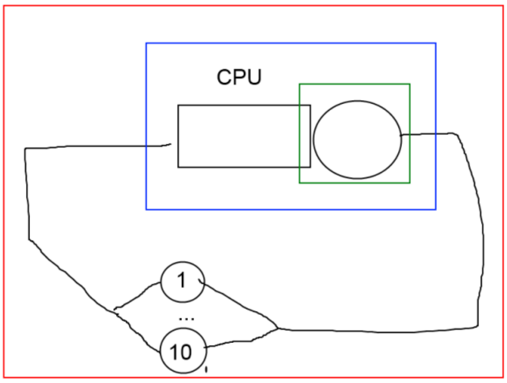
Esercizio 1

Monitoriamo un sistema ottenendo un utilizzo CPU di 0.75, un CPU service demand di 3 s, un response tme di 15 s e 10 utenti attivi. Qual è il think time medio degli utenti?



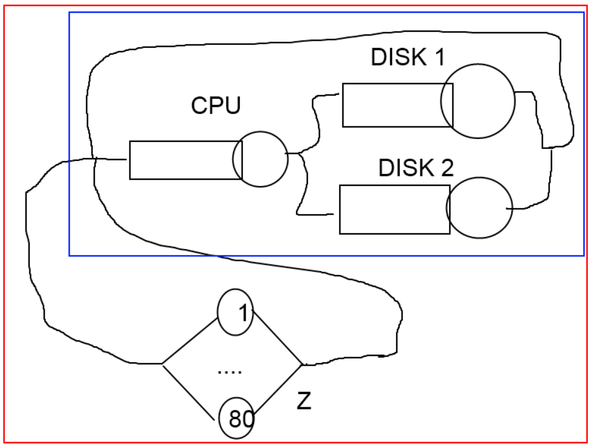
Legge di Little: N = X R

Box verde -> Legge dell'utilizzo: U = X D -> X = U / D = 0.75 / 3 s = 0.25 job/s
Box rosso -> Legge del tempo di risposta: N = X (R + Z) - > N/X = R + Z -> Z = N/X – R = 10 / (0.25 job/s) – 15 s = 40 s – 15 s = 25 s

Esercizio 2

Un sistema interattivo con una memoria limitata

- Monitorato per 1 ora
- Numero di utenti è 80
- Tempo medio di risposta per una transazione è 1 secondo
- Transazioni completate è 36000
- Utilizzo CPU è 75%
- Utilizzo del disco 1 è del 50%
- Utilizzo del disco 2 è anch'esso del 50%



<u>In media quanti utenti NON stanno pensando?</u>

$X = C / T = 36000 / 3600$ s = 10 job/s
Legge di Little: $N = X R$ = 10 job/s * 1 s = 10

Esercizio 3

In un sistema batch, uno specifico disco sta effettuando, in media, 12 operazioni per secondo. Inoltre, noi sappiamo che ogni transazione batch richiede, in media, 6 accessi a questo disco. Un secondo disco nel sistema sta gestendo, in media, 18 operazione al secondo. Qual è il numero medio di accessi a questo secondo disco richiesti per ogni transazione batch?

X: throughput del sistema
V_k: visite alla service station k
X_k: throughput della service station

Legge dei flussi forzati: $X_k = X V_k$

$X_{disk1} = 12$ op/s
$V_{disk1} = 6$
$X_{disk2} = 18$ op/s
$V_{disk2} = ?$

$X_{disk1} = X V_{disk1} \to X = X_{disk1} / V_{disk1}$ = 12 op/s / 6 = 2 tran/s
$X_{disk2} = X V_{disk2} \to V_{disk2} = X_{disk2} / X$ = 18 op/s / 2 tran/s = 9

