Εικονικοποίηση

Αρχιτεκτονική Υπολογιστών

5ο Εξάμηνο, 2016-2017

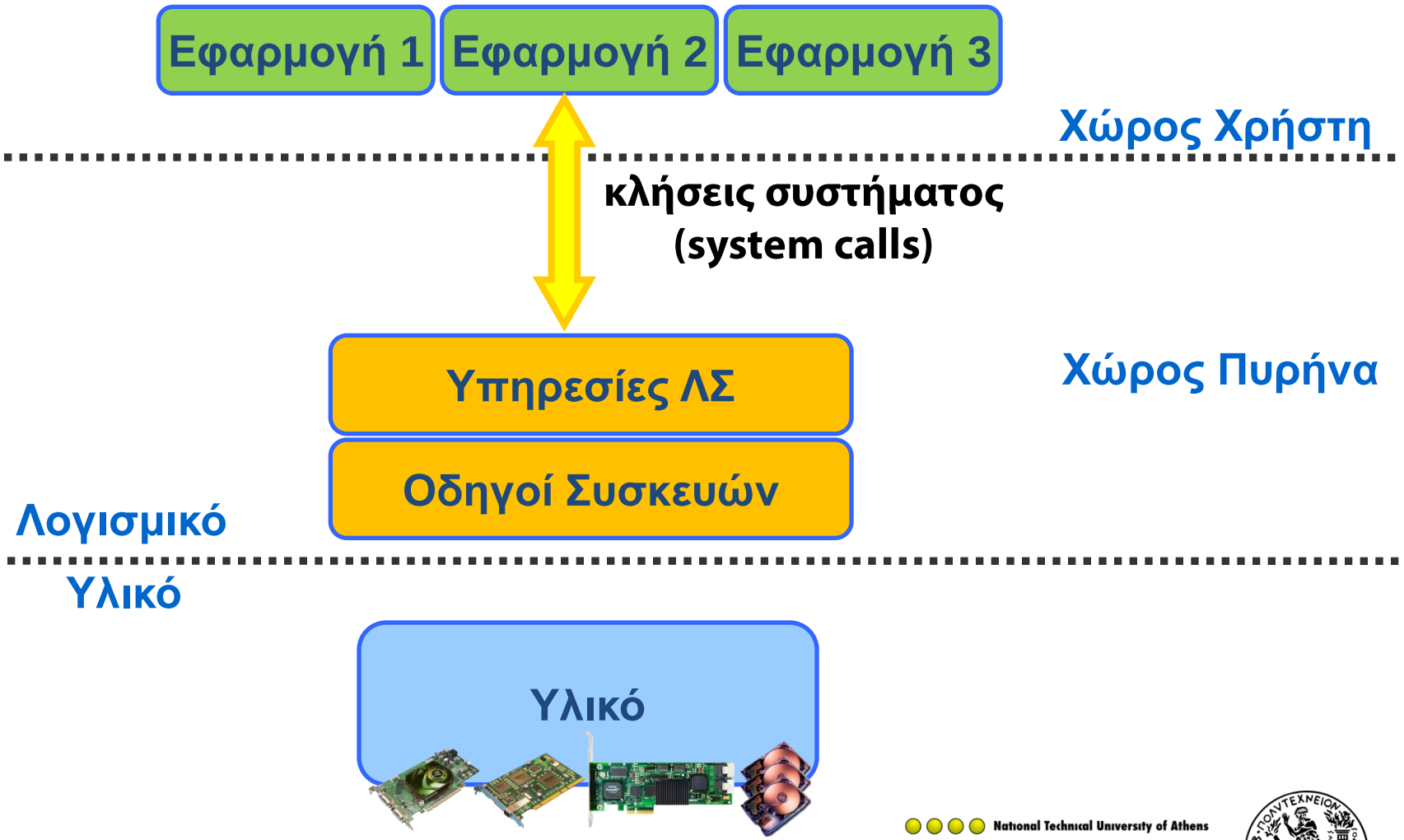
Εικονικοποίηση - Σύνοψη

- Γενικά
- Οργάνωση VMM
- Τεχνικές Εικονικοποίησης
- Εικονικοποίηση Μνήμης
- Live Migration

Εικονικοποίηση - Σύνοψη

- Γενικά
- Οργάνωση VMM
- Τεχνικές Εικονικοποίησης
- Εικονικοποίηση Μνήμης
- Live Migration

Αρχιτεκτονική Συστήματος Εγγενούς Εκτέλεσης (Native execution) (1)



Αρχιτεκτονική Συστήματος Εγγενούς Εκτέλεσης (Native execution) (2)

- Οπότε, από την πλευρά του λογισμικού προκύπτουν πλέον 2 καταστάσεις:
  - Χώρος χρήστη
  - Χώρος πυρήνα
- Ο επεξεργαστής (CPU) αναγνωρίζει τις 2 καταστάσεις για να παρέχει ασφάλεια ανάμεσα στις εφαρμογές και το ΛΣ
  - Παρέχει μηχανισμούς

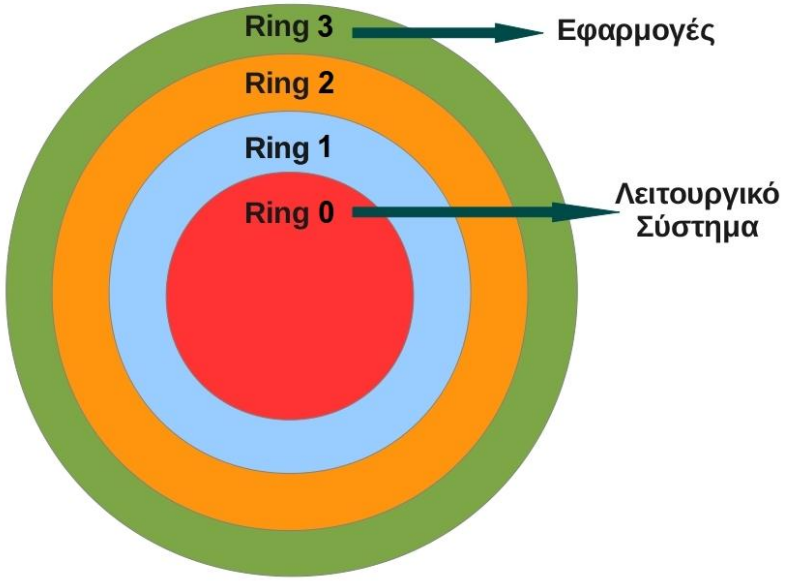
Απαιτήσεις Αρχιτεκτονικής

- Τουλάχιστον 2 καταστάσεις επεξεργαστή:
  - Κατάσταση χρήστη
  - Κατάσταση πυρήνα

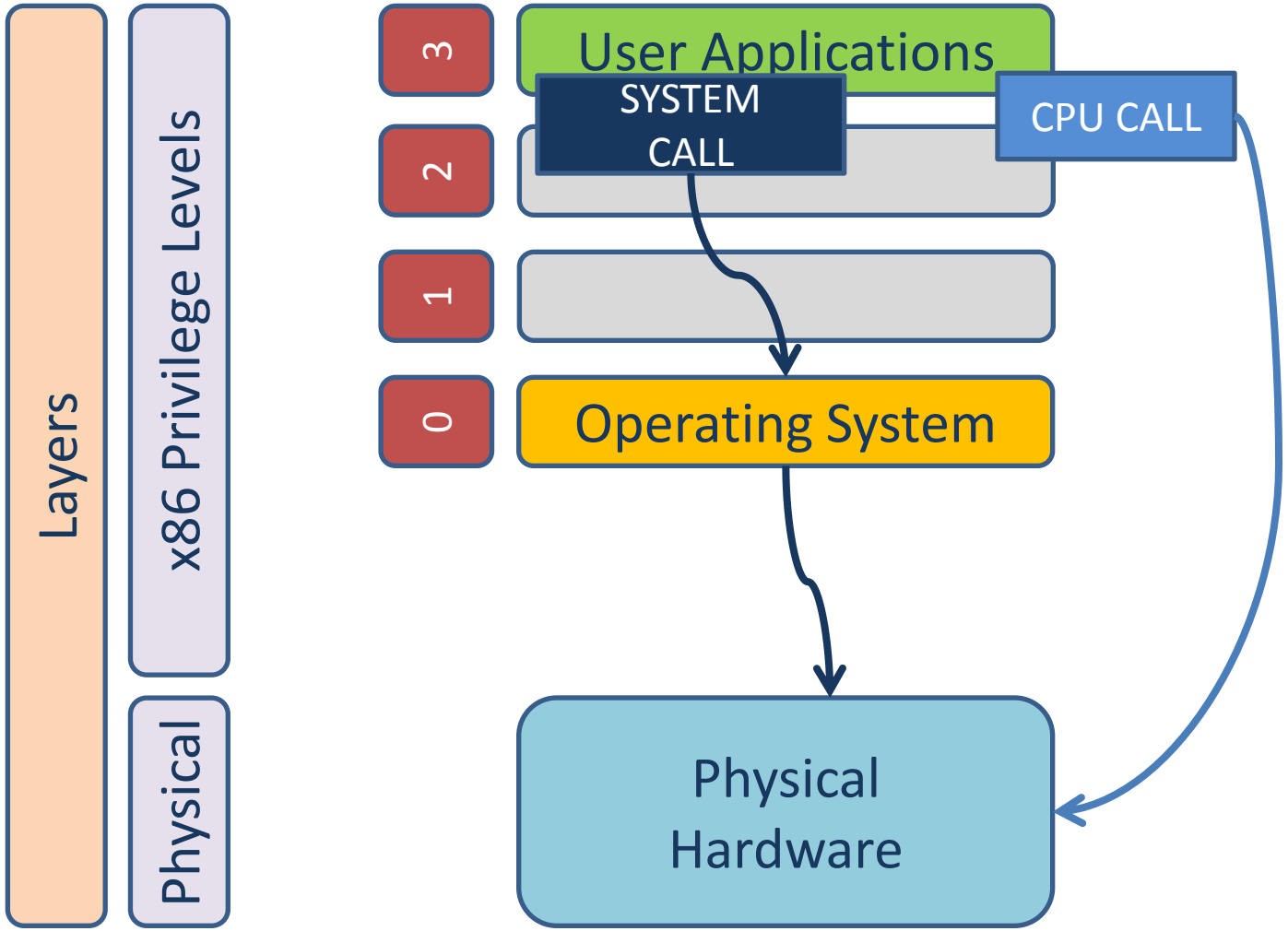
- Ένα προνομιούχο υποσύνολο εντολών που είναι διαθέσιμες μόνο σε κατάσταση πυρήνα
  - Αν εκτελεστούν σε κατάσταση χρήστη → **παγίδευση** (trap)
  - 'Ολοι οι πόροι του συστήματος ελέγχονται μέσω αυτών των εντολών

Privilege levels (Rings)

