CSIE52400/CSIEM0140 Distributed Systems, Spring 2020
Final Exam

ID: _____ Dept:_____ Name: _____

1. **(30%) Basic concepts**.

(a) What is a distributed system? For each of the following systems, briefly explain why (or why not) it satisfies the definition of a distributed system.

[1] World Wide Web

[2] Facebook

[3] Alpha Go

(b) What does it mean by transparency in distributed systems? How to achieve it? Full transparency may not be a good thing. Why?

(c) What is openness in distributed systems? How to achieve it? Give an example of a open distributed system.

(d) What are the main mechanisms in RPC and RMI, respectively, which enable the calling/invocation of remote procedures/methods as if they were local? Briefly describe the similarities and differences between them.

.

(e) What is the relationship between blocking/non-blocking and synchronous/asynchronous primitives? Is it possible to have a blocking asynchronous operation? What about nonblocking synchronous operation?

2. **(30%)** Answer the following TRUE/FALSE questions.

No.	Problem Description	TRUE/FALSE
(a)	In a layered architecture, a layer provides services to the layer immediate below it.	
(b)	In MapReduce system, after the reducers are executed, the shuffle & sort process is executed to sort the output of the reducers.	
(c)	In a synchronous distributed system, since the bounds on execution step, message transmission and clock drift are known, it is possible to use timeouts in system design.	
(d)	In RMI mechanism, proxy and skeleton can be automatically generated while the remote interfaces must be defined by the programmers.	
(e)	In data streaming systems, records in data streams are usually accessed on-demand.	
(f)	In the Internet protocol layers, only the physical layer is hardware-dependent. All other layers are hardware-independent.	
(g)	In a time-uncoupled indirect communication mechanism, the sender and the receiver do not need to exist at the same time but must know each other's identities.	
(h)	Concurrent threads within the same process share the same execution environment and therefore do not need to synchronize with each other.	
(i)	In a publish-subscribe system, a node cannot be both a publisher and a subscribe at the same time.	
(j)	In general, asynchronous communication allows higher degree of parallelism while synchronous communication leads to easier programming.	
(k)	A distributed system with higher degree of transparency is usually more user friendly than the case with lower degree of transparency.	
(l)	A cluster computing system is a type of distributed system normally with heterogeneous nodes running on high speed networks.	
(m)	The UDP protocol is for reliable communication while the TCP protocol is for fast and best-effort communication.	
(n)	In IoT systems, IPv6 is usually used to give each device a unique address.	
(o)	Apache Kafka is usually used in an IoT pipeline for streaming data analytics and visualization.	

3. (15%) **Operating System Support:** Answer the following questions.
   (a) What are main differences between user-level threads and kernel-level threads?
   (b) Give two circumstances where user-level thread is better than the kernel-level thread.
   (c) Give two circumstances where kernel-level thread is better than the user-level thread.
   (d) In some systems, e.g., web servers, multithreading may achieve better performance than single threading. However, thread creation incurs overhead. Describe an approach to improve the service response time by reducing the delay due to thread creation.

4. **(10%) RPC**: A client makes remote procedure calls to a server. The client takes 5 milliseconds to compute the arguments for each request, and the server takes 10 milliseconds to process each request. The local operating system processing time for each send or receive operation is 0.5 milliseconds, and the network time to transmit each request or reply message is 3 milliseconds. Marshalling or unmarshalling takes 0.5 milliseconds per message.

   Calculate the time taken by the client to generate and return from two requests (you can ignore context-switching times):

   (a) if it is single-threaded, and

   (b) if it has two threads that can make requests concurrently on a single processor.

5. **(10%)** Answer the following questions about MapReduce and Spark.

(a) In the design and execution of a MapReduce application, determine who (user, OS, MapReduce system, …) is responsible for each of the following tasks.

[1] Splitting of input among mappers.

[2] Map function

[3] Shuffle and sort

[4] Reduce function

[5] Save results to disks

(b) In a class status database for Covid-19 tracking, each record is the set of students/teachers that participate in a particular class. Design a Spark algorithm (in pseudo code or any compatible language) to find all students/teachers that must be placed on alert given a positively infected student/teacher.

6. **(10%)** In a publish-subscribe (PS) system, for each type of information below, describe the best subscription model for that type of information. If you think that a particular type of information can be best subscribed by more than one model, briefly explain why it is the case.

Desired information	Subscription model(s)
The 2018 MLB schedule and scores	
News articles about Donald Trump on Twitter	
All stocks with price > $50 and keep rising for at least two days	
All information about the CSIE department of NDHU	
Sales advertisement of iPhone X around Hualien	