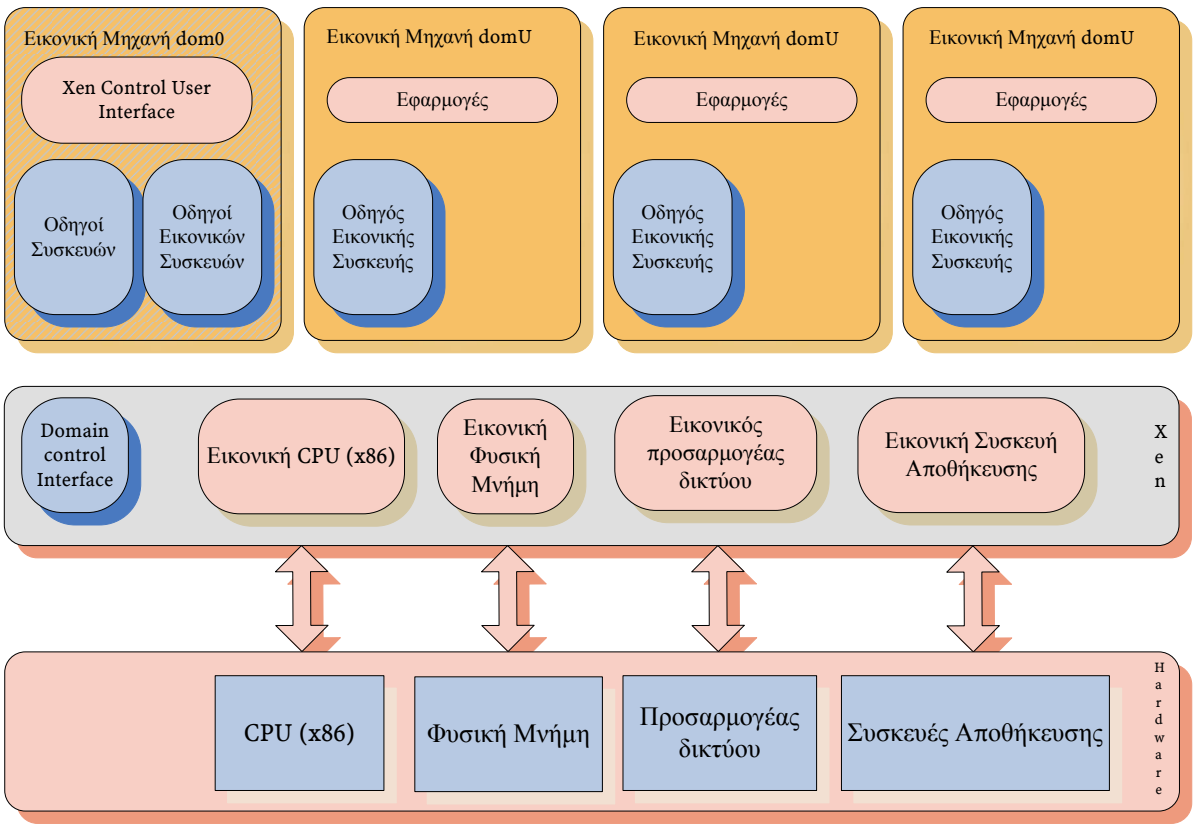
Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχ. και Μηχ. Υπολογιστών
Εργαστήριο Υπολογιστικών Συστημάτων
http://www.cslab.ece.ntua.gr

Υλοποίηση Στρώματος Μοιραζόμενης Πρόσβασης σε Συσκευές Δικτύου για Εικονικές Μηχανές

Εισαγωγή

Πολλά σημερινά συστήματα σχεδιάζονται με βάση τεχνικές virtualization με στόχο να επιμεριστεί η χρήση των πόρων που διαθέτουν. Η έννοια του virtualization αναφέρεται στην από κοινού χρήση συσκευών (CPU, Μνήμης, συσκευών I/O κλπ) από λειτουργικά συστήματα που εκτελούνται σε εικονικές μηχανές, με διαφανή τρόπο, χωρίς αυτά να αντιλαμβάνονται ότι μοιράζονται μία συσκευή.

Για να επιτευχθεί η σωστή (ασφαλής) και δίκαιη από κοινού χρήση των πόρων ενός συστήματος απαιτείται η ύπαρξη ενός συντονιστή, που θα καθορίζει ποια εικονική μηχανή μπορεί να αποκτήσει πρόσβαση σε κάποια συσκευή ή να διατηρεί πίνακες αντιστοίχισης εικονικής-φυσικής μνήμης των μηχανών κλπ. Το ρόλο του συντονιστή αναλαμβάνει να διεκπεραιώσει ο hypervisor ή Virtual Machine Monitor (VMM).



