Identification and Rules:

In this question you must insert your given exam code. You exam is not valid without the exame code.

The exam has a time limit of 80 minutes. You can only leave after submiting your exam and only after 45 minutes. We take the best grade of the two exams.

There will be no further information about the exam questions (*dúvidas acerca do enunciado*) provided during the exam, besides the question text itself.

Answer the questions given the context in the question.

There are 24 actual questions in total in the exam, all graded with the same weight except the Map-Reduce problem that has a weight twice a normal (all others) question (1 mark -> 0.8 in 1-20 scale).

E.g. for a final grade of 8/20 ("valores") you need to sum up 10 marks, for a final grade of 15/20 ("valores") you need to sum up 18.75 marks. For a grade of 20/20 ("valores"), you need to sum up the 25 complete marks.

(Only if strictly needed, you can also insert here any other notes regarding the answering of some question(s) in the exam clearly identifying them in your comments).

Good work.

LV.

```
CNV-23-24: sample exam with questions from previous exams, addressing the topics covered
this year, and showing typical question formats. Some printing artifacts result from moodle
extraction. Specific topics, mix of topics, and question format about each topic may vary.
Good work. LV.
```

Note:
- Questions are presented in randomized order. Make sure you visit all of them. they cover different difficulty levels.
- In all questions where wrong choice may lead to penalty there is an option (no response/no penalty).
- When there is only one correct choice, incorrect choice leads to penalty (1/number-of-incorrect-or-not-fully-correct-choices).
- In questions where there are multiple correct and incorrect options, the correct and incorrect options are graded in a way that selecting all of them leads to zero score.

Question 2

Not answered

Marked out of 1.00

In the context of Cloud Computing, identify the available deployment models for a cloud.

(select all the correct options).

Select one or more:

- a. **No response (no penalty).**
- b. Software cloud.
- c. Managed cloud.
- d. Hybrid cloud.
- e. Public cloud.

Your answer is incorrect.

The correct answers are: Public cloud., Hybrid cloud.

Question 3

Not answered

Marked out of 1.00

Considering emulation by interpretation with decode-and-dispatch (D&D), there are a number of specific features to D&D that hinder its performance when compared to other interpretation-based approaches.

What are the causes for this inefficiency?

(select all the correct options).

Select one or more:

- a. Excessive number of accesses to register context blocks.
- b. **No response (no penalty).**
- c. Excessive number of branches/jumps in the emulator code.
- d. High memory cost.
- e. Low emulation speed in steady-state.

Your answer is incorrect.

The correct answers are: Excessive number of branches/jumps in the emulator code., Low emulation speed in steady-state.

Question 4

Not answered

Marked out of 1.00

In processor virtualization (System VMs), Theorem 1 addresses ( ) VMM construction.

From that theorem we can conclude the following:

- the set of privileged instructions for that computer needs to be the set of sensitive instructions.

- there can be any number of instructions.

- there must not be any instructions.

(no penalty for wrong answers).

Your answer is incorrect.

The correct answer is:
In processor virtualization (System VMs), Theorem 1 addresses ([efficient]) VMM construction.

From that theorem we can conclude the following:

- the set of privileged instructions for that computer needs to be [equal to or a superset of] the set of sensitive instructions.

- there can be any number of [innocuous] instructions.

- there must not be any [critical] instructions.

(no penalty for wrong answers).

Question 5

Not answered

Marked out of 1.00

What mappings of page addresses are employed in Memory Virtualization carried out by system VMs?

(select all the correct options).

Select one or more:

- a. user addresses to OS addresses.
- b. real addresses to physical addresses.
- c. guest addresses to virtual addresses.
- d. **No response (no penalty).**
- e. virtual addresses to real addresses.

Your answer is incorrect.

The correct answers are: virtual addresses to real addresses., real addresses to physical addresses.

Question 6

Not answered

Marked out of 1.00

When deploying instances in Amazon Web Services (AWS), the user can make selections based on geographical and fault-tolerance criteria.

What are they?

(select the option that best fits the question).

- a. Regions and Availability Zones.
- b. **No response (no penalty if only option answered).**
- c. Domains and Regions.
- d. Types and Availability Zones.
- e. Areas and Locations.