- Μηχανισμός ιεραρχίας του υλικού
  - Εξασφάλιση προστασίας & διαχωρισμού εφαρμογών
- Προνομιακότερο το χαμηλότερο ring
- Π.χ. Linux σε x86:
  - εφαρμογές (user-mode) → Ring 3
  - ΛΣ (kernel-mode) → Ring 0



Privilege levels (Rings) (2)



Εικονικοποίηση – Έννοιες

- **Εικονική Μηχανή (Virtual Machine - VM)**
  - Λογισμικό προσομοίωσης ενός φυσικού συστήματος (με πραγματικό/απτό υλικό)

- **Ελεγκτής Εικονικών Μηχανών (Virtual Machine Monitor – VMM – Hypervisor)**
  - Λογισμικό το οποίο είναι υπεύθυνο για τη δημιουργία, διαχείριση και ασφαλή εκτέλεση των εικονικών μηχανών

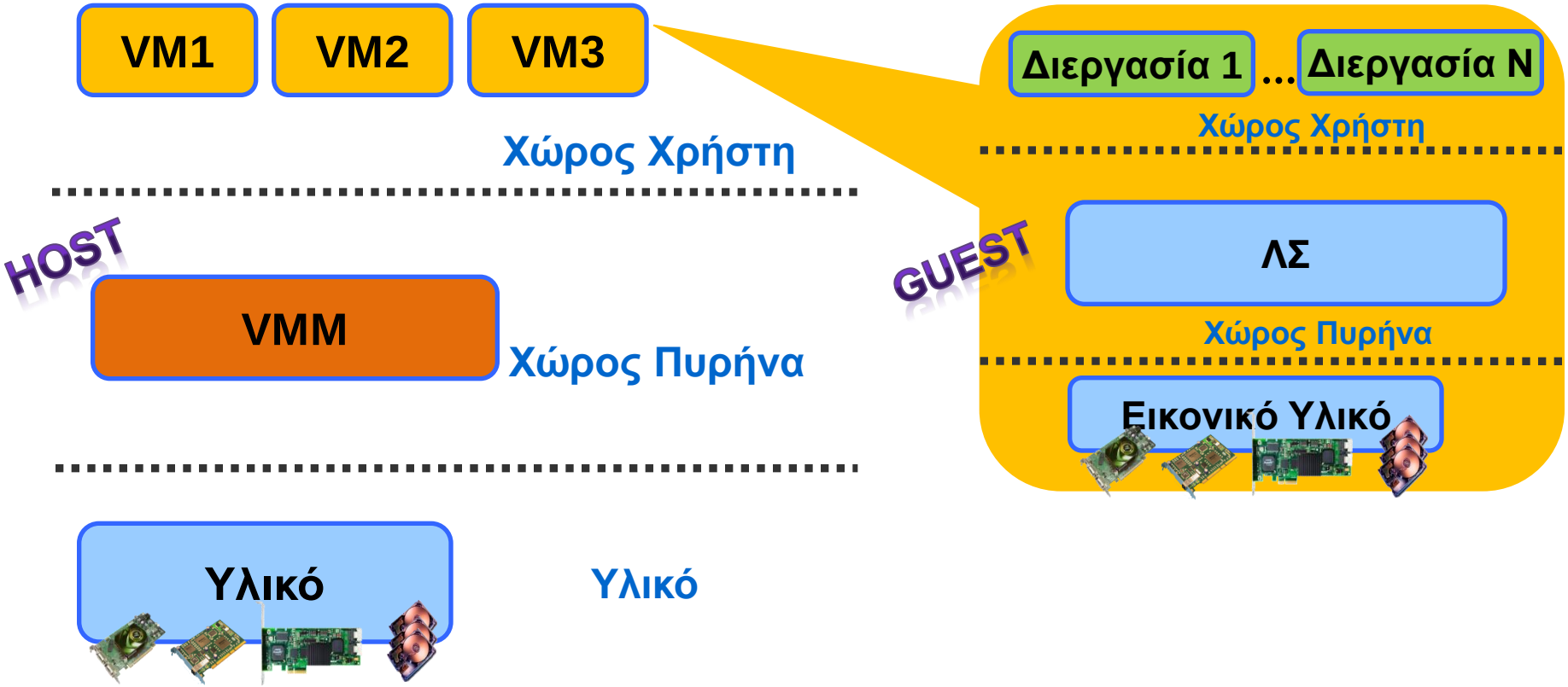
➡ **Guest,** για το εικονικό μηχάνημα

➡ **Host,** για το φυσικό μηχάνημα

Εικονικοποίηση (Virtualization)

- Δυνατότητα ταυτόχρονης εκτέλεσης πολλών *εικονικών μηχανών* σε ένα φυσικό μηχάνημα.
- Κάθε εικονική μηχανή έχει την ψευδαίσθηση ότι έχει πλήρη και αποκλειστική πρόσβαση στο υλικό του συστήματος (εικονική CPU (VCPU), εικονική μνήμη, εικονικές συσκευές).
- Ο VMM εγγυάται ασφαλή πρόσβαση των εικονικών μηχανών στους φυσικούς πόρους του συστήματος.

Εικονικοποίηση (Virtualization) (2)



VMM/Hypervisor

- Παρέχει την ψευδαίσθηση στα guest VMs ότι εκτελούνται στο εγγενές υλικό
- Ελέγχει την πρόσβαση σε πόρους συστήματος
  - Μεταφορά σε προνομιούχο κατάσταση
  - Μετάφραση διευθύνσεων
  - Είσοδο/έξοδο
  - Εξαιρέσεις, Διακοπές
- Χρησιμοποιεί ένα προνομιούχο υποσύνολο εντολών
  - Παγίδευση αν εκτελεστούν σε κατάσταση χρήστη

Οφέλη Εικονικοποίησης (1)

- Μείωση κόστους/ενέργειας
  - Π.χ. σε datacenters: λιγότεροι φυσικοί servers φιλοξενούν εικονικές μηχανές → μικρότερη υποδομή, λιγότερες ανάγκες για ενέργεια λειτουργίας/ψύξης.

- Απομόνωση + προστασία μεταξύ των εικονικών μηχανών

Οφέλη Εικονικοποίησης (2)

- Ευκολότερη ανάπτυξη και δοκιμή εφαρμογών σε διαφορετικά ΛΣ
- Μεγαλύτερη κλιμάκωση πλήθους εικονικών μηχανών ⟶ συστατικό στοιχείο του cloud computing

Οφέλη Εικονικοποίησης (3)

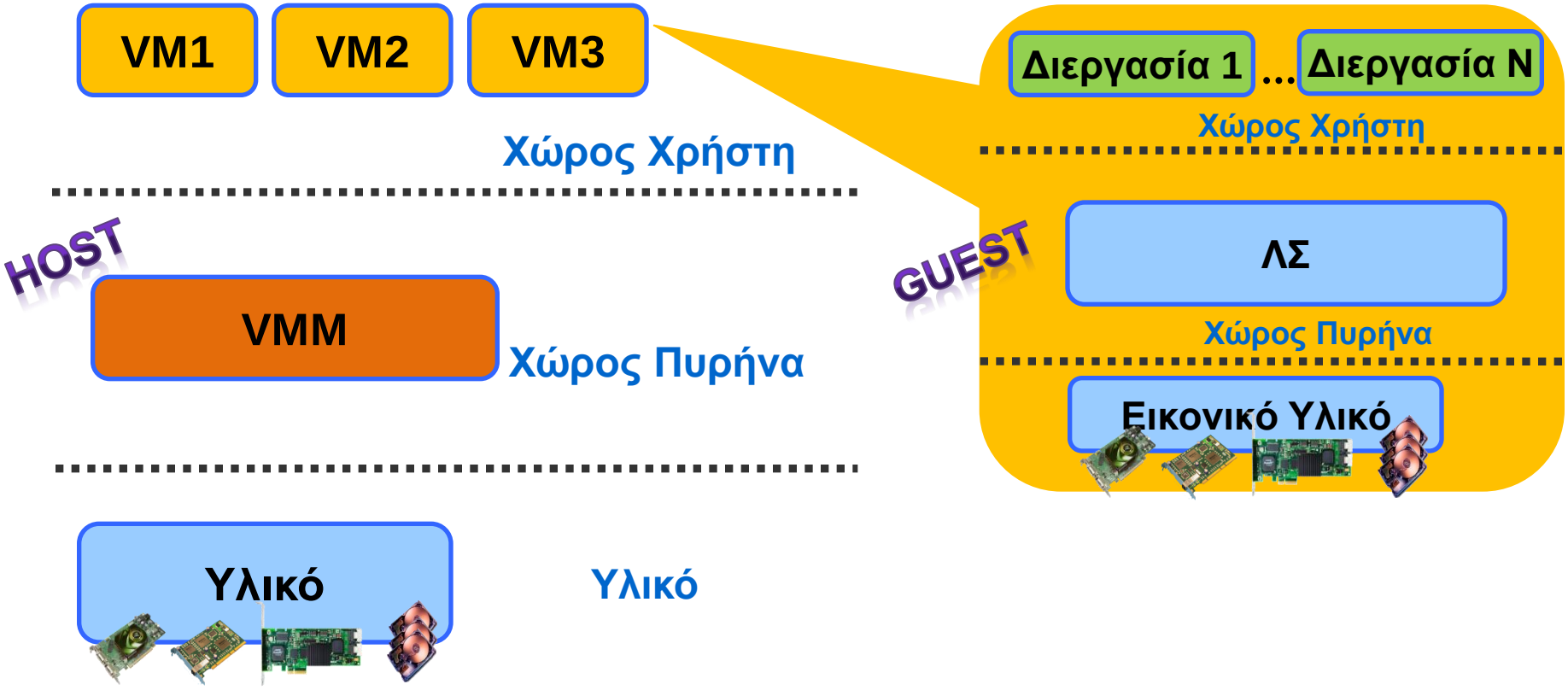
- Αστοχίες στην ασφάλεια και αξιοπιστία των συνηθισμένων λειτουργικών συστημάτων ⟶ αυξημένη φερεγγυότητα

- Δραματική αύξηση της ταχύτητας των επεξεργαστών ⟶ αποδεκτό το κόστος εικονικοποίησης στην επίδοση

Εικονικοποίηση - Σύνοψη

- Γενικά
- Οργάνωση VMM
- Τεχνικές Εικονικοποίησης
- Εικονικοποίηση Μνήμης
- Live Migration

