ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ ΣΧΕΔΙΑΣΜΟΥ ΔΙΚΤΥΩΝ ΜΕ ΟΠΤΙΚΗ ΔΡΟΜΟΛΟΓΗΣΗ ΣΕ ΚΕΝΤΡΑ ΔΕΔΟΜΕΝΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γρηγόρης Λ. Κανέλλος

Επιβλέπων: Ηρακλής Β. Αβραμόπουλος
Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ ΣΧΕΔΙΑΣΜΟΥ ΔΙΚΤΥΩΝ ΜΕ ΟΠΤΙΚΗ ΔΡΟΜΟΛΟΓΗΣΗ ΣΕ ΚΕΝΤΡΑ ΔΕΔΟΜΕΝΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γρηγόρης Λ. Κανέλλος

Επιβλέπων: Ηρακλής Β. Αβραμόπουλος
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Ιουλίου 2015.

.....	.....	.....
Ηρακλής Αβραμόπουλος	Χρήστος Καψάλης	Νικόλαος Ουζούνογλου
Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2015

.....
Γρηγόρης Λ. Κανέλλος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη:

Οι μεγάλες ροές πληροφορίας στα σύγχρονα κέντρα δεδομένων, η ευρεία χρήση εφαρμογών cloud, τα ραγδαία αυξανόμενα δεδομένα κινητής τηλεφωνίας αλλά και η ανάγκη για αποθήκευση, αναπαραγωγή και αποδοτική επεξεργασία όλων των παραπάνω δεδομένων έχει οδηγήσει σε μια κατάσταση όπου οι ροές πληροφορίας και τα κλασσικά μοτίβα της κίνησης της πληροφορίας σε ένα κέντρο δεδομένων αλλάζουνε δυναμικά. Παρατηρείται μία ραγδαία αύξηση στην κίνηση των κέντρων δεδομένων και είναι πιθανό ότι η τεχνολογία και η υποδομή των σημερινών αρχιτεκτονικών δικτύου στα κέντρα δεν θα μπορεί να ανταπεξέλθει στις ολοένα αυξανόμενες απαιτήσεις της κίνησης. Συνεπώς υπάρχει ανάγκη για εξέλιξη και αναβάθμιση των κέντρων δεδομένων με νέες αρχιτεκτονικές βασισμένες σε οπτική τεχνολογία έτσι ώστε να βελτιωθούν σημαντικά οι επιδόσεις των κέντρων δεδομένων και να αναπτυχθούν έτσι ώστε να υποστηρίζουν την ραγδαία αυτή αύξηση της κίνησης στα κέντρα δεδομένων.

Αντικείμενο της εργασίας αυτής είναι αρχικά η παρουσίαση της μορφής των σημερινών κέντρων δεδομένων και η ανάλυση των παραγόντων που εμποδίζουν την περαιτέρω επέκταση τους. Στην συνέχεια παρουσιάζονται κάποιες κύριες υλοποιήσεις οπτικών αρχιτεκτονικών βασισμένες σε διαφορετικές τεχνολογίες και τέλος αναφέρονται τα αναγκαία πρωτόκολλα που αφορούν το πεδίο ελέγχου των οπτικών αυτών αρχιτεκτονικών.

Abstract:

Big data flows along with server virtualization at current data centers, cloud applications, mobile data, and the need to store and replicate vast amounts of data has led to a situation where there are large dynamically changing data traffic patterns and flows across modern datacenters. There is a rapid increase in data traffic and it is likely that the technology and infrastructure of today's architecture in data centers will not be able to cope with the growing traffic demands. Therefore there is a need for disruptive development and migration to new architectures based on optical technology to significantly improve the performance of data centers and to support this rapid increase in traffic in data centers.

The purpose of this thesis is firstly the presentation of the form of today's data centers and the investigation of the factors that hinder their further scaling so as to meet the enormous needs of increasing traffic in data centers. Then we present some major implementations of optical architectures based on different technologies aimed for data centers and lastly we present the overarching protocols concerning the control plane of these optical architectures.

Δομή εργασίας:

Στο πρώτο κεφάλαιο της παρούσας εργασίας, παρουσιάζονται μερικές βασικές έννοιες και ορισμοί των δικτύων και των βασικών δομών με τις οποίες υλοποιούνται. Επίσης γίνεται και η παρουσίαση των βασικότερων μετρικών απόδοσης τα οποία καθορίζουν τη λειτουργία του δικτύου.

Στο δεύτερο κεφάλαιο παρουσιάζεται αναλυτικά η παρούσα κατάσταση και δομή των κέντρων δεδομένων σήμερα και αναλύεται διεξοδικά το γιατί οι δομές αυτές δεν μπορούν να επεκταθούν και να ανταποκριθούν στις μελλοντικές απαιτήσεις των δικτύων, με την οπτική τεχνολογία να παρουσιάζεται ως μία υποσχόμενη λύση.

Στο τρίτο κεφάλαιο γίνεται αναλυτικά η παρουσίαση των κυρίων αρχιτεκτονικών οι οποίες κάνουν χρήση οπτικών τεχνολογιών και αποτελούν πολλά υποσχόμενες αρχιτεκτονικές. Δίνεται έμφαση στα εξής:Plexxi, AWG-based, 3D-MEMS και WSS-based αρχιτεκτονικές.

Στο τέταρτο κεφάλαιο της εργασίας αναφερόμαστε στην ανάπτυξη του cloud computing και στην πρόσφατη ανάπτυξη των SDN αρχιτεκτονικών και με ποιους τρόπους μπορεί να ελεχθεί ορθά και αποτελεσματικά μία αρχιτεκτονική. Επίσης παρουσιάζονται μερικά πρωτόκολλα κίνησης και τέλος αναφέρονται μερικές μελλοντικές ερευνητικές διεργασίες.

Στο πέμπτο και τελευταίο κεφάλαιο αναφέρονται τα συμπεράσματα που προκύπτουν από τη παρούσα διπλωματική εργασία.

Λέξεις Κλειδία:
Data Centers
Οπτικές Αρχιτεκτονικές Δομές
Switches
Plexxi
AWG
3D-MEMS
WSS
SDN
Open Flow

Θα ήθελα να εκφράσω τις θερμότερες ευχαριστίες μου στον καθηγητή και επιβλέποντα της συγκεκριμένης εργασίας κ. Ηρακλή Αβραμόπουλο για την άριστη συνεργασία μας, την πολύτιμη βοήθεια του και τις γνώσεις που μου πρόσφερε κατά τη διάρκεια των σπουδών μου στη σχολή. Επίσης, θα ήθελα πολύ να ευχαριστήσω τον ερευνητή κ. Μπακόπουλο Παρασκευά για τις πολύτιμες υποδείξεις και την ουσιαστική βοήθεια του στην εκπόνηση της παρούσας εργασίας.

Περιεχόμενα

Εισαγωγή

Η παρούσα εργασία έχει ως σκοπό τη μελέτη των τεχνολογιών και αρχιτεκτονικών δομών που απαρτίζουν τα σημερινά κέντρα δεδομένων αλλά κυρίως δίνεται έμφαση στις μελλοντικές δομές. Πιο συγκεκριμένα δίνεται έμφαση σε αρχιτεκτονικές δομές οι οποίες κάνουν χρήση οπτικών στοιχείων και τεχνολογιών έτσι ώστε να εξυπηρετήσουν καλύτερα τις ολοένα και αυξανόμενες ανάγκες και εφαρμογές του κάθε κέντρου δεδομένων.

Ένα κέντρο δεδομένων είναι ένας συγκεντρωτικός χώρος αποθήκευσης, είτε φυσικός ή εικονικός (**virtual**), για την αποθήκευση, επεξεργασία και διάδοση των δεδομένων και των πληροφοριών που οργανώνονται γύρω από ένα συγκεκριμένο αντικείμενο γνώσεων και πληροφοριών ή σχετίζονται με τα δεδομένα και τις εφαρμογές μιας συγκεκριμένης επιχείρησης. Ως φυσικός χώρος αποθήκευσης ορίζεται ένα σύνολο από διακομιστές (**servers**) και υπολογιστές οργανωμένους σε μία φυσική διάταξη ενώ ένα virtual κέντρο δεδομένων είναι ένα εικονικό μοντέλο υπηρεσιών **cloud** εφαρμογών τα δεδομένα του οποίου έχουν αποθηκευτεί στο διαδίκτυο. Φυσικά και στη περίπτωση του cloud τα δεδομένα διακινούνται και διαχειρίζονται διακομιστές οι οποίοι λόγω της virtual φύσης του cloud δεν στεγάζονται απαραίτητα στον ίδιο χώρο. Τα κέντρα δεδομένων, είτε φυσικής ή εικονικής υποδομής, χρησιμοποιούνται από τις επιχειρήσεις για να στεγάσουν τους διακομιστές, το σύνολο των υπολογιστών και τα συστήματα δικτύωσης για τις ανάγκες της κάθε επιχείρησης οι οποίες συνήθως περιλαμβάνουν την αποθήκευση, επεξεργασία και διάθεση μεγάλων ποσοτήτων κρίσιμων δεδομένων σε πελάτες.

Οι ειδικοί διαχωρίζουν τα κέντρα δεδομένων σε κατηγορίες ανάλογα με τις λειτουργίες που εκτελούν και τις υπηρεσίες που παρέχουν [1]. Υπάρχουν πάρα πολλά κριτήρια σύμφωνα με τα οποία ταξινομούνται τα κέντρα δεδομένων και στη συνέχεια θα αναφέρουμε μερικά μοντέλα σύμφωνα με την χωρική οργάνωση. Η παροχή ανοιχτού cloud – **public cloud providers** είναι κέντρα δεδομένων που προφέρουν ανοιχτά και ελεύθερα τις υπηρεσίες τους στο κοινό, χαρακτηριστικό παράδειγμα οι μεταμηχανές διαδικτυακής αναζήτησης. Τα **ιδιωτικά in-home κέντρα δεδομένων** στη συνέχεια είναι το ακριβώς αντίθετο, δηλαδή εγκαταστάσεις που ανήκουν και λειτουργούν από μία εταιρία ή επιχείρηση η οποία διαχειρίζεται αποκλειστικά τους διακομιστές. Άξια αναφοράς είναι και τα επιστημονικά κέντρα υπολογιστών – **scientific computing centers** αλλά και τα **co-location datacenters** όπου τα τελευταία αποτελούν μία ιδιωτική διαδικτυακή εφαρμογή cloud [1]. Ένας άλλος διαχωρισμός είναι με βάση τις λειτουργίες τους, σε **commercial datacenters** – διαθέσιμα κέντρα δεδομένων όπως είναι δηλαδή τα public cloud providers για αναζητήσεις στο διαδίκτυο και σε **high performance computing (HPC)** κέντρα δεδομένων δηλαδή υπολογιστικά συστήματα υψηλών επιδόσεων τα οποία χρησιμοποιούνται από επιχειρήσεις ή ερευντηκά κέντρα τα οποία έχουν ανάγκη την συνεχή, σταθερή και γρήγορη ροή και επεξεργασία πληροφοριών και δεδομένων [1].

Προφανώς ένα κέντρο δεδομένων έχει κάποιες προδιαγραφές ή απαιτήσεις για την αποτελεσματική λειτουργία του και κάθε σχεδιαστής δικτύου πρέπει να λαμβάνει υπόψη το είδος του κέντρου δεδομένου καθώς κάθε κέντρο θα έχει διαφορετικές απαιτήσεις ανάλογα με την υποδομή του, τις λειτουργίες που θα εκτελεί και το είδος των εφαρμογών που θα εξυπηρετεί. Στην συνέχεια θα αναφερθούμε στα γενικά κοινά χαρακτηριστικά που απαιτούν όλα τα κέντρα ανεξαρτήτως λειτουργιών. Ένα από τα κυριότερα προβλήματα είναι η διασυνέχεια της επιχειρηματικής δραστηριότητας. Οι εταιρείες βασίζονται στα πληροφοριακά τους συστήματα για να διευθύνουν τις επιχειρήσεις τους. Εάν ένα σύστημα δεν είναι πλέον διαθέσιμο, οι εργασίες της εταιρείας μπορεί να εξασθενήσουν ή να σταματήσουν και εντελώς. Είναι αναγκαίο

να προβλεφθεί μια αξιόπιστη υποδομή για τις επιχειρήσεις πληροφορικής, προκειμένου να ελαχιστοποιηθεί οποιαδήποτε πιθανότητα κατάρευσσης του συστήματος. Η ασφάλεια των πληροφοριών είναι επίσης μια ανησυχία, και για το λόγο αυτό ένα κέντρο δεδομένων επιβάλλεται να προσφέρει ένα ασφαλές περιβάλλον που να ελαχιστοποιεί τις πιθανότητες παραβίασης της ασφάλειας. Ένα κέντρο δεδομένων θα πρέπει, επομένως, να κρατήσει υψηλά πρότυπα για την εξασφάλιση της ακεραιότητας και της λειτουργικότητας των εφαρμογών του. Επιπλέον στα κέντρα δεδομένων είναι απαραίτητες οι προσθήκες πολλών μηχανικών συστημάτων ψύξης, πράγμα σημαντικό καθώς οι υπολογιστές δεν σταματούν να λειτουργούν και υπάρχει μεγάλος κίνδυνος υπερθέρμανσης. Όλοι οι σχεδιαστές εφοδιάζουν τις υποδομές πάντα με περισσότερους μηχανισμούς ψύξης από ότι είναι απαραίτητο, ως πλεονασμό για την ελαχιστοποίηση της πιθανότητας υπερθέρμανσης, ακόμα και μετά από επέκταση του κέντρου δεδομένων με περισσότερους διακομιστές. Μία άλλη σημαντική προδιαγραφή είναι η παροχή της κατάλληλης ηλεκτρικής ενέργειας για τη λειτουργία του κέντρου το οποίο συνήθως έχει μεγάλες απαιτήσεις ηλεκτρικής ενέργειας. Επίσης σε ένα κέντρο δεδομένων πρέπει η υποδομή των υπολογιστικών πόρων του να σχεδιάζεται με τέτοιο τρόπο ώστε να είναι εύελικτο και εύκολα επεμβάσιμο είτε για πιθανή επέκταση του είτε για αντικατάσταση/αναβάθμιση διακομιστών.

Τα κυριότερα πράγματα που απασχολούν τα σημερινά κέντρα δεδομένων είναι η ραγδαία αύξηση της κίνησης και μεταφοράς δεδομένων και το γεγονός ότι στο άμεσο μέλλον οι περισσότερες εφαρμογές θα λαμβάνουν μέρος σε cloud υλοποιήσεις και ότι τα κέντρα δεδομένων θα μετατραπούν στις virtual εκδοχές τους. Σύμφωνα με μετρήσεις και προβλέψεις ειδικών η κίνηση (**traffic**) και η μεταφορά δεδομένων μέσα στα κέντρα δεδομένων είναι μεγάλη, της τάξεως των 3,1 zettabytes ανά χρόνο με τις μετρήσεις αυτές να αναφέρονται το 2013. Υπολογίζεται ότι μέχρι το 2018 η κίνηση θα έχει φτάσει στα 8,6 zettabytes ανά χρόνο, δηλαδή θα έχουμε σχεδόν τριπλασιασμό της κινητικότητας και μεταφοράς δεδομένων στα δίκτυα μέσα σε 5 χρόνια [1]. Υπολογίζεται ακόμη ότι με την ανάπτυξη των cloud εφαρμογών το 78% των εφαρμογών θα διαχειρίζεται από virtual κέντρα δεδομένων και μόνο το 22% των εφαρμογών θα διαχειρίζεται από τα παραδοσιακά κέντρα [1]. Αυτό συμβαίνει καθώς το virtualization των κέντρων δεδομένων επιτρέπει τη δυναμική ανάπτυξη και επεξεργασία των εφαρμογών στο cloud για την αντιμετώπιση των ολοένα αυξανόμενων απαιτήσεων των υπηρεσιών cloud, ενώ το παραδοσιακό datacenter δεν θα μπορεί να ανταπεξέλθει. Ακόμη το μεγαλύτερο κομμάτι της κίνησης θα παραμένει μέσα στο κέντρο δεδομένων, αναμένεται ότι περισσότερο από τα 2/3 της συνολικής κίνησης (περίπου το 76%) θα παραμένει μέσα στο κέντρο δεδομένων παρά θα χρησιμοποιείται για την επικοινωνία με εξωτερικούς χρήστες ή μεταξύ των κέντρων. Επίσης η κίνηση αυτή θα έχει east-west προσανατολισμό, δηλαδή θα ανταλλάσουν στοιχεία και δεδομένα οι servers και οι hosts μεταξύ τους και όχι τα επίπεδα- layers του δικτύου. Η ραγδαία αύξηση του όγκου των δεδομένων και της κίνησης δημιουργεί υπέρμετρες απαιτήσεις στον εξοπλισμό δρομολόγησης, έτσι ώστε ήδη διαφαίνονται τα όρια των τεχνολογιών που χρησιμοποιούνται στα σύγχρονα δίκτυα κέντρων δεδομένων. Ο προσανατολισμός της κίνησης (east – west) αναδεικνύει την αδυναμία των σύγχρονων αρχιτεκτονικών δικτύου που είναι σχεδιασμένες και βελτιστοποιημένες για north-south κίνηση, περιορίζοντας την απόδοση του δικτύου και αυξάνοντας την καθυστέρηση, που είναι καθοριστική για αρκετές σύγχρονες εφαρμογές που τρέχουν σε δίκτυα δεδομένων. Συμπερασματικά από τα παραπάνω, λίγα από τα σημερινά δίκτυα ηλεκτρικής μεταγωγής πακέτων θα μπορούν να ανταπεξέλθουν στις μελλοντικές ανάγκες της κίνησης στα δίκτυα και πιθανώς οι σχεδιαστές δικτύου θα πρέπει να στραφούν σε άλλου είδους τεχνολογίες οι οποίες θα καθιστούν ικανές τις παραπάνω απαιτήσεις [1].

Κεφάλαιο 1
Τοπολογίες Δικτύων και Μετρικά Απόδοσης

Στην αρχική ενότητα της εργασίας παρουσιάζονται και γνωστοποιούνται βασικές και γενικές αρχές περί των δικτύων και των τοπολογιών τους. Αρχικά γίνεται η παρουσίαση των βασικών μορφών και τοπολογιών στις οποίες είναι βασισμένα τα περισσότερα δίκτυα, είτε αυτά είναι δίκτυα τηλεφωνίας, είτε ακόμα και εγκατεστημένα δίκτυα στα σύγχρονα κέντρα δεδομένων. Στην συνέχεια, παρατίθονται μερικές τοπολογίες μεταγωγής δικτύου (**switching fabric topologies**) με τα βασικότερα τους στοιχεία. Τέλος παρουσιάζουμε τα μετρικά απόδοσης ενός δικτύου, δηλαδή από τι μεγέθη καθορίζεται ένα ορθά λειτουργικό δίκτυο αλλά και το πως υπολογίζονται τα κυριότερα από αυτά.

1.1 Βασικές τοπολογίες δικτύων

Υπάρχουν δύο κατηγορίες τοπολογίας δικτύων: φυσικές τοπολογίες (**physical topologies**) και λογικές τοπολογίες (**logical topologies**) [2].

Η διάταξη καλωδίωσης που χρησιμοποιείται για τη σύνδεση συσκευών είναι η φυσική τοπολογία του δικτύου. Αυτό αναφέρεται στην διάταξη της καλωδίωσης στο χώρο, τις θέσεις των κόμβων, καθώς και τις διασυνδέσεις μεταξύ των κόμβων και της καλωδίωσης. Η φυσική τοπολογία του δικτύου καθορίζεται από τις δυνατότητες των συσκευών πρόσβασης στο δίκτυο και τα τεχνολογικά μέσα, το επίπεδο ελέγχου, την ανοχή σε σφάλματα και το κόστος που σχετίζεται με την καλωδίωση ή τα τηλεπικοινωνιακά κυκλώματα [2].

Η λογική τοπολογία αντίθετα, είναι ο τρόπος που τα σήματα ενεργούν στο μέσο δικτύου ή ο τρόπος με τον οποίο τα δεδομένα διέρχονται διαμέσου του δικτύου από τη μία συσκευή στην άλλη χωρίς να λαμβάνεται υπόψη η φυσική διασύνδεση των συσκευών. Η λογική τοπολογία ενός δικτύου δεν είναι αναγκαστικά ίδια με τη φυσική τοπολογία του. Η λογική κατάταξη των τοπολογίων δικτύου ακολουθεί γενικά τις ίδιες ταξινομήσεις με αυτές των φυσικών ταξινομήσεων των τοπολογιών δικτύου, αλλά περιγράφει τη διαδρομή που τα δεδομένα κάνουνε μεταξύ των κόμβων που χρησιμοποιούνται σε αντίθεση με τις πραγματικές φυσικές συνδέσεις μεταξύ των κόμβων [2].

Στο σημείο αυτό, αφού εξηγήθηκε ο διαχωρισμός της φυσικής και λογικής τοπολογίας, να αναφερθούμε στα διάφορα βασικά είδη των τοπολογιών δικτύου. Η μελέτη των δικτύων αναγνωρίζει επτά βασικές τοπολογίες: σημείο-προς-σημείο – **point-to-point**, διάυλου – **bus**, αστέρα – **star**, δακτυλίου – **ring**, κατανεμημένη – **mesh,** δένδρου – **tree** και **υβριδικούς συνδυασμούς** [3].

Point – to – point

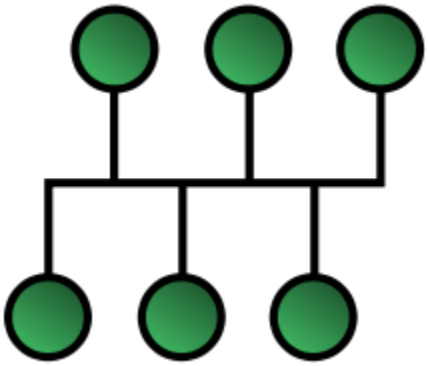
Πρόκειται για την απλούστερη τοπολογία με μια μόνιμη σύνδεση μεταξύ δύο τελικών σημείων. Οι τοπολογίες μεταγωγής point-to-point είναι το βασικό μοντέλο της συμβατικής τηλεφωνίας. Η αξία ενός μόνιμου δικτύου σημείο-προς-σημείο είναι ανεμπόδιστη επικοινωνία μεταξύ των δύο τελικών σημείων ή κόμβων. Όσο μεγαλύτερος είναι ο αριθμός των πιθανών

ζευγών των κόμβων μίας σημείου-προς-σημείο σύνδεσης κατ' αίτηση τόσο πιο σημαντική είναι η σύνδεση αυτή για το δίκτυο [3].

Τοπολογία Bus

Σε τοπικά δίκτυα όπου χρησιμοποιείται τοπολογία διαύλου, κάθε κόμβος συνδέεται με ένα μόνο καλώδιο. Κάθε υπολογιστής ή server είναι συνδεδεμένος με το ενιαίο καλώδιο διαύλου. Ένα σήμα από την πηγή ταξιδεύει σε δύο διευθύνσεις σε όλες τις μηχανές που συνδέονται στο καλώδιο διαύλου μέχρι να βρει τον αποδέκτη. Εάν η διεύθυνση μηχάνημα δεν ταιριάζει με τη διεύθυνση που προορίζονται για τα δεδομένα, η μηχανή αγνοεί τα δεδομένα. Εναλλακτικά, εάν τα δεδομένα ταιριάζουνε με τη διεύθυνση του μηχανήματος, τα δεδομένα είναι αποδεκτά. Επειδή η τοπολογία διαύλου αποτελείται από ένα μόνο καλώδιο, είναι μάλλον φθηνή για την εφαρμογή σε σύγκριση με άλλες τοπολογίες. Ωστόσο, το χαμηλό κόστος της εφαρμογής της τεχνολογίας αντισταθμίζεται από το υψηλό κόστος διαχείρισης του δικτύου. Επιπλέον, επειδή χρησιμοποιείται μόνο ένα καλώδιο, μπορεί να είναι το μοναδικό σημείο αποτυχίας και κατάρρευσης του όλου συστήματος. Γνωστό παράδειγμα της συγκεκριμένης τοπολογίας είναι το πρότυπο IEEE 802.3 για τα LAN. Το IEEE 802.3 περιγράφει ένα δίκτυο εκπομπής, που χρησιμοποιεί ως κοινό κανάλι επικοινωνίας ένα καλώδιο Ethernet, το οποίο προσφέρει ταχύτητες από 1Mbps έως και 100 Gbps.

Θα πρέπει να επισημανθεί ωστόσο, πως αν συμβεί κάποια σύγκρουση στο δίκτυο που έχει ως αποτέλεσμα την απώλεια κάποιου πακέτου, επειδή έτυχε δύο ή περισσότεροι υπολογιστές να ξεκινήσουν την μετάδοσή τους την ίδια χρονική στιγμή, τότε οι υπολογιστές που έστειλαν τα πακέτα απλά περιμένουν ένα σύντομο τυχαίο χρονικό διάστημα και ξαναδοκιμάζουν να στείλουν τα πακέτα [3].

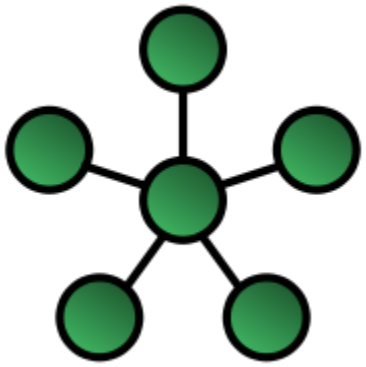


Εικόνα 1.1 – Τοπολογία Bus

Τοπολογία Star

Στα δίκτυα με τοπολογία αστέρα κάθε διακομιστής-**server** δικτύου συνδέεται σε ένα κεντρικό σταθμό-**hub** με μία σύνδεση σημείου-προς-σημείο. Στη Star τοπολογία κάθε κόμβος συνδέεται

με ένα κεντρικό κόμβο που ονομάζεται διανομέας ή μεταγωγέας. Το δίκτυο δεν πρέπει αναγκαστικά να μοιάζει με ένα αστέρι για να χαρακτηριστεί ως ένα δίκτυο αστέρα, αλλά όλοι οι κόμβοι του δικτύου πρέπει να είναι συνδεδεμένοι με μία κεντρική συσκευή. Όλη η κίνηση που διασχίζει το δίκτυο περνά μέσα από τον κεντρικό κόμβο. Το κέντρο λειτουργεί ως αναμεταδότης σήματος. Η τοπολογία αστέρα θεωρείται η ευκολότερη τοπολογία για να σχεδιαστεί και να εφαρμοστεί μετά την point to point. Ένα πλεονέκτημα της τοπολογίας αστέρα είναι η απλότητα της προσθήκης επιπρόσθετων κόμβων. Το κύριο μειονέκτημα της τοπολογίας αστέρα είναι ότι και εδώ ο κεντρικός κόμβος αντιπροσωπεύει ένα μοναδικό σημείο ολικής αποτυχίας [3].



Εικόνα 1.2 – Τοπολογία Star

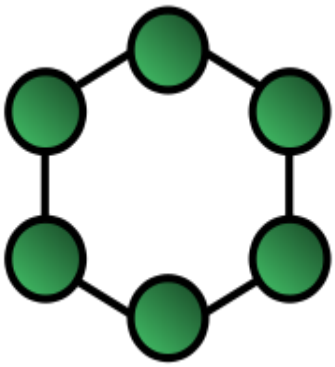
Τοπολογία Ring

Σε αντίθεση με τα δίκτυα που χρησιμοποιούν ένα κοινό δίαυλο επικοινωνίας, οι κόμβοι των δικτύων δακτυλίου συνδέονται μεταξύ τους. Ο κάθε κόμβος, που απαρτίζει το δίκτυο, ενώνεται ακριβώς με δύο διαφορετικούς κόμβους του δικτύου, σχηματίζοντας έτσι μία συνεχόμενη διαδρομή που διέρχεται από όλο το δίκτυο. Έτσι ο δακτύλιος είναι μια τοπολογία του δικτύου που έχει συσταθεί σε ένα κυκλικό τρόπο με τον οποίο τα δεδομένα ταξιδεύουν γύρω από τον δακτύλιο σε μία κατεύθυνση και κάθε συσκευή στο δακτύλιο λειτουργεί ως αναμεταδότης για να κρατήσει το σήμα ισχυρό καθώς ταξιδεύει. Κάθε συσκευή ενσωματώνει ένα δέκτη για το εισερχόμενο σήμα και ένα πομπό για την αποστολή των δεδομένων στην επόμενη συσκευή στον δακτύλιο. Το δίκτυο εξαρτάται από την ικανότητα του σήματος να ταξιδέψει γύρω από το δακτύλιο. Όταν μια συσκευή στέλνει δεδομένα, θα πρέπει να ταξιδέψουν μέσα από κάθε συσκευή στο δακτύλιο μέχρι να φτάσουν στον προορισμό τους. Κάθε κόμβος είναι ένα κρίσιμο σημείο, δηλαδή όλοι οι κόμβοι λειτουργούν ως διακομιστές στην τοπολογία ring και πρέπει να προωθήσουνε το σήμα.

Τα δίκτυα δακτυλίου διαθέτουν αρκετά μειονεκτήματα αλλά και πλεονεκτήματα έναντι των δικτύων, που χρησιμοποιούν ένα κοινό δίαυλο. Συγκεκριμένα, επειδή τα δίκτυα δακτυλίων στην πραγματικότητα συνίστανται από πολλές point to point συνδέσεις, μας είναι γενικά πιο εύκολο να απομονώσουμε και να αναγνωρίσουμε τυχόν προβλήματα στο δίκτυο. Περαιτέρω, σε περίπτωση πολλής κίνησης και γενικότερα υψηλού φόρτου στο δίκτυο, έχει παρατηρηθεί πως τα δίκτυα δακτυλίων αποδίδουν καλύτερα, εξαιτίας κανόνων διαιτησίας που εφαρμόζονται από όλους τους κόμβους του δικτύου. Επίσης, σε αντίθεση με την αρχιτεκτονική κοινού διαύλου, όπου ο κάθε κόμβος μπορεί να ξεκινήσει τη μετάδοση ενός πακέτου οποιαδήποτε χρονική

στιγμή, στα δίκτυα δακτυλίου εφαρμόζονται διάφορες μέθοδοι που καθορίζουν την σειρά που θα μεταδίδουν οι κόμβοι.

Το μειονέκτημα αυτής της τοπολογίας είναι ότι αν ένας κόμβος σταματήσει να λειτουργεί, όλο το δίκτυο μπορεί να επηρεαστεί ή ακόμα και να σταματήσει να λειτουργεί. Επίσης, η επεκτασιμότητα του δικτύου, δηλαδή η προσθήκη νέων κόμβων μέσα στο δίκτυο, είναι δυσκολότερη, καθότι η εμφάνιση ενός νέου κόμβου θα σήμαινε την αναθεώρηση όλων των μηχανισμών αστυνόμευσης και κανόνων διαιτησίας, που έχουν καθολική ισχύ σε όλο το εύρος του δικτύου [3].



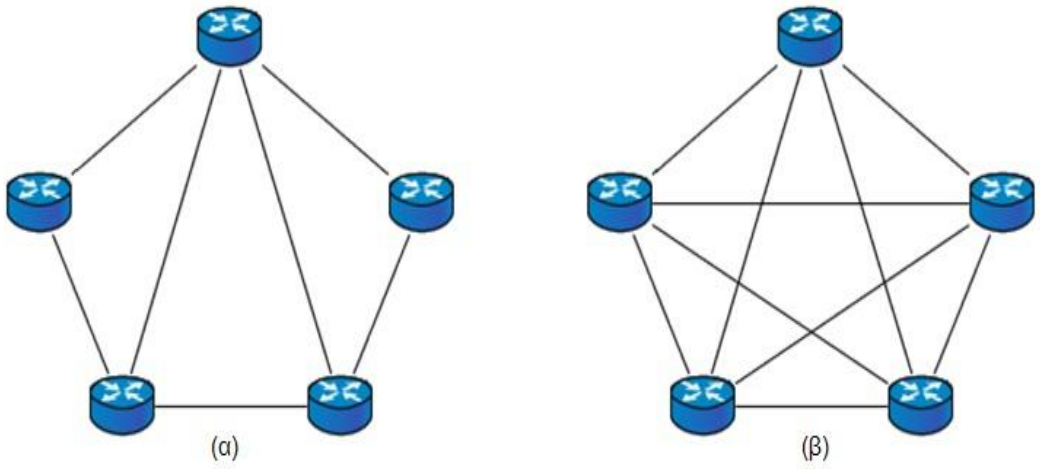
Εικόνα 1.3 – Τοπολογία Ring

Τοπολογία mesh

Ως κατανεμημένη τοπολογία ορίζουμε τη διάταξη των οντοτήτων ενός δικτύου, έτσι ώστε η κάθε οντότητα να μη χρησιμοποιείται απλά για τη μεταφορά των δικών της δεδομένων, αλλά να λειτουργεί και σαν μεσάζοντας, προωθώντας δεδομένα που ανταλλάσσονται μεταξύ άλλων οντοτήτων του δικτύου. Η πληροφορία μέσα σε ένα δίκτυο, που ακολουθεί τις αρχιτεκτονικές προδιαγραφές της κατανεμημένης τοπολογίας, μεταφέρεται με την μορφή πακέτων. Για να προωθηθούν τα πακέτα προς τον κατάλληλο προορισμό, γίνεται χρήση είτε αλγορίθμων πλημμύρας είτε παραδοσιακών αλγορίθμων δρομολόγησης. Οι αλγόριθμοι δρομολόγησης, που έχουν εφαρμογή στα εν λόγω δίκτυα, καθορίζουν μία διαδρομή που ακολουθούν τα πακέτα, με άλματα από κόμβο σε κόμβο μέχρι να επιτευχθεί ο τελικός προορισμός. Αντίθετα, όταν γίνεται εφαρμογή του αλγόριθμου πλημμύρας, ο κάθε κόμβος του δικτύου προωθεί τα εισερχόμενα πακέτα προς κάθε δυνατό προορισμό, με αποτέλεσμα τα πακέτα κάποια στιγμή να φτάσουν στον τελικό τους προορισμό. Προφανώς, ο αλγόριθμος πλημμύρας δεν αποτελεί τη βέλτιστη λύση για τη δρομολόγηση των πακέτων, ωστόσο χρησιμοποιείται σε πολύ συγκεκριμένες περιπτώσεις εξαιτίας της απλότητάς του.

Στην παρακάτω εικόνα διαφαίνονται οι δύο διαφορετικές εκδοχές μίας κατανεμημένης τοπολογίας. Συγκεκριμένα, η εικόνα α παρουσιάζει μια **μερικώς κατανεμημένη (partially mesh)** τοπολογία ενός δικτύου, ενώ η εικόνα β παρουσιάζει μία **πλήρως κατανεμημένη (fully mesh)** τοπολογία. Η διαφορά μεταξύ των τοπολογιών έγκειται στο γεγονός ότι μία πλήρως κατανεμημένη τοπολογία απαρτίζεται ακριβώς από n(n-1)/2 συνδέσμους, οπού n ο αριθμός των κόμβων του δικτύου. Αντίθετα, οι μερικώς κατανεμημένες τοπολογίες διαθέτουν λιγότερους

συνδέσμους.



Εικόνα 1.4 – α) Partal Mesh β) Fully Mesh

Η κατανεμημένη τοπολογία ενδείκνυται κυρίως για λόγους αξιοπιστίας και φερεγγυότητας. Η ακεραιότητα των κατανεμημένων τοπολογιών απορρέει από το γεγονότος ότι πρακτικά απαρτίζονται από πολλές point to point συνδέσεις μεταξύ των οντοτήτων του δικτύου. Το παραπάνω φαινόμενο καθιστά εύκολη την απομόνωση και εύρεση διάφορων βλαβών που μπορεί να συμβούν στο δίκτυο. Επίσης, αν παρουσιαστεί κάποιο πρόβλημα σε μία γραμμή, θα καταρρεύσει μονάχα μία σύνδεση, με αποτέλεσμα να μην επηρεαστεί η επικοινωνία όλων των κόμβων του δικτύου. Ωστόσο, πρακτικά οι πλήρως κατανεμημένες τοπολογίες εφαρμόζονται μόνο σε δίκτυα που καλύπτουν μικρή έκταση και διαθέτουν λίγους κόμβους, καθώς για ευρέα δίκτυα συνήθως εφαρμόζονται μερικώς κατανεμημένες τοπολογίες με λίγες συνδέσεις, εξαιτίας του αυξημένου κόστους του πλήρως κατανεμημένου δικτύου [3].

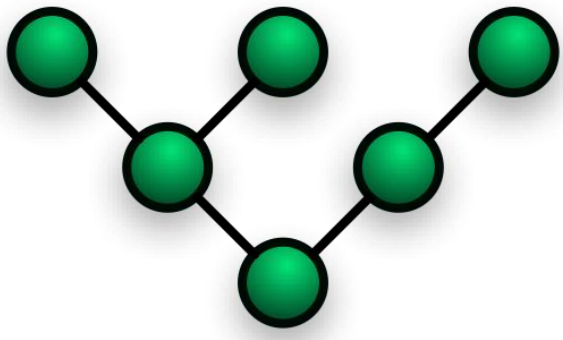
Τοπολογία δένδρου

Η τοπολογία δένδρου αποτελεί συνδυασμό της τοπολογίας αστέρα και της τοπολογίας διαύλου και είναι ουσιαστικά μια καινοτομική υβριδική αρχιτεκτονική. Πιο συγκεκριμένα στη τοπολογία δένδρου, οι κόμβοι της τοπολογίας bus αντικαθίστανται πλήρως με πιο μικρά δίκτυα τοπολογίας αστέρα. Με άλλα λόγια πολλά μικρότερα υπο-δίκτυα αστέρα συνδέονται σε ένα μία bus τοπολογία και σχηματίζουν ένα μεγαλύτερο. Η τοπολογία δένδρου εμπεριέχει τόσο τα πλεονεκτήματα της τοπολογίας αστέρα όσο και τα μειονεκτήματα της τοπολογίας αρτηρίας. Για παράδειγμα, εάν η σύνδεση μεταξύ των δύο ομάδων ή δύο κόμβων του δικτύου οι οποίες δεν ανήκουν στην ίδια συστάδα (**pod**) χαθεί ή χαλάσει λόγω μίας απρόσμενης θραύσης της σύνδεσης στον κεντρικό γραμμικό πυρήνα, τότε οι δύο αυτές ομάδες δεν μπορούν να επικοινωνήσουν όπως ακριβώς θα γινόταν σε μία bus τοπολογία. Ωστόσο, οι κόμβοι τοπολογίας αστέρα θα εξακολουθήσουν να επικοινωνούν αποτελεσματικά μεταξύ τους. Οι δομές αυτές συνήθως περιλαμβάνουν ριζικούς κόμβους (root nodes), ενδιάμεσους κόμβους (intermediate node) και κεντρικούς κόμβους (core). Αυτή η δομή είναι διατεταγμένη σε μια ιεραρχική μορφή και κάθε ενδιάμεσος κόμβος μπορεί να έχει οποιοδήποτε αριθμό από τεκνικούς κόμβους (**child**

nodes). Η τοπολογία δένδρου είναι συνήθως οργανωμένη σε ιεραρχικά επίπεδα (**levels** ή **layers**). Ο κεντρικός κόμβος συνδέεται με point to point συνδέσεις με έναν ή περισσότερους κόμβους που ανήκουν στο δεύτερο επίπεδο ιεραρχίας του δικτύου. Αντίστοιχα, οι κόμβοι του δεύτερου επιπέδου συνδέονται με ανάλογο τρόπο με τους κόμβους του τρίτου επιπέδου και ούτω κάθε εξής. Το παραπάνω φαινόμενο συνεχίζεται, με αποτέλεσμα η λογική τοπολογία του δικτύου να θυμίζει την γραφική απεικόνιση ενός δένδρου. Γενικότερα, για να θεωρηθεί ότι η τοπολογία ενός δικτύου ανήκει σε αυτήν την κατηγορία θα πρέπει να απαρτίζεται το λιγότερο από τρία επίπεδα. Επίσης, τουλάχιστον ένας κόμβος ενός επιπέδου n θα πρέπει να συνδέεται με τουλάχιστον δύο κόμβους που ανήκουν στο επίπεδο n+1, δηλαδή στο επόμενο επίπεδο. Αυτό σημαίνει ότι ο παράγοντας διακλάδωσης (branching factor) του κάθε κόμβου πρέπει να είναι πάντα μεγαλύτερος του ένα. Επίσης, ένα μοναδικό χαρακτηριστικό της συγκεκριμένης τοπολογίας είναι το γεγονός ότι οι κόμβοι που βρίσκονται υψηλότερα στην ιεραρχία εκτελούν περισσότερες επεξεργαστικές διεργασίες και συχνά παρέχουν υπηρεσίες που συσχετίζονται με την επεξεργασία δεδομένων προς τους κόμβους των κατώτερων επιπέδων.

Εξαιτίας των αρχιτεκτονικών της προδιαγραφών η τοπολογία δένδρου διαθέτει αρκετά πλεονεκτήματα και γενικότερα ενδείκνυται για την υλοποίηση μεγάλων δικτύων οργανισμών ακόμα και σε κέντρα δεδομένων. Όπως και με την κατανεμημένη τοπολογία, έτσι και η τοπολογία δένδρου προσφέρει εύκολη απομόνωση λαθών, επειδή απαρτίζεται από point to point συνδέσεις. Επίσης, ένας ακόμα παράγοντας που διευκολύνει στον εντοπισμό των λαθών, είναι το γεγονός ότι το πλήθος των διεργασιών που λαμβάνουν τόπο στο δίκτυο δεν είναι ισάριθμα κατανεμημένος, αλλά ο κάθε κόμβος εκτελεί ένα συγκεκριμένο πλήθος από διεργασίες, που συχνά διαφέρει από τις διεργασίες των κόμβων των άλλων επιπέδων.

Όσον αφορά την επεκτασιμότητα ενός δικτύου, συγκεκριμένα, αν και η ενσωμάτωση ενός νέου κόμβου στο κατώτερο επίπεδο του δικτύου συνήθως αποτελεί μία απλή διαδικασία, η ίδια διαδικασία μπορεί να αποβεί ιδιαίτερα πολύπλοκη και χρονοβόρα, αν εφαρμοστεί σε ένα ανώτερο επίπεδο. Το παραπάνω συμβαίνει, διότι οι κύριες διεργασίες που συσχετίζονται με τις υπηρεσίες που παρέχει το δίκτυο, εκτελούνται στους κόμβους που βρίσκονται υψηλά στην ιεραρχία και συνεπώς μία αλλαγή σε ένα τέτοιο επίπεδο μπορεί να επηρεάσει δραματικά την απόδοση ολόκληρου του δικτύου [3].



Εικόνα 1.5 – Τοπολογία δένδρου

Υβριδικές τοπολογίες

Τα υβριδικά δίκτυα χρησιμοποιούν ένα συνδυασμό οποιωνδήποτε δύο ή περισσοτέρων

τοπολογιών, κατά τέτοιο τρόπο ώστε το προκύπτον δίκτυο δεν παρουσιάζει καμία από τις τυπικές τοπολογίες. Μια υβριδική τοπολογία παράγεται όταν δύο διαφορετικές βασικές τοπολογίες δικτύου διασυνδέονται μεταξύ τους. Ενδεικτικά παρατίθονται δύο γνωστά υβριδικά δίκτυα των οποίων η ανάλυση δεν θα αναφερθεί στα πλαίσια της εν λόγω εργασίας: το **star-ring** network καθώς και το **star-bus** network [3].

1.2 Βασικά στοιχεία μεταγωγής και τοπολογίες μεταγωγής δικτύου

Τα 2 κυριότερα στοιχεία μεταγωγής στα σύγχρονα δίκτυα ειναι οι μεταγωγείς (**switches**) και οι δρομολογητές (**routers**) [4].

Οι μεταγωγείς αποτελούν το θεμέλιο των περισσότερων δικτύων. Ένας μεταγωγέας ενεργεί ως ελεγκτής, που συνδέει υπολογιστές, εκτυπωτές και servers σε ένα δίκτυο. Οι μεταγωγείς επιτρέπουν οι συσκευές του δικτύου να επικοινωνούν μεταξύ τους καθώς και με άλλα δίκτυα. Υπάρχουν δύο τύποι μεταγωγέων ως μέρος των βασικών δομικών στοιχείων δικτύωσης : διαχειριζόμενοι και μη διαχειριζόμενοι. Ένας μη διαχειριζόμενος μεταγωγέας λειτουργεί αλλά δεν μπορεί να ρυθμιστεί. Ένας διαχειριζόμενος μεταγωγέας μπορεί να ρυθμιστεί τοπικά είτε εξ αποστάσεως και προσφέρει καλύτερο έλεγχο της κυκλοφορίας μέσα στο δίκτυο καθώς και την πρόσβαση σε αυτό [4].

Οι δρομολογητές συνδέουν διαφορετικά δίκτυα. Μπορούν επίσης να συνδέουν τους υπολογιστές των δικτύων αυτών στο Διαδίκτυο. Οι δρομολογητές επιτρέπουν σε όλους τους δικτυωμένους υπολογιστές να μοιράζονται μια ενιαία σύνδεση στο Internet. Ένας δρομολογητής λειτουργεί ως αποστολέας, αναλύοντας τα δεδομένα που αποστέλλονται μέσω ενός δικτύου, επιλέγοντας την καλύτερη διαδρομή που θα πάρουν τα δεδομένα αυτά και είναι υπεύθυνος και για την αποστολή τους. Οι δρομολογητές επίσης φροντίζουν και για την προστασία των πληροφοριών από απειλές της ασφάλειας και μπορεί ακόμη και να αποφασίζουν ποιοι υπολογιστές λαμβάνουν προτεραιότητα έναντι άλλων, ανάλογα με τις προδιαγραφές του αντίσοιχου πρωτοκόλλου [4].

