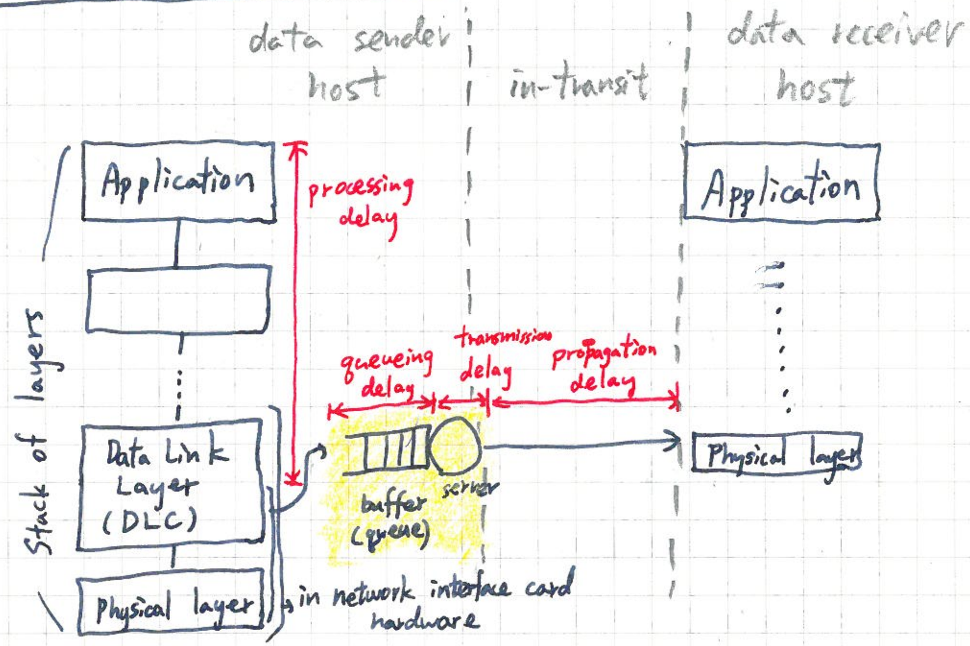
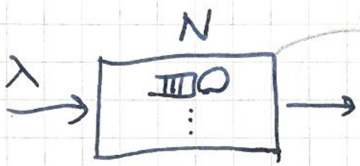
Note for lecture 06



Little's Theorem and queueing systems



→ a system can be one or multiple queues plus other elements; see the following examples.

$$N = \lambda \cdot T$$

N: The average number of customers (data packets) in the system.

λ: The customer arrival rate

T: The average delay per customer (time spent in the system)

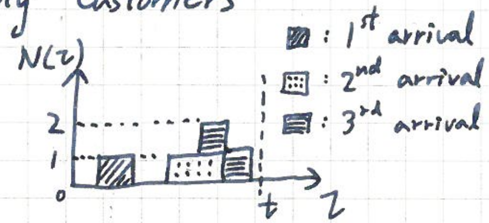
Deriving Little's Theorem:

Let $N(t)$ = # of customers in the system at time t

$\alpha(t)$ = # of customers arrived in interval $[0, t]$

T_i = Time spent in the system by the i-th arriving customers

$$N_t = \frac{1}{t}\int_0^t N(\tau)\cdot d\tau$$

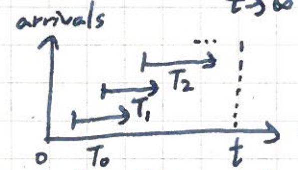


at steady state,

N_t converges to $N = \lim_{t\to\infty} N_t$

$$\lambda_t = \frac{\alpha(t)}{t}$$ at steady state, $\lambda = \lim_{t\to\infty} \lambda_t$

$$T_t = \frac{1}{\alpha(t)}\sum_{i=0}^{\alpha(t)} T_i$$, $T = \lim_{t\to\infty} T_t$



let $\beta(t)$ = # of customers departed in $[0,t]$

$\Rightarrow N(\tau) = \alpha(\tau) - \beta(\tau)$

The area between curves $\alpha(\tau)$ and $\beta(\tau)$ is $\int_0^t N(\tau)\,d\tau$.

The area is also equal to $\sum_{i=0}^{\alpha(t)} T_i$ if $N(t)=0$ if $N(t)>0$

$$\sum_{i=0}^{\beta(t)} T_i \le \int_0^t N(\tau)\,d\tau \le \sum_{i=0}^{\alpha(t)} T_i$$

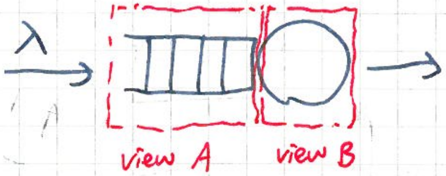
assuming $\underbrace{\frac{\beta(t)}{t}}_{=\lambda} = \underbrace{\frac{\alpha(t)}{t}}_{=\lambda}$ when $t\to\infty$

$$\Rightarrow \underbrace{\frac{1}{t}\cdot\int_0^t N(\tau)\,d\tau}_{N} = \underbrace{\frac{\alpha(t)}{t}}_{\lambda}\cdot\underbrace{\frac{\sum_{i=0}^{\alpha(t)} T_i}{\alpha(t)}}_{T}$$

at steady state

$$N = \lambda \cdot T$$

Example 1 (Exp 3.1 in the textbook)



