ΗΥ 134
Εισαγωγή στην Οργάνωση και στον Σχεδιασμό Υπολογιστών Ι

Διάλεξη 11

Αριθμητική Κινητής Υποδιαστολής
Πρόσθεση Αριθμών Κινητής Υποδιαστολής

Νίκος Μπέλλας
Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων

Αναπαράσταση πολύ μεγάλων και πολύ μικρών αριθμών

- Πως θα μπορούσαμε να αναπαραστήσουμε την μέση ηλικία της γης?

  4.600.000.000 ή $4,6 \times 10^9$ έτη

- ή την μονάδα ατομικής μάζας

  0,000000000000000000000000166 or $1,6 \times 10^{-27}$ Kg

Αυτοί οι αριθμοί δεν μπορούν να αναπαρασταθούν με 32-bit ή 64-bit ακεραίους

- Χρησιμοποιούμε αριθμούς κινητής υποδιαστολής για την αναπαράσταση πραγματικών αριθμών
  - float x, y
  - double x, y

Κινητή υποδιαστολή

- Αναπαράσταση μη-ακεραίων
  - Καθώς και πολύ μικρών/μεγάλων αριθμών
- Παρόμοιο με επιστημονική σημειογραφεία
  - –+987,02 × 10^9 Όχι ΕΣ
  - 2,34 × 10^{56} ΕΣ, Κ
  - +0,002 × 10^{-4} ΕΣ, Όχι Κ

Επιστημονική σημειογραφία: μόνο ένα ψηφίο πρίν την υποδιαστολή

Κανονικοποιημένος εάν είναι σε επιστημονική σημειογραφία και το ψηφίο αριστερά της υποδιαστολής δεν είναι 0.

- Στο δυαδικό σύστημα χρησιμοποιούμε βάση 2:
  - Σημαντικό Significand, Εκθέτης Exponent
  - $\pm 1,xxxxxxx_2 \times 2^{yyyy}$
  - Όταν ο αριθμός είναι κανονικοποιημένος, 1,0 ≤ |significand| < 2,0
  - Significand = “1,xxxx…”

IEEE 754 FP Standard

- Οι περισσότεροι υπολογιστές σήμερα χρησιμοποιούν το standard της IEEE 754 για αναπαράσταση αριθμών κινητής υποδιαστολής
  - Τι είναι η IEEE?
- Αριθμός = $(-1)^{\text{Πρόσημο}} \times (1+\text{Κλάσμα}) \times 2^{\text{Εκθέτης-Πόλωση}}$

  Απλή ακρίβεια : συνολικά 32 bits, Διπλή ακρίβεια : συνολικά 64 bits

  Απλή ακρίβεια: 8 bits
  Διπλή ακρίβεια: 11 bits

  Απλή ακρίβεια: 23 bits
  Διπλή ακρίβεια: 52 bits

- Υπάρχει συμβιβασμός μεταξύ μεγέθους (σε bits) του εκθέτη και του κλάσματος
  - Μεγαλύτερος εκθέτης → μεγαλύτερο εύρος τιμών (range)
  - Μεγαλύτερο κλάσμα → μεγαλύτερη ακρίβεια (precision)

IEEE 754 FP Standard

$$\text{Αριθμός} = (-1)^{\text{Πρόσημο}} \times (1+\text{Κλάσμα}) \times 2^{\text{Εκθέτης-Πόλωση}}$$

- Π: πρόσημο (0 ⇒ μη-αρνητικός, 1 ⇒ αρνητικός)
- Κανονικοποίηση του significand: 1.0 ≤ |significand| < 2.0
  - Πάντα έχει ένα αρχικό bit 1 οπότε δε χρειάζεται να αποθηκευτεί (κρυφό bit)
- Εκθέτης: πραγματικός εκθέτης + Πόλωση (Bias)
  - Εξασφαλίζει ότι ο εκθέτης είναι **απρόσημος**
  - Απλή ακρίβεια: Πόλωση = 127; Διπλή ακρίβεια: Πόλωση = 1023

