ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Νικολουδάκης Μιχαήλ

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Αναπλ. Καθηγητής, Π. Πρατικάκης

Δρ. Μ. Πλουμίδης (επιβλέπων)

Δευτέρα, 11 Απριλίου 2022, ώρα 16:00 μ.μ.

Join Zoom Meeting

https://zoom.us/j/93276154642

" Σχεδιασμός και υλοποίηση μιας έκδοσης του Exanest MPI βασισμένης σε εγγραφή"

ΠΕΡΙΛΗΨΗ

Το MPI είναι ένα από τα κυρίαρχα πρωτόκολλα επικοινωνίας που χρησιμοποιούνται στις πλατφόρμες HPC (Υπολογισμού Υψηλής απόδοσης) σήμερα λόγω της φορητότητας και της κλιμακοσιμότητας του. Πολλές HPC εφαρμογές κάνουν χρήση του MPI για να καταστήσουν εφικτή την επικοινωνία μεταξύ διαφορετικών διεργασιών. Στα πλαίσια του έργου ExaNeST, ένα νέο HPC πρωτότυπο αναπτύχθηκε στο εργαστήριο CARV του I.T.E αποτελούμενο από συνολικά 512 ARMv8 πυρήνες συνδυασμένους με λογική FPGA. Το πρωτότυπο αυτό κάνει χρήση ειδικών μονάδων επικοινωνίας σχεδιασμένων να επιτρέπουν τη διάδοση μηνυμάτων συγχρονισμού με πολύ μικρή καθυστέρηση (latency) καθώς και την αποδοτική μεταφορά μεγάλων δεδομένων μέσω του δικτύου Exanet.

Προκειμένου να αξιοποιηθούν οι προαναφερθείσες ικανότητες του πρωτοτύπου, αναπτύχθηκε μια υψηλά βελτιστοποιημένη MPI υλοποίηση (Exanet MPI) στα πλαίσια του ίδιου έργου, πριν από την αρχή της δουλειάς μας. Η υλοποίηση αυτή κάνει χρήση των μονάδων επικοινωνίας του HPC πρωτοτύπου και καταφέρνει να επιτύχει καλύτερη απόδοση από την γνωστή MPI υλοποίηση MPICH επιτυγχάνοντας μέχρι και 30x μικρότερη καθυστέρηση διάδοσης δεδομένων (latency). Το Exanet MPI υποστηρίζει ένα eager και ένα long πρωτόκολλο επικοινωνίας για μικρές και μεγάλες

μεταφορές δεδομένων αντίστοιχα. Το πρωτόκολλο long βασίζεται σε προσομοιωμένα DMA reads (read based) και υποστηρίζει εκκίνηση της επικοινωνίας αποκλειστικά από τον αποστολέα. Η “εκκίνηση από τον αποστολέα” ορίζεται ως η ικανότητα του αποστολέα ενός MPI μηνύματος να ξεκινήσει την επικοινωνία με τον παραλήπτη εκδίδοντας το κατάλληλο μήνυμα συγχρονισμού. Παρά την απλότητα της, η εκκίνηση αποκλειστικά από τον αποστολέα δεν μας επιτρέπει να εκμεταλλευτούμε περιπτώσεις όπου ο παραλήπτης καταχωρεί την αίτηση του νωρίτερα από τον αποστολέα. Επιπρόσθετα, η χρήση προσομοιωμένων DMA reads απαιτεί από τον παραλήπτη να ενημερώσει τον αποστολέα σχετικά με το τέλος μιας DMA μεταφοράς μέσω ενός Ack μηνύματος συγχρονισμού, κάτι που επιφέρει επιπλέον καθυστέρηση (latency).

