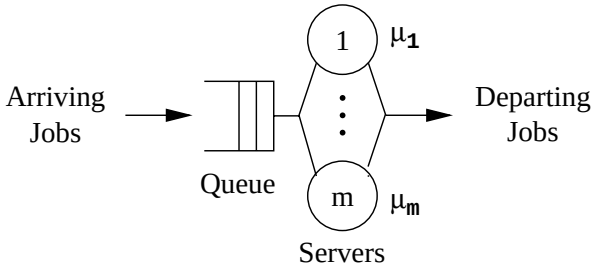
D.6 Heterogeneous Queueing Systems



- G/G/m - Systems with different service rates μ_i ($1 \le i \le m$)
- Strategies for the next empty server if more then one server is empty (If the servers are identical no strategy is necessary !):
  - Selection of the **fastest** free server: **FFS** (Fastest Free Server)
  - **Random** Selection: **Random**
- Very important queueing system, e.g. in manufacturing

1 Approximate Analysis

- "Heavy traffic"-Aprroximation, ($\rho \to 1$)
  - Usually no or at most one server is empty
  - Strategy has no influence
- Probability p_k ($1 \quad k \le m$), a job is served by the server k is proportional to the service rate μ_k:

$$p_k \approx \frac{\mu_k}{\sum_{i=1}^{m} \mu_i}$$

- Utilization of server k:

$$\rho_k \approx \frac{\lambda}{\mu_k} \cdot p_k = \frac{\lambda}{\mu_k} \cdot \frac{\mu_k}{\sum_{i=1}^{m} \mu_i} = \frac{\lambda}{\sum_{i=1}^{m} \mu_i}$$

◆ The utilization of all servers is identical and the utilization of the whole system is given by:

$$\rho = \rho_k$$

◆ Using the formulas for homogeneous M/M/m-, M/G/m- und G/G/m-systems the performance measures for heterogeneous systems can be calculated approximately, e.g. we have for homogeneous M/M/m - systems:

$$\overline{K} = m\rho + \frac{\rho}{1-\rho} \cdot P_m$$

with:

$$P_m = P(K \geq m) = \sum_{k=m}^{\infty} \pi_k$$
$$= \frac{(m\rho)^m}{m!(1-\rho)} \cdot \pi_0$$

◆ In the table approximate and also exact values for performance measures of heterogeneous a) M/M/5- und b) M/E_2/5- systems are given:

	λ	μ_k	$\rho(A)$	$\rho(S_1)$	$\rho(S_2)$	$\overline{Q}(A)$	$\overline{Q}(S_1)$	$\overline{Q}(S_2)$
(a)	73	10,15,20,20,25	0.811	0.799	0.819	2.472	2.367	2.495
	73	16,17,18,19,20	0.811	0.807	0.812	2.472	2.476	2.500
	73	5,10,18,22,35	0.811	0.808	0.844	2.472	2.443	2.626
	81.11	4,8,16,32,40	0.811	0.826	0.858	2.472	2.549	2.713
	81.11	8,9,20,31,32	0.811	0.807	0.839	2.472	2.456	2.606
	81.11	8,14,20,26,32	0.811	0.801	0.830	2.472	2.442	2.569
	81.11	10,15,20,25,30	0.811	0.799	0.824	2.472	2.425	2.523
(b)	81.11	10,15,20,25,30	0.811	0.800	0.825	1.854	1.854	1.962

A: Approximation, S_1: FFS, S_2: Random

2 Heterogeneous M/M/2 - System

- Strategy: **FFS**
- State of the system: (k_1, k_2)
  - $k_1 \geq 0$: Number of jobs in the Queue + job in the fast server
  - $k_2 \in \{0, 1\}$: Number of jobs in the slow server
- Steady state probalities (using Markov analysis):

$$\pi(k,1) = c\pi(k-1,1) = c^{k-1}\pi(1,1), \qquad \text{for} \qquad k > 1$$

und:

$$c = \frac{\lambda}{\mu_1 + \mu_2}$$

and

$$\pi(0,1) = \frac{c}{1+2c}\frac{\lambda}{\mu_2}\pi(0,0)\,,$$

$$\pi(1,0) = \frac{1+c}{1+2c}\frac{\lambda}{\mu_1}\pi(0,0)\,,$$

$$\pi(1,1) = \frac{c}{1+2c}\frac{\lambda(\lambda+\mu_2)}{\mu_1\mu_2}\pi(0,0)\,.$$

