Queueing disciplines

- Nonpreemptive
  - ✓ First come first served (FCFS)
  - ✓ Head-of-the-line (HOL) priority
  - ✓ Shortest Processing Time (SPT) first
- Round-robin (RR) and Processor-sharing (PS)

M/G/1 queue (FCFS discipline)

- Poisson arrivals at λ customers per second
- Service times with a general probability distribution

  mean value $\overline{x}$ and second moment $\overline{x^2}$
- Define $\rho = \lambda \overline{x}$

- Another derivation of P-K formula using mean residual life and Little's Law

M/G/1 queue (FCFS)

$$W = E[\text{ wait } | \text{ customer finds empty system }] \cdot Prob[\text{ customer finds empty system }] + \sum_{n=1}^{\infty} E[\text{ wait } | \text{ customer finds } n \text{ in the system }] \cdot P_n$$

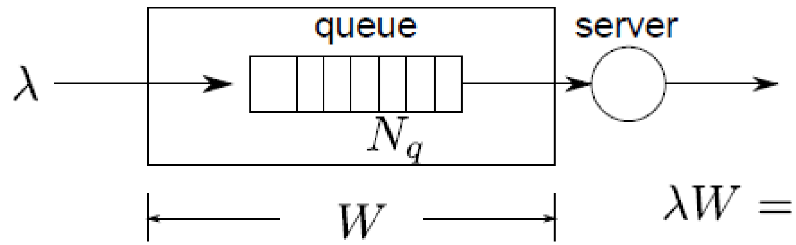
$$= 0 \cdot (1-\rho) + \sum_{n=1}^{\infty} P_n[(n-1)\bar{x} + \frac{\overline{x^2}}{2\bar{x}}]$$

$$= Prob[\text{ busy system }]\frac{\overline{x^2}}{2\bar{x}} + \bar{x}\overbrace{\sum_{n=1}^{\infty}(n-1)\cdot P_n}^{\text{mean \# in the queue}}$$

$$= \boxed{\rho\frac{\overline{x^2}}{2\bar{x}} + \bar{x}N_q}$$

Now apply Little's law

M/G/1 queue (FCFS)



$\lambda W = N_q$

$$\boxed{W = \rho\frac{\overline{x^2}}{2\bar{x}} + \bar{x}\lambda W} = \rho\frac{\overline{x^2}}{2\bar{x}} + \rho W$$

$$W = \frac{\rho\overline{x^2}}{2\bar{x}(1-\rho)} = \frac{\lambda\bar{x}\overline{x^2}}{2\bar{x}(1-\rho)} = \frac{\lambda\overline{x^2}}{2(1-\rho)}$$

Pollaczek-Khinchin (P-K) mean value formula

mean delay $T = W + \bar{x} = \frac{\lambda\overline{x^2}}{2(1-\rho)} + \bar{x}$

M/G/1 Head-of-the-Line (HOL) nonpreemptive

- R classes
  - class 1: highest priority
  - class R: lowest priority
- Poisson arrivals at rate λ_r customers/sec, for $r = 1, 2, \ldots, R$
- mean service time $\overline{x_r}$

  $\rho_r = \lambda_r \overline{x_r}$

  second moment of service time $\overline{x_r^2}$

M/G/1 (HOL) nonpreemptive

$\boxed{\text{Case 1 } \rho = \sum_{r=1}^{R} \rho_r < 1}$

- mean waiting time of class r customers

$$\boxed{W_r = U_s + \sum_{k=1}^{r} N_{q,k}\overline{x_k} + \sum_{k=1}^{r-1} \lambda_k W_r \overline{x_k}, \;\; r > 1}$$

$\boxed{U_s}$: mean time to finish customer being served at arrival instant

$$U_s = \boxed{\sum_{r=1}^{R} \rho_r \frac{\overline{x_r^2}}{2\overline{x_r}}} = \sum_{r=1}^{R} \frac{\lambda_r \overline{x_r^2}}{2} = \frac{\lambda}{2} \sum_{r=1}^{R} \frac{\lambda_r}{\lambda} \overline{x_r^2} = \boxed{\frac{\lambda}{2} \overline{x^2}},$$

$$\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_R$$

M/G/1 (HOL) nonpreemptive

apply Little's law

$$\boxed{\begin{aligned} W_1 &= U_s + \lambda_1 W_1 \overline{x_1} \\ W_r &= U_s + \sum_{k=1}^{r} \lambda_k W_k \overline{x_k} + W_r \sum_{k=1}^{r-1} \rho_k,\ r = 2, 3, \ldots, R \end{aligned}}$$

