

MI-ARI

(Computer arithmetics)

winter semester 2017/18

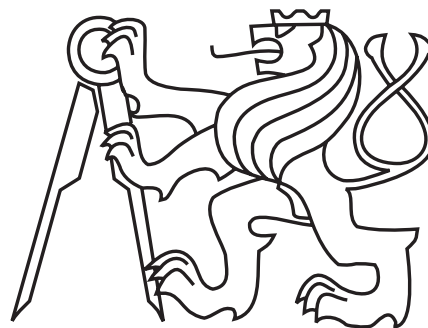
CS. Number systems and basic operations

© Alois Pluháček Pavel Kubalík, 2017

Department of digital design

Faculty of Information technology

Czech Technical University in Prague



<https://edux.fit.cvut.cz/courses/MIE-ARI/lectures>

CS. Number systems and basic operations

- Number systems
- Number formats
 - Basic operation in general format
- Binary adders
- Format respecting addition
- Subtractor
- Subtractor using adder
- Signed number representation
 - Sign and magnitude
 - Radix complement
 - Diminished radix complement
 - Biased code
 - * Type 0 of biased code
 - * Type 1 of biased code

Number systems

number systems

- **position number systems**

- **standard**

- * **decimal**

- * **binary (or dyadic)**

- * **octal**

- * **hexadecimal**

- * **ternary (or triadic)**

- etc.

- **non-standard number systems**

- * **negative radix number systems** (Polish system)

- * **signed digit number systems**

- * **multiple radix number systems**

- and others.

- **non-position number systems**

- **so called Roman numerals**

- **residue number systems (RNS)**, (Czech system)

- etc.

Standard number system

Standard number system –position number system

$$A \sim a_n a_{n-1} \dots a_0, a_{-1} \dots a_{-m}$$

$$A = \sum_{i=-m}^n a_i z^i$$

$$z \geq 2$$

radix (or base) of system — natural number

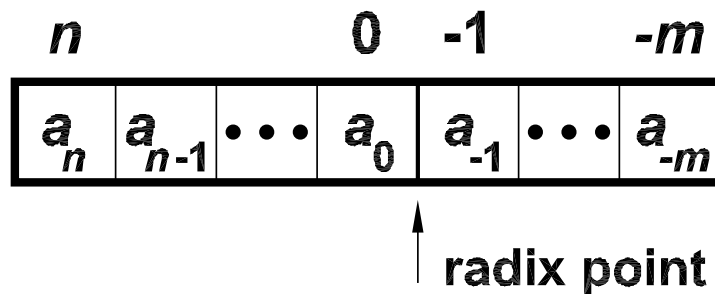
$$a_i \in \langle 0; z \rangle$$

digits — non-negative integers

$$0 \leq A \leq z^{n+1} - z^{-m}$$

!!! Negative numbers can not be represented !!!

Number formats



$n \dots$ highest position
 $-m \dots$ lowest position

$$A \sim a_n a_{n-1} \dots a_0, a_{-1} \dots a_{-m}$$

$$A = a_n z^n + a_{n-1} z^{n-1} + \dots + a_0 + a_{-1} z^{-1} \dots a_{-m} z^{-m}$$

$z \dots$ **radix** (or base) of the number system

$$\mathcal{Z} = z^{n+1} \quad \text{module of the format — it is out of format}$$

$$\varepsilon = z^{-m} \quad \text{unit of the format — smallest positive number in the format}$$

representable numbers A :
$$0 \leq A = k \cdot \varepsilon < \mathcal{Z} ,$$

k is an integer

$$k = A/\varepsilon = A^* \implies A = A^* \text{ of units } \varepsilon$$

A^*  digit number at notation of A

ε  radix point position

Basic operation in general format

Addition and subtraction

the same format of both operands and the result:

$$\left. \begin{array}{l} A = A^* \cdot \varepsilon \\ B = B^* \cdot \varepsilon \end{array} \right\} \Rightarrow A \pm B = A^* \cdot \varepsilon \pm B^* \cdot \varepsilon = (A^* \pm B^*) \cdot \varepsilon$$

Ex.: $z = 10$, $\mathcal{Z} = 10 = 10^{n+1}$, $\varepsilon = 0,01 = 10^{-m}$
 $n = 0$, $m = 2$ (or $-m = -2$)

$$A = 1,23 \Rightarrow A^* = 1,23/0,01 = 123$$

$$B = 4,56 \Rightarrow B^* = 4,56/0,01 = 456$$

$$\begin{aligned} A + B &= 1,23 + 4,56 = (123 + 456) \cdot 0,01 = \\ &= 579 \cdot 0,01 = 5,79 \end{aligned}$$

different formats:

transformation numbers into suitable format — zeroes adding

$$\text{Ex.: } 1,234 + 56,7 = 01,234 + 56,700 = 57,934$$

Conclusion: Addition and subtraction (in general format) can be easily transformed on addition and subtraction of integers.

Multiplication:

$$\left. \begin{array}{l} A = A^* \cdot \varepsilon_A \\ B = B^* \cdot \varepsilon_B \end{array} \right\} \Rightarrow \begin{array}{l} A \cdot B = A^* \cdot \varepsilon_A \cdot B^* \cdot \varepsilon_B = \\ = (A^* \cdot B^*) \cdot \varepsilon_A \cdot \varepsilon_B \end{array}$$

Ex.: $z = 10$

$$Z_A = 10, \quad \varepsilon_A = 0,01, \quad n_A = 0, \quad m_A = 2$$

$$Z_B = 100, \quad \varepsilon_B = 0,1, \quad n_B = 1, \quad m_B = 1$$

$$\begin{aligned} 7,01 \cdot 80,3 &= (701 \cdot 803) \cdot 0,001 = \\ &= 562\,903 \cdot 0,001 = 562,903 \end{aligned}$$

Conclusion: Multiplication(in general format) can be easy transformed on addition a subtraction of integers.

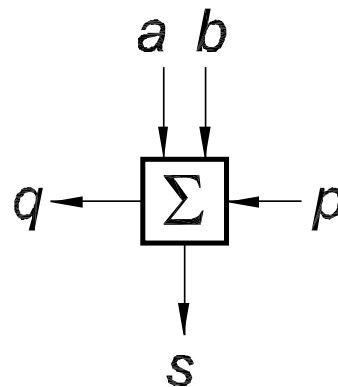
Binary adders

Full-adder

a	b	p	q	s
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

$$\begin{aligned}s &= a \oplus b \oplus p = \\ &= \bar{a}\bar{b}p + \bar{a}b\bar{p} + a\bar{b}\bar{p} + abp\end{aligned}$$

$$\begin{aligned}q &= M_3(a, b, p) = \\ &= ab + ap + bp = \\ &= ab \oplus ap \oplus bp = \\ &= ab + (ap \oplus bp)\end{aligned}$$



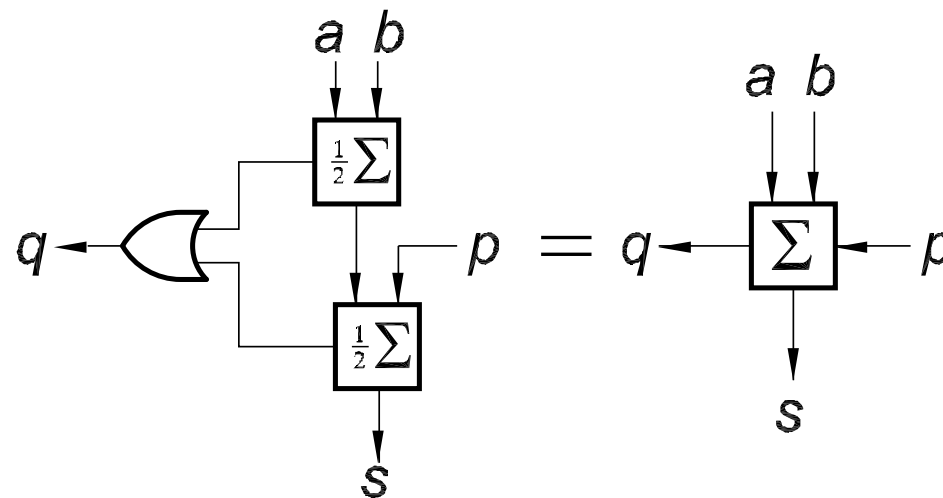
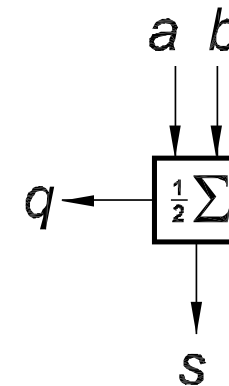
Half-adder

a	b	q	s
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

$$s = a \oplus b$$

$$= \bar{a}b + a\bar{b}$$

$$q = a \cdot b$$



Ripple-carry adder (also *parallel adder*)