Your answer is incorrect.

The correct answer is:
Regions and Availability Zones.

Question 7

Not answered

Marked out of 1.00

When deploying workloads in the cloud, there is a sharing of responsibilities between the user and the cloud service provider (CSP), in the context of a service-level agreement (SLA).

What is the correct assessment, when a given problem arises, considering the sharing of responsibilities is used to determine who (user, CSP) should correct the situation?

(select the option that best fits the question).

a. In IaaS and PaaS, when there is networking overload, the CSP is responsible for correcting the situation.

b. In IaaS and PaaS, when there is a security problem in the guest OS, the CSP is responsible for correcting the situation.

c. In IaaS and SaaS, when there is a large scale bug in application software, the CSP is responsible for correcting the situation.

d. In PaaS and FaaS, when there is a problem with the execution environment, the user is responsible for correcting the situation.

e. **No response (no penalty if only option answered).**

Your answer is incorrect.

The correct answer is:
In IaaS and PaaS, when there is networking overload, the CSP is responsible for correcting the situation.

Question 8

Not answered

Marked out of 1.00

Nova Scheduler is the component in OpenStack that is responsible for launching instances (virtual servers) and deciding in what physical machines (hosts, nodes) they should be launched.

The built-in scheduler is the Filter Scheduler. It is parametrizable and operates in phases: i) gathering state of all compute nodes; ii) finding suitable hosts; iii) choosing the best nodes.

Consider a scenario where the developer wants to launch a VM instance that requires:

- 4 VCPUs and

- 4 GB of memory

Furthermore, he/she wants to:

- employ a balancing policy (i.e. attempt at making the new VM instance to run on one of the least loaded nodes, so as to balance the load overall).

- give priority to free memory over available CPU when selecting hosts.

How can the custom filters and weights of this algorithm be parametrized/configured in order to achieve these goals ?

(select the option that best fits the question).

a. **No response (no penalty).**

b. Use filters (vcpus_total >= 4; free_ram_mb >= 4096), use weights (RAMWeigher = -2; CPUWeigher= -1).

c. Use filters (vcpus_total >= 4; free_ram_mb >= 8192), use weights (RAMWeigher = 1; CPUWeigher= 1).

d. Use filters (vcpus_total >= 4; free_ram_mb >= 4096), use weights (RAMWeigher = 1; CPUWeigher= -2).

e. Use filters (vcpus_total >= 4; free_ram_mb >= 4096), use weights (RAMWeigher = 2; CPUWeigher= 1).

Your answer is incorrect.

The correct answer is: Use filters (vcpus_total >= 4; free_ram_mb >= 4096), use weights (RAMWeigher = 2; CPUWeigher= 1).

Question 9

Not answered

Marked out of 1.00

The interaction of AWS instances with other AWS instances and with the outside world needs to be monitored and allowed only when that is intended by the application developers/administrators.

How is this carried out in AWS?

(select the option that best fits the question).

a. DNS can be set up to enforce access control to AWS instances.

b. IP rules can allow assigning fixed network addresses to AWS instances.

c. **No response (no penalty).**

d. Security groups can be added to or removed from each AWS instance.

e. Security groups can limit network traffic in and out of AWS instances.

Your answer is incorrect.

partial correct: Security groups can limit network traffic in and out of AWS instances.

(serve to limit inbound traffic)

The correct answer is: Security groups can be added to or removed from each AWS instance.

Question 10

Not answered

Marked out of 1.00

What sequence of byte-code could illustrate the invocation of an instance method (receiving two integer arguments) on the current object's context ?

(select the option that best fits the question).

a.

iload_1
iload_2
aload_0
invokevirtual #4

b.

aload_0
ldc #3
iload_1
invokevirtual #1

c. **No response (no penalty)**

d.

aload_0
dup
iload_2
iload_1
invokespecial #7

e.

iload_1
iload_2
invokevirtual #1

Your answer is incorrect.

partially correct (using relevant byte-codes but in wrong order):

iload_1

iload_2

aload_0

invokevirtual #4

The correct answer is:

aload_0
ldc #3
iload_1
invokevirtual #1

Question 11

Not answered

Marked out of 1.00

In the context of garbage collection (GC), consider the following hypothetical and simplified application execution scenario:

- first, a Java application starts execution and initially creates 10000 small objects (e.g., of class C1, smaller than 100 bytes in size each).

