Hashing Solution

H1. Give an example of a bad hash function for strings (that generates many collisions). Justify why it is bad: find some strings that will hash to the same cell.

H2. A hash table has 1000 slots and 200 items have already been hashed in it. What is the load factor?

H3. In the hash table below, two items are originally hashed to the same slot. The slots examined for inserting one of them are shown by a star and the slots examined for inserting the other are shown by a plus. You can assume that the table size is very big (e.g. more than 1000) and the slots shown did not require a mod (%) operation (that is we did not have to wrap around).

a) What type of open addressing was used? Justify.
b) Give the next slots to be checked for each item (show where the next star and where the next plus will be).

Index	
…	
5	+ *
…	
8	+
9	*
…	
11	+
12	
13	*
14	+
…	
17	*
…	

H4. The images below show the occupancy of two hash tables. Both tables have **the same size**, the **same items** hashed in them, and both use **open addressing**. However they differ in the way they find an available slot in the table. Which one is a better hash table and why? (You do NOT have to deduce how they find the next available slot. You simply have to judge which one would behave better based on this image.)

P5. You want to hash integers in a table of size 9 and use quadratic $_{probing}$. Give an appropriate hash function for this table (give the math function, not the C code for it). Give 4 different integers that hash to the same value. Draw the table (with all its cells) and show where these numbers are inserted (hashed) in the table.

P6. A hash table of size 11 with open addressing, probes the following cells in order to hash an item: **6, 9, 3, 10**. What kind of open addressing method does it use? (You do not need to give the exact formula, just the type (name) for the open addressing used.) Justify your answer (specify what made you draw the conclusion that you did).

P7. Assume you have a hash table of **size 20**, that uses **double hashing** with the first hash function $\mathbf{h_1(x) = x \%20}$ and the second hash function $\mathbf{h_2(x) = 1+(x\%9)}$. Give the first 4 probes (indexes) in the sequence of probes generated by double hashing for the key **x = 13**. Show the index and the calculations in the table below.

$h(x,i) = (h_1(x) + i*h_2(x))\%20$ where $h_1(13) = 13\%20 = 13$ and $h_2(13) = 1+(13\%9) = 1+ 4 = 5$

Probe	Index	Calculation that resulted in the Index in the middle column.
1st	**13**	**$(h_1(x) + i*h_2(x))\%20 =(13+0)\%20 = 13\%20 = 13$**
2nd	**18**	**$(h_1(x) + i*h_2(x))\%20 =(13+1*5)\%20 = 18\%20 = 18$**
3rd	**3**	**$(h_1(x) + i*h_2(x))\%20 =(13+2*5)\%20 = 23\%20 = 3$**
4th	**8**	**$(h_1(x) + i*h_2(x))\%20 =(13+3*5)\%20 = 28\%20 = 8$**

P8. You have a hash table of size M that uses linear probing and is half full.

a) What is the load factor of this table? **$\alpha = 1/2$**
b) What is the average time to insert an item?
   a. All occupied cells are consecutive (e.g. left half is full, right half is empty)

$[M + (1+2+3+\ldots. +((M/2)-1)+M/2)]/M = \Theta((M^2/2 + M)/M) = \Theta(M)$

   b. Occupied and empty cells alternate.

$(M/2 + (M/2)*2)/M = (3M/2)/M = \Theta(1)$

c) How long does it take to search for an item. Note that you must now differentiate between EMPTY and DELETED cells. (A DELTED cell is one where there was an item and later that item was removed from the table). **We cannot tell. Search after DELETE takes at least as long as INSERT since the search will continue after a deleted cell is encountered, but insert would stop there. If all the non-occupied cells are "deleted" cells, the search can take linear time, $\Theta(M)$.**

P9. We know that linear probing is bad because it creates long chains and moreover, the longer the chain, the more likely it is to grow. What exactly is so bad about long chains? Wouldn't the data get in the table anyway? Why is it worse that it is grouped in these long chains rather than be more spread out? You answer cannot be along the lines "because if the data is more spread, the table is better". It has to be a justification involving the expected time needed to insert a new item. **P8 b) is the answer for P9. See there two examples that result in $\Theta(1)$ and $\Theta(M)$ average time.**

P10 was a repeat of P5 so I removed it.