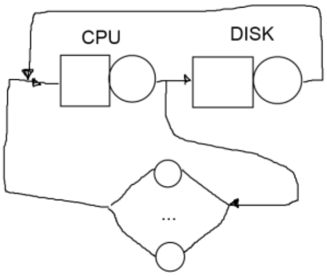
Esercizio 4

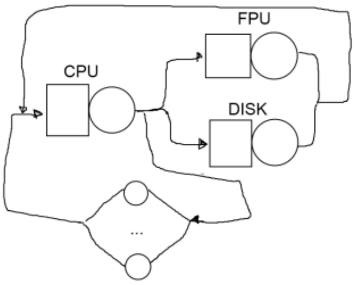
A session of a graphical multi-user workstation, using a disk with an average service time $S_{disk} = 25$ ms, yields the following measurements:

- average think-time, z = 10 s
- average CPU service demand, $D_{cpu} = 4$ s
- average disk service demand, $D_{disk} = 5$ s
- fraction of the busy time in which the CPU performs floating point operations, 75%

Evaluate, using asymptotic bounds, which of the following modifications is more advantageous:

- adding a FPU, which is 10 times as fast as the CPU, to offload floating point operations
- replacing the disk with a new one with $S'_{disk} = 15$ ms





First modification:
$D'_{cpu} = (1 – 0.75)\ D_{cpu} = ¼\ 4$ s = 1 s
$D_{fpu} = 0.75\ D_{cpu} / 10 = ¾\ 4$ s / 10 = 3 s / 10 = 0.3 s

Second modification:

S_{disk} = 25 ms = 0.025 s

D_{disk} = 5 s

$V_{disk} = D_{disk} / S_{disk}$ = 5 / 0.025 = 200

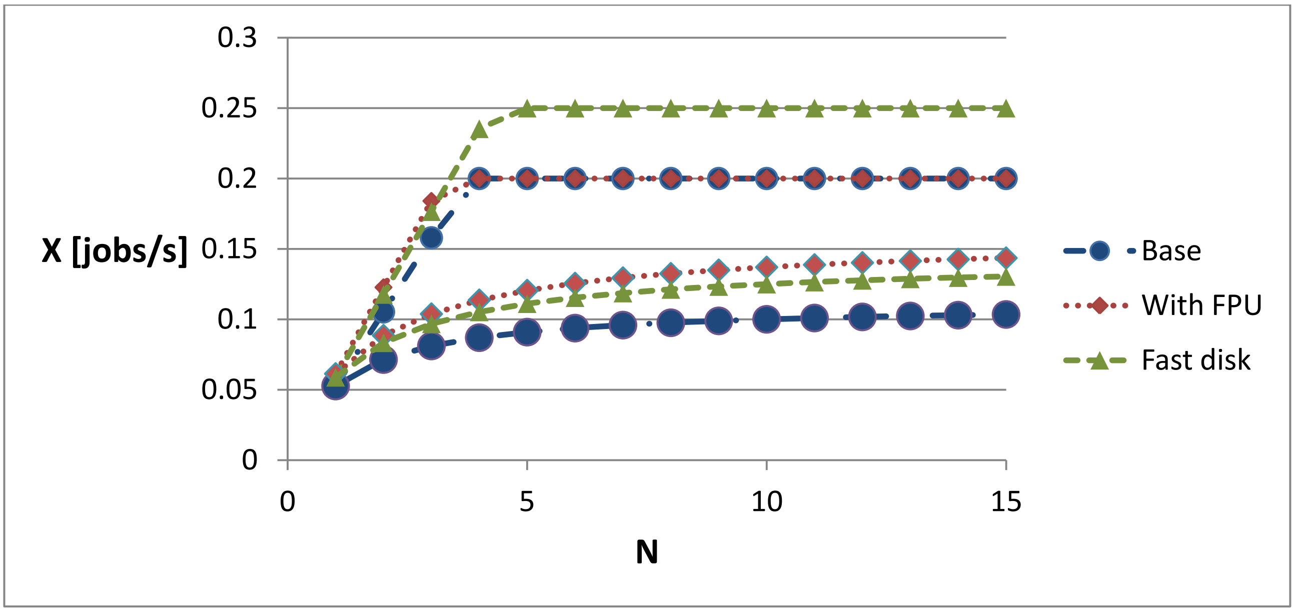
S'_{disk} = 15 ms = 0.015 s

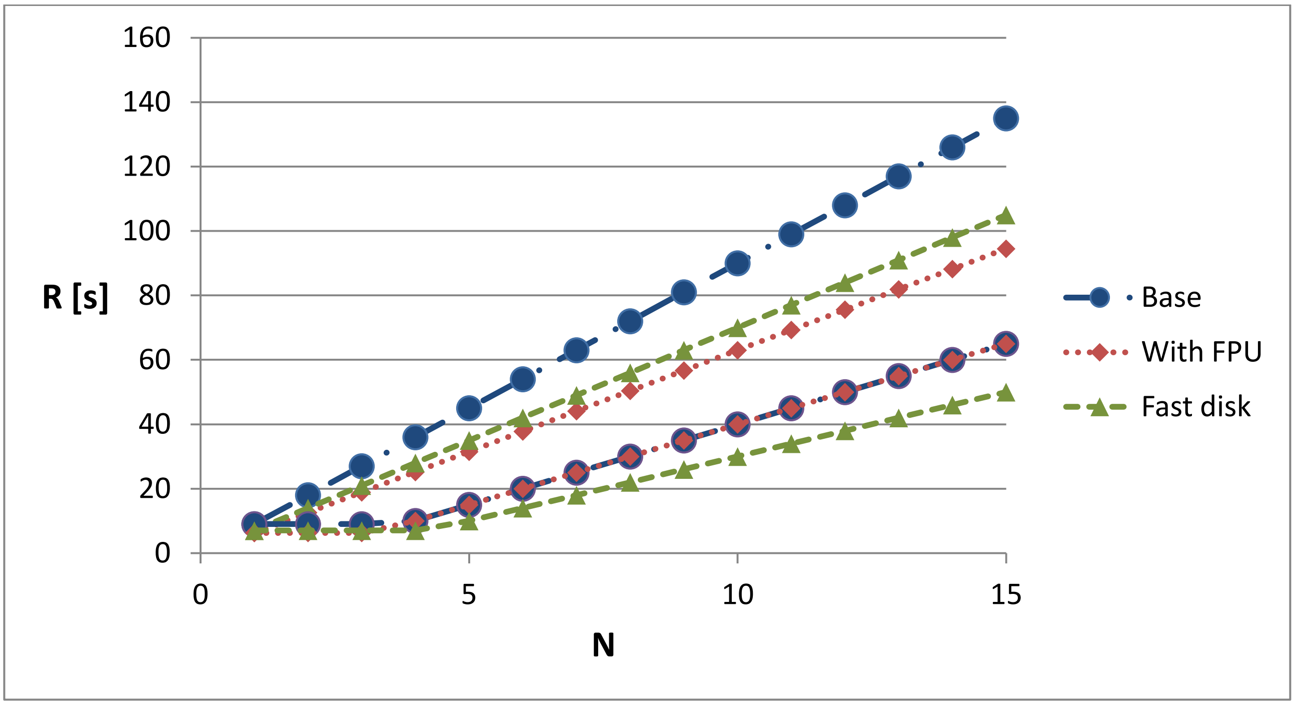
$D'_{disk} = V_{disk} S'_{disk}$ = 200 * 0.015 s = 3 s

System	**D_{cpu} [s]**	**D_{fpu} [s]**	**D_{disk} [s]**	**Z [s]**	**D [s]**	**D_{max} [s]**
Base	4	0	5	10	9	5
With FPU	1	0.3	5	10	6.3	5
Fast disk	4	0	3	10	7	4

Response time law:	N = X (R + Z)		
	R = N/X - Z		
	X = N / (R + Z)		
Optimistic bound:	R >= D	N/X - Z >= D	X <= N / (D + Z)
Pessimistic bound:	R <= ND	N/X - Z <= ND	X >= N / (ND + Z)
Optimistic bound:	X <= $1/D_{max}$	N/(R+Z) <= $1/D_{max}$	R >= N D_{max} - Z

System	**1 / Dmax [jobs/s]**	**N***	**1 / (D + Z) [jobs/s]**	**1 / D [jobs/s]**
Base	0.2	3.8	0.053	0.111
With FPU	0.2	3.26	0.061	0.159
Fast disk	0.25	4.25	0.059	0.143




160
140
120
100
80
60
40
20
0
R [s]
0
5
10
15
N
Base
With FPU
Fast disk

Exercises – Performance Evaluation

10 / 5 / 2013

Exercise 1

The storage server of an intranet consists of two groups of disks, A and B, each having exponential distributed service times, with means S_A = 5ms and S_B = 3ms. The mean number of visits for the two components are V_A= 20 and V_B = 30. The throughput of A is 150 op./sec. The above data were collected when the system is processing a workload generated by 300 users with think time Z = 15 sec.