Define $\boxed{\sigma_r = \sum_{k=1}^{r} \rho_k}$,

$$\boxed{\begin{aligned} W_1 &= \frac{U_s}{1 - \rho_1} \\ W_r &= \frac{U_s + \sum_{k=1}^{r} \rho_k W_k}{1 - \sigma_{r-1}},\ r = 2, 3, \ldots, R \end{aligned}}$$

Solving recursively, we get

$$W_r = \frac{U_s}{(1 - \sigma_r)(1 - \sigma_{r-1})},\ r = 2, 3, \ldots, R$$

M/G/1 (HOL) nonpreemptive

Case 2 $\sigma_q < 1$, but $\sigma_{q+1} \geq 1$

- $W_r = \infty$ for $r \geq q+1$
- $W_r < \infty$ for $r = 1, 2, \ldots, q$

$$W_1 = \frac{U'_s}{1-\rho_1}$$

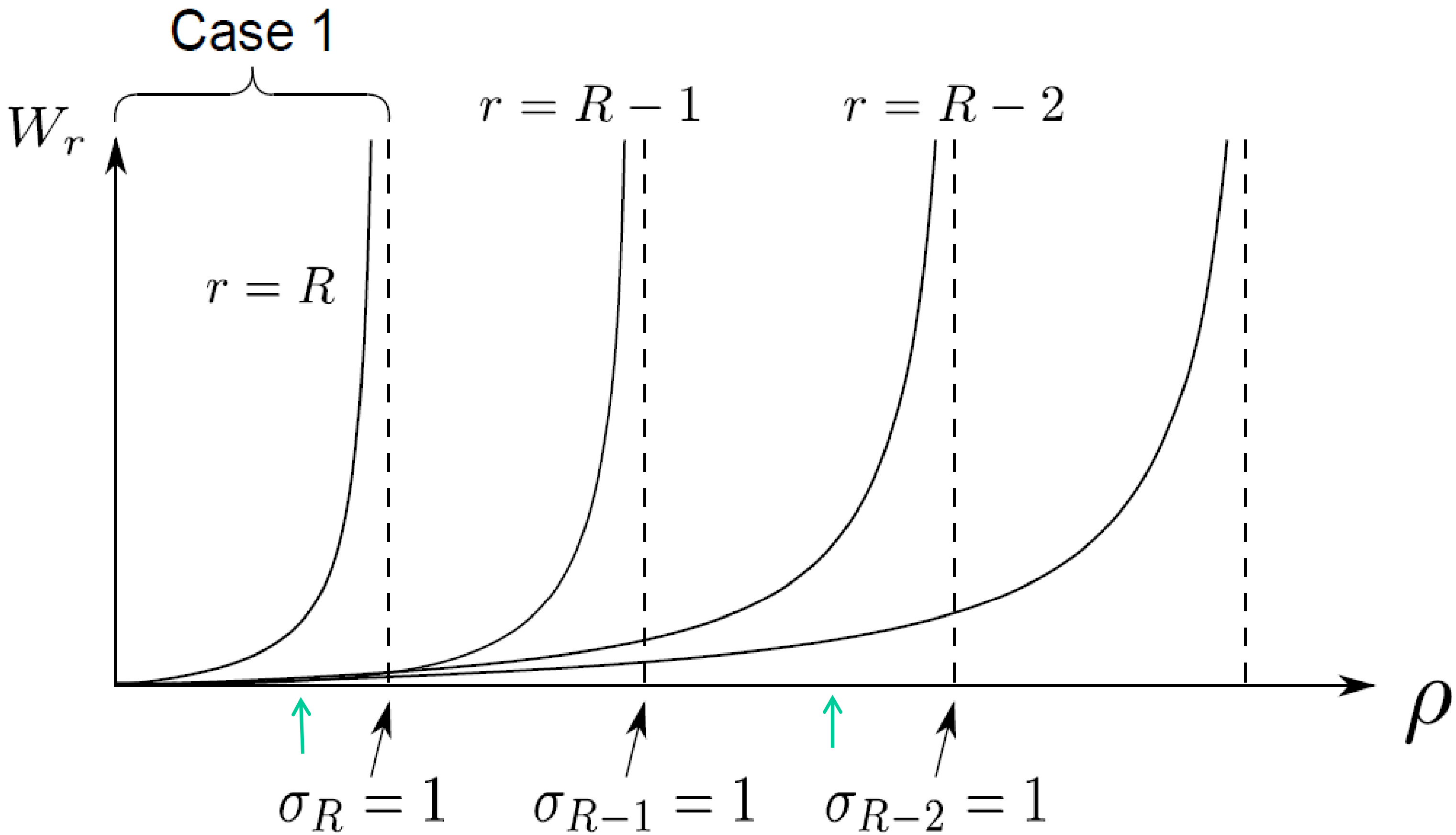
$$W_r = \frac{U'_s}{(1-\sigma_r)(1-\sigma_{r-1})}, r = 2, 3, \ldots, q$$

