Hash Lab COSC 3319 Fall 2022 Burris

Due: prior to the start of class November 14 for MWF and November 15 for TTh classes. Early submissions are appreciated.

For your convenience a file of 200+ random words and names (Words200D16) has been provided on Blackboard. Each line of text in the file is exactly 16 characters in length with the word left justified. You may utilize any code developed in class directly in the lab. While not initially apparent, this lab will be more of a thought problem than coding problem. A careful examination of the sample code utilized in class will reveal a great deal of the code required for the lab has already been developed in class. Indeed, with so many examples*, I assume you should be able to complete the “A” option code with little trouble. Hence grading will emphasize evaluation and presentation of results for correctly coded programs.* ***Be professional grade!***

**"C" Option:**

Assume a **main memory hash table of size 128**. For convenience, assume the characters in the key are numbered 1 through 16. “My” hash address must be calculated as follows for “STR” using Wallgol:

HA = ( STR(1) + STR(3) ) / 256 + STR(6) + 30.

For STR = “ABCDEFGHIJKLMNOP” this would be ( ‘A’ + ‘C’ ) / 256 + ‘F’ + 30.

All alphabetic/numeric characters in a key are left justified. Keys containing less than 16 alphabetic/numeric characters are padded with spaces on the right to a total of 16 characters.

I have never used this function but it should be substantially less than optimal generating a plethora of collisions. *All characters must be treated as integers in all steps of the calculation*! You may use the slice operator or simple subscripting depending on how you define the keys.

I recommend using a long\_integer to hold the intermediate results of the hash calculation. Remember a “slice” in Ada allows the user to treat a substring as a single unit. If “str” has the value “ABCD” then str(2..3) refers to the characters in positions 2 through 3 as a group from the string str, i.e., ”BC.” “str(2..2) would be the second character in the string, i.e., “B”. I am assuming you have instantiated appropriate functions (coercion) which converts character or strings (8 bits per ASCII character) to long integers (32 bits) so overflow will not occur in the sum.

**Each entry placed in the hash table must be a record with 3 fields. The first filed is the key, the second the initial hash address, and the third field is the number of probes to insert the key in the table counting the initial hash address as the first probe.**

A) Create a hash table in main memory and fill it 40% full. Use the linear probe technique developed in class to handle collisions. Now look up the first 25 keys placed in the table. Print the minimum, maximum, and average number of probes required to locate the first 30 keys placed in the table. Now search for the last 30 keys placed in the table. Print the minimum, maximum, and average number of probes required to locate the last 30 keys placed in the table. Print the contents of the hash table clearly indicating open entries (this should allow you to see the primary/secondary clustering effect). Calculate and print the theoretical expected number of probes to locate a random item in the table. Explain your empirical results in light of the theoretical results. **Your grade points will be highly dependent on your explanation!**

B) Create a hash table in main memory and fill it 40% full. Use the linear probe technique developed in class to handle collisions. Now look up the first 30 keys placed in the table. Print the minimum, maximum, and average number of probes required to locate the first 30 keys placed in the table. Now search for the last 30 keys placed in the table. YOU MUST PHYSICALLY SEARCH FOR EACH KEY TO CALCULATE THE STATISTICS! Print the minimum, maximum, and average number of probes required to locate the last 30 keys placed in the table. You must start filling the table with the same keys in the same order as used when only filling the table 40% full. Print the contents of the table clearly indicating open entries (this should allow you to see the primary clustering effect). Calculate and print the theoretical expected number of probes to locate a random item in the table. Discuss your empirical results in light of the theoretical results. **Your grade points will be highly dependent on your explanation!**

C) Repeat A and B above but use the random probe for handling collisions as developed in class. When you print the contents of the table you should be able to visually see the secondary clustering affect. Calculate and print the theoretical expected number of probes to locate a random key in the table. Discuss your empirical results in light of the theoretical results. **Your grade points will be highly dependent on your explanation!**

D) ***Present all your results in a neat tabular format (typed). Compare your results to the theoretical results. If there are differences, please explain why***. **I expect you to physically calculate the theoretical values for comparison to your empirical results for both the linear and random probes**! You may not ask a friend, enemy or anyone else for help. You may not look them up in some table or the internet! Providing someone else with the results will not only cause them to fail the lab but you as well. You may however teach them to use a calculator, spreadsheet, log tables, or other techniques. You may help others with coding errors.

E) My required hash function has multiple weaknesses. Criticize the hash function on a technical basis. **To receive full credit, you must state clearly and explicitly state why my hash function should fail!**

Based on your criticism, write a better hash function. **You must explain explicitly why your hash function should be better from both a theoretical and empirical standpoint based on your previous explanation of why my function is bad. You will not receive credit for simply writing a hash function that performs better**. Implement your hash function and generate the same results as required for parts A thru E presented in tabular format for comparison. Formally evaluate the results as part of your lab (typed evaluation). ***The results for all parts of the lab should appear in a single table***. Shame on you if your hash function’s performance does not exceed the theoretical values for both the linear and random probe but at least my hash function.

**"B" Option:**

You need not complete the "C" option using a main memory hash table. Rather, complete the "C" option **using a random access (relative) file for the hash table**. **Main memory may not be used to store any portion of the hash table!**

Generate the same results as required for parts A thru E presented in tabular format for comparison.

Script Kiddies may have trouble evaluating the failures of my hash function and creating a better function.

**"A" Option:**

Do both the "C" and "B" options. **Professionally compare and explain the results.**

***The results for all parts of the lab should appear in a single table***. Shame on you if your hash function’s performance does not exceed the theoretical values for both the linear and random probe collision handling techniques.

**REPORT ALL GRADING OPTIONS**

**The exact format of the hash report is left to you. A professional report must at a minimum contain a single table with all results conveniently displayed for the reader. It should contain the results for my hash function for the first and last 30 keys inserted using the linear probe. This should be followed by the results using the random probe. Next exhibit the same material using your hash function. If you do the “A” option, you must include both the “C” and “B” option results in the table.**

**Your evaluation of results should appear next.**

**Do not forget to include the required memory contents including the location, key and number of probes to store/locate the key for both the linear and random probes.**

**YOU ARE THE EXPERT - I AM THE CLIENT. Convenience me you have done a professional job!**

**Partial Table**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **My Hash Linear** | | **Your Hash Linear** | | **My Hash Random** | | **Your Hash Random** | |  |
| **First 30** | **Low 40%** | **High 85%** | **Low 40%** | **High 85%** | **Low 40%** | **High 85%** | **Low 40%** | **High 85%** |  |
| **Min Probe** |  |  |  |  |  |  |  |  |  |
| **Max** |  |  |  |  |  |  |  |  |  |
| **Expected** |  |  |  |  |  |  |  |  |  |
| **Theoretical** |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| **Last 30** | **Low 40%** | **High 85%** | **Low 40%** | **High 85%** | **Low 40%** | **High 85%** | **Low 40%** | **High 85%** |  |
| **Min Probe** |  |  |  |  |  |  |  |  |  |
| **Max** |  |  |  |  |  |  |  |  |  |
| **Expected** |  |  |  |  |  |  |  |  |  |
| **Theoretical** |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

**A professional verifies their hash function to be sure it is calculating hash addresses correctly. Hint: You do not write code to be sure the correct answers are being produced. Use your knowledge of the character set to do the calculations by hand. If you are not verifying results, you are not a professional.**