

Algorithms and Data Structures I.

Lecture Notes

Tibor Ásványi

Department of Computer Science
Eötvös Loránd University, Budapest
asvanyi@inf.elte.hu

February 8, 2021

Contents

1	Notations	4
1.1	Time complexity of algorithms	5
2	Elementary Data Structures and Data Types	8
2.1	Arrays, memory allocation and deallocation	8
2.2	Stacks	10
2.3	Queues	11
3	Algorithms in computing: Insertion Sort	13
4	Fast sorting algorithms based on the <i>divide and conquer</i> approach	16
4.1	Merge sort	16
4.1.1	The efficiency of merge sort	18
4.2	Quicksort	20
5	Linked Lists	24
5.1	One-way or singly linked lists	24
5.1.1	Simple one-way lists (S1L)	24
5.1.2	One-way lists with header node (H1L)	25
5.1.3	One-way lists with trailer node	25
5.1.4	Handling one-way lists	26
5.1.5	Sorting one-way lists: Insertion sort	27
5.1.6	Cyclic one-way lists	27
5.2	Two-way or doubly linked lists	28
5.2.1	Simple two-way lists (S2L)	28
5.2.2	Cyclic two-way lists (C2L)	28
5.2.3	Example programs on C2Ls	31
6	Trees, binary trees	35
6.1	General notions	35
6.2	Binary trees	37
6.3	Linked representations of binary trees	38
6.4	Binary tree traversals	40
6.4.1	An application of traversals: the height of a binary tree	41
6.4.2	Using parent pointers	42
6.5	Parenthesized, i.e. textual form of binary trees	42
6.6	Binary search trees	42
6.7	Level-continuous binary trees, and heaps	47
6.8	Arithmetic representation of level-continuous binary trees . . .	48

6.9	Heaps and priority queues	49
6.10	Heapsort	52
7	Lower bounds for sorting	55
7.1	Comparison sorts and the decision tree model	55
7.2	A lower bound for the worst case	56
8	Sorting in Linear Time	58
8.1	Radix sort	58
8.2	Distributing sort	58
8.3	Radix-Sort on lists	59
8.4	Counting sort	61
8.5	Radix-Sort on arrays ([1] 8.3)	64
8.6	Bucket sort	64
9	Hash Tables	66
9.1	Direct-address tables	66
9.2	Hash tables	67
9.3	Collision resolution by chaining	67
9.4	Good hash functions	69
9.5	Open addressing	70
9.5.1	Open addressing: insertion and search, without deletion	70
9.5.2	Open addressing: insertion, search, and deletion	71
9.5.3	Linear probing	72
9.5.4	Quadratic probing	73
9.5.5	Double hashing	74

References

- [1] CORMEN, T.H., LEISERSON, C.E., RIVEST, R.L., STEIN, C.,
Introduction to Algorithms (Third Edititon), *The MIT Press*, 2009.
- [2] CORMEN, THOMAS H., Algorithms Unlocked, *The MIT Press*, 2013.
- [3] NARASHIMA KARUMANCHI,
Data Structures and Algorithms Made Easy, *CareerMonk Publication*,
2016.
- [4] NEAPOLITAN, RICHARD E., Foundations of algorithms (Fifth edition),
Jones & Bartlett Learning, 2015. ISBN 978-1-284-04919-0 (pbk.)
- [5] WEISS, MARK ALLEN, Data Structures and Algorithm Analysis in C++
(Fourth Edition),
Pearson, 2014.
http://aszt.inf.elte.hu/~asvanyi/ds/DataStructuresAndAlgorithmAnalysisInCpp_2
- [6] WIRTH, N., Algorithms and Data Structures,
Prentice-Hall Inc., 1976, 1985, 2004.
<http://aszt.inf.elte.hu/~asvanyi/ds/AD.pdf>
- [7] BURCH, CARL, B+ trees
<http://aszt.inf.elte.hu/~asvanyi/ds/B+trees.zip>

1 Notations

$\mathbb{N} = \{0; 1; 2; 3; \dots\}$ = natural numbers.

$\mathbb{Z} = \{\dots - 3; -2, -1; 0; 1; 2; 3; \dots\}$ = integer numbers

\mathbb{R} = real numbers

\mathbb{P} = positive real numbers

\mathbb{P}_0 = nonnegative real numbers

$\log n = \begin{cases} \log_2 n & \text{ha } n > 0 \\ 0 & \text{ha } n = 0 \end{cases}$

$half(n) = \lfloor \frac{n}{2} \rfloor$, ahol $n \in \mathbb{N}$

$Half(n) = \lceil \frac{n}{2} \rceil$, ahol $n \in \mathbb{N}$

We use the structogram notation in our pseudo codes. In the structograms we use a UML-like notation with some C++/Java/Pascal flavour. Main points:

1. A program is made up of declarations. Each of them is represented by a structogram, but we do not care about their order.

2. We can declare subprograms, classes, variables and constants.
3. The subprograms are the functions, the procedures (i.e. void functions), and the methods of the classes.
4. **while** loop is the default loop ; **for** loops are also used..
5. The operators \neq, \geq, \leq are written by the usual mathematical notation. The assignment statements and the “is equal to” comparisons are written in Pascal style. For example, $x := y$ assigns the value of y to x and $x = y$ checks their equality.
6. Only scalar parameters can be passed by value (and it is their default). Non-scalar parameters are always passed by reference. For example, an object cannot be passed by value, just by reference.
7. An assignment statement can copy just a scalar value.
8. In the definitions of classes we use a simple UML notation. The name of the class comes in the first part. The data members in the second part, and the methods, if any, in the third part. A private declaration/specification is prefixed by a “-” sign. A public declaration/specification is prefixed by a “+” sign. We use neither templates nor inheritance (nor protected members/methods).
9. We do not use libraries.
10. We do not use exception handling.

1.1 Time complexity of algorithms

Time complexity (operational complexity) of an algorithm reflects some kind of abstract time. We count **the subprogram calls + the number of loop iterations** during the run of the program. (This measure is approximately proportional to the real runtime of the program, and we can omit constant factors here, because they are useful only if we know the programming environment [for example, (the speed of) the hardware, the operation system, the compiler etc.]).

We calculate the time complexity of an algorithm typically as a function of the size of the input data structure(s) (for example, the length of the input array). We distinguish $MT(n)$ (Maximum Time), $AT(n)$ (Average or expected Time), and $mT(n)$ (minimum Time). Clearly $MT(n) \geq AT(n) \geq$

$mT(n)$. If $MT(n) = mT(n)$ then we can speak of a general time complexity $T(n)$ where $T(n) = MT(n) = mT(n)$.

Definition 1.1 Given $g : \mathbb{N} \rightarrow \mathbb{R}$;
 g is asymptotically positive, **iff**
there exists an $N \in \mathbb{N}$ so that $g(n) > 0$ for each $n \geq N$.

In this chapter we suppose that f, g, h denote *asymptotically positive* functions of type $\mathbb{N} \rightarrow \mathbb{R}$ in each case, because they denote time (or space) complexity functions and these satisfy this property. Usually it is not easy to give such a function exactly. We just make estimations.

When we make an *asymptotic upper estimate* g of f , then we say that “ f is big-O g ”, and write $f \in O(g)$. Informally $f \in O(g)$ means that function f is at most proportional to g . ($f(n) \leq d * g(n)$ for some $d > 0$, if n is large enough.)

When we make an *asymptotic lower estimate* g of f , then we say that “ f is Omega g ”, and write $f \in \Omega(g)$. Informally $f \in \Omega(g)$ means that function f is at most proportional to g . ($f(n) \geq c * g(n)$ for some $c > 0$, if n is large enough.)

When we make an *asymptotic upper and lower estimate* g of f , then we say that “ f is Theta g ”, and write $f \in \Theta(g)$ which means that $f \in O(g) \wedge f \in \Omega(g)$.

Definition 1.2

$$f \prec g \iff \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$$

“ $f \prec g$ ” is read as “ f is asymptotically less than g ”. It can also be written as “ $f \in o(g)$ ” which means that

$$o(g) = \{h : \mathbb{N} \rightarrow \mathbb{P}_0 \mid h \prec g\}$$

Definition 1.3

$$g \succ f \iff f \prec g$$

“ $g \succ f$ ” is read as “ g is asymptotically greater than f ”.

Definition 1.4 $O(g) = \{f \mid \text{there exist positive constants } d \text{ and } n_0$
such that $f(n) \leq d * g(n)$ for all $n \geq n_0\}$

(“ $f \in O(g)$ ” can be read as “ f is at most proportional to g ”.)

Definition 1.5 $\Omega(g) = \{f \mid \text{there exist positive constants } c \text{ and } n_0$
such that $f(n) \geq c * g(n)$ for all $n \geq n_0\}$

(“ $f \in \Omega(g)$ ” can be read as “ f is at least proportional to g ”.)

Definition 1.6

$$\Theta(g) = O(g) \cap \Omega(g)$$

(“ $f \in \Theta(g)$ ” can be read as “ f is proportional to g ”.)

Consequence 1.7 .

$f \in O(g) \iff \exists d, n_0 > 0$, and $\psi : \mathbb{N} \rightarrow \mathbb{R}$ so that $\lim_{n \rightarrow \infty} \frac{\psi(n)}{g(n)} = 0$, and

$$f(n) \leq d * g(n) + \psi(n)$$

for each $n \geq n_0$.

Consequence 1.8 .

$f \in \Omega(g) \iff \exists c, n_0 > 0$, and $\varphi : \mathbb{N} \rightarrow \mathbb{R}$ so that $\lim_{n \rightarrow \infty} \frac{\varphi(n)}{g(n)} = 0$ and

$$c * g(n) + \varphi(n) \leq f(n)$$

for each $n \geq n_0$.

Consequence 1.9 $f \in \Theta(g) \iff f \in O(g) \wedge f \in \Omega(g)$.

Note 1.10 .

If $f \in O(g)$, we can say that g is asymptotic upper bound of f .

If $f \in \Omega(g)$, we can say that g is asymptotic lower bound of f .

If $f \in \Theta(g)$, we can say that f and g are asymptotically equivalent.

Theorem 1.11

$$\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0 \implies f \prec g \implies f \in O(g)$$

$$\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = c \in \mathbb{P} \implies f \in \Theta(g)$$

$$\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = \infty \implies f \succ g \implies f \in \Omega(g)$$

Consequence 1.12

$$k \in \mathbb{N} \wedge a_0, a_1, \dots, a_k \in \mathbb{R} \wedge a_k > 0 \implies a_k n^k + a_{k-1} n^{k-1} + \dots + a_1 n + a_0 \in \Theta(n^k)$$

Proof.

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \frac{a_k n^k + a_{k-1} n^{k-1} + \dots + a_1 n + a_0}{n^k} = \\
& \lim_{n \rightarrow \infty} \left(\frac{a_k n^k}{n^k} + \frac{a_{k-1} n^{k-1}}{n^k} + \dots + \frac{a_1 n}{n^k} + \frac{a_0}{n^k} \right) = \\
& \lim_{n \rightarrow \infty} \left(a_k + \frac{a_{k-1}}{n} + \dots + \frac{a_1}{n^{k-1}} + \frac{a_0}{n^k} \right) = \\
& \lim_{n \rightarrow \infty} a_k + \lim_{n \rightarrow \infty} \frac{a_{k-1}}{n} + \dots + \lim_{n \rightarrow \infty} \frac{a_1}{n^{k-1}} + \lim_{n \rightarrow \infty} \frac{a_0}{n^k} = \\
& a_k + 0 + \dots + 0 + 0 = a_k \in \mathbb{P} \implies \\
& a_k n^k + a_{k-1} n^{k-1} + \dots + a_1 n + a_0 \in \Theta(n^k)
\end{aligned}$$

□

2 Elementary Data Structures and Data Types

A **data structure** (DS) is a way to store and organize data in order to facilitate access and modifications. No single data structure works well for all purposes, and so it is important to know the strengths and limitations of several of them ([1] 1.1).

A **data type** (DT) is a **data structure** + its **operations**.

An **abstract data type** (ADT) is a mathematical or informal structure + its operations described mathematically or informally.

A *representation* of an ADT is an appropriate DS.

An *implementation* of an ADT is some program code of its operations.

[1] Chapter 10; [6] 4.1 - 4.4.2; [3] 3-5

2.1 Arrays, memory allocation and deallocation

The most common data type is the **array**. It is a finite sequence of data. Its data elements can be accessed and updated directly and efficiently through indexing. Most programming languages support arrays and we can access any element of an array in $\Theta(1)$ time.

In this book arrays must be declared in the following way.

$$A, Z : \mathcal{T}[n]$$

In this example A and Z are two arrays of element type \mathcal{T} and of size n .

In our model, the array data structure is an *array object* containing its size and elements, and it is accessed through a so called *array pointer* containing

its memory address. The operations of the array can

- read its size like $A.length$ and $Z.length$: $A.length = Z.length = n$ here,
- access (read and write) its elements through indexing as it is usual, for example $A[i]$ and $Z[j]$ here.

Arrays are indexed from 0.

Provided that an object or variable is created by declaring it, this object or variable is deleted automatically when the subprogram containing it finishes. Then the memory area reserved by it can be reused for other purposes.

If we want to declare array pointers, we can do it in the following way.

$$P : \mathcal{T}[]$$

Now, given the declarations above, after the assignment statement $P := Z$, P and Z refer to the same array object, $P[0]$ is identical with $Z[0]$, and so on, $P[n - 1]$ is identical with $Z[n - 1]$.

Array objects can be created (i.e. allocated) dynamically like in C++, but our array objects always contain their size, unlike in C++. For example, the statement

$$P := \mathbf{new} \mathcal{T}[m]$$

creates a new array object, pointer P refers to it, $P.length = m$.

Note that any object (and especially array object) generated dynamically must be deleted (i.e. deallocated) explicitly when the object is not needed any more. Deletion is done like in C++, in order to avoid memory leaking.

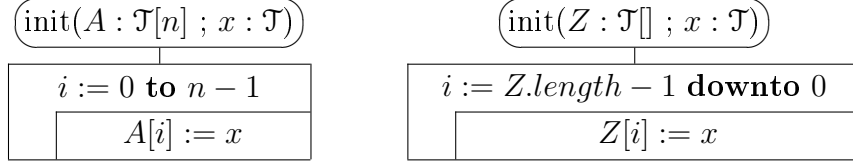
delete P

suffices here. It does not delete the pointer but the pointed object. Having deleted the object the memory area reserved by it can be reused for other purposes.

Unfortunately we cannot tell anything about efficiency of memory allocation and deallocation in general. Sometimes these can be performed with $\Theta(1)$ time complexity, but usually they need much more time, and their efficiency is often unpredictable. Therefore we avoid the overuse of them and we apply them only when it is really necessary.

If we want to pass an array parameter to a subprogram, the actual parameter must be the identifier of the array, the formal parameter has to be specified as an array pointer and the parameter passing copies the address of the actual array object into the formal parameter. If we write an identifier between the square brackets, it is a short notation for the length of the array.

The following two procedures are equivalent, because in the left strug-
togram $n = A.length$.



Given an array A , subarray $A[u..v]$ denotes the following sequence of elements of A : $\langle A[u], \dots, A[v] \rangle$ where u and v are valid indexes in the array. If $u > v$, the subarray is empty (i.e. $\langle \rangle$). In this case u or v may be invalid index.

Given an array A , subarray $A[u..v)$ denotes the following sequence of elements of A : $\langle A[u], \dots, A[v - 1] \rangle$ where u and $v - 1$ are valid indexes in the array. If $u \geq v$, the subarray is empty (i.e. $\langle \rangle$).

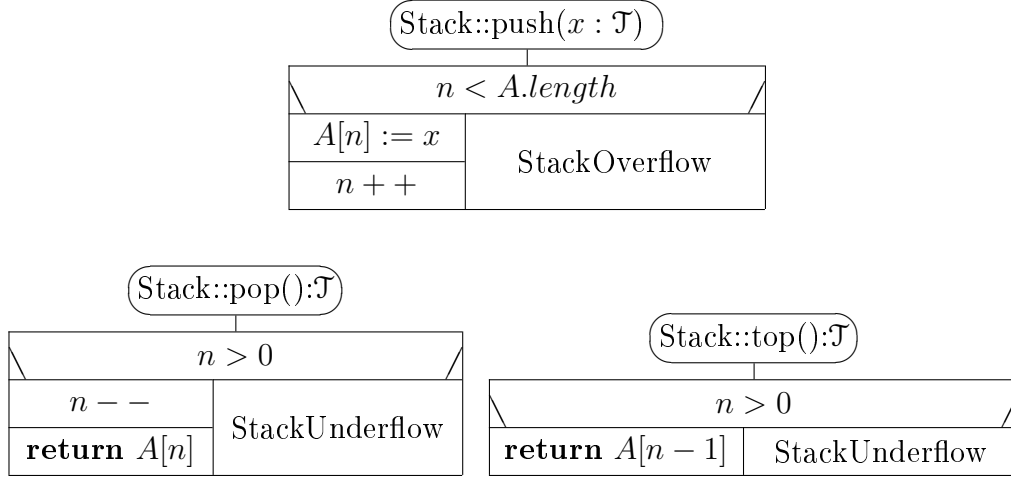
2.2 Stacks

A stack is a LIFO (Last-In First-Out) data storage. It can be imagined as a vertical sequence of items similar to a tower of plates on a table. We can push a new item at the top, and we can check or remove (i.e. pop) the topmost item.

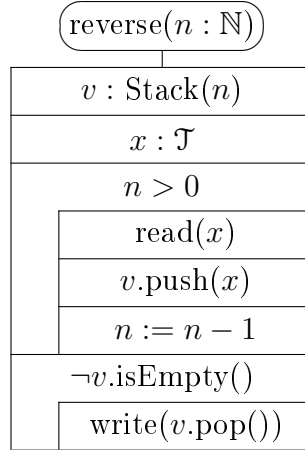
In the following representation the elements of the stack are stored in subarray $A[0..n)$. Provided that $n > 0$, $A[n - 1]$ is the top of the stack. Provided that $n = 0$, the stack is empty.

$T(n) \in \Theta(1)$ for each method, because there is neither iteration nor sub-program invocation in their code. The time complexities of the constructor and of the destructor depend on that of the **new** and **delete** expressions.

Stack
- $A : \mathcal{T}[]$ // \mathcal{T} is some known type ; $A.length$ is the max. size of the stack - $n : 0..A.length$ // n is the actual size of the stack
+ Stack($m : \mathbb{N}$) { $A := \text{new } \mathcal{T}[m] ; n := 0$ } // create an empty stack + push($x : \mathcal{T}$) // push x onto the top of the stack + pop() : \mathcal{T} // remove and return the top element of the stack + top() : \mathcal{T} // return the top element of the stack + isFull() : \mathbb{B} { return $n = A.length$ } + isEmpty() : \mathbb{B} { return $n = 0$ } + setEmpty() { $n := 0$ } // reinitialize the stack + \sim Stack() { delete A }



Example for a simple use of a stack: printing n input items in reversed order. We suppose that $read(x)$ reads from the current input the next piece of data and $write(x)$ prints to the current output the value of x . $T_{\text{reverse}}(n) \in \Theta(n)$.



2.3 Queues

A queue is a FIFO (First-In First-Out) data storage. It can be imagined as a horizontal sequence of items similar to a queue at the cashier's desk. We can add a new item to the end of the queue, and we can check or remove the first item.

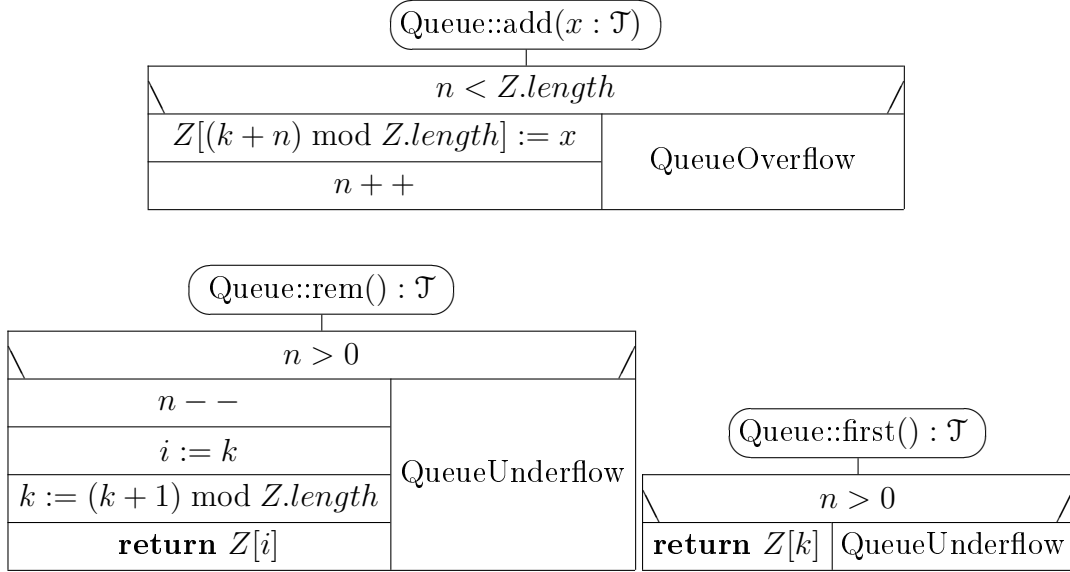
In the following representation the elements of the queue are stored in

$$\langle Z[k], Z[(k+1) \bmod Z.length], \dots, Z[(k+n-1) \bmod Z.length] \rangle$$

Provided that $n > 0$, $Z[k]$ is the first element of the queue where n is the length of the queue. Provided that $n = 0$, the queue is empty.

$T(n) \in \Theta(1)$ for each method, because there is neither iteration nor sub-program invocation in their code. The time complexities of the constructor and of the destructor depend on that of the **new** and **delete** expressions.

