Section 1.2–3, Direct Files

1. A telephone book of 3.2 million 100-byte records is organized as a direct-access file with 3200-byte blocks and probe factor π=1.2. Assuming π to give an accurate measure of the number of accesses required, give the expected time required to find one entry, searching by name, on a disk with a transfer time of 2 microseconds (μs.) per byte and an average seek time of 20 milliseconds (ms.).

   (a) 20.00 milliseconds (ms.)

   (b) 24.00 milliseconds (ms.)

   (c) 28.56 milliseconds (ms.)

   (d) 31.68 milliseconds (ms.)

   (e) None of the above.

2. How is the probe factor, π, inaccurate in giving the expected number of accesses to a direct file?

   (a) π gives the best-case number of accesses.

   (b) π is 1 + (the number of records which overflow)/(total number of records).

   (c) Using π as the expected number of accesses assumes every overflow record requires exactly one extra probe.

   (d) π gives the number of accesses for a high-activity query.

   (e) More than one of the above explains the inaccuracy.

3. Blocks with block numbers 0, 1, 2, 3 and 4 have been created by linear hashing on integer keys using division/remainder with separate chaining, and starting with a single block. The key, 12, is presented to the file. What block will it hash to?

   (a) Block 0

   (b) Block 2

   (c) Block 3

   (d) Block 0 then block 4

   (e) The result depends on whether 12 is already present or on which blocks have overflowed.

4. Data is distributed linearly on a range of values, with few records at the low end of the range and many at the high end. (That is, the distribution function is a straight line with non-zero slope.) It is mapped into three blocks by a one-dimensional tidy function which is a single linear piece approximating the true cumulative distribution. (That is, the approximation treats the data as though it were uniformly distributed instead of linearly.)

   What is true about overcrowded and undercrowded blocks? Call the blocks 0, 1 and 2, in order from low end of the range to high end. Assume α=1.

   HINT. Plot the distribution and the cumulative distribution.

   (a) Block 0 is overcrowded, block 1 is just right, block 2 is undercrowded.

   (b) Block 0 is undercrowded, block 1 is just right, block 2 is overcrowded.

   (c) All blocks are equally populated, just right.

(d) Blocks 0 and 1 are undercrowded, block 2 is overcrowded.

(e) None of the above.

5. Thirty-three records are distributed in a multi-dimensional space with the following axial distribution for one axis.

   2 3 1 5 7 8 4 3

   If this axis is to be divided into five segments of seven records each, what is α and what are the best and next-best number of intersegment overflows that can be obtained by analyzing the distribution?

   (a) $\alpha = 1$, best = 0, next = 1

   (b) $\alpha = 33/35$, best = 1, next = 2

   (c) $\alpha = 33/35$, best = 1, next = 5

   (d) $\alpha = 33/35$, best = 0, next = 1

   (e) None of the above.

6. What are the practical differences between separate chaining and linear probing for use in RAM and on secondary storage?

7. What are two ways in which linear hashing improves virtual hashing?

8. A file has a 20-bit field, a, (whose values range from 1 to 1048576). The file is mapped to blocks 1 .. 1024 by a tidy function derived from the cumulative distribution function $D(a) = a/1048576$. If the tidy function uses four partitions, equally dividing the range of a-values, what block does it store the record with $a = 1000$ on? (Assume the record does not overflow its home block.)

9. A file of $N = 320$ records is to be stored in a two-dimensional static multipage structure. If the blocksize is to be $b = 3$ and there are 40 different values of the first field (V_1) and 70 different values of the second field (V_2), what are the best values for n, the number of blocks, and the factors f_1 and f_2? What is the load factor, α (if you do not have a calculator, give α as a ratio)?

   (Some factorizations for non-primes between 100 and 115: $102 = 6 \times 17, 104 = 8 \times 13, 105 = 15 \times 7, 106 = 2 \times 53, 108 = 4 \times 27, 112 = 16 \times 7, 114 = 6 \times 19, 115 = 5 \times 23$.)

10. A file of 32 records is to be multipaged statically. One axis is to be partitioned into four segments with a capacity for nine records each, and has the following axial distribution. Find the number of overflows for the best partitioning(s), and find all partitionings that give this many overflows.

    5 7 3 2 4 4 6 1

11. A two-dimensional dynamically multipaged file has the following axial directories. Give the two-dimensional array showing all the page numbers and including them all in the nine rectangles that indicate the sequence in which the rows and columns were added.

    Horizontal: 0 2 4 9 16 20

    Vertical: 0 1 6 12

12. A direct access file is stored in 4000-byte blocks on a device with access-transfer ratio 36,000 and transfer time of 1 μsec. per byte. Most (95%) of the records are on their home block, but 4% require an extra access and 1% require two extra accesses. What are (a) the probe factor and (b) the expected access time?

13. A two-dimensional data space with x-values 0 .. 7 and y-values 0 .. 15 is Z-ordered by bit interleaving. An orthogonal range query is posed for the values $x = 2, 3$ and $y = 3 \ldots 9$ inclusive. What is the maximum number of direct accesses needed?

14. Quadruples of the first ten integers (4, 8 .. 40) are hashed by division-remainder hashing with linear probing into a 10-block file of blocksize $b = 1$, (A) in ascending order and (B) in descending order. What is the ratio of probe factors that result, A:B?

15. If 6 is deleted from the following file, which was hashed by division-remainder with linear probing, what does the file look like after execution of Algorithm LD (*hash delete with linear probing*)? (Blocksize $b = 1$.)

location	0	1	2	3
value	7	2	6	3

16. Two volatile hashing methods are being run, (a) virtual hashing and (b) linear hashing. Each has just completed its fourth split. How many blocks has space been set aside for in each case?

17. Give an example of multidimensional hashing.

18. A tidy function is used to allocate the squares of the first twelve integers to two partitions of three pages each (blocksize $b = 2$). The first partition lies between 1 and 36, and so contains 1, 4, 9, 16, 25 and 36. The second partition lies between 37 and 144, and so contains 49, 64, 81, 100, 121 and 144. What is the probe factor (as a fraction)?

19. Find good values for f_1, f_2, n, b, and α for the following static multipaging problem. $N = 1000, V_1 = 800, V_2 = 600, b \approx 20$. Keep α as close to 1 as possible.

20. Given that $f_1 = 7, f_2 = 3, b = 2$ and the axial distribution for axis 2 is

    13 2 5 9 6 4

    find the best probe factor predicted by the analysis of this axis for static multipaging.

21. Axial analysis of a 2×2 static multipaging problem suggests two possible boundary placements for the one internal boundary of each axis: A or B for one axis and a or b for the other, as shown. Of the four possible combinations, use the histogram shown to pick the best combination(s) and show for it (them) the complete relationship between π and α.

2	1	2
2	0	1
1	0	1

22. A two-dimensional dynamic multipage space has the following indexes to its page numbers, used to calculate the page addresses

horizontal axis	0	1	2	6	8
vertical axis	0	3	10	15	

Draw a schematic diagram or otherwise indicate the two-dimensional ordering of the pages.

23. In a four-dimensional dynamic multipage space, just when a split is needed, $V_1 = 404, V_2 = 330, V_3 = 396, V_4 = 550, f_1 = 4, f_2 = 3, f_3 = 4$ and $f_4 = 5$. Which axis should be split next? (I.e., which f_i should be increased?) Assume we are more concerned about α than about π.

24. Describe three key-to-address transformations for direct-access files.

25. Describe two collision-resolution methods for direct-access files.

26. The items `3, 13, 8, 2, 1` are hashed, in that order, to a file of five blocks of size 1, using division-remainder with linear probing, then `13` is deleted, using Algorithm LD. What is the final disposition of the items in the blocks?

27. a) What principle applies to designing a file structure for volatile data? b) What results from this principle in the case of virtual hashing?

28. Discuss two ways in which linear hashing improves on virtual hashing.

29. A linearly hashed file has five blocks of capacity three records each, and controls splitting according to load factor, with $\alpha_0 = 0.75$. A fourteenth record is just about to be added. Will a block be split? Why?

30. How would you efficiently run a high-activity transaction on a hashed file?

31. A partition of a tidy function corresponds to blocks 5 .. 8 and key values 513 .. 768. What is the address of the item with key 580?

32. One thousand data records, with 300 different values for field X and 500 different values for field Y, is to be multipaged statically, with at most five records per page and a load factor of at least 80%. Find suitable factors, f_X and f_Y, of the total number of pages, for the axis partitions.

33. The following shows numbers of records allocated to pages in 4×4 static multipaging. The lines show boundaries which cannot be changed. In particular, the first and third vertical boundaries cannot be changed. The second vertical boundary may be placed on either side of the middle column. For all practical $\alpha < 1$ and all $\pi > 1$ show π as a function of α for *both* possible positions of the second boundary and thus determine the better position. There are 61 records in all, and you should leave all numbers as fractions.

6	4	1	2	3
5	6	2	3	4
2	3	0	5	2
2	1	2	5	3

34. A two-dimensional dynamic array is supported by the following axial arrays.

$i =$	0	1	2	3	4	$j =$	0	1	2	3
	`0`	`1`	`4`	`6`	`12`		`0`	`2`	`8`	`15`

What is the address of element $(i, j) = (2, 2)$?

35. Give two good hash functions, indicating which circumstances each is most useful for.

36. Seven records, with hash addresses h(k) = 0, 1, 2, 3, 3, 3, 3, respectively, are hashed to seven locations, each able to hold one record. What are the exact probe factors in the cases of linear probing and separate chaining? ("Exact" means the expected number of probes required, ot an optimistic approximation. Assume that separate chaining uses overflow locations also of capacity 1 record.)

37. Three records, "1", "2", and "5", hash to addresses h(k) = 1, 2, and 2, respectively, with blocksize 1. Only locations 0, 1, and 2 are available. Collisions are resolved by linear probing. Show all of the arrangements of records and their locations that can result.

38. Name two disadvantages of linear hashing which are not also disadvantages of conventional hashing (say with linear probing).

39. A file is stored with a tidy function, using the single field, *a*. Sixteen pages are needed, partitioned into four groups of four pages each. The cumulative distribution is given by the function

$$D(a) = \sqrt{\frac{a}{1024}}$$

and a has values from 1 to 1024.

What values of a give the sixteen page boundaries? (Find a way to indicate the answers without doing all the arithmetic, unless you have lots of time.)

40. A d-dimensional file has N records with V_i values of the ith field ($i = 1 \ldots d$). What are the smallest and largest possible values for N?

41. A 3-dimensional multipaged file has a million pages. In the absence of other information, estimate a) the *total* number of entries for the axial directories, and b) the number of pages that will be split when it is necessary to expand the file.

42. In a 4-bit computer, the factor for multiplicative hashing is $A = 9 = 1001_2$. What is the hash order of the first eight keys, $k = 1 \ldots 8$? (That is, which will hash to address 0, to address 1, .., to address 7?)

43. The algorithm for deletion from a linearly probed file, using division-remainder hashing modulo 3, would change the file

```
9 0 1     3 4 2     11 5 8
```

to

```
9 0       1 4 2     3 5 8
```

if `11` were deleted. But this leaves `3` inaccessible. What is wrong with this example?

44. Both the load factor and the probe factor have true and approximate variants. We indicate the approximation with $\sim$. Here they are illustrated for a two-block file, with blocksize 2 records, hashed by division-remainder with separate chaining, and with overflow blocks holding one record each. The keys in the file are `2, 3, 7, 5, 13`.

- α = (# records)/(*total* capacity) = 5/6 = 0.83
- $\sim\alpha$ = (# records)/(capacity of non-overflow blocks only) = 5/4 = 1.25
- π = (# probes)/(total records) = 8/5 = 1.6
- $\sim\pi$ = (# overflows)/(total records) = 7/5 = 1.4

If this file were built by virtual hashing and a split were made, what are the new values for each of the four quantities?

45. The keys `3, 7, 2, 5, 8, 13` have been hashed by division-remaindermodulo 3 into blocks of size 2. Linear probing was used for collision resolution and the keys were loaded in the above order.

    A request is made to find all values from `7` to `14` in this file, and we will suppose that these are too many values to fit into RAM. So the request is run as a merge, followed by direct-access lookups of all keys not found by the merge.

    How many keys, and which ones, will be found by the merge?

46. How should overflows be handled when using 1-dimensional tidy functions?

47. Multipage the following 2-dimensional data into 3×3 pages. Show the resulting values of $\sim\pi$ for $\alpha = 0.5, 1.0$

$$(1,1),(2,3),(3,1),(4,2),(4,7),(5,4),(6,5),(6,7),(7,6)$$

48. The following shows the two axial indexes for a variable 2-dimensional array, as in dynamic multipaging. Show the sequence in which the elements were added, by drawing boxes or by putting down all the addresses.

	0	1	4	6	20
0					
2					
8					
12					
16					

49. A multipaged file of n pages has been built up so that there are no overflows. Independent requests are made for r records. Show that the number of pages retrieved is *approximately*

$$r\left(1+\frac{1-r}{2n}\right)$$

50. The following integer values are hashed using division-remainder to five blocks of secondary storage, of capacity $b = 2$. What is the probe factor a) optimistically? b) with linear probing?

51. The integer values from the last question are hashed to seven blocks, still using division-remainder and with $b = 2$. What is the load factor?

52. Nine records are inserted into a linearly hashed file, initially empty.If the split criterion is *split unless this makes* $\alpha < 0.9$, how many pages will have been allocated for the nine records? (Assume that α is calculated using *all* the records that have been inserted, and $b = 1$.)

53. A file of five blocks has been created using division-remainder hashingwith $R = 100$ bytes, $b = 25$, and $\pi = 1.0$, on a device with access-transfer ratio $\rho = 14100$. a) How would you sort the following search keys to use them in a high-activity search? b) Is this the best way to search this file?

    `280, 531, 427, 329, 191, 350, 346, 189, 313`

54. A 40-block file is to be accessed by tidy function on a field, f, which has values from 1 to 1,000,000. A quarter of the records are uniformly distributed between 1 and 500,000; a quarter are uniformly distributed between 500,000 and 707,106; a third quarter are uniform between 707,107 and 866,025; and the final quarter are uniform between 866,026 and 1,000,000.

    a) Describe the tidy function, using mathematical expressions, words, and drawing. (This may help you: $\sqrt{2}/2 = 0.7071067$, $\sqrt{3}/2 = 0.8660254$)

55. Design static multipaging, in two dimensions and using blocksize $b = 1$, for the following sixteen tuples.

$$(\pm 0.20, \pm 0.98), (\pm 0.56, \pm 0.83), (\pm 0.83, \pm 0.56), (\pm 0.98, \pm 0.20)$$

    (That is, $(\pm 0.20, \pm 0.98)$ means the four tuples, (0.20, 0.98), (0.20, −0.98), (−0.20, 0.98), (−0.20, −0.98), and so on.)

    a) What are n, the number of blocks, V_i, the number of values for the ith attribute, and f_i, the number of segments for the ith attribute?

    b) Draw the tuples and the optimal segment boundaries.

    c) What is pi?

    d) Can this multipaging be improved? How?

56. Doing multiplicative hashing to a file of 16 blocks on an 8-bit computer, the product Ak works out to be `1101001011110110`. What block will the value of k that produced this product hash to?

57. The keys, `27, 8, 13, 16, 37`, and `42` are hashed (division-remainder) to a 7-block file with capacity $b = 1$ record per block. What is the resulting probe factor (a) using linear probing, and (b) using separate chaining?

58. (a) Why should we *prefer* multiplicative hashing over division-remainder hashing for linear hashing?

    (b) Why do we *use* division-remainder hashing for linear hashing?

59. The keys `8, 13, 16, 27, 37`, and `42` are stored by tidy function on blocks, of capacity one record each, numbered 0 to 5. If the tidy function gives the addresses 0, 1, 2, 3, 4, and 5, respectively, what is the probe factor for these records?

60. Find the best set(s) of boundaries for 6 segments to multipage an axis with the following axial distribution, given a load factor of 7/8. (There are 42 records.)

    `6 2 5 1 2 4 3 4 2 6 2 2 3`

    What is(are) the cost(s), in overflows?

61. In two-dimensional dynamic multipaging, $f_1 = 4$ and $f_2 = 5$, so there are 20 pages. A new record is added, after which $V_1 = 300$ and $V_2 = 375$. Adding this record triggers the splitting criterion, so we must split. Which axis should be split to give the smallest decrease in load factor?

62. In two-dimensional dynamic multipaging, the base page numbers, stored along the axes, are

x axis	0 1 2 9 16
y axis	0 3 6 12

    What is the address of the page with coordinates $(x, y) = (3, 1)$?

63. What is the worst-case computational complexity of direct access? Explain.

64. The following eight numbers are hashed modulo 5 to addresses of size $b = 1$. Assuming separate chaining to blocks of the same size, what are the probe factors, π and π_{op}?

    `3, 7, 24, 8, 6, 12, 18, 11`

65. The (3-D) record, (`27, 4, 12`) is hashed using a different hash function for each field: respectively (all division-remainder), modulo 5, 3, and 11. What is the (one-dimensional) address of the home block that will be searched? There are six possible answers: give two different ones and explain.

66. The following data are stored on two pages, $b = 4$, and searched using a tidy function which approximates the cumulative distribution from the smallest to the largest by a single linear piece. What is the probe factor for successful searches?

    `3, 6, 7, 8, 11, 12, 18, 24`

67. Static multipaging is needed for data with the following distribution along one of the axes.

    `122112122122`

    The segment capacity is 7 records.

    What are the best and worst-case overflows, and how many ways can the best be achieved?

68. The seven prime numbers starting at 7 are hashed in two different ways: first, using division-remainder modulo 7, second using division-remainder modulo 8. Which is the better hash function? Why? (The primes are: 7, 11, 13, 17, 19, 23, 29.)

69. Use linear probing to resolve the collisions in the two cases of the previous question. Assume a blocksize of 1. Show where each prime is stored. What are the probe factors? What are the "optimistic" probe factors? What are the load factors?

70. What are the "lazy" and the "greedy" criteria for splitting blocks in dynamic direct-access file structures? Discuss their relative advantages and disadvantages.

71. A set of letters hash (mod 7) as follows

0	1	2	3	4	5	6
`ip`	`qcxj`	`krd`	`es`	`tfm`	`u`	`hov`

Suppose each address can fit three letters and collisions are resolved by linear probing. a) How many accesses will be required by a direct access algorithm to retrieve all the above letters that are in the range `e..o`? b) How many accesses would be required by a hash merge algorithm? c) How many accesses would be required by a perfect tidy function, using the same blocksize? d) How might this change with an imperfect tidy function?

72. A one-piece linear tidy function is used for the "records" (1 letter each)

    `C,D,E,F,H,I,K,O,Q,R,T,U,X`.

    If `F` overflows, will it be found on a block with a higher or lower address than the block expected by the tidy function? (Hint. Draw the tidy function.)

73. A three-dimensional dynamic multipage structure, with page coordinates in the range (0,0,0) to (2,2,1), has three axial indexes giving start pages in the usual way:

    `0 1 8` `0 2 12` `0 4`

    What is the address of the *page* with coordinates (1,1,1)? (Pages on secondary storage have one-dimensional addresses.) 7 = max(1,2,4) + 2*1 + 1 or max(1,2,4) + 1 + 2*1

74. Using the keys `1, 3, 8, 13, 15`, demonstrate that division-remainder hashing using an even total number of addresses cannot produce a uniform distribution of addresses.

75. If 100 records are sought uniformly in 1000 blocks by a direct-access method, how many accesses are saved by a method that fetches any block only once, as opposed to independent retrieval of every record? (Hint. 0.999^{100}=0.90479)

76. In a linearly hashed file, just after an insertion, $b = 2, n = 2$, and $N = 5$, with no overflows in block 0. a) If $\pi_0 = 1.2$ and $\alpha_0 = 0.8$, which splitting criterion will cause the next block to split: "lazy" or "greedy"? Why? b) What will α and π_{opt} be after the split. Why?

77. The keys `1, 4, 9, 16, 25`, and `36` are stored on three blocks and accessed by a tidy function of a single linear approximation. If the load factor is 1.0, what is the probe factor?

78. Find all the optimal ways to partition the following distribution of keys into three blocks of three records each.

    1 1 2 1 1 1

79. In a dynamically multipaged file in two dimensions, the axial indexes show the following page numbers.

    `0 1 2 6 8` `0 3 10`

    Draw the page array, with all fifteen page numbers shown.

80. The keys `24,9,15,12,21,6,3,18,27,0` are to be stored by direct access, without concern for volatility. Write the simplest function that gives the best key-to-address transformation for this data.

81. In division-remainder hashing modulo 6, (a) how many keys that are successive multiples of 2 does it take to produce one collision? (b) how many that are successive multiples of 3?

82. If we are hashing twelve records into three pages, what is the probability that exactly 2 records are mapped to page 0? Give the expression, and work it out, approximately, to the form <integer>/1000.

83. The keys `1,2,3,4,5,6,..` are hashed, in that order, modulo 2, into blocks of size 2. Suppose a splitting mechanism is being used and the criterion is a) α-based, with $\alpha_0 = 0.6$, or b) π_{opt}-based, with $\pi_0 = 1.3$. Which of the keys will cause the first split in the two cases?

84. Why can we not use linear dynamic programming, as multipaging does, to find one-dimensional tidy functions?

85. A three-dimensional multipaged file has the following numbers of different values along the three axes: 444, 236, 503. It would fit into 36 pages with a load factor very close to 1.0. Suggest good possible values for f_1, f_2, and f_3 and say what these values will cause the load factor to be.

86. A matrix with indices $(a, b), a = 0, .., 3, b = 0, ..3$ is stored in two ways (below). For each, what is the address of element $(a, b) = (2, 1)$? i) Static storage: elements are allocated in order of increasing a within increasing b (a "runs faster"). ii) Dynamic storage: the axial indexes are, for a, `0,1,4,9`, and, for b, `0,2,6,12`.

87. How do direct-access files get around the impossibility of finding a record in a file of N records with fewer than $\mathcal{O}(\log N)$ comparisons?

88. The following five numbers were generated randomly from integers between 0 and 29. Hash them with division-remainder to (a) 6 addresses and (b) 5 addresses. Calculate α and π_{op} for each case (assuming b=1). Say which is better and discuss why.

```
28, 27, 4, 15, 12
```

89. Using linear probing to resolve the collisions of question 88, calculate π for each case.

90. With division-remainder hashing to addresses holding only one key, and linear probing for collision resolution, the following keys are stored as shown.

Key	239	60	298	220	102	285	13
Home	(1)	(4)	(4)	(3)	(4)	(5)	(6)
Address	0	1	2	3	4	5	6

(The home address for each key is shown to save you arithmetic.)

Where are all the keys stored after `102` has been deleted?

91. The following keys are inserted, one after the other, into a linearly hashed file using the "greedy" criterion with α_0=0.8 and blocksize 2. What is the final result?

```
220, 13, 102, 298, 60, 239, 285
```

92. Explain the activity, volatility, or symmetry requirements that led to the introduction of tidy functions as a direct-access data structure, after we investigated hashing. Which of low or high activity, volatility, or symmetry are supported by tidy functions of one sort or another?

93. Using all the space provided in the following six lines, draw the cumulative distribution of

`13, 60, 102, 220, 239, 285, 298`

as a step function over the range 0 to 299.

Assuming blocksize = 3, show the keys in the tidy file also on your drawing, in the appropriate positions; show the two-piece linear approximation and place page boundaries so that the approximation finds all the keys stored with probe factor 1.0. Write the value of the key at which the two linear pieces meet on your drawing.

94. The following two histograms are input to step MP4 of the algorithm to construct a static multipage file. Compute π_{op} from each of them for α= 1.0, 0.75, and 0.6, and state which is better. *Leave each value for* π_{op} *as a* fraction *unless it is an integer.*

3	1	5
4	3	2
2	5	2

(a)

4	2	4
3	4	2
2	4	2

(b)

95. Dynamic multipaging has built up an address space with axial indexes

i	0	1	2
index	0	1	8

j	0	1
index	0	4

k	0	1	2
index	0	2	12

Addressing within the two-dimensional slabs is done such that i, j and k increase in that order. That is, for example, if a slab is accessed by i and k, i runs through all its values before k is incremented.

What is the address for $(i, j, k) = (1, 0, 2)$? (For partial marks, what is the lowest address in the slab containing (1,0,2)?)

96. Suppose you have a perfect direct-access file structure (which never costs more than one probe to find a requested record). a) Arrange the following keys on four blocks of size 2 so that finding any subset of the keys in ascending order is maximally *expensive.*

`3, 8, 11, 24, 27, 30, 32, 35`

b) Which keys give the worst of these costs? What would the cost be if the file were optimally organized?

97. a) How does direct access manage constant expected retrieval cost? b) Why is direct access at best $\mathcal{O}(\log n)$ in complexity? Assuming no duplicate keys, say exactly when such worst-case performance is encountered for *any* direct-access file structure.

98. What are the load, probe, and optimistic probe factors after records with the following keys are hashed **mod** 7, using linear probing, and in the order shown, to a file with blocksize 2? (Just write the fractions; do not spend time working them out.)

`8, 9, 16, 23, 29, 30`

99. If the `23` were deleted from the keys in the above question, show the resulting file. (If there is more than one possibility, show all possibilities.)

100. Suppose a linearly hashed file has grown from one block to five ($b = 2$) and contains keys `0, 1, 3, 4, 5, 7, 9,` and `14`. a) Using α_0=0.8, should a block be split? Which one? b) Using π_0=1.1 (optimistic π), should a block be split? Which one?

101. A one-dimensional tidy function stores ten records per block, sorted according to two-dimensional Z-order on numeric fields with values from 0 to 255. At the beginning of the file, there are 13,000 records; at the end there are 15,000 records; and in between there are the numbers of records shown for the following ranges (each `x` ranges from `0` to `1`)

range	*range in binary*	*# records*
`(96,176)–(103,191)`	`(01100xxx,1011xxxx)`	50
`(104,176)–(111,183)`	`(01101xxx,10110xxx)`	21
`(104,184)–(107,187)`	`(011010xx,101110xx)`	7
`(104,188)–(107,191)`	`(011010xx,101111xx)`	8
`(108,184)–(111,187)`	`(011011xx,101110xx)`	13
`(108,188)–(111,191)`	`(011011xx,101111xx)`	6

To answer an orthogonal range query for all records in the range `(104,184)` to `(111,187)`, assuming no overflows, how many probes and how many scans should be done? How many blocks will be retrieved?

102. Records with keys `1, 2, 4, 8, 16,` and `32` are stored, two per block, by a one-dimensional, one-piece linear tidy function. As an average per record, how many accesses are needed to find all of these records by independent searches?

103. A three-dimensional multipaged file has equal numbers of different values for each of the indexed fields, and a million pages. Assuming no overflows, how many accesses must be done to retrieve all records for a) an exact-match query, b) a partial-match query given one field, c) a partial-match query given two fields, and d) an orthogonal range query on ten percent of the values of each field?

104. What is wrong with the page numbers shown in the following two axial indexes for a dynamic array? Say which numbers must be changed, to what, and give the general rule that has been violated.

`0, 2` `0, 1, 5`

105. Suppose a large set of records, whose keys are all multiples of 3, are mapped, using division-remainder hashing, to 6 blocks which have enough room for all the records. What is the optimistic probe factor?

106. What are the first eight bits of ϕw where lg w is the number of bits in a computer word and $\phi = (\sqrt{5}-1)/2 = 0.6180333$, the "golden ratio"? Hint: try 0.6180333 = .5(a + .5(b + .5(c + .. +.5(h + ..)..)))

107. The letters (but not the blanks) in "`files and databases`" are hashed to five blocks ($b = 2$) with division-remainder and linear probing. Duplicates are entered only once. From this, the letters `f, i, l` and `n` are deleted (leaving only the letters in "`databases`"). For computing hash addresses, assume the alphabet is represented by the integers 1 to 26. Show the resulting file.

108. Suppose we are in the process of inserting key `8` into the linearly hashed file

`4`	`1,5`	`2,6`	`3,7`

Given $\pi_0 = 1.2$ and $\alpha_0 = 0.8$, use the greedy splitting criterion to determine the new state of the file. (Show the new file.)

109. A file holds the records `3,4,5,6,7,8,12,13,14,15,16,17` on 6 pages. Draw the tidy function that addresses it with $\pi = 1.0$. What data must be stored for each segment?

110. Find the optimal partitioning into three parts of the axial distribution

`3 2 4 1 5 2 3 3 1`

What are the best-case and next-best-case overflows?

111. A two-dimensional dynamically multipaged file has the following page numbers in its axial indexes. In a drawing, show the numbering of all the pages and hence the order in which they were created.

`0 2 8` `0 1 4 6`

112. A very large number of records, with keys that are even integers but otherwise random, is hashed by division-remainder to an even number of pages with just enough total space to store all the records. a) What is the optimistic probe factor? b) If collision resolution is by linear probing, what is the probe factor? (You may need to experiment a little with some examples.)

