



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ
ΥΠΟΛΟΓΙΣΤΩΝ

-ΕΙΣΑΓΕΤΕ ΤΟΝ ΤΟΜΕΑ-
-ΕΙΣΑΓΕΤΕ ΤΟ ΕΡΓΑΣΤΗΡΙΟ-

-Εισάγετε τον τίτλο της διπλωματικής
εργασίας-

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

-Εισάγετε το σωστό άρθρο (της/του/των)-

-Εισάγετε το όνομα,	
αρχικό πατρώνυμου	-another author,
και επώνυμο των	whose name is here-
συγγραφέων-	

Επιβλέπων: **-Εισάγετε το όνομα, αρχικό πατρώνυμου και επίθετο-**
-Εισάγετε τον τίτλο του επιβλέποντα-

Αθήνα, -Εισάγετε το μήνα και το έτος κατάθεσης της
εργασίας-



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
-ΕΙΣΑΓΕΤΕ ΤΟΝ ΤΟΜΕΑ-
-ΕΙΣΑΓΕΤΕ ΤΟ ΕΡΓΑΣΤΗΡΙΟ-

**-Εισάγετε τον τίτλο της διπλωματικής
εργασίας-**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

-Εισάγετε το σωστό άρθρο (της/του/των)-

**-Εισάγετε το όνομα,
αρχικό πατρώνυμου
και επώνυμο των
συγγραφέων-**

Επιβλέπων: -Εισάγετε το όνομα, αρχικό πατρώνυμου και επίθετο-
-Εισάγετε τον τίτλο του επιβλέποντα-

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την -εισάγετε
ημερομηνία-.

.....
-Εισάγετε ονοματεπώνυμο- -Εισάγετε ονοματεπώνυμο- -Εισάγετε ονοματεπώνυμο-
-Εισάγετε τίτλο- -Εισάγετε τίτλο- -Εισάγετε τίτλο-

Αθήνα, (Εισάγετε το μήνα και το έτος κατάθεσης της εργα-
σίας).

.....

(Εισάγετε όνομα, αρχικό πατρώνυμου και επίθετο συγγραφέα)

(Εισάγετε τον απονεμηθέντα τίτλο του συγγραφέα)

© (Εισάγετε έτος έκδοσης) Εθνικό Μετσόβιο Πολυτεχνείο. All rights reserved.

Περίληψη

Τα τελευταία χρόνια το cloud computing αποτελεί ένα σημαντικό κεφάλαιο στη σύγχρονη επιστήμη υπολογιστών. Η κύρια τεχνολογία που χρησιμοποιείται, προκειμένου να μπορεί να υποστηριχθεί το cloud computing είναι αυτή της εικονικοποίησης. Με αυτό τον τρόπο ένα φυσικό μηχάνημα, μπορεί να φιλοξενήσει πολλά εικονικά μηχανήματα, κάθε ένα από τα οποία αποτελεί έναν αυτοδύναμο υπολογιστή. Ωστόσο, αποτελεί συχνό φαινόμενο οι εικονικές αυτές μηχανές να χρησιμοποιούνται για την εκτέλεση μίας και μόνο εφαρμογής. Αυτό έχει ως αποτέλεσμα, να χαρραμίζονται πόροι σε ενέργειες που δε χρειάζονται από την εφαρμογή, αλλά είναι απαραίτητες για το λειτουργικό σύστημα στο οποίο τρέχουν αυτές οι εφαρμογές.

Μία νεότερη τάση για την υποστήριξη του cloud computing είναι τα containers, που προσφέρουν ελαφρύτερη εικονικοποίηση με γρηγορους χρόνους εκτέλεσης, μικρή κατανάλωση μνήμης και άλλα πλεονεκτήματα. Από την άλλη, παρουσιάζουν αρκετά σημαντικά ζητήματα που αφορούν την ασφάλεια. Ένα από τα ζητήματα αυτά είναι, εκείνο της απομόνωσης το οποίο αναγκάζει σε αρκετές περιπτώσεις να οδηγεί στη χρήση εικονικών μηχανών για τη φιλοξενία των containers, χάνοντας αρκετά από τα πλεονεκτήματά τους.

Μία ακόμη προσέγγιση στο θέμα είναι οι unikernels. Πρόκειται για μία εικόνα εικονικής μηχανής, με ένα μόνο address space το οποίο κατασκευάζεται από library operating systems και είναι ειδικευμένο για μία συγκεκριμένη εφαρμογή. Πιο απλά, περιέχει τον κώδικα της εφαρμογής και ακριβώς ό,τι κομμάτι του λειτουργικού συστήματος χρειάζεται η εφαρμογή για να λειτουργήσει η διεργασία (drivers, βιβλιοθήκες, κ.λ.π.), ενοποιημένα σαν ένα αυτόνομο πρόγραμμα που μπορεί να τρέξει ως εικονική μηχανή. Οι unikernels καταφέρνουν να έχουν γρήγορους χρόνους εκκίνησης και μικρή κατανάλωση μνήμης, χωρίς να θυσιάζεται η ασφάλεια. Εν τούτοις, ένα πρόβλημα είναι ότι οι unikernels υποστηρίζουν μία και μόνο διεργασία, με αποτέλεσμα να μην μπορούν εφαρμογές με παραπάνω από μία διεργασίες να εκτελεστούν σε unikernels.

Σκοπός, λοιπόν, αυτής της εργασίας είναι η υλοποίηση ενός μηχανισμού που θα επιτρέπει σε εφαρμογές με περισσότερες από μία διεργασίες να μπορούν να εκτελεστούν και σε unikernels. Επιπλέον, υλοποιείται και ένας απλός μηχανισμός για επικοινωνία μεταξύ των εικονικών μηχανών, στα πρότυπα του pipe.

Λέξεις-Κλειδιά: εικονικοποίηση, εικονικές μηχανές, ενδοεπικοινωνία εικονικών μηχανών, unikernel, kvm, QEMU

Abstract

In recent years cloud computing is one important chapter of modern computer science. The main technology used in order to support cloud computing is virtualization. Virtualization makes possible for a physical machine to host many virtual machines, each one of which is a self-sufficient computer. However, virtual machines are often used to execute a single application. As a result, resources are devoted to actions that are not needed by the application, but they are necessary for the operating system in which the applications run.

Another technology to support cloud computing is containers which offer lightweight virtualization, fast instantiation times and small per-instance memory footprints among other features. On the other hand, containers have several important security issues. Isolation is one of these issues, which in several cases leads to the use of virtual machines to host the containers, losing several of their advantages.

A further approach in cloud computing is unikernels. Unikernels are specialised, single-address-space machine images constructed by using library operating systems and are specialised for one application. In somewhat simplified terms, unikernels consist of the application's source code and the parts of an operating system that are necessary for the process to run (drivers, libraries, etc.) consolidated as a stand-alone virtual machine. The Unikernels manage to have fast instantiation times, small memory footprints, without sacrificing security. However, one of the problems of unikernels is that they are single-process and as a result multi-process applications are not able to run on unikernels.

The purpose of this thesis is to implement a mechanism that will enable the execution of multi-process applications on unikernels. Furthermore, a pipe-like mechanism for inter-vm communication is implemented.

Keywords: virtualization, virtual machines, inter-vm communication, unikernel, kvm, QEMU

Περιεχόμενα

1	Εισαγωγή	3
1.1	Σκοπός	4
1.2	Οργάνωση Κειμένου	5
2	Θεωρητικό Υπόβαθρο	6
2.1	Cloud computing	6
2.1.1	Χαρακτηριστικά του cloud computing	6
2.1.2	Μοντέλα υπηρεσιών	7
2.2	Εικονικοποίηση	8
2.2.1	Εικονικοποίηση σε επίπεδο λειτουργικού συστή- ματος	9
2.2.2	Εικονικοποίηση σε επίπεδο υλικού	10
2.2.3	QEMU/KVM	14
2.3	Λειτουργικά συστήματα	15
3	Unikernel frameworks	18
3.1	Rumpun	20
3.2	OSv	22
3.3	Linux Kernel Library - lkl	24
3.4	IncludeOS	26
3.5	ClickOS	28
3.6	MirageOS	29
3.7	Solo5	31
3.8	Mini-OS	32
3.9	Και πολλά ακόμα	33
4	Σχεδιασμός και υλοποίηση	34
4.1	Μηχανισμός pipe	35
4.1.1	Στάδια υλοποίησης	35
4.1.2	Στάδια υλοποίησης	51
5	Επίλογος	53
5.1	Mpla	53
5.2	Mpla	53
5.2.1	Mpla	53

Κατάλογος σχημάτων

2.1	Cloud services	8
2.2	Εικονικοποίηση σε επίπεδο λειτουργικού συστήματος .	10
2.3	Εικονικοποίηση σε επίπεδο υλικού	12
3.1	Unikerne vs OS software stack	19
3.2	Rumprun software stack	21
3.3	Σχέση μεταξύ των εννοιών anykernel, rump kernel και rumprun	22
3.4	Osn unikernel stack	23
3.5	Η αρχιτεκτονική του lkl	25
3.6	IncludeOS building process	27
3.7	Clickos design	29
3.8	mirage compiler	30
3.9	a) γενικού σκοπού monitor b) ειδικού σκοπού monitor .	32
4.1	Πρώτο στάδιο υλοποίησης μηχανισμού pipe	36
4.2	Αλληλουχία κλήσεων για επικοινωνία μέσω TCP/IP sockets	37
4.3	Δεύτερο στάδιο υλοποίησης μηχανισμού pipe	38
4.4	Τρίτο στάδιο υλοποίησης μηχανισμού pipe	41
4.5	shared memory layout	46
4.6	pipe read flow chart	50
4.7	pipe write flow chart	52

Κεφάλαιο 1

Εισαγωγή

Στις μέρες μας το cloud computing έχει γίνει ένα από τα ταχύτερα αναπτυσσόμενα και ενδιαφέροντα θέματα στην επιστήμη των υπολογιστών. Το cloud computing δίνει τη δυνατότητα σε απομακρυσμένους χρήστες να αποκτάνε πρόσβαση σε υπολογιστικούς πόρους (αποθηκευτικός χώρος, εφαρμογές, υπηρεσίες κ.λ.π.) όταν χρειαστούν από τους χρήστες. Ιδιαίτερα, τα τελευταία χρόνια το cloud έχει μπει και στις ζωές των απλών χρηστών. Η χρήση των προσωπικών υπολογιστών αρχίζει να αλλάζει σημαντικά, καθώς πλέον προγράμματα και δεδομένα απομακρύνονται από τους προσωπικούς υπολογιστές για να εκτελεστούν και να αποθηκευτούν στο λεγόμενο cloud.

Μία από τις βασικές τεχνολογίες που κρύβονται πίσω από το cloud είναι αυτή της εικονικοποίησης. Χάρη την εικονικοποίηση, μπορούμε σε ένα φυσικό μηχάνημα να φιλοξενήσουμε πολλά εικονικά μηχανήματα κάθε ένα από τα οποία είναι ανεξάρτητο. Με αυτό τον τρόπο, μπορούμε να αξιοποιήσουμε καλύτερα τους φυσικούς πόρους του μηχανήματος και να τους διαμοιράσουμε όπως επιθυμούμε μεταξύ των εικονικών μηχανών. Ακόμα, η εικονικοποίηση δημιουργήσε και κάποιες νέες δυνατότητες, όπως αυτή της μεταφοράς εικονικών μηχανών σε διαφορετικό φυσικό μηχάνημα, η δημιουργία αντιγράφων εικονικών μηχανών, αλλά και περισσότερη ασφάλεια, αφού ένα πρόβλημα σε ένα εικονικό μηχάνημα δε θα επηρεάσει ούτε το φυσικό ούτε τα υπόλοιπα εικονικά μηχανήματα.

Ένα συχνό φαινόμενο κατά τη χρήση της εικονικοποίησης, είναι η χρήση μία εικονικής μηχανής με συμβατικά λειτουργικά συστήματα για την υποστήριξη μίας και μόνο υπηρεσίας. Όπως είναι φυσικό το λειτουργικό σύστημα μέσα στην εικονική μηχανή χρειάζεται κάποιους πόρους για να λειτουργήσει, ενώ συχνά μπορεί να εκτελεί λειτουργίες οι οποίες δε χρειάζονται από την εφαρμογή. Ακόμα, ο κώδικας όλου του λειτουργικού συστήματος αυξάνει αρκετά και το μέγεθος σε μνήμη της εικονικής μηχανής. Γίνεται λοιπόν, κατανοητό ότι σπαταλούνται πόροι, οι οποίοι θα μπορούσα να διατεθούν είτε στην ίδια την υπηρεσία είτε σε κάποια άλλη.

Για την επίλυση του παραπάνω προβλήματος αναζητήθηκαν λύσεις για να γίνει η εικονικοποίηση πιο ελαφρύα, αλλά και να μη σπαταλούνται πόροι σε αχρείαστες λειτουργίες. Μία από αυτές τις λύσεις, ήταν η εικονικοποίηση σε επίπεδο λειτουργικού συστήματος ή διαφορετικά containerization. Με τη συγκεκριμένη μέθοδο ο πυρήνας επιτρέπει την ύπαρξη πολλαπλών απομονωμένων user-space instances, που ονομάζονται containers. Τα containers μοιράζονται μεταξύ τους το λειτουργικό σύστημα στα οποία εκτελούνται, ενώ ταυτόχρονα παρέχουν ένα απομονωμένο περιβάλλον για τις διεργασίες μέσα σε αυτό.

Η τεχνολογία των containers, όπως κάθε άλλη τεχνολογία δε θα μπορούσε να μην έχει και κάποια μειονεκτήματα. Ένα από τα κύρια μειονεκτήματα, είναι αυτό της ασφάλειας. Τα containers μοιράζονται τον ίδιο πυρήνα του host, ενώ οι μηχανισμοί απομόνωσης δεν είναι το ίδιο ισχυροί με αυτούς στα εικονικά μηχανήματα. Μάλιστα, έχουν αναφερθεί περιπτώσεις όπου από ένα container μπορούσαν να παρθούν πληροφορίες και δεδομένα τόσο για το host, όσο και για άλλα containers [9]. Όλα αυτά έχουν οδηγήσει σε περιπτώσεις όπου τα containers χρησιμοποιούνται πάνω από ένα πλήρη λειτουργικό σύστημα το οποίο τρέχει μέσα σε μία εικονική μηχανή.

Μία ιδέα που αρχίζει να αποκτά όλο και περισσότερο ενδιαφέρον και προσοχή είναι αυτή των unikernels. Η βάση της ιδέας, είναι η κατασκευή ειδικευμένων εικονικών μηχανών για κάθε ξεχωριστή υπηρεσία. Στην εικονική αυτή μηχανή δε χρειάζεται να υπάρχει ένα πλήρη λειτουργικό σύστημα, αλλά μόνο τα κομμάτια αυτού τα οποία είναι απαραίτητα για την εκτέλεση της υπηρεσίας. Με αυτό τον τρόπο, μειώνεται σημαντικά το μέγεθος των εικονικών μηχανών, ενώ πλέον οι πόροι χρησιμοποιούνται ακριβώς για τις λειτουργίες που χρειάζεται η υπηρεσία. Τέλος, εφόσον αναφερόμαστε σε εικονικές μηχανές, η απομόνωση τους είναι κάτι το οποίο έχουν φροντίσει ήδη οι ελεγκτές.

1.1 Σκοπός

Η φιλοσοφία των unikernels είναι ότι κάθε εικονική μηχανή θα έχει ένα συγκεκριμένο σκοπό. Για το λόγο αυτό οι unikernels δεν υποστηρίζουν παραπάνω από μία διεργασία σε κάθε εικονική μηχανή. Από την άλλη, υπάρχουν unikernel frameworks τα οποία επιτρέπουν την υποστήριξη περισσότερων από ένα νήμα. Ωστόσο οι περισσότερες εφαρμογές και υπηρεσίες στο cloud είναι φτιαγμένες για ένα πλήρη λειτουργικό σύστημα. Επομένως, προκειμένου να μπορούν να τρέξουν σε ένα unikernel θα πρέπει να αλλάξει άλλοτε περισσότερο και άλλοτε λιγότερο ο σχεδιασμός τους. Ιδιαίτερα, οι εφαρμογές που χρησιμοποιούν παραπάνω από μία διεργασίες θα πρέπει να αλλάξουν

σε ένα μοντέλο με μία μόνο διεργασία.

Ο σκοπός της συγκεκριμένης εργασίας είναι να υλοποιήσει, κρατώντας το `single-process` χαρακτηριστικό των `unikernels`, ένα μηχανισμό για την υποστήριξη των εφαρμογών που απαιτούν παραπάνω από μία διεργασία.

Αρχικά έγινε μία μελέτη γύρω από τα υπάρχοντα `unikernel frameworks`, παρατηρώντας τα χαρακτηριστικά του κάθε ενός. Στη συνέχεια σε ένα από αυτά τα `unikernels` (`gump-run`) υλοποιήθηκαν οι δύο παρακάτω λειτουργίες:

- ένας μηχανισμός για επικοινωνία μεταξύ των `vms`, ο οποίος είναι στα πρότυπα της κλήσης συστήματος `pipe` (`POSIX`). Ουσιαστικά πρόκειται για την υλοποίηση της `pipe` σε επίπεδο εικονικών μηχανών.
- ένας μηχανισμός για την υποστήριξη της κλήσης συστήματος `fork` από τους `unikernels`. Όταν μία εφαρμογή θα χρησιμοποιεί τη συγκεκριμένη κλήση συστήματος, θα δημιουργείται μία νέα εικονική μηχανή κλώνος της αρχικής και θα ξεκινά την εκτέλεση της, ακριβώς μετά την κλήση συστήματος `fork`.

1.2 Οργάνωση Κειμένου

Στη συνέχεια παρουσιάζεται αναλυτικά η περιγραφή του σχεδιασμού και της υλοποίησης των δύο λειτουργιών `pipe`, `fork`, γίνεται μία ανασκόπηση στα ήδη υπάρχοντα `unikernel frameworks` και περιγράφεται το απαραίτητο θεωρητικό υπόβαθρο.

Πιο συγκεκριμένα, στο Κεφάλαιο 2 καλύπτεται το απαραίτητο θεωρητικό υπόβαθρο, κάνοντας μία μικρή αναφορά για το `cloud computing`, μία εισαγωγή στην εικονικοποίηση και στα λειτουργικά συστήματα.

Στο Κεφάλαιο 3 παρουσιάζονται τα `unikernel frameworks` που μελετήθηκαν, τα χαρακτηριστικά τους, η φιλοσοφία τους κ.λ.π.

Στο Κεφάλαιο 4 γίνεται αναλυτική περιγραφή του σχεδιασμού και της υλοποίησης των δύο λειτουργιών που αναφέρθηκαν προηγουμένως στο σκοπό της εργασίας (`pipe`, `fork`).

Τέλος, στο Κεφάλαιο 5 αναφέρονται τα τελικά συμπεράσματα της παρούσας μελέτης όπως επίσης και πιθανές μελλοντικές επεκτάσεις αυτής της διπλωματικής εργασίας.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Cloud computing

Εδώ και αρκετό καιρό γίνεται συχνά αναφορά στο "Cloud", αλλά και σε αυτή την εργασία έχει αναφερθεί αρκετές φορές ο όρος "cloud computing". Παρά τη δημοφιλία του όρου, δεν υπάρχει ακριβής ορισμός. Εντούτοις, θα χρησιμοποιηθεί ο ορισμός που έχει δωθεί από το NIST (National Institute of Standards and Technology of USA) [20]. Το cloud computing (υπολογιστικό νέφος με ελληνικούς όρους) είναι η κατ' αίτηση διαδικτυακή κεντρική διάθεση υπολογιστικών πόρων (όπως δίκτυο, εξυπηρετητές, εφαρμογές και υπηρεσίες) με υψηλή ευελιξία, ελάχιστη προσπάθεια από τον χρήστη και υψηλή αυτοματοποίηση. Ωστόσο, είναι συχνό φαινόμενο να αναφερόμαστε με το συγκεκριμένο όρο τόσο στις υπηρεσίες που είναι διαθέσιμες μέσω διαδικτύου με τη μορφή μίας υπηρεσίας ιστού όσο και στο υλικό και λογισμικό που απαρτίζουν την υποδομή που προσφέρει αυτές τις υπηρεσίες.

2.1.1 Χαρακτηριστικά του cloud computing

Πέρα από τον ορισμό το NIST [20], περιγράφει και τα ακόλουθα βασικά χαρακτηριστικά του cloud computing:

- παροχή υπηρεσίας κατ' απαίτηση: Οι χρήστες έχουν τη δυνατότητα να τροποποιήσουν τους πόρους που προσφέρονται, χωρίς την ανθρώπινη διαμεσολάβηση από την πλευρά του παρόχου του cloud.
- ευρυζωνική δικτυακή πρόσβαση: Οι υπηρεσίες είναι διαθέσιμες από το διαδίκτυο.
- Διάθεση πόρων σε περισσότερους χρήστες: Οι πόροι μπορούν

να διατεθούν ανάμεσα σε πολλούς χρήστες χωρίς να δημιουργούνται προβλήματα.

