ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Κατεβαίνης – Μπίτζος Γεώργιος
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Καθηγητής, Α. Μπίλας

Μ. Μαραζάκης, (επιβλέπων)

Παρασκευή 3 Σεπτεμβρίου 2021, ώρα 15:00 μ.μ.

Join Zoom Meeting

https://zoom.us/j/97885263312

" Ιεραρχικά και Κοινού Εύρους Διευθύνσεων MPI Collectives, για Multi/Many-Core Επεξεργαστές"

Περίληψη

Ένας από τους βασικούς στόχους του MPI και των υλοποιήσεων του, είναι το performance portability. Με την ταχεία διάδοση των muli/many-core επεξεργαστών, εμφανίζονται εμπόδια που περιορίζουν την κλιμάκωση απόδοσης, και θέτουν σε κίνδυνο αυτό τον στόχο. Για να εγγυηθούμε τις προσδοκίες υψηλής απόδοσης και σε αυτούς τους επεξεργαστές, υλοποιούμε εντός του Open MPI, το XHC (XPMEM-based Hierarchical Collectives), ένα component για intra-node collectives, που λαμβάνει υπόψει του τις ιδιαιτερότητες των μοντέρνων επεξεργαστών.

Το XHC κατασκευάζει μια πολυεπίπεδη ιεραρχία που ανταποκρίνεται στα τοπολογικά χαρακτηριστικά του εκάστοτε επεξεργαστή, και η οποία καθορίζει τη ροή δεδομένων των αλγόριθμων του. Τα δεδομένα αποστέλλονται μέσω μνήμης, χωρίς περιττές αντιγραφές, μέσω user-level shared address space απεικονίσεων, κατασκευασμένων με χρήση του XPMEM. Ο συγχρονισμός επίσης επιτυγχάνεται μέσω της μνήμης, με την αξιοποίηση υψηλά αποδοτικών lock-free μεθόδων.

Παρέχουμε υλοποιήσεις για τα Broadcast, Barrier, και Allreduce, και τις αξιολογούμε σε τρία διαφορετικά υπολογιστικά συστήματα με μεγάλο αριθμό πυρήνων και με διάφορες τοπολογίες, συμπεριλαμβανομένου πολλαπλών πακέτων επεξεργαστών και πολλαπλών κόμβων NUMA. Χρησιμοποιούμε τα MPI microbenchmarks του OSU, αλλά και πραγματικές εφαρμογές HPC, για να παρέχουμε συγκρίσεις με τις υπόλοιπες υλοποιήσεις collectives εντός του Open MPI. Παρουσιάζουμε σημαντικά ευρήματα σχετικά με τη λειτουργία των microbenchmark, και τους κεντρικούς παράγοντες που επηρεάζουν τα αποτελέσματα τους. Τέλος, παραθέτουμε στατιστικά σχετικά με το ποσοστό χρήσης της κάθε λειτουργίας colletive στις HPC εφαρμογές μας.

Η προτεινόμενη μας υλοποίηση, επιτυγχάνει σε microbenchmarks, σε σχέση με την αμέσως καλύτερη, speedups 3x, 2x, και 8x για τα Broadcast, Barrier, και Allreduce αντίστοιχα. Στο miniAMR κατορθώνουμε speedup 4.8x, και στο CTNK μειώνουμε τον χρόνο απαιτούμενο για την εκμάθηση του νευρωνικού δικτύου AlexNet κατά 8%.

Τα αποτελέσματα της πειραματικής μας αποτίμησης επισημάνουν την ανάγκη για συγκεκριμένες τεχνικές, όπως την επίγνωση της τοπολογίας του επεξεργαστή, και τις μεταφορές δεδομένων με τεχνικές single-copy, οποτεδήποτε αναζητείται μέγιστη απόδοση για MPI collectives σε multi/many-core επεξεργαστές.

University of Crete

Computer Science Department

M.Sc. Thesis

Katevenis – Bitzos Georgios

Master's Thesis Supervisor: Professor, A. Bilas

M. Marazakis (Thesis Co- Advisor)

Friday, 3 September 2021, 15:00 p.m.

Join Zoom Meeting

https://zoom.us/j/97885263312

" Hierarchical Shared Address Space MPI Collectives, for Multi/Many-Core Processors"

Abstract

Performance portability is at the forefront of MPI's feature-set. With the ever-growing adoption of multi/many-core CPUs, obstacles arise that threaten to throttle performance and restrict scaling, jeopardizing this objective. In order to extend the expectations of high performance on these platforms, we implement in Open MPI XPMEM-based Hierarchical Collectives (XHC), a component for intra-node collectives that addresses the challenges inherent to modern processors.

XHC constructs a multi-level hierarchy that conforms to the processor's topological features and dictates the algorithms' data movement patterns. Data is exchanged over memory, without any redundant copies, via XPMEM-created user-level shared address space mappings. Synchronization is also realized through memory, with utilization of highly efficient lock-free techniques.

We provide implementations for the Broadcast, Barrier, and Allreduce collective operations, and evaluate them on three different machines with high core counts and varying topologies, including multiple CPU sockets and/or NUMA nodes. OSU's MPI microbenchmarks, as well as real-world HPC applications are used to produce comparisons with Open MPI's built-in components. We offer notable insight into the workings of microbenchmarks and core factors that affect their results. Finally, we present statistics on the usage of collective primitives in our HPC applications of choice.

Our proposed collectives achieve in microbenchmarks, over the next best option, speedups of up to 3x, 2x, and 8x for the Broadcast, Barrier, and Allreduce operations respectively. In miniAMR we attain up to 4.8x speedup, and in CNTK we reduce the training time of the AlexNet deep neural network by 8%.

The results of our experimental evaluation clearly highlight the necessity for certain techniques, like topology awareness and single-copy data transfers, when seeking maximum efficiency from MPI collectives on multi/many-core processors.