$$n = 3$$

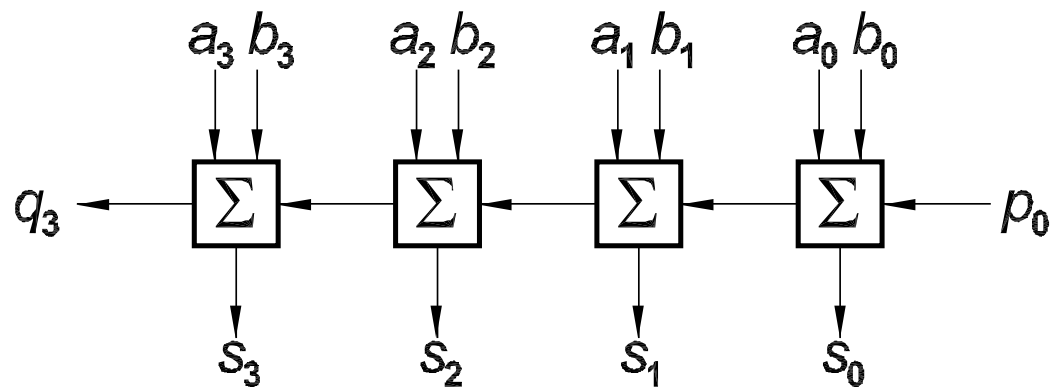
$$\mathcal{Z} = 16$$

$$A \sim a_3 a_2 a_1 a_0$$

$$B \sim b_3 b_2 b_1 b_0$$

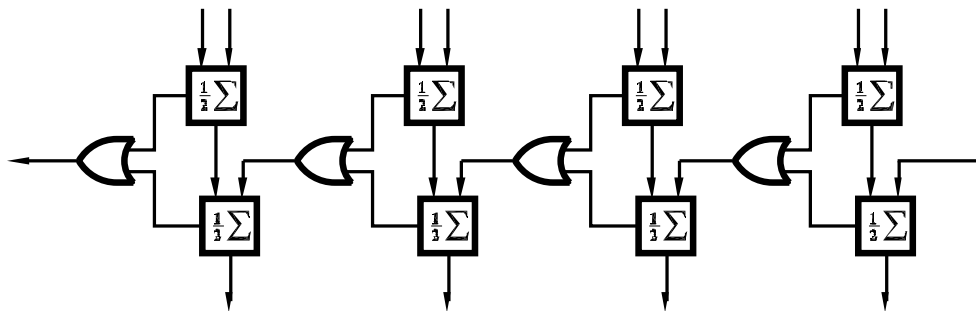
$$S \sim s_3 s_2 s_1 s_0$$

$$p_{i+1} = q_i$$



$$S = A + B + p_0 - q_n \cdot \mathcal{Z}$$

the same using half-adders



Format respecting addition

Output of adder: $S = A + B + p_0 - q_n \cdot Z$

Let $p_0 = 0$

(or half adder is used in zero order of adder):

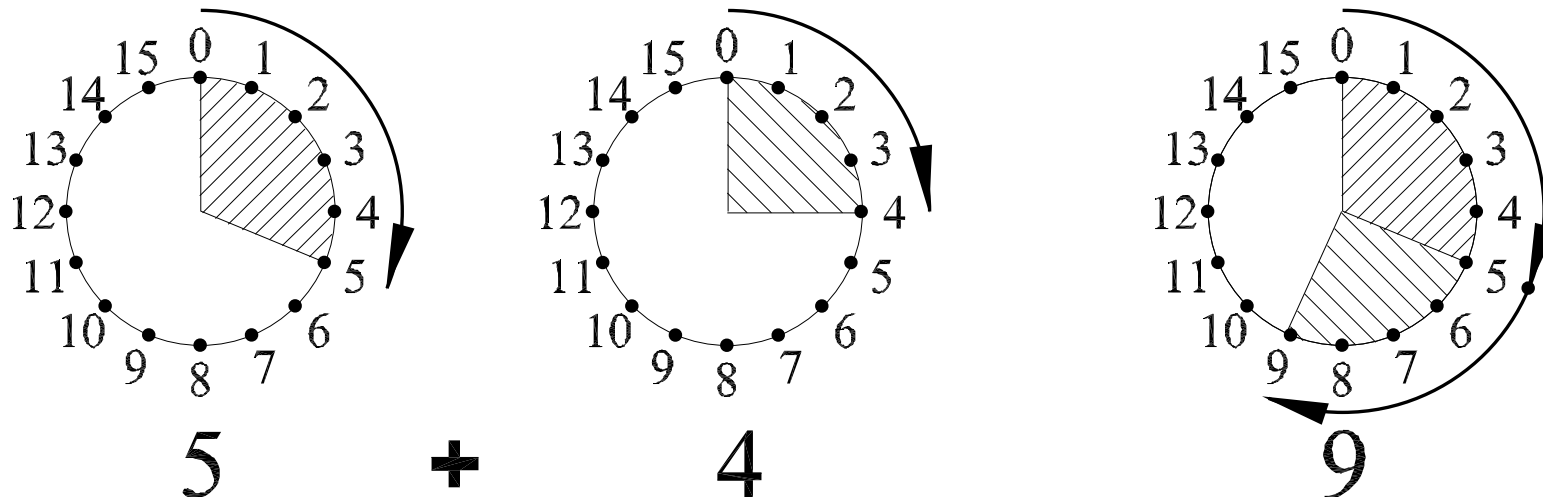
$$S = A + B - q_n \cdot Z$$

S differ from $A + B$ by multiple of Z so that
 $S \equiv A + B \pmod{Z}$

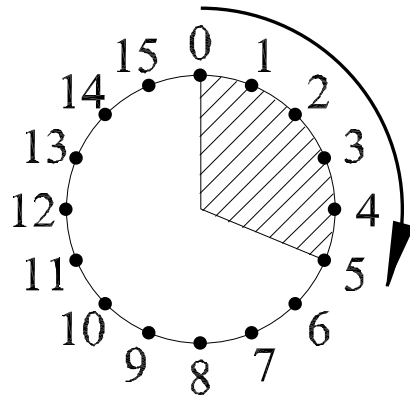
\mapsto

graphic view (analogy of clock face):

$$0101 + 0100 = {}_01001 \rightarrow 1001$$

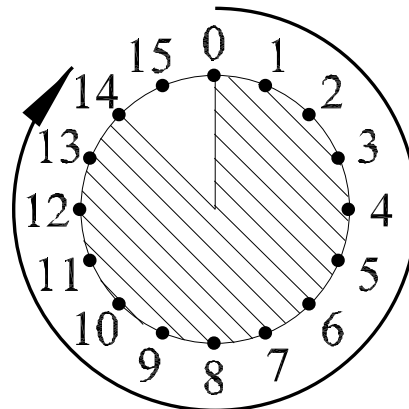


$$0101 + 1110 = {}_10011 \rightarrow 0011$$

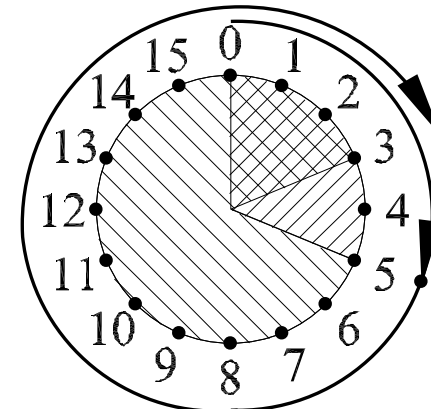


5

+



14



3

passing thru \Leftrightarrow carry from the highest order $q_n = 1$

in this case (addition using unsigned numbers):

$$q_n = 1 \Rightarrow A + B \geq \mathcal{Z} \quad (\mathcal{Z} = 1\,0000_2 = 16_{10})$$

$q_n = 1 \Rightarrow$ overflow \sim the result is out of format

however generally: **!!! carry \neq overflow !!!**

Subtractor

Full adder (! How to do it in wrong way. !)

a	b	v	u	r
0	0	0	0	0
0	0	1	1	1
0	1	0	1	1
0	1	1	1	0
1	0	0	0	1
1	0	1	0	0
1	1	0	0	0
1	1	1	1	1

$$\begin{aligned} r &= a \oplus b \oplus v = \\ &= \bar{a}\bar{b}v + \bar{a}b\bar{v} + a\bar{b}\bar{v} + abv \end{aligned}$$