- ταχεία ελαστικότητα/επεκτασιμότητα: Οι πόροι μπορούν να ανακαταμεθθούν, είτε αυτόματα είτε κατ' απαίτηση, χωρίς περιορισμούς και χωρίς να επηρεάζεται η ομαλή λειτουργία άλλων υπηρεσιών.
- μετρήσιμη υπηρεσία: Η χρήση των πόρων καταγράφεται και παρουσιάζεται τόσο στον πάροχο όσο και στον πελάτη.

2.1.2 Μοντέλα υπηρεσιών

Σύμφωνα με το NIST [20], τα μοντέλα υπηρεσιών του cloud είναι τα εξής:

Λογισμικό ως Υπηρεσία (SaaS)

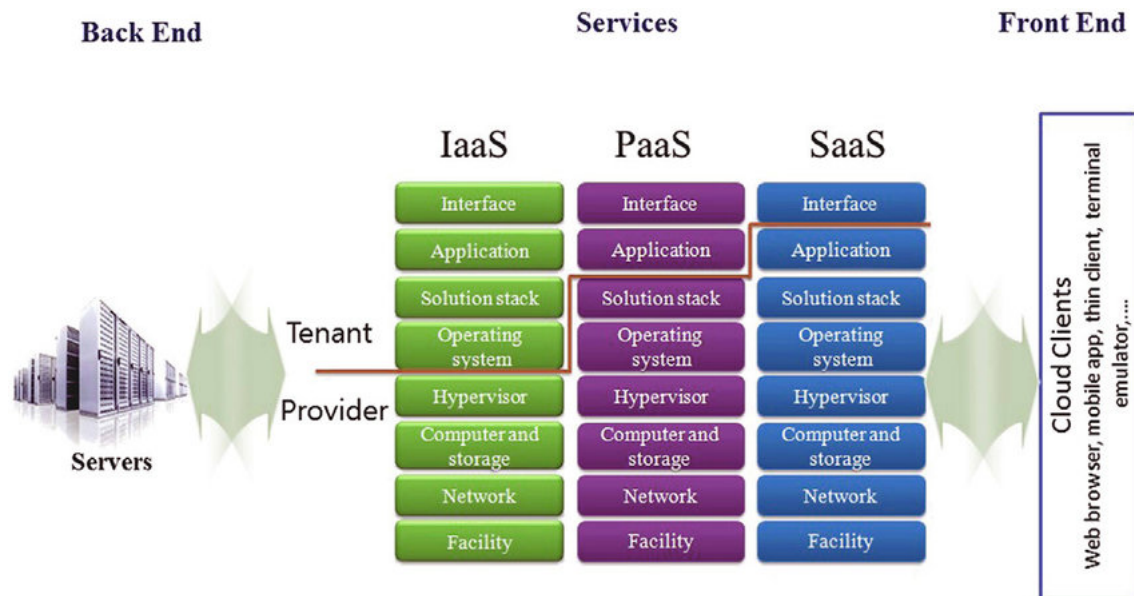
Η υπηρεσία που παρέχεται στο χρήστη είναι αυτή μία εφαρμογή που τρέχει στην υποδομή του παρόχου. Ο χρήστης δε χρειάζεται να γνωρίζει τίποτα σχετικά με την υποδομή που υποστηρίζει την εφαρμογή, ούτε έχει τη δυνατότητα να τροποποιήσει πολλά πράγματα πέρα ίσως από κάποιες ρυθμίσεις που προσφέρει η εφαρμογή. Οι εφαρμογές αυτές είναι προσβάσιμες μέσα από κάποιο λογισμικό πελάτη που προσφέρει ο πάροχος ή ακόμα και από ένα πειληγητή ιστού (web browser).

Πλατφόρμα ως υπηρεσία (PaaS)

Σε αυτή την περίπτωση ο χρήστης έχει τη δυνατότητα της δημιουργίας εφαρμογών ή/και περιβάλλοντων εφαρμογών, χρησιμοποιώντας προγραμματιστικά εργαλεία και υπηρεσίες που παρέχονται από τον πάροχο. Ο χρήστης δεν μπορεί να ελέγξει και να διαχειριστεί την υποδομή του cloud, ωστόσο έχει τον έλεγχο των αναπτυγμένων εφαρμογών και ενδεχομένως των ρυθμίσεων διαμόρφωσης για το περιβάλλον φιλοξενίας αυτών των εφαρμογών.

Υποδομή ως υπηρεσία (IaaS)

Οι παροχές προς το χρήστη είναι υπολογιστικοί πόροι όπως επεξεργαστές, αποθηκευτικός χώρος και δίκτυο στα οποία μπορεί αναπτύξει και να τρέξει λειτουργικά συστήματα και εφαρμογές. Παρά τη δυνατότητα να τροποποιήσει την ποσότητα των πόρων, ο χρήστης δεν έχει παραπάνω έλεγχο ως προς την υποδομή του cloud.



Σχήμα 2.1: Cloud services

Όπως βλέπουμε και από το σχήμα 2.1, όσο περισσότερη ευελιξία προσφέρει ένα μοντέλο τόσο περισσότερο έλεγχο έχει και ο χρήστης στις υπηρεσίες που λαμβάνει. Για παράδειγμα στην περίπτωση SaaS, ο χρήστης έχει το λιγότερο έλεγχο, σε αντίθεση με την περίπτωση IaaS όπου ο χρήστης μπορεί πλέον να αποκτήσει έλεγχο ακόμα και στο λειτουργικό σύστημα.

2.2 Εικονικοποίηση

Στην επιστήμη των υπολογιστών ο όρος εικονικοποίηση (virtualization) περιγράφει ένα μηχανισμό αφαίρεσης, που με τη χρήση "εικονικών υπολογιστικών πόρων", έχει ως σκοπό την απόκρυψη λεπτομερειών για την υλοποίηση ή την κατάσταση των φυσικών πόρων. Με το μηχανισμό αυτό ένας φυσικός πόρος μπορεί να παρουσιαστεί ως μία πλειάδα εικονικών φυσικών πόρων, ή αντίστροφα μία πλειάδα φυσικών πόρων να παρουσιαστούν ως ένας ενιαίος εικονικός πόρος. Ανάλογα σε ποιο επίπεδο στη στοίβα του λογισμικού υλοποιείται η εικονικοποίηση, υπάρχουν και τα αντίστοιχα πλεονεκτήματα, αλλά σε γενικές γραμμές κάποια από τα πλεονεκτήματα της εικονικοποίησης είναι:

- Καλύτερη αξιοποίηση και διαμοιρασμός των πόρων.
- Ασφάλεια μέσω της απομόνωσης των υπηρεσιών.

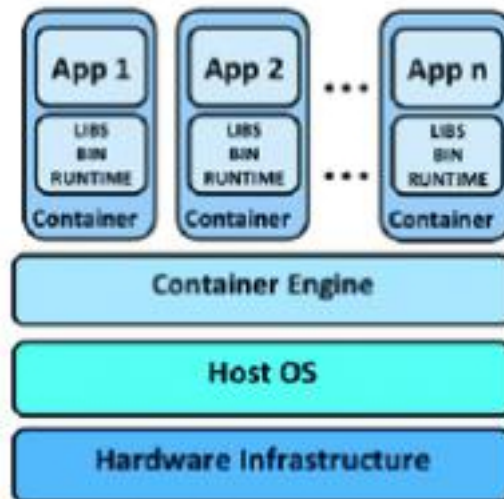
- Δυνατότητα προσομοίωσης περιβαλλόντων που δεν είναι φυσικά διαθέσιμα.
- Δυνατότητα αποθήκευσης, μεταφοράς και επαναφοράς της κατάστασης των υπηρεσιών που προσφέρονται μέσω της εικονικοποίησης.

Η εικονικοποίηση μπορεί να υλοποιηθεί σε διάφορα επίπεδα, όπως στο δίκτυο, στον αποθηκευτικό χώρο, στο hardware, στο λειτουργικό σύστημα κ.α. Η συγκεκριμένη εργασία ασχολείται κυρίως με την εικονικοποίηση σε επίπεδο υλικού, ωστόσο είναι συχνό φαινόμενο να συγκρίνονται οι unikernels με τα containers. Για το λόγο αυτό θα γίνει μία σύντομη παρουσίαση της εικονικοποίησης σε επίπεδο λειτουργικού συστήματος, ενώ μετά από αυτή την περιγραφή, όταν χρησιμοποιείται ο όρος εικονικοποίηση θα εννοείται εικονικοποίηση επιπέδου υλικού.

2.2.1 Εικονικοποίηση σε επίπεδο λειτουργικού συστήματος

Η εικονικοποίηση σε επίπεδο λειτουργικού συστήματος είναι νεότερη σε σχέση με αυτή σε επίπεδο υλικού, ωστόσο έχει καταφέρει να συγκεντρώσει αρκετή προσοχή τα τελευταία χρόνια. Ο πυρήνας του λειτουργικού συστήματος επιτρέπει την ύπαρξη παραπάνω από ένα, απομονωμένα userspace instances τα οποία ονομάζονται containers, virtualization engines ή jails. Όπως φαίνεται στο σχήμα 2.2 όλα τα containers μοιράζονται τον ίδιο πυρήνα πάνω από τον οποίο βρίσκεται το στρώμα που είναι υπεύθυνο για την εικονικοποίηση.

Το γεγονός ότι τα containers μοιράζονται τον ίδιο πυρήνα κάνει το μέγεθος τους αρκετά μικρό, αφού δε χρειάζεται να συμπεριλάβουν κάποιο λειτουργικό σύστημα. Επιπλέον, οι χρόνοι εκκίνησης τους είναι αρκετά μικροί, χωρίς να χρειάζεται να καταναλωθεί πολύ μνήμη για τη λειτουργία των containers. Επιπροσθέτως, τα containers είναι ανεξάρτητα από το φυσικό μηχάνημα, καθώς η εικονικοποίηση γίνεται με τη χρήση του κατάλληλου λογισμικού χωρίς να χρειάζεται κάποια υποστήριξη από το hardware. Τέλος, τα containers επιτρέπουν το εύκολο πακετάρισμα τόσο των εφαρμογών όσο και των κατάλληλων ρυθμίσεων, που σε συνδυασμό με την ανεξαρτησία από το hardware, δίνουν τη δυνατότητα για την εύκολη μεταφορά τους σε διαφορετικά μηχανήματα.



Σχήμα 2.2: Εικονικοποίηση σε επίπεδο λειτουργικού συστήματος

Δυστυχώς, όλα αυτά τα πλεονεκτήματα έρχονται με κάποιο κόστος, με το μεγαλύτερο να είναι αυτό της ασφάλειας. Αν και οι υποστηρικτές των containers θεωρούν ότι με τις κατάλληλες ρυθμίσεις τα containers μπορούν να γίνουν ασφαλή, δεν προσφέρουν την ίδια ασφάλεια με την εικονικοποίηση σε επίπεδο υλικού [10]. Τα containers δεν έχουν τη δυνατότητα να προσφέρουν τόσο ισχυρή απομόνωση όσο προσφέρει η εικονικοποίηση σε επίπεδο υλικού. Τα αποτελέσματα από την παραβίαση της απομόνωσης μπορεί να είναι από την απόκτηση πληροφοριών σχετικά με το host μηχανήμα ή άλλων containers, μέχρι την απόκτηση ελέγχου του λειτουργικού συστήματος στον host [9].

Η σύγκριση των δύο ειδών εικονικοποίησης είναι ένα ανοιχτό ζήτημα το οποίο δύσκολα θα λυθεί σύντομα [8]. Αλλωστε και τα δύο είδη, αναπτύσσονται ταχύτατα, προσφέροντας συνεχώς νέες δυνατότητες και μειώνοντας τα προβλήματα που υπάρχουν. Σε μία προσπάθεια να ενοποιηθούν τα θετικά των δύο ειδών, είναι συχνή η χρήση εικονικών μηχανών μέσα στις οποίες τρέχουν containers. Με αυτό τον τρόπο βελτιώνεται η ασφάλεια σε βάρος όμως της απόδοσης, καθώς πλέον χρειάζονται πόροι και για την εκτέλεση του απαραίτητου λογισμικού που θα επιτρέψει την υπέρβαση εικονικών μηχανών.

2.2.2 Εικονικοποίηση σε επίπεδο υλικού

Η εικονικοποίηση σε επίπεδο υλικού, ή απλά εικονικοποίηση για το υπόλοιπο της συγκεκριμένης εργασίας, επιτρέπει τη λειτουργία ενός ολόκληρου υπολογιστικού συστήματος μέσα από έναν άλλον. Τα εμφωλευμένα υπολογιστικά συστήματα αναφέρονται ως guests,

ενώ το υπολογιστικό σύστημα στο οποίο πραγματοποιείται η εικονικοποίηση αναφέρεται ως host. Ένας guest εκτελείται μέσα σε μία εικονική μηχανή (virtual machine - VM), την οποία οι Porek και Goldberg [21] ορίζουν ως εξής: “ένα αποδοτικό και απομονωμένο αντίγραφο μιας πραγματικής μηχανής”. Κάθε εικονική μηχανή είναι ανεξάρτητη τόσο από το host, όσο και από άλλες τυχόν εικονικές μηχανές που μπορεί να υπάρχουν. Με αυτού του είδους την εικονικοποίηση, κάθε εικονική μηχανή έχει την ψαυδαίσθηση ότι κατέχει αποκλειστικά τους πόρους που της έχουν διατεθεί.

Απαραίτητο συστατικό για να επιτευχθούν όλα τα παραπάνω είναι ο επόπτης (hypervisor). Ο επόπτης, σύμφωνα με τους Porek και Goldberg έχει τρία συγκεκριμένα χαρακτηριστικά [21]:

1. ο επόπτης παρέχει ένα περιβάλλον όμοιο με το πραγματικό υπολογιστικό σύστημα,
2. τα προγράμματα που τρέχουν στο περιβάλλον που δημιουργείται θα πρέπει να έχουν στη χειρότερη περίπτωση μικρές μειώσεις στην ταχύτητα εκτέλεσης,
3. ο επόπτης έχει πλήρη έλεγχο των πόρων.

Ο ρόλος του επόπτη μοιάζει με το ρόλο του λειτουργικού συστήματος, με τις εικονικές μηχανές να έχουν το ρόλο των διεργασιών. Ο επόπτης χρονοδρομολογεί τις εικονικές μηχανές στη CPU, μοιράζει γενικότερα τους διαθέσιμους πόρους στα εικονικά μηχανήματα, εκτελεί εκ μέρους των εικονικών μηχανών “προνομιούχες εντολές” και είναι αυτός που διαχειρίζεται τις εικονικές μηχανές (δημιουργία, εκκίνηση, τερματισμός κ.λ.π).

Όπως βλέπουμε στο σχήμα 2.3 υπάρχουν δύο διαφορετικά είδη εποπτών, ανάλογα με το που βρίσκονται στο σύστημα.

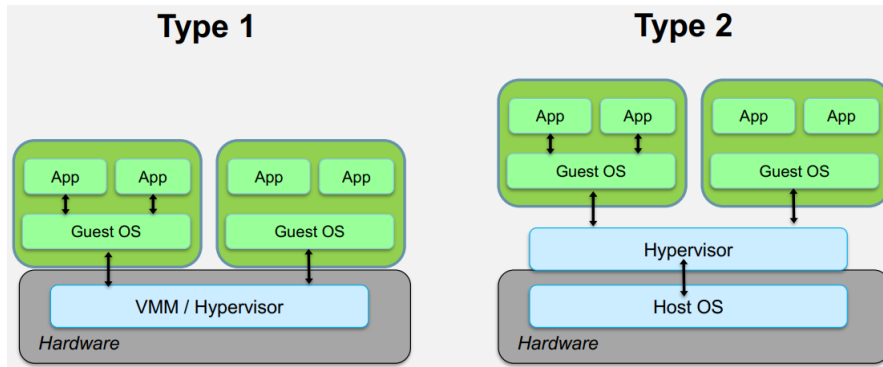
- **Type 1 - Bare metal/native**

Σε αυτή την περίπτωση ο επόπτης εκτελείται πάνω από το υλικό του συστήματος, χωρίς να παρεμβάλλεται κάποιο λειτουργικό σύστημα. Είναι υπεύθυνος για τη χρονοδρομολόγηση των εικονικών μηχανών, τη διαχείριση της μνήμης και των πόρων καθώς και άλλων λειτουργιών. Χρησιμοποιεί δικούς του drivers για τη διαχείριση του υλικού και για την εξυπηρέτηση των I/O λειτουργιών των guest.

- **Type 2 - Hosted**

Σε αυτή την περίπτωση ο επόπτης εκτελείται πάνω από ένα λειτουργικό σύστημα, σαν ένα συνηθισμένο πρόγραμμα στο userspace. Αποτελούν ένα ενδιάμεσο στρώμα μεταξύ του host και του guest, δημιουργώντας το κατάλληλο περιβάλλον για τις εικονικές μηχανές. Δε χρειάζονται ειδικοί drivers για το υλικό, καθώς αυτοί παρέχονται από το λειτουργικό σύστημα του host. Από την

άλλη, οι I/O λειτουργίες και διάφορες άλλες "προνομιούχες εντολές" του guest θα πρέπει να περάσουν μέσα από τον hypervisor και μετά να εξυπηρετηθούν από τον host, δημιουργώντας επιπλέον κόστος στην εκτέλεση τους.



Σχήμα 2.3: Εικονικοποίηση σε επίπεδο υλικού

Εκτός από το διαχωρισμό των εποπτών, γίνεται και διαχωρισμός και στις τεχνικές που χρησιμοποιούνται για να επιτευχθεί η εικονικοποίηση. Οι τεχνικές αυτές βασίζονται τόσο στη δυνατότητα που προσφέρει ειδικό υλικό για την υποστήριξη της εικονικοποίησης, όσο και στο βαθμό συνεργασίας μεταξύ του επόπτη με το εικονικό μηχάνημα. Θα αναφερθούμε σε τρεις από αυτές τις τεχνικές, την πλήρη εικονικοποίηση (full virtualization), την παραεικονικοποίηση (paravirtualization) και τέλος την εικονικοποίηση υποστηριζόμενη από το υλικό (hardware-assisted virtualization).

Full virtualization

Στην πλήρη εικονικοποίηση ο guest δεν έχει υποστεί καμία αλλαγή και εκτελείται όπως και στην περίπτωση που εκτελείται απευθείας σε ένα φυσικό υπολογιστικό σύστημα. Απο τη μεριά του ελεγκτή, θα πρέπει να γίνει πλήρης προσομοίωση του υλικού ώστε να δίνει την ψευδαίσθηση στον guest ότι επικοινωνούν απευθείας με το υλικό. Ωστόσο δημιουργείται ένα πρόβλημα, το λειτουργικό σύστημα του guest εκτελείται σε unprivileged mode, οπότε δεν μπορεί να εκτελέσει προνομιούχες εντολές. Η λύση σε αυτό το πρόβλημα είναι η τεχνική trap-and-emulate [4], κατά την οποία όταν ο guest προσπαθεί να εκτελέσει μία προνομιούχα εντολή, δημιουργείται μία εξαίρεση (trap) στον hypervisor, καθώς ο guest βρίσκεται σε unprivileged mode. Ύστερα, ο hypervisor παίρνει τον έλεγχο και προσομοιώνει ή εκτελεί την προνομιούχα εντολή για λογαριασμό του guest.

Η συγκεκριμένη λύση δεν είναι αρκετή για όλες τις αρχιτεκτονικές. Μία από αυτές τις αρχιτεκτονικές είναι και η x86. Στην x86 αρχιτεκτονική ορισμένες εντολές δεν προκαλούν trap όταν εκτελούνται

σε unprivileged mode, παρά το γεγονός ότι για την εκτέλεση τους είναι ανάγκη να βρίσκονται σε privileged mode (non-virtualizable instructions). Μία από αυτές τις εντολές είναι η `ropf`, η οποία ενώ αλλάζει τόσο τα ALU flags όσο και τα flags του επεξεργαστή δεν προκαλεί κάποιο trap με αποτέλεσμα να αγνοούνται οι αλλαγές στα flags του επεξεργαστή. Για την αντιμετώπιση του συγκεκριμένου προβλήματος χρησιμοποιείται από τους επόπτες η τεχνική του binary translation [4]. Όταν ο guest βρίσκεται σε unprivileged mode όλες οι εντολές εκτελούνται κανονικά, ενώ όταν μεταβεί στον πυρήνα και σε privileged mode τότε ο hypervisor μεταφράζει τις εντολές που δεν προκαλούν trap σε ένα νέο σύνολο εντολών που θα επιφέρουν τις επιθυμητές αλλαγές στην εικονική μηχανή.