- View A, applying Little's Theorem

$$N_Q = \lambda \cdot W$$

N_Q : the # of packets waiting in the queue

W : time spent by a packet waiting in queue

- View B, applying Little's Theorem

assuming the arrival rate at server equals the departure rate at queue

$$\rho = \lambda \cdot \overline{X}$$ (compare to the notion of link utilization in lecture 01)

ρ : the # of packets in service

$\overline{X}$: the transmission time

by definition, $\rho \leq 1$

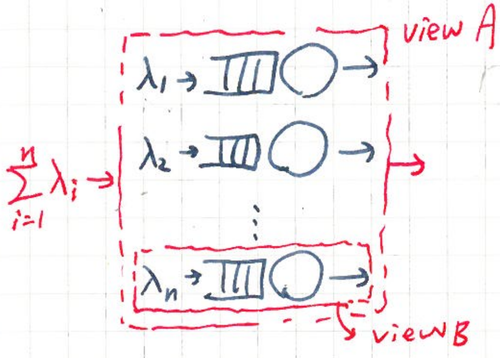
⇒ given $\overline{X}$, then λ' is at most $\frac{1}{\overline{X}}$

which implies accumulation in queue if $\lambda > \lambda' = \frac{1}{\overline{X}}$

⇒ given λ, then we know to what degree we should improve $\overline{X}$ to prevent buffer overflow.

e.g., get a faster NIC

Example 2 (Exp 3.2 in the textbook)



View A, applying Little's Theorem

$$N = \sum_{i=1}^{n} \lambda_i \cdot T$$

$$\Rightarrow T = \frac{N}{\sum_{i=1}^{n} \lambda_i}$$

View B, applying Little's Theorem to each queueing system

$$\sum_{i=1}^{n} N_i = \sum_{i=1}^{n} \lambda_i \cdot T_i$$

$$N = \sum_{i=1}^{n} \lambda_i \cdot T_i$$

compare with View A, one configuration to make

$\sum_{i=1}^{n} \lambda_i \cdot T_i = \sum_{i=1}^{n} \lambda_i \cdot T$ is that $T_i = T$ for all i.

Example 3 (Exp 3.4 in the textbook)

in window flow control (e.g., in go-back-N ARQ)

Little's Theorem tells us that $W \geq \lambda T$

where W is the window size.

⇒ if congestion occurs (i.e, $T \uparrow$)

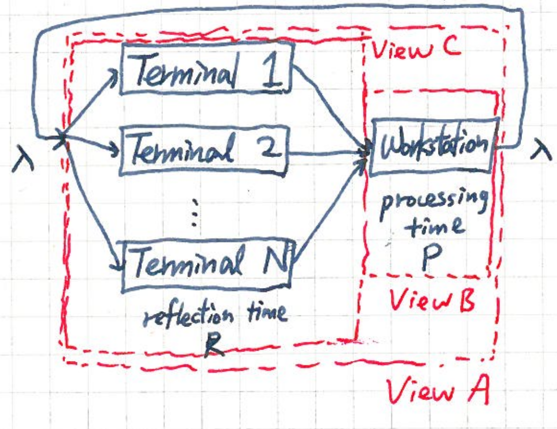
then the control will slow down accepting packets from upper layer. (i.e., $\lambda \downarrow$)

⇒ if the transmission line has 100% link utilization

then $W = \lambda T$ λ is fixed

suggesting that increasing W would only increase T!

Example 4 (Exp 3.7) Bounding the attainable system throughput λ



View A

$\Rightarrow \lambda = \frac{N}{T}$, $T = R + D$

where

$P \le D \le NP$

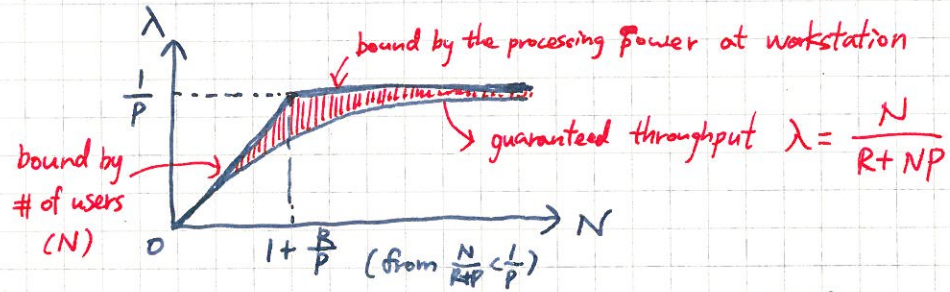
$\Rightarrow \frac{N}{R+NP} \le \lambda \le \frac{N}{R+P}$

View B $\Rightarrow 1 \ge \lambda P \Rightarrow \lambda \le \frac{1}{P}$

(View C $\Rightarrow N' = \lambda R \Rightarrow N' \le \frac{R}{P}$)

From Views A and B, we have

$$\frac{N}{R+NP} \le \lambda \le \min\left\{\frac{1}{P}, \frac{N}{R+P}\right\}$$



Further, by $T = \frac{N}{\lambda}$

$\Rightarrow \max\{NP, R+P\} \le T \le R + NP$