$$u = \bar{a}b + \bar{a}v + bv$$

v_i borrow for order i

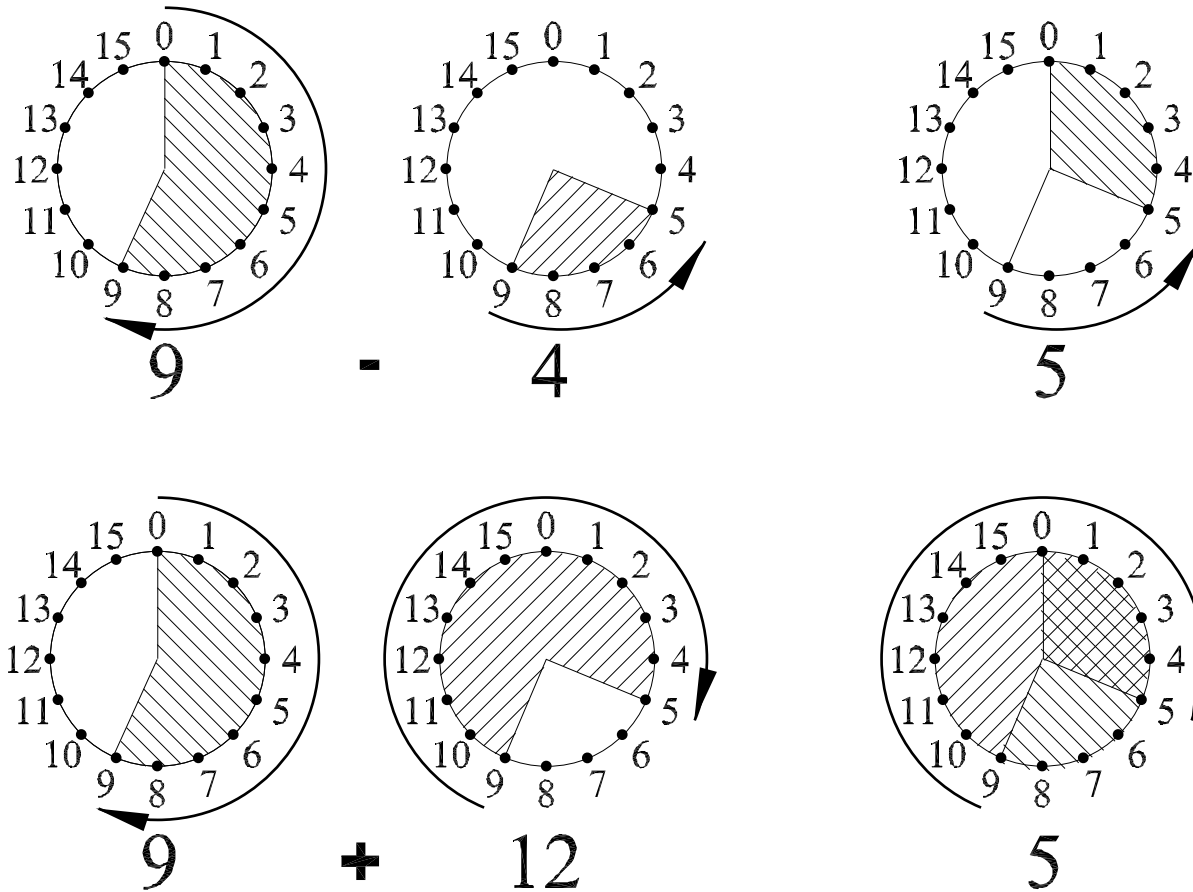
u_i borrow from order i

etc. - similarly as for addition

Is it possible modify an adder for subtraction?

Subtractor using adder

$$A - B \equiv A + (Z - B) \pmod{Z}$$



How to find $Z - B$?

$$X = \sum_{i=0}^n x_i z^i \quad x_i \in \langle 0, z - 1 \rangle$$

$$\begin{aligned} X_{max} &= \sum_{i=0}^n (z - 1) z^i = \sum_{j=1}^{n+1} z^j - \sum_{i=0}^n z^i = z^{n+1} - 1 = \\ &= Z - 1 \end{aligned}$$

$$z = 2: \quad X_{max} = \boxed{11\dots 11} = Z - 1$$

$$Z = \boxed{11\dots 11} + 1$$

$$Z - B = \boxed{11\dots 11} - B + 1$$

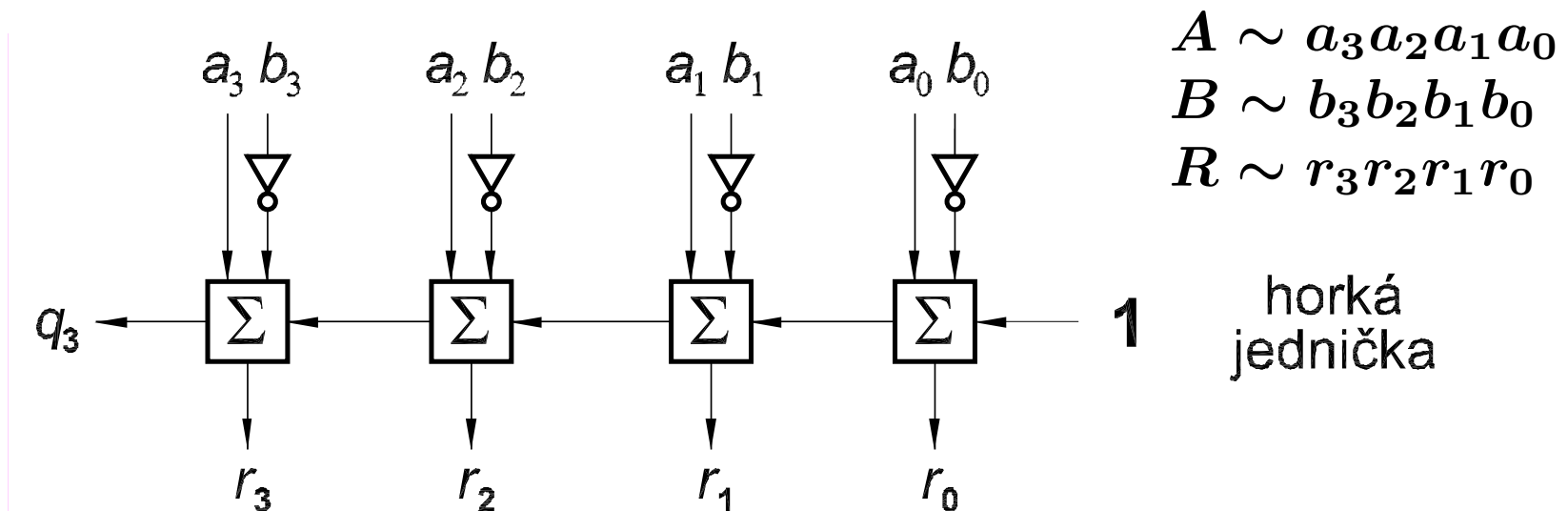
$$\boxed{Z - B = \overline{B} + 1} \quad \overline{B} \dots \text{negation of all bits}$$

$\dots + 1$ — so called hot one

note.: $B = 0 \Rightarrow Z - B = Z \equiv 0 \pmod{Z}$

$$\overline{B} + 1 = \boxed{11\dots 11} + 1 = 1\boxed{00\dots 00}$$

Subtractor using adder iii



$$R = A + (\mathcal{Z} - B) - q_n \cdot \mathcal{Z} = A - B + (1 - q_n) \cdot \mathcal{Z}$$

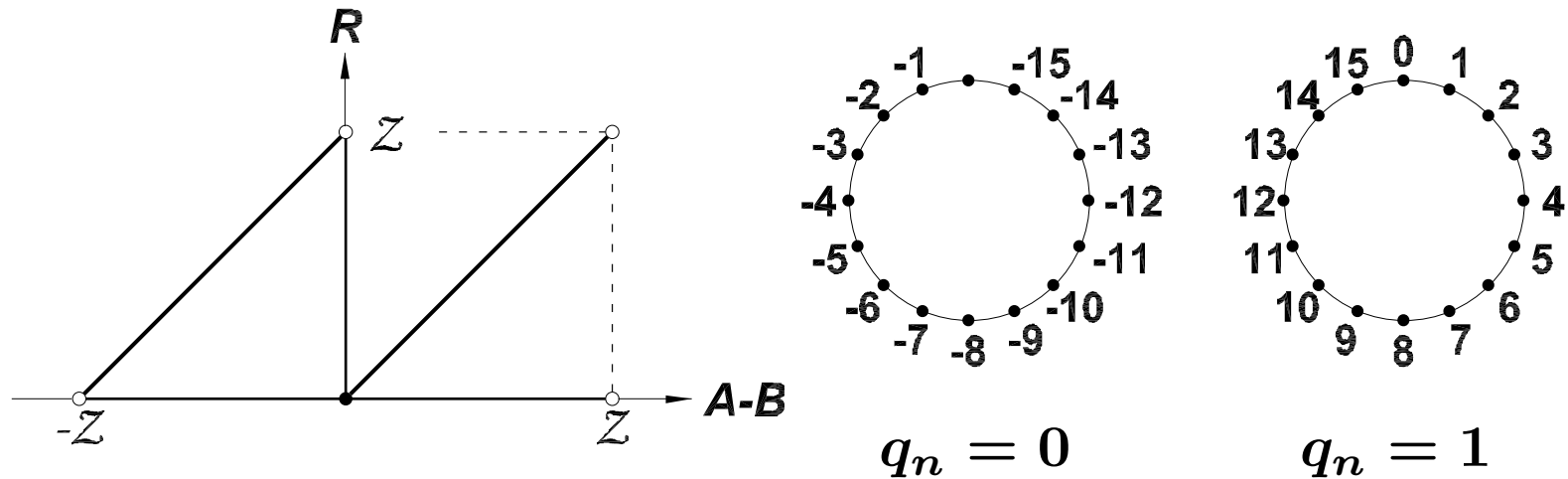
$$R = A - B + \overline{q_n} \cdot \mathcal{Z} \quad 0 \leq R < \mathcal{Z}$$

$$q_n = 1 \Rightarrow R = A - B \geq 0$$

$$q_n = 0 \Rightarrow R = A - B + \mathcal{Z} < \mathcal{Z} \Rightarrow A - B < 0$$

