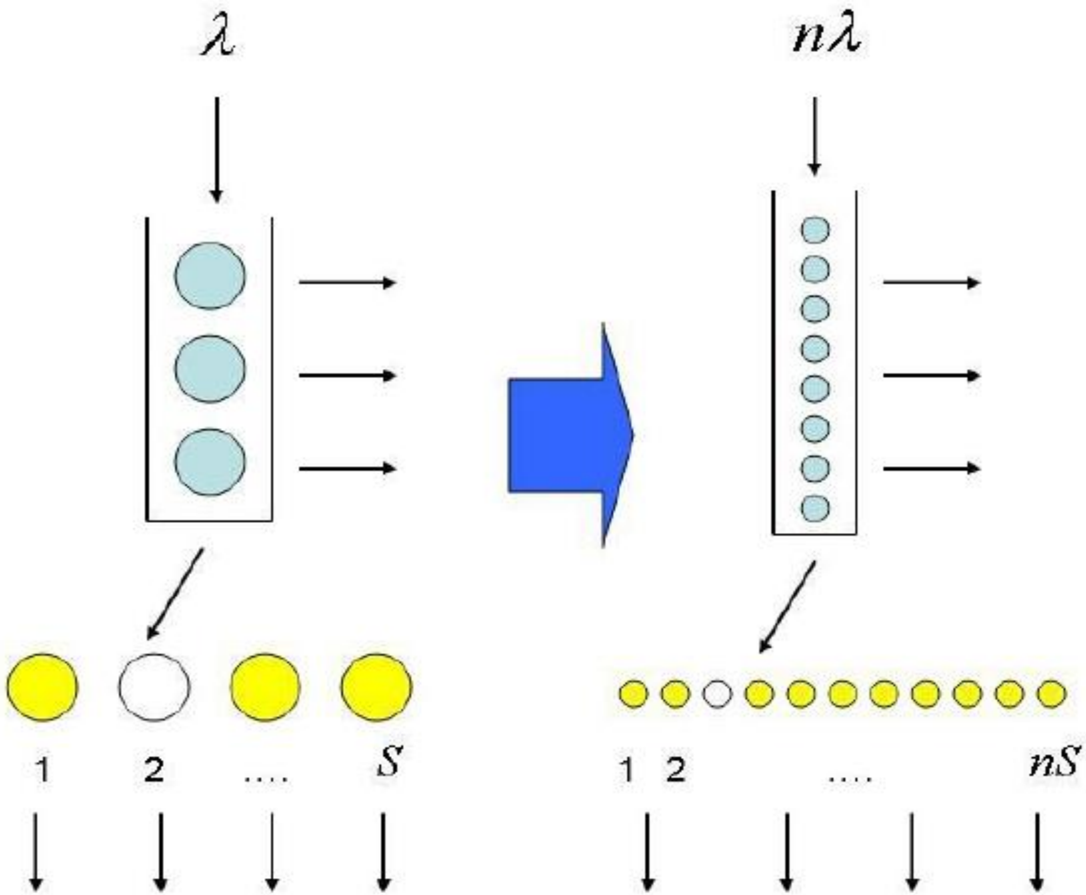
Stochastic Processes:

- $A_n(t)$ number of arrivals in system n in $[0,t]$
- $D_n(t)$ number of departures in system n in $[0,t]$
- $Q_n(t)$ number of customers in system n at time t
- $W_n(t)$ potential waiting for arrival at time t (with infinite patience)