- secondly, the application creates 1000 very large objects (e.g., photographs inside objects of class C2, larger than 10 MB in size each). These objects are accessed frequently and the photographs are fully analyzed. No new objects are created afterwards.

- third, due to application functionality in continuous operation, some objects of class C1 end up becoming garbage, at an hypothetical rate of 1 for each second of execution.

- finally, as garbage is created, the garbage collector is triggered and executes a full collection every 100 seconds. When all C1 objects have become garbage, the application terminates.

Based on this scenario, state what is the GC algorithm that would enable the application to achieve the best resource efficiency, i.e., with a good balance between:

- shorter total total execution time, and

- smaller memory occupation.

(select the option that best fits the question).

a. Compacting GC as it improves memory locality, reducing object access time.

b. Any of them, since Mark-and-Sweep, Compacting GC and Copying GC will impose similar overheads in this specific scenario.

c. Copying GC as it has faster collection time, freeing memory faster; and improves memory locality, reducing object access time.

d. Mark-and-Sweep as it has a relatively short collection time.

e. **No response (no penalty).**

Your answer is incorrect.

partial-correct: "Compacting GC as it improves memory locality, reducing object access time. "

(Not true in this case, as it would crate excessive movement of large objects in memory, where actually no garbage of large objects is ever created.)

The correct answer is: Mark-and-Sweep as it has a relatively short collection time.

Question 12

Not answered

Marked out of 1.00

Method in-lining is a key and frequently employed optimization to achieve high-performance emulation in Java.

Consider an hypothetical and simplified application scenario, where a static method A.main() invokes two methods in sequence in its code, for a very high number of iterations:

- static method int A.f1(int, int, int, int) that is 1000 bytecodes long;

- static method void A.f2(int, int) that is 8 bytecodes long;

Assume for simplicity only that:

- pushing each argument to the operand stack requires executing one bytecode;

- invoking a method requires executing one bytecode;

- returning a value from a method requires executing one bytecode;

What is the method (A.f1() or A.f2()) that would gain more from being in-lined in the code of A.main()?

Justify your answer to value your reasoning, with a brief outline of the calculation of the gains/savings introduced by in-lining in each case.

(No penalty for wrong answer).

In these questions, your reasoning and assumptions are relevant.

Aspects considered: expression of calculation of relative and absolute gains/reduction and choice reasoning.

In general, it is better to optimize small and hot methods because the relative gains are greater (i.e., when the bytecodes necessary to push arguments, invoke method and push/pop return value represent a greater proportion of the bytecodes in the method body).

Important: that the bytecodes of the method body are always executed with in-lining or not).

In the specific case:

=== Cost of invoking method A.f1():

- current: 4 params + 1 invoke + 1 value returned + 1000 bytecode instructions in body = 4 +1 + 1 + 1000 = 1006

- inlined: 1000

Absolute reduction in bytecodes is 6.

Relative code reduction is slim (~0.6%), and speedup is 1.006 (0.6% faster)

=== Cost of invoking method A.f2():

- current: 2 params + 1 invoke + 0 value retuned + 8 bytecode instructions in body: 2 +1 + 0 + 8 = 11

- inlined: 8

Absolute reduction in bytecode is 3.

Relative code reduction is significant (3/11 -> 27.7%) and speedup is 1.375 (37.5%) faster.

=== Best option (for relative and absolute gains):

- In general, the gains of inlining A.f2() are *relatively* much much higher, and the overhead of inlining A.f2() is minimal

[this is the standard approach followed].

- The gains of inlining A.f1() specifically *just inside main*, if we only account for the *total absolute reduction of bytecodes executed in each iteration* are higher (6 vs 3), albeit this would generate significant code duplication.

[this reasoning is also considered when demonstrated with calculations].

Question 13

Not answered

Marked out of 1.00

The Amazon Simple Queue Service (SQS) manages to combine fault-tolerance (ensuring no messages are lost) with efficiency (trying to avoid repeated/redundant work carried out by applications).

What is the sequence of events that best reflects a situation where there is an application failure (during processing of a given message) that is correctly handled and recovered from?

