OUTLINE

- Heavy-Traffic (HT) Approximations for Queueing Models
  1. Many-Server HT Limits for Multi-Server Models ($\lambda \uparrow \infty$ and $s \uparrow \infty$)
     - How to Apply the Simple Formulas
     - Statement of the Limit Theorem
  2. Conventional HT Limits for Multi-Server Models ($\rho \uparrow 1$ or $\lambda \uparrow s\mu$)
     - How to Apply the Simple Formulas
     - Statement of the Limit Theorem

- Sketch of One Proof (The Conventional HT Limit)
  - The $M/M/1$ Queue Length Process
  - FCLT for Poisson processes
  - Application to get HT Limit
1. Question for $M/M/s + M$ Model

- Suppose that you are considering a call center that can be modeled as an $M/M/s + M$ queue (birth-and-death process model).

- Suppose that the parameter values are initially:
  \[(\lambda, s, \mu, \theta) = (25, 30, 1, 2)\].

- Suppose that an expansion is planned in which the arrival rate is increased from $\lambda = 25$ to $\lambda = 100$.

  - What new level of staffing, $s_{\text{new}}$, is needed to provide the same quality of service?

  - Is $s_{\text{new}} = s_{\text{old}} \times \frac{\lambda_{\text{new}}}{\lambda_{\text{old}}} = 30 \times \frac{100}{25} = 120$ good?
Answer: No, 120 is too high.

- There is an important economy of scale.
- We should use the Square Root Staffing (SRS) formula:
- Quality of Service (QoS) is initially $\beta = 1.0$, because

$$s = \left( \frac{\lambda}{\mu} \right) + \beta \left( \frac{\lambda}{\mu} \right)^{1/2} \quad \text{for} \quad \beta = 1.0$$

$$30 = 25 + 1.0 \times \sqrt{25}.$$

- When the arrival rate is increased from $\lambda = 25$ to $\lambda = 100$,
  - $s_{new} = 100 + 1.0\sqrt{100} = 110 < 120$.
  - The staffing needs to be increased less than proportionally.
Let $\lambda \rightarrow \infty$ and $s \rightarrow \infty$, so that $\rho \equiv \lambda/s\mu \rightarrow 1$ and

- $(1 - \rho)\sqrt{s} \rightarrow \beta$, $-\infty < \beta < \infty$ (Quality-and-Efficiency-Driven (QED)).

Let $Q(s)$ and $W(s)$ be steady-state number in system and waiting time.

**HT limit:** As $s \rightarrow \infty$, $\hat{Q}_s \equiv (Q(s) - s)/\sqrt{s} \Rightarrow Q^*$ (Garnett et al. 2002)

and $P(W(s) > 0) = P(Q(s) \geq s) = P(\hat{Q}_s > 0) \rightarrow P(Q^* > 0) \equiv \alpha$,

- where $\alpha \equiv \alpha(\beta, \gamma) \equiv 1/[1 + \gamma h(\beta/\gamma)/h(-\beta)]$, (Garnett function)
- $\gamma \equiv (\theta/\mu)^{1/2}$, $h(x) \equiv \phi(x)/[1 - \Phi(x)]$, $\Phi(x) \equiv P(N(0, 1) \leq x)$ is the standard normal cdf and $\phi(x)$ is the associated pdf

$P(Q^* > x|Q^* > 0)$ and $P(Q^* \leq x|Q^* \leq 0)$ are truncated normal dists.
Staff (choose $s$) to Meet QoS Target

- Decide upon Quality of Service (QoS) target: $P(W > 0) \equiv \alpha^*$. 

- Choose $\beta$ so that $\alpha(\beta, \gamma) \equiv 1/[1 + \gamma h(\beta/\gamma)/h(-\beta)] = \alpha^*$. 

- Given $\beta$, let $s$ be such that $(1 - \rho)\sqrt{s} \equiv (1 - (\lambda/s\mu))\sqrt{s} = \beta$. 

