ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Γιορταμής Εμμανουήλ
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Επικ. Καθηγητής, Π. Πρατικάκης

Παρασκευή, 25 Ιουνίου 2021, ώρα 10:00 π.μ.

Join Zoom Meeting

https://zoom.us/j/92869781408

" Ελαστική διανομή πόρων για εφαρμογή στατικής ανάλυσης κτιρίων"

Περίληψη

Ο υπολογισμός νέφους παρέχει υπολογιστικές υπηρεσίες όπως εξυπηρετητές, αποθηκευτικό χώρο, υπολογιστική ισχύ, βάσεις δεδομένων και δικτύωση, μέσω του διαδικτύου, κατ' απαίτηση. Αντί να αγοράζουν τις δικές τους υποδομές και τα δικά τους υπολογιστικά συστήματα, ιδιώτες και εταιρείες έχουν την δυνατότητα να ενοικιάζουν υπολογιστικούς πόρους υψηλής απόδοσης από παρόχους υπολογισμού νέφους. Αυτοί οι πάροχοι συνήθως προσφέρουν μοντέλα κόστους τύπου "πλήρωσε όσο χρησιμοποιείς", τα οποία χρεώνουν τους χρήστες μόνο για τους πόρους που χρησιμοποιούν. Όμως, οι εφαρμογές τείνουν να έχουν μεταβλητές απαιτήσεις σε πόρους κατά τη διάρκεια εκτέλεσης τους για δύο λόγους: μεταβλητά εισερχόμενα ποσοστά κυκλοφορίας και διαφορετικοί τύποι των εισερχόμενων εργασιών. Η διανομή περιττών πόρων οδηγεί στην μη-αξιοποίηση τους και στην σπατάλη χρημάτων, ενώ η διανομή λιγότερων από τους αναγκαίους πόρους οδηγεί σε παραβιάσεις συμφωνιών επιπέδου υπηρεσίας. Για αυτό τον λόγο, οι πλατφόρμες υπολογισμού νέφους εφαρμόζουν τεχνικές οριζόντιας και κάθετης ελαστικότητας προκειμένου να κλιμακώνουν τους πόρους μιας εφαρμογής σύμφωνα με τις απαιτήσεις της. Οριζόντια ελαστικότητα

σημαίνει προσθήκη περισσότερων πόρων προκειμένου να εκτελεστούν περισσότερες εφαρμογές παράλληλα, ενώ κάθετη σημαίνει αλλαγή στο μέγεθος των υπαρχόντων πόρων προκειμένου να έχουν περισσότερους οι εφαρμογές που ήδη εκτελούνται. Τόσο στην βιομηχανία όσο και στην ακαδημαϊκή βιβλιογραφία υπάρχει σαφώς περισσότερη δουλειά στην οριζόντια παρά στην κάθετη ελαστικότητα. Όμως, η κάθετα ελαστική διανομή πόρων είναι αναγκαία για εφαρμογές των οποίων οι απαιτήσεις σε πόρους αλλάζουν απότομα και εξαρτώνται από τον τύπο των εργασιών που δέχονται.

Σε αυτή την εργασία παρουσιάζουμε έναν κάθετα ελαστικό διανομέα πόρων που διανέμει χρόνο των επεξεργαστών με υψηλή ακρίβεια. Απευθύνεται σε εφαρμογές των οποίων ο παραλληλισμός είναι πολυποίκιλος και εξαρτάται από τον τύπο των εργασιών και δεδομένων που έρχονται. Ο διανομέας συνυπολογίζει τόσο τα μεταβλητά ποσοστά κυκλοφορίας όσο και τις διαφορετικές απαιτήσεις σε πόρους του κάθε φόρτου εργασίας. Εφαρμόσαμε τον διανομέα σε μια Ελληνική εφαρμογή στατικής ανάλυσης ονόματι ΡΑΦ, η οποία χρησιμοποιείται από πολιτικούς μηχανικούς για μελέτες κτηρίων και κατασκευών. Στον πυρήνα της, ο επιλυτής ΡΑΦ υπολογίζει τις στατικές αναλύσεις των κτιρίων λύνοντας εξισώσεις γραμμικής άλγεβρας, η επίλυση των οποίων γίνεται με παράλληλη παραγοντοποίηση τύπου Cholesky. Μέρος της εργασίας ήταν η μεταφορά του επιλυτή από το λειτουργικό σύστημα Windows στο Linux και η μεταφορά του στο νέφος ως υπηρεσία. Στη συνέχεια, η μεθοδική ανάλυση μας κατά τον χρόνο εκτέλεσης της εφαρμογής έδειξε ότι οι μοναδικές ιδιότητες του εκάστοτε κτιρίου οδηγούν σε διαφορετικό δυνατό εύρος παραλληλοποίησης και έτσι, ανάγκες για υπολογιστικούς πόρους. Βάσει αυτών των παρατηρήσεων υλοποιήσαμε τόσο στατικούς όσο και ελαστικούς διανομείς χρόνου του επεξεργαστή. Η εκτενής αξιολόγηση της απόδοσης που εφαρμόσαμε δείχνει ότι ο υψηλής ακρίβειας και κάθετα ελαστικός διανομέας μας οδηγεί σε μεγαλύτερη εκμετάλλευση του παραλληλισμού, αποδοτικότερη αξιοποίηση των πόρων έως και 77%, και έως 10 φορές λιγότερες παραβιάσεις συμφωνιών επιπέδου υπηρεσίας, συγκριτικά με τους στατικούς διανομείς.

University of Crete

Computer Science Department

M.Sc. Thesis presentation / examination

Giortamis Emmanouil

Master's Thesis Supervisor: Assistant Professor, P. Pratikakis

Friday, 25 June 2021, 10:00 a.m.

Join Zoom Meeting

https://zoom.us/j/92869781408

“Elastic resource allocation for a structural design application”

Abstract

Cloud computing is the on-demand delivery of computing services such as servers, storage, computing power, databases and networking, through the internet. Rather that owning their own infrastructure, individuals or companies can rent access to high-performant computing resources from a cloud service provider. Cloud providers typically offer pay-as-you-go pricing models which charge users only for the resources they use. However, applications tend to have varying resource demands depending on both incoming traffic rates and incoming workload types. Resource over-allocation leads to wasted resources and thus, money, while under-allocation leads to Service-Level-Objective (SLO) violations. To this end, cloud computing platforms adopt horizontal and vertical elasticity in order to timely scale the application's resources on demand. Horizontal elasticity replicates the application's resources while vertical elasticity resizes them. It came to our attention that both industry and scientific literature focus more on horizontal elasticity than on vertical elasticity. Vertically elastic resource scaling is essential for applications with workload-dependent and spiky resource demands, however.

In this thesis we present a vertically elastic resource allocator for fine-grained CPU-time allocation. Our proposed algorithm targets applications with job dependent parallelization spikes and accounts for variable traffic rates. Our example application is a Greek commercial structural design application used by civil engineers, named RAF. Its back-end, RAF::Solver, computes a building's static analyses by solving linear algebra equations and factorizing matrices using parallel Cholesky decomposition. Part of our work was to port the RAF::Solver to Linux, containerize, and deploy it as a cloud service. Then, our methodical profiling and benchmark analysis showed that each RAF::Solver instance has different parallelization speedup margins and thus CPU demands, due to each building's unique properties. Based on these observations we implemented both static and elastic CPU-time allocation schemes. Our evaluation analysis indicates that our fine-grained, vertically elastic CPU-time allocator yields better parallelization exploitation, up to 77% higher resource utilization and up to x10 less SLO violations, compared to the static allocation approaches.