eCS5239 COMPUTER SYSTEM PERFORMANCE ANALYSIS

Assignment 2: Queueing Theory

Liu Yuancheng(A0159354X)

1. **As a performance analyst, you are asked to investigate the performance of a simple database server system - question (b) to (e) below. The database server consists of a CPU and two disks and receives requests at a rate of 18,000 requests per hour. Each request requires 0.15 second of CPU service and performs on average 5 I/Os on disk 1 and 3 I/Os on disk 2. Each I/O request takes on average 0.03 second and 0.02 second on disk 1 and disk 2 respectively.**

$\rho = \lambda/\mu$

1.1 Briefly explain where and how you obtain the input data as listed above.

The receives request rate is obtained from Database request Logfile.

request rate = number of request/ time

The CPU, Disk1 and Disk2 data is obtained from the computer system Logfile.

Receives requests at a rate of 18000 requests per hour.

⇨ $\lambda = \frac{18000}{h=3600s} = 5 \text{ request/ sec}$

Each request requires 0.15 second of CPU service.

⇨ S_{cpu} = 0.15 s

Average 5 I/Os on disk 1, Each I/O request takes on average 0.03 second on on disk 1:

⇨ S_{Disk1} = 0.03s x 5 = 0.15 s

Average 3 I/Os on disk 2, Each I/O request takes on average 0.02 second on on disk 2:

⇨ S_{Disk2} = 0.02 x 3 = 0.06 s

1.2 What are the average response time per request, average throughput of the database server and the utilizations of the CPU and disks?

We assume system is job flow balanced, $A_i = C_i$

Average response time per request: $S = S_{cpu} + S_{Disk1} + S_{Disk2}$ = 0.36 S

Average throughput: $X_i = C_i/T = 5 \text{ request/ sec}$

Utilizations of the CPU: $U_{cpu} = B_{cpu}/T = X_i S_{cpu}$ = 0.75

Utilizations of the Disk1: $U_{Disk1} = B_{Disk1}/T = X_i S_{Disk1}$= 0.75

Utilizations of the Disk2: $U_{Disk2} = B_{Disk2}/T = X_i S_{Disk2}$= 0.30

1.3 What is the maximum theoretical request arrival rate that can be sustained by the database server?

If the system is job flow balance, the maximum arrival rate should be less that the departures rate $X(N) <= \{1/D_{max}, N/(D+Z)\}$

$X(N)_{max} = 1/0.15 = 6.7$requst/s = 24000 reqs/ hour

1.4 What are the top two bottleneck devices? Given a choice to relieve only one bottleneck device, which device will you choose and does it matter? Justify your answer.

Bottleneck devices: CPU and Disk1

My answer is relieving CPU: As the data show the disk2 I/O is faster than the disk1, so if we swap the contents between disk1 and disk2 can increase the performance but if we want to increase the performance of the CPU we can only relieve it.

1.5 If we improve the performance of the bottleneck device(s) by 50%, i.e., reduce the service time by half, re-compute the average response time per request and the maximum theoretical database service request arrival rate and comment on the results.

$S_{cpu} = 0.15$ s/2 = 0.075s

$S_{Disk1} = 0.03$s x 5 x 0.5 = 0.075s

Average response time per request: $S = S_{cpu} + S_{Disk1} + S_{Disk2} = 0.21$s

$X(N) <= \{1/D_{max}, N/(D+Z)\}$

$X(N)_{max} = 1/0.075 = 13.3$requst/s = 48000 reqs/ hour

Comments: If we increase the bottleneck device twice and the device is still bottleneck device after the incensement, the system performance will be doubled. (Doubled max theoretical arrival rate)

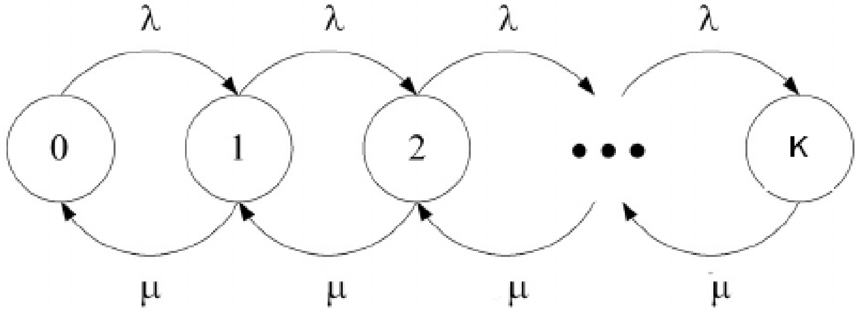
2. **A two-stage cyclic queueing system consisting of two servers, S1 and S2, with exponential service rates of λ and μ. There are K jobs in this closed system. If N(t) is the number of jobs at S2 at time t then the number of jobs at S1 is K – N(t).**

2.1 Show that N(t) can be modeled as a birth-and-death (BD) process and find the BD rates. State your assumption if any.

Assumption: The inter-arrival time and service time of the system S1 and S2 are exponential. There are no new jobs join in the system and the jobs will not leave the system.