(Drag-and-Drop the small sentences to their appropriate place)

Action 1:

Action 2:

Action 3:

Action 4:

Action 5:

The message is sent to the queue.	The message is re-read from the queue.
The message in the queue becomes temporarily invisible.	The message is deleted from the queue.
The message is read from the queue.	

Your answer is incorrect.

1 correct -> 25%

2 correct -> 50%

3 correct -> 75%

5 correct -> 100%

(it is impossible to have 4 correct and 1 incorrect in this scheme, hence the adjustment).

The correct answer is:
The Amazon Simple Queue Service (SQS) manages to combine fault-tolerance (ensuring no messages are lost) with efficiency (trying to avoid repeated/redundant work carried out by applications).

What is the sequence of events that best reflects a situation where there is an application failure (during processing of a given message) that is correctly handled and recovered from?

(Drag-and-Drop the small sentences to their appropriate place)

Action 1: [The message is sent to the queue.]

Action 2: [The message is read from the queue.]

Action 3: [The message in the queue becomes temporarily invisible.]

Action 4: [The message is re-read from the queue.]

Action 5: [The message is deleted from the queue.]

Question 14

Not answered

Marked out of 1.00

When deploying an application on the Google App Engine (GAE) PaaS, how does the GAE achieve scalability, i.e. by adhering to what specific behavior?

(select the option that best fits the question).

- a. Application state in static variables is automatically managed.
- b. Web apps are permanently kept in memory.
- c. Internal application data is managed using a LRU (least recently used) policy.
- d. **No response (no penalty)**
- e. Requests are sent to the servers where the application is already running.

Your answer is incorrect.

partial correct: "Web apps are permanently kept in memory."

(Applications are **not** permanently kept in memory, they are recycled using LRU)

The correct answer is: Requests are sent to the servers where the application is already running.

Question 15

Not answered

Marked out of 1.00

Function-as-a-Service (FaaS) aims to improve on the delivery model provided by Infrastructure-as-a-Service (IaaS).

One such example is AWS Lambda that follows a specific pricing model.

Identify the key aspects that are considered when charging the user for the actual resource utilization of a function over a given period of time.

(select all the correct options).

- a. Total number of cores reserved for the function.
- b. **No response (no penalty if only option answered).**
- c. Total Input/Output bandwidth consumed by the function.
- d. Number of function invocations.
- e. Memory reserved for the function.

Your answer is incorrect.

The correct answers are:
Number of function invocations.,

Memory reserved for the function.

Question 16

Not answered

Marked out of 1.00

Function-as-a-Service (FaaS) brings to programmers significant advantages.

Nonetheless, cold starts resulting from VM/container and runtime startup introduce significant latency penalties in short invocations.

What approaches can the providers implement to address cold starts and make them less frequent?

(select all the correct options).

- a. **No response (no penalty if only option answered).**
- b. Providers could allocate machines to exclusively execute functions.
- c. Functions could reuse already started execution environments.
- d. Functions execution could be triggered based on events triggered or new or updated data.
- e. Providers could execute concurrently several lambdas in the same runtime.

Your answer is incorrect.

The correct answers are:
Functions could reuse already started execution environments.,

Providers could execute concurrently several lambdas in the same runtime.

Question 17

Not answered

Marked out of 1.00

The architecture of the Google File System/Hadoop File System (GFS/HDFS) is tuned/optimized for specific operation scenarios.

From the following, what is the type of operation that GFS/HDFS was specially designed to support?

(select the option that best fits the question).

- a. Random/direct read access to very large files.
- b. Random/direct read access to small files.
- c. **No response (no penalty if only option answered).**
- d. Sequential read access to very large files.
- e. Random/direct write access to very large files.

Your answer is incorrect.

The correct answer is:
Sequential read access to very large files.

Question 18

Not answered

Marked out of 1.00

BigTable offers developers tabular storage with scalability and aimed for high throughput in data access.

What are the BigTable design decisions that contribute the most to making clients access the data fast, in the typical use case of scanning the rows of a (big-)table within a range/interval?

(select all the correct options).

- a. **No response (no penalty).**
- b. Tablets that become too large are partitioned.
- c. Clients contact tablet servers directly to access data.
- d. Metadata is stored in (a tree of) metadata tablets.
- e. Tablets contain contiguous sequences of entities.

Your answer is incorrect.

The correct answers are: Clients contact tablet servers directly to access data., Tablets contain contiguous sequences of entities.

Question 19

Not answered

Marked out of 1.00

Considering the cloud storage Dynamo from Amazon, identify a specific key design decision, adopted by Dynamo creators.

(select the option that best fits the question).

- a. Availability is ensured using vector-clocks with sizes adapted to update rates.
- b. **No response (no penalty)**
- c. Membership information is managed using Merkle trees.
- d. Partitioning of data is decided using relaxed quorums for eventual consistency.
- e. Permanent faults are handled asynchronously with anti-entropy.

Your answer is incorrect.

The correct answer is: Permanent faults are handled asynchronously with anti-entropy.

Question 20

Not answered

Marked out of 1.00

In the context of a MapReduce deployment in Hadoop, what is the default size of a *split* (the unit of input data given to a Mapper process) ?

(select the option that best fits the question).

- a. Size of local file system blocks.
- b. **No response (no penalty)**
- c. Average size of files in the input.
- d. Size of HDFS blocks.
- e. Total input size divided by number of workers.

Your answer is incorrect.

The correct answer is: Size of HDFS blocks.

Question 21

Not answered

Marked out of 1.00

Consider a MapReduce job where:

- there are 20 Mapper processes;

- there is 1 Reducer process;

- across the whole input, the map() function will emit intermediate output to reduce() function, where the same 100 individual distinct keys occur in the intermediate output of every mapper.

There is a combine() function defined to optimize the execution of this job.

State how many times the combine() function will be invoked.

(select the option that best fits the question).

- a. 10000
- b. 100
- c. 2000
- d. 20
- e. **No response (no penalty).**

Your answer is incorrect.

The correct answer is: 2000

Question 22

Not answered

Marked out of 2.00

A supervising authority is in charge of monitoring fraud in stock markets (where shares, participation in the capital of companies, are publicly traded).

For instance, the supervising authority wants to determine the companies where there is suspicion of fraudulent trading, e.g. with undue privileged information.

In general, the authority wants to detect companies that have little presence in the news but end up having a very high volume of trade, which is regarded as strange.

Therefore, the trade of shares of a company is considered suspect if, on the data of a given period that is processed:

- the number of shares traded is higher than 1000 times the number of times it was viewed by in news pages, discarding very small operations (i.e., where less than 10 shares are traded).

The input consists of two text files.

In the first one, trades.txt, each line stores one trade operation regarding a given company:

<tradeID>, <company_name>, <number_of_shares_traded>

In the second one, news.txt, each line stores one news page with all the companies mentioned:

<newsID>, <number_of_page_views>, <company_name_1>, <company_name_2>, ..., <company_name_N>

Assume the standard signature for the **mapper**: List<k2,v2> map(string, int, string) and **reducer**: List<k3, v3> reduce(k2, List<v2>) functions, where map has input (filename, cursor/line number, and text line content) and k2, v2, k3 and v3 can all be replaced with types of your choice.

Write the pseudocode for a MapReduce application, with a single map and a single reduce method.

The final output of the reduce function should produce, if adequate, a single record for each company that the supervising authority should investigate further.

(if you need, you can briefly state any assumptions at the end of the answer after the pseudocode).

Note: MapReduce is not natural language. Pseudocode needs not be Java but it must include clear operating steps well identified, it should not be a paragraph rephrasing the question.

Solution sketch:

Mapper:

- The first element of the solution is determining the key for the mapper and reducer.

- It should be the company field(s) in each of the files as it is the only one that allows to cross-reference the information about each company in the two files (in essence, a join between the relevant information in two tables).

(This allows the reduce function to receive and process all tuples sharing (i.e. concerning) the same company regardless of their additional information or file provenance. Only this will allow summing the number of trades of each company - what appears in trades records and the number of page views of the news that covered it – as it was included in each news page).

Mappers cannot access external information besides function input. There is no other way for the reducer to retrieve information regarding a given company besides the arguments of the reduce function receives, i.e., no access to external storage, global hash-table with info of all companies, etc.)