Παραδείγματα IEEE 754 FP Standard

$$\text{Αριθμός} = (-1)^{\text{Πρόσημο}} \times (1+\text{Κλάσμα}) \times 2^{\text{Εκθέτης-Πόλωση}}$$

Π	Εκθέτης	Κλάσμα

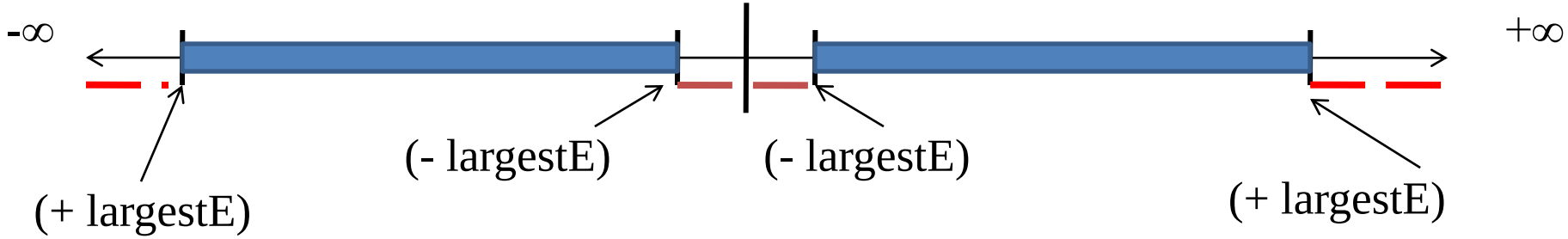
- Παραδείγματα (κανονικοποιημένη μορφή)
- Πρώτα πρέπει να γράψουμε τον αριθμό στην μορφή $\pm 1,xxxxxxx_2 \times 2^{yyyy}$

- $0,5 = 1,0_2 \times 2^{-1} = 1,0_2 \times 2^{126-127} =$ 0 01111110 00000000000000000000000
- $-0,875_{10} \times 2^4 = -0,111_2 \times 2^4 = -1, 11_2 \times 2^3 = -1,11_2 \times 2^{130-127} =$
  1 10000010 11000000000000000000000
- Μικρότερος +: 0 00000001 00000000000000000000000 = $1 \times 2^{1-127} = 2^{-126}$
- Zero: 0 00000000 00000000000000000000000 = true 0
- Μεγαλύτερος +: 0 11111110 11111111111111111111111 =
  $(2-2^{-23}) \times 2^{254-127} \sim 2 \times 2^{127} = 2^{128}$

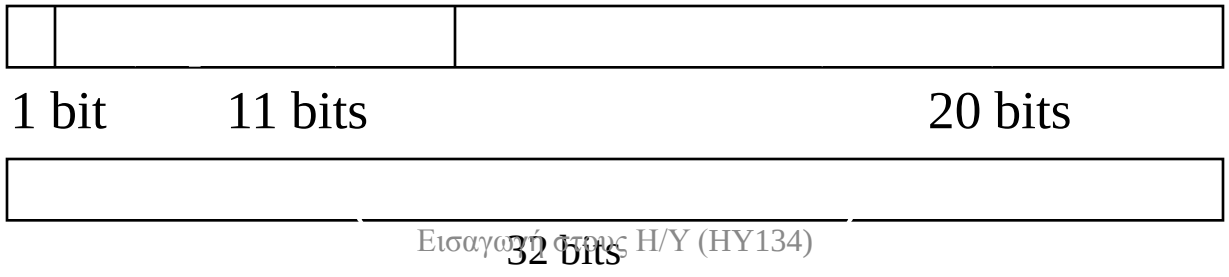
Εξαιρέσεις στους αριθμούς κινητής υποδιαστολής

- Υπερχείλιση (overflow) όταν ο αριθμός που θέλουμε να αναπαραστίσουμε έχει πολύ μεγάλο θετικό εκθέτη που δεν μπορεί να χωρέσει στο πεδίο του εκθέτη

- Ανεπάρκεια (underflow) όταν ο αριθμός που θέλουμε να αναπαραστίσουμε έχει πολύ μεγάλο αρνητικό εκθέτη που δεν μπορεί να χωρέσει στο πεδίο του εκθέτη