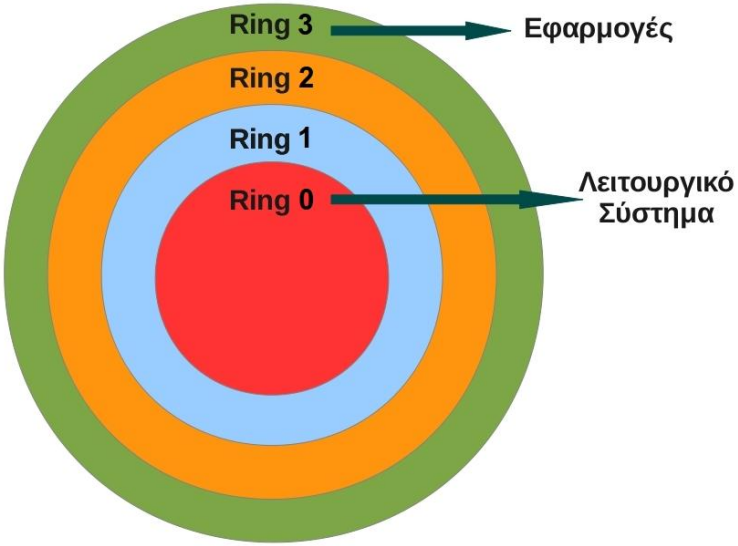
Οργάνωση VMM (1)



- Ο VMM δίνει την ψευδαίσθηση στις εικονικές μηχανές ότι εκτελούνται στο φυσικό υλικό

Οργάνωση VMM (2)

- Οπότε, από την πλευρά του λογισμικού προκύπτουν πλέον 4 καταστάσεις:
  - Χώρος χρήστη guest
  - Χώρος πυρήνα guest
  - Χώρος χρήστη host
  - Χώρος πυρήνα host



- Ο επεξεργαστής (CPU), όμως, γνωρίζει μόνο 2 καταστάσεις: χώρος χρήστη (μη-προνομιούχος) και χώρος πυρήνα (προνομιούχος)

Εικονικοποίηση - Σύνοψη

- Γενικά
- Οργάνωση VMM
- Τεχνικές Εικονικοποίησης
- Εικονικοποίηση Μνήμης
- Live Migration

Κατηγορίες Εντολών (0)

- Προνομιούχες εντολές
- Μη-προνομιούχες εντολές
- Ευαίσθητες εντολές

Κατηγορίες Εντολών (1)

- Προνομιούχες εντολές
  - Μπορούν να εκτελεστούν απευθείας μόνο αν η CPU βρίσκεται σε προνομιούχο κατάσταση (χώρο πυρήνα).
  - Αν η CPU βρίσκεται σε μη-προνομιούχο κατάσταση (χώρο χρήστη), προκαλείται **trap**, η CPU μεταβαίνει σε προνομιούχο κατάσταση και η εκτέλεση συνεχίζεται από προκαθορισμένη ρουτίνα χειρισμού στο ΛΣ ή στο VMM αντίστοιχα.

Κατηγορίες Εντολών (2)

- Μη-προνομιούχες εντολές
  - Μπορούν να εκτελεστούν απευθείας σε οποιαδήποτε κατάσταση βρίσκεται η CPU.
- Ευαίσθητες εντολές

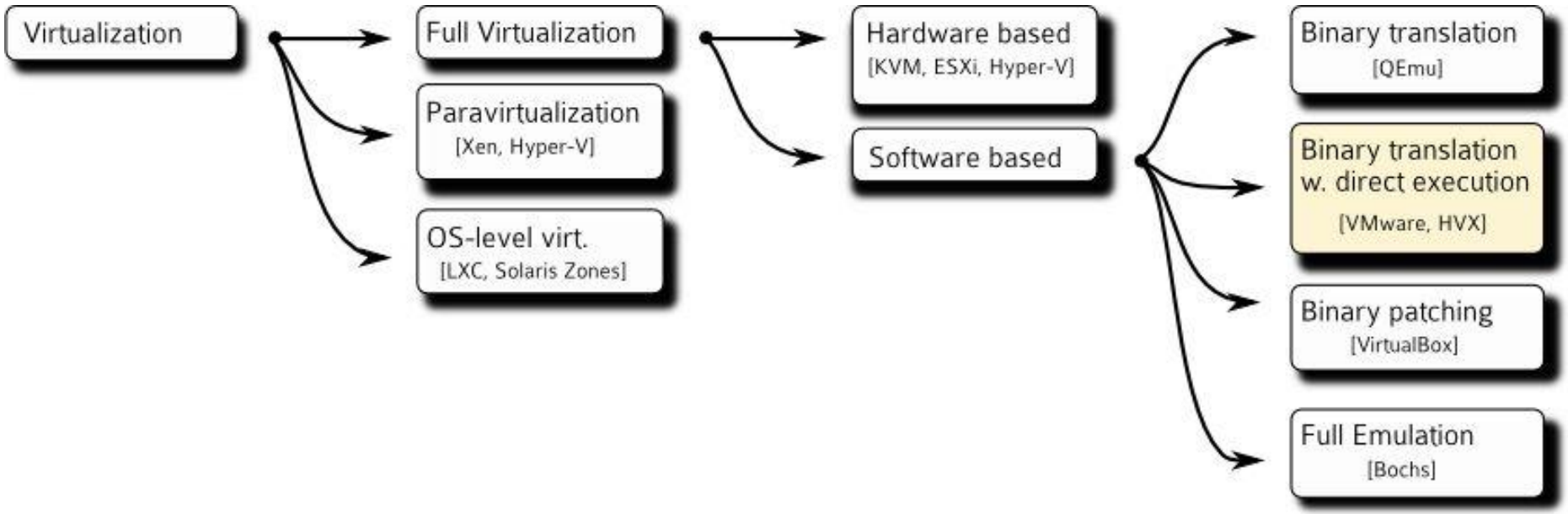
Κατηγορίες Εντολών (3)

- Κατηγοριοποίηση ευαίσθητων εντολών:
  - **Ευαίσθητες εντολές ελέγχου**: Εκείνες που προσπαθούν να αλλάξουν τις ρυθμίσεις των πόρων του συστήματος.
  - **Ευαίσθητες εντολές συμπεριφοράς**: Η συμπεριφορά ή το αποτέλεσμά τους εξαρτάται από τις ρυθμίσεις των πόρων (π.χ. την κατάσταση στην οποία βρίσκεται η CPU).

Κατηγορίες Εντολών (4)

- Ανάλογα με την αρχιτεκτονική του συστήματος, συγκεκριμένες ευαίσθητες εντολές μπορεί να παράγουν **trap** όταν εκτελούνται σε χώρο χρήστη και άρα να είναι και *προνομιούχες*.
- Όμως, δεν είναι όλες οι ευαίσθητες εντολές προνομιούχες. Εξαρτάται από την αρχιτεκτονική.
  - Π.χ. η ***popf*** σε x86 εκτελείται χωρίς trap και σε χώρο χρήστη παράγοντας διαφορετικό αποτέλεσμα.

Κατηγοριοποίηση Εικονικοποίησης



Μέθοδοι Εικονικοποίησης

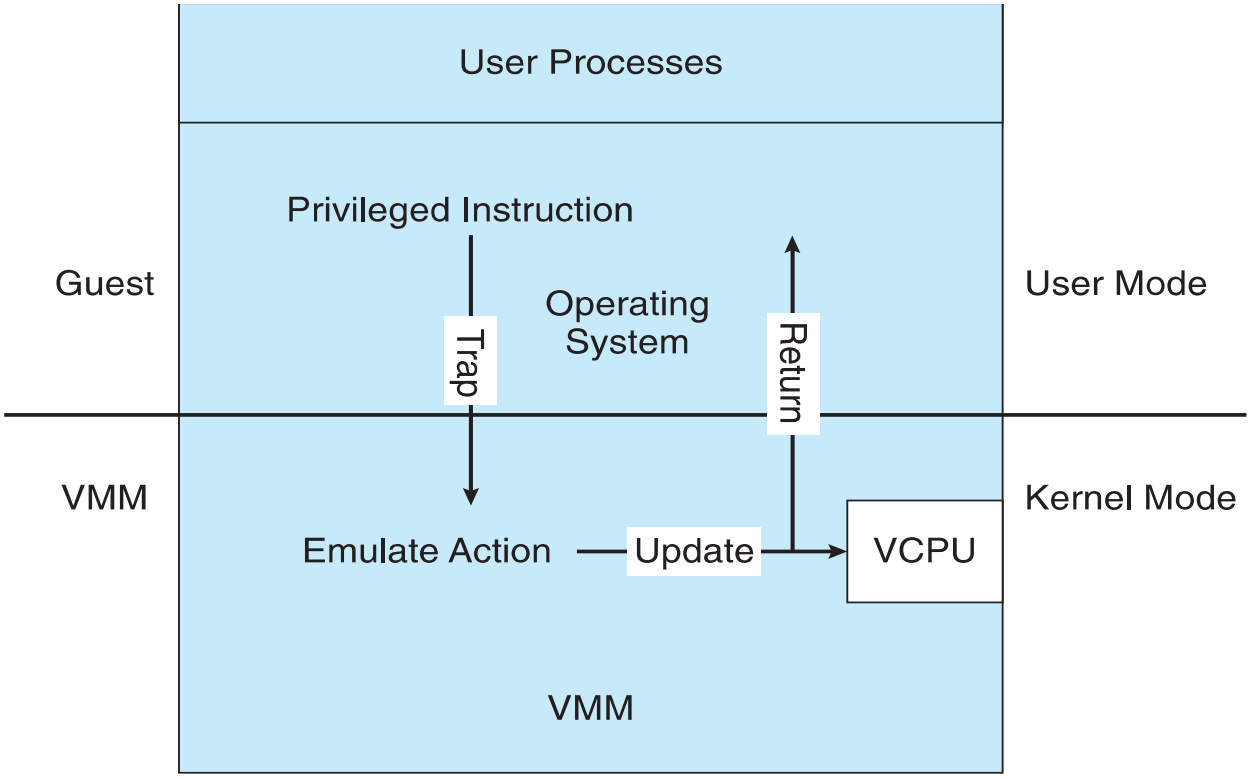
- Trap & Emulate
- Binary Translation
- Hardware-assisted
- Paravirtualization

Μέθοδος trap & emulate (1)

- Η εικονική μηχανή εκτελείται στο χώρο χρήστη του host.
- Ο VMM εκτελείται στο χώρο πυρήνα του host.

- Οι μη-προνομιούχες εντολές (διεργασιών ή πυρήνα) εκτελούνται απευθείας από τη CPU.
- Οι προνομιούχες εντολές προκαλούν trap.
  - Παρεμβαίνει ο VMM και προσομοιώνει τη συμπεριφορά της εντολής που θα ανέμενε η εικονική μηχανή.

Μέθοδος trap & emulate (2)



- **Trap:** Η CPU μεταβαίνει σε προνομιούχο κατάσταση
- **Return:** Η CPU μεταβαίνει σε μη-προνομιούχο κατάσταση

Μέθοδος trap & emulate (3)

- **Θεώρημα:** "Για να είναι δυνατή η εικονικοποίηση με τη μέθοδο *trap & emulate*, πρέπει το σύνολο των ευαίσθητων εντολών να είναι υποσύνολο των προνομιούχων εντολών".

Popek, G. J., Goldberg, R. P. (July 1974)."Formal requirements for virtualizable third generation architectures".

