Ph.D. Qualify Examination 2014
Theory of Computation

- This examination is closed books.
- Please turn off your cell phones.
- Remember that there are 2 pages (7 questions) of the qualify examination.
- Answer all questions as possible. You may have a partial score if you answer with the correct direction.

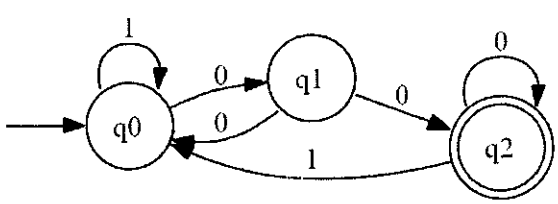
1. Deterministic Finite Acceptors (DFAs) (10 pts)
   For $\Sigma = \{a, b\}$, construct a DFA that accepts the set consisting of:
   All strings with at least one "a" and exactly two "b".

2. Nondeterministic Finite Acceptors (NFAs) (20pts. 10 pts each)
   Draw NFA to accept the following sets of strings over {0, 1}:
   (a) Draw NFA with the specified number of states to accept the following set:
   All strings that the third symbol from the right end is a "1". (use a "4 states" solution)

   (b) Draw NFA with the specified number of states to accept the following set:
   The language 0*1*0*0. (use a "3 states" solution)

3. In the derivation of using pumping lemma to prove a language L is not regular, we will give an assumption "A DFA M with the number of states $|M|$ exists for L". Can you replace the assumption with "An NFA M' with the number of states $|M'|$"? Please justify your answer. (15 pts)

4. Convert the following NFA into an equivalent DFA: (15 pts)



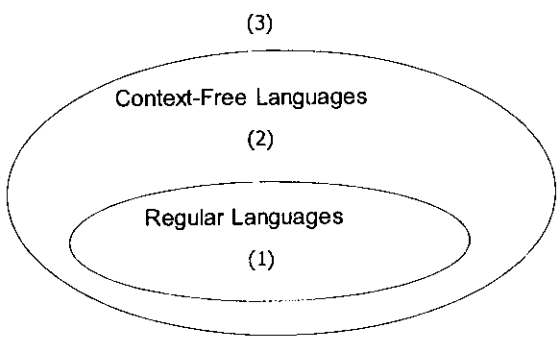
5. Show that the following grammar is ambiguous. (10 pts)
   $S \rightarrow AB|aaB$
   $A \rightarrow a|Aa$
   $B \rightarrow b$

6. Construct an NPDA that accepts the following language (use a NPDA with 5 states): (10 pts)

$L = \{a^n b^{n+m} c^m : n \geq 0, m \geq 1\}$

7. Fill the following languages into the language hierarchy (If L_i is a regular language and also a context-free language, please fill L_i in the set of regular languages): (20 pts)

$L_1 = \{a^n b^m : n \geq m\}$,
$L_2 = L(a^* b^*)$,
$L_3 = \{a^n b^n c^n : n \geq 0\}$,
$L_4 = \{a^n w w^R a^n : n \geq 0, w \in \{a, b\}^*\}$,
$L_5 = \{ab, ad, a\}$,
$L_6 = \{ww : w \in \{a, b\}^*\}$,
$L_7 = \{a^{n!} : n \geq 0\}$,
$L_8 = \{a^n b^j a^j b^n : n \geq 0, j \geq 0\}$,
$L_9 = \{a^n b^m c^{n+m} : n \geq 0, m \geq 0\}$,
$L_{10} = \{a^3 b^n c^n : n >= 0\}$.



2014 Oct. NCKU CSIE PH.D. Qualification Examination
Computer Archtecture

1. With dynamic hardware prediction for reducing branch costs, what is the disadvantage of a simple 1-bit branch-prediction buffer for a branch that is almost always taken. Explain why the 2-bit prediction scheme can remedy this disadvantage. Also, explain what is correlated predictors by illustrating an example. (20 points)

2. Explain the following synchronization primitives: atomic exchange, test-and-set, and fetch-and-increment. Also, explain what is the pair of instructions, load linked (LL) and store conditional (SC) and how this pair of instructions can be used to implement atomic exchange and fetch-and-increment. (15 points)

3. The classical approach to improving cache behavior is to reduce miss rates. Please summarize the techniques that can reduce miss rates. (20 points)

4. Describe two major instruction set characteristics that can further divide general purpose register (GPR) instruction set architecture into three classes, based on whether the instruction operands are used explicitly or implicitly. And show the advantages and disadvantages of these three further divided classes. (15 points)