Σχήμα 1: Δομή της πλατφόρμας virtualization Xen

Συσκευές I/O

Η χρήση συσκευών I/O από λειτουργικά συστήματα που εκτελούνται σε εικονικές μηχανές επιβάλλει περιορισμούς που οφείλονται τόσο στο σχεδιασμό του hardware της συσκευής, όσο και στο λογισμικό που αναλαμβάνει να εξυπηρετήσει αιτήσεις για αποστολή ή λήψη δεδομένων από / προς τη συσκευή. Στην πλειοψηφία τους, οι συσκευές I/O διαθέτουν μηχανές DMA για να ανταλλάζουν δεδομένα με το Λειτουργικό Σύστημα. Ο hypervisor μιας πλατφόρμας virtualization θα πρέπει να διασφαλίζει ότι οι εικονικές μηχανές (Virtual Machines) δεν μπορούν να χρησιμοποιούν αυτές τις μηχανές DMA για να γράφουν ή να διαβάζουν δεδομένα χωρίς έλεγχο.

Πιο συγκεκριμένα, για μια διεπαφή δικτύου, το VMM πρέπει να διαχωρίζει την πρόσβαση στη συσκευή από τα επιμέρους VMs είτε για εξερχόμενη, είτε για εισερχόμενη κίνηση. Για παράδειγμα, μία κακόβουλη VM μπορεί να δώσει εντολή στη συσκευή για μεταφορά δεδομένων σε χώρο μνήμης μιας άλλης VM με αποτέλεσμα να αλλοιώσει τη μνήμη της τελευταίας. Η συσκευή θα πρέπει να είναι αρκετά έξυπνη για να αποτρέψει κάτι τέτοιο (πράγμα που απαιτεί εξειδικευμένο hardware). Συνήθως το ρόλο του ελεγκτή για την ορθότητα των μεταφορών δεδομένων τον τηρεί το Λειτουργικό Σύστημα. Σε περιβάλλοντα virtualization, αυτόν τον ρόλο τον παίζει το VMM (ο hypervisor).

Myrinet

Το Myrinet είναι ένα προηγμένο δίκτυο διασύνδεσης υψηλών επιδόσεων για την δημιουργία συστοιχιών υπολογιστών (compute clusters). Η σχεδίασή του προσφέρει ανταλλαγή μηνυμάτων με μεγάλο ρυθμό διαμεταγωγής (2-10Gb/s throughput) και πολύ μικρό χρόνο αρχικής απόκρισης (<5μs latency).

Το λογισμικό που χρησιμοποιείται σε προσαρμογείς Myrinet επιτρέπει αποστολή και λήψη μηνυμάτων από το επίπεδο χρήστη, μετακινώντας δεδομένα απευθείας από και προς απομονωτές της εφαρμογής, χωρίς να απαιτείται η εκτέλεση κλήσεων συστήματος για την πραγματοποίηση της επικοινωνίας.

Περιγραφή της προτεινόμενης διπλωματικής εργασίας

Μια από τις ερευνητικές δραστηριότητες του εργαστηρίου υπολογιστικών συστημάτων αφορά στη μελέτη τεχνικών virtualization και έξυπνων δικτύων διασύνδεσης καθώς και την αποδοτική χρήση αυτών.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η σχεδίαση ενός μηχανισμού πρόσβασης στη διεπαφή δικτύου από πολλές εικονικές μηχανές που βρίσκονται στο ίδιο υπολογιστικό σύστημα, με διαφανή τρόπο, ενώ παράλληλα διατηρείται η σημασιολογία πρόσβασης στη συσκευή καθώς και ο δίκαιος διαμοιρασμός των πόρων της συσκευής.

Για την αποδοτικότερη αξιοποίηση συσκευών δικτύου σε εικονικές μηχανές θα πρέπει να λάβουμε υπόψη τεχνικές μοιραζόμενης πρόσβασης σε συσκευές, που να προσανατολίζονται σε υψηλές επιδόσεις.