- With that staffing level $s$, we closely approximate our goal:

  **square root staffing formula:**

  $$s = \frac{\lambda}{\mu} + \beta\sqrt{s} \approx \left(\frac{\lambda}{\mu}\right) + \beta \left(\frac{\lambda}{\mu}\right)^{1/2}.$$ 

($\lambda/\mu$ is the offered load, i.e., the mean number of busy servers in the associated infinite-server model. Now we get closer to QoS $\alpha^*$.)
The Garnett Function $\alpha \equiv \alpha(\beta, \gamma)$

Curves with abandonment rate ratio $\theta/\mu$ increasing to the left:
2. Questions for $GI/GI/2/\infty$ Model

- Suppose that you are considering a medical clinic staffed by two 
  doctors that can be modeled as a $GI/GI/2/\infty$ queue.
- Suppose that the parameter are: $(\lambda, c_a^2, \mu, c_s^2) = (4.0, 2.0, 2.5, 0.5)$.
- **Question 1:** A new appointment system is being considered that can 
  reduce the arrival process variability from $c_a^2 = 2.0$ to $c_a^2 = 0.5$.
  
  - By approximately how much will the expected steady-state waiting time 
    (before starting service) $E[W]$ be reduced? By at least 25%?
- **Question 2:** What happens if instead only the arrival rate increases 
  from $\lambda = 4.0$ to $\lambda_{new} = 4.5$?
  
  - Will $E[W]$ increase by a factor of $\lambda_{new}/\lambda = 4.5/4.0 = 1.125$? Or by less? 
  Or by more?
Use Simple Heavy-Traffic Approximation

- \( E[W] \approx \frac{\rho E[S](c_a^2 + c_s^2)}{2s(1-\rho)} \)

**Question 1:** The mean waiting time is directly proportional to \((c_a^2 + c_s^2)\).

- This goes from \((c_a^2 + c_s^2) = 2.0 + 0.5 = 2.5\) to \((c_a^2 + c_s^2) = 0.5 + 0.5 = 1.0\)
- The new value is \(1.0/2.5 = 0.40\) times or 40% of the original value. The reduction is by more than 25%!

**Question 2:** The mean waiting time is sharply increasing in the traffic intensity \(\rho\). The mean \(E[W]\) is proportional to \(\rho/(1-\rho)\).

- Note that \(\rho\) increases from \(\rho = \lambda/s\mu = 4.0/5.0 = 0.8\) to 
  \[\rho = \lambda/s\mu = 4.5/5.0 = 0.9.\]
- The mean \(E[W]\) is now proportional to \(0.9/0.1 = 9\) instead of \(0.8/0.2 = 4.0\). The mean becomes \(9/4 = 2.25\) times larger!
Conventional HT Limit in $GI/GI/s/\infty$ Model

- i.i.d. interarrival times $T_k$: $E[T] \equiv \frac{1}{\lambda}$, $c_a^2 \equiv \frac{\text{Var}(T)}{E[T]^2}$
- i.i.d. service times $S_k$: $E[S] \equiv \frac{1}{\mu}$, $c_s^2 \equiv \frac{\text{Var}(S)}{E[S]^2}$
- Let traffic intensity $\rho \equiv \frac{\lambda}{s\mu} \uparrow 1$ (by multiplying $T_k$ be constants).
- Let $W(\rho)$ be the steady-state waiting time before starting service.
- The distribution of $W(\rho)$ is complicated except for special cases.
- HT limit: $(1 - \rho)W(\rho) \Rightarrow W^*$ (with exponential distribution)
- $E[W(\rho)] \approx \frac{\rho E[S](c_a^2 + c_s^2)}{2s(1-\rho)}$ and $P(W(\rho) > x) \approx e^{-2s(1-\rho)x/(c_a^2 + c_s^2)}$, $x \geq 0$.
- (The mean is exact for $M/M/1/\infty$ and $M/GI/1/\infty$ special cases.)
- Refined approx.: $E[W(\rho)] \approx \left(\frac{c_a^2 + c_s^2}{2}\right) E[W(\rho; M/M/s/\infty)]$ (QNA)
Heavy-Traffic Limit for $M/M/1/\infty$ Queue Length Process

Sketch of the Proof
Construction of $M/M/1/\infty$ Queue Length Process

- **arrival process** $A(t)$: Poisson process with arrival rate $\lambda$
- **potential service process** $S(t)$: Poisson process with rate $\mu$
- **net input process** $X(t) \equiv A(t) - S(t)$, $t \geq 0$
- **queue length process** (number in system)

\[
Q(t) \equiv X(t) - \inf_{0 \leq s \leq t} \{X(s)\}, \quad t \geq 0.
\]

- Starting empty $Q(0) \equiv 0$.
- The process $Q$ is a reflection of the process $X$. 


