Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Τεχνικές Εισόδου/Εξόδου και Χρονοδρομολόγησης για την Αποδοτική Χρήση Μοιραζόμενων Αρχιτεκτονικών Πόρων σε Συστοιχίες Κόμβων Συμμετρικής Πολυεπεξεργασίας

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ευάγγελος Λ. Κούκης

Αθήνα, Ιανουάριος 2010

Περίληψη

Οι συστοιχίες (clusters) έχουν επικρατήσει ως οικονομική λύση για την κατασκευή κλιμακούμενων παράλληλων αρχιτεκτονικών, παρέχοντας υπολογιστική ισχύ σε ποικίλες εφαρμογές. Συστήματα Συμμετρικής Πολυεπεξεργασίας (SMPs) από πολυπύρηνους επεξεργαστές χρησιμοποιούνται συχνά ως δομικοί λίθοι στην κατασκευή συστοιχιών, σε συνδυασμό με δίκτυα διασύνδεσης υψηλής επίδοσης, όπως το Myrinet. Τα συστήματα SMP χαρακτηρίζονται από διαμοιρασμό πόρων σε πολλά επίπεδα· στους μοιραζόμενους πόρους περιλαμβάνονται ο χρόνος CPU, επίπεδα της ιεραρχίας κρυφών μνημών, το εύρος ζώνης προς την κύρια μνήμη και το εύρος ζώνης στον περιφερειακό διάδρομο.

Η αυξανόμενη χρήση των συστοιχιών για εφαρμογές απαιτητικές σε δεδομένα, σε συνδυασμό με την τάση για περισσότερους υπολογιστικούς πυρήνες ανά επεξεργαστή, αυξάνει τον φόρτο του υποσυστήματος Εισόδου/Εξόδου. Η επίδοσή του είναι καθοριστική στο συνολικό ρυθμό εξυπηρέτησης του συστήματος. Για το λόγο αυτό, χρειαζόμαστε μηχανισμούς χαμηλής επιβάρυνσης για την αποδοτική μετακίνηση μεγάλων συνόλων δεδομένων ανάμεσα σε υπολογιστικούς πυρήνες και αποθηκευτικά μέσα. Στην περίπτωση των SMP, η απαίτηση αυτή μεταφράζεται σε μειωμένη κατανάλωση χρόνου CPU και εύρους ζώνης στον διάδρομο μνήμης και τον περιφερειακό διάδρομο.

Η παρούσα διατριβή εξερευνά τις επιπτώσεις του ανταγωνισμού για μοιραζόμενους πόρους σε εξυπηρετητές αποθήκευσης. Μελετάμε την κίνηση των δεδομένων σε σύστημα μοιραζόμενης πρόσβασης επιπέδου μπλοκ πάνω από Myrinet και βρίσκουμε ότι ο κορεσμός του διαδρόμου κύριας μνήμης και του περιφερειακού διαδρόμου επιβαρύνει σημαντικά τη λειτουργία του. Για την αντιμετώπιση του προβλήματος, προτείνουμε τεχνικές για την κατασκευή

αποδοτικών μονοπατιών δεδομένων ανάμεσα σε αποθηκευτικά μέσα και το δίκτυο, στην πλευρά του εξυπηρετητή, και το δίκτυο και τους υπολογιστικούς πυρήνες, στην πλευρά των πελατών. Παρουσιάζουμε το gmblock, ένα σύστημα μοιραζόμενης πρόσβασης επιπέδου μπλοκ το οποίο υποστηρίζει απευθείας μονοπάτι δεδομένων από το δίσκο σε προσαρμογέα Myrinet, παρακάμπτοντας τον επεξεργαστή και το διάδρομο μνήμης. Για βελτιωμένη υποστήριξη αιτήσεων μεγάλου μήκους και υποστήριξη επικάλυψης των φάσεων ανάγνωσης και δικτυακής αποστολής με ελάχιστη εμπλοκή της CPU, εισάγουμε συγχρονισμένες λειτουργίες αποστολής ως επεκτάσεις στο Myrinet/GM. Η σημασιολογία τους επιτρέπει συγχρονισμό της κάρτας δικτύου με εξωτερικό παράγοντα, π.χ. έναν ελεγκτή αποθήκευσης που χρησιμοποιεί το άμεσο μονοπάτι.

Στην πλευρά του πελάτη, το προτεινόμενο σύστημα εκμεταλλεύεται την προγραμματισιμότητα της κάρτας δικτύου για να υποστηρίξει άμεση τοποθέτηση εισερχόμενων τεμαχίων δεδομένων σε απομονωτές διάσπαρτους στη φυσική μνήμη. Η σχεδίαση αυτή καθιστά δυνατή την κίνηση μπλοκ από άκρο σε άκρο χωρίς αντίγραφα, απευθείας από απομακρυσμένο αποθηκευτικό μέσο στο δίκτυο και τελικά στη μνήμη του πελάτη.

Η πειραματική αποτίμηση των προτεινόμενων τεχνικών δείχνει σημαντική αύξηση του ρυθμού απομακρυσμένης Ε/Ε και μειωμένη παρεμβολή στον τοπικό υπολογισμό στην πλευρά του εξυπηρετητή. Από την εκτέλεση διαφόρων μετροπρογραμμάτων σε εγκατάσταση του παράλληλου συστήματος αρχείων OCFS2 πάνω από το gmblock προκύπτει βελτίωση της απόδοσης του συστήματος, με την προϋπόθεση ότι η χρονοδρομολόγηση Ε/Ε στην πλευρά του εξυπηρετητή εξαλείφει τη στενωπό στους δίσκους λόγω ταυτόχρονης πρόσβασης από πολλούς πελάτες.