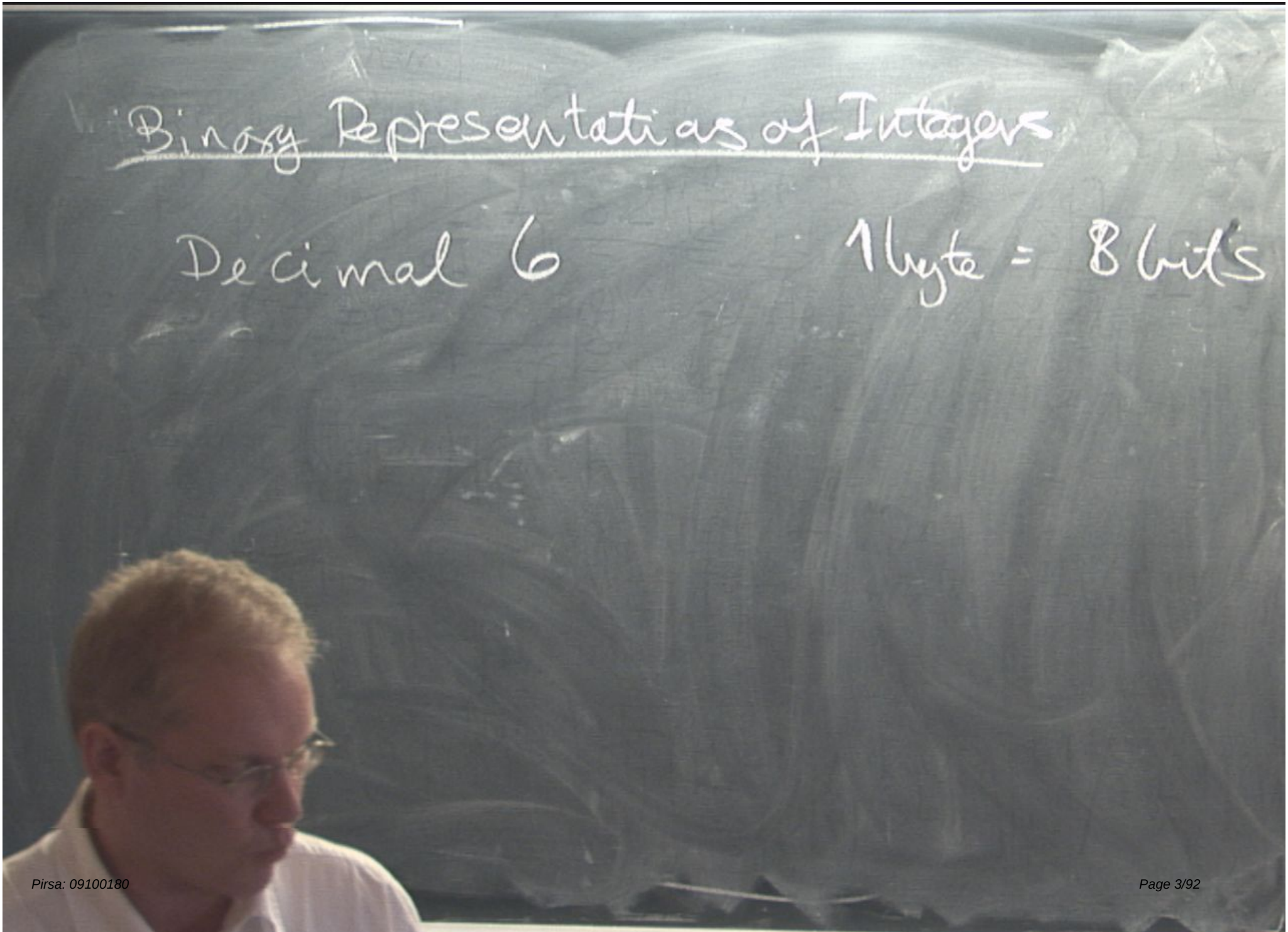
Title: Scientific Computation (PHYS 608) - Lecture 4

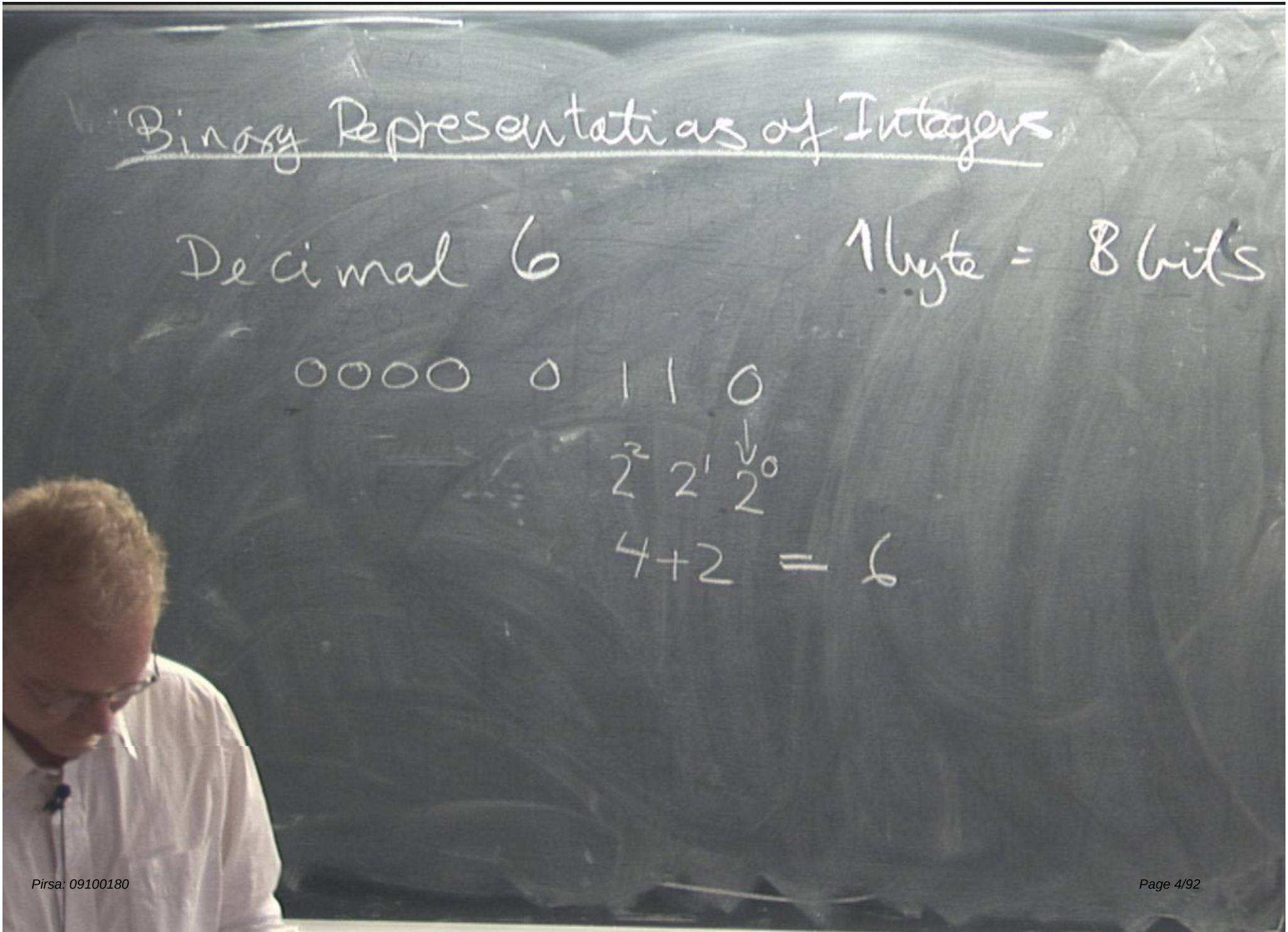
Date: Oct 29, 2009 10:30 AM

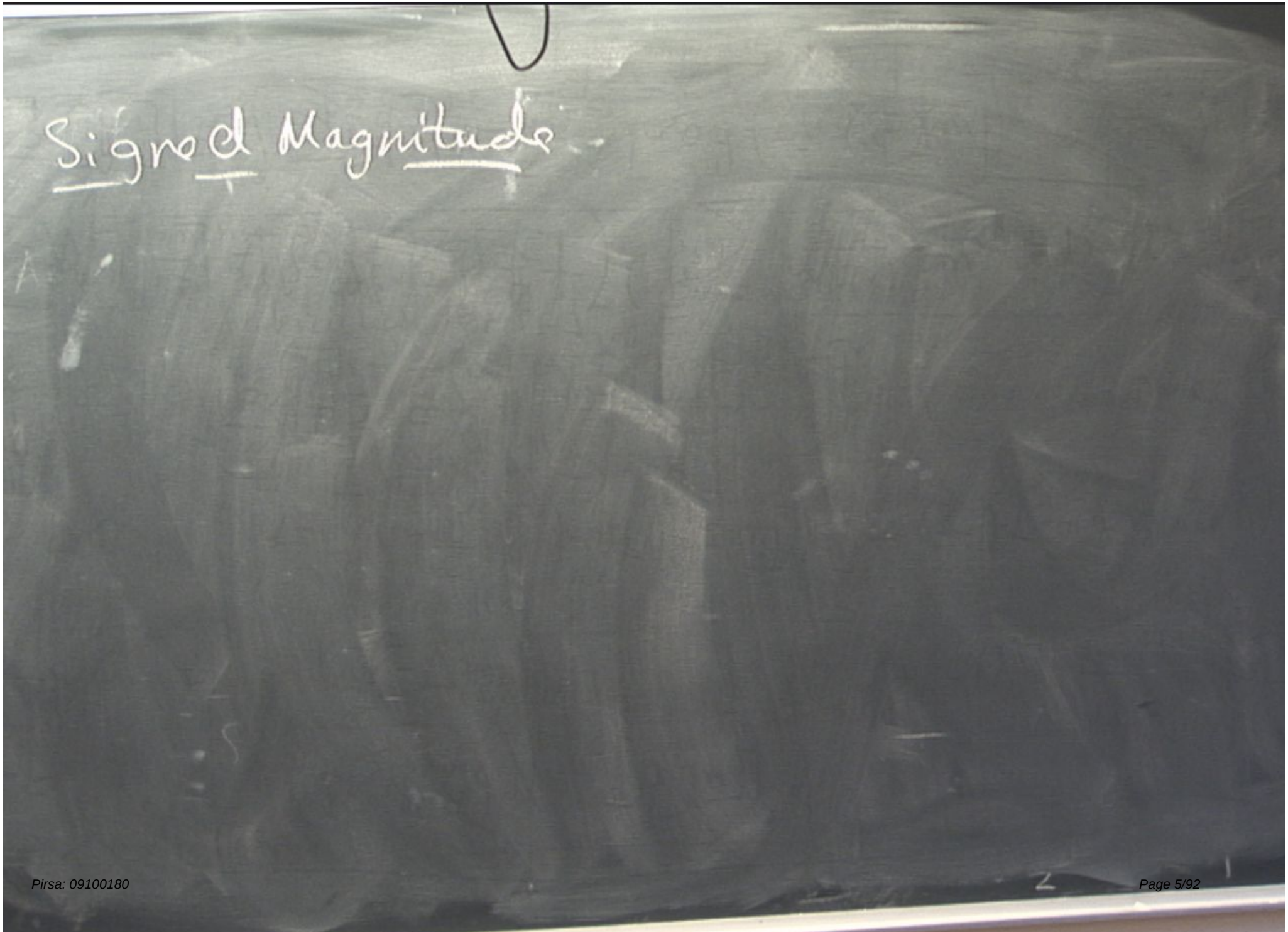
URL: http://pirsa.org/09100180

Abstract:

Binary Representations of Integers


Binary Representations of Integers
Decimal 6
1 byte = 8 bits


Binary Representation of Integers
Decimal 6
1 byte = 8 bits
0000 0 1 1 0
2^2 2^1 2^0
4+2 = 6


Signed Magnitude

Signed Magnitude

1 0000 0110 = −6

sign bit

1000 0000 = −0

0000 0000 = +0

Signed Magnitude

1 0000 0110 = −6

sign bit

1000 0000 = −0

0000 0000 = +0

One's Complement

0000 0110

Signed Magnitude

1 0000 0110 = −6

sign bit

1 0000 = −0

0 0000 = +0

One's Complement

0000 0110

1111 1001

Signed Magnitude

1 0000 0110 = -6

sign bit

1 0000 0000 = -0

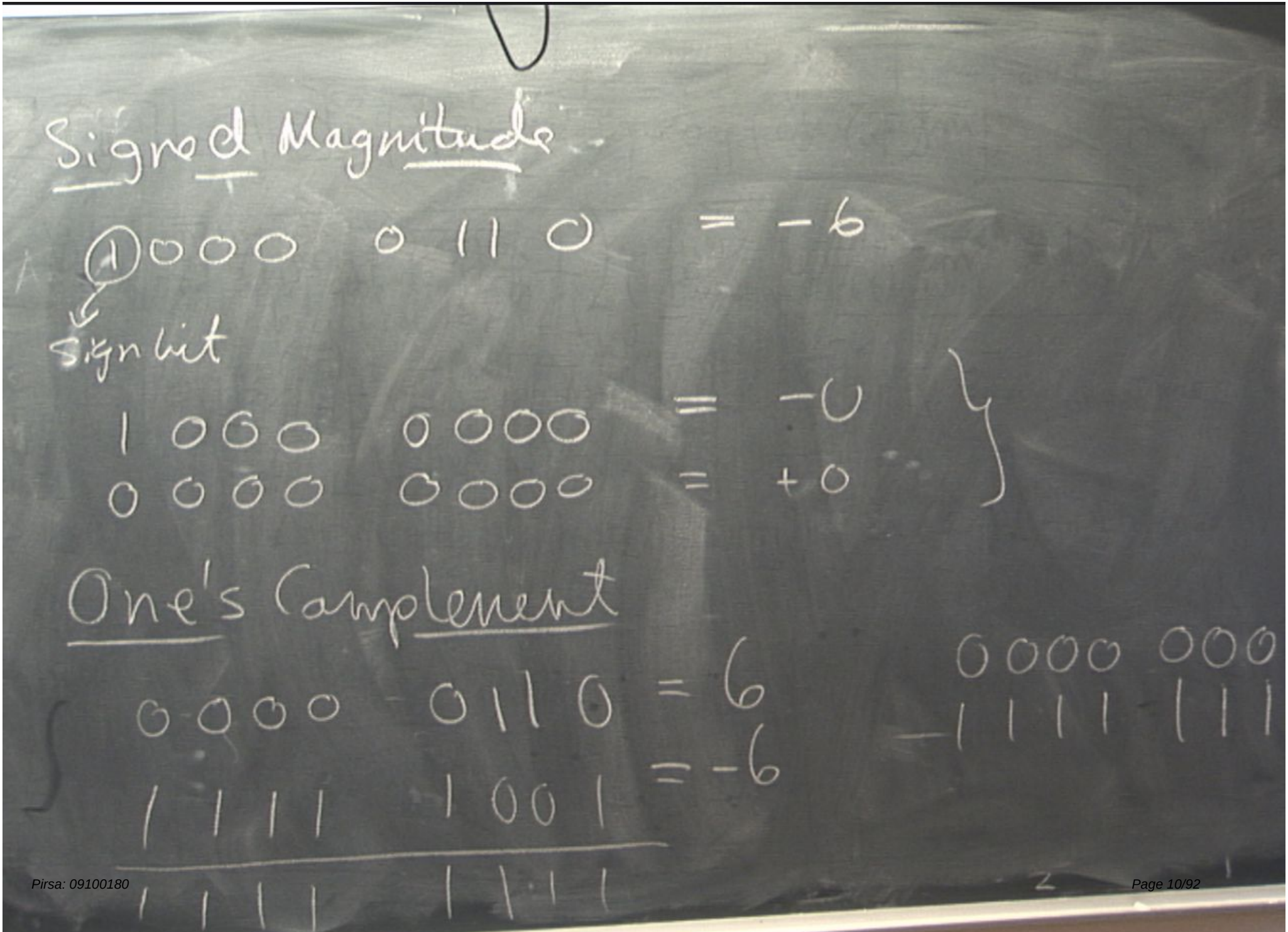
0 0000 0000 = +0

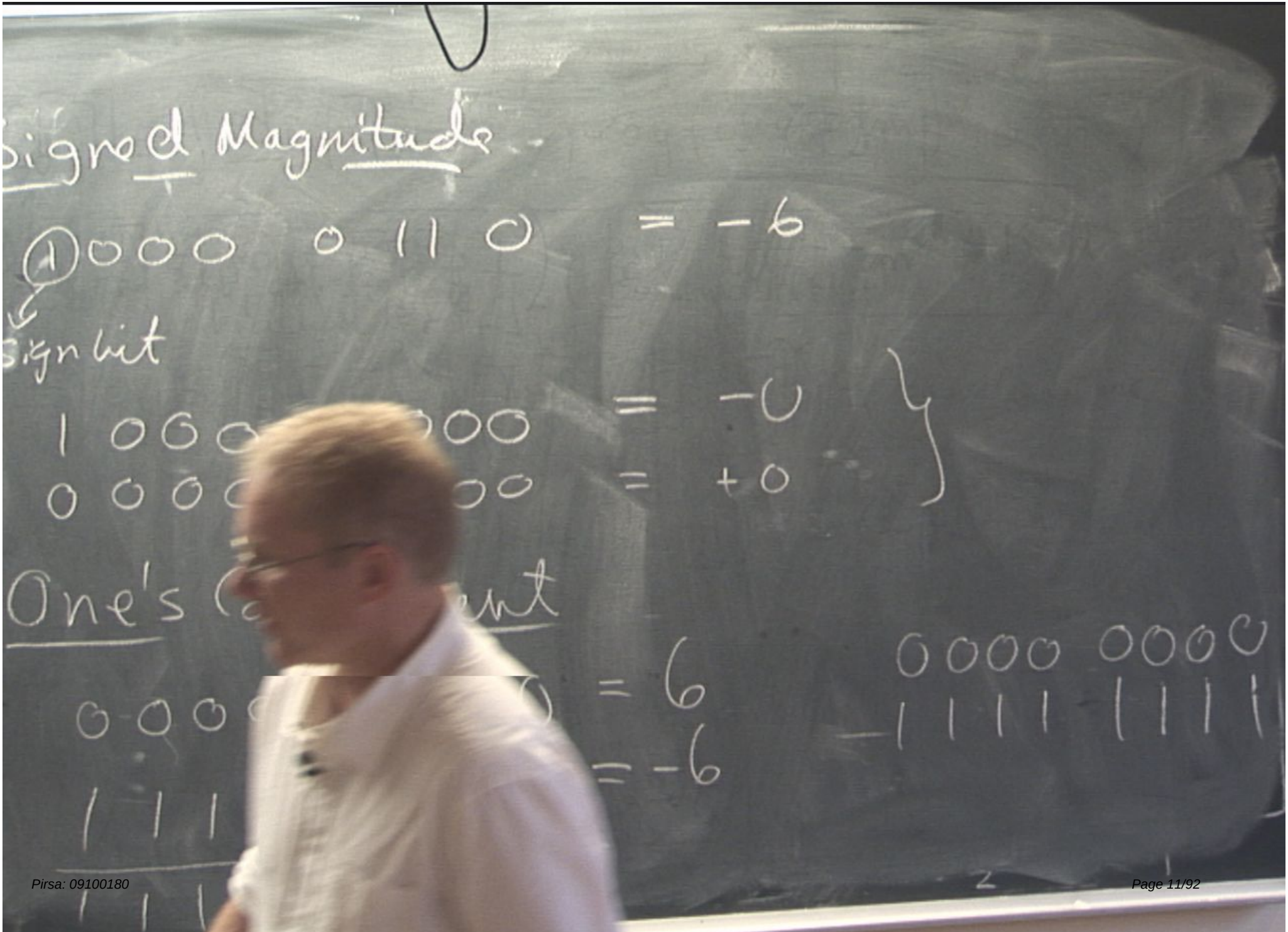
One's Complement

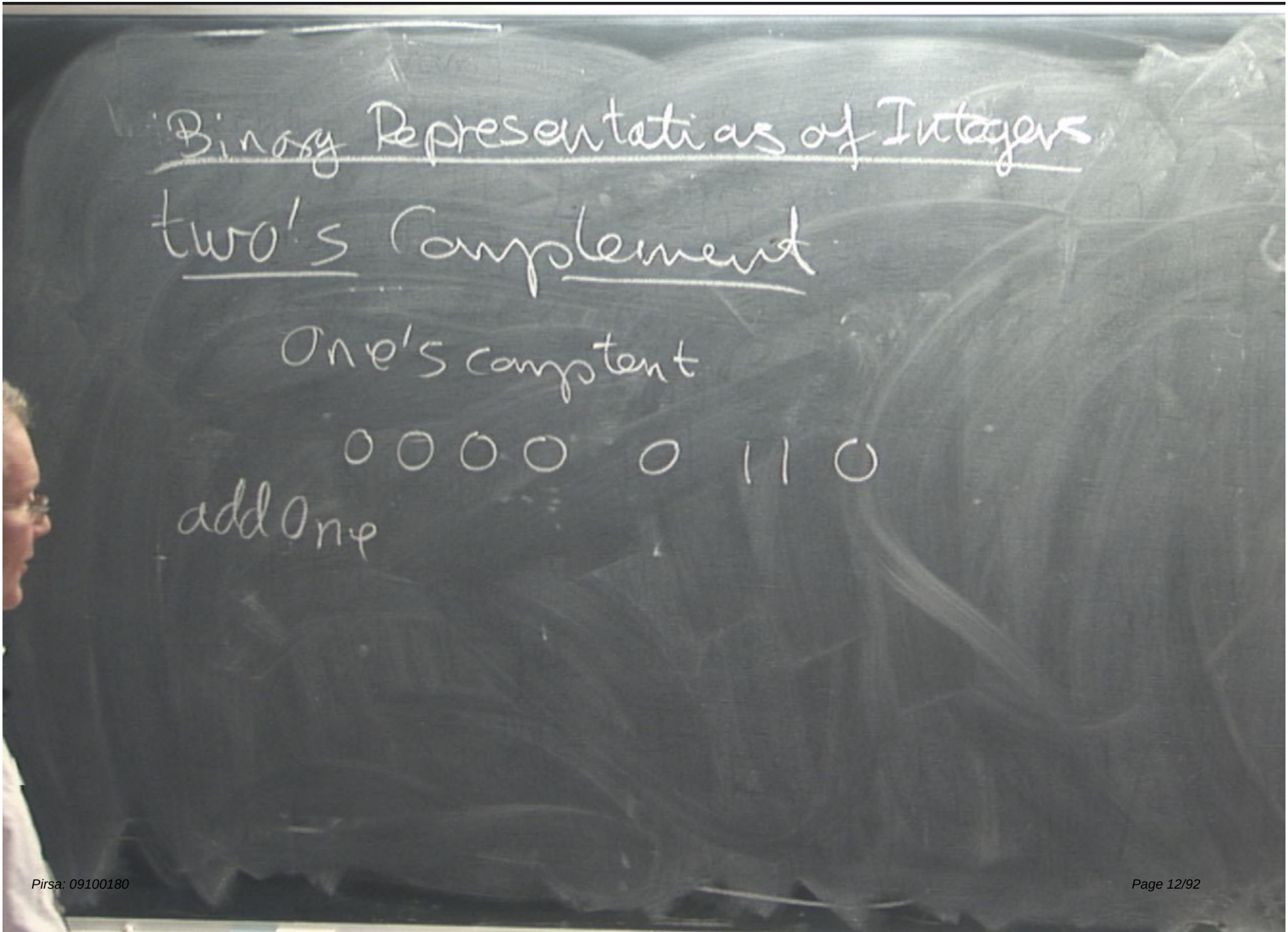
0000 0110 = 6

1111 1001 = -6

1111 1111


Signed Magnitude
1 0000 0110 = -6
sign bit
1000 0000 = -0
0000 0000 = +0
One's Complement
0000 0110 = 6
1111 1001 = -6
1111 1111
0000 000
1111 111


Signed Magnitude
1 0 0 0 0 0 1 1 0 = -6
sign bit
1 0 0 0 0 0 0 0 = -0
0 0 0 0 0 0 0 0 = +0
One's
= 6
= -6
0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1


Binary Representations of Integers
two's Complement
One's complement
0 0 0 0 0 1 1 0
add One

