Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων
Εργαστηριο Υπολογιστικων Συστηματων

Διπλωματική Εργασία 2013-2014

Αποδοτική συγκέντρωση και αποθήκευση δεδομένων χρήσης παρόχων υπολογιστικών νεφών

Εισαγωγή

Τον τελευταίο καιρό οι υπηρεσίες που προσφέρονται πάνω από υπολογιστικά νέφη έχουν γνωρίσει ιδιαίτερη άνθηση [4]. Εταιρίες που εξυπηρετούν εκατομμύρια χρήστες όπως κοινωνικά δίκτυα (πχ Foursquare [2]) ή παροχείς video on demand όπως η Netflix [3] χρησιμοποιούν υπολογιστικά νέφη πάνω στα οποία εγκαθιστούν τις υποδομές τους και παρέχουν τις υπηρεσίες τους.

Οι πάροχοι υπηρεσιών υπολογιστικών νεφών (cloud providers) διαθέτουν υπολογιστικά κέντρα με συστοιχίες από φυσικές μηχανές (physical machines) διασυνδεδεμένες σε τοπικά δίκτυα μεγάλης ταχύτητας (gigabit Ethernet) και μέσω της τεχνολογίας του virtualization παρέχουν τους πόρους των φυσικών μηχανών (υπολογιστική ισχύς και αποθηκευτικός χώρος) στους πελάτες τους μέσω του διαδικτύου.

Οι cloud providers όπως η Amazon [1], η RackSpace [9] ή ο okeanos [8] που προσφέρουν την υποδομή (εικονικές μηχανές, αποθηκευτικό χώρο και δικτυακές υπηρεσίες) χρειάζονται βέλτιστη αξιοποίηση της υποδομής τους με σκοπό την απρόσκοπτη παροχή πόρων στους πελάτες τους.

Αυτή η βελτιστοποίηση μπορεί να επιτευχθεί με την αξιοποίηση των δεδομένων χρήσης της υπηρεσίας (monitoring data). Η χρήση της υποδομής δημιουργεί ένα μεγάλο μέγεθος από δεδομένα τα οποία μπορούν να αξιοποιηθούν από τους παρόχους για να βελτιώσουν την ποιότητα των υπηρεσιών τους. Τέτοια δεδομένα μπορεί να είναι monitoring information της χρήσης της φυσικής και εικονικής υποδομής των παρόχων, δείγματα δικτυακής κίνησης εντός ή εκτός του data center (με την χρήση της τεχνολογίας sflow [10]), πληροφορία για την φυσική τοποθεσία των εικονικών μηχανών (δηλαδή σε πιο φυσικό μηχάνημα εκτελείται μια εικονική μηχανή, κλπ). Η ευελιξία που προσφέρει η τεχνολογία του virtualization (πχ μετακίνηση εικονικών μηχανών μεταξύ φυσικών, κλπ) μπορεί να αυξήσει ακόμα περισσότερο την βελτίωση της ποιότητας.

Ο μεγάλος όγκος και ρυθμός των δεδομένων χρήσης που παράγονται δεν επιτρέπει την παραδοσιακή μέθοδο συλλογής και αποθήκευσης: η αποθήκευση των μετρικών σε απλά log files δεν είναι αποδοτική, δεν μπορεί εύκολα να κλιμακωθεί και δεν επιτρέπει την δημιουργία ενός κεντρικού αποθετηρίου.

Σκοπός

Ο σκοπός της παρούσας διπλωματικής είναι η δημιουργία ενός κατανεμημένου συστήματος συλλογής των δεδομένων χρήσης που παράγονται από την λειτουργία ενός υπολογιστικού νέφους. Ο στόχος είναι να έχουμε συλλογή και μερική ανάλυση δεδομένων χρήσης σε πραγματικό χρόνο (real-time, low latency), καθώς και αποθήκευση των μετρικών σε κεντρικά αποθετήρια. Το υπολογιστικό νέφος που θα χρησιμοποιήσουμε είναι το okeanos cloud (`http://okeanos.grnet.gr`).

Για την αποδοτική συλλογή δεδομένων θα χρησιμοποιηθούν κατανεμημένες τεχνικές συλλογής ροών δεδομένων, όπως το σύστημα Storm [11] της twitter, το Apache Flume [5], το Apache Kafka [7] της LinkedIn, κλπ. Για την αποδοτική αποθήκευση των συλλεγόμενων δεδομένων θα χρησιμοποιηθούν κατανεμημένες τεχνικές αποθήκευσης σε συστοιχίες υπολογιστών, όπως το σύστημα Apache Hadoop [6].

Πορεία υλοποίησης

Η πορεία υλοποίησης έχει ως εξής:

- **Μελέτη πιθανών μετρικών.** Στο στάδιο αυτό θα εντοπιστούν οι πάροχοι των μετρικών καθώς και τα μετρικά που μπορούν να συγκεντρωθούν. Τέτοιοι πάροχοι μπορεί να είναι ο δικτυακός εξοπλισμός του data center, τα φυσικά μηχανήματα του cloud, οι hypervisors των φυσικών μηχανημάτων, ο scheduler του λογισμικού του υπολογιστικού νέφους (synnefo), κλπ.
- **Επιλογή πλατφόρμας συγκέντρωσης μετρικών.** Στο στάδιο αυτό θα γίνει μια συγκριτική μελέτη των συστημάτων Storm [11], Flume [5] και Kafka [7]. Ανάλογα με τις υπηρεσίες που προσφέρουν, την ευελιξία χρήσης και την δυνατότητα συλλογής των μετρικών του okeanos θα γίνει η επιλογή της πλατφόρμας που θα χρησιμοποιηθεί στην συνέχεια.
- **Ανάπτυξη συστήματος συγκέντρωσης και αποθήκευσης των μετρικών.** Στο στάδιο θα γίνει η ανάπτυξη των απαραίτητων monitoring hooks που θα συνδεθούν με τους παρόχους δεδομένων του okeanos cloud. Τα monitoring hooks θα στέλνουν (push) την πληροφορία που μαζεύεται στο σύστημα συγκέντρωσης μετρικών. Σε αυτή την φάση μπορεί να γίνεται μια πρώτη ανάλυση για την εξαγωγή κάποιας άμεσης πληροφορίας. Κατόπιν, το σύστημα θα αποθηκεύει τα δεδομένα στο αποθετήριο (hadoop) για περαιτέρω offline επεξεργασία.

Επικοινωνία:
Νεκτάριος Κοζύρης, Καθηγητής nkoziris@cslab.ece.ntua.gr
Γιάννης Κωνσταντίνου, Μεταδ. Ερευνητής ikons@cslab.ece.ntua.gr

Βιβλιογραφία

[1] Amazon Elastic Compute Cloud (Amazon EC2). http://aws.amazon.com/ec2/.

[2] AWS Case Study: Foursquare. http://aws.amazon.com/solutions/case-studies/foursquare/.

[3] AWS Case Study: Netflix. https://aws.amazon.com/solutions/case-studies/netflix/.

[4] Customer Success. Powered by the AWS Cloud. https://aws.amazon.com/solutions/case-studies/.

[5] Flume. http://flume.apache.org/.

[6] Hadoop. http://hadoop.apache.org/.

[7] Kafka. http://kafka.apache.org/.

[8] Okeanos IAAS. https://okeanos.grnet.gr/home/.

[9] Rackspace public cloud. http://www.rackspace.com/cloud/.

[10] sFlow. http://www.sflow.org/.

[11] Storm. http://storm-project.net/.