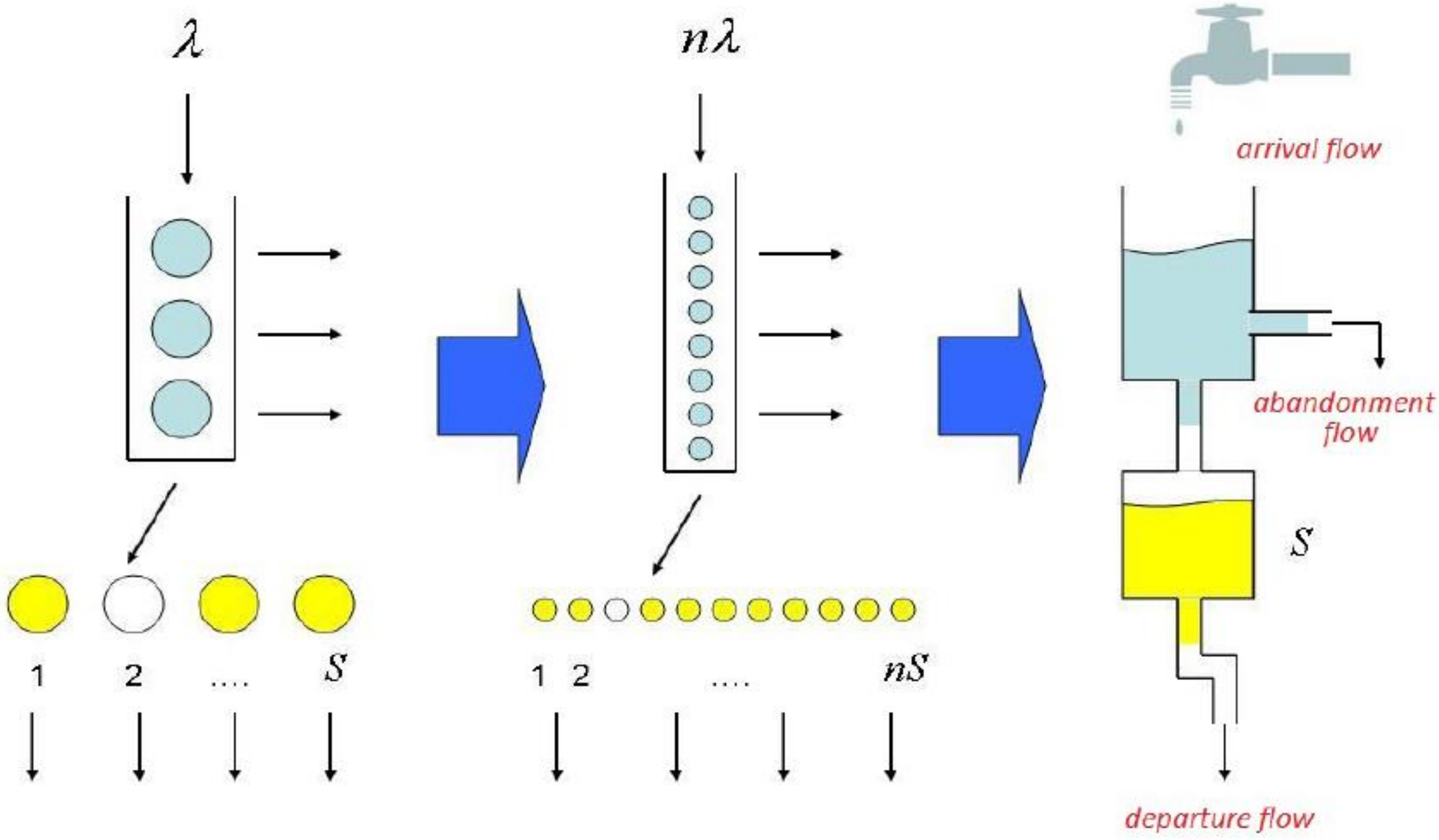
Fluid Approximation from Many-Server Heavy-Traffic Limit



Fluid Approximation from Many-Server Heavy-Traffic Limit



Fluid Approximation from Many-Server Heavy-Traffic Limit



Many-Server Heavy-Traffic Limits
Sequence of $M_t/M/s_t+M$ Models Indexed by n

Parameters:

- $\lambda_n(t) = n\lambda(t)$ arrival rate at time t **[heavy-traffic]**
- $s_n(t) = ns(t)$ number of servers at time t **[large scale]**
- $\mu_n(t) = \mu$ individual service rate (constant)
- $\theta_n(t) = \theta$ individual abandonment rate (constant)

Stochastic Processes:

- $A_n(t)$ number of arrivals in system n in $[0,t]$
- $D_n(t)$ number of departures in system n in $[0,t]$
- $Q_n(t)$ number of customers in system n at time t
- $W_n(t)$ potential waiting for arrival (with infinite patience)

Limits (Fluid Limit = Law of Large Numbers): As n $\rightarrow\infty$,

- $A_n(t)/n \rightarrow \Lambda(t) = \int^t \lambda(s)\, ds$ (integral of fluid arrival rate)
- $D_n(t)/n \rightarrow D(t) = \int^t \delta(s)\, ds$ (integral of fluid departure rate)
- $Q_n(t)/n \rightarrow q(t)$ fluid content at time t
- $W_n(t) \rightarrow w(t)$ potential waiting time for atom of fluid

Refined Many-Server Heavy-Traffic Limit
Sequence of Queueing Models Indexed by *n*
$M_t/M/s_t+M$

Fluid Limits (Fluid Limit = Law of Large Numbers): As $n \to \infty$,

- $A_n(t)/n \to \Lambda(t) = \int^t \lambda(s)\, ds$ (integral of fluid arrival rate)
- $D_n(t)/n \to D(t) = \int^t \delta(s)\, ds$ (integral of fluid departure rate)
- $Q_n(t)/n \to q(t)$ fluid content at time t
- $W_n(t) \to w(t)$ waiting time