$q_n = 1$	\Rightarrow	$A \geq B$	$R = A - B$
$q_n = 0$	\Rightarrow	$A < B$	$R = \mathcal{Z} - (B - A)$

complementary pseudocode



If it is $q_n = 0$, it means $B > A$, then

$$R = Z - (B - A) \Rightarrow B - A = Z - R$$

$$\Rightarrow \boxed{B - A = \overline{R} + 1}$$

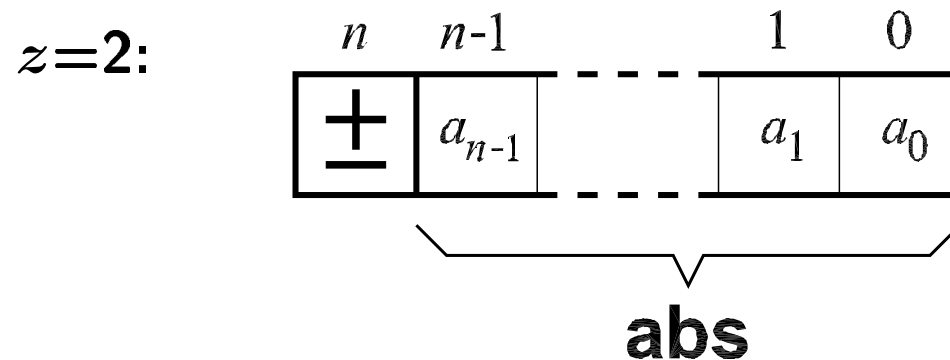
Signed number representation

*5 ways to represent
negative (as well as non-negative) number X :*

1. sign and magnitude $\dots \mathcal{P}(X)$
2. radix complement $\dots \mathcal{D}(X)$
2's complement - $z = 2$
10's complement - $z = 10$
3. diminished radix complement $\dots \mathcal{I}(X)$
1's complement - $z = 2$
9's complement - $z = 10$
4. biased code (or excess K) $\dots \mathcal{A}(X)$
 - a. type 0 $\dots \mathcal{A}_0(X)$
 - b. type 1 $\dots \mathcal{A}_1(X)$

Sign and magnitude

sign & abs value (magnitude):



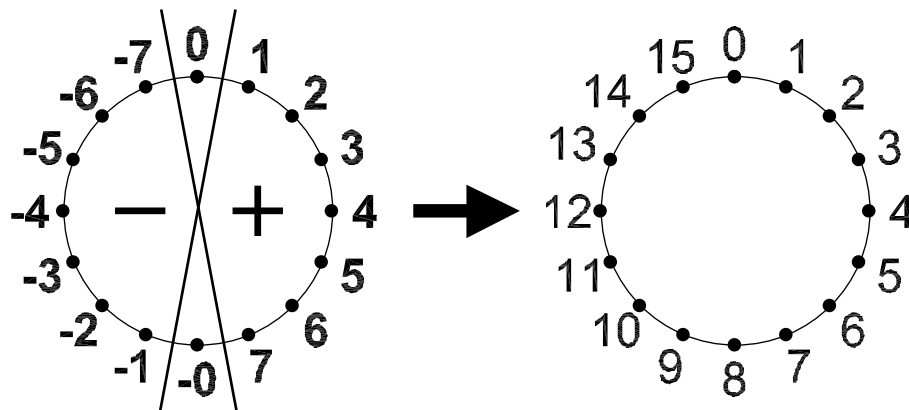
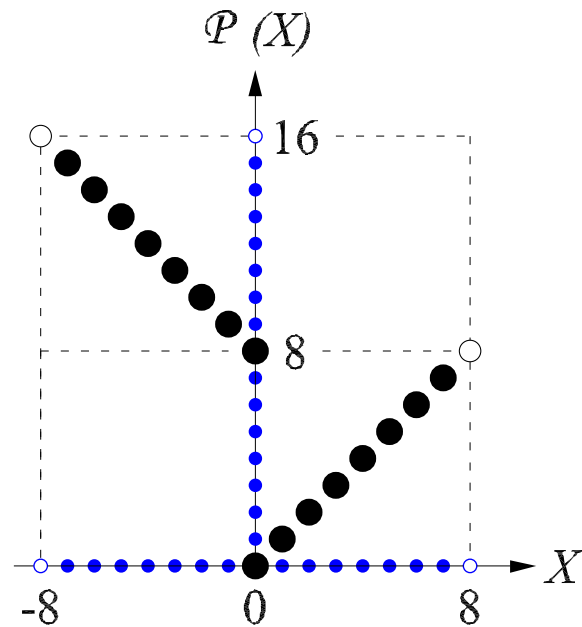
$$\text{sign bit} = \begin{cases} 0 & \text{for } X \geq 0 \\ 1 & \text{for } X \leq 0 \end{cases}$$

$$\mathcal{P}(X) = \begin{cases} X & \text{for } X \geq 0 \\ 2^n + |X| & \text{for } X \leq 0 \end{cases}$$

$$-\frac{1}{2}\mathcal{Z} < X < \frac{1}{2}\mathcal{Z} \quad \text{👉 symmetric range}$$

2 zero representation: $\begin{cases} \text{so called „positive zero“} & 0\ 0\dots 0 \\ \text{so called „negative zero“} & 1\ 0\dots 0 \end{cases}$

Sign and magnitude ii



X	$\mathcal{P}(X)$
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
-0	1000
-1	1001
-2	1010
-3	1011
-4	1100
-5	1101
-6	1110
-7	1111

addition :

$$(zA, aA) + (zB, aB) \rightarrow (zC, aC) ,$$

where zA , zB and zC are the sign bits and
 aA , aB and aC are the magnitudes

if ($zA=zB$)	$\{ aA + aB \rightarrow aC;$ $zA \rightarrow zC;$ if ($q = 1$) overflow; }
else	$\{ aA + \overline{aB} + 1 \rightarrow aC;$ $zA \rightarrow zC;$ if ($q = 0$) $\{ \overline{aC} + 1 \rightarrow aC;$ $\overline{zC} \rightarrow zC; \}$ }

q carry-out from higher order

sign change :

$$\mathcal{P}(-X) \rightarrow \mathcal{P}(X) \iff \overline{\text{MSB}} \rightarrow \text{MSB}$$

MSB - bit in higher order (first from left)
- **Most Significant Bit**

subtraction :

$$A - B = A + (-B)$$

to swap \overline{zB} by zB
else alike the addition

absolute value :

$$\mathcal{P}(X) \rightarrow \mathcal{P}(|X|) \iff 0 \rightarrow \text{MSB}$$

Radix complement

radix complement $\begin{cases} 2\text{'s complement for } z = 2 \\ 10\text{'s complement for } z = 10 \end{cases}$

$$\mathcal{D}(X) = \begin{cases} X & \text{pro } X \geq 0 \\ z + X = z - |X| & \text{pro } X < 0 \end{cases}$$

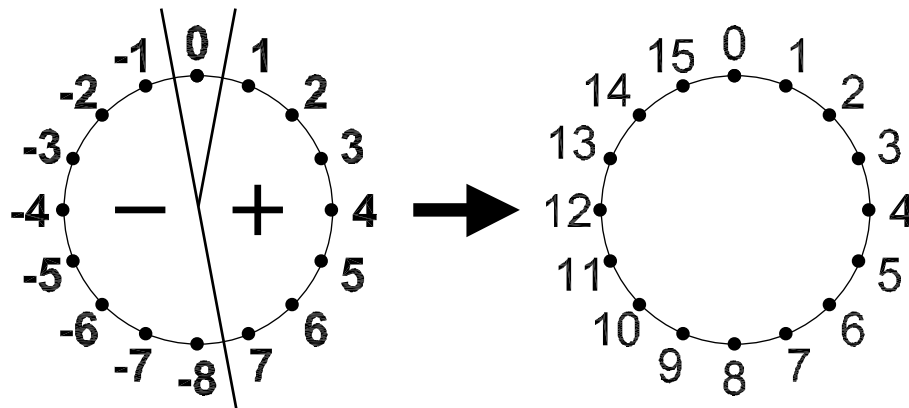
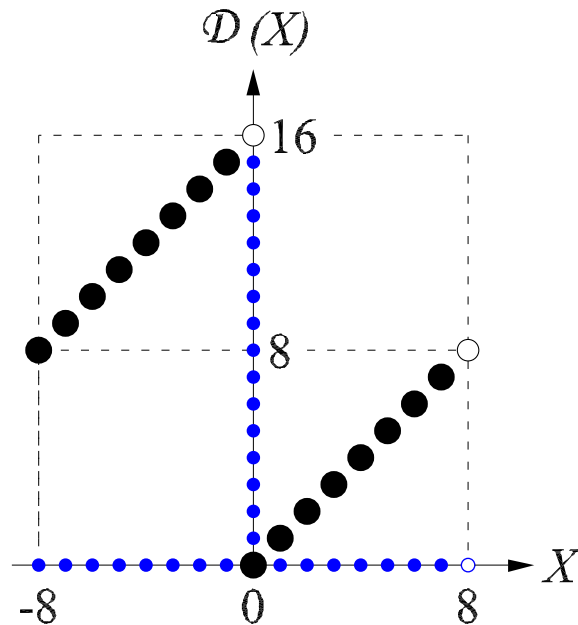
$$-\frac{1}{2}z \leq X < \frac{1}{2}z \quad \text{☞ asymmetric range}$$

$$\begin{array}{ll} \text{MSB} = 0 & \iff X \geq 0 \\ \text{MSB} = 1 & \iff X < 0 \end{array}$$

$$\mathcal{D}(X) \equiv X \pmod{z}$$

Radix complement ii

$z=2$:



X	$\mathcal{D}(X)$
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
-8	1000
-7	1001
-6	1010
-5	1011
-4	1100
-3	1101
-2	1110
-1	1111

addition: $\mathcal{D}(A+B) \equiv A+B \equiv \mathcal{D}(A) + \mathcal{D}(B) \pmod{\mathcal{Z}}$



add up $\mathcal{D}(A) + \mathcal{D}(B)$ and
ignore carry q_n from higher order

subtraction: $\mathcal{D}(A-B) \equiv A-B \equiv \mathcal{D}(A) - \mathcal{D}(B) \pmod{\mathcal{Z}}$



subtract $\mathcal{D}(B)$ from $\mathcal{D}(A)$ and
ignore borrow v_n from higher order

or



convert subtraction on addition, that is
add up $\mathcal{D}(A) + \overline{\mathcal{D}(B)} + 1$ and
ignore carry q_n from higher order

Carry (eventually borrow) is ignored.