Μπορούμε να παρατηρήσουμε εύκολα ότι και στις δύο περιπτώσεις η απόδοση θα μειωθεί σημαντικά, λόγω της ανάμειξης του hypervisor. Ιδιαίτερα στην περίπτωση του binary translation το κόστος της παρακολούθησης και μετάφρασης των εντολών του guest είναι αρκετά μεγάλο.

Paravirtualization

Στην παραεικονικοποίηση ο guest γνωρίζει ότι δεν εκτελείται απευθείας στο υλικό και "συνεργάζεται" κατάλληλα με τον hypervisor. Απαιτείται λοιπόν, η τροποποίηση του guest, ώστε να αντικατασταθούν οι non-virtualizable εντολές, με κατάλληλες εντολές (hypercalls) που δίνουν τον έλεγχο στο hypervisor και αυτός με τη σειρά του διεκπαιρεύει την αντίστοιχη λειτουργία. Επιπλέον ο hypervisor μπορεί να υποστηρίξει και κάποιες άλλες εντολές, που χρησιμοποιούνται για άλλες σημαντικές λειτουργίες του πυρήνα, όπως διαχείριση μνήμης, διαχείριση διακοπών κ.α.

Με αυτό τον τρόπο μειώνεται σημαντικά το κόστος της εικονικοποίησης, αφού ο guest γνωρίζει ότι εκτελείται σε εικονικό περιβάλλον και μπορεί να βελτιστοποιηθεί για τις συγκεκριμένες συνθήκες. Παράλληλα γίνεται δυνατόν να εικονικοποιηθούν αρχιτεκτονικές όπως η x86 χωρίς σημαντική μείωση της απόδοσης αποφεύγοντας το binary translation. Από την άλλη, ο guest τροποποιείται αρκετά για να μπορέσει να "συνεργαστεί" με τον επόπτη. Συνεπώς, μειώνεται η δυνατότητα μεταφοράς του σε διαφορετικά συστήματα (π.χ. διαφορετικό επόπτη), ενώ ταυτόχρονα αυξάνει το κόστος συντήρησης του καθώς οποιεσδήποτε αλλαγές ή αναβαθμίσεις στον πυρήνα του guest θα πρέπει να τροποποιηθούν κατάλληλα για την υποστήριξη της παραεικονικοποίησης.

Hardware-assisted virtualization

Η συγκεκριμένη μέθοδος ουσιαστικά έχει τα ίδια χαρακτηριστικά με τη μέθοδο της πλήρους εικονικοποίησης. Ο guest δε χρειάζεται να υποστεί καμία αλλαγή, ενώ εισάγεται κατάλληλο hardware το οποίο

διευκολύνει την εικονικοποίηση. Πρόκειται για επεκτάσεις των επεξεργαστών με στόχο την αύξηση της απόδοσης και της ασφάλειας της εικονικοποίησης. Αυτές οι επεκτάσεις είναι γνωστές ως επεκτάσεις εικονικοποίησης (virtualization extensions).

Ένα από τα βασικά στοιχεία αυτών των επεκτάσεων είναι η δημιουργία ενός νέου επιπέδου εκτέλεσης, αυτό του guest στο οποίο μπορούν να εκτελεστούν τόσο privileged όσο και unprivileged εντολές. Συνεπώς, δε χρειάζεται πλέον να γίνονται trap οι privileged εντολές του guest, αλλά αυτές μπορούν να εκτελεστούν απευθείας, χωρίς μάλιστα να επηρεάζεται ο host από την εκτέλεση τους. Οι επεκτάσεις αφορούν ακόμα τη διαχείριση μνήμης, αλλά και των διακοπών. Πλέον η εκτέλεση του guest συνεχίζεται αδιάκοπα μέχρι την ικανοποίηση κάποιας συνθήκης, όπως για παράδειγμα ένα page fault, όπου τότε ο έλεγχος μεταφέρεται στον host (VM exit).

Με αυτό τον τρόπο επιτυγχάνεται υψηλή απόδοση για την εικονικοποίηση χωρίς να είναι απαραίτητες οι αλλαγές στον guest. Ο πιο σημαντικός παράγοντας απόδοσης για την εν λόγω τεχνική είναι το πλήθος των VM exits, καθώς πρόκειται για μία χρονοβόρα διαδικασία [5]. Μάλιστα πολλές από αυτές τις επεκτάσεις έχουν δημιουργηθεί με ακριβώς αυτό το σκοπό, δηλαδή τη μείωση των VM exits.

2.2.3 QEMU/KVM

Το Qemu [6] είναι ένας επόπτης τύπου 2 και πρόκειται για έναν από τους πιο γνωστούς επόπτες. Με την τεχνική του full virtualization και συγκεκριμένα με τη μέθοδο του dynamic binary translation, υποστηρίζει τη φιλοξενία πολλών λειτουργικών συστημάτων χωρίς να απαιτείται κάποια αλλαγή σε αυτά. Πέρα από τη λειτουργία του επόπτη, μπορεί να λειτουργήσει και ως προσομοιωτής υλικού. Το Qemu έχει τη δυνατότητα να προσομοιώσει διαφορετικές αρχιτεκτονικές από αυτή που διαθέτει ο host, πληθώρα από κάρτες δικτύου, σκληρούς δίσκους, περιφερειακές συσκευές κ.λ.π.. Με αυτό τον τρόπο μπορεί να χρησιμοποιηθεί και για εφαρμογές που έχουν υλοποιηθεί για διαφορετική αρχιτεκτονική να εκτελεστούν στην αρχιτεκτονική που διαθέτει ο host.

Οι εικονικές μηχανές που εκτελούνται στον ίδιο host, συχνά επικοινωνούν μεταξύ τους. Συνήθως πρόκειται για πλήρη υπολογιστικά συστήματα, συνεπώς για την επικοινωνία μεταξύ τους θα μπορούσε να χρησιμοποιηθεί ό,τι χρησιμοποιείται μεταξύ φυσικών υπολογιστικών συστημάτων (π.χ. TCP/IP sockets). Μάλιστα ειδικά στη συγκεκριμένη περίπτωση οι επόπτες μπορούν να αντιληφθούν ότι πρόκειται για επικοινωνία μεταξύ εικονικών μηχανών που ελέγχουν οπότε και χρησιμοποιούν διάφορες βελτιστοποιήσεις, προκειμένου να μη χρειαστεί να περάσουν τα πακέτα πέρα από το host. Εντούτοις, συχνά πολλές εικονικές μηχανές δε χρειάζονται δίκτυο και δημιουργεί-

ται η ανάγκη για επικοινωνία μεταξύ άλλων εικονικών μηχανών λη ακόμα και με το host. Επιπλέον πολλές φορές είναι χρήσιμο εικονικές μηχανές να μοιράζονται μνήμη μεταξύ τους . Έτσι έχουν δημιουργηθεί αρκετές μέθοδοι για τον διαμοιρασμό μνήμης και ενδοεπικοινωνίας μεταξύ εικονικών μηχανών, αλλά και μεταξύ του host και των εικονικών μηχανών [24].

Ένας από αυτούς τους μηχανισμούς είναι ο nahanni ή ivshmem (inter-VM shared memory) [17]. Το ivshmem είναι ένας μηχανισμός για το διαμοιρασμό μνήμης του host με τις εικονικές μηχανές που εκτελούνται στο host. Μπορεί να χρησιμοποιηθεί τόσο για διαμοιρασμό μνήμης μεταξύ host και guest, όσο και μεταξύ των guests. Το Qemu επιτρέπει τη χρήση αυτού του μηχανισμού μέσω μίας PCI συσκευής την οποία πρέπει να υποστηρίξουν οι guests για να χρησιμοποιήσουν τη συγκεκριμένη περιοχή μνήμης.

Πολλές φορές το Qemu χρησιμοποιείται από άλλους επόπτες για την εξομίωση του υλικού, όπως για παράδειγμα γίνεται από το Xen. Επιπροσθέτως συχνά χρησιμοποιείται σε συνδυασμό με το kvm (στην περίπτωση υποστήριξης hardware-assisted virtualization), πετυχαίνοντας υψηλή απόδοση. Ωστόσο, από μόνο του το kvm δε λειτουργεί ως επόπτης, αλλά ανοίγει μία διεπαφή (/dev/kvm) μέσω της οποίας μπορούν να χρησιμοποιηθούν οι διάφορες λειτουργίες που προσφέρει. Το kvm είναι ένα kernel module του Linux, που μετατρέπει τον πυρήνα σε έναν επόπτη. Σε αυτό τον συνδυασμό (Qemu/Kvm) κάθε εικονική μηχανή εκτελείται ως μία διεργασία.

2.3 Λειτουργικά συστήματα

Το λειτουργικό σύστημα είναι ένα πρόγραμμα το οποίο εκτελείται σε ένα υπολογιστικό σύστημα. Η ειδοποιός διαφορά του με οποιοδήποτε άλλο πρόγραμμα είναι ότι το πρώτο υπάρχει για να εξυπηρετεί το δεύτερο. Το λειτουργικό σύστημα διαχειρίζεται το υλικό του υπολογιστικού συστήματος και παρέχει στους προγραμματιστές των εφαρμογών ένα σαφές αφηρημένο σύνολο πόρων, αντί για το μπερδεμένο σύνολο που υπάρχει στο υπολογιστικό σύστημα. Χωρίς αυτό, ο προγραμματιστής μίας οποιασδήποτε εφαρμογής θα έπρεπε να γράψει και τον κατάλληλο κώδικα για το υλικό που χρησιμοποιεί η συγκεκριμένη εφαρμογή (π.χ. αποθήκευση σε σκληρό δίσκο). Ως εκ τούτου θα σπαταλούσε αρκετό χρόνο σε μη βασικές λειτουργίες του προγράμματος. Επιπροσθέτως πολλές από αυτές τις λειτουργίες που αφορούν το υλικό μπορεί να έχουν ήδη υλοποιηθεί από άλλα προγράμματα, σπαταλώντας έτσι χρόνο για ήδη υπάρχουσες λειτουργίες.

Η πληθώρα και οι σημαντικές διαφορές στο σκοπό και στις δυνατότητες των υπολογιστικών συστημάτων οδήγησε στη δημιουργία διαφορετικών ειδών λειτουργικών συστημάτων. Για παράδειγμα,

ένα λειτουργικό σύστημα που έχει δημιουργηθεί με σκοπό την εκτέλεση του σε προσωπικούς υπολογιστές διαφέρει αρκετά από ένα λειτουργικό σύστημα που χρησιμοποιείται σε ενσωματωμένα συστήματα. Το πιο γνωστό είδος λειτουργικών συστημάτων είναι αυτά του γενικού σκοπού, τα οποία έχουν ως στόχο να μπορούν να εκτελεστούν σε διάφορα υπολογιστικά συστήματα προσφέροντας όσο το δυνατόν περισσότερες λειτουργίες.

Ένας διαφορετικός τρόπος κατηγοριοποίησης των λειτουργικών συστημάτων γίνεται με τη δομή τους. Θα αναφερθούν οι κυριότερες από αυτές τις δομές, μονολιθικά συστήματα, πολυεπίπεδα συστήματα, μικροπυρήνες.

- **μονολιθικά συστήματα:** η πιο διαδεδομένη οργάνωση, στην οποία ολόκληρουόκληρο το λειτουργικό σύστημα εκτελείται ως ένα μοναδικό πρόγραμμα σε κατάσταση πυρήνα. Ουσιαστικά πρόκειται για ένα μεγάλο εκτελέσιμο δυαδικό πρόγραμμα. Κάθε διαδικασία του συστήματος μπορεί να καλέσει οποιαδήποτε άλλη.
- **πολυεπίπεδα συστήματα:** το λειτουργικό σύστημα οργανώνεται σε μία ιεραρχία επιπέδων, όπου το καθένα δομείται πάνω στο αμέσως κατώτερο του. Κάθε επίπεδο αναλαμβάνει μία διαφορετική λειτουργία και μπορεί να χρησιμοποιήσει μόνο λειτουργίες που υποστηρίζουν κατώτερα επίπεδα.
- **μικροπυρήνες:** το λειτουργικό σύστημα διαιρείται σε μικρές καλά ορισμένες υπομονάδες, από τις οποίες μόνο μία, ο μικροπυρήνας, εκτελείται σε κατάσταση πυρήνα ενώ οι υπόλοιπες εκτελούνται σε σχετικά λιγότερο ισχυρές κοινές διεργασίες χρήστη.

Σε ένα μονολιθικό λειτουργικό σύστημα διακρίνονται δύο ξεχωριστά επίπεδα εκτέλεσης κώδικα. Κάθε επίπεδο έχει διαφορετικούς χώρους μνήμης και διαφορετικά δικαιώματα πρόσβασης στο υλικό. Τα δύο αυτά επίπεδα είναι ο χώρος χρήστη (userspace), στο οποίο εκτελούνται οι διεργασίες του χρήστη και ο χώρος πυρήνα (kernel space), στο οποίο εκτελείται ο κώδικας του πυρήνα. Προκειμένου μία διεργασία να εκτελέσει μία λειτουργία για την οποία δε διαθέτει τα ανάλογα δικαιώματα, υπάρχει ένας μηχανισμός μέσω του οποίου η διεργασία ζητά από τον πυρήνα να εκτελέσει τη συγκεκριμένη λειτουργία για λογαριασμό της. Ο μηχανισμός αυτός ονομάζεται κλήση συστήματος (system call).

Ένα από τα πρώτα και πιο ιστορικά λειτουργικά συστήματα είναι το UNIX. Το UNIX επιλέχτηκε ως η βάση για μία τυποποιημένη διεπαφή συστήματος, ωστόσο υπήρχαν και δημιουργήθηκαν πολλές παραλλαγές του. Συνεπώς, δημιουργήθηκε η ανάγκη για τη δημιουργία ενός κοινού Standard, ώστε να διατηρείται η συμβατότητα μεταξύ

αυτών των διάφορων εκδόσεων του UNIX. Το standard που δημιουργήθηκε ήταν το POSIX, το οποίο ορίζει μία ελάχιστη διασύνδεση κλήσεων συστήματος που πρέπει να υποστηρίζουν τα συμβατά συστήματα UNIX. Μάλιστα το POSIX χρησιμοποιείται ακόμα και στις μέρες μας, ενώ έχουν δημιουργηθεί και κάποιες παραλλαγές του. Το UNIX επηρέασε σε μεγάλο βαθμό και κάποια από τα πιο δημοφιλή σύγχρονα λειτουργικά συστήματα όπως το Linux και αυτά της BSD οικογένειας. Ένα από αυτά τα λειτουργικά συστήματα είναι και το NetBSD, το οποίο είναι ένα γενικού σκοπού λειτουργικό σύστημα. Χρησιμοποιείται συνήθως σε servers και λιγότερο σε προσωπικούς υπολογιστές, ενώ υποστηρίζει αρκετές αριτεκτονικές.

Κεφάλαιο 3

Unikernel frameworks

Σε αυτό το κεφάλαιο θα γίνει μία εισαγωγή στην έννοια των unikernels και το σκοπό τους, ενώ θα παρουσιαστούν ορισμένα unikernel frameworks, τα οποία μελετήθηκαν κατά τη διάρκεια αυτής της εργασίας.

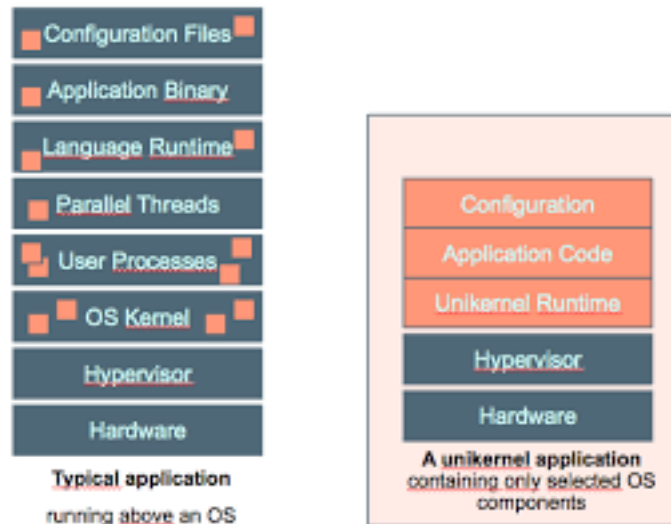
Ένας τυπικός ορισμός των unikernels αναφέρεται στη σελίδα unikernel.org και είναι ο εξής:

Τα Unikernels είναι εξειδικευμένες εικόνες μηχανής, με ένα μοναδικό χώρο διευθύνσεων, τα οποία κατασκευάζονται χρησιμοποιώντας library operating systems

Αναλύοντας τον παραπάνω ορισμό μπορούμε εύκολα να συμπεράνουμε τα βασικά χαρακτηριστικά των unikernels:

- εξειδικευμένες εικονικές μηχανές: κάθε unikernel μπορεί να είναι διαφορετικός από τους υπόλοιπους. Κάθε ένας αφορά μία συγκεκριμένη εφαρμογή και φτιάχνεται γύρω από αυτήν.
- ένας μοναδικός χώρος διευθύνσεων: στην περίπτωση των unikernels δεν υπάρχει ο διαχωρισμός μεταξύ userspace και kernelspace, όπως στα συμβατικά λειτουργικά συστήματα.
- library operating systems [22]: πρόκειται για λειτουργικά συστήματα τα οποία παρέχουν τις υπηρεσίες που υποστηρίζουν (networking κ.λ.π.) στη μορφή βιβλιοθηκών, οι οποίες στη συνέχεια μπορούν να χρησιμοποιηθούν (link) από τις εφαρμογές.

Με λίγα λόγια ένας unikernel αποτελείται από τον κώδικα της εφαρμογής την οποία θα εκτελέσει, τα απαραίτητα κομμάτια του λειτουργικού συστήματος που η συγκεκριμένη εφαρμογή χρειάζεται και τις κατάλληλες παραμετροποιήσεις. Τα unikernels δεδομένου ότι αφορούν μία και μόνο εφαρμογή δεν υποστηρίζουν περισσότερες από μία διεργασίες, ούτε περισσότερους από έναν χρήστες. Αυτόματα όλο το κομμάτι των συμβατικών λειτουργικών συστημάτων που αφορούν τη



Σχήμα 3.1: Unikernel vs OS software stack

διαχείριση και υποστήριξη πολλών διεργασιών και χρηστών καθίσταται άχρηστο και δεν συμπεριλαμβάνεται. Τα πράγματα είναι διαφορετικά όσον αφορά τα νήματα, καθώς άλλα unikernel frameworks τα υποστηρίζουν ενώ άλλα όχι.

Στην εικόνα 3.1 φαίνεται η διαφορά μεταξύ ενός συμβατικού λειτουργικού συστήματος και ενός unikernel. Τα παραπάνω ενοποιούνται σε μία εικόνα μηχανής που μπορεί να εκτελεστεί από έναν επόπτη. Γίνεται εύκολα αντιληπτό ότι το μέγεθος της τελικής εικόνας θα είναι πολύ μικρότερο από αυτή ενός συμβατικού λειτουργικού συστήματος. Μία ακόμη συνέπεια αυτού είναι ότι οι χρόνοι εκκίνησης θα είναι αρκετά μικρότεροι, αφού πλέον δε χρειάζεται να φορτωθούν και να εκκινήσουν όλες οι υπηρεσίες που προσφέρει ένα λειτουργικό σύστημα.

Από άποψη ασφάλειας, ο σχεδιασμός των unikernels από μόνος του προσφέρει σημαντική ασφάλεια. Αρχικά το μικρό μέγεθος των unikernels συνεπάγεται αυτόματα και μικρότερο εύρος επίθεσης από κάποιον κακόβουλο χρήστη. Επιπλέον πρόκειται για εξειδικευμένες εικόνες που αφορούν μία και μόνο εφαρμογή, οπότε στην περίπτωση που αποκτηθεί πρόσβαση στην εικονική μηχανή, το περιθώριο εκμετάλλευσης θα είναι αρκετά μικρό. Τέλος, υπάρχει απομόνωση μεταξύ των εικονικών μηχανών η οποία προσφέρεται από τον επόπτη.

Η έλλειψη διαχωρισμού kernelspace και userspace, μειώνει το κόστος μετάβασης από τον ένα χώρο διευθύνσεων στον άλλο, κάνοντας έτσι τις κλήσεις συστήματος να λειτουργούν σαν απλές κλήσεις συναρτήσεων. Ως αποτέλεσμα, η ταχύτητα εκτέλεσης αυξάνει σημαντικά, καθώς δε χρειάζονται πλέον οι μεταβάσεις από τον ένα χώρο διευθύνσεων στον άλλο, μία αρκετά δαπανηρή λειτουργία.

Η έννοια των library operating systems δεν είναι καινούρια, ωστόσο δεν μπορούσαν να χρησιμοποιηθούν λόγω του μεγάλου πλήθους των συσκευών και συνεπώς των drivers που θα έπρεπε να υποστηρίζουν. Στις μέρες μας όμως, οι επόπτες έχουν περιορίσει σημαντικά το πλήθος των συσκευών κάνοντας εφικτή τη χρήση των συγκεκριμένων λειτουργικών συστημάτων.

