ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Πούλιος Δημήτριος

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Καθηγητής, Μ. Κατεβαίνης

Παρασκευή, 3 Απριλίου 2015, 14:00

Αίθουσα Ε313, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Υλοποίηση Δικτυακών Sockets Χαμηλής Καθυστέρησης μέσω Απομακρυσμένου DMA "

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια, οι αλλαγές στην αγορά των servers έχουν φέρει στο προσκήνιο νέες υλοποιήσεις, όπως ο Microserver, οι οποίες στοχεύουν σε μειωμένη κατανάλωση ενέργειας και οικονομία χώρου. Τέτοιες υλοποιήσεις χρησιμοποιούν μεγάλο πλήθος όχι ιδιαίτερα ισχυρών υπολογιστικών κόμβων, ομαδοποιημένων ώστε να εξυπηρετούν κλιμακώσιμες εφαρμογές προορισμένες για Data Centers. Δυστυχώς όμως, πολλές φορές αυτή η κλιμάκωση περιορίζεται από την ποιότητα της εσωτερικής επικοινωνίας μεταξύ των κόμβων, όπου η χαμηλή παροχή (throughput) και, ακόμα χειρότερα, η μεγάλη καθυστέρηση (latency), μπορεί να οδηγήσει σε κακή απόδοση. Σε αυτήν τη δουλειά, εξερευνούμε την επίδραση που μπορεί να έχει σε ένα περιβάλλον Microserver, η ύπαρξη ενός εσωτερικού δικτύου το οποίο έχει τη δυνατότητα να εκτελεί μεταφορές δεδομένων με πράξεις απομακρυσμένου DMA (RDMA).

Κατά κύριο λόγο, οι εφαρμογές χρησιμοποιούν το Socket API για επικοινωνήσουν μεταξύ τους μέσω δικτύων. Συνεπώς, για να μπορέσουμε να εκμεταλλευτούμε το προαναφερθέν εσωτερικό δίκτυο χωρίς να χρειαστεί να τροποποιήσουμε τις υπάρχουσες εφαρμογές, οι κλήσεις συστήματος (system calls) σχετικές με τα Sockets πρέπει να αναχαιτιστούν (intercepted). Πραγματοποιούμε την αναχαίτιση αυτή στο επίπεδο του χρήστη (user space), χρησιμοποιώντας μια τροποποιημένη έκδοση της Standard C Library, με σκοπό να παρακάμψουμε την επιβάρυνση του πρωτοκόλλου TCP / IP. Επιπλέον, υλοποιήσαμε έναν driver στον πυρήνα, ο οποίος πραγματοποιεί ασφαλείς μεταφορές δεδομένων μέσω πράξεων RDMA, οι οποίες χρειάζονται φυσικές διευθύνσεις. Η απομακρυσμένη ειδοποίηση της ολοκλήρωσης τέτοιων μεταφορών γίνεται με τη βοήθεια ενός μηχανισμού Mailbox, ο οποίος χρησιμοποιείται επίσης για την επικοινωνία που χρειάζονται οι κόμβοι ώστε να δημιουργήσουν ή να τελειώσουν τοπικές συνδέσεις.

Συνδυάζοντας τα παραπάνω στοιχεία, είτε στο επίπεδο του χρήστη ή του πυρήνα, κατευθύνουμε τις εφαρμογές να χρησιμοποιούν το εσωτερικό δίκτυο για τοπικές συνδέσεις TCP. Η αξιολόγηση του συστήματός μας, σε σχέση με μια τυπική διάταξη ethernet, έδειξε βελτίωση από 3 μέχρι 5 φορές στο χρόνο.

Poulios Dimitrios

M.Sc. Thesis

Computer Science Department

University of Crete

Master's Thesis Supervisor: Professor M. Katevenis

Friday, 03/04/2015, 14:00

Room E313, Computer Science dept., University of Crete

"Low-Latency Implementation of Network Sockets over Remote DMA"

ABSTRACT

In recent years, changes in the server market have brought power and space efficient server designs, like the Microserver. Such designs utilize large numbers of lightweight compute nodes bundled together to serve scale-out data center workloads. Unfortunately, scalability can often be limited by the quality of internal communication among running nodes, where low throughput and, even more critically, high latency can lead to poor performance. In this work, we explore the

efficiency of a Remote Direct Memory Access (RDMA) capable internal network in a Microserver environment.

Applications commonly use the standard Socket API for interprocess communication across networks. Therefore, to take advantage of the aforementioned internal network without modifying existing applications, socket-related system calls have to be intercepted. We implement system call interception in user space, using a modified Standard C Library, in order to bypass the kernel TCP/IP stack. A kernel driver has also been developed to securely perform data transfers via RDMA operations, which require physical addresses. Remote completion notifications of RDMA operations are triggered by a custom hardware Mailbox Mechanism, which also handles communication among nodes, necessary to initiate and close local connections.

By combining these user and kernel space elements, we direct local TCP traffic through our internal network. Evaluation results show a 3x to 5x improvement to the latency, using our system compared to a typical ethernet configuration.