where

$$\boxed{U'_s = \sum_{r=1}^{q} \rho_r \frac{\overline{x_r^2}}{2\overline{x_r}} + (1-\sigma_q)\frac{\overline{x_{q+1}^2}}{2\overline{x_{q+1}}}}$$

M/G/1 (HOL) nonpreemptive

mean delay $T_r = W_r + \overline{x_r}$



M/G/1 Shortest processing time first (SPT) nonpreemptive

- Application of HOL nonpreemptive
- service time distribution function

$$B(x) = Prob[\text{service time} \leq x]$$

service time density function

$$b(x) = \frac{\mathrm{d}B(x)}{\mathrm{d}x}$$

M/G/1 SPT nonpreemptive

- Assume $\lambda\bar{x} < 1$ (analogous to case 1 of discrete case)

$$\boxed{W_{SPT}(x) = E[\text{wait}|\text{service time } = x]}$$

Recall: $\boxed{W_r = \dfrac{\lambda\overline{x^2}}{2(1-\sigma_r)(1-\sigma_{r-1})}}$

$$\boxed{W_{SPT}(x) = \begin{cases} \dfrac{\lambda\overline{x^2}}{2(1-\lambda\int_0^{x+} y\,\mathrm{d}B(y))(1-\lambda\int_0^{x-} y\,\mathrm{d}B(y))} \\ \dfrac{\lambda\overline{x^2}}{2(1-\lambda\int_0^{x} y\,\mathrm{d}B(y))^2} & \text{if } B(y) \text{ is continuous at } y = x \end{cases}}$$

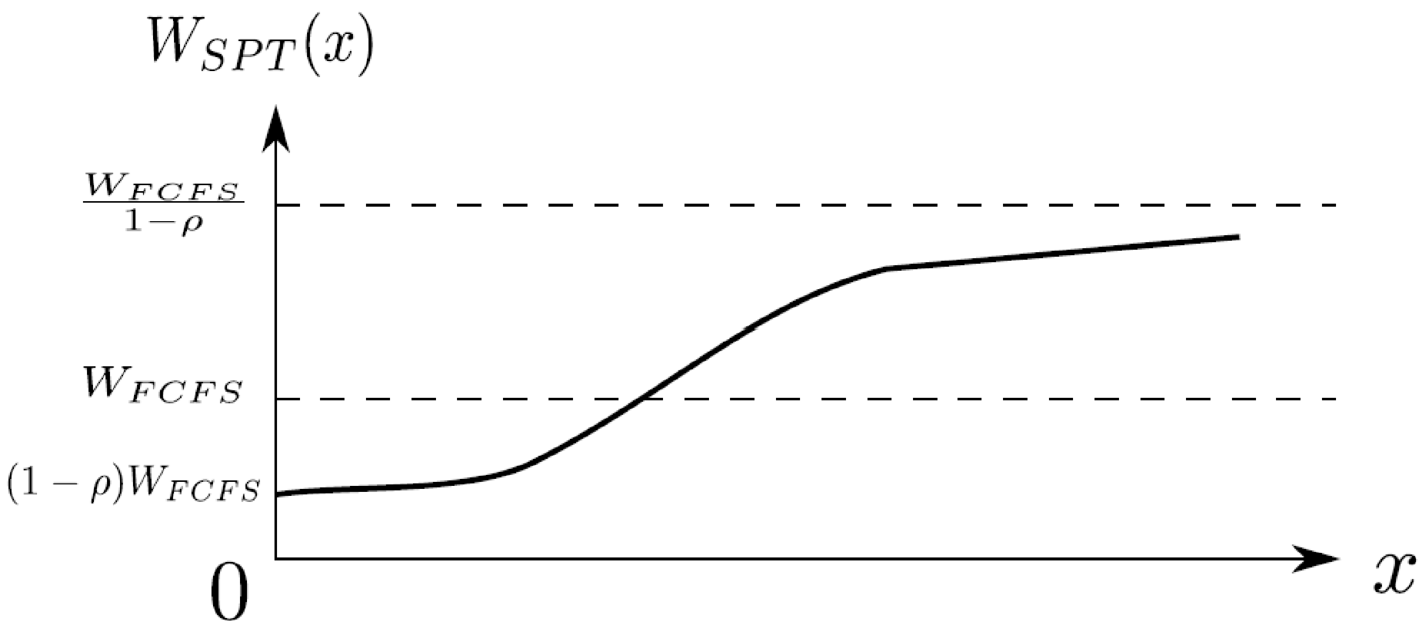
- Mean response time $T_{SPT}(x) = W_{SPT}(x) + x$