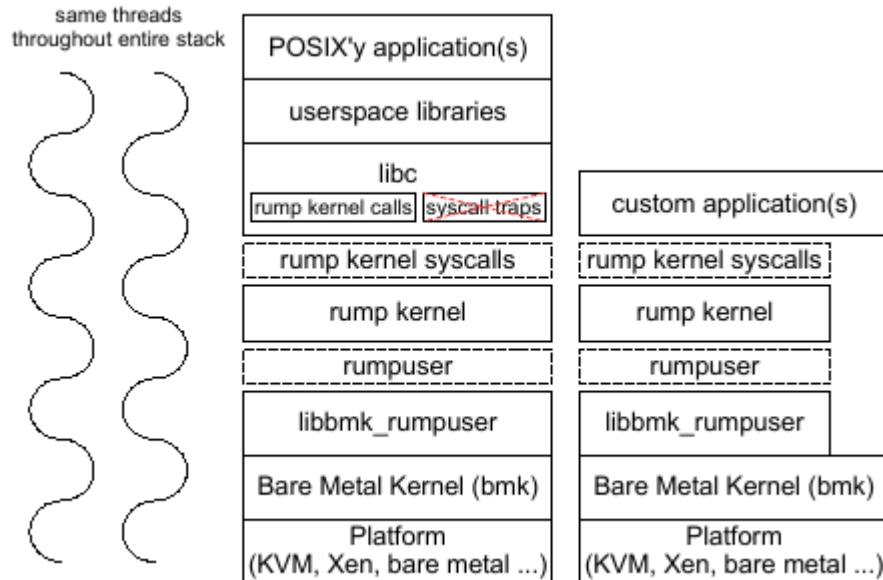
Αφ' ετέρου η απομάκρυνση όλων αυτών των υπηρεσιών από το λειτουργικό σύστημα, δημιουργεί αρκετούς περιορισμούς. Οι περισσότερες εφαρμογές έχουν φτιαχτεί για συμβατικά λειτουργικά συστήματα λαμβάνοντας υπόψιν και χρησιμοποιώντας τις συγκεκριμένες υπηρεσίες. Απαιτείται, λοιπόν, να γίνουν σημαντικές αλλαγές στις εφαρμογές για να μπορέσουν να εκτελεστούν σε unikernels. Επιπλέον γίνεται πιο δύσκολη η αποσφαλμάτωση των unikernels δεδομένου των λίγιστων υπηρεσιών που προσφέρουν.

3.1 Rumprun

Το rumprun είναι ένα unikernel framework, το οποίο έχει χτιστεί με βάση τα συστατικά των drivers που προσφέρουν οι rump kernels. Το rumprun μπορεί να προσφέρει ένα POSIX-like περιβάλλον, το οποίο μπορεί να χρησιμοποιηθεί από POSIX εφαρμογές ώστε να μετατραπούν χωρίς καμία αλλαγή σε unikernel εικόνες. Αυτός ήταν και ένας από τους βασικούς στόχους του εγχειρήματος, η δυνατότητα δηλαδή σε POSIX κώδικα να μπορεί να εκτελεστεί χωρίς να υποστεί καμία αλλαγή. Στην εικόνα 3.2 φαίνεται η δομή του rumprun για POSIX εφαρμογές (αριστερά) και προσαρμοσμένες εφαρμογές (δεξιά). Προφανώς το δεξί μοντέλο φαίνεται αρκετά μικρότερο, αλλά είναι απαραίτητες να γίνουν συγκεκριμένες αλλαγές στην εφαρμογή. Ο συγκεκριμένος unikernel μπορεί να εκτελεστεί χρησιμοποιώντας Xen ή KVM, ενώ υποστηρίζει arm και x86 αρχιτεκτονικές. Ο κώδικας του rumprun είναι διαθέσιμος στο <http://repo.rumpkernel.org/rumprun>.

Όπως κάθε unikernel framework το rumprun υποστηρίζει μία και μόνο διεργασία. Μολαταύτα υποστηρίζει POSIX threads, δίνοντας τη δυνατότητα για την ύπαρξη περισσότερων από ένα νήματα. Μάλιστα ο χρονοδρομολογητής που διαθέτει είναι cooperative, οπότε αν ένα νήμα αποκτήσει τον έλεγχο δεν πρόκειται να διακοπεί από το χρονοδρομολογητή αλλά θα αφήσει το έλεγχο όταν εκείνο τερματίσει ή πρόκειται να μπλοκάρει. Κάποιοι ακόμα περιορισμοί, είναι η μη υποστήριξη εικονικής μνήμης (virtual memory) και σημάτων (signals), Συνεπώς εφαρμογές που χρησιμοποιούν σήματα ή κλήσεις συστήματος όπως η mmap() δε θα λειτουργούν χωρίς αλλαγές στον κώδικα τους.

Ένας rumprun unikernel αποτελείται από τα απαραίτητα κομ-



Σχήμα 3.2: Rumprun software stack

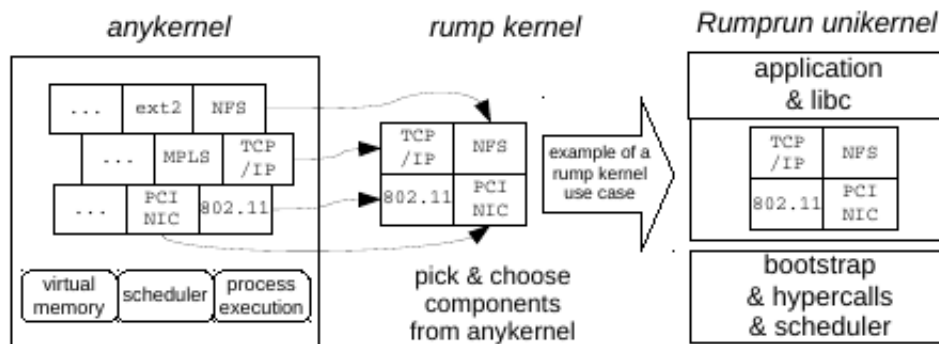
μάτια των rump kernels που χρειάζεται η εφαρμογή και την ίδια την εφαρμογή. Για τη δημιουργία του unikernel γίνεται πάντα cross-compilation, οπότε δε χρειάζεται κάποιος ήδη έτοιμος rumprun unikernels, αλλά μία σειρά από εργαλεία. Η διαδικασία έχει ως εξής:

1. Αρχικά γίνεται η μεταγλώττιση του κώδικα της εφαρμογής χρησιμοποιώντας έναν cross-compiler.
2. Ύστερα, αντί για linking γίνεται pseudo-linking όπως ονομάζουν τη διαδικασία κατά την οποία απλά ελέγχονται αν ικανοποιούνται όλες οι εξαρτήσεις συμβόλων, χωρίς να γίνεται κάποια σύνδεση με κάποιο συστατικό του λειτουργικού συστήματος.
3. Τέλος, το παράγωγο του προηγούμενου βήματος γίνεται "bake", όπου εισάγονται και τα κομμάτια του λειτουργικού συστήματος που απαιτούνται, όπως έχουν οριστεί κατά την εκτέλεση της εντολής.

Όπως έχει αναφερθεί ο rumprun unikernel βασίζεται στα rump kernels [13]. Τα rump kernels παρέχουν drivers του NetBSD ως φορητά εξαρτήματα με τα οποία μπορείς να εκτελέσεις εφαρμογές χωρίς να ναι απαραίτητη η ύπαρξη λειτουργικού συστήματος [14]. Η αρχική ιδέα ήταν να υπάρχει η δυνατότητα να εκτελούνται αμετάβλητοι drivers του NetBSD ως απλά προγράμματα σε userspace, ώστε να είναι πιο εύκολος ο έλεγχος και η ανάπτυξη NetBSD drivers. Ένας

rumprun kernel είναι ένας πυρήνας χρονομερισμού (timesharing) από τον οποίο έχουν αφαιρεθεί ορισμένα κομμάτια [12]. Τα κομμάτια που κρατήθηκαν είναι οι drivers και οι απαραίτητες ρουτίνες που χρειάζονται οι συγκεκριμένοι drivers για να λειτουργήσουν (συγχρονισμός, κατανομή μνήμης κ.λ.π.). Τα κομμάτια που αφαιρέθηκαν είναι αυτά που αφορούν τις διεργασίες, την εικονική μνήμη κ.α.. Επιπλέον τα rump kernels έχουν και ένα καλώς ορισμένο στρώμα φορητότητας, ώστε να είναι εύκολο να ενσωματωθούν σε διάφορα περιβάλλοντα.

Πίσω από τα rump kernels υπάρχει μία ακόμα έννοια αυτή του anykernel [12]. Ο συγκεκριμένος όρος δημιουργήθηκε ως απάντηση στην όλο και αυξανόμενο αριθμό μοντέλων λειτουργικών συστημάτων (monolithic, exokernel, mikrokernel, unikernel κ.λ.π.). Ένα πολύ μεγάλο ποσοστό ενός λειτουργικού συστήματος ανεξάρτητα από το μοντέλο που χρησιμοποιεί αποτελείται από drivers. Ο όρος anykernel περιγράφει ένα κώδικα βάσης (codebase) τύπου κώδικα πυρήνα από τον οποίο οι οδηγοί μπορούν να εξαχθούν και να ενσωματωθούν σε οποιοδήποτε μοντέλο λειτουργικού συστήματος, χωρίς να χρειάζονται αλλαγές και συντήρηση. Στην εικόνα 3.3 φαίνεται η σχέση μεταξύ των εννοιών anykernel, rump kernels και rumprun.

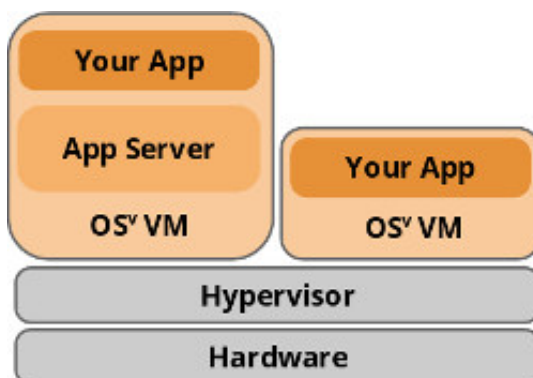


Σχήμα 3.3: Σχέση μεταξύ των εννοιών anykernel, rump kernel και rumprun

3.2 OSv

Το OSv δημιουργήθηκε από το μηδέν, με σκοπό να μπορεί να εκτελέσει μία εφαρμογή πάνω από έναν επόπτη. Ως γλώσσα προγραμματισμού χρησιμοποιεί τη C++ και ο κώδικας του είναι διαθέσιμος στο <https://github.com/cloudius-systems/osv>. Υποστηρίζει διάφορους επόπτες και επεξεργαστές, όπως arm και x86, ενώ στην

x86 αρχιτεκτονική υποστηρίζει τους επόπτες Xen, KVM, VMWare, Virtualbox. Παρέχει τη δυνατότητα να μπορεί να υποστηρίξει σχεδόν οποιοδήποτε πρόγραμμα, που δε χρειάζεται παραπάνω από μία διεργασία. Μάλιστα παρέχει υποστήριξη για αρκετές γλώσσες προγραμματισμού όπως C, Java, Ruby, Node.js, Perl και άλλες.



Σχήμα 3.4: Osv unikernel stack

Η επιλογή των δημιουργών του OSv να δημιουργήσουν από την αρχή ένα νέο λειτουργικό σύστημα τους έδωσε τη δυνατότητα να εισάγουν ορισμένες νέες τεχνικές [15]. Σε αυτό το πλαίσιο στο OSv δεν υπάρχουν spinlocks. Στα περισσότερα SMP (Symmetric multiprocessing) λειτουργικά συστήματα χρησιμοποιούνται spinlocks για την προστασία κοινών δεδομένων από πολλά νήματα. Όταν ένα νήμα κατέχει το spinlock για μία δομή δεδομένων τα υπόλοιπα νήματα δεν μπορούν να αποκτήσουν πρόσβαση στη συγκεκριμένη δομή και περιμένουν, κάνοντας συνεχώς ερωτήματα αν το κλείδωμα είναι ελεύθερο. Τα spinlocks χρησιμοποιούνται κυρίως για νήματα που δεν μπορούν να μπλοκάρουν (κυρίως νήματα του πυρήνα). Ωστόσο, στην περίπτωση της εικονικοποίησης τα spinlocks δημιουργούν το πρόβλημα του "lock-holder preemption" [25]. Αν μία εικονική CPU σταματήσει να εκτελείται και κρατάει ένα spinlock, οι άλλες CPU που χρειάζονται το συγκεκριμένο spinlock, είναι αναγκασμένες να περιμένουν σπαταλώντας κύκλους της CPU. Υπό αυτές τις συνθήκες στο OSv αποφάσισαν όλη η δουλειά του πυρήνα να γίνεται από νήματα που μπορούν να "κοιμηθούν". Τα παραπάνω δεν ισχύουν για το χρονοδρομολογητή, ο οποίος δεν μπορεί να "κοιμηθεί", οπότε χρησιμοποιεί ουρές εκτέλεσης για κάθε επεξεργαστή, ώστε να μην απαιτείται συνεννόηση μεταξύ των επεξεργαστών για τη χρονοδρομολόγηση, ενώ αν ένα νήμα αλλάξει επεξεργαστή χρησιμοποιούνται lock-free αλγόριθμοι.

Όπως αναφέρθηκε προηγουμένως το OSv διαθέτει χρονοδρομολογητή. Μερικά από τα χαρακτηριστικά του είναι: διακοπτόμενη χρονοδρομολόγηση (preemptive scheduling), η μη χρήση spinlocks και χω-

ρίς timer interrupts, ενώ τέλος δε γίνεται κανένας διαχωρισμός μεταξύ των νημάτων του πυρήνα και της εφαρμογής. Ένα εξίσου ενδιαφέρον κομμάτι στο σχεδιασμό του OSv είναι το network stack. Λαμβάνοντας υπόψιν ότι πρόκειται για ένα λειτουργικό σύστημα για το cloud, το OSv έχει δώσει μεγάλη βαρύτητα στην υποστήριξη εφαρμογών που χρησιμοποιούν πολύ το δίκτυο. Το network stack λοιπόν, είναι σχεδιασμένο με βάση τις ιδέες του Van Jacobson [11], με βάση τις οποίες το OSv υποστηρίζει ότι πετυχαίνει πολύ υψηλότερη απόδοση από τα network stacks των συμβατικών λειτουργικών συστημάτων.

Η εκτέλεση μίας εφαρμογής στο OSv δε διαφέρει αρκετά από την εκτέλεση της σε ένα συμβατικό λειτουργικό σύστημα, όπως π.χ. το Linux. Η διαφορά είναι ότι στο OSv δεν εκτελούνται τα συνηθισμένα fixed-position εκτελέσιμα αλλά απαιτούνται shared objects. Για την εκτέλεση του το OSv αναζητά την main συνάρτηση της εφαρμογής και την εκτελεί. Για το πως θα τοποθετηθεί η εφαρμογή στο unikernel, προσφέρονται δύο τρόποι. Είτε όταν χτίζεται το unikernel ορίζουμε να συμπεριληφθεί και το αντίστοιχο shared object, είτε μπορούμε να το φορτώσουμε στο unikernel μεταγενέστερα, μέσω του REST API που προσφέρει το OSv. Μάλιστα, υπάρχει η επιλογή για διαχείριση του unikernel μέσω μία ιστοσελίδας, ενώ υπάρχει η δυνατότητα να υπάρχει ένα απλό κέλυφος (shell). Αξιοσημείωτη είναι και η προσπάθεια δημιουργίας ενός εργαλείου για τη δημιουργία unikernels με OSv. Το εργαλείο αυτό ονομάζεται Capstan και μοιάζει αρκετά στη χρήση του με το Docker.

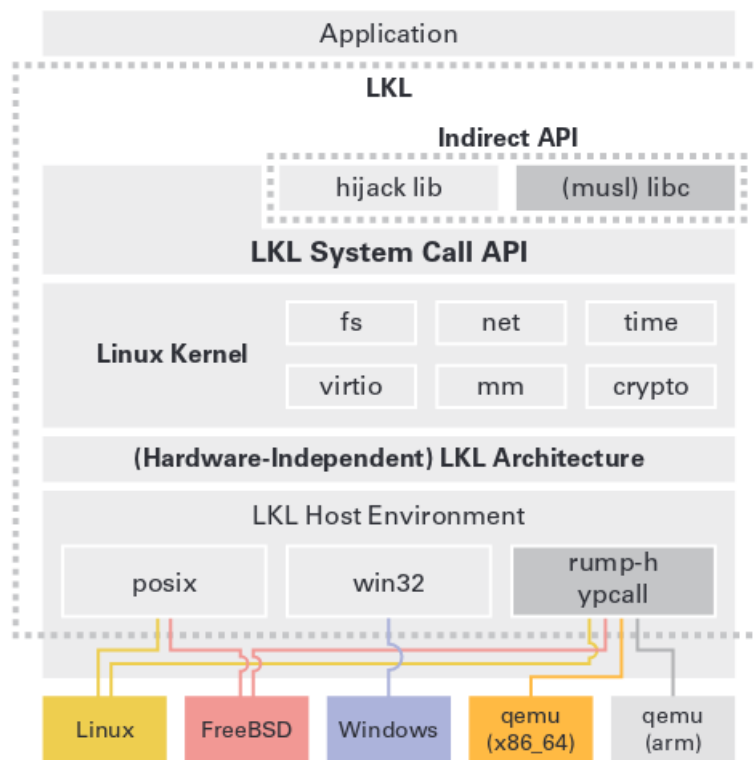
Σαν γενική εικόνα το OSv μοιάζει με ένα library operating system, το οποίο υποστηρίζει πολλές λειτουργίες των συμβατικών λειτουργικών συστημάτων. Ενδεικτικό είναι το γεγονός ότι διαθέτει εικονική μνήμη, καθώς και ένα εικονικό σύστημα αρχείων. Χάρη όλα αυτά τα χαρακτηριστικά μπορεί να υποστηρίξει ένα πολύ μεγάλο εύρος εφαρμογών ακόμα και αυτές που κάνουν χρήση της mmap(). Ο μόνος περιορισμός που υπάρχει είναι ότι δεν υποστηρίζεται η fork(), καθώς παραβιάζει το single-process χαρακτήρα. Από την άλλη μεριά, όλες αυτές οι λειτουργίες που προσφέρονται αυξάνουν το μέγεθος του unikernel, κάνοντας το OSv ένα από τα πιο "βαριά" unikernel frameworks.

3.3 Linux Kernel Library - lkl

Το lkl [23] φέρνει τις ιδέες του rumpkernel και του anykernel στο Linux. Βασικός στόχος του είναι η επαναχρησιμοποίηση του κώδικα του Linux με όσο το δυνατόν λιγότερη προσπάθεια και μειωμένο κόστος συντήρησης. Έχει υλοποιηθεί ως ένα port του Linux σε μία εικονική αρχιτεκτονική (lkl), προκειμένου να μειωθούν οι ενοχλητικές τροποποιήσεις του πυρήνα. Το lkl μπορεί να χρησιμοποιηθεί για

τρεις διαφορετικούς λόγους, α) στιγμιαία παράκαμψη του πυρήνα, β) επαναχρησιμοποίηση κώδικα πυρήνα στο χώρο χρήστη και τέλος γ) ως unikernel. Βασικά χρησιμοποιείται περισσότερο για τη χρήση ορισμένων λειτουργιών του πυρήνα του Linux σε άλλα λειτουργικά συστήματα ή στο χώρο χρήστη, παρά ως unikernel.

Το lkl στην ουσία είναι βιβλιοθήκες τις οποίες οι εφαρμογές μπορούν να χρησιμοποιήσουν. Για τη χρήση του, λοιπόν, προσφέρει ένα δικό του API, που αποτελεί ένα υποσύνολο των κλήσεων συστήματος του Linux. Στην εικόνα 3.7 φαίνονται όλα τα συστατικά του lkl. Όπως φαίνεται υπάρχουν δύο APIs 1) lkl system call API, που είναι βασισμένο στις κλήσεις συστήματος του Linux και 2) το API που χρησιμοποιεί το πρώτο έμμεσα. Το δεύτερο API αποτελείται από τη hijack library και μία βασική υλοποίηση βιβλιοθήκης και συγκεκριμένα μία τροποποίηση της musl libc ώστε να χρησιμοποιεί το lkl system call API. Η hijack library χρησιμοποιείται για την επί τόπου αντικατάσταση των κλήσεων συστήματος που εκτελεί μία εφαρμογή, ώστε να μπορεί η εφαρμογή να χρησιμοποιεί το lkl αντί για τον πυρήνα του host.



