- Given the source statistics, we can determine the required link rate C such that $Pr(\sum_{i=1}^{m} B^{(i)} > C) < \epsilon$, where $B^{(i)}$ are the independent marginal r.v.s of the sources $i = 1, 2, .., m$, & $0 < \epsilon < 1$
  - Assume the sources are identically distributed with mean $E(B)$ & max value $R \geq B^{(j)}$
  - For a single source ($m=1$) & $\epsilon < Pr(B^{(1)} = R)$, we need $C = R$
  - For larger m, let $C = m\gamma$ (γ is the link capacity per source); we need $E(B) < \gamma \leq R$
  - For very large m, γ can be very close to but strictly larger than $E(B)$
- Statistical multiplexing gain: as the no. of sources ↑s, the bandwidth requirement per source ↓s for the same QoS objective
  - gain obtained at the cost of a nondeterministic QoS

• Analysis using Chernoff's Bound

- consider iid r.v.s $B^{(1)}, \ldots, B^{(m)}$
- want to bound $\Pr\left(\sum_{i=1}^{m} B^{(i)} \geq m\gamma\right) = \Pr\left(\sum_{i=1}^{m} (B^{(i)} - \gamma) \geq 0\right)$

  for $\gamma > E(B)$

- **Chernoff's Bound: For any r.v. Y,**

  $$\Pr(Y \geq a) \leq e^{-\theta a} E(e^{\theta Y}) \quad \forall\ \theta \geq 0$$

- Applying Chernoff's Bound to r.v. $\sum_{i=1}^{m} (B^{(i)} - \gamma)$ with $a = 0$, we have

$$\Pr\left(\sum_{i=1}^{m} (B^{(i)} - \gamma) \geq 0\right) \leq E\left(e^{\theta\left(\sum_{i=1}^{m}(B^{(i)} - \gamma)\right)}\right) \quad \forall\ \theta \geq 0$$

$$= \left(e^{-\theta\gamma} E(e^{\theta B^{(i)}})\right)^m \quad \text{(since iid)}$$

$$= e^{-m(\theta\gamma - \ln M(\theta))}$$

where $M(\theta) \triangleq E(e^{\theta B^{(i)}})$

$$\leq \inf_{\theta \geq 0} e^{-m(\theta\gamma - \ln M(\theta))}$$

$$= e^{-m \underbrace{\sup_{\theta \geq 0} (\theta\gamma - \ln M(\theta))}_{\triangleq\ I(\gamma)}}$$

- Application to admission control problem — for a given type of source B & a given probability e^{-s} of capacity overflow, up to N sources can be handled by a link of capacity C, where

$$N = \left\lfloor \frac{C}{\gamma^*} \right\rfloor, \quad s = \frac{C}{\gamma^*} I(\gamma^*), \quad E(B) < \gamma^* \leq \max(B)$$

→ Probabilistic QoS

- end-to-end packet delay below some upper bound, and the fraction of lost or late packets below some small value
- call-blocking probability

- **Arbitrary buffering**
  - having more than marginal buffering allows more efficient use of link bandwidth when sources are bursty / time-varying
  - data that the link capacity cannot accommodate in a slot is queued for transmission in later slots
  - if the total arrival rate exceeds the link transmission rate, the queue length increases; if subsequently the arrival rate drops below the link rate, the queue length decreases
  - the average queue length, average arrival rate & average delay are related by Little's Theorem

Queueing models & Little's Theorem

- Customers arrive at random times to obtain service, eg. packets arriving at a communication link
- Suppose we observe a sample path of the system for time $t \geq 0$
- $N(t)$ = no. of customers in system at time t

  $\alpha(t)$ = no. of customers who arrived in the interval $[0, t]$

  T_i = time spent in system by ith arriving customer
- time average no. of customers in system up to time t = $N_t = \frac{1}{t}\int_0^t N(\tau)\, d\tau$

  time average arrival rate over $[0, t]$ = $\lambda_t = \frac{\alpha(t)}{t}$

  time average customer delay up to t = $T_t = \frac{\sum_{i=0}^{\alpha(t)} T_i}{\alpha(t)}$
- steady state time averages (assuming limits exist): $N = \lim_{t\to\infty} N_t$