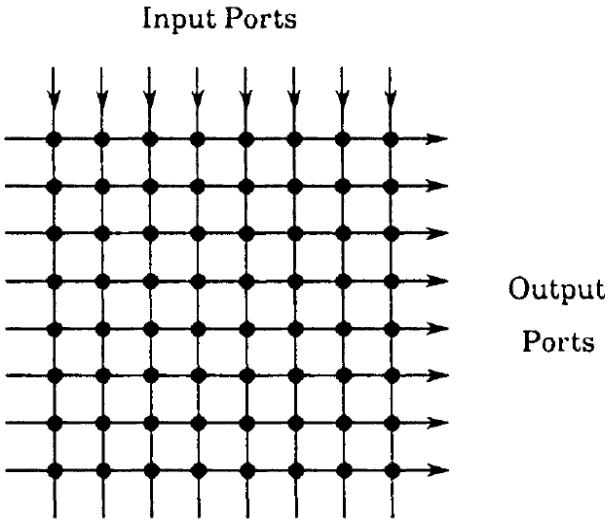
Στο σημείο αυτό έχοντας παρουσιάσει τα βασικά στοιχεία μεταγωγής, θα ακολουθήσουν οι κυριότερες τοπολογίες μεταγωγής δικτύου (**switching fabric topologies**) [5]. Οι τοπολογίες μεταγωγής δικτύου είναι ουσιαστικά το hardware και το software που χρησιμοποιείται στους κόμβους των δικτύων για τη μεταφορά των δεδομένων που φτάνουν στις εισόδους, προς τις σωστές εξόδους. Οι τοπολογίες που θα παρουσιαστούν είναι το **Crossbar** δίκτυο[5], τα δίκτυα **Banyan** [5], η **Clos** τοπολογία [5] και το δίκτυο **Benes** [5]. Πριν προχωρήσουμε όμως στην ανάλυση των τοπολογιών αυτών, κρίνεται απαραίτητη η επεξήγηση της ταξινόμησης τους ανάλογα με τα χαρακτηριστικά φραγής τους (**blocking**), δηλαδή εάν τα εισερχόμενα πακέτα σε ένα δίκτυο καταφέρουν να δρομολογηθούν σε μία έξοδο ανεξαρτήτως της εισόδου από την οποία εισχωρήσαν στο δίκτυο. Ένα δίκτυο που είναι πάντα σε θέση να συνδέει μια ελεύθερη είσοδο σε μια ελεύθερη έξοδο, ανεξάρτητα από τις συνδέσεις που έχουν ήδη εγκαταστηθεί σε όλο το δίκτυο, λέγεται ότι είναι **non-blocking,** δηλαδή χωρίς φραγή. Τα crossbar και Clos δίκτυα είναι παραδείγματα non-blocking δικτύων. Ένα δίκτυο που είναι πάντα σε θέση να συνδέει μια ελεύθερη είσοδο σε μια ελεύθερη έξοδο, αλλά η οποία μπορεί να απαιτήσει αλλαγές στις υπάρχουσες συνδέσεις καλείται επαναδιευθετήσιμο δίκτυο χωρις φραγή (**rearrangeable non-blocking**). Να σημειωθεί πάντως ότι μία αίτηση για αποσύνδεση δεν επηρεάζει τα επαναδιευθετήσιμα δίκτυα να αλλάξουν ξανά την κατάσταση τους. Το Benes δίκτυο είναι ένα παράδειγμα ενός rearrangeable non-blocking δικτύου. Ένα δίκτυο χαρακτηρίζεται ως **blocking** –

δίκτυο φραγής εάν οποιαδήποτε τυχόν ελεύθερη έξοδος μπορεί να μην είναι διαθέσιμη σε οποιαδήποτε ελεύθερη είσοδο, επειδή οι υπάρχουσες συνδέσεις αποτρέπουν την εγκατάσταση ενός μονοπατιού (path) μεταξύ της εισόδου αυτής και της ελεύθερης εξόδου. Το δίκτυο Banyan χαρακτηρίζεται ως blocking για την κίνηση με τυχαία κατανομή προορισμού [5].

Όπως θα αναμένεται ένα μη-blocking δίκτυο απαιτεί περισσότερα στοιχεία μεταγωγής και διασυνδέσεις ό, τι ένα rearrangeable non-blocking δίκτυο η οποία με τη σειρά του απαιτεί περισσότερα στοιχεία μεταγωγής από ένα απλό δίκτυο φραγής. Επίσης, η απόδοση μιας ταχείας μεταγωγής πακέτων εξαρτάται από τα χαρακτηριστικά φραγής ενός δικτύου. Τέλος ένα rearrangeable non-blocking δίκτυο παρέχει υπηρεσίες και αποδόσεις χωρίς φραγή μόνο εάν ένας αλγόριθμος ελέγχου είναι διαθέσιμος για να εκτελέσει την αναδιάταξη των συνδέσεων. Για γρήγορη μεταγωγή πακέτων μπορεί να γίνει χρήση μόνο κατανεμημένων αλγορίθμων (distributed algorithms).

Crossbar Δίκτυο

Το δίκτυο crossbar, είναι ένα non-blocking δίκτυο διασύνδεσης που έχει ονομαστεί έτσι από μια συγκεκριμένη εφαρμογή μεταγωγέα που αναπτύχθηκε για εφαρμογές τηλεφωνικής μεταγωγής. Το όνομα παραπέμπει σε non-blocking δίκτυα γενικότερα. Η δομή του δικτύου αυτού απεικονίζεται στην ακόλουθη εικόνα:



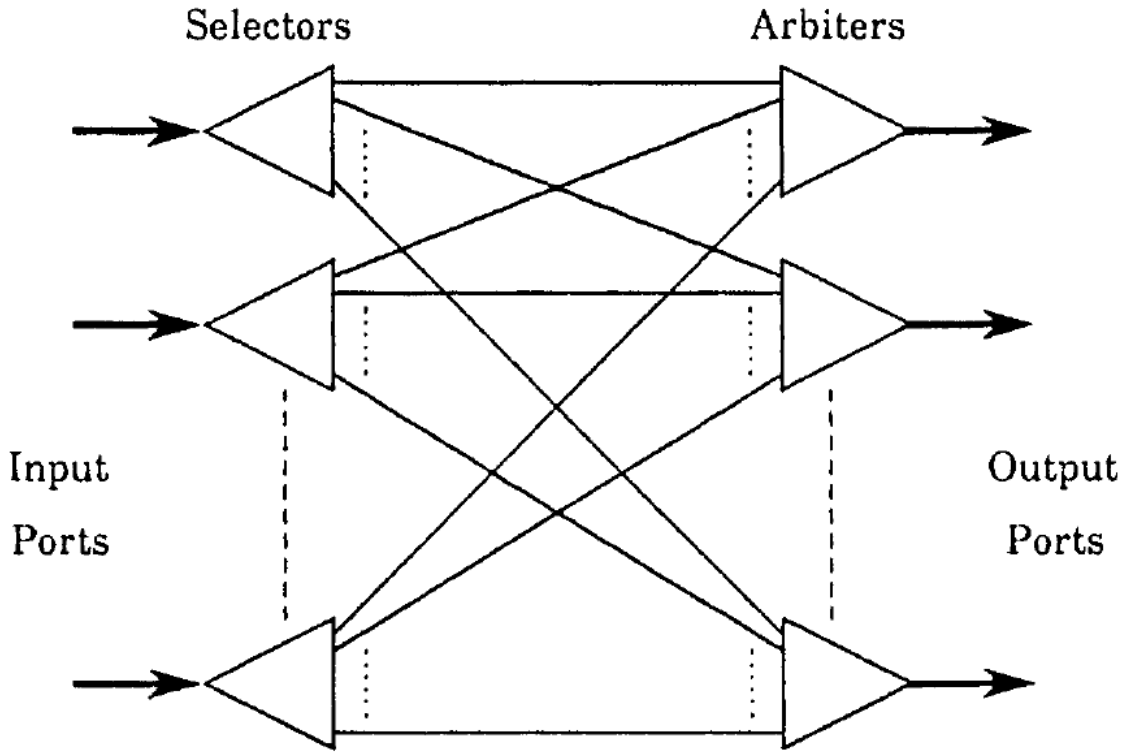
Εικόνα 1.6 – Δομή δικτύου Crossbar

Κάθε κόμβος του δικτύου χαρακτηρίζεται ως μία διασταύρωση (**crosspoint**) και είναι ένας απλός μεταγωγέας που έχει δύο καταστάσεις, ανοιχτά (open) και κλειστά (closed). Χρησιμοποιώντας μηχανισμούς κεντρικού ελέγχου το δίκτυο μπορεί να ικανοποιήσει συνδέσεις με hosts τύπου είτε ένα προς ένα (unicast) ή και ένας προς πολλούς (multicast). Απαιτεί $O(N^2)$ crosspoints και ως εκ τούτου το υλικό που απαιτείται για την υλοποίηση του δικτύου αναπτύσσεται γρήγορα και αυξητικά με το μέγεθος του δικτύου.

Δίκτυα διασύνδεσης πολλαπλών σταδίων κατασκευάζονται από τα στάδια των διασυνδεδεμένων crossbar στοιχείων μεταγωγής χαμηλού βαθμού. Η crossbar τοπολογία μπορεί να υλοποιηθεί με την bus τοπολογία μία διάταξη στην οποία όλες οι συσκευές του δικτύου θα

συνδέονται σε ένα κεντρικό δίαυλο. Ένα σημαντικό πλεονέκτημα της crossbar μεταγωγής είναι ότι, καθώς η κίνηση μεταξύ οποιωνδήποτε δύο συσκευών αυξάνεται, δεν επηρεάζει την κυκλοφορία μεταξύ άλλων συσκευών. Εκτός από την προσφορά μεγαλύτερης ευελιξίας, ένας crossbar μεταγωγέας προσφέρει καλή δυνατότητα κλιμάκωσης.

Στην επόμενη απεικόνιση παρουσιάζεται μια εναλλακτική υλοποίηση του crossbar στοιχείου μεταγωγής κατάλληλο για χρήση με κατανεμημένο έλεγχο μέσα στο περιβάλλον ενός δικτύου διασύνδεσης πολλαπλών σταδίων. Ας υποθέσουμε ότι προκύπτει ένα μη αναμενόμενο πακέτο στο οποίο προηγείται η ετικέτα του που δείχνει τον απαιτούμενο προορισμό. Ο επιλογέας της θύρας εισόδου εξετάζει αυτή την ετικέτα και ελέγχει την κατάσταση του διαιτητή του στη θύρα εξόδου. Εάν ο συγκεκριμένος διαιτητής δείχνει ότι η απαιτούμενη θύρα εξόδου είναι ελεύθερη εγκαθίσταται μια σύνδεση για τη μεταφορά του πακέτου αλλά εάν δεν είναι τότε του αρνείται η πρόσβαση. Όλοι οι επιλογείς (**selectors**) μπορούν έτσι να λειτουργούν ταυτόχρονα και ασύγχρονα. Συνδέσεις πολλαπλής διανομής δεν υποστηρίζονται από αυτό το σχεδιασμό crossbar δικτύου [5].

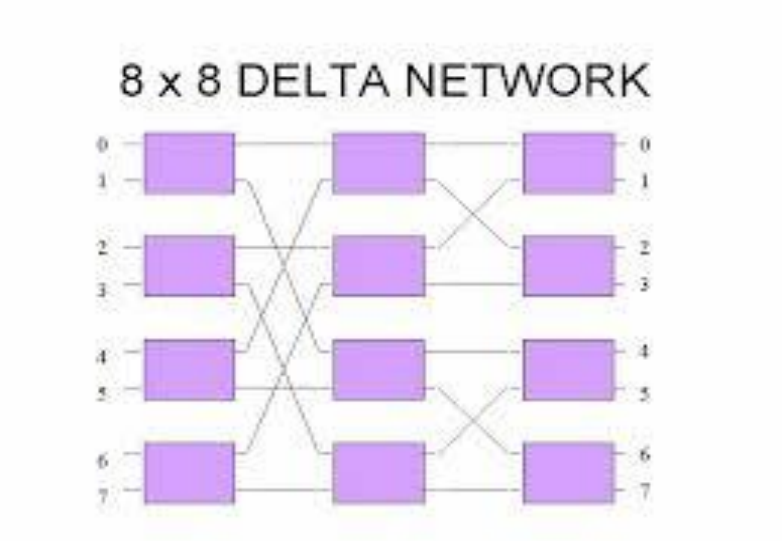


Εικόνα 1.7 – Crossbar Switch

Banyan δίκτυα

Το δίκτυο Banyan είναι ένα δίκτυο πολλαπλών σταδίων αποτελούμενο από διασυνδεδεμένα crossbar στοιχεία μεταγωγής και έχει ονομαστεί από το East Indian fig tree του οποίου η δομή

υποτίθεται ότι μοιάζει. Τα banyan δίκτυα έχουν ένα και μόνο μονοπάτι από οποιαδήποτε είσοδο σε οποιαδήποτε έξοδο και έτσι καλύπτεται μια πολύ μεγάλη κατηγορία των πιθανών δομών του δικτύου. Εάν οι συνδέσεις στο Banyan δίκτυο περιορίζεται σε συνδέσεις στοιχείων μεταγωγής με γειτονικά στάδια μεταγωγής τότε προκύπτει ένα L-level banyan δίκτυο και αν επιπλέον όλα τα στοιχεία μεταγωγής στο δίκτυο είναι πανομοιότυπα έχουμε ένα κλασσικό δίκτυο banyan. Ένα δίκτυο Banyan με τετραγωνικά στοιχεία μεταγωγής δηλαδή με στοιχεία που έχουν τον ίδιο αριθμό εισόδων και εξόδων, καλείται ορθογώνιο. Δύο τάξεις του banyan αποτελούν ιδαίτερο ενδιαφέρον τα SW και CC banyans. Το CC-Banyan είναι ορθογώνιο δίκτυο [5]. Το SW-Banyan μπορεί να δειχθεί ότι είναι αυτο-δρομολόγησης και ως εκ τούτου το ορθογώνιο SW-banyan που είναι κατασκευασμένο από 22 crossbar στοιχεία μεταγωγής, είναι σχεδόν πάντα το δίκτυο που προβλέπεται και η κλασσική αναφορά σε δίκτυα banyan παραπέμπει στο SW.
Τα δίκτυα Banyan χαρακτηρίζονται από μία μεγάλη ποικιλία δικτύων σε παραλλαγές. Άξια αναφοράς είναι τα δίκτυα Delta αλλά και τα Omega [5]. Στην επόμενη εικόνα διασαφηνίζεται μία από τις παραλλαγές αυτές, ένα 8 x 8 Delta δίκτυο [5].



Εικόνα 1.8 – Δίκτυο Banyan

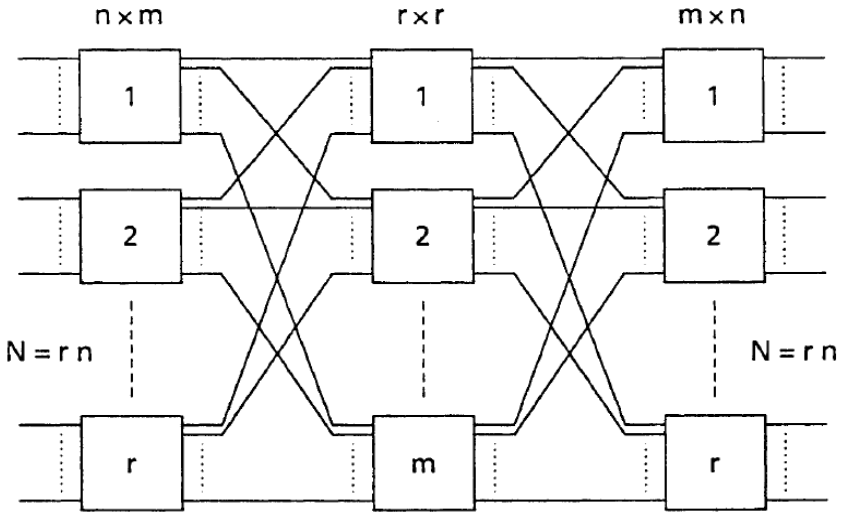
Clos δίκτυα

Το Clos δίκτυο είναι ένα δίκτυο μεταγωγής πολλαπλών σταδίων. Στην επόμενη εικόνα απεικονίζεται ένα παράδειγμα ενός δικτύου Clos 3 σταδίων. Το πλεονέκτημα αυτών των δικτύων είναι ότι η σύνδεση μεταξύ ενός μεγάλου αριθμού των θυρών εισόδου και εξόδου μπορεί να γίνει με τη χρήση μόνο μικρού μεγέθους μεταγωγέων. Στο σχήμα το n παριστά τον αριθμό των πηγών που τροφοδοτούν κάθε μία από τις m εισόδους των crossbar μεταγωγέων. Όπως μπορεί να φανεί, υπάρχει ακριβώς μία σύνδεση μεταξύ κάθε μεταγωγέα εισόδου και κάθε μεταγωγέα στο μεσαίο στάδιο της αρχιτεκτονικής. Και κάθε μεταγωγέας μεσαίου στάδιου συνδέεται ακριβώς μία φορά σε κάθε μεταγωγέα στο στάδιο εξόδου.

Μπορεί να δειχθεί ότι με $m \geq n$, το δίκτυο Clos μπορεί να είναι non-blocking δίκτυο όπως και το crossbar, όπου τα μεγέθη των m και n απεικονίζονται στο σχήμα 1.9 ως είσοδοι και έξοδοι (ή αντίθετα). Δηλαδή όπως αναλύθηκε και προηγουμένως για ένα ζευγάρι εισόδου εξόδου

μπορούμε να βρούμε μια διάταξη μονοπατιών για τη σύνδεση των εισόδων και εξόδων μέσω των μεταγωγέων του μεσαίου σταδίου.

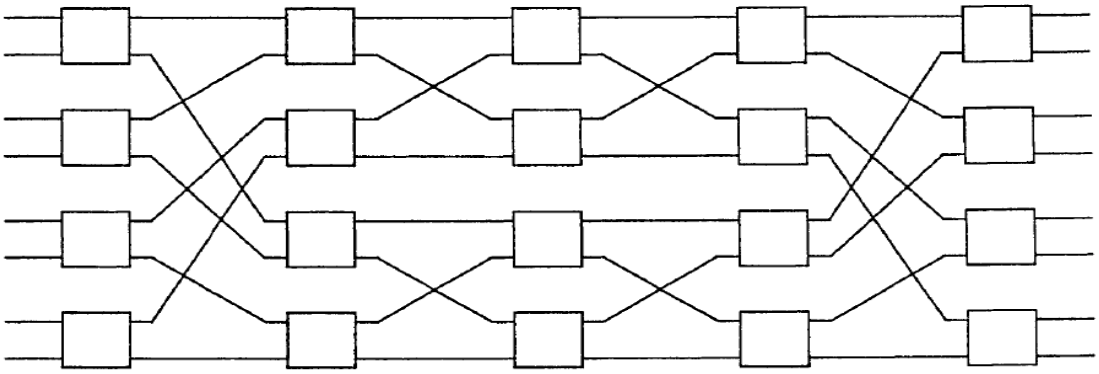
Επίσης μπορεί να δειχθεί με μαθηματικές αποδείξεις ότι αν η συνθήκη $m \geq 2n-1$ ικανοποείται τότε μια νέα σύνδεση μπορεί πάντα να προστεθεί χωρίς αναδιάταξη του όλου συστήματος, πράγμα που καθιστά τα Clos δίκτυα σχετικα ευέλικτα. Το Clos δίκτυο μπορεί αναδρομικά να ανακατασκευαστεί σε δίκτυο με 5 στάδια ή και σε δίκτυο με 7 στάδια και ούτω καθεξής εάν υπάρχει ανάγκη για επέκταση, αντικαθιστώντας κάθε διακόπτη στο κεντρικό στάδιο με ένα άλλο Clos δίκτυο 3 σταδίων [5].



Εικόνα 1.9 – Clos δικτυα

Benes Δίκτυα

Το δίκτυο Benes είναι μια ειδική περίπτωση του δικτύου Clos για το οποίο έχει δειχτεί ότι εάν ισχύει $m > n$ το δίκτυο είναι rearrangeable non-blocking. Ένα δίκτυο benes 8x8 απεικονίζεται στην επόμενη εικόνα, αποτελούμενο από 2 x 2 στοιχεία μεταγωγής. Ένα N x N δίκτυο Benes απαιτεί $2 * \log_d (N - 1)$ στάδια από στοιχεία μεταγωγής βαθμού d, με N / d στοιχεία μεταγωγής ανά στάδιο.



Εικόνα 1.10 – Δίκτυα Benes

Μόνο το τα τελικά $\log_d N$ στάδια μεταγωγής χρειάζεται να παρέχουν τη λειτουργία της δρομολόγησης σε ένα δίκτυο Benes και έτσι τα υπόλοιπα στάδια προσφέρουν απλά πολλαπλές διαδρομές μέσα στο δίκτυο. Πράγματι, το δίκτυο Benes μπορεί να θεωρηθεί ως ένα δίκτυο Delta στο οποίο προηγούνται στάδια μεταγωγής τα οποία διανέμουν την μη αναμενόμενη κίνηση σε όλο το δικτυδόμημα κάνοντας χρήση των πολλαπλών διαδρομών έτσι ώστε να προσφέρει ανοχή σε σφάλματα και να μειωθεί το blocking. Αν αναπτυχθεί ένας κεντρικοποιημένος αλγόριθμος ελέγχου το δίκτυο μπορεί να λειτουργήσει σαν ένα rearrangeable non-blocking δίκτυο αλλά ο αλγόριθμος που απαιτείται είναι και κεντρικοποιημένος και χρονοβόρος.

Διάφορες επιλογές είναι δυνατές για τη χρήση ενός κατανεμημένου αλγορίθμου σε ένα Benes δίκτυο. Τα στάδια δρομολόγησης του δικτύου μπορεί να χρησιμοποιούν το ίδιο αυτο-αλγόριθμο δρομολόγησης όπως και με τα δίκτυα Delta που βασίζεται στη χρήση μιας ετικέτας δρομολόγησης προορισμού. Τα στάδια της διανομής του δικτύου μπορεί να αλλάζουν ανάλογα με τρεις πιθανούς αλγόριθμους: δρομολόγηση προέλευσης, τυχαία δρομολόγηση ή αλγορίθμους πλημμύρας [5].

1.3 Μετρικά Απόδοσης Δικτύου

Σε αυτήν την ενότητα θα παρουσιάσουμε τα μετρικά απόδοσης ή επίδοσης ενός δικτύου. Ο καθορισμός αυτών των τεχνικών στόχων αποτελεί ένα σημαντικό στάδιο κατά την υλοποίηση ενός δικτύου, αφού καθορίζει την αρχιτεκτονική δομή του δικτύου σε επίπεδο τοπολογίας, υλικού και λογισμικού, την πολιτική δρομολόγησης, τα χαρακτηριστικά της κίνησης και συσχετίζεται άρρηκτα με την ποιότητα των υπηρεσιών, που θα προσφέρει το δίκτυο. Τα δυο βασικότερα αλλά και σημαντικότερα μετρικά τα οποία καθορίζουν την επίδοση του δικτύου ειναι η λανθάνουσα καθυστέρηση – **latency** και η διεκπεραιωτικότητα – **throughput**. Πριν προχωρήσουμε όμως στην ανάλυση και μελέτη αυτών θα αναφερθούμε σε ένα άλλο πολύ σημαντικό μετρικό, το εύρος ζώνης – **bandwidth**.

1.3.1 Bandwidth

Το εύρος ζώνης αναφέρεται στην ποσότητα των πληροφοριών που μπορούν να μεταδοθούν μέσω μίας ζεύξης σε ένα δεδομένο χρονικό διάστημα από ένα σημείο του δικτύου σε ένα άλλο και συνήθως εκφράζεται σε bits ανά δευτερόλεπτο ή **bps** [6] [10]. Συνήθως, το εύρος ζώνης εξαρτάται από διάφορα τεχνικά χαρακτηριστικά, όπως για παράδειγμα από τι υλικό είναι το μέσο ή τη φυσική δομή του γενικότερα. Επίσης είναι σημαντικό να αναφερθεί ότι όταν σε ένα κέντρο δεδομένων θέλουν να επικοινωνήσουν 2 hosts μεταξύ τους, και το δίκτυο εξασφαλίσει ένα μονοπάτι μεταξύ τους, το μονοπάτι αυτό στην ουσία είναι μια διαδοχή απο συνδέσεις, η κάθε μία σύνδεση με το δικό της εύρος ζώνης, και έτσι το συνολικό εύρος ζώνης περιορίζεται στο εύρος ζώνης του συνδέσμου με τη χαμηλότερη ταχύτητα. Αντίστοιχα, διαφορετικές εφαρμογές απαιτούν διαφορετικά εύρη ζώνης για την εκτέλεση τους. Στην συνέχεια θα παρουσιάσουμε μερικούς χρήσιμους ορισμούς.

Η διχοτόμηση (**bisection**) ενός δικτύου είναι ο διαμερισμός του σε δύο ισομέγεθα σύνολα από κόμβους [7]. Το πλάτος διχοτόμησης (**Bisection Width**) ενός δικτύου ορίζεται ως ο ελάχιστος αριθμός των συνδέσεων που πρέπει να κοπούν προκειμένου να διχοτομηθεί η τοπολογία με τέτοιο τρόπο. Για μια συγκεκριμένη διατομή, το άθροισμα των χωρητικοτήτων των δεσμών μεταξύ των δύο διχοτομημένων τμημάτων ονομάζεται το εύρος ζώνης της διχοτόμησης. Το **εύρος ζώνης διχοτόμησης – Bisection Bandwidth** ενός δικτύου είναι το ελάχιστο εύρος ζώνης (το άθροισμα των χωρητικοτήτων των δεσμών μεταξύ των δύο τμημάτων) κατά μήκος όλων των πιθανών bisections. Για ένα δίκτυο όπου όλοι οι σύνδεσμοι έχουν εύρος ζώνης ίσο με b τότε το εύρος ζώνης διχοτόμησης μετριέται με τον τύπο:

$$B_b = b * B_w \quad (1)$$

όπου B_b είναι το εύρος ζώνης διχοτόμησης και B_w το πλάτος διχοτόμησης που ορίστηκε προηγουμένως [11].

Ένα δίκτυο έχει πλήρες εύρος ζώνης διχοτόμησης (**Full Bisection Bandwidth*)*** εάν το εύρος ζώνης διχοτόμησης του είναι αρκετό για να υποστηρίξει τη χειρότερη δυνατή περίπτωση της κίνησης στο δίκτυο. Εάν υποτεθεί η διχοτόμηση σε ένα δίκτυο, ένα παράδειγμα για τη χειρότερη περίπτωση κίνησης είναι όταν το σύνολο της κίνησης διασχίζει τη διχοτόμηση, δηλαδή κάθε κόμβος από ένα σύνολο επικοινωνεί με έναν κόμβο του άλλου σετ σε ένα πλήρες εύρος ζώνης. Το Bisection Bandwidth αποτελεί μόνο ένα άνω φράγμα για το συνολικό εύρος ζώνης που η αίτηση μπορεί να λάβει για αυτή τη χειρότερη δυνατή περίπτωση. Αυτό το ανώτερο όριο επιτυγχάνεται μόνο όταν παρέχεται τέλεια εξισορρόπηση φορτίου από το σύστημα. Το Bisection Width και το Bisection Bandwidth είναι μετρικά που εξαρτώνται από την τοπολογία και το εύρος ζώνης των συνδέσεων [7].

1.3.2 Throughput

Η διεκπεραιωτικότητα ή throughput ενός δικτύου για ένα συγκεκριμένο μοτίβο της κυκλοφορίας είναι η ταχύτητα με την οποία τα πακέτα παραδίδονται επιτυχώς στους προορισμούς τους από το δίκτυο και μετράται σε bits ανά δευτερόλεπτο (bps) ή ορισμένες φορές σε πακέτα δεδομένων ανά δευτερόλεπτο (data packets per sec) [8] [10]. Πιο συγκεκριμένα το throughput αποτελεί την ποσότητα της πληροφορίας που μπορεί να μεταδοθεί με επιτυχία μέσω

του δικτύου. Συχνά το throughput ταυτίζεται λανθασμένα με το εύρος ζώνης, η σημαντική διαφορά των δύο είναι ότι το bandwidth είναι πιο πολύ το τι προδιαγραφές έχει το δίκτυο ενώ το throughput είναι η ταχύτητα της επιτυχημένης παράδοσης των πακέτων και το πόσο καλά αξιοποιείται το bandwidth από το δίκτυο. Συνεπώς, θα ήταν επιθυμητό το throughput και το εύρος ζώνης να είχαν τις ίδιες τιμές. Ωστόσο, το παραπάνω αποτελεί μια ιδανική κατάσταση, καθότι στην πράξη πάντα κάποια πακέτα θα απορρίπτονται για διάφορους λόγους.

Το Throughput είναι στενά συνδεδεμένο με το προσφερόμενο φορτίο, δεδομένης της αρχιτεκτονικής του δικτύου, το μοτίβο της κυκλοφορίας και τη στρατηγική δρομολόγησης. Μπορούμε να μετρήσουμε το μέσο throughput όλων των συνδέσεων του δικτύου ή της κάθε σύνδεσης ξεχωριστά με τη σύνδεση που εμφανίζει τη χαμηλότερη ταχύτητα ή απόδοση.

Στη συνέχεια θα περιγράψουμε μια ειδική περίπτωση του throughput, το ιδανικό throughput (ideal) το οποίο μπορεί να υπολογιστεί για ένα αυθαίρετο μοτίβο κίνησης και μια αυθαίρετη τοπολογία με την επίλυση ενός προβλήματος σε σχέση με τη ροή του φορτίου [9]. Για ορισμένα μοτίβα κυκλοφορίας και τοπολογίες το ιδανικό throughput μπορεί να υπολογιστεί και στο χαρτί. Η μέση και η χειρότερη περίπτωση throughput (μη-ιδανικό) που λαμβάνει υπόψη τις αποφάσεις δρομολόγησης και τον έλεγχο ροής μπορεί να υπολογιστεί με προσομοιώσεις.

Το ιδανικό throughput είναι το throughput του δικτύου για ένα συγκεκριμένο μοτίβο κίνησης, θεωρώντας ιδανικό έλεγχο της ροής και της δρομολόγησης. Αυτή είναι η διεκπεραιωτικότητα που θα προέκυπτε αν η δρομολόγηση εξισσοροπούσε τέλεια το φορτίο σε εναλλακτικές διαδρομές στο δίκτυο και εάν ο έλεγχος της ροής δεν παρουσίαζε ελαττώματα στα σημεία συμφόρησης (**bottleneck**). Για ένα συγκεκριμένο μοτίβο κίνησης η ιδανική δρομολόγηση είναι η στρατηγική δρομολόγησης η οποία θα μεγιστοποιήσει τη κυκλοφορία πακέτων που παραδίδονται επιτυχώς μέσα στο δίκτυο, δηλαδή θα επιτύχει να εξισορροπήσει το φορτίο και να αποφεύγει τον κορεσμό των συνδέσμων.

Για ένα δίκτυο στο οποίο όλες του οι συνδέσεις έχουν ίσο εύρο ζώνης b, το ιδανικό throughput μπορεί να οριστεί ως το εύρος ζώνης εισόδου που κορεννύει (saturates) το σύνδεσμο ή σημείο συμφόρησης – bottleneck. Το ιδανικό throughput μπορεί να οριστεί και υπολογιστεί ως εξής:

$$\Theta_{ideal} = b / Y_{max} \quad (2)$$

όπου το Y_{max} είναι ένας αδιάστατος αριθμός, ίσος με το λόγο της διέλευσης φορτίου από το δίκτυο και b είναι το εύρος ζώνης του συνδέσμου συμφόρησης [11].

Έχοντας πλέον αναλύσει τι είναι το bandwidth, τι είναι το throughput και πώς συσχετίζονται, στην συνέχεια θα αναφέρουμε τι είναι το **goodput**. Το goodput αποτελεί την ποσότητα του ωφέλιμου φορτίου που μεταφέρεται στο δίκτυο εκφρασμένη και αυτή σε bits/sec. Με την έννοια ωφέλιμο φορτίο, εννοούμε την πληροφορία που προέρχεται από το application layer ή το υψηλότερο επίπεδο πρωτοκόλλου σε ένα δίκτυο, αποκλείοντας τα bits του πρωτοκόλλου καθώς και τα αναμεταδιδόμενα πακέτα δεδομένων.

Συνοψίζοντας, αν και το bandwidth εξαρτάται κυρίως από την τεχνολογία και από τα φυσικά χαρακτηριστικά του μέσου μετάδοσης, το throughput και το goodput δεν επηρεάζονται μόνο από το bandwidth, αλλά εξαρτώνται και από άλλες πιο ευμετάβλητες παραμέτρους. Οι πιο σημαντικές από αυτές είναι: το μέγεθος των πακέτων, ο ρυθμός αποστολής των πακέτων, η επεξεργαστική ισχύς των υπολογιστών ή μηχανημάτων, η τοπολογία του δικτύου και τα πρωτόκολλα σύμφωνα με τα οποία γίνεται η επικοινωνία στο εκάστοτε δίκτυο [8].

1.3.3 Latency

Η λανθάνουσα καθυστέρηση ενός δικτύου καθορίζει την χρονική διάρκεια που απαιτείται, ώστε τουλάχιστον 1 bit δεδομένων να ταξιδέψει σε όλο το δίκτυο από έναν κόμβο ή τερματικό σε ένα άλλο και συνήθως εκφράζεται σε πολλαπλάσια ενός κλάσματος δευτερολέπτου [10]. Προφανώς, η τιμή της καθυστέρησης μπορεί να διαφοροποιηθεί αναλόγως με τη θέση των κόμβων που συμμετέχουν στην μεταφορά των δεδομένων. Οι ειδικοί δικτύων διαφοροποιούν τη συνολική καθυστέρηση που υφίστανται τα δεδομένα στην καθυστέρηση μετάδοσης (**transmission delay**), καθυστέρηση διάδοσης (**propagation delay**), καθυστέρηση επεξεργασίας (**processing delay**) και καθυστέρηση αναμονής (**queuing delay)** καθώς τα πακέτα μπορεί να υπόκεινται σε αποθήκευση και πρόσβαση στο σκληρό δίσκο ή καθυστερήσεις σε ενδιάμεσες συσκευές όπως για παράδειγμα στους μεταγωγείς ή τους δρομολογητές. Το latency επηρεάζει σε μεγάλο βαθμό το πώς μπορούν να χρησιμοποιηθούν ηλεκτρονικές και μηχανικές συσκευές.

Η καθυστέρηση μετάδοσης αποτελεί τη χρονική διάρκεια που είναι απαραίτητη για την προώθηση των δεδομένων ενός πακέτου από έναν κόμβο στον δίαυλο επικοινωνίας και είναι ανάλογη με το μέγεθος του πακέτου, προς το εύρος ζώνης του διαύλου. Αντίστοιχα, η καθυστέρηση διάδοσης συσχετίζεται με τη χρονική διάρκεια που χρειάζεται 1 bit για να ταξιδέψει από την μία άκρη ενός μέσου μετάδοσης στην άλλη άκρη και συνεπώς είναι ανάλογη με το μήκος του μέσου, προς την ταχύτητα των δεδομένων στο μέσο. Περαιτέρω, η καθυστέρηση επεξεργασίας και η καθυστέρηση αναμονής συσχετίζονται με την υπολογιστική ισχύ των κόμβων, αφού εκφράζουν την χρονική καθυστέρηση πού είναι απαραίτητη για την επεξεργασία των headers των πακέτων, καθώς και την αναμονή των πακέτων στους buffers των κόμβων.

Συνήθως στα δίκτυα,μετριέται η μέση λανθάνουσα καθυστέρηση του δικτύου, αλλά και η χειρότερη περίπτωση καθυστέρησης αποτελεί ένα ενδιαφέρον καθώς αυτή μπορεί να είναι υπεύθυνη για την ολική καθυστέρηση του δικτύου και να καθυστερεί και τις μελλοντικές προγραμματισμένες αιτήσεις και την είσοδο νέων πακέτων στο δίκτυο. Οι ειδικοί των δικτύων διαχωρίζουνε για αυτό το λόγο το latency σε καθυστέρηση μηδενικού ή ελάχιστου φορτίου (**zero low-load latency**) και σε καθυστέρηση μη μηδενικού φορτίου (**non-zero load latency**). Το zero load average latency είναι η καθυστέρηση που υφίστανται τα πακέτα σε ένα δίκτυο κατά μέσο όρο, όταν το εκάστοτε φορτίο δεν δημιουργεί **διαμάχη πρόσβασης – contention** στο δίκτυο, δηλαδή όταν τα πακέτα δεν ανταγωνίζονται μεταξύ τους. Το zero load latency T_0 υπολογίζεται από τον τύπο:

$$T_0 = h_{av} * (t_r + t_{trans}) + T_{prop} \quad (3)$$

όπου το $h_{αν}$ είναι μέση απόσταση για το μοτίβο κίνησης στο δίκτυο, το t_r είναι η μέση καθυστέρηση δρομολογητή, το t_{trans} είναι η καθυστέρηση μετάδοσης και τέλος το T_{prop} είναι η συνολική καθυστέρηση διάδοσης για το μέσο μονοπάτι που ακολουθούν τα πακέτα στο δίκτυο. Έτσι το zero load average latency μπορεί να υπολογιστεί με τον πάνω μαθηματικό τύπο, ενώ το non-zero load latency περιλαμβάνει χρόνους αναμονής λόγω του contention που δημιουργείται μεταξύ των πακέτων και τυπικά υπολογίζεται χρησιμοποιώντας προσομοιώσεις ή θεωρίες βασισμένες σε προηγούμενα αποτελέσματα [11].

Οι δοκιμές υπολογισμού του latency μπορεί να ποικίλουν από εφαρμογή σε εφαρμογή. Σε ορισμένες εφαρμογές, η μέτρηση της λανθάνουσας καθυστέρησης απαιτεί ειδικό και πολύπλοκο εξοπλισμό ή γνώση των ειδικών εντολών προγραμματισμού και προγραμμάτων από τους

χρήστες. Σε άλλες περιπτώσεις, η λανθάνουσα κατάσταση μπορεί να μετρηθεί ακόμα και με ένα χρονόμετρο εάν αυτό είναι εύκολο αλλά οι περιπτώσεις αυτές είναι σπάνιες καθώς οι χρόνοι του latency είναι συνήθως πολλαπλάσια ενός κλάσματος δευτερολέπτου όπως αναφέρθηκε και προηγουμένως. Σε ορισμένα δίκτυα, το latency μπορεί να προσδιοριστεί εκτελώντας μια εντολή **ping**, όπως γίνεται και στο διαδίκτυο.

Να σημειωθεί στο σημείο αυτό ότι υπάρχει μια σημαντική διαφορά μεταξύ του μέσου latency και του latency της χειρότερης περίπτωσης. Για τον υπολογισμό του zero load worst latency, αντί για τη μέση απόσταση $h_{αν}$ και τη καθυστέρηση διάδοσης για το μέσο μονοπάτι $T_{prop,}$ χρειαζόμαστε τη απόσταση d του μονοπατιού αυτού στο οποίο παρουσιάζεται το latency της χειρότερης περίπτωσης αλλά και η σχετική καθυστέρηση διάδοσης της συγκεκριμένης διαδρομής $T_{d\text{-}prop}$. Αντικαθιστώντας τα μεγέθη αυτά στον τύπο (3) μπορούμε να έχουμε ένα αρκετά καλό προσεγγιστικό υπολογισμό του zero load worst latency.

Τέλος η μείωση του latency μπορεί να συμβεί μέσως μιας λειτουργίας συντονισμού των συστημάτων, αλλαγών ή και αναβαθμίσεων τόσο του υλικού όσο και του λογισμικού των ηλεκτρονικών υπολογιστών και των μηχανικών συστημάτων. Άλλα μέτρα για τη μείωση του latency και την αύξηση της απόδοσης περιλαμβάνουν την απεγκατάσταση πιθανών περιττών προγραμμάτων, βελτιστοποίηση των πόρων του δικτύου και του λογισμικού, αναβάθμιση του hardware του δικτύου (για παράδειγμα νέα και γρηγορότερα switches) αλλά και μέσω της τεχνικής του **overclocking** σε ορισμένα δίκτυα.

1.3.4 Επεκτασιμότητα

Όταν κάνουμε λόγο για στόχους που αφορούν την επεκτασιμότητα ενός δικτύου, αναφερόμαστε στην λήψη των αποφάσεων που θα κρίνουν κατά πόσο ένα δίκτυο έχει τη δυνατότητα να επεκταθεί και να μεγαλώσει [10]. Συνεπώς, σε αυτήν την κατηγορία στόχων περιλαμβάνονται όλες οι αποφάσεις που συσχετίζονται με επιλογή των κατάλληλων τεχνολογιών που θα καθορίσουν, αν το δίκτυο έχει τη δυνατότητα να συνδεθεί με περισσότερους υπολογιστές, νέους χρήστες, νέους εξυπηρετητές ή ακόμα και νέα δίκτυα.

Κεφάλαιο 2
Αρχιτεκτονική σύγχρονων δικτύων σε κέντρα δεδομένων και περιορισμοί

2.1 Εισαγωγή

Στο κεφάλαιο αυτό, αρχικά γίνεται μία περιγραφή των δομών οι οποίες χρησιμοποιούνται και επικρατούν στα σύχρονα κέντρα δεδομένων. Συγκεκριμένα, προτιμούνται ιεραρχικές δομές με υλοποιήσεις δένδρου, στις οποίες παρατηρούνται και μερικές παραλλαγές. Στην συνέχεια ακολουθεί η ανάλυση της επιλογής της χρήσης ToR (top of rack) μεταγωγέων από αρκετά κέντρα δεδομένων, καθώς αποτελεί μία επικρατούσα τάση, αλλά και τα πλεονεκτήματα που αυτό επιφέρει. Έπειτα εξηγείται διεξοδικά το γιατί αυτές οι ιεραρχικές δομές δένδρων δεν μπορούν να καλύψουν τις μελλοντικές ανάγκες και αυξανόμενες απαιτήσεις των κέντρων δεδομένων. Τέλος εξηγείται η πρόκληση της αποσυνάθροισης των στοιχείων ενό κέντρου δεδομένων και γιατί ότι αυτή αποτελεί έναν ακόμη λόγο για επιπλέον έρευνα και αναβάμιση των κέντρων δεδομένων.

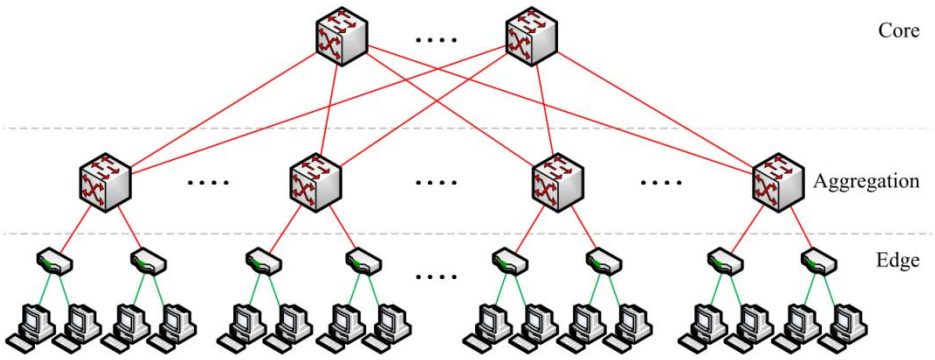
2.2 Σύγχρονες Δομές Δικτύων στα κέντρα δεδομένων

Στην ενότητα αυτή,παρουσιάζονται και αναλύονται οι κύριες δομές δικτύων γύρω από τις οποίες έχουν χτιστεί και οργανωθεί τα κέντρα δεδομένων.

Είναι γεγονός ότι οι πιο επικρατείς τύποι αρχιτεκτονικών στα κέντρα δεδομένων σήμερα έχουν ως βάση τους τη τοπολογία δένδρου (**tree topology**) η οποία αποτελεί συνδυασμό της τοπολογίας αστέρα (**star**) και της τοπολογίας διαύλου (**bus**) όπως αναλύθηκε και στην ενότητα 1.1. Υπευνθυμίζουμε ενδεικτικά ότι οι δομές αυτές συνήθως περιλαμβάνουν ριζικούς κόμβους (root nodes), ενδιάμεσους κόμβους (intermediate nodes) και κεντρικούς κόμβους (core) καθώς και ότι η τοπολογίες δένδρου συνήθως υλοποιούνται σε επίπεδα. Δύο είναι οι πιο διαδεδομένες και χρησιμοποιημένες αρχιτεκτονικές: η παραδοσιακή τοπολογία δένδρου με επίπεδα γνωστή και ως **fat tree** η οποία αξιοποιεί ολοένα και υψηλότερου ρυθμού μετάδοσης μεταγωγείς προς τις ρίζες του δένδρου και η δεύτερη είναι η χρήση φτηνών εμπορικών μεταγωγέων σε ένα fat tree δίκτυο κάνοντας χρήση της **Clos τοπολογίας** έτσι ώστε να παρέχει πολλαπλά μονοπάτια με χαμηλότερο ρυθμό για την επικοινωνία μεταξύ των διακομιστών [12].

2.2.1 Παραδοσιακή τοπολογία δένδρου 3 επιπέδων σε κέντρα δεδομένων

Οι τυπικές αρχιτεκτονικές κέντρων δεδομένων σήμερα αποτελούνται από ένα δένδρο δύο ή τριών επιπέδων αποτελούμενο από μεταγωγείς ή δρομολογητές [12]. Στην εικόνα που ακολουθεί απεικονίζεται ένα δένδρο 3 επιπέδων καθώς είναι η πιο καθιερωμένη μορφή, ενώ τα δένδρα 2 επιπέδων είναι πιο σπάνιες περιπτώσεις.



Εικόνα 2. 1 - 3- Layer Tree Topology

Ένα 3-επίπεδο δένδρο αποτελείται από το κύριο επίπεδο (core layer) το οποίο βρίσκεται στη ρίζα (ή αρχή) του δένδρου, ένα επίπεδο συνάθροισης (**aggregation layer**) το οποίο είναι στη μέση του δένδρου και το επίπεδο των άκρων (edge layer) το οποίο βρίσκεται προς τις φυσικές καταλήξεις του δένδρου ή αλλιώς φύλλα. Το aggregation επίπεδο είναι υπεύθυνο για τη διασύνδεση και την επικοινωνία των edge μεταγωγέων οι οποίοι με τη σειρά τους φροντίζουν για την επικοινωνία των hosts από κάτω τους. Οι μεταγωγείς του aggregation επιπέδου συνδέονται μεταξύ τους μέσω των μεταγωγέων του core επιπέδου. Επίσης οι μεταγωγείς στο core επίπεδο είναι και υπεύθυνοι για τη σύνδεση όλου του κέντρου δεδομένων με το διαδίκτυο – Internet. Οι πιο σπάνιοι σχεδιασμοί της δομής αυτής με 2 επίπεδα περιλαμβάνουν τα core και edge επίπεδα. Ενδεικτικά, ένα δένδρο δύο επιπέδων μπορεί να συντηρεί μεταξύ πέντε με οκτώ χιλιάδες διακομιστές ενώ αντιθέτως ένα δένδρο τριών επιπέδων μπορεί να περιλαμβάνει μέχρι και 25.000 διακομιστές, για αυτό και τα μεγάλα κέντρα δεδομένων επιλέγουν αυτό [12].