Σχήμα 3.5: Η αρχιτεκτονική του lkl

Τα υπόλοιπα κομμάτια του lkl έχουν να κάνουν με το περιβάλλον του host και την generic lkl architecture. Δεδομένου, ότι το lkl δε στο-

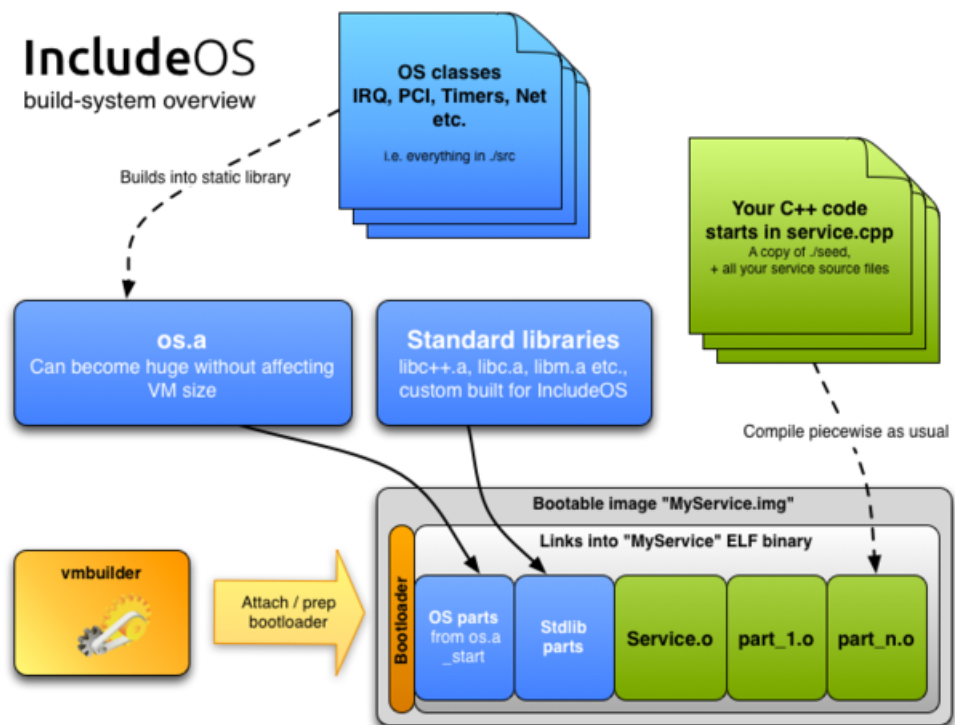
χεύει μόνο σε ένα περιβάλλον, δε χρησιμοποιεί κώδικα που εξαρτάται από την πλατφόρμα. Αντ' αυτού η εφαρμογή θα πρέπει να υλοποιεί ορισμένα θεμελιακά στοιχεία που εξαρτώνται από το περιβάλλον. Αυτά τα θεμελιακά στοιχεία χρησιμοποιούνται από την lkl generic architecture για τη δημιουργία της εικονικής μηχανής πάνω στην οποία θα εκτελεστεί ο πυρήνας του Linux. Ωστόσο, υπάρχουν ορισμένα περιβάλλοντα για τα οποία το lkl παρέχει αυτά τα θεμελιακά στοιχεία. Τα περιβάλλοντα αυτά είναι το POSIX, το NT (windows userspace), το NTK (Windows kernel), το Apache Portable Runtime, ενώ υπάρχει υποστήριξη και για το rumpun.

Το lkl επιτρέπει την εκτέλεση μίας και μόνο διεργασίας, ενώ παρέχει υποστήριξη για πολλαπλά νήματα με ορισμένους περιορισμούς. Αρχικά δεν υπάρχει υποστήριξη για SMP και επιπλέον μία εφαρμογή περιορίζεται απλά στη δημιουργία και τον τερματισμό ενός νήματος. Όσον αφορά τη διαχείριση της μνήμης, αφού δε χρειάζεται κάποια προστασία μνήμης το lkl απλά χρειάζεται ένα χώρο μνήμης από τον οποίο μπορεί ο πυρήνας του Linux να κατανείμει μνήμη που χρειάζεται. Επιπλέον χάρις τη δυνατότητα του Linux να παρέχει εικονική μνήμη ακόμα και για non-MMU αρχιτεκτονικές μπορεί να υποστηριχθεί και εικονική μνήμη.

3.4 IncludeOS

Το IncludeOS [7] είναι ένα από τα νεώτερα Unikernel frameworks. Είναι γραμμένο σε C++ και υποστηρίζει εφαρμογές που χρησιμοποιούν την ίδια γλώσσα προγραμματισμού. Βασικός στόχος για το συγκεκριμένο unikernel είναι να έχει τόσο απλή χρήση όσο η χρήση μίας βιβλιοθήκης στη C++. Πιο συγκεκριμένα μία εντολή της μορφής `#include <os>` στην αρχή μίας εφαρμογής θα έχει ως αποτέλεσμα το χτίσιμο μίας εικόνας δίσκου (disk image) που μπορεί να εκτελεστεί από τους περισσότερους επόπτες (kvm, virtualbox). Για να υπάρχει αυτή η δυνατότητα οι δημιουργοί του IncludeOS, έχουν αναπτύξει και μία σειρά εργαλείων. Χρησιμοποιώντας τα, η διαδικασία του "χτίσματος" περιλαμβάνει τρία στάδια. Αρχικά, γίνεται η σύνδεση της εφαρμογής με τα μέρη του λειτουργικού συστήματος που χρειάζονται. Στη συνέχεια, προσάπτεται ο boot loader και τέλος όλα αυτά ενοποιούνται σε ένα disk image. Η διαδικασία του χτίσματος φαίνεται και στην εικόνα 3.6.

Αρκετά ενδιαφέρουσες είναι και ορισμένες σχεδιαστικές επιλογές των δημιουργών που αφορούν κυρίως το network stack. Χρησιμοποιεί έναν virtio driver για την επικοινωνία με τον επόπτη, συνεπώς δε χρειάζεται ο τελευταίος να προσομοιώνει διάφορες συσκευές αλλά απλά να χρησιμοποιεί το virtio. Επιπροσθέτως, η επιλογή να φτιαχτεί από την αρχή το συγκεκριμένο framework, έδωσε τη δυνατό-



Σχήμα 3.6: IncludeOS building process

τητα στη δημιουργία ενός νέου TCP-IP stack, χρησιμοποιώντας σε μεγάλο βαθμό τις δυνατότητες που έδινε η C++. Τέλος, το IncludeOS δεν έχει κάποιο τρόπο να φορτώσει ένα πρόγραμμα, συνεπώς δεν υπάρχει η κλασική main συνάρτηση. Αντ' αυτού, υπάρχει η κλάση Service και ο προγραμματιστής πρέπει να υλοποιήσει την Service::start η οποία θα κληθεί μετά την αρχικοποίηση του λειτουργικού συστήματος.

Όπως έχει αναφερθεί προηγουμένως, μόνο C++ προγράμματα μπορούν να εκτελεστούν στο IncludeOS, ενώ μπορεί να χρησιμοποιηθεί και η standard βιβλιοθήκη της C. Με αυτό τον τρόπο υπάρχει μία μερική υποστήριξη για POSIX εφαρμογές. Ιδιαίτερα λειτουργίες που αφορούν την είσοδο/έξοδο (I/O) και τα νήματα δεν υποστηρίζονται. Στο IncludeOS όλη η εργασία γίνεται από ένα και μόνο νήμα, χωρίς I/O μπλοκάρισμα ή context switching μέσα στο guest, με το συγκεκριμένο μοντέλο να θυμίζει εκείνο του Node.js. Όπως και στα υπόλοιπα unikernel frameworks μόνο μία διεργασία μπορεί να υπάρχει ανά unikernel, ενώ δεν υπάρχει και εικονική μνήμη.

3.5 ClickOS

Το ClickOS είναι ένα unikernel framework το οποίο δημιουργήθηκε με γνώμονα την υποστήριξη εφαρμογών Network Function Virtualization (NFV). Η ιδέα του NFV είναι η μεταφορά της εργασίας που γίνεται στα middleboxes από συσκευές υλικού σε ειδικό λογισμικό που μπορεί να εκτελεστεί σε οποιοδήποτε υλικό σαν μία εφαρμογή. Τα middleboxes είναι συσκευές δικτύου, οι οποίες μετατρέπουν, ελέγχουν, φιλτράρουν και γενικότερα χειρίζονται διαδικτυακή κίνηση για σκοπούς διαφορετικούς από την προώθηση πακέτων. Οι συσκευές αυτές συνήθως είναι ακριβές, δύσκολα διαχειρίσιμες και είναι αρκετά δύσκολο να αλλάξει η λειτουργία τους. Χρησιμοποιώντας τις τεχνολογίες εικονικοποίησης το NFV αναπτύσσει τις δικτυακές λειτουργίες σε εικονικές μηχανές. Με αυτό τον τρόπο αντιμετωπίζονται τα παραπάνω προβλήματα των middleboxes, ενώ παράλληλα επιτρέπει σε παραπάνω από μία λειτουργία να εκτελεστεί στο ίδιο υλικό.

Το ClickOS μετατρέπει το Click modular router [16] σε μία εικονική μηχανή που μπορεί να εκτελεστεί στο Xen. Για να γίνει αυτό εφικτό χρησιμοποιείται το mini-os (βλέπε ενότητα 3.8). Η δομή του ClickOS είναι αρκετά απλή με το click να εκτελείται στο mini-os όπως φαίνεται και στην εικόνα 3.7. Το mini-os διαθέτει τα βασικά χαρακτηριστικά που χρειάζονται για να μπορέσει να εκτελεστεί το Click, όπως διαχείριση μνήμης, τους απαραίτητους drivers και έναν απλό χρονοδρομολογητή νημάτων. Έτσι σε κάθε εικόνα του ClickOS μπορούν αν συνυπάρχουν περισσότερα από ένα click modular routers, κάθε ένα από τα οποία εκτελείται σε διαφορετικό νήμα. Το configuration

για κάθε router γράφεται στο xenstore και ένα ειδικό νήμα είναι υπεύθυνο για την παρακολούθηση τυχόν αλλαγών σε αυτό και την εφαρμογή του configuration στο αντίστοιχο router. Ωστόσο κάθε φορά ένα νήμα εκτελείται, ενώ η χρονοδρομολόγηση είναι non-preemptive.

Το ClickOS παρουσιάζει αρκετά υψηλές δικτυακές επιδόσεις, ενώ παράλληλα διαθέτει και άλλα θετικά στοιχεία των unikernels, όπως μικρό μέγεθος, γρήγορη εκκίνηση της εικονικής μηχανής και μικρή κατανάλωση μνήμης. Σημαντικό ρόλο στις επιδόσεις του δικτύου είχαν οι αλλαγές που έγιναν σε όλη την αρχιτεκτονική του δικτύου στο Xen. Το Xen ακολουθεί το μοντέλο του split driver για το δίκτυο, με ένα driver (netback) να υπάρχει στο driver domain και έναν άλλο driver (netfront) στο παραεικονικοποιημένο λειτουργικό σύστημα, οι οποίοι drivers επικοινωνούν μεταξύ τους μέσω κοινής μνήμης. Προκειμένου λοιπόν, να πετύχει τόσο υψηλή απόδοση το ClickOS έγιναν σημαντικές αλλαγές τόσο στο netfront και netback driver όσο και σε άλλα κομμάτια του Xen που εμπλέκονται για την εξυπηρέτηση των δικτυακών λειτουργιών όπως το switch [19].



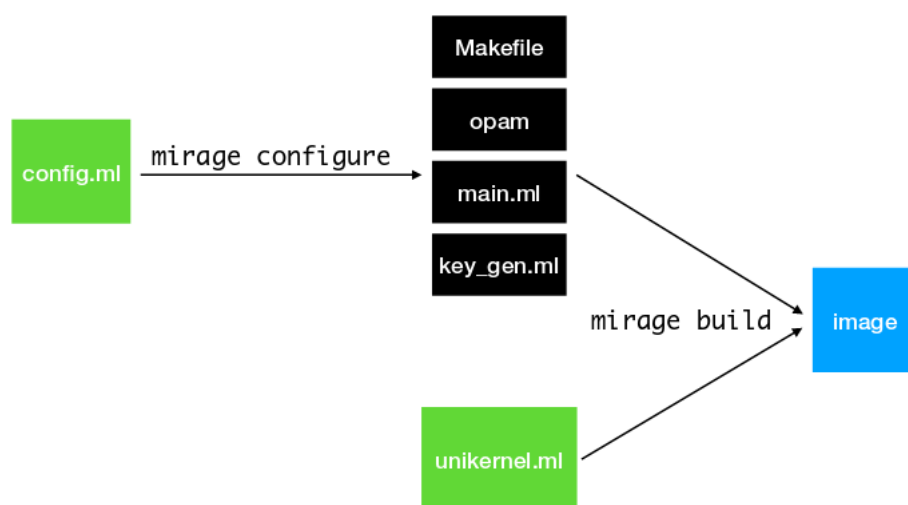
Σχήμα 3.7: Clickos design

3.6 MirageOS

Το MirageOS είναι ένα από τα παλιότερα και πιο διαδεδομένα unikernel frameworks. Σε αντίθεση με όσα είδαμε προηγουμένως, το MirageOS είναι γραμμένο σε μία υψηλού επιπέδου γλώσσα την OCaml. Μερικά από τα βασικά πλεονεκτήματα που προσφέρει η χρήση μία τέτοιας γλώσσας είναι η δημιουργία type safe προγραμμάτων και η μείωση των memory leaks. Η έννοια του type safety αφορά την αποθάρρυνση ή αποτροπή σφαλμάτων τύπων (type errors). Τα σφάλματα τύπων δημιουργούνται όταν υπάρχει ασυμφωνία μεταξύ των τύπων δεδομένων σε μία εντολή και μπορεί να οδηγήσουν σε λανθασμένη ή ανεπιθύμητη συμπεριφορά του προγράμματος. Η OCaml κάνει στατικό έλεγχο τύπων (static type checking) οπότε βρίσκει σφάλματα τύπων κατά τη διάρκεια της μεταγλώττισης και όχι κατά τη διάρκεια της εκτέλεσης. Επιπλέον διαχειρίζεται μόνη της τη μνήμη ενώ ο

συλλέκτης σκουπιδιών (garbage collector) έχει σχεδιαστεί με τέτοιο τρόπο ώστε να μειώνονται οι διακοπές.

Η επιλογή της γλώσσας είχε ως συνέπεια και τη δημιουργία από την αρχή πολλών συστατικών ενός λειτουργικού συστήματος. Με αυτό τον τρόπο δημιουργήθηκαν βιβλιοθήκες για το network και storage stack, αλλά και για διάφορα πρωτόκολλα όπως τα TCP/IP, SSH, HTTP, DNS, Openflow, XMPP και Xen inter-VM transports. Γενικότερα η αρχιτεκτονική του MirageOS είναι αρθρωτή και κάθε κομμάτι του είναι σε μορφή βιβλιοθηκών και κάθε φορά οι απαραίτητες βιβλιοθήκες συνδέονται μεταξύ τους για τη δημιουργία μία συγκεκριμένης εφαρμογής (π.χ. μία ιστοσελίδα). Η διαδικασία για τη δημιουργία και εκτέλεση εφαρμογών στο MirageOS είναι αρκετά απλή. Αρχικά, η δημιουργία και ο έλεγχος της εφαρμογής μπορεί να γίνει χρησιμοποιώντας τον mirage compiler, ο οποίος μπορεί να δημιουργήσει εκτελέσιμα τόσο για unix συστήματα όσο και για το MirageOS. Συνεπώς, αρχικά μπορεί η εφαρμογή να εκτελεστεί σε unix συστήματα, ώστε να γίνει πιο εύκολα και η αποσφαλμάτωση της και στη συνέχεια να ξαναμεταγλωττιστεί για να αποτελέσει ένα πλήρη unikernel. Η εικόνα αυτή περιέχει την εφαρμογή, το configuration αυτής, τις κατάλληλες βιβλιοθήκες, υποστήριξη για την εκκίνηση του συστήματος και τον garbage collector. Το MirageOS μπορεί να εκτελεστεί σε Xen, KVM αλλά και σε arm αρχιτεκτονικές με τη χρήση του Solo5.



Σχήμα 3.8: mirage compiler

Σε μετρήσεις που έχουν πραγματοποιήσει οι δημιουργοί του, δείχνουν ότι παρά την επιλογή μίας υψηλού επιπέδου γλώσσας, το MirageOS παρουσιάζει υψηλή απόδοση τόσο σε θέματα δικτύου όσο και σε μνήμη και χρόνους εκκίνησης [18]. Μάλιστα παρουσιάζει αρκετά καλή συ-

μπεριφορά και στη χρήση περισσότερων του ενός νήματος. Η λειτουργία αυτή γίνεται εφικτή με τη χρήση μίας ξεχωριστής βιβλιοθήκης η οποία διαχειρίζεται τα νήματα. Επιπλέον, σε αντίθεση με τα περισσότερα λειτουργικά συστήματα το μέγεθος του κώδικα είναι αρκετά μικρότερο. Από την άλλη βέβαια η επιλογή της συσγκεκριμένης γλώσσας επιτρέπει την υποστήριξη εφαρμογών που είναι γραμμένες σε αυτήν και μόνο. Ωστόσο, οι περισσότερες εφαρμογές που υπάρχουν δεν είναι γραμμένες σε OCaml και συνεπώς η εκτέλεση τους στο MirageOS απαιτεί την ολική επανασχεδίαση τους και μεταφορά τους στην OCaml.

3.7 Solo5

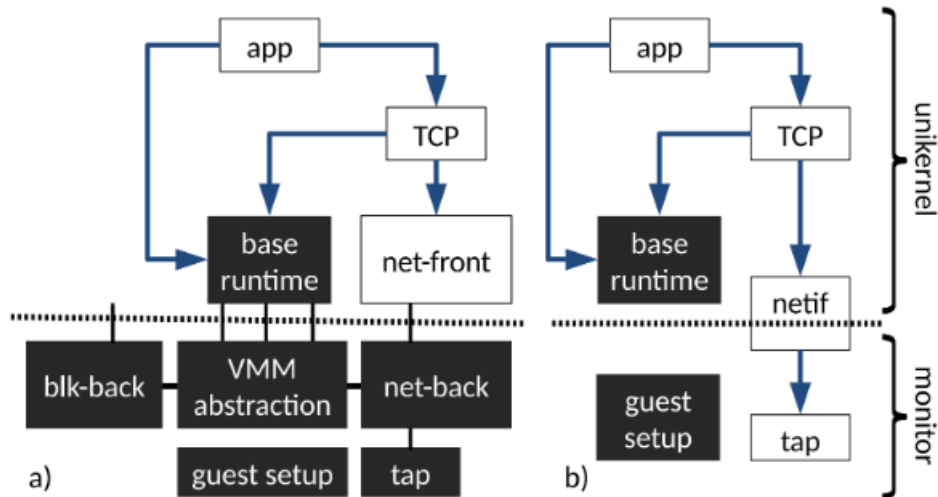
Το Solo5 δεν είναι ένα πλήρες unikernel framework, αλλά μία βάση πάνω στην οποία μπορεί να φτιαχτεί ένα unikernel. Αρχικά σχεδιάστηκε με σκοπό την εκτέλεση του MirageOS στο KVM, αλλά σύντομα εξελίχθηκε σε ένα γενικότερο περιβάλλον εκτέλεσης που προσφέρει απομόνωση (sandbox), κατάλληλο για την εκτέλεση εφαρμογών που έχουν χτιστεί χρησιμοποιώντας διάφορα unikernel frameworks, με στόχο διάφορες τεχνολογίες sandboxing σε διάφορα host λειτουργικά συστήματα και hypervisors [2]. Σκεφτόμενοι τα unikernels ως διεργασίες, αυτό που κάνει το Solo5 είναι να επανασχεδιάζει τη διεπαφή μεταξύ της διεργασίας και του πυρήνα ή του hypervisor [3] με στόχο η νέα διεπαφή να είναι:

- όσο το δυνατόν πιο legacy-free και "λεπτό" σε σχέση με τις υπάρχουσες υλοποιήσεις. Δηλαδή να εκθέτει όσο το δυνατόν λιγότερα σε σχέση με τον πυρήνα ή με έναν hypervisor.
- να είναι πιο εύκολη η υποστήριξη νέων τεχνολογιών sandboxing, λειτουργικά συστήματα και hypervisors.

Μερικά από τα unikernels που χρησιμοποιούν το Solo5 είναι το MirageOS και το IncludeOS.

