Data Processing

Lau Chi Yung

HKOI

2017/02/18

Outline

1. Data storage
   - Integers
   - Floating points
   - Character and strings
2. Data I/O
   - Standard I/O
   - File I/O
3. Data manipulation
   - Bitwise operation
   - Type conversion
   - String functions

Outline

Integers

unsigned 8-bit integer	
0000000	0
0000001	1
0000010	2
0000011	3
0000100	4
...	
0111110	126
0111111	127
1000000	128
1000001	129
...	
1111110	254
1111111	255

signed 8-bit integer	
0000000	0
0000001	1
0000010	2
0000011	3
0000100	4
...	
0111110	126
0111111	127
1000000	-128
1000001	-127
...	
1111110	-2
1111111	-1

Integers

- 1 bit can represent 2 elements
- 8 bits can represent $2^8 = 256$ elements
- 8-bit unsigned integer represents the 256 integers in $[0, 255]$
- 8-bit signed integer represents
  - 128 negative integers $[-128, -1]$ and
  - 128 non-negative integers $[0, 127]$

Integers

- underlying addition procedure is the same for both types, the only difference is the representation of the outcome

signed			unsigned
129		10000001	-127
1	+	00000001	1
130		10000010	-126

Integers

- underlying addition procedure is the same for both types, the only difference is the representation of the outcome

signed			unsigned
129		10000001	-127
1	+	00000001	1
130		10000010	-126

- applying two's-complement flips the sign of a signed integer
- Two's-complement: flip all bits and add one

	01111110	(126)
$\Rightarrow$	10000001	
$\Rightarrow$	10000010	(-126)

Integers

- beware of overflow

8-bit unsigned	8-bit signed
9 - 10 = 255	127 + 127 = -2

Integers

C/C++	Pascal	Range
signed char	shortint	-128 127
short	smallint	-32768 32767
int long	**longint**	-2147483648 2147483647
long long	int64	-9223372036854775808 9223372036854775807

- Use **int** and **longint** for most cases
- Use **long long** and **int64** to avoid overflow

Integers

C/C++	Pascal	Range
unsigned char	byte	0 255
unsigned short	word	0 65535
unsigned int unsigned	longword	0 4294967295
unsigned long long	qword	0 18446744073709551615

- unsigned integers are rarely used, unless the constraints are deliberately designed

Outline

Floating points

Size	C/C++	Pascal	Range
32 bits	float	single	$\pm 3.40 \times 10^{38}$
64 bits	double	double	$\pm 1.78 \times 10^{308}$

Floating points

Size	C/C++	Pascal	Range
32 bits	float	single	$\pm 3.40 \times 10^{38}$
64 bits	double	double	$\pm 1.78 \times 10^{308}$

How many integers are in the range [0, 1]?

Floating points

Size	C/C++	Pascal	Range
32 bits	float	single	$\pm 3.40 \times 10^{38}$
64 bits	double	double	$\pm 1.78 \times 10^{308}$

How many integers are in the range [0, 1]?

2

How many real numbers are in the range [0, 1]?

Floating points

Size	C/C++	Pascal	Range
32 bits	float	single	$\pm 3.40 \times 10^{38}$
64 bits	double	double	$\pm 1.78 \times 10^{308}$

How many integers are in the range [0, 1]?

2

How many real numbers are in the range [0, 1]?

Infinite

Then how to store $\pm 3.40 \times 10^{38}$ in 32 bits?

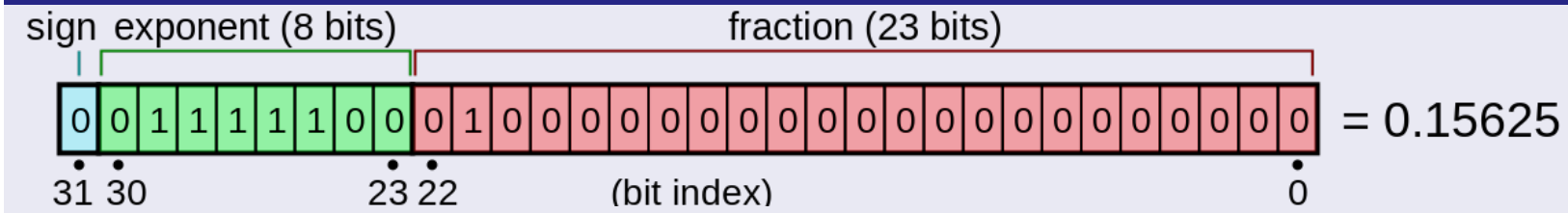
Floating points

- break a real number into **sign**, **significand** and **exponent**
- $3.14159 = 314159 \times 10^{-5}$
-

