

Πανεπιστήμιο Μακεδονίας

University of Macedonia

ΔΠΜΣ Πληροφοριακά Συστήματα

Master Information Systems

Δίκτυα Υπολογιστών

Computer Networks

Καθηγητής: Α.Α. Οικονομίδης

Professor: A.A. Economides

Αλγόριθμοι Εξισορρόπησης Φορτίου σε Περιβάλλον «Σύννεφου»

Load Balancing Algorithms in Cloud Computing



Πλασταράς Ευριπίδης Α.Μ. 18/11

Φεβρουάριος 2013

Περιεχόμενα

Περίληψη	iii
Abstract	iii
Εισαγωγή	1
1. Εξισορρόπηση Φορτίου (Load Balancing)	3
1.1 Ορισμός και Στόχοι	3
1.2 Βασικοί Πόροι Συστήματος	4
1.3 Συστατικά Στοιχεία και Διαδικασία Εξισορρόπησης Φορτίου	6
1.4 Τρόποι Επίτευξης Εξισορρόπησης Φορτίου	8
1.5 Οφέλη και Πλεονεκτήματα του Server Load Balancing	9
1.6 Τύποι Αλγορίθμων Εξισορρόπησης Φορτίου	10
2. Υπολογιστική «Σύννεφου» (Cloud Computing)	11
2.1 Ορισμός	11
2.2 Συστατικά μέρη του Cloud	12
2.3 Τεχνολογίες που Χρησιμοποιεί το Cloud	13
2.4 Μοντέλο Υπηρεσιών Cloud	14
2.5 Τύποι Cloud	16
2.6 Βασικά Χαρακτηριστικά	18
2.7 Πλεονεκτήματα και Προκλήσεις	20
3. Αλγόριθμοι Εξισορρόπησης Φορτίου σε Περιβάλλον «Σύννεφου»	22
3.1 Εξισορρόπηση Φορτίου σε Περιβάλλον «Σύννεφου»	22
3.2 Απλοί Αλγόριθμοι Εξισορρόπησης Φορτίου	22
3.3 Αλγόριθμος Ισόποσης Διάδοσης Τρέχουσας Εκτέλεσης (ESCE)	24
3.4 Αλγόριθμος Εξισορρόπησης Φορτίου LB3M	25
3.5 Αλγόριθμος Εξισορρόπησης Φορτίου CMLB	26
3.6 Αλγόριθμος Εξισορρόπησης Φορτίου με Επίγνωση Ενέργειας	28
3.7 Cloud Load Balancing μια προέκταση του Global Server Load Balancing	29
4. Συμπεράσματα και Προτάσεις για Μελλοντική Έρευνα	31
Βιβλιογραφία	32

Περίληψη

Η εξισορρόπηση φορτίου αποτελεί αναπόσπαστο κομμάτι ενός δικτύου ή ενός κατανεμημένου συστήματος γενικότερα, προκειμένου να επιτευχθεί η εύρυθμη λειτουργία του, ιδίως όταν σχετίζεται με αυξημένο όγκο δεδομένων. Η τεχνολογία της υπολογιστικής «σύννεφου» έχει σημειώσει τα τελευταία χρόνια ραγδαία διάδοση και συνεχώς εξελίσσεται, πράγμα που σημαίνει ότι ένα τέτοιου είδους σύστημα διαχειρίζεται μεγάλο όγκο δεδομένων. Το γεγονός αυτό καθιστά σαν απαραίτητο στόχο την εξισορρόπηση φορτίου, προκειμένου το σύστημα να λειτουργεί αποδοτικά. Έχουν αναπτυχθεί αρκετοί αλγόριθμοι που επιτυγχάνουν αυτό το στόχο, είτε απλοί, είτε σύνθετοι. Σκοπός της εργασίας αυτής είναι να παρουσιάσει ορισμένους από αυτούς, καθώς επίσης, και να κάνει σαφή την έννοια της εξισορρόπησης φορτίου και πόσο χρήσιμη είναι σε ένα περιβάλλον «σύννεφου».

Abstract

Load Balancing is an integral part of a network or a distributed system in order to operate efficiently, especially when it is associated with an increased volume of data. The technology of Cloud Computing has spread rapidly in recent years and it constantly evolving, which means that such systems handle large amounts of data. This fact, makes it necessary for the system to balance the load, in order to operate efficiently. There have been developed several algorithms which achieve this goal, both simple and complex. This paper means to present some of these algorithms, as well to make clear the concept of load balancing and how useful it is in Cloud Computing.

Εισαγωγή

Έναν απλό ορισμό που θα μπορούσαμε να δώσουμε στον όρο «σύννεφο» (**cloud**) ή υπολογιστική «σύννεφου» (**cloud computing**) είναι ότι λειτουργώντας σε περιβάλλον cloud μας δίνεται η δυνατότητα πρόσβασης σε αρχεία, δεδομένα, εφαρμογές και υπηρεσίες από τον φυλλομετρητή μας (**web browser**) μέσω του Ίντερνετ, τα οποία προσφέρονται από κάποιον τρίτο πάροχο, πληρώνοντας μόνο για τους υπολογιστικούς πόρους και υπηρεσίες που χρησιμοποιήσαμε. (Won, 2009). Βασικά συστατικά του cloud computing είναι οι εξυπηρετητές (**servers**) και τα κέντρα δεδομένων (**data centers**). Οι servers είναι υπεύθυνοι για την διαχείριση των αιτημάτων των πελατών (**clients**), ενώ στα κέντρα δεδομένων είναι αποθηκευμένη όλη η πληροφορία. Είναι προφανές ότι η εκτεταμένη χρήση του Ίντερνετ σε συνδυασμό με το γεγονός ότι πλέον τα πάντα βρίσκονται στον παγκόσμιο ιστό, εφόσον μιλάμε για περιβάλλον «σύννεφου», δημιουργεί ένα τεράστιο φόρτο αιτημάτων και εργασιών τόσο για τους servers, όσο και για τα data centers. Επομένως κρίνεται απαραίτητη η εξισορρόπηση αυτού του φορτίου.

Η έννοια φόρτος και η διαχείρισή του είναι άμεσα συνδεδεμένη με τα δίκτυα, όπου για την ομαλή λειτουργία ενός δικτύου και τη σωστή κατανομή των πόρων απαιτείται η εξισορρόπηση του φόρτου αυτού. Εξισορρόπηση φορτίου είναι η διαδικασία διανομής του φόρτου μεταξύ των κόμβων ενός δικτύου, με σκοπό να βελτιωθεί τόσο η αξιοποίηση των πόρων, όσο και ο χρόνος απόκρισης των εργασιών, αποφεύγοντας παράλληλα καταστάσεις όπου κάποιοι κόμβοι είναι βαριά φορτωμένοι, ενώ άλλοι είναι είτε σε αδράνεια, είτε υπολειτουργούν. Με την εξισορρόπηση φορτίου εξασφαλίζουμε ότι κάθε κόμβος του δικτύου εργάζεται περίπου το ίδιο, οποιαδήποτε χρονική στιγμή. (Kaur, 2012).

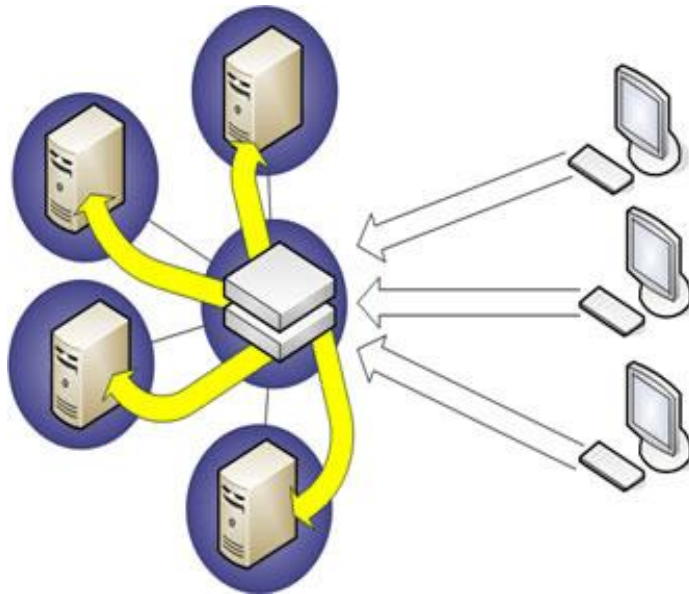
Στην περίπτωση των συστημάτων cloud, ο ρόλος των εξισορροποιητών φορτίου (**load balancers**) είναι να βελτιώσουν την απόδοση και τη διαθεσιμότητα των πόρων και των υπηρεσιών που προσφέρονται. Είναι γεγονός ότι η εξισορρόπηση φορτίου σε συστήματα cloud αποτελεί πρόκληση, καθώς δεν είναι πάντα πρακτικά εφικτό ή οικονομικά αποδοτικό να διατηρούμε έναν ή περισσότερους πόρους αδρανείς, μόνο για να ικανοποιηθεί η απαιτούμενη ζήτηση. (Padhy & Rao, 2011).

Σκοπός της εργασίας είναι να παρουσιάσει τους διάφορους αλγόριθμους εξισορρόπησης φορτίου σε συστήματα cloud, ρίχνοντας το βάρος σε αυτούς που επικρατούν περισσότερο σήμερα. Παράλληλα, γίνεται και μια προσπάθεια αξιολόγησης και σύγκρισης αυτών με σκοπό την εξαγωγή κάποιου γενικού συμπεράσματος. Έτσι, αρχικά προσδιορίζουμε την έννοια εξισορρόπηση φορτίου και την αναγκαιότητά της σε ένα περιβάλλον δικτύου. Έπειτα, αναλύουμε τα κύρια στοιχεία, καθώς και τις βασικές τεχνολογίες που χρησιμοποιεί ένα σύστημα «Σύννεφου». Τέλος, παρουσιάζουμε την χρησιμότητα της εξισορρόπησης φορτίου σε ένα περιβάλλον cloud, μέσω της παρουσίασης βασικών αλγορίθμων.

1. Εξισορρόπηση Φορτίου (Load Balancing)

1.1 Ορισμός και Στόχοι

Με τον όρο εξισορρόπηση φορτίου εννοούμε την ικανότητα να διαμοιράζουμε τον φόρτο εργασίας μιας εφαρμογής σε ένα πλήθος ξεχωριστών συστημάτων με σκοπό την αύξηση της συνολικής απόδοσης και της διαχείρισης των εισερχόμενων αιτήσεων προς την συγκεκριμένη εφαρμογή. Είναι μια διαδικασία που αρχικά φαίνεται απλή και βοηθάει όλα να γίνουν γρηγορότερα, όμως υπάρχουν και αρκετοί παράγοντες που αυξάνουν το επίπεδο πολυπλοκότητας. (Membrey, Plugge, & Hows, 2012).