Σε συνδυασμό με το ukvm [26] το Solo5 λειτουργεί ως ένα πρόγραμμα φύλακα (tender) αποτελώντας ένα εξειδικευμένο monitor για unikernels. Ο ρόλος αυτός μοιάζει με το ρόλο του QEMU στην περίπτωση του συνδυασμού QEMU/KVM. Ωστόσο το QEMU είναι ένα monitor γενικού σκοπού σε αντίθεση με το Solo5 που είναι μόνο για unikernels. Η λογική πίσω από αυτό είναι ίδια με αυτή των unikernels, τα οποία αποτελούνται μόνο από τα απαραίτητα στοιχεία που χρειάζεται η εφαρμογή. το Solo5 αντίστοιχα εκθέτει μόνο τα απαραίτητα στοιχεία που χρειάζεται ένα Unikernel για να εκτελεστεί. Τα πλεονεκτήματα που προσφέρονται είναι περισσότερη ασφάλεια, καθώς μειώνεται το attack surface, ταχύτεροι χρόνοι εκκίνησης καθώς δε

χρειάζεται να αρχικοποιηθούν συσκευές που δε θα χρειαστούν. Στην παρακάτω εικόνα 3.9 βλέπουμε τις διαφορές μεταξύ ενός ειδικού και γενικού σκοπού monitor. Το Solo5 ως tender μπορεί να εκτελεστεί σε x86_64 και arm 64-bit (aarch64 αρχιτεκτονικές μαζί με Linux/KVM. Έτσι τα unikernels που χρησιμοποιούν ως βάση το solo5 έχουν τη δυνατότητα να εκτελούνται σε αυτές τις αρχιτεκτονικές.



Σχήμα 3.9: a) γενικού σκοπού monitor b) ειδικού σκοπού monitor

3.8 Mini-OS

Το Mini-OS είναι ένας μικρός πυρήνας που παρέχεται μαζί με το Xen και μπορεί να αποτελέσει τη βάση για τη δημιουργία παραεικονικοποιημένων λειτουργικών συστημάτων. Μία από τις βασικές του χρήσεις είναι η δημιουργία stub domains στο xen. Ένα stub domain είναι ένα ελαφρύ service ή driver domain που έχει ως στόχο να μειώσει το φόρτο εργασίας του QEMU από το dom0 σε ένα διαφορετικό domain. Με αυτό τον τρόπο επιτυγχάνεται καλύτερη απόδοση και περισσότερη ασφάλεια. Μπορεί να εκτελεστεί τόσο σε x86 αρχιτεκτονικές όσο και σε arm. Πολλά από τα γνώμοναωστά λειτουργικά συστήματα χρησιμοποιούν το Mini-OS για να μπορούν να είναι είναι συμβατά με το Xen.

Εκτός από λειτουργικά συστήματα, πολλά unikernel frameworks χρησιμοποιούν το Mini-OS ως βάση ώστε να μπορούν να εκτελεστούν

στο Xen, όπως το MirageOS, το RumpRun, το ClickOS και άλλα. Ωστόσο μπορεί να χρησιμοποιηθεί και μόνο του για τη δημιουργία unikernels. Διαθέτει TCP/IP stack χρησιμοποιώντας την LWIP βιβλιοθήκη και τη newlib ως βιβλιοθήκη για την C. Επιπλέον διαθέτει υποστήριξη για περισσότερα από ένα νήματα. Από την άλλη, το μικρό εύρος δυνατοτήτων που παρέχει κάνει δύσκολη τη μεταφορά μιας σύνθετης εφαρμογής σε MInI-OS και κατά συνέπεια τη δημιουργία unikernels.

3.9 Και πολλά ακόμα

Εκτός από τα παραπάνω υπάρχουν πολλά ακόμα unikernel frameworks. Μερικά ακολουθώντας το παράδειγμα του MirageOS, χρησιμοποιούν γλώσσες υψηλού επιπέδου όπως το Clive που χρησιμοποιεί την Go, το Runtime.js που χρησιμοποιεί Javascript, το HalVM που χρησιμοποιεί Haskell και το LING που χρησιμοποιεί Erlang. Επιπλέον υπάρχει και το Drawbridge που αποτελεί ένα library operating system βασισμένο στο windows.

Πέρα από τα unikernel frameworks, υπάρχουν και πολλά projects τα οποία έχουν άμεση σχέση με αυτά. Μερικά τέτοια παραδείγματα είναι το Unik, το οποίο αποτελεί ένα εργαλείο που βοηθάει στη μεταγλώττιση και στην εντοπισμό των unikernels. Ουσιαστικά πρόκειται για κάτι ανάλογο με το docker αλλά για unikernels. Ένα πολύ ενδιαφέρον project είναι το Jitsu, ένα forwarding DNS server που εκκινεί αυτόματα unikernels όταν χρειαστεί.

Κεφάλαιο 4

Σχεδιασμός και υλοποίηση

Όπως είδαμε στο προηγούμενο κεφάλαιο κανένα unikernel framework δεν υποστηρίζει την κλήση συστήματος `fork` και επομένως εφαρμογές που χρησιμοποιούν τη συγκεκριμένη κλήση δεν μπορούν να υποστηριχτούν από αυτά. Στο πλαίσιο, λοιπόν, της παρούσας διπλωματικής εργασίας δημιουργήσαμε ένα μηχανισμό που υλοποιεί τις κλήσεις συστήματος `fork` και `pipe` σε KVM hypervisor. Ο στόχος ήταν να διατηρηθεί το `single process` χαρακτηριστικό των unikernels. Για το λόγο αυτό το αποτέλεσμα της κλήσης `fork`, δεν είναι η δημιουργία μία νέας διεργασίας στο υπάρχον unikernel, αλλά η εκκίνηση ενός unikernel, ίδιου με το αρχικό.

Επιπλέον, υλοποιήθηκε και ένας `inter-vm communication` μηχανισμός, στα πρότυπα της κλήσης συστήματος `pipe`. Δηλαδή δύο εικονικά μηχανήματα μπορούν να επικοινωνούν μεταξύ τους, όπως δύο διεργασίες επικοινωνούν μεταξύ τους, μέσω της κλήσης συστήματος `pipe`, σε ένα λειτουργικό σύστημα γενικού σκοπού.

Στο παρών κεφάλαιο περιγράφουμε αναλυτικά τους δύο αυτούς μηχανισμούς. Το κεφάλαιο χωρίζεται σε δύο μέρη, ένα για το μηχανισμό `pipe` και ένα για το μηχανισμό `fork`. Σε κάθε μέρος περιγράφεται αναλυτικά τόσο ο μηχανισμός, όσο και η διαδικασία και τα στάδια μέχρι την τελική υλοποίηση τους.

4.1 Μηχανισμός pipe

Σύμφωνα με το POSIX ο ορισμός της κλήσης συστήματος pipe έχει ως εξής:

```
int pipe(int fildes[2]);
```

Ο μηχανισμός pipe που υλοποιήθηκε ακολουθεί τον ορισμό του POSIX. Πιο συγκεκριμένα μία κλήση στη συγκεκριμένη συνάρτηση δημιουργεί ένα pipe μεταξύ των δύο εικονικών μηχανημάτων και δημιουργεί δύο νέους file descriptors. Οι δύο αυτοί file descriptors αποθηκεύονται στις παραμέτρους fildes[0] και fildes[1]. Ο πρώτος file descriptor αφορά το read κομμάτι του pipe, ενώ ο δεύτερος το write κομμάτι του pipe. Η τιμή που επιστρέφει η κλήση συστήματος είναι 0, αν η δημιουργία του pipe ήταν επιτυχής, ενώ διαφορετικά θα επιστραφεί -1 και η μεταβλητή errno περιέχει την τιμή του σφάλματος.

Η ανάγνωση από το pipe γίνεται χρησιμοποιώντας τον πρώτο file descriptor από τις παραμέτρους της κλήσης. Τα δεδομένα διαβάζονται με FIFO (first in first out) σειρά. Ενώ αν δεν υπάρχουν δεδομένα στο pipe τότε η κλήση "μπλοκάρει" μέχρι να προκύψουν δεδομένα από το write άκρο του pipe. Η εγγραφή στο pipe γίνεται χρησιμοποιώντας το δεύτερο file descriptor από τις παραμέτρους της κλήσης. Η εγγραφή στο pipe μπορεί να μπλοκάρει αν δεν υπάρχει χώρος στο pipe. Επιπλέον μπορεί να αποτύχει αν όλα τα file descriptors για ανάγνωση από το pipe έχουν κλείσει με το σφάλμα EPIPE.

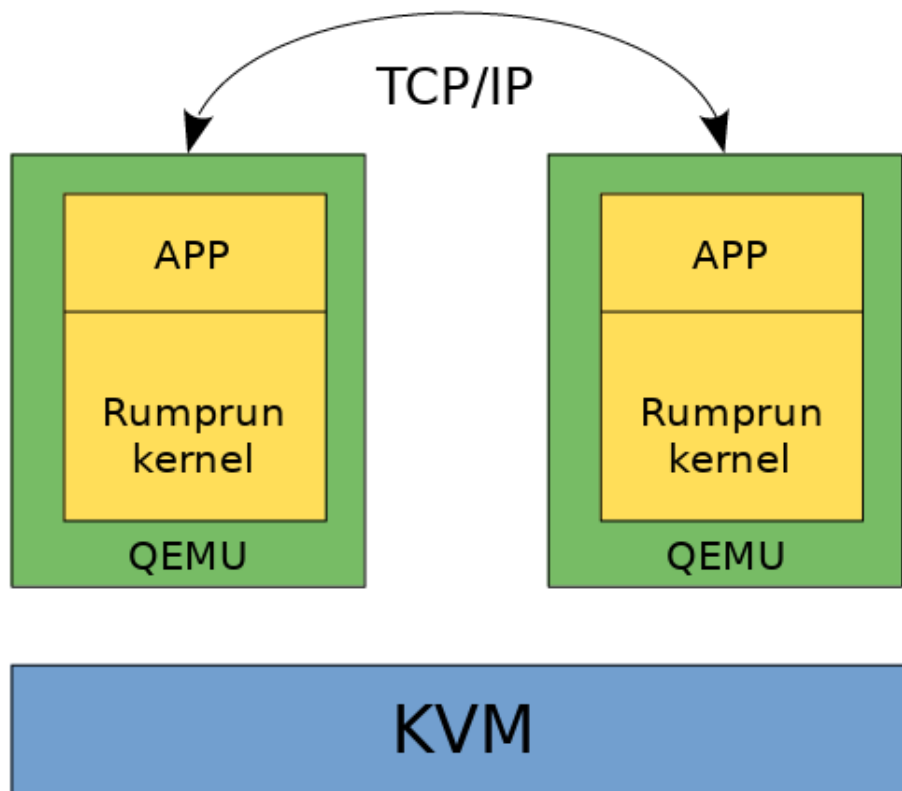
4.1.1 Στάδια υλοποίησης

Η υλοποίηση του συγκεκριμένου μηχανισμού έγινε σε 3 στάδια, τα οποία αναλύονται παρακάτω.

4.1.1.1 Στάδιο 1 - επίπεδο εφαρμογής

Στο πρώτο στάδιο, υλοποιήσαμε το pipe ως μία συνάρτηση που καλείται από την εφαρμογή. Η συνάρτηση αυτή χρησιμοποιεί TCP/IP sockets για να εγκαθιδρύσει την επικοινωνία μεταξύ των δύο εφαρμογών. Συγκεκριμένα χρησιμοποιούνται δύο sockets, ένα για την αποστολή δεδομένων και ένα για την παραλαβή. Οι file descriptors των δύο αυτών sockets, είναι και οι τιμές που αποθηκεύονται στις μεταβλητές fildes[0] και fildes[1]. Στην παρακάτω εικόνα 4.1 φαίνεται σχηματικά η υλοποίηση.

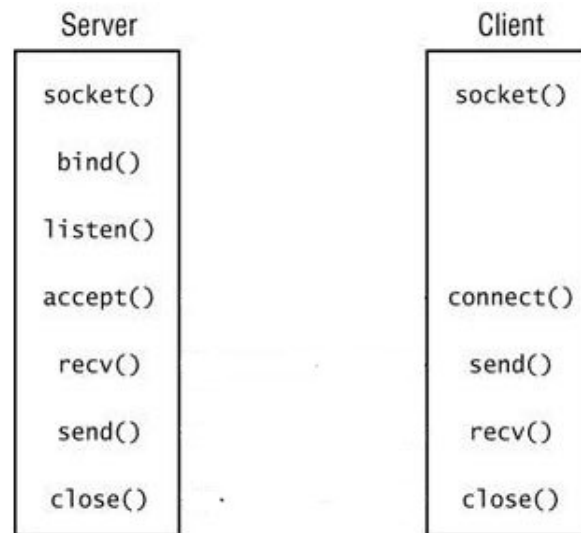
Όπως γίνεται εύκολα κατανοητό η συγκεκριμένη υλοποίηση απαιτεί την ύπαρξη δικτύου ανάμεσα στα δύο εικονικά μηχανήματα. Επιπλέον, λόγω της χρήσης των sockets, δεν υλοποιήθηκε όλα τα semantics



Σχήμα 4.1: Πρώτο στάδιο υλοποίησης μηχανισμού pipe

του pipe. Έτσι ακόμα και αν δεν υπάρχει ανοιχτό read άκρο στο pipe, οποιαδήποτε εγγραφή στο pipe θα είναι επιτυχής. Ακόμα, οποιαδήποτε εγγραφή μπορεί να μπλοκάρει μόνο αν ο παραλήπτης δεν μπορεί να δεχτεί άλλα δεδομένα. Αντιθέτως τα semantics του άκρου ανάγνωσης του pipe διατηρήθηκαν καθώς και στην περίπτωση των sockets, αν δεν υπάρχουν δεδομένα προς ανάγνωση η αντίστοιχη κλήση "μπλοκάρει".

Για να υλοποιηθεί η συνάρτηση pipe, ακολουθήθηκε η ίδια διαδικασία που ακολουθείται από μία εφαρμογή για να επικοινωνήσει με κάποια άλλη μέσω δικτύου. Η διαφορά είναι ότι η εφαρμογή έχει το ρόλο του server και του client ταυτόχρονα. Για αυτό το λόγο χρησιμοποιήθηκαν δύο sockets, ένα για κάθε ρόλο. Στην παρακάτω εικόνα 4.2 φαίνεται η αλληλουχία κλήσεων συστήματος για την εγκαθίδρυση της επικοινωνίας (server/client).

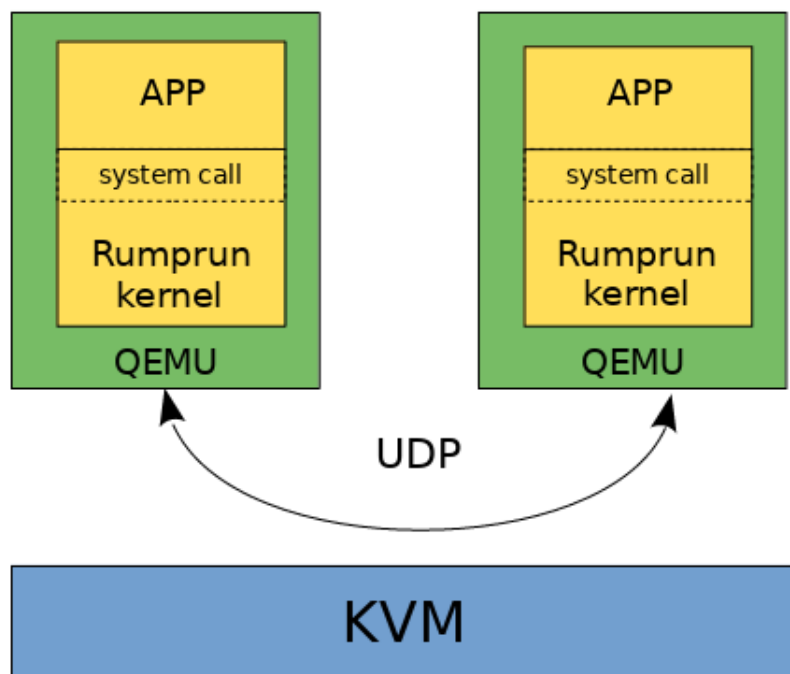


Σχήμα 4.2: Αλληλουχία κλήσεων για επικοινωνία μέσω TCP/IP sockets

Η συνάρτηση pipe, λοιπόν ακολουθεί και τις δύο διαδικασίες που φαίνονται (server/client). Αυτό όμως προξενεί ένα πρόβλημα κατά τη δημιουργία της επικοινωνίας. Αν και οι δύο εφαρμογές ακολουθήσουν την παραπάνω ακολουθία κλήσεων τότε θα δημιουργηθεί αδιέξοδο, καθώς και οι 2 θα έχουν κολλήσει στην κλήση accept, περιμένοντας κάποια σύνδεση. Ο τρόπος με τον οποίο επιλύθηκε το παραπάνω πρόβλημα είναι ως εξής. Μία από τις δύο εφαρμογές να εκτελεί πρώτα τις κλήσεις που αφορούν το client κομμάτι και μετά αυτές που αφορούν το server. Αντίθετα η άλλη εφαρμογή, πρώτα εκτελεί τις κλήσεις που αφορούν το server και ύστερα αυτές που αφορούν το client μέρος.

4.1.1.2 Στάδιο 2 - pipe ως system call με UDP sockets

Στο δεύτερο στάδιο, υλοποιήσαμε το μηχανισμό του pipe μέσα στον πυρήνα του rumprun. Ουσιαστικά μετατρέψαμε το function call του προηγούμενου σταδίου σε system call. Σε αντίθεση με πριν, η επικοινωνία γίνεται μέσω UDP sockets, οπότε και είναι απαραίτητη η ύπαρξη δικτύου. Η επιλογή του πρωτοκόλλου UDP έναντι του TCP, έγινε για την πιο εύκολη δημιουργία και χρήση sockets στον πυρήνα του rumprun. Ο συγκεκριμένος μηχανισμός δίνει τη δυνατότητα να υπάρχει επικοινωνία μεταξύ δύο ξεχωριστών unikernels, ακόμα και αν αυτά δε μοιράζονται τον ίδιο host, μέσω μίας κλήσης συστήματος παρόμοια με την pipe(). Με αυτό τον τρόπο, οι εφαρμογές δε χρειάζονται να τροποποιηθούν σε μεγάλο βαθμό, από την αρχική τους έκδοση. Στην παρακάτω εικόνα 4.3 φαίνεται σχηματικά η υλοποίηση.



Σχήμα 4.3: Δεύτερο στάδιο υλοποίησης μηχανισμού pipe

Όσον αφορά τη χρήση του μηχανισμού, αυτή γίνεται ακολουθώντας την ίδια διαδικασία με τη χρήση της κλήσης pipe σε ένα UNIX σύστημα. Στις παραμέτρους `filides[0]`, `filides[1]`, αποθηκεύονται οι δύο file descriptors που θα χρησιμοποιηθούν για την ανάγνωση και εγγραφή αντίστοιχα. Σε περίπτωση επιτυχίας επιστρέφεται η τιμή 0, ενώ σε αντίθετη περίπτωση, στη μεταβλητή `errno` αποθηκεύεται ο κωδικός του σφάλματος. Δεδομένου, ότι η επικοινωνία γίνεται με UDP

sockets, τα semantics είναι ίδια με το προηγούμενο στάδιο. Μία σημαντική διαφορά, είναι ότι πρέπει να καθοριστεί η διεύθυνση IP του unikernel με το οποίο θα εγκαθιδρυθεί η επικοινωνία. Για το λόγο αυτό, μετά την κλήση της pipe() θα πρέπει μέσω της ioctl στο file descriptor της εγγραφής

```
ioctl(filides[1], SETIPADDR, htonl(IP_ADDRESS));
```

Η μορφή της IP διεύθυνσης θα πρέπει να είναι σε network byte order και για αυτό το σκοπό μπορεί να χρησιμοποιηθεί η συνάρτηση htonl. Ο τρόπος που λειτουργεί ο μηχανισμός έχει ως εξής:

1. Αρχικά καλώντας την pipe, επιστρέφονται τα δύο απαραίτητα file descriptors, ένα για εγγραφή και ένα για ανάγνωση.
2. Στη συνέχεια, πρέπει να οριστεί η διεύθυνση IP, στην οποία θέλουμε να στείλουμε τα δεδομένα. Αυτό γίνεται μέσω της προαναφερθείσας ioctl κλήσης.
3. Η εφαρμογή στέλνει δεδομένα χρησιμοποιώντας την κλήση write() και το file descriptor που αντιπροσωπεύει το άκρο εγγραφής του pipe.
4. Η εφαρμογή λαμβάνει δεδομένα χρησιμοποιώντας την κλήση read() και το file descriptor που αντιπροσωπεύει το άκρο ανάγνωσης του pipe.

Η μεταφορά του μηχανισμού στον πυρήνα του rumpun απαιτούσε και τις ανάλογες αλλαγές στις συναρτήσεις που χρησιμοποιήθηκαν, καθώς πλέον ο προγραμματισμός γινόταν εντός του πυρήνα. Ο πυρήνας που χρησιμοποιεί το rumpun είναι αυτός του NetBSD και σε σχέση με τον τρόπο που χρησιμοποιούνται τα sockets σε userspace, διαφέρει με τον αντίστοιχο σε kernelspace. Ο συγκεκριμένος μηχανισμός χρησιμοποιεί ένα struct, στο οποίο αποθηκεύονται το socket, το file descriptor και η διεύθυνση IP του παραλήπτη.

```
struct pipe_data {  
    struct socket *so;  
    uint32_t ip;  
    int fd;  
};
```

Παρακάτω, περιγράφεται τι γίνεται μέσα στον πυρήνα όταν εκτελούνται οι παραπάνω κλήσεις.