HT Limit for $M/M/1/\infty$ Queue Length Process ($\rho = 1$)

- As $n \to \infty$,
  
- arrival process: $\frac{[A(nt) - \lambda nt]}{\sqrt{n}} \Rightarrow \sqrt{\lambda} B_a(t)$ (BM limit)

- potential service process: $\frac{[S(nt) - \mu nt]}{\sqrt{n}} \Rightarrow \sqrt{\mu} B_s(t)$

- If $\lambda = \mu$ or, equivalently, if $\rho = 1$, then
  
  - net input process: $\frac{X(nt)}{\sqrt{n}} \equiv \frac{A(nt) - S(nt)}{\sqrt{n}}$
    
    $\Rightarrow \sqrt{\lambda} B_a(t) - \sqrt{\lambda} B_s(t) \overset{d}{=} \sqrt{2\lambda} B(t)$. (again BM limit)

  - queue length process: $\frac{Q(nt)}{\sqrt{n}} \Rightarrow Q(t) \equiv X(t) - \inf_{0 \leq s \leq t} \{X(s)\}$

  - where $X(t) \equiv \sqrt{2\lambda} B(t)$.

- $Q(t)$ is reflected Brownian motion (RBM).
HT Limit for $M/M/1/\infty$ Queue Length Process with Drift

- As $n \to \infty$, with $\lambda_n$ function of $n$,

- If $(\lambda_n - \mu)\sqrt{n} \to c$, i.e., if $\rho_n \equiv 1 - (c/\sqrt{n})$, then

arrival process: $[A_n(nt) - \lambda_n nt]/\sqrt{n} \Rightarrow \sqrt{\mu}B_a(t)$ \hspace{1cm} (BM limit)

potential service process: $[S(nt) - \mu nt]/\sqrt{n} \Rightarrow \sqrt{\mu}B_s(t)$

- net input process: $\frac{X_n(nt)}{\sqrt{n}} \equiv \frac{A_n(nt) - S(nt) - (\lambda_n - \mu)nt}{\sqrt{n}}$

  $\Rightarrow \sqrt{\mu}B_a(t) - \sqrt{\mu}B_s(t) - ct \overset{d}{=} \sqrt{2\mu}B(t) - ct$. \hspace{1cm} (BM with drift)

queue length process: $\frac{Q(nt)}{\sqrt{n}} \Rightarrow Q(t) \equiv X(t) - \inf_{0 \leq s \leq t} \{X(s)\}$

- where $X(t) \equiv \sqrt{2\mu}B(t) - ct$.

- $Q(t)$ is reflected Brownian motion (RBM) with drift.

- The steady-state distribution of RBM with drift is exponential!
HT Limit for $GI/GI/1/\infty$ Queue Length Process with Drift

- As $n \to \infty$, variation of same reasoning applies:
  - If $(\lambda_n - \mu)\sqrt{n} \to c$, then
  - arrival process: $[A_n(nt) - \lambda_n nt]/\sqrt{n} \Rightarrow \sqrt{\mu c_a^2} B_a(t)$ (BM limit)
  - potential service process: $[S(nt) - \mu nt]/\sqrt{n} \Rightarrow \sqrt{\mu c_s^2} B_s(t)$
  - net input process: $\frac{X_n(nt)}{\sqrt{n}} \equiv \frac{A_n(nt) - S(nt) - (\lambda_n - \mu)nt}{\sqrt{n}}$
    $\Rightarrow \sqrt{\mu c_a^2} B_a(t) - \sqrt{\mu c_s^2} B_s(t) - ct \overset{d}{=} \sqrt{\mu (c_a^2 + c_s^2)} B(t) - ct$.
  - queue length process: $\frac{Q(nt)}{\sqrt{n}} \Rightarrow Q(t) \equiv X(t) - \inf_{0 \leq s \leq t} \{X(s)\}$
  - where $X(t) \equiv \sqrt{\mu (c_a^2 + c_s^2)} B(t) - ct$.
  - $Q(t)$ is reflected Brownian motion (RBM) with drift.
  - The steady-state distribution is again exponential, but with $(c_a^2 + c_s^2)$!
Many-Server Heavy-Traffic Limits and Approximations


Conventional Heavy-Traffic Limits and Approximations

  
  http://www.columbia.edu/~ww2040/book.html (See Chapters 1, 2, 5 and 9 plus §7.3.