Binary Representations of Integers

two's Complement

One's complement

1111 1001 = -6

add One

0000 0001

1111 1010

Binary Representations of Integers

two's complement

one's complement

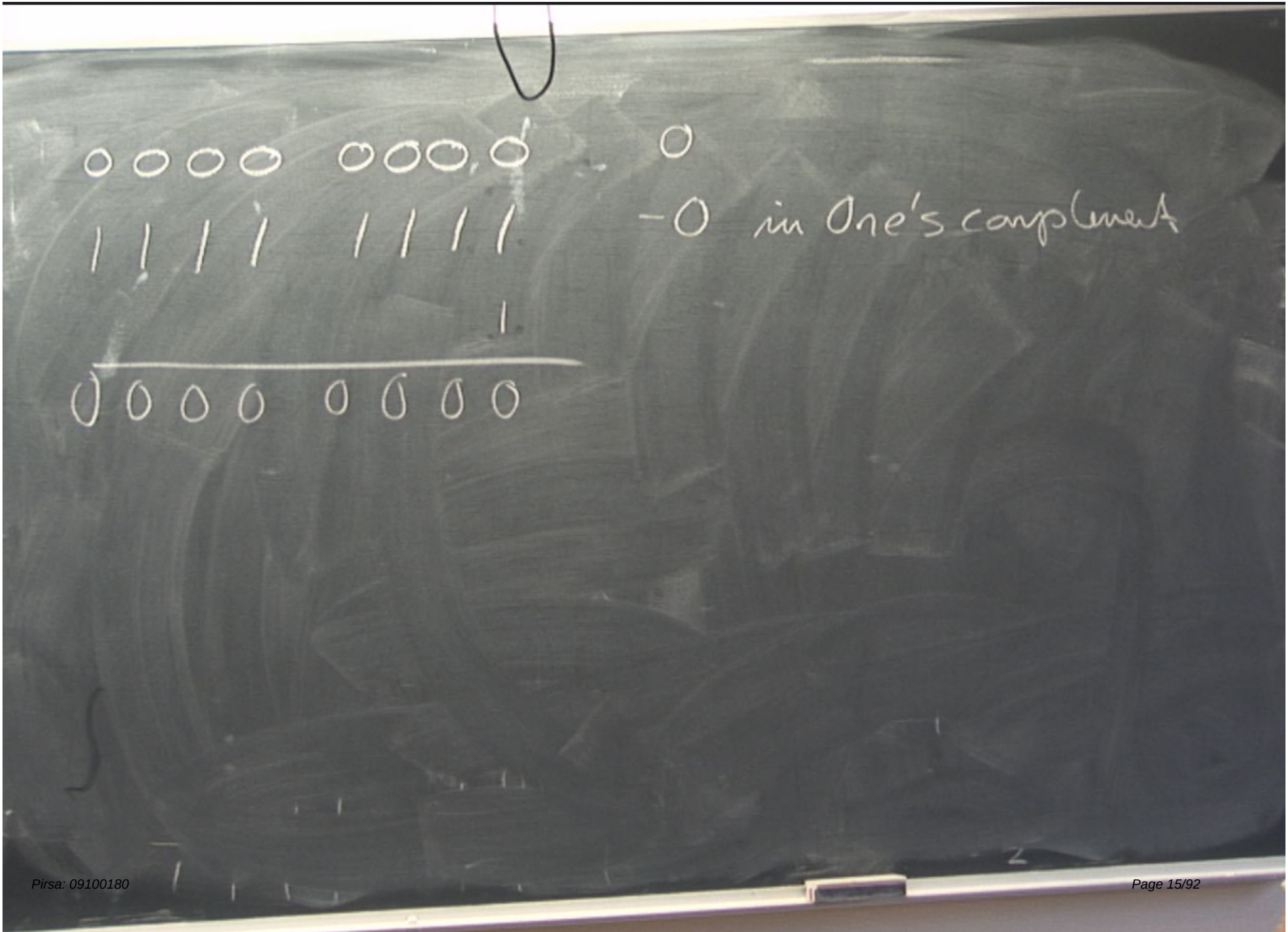
1 1 1 1 1 0 0 1 = -6

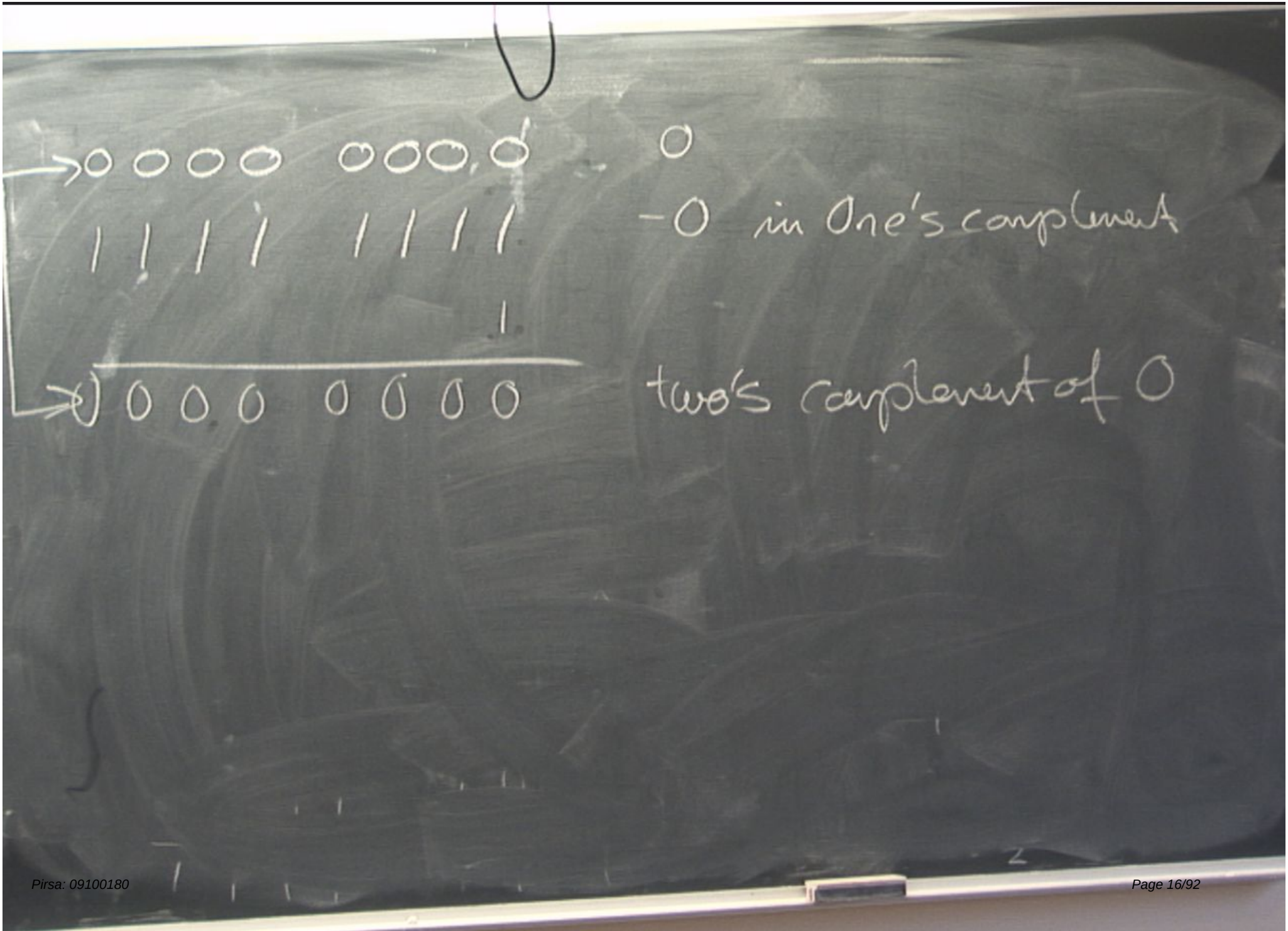
add one

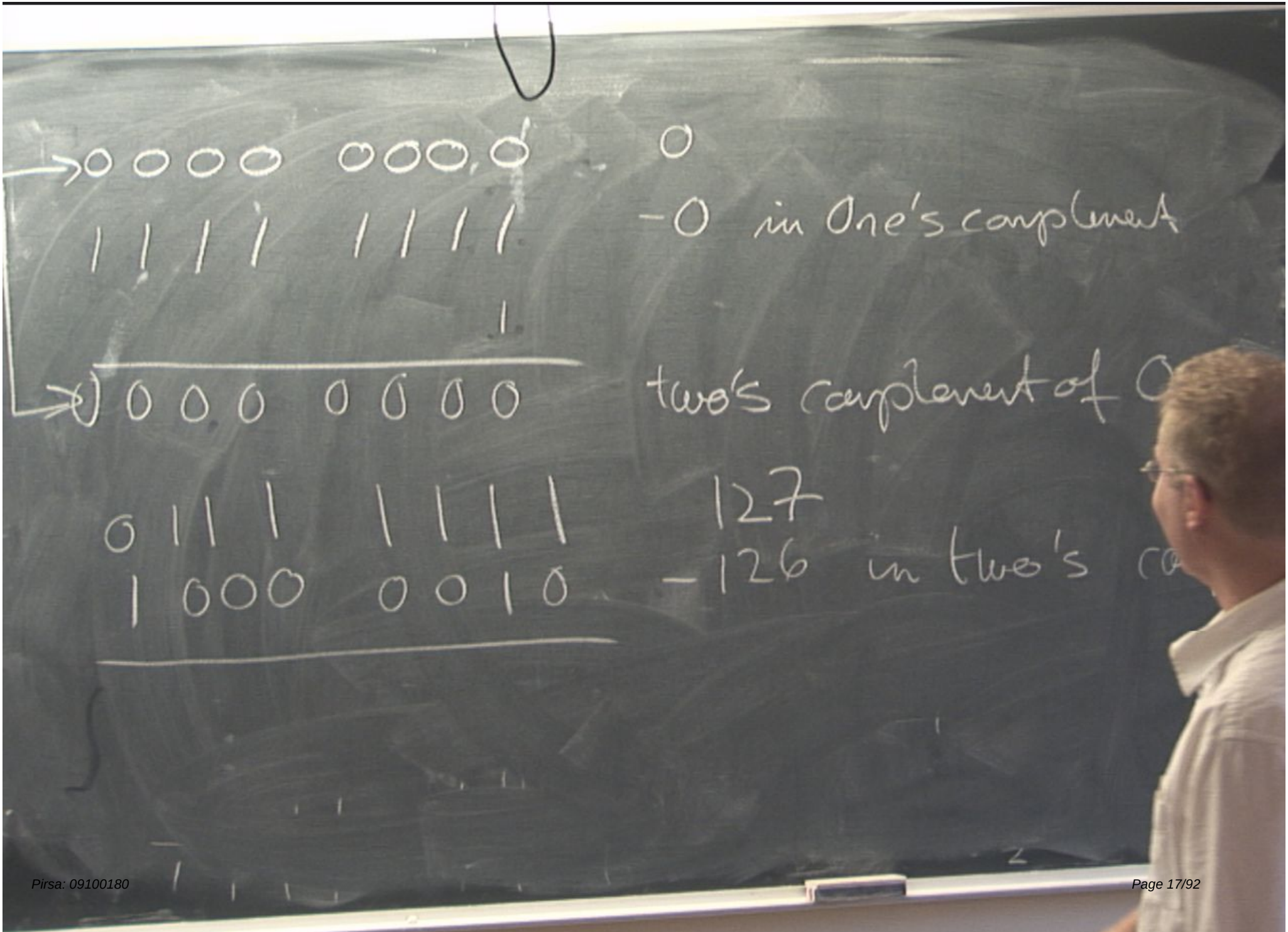
0 0 0 0 0 0 0 1

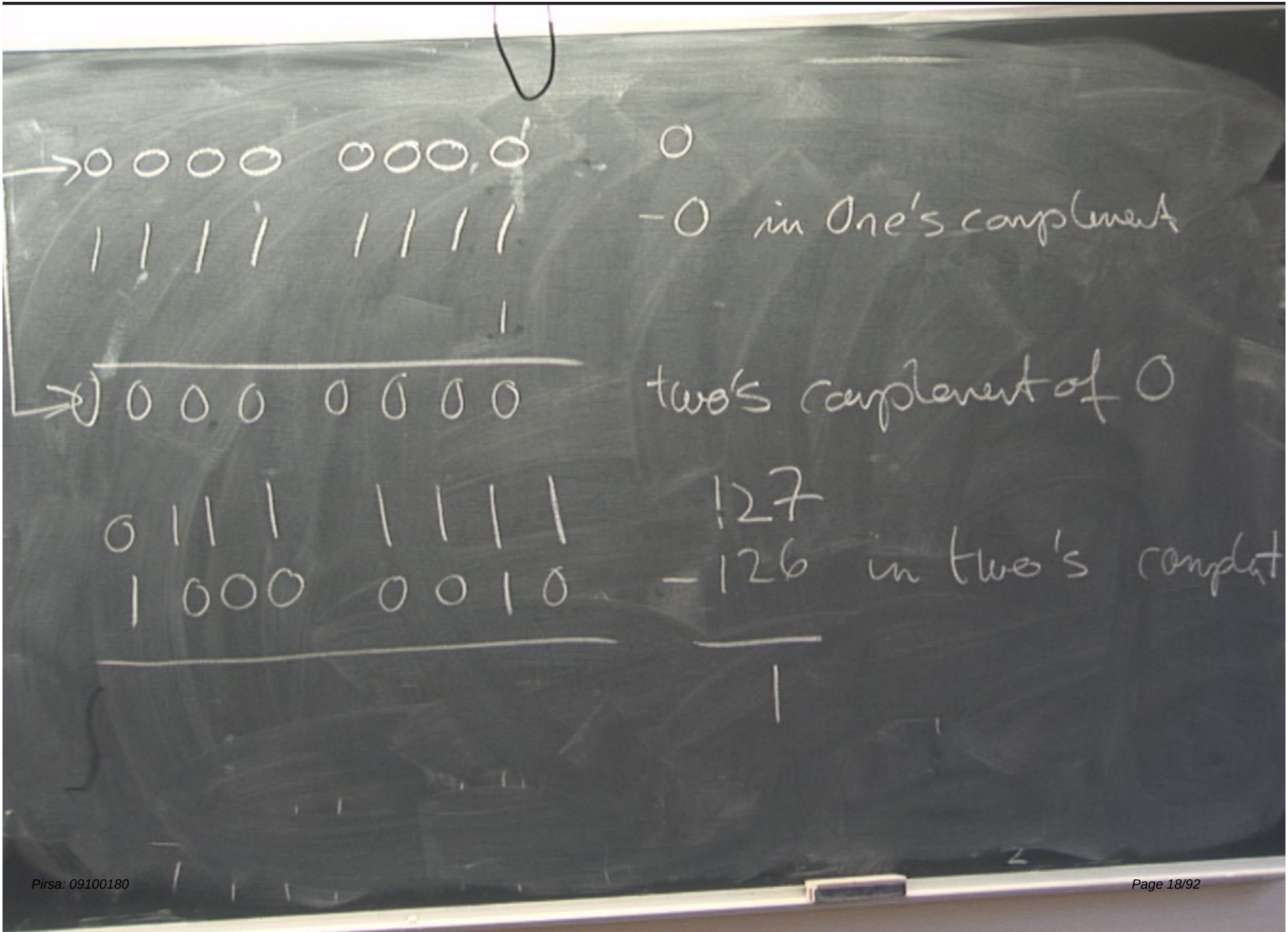
---

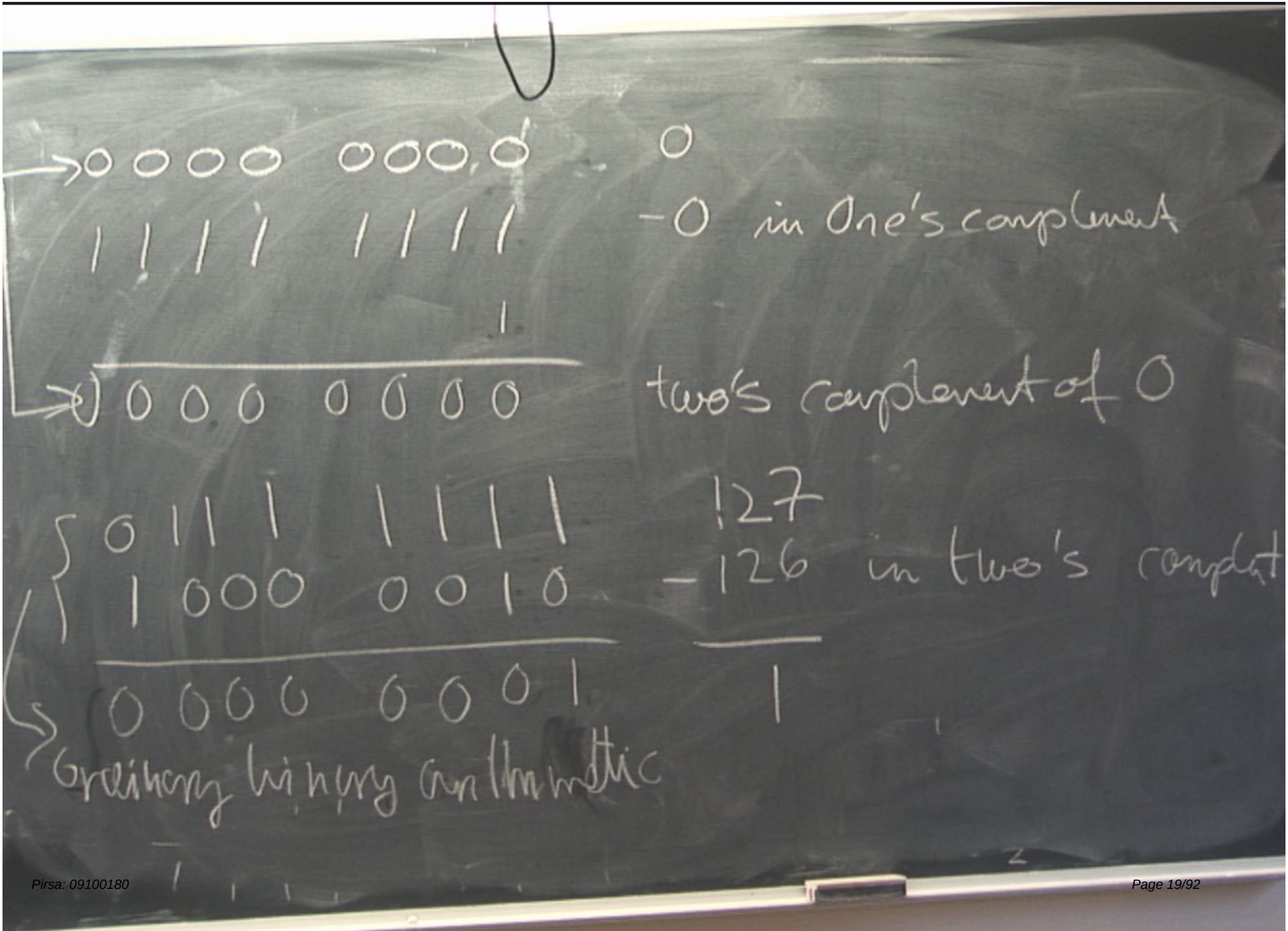
1 1 1 1 1 0 1 0 = -6 in two's complement

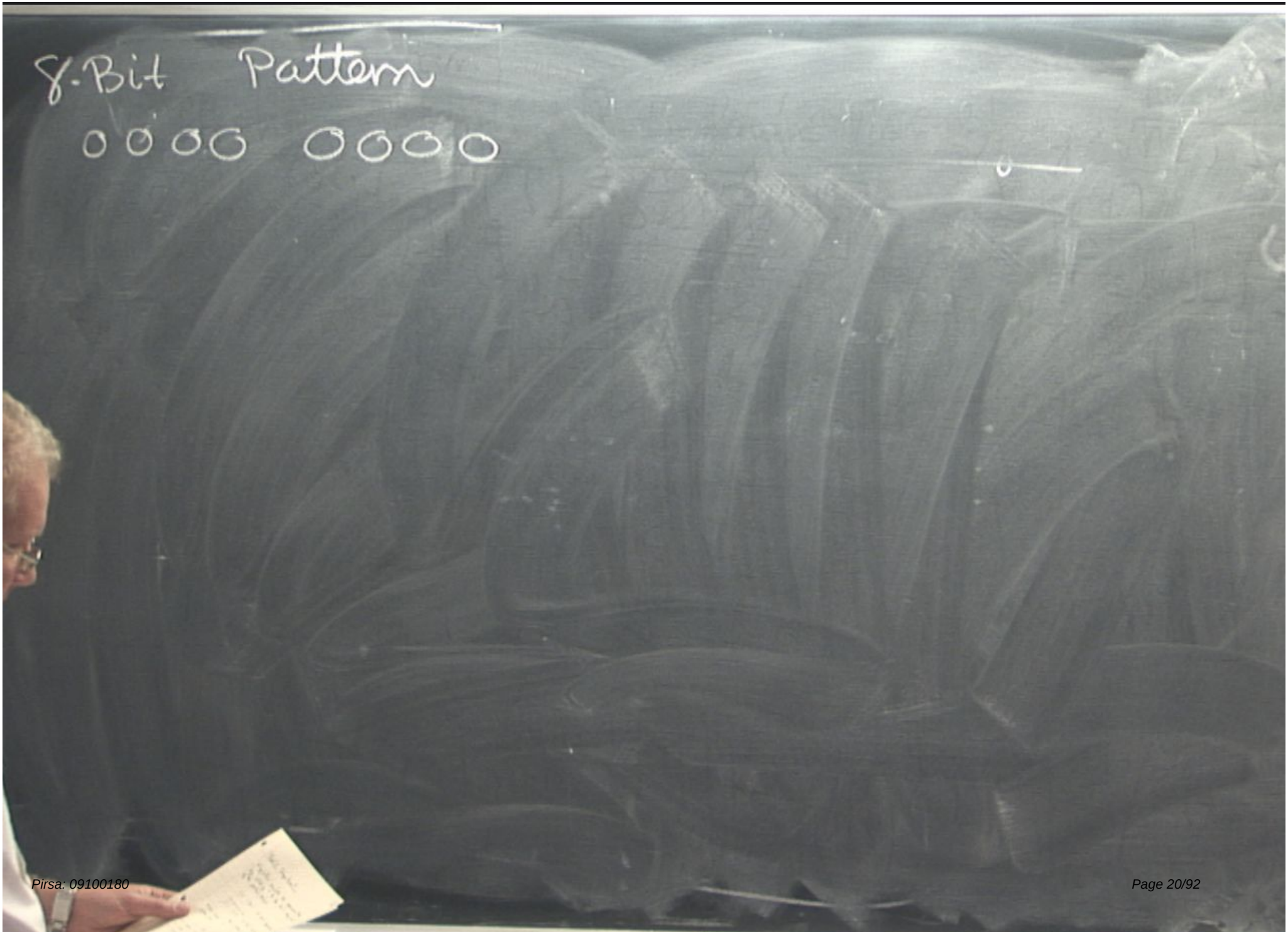

0000 0000 0
1111 1111 -0 in One's complement
1
0000 0000

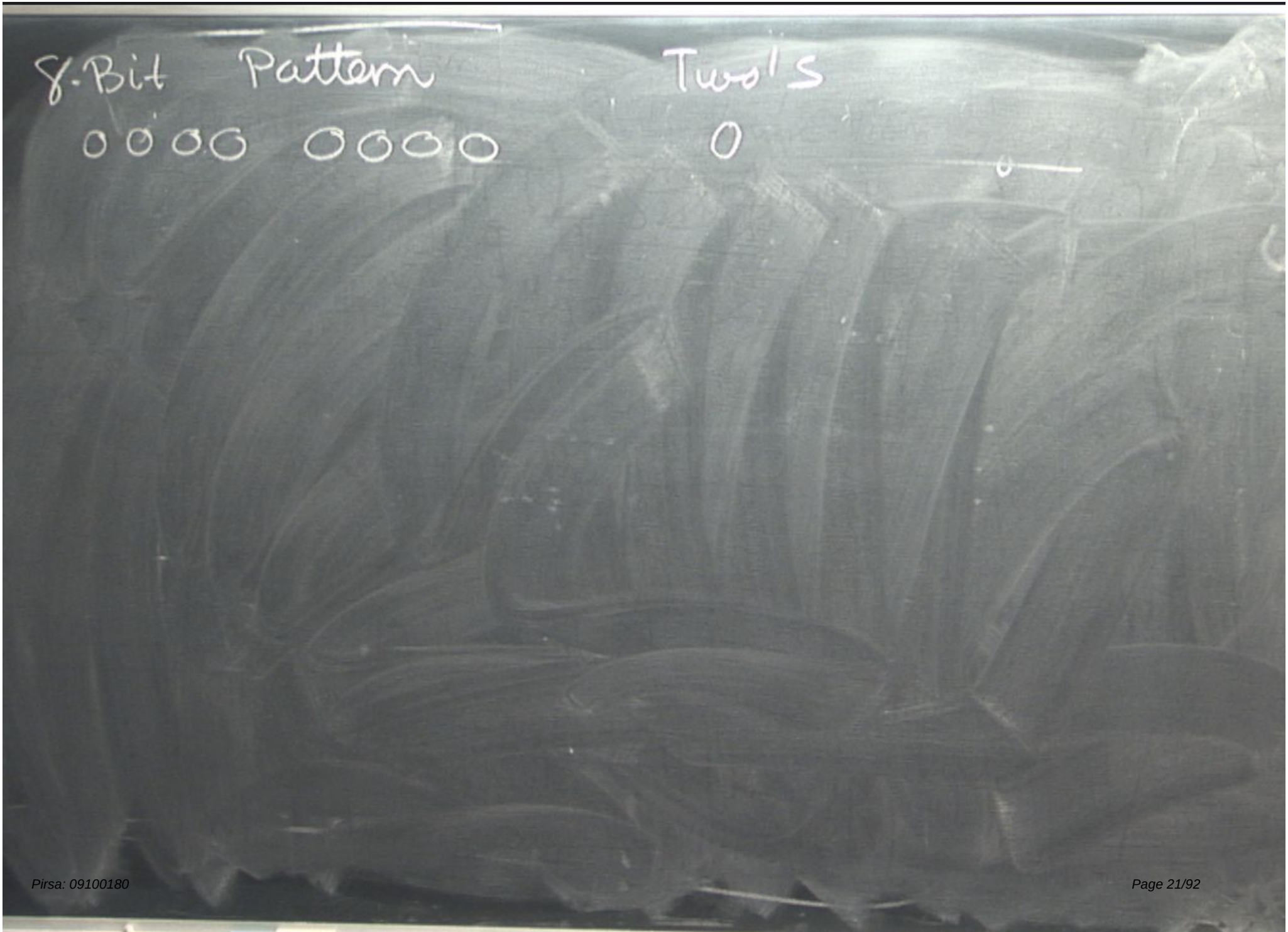

0000 0000 0
1111 1111 -0 in One's complement
1
0000 0000 two's complement of 0

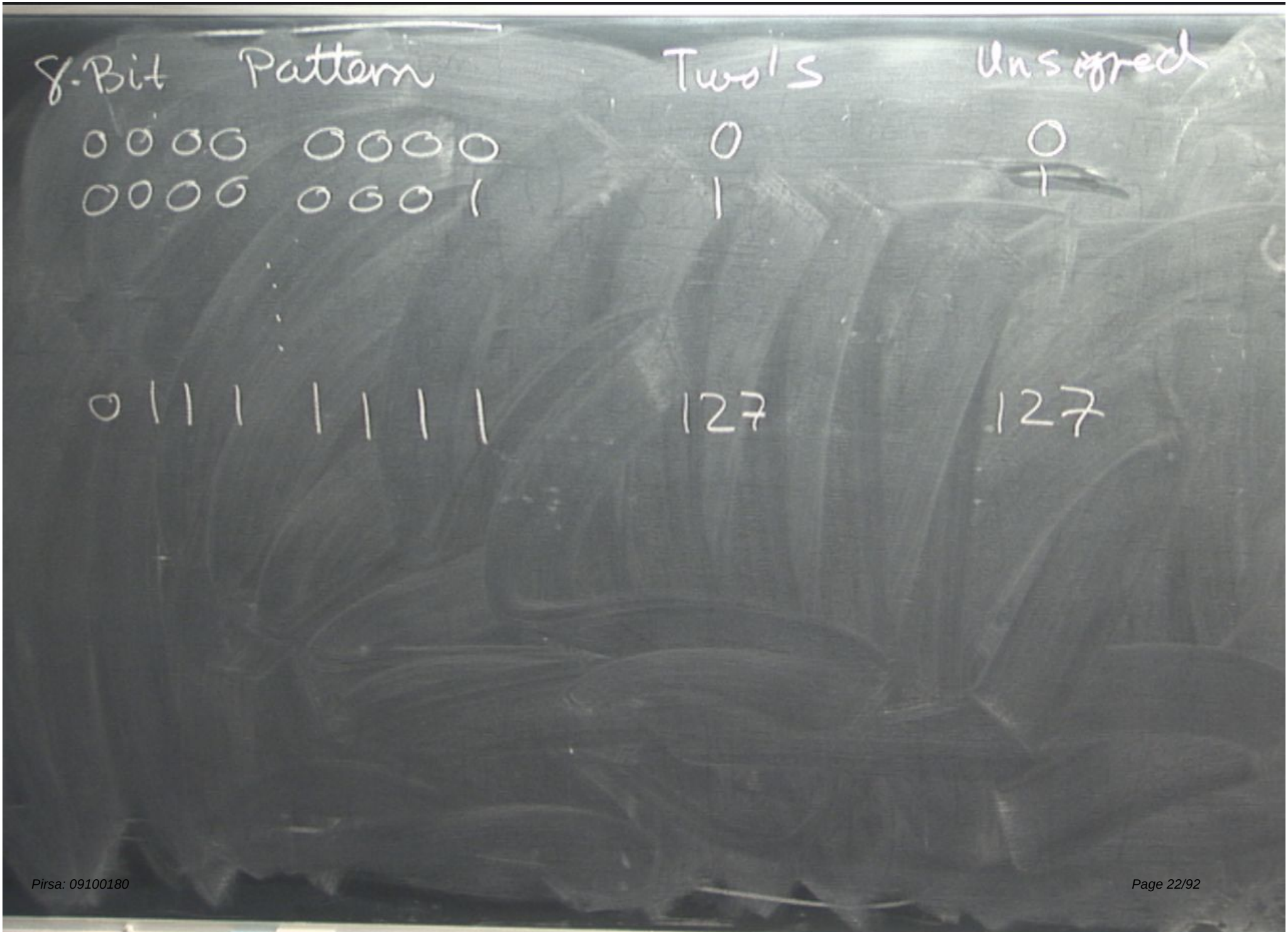

0000 0000 0 0
1111 1111 -0 in One's complement
1
0000 0000 two's complement of 0
0111 1111 127
1000 0010 -126 in two's

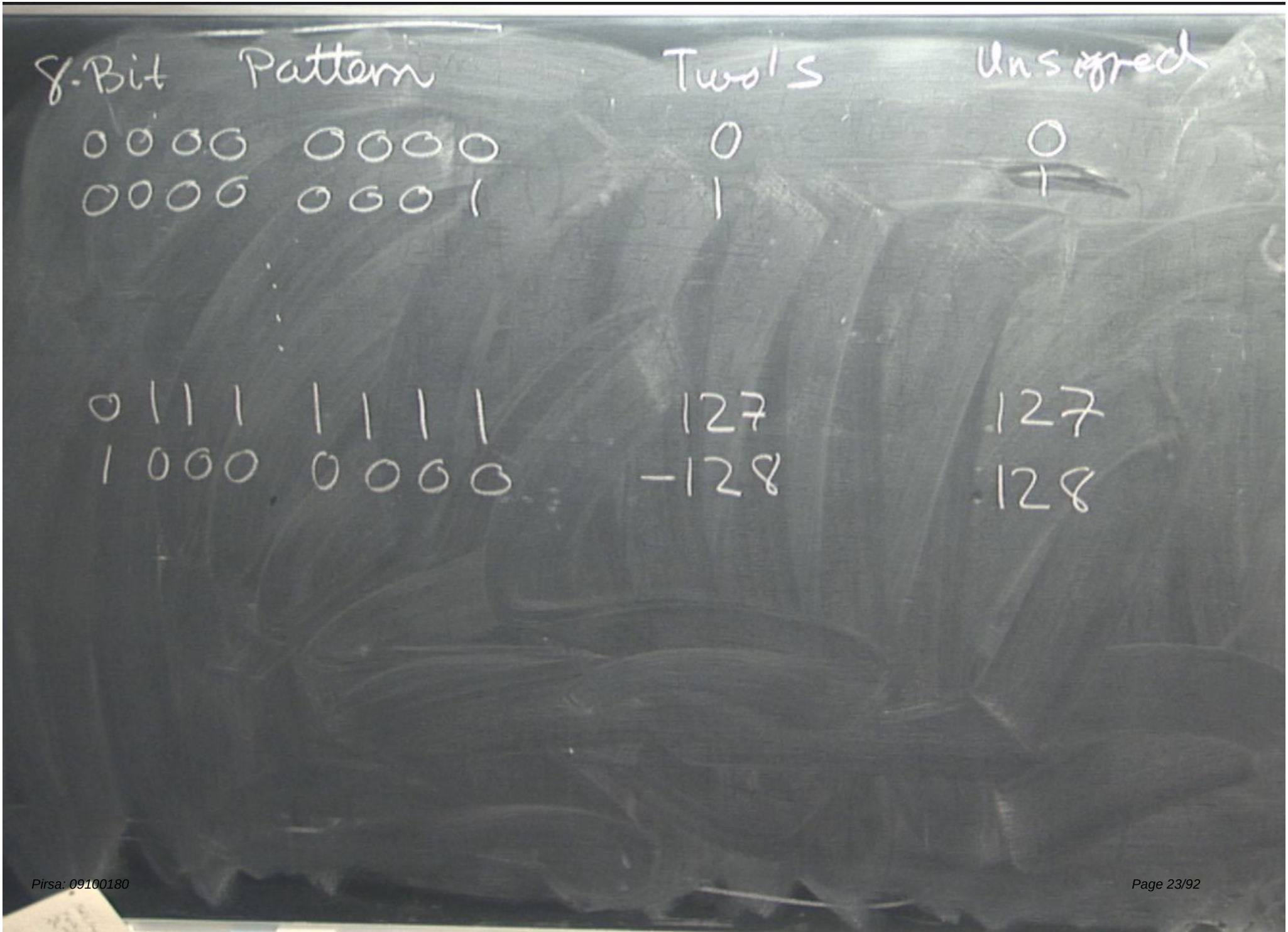

0 0 0 0 0 0 0 0 0
0
-0 in One's complement
1 1 1 1 1 1 1 1
1
0 0 0 0 0 0 0 0 0
two's complement of 0
0 1 1 1 1 1 1 1 1
127
1 0 0 0 0 0 0 1 0
-126 in two's complet
1

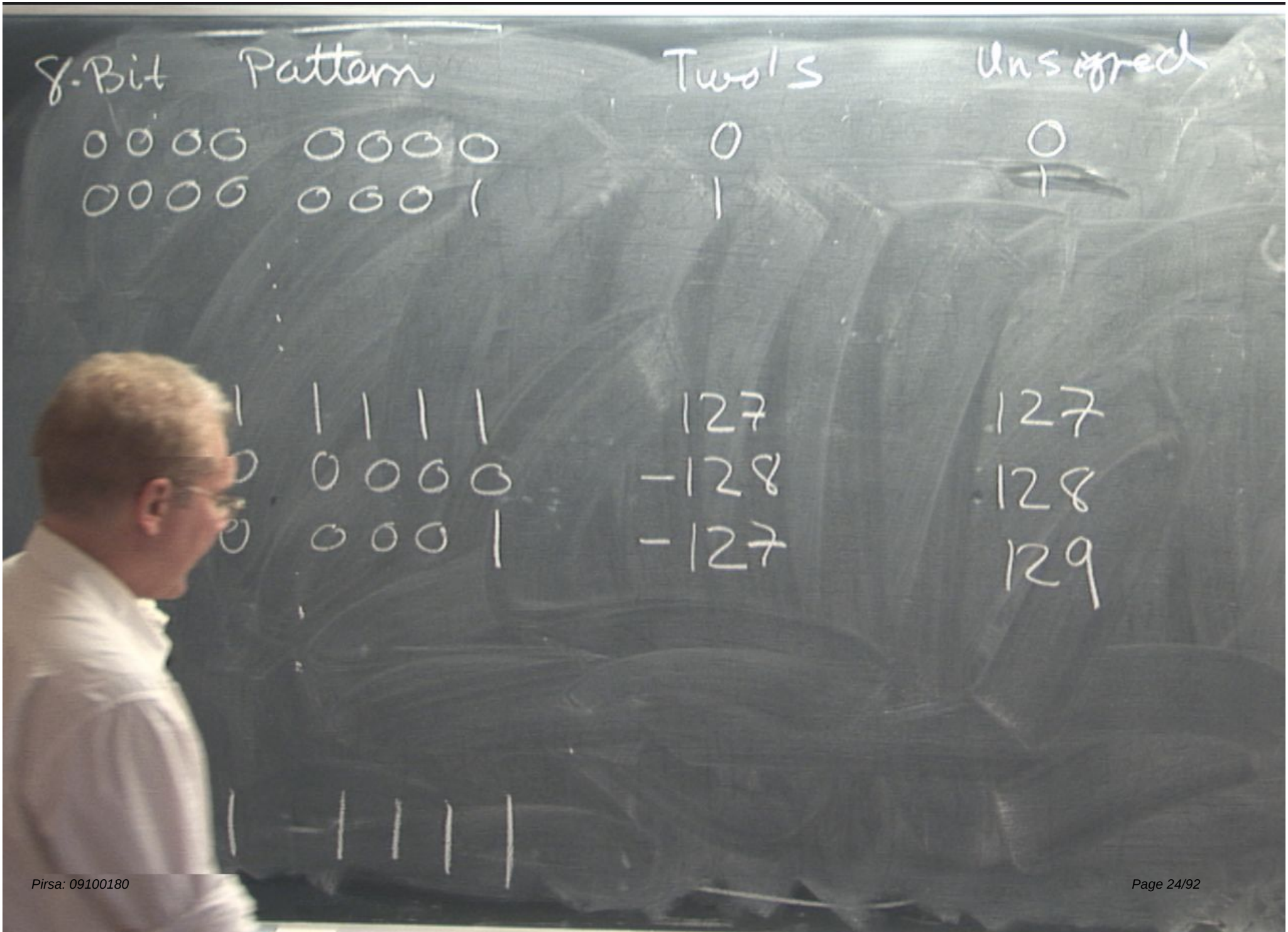

0000 0000 0 0
1111 1111 -0 in One's complement
1
0000 0000 two's complement of 0
0111 1111 127
1000 0010 -126 in two's complet
1
0000 0001
Ordinary binary arithmetic