1. pipe: Αρχικά, δημιουργούνται δύο sockets με την socreate, ένα για αποστολή και ένα για παραλαβή δεδομένων και αποθηκεύονται στα αντίστοιχα πεδία του struct pipe_data. Το socket που

θα χρησιμοποιηθεί για ανάγνωση γίνεται bind στη θύρα 23456 με την sobind, επιτρέποντας συνδέσεις από κάθε IP. Στη συνέχεια, δημιουργούνται τα 2 file descriptors που θα χρησιμοποιούνται από την εφαρμογή. Τα file descriptors, από τη μεριά του πυρήνα αντιπροσωπεύονται από τις δομές file_t. Στις δομές αυτές αποθηκεύεται και το struct pipe_data που αναφέρθηκε προηγουμένως. Αν όλα έχουν πάει καλά, προστίθονται τα δύο file descriptors στην εφαρμογή και επιστρέφει με επιτυχία η pipe.

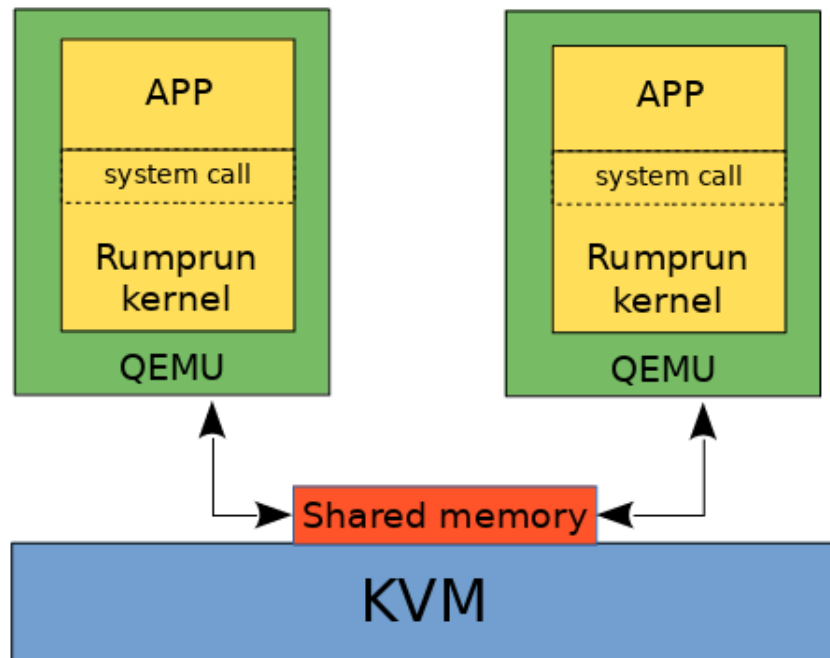
2. ioctl: Όταν καλείται η ioctl με το command SETIPADDR, αποθηκεύεται στο πεδίο ip του pipe_data η τιμή που έχει περαστεί ως τρίτη παράμετρος στην ioctl.
3. write: Αρχικά αρχικοποιείται το struct sockaddr_in, το οποίο λαμβάνει την IP του παραλήπτη από το πεδίο ip του pipe_data. Στη συνέχεια χρησιμοποιώντας τη sosend, στέλνονται τα δεδομένα μέσα από το socket.
4. read: Με τη χρήση της soreceive, διαβάζονται τυχόν δεδομένα στο socket. Αν δεν υπάρχουν δεδομένα η soreceive μπλοκάρει μέχρι να προκύψουν.

Στον πυρήνα του NetBSD, τυχόν δεδομένα που μεταφέρονται από userspace σε kernelspace αποθηκεύονται στο struct uio. Τόσο η soreceive, όσο και η sosend, δίνουν τη δυνατότητα να χρησιμοποιηθεί το συγκεκριμένο struct και συνεπώς δεν υπάρχει ανάγκη για παραπάνω αντιγραφές των δεδομένων. Η διαδικασία για την εισαγωγή μία νέας κλήσης συστήματος στο rump, είναι ακριβώς ίδια με αυτή που θα χρησιμοποιηθεί για την εισαγωγή μία κλήσης συστήματος στον πυρήνα του NetBSD, με ένα επιπλέον βήμα. Αφού εισαχθεί η κλήση συστήματος στο NetBSD, πρέπει να εισαχθεί στο αρχείο src-netbsd/sys/rump/librump/rumpkern η κλήση συστήματος. Σε αυτό το σημείο, καλό είναι να ξανααναφερθεί ότι στα unikernels δεν υπάρχει διαχωρισμός μεταξύ userspace και kernelspace, ωστόσο οι συγκεκριμένοι όροι χρησιμοποιούνται για να γίνει διαχωρισμός μεταξύ του κώδικα μίας οποιασδήποτε εφαρμογής και του κώδικα του πυρήνα του unikernel.

4.1.1.3 Στάδιο 3 - pipe ως function call με shared memory

Στο τελευταίο στάδιο, υλοποιήσαμε το μηχανισμό pipe, πάλι ως system call με τη διαφορά ότι πλέον δε χρησιμοποιούνται sockets, αλλά κοινή μνήμη μεταξύ των εικονικών μηχανημάτων. Πλέον δεν υπάρχει ανάγκη για ύπαρξη δικτύου μεταξύ των εικονικών μηχανημάτων, ωστόσο η συγκεκριμένη υλοποίηση μπορεί να χρησιμοποιηθεί μόνο για unikernels, που μοιράζονται τον ίδιο host. Για την κοινή μνήμη χρησιμοποιήθηκε το ivshmem [17]. Όπως έχει ήδη αναφερθεί

πρόκειται για ένα μηχανισμό, που επιτρέπει το διαμοιρασμό μνήμης μεταξύ του host και των εικονικών μηχανών που εκτελούνται στο host. Η χρήση του γίνεται μέσω μίας PCI συσκευής. Συνεπώς έπρεπε να δημιουργήσουμε τον κατάλληλο driver που θα μας επιστρέψει να το χρησιμοποιήσουμε. Στην παρακάτω εικόνα 4.4 φαίνεται σχηματικά η υλοποίηση.



Σχήμα 4.4: Τρίτο στάδιο υλοποίησης μηχανισμού pipe

Η χρήση του pipe, από την εφαρμογή γίνεται πλέον όπως σε κάθε UNIX λειτουργικό σύστημα. Καλώντας την pipe, αν όλα πάνε καλά αποθηκεύονται στις παραμέτρους `filides[0]`, `filides[1]`, οι δύο file descriptors που θα χρησιμοποιηθούν για την ανάγνωση και εγγραφή αντίστοιχα. Σε περίπτωση επιτυχίας επιστρέφεται η τιμή 0, ενώ σε αντίθετη περίπτωση, στη μεταβλητή `errno` αποθηκεύεται ο κωδικός του σφάλματος. Επιπλέον έχουν υλοποιηθεί και όσα semantics δεν είχαν υλοποιηθεί στα προηγούμενα στάδια. Οπότε, σε περίπτωση που δεν υπάρχουν ανοιχτά άκρα ανάγνωσης στο pipe, η εγγραφή δε θα είναι επιτυχής. Αν δεν υπάρχει αρκετός χώρος για την εγγραφή δεδομένων, τότε η write "μπλοκάρει".

Θα ξεκινήσουμε την περιγραφή της υλοποίησης από την υλοποίηση του PCI driver για το `ivshmem`. Αρχικά πρέπει να ενημερώσουμε το `qemu` για τη χρήση του συγκεκριμένου μηχανισμού. Για να γίνει αυτό χρησιμοποιούμε τις ακόλουθες παραμέτρους για το `qemu`.


```
-device ivshmem-plain,memdev=hostmem -object  
memory-backend-file,size=1M,share,mem-path=/dev/shm/ivshmem,id=hostmem
```

Στην παράμετρο `size`, ορίζουμε τη χωρητικότητα του pipe και για κάθε pipe πρέπει να ορίσουμε διαφορετικό `mem-path`. Κάθε unikernel που ξεκινάει με τις συγκεκριμένες παραμέτρους μπορεί να χρησιμοποιήσει το pipe.

Από τη μεριά του unikernel, χρειάζεται να κατασκευάσουμε έναν PCI driver για το μηχανισμό `ivshmem`. Όπως και στην περίπτωση του system call έτσι και στην περίπτωση του driver, αρχικά πρέπει να φτιάξουμε το driver για το NetBSD. Ακολουθώντας το Kernel Development Manual του NetBSD [1], για τον pci device driver υλοποιήσαμε τρεις συναρτήσεις:

1. `match`: Είναι η συνάρτηση που καλείται από το autoconfiguration framework του NetBSD, όταν το σύστημα εκκινεί, ώστε να ταιριαστεί ο driver με τη συσκευή. Οπότε στην περίπτωση του δικού μας driver η συγκεκριμένη συνάρτηση ελέγχει αν πρόκειται για `ivshmem pci device`, ελέγχοντας τις τιμές του κατασκευαστή και το `id` της συσκευής.
2. `attach`: Καλείται μόνο αν η ανίχνευση της συσκευής ήταν επιτυχής (αν η `match` επέστρεψε 1). Ο ρόλος της είναι να αρχικοποιήσει τη συσκευή. Δεδομένου, ότι εμείς χρησιμοποιούμε το συγκεκριμένο μηχανισμό απλά για διαμοιρασμό μνήμης, αρκεί να κάνουμε `map` το `PCI_BAR(2)` του `ivshmem` και να αποθηκεύσουμε τις απαραίτητες μεταβλητές για τη χρήση του μηχανισμού.
3. `detach`: Καλείται σε περίπτωση που η συσκευή αποσυνδεθεί. Στη δική μας περίπτωση είναι αρκετό να κάνουμε `unmap` το `bus_space` που χρησιμοποιούμε και να μηδενίσουμε τις μεταβλητές που χρησιμοποιούμε για τη συσκευή.

Προκειμένου να μπορεί να χρησιμοποιηθεί η συσκευή από τον υπόλοιπο πυρήνα του NetBSD, είναι απαραίτητο να αποθηκευτούν οι απαραίτητες μεταβλητές που αναφέρθηκαν στις συναρτήσεις `attach` και `detach`. Για το λόγο αυτό δημιουργήθηκε το `struct ivshm`, που έχει ως πεδία τις μεταβλητές αυτές.

```
struct ivshm {  
    bus_size_t          data_s; /* size of shared  
        memory */  
    bus_addr_t          data_b; /* base address  
        of shared memory */  
    bus_space_tag_t     data_t; /* bus tag for  
        shared memory */  
};
```

```

        bus_space_handle_t      data_h; /* bus handle for
        shared memory */
};

```

Ιδανικά θα θέλαμε να κάνουμε map όλη την κοινή μνήμη του ivshmem, ώστε να μπορεί να χρησιμοποιηθεί σαν μνήμη του πυρήνα. Ωστόσο, το rumpun δεν υποστηρίζει το mapping μνήμης μέσω του bus. Ως εκ τούτου, η πρόσβαση στην κοινή μνήμη γίνεται κάθε φορά με εγγραφή και ανάγνωση bytes από αυτήν χρησιμοποιώντας το bus. Υπό αυτές τις συνθήκες, υπήρχαν τέσσερις μεταβλητές που ήταν απαραίτητες:

1. το μέγεθος της κοινής μνήμης.
2. η διεύθυνση βάσης της κοινής μνήμης
3. το bus tag της κοινής μνήμης
4. και το bus handle της κοινής μνήμης.

Οι 2 τελευταίες μεταβλητές, είναι αυτές που χρησιμοποιούνται από τις συναρτήσεις που παρέχει το NetBSD για να μπορέσει χρησιμοποιώντας το bus, να επικοινωνήσει με τη συσκευή. Ο ρόλος των υπόλοιπων δύο μεταβλητών θα γίνει πιο ξεκάθαρος παρακάτω.

Αφού λοιπόν, δημιουργήσαμε το driver για το NetBSD και ενημερώσαμε το autoconfiguration framework για την ύπαρξη του, έπρεπε να κάνουμε γνωστή την ύπαρξη του και στο rumpun. Για να γίνει αυτό απαιτούνται δύο ενέργειες. Πρώτον προσθέτουμε το pci driver στο src-netbsd/sys/rump/dev/Makefile.rumpdevcomp. Δεύτερον, να δημιουργηθεί ένα νέο directory στο src-netbsd/sys/rump/dev/lib/ για τη νέα συσκευή. Το directory νατό περιέχει το ioconf και το απαραίτητο Makefile για να προστεθεί ο οδηγός μας στο build σύστημα του rumpun.

```

1 ioconf ivshmem
2
3 include "conf/files"
4 include "dev/pci/files.pci"
5 include "rump/dev/files.rump"
6
7 pseudo-root pci*
8
9 ivshmem* at pci? dev ? function ?

```

Code 4.1: IVSHMEM.ioconf

```

1 RUMPTOP=${TOPRUMP}
2
3 .PATH:  ${RUMPTOP}/../dev/pci
4
5 LIB=    rumpdev_ivshmem
6 COMMENT=ivshmem
7
8 IOCONF= IVSHMEM.ioconf
9 RUMP_COMPONENT=ioconf
10
11 SRCS+=  ivshmem.c
12 ##SRCS+=      ivshmem_component.c
13
14 CPPFLAGS+= -I${RUMPTOP}/librump/rumpkern
15
16 .include  "${RUMPTOP}/Makefile.rump"
17 .include  <bsd.lib.mk>
18 .include  <bsd.klinks.mk>

```

Code 4.2: rumprun Makefile για το ivshmem

Η αρθρωτή οργάνωση των unikernels, δίνει τη δυνατότητα κάθε φορά να επιλέγονται μόνο τα απαραίτητα συστατικά για κάθε unikernel. Σε αυτό αποσκοπεί και η δημιουργία του παραπάνω directory, ώστε να μπορεί ο συγκεκριμένος driver να ενσωματωθεί στο unikernel, κατά τη διάρκεια του baking. Για τη χρήση του συγκεκριμένου οδηγού είναι απαραίτητη η εισαγωγή του `-lrumpdev_ivshmem` στο config με το οποίο θα γίνει bake η εφαρμογή.

Αφού δημιουργήσαμε τον driver, το επόμενο βήμα ήταν η αλλαγή του system call, ώστε αντί για sockets να χρησιμοποιείται ο μηχανισμός του ivshmem. Αρχικά, δε χρειαζόμαστε την ioctl κλήση και δημιουργούμε δύο structs, το struct pipe και το struct pipe_op. Οι ορισμοί τους φαίνονται παρακάτω:

```

1 struct pipe {
2     bus_size_t    init;           /* is shared
3     memory initialized? */
4     bus_size_t    lock;           /* pipe lock */
5     bus_size_t    wr_lock;        /* writers lock */
6     bus_size_t    nreaders;       /* number of
7     readers in pipe */
8     bus_size_t    nwriters;       /* number of
9     writers in pipe */
10    bus_size_t    len;             /* size of pipe
11    buffer */
12    bus_size_t    in;              /* pointer for

```

```

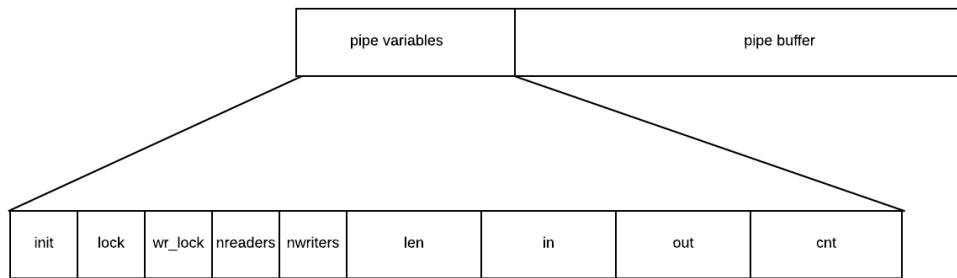
        next write */
9      bus_size_t    out;          /* pointer for
        next read */
10     bus_size_t    cnt;          /* number of
        bytes in pipe */
11     bus_size_t    buf;          /* pipe buffer */
12     int            pr_readers;   /* readers from
        this process */
13     int            pr_writers;   /* writers from
        curr process */
14 };
15
16 struct pipe_op {
17     int            oper;          /* operation in
        pipe 0 for read,
18                                     1 for write*/
19     struct pipe     *pipe;
20 };

```

Το struct pipe_op συνδέεται με τα file descriptors που δημιουργεί η κλήση pipe, ένα για κάθε file descriptor. Η μεταβλητή oper, καθορίζει αν πρόκειται για το file descriptor που αφορά την εγγραφή στο pipe ή αυτό που αφορά στην ανάγνωση. η άλλη τιμή είναι το struct pipe που είναι κοινό και για τα δύο άκρα του pipe και περιέχει βοηθητικές μεταβλητές για τη διαχείριση του pipe. Όπως έχει ήδη αναφερθεί η χρήση της κοινής μνήμης γίνεται μέσω του bus και οι συναρτήσεις εγγραφής ή ανάγνωσης από το bus απαιτούν ως παράμετρο τη διεύθυνση βάσης από την οποία θα διαβαστούν τα δεδομένα. Έτσι για να αναφερθούμε σε διαφορετική μεταβλητή μέσα σε αυτή τη μνήμη θα πρέπει να χρησιμοποιήσουμε διαφορετική διεύθυνση βάσης. Συνεπώς, το struct pipe περιέχει τις διευθύνσεις των κοινών μεταβλητών στην κοινή μνήμη. Επιπλέον, χρησιμοποιούνται και οι μεταβλητές pr_readers και pr_writers που αφορούν τα ανοιχτά άκρα στο pipe από τη συγκεκριμένη διεργασία. Η δύο τελευταίες μεταβλητές είναι απαραίτητες για τη διαχείριση του pipe όταν η διεργασία που έχει δημιουργήσει το pipe κάνει fork.

Αν θεωρήσουμε την κοινή μνήμη σαν ένα μεγάλο συνεχόμενο κομμάτι μνήμης, τότε στην παραπάνω εικόνα 4.5 φαίνεται πως την έχουμε διαχωρίσει εσωτερικά, με ένα κομμάτι της να αφορά τις μοιραζόμενες μεταβλητές και ένα άλλο τον buffer του pipe. Οι αρχικές μεταβλητές, έχουν μικρότερο μέγεθος από τις τελευταίες και η λειτουργία της κάθε μίας έχει ως εξής:

- init: Πρόκειται για μία μεταβλητή ελέγχου, που καθορίζει αν έχουν αρχικοποιηθεί οι κοινές μεταβλητές του pipe. Αλλάζεις



Σχήμα 4.5: shared memory layout

μόνο δύο φορές, όταν πρωτοδημιουργείται το pipe και όταν αυτό καταστρέφεται. Ουσιαστικά δηλώνει αν οι υπόλοιπες μεταβλητές περιέχουν "σκουπίδια", ή αν έχουν χρήσιμες τιμές.