For service S2, its arrival rate is as same as the S1's departure rate, the arrival rate of S2 is λ , the service rate of S2 is μ.

The state-transition diagram of the N(t) BD process is:



2.2 What is the steady-state probability distribution of N(t)?

According to the state-transition diagram the system is a M/M/1/K system with $\rho = \lambda/\mu$.

n = 0 $P_0 = (1- \rho)/(1- \rho^{K+1}) = (1- \lambda/\mu)/[1- (\lambda/\mu)^{K+1}]$, (ρ != 1)

$P_0 = 1/(K+1)$, (ρ == 1)

n > 0 $P_n = \rho^n \times P_0 = (1-\rho)\, \rho^n /(1- \rho^{K+1}) = (1- \lambda/\mu) x (\lambda/\mu)^n/[1- (\lambda/\mu)^{K+1}]$, (ρ != 1)

$P_n = 1/(K+1)$, (ρ == 1)

2.3 Derive expressions for the expected number of jobs in S1 and S2.

For system S2: $\rho = \lambda/\mu$

Mean number of jobs in S2:

$E[N(t)] = \sum_0^k nP_n = \rho/(1- \rho) – (K+1)\, \rho^{K+1}/(1- \rho^{K+1})$

$= \mu/(\lambda- \mu)- (K+1)\, (\lambda/\mu)^{K+1}/[1-(\lambda/\mu)^{K+1}]$

There are total K jobs in the closed system, $E[n\_S_1]$ = K- E[N(t)], the mean number of jobs in S1:

$E[n\_S_1] = E[K-N(t)] = K - E[N(t)] = K-(\mu/(\lambda- \mu) - (K+1)\, (\lambda/\mu)^{K+1}/[1-(\lambda/\mu)^{K+1}])$

3. As a system manager, you are evaluating three system configurations proposed by different computer vendors as shown below. For simplicity, assume your total organization workload has an arrival rate of λ, and the expected total service rate is kμ. Thus, ρ = λ/(kμ).

3.1 Which system is better: M/M/1 or k M/M/1? Explain your answer.

M/M/1 Queue: $\rho = \lambda/(k\mu)$

Average jobs in the system: $E[n] = \rho/(1-\rho) = \lambda/(k\mu-\lambda)$

Average response time: $E[r] = (1/ \mu)/1- \rho = 1/k\, \mu- \lambda$

Average Waiting time: $E[w] = \rho \times (1/\mu)/(1-\rho) = \lambda/(k\mu-\lambda)\, k\mu$

K M/M/1 Queue: $\rho = \lambda/(k\mu)$

Average jobs in the system: $E[n] = k \times E[n'] = k \times \lambda/k\mu / (1-\lambda/(k\mu)) = k\lambda/(k\mu-\lambda)$

Average response time: $E[r] = (1/ \mu)/1- \rho = k/k\, \mu- \lambda$

Average Waiting time: $E[w] = \rho \times (1/\mu)/(1-\rho) = \lambda/(k\mu-\lambda)\, \mu$

So with the same throughput rate, for the average waiting time and average response time in the k M/M/1 Queue is longer than the M/M/1 Queue. So the M/M/1 queue is better.

3.2 How does the M/M/1 system compare with the M/M/k system when the load is low and the load is high?

From 3.1 the waiting time of M/M/1 queue is: $E[w] = \rho x(1/\mu)/(1-\rho) = \lambda/(k\mu-\lambda)\, k\mu$

For the M/M/K system: $\rho = \lambda/(k\mu)$

$\varepsilon = P_0\, x(k\rho)^k/k!(1-\rho)$

$P_0 = [1+(k\rho)^k/k!(1-\rho) + \sum_1^{k-1} (k\,\rho)\hat{}k/n!]^{-1}$

Average Waiting time: $E[w'] = \varepsilon/[k\mu((1-\rho))]$

Compare the average waiting time between M/M/1 queue and M/M/K queue
E[w]/E[w'] = ε/kρ = P_0 x$(k\rho)^{k-1}$/k!(1-ρ)

E[w]/E[w'] = $(k\rho)^{k-1}$/k!(1-ρ)/ [1+$(k\rho)^{k}$/k!(1-ρ) + $\sum_{1}^{k-1}(\mathrm{k}\,\rho)\hat{}\mathrm{k}/\mathrm{n}!$] $\approx$ θ(n!/k!)

when the work load is low n < K
E[w] / E[w'] < 1 => E[w] < E[w']
Waiting time: M/M/1 is less than M/M/k => We select the M/M/1 system
When the work load is high n >= K
E[w] / E[w'] > 1 => E[w] < E[w']
Waiting time: M/M/1 is more than M/M/k => We select the M/M/k system

So when the work load is low we select the M/M/1 queue system and when the workload is high we select the M/M/k queue system.