- Η αριθμητική διπλής ακρίβειας χρησιμοποιεί 64 bits για να αυξήσει τα bits εκθέτη και κλάσματος και να μειώσει την πιθανότητα overflow ή underflow



Παραδείγματα IEEE 754 FP Standard

- Ποιος αριθμός αναπαρίσταται σε απλή ακρίβεια από το

  1 10000001 01000…00

  – Π = 1

  – Κλάσμα = $01000\ldots00_2$

  – Εκθέτης = $10000001_2 = 129$

- $x = (-1)^1 \times (1 + ,01_2) \times 2^{(129 - 127)}$

  $= (-1) \times 1,25 \times 2^2$

  $= -5,0$

IEEE 754 FP Special Encoding

- Ειδικοί κώδικες στον IEEE 754 FP χρησιμοποιούνται για “ασυνήθιστα» γεγονότα
  - +- άπειρο μετά από διαίρεση με 0
  - NαN (not a number) για άκυρα αποτελέσματα όπως διαίρεση 0/0
  - Το 0 αναπαριστάται με όλα τα bits ίσα με 0 (γιά λόγους απλότητας)

Απλή ακρίβεια		Double Precision		Object Represented
E (8)	F (23)	E (11)	F (52)	
0	0	0	0	true zero (0)
0	nonzero	0	nonzero	± denormalized number
1-254	anything	1-2046	anything	± floating point number
255	0	2047	0	± infinity
255	nonzero	2047	nonzero	not a number (NaN)

Αθροιστής κινητής υποδιαστολής

Παράδειγμα 4-ψήφιου δεκαδικού. Θεωρούμε ότι μόνο 4 σημαντικά ψηφία μπορούν να αποθηκευθούν:

$9{,}999 \times 10^{1} + 1{,}610 \times 10^{-1}$

1. Ευθυγράμμιση υποδιαστολής

   Ολίσθηση αριθμού με μικρότερο εκθέτη

   $9{,}999 \times 10^{1} + 0{,}016 \times 10^{1}$

2. Πρόσθεση significands

   $9{,}999 \times 10^{1} + 0{,}016 \times 10^{1} = 10{,}015 \times 10^{1}$

3. Κανονικοποίηση αποτελέσματος και έλεγχος υπερείλισης ή υποχείλισης

   $10{,}015 \times 10^{1} \rightarrow 1{,}0015 \times 10^{2}$

4. Στρογγυλοποίηση και (αν ξαναχρειαστεί) κανονικοποίηση

   $1{,}002 \times 10^{2}$

Αθροιστής κινητής υποδιαστολής

Παράδειγμα 4-ψήφιου δυαδικού: $(\pm F1 \times 2^{E1}) + (\pm F2 \times 2^{E2}) = \pm F3 \times 2^{E3}$

$0{,}5 + -0{,}4375 = 1{,}000_2 \times 2^{-1} + -1{,}110_2 \times 2^{-2}$

1. Ευθυγράμμιση υποδιαστολής

   Ολίσθηση αριθμού με μικρότερο εκθέτη (E2) κατά E1-E2 θέσεις

   $1{,}000_2 \times 2^{-1} + -0{,}111_2 \times 2^{-1}$

2. Πρόσθεση significands ($\pm F1 + \pm F2 = F3$), θεωρώντας άπειρη ακρίβεια στο κλασματικό μέρος

   $1{,}000_2 \times 2^{-1} + -0{,}111_2 \times 2^{-1} = 0{,}001_2 \times 2^{-1}$

Αθροιστής κινητής υποδιαστολής

3. Κανονικοποίηση αποτελέσματος και έλεγχος overflow/underflow

   $0{,}001_2 \times 2^{-1} = 0{,}01_2 \times 2^{-2} = \ldots = 1{,}000_2 \times 2^{-4}$, χωρίς overlflow/underflow

