ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Μπάρτζης Σωκράτης

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Καθηγητής, Μ. Κατεβαίνης

Δρ. Ν. Χρυσός (επιβλέπων)

Παρασκευή, 27 Μαΐου 2022, ώρα 14:00 μ.μ.

Join Zoom Meeting

https://us02web.zoom.us/j/88012293674

"Προηγμένη υποστήριξη για την Βελτίωση της Ποιότητας Υπηρεσίας μιας μηχανής υλικού για Απομακρυσμένες Άμεσες Προσπελάσεις Μνήμης"

ΠΕΡΙΛΗΨΗ

Τις τελευταίες δεκαετίες, τόσο ο κλάδος της έρευνας, όσο και αυτός της βιομηχανίας έχουν στραφεί προς την Υπολογιστική Υψηλών Αποδόσεων για να καλύψουν τις αυξανόμενες ανάγκες τους για υπολογιστική ισχύ. Σε μία προσπάθεια να υλοποιήσουμε ένα πλαίσιο επικοινωνίας υψηλής απόδοσης για ευρωπαϊκούς υπερυπολογιστές, στα πλαίσια των ευρωπαϊκών προγραμμάτων ExaNeSt και RED-SEA, σχεδιάζουμε μια νέα διεπαφή δικτύου χαμηλής καθυστέρησης (λιγότερο από 0,5 μs) και υψηλής παροχής (100 Gb/s), ικανή για απομακρυσμένες άμεσες προσπελάσεις μνήμης.

Σε αυτήν την εργασία σχεδιάζουμε μια μηχανή υλικού για την βελτίωση της παροχής υπηρεσιών (Quality of Service, QoS) μιας μηχανής Απομακρυσμένων Άμεσων Προσπελάσεων Μνήμης (Remote Direct Memory Access, RDMA). Οι μεγάλες μεταφορές δεδομένων χωρίζονται σε μικρότερα τμήματα, έτσι ώστε να επιτραπεί η επιλεκτική αναμετάδοση δεδομένων, η χρήση πολλαπλών διαδρομών μέσα στο δίκτυο, καθώς και να αποφευχθεί ο επιπλέον φόρτος που προκύπτει από επιβεβαιώσεις λήψεων σε επίπεδο πακέτων. Οι μεταφορές μικρού μεγέθους μπορούν να παρακάμψουν την διαδρομή RDMA-DRAM, ελαχιστοποιώντας περαιτέρω τον χρόνο ολοκλήρωσής τους. Προγραμματίζουμε τις μεταφορές σε επίπεδο τμημάτων, βασιζόμενοι σε σειρά προτεραιότητας που καθορίζεται από τον χρήστη, και υποστηρίζουμε διαχείριση συμφόρησης του δικτύου. Επιπροσθέτως,

παρέχουμε 2048 εικονικά κανάλια στον χρήστη για την έκδοση πολλαπλών εκκρεμών αιτημάτων μεταφοράς δεδομένων, υλοποιούμε μια μηχανή ειδοποίησης ολοκλήρωσης σε υλικό και εισάγουμε έναν νέο τρόπο μαζικής, διαδοχικής διερεύνησης της κατάστασης πολλαπλών καναλιών.

Η υλοποίησή μας σε επίπεδο μεταφοράς καταχωρητών χρησιμοποιεί ομοχειρία για να επιτύχει υψηλή συχνότητα ρολογιού και υψηλό ρυθμό αποστολής μηνυμάτων (1 πράξη/κύκλο ρολογιού ή 150 MOP/s για υλοποίηση στην συστοιχία επιτόπια προγραμματιζόμενων πυλών (Field Programmable Gate Array, FPGA) που χρησιμοποιήσαμε), ενώ παράλληλα διατηρεί χαμηλούς χρόνους καθυστέρησης, 4 κύκλους ρολογιού για μεταφορές του ενός (1) τμήματος. Για να μειώσουμε περαιτέρω τον χρόνο καθυστέρησης, υλοποιήσαμε πολλαπλές ουρές (32) προγραμματισμού μεταφορών, σε κοινόχρηστο χώρο, οι οποίες υποστηρίζουν μια (1) πράξη εξαγωγής και μία (1) εισαγωγής κόμβου από/στις ουρές ανά κύκλο ρολογιού, καθώς και πράξεις εξαγωγής σε διαδοχικούς κύκλους ρολογιού.

Υλοποιήσαμε την εργασία στην FPGA του Zynq Ultrascale+ MPSoC της Xilinx. Για την μηχανή βελτίωσης Ποιότητας Υπηρεσίας χρησιμοποιήθηκαν 13,3K Προγραμματιζόμενες Πύλες (LUTs), 5,1K καταχωρητές και 23 μνήμες τυχαίας προσπέλασης (848 kbits). Η μέγιστη συχνότητα που επετεύχθη ήταν 150 MHz, μπορεί, ωστόσο, να βελτιωθεί περαιτέρω, ιδιαίτερα σε μία υλοποίηση πολύ μεγάλης κλίμακας ολοκλήρωσης (Very Large Scale Integration, VLSI).

Εκτενείς δοκιμές για την επαλήθευση της λειτουργικότητας της μηχανής πραγματοποιήθηκαν χρησιμοποιώντας το Vivado Design Suite. Η μηχανή QoS που αναπτύχθηκε σε αυτή την διατριβή ολοκλήρωσε σε προσομοίωση 100K μεταφορές δεδομένων, μεταβλητού μεγέθους, έως 1 MB. Επιπρόσθετα, ενσωματώσαμε την μηχανή QoS με την μονάδα αποστολής σε έναν προσομοιωμένο πάγκο δοκιμών, εκδίδοντας 5K εκκρεμείς μεταφορές, μεγίστου μεγέθους 256 KB (256 πακέτων), οι οποίες ολοκληρώθηκαν και αυτές με επιτυχία. Σε αυτές τις δοκιμές εξετάσαμε κάθε είδους μεταφορά, συμπεριλαμβανομένων των ροών υπό διαχείριση συμφορήσεως και των ροών γρήγορης διαδρομής, και επαληθεύσαμε τον μηχανισμό ειδοποίησης ολοκλήρωσης.

Η μηχανή RDMA υλοποιήθηκε στην FPGA του Zynq και ελήφθησαν μετρήσεις απόδοσης από προγράμματα σε επίπεδο χρήστη, εκτελεσμένα στον επεξεργαστή ARM A53 του Zynq. Ο χρόνος ολοκλήρωσης για μικρές μεταφορές έως 512 Byte ανέρχεται στα 360 ns, κατά τη μεταφορά εντός κόμβου, από BRAM σε BRAM (εξαιρουμένων των καθυστερήσεων δικτύου και DRAM), δέκα φορές χαμηλότερο από τον αντίστοιχο χρόνο της μηχανής ExaNeSt RDMA, μιας προηγούμενης υλοποίησης λογισμικού-υλισμικού στο ίδιο MPSoC, χρησιμοποιώντας τον συνεπεξεργαστής ARM Cortex-R5 για να υποστηρίξει QoS. Επιπλέον, βελτιώσαμε δραματικά τον ρυθμό μεταφοράς δεδομένων, επιτυγχάνοντας την μέγιστη θεωρητική παροχή με μεταφορές των 16 KB, ενώ στην προηγούμενη υλοποίηση απαιτούνταν μεταφορές των 4 MB. Τέλος, παρ΄ότι η μηχανή RDMA έχει δοκιμαστεί και βελτιστοποιηθεί για διασυνδέσεις κεντρικού επεξεργαστή τύπου AXI, μπορεί επίσης να συνδεθεί και με διασυνδέσεις τύπου PCI και CHI.

University of Crete

Computer Science Department

M.Sc. Thesis

Mpartzis Sokratis

Master's Thesis Supervisor: Professor M. Katevenis
Dr. N. Chrysos (Thesis Co-Advisor)

Friday, 27 May 2022, 14:00 p.m.

Join Zoom Meeting

https://us02web.zoom.us/j/88012293674

"Advanced Quality of Service support for hardware RDMA engine"

ABSTRACT

In recent decades, both research and industry have turned to High Performance Computing (HPC) for their ever-increasing computational needs. In an attempt to provide a high-performance communication framework for European supercomputers, under the EU-funded ExaNeSt and RED-SEA projects, we design a novel Remote Direct Memory Access (RDMA) engine, capable of low latency (less than 0.5 μs) and high throughput communication (100 Gb/s).

In this thesis, we design the Quality of Service (QoS) hardware of our RDMA engine. Transfers are segmented into blocks, so as to enable selective retransmissions, multi-path routing and to avoid per packet acknowledgment overheads. Small-sized transfers can bypass the RDMA-DRAM path, to further minimize latency. We schedule transfers at block level, based on a user-defined priority, we support end-to-end flow control and we enable network multi-pathing and congestion management options. We also implement a completion notification engine in hardware. We expose 2048 virtual channels to users supporting multiple outstanding data transfer requests. Finally, we introduce a novel way of collectively polling the status of multiple channels.

Our register-transfer-level (RTL) hardware implementation is pipelined in order to achieve higher clock and message rates (1 operation/clock cycle, or 150 MOP/s in our FPGA implementation), while maintaining a low latency of 4 clock cycles for single block transfers. To further reduce latency, we implement multiple (32) scheduling queues in shared space, that support one (1) enqueue and one (1) dequeue operation per clock cycle, as well as back-to-back dequeue operations.

We synthesized our design for the Zynq Ultrascale+ MPSoC. The RDMA's QoS engine leverages 13.3K Look-Up Tables (LUTs), 5.1K register and 23 BRAM blocks (848

kbits). The maximum frequency achieved in this FPGA was 150 MHz, but this can be further improved, especially in a VLSI implementation.

Extensive functional verification tests were performed using the Vivado Design Suite. The QoS engine developed in this thesis completed in simulation 100K outstanding transfers of varying size, up to 1 MB. Additionally, we integrated our QoS implementation with the RDMA send unit in another simulated test-bench, issuing 5K transfers of maximum 256 KB (256 packets), which the design also completed successfully. In these tests, we examined every possible transfer type, including congestion managed and fast-path flows, as well as completion notifications.

The design was implemented on the Zynq's FPGA and performance measurements were taken from user-level programs on the Zynq's A53 ARM core. Completion time for small transfers of up to 512 Bytes was measured at 360 ns, when transferring intra-node, BRAM to BRAM (excluding network and DRAM latencies), ten times lower than the latency of the ExaNeSt RDMA, a previous implementation on the same MPSoC, using the ARM Cortex-R5 co-processor for QoS support. Moreover, we significantly improved the transfer rate that can be achieved, reaching the theoretical maximum (line) throughput as early as with 16KB transfers, whereas using the previous implementation the corresponding transfer size was 4MB. Finally, although the RDMA engine is optimized for and tested using AXI processor interconnects, it can also be connected to PCI or CHI host-processor interconnects.