Queue
$-Z : \mathcal{T}[]$ \mathcal{T} is some known type $-n : 0..Z.length$ // n is the actual length of the queue $-k : 0..(Z.length - 1)$ // k is the starting position of the queue in array Z
$+ \text{Queue}(m : \mathbb{N}) \{ Z := \text{new } \mathcal{T}[m] ; n := 0 ; k := 0 \}$ // create an empty queue $+ \text{add}(x : \mathcal{T})$ // join x to the end of the queue $+ \text{rem}() : \mathcal{T}$ // remove and return the first element of the queue $+ \text{first}() : \mathcal{T}$ // return the first element of the queue $+ \text{length}() : \mathbb{N} \{ \text{return } n \}$ $+ \text{isFull}() : \mathbb{B} \{ \text{return } n = Z.length \}$ $+ \text{isEmpty}() : \mathbb{B} \{ \text{return } n = 0 \}$ $+ \sim \text{Queue}() \{ \text{delete } Z \}$ $+ \text{setEmpty}() \{ n := 0 \}$ // reinitialize the queue

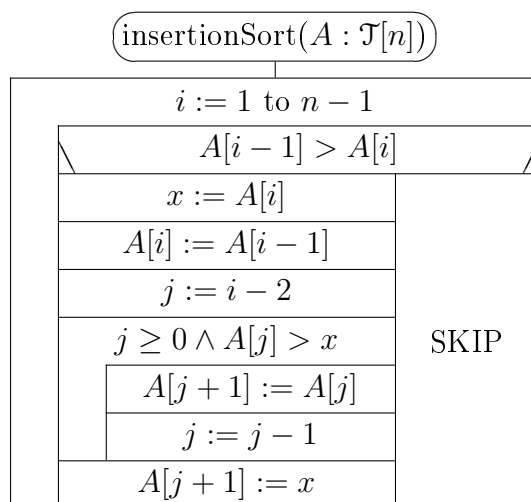


We described the methods of stacks and queues with simple codes: we applied neither iterations nor recursions. Therefore the time complexity of each method is $\Theta(1)$. This is a fundamental requirement for each implementation of stacks and queues.

Note that with linked list representations this constraint can be guaranteed only if $MT_{\text{new}}, MT_{\text{delete}} \in \Theta(1)$.

3 Algorithms in computing: Insertion Sort

[1] Chapter 1-3; [4] Chapter 1, 2, 7



$$mT_{IS}(n) = 1 + (n - 1) = n$$

$$MT_{IS}(n) = 1 + (n-1) + \sum_{i=1}^{n-1} (i-1) = n + \sum_{j=0}^{n-2} j = n + \frac{(n-1) * (n-2)}{2}$$

$$MT_{IS}(n) = \frac{1}{2}n^2 - \frac{1}{2}n + 1$$

$$MT_{IS}(n) \approx (1/2) * n^2, \text{ if } n \text{ is large.}$$

Let us suppose that our computer can perform $2 * 10^9$ elementary operations /second.

Considering the code of insertion sort above, and counting these operations as a function of n , we receive that $mT(n) \geq 8 * n$ and $MT(n) > 6 * n^2$ elementary operations. Counting with $mT(n) \approx 8 * n$ and $MT(n) \approx 6 * n^2$, we receive the following table on running times as lower estimates:

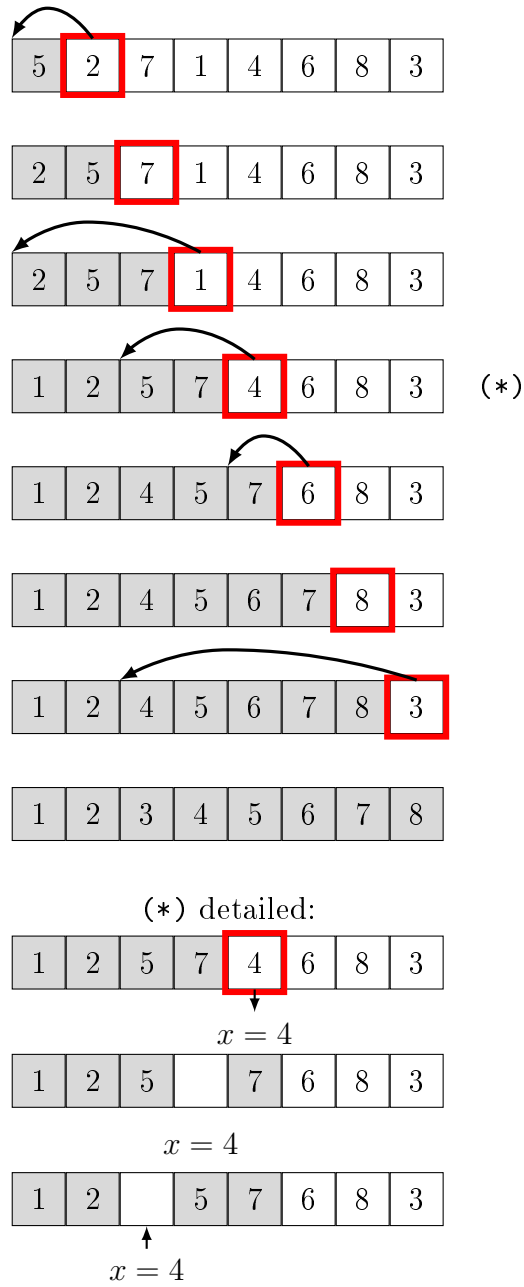


Figure 1: An illustration of Insertion Sort

n	$mT(n)$	in secs	$MT(n)$	in time
1000	8000	$4 * 10^{-6}$	$6 * 10^6$	0.003 sec
10^6	$8 * 10^6$	0.004	$6 * 10^{12}$	50 min
10^7	$8 * 10^7$	0.04	$6 * 10^{14}$	≈ 3.5 days
10^8	$8 * 10^8$	0.4	$6 * 10^{16}$	≈ 347 days
10^9	$8 * 10^9$	4	$6 * 10^{18}$	≈ 95 years

This means that in the worst case insertion sort is too slow to sort one million element, and it becomes completely impractical, if we try to sort huge amount of data. Let us consider the average case:

$$\begin{aligned}
AT_{IS}(n) &\approx 1 + (n-1) + \sum_{i=1}^{n-1} \left(\frac{i-1}{2} \right) = n + \frac{1}{2} * \sum_{j=0}^{n-2} j = \\
&= n + \frac{1}{2} * \frac{(n-1) * (n-2)}{2} = \frac{1}{4}n^2 + \frac{1}{4}n + \frac{1}{2}
\end{aligned}$$

This calculation shows, that the expected or average running time of insertion sort is roughly the half of the time needed in the worst case, so even the expected running time of insertion sort is too long to sort one million element, and it becomes completely impractical, if we try to sort huge amount of data. The asymptotic time complexities:

$$\begin{aligned}
mT_{IS}(n) &\in \Theta(n) \\
AT_{IS}(n), MT_{IS}(n) &\in \Theta(n^2)
\end{aligned}$$

Let us notice that the minimum running time is very good. There is no chance to sort elements faster than in linear time, because each piece of data must be checked. One may say that this best case is not much gain, because in this case the items are already sorted. The fact is, that if the input is *nearly sorted*, we can remain close to the best case, and insertion sort turns out to be the best to sort such data.

Insertion sort is also *stable*: Stable sorting algorithms maintain the relative order of records with equal keys (i.e. values). That is, a sorting algorithm is stable if whenever there are two records R and S with the same key and with R appearing before S in the original list, R will appear before S in the sorted list. (Stability is an important property of sorting methods in some applications.)

4 Fast sorting algorithms based on the *divide and conquer* approach

In computer science, divide and conquer is an algorithm design paradigm based on multi-branched recursion. A divide and conquer algorithm works by recursively breaking down a problem into two or more sub-problems of the same or related type, until these become simple enough to be solved directly. The solutions to the sub-problems are then combined to give a solution to the original problem.

This divide and conquer technique is the basis of efficient algorithms many kinds of problems, for example sorting (e.g., quicksort, mergesort).

4.1 Merge sort

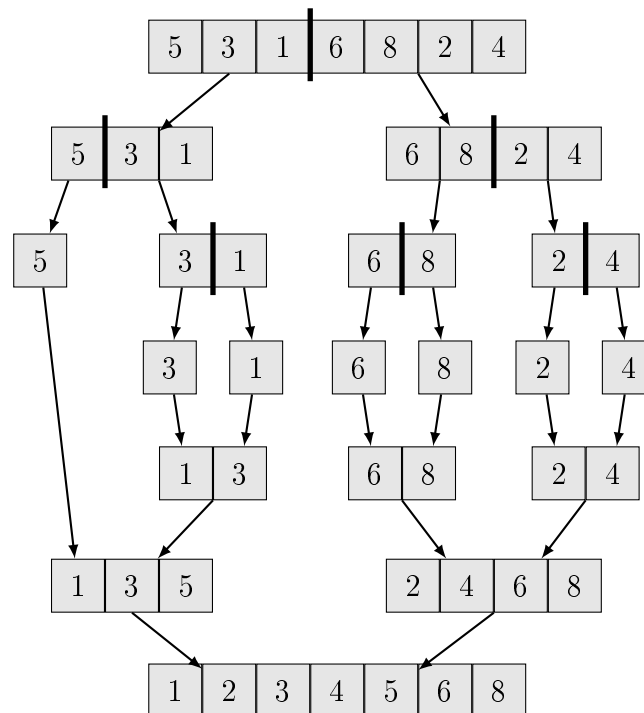


Figure 2: An illustration of merge sort.

The divide and conquer paradigm is often used to find the optimal solution of a problem. Its basic idea is to decompose a given problem into two or more similar, but simpler, subproblems, to solve them in turn, and to compose their solutions to solve the given problem. Problems of sufficient

simplicity are solved directly. For example, to sort a given list of n keys, split it into two lists of about $n/2$ keys each, sort each of them in turn, and interleave (i.e. merge) both results appropriately to obtain the sorted version of the given list. (See Figure 2.) This approach is known as the merge sort algorithm.

Merge sort is stable (preserving the input order of items with equal keys) and its worst case time complexity is asymptotically optimal among comparison sorts. (See section 7 for details.)

$\text{mergeSort}(A : \mathcal{T}[n])$
$B : \mathcal{T}[n] ; B[0..n) := A[0..n)$
// Using $B[0..n)$ sort $A[0..n)$ non-decreasingly:
$\text{ms}(B, A, 0, n)$

$\text{ms}(B, A : \mathcal{T}[] ; u, v : \mathbb{N})$
// Initially $B[u..v) = A[u..v)$.
// Using $B[u..v)$ sort $A[u..v)$ non-decreasingly:
$v - u > 1$
$m := \lfloor \frac{u+v}{2} \rfloor$
$\text{ms}(A, B, u, m)$ // Using $A[u..m)$ sort $B[u..m)$
$\text{ms}(A, B, m, v)$ // Using $A[m..v)$ sort $B[m..v)$
$\text{merge}(B, A, u, m, v)$ // merge $B[u..m)$ and $B[m..v)$ into $A[u..v)$
SKIP

Using $m = \lfloor \frac{u+v}{2} \rfloor$, $A[u..m)$ and $A[m..v)$ have the same length, if the length of $A[u..v)$ is even number; and $A[u..m)$ is shorter by one than $A[m..v)$, if the length of $A[u..v)$ is odd number; because

$$\text{length}(A[u..m)) = m - u = \left\lfloor \frac{u+v}{2} \right\rfloor - u =$$

$$\left\lfloor \frac{u+v}{2} - u \right\rfloor = \left\lfloor \frac{v-u}{2} \right\rfloor = \left\lfloor \frac{\text{length}(A[u..v))}{2} \right\rfloor$$

merge($B, A : \mathcal{T}[] ; u, m, v : \mathbb{N}$)	
// sorted merge of $B[u..m)$ and $B[m..v)$ into $A[u..v)$	
$k := u$ // in loop, copy into $A[k]$	
$i := u ; j := m$ // from $B[i]$ or $B[j]$	
$i < m \wedge j < v$	
$B[i] \leq B[j]$	
$A[k] := B[i]$	$A[k] := B[j]$
$i := i + 1$	$j := j + 1$
$k := k + 1$	
$i < m$	
$A[k..v) := B[i..m)$	$A[k..v) := B[j..m)$

The stability of merge sort is ensured in the explicit loop of merge, because in case of $B[i] = B[j]$, $B[i]$ is copied into $A[k]$, and $B[i]$ came earlier in the input.

Merge makes l loop iterations where $l = v - u$ is the length of the actual subarrays $A[u..v)$ and $B[u..v)$. In order to prove this statement consider the value of k after the explicit loop of merge. Clearly, this loop fills subarray $A[u..k)$, and it copies one element / iteration. Thus it iterates $k - u$ times. In addition, the implicit loop hidden into $A[k..v) := \dots$ iterates $v - k$ times. Therefore the sum of the loop iterations is $(k - u) + (v - k) = v - u = l$. Consequently for the body of procedure merge (mb) we have:

$$T_{\text{mb}}(l) = l \quad (l = v - u)$$

4.1.1 The efficiency of merge sort

Merge sort is one of the fastest sorting algorithms and there is not a big difference between its worst-case and best-case (i.e. maximal and minimal) running time. We state:

$$MT_{\text{mergeSort}}(n), mT_{\text{mergeSort}}(n) \in \Theta(n \log n)$$

Proof: Clearly $T_{\text{mergeSort}}(0) = 2$. We suppose that $n > 0$. First we count all the loop iterations of procedure merge. Next we count the procedure calls of merge sort.

Loop iterations: We proved above that a single call of merge(B, A, u, m, v) makes $T_{\text{mb}}(l) = l$ iterations where $l = v - u$. Let us consider the levels of recursion of procedure ms(B, A, u, v). At level 0 of the recursion ms is

called for the whole array $A[0..n]$. Considering all the recursive calls and the corresponding subarrays at a given depth of recursion, at level 1 this array is divided into two halves, at level 2 into 4 parts and so on. Let n_{ij} be the length of the j th subarray of the form $A[u..v]$ at level i , and let $m = \lfloor \log n \rfloor$. We have:

At level 0: $2^m \leq n_{01} = n < 2^{m+1}$
 At level 1: $2^{m-1} \leq n_{1j} \leq 2^{m+1-1}$ ($j \in [1..2^1]$)
 At level 2: $2^{m-2} \leq n_{2j} \leq 2^{m+1-2}$ ($j \in [1..2^2]$)
 ...
 At level i : $2^{m-i} \leq n_{ij} \leq 2^{m+1-i}$ ($i \in [1..m], j \in [1..2^i]$)
 ...
 At level $m-1$: $2 \leq n_{(m-1)j} \leq 4 = 2^2$ ($j \in [1..2^{m-1}]$)
 At level m : $1 \leq n_{mj} \leq 2 = 2^1$ ($j \in [1..2^m]$)
 At level $m+1$: $n_{(m+1)j} = 1$ ($j \in [1..(n-2^m)]$)

Thus at levels $[0..m]$ these subarrays cover the whole array. At levels $[0..m]$ merge is called for each subarray, but at level m it is called only for those subarrays with length 2, and the number of these subarrays is clearly $n - 2^m$. Merge makes as many iterations as the length of the actual $A[u..v]$. Consequently at each level in $[0..m)$ merge makes n iterations in all the merge calls of the level altogether. At level m the sum of the iterations is $2 * (n - 2^m)$, and there is no iteration at level $m+1$. Therefore the sum total of all the iterations during merge sort is

$$T_{mb[0..m]}(n) = n * m + 2 * (n - 2^m).$$

The number of procedure calls: Clearly, the ms calls form a strictly binary tree. (See section 6.2.) The leaves of this tree correspond to the subarrays with length 1. Thus this strictly binary tree has n leaves and $n - 1$ internal nodes. Consequently we have $2n - 1$ calls of ms, and $n - 1$ calls of merge. Adding to this the single call of mergeSort we receive $3n - 1$ procedure calls altogether. Thus the number of steps of mergeSort is

$$T(n) = (n * m + 2 * (n - 2^m)) + (3 * n - 1) = n * \lfloor \log n \rfloor + 2 * (n - 2^{\lfloor \log n \rfloor}) + 3 * n - 1$$

$$n * \log n + n \leq n * (\log n - 1) + 3 * n - 1 \leq T(n) < n * \log n + 5 * n.$$

Thus $T(n) \in \Omega(n \log n)$ (see Consequence 1.8) and $T(n) \in O(n \log n)$ (see Consequence 1.7). After all we have for mergeSort($A : \mathcal{T}[n]$)

$$T(n) \in \Theta(n \log n).$$

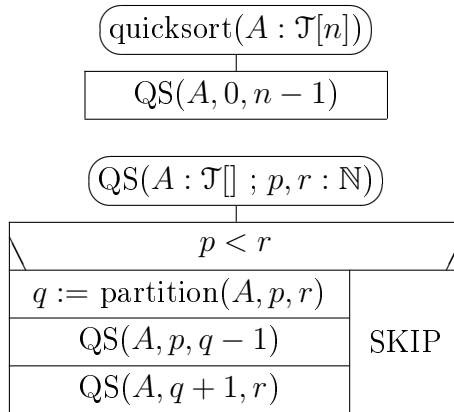
4.2 Quicksort

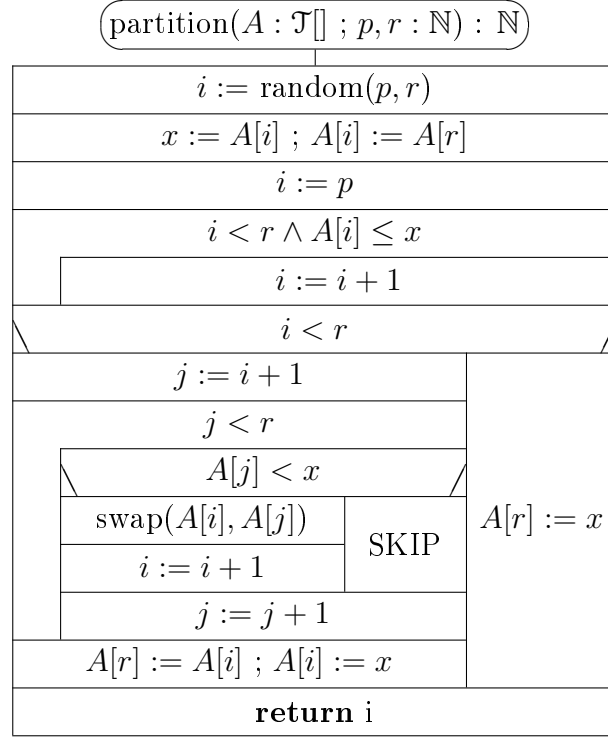
Quicksort is a divide and conquer algorithm. Quicksort first divides a large array into two smaller sub-arrays: the low elements and the high elements. Quicksort can then recursively sort the sub-arrays.

The steps are:

- Pick an element, called a pivot, from the array.
- Partitioning: reorder the array so that all elements with values less than the pivot come before the pivot, while all elements with values greater than the pivot come after it (equal values can go either way). After this partitioning, the pivot is in its final position. This is called the partition operation.
- Recursively apply the above steps to the sub-array of elements with smaller values and separately to the sub-array of elements with greater values.
- The base case of the recursion is arrays of size zero or one, which are in order by definition, so they never need to be sorted.

The pivot selection and partitioning steps can be done in several different ways; the choice of specific implementation schemes greatly affects the algorithm's performance.





Explanation of function partition: In order to illustrate the operation of function partition, let us introduce the following notation:

- $A[k..m] \leq x$, iff
for each $l, k \leq l \leq m$ implies $A[l] \leq x$
- $A[k..m] \geq x$, iff
for each $l, k \leq l \leq m$ implies $A[l] \geq x$

We suppose that subarray $A[p..r]$ is cut into pieces and the pivot is the second 5, i.e. the 4. element of this subarray (with index $p+3$).

Input:

	p							r
A :	5	3	8	5	6	4	7	1

Preparations for the first loop:

	p							r	
A :	5	3	8	1	6	4	7		$x = 5$

The first loop searches for the first item being greater than the pivot (x), if there is such element.

	i=p	i	i					r	
A :	5	3	8	1	6	4	7		$x = 5$

We have found an item greater than the pivot. Variable j starts at the next item. From this moment $p \leq i < j \leq r$. We have cut $A[p..r]$ into four sections:

$A[p..(i-1)]$, $A[i..(j-1)]$, $A[j..(r-1)]$, and $A[r]$.

We have the following properties.

$A[p..(i-1)] \leq x \wedge A[i..(j-1)] \geq x$, $A[j..(r-1)]$ is unknown and $A[r]$ is undefined. Together with $p \leq i < j \leq r$ this is an invariant of the second loop. (Notice that items equal to the pivot may be either in the first, and in the second section of $A[p..r]$.) Starting the run of the second loop, it turns out that $A[j]$ must be exchanged with $A[i]$ in order to add it to the first section of $A[p..r]$ containing items \leq than the pivot. Based on the invariant $A[i] \geq x$, it can go to the end of the second section (containing items \geq than the pivot). (The items to be exchanged are printed in bold.)

	p		i	j				r	
A :	5	3	8	1	6	4	7		$x = 5$

Now we exchange $A[i]$ and $A[j]$. Thus the length of the first section of $A[p..r]$ (containing items \leq than the pivot) has been increased by 1, while its second section (containing items \geq than the pivot) has been moved by one position. So we increment variables i and j by 1, in order to keep the invariant of the loop.

Now $A[j] = 6 \geq x = 5$, so we add $A[j]$ to the second section of $A[p..r]$, i.e. we increment j by 1.

And now $A[j] = 4 < x = 5$, so $A[j]$ must be exchanged with $A[i]$. (The items to be exchanged are printed in bold.)

	p			i	j	j		r	
A :	5	3	1	8	6	4	7		$x = 5$

Now we exchange $A[i]$ and $A[j]$. Thus the length of the first section of $A[p..r]$ (containing items \leq than the pivot) has been increased by 1, while its second section (containing items \geq than the pivot) has been moved by one position. So we increment variables i and j by 1, in order to keep the invariant of the loop.

Now $A[j] = 7 \geq x = 5$, so we add $A[j]$ to the second section of $A[p..r]$, i.e. we increment j by 1.