The mapper will emit two types of tuples supporting any range of values in each of the significant variables:

- i) once for each trade that buys/sells at least 10 shares (this filtering can and should be done in the mapper);

- ii) zero, one, or many for each news item, each time a company shows up the in a page view log.

Mapper (path i) – trades.txt database) –

- When the file argument is trades.txt: *(this means we are reading from the trades database.)*

- Mapper should output the company and shares traded and record this type of information explicitly in the tuple, so that it enables the reducer to interpret it correctly, in any circumstances, without resorting to special cases where only one of the fields can make sense.

- E.g.

```
if (file is "trades.txt") {
        if (line[2] >= 10) {
            output(line[1], TRADE, line[2]);
            }
    }
```

Mapper (path ii) – news.txt database) –

- When the file argument is news.txt: *(this means we are reading from the news database.)*

- Mapper should iterate the company names in the line/list (i.e., line[2] -> line[line.length]) and output, for each one, the company name and the number of page views of that news item. Once again, it must record explicitly the type of the tuple so that it enables the reducer to interpret it correctly:

- E.g.

```
if (file is "news.txt") {
  for (i = 2; i < line.length; i++) {
      output(line[i], NEWS, line[1]);
   }
}
```

Reducer:

- At the reducer function, the key is the company name, and the tuples carry the type information in the first field (v[0]) about TRADE or NEWS, so that they can be added together.
- the company is suspected (and should only be once) if and only if the total of shares traded is higher than 1000 times the number of new views regarding that company.

-

```
n_page_view = 0;
n_shares_traded= 0;
foreach (v in values) {
    if (v[0] == TRADE { v_shared_traded += v[1]; }
    else if {v[0] == NEWS {n_page_view += v[1]; }
}
if (n_shared_traded > 1000 * n_page_view) {
     output(key);
}
```

Note: *There is only one map and reduce function. The map can receive splits with (sets of) lines from any of the two giant files. I considered your interpretations when deciding about each file, or simply having a code path to handle "file1" and another code path to handle "file2", etc.*

Typical errors:

- *wrong key selection.*
- *incorrect filtering of data (not at the mapper).*
- *assuming shares traded and page views are present both in all tuples.*
- *accessing files or external information in the reduce function.*
- *wrong decision criteria in conditions/ifs.*
- *too generic code or English text.*

Question 23

Not answered

Marked out of 1.00

When comparing Spark against MapReduce, identify Spark's features that are specifically advantageous when compared to MapReduce?

(select all the correct options).

Select one or more:

- a. Spark has special support for the reuse (e.g., iterative, interactive) of working sets.
- b. Spark allows eliminating irrelevant input.
- c. **No response (no penalty if only option answered).**
- d. Spark allows in-memory processing.
- e. Spark provides automatic fault-tolerance.

Your answer is incorrect.

The correct answers are: Spark has special support for the reuse (e.g., iterative, interactive) of working sets., Spark allows in-memory processing.

Question 24

Not answered

Marked out of 1.00

Consider the Big-Data processing systems Spark, Spark Streaming and Flink.

Taking into account how they handle fault-tolerance and input management, identify the true analysis/assessments from the following sentences.

(select all the correct options).

Select one or more:

- a. Spark Streaming and Flink provide quick recovery to faults by avoiding reprocessing all the data already processed in current windows.
- b. Spark Streaming and Flink cannot both take into account event time when constructing windows.
- c. Spark, Spark Streaming and Flink provide sliding windows semantics.
- d. **No response (no penalty if only option answered).**
- e. Spark Streaming and Flink are able to process events as soon as they arrive.

Your answer is incorrect.

The correct answers are: Spark Streaming and Flink provide quick recovery to faults by avoiding reprocessing all the data already processed in current windows., Spark Streaming and Flink cannot both take into account event time when constructing windows.

Question 25

Not answered

Marked out of 1.00

In a cloud data center facility the total energy consumption is 5000 kW and the energy delivered to computing equipment is 4000 kW.

What is the PUE (power usage effectiveness) of this facility?

(select the option that best fits the question).

- a. 1.10
- b. **No response (no penalty if only option answered).**
- c. 0.80
- d. 1.25
- e. 0.10

The correct answer is: 1.25