Μέθοδος trap & emulate (3)

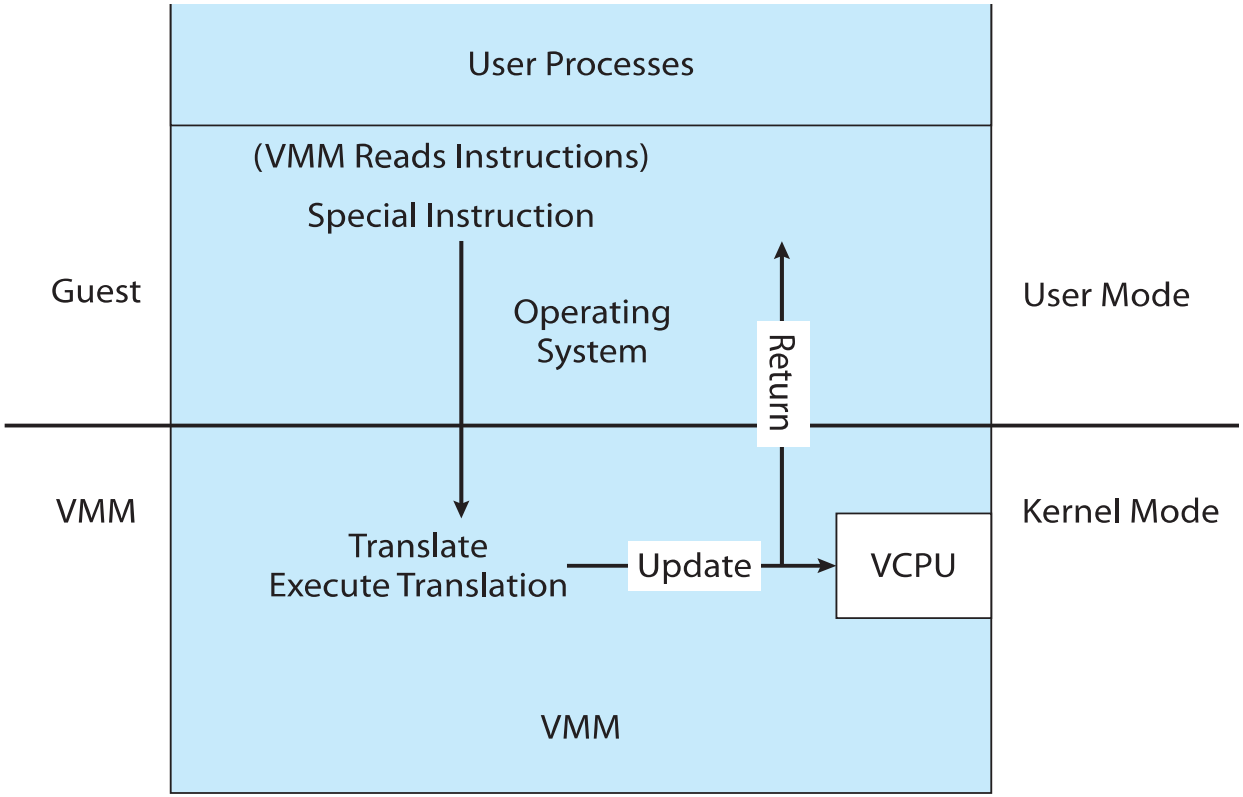
- Δεν υπάρχει επιπλέον κόστος εκτέλεσης των μη-προνομιούχων εντολών.
- Αντίθετα, οι προνομιούχες εντολές κοστίζουν λόγω emulation.
- Δεν είναι εφικτή μέθοδος σε αρχιτεκτονικές με ευαίσθητες εντολές που δεν προκαλούν trap όταν εκτελούνται στο χώρο χρήστη.

Μέθοδος binary translation (1)

- Πρέπει ο VMM να είναι σε θέση να “πιάσει” ευαίσθητες εντολές που εκτελούνται σε guest χώρο πυρήνα (άρα σε host χώρο χρήστη).

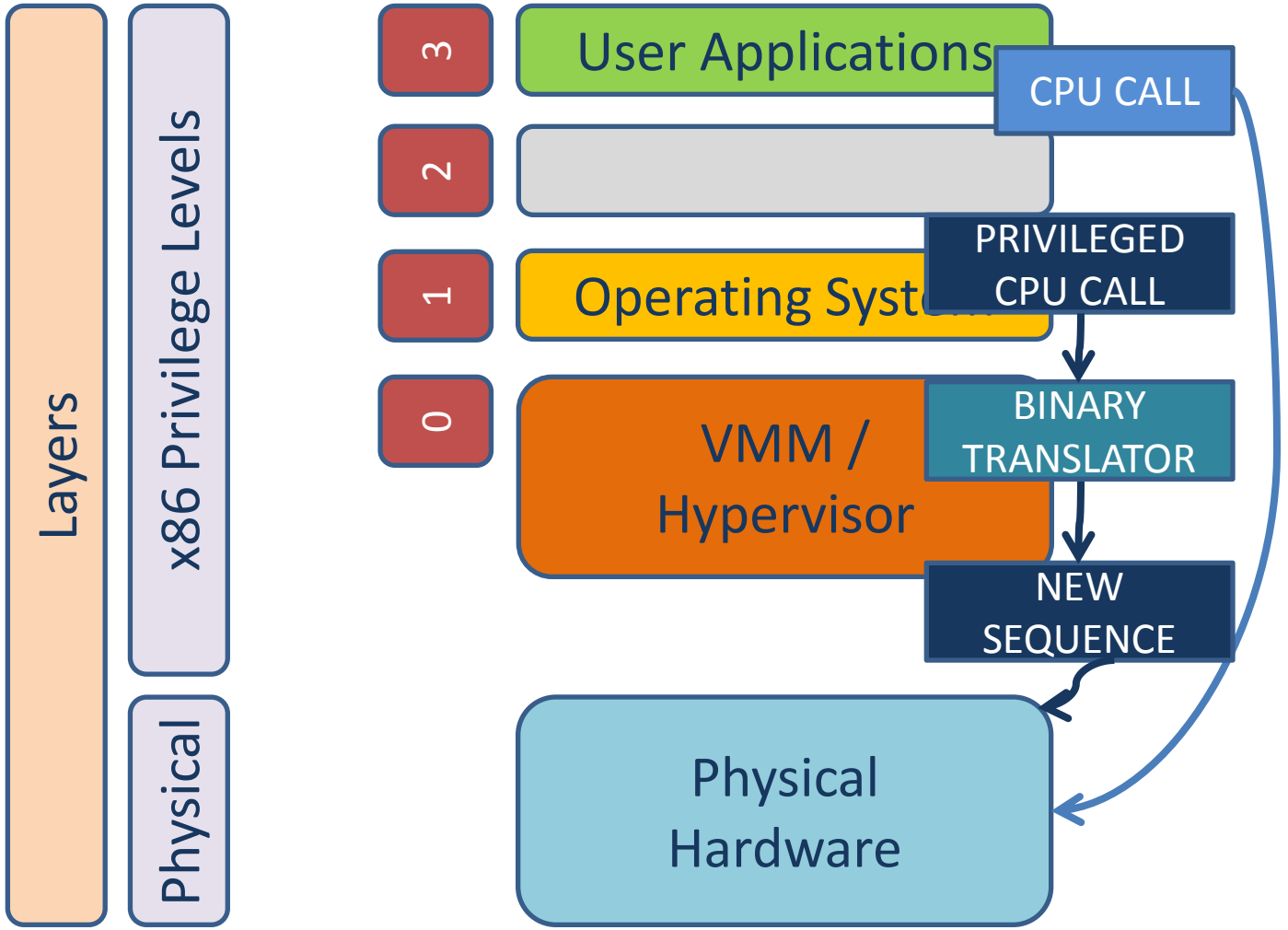
- Αν αυτές δεν προκαλούν trap στο χώρο χρήστη;
- Λύση:
  - Οι εντολές στο guest χώρο χρήστη εκτελούνται απευθείας στη CPU.
  - Ο VMM διαβάζει τις εντολές στο guest χώρο πύρηνα και τις μεταφράζει (στατικά ή δυναμικά) με άλλες που προσομοιώνουν την αντίστοιχη λειτουργία.

Μέθοδος binary translation (2)



- Συνήθως ο VMM δε μεταφράζει μία-μία τις εντολές, αλλά σε block, τα οποία αποθηκεύει σε cache.
- Ακόμα και έτσι, όμως, το κόστος της μετάφρασης είναι μεγάλο.

Μέθοδος binary translation (3)



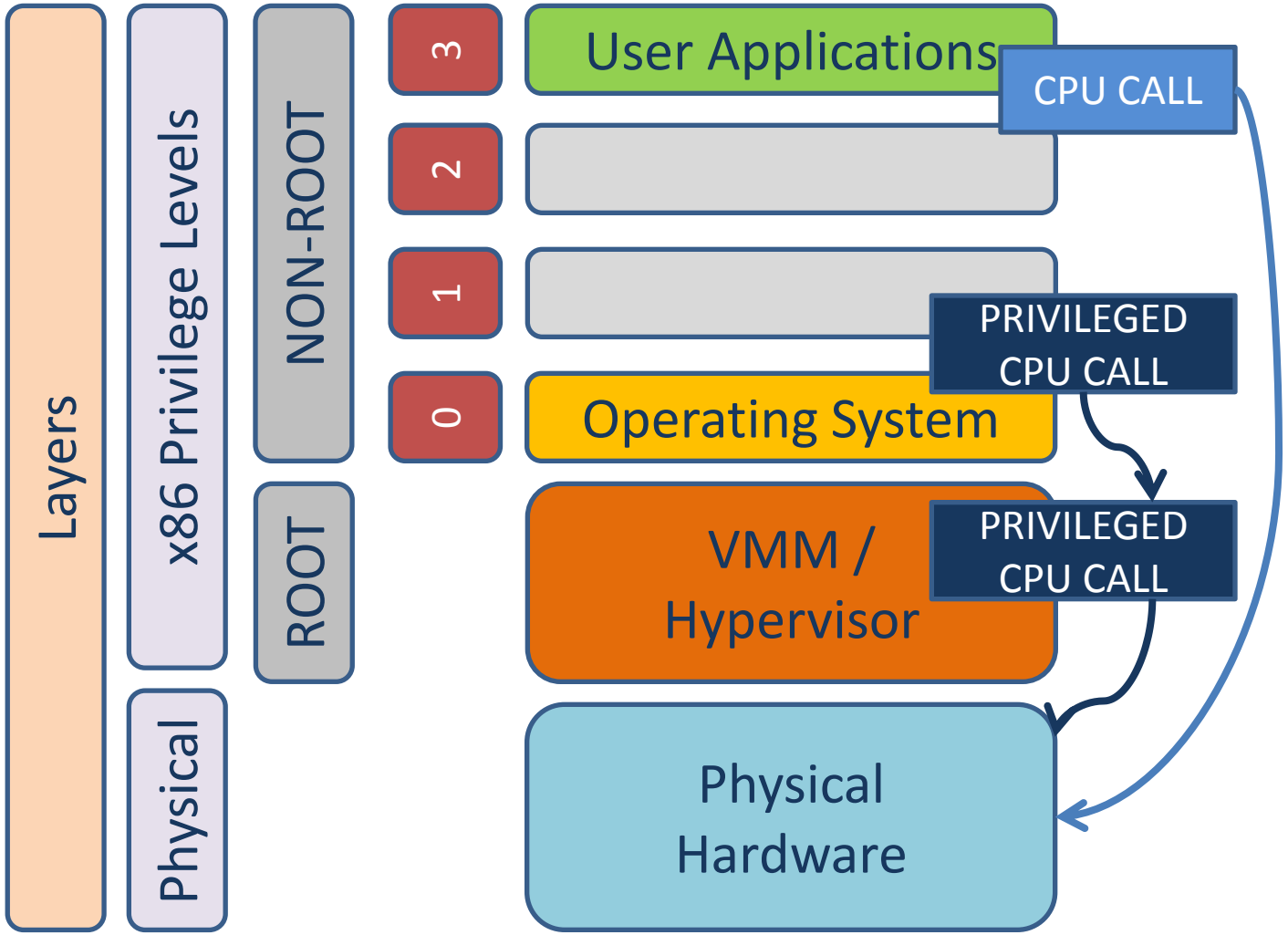
Μέθοδος εικονικοποίησης υποβοηθούμενη από το υλικό (1)