Ο λόγος για τον οποίο η συγκεκριμένη δομή έχει μείνει γνωστή ως fat tree είναι επειδή όσο ανεβαίνουμε σε υψηλότερα επίπεδα οι μεταγωγείς εκεί είναι όλο και πιο απαιτητικοί σε ρυθμούς μεταδοσης σε σχέση με το προηγούμενο επίπεδο. Αυτό συμβαίνει γιατί στη δομή αυτή οι σχεδιαστές δικτύων πρέπει να προνοούν και να μπορούν να διαθέτουν πάντα το απαραίτητο εύρος ζώνης για τη περίπτωση που όλοι οι hosts μιας ίδιας συστάδας θέλουν να μεταδόσουν πληροφορία. Για παράδειγμα εάν αναλογιστούμε το δίκτυο της εικόνας 2.1, ο κάθε edge μεταγωγέας διασυνδεέται με 2 hosts. Εάν για παράδειγμα οι hosts αυτοί ανεβάζουνε ταυτόχρονα δεδομένα στα ανώτερα επίπεδα του δικτύου με ρυθμό της τάξεως του 1 Gb/s ο καθένας τότε ο edge μεταγωγέας θα πρέπει να μπορεί να μεταδίδει τουλάχιστον με ρυθμό των 2 Gb/s. Αντιστοίχως ισχύει και για τους μεταγωγείς των ανώτερων επιπέδων.

Οι μεταγωγείς στα φύλλα του δένδρου έχουν ένα σημαντικό αριθμό από GigE θύρες (περίπου 48–288) καθώς επίσης και μερικές 10GbE ανερχόμενες ζεύξεις σε ένα ή περισσότερα επίπεδα του δικτύου, τα οποία συγκεντρώνουν και μεταφέρουν πακέτα μεταξύ των μεταγωγέων στο edge επίπεδο. Στα υψηλότερα επίπεδα της ιεραρχίας, υπάρχουν μεταγωγείς με 10 GigE θύρες (τυπικά 32-128) και αρκετή χωρητικότητα μεταγωγής για την συνάθροιση της κίνησης μεταξύ των edge μεταγωγέων.

Πολλά από τα κέντρα δεδομένων, ιδιαιτέρως τα μεγαλύτερα, εισάγουν τον όρο της υπερκάλυψης (**oversubscription**) ως μέσο μείωσης του συνολικού κόστος του σχεδιασμού τους. Ο όρος oversubscription ορίζεται ως η αναλογία του συναθροισμένου εύρου ζώνης στη χειρότερη αλλά εφικτή περίπτωση προς το συνολικό εύρος ζώνης μιας συγκεκριμένης τοπολογίας επικοινωνίας. Για παράδειγμα, μία υπερκάλυψη 1:1 δείχνει ότι όλοι οι hosts μπορούν δυνητικά να επικοινωνούν αυθαίρετα με άλλους hosts στο πλήρες εύρος ζώνης του δικτύου διεπαφής τους, ενώ μία υπερκάλυψη 5:1 σημαίνει ότι μόνο το 20% του διαθέσιμου εύρους ζώνης του host είναι διαθέσιμο, για ορισμένα πρότυπα επικοινωνίας. Οι τυπικές τιμές της υπερκάλυψης κυμαίνονται συνήθως από 2,5:1 έως και 8:1, ενώ όσο μεγαλύτερος είναι ο όγκος πληροφορίας ενός κέντρου δεδομένων τόσο μεγαλύτερη είναι και η αναλογία της υπερκάλυψης για να μην ανέβει ο προϋπολογισμός του κέντρου δεδομένων στα ύψη.

Τέλος θα γίνει μία αναφορά στη δρομολόγηση που γίνεται σε μία τέτοια αρχιτεκτονική. Η παροχή πλήρους εύρος ζώνης μεταξύ δύο hosts σε μεγάλες συστάδες απαιτεί ένα πολυδιάστατο δένδρο με πολλαπλούς core μεταγωγείς. Αυτό με τη σειρά του απαιτεί μια τεχνική δρομολόγησης πολλαπλών διαδρομών, όπως είναι η ECMP [12]. Επί του παρόντος, οι περισσότεροι core μεταγωγείς στο πρώτο επίπεδο του δένδρου υποστηρίζουν την τεχνική ECMP.

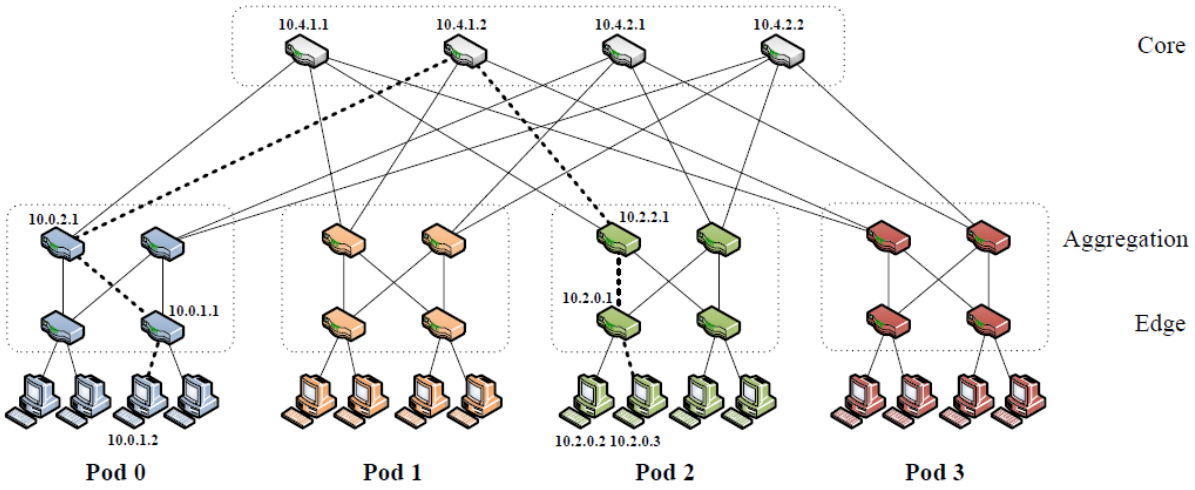
Για να επωφεληθεί από τις πιθανές πολλαπλές διαδρομές η ECMP εκτελεί διαχωρισμό στατικού φορτίου μεταξύ των ροών. Αυτό δεν περιλαμβάνει και την ροή του εύρους ζώνης στη λήψη των αποφάσεων κατανομής, η οποία θα μπορούσε να οδηγήσει σε περαιτέρω αύξηση της υπερκάλυψης ακόμη και για απλά σχέδια επικοινωνίας. Επιπλέον, οι τρέχουσες ECMP υλοποιήσεις περιορίζουν την πολλαπλότητα των διαδρομών σε αριθμό των 8-16, η οποίες είναι συχνά λιγότερο απαραίτητες για την παροχή υψηλού εύρους ζώνης για τα μεγαλύτερα κέντρα δεδομένων. Επιπλέον, ο αριθμός καταχωρήσεων του πίνακα δρομολόγησης μεγαλώνει πολλαπλασιαστικά με τον αριθμό των διαδρομών που προσδιορίζονται γεγονός που αυξάνει το κόστος και μπορεί επίσης να αυξήσει και την λανθάνουσα καθυστέρηση.

2.2.2 Fat trees με Clos topology

Σήμερα η διαφορά στη τιμή μεταξύ των απλών και φτηνών εμπορικών μεταγωγέων (commodity switches) και των ακριβότερων, που χρησιμοποιούνται στα ανώτερα επίπεδα των fat tree δομών όπως αναλύθηκε στην ενότητα 2.2.2, παρέχει ένα ισχυρό κίνητρο για τη κατασκευή δικτύων επικοινωνίας μεγάλης κλίμακας, αποτελούμενα από φτηνούς και απλούς μεταγωγείς παρά από τη χρήση μεγάλων και ακριβών. Το 1952 ο Charles Clos υποκινούμενος από παρόμοιες καταστάσεις στον τομέα των τηλεπικοινωνιών, σχεδίασε μία τοπολογία δικτύου η οποία προσφέρει υψηλά επίπεδα εύρους ζώνης για πολλές τερματικές συσκευές, με την κατάλληλη διασύνδεση μικότερων και φτηνότερων εμπορικών μεταγωγέων. Οι προδιαγραφές της τοπολογίας Clos αναλύθηκαν στην ενότητα 1.2.

Για τις περισσότερες υλοποιήσεις της Clos τοπολογίας με fat tree, ισχύουν τα ακόλουθα. Θεωρούμε την ύπαρξη και οργάνωση των διακομιστών σε συστάδες (pods) [12]. Ο συνολικός αριθμός των pods είναι k και το κάθε pod αποτελείται από 2 επίπεδα όπου κάθε επίπεδο εμπεριέχει k/2 μεταγωγείς. Ο καθε μεταγωγέας k-θυρών από το κατώτερο προαναφερόμενο επίπεδο εκ των δύο συνδέεται απευθείας με k/2 hosts. Οι υπόλοιπες k/2 θύρες που περισσεύουν στους μεταγωγείς του επιπέδου αυτού συνδέονται με k/2 από τις k θύρες των μεταγωγέων του

aggregation επιπέδου. Υπάρχουνε $(k/2)^2$ core μεταγωγείς k θυρών. Κάθε core μεταγωγέας έχει μία θύρα συνδεδεμένη σε καθένα από τα k pods. Μία i θύρα οποιουδήποτε core μεταγωγέα συνδέεται με το pod i με τέτοιο τρόπο έτσι ώστε οι διαδοχικές θύρες του κάθε μεταγωγέα στο aggregation επίπεδο να συνδέονται με τους core μεταγωγείς με k/2 βήματα-hops. Σε γενικές γραμμές μια τέτοια διάταξη δικτύου με μεταγωγείς k θυρών μπορεί να συντηρήσει k^3 / 4 hosts συνολικά. 'Ολα τα προαναφεθέντα απεικονίζονται στην επόμενη εικόνα όπου παρουσιάζουμε μία fat tree Clos τοπολογίας αρχιτεκτονική με k = 4:



Εικόνα 2. 2 - Clos τοπολογία με k=4

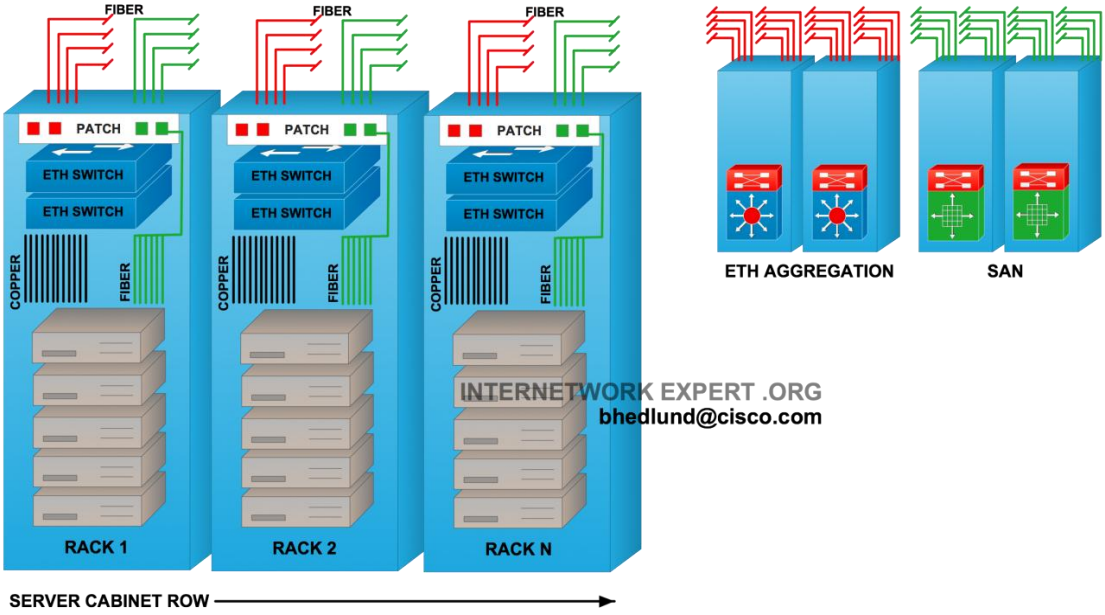
Ένα πλεονέκτημα της αρχιτεκτονικής αυτής είναι ότι όλα τα στοιχεία μεταγωγής είναι πανομοιότυπα, σε αντίθεση με άλλες υλοποιήσεις δένδρου, δίνοντας έτσι την δυνατότητα να αξιοποιηθούν μόνο φτηνοί και εμπορικοί μεταγωγείς. Επιπλέον τα fat trees είναι επαναδιευθετήσιμα δίκτυα χωρίς φραγή (rearrangeably non-blocking networks), πράγμα που σημαίνει ότι για αυθαίρετα πρότυπα επικοινωνίας υπάρχει κάποιο σύνολο διαδρομών το οποίο θα διαθέσει όλο το εύρος ζώνης στους hosts της τοπολογίας. Θεωρητικά μία τέτοια τοπολογία προσφέρει 1:1 λόγο υπερκάλυψης και πλήρες bisection εύρος ζώνης. Πρακτικά η επίτευξη του λόγου 1:1 μπορεί να είναι δύσκολη λόγω της ανάγκης να αποφευχθούν ανεπιθύμητες επαναδιατάξεις πακέτων για τις TCP ροές [12].

Τέλος είναι σημαντικό να αναφερθεί ότι η Clos τοπολογία χρησιμοποιεί σχεδόν πάντα μια εξοτομικευμένη διάταξη διευθυνσιοδότησης και διάφορους αλγορίθμους δρομολόγησης, πράγματα τα οποία διαφέρουν από δίκτυο σε δίκτυο ανάλογα με τις ανάγκες του κάθε κέντρου δεδομένων.

2.3 ToR Switch

Λύση σε πολλά από τα προαναφερθέντα έρχεται να δώσει το **ToR (top of the rack)** switch. Ένα ''top of the rack switch'' είναι ένας μικρός μεταγωγέας ο οποίος εξυπηρετεί ένα ικρίωμα (**rack**) και ο οποίος έχει ένα μικρό αριθμό θυρών έτσι ώστε να συνδέεται με την υπόλοιπη στοίβα [14].

Στην ToR αρχιτεκτονική οι διακομιστές ενός ικριώματος συνδέονται σε 1 ή 2 Ethernet μεταγωγείς που είναι εγκατεστημένοι μέσα στην στοίβα. Ο όρος ToR δεν είναι πάντα αντιπροσωπευτικός καθώς δεν είναι υποχρεωτικό ο μεταγωγέας να βρίσκεται στη κορυφή του rack. Η πραγματική τοποθεσία του μεταγωγέα μπορεί να είναι είτε στο τέλος του rack είτε στη μέση, απλώς το συνηθέστερο είναι στη κορυφή καθώς έτσι υπάρχει ευκολότερη προσβασιμότητα στον μεταγωγέα και ευκολότερη διαχείριση των καλωδίων. Το ToR συνήθως δεν έχει υψηλές απαιτήσεις (1 RU – 2 RU) και έχει σταθερή διαμόρφωση. Το κύριο χαρακτηριστικό αυτής της αρχιτεκτονικής είναι ότι όλες οι χάλκινες καλωδιώσεις για διακομιστές μένουν μέσα στο rack, η μορφή των οποίων είναι σχετικά κοντά RJ45 καλώδια από τους αντίστοιχους διακομιστές προς το ToR. Ο ToR μεταγωγέας ουσιαστικά συνδέει το rack στο δίκτυο του κέντρου δεδομένων με μία οπτική ίνα απευθείας από το rack σε μία περιοχή συνάθροισης [14]. Τα προαναφερθέντα απεικονίζονται στη παρακάτω εικόνα:



Εικόνα 2. 2 - ToR Switches μέσα σε racks

Κάθε rack συνδέεται στο κέντρο δεδομένων με οπτική ίνα όπως αναφέρθηκε. Συνεπώς δεν υπάρχει ανάγκη για μία πυκνή και ακριβή υποδομή καλωδίωσης (χαλκού) η οποία να περιπλέκει περισσότερο την κατάσταση. Μεγάλες ποσότητες χάλκινων καλωδίων προσθέτουν ένα επιπλέον βάρος στα κέντρα δεδομένων καθώς πολλά καλώδια μπορεί να περιπλέκονται με αποτέλεσμα να είναι δύσκολο να εντοπιστεί η συνδεσιμότητα τους, να εμποδίζεται η ροή του αέρα, να μην κλιματίζονται όσο πρέπει τα συστήματα και γενικά απαιτούνται και περισσότερα rack και άλλα είδη υποδομής τα οποία θα υπάρχουν μόνο για τη διαχείριση των καλωδίων [14]. Το πιθανό μπέρδεμα των καλωδίων επίσης μπορεί να δημιουργήσει περιορισμούς στη ταχύτητα με την οποία λειτουργεί ο διακομιστής αλλά και στη τεχνολογία του δικτύου. Με την αρχιτεκτονική ToR αποφεύγονται όλα τα προαναφερθέντα καθώς είναι εύκολο το συμπέρασμα ότι μια μεγάλη υποδομή χάλκινης καλωδίωσης είναι μη αποδοτική.

Κάθε rack μέσα στο κέντρο δεδομένων μπορεί να οργανωθεί και να διαχειριστεί σαν μία αυτόνομη μονάδα. Είναι έτσι πολύ ευκολότερο το να γίνει μία αλλαγή σε ένα rack ή και ακόμα

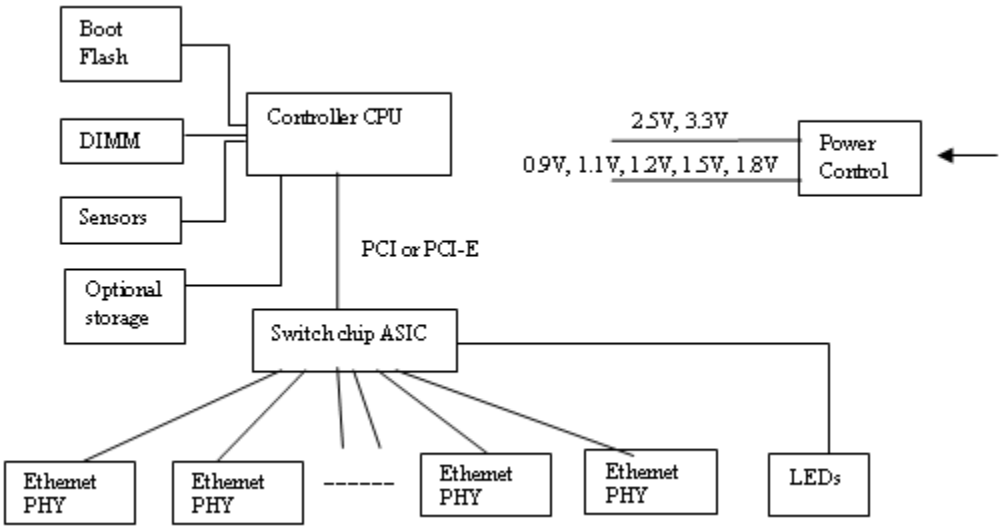
μία πιθανώς επιθυμητή αναβάθμιση της τεχνολογίας του διακομιστή. Οποιαδήποτε πάντως αλλαγή είτε αναβάθμιση μέσα στους μεταγωγείς του rack θα επηρεάζει μόνο τους διακομιστές του rack αυτού και όχι το σύνολο του δικτύου. Δεδομένου ότι ο διακομιστής συνδέεται με πολύ μικρά χάλκινα καλώδια μέσα στη στοίβα, υπάρχει μεγαλύτερη ευελιξία και περισσότερες επιλογές όσον αφορά την επιλογή των καλωδίων συνδεσιμότητας όσον αφορά τον τύπο ή την ταχύτητα της σύνδεσης που μπορεί να υποστηριχτεί. Για παράδειγμα, ένα 10 GBASE – CX1 χάλκινο καλώδιο μπορεί να χρησιμοποιηθεί για να παρέχει μια χαμηλού κόστους και χαμηλής ενεργειακής κατανάλωσης συνδεσιμότητα. Το συγκεκριμένο καλώδιο υποστηρίζει αποστάσεις μέχρι και 7 μέτρα το οποίο το καθιστά ιδανικό για μία ToR αρχιτεκτονική.

Η χρήση της οπτικής ίνας σε κάθε rack παρέχει πολύ καλύτερη ευελιξία και προστασία από ότι οι χάλκινες συνδέσεις εξαιτίας της μοναδικής ικανότητας της οπτικής ίνας να μεταφέρει υψηλότερο εύρους ζώνης σήματα σε πιο μακρινές αποστάσεις, λύνοντας έτσι εν μέρη ένα μεγάλο πρόβλημα των σημερινών κέντρων δεδομένων. Επίσης με την οικειοποίηση της οπτικής ίνας οι μελλοντικές μεταβάσεις του κέντρου δεδομένων σε 40 GE και σε 100 GE συνδεσιμότητες δικτύου θα υποστηρίζονται εύκολα σε μία τέτοια υποδομή. Με αυτά η λύση του ToR για τα κέντρα δεδομένων γίνεται πολύ ελκυστική καθώς οι σχεδιαστές δεν θα χρειάζεται να κάνουν ριζικές αλλαγές σε όλο το κέντρο όταν θα πρέπει να γίνουν οι προαναφερθείσες μεταβάσεις.

Ένα σημαντικό μειονέκτημα της ToR αρχιτεκτονικής είναι η αυξημένη ανάγκη στον τομέα διαχείρισης και διεύθυνσης, καθώς κάθε ToR μεταγωγέας θα αποτελεί μία επιπλέον μονάδα η οποία πρέπει να διαχειρίζεται ή ακόμα πιο σωστά μετατρέπεται σε μια επιπλέον μονάδα ελέγχου. Σε ένα μεγάλο κέντρο δεδομένων με πολλές στοίβες μία ToR αρχιτεκτονική μπορεί πολύ γρήγορα να δημιουργήσει πρόβλημα στη διαχείριση του κέντρου με την προσθήκη πολλών μεταγωγέων όπου θα πρέπει να διαχειρίζονται σε ατομικό επίπεδο. Για παράδειγμα σε ένα κέντρο δεδομένων με 40 racks όπου κάθε rack θα περιείχε 2 ToRs το αποτέλεσμα θα ήταν 80 μεταγωγείς οι οποίοι χρησιμοποιούνται για τη συνδεσιμότητα του κέντρου και τη πρόσβαση και επικοινωνία από μεταγωγέα σε μεταγωγέα. Αυτό σημαίνει 80 αντιγραφές του λογισμικού του μεταγωγέα οι οποίες θα πρέπει να αναβαθμίζονται, 80 αρχεία διαμόρφωσης τα οποία πρέπει να δημιουργηθούν και να αποθηκεύονται, 80 διαφορετικοί μεταγωγείς που συμμετέχουν στο δεύτερο επίπεδο ενός tree topology, θεωρώντας ότι αυτή η δομή επικρατεί τώρα στα κέντρα δεδομένων, καθώς και 80 επιπλέον μέρη στο κέντρο όπου μία ρύθμιση μπορεί να πάει στραβά. Όταν ένας ToR μεταγωγέας δεν λειτουργεί σωστά το υπεύθυνο άτομο που θα αντικαθιστά τον μεταγωγέα πρέπει να γνωρίζει πώς να αποκτήσει πρόσβαση και να αντικαταστήσει την αρχειοθετημένη ρύθμιση του μεταγωγέα αυτού. Επίσης θα είναι απαραίτητες μερικές δοκιμές επιβεβαίωσης ακόμα και δοκιμές του καινούργιου μεταγωγέα στο αν μπορεί να ανταποκριθεί σωστά σε απαιτητικές συνθήκες, όπως για παράδειγμα η ορθή λειτουργία του σε ένα υπερφορτωμένο δίκτυο. Όλα αυτά για την πραγματοποίηση τους χρειάζονται ένα αρκετά εκπαιδευμένο άτομο που να γνωρίζει τις διαδικασίες αυτές, πράγμα που δεν θα είναι πάντα διαθέσιμο. Επίσης δεν υπάρχουν τα συγκεκριμένα άτομα σε αφθονία ακόμη.

Ένα άλλο μειονέκτημα της ToR αρχιτεκτονικής είναι ότι απαιτούνται περισσότερες θύρες στους μεταγωγείς συνάθροισης. Επιστρέφοντας στο προαναφερθέν παράδειγμα μας, 80 ToR μεταγωγείς που θα συνδέονται με οπτική ίνα στο aggregation επίπεδο απαιτούν ότι ο κάθε μεταγωγέας συνάθροισης θα έχει 80 θύρες. Όσο περισσότερες θύρες υπάρχουν στους aggregation μεταγωγείς, τόσο περισσότερο πιθανό είναι να δημιουργηθούν περιορισμοί στην επεκτασιμότητα του κέντρου [14].

Όσο αφορά την αρχιτεκτονική του φυσικού μέρους (hardware design) του μεταγωγέα ToR είναι αυτή που απεικονίζεται στο παρακάτω σχήμα [15]:



Εικόνα 2. 3 – Αρχικτεκτονική ενός ToR switch

Ένας ελεγκτής χαμηλής ισχύος CPU. Αυτή η CPU χειρίζεται τα πρωτόκολλα από το πεδίο ελέγχου και είναι συνήθως ελαφριά φορτωμένη το μεγαλύτερο μέρος του χρόνου. Οι περισσότεροι ToR μεταγωγείς χρησιμοποιούνε Power PC SoC chip, ενώ άλλοι χρησιμοποιούνε ARM ή MIPS ως επεξεργαστή.

- Ένα ή πολλαπλά κυκλώματα ASIC διακόπτη για να χειριστεί την κυκλοφορία του πεδίου δεδομένων. Αυτά τα κυκλώματα ASIC διαχειρίζονται την ανταλλαγή πακέτων, τη δρομολόγησή τους καθώς και την ασφάλεια της κυκλοφορίας δεδομένων και φιλτράρισμα

-Ένα PCI ή ένα δίαυλο PCI-E για τη σύνδεση του ελεγκτή CPU και του διακόπτη ASIC. Αυτός ο δίαυλος έχει συνήθως πολύ χαμηλότερο εύρος ζώνης από τις θύρες δεδομένων στο ASIC. Χρησιμοποιείται για να περάσει το PDU (Protocol Data Unit, Μονάδες Δεδομένων Πρωτοκόλλου) μεταξύ της CPU και του ASIC. Να προστεθεί ότι δεν είναι αρκετά γρήγορο για να χειριστεί οποιαδήποτε κίνηση δεδομένων.

-Ethernet PHY. Το chip PHY μεταφράζει σειριακά σήματα σε διαφορετικά πρωτόκολλα των μέσων ενημέρωσης, όπως GBT, οπτικές ίνες, ή 10GbT. Το κάθε PHY μπορεί να έχει 1, 2, 4 ή ακόμα και 8 θύρες. Ανάλογα με τον αριθμό των θυρών στο διακόπτη και την ταχύτητα των θυρών, κάθε μεταγωγέας θα απαιτεί διαφορετικό αριθμό PHYs.

-Αποθήκευση. Υπάρχουν συνήθως δύο είδη αποθήκευσης, η ROM και ένα USB / Compact-Flash της κάρτας αποθήκευσης. Η ROM είναι συνήθως περίπου 32M ή 64M ενώ τα USB / CF μπορούν πλέον εύκολα να επεκταθούν μέχρι και σε δεκάδες GB [15].

Σε μερικές ειδικές περιπτώσεις παρατηρούνται και μικρές αλλαγές στη δομή της για να καλύψουνε τις αντίστοιχες ανάγκες.

Στη συνέχεια αποδίδονται τα κυριότερα πλεονεκτήματα και μειονεκτήματα της χρήσης του ToR switch στα κέντρα δεδομένων:

Πλεονεκτήματα:
- Οι χάλκινες καλωδιώσεις περιορίζονται στο rack. Δεν απαιτείται μεγάλη υποδομή χάλκινων καλωδίων. Η πολυπλοκότητα της καλωδίωσης ελαχιστοποιείται [14]
- Χαμηλότερο κόστος καλωδίωσης. Λιγότερη υποδομή αφιερωμένη στην καλωδίωση και σε αναβαθμίσεις λογισμικού. Καθαρότερη διαχείριση των καλωδίων.
- Ευέλικτη ανά ικρίωμα αρχιτεκτονική. Εύκολες αναβαθμίσεις / αλλαγές ανά rack. Όταν παρουσιαστεί πρόβλημα κάπου δεν απαιτούνται πολλές αλλαγές παρά μόνο στο ίδιο το rack.
- Αυτός ο σχεδιασμός επιτρέπει την επέκταση του κέντρου, διότι το δίκτυο μπορεί να τρέξει σε 1GE / 10GE σήμερα και μπορεί να αναβαθμιστεί για να τρέξει σε 10GE / 40GE στο μέλλον με το ελάχιστο κόστος και αλλαγές στην καλωδίωση.
- Μικρό μήκος καλωδίων χαλκού στους διακομιστές επιτρέπει χαμηλής ισχύος, χαμηλού κόστους 10GE (10GBASE-CX1), και 40G στο μέλλον.
- Αν τα racks είναι μικρά, θα μπορούσε να υπάρχει ένας διακόπτης δικτύου για 2-3 racks.
- Είναι υλοποιήσιμο με τον τεχνολογικό εξοπλισμό και τα μέσα που έχουμε σήμερα.

Μειονεκτήματα:
- Περισσότεροι μεταγωγείς απαιτούνται σε τέτοιες εγκαταστάσεις και κάθε μεταγωγέας πρέπει να αντιμετωπιστεί ανεξάρτητα. Έτσι, το κόστος επένδυσης και συντήρησης ενδέχεται να είναι υψηλότερο σε εγκαταστάσεις με ToR. Περισσότερες θύρες που απαιτούνται για το aggregation επίπεδο [14].
- Πιθανά προβλήματα επεκτασιμότητας (STP λογικές θύρες, μεγαλύτερο φυσικό μέγεθος των μεταγωγέων συνάθροισης)
- Πιθανώς να υπάρχουν αχρησιμοποίητες θύρες σε κάθε rack (καθώς οι μεταγωγείς έχουν σταθερές διαμορφώσεις και ο αριθμός των διακομιστών ποικίλλει) και είναι πολύ δύσκολο να υπολογιστεί με ακρίβεια ο απαιτούμενος αριθμός των θυρών. Αυτό οδηγεί σε αυξημένη αναξιοποίητη ισχύ, ψύξη και αριθμό θυρών.
- Μοναδικό επίπεδο ελέγχου ανά διακόπτη, υψηλότερο σύνολο ικανοτήτων που απαιτούνται από εκπαιδευμένα άτομα για την αντικατάσταση του διακόπτη.
- Σε πολύ σπάνιες περιπτώσεις μη προγραμματισμένες επεκτάσεις (μέσα σε ένα rack), μπορεί να είναι δύσκολο να επιτευχθούν χρησιμοποιώντας την προσέγγιση TOR.

2.4 Οι ανάγκες των κέντρων δεδομένων και οι περιορισμοί των αρχιτεκτονικών

Τα IT τμήματα (information technology) καθώς και οι σχεδιαστές δικτύων προετοιμάζουν τα κέντρα δεδομένων τους για το μέλλον, ενσωματώνοντας υποστήριξη για 10 Gigabit Ethernet (**10GE**, **10GbE**, or **10 GigE**) και μία ενοποιημένη δομή δικτύων στις στρατηγικές καλωδίωσής τους [16]. Δεδομένου ότι, ο κύκλος ζωής ενός τυπικού κέντρου δεδομένων είναι 10 έως 15 έτη, ο τρόπος με τον οποίο έχει γίνει η καλωδίωση στο κέντρο δεδομένων έχει τεράστια επίδραση στην ικανότητα του κέντρου να προσαρμοστεί σε τυχόν αλλαγές της αρχιτεκτονικής του δικτύου, σε αλλαγές που οφείλονται στην εξέλιξη της τεχνολογίας καθώς και στις ολοένα αυξανόμενες ανάγκες εύρους ζώνης. Οι αρχιτεκτονικές καλωδίωσης εάν δεν επιλεχτούν σωστά, θα μπορούσαν να αναγκάσουν μια πρόωρη αντικατάσταση ολόκληρης της υποδομής της καλωδίωσης έτσι ώστε να καλυφτούν οι απαιτήσεις συνδεσιμότητας. Αυτό μπορεί να

αποτελέσει ένα επαναλαμβανόμενο πρόβλημα καθώς το δίκτυο και οι τεχνολογίες πληροφορικής εξελίσσονται διαρκώς.

Τα σημερινά κέντρα δεδομένων εμπεριέχουν μια ποικιλία από μοντέλα καλωδίωσης και αρχιτεκτονικές. Με τη μετάβαση από το Gigabit Ethernet στο 10 Gigabit Ethernet, τόσο οι αρχιτεκτονικές καλωδίωσης όσο και οι ίδιες οι αρχιτεκτονικές των δικτύων επαναξιολογούνται ώστε να διασφαλιστεί μια οικονομικά αποδοτική και ομαλή αλλαγή του κέντρου δεδομένων [16]. Η επιλογή της αρχιτεκτονικής καλωδίωσης θα επηρεάσει την απόδοση, την επεκτασιμότητα του κέντρου, τη βιωσιμότητα του κύκλου ζωής του, τη βέλτιστη διαχείριση της ενέργειας σε αυτό, το συνολικό κόστος ιδιοκτησίας (total cost of ownership TCO) καθώς και την απόδοση της επένδυσης (return on investment ROI). Η πρόβλεψη της ανάπτυξης και των τεχνολογικών αλλαγών μπορεί να είναι δύσκολη, αλλά το κέντρο δεδομένων θα πρέπει να είναι σε θέση να ανταποκριθεί στην ανάπτυξη, στις αλλαγές στον εξοπλισμό, στα πρότυπα και τις απαιτήσεις ενώ παράλληλα να παραμένει διαχειρίσιμο και αξιόπιστο.

Άμεση συνέπεια από τα προαναφερθέντα είναι ότι η ‘’εικόνα’’ του κέντρου δεδομένων αλλάζει ραγδαία. Τα IT τμήματα είτε κατασκευάζουν νέα κέντρα δεδομένων, είτε σχεδιάζουν την επέκτασή τους, είτε αναβαθμίζουν τον εξοπλισμό των κέντρων, όλα πρέπει να σχεδιάσουν μια αρχιτεκτονική καλωδίωσης και μετάβασης η οποία να είναι ευέλικτη, να μπορεί να ανταπεξέλθει στις γρήγορες αλλαγές και κυρίως να μπορεί να υποστηρίξει μεταβάσεις σε 10, 40, και 100 Gigabit Ethernet, άμεσα στην πρώτη περίπτωση και στις δύο τελευταίες σε βάθος χρόνου. Οι κύριοι παράγοντες που τα IT departments πρέπει να αντιμετωπίσουν περιλαμβάνουν τα εξής:

- Η δομοστοιχείωση και η ευελιξία είναι υψίστης σημασίας. Η ανάγκη για την ταχεία ανάπτυξη νέων εφαρμογών καθώς και για την επέκταση των ήδη υπαρχόντων αρχιτεκτονικών έχει προκαλέσει την αντικατάσταση του ‘’server-at-a-time’’ μοντέλου με το rack-at-a-time μοντέλο. Με άλλα λόγια εξετάζεται το κέντρο δεδομένων με το rack ως βασική μονάδα. Πολλά τμήματα IT παραγγέλνουν ήδη προδιαμορφωμένα racks με ενσωματωμένη καλωδίωση και μεταγωγή μέχρι και για 96 διακομιστές ανά rack. Ο χρόνος που απαιτείται για την εγκατάσταση των καινούργιων rack και την επανενσωμάτωση των παλιών είναι πλέον ζήτημα ωρών αντί για ημέρες ή εβδομάδες. Επειδή διαφορετικά racks έχουν διαφορετικές I/O (input/ output) απαιτήσεις, οι στρατηγικές μεταγωγής και καλωδίωσης του κέντρου δεδομένων πρέπει να υποστηρίζουν μια ευρεία ποικιλία απαιτήσεων συνδεσιμότητας σε οποιαδήποτε θέση του rack [16].
- Ολοένα Αυξανόμενες απαιτήσεις εύρους ζώνης. Οι σημερινοί πολυπύρηνοι διακομιστές και όλοι οι ενσωματωμένοι διακομιστές σε μία στοίβα έχουν μεγαλύτερη επεξεργαστική ικανότητα με άμεσο αποτέλεσμα να χρειάζονται και περισσότερο εύρος ζώνης. Αρκετές στοίβες περιέχουν διακομιστές οι οποίοι απαιτούν για τη λειτουργία τους από 5 – 7 GigabitEthernet συνδέσεις καθώς και 2 FibreChannel συνδέσεις προς το δίκτυο περιοχής αποθήκευσης (storageareanetworkSAN) [16].
- Οι τρόποι συνδεσιμότητας εισόδου/εξόδου εξελίσσονται. Οι τρόποι συνδεσιμότητας εισόδου / εξόδου εξελίσσονται έτσι ώστε να καλύψουν την ανάγκη για το αυξημένο εύρος ζώνης και οι στρατηγικές καλωδίωσης και μεταγωγής πρέπει να μπορούν να συντηρούν όλες τις απαιτήσεις συνδεσιμότητας σε οποιαδήποτε θέση της στοίβας. Οι σημερινές επιλογές καλωδίωσης περιλαμβάνουν GE, 10 GE ή ακόμα και μία ενοποιημένη δομή δικτύου με Fibre channel over Ethernet (FCoE). Η πρόκληση με την οποία έρχονται αντιμέτωπα τα κέντρα δεδομένων είναι το πώς να υποστηρίζουν τη δόμηση και ευελιξία που χρειάζεται έτσι ώστε το κέντρο δεδομένων να παράγει αποδοτικό έργο και να μπορεί να χειριστεί το σημερινό φορτίο εργασίας των

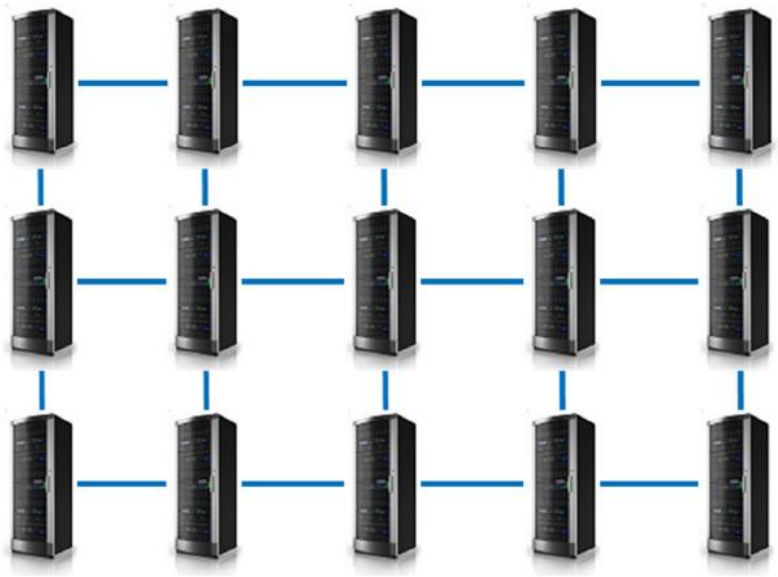
ζητούμενων εφαρμογών. Επίσης πρέπει να μπορεί να υποστηρίζει ένα μεγάλο εύρος για διαφορετικές συνδέσεις, GE, 10 GE, FCoE [16].

Οι δομές που επικρατούν στα κέντρα δεδομένων αυτή τη στιγμή αναλύθηκαν στην ενότητα 2.2. Στη συνέχεια θα εξηγήσουμε γιατί οι παραλλαγές της ιεραρχικής δομής του fat tree δεν μπορούν να καλύψουνε τις ανάγκες ενός κέντρου δεδομένων, οι οποίες προαναφέρθηκαν, αλλά και να επεκταθούν περαιτέρω.

Όσο αφορά τη παραδοσιακή εκδοχή του fat tree, όπου όσο ανεβαίνουμε επίπεδα στο δένδρο τόσο πιο απαιτητικοί και ακριβοί γίνονται οι μεταγωγείς, αυτή κρίνεται ως τελείως ακατάλληλη για τα μελλοντικά μεγάλα κέντρα δεδομένων, δεδομένου ότι η υποστήριξη πολλών hosts απαιτεί οι θύρες των μεταγωγέων του δένδρου στα ανώτερα επίπεδα και ειδικά στις ρίζες του δένδρου να έχουν τεράστιες απαιτήσεις σε εύρος ζώνης. Στο παράδειγμα που αναφέρθηκε στην ενότητα 2.2 όπου το κατώτερο επίπεδο ήτανε υπεύθυνο για τη διασύνδεση 2 hosts, για την σύνδεση σε ανώτερα επίπεδα απαιτούνταν εκθετική αύξηση του εύρους ζώνης. Αν αναλογιστούμε ότι το νούμερο των hosts είναι 10, τότε οι απαιτήσεις για το εύρος ζώνης στους μεταγωγείς στη ρίζα του δένδρου ξεφεύγουν υπερβολικά. Και για τα μεγάλα κέντρα δεδομένων ο αριθμός των 10 hosts δεν προσεγγίζει καν τη πραγματική εικόνα, όπου ανάλογα με το μέγεθος του κέντρου το κάθε rack περιέχει 20-40 διακομιστές. Επιπλέον, ένα δίκτυο που κατασκευάστηκε με αυτόν τον τρόπο δεν μπορεί να επεκταθεί και να υποστήριξει περισσότερους endhosts. Με βάση όλα τα προηγούμενα η δομή Clos κρίνεται ως αρκετά καταλληλότερη εάν αναλογιστεί κανείς το γεγονός ότι όλοι οι μεταγωγείς της είναι φτηνοί και δεν χρειάζονται τεράστια ποσά εύρους ζώνης, ακόμα και στις ρίζες του δένδρου. Σε μερικές περιπτώσεις, είναι αποδεκτές και αρχιτεκτονικές οι οποίες αποτελούν συνδυασμό και των 2 αυτών δομών, για παράδειγμα χρήση λίγο πιο ακριβών μεταγωγέων στις ρίζες του δένδρου.

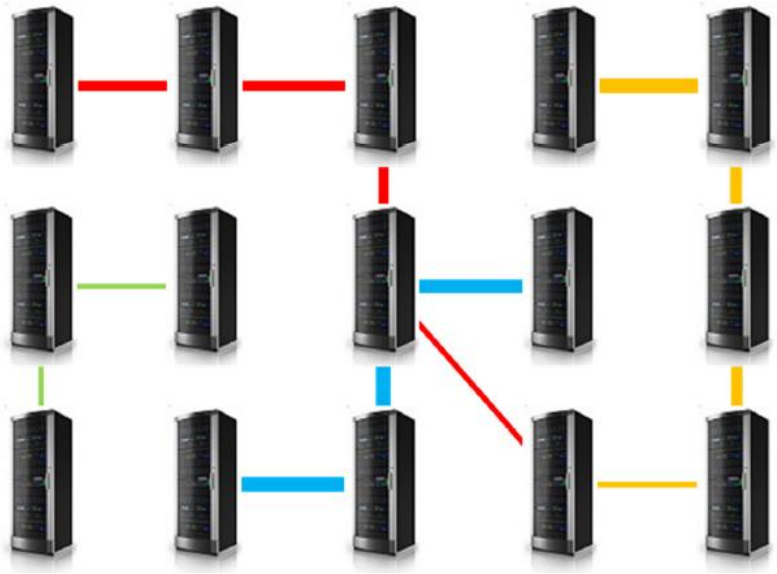
Ας υποθέσουμε μία fat tree υλοποίηση αρχιτεκτονικής. Έστω ότι ο αριθμός των hosts στην αρχιτεκτονική αυτή είναι ίσος με N και k είναι ο αριθμός των ηλεκτρικών μεταγωγέων πακέτων, τότε το βάθος του δένδρου είναι ίσο με $D = [(\log N - 1) / (\log k - 1)]$. Συνεπάγεται ότι η κλιμάκωση του αριθμού των endhosts σε μία τέτοια αρχιτεκτονική έρχεται με την απαίτηση να εγκαταστηθεί ένα πρόσθετο επίπεδο fat tree τοπολογίας κάθε φορά που συμπληρώνεται ένας αριθμός από διακομιστές και η αρχιτεκτονική φτάνει σε σημείο κορεσμού. Σήμερα τα κέντρα δεδομένων έχουν δεκάδες έως εκατοντάδες με χιλιάδες διακομιστές και απαιτούν 3 έως 4 επίπεδα fat tree για να επιτύχουν πλήρες εύρος ζώνης διχοτόμησης. Ως εκ τούτου, αν και το κόστος της Clos τοπολογίας είναι χαμηλότερο από την παραδοσιακή προσέγγιση fat tree, η επεκτασιμότητα της είναι ακόμη μη-γραμμική. Επιπλέον, η καλωδίωση ενός τέτοιου τεράστιου αριθμού μεταγωγέων γίνεται αρκετά περίπλοκη και είναι επιρρεπής σε σφάλματα κατά την εγκατάσταση και τη συντήρηση του κέντρου δεδομένων. Η χρήση ενός μεγάλου αριθμού ηλεκτρικών μεταγωγέων πακέτων συμβάλλει σημαντικά στην κατανάλωση ενέργειας του όλου συστήματος (10GbE μεταγωγείς 64 θυρών καταναλώνουν από 150 έως 350 Watt ο καθένας). Να σημειωθεί ότι περίπου το 90% αυτής της κατανάλωσης ενέργειας είναι ανεξάρτητη από το φορτίο του δικτύου και κατά συνέπεια εξοικονομήσεις είναι αδύνατον να προέρχονται από οποιαδήποτε εξισορρόπηση φορτίου ή κάποια μέθοδο χρονοπρογραμματισμού. Τέλος, η αναβάθμιση είναι ένα μεγάλο θέμα με τις αρχιτεκτονικές fat tree καθώς η προσθήκη περισσότερων racks-servers απαιτεί τη σύνδεση ελεύθερων θυρών πολλών και διάσπαρτων μεταγωγέων και αυτό υποθέτοντας ότι ελεύθερες θύρες υπάρχουν και είναι διαθέσιμες αλλά και η αναβάθμιση του ρυθμού επικοινωνίας των διακομιστών απαιτεί ένα πλήρες νέο δίκτυο fat tree και στις περισσότερες περιπτώσεις δεν μπορούν να επαναχρησιμοποιηθούν οι προηγούμενοι μεταγωγείς.