4. Στρογγυλοποίηση και (αν ξαναχρειαστεί) κανονικοποίηση

   $1{,}000_2 \times 2^{-4}$ (καμμία αλλαγή) $= 0{,}0625$

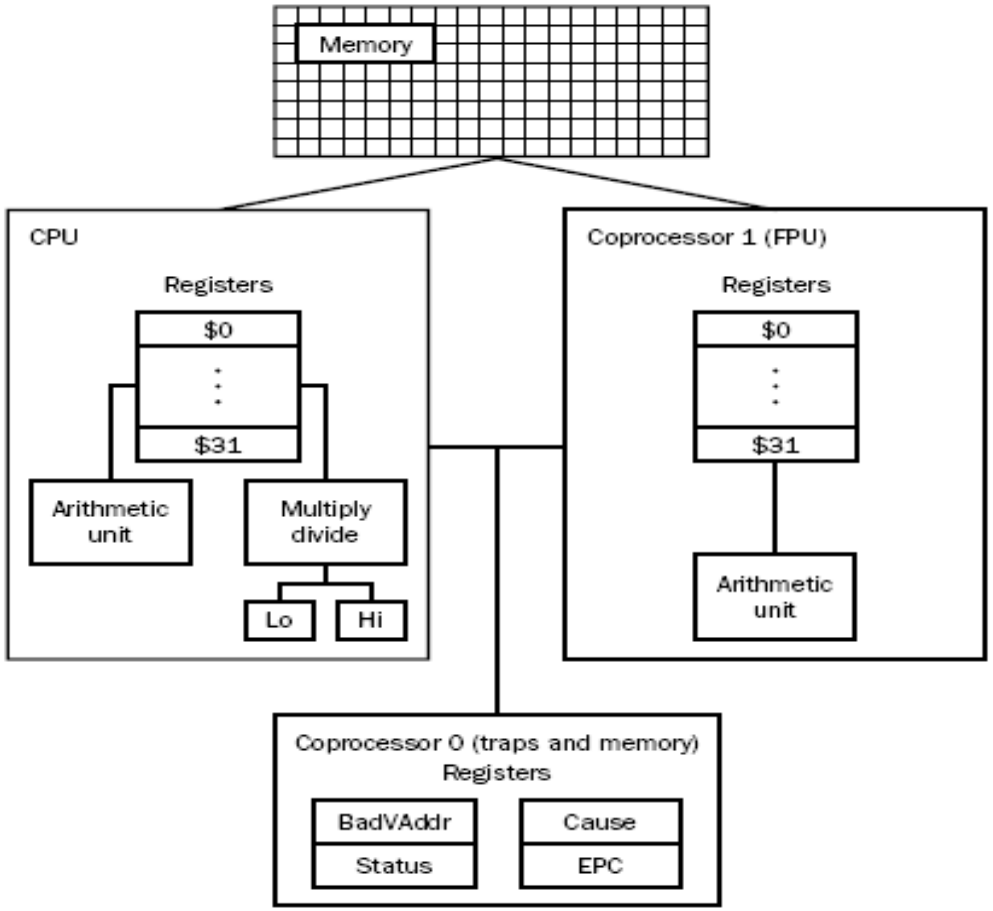
Παραδείγματα

- A + B = 2850,0 + -9840,0 = ? σε IEEE FP 754
  - A = 2850,0 = $(101100100010,0)_2 \times 2^0 = 1,01100100010 \times 2^{11}$
  - B = -9840,0 = $-(10011001110000,0)_2 = -1,0011001110000 \times 2^{13}$
  - $A = 1,01100100010 \times 2^{11} = 0,0101100100010 \times 2^{13}$
  - Πρόσθεση σημαντικών

    0,0101100100010 + -1,0011001110000 = 0,0101100100010 + 10,1100110010000 = 11,0010010110010 = -0,1101101001110
  - Sum = A+B = $-0,1101101001110 \times 2^{13} =$ **$-1,101101001110 \times 2^{12}$**
  - Πραγματικό αποτέλεσμα είναι: 2850,0 + -9840,0 = -6990,0
  - Αποτέλεσμα σε IEEE 754 είναι: $-1,101101001110 \times 2^{12} =$ -1,70654296875 x 4096 = -6990,0
  - Άρα Error = 0

