>

Background know-how

Limits of Floating-Point arithmetic in C

```
#include <stdio.h>
int main(void) {
   double x=0.7;
   int i = 0;
   while(i < 10) {
         x = 11.0 * x – 7.0;
         printf("%d: %.20lf\n",i,x);
         i=i+1;
   }
}
```

The result of the C-program is rubbish. In the last round it is
y = -1127140547773912.5

Limits of Floating-Point arithmetic in Maple

> $restart; x := \frac{7.0}{10};$

$$x := 0.7000000000 \quad (1)$$

```
> for i from 1 to 30 do
    x := 11·x − 7;
  end do:
```

> $x;$

$$0.700000000 \quad (2)$$

> $restart; x := \frac{1.0}{3}:$

> **for** i **from** 1 **to** 30 **do**
  $x := 3 \cdot x - \frac{2}{3};$
  end do:

> $x;$

$$-10294.22328 \quad (3)$$

>

```
> x := 0 : t := time( ) :
  for i from 1 to 5000000 do
    r := rand( ) mod 10;
    for j from 1 to r do
      x := x + 1;
    end do:
  end do:
```

$x, time() - t;$

$$22492822, 91.588 \qquad (4)$$

Numbers, their representations and
more and less native number representations for a digital computer

numbers can be elements from various sets. e.g. $x \in \mathbb{Z}$, $x \in \mathbb{N}$.
each number has various representations. e.g.
17
XVII
IIIII IIIII IIIII II

usually, we encode numbers with the help of base-10 digits, i.e. the alphabet
$\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
A string $s = (a_{n-1} a_{n-1} \ldots a_1 a_0) \in \Sigma_n$ is then interpreted as

$$\sum_{i=0}^{n-1} a_i \cdot 10^i \ : \ \text{Example: } 17 = 1 \cdot 10^1 + 7 \cdot 10^0$$

What happens, if we use another base, another alphabet?
E.g. with "bits", we have:
$\Sigma_2 = \{0, 1\} \quad 17_{10} = 1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 10001_2$

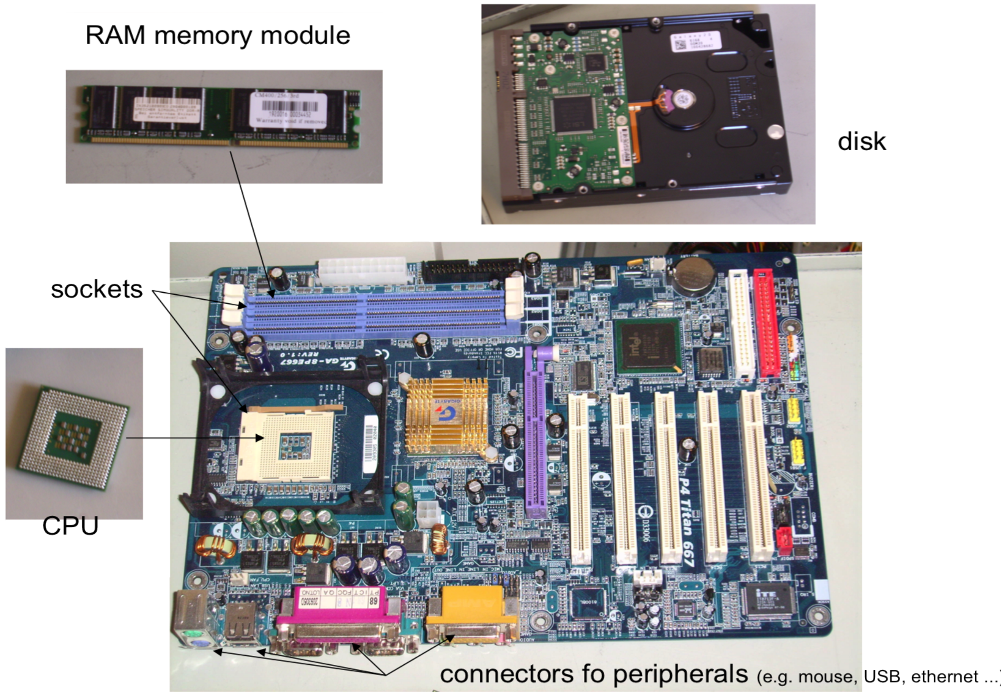
$\Sigma_{16} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, a, b, c, d, e, f\}$
$17_{10} = 1 \cdot 16^1 + 1 \cdot 16^0 = 0x11$ (so called hex numbers)

integer variables of fixed length are the most natural and mostly used kind of variables