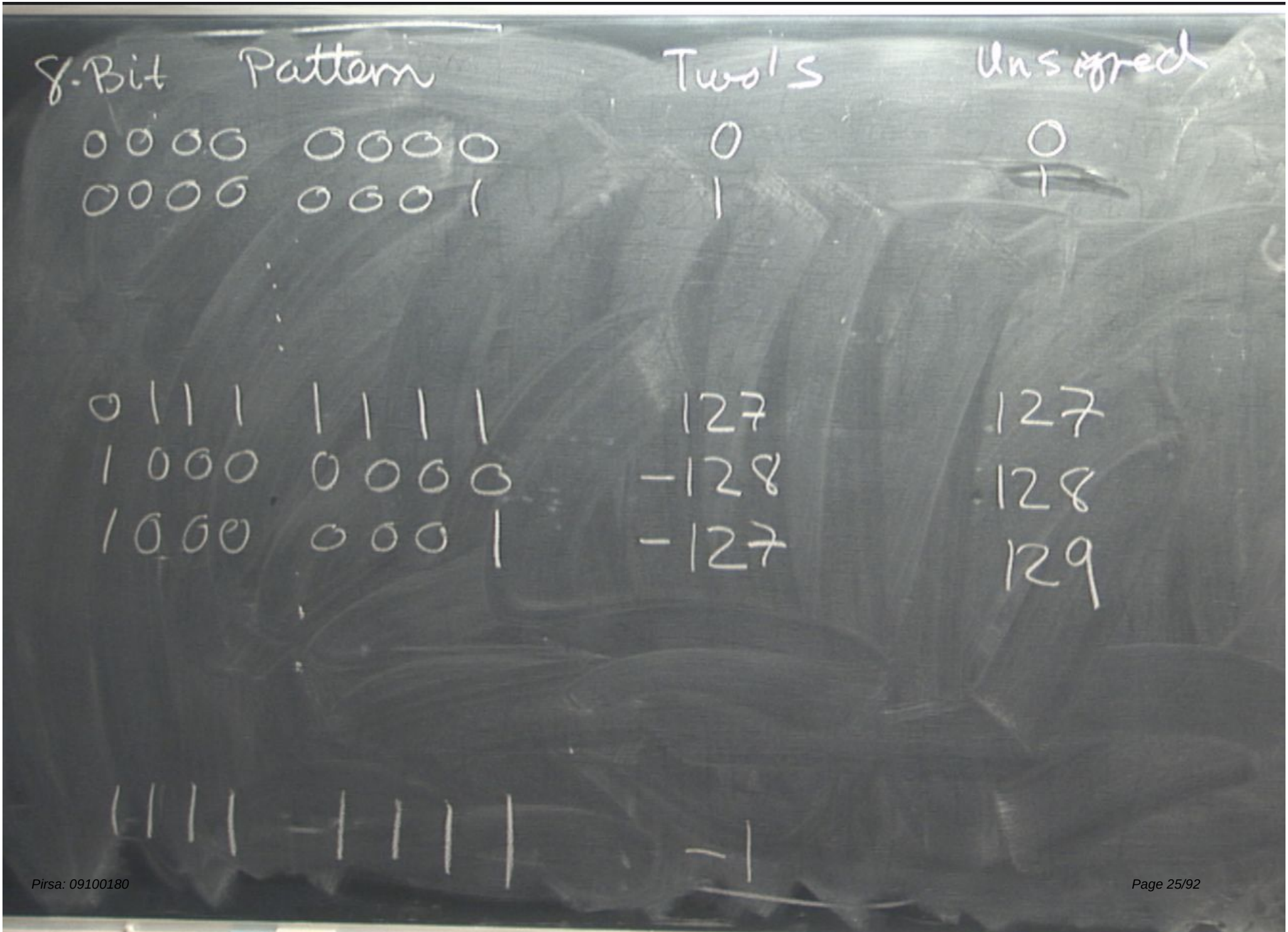

8-Bit Pattern
0000 0000

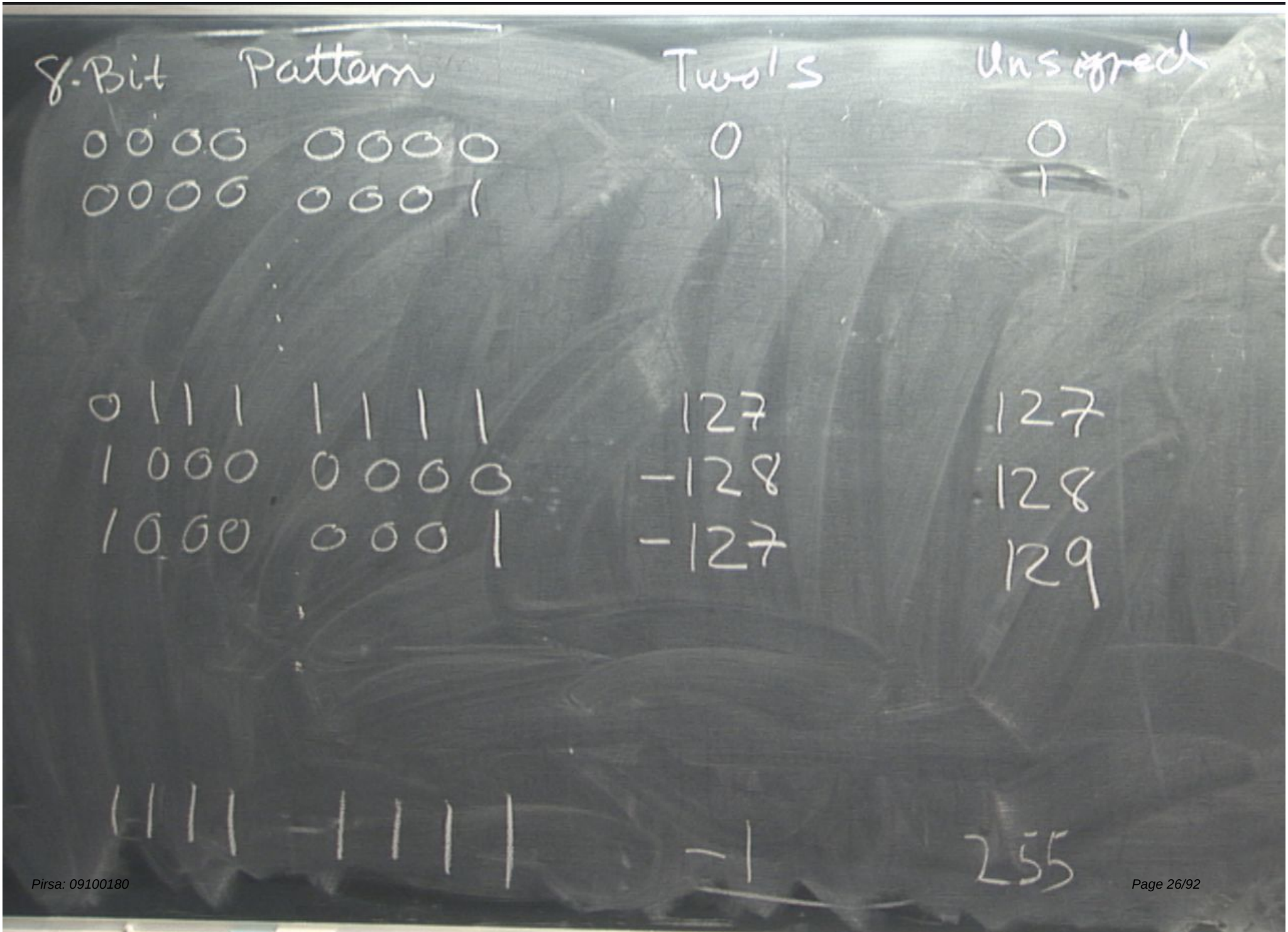

8-Bit Pattern
0000 0000
Two's
0


8-Bit Pattern
Two's
Unsigned
0000 0000
0
0
0000 0001
1
1
0111 1111
127
127

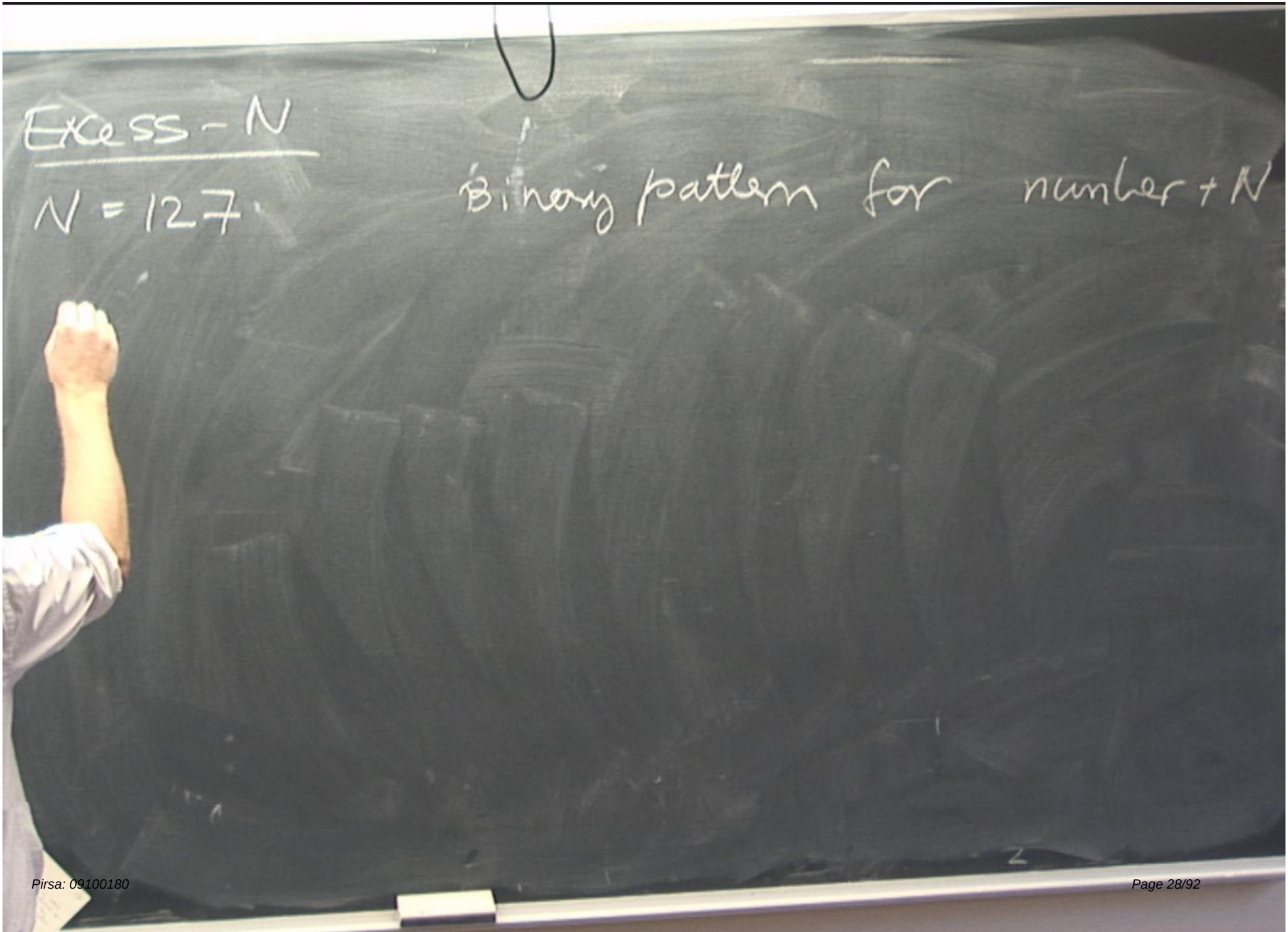

8-Bit Pattern
Two's
Unsigned
0000 0000
0
0
0000 0001
1
1
0111 1111
127
127
1000 0000
-128
128

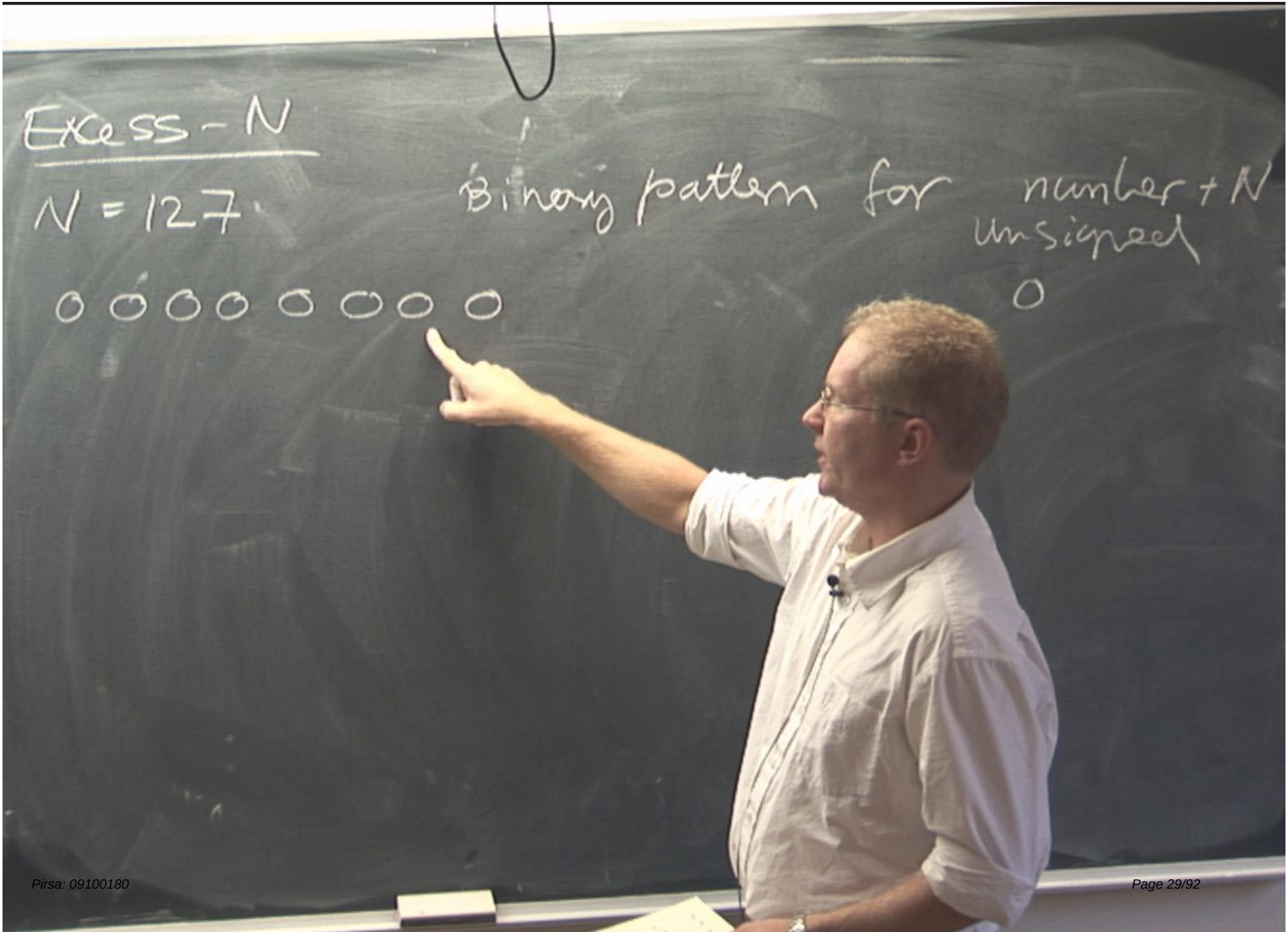

8-Bit Pattern
Two's
Unsigned
0000 0000
0
0
0000 0001
1
1
0111 1111
127
127
1000 0000
-128
128
1000 0001
-127
129
1111 1111

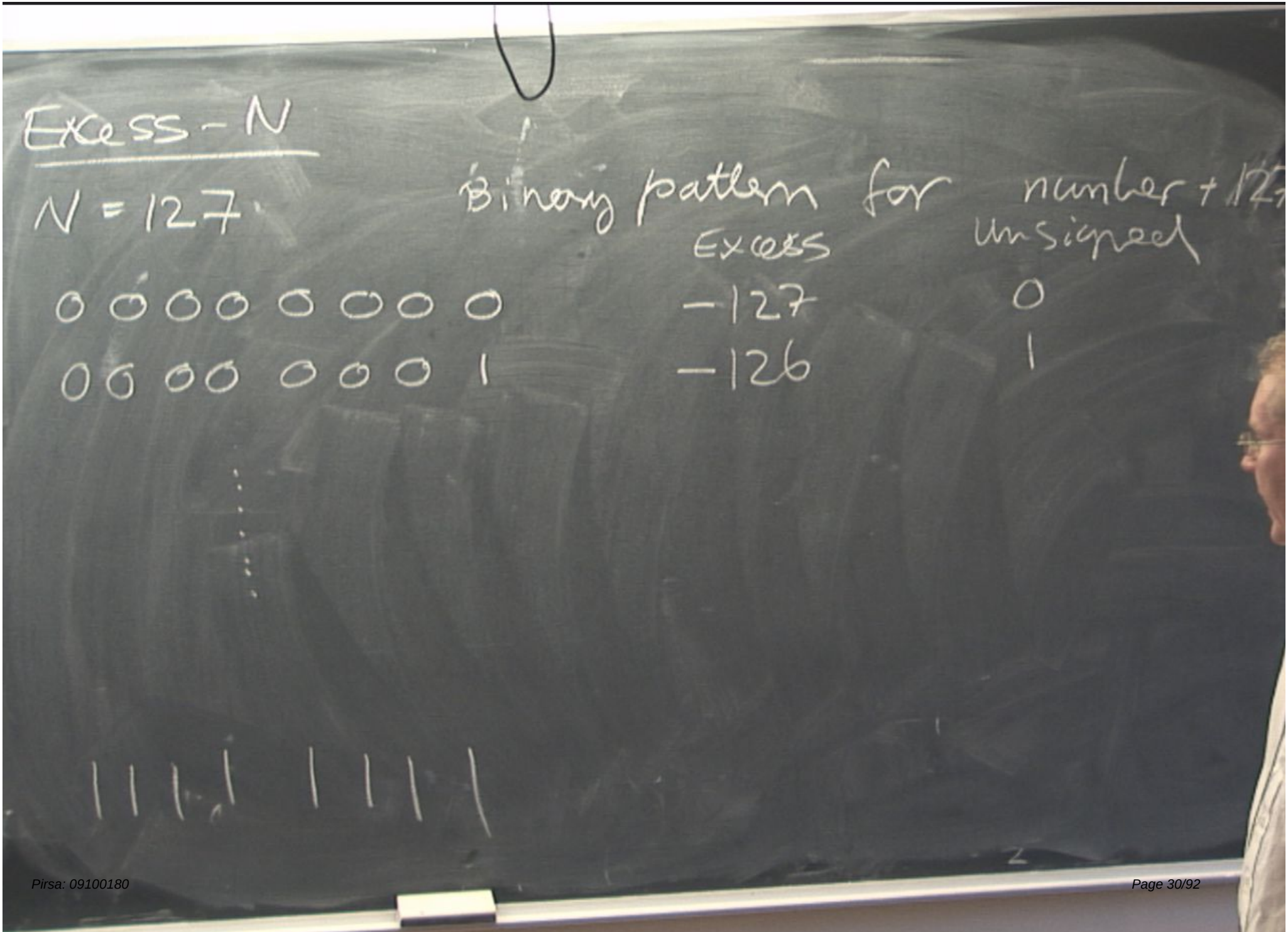

8-Bit Pattern
Two's
Unsigned
0000 0000
0
0
0000 0001
1
1
0111 1111
127
127
1000 0000
-128
128
1000 0001
-127
129
1111 1111
-1

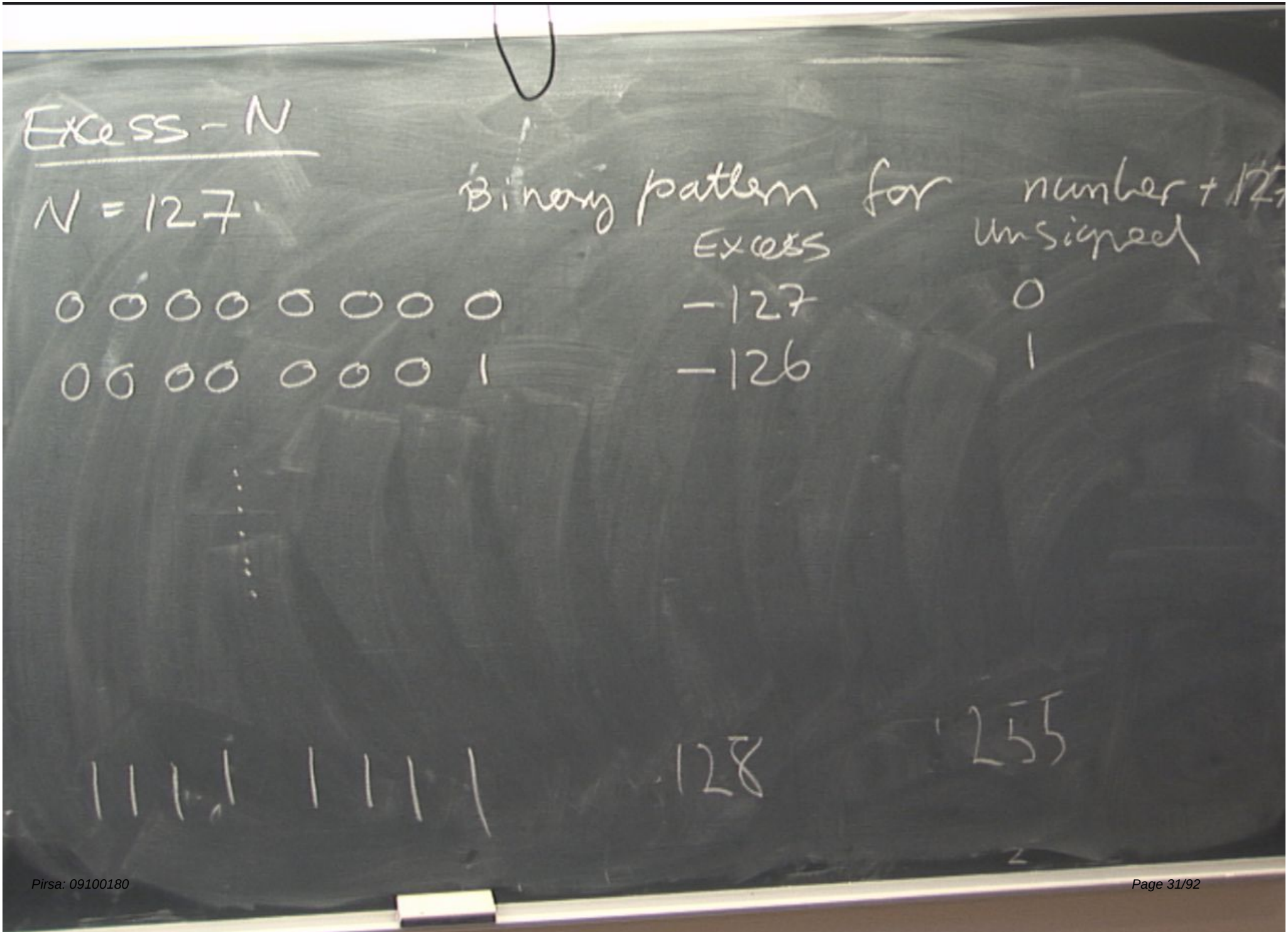

8-Bit Pattern
Two's
Unsigned
0000 0000
0
0
0000 0001
1
1
0111 1111
127
127
1000 0000
-128
128
1000 0001
-127
129
1111 1111
-1
255

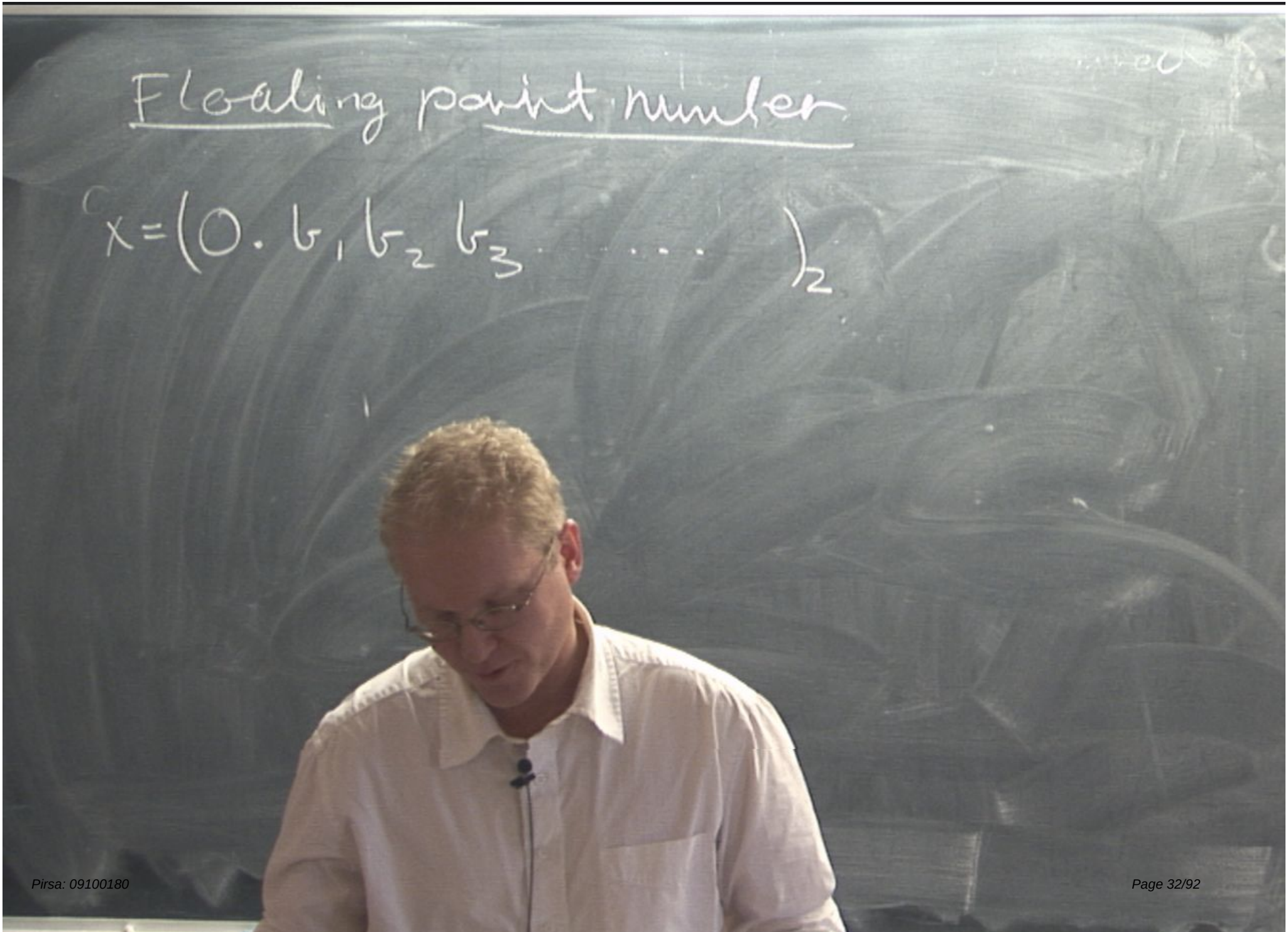

Excess - N
N = 127


Excess - N
N = 127
Binary pattern for number + N


Excess - N
N = 127
Binary pattern for number + N
Unsigned
0 0 0 0 0 0 0 0
0


Excess - N
N = 127
Binary pattern for number + 127
Excess
Unsigned
0 0 0 0 0 0 0 0 0
−127
0
0 0 0 0 0 0 0 0 1
−126
1


Excess-N
N = 127
Binary pattern for number + 127
Excess
Unsigned
0 0 0 0 0 0 0 0 0
−127
0
0 0 0 0 0 0 0 0 1
−126
1
1 1 1 1 1 1 1 1
128
255