Stochastic Limits (Gaussian Limit = Central limit Theorem):

As $n \to \infty$,

- $\sqrt{n}[(A_n(t)/n) - \Lambda(t)] \to X_A(t)$
- $\sqrt{n}[(D_n(t)/n) - D(t)] \to X_D(t)$ Gaussian limit processes
- $\sqrt{n}[(Q_n(t)/n) - q(t)] \to X_Q(t)$
- $\sqrt{n}[W_n(t) - w(t)] \to X_W(t)$

Three Many-Server Heavy-Traffic Limiting Regimes
Sequence of Stationary *M/M/s+M* Models

Parameters:

- $\lambda_n = n\lambda - c\sqrt{n}$ arrival rate at time t **[No time-varying parameters]**
- $s_n = ns$ number of servers at time t **[large scale]**
- $\mu_n = \mu$ individual service rate (constant)
- $\theta_n = \theta$ individual abandonment rate (constant)

Limiting Regimes

- $\lambda > s\mu$ overloaded or Efficiency-Driven (**ED**)
- $\lambda < s\mu$ underloaded or Quality-Driven (**QD**)
- $\lambda = s\mu$ critically loaded – need to look more closely

Expanding the Critically Loaded Regime: $\lambda = s\mu$

- More general arrival rate scaling:
- Quality-and-Efficiency-Driven (**QED**) regime = Halfin-Whitt (1981) regime
  - $\lambda_n = ns\mu[1 - (\beta/\sqrt{n})]$ ($\lambda = s\mu$ and $c = -s\mu\beta$)
  - $(1-\rho_n)\sqrt{n} = \beta$ where $\rho_n = \lambda_n/s_n\mu = \lambda_n/ns\mu$

Delay Probability Approximation in the $M/M/s/\infty$ Model in the QED Regime

- **Q**uality-and-**E**fficiency-**D**riven (**QED**) regime = Halfin-Whitt (1981) regime
  - $\lambda_n = ns\,\mu[1 - (\beta/\sqrt{n})]$
  - $(1-\rho_n)\sqrt{n} = \beta$ where $\rho_n = \lambda_n/s_n\mu = \lambda_n/ns\mu$
  - $P(W_n > 0) \to \alpha$ with $0 < \alpha < 1$. (W_n steady state wait before starting service in model n)

$$P(W_n > 0) \approx \alpha(\beta) = HW(\beta) = 1/[\beta\Phi(\beta)/\phi(\beta)]$$

Where $\Phi(x) = P(N(0,1) < x)$ standard normal cdf and $\phi(\beta)$ is the associated density function

Implications for Staffing in the $M/M/s/\infty$ Model

$$P(W_n > 0) \approx \text{Target} = \alpha = \alpha\,(\beta)$$
$$= \ HW(\beta) = \ 1/[\beta\Phi(\beta)/\phi(\beta)]$$

Where $\Phi(x) = P(N(0,1) < x)$ standard normal cdf and $\phi(\beta)$ is the associated density function

Use **Square-Root Staffing Formula**: Set

$$S = S(\lambda/\mu) = (\lambda/\mu) + \beta\,(\lambda/\mu)^{1/2}$$

For $\beta = HW^{-1}(\alpha)$ (inverse function)

λ/μ = offered load (infinite-server model)

Implications for Staffing in the $M/M/s+M$ Model

$$P(W_n > 0) \approx \text{Target} = \alpha = \alpha(\beta, \gamma)$$
$$= G(\beta, \gamma) = 1/[1 + \gamma h(\beta/\gamma) h(-\beta)]$$

Garnett function from Garnett et al. (2002)