And now $j = r$, therefore the third section of $A[p..r]$ disappeared.

	p				i		j	j=r	
A :	5	3	1	4	6	8	7		$x = 5$

Now the first two sections cover $A[p..(r-1)]$, and the second section is not empty because of invariant $i < j$. Thus the pivot can be put in between the

items \leq than it, and the items \geq than it: first we move the first element of the second section to the end of $A[p..r]$, next we put the pivot into its previous place. (The pivot is printed in bold.)

	p				i			j=r
A :	5	3	1	4	5	8	7	6

Now the partitioning of subarray $A[p..r]$ has been completed. We return the position of the pivot (i), in order to inform procedure $\text{Quicksort}(A, p, r)$, which subarrays of $A[p..r]$ must be sorted recursively.

In the other case of procedure partition the pivot is a maximum of $A[p..r]$. This case is trivial. Let the reader consider it.

The time complexity of function partition is linear, because the two loops together perform $r-p-1$ or $r-p$ iterations.

$$mT_{\text{quicksort}}(n), AT_{\text{quicksort}}(n) \in \Theta(n \log n)$$

$$MT_{\text{quicksort}}(n) \in \Theta(n^2)$$

It is known that on short arrays insertion sort is more efficient than the fast sorting methods (merge sort, heap sort, quicksort). Thus procedure $\text{quicksort}(A, p, r)$ can be optimized by switching to insertion sort on short arrays. Because during partitioning we create a lot of short subarrays, the speed-up can be significant.

$\text{QS}(A : \mathcal{T}[] ; p, r : \mathbb{N})$	
$p + c < r$	
$q := \text{partition}(A, p, r)$	$\text{insertionSort}(A, p, r)$
$\text{QS}(A, p, q - 1)$	
$\text{QS}(A, q + 1, r)$	

$c \in \mathbb{N}$ is a constant. Its optimal value depends on many factors, but usually it is between 20 and 40.

Exercise 4.1 *How to speed-up merge sort in a similar way?*

Considering its expected running time, quicksort is one of the most efficient sorting methods. However, it slows down, if (for example) by chance function partition always selects the maximal element of the actual subarray. (Its time complexity is $\Theta(n^2)$ in this case.) The probability of such cases is low. However, if we want to avoid such cases, we can pay attention to the depth of recursion, and switch to heap sort or merge sort when recursion becomes too deep (for example, deeper than $2 * \log n$).

5 Linked Lists

One-way lists can represent finite sequences of data. They consist of zero or more elements (i.e. nodes) where each element of the list contains some data and linking information.

5.1 One-way or singly linked lists

In singly linked lists each element contains a *key*, and a *next* pointer referring to the next element of the list. A pointer referring to the front of the list identifies the list. The *next* pointer of the last element typically contains a so called NULL (i.e. \oslash) pointer¹ which is the address of no object. (With some abstraction, the nonexisting object with address \oslash can be called NO object.) In this book E1 is the element type of one-way lists.

E1
+ <i>key</i> : \mathcal{T} ... // satellite data may come here + <i>next</i> : E1 *
+ E1 () { <i>next</i> := \oslash }

5.1.1 Simple one-way lists (S1L)

An S1L is identified by a pointer referring to its first element, or this pointer is \oslash , if the list is empty.

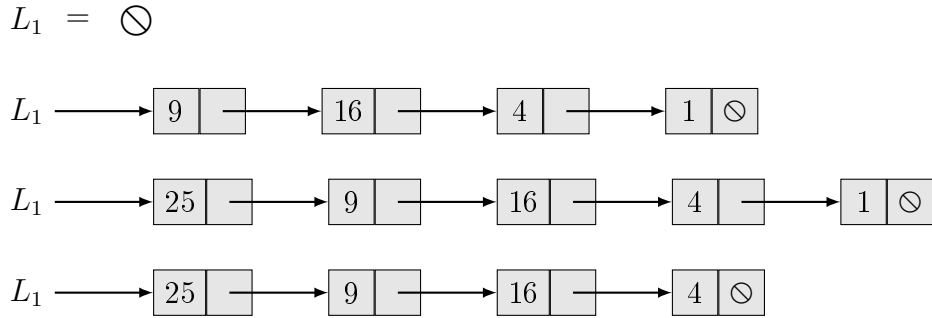
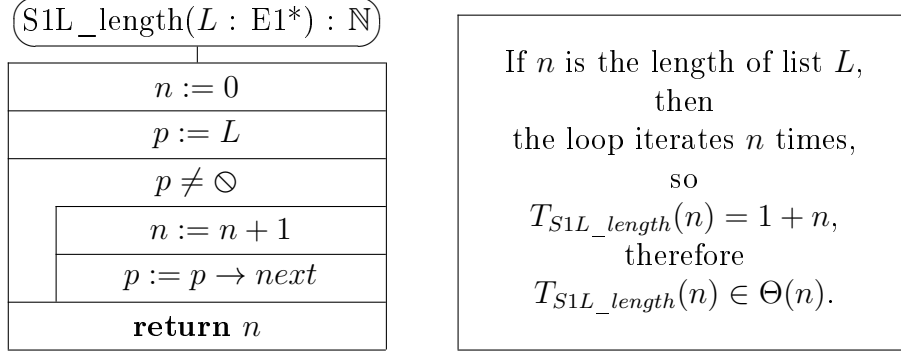


Figure 3: Pointer L_1 identifies simple one-way lists at different stages of an imaginary program. In the first line the S1L is empty.

¹except in cyclic lists



5.1.2 One-way lists with header node (H1L)

The header node or simply header of a list is an extra *zeroth* element of the list with undefined *key*. A list with header cannot be \ominus , it always contains a header. The pointer identifying the list always refers to its header.

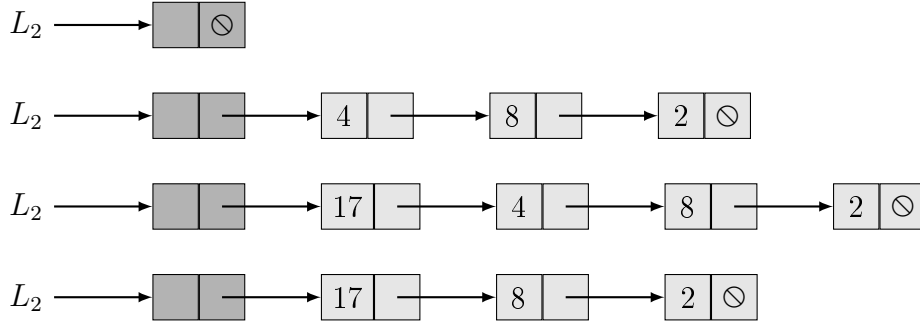
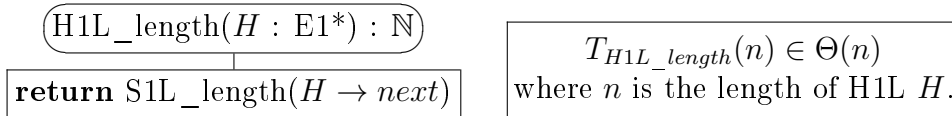


Figure 4: Pointer L_2 identifies one-way lists with heather at different stages of an imaginary program. In the first line the H1L is empty.

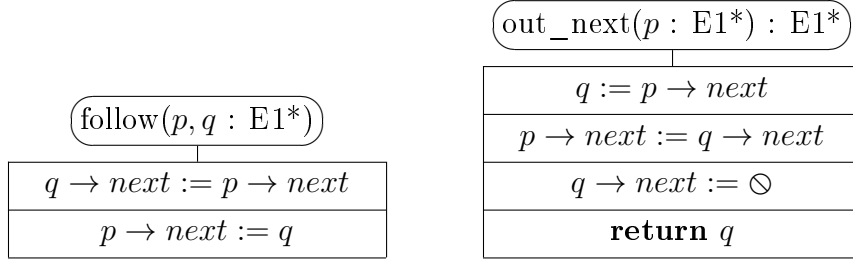


5.1.3 One-way lists with trailer node

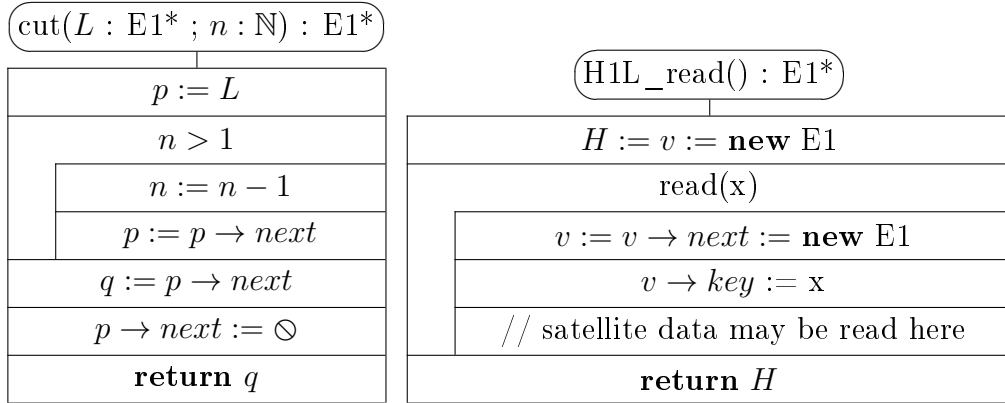
Some one-way lists contain a trailer node instead of header. A trailer node is always at the end of such lists. The data members (*key* and *next*) of the trailer node are typically undefined, and we have two pointers identifying the list: One of them refers to its first element, and the other to its trailer node. If the list is empty, both identifier pointers refer to its trailer node.

Such a list is ideal for representing a queue. Its method $\text{add}(x : \mathcal{T})$ copies x into the *key* of the trailer node, and joins a new trailer node to the end of the list.

5.1.4 Handling one-way lists



$$T_{\text{follow}}, T_{\text{out_next}} \in \Theta(1)$$

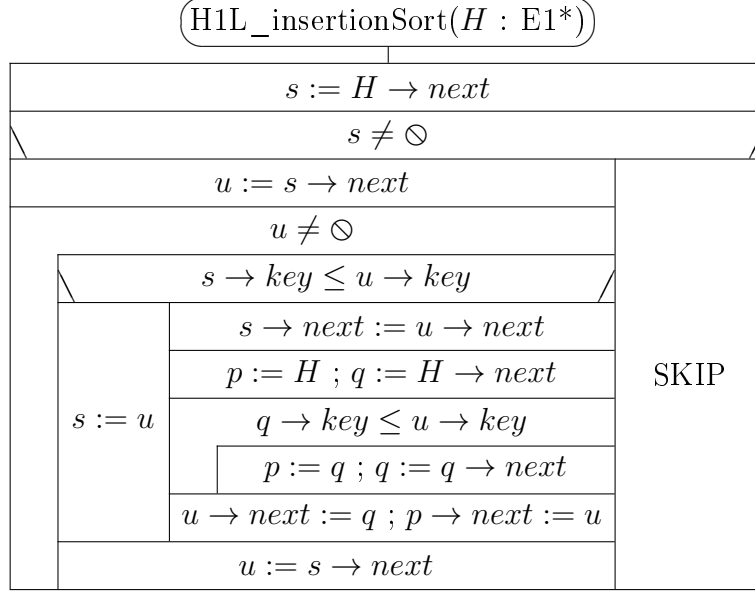


$$T_{\text{cut}}(n) \in \Theta(n)$$

$$T_{\text{H1L_read}}(n) \in \Theta(n)$$

where n is the length of the list.

5.1.5 Sorting one-way lists: Insertion sort



$$mT_{IS}(n) \in \Theta(n)$$

$$AT_{IS}(n), MT_{IS}(n) \in \Theta(n^2)$$

where n is the length of H1L H . Clearly procedure insertionSort($H : E1^*$) is stable.

5.1.6 Cyclic one-way lists

The last element of the list does not contain \ominus in its *next* field, but this pointer points back to the beginning of the list.

If a cyclic one-way list does not contain header node, and it is not empty, the *next* field of its last element points back to its first element. If it is empty, it is represented by the \ominus pointer. A pointer identifying a nonempty, cyclic one-way list typically refers to its last element.

If a cyclic one-way list has a header node, the *next* field of its last element points back to its header node. If it is empty, the *next* field of its header node points to itself. Notice that the header of a cyclic list is also the trailer node of that list.

Such lists are also good choices for representing queues. Given a queue represented by a cyclic list with header the method add($x : \mathcal{T}$) copies x into the *key* of the trailer/header node, and inserts a new trailer/header node into the list.

5.2 Two-way or doubly linked lists

In a two-way list each element contains a *key*, and two pointers: a *next* and a *prev* pointer referring to the next and previous elements of the list. A pointer referring to the front of the list identifies the list.

5.2.1 Simple two-way lists (S2L)

An S2L is identified by a pointer referring to its first element, or this pointer is \ominus , if the list is empty.

The *prev* pointer of the first element, and the *next* pointer of the last element contain the \ominus pointer.

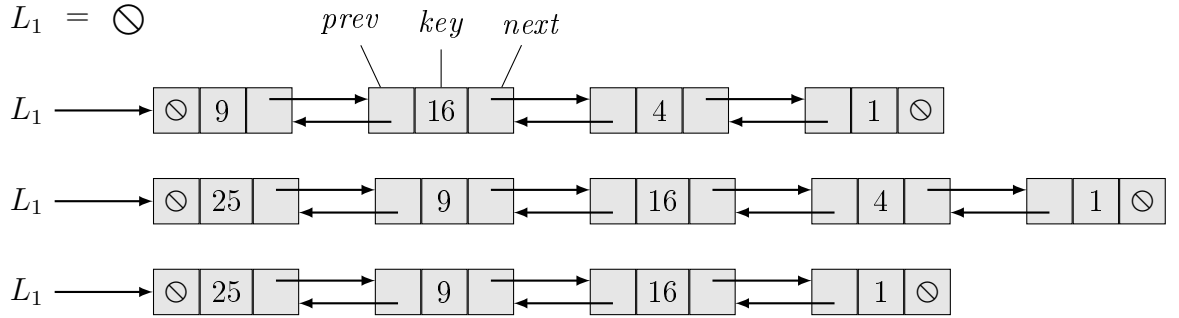


Figure 5: Pointer L_1 identifies simple two-way lists (S2L) at different stages of an imaginary program. In the first line list L_1 is initialized to empty.

Handling of S2Ls is inconvenient because all the modifications of an S2L must be done differently at the front of a list, at the end of the list, and in between. Consequently we prefer cyclic two-way lists (C2L) in general. (In hash tables, however S2Ls are usually preferred to C2Ls.)

5.2.2 Cyclic two-way lists (C2L)

A cyclic two-way list (C2L) contains a header node (i.e. header) by default, and it is identified by a pointer referring to its header. The *next* field of its last element points back to its header which is considered a zeroth element. The *prev* pointer of its first element points back to its header, and the *prev* pointer of its header points back to its last element. If a C2L is empty, both of the *prev*, and *next* fields of its header point to itself. Notice that the header of a cyclic list is also the trailer node of the list.

The element type of C2Ls follows.

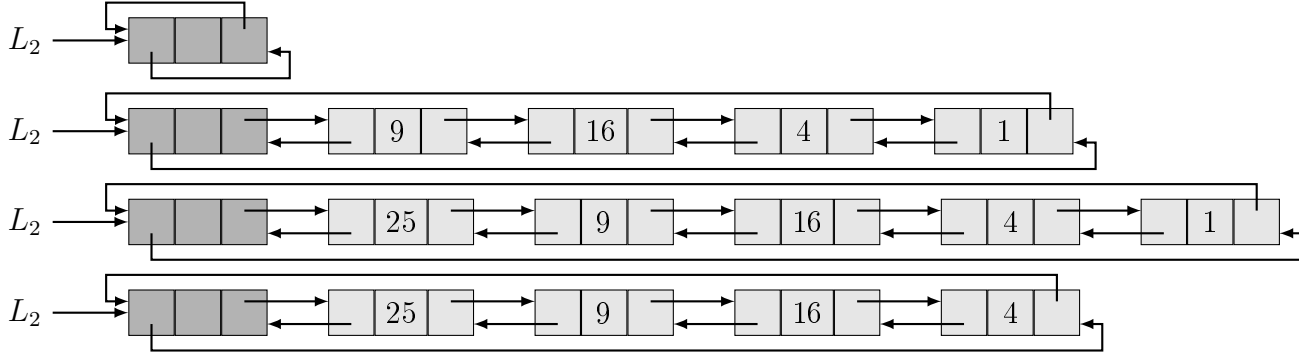


Figure 6: Pointer L_2 identifies cyclic two-way lists (C2L) at different stages of an imaginary program. In the first line list L_2 is empty.

E2
$+prev, next : \mathbf{E2}^*$ // refer to the previous and next neighbour or be this
$+key : \mathcal{T}$
$+ E2() \{ prev := next := \mathbf{this} \}$

Some basic operations on C2Ls follow. Notice that these are simpler than the appropriate operations of S2Ls.

$\text{precede}(q, r : \mathbf{E2}^*)$	$\text{follow}(p, q : \mathbf{E2}^*)$
// (*q) will precede (*r)	// (*q) will follow (*p)
$p := r \rightarrow prev$	$r := p \rightarrow next$
$q \rightarrow prev := p ; q \rightarrow next := r$	$q \rightarrow prev := p ; q \rightarrow next := r$
$p \rightarrow next := r \rightarrow prev := q$	$p \rightarrow next := r \rightarrow prev := q$

$\text{out}(q : \mathbf{E2}^*)$
// remove (*q)
$p := q \rightarrow prev ; r := q \rightarrow next$
$p \rightarrow next := r ; r \rightarrow prev := p$
$q \rightarrow prev := q \rightarrow next := q$

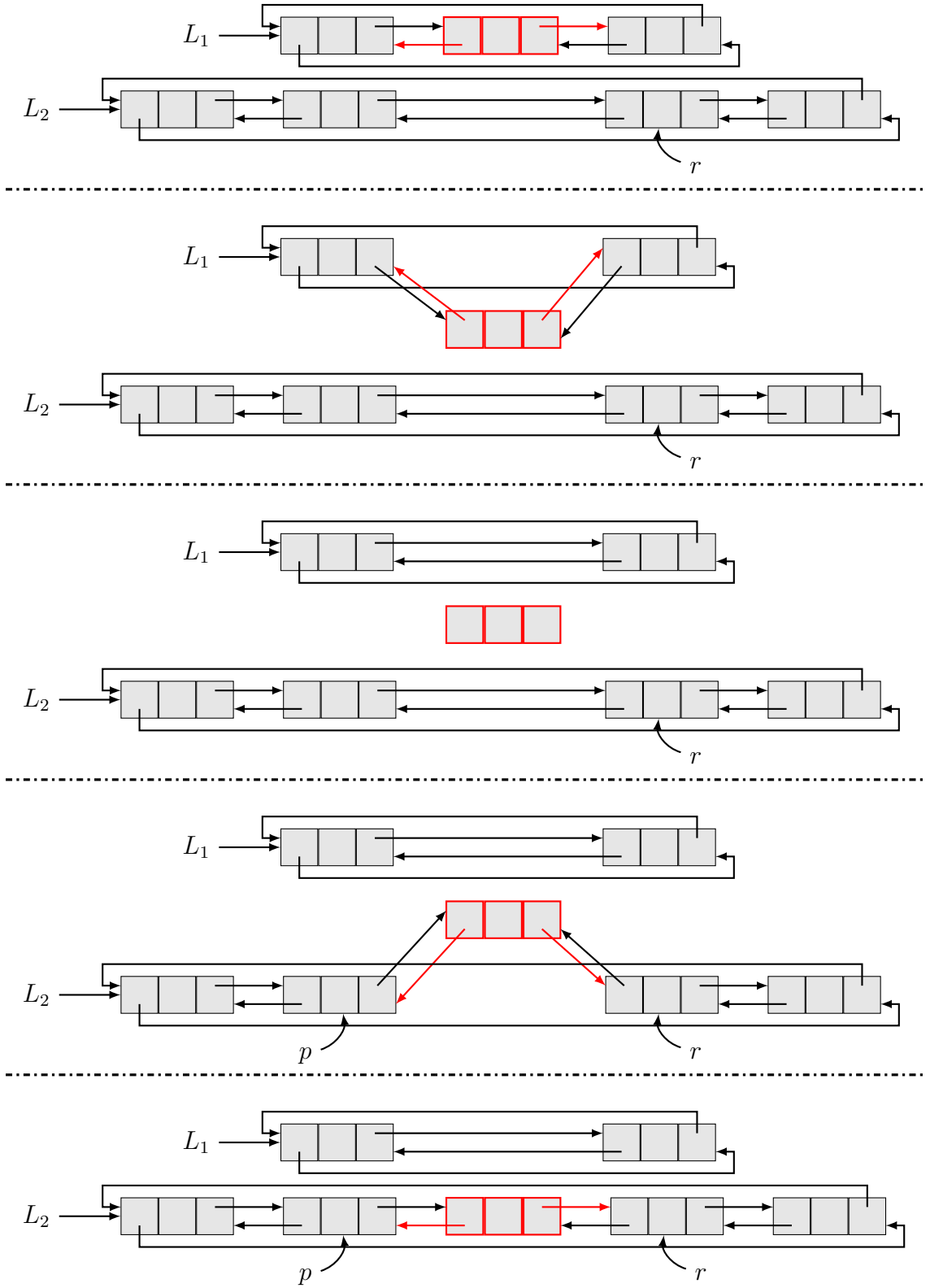


Figure 7: Illustrating $\text{out}(q)$ and $\text{precede}(q, r)$. Insertion is called on the red element ($*q$), and it is inserted left to ($*r$) into list L_2 .

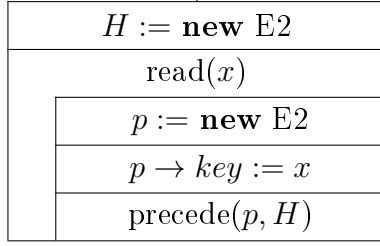
5.2.3 Example programs on C2Ls

We can imagine a C2L straightened, for example

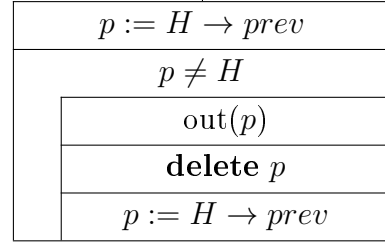
$H \rightarrow [/] - [5] - [2] - [7] - [/] \leftarrow H$ representing $\langle 5; 2; 7 \rangle$

or empty: $H \rightarrow [/] - [/] \leftarrow H$ representing $\langle \rangle$.

$\text{C2L_read}(\&H : \text{E2}^*)$



$\text{setEmpty}(H : \text{E2}^*)$



$H \rightarrow [/] - [/] \leftarrow H$

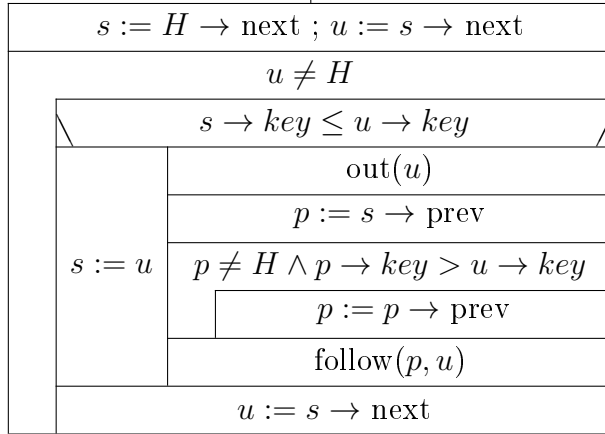
$H \rightarrow [/] - [5] - [/] \leftarrow H$

$H \rightarrow [/] - [5] - [2] - [/] \leftarrow H$

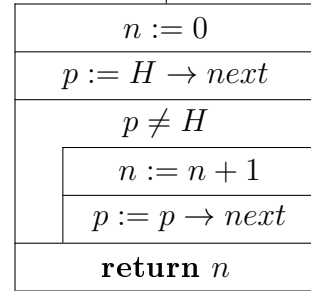
$H \rightarrow [/] - [5] - [2] - [7] - [/] \leftarrow H$

$$T_{\text{C2L_read}}(n), T_{\text{setEmpty}}(n), T_{\text{length}}(n) \in \Theta(n)$$

$\text{insertionSort}(H : \text{E2}^*)$



$\text{length}(H : \text{E2}^*) : \mathbb{N}$



$H \rightarrow [/] - [5] - [2] - [7] - [2] - [/] \leftarrow H$

$H \rightarrow [/] - [2] - [5] - [7] - [2] - [/] \leftarrow H$

$H \rightarrow [/] - [2] - [5] - [7] - [2] - [/] \leftarrow H$

$H \rightarrow [/] - [2] - [2] - [5] - [7] - [/] \leftarrow H$

$$mT_{IS}(n) \in \Theta(n); AT_{IS}(n), MT_{IS}(n) \in \Theta(n^2)$$

where n is the length of C2L H . Clearly procedure $\text{insertionSort}(H : E2^*)$ is stable.

Example 5.1 *Let us suppose that H_u and H_i are strictly increasing C2Ls. We write procedure $\text{unionIntersection}(H_u, H_i : E2^*)$ which calculates the strictly increasing union of the two lists in H_u and the strictly increasing intersection of them in H_i .*

*We use neither memory allocation (**new**) nor deallocation (**delete**) statements nor explicit assignment statements to data members. We rearrange the lists only with $\text{out}(q)$, $\text{precede}(q, r)$, $\text{follow}(p, q)$. $MT(n, m) \in O(n + m)$ where $n = \text{length}(H_u)$ and $m = \text{length}(H_i)$.*

Let us have $q, r : E2^*$ initialized to point to the first items of H_u and H_i respectively. For example:

$$\begin{aligned} H_u &\rightarrow [/] - \overset{q}{[2]} - [4] - [6] - [/] \leftarrow H_u \\ H_i &\rightarrow [/] - \underset{r}{[1]} - [4] - [8] - [9] - [/] \leftarrow H_i \end{aligned}$$

We work with the following invariant where

$$\text{key}(q, H) = \begin{cases} q \rightarrow \text{key} & \text{if } q \neq H \\ \infty & \text{if } q = H \end{cases}$$

(H, q) is the sublist of the items between H and q ,

$[q, H)$ is the sublist of the items starting with q but before H .

- C2Ls H_u and H_i are strictly increasing consisting of the original bag of list items, q is pointer on H_u , and r on H_i ,
- (H_u, q) is the prefix of the sorted union containing the keys less than $\min(\text{key}(q, H_u), \text{key}(r, H_i))$,
- (H_i, r) is the prefix of the sorted intersection containing the keys less than $\min(\text{key}(q, H_u), \text{key}(r, H_i))$,
- $[q, H_u)$ and $[r, H_i)$ are still unaltered.

Illustration of the run of the program:

$$\begin{aligned} H_u &\rightarrow [/] - [1] - \overset{q}{[2]} - [4] - [6] - [/] \leftarrow H_u \\ H_i &\rightarrow [/] - \underset{r}{[4]} - [8] - [9] - [/] \leftarrow H_i \\ \\ H_u &\rightarrow [/] - [1] - [2] - \overset{q}{[4]} - [6] - [/] \leftarrow H_u \\ H_i &\rightarrow [/] - \underset{r}{[4]} - [8] - [9] - [/] \leftarrow H_i \end{aligned}$$

$$H_u \rightarrow [/] - [1] - [2] - [4] - [\overset{q}{6}] - [/] \leftarrow H_u$$

$$H_i \rightarrow [/] - [4] - [\underset{r}{8}] - [9] - [/] \leftarrow H_i$$

$$H_u \rightarrow [/] - [1] - [2] - [4] - [6] - [\overset{q}{/}] \leftarrow H_u$$

$$H_i \rightarrow [/] - [4] - [\underset{r}{8}] - [9] - [/] \leftarrow H_i$$

$$H_u \rightarrow [/] - [1] - [2] - [4] - [6] - [8] - [\overset{q}{/}] \leftarrow H_u$$

$$H_i \rightarrow [/] - [4] - [\underset{r}{9}] - [/] \leftarrow H_i$$

$$H_u \rightarrow [/] - [1] - [2] - [4] - [6] - [8] - [9] - [\overset{q}{/}] \leftarrow H_u$$

$$H_i \rightarrow [/] - [4] - [\underset{r}{/}] \leftarrow H_i$$

unionIntersection($H_u, H_i : E2*$)		
$q := H_u \rightarrow next ; r := H_i \rightarrow next$		
$q \neq H_u \wedge r \neq H_i$		
$\backslash q \rightarrow key < r \rightarrow key$	$\backslash q \rightarrow key > r \rightarrow key$	$\backslash q \rightarrow key = r \rightarrow key$
$q := q \rightarrow next$	$p := r$	$q := q \rightarrow next$
	$r := r \rightarrow next$	
	out(p)	$r := r \rightarrow next$
	precede(p, q)	
$r \neq H_i$		
$p := r ; r := r \rightarrow next ; out(p)$		
precede(p, H_u)		

Exercise 5.2 Write the structograms of $quickSort(H:E2*)\{ QS(H, H) \}$ which sort C2L H increasingly. $QS(p, s:E2*)$ can be a recursive procedure sorting sublist (p, s) with quicksort. It starts with partitioning sublist (p, s) . Let the pivot (q) be the first element of the sublist. Let your version of quicksort be stable.

Illustration of the partition part of $QS(H, H)$ on C2L H below:
(Partition of the sublist strictly between p and s , i.e. partition of sublist (p, s) . The pivot is q , r goes on sublist (q, s) , elements smaller then the pivot are moved before the pivot.)

$$H \rightarrow [\overset{p}{/}] - [\underset{q}{5}] - [\underset{r}{2}] - [4] - [6] - [5] - [2] - [\overset{s}{/}] \leftarrow H$$

$$H \rightarrow \overset{p}{[/]} - [2] - [5] - \underset{q}{[4]} - \underset{r}{[6]} - [5] - [2] - \overset{s}{[/]} \leftarrow H$$

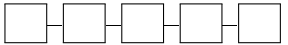
$$H \rightarrow \overset{p}{[/]} - [2] - [4] - \underset{q}{[5]} - \underset{r}{[6]} - [5] - [2] - \overset{s}{[/]} \leftarrow H$$

$$H \rightarrow \overset{p}{[/]} - [2] - [4] - [5] - [6] - \underset{q}{[5]} - \underset{r}{[2]} - \overset{s}{[/]} \leftarrow H$$

$$H \rightarrow \overset{p}{[/]} - [2] - [4] - \underset{q}{[5]} - [6] - [5] - \underset{r}{[2]} - \overset{s}{[/]} \leftarrow H$$

$$H \rightarrow \overset{p}{[/]} - [2] - [4] - [2] - \underset{q}{[5]} - [6] - [5] - \underset{r}{[/]} \leftarrow H$$

6 Trees, binary trees

Up till now, we worked with arrays and linked lists. They have a common property: There is a first item in the data structure and any element has exactly one other item next to it except of the last element which has no successor, according to the following scheme: .

Therefore the arrays and linked lists are called *linear data structures*. (Although in the case of cyclic lists the linear data structure is circular.) Thus the arrays and linked lists can be considered representations of linear graphs.

6.1 General notions

At an abstract level, *trees* are also special finite graphs. They consist of nodes (i.e. vertices) and edges (i.e. arcs). If the tree is empty, it contains neither node nor edge. The empty tree is denoted by \oslash . If it is nonempty, it always contains a *root node*.² Each node of the graph has zero or more immediate successor nodes which are called its *children* and it is called their *parent*. There is a directed edge going from the parent to each of its children. If a node has no child, it is called a *leaf*. The root has no parent. Each other node of the tree has exactly one parent. The non-leaf nodes are also called *internal nodes*.

Notice that a linear data structure can be considered a tree where each internal node has a single child. Such trees will be called *linear trees*.

The *descendants* of a node are its children, and the descendants of its children. The *ancestors* of a node is its parent, and the ancestors of its parent. The root (node) has no ancestor, but each other node of the tree is descendant of it, thus the root is ancestor of each other node in the tree. Thus the root of a tree identifies the whole tree. The leaves have no descendant.

Traditionally a tree is drawn in reversed manner: The topmost node is the root, and the edges go downwards (so the arrows at the ends of the edges are often omitted). See figure 8.

Given a tree, it is *subtree* of itself. If the tree is not empty, its other subtrees are the subtrees of the trees rooted by its children (recursively).

The *size* of a tree is the number of nodes in the tree. The size of tree t is denoted by $n(t)$ or $|t|$ ($n(t) = |t|$). The number of internal nodes of t is $i(t)$. The number of leaves of t is $l(t)$. Clearly $n(t) = i(t) + l(t)$ for a tree. For example, $n(\oslash) = 0$, and considering figure 8, $n(t_1) = i(t_1) + l(t_1) = 1 + 2 = 3$,

²In this chapter the trees are always *rooted, directed trees*. We do not consider *free trees* here which are undirected connected graphs without cycles.

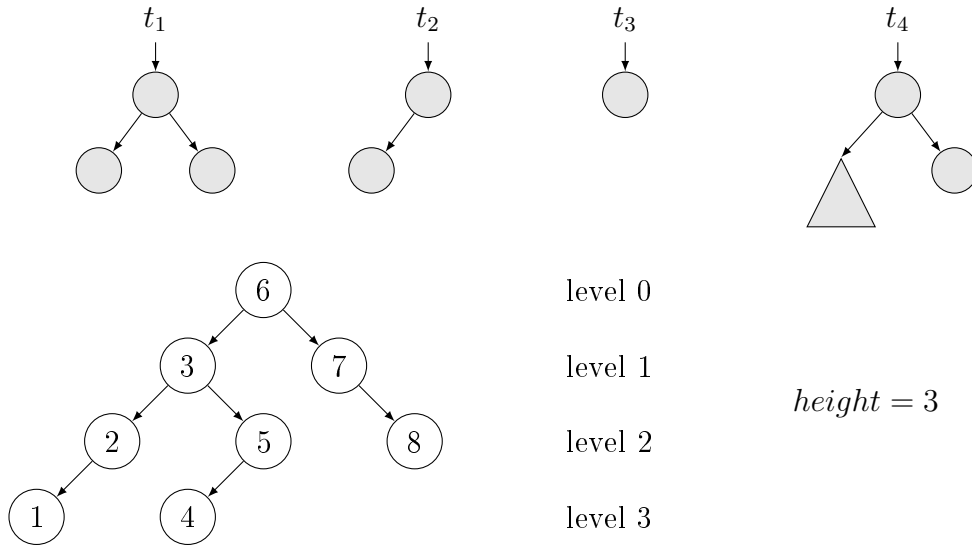


Figure 8: Simple binary trees. The nodes of the trees are represented by circles. If the structure of a subtree is not important or unknown, we denote it with a triangle.

$n(t_2) = i(t_2) + l(t_2) = 1 + 1 = 2$, $n(t_3) = i(t_3) + l(t_3) = 0 + 1 = 1$, and $n(t_4) = 2 + n(t_4 \rightarrow \text{left})$ where “ $t_4 \rightarrow \text{left}$ ” is the unknown left subtree of t_4 .

We can speak of the *levels* of a nonempty tree. The root is at level zero (which is the topmost level), its children are at level one, its grandchildren at level two and so on. Given a node at level i , its children are at level $i + 1$ (always in downwards direction).

The *height* of a nonempty tree is equal to the level of the leaves at its lowest level. The height of the empty tree is -1 . The height of tree t is denoted by $h(t)$. For example, $h(\emptyset) = -1$, and considering figure 8, $h(t_1) = 1$, $h(t_2) = 1$, $h(t_3) = 0$, and $h(t_4) = 1 + \max(h(t_4 \rightarrow \text{left}), 0)$ where “ $t_4 \rightarrow \text{left}$ ” is the unknown left subtree of t_4 .

All the hierarchical structures can be modeled by trees, for example the directory hierarchy of a computer. Typically, each node of a tree is labeled by some *key*, and maybe with other data.

Sources: [1] Chapter 10, 6, 12; [6] 4.4; [3] 6-7

6.2 Binary trees

Binary trees are useful, for example for representing sets and multisets (i.e. bags) like dictionaries and priority queues.

A *binary tree* is a tree where each internal (i.e. non-leaf) node has at most two children. If a node has two children, they are the *left* child, and the *right* child of their parent. If a node has exactly one child, then this child is the *left* or *right* child of its parent. This distinction is important.

If t is a nonempty binary tree (i.e. $t \neq \emptyset$), then $*t$ is the root node of the tree, $t \rightarrow \text{key}$ is the key labeling $*t$, $t \rightarrow \text{left}$ is the left subtree of t , and $t \rightarrow \text{right}$ is the right subtree of t . If $*t$ does not have left child, $t \rightarrow \text{left} = \emptyset$, i.e. the left subtree of t is empty. Similarly, if $*t$ does not have right child, $t \rightarrow \text{right} = \emptyset$, i.e. the right subtree of t is empty.

If $*p$ is a node of a binary tree, p is the (sub)tree rooted by $*p$, thus $p \rightarrow \text{key}$ is its key, $p \rightarrow \text{left}$ is its left subtree, and $p \rightarrow \text{right}$ is its right subtree. If $*p$ has left child, it is $*p \rightarrow \text{left}$. If $*p$ has right child, it is $*p \rightarrow \text{right}$. If $*p$ has parent, it is $*p \rightarrow \text{parent}$. The tree rooted by the parent is $p \rightarrow \text{parent}$. (Notice that the infix operator \rightarrow binds stronger than the prefix operator $*$. For example, $*p \rightarrow \text{left} = *(p \rightarrow \text{left})$.) If $*p$ does not have parent, $p \rightarrow \text{parent} = \emptyset$.

If $p = \emptyset$, all the expressions $*p$, $p \rightarrow \text{key}$, $p \rightarrow \text{left}$, $p \rightarrow \text{right}$, $p \rightarrow \text{parent}$ etc. are erroneous.

Properties 6.1

$$h(\emptyset) = -1$$

$$t \neq \emptyset \text{ binary tree} \implies h(t) = 1 + \max(h(t \rightarrow \text{left}), h(t \rightarrow \text{right}))$$

A *strictly binary tree* is a binary tree where each internal node has two children. The following property can be proved easily by induction on $i(t)$.

Property 6.2 Given a $t \neq \emptyset$ strictly binary tree,

$$l(t) = i(t) + 1.$$

$$\text{Thus } n(t) = 2i(t) + 1 \wedge n(t) = 2l(t) - 1$$

A *complete binary tree* is a strictly binary tree where all the leaves are at the same level. It follows that in a complete nonempty binary tree of height h we have 2^0 nodes at level 0, 2^1 nodes at level 1, and 2^i nodes at level i ($0 \leq i \leq h$). Therefore we get the following property.

Property 6.3 Given a $t \neq \emptyset$ complete binary tree ($n = n(t)$, $h = h(t)$)

$$n = 2^0 + 2^1 + \dots + 2^h = 2^{h+1} - 1$$

A nonempty *nearly complete binary tree* is a binary tree which becomes complete if we delete its lowest level. The \odot is also nearly complete binary tree. (An **equivalent definition**: A *nearly complete binary tree* is a binary tree which can be received from a complete binary tree by deleting zero or more nodes from its lowest level.) Notice that the complete binary trees are also nearly complete, according to this definition.

Because a nonempty nearly complete binary tree is complete, possibly except at its lowest level h , it has $2^h - 1$ nodes at its first $h - 1$ levels, at least 1 node and at most 2^h nodes at its lowest level. Thus $n \geq 2^h - 1 + 1 \wedge n \leq 2^h - 1 + 2^h$ where n is the size of the tree.

Properties 6.4 *Given a $t \neq \odot$ nearly complete binary tree ($n = n(t), h = h(t)$). Then $n \in 2^h..(2^{h+1}-1)$. Thus $h = \lfloor \log n \rfloor$.*

Now we consider the relations between the size n and height h of a nonempty binary tree. These will be important at binary search trees which are good representations of data sets stored in the RAM.

Clearly, a nonempty binary tree of height h has the most number of nodes, if it is complete. Therefore from property 6.3 we receive that for the size n of any nonempty binary tree $n < 2^{h+1}$. Thus $\log n < \log(2^{h+1}) = h + 1$,³ so $\lfloor \log n \rfloor < h + 1$. Finally $\lfloor \log n \rfloor \leq h$.

On the other hand, a nonempty binary tree of height h has the least number of nodes, if it contains just one node at each level (i.e. it is linear), and so for the size n of any nonempty binary tree $n \geq h + 1$. Thus $h \leq n - 1$.

Property 6.5 *Given a $t \neq \odot$ binary tree ($n = n(t), h = h(t)$) $\lfloor \log n \rfloor \leq h \leq n - 1$ where the $h = \lfloor \log n \rfloor$ if the tree is nearly complete, and $h = n - 1$ iff the tree is linear (see page 35).*

Notice that there are binary trees with $h = \lfloor \log n \rfloor$, although they are not nearly complete.

6.3 Linked representations of binary trees

The empty tree is represented by a null pointer, i.e. \odot . A nonempty binary tree is identified by a pointer referring to a **Node** which represents the root of the tree:

³Remember the definition of $\log n$ given in section 1: $\log n = \log_2(n)$ if $n > 0$, and $\log 0 = 0$.

Node
+ $key : \mathcal{T}$ // \mathcal{T} is some known type
+ $left, right : Node^*$
+ $Node() \{ left := right := \oslash \}$ // generate a tree of a single node.
+ $Node(x : \mathcal{T}) \{ left := right := \oslash ; key := x \}$

Sometimes it is useful, if we also have a parent pointer in the nodes of the tree. We can see a binary tree with parent pointers on figure 9.

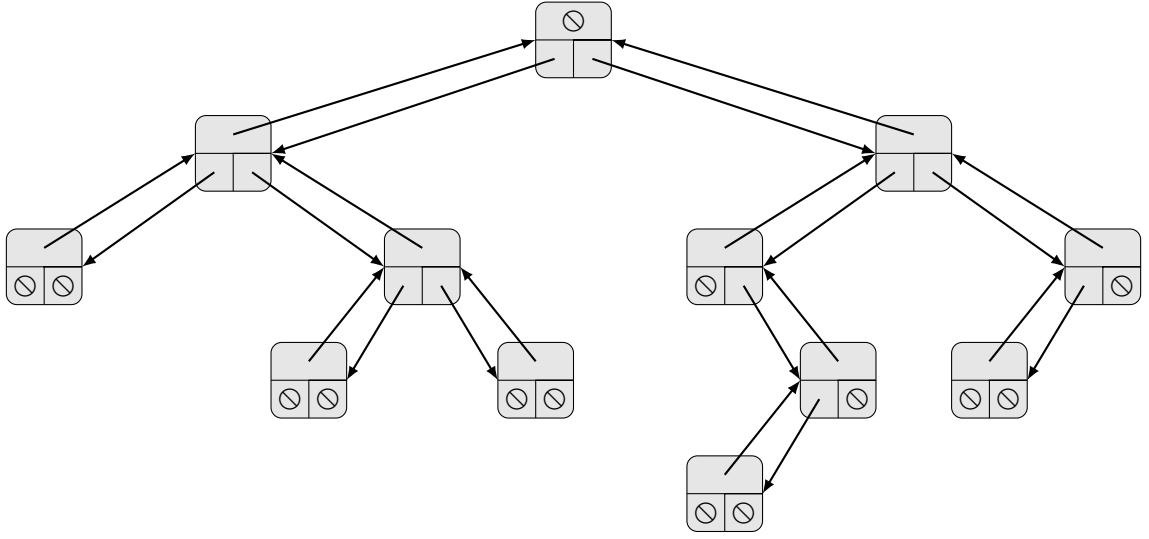
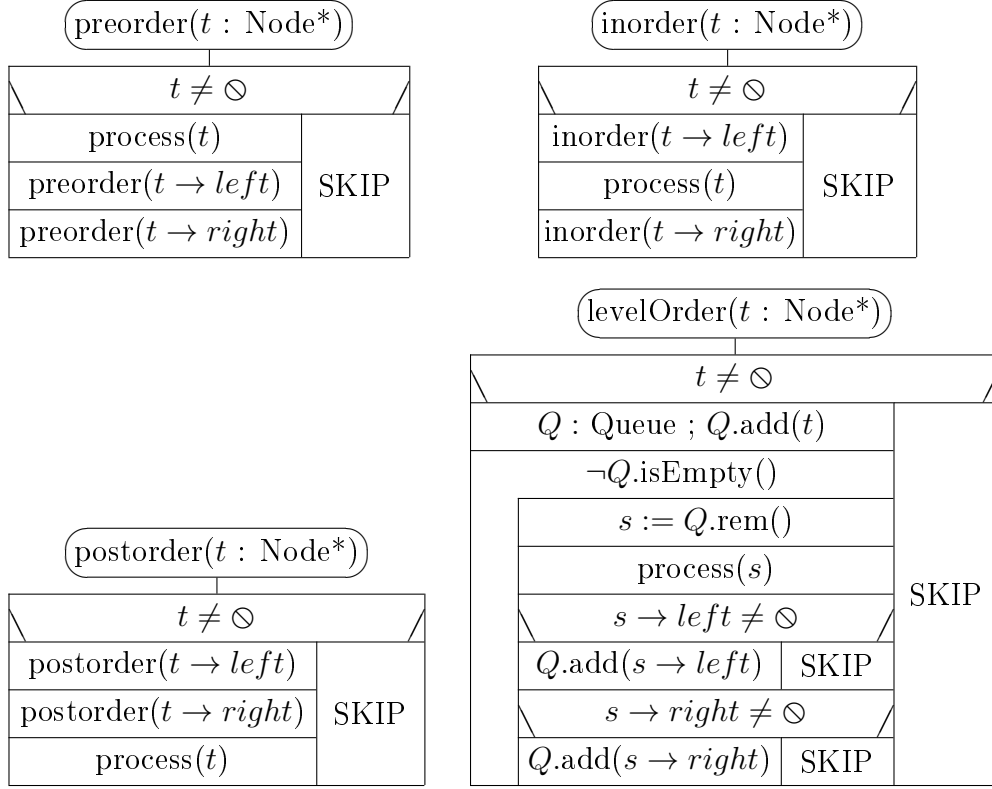


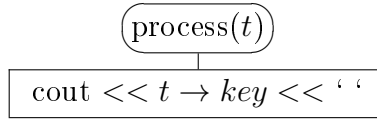
Figure 9: Binary tree with parent pointers. (We omitted the keys of the nodes here. The null pointers are represented by “/” signs in this figure.)

Node3
+ $key : \mathcal{T}$ // \mathcal{T} is some known type
+ $left, right, parent : Node3^*$
+ $Node3(p:Node3^*) \{ left := right := \oslash ; parent := p \}$
+ $Node3(x : \mathcal{T}, p:Node3^*) \{ left := right := \oslash ; parent := p ; key := x \}$

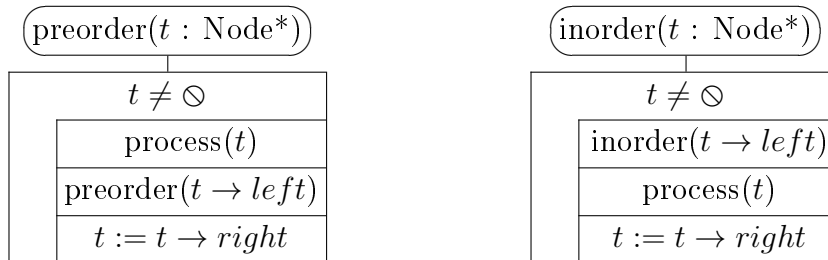
6.4 Binary tree traversals



$T_{\text{preorder}}(n), T_{\text{inorder}}(n), T_{\text{postorder}}(n), T_{\text{levelOrder}}(n) \in \Theta(n)$ where $n = n(t)$ and the time complexity of process(t) is $\Theta(1)$. For example, process(t) can print $t \rightarrow \text{key}$:



We can slightly reduce the running times of preorder and inorder traversals, if we eliminate the tail recursions. (It does not affect the time complexities.)



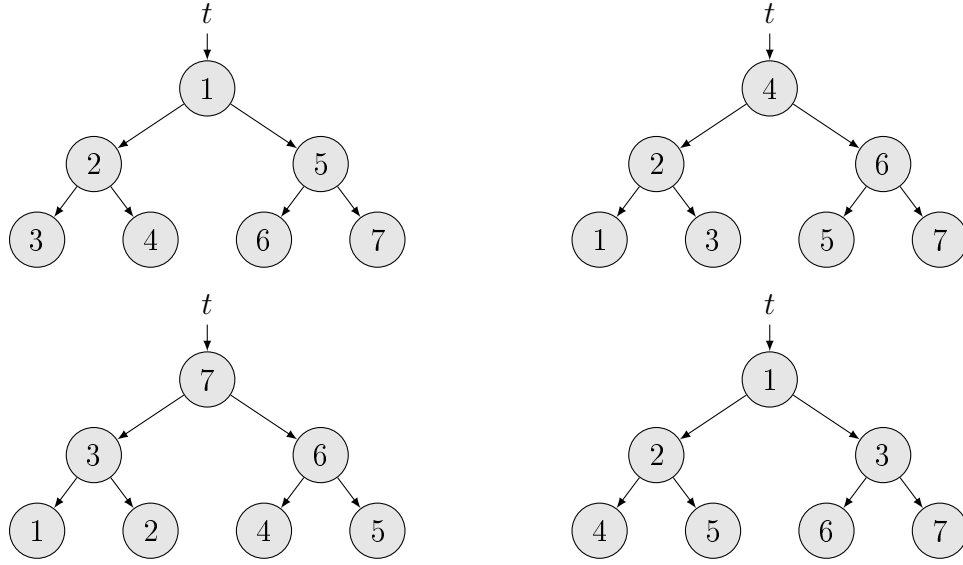
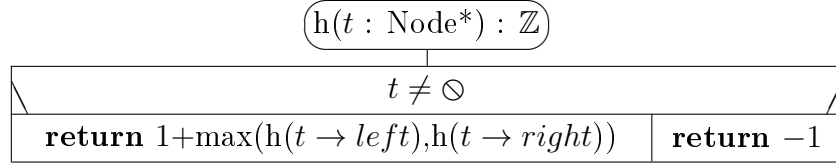


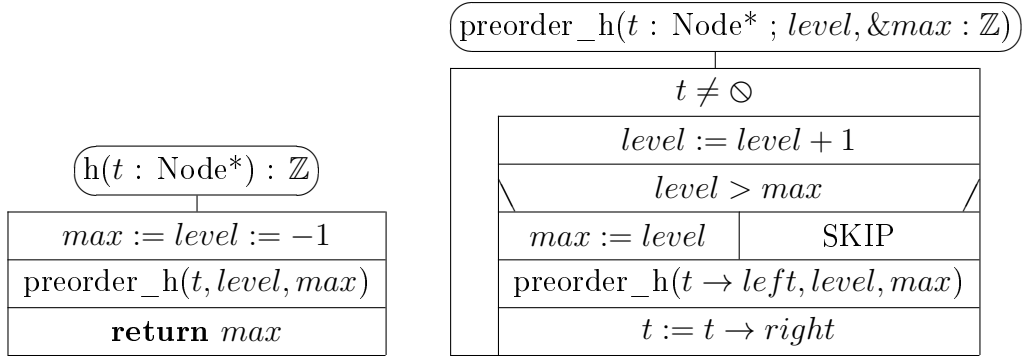
Figure 10: left upper part: preorder, right upper part: inorder, left lower part: postorder, right lower part: level order traversal of binary tree t .

6.4.1 An application of traversals: the height of a binary tree

Postorder:



Preorder:



6.4.2 Using parent pointers

$\text{inorder_next}(p : \text{Node3}^*) : \text{Node3}^*$	
$q := p \rightarrow \text{right}$	
$q \neq \emptyset$	
$q \rightarrow \text{left} \neq \emptyset$	$q := p \rightarrow \text{parent}$
$q := q \rightarrow \text{left}$	$q \neq \emptyset \wedge q \rightarrow \text{left} \neq p$
	$p := q ; q := q \rightarrow \text{parent}$
return q	

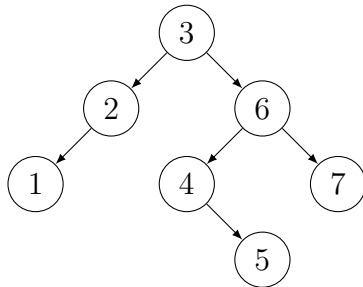
$MT(h) \in O(h)$ where $h = h(t)$

6.5 Parenthesized, i.e. textual form of binary trees

Parenthesized, i.e. textual form of a nonempty binary tree:

(leftSubtree Root rightSubtree)

The empty tree is represented by the empty string. We can use different kinds of parentheses for easier reading. For example, see Figure 11.



The binary tree on the left in

- simple textual form:
 $(((1) 2) 3 ((4 (5)) 6 (7)))$
- elegant parenthesized form:
 $\{ [(1) 2] 3 [(4 \langle 5 \rangle) 6 (7)] \}$

Figure 11: The same binary tree in graphical and textual representations. Notice that omitting the parenthesis from a textual form of a tree we receive the inorder traversal of that tree.

6.6 Binary search trees

» *Binary search trees* (BST) are a particular type of containers: data structures that store "items" (such as numbers, names etc.) in memory. They

allow fast lookup, addition and removal of items, and can be used to implement either dynamic sets of items, or lookup tables that allow finding an item by its key (e.g., finding the phone number of a person by name).

Binary search trees keep their keys in sorted order, so that lookup and other operations can use the principle of binary search: when looking for a key in a tree (or a place to insert a new key), they traverse the tree from root to leaf, making comparisons to keys stored in the nodes of the tree and deciding, on the basis of the comparison, to continue searching in the left or right subtrees. On average, this means that each comparison allows the operations to skip about half of the tree, so that each lookup, insertion or deletion takes time proportional to the logarithm of the number of items stored in the tree. This is much better than the linear time required to find items by key in an (unsorted) array, but slower than the corresponding operations on hash tables.«

(The source of the text above is Wikipedia.)

A *binary search tree* (BST) is a binary tree, whose nodes each store a *key* (and optionally, some associated value). The tree additionally satisfies the binary search tree property, which states that the key in each node must be greater than any key stored in the left sub-tree, and less than any key stored in the right sub-tree. (See Figure 12. Notice that there is no duplicated key in a BST.)

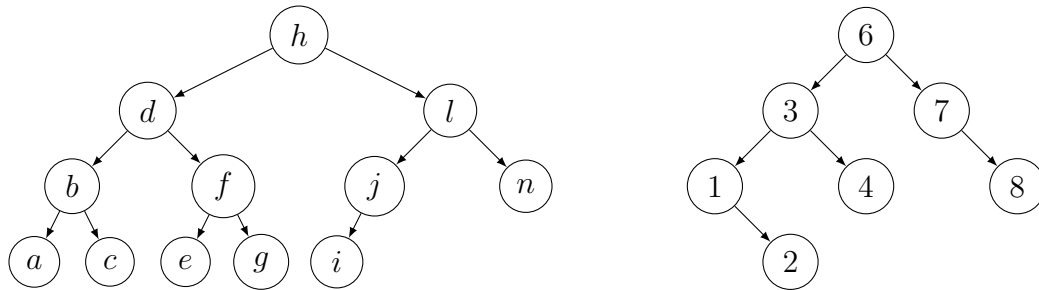
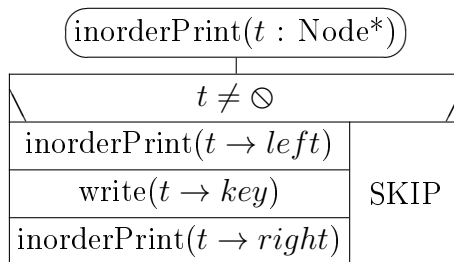


Figure 12: Two *binary search trees* (BSTs)



Property 6.6 *Procedure $\text{inorderPrint}(t : \text{Node}^*)$ prints the keys of binary tree t in strictly increasing order, iff binary tree t is a search tree.*

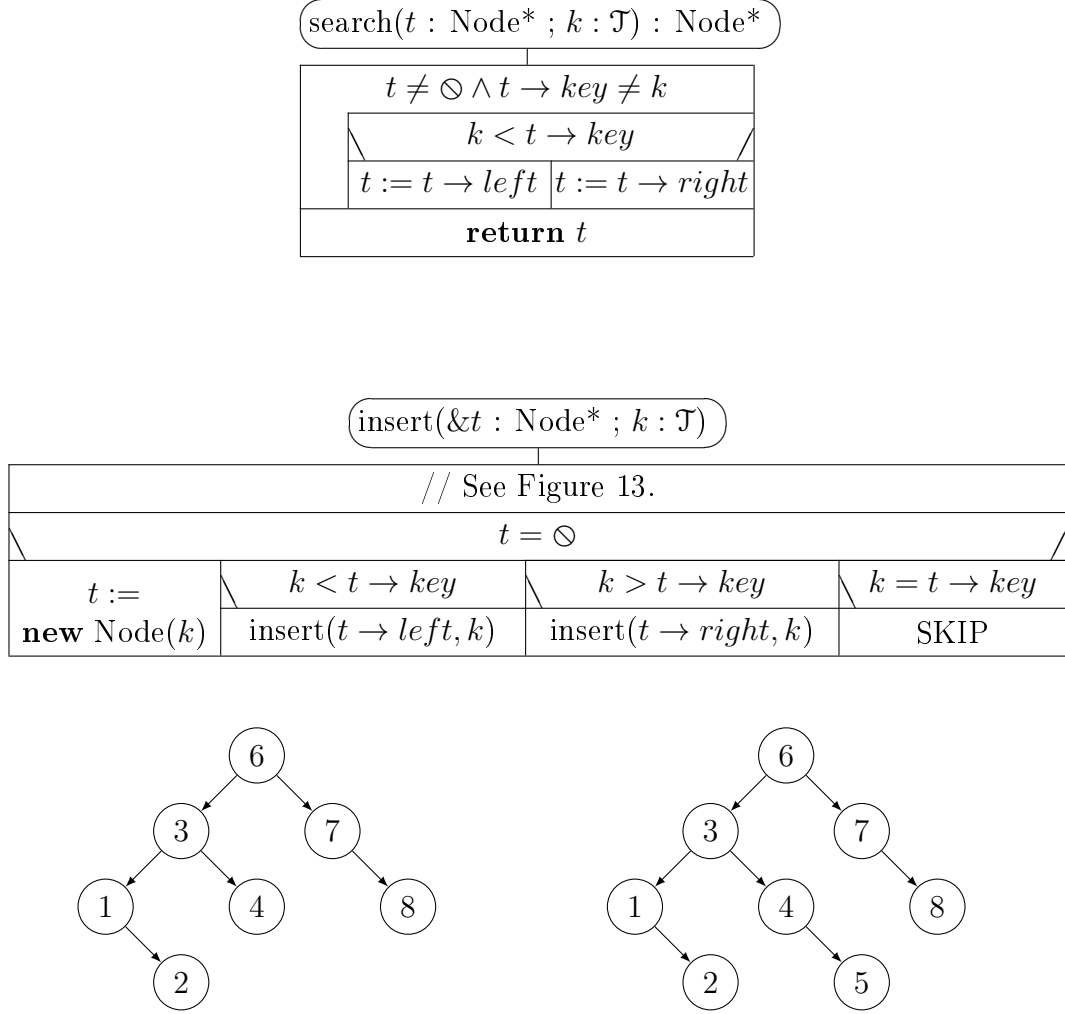


Figure 13: Inserting 5 into the BST on the left gives the BST on the right: First we compare 5 with the key of the root, which is 6, and $5 < 6$. Thus we insert 5 into the left subtree. Next we compare 5 with 3, and $5 > 3$, so we insert 5 into the right subtree of node 3, and again $5 > 4$, therefore we insert 5 into the right subtree of node 4. The right subtree of this node is empty. Consequently we create a new node with key 5, and we substitute that empty right subtree with this new subtree consisting of this single new node.

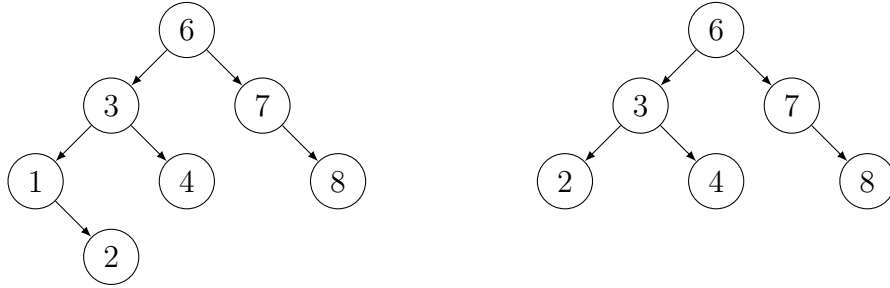
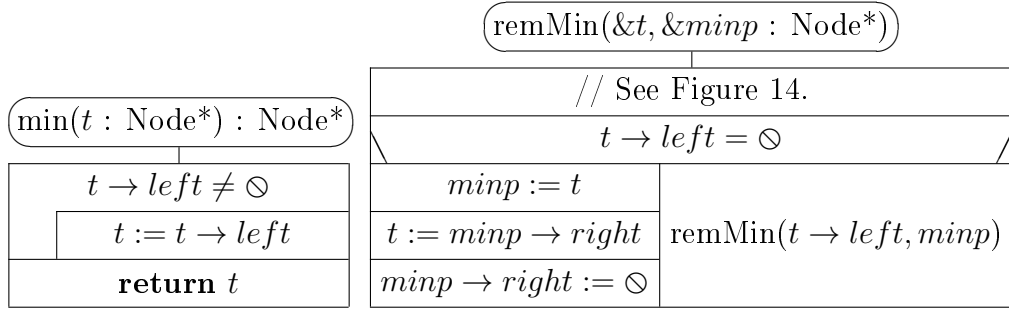


Figure 14: Removing the node with minimal key from the BST on the left gives the BST on the right: The leftmost node of the original BST is substituted by its right subtree. Notice that deleting the node with key 1 from the BST on the left gives the same BST as result: The left subtree of this node is empty. Thus deleting it means substituting it with its right subtree.

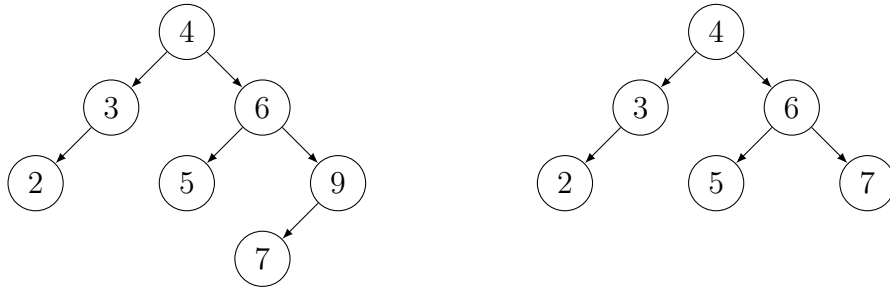


Figure 15: Removing the node with maximal key from the BST on the left gives the BST on the right: The rightmost node of the original BST is substituted by its left subtree. Notice that deleting the node with key 9 from the BST on the left gives the same BST as result: The right subtree of this node is empty. Thus deleting it means substituting it with its left subtree.

del(&t : Node* ; k : T)			
// See Figure 16.			
$t \neq \ominus$			
$k < t \rightarrow key$	$k > t \rightarrow key$	$k = t \rightarrow key$	SKIP
del($t \rightarrow left, k$)	del($t \rightarrow right, k$)	delRoot(t)	

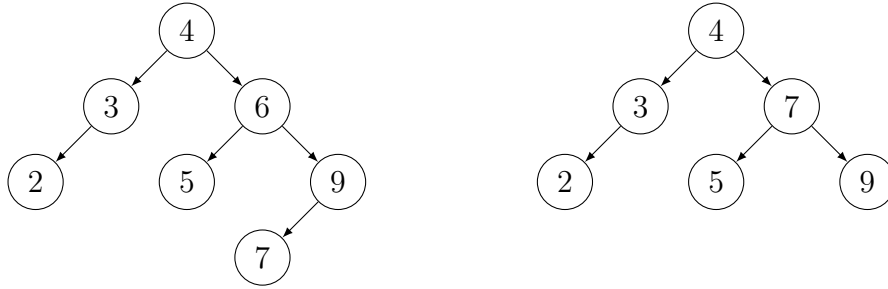


Figure 16: Deleting node 6 of the BST on the left gives the BST on the right: Node 6 has two children. Thus we remove the minimal node of its right subtree and substitute node 6 with it.

delRoot(&t : Node*)		
// See Figure 17.		
$t \rightarrow left = \ominus$	$t \rightarrow right = \ominus$	$t \rightarrow left \neq \ominus \wedge t \rightarrow right \neq \ominus$
$p := t$	$p := t$	remMin($t \rightarrow right, p$)
$t := p \rightarrow right$	$t := p \rightarrow left$	$p \rightarrow left := t \rightarrow left ; p \rightarrow right := t \rightarrow right$
delete p	delete p	delete $t ; t := p$

$MT_{\text{search}}(h), MT_{\text{insert}}(h), MT_{\text{min}}(h) \in \Theta(h)$
 $MT_{\text{remMin}}(h), MT_{\text{del}}(h), MT_{\text{delRoot}}(h) \in \Theta(h)$
 where $h = h(t)$.

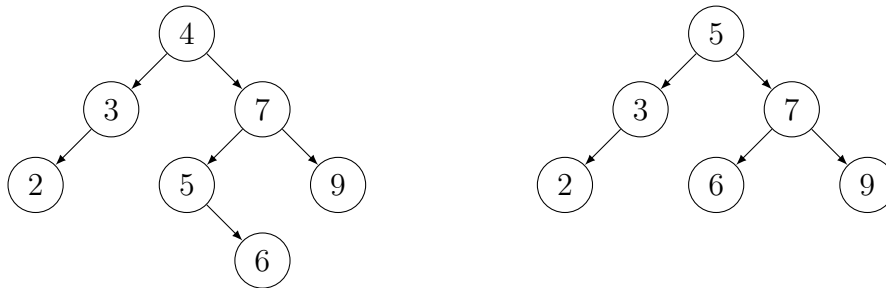


Figure 17: Deleting the root node of the BST on the left gives the BST on the right: The root node has two children. Thus we remove the minimal node of its right subtree and substitute the root node with it. (Notice that we could remove the maximal node of the left subtree and substitute the root node with it.)

6.7 Level-continuous binary trees, and heaps

A binary tree is *level-continuous*, **iff** all levels of the tree, except possibly the last (deepest) one are completely filled (i.e. it is nearly complete), and, provided that the last level of the tree is not complete, the nodes of that level are filled from left to right. Clearly a node is internal node of a *level-continuous* tree, iff it has left child in the tree.

A *binary maximum heap* is a level-continuous binary tree where the key stored in each node is greater than or equal to (\geq) the keys in the node's children. (See Figure 18.)

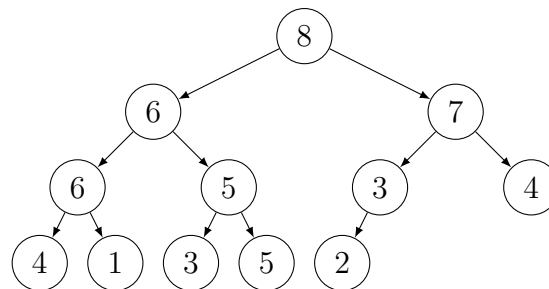


Figure 18: A (*binary maximum*) *heap*. It is a nearly complete binary tree, and the last level is filled from left to right. The key of each parent \geq the key of the child.

A *binary minimum heap* is a level-continuous binary tree where the key stored in each node is less than or equal to (\leq) the keys in the node's children.

In this lecture notes a *heap* is a *binary maximum heap* by default.

Priority queues are usually represented by heaps in arithmetic representation.

6.8 Arithmetic representation of level-continuous binary trees

A level-continuous binary tree of size n is typically represented by the first n elements of an array A : We put the nodes of the tree into the array in level order. (See Figure 19.)

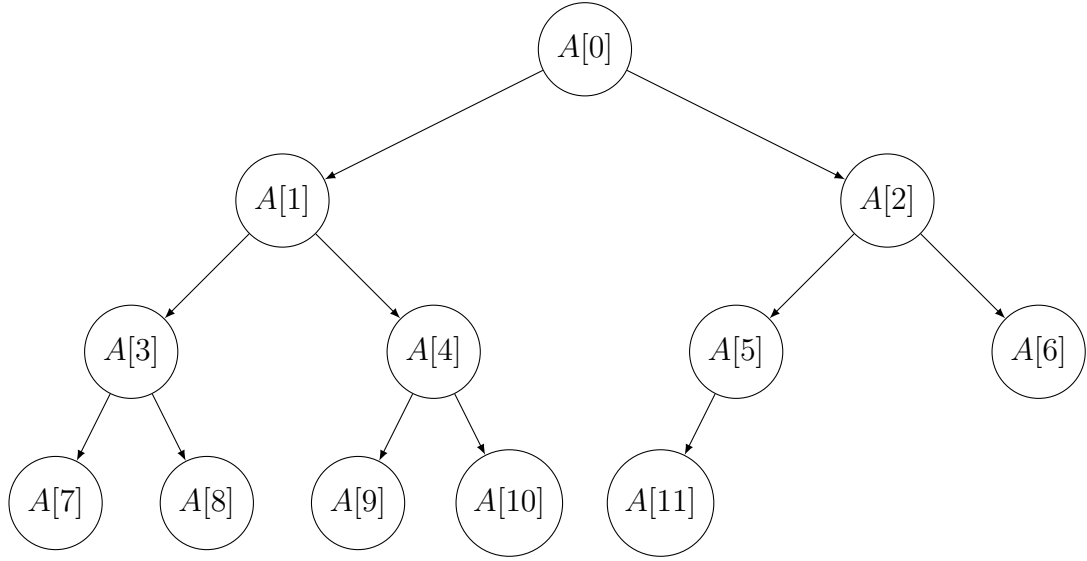


Figure 19: A level-continuous binary tree of size 12 represented by the first 12 elements of an array A : We put the nodes of the tree into the array in level order.

If an internal node has index i in array A , let $left(i)$ be the index of its left child. If it has right child, too, let $right(i)$ be the index of it. Clearly $right(i) = left(i) + 1$ in the latter case. If $A[j]$ represents a node different from the root, Let $parent(j)$ be the index of its parent.

Let us suppose that we use the first n elements of an array (indexed from zero), for example $A : \mathcal{T}[m]$, i.e. we use the subarray $A[0..n)$. Then the root node of a nonempty tree is $A[0]$. Node $A[i]$ is internal node, iff $left(i) < n$. Then $left(i) = 2i + 1$. The right child of node $A[i]$ exists, iff $right(i) < n$. Then $right(i) = 2i + 2$. Node $A[j]$ is not the root node of the tree, iff $j > 0$. Then $parent(j) = \lfloor \frac{j-1}{2} \rfloor$.

6.9 Heaps and priority queues

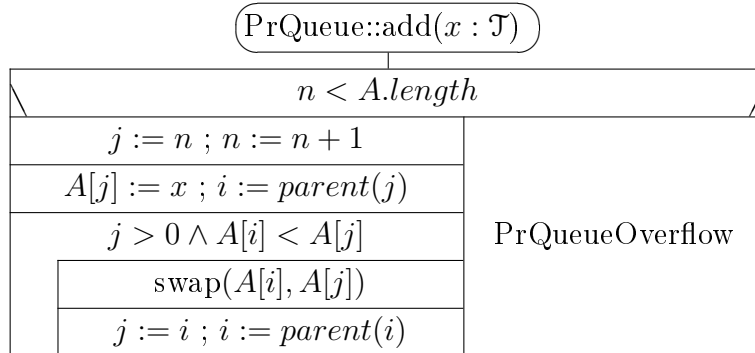
A priority queue is a bag (i.e. multiset). We can add a new item to it, and we can check or remove a maximal item of it.⁴ Often there is a priority function: $\mathcal{T} \rightarrow \text{some number type}$, where \mathcal{T} is the element type of the priority queue, and the items are compared according to their priorities. For example, the processes of a computer are often scheduled according to their priorities.

Our representation is similar to our representation of the Stack type, and some operations are also the same, except of the names. The actual elements of the priority queue are in the subarray $A[0..n)$ containing a (binary maximum) heap.

PrQueue
- $A : \mathcal{T}[]$ // \mathcal{T} is some known type - $n : 0..A.length$ // n is the actual length of the priority queue
+ PrQueue($m : \mathbb{N}$) { $A := \text{new } \mathcal{T}[m]; n := 0$ } // create an empty priority queue + add($x : \mathcal{T}$) // insert x into the priority queue + remMax(): \mathcal{T} // remove and return the maximal element of the priority queue + max(): \mathcal{T} // return the maximal element of the priority queue + isFull(): \mathbb{B} { return $n = A.length$ } + isEmpty(): \mathbb{B} { return $n = 0$ } + \sim PrQueue() { delete A } + setEmpty() { $n := 0$ } // reinitialize the priority queue

Provided that subarray $A[0..n)$ is the arithmetic representation of a heap, $MT_{\text{add}}(n) \in \Theta(\log n)$, $mT_{\text{add}}(n) \in \Theta(1)$, $MT_{\text{remMax}}(n) \in \Theta(\log n)$, $mT_{\text{remMax}}(n) \in \Theta(1)$, $T_{\text{max}}(n) \in \Theta(1)$.

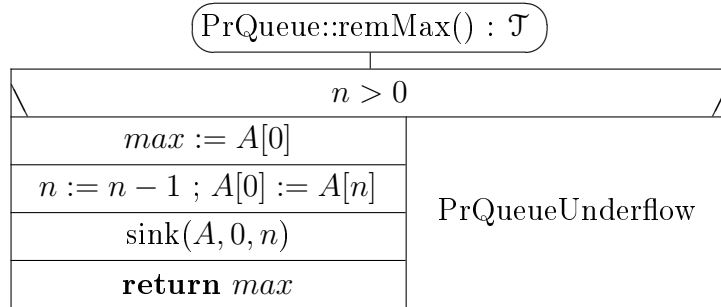
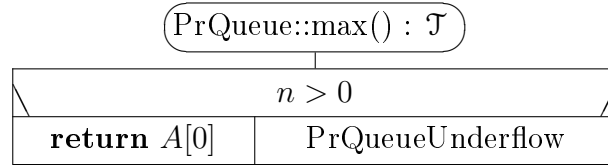
$MT_{\text{add}}(n) \in \Theta(\log n)$, and $MT_{\text{remMax}}(n) \in \Theta(\log n)$, because the main loops of subprograms “add” and “sink” below iterates maximum h times if h is the height of the heap, and $h = \lfloor \log n \rfloor$.



⁴There are also min priority queues where we can check or remove a minimal item.

Changes of $A : \mathbb{Z}[16]$ while applying $\text{add}()$ and $\text{remMax}()$ operations to it.

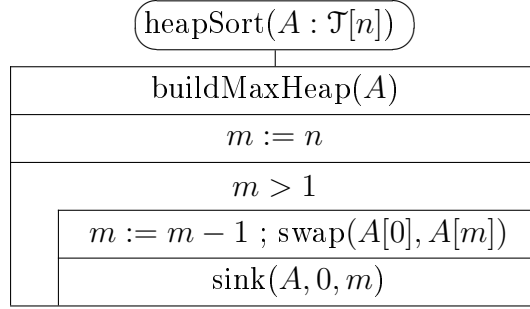
op	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
-	8	6	7	6	5	3	4	4	1	3	5	2				
add(8)	8	6	7	6	5	<u>3</u>	4	4	1	3	5	2	<u>8</u>			
...	8	6	<u>7</u>	6	5	<u>8</u>	4	4	1	3	5	2	3			
...	<u>8</u>	6	<u>8</u>	6	5	7	4	4	1	3	5	2	3			
.	8	6	8	6	5	7	4	4	1	3	5	2	3			
add(2)	8	6	8	6	5	7	<u>4</u>	4	1	3	5	2	3	<u>2</u>		
.	8	6	8	6	5	7	4	4	1	3	5	2	3	2		
add(9)	8	6	8	6	5	7	<u>4</u>	4	1	3	5	2	3	2	<u>9</u>	
...	8	6	<u>8</u>	6	5	7	<u>9</u>	4	1	3	5	2	3	2	4	
...	<u>8</u>	6	<u>9</u>	6	5	7	8	4	1	3	5	2	3	2	4	
.	9	6	8	6	5	7	8	4	1	3	5	2	3	2	4	
remMax()	<u>9</u>	6	8	6	5	7	8	4	1	3	5	2	3	2	<u>4</u>	
$max := 9$	<u>4</u>	6	<u>8</u>	6	5	7	8	4	1	3	5	2	3	2		
...	8	6	<u>4</u>	6	5	7	<u>8</u>	4	1	3	5	2	3	2		
...	8	6	8	6	5	7	<u>4</u>	4	1	3	5	2	3	<u>2</u>		
return 9	8	6	8	6	5	7	4	4	1	3	5	2	3	2		
remMax()	<u>8</u>	6	8	6	5	7	4	4	1	3	5	2	3	<u>2</u>		
$max := 8$	<u>2</u>	6	<u>8</u>	6	5	7	4	4	1	3	5	2	3			
...	8	6	<u>2</u>	6	5	<u>7</u>	4	4	1	3	5	2	3			
...	8	6	7	6	5	<u>2</u>	4	4	1	3	5	2	<u>3</u>			
return 8	8	6	7	6	5	3	4	4	1	3	5	2	2			



$\text{sink}(A : \mathcal{T}[] ; k, n : \mathbb{N})$	
$i := k ; j := \text{left}(k) ; b := \text{true}$	
$j < n \wedge b$	
// $A[j]$ is the left child of $A[i]$	
$j + 1 < n \wedge A[j + 1] > A[j]$	
$j := j + 1$	SKIP
// $A[j]$ is the greater child of $A[i]$	
$A[i] < A[j]$	
$\text{swap}(A[i], A[j])$	$b := \text{false}$
$i := j ; j := \text{left}(j)$	

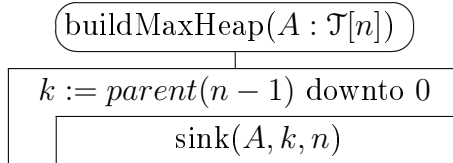
6.10 Heapsort

$MT_{\text{heapSort}}(n) \in O(n \log n)$ because $MT_{\text{buildMaxHeap}}(n) \in O(n \log n)$ and the time complexity of the main loop of heapSort is also $O(n \log n)$.



While building the heap we consider the level order of the tree, and we start from the last internal node of it. We go back in level order and sink the root node of each subtree making a heap out of it.

The time complexity of each sinking is $O(\log n)$ because $h \leq \lfloor \log n \rfloor$ for each subtree. Thus $MT_{\text{buildMaxHeap}}(n) \in O(n \log n)$ and the time complexity of the main loop of heapSort is also $O(n \log n)$.



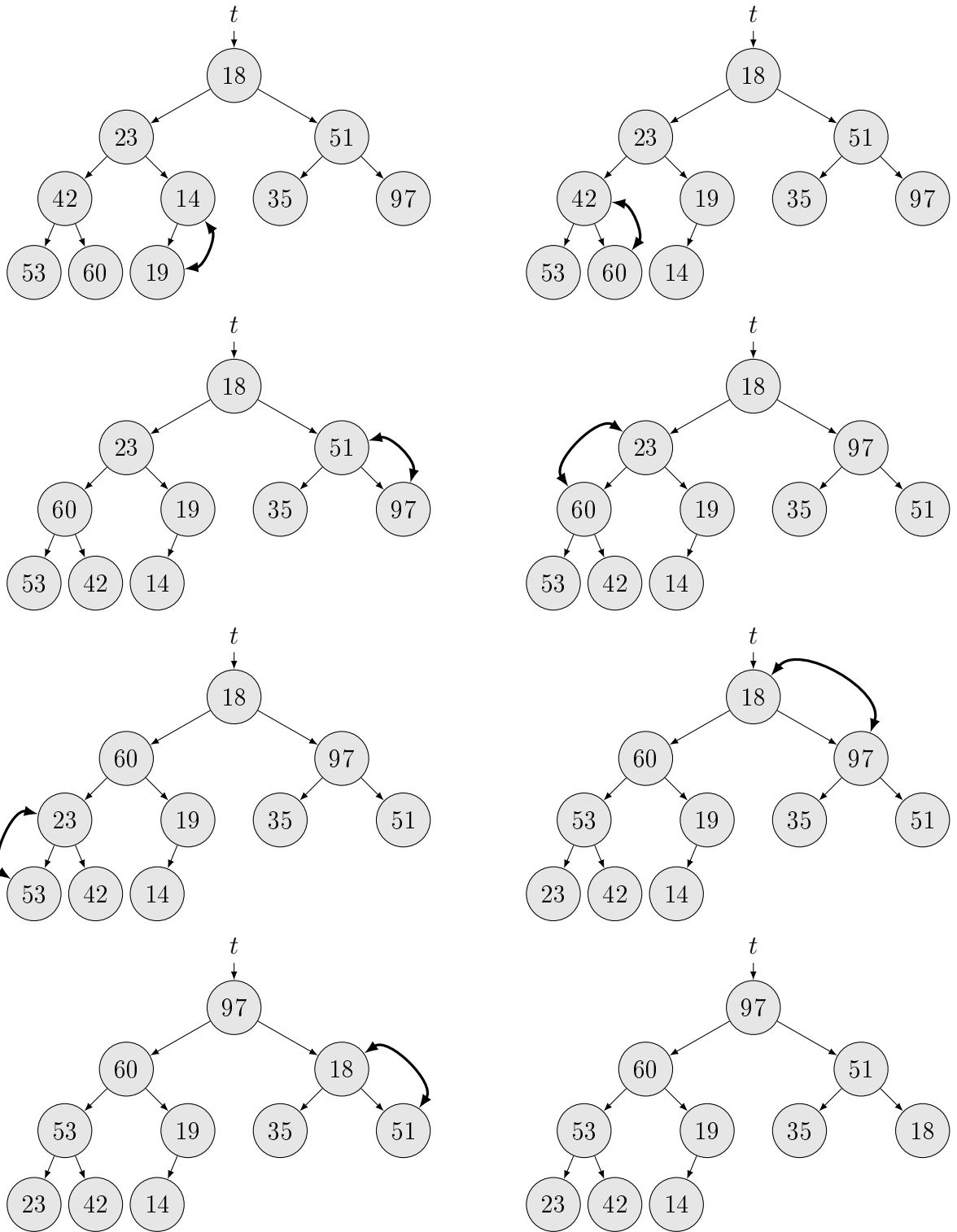


Figure 20: From array $\langle 18, 23, 51, 42, 14, 35, 97, 53, 61, 19 \rangle$ we build a maximum heap. Notice that we work on the array from the beginning to the end. The binary trees show the logical structure of the array. Finally we receive array $\langle 97, 60, 51, 53, 19, 35, 18, 23, 42, 14 \rangle$.

Full example of Heapsort on $A : \mathbb{Z}[10]$

op	0	1	2	3	4	5	6	7	8	9
sink	68	23	51	42	<u>19</u>	35	97	53	60	<u>14</u>
sink	68	23	51	<u>42</u>	19	35	97	53	<u>60</u>	14
sink	68	23	51	<u>42</u>	19	35	97	53	<u>60</u>	14
sink	68	23	<u>51</u>	60	19	35	<u>97</u>	53	42	14
sink	68	<u>23</u>	97	<u>60</u>	19	35	51	53	42	14
...	68	60	97	<u>23</u>	19	35	51	<u>53</u>	42	14
sink	<u>68</u>	60	<u>97</u>	53	19	35	51	23	42	14
...	97	60	<u>68</u>	53	19	35	<u>51</u>	23	42	14
.	97	60	68	53	19	35	51	23	42	14
swap	<u>97</u>	60	68	53	19	35	51	23	42	<u>14</u>
sink	<u>14</u>	60	<u>68</u>	53	19	35	51	23	42	<u>97</u>
...	68	60	<u>14</u>	53	19	35	<u>51</u>	23	42	<u>97</u>
swap	<u>68</u>	60	51	53	19	35	14	23	<u>42</u>	<u>97</u>
sink	<u>42</u>	<u>60</u>	51	53	19	35	14	23	<u>68</u>	<u>97</u>
...	60	<u>42</u>	51	<u>53</u>	19	35	14	23	<u>68</u>	<u>97</u>
.	60	53	51	<u>42</u>	19	35	14	<u>23</u>	<u>68</u>	<u>97</u>
swap	<u>60</u>	53	51	42	19	35	14	<u>23</u>	<u>68</u>	<u>97</u>
sink	<u>23</u>	<u>53</u>	51	42	19	35	14	<u>60</u>	<u>68</u>	<u>97</u>
...	53	<u>23</u>	51	<u>42</u>	19	35	14	<u>60</u>	<u>68</u>	<u>97</u>
swap	<u>53</u>	42	51	23	19	35	14	<u>60</u>	<u>68</u>	<u>97</u>
sink	<u>14</u>	42	<u>51</u>	23	19	35	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
...	51	42	<u>14</u>	23	19	35	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
swap	<u>51</u>	42	35	23	19	14	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
sink	<u>14</u>	<u>42</u>	35	23	19	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
...	42	<u>14</u>	35	<u>23</u>	19	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
swap	<u>42</u>	23	35	14	<u>19</u>	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
sink	<u>19</u>	23	<u>35</u>	14	<u>42</u>	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
swap	<u>35</u>	23	19	<u>14</u>	<u>42</u>	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
sink	<u>14</u>	<u>23</u>	19	<u>35</u>	<u>42</u>	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
swap	<u>23</u>	14	<u>19</u>	<u>35</u>	<u>42</u>	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
sink	<u>19</u>	<u>14</u>	<u>23</u>	<u>35</u>	<u>42</u>	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
swap	<u>19</u>	<u>14</u>	<u>23</u>	<u>35</u>	<u>42</u>	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
sink	<u>14</u>	<u>19</u>	<u>23</u>	<u>35</u>	<u>42</u>	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>
.	<u>14</u>	<u>19</u>	<u>23</u>	<u>35</u>	<u>42</u>	<u>51</u>	<u>53</u>	<u>60</u>	<u>68</u>	<u>97</u>

7 Lower bounds for sorting

Theorem 7.1 *For any sorting algorithm $mT(n) \in \Omega(n)$.*

Proof. Clearly we have to check all the n items and only a limited number of items is checked in a subprogram call or loop iteration (without the embedded subprogram calls and loop iterations). Let this limit be k . Thus $mT(n) * k \geq n \implies mT(n) \geq \frac{1}{k}n \implies mT(n) \in \Omega(n)$. \square

7.1 Comparison sorts and the decision tree model

Definition 7.2 *A sorting algorithm is comparison sort, iff it gains information about the sorted order of the input items only by comparing them. That is, given two items a_i and a_j , in order to compare them it performs one of the tests $a_i < a_j$, $a_i \leq a_j$, $a_i = a_j$, $a_i \neq a_j$, $a_i \geq a_j$, or $a_i > a_j$. We may not inspect the values of the elements or gain order information about them in any other way. [1]*

The sorting algorithms we have studied before, that is insertion sort, heap sort, merge sort, and quick sort are comparison sorts.

In this section, we assume without loss of generality that all the input elements are distinct⁵. Given this assumption, comparisons of the form $a_i = a_j$ and $a_i \neq a_j$ are useless, so we can assume that no comparisons of this form are made⁶. We also note that the comparisons $a_i < a_j$, $a_i \leq a_j$, $a_i \geq a_j$, and $a_i > a_j$ are all equivalent in that they yield identical information about the relative order of a_i and a_j . We therefore assume that all comparisons have the form $a_i \leq a_j$. [1]

We can view comparison sorts abstractly in terms of decision trees. A decision tree is a strictly binary tree that represents the comparisons between elements that are performed by a particular sorting algorithm operating on an input of a given size. Control, data movement, and all other aspects of the algorithm are ignored. [1]

Figure 21 shows the decision tree corresponding to the insertion sort algorithm from Section 3 operating on an input sequence of three elements.

⁵In this way we restrict the set of possible inputs and we are going to give a lower bound for the worst case of comparison sorts. Thus, if we give a lower bound for the maximum number of key comparisons ($MC(n)$) and for maximum time complexity ($MT(n)$) on this restricted set of input sequences, it is also a lower bound for them on the whole set of input sequences, because $MC(n)$ and $MT(n)$ are surely \geq on a larger set than on a smaller set.

⁶Anyway, if such comparisons are made, neither $MC(n)$ nor $MT(n)$ are decreased.

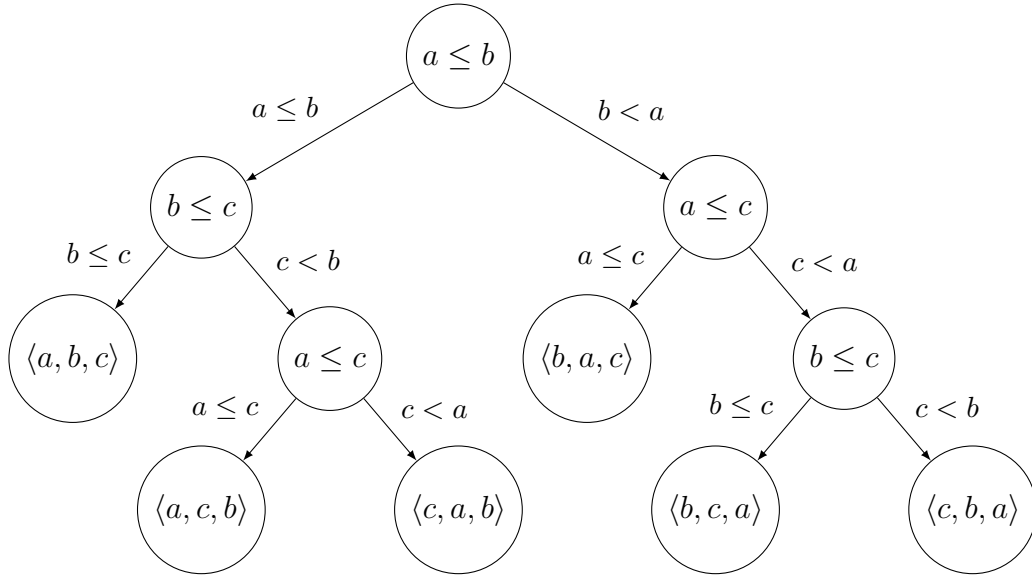


Figure 21: The decision tree for insertion sort operating on the input sequence $\langle a, b, c \rangle$. There are $3! = 6$ permutations of the 3 input items. Thus the decision tree must have at least $3! = 6$ leaves.

Let us suppose that $\langle a_1, a_2, \dots, a_n \rangle$ is the input sequence to be sorted. In a decision tree, each internal node is labeled by $a_i \leq a_j$ for some elements of the input. We also annotate each leaf by a permutation of $\langle a_1, a_2, \dots, a_n \rangle$. The execution of the sorting algorithm corresponds to tracing a simple path from the root of the decision tree down to a leaf. Each internal node indicates a comparison $a_i \leq a_j$. The left subtree then dictates subsequent comparisons once we know that $a_i \leq a_j$, and the right subtree dictates subsequent comparisons knowing that $a_i > a_j$. When we come to a leaf, the sorting algorithm has established the appropriate ordering of $\langle a_1, a_2, \dots, a_n \rangle$. Because any correct sorting algorithm must be able to produce each permutation of its input, each of the $n!$ permutations on n elements must appear as one of the leaves of the decision tree for a comparison sort to be correct. Thus, we shall consider only decision trees in which each permutation appears as a leaf of the tree. [1]

7.2 A lower bound for the worst case

Theorem 7.3 *Any comparison sort algorithm requires $MC(n) \in \Omega(n \log n)$ comparisons in the worst case.*

Proof. From the preceding discussion, it suffices to determine the height $h = MC(n)$ of a decision tree in which each permutation appears as a reachable leaf. Consider a decision tree of height h with l leaves corresponding to a comparison sort on n elements. Because each of the $n!$ permutations of the input appears as some leaf, we have $n! \leq l$. Since a binary tree of height h has no more than 2^h leaves, we have $n! \leq l \leq 2^h$. [1]

Consequently

$$MC(n) = h \geq \log n! = \sum_{i=1}^n \log i \geq \sum_{i=\lceil \frac{n}{2} \rceil}^n \log i \geq \sum_{i=\lceil \frac{n}{2} \rceil}^n \log \left\lceil \frac{n}{2} \right\rceil \geq \left\lceil \frac{n}{2} \right\rceil * \log \left\lceil \frac{n}{2} \right\rceil \geq$$

$$\geq \frac{n}{2} * \log \frac{n}{2} = \frac{n}{2} * (\log n - \log 2) = \frac{n}{2} * (\log n - 1) = \frac{n}{2} \log n - \frac{n}{2} \in \Omega(n \log n)$$

□

Theorem 7.4 *For any comparison sort algorithm $MT(n) \in \Omega(n \log n)$.*

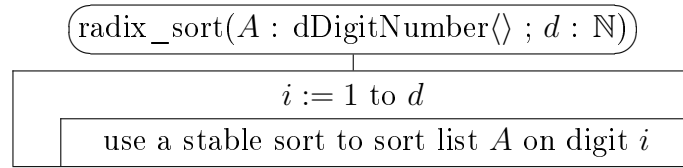
Proof. Only a limited number of key comparisons are performed in a subprogram call or loop iteration (without the embedded subprogram calls and loop iterations). Let this upper limit be k . Then $MT(n) * k \geq MC(n) \implies MT(n) \geq \frac{1}{k} MC(n) \implies MT(n) \in \Omega(MC(n))$. Together with theorem 7.3 ($MC(n) \in \Omega(n \log n)$) and transitivity of relation “ $\cdot \in \Omega(\cdot)$ ” we receive this theorem ($MT(n) \in \Omega(n \log n)$). □

Let us notice that *heap sort* and *merge sort* are *asymptotically optimal* in the sense that their $MT(n) \in O(n \log n)$ asymptotic upper bound meets the $MT(n) \in \Omega(n \log n)$ asymptotic lower bound from Theorem 7.4. This proves that $MT(n) \in \Theta(n \log n)$ for both of them.

8 Sorting in Linear Time

8.1 Radix sort

The abstract code for radix sort is straightforward. We assume that each element of abstract list A is a natural number of d digits, where digit 1 is the lowest-order digit and digit d is the highest-order digit.



Radix sort solves the problem of sorting – counterintuitively – by sorting on the least significant (i.e. first) digit in its first pass.

In the second pass it takes the result of the first pass and applies a stable sort to sort it on the second digit. It is important to apply stable sort so that those numbers with equal second digit remain sorted according to their first digit. Consequently, after the second pass the numbers are already sorted on the first two digits (which are the two lowest order digits).

In the third pass Radix sort takes the result of the second pass and applies a stable sort to sort it on the third digit. It is important to apply stable sort so that those numbers with equal third digit remain sorted according to their first two digits. Consequently, after the third pass the numbers are already sorted on the first three digits (which are the three lowest order digits).

This process continues until the items have been sorted on all d digits. Remarkably, at that point the items are fully sorted on the d -digit number. Thus, only d passes through the numbers are required to sort. [1]

The sorting method used in each pass can be any stable sort, but it should run in linear time in order to maintain efficiency.

Distributing sort works well on linked lists, and *counting sort* on arrays. Both of them is stable and works in linear time.

8.2 Distributing sort

Distributing sort is efficient on linked lists, and a version of radix sort can be built on it.

Remember that stable sorting algorithms maintain the relative order of records with equal keys (i.e. values).

Distributing sort is ideal auxiliary method of radix sort because of its stability and linear time complexity.

Here comes a bit more general study than needed for radix sort. When distributing sort is used for radix sort, its key function φ must select the appropriate digit.

The sorting problem: Given abstract list L of length n with element type \mathcal{T} , $r \in O(n)$ positive integer,
 $\varphi : \mathcal{T} \rightarrow 0..(r-1)$ key selection function.

Let us sort list L with stable sort, and with linear time complexity.

$\text{distributing_sort}(L : \mathcal{T}\langle \rangle ; r : \mathbb{N} ; \varphi : \mathcal{T} \rightarrow 0..(r-1))$	
$B : \mathcal{T}\langle \rangle[r]$ // array of lists, i.e. bins	
$k := 0$ to $r-1$	
	Let $B[k]$ be empty list // init the array of bins
L is not empty	
	Remove the first element x of list L
	Insert x at the end of $B[\varphi(x)]$ // stable distribution
$k := r-1$ downto 0	
	$L := B[k] + L$ // append $B[k]$ before L

Efficiency of distributing sort: The first and the last loop iterates r times, and the middle loop n times. Provided that insertion and concatenation can be performed in $\Theta(1)$ time, the time complexity of distributing sort is consequently $\Theta(n + r) = \Theta(n)$ because of the natural condition $r \in O(n)$.

8.3 Radix-Sort on lists

The next example shows how radix sort operates on an abstract list of seven 3-digit numbers with base (i.e. radix) 4. In each pass we apply distributing sort.

The input list (with a symbolic notation):

$L = \langle 103, 232, 111, 013, 211, 002, 012 \rangle$

First pass (according to the rightmost digits of the numbers):

$B_0 = \langle \rangle$

$B_1 = \langle 111, 211 \rangle$

$B_2 = \langle 232, 002, 012 \rangle$

$B_3 = \langle 103, 013 \rangle$
 $L = \langle 111, 211, 232, 002, 012, 103, 013 \rangle$

Second pass (according to the middle digits of the numbers):

$B_0 = \langle 002, 103 \rangle$
 $B_1 = \langle 111, 211, 012, 013 \rangle$
 $B_2 = \langle \rangle$
 $B_3 = \langle 232 \rangle$
 $L = \langle 002, 103, 111, 211, 012, 013, 232 \rangle$

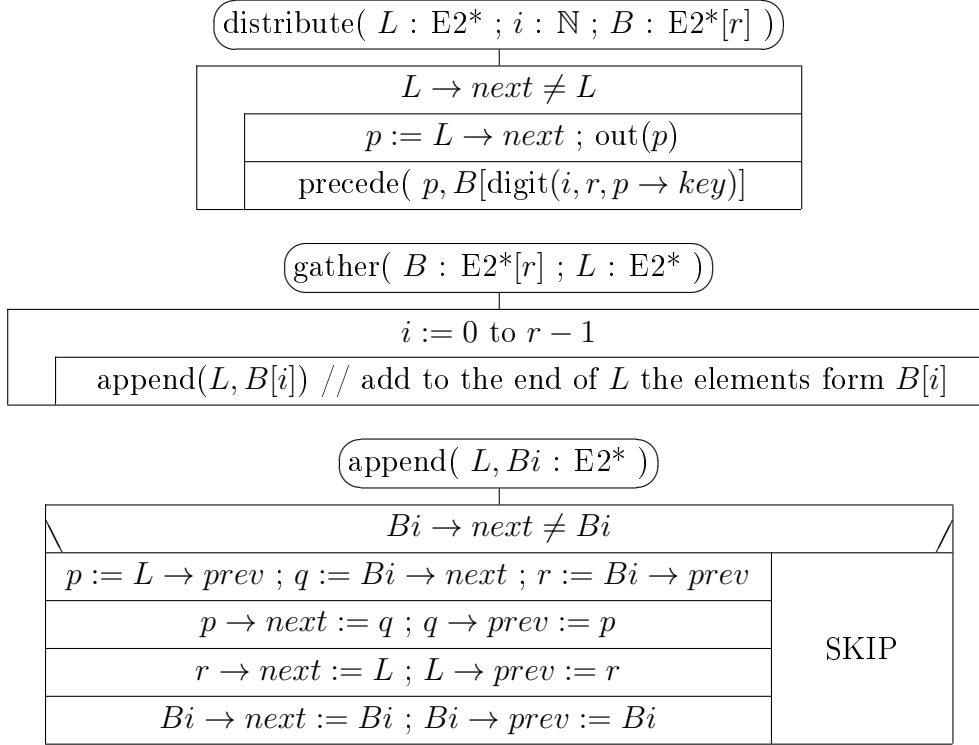
Third pass (according to the leftmost digits of the numbers):

$B_0 = \langle 002, 012, 013 \rangle$
 $B_1 = \langle 103, 111 \rangle$
 $B_2 = \langle 211, 232 \rangle$
 $B_3 = \langle \rangle$
 $L = \langle 002, 012, 013, 103, 111, 211, 232 \rangle$

In order for radix sort to work correctly, the digit sorts must be stable. Distributing sort satisfies this requirement. If distributing sort runs in linear time ($\Theta(n)$ where n is the length of the input list), and the number of digits is constant d , radix sort also runs in linear time $\Theta(d * n) = \Theta(n)$.

Provided that we have to sort linked lists where the keys of the list-elements are d -digit natural numbers with number base r , the implementation of the algorithm above is straightforward. For example, let us suppose that we have an L C2L with header, and function $\text{digit}(i, r, x)$ can extract the i th digit of number x , where digit 1 is the lowest-order digit and digit d is the highest-order digit, in $\Theta(1)$ time.

radix_sort($L : E2^*$; $d, r : \mathbb{N}$)	
$BinHead : E2[r]$ // the headers of the lists representing the bins	
$B : E2^*[r]$ // pointers to the headers	
$i := 0$ to $r - 1$	
	$B[i] := \&BinHead[i]$ // Initialize the i th pointer.
$i := 1$ to d	
	distribute(L, i, B) // Distribute L on the i th digits of keys.
	gather(B, L) // Gather form the bins back into L



Clearly, $T_{\text{append}} \in \Theta(1)$, so $T_{\text{gather}} \in \Theta(r)$ where $r = B.length$.

And $T_{\text{distribute}}(n) \in \Theta(n)$ where $n = |L|$.

Thus $T_{\text{radix_sort}}(n, d, r) \in \Theta(r + d(n + r))$.

Consequently, if d is constant and $r \in O(n)$, then $T_{\text{radix_sort}}(n) \in \Theta(n)$.

In a typical computer, which is a sequential random-access machine, we sometimes use radix sort to sort records of information that are keyed by multiple fields. For example, we might wish to sort dates by three keys: year, month, and day. We could run a sorting algorithm with a comparison function that, given two dates, compares years, and if there is a tie, compares months, and if another tie occurs, compares days. Alternatively, we could sort the information three times with a stable sort: first on day, next on month, and finally on year. [1]

8.4 Counting sort

While the previous version of radix sort is efficient on linked lists, counting sort can be applied efficiently to arrays, and another version of radix sort can be built on it.

Remember that stable sorting algorithms maintain the relative order of records with equal keys (i.e. values).

Counting sort is stable, and it is ideal auxiliary method of radix sort because of its linear time complexity.

Here comes a bit more general study than needed for radix sort. When counting sort is used for radix sort, its key function φ must select the appropriate digit.

The sorting problem: Given array $A: \mathcal{T}[n]$, $r \in O(n)$ positive integer, $\varphi: \mathcal{T} \rightarrow 0..(r-1)$ key selection function.

Let us sort array A with stable sort, and with linear time complexity so that the result is produced in array B .

counting_sort($A, B: \mathcal{T}[n]; r: \mathbb{N}; \varphi: \mathcal{T} \rightarrow 0..(r-1)$)	
$C: \mathbb{N}[r]$ // counter array	
$k := 0$ to $r-1$	
	$C[k] := 0$ // init the counter array
$i := 0$ to $n-1$	
	$C[\varphi(A[i])]++$ // count the items with the given key
$k := 1$ to $r-1$	
	$C[k] += C[k-1]$ // $C[k] :=$ the number of items with key $\leq k$
$i := n-1$ downto 0	
	$k := \varphi(A[i])$ // $k :=$ the key of $A[i]$
	$C[k]--$ // The next one with key k must be put before $A[i]$ where
	$B[C[k]] := A[i]$ // Let $A[i]$ be the last of the {unprocessed items with key k }

The first loop of the procedure above assigns zero to each element of the counting array C .

The second loop counts the number of occurrences of key k in $C[k]$, for each possible key k .

The third loop sums the number of occurrences of keys \leq than k , considering each possible key k .

The number of keys ≤ 0 is the same as the number of keys $= 0$, so the value of $C[0]$ is not changed in the third loop. Considering greater keys we have that the number of keys $\leq k$ equals to the number of keys $= k$ + the number of keys $\leq k-1$. Thus the new value of $C[k]$ can be counted with adding the new value of $C[k-1]$ to the old value $C[k]$.

The fourth loop goes on the input array in reverse direction. We put the elements of input array A into output array B : Considering any key k in the

input array, first we process its last occurrence. The element containing it is put into the last place reserved for keys $= k$, i.e. into $B[C[k] - 1]$: First we decrease $C[k]$ by one, and put this element into $B[C[k]]$. According to this reverse direction the next occurrence of key k is going to be the immediate predecessor of the actual item etc. Thus the elements with the same key remain in their original order, and we receive a stable sort.

The time complexity is clearly $\Theta(n+r)$. Provided that $r \in O(n)$, $\Theta(n+r) = \Theta(n)$, and so $T(n) \in \Theta(n)$.

Illustration of counting sort: We suppose that we have to sort numbers of two digits with number base 4 according to their right-side digit, i.e. function φ selects the rightmost digit.

The input:

	0	1	2	3	4	5
$A :$	02	32	30	13	10	12

The changes of counter array C [the first column reflects the first loop initializing counter array C to zero, the next six columns reflect the second loop counting the items with key k , for each possible key; the column labeled by Σ reflects the third loop which sums up the number of items with keys $\leq k$; and the last six columns reflect the fourth loop placing each item of the input array into its place in the output array]:

	C	02	32	30	13	10	12	Σ	12	10	13	30	32	02
0	0			1		2		2		1		0		
1	0							2						
2	0	1	2				3	5	4				3	2
3	0				1			6			5			

The output:

	0	1	2	3	4	5
$B :$	30	10	02	32	12	13

Now we suppose that the result of the previous counting sort is to be sorted according to the left-side digits of the numbers (of number base 4), i.e. function φ selects the leftmost digits of the numbers.

The input:

	0	1	2	3	4	5
$B :$	30	10	02	32	12	13

The changes of counter array $C:\mathbb{N}[4]$:

	C	30	10	02	32	12	13	Σ	13	12	32	02	10	30
0	0			1				1				0		
1	0		1			2	3	4	3	2			1	
2	0							4						
3	0	1			2			6			5			4

The output:

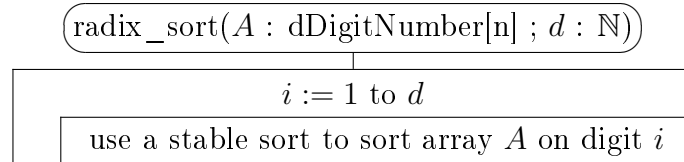
	0	1	2	3	4	5
$A :$	02	10	12	13	30	32

The first counting sort ordered the input according to the right-side digits of the numbers. The second counting sort ordered the result of the first sort according to the left-side digits of the numbers using a stable sort. Thus in the final result the numbers with the same left-side digits remained in order according to their right-side digits. Consequently in the final result the numbers are already sorted according to their both digits.

Therefore the two counting sorts illustrated above form the first and second passes of a radix sort. And our numbers have just two digits now, so we have performed a complete radix sort in this example.

8.5 Radix-Sort on arrays ([1] 8.3)

The keys of array A are d -digit natural numbers with number base r . The rightmost digit is the least significant (digit 1), and the leftmost digit is the most significant (digit d).

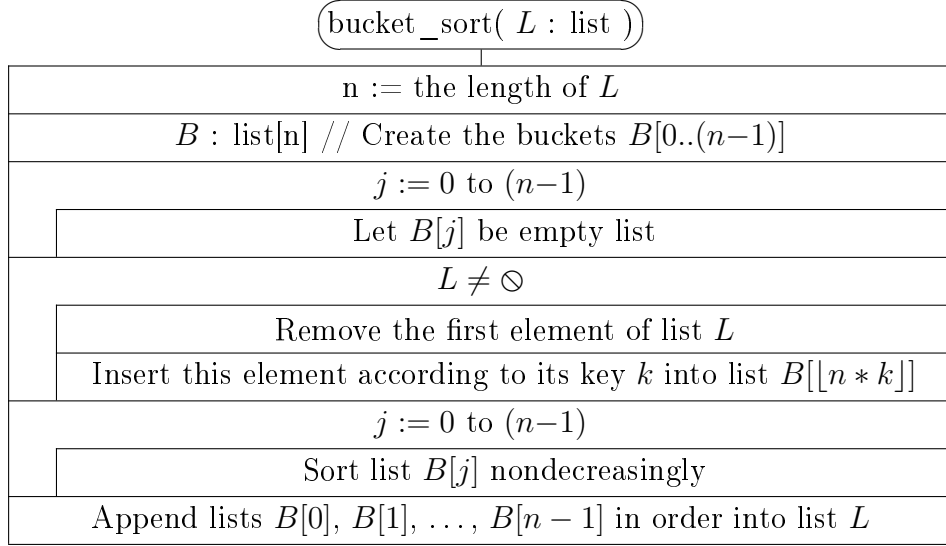


Provided that the stable sort is counting sort, the time complexity of Radix sort is $\Theta(d(n+r))$. If d is a constant and $r \in O(n)$, $\Theta(d(n+r)) = \Theta(n)$, i.e. $T(n) \in \Theta(n)$.

8.6 Bucket sort

We suppose that the items to be sorted are elements of the real interval $[0; 1)$.

This algorithm is efficient, if the keys of the input are equally distributed on $[0; 1)$. We sort the buckets with some known sorting method like insertion sort or merge sort.



Clearly $mT(n) \in \Theta(n)$. If the keys of the input are equally distributed on $[0; 1)$, $AT(n) \in \Theta(n)$. $MT(n)$ depends on the sorting method we use when sorting lists $B[j]$ nondecreasingly. For example, using insertion sort $MT(n) \in \Theta(n^2)$; using merge sort $MT(n) \in \Theta(n \log n)$.

9 Hash Tables

In everyday programming practice we often need so called dictionaries. These are collections of records with unique keys. Their operations: (1) inserting a new record into the dictionary, (2) searching for a record identified by a key, (3) removing a record identified by a key (or localized by a previous search).

Besides AVL trees, B+ trees (see them in the next semester), and other kinds of balanced search trees dictionaries are often represented and implemented with hash tables, provided that we would like to optimize the average running time of the operations. Using hash tables the average running time of the operations above is the ideal $\Theta(1)$, while the maximum of it is $\Theta(n)$. (With balanced search trees the worst case and average case performance of each key-based operation are $\Theta(\log n)$.)