- **lock:** Πρόκειται για το γενικό lock του pipe και προστατεύει όλες τις κοινές μεταβλητές του pipe.
- **wr_lock:** Πρόκειται για ένα lock που κάθε φορά επιτρέπει σε ένα και μόνο unikernel να γράφει στο pipe.
- **nreaders:** Ο αριθμός των ανοιχτών άκρων ανάγνωσης για το pipe
- **writers:** Ο αριθμός των ανοιχτών άκρων εγγραφής για το pipe
- **len:** Το μέγεθος του pipe buffer.
- **in:** Ο δείκτης που καθορίζει σε ποιο σημείο θα γραφτούν τα νέα δεδομένα στο pipe buffer.
- **out:** Ο δείκτης που καθορίζει από ποιο σημείο θα διαβαστούν δεδομένα από το pipe buffer
- **cnt:** Μετρητής των bytes που υπάρχουν στο pipe, κάθε στιγμή

Για την πρόσβαση στις παραπάνω μεταβλητές και γενικά για την κοινή μνήμη χρησιμοποιήθηκαν οι συναρτήσεις `bus_space_read_1`, `bus_space_read_4`, `bus_space_write_1`, `bus_space_write_4`. Ο αριθμός στο τέλος της κάθε συνάρτησης υποδηλώνει το μέγεθος των δεδομένων που θα διαβαστούν ή θα γραφτούν. Για την πιο εύκολη χρήση της κοινής μνήμης δημιουργήθηκαν 4 βοηθητικές συναρτήσεις, που φαίνονται παρακάτω.

```
1 /*
2  * Read a region of bytes in bus (data is 1 byte)
```

```

3  */
4 void read_region_1(bus_size_t offset, uint8_t *datap,
   bus_size_t count)
5 {
6     int i;
7     for (i=0; i<count; i++) {
8         datap[i] =
           bus_space_read_1(sharme.data_t,
           sharme.data_h,
           offset + i);
9     }
10 }
11 return;
12 }
13
14 /*
15  * Read a region of bytes in bus (data is 4 byte)
16  */
17 void read_region_4(bus_size_t offset, uint32_t *datap,
   bus_size_t count)
18 {
19     int i;
20     for (i=0; i<count; i++) {
21         datap[i] =
           bus_space_read_4(sharme.data_t,
           sharme.data_h,
           offset + i*4);
22     }
23 }
24 return;
25 }
26
27 /*
28  * Write in a region of bytes in bus (data is 1 byte)
29  */
30 void write_region_1(bus_size_t offset, uint8_t *datap,
   bus_size_t count)
31 {
32     int i;
33     for (i=0; i<count; i++) {
34         bus_space_write_1(sharme.data_t,
           sharme.data_h, offset + i,
           datap[i]);
35     }
36 }
37 return;
38 }
39

```

```

40 /*
41  * Write in a region of bytes in bus (data is 4 byte)
42  */
43 void write_region_4(bus_size_t offset, uint32_t *datap,
    bus_size_t count)
44 {
45     int i;
46     for (i=0; i<count; i++) {
47         bus_space_write_4(sharme.data_t,
            sharme.data_h, offset + i*4,
48                             datap[i]);
49     }
50     return;
51 }

```

Οι παραπάνω συναρτήσεις διαβάζουν ή γράφουν σε μία περιοχή μνήμης, είτε 1 είτε 4 bytes δεδομένα. Το μέγεθος της περιοχής που θα γίνει η πρόσβαση καθορίζεται από την παράμετρο count και το μέγεθος των bytes (count * bytes_of_data). Οι άλλες 2 παράμετροι των συναρτήσεων είναι η διεύθυνση βάσης για το bus και ένας πίνακας, στον οποίο αποθηκεύονται τα δεδομένα που διαβάστηκαν από μνήμη είτε τα δεδομένα που θα γραφτούν στην περίπτωση, στις λειτουργίες read και write αντίστοιχα. Οι συναρτήσεις bus_space, χρησιμοποιούν τη δομή struct ivshmem, για τις μεταβλητές bus tag και bus handle. Όλες οι αναφορές στην κοινή μνήμη, πέρα από αυτές των locks, γίνονται χρησιμοποιώντας αυτές τις συναρτήσεις.

Η ύπαρξη κοινής μνήμης μεταξύ διαφορετικών unikernels, απαιτεί και την ύπαρξη κάποιου συγχρονισμού μεταξύ αυτών. Για αυτό σκοπό χρησιμοποιούνται τα δύο locks στο struct pipe (lock, wr_lock). Τα δύο αυτά locks, βρίσκονται εντός της κοινής μνήμης και απαιτούν ειδική μεταχείριση. Ο μηχανισμός ivshmem μπορεί να υποστηρίξει τη χρήση ατομικών εντολών του gcc. Για το λόγο, αυτό δημιουργήσαμε δύο ακόμα συναρτήσεις για το locking, μία για την απόκτηση του κλειδώματος και μία για την απελευθέρωση του. Οι συναρτήσεις αυτές φαίνονται παρακάτω.

```

1 /*
2  * Spinlock for pipe
3  */
4 void pipe_lock(bus_size_t lock)
5 {
6     while(__sync_val_compare_and_swap((uint8_t
        *)sharme.data_b + lock, 0, 1) == 1)
7         /* do nothing */;

```

```

8         return;
9     }
10
11 /*
12  * Release the lock
13  */
14 void pipe_unlock(bus_size_t lock)
15 {
16     __sync_lock_release((uint8_t *)sharme.data_b +
17                          lock);
18     return;
19 }

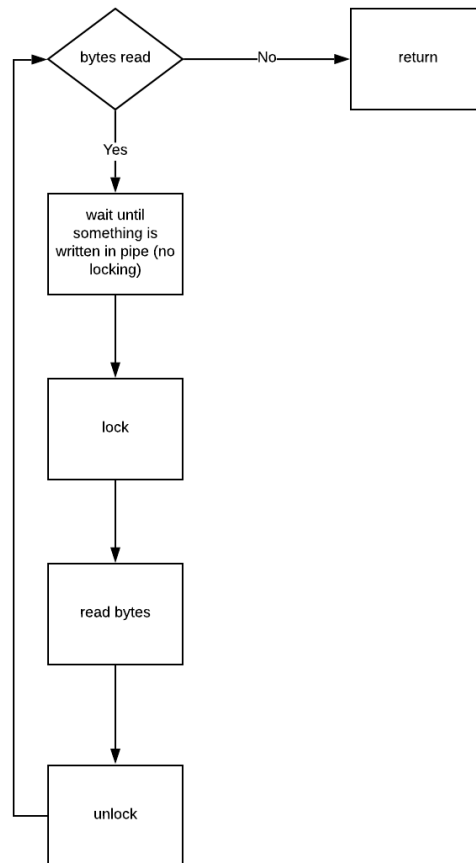
```

Οι ατομικές εντολές που χρησιμοποιήθηκαν για τη χρήση του lock, είναι οι `__sync_val_compare_and_swap` και `__sync_lock_release`. Οι εντολές αυτές δέχονται ως παράμετρο τη διεύθυνση της μνήμης στην οποία θα γίνει η πρόσβαση. Η διεύθυνση αυτή καθορίζεται από τη διεύθυνση βάσης των δεδομένων (`data_b`) στο struct `ivshmem` και το offset του lock μέσα στην κοινή μνήμη που είναι αποθηκευμένο στο struct `pipe`. Οποιαδήποτε αλλαγή σε κάποιο κομμάτι της κοινή μνήμης απαιτεί την απόκτηση του γενικού lock, ενώ οποιαδήποτε εγγραφή στο `pipe` θα πρέπει να έχει πάρει τον έλεγχο του `wr_lock`.

Όταν λοιπόν, η εφαρμογή που εκτελείται σε ένα `unikernel` εκτελέσει την κλήση `pipe`, από τη μεριά του πυρήνα πέρα από τη δημιουργία των `file descriptors` και τη σύνδεση τους με το εκάστοτε struct `pipe_op` συμβαίνουν τα εξής. Αρχικά, δημιουργείται το struct `pipe` και αρχικοποιούνται όλες οι μεταβλητές τους με το offset της κάθε μίας στην κοινή μνήμη. Έστερα, ξεκινά ο έλεγχος της κοινής μνήμης. Ελέγχεται η τιμή της `init` στην κοινή μνήμη και ανάλογα με την τιμή της η κοινή μνήμη, είτε θα αρχικοποιηθεί είτε όχι. Στη συνέχεια, αφού η περιοχή μνήμης έχει αρχικοποιηθεί, αυξάνονται οι μετρητές των άκρων του `pipe` και τέλος συνδέεται το κοινό struct `pipe` με τα struct `pipe_op`. Η κλήση επιστρέφει τους `file descriptors` στην εφαρμογή και το `pipe` είναι πλέον έτοιμο να χρησιμοποιηθεί. Ειδική σημείωση πρέπει να γίνει για τον έλεγχο της τιμής `init` στην κοινή μνήμη. Η κοινή μνήμη πρέπει να καταστραφεί, αν δεν υπάρχουν `pipes` που τη χρησιμοποιούν καθώς σε διαφορετική περίπτωση μπορεί λόγω του τρόπου ελέγχου της `init`, να μην αρχικοποιηθεί.

Σημαντικό ρόλο στον παραπάνω σκοπό έχει η κλήση `close` σε ένα `file descriptor` του `pipe`. Στην απλή περίπτωση που δεν πρόκειται για το τελευταίο `file descriptor` του `pipe`, απλά ενημερώνονται οι μετρητές των ανοιχτών άκρων του `pipe`. Αν όμως, πρόκειται να κλείσει το τελευταίο `file descriptor` που αφορά το συγκεκριμένο `pipe` τότε μηδενίζεται όλη η κοινή μνήμη που χρησιμοποιήθηκε. Με αυτό τον τρόπο

την επόμενη φορά που θα χρησιμοποιηθεί το συγκεκριμένο κομμάτι μνήμης η τιμή `init` θα υποδηλώνει ότι πρέπει η κοινή μνήμη να αρχικοποιηθεί.



Σχήμα 4.6: pipe read flow chart

Στην παραπάνω εικόνα 4.6 φαίνεται το διάγραμμα ροής, όταν εκτελείται η κλήση `read` σε άκρο ανάγνωσης του `pipe`. Ουσιαστικά πρόκειται για ένα `while loop`, το οποίο τερματίζει υπό τέσσερις ορισμένες συνθήκες:

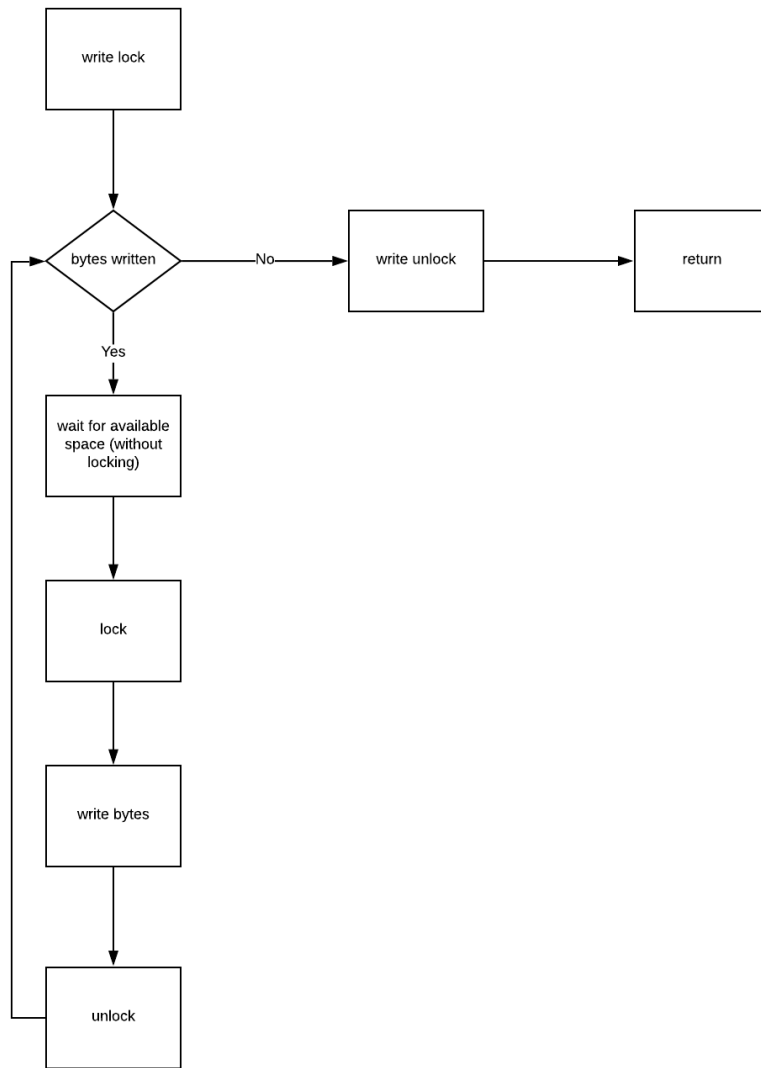
1. Αν έχουν διαβαστεί όλα bytes ζήτησε η εφαρμογή
2. Αν έχουν διαβαστεί όλα τα bytes από το `pipe` (ανεξάρτητα από τα πόσα ζήτησε η εφαρμογή)
3. Αν δεν υπάρχουν δεδομένα να διαβαστούν και όλα τα άκρα εγγραφής έχουν κλείσει
4. Αν κάτι πάει στραβά κατά τη διαδικασία μεταφοράς δεδομένων από την κοινή μνήμη στην εφαρμογή.

Για την αποφυγή deadlocks, αρχικά γίνεται έλεγχος για τη διαθεσιμότητα δεδομένων στο pipe χωρίς την απόκτηση του γενικού κλειδώματος. Σε περίπτωση, που υπάρχουν δεδομένα, τότε αποκτάται το κλείδωμα και επαναλαμβάνεται ο έλεγχος με την κατοχή του κλειδώματος. Αφού λοιπόν μεταφερθούν τα δεδομένα από την κοινή μνήμη στην εφαρμογή, τότε ενημερώνονται οι κοινές μεταβλητές του pipe (out, cnt) και ελευθερώνεται το lock. Καθώς ο buffer του pipe είναι κυκλικός, πρέπει να λαμβάνουμε υπόψιν την περίπτωση που έχουμε φτάσει στο τέλος του buffer, οπότε πρέπει να ξεκινήσουμε να διαβάζουμε τα υπόλοιπα δεδομένα από την αρχή. Τέλος, τόσο κατά τη διάρκεια του ελέγχου δεδομένων χωρίς αλλά και με το κλείδωμα, ελέγχονται και τα ανοιχτά άκρα εγγραφής στο pipe. Αν όλα έχουν κλείσει τότε πρέπει να επιστρέψει η κλήση pipe καθώς δεν πρόκειται να εγγραφούν νέα δεδομένα.

Στην παραπάνω εικόνα 4.7 φαίνεται το διάγραμμα ροής της κλήσης write του pipe. Όπως φαίνεται δεν έχει κάποια ιδιαίτερη διαφορά με το διάγραμμα ροής της κλήσης read, εκτός από το γεγονός ότι από την αρχή μέχρι το τέλος της εκτέλεσης της δεσμεύεται το wr_lock. Ο λόγος ύπαρξης ενός τέτοιου κλειδώματος, είναι για να επιτευχθεί η ατομικότητα της κλήσης write σε ένα pipe. Από εκεί και πέρα, όπως και πριν αρχικά ελέγχουμε αν υπάρχει διαθέσιμος χώρος στην κοινή μνήμη για την εισαγωγή νέων δεδομένων. Έπειτα αποκτάται το κλείδωμα και επαναλαμβάνεται ο έλεγχος διαθέσιμου χώρου. Αν είναι επιτυχής, γίνεται η εγγραφή δεδομένων στην κοινή μνήμη και ενημερώνονται οι αντίστοιχες μεταβλητές (in, cnt). Καθώς obuffer του pipe είναι κυκλικός πρέπει να λαμβάνουμε υπόψιν την περίπτωση που ο διαθέσιμος χώρος βρίσκεται την αρχή του buffer. Τόσο κατά τη διαδικασία ελέγχου διαθέσιμου χώρου χωρίς κλείδωμα όσο και με κλείδωμα, ελέγχονται και τα ανοιχτά άκρα ανάγνωσης του pipe. Αν όλα έχουν κλείσει τότε πρέπει να επιστρέψει η κλήση pipe με το error EPIPE, καθώς δεν υπάρχουν αναγνώστες να διαβάσουν τα δεδομένα. Τέλος, όλα τα παραπάνω βρίσκονται μέσα σε ένα while loop το οποίο μπορεί να τερματίσει υπό τις τέσσερις παρακάτω συνθήκες.

1. Αν έχουν γραφτεί όλα bytes ζήτησε η εφαρμογή
2. Αν δεν υπάρχει κανένα ανοιχτό άκρο ανάγνωσης στο pipe.
3. Αν κάτι πάει στραβά κατά τη διαδικασία μεταφοράς δεδομένων από την την εφαρμογή στην κοινή μνήμη.

4.1.2 Στάδια υλοποίησης



Σχήμα 4.7: pipe write flow chart

Κεφάλαιο 5

Επίλογος

μπλα μπλα μπλα

5.1 Mpla

dsfdsgfdsgdfs

5.2 Mpla

dsfdsgfdsgdfs

5.2.1 Mpla

dsfdsgfdsgdfs

mpla [9]

Βιβλιογραφία

- [1] Netbsd kernel developer's manual driver(9). <http://netbsd.gw.com/cgi-bin/man-cgi?driver+9+NetBSD-current>, Last accessed on 08-01-2019.
- [2] Solo5 on github. <https://github.com/Solo5/solo5>, Last accessed on 08-01-2019.
- [3] Solo5 technical overview. <https://github.com/Solo5/solo5/blob/master/docs/architecture.md>, Last accessed on 08-01-2019.
- [4] Keith Adams and Ole Agesen. A comparison of software and hardware techniques for x86 virtualization. *ACM SIGARCH Computer Architecture News*, 34(5):2-13, 2006.
- [5] Ole Agesen, Jim Mattson, Radu Rugina, and Jeffrey Sheldon. Software techniques for avoiding hardware virtualization exits. In *USENIX Annual Technical Conference*, pages 373-385, 2012.
- [6] Fabrice Bellard. Qemu, a fast and portable dynamic translator. In *USENIX Annual Technical Conference, FREENIX Track*, volume 41, page 46, 2005.
- [7] Alfred Bratterud, Alf-Andre Walla, Hårek Haugerud, Paal E Engelstad, and Kyrre Begnum. Includeos: A minimal, resource efficient unikernel for cloud services. In *Cloud Computing Technology and Science (CloudCom), 2015 IEEE 7th International Conference on*, pages 250-257. IEEE, 2015.
- [8] Michael Eder. Hypervisor-vs. container-based virtualization. *Future Internet (FI) and Innovative Internet Technologies and Mobile Communications (IITM)*, 1, 2016.
- [9] Xing Gao, Zhongshu Gu, Mehmet Kayaalp, Dimitrios Pendarakis, and Haining Wang. Containerleaks: emerging security threats of information leakages in container clouds. In *Dependable Systems and Networks (DSN), 2017 47th Annual IEEE/IFIP International Conference on*, pages 237-248. IEEE, 2017.

- [10] Major Hayden and Richard Carbone. Securing linux containers. *GIAC (GCUX) Gold Certification, Creative Commons Attribution-ShareAlike 4.0 International License*, 19, 2015.
- [11] Van Jacobson and Bob Felderman. Speeding up networking. In *Linux Conference Au*, <http://www.lemis.com/grog/Documentation/vj/lca06vj.pdf>, 2006.
- [12] Antti Kantee. *The Design and Implementation of the Anykernel and Rump Kernels*. PhD thesis, doctoral dissertation. Department of Computer Science and Engineering, Aalto University, 2012.
- [13] Antti Kantee. On rump kernels and the rumprun unikernel, 2015. <https://blog.xenproject.org/2015/08/06/on-rump-kernels-and-the-rumprun-unikernel/>, Last accessed on 08-01-2019.
- [14] Antti Kantee and Justin Cormack. Rump kernels no os? no problem! *USENIX; login: magazine*, 2014.
- [15] Avi Kivity, Dor Laor Glauber Costa, and Pekka Enberg. Os v—optimizing the operating system for virtual machines. In *Proceedings of USENIX ATC’14: 2014 USENIX Annual Technical Conference*, page 61, 2014.
- [16] Eddie Kohler, Robert Morris, Benjie Chen, John Jannotti, and M Frans Kaashoek. The click modular router. *ACM Transactions on Computer Systems (TOCS)*, 18(3):263–297, 2000.
- [17] A Cameron Macdonell et al. *Shared-memory optimizations for virtual machines*. University of Alberta Edmonton, Canada, 2011.
- [18] Anil Madhavapeddy, Richard Mortier, Charalampos Rotsos, David Scott, Balraj Singh, Thomas Gazagnaire, Steven Smith, Steven Hand, and Jon Crowcroft. Unikernels: Library operating systems for the cloud. In *Acm Sigplan Notices*, volume 48, pages 461–472. ACM, 2013.
- [19] Joao Martins, Mohamed Ahmed, Costin Raiciu, Vladimir Olteanu, Michio Honda, Roberto Bifulco, and Felipe Huici. Clickos and the art of network function virtualization. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, pages 459–473. USENIX Association, 2014.
- [20] Peter Mell, Tim Grance, et al. The nist definition of cloud computing. 2011.

- [21] Gerald J Popek and Robert P Goldberg. Formal requirements for virtualizable third generation architectures. *Communications of the ACM*, 17(7):412–421, 1974.
- [22] Donald E Porter, Silas Boyd-Wickizer, Jon Howell, Reuben Olinsky, and Galen C Hunt. Rethinking the library os from the top down. In *ACM SIGPLAN Notices*, volume 46, pages 291–304. ACM, 2011.
- [23] Octavian Purdila, Lucian Adrian Grijincu, and Nicolae Tapus. Lkl: The linux kernel library. In *Roedunet International Conference (RoEduNet), 2010 9th*, pages 328–333. IEEE, 2010.
- [24] Yi Ren, Ling Liu, Qi Zhang, Qingbo Wu, Jianbo Guan, Jinzhu Kong, Huadong Dai, and Lisong Shao. Shared-memory optimizations for inter-virtual-machine communication. *ACM Computing Surveys (CSUR)*, 48(4):49, 2016.
- [25] Volkmar Uhlig, Joshua LeVasseur, Espen Skoglund, and Uwe Dannowski. Towards scalable multiprocessor virtual machines. In *Virtual Machine Research and Technology Symposium*, pages 43–56, 2004.
- [26] Dan Williams and Ricardo Koller. Unikernel monitors: Extending minimalism outside of the box. In *HotCloud*, 2016.