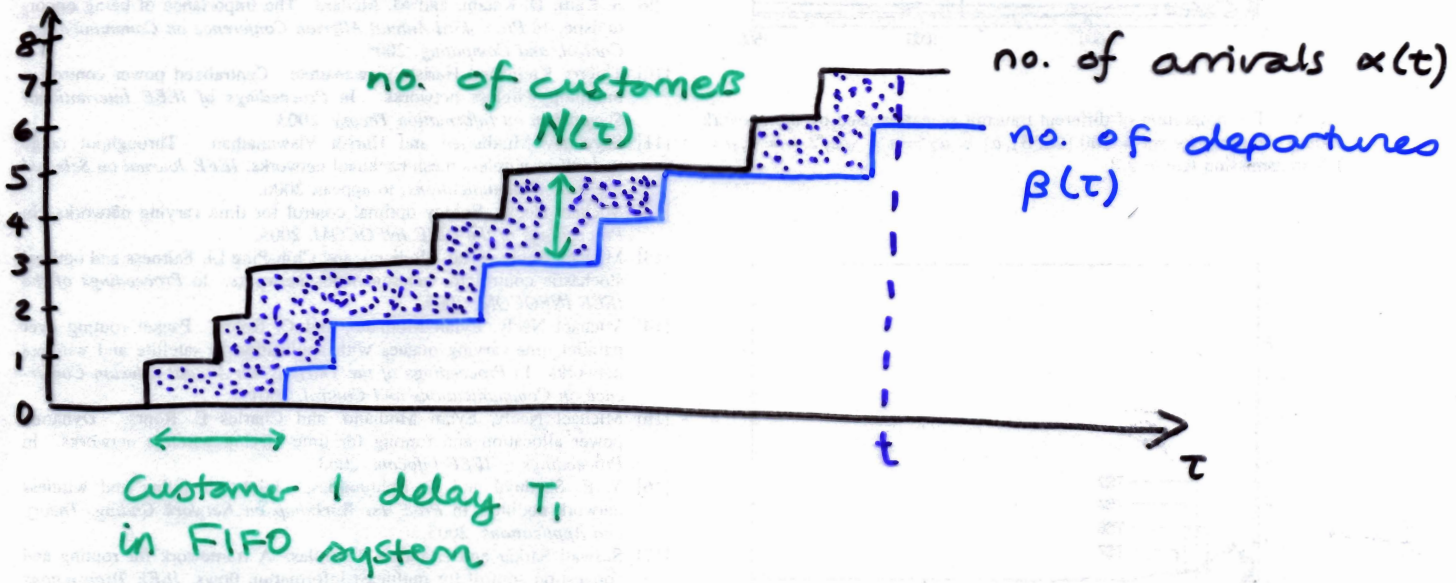
  $\lambda = \lim_{t\to\infty} \lambda_t$

  $T = \lim_{t\to\infty} T_t$

Little's Thm: $N = \lambda T$

- the higher the arrival rate (λ) & the longer the average wait (T), the larger the average queue length (N)

Graphical interpretation for $N(0) = 0$, FCFS

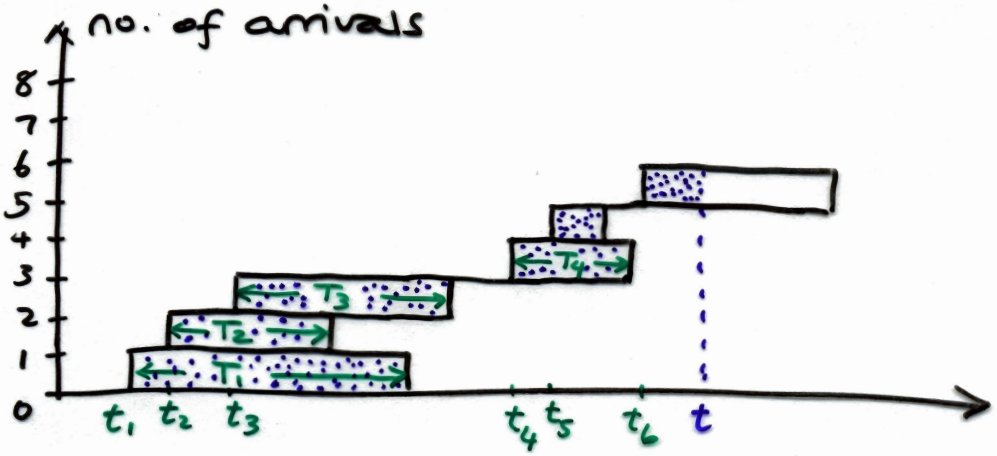


- shaded area between $\alpha(\tau)$ & $\beta(\tau)$ graphs up to $\tau = t$ is equal to $\int_0^t N(\tau)\, d\tau$ & lies between $\sum_{i=1}^{\beta(t)} T_i$ & $\sum_{i=1}^{\alpha(t)} T_i$
- dividing by t & assuming limits $N_t \to N$, $\frac{\alpha(t)}{t} = \lambda_t \to \lambda$, $\frac{\beta(t)}{t} \to \lambda$, $T_t \to T$, as $t \to \infty$:

$$\frac{\beta(t)}{t} \frac{\sum_{i=1}^{\beta(t)} T_i}{\beta(t)} \leq \frac{\int_0^t N(\tau)\, d\tau}{t} \leq \frac{\alpha(t)}{t} \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

$$\lambda T \leq N \leq \lambda T$$

Graphical interpretation for $N(0) = 0$, non-FCFS



- shaded area is equal to $\int_0^t N(\tau)\,d\tau$ as before, & also equal to $\sum_{\text{departed customers } i} T_i + \sum_{\text{remaining customers } i} t - t_i$
- dividing by t & taking $t \to \infty$ gives $\lambda T = N$

Probabilistic version

- in ergodic systems, the time average of a sample path is equal to the steady state statistical / ensemble average

$$N = \lim_{t\to\infty} N_t = \lim_{t\to\infty} \bar{N}(t) = \bar{N}$$

$$T = \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} T_i = \lim_{k\to\infty} \bar{T}_k = \bar{T}$$

where $\bar{N}(t) = \sum_{n=0}^{\infty} n\, p_n(t)$

$p_n(t)$ = prob of n customers in system at t

$\bar{T}_k$ = expected delay of k^{th} customer

- $N = \lambda T$ where $\lambda = \lim_{t\to\infty} \dfrac{\text{expected no. of arrivals in } [0,t]}{t}$