M/G/1 SPT nonpreemptive

- Assume that $B(y)$ is continuous in y

$$\boxed{\lim_{x\to\infty} W_{SPT}(x) = \frac{\lambda \overline{x^2}}{2(1-\rho)^2} = \frac{W_{FCFS}}{1-\rho}}$$

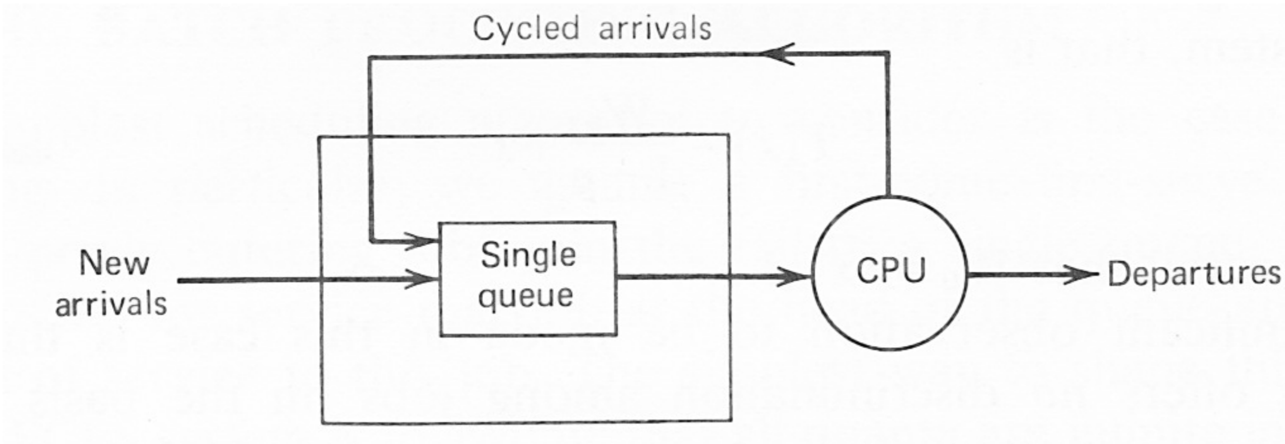
$$\boxed{\lim_{x\to 0} W_{SPT}(x) = \frac{\lambda \overline{x^2}}{2} = (1-\rho)W_{FCFS}}$$



Queueing disciplines

- **Nonpreemptive**
  - First come first served (FCFS)
  - Head-of-the-line (HOL) priority
  - Shortest Processing Time (SPT) first
- **Round-robin (RR) and Processor-sharing (PS)**

Round-robin (RR) scheduling



- The job first in queue gets a quantum *q* of service. Then if it needs more service, it is returned to the end of the queue.
  - *Good for CPU scheduling because job size is unknown a priori.*
- In packet switching, a packet's size is known
  - But size of application data unit may not be known

Processor-Sharing (PS) discipline for M/G/1

- What is the average delay and wait of a job with service time x in the limit as $q \rightarrow 0$?
- T(x) = ave. delay of a job with service time x

$$= \frac{x}{1-\rho}$$

- W(x) = ave. wait of a job with service time x

$$= \frac{\rho x}{1-\rho}$$

[From Kleinrock, Vol. 2, page 168]

Packet scheduling in networks

- What if the app. user is willing to pay more money for priority service?
    - Network neutrality advocates do not like this
- RR and PS scheduling – Are they more fair?
    - How to implement PS scheduling?
- A packet can be thought of as a quantum in RR for an application data unit.
    - Delay of an application data unit is more important than packet delay.
- We will return to these issues when we study deterministic delay guarantees for a packet-switching network.

The end