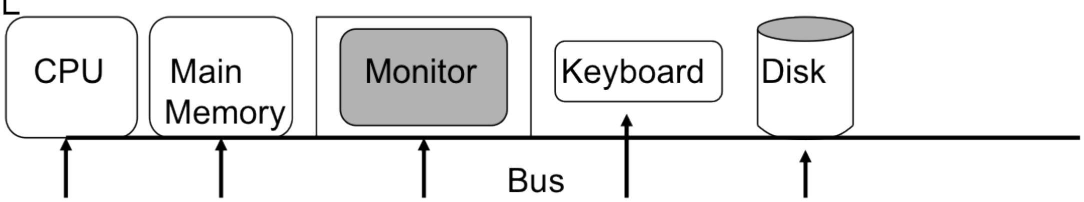
Bitstrings are interpreted as numbers in the dual number system.

```
0 1 0 0 0 1 1 0 0 0 1 0 1 0 1 1 0 1 0 1 0 0 0 0 1 0 0 0 0 1 0 1
: :                                                         : :
: :                                                         : :
: :                                                         : '- - - -  bit 0
: :                                                         '- - - - -  bit 1
: :
: '- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -  bit 30
'- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -  bit 31 (MSB)
```

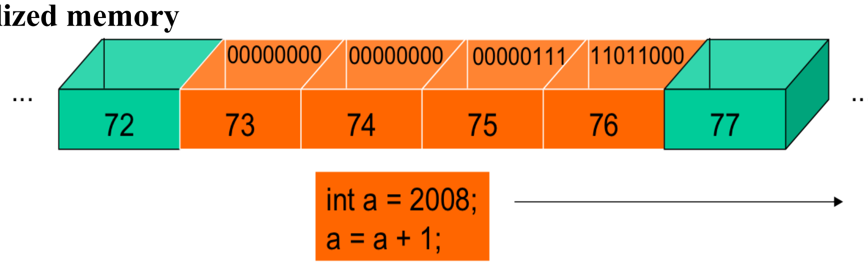
The value then is $bit_{31} \cdot 2^{31} + bit_{30} \cdot 2^{30} + \ldots + bit_0 \cdot 2^0$.



My idealized computer



My idealized memory



How to compute with binary numbers?

□ base-2 base-10

sum:

```
      1 0 1 1           9 9
   +    1 1 1 1       + 1 1 3
      1 1 1 0         1 0 2
```

□ product:

```
1011 · 101
      1011
     0000
    1011
    110111
```

Generalized binary fixed-point and floating-point numbers

0.75

$$0.75 = 1\cdot\frac{1}{2} + 1\cdot\frac{1}{4} = 0.11_2$$

0.7

$$0.7 = 1\cdot\frac{1}{2} + 0\cdot\frac{1}{4} + 1\cdot\frac{1}{8} + 1\cdot\frac{1}{16} + ...$$

the first 64 bits:
0.1011001100110011001100110011001100110011001100110011001100110011

0.7 is a periodic number in the binary system.

floating point variables

0/1 sequences are interpreted as sign (s) , mantissa (m) and exponent (p)

0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1

s p m

the resulting number then is $s\cdot m\cdot 2^p$

In the IEEE-754 standard, 127 is added to the exponent, and the leading 1 of the mantissa is not stored. The exponent has 8 bits and the mantissa 23 explicit bits, thus 24 implicit bits.

0 1 0 0 0 0 0 0 1 1 0 1 0 1 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1

s p' m'

-> representation errors in IEEE format is not avoidable
-> x = 0.7; x = 11.0 · x – 7.0; increases the error by a factor of 10

Wrong results in spite of exact computations

Expand

$$expand\left(\frac{x\cdot(x^3+3)}{x\cdot(x+1)}\right);$$

$$\frac{113797268899788329962 51}{22492823} \quad (5)$$

The case

The fibonacci series is defined as follows:
fib(0) = 0, fib(1) = 1 and fib(n+1) = fib(n-1) + fib(n)
We would like to know whether f(n) might be expressible as

$$fib(n) = \frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right)$$

We would like to get some information fast and without lots of hand work.
How can we start working at the exercise? How can Maple help us?

Solution:
Relativly soon, it is clear that:

$$\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^0 - \left(\frac{1-\sqrt{5}}{2}\right)^0\right)$$

$$0 \quad (6)$$

and

$$\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^1 - \left(\frac{1-\sqrt{5}}{2}\right)^1\right)$$

$$1 \quad (7)$$

Additionally, it must be true that

$$\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n-1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n-1}\right) + \frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n} - \left(\frac{1-\sqrt{5}}{2}\right)^{n}\right) = \frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right)$$

Some large numbers can quickly be tested, the expression may be simpified via the command simlify. An example is 876:

$$simplify\left(\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{876-1}-\left(\frac{1-\sqrt{5}}{2}\right)^{876-1}\right)+\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{876}-\left(\frac{1-\sqrt{5}}{2}\right)^{876}\right)-\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{876+1}-\left(\frac{1-\sqrt{5}}{2}\right)^{876+1}\right)\right)$$

(8)