sign	significand	exponent
$+$	314159	-5

Floating points

IEEE-754 Single-precision floating-point format

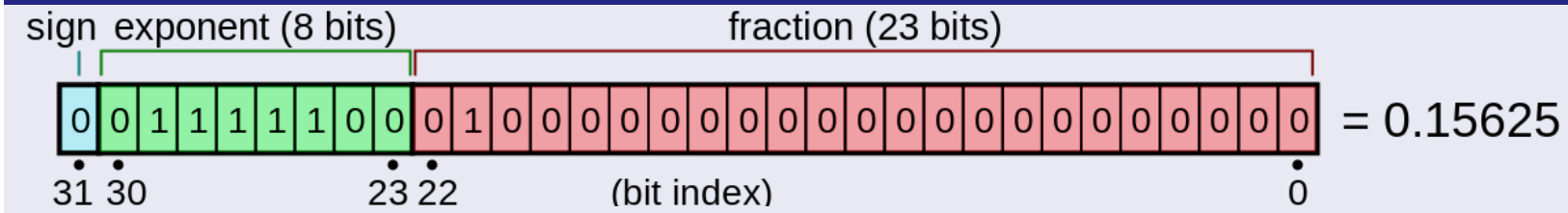


- only approximate the first 24 significant bits of a real number[1]
- 7.22 significant decimal digits
- not recommended; use **double** instead
- can represent integers in ± 16777215 exactly?
- can represent 0.5 exactly?
- can represent 0.1 exactly?

[1] but there are 23 bits in storage; go wikipedia for details

Floating points

IEEE-754 Single-precision floating-point format

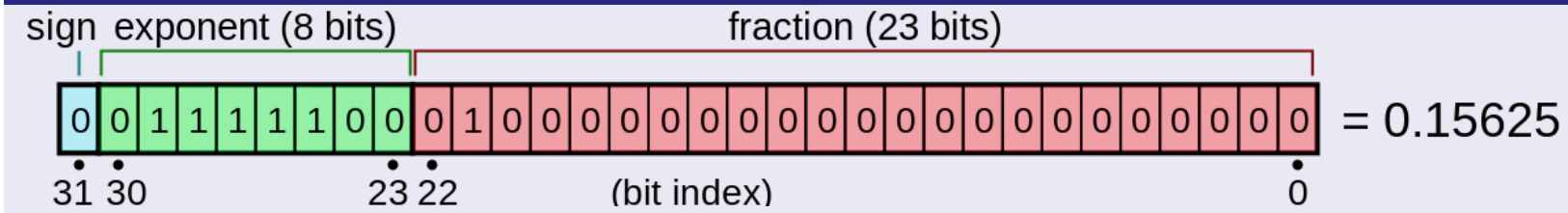


- only approximate the first 24 significant bits of a real number[1]
- 7.22 significant decimal digits
- not recommended; use **double** instead
- can represent integers in ± 16777215 exactly? yes
- can represent 0.5 exactly?
- can represent 0.1 exactly?

[1] but there are 23 bits in storage; go wikipedia for details

Floating points

IEEE-754 Single-precision floating-point format

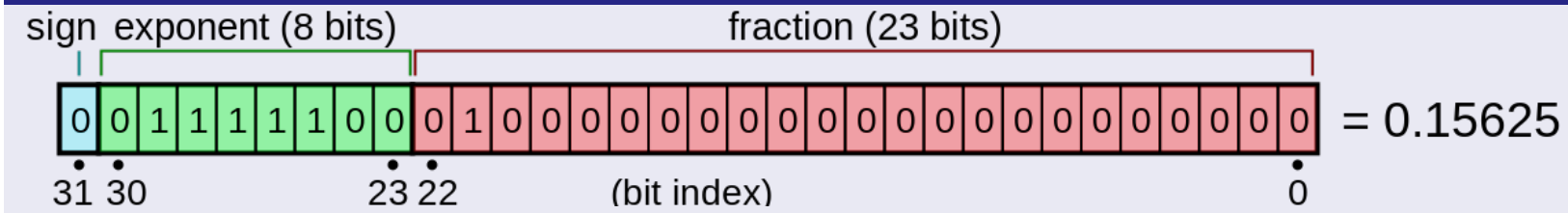


- only approximate the first 24 significant bits of a real number[1]
- 7.22 significant decimal digits
- not recommended; use **double** instead
- can represent integers in ± 16777215 exactly? yes
- can represent 0.5 exactly? yes
- can represent 0.1 exactly?

[1] but there are 23 bits in storage; go wikipedia for details

Floating points

IEEE-754 Single-precision floating-point format

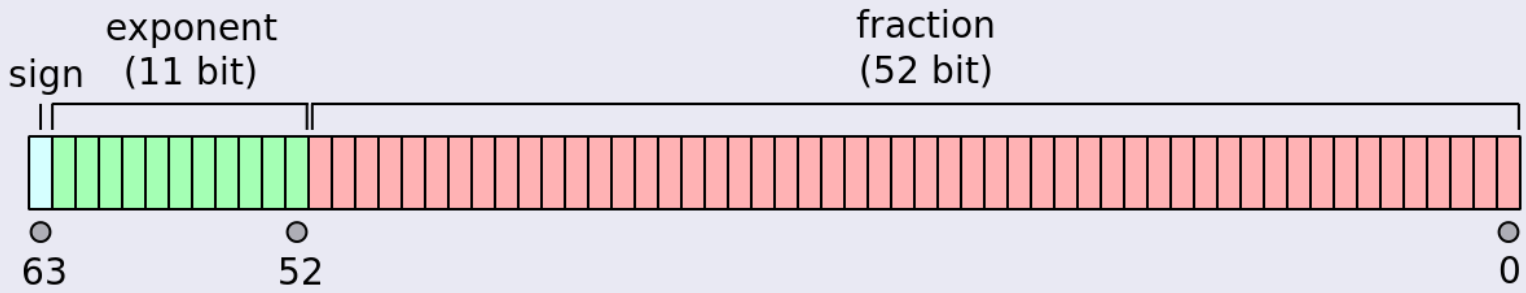


- only approximate the first 24 significant bits of a real number[1]
- 7.22 significant decimal digits
- not recommended; use **double** instead
- can represent integers in ± 16777215 exactly? yes
- can represent 0.5 exactly? yes
- can represent 0.1 exactly? no

[1] but there are 23 bits in storage; go wikipedia for details

Floating points

IEEE-754 Double-precision floating-point format



- only approximate the first 53 significant bits of a real number[2]
- 15.95 significant decimal digits
- can represent integers in ± 9007199254740991 exactly

[2] but there are 52 bits in storage; go wikipedia for details

Floating points

- usually, the task will accept a floating point answer within an error range
- beware of precision error
  - $0.5 * 2 = 1$?

Floating points

- usually, the task will accept a floating point answer within an error range
- beware of precision error
  - 0.5 * 2 = 1? yes
  - 0.1 * 10 = 1?

Floating points

- usually, the task will accept a floating point answer within an error range
- beware of precision error
  - 0.5 * 2 = 1? yes
  - 0.1 * 10 = 1? no

Floating points

- usually, the task will accept a floating point answer within an error range
- beware of precision error
  - 0.5 * 2 = 1? yes
  - 0.1 * 10 = 1? no
- use epsilon when comparing equality
  - let $\epsilon = 0.000000001$

Comparison	Expression
a = b	abs(a - b) $< \epsilon$
a $>$ b	a $>$ b + ϵ
a $\geq$ b	a $\geq$ b + ϵ
a $<$ b	a + $\epsilon <$ b
a $\leq$ b	a + $\epsilon \leq$ b

Summary

Size	C/C++	Pascal	Range
32 bits	int	longint	-2147483648 2147483647
64 bits	long long	int64	-9223372036854775808 9223372036854775807
64 bits	double	double	real number: $\pm 1.78 \times 10^{308}$ integer: ± 9007199254740991

Literals

		C/C++	Pascal
Integer	Decimal	102	102
	Octal	0146	&146
	Hexadecimal	0x66	$66
	Binary notation	0b1100110	%1100110
Floating point		1.2345	1.2345
		12345e-4	12345e-4
		0.012345e2	0.012345e2
Character		'a'	'a'
String		"apple"	'apple'

- for C/C++ **long long**, append "LL"
- e.g. "**long long** x = 9223372036854775807LL;"
- C/C++ binary notation is GCC extension

Outline

Character

- in C, **char** is just an 8-bit integer (signed or unsigned depending on machine)
- in Pascal, **char** is also 8-bit but not an integer
- i is 990

```c
char c = 'c';
char d = 99;
int i = c * 10;
```

```pascal
var
  c, d : char;
  i : longint;
begin
  c := 'c';
  d := chr(99);
  i := ord(c) * 10
end.
```

String

- in C, string is just an array of **char** with a null character at the end
- in Pascal, there are two types of strings
  - **string**: stores at most 255 characters (don't use this)
  - **ansistring**: no limit
- theoretically you can use an array of **char** in Pascal too, but it will be troublesome
- **ansistring** is null-terminated (though you don't really need to use this fact)

```
char s[] = "abc";
char t[4] = "abc";
char u[4] =
  {'a', 'b', 'c', '\0'};
```

```
var s : ansistring =
  'hello';
```

Outline

Header files for C/C++

- In C, add "**#include**<stdio.h>"
- In C++, add "**#include**<cstdio>"

One character


```c
char c;
scanf("%c", &c);
printf("%c", c);

int d = getchar();
putchar(d);
```

```
var
   c : char;
begin
   read(c);
   write(c)
end.
```

- note that `getchar()` returns **int**

One line


```c
char s[1001];
gets(s);
puts(s);
```

```pascal
var
  s : ansistring;
begin
  readln(s);
  writeln(s)
end.
```

- on Windows, EOL (end-of-line) is the two characters "\r\n" ('#13#10' in Pascal)
- on Linux, EOL is '\n' ('#10' in Pascal)
- `gets()` and `read()` reads and discards EOL, so `s` does not contain EOL
- `puts()` appends an EOL
- in C, remember to allocate one more cell for the null character
- in Pascal, **ansistring** is dynamically allocated

N lines


```c
char s[1001];
int i, n = 10;
for (i = 0; i < n; i++) {
    gets(s);
    puts(s);
}
```

```
var
  s : ansistring;
  i : longint;
  n : longint = 10;
begin
  for i := 1 to n do
  begin
    readln(s);
    writeln(s)
  end
end.
```

All lines until EOF


```c
char s[1001];
while (gets(s) != NULL)
    puts(s);
```

```
var
  s : ansistring;
begin
  while not eof do
  begin
    readln(s);
    writeln(s)
  end
end.
```

- gets returns NULL upon EOF
- **eof** is a function with no arguments

Numbers

Type	C/C++	Pascal
int	scanf("%d", &x) printf("%d", x)	read(x) write(x)
double	scanf("%lf", &x) printf("%f", x)	read(x) write(x)
long long (Linux)	scanf("%lld", &x) printf("%lld", x)	read(x) write(x)
long long (Windows)	scanf("%I64d", &x) printf("%I64d", x)	read(x) write(x)

- notice the use of %lf in scanf and %f in printf

Space separated numbers

Input

```
3    5  7
```

```c
int x;
while (scanf("%d", &x) == 1)
    printf("%d\n", x);
```

```
var
  x : longint;
begin
  while not eof do
  begin
    read(x);
    writeln(x)
  end
end.
```

- `scanf("%d")` skips all whitespaces and EOLs
- `scanf` returns the number of variables successfully read
- `read` skips only whitespaces

Formatted output


```c
char s[] = "abc";
int d = 3;
double f = -3.1875;
printf("%s\n", s);
printf("%10s\n", s);
printf("%d\n", d);
printf("%10d\n", d);
printf("%10.2f\n", f);
```

```
var
  s : ansistring = 'abc';
  d : longint = 3;
  f : double = -3.1875;
begin
  writeln(s);
  writeln(s:10);
  writeln(d);
  writeln(d:10);
  writeln(f:10:2)
end.
```

Output

```
abc
       abc
3
         3
     -3.19
```

Advanced C formatted output


```
char s[] = "abc";
printf("%s$\n", s);      //normal
printf("%1s$\n", s);     //at least 1 char
printf("%5s$\n", s);     //at least 5 char
printf("%.2s$\n", s);    //at most 2 char from s
printf("%5.2s$\n", s);   //5 char, 2 char from s
printf("%-5.2s$\n", s); //left justified
```