?

How to detect overflow?

(that is $A+B \geq \frac{1}{2}\mathcal{Z}$ or $A+B < -\frac{1}{2}\mathcal{Z}$)

?

overflow during addition in 2's complement code:

$$1. A < 0 \text{ a } B \geq 0 \Rightarrow A \leq A + B < B$$

$$B < 0 \text{ a } A \geq 0 \Rightarrow B \leq A + B < A$$

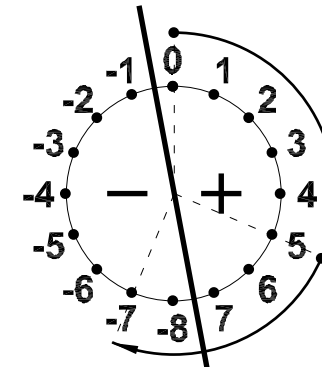
! in this case overflow can not occur !

$$2. A \geq 0 \text{ a } B \geq 0$$

$$\text{overflow: } A + B \geq \frac{1}{2}Z$$

result has opposite sign

see example $5 + 4 \rightarrow -7$

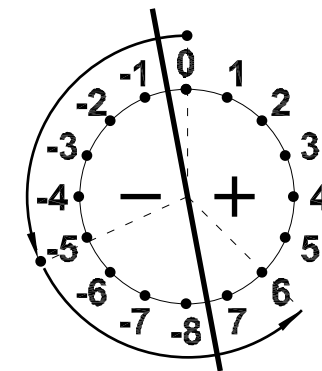


$$3. A < 0 \text{ a } B < 0$$

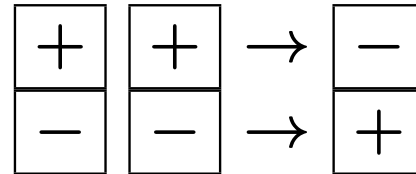
$$\text{overflow: } A + B < -\frac{1}{2}Z$$

result has opposite sign

see example $(-5) + (-5) \rightarrow 6$



overflow during addition: same sign of both operands and opposite sign of result



overflow during addition:

subtraction converted on addition

⇒ There is nothing to solve.

will be referred to: $\mathcal{D}(A) + \mathcal{D}(B) - q_n \cdot \mathcal{Z} = S$

$$\mathcal{D}(A) \sim a_n^{\mathcal{D}} a_{n-1}^{\mathcal{D}} \dots a_1^{\mathcal{D}} a_0^{\mathcal{D}}$$

$$\mathcal{D}(B) \sim b_n^{\mathcal{D}} b_{n-1}^{\mathcal{D}} \dots b_1^{\mathcal{D}} b_0^{\mathcal{D}}$$

$$S \sim s_n^{\mathcal{D}} s_{n-1}^{\mathcal{D}} \dots s_1^{\mathcal{D}} s_0^{\mathcal{D}}$$

over — overflow

detection of overflow:

$$\boxed{1} \quad a_n^{\mathcal{D}} = b_n^{\mathcal{D}} = 0 \quad \text{a} \quad s_n^{\mathcal{D}} = 1 \quad \text{or} \\ a_n^{\mathcal{D}} = b_n^{\mathcal{D}} = 1 \quad \text{a} \quad s_n^{\mathcal{D}} = 0$$

$$over = \overline{a_n^{\mathcal{D}}} \cdot \overline{b_n^{\mathcal{D}}} \cdot s_n^{\mathcal{D}} + a_n^{\mathcal{D}} \cdot b_n^{\mathcal{D}} \cdot \overline{s_n^{\mathcal{D}}}$$

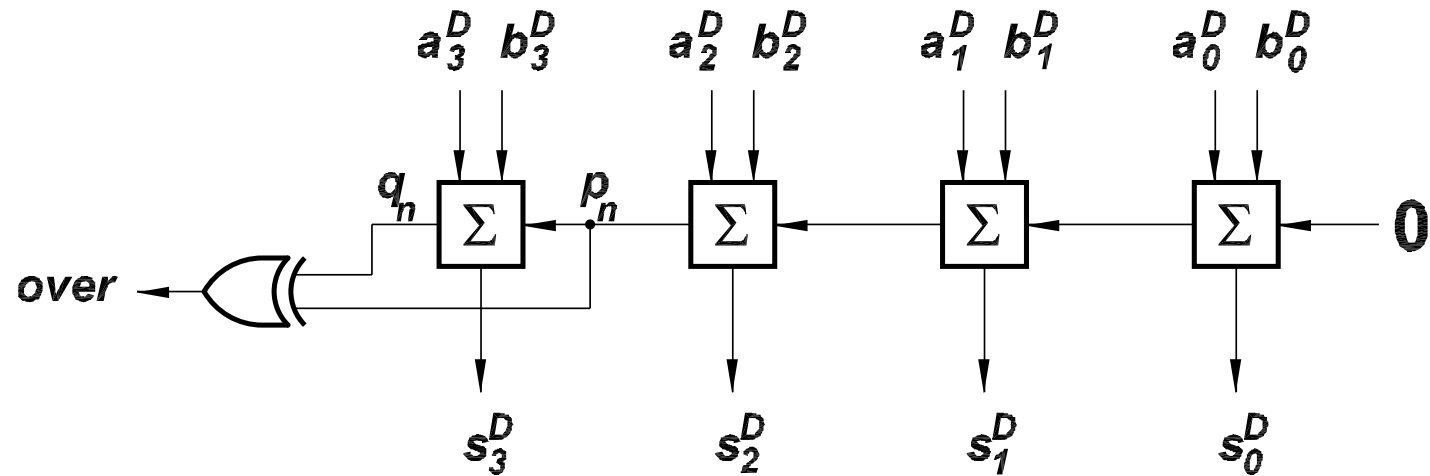
$a_n^{\mathcal{D}}$	$b_n^{\mathcal{D}}$	$p_n^{\mathcal{D}}$	$q_n^{\mathcal{D}}$	$s_n^{\mathcal{D}}$	
0	0	0	0	0	$q_n = p_n$
0	0	1	0	1	$q_n \neq p_n$
0	1	0	0	1	$q_n = p_n$
0	1	1	1	0	$q_n = p_n$
1	0	0	0	1	$q_n = p_n$
1	0	1	1	0	$q_n = p_n$
1	1	0	1	0	$q_n \neq p_n$
1	1	1	1	1	$q_n = p_n$

$$over = q_n \oplus p_n$$

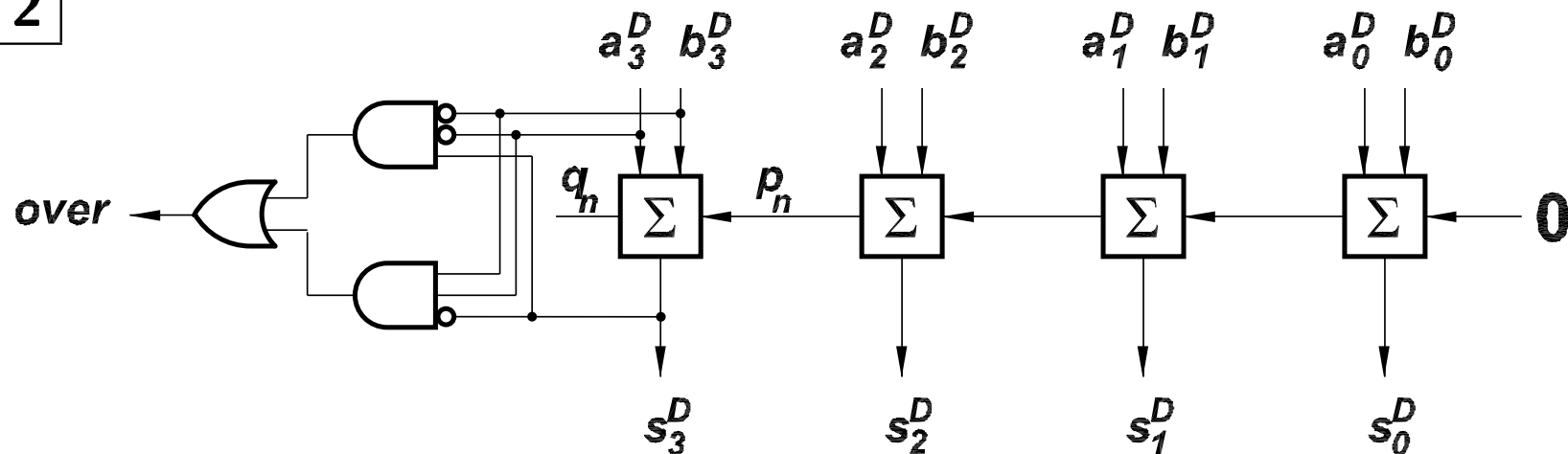
$z=2$:

addition : $\mathcal{D}(A + B) \equiv \mathcal{D}(A) + \mathcal{D}(B) \pmod{\mathcal{Z}}$