Notations:

m : the size of the hash table

$T[0..(m-1)]$: the hash table

$T[0], T[1], \dots, T[m-1]$: the slots of the hash table

\odot : empty slot in the hash table, if we use direct-address tables or key collisions are resolved by chaining

E : the key of empty slots in case of open addressing

D : the key of deleted slots in case of open addressing

n : the number of records stored in the hash table

$\alpha = n/m$: *load factor*

U : the universe of keys; $k, k', k_i \in U$

$h : U \rightarrow 0..(m-1)$: hash function

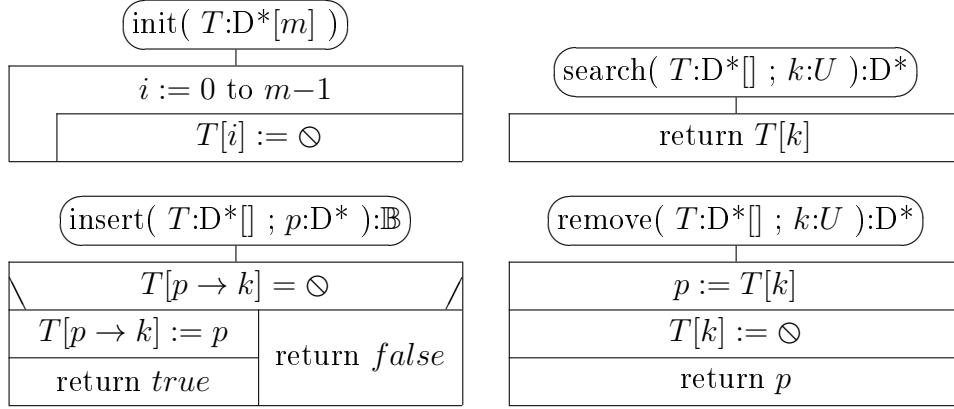
We suppose that the hash table does not contain two or more records with the same key, and that $h(k)$ can be calculated in $\Theta(1)$ time.

9.1 Direct-address tables

In case of direct address tables we do not have hash functions. We suppose that $U = 0..(m-1)$ where $m \geq n$, but m is not too big compared to n .

$T : D*[m]$ is the direct-address table. Its slots are pointers referring to records of type D. Each record has a key $k : U$ and contains satellite data. The direct-address table is initialized with \odot pointers.

D
+ $k : U$ // k is the key
+ ... // satellite data



Clearly $T_{\text{init}}(m) \in \Theta(m)$. And for the other three operations we have $T \in \Theta(1)$.

9.2 Hash tables

Hash function: Provided that $|U| \gg n$, direct address tables cannot be applied or at least applying them is wasting space. Thus we use a hash function $h : U \rightarrow 0..(m-1)$ where typically $|U| \gg m$ (the size of the universe U of keys is much greater than the size m of the hash table). The record with key k is to be stored in slot $T[h(k)]$ of hash table $T[0..(m-1)]$.

Remember that the hash table should not contain two or more records with the same key, and that $h(k)$ must be calculated in $\Theta(1)$ time.

Function $h : U \rightarrow 0..(m-1)$ is *simple uniform hashing*, if it distributes the keys evenly into the slots, i.e. any given key is equally likely to hash into any of the m slots, independently of where other items have hashed to. Simple uniform hashing is a general requirement about hash functions.

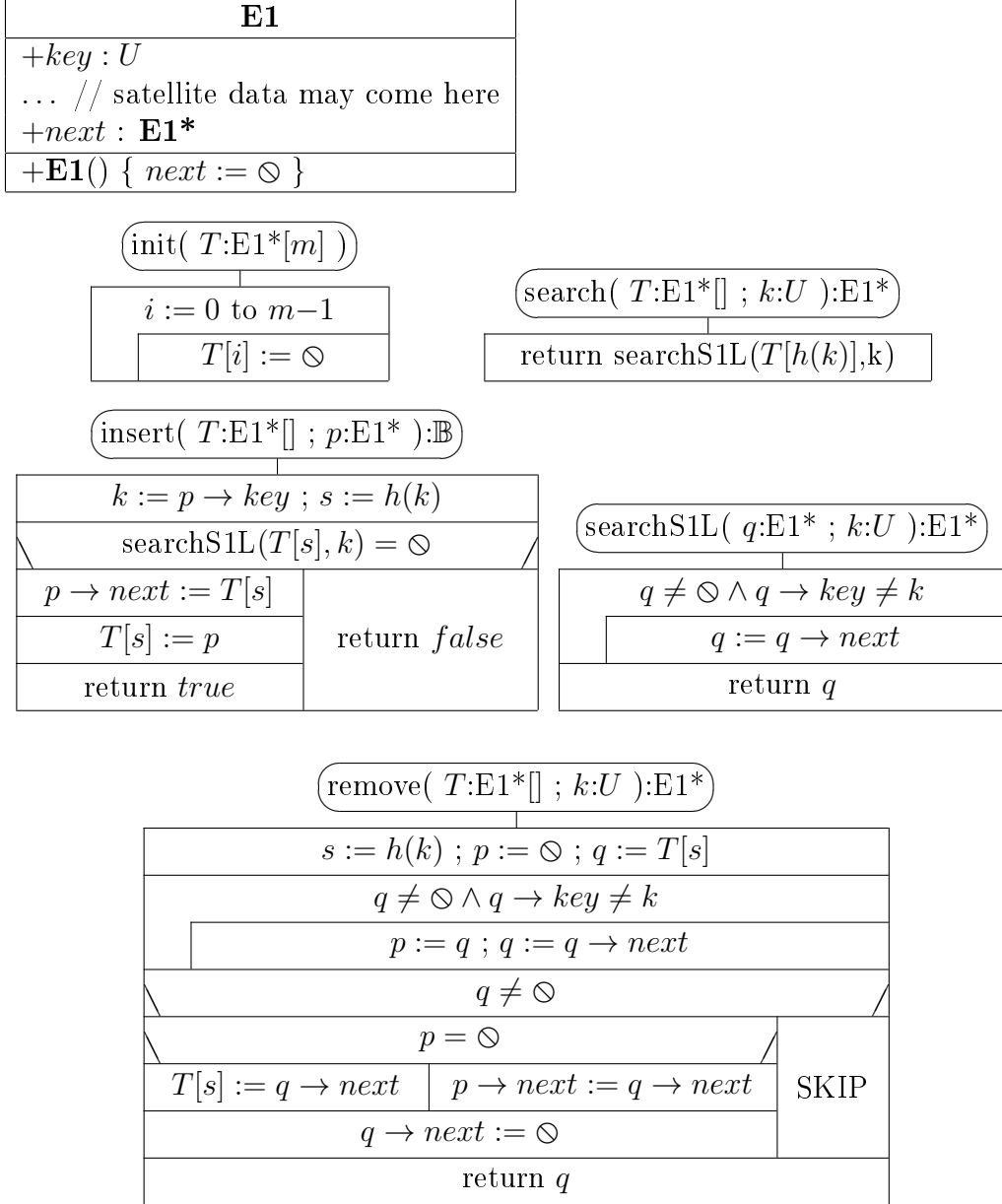
Collision of keys: Provided that $h(k_1) = h(k_2)$ for keys $k_1 \neq k_2$ we speak of key collision. Usually $|U| \gg m$, so key collisions probably happen, and we have to handle this situation.

For example, let us suppose that the keys are integer numbers and $h(k) = k \bmod m$. Then exactly those keys are hashed to slot s for which $s = k \bmod m$.

9.3 Collision resolution by chaining

We suppose that the slots of the hash table identify simple linked lists (SLL) that is $T:E1^*[m]$ where the elements of the lists contain the regular fields *key* and *next*, plus usually additional fields (satellite data). Provided that

the hash function maps two or more keys to the same slot, the corresponding records are stored in the list identified by this slot.



Clearly $T_{\text{init}}(m) \in \Theta(m)$. For the other three operations $mT \in \Theta(1)$, $MT(n) \in \Theta(n)$, $AT(n, m) \in \Theta(1 + \frac{n}{m})$.

$AT(n, m) \in \Theta(1 + \frac{n}{m})$ is satisfied, if function $h : U \rightarrow 0..(m-1)$ is *simple uniform hashing*, because the average length of the lists of the slots is equal to $\frac{n}{m} = \alpha$.

Usually $\frac{n}{m} \in O(1)$ is required. In this case $AT(n, m) \in \Theta(1)$ is also satisfied for insertion, search, and removal.

9.4 Good hash functions

Division method: Provided that the keys are integer numbers,

$$h(k) = k \bmod m$$

is often a good choice, because it can be calculated simply and efficiently. And if m is a prime number not too close to a power of two, it usually distributes the keys evenly among the slots, i.e. on the integer interval $0..(m-1)$.

For example, if we want to resolve key collision by chaining, and we would like to store approximately 2000 records with maximum load factor $\alpha \approx 3$, then $m = 701$ is a good choice: 701 is a prime number which is close to $2000/3$, and it is far enough from the neighboring powers of two, i.e. from 512, and 1024.

Keys in interval $[0; 1)$: Provided that the keys are evenly distributed on $[0; 1)$, function

$$h(k) = \lfloor k * m \rfloor$$

is also simple uniform hashing.

Multiplication method: Provided that the keys are real numbers, and $A \in (0; 1)$ is a constant,

$$h(k) = \lfloor \{k * A\} * m \rfloor$$

is a hash function. ($\{x\}$ is the fraction part of x .) It does not distribute the keys equally well with all the possible values of A .

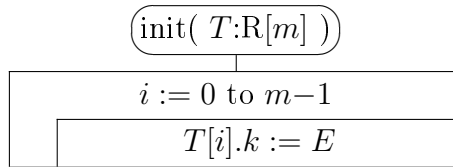
Knuth proposes $A = \frac{\sqrt{5}-1}{2} \approx 0.618$, because it is likely to work reasonably well. Compared to the division method it has the advantage that the value of m is not critical.

Each method above supposes that the keys are numbers. If the keys are strings, the characters can be considered digits of unsigned integers with the appropriate number base. Thus the strings can be interpreted as big natural numbers.

9.5 Open addressing

The hash table is $T : R[m]$. The records of type R are directly in the slots. Each record has a key field $k : U \cup \{E, D\}$ where $E \neq D$; $E, D \notin U$ are global constants in order to indicate empty (E) and deleted (D) slots.

R
+ $k : U \cup \{E, D\}$ // k is a key or it is Empty or Deleted
+ ... // satellite data



Notations for open addressing:

$h : U \times 0..(m-1) \rightarrow 0..(m-1)$: probing function

$\langle h(k, 0), h(k, 1), \dots, h(k, m-1) \rangle$: potential probing sequence

The hash table does not contain double keys.

The empty and deleted slots together are *free* slots. The other slots are *occupied*.) Instead of a single hash function we have m hash functions now:

$$h(\cdot, i) : U \rightarrow 0..(m-1) \quad (i \in 0..(m-1))$$

In open addressing we try these in this order one after the other if needed.

9.5.1 Open addressing: insertion and search, without deletion

In many applications of dictionaries (and hash tables) we do not need deletion. Insertion and search are sufficient. In this case insertion is simpler.

Let us suppose that we want to insert record r with key k into the hash table. First we probe slot $h(k, 0)$. If it is occupied and its key is not k , we try $h(k, 1)$, and so on, throughout $\langle h(k, 0), h(k, 1), \dots, h(k, m-1) \rangle$ until

- we find an empty slot, or
 - we find an occupied slot with key k , or
 - all the slots of the *potential probing sequence* have been considered but found neither empty slot nor occupied slot with key k .
- + If we find an empty slot, we put r into it. Otherwise insertion fails.

$\langle h(k, 0), h(k, 1), \dots, h(k, m-1) \rangle$ is called *potential probing sequence* because during insertion, or search (or deletion) only a prefix of it is actually generated. This prefix is called *actual probing sequence*.

The potential probing sequence must be a permutation of $\langle 0, 1, \dots, (m-1) \rangle$, which means, that it covers the whole hash table, i.e. it does not refer twice to the same slot.

The length of the actual probing sequence of insertion/search/deletion is i , iff this operation stops at probe $h(k, i-1)$.

When we search for the record with key k , again we follow the potential probing sequence $\langle h(k, 0), h(k, 1), \dots, h(k, m-1) \rangle$.

- We stop successfully when we find an occupied slot with key k .
- The search fails, if we find an empty slot, or we use up the potential probing sequence unsuccessfully.

In ideal case we have *uniform hashing*: the potential probe sequence of each key is equally likely to be any of the $m!$ permutations of $\langle 0, 1, \dots, (m-1) \rangle$.

Provided that

- the hash table does not contain deleted slots,
 - its load factor $\alpha = n/m$ satisfies $0 < \alpha < 1$, and
 - we have uniform hashing,
- + the expected length of an unsuccessful search / successful insertion is at most

$$\frac{1}{1 - \alpha}$$

- + and the expected length of a successful search / unsuccessful insertion is at most

$$\frac{1}{\alpha} \ln \frac{1}{1 - \alpha}$$

For example, the first result above implies that with uniform hashing, if the hash table is half full, the expected number of probes in a unsuccessful search (or in a successful insertion) is less than 2. If the hash table is 90% full, the expected number of probes is less than 10.

Similarly, the second result above implies that with uniform hashing, if the hash table is half full, the expected number of probes in a successful search (or in an unsuccessful insertion) is less than 1.387. If the hash table is 90% full, the expected number of probes is less than 2.559. [1].

9.5.2 Open addressing: insertion, search, and deletion

A successful deletion consists of a successful search for the slot $T[s]$ containing a given key + the assignment $T[s].k := D$ (let the slot be deleted). $T[s].k :=$

E (let the slot be empty) is not correct. For example, let us suppose that we inserted record r with key k into the hash table, but it could not be put into $T[h(k, 0)]$ because of key conflict, and we put it into $T[h(k, 1)]$. And then we delete the record at $T[h(k, 0)]$. If this deletion performed $T[h(k, 0)].k := E$, a subsequent search for key k would stop at the empty slot $T[h(k, 0)]$, and it would not find record r with key k in $T[h(k, 1)]$. Instead, deletion performs $T[h(k, 0)].k := D$. Then the subsequent search for key k does not stop at slot $T[h(k, 0)]$ (because neither it is empty nor it contains key k), and it finds record r with key k in $T[h(k, 1)]$. (Clearly the procedures of search and deletion are not changed in spite of the presence of deleted slots.)

Thus during search we go through deleted slots and we stop when

- we find the slot with the key we search for (successful search), or
- we find an empty slot or use up the the potential probe sequence of the given key (unsuccessful search).

Insertion becomes more complex because of the presence of deleted slots. During the insertion of record r with key k , we perform a full search for k but we also remember the first empty slot found during the search.

- If the search is successful, then the insertion fails (because we do not allow duplicated keys).
- If the search is unsuccessful, but some deleted slot is remembered, then we put r into it.
- If the search is unsuccessful, no deleted slot is remembered, but the search stops at an empty slot, then we put r into it.
- If the search is unsuccessful, and neither deleted nor empty slot is found, then the insertion fails, because the hash table is full.

If we use a hash table for a long time, there may be many deleted slots and no empty slot, although the table is far from being full. This means that the unsuccessful searches will check all the slots, and also the other operations slow down. So we have to get rid of the deleted slots, for example, by rebuilding the whole table.

9.5.3 Linear probing

In this subsection, and in the next two we consider three strategies in order to generate actual probing sequences, that is the adequate prefix of $\langle h(k, 0), h(k, 1), \dots, h(k, m-1) \rangle$. In each case we have primary hash function $h_1 : U \rightarrow 0..(m-1)$ where $h(k, 0) = h_1(k)$. If it is needed, starting from this slot we go on step by step throughout the slots of the hash table, according

to a well defined rule, until we find the appropriate slot, or we find that the actual operation is impossible. h_1 must be simple uniform hash function.

The simplest strategy is linear probing:

$$h(k, i) = (h_1(k) + i) \bmod m \quad (i \in 0..(m-1))$$

It is easy to implement linear probing, but we have only m different probing sequences instead of the $m!$ probing sequences needed for uniform hashing: Given two keys, k_1 and k_2 ; if $h(k_1, 0) = h(k_2, 0)$ then their whole probing sequences are the same. In addition, different probing sequences tend to be linked into continuous, long runs of occupied slots, increasing the expected time of searching. This problem is called *primary clustering*. The longer such a cluster is the the more probable that it becomes even longer after the next insertion. For example, let we have two free slots with i occupied slots between them. Then the probability that its length will be increased by the next insertion is at least $(i+2)/m$. And it may even be linked with another cluster. Linear probing may be selected only if the probability of key collision is extremely low.

9.5.4 Quadratic probing

$$h(k, i) = (h_1(k) + c_1i + c_2i^2) \bmod m \quad (i \in 0..m-1)$$

where $h_1 : U \rightarrow 0..(m-1)$ is the primary hash function; $c_1, c_2 \in \mathbb{R}; c_2 \neq 0$. The different probing sequences are not linked together, but we have only m different probing sequences instead of the $m!$ probing sequences needed for uniform hashing: Given two keys, k_1 and k_2 ; if $h(k_1, 0) = h(k_2, 0)$ then their whole probing sequences are the same. This problem is called *secondary clustering*.

Choosing the constants of quadratic probing: In this case the potential probing sequence, i.e. $\langle h(k, 0), h(k, 1), \dots, h(k, m-1) \rangle$ may have equal members which implies that it does not cover the hash table. Therefore we must be careful about selecting the constants of quadratic probing.

For example, if size m of the hash table is a power of 2, then $c_1 = c_2 = 1/2$ is appropriate. In this case

$$h(k, i) = \left(h_1(k) + \frac{i + i^2}{2} \right) \bmod m \quad (i \in 0..m-1)$$

Thus

$$(h(k, i+1) - h(k, i)) \bmod m = \left(\frac{(i+1) + (i+1)^2}{2} - \frac{i + i^2}{2} \right) \bmod m =$$

$$(i + 1) \bmod m$$

So it is easy to compute the slots of the probing sequences recursively:

$$h(k, i + 1) = (h(k, i) + i + 1) \bmod m$$

Exercise 9.1 Write the structograms of the operations of hash tables with quadratic probing ($c_1 = c_2 = 1/2$) applying the previous recursive formula.

9.5.5 Double hashing

$$h(k, i) = (h_1(k) + ih_2(k)) \bmod m \quad (i \in 0..(m-1))$$

where $h_1 : U \rightarrow 0..(m-1)$ and $h_2 : U \rightarrow 1..(m-1)$ are hash functions. The probing sequence covers the hash table, iff $h_2(k)$ and m are relative primes. It is satisfied, for example, if $m > 1$ is a power of 2 and $h_2(k)$ is odd number for each $k \in U$, or if m is prime number. For example, if m is prime number (which should not be close to powers of 2) and m' is a bit smaller (let $m' = m - 1$ or $m' = m - 2$) then

$$h_1(k) = k \bmod m$$

$$h_2(k) = 1 + (k \bmod m')$$

is an eligible choice.

In case of double hashing for each different pairs of $(h_1(k), h_2(k))$ there is a different probing sequence, and so we have $\Theta(m^2)$ different probing sequences.

Although the number of the probing sequences of double hashing is far from the ideal number $m!$ of probing sequences, its performance appears to be very close to that of the ideal scheme of uniform hashing.

Illustration of the operations of double hashing: Because $h(k, i) = (h_1(k) + ih_2(k)) \bmod m \quad (i \in 0..(m-1))$, therefore $h(k, 0) = h_1(k)$ and $h(k, i + 1) = (h(k, i) + d) \bmod m$ where $d = h_2(k)$. After calculating the place of the first probing ($h_1(k)$) we always make a step of distance d cyclically around the table..

Example: $m = 11 \quad h_1(k) = k \bmod 11 \quad h_2(k) = 1 + (k \bmod 10)$.

In the “operations” (op) column of the next table, *ins*=insert, *src*=search, *ddel*=delete. Next the *key* of the operation comes (being neither *E* nor *D*). In this table we do not handle satellite data. We show just the keys. In the next column of the table there is $h_2(key)$, but only if it is needed. Next we find the actual probing sequence. Insertion remembers the first deleted slot of the actual probing sequence if any. In such cases we underlined the index

of this slot. In column “s” we have a “+” sign for a successful, and an “−” sign for an unsuccessful operation. In the table we do not handle satellite data, we process only the keys. (See the details in section 9.5.2.)

In the last 11 columns of the table we represent the actual state of the hash table. The cells representing empty slots are simply left empty. We wrote the reserving key into each cell of occupied slots, while the cells of the deleted slots contain letter *D*.

op	key	h_2	probes	s	0	1	2	3	4	5	6	7	8	9	10
init				+											
ins	32		10	+											32
ins	40		7	+								40			32
ins	37		4	+					37			40			32
ins	15	6	4; 10; 5	+					37	15		40			32
ins	70	1	4; 5; 6	+					37	15	70	40			32
src	15	6	4; 10; 5	+					37	15	70	40			32
src	104	5	5; 10; 4; 9	−					37	15	70	40			32
del	15	6	4; 10; 5	+					37	D	70	40			32
src	70	1	4; 5; 6	+					37	D	70	40			32
ins	70	1	4; <u>5</u> ; 6	−					37	D	70	40			32
del	37		4	+					D	D	70	40			32
ins	104	5	<u>5</u> ; 10; 4; 9	+					D	104	70	40			32
src	15	6	4; 10; 5; 0	−					D	104	70	40			32

Exercise 9.2 (Programming of double hashing) Write the structograms of insertion, search, and deletion where x is the record to be inserted, and k is the key we search for, and it is also the key of the record we want to delete.

The hash table is $T[0..(m-1)]$.

In a search we try to find the record identified by the given key. After a successful search we return the position (index of the slot) of the appropriate record. After an unsuccessful search we return “−1”.

After a successful insertion we return the position of the insertion. At an unsuccessful insertion there are two cases. If there is no free place in the hash table, we return “−($m+1$)”. If the key of the record to be inserted is found at slot j , we return “−($j+1$)”.

In a deletion we try to find and delete the record identified by the given key. After a successful deletion we return the position (index of the slot) of the appropriate record. After an unsuccessful deletion we return “−1”.

Solution:

