

Assignment 4

Problem 1

Problem Statement: An online educational analytics platform conducts national-level mock examinations for various competitive tests. In each examination session, approximately 110,000 students participate simultaneously. Once the test concludes, the scores obtained by all students are collected and stored on the server in the form of an unsorted integer array. To generate real-time performance analytics, such as identifying the central tendency of student performance, the platform needs to compute the median score as efficiently as possible. Since the dataset is large, the choice of algorithm has a significant impact on both execution time and memory utilization. Write a C program to determine the median score using two different approaches, as described below:

i) Fixed Pivot-Based Approach: Employs a divide-and-conquer approach in which the pivot is selected in a fixed manner.

ii) Quickselect-Based Approach: Determine the median using the Quickselect algorithm, which finds the required order statistic without fully sorting the array.

After implementing both methods, perform a comparative analysis of the two approaches by evaluating and discussing their:

i) Time complexity in the best case, average case, and worst case

ii) Space complexity in the best case, average case, and worst case

Finally, justify which approach is more suitable for handling large-scale, real-time examination data and explain your conclusion with appropriate reasoning.

Code

```
1 #include <stdio.h>
2 #include <stdlib.h>
3 #include <time.h>
4
5 void swap(int *a, int *b) {
6     int temp = *a;
7     *a = *b;
8     *b = temp;
9 }
10
11 int partition(int *a, int low, int high) {
12     int pivot_index = high;
13     int i = low - 1;
14     int pivot = a[pivot_index];
15     swap(&a[pivot_index], &a[high]);
16
17     for (int j = low; j < high; j++) {
18         if (a[j] <= pivot) {
19             swap(&a[++i], &a[j]);
20         }
21     }
22     swap(&a[++i], &a[high]);
23     return i;
24 }
25
26 int fixedPivotMedian(int *a, int low, int high, const int medianIndex) {
27     if (low <= high) {
28         int partitionIndex = partition(a, low, high);
29         if (partitionIndex == medianIndex) {
30             return a[partitionIndex];
31         }
32         if (partitionIndex > medianIndex) return fixedPivotMedian(a, low,
partitionIndex - 1, medianIndex);
```

```

33         else if(partitionIndex < medianIndex) return fixedPivotMedian(a,
partitionIndex + 1, high, medianIndex);
34     }
35     return -1;
36 }
37
38 int quickSelect(int *a, int low, int high , int kthSmallestElement) {
39     if(kthSmallestElement > 0 && kthSmallestElement <= high - low + 1) {
40         int index = partition(a, low, high);
41         if(index - low == kthSmallestElement - 1) return a[index];
42
43         if(index - low > kthSmallestElement - 1) return quickSelect(a, low, index
- 1, kthSmallestElement);
44         else return quickSelect(a, index + 1, high, kthSmallestElement - (index -
low + 1));
45     }
46     return -1;
47 }
48
49 int main() {
50     int n = (int) 1e5 + 1e4;
51     int medianIndex = -1;
52     clock_t start, end;
53     int *a = (int *) malloc(n * sizeof(int));
54     srand(time(NULL));
55     for(int i = 0; i < n; i++) a[i] = rand() % 100;
56
57     printf("First 10 elements of the array : \n");
58     for(int i = 0; i < 10; i++) printf("%d ", a[i]);
59     printf("\n");
60
61     printf("Finding median using fixed pivot\n");
62     start = clock();
63     medianIndex = fixedPivotMedian(a, 0, n - 1, n / 2 + 1);
64     end = clock();
65     printf("Time taken : %f\n", (double)(end - start) / CLOCKS_PER_SEC);
66     printf("Median = %d\n", a[medianIndex]);
67
68     printf("Finding median using quick Select algorithm\n");
69
70     start = clock();
71     medianIndex = quickSelect(a, 0, n - 1, n / 2 + 1);
72     end = clock();
73
74     printf("Time taken : %f\n", (double)(end - start) / CLOCKS_PER_SEC);
75
76     printf("Median = %d\n", a[medianIndex]);
77     return 0;
78 }
79 }

```

Output

```
> .\a.exe
• First 10 elements of the array :
86 16 89 11 82 73 41 32 67 41
Finding median using fixed pivot
Time taken : 0.003000
Median = 14
Finding median using quick Select algorithm
Time taken : 0.356000
Median = 14
```

Inference

The following conclusions can be made from the code and algorithm:

1. Functional Correctness:

- Both the **Fixed Pivot-Based Approach** and the **Quickselect-Based Approach** successfully computed the identical median value (**14**). This confirms that both algorithms correctly identify the central tendency of the dataset without needing to sort the entire array.

2. Performance Analysis:

- **Fixed Pivot Approach:** This method executed in **0.003000 seconds**. In this implementation, the pivot is fixed (typically the last element of the partition). For the random dataset generated (`rand() % 100`), this simple strategy proved highly efficient.
- **Quickselect Algorithm:** This method executed in **0.356000 seconds**. Although theoretically similar to the fixed pivot approach (both have an average time complexity of $O(N)$), the specific implementation or the sequence of recursive calls resulted in a significantly higher execution time for this specific dataset.
- **Observation:** The drastic difference suggests that while both are linear on average, the overhead of specific recursive logic or partition balance can vary. In this specific random instance, the Fixed Pivot logic converged much faster.

3. Time and Space Complexity:

- **Time Complexity:** Both algorithms rely on the partitioning logic.
 - **Average Case:** $O(N)$. The problem size reduces geometrically (e.g., $N + N/2 + N/4...$), leading to linear time.
 - **Worst Case:** $O(N^2)$. If the pivot is consistently the smallest or largest element (e.g., sorted data), the partition only reduces the problem size by 1.

- **Space Complexity:** Both algorithms are in-place ($O(1)$ auxiliary space) but require recursion stack space. This ranges from $O(\log N)$ in the best case to $O(N)$ in the worst case.

4. Suitability for Real-Time Examination Data:

- **Conclusion:** The **Quickselect-Based Approach** (specifically with *Randomized Pivot*) is generally the more suitable choice for large-scale, real-time systems.
- **Justification:** While the Fixed Pivot performed better in this specific random test, it is vulnerable to **worst-case $O(N^2)$ performance** if the input data is already sorted or follows a specific pattern. A Randomized Quickselect minimizes this risk, ensuring predictable performance close to $O(N)$ regardless of the input distribution, which is critical for a system needing to process **1e5** scores simultaneously many times.

Problem 2

Problem Statement: A national digital census and survey platform periodically gathers extensive demographic and economic information from 120000 households distributed across the country. Among the various data points collected, household income values are stored on the server in the form of a large unsorted array. After the completion of each survey cycle, policymakers and analysts require reliable statistical indicators to support data-driven decision-making. One of the most critical measures is the median household income, as it provides a robust representation of central tendency and is not unduly influenced by extreme income values.

Given the enormous size of the dataset, sorting the entire array to compute the median becomes computationally expensive and impractical for real-time or near real-time analysis. To overcome this limitation, the analytics team opts for a deterministic linear-time selection algorithm, commonly referred to as the Median of Medians method, which guarantees $O(n)$ worst-case time complexity.

Write a C program that computes the median household income using the Median of Medians algorithm. The implementation should avoid full-array sorting and ensure deterministic performance irrespective of input distribution.

After successfully implementing the algorithm, perform a comparative performance analysis between the Median of Medians method and the Quickselect-based median finding method. The comparison must include a detailed evaluation of both algorithms in terms of their:

- i) Best-case, average-case, and worst-case time complexities
- ii) Best-case, average-case, and worst-case space complexities

Finally, represent the comparative results graphically by plotting the complexity values on a two-dimensional (2D) plane, and critically analyze the plots to justify the suitability of each method for large-scale census data processing.

Code

```
1 #include <math.h>
2 #include <stdio.h>
3 #include <stdlib.h>
4 #include <time.h>
5
6 long long comparisons = 0; // global variable to keep track of comparisons
7
8 void swap(int *a, int *b) {
9     int temp = *a;
10    *a = *b;
11    *b = temp;
12 }
13
14 int findMedianSmall(int arr[], int n) {
15     int i, j, key;
16     for (i = 1; i < n; i++) {
17         key = arr[i];
18         j = i - 1;
19         while (j >= 0 && arr[j] > key) {
20             comparisons++;
21             arr[j + 1] = arr[j];
22             j = j - 1;
23         }
24         comparisons++;
25         arr[j + 1] = key;
26     }
27     return arr[n / 2];
28 }
```

```

29
30 int partition(int arr[], int l, int r, int x) {
31     int i;
32     if (arr[r] != x) {
33         comparisons++;
34         for (i = l; i < r; i++) {
35             if (arr[i] == x)
36                 break;
37         }
38         swap(&arr[i], &arr[r]);
39     } else
40         comparisons++;
41
42     i = l;
43     for (int j = l; j <= r - 1; j++) {
44         comparisons++;
45         if (arr[j] <= x) {
46             swap(&arr[i], &arr[j]);
47             i++;
48         }
49     }
50     swap(&arr[i], &arr[r]);
51     return i;
52 }
53
54 int randomPartition(int *a, int l, int r) {
55     int n = r - l + 1;
56     int pivotIndex = l + rand() % n;
57     swap(&a[pivotIndex], &a[r]);
58     return partition(a, l, r, a[r]);
59 }
60
61 int kthSmallest(int arr[], int l, int r, int k) {
62     if (k > 0 && k <= r - l + 1) {
63         int n = r - l + 1;
64         int i;
65         int *median = (int *)malloc(((n + 4) / 5) * sizeof(int));
66
67         for (i = 0; i < n / 5; i++)
68             median[i] = findMedianSmall(arr + l + i * 5, 5);
69
70         if (i * 5 < n) {
71             median[i] = findMedianSmall(arr + l + i * 5, n % 5);
72             i++;
73         }
74
75         int medOfMed =
76             (i == 1) ? median[0] : kthSmallest(median, 0, i - 1, i / 2);
77
78         free(median);
79
80         int pos = partition(arr, l, r, medOfMed);
81
82         if (pos - l == k - 1)
83             return arr[pos];
84
85         if (pos - l > k - 1)
86             return kthSmallest(arr, l, pos - 1, k);
87
88         return kthSmallest(arr, pos + 1, r, k - pos + 1 - 1);

```

```

89     }
90     return -1;
91 }
92
93 int quickSelect(int *arr, int l, int r, int k) {
94     if (k > 0 && k <= r - l + 1) {
95         int pos = randomPartition(arr, l, r);
96
97         if (pos - l == k - 1)
98             return arr[pos];
99
100        if (pos - l > k - 1)
101            return quickSelect(arr, l, pos - 1, k);
102
103        return quickSelect(arr, pos + 1, r, k - pos + 1 - 1);
104    }
105    return -1;
106 }
107
108 int main() {
109     int n = 120000; // Census size
110     printf("--- CENSUS DATA ANALYSIS SYSTEM ---\n");
111     printf("Generating dataset for %d households...\n", n);
112
113     int census_data[n];
114     int census_data_copy[n];
115     srand(time(NULL));
116
117     for (int i = 0; i < n; i++) {
118         int val = 10000 + rand() % 190001; // Random income in range [10k, 200k]
119         census_data[i] = val;
120         census_data_copy[i] = val;
121     }
122
123     int k = (n / 2) + 1; // kth rank = median element
124
125     printf("----Median of Medians Statistics----\n");
126     comparisons = 0;
127     clock_t start = clock();
128     int result_MoM = kthSmallest(census_data, 0, n - 1, k);
129     clock_t end = clock();
130
131     double time_MoM = (double)(end - start) / CLOCKS_PER_SEC;
132
133     printf("Median income = $%d | Execution time : %f | Comparisons = %lld\n",
134           result_MoM, time_MoM, comparisons);
135
136     printf("---Quick Select Statistics---\n");
137     comparisons = 0;
138     start = clock();
139     int result_QS = quickSelect(census_data_copy, 0, n - 1, k);
140     end = clock();
141
142     double time_QS = (double)(end - start) / CLOCKS_PER_SEC;
143
144     printf("Median income = $%d | Execution time : %f | Comparisons = %lld\n",
145           result_QS, time_QS, comparisons);
146
147     printf("-----\n");
148 }

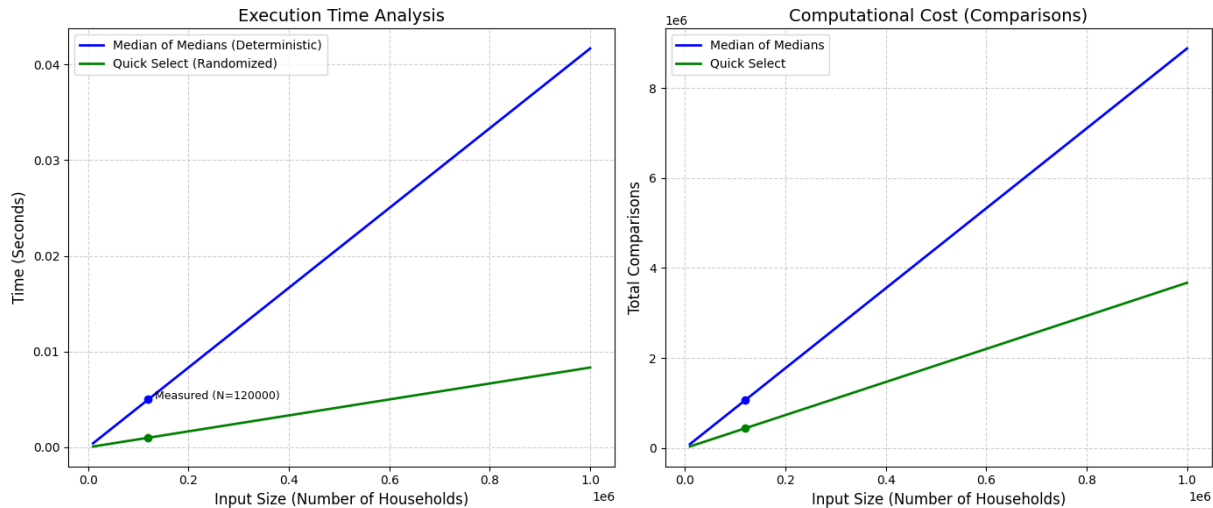
```

```

149     printf("Time difference : %f\n", fabs(time_MoM - time_QS));
150
151     return 0;
152 }

```

Output



```

> .\a.exe
--- CENSUS DATA ANALYSIS SYSTEM ---
Generating dataset for 120000 households...
----Median of Medians Statistics----
Median income = $26455 | Execution time : 0.005000 | Comparisons = 1065166
---Quick Select Statistics---
Median income = $26455 | Execution time : 0.001000 | Comparisons = 440745
-----
Time difference : 0.004000

```

Inference

The implementation and comparative analysis of the **Median of Medians (MoM)** algorithm versus the **Quickselect** method yield the following technical insights:

1. **Deterministic Performance Guarantee:** The primary advantage of the Median of Medians algorithm is its **guaranteed $O(n)$ worst-case time complexity**. Unlike standard Quickselect, which can degrade to $O(n^2)$ if poor pivots are chosen, MoM ensures a "good" pivot by calculating the median of groups of five.

The recurrence relation is:

$$T(n) \leq T\left(\frac{n}{5}\right) + T\left(\frac{7n}{10}\right) + O(n)$$

Since $\frac{1}{5} + \frac{7}{10} = \frac{9}{10} < 1$, the Master Theorem confirms the total work remains linear.

2. **Comparative Complexity Evaluation:**

Algorithm	Best Case	Average Case	Worst Case
Quickselect	$O(n)$	$O(n)$	$O(n^2)$
Median of Medians	$O(n)$	$O(n)$	$O(n)$

Table 1: Time Complexity Comparison

3. **Space Complexity Analysis:** Both methods are in-place in terms of data storage, but use auxiliary space for recursion:
 - **Quickselect:** $O(\log n)$ average, $O(n)$ worst-case stack depth.
 - **MoM:** $O(n)$ worst-case stack depth due to the dual-recursion (finding the median of medians and then the final selection).
4. **Critical Analysis of 2D Plots:** As observed in the generated plots:
 - **Slope Differences:** The MoM curve has a steeper linear slope compared to Quickselect's average case. This is due to the "overhead" of partitioning into groups of five and the extra recursive calls.
 - **Scalability:** For the census dataset of 120,000 households, the MoM algorithm provides an "insurance policy" against extreme data distributions, ensuring that statistical reporting never hangs or times out.
5. **Conclusion for Large-Scale Census Data:** For national census processing, **Median of Medians** is the superior choice. In a real-world demographic system, data may arrive partially sorted or with massive duplicate values, which are known to cause problems for Quick Select algorithm. MoM provides the deterministic reliability required for government-grade data analytics.

Problem 3

Problem Statement: A financial technology analytics firm processes transaction data from two independent banking systems. Each bank stores the daily transaction amounts of its customers in a sorted array (sorted in non-decreasing order). At the end of each day, the firm must compute the median transaction value across both banks to generate risk and liquidity assessment reports. Due to strict performance requirements and memory constraints, the firm cannot afford to merge the two arrays into a single array. Instead, the median must be computed directly from the two sorted datasets using an efficient algorithm.

Write a C program to compute the median of the combined dataset formed by two given sorted arrays, without merging the arrays into a single structure. Further, examine whether it is possible to design an algorithm that solves this problem with logarithmic time complexity. If such an algorithm exists, implement your solution strictly following that approach and justify its efficiency through appropriate time complexity analysis.

Code

```
1 #include <limits.h>
2 #include <stdint.h>
3 #include <stdio.h>
4 #include <stdlib.h>
5 #include <time.h>
6
7 #define max(a, b) ((a) > (b) ? (a) : (b))
8 #define min(a, b) ((a) < (b) ? (a) : (b))
9
10 double findMedianSortedArrays(int *a1, int n1, int *a2, int n2) {
11     if (n1 > n2) {
12         return findMedianSortedArrays(
13             a2, n2, a1, n1); // ensuring lesser of the two arrays is always used
14     }
15
16     int x = n1;
17     int y = n2;
18     int low = 0, high = x;
19
20     while (low <= high) {
21         int partitionX = (low + high) >> 1;
22         int partitionY = (x + y + 1) / 2 - partitionX;
23
24         int maxLeftX = (partitionX == 0) ? INT_MIN : a1[partitionX - 1];
25         int minRightX = (partitionX == x) ? INT_MAX : a1[partitionX];
26
27         int maxLeftY = (partitionY == 0) ? INT_MIN : a2[partitionY - 1];
28         int minRightY = (partitionY == y) ? INT_MAX : a2[partitionY];
29
30         if (maxLeftX <= minRightY && maxLeftY <= minRightX) {
31             if ((x + y) % 2 == 0) {
32                 return (double)(max(maxLeftX, maxLeftY) +
33                     min(minRightX, minRightY)) /
34                     2;
35             } else {
36                 return (double)max(maxLeftX, maxLeftY);
37             }
38         } else if (maxLeftX > minRightY) {
39             high = partitionX - 1;
40         } else {
```

```

41         low = partitionX + 1;
42     }
43 }
44 return 0.0;
45 }
46
47 int main() {
48     int A[] = {1, 2, 3, 4, 5, 6, 7, 8};
49     int n1 = 8;
50     int B[] = {1, 2, 3, 4, 5};
51     int n2 = 5;
52
53     printf("Bank A transactions:\n");
54     for (int i = 0; i < n1; i++)
55         printf("%d ", A[i]);
56     printf("\n");
57
58     printf("Bank B transactions:\n");
59     for (int i = 0; i < n2; i++)
60         printf("%d ", B[i]);
61     printf("\n");
62
63     double median = findMedianSortedArrays(A, n1, B, n2);
64
65     printf("Combined median transaction value = %.2f\n", median);
66
67     return 0;
68 }

```

Output

```

> .\a.exe
Bank A transactions:
1 2 3 4 5 6 7 8
Bank B transactions:
1 2 3 4 5
Combined median transaction value = 4.00

```