Εικόνα 1: Load Balancing¹

Αν θέλαμε να δώσουμε ένα ορισμό σε αυτήν την διαδικασία θα ήταν: «Η εξισορρόπηση φορτίου είναι μια μέθοδος δικτύωσης υπολογιστών για το διαμοιρασμό του φόρτου εργασίας σε πολλούς υπολογιστές ή σε μια συστάδα υπολογιστών, σε συνδέσεις δικτύου, κεντρικές μονάδες

¹ Πηγή: <http://www.provider-one.net/products-and-services/dedicated/load-balancing.html>

επεξεργασίας, δίσκους, ή άλλους πόρους, για να επιτευχθεί η βέλτιστη αξιοποίηση των πόρων, η μεγιστοποίηση της απόδοσης, η ελαχιστοποίηση του χρόνου απόκρισης και η αποφυγή υπερφόρτωσης. Η υπηρεσία αυτή, συνήθως παρέχεται από ειδικό λογισμικό (**software**) ή υλικό (**hardware**).» (Wikipedia, 2012).

Οι στόχοι της εξισορρόπησης φορτίου σύμφωνα με τους Padhy και Rao (2011) είναι η ουσιαστική βελτίωση της απόδοσης, η ύπαρξη σχεδίου back up σε περίπτωση που το σύστημα αποτύχει, η διατήρηση της σταθερότητας του συστήματος και να διευκολυνθεί η μελλοντική τροποποίηση.

1.2 Βασικοί Πόροι Συστήματος

Όπως έχει αναφερθεί παραπάνω η εξισορρόπηση φορτίου έχει να κάνει με τον διαμοιρασμό των πόρων του συστήματος προκειμένου να ολοκληρωθεί μια διαδικασία. Υπάρχουν τέσσερις βασικοί πόροι, εξίσου σημαντικοί, που συνθέτουν έναν εξυπηρετητή (**server**), καθένας από τους οποίους επηρεάζει την απόδοση του συστήματος. Αυτοί είναι: ο επεξεργαστής, η μνήμη, το δίκτυο και ο χώρος αποθήκευσης. (Membrey, Plugge, & Hows, 2012).

Η κύρια δουλειά του επεξεργαστή (**processor**) είναι να επεξεργάζεται ότι εισέρχεται στο σύστημα μας, επιστρέφοντας κάποια απάντηση. Βρίσκεται σε συνεχή λειτουργία και είναι ένα από τα βασικά μέτρα απόδοσης σε οποιοδήποτε σύστημα. Για τη χρήση και διατήρηση μιας βάσης δεδομένων (**database**) μπορεί να χρησιμοποιηθεί μεγάλη ποσότητα επεξεργασίας. Δίνοντας έναν server σε μια βάση αφαιρείτε ο ανταγωνισμός μεταξύ web server και βάσης, επιτρέποντας την επεξεργασία χωρίς παρεμβολές. Η διάσπαση διαφορετικών εφαρμογών για την αύξηση της επεξεργαστικής ισχύς αποτελεί τον πυρήνα της εξισορρόπησης φορτίου. (Membrey, Plugge, & Hows, 2012)

Η *μνήμη (memory)* αποτελεί τις πληροφορίες που αποθηκεύονται από το λειτουργικό σύστημα βραχυχρόνια. Ένα μεγάλο μέρος στοιχείων του συστήματος μας βρίσκεται εδώ, επειδή μπορεί να προσεγγιστεί πολύ πιο γρήγορα. Η βελτιστοποίηση της απόδοσης της μνήμης δεν είναι τόσο δύσκολη, αλλά μπορεί να επηρεάσει την συνολική απόδοση. Οι περισσότερες βάσεις δεδομένων καταναλώνουν μεγάλο μέρος μνήμης, διότι επιθυμούν να κρατήσουν όσο το δυνατόν περισσότερες πληροφορίες στη μνήμη, ώστε να υπάρχει γρηγορότερη πρόσβαση σε αυτές. Ο διαμοιρασμός μιας διαδικασίας σε δύο διαφορετικούς servers αυξάνει την απόδοση τόσο της βάσης δεδομένων, όσο και του web server. (Membrey, Plugge, & Hows, 2012)

Το *δίκτυο (network)* αντιπροσωπεύει την τρέχουσα απόδοση μιας συσκευής δικτύου και το ποσοστό που χρησιμοποιείται τη δεδομένη στιγμή για μεταφορά δεδομένων. Εκφράζει επίσης, το χρόνο που απαιτείται για τη μεταφορά δεδομένων μεταξύ δύο σημείων. Αν αρκετή χωρητικότητα δεν είναι διαθέσιμη, τα δεδομένα δεν μπορούν να μεταφερθούν από σημείο σε σημείο. Είναι εμφανές ότι το δίκτυο παίζει σημαντικό ρόλο στη συνολική απόδοση ενός συστήματος. (Membrey, Plugge, & Hows, 2012).

Η έννοια *αποθηκευτικός χώρος (storage)* αναφέρεται στην ποσότητα της φυσικής αποθήκευσης που υπάρχει στον εξυπηρετητή μας. Μπορούμε να την μετρήσουμε, είτε με τη ποσότητα που εισέρχεται/εξέρχεται στο σύστημα ανά δευτερόλεπτο, είτε με τον συνολικό όγκο που καταναλώνεται. Γενικά, ο αποθηκευτικός χώρος είναι σημαντικός, καθώς αν ο server μείνει χωρίς χώρο τότε θα δημιουργηθούν πληθώρα προβλημάτων. (Membrey, Plugge, & Hows, 2012)

1.3 Συστατικά Στοιχεία και Διαδικασία Εξισορρόπησης Φορτίου

Προτού αναφέρουμε ποια είναι η βασική διαδικασία εξισορρόπησης φορτίου, καλό θα ήταν να παραθέσουμε τα κύρια συστατικά της διαδικασίας αυτής. Αξίζει να πούμε ότι έχουμε να κάνουμε με client-server αρχιτεκτονική.

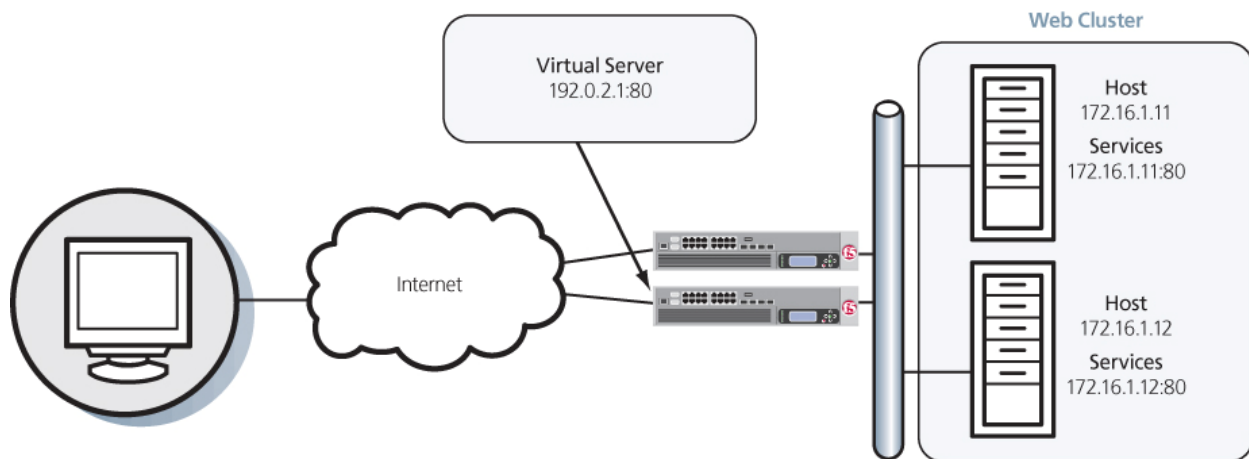
Αρχικά, χρησιμοποιούμε τους όρους κόμβος (**node**), πάροχος (**host**), και εξυπηρετητής (**server**). Οι τρεις όροι πολλές φορές εννοούν το ίδιο πράγμα και άλλες πάλι φορές υφίστανται σαν τρεις διαφορετικές οντότητες, και οι τρεις όμως δέχονται το φορτίο από έναν εξισορροποιητή φορτίου (**load balancer**). Κάθε ένας load balancer αναπαριστά διαφορετικά τον host ή τον server, και διαφορετικά τις υπηρεσίες που αυτοί παρέχουν.

Σε αρκετές περιπτώσεις, ο διαμοιρασμός του φόρτου γίνεται με τη χρήση παραπάνω του ενός server με σκοπό την αύξηση της απόδοσης. Αυτό προϋποθέτει ότι το σύστημα μας διαθέτει μια συστάδα (**cluster**) ή μια «φάρμα» (**farm**) από servers.

Επίσης, σημαντική είναι και η έννοια του εικονικού διακομιστή (**virtual server**). Ένας Virtual server είναι πολλαπλοί servers που εμφανίζονται ως ένας, με σκοπό την εξισορρόπηση φορτίου. (PC Magazine, 2012).

Συνδυάζοντας όλες τις παραπάνω έννοιες προκύπτουν τα βασικά βήματα για την εξισορρόπηση φορτίου. Ένας load balancer παρουσιάζει έναν virtual server στον έξω κόσμο. Ο virtual server από την μεριά του υποδεικνύει σε μια συστάδα υπηρεσίες παρεχόμενες από έναν host και ούτω καθεξής. (Salchow, 2007)

Ο εξισορροποιητής φορτίου βρίσκεται συνήθως μεταξύ του πελάτη και του παρόχου των υπηρεσιών που ο πρώτος επιθυμεί να χρησιμοποιήσει. Υπάρχει ο εικονικός διακομιστής που ρυθμίζει σε ποιους server θα μοιραστεί η επεξεργασία του αιτήματος του πελάτη.



Εικόνα 2: Βασικά συστατικά εξισορρόπησης φορτίου²

Αρχικά, ο πελάτης (**client**) προσπαθεί να συνδεθεί με την υπηρεσία στον load balancer. Ο load balancer αποδέχεται τη σύνδεση και αφού αποφασίσει ποιος host θα πάρει τη σύνδεση, αλλάζει τη IP διεύθυνση προορισμού προκειμένου να ταιριάζει με την υπηρεσία του επιθυμητού παρόχου. Έπειτα ο πάροχος αποδέχεται τη σύνδεση και απαντάει στον πελάτη μέσω του load balancer.