113. The "golden ratio" used as multiplier in multiplicative hashing is $A = 2^8(\sqrt{5}-1)/2 = 10011100$ in binary, to eight bits. Eight records, with keys that are the eight powers of two, $2^1..2^8$, are hashed using it to eight addresses with space for one record each. a) What is the optimistic probe factor? b) If collision resolution is by linear probing, what is the probe factor?

114. After hashing by division-remainder modulo 5, the following storage is allocated, using linear probing.

contents	`512`	`16,256`	`2,32`	`8,128`	`4,64`
block	0	1	2	3	4

What will this look like after `32` is deleted?

115. A file is being built up using linear hashing and has 62 records in seven blocks of capacity 10 records each. The load-factor splitting criterion is being used with a threshold of 0.8. The last block to be split was block 2 (addresses begin at 0). A new record is added to block 4, causing an overflow. Which block will now be split?

116. In the previous question, which hash function(s) was (were) used to add the new record?

117. In a 3×2 two-dimensional hash function, ranging from indexes `(0,0)` to `(2,1)`, the block at address 4 (addresses start at 0) has index `(1,1)`. What is the address of the block with index `(1,0)`?

118. In a (one-dimensional) tidy function, what must be true if the "overflow" probing is to behave like linear probing collision resolution in hashing?

119. Prove that the "error" in one-dimensional tidy functions is proportional to the area between the cumulative distribution and the linear approximation.

120. A range query for names `Mac` to `Tom` and ages `17` to `22` is performed on a two-dimensional dynamic multipage structure with axial indexes (data values in indexes are *high* values)

(*basepage, name*): `(0, Joe), (2, Sue), (6, zzz)`
(*basepage, age*): `(0, 15), (1, 20), (4, 99)`

What are the addresses of the pages that will be returned (addresses start at 0)?

Section 1.3–3, Direct Files; Section 1.3–5, Sequential Files

1. In comparison with uniform usage, a non-uniform usage distribution
   (a) is always better for a sequential file.
   (b) gives a higher probability that the low-numbered blocks of a direct file are accessed.
   (c) gives a breakeven activity of $R/(R+\rho)$ where R is the record size and ρ is the access/transfer ratio.
   (d) gives a lower hit rate for direct-access files.
   (e) More than one of the above.
2. Assuming uniform usage distribution, what is the expected number of blocks accessed by a request for r independently located records in a direct access file of n blocks?
3. Assuming the 80-20 family of usage distributions, what is the expected number of blocks accessed by a request for r independently located records in a sequential file of n blocks? Explain any symbols you introduce and how the 80-20 distribution itself can be obtained.
4. What is the worst case complexity of merge sorting? Prove your answer.
5. (a) What is the expected size of initial runs generated by replacement selection?
   (b) The first twelve integers are read, in backwards order, into 4 RAM locations by the replacement selection algorithm. How many runs are generated?
6. Derive the binomial distribution for the probability that an ideal hash function places k out of N records in one of n blocks.
7. The usage of a sequential file is a distribution in the 80-20 family, with $\theta = 0.5$. The most used records are at the beginning of the file. a) How far down the file would the average search for a single record go? b) For two records?
8. Five records are hashed to two blocks of capacity two records each. What is the probability that exactly two records will be placed on the first block? (Assume an ideal hash function.)
9. For a disk drive with access/transfer ratio = 100,000, what record size gives a breakeven activity (between direct and sequential access) of 1%? Which access method is better if the records are larger than this size?
10. The formula for the breakeven activity between sequential and direct access files was derived in class for uniform usage. Qualitatively, how does this change when the usage is in the 80-20 family, with the most used records at the end of the sequential file?

Copyright ©2006 Timothy Howard Merrett
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation in a prominent place. Copyright for components of this work owned by others than T. H. Merrett must be honoured. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or fee. Request permission to republish from: T. H. Merrett, School of Computer Science, McGill University, fax 514 398 3883.

The author gratefully acknowledges support from the taxpayers of Québec and of Canada who have paid his salary and research grants while this work was developed at McGill University, and from his students (who built the implementations and investigated the data structures and algorithms) and their funding agencies.