5. Describe what are the RAW, WAW, and WAR hazards. (15 points)

6. For the memory-hierarchy design, please answer the following questions: (Assuming the cache is n-way set associative and there are $S = 2^s$ sets, and each block is of size $B = 2^b$ bytes) (15 points)
   A. Where can a block with variable *v* be placed in a cache? (assume v has the address *addr*)
   B. How is a block with variable v is found if it is in the cache? (describe the general tag design)
   C. Which block should be replaced on a cache miss? (make your assumption)
   D. What happens for a write operation? (describe two basic write policies)

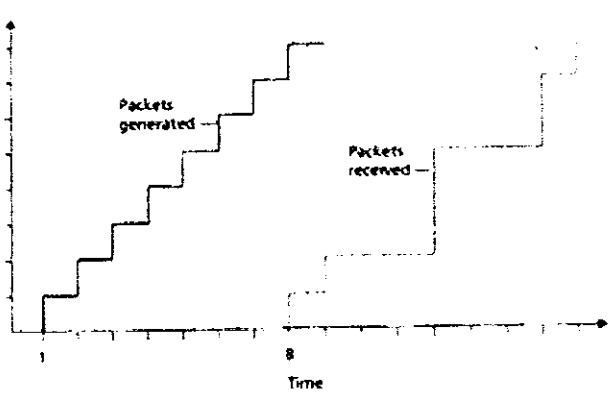
Computer Networks – Qualified Exam

(2014/Fall)

系別 _____ 年級 _____學號 _____ 姓名 _____

1. (10%)
   Let a link's distance be 2,500 km, propagation speed be 2.5 • 108 m/s, and transmission rate be 2 Mbps. How long does a packet of length 1,000 bytes take to propagate over this link? Let's derive the general form: how long does it take a packet of length L to propagate over a link of distance d, propagation speed s, and transmission rate R bps? Does this delay depend on the packet length? Does this delay depend on the transmission rate?
2. (10%)
   i) HTTP, FTP, SMTP, and POP3 run on top of TCP. Why aren't they on the top of UDP?
   ii) Let Alice provide chunks to Bob throughout a 30-second interval using BitTorrent. Will Bob necessarily return the favor and provide chunks to Alice in this same interval? Why or why not?

3. (10%)
   Consider the figure below. A sender begins sending packetized audio periodically at t = 1. The first packet arrives at the receiver at t = 8.



   i) What are the delays (from sender to receiver, ignoring any playout delays) of packets 2 through 8? Note that each vertical and horizontal line segment in the figure has a length of 1, 2, or 3 time units.
   ii)If audio playout begins as soon as the first packet arrives at the receiver at t = 8, which of the first eight packets sent will not arrive in time for playout?
   iii)If audio playout begins at t = 9, which of the first eight packets sent will not arrive in time for playout?

iv)What is the minimum playout delay at the receiver that results in all of the first eight packets arriving time for their playout?

4. (10%)
Consider a datagram network using 8-bit host addresses. Suppose a router uses longest prefix matching and has the following forwarding table:

Prefix Match	Interface
00	0
01	1
10	2
11	3

For each of the four interfaces, give the associated range of destination host addresses and the number of addresses in the range.

5. (10%)
Suppose three active nodes—nodes A, B, and C—are competing for access to a channel using slotted ALOHA. Assume each node has an infinite number of packets to send. Each node attempts to transmit in each slot with probability p. The first slot is numbered slot 1, the second slot is numbered slot 2, and so on.
   a. What is the probability that node A succeeds for the first time in slot 4?
   b. What is the probability that some node (either A, B or C) succeeds in slot 2?
   c. What is the probability that the first success occurs in slot 4?
   d. What is the efficiency of this three-node system?

6. (10%)
i) In classless addressing, can two blocks have the same prefix length? Explain.

ii) In classless addressing, we know the first address and one of the addresses in the block (not necessarily the last address). Can we find the prefix length? Explain.

7. (10%)
Using 5-bit sequence numbers, what is the maximum size of the send and receive windows for each of the following protocols?
a. Stop-and-Wait
b. Go-Back-N
c. Selective-Repeat

8. (10%)

i) In cases where reliability is not of primary importance, UDP would make a good transport protocol. Give examples of specific cases.

ii) Are both UDP and IP unreliable to the same degree? Why or why not?

9. (10%)

i)What is the maximum number of routers that can be recorded if the timestamp option has a flag value of 1? Why?

ii)The value of HLEN in an IP datagram is 7. How many option bytes are present?

10. (10%)

A client uses UDP to send data to a server. The data is 16 bytes. Calculate the efficiency of this transmission at the UDP level (ratio of useful bytes to total bytes).

2014 Fall DBMS Qualify Exam

1. (15%) 假設你要去 model 一個醫療系統，經過訪談，你得到以下的需求，請依需求，畫出其 E-R diagram：
   - 有三種資料: 醫師，病人，檢驗項目。
   - 醫師有三個 attributes: 代號，姓名，性別。其中代號是唯一的。
   - 病人有三個 attributes: 病人代號，姓名，緊急聯絡人，其中緊急聯絡人可以有多位，且必須記載緊急聯絡人之姓名與電話。其中病人代號是唯一的。
   - 檢驗項目有兩個 attributes: 項目代號，名稱。其中項目代號是唯一的。
   - 有些醫師會有一位資深醫師來作他的指導醫師。
   - 一位醫師診療一個病人時可能會要其做零或多項檢驗，一位醫師可以診療多位病人，一位病人也可能被一或多位醫師診療，此外，診療日期必須記載。

2. (40%) Given a database schema as follows.

   S(S#, Sname, Status, City) /* This is a relation for Supplier */

   P(P#, Pname, Color, Weight, City) /* This is a Part relation */

   J(J#, Jname, City) /* This is a Project relation */

   SPJ(S#, P#, J#, Quantity)

   Answer the following queries in SQL.

   (a) Get the names of the suppliers that supply at least one or more parts to each project.

   (b) Get the names of the parts that are used in more than 2 projects or the parts used in a project are white parts.

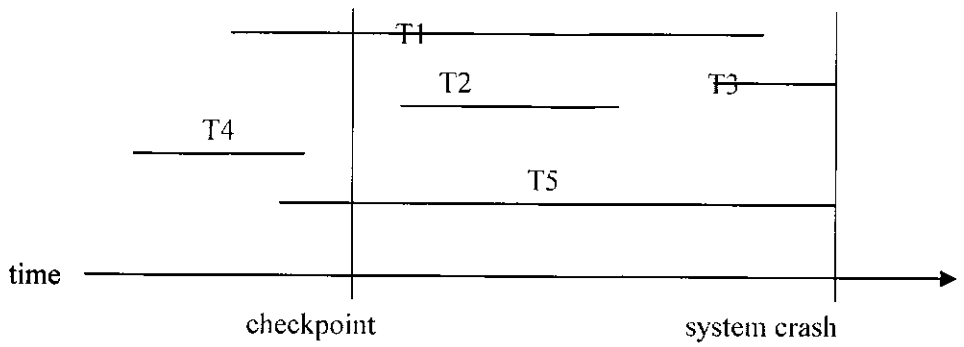
   Answer the following queries in relational algebra.

   (c) Get the names of the suppliers supplying parts that contains all the parts project P3 uses.

   (d) For each project, list the maximum weight of the parts used in this project and the number of suppliers that supply this project.

3. (10%) Describe the two-phase locking (2PL) protocol. Also explain the disadvantages of 2PL.

4. (10%) Given a time map of execution of several transactions.



(a) What do we need to do to these transactions if the deferred update protocol is used.

(b) What do we need to do to these transactions if the immediate update protocol is used.

5. (10%) Relation R(A, B, C, D, E, F, G, H) has the set of functional dependencies

$A,B \rightarrow C, D$

$D \rightarrow E, F$

$F \rightarrow G, H$

$A \rightarrow H$

(a) (5%) Normalize the relation into 2NF.

(b) (5%) Normalize the relation into 3NF.

/* Don't do unnecessary normalization! */

6. Answer the following questions.

(a) (5%) We all have seen a lot of definitions in papers and in mathematics textbooks. Now could you give a formal definition (just as those given in the literature) for the relational join operation, including giving a notation for the join operation?

(b) (10%) In performing an update operation, the record to be updated must be locked (as indicated in concurrency control protocols). Then, can a query access these locked records during query processing? If no, how can a query result be accurate with these records being locked and inaccessible? If yes, then the access to these locked items would violate the concurrency control protocol (because the lock on the records is broken). How does a DBMS manage this problem?