Ο load balancer παρεμβάλλεται στο επιστρεφόμενο πακέτο από τον host και αλλάζει την IP της πηγής για να ταιριάζει με την IP του virtual server. Έπειτα προωθεί το πακέτο πίσω στον πελάτη, ο οποίος το παραλαμβάνει, πιστεύοντας ότι προήλθε από τον virtual server. (Salchow, 2007)

² Salchow, K. j. (2007). Load Balancing 101: Nuts and Bolts.

1.4 Τρόποι Επίτευξης Εξισορρόπησης Φορτίου

Προκειμένου να αυξηθεί η συνολική απόδοση του συστήματος, πρέπει να διανείμουμε αποδοτικά παραπάνω συστατικά.

Αρχικά, μπορούμε να εφαρμόσουμε διαδικασίες διαχωρισμού σε ένα πλήθος server, έτσι ώστε να μοιράζεται ο φόρτος σε διαφορετικούς server, πράγμα που μπορεί σε μεγάλο βαθμό να αυξήσει το ποσό της διαθέσιμης υπολογιστικής επεξεργαστικής ισχύς. Για παράδειγμα, ξεχωριστοί web servers από τους database servers. Επίσης, η προσθήκη ειδικών εργαλείων σε εφαρμογές, προκειμένου να εξισορροπήσουν το ύψος της επεξεργασίας που απαιτείται για κάθε λειτουργία είναι ένας ακόμη τρόπος. Για παράδειγμα, η προσωρινή αποθήκευση των κοινών δεδομένων μιας ιστοσελίδας οδηγεί στην εξοικονόμηση χρόνου. (Membrey, Plugge, & Hows, 2012).

Παράλληλα, μπορούμε να ρυθμίσουμε τον εξισορροποιητή φορτίου να εργάζεται ανά προορισμό (**per-destination**) ή ανά πακέτο (**per-packet**). Η ανά προορισμό λειτουργία σημαίνει ότι ο εκάστοτε δρομολογητής διανέμει τα πακέτα ανάλογα με τη διεύθυνση προορισμού. Διατηρείται έτσι η σειρά των πακέτων, με πιθανότητα όμως, ένας server να λαμβάνει το μεγαλύτερο μέρος της κυκλοφορίας και κάποιος άλλος να μένει αδρανής. Αντίθετα, η ανά πακέτο εξισορρόπηση φορτίου σημαίνει ότι ο εκάστοτε δρομολογητής στέλνει το πρώτο πακέτο από μία διαδρομή, το δεύτερο από μία άλλη και ούτω καθεξής. Εγγύαται ίσο φορτίο για όλες τις συνδέσεις. Ωστόσο, υπάρχει η πιθανότητα τα πακέτα να φθάσουν με λάθος σειρά, καθώς μπορεί να υπάρξει διαφορετική καθυστέρηση εντός του δικτύου. (CISCO, 2005).

1.5 Οφέλη και Πλεονεκτήματα του Server Load Balancing

Το Server Load Balancing είναι η διαδικασία διαμοιρασμού αιτημάτων για υπηρεσίες σε μια ομάδα από εξυπηρετητές. Είναι συνδεδεμένη με σημαντικές απαιτήσεις ενός δικτύου, όπως η κλιμάκωση, η υψηλή απόδοση, η υψηλή διαθεσιμότητα και η αποκατάσταση καταστροφών. Η διαδικασία αυτή επιτρέπει σε πολλούς εξυπηρετητές να εμφανίζονται σαν ένας (virtual server).

Η υψηλή απόδοση επιτυγχάνεται όταν η υπολογιστική δύναμη των servers χρησιμοποιείται έξυπνα. Οι εξισορροποιητές κατευθύνουν τις αιτήσεις του χρήστη σε διακομιστές (servers) που είναι λιγότερο απασχολημένοι και άρα σε θέση να παρέχουν ταχύτερους χρόνους απόκρισης.

Παράλληλα αν μια υπηρεσία ή ένας server παρουσιάσει πρόβλημα, τότε ο εξισορροποιητής φορτίου ανακατανέμει τα αιτήματα των τελικών χρηστών σε άλλους servers, προκειμένου να διατηρηθεί η διαθεσιμότητα του συστήματος. (Radharc-Communications, 2012).

Τεχνικές όπως η εξισορρόπηση φορτίου server χρησιμοποιούνται συνήθως σε ένα ενιαίο κέντρο δεδομένων (**datacenter**) για τη δημιουργία ομάδας εφαρμογών και / ή servers ανεκτικούς σε σφάλματα και κλιμακούμενους. Οι οργανισμοί που επιθυμούν να προστατεύουν από τους κινδύνους, φυσικές καταστροφές, άρνηση εξυπηρέτησης, κ.α. επιλέγουν συχνά να αναπτύξουν μια παγκόσμια λύση εξισορρόπησης φόρτου (**Global Server Load Balancing**) που δρομολογεί τις εφαρμογές σε πολλαπλά κέντρα δεδομένων και αφαιρεί έτσι σημεία της αποτυχίας. Υπάρχουν δύο βασικά πλεονεκτήματα από την υλοποίηση της παγκόσμια εξισορρόπησης φορτίου, η επιχειρησιακή συνέχεια και η βελτίωση της πελατειακής εμπειρίας, το οποίο επιτυγχάνεται με τη χρήση του κοντινότερου και αποδοτικότερου datacenter ως προς το τον πελάτη. (Adler, 2010)

1.6 Τύποι Αλγορίθμων Εξισορρόπησης Φορτίου

Ανάλογα με την τρέχουσα κατάσταση του συστήματος, οι αλγόριθμοι εξισορρόπησης φορτίου μπορεί να διαιρεθεί σε δύο κατηγορίες, στους **στατικούς** που δεν λαμβάνουν υπόψη την τρέχουσα κατάσταση του συστήματος, οπότε απαιτείται προγενέστερη γνώση του συστήματος, και στους **δυναμικούς**, όπου η εξισορρόπηση φορτίου βασίζεται στην τρέχουσα κατάσταση του συστήματος και δεν απαιτείται προγενέστερη γνώση.

Ένας δυναμικός αλγόριθμος εξισορρόπησης φορτίου, όπως είπαμε, δεν υπολογίζει την προηγούμενη κατάσταση ή τη συμπεριφορά του συστήματος, επειδή εξαρτάται από την παρούσα συμπεριφορά του. Παράγοντες που λαμβάνονται υπόψη κατά την ανάπτυξη τέτοιων αλγορίθμων είναι: η εκτίμηση του φόρτου, η σύγκριση του φόρτου, η σταθερότητα των διάφορων συστημάτων, οι επιδόσεις του συστήματος, η αλληλεπίδραση μεταξύ των κόμβων, η φύση της επεξεργασίας που απαιτείται, η επιλογή των κόμβων κ.α. (Padhy & Rao, 2011)

Παράλληλα, ανάλογα με το ποιος ξεκίνησε τη διαδικασία, οι αλγόριθμοι εξισορρόπησης φορτίου μπορούν να χωριστούν σε sender initiated (αποστολέας), receiver initiated (δέκτης) και symmetric (συνδυασμός των δύο).

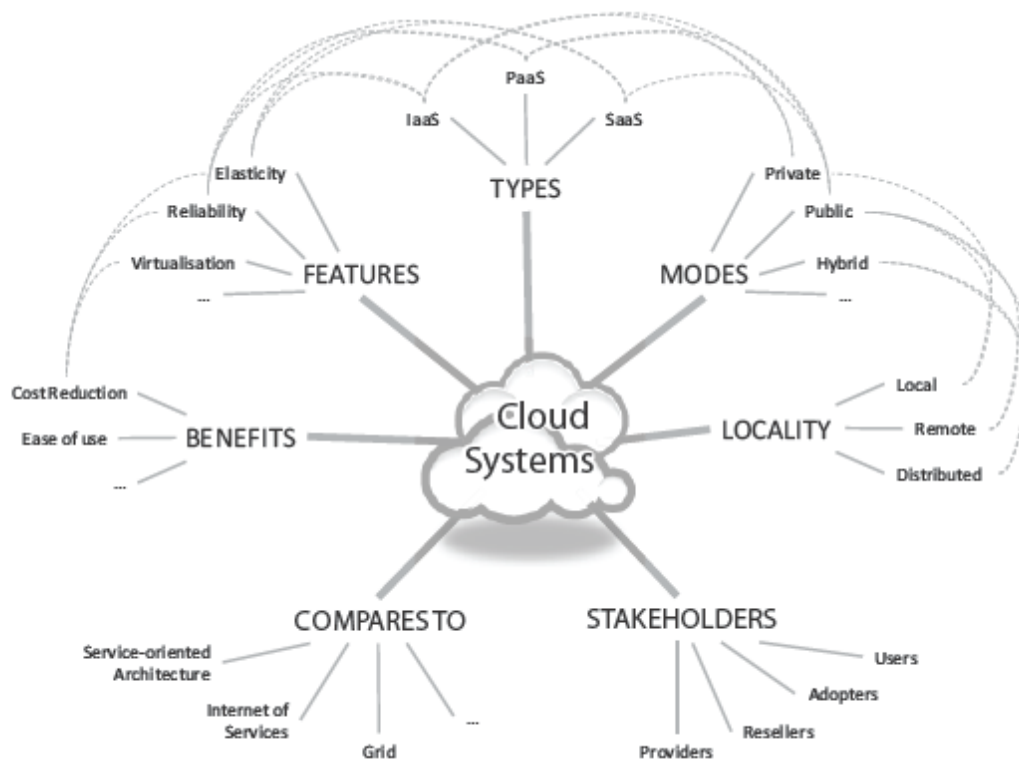
2. Υπολογιστική «Σύννεφου» (Cloud Computing)

2.1 Ορισμός

Η έννοια της υπολογιστικής «σύννεφου» (**cloud computing**) γενικά είναι πολύ μπερδεμένη και πολύ μεταβαλλόμενη. Ξεκίνησε σαν επιστημονικό υπολογιστικό εργαλείο, αλλά η αρχιτεκτονική του το μετέτρεψε σε μια νέα τεχνολογία. Η δυνατότητα πρόσβασης σε Internet υψηλής ταχύτητας, το ισχυρό δίκτυο server και αποθήκευσης σε data centers, αλλά και οι εξελίξεις στη δημιουργία υψηλής απόδοσης λογισμικού για data centers και το Web, μας δείχνουν ότι το cloud computing θα επικρατήσει.