Output

```
abc$
abc$
  abc$
ab$
   ab$
ab   $
```

Advanced C formatted output


```c
printf("%5d$\n", 1);  //at least 5 char
printf("%-5d$\n", 1); //left justified
printf("%05d\n", 1);  //at least 5 char, 0 padded
printf("%.5d\n", 1);  //at least 5 digits
printf("%05d\n", -1); //at least 5 char, 0 padded
printf("%.5d\n", -1); //at least 5 digits
```

Output

```
    1$
1    $
00001
00001
-0001
-00001
```

Outline

File I/O

- read from a file, write to a file
- required in some contests, e.g. NOIP
- do not use file I/O in HKOI

File I/O
Method 1


```c
FILE *fin, *fout;
int x;
fin = fopen("in.txt", "r");
fout = fopen("out.txt", "w");
fscanf(fin, "%d", &x);
fprintf(fout, "%d", x);
fclose(fin);
fclose(fout);
```

```
var
  fin, fout : text;
  x : longint;
begin
  assign(fin, 'in.txt');
  reset(fin);
  assign(fout, 'out.txt');
  rewrite(fout);
  read(fin, x);
  write(fout, x);
  close(fin);
  close(fout)
end.
```

- notice the different use of I/O functions
- remember to close the files

File I/O
Method 2


```c
int x;
freopen("in.txt", "r",stdin);
freopen("out.txt", "w", stdout);
scanf("%d", &x);
printf("%d", x);
```

```
var
  x : longint;
begin
  assign(input, 'in.txt');
  reset(input);
  assign(output, 'out.txt');
  rewrite(output);
  read(x);
  write(x)
end.
```

- redirect standard I/O to the files and then use ordinary I/O functions

Outline

Bitwise operation

Operation	C/C++	Pascal
AND	&	and
OR	|	or
XOR	^	xor
NOT	~	not
Shift left	<<	<<
Shift right	>>	>>

AND

- used to test if a bit is one
  - test the third bit:

```
  10000100
AND 00000100
------------
  00000100
```

  - test the fourth bit:

```
  10000100
AND 00001000
------------
  00000000
```

- used to clear a bit
  - clear the third bit:

```
  10000100
AND 11111011
------------
  10000000
```

OR

- used to set a bit to one
- set the third and fourth bit:

$$\begin{array}{rr} & 10000100 \\ \text{OR} & 00001100 \\ \hline & 10001100 \end{array}$$

XOR

- used to flip a bit
- flip the third and fourth bit:

$$\begin{array}{rr} & 10000100 \\ \text{XOR} & 00001100 \\ \hline & 10001000 \end{array}$$

NOT

- used to flip all bits

$$\begin{array}{ll} \text{NOT} & 10000100 \\ \hline & 01111011 \end{array}$$

Shift left

- a quick way to multiply by a power of 2

$$\begin{array}{rr} & 12 \\ << & 1 \\ \hline & 24 \end{array}$$

Shift right

- a quick way to divide by a power of 2

$$\begin{array}{rr} & 12 \\ >> & 2 \\ \hline & 3 \end{array}$$

- better not use this on negative numbers
  - -2 >> 1 = -1?
  - -2 >> 1 = 127?

Two's complement on signed integers

still remember?

C/C++	Pascal
(~x) + 1 == -x	(not x) + 1 = -x

Outline

Type conversion


```
var
  a : longint;
  b : double;
  c : int64;
begin
  a := 100000;
  b := a;
  c := a;
  writeln(a, ',', b:19:0, ',', c);

  b := 90071992547409 91;
  a := trunc(b); {a = -1}
  c := trunc(b);
  writeln(a, ',', b:19:0, ',', c);

  c := 9223372036854775807;
  a := c; {a = -1}
  b := c; {b = 9223372036854775800}
  writeln(a, ',', b:19:0, ',', c);
end.
```

- when assigning to a different type, prescision may lose

Number to string


```c
char s[10];
sprintf(s, "%d", 123);
```

```pascal
var
  s : ansistring;
begin
  str(123, s)
end.
```

String to number


```c
int x;
sscanf("123", "%d", &x);
```

```pascal
var
  x : longint;
begin
  val('123', x)
end.
```

Outline

Header files for C/C++

- In C, add "**#include**<string.h>"
- In C++, add "**#include**<cstring>"
  - note: `cstring` is different from `string`