H1-H4 are in the homework. See the homework solution in Blackboard.

For P5 and P6 the answer is straightforward. If in doubt, or to verify your answer see the work-out slides on Hashing with Open Addressing, especially slides 10-12.

Open Addressing
Example: insert 25

M = 10, $h_1(k) = k\%10$.

Table already contains keys: 46, 15, 20, 37, 23

Next want to insert 25:

$h_1(25) = 5$ (collision: 25 with 15)

Linear probing
- $h(k,i,M) = (h_1(k) + i)\% M$
  (try slots: 5,6,7,8)

Quadratic probing example:
- $h(k,i,M) = (h_1(k) + 2i+i^2)\% M$
  (try slots: 5, 8)
- Inserting 35(not shown in table):
  (try slots: 5, 8, 3,0)

i	$h_1(k) + 2i+i^2$	%10
0		
1		
2		
3		
4		

Index	Linear	Quadratic	Double hashing $h_1(k) = 14(k\%7)$	Double hashing $h_2(k) = 14(k\%9)$
0	20	20	20	20
1				
2				
3	23	23	23	23
4				
5	15	15	15	15
6	46	46	46	46
7	37	37	37	37
8				
9				

Where will 9 be inserted now (after 35)?

Open Addressing
Example: insert 25

M = 10, $h_1(k) = k\%10$.

Table already contains keys: 46, 15, 20, 37, 23

Next want to insert 25:

$h_1(25) = 5$ (collision: 25 with 15)

Linear probing
- $h(k,i,M) = (h_1(k) + i)\% M$
  (try slots: 5,6,7,8)

Quadratic probing example:
- $h(k,i,M) = (h_1(k) + 2i+i^2)\% M$
  (try slots: 5, 8)
- Inserting 35(not shown in table):
  (try slots: 5, 8, 3,0)

i	$h_1(k) + 2i+i^2$	%10
0	5+0=5	5
1	5+3=8	8
2	5+8=13	3
3	5+15=20	0
4		

Index	Linear	Quadratic	Double hashing $h_1(k) = 14(k\%7)$	Double hashing $h_2(k) = 14(k\%9)$
0	20	20	20	20
1				25
2				
3	23	23	23	23
4				
5	15	15	15	15
6	46	46	46	46
7	37	37	37	37
8	25	25		
9				

Where will 9 be inserted now (after 35)?

Open Addressing
Example: insert 25

M = 10, $h_1(k) = k\%10$.

Table already contains keys: 46, 15, 20, 37, 23

Next want to insert 25:

$h_1(25) = 5$ (collision: 25 with 15)

Double hashing example
- $h(k,i,M) = (h_1(k) + i * h_2(k))\% M$
- Choice of h_2 matters:
- $h_2(k) = 1+(k\%67)$: try slots: 5, 9,
  - $h_2(25) = 1 + 4 = 5$ => slots: 5,0,5,0,...
    Cannot insert 25.
- $h_2(k) = 1+(k\%69)$:
  - $h_2(25) = 1 + 7 = 8$ => slots: 5,3,1, (the cycle would have length 5 and give cells:
    5 (from (5+0)%10),
    3 (from (5+8)%10) ,
    1 (from (5+16)%10),
    9 (from (5+24)%10),
    7 (from (5+32)%10), and then it cycles back to 5 (from (5+40)%10).

Index	Linear	Quadratic	Double hashing $h_1(k) = 14(k\%7)$	Double hashing $h_2(k) = 14(k\%9)$
0	20	20	20	20
1				25
2				
3	23	23	23	23
4				
5	15	15	15	15
6	46	46	46	46
7	37	37	37	37
8	25	25		
9				

Where will 9 be inserted now?