- Αλλαγές στο υλικό, προκειμένου η CPU να “αναγνωρίζει” και τις 4 καταστάσεις guest/host.
  - Πώς → Ring -1 για τον VMM

- Έτσι, οι αντίστοιχες ευαίσθητες εντολές προκαλούν πάντα **trap** όταν εκτελούνται σε guest χώρο πυρήνα και παρεμβαίνει ο VMM.

- Σαφώς βελτιωμένη επίδοση - απαιτούνται επεκτάσεις υλικού (virtualization extensions), π.χ. Intel VT-x, AMD-V.

Μέθοδος εικονικοποίησης υποβοηθούμενη από το υλικό (2)



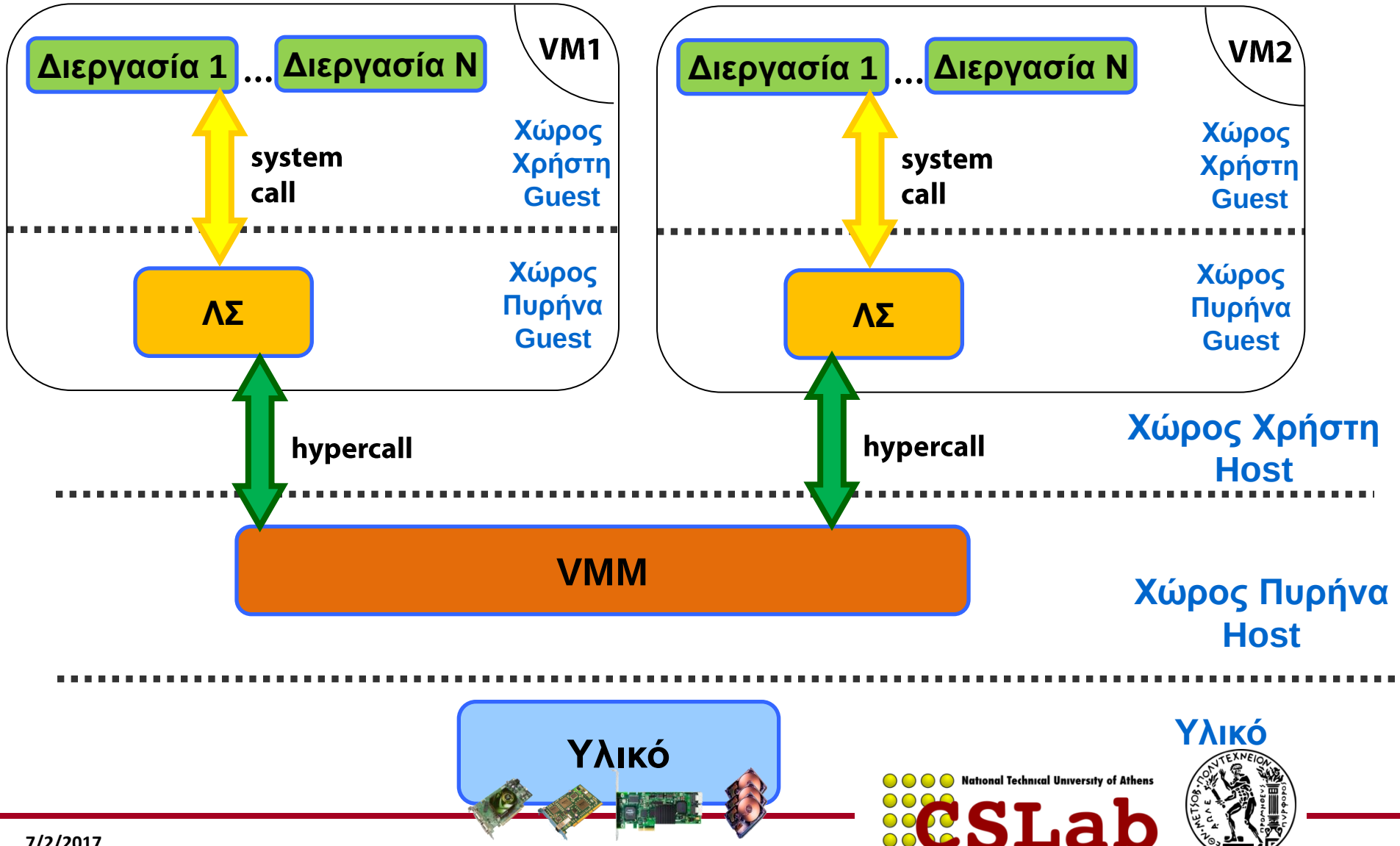
Paravirtualization (1)

- Έρχεται σε αντίθεση με τις προηγούμενες τεχνικές πλήρους εικονικοποίησης (full virtualization).

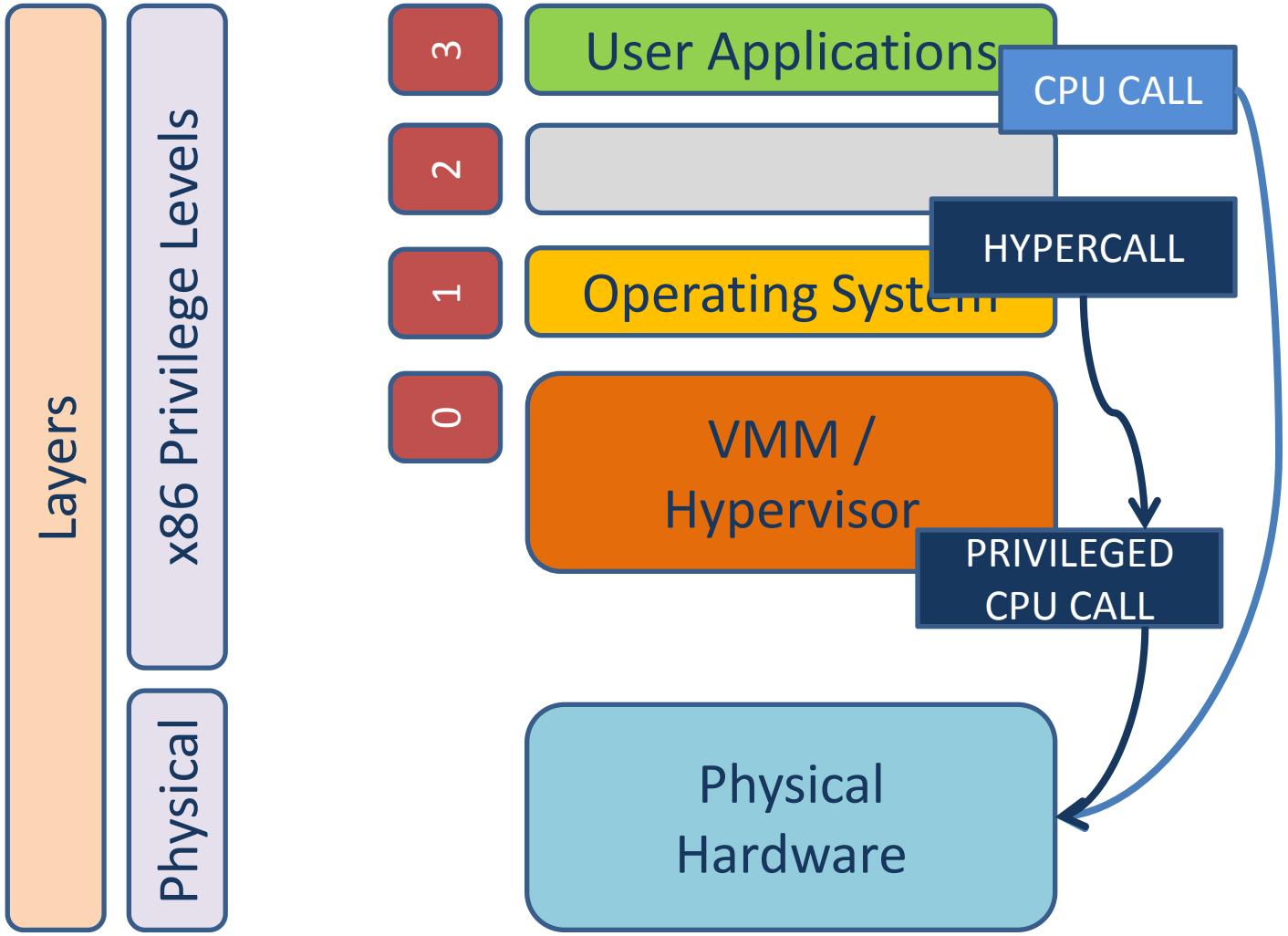
- Τμήματα του guest ΛΣ πλέον “γνωρίζουν” ότι εκτελούνται σε εικονικό περιβάλλον και πραγματοποιούν άμεσα αιτήσεις στον VMM μέσω ειδικού API, τις *υπερκλήσεις* **(hypercalls)**.

- Απαιτούνται αλλαγές στον guest πυρήνα, με σκοπό συγκεκριμένες λειτουργίες να πραγματοποιούνται ταχύτερα.