1



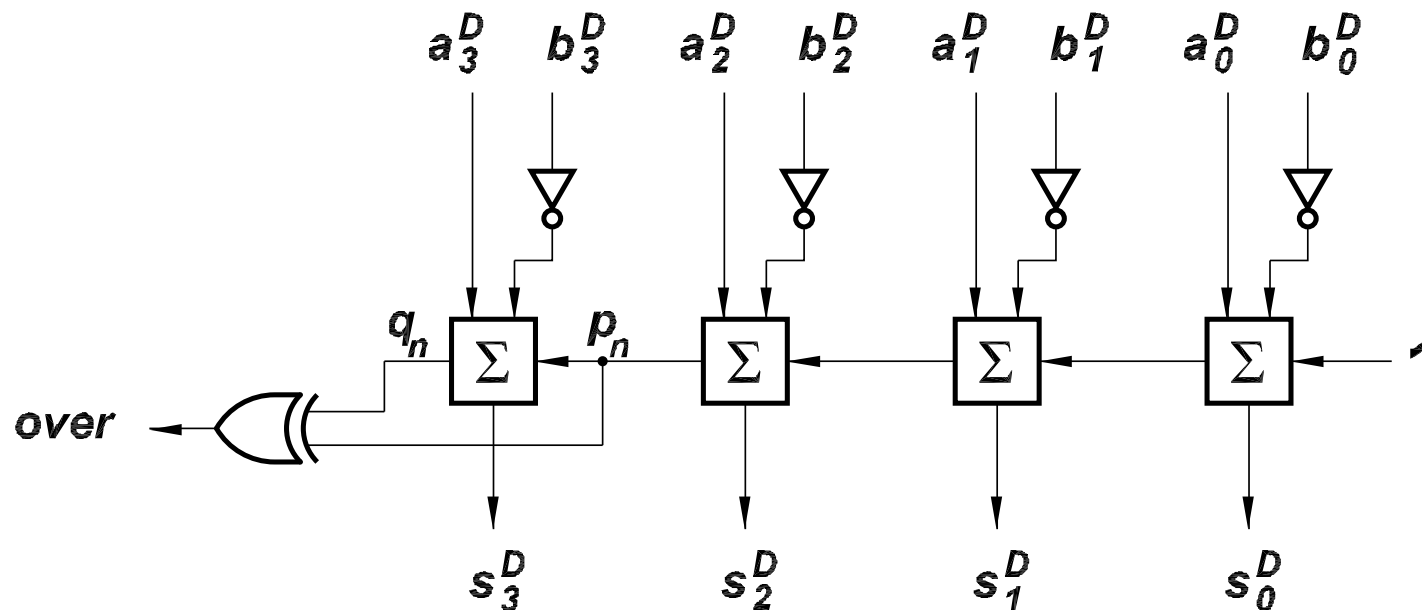
2



$z=2$:

sign change :

$$\mathcal{D}(-X) \equiv \overline{\mathcal{D}(X)} + 1 \pmod{\mathcal{Z}}$$

subtraction : $A - B = A + (-B)$ 

Diminished radix complement

radix complement $\begin{cases} 1\text{'s complement for } z = 2 \\ 9\text{'s complement for } z = 10 \end{cases}$

$$\mathcal{I}(X) = \begin{cases} X & \text{for } X \geq 0 \\ \overline{|X|} & \text{for } X \leq 0 \end{cases}$$

$$-\frac{1}{2}Z < X < \frac{1}{2}Z \quad \text{👉 symmetric range}$$

two zero representation: $\begin{cases} \text{so called „positive zero“} & 0\ 0\dots 0 \\ \text{so called „negative zero“} & 1\ 1\dots 1 \end{cases}$

$$\text{MSB} = 0 \implies X \geq 0$$

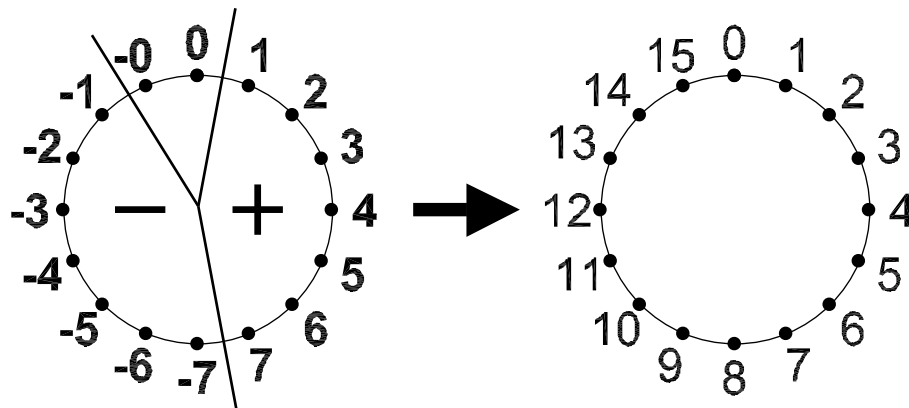
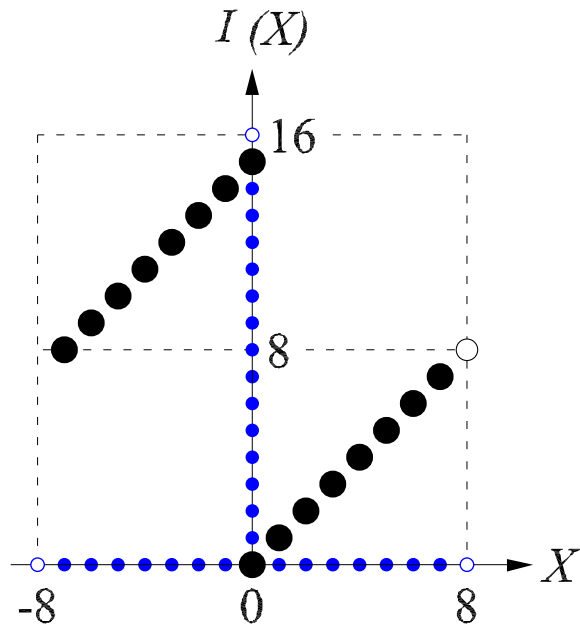
$$\text{MSB} = 1 \implies X \leq 0$$

$$T + \overline{T} = 11\dots 11 = Z-1$$

$$\mathcal{I}(X) \equiv X \pmod{Z-1}$$

Diminished radix complement ii

$z=2$:

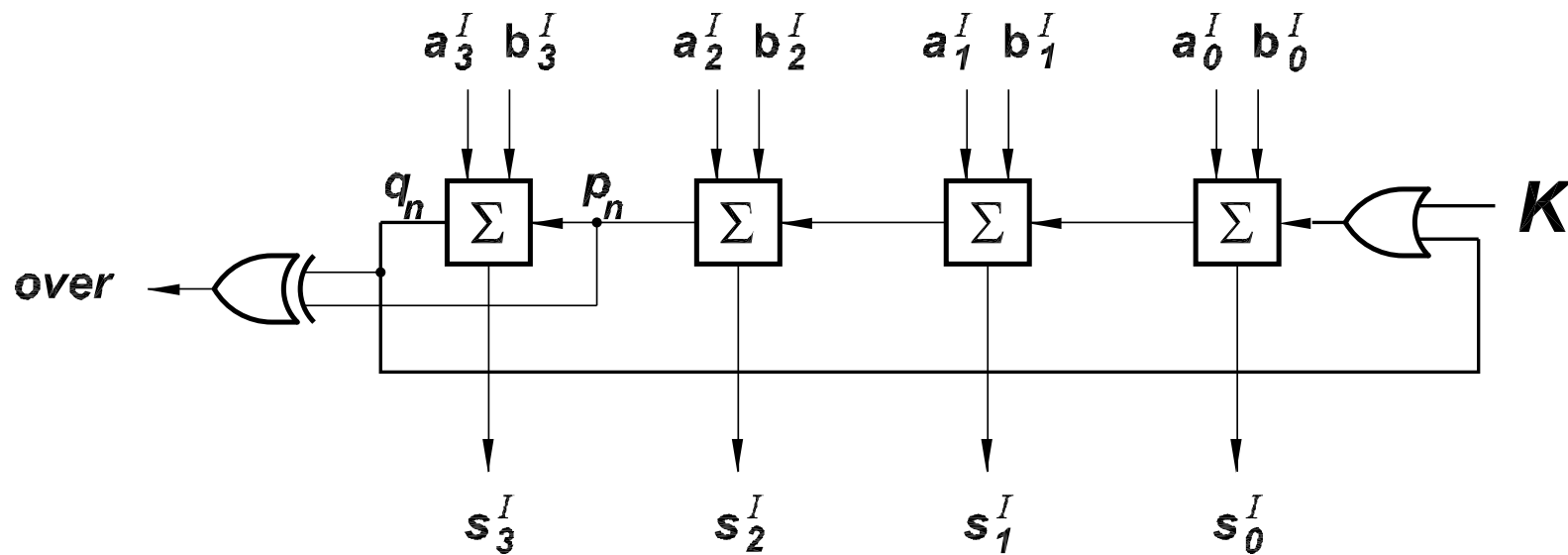


X	$\mathcal{I}(X)$
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
-7	1000
-6	1001
-5	1010
-4	1011
-3	1100
-2	1101
-1	1110
- 0	1111

addition : $\mathcal{I}(A + B) \equiv \mathcal{I}(A) + \mathcal{I}(B) \pmod{\mathcal{Z}-1}$