OS 資格考題 (103 學年度第一學期)

1. (20%) Assume that there are 4 processes, P1 to P4, in the system. The process information is shown in the table below. Assume that the system has only a single core and the context switch time can be ignored, please draw the **Gantt chart** and compute the **average waiting time** for each of the following scheduling algorithms: (a) FCFS scheduling, (b) non-preemptive SJF scheduling, (c) preemptive SJF scheduling, (d) RR scheduling (time slice = 3 ms)

Processes	Arrival Time (ms)	Burst Time (ms)
P1	1	6
P2	2	2
P3	0	7
P4	3	10

2. (15%) In a two-level paging scheme, please describe the format of a **virtual address** and the typical format of the **page table entry**.

3. (15%) What are the advantages and disadvantages between a single-level paging scheme and a multiple-level paging scheme?

4. (20%) In the **many-to-1** threading model, please discuss
   (a) why the thread creation is typically faster than the 1-to-1 model
   (b) why the process blocks when one of its threads makes a blocking system call

5. (10%) What is a **signal** in an UNIX system?

6. (20%) In a file system with block size 4 KB. Assume that an inode (index node) has 10 direct data pointers, 1 single-indirect pointer, 1 double-indirect pointer and 1 triple-indirect pointer. Each pointer is 4 bytes in size. For a 40MB file F, please describe how the inode of F points to the data blocks of F.

1 牙醫用解剖生理學

1. What are the organ systems of the human body? And name two organs of each system. (15%)

2. Describe the process of transcription. (15%)

3. Write down the Goldman Equation in a system where sodium, potassium and chloride ions are involved. Meanings of the variables involved shall be specified. (15%)

4. Describe the ossification process of a long bone. (10%)

5. Describe the events that cause an action potential. (10%)

6. Describe the nervous pathway of a Flexor Reflex (10%)

7. What cause the 2^{nd} and the 4^{th} heart sounds (10%)

8. Describe the pathway of Hepatic Portal Circulation. And give two examples to show the importance of this pathway (10%)

9. Give 5 physiological factors that can decrease insulin secretion of a normal human body. (5%)

12/2014 博士班資格考: 機率與統計 Show All Details.

1. (20%) Let $x_1, x_2, \ldots, x_n, \ldots$ be a sequence of i.i.d. random variables with uniform distributed pdf in the interval [-1, 1].

   (a) Let $z = x_1 + x_2$, find the *pdf* of z.

   (b) Let $z = \sum_{i=1}^{\infty} x_i$, find the *pdf* of z

2. (20%) Given $x_1, x_2, \ldots, x_n$ be a sequence of i.i.d. random variables with uniform distributed pdf in the interval [0, θ], where θ is unknown.
   (a) Give the likelihood function of
   (b) Find the maximum likelihood estimate of θ 。

3. (20%)The hypothesis that a coin is fair is tested against the hypothesis that head is favored. If we toss this coin 200 times and 125 heads are obtained, do we reject the null hypothesis with significance level equal to 0.05?

4. (20%) Describe the Tchbecheff Inequality and Prove it.

5. (20%) Let X and Y be two RVs. X is $N(\mu, \sigma)$. If $y = e^x$, find the *p.d.f.* of Y.

Algorithms 資格考 October 2014

1. (30%) Use the recursion tree method to show au upper bound of
   $T(n) = T(\frac{n}{3}) + T(\frac{2n}{3}) + O(n)$ as tight as possible.
2. (10%) Express the function $\frac{n^4}{30} - 8n^3 - 6n + 9$ in terms of Θ.
3. (10%) Using the master theorem to solve the recurrence $T(n) = T(\frac{2n}{3}) + 1$.
4. (20%) Describe a $\Theta(n \lg n)$-time algorithm that, given a set S of n integers and another integer x, determine whether or not there exist two elements in S whose sum is exactly x.
5. (10%) Describe the merge sort algorithm and analyze its time complexity.
6. (20%) Describe the quick sort algorithm and analyze its time complexity.

Information Retrieval Qualification (資訊檢索)

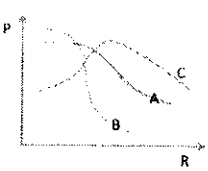
1. (20 points) Please briefly describe the following terminologies. (1) PageRank (2) Bag-of-Word (3) Latent Semantic Index (4) Signature file (in the area of information retrieval) (5) stemming

2. (20 points) For **each** following evaluation criteria, please **describe** and **explain ONE** retrieval system in which this criterion is important. (1) R-Precision (2) NDCG (3) P@3 (4) AUC (5) Specificity.

3. (20 points) Given a dataset with 50 pages labeled with A and 500 pages labeled with B. Now, you want to design a classifier to judge the label of a page is A or not. (1) Please describe the learning flow when you apply a 5-fold cross validation, (2) In your best two results shown in the right table ($1^{st}/2^{nd}$), please describe which result is better? (3) You find there have 3 specific terms (a1, a2, a3) that appear in class A but not in class B and also have 2 specific terms (b1, b2) that are in class B and not in class A. How to use them to improve your first classifier and second classifier?