Paravirtualization (2)



Paravirtualization (3)



Κόστος Εικονικοποίησης (1)

- Εξαρτάται από το φορτίο εργασίας
  - Λιγότερες παρεμβάσεις VMM → μικρότερη επιβάρυνση
- Processor-bound εφαρμογές
  - Εκτελούνται σε εγγενείς ταχύτητες
- Συχνή κλήση system calls και προνομιούχων εντολών
  - Παρέμβαση VMM → επιβάρυνση

Κόστος Εικονικοποίησης (2)

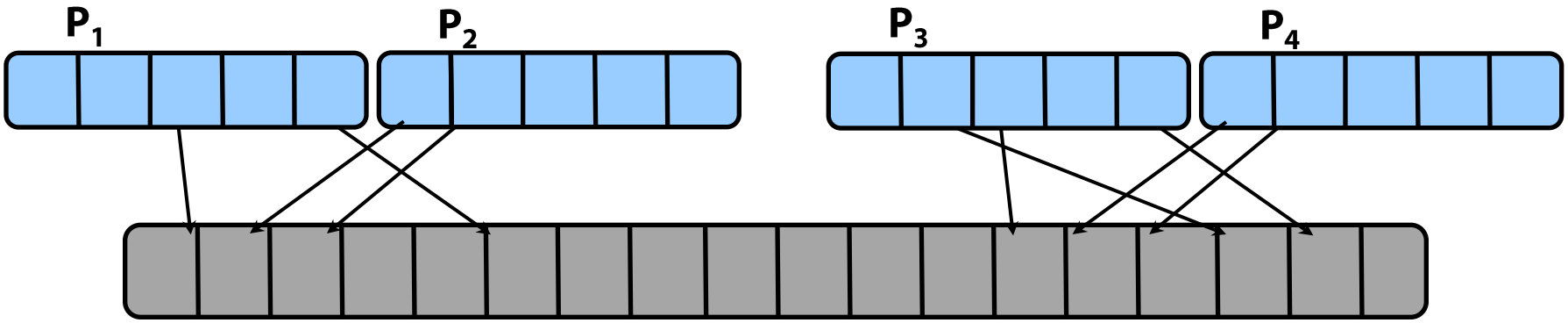
- I/O-bound εφαρμογές
  - Παρέμβαση VMM → επιβάρυνση
- Εφαρμογές με μεγάλα working sets
  - Επιπλέον επίπεδο αφαίρεσης → επιβάρυνση

  (θυμηθείτε virtual memory)

Εικονικοποίηση - Σύνοψη

- Γενικά
- Οργάνωση VMM
- Τεχνικές Εικονικοποίησης
- Εικονικοποίηση Μνήμης
- Live Migration

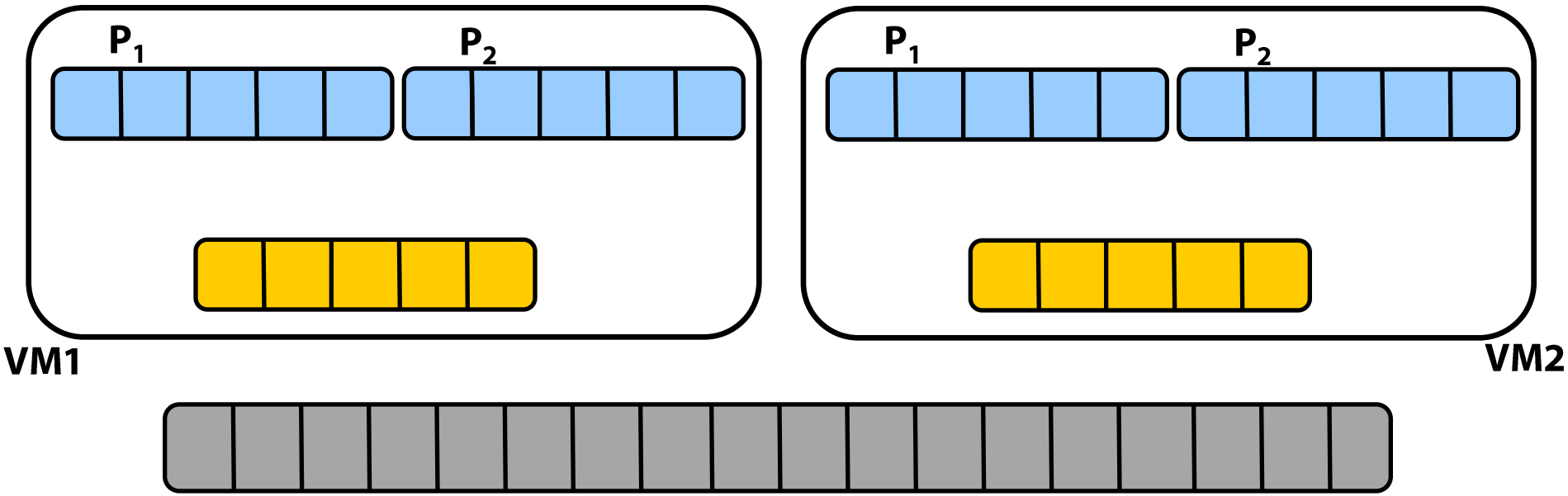
Εικονική Μνήμη



- Σε κλασσικό μη-εικονικοποιημένο σύστημα διατηρείται πίνακας σελίδων για μεταφράσεις από εικονικό σε φυσικό χώρο διευθύνσεων.

- Η MMU (Memory Management Unit) είναι υπεύθυνη για να διατρέξει τον πίνακα και να βρει την αντίστοιχη εγγραφή.

Εικονικοποίηση Μνήμης



- Σε σύστημα εικονικοποίησης πρέπει πλέον να γίνουν 2 επίπεδα μεταφράσεων:
  - Guest virtual → Guest physical
  - Guest physical → Host physical

Εικονική Μνήμη Guest

Φυσική Μνήμη Guest

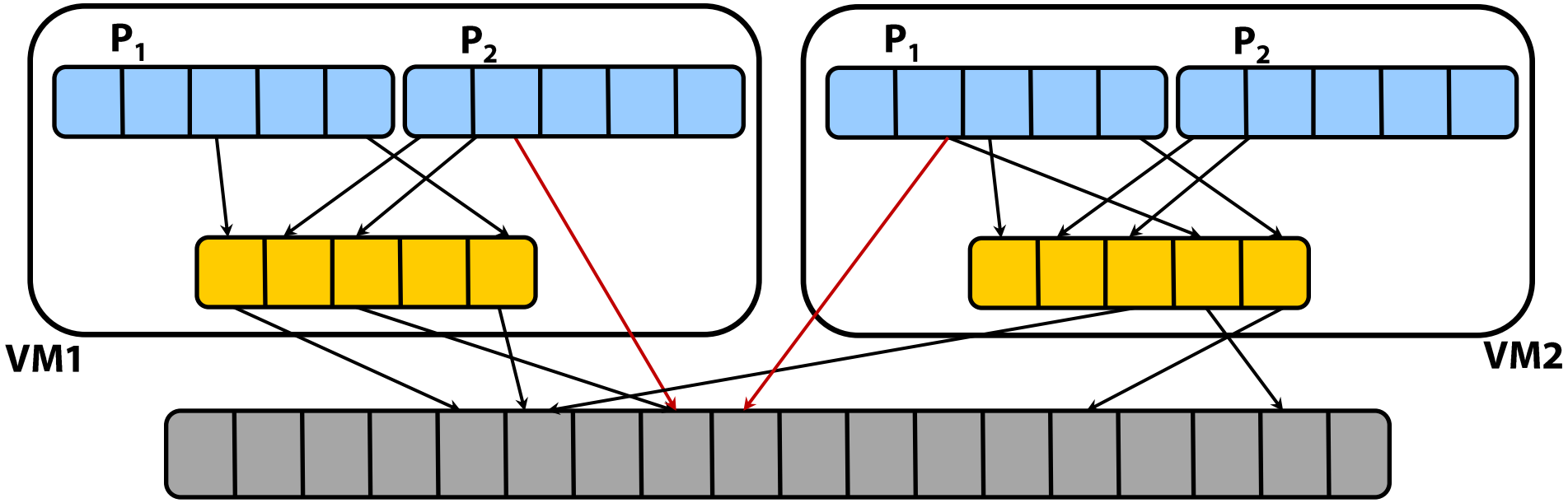
Φυσική Μνήμη Host

Δύο Τεχνικές Εικονικοποίησης Μνήμης

1. Shadow Page Tables

2. Nested Page Tables

Shadow Page Tables (1)



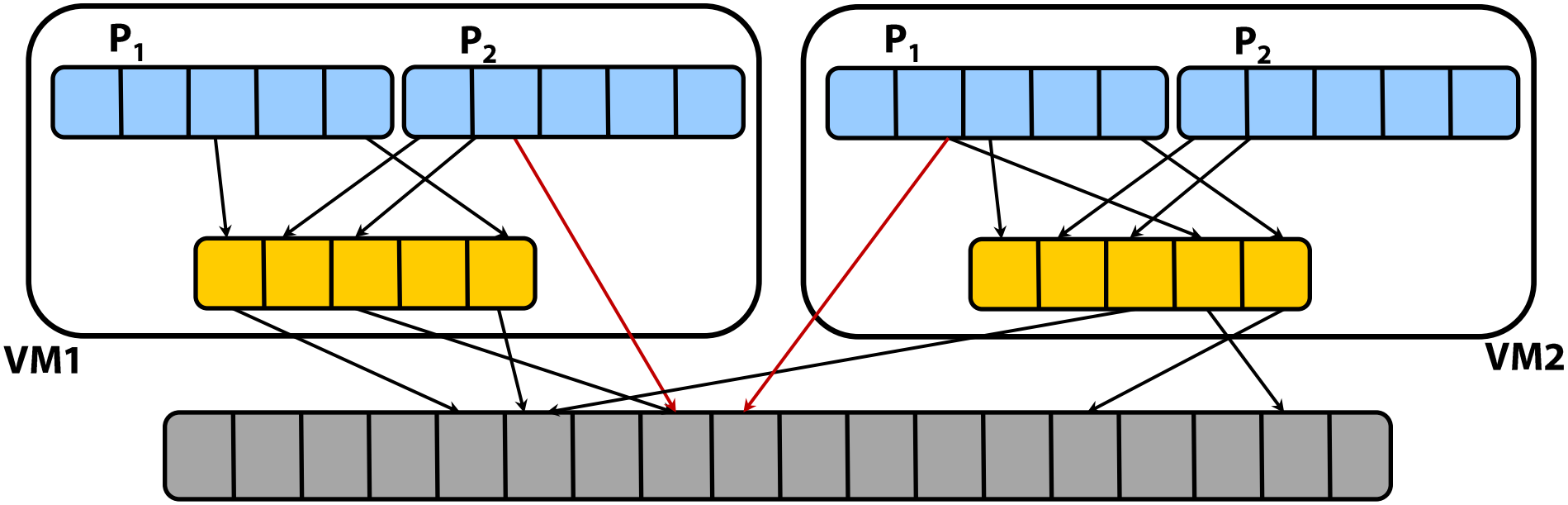
- Με τους σκιώδεις πίνακες σελίδων (shadow page tables) διατηρούνται μεταφράσεις απευθείας από το χώρο εικονικών διευθύνσεων του guest στο χώρο φυσικών διευθύνσεων του host (κόκκινο βέλος).

