# Algorithms ROBERT SEDGEWICK | KEVIN WAYNE





# Two classic sorting algorithms: mergesort and quicksort

#### Critical components in our computational infrastructure.

#### Mergesort. [this lecture]

#### Quicksort. [next lecture]







# 2.2 MERGESORT

**‣** *mergesort* 

**‣** *bottom-up mergesort* 

**‣** *sorting complexity* 



# Algorithms

ROBERT SEDGEWICK | KEVIN WAYNE

**<https://algs4.cs.princeton.edu>**



#### Basic plan.

# **input** M E R G E S O R T E X A M P L E **sort left half** E E G M O R R S T E X A M P L E E L M P T X



- ・Divide array into two halves.
- ・Recursively sort left half.
- ・Recursively sort right half.
- ・Merge two sorted halves.

4



#### **First Draft**  $of \alpha$ Report on the **EDVAC**

John von Neumann



Goal. Given two sorted subarrays a[lo] to a[mid] and a[mid+1] to a[hi], replace with sorted subarray a[lo] to a[hi].







### Merging: Java implementation













**How many calls does** merge() **make to** less() **when merging two sorted subarrays, each of length** *n* / 2**, into a sorted array of length** *n* **?**

7



 $D. \sim n$ 





#### Mergesort: Java implementation



```
public class Merge {
    private static void merge(...) {
       /* as before */ 
 }
    private static void sort(Comparable[] a, Comparable[] aux, int lo, int hi) {
      if (hi \leq 10) return;
      int mid = 10 + (hi - 10) / 2; sort(a, aux, lo, mid);
      sort(a, aux, mid+1, hi); merge(a, aux, lo, mid, hi);
 }
    public static void sort(Comparable[] a) {
      Comparable \Box aux = new Comparable [a length]; \leftarrowsort(a, aux, 0, a.length - 1); }
<u>}</u>
```


*avoid allocating arrays within recursive function calls*

## Mergesort: trace





#### **Which subarray lengths will arise when mergesorting an array of length 12?**

- **A.** { 1, 2, 3, 4, 6, 8, 12 }
- **B.** { 1, 2, 3, 6, 12 }
- **C.** { 1, 2, 4, 8, 12 }
- **D.** { 1, 3, 6, 9, 12 }





## Mergesort: animation



**<https://www.toptal.com/developers/sorting-algorithms/merge-sort>**



#### **50 random items**



**in order**

**current subarray**

**algorithm position**

**not in order**

### Mergesort: animation







**in order**

**current subarray**

**algorithm position**

**not in order**

**<https://www.toptal.com/developers/sorting-algorithms/merge-sort>**



Running time estimates:

- Laptop executes  $10^8$ compares/second.
- Supercomputer executes  $10^{12}$  compares/second.

Bottom line. Good algorithms are better than supercomputers.





### Mergesort analysis: number of compares

Proposition. Mergesort uses  $\leq n \log_2 n$  compares to sort any array of length *n*.

Pf sketch. The number of compares *C*(*n*) to mergesort any array of length *n*  satisfies the recurrence:

 $C(n) \leq C(\lceil n/2 \rceil) + C(\lfloor n/2 \rfloor) + n - 1$  for  $n > 1$ , with  $C(1) = 0$ . *sort left half sort right half merge*



*proposition holds even when n is not a power of* 2 (*but analysis cleaner in this case*)

For simplicity. Assume *n* is a power of 2 and solve this recurrence:

 $D(n) = 2 D(n/2) + n$ , for  $n > 1$ , with  $D(1) = 0$ .

## Divide-and-conquer recurrence

Proposition. If  $D(n)$  satisfies  $D(n) = 2 D(n/2) + n$  for  $n > 1$ , with  $D(1) = 0$ , then  $D(n) = n \log_2 n$ .

Pf by picture. [assuming *n* is a power of 2]

15

$$
n = n
$$
\n
$$
(n/2)
$$
\n
$$
2(n/2) = n
$$
\n
$$
D(n/4)
$$
\n
$$
4(n/4) = n
$$
\n
$$
D(n/8) D(n/8) = 8(n/8) = n
$$
\n
$$
\vdots
$$

 $D(n) = n \log_2 n$ 



### Mergesort analysis: number of array accesses

Proposition. Mergesort makes Θ(*n* log *n*) array accesses.

Pf sketch. The number of array accesses *A*(*n*) satisfies the recurrence:

 $A(n) = A(\lceil n/2 \rceil) + A(\lfloor n/2 \rfloor) + \Theta(n)$  for  $n > 1$ , with  $A(1) = 0$ .

Key point. Any algorithm with the following structure takes Θ(*n* log *n*) time:

Famous examples. FFT, closest pair, hidden-line removal, Kendall-tau distance, …



```
\npublic static void f(int n) {\n    if (n == 0) return;\n    f(n/2);\n    f(n/2);\n    linear(n);\n    
$$
\leftarrow do \Theta(n) \text{ work}
$$
\n}\n
```

*what happens if we call*  f(n/3) *three times?* 

## Mergesort analysis: memory

Proposition. Mergesort uses Θ(*n*) extra space.

Pf. The length of the aux[] array is *n*, to handle the last merge.

Challenge 1 (not hard). Get by with an  $aux[]$  array of length  $\sim \frac{1}{2} n$  (instead of *n*). Challenge 2 (very hard). In-place merge. [Kronrod 1969]



Def. A sorting algorithm is in-place if it uses Θ(log *n*) extra space (or less). Ex. Insertion sort and selection sort.





*essentially negligible*

## **Consider the following modified version of mergesort. How much total memory is allocated over all recursive calls?**

```
private static void sort(Comparable[] a, int lo, int hi) {
  int mid = 10 + (hi - 10) / 2;
   Comparable[] aux = new Comparable[n];
```




if  $(hi \leq 10)$  return; int  $n = hi - Io + 1;$  sort(a, lo, mid);  $sort(a, mid+1, hi);$  merge(a, aux, lo, mid, hi); <u>}</u>

## Mergesort: practical improvement

- ・Mergesort has too much overhead for tiny subarrays.
- Cutoff to insertion sort for  $\approx 10$  items.

#### Use insertion sort for small subarrays.



```
private static void sort(...) { 
   if (hi \le 1o + CUTOFF - 1) {
       Insertion.sort(a, lo, hi);
       return;
   }<br>}
   int mid = 10 + (hi - 10) / 2; sort (a, aux, lo, mid);
   sort (a, aux, mid+1, hi);
    merge(a, aux, lo, mid, hi);
}
```
*makes mergesort about* 20% *faster*



**‣** *mergesort* 



# **‣** *bottom-up mergesort*

**‣** *sorting complexity* 

# Algorithms

ROBERT SEDGEWICK | KEVIN WAYNE

**<https://algs4.cs.princeton.edu>**

#### Basic plan.

- ・Pass through array, merging subarrays of length 1.
- ・Repeat for subarrays of length 2, 4, 8, …





Proposition. At most  $n \log_2 n$  compares;  $\Theta(n)$  extra space. Bottom line. Simple and non-recursive version of mergesort.





```
public class MergeBU {
   private static void merge(...) {
      /* as before */
 }
   public static void sort(Comparable[] a) {
     int n = a. length;
     Comparable[] aux = new Comparable[n];
     for (int sz = 1; sz < n; sz = sz+sz)
        for (int lo = 0; lo < n-sz; lo \leftarrow sz+sz)
          merge(a, aux, lo, lo+sz-1, Math.min(lo+sz+sz-1, n-1));
   }
} mid hi
```
Which is faster in practice for  $n = 2^{20}$ , top-down mergesort or bottom-up mergesort?

- **A.** Top-down (recursive) mergesort.
- **B.** Bottom-up (non-recursive) mergesort.
- **C.** No difference.
- **D.** *I don't know.*





#### Idea. Exploit pre-existing order by identifying naturally occurring runs.

Tradeoff. Fewer passes vs. extra compares per pass to identify runs.





# Timsort (2002)

- ・Natural mergesort.
- ・Use binary insertion sort to make initial runs (if needed).
- ・A few more clever optimizations.

Consequence. Only Θ(*n*) compares on many arrays with pre-existing order. Widely used. Python, Java, GNU Octave, Android, …

25

This describes an adaptive, stable, natural mergesort, modestly called timsort (hey, I earned it <wink>). It has supernatural performance on many kinds of partially ordered arrays (less than lg(n!) comparisons needed, and as few as n-1), yet as fast as Python's previous highly tuned samplesort hybrid on random arrays.

In a nutshell, the main routine marches over the array once, left to right, alternately identifying the next run, then merging it into the previous runs "intelligently". Everything else is complication for speed, and some hard-won measure of memory efficiency.

...



**Tim Peters**

# Timsort bug (February 2015)



Envisage

**About Envisage** 

**Follow Envisage** 

**Dissemination** 

# Proving that Android's, Java's and Python's sorting algorithm is broken (and showing how to fix it)

*© February 24, 2015* Written by Stijn de Gouw. & \$s E Envisage

Tim Peters developed the Timsort hybrid sorting algorithm in 2002. It is a clever combination of ideas from merge sort and insertion sort, and designed to perform well on real world data. TimSort was first developed for Python, but later ported to Java (where it appears as java.util.Collections.sort and java.util.Arrays.sort) by Joshua Bloch (the designer of Java Collections who also pointed out that most binary search algorithms were broken). TimSort is today used as the default sorting algorithm for Android SDK, Sun's JDK and OpenJDK. Given the popularity of these platforms this means that the number of computers, cloud services and mobile phones that use TimSort for sorting is well into the billions.



Log in

Q

# Timsort bug (May 2018)



#### **Details**



#### au wrote:

g on a proper complexity analysis of , we realised that there was an error in reporting such a bug (http://envisage-0-content/uploads/2015/02/sorting.pdf). that the correction implemented in the code (changing Timsort stack size) is at it is still possible to make it break. ined in full details in our analysis: org/pdf/1805.08612.pdf.

nd that coming upon data that actually rror is very unlikely, but we thought to know and do something about it. rs of the previous article advocated for, elieve that you should consider algorithm as explained in their article done in Python) rather than trying to fix

# Sorting summary





**number of compares to sort an array of n elements**



**‣** *mergesort* 



**‣** *bottom-up mergesort* 

**‣** *sorting complexity* 

# Algorithms

ROBERT SEDGEWICK | KEVIN WAYNE

**<https://algs4.cs.princeton.edu>**

# Computational complexity

A framework to study efficiency of algorithms for solving a particular problem *X*.







## Comparison tree (for 3 distinct keys a, b, and c)

31



*one* (*and only one*) *reachable leaf corresponds to each each possible ordering*



# Compare-based lower bound for sorting

Proposition. In the worst case, any compare-based sorting algorithm must make at least  $\log_2(n!) \sim n \log_2 n$  compares.

#### Pf.

- Assume array consists of *n* distinct values  $a_1$  through  $a_n$ .
- $n!$  different orderings  $\Rightarrow$   $n!$  reachable leaves.
- ・Worst-case number of compares = height *<sup>h</sup>* of pruned comparison tree.
- Binary tree of height *h* has  $\leq 2^h$  leaves.





## Compare-based lower bound for sorting

Proposition. In the worst case, any compare-based sorting algorithm must make at least  $\log_2(n!) \sim n \log_2 n$  compares.

#### Pf.

- Assume array consists of *n* distinct values  $a_1$  through  $a_n$ .
- $n!$  different orderings  $\Rightarrow$   $n!$  reachable leaves.
- ・Worst-case number of compares = height *<sup>h</sup>* of pruned comparison tree.
- Binary tree of height *h* has  $\leq 2^h$  leaves.

$$
2^h \geq #
$$
 reachable leaves

33

$$
\Rightarrow h \ge \log_2(n!)
$$
  
 
$$
\sim n \log_2 n
$$

*Stirling's formula*

 $= n!$ 



## Computational complexity

A framework to study efficiency of algorithms for solving a particular problem *X*.

First goal of algorithm design: optimal algorithms.

**example**  $(X =$  **sorting**)

comparison tree

# compares

 $\sim n \log_2 n$ 

 $\sim n \log_2 n$ 





mergesort

# Computational complexity results in context

Compares? Mergesort is optimal with respect to number compares. Space? Mergesort is not optimal with respect to space usage.



- Ex. Design sorting algorithm that makes  $\sim \frac{1}{2} n \log_2 n$  compares in worst case?
- Ex. Design sorting algorithm that makes Θ(*n* log *n*) compares and uses Θ(1) extra space.

Lesson. Use theory as a guide.

# Commercial break (sponsored by July)

#### Q. Why doesn't this Skittles sorter violate the sorting lower bound?



<sup>36</sup> **[https://www.youtube.com/watch?v=tSEHDBSynVo](http://www.apple.com)**





# Complexity results in context (continued)

Lower bound may not hold if the algorithm can exploit:

- ・The initial order of the input array. Ex: insertion sort makes only Θ(*n*) compares on partially sorted arrays.
- ・The distribution of key values. Ex: 3-way quicksort makes only Θ(*n*) compares on arrays with a small number of distinct keys. [next lecture]
- ・The representation of the keys.

Ex: radix sorts do not make any key compares; they access the data via individual characters/digits.



### A better Skittles sorter





**[https://www.youtube.com/watch?v=-\\_JdQZTQuTI#ws](https://www.youtube.com/watch?v=-_JdQZTQuTI#ws)**



## Asymptotic notations







#### Which of the following correctly describes the function  $f(n) = 3n^2 + 30n$  ?

- A.  $\sim n^2$
- **B.** Θ(*n*)
- **C.**  $O(n^3)$
- **D.** *All of the above.*
- **E.** *None of the above.*





#### Which of the following correctly describes the function  $f(n) = 10 \log n + 2 n \log n + 0.1 n$ ?

- **A.** Θ(*n* log *n*)
- **B.**  $O(2^n)$
- $C.$   $\Omega(n)$
- $D.$   $\Omega(n \log n)$
- **E.** *All of the above.*





Interviewer. Give a formal description of the sorting lower bound for sorting arrays of *n* elements.







Mergesort. Makes Θ(*n* log *n*) compares (and array accesses) in the worst case.

Sorting lower bound. No compare-based sorting algorithm makes fewer than Θ(*n* log *n*) compares in the worst case.

Divide-and-conquer. Divide a problem into two (or more) subproblems; solve each subproblem independently; combine results.









## **Credits**

#### $image/video$

#### Lecture Slides © Copyright 2024 Robert Sedgewick and Kevin Wayne





# A final thought





 $\overline{O}$ 

O

 $\overline{\circ}$