COMPUTING INFRASTRUCTURES

EXERCISES ON BOUNDS ON PERFORMANCE

Stefano Cereda

stefano.cereda@polimi.it

20/05/2019

Politecnico di Milano

RECAP - UNBALANCED SYSTEMS

1. Calculate $D = \sum_{k=1}^{K} D_k$ and $D_{max} = \max_k D_k$
2. Calculate the intersection point $N^* = \frac{D+Z}{D_{max}}$
3. Compute bounds for throughput and response time

batch	$\frac{1}{D} \leq X(N) \leq \min\left(\frac{N}{D}, \frac{1}{D_{max}}\right)$
terminal	$\frac{N}{ND+Z} \leq X(N) \leq \min\left(\frac{N}{D+Z}, \frac{1}{D_{max}}\right)$
transaction	$X(\lambda) \leq \frac{1}{D_{max}}$

batch	$\max\left(D, ND_{max}\right) \leq R(N) \leq ND$
terminal	$\max\left(D, ND_{max} - Z\right) \leq R(N) \leq ND$
transaction	$D \leq R(\lambda)$

Tighter bounds hold for balanced systems.

EXERCISE 1

An intranet is composed of 5 web servers used in parallel, 3 application servers used in parallel, and 1 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 servers 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.

Derive bounds on throughput and response time assuming $Z = 0$

SOLUTION

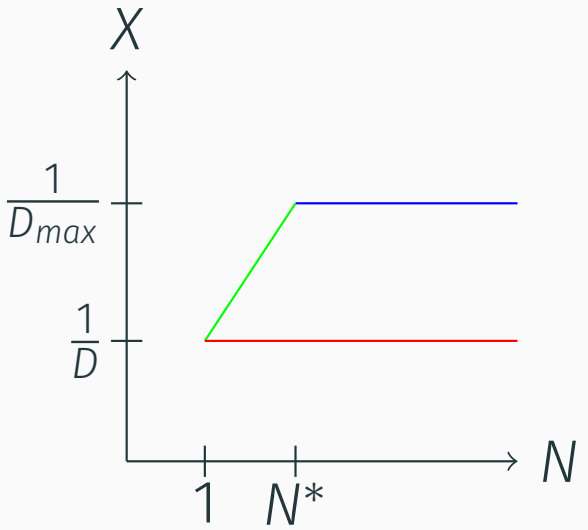
$D_{ws} = \frac{750ms}{5} = 150ms\ D_{as} = \frac{600ms}{3} = 200ms\ D_{ss} = 300ms$

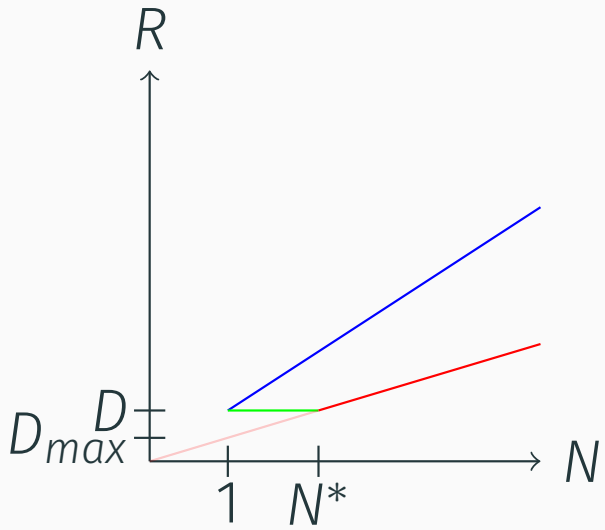
$X_{max} = \frac{1}{D_{max}} = \frac{1}{0.3sec} = 3.3sec$

$\frac{1}{D} \leq X(N) \leq \max(\frac{N}{D}, \frac{1}{D_{max}})\ \max(D, ND_{max}) \leq R(N) \leq ND$

$D = \sum D_i = 650ms\ D_{max} = 300ms\ N^* = \frac{D+Z}{D_{max}} = 2.16\ \frac{1}{D} = 1.54$

$\frac{1}{D_{max}} = 3.3$





Note on Exercise I:

As explained during the live exercises sessions, this solution is valid if the servers are considered to work in parallel on the request, so that their combined effort will be able to execute the requests in the corresponding fraction of time.