The procedure becomes by far more tricky, if we want Maple to show equality for general n. Sometimes, it helps to expand the expression.

$$expand\left(\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n-1}-\left(\frac{1-\sqrt{5}}{2}\right)^{n-1}\right)+\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n}-\left(\frac{1-\sqrt{5}}{2}\right)^{n}\right)-\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1}-\left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right)\right);$$

$$\frac{1}{5}\,\frac{\sqrt{5}\left(\frac{1}{2}+\frac{1}{2}\sqrt{5}\right)^{n}}{\frac{1}{2}+\frac{1}{2}\sqrt{5}}-\frac{1}{5}\,\frac{\sqrt{5}\left(\frac{1}{2}-\frac{1}{2}\sqrt{5}\right)^{n}}{\frac{1}{2}-\frac{1}{2}\sqrt{5}}+\frac{1}{10}\sqrt{5}\left(\frac{1}{2}+\frac{1}{2}\sqrt{5}\right)^{n}-\frac{1}{10}\sqrt{5}\left(\frac{1}{2}-\frac{1}{2}\sqrt{5}\right)^{n}-\frac{1}{2}\left(\frac{1}{2}+\frac{1}{2}\sqrt{5}\right)^{n}-\frac{1}{2}\left(\frac{1}{2}-\frac{1}{2}\sqrt{5}\right)^{n}$$ **(9)**

$$g := simplify\left(expand\left(\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n-1}-\left(\frac{1-\sqrt{5}}{2}\right)^{n-1}\right)+\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n}-\left(\frac{1-\sqrt{5}}{2}\right)^{n}\right)-\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1}-\left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right)\right)\right);$$

$$0$$ **(10)**

$$h := simplify\left(\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n-1}-\left(\frac{1-\sqrt{5}}{2}\right)^{n-1}\right)+\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n}-\left(\frac{1-\sqrt{5}}{2}\right)^{n}\right)-\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1}-\left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right)\right);$$

$$-\frac{1}{10}\sqrt{5}\left(-\left(\frac{1}{2}+\frac{1}{2}\sqrt{5}\right)^{n}-\sqrt{5}\left(\frac{1}{2}+\frac{1}{2}\sqrt{5}\right)^{n}+\left(\frac{1}{2}-\frac{1}{2}\sqrt{5}\right)^{n}-\sqrt{5}\left(\frac{1}{2}-\frac{1}{2}\sqrt{5}\right)^{n}+2\left(\frac{1}{2}+\frac{1}{2}\sqrt{5}\right)^{n+1}-2\left(\frac{1}{2}-\frac{1}{2}\sqrt{5}\right)^{n+1}\right)$$ **(11)**

$$is\left(simplify\left(\left(\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n-1}-\left(\frac{1-\sqrt{5}}{2}\right)^{n-1}\right)+\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n}\right.\right.\right.\right.$$

$$-\left(\frac{1-\sqrt{5}}{2}\right)^{n}\Bigg)\Bigg) - \left(\frac{1}{\sqrt{5}}\cdot\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right)\right) = 0\Bigg);$$

false **(12)**

\>

\>

\>

\> *restart*; *sff* := [*seq*(0, *i* = 0 ..99)] :

\> *sff*[0] := 0; *sff*[1] := 1;

Error, out of bound assignment to a list

$sff_1 := 1$ **(13)**

\> *sff*[0 + 1] := 0; *sff*[1 + 1] := 1;

$sff_1 := 0$

$sff_2 := 1$ **(14)**

\> *sff*[2 + 1] := *sff*[0 + 1] + *sff*[1 + 1];

$sff_3 := 1$ **(15)**

```
> for i from 3 to 99 do
    sff[i + 1] := sff[i − 1 + 1] + sff[i − 2 + 1];
    if i = 97 then print(sff[i + 1]) fi;
  end do:
```

83621143489848422977 **(16)**

\> *restart*; *sff* := *Array*(0 ..10000, *fill* = 0) :

\>

\> *sff*[0] := 0; *sff*[1] := 1;

$sff_0 := 0$

$sff_1 := 1$ **(17)**

\> *sff*[2] := *sff*[0] + *sff*[1];

$sff_2 := 1$ **(18)**

```
> for i from 3 to 10000 do
    sff[i] := sff[i − 1] + sff[i − 2];
    if i = 97 then print(sff[i]) fi;
  end do:
```

83621143489848422977 **(19)**

\> *sff1* := *sff*[0]; *sff2* := *sff*[1]; *sff3* := *sff*[2];

sff1 := 0

sff2 := 1

sff3 := 1 **(20)**

```
> for i from 3 to 10000 do
    sff4 := sff3 + sff2;
    sff2 := sff3 : sff3 := sff4;
    if i = 97 then print(sff3) fi;
  end do:
```

$$8362114348984842297 \tag{21}$$

>