	Answer (ground truth)	
Predict	A	B
A	40/30	150/100
Not A	10/20	350/400

4. (20 points) The precision-recall results of three retrieval methods A, B, and C are listed in the right figure. Please (1) the difference between A and B, (2) what's happened to C, and (3) describe which system is the best?



5. (20 points) You have designed a retrieval system on a data set with 1000 web pages. The test results of 2 queries are listed below. Please answer the following questions.
   (1). (8 points) If the numbers of positive pages for these 2 queries are 5 and 6, please answer the R-precision and MAP of M2 (just list your result, don't need to calculate out).
   (2). (5 points) Use the result of Q1 to draw the ROC curve of M1 and M2.
   (3). (7 points) Use ROC curve to judge which method is better? How to improve it?

Top-10 results	Result (M1)	Result (M2)
Q1	OOXXO XXXOO	XXXOX OOOOX
Q2	OXOXO XOXOX	OOOOX XXOOX

Qualify Exam - Data Mining

Notice: Close book. 滿分: 100 分

(共 2 頁)

1. (30%; 6% for each of A-E) Answer the following questions:
   A. Explain what are *maximal itemsets* and *closed itemsets* in the topic of frequent itemset mining, respectively.
   B. In mining time series data, one problem encountered often is the high dimension in term of large time points. Describe an effective way to reduce the dimension of a large time series data and explain the tradeoff in doing this kind of dimension reduction.
   C. Give the definition of *Accuracy* and *Recall* using the *Confusion Matrix*.
   D. Explain what is "overfitting" problem in classification modeling and how to avoid it in using decision tree for classification.
   E. Describe how *"Ensemble"* method works for building a classifier and explain why it can reach higher accuracy than using only one classification method normally.

2. (25%) Answer the following questions on data clustering:
   A. Describe how *k-means* and *DBSCAN* work, respectively.
   B. Compare the above two clustering methods in terms of accuracy and efficiency.
   C. Given a dataset *D*, suppose two clustering results *R1* and *R2* are obtained using *k-means* and *DBSCAN*, respectively. Explain how to compare the quality of the clustering results *R1* and *R2* through cluster validation methods.

3. (25%) Answer the following questions:
   A. Given a database *D* consisting of a set of transactions T_i: $\{t_i, C_{id}, (I_a, I_b, ..., I_n)\}$, where t_i is the purchase time of transaction T_i, C_{id} is the customer id and $(I_a, I_b, ..., I_n)$ are the items contained in this transaction. Defining *Sequential Pattern Mining* as "Finding all maximal sequences that meet the user-specified minimum support S_{min}", please describe the main processes involved in conducting the *Sequential Pattern Mining*.
   B. *AprioriAll* and *AprioriSome* are the well-known methods for mining sequential patterns. Describe how *AprioriAll* and *AprioriSome* work, respectively, and point out the main differences between them briefly.
   C. Give an approach for mining sequential patterns without generating candidates.

4. (20%) Given a dataset *D* of *m* records, where each record contains attributes $\{A_1, A_2, ...,$

$A_n\}$ and a class C. Suppose the value of any A_i ($1 \leq i \leq n$) is in numerical type and the value of C belongs to "Positive" or "Negative". Defining ratio R as $count(C_{Positive})/count(C_{Negative})$, where $count(C_{Positive})$ and $count(C_{Negative})$ means the number of records with class value as "Positive" and "Negative", respectively. Answer the following questions about classification modeling:

A. If the value of R is 1, in general cases, how would you rank *Decision Tree, SVM (Support Vector Machine)* and *CBA (Classification Based on Association)* in terms of the accuracy and execution efficiency in building a classification model on D? Explain why.

B. Suppose the value of R is 0.05 and the *Recall* of "Positive" class for the classification result is low. Without changing the classification methods, given an effective approach to deal with dataset D for improving the *Recall* ?

103. 10. 31 VLSI

1. Explain the following terms in detail: (60%)
   (a) critical path (b) setup time
   (c) functional simulation (d) Design for testability
   (e) hold time (f) soft IP

2. Describe the difference between full custom and Cell-based design flow. (20%)

3. Suppose you have completed a circuit design with hardware description language. Please describe the advantages and disadvantages of implementing your circuit with (a) ASIC and (b) CPLD/FPGA, respectively.