Floating point number
$x = (0.b_1 b_2 b_3 \ldots)_2$

Floating point number

$$x = (0.b_1 b_2 b_3 \ldots\ldots)_2$$

$$= b_1 \times 2^{-1} + b_2 \times 2^{-2} \ldots\ldots$$

$$= (4b_1 + 2b_2 + b_3) \times 8^{-1} + (4b_4 + 2b_5 + b_6) \times$$

↖ Octal form

Floating point number

$= (0.b_1 b_2 b_3 \ldots)_2$

$= b_1 \times 2^{-1} + b_2 \times 2^{-2} \ldots$

$= (4b_1 + 2b_2 + b_3) \times 8^{-1} + (4b_4 + 2b_5 + b_6) \times 8^{-2}$

↖ Octal form

Floating point number

$$= (0.b_1 b_2 b_3 \ldots)_2$$

$$= b_1 \times 2^{-1} + b_2 \times 2^{-2} \ldots$$

$$= (4b_1 + 2b_2 + b_3) \times 8^{-1} + (4b_4 + 2b_5 + b_6) \times 8^{-2}$$

↖ Octal form

Floating point number

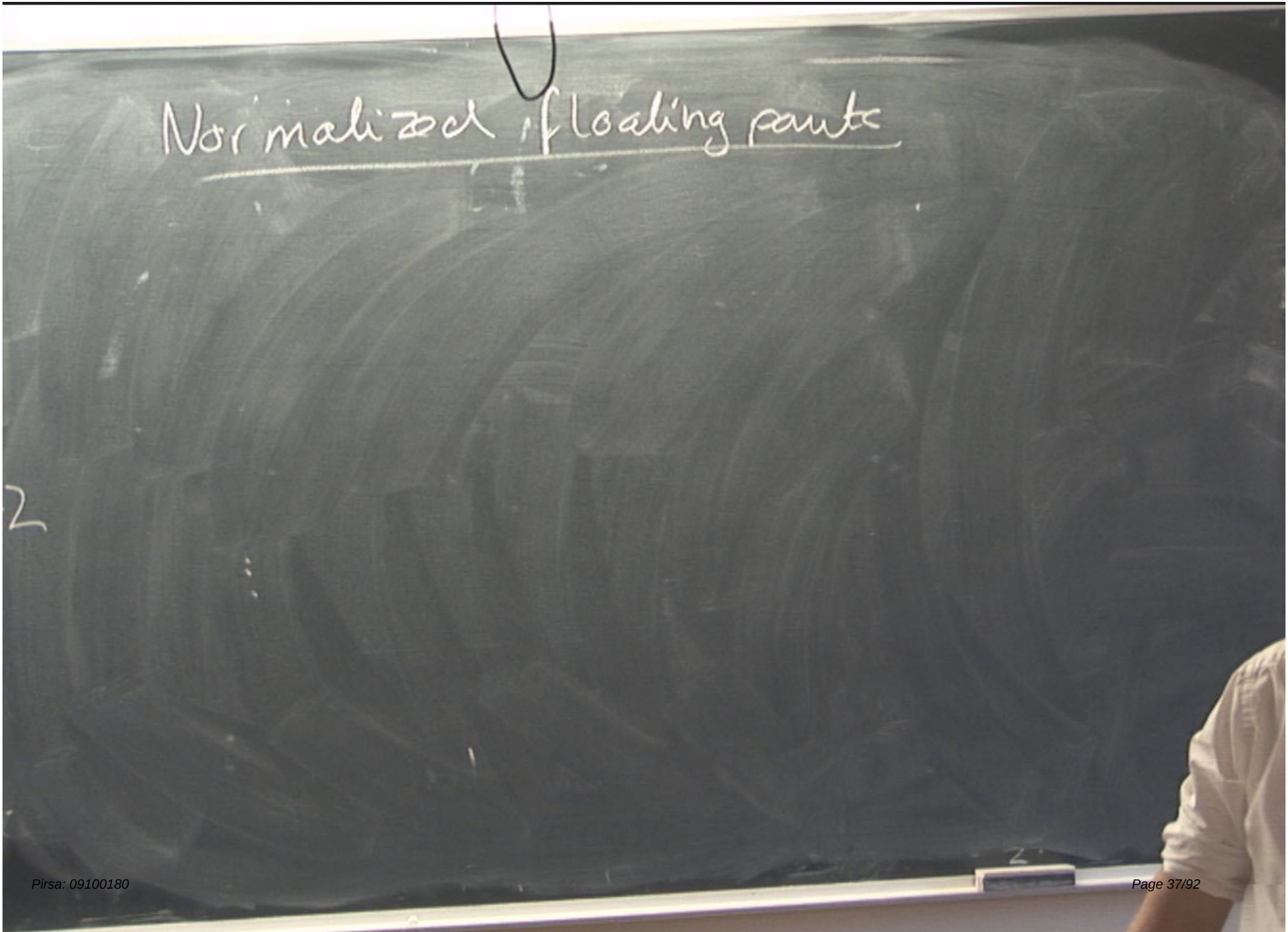
$= (0.b_1 b_2 b_3 \ldots)_2$

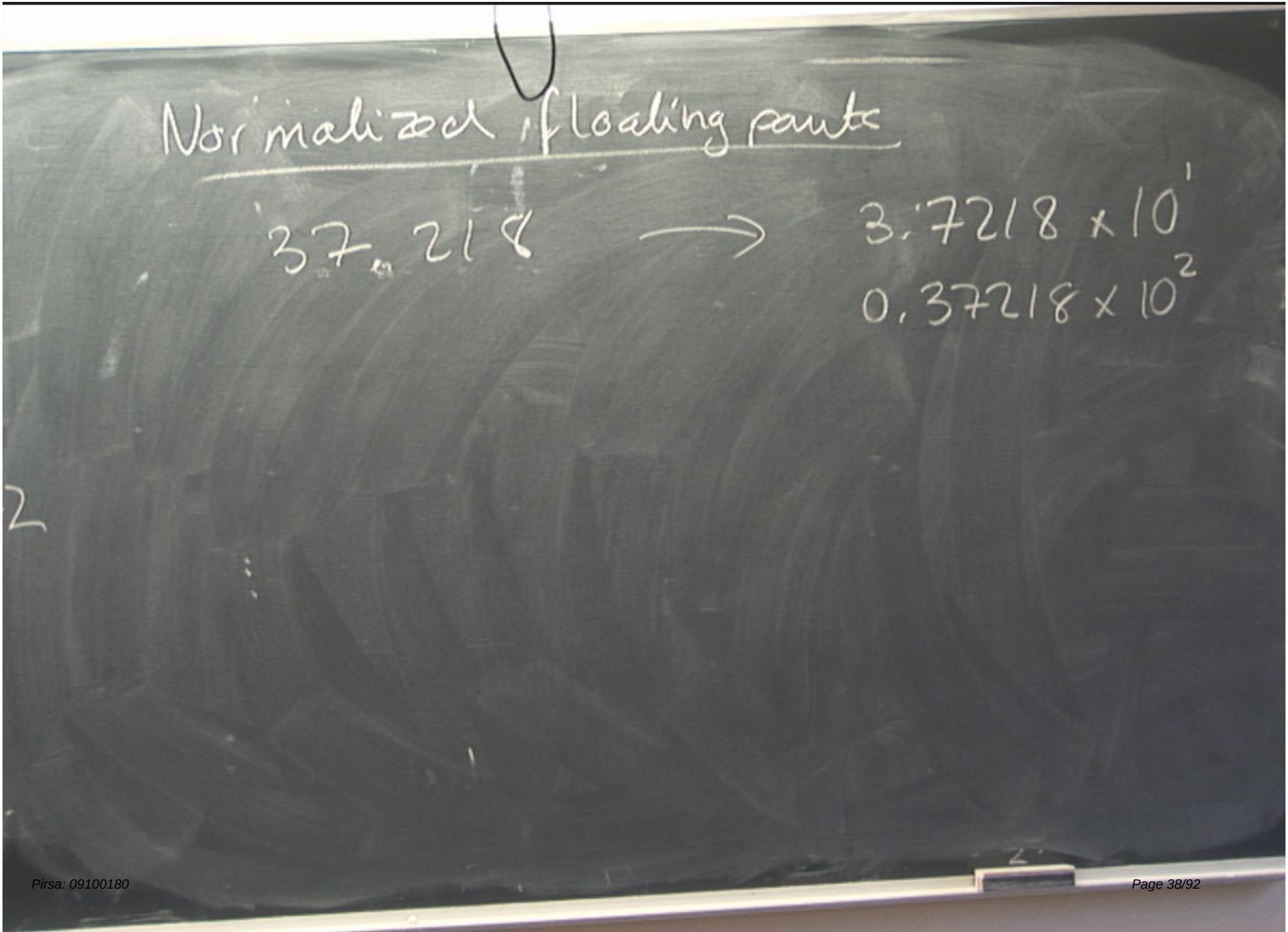
$= b_1 \times 2^{-1} + b_2 \times 2^{-2} \ldots$

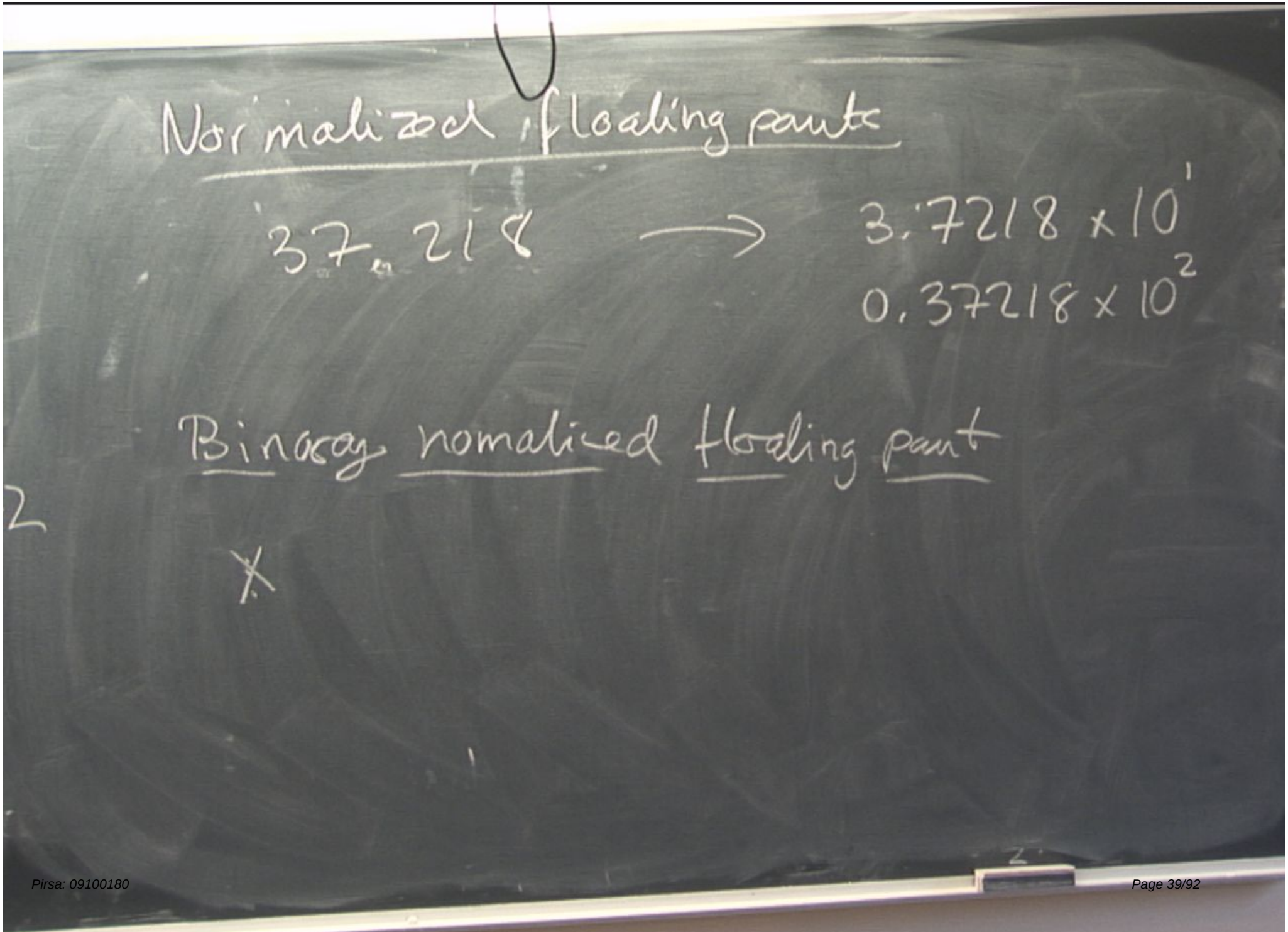
$= (4b_1 + 2b_2 + b_3) \times 8^{-1} + (4b_4 + 2b_5 + b_6) \times 8^{-2}$

↖ Octal form

$(101\ 011\ .\ 110\ 011)_2$


Normalized floating point


Normalized floating point
37.218
3.7218 x 10^1
0.37218 x 10^2


Normalized floating point
37.218 → 3.7218 × 10^1
0.37218 × 10^2
Binary normalized floating point
2
X

Normalized floating point

$37.218 \longrightarrow 3.7218 \times 10^{1}$

0.37218×10^{2}

Binary normalized floating point

$$x = \pm q \times 2^{m}$$

$$q = (0.b_1 b_2 \cdots)_2$$

Normalized floating point

$37.218 \longrightarrow 3.7218 \times 10^{1}$

0.37218×10^{2}

Binary normalized floating point

$$x = \pm q \times 2^{m}$$

$$q = (0.b_1 b_2 \cdots)_2 \qquad b_1 \neq 0$$

Normalized floating point

$37.218 \longrightarrow 3.7218 \times 10^{1}$

0.37218×10^{2}

Binary normalized floating point

$$x = \pm q \times 2^{m}$$

$$q = (0.b_1 b_2 \cdots)_2 \qquad b_1 \neq 0 \qquad \frac{1}{2} \leq q < 1$$

or

Normalized floating point

$37.218 \longrightarrow 3.7218 \times 10^{1}$

0.37218×10^{2}

Binary normalized floating point

$$x = \pm q \times 2^{m}$$

$$q = (0.b_1 b_2 \cdots)_2 \qquad b_1 \neq 0 \qquad \frac{1}{2} \leq q < 1$$

or $q = (1.b_1 b_2 \cdots)_2$

Normalized floating point

$37.218 \longrightarrow 3.7218 \times 10^{1}$

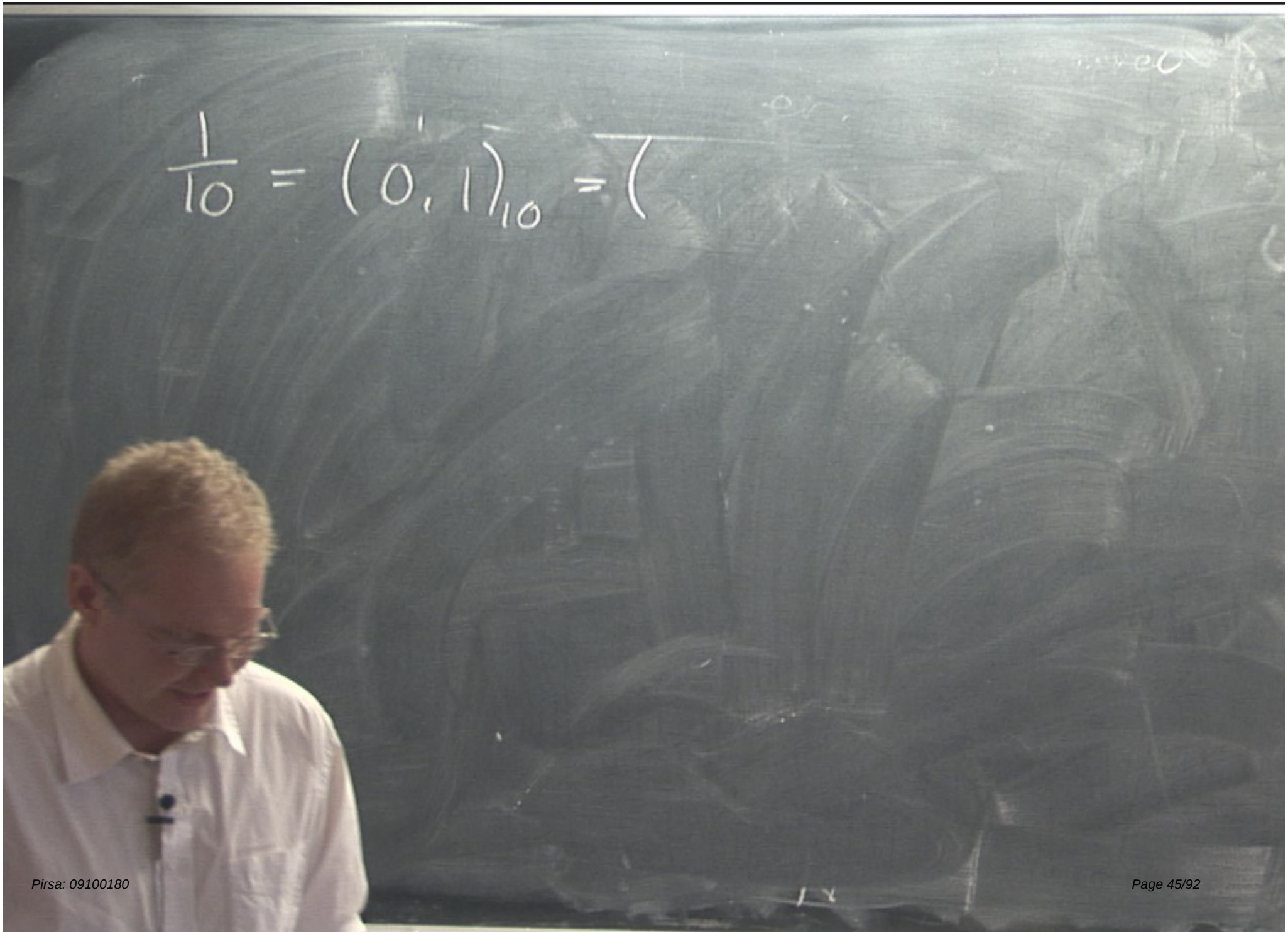
0.37218×10^{2}

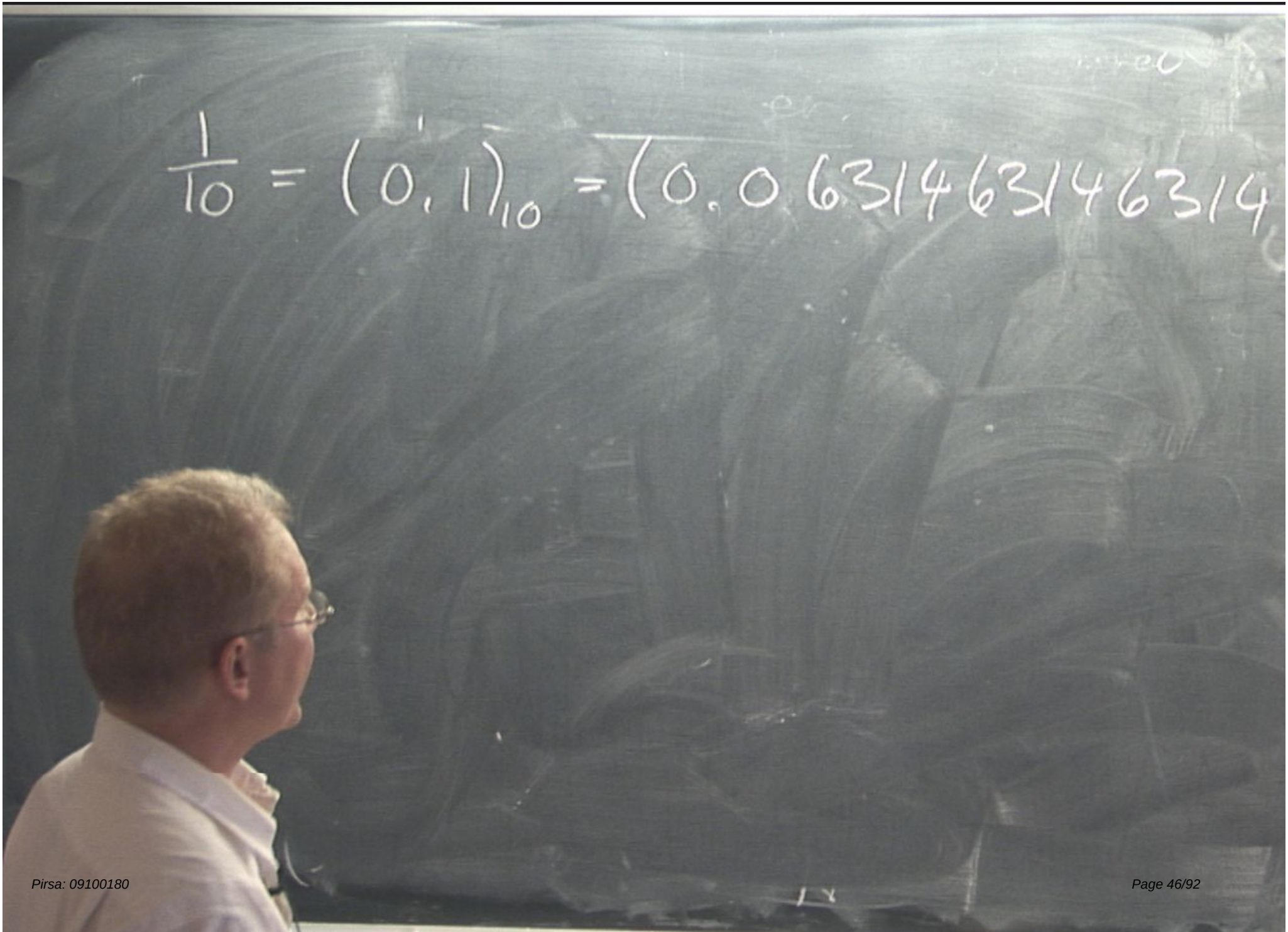
Binary normalized floating point

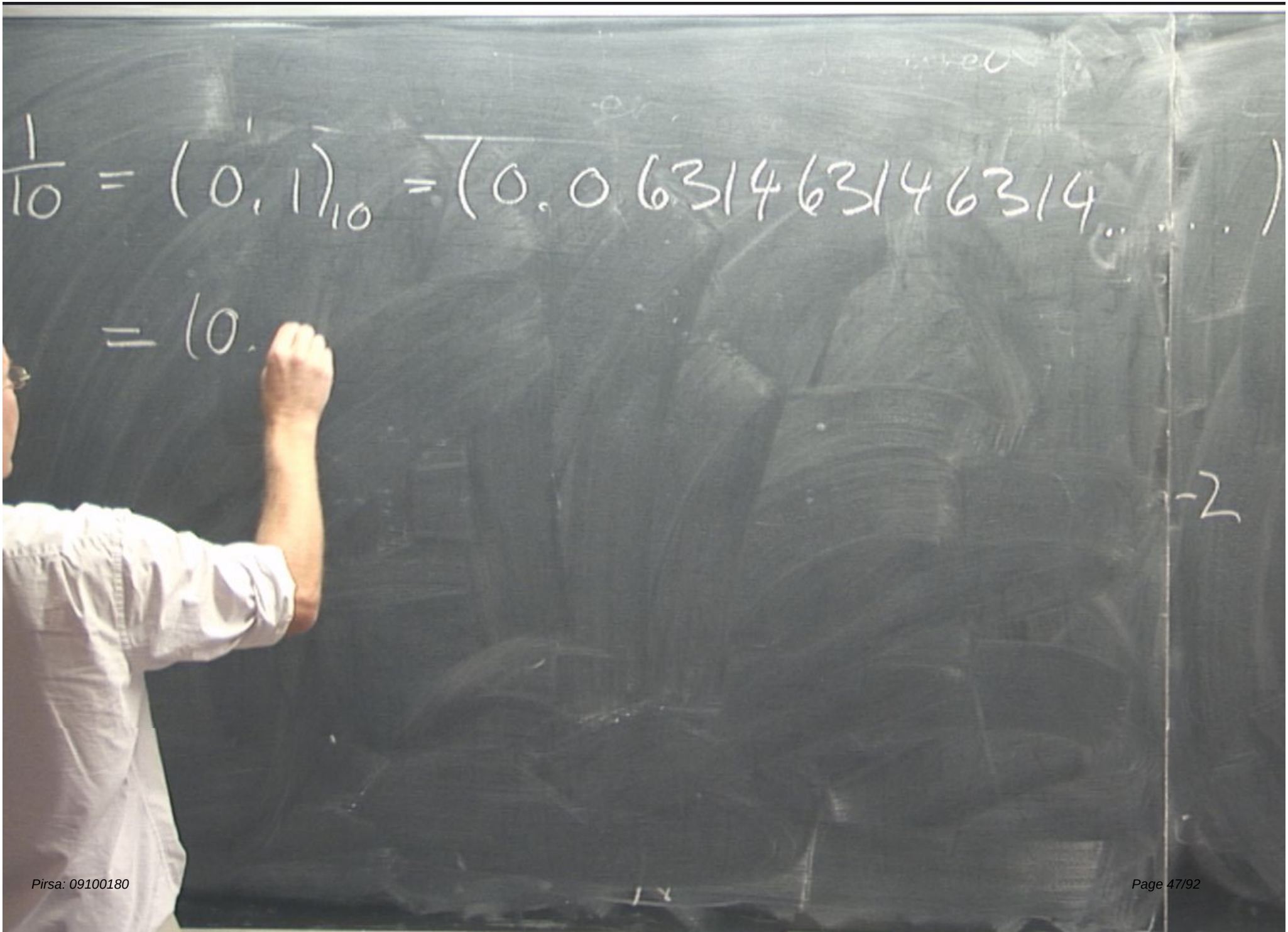
$$x = \pm q \times 2^{m}$$

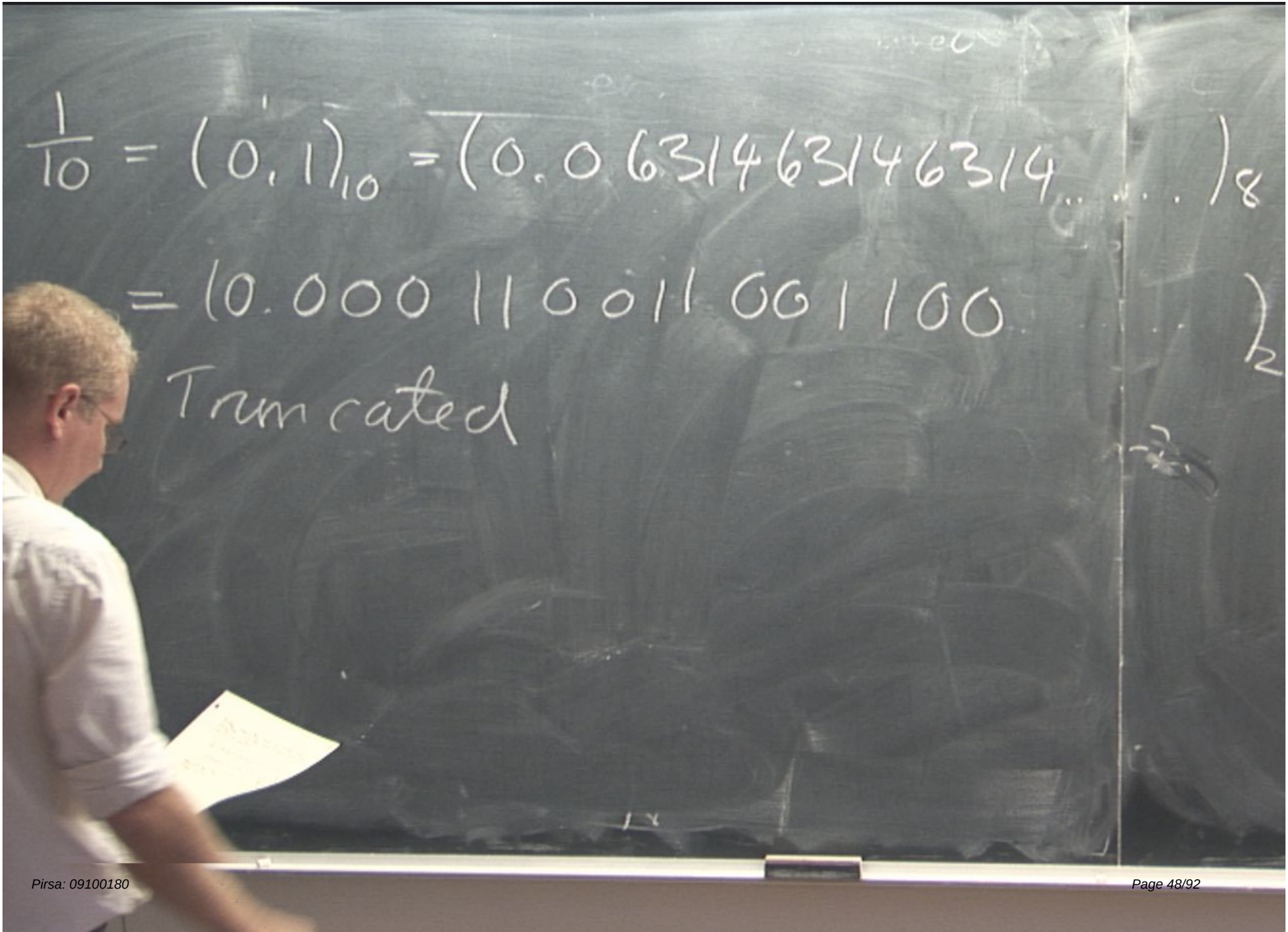
$$q = (0.b_1 b_2 \cdots)_2 \qquad b_1 \neq 0 \qquad \frac{1}{2} \leq q < 1$$

or $q = (1.b_1 b_2 \cdots)_2 \qquad 1 \leq q < 2$


$\frac{1}{10} = (0,1)_{10} = ($


$\frac{1}{10} = (0.1)_{10} = (0.0\overline{0011})_2$

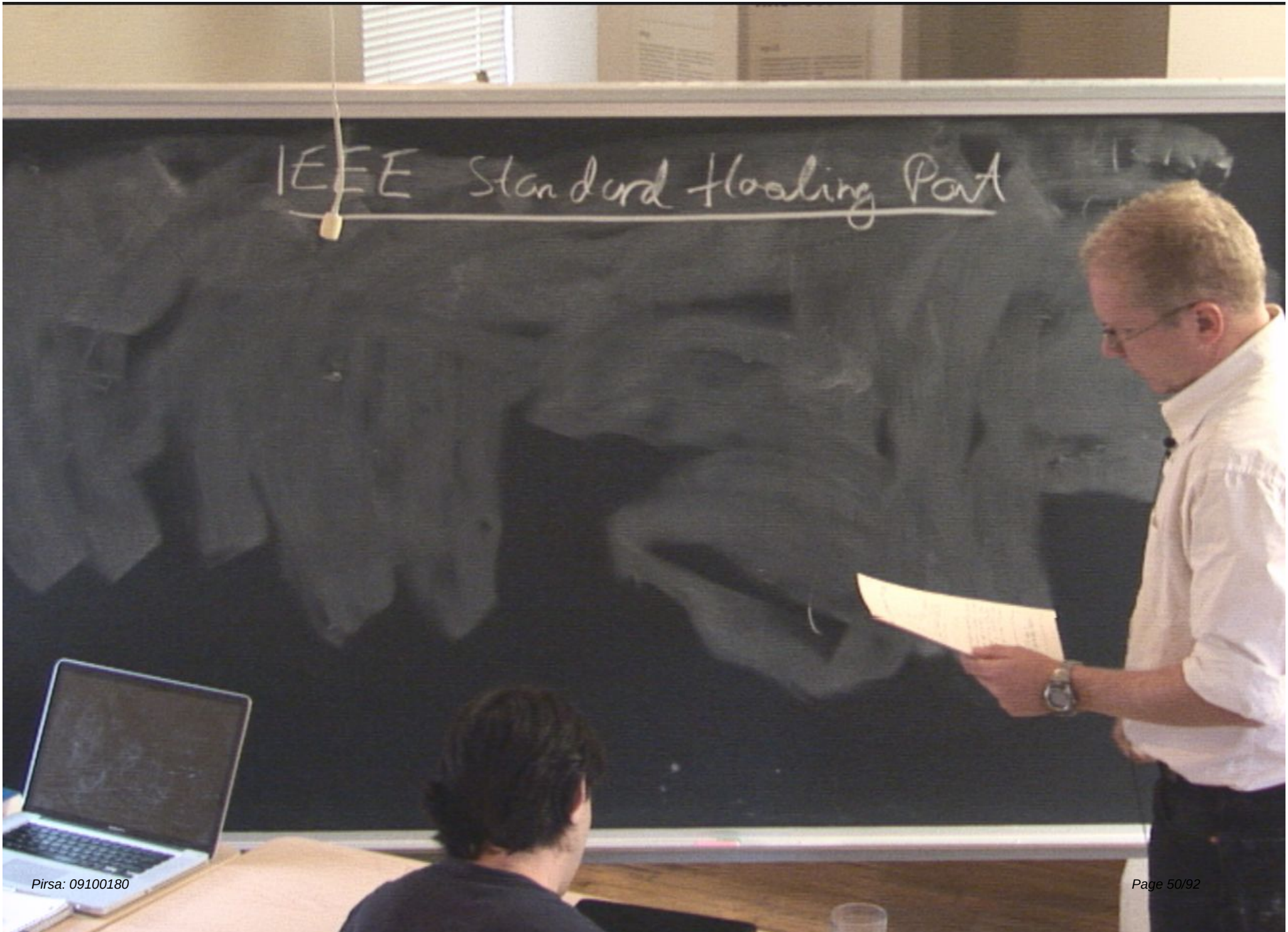

$\frac{1}{10} = (0.1)_{10} = (0.0\overline{6314})$
$= 10.$
-2

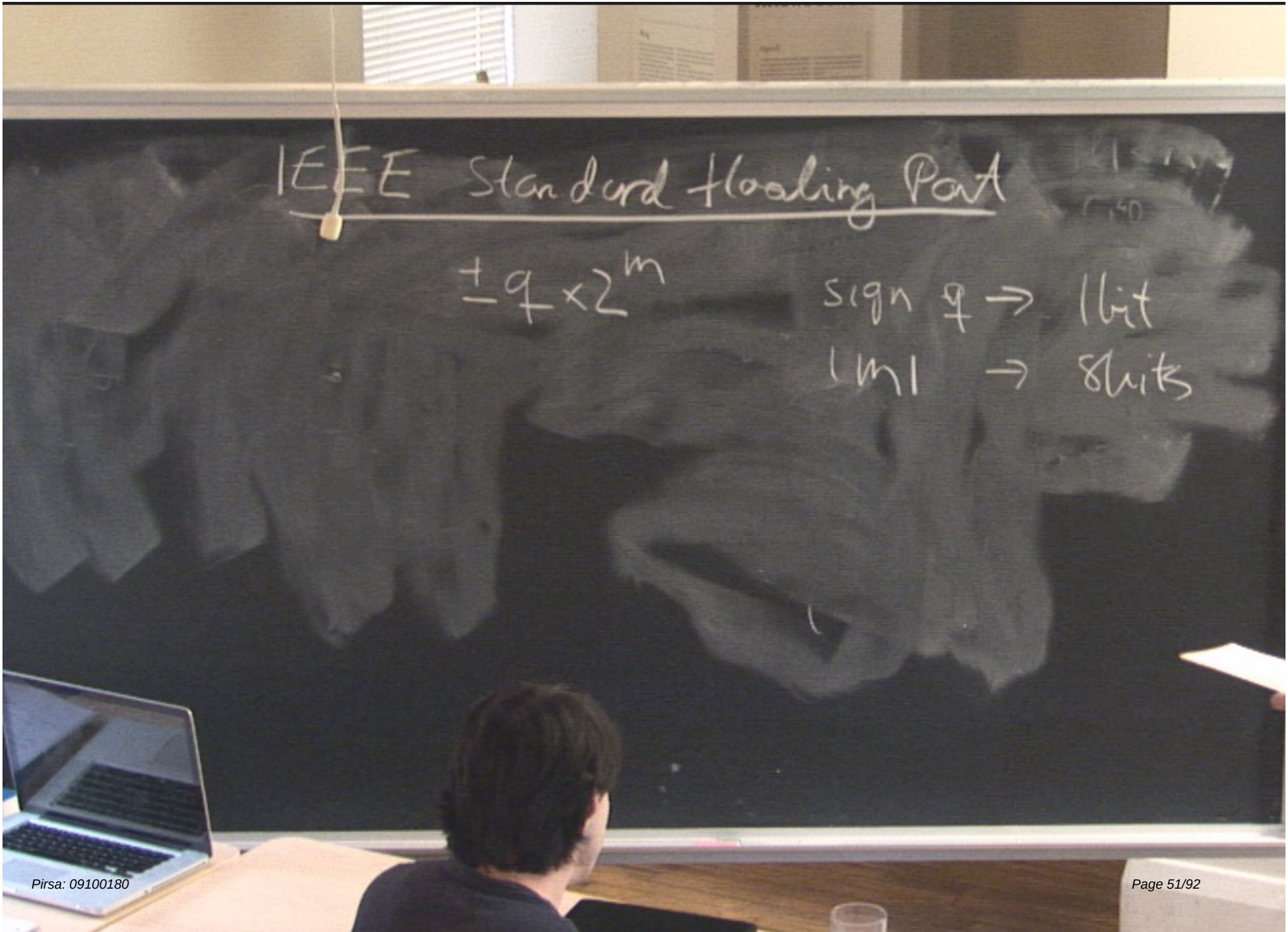

$\frac{1}{10} = (0.1)_{10} = (0.0\overline{6314}\ldots)_8$
$= (0.000\ 1100\ 1100\ 1100\ldots)_2$
Truncated

