D.5 Priority Systems

- Jobs have priorities r (r = 1, 2,.... , R)
- 1 is the lowest priority und R the highest !
- The next job to be served is the job with the highest priority number
- Inside a priority class the queueing discipline is FCFS

1 Priority Systems without Preemption

◆ The mean waiting time of a job of priority class r W_r has three components:

➤ The mean remaining service time W_0 of the job in service

➤ Mean service time of jobs in the queue that are served before the tagged job. These are the jobs in the queue of the same and higher priority as the tagged job.

➤ Mean service time of jobs that arrive at the system while the tagged job is in the queue and are served before it.These are jobs with higher priority than the tagged job.

- Definition:
  - ➤ $\overline{N}_{ir}$: Mean number of jobs of class i found in the queue by the tagged (priority r) job and receiving service before it,
  - ➤ $\overline{M}_{ir}$: Mean number of jobs of class i who arrive during the waiting time of the tagged job and receive service before it

- The mean waiting time of class r jobs can be written as the sum of three components:

$$\overline{W}_r = \overline{W}_0 + \sum_{i=1}^{R} \overline{N}_{ir} \cdot \frac{1}{\mu_i} + \sum_{i=1}^{R} \overline{M}_{ir} \cdot \frac{1}{\mu_i}$$

Multiple server systems (m > 1):

$$\overline{W}_r = \overline{W}_0 + \sum_{i=1}^{R} \frac{\overline{N}_{ir}}{m} \frac{1}{\mu_i} + \sum_{i=1}^{R} \frac{\overline{M}_{ir}}{m} \cdot \frac{1}{\mu_i}$$

- $\overline{N}_{ir}$ and $\overline{M}_{ir}$:

$$\overline{N}_{ir} = 0 \quad i < r$$

$$\overline{M}_{ir} = 0 \quad i \leq r$$

With Little's law:

$$\overline{N}_{ir} = \lambda_i \overline{W}_i \quad i \geq r$$

Mean number of arriving jobs of class i $\overline{M}_{ir}$ during the mean waiting time $\overline{W}_r$:

$$\overline{M}_{ir} = \lambda_i \overline{W}_r \quad i > r.$$

- Mean waiting time of a job of priority r:

$$\overline{W}_r = \frac{\overline{W}_0}{(1-\sigma_r)(1-\sigma_{r+1})}$$

where:

$$\sigma_r = \sum_{i=r}^{R} \rho_i$$

- Mean overall waiting time:

$$\overline{W} = \sum_{i=1}^{R} \frac{\lambda_i}{\lambda} \cdot \overline{W}_i$$

- Mean remaining service time:
  - ➤ M/M/1:

$$\overline{W}_{0,\mathrm{M/M/1}} = \sum_{i=1}^{R} \rho_i \frac{1}{\mu_i}$$

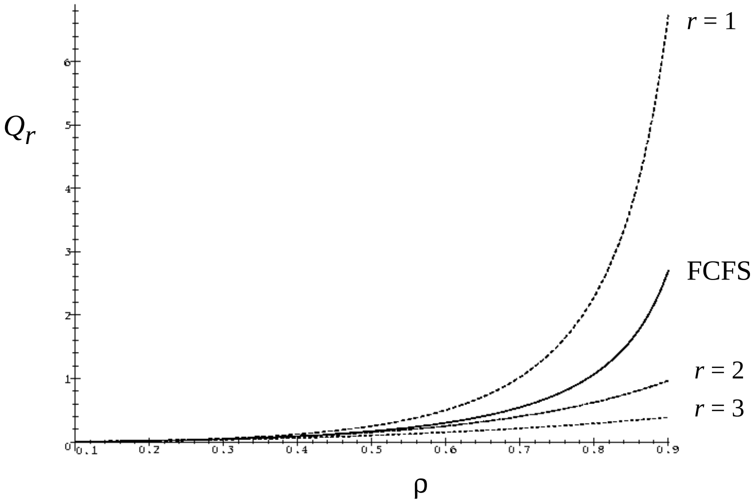
  - ➤ M/G/1:

$$\overline{W}_{0,\mathrm{M/G/1}} = \sum_{i=1}^{R} \rho_i \cdot \frac{1+c_{B_i}^2}{2\mu_i}$$

  - ➤ M/M/m:

$$\overline{W}_{0,\mathrm{M/M/m}} = \frac{P_m}{m\rho} \sum_{i=1}^{R} \rho_i \cdot \frac{1}{\mu_i}$$