Εικονική Μνήμη Guest

Φυσική Μνήμη Guest

Φυσική Μνήμη Host

Shadow Page Tables (2)



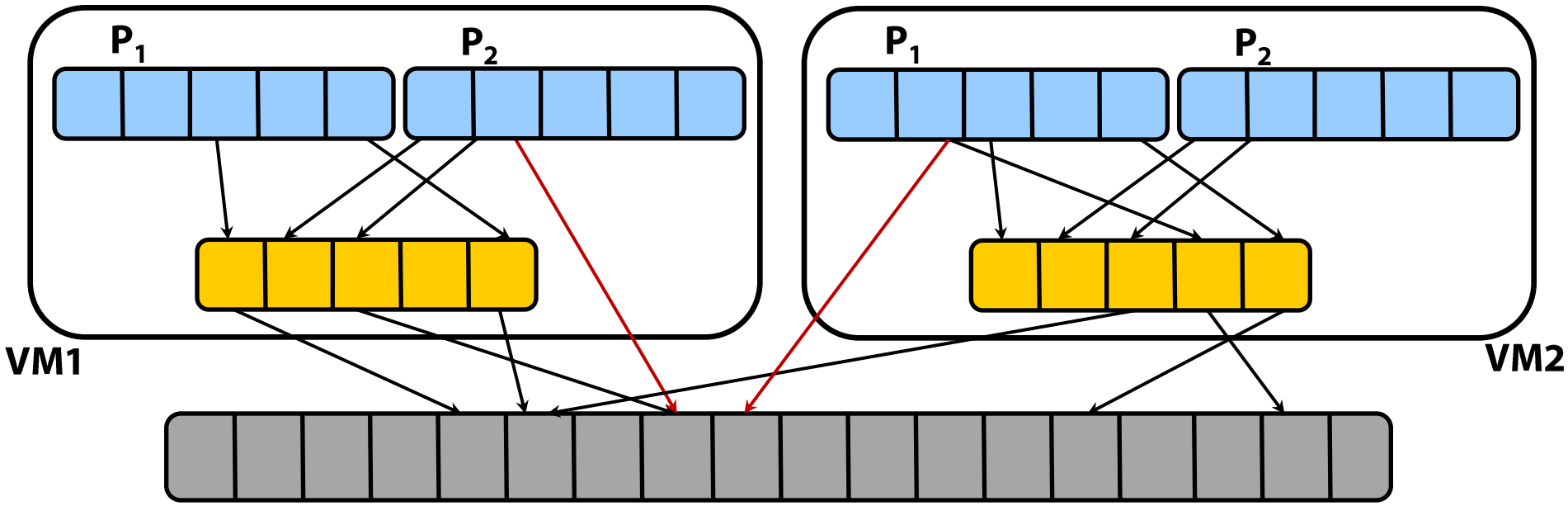
- Σφάλμα σελίδας (page fault): τα guest page faults χειρίζεται το guest ΛΣ, ενώ τα host page faults ο VMM.

Εικονική Μνήμη Guest

Φυσική Μνήμη Guest

Φυσική Μνήμη Host

Shadow Page Tables (3)



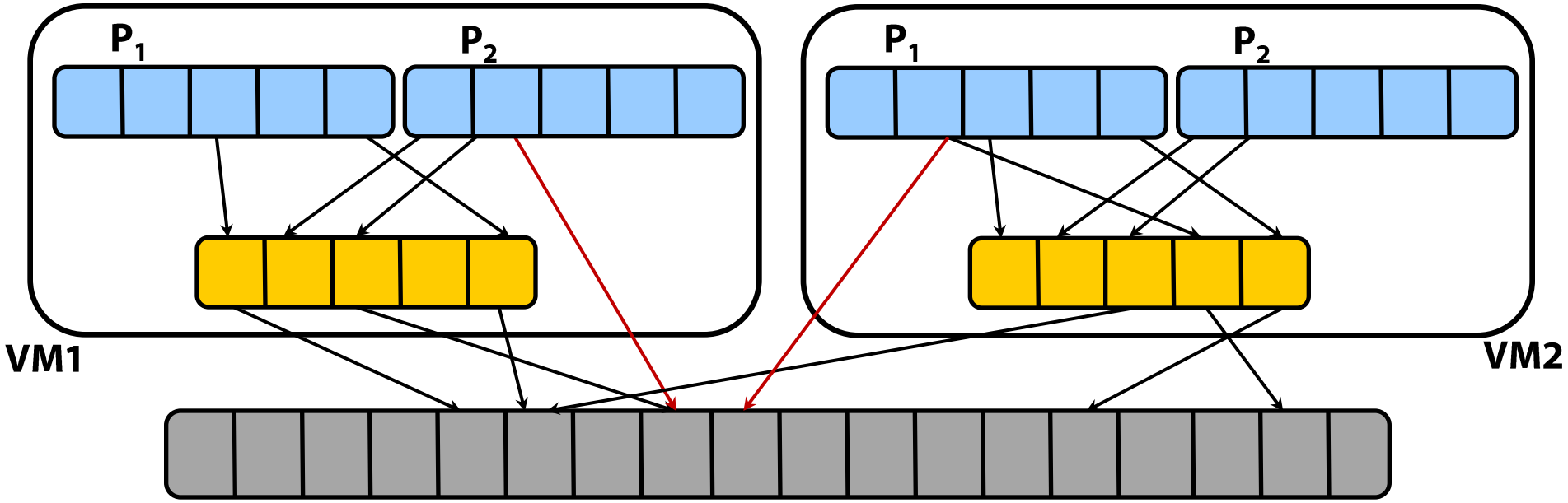
- Σε κάθε ενημέρωση εγγραφής του guest πίνακα σελίδων, πρέπει να ενημερώνεται και ο αντίστοιχος shadow table. Πώς;

Εικονική Μνήμη Guest

Φυσική Μνήμη Guest

Φυσική Μνήμη Host

Shadow Page Tables (4)



- Ο VMM "μαρκάρει" όλες τις εγγραφές του guest πίνακα σελίδων ως ***read-only***. Όταν ο guest προσπαθήσει να γράψει => page fault και ο VMM διαχειρίζεται τον αντίστοιχο shadow page table.

Εικονική Μνήμη Guest

Φυσική Μνήμη Guest

Φυσική Μνήμη Host

Shadow Page Tables (5)

- Χρησιμοποιεί έναν πίνακα σελίδων: Shadow page table
  - Guest Virtual → Host physical
- Δεν απαιτεί υποστήριξη από το υλικό

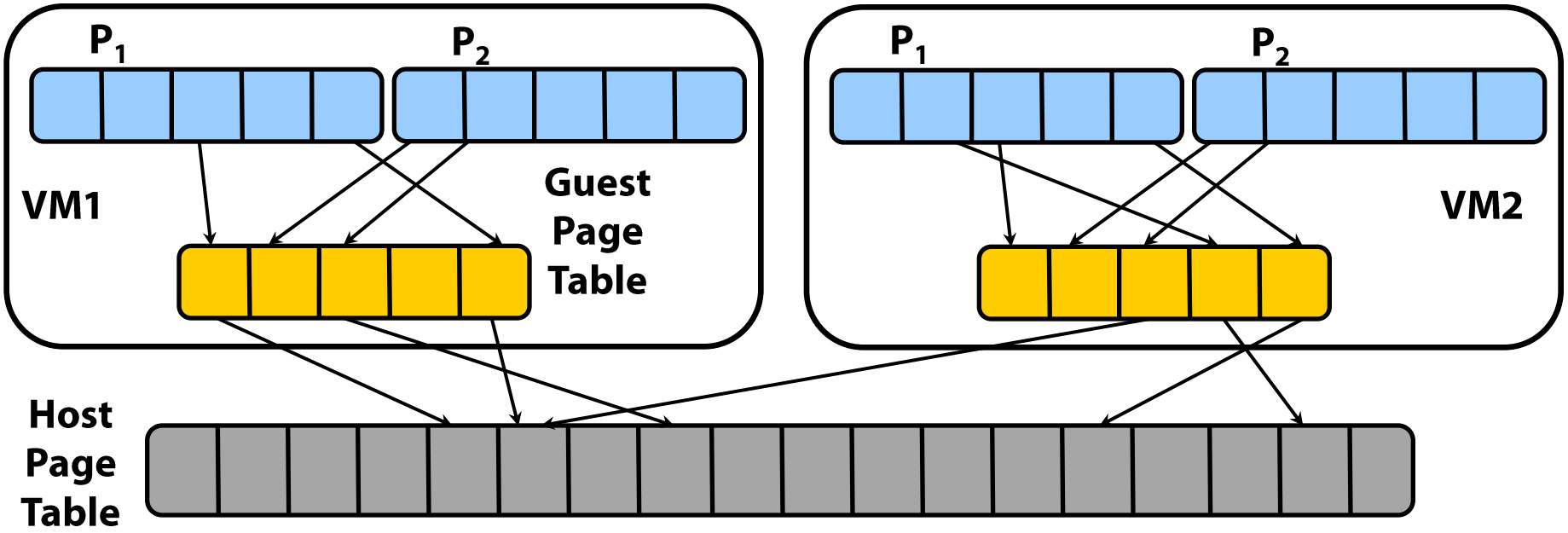
- Γρήγορα TLB misses (+)
- Αργές ενημερώσεις στον page table (-)

Δύο Τεχνικές Εικονικοποίησης Μνήμης

1. Shadow Page Tables

2. Nested Page Tables

Nested Page Tables (1)



- Χρησιμοποιεί δύο πίνακες σελίδων
- Guest page table: Guest Virtual → Guest Physical
- Host page table: Guest Physical → Host Physical