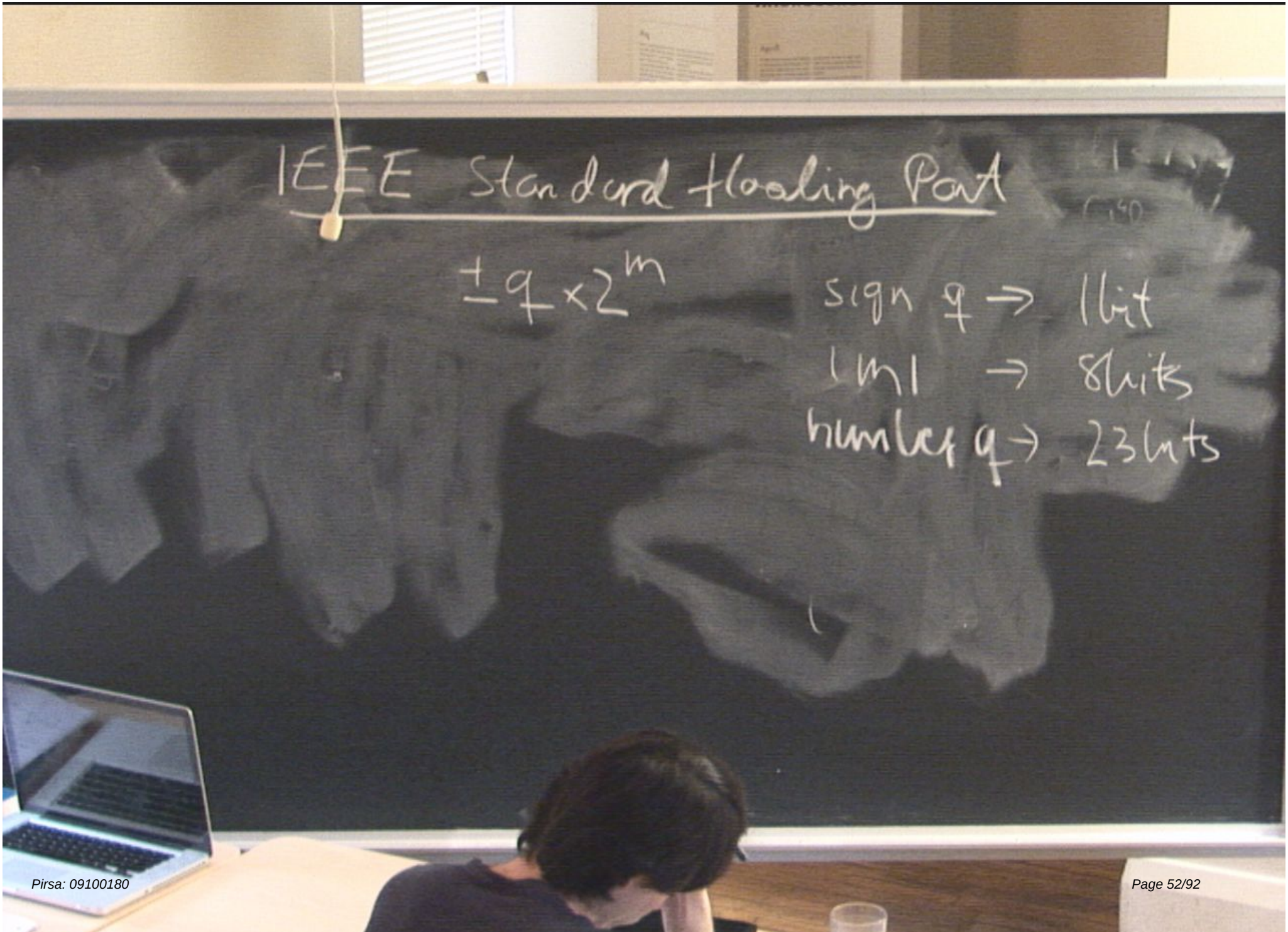
$$\frac{1}{10} = (0.1)_{10} = (0.0\overline{6314})_8$$

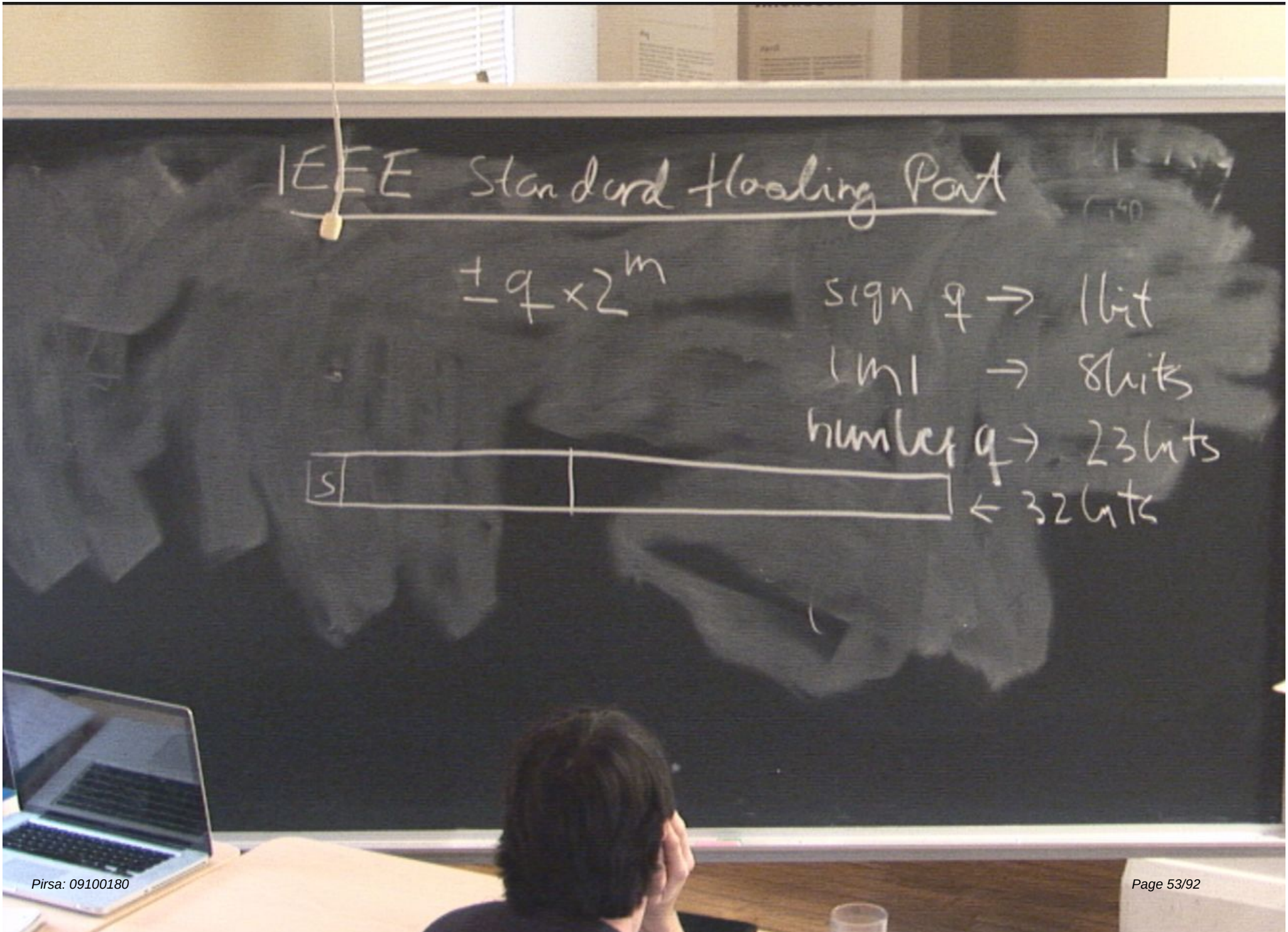
$$= (0.000\,\overline{1100\,1100\,1100})_2$$

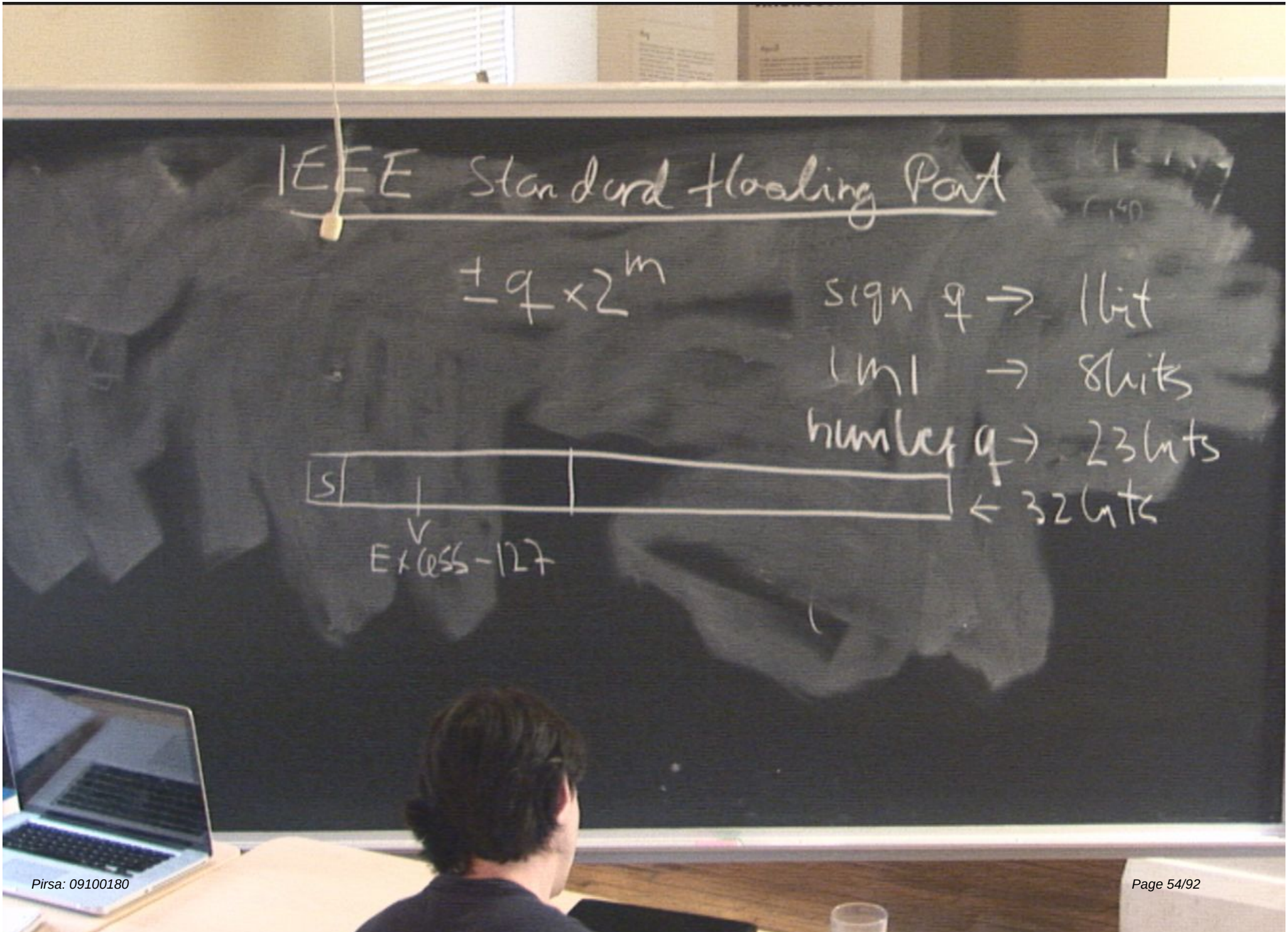
Truncated

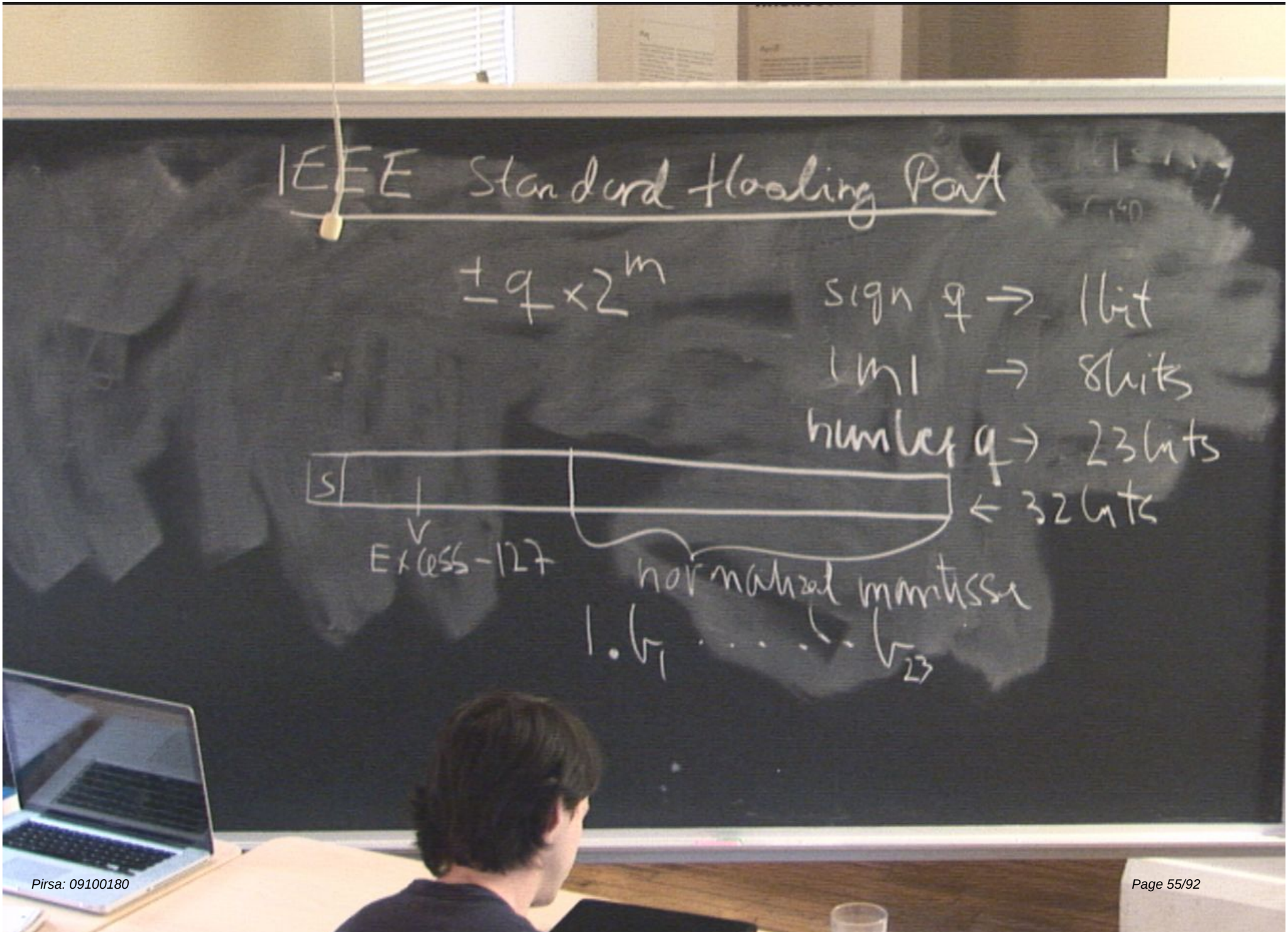

IEEE Standard Hooking Part

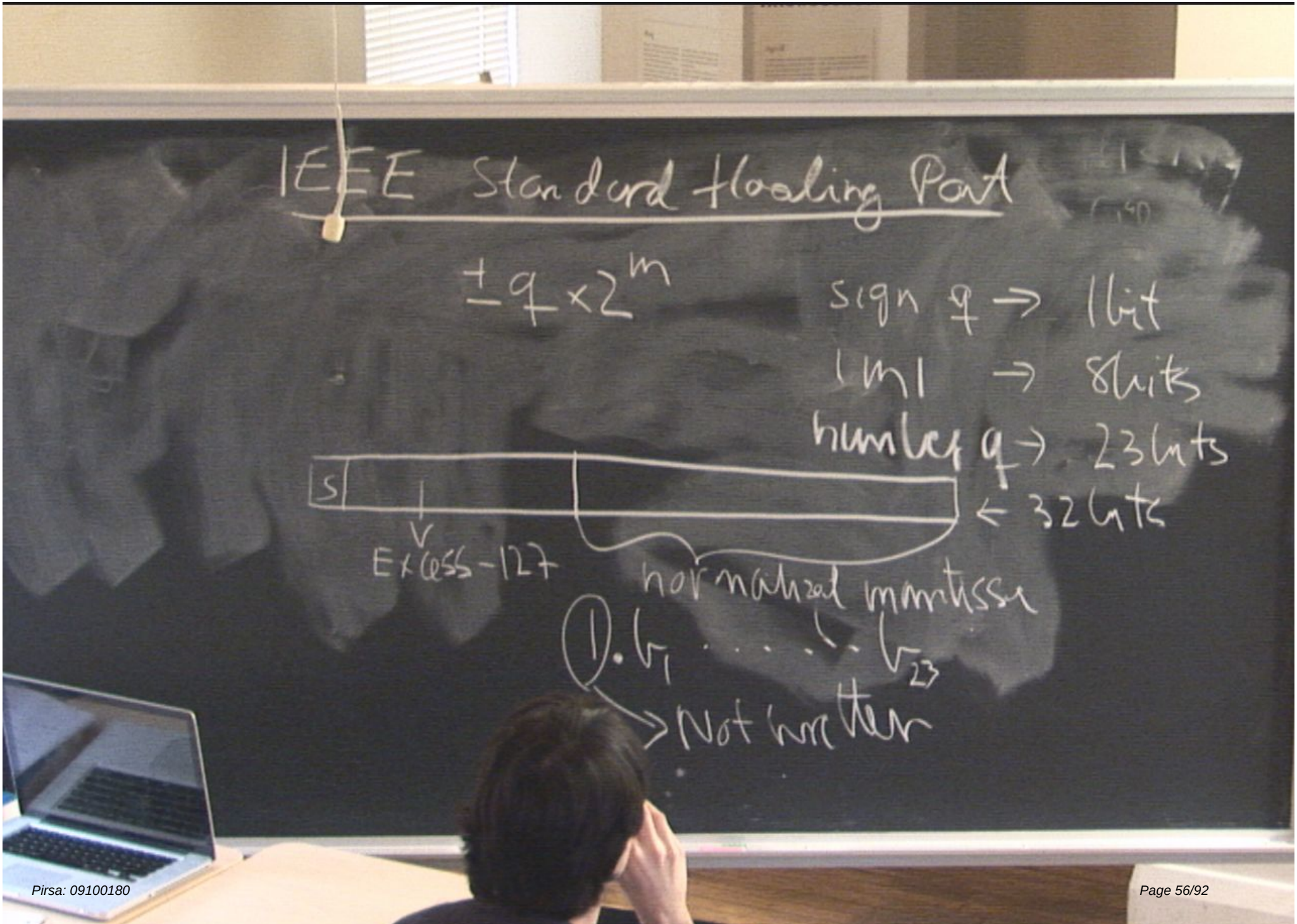

IEEE Standard floating Part
$\pm q \times 2^m$
sign q → 1bit
|m| → 8bits

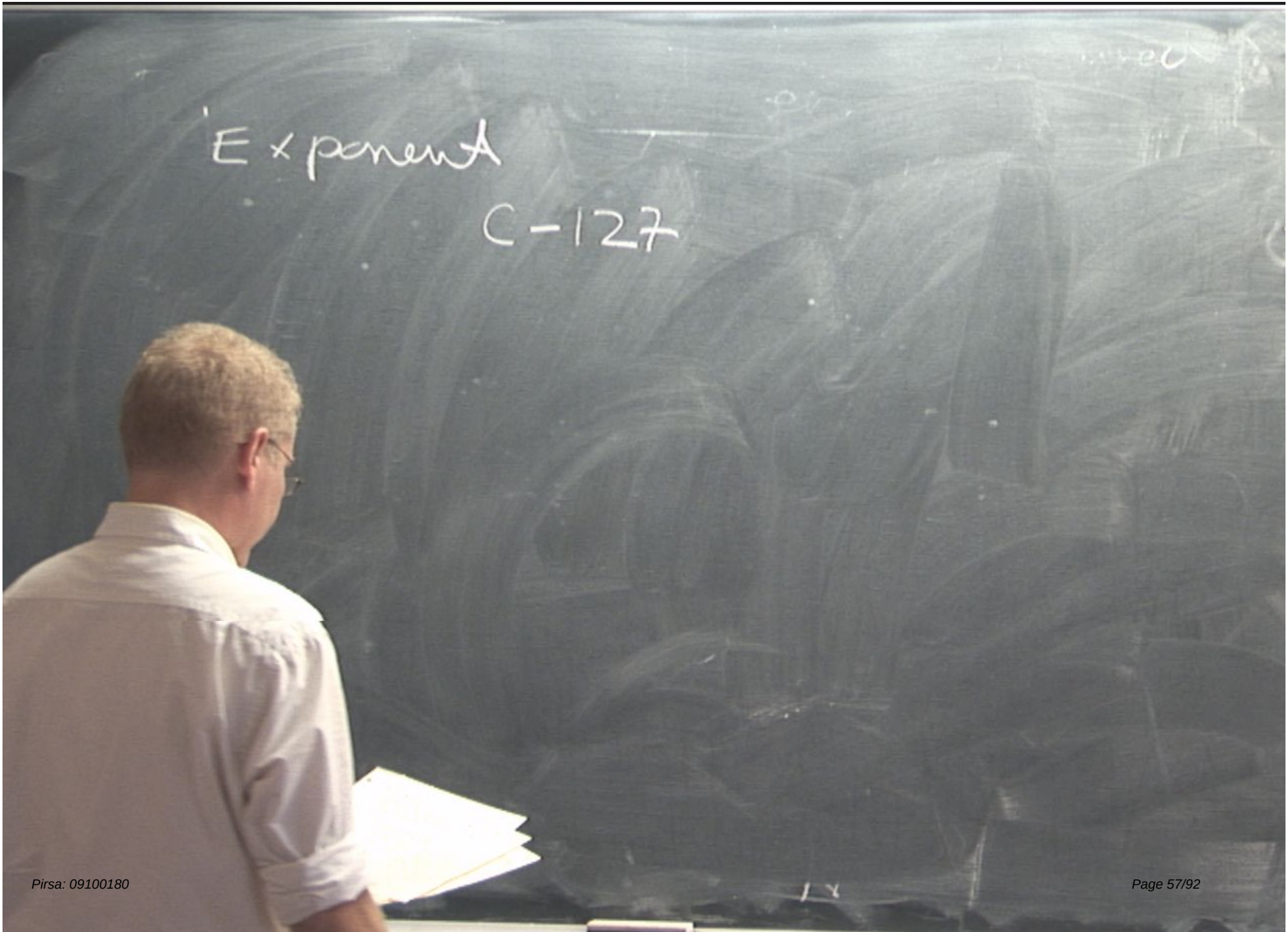

IEEE Standard Floating Point
$\pm q \times 2^m$
sign q → 1bit
|m| → 8bits
number q → 23bits

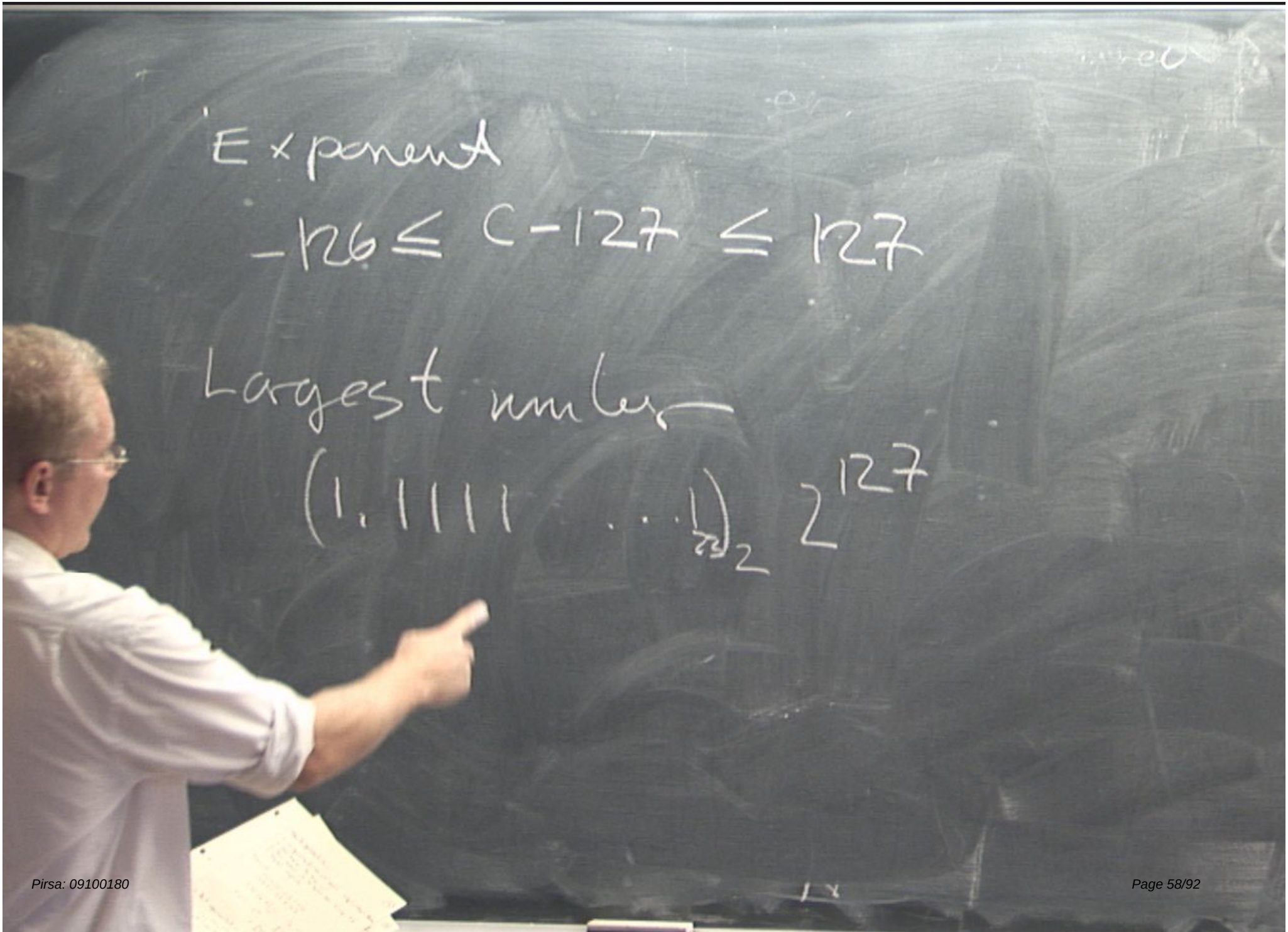

IEEE Standard Floating Point
$\pm q \times 2^m$
sign q → 1bit
|m| → 8bits
number q → 23bits
← 32bits
S


IEEE Standard floating Part
$\pm q \times 2^m$
sign q → 1bit
|m| → 8bits
number q → 23bits
← 32bits
S
Excess-127


IEEE Standard floating Part
$\pm q \times 2^m$
sign q → 1bit
|m| → 8bits
number q → 23bits
← 32bits
S
Excess-127
normalized mantissa
$1.b_1 \ldots b_{23}$


IEEE Standard floating Part
$\pm q \times 2^m$
sign q → 1bit
|m| → 8bits
number q → 23bits
← 32bits
s
Excess-127
normalized mantissa
$1.b_1 \ldots b_{23}$
Not hidden


Exponents
C-127


Exponent
$-126 \leq e - 127 \leq 127$
Largest number
$(1.11111\ldots1)_2\ 2^{127}$