$\mathcal{I}(A) + \mathcal{I}(B) \geq \mathcal{Z} \Rightarrow q=1$

necessary to subtract \mathcal{Z} and add 1 \Rightarrow **circular carry**



feedback  sequential circuit

solution: correction K on input carry into lower order

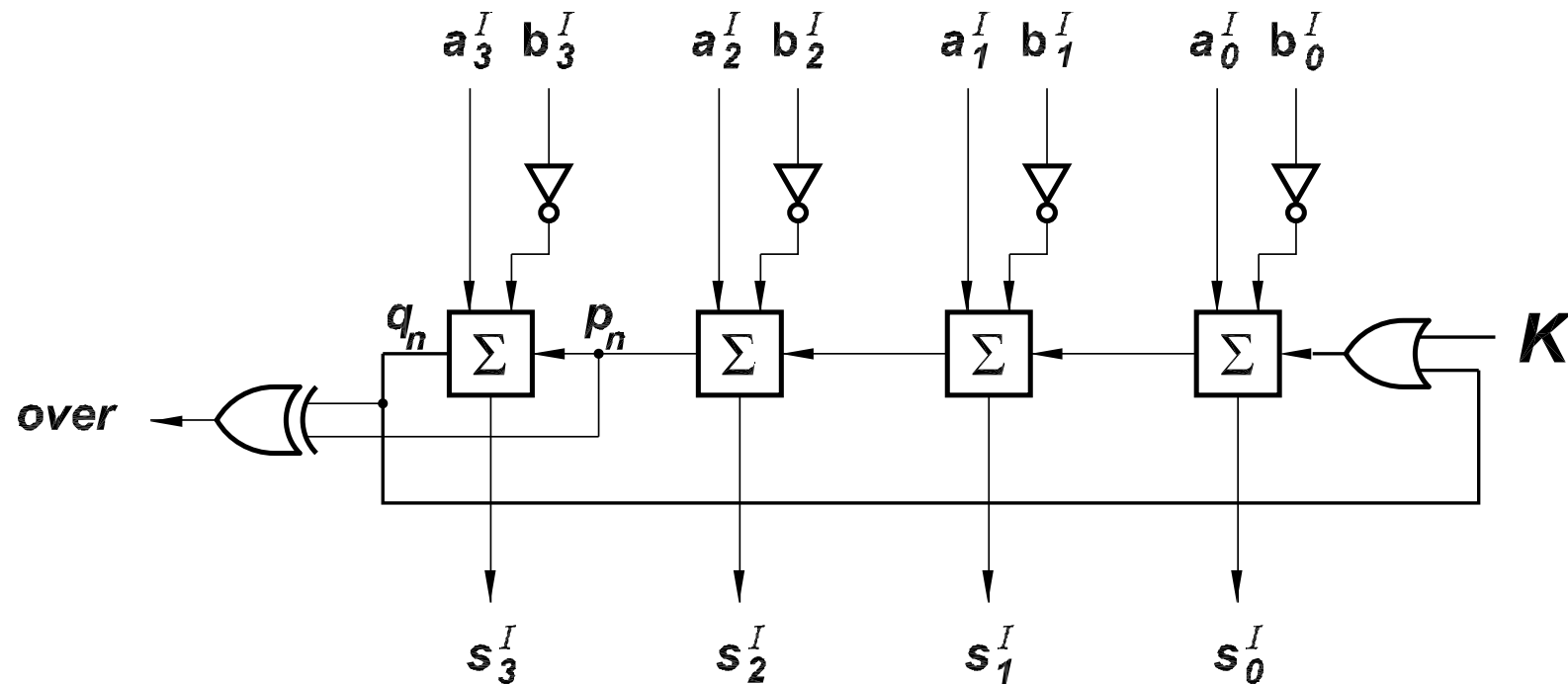
$K = 1$, in a case, when carry go over all orders

$z=2$:

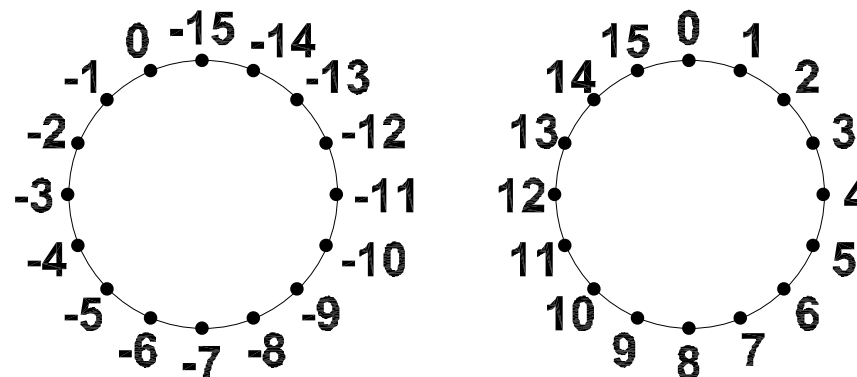
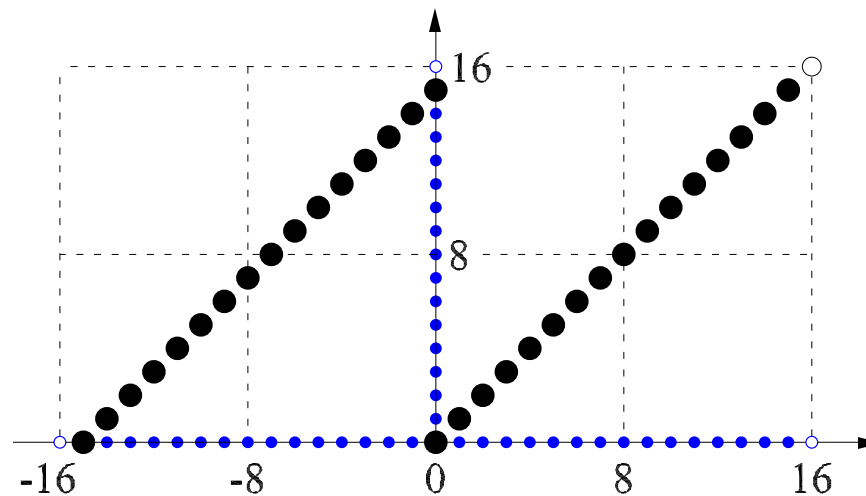
sign change :

$$\mathcal{I}(-X) \equiv \overline{\mathcal{I}(X)}$$

subtraction : $A - B = A + (-B)$



unsigned numbers + adder for radix complement code \Rightarrow
 \Rightarrow **radix complement pseudo-code** (comparison
with 2's complement pseudo-code)



Biased code

$$\mathcal{A}(X) = X + K ,$$

where K is „suitable“ constant

„suitable“ constants:

$$\begin{array}{llll} \frac{1}{2}Z & \Rightarrow & \mathcal{A}_0(X) & \dots \text{ biased code type 0} \\ \frac{1}{2}Z - 1 & \Rightarrow & \mathcal{A}_1(X) & \dots \text{ biased code type 1} \end{array}$$

Code is monotonous — growing.

addition :

$$\mathcal{A}(A + B) = \mathcal{A}(A) + \mathcal{A}(B) - K$$

subtraction :

$$\mathcal{A}(A - B) = \mathcal{A}(A) - \mathcal{A}(B) + K$$

It is true (of course) *iff overflow doesn't occurs !*

Biased code - type 0

$$\mathcal{A}_0(X) = X + \frac{1}{2}\mathcal{Z}$$

$$-\frac{1}{2}\mathcal{Z} \leq X < \frac{1}{2}\mathcal{Z} \quad \text{☞ asymmetric range}$$

$$\text{MSB} = 1 \iff X \geq 0$$

$$\text{MSB} = 0 \iff X < 0$$

$$\mathcal{A}_0(X) \equiv \mathcal{D}(X) \pmod{\frac{1}{2}\mathcal{Z}}$$

! $\mathcal{A}_0(X)$ and $\mathcal{D}(X)$ differ in the MSB only !

$$\text{MSB}(\mathcal{A}_0(X)) = \overline{\text{MSB}(\mathcal{D}(X))}$$

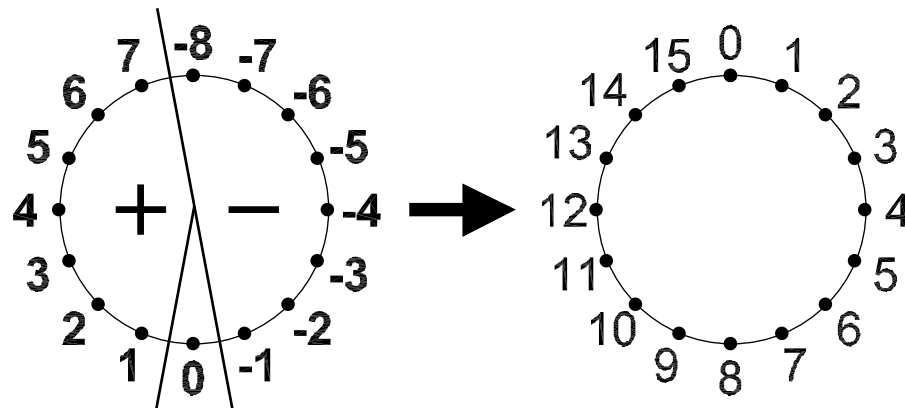
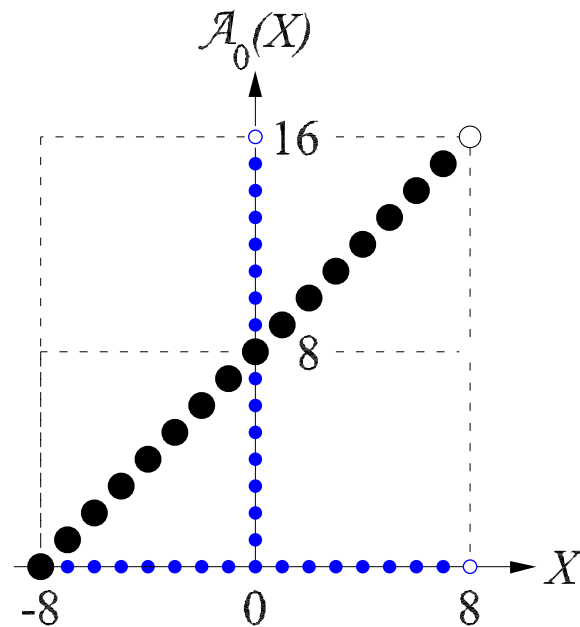
\implies **operations** (addition, subtraction, sign change etc.):

$$\mathcal{A}_0(\text{operands}) \rightarrow \mathcal{D}(\text{operands})$$

$$\mathcal{D}(\text{result}) \rightarrow \mathcal{A}_0(\text{result})$$

Biased code - type 0 ii

$z=2$:



X	$A_0(X)$
-8	0000
-7	0001
-6	0010
-5	0011
-4	0100
-3	0101
-2	0110
-1	0111
0	1000
1	1001
2	1010
3	1011
4	1100
5	1101
6	1110
7	1111

Biased code - type 1

$$\mathcal{A}_1(X) = X + \frac{1}{2}\mathcal{Z} - 1$$

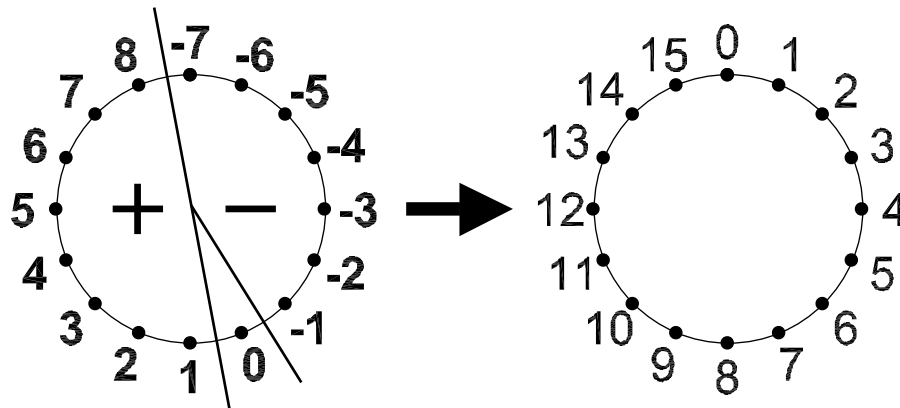
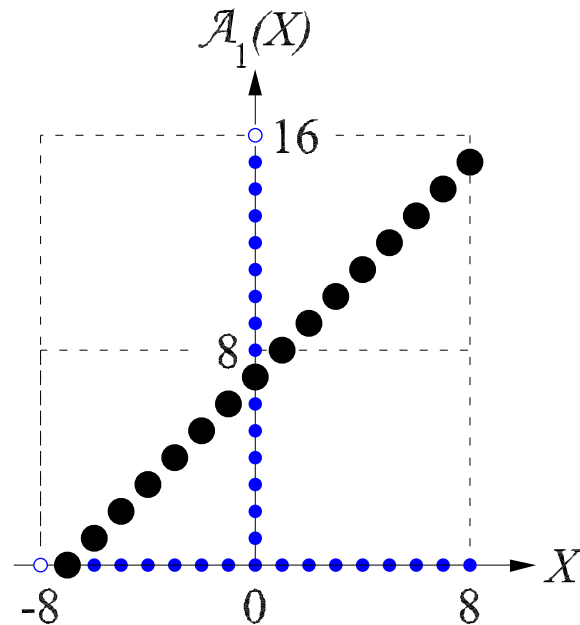
$$-\frac{1}{2}\mathcal{Z} < X \leq \frac{1}{2}\mathcal{Z} \quad \text{☞ asymmetric range}$$

$$\text{MSB} = 1 \iff X > 0$$

$$\text{MSB} = 0 \iff X \leq 0$$

Biased code - type 1 ii

$z=2$:



X	$\mathcal{A}(1)X$
-7	0000
-6	0001
-5	0010
-4	0011
-3	0100
-2	0101
-1	0110
0	0111
1	1000
2	1001
3	1010
4	1011
5	1100
6	1101
7	1110
8	1111

addition and subtraction :

$$\mathcal{A}(A + B) = \mathcal{A}(A) + \mathcal{A}(B) - K$$

$$\mathcal{A}(A - B) = \mathcal{A}(A) - \mathcal{A}(B) + K$$

$$\mathcal{A}(A + B) = \mathcal{A}(A) + \mathcal{A}(B) - \frac{1}{2}Z + 1$$

$$\mathcal{A}(A - B) = \mathcal{A}(A) - \mathcal{A}(B) + \frac{1}{2}Z - 1$$

$$-\frac{1}{2}Z \equiv +\frac{1}{2}Z \pmod{Z} \qquad \frac{1}{2}Z \sim \boxed{100\dots 00}$$

$$-\frac{1}{2}Z \equiv +\frac{1}{2}Z \pmod{Z} \rightsquigarrow \text{negation of bit in higher order}$$

overflow:

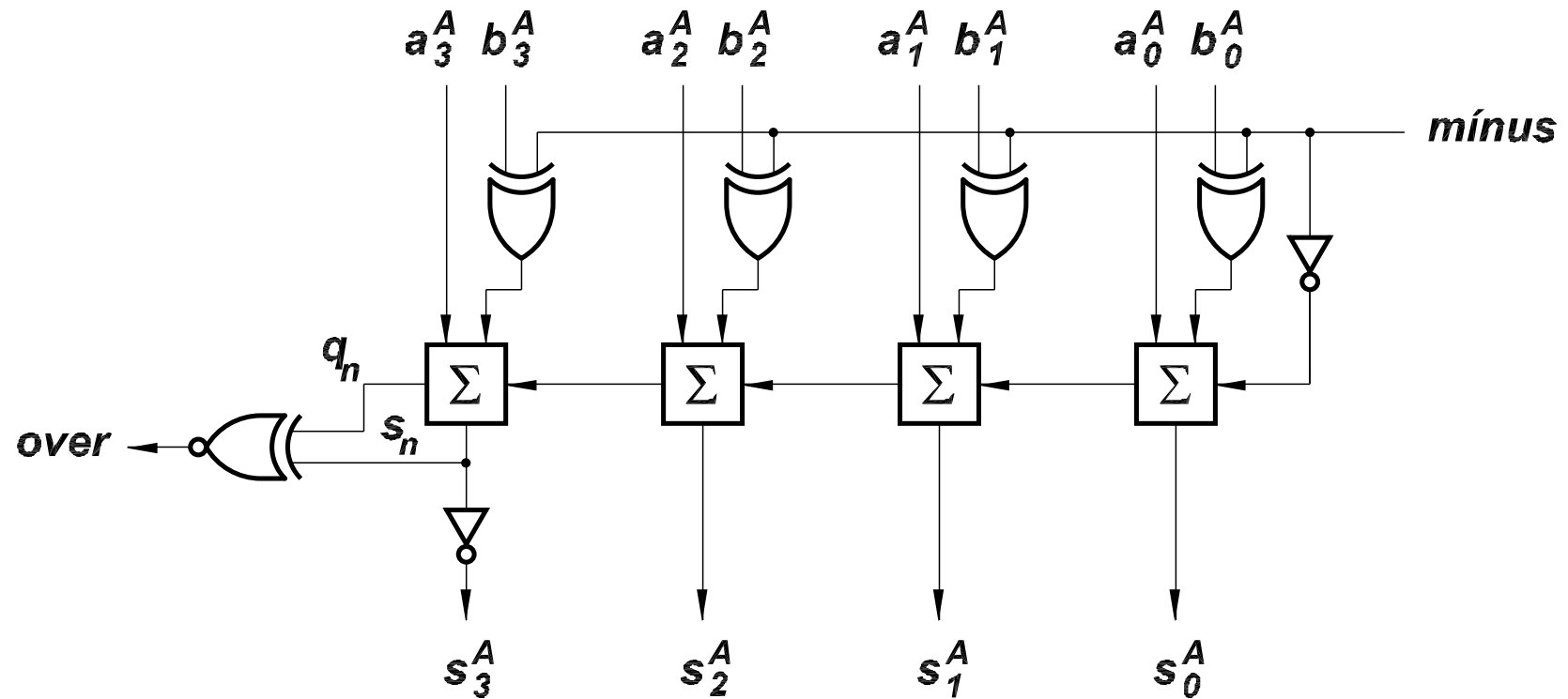
addition: same sign of addend
 and different sign of result

subtraction: sign of minuend and subtrahend are different
 signs of minuend and result are different

(derivation is analogous as in 2's complement code)

$z=2$:

adder / subtractor



sign change: $-X = 0 - X$

overflow detection

a_n	b_n	p_n	q_n	s_n
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

$$q_n = s_n$$

$$q_n \neq s_n$$

$$q_n \neq s_n$$

$$q_n \neq s_n$$

$$q_n \neq s_n$$

$$q_n \neq s_n$$

$$q_n \neq s_n$$

$$q_n = s_n$$

$$over = \overline{q_n \oplus s_n}$$