I) Compute the System Throughput X_0 and the Utilization of B. Which one of the two groups is the bottleneck?

II) Compute the system response time.

III) If the number of users increases to 400, which will be the new response time?

Solution

I) $D_A = V_A \cdot S_A = 20 \cdot 0.005 = 0.1s$;

$U_A = X_A \cdot S_A = 150 \cdot 0.005 = 0.75$;

$$U_A = X_0 \cdot D_A \rightarrow X_0 = \frac{U_A}{D_A} = \frac{0.75}{0.1} = 7.5\, int/sec$$

$U_B = X_0 \cdot D_B = X_0 \cdot (S_B \cdot V_B) = 7.5 \cdot 30 \cdot 0.003 = 0.675$

II) $R = \frac{N}{X_0} - z = \frac{300}{7.5} - 15 = 40 - 15 = 25\, s;$

III) The response time **R when the number of users increases to400cannot be computed** with the operational analysis equations since we cannot use the value of system throughput obtained when the N was 300. In this new case we do not know the new throughput.

Exercise 2

An intranet is composed of 5 web servers used in parallel, 3 application servers used in parallel, and one storage server. The other components on the intranet (e.g., switches, gateways, load balancers, firewalls, network) are not considered since their utilization is very low. The server connected in parallel are used in a balanced way. The complete execution of a transaction requires (service demands) 750 ms to the web server, 600 ms to the application server and 300 ms to the storage server.

Compute the maximum throughput of the intranet.

Solution

$$D_{ws} = \frac{750}{5} = 150\ ms\ for\ each\ web\ server;\ D_{as} = \frac{600}{3} = 200\ ms\ for\ each\ application\ server;$$

$$D_{ss} = 300\ ms\ for\ the\ storage\ server;$$

$$X_{max} = \frac{1}{D_{max}} = \frac{1}{300} = 0{,}003\ int/sec$$

Exercise 3

The throughput of a disk is 100 I/O operations per second. To complete a given request 20 visits to the disk are required. The number of users is 100 and the response time is 15 seconds.

Compute the users think time.

Solution

$$\mathrm{R} = \frac{\mathrm{N}}{\mathrm{X_0}} - \mathrm{z} \rightarrow \mathrm{Z} = \frac{\mathrm{N}}{\mathrm{X_0}} - \mathrm{R} = \frac{100}{\mathrm{X_0}} - 15$$

$$\mathrm{X_D} = \mathrm{V_D} \cdot \mathrm{X_0} \rightarrow \mathrm{X} = \frac{\mathrm{X_D}}{\mathrm{V_D}} = \frac{100}{20} = 5\ \mathrm{int/sec}$$

$$\mathrm{Z} = \frac{100}{5} - 15 = 20 - 15 = 5\ \mathrm{sec}$$

Exercise 4

Let's consider an intranet that can be accessed by a large number of users. The execution of a single request pass through an application server (AS), which has a service time S = 300 ms, then through a database server (DS), which has a service time S = 250 ms, and then back through the application server. A request must pass through the system firewall before entering the intranet and before exiting from it. The firewall service time per visit is S = 10 ms.

I) Compute the maximum throughput of the system.

II) It is possible to have a Response Time R < 9 s? At which conditions?

Solution

$By\ drawing\ the\ intranet\ model, the\ visits\ V_{as} = 2;\ V_{ds} = 1;\ V_{fw} = 2;\ can\ be\ obtained.$

$Demands\ can\ then\ be\ computed. D_{as} = 600;\ D_{ds} = 250;\ D_{fw} = 20;$

I) $X_{max} = \frac{1}{D_{max}} = \frac{1}{600};$

II) If we assume that the intranet is modeled by a **closed model** with a think time $Z = 0$, we have:

$N\,D_{max} - Z \le R$; $N\,D_{max} \le R$; $N\ 600 \le 9000$; $N \le \frac{9000}{600}$; $N \le 15$

If we assume that the intranet is modeled by an **open model,** even if:

$X_0 \le \frac{1}{D_{max}}; X_0 \le \frac{1}{600}$

thus, the system is not saturated, there is no pessimistic bound on the response time. For such reason, we can not provide conditions at which the constraint was satisfied, as done in the **closed model**.

Exercise 5

A web server of a company is connected to an intranet and is accessed by the employees that work internally in the company resulting in a population of fixed size: N=21 users. The average think time of the users is Z=20 sec. A complete execution of a request generate a load of V_s=20 operations to a specific storage device whose utilization is U_s= 0.30. The service time of the storage device per each visit is S_s =0.025 sec.

I) Determine the average system response time R
II) Compute the average throughput and system response time with N=40 users.

Solution

I) $U_s = X_s \cdot S_s \rightarrow X_s = \frac{U_s}{S_s} = \frac{0.3}{0.025} = 12$ int/sec;

$X_s = V_s \cdot X_0 \rightarrow 12 = 20 \cdot X \rightarrow X = \frac{12}{20} = 0.6$ int/sec;

$R = \frac{N}{X_0} - z = \frac{21}{0.6} - 20 = 35 - 20 = 15$ sec;

II) The response time **R when the number of users increases to 40 cannot be computed** with the operational analysis equations since we cannot use the value of system throughput obtained when the N was 21. Indeed, in this case we do not know the new throughput.

Exercise 6

The Intranet of a medium scale company consists of three servers, namely A, B and C, which represent the web server of the clients, the application server and the database server, respectively. The number of users is constant, N=20 users. In order to evaluate the performance of the system a 10 minutes monitoring phase has been performed. The following data have been collected:

Server B number of completions, C_B = 150 op.
Server C number of completions, C_c = 300 op.
Network (intranet) completions, C= 100 op.
Server B busy time, B_B = 300s
Server C busy time, B_C = 100s

It is also known that the maximum throughput achievable by the intranet is 0.2 trans/sec. Compute:

I) the system throughput during the measurement phase
II) the service demands of all the servers and determine the server which should be upgraded to achieve the maximum gain of the network performance
III) the utilizations of all the servers
IV) the number of visits at server B
V) the system response time if the users have a mean think time of 30 sec.

Solution

I) $X_0 = \frac{C}{T} = \frac{100}{600} = \frac{1}{6} =$ 0.1666 int/sec

II) $D_B = \frac{B_B}{C} = \frac{300}{100} = 3\ sec; D_C = \frac{B_C}{C} = \frac{100}{100} = 1\ sec;$

Given that $X_{max} < \min\left\{\frac{1}{D_B}, \frac{1}{D_C}\right\}$ then $D_{max} = D_A = \frac{1}{X_{max}} = \frac{1}{0.2} = 5$ sec, A is the bottleneck.

III) $D_k = \frac{U_k T}{C} \rightarrow U_A = \frac{D_A C}{T} = \frac{5 \cdot 100}{600} = 0.833; U_B = \frac{D_B C}{T} = \frac{3 \cdot 100}{600} = \frac{3}{6} = {}^{1}/_{2} = 0.5;$

$U_C = \frac{D_C C}{T} = \frac{1 \cdot 100}{600} = \frac{1}{6} = 0.1667$

IV) $V_B = \frac{C_B}{C} = \frac{150}{100} = 1.5 \text{ visits}$

V) $R = \frac{N}{X_0} - z = \frac{20}{\frac{1}{6}} - 30 = 120 - 30 = 90 \text{ sec}$

Exercise 7 - Performance of an IT infrastructure

Let's consider an IT infrastructure consisting of a Web Server (WS), an Application Server (AS) and a Storage Server (SS).After 1 hour measurement, during which N = 50 users were working continuously, the following data have been collected:

C_0 total number of jobs executed by the system: 5400 j

C_{WS}Number of WS completed operations: 54000 op

C_{AS}Number of AS completed operations: 32400 op

C_{SS} Number of SS completed operations: 10800 op

B_{WS} WS total *activity* time: 1800 sec

B_{AS} AS total *activity* time: 720 sec

B_{SS} SS total *activity* time: 900 sec

Z Mean think time 5 sec

Using Operational Analysis equations:

1. Compute the visits V_i to the three servers during a complete job execution, their global service requests D_i and determine the bottleneck resource of the IT infrastructure

2. Compute response time when N = 50 users are connected, as well as the maximum throughput when the number of users tends to infinity (asymptotic value)

3. Let's substitute the bottleneck resource determined at point 1) with another, two times (2x) more powerful. Does the bottleneck migrate to another resource? If so, which one? Compute the new value of the asymptotic throughput.

Solution:

1)$V_{WS} = C_{WS}/ C_0 = 54000 / 5400 = 10; V_{AS} = C_{AS}/C = 32400/5400 = 6; V_{SS} = 2;$
$U_{WS} = B_{WS}/T = 1800/3600 = 1/2; U_{AS} = B_{AS}/T = 1/5; U_{SS} = 1/4;$
$X = C/T = 5400/3600 = 3/2$ j/s;
$D_{WS} = U_{WS}/X_0 = (1/2)/(3/2) = 1/3s; D_{AS} = U_{AS}/X_0 = 2/15s; D_{SS} = U_{SS}/X_0 = 1/6sec;$
bottleneck = WebServerD_{max} = (1/3)s;

2)R = (N/X) - Z = (50/3/2) - 5 = 85/3s; $X_{max} = 1/D_{max} = 3$ j/s;

3)new D_{WS} = 1/6 sec hence Web Server and Storage Server are now both bottlenecks, new volume asymptote of throughput $X_{max} = 1/D_{max} = 6$ j/s;

Performance exercises

An intranet of a company consists of a web server WS, an application server AS and a storage server SS. The system has been measured for an interval of time of T sec. and the following data were detected:

C number of interactions executed by the system: 3600 job
C_{WEB} number of operations completed by the web server 5400 op
C_{APP} number of operations completed by the application server 7200 op
C_{SS} number of operations completed by the storage server 10800 op
B_{WEB} busy time of web server 360 s
B_{APP} busy time of application server 720 s
B_{SS} busy time of storage server 1080 s
Z think time 4 s

The utilization of the local area network is negligible and thus should not be considered. Applying the operational analysis technique compute:

(a) the service demands of the three servers D_{WEB},D_{APP},D_{SS}
(b) knowing from the measurements that the utilization of the bottleneck is 0.6, compute the throughput of the system X_0
(c) compute the utilization of the two servers that are not bottleneck
(d) write the equations of the asymptotes of system throughput and system response time (assume that it is a closed system)
(e) if the bottleneck server will be substituted with a new one twice as fast, which will be the new bottleneck of the system? With this new configuration, would it be possible to have a system throughput of 4 j/s or greater?
(f) in the original system would it be possible to have a response time of 0.7s with 16 users?
(g) which is the duration of the observation interval T?

A cloud SaaS application has a fixed number N = 100 of registered users, having a mean think time Z = 15sec. During its execution, the application utilizes a web-server (WS), having a mean service time S_{WS} = 0.1sec, a database (DB) with mean service time S_{DB} = 0.05sec and a storage server (SS) with S_{SS} = 0.01sec. The complete execution of a transaction requires one access to the WS, 3 accesses to the DB, and 6 accesses to the SS. The utilization of the DB has been measured, its value is U_{DB}=0.9.

Compute:
(a) the throughput of WS, of DB, of SS and of the system
(b) the utilizations of the WS and of the SS
(c) the system response time R
(d) The system response time with N=200 users

A computing infrastructure consists of a web server (WS), an application server (AS), and a storage server (SS).
The service demands D_k are: D_{WS} = 10ms; D_{AS} =20ms; D_{SS} = 30ms.
The LAN and other components of the intranet are very lightly loaded and are not considered in the study.

(a) In this intranet will be possible to have a throughput of X = 40tr/sec? (show the computations)

(b) The management decides to use the storage of a cloud infrastructure. Half of the data stored in the local storage server will be allocated in this new cloud storage. At the end of the migration it will be: $D_{SS, int} = D_{SS,cloud}$ = 15ms. Which will be the maximum throughput of this new intranet? Which is the bottleneck resource?

(c) With a workload consisting of a constant number of users N = 500 users with think time Z = 5sec the throughput of the intranet is X = 25tr/sec. Compute the response time R of the intranet and the utilization of the servers in the configuration with the cloud storage (new intranet).

(d) Compute the minimum (theoretical) response time R of the new intranet with N = 600 users with think time Z = 5sec

A computing digital infrastructure consists of a web server (WS), an application server (AS), and a storage server (SS).
The service demands D_k of the web server and of the application server are: D_{WS}=100ms; D_{AS} = 150ms.
The LAN and other components of the intranet are very lightly loaded and are not considered in the study.

The service time of the Storage Server is S_{SS} = 2ms and its throughput is X_{SS}=0.4op/ms. The measured throughput of the network is X_0 = 4tr/sec.

(a) How many operations (visits) a request executes on the Storage Server during a complete execution? Compute the service demand D_{SS} of the Storage Server

(b) Compute the utilization of the three servers with the measured throughput.

(c) Which server is the bottleneck of the network, why? Compute the maximum throughput of the network.

(d) To support the business objective, a throughput of the network of X_0 = 8tr/sec is required. Which are the servers that must be replicated? Compute the number of replicas for each server that must be used.

(e) With the measured throughput of the network of 4tr/sec the number of customers in the network is N = 6.1666jobs. Which is the network response time R? (consider Z=0)

The intranet of a company is accessed by N = 100 employees that have a mean think time Z = 20sec. The execution of a typical transaction requires 10 accesses to the web server ws, whose service time is S_{ws} = 30ms, utilized at 60%.

(a) Compute the throughput X and the mean response time R of the intranet

(b) How many accesses to the storage server ss are generated by a complete execution of a transaction knowing that its throughput is X_{ss} = 20 op/sec?

(c) It is known that the mean service time of the storage server S_{ss} is 45 ms. Compute its utilization U_{ss}.

(d) Compute the maximum throughput that a system consisting of the web server and the storage server may obtain when the number of customers grows to infinite.

A computing digital infrastructure consists of a web server (WS), an application server (AS), and a storage server (SS).

The service demands D_k are: $D_{WS} = 10ms$; $D_{SS} = 15ms$.

The LAN and other components of the intranet are very lightly loaded and are not considered in the study. The maximum throughput of the network is 50 transactions/sec (tr/s).

(a) Which server is the bottleneck of the network? What is its D_k?
(b) Compute the utilization of the Web Server and of the Storage Server when the throughput is maximum.
(c) Consider a production workload whose execution causes a Storage Server utilization of 50%. Which will be the utilization of the Application Server with this production workload?
(d) When the throughput of the network is 45 tr/s the number of customers in the network is 11.895. Which is the network response time R? (consider Z=0)

The technical documentation of the products of a company is distributed through a system that consists of a web server (with a service time $S_{WS} = 300ms$), a DB (service time $S_{DB} = 500ms$), and two storages: ST1 (service time $S_{ST1} = 900ms$) and ST2 (service time $S_{ST2} = 750ms$). The number of customers is constant N = 100 users and the think time is Z = 50s.

Each request is always served by the web server WS. Then, it is directed to the DB for the 40% of the times, to the storage ST1 for the 30% and to the storage ST2 for the 30% of the time. The measured system throughput is X = 1.9 req/s.

Compute:
(a) The service demands of the four stations
(b) The utilizations of the four stations
(c) The system response time of each request
(d) Which component is the bottleneck? Which is the minimum response time that can be obtained with N = 150 users (which is the value of N^*)?