Exponent

$-126 \leq e - 127 \leq 127$

Largest number

$(1.11111\ldots1)_2\ 2^{127} = (2 - 2^{-23}) \times 2^{127}$

$\simeq 2^{128}$

Exponent

$-126 \leq C - 127 \leq 127$

Largest number

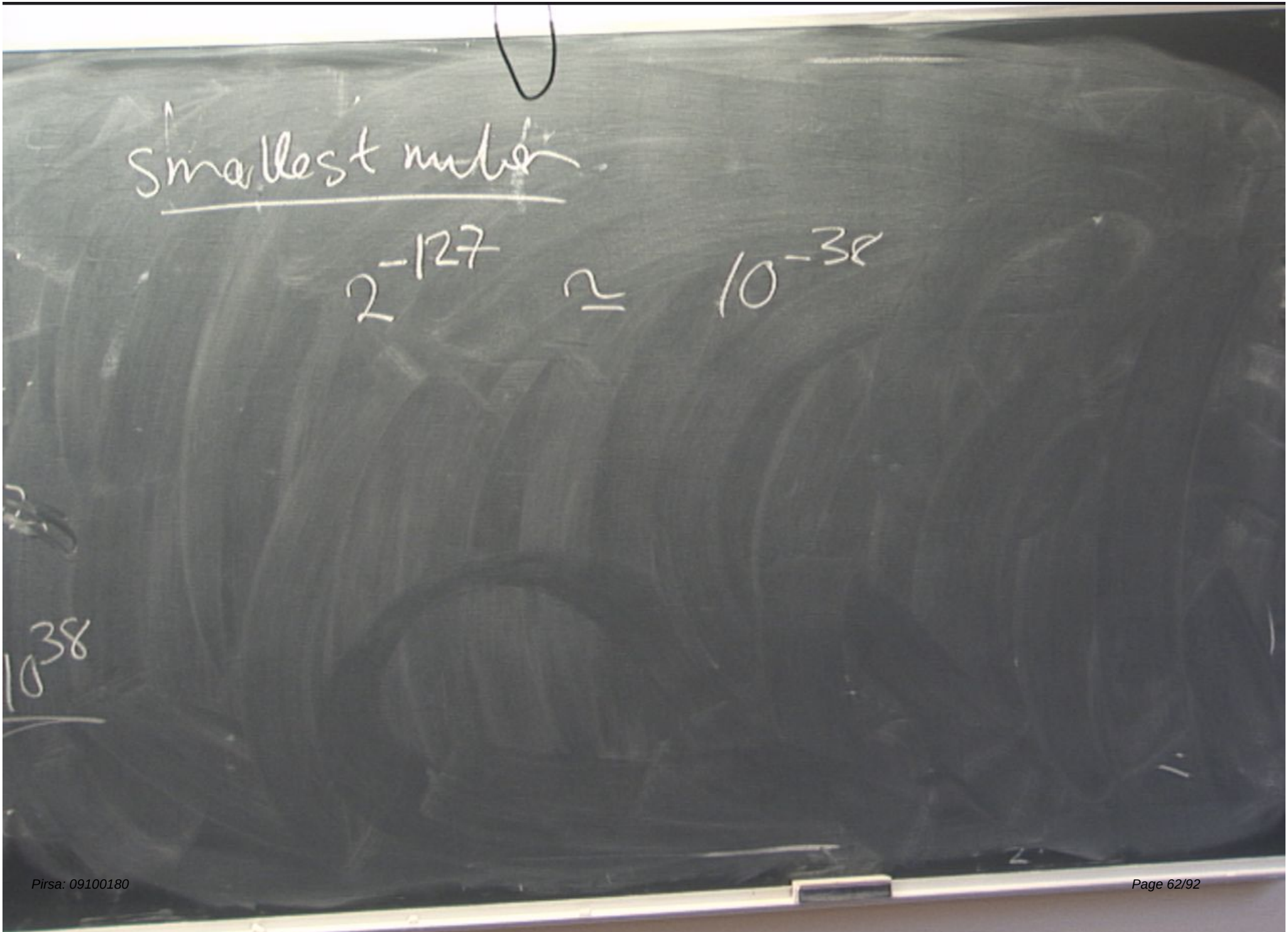
$(1.1111\ldots)_2\, 2^{127} = (2 - 2^{-23}) \times 2^{127} = 2^{128} \simeq 3.4 \times 10^{38}$

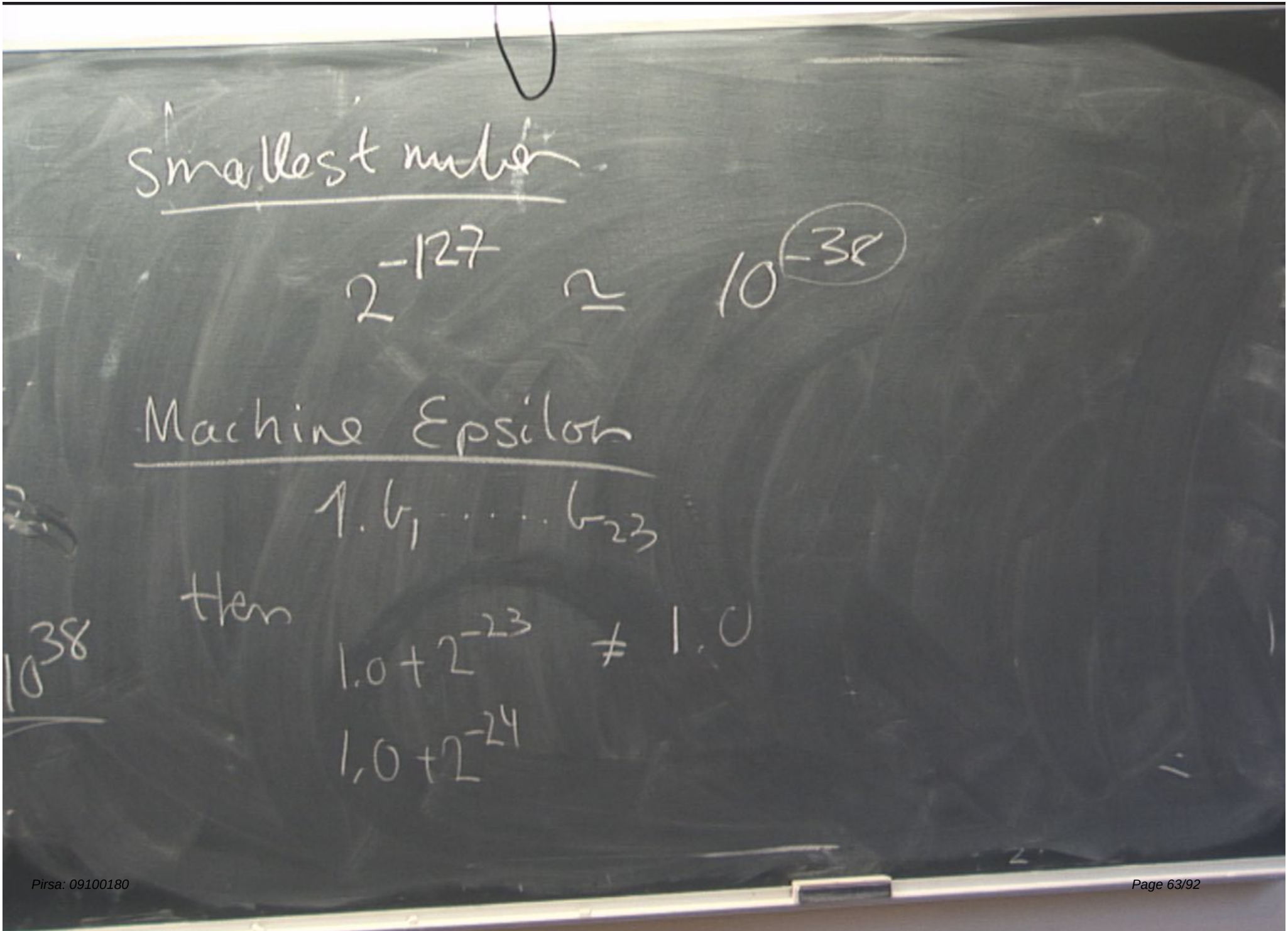
Exponent

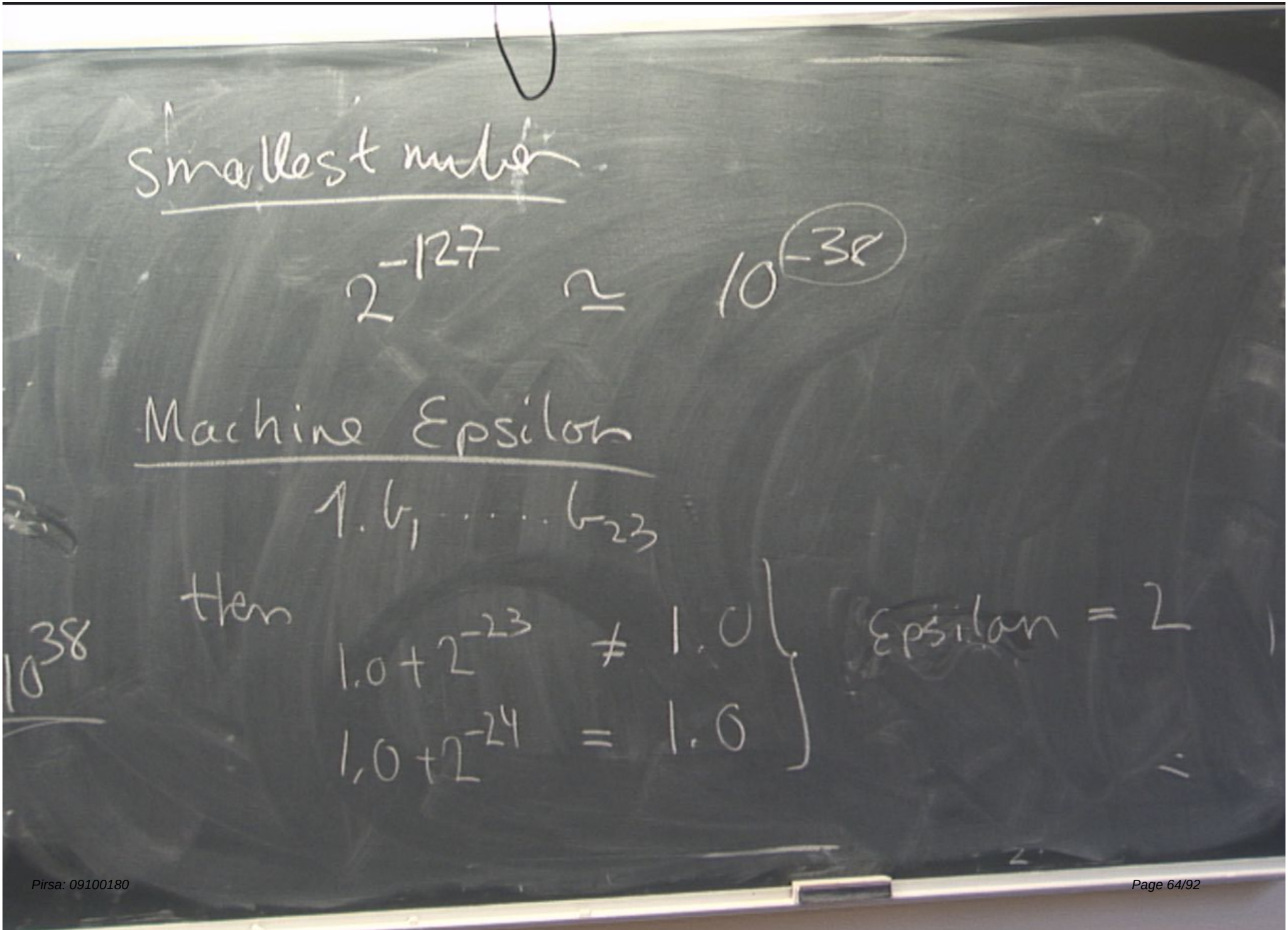
$-127 \leq C - 127 \leq 128$

Largest number

$(1.11111\ldots1)_2\, 2^{127} = (2 - 2^{-23}) \times 2^{127} \simeq 2^{128} \simeq 3{,}4 \times 10^{38}$


Smallest number
$2^{-127} \simeq 10^{-38}$
10^{38}


Smallest number
$2^{-127} \simeq 10^{-38}$
Machine Epsilon
$1.b_1 \ldots b_{23}$
then
$1.0 + 2^{-23} \neq 1.0$
$1.0 + 2^{-24}$
10^{38}


Smallest number
$2^{-127} \simeq 10^{-38}$
Machine Epsilon
$1.b_1 \ldots b_{23}$
then
$1.0 + 2^{-23} \neq 1.0$
$1.0 + 2^{-24} = 1.0$
Epsilon = 2
10^{38}

Smallest number

$2^{-127} \simeq 10^{-38}$

Machine Epsilon

$1.b_1 \ldots b_{23}$

then

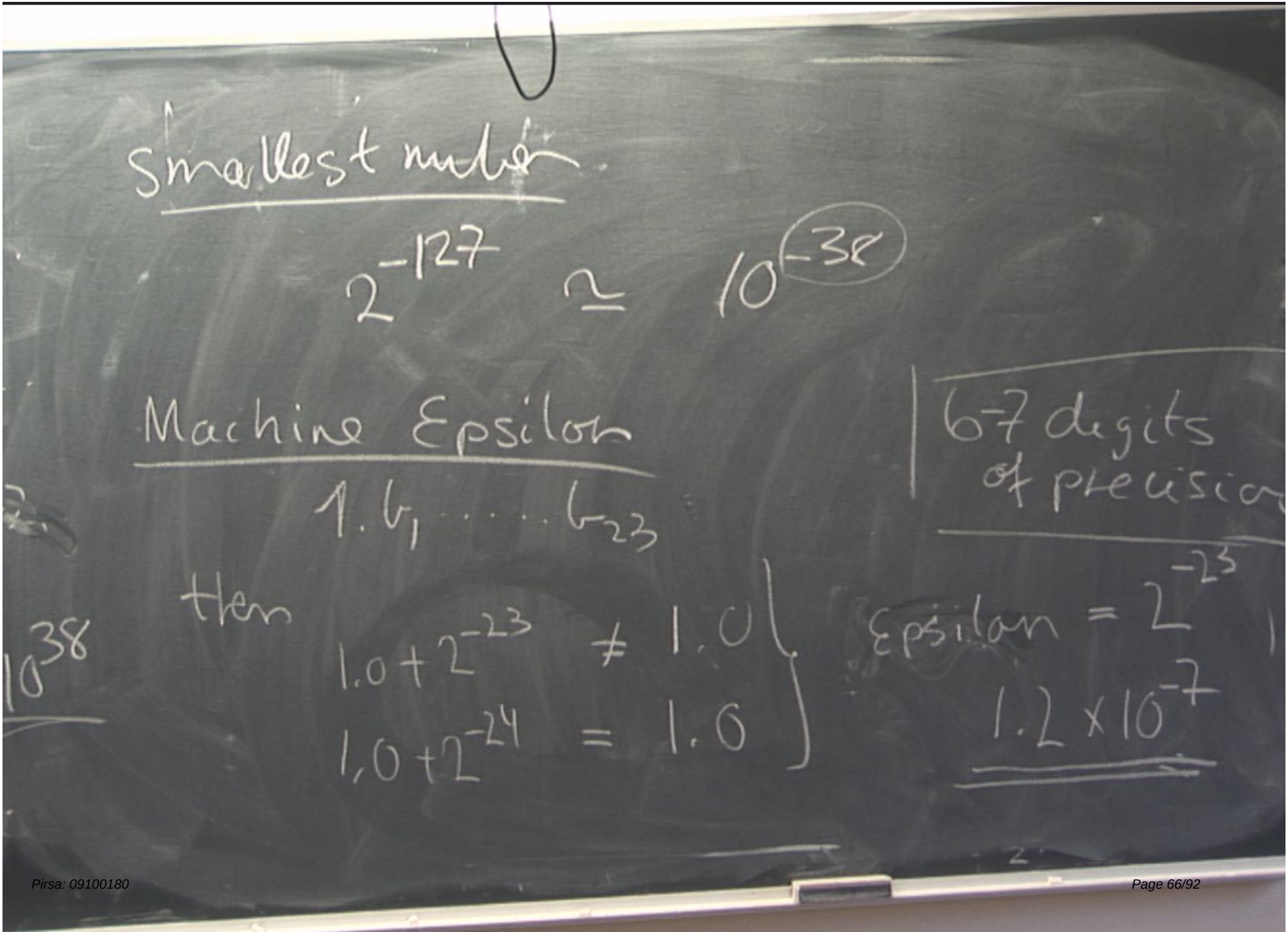
$1.0 + 2^{-23} \neq 1.0$

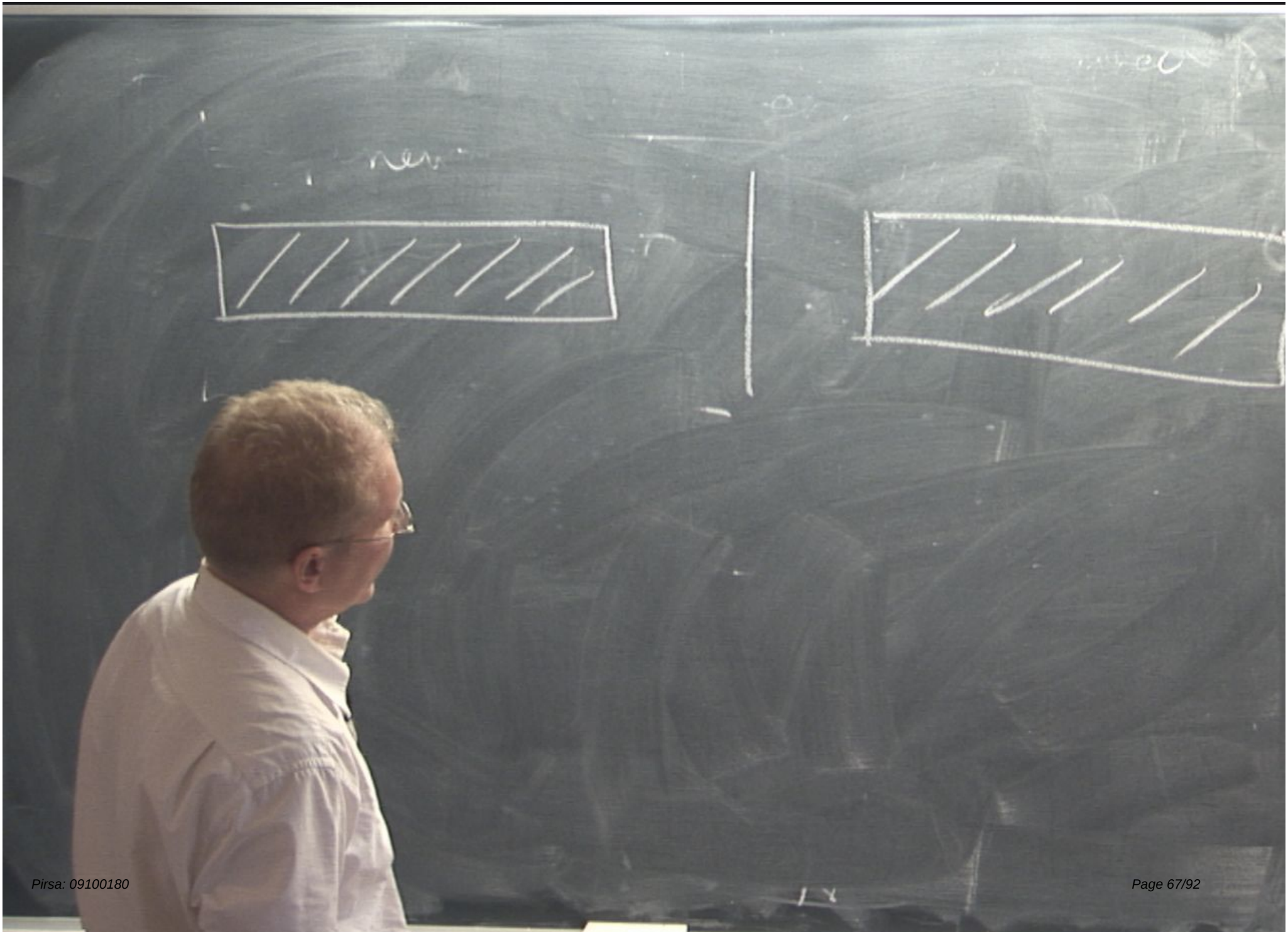
$1.0 + 2^{-24} = 1.0$

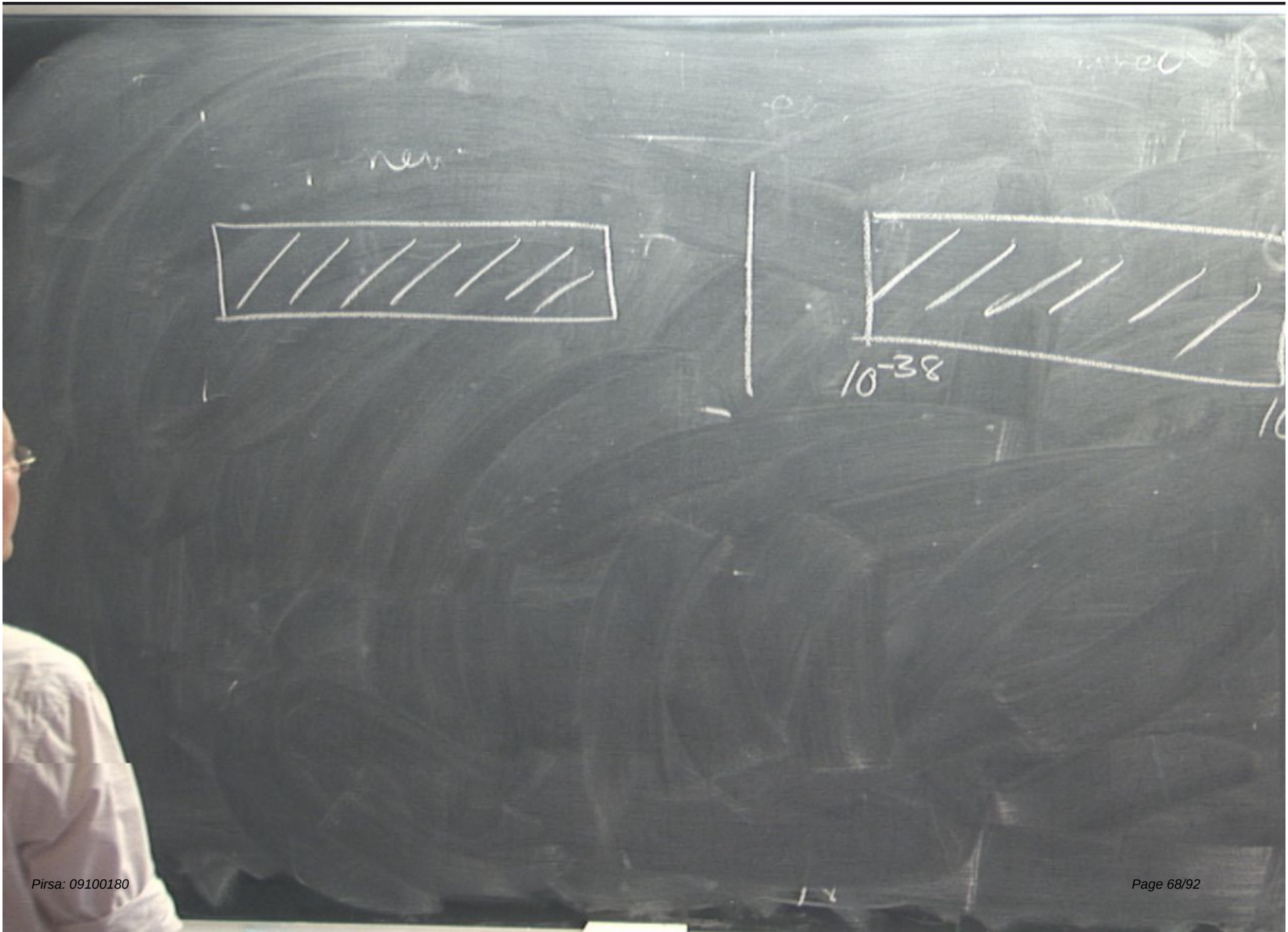
Epsilon $= 2^{-23}$

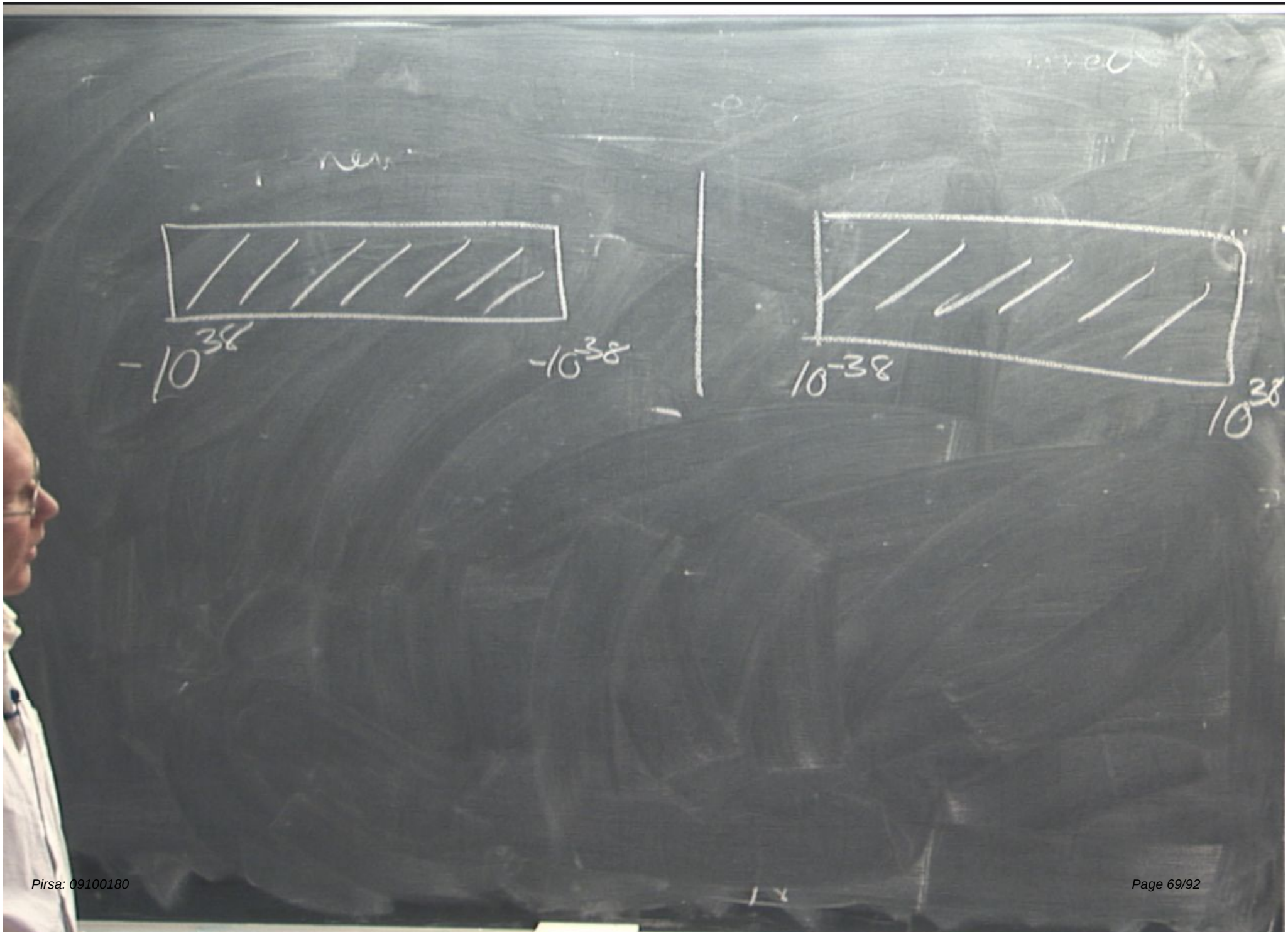
1.2×10^{-7}

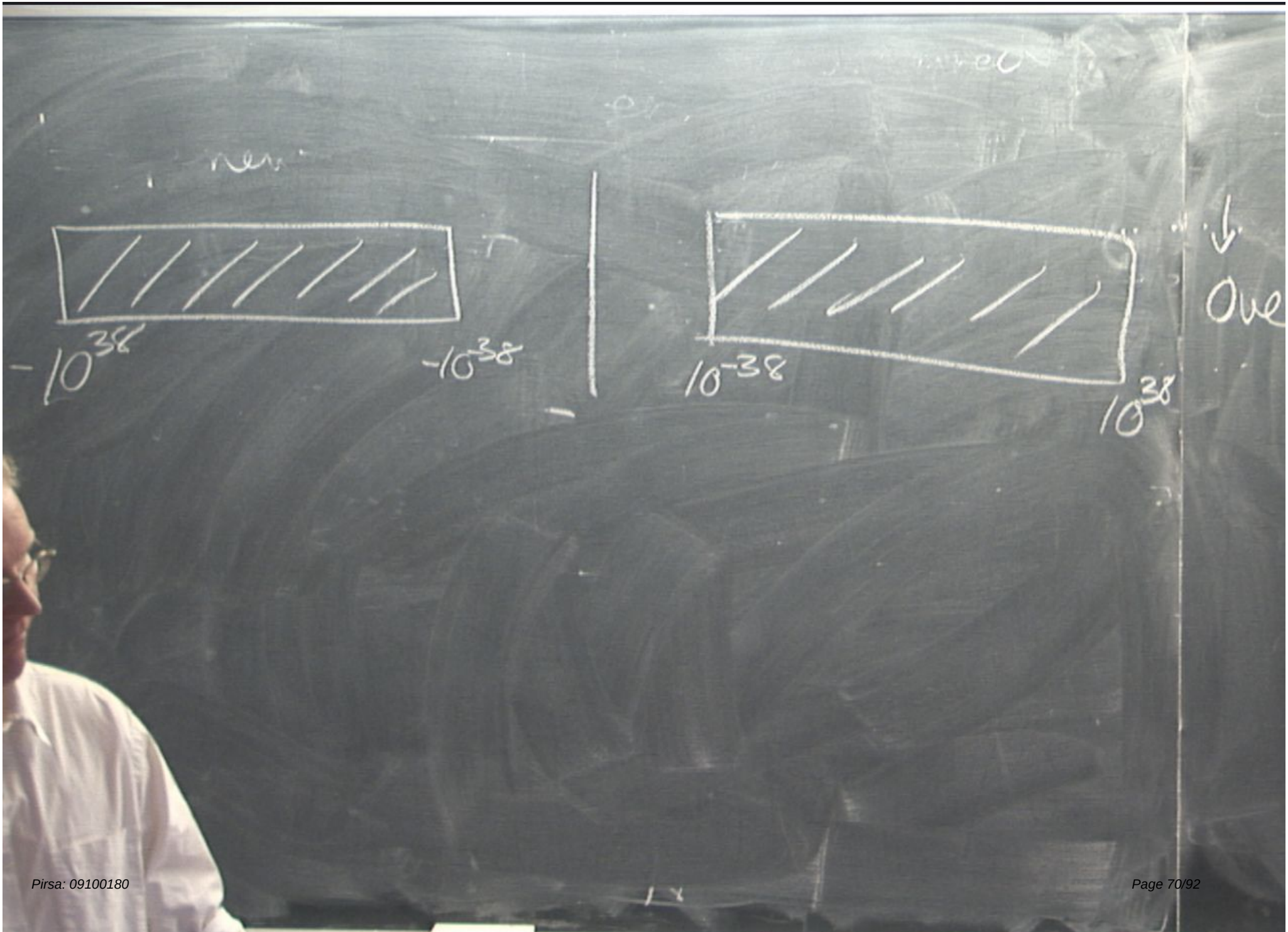
10^{38}

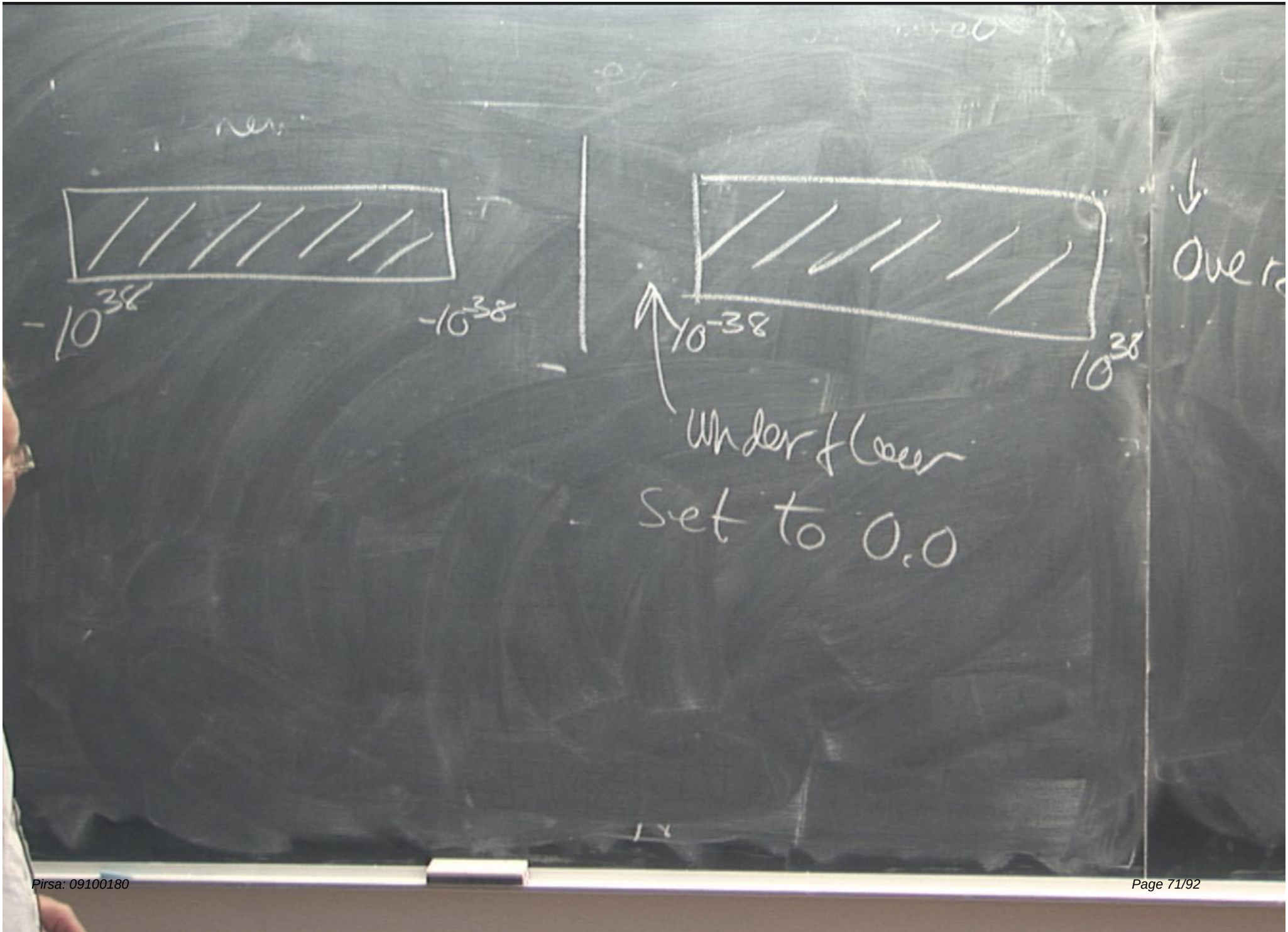

Smallest number
$2^{-127} \simeq 10^{-38}$
Machine Epsilon
$1.b_1 \ldots b_{23}$
then
$1.0 + 2^{-23} \neq 1.0$
$1.0 + 2^{-24} = 1.0$
Epsilon $= 2^{-23}$
1.2×10^{-7}
6-7 digits of precision
10^{38}