Εντολές κινητής υποδιαστολής στο MIPS

- Το υλικό κινητής υποδιαστολής είναι ο συνεπεξεργαστής 1
  - Επεκτείνει το ISA
- Ξεχωριστοί κατ/τές κινητής υποδιαστολής
  - 32 απλής ακρίβειας: $f0, $f1, … $f31
  - Σε ζεύγη για διπλή ακρίβεια: $f0/$f1, $f2/$f3,



Ακριβής Αριθμητική

- Υπάρχουν άπειροι αριθμοί μεταξύ δύο οποιονδήποτε πραγματικών αριθμών
- Συνεπώς, οι αριθμοί κινητής υποδιαστολής είναι απλά προσεγγίσεις ενός αριθμού που δεν μπορεί να παρασταθεί πραγματικά
- Μικρότερος +: A = 0 00000001 00000000000000000000000 = $1 \times 2^{1-127} = 2^{-126}$
- Αμέσως μεγαλύτερος: B= 0 00000001 00000000000000000000001 = $(1+ 2^{-23}) \times 2^{-126}$
- $\Delta = B-A = 2^{-23} \times 2^{-126} = 2^{-149}$
- Αυτή είναι και η μεγαλύτερη δυνατή ακρίβεια στην αναπαράσταση μεταξύ διαδοχικών αριθμών απλής ακρίβειας
- Αλλά και πάλι ο αριθμοί : $(1+ 2^{-24}) \times 2^{-126}$, $(1+ 2^{-24} +2^{-25}) \times 2^{-126}$ κοκ. δεν μπορούν να αναπαρασταθούν

Ακριβής Αριθμητική

- Η στρογγυλοποίηση (rounding) χρησιμοποιείται για προσέγγιση πραγματικών αριθμών από το πεπερασμένο πλήθος των αριθμών κινητής υποδιαστολής
  - Η στρογγυλοποίηση ενδιάμεσων αποτελεσμάτων στην πρόσθεση και πολλαπλασιασμό μπορεί να επιφέρει χαμηλότερη προσέγγιση
  - Υποθέστε μόνο 3 σημαντικά ψηφία στην παρακάτω πρόσθεση