Mean queue length Q_r of a M/M/1 priority system without preemption:



2 Conservation Law

In priority systems have jobs with higher priority a shorter mean queue length than jobs with lower priority. There exists a conservation law:

$$\frac{1}{\rho}\sum_{i=1}^{R}\rho_i\overline{W}_i = \frac{\overline{W}_0}{1-\rho} = \overline{W}_{\mathrm{FCFS}}$$

- The conservation law to apply the following restrictions must be satisfied:
  - No service facility is idle as long as there are jobs in the queue.
  - No job leaves the system before its service is completed.
  - The distributions of the interarrival times and the service times are arbitrary with the restriction that the first moments of both the distributionsand the second moment of the service time distribution exist.
  - The service times of the jobs are independent of the queueing discipline.
  - Preemption is allowed only when all jobs have the same exponential service time distribution and preemption is of the type preemptive resume.
  - For GI/G/m systems all classes have the same service times. This restriction is not necessary for GI/G/1 systems.

3 Priority Systems with Preemption

- Preemptive resume
- Preemption needs no time..
- Job of priority r is not influenced by jobs of priorities 1, 2, 3, ../.
- To determine the mean waiting time $\overline{W}_r$, only the piorities $r, r+1, \ldots, R$ have to be considered
- We replace:

$$\rho = \sum_{i=1}^{R} \rho_i \qquad \sigma_r = \sum_{i=r}^{R} \rho_i$$

and

$$\overline{W}_{\mathrm{FCFS}} \qquad \overline{W}^r = \frac{\overline{W}_0^r}{1-\sigma_r}$$

- Mean remaining service time:

$$\overline{W}_0^r = \frac{P_m^r}{2m\sigma_r}\cdot\sum_{i=r}^{R}\rho_i\frac{1+c_B^2}{\mu_i}$$

- Application of the conservation law to get the mean waiting time $\overline{W}_r$ of a job of priority r :

$$\sigma_r\cdot\overline{W}^r = \sum_{i=r}^{R}\rho_i\overline{W}_i\,,$$

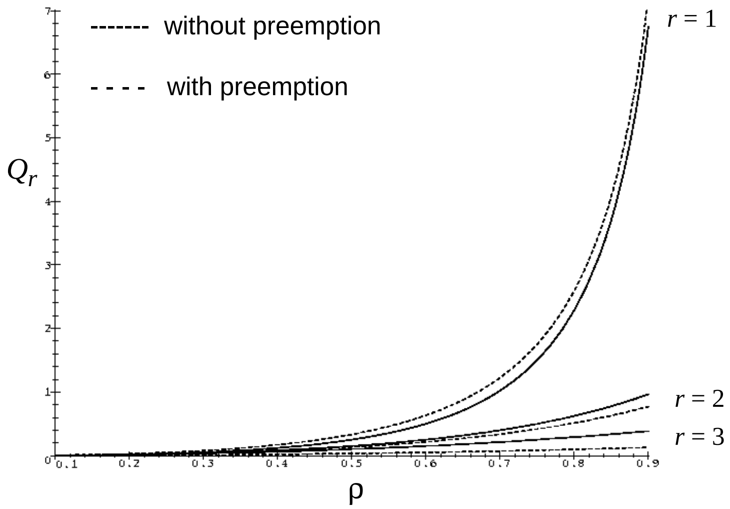
$$\sigma_{r+1}\cdot\overline{W}^{r+1} = \sum_{i=r+1}^{R}\rho_i\cdot\overline{W}_i$$

- Mean waiting time $\overline{W}_r$ of a job of priority r:

$$\overline{W}_r = \frac{1}{\rho_r}\left(\sigma_r\overline{W}^r - \sigma_{r+1}\overline{W}^{r+1}\right)$$

- Exact results:
  - M/M/1
  - M/G/1
  - M/M/m
- For other systems good approximation

Mean queue length Q_r of a M/M/1 priority system with and without preemption



4 Time Dependent Priorities

- In many systems the priorities are time dependent to prefer jobs with a long waiting time:
  - ➤ Real time systems
  - ➤ Mobil networks
  - ➤ Internet

- Priority function:

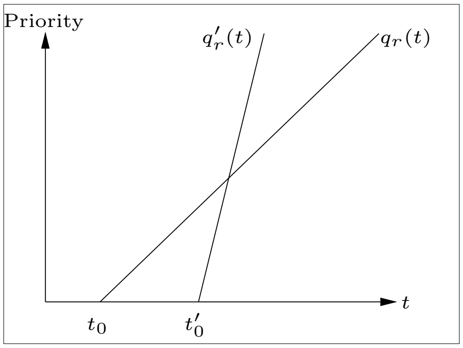
$$q_r(t) = \text{Priority of class } r \text{ at time } t$$