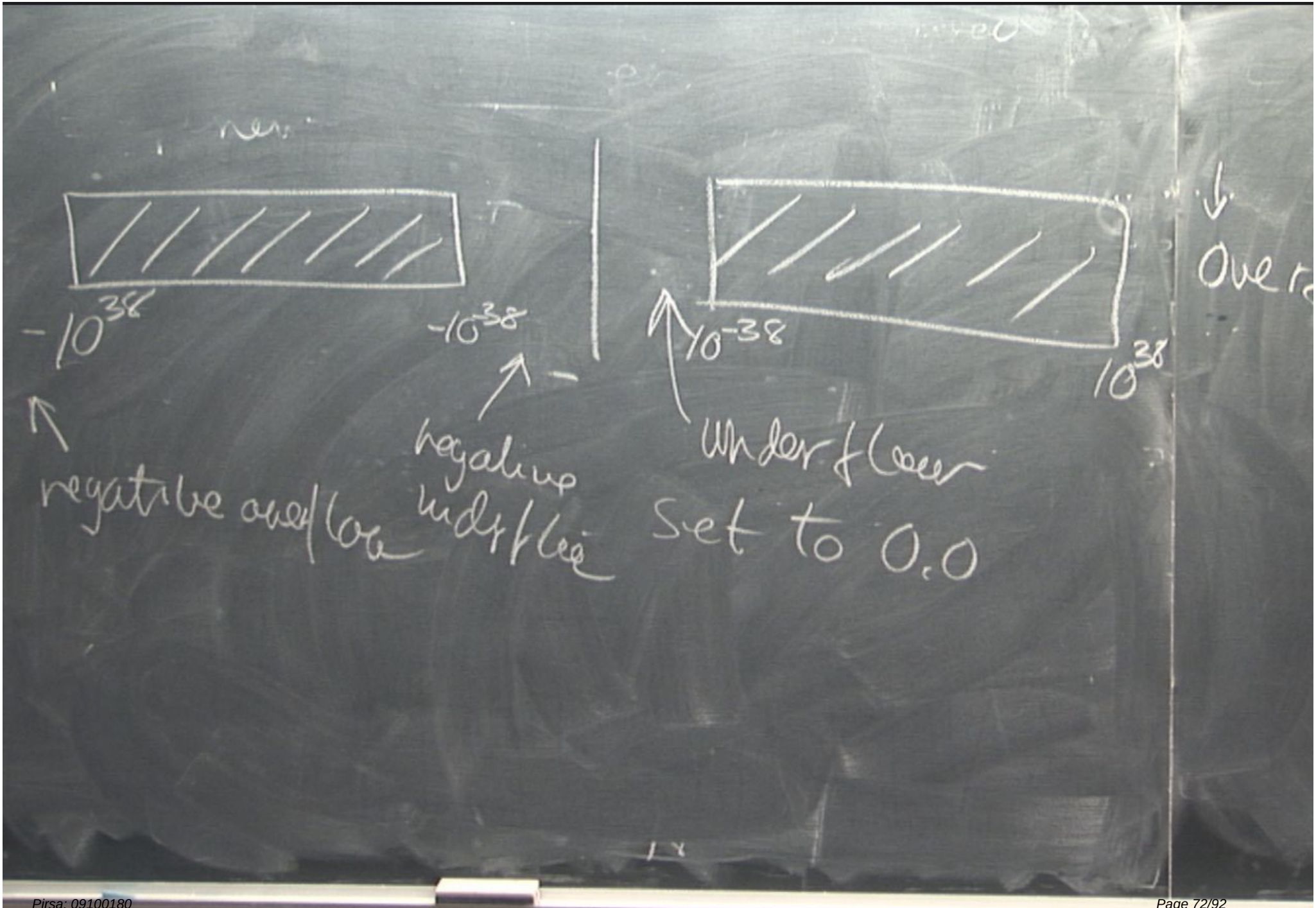


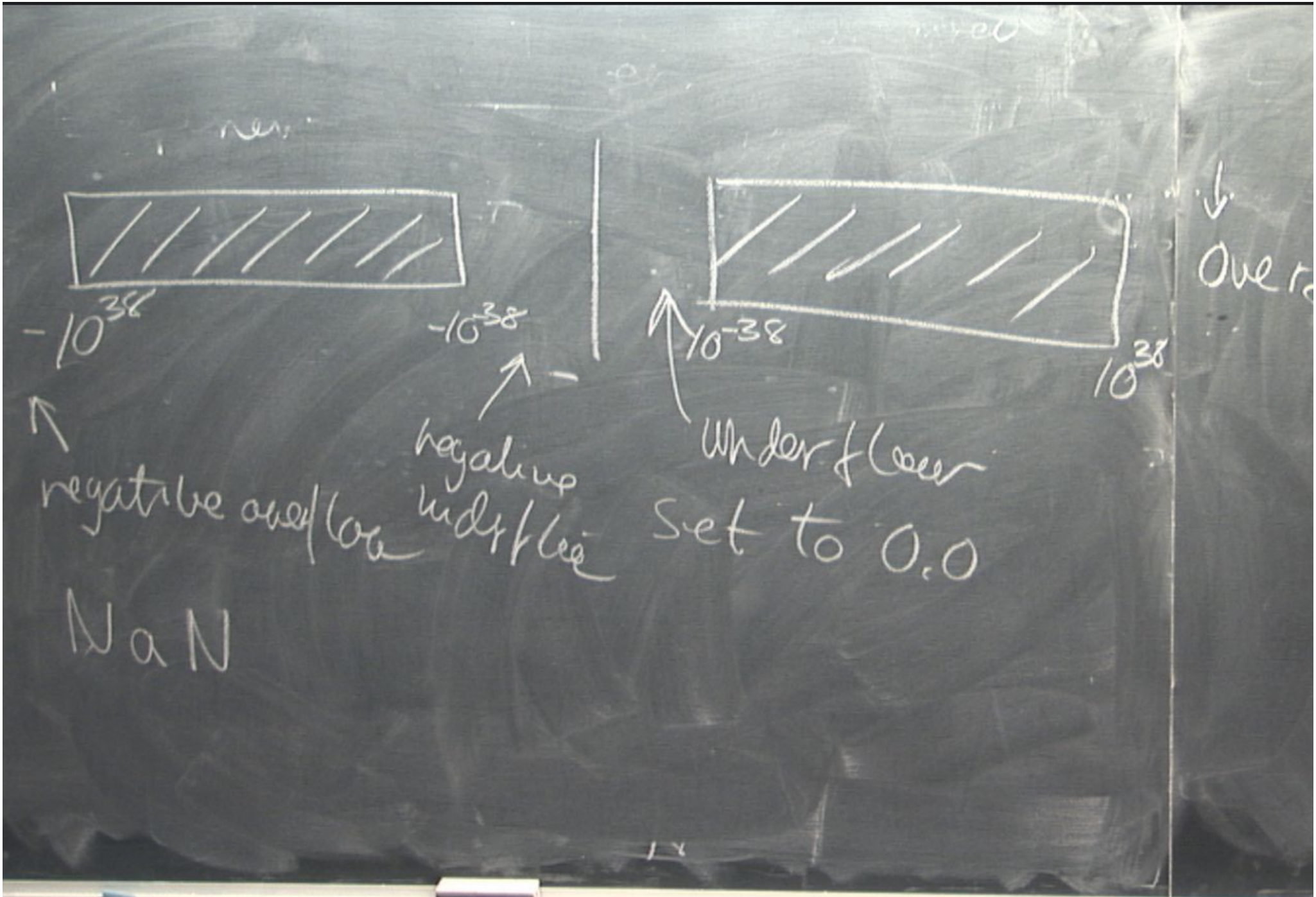

10^{-38}

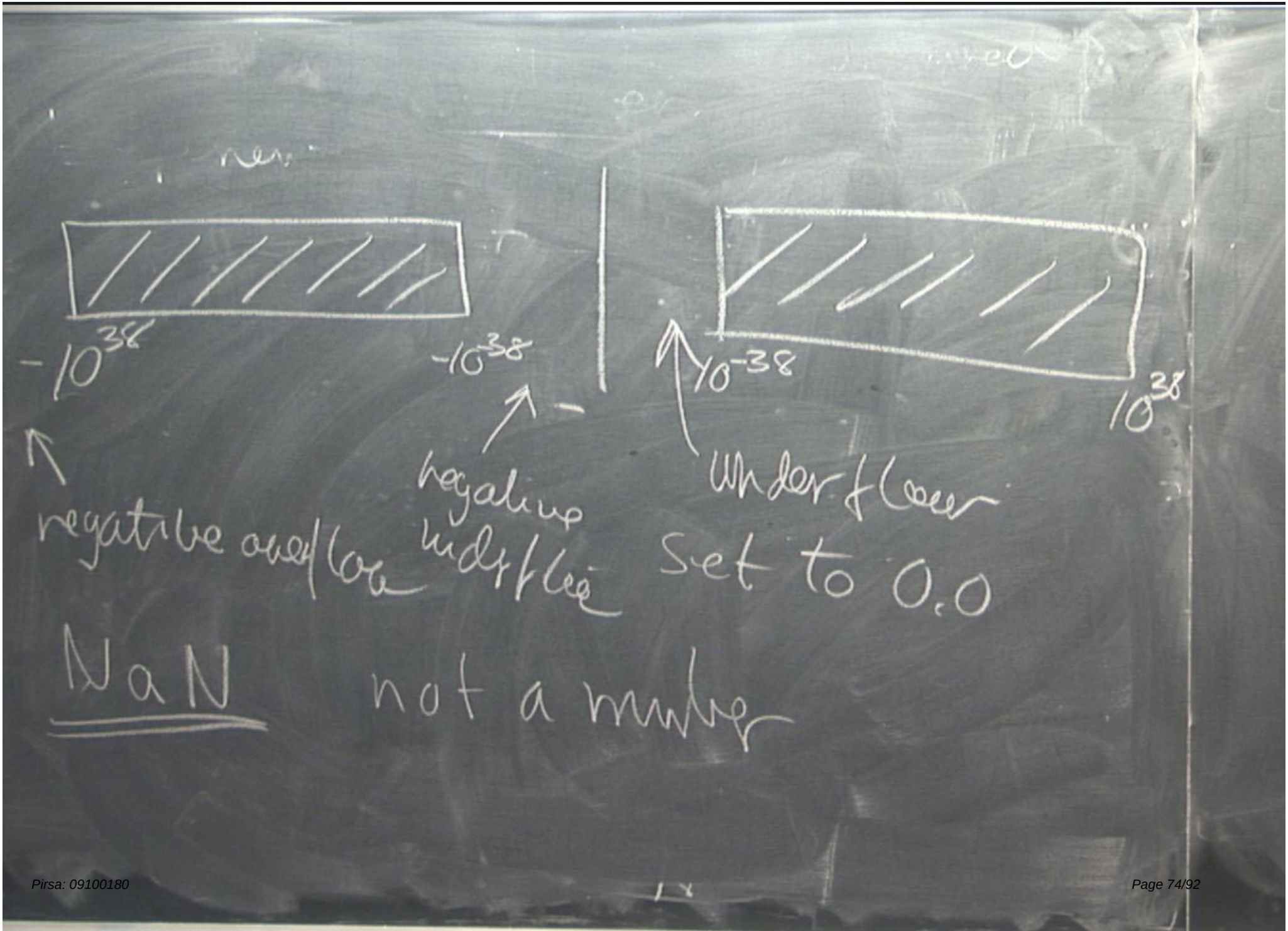

-10^{38}
-10^{-38}
10^{-38}
10^{38}

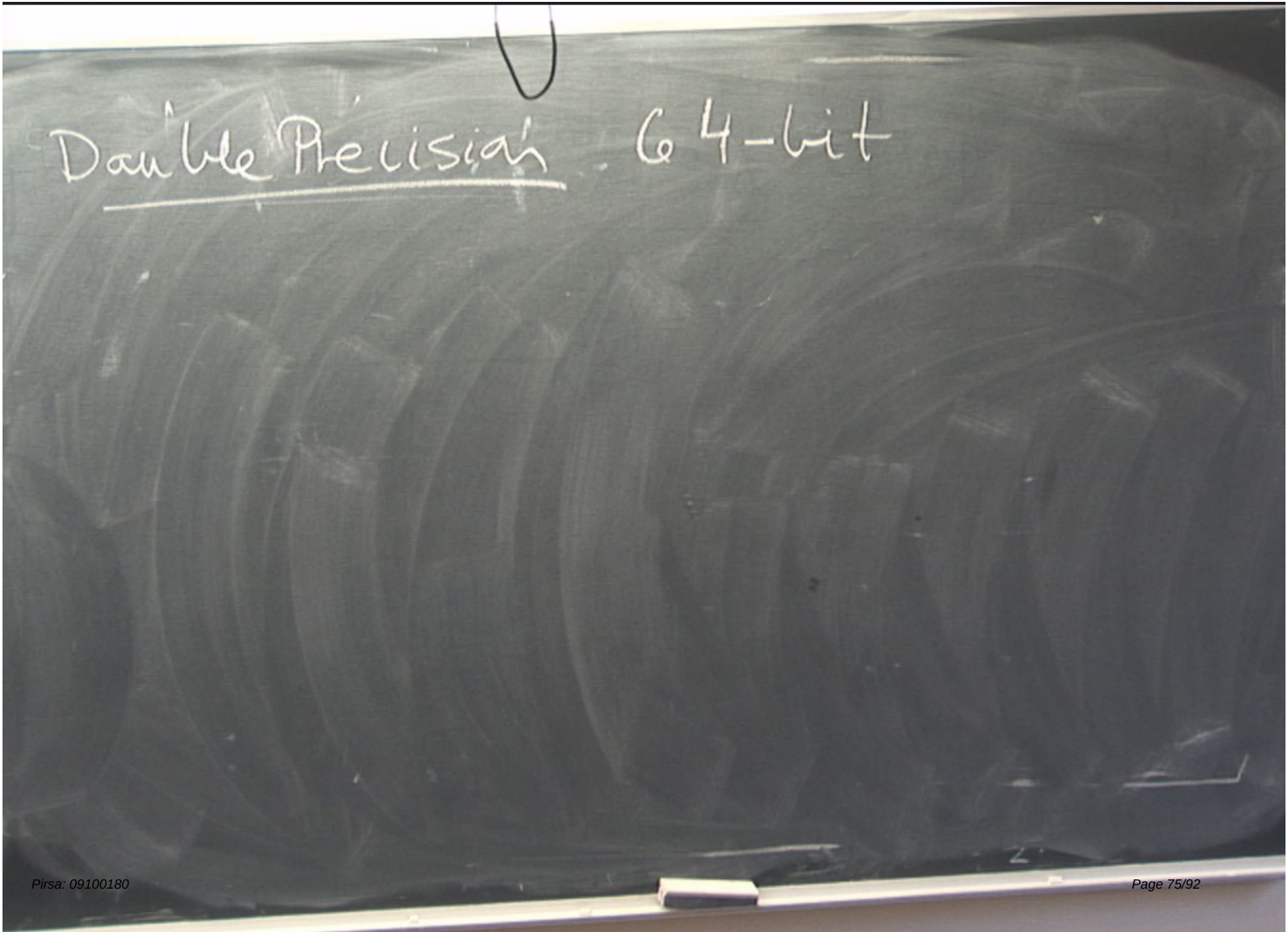

-10^{38}
-10^{38}
10^{-38}
10^{38}


-10^{38}
-10^{-38}
10^{-38}
10^{38}
Under flow
Set to 0.0


-10^{38}
-10^{-38}
10^{-38}
10^{38}
negative overflow
negative underflow
Under flow
Set to 0.0


-10^{38}
-10^{-38}
10^{-38}
10^{38}
negative overflow
negative underflow
underflow
Set to 0.0
NaN


-10^{38}
-10^{-38}
10^{-38}
10^{38}
negative overflow
negative underflow
positive underflow
Set to 0.0
NaN
not a number


Double Precision . 64-bit

Double Precision: 64-bit

- 52 bits for the mantissa
- 11 bits for the exponent

$-1022 \quad \text{---} \quad 1023$

Double Precision: 64-bit

- 52 bits for the mantissa
- 11 bits for the exponent

$-1022 \quad \text{---} \quad 1023$

$\varepsilon = 2.2 \times 10^{-16}$

Double Precision: 64-bit

- 52 bits for the mantissa
- 11 bits for the exponent

$-1022 \quad \text{---} \quad 1023$

$\varepsilon = 2.2 \times 10^{-16}$

Largest number 1.8×10^{308}

Double Precision: 64-bit

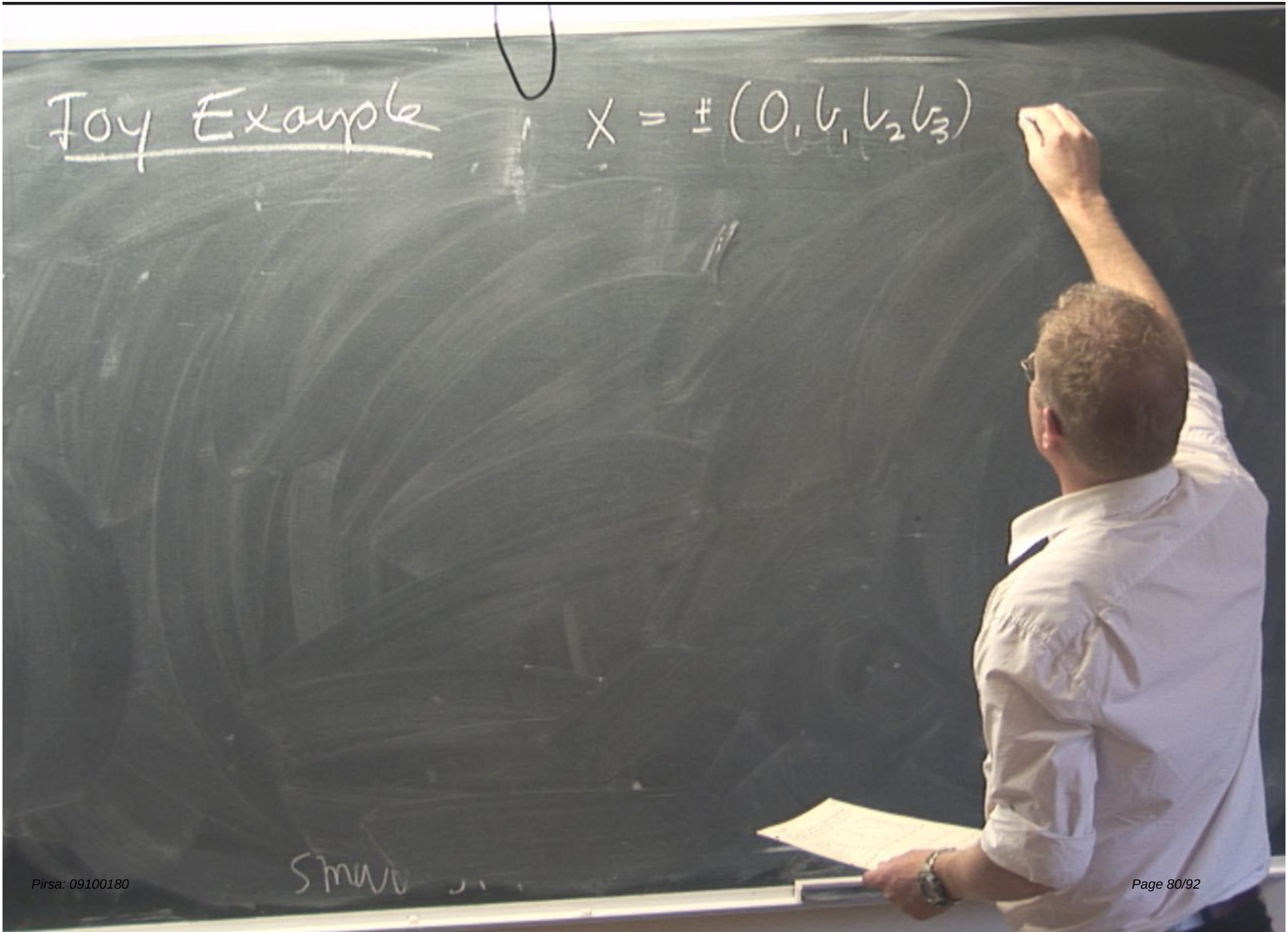
- 52 bits for the mantissa
- 11 bits for the exponent