String functions

- it is not a must to learn these functions, they can easily be implemented in one or two for-loops
- learning these functions can help you shorten your code to make it more readable
- avoid mistakes in implementing these functions by yourselves

strlen / length


```
int len = strlen("abc");
```

```
var
  len : longint;
begin
  len := length('abc')
end.
```

- `len` is 3

strlen / length

- time complexity of `length()`: $O(1)$ on **ansistring**
- time complexity of `strlen()`:

strlen / length

- time complexity of `length()`: $O(1)$ on **ansistring**
- time complexity of `strlen()`: $O(N)$
- do not:

```
while (i < strlen(s)) {
    ...
}
```

- rather, save the length in a varible first
- note that `strlen()` returns an unsigned integer, beware of overflow

strcmp / strncmp

`strcmp("abc", "def")` | `'abc' < 'def'`

- dictionary ordering
- `strcmp(s, t)` < 0 when s $<$ t returns
- `strcmp(s, t)` $== 0$ when s $==$ t returns
- `strcmp(s, t)` > 0 when s $>$ t returns
- `strncmp(s, t, n)` compares at most n characters
- in Pascal, simply use $<$, $>$, $<>$, $=$ etc.

strstr / strchr / pos


```c
char target[] = "abc";
char *p =
  strstr(target, "bc");
char *q =
  strchr(target, 'c');
//p == &target[1]
//q == &target[2]
```

```
var
  target : ansistring;
  i, j : longint;
begin
  target := 'abc';
  i := pos('bc', target);
  j := pos('c', target)
  {i = 2, j = 3}
end.
```

- `strstr`(target, s) returns
  - the pointer of the first occurrence of `s` in `target`
  - `NULL` if not found
- pos(s, target) returns
  - the index of the first occurrence of `s` in `target`
  - 0 if not found
- note that **ansistring** is 1-based

strcpy / strncpy / copy


```c
char target1[10];
char target2[10];
char s[] = "abc";
strcpy(target1, s);
strncpy(target2, s + 1, 2);
```

```pascal
var
  target1 : ansistring;
  target2 : ansistring;
  s : ansistring = 'abc';
begin
  target1 := copy(s, 1,
                  length(s));
  target2 := copy(s, 2, 2)
end.
```

- `target1` = "abc"
- `target2` = "ab" (not null-terminated in C/C++)
- `target` and `s` should not overlap
- if `n` $\leq$ `strlen(s)`, `target` may not be null-terminated
- if `n` $>$ `strlen(s)`, `target` will be padded with zeroes
- `strcpy` and `strncpy` return `target`

strcat / concat


```
char s[10] = "ab";
strcat(strcat(s, "cd"), "ef");
```

```
var
  s : ansistring = 'ab';
begin
  s := concat(s, 'cd', 'ef')
end.
```

- concatenate
- for `strcat`, arguments should not overlap
- `strcat` returns first argument
- `concat` accepts many arguments and creates a new string

insert


```c
char s[10] = "ad";
char t[] = "bc";
int tlen = strlen(t);
memmove(s + 1 + tlen,
        s + 1,
        strlen(s + 1));
memcpy(s + 1, t, tlen);
```

```
var
  s : ansistring = 'ad';
begin
  insert('bc', s, 2)
end.
```

- s is "abcd"
- `memmove` allows memory overlap, while `strcpy` doesn't

delete


```
var
   s : ansistring = 'azzbc';
begin
 delete(s, 2, 2)
end.
```

- s is now 'abc'

Summary

C/C++		Pascal	
return value	function	return value	function
unsigned integer	strlen(s)	integer	length(s)
-1, 0, 1	strcmp(s, t)	boolean	$<$, $>$, $<>$, $=$, etc
-1, 0, 1	strncmp(s, t, n)	-	-
char*	strstr(target, s)	longint	pos(s, target)
char*	strchr(target, c)	longint	pos(c, target)
char*	strcpy(target, s)	ansistring	copy(s, index, count)
char*	strncpy(target, s, n)		
char*	strcat(s1, s2)	ansistring	concat(s1, s2, ..., sn)
-	-	-	insert(s, target, index)
-	-	-	delete(target, index, count)

Practice

HKOJ01001	TeX Processing
HKOJ01013	Internet Usage Bills
HKOJ01007	Packet Re-assembly
HKOJ10001	Data, Data, Everywhere