- Priority function with time dependent slope:

$$q_r(t) = (t - t_0) \cdot b_r$$

with:

$$0 \le b_1 \le b_2 \le \ldots \le b_r$$

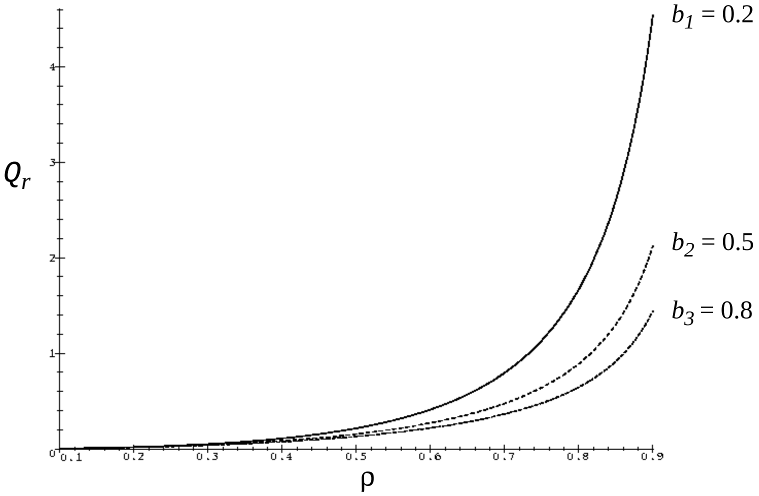


➤ Mean waiting time of a job with priority *r* :

$$\overline{W}_r = \frac{\dfrac{\overline{W}_0}{1-\rho} - \sum_{i=1}^{r-1} \rho_i \overline{W}_i \left(1 - \dfrac{b_i}{b_r}\right)}{1 - \sum_{i=r+1}^{R} \rho_i \left(1 - \dfrac{b_r}{b_i}\right)}$$

Arbitrary G/G/m-system

Mean queue length Q_r for an M/M/1 priority system with time dependent priorities and slope b_r

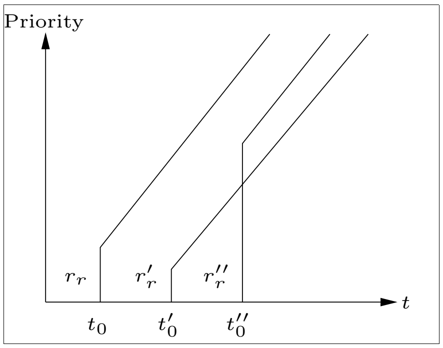


- Priority function with starting priority r_r:

$$q_r(t) = r_r + t - t_0$$

mit:

$$0 \leq r_1 \leq r_2 \leq \ldots \leq r_r$$



- Mean waiting time of a job with priority r :

  - Heavy traffic approximation (ρ --> 1):

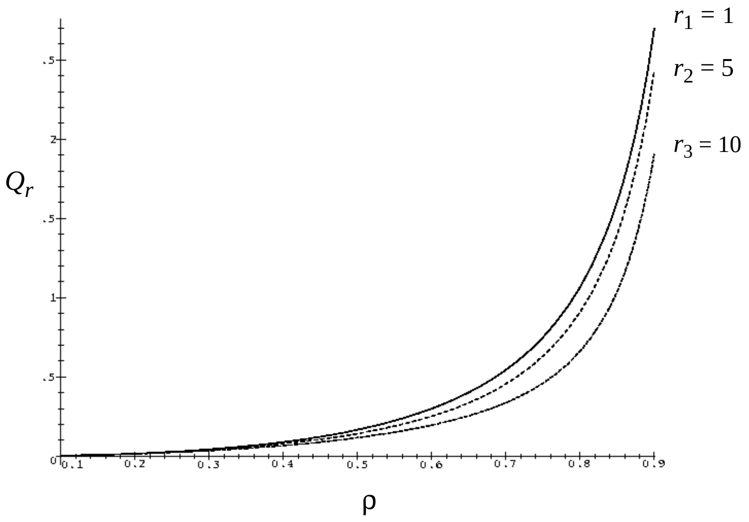
$$\overline{W}_r \approx \frac{\overline{W}_0}{1-\rho} - P_m \cdot \sum_{i=1}^{R} \rho_i (r_r - r_i)$$