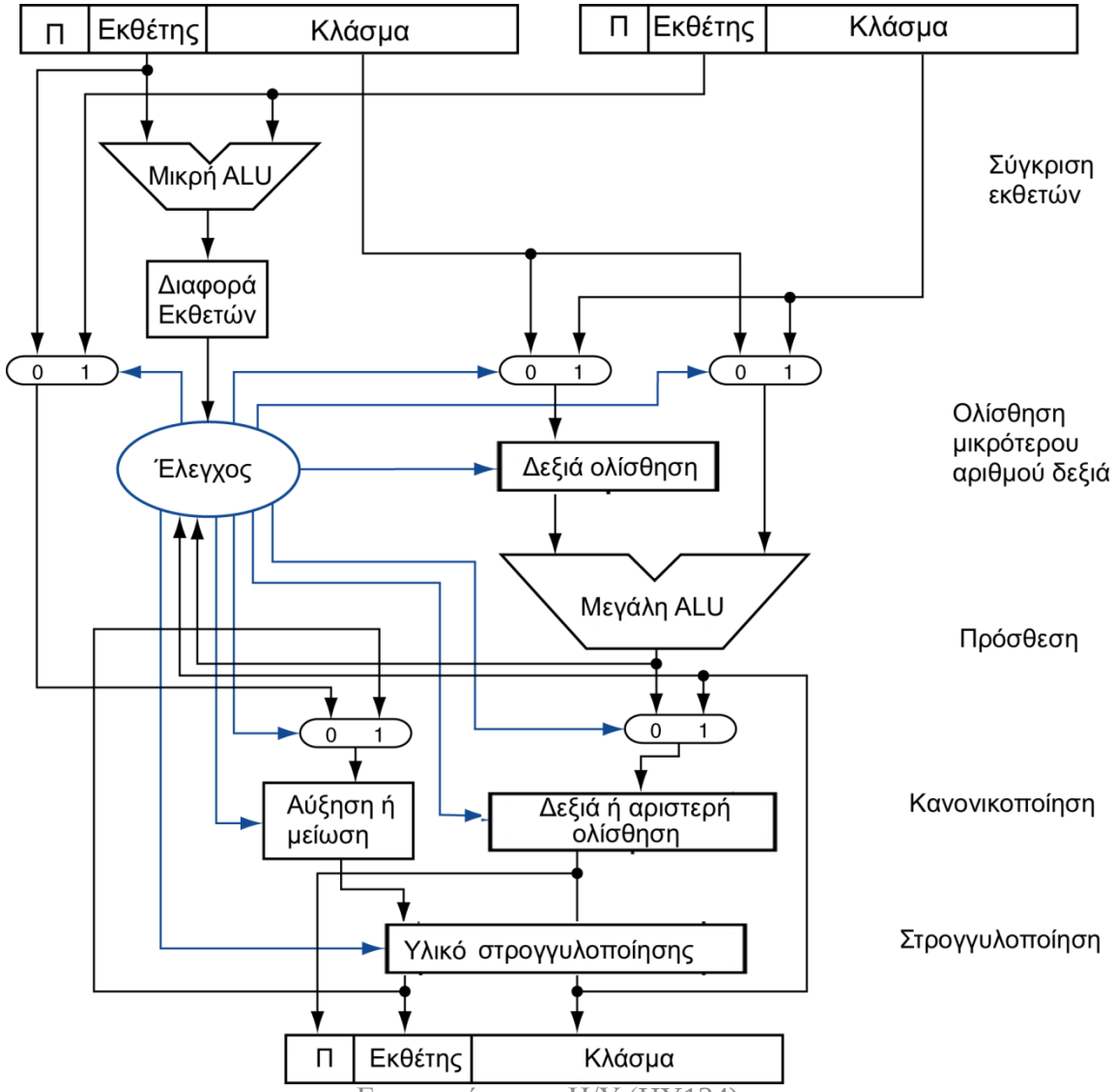
Στρογγυλοποίηση χωρίς extra bits	**Στρογγυλοποίηση με guard και round bits για ενδιάμεσα αποτελέσματα**
$2,56 \times 10^0 + 2,34 \times 10^2$ = $0,0256 \times 10^2 + 2,34 \times 10^2$ (Το 0,0256 πρέπει να στρογγυλοποιηθεί στο 0,02, επειδή τα ψηφία 5,6 χάνονται στην ολίσθηση) Άρα : $0,02 \times 10^2 + 2,34 \times 10^2 =$ **$2,36 \times 10^2$**	$2,56 \times 10^0 + 2,34 \times 10^2$ = $0,0256 \times 10^2 + 2,34 \times 10^2$ = $2,3656 \times 10^2$ Στρογγυλοποίηση στο **$2,37 \times 10^2$** για 3 σημαντικά ψηφία

Προσεταιριστικότητα (Associativity)

- Η πρόσθεση κινητής υποδιαστολής ΔΕΝ είναι προσεταιριστική πράξη
  - x + (y+z) != x+(y+z)
  - X = $-1,5 \times 10^{38}$
  - Y= $1,5 \times 10^{38}$
  - Z = 1,0

x+(y+z)	**(x+y)+z**
Y+Z = $1,5 \times 10^{38} + 1,0 = 1,5 \times 10^{38}$ X+(Y+Z) = $-1,5 \times 10^{38} + 1,5 \times 10^{38}$ = **0,0**	X+Y = $-1,5 \times 10^{38} + 1,5 \times 10^{38}$ = 0,0 (X+Y)+Z = 0,0 + 1,0 = **1,0**

Υλοποίηση αθροιστή κινητής υποδιαστολής



Αθροιστής κινητής υποδιαστολής

- Πολύ πιο πολύπλοκος από αθροιστή ακεραίων
- Συνήθως απαιτεί πολλαπλούς κύκλους μηχανής
- Τμήμα του Floating Point Unit (FPU) του επεξεργαστή