Εικονική Μνήμη Guest

Φυσική Μνήμη Guest

Φυσική Μνήμη Host

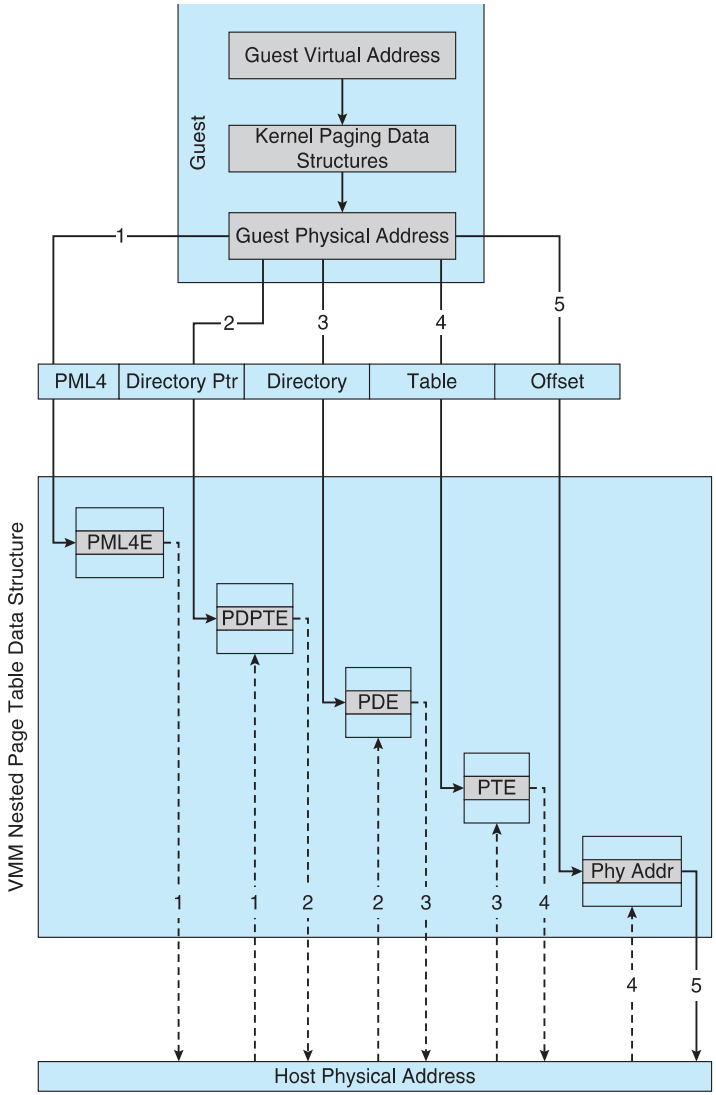
Nested Page Tables (2)

- Η τεχνική των εμφωλιασμένων πινάκων σελίδων (nested page tables) απαιτεί αλλαγές στο υλικό.
- Υποστήριξη 2 (ή και περισσότερων) επιπέδων μετάφρασης από το υλικό.
- Το υλικό αναλαμβάνει να διατρέξει τους πίνακες σελίδων, προκειμένου να βρει την τελική διεύθυνση στο φυσικό χώρο διευθύνσεων του host.

Nested Page Tables (3)

- TLB (Translation Lookaside Buffer): cache μεταφράσεων για γρήγορη αναζήτηση διευθύνσεων.
- Αυξημένο μέγεθος TLB, ώστε να αποφεύγεται η αναζήτηση στους πίνακες σελίδων (TLB miss).
- Παραδείγματα υλοποιήσεων: Intel EPT (Extended Page Tables), AMD NPT (Nested Page Tables).

Nested Page Tables (4)



Nested Page Tables (5)

- Χρησιμοποιεί δύο πίνακες σελίδων
- Guest page table
  - Guest Virtual → Guest Physical
- Host page table
  - Guest Physical → Host Physical
- Απαιτεί υποστήριξη από το υλικό

- Αργά TLB misses (-)
- Γρήγορες ενημερώσεις στον page table (+)

Εικονικοποίηση - Σύνοψη

- Γενικά
- Οργάνωση VMM
- Τεχνικές Εικονικοποίησης
- Εικονικοποίηση Μνήμης
- Live Migration

Live Migration (1)

- Μετακίνηση εικονικής μηχανής από ένα φυσικό μηχάνημα σε κάποιο άλλο.

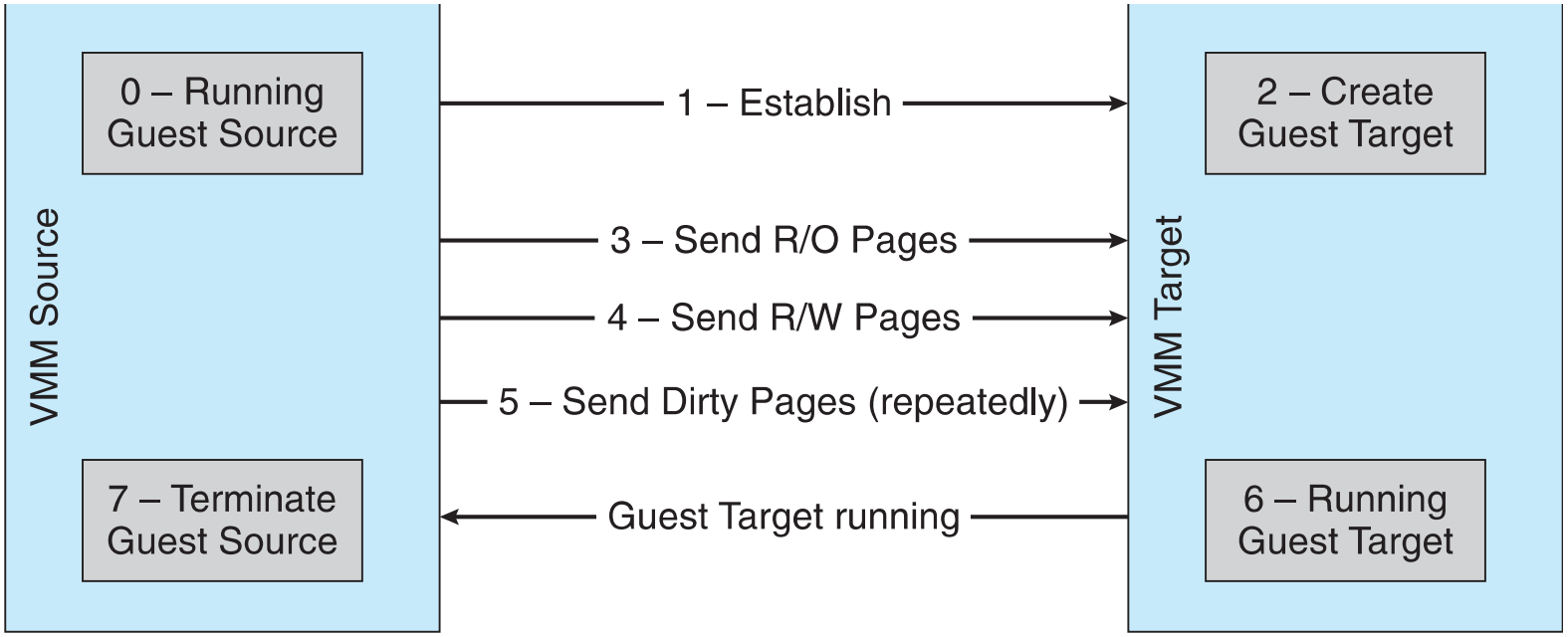
- Γενική απαίτηση: ο χρήστης να μην καταλάβει κάτι.

- Σε τι χρησιμεύει το migration;
  - Εξισσορρόπηση φόρτου μεταξύ φυσικών μηχανημάτων (π.χ. servers).
  - Συντήρηση του φυσικού μηχανήματος + πιθανή ανάγκη επανεκκίνησής του (reboot).

Live Migration (2)

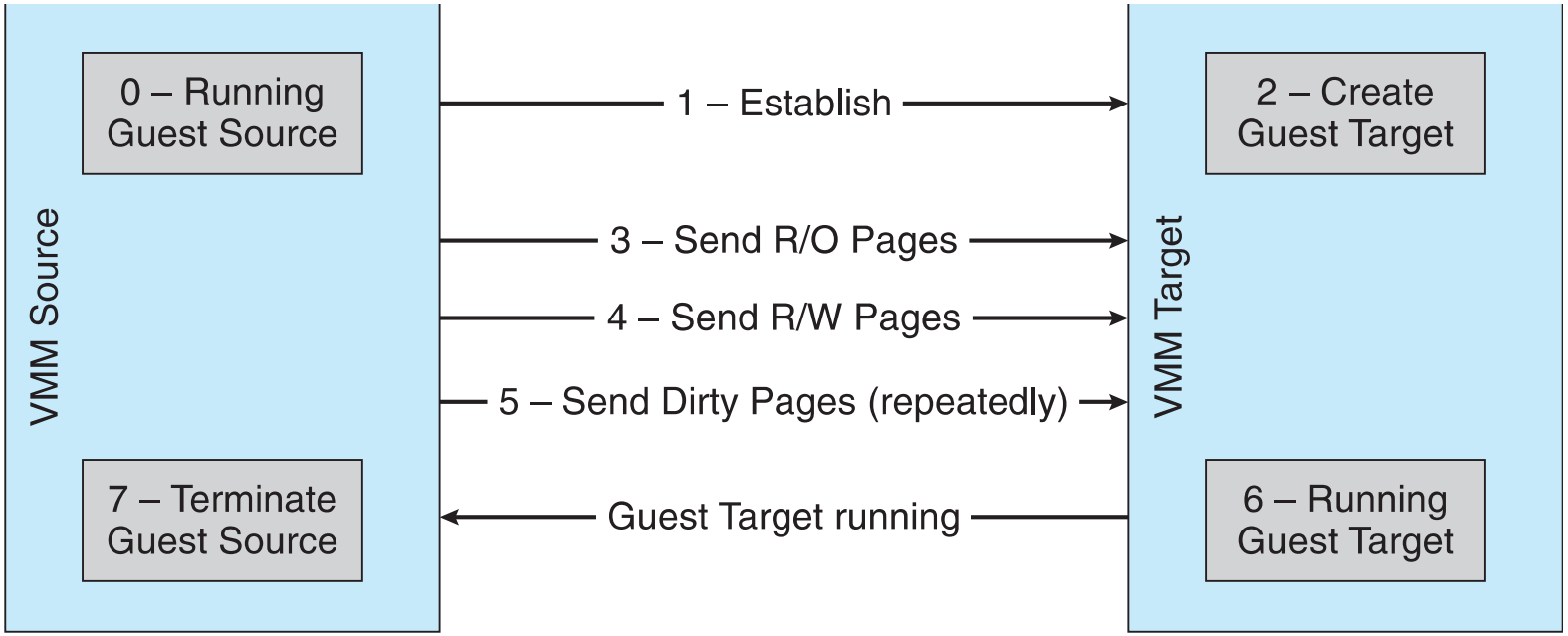
- Τι πρέπει να μεταφερθεί από το ένα φυσικό μηχάνημα προς το άλλο:
- Κατάσταση εικονικών επεξεργαστών (π.χ καταχωρητές)
  - Λύση: μικρά δεδομένα σε μέγεθος, μεταφέρονται γρήγορα.
- Εικονικοί δίσκοι αποθήκευσης
  - Λύση: μοιραζόμενο σύστημα αρχείων σε datacenters.
- Μνήμη εικονικής μηχανής

Γενικός Αλγόριθμος Live Migration (1)



- Στόχος: να μεταφερθεί η εικονική μηχανή (Guest) από τον VMM1 στον VMM2.
- Πρέπει να μεταφερθεί η μνήμη του (σελίδες).

Γενικός Αλγόριθμος Live Migration (2)



- Αποστέλλονται επανειλημμένα μόνο οι σελίδες που έχουν αλλάξει περιεχόμενο (dirty).
- Όταν ο αριθμός των dirty σελίδων γίνει αρκετά μικρός, ο VMM1 “παγώνει” τον Guest 1, αποστέλλει τις τελευταίες dirty σελίδες και εκκινεί τον Guest 2

Ερωτήσεις;