  - More accurate approximation (0 < ρ < 1):

$$\overline{W}_r \approx \frac{\overline{W}_0}{1-\rho} - \sum_{i=1}^{r-1} \rho_i \overline{W}_i \left(1 - \exp\left(\frac{P_m(r_i - r_r)}{\overline{W}_i}\right)\right)$$

  - Arbitrary G/G/m-systems ($m = 1$ --> $P_m = \rho$)

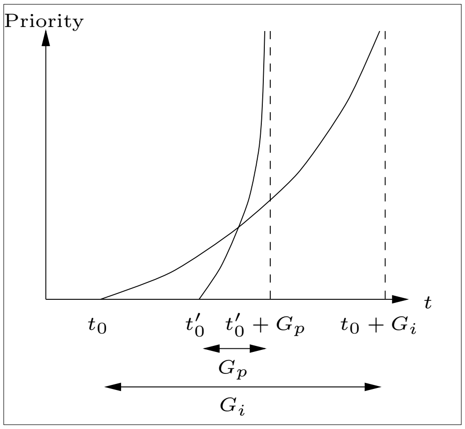
Mean queue length $\overline{Q}_r$ for an M/M/1 priority system with time dependent priorities having starting priorities r_r



- Priority function with upper time limit u_r :
  - In many real time systems, a job has to be serviced within a **upper time limit**.
  - Then it is advantageous to use a priority function that increases from 0 to ∞ between the arrival time t_0 and the upper time limit u_r:
  - Priority function:

$$q_r(t) = \begin{cases} (t-t_0)/(u_r - t + t_0) & t_0 < t \le u_r + t_0\,, \\ \infty & u_r + t_0 \le t\,. \end{cases}$$

- Priority function with upper time limit u_r:



$(G_r = u_r)$

◆ Mittlere Wartezeit eines Auftrags der Prioritätsklasse *r* :

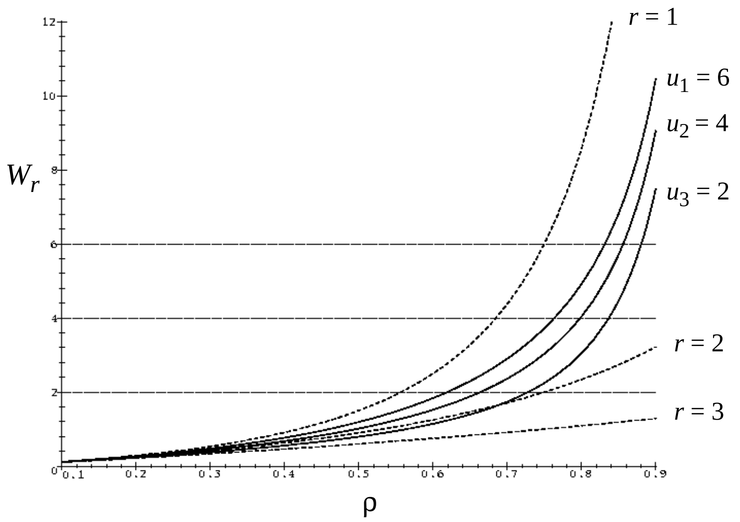
➤ Heavy traffic approximation (ρ --> 1):

$$\overline{W}_r \approx \left(\frac{\overline{W}_0}{1-\rho} - P_m \sum_{i=1}^{r-1} \rho_i (u_i - u_r)\right)\left(1-(1-P_m)\sum_{i=r+1}^{R}\rho_i\left(1-\frac{u_i}{u_r}\right)\right)^{-1}$$

➤ More accurate approximation (0 < ρ < 1):

$$\overline{W}_r \approx \left(\frac{\overline{W}_0}{1-\rho} - \sum_{i=1}^{r-1}\rho_i \cdot \overline{W}_i\left(1-\frac{u_r}{u_i}\right)\left(1-P_m\exp\left(-\frac{\rho u_i}{\overline{W}_i}\right)\right)\right) \cdot \left(1-\sum_{i=r+1}^{R}\rho_i\left(1-\frac{u_i}{u_r}\right)\left(1-P_m\cdot\exp\left(-\frac{\rho u_r}{\overline{W}_r}\right)\right)\right)^{-1} .$$

Mean waiting time W_r for an M/M/1 priority system with upper time limits and static priorities:
:


$r = 1$
$u_1 = 6$
$u_2 = 4$
$u_3 = 2$
$r = 2$
$r = 3$
W_r
12
10
8
6
4
2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
ρ