Δεν είναι εφικτό να δώσουμε έναν ξεκάθαρο ορισμό, όμως ένας ορισμός που ερμηνεύει την έννοια cloud θα μπορούσε να είναι: «Το cloud computing είναι ένα μοντέλο επεξεργασίας πληροφοριών στο οποίο οι κεντρικές διαχειριστικές υπολογιστικές δυνατότητες παρέχονται σαν υπηρεσίες, ανάλογα με τις ανάγκες, μέσω ενός δικτύου σε μια πληθώρα από συσκευές χρηστών» (Chee & Franklin, 2010).

Αν θέλουμε να δώσουμε έναν πιο απλό ορισμό, σύμφωνα με τον Bobrowski (2010), το «σύννεφο» είναι ένα μέρος όπου πόροι όπως hardware, λειτουργικά συστήματα, δίκτυα, αποθηκευτικός χώρος, βάσεις δεδομένων, αλλά και ολοκληρωμένες εφαρμογές γίνονται άμεσα και κατόπιν αιτήσεως διαθέσιμοι.



Εικόνα 3: Το περιβάλλον του Cloud Computing³

2.2 Συστατικά μέρη του Cloud

Ένα περιβάλλον «σύννεφου» αποτελείται από τρία βασικά συστατικά: τους πελάτες (**clients**), τα datacenters και τους servers (Padhy & Rao, 2011).

Οι clients δεν είναι τίποτα άλλο από τις εφαρμογές που χρησιμοποιούν οι τελικοί χρήστες προκειμένου να διαχειριστούν οποιαδήποτε πληροφορία σχετική με το cloud. Οι εφαρμογές μπορεί να είναι μικρές και απλές για χρήση ακόμη και μέσω κινητού τηλεφώνου ή tablet. Η λειτουργία τους επικεντρώνεται στην εμφάνιση της πληροφορίας ή ακόμα και στη διαχείριση του ίδιου του server.

³ Danchev, S., Tsakanikas, A., & Ventouris, N. (2011). Cloud Computing: A Driver for Greek Economy Competitiveness.

Από την άλλη πλευρά, τα datacenters δεν είναι τίποτα άλλο από μια συλλογή από servers όπου είναι εγκατεστημένο πλήθος εφαρμογών. Ο τελικός χρήστης συνδέεται μέσω του client για να αποκτήσει καινούριες εφαρμογές ή απλά να διαχειριστεί τις ήδη υπάρχουσες.

Τέλος, οι servers είναι υπεύθυνοι για όλη τη δουλειά πίσω από το cloud. Συνήθως, βρίσκονται κατανεμημένοι σε διάφορα σημεία και επικοινωνούν μεταξύ τους μέσω του Internet. Ο τελικός χρήστης χρησιμοποιεί τις εφαρμογές που είναι εγκατεστημένες σε αυτούς, όμως είναι σαν να βρίσκονται στο μηχάνημα του.

2.3 Τεχνολογίες που Χρησιμοποιεί το Cloud

Το cloud computing αξιοποιεί υπάρχουσες τεχνολογίες και υπηρεσίες που είναι συνδεδεμένες με το Internet. Είναι βασισμένο στην έννοια της αφαίρεσης (**abstraction**) και στην έννοια της εικονικοποίησης (**virtualization**).

Με τον όρο αφαίρεσης (**abstraction**) εννοούμε ότι οι λεπτομέρειες του πώς υλοποιείται το σύστημα μας αποκρύπτονται από τους χρήστες και τους προγραμματιστές. Η εφαρμογή τρέχει σε συστήματα τα οποία δεν είναι αδύνατο να προσδιοριστούν, όπως επίσης και τα δεδομένα είναι αποθηκευμένα σε άγνωστες τοποθεσίες. Συγχρόνως, η διαχείριση των συστημάτων γίνεται από τρίτους και συνήθως τον πάροχο του συστήματος. Η πρόσβαση στην εφαρμογή από τους τελικούς χρήστες μπορεί να γίνει από οπουδήποτε (Danchev, Tsakanikas, & Ventouris, 2011).

Παράλληλα, με την έννοια της αφαίρεσης υπάρχει και η έννοια της εικονικοποίησης (**virtualization**). Το cloud computing δημιουργεί εικονικά συστήματα μέσω της συγκέντρωσης και του διαμοιρασμού πόρων. Συστήματα και αποθηκευτικός χώρος παρέχονται από μια κεντρική υποδομή, σύμφωνα με το πόσο χρειάζονται. Οι πόροι που παρέχονται είναι κλιμακούμενοι, ανάλογα με λειτουργία της εφαρμογής. Το κόστος αξιολογείται ανάλογα με το

πόσο χρησιμοποιείται η υπηρεσία, δηλαδή αν χρησιμοποιήσω μια εφαρμογή cloud θα πληρώσω ανάλογα με το πόσο τη χρησιμοποίησα (Danchev, Tsakanikas, & Ventouris, 2011). Σύμφωνα και με τους Chee και Franklin (2010), ένα ολόκληρο λειτουργικό σύστημα ή μια εφαρμογή μπορεί να ενσωματωθεί σε ένα πακέτο αρχείων, τα οποία φαινομενικά θα τρέχουν στο μηχάνημα μας, αλλά στην ουσία θα τρέχουν σε ένα εικονικό περιβάλλον.

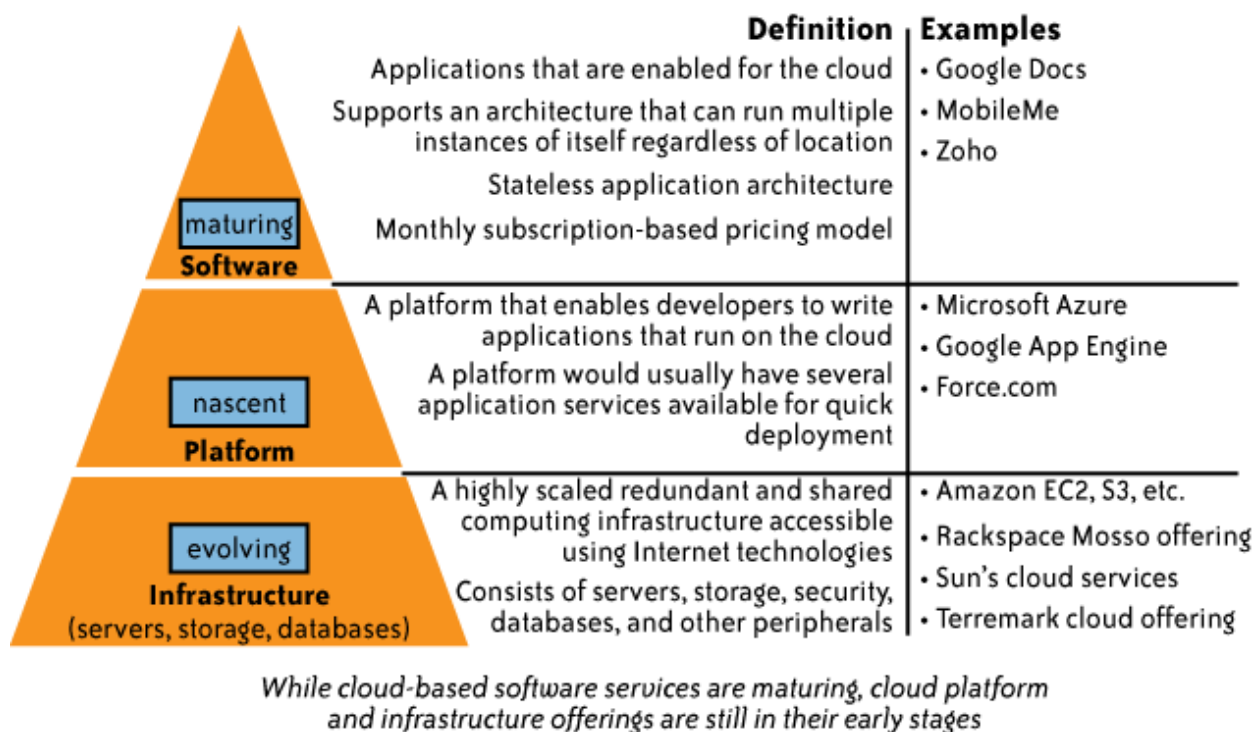
Επίσης, το virtualization μπορεί να πάρει δύο μορφές, είτε πλήρης εικονικοποίηση, είτε μερική. Στην πρώτη περίπτωση ένα ολόκληρο μηχάνημα εγκαθίστατε στο μηχάνημα μας, ενώ η δεύτερη είναι η περίπτωση όπου διαφορες εφαρμογές τρέχουν σε ένα απομακρυσμένο μηχάνημα, για λόγους εξοικονόμησης χώρου ή πιο σημαντικά για πρόληψη από καταστροφές (**disaster recovery**) (Padhy & Rao, 2011).

2.4 Μοντέλο Υπηρεσιών Cloud

Το μοντέλο υπηρεσιών cloud, σύμφωνα με τους Mather, Kumaraswamy, και Latif (2009), χωρίζεται σε τρεις κατηγορίες: λογισμικό-σαν-υπηρεσία (Software-as-aService – SaaS), πλατφόρμα-σαν-υπηρεσία (Platform-as-a-Service – PaaS), υποδομή σαν υπηρεσία (Infrastructure-as-a-Service – IaaS).

Το λογισμικό-σαν-υπηρεσία (**SaaS**) παρέχει τη δυνατότητα στον τελικό χρήστη να χρησιμοποιεί εφαρμογές σε περιβάλλον «σύννεφου». Οι εφαρμογές αυτές είναι προσβάσιμες από διάφορες συσκευές, μέσω μιας διασύνδεσης πελάτη (**client**), όπως ένας web browser ή μια διεπαφή προγράμματος. Η βασική υποδομή του cloud, όπως το δίκτυο, οι servers, το λειτουργικό σύστημα, ο αποθηκευτικός χώρος, ακόμη και κάποιες μεμονωμένες δυνατότητες της εφαρμογής, διαχειρίζονται και ελέγχονται από έναν τρίτο, συνήθως τον πάροχο της εφαρμογής και όχι από τον τελικό χρήστη (Mell & Grance, 2011). Γενικά, σε μια τέτοια μορφής υπηρεσία ο

χρήστης δεν αγοράζει το λογισμικό, αλλά το μισθώνει με βάση μια συνδρομή ή ανάλογα με τη χρήση. Σε πολλές περιπτώσεις η εφαρμογή είναι δωρεάν για περιορισμένη χρήση (Mather, Kumaraswamy, & Latif, 2009).



Εικόνα 4: Μοντέλο υπηρεσιών Cloud⁴

Σε πολλές περιπτώσεις η εφαρμογή είναι δωρεάν για περιορισμένη χρήση (Mather, Kumaraswamy, & Latif, 2009).

Από την άλλη πλευρά, με την πλατφόρμα-σαν-υπηρεσία (**PaaS**) η χρήστης έχει τη δυνατότητα να αναπτύξει και να φιλοξενήσει, πάνω στη υποδομή cloud που του προσφέρεται τις δικές του εφαρμογές. Οι εφαρμογές αυτές δημιουργήθηκαν με τη χρήση γλωσσών προγραμματισμού, βιβλιοθηκών, υπηρεσιών και εργαλείων που υποστηρίζει ο πάροχος. Ο τελικός χρήστης δεν διαχειρίζεται την βασική υποδομή του cloud, αλλά ελέγχει τις

⁴ Mather, T., Kumaraswamy, S., & Latif, S. (2009). Cloud Security and Privacy. O'Reilly.

εγκατεστημένες εφαρμογές και τις ρυθμίσεις για περιβάλλον φιλοξενίας εφαρμογών (hosting) (Mell & Grance, 2011). Κατά τους Mather, Kumaraswamy, και Latif (2009), το PaaS είναι στην ουσία μια παραλλαγή του SaaS όπου το περιβάλλον ανάπτυξης των εφαρμογών προσφέρεται ως υπηρεσία.

Τέλος, με την υποδομή-σαν-υπηρεσία (**IaaS**) η δυνατότητα που προσφέρεται στον χρήστη είναι η παροχή θεμελιωδών υπολογιστικών πόρων, όπως επεξεργαστές, δίκτυα, αποθηκευτικός χώρος κ.α. Ο τελικός χρήστης μπορεί να αναπτύξει και να εκτελέσει αυθαίρετο λογισμικό, το οποίο περιλαμβάνει εφαρμογές και λειτουργικά συστήματα. Για ακόμη μια φορά δεν έχει τον έλεγχο και τη διαχείριση της υποδομής cloud, αλλά ελέγχει τα λειτουργικά συστήματα, τον αποθηκευτικό χώρο και τις εφαρμογές, και ενδεχομένως περιορισμένο έλεγχο της επιλογής εξαρτημάτων δικτύωσης (Mell & Grance, 2011). Επίσης, σύμφωνα με τους Mather, Kumaraswamy, και Latif (2009), «ο χρήστης πληρώσει για το ποσό της επεξεργαστικής δύναμης, του αποθηκευτικού χώρου, και ούτω καθεξής που πραγματικά καταναλώνει».

2.5 Τύποι Cloud

Παρόμοια με τις υπηρεσίες SaaS, PaaS, IaaS που προσφέρονται στον τελικό χρήστη, το περιβάλλον «σύννεφου» μπορούμε να το κατηγοριοποιήσουμε ανάλογα με τους τύπους που το συναντάμε.

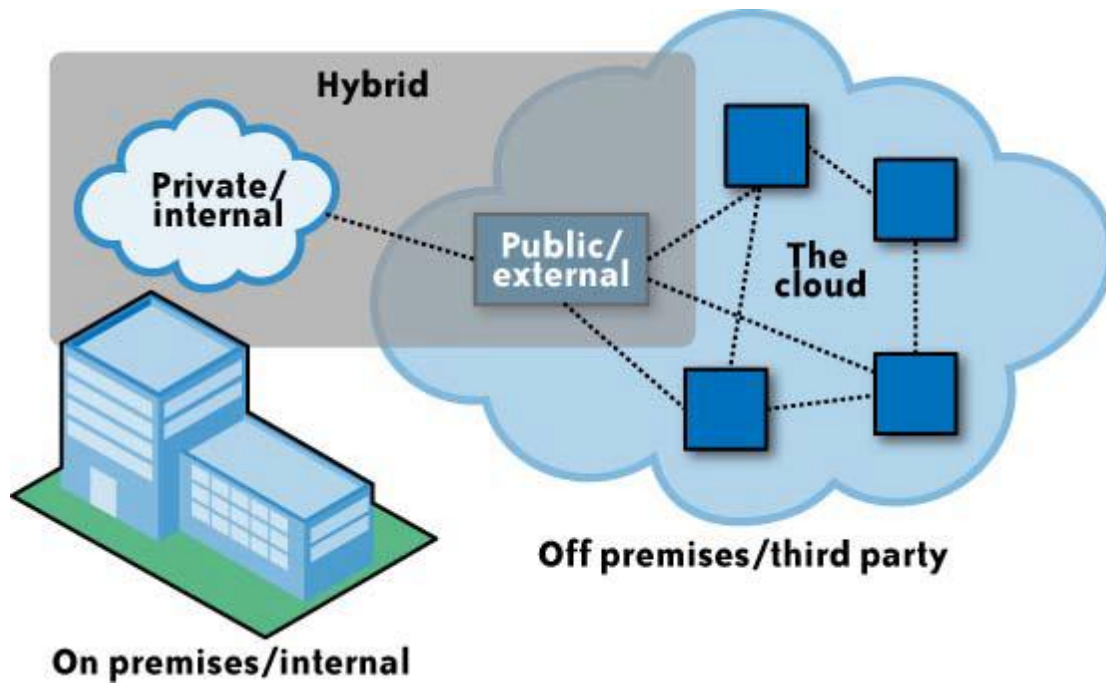
Αρχικά, υπάρχει το ιδιωτικό «σύννεφο» (**Private Cloud**), όπου το περιβάλλον cloud λειτουργεί αποκλειστικά για έναν μόνο οργανισμό που περιλαμβάνει πολλαπλούς χρήστες. Οι λειτουργικότητές του δεν είναι άμεσα εκτεθειμένες, με σκοπό την ιδιωτικότητα και την ασφάλεια. Η διαχείριση και η λειτουργία του μπορεί να ανήκει στον οργανισμό, σε κάποιον

τρίτο, ή σε κάποιο συνδυασμό των δύο (Mell & Grance, 2011), (Danchev, Tsakanikas, & Ventouris, 2011).

Παράλληλα, με το ιδιωτικό «σύννεφο», υπάρχει και το κοινοτικό «σύννεφο» (**Community Cloud**), που αποτελεί στην ουσία μια παραλλαγή του πρώτου. Σύμφωνα με τους Mell και Grance (2011) αυτή η μορφή cloud προσφέρεται για αποκλειστική χρήση από μια συγκεκριμένη κοινότητα, ή από οργανώσεις με κοινές «ανησυχίες». Ανήκει, διαχειρίζεται, και λειτουργεί από έναν ή περισσότερους από τους οργανισμούς της κοινότητας, από τρίτους, ή και κάποιο συνδυασμό τους.

Αντίθετα με τους προηγούμενους δύο τύπους cloud, το δημόσιο cloud (**Public Cloud**) προσφέρεται για ανοικτή χρήση από το ευρύ κοινό. Επιχειρήσεις, ακαδημαϊκά ιδρύματα, κυβερνητικές οργανώσεις κατέχουν, διαχειρίζονται και λειτουργούν ένα δημόσιο «σύννεφο» (Mell & Grance, 2011). Οι Danchev, Tsakanikas, και Ventouris (2011) χαρακτηριστικά αναφέρουν ότι «αυτό το είδος της εγκατάστασης «σύννεφου» απεικονίζεται στην παραδοσιακή γενική έννοια του «Ο καθένας μπορεί να έχει πρόσβαση»».

Ο τελευταίος τύπος «σύννεφου» είναι το υβριδικό cloud (**Hybrid Cloud**), το οποίο αποτελεί στην ουσία μια σύνθεση από δύο ή περισσότερους τύπους «σύννεφου».



Εικόνα 5: Υβριδικό Cloud⁵

2.6 Βασικά Χαρακτηριστικά

Όπως έχουμε αναφέρει το cloud computing είναι ένα μοντέλο το οποίο επιτρέπει την πανταχού και εύκολη πρόσβαση κατόπιν αιτήσεως σε μια πληθώρα υπολογιστικών πόρων. Υπάρχουν πολλά βασικά χαρακτηριστικά που συνθέτουν το μοντέλο του cloud computing (Mell & Grance, 2011).

Το πρώτο χαρακτηριστικό είναι η αυτοεξυπηρέτηση κατόπιν αιτήσεως (on-demand self-service). Ο τελικός χρήστης μπορεί εφοδιαστεί με θεμελιώδεις υπολογιστικές δυνατότητες, όπως τη χρήση ενός server ή αποθηκευτικού χώρου, χωρίς να υπάρχει η απαίτηση για αλληλεπίδραση με τον πάροχο των υπηρεσιών.

⁵ Mather, T., Kumaraswamy, S., & Latif, S. (2009). Cloud Security and Privacy. O'Reilly.

Ένα ακόμη χαρακτηριστικό είναι η ευρεία πρόσβαση δικτύου (**broad network access**).

Αυτό σημαίνει ότι οι υπολογιστικοί πόροι είναι διαθέσιμοι από μηχανήματα που επιτρέπουν την πρόσβαση από κάθε είδους πλατφόρμα, όπως smartphones, tablets, laptops, workstations, κτλ.

Επίσης, οι υπολογιστικοί πόροι που παρέχονται είναι συγκεντρωμένοι με τέτοιο τρόπο, ώστε να εξυπηρετούν καταναλωτές χρησιμοποιώντας ένα μοντέλο πολλαπλής μίσθωσης, με διαφορετικούς φυσικούς και εικονικούς πόρους να ανατίθενται ξανά και ξανά ανάλογα με την ζήτηση του καταναλωτή (**resource pooling**). Υπάρχει γενικά, μια αίσθηση ανεξαρτησίας, δηλαδή ο τελικός χρήστης δεν έχει κανένα έλεγχο ή γνώση για την ακριβή τοποθεσία των πόρων που του παρέχονται.

Παράλληλα, βασικό χαρακτηριστικό αποτελεί και η ταχεία ελαστικότητα (**rapid elasticity**).

Οι υπολογιστικές δυνατότητες που προσφέρονται χαρακτηρίζονται από ελαστικότητα, δηλαδή μπορούν να αναβαθμιστούν γρήγορα και ορισμένες φορές αυτόματα, ανάλογα με την υπάρχουσα ζήτηση. Για τον χρήστη οι διαθέσιμες υπηρεσίες συχνά φαίνεται να είναι απεριόριστες.

Ένα τελευταίο χαρακτηριστικό είναι η μέτρηση της υπηρεσίας (**measured service**), δηλαδή η δυνατότητα παρακολούθησης, ελέγχου και αναφοράς της χρήσης των πόρων, προκειμένου να παρέχεται διαφάνεια τόσο για τον χρήστη, όσο και για τον πάροχο της χρησιμοποιούμενης υπηρεσίας.

Μια υποδομή cloud, είναι η συλλογή του υλικού και λογισμικού που εφαρμόζουν τα παραπάνω βασικά χαρακτηριστικά του cloud computing. Η υποδομή cloud μπορεί να θεωρηθεί ότι περιέχει τόσο φυσικό επίπεδο, όσο και επίπεδο αφαίρεσης. Το φυσικό επίπεδο αποτελείται από πόρους σε υλικό, όπως server, αποθηκευτικός χώρος και συστατικά δικτύου. Το επίπεδο αφαίρεσης αποτελείται από το λογισμικό που αναπτύχθηκε στο φυσικό επίπεδο, το οποίο

εμπεριέχει όλα τα βασικά χαρακτηριστικά του cloud computing. Εννοιολογικά το επίπεδο αφαίρεσης βρίσκεται πάνω από το φυσικό επίπεδο.

2.7 Πλεονεκτήματα και Προκλήσεις

Σύμφωνα με τον Won (2009) το cloud computing προσφέρει μια σειρά από σημαντικά πλεονεκτήματα που το καθιστούν τέλειο παράδειγμα για εξυπηρέτηση σε υπολογιστικές ανάγκες, για κάθε τελικό χρήστη.

Το αρχικό πλεονέκτημα που παρέχει το «σύννεφο» είναι ότι ο πάροχος της υπηρεσίας κατέχει και διαχειρίζεται όλους τους υπολογιστικούς πόρους, όπως servers, λογισμικό, αποθηκευτικό χώρο και δίκτυα, που απαιτούν οι παρεχόμενες υπηρεσίες. Το μόνο που χρειάζεται από την πλευρά του χρήστη είναι να συνδεθεί στο περιβάλλον του cloud.

Επίσης, οι τελικοί χρήστες έχουν τη δυνατότητα να αυξήσουν ή να μειώσουν το επίπεδο χρήσης των υπολογιστικών πόρων, ευέλικτα και εύκολα, ανάλογα με τις απαιτήσεις τους, αντιμετωπίζοντας έτσι περιστασιακές περιόδους αιχμής. Το κόστος της υπηρεσίας είναι πιθανώς πολύ λιγότερο, καθώς πληρώνουν μόνο για τους πόρους και τις υπηρεσίες που πραγματικά χρησιμοποιούν.

Παρά το γεγονός ότι η υπολογιστική «σύννεφου» παρέχει τεράστια πλεονεκτήματα και προσφέρει πολλές λύσεις, υπάρχουν και ορισμένες προκλήσεις που οφείλει να αντιμετωπίσει. Η εταιρεία CITRIX (2010) σε μια από τις αναφορές της, παραθέτει μια σειρά από τέτοιου είδους προκλήσεις.

Ένα βασικό θέμα είναι η ασφάλεια των ευαίσθητων δεδομένων που έχουν ανέβει στο cloud, όπως επίσης και η αξιοπιστία των υπηρεσιών. Ο πάροχος ενδεχομένως να μην είναι αρκετά ώριμος ώστε να εξυπηρετήσει το κοινό στο οποίο απευθύνεται, με αποτέλεσμα να πληγεί η αξιοπιστία του.

Συγχρόνως, είναι δύσκολη η συμμόρφωση της cloud υποδομής, με την ήδη υπάρχουσα IT οργάνωση των χρηστών. Το γεγονός αυτό δυσχεραίνει την αποτελεσματική μετάβαση των εφαρμογών από τα datacenters των επιχειρήσεων-χρηστών στο «σύννεφο». Σημαντική πρόκληση αποτελεί και η έλλειψη διαλειτουργικότητας και συνοχής μεταξύ των διάφορων πάροχων υπηρεσιών cloud, πράγμα που δυσχεραίνει την ενοποίηση. Υπάρχει, γενικά, έλλειψη ώριμων εργαλείων για τη διαχείριση των στοιχείων της υποδομής cloud ή ενδεχόμενες υβριδικές μορφές.

3. Αλγόριθμοι Εξισορρόπησης Φορτίου σε Περιβάλλον «Σύννεφου»

3.1 Εξισορρόπηση Φορτίου σε Περιβάλλον «Σύννεφου»

Όπως έχει γίνει σαφές από το πρώτο κεφάλαιο, η εξισορρόπηση φορτίου αποτελεί μια αρκετά σημαντική διαδικασία και αποτελεί κομμάτι του cloud computing. Η τυχαία άφιξη φορτίου σε ένα τέτοιο περιβάλλον, όπως αυτό του «σύννεφου», μπορεί να προκαλέσει την υπερφόρτωση ορισμένων servers, ενώ κάποιοι άλλοι παραμένουν αδρανείς. Ο αποτελεσματικός προγραμματισμός και η σωστή κατανομή των πόρων είναι ένα κρίσιμο χαρακτηριστικό, με βάση το οποίο εκτιμάται το σύστημα (Kaur, 2012).

Σήμερα, το εύρος ζώνης του δικτύου, μικρότεροι χρόνοι απόκρισης, ελάχιστη καθυστέρηση στη μεταφορά δεδομένων και ελάχιστο κόστος μεταφοράς δεδομένων είναι προκλήσεις του cloud computing όσον αφορά την εξισορρόπηση φορτίου. Η βασική ιδέα για μια τέτοιου είδους πλατφόρμα είναι να έχουμε αρκετούς servers που τρέχουν, αλλά δεν υπερφορτώνονται. Οι διάφορες εικονικές μηχανές (virtual machines) της υποδομής του cloud, οι οποίες ενδεχομένως λειτουργούν και με διαφορετική απόδοση, μοιράζονται ανάλογα με το χρόνο και το φόρτο (Membrey, Plugge, & Hows, 2012). Σύμφωνα με τους Chaczko, Mahadevan, Aslanzadeh, και Mcdermid (2011), οι εξισορροποιητές φορτίου σε ένα περιβάλλον «σύννεφου» εξυπηρετούν δύο σημαντικές ανάγκες, την προώθηση της διαθεσιμότητας των πόρων του «σύννεφου» και τη βελτίωση της απόδοσης.

3.2 Απλοί Αλγόριθμοι Εξισορρόπησης Φορτίου

Το περιβάλλον του cloud computing χρησιμοποιεί ως επί τι πλείστον δυναμικούς αλγόριθμους για την εξισορρόπηση φορτίου, χωρίς να αποκλείονται και οι στατικοί αλγόριθμοι.

Οι δυναμικοί αλγόριθμοι ορίζουν βάρη στους servers και αναζητούν αυτόν με το ελαφρύτερο για να διοχετεύσουν τον φόρτο. Ωστόσο, αυτή η επιπλέον επικοινωνία με τους διακομιστές, επιβαρύνουν το δίκτυο, δημιουργώντας παραπάνω φόρτο. Αντίθετα, οι στατικοί αλγόριθμοι χωρίζουν την κυκλοφορία ισόποσα μεταξύ των διακομιστών. Με την προσέγγιση αυτή όμως, παραβλέπεται ο ήδη υπάρχον φόρτος του server, κατά συνέπεια η κατάσταση γίνεται πιο δυσμενής (Chaczko, Mahadevan, Aslanzadeh, & Mcdermid, 2011). Συγκρίνοντας τις δύο κατηγορίες αλγορίθμων βλέπουμε ότι, οι στατικοί αλγόριθμοι είναι απλοί, αλλά πολλές φορές επιβαρύνουν τους servers. Ενώ, οι δυναμικοί χρησιμοποιούν ερωτήματα προς τους servers για να μην τους επιβαρύνουν, αλλά ορισμένες φορές προσθέτουν επιπλέον φόρτο στο δίκτυο.

Υπάρχουν αρκετοί απλοί αλγόριθμοι, που επιτυγχάνουν την εξισορρόπηση φορτίου, όχι τόσο αποτελεσματικά όσο αυτοί που θα αναλύσουμε στη συνέχεια. Αξίζει, να αναφέρουμε κάποιους βασικούς, όπως ο αλγόριθμος round-robin, ο οποίος διανέμει ισόποσα το φορτίο σε όλους τους servers τους συστήματος. Άλλος αλγόριθμος βασίζεται στον κατακερματισμό, όπως modulo-N κατακερματισμός, το υψηλότερο τυχαίο βάρος κ.α. Επίσης, άλλος αλγόριθμος λαμβάνει υπόψη το φόρτο εργασίας που έχει κάθε server, όπως για παράδειγμα τον αριθμό των συνδέσεων, τη χρήση του επεξεργαστή, τη διαθέσιμη μνήμη κ.α.

Αν θέλουμε να βελτιώσουμε την αξιοποίηση των πόρων, να μεγιστοποιήσουμε την απόδοση, να μειώσουμε τον χρόνο απόκρισης και να αποφύγουμε την υπερφόρτωση, τότε θα χρησιμοποιήσουμε αλγόριθμους που λαμβάνουν υπόψη και τους υπόλοιπους πόρους της υποδομής cloud. Αυτού του είδους οι αλγόριθμοι χωρίζονται σε δύο κατηγορίες, τους προσαρμοστικούς (**adaptive**) και στους μη-προσαρμοστικούς (**non-adaptive**). Οι πρώτοι προσμετράνε τη τρέχουσα κατάσταση του συστήματος προτού κατανέμουν το φόρτο, ενώ οι δεύτεροι δεν λαμβάνουν την τρέχουσα κατάσταση υπόψη, αλλά χρησιμοποιούν κάποιες

ευρετικές μεθόδους (heuristics) για την επιλογή κόμβων (Wen, Hai-ying, Chuang, & Yang, 2011).

3.3 Αλγόριθμος Ισόποσης Διάδοσης Τρέχουσας Εκτέλεσης (ESCE)

Το πρόβλημα μας, όπως σε κάθε περίπτωση, είναι η τυχαία άφιξη φόρτου εργασίας σε ένα περιβάλλον υποδομής cloud που επιβαρύνει τους servers μας και άρα το σύστημα μας.

Παρακάτω αναλύουμε πως λειτουργεί ο Αλγόριθμος Ισόποσης Διάδοσης Τρέχουσας Εκτέλεσης (Equally Spread Current Execution - **ESCE**).

Αρχικά, καθώς καταυθάνουν στο «σύννεφο» μας οι διάφορες εργασίες, στοιβάζονται σε μια ουρά. Ο διαχειριστής του «σύννεφου» εκτιμά το μέγεθος της κάθε εργασίας και ελέγχει για αν υπάρχει διαθέσιμη εικονική μηχανή και τη χωριτικότητα της. Επειτα, αφού βρεθεί η κατάλληλη μηχανή, ο διαχειριστής των εργασιών παρέχει τους απαιτούμενους πόρους στην εργασία που βρίσεται στην ουρά.

Το πλεονέκτημα αυτού του αλγόριθμου είναι ότι προσφέρει μια βελτίωση στο χρόνο απόκρισης και του χρόνου επεξεργασίας, σε σύγκριση με άλλους αλγόριθμους, όπως για παράδειγμα ο round-robin. Οι εργασίες μοιράζονται ισόποσα, το υπολογιστικό μας σύστημα είναι εξισορροπημένο και δεν υπάρχει εικονική μηχανή που να υπολειτουργεί. Λόγω αυτού του πλεονεκτήματος, υπάρχει μια μείωση στο κόστος των εικονικών μηχανών κόστος και το κόστος μεταφοράς δεδομένων (Kaur, 2012).

ACTIVE VM LOAD BALANCER

[START]

Step1:- find the next available VM

Step2:-check for all current allocation count is less than max
length of VM list allocate the VM

Step3:- if available VM is not allocated create a new one in

Step 4:- count the active load on each VM

Step5:- return the id of those VM which is having least load

[END]

Εικόνα 6: Αλγόριθμος ESCE⁶

3.4 Αλγόριθμος Εξισορρόπησης Φορτίου LB3M

Ο αλγόριθμος **LB3M** συνδυάζει στρατηγικές εξισορρόπησης φορτίου με στρατηγικές ελάχιστου χρόνου ολοκλήρωσης. Είναι αποτέλεσμα συνδυασμού ορισμένων απλών αλγορίθμων τους οποίους αναφέρουμε παρακάτω.

Ο Ευκαιριακός Αλγόριθμος Εξισορρόπησης Φορτίου (Opportunistic Load Balancing algorithm - **OLB**) διατηρεί τον κάθε κόμβο απασχολημένο ανεξάρτητα από τρέχοντο φόρτο εργασίας. Αναθέτει τις διεργασίες με τυχαία σειρά. Ο αλγόριθμος Ελάχιστου Χρόνου Ολοκλήρωσης (Minimum Completion Time algorithm - **MCT**), αναθέτει μια εργασία σε εκείνο τον κόμβο που πιστεύει ότι έχει τον αναμενόμενο ελάχιστο χρόνο ολοκλήρωσης. Ο αλγόριθμος προγραμματισμού Min-Min (**MM** algorithm) ακολουθεί την ίδια προσέγγιση με τον MCT. Ο αλγόριθμος εξισορρόπησης φορτίου Min-Min (**LBMM**), από την άλλη πλευρά, υιοθετεί και στρατηγικές εξισορρόπησης, αποφεύγοντας έτσι την διπλή δουλειά. Ο αλγόριθμος Load

⁶ Kaur, J. (2012). Comparison of load balancing algorithms in a Cloud

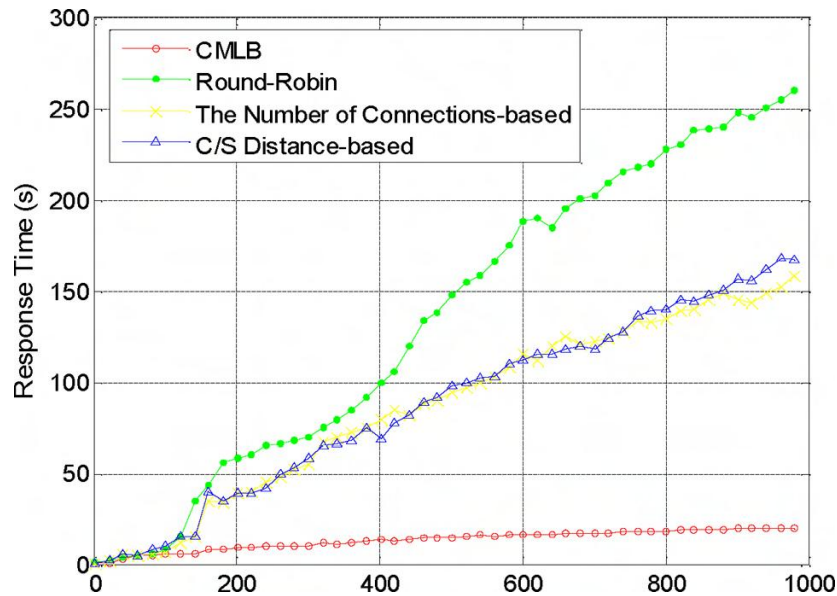
Balance Max-Min and Max (**LB3M**) επιτυγχάνει καλύτερη εξισορρόπηση φορτίου και καλύτερο ελάχιστο χρόνο ολοκλήρωσης σε σχέση με τους MM και LBMM.

Η λειτουργία του αλγορίθμου LB3M έχει ως εξής, αρχικά υπολογίζεται ο μέσος χρόνος ολοκλήρωσης της κάθε εργασίας για κάθε κόμβο και επιλέγεται ο μέγιστος. Στη συνέχεια, εντοπίζεται ο ελεύθερος κόμβος με τον ελάχιστο χρόνο ολοκλήρωσης και του αποστέλλεται η εργασία. Έπειτα, αν δεν υπάρχει ελεύθερος κόμβος, τότε όλοι οι κόμβοι επαναξιολογούνται και υπολογίζονται οι νέοι ελάχιστοι χρόνοι ολοκλήρωσης. Οι χρόνοι αυτοί για μη ελεύθερους κόμβους είναι το άθροισμα του χρόνου ολοκλήρωσης της τρέχουσας εργασίας και της εργασίας που πρόκειται να εκτελεστεί. Αφού υπολογιστούν οι νέοι χρόνοι ανατίθεται η εργασία στον κόμβο με τον ελάχιστο χρόνο ολοκλήρωσης. Τέλος, επαναλαμβάνεται η διαδικασία αυτή μέχρι να τελειώσουν όλες οι εργασίες (Hung, Wang, & Hu, 2011).

Γενικά, ο αλγόριθμος LB3M είναι αρκετά αποδοτικός και επιτυγχάνει καλύτερη εξισορρόπηση φορτίου και απόδοση, συγκριτικά με άλλους αλγορίθμους.

3.5 Αλγόριθμος Εξισορρόπησης Φορτίου CMLB

Ο αλγόριθμος **CMLB** έχει πάρει το όνομα του από τα αρχικά του Cloud-based Multimedia Load Balancing. Έχει αναπτυχθεί για multimedia εφαρμογές σε περιβάλλον cloud. Βασικό του πλεονέκτημα είναι ότι μειώνει τον χρόνο απόκρισης του συστήματος. Επίσης, λαμβάνει υπόψη πλήρως την τρέχουσα κατάσταση του κόμβου, δηλαδή τον υπάρχοντα φόρτο, και κατανέμει έτσι τους πόρους πιο αποτελεσματικά.



Εικόνα 7: Σύγκριση CMLB με άλλους αλγόριθμους με βάση το χρόνο απόκρισης⁷

Η διαδικασία που ακολουθεί είναι στην ουσία μια προσπάθεια για παγκόσμια εξισορρόπηση φορτίου μεταξύ όλων των κόμβων, όταν οι τελικοί χρήστες ζητούν υπηρεσίες πολυμέσων. Αρχικά, μετριοούνται οι τοποθεσίες των clients, όπως επίσης και οι τοποθεσίες των κόμβων. Με βάση τη σειρά ορόσημο που πήραμε για τους πελάτες, επιλέγουμε τους πιο κοντινούς κόμβους σαν υποψήφιους. Υπολογίζουμε τον τρέχοντα φόρτο και την κατάσταση του δικτύου για αυτούς τους κόμβους και επιλέγουμε τους βέλτιστους με βάση το ελάχιστο κόστος. Αυτοί οι κόμβοι προστίθενται σε μια λίστα βέλτιστων κόμβων, όπου αποθηκεύονται οι πελάτες και οι βέλτιστοι κόμβοι. Αν για κάθε πελάτη υπάρχει και ένας κόμβος, τότε μπορεί να του ανατεθεί. Αν υπάρχουν δύο κόμβοι, τότε επιλέγεται ένας από τους δύο. Αφού κατανεμηθούν οι πελάτες στους βέλτιστους κόμβους, τότε διατίθενται και οι servers ανάλογα με τις απαιτήσεις (Wen, Hai-ying, Chuang, & Yang, 2011).

⁷ Wen, H., Hai-ying, Z., Chuang, L., & Yang, Y. (2011). Effective Load Balancing for Cloud-based Multimedia System

Αν και ο συγκεκριμένος αλγόριθμος είναι αρκετά αποτελεσματικός, παρουσιάζει ορισμένα προβλήματα. Συγκεκριμένα, δεν μπορεί να αξιοποιήσει αρκετά αποτελεσματικά τους αυξημένους πόρους του συστήματος, προκειμένου να πετύχει αυξημένο throughput.

3.6 Αλγόριθμος Εξισορρόπησης Φορτίου με Επίγνωση Ενέργειας

Ο αλγόριθμος εξισορρόπησης φορτίου με επίγνωση ενέργειας (Power Aware Load Balancing algorithm - **PALB**) διατηρεί την κατάσταση των κόμβων και ανάλογα με το βαθμό χρήσης αποφασίζει πόσοι κόμβοι πρέπει να λειτουργούν. Επιτυγχάνει τη διάθεση επαρκών υπολογιστικών πόρων, αλλά παράλληλα εξοικονομείται και ενέργεια.

Ο αλγόριθμος PALB χωρίζεται σε τρεις βασικούς τομείς. Ο πρώτος τομέας αφορά την εξισορρόπηση και είναι υπεύθυνος για το που θα πρέπει οι εικονικές μηχανές να αρχικοποιηθούν. Αυτό επιτυγχάνεται αρχικά συγκεντρώνοντας το ποσοστό χρήσης του κάθε ενεργού κόμβου. Αν όλοι οι κόμβοι έχουν βαθμό χρήσης πάνω από 75%, αρχικοποιεί μια καινούρια εικονική μηχανή στον κόμβο με τον χαμηλότερο βαθμό χρήσης. Σε αντίθετη περίπτωση, η νέα εικονική μηχανή εγκαθίσταται στον κόμβο με τον μεγαλύτερο βαθμό χρήσης.

Ο δεύτερος τομέας του αλγορίθμου χρησιμοποιείται με σκοπό την ενεργοποίηση επιπλέον κόμβων. Αυτό το επιτυγχάνει αν όλοι οι ενεργοί κόμβοι έχουν βαθμό χρήσης πάνω από 75%.

Από τη άλλη πλευρά, ο τρίτος τομέας είναι υπεύθυνος για την απενεργοποίηση αδρανών κόμβων. Στην περίπτωση που κάποιος κόμβος χρησιμοποιεί λιγότερο από το 25% των πόρων του, τότε ο αλγόριθμος των απενεργοποιεί (Galloway, Smith, & Vrbsky, 2011).

Η μεγαλύτερη διαφορά όσον αφορά την απόδοση του PALB από τους υπόλοιπους αλγορίθμους εξισορρόπησης φορτίου είναι ότι απενεργοποιεί όσους υπολογιστικούς κόμβους βρίσκονται σε αδράνεια.

3.7 Cloud Load Balancing μια προέκταση του Global Server Load Balancing

Σύμφωνα με την MacVittie (2010), «το Cloud Load Balancing επεκτείνει το αρχιτεκτονικό μοντέλο ανάπτυξης που χρησιμοποιείται, σε συνδυασμό με το Global Server Load Balancing, το οποίο αυξάνει τις διαθέσιμες επιλογές για έναν οργανισμό, όταν καθορίζει από που μια συγκεκριμένη εφαρμογή θα πρέπει να παραδίδεται». Μπορεί να εφαρμοστεί είτε από παραδοσιακά datacenters, είτε μέσω δημόσιων, ιδιωτικών και υβριδικών cloud.

Συνδυάζοντας στοιχεία της τοπικής και της παγκόσμιας εξισορρόπησης φορτίου διακομιστή, η εξισορρόπηση φορτίου «σύννεφου» βοηθάει τους οργανισμούς να διανέμουν τα αιτήματα των χρηστών σε ένα πλήθος εφαρμογών, τοποθετημένες σε πολλαπλά datacenters ή σε υποδομές cloud. Με την επέκταση των αρχικών αρχιτεκτονικών προτύπων της τοπικής και παγκόσμιας εξισορρόπησης φορτίου σε περιβάλλον «σύννεφου», αυξάνονται οι επιλογές για έναν οργανισμό να προσδιορίσει τον τόπο από όπου θα παραδίδεται μια αίτηση.

Ο βασικός στόχος της εξισορρόπησης φορτίου «σύννεφου» είναι να δίνει στον τελικό χρήστη μια εφαρμογή όσο το δυνατόν γρηγορότερα, με τη λιγότερη ποσότητα πόρων, για το χαμηλότερο κόστος. Παράγοντες οι οποίοι πρέπει να ληφθούν υπόψη είναι το κόστος εκτέλεσης μιας αίτησης από μια δεδομένη θέση και το συνολικό κόστος που απαιτείται για την παράδοση της αίτησης αυτής. Επίσης, η συμμόρφωση προς τους κανόνες και η νομικοί περιορισμοί, όπως και ο χρόνος απόκρισης (Nebula 6, 2011).

Η εξισορρόπηση φορτίου «σύννεφου» προκειμένου να καθορίσει την καλύτερη θέση από όπου θα παραδώσει την αίτηση στον χρήστη, χρησιμοποιεί μια παγκόσμια μέθοδο παροχής εφαρμογών. Η διαδικασία επιλογής περιλαμβάνει παράγοντες της παγκόσμιας εξισορρόπησης φορτίου server, όπως ο χρόνος απόκρισης της εφαρμογής, η θέση του χρήστη, η διαθεσιμότητα της εφαρμογής σε μια δεδομένη θέση, η ώρα της ημέρας και η τρέχουσα και συνολική

χωρητικότητα του datacenter ή του cloud περιβάλλοντος στο οποίο έχει αναπτυχθεί η εφαρμογή (MacVittie, 2010).

4. Συμπεράσματα και Προτάσεις για Μελλοντική Έρευνα

Στην παρούσα εργασία αναπτύχθηκαν θέματα που έχουν να κάνουν με την εξισορρόπηση φορτίου εντός ενός δικτύου και πως μπορεί να εφαρμοστεί μέσα σε ένα περιβάλλον «σύννεφου». Το cloud computing αποτελεί, όπως είδαμε, μια νέα τεχνολογία με ραγδαία διάδοση και εξέλιξη. Ο μεγάλος όγκος δεδομένων που διαχειρίζονται, καθιστά αναγκαία την εξισορρόπηση φορτίου εντός της υποδομής του cloud.

Μέχρι στιγμής, έχουν αναπτυχθεί αρκετοί αλγόριθμοι, είτε απλοί, είτε πιο πολύπλοκοι, που επιτυγχάνουν την εξισορρόπηση φορτίου. Κάθε αλγόριθμος ξεχωριστά, υπερτερεί αλλά και συγχρόνως υστερεί έναντι κάποιου άλλου. Το πεδίο εφαρμογής του καθενός είναι λίγο πολύ συγκεκριμένο, οπότε και ταιριάζει καλύτερα σε συγκεκριμένες εφαρμογές.

Ξεχωριστό κομμάτι αποτελεί το Global Server Load Balancing, το οποίο ουσιαστικά είναι η αφετηρία του Cloud Load Balancing, όπου η εξισορρόπηση φορτίου επιτυγχάνεται αξιοποιώντας όλους τους servers ανά τον κόσμο. Στο συγκεκριμένο κομμάτι αρκετές εταιρείες έχουν δώσει βαρύτητα, προκειμένου να επιτευχθεί η επιθυμητή εξισορρόπηση φορτίου.

Συνοψίζοντας, θα μπορούσαμε να πούμε ότι η τομή μεταξύ Cloud Computing και Load Balancing, μπορεί να υπάρξει, η μορφή της όμως εξαρτάται από τις απαιτήσεις και επιθυμίες του κάθε χρήστη ενός περιβάλλοντος Cloud.

Μελλοντικά θα ήταν καλό να γίνουν μελέτες και βελτιώσεις των υπάρχοντων αλγορίθμων, με απώτερο σκοπό την αύξηση της αποδοτικότητας, καθώς ο όγκος των δεδομένων που θα διαχειρίζεται ένα περιβάλλον cloud θα είναι πολύ μεγαλύτερος. Η δημιουργία ίσως ενός πιο γενικού αλγορίθμου, που θα καλύπτει ένα ευρύ φάσμα εφαρμογών, θα ξεκαθαρίσει κάπως το τοπίο για ποιος αλγόριθμος είναι καλύτερος κάθε φορά.

Βιβλιογραφία

- Adler, B. (2010). *Load Balancing in the Cloud: Tools, Tips, and Techniques*. Rightscale. Retrieved from <http://www.rightscale.com/lp/load-balancing-in-the-cloud-whitepaper.php>
- Bobrowski, S. (2010). *Cloud Computing Defined*. Retrieved November 28, 2012, from <http://thecloudview.com/cloud-computing-defined/>
- Chaczko, Z., Mahadevan, V., Aslanzadeh, S., & Mcdermid, C. (2011). Availability and Load Balancing in Cloud Computing. *International Conference on Computer and Software Modeling* (pp. 134-140). Singapore: IACSIT.
- Chee, B. J., & Franklin, C. (2010). *Cloud computing : technologies and strategies of the ubiquitous data center*. New York: CRC.
- CISCO. (2005). *CISCO*. Retrieved October 28, 2012, from http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094820.shtml
- CITRIX. (2010). *Is Your Load Balancer Cloud Ready?* CITRIX. Retrieved from https://www.insight.com/content/dam/insight/en_US/pdfs/citrix/Whitepaper-2-Net-Scaler-Is-Your-Load-Balancer-Cloud-Ready.pdf
- Danchev, S., Tsakanikas, A., & Ventouris, N. (2011). *Cloud Computing: A Driver for Greek Economy Competitiveness*. Αθήνα: I.O.B.E. Retrieved from <http://greece.micnetwork.org/Portals/16/Docs/publications/IOBE%20Cloud%20Computing%20A%20Driver%20for%20Greek%20Economy%20Competitiveness.pdf>
- Galloway, J. M., Smith, K. L., & Vrbsky, S. S. (2011). Power Aware Load Balancing for Cloud Computing. *Proceedings of the World Congress on Engineering and Computer Science. I*, pp. 127-132. San Fransisco: WCECS.
- Hung, C.-L., Wang, H.-h., & Hu, Y.-C. (2011). *Efficient Load Balancing Algorithm for Cloud*. Retrieved from http://onlinepresent.org/proceedings/vol2_2012/80.pdf
- Kaur, J. (2012). Comparison of load balancing algorithms in a Cloud. *International Journal of Engineering Research and Applications, II*(3), 1169-1173. Retrieved from http://www.ijera.com/papers/Vol2_issue3/GV2311691173.pdf
- MacVittie, L. (2010). *Cloud Balancing: The Evolution of Global Server Load Balancing*. F5 Networks. Retrieved from <http://www.f5.com/pdf/white-papers/cloud-balancing-white-paper.pdf>
- Mather, T., Kumaraswamy, S., & Latif, S. (2009). *Cloud Security and Privacy*. Sebastopol: O'Reilly.

- Mell, P., & Grance, T. (2011). *The NIST Definition of Cloud Computing*. N.I.S.T. Retrieved from <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>
- Membrey, P., Plugge, E., & Hows, D. (2012). *Practical Load Balancing: Ride The Performance Tiger*. New York: Apress.
- Nebula6. (2011). *Introduction to Global Load Balancing & Cloud Load Balancing*. Nebula 6. Retrieved from http://www.nebula6.com/downloads/Whitepaper_Global_Load_Balancing_whitepaper.pdf
- Padhy, R. P., & Rao, P. G. (2011). Retrieved from http://ethesis.nitrkl.ac.in/2545/1/load_balancing_in_cloud_computing_systems.pdf
- PC Magazine. (2012). Retrieved October 30, 2012, from http://www.pcmag.com/encyclopedia_term/0,1237,t=virtual+server&i=53950,00.asp
- Radharc-Communications. (2012). *Server Load Balancing*. Radharc Communications. Retrieved from http://www.radharc.com/whitepapers/Server_Load_Balancing.pdf
- Salchow, K. j. (2007). *Load Balancing 101: Nuts and Bolts*. F5 Networks. Retrieved from <http://www.f5.com/pdf/white-papers/load-balancing101-wp.pdf>
- Wen, H., Hai-ying, Z., Chuang, L., & Yang, Y. (2011). Effective Load Balancing for Cloud-based Multimedia System. *International Conference on Electronic & Mechanical Engineering and Information Technology* (pp. 165-168). IEEE.
- Wikipedia. (2012). *Wikipedia*. Retrieved November 1, 2012, from [http://en.wikipedia.org/wiki/Load_balancing_\(computing\)](http://en.wikipedia.org/wiki/Load_balancing_(computing))
- Won, K. (2009). Cloud Computing: Today and Tomorrow. *Journal of Object Technology*, 8(1), 67-72. doi:10.5381/jot.2009.8.1.c4.