EXERCISE 2 I

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 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

EXERCISE 2 II

- 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)

EXERCISE 2 III

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 I

Point 1:

$V_i = \frac{C_i}{C} \rightarrow V_{ws} = 10 \quad V_{as} = 6 \quad V_{ss} = 2$

$D_i = \frac{U_i}{X} \qquad X = \frac{C}{T} = 1.5$

$U_i = \frac{B_i}{T} \rightarrow U_{ws} = 0.5 \quad U_{as} = 0.2 \quad U_{ss} = 0.25$

$D_{ws} = 0.33 \quad D_{as} = 0.13 \quad D_{ss} = 0.16$

$D_{max} = D_{ws} = 0.33$

Point 2:

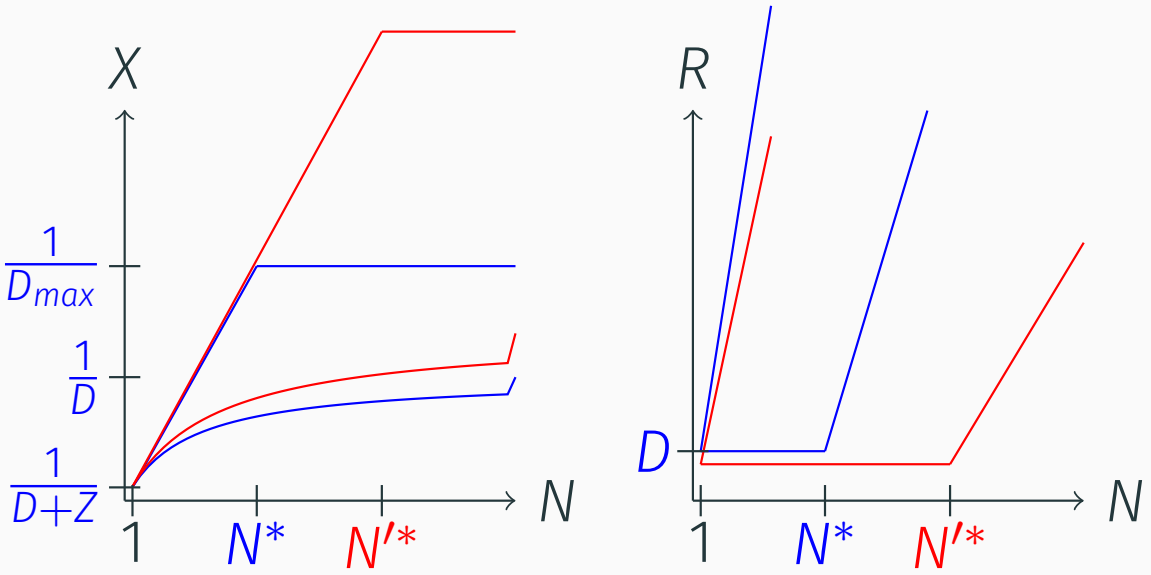
$R(50) = \frac{N}{X} - Z = 28.3sec \quad X_{max} = \frac{1}{D_{max}} = 3\frac{job}{sec}$

Point 3:

$D'_{ws} = 0.16sec \rightarrow$ new bottlenecks are ws and ss. $\quad X'_{max} = 6\frac{job}{sec}$

SOLUTION II

System	D	D_{max}	N^*	$\frac{1}{D}$	$\frac{1}{D+Z}$	$\frac{1}{D_{max}}$
Original	0.633	0.333	16.9	1.58	0.17	3
Modified	0.466	0.166	32.9	2.14	0.18	6



EXERCISE 3

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 $S = 300ms$, then through a database server (DS), which has a service time $S = 250ms$, 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 $S = 10ms$.

1. Compute the maximum throughput of the system.
2. Is it possible to have a Response Time $R < 9s$? At which conditions?

$D_i = S_i V_i \rightarrow D_{as} = 600ms \quad D_{ds} = 250ms \quad D_{fw} = 20ms$

Assuming a closed system with $Z = 0$ we need:
$\mathbf{max}(D, ND_{max}) \leq R(N) \rightarrow ND_{max} \leq 9 \rightarrow N \leq 15$

EXERCISE 4

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

- average think-time $z = 10s$
- average CPU service demand, $D_{cpu} = 4s$
- average disk service demand, $D_{disk} = 5s$
- 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:

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

SOLUTION I

Adding a fpu:

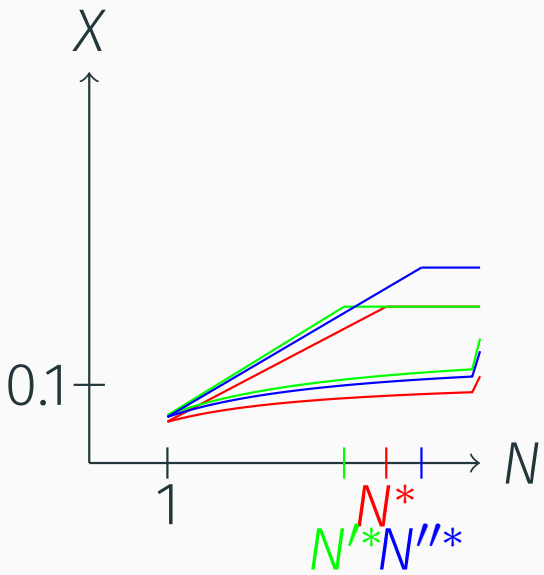
$$D'_{cpu} = \frac{D_{cpu}}{4} = 1sec \qquad D_{fpu} = \frac{\frac{3D_{cpu}}{4}}{10} = 0.3sec$$

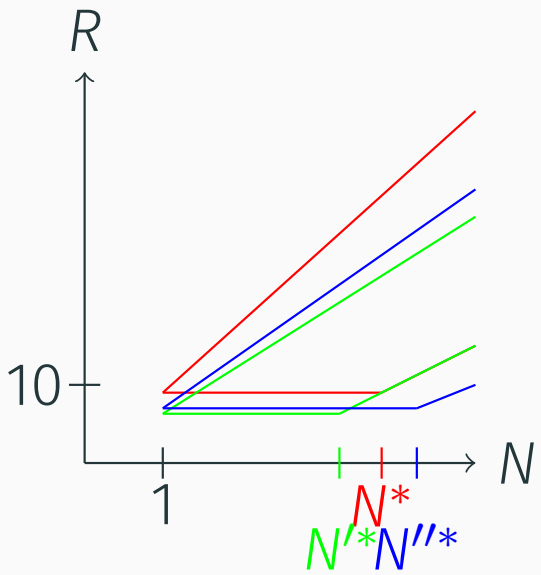
Replacing the disk:

$$V_{disk} = \frac{D_{disk}}{S_{disk}} = 200 \qquad D'_{disk} = V_{disk}S'_{disk} = 3s$$

SOLUTION II

System	D	D_{max}	N^*	$\frac{1}{D}$	$\frac{1}{D+Z}$	$\frac{1}{D_{max}}$
Original	9	5	3.8	0.111	0.053	0.2
FPU	6.3	5	3.26	0.159	0.061	0.2
Disk	7	4	4.25	0.143	0.059	0.25





EXERCISE 5 I

Consider a web application deployment structured as follows:

- one web server (WS) that renders dynamic pages,
- one application server (AS) that constructs such dynamic pages,
- one database server (DB) that holds the data needed by the application server.

EXERCISE 5 II

The web server is the most demanded center, it has indeed $D_{WS} = 5ms$. The application server and the database server have a service demand of, respectively, $D_{AS} = 4ms$ and $D_{DB} = 3ms$.

The we application is tested in a closed deployment with a terminal workload of 20 customers with a think time of 7 milliseconds.

1. What is the primary bottleneck center?
2. What is the secondary bottleneck center?
3. What is the minimum response time that can be achieved?

EXERCISE 5 III

4. What is the effect of replacing the application server with a new one that is twice as fast?
5. What is the modification, if any, that would allow the application server to run at its maximum speed?

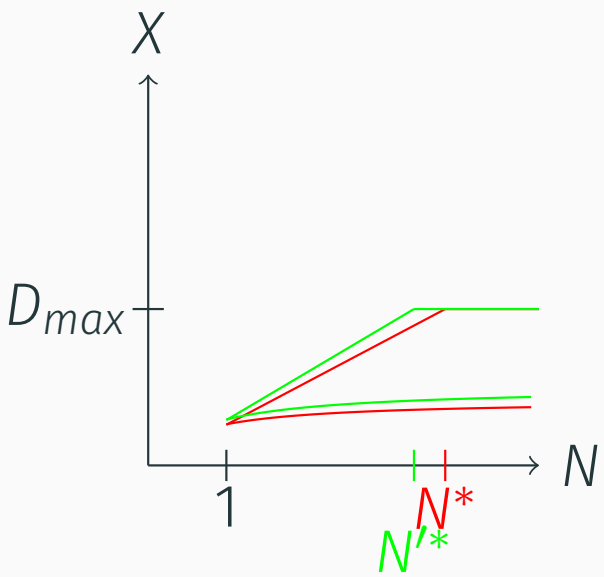
SOLUTIONS I

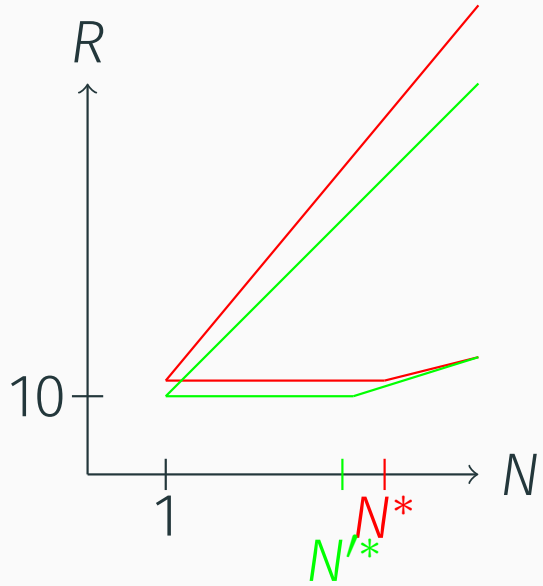
$$D_{max} = D_{ws} \rightarrow X_{max} = \frac{1}{D_{max}} = 200r/s$$

$$D'_{max} = D_{as} \rightarrow X'_{max} = \frac{1}{D'_{max}} = 250r/s$$

$$R_{min} = D = 12ms \quad (R \geq \mathbf{max}(D, ND_{max} - Z))$$

$D'_{as} = 2ms$ does not modify X_{max} as the AS is not the bottleneck. However, D changes and therefore there is a (small) difference in the graphs:





EXERCISE 6 I

A website is deployed onto a machine that runs one instance of the Nginx3 web server and two parallel instances of the PostgreSQL4 database server (each on a separate disk).

In our closed, terminal workload environment, we simulated the activity of N clients with 15 seconds of think time and we measured the busy time and the number of jobs completed at each center (except for the web server) for a time interval of 900 seconds. During this time interval the system rendered 200 HTTP responses back to the clients.

The measurements were as follows:

- the web server has a busy time of 400 seconds,

EXERCISE 6 II

- the first database server has a busy time of 100 seconds, and completed 2000 operations,
- the second database server has a busy time of 600 seconds, and completed 20000 operations.

1. Can you spot the bottleneck?
2. How many visits, to each database server, are required per HTTP request?
3. How much service time does a visit take on each database server?
4. What is the effect of replacing the Nginx web server with the Cherokee5 web server, which is twice as fast?

EXERCISE 6 III

5. The database administrator claims that the load between the two database servers is perfectly balanced. Is he correct? Why?
6. If you are given enough money to buy a new disk to run an additional database server, how would you modify the system?
7. What would be a non-obtrusive modification (i.e., involving no costs at all) to alleviate the bottleneck?

1. $D_i = \frac{B_i}{C} \rightarrow D_{ws} = 2sec \quad D_{db1} = 0.5sec \quad D_{db2} = 3sec$
   $D_{max} = D_{db2} = 3sec \rightarrow X_{max} = 0.3j/sec$
2. $V_i = \frac{C_i}{C} \rightarrow V_{db1} = 10 \quad V_{db2} = 100$
3. $S_i = \frac{D_i}{V_i} \rightarrow S_{db1} = 50ms \quad S_{db2} = 30ms$
4. $D'_{ws} = 1sec$ but D_{max} does not change.
5. The second database has 6 times the demand of the first one, so they are by no means balanced.
6. We could split the load of the second database
7. We want $D'_{db1} = D'_{db2} = d$ and
   $V_{db1} + V_{db2} = V'_{db1} + v'_{db2} = 110 \rightarrow V'_{db1}\frac{S_{db1}}{S_{db1}} + V'_{db2}\frac{S_{db2}}{S_{db2}} = 110$
   $\rightarrow \frac{D'_{db1}}{S_{db1}} + \frac{D'_{db2}}{S_{db2}} = 110 \rightarrow d(\frac{1}{S_{db1}} + \frac{1}{S_{db2}}) = 110 \rightarrow d = 2.07sec$
   and $X_{max} = 0.48j/sec$