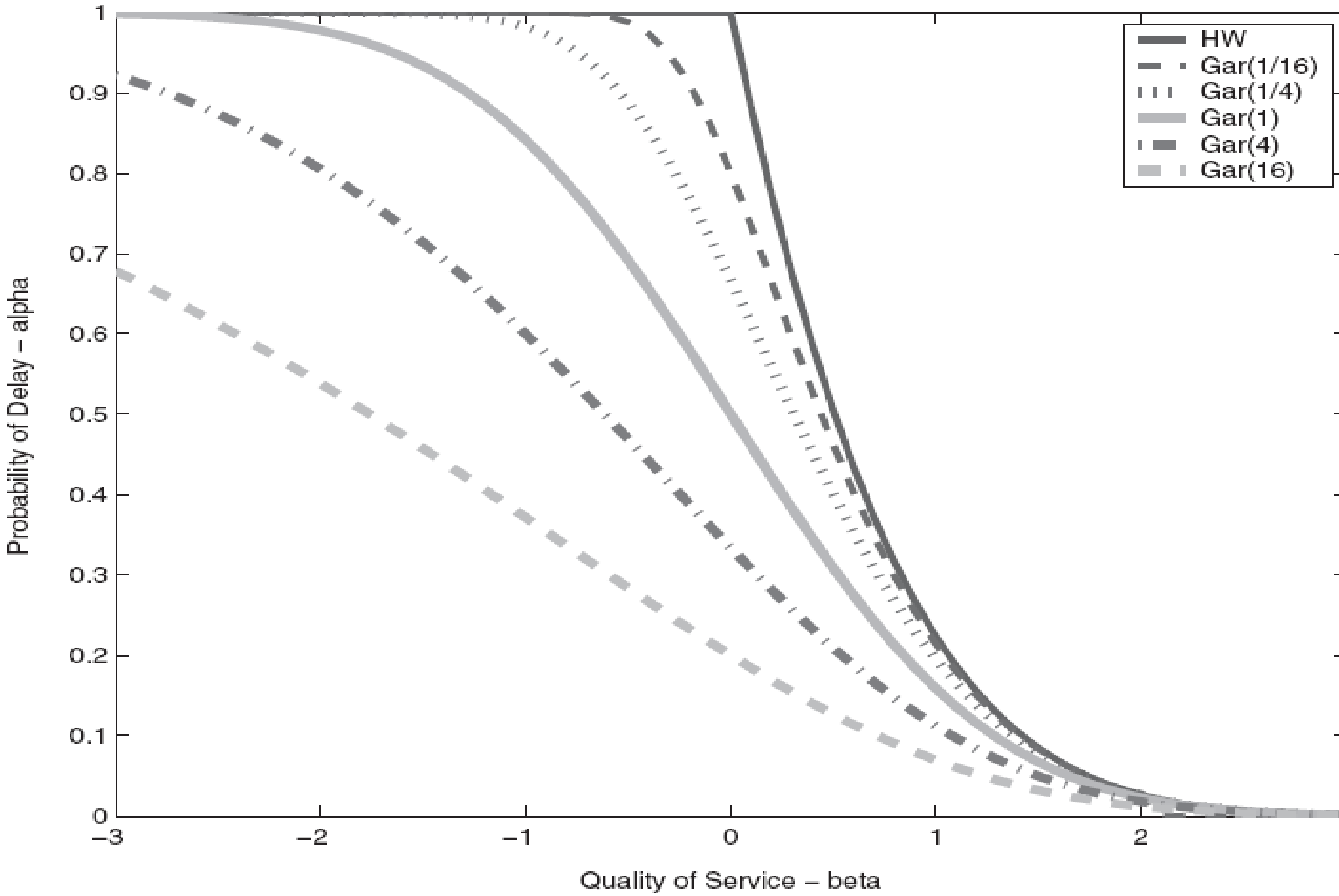
Where $\gamma = (\theta/\mu)^{1/2}$, $h(x) = \phi(x)/[1- \Phi(x)]$, $\Phi(x) = P(N(0,1) < x)$ standard normal cdf and $\phi(x)$ is the associated density function.

Use **Square-Root Staffing Formula**: Set

$$s = s(\lambda/\mu) = (\lambda/\mu) + \beta\,(\lambda/\mu)^{1/2}$$

For $\beta = G^{-1}(\text{Target}, \gamma)$

Figure **The Halfin–Whitt and Garnett Functions Mapping the QoS Parameter β into the Steady-State Delay Probability α. Five Different Values are Considered for the Parameter $\theta_{rat} \equiv \theta/\mu$: $\frac{1}{16}$, $\frac{1}{4}$, 1, 4, and 16.**



References
Fluid Models

1. Chapter 6, especially Section 6.4, of Hall (1991) *Queueing Methods for Services and Manufacturing,* Prentice Hall.
2. Chapters 1 and 2 of Newell (1982) *Applications of Queueing Theory,* second edition, Chapman and Hall.

Many-Server heavy-Traffic Limits

3. Halfin, S., W. Whitt. (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29, 567-588.
4. Whitt, W. (1992) Understanding the Efficiency of Multi-Server Service Systems. *Management Science*, 38, 708-723.

More Advanced References

Fluid Models

1. Hampshire, R. C., W. A. Massey. (2010). A tutorial on dynamic optimization with application to dynamic rate queues. *TutORials in Operations Research,* presented at the INFORMS National Meeting in Austin Texas. (www.princeton.edu~wmassey)

Many-Server Heavy-Traffic Limits

2. Garnett, O., A. Mandelbaum, M. I. Reiman. (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management 4, 208–227. (http://iew3.technion.ac.il/serveng)*

3. Pang, G., R. Talreja, W. Whitt. (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, 4, 193-267.