Normalizing condition ($\Sigma\, \pi(k_1, k_2) = 1$):

$$\pi(0,0) = [1 + \frac{\lambda(\lambda+\mu_2)}{\mu_1\mu_2(1+2c)(1-c)}]^{-1}$$

◆ Utilization of the servers:

$$\rho_1 = 1 - \pi(0,0) - \pi(0,1)$$
$$\rho_2 = 1 - \pi(0,0) - \pi(1,0)$$

◆ Mean number of jobs:

$$\overline{K} = \frac{1}{A(1-c)^2}$$

with:

$$A = \left[\frac{\mu_1\mu_2(1+2c)}{\lambda(\lambda+\mu_2)} + \frac{1}{1-c}\right]$$

3 Heterogeneous M/M/m - System

■ **Two Steps:**

◆ Analysis of a heterogeneous **loss system**

➤ Loss system has no queue

➤ Jobs, which arrive, when all servers are active, are lost

◆ The performance measures of the heterogenous queueing system can be calculated from the performance measures of the corresponding heterogeneous loss system.

- Heterogeneous M/M/m - loss system:
  - Strategy: **Random**
    - Steady state probabilities:

$$\pi_g = \frac{(m - |g|)!}{m!} \cdot \prod_{k \in g} \frac{\lambda}{\mu_k} \cdot \pi_\emptyset$$

$$\text{for all } g \subseteq G \text{ with } g \neq \emptyset \text{ and } G = \{1, 2, \ldots, m\}$$

π_0 is calculated using the normalizing condition:

$$\sum_{g \subseteq G} \pi_g = 1 \qquad \text{with} \qquad G = \{1, 2, \ldots, m\}$$

- Loss probability (P(all servers are occupied)):

$$\pi_m^{(L)} = \pi_{\{1,2,\ldots,m\}} = \pi_G$$

- Utilization of the k th server: Sum of the state probabilities of all states in which the k th server is occupied:

$$\rho_k^L = \sum_{g\,:\,k \in g} \pi_g$$

◆ Strategy: **FFS**

➤ Abbreviation:

$$\mu^m = (\mu_1, \mu_2, \cdots, \mu_m)$$

and

$$(\mu^{k-1}, \mu_m) = (\mu_1, \mu_2, \cdots, \mu_{k-1}, \mu_m)$$

➤ Loss probability:

$$\pi_m^{(L)} = \pi_{\{1,\ldots,m\}}(\mu^m) = B_m(\mu^m)$$
$$= B_{m-1}(\mu^{m-1}) \cdot \left[1 + \frac{\mu_m}{\lambda} \cdot \prod_{k=1}^{m-1} \frac{B_k(\mu^{k-1}, \mu_m)}{B_k(\mu^{k-1}, \mu_k + \mu_m)}\right]^{-1}$$

➤ with:

$$B_1(\mu^1) = \pi_{\{1\}}(\mu_1) = \frac{\lambda}{\lambda + \mu_1}$$

➤ Utilization of the servers:

$$\rho_k = \frac{\lambda}{\mu_k}[B_{k-1}(\mu^{k-1}) - B_k(\mu^k)] \qquad \text{with} \qquad B_0(\mu^0) = 1$$

Heterogeneous M/M/m - queueing system:

◆ Steady state probabilities:

$$\pi_i^{(W)} = \left(\sum_{k=1}^{m} \frac{\mu_k}{\lambda} \right)^{m-i} \cdot \pi_m^{(W)} \qquad \text{for} \qquad i > m$$

with

$$\pi_m^{(W)} = \frac{\pi_m^{(L)}}{N}$$

and

$$N = 1 + \frac{\pi_m^{(L)} \cdot c}{1-c} \qquad \text{with} \qquad c = \frac{\lambda}{\sum_{k=1}^{m} \mu_k}$$

◆ Probability of waiting:

$$P_m = \sum_{i=m}^{\infty} \pi_i^{(W)} = \frac{1}{1-c} \cdot \pi_m^{(W)} = \frac{1}{1-c} \cdot \frac{\pi_m^{(L)}}{N}$$

◆ Mean queue length:

$$\overline{Q} = \sum_{i=m+1}^{\infty} (i-m) \cdot \pi_i^{(W)} = \frac{P_m \cdot c}{1-c}$$

◆ Utilization of the servers:

$$\rho_k^{(W)} = \frac{\rho_k^{(L)}}{N} + \sum_{i=m+1}^{\infty} \pi_i^{(W)} = \frac{\rho_k^{(L)} + \pi_m^{(L)} \cdot c/(1-c)}{N}$$

◆ Total utilization (Mean proportion of active servers aktiver):

$$\rho = \frac{m}{\sum_{k=1}^{m} \rho_k^{-1}}$$

◆ Performance measures of a heterogeneous M/M/2 - system with $\lambda = 0.2$, $\mu_1 = 0.5$ and $\mu_2 = 0.25$:

Strategy	Appr.	FFS(M/M/2)	FFS(M/M/m)	Random
ρ	0.267	(0.267)	0.234	0.275
ρ_1	0.267	0.305	0.306	0.230
ρ_2	0.267	0.189	0.189	0.341
$\overline{K}$	0.575	0.533	0.533	0.615
$\overline{T}$	2.875	2.663	2.665	3.074

FFS: Fastest Free Server

◆ Mean response time $\overline{T}$ of a heterogeneous M/M/2 -systems (Strategy: **FFS**) and mean response times $\overline{T}_1$ und $\overline{T}_2$ of two M/M/1 - Systeme $\overline{T}_1$ und $\overline{T}_2$ ($\lambda = 0.2$, $\mu_2 = 0.25$ und $\mu_1 = \alpha\mu_2$):

α	1	2	3	4	5
$\overline{T}_1$	20	3.33	1.818	1.25	0.55
$\overline{T}_2$	20	20	20	20	20
$\overline{T}$	4.762	2.662	1.875	1.459	1.20

➤ Sometimes it is better to disconnect the slower server if we wish to minimize the mean response time.

➤ This is the case only for low utilization of the servers and if the service rates differ considerably.

D.7 Batch-Systems

◆ Several jobs (**batch**) are started and executed in parallel.

◆ **Full-Batch**-strategy (FB):
A batch is started, when all b jobs of the batch are arrived.

◆ **Minimum-Batch**-strategy (MB):
A batch is started, when at least a jobs of the batch are arrived.
If more than b jobs are waiting, then b jobs were merged to a batch and executed in parallel. .

◆ A special case of the MB-Strategy is the **GREEDY**-strategy:
The executing starts, when $a = 1$ job is in the queue.

◆ Kendall's Notation for Batch-Systems:

$$\mathrm{G/G^{[b,b]}/m}, \text{ Multiserver system with FB policy}$$

$$\mathrm{G/G^{[a,b]}/m}, \text{ Multiserver system with MB policy}$$

◆ Full-Batch (FB):

➤ Calculation of the mean queue length $\hat{Q}$ for the batches using the wellknown formulas for G/G/m-systems:

➤ Arrival rate of the batches (λ : arrival rate of the jobs):

$$\hat{\lambda} = \frac{\lambda}{b}$$

➤ Coefficient of variation of the interarrival time of the batches (c_A^2: coefficient of variation of the interarrival time of the jobs):

$$\hat{c}_A^2 = \frac{c_A^2}{b}$$

➤ Mean queue length of a single job in a Full-Batch-System:

$$\overline{Q} \approx b \cdot \hat{Q} + \frac{b-1}{2}$$

◆ Minimum-Batch (MB):

➤ Mean queue length of a sinle job:

$$\overline{Q}^{[a,b]} \approx b \cdot \hat{Q} + P_m \cdot \frac{b-1}{2} + (1 - P_m) \cdot \frac{a-1}{2}$$

P_m: Probability of waiting

◆ Example: M/M$^{[b,b]}$/m - System

➤ Full-Batch system (batch length b)

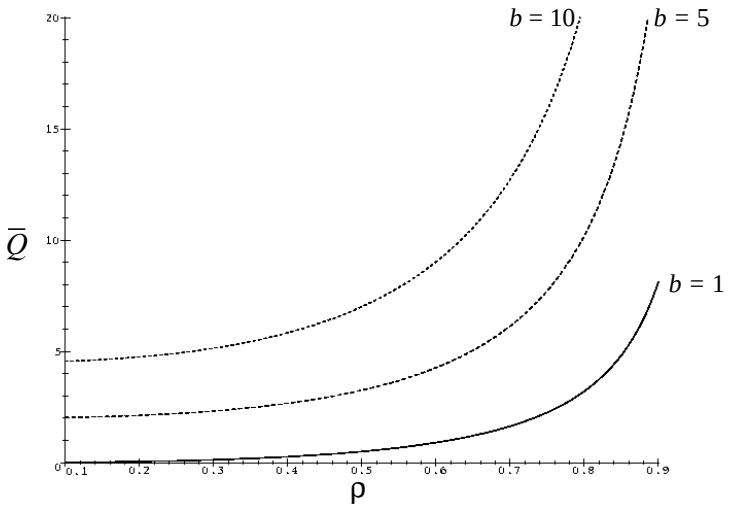
➤ Exponentially distributed interarrival time and service time of the jobs of the batch

➤ Service time of the batch is also exponentially distributed , since the jobs are executed in parallel.

➤ Interarrival time of a batch is E_b- distributed (b jobs have to be arrived to start with the execution). The sum of exponentially distributed times has a Erlang distribution.

➤ To calculate the mean queue length for the batches $\hat{Q}$ a E_b/M/m-system has to be used.

Mean queue length $\overline{Q}$ for a single job in a FB-System



Mean queue length $\overline{Q}$ for a single job in a MB-System ($b = 10$)