Σε αυτή την εργασία, σχεδιάζουμε και υλοποιούμε εξ αρχής μια έκδοση του Exanet MPI βασισμένη σε εγγραφή (write based) η οποία υποστηρίζει εκκίνηση επικοινωνίας από τον αποστολέα αλλά και από τον παραλήπτη. Με τη χρήση DMA writes, καθιστούμε τον αποστολέα ικανό να αντιληφθεί το τέλος μιας DMA μεταφοράς δίχως να χρειάζεται επιβεβαίωση από τον παραλήπτη. Επιπρόσθετα, εκμεταλλευόμαστε περιπτώσεις όπου η αίτηση παραλαβής καταχωρείται νωρίτερα από την αίτηση αποστολής επιτρέποντας στον παραλήπτη να μεταφέρει ασύγχρονα στον αποστολέα όλες τις πληροφορίες που χρειάζονται για μια DMA μεταφορά. Συνεπώς, ο αποστολέας που θα καταχωρήσει την αίτηση αποστολής μετά τον παραλήπτη, μπορεί να μεταφέρει άμεσα τα δεδομένα του χωρίς να υπάρχει ανάγκη για περαιτέρω συγχρονισμό με τον παραλήπτη. Παρ’ όλα αυτά, η απλή προσθήκη της δυνατότητας εκκίνησης από τον παραλήπτη στο long πρωτόκολλο επιφέρει κάποιες επιπλοκές συμπεριλαμβανομένης και της μεγάλης αύξησης της καθυστέρησης στο eager πρωτόκολλο. Προτείνουμε δικές μας μεθόδους για την επιτυχή αντιμετώπιση των επιπλοκών που προκύπτουν από την υποστήριξη εκκίνησης από τον παραλήπτη καθώς επίσης βελτιστοποιούμε περαιτέρω το long πρωτόκολλο εξουδετερώνοντας την ανάγκη χρήσης κάποιων μηνυμάτων συγχρονισμού. Προκειμένου να αναλύσουμε το όφελος απόδοσης των βελτιστοποιήσεων μας, αναπτύσσουμε συνολικά τέσσερις παραλλαγές του Exanet MPI βασισμένου σε εγγραφή. Σε κάθε παραλλαγή, παρέχουμε υλοποιήσεις για τις περισσότερες εκ των point-to-point, collective, και communicator manipulating συναρτήσεων. Περιγράφουμε τις περιπτώσεις χρήσης της κάθε παραλλαγής και αξιολογούμε τις παραλλαγές της υλοποίησής μας έναντι της αρχικής, ήδη βελτιστοποιημένης, read based έκδοσης του Exanet MPI στο HPC πρωτότυπο. Εμβαθύνουμε στους τρόπους με τους οποίους οι βελτιστοποιήσεις μας στο μονοπάτι συγχρονισμού βελτιώνουν την απόδοση καθώς και στους παράγοντες που επιτρέπουν στην υλοποίησή μας να δείξει περισσότερο όφελος. Για την αξιολόγηση χρησιμοποιούμε τόσο microbenchmarks όσο και πραγματικές επιστημονικές εφαρμογές. Δείχνουμε ότι η υλοποίησή μας μπορεί να ξεπεράσει σε απόδοση το read based πρωτόκολλο προσφέροντας έως και 50% μικρότερη καθυστέρηση (latency) ενώ μπορεί να μειώσει το συνολικό χρόνο εκτέλεσης ορισμένων εφαρμογών έως και κατά 10%. (ανάλογα με το ποσοστό χρόνου επικοινωνίας που περιέχουν).

University of Crete

Computer Science Department

M.Sc. Thesis

Nikoloudakis Mixail

Master's Thesis Supervisor: Associate Professor, P. Pratikakis
Dr. M. Ploumidis (Thesis Co-Advisor)

Monday, 11 April 2022, 16:00 p.m.

Join Zoom Meeting

https://zoom.us/j/93276154642

"Design and Implementation of a Write Based version of the Exanet MPI"

ABSTRACT

MPI is one of the leading communication protocols used in HPC (High Performance Computing) suites today due to its portability and scalability. Many HPC applications make use of MPI in order to enable communication between different processes. In the scope of the ExaNeST project, an HPC prototype was deployed in the CARV Laboratory of FORTH consisting of 512 ARMv8 cores coupled with FPGA logic. This prototype makes use of special network primitives designed to allow the low latency transmission of control messages as well as the efficient transfer of large data through the Exanet network.

In order to exploit the aforementioned capabilities of the prototype, a highly optimized MPI implementation (Exanet MPI) was developed in the scope of the same project prior to our work. This implementation makes use of the prototype's communication primitives and manages to outperform the well known MPI implementation, MPICH by achieving up to 30x lower latency. Exanet MPI supports both an eager and a long communication protocol used for short and large MPI transfers respectively. The long protocol depends on emulated DMA reads and supports exclusively sender initiation.

Sender initiation is defined as the ability of the sender of an MPI message to initiate the communication with the receiver by issuing an appropriate control message. Despite its simplicity, sender initiation does not let us exploit scenarios in which the receiver posts its request earlier than the sender. In addition, the use of emulated reads requires the receiver to notify the sender about the end of a DMA transfer through the use of an Ack control message which incurs extra latency.

In this thesis, we design and implement from scratch a write-based version of the Exanet MPI that supports both sender and receiver initiation. With the use of DMA writes, we render the sender able to determine the end of a DMA transfer by itself without the need of acknowledgment from the receiver. Additionally, we take advantage of cases where a receive request gets posted earlier than a matching send request by letting the receiver initiate communication by asynchronously transferring its DMA related information to the sender. Consequently, a sender that posts its send request after the receiver, can immediately transfer data without the need of further synchronization with the receiver. However, simply adding receiver initiation support to the long protocol also infers some complications including (but not limited to) the significant increase of the eager protocol's latency. We propose our method for successfully facing the complications that arise from the support of receiver initiation and we also further optimize the long protocol by eliminating the need of some control messages. In order to break down the performance gain caused by our optimizations we develop in total 4 variants of the write based Exanet MPI. In each variant, we provide implementations for most point-to-point, collective as well as communicator manipulating functions. We describe the use cases of each developed variant and evaluate them against the already optimized read based original version of Exanet MPI on the HPC prototype. We offer insight into the ways our control path optimizations improve performance and the factors that let our implementation show more benefit. For the evaluation we use both microbenchmarks and real scientific applications. We show that our implementation can outperform the read based protocol by up to 50% in communication latency while also reduce the total execution time of specific applications by up to 10% (depending on the percentage of communication time they contain).