$-1022 \quad \text{---} \quad 1023$

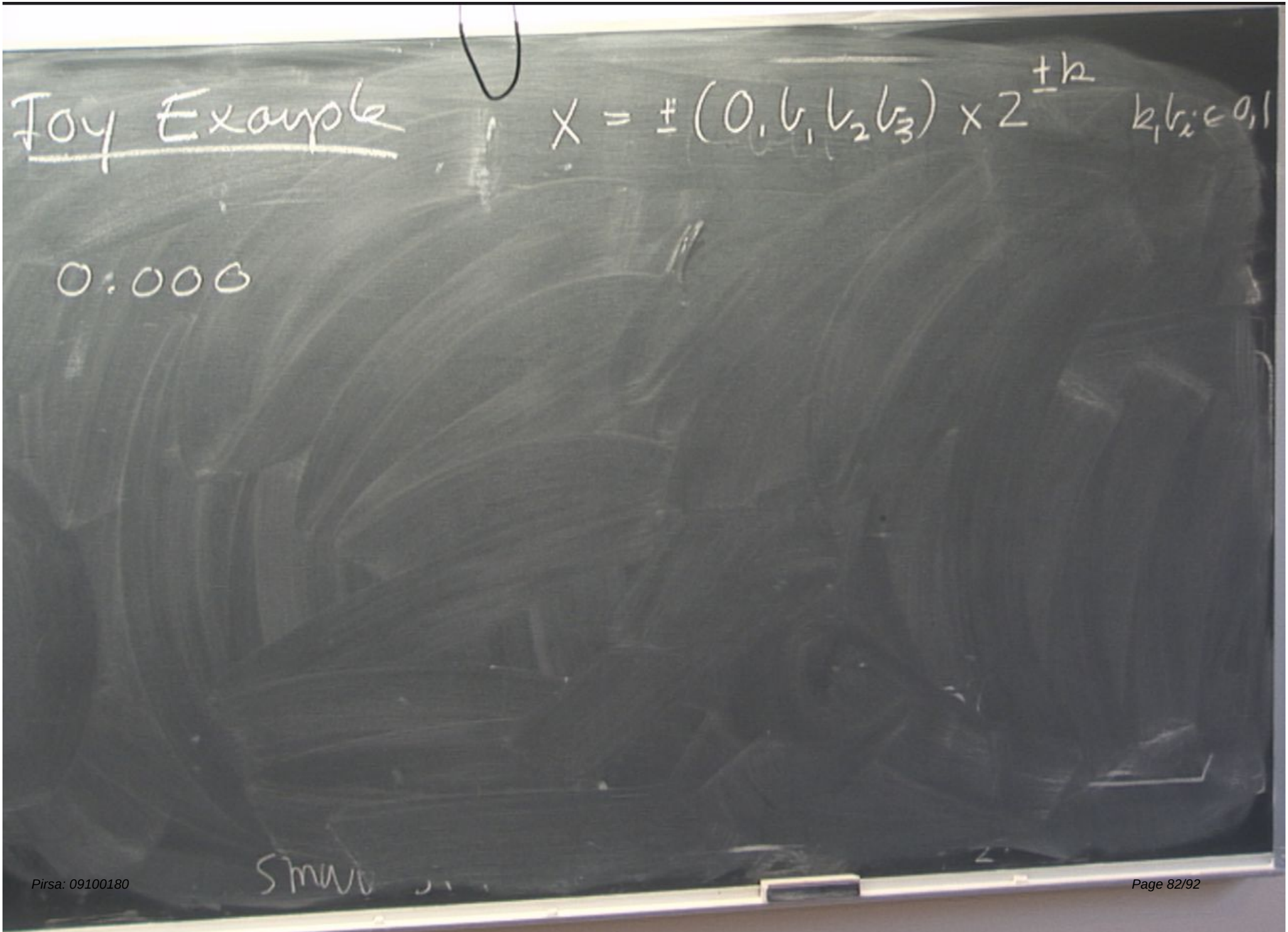
$\varepsilon = 2.2 \times 10^{-16}$

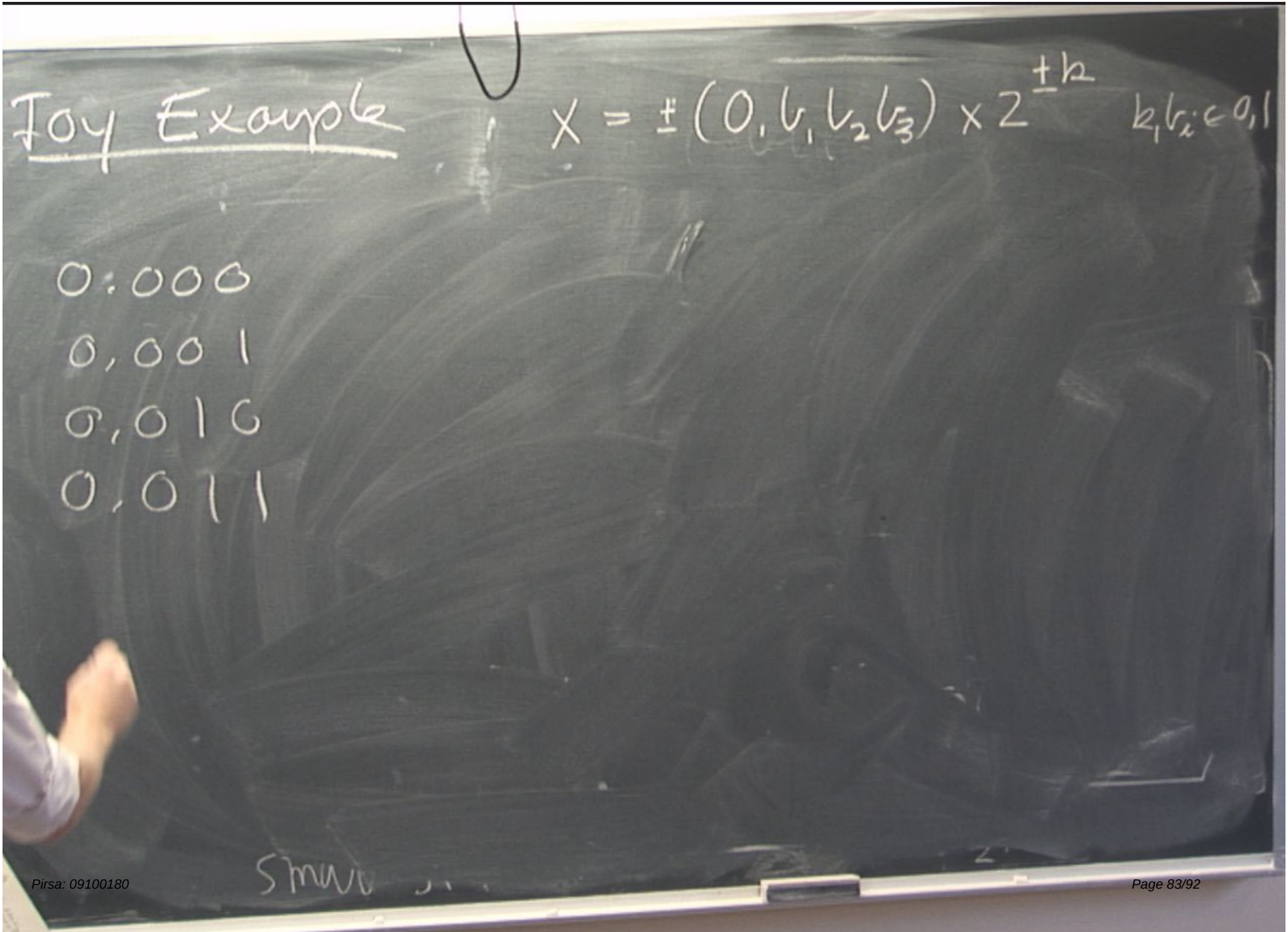
Largest number: 1.8×10^{308}

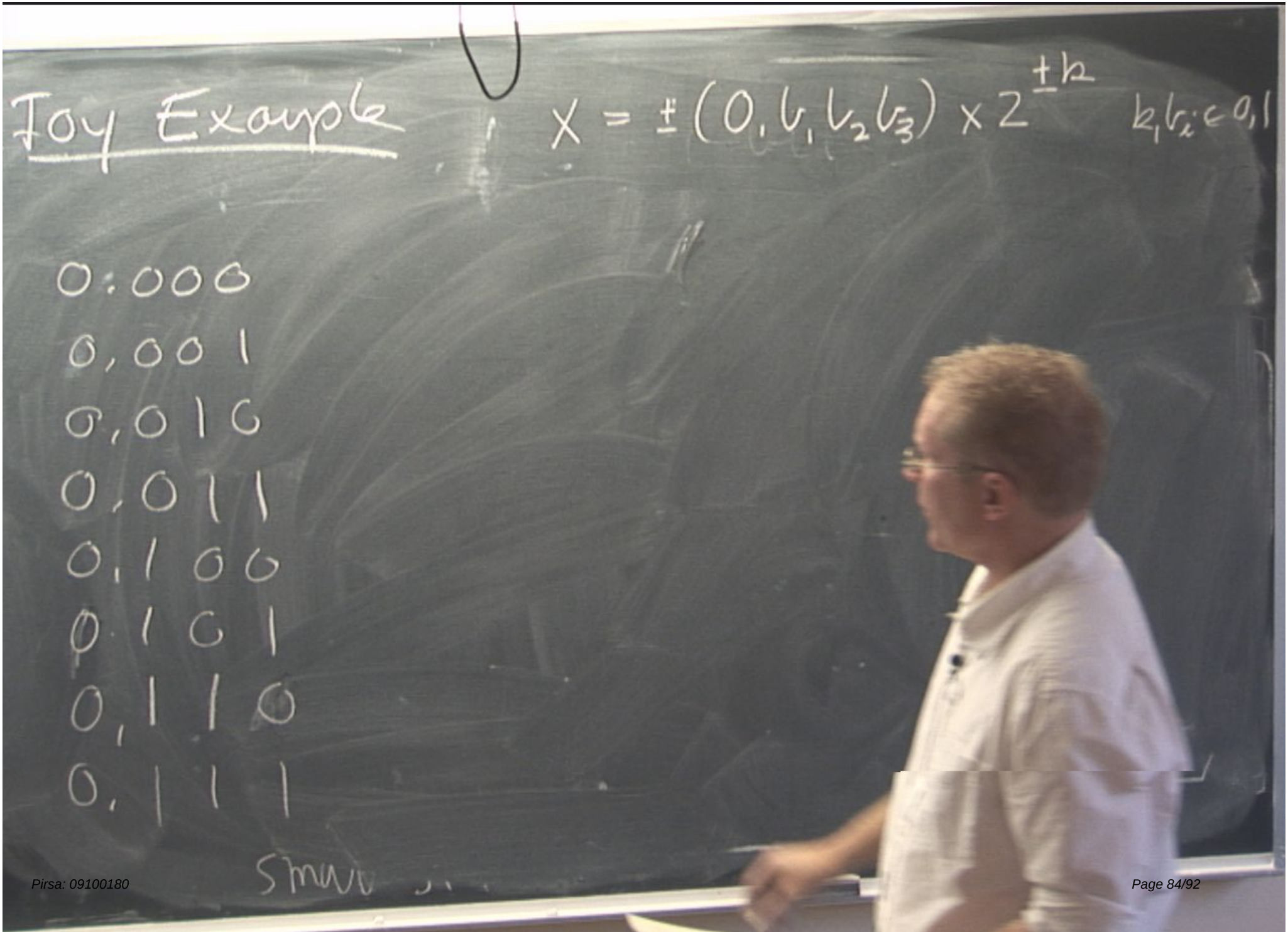
Smallest number: 10^{-308}

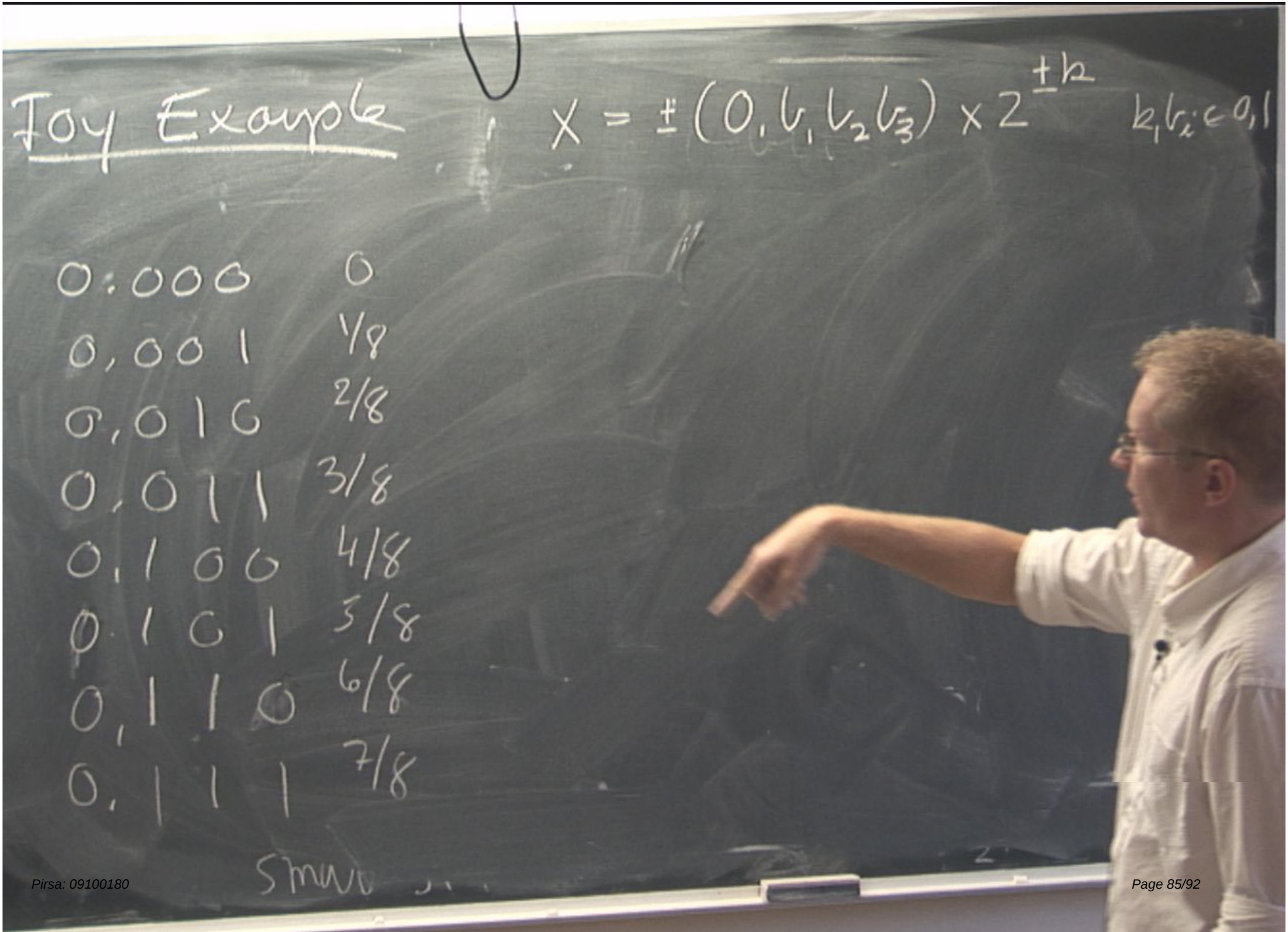

Toy Example
$X = \pm(0, v_1, v_2, v_3)$

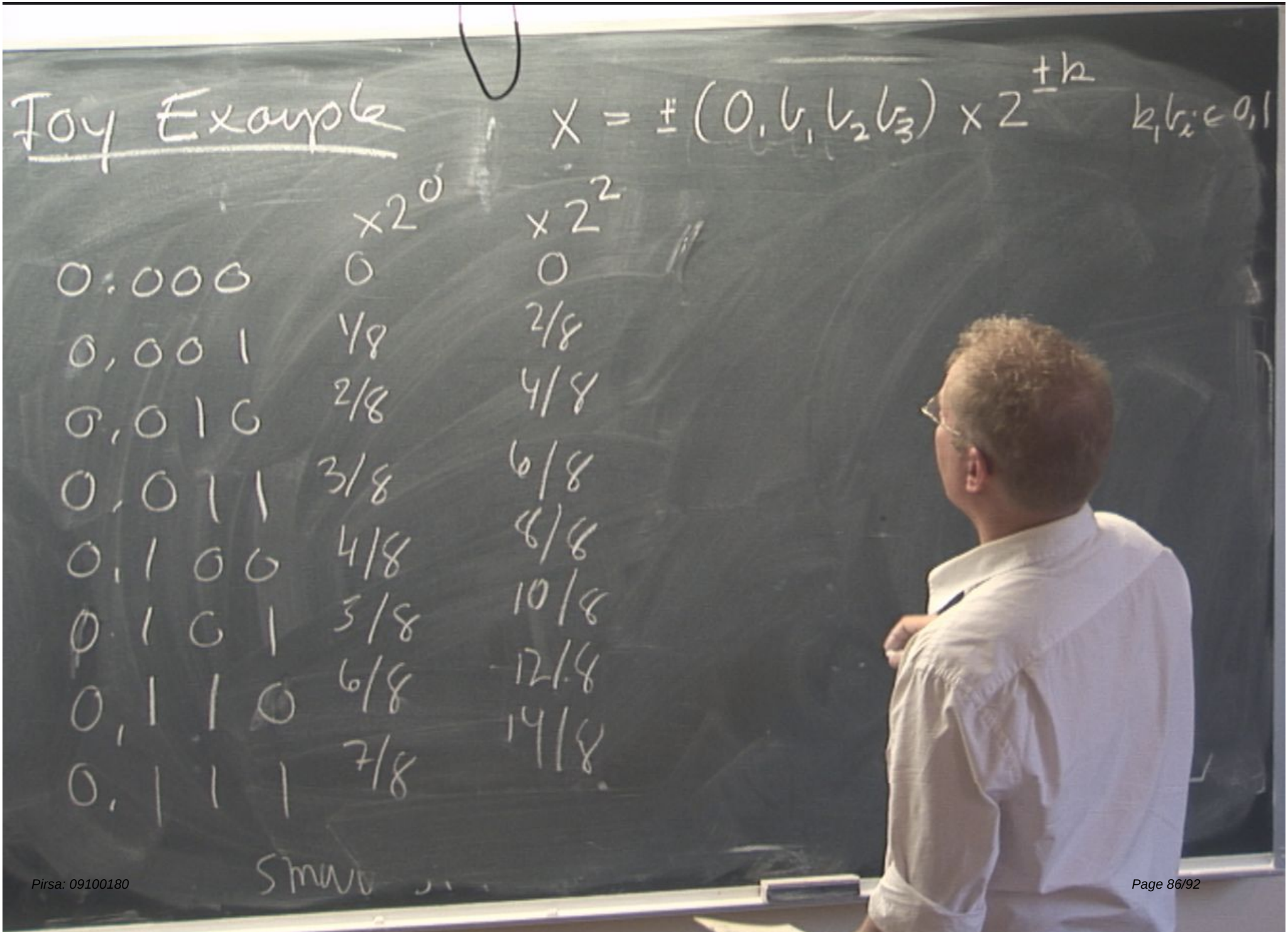
Toy Example : $X = \pm(0.b_1 b_2 b_3) \times 2^{\pm b}$

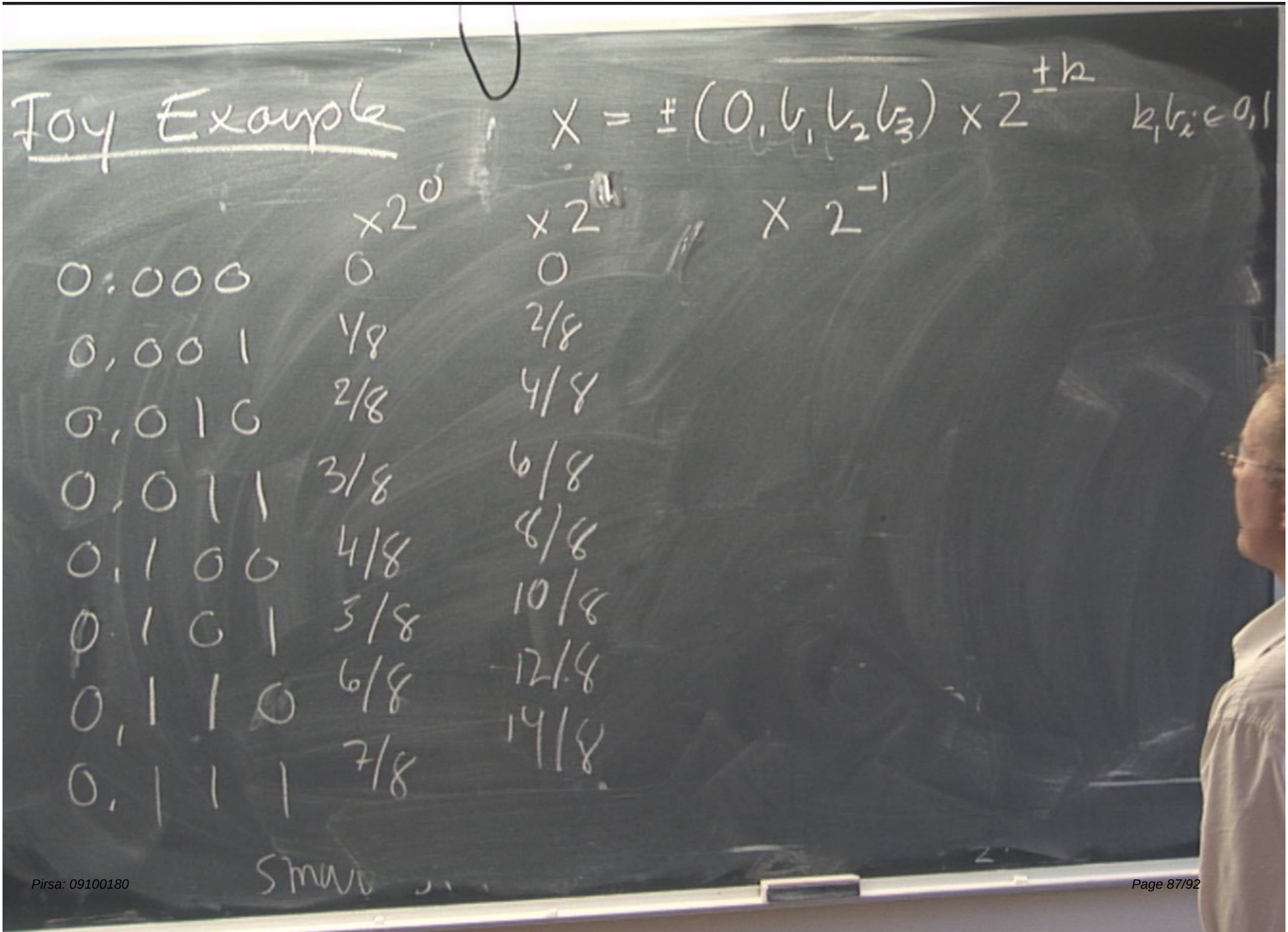

Toy Example
$X = \pm(0.b_1b_2b_3) \times 2^{\pm k}$ $k, b_i \in 0,1$
0.000


Toy Example
$X = \pm(0.b_1b_2b_3) \times 2^{\pm b_2}$ $b_1b_2 \in 0,1$
0.000
0,001
0,010
0,011


Toy Example
$X = \pm(0.b_1b_2b_3) \times 2^{\pm b}$ $b_1b_2 \in 0,1$
0:000
0,001
0,010
0,011
0,100
0,101
0,110
0,111


Toy Example
$X = \pm(0.b_1b_2b_3) \times 2^{\pm b_2}$ $b_1b_2 \in 0,1$
0.000 0
0.001 1/8
0.010 2/8
0.011 3/8
0.100 4/8
0.101 5/8
0.110 6/8
0.111 7/8

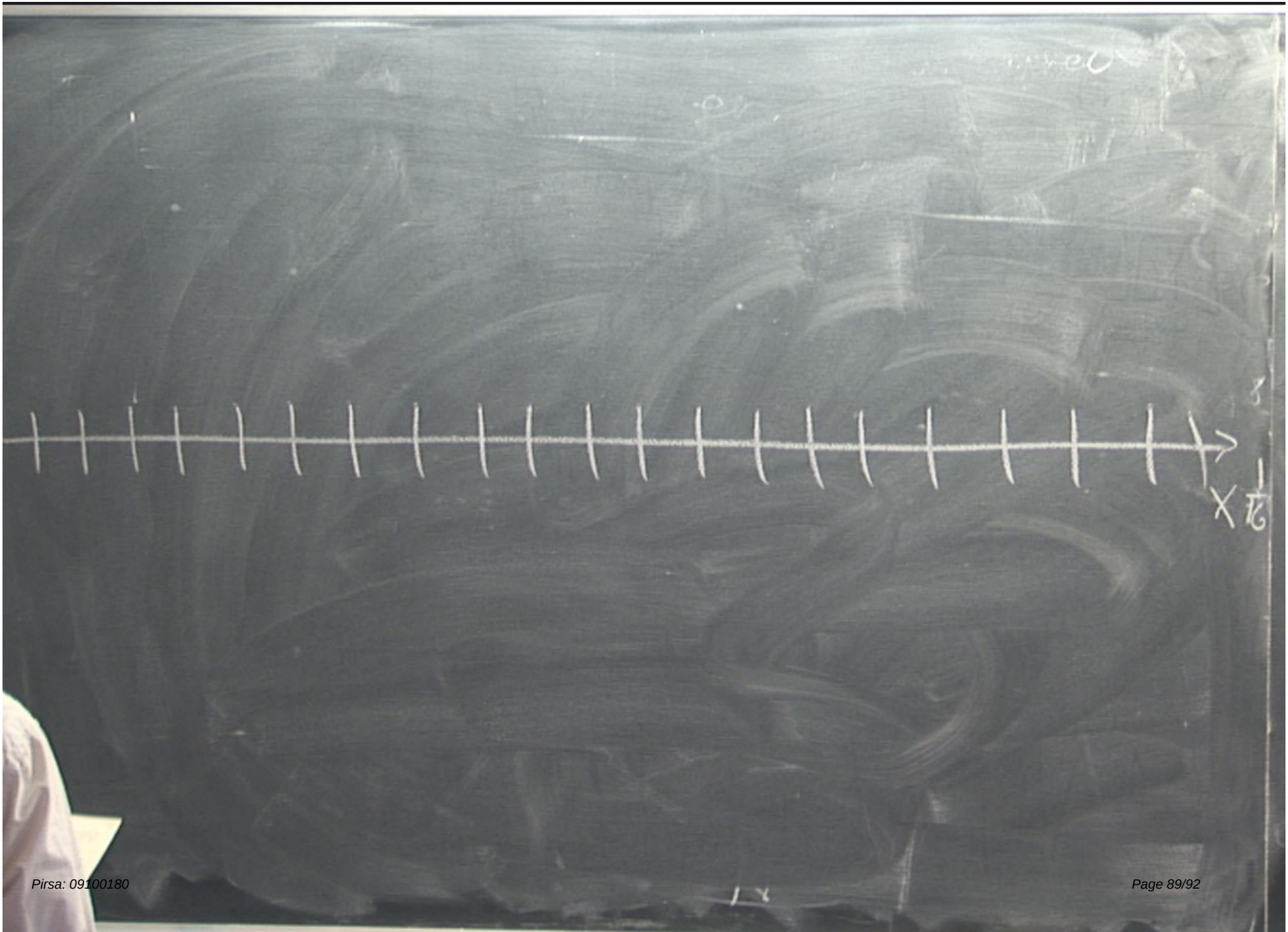

Toy Example
$X = \pm(0.b_1b_2b_3) \times 2^{\pm b_2}$ $b_1b_2 \in 0,1$
$\times 2^0$ $\times 2^2$
0.000 0 0
0.001 1/8 2/8
0.010 2/8 4/8
0.011 3/8 6/8
0.100 4/8 8/8
0.101 5/8 10/8
0.110 6/8 12/8
0.111 7/8 14/8

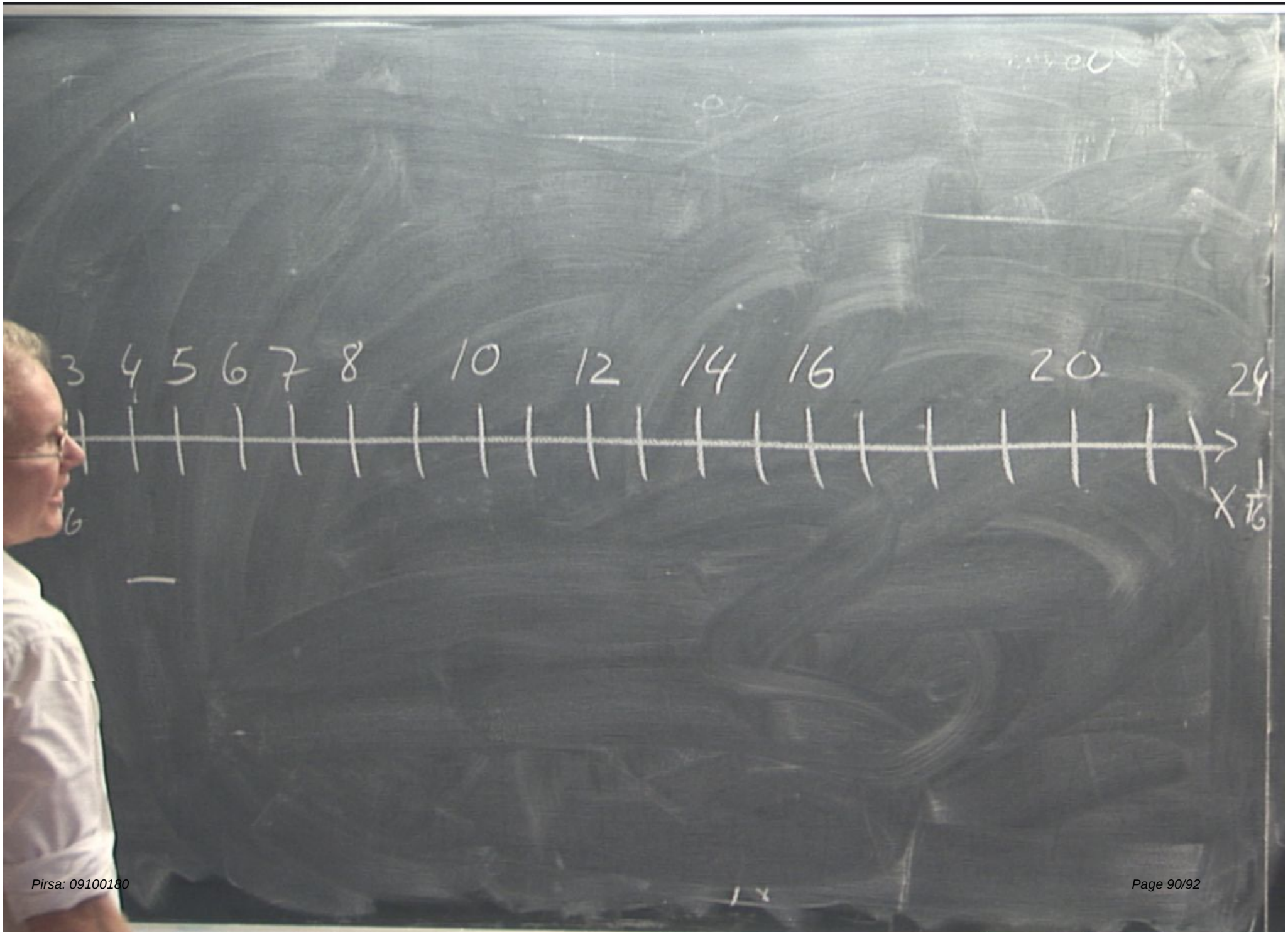

Toy Example
$X = \pm(0.b_1b_2b_3) \times 2^{\pm k}$ $k_1k_2 \in 0,1$
$\times 2^0$ $\times 2^1$ $\times 2^{-1}$
0.000 0 0
0.001 1/8 2/8
0.010 2/8 4/8
0.011 3/8 6/8
0.100 4/8 8/6
0.101 5/8 10/8
0.110 6/8 12/8
0.111 7/8 14/8

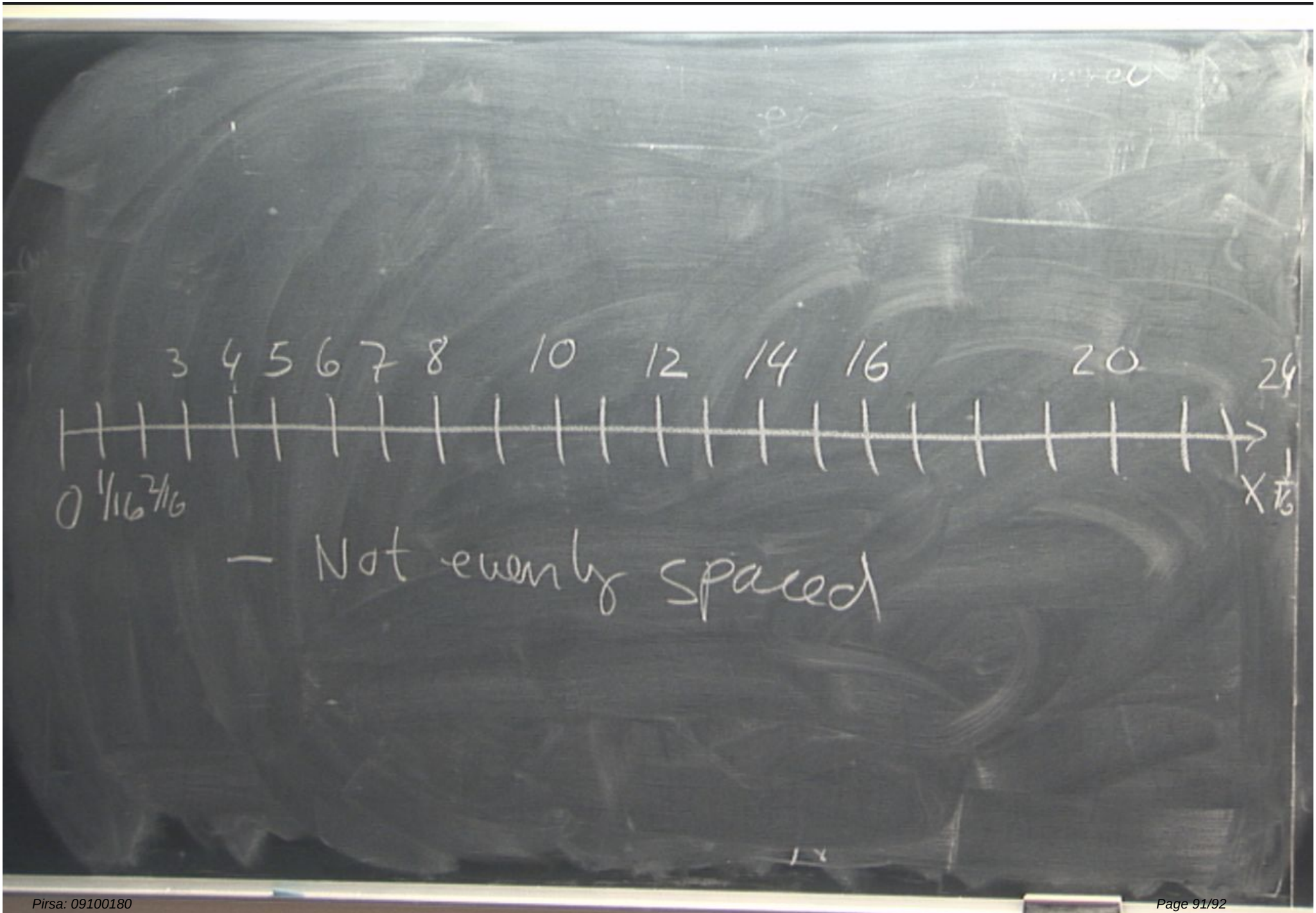
Toy Example

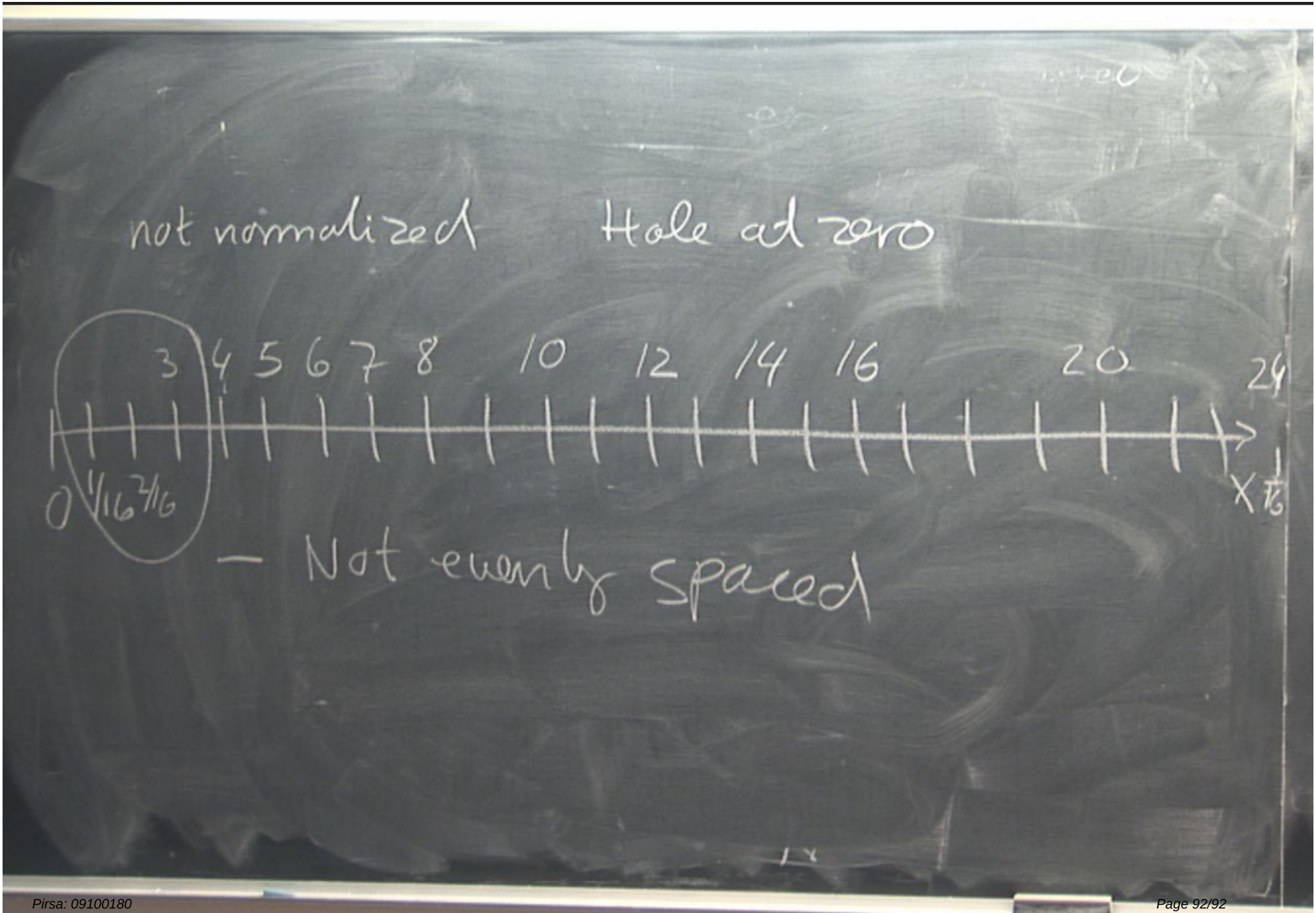
$X = \pm(0.b_1b_2b_3) \times 2^{\pm k}$ $\quad k_1k_2 \in 0,1$

	$\times 2^0$	$\times 2^1$	$\times 2^{-1}$
0.000	0	0	0
0.001	1/8	2/8	1/16
0.010	2/8	4/8	2/16
0.011	3/8	6/8	3/16
0.100	4/8	8/8	4/16
0.101	5/8	10/8	5/16
0.110	6/8	12/8	6/16
0.111	7/8	14/8	7/16


X


3 4 5 6 7 8 10 12 14 16 20 24
X
1/16


3 4 5 6 7 8 10 12 14 16 20 24
0 1/16 2/16
X 1/16
– Not evenly spaced


not normalized
Hole at zero
3
4 5 6 7 8 10 12 14 16 20 24
0 1/16 3/16
X 1/16
– Not evenly spaced