Στο σημείο αυτό πρέπει να γίνει λόγος για ένα άλλο μεγάλο μειονέκτημα των δομών αυτών. Τα fat tree υπο-χρησιμοποιούνται τις περισσότερες φορές και δεν ανταποκρίνονται στο κόστος κατασκευής και εγκατάστασης τους. Αυτό ισχύει γιατί οι σχεδιαστές δικτύων, σχεδιάζουνε το δίκτυο έτσι ώστε να μπορεί να ανταπεξέλθει στη χειρότερη δυνατή υπάρχουσα κατάσταση. Για παράδειγμα, ένας edge μεταγωγέας ο οποίος είναι υπεύθυνος για τη διασύνδεση 20-40 hosts στο υπόλοιπο δίκτυο, θα πρέπει να συνδέεται με το aggregation επίπεδο με ζεύξη της τάξεως 20-40 Gb/s αντιστοίχως (υποθέτουμε ότι κάθε host απαιτεί 1 Gb/s). Το σενάριο αυτό όμως είτε θα πραγματοποιηθεί σπάνια είτε ακόμα και ποτέ και έτσι για τη πρόληψη της χειρότερης περίπτωσης οι σχεδιαστές δικτύων είναι αναγκασμένοι να υλοποιήσουν ένα πολύ ακριβό δίκτυο το οποίο θα αξιοποίησει το 100% των δυνατοτήτων του σπάνια. Με αυτό τον τρόπο και αυξάνεται υπερβολικά το κόστος και υπάρχει κίνδυνος αύξησης του latency και μείωση της απόδοσης των εφαρμογών που θα εκτελεί. Στις επόμενες εικόνες απεικονίζεται αρχικά η ολική συνδεσιμότητα που υπάρχει μεταξύ των racks ενός κέντρου δεδομένων σήμερα και προνοοεί για τη χειρότερη περίπτωση ενώ στην δεύτερη απεικόνιση δίνεται ένα παράδειγμα για το τι πραγματικά χρειάζεται σε ένα δίκτυο [17]:



Εικόνα 2. 4 - Ολική συνδεσιμότητα των racks

Χωρητικότητα και συνδεσιμότητα όπου χρειάζεται αλλά και να είναι επεκτάσιμο και επαναδιαρθρώσιμο κατ’ αίτηση και όχι να σχεδιάζεται εξαρχής για τις πολύ σπάνιες χειρότερες δυνατές καταστάσεις:



Εικόνα 2. 5 - Η συνδεσιμότητα που όντως απαιτείται

Τέλος μία αναφορά στο θέμα της λανθάνουσας καθυστέρησης-latency που προαναφέρθηκε, το latency εξαρτάται από το φορτίο του συστήματος, δεδομένου ότι η συμφόρηση (**congestion**) προσθέτει καθυστερήσεις αναμονής και μπορεί να οδηγήσει σε πτώσεις πακέτων τα οποία διαχειρίζονται από τα ανώτερα στρώματα (TCP) ή την τρέχουσα εφαρμογή. Ακόμα και όταν είναι εγγυημένη η λειτουργία χωρίς απώλειες, η συμφόρηση εξακολουθεί να αυξάνει την λανθάνουσα καθυστέρηση. Τυπικές τιμές της λανθάνουσας καθυστέρησης ενός Ethernet μεταγωγέα είναι μεταξύ 500 nsec και 1,5 μsec [17].

Εν συντομία οι κύριοι περιορισμοί των ιεραρχικών αρχιτεκτονικών στα σύγχρονα κέντρα δεδομένων, είναι οι εξής [16][17]:

1) Η επεκτασιμότητα τους είναι μη-γραμμική και από ένα σημείο και μετά ίσως και να απαιτούνται περισσότερα επίπεδα-layers στο fat tree.
2) Υψηλά επίπεδα κατανάλωσης ενέργειας ανεξάρτητως του αν υπάρχει ενεργό φορτίο στο δίκτυο.
3) Πολύπλοκες μορφές καλωδίωσης και καθίσταται δύσκολη μια πιθανή επέμβαση για αλλαγή ή αναβάθμιση των καλωδίων.
4) Η παραδοσιακή fat tree υλοποίηση είναι υπερβολικά ακριβή αλλά και οποιαδήποτε άλλη παραλλαγή της δημιουργεί προβλήματα συμφόρησης στο δίκτυο.
5) Άκαμπτη κατανομή του εύρους ζώνης.
6) Το 75% της κίνησης σε ένα κέντρο δεδομένων είναι east-west μεταξύ των διακομιστών και όχι north-south όπως συμβαίνει στις fat tree αρχιτεκτονικές.
7) Οι δομές αυτές δεν θα μπορούν να συντηρήσουν τις μεταβάσεις του κέντρου δεδομένων στις ταχύτητες των 40 ή ακόμα και 100 Gb/s με ρεαλιστικό κόστος και κατανάλωση ενέργειας.
8) Τα πολύ μεγάλα κέντρα δεν θα μπορούν να εξηπηρετούν τον ολοένα αυξανόμενο αριθμό των racks τα οποία θα προστίθονται σε βάθος χρόνου. Αυτό έχει επίπτωση και στην απόδοση του

δικτύου αλλά και στην αύξηση του latency πράγμα που δεν ειναι επιθυμητό. Πολλές εφαρμογές επίσης είναι ευαίσθητες στο latency.

2.5 Η πρόκληση της αποσυνάθροισης (disaggregation)

Όλοι αυτοί οι περιορισμοί που αναλύθηκαν στην πρηγούμενη ενότητα 2.4 αλλά και η ανάγκη για ταχύτερα δίκτυα επιδεινώνονται περαιτέρω από την καινοτομική έννοια της αποσυνάθροισης των πόρων ενός δικτύου (**Resource Disaggregation**) [13]. Μέχρι στιγμής, στο συμβατικό μοντέλο του σημερινού κέντρου δεδομένων οι μονάδες υπολογισμού, μνήμης, αποθήκευσης και επικοινωνίας είναι προκαθορισμένες για τους διακομιστές που συνθέτουν τα κέντρα δεδομένων. Στην πραγματικότητα, κάθε διακομιστής αποτελείται από ένα σταθερό συνδυασμό των μονάδων υπολογισμού, μνήμης, αποθήκευσης και επικοινωνίας που όλα ενσωματώνονται, ή συναθροίζονται στον ίδιο έγκλειστο χώρο. Η βασική ιδέα της αποσυνάθροισης είναι ο διαχωρισμός των μονάδων αυτών και ο μοιρασμός τους μεταξύ των racks ενός κέντρου δεδομένων και η άμεση χρήση τους κατ᾽ αίτηση.

Υπάρχουν πολλαπλά οφέλη από τη μετάβαση σε αποσυναθροισμένα κέντρα δεδομένων. Η δομοστοιχείωση αυτή της υποδομής οδηγεί σε πιο αποδοτική λειτουργία και βελτιωμένη απόδοση. Επίσης αυτή η φυσική αποσύνδεση των πόρων επιτρέπει πιο αποδοτική παροχή στους πόρους ενός δικτύου καθώς και υψηλότερη χρησιμοποίηση με στατιστική πολυπλεξία των διαθέσιμων πόρων. Επιπλέον, η οργάνωση αυτή επιτρέπει επίσης ανεξάρτητη εξέλιξη ή αναβάθμιση του κάθε πόρου καθώς η κάθε διαφορετική κατηγορία πόρων ακολουθεί νέες τεχνολογικές εξελίξεις και τεχνολογικές τάσεις ή και περιορισμούς. Για παράδειγμα εάν χρειάζεται αναβάθμιση η μονάδα μνήμης ενός διακομιστή θα γίνει άμεση επέμβαση αναβάθμισης ή αλλαγής μονάχα στο κομμάτι της μνήμης και όχι ολόκληρου του διακομιστή. Έτσι θα υιοθετείται εύκολα το state-of-the-art σε οποιαδήποτε μεμονωμένη μονάδα ανεξάρτητα από τους άλλους πόρους και έτσι στο παράδειγμα που θίξαμε η μνήμη θα αναβαθμιστεί αφήνοντας τους υπόλοιπους πόρους του computing, αποθήκευσης και των υπόλοιπων στοιχείων του δικτύου ανεπηρέαστους και χωρίς περιττά έξοδα. Επιπλέον δίνεται η δυνατότητα στους σχεδιαστές υλισμικού να είναι πιο εύελικτοι και άρα να αναπτύσσουν καινοτομικές ιδέες.

Η ιδέα της αποσυνάθροισης υποστηρίζεται δυναμικά από εταιρίες όπως η Intel και το Facebook, οι οποίες είναι και υπεύθυνες για την πρωτοβουλία Open Compute Project (OCP). Η προτεινόμενη αρχιτεκτονική, που ονομάζεται Disaggregated Rack-Scale Server (DRS) προτείνεται για μεγάλα κέντρα δεδομένων. Ο στόχος του OCP-DRS είναι ο διαχωρισμός των υπολογιστικών, αποθηκευτικών και στοιχείων επικοινωνίας μέσα στο rack αλλά και και η διασύνδεση μεταξύ τους με ειδικές λειτουργίες μεταγωγής. Το OCP προβλέπει μείωση κατά 24% του κόστους και μια αύξηση της αποτελεσματικότητας περίπου της τάξεως 38% με την καινοτομική αυτή ιδέα της αποσυνάθροισης. Παρά το γεγονός ότι η ιδέα της αποσυνάθροισης είναι ακόμα σε πρώιμο στάδιο και μόλις πρόσφατα έγιναν οι πρώτες της υλοποιήσεις, οι προσδοκίες είναι ότι γρήγορα θα υιοθετηθεί από πολλά μεγάλα κέντρα δεδομένων.

Παρά τα μοναδικά οφέλη της αποσυνάθροισης, από την άποψη της βέλτιστης τροφοδότησης των πόρων και αναβάθμισης των υποδομών, αυτή έρχεται με ένα μεγάλο εμπόδιο: Ο διαχωρισμός των πόρων του συστήματος προκαλεί υπέρμετρες απαιτήσεις για το δίκτυο το οποίο θα διασυνδέει αυτές τις διαχωρισμένες μονάδες. Ως αποτέλεσμα, οι διασυνδέσεις του δικτύου θα αντιμετωπίζουν την πρόκληση να καλύψουν τις αρκετά αυξημένες απαιτήσεις σε εύρος ζώνης και χαμηλή λανθάνουσα καθυστέρηση κατά μήκος όλης της φυσικής απόστασης

των κατανεμημένων στοιχείων στο κέντρο δεδομένων. Οι υπάρχουσες υλοποιήσεις που βασίζονται στις fat tree δομές και σε απλούς μεταγωγείς αυτής της τεχνολογικής γενιάς δεν μπορεί να επεκταθούν για να υποστηρίξουν τις απαιτήσεις της ολοένα αυξανόμενης κίνησης μέσα στο κέντρο δεδμένων. Για τη μεταφορά δεδομένων μεγάλου εύρους ζώνης συνίστανται υλοποιήσεις με οπτικά στοιχεία οι οποίες είναι και πιο γρήγορες και επεκτάσιμες. Για ακόμα μία φορά τα οπτικά δίκτυα εμφανίζονται ως η τεχνολογία η οποία μπορεί να υλοποιήσει τον διαχωρισμό αυτό. Χρησιμοποιώντας τα οπτικά στοιχεία όχι μόνο ως στοιχεία προώθησης αλλά και ως στοιχεία μεταγωγής μπορεί να ελαφρύνει το κέντρο δεδομένων από τους τρομερά μεγάλους ηλεκτρονικούς μεταγωγείς οι οποίοι καταναλώνουν πολύ μεγάλη ισχύ και δεν παράγουν την αντίστοιχη αποδοτικότητα.

Από όλα τα παραπάνω καταλήγουμε στο συμπέρασμα ότι υπάρχει ανάγκη για καλύτερα και πιο αποδοτικά δίκτυα και εδώ προτείνεται η χρήση οπτικής τεχνολογίας όχι μόνο για τη μεταφορά της πληροφορίας αλλά και για τη δρομολόγησή της με γρηγορότερο και λιγότερο ενεργοβόρο τρόπο. Η ιδέα δεν είναι να αντικαταστηθούν πλήρως τα ηλεκτρονικά στοιχεία και μεταγωγείς με οπτικά αλλά να χρησιμοποιηθούν μαζί και με αυτό τον τρόπο να προσφέρουν στο δίκτυο τα θετικά και των δύο. Να σημειωθεί σε αυτό το σημείο ότι υπάρχει και ένας μικρός αριθμός από προτάσεις με μόνο οπτικά στοιχεία. Στην ενότητα που ακολουθεί γίνεται η παρουσίαση και η ανάλυση των κυριότερων αρχιτεκτονικών.

Κεφάλαιο 3
Επεκτάσιμες πρόσφατες αρχιτεκτονικές οπτικής σε κέντρα δεδομένων

Στο τρίτο κεφάλαιο της εργασίας προχωρούμε στη μελέτη και παρουσίαση μερικών νέων αρχιτεκτονικών δικτύων που βασίζονται στην (υβριδική) οπτική δρομολόγηση οι οποίες είτε χρησιμοποιούνται ήδη από κάποιακέντρα δεδομένων είτε είναι σε πειραματικό στάδιο. Στο κεφάλαιο αυτό ερευνούμε και παρουσιάζουμε τις τεχνολογίες από τις οποίες απαρτίζονται αλλά και διασαφηνίζουμε τις υπηρεσίες που προσφέρουν. Πιο συγκεκριμένα παρουσιάζουμε τις διάφορες δομές δικτύων με τις πιθανές παραλλαγές τους σε δομικά επίπεδα και επεξηγούμε τον τρόπο λειτουργίας τους. Οι αρχιτεκτονικές που εξετάζονται στη παρούσα εργασία είναι οι ακόλουθες: Plexxi, MEMS-based, Arrayed Wavelength-Grating (AWG) based και WSSbased.

3.1 Δομές δικτύου Plexxi

Ξεκινάμε την περιγραφή αρχιτεκτονικών οπτικής από το δίκτυο Plexxi, μία δομή η οποία έχει ήδη υιοθετηθεί από ορισμένα κέντρα δεδομένων και είναι διαθέσιμη στην αγορά εδώ και 3 χρόνια καθώς όλη η τεχνολογία που απαιτεί είναι διαθέσιμη σήμερα [18]. Για τη μελέτη της αρχιτεκτονικής Plexxi θα γίνει αρχικά η παρουσίαση της δομής της και στη συνέχεια ακολουθεί η αναφορά και παρουσίαση των οπτικών διακοπτών που χρησιμοποιεί. Ο πρώτος διακόπτης ονομάζεται plexxi switch 1 και στη συνέχεια μετά από ορισμένες τροποποιήσεις και τεχνολογικές αναβαθμίσεις του υλικού και του λογισμικού προέκυψε ο plexxi switch 2 [19].

3.1.1 Αναλυτική Παρουσίαση αρχιτεκτονικής Plexxi

Σε αυτό το στάδιο θα ακολουθήσει η περιγραφή και παρουσίαση ενός δικτύου Plexxi. Οι μεταγωγείς Plexxi είναι 10GbE Ethernet μεταγωγείς. Οι μεταγωγείς είναι βασισμένοι σε εμπορική σιλικόνη, με όλα τα θετικά και αρνητικά που συνεπάγεται αυτό, κυρίως στον τομέα προώθησης των πακέτων του μεταγωγέα (packet forwarding sector) [20]. Οι θύρες σε ένα μεταγωγέα Plexxi χωρίζονται σε 2 ομάδες: θύρες πρόσβασης (Access ports) και lightrail θύρες. Οι θύρες πρόσβασης είναι 10GbE θύρες οι οποίες χρησιμοποιούνται για να συνδέουν τους διακομιστές, τη μνήμη – αποθήκευση, άλλους δρομολογητές, διακόπτες για επικοινωνία με άλλα δίκτυα ή οποιαδήποτε άλλου είδους ανάγκη απαιτεί το κέντρο δεδομένων. Ανάλογα με την ακριβή έκδοση του μεταγωγέα μπορεί να υπάρχουν έως και 72 από αυτούς. Οι lightrail θύρες είναι επίσης 10GbE θύρες αλλά χρησιμοποιούνται για να δημιουργήσουν το πλέγμα του δικτύου Plexxi, δηλαδή μέσω αυτών των θυρών συνδέονται οι μεταγωγείς Plexxi με τους γειτονικούς τους μεταγωγείς στο δακτύλιο. Και πάλι, ανάλογα με την ακριβή έκδοση του μεταγωγέα, υπάρχουν τουλάχιστον 24 από αυτές τις θύρες, έως και 48 θύρες στους πιο πρόσφατους μεταγωγείς. Τα lightrail είναι υψηλής πυκνότητας καλώδια που χρησιμοποιούνται από το δίκτυο Plexxi ως διεπαφές σύνδεσης. Αυτές οι θύρες lightrail δεν εμφανίζονται ως κανονικές SFP + ή QSFP + στο διακόπτη, αλλά μεταφέρονται χρησιμοποιώντας Wavelength Division Multiplexing (WDM) μέσω των υποδοχών lightrail. Η χρήση του WDM επιτρέπει τη μεταφορά πολλαπλών συνδέσεων 10GbE σε μια ενιαία ίνα lightrail χρησιμοποιώντας διαφορετικά μήκη κύματος για

κάθε σύνδεση. Με αυτό τον τρόπο τα καλώδια lightrail χαρακτηρίζονται από υψηλή χωρητικότητα και ένα καλώδιο lightrail μπορεί να αντικαταστήσει φυσικά μέχρι και 12 κανονικά καλώδια. Στην επόμενη εικόνα απεικονίζεται ένας μεταγωγέας Plexxi με τα lightrail καλώδια του τα οποία εκτείνονται λογικά αριστερά και δεξιά [20]:

Εικόνα 3. 1 – Συνδέσεις ενός Plexxi Switch με lightrail

Οι μεταγωγείς Plexxi συνδέονται μεταξύ τους και δημιουργούν ένα δακτύλιο (ring topology). Με αυτό τον τρόπο, κάθε διακόπτης έχει ένα καλώδιο lightrail το οποίο στρέφεται λογικά αριστερά (logical East) και ένα ακόμα καλώδιο προς τα λογικά δεξιά (logical West). Κάθε καλώδιο lightrail μεταφέρει 12 από τις θύρες 10GbE lightrail προς την αντίστοιχη κατεύθυνση. Παρά το γεγονός ότι τα καλώδια lightrail συνδέονται μεταξύ τους σαν ένα λογικό δαχτυλίδι (logical ring topology) τα μήκη κύματος 10GbE δεν τερματίζουν αναγκαστικά σε όλους τους γειτονικούς μεταγωγείς. Εξαιτίας της χρήσης του WDM, πολλά από τα μήκη κύματος περνούν από το γειτονικό μεταγωγέα, χρησιμοποιώντας οπτική τεχνολογία, και προωθούνται σε μεταγωγείς που είναι 2, 3, 4 ή ακόμα και 5 μεταγωγείς λογικά αριστερά ή λογικά δεξιά του αρχικού μεταγωγέα που τα μεταδίδει (transmitting switch). Με τη χρήση τεχνολογίας παθητικών στοιχείων διέλευσης, τα μήκη κύματος μπορούν να προωθηθούν ακόμα και αν ένας ενδιάμεσος διακόπτης είναι κλειστός, σε κατάσταση επανεκκίνησης ή ακόμα και είναι υπερφορτωμένος λόγω αυξημένης κίνησης.

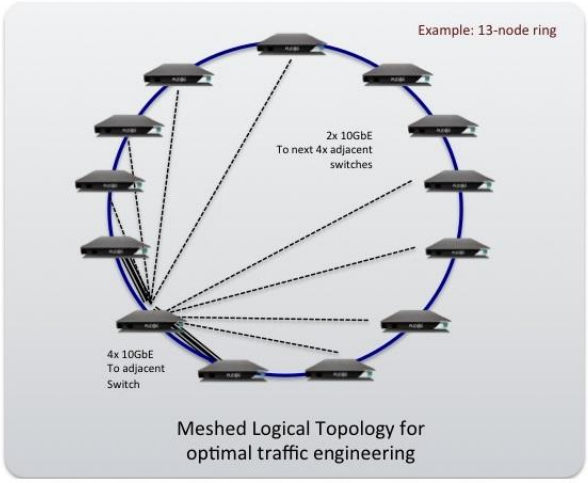
Κάθε μεταγωγέας παίρνει ένα σύνολο από τα μήκη κύματος 10GbE WDM και τα διαβιβάζει παθητικά αλλά μερικά από αυτά τα μήκη κύματος τερματίζονται σε αυτόν τον μεταγωγέα προσαρτώντας τα με το Ethernet πρωτόκολλο μεταγωγής ASIC. Με αυτό τον τρόπο οι μεταγωγείς έχουν δημιουργήσει μία δισημειακή 10GbE σύνδεση (10GbE **point to point connection**) Ethernet, μεταξύ του διακόπτη από όπου προέρχεται αυτό το κύμα 10GbE, και του διακόπτη που το τερματίζει. Κατά την σύνδεση των μεταγωγέων Plexxi μεταξύ τους δεν απαιτείται κάποια συγκεκριμένη ρύθμιση για να επιτευχθεί αυτό, οι μεταγωγείς αυτομάτως θα δημιουργήσουν μία προεπιλεγμένη τοπολογία. Αυτή η προεπιλεγμένη ρύθμιση δημιουργεί 4 10GbE σημείο σε σημείο συνδέσεις μεταξύ οποιωνδήποτε δύο γειτονικών μεταγωγέων και 2 10GbE συνδέσεις μεταξύ οποιωνδήποτε δύο μεταγωγέων που είναι 2, 3, 4 ή 5 μεταγωγείς απομακρυσμένοι μεταξύ τους.

Το σύνολο όλων αυτών των σημείο σε σημείο συνδέσεων δημιουργεί μία μερικώς ή πλήρως κατανεμημένη τοπολογία (full or partial mesh) μεταξύ όλων των μεταγωγέων οι οποίοι συνδέονται σε ένα δακτύλιο. Οδηγούμαστε στο συμπέρασμα ότι ενώ η φυσική τοπολογία του δικτύου Plexxi,δηλαδή το πώς είναι συνδεδεμένο το υλικό (hardware), είναι ένας δακτύλιος

(ring physical topology) καθώς συνδέονται μόνο οι γειτονικοί μεταγωγείς ενώ η λογική τοπολογία του δικτύου, δηλαδή ο τρόπος που λειτουργεί, είναι μία κατανεμημένη τοπολογία με τον περιορισμό που αναφέρθηκε στην προηγούμενη παράγραφο.

Χωρίς να απαιτείται κάποια ρύθμιση, οι μεταγωγείς χρησιμοποιούν έναν μηχανισμό αναζήτησης για να καθορίσει ποιος μεταγωγέας είναι στην άλλη πλευρά της κάθε σημείο σε σημείο σύνδεσης. Ακόμα κι αν αυτή είναι η προεπιλεγμένη τοπολογία τους, οι μεταγωγείς δεν προβαίνουν σε υποθέσεις για το τι μεταγωγέας είναι στην άλλη πλευρά, αντιθέτως ενεργά ανιχνεύουν και ανακαλύπτουν. Έτσι ως αποτέλεσμα κάθε μεταγωγέας θα έχει 24 ή 48 (ή στην πραγματικότητα οποιονδήποτε αριθμό απαιτείται) συνδέσεις προς τους υπόλοιπους μεταγωγείς του δικτύου. Ο συνδυασμός όλων αυτών των συνδέσεων από όλους τους μεταγωγείς, είναι η προεπιλεγμένη τοπογραφία για αυτό το δίκτυο. Ο αριθμός των τρόπων σύνδεσης είναι πολύ μεγάλος,για παράδειγμα με 20 μεταγωγείς σε ένα δίκτυο Plexxi, υπάρχουν τουλάχιστον (20 * 24) / 2 = 240 10GbE συνδέσεις για τους μεταγωγείς (μερικώς κατανεμημένη τοπολογία).

Σε ένα δίκτυο Plexxi, ο αριθμός των διαφορετικών τρόπων για να επικοινωνήσουν 2 μεταγωγείς είναι τεράστιος. Ας υποθέσουμε για παράδειγμα ένα δίκτυο Plexxi με 11 μεταγωγείς και ότι ο μεταγωγέας 1 θέλει να επικοινωνήσει με τον μεταγωγέα 5. Οι μεταγωγείς 1 και 5 έχουνε 2 άμεσες σημείο σε σημείο 10GbE συνδέσεις. Υποθέτοντας όμως ότι το μονοπάτι αυτό χρησιμοποιείται ήδη ή ότι υπάρχει ήδη αρκετή κίνηση, ο μεταγωγέας 1 έχει 4 10GbE με τους γειτονικούς του, δηλαδή τον μεταγωγέα 2 και 11. Με τη σειρά τους οι μεταγωγείς 2 και 11 έχουν ο καθένας 2 10GbE συνδέσεις με τον μεταγωγέα 5 δημιουργώντας έτσι 16 έμμεσα μοναδικά 10GbE μονοπάτια μεταξύ του μεταγωγέα 1 και 5. Ανάλογα με το μέγεθος του δικτύου, σε ένα μεγάλο κέντρο δεδομένων θα μπορούσαν να υπάρχουν εκατοντάδες διαφορετικά μονοπάτια μεταξύ 2 μεταγωγέων, μερικά άμεσα και τα περισσότερα έμμεσα με πολλές αναπηδήσεις (**hops**) ανάμεσα στους μεταγωγείς του δακτυλίου [20].



Εικόνα 3. 2 – Φυσική τοπολογία ενός Plexxi Ring

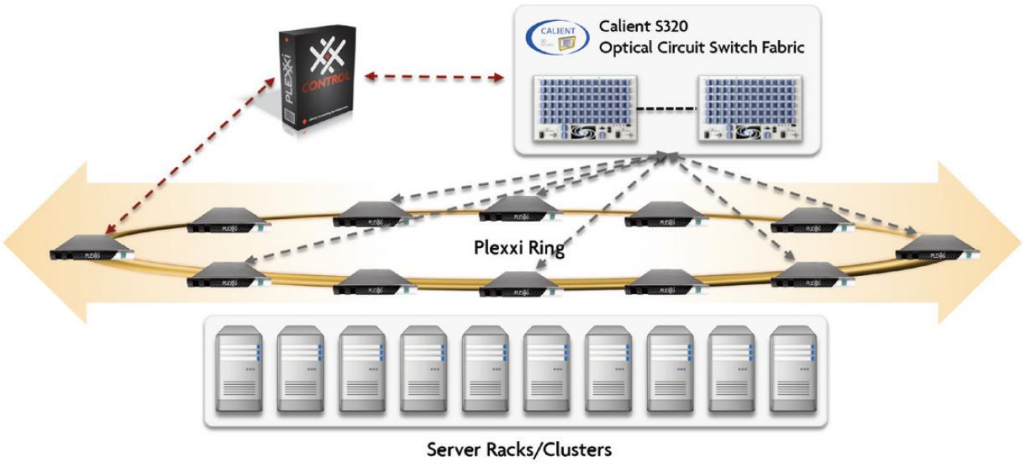
Τέλος, όταν υπάρχουν τόσες πολλές διαδρομές μεταξύ 2 μεταγωγέων στο δίκτυο, πρέπει το δίκτυο να είναι εξαιρετικά επιλεκτικό στην επιλογή των διαδρομών που θα επιλέγονται για τα πακέτα έτσι ώστε να μην δημιουργείται συμφόρηση. Ο τομέας ελέγχου Plexxi (Plexxi Control) χρησιμοποιεί τις συγγένειες (**affinities**) για να υπολογίσει ποια κίνηση θα πρέπει να πάει που ώστε να λειτουργεί το δίκτυο ορθά.

3.1.2 Υβριδική αρχιτεκτονική Plexxi με οπτικούς μεταγωγείς Calient

Σε αυτό το σημείο θα ακολουθήσει η παρουσίαση της υβριδικής αρχιτεκτονικής Plexxi and Calient [21]. Το δίκτυο plexxi παρουσιάστηκε και αναλύθηκε στην προηγούμενη ενότητα.

Η Calient παρασκευάζει υψηλής πυκνότητας οπτικούς μεταγωγείς οι οποίοι χαρακτηρίζονται και από χαμηλή λανθάνουσα καθυστέρηση (latency). Στην παρούσα αρχιτεκτονική γίνεται χρήση του μεταγωγέα S320 Photonic Switch της Calient, ο οποίος χαρακτηρίζεται από υψηλή αξιοπιστία και ποιότητα λειτουργίας, χαμηλή κατανάλωση ενέργειας και ευκολία στη χρήση. Τα παραπάνω χαρακτηριστικά του τον καθιστούν ιδανικό για την αρχιτεκτονική αυτή και φέρνει τα πραγματικά πλεονεκτήματα της φωτονικής μεταγωγής στα κέντρα δεδομένων.

Ο μεταγωγέας S320 Photonic Switch καθιστά δυνατή την σύνθεση μιας υβριδικής αρχιτεκτονικής, ηλεκτρικών πακέτων και οπτικής (**hybrid packet-optical architecture)** η οποία μπορεί να αναπτυχθεί κλιμακωτά ώστε να υποστηρίξει την απαιτούμενη ανάπτυξη των κέντρων δεδομένων [21]. Αυτή η αρχιτεκτονική εισάγει επιπλέον οπτική χωρητικότητα και επιτρέπει στο δίκτυο να αναδιαρθρώνεται δυναμικά. Στη παρακάτω εικόνα απεικονίζεται η υβριδική αρχιτεκτονική:



Εικόνα 3. 3 - Αναπαράσταση Plexxi - Calient υβριδικής αρχιτεκτονικής

Στο κέντρο της υβριδικής packet-optical αρχιτεκτονικής παραμένει το ηλεκτρικό κύκλωμα των πακέτων. Το κύκλωμα αυτό παρέχει ολική συνδεσιμότητα μεταξύ οποιουδήποτε rack ή τομέα-rack στο κέντρο δεδομένων. Παράλληλα με το δίκτυο πακέτων η υβριδική αρχιτεκτονική περιέχει ένα οπτικό κύκλωμα μεταγωγής (**optical switch circuit OSC**). Ο κάθε ToR μεταγωγέας στο δίκτυο πακέτων συνδέεται με το οπτικό δικτυοδόμημα (fabric). Το κλειδί που απαιτείται ώστε να λειτουργεί η υβριδική αρχιτεκτονική είναι ότι τα ηλεκτρικά και τα οπτικά στοιχεία

πρέπει να συνυπάρχουν και να συντονίζονται από υψηλότερου επιπέδου στρώματα διαχείρισης. Αυτό καθίσταται δυνατό με τη μονάδα ελέγχου Plexxi και σε αυτή την αρχιτεκτονική η οποία χρησιμοποιεί SDN.

Στην υβριδική αρχιτεκτονική τμήματα του Plexxi ring συνδέονται άμεσα με το Calient οπτικό δικτυοδόμημα μέσω 10GbE ή 40GbE θυρών [21]. Οι μεταγωγείς Plexxi συνδέονται μεταξύ τους μέσω των καλωδίων lightrail. Όταν υπάρχει μία μεγάλη ροή δεδομένων ή εφαρμογών για να καταπολεμηθεί η κίνηση στο δίκτυο οι Plexxi διακόπτες περνάνε την μεταφορά των πακέτων προς το διακόπτη εισόδου του Calient και στη συνέχεια το οπτικό δικτυοδόμημα προωθεί τη κίνηση στο διακόπτη εξόδου, το οποίο τελικά στέλνει στον προορισμό τους τα πακέτα. Με αυτό τον τρόπο επιτυγχάνεται η προστασία του δικτύου από την κυκλοφοριακή συμφόρηση των πακέτων και εγγυάται μία υψηλού εύρους ζώνης, χαμηλής καθυστέρησης διαδρομή για τη ροή. Η αρχιτεκτονική αυτή είναι αρκετά παρόμοια με αυτή των 3D MEMS η οποία θα αναπτυχθεί σε επόμενη ενότητα.

Για να συνεργαστούν τα δύο συστήματα, πρέπει να υπάρχει ένα μέσο ανταλλαγής πληροφοριών. Η ανταλλαγή πληροφοριών μεταξύ ενός δικτύου Plexxi και μίας άλλης υποδομής γίνεται μέσω των Plexxi connectors. Ένας Plexxi connector είναι μία περιορισμένη υπηρεσία συλλογής δεδομένων που μπορεί να βρίσκεται είτε στο εσωτερικό του ελεγκτή Plexxi είτε οπουδήποτε μέσα ή γύρω από το δίκτυο με μοναδικό περιορισμό να έχει πρόσβαση εκεί ο Plexxi controller.

3.1.3 Plexxi switch 1

Εικόνα 3. 4 – Μεταγωγέας Plexxi

Ο διακόπτης Plexxi 1 παρέχει μια καινοτόμο λύση,στον τομέα των δικτύων, σχεδιασμένη για να καλύψει τις αυξανόμενες ανάγκες των σημερινών κέντρων δεδομένων που όπως προαναφέρθηκε αντιμετωπίζουν μία πληθώρα προβλημάτων. Η προηγμένη αρχιτεκτονική του διακόπτη ενοποιεί ένα Ethernet μεταγωγέα μαζί με έναν πρωτοποριακό κεντρικό και ομόσπονδο μηχανισμό ελέγχου ο οποίος είναι SDN βασισμένος καθώς και με ένα μοναδικό οπτικό τρόπο διασύνδεσης με πολυπλεξία. Αυτό αντικαθιστά τις παραδοσιακές ιεραρχικές δομές με ένα κλιμακούμενο, υψηλού εύρους ζώνης, χαμηλής καθυστέρησης μεταφοράς δεδομένων και προγραμματιστικά προσαρμόσιμο Εικονικό Πολυπύρηνο (Virtual Multicore) δίκτυο και επιτρέπει την συγγενική δικτύωση (affinity networking), το νέο πρότυπο στις υποδομές κέντρων δεδομένων [18].

Ο κάθε μεταγωγέας Plexxi 1 διαθέτει δύο lightrail οπτικές διεπαφές σύνδεσης, παρέχοντας έως και 240 Gbps εύρος ζώνης για επικοινωνία εξ ολοκλήρου διπλής κατεύθυνσης, δημιουργώντας έτσι επεκτάσιμα και προγραμματιζόμενα δίκτυα με πλήρως κατανεμημένη τοπολογία (**full meshed - meshed networks**). Οι Plexxi μεταγωγείς χρησιμοποιούνε διαίρεση μήκους κύματος πολυπλεξίας (**wavelength division multiplexing – WDM**) και οπτική cross connect τεχνολογία έτσι ώστε να δημιουργήσουν ένα οπτικό δίκτυο επιπέδου 1,το οποίο διαθέτει πλήρως κατανεμημένη τοπολογία (mesh) μεταξύ των διακοπτών σε ένα δαχτυλίδι Plexxi και είναι εντελώς ελεγχόμενο από λογισμικό. Ένα cross-connect είναι οποιαδήποτε σύνδεση μεταξύ των εγκαταστάσεων που παρέχονται ως ξεχωριστές μονάδες από το κέντρο δεδομένων. Η καλωδίωση από μεταγωγέα σε μεταγωγέα απλοποιείται σε μεγάλο βαθμό με μόνο 2 συνδέσεις ανά μεταγωγέα.

-Virtual Multicore

Το πλήρες δυναμικό της οπτικής μεταγωγής γίνεται εμφανές από το Virtual multicore της αρχιτεκτονικής Plexxi το οποίο παρέχει αποτελεσματικές τοπολογίες δικτύου στις κρίσιμες και φορτωμένες απαιτήσεις των εφαρμογών. Η οπτική τεχνολογία πολυπλεξίας δημιουργεί ένα πλέγμα δικτύου πολλαπλών διαδρομών με πολλές άμεσες και έμμεσες διαδρομές μεταξύ των διακοπτών σε ένα δίκτυο Plexxi. Η οργάνωση αυτών των διαδρομών διαχειρίζεται από το λογισμικό της μονάδας ελέγχου Plexxi η οποία κατανοεί συγκεκριμένα τη σχέση μεταξύ των πόρων που περιλαμβάνουν ένα φόρτο απαιτήσεων και τις ειδικές ανάγκες ή τους περιορισμούς μεταξύ των πόρων αυτών ή με άλλους εξωτερικούς παράγοντες. Το αποτέλεσμα είναι ένα δίκτυο όπου κάθε μονάδα η οποία έχει φόρτο απαιτήσεων λαμβάνει το δικό της εικονικό πυρήνα του δικτύου (virtual multicore).

-Affinity Smartpath

Το Affinity- SmartPath είναι μια έξυπνη και προσαρμοστική τεχνολογία που εξασφαλίζει ότι ο φόρτος αιτήσεων παίρνει πάντα πρόσβαση στις πιο βέλτιστες και διαθέσιμες διαδρομές δικτύου. Το Affinity SmartPath επιλέγει έξυπνα το καλύτερο μονοπάτι του δικτύου για αιτήσεις που έχουν σαφείς περιορισμούς ή συγκεκριμένες ανάγκες από το δίκτυο. Οι αιτήσεις που δέχεται το δίκτυο και οι οποίες δεν απαιτούνε κάποια ειδική μεταχείριση ή δεν περιλαμβάνουν περιορισμούς, φορτώνονται ισότοπα και αποτελεσματικά φτάνουν στον προορισμό τους μέσω πολλαπλών άμεσων είτε έμμεσων διαδρομών που δημιουργούνται από το οπτικό πλέγμα, ανάλογα με τη κατάσταση του δικτύου την κάθε στιγμή (υπερφορτωμένο ή όχι). Το λογισμικό ελέγχου Plexxi είναι ακόμη έξυπνο αρκετά έτσι ώστε να ανιχνεύει κίνηση υψηλού επιπέδου και να ανακατευθύνει και να προωθεί ένα μέρος της είτε σε άμεσα μονοπάτια είτε σε ελεύθερα μονοπάτια υψηλού εύρους ζώνης,χωρίς να χρειάζεται εξουσιοδότηση.

Αντίθετα με τα τυπικά δίκτυα πολλαπλών διαδρομών τα οποία πιθανώς να μπορούνε να διαχειριστούνε το πολύ 16 ή 32 διαδρομές μεταξύ των μεταγωγέων η μονάδα ελέγχου Plexxi μπορεί έξυπνα να επιλέξει από εκατοντάδες έμμεσα μονοπάτια, τα οποία να μην δημιουργούνε κίνηση ή υπερφόρτωση στο δίκτυο, ανάμεσα σε όλη την μεγάλη ποικιλομορφία του πλέγματος. Ως αποτέλεσμα το Affinity SmartPath δημιουργεί πρωτοφανή αποτελεσματικότητα, παρέχοντας υψηλότερες επιδόσεις με μεγαλύτερη ευελιξία από ό, τι παραδοσιακές λύσεις δικτύων όπως για παράδειγμα ή ιεραρχική δομή.

Στη συνέχεια παρατίθεται ένα πλαίσιο το οποίο περιλαμβάνει μερικά βασικά χαρακτηριστικά του Plexxi switch 1 [18]:

Επιφάνεια	1RU / USB console
Διεπαφές πρόσβασης	32 x 10GbE SFP+, 2 x 40GbE QSFP+
Χωρητικότητα μεταγωγής (Switching capacity)	1.28 Tbps
Lightrail ™ διεπαφές	2 MPO συμβατοί connectors 24 μήκη κύματος (10 GbE)
Κατανάλωση ρεύματος	Μέγιστη κατανάλωση:250 W Τυπική κατανάλωση:120 W
Διαστάσεις	17.29"W x 28.00"D x 1.73"H
Βάρος	27 lbs
Ασφάλεια	UL/CSA/EN 60950
SDN βασισμένη αρχιτεκτονική	

3.1.4 Plexxi Switch 2

Ο μεταγωγέας Plexxi 2 αποτελεί την νέα γενιά δικτύωσης στις καινοτομικές προσπάθειες για λύση των μη επεκτάσιμων δικτύων [19].

Ο κάθε μεταγωγέας Plexxi 2 διαθέτει τέσσερις οπτικές διεπαφές lightrail οι οποίες παρέχουν έως και 480 Gbps εύρος ζώνης για επικοινωνία διπλής κατεύθυνσης, αριστερά και δεξιά του μεταγωγέα σε ένα δίκτυο Plexxi. Όσο αφορά τον τρόπο διασύνδεσης των μεταγωγέων χρησιμοποιούνται 2 ή 4 συνδέσεις ανά μεταγωγέα, ανάλογα με τις ανάγκες του δικτύου. Ο μεταγωγέας Plexxi 2 μπορεί να συνυπάρχει πλήρως με τον μεταγωγέα Plexxi 1 και έτσι είναι πιθανή η δημιουργία δικτύου Plexxi η οποία μπορεί να περιέχει οποιονδήποτε συνδυασμό μεταγωγέων 1 και 2.

Οι διακόπτες Plexxi διασυνδέονται σε μία τοπολογία δακτυλίου χρησιμοποιώντας lightrail υψηλής πυκνότητας οπτική τεχνολογία δημιουργώντας μια πλήρη ή μερικώς κατανεμημένη τοπολογία (mesh) μεταξύ των μεταγωγέων. Αυτό δημιουργεί μία πιο αποδοτική και αποτελεσματική αρχιτεκτονική δικτύου από ότι οι υπάρχουσες αρχιτεκτονικές δένδρου ή σπονδυλικές (spin) μπορούν να επιτύχουν. Η κατανεμημένη αρχιτεκτονική επιτρέπει γραμμική επεκτασιμότητα καθώς η προσθήκη κάθε μεταγωγέα στο δακτύλιο γίνεται εύκολα και ο μεταγωγέας συμβάλει στην αύξηση της χωρητικότητας του πλέγματος. Οι δακτύλιοι Plexxi μπορούν να υποστηρίξουν δίκτυα μεγέθους από μερικές στοίβες διακομιστών μέχρι και ένα πολύ μεγάλο κέντρο δεδομένων. Η γραμμική κατασκευή προσφέρει προβλέψιμη οικονομία και εύκολη ανάπτυξη της χωρητικότητας.

Με την εισαγωγή των 2 επιπλέον συνδέσεων lightrail και τον διπλασιασμό του εύρους ζώνης ο μεταγωγέας 2 δημιουργεί το επόμενο επίπεδο της επεκτασιμότητας στα κέντρα δεδομένων. Οι πρόσθετες συνδέσεις επιτρέπουν τη δημιουργία των ομόκεντρων δακτυλίων για την ενίσχυση του εύρους ζώνης μέσα στο δακτύλιο ή ακόμη και πολυδιάστατο δαχτυλίδια μεταγωγέων, αρχιτεκτονικές που μπορούν να επεκταθούν πάρα πολύ και να περιέχουν μέχρι και χιλιάδες μεταγωγείς και αντίστοιχα πολύ περισσότερες θύρες.

Flexxports

Ο μεταγωγέας Plexxi 2 μπορεί να παρέχει μέχρι και 28 FlexxPorts (4 QSFP+, 12 SFP+) [19]. Τα flexxports είναι θύρες πρόσβασης που είναι άμεσα συνδεδεμένες με το οπτικό πλέγμα Plexxi. Σε ταχύτητες έως 11Gbit / sec, μπορούν να παρέχουν ανεμπόδιστη non-Ethernet μεταφορά μέσω του οπτικού πλέγματος Plexxi σε οποιοδήποτε άλλο flexxport σε ένα δακτύλιο Plexxi. Τα FlexxPorts επιτρέπουν επίσης συνδέσεις του πλέγματος να ανακατευθυνθούν σε θύρες πρόσβασης για να επιτραπεί επικοινωνία μεγάλης εμβέλειας.

