Student ID (*Matricola*)

0	0	0	0	0	0
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	4	4	4	4	4
5	5	5	5	5	5
6	6	6	6	6	6
7	7	7	7	7	7
8	8	8	8	8	8
9	9	9	9	9	9

Computing Infrastructures

Course 095897

M. Gribaudo, R.Mirandola

25-06-2016

Last Name / Cognome:

....

First Name / Nome:

....

Answers must be given exclusively on the answer sheet (last sheet): DO NOT FILL ANY BOX IN THIS SHEET

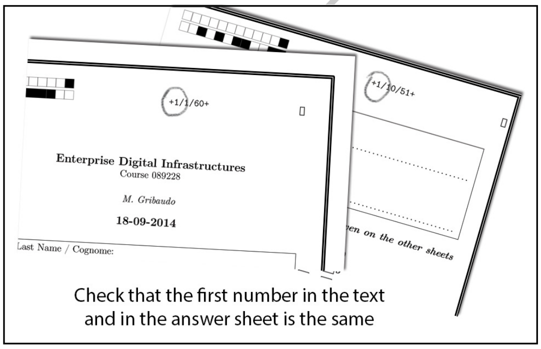
Students must use pen (black or blue) to mark answers (no pencil).
Students are permitted to use a non-programmable calculator.

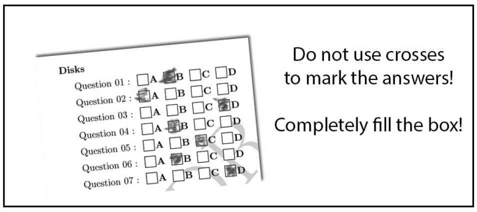
Students are NOT permitted to copy anyone else's answers, pass notes amongst themselves, or engage in other forms of misconduct at any time during the exam.

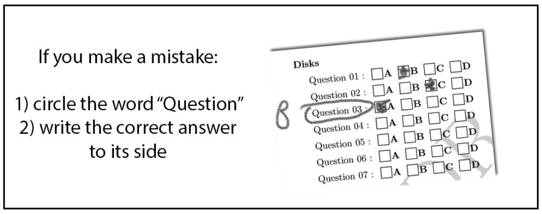
Students are NOT permitted to use mobile phones and similar connected devices.

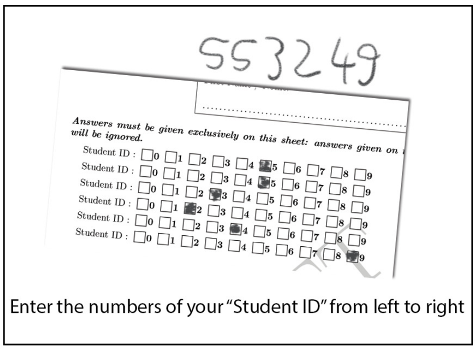
Scores for the multiple-choice part: correct answers +1 point, unanswered questions 0 points, wrong answers -0.333 points.

You cannot keep a copy of the exam when you leave the room.









Disks

A HDD has a rotation speed of 10000 RPM, an average seek time of 4 ms, a negligible controller overhead and transfer time of 256 MB/s. Files are stored into blocks whose size is 4 KB.

Question 1 The rotational latency of the disk is:

[A] 0.015259 ms [B] 3 ms [C] 6 ms [D] 0.038147

SOLUTION:
The rotational latency is half of the time required to perform one rotation:
$T_l = \frac{60000}{2 \cdot 10000} = 3$ ms.

Question 2 The time required to read a 400 KB file divided into 5 sets of contiguous blocks is:

[A] 36.526 ms [B] 51.526 ms [C] 701.53 ms [D] 1001.5 ms

SOLUTION:
Since the file is divided into 5 sets of contiguous blocks, latency and seek times are spent 5 times. The total average transfer time is thus:
$T_a = 5 \cdot (T_l + T_s) + F/r_t = 5 \cdot (3 + 4) + 400/(256 * 1024) * 1000 = 36.526$ ms

Question 3 The time required to read a 400 KB file with a locality of 95% is:

[A] 51.526 ms [B] 36.526 ms [C] 951.53 ms [D] 666.526 ms

SOLUTION:
The file is composed by $400/4 = 100$ blocks. Then we have:
$T_a = 100 \cdot ((1 - l) \cdot (T_l + T_s) + B/r_t) = 5 \cdot ((1 - 0.95) \cdot (3 + 4) + 40/(256 * 1024) * 1000) = 36.526$ ms

Question 4 The answers to questions two and three are:

[A] Identical, because the average transfer time for a file depends only on its length.

[B] Different, because the number of blocks equal to 5 does not corresponds to a 95% locality for a 400 KB file.

[C] Due to the write amplification factor, they could either be identical or different.

[D] Identical, because the number of contiguous set of blocks equal to 5 corresponds to a 95% locality for a 400 KB file.

SOLUTION:
Identical, because the number of contiguous set of blocks equal to 5 corresponds to a 95% locality for a 400 KB file and 4 KB blocks. The Write Amplification Factor is a characteristic of SSD, not of HDD, so this answer was misleading.

Virtualization and IaaS

Question 5 The paravirtualization is:

- [A] A type of virtualization that at loading time, converts the instructions of a Guest O.S. to correctly work under a Virtual Machine Monitor.
- [B] A type of virtualization that requires a modified Guest O.S.
- [C] It is the type of virtualization done at the application level, such as in the Java Virtual Machine.
- [D] A type of virtualization that exploits special instructions of a CPU to support virtual machines.

SOLUTION:
A type of virtualization that requires a modified Guest O.S.

Question 6 Which of the following answer **is not** a typical component of a modern data-center:

[A] Solar panels [B] Racks [C] Cooling system [D] Diesel generators

SOLUTION:
Racks are where IT units are stored. Cooling system is necessary since servers produce a lot of heat. Diesel generators are required to support the system in case of long energy outages. Although solar panels can be a good enhancement for a data-center, they are not strictly required for its operation: thus they cannot be considered as a typical component of a modern data-center.

An application run in a virtual environment requires 101.9 sec, is characterized by fraction of privileged instructions equal to 0.1% and an execution overhead of 1900%.

Question 7 The run time of the same application in a physical environment is:

[A] 295.51 s [B] 100 s [C] 103.84 s [D] 35.138 s

SOLUTION:
$T_p = \frac{T_v}{1+p_p \cdot o_p} = \frac{101.9}{1+0.001 \cdot 19} = 100$ s.

Question 8 The same application is run in an environment where the overhead is reduced to 800%. The execution time in this case will be:

[A] 104.67 s [B] 102.72 s [C] None of the other answers [D] 100.8 s

SOLUTION:
$T_v = T_v \cdot (1 + p_p \cdot o_p) = \frac{100}{1+0.001 \cdot 8} = 100.8$ s.

Big Data - (4 points)

The following Apache Spark code processes tweets with the aim to understand if they are positive or not

IMPORTANT NOTES:

Comments are as in Java. *Use them to understand what the content of an RDD or the outcome of an instruction should be.*

```

case class Tweet(Num:Int,
Date: String,
Time: String,
Text: String)

case class ClassifiedTweet(Num:Int,
Sentiment: String)

/* Analyse a text and detect if it is positive negative or neutral */

def sentiment(s:String) : String = {
val positive = Array("like", "love", "good", "great", "happy",
"cool", "amazing")
val negative = Array("hate", "bad", "stupid")

var st = 0;

val words = s.split(" ")

positive.foreach(p =>
words.foreach(w =>
if(p==w) st = st+1
)
)

/* Suggestion: numNeg can be determined as the number of words
contained in the
negative array */
val numNeg= FILL IN

st=st-numNeg;

if(st>0)
"positive"
else if(st<0)
"negative"
else
"neutral"
}

val tweet1 = Tweet(1, "22/06/2016","08:00:00","I love the new phone
by YYYY")
val tweet2 = Tweet(2, "22/06/2016","08:10:00","The new camera by ZZZZ
is amazing")
val tweet3 = Tweet(3, "23/06/2016","08:30:00","I heard about the
strike but it is unbelivable we don’t move for more than one hour. I
hate traffic jams")

val tweetsRDD=sc.parallelize(List(tweet1,tweet2,tweet3))

val classifiedTweets= FILL IN
```

```
53
54
55 classifiedTweets.collect
56
57 /* Array[ClassifiedTweet] = Array(ClassifiedTweet(1,positive),
58 ClassifiedTweet(2,positive), ClassifiedTweet(3,negative)) */
59
60 val t1=tweetsRDD.map(t => (t.Num,
61 t.Date)).join(classifiedTweets.map(t => (t.Num, t.Sentiment)))
62
63 t1.collect
64
65 /* FILL */
66
67 case class ClassifiedTweetDay(Num:Int,
68 Date: String,
69 Sentiment: String)
70
71
72 val t2= t1.map( {case (num, (date, sentiment)) =>
73 ClassifiedTweetDay(num, date, sentiment)})
74
75
76 t2.collect
77
78 /* res714: Array[ClassifiedTweetDay] =
79 Array(ClassifiedTweetDay(2,22/06/2016,positive),
80 ClassifiedTweetDay(1,22/06/2016,positive),
81 ClassifiedTweetDay(3,23/06/2016,negative)) */
82
83 /* Determine the number of positive tweets on 23/06/2016 */
84 val dayOfInterestPositive= FILL IN
```


Question 9 Complete line 30.

- [A] val numNeg=0
  negative.foreach(p=>
  words.foreach(w=>
  if(p==w) numNeg = numNeg+1
)
- [B] val numNeg=words.filter(w => positive contains w).length
- [C] val numNeg=words.filter(w => negative contains w).length
- [D] val numNeg=words.map(w => negative contains w).length

SOLUTION:
val numNeg=words.filter(w => negative contains w).length
Note that in the other "iterative version" numNeg cannot be incremented

Question 10 Complete line 52.

- [A] val classifiedTweets=tweetsRDD.map(t => ClassifiedTweet(t.Num,sentiment(t.Text)))
- [B] val classifiedTweets=tweetsRDD.map(ClassifiedTweet(Num,sentiment(Text)))
- [C] val classifiedTweets=tweetsRDD.reduce(t => ClassifiedTweet(t.Num,sentiment(t.Text)))
- [D] val classifiedTweets=tweetsRDD.map(t => (t.Num,sentiment(t.Text)))

SOLUTION:
val classifiedTweets=tweetsRDD.map(t => ClassifiedTweet(t.Num,sentiment(t.Text)))

Question 11 Select the output compliant with line 63.

- [A] Array[(Int, String, String)] = Array((2,22/06/2016,positive), (1,22/06/2016,positive), (3,23/06/2016,negative))
- [B] Array[(Int, (String, String))] = Array((2,(22/06/2016,positive)), (1,(22/06/2016,positive)), (3,(23/06/2016,negative)))
- [C] Array[(Int, String, String)] = Array((2,22/06/2016,positive), (1,22/06/2016,negative), (3,23/06/2016,negative))
- [D] Array[(Int, String, String)] = Array((2,22/06/2016,positive), (1,22/06/2016,positive), (3,23/06/2016,positive))

SOLUTION:
Array[(Int, (String, String))] = Array((2,(22/06/2016,positive)), (1,(22/06/2016,positive)), (3,(23/06/2016,negative)))

Question 12 Complete line 84.

- [A] val dayOfInterestPositive=t2.filter(c => c.Date == "22/06/2016").filter(c => c.Sentiment == "positive").count
- [B] val dayOfInterestPositive=t2.filter(Date == "22/06/2016").count
- [C] val dayOfInterestPositive=t2.filter(c => c.Date == "22/06/2016" || c.Sentiment == "positive").count
- [D] val dayOfInterestPositive=t2.filter(Date == "22/06/2016").filter(Sentiment == "positive").count

SOLUTION:
val dayOfInterestPositive=t2.filter(c => c.Date == "22/06/2016").filter(c => c.Sentiment == "positive").count

Big Data and PaaS - (4 points)

Compare Pig and HIVE. Are they alternative technologies or might it be useful to integrate them in business intelligence pipeline?

SOLUTION:

See slides 261-263 and 156

Performance - (3 points)

Describe the main benefits of the simulation.
SOLUTION:

See slides 10-11 L07 Simulation

Performance - (7 points)

Let us consider a computing Infrastructure composed by servers A, B, C and D and that can be accessed by a large number of users. The execution of a single request must pass through: server A, which has a service time $S_A = 300$ ms, then through server B, which has a service time $S_B = 250$ ms. Then it directed to server C (with service time $S_C = 500$ ms) for the 40% of the times and to server D (with service time $S_D = 400$ ms) for the 60% of the times and then back to server A before leaving the system.

1. Define the system model
2. Compute:
   (a) The visit numbers for servers A,B,C and D
   (b) The demands of servers A,B,C and D
   (c) The maximum throughput of the system
   (d) To allow the possibility of a maximum throughput $X_{max} = 4job/sec$ which kind of modification should we implement in the original system?
   (e) In this modified system is it possible to have a Response Time $R < 5$ s? At which conditions?

SOLUTION:

1) We can use an open model with four stations with the following characteristics:
$S_A = 0.3$sec
$S_B = 0.25$sec
$S_C = 0.5$sec
$S_D = 0.4$sec
2)

a) From the description of the job flow we can derive the visit numbers as: $v_A = 2$, $v_B = 1$, $v_C = 0.4$, $v_C = 0.6$
b) Applying the definition of demand $D_k = S_k \cdot v_k$, we have: $D_A = 0.3 \cdot 2 = 0.6$ sec, $D_B = 0.25 \cdot 1 = 0.25$ sec, $D_C = 0.5 \cdot 0.4 = 0.2$ sec., $D_D = 0.4 \cdot 0.6 = 0.24$ sec.
c) The bottleneck of the system is server A since it has the highest demand, so we have $D_{max} = 0.6$sec. The maximum throughput can be obtained as: $X_{\max} = \frac{1}{D_{\max}} = 1.66job/s$
d) We know that $X_{\max} = \frac{1}{D_{\max}}$, so if we have $X_{\max} = 4j/s$ this implies that $\frac{1}{D_{\max}} = 4job/s$ so $D_{\max} = \frac{1}{4} = 0.25$s. So to be able to obtain this maximum throughput the system should be modified and server A substituted with a new server A_1 with $D_{A_1} = 0.25$s
e) The model is open the only thing that can be said is that $R > D$, so $R > 0.94$ sec, so nothing can be said about the upper bound. We can model the system with a closed model with Z=0. In this case the bounds for closed models can be applied. To guarantee the possibility that R could be lesser than 5 sec, we can work on the lower bound and guarantee that $max(D, NDmax - Z) < R(N) < 5$ s. So N*250<5000, the possibility is given for N< 5000/250=20, so for N<20.

Dependability - (6 points)

In the following questions we will assume that both failure and repair events follow exponential distributions.

We have a two-component system in series. The failure rates of both components is the same: $\lambda_A = \lambda_B$.

1. Calculate the maximum possible value of the failure rate of each component (λ_A and λ_B) to have a system whose reliability at time $t = 20$ days is, at least 0.9.
2. Calculate the MTTF of the two-component system using the failure rates values previously calculated.
3. If we decide to use components whose failure rate is 0.001 days^{-1}, calculate how many of them we can put in series while the system still keeps a reliability at time $t = 20$ higher than 0.9.
4. In the two-component system in series, having $\lambda_A = \lambda_B = 0.001$ days^{-1} and $MTTR_A = 10$ days and $MTTR_B = 15$ days; calculate the availability of the system.
5. In the two-component system in **parallel**, having $\lambda_A = \lambda_B = 0.001$ days^{-1} and $MTTR_A = 10$ days and $MTTR_B = 15$ days; calculate the availability of the system.
6. Calculate the MTTF of the system in question 5)

SOLUTION:

1) $R(20) = e^{-\lambda_A 20} e^{-\lambda_B 20} = 0.9$. Since $\lambda_A = \lambda_B$, then $e^{-\lambda_A 20} e^{-\lambda_A 20} = 0.9 \implies e^{-\lambda_A 40} = 0.9 \implies \frac{ln(0.9)}{-40} = \lambda_A = \lambda_B = 0.002634$ days^{-1}
2) $MTTF = \frac{1}{\lambda_A + \lambda_B} = \frac{1}{0.002634 + 0.002634} = 189.82$ days.
3) $R(20) = 0.9 = (e^{-0.001 \cdot 20})^n \implies 0.9 = e^{-0.001 \cdot 20 n} \implies -0.001 \cdot 20n = ln(0.9) \implies n = \frac{ln(0.9)}{-0.001 \cdot 20} = 5.26$. Then, the maximum number of components to put in series is 5.
4) $MTTF_A = MTTF_B = \frac{1}{\lambda_A} = \frac{1}{\lambda_B} = 1000$, $Availability = \frac{1000}{1000+10} \cdot \frac{1000}{1000+15} = 0.97547$
5) $MTTF_A = MTTF_B = \frac{1}{\lambda_A} = \frac{1}{\lambda_B} = 1000$, $Availability = 1 - (1 - \frac{1000}{1010})(1 - \frac{1000}{1015}) = 0,999853$
6) $Availability = \frac{MTTF}{MTTF + MTTR}$, Availability=0,999853 (calculated in 5)). $MTTR = \frac{1}{SystemRepairRate} = \frac{1}{\frac{1}{10} + \frac{1}{15}} = 6$ days. Then, $0,999853 = \frac{MTTF}{MTTF+6} \implies MTTF = 41000$days

Answer sheet:

Last Name / Cognome:

....

First Name / Nome:

....

Answers of the multiple-choice part of the exam must be given exclusively on this sheet

Student ID : ☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9
Student ID : ☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9
Student ID : ☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9
Student ID : ☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9
Student ID : ☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9
Student ID : ☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9

Disks

Question 01 : ☐A ☐B ☐C ☐D
Question 02 : ☐A ☐B ☐C ☐D
Question 03 : ☐A ☐B ☐C ☐D
Question 04 : ☐A ☐B ☐C ☐D

Virtualization and Iaas

Question 05 : ☐A ☐B ☐C ☐D
Question 06 : ☐A ☐B ☐C ☐D
Question 07 : ☐A ☐B ☐C ☐D
Question 08 : ☐A ☐B ☐C ☐D

Spark

Question 09 : ☐A ☐B ☐C ☐D
Question 10 : ☐A ☐B ☐C ☐D
Question 11 : ☐A ☐B ☐C ☐D
Question 12 : ☐A ☐B ☐C ☐D