Μία από αυτές τις τεχνικές είναι το VMM-bypass I/O [3] που αξιοποιεί μια μορφή του user-level communication. Η μοιραζόμενη πρόσβαση στη συσκευή υλοποιείται σε δύο επιμέρους στάδια. Με έναν οδηγό συσκευής δικτύου για την εικονική μηχανή και έναν μηχανισμό ελέγχου που υλοποιείται στον hypervisor και διασφαλίζει την από κοινού πρόσβαση όλων των εικονικών μηχανών στη συσκευή με ασφάλεια.

Η προτεινόμενη διπλωματική εργασία αφορά στην υλοποίηση ενός στώματος διεπαφής ανάμεσα στην εικονική μηχανή και την πραγματική συσκευή έτσι, ώστε οι μεταφορές δεδομένων από και προς τη συσκευή να γίνονται χωρίς να παρεμβάλλεται ο hypervisor.

Στάδια Υλοποίησης

Η προτεινόμενη διπλωματική εργασία μπορεί ενδεικτικά να υλοποιηθεί στα εξής επιμέρους στάδια:

- Μελέτη και εξοικείωση με τα στρώματα συσκευών του Linux Kernel
  - επίπεδο συσκευών δικτύου
  - υποσύστημα μνήμης (σελιδοποίηση, διευθυνσιοδότηση, επικοινωνία με συσκευές Εισόδου / Εξόδου)
- Μελέτη του Xen Hypervisor για τον πυρήνα του Linux (μνήμη, εικονικές συσκευές)
- Εξοικείωση με τον οδηγό συσκευής και το firmware της διεπαφής δικτύου Myrinet 10G-PCIE-8A
- Υλοποίηση ενός εικονικού οδηγού συσκευής για τη διεπαφή δικτύου σε Εικονική Μηχανή
- Υλοποίηση μεθόδων ασφαλούς πρόσβασης της Εικονικής Μηχανής στο χώρο διευθύνσεων της διεπαφής δικτύου μέσα στο VMM (hypervisor)
- Σύνδεση του εικονικού οδηγού συσκευής με την συσκευή με παράκαμψη του hypervisor για τη μεταφορά δεδομένων
- Πειραματική αποτίμηση της παραπάνω υλοποίησης
- Συγγραφή της διπλωματικής

Γνώσεις που απαιτούνται

- βασικές αρχές σύγχρονης αρχιτεκτονικής υπολογιστών
- εμπειρία με το λειτουργικό σύστημα Linux και με το προγραμματιστικό περιβάλλον του
- *πολύ* καλή γνώση προγραμματισμού σε C

Γνώσεις που θα αποκτηθούν

- εξοικείωση με τον πυρήνα ενός σύγχρονου Λ.Σ. και τις σχεδιαστικές του αρχές.
- εμπειρία σε τεχνικές virtualization.
- εξοικείωση με το στρώμα συσκευών δικτύου του Linux.
- εξοικείωση με ένα σύγχρονο, έξυπνο δίκτυο διασύνδεσης.

Ενδιαφέροντες σύνδεσμοι

- *Myrinet-2000* [http://www.myri.com/myrinet]/
- *Myri10g* [http://www.myri.com/myri-10g/overview]
- Linux device drivers v3 (2.6.x πυρήνες) [http://lwn.net/kernel/ldd3/]

Επικοινωνία

Νεκτάριος Κοζύρης	Αναπληρωτής Καθηγητής	nkoziris@cslab.ece.ntua.gr
Ευάγγελος Κούκης	υπ. διδάκτορας	vkoukis@cslab.ece.ntua.gr
Αναστάσιος Νάνος	υπ. διδάκτορας	ananos@cslab.ece.ntua.gr

Αναφορές

[1] Paul Barham, Boris Dragovic, Keir Fraser, Steven Hand, Tim Harris, Alex Ho, Rolf Neugebauer, Ian Pratt, and Andrew Warfield. Xen and the art of virtualization. In *SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles*, pages 164–177, New York, NY, USA, 2003. ACM.

[2] David Chisnall. *The definitive guide to the xen hypervisor*. Prentice Hall Press, Upper Saddle River, NJ, USA, 2007.

[3] Jiuxing Liu, Wei Huang, Bulent Abali, and Dhabaleswar K. Panda. High performance VMM-bypass I/O in virtual machines. In *ATEC '06: Proceedings of the annual conference on USENIX '06 Annual Technical Conference*, pages 3–3, Berkeley, CA, USA, 2006. USENIX Association.

[4] Sreekrishnan Venkateswaran. *Essential Linux Device Drivers*. Prentice Hall, 2008.