Εικόνα 3. 5 - Flexxports

Στη συνέχεια παρατηθούμε έναν πίνακα με μερικά βασικά χαρακτηριστικά του Plexxi μεταγωγέα 2 [19]:

Επιφάνεια	2 RU Form / Rj45
Διεπαφές πρόσβασης	48 x 10GbE: 12 x QSFP+ (1x40GbE, 4x10Gbe) 4 x QSFP+ FlexxPorts 12 x SFP+ FlexxPorts
Χωρητικότητα μεταγωγής (Switching capacity)	2.56 Tpbs (1.92 Tbps User)
Lightrail ™ διεπαφές	4 MPO συμβατοί connectors 48 x 10GbE μήκη κύματος
Κατανάλωση ρεύματος	Μέγιστη κατανάλωση:500 W Τυπική κατανάλωση:400W
Διαστάσεις	19.00”W x 28.00”D x 3.375”H
Βάρος	41.65 Lbs
Ασφάλεια	UL/CSA/EN 60950
SDN βασισμένη αρχιτεκτονική	
Ταξινόμηση Laser	Class 1 laser

Στο σημείο αυτό αξίζει να αναφερθούμε στο γεγονός ότι έχουνε σχεδιαστεί και άλλες μορφές του μεταγωγέα Plexxi 2 έτσι ώστε να καλύπτουν διαφορετικές ανάγκες στα κέντρα δεδομένων [22]. Η ‘’οικογένεια’’ Plexxi 2 αποτελείται από τέσσερα διαφορετικά μοντέλα που κατασκευάστηκαν με το ίδιο υλικό αλλά προσαρμοσμένα στο να εξυπηρετούν διαφορετικές εφαρμογές όπως κέντρα δεδομένων χτισμένα γύρω από συστάδες εξυπηρετητών (**Pods**). Τα μοντέλα είναι:

Μεταγωγέας Plexxi 2: Πολυδιάστατος μεταγωγέας διασύνδεσης
Μεταγωγέας Plexxi 2s: Μεταγωγέας πρόσβασης πολλαπλών θυρών 10GbE
Μεταγωγέας Plexxi 2p: Πολυδιάστατος μεταγωγέας διασύνδεσης Pod
Μεταγωγέας Plexxi 2sp: 10GbE pod μεταγωγέας πολλαπλών θυρών

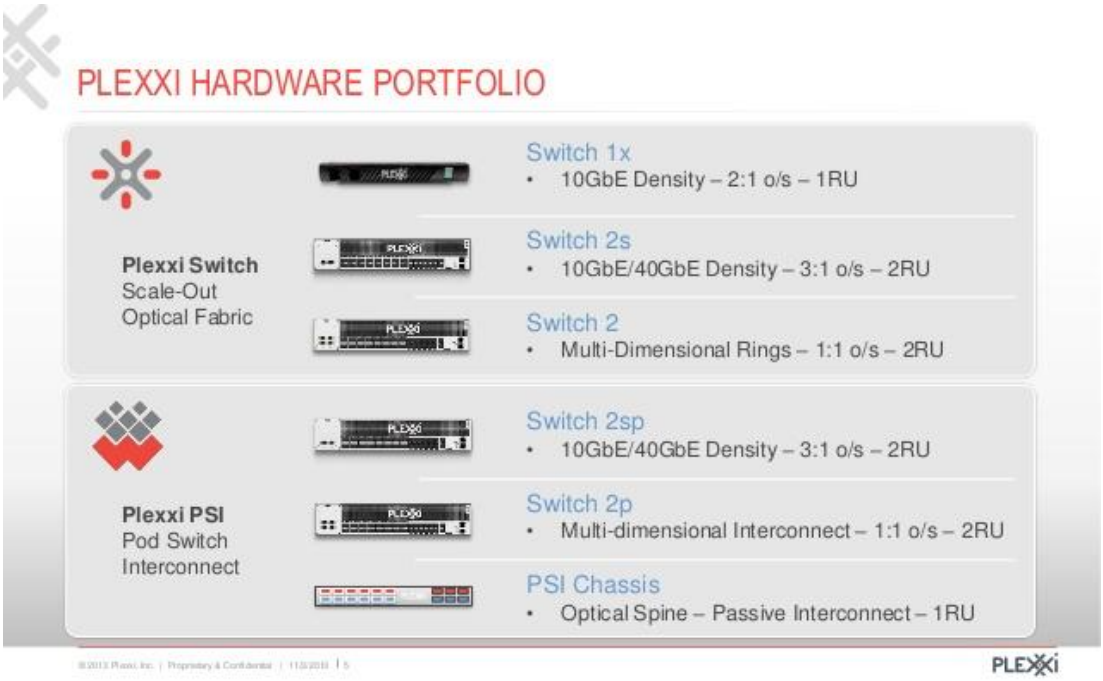
Μεταγωγέας Plexxi 2: Όπως έχει αναφερθεί και στην προηγούμενη ενότητα, ο μεταγωγέας 2 είναι ένα πολυδιάστατος μεταγωγέας διασύνδεσης σχεδιασμένος για μαζική επεκτασιμότητα του δικτύου. Η συσκευή είναι ένας 2RU μεταγωγέας με 12 QSFP + θύρες πρόσβασης (12x40GBE ή 48x10GbE). Ο μεταγωγέας έχει 4 lightrail διεπαφές που παρέχουν οπτική διασύνδεση μεταξύ των μεταγωγέων σε ένα Plexxi δακτύλιο. Οι οπτικές διεπαφές υποστηρίζουν οπτική τεχνολογία 10 χιλιομέτρων ‘’single-mode WDM’’ με 120Gbps επικοινωνία εξ ολοκλήρου διπλής κατεύθυνσης ανά διεπαφή (συνολικά 480Gbps). Ο μεταγωγέας 2 περιλαμβάνει συνήθως 16 FlexxPorts (4xQSFP+ και 12xSFP+). Επίσης το βασικό μοντέλο Plexxi 2 λειτουργεί με 1:1 υπερκάλυψη. Ο μεταγωγέας 2 είναι σχεδιασμένος να λειτουργεί ως μεταγωγέας διασύνδεσης μεταξύ των Plexxi δαχτυλιδιών. Οι πρόσθετες διασυνδέσεις lightrail επιτρέπουν στους αρχιτέκτονες να σχεδιάσουν τεμνόμενα δαχτυλίδια είτε για μαζικά επεκτάσιμες τοπολογίες είτε για μελλοντικούς σχεδιασμούς. Οι επιπλέον διεπαφές διπλασιάζουν την χωρητικότητα του οπτικού πλέγματος και αυξάνουν σημαντικά το εύρος ζώνης.

Μεταγωγέας Plexxi 2s: Ο μεταγωγέας 2s είναι ένας 10GbE μεταγωγέας πρόσβασης πολλαπλών θυρών που λειτουργεί με υπερκάλυψη 3:1. Ο διακόπτης 2RU έχει 16xQSFP + και 8xSFP + θύρες, οι οποίες επιτρέπουν μέχρι και 72 10GbE συνδέσεις πρόσβασης. Ο μεταγωγέας 2s έχει 2 διεπαφές lightrail (ίδιες με τις διεπαφές lightrail του μεταγωγέα 2) και 4xSFP + FlexxPorts. Με βάση το ίδιο υποκείμενο υλικό, οι μεταγωγείς 2s αντιπροσωπεύουν μια εκ νέου κατανομή της συνολικής δυναμικότητας μεταγωγής. Αυτή η διαμόρφωση βελτιστοποιείται για πολλαπλή πρόσβαση, κάνοντας το μεταγωγέα 2s ένα ιδανικό ακριανό μεταγωγέα σε ένα δαχτυλίδι Plexxi. Όταν τοποθετηθεί παράλληλα με τον μεταγωγέα 2, ο 2s γίνεται ένα υποστηριχτικό στοιχείο σε ενδεχομένως τεράστια τοπολογίες πολύ-δακτυλίων.

Μεταγωγέας Plexxi 2p: Ο μεταγωγέας 2p είναι ένα πολυδιάστατος μεταγωγέας διασύνδεσης, σχεδιασμένος ειδικά για αναπτύξεις pod. Ο σχεδιασμός του είναι παρόμοιος με αυτόν του μεταγωγέα 2 εκτός από τις διεπαφές lightrail. Ενώ ο μεταγωγέας 2 χρησιμοποιεί οπτική τεχνολογία 10 χιλιομέτρων WDM ο μεταγωγέας 2ρ χρησιμοποιεί ενός χιλιομέτρου Fabry-Perot (FP) οπτικά 1310nm. Αυτό συμβαίνει επειδή τα pods έχουν την τάση να αναπτύσσονται κατά σειρές σε κοντινές αποστάσεις και τα οπτικά του μεταγωγέα 2p είναι αρκετά μικρότερα. Οι 4 διεπαφές lightrail επιτρέπουν στον μεταγωγέα 2p να λειτουργεί ως ένα σημείο αλληλοσύνδεσης

για αναπτύξεις pod, με σκοπό ουσιαστικά την ενοποίηση πολλαπλών pod σε αυθαίρετα μεγάλες διαμορφώσεις, τα οποία διαχειρίζονται από μια ενιαία Plexxi μονάδα ελέγχου.

Μεταγωγέας Plexxi 2sp: Ο μεταγωγέας 2sp είναι ένας 10GbE μεταγωγέας πρόσβασης πολλαπλών θυρών, σχεδιασμένος αποκλειστικά για αναπτύξεις pod είτε ατομικές (single-pod) είτε ομαδικές (multi-pod). Ο σχεδιασμός του είναι παρόμοιος με αυτόν του 2s εκτός από τις διεπαφές lightrail. Χρησιμοποιεί την ίδια οπτική τεχνολογία με τον διακόπτη 2p. Ο μεταγωγέας 2sp σχεδιάστηκε ως διακόπτης πρόσβασης pod. Σε single-pod αναπτύξεις μπορεί να τοποθετηθεί παράλληλα με άλλες συσκευές 2sp. Σε multi-pod αναπτύξεις θα λειτουργήσει δίπλα στο μεταγωγέα διασύνδεσης 2p.



Εικόνα 3. 6 - Plexxi Μεταγωγείς

Σε αυτό το σημείο πρέπει να αναφέρουμε ότι όλοι οι μεταγωγείς του Plexxi, τόσο πρώτης όσο και δεύτερης γενιάς, είναι διαλειτουργικοί και μπορούν να τοποθετηθούν ο ένας δίπλα στον άλλο σε οποιαδήποτε επιθυμητή διαμόρφωση [22]. Δεν υπάρχουν περιορισμοί επί των οποίων οι συσκευές να μην μπορούν να συνδεθούν μεταξύ τους, ούτε υπάρχουν περιορισμοί με τρόπο που αλληλεπιδρούν με το λογισμικό της μονάδας ελέγχου Plexxi.

Plexxi Control Unit: Τέλος θα γίνει μία σύντομη αναφορά στη μονάδα ελέγχου Plexxi. Κάθε μεταγωγέας 2 περιέχει μία CPU μαζί με έναν Quad Core i5 επεξεργαστή και 16GB DRAM [22]. Η CPU είναι υπεύθυνη για τη λειτουργία του λογισμικού στον μεταγωγέα συμπεριλαμβανομένου και του λογισμικού στην μονάδα ελεγκτή του Plexxi. Ο ελεγκτής SDN

του Plexxi είναι χτισμένος με τη χρήση ιεραρχικής αρχιτεκτονικής ελεγκτή. Το λογισμικό του κεντρικού ελεγκτή (**central controller**) λειτουργεί ανεξάρτητα, και επικοινωνεί με τον κάθε μεταγωγέα με διακριτές υποστάσεις του co-controller (**discrete instances**) οι οποίες εκτελούνται στην CPU του κάθε μεταγωγέα. Ο συν-ελεγκτή (**co-controller**) είναι υπεύθυνος για πράγματα όπως η ανακάλυψη της φυσικής τοπολογίας, η διατήρηση του δικτύου και της συσκευής και η αυτόνομη προώθηση πακέτων.

3.2 Αρχιτεκτονικές δικτύου βασισμένες σε οπτικό αποπολυπλέκτη μήκους κύματος τύπου Arrayed Waveguide Grating (AWG)

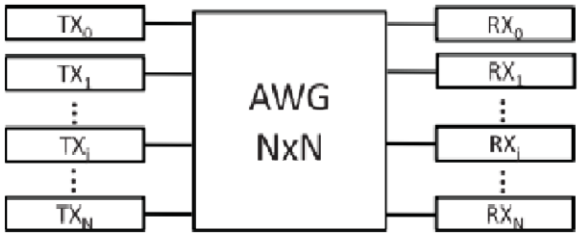
Στο συγκεκριμένο κεφάλαιο συνεχίζουμε την παρουσίαση των οπτικών αρχιτεκτονικών με τις δομές οι οποίες είναι βασισμένες σε δρομολόγηση μήκους κύματος με χρήση αποπολυπλεκτών μήκους κύματος τύπου AWG, οι οποίοι είναι γνωστοί και ως ‘’optical phased-arrays’’ [23]. Αρχικά γίνεται μία αναφορά και μία σύντομη παρουσίαση ενός κλασσικού AWG διευκρινίζοντας τη δομή και τη λειτουργία του. Στη συνέχεια ακολουθεί μία εκτενής παρουσίαση των σημαντικότερων αρχιτεκτονικών οι οποίες είναι βασισμένες σε αυτό.

3.2.1 Βασική μονάδα AWG

Τα AWG χρησιμοποιούνται συχνά ως οπτικοί πολυπλέκτες (**multiplexers**) ή αποπολυπλέκτες (**demultiplexers**) σε συστήματα τα οποία χρησιμοποιούν WDM [23]. Πιο συγκεκριμένα, οι συσκευές βασίζονται σε μια θεμελιώδη αρχή της οπτικής ότι τα κύματα φωτός που προέρχονται από διαφορετικά μήκη κύματος αλληλεπιδρούν γραμμικά μεταξύ τους. Αυτό σημαίνει ότι, εάν κάθε δίαυλος σε ένα οπτικό δίκτυο επικοινωνίας κάνει χρήση φωτός ενός ελαφρώς διαφορετικού μήκους κύματος, τότε το φως από ένα μεγάλο αριθμό αυτών των διαύλων μπορεί να μεταφερθεί από μία μόνο οπτική ίνα με σχετικά αμελητέες παρεμβολές μεταξύ των καναλιών-διαύλων. Τα AWGs χρησιμοποιούνται για την πολυπλεξία καναλιών διαφόρων μήκων κύματος σε μια ενιαία οπτική ίνα και μεταφορά στο άκρο μεταδόσεως (**transmission** end) και χρησιμοποιούνται επίσης ως αποπολυπλέκτες για να λαμβάνουν μεμονωμένα κανάλια διαφορετικών μηκών κύματος στο άκρο λήψης (**receiving** end) ενός οπτικού δικτύου επικοινωνίας. Αυτές οι συσκευές είναι ικανές να πολυπλέξουν ένα μεγάλο αριθμό μηκών κύματος σε μια ενιαία οπτική ίνα, αυξάνοντας έτσι την ικανότητα μεταφοράς των οπτικών δικτύων σημαντικά. Για να επιτύχουμε μεταγωγή μέσω του AWG κάνουμε χρήση της τεχνικής **wavelengthswitching,** δηλαδή μεταγωγή με βάση το μήκος κύματος του σήματος εισόδου. Με την τεχνική αυτή εκμεταλλευόμαστε την ευμεταβλητότητα του μήκους κύματος εισόδου στο AWG (wavelengthagility) έτσι ώστε να επιλέξουμε δυναμικά τη θύρα εξόδου (signalswitching) [23].

Ένα κέντρο δεδομένων αποτελείται συνήθως από πολλά racks. Το κάθε rack περιέχει ένα σύνολο από πολλές κάρτες γραμμών (**line cards**) οι οποίες συνδέονται μεταξύ τους μέσω ενός οπισθεπιπέδου (**backplane**) [24]. Οι line cards εμπεριέχουν τις διεπαφές του δικτύου (πομπούς και δέκτες) και εφαρμόζουν χαμηλού επιπέδου λειτουργίες επεξεργασίας του δικτύου.

Με βάση όλα τα παραπάνω προχωρούμε στη παρουσίαση της βασικής αρχής δρομολόγησης με AWG.



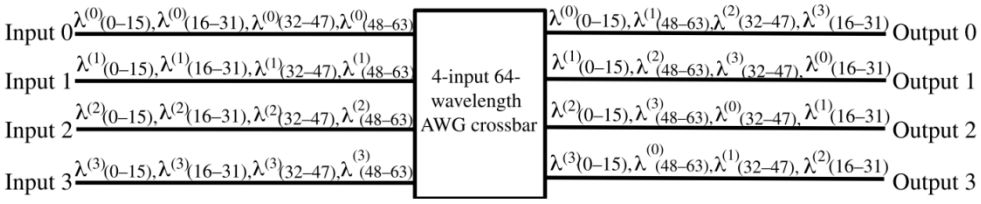
Εικόνα 3. 7- βασική αρχή δρομολόγησης με χρήση αποπολυπλέκτη μήκους κύματος AWG

Εδώ όλοι οι πομποί (**transmitters TX**) των καρτών γραμμής συνδέονται κατευθείαν σε ένα AWG και αυτό με τη σειρά του συνδέεται με τους δέκτες (**receivers RX**). Εάν *N* είναι ο αριθμός των line cards τότε *N* διαφορετικά μήκη κύματος απαιτούνται έτσι ώστε ο κάθε πομπός να μπορεί να επικοινωνεί με τον κάθε δέκτη, όπου είναι απαραίτητο ο κάθε πομπός να είναι εξοπλισμένος με ένα ρυθμιζόμενο λέϊζερ (tunable laser - fixed RX - tunable TX system). Η ιδιότητα δρομολόγησης του AWG αξιοποιείται έτσι ώστε να συνδεθούν οι TX και RX σύμφωνα με την σχέση:

$$j = (i + f) \bmod N$$

όπου τα i και j δείχνουν το TX και το RX, αντίστοιχα (i, j ∈ [0, N-1]) και f είναι ο δείκτης του το μήκος κύματος λf στο διακριτό σύνολο $\lambda f = \lambda 0 + f\Delta\lambda \ (f \in [0, N-1])$. Τέλος πρέπει να σημειωθεί ότι οι δέκτες και πομποί απεικονίζονται στο σχήμα σε ξεχωριστές στήλες στην είσοδο και έξοδο της αρχιτεκτονικής αλλά στην πραγματικότητα τα RX j και TX i είναι εγκατεστημένα στην ίδια line card *i* [24].

Στο σημείο αυτό αξίζει να σημειωθεί ότι είναι πιθανό και σε αρκετές περιπτώσεις αποδοτικό να γίνει μέσω της τεχνολογίας WDM και μίας αντίστοιχης ρύθμισης στα λέιζερ μεταφορά πολλαπλών μήκων κύματος σε μία line card ενός AWG. Ο σκοπός που εξυπηρετείται με αυτό είναι ότι έτσι το δίκτυο θα απαιτεί λιγότερες συνδέσεις και λιγότερο υλισμικό hardware για την υλοποίησή του. Για παράδειγμα εάν κάπου στο δίκτυο απαιτείται ένα 64x64 AWG, θα μπορούσαμε αντί αυτού να τοποθετήσουμε ένα 4x4 AWG το οποίο θα μεταφέρει 16 μήκη κύματος σε κάθε του line card. Σε ένα μεγάλο κέντρο δεδομένων αυτό θα μπορούσε να μειώσει σημαντικότατα το μέγεθος του. Τέλος θα ακολουθήσει μία σύντομη απεικόνιση του προαναφερθέντος παραδείγματος. Τα εισερχόμενα μήκη κύματος απεικονίζονται ως $\lambda^{c}_{a\text{-}b}$ όπου ο εκθέτης c απεικονίζει την θύρα εισόδου και τα a-b τον αριθμό των μήκων κύματος ανά γραμμή και ομαδοποιημένα.



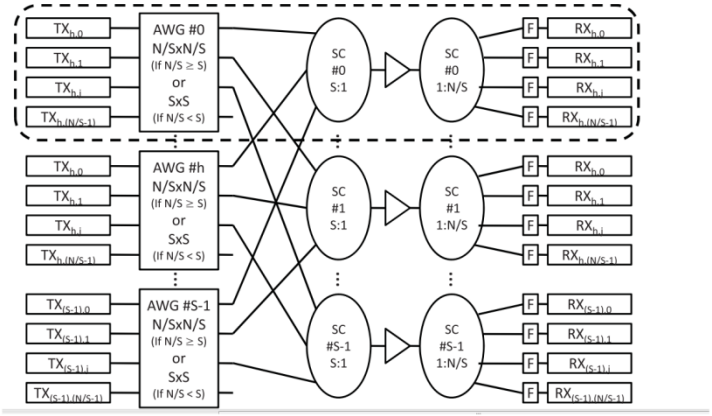
Εικόνα 3. 8 – Απεικόνιση AWG λειτουργικότητας

Στο παραπάνω σχήμα όπου απεικονίζεται ένα 4x4 AWG 64 μήκων κύματος θέτουμε προς εξέταση την θύρα εισόδου 0. Όλα τα μήκη κύματος που εισέρχονται στη θύρα 0 είναι τα εξής: $\lambda^0_{0\text{-}15}$ $\lambda^0_{16\text{-}31}$ $\lambda^0_{32\text{-}47}$ και $\lambda^0_{48\text{-}63}$. Ασχέτως με τον χρόνο άφιξης τους τα μήκη κύματος θα δρομολογηθούν ως εξής: τα μήκη κύματος $\lambda^0_{0\text{-}15}$ θα πάνε στη θύρα εξόδου 0, τα $\lambda^0_{16\text{-}31}$ στη θύρα εξόδου 1, τα $\lambda^0_{32\text{-}47}$ στη θύρα εξόδου 2 και τέλος τα $\lambda^0_{48\text{-}63}$ τη θύρα εξόδου 3. Τα αντίστοιχα συμβαίνουν και για τα υπόλοιπα μήκη κύματος που εισέρχονται σε διαφορετική είσοδο.

Εύκολα προκύπτει από όλα τα παραπάνω το πόρισμα ότι το AWG είναι ένα παθητικό στοιχείο (**passive element**) καθώς απλώς ανακατευθύνει τα διάφορα μήκη κύματος στις εξόδους στις οποίες έχουν ρυθμιστεί, το οποίο μπορεί να επιτύχει μεταγωγή αν συνδυαστεί με κατάλληλα ενεργά στοιχεία λέιζερ ρυθμιζόμενου μήκους κύματος εκπομπής.

3.2.2 AWG πολύ-επίπεδη αρχιτεκτονική (multi-plane architecture)

Σε αυτό το σημείο παραθέτουμε μια αρχιτεκτονική η οποία είναι βασισμένη στα AWG και αποσκοπεί στη δημιουργία επεκτάσιμου δικτύου μεγάλης κλίμακας με χρήση οπτικών στοιχείων με ρεαλιστικές διαστάσεις [24]. Η αρχιτεκτονική παρουσιάζεται στο παρακάτω σχήμα:



Εικόνα 3. 9 - Multiplane AWG αρχιτεκτονική

Όπως φαίνεται από το σχήμα οι N line cards διαχωρίζονται σε S ομάδες όπου η κάθε μια περιέχει N/S line cards. Οι N/S πομποί TX μίας ομάδας συνδέονται σε ένα D x D AWG όπου το D ορίζεται ως D = max { N/S, S }. Οι N/S δέκτες RX της ίδιας ομάδας των καρτών γραμμής συνδέονται μέσω N/S φίλτρων F σε ένα υποσύστημα το οποίο αποτελείται από: έναν S:1 οπτικό συζεύκτη,έναν οπτικό ενισχυτή και έναν 1:N/S οπτικό διαιρέτη. Το σύνολο όλων των παραπάνω στοιχείων και συσκευών αποκαλείται επίπεδο (**plane**) και απεικονίζεται ως κυκλωμένο στο αντίστοιχο σχήμα. Έτσι η αρχιτεκτονική αποκαλείται πολύ-επίπεδη (**multi-plane**) με S αριθμό επιπέδων. Το πρώτο ημι-επίπεδο εισόδου του κάθε επιπέδου (TX + AWG) συνδέεται με το δεύτερο ημι-επίπεδο εξόδου (ζεύκτης + ενισχυτής + διαιρέτης + φίλτρο + RX) μέσω των S εξόδων του D x D AWG. Ο συνολικός αριθμός των S planes μπορεί να κυμανθεί από 1 έως N.

Η αρχιτεκτονική προϋποθέτει ότι μόνο διασημειακές συνδέσεις απαιτούνται και χρειάζεται να καθιερωθούν μεταξύ των TX και RX χωρίς να δημιουργείται διαμάχη εξόδου (**output contention**). Δηλαδή ο κάθε δέκτης θα συνδέεται με το πολύ έναν πομπό χρησιμοποιώντας ένα μοναδικό μήκος κύματος. Ο κανόνας ανάθεσης ώστε το κάθε μήκος κύματος να βρεθεί στο σωστό δέκτη έχει ως εξής: για να επικοινωνήσει ο j-TX του h-plane με τον i-δέκτη του k-plane ο πομπός πρέπει να ρυθμιστεί στο αντίστοιχο μήκος κύματος που προκύπτει σύμφωνα με τη σχέση:

$$TXh,j \rightarrow RXk,i:\lambda f \, με \, f = (k - j) \, mod \, D + i \cdot D \quad (1)$$

όπου (h, k) ∈ [0, S −1] και (j, i) ∈ [0, N/S −1]. Το σύνολο των απαιτούμενων μήκων κύματος είναι ίσο με D * S. Η εξίσωση (1) δείχνει ότι ο κάθε δέκτης RX μπορεί να προσεγγιστεί από το πολύ D γειτονικά μήκη κύματος τα οποία έχουνε απόσταση Δλ ανά 2 μεταξύ τους. Έτσι το κάθε φίλτρο F πρέπει να έχει εύρος ζώνης D * Δλ.

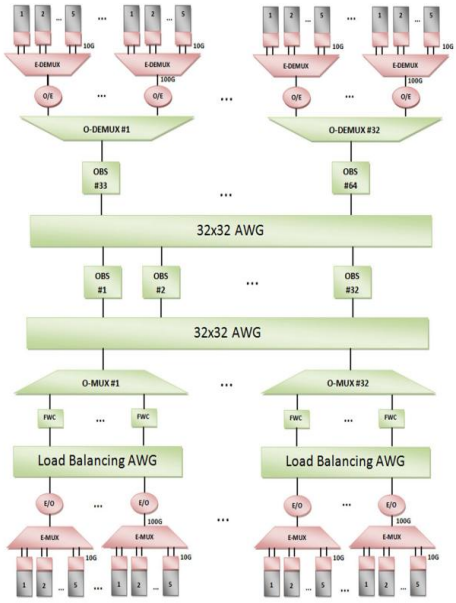
Η αρχιτεκτονική αυτή με τη συγκεκριμένη διάταξη των planes προσφέρει τη δυνατότητα επεκτασιμότητας της βασικής αρχιτεκτονικής AWG για μεγαλύτερα δίκτυα,ενώ παράλληλα εξασφαλίζει χαμηλή διακαναλική παρεμβολή (**crosstalk**) ακόμα και για δίκτυα με πολλούς κόμβους. Η συγκεκριμένη αρχιτεκτονική είναι σε θέση να επιτύχει υψηλό throughput για τη διασύνδεση καρτών των τηλεπικοινωνιακών συστημάτων μεταγωγής αλλα και είναι ικανή για την επίτευξη υψηλού συνολικού εύρους ζώνης για τον κάθε φορά εισερχόμενο ρυθμό bit (bit rate).

3.2.3 Αρχιτεκτονική υψηλής απόδοσης για επεκτασιμότητα των κέντρων δεδομένων

Στη συνέχεια παρουσιάζουμε μία ακόμα αρχιτεκτονική βασισμένη στα AWG [25]. Η συγκεκριμένη αρχιτεκτονική χρησιμοποιεί μόνο οπτικά στοιχεία και προορίζεται για μελλοντική χρήση στα κέντρα δεδομένων όταν δεν θα μπορούν να επεκταθούν περαιτέρω. Βασισμένη στην τεχνολογία του WDM,που επιτρέπει στην ίδια οπτική ίνα να μεταφέρει ταυτόχρονα διαφορετικά μήκη κύματος, η αρχιτεκτονική αυτή προσφέρει σημαντικότατη μείωση στον αριθμό των συνδέσεων που απαιτούνται σε ένα κέντρο δεδομένων, σε κλίμακα τάξης μεγέθους. Η αρχιτεκτονική πέρα από τα AWG χρησιμοποιεί εκτενώς και οπτικούς μεταγωγείς ριπής (**optical burst switches OBS**).

Η αρχιτεκτονική αυτή περιέχει μόνο οπτικές συνδέσεις και συνεπώς παρέχει ολικά οπτική συνδεσιμότητα μεταξύ οποιοδήποτε δυνατού ζευγαριού οπτικών κόμβων (**optical node**) μέσω εξισορρόπησης φορτίου (**load balancing**) και ιεραρχικής δρομολόγησης (**hierarchical routing**)

των οπτικών ριπών. Ο οπτικός κόμβος – optical node σε ένα κέντρο δεδομένων ορίζεται ως ένα ακροσημείο (**endpoint**) του οποίου η κίνηση μπορεί να μεταφερθεί σε μία οπτική ίνα. Ανάλογα με το κέντρο δεδομένων το optical node μπορεί να είναι ένα server rack,μία συστάδα από server racks, ή ακόμα και ένα σύνολο από servers. Με αυτά εν γνώσει παραθέτουμε τώρα την αρχιτεκτονική αυτή:



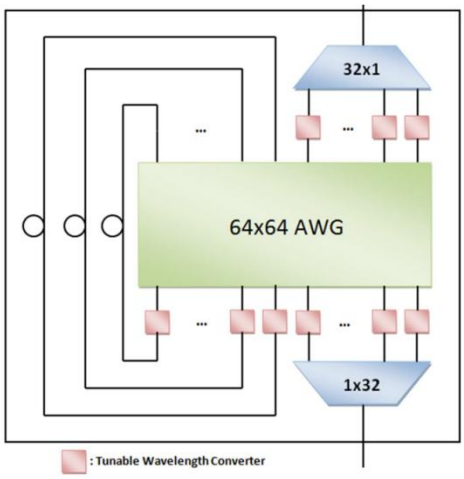
Εικόνα 3. 10 - AWG αρχιτεκτονική πολλαπλών διακομιστών

Στην παρούσα αρχιτεκτονική εξετάζονται περιπτώσεις κέντρων δεδομένων τα οποία αποτελούνται από υπερβολικά μεγάλο αριθμό διακομιστών. Στη συγκεκριμένη εξετάζουμε μία περίπτωση ενός κέντρου με 100.000+ διακομιστές, όπου οι servers είναι οργανωμένοι σε 5.120 racks και υποθέτουμε ότι κάθε rack περιέχει 20 διακομιστές. Υποθέτουμε ότι η μέγιστη διεκπεραιωτικότητα (**throughput**) του κάθε rack από servers είναι ίση με 20 Gbps και η κίνηση κάθε 5 rack συν-πλέκεται σε ένα 100 Gbps σήμα μέσω του WDM. Άρα ένα node στο συγκεκριμένο κέντρο δεδομένων αποτελείται από 5 server racks με μέγιστη διεκπεραιωτικότητα ίση με 100 Gbps.

Όπως απεικονίζεται το δίκτυο αυτό συγκροτείται από τον τομέα εξισορρόπησης φορτίου – load balancing section, όπου μετά ακολουθούνε 2 επίπεδα δρομολόγησης αποτελούμενα από OBS και AWG συσκευές. Το πρώτο βήμα για την εγκατάσταση της επιθυμητής λειτουργίας του δικτύου είναι η εξισορρόπηση του φορτίου από τα nodes. Ο σκοπός του load balancing είναι να ελαχιστοποιήσει την πιθανότητα διαμάχης (**contention probability**) και να ελαφρύνει το βάρος του χρονοπρογράμματος (**scheduling burden**) στα OBS. Στη συνέχεια γίνεται χρήση της κυκλικής ιδιότητας δρομολόγησης του AWG, η οποία έχει αναλυθεί σε προηγούμενη ενότητα, με σκοπό την ορθή δρομολόγηση. Ρυθμιζόμενα λέιζερ που προηγούνται των AWG εξισορρόπησης φορτίου είναι υπεύθυνα για το συντονισμό σε διαφορετικά μήκη κύματος έτσι ώστε να εξισορροπήσουν την κίνηση πριν την είσοδο στα πρώτα OBS. Η πρώτη ομάδα OBS μαζί με τα AWG που βρίσκονται μετά από αυτά, αναλαμβάνουν τη δρομολόγηση των οπτικών ριπών ανάμεσα σε 32 πιθανές συστάδες από server racks. Η κάθε συστάδα από racks περιέχει 32

nodes. Μόλις τα μήκη κύματος κατευθυνθούν στην αντίστοιχη συστάδα για την οποία προορίζονται, μέσω του 32 x 32 AWG, το δεύτερο στάδιο των OBS ολοκληρώνουν τη δρομολόγηση με τη βοήθεια οπτικών αποπολυπλεκτών και παραδίδουν τα optical bursts στο αντίστοιχο node και στον προορισμό τους.

Ένα από τα σημαντικότερα στοιχεία για την λειτουργία του δικτύου αυτού είναι τα OBS τα οποία επιτρέπουν την περιοριοθέτηση (**buffering**) δηλαδή την προσωρινή αποθήκευση των δεδομένων. Το κάθε OBS εμπεριέχει ένα AWG, ρυθμιζόμενους μετατροπείς μήκους κύματος (tunable wavelength converters **TWC**) για δρομολόγηση και αντιμετώπιση της κίνησης κατά την έξοδο και γραμμές καθυστέρησης (delay lines). Στο σχήμα που ακολουθεί παρουσιάζεται ένα από τα 64 32 x 32 OBS που χρησιμοποιούνται στη παρούσα αρχιτεκτονική [25]:



Εικόνα 3. 11 - Αρχιτεκτονική του OBS

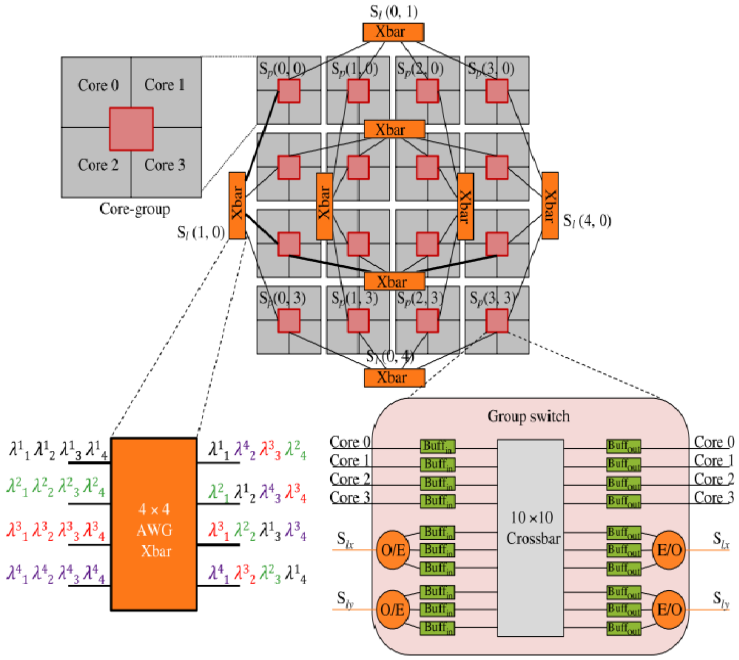
Τα 64 TWC στην είσοδο του AWG ρυθμίζονται να δρομολογούνε τα bursts και τα 32 TWC στην έξοδο μετατρέπουν τα μήκη κύματος των εξερχόμενων bursts στα αντίστοιχα μήκη κύματος μετάδοσης. Οι 32 ανακυκλικοί βρόχοι χρησιμοποιούνται για την αποθήκευση των bursts μέσα στον μεταγωγέα και εξασφαλίζουν ότι δεν θα υπάρχει contention στην έξοδο του OBS καθώς εάν παρευρεθούν πολλά διαφορετικά bursts στο ίδιο OBS αποθηκεύονται και οδεύουν προς τις γραμμές καθυστέρησης. Τέλος οποιοδήποτε εισερχόμενο burst μπορεί να τοποθετηθεί σε οποιαδήποτε γραμμή καθυστέρησης εάν παρουσιαστεί η ανάγκη.

Καθώς η αρχιτεκτονική αυτή που παρουσιάστηκε χρησιμοποιεί ιεραρχική δρομολόγηση και χρονοπρογραμματισμό, το κέντρο δεδομένων μπορεί να επιτύχει πολύ μεγάλο αριθμό από διακομιστές και να επεκταθεί εύκολα χωρίς να παρουσιάσει μείωση της αποδοτικότητας του. Επίσης απαιτεί σημαντικά λιγότερες φυσικές ζεύξεις (**physical links**) και συνδέσεις από ένα ηλεκτρονικό δίκτυο. Για παράδειγμα κοιτώντας από την εικόνα που απεικονίζεται η αρχιτεκτονική απαιτούνται 64 OBS, 32 multiplexers, 32 de- multiplexers,34 32x32 AWGs και 160 οπτικές ίνες για τη μεταφορά των WDM σημάτων. Ένα αντίστοιχο ηλεκτρονικό δίκτυο με τις ίδιες προδιαγραφές θα απαιτούσε 20.840 Ethernet links, 160 128x128 μεταγωγείς συνάθροισης και 64 160x160 ενδιάμεσους μεταγωγείς, συνολικά 2 παραπάνω τάξεις μεγέθους. Η αρχιτεκτονική αυτή καθίσταται μια ιδανική επεκτάσιμη λύση για κέντρα δεδομένων υψηλής απόδοσης άνω των 10 εκατομμυρίων πυρήνων μικροεπεξεργαστών. Το αρνητικό αυτής της

αρχιτεκτονικής είναι ότι η υλοποίηση της είναι εξαιρετικά ακριβή καθώς δεν είναι ακόμα διαθέσιμα σε αφθονία τα τεχνολογικά προϊόντα που απαιτεί.

3.2.4 Αρχιτεκτονική SPRINT (scalable photonic re-configurable interconnect)

Στο κεφάλαιο αυτό παρουσιάζουμε την αρχιτεκτονική SPRINT [26], μια κλιμακοθετήσιμη (**scalable**) οπτικό-ηλεκτρονική αρχιτεκτονική με υψηλό εύρος ζώνης περίπου 10 Gb/s ανά μήκος κύματος, χαμηλή λανθάνουσα καθυστέρηση (latency) της τάξεως των 2-3 ns και μεταγωγείς που χαρακτηρίζονται από μειωμένη πολυπλοκότητα. Η αρχιτεκτονική αυτή είναι βασισμένη στα AWG και εκμεταλλεύεται ταυτόχρονα τα πρόσφατα τεχνολογικά ευρήματα της νάνο-τεχνολογίας, τα MRR (**micro-ring resonators**) τα οποία μπορούν να χρησιμοποιηθούν για τη κατασκευή ενός AWG. Τα MRR είναι χαμηλής ενέργειας,υψηλού εύρους ζώνης φωτονικής τεχνολογίας ρυθμιζόμενα οπτικά φίλτρα τα οποία μπορούν να ρυθμιστούν ως συσκευές μεταγωγής ώστε να λειτουργούν παρόμοια με δρομολογητές (routers) χαμηλής κατανάλωσης ενέργειας. Στην αρχιτεκτονική αυτή οι πυρήνες (**cores**) που αποτελούν το κέντρο δεδομένων, ομαδοποιούνται σε ομάδες και σχηματίζουν ένα core group. Η ομαδοποίηση μπορεί να γίνει με οποιοδήποτε κριτήριο και ανάλογα με τις ανάγκες του αντίστοιχου κέντρου δεδομένων, εδώ προτείνεται και εξετάζεται η περίπτωση όπου ένα core group αποτελείται από 4 cores. Με την ομαδοποίηση επιτυγχάνεται μείωση του κόστους για την διασύνδεση των οπτικών πομπών σε κάθε core επειδή μειώνονται ο αριθμός των διαμορφωτών και το ποσό των O/E (οπτικό σε ηλεκτρικό) κυκλωμάτων. Έτσι η βασική θεμελιώδης μονάδα του δικτύου είναι ένα 4x4 block (ομάδα) το οποίο επιτρέπει 4 links να λειτουργούνε σαν να ήτανε 16 links χωρίς να χρειάζεται επιπλέον ένα ηλεκτρικό σύστημα μεταγωγής πακέτων. Να σημειωθεί ότι και εδώ γίνεται χρήση της τεχνολογίας WDM για τη μεταφορά πολλαπλών μήκων κύματος σε ένα μόνο physical link. Στη συνέχεια τα core groups ομαδοποιούνται και δημιουργούν μία συστάδα (**cluster**) και οι συστάδες με τη σειρά τους συνθέτουν ένα τομέα (**domain**). Η αρχιτεκτονική SPRINT επεκτείνεται σε 2 κατευθύνσεις (x, y) και οι μεταγωγείς διασυνδέονται και προς τις 2 αυτές κατευθύνσεις οι οποίες διακλαδίζονται και σε 2 επιπλέον διαστάσεις. Τα 2 αυτά επίπεδα διευθύνσεων συμβολίζονται ως C για τις θέσεις μέσα στην συστάδα και D για τις θέσεις μέσα στον τομέα. Με αυτό τον τρόπο μπορεί να αναγνωρίζεται εύκολα η θέση του κάθε πυρήνα μέσα στο δίκτυο και συμβολίζεται με Sp(⟨Cx,Cy,Dx,Dy⟩) όπου Cx και Cy είναι οι θέσεις μέσα στην συστάδα με τις αντίστοιχες κατευθύνσεις x,y και Dx και Dy οι αντίστοιχες θέσεις στον τομέα. Στην αρχιτεκτονική αυτή χρησιμοποιούνται 2 επίπεδα μεταγωγής, τοπικά (local) μέσα στο ίδιο cluster και σφαιρικά (global) για την διασύνδεση μεταξύ των cluster. Η τοπική μεταγωγή συμβολίζεται ως S_l (x, y) και η σφαιρική ως S_g (x, y) και γίνεται με AWG συσκευές. Στη συνέχεια παρουσιάζεται αναλυτικά η αρχιτεκτονική SPRINT 64-core και στη συνέχεια το πώς αυτή μπορεί να επεκταθεί σε 256, 512 και 1024 cores [26].

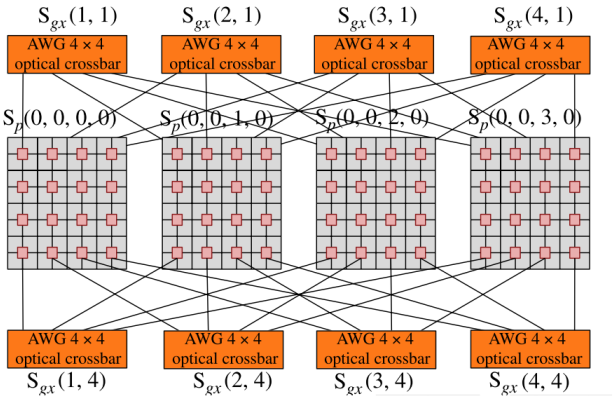


Εικόνα 3. 12 – Αρχιτεκτονική SPRINT

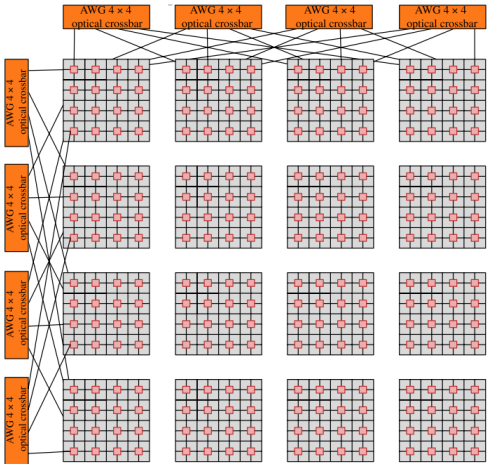
Στο σχήμα παρουσιάζεται η μορφή του SPRINT για 64-core αποτελούμενη από μία μόνο συστάδα με τιμές Cx = 4, Cy = 4, Dx = 0, και Dy = 0. Στο πάνω μέρος της εικόνας παρατηρούμε την ομαδοποίηση των cores και την αρχιτεκτονική του SPRINT καθώς και τη δομή των στοιχείων που την αποτελούν. Στη περίπτωση αυτή γίνεται χρήση μόνο των μεταγωγέων S_l (x, y) οι οποίοι διασυνδέουν τα core groups κατά το μήκος και των 2 κατευθύνσεων όπου $1 \leq x \leq 4$ και $1 \leq y \leq 4$. Η αύξηση του x συμβολίζει οριζόντια κίνηση ενώ του y κάθετη. Στο αριστερό κομμάτι του σχήματος παρουσιάζεται ένα AWG και το πώς δρομολογεί κυκλικά τα διάφορα μήκη κύματος ανάλογα με την είσοδο τους προς την θύρα εξόδου όπως και έχει προαναφερθεί. Αυτή η κυκλική δρομολόγηση επιτρέπει την οπτική επικοινωνία κατά μήκος της σειράς η της στήλης. Συνεπώς η επικοινωνία μέσα σε μία συστάδα θα περιορίζεται μόνο μέσα σε 2 αναπηδήσεις (**hops**).

Ας υποθέσουμε για παράδειγμα ότι το Sp(0,0) θέλει να επικοινωνήσει με το Sp(3,2), οι συντεταγμένες του τομέα είναι αδιάφορες μέσα σε μία συστάδα. Αρχικά δρομολογείται η y κατεύθυνση και έπειτα η x. Το πακέτο μετακινείται αρχικά προς το S_l(1,0) AWG και μέσω αυτού δρομολογείται στο Sp(0,2). Στη συνέχει το πακέτο δρομολογείται μέσω του S_l (0.3) στο προορισμό του Sp(3,2). Υπάρχουν έτσι το πολύ 2 hops για να φτάσει το πακέτο στον προορισμό του από οποιοδήποτε core σε ένα άλλο. Τέλος στην εικόνα απεικονίζεται τη μικρο-αρχιτεκτονική του router.

Στη συνέχεια παραθέτουμε 2 πιθανές επεκτάσεις της προηγούμενης αρχιτεκτονικής, μία για 256 cores και μία αρκετά μεγαλύτερη με 1024 cores.



Εικόνα 3. 12 – SPRINT με 256 cores



Εικόνα 3. 13 – SPRINT με 1024 cores

Στη πρώτη εικόνα απεικονίζεται το σύστημα των 256 cores όπου έχουν τοποθετηθεί παράλληλα 4 συστάδες από 64-cores των οποίων η λειτουργία αναλύθηκε στις προηγούμενες παραγράφους. Η επικοινωνία μέσα στο ίδιο το cluster γίνεται ακριβώς με τον ίδιο τρόπο. Εάν πρέπει 2 cores που βρίσκονται σε διαφορετικές συστάδες να επικοινωνήσουν τότε γίνεται χρήση των S_g (x, y) AWG. Ο μέγιστος αριθμός των απαιτούμενων hops είναι 4, 2 για την εύρεση της σωστής συστάδας και 2 για την εύρεση του core μέσα στην συστάδα. Στην δεύτερη εικόνα απεικονίζεται το SPRINT με 1024 cores. Εδώ τα core groups Sp(0,0,0,0), Sp(0,1,0,0), Sp(0,0,0,1) και Sp(0,1,0,1) συνδέονται με τον μεταγωγέα S_{gx} (0,1) και το ίδιο ισχύει και για τα υπόλοιπα core groups με τον αντίστοιχο κυκλικό συμβολισμό. Και εδώ ο μέγιστος αριθμός των hops είναι 4.

Η αρχιτεκτονική αυτή χαρακτηρίζεται ως ένα επεκτάσιμο φωτονικό δίκτυο για κέντρα δεδομένων υψηλών επιδόσεων η οποία προσφέρει λύσεις στα προβλήματα της κατασκευής και του σχεδιασμού των δικτύων για αρκετά μεγάλους αριθμούς από cores. Επίσης η συγκεκριμένη αρχιτεκτονική παρέχει τα πακέτα πολύ γρήγορα στον προορισμό τους και μπορεί πολύ εύκολα

να διαχειριστεί πολλαπλές μεταφορές πακέτων ταυτόχρονα. Ένα πρόβλημα που αντιμετωπίζει είναι ότι απαιτούνται πολλοί μεταγωγείς για να λειτουργεί, οι οποίοι παρότι αυξάνουν το εύρος ζώνης και μειώνουν σημαντικά το latency αυξάνουν παράλληλα το κόστος υλοποίησης-επίσης η αρχιτεκτονική αυτή περιλαμβάνει μεγάλο αριθμό από φυσικές ζεύξεις με αντίστοιχα αποτελέσματα στο κόστος υλοποίησης και στην πρακτική δυσκολία διασύνδεσης κάθε στοιχείου στο δίκτυο.

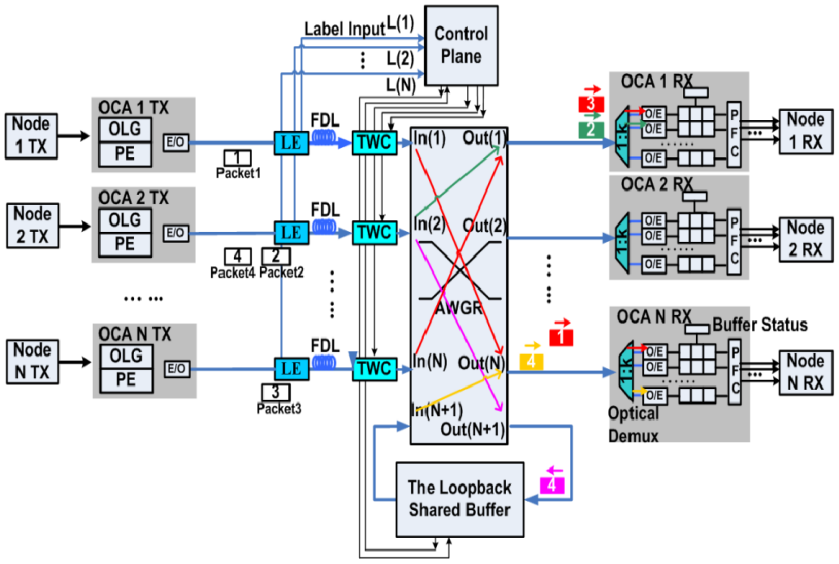
Ένας τρόπος να περιοριστεί αυτό είναι η χρήση λιγότερων global switches γύρω από τις συστάδες. Παρότι αυτό είναι εφικτό και έτσι θα μειώνονται σημαντικά ο υπερβολικός αριθμός των συνδέσεων υπάρχει μεγάλη περίπτωση να υπάρχει contention και μεγάλη κίνηση μέσα στο δίκτυο. Είναι επίσης εφικτό το να μην υπάρχει all-to-all connectivity μεταξύ όλων των clusters απευθείας αλλά μέσω τρίτων. Οι παραπάνω τρόποι ενώ είναι υλοποιήσιμοι δημιουργούν προβλήματα στο δίκτυο και προτείνονται κυρίως μόνο σε δίκτυα των οποίων η επικοινωνία θα γίνεται κατά κύριο λόγο μέσα στις συστάδες, ενώ στις πιο σπάνιες περιπτώσεις όπου θα είναι απαραίτητη η επικοινωνία μεταξύ των clusters δεν θα υπάρχει contention.

Ένας πιο αποδοτικός τρόπος, μέσω της χρήσης της WDM τεχνολογίας είναι η αύξηση του αριθμού των μήκων κύματος που μεταφέρονται ανά σύνδεση. Πιθανές επιλογές είναι 4, 16, 32 ή ακόμα και 64 μήκη κύματος ανά γραμμή. Αυτή η λύση βέβαια απαιτεί περισσότερα MRRs μέσα στα AWG και λέιζερ μέσα στους μεταγωγείς των core groups αλλά δεν επηρεάζει καθόλου τη τοπολογία του δικτύου. Το μειονέκτημα αυτής της λύσης είναι το αυξημένο κόστος που απαιτείται.

3.2.5 DOS (Datacenter Optical Switch) για AWG αρχιτεκτονικές σε κέντρα δεδομένων

Σε αυτήν την ενότητα θα γίνει παρουσίαση του μεταγωγέα DOS [27] ο οποίος είναι σχεδιασμένος για κλιμακοθετήσιμα κέντρα δεδομένων τα οποία απαιτούν ένα μεγάλο αριθμό συνδέσεων και τα οποία χρησιμοποιούν AWG μονάδες. Η κυκλική λειτουργία του AWG που επιτρέπει διαφορετικά πακέτα στις εισόδους του AWG να φτάνουν στην ίδια έξοδο με διαφορετικά μήκη κύματος καθώς και τα υπόλοιπα πλεονεκτήματα του έχουν παρουσιαστεί εκτενώς σε προηγούμενη ενότητα. Το DOS κάνει επίσης χρήση της τεχνικής label switching πράγμα που έχει ως αποτέλεσμα γρηγορότερη και ορθή λειτουργία του control plane. Σύμφωνα με την τεχνική label switching, σε κάθε πακέτο έχει εκχωρηθεί ένας αριθμός ετικέτας και η μετάβαση λαμβάνει χώρα μετά από την εξέταση της ετικέτας, που έχει αποδοθεί σε κάθε πακέτο, από το control plane. Η αρχιτεκτονική που είναι βασισμένη στο DOS παρουσιάζει χαμηλή λανθάνουσα καθυστέρηση και υψηλή διεκπεραιωτικότητα ακόμα και σε αρκετά μεγάλο αριθμό εισερχόμενων πακέτων. Είναι σημαντικό ακόμη να σημειωθεί ότι η κατανάλωση ενέργειας στο δίκτυο DOS σχετίζεται ανάλογα και γραμμικά με τον αριθμό των θυρών στο δίκτυο σε αντίθεση με άλλες αρχιτεκτονικές. Αυτά τα χαρακτηριστικά μαζί με τον σχετικά μεγάλο αριθμό θυρών που μπορεί να υπάρχει σε μία DOS-based αρχιτεκτονική καθιστούν το DOS ως μία ελκυστική λύση για όλα τα κέντρα δεδομένων που πάσχουν από contention.

Στην επόμενη εικόνα παρουσιάζεται λεπτομερέστατα η δομή της DOS αρχιτεκτονικής:



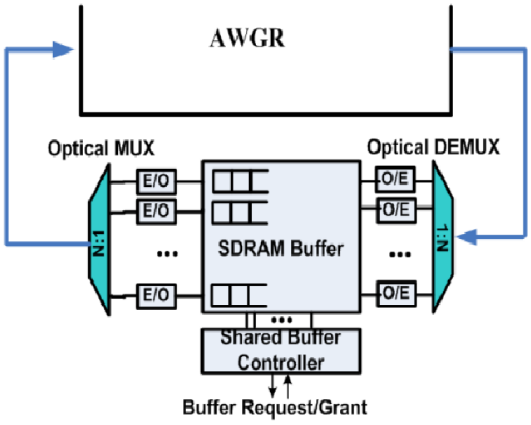
Εικόνα 3. 14 - Αρχιτεκτονική DOS

Στο κεντρικότερο κομμάτι της αρχιτεκτονικής βρίσκεται ένα οπτικό δικτυοδόμημα μεταγωγής το οποίο περιλαμβάνει συντονισμένους μετατροπείς μήκους κύματος (TWCs), ένα ULCF-AWG (uniform-lossandcyclic-frequency AWG) και ένα σύστημα αποθηκευτικού βρόχου (**loopback shared buffer system**) [27]. Επιπλέον, η αρχιτεκτονική περιέχει ένα επίπεδο ελέγχου το οποίο είναι υπεύθυνο για τη επεξεργασία της ετικέτας (**label**) του πακέτου καθώς και της διαιτήτευσης (**arbitration**) του κάθε πακέτου,ελέγχοντας πρώτα για διαθεσιμότητα σε θύρα εξόδου. Οι οπτικοί προσαρμογείς καναλιού (**optical channel adapter OCA**) λειτουργούν ως μεσάζοντες ανάμεσα στο AWG και στους ακραίους κόμβους.

Όπως είναι γνωστό το ULCF-AWG επιτρέπει την ολική διασύνδεση μεταξύ θυρών εισόδου και θυρών εξόδου με βάση το μήκος κύματος και παρέχει χαρακτηριστικά κυκλικής δρομολόγησης. Η δρομολόγηση του κάθε πακέτου μέσα στο AWG γίνεται με βάση το μήκος κύματος που μεταφέρει το σήμα και οποιαδήποτε θύρα εξόδου συνδέεται με οποιαδήποτε είσοδο με διαφορετικά μήκη κύματος (WDM). Με κατάλληλη ρύθμιση των TWCs και όσα προαναφέρθηκαν το εισερχόμενο πακέτο δρομολογείται στο προορισμό του χωρίς να υπάρχει μεγάλο contention. Κάθε θύρα εξόδου μπορεί να περιέχει πολλά διαφορετικά μήκη κύματος και τα σήματα αυτά μπορούν να απομονωθούν απλά με τη χρήση ενός αποπολυπλέκτη. Τέλος να σημειωθεί ότι οι υπάρχουσες πρακτικές εφαρμογές του DOS μπορούνε να χρησιμοποιήσουν και να λειτουργούν χωρίς απώλειες μέχρι και 512 x 512 AWGs με τη σημερινή τεχνολογική υποδομή.

Σε περίπτωση που δεν γίνεται χρήση της WDM ή οι θύρες εξόδου είναι λιγότερες από τις επιθυμητές ή τυχαίνει πάρα πολλά πακέτα να προορίζονται ταυτόχρονα για την ίδια θύρα εξόδου, τότε παρουσιάζεται contention στο δίκτυο. Αυτό αντιμετωπίζεται στην DOS αρχιτεκτονική με τον loopback shared synchronous dynamic random access memory (SDRAM) buffer. Ο SDRAM είναι υπεύθυνος για την αποθήκευση των πακέτων τα οποία για οποιοδήποτε

λόγο δεν πήρανε πρόσβαση προς τις θύρες εξόδου του AWG και για την επαναδρομολόγηση τους προς το AWG. Στα κέντρα δεδομένων η επανα-μετάδοση των πακέτων μπορεί να δημιουργήσει μεγάλα προβλήματα στο latency του δικτύου και ενώ αυτό το πρόβλημα μπορεί να αντιμετωπιστεί με γραμμές καθυστέρησης, σε ένα αρκετά μεγάλο κέντρο δεδομένων αυτό είναι μη αποτελεσματικό. Στην παρακάτω εικόνα απεικονίζεται ο SDRAM ο οποίος αποτελείται επιπλέον από οπτικούς αποπολυπλέκτες και πολυπλέκτες στα άκρα και O/E καθώς και E/O μετατροπείς (optical- electrical και electrical- optical):



Εικόνα 3. 15 - Δομή του SDRAM της αρχιτεκτονικής DOS

Ο SDRAM λαμβάνει πακέτα τα οποία απέτυχαν να εκλάβουν επιτρεπτή πρόσβαση προς τις εξόδους του AWG. Όλα τα πακέτα αυτά, κατευθύνονται με διαφορετικά μήκη κύματος και δρομολογούνται στην ίδια έξοδο (N + 1) του AWG. Στην συνέχεια διαχωρίζονται από τον οπτικό αποπολυπλέκτη (optical DEMUX στο σχήμα), μετατρέπονται σε ηλεκτρικά πακέτα αφού περάσουν από τον O/E μετατροπέα και αποθηκεύονται στο SDRAM. Η επιλογή του οπτικού αποπολυπλέκτη είναι 1:N με N παράλληλους δέκτες έτσι ώστε καθυστερημένα πακέτα που ταξιδεύουν σε διαφορετικά μήκη κύματος από διαφορετικές εισόδους να μπορούν να βγουν από διαφορετικές εξόδους του οπτικού DEMUX και να ληφθούν σωστά από ξεχωριστούς δέκτες. Ο SDRAM buffer έχει πολλές εξόδους κάθε μία με έναν E/O μετατροπέα ο οποίος μπορεί να παράγει ένα συγκεκριμένο μήκος κύματος το οποίο να επιτρέπει τη δρομολόγηση των πακέτων από την είσοδο AWG, η οποία είναι συνδεδεμένη με τον SDRAM, προς την έξοδο του AWG για την οποία προορίζεται. Ο SDRAM μπορεί να στείλει τα καθυστερημένα πακέτα σε πολλαπλές AWG θύρες εξόδου ταυτόχρονα με τη χρήση διαφορετικών μηκών κύματος. Ο SDRAM buffer controller, τον οποίο παρατηρούμε ακριβώς κάτω από τον SDRAM, παρέχει σήματα στον SDRAM τα οποία δείχνουν εάν η ουρά αναμονής είναι άδεια ή όχι. Ο SDRAM buffer controller επίσης στέλνει αιτήσεις προς το control plane της αρχιτεκτονικής ανάλογα με την κατάσταση της ουράς αναμονής. Τέλος να σημειωθεί ότι ο SDRAM buffer controller μπορεί να δημιουργήσει πολλαπλά αιτήματα εάν περισσότερες από μία ουρά αναμονής δεν είναι άδεια και είναι ικανή να αποδεχτεί πολλαπλές αιτήσεις και να εκκινήσει πολλαπλές ταυτόχρονες μεταδόσεις. Ο έλεγχος της ροής (**flow control**) είναι απαραίτητος σε περίπτωση που ο SDRAM είναι γεμάτος και δεν μπορεί να εισάγει ένα νέο πακέτο, με αποτέλεσμα να υπάρχει πιθανότητα διαμάχης και στον SDRAM. Για τη καταπολέμηση αυτού προτείνεται η χρήση ενός ενδοζωνικού ON-OFF πρωτοκόλλου ροής (**in-band ON-OFF flow control**) το οποίο δεν επιβαρύνει

ιδιαίτερα το σύστημα. Το πρωτόκολλο αυτό μπορεί εύκολα να εγκατασταθεί αξιοποιώντας απλά κάποια αχρησιμοποίητα bits στην κεφαλίδα του πακέτου. Όταν η χωρητικότητα του SDRAM υπερβεί ένα ορισμένο όριο, τα αντίστοιχα bits στην κεφαλίδα του καθυστερημένου πακέτου που πρόκειται να σταλεί θα αλλάξουν στην επιλογή OFF. Έτσι, όταν ο τελικός κόμβος λαμβάνει και διαβάζει την κεφαλίδα αυτού του πακέτου θα αναστείλει προσωρινά τη μετάδοση του και θα περιμένει για άλλο πακέτο που θα υποδεικνύει ότι ο buffer είναι έτοιμος να λάβει καθυστερημένα πακέτα και πάλι.

Οι οπτικοί προσαρμογείς καναλιού - **OCA** παρέχουν την ενδιάμεση διεπαφή ανάμεσα στο DOS και στους ακραίους κόμβους όπως έχει προαναφερθεί. Αυτό επιτρέπει στους τελικούς κόμβους να χρησιμοποιούν οποιοδήποτε πρωτόκολλο τους αρέσει όπως InfiniBand, 10G Ethernet ή PCI express. Το OCA εξάγει τα σχετικά πεδία του πακέτου και δημιουργεί μια οπτική ετικέτα που μεταδίδεται σε ένα μήκος κύματος διαφορετικό από το υπόλοιπο του πακέτου. Το μήκος κύματος που φέρει το οπτικό σήμα διαχωρίζεται αργότερα από το μήκος κύματος που φέρει το ωφέλιμο φορτίο πακέτου με ένα φίλτρο, και η ετικέτα παραδίδεται στο επίπεδο ελέγχου μετά από O / E μετατροπή. Η οπτική ετικέτα περιέχει ένα προοίμιο, πέρα από τα πεδία προορισμού και το μήκος του πακέτου έτσι ώστε να επιτρέψει τον συγχρονισμό των κύκλων του ρολογιού και το συγχρονισμό των δεδομένων στο επίπεδο ελέγχου. Στην αντίθετη κατεύθυνση το OCA διεφάπτεται μεταξύ μιας θύρας εξόδου AWG και ενός δέκτη κόμβου. Το κάθε OCA έχει έναν 1:k optical DEMUX και k παράλληλους δέκτες για να μπορεί να περιέχει k ταυτόχρονα πακέτα που ταξιδεύουν χρησιμοποιώντας k μήκη κύματος για k διαφορετικές ομάδες κυματομορφών (**wave-groups**). Επίσης υπάρχει ένας ηλεκτρικός buffer ο οποίος συνδέει κάθε δέκτη με τα αποθηκευμένα πακέτα που έχουν παραληφθεί και περιμένουν να παραδοθούν στο τέλος του κόμβου.

Η αρχιτεκτονική αυτή με την αξιοποίηση του παραλληλισμού του μήκους κύματος από το δικτυοδόμημα μεταγωγής που δημιουργείται από τα AWG μειώνει αποτελεσματικά το contention σε κάθε έξοδο. Το DOS επίσης παρέχει χαμηλό latency και υψηλή διεκπεραιωτικότητα-throughput και δεν θα κορεστεί σε πολύ υψηλά επίπεδα του φορτίου εισόδου, ακόμα και σε επίπεδα μέχρι περίπου 90% [27]. Επιπλέον, η λανθάνουσα καθυστέρηση του DOS είναι σχεδόν ανεξάρτητη από τον αριθμό των θυρών εισόδου.

Τέλος, αξίζει να αναφερθούμε στην νεότερη γενιά των DOS μεταγωγέων οι οποίοι ονομάζονται TONAK-LION [28] και βελτιώνουν την αντίστοιχη αρχιτεκτονική. Συγκεκριμένα οι TONAK-LION συνδυάζουν μία αποκλειστικά οπτική τεχνική ονομαζόμενη NACK, η οποία καταργεί την ανάγκη για ηλεκτρικούς buffers στις θύρες εισόδου και εξόδου του κάθε μεταγωγέα, με την τεχνική TOKEN η οποία επιτρέπει σε έναν κατανεμημένο αποκλειστικά οπτικό arbiter να διαχειριστεί αποκλειστικά την διαμάχη-contention των πακέτων [28].

3.3 Αρχιτεκτονικές βασισμένες στην τεχνολογία οπτικών διακοπτών τύπου MEMS (Micro-Electro-Mechanical Switch)

3.3.1 Αρχές λειτουργίας κυκλωμάτων MEMS

Στο ακόλουθο κεφάλαιο παρουσιάζονται οι αρχιτεκτονικές οι οποίες στηρίζονται στην τεχνολογία των οπτικών διακοπτών τύπου MEMS. Η MEMS είναι μια τεχνολογία που στην γενικότερη μορφή της μπορεί να οριστεί ως μικρογραφία μηχανικών και ηλεκτρό-μηχανικών στοιχείων τα οποία προκύπτουν από τεχνικές μικροκατασκευής [29]. Οι φυσικές διαστάσεις των MEMS μπορεί να κυμαίνονται από εξαιρετικά μικρές διαστάσεις μέχρι και μερικά millimeters. Τα είδη των συσκευών MEMS μπορεί να κυμαίνονται από σχετικά απλές κατασκευές που δεν έχουν κινούμενα στοιχεία μέχρι και εξαιρετικά πολύπλοκα ηλεκτρομηχανικά συστήματα με πολλαπλά κινούμενα στοιχεία τα οποία ελέγχονται από ενοποιημένα μικρό-ηλεκτρονικά. Το ένα βασικό χαρακτηριστικό των MEMS είναι ότι υπάρχουν τουλάχιστον μερικά στοιχεία που έχουν κάποιο είδος μηχανικής λειτουργικότητας ασχέτως του αν τα στοιχεία αυτά μπορούν να κινηθούν από μόνα τους. Ενώ τα λειτουργικά στοιχεία των MEMS είναι μικρογραφημένες δομές, εκκινητήρες, αισθητήρες και μίκρο-ηλεκτρονικά στοιχεία, τα πιο αξιοσημείωτα στοιχεία είναι οι μίκρο- εκκινητήρες και οι μίκρο- αισθητήρες. Οι μίκρο- εκκινητήρες και οι μίκρο-αισθητήρες κατηγοριοποιούνται ως μορφοτροπέες (**transducers**) οι οποίες ορίζονται ως οι συσκευές που μετατρέπουν την ενέργεια από μια μορφή σε άλλη. Περισσότερες πληροφορίες για την τεχνολογική εξέλιξη των MEMS μπορούν να βρεθούν στο συγκεκριμένο άρθρο [29].

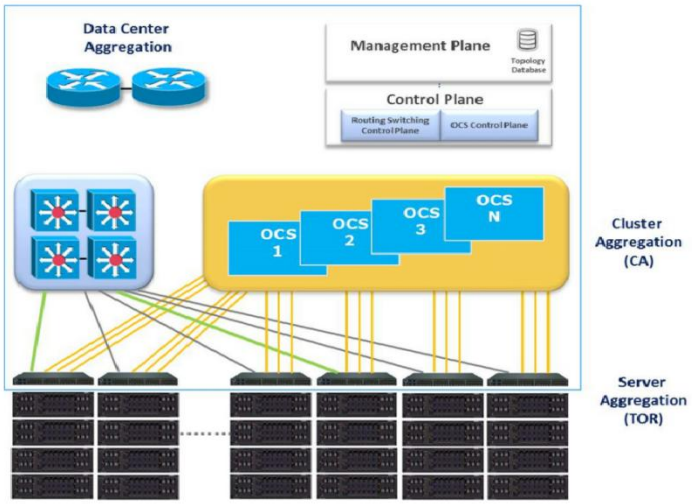
Όσο αφορά τις οπτικές αρχιτεκτονικές η τεχνολογία MEMS χρησιμοποιείται συχνά για την κατασκευή οπτικών κυκλωμάτων μεταγωγής (optical circuit switch **OCS**). Ένα MEMS-based OCS είναι θεμελιωδώς διαφορετικό από έναν μεταγωγέα ηλεκτρικών πακέτων. Το OCS λειτουργεί αποκλειστικά σε ακτίνες φωτός χωρίς να χρειάζεται να αποκωδικοποιήσει κανένα πακέτο. Το OCS υλοποιεί μία N x N ράβδο (**crossbar**) από μίκρο-κάτοπτρα (ή καθρέφτες) για να κατευθύνει τις δέσμες φωτός από μία οποιαδήποτε θύρα εισόδου σε οποιαδήποτε θύρα εξόδου [30]. Τα ίδια τα κάτοπτρα είναι συνδεδεμένα σε μικροσκοπικούς κινητήρες (**motors**) των οποίων η διάσταση είναι περίπου 1 mm^2. Ένας ενσωματωμένος επεξεργαστής ελέγχου τοποθετεί τα κάτοπτρα ώστε να υλοποιήσουν μια συγκεκριμένη μήτρα σύνδεσης (**connection matrix**) και αποδέχεται απομακρυσμένες εντολές έτσι ώστε να επαναρυθμίσει τον προσανατολισμό των κατόπτρων σε μία νέα μήτρα σύνδεσης όταν απαιτείται. Η μηχανική επανατοποθέτηση των κατόπτρων επιβάλλει χρόνο μεταγωγής στο κύκλωμα, τυπικά της τάξεως των χιλιοστών του δευτερολέπτου (ms). Παρά τον χρόνο μεταγωγής που επιβάλλεται στο κέντρο δεδομένων έτσι, το MEMS-based OCS προσφέρει κάποια πλεονεκτήματα. Αρχικά ένα OCS δεν χρειάζεται επιπλέον πομποδέκτες (**transceivers**) στους ενδιάμεσους κόμβους αφού δεν υπάρχει ανάγκη για μετατροπή του φωτός σε ηλεκτρική ενέργεια ώστε να γίνει η μεταγωγή. Δεύτερον, ένα OCS χρησιμοποιεί σημαντικά λιγότερη ενέργεια από ότι ένας ηλεκτρικός μεταγωγέας πακέτων. Τρίτον, δεδομένου ότι OCS δεν επεξεργάζεται τα πακέτα, όσο το κέντρο δεδομένων αναβαθμίζεται σε 40 GigE και 100 GigE, το OCS δεν χρειάζεται να αναβαθμιστεί, διότι είναι διαφανές στον τύπο και το ρυθμό των δεδομένων. Επιπλέον μπορεί να γίνει χρήση της WDM στο οπτικό κύκλωμα και να εξυπηρετήσει πολύ καλύτερα το κέντρο δεδομένων με τη πολλαπλή μεταφορά πολλών channel ταυτόχρονα ενώ ένας ηλεκτρικός μεταγωγέας πακέτων θα πρέπει πρώτα να αποπολυπλέξει όλα τα channels και στη συνέχεια να κατευθύνει κάθε channel σε κάθε θύρα εξόδου [30].

Τέλος είναι σημαντικό να αναφερθεί ότι είναι διαθέσιμα στην αγορά οπτικά κυκλώματα MEMS-based τα οποία περιέχουν μέχρι και μερικές εκατοντάδες θύρες τα οποία λειτουργούν σε επιθυμητά επίπεδα, ενώ ταυτόχρονα γίνονται έρευνες και μελέτες για επέκταση των MEMS σε περισσότερες από 1000 θύρες [31]. Στη συνέχεια θα ακολουθήσει η παρουσίαση μερικών αρχιτεκτονικών οι οποίες χρησιμοποιούν MEMS-based οπτικά κυκλώματα.

3.3.2 Calient 3D MEMS

Ξεκινάμε την παρουσίαση των αρχιτεκτονικών βασισμένες στα MEMS με την αρχιτεκτονική που προτείνεται από την Calient [17]. Σύμφωνα με αυτή, ένα OCS μπορεί να τοποθετηθεί και να αναπτυχθεί μαζί με ένα ηλεκτρικό κύκλωμα μεταγωγής πακέτων, ώστε να σχηματίσουν ένα υβριδικό δίκτυο για να βελτιώσει σημαντικά τις επιδόσεις και να μπορεί να υποστηρίξει τις απαιτήσεις των σύγχρονων και μελλοντικών κέντρων δεδομένων. Η υψηλού επιπέδου υβριδική αυτή δομή πακέτων-οπτικού κυκλώματος εξίσου υποστηρίζει τόσο σύντομη ριπαία κίνηση (bursty traffic) όσο και ροές υψηλής χωρητικότητας μέσα στο κέντρο δεδομένων.

Ένα από τα λάθη των τωρινών κέντρων δεδομένων είναι το γεγονός ότι υπάρχει ολική συνδεσιμότητα μεταξύ όλων των racks / pods ενώ αυτό είτε δεν απαιτείται είτε απαιτείται σπάνια. Ο λόγος για τον οποίο συνδέονται έτσι είναι για να είναι τα κέντρα δεδομένων προετοιμασμένα για τη χειρότερη και σπανιότερη δυνατή περίσταση. Αυτό που πραγματικά χρειάζεται και προσφέρει η παρούσα αρχιτεκτονική, είναι η δυνατότητα να επαναδιαρθρώνεται (**reconfiguration**) το δίκτυο κατ΄ αίτηση έτσι ώστε να επεμβαίνει δυναμικά ενεργοποιώντας τα απαιτούμενα μονοπάτια και να μην δημιουργείται latency στο δίκτυο. Έτσι έχουμε ένα δίκτυο υψηλών επιδόσεων που κάνει τη καλύτερη δυνατή χρήση των διαθέσιμων πόρων με χαμηλή επένδυση. Στην επόμενη εικόνα απεικονίζεται η αρχιτεκτονική Calient:



Εικόνα 3. 16 - Αρχιτεκτονική Calient

Σε αυτή την υβριδική λύση το ηλεκτρικό δίκτυο πακέτων (δεξιά πάνω) συνεχίζει να υπάρχει με ολική συνδεσιμότητα μεταξύ των συστάδων- ToRs. Το δίκτυο πακέτων εστιάζει κυρίως στον

χειρισμό χαμηλής ριπής ροών δεδομένων όπως αυτά μεταφέρονται μεταξύ των συστάδων του κέντρου δεδομένων.

Παράλληλα με το δίκτυο πακέτων, τοποθετείται ένα οπτικό δίκτυο μεταγωγής (OCS) Trunk δίκτυο (αριστερά πάνω) που αποτελείται από ένα δικτυοδόμημα οπτικών στοιχείων μεταγωγής κυκλωμάτων. Ο ρόλος αυτού του δικτύου μεταγωγής κυκλώματος είναι να αναδιαρθρώνεται κατ΄ αίτηση και να αλλάζει όπως απαιτείται ώστε να υποστηρίξει μεγάλες ροές δεδομένων, ελευθερώνοντας έτσι το ηλεκτρικό δίκτυο πακέτων και αφαιρώντας οποιαδήποτε πιθανότητα δημιουργίας contention.

Το OCS δικτυοδόμημα παρέχει ουσιαστικά απεριόριστο εύρος ζώνης που θα κλιμακώνεται χωρίς ανάγκη για αναβάθμιση καθώς οι ταχύτητες των δικτύων θα αυξάνουν μελλοντικά σε 100G και πάνω. Αυτό είναι επειδή καθαρά φωτονικές 3D MEMS βασισμένα OCS λύσεις όπως ο μεταγωγέας Calient 320 port S320, ο οποίος χρησιμοποιείται σε ένα Plexxi δίκτυο [21], είναι εντελώς διαφανής προς το ρυθμό δεδομένων και το πρωτόκολλο και δεν χρησιμοποιεί οπτικούς πομποδέκτες. Επιπλέον, το OCS δικτυοδόμημα προσφέρει εξαιρετικά χαμηλό latency (λιγότερο από 60ns) μεταξύ των μονοπατιών των ToRs παρέχοντας έτσι άριστη υποστήριξη για εφαρμογές ευαίσθητες στο latency.

Ο χρόνος αποκατάστασης ενός μεταγωγέα οπτικού κυκλώματος όπως ο S320 CALIENT είναι συνήθως 25 χιλιοστά του δευτερολέπτου (50ms μέγιστο) [17]. Αυτό υπαγορεύεται από το γεγονός ότι απαιτούνται ηλεκτροστατικές επανατοποθετήσεις των μίκρο-κατόπτρων για την επίτευξη της μεταγωγής και οι νόμοι της φυσικής επιβάλλουν περιορισμούς. Στα ηλεκτρικά κυκλώματα μεταγωγής πακέτων 25ms είναι αρκετά υψηλός και μη αποδεκτός χρόνος μεταγωγής αλλά σε μια υβριδική αρχιτεκτονική όπου το οπτικό κύκλωμα χρησιμοποιείται για να εξυπηρετήσει μεγάλες ροές δεδομένων είναι αποδεκτό. Ο λόγος είναι ότι οι περισσότερες μεγάλες ροές εκπέμπουν και παραμένουν για λεπτά ή περισσότερο και έτσι ο χρόνος αποκατάστασης του OCS είναι ουσιαστικά άνευ σημασίας. Κατά τη μεταβατική περίοδο πριν από την πραγματοποίηση μίας OCS σύνδεσης το δίκτυο πακέτων θα συνεχίσει να μεταφέρει τη ροή της κυκλοφορίας και τα δεδομένα δεν χάνονται.

Κανονικά στην αρχιτεκτονική αυτή όταν δεν υπάρχει μεγάλη ροή δεδομένων η οποία να δημιουργεί προβλήματα και να υπερφορτώνει το δίκτυο χρησιμοποιείται μόνο το ηλεκτρικό κύκλωμα πακέτων. Ας υποθέσουμε ότι παρουσιάζεται μία υψηλή ροή δεδομένων στο δίκτυο η οποία δεν μπορεί να διαχειριστεί από το ηλεκτρικό κύκλωμα. Στην αρχική κατάσταση όλα τα ToRs συνδέονται και στο οπτικό και στο ηλεκτρικό σύστημα. Μεταξύ δύο συστάδων παρουσιάζεται μία υψηλή ροή δεδομένων και οι buffers των μεταγωγέων δεν έχουν περαιτέρω αποθηκευτικό χώρο και απαιτείται μία διαδρομή υψηλότερης χωρητικότητας προκειμένου να περάσουν τα δεδομένα χωρίς προβλήματα. Στη συνέχεια το δίκτυο πακέτων ειδοποιεί το επίπεδο ελέγχου πως παρουσιάζεται πρόβλημα με τη κίνηση στο δίκτυο και αυτό με τη σειρά του αποκρίνεται με την έκδοση μιας εντολής για το OCS ώστε να ανοίξει ένα άμεσο οπτικό μονοπάτι μεταξύ των racks που έχουν τη μεγάλη ροή δεδομένων. Η ροή δεδομένων έχει πλέον πρόσβαση σε ένα πολύ υψηλό εύρος ζώνης, χαμηλό latency, και το δίκτυο πακέτων δεν είναι πλέον φορτωμένο. Το επίπεδο διαχείρισης / ελέγχου, μπορεί να ξεκινήσει από μόνο του την εγκατάσταση και απόσυρση των οπτικών μονοπατιών που βασίζονται σε μια σειρά από κριτήρια, όπως ανταπόκριση στη ζήτηση του δικτύου σε πραγματικό χρόνο, προγραμματισμένο χρονοδιάγραμμα, την ώρα της ημέρας, ή ενδεχομένως συμπεριφορά ανάλογα με έναν πολύ έξυπνο αλγόριθμο. Τέλος αξίζει να σημειωθεί ότι το επίπεδο διαχείρισης / ελέγχου κάνει χρήση του μοντέλου SDN για καλύτερη και πιο άμεσα λειτουργία.

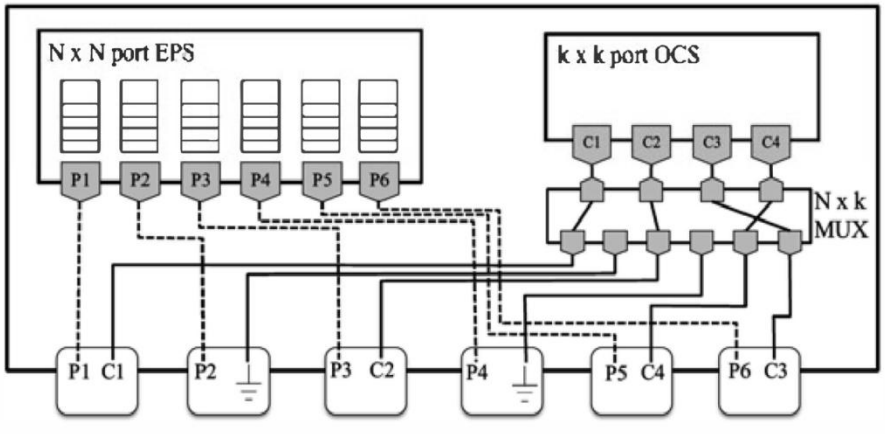
Γενικά, αυτό το υβριδικό δίκτυο παρέχει μια κλιμακούμενη, με χαμηλό κόστος, υψηλής χωρητικότητας, χαμηλής λανθάνουσας καθυστέρησης, μελλοντική λύση που λύνει τα σημαντικότερα προβλήματα με τα οποία έρχεται αντιμέτωπο ένα σημερινό κέντρο δεδομένων. Να συμπληρωθεί ότι η συγκεκριμένη αρχιτεκτονική μπορεί να να διαχειριστεί μεγάλες και επίμονες ροές δεδομένων με απεριόριστο εύρος ζώνης σε χαμηλό κόστος, ελευθερώνοντας έτσι το κύκλωμα πακέτων. Τέλος να αναφερθεί ότι η αρχιτεκτονική αυτή είναι εφαρμόσιμη και σε μελλοντικές ταχύτητες της τάξης των 100G καθώς οι διεπαφές του δικτύου αναβαθμίζονται στα ToRs.

3.3.3 REACToR: A REconfigurable pAcket and Circuit ToR Switch

Συνεχίζουμε τη παρουσίαση των δομών βασισμένων στα MEMS με την αρχιτεκτονική που ονομάζεται REACToR [32]. Το REACToR είναι ένας υβριδικός ToR μεταγωγέας που συνδέει ένα ταχέως επαναδιαρθρώσιμο κύκλωμα έμφραξης μεταγωγής (blocking optical circuit switch) απευθείας με τους servers ή end-hosts των racks. Η υλοποίηση του REACToR αποτελείται από ένα οπτικό κύκλωμα μεταγωγής υψηλής χωρητικότητας (OCS) που έχει προγραμματιστεί σύμφωνα με το TDMA μοντέλο μαζί με ένα σχετικά υπό-παρεχόμενο ηλεκτρικό κύκλωμα πακέτων χωρίς φραγή (non blocking EPS). Το κλειδί στον σχεδιασμό του REACToR είναι ένας χρονοπρογραμματιστής μεταγωγέων ο οποίος προγραμματίζει τη κίνηση που προέρχεται από τον κάθε end-host μέσα από τα racks, χρησιμοποιώντας ένα πρωτόκολλο σηματοδοσίας που εξασφαλίζει ότι το στιγμιαίο προσφερόμενο φορτίο μπορεί να διαχειριστεί από την τρέχουσα διαμόρφωση του μεταγωγέα REACToR. Οι ριπές πακέτων αποθηκεύονται στους καταχωρητές των end-host μέχρι να λάβουν ένα σήμα ότι έχει αναδιαμορφωθεί και σχηματιστεί το επιθυμητό κύκλωμα για την μεταφορά τους, ενώ οι κινήσεις μικρής έντασης ή αρκετά ευαίσθητες στο latency προωθούνται προς το ηλεκτρικό κύκλωμα πακέτων. Επιπλέον το ηλεκτρικό κύκλωμα πακέτων προορίζεται να εξυπηρετήσει όλες τις αναπάντεχες αιτήσεις (applications) που οφείλονται σε λάθη εκτίμησης ή σε λάθη του χρονοπρογραμματιστή του κυκλώματος.

Η κύρια αρχή για τον σχεδιασμό του REACToR είναι ότι μεταφέροντας τη συντριπτική πλειοψηφία, όχι πάντως ολόκληρη, του buffering έξω από τον μεταγωγέα και μέσα στους end-hosts το δίκτυο θα έχει πολύ λιγότερο latency και θα είναι πολύ ευκολότερη η επεκτασιμότητα του. Εάν η μεταγωγή κυκλώματος είναι αρκετά γρήγορη, τότε καθυστερήσεις που οφείλονται στο buffering στον end-host, δεν θα υποβαθμίσουν την απόδοση των πρωτοκόλλων βασισμένων στα πακέτα. Το REACToRσυνδυάζοντας τις δυνάμεις και των 2 τεχνολογιών μπορεί να φτάσει υψηλά επίπεδα απόδοσης.

Στην συνέχεια προχωρούμε με την παρουσίαση της αρχιτεκτονικής του REACToR:



Εικόνα 3. 18 - REACToR αρχιτεκοτινκή

Στην παραπάνω εικόνα απεικονίζεται μία από τις πιθανές μορφές που μπορεί να έχει ο REACToR μεταγωγέας. Ένα REACToR n-θυρών αποτελείται από ένα οπτικό κύκλωμα OCS k-θυρών όπου k < n (πάνω δεξιά) και ένα ηλεκτρικό κύκλωμα EPS n-θυρών (πάνω αριστερά). Κάθε ένα από τα εξωτερικά ToRs συνδέεται και με το OCS και με το EPS όπως είναι εμφανές και στο σχήμα. Τα ToRs ανταλλάσουν δεδομένα μεταξύ τους εκπέμποντας πακέτα τα οποία δρομολογούνται στους προορισμούς τους είτε μέσω του OCS ή του EPS. Σε μία πλήρως παρεχομένη περίσταση θα ίσχυε k = n, δηλαδή το οπτικό κύκλωμα θα είχε ίδιο αριθμό θυρών με το ηλεκτρικό, αλλά εξαιτίας της μικρής πιθανότητας για επεκτασιμότητα από την οποία χαρακτηρίζονται τα σημερινά OCS, ένα πλήρως παρεχόμενο REACToR είναι απίθανο να είναι οικονομικά αποδοτικό όπως επίσης είναι και πολύ σπάνιο να εμφανιστούν τόσο μεγάλες ριπές κίνησης ταυτόχρονα. Γενικά ο REACToR μεταχειρίζεται ένα OCS φραγής (**blocking**) n-θυρών το οποίο μπορεί αν συνδέσει μόνο k θύρες τη φορά. Όπως φαίνεται και από την εικόνα αρχιτεκτονικής του REACToR, πριν από το OCS προηγείται ένας n x k multiplexer. Με αυτό τον τρόπο όλοι οι end-hosts των ToRs συνδέονται και έχουν πρόσβαση στο OCS για την περίπτωση εμφάνισης μεγάλης ριπής και ο πολυπλέκτης χρησιμοποιείται για να προωθεί την κίνηση προς το OCS που έχει μόνο k θύρες. Με τη σημερινή και παρούσα τεχνολογία των MEMs, μπορούν να κατασκευαστούν REACToR με δεκάδες αριθμό θυρών χωρίς να επηρεάζουν τον χρόνο αναδιαμόρφωσης του κυκλώματος, ο οποίος είναι η σημαντικότερη παράμετρος για την ορθή λειτουργία του REACToR [32].

Σημαντικά και άξια αναφοράς επίσης πέρα από την κατασκευή του υβριδικού μεταγωγέα, είναι η εφαρμογή ενός χαμηλού latency πρωτοκόλλου ανάμεσα στον REACToR και στους end-hosts έτσι ώστε να προγραμματίζονται οι μεταδόσεις των πακέτων και να ανασχηματίζεται το κύκλωμα ανάλογα, καθώς και η διαχείριση του buffering στους end-hosts έτσι ώστε να επάγει bursty μεταδόσεις για μέγιστη αποδοτικότητα του κυκλώματος.

Για να γίνει αποτελεσματική χρήση της χωρητικότητας του κυκλώματος μεταγωγής, ο REACToR πρέπει να καθορίσει ένα κατάλληλο πρόγραμμα αντιστοίχισης και χαρτογράφησης κυκλώματος (**mapping**) για να εξυπηρετήσει τις ερχόμενες αιτήσεις. Ένας αποτελεσματικός χρονοπρογραμματιστής χρειάζεται μόνο να αναγνωρίζει τις μεγάλες ροές καθώς οι μικρές εξυπηρετούνται καλύτερα από το ηλεκτρικό κύκλωμα πακέτων. Ο REACToR χρησιμοποιεί το πρωτόκολλο IEEE 802.11 Qbb Priority Flow Control (PFC) στο επίπεδο έλεγχου του για να επιτρέψει επιλεκτικά την κίνηση που προορίζεται για κάθε θύρα του REACToR. Εκτιμάται ότι

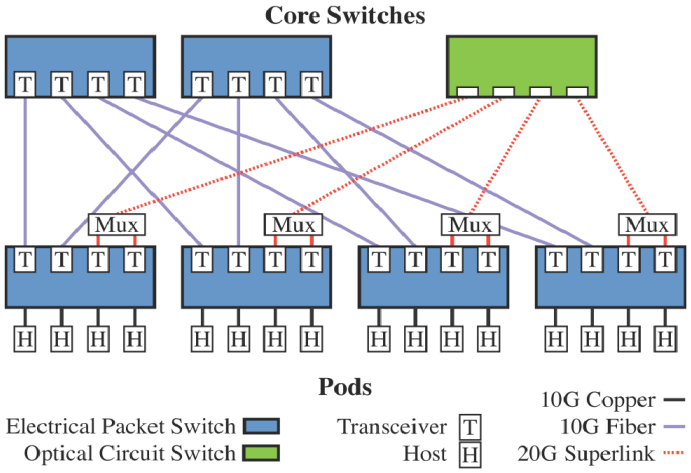
το latency του control plane είναι μόνο μόλις 2,4 μs για REACToR με δεκάδες αριθμό θυρών.

Τέλος όσο αφορά το buffering στους end-hosts, η επιλογή αυτή γίνεται γιατί όταν στο μέλλον οι ταχύτητες των κέντρων δεδομένων γίνουν 100 Gbps και περαιτέρω θα είναι πολύ δύσκολο να μην δημιουργείται υψηλό latency στο κύκλωμα εάν αυτό πρέπει να ασχοληθεί και με το buffering των πακέτων. Το REACToR χρησιμοποιεί TDMA μοντέλο χρονοπρογραμματισμού για να αναγνωρίζει τα bursts των πακέτων στους end-hosts (30+ πακέτα) και το TCP μοντέλο ως το πρωτόκολλο μεταφοράς (**transport protocol**) για τη μεταφορά των πακέτων προς τον προορισμό τους.

3.3.4 Helios: Hybrid Electrical/Optical Switch Architecture

Σε αυτή την ενότητα παρουσιάζεται η δομή της αρχιτεκτονικής Helios [30]. Πολλά σημερινά κέντρα δεδομένων στηρίζουν την δομή τους γύρω από την ομαδοποίηση και δημιουργία συστάδων εξυπηρετών (**pods**), τα οποία pods αποτελούνται από εξυπηρετητές (servers), στοιχεία δικτύων και στοιχεία ψύξης. Το κάθε pod περιέχει συνήθως από 250 έως και 1000 servers και το να κατασκευαστεί ένα δικτυοδόμημα μεταγωγής το οποίο να διασυνδέει τους servers μεταξύ τους μέσα στο ίδιο το pod είναι πραγματοποιήσιμο. Αυτό όμως που παραμένει δύσκολη πρόκληση είναι η διασύνδεση εκατοντάδων ή ακόμα και χιλιάδων pods για να σχηματιστεί ένα μεγάλο κέντρο δεδομένων. Για κέντρα δεδομένων που στηρίζουν τη δομή τους στα pods η αρχιτεκτονική Helios είναι ελκυστική καθώς είναι μία ακόμα υβριδική αρχιτεκτονική η οποία συνδυάζει δυναμικά τα πλεονεκτήματα του ηλεκτρικού και του οπτικού κυκλώματος (MEMS-based) και είναι σχεδιασμένη αποκλειστικά για pod-based κέντρα δεδομένων.

Στην επόμενη εικόνα παρουσιάζουμε την αρχιτεκτονική του Helios:



Εικόνα 3. 17 - Αρχιτεκτονική Helios

Εδώ απεικονίζεται μία μικρογραφία της αρχιτεκτονικής. Το Helios είναι ένα δύο επιπέδων και βασικό δένδρο πολλαπλών διακλαδώσεων σε pod μεταγωγείς και σε core μεταγωγείς [30]. Ως core μεταγωγείς ορίζονται το ηλεκτρικό κύκλωμα μεταγωγής το οποίο απεικονίζεται πάνω αριστερά και επειδή συνήθως τα pod-based κέντρα δεδομένων είναι αρκετά μεγάλα σε έκταση μπορεί να απαιτούνται πολλοί ηλεκτρικοί 10 GigE μεταγωγείς ώστε να μην υπερφορτώνεται το

δίκτυο και το οπτικό κύκλωμα του οποίου η τεχνολογία είναι βασισμένη και εδώ στα 3D-MEMS και βρίσκεται πάνω δεξιά. Τα πλεονεκτήματα που υπάρχουν με την χρήση και των 2 κυκλωμάτων έχουν αναλυθεί πλήρως σε προηγούμενες αρχιτεκτονικές. Εδώ το οπτικό κύκλωμα OCS αναλαμβάνει τις μεγάλες και αργές ροές της επικοινωνίας μεταξύ των pod,που θα δημιουργούσαν πρόβλημα. Το ηλεκτρικό κύκλωμα παρέχει ολικό εύρος ζώνης για τις bursty εκπομπές μέσα στο ίδιο το pod. Ένας από τους κύριους στόχους του Helios είναι ότι αντί να προετοιμαστεί ένα τεράστιο κέντρο δεδομένων για τη χειρότερη δυνατή περίσταση, είναι προτιμότερο να παρέχεται μία ‘’δεξαμενή’’ (**pool**) από εύρος ζώνης και να εκχωρείται όπου και όταν απαιτείται χάρις σε δυναμικές αλλαγές των σχηματομορφών επικοινωνίας (**communication patterns**).

Συνεχίζοντας με την ανάλυση της αρχιτεκτονικής το κάθε pod εμπεριέχει ένα αριθμό από hosts, οι οποίοι συμβολίζονται ως H και συνδέονται με τον μεταγωγέα του pod από μικρού μήκους χάλκινα καλώδια. Το πλήθος των hosts εξαρτάται από το μέγεθος του κέντρου δεδομένων. Το μήκος των χάλκινων καλωδίων επιλέγεται να είναι το πολύ 10 μέτρα καθώς μέσα σε ένα pod τόσες είναι οι μεγαλύτερες αποστάσεις και παραπάνω επένδυση σε καλώδια χαλκού κρίνεται μη αποδοτική για το κέντρο δεδομένων. Ο pod μεταγωγέας περιέχει έναν αριθμό από οπτικούς πομποδέκτες (optical transceiver),οι οποίοι απεικονίζονται με T, έτσι ώστε να συνδέεται το pod με τα core switches. Σε αυτή τη μορφή του Helios οι μισές ανερχόμενες ζεύξεις (**uplinks**) από το κάθε pod συνδέονται με τους μεταγωγείς πακέτων,οι οποίοι απαιτούν επίσης από έναν πομποδέκτη, μέσω μίας 10G οπτικής ίνας. Τα άλλα μισά uplinks από το κάθε pod περνούν μέσα από ένα παθητικό οπτικό πολυπλέκτη (απεικονιζόμενος ως M) πριν συνδεθούν σε ένα OCS. Οι ζεύξεις αυτές αποκαλούνται από τους σχεδιαστές ως superlinks και στη συγκεκριμένη παραλλαγή του δικτύου μεταφέρουν 20G δεδομένων. Να αναφερθεί σε αυτό το σημείο ότι είναι πολύ σημαντική η χρήση της τεχνολογίας WDM σε αυτή την αρχιτεκτονική έτσι ώστε να μεταφέρονται πολλά μήκη κύματος μαζί και να εξοικονομούν συνδέσμους οπτικών καλωδίων. Ορίζεται ως w το μέγεθος του κάθε superlink και εξαρτάται από τον αριθμό των WDM μήκων κύματος που μεταφέρει. Παραλλαγές του Helios με w = 1,2,4,8,16 και 32 έχουν δοκιμαστεί και λειτουργούν με θετικά αποτελέσματα.

Η αρχιτεκτονική Helios αναγνωρίζει το υποσύνολο της κίνησης η οποία ταιριάζει καλύτερα στη μεταγωγή του κυκλώματος και διαρθρώνει δυναμικά την τοπολογία του δικτύου σε χρόνο λειτουργίας που καθορίζεται από τις άμεσες αλλαγές των σχηματομορφών επικοινωνίας. Ένας από τους σημαντικότερους στόχους του Helios επειδή προορίζεται για μεγάλα κέντρα δεδομένων είναι να παρουσιάζει την ίδια απόδοση με πλήρως συνδεδεμένα και παρεχόμενα σημερινά ηλεκτρικά δίκτυα αλλά με σημαντικά χαμηλότερο κόστος, λιγότερη πολυπλοκότητα σύνδεσης και λιγότερη κατανάλωση ενέργειας.

Επίσης, να αναφέρουμε ότι η αρχιτεκτονική Helios δεν απαιτεί καμία τροποποίηση στους end-hosts αλλά μόνο απλή και άμεση τροποποίηση λογισμικού στους μεταγωγείς. Το λογισμικό που απαιτούν τα μοντέλα Helios είναι ένα λογισμικό διαχείρισης της τοπολογίας του Helios (topology manager TM) σύμφωνα με την οποία αλλάζει και ελευθερώνει μονοπάτια, ένα λογισμικό διαχείρισης για το οπτικό κύκλωμα (CSM) και ένα λογισμικό διαχείρισης για το κάθε pod μεταγωγέα.

Συνολικά, ο συνδυασμός των οπτικών κυκλωμάτων MEMS μαζί με τους WDM πομποδέκτες στο Helios, προσφέρουν μία αρχιτεκτονική με επεκτάσιμο εύρος ζώνης ανά θύρα με σημαντικά λιγότερο κόστος και κατανάλωση ενέργειας από τους μεταγωγείς ηλεκτρικών πακέτων. Επιπλέον η δομή της αρχιτεκτονικής Helios είναι βασισμένη στα pods, χαρακτηριστικό που την

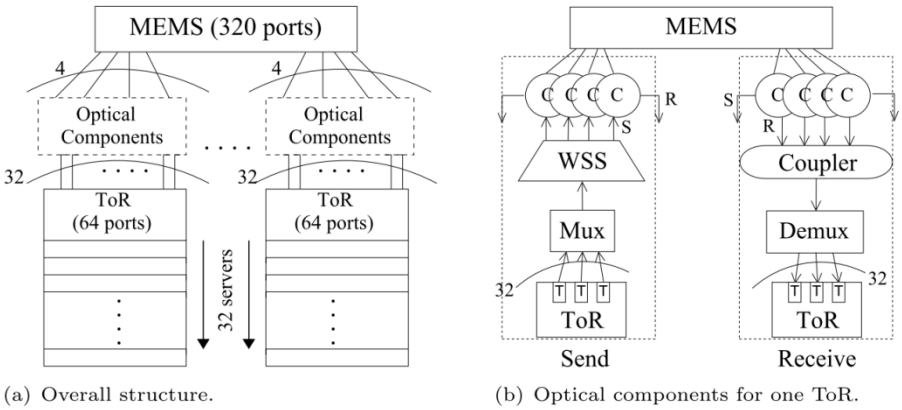
κάνει συμβατή με σύγχρονα κέντρα δεδομένων που κατανέμουν εκατοντάδες έως χιλιάδες servers σε pods.

3.3.5 Αρχιτεκτονική Proteus: Μία μορφοποιήσιμη αρχιτεκτονική για κέντρα δεδομένων

Ολοκληρώνουμε την παρουσίαση των MEMs-based αρχιτεκτονικών με την αρχιτεκτονική Proteus [31]. Πρόκειται για μία αρχιτεκτονική αποκλειστικά αποτελούμενη από οπτικά στοιχεία η οποία εστιάζει στην πολύ εύκολα μορφοποιήσιμη δομή της και στη χαμηλή πολυπλοκότητα των οπτικών λειτουργικών στοιχείων που την αποτελούν. Η αρχιτεκτονική Proteus προσαρμόζει δυναμικά την τοπολογία της έτσι ώστε να πληρεί τις απαιτήσεις της κίνησης όπου αυτή παρουσιαστεί. Η αρχιτεκτονική Proteus επιτυγχάνει υψηλά επίπεδα μορφοποίησης εκμεταλλεύοντας την επαναδιαρθρωσιμότητα (reconfigurability) της οπτικής τεχνολογίας, δηλαδή και την ικανότητα για αλλαγή της μορφής του οπτικού κυκλώματος και της παροχής του οπτικού μήκους κύματος στο χρόνο εκτέλεσης. Η αρχιτεκτονική Proteus αποφεύγει εντελώς τη χρήση ηλεκτρικού εξοπλισμού πέρα από τους ToR μεταγωγείς, εξασφαλίζοντας με αυτό τον τρόπο καλύτερη απόδοση ενέργειας, καλύτερη μετάβαση στις 40-GigE ταχύτητες και παραπέρα καθώς και σημαντικά απλοποιημένη καλωδίωση στο κέντρο δεδομένων.

Πριν προχωρήσουμε στην παρουσίαση της αρχιτεκτονικής του Proteus, θα ακολουθήσει μία σύντομη αναφορά των κυρίων οπτικών στοιχείων που το αποτελούνε. Το Proteus είναι βασισμένο στη λειτουργία των MEMS για τη μεταφορά δεδομένων από το ένα ToR στο άλλο. Το Proteus επίσης κάνει χρήση και μίας άλλης θεμελιώδους μονάδας για την κατασκευή οπτικών αρχιτεκτονικών, το WSS (**Wavelength Selective Switch**) του οποίου η λειτουργία αναπτύσσεται λεπτομερώς σε επόμενη ενότητα, αλλά προς το παρών εδώ χρησιμοποιείται ως ένα 1 × N οπτικό στοιχείο. Επίσης το Proteus περιλαμβάνει οπτικούς πομποδέκτες (**Optical Transceivers**) και οπτικούς κυκλοφορητές (**Optical Circulators**) οι οποίοι επιτρέπουν δικατευθυντική (**bidirectional**) οπτική μετάδοση μέσα σε μία οπτική ίνα. Τέλος γίνεται χρήση της τεχνολογίας WDM.

Στην επόμενη εικόνα απεικονίζεται στο αριστερό κομμάτι η γενικότερη μορφή της αρχιτεκτονικής του Proteus και στο δεξιά λεπτομερέστερα το πώς είναι δομημένα και τοποθετημένα τα οπτικά στοιχεία που αποτελούν την αρχιτεκτονική:



(a) Overall structure. (b) Optical components for one ToR.

Εικόνα 3. 18 - Αρχιτεκτονική Proteus

Το μοντέλο αυτό αποκαλείται Proteus-2560 και περιέχει ένα κύκλωμα MEMS 320 θυρών και μπορεί να συνδέει με άνεση 80 ToRs τα οποία περιέχουν 2560 servers. Το κάθε ToR είναι ένας ηλεκτρικός μεταγωγέας (τα μόνα ηλεκτρικά στοιχεία της αρχιτεκτονικής) με 64 10-GigE θύρες χωρίς φραγή. 32 από αυτές τις θύρες συνδέονται με διακομιστές ενώ οι υπόλοιπες συνδέονται με τα οπτικά στοιχεία και μεταφέρουν τα δεδομένα. Κάθε μία από τις 32 θύρες που συνδέεται με το οπτικό κομμάτι έχει έναν πομποδέκτη που σχετίζεται με ένα μοναδικό και προκαθορισμένο μήκος κύματος για την αποστολή και την αποδοχή δεδομένων. Η τεχνολογία WDM επιτρέπει στα δεδομένα από διαφορετικές θύρες να πολυπλεχθούν σε μία οπτική ίνα χωρίς να δημιουργείται contention των μήκων κύματος. Ο πομποδέκτης χρησιμοποιεί ξεχωριστές οπτικές ίνες για να συνδεθεί με τις υποδομές αποστολής και λήψης.

Προχωρώντας στην ανάλυση των οπτικών κομματιών της αρχιτεκτονικής παρατηρούμε από την εικόνα ότι η ίνα αποστολής από τους πομποδέκτες των 32 θυρών των ToR, που συνδέονται με το οπτικό κομμάτι, συνδέεται συγκεκριμένα με έναν οπτικό πολυπλέκτη. Ο πολυπλέκτης τροφοδοτεί ένα 1 × 4 WSS. Το WSS με τη σειρά του χωρίζει το σύνολο των 32 μήκων κύματος που βλέπει σε 4 ομάδες, όπου κάθε ομάδα μεταδίδεται με ξεχωριστή οπτική ίνα. Αυτές οι οπτικές ίνες συνδέονται με τον μεταγωγέα MEMS μέσω κυκλοφορητών έτσι ώστε να επιτρέπεται μέσω αυτών η δικατευθυντική κίνηση.

Στο δεξιότερο κομμάτι της απεικονιζόμενης αρχιτεκτονικής είναι η υποδομή λήψης. Οι 4 οπτικές ίνες λήψης που προέρχονται από κάθε ένα από τους 4 κυκλοφορητές λήψης, συνδέονται σε ένα συζεύκτη ισχύος (**power coupler**) ο οποίος συνδυάζει τα μήκη κύματος τους σε μία οπτική ίνα. Ο συζεύκτης ισχύος είναι μία οπτική συσκευή του οποίου η λειτουργία είναι παρόμοια με ενός πολυπλέκτη αλλά απλούστερη. Αυτή η οπτική ίνα τροφοδοτεί ένα αποπολυπλέκτη, ο οποίος διαχωρίζει κάθε εισερχόμενο μήκος κύματος στην αντίστοιχη θύρα για την οποία προορίζεται.

Είναι σημαντικό σε αυτό το σημείο να αναφερθεί ότι κάθε ToR μπορεί να επικοινωνήσει ταυτόχρονα με άλλα 4 ToRs. Διαφορετικά πρωτότυπα της αρχιτεκτονικής Proteus έχουν δοκιμαστεί με διαφορετικό μέγεθος της παραμέτρου k όπου ως k ορίζεται ο αριθμός των ToRs με τον οποίο μπορεί να συνδεθεί ένα μόνο ToR. Για παράδειγμα, εάν συνδεθούν N ToRs σε μία θύρα του MEMS τότε το κάθε ToR μπορεί να επικοινωνήσει αποκλειστικά με ένα μόνο ToR από τα N αυτά. Εάν πάλι συνδεθούν N/k ToRs σε k θύρες του MEMS (όπου k > 1) τότε το κάθε ToR μπορεί να επικοινωνήσει με k ToRs ταυτόχρονα. Το συγκεκριμένο μοντέλο Proteus χρησιμοποιεί k = 4 το οποίο έχει αξιοποιηθεί ως αρκετά λειτουργικό και επίσης η αναδιαμόρφωση του μεταγωγέα MEMs γίνεται σε μικρό και επιτρεπτό χρόνο καθυστέρησης. Αυτές οι αναδιαμορφώσεις στην αρχιτεκτονική γίνονται από ένα διαχειριστή τοπολογίας (**TM**). Ο TM αποκτάει τη μήτρα κίνησης από τους ToR μεταγωγείς, υπολογίζει τις κατάλληλες αναδιαμορφώσεις και τις προωθεί όπου χρειάζεται αντίστοιχα στα MEMS, WSS, και ToRs.

Δεδομένου τώρα ότι υπάρχει μία προς διαμόρφωση τοπολογία γίνεται χρήση επικοινωνίας τμήμα προς τμήμα (**hop-by-hop**) για να επιτευχθεί η συνδεσιμότητα. Για να επικοινωνήσει ένα ToR με άλλα ToRs τα οποία δεν συνδέονται μαζί του μέσω του MEMS, το πρώτο ToR χρησιμοποιεί μία από τις k συνδέσεις του. Αυτό το πρώτο-hop ToR (μία από τις προαναφερθέντες k συνδέσεις) λαμβάνει την μετάδοση μέσω οπτικής ίνας, τη μετατρέπει σε ηλεκτρικά σήματα, διαβάζει την κεφαλίδα του πακέτου και το αναμεταδίδει προς τον προορισμό του. Γενικά στην αρχιτεκτονική Proteus, σε κάθε hop, κάθε πακέτο βιώνει τη μετατροπή από οπτικό σε ηλεκτρικό και στη συνέχεια ξανά οπτικό (O-E-O). Η πρόσθετη λανθάνουσα

καθυστέρηση που επιβάλλεται από αυτή την O-E-O μετατροπή είναι αρκετά μικρή για να αγνοηθεί της τάξεως των nanoseconds.

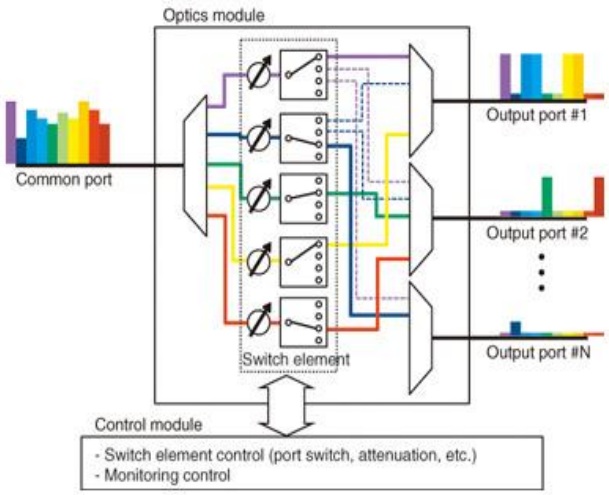
Τέλος να αναφερθεί ότι η αρχιτεκτονική Proteus είναι μία εύκολα διαμορφώσιμη αρχιτεκτονική η οποία καθίσταται κατάλληλη για κέντρα δεδομένων των οποίων η διαμόρφωση χρειάζεται συχνά αλλαγές στη δομή της. Χαρακτηρίζεται επίσης από χαμηλό κόστος κατασκευής, χαμηλή κατανάλωση ενέργειας, χαμηλή πολυπλοκότητα σύνδεσης και πολύ λιγότερα καλώδια από ηλεκτρονικά δίκτυα και μπορεί να συντηρήσει τις μελλοντικές ταχύτητες της τάξεως των 40-GigE και 100-GigE που θα αποκτήσουν τα κέντρα δεδομένων [31].

3.4 Αρχιτεκτονικές δικτύου βασισμένες σε διακόπτες επιλογής μήκους κύματοςWSS (Wavelength Selective Switch)

Το κεφάλαιο αυτό επικεντρώνεται στην χρήση των WSS – wavelength selective switches στα κέντρα δεδομένων και στις αρχιτεκτονικές που είναι βασισμένες σε αυτά τα στοιχεία. Αρχικά γίνεται η παρουσίαση ενός WSS και της βασικής του λειτουργίας για να ολοκληρώσουμε στη συνέχεια με τους διάφορους τρόπους υλοποίησης του. Αναφέρονται εν συντομία μερικά πλεονεκτήματα και μειονεκτήματα του κάθε τρόπου υλοποίησης. Στο δεύτερο μέρος γίνεται η παρουσίαση και η ανάλυση των αρχιτεκτονικών που χρησιμοποιούν WSS και προορίζονται για οικειοποίηση από τα κέντρα δεδομένων.

3.4.1 Ανάλυση λειτουργίας WSS

Τα WSS είναι στοιχεία τα οποία χρησιμοποιούνται σε WDM (wavelength division multiplexing) οπτικά δίκτυα για να δρομολογήσουν σήματα μεταξύ των οπτικών ινών, σύμφωνα με το μήκος κύματος [33]. Ένα WSS μπορεί δυναμικά να δρομολογήσει, να μπλοκάρει ακόμα και να εξασθενίσει την ισχύ καθενός από τα WDM πολυπλεγμένα κύματα μέσα σε ένα node του δικτύου. Στην εικόνα όπου ακολουθεί φαίνεται η γενική μορφή ενός WSS και διαγράφονται οι λειτουργίες του:



Εικόνα 3. 19 - WSS λειτουργικότητα

Όπως παρατηρούμε,ένα WSS αποτελείται από μια ενιαία κοινή οπτική θύρα (**common port**) και N θύρες εξόδου πολλαπλών μηκών κύματος, όπου κάθε WDM κανάλι εισόδου από την κοινή θύρα μπορεί να δρομολογηθεί σε κάθε μία από τις N θύρες πολλαπλών μηκών κύματος, ανεξάρτητα από το πώς δρομολογούνται όλα τα άλλα κανάλια. Η μεταγωγή με βάση το μήκος κύματος είναι η μία από τις 2 βασικότερες λειτουργίες ενός WSS. Η δεύτερη είναι η εξασθένιση (attenuation) σύμφωνα με την οποία το επίπεδο ισχύος του μεταδιδόμενου φωτός ρυθμίζεται ξεχωριστά για κάθε μήκος κύματος.

Το hardware χωρίζεται σε μία μονάδα οπτικού υλικού και τη μονάδα ελέγχου όπως βλέπουμε και από το σχήμα. Αυτές οι διαδικασίες εξασθένισης και δρομολόγησης του μήκους κύματος μπορούν να ελεγχτούν και να αλλάξουν δυναμικά μέσω μιας ηλεκτρονικής διεπαφής ελέγχου επικοινωνίας στο WSS η οποία συνδέεται με τη μονάδα ελέγχου. Ένα WSS έχει πολλές θύρες εισόδου και εξόδου, έτσι χρησιμοποιείται ένα φράγμα περίθλασης (diffraction grating) το οποίο είναι ικανό να πολυπλέξει και να αποπολυπλέξει ταυτόχρονα τα σήματα σε μία θύρα [33]. Τα ενεργά στοιχεία περιλαμβάνουν ένα χωρικό διαμορφωτή φωτός (**spatial light modulator**), ο οποίος μπορεί να είναι είτε ένα σύστημα MEMS είτε υγροί κρύσταλλοι σε πυρίτιο (**liquid crystal on silicon – LcoS**), ο οποίος μπορεί να αλλάξει την αντανακλώμενη κατεύθυνση της δέσμης φωτός εισόδου. Τα μονοπάτια των ακτινών του φωτός είναι διαφορετικά για κάθε συνδυασμό μήκους κύματος και θύρας, έτσι οι ακτίνες φωτός διασχίζουν η μία την άλλην μέσα στην οπτική μονάδα του WSS.

Σχετικά με το τρόπο λειτουργίας του WSS, το φως από μία ίνα ευθυγραμμίζεται με τη χρήση ενός φακού-lens με μήκος εστίασης f και αποπολυπλέκεται με περίθλαση από το φράγμα. Ένας φακός είναι μία διαπερατή οπτική συσκευή η οποία επηρεάζει την εστίαση μιας δέσμης φωτός διαμέσου της διάθλασης. Η κατεύθυνση της δέσμης μετά το φράγμα θα εξαρτάται από το μήκος κύματος λ_0 της ακτίνας φωτός. Στη συνέχεια οι διαθλασμένες ακτίνες περνάνε μέσα από το lens για δεύτερη φορά, και το φασματικά διαχωρισμένο φώς εστιάζεται στον χωρικό διαμορφωτή φωτός. Επειδή σε κάθε 1 x N WSS μεταγωγέα το κάθε μήκος κύματος μπορεί να δρομολογηθεί σε κάθε μία από τις N θύρες εξόδου, αυτό καθιστά το WSS ως ένα αρκετά ευέλικτο μεταγωγέα για την υλοποίηση ενός OADM (optical add drop multiplexer) συστήματος με πολλαπλές add/drop θύρες [34].

Τέλος, να αναφερθεί ότι η ταχύτητα μεταγωγής των WSS εξαρτάται από την τεχνολογία υλοποίησης του χωρικού διαμορφωτή φωτός και είναι συνήθως της τάξης των ms, αν και υπάρχουν διαθέσιμες τεχνολογίες με δυνατότητα γρηγορότερης μεταγωγής της τάξης του 1 ms όπως για παράδειγμα η τεχνολογία DLP της εταιρείας TexasInstruments.

Οι τρόποι υλοποίησης του WSS όπως προαναφέρθηκε είναι οι ακόλουθοι:

- Microelectromechanical Mirrors (MEMS) :

Το απλούστερο και παλαιότερο εμπορικό WSS βασίστηκε σε κινητά κάτοπτρα [35] χρησιμοποιώντας MEMS. Το εισερχόμενο φως διαχωρίζεται σε ένα φάσμα από ένα πλέγμα περίθλασης και κάθε κανάλι μήκους κύματος, στη συνέχεια, επικεντρώνεται σε ένα ξεχωριστό MEMS καθρέφτη. Με την κλίση του κατόπτρου σε μία διάσταση, το κανάλι μπορεί να κατευθύνεται πίσω σε οποιαδήποτε από τις ίνες στη συστοιχία. Η κλίση σε δεύτερο άξονα επιτρέπει την ελαχιστοποίηση της διαφωνίας- crosstalk η οποία οφείλεται σε γειτονικά MEMS.

Η τεχνολογία αυτή έχει το πλεονέκτημα μίας ενιαίας επιφάνειας καθοδήγησης που δεν απαιτεί οπτική ποικιλομορφία πόλωσης (polarization diversity). Λειτουργεί καλά υπό την παρουσία ενός συνεχούς σήματος, επιτρέποντας τα κυκλώματα παρακολούθησης του καθρέφτη να περιορίζουν τα λάθη που οφείλονται στο γειτονικό crosstalk και να επιτυγχάνουν μεγιστοποίηση της σύζευξης. Τα MEMS based WSS παράγουν συνήθως καλό λόγο σβέσης (**extinction ratio**) αλλά κακή απόδοση ανοικτού βρόχου για τον καθορισμό ενός συγκεκριμένου επιπέδου εξασθένησης.

-Binary Liquid Crystal (LC):

Το υγρό κρύσταλλο μεταγωγής αποφεύγει τόσο το υψηλό κόστος κατασκευής των MEMS και ενδεχομένως μερικούς από τους περιορισμούς τους [36].

Ένα φράγμα περίθλασης διαθλά το εισερχόμενο φως σε ένα φάσμα. Μία στοίβα από δυαδικό υλικό υγρών κρυστάλλων ελεγχόμενη από λογισμικό, κλείνει ξεχωριστά κάθε οπτικό κανάλι και ένα δεύτερο φράγμα χρησιμοποιείται για να επανασυνδέσει φασματικά τις ακτίνες. Η Οπτική ποικιλομορφία πόλωσης απαιτείται και εξασφαλίζει χαμηλές απώλειες που οφείλονται στη πόλωση.

Αυτή η τεχνολογία έχει τα πλεονεκτήματα του σχετικά χαμηλού κόστους εξαρτημάτων, απλό ηλεκτρονικό έλεγχο και σταθερές θέσεις των ακτινών χωρίς ενεργό ανάδραση. Αυτή η τεχνολογία είναι επίσης ικανή να διαμορφώνει σε ένα εύκαμπτο φάσμα πλέγματος με τη χρήση ενός λεπτού πλέγματος pixel. Το κύριο μειονέκτημα αυτής της τεχνολογίας προκύπτει από το πάχος των στοιβαγμένων στοιχείων μεταγωγής. Κρατώντας την οπτική δέσμη αυστηρά επικεντρωμένη πάνω από αυτό το βάθος είναι δύσκολο και έχει, μέχρι στιγμής περιορίσει την ικανότητα των WSS πολλαπλών θυρών να επιτύχουν πολύ καλή κοκκιότητα (**granularity**), συνήθως 12.5 GHz ή χαμηλότερα.

-Liquid Crystal on Silicon (LcoS):

Το LcoS είναι ιδιαίτερα ελκυστική ως μηχανισμός μεταγωγής σε ένα WSS λόγω της σχεδόν συνεχής ικανότητα επέμβασης, που επιτρέπει πολύ νέες λειτουργίες [37]. Ειδικότερα, οι ζώνες μηκών κύματος τα οποία μεταγωγούνται μαζί δεν χρειάζεται να έχουν διαμορφωθεί στο οπτικό υλικό αλλά μπορούν να προγραμματιστούν μέσα στο μεταγωγέα μέσω του ελέγχου του λογισμικού. Επιπλέον, είναι δυνατόν να επωφεληθούν από αυτή την ικανότητα για την αναδιαμόρφωση των καναλιών ενώ ταυτόχρονα η συσκευή λειτουργεί.

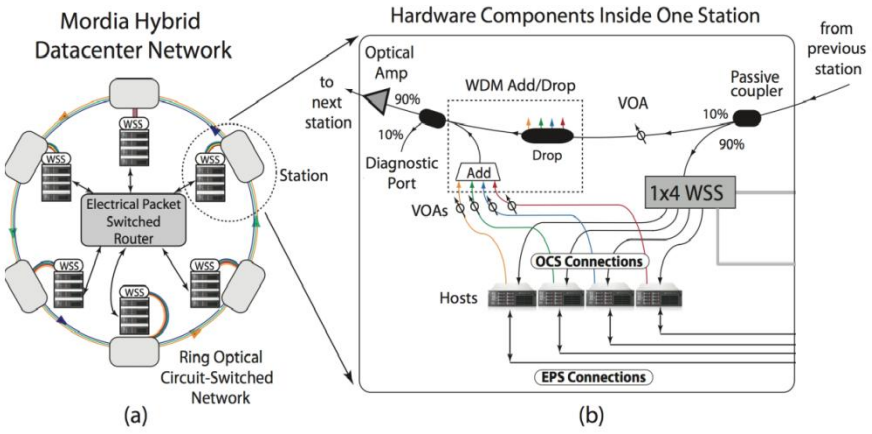
Η τεχνολογία LcoS επέτρεψε την εισαγωγή πιο ευέλικτων δικτύων πολυπλεξίας μήκους κύματος που βοηθούν να απελευθερωθεί η πλήρη φασματική ικανότητα των οπτικών ινών.

3.4.2 Mordia WSS-based Network

Το δίκτυο Mordia [38] [39] πρόκειται για μία υβριδική αρχιτεκτονική η οποία περιέχει και ηλεκτρικό κύκλωμα μεταγωγής και οπτικό κύκλωμα μεταγωγής το οποίο στηρίζεται σε μεταγωγείς WSS. Το πρωτότυπο Mordia μοντέλο περιέχει 24 hosts και λειτουργεί ταχύτατα με χρόνο μεταγωγής μόνο στα 11,5 μs, χάρη στην πολύ γρήγορη μεταγωγή των WSS. Επίσης σε μερικές του υλοποιήσεις το δίκτυο Mordia σχεδιάζεται έτσι ώστε η επεξεργασία των δεδομένων να λαμβάνει μέρος στο ToR επίπεδο, πράγμα που εξυπηρετεί αρκετά στη μείωση του latency. Αρχικά θα γίνει η παρουσίαση και η μελέτη της αρχιτεκτονικής και δομής του Mordia, έπειτα

μία σύντομη ανάλυση των δομικών του στοιχείων και τέλος θα αναφερθούμε στο πρωτόκολλο σχεδιασμού γύρω από το οποίο λειτουργεί το Mordia καθώς είναι από τους κύριους παράγοντες που το καθιστούν τόσο ευέλικτο [38].

Στην επόμενη εικόνα ακολουθεί η αρχιτεκτονική του Mordia:



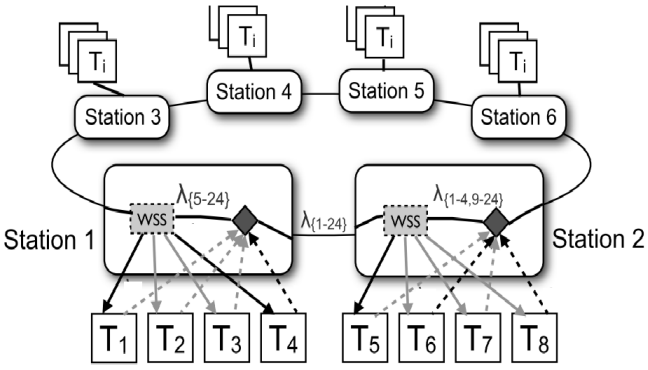
Εικόνα 3. 22 - a) Mordia RIng b) αναλυτική δομή ενός από τα Mordia Stations

Στο αριστερό κομμάτι παρατηρούμε την υβριδική αρχιτεκτονική σε μορφή δακτυλίου ενώ δεξιά την υλική υλοποίηση ενός από τους 6 σταθμούς (stations) που αποτελούν το δίκτυο. Ο κάθε host αποτελείται από μία 10G Ethernet Network Interface Card (NIC) διπλής θύρας με 2 SFP + συνδέσεις. Μία από τις προαναφερθείσες θύρες συνδέεται στο ηλεκτρικό κύκλωμα, συνήθως 10G Ethernet EPS. Η δεύτερη θύρα συνδέεται στο οπτικό κύκλωμα OCS το οποίο είναι βασισμένο στα WSS. Αυτός ο μεταγωγέας WSS χρησιμοποιείται για να δρομολογήσει ένα προκαθορισμένο κανάλι μήκους κύματος από μία θύρα σε ένα host σε μία άλλη θύρα σε ένα ή και πολλαπλούς hosts. Τα δύο δίκτυα λειτουργούν παράλληλα παράγοντας έτσι ένα υβριδικό δίκτυο.

Η φυσική τοπολογία του Mordia, όπως διαγράφεται και από την εικόνα, είναι ένας μονοκατευθυντικός δακτύλιος που αποτελείται από N μήκη κύματος μέσα σε μία μόνο οπτική ίνα [39]. Αντιθέτως η λογική τοπολογία του Mordia είναι ένα mesh αφού όλα τα stations επικοινωνούν μεταξύ τους. Συνεπώς αυτή η αρχιτεκτονική του OCS υποστηρίζει μονοεκπομπή του κυκλώματος, πολυεκπομπή, ευρυεκπομπή και λειτουργία βρόχου.

Σε κάθε host ανατίθεται το δικό του προκαθορισμένο μήκος κύματος μετάδοσης,χρησιμοποιώντας εμπορικά διαθέσιμα DWDM SFP + μοντέλα, στο πλέγμα των 100 GHz ITU. Στην αρχιτεκτονική του Mordia ο αριθμός των hosts είναι 24.

Τα μήκη κύματος προστίθενται ή αφαιρούνται από το δαχτυλίδι σε έξι σταθμούς-stations. Μία καλύτερη απεικόνιση των 6 stations, των ToR που ανταλλάσουν δεδομένα και του OCS φαίνεται στην παρακάτω απεικόνιση:



Εικόνα 3. 20 – Αναπαράσταση Mordia Network

Ο κάθε station μπορεί να υποστηρίξει μέχρι και 4 hosts. Τα μήκη κύματος για κάθε host σε ένα σταθμό απέχουν μεταξύ τους κατά 100 GHz. Το σύνολο των προκαθορισμένων μήκων κύματος μετάδοσης που έχουν ανατεθεί στον επόμενο σταθμό αντισταθμίζεται κατά 400 GHz. Τα 4 μήκη κύματος αυτά που προέρχονται από τον ίδιο σταθμό εισέρχονται στο δακτύλιο χρησιμοποιώντας ένα ζωνοπερατό φίλτρο προσθήκης/απόρριψης (**add / drop**). Η οπτική ισχύ από κάθε host προσαρμόζεται πριν από την προώθηση των κυμάτων στο δακτύλιο με τη χρήση ενός οπτικού μεταβλητού εξασθενητή VOA (variable optical attenuator). Η οπτική μεταγωγή σε κάθε σταθμό γίνεται έξω από τον δακτύλιο με τη χρήση ενός WSS το οποίο στη συγκεκριμένη αρχιτεκτονική είναι μια παραλλαγή του 1 × 4-port Nistica Full Fledge 100 WSS μοντέλου. Στην είσοδο του κάθε σταθμού ένας παθητικός διαιρέτης ισχύος, ο οποίος είναι ανεπηρέαστος από το μήκος κύματος, κατευθύνει το 90% της ισχύος έξω από το δακτύλιο και μέσα σε ένα WSS. Το εναπομείναν 10% του σήματος παραμένει και προωθείται στον δακτύλιο. Επειδή ο διαιρέτης είναι παθητικός, η είσοδος προς την μονάδα WSS σε καθένα από τους έξι σταθμούς περιέχει και τα 24 μήκη κύματος [39].

Το κάθε WSS στοιχείο μπορεί να δρομολογήσει ένα ελεγχόμενο κλάσμα οποιουδήποτε καναλιού μήκους κύματος σε κάθε μία από τις 4 θύρες εξόδου. Η αρχιτεκτονική που προκύπτει με αυτό τον τρόπο είναι ένα δίκτυο επιλογής και ευρυεκπομπής (broadcast and select) στο οποίο κάθε προκαθορισμένο σήμα μήκους κύματος από κάθε θύρα-πομπό μπορεί να δρομολογηθεί σε οποιαδήποτε θύρα-δέκτη. Αυτού του είδους η αρχιτεκτονική απαιτεί ο δέκτης να είναι ανεπηρέαστος από το μήκος κύματος σε όλο το εύρος των μηκών κύματος μετάδοσης.

Τα τέσσερα μήκη κύματος που εισέρχονται σε κάθε σταθμό ταξιδεύουν μία φορά στον δακτύλιο και δεν τους επιτρέπεται παραπάνω από ένα πέρασμα από το add/drop φίλτρο. Αυτό το φίλτρο διοχετεύει επίσης το σήμα σε αυτό το σταθμό στον δακτύλιο. Σε κάθε σταθμό, το σήμα στον δακτύλιο ενισχύεται χρησιμοποιώντας έναν 23 dBm οπτικό ενισχυτή (Optilab EDFA-B-23-M). Όλες οι ενισχύσεις λαμβάνουν χώρα μέσα στον δακτύλιο και όλες οι μεταγωγές γίνονται έξω από τον δακτύλιο. Αυτός ο σχεδιασμός αποτρέπει τις μεταβατικές διακυμάνσεις ισχύος στους οπτικούς ενισχυτές κατά τη διάρκεια της αναδιάρθρωσης του κυκλώματος. Όλα τα παραπάνω απεικονίζονται και στο δεξιό μέρος της εικόνας 3.22.

Στην συνέχεια, θα γίνει μία σύντομη παρουσίαση των δομικών στοιχείων και της επιλογής τους. Η επιλογή των WSS έγινε καθώς τα συγκεκριμένα οπτικά στοιχεία που χρησιμοποιήθηκαν (τεχνολογίας DLP) αποτελούν πολύ γρήγορα στοιχεία μεταγωγής σε αντίθεση με στοιχεία όπως τα MEMS. Επίσης το κάθε WSS πέρα από το common port και τις θύρες εξόδου μπορεί να δομηθεί έτσι ώστε να έχει και ένα bypass port και στην συγκεκριμένη αρχιτεκτονική είναι πολύ σημαντική καθώς χρησιμοποιείται ως η θύρα που θα μεταφέρει τα μήκη κύματος που δεν απορροφήθηκαν από τους hosts του συγκεκριμένου σταθμού.

Όσο αφορά τα ToRs (hosts), το κάθε ToR συνδέεται με το OCS μέσω μίας η παραπάνω οπτικών ζεύξεων και εσωτερικά εμπεριέχει N-1 ουρές από εξερχόμενα πακέτα, μία για κάθε μία από τις N-1 θύρες εξόδου [39]. Το κάθε ToR συμμετέχει σε ένα επίπεδο ελέγχου, το οποίο ενημερώνει το ToR για το βραχυπρόθεσμο πρόγραμμα των επικείμενων διαρθρώσεων του κυκλώματος Με τον τρόπο αυτό, τα ToR γνωρίζουν ποια κυκλώματα θα δημιουργηθούν στο άμεσο μέλλον, και μπορεί να χρησιμοποιήσουν αυτή τη γνώση για να κάνουν αποτελεσματική χρήση των κυκλωμάτων από τη στιγμή που αυτά εγκατασταθούν. Αρχικά το ToR δεν αποστέλλει κανένα πακέτο στο δίκτυο, και απλά περιμένει να συγχρονιστεί με το Mordia OCS. Αυτός ο συγχρονισμός είναι απαραίτητος καθώς τα OCS δεν μπορούν να αποθηκεύσουν τα πακέτα και έτσι τα ToR πρέπει να τραβήξουν τα πακέτα από την κατάλληλη σειρά σε τέλειο συγχρονισμό με την επανεγκατάσταση του OCS. Αυτός ο συγχρονισμός αποτελείται από 2 βήματα:

1) την λήψη του προγράμματος από έναν χρονοπρογραμματιστή μέσω ενός ανεξάρτητου καναλιού
2) έχοντας πλήρη επίγνωση της παρούσας κατάστασης του OCS

Control Plane:

Το πεδίο ελέγχου του Mordia αποτελείται από έναν Linux Host για να εκτελεί μη πραγματικόχρονες επεξεργασίες, ένα FPGA board για την εκτέλεση των πραγματικόχρονων επεξεργασιών, τα 6 WSS modules και έναν 10G Ethernet μεταγωγέα. Το δίκτυο Mordia χρησιμοποιεί TDMA για τον συγχρονισμό των hosts.

Το FPGA συγχρονίζει τους hosts και τα WSS μεταδίδοντας ένα πακέτο συγχρονισμού σε όλους τους κεντρικούς υπολογιστές και στο EPS.

Σε αυτό το σημείο κρίνεται σημαντικό να γίνει μία παρουσίαση του πρωτοκόλλου που ακολουθεί η αρχιτεκτονική Mordia καθώς είναι ένας καθοριστικός παράγοντας στην τόσο γρήγορη μεταγωγή του κυκλώματος της τάξεως των 10-11,5 μs [40].

Σε γενικές γραμμές οι προηγούμενες υβριδικές αρχιτεκτονικές κάνανε χρήση παρόμοιων μοντέλων σχεδιασμού κυκλωμάτων που ονομάζονται HSS (hotspot scheduling) [40]. Σύμφωνα με το μοντέλο HSS αρχικά γίνεται η μέτρηση της μήτρας κίνησης μέσα στα pods, ύστερα γίνεται ο υπολογισμός της μήτρας ζήτησης, μετά η αναγνώριση των hotspots και τέλος ένας χρονοπρογραμματιστής εγκαθιστά τα κυκλώματα που απαιτούνται για την αντιμετώπιση της κίνησης. Το HSS όμως απαιτεί και απαιτητικούς αλγόριθμους για να λειτουργήσει πριν από κάθε αναδιαμόρφωση του κυκλώματος και δεν είναι πάντα απόλυτα ακριβές.

Στο δίκτυο Mordia γίνεται χρήση ενός άλλου μοντέλου το οποίο ονομάζεται TMS – traffic matrix scheduling (σχεδιασμός μήτρας κίνησης), όπου το μεγαλύτερο κομμάτι της κίνησης, εάν όχι όλο, δρομολογείται μέσω κυκλωμάτων. Το TMS χρονομερίζει πολύ γρήγορα κυκλώματα σε

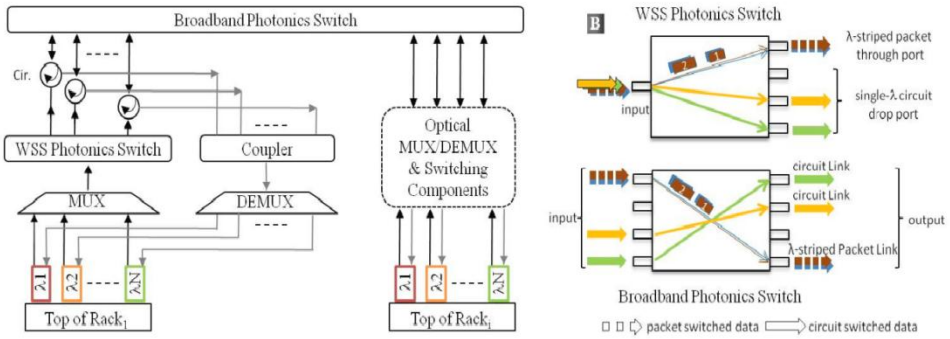
πολλούς διαφορετικούς προορισμούς σε χρόνο της τάξεως των microseconds. Ένας λόγος που το TMS δεν έχει προταθεί πριν είναι ότι η εφαρμογή του ήταν ανέφικτη αν αναλογιστούμε ότι η χρόνο κλίμακα τους ήταν της τάξεως των milliseconds. Μόνο σε χρονικά πλαίσια της τάξεως των microseconds μπορεί να γίνει εφικτή η χρήση του TMS. Το TMS κάνει τη μεταγωγή των κυκλωμάτων πολύ πιο αποτελεσματική από ό, τι το HSS πρώτης γενιάς επειδή πολύ περισσότερη από τη κίνηση του δικτύου μπορεί να οδηγηθεί προς τα κυκλώματα. Ενώ το HSS συζευγνύει τον σχεδιασμό και τη μεταγωγή μαζί, το TMS τα αποσυζευγνύει. Το TMS χρησιμοποιεί απαιτητικούς αλγόριθμους για να κατασκευάσει ένα χρονοδιάγραμμα μεταγωγής κυκλωμάτων, αλλά σε αντίθεση με το HSS, όταν το χρονοδιάγραμμα έχει κατασκευαστεί, τότε εφαρμόζεται σε hardware επίπεδο σε κλίμακα χρόνου της τάξεως των microseconds. Αυτός ο διαχωρισμός του σχεδιασμού και της μεταγωγής είναι που επιτρέπει στο TMS να χρονοπρογραμματίσει ένα μεγαλύτερο μέρος της κίνησης από το HSS παρά το γεγονός ότι οι αλγόριθμοι δεν είναι πιο γρήγοροι ή καλύτεροι.

Ενώ ο μεταγωγέας του κυκλώματος είναι απασχολημένος με το να δημιουργεί και να εγκαθιστά πολύ γρήγορα κυκλώματα δρομολόγησης (ή το ανάποδο) οι hosts σε κάθε pod έξυπνα παρατηρούν την κατάσταση του μεταγωγέα κυκλώματος ώστε να ξέρουν πότε είναι η κατάλληλη στιγμή για να μεταδώσουν πακέτα. Με άλλα λόγια οι hosts χρησιμοποιούν ένα πρωτόκολλο γνωστό ως TDMA – time division multiple access, αντί να χρησιμοποιούν καθιερωμένα πρωτόκολλα πακέτων. Με το TDMA οι hosts αποκτούνε πρόσβαση στο δίκτυο και στα κυκλώματα που αυτό δημιουργεί. Το TDMA επιτρέπει σε πολλούς χρήστες να μοιράζονται το ίδιο κανάλι συχνότητας διαιρώντας το σήμα σε διαφορετικές χρονοθυρίδες (time slots). Οι χρήστες μεταδίδουν σε γρήγορη διαδοχή, ο ένας μετά τον άλλο, και ο καθένας με τη δική του χρονοθυρίδα. Όλα τα προηγούμενα συμβάλουν εξαιρετικά στη μείωση του χρόνου μεταγωγής του κυκλώματος και γενικά του δικτύου και προσφέρουν μία επιπλέον δυναμικότητα στην αρχιτεκτονική Mordia. Περισσότερες πληροφορίες για το TMS και τον αλγόριθμο που χρησιμοποιεί μπορούν να βρεθούν στο [40].

3.4.3 Μια υβριδική αρχιτεκτονική με WSS για κέντρα δεδομένων με υψηλές αποδόσεις

Στην ενότητα αυτή παρουσιάζεται μία υβριδική πλατφόρμα η οποία κάνει χρήση και των WSS αλλά και την επεξεργασία μικρών ροών δεδομένων μέσω ενός επιπλέον οπτικού δικτύου μεταγωγής με γρηγορότερη απόκριση και έτσι επιτυγχάνει μεγάλες αποδόσεις στα κέντρα δεδομένων [41]. Αυτή η αρχιτεκτονική η οποία είναι βασισμένη σε διαδοχικούς συνδεόμενους μίκρο- δακτυλίους (microrings) από πυρίτιο και σε ημιαγώγιμους οπτικούς ενισχυτές (**SOAs**), υποστηρίζει την αναδιάρθρωση των πακέτων και της μεταγωγής του κυκλώματος και είναι επεκτάσιμη και ενεργειακά αποτελεσματική. Συνδυάζοντας την επιλεκτική ικανότητα με βάση το μήκος κύματος των microrings και την ευρυζωνική συμπεριφορά του μεταγωγέα SOA, επιτυγχάνεται γρήγορες μεταβάσεις μεταγωγής, υψηλός λόγος σβέσης και χαμηλότερη κατανάλωση ενέργειας, πράγματα τα οποία είναι όλα απαραίτητα για τα μελλοντικά κέντρα δεδομένων.

Η αρχιτεκτονική της συγκεκριμένης πλατφόρμας περιγράφεται στην επόμενη εικόνα:



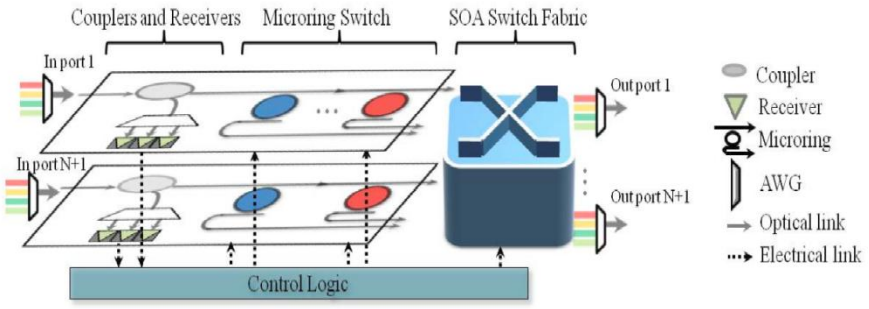
Εικόνα 3. 21 – WSS αρχιτεκοτνική υψηλών επιδόσεων

Η οπτική πλατφόρμα διασύνδεσης αποτελείται από οπτικούς πολυπλέκτες (Mux) και αποπολυπλέκτες (Demux), κυκλοφορητές, συζεύκτες, και υβριδικές μονάδες μεταγωγής οι οποίες αποτελούνται από ένα WSS και ένα ευρυζωνικό μεταγωγέα (**broadband switch**). Σε αυτό το σχεδιασμό, ένας WSS φωτονικός μεταγωγέας διαμερίζει μία εισερχόμενη δέσμη από μήκη κύματος (τα πολυπλεγμένα WDM σήματα από τους διακοσμιτές ή τα racks) από την θύρα εισόδου common port σε διαφορετικές θύρες εξόδου. Στη συνέχεια, όλα τα οπτικά σήματα συνδέονται με ένα ευρυζωνικό φωτονικό μεταγωγέα για να επιτύχουν ένα προς ένα οπτικό/ηλεκτρικό κύκλωμα μεταγωγής. Όπως παρατηρούμε στο δεξιότερο κομμάτι της προηγούμενης εικόνας, κάθε drop-θύρα του WSS μεταγωγέα κάνει χρήση ενός ξεχωριστού και μοναδικού καναλιού μήκους κύματος. Οπτικά δεδομένα εμφανίζονται σε κάθε μία από αυτές τις drop-θύρες όταν ένας σύνδεσμος του κυκλώματος ενεργοποιείται για την αντίστοιχη θύρα. Διαφορετικά το κανάλι αυτό μήκους κύματος κατευθύνεται προς τη θύρα μεταγωγής πακέτων (packet-switched through port). Τα συστήματα δρομολόγησης και των δύο μεταγωγέων αλληλοσυμπληρώνονται: Ο WSS μεταγωγέας υποστηρίζει την ταυτόχρονη μεταγωγή των πακέτων και των κυκλωμάτων ενός προεπιλεγμένου συνδυασμού μήκων κύματος από μία μόνο θύρα εισόδου σε διαφορετικές θύρες εξόδου, καταλήγοντας έτσι σε αποτελεσματική δρομολόγηση του μήκους κύματος. Ο ευρυζωνικός φωτονικός μεταγωγέας υποστηρίζει ταυτόχρονα τη μεταγωγή και των πακέτων και του κυκλώματος των καναλιών που έχει αποδεσμευτεί το μήκος κύματος τους. Η προτεινόμενη αρχιτεκτονική χαρακτηρίζεται όχι μόνο από την ταυτόχρονη δρομολόγηση των οπτικών πακέτων και την αναγνώριση της κίνησης στο κύκλωμα αλλά υποστηρίζει επίσης την αναδιάρθρωση του μήκους κύματος χρησιμοποιώντας τα WSS.

Η προτεινόμενη πλατφόρμα είναι υλοποιήσιμη χρησιμοποιώντας μεταγωγείς βασισμένους σε silicon microrings και ημιαγώγιμους οπτικούς ενισχυτές (SOA). Και οι δύο τύποι μεταγωγέων είναι συμπαγείς σε μέγεθος και ικανοί για τη δρομολόγηση υψηλού εύρους ζώνης μήκους κύματος με ταχύτητα μεταγωγής στη κλίμακα των ns, μέσω ενός φωτονικού δικτύου διασύνδεσης. Πιο συγκεκριμένα οι silicon microring μεταγωγείς είναι ανταποκρίσιμοι στο μήκος κύματος, επιτρέποντας έτσι την ευελιξία της επιλογής διαφορετικών μήκων κύματος για τις θύρες μεταγωγής, ενώ οι SOA-based μεταγωγείς προσφέρουν εξαιρετικά μεγάλο εύρος ζώνης. Η

κλιμακοθετησιμότητα μπορεί να επιτευχτεί με την επέκταση της χαμηλής ισχύος και χαμηλού κόστους μεταγωγής πλατφόρμας.

Τέλος θα ακολουθήσει μία αναφορά στα δομικά στοιχεία της αρχιτεκτονικής [41]. Στην επόμενη εικόνα απεικονίζεται ένα σχηματικό διάγραμμα του υβριδικού μεταγωγέα:



Εικόνα 3. 22- Υβριδικός μεταγωγέας SOA

Το δικτυοδόμημα του μεταγωγέα αποτελείται από microring συντονιστές, SOAs, Φωτο-ανιχνευτές PIN, λογικά κυκλώματα ηλεκτρονικού ελέγχου καθώς και παθητικά οπτικά στοιχεία (π.χ. συζεύκτες και φίλτρα μήκους κύματος). Ο αριθμός των καναλιών μήκους κύματος που χρησιμοποιούνται για τη μεταγωγή των πακέτων μπορεί να ρυθμιστεί δυναμικά. Αντί οι έξοδοι των lasers να στέλνονται απευθείας στον broadband μεταγωγέα, το να περνάνε πρώτα από το WSS μπορεί να μειώσει τον αριθμό των θυρών στον broadband μεταγωγέα. Ωστόσο, πιο σημαντικά, ο WSS επιτρέπει μια δυναμική κατανομή των καναλιών μήκους κύματος που μπορεί να προσαρμόζεται ανάλογα με τις ανάγκες του κάθε κέντρου δεδομένων. Το σύστημα μπορεί να χρησιμοποιήσει τις συνδέσεις υψηλού εύρους ζώνης της μεταγωγής κυκλώματος για εφαρμογές με μεγάλες ροές δεδομένων ή μπορεί να αυξήσει την ευελιξία της μεταγωγής πακέτου για εφαρμογές με μικρά bursty πακέτα. Επιπλέον, κανάλια μήκους κύματος μπορούν να προστίθεται δυναμικά στον wavelength-striped μεταγωγέα πακέτων εφόσον απαιτείται επιπλέον απόδοση. Αυτή η λειτουργικότητα της κατανομής του καναλιού μήκους κύματος δεν θα ήταν εφικτή μόνο με τον broadband μεταγωγέα. Ο microring μεταγωγέας επιλέγει πρώτα διαφορετικά δεδομένα μήκους κύματος πάνω στις αντίστοιχες εξόδους με τον τρόπο του WSS πριν από την είσοδο στον broadband SOA μεταγωγέα ο οποίος περαιτέρω κατευθύνει τα εισερχόμενα δεδομένα στις εξόδους προορισμού τους. Με αυτό τον σχεδιασμό κάθε κανάλι μήκους κύματος της θύρας εισόδου του υβριδικού μεταγωγέα μπορεί να μεταχθεί σε κάθε θύρα εξόδου ανεξάρτητα, είτε σε πακέτα ή ροές ανάλογα με τις απαιτήσεις κίνησης. Ο αριθμός των καναλιών μήκους κύματος που χρησιμοποιείται για την μεταγωγή των πακέτων μπορεί και αυτός να ρυθμιστεί δυναμικά. Η χρήση των WSS έχει το πλεονέκτημα ότι ο αριθμός των θυρών του broadband μεταγωγέα θα είναι σημαντικά μικρότερος πράγμα που θα ελαχιστοποιήσει την κατανάλωση ενέργειας στα SOA καθώς επίσης θα ελαχιστοποιηθεί και η κεφαλίδα στα πακέτα. Και οι δύο μεταγωγείς ελέγχονται από ηλεκτρονικό λογικό κύκλωμα ελέγχου που προκύπτει από προγραμματιστικά λογικά στοιχεία ενώ το σήμα ελέγχου μπορεί να αποσπαστεί από ένα υποσύνολο ειδικών μηκών κύματος με τον τρόπο του ελέγχου διανομή.

Κεφάλαιο 4
Πεδίο ελέγχου και πρωτόκολλα οπτικών αρχιτεκτονικών

Το κεφάλαιο αυτό της εργασίας αφιερώνεται αποκλειστικά στο πεδίο ελέγχου – **control plane** των οπτικών αρχιτεκτονικών, καθώς ένα δίκτυο για να είναι αποδοτικό και αποτελεσματικό δεν απαιτείται μονάχα μια καλή αρχιτεκτονική δομή αλλά και ένα σωστό πεδίο ελέγχου το οποίο συμβάλει άμεσα και δυναμικά στη γρήγορη λειτουργία των αρχιτεκτονικών που αναφέρθηκαν στο προηγούμενο κεφάλαιο. Αρχικά, παρουσιάζονται οι αιτίες, που οδήγησαν στη δημιουργία της τεχνολογίας **SDN**, η χρήση της οποίας είναι πολύ διαδεδομένη, οι οποίες σχετίζονται με τη σταδιακή απαξίωση των παραδοσιακών πρωτοκόλλων και με την ανάδειξη του **Cloud Computing.** Στη συνέχεια ακολουθεί μία εκτενής αναφορά στην έννοια του SDN και στη λειτουργία του. Ακολουθεί μετά από αυτό η παρουσίαση διαφόρων πρωτοκόλλων με μεγαλύτερη έμφαση στο **Open-Flow**. Τέλος αναφέρονται μερικά μελλοντικά σχέδια και προκλήσεις για τα control planes στα κέντρα δεδομένων.

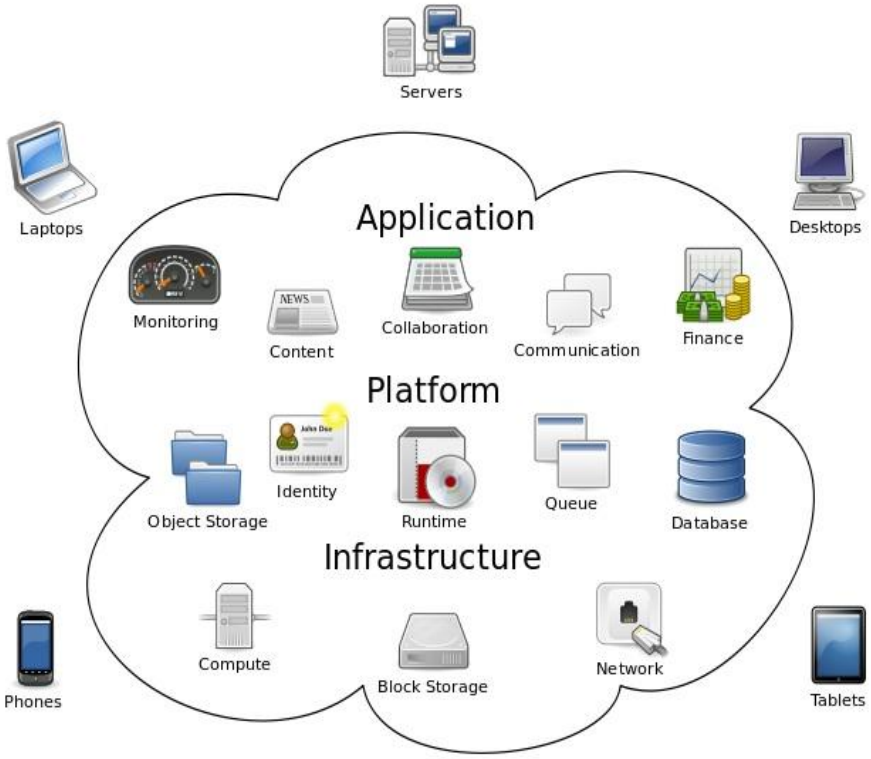
4.1 Cloud Computing και ανάγκη για πιο αποτελεσματικά πρωτόκολλα

Το cloud computing αποτελεί την πρακτική κατά την οποίαν γίνεται χρήση υπολογιστικών πόρων (υλικού και λογισμικού), που χωροταξικά βρίσκονται σε απομακρυσμένα δίκτυα, με απώτερο σκοπό την προσφορά υπηρεσιών. Ένα σύνηθες παράδειγμα διαδικτυακής εφαρμογής, που υπάγεται στην κατηγορία του cloud computing είναι η χρήση πλατφόρμων, που προσφέρουν δωρεάν υπηρεσίες ηλεκτρονικού ταχυδρομείου για την αποθήκευση και μεταφορά προσωπικών δεδομένων. Συνεπώς, κατά το cloud computing, η παροχή υπηρεσιών προς τους χρήστες επιτυγχάνεται με την ανάθεση των υπολογιστικών πόρων, του λογισμικού, αλλά και των προσωπικών δεδομένων των χρηστών στο cloud, δηλαδή το συνονθύλευμα που αποτελεί το δίκτυο ή το ευρύτερο κατανεμημένο σύστημα, που αξιοποιεί τα εν λόγω αγαθά [42].

Παραδοσιακά, το μοντέλο του cloud computing λειτουργεί ως εξής: Οι πάροχοι του cloud (cloud providers) διαχειρίζονται τη γενικότερη υποδομή και τις πλατφόρμες εφαρμογών, που συνιστούν το cloud. Περαιτέρω, ένας χρήστης ή γενικότερα ένας οργανισμός, που επιθυμεί να αποκτήσει πρόσβαση στο λογισμικό εφαρμογών και τις βάσεις δεδομένων, που προσφέρονται από ένα σύστημα cloud, έρχεται σε επικοινωνία με τους cloud providers και ανάλογα με τις ανάγκες των χρηστών ή των οργανισμών οι cloud providers παρέχουν τους αντίστοιχους υπολογιστικούς πόρους. Βέβαια, στην περίπτωση απλοϊκών εφαρμογών, που συνήθως χρησιμοποιούνται από μεμονωμένους χρήστες, όπως για παράδειγμα η υπηρεσία ηλεκτρονικού ταχυδρομείου, που αναφέραμε στην προηγούμενη παράγραφο, οι χρήστες μπορούν να αποκτήσουν πρόσβαση σε cloud-based εφαρμογές μέσω ενός web browser, ενός υπολογιστή ή ενός smart phone, ενώ τα στοιχεία του λογισμικού των χρηστών αποθηκεύονται σε απομακρυσμένους διακομιστές που βρίσκονται στο cloud.

Οι λόγοι, που επέφεραν την ανάπτυξη του cloud computing, είναι κυρίως οικονομικοί και συσχετίζονται άρρηκτα με την βέλτιστη κατανομή των πόρων ενός οργανισμού, που είναι απαραίτητοι για την υποστήριξη της κατάλληλης πληροφοριακής υποδομής. Συγκεκριμένα, το μοντέλο του cloud computing επιφέρει τη μείωση των εξόδων ενός οργανισμού που συσχετίζονται με τη συνεχή υποστήριξη και συντήρηση μίας ολόκληρης πληροφοριακής υποδομής, καθώς και την συνεχή ενημέρωση των λογιστικών εφαρμογών, που είναι αναγκαίες

στον οργανισμό, δεδομένου ότι οι εν λόγω λειτουργίες πραγματοποιούνται από τους παρόχους του cloud.



Εικόνα 4. 1 - Αναπαράσταση λογικής τοπολογίας ενός cloud

Έτσι, δεδομένου ότι οι ανάγκες των οργανισμών σε εφαρμογές λογισμικού μπορεί να διαφέρουν ανά τυχαία χρονικά διαστήματα, μπορούμε να πούμε πως το γεγονός ότι το cloud computing προσφέρει ‘’**κατά απαίτηση λογισμικό**’’ (**on demand software**) και έτσι συνεισφέρει στην ανακατανομή του λειτουργικού κόστους για τους υπολογιστικούς πόρους ενός οργανισμού. Ωστόσο, προσοχή πρέπει να δοθεί στο ζήτημα ότι με τη χρήση των υπηρεσιών ενός cloud τα δεδομένα του οργανισμού ή των χρηστών βρίσκονται στη δικαιοδοσία των cloud providers, κάτι το οποίο σημαίνει πως οι απαιτήσεις σε ασφάλεια είναι ιδιαίτερα υψηλές.

Πέρα τώρα από την ανάπτυξη και ανάδειξη του cloud computing,τα περισσότερα δίκτυα στις μέρες μας, συμπεριλαμβανομένου και του Internet, έχουν δημιουργηθεί με γνώμονα τις προδιαγραφές που ορίζει το πρωτόκολλο IP και λειτουργούν σύμφωνα με τη φιλοσοφία των αυτόνομων συστημάτων AS (δηλαδή οργάνωση ενός δικτύου σε περιφέρειες όπου όλες οι περιφέρειες συνδέονται σε ένα δίκτυο σπονδυλικής στήλης, το οποίο λειτουργεί σαν μεσολαβητής για την μεταφορά πληροφορίας από μία περιφέρεια σε μίαν άλλη). Η μεταφορά δεδομένων εντός ή εκτός των AS επιτυγχάνεται με την προώθηση των πακέτων από κόμβο σε κόμβο, με τον κάθε κόμβο που συμμετέχει στη διαδικασία προώθησης να γνωρίζει μονάχα ποιο θα πρέπει να είναι το επόμενο βήμα (next-hop), δηλαδή ποιος είναι ο επόμενος κόμβος στον οποίο θα πρέπει να μεταφερθεί το πακέτο, έως ότου τελικά το πακέτο αυτό να φτάσει στον τελικό του προορισμό. Κατά τον τρόπο αυτόν τα δίκτυα, τα οποία ενστερνίζονται τη λειτουργία

των αυτόνομων πακέτων μπορούν να επεκταθούν εύκολα, δεδομένου ότι η προσθήκη ενός νέου κόμβου δεν αλλάζει καθοριστικά τα δεδομένα που πρέπει να γνωρίζουν όλοι οι κόμβοι των δικτύων.

Το πέρασμα των χρόνων έχει αποδείξει πως οι τεχνολογίες και τα πρωτόκολλα των δικτύων αυτοδύναμων πακέτων έχουν αποδειχτεί αποτελεσματικά, καθότι η απλή λογική με την οποίαν λειτουργούνε έχουν καταστήσει τα δίκτυα ανθεκτικά και με καλή επεκτατική ικανότητα. Ωστόσο, παρόλα τα πλεονεκτήματά τους, τα αυτόνομα δίκτυα IP διαθέτουν και πολλά μειονεκτήματα, τα οποία άρχισαν να διαφαίνονται με την αυξανόμενη χρήση του Internet. Τα μειονεκτήματα των δικτύων αυτοδύναμων πακέτων συσχετίζονται κυρίως με την αδυναμία τους να παρέχουν στους διαχειριστές των δικτύων εύκολη παραμετροποίηση των διαφορετικών ροών από πακέτα που διέρχονται μέσα στο δίκτυο. Αιτία για το παραπάνω φαινόμενο είναι το γεγονός ότι οι παραδοσιακές τεχνολογίες, που στηρίζουν τη λειτουργία τους στο πρωτόκολλο IP κατέχουν καθολική ισχύ κατά μήκος όλων των οντοτήτων ενός δικτύου, εννοώντας πως οι κανόνες και λειτουργίες που ορίζονται από τα πρωτόκολλα, που συνοδεύον το IP, όπως για παράδειγμα το UTP και TCP, εφαρμόζονται κατά τον ίδιο τρόπο για όλες τις οντότητες του δικτύου. Να σημειωθεί επίσης ότι δεν έχουν όλες οι ροές ενός δικτύου τις ίδιες απαιτήσεις σε QoS, κάτι που σημαίνει ότι για την εκάστοτε ροή απαιτείται διαφορετική διαχείριση από τους ελεγκτικούς μηχανισμούς του δικτύου. Ωστόσο, σύμφωνα με τα παραπάνω, κάτι τέτοιο είναι δύσκολο, αφού παραβιάζεται η αρχή της καθολικότητας που διέπει τις τεχνολογίες IP. Έτσι, οι ειδικοί δικτύων δημιούργησαν μία σειρά συμπληρωματικών πρωτοκόλλων και εξειδικευμένων μηχανισμών αστυνόμευσης, που ενσωματώθηκαν στις τεχνολογίες IP, με σκοπό τα δίκτυα αυτοδύναμων πακέτων, να μπορούν να εφαρμόσουν διαφορετικές πολιτικές δρομολόγησης σε ροές με διαφορετικές απαιτήσεις σε QoS [42].

Η δυσκολία εγκαθίδρυσης ροών με διαφορετικές απαιτήσεις σε καθυστέρηση, bandwidth κτλ στα δίκτυα IP,, απορρέει και από το γεγονός ότι η δρομολόγηση στα δίκτυα αυτόνομων πακέτων υπάγεται στις αρχιτεκτονικές προδιαγραφές, που ορίζει η τοπολογία του δικτύου. Επιπλέον, η εξάρτηση της λειτουργικότητας των δικτύων αυτόνομων πακέτων από την εκάστοτε τοπολογία, που εφαρμόζεται, έχει προσδώσει περαιτέρω αρνητικά χαρακτηριστικά στα AS. Συγκεκριμένα, η ανελαστικότητα των δικτύων IP κατέστησε δύσκολη την προσφορά φερέγγυων υπηρεσιών σε κινητούς χρήστες, καθότι η διεύθυνση ενός κινητού παραλήπτη ή αποστολέα συσχετίζεται άρρηκτα με τοπολογία του δικτύου. Επίσης, ένα ακόμα πρόβλημα, που αξίζει να σημειωθεί, είναι το φαινόμενο κατά το οποίο τα παραδοσιακά πρωτοκόλλα παρουσιάζουν πρόβλημα κατά την εκπομπή πολυδιανομής (multicast), δεδομένου ότι η ομαδοποίηση των διευθύνσεων και γενικά ο καθορισμός ομάδων αποστολής είναι μία πολύπλοκη διαδικασία, ιδικά εάν τα μέλη της ομάδας βρίσκονταν σε διαφορετικά δίκτυα ή ακόμα χειρότερα αν είναι και κινητοί χρήστες που μπορεί να άλλαζαν θέση σε διαφορετικό δίκτυο. Λύση στα εν λόγω προβλήματα πρόσφερε η δημιουργία και εγκαθίδρυση των εικονικών τοπικών δικτύων (VLAN) και εικονικών προσωπικών δικτύων (VPN), που απέδωσαν μία ελαστικότητα στην κατά τα άλλα αδιαλλαξία που χαρακτηρίζει τα παραδοσιακά πρωτόκολλα επιτρέποντας, σε συνδυασμό φυσικά με συμπληρωματικά πρωτόκολλα στους διαχειριστές δικτύων, τον καθορισμό ροών δεδομένων που διαθέτουν διαφορετικά χαρακτηριστικά [42] [43].

Αν και φαινομενικά η δημιουργία των εικονικών κυκλωμάτων και η ενσωμάτωση ειδικών πρωτοκόλλων, που επιτρέπουν τον καθορισμό ροών με συγκεκριμένα χαρακτηριστικά, έχουν επιλύσει τα προβλήματα που αναφέρθηκαν, πολλοί ειδικοί δικτύων θεωρούν ότι οι αυξανόμενες απαιτήσεις των υπηρεσιών των δικτυακών εφαρμογών και η ανάδειξη νέων αρχιτεκτονικών όπως του cloud, απαιτούν την εγκαθίδρυση μίας νέας αρχιτεκτονικής. Επιπλέον, το νέο

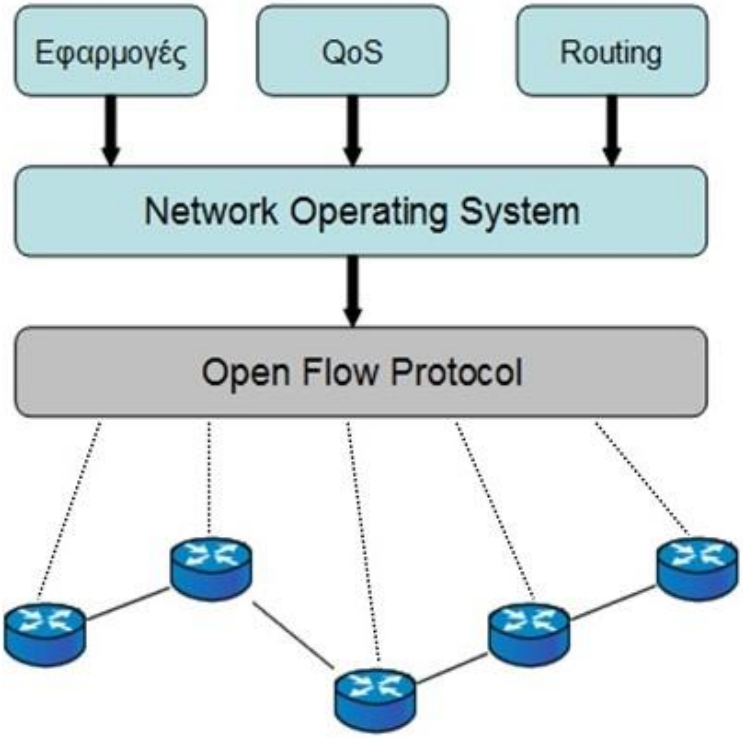
πρόβλημα, που δημιουργήθηκε με την προσθήκη και ενσωμάτωση των ειδικών πρωτοκόλλων στις αξιόπιστες τεχνολογίες IP, είναι η αύξηση της πολυπλοκότητας των σύγχρονων δικτυακών υποδομών και η μετατροπή τους σε δυσπροσάρμοστα συστήματα ως προς επικείμενες μελλοντικές αλλαγές.

Συνεπώς, με την αυξανόμενη χρήση των εικονικών κυκλωμάτων και την ανάδειξη νέων αρχιτεκτονικών δομών, που αποσκοπούνε στην εξυπηρέτηση συγκεκριμένων απαιτήσεων, όπως το cloud, απαιτείται μία νέα αρχιτεκτονική δικτύων [43]. Η φιλοσοφία λειτουργίας της επικείμενης αρχιτεκτονικής θα πρέπει να προσφέρει δυναμική κατανομή των πόρων ενός δικτύου, ενώ παράλληλα να επιτρέπει στους διαχειριστές δικτύων να προσδιορίζουν τις υπηρεσίες του δικτύου χωρίς να συσχετίζονται με τις προδιαγραφές των διεπαφών του δικτύου. Η νέα αυτή αρχιτεκτονική ονομάστηκε **software defined networking** ή **SDN** και υπολογίζεται ότι θα αναιρέσει την πολυπλοκότητα των δικτυακών μηχανισμών.

4.2 SDN αρχιτεκτονική

Η αρχιτεκτονική SDN ακολουθεί μια τελείως διαφορετική προσέγγιση από προηγούμενες αρχιτεκτονικές σχετικά με την οργάνωση και διαχείριση των δικτυακών οντοτήτων και πόρων. Συγκεκριμένα, ορίζει πως η δικτυακή υποδομή πρέπει να διαχωρίζεται στο επίπεδο ελέγχου (control plane) και στο επίπεδο δεδομένων (data plane) [44]. Κατά το SDN, το επίπεδο ελέγχου είναι υπεύθυνο για την λήψη των αποφάσεων, που καθορίζουν την κατεύθυνση και τις παραμέτρους των ροών από πακέτα, ενώ το επίπεδο δεδομένων περιλαμβάνει τις κατάλληλες υποδομές για την φυσική προώθηση των πακέτων στο δίκτυο.

Το επίπεδο ελέγχου υλοποιείται από εικονικές μηχανές, που επιτρέπουν στους διαχειριστές δικτύων να παραμετροποιήσουν τις ροές και γενικότερα να διαχειριστούν τους πόρους των δικτύων, μέσω ενός ειδικού λογισμικού που αποκαλείται **λογισμικό δικτύου** ή **NOS** (**Network Operating System**). Το NOS παρέχει στους διαχειριστές ένα προγραμματιστικό περιβάλλον, από το οποίο μπορούν να ασκήσουν τις απαραίτητες ελεγκτικές διαδικασίες για την ορθή δρομολόγηση των πακέτων. Επιπλέον, το NOS είναι ενσωματωμένο στη δικτυακή υποδομή έτσι ώστε να είναι εφικτή από τους διαχειριστές των δικτύων η διαχείριση όλων των δικτυακών υποδομών και πόρων. Δηλαδή, όπως διαφαίνεται και από την ακόλουθη εικόνα το NOS αποτελεί ένα κατανεμημένο λογισμικό ελέγχου που καλύπτει όλο το εύρος του δικτύου, ασκώντας τις ελεγκτικές του διαδικασίες στα υποεπίπεδα, που αποτελούν κυρίως την υλικοτεχνική υποδομή του δικτύου.



Εικόνα 4. 2 - Αναπαράσταση αρχιτεκτονικής SDN

Η ενσωμάτωση και διασύνδεση του λογισμικού ελέγχου με τις δικτυακές υποδομές επιτυγχάνεται μέσω των άκρων των δικτύων (**network edges**). Το δίκτυο, που απαρτίζεται από το σύνολο των άκρων, συνιστά ένα δίκτυο που αντικαθιστά τις λειτουργίες ενός δικτύου κορμού. Κατά αυτόν τον τρόπο, η αρχιτεκτονική SDN διασφαλίζει την υψηλή διαθεσιμότητα των υπηρεσιών του δικτύου και επιτρέπει την επεκτασιμότητα του δικτύου, όπως ένα αντίστοιχο παραδοσιακό δίκτυο κορμού.

Η αρχιτεκτονική SDN ορίζει πως κατά την επικοινωνία οποιονδήποτε οντοτήτων του δικτύου, τα άκρα λειτουργούν σαν διαμεσολαβητές συγκεντρώνοντας πακέτα, τα οποία προωθούνται σύμφωνα με την πλέον βέλτιστη διαδρομή, που έχει καθοριστεί μέσω των ελεγκτικών μηχανισμών, που εφαρμόζονται από το λογισμικό δικτύου. Προφανώς, οι εν λόγω μηχανισμοί συσχετίζονται άρρηκτα με τις προγραμματιστικές εντολές που έχουν καταχωρήσει στο σύστημα οι διαχειριστές του δικτύου. Όταν, λοιπόν, ένα πακέτο ξεκινήσει από έναν κόμβο-αφετηρία, αρχικά προωθείται προς ένα άκρο του δικτύου SDN, ασχέτως του τελικού προορισμού που έχει επιλεγεί από τον χρήστη ή γενικότερα από την ανάλογη αρμόζουσα οντότητα. Εκεί, το άκρο βασισμένο σε ένα σύνολο από μέτρα και παράγοντες, που έχουν θεσπιστεί από τους διαχειριστές του δικτύου, αποθηκεύει στο μπροστινό μέρος του πακέτου μία επικεφαλίδα με ένα σύνολο από πληροφορίες και στην συνέχεια το προωθεί και πάλι στο δίκτυο [44][45].

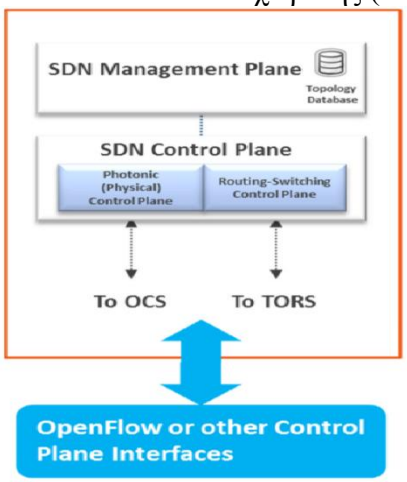
Στη συνέχεια, το πακέτο ενδέχεται να περάσει μέσα από ένα σύνολο από κόμβους που αξιοποιούν τις παραδοσιακές τεχνολογίες IP, μέχρι να φτάσει στο επόμενο άκρο SDN. Προφανώς, κατά την δρομολόγηση και προώθηση του πακέτου από τους εν λόγω κόμβους, γίνεται χρήση των παραδοσιακών τρόπων δρομολόγησης που υπαγορεύουν οι τεχνολογίες και συνηθέστερα τα πρωτόκολλα IP. Επίσης, πρέπει να διευκρινιστεί ότι η επικεφαλίδα, που προστίθεται από το αρχικό άκρο SDN σε οποιοδήποτε πακέτο που καταφτάνει για πρώτη φορά,

επεξεργάζεται μόνο από τα άκρα SDN και όχι από τους κόμβους IP, οι οποίοι αξιοποιούν τα υπόλοιπα πλαίσια του πακέτου για την εφαρμογή των λειτουργιών τους.

Τα πλαίσια, που προστίθενται από τα άκρα SDN, περιέχουν πληροφορίες που αφορούν συγκεκριμένα την εκάστοτε ροή στην οποία ανήκει το κάθε πακέτο. Οι πληροφορίες αυτές ενημερώνουν τα άκρα του δικτύου SDN για τις απαιτήσεις των ροών σε δικτυακούς πόρους, όπως bandwidth, καθυστέρηση ή ακόμα και μέγιστο bit error rate. Όπως αναφέρθηκε και προηγουμένως, οι πληροφορίες που επεξεργάζονται από τους κόμβους SDN είναι απόρροια των εντολών που προέρχονται από τους διαχειριστές του δικτύου. Ωστόσο, πρέπει να αναφερθεί το γεγονός ότι το κριτήριο, με το οποίο καθορίζουν τα άκρα του SDN σε ποιες ροές πρέπει να εφαρμόσουν τις ανάλογες υπηρεσίες, αποτελεί η αφετηρία και ο προορισμός των πακέτων [45].

Αναφέραμε προηγουμένως, πως τα άκρα αποτελούν τη διασύνδεση του επιπέδου ελέγχου και του επιπέδου δεδομένων. Το επίπεδο δεδομένων αποτελεί όλη την υλικοτεχνική υποδομή του δικτύου, που απλά εκτελεί ένα σύνολο ενεργειών για την προώθηση των πακέτων. Ένα αρκετά μεγάλο μέρος του hardware ενός δικτύου SDN απαρτίζεται από παραδοσιακούς δρομολογητές IP, οι οποίοι και βρίσκονται ενδιάμεσα στα άκρα SDN. Οι οντότητες των άκρων που είναι υπεύθυνες για την εφαρμογή των εκλεκτικών εντολών και για την γενικότερη καθοδήγηση του υλικού, λειτουργούν ιδανικά σύμφωνα με τις προδιαγραφές του πρωτόκολλου Ανοιχτής Ροής ή Open Flow [46]. Στην επόμενη ενότητα, θα εξεταστούν ενδελεχώς οι οντότητες που απαρτίζουν το πρωτόκολλο Open Flow και θα αναφερθούν οι εκάστοτε λειτουργίες του.

Στην επόμενη εικόνα απεικονίζεται το πώς μπορεί στα μεγάλα κέντρα δεδομένων που περιέχουν οπτικά και ηλεκτρικά κυκλώματα να εφαρμοστεί αποτελεσματικά η αρχιτεκτονική SDN [17]. Τονίζεται ότι όπως αναφέρθηκε και στις προηγούμενες παραγράφους είναι απαραίτητο ένα επίπεδο έλεγχου και ένα επίπεδο διαχείρισης (management plane).



Εικόνα 4. 3 - Εφαρμογή SDN σε οπτικές αρχιτεκτονικές

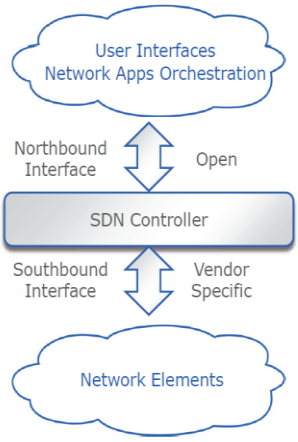
Το management plane δημιουργεί και διαχειρίζεται τοπολογίες και συναφείς ρυθμίσεις, και αναλύει τις διάφορες ροές στο δίκτυο κατά την διάρκεια εκτέλεσης σε συντονισμό με τα επίπεδα ελέγχου δρομολόγησης και μεταγωγής. Με βάση τις λειτουργικές ανάγκες του δικτύου (όπως πχ ενεργοποίηση χρόνου λειτουργίας, προγραμματισμένων και προσχεδιασμένων μοτίβων, μέχρι και δραστηριότητες συντήρησης δικτύου) δημιουργεί νέες τοπολογίες και συναφείς ρυθμίσεις του δικτύου και διαδίδει τις ρυθμίσεις αυτές στα αντίστοιχα επίπεδα ελέγχου για ασύγχρονη

εκτέλεση αυτών. Με τον τρόπο αυτό μπορούν να πραγματοποιούνται αποτελεσματικά οι επαναδιαρθρώσεις όλων των δικτύων που αναλύθηκαν διεξοδικά στην προηγούμενη ενότητα 3.

Τα επίπεδα ελέγχου οργανώνουν τις αλλαγές που αφορούν την ρύθμιση της τοπολογίας του δικτύου:οι φωτονικής φύσεως μηχανές διαχειρίζονται τις αλλαγές της τοπολογίας στο οπτικό δικτυοδόμημα ενώ οι μηχανές δρομολόγησης και μεταγωγής διαχειρίζονται τις αλλαγές τοπολογίας στα διάφορα στοιχεία δρομολόγησης και μεταγωγής πακέτων.

Και οι δύο ειδών μηχανές, διατηρούν την κατάσταση της εκτέλεσης του κάθε σταδίου στην ροή διαμόρφωσης και επικοινωνούν με το πεδίο διαχείρισης για τους χρόνους εκτέλεσης εντολών και για την επεξεργασία εντολών με τις οποίες επαναδιαρθρώνεται το δίκτυο.

Στο σημείο αυτό θα πρέπει να τονιστεί η διαφοροποίηση μεταξύ δύο διεπαφών (**interface**) της νότιας διεπαφής-**southbound interface** και της βόρειας-**northbound interface** του SDN controller [47]. Η southbound interface είναι η σύνδεση μεταξύ του ελεγκτή SDN και των στοιχείων του δικτύου. Η διεπαφή μεταξύ του ελεγκτή SDN και του hypervisor είναι ένα χαρακτηριστικό παράδειγμα μιας southbound interface,όπως και το πρωτόκολλο **OpenFlow** το οποίο αναλύεται στη συνέχεια. Οι southbound interfaces δεν είναι απαραίτητα όλες προσβάσιμες και ελεύθερες και πράγματι μπορεί να είναι και ιδιόκτητες (proprietary). Η northbound interface είναι η σύνδεση του ελεγκτή SDN σε εφαρμογές και υπηρεσίες δικτύου. Αυτό μερικές φορές ονομάζεται το επίπεδο εφαρμογής (application layer), καθώς εδώ είναι που οι εφαρμογές τρίτων προσώπων μπορούν να γραφτούν και οι διεπαφές των χρήστών είναι διαθέσιμες. Εδώ είναι επίσης που η κανονική ενορχήστρωση του δικτύου μεταξύ δύο οποιωδήποτε τερματικών μπορεί να πάρει μέρος καθώς οι αιτήσεις σε αυτό το επίπεδο έχουνε μια πιο ολοκληρωμένη και ολιστική άποψη του συνόλου του δικτύου. Στην εικόνα που ακολουθεί απεικονίζεται διευκρινιστικά η διαφορά μεταξύ των southbound και northbound interfaces [47].



Εικόνα 4. 4 – Northbound και Southbound διεπαφές

Τέλος θα ακολουθήσουν επιγραμματικά μερικά πλεονεκτήματα που προσφέρει η οικειοποίηση του SDN από τα κέντρα δεδομένων [48]:

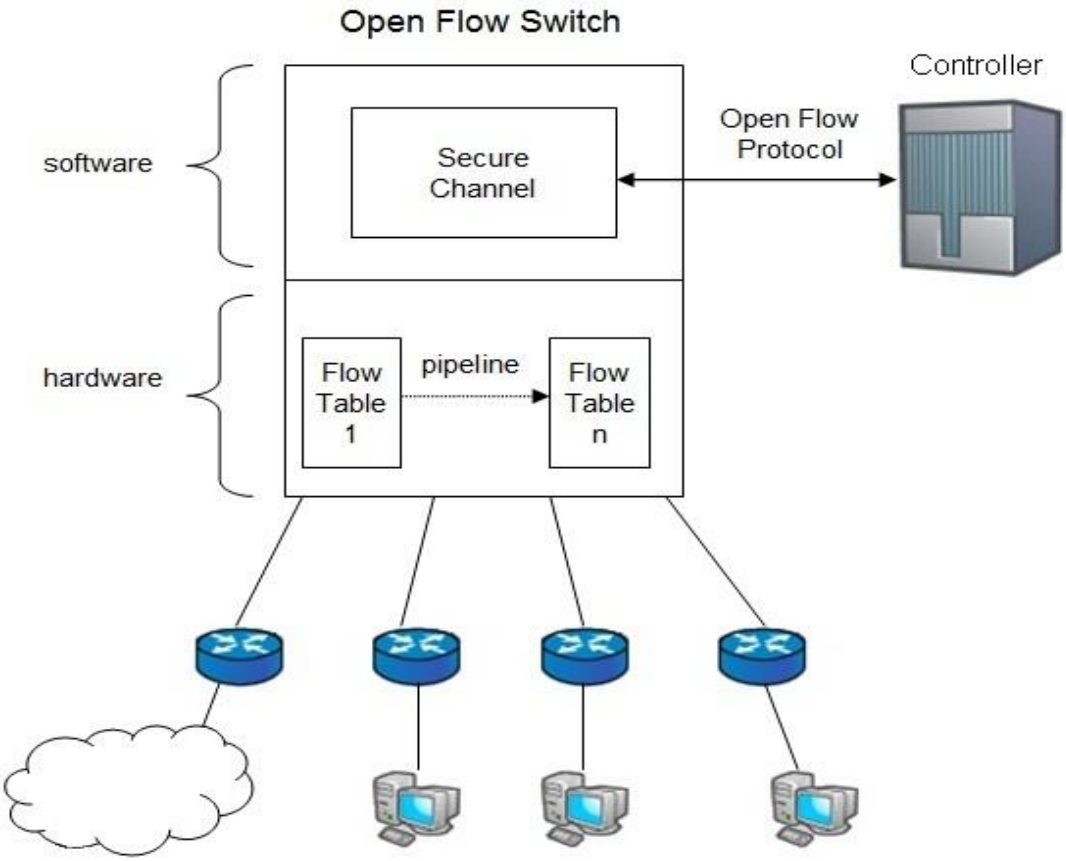
- κατά-απαίτηση παροχή
- αυτοματοποιημένη εξισορρόπηση φορτίου
- άμεσα προγραμματιζόμενο, καθώς το λογισμικό είναι ανεξάρτητο από τις εντολές προώθησης
- ευέλικτο, καθώς η ανεξαρτητοποίηση του ελέγχου από την προώθηση των πακέτων αφήνει τους διαχειριστές να επέμβουν στο δίκτυο και να προσαρμόσουν τη ροή της κίνησης
- ικανότητα επέκτασης των πόρων του δικτύου

4.3 Open Flow πρωτόκολλο

Το πρωτόκολλο **Ανοιχτής Ροής** ή **OF** (**Open Flow**) αποτελεί ένα σύνολο από προγραμματιστικούς κανόνες, που καθορίζουν την λειτουργία των άκρων ενός δικτύου SDN και είναι το πιο διαδεδομένο από τα πρωτόκολλα που συνοδεύουν μία SDN αρχιτεκτονική [46][44]. Συγκεκριμένα, το πρωτόκολλο Open Flow παρέχει υπηρεσίες γεφύρωσης και προώθησης στα πακέτα IP του επιπέδου δεδομένων, προσδιορίζοντας τις ροές στις οποίες ανήκουν με διάφορες προγραμματιστικές εντολές. Στην ενότητα αυτή θα ακολουθήσει διεξοδική παρουσίαση του Open Flow ενώ στην επόμενη θα γίνει αναφορά στα υπόλοιπα πρωτόκολλα που μπορεί να εξυπηρετούν μία SDN αρχιτεκτονική.

4.3.1 Το μοντέλο πρωτοκόλλου Open Flow

Οι συσκευές, που είναι απαραίτητες για τη λειτουργία του πρωτοκόλλου Open Flow βρίσκονται σε συσκευές που αποκαλούνται Open Flow Switches. Όπως διαφαίνεται στην επόμενη εικόνα τα Open Flow Switches συνδέονται με έναν Controller μέσω ενός ασφαλούς καναλιού [49]. Οι Controllers αποτελούν τερματικά, που μέσω των διαδικασιών του λογισμικού δικτύου, προσφέρουν στους διαχειριστές δικτύων τη δυνατότητα να διαχειριστούν τα Open Flow Switches. Σε αυτό το σημείο, πρέπει να διασαφηνιστεί ότι τα edges για τα οποία έγινε λόγος στην ενότητα περιγραφής της αρχιτεκτονικής SDN, στην πραγματικότητα απαρτίζονται από Controllers και Open Flow Switches. Για την επικοινωνία μεταξύ ενός Controller και ενός Open Flow Switch χρησιμοποιείται το πρωτόκολλο Open Flow, το οποίο είναι ένα σύνολο από προγραμματιστικές εντολές, που όπως θα παρουσιαστεί στην συνέχεια αποσκοπεί στην προσθήκη, ενημέρωση ή διαγραφή δεδομένων, που αποθηκεύονται στους πίνακες ροής (**flow tables**) του Open Flow Switch. Επιπλέον, όπως φαίνεται στην παρακάτω εικόνα, τα Open Flow Switches συνιστώνται από ένα σύνολο από οντότητες,που αποκαλούνται Flow Tables και αποτελούν το κοντινότερο τμήμα λογισμικού στην υλικοτεχνική (hardware) υποδομή του δικτύου IP (δεδομένου ότι οι λειτουργίες του συσχετίζονται άμεσα με την προώθηση και δρομολόγηση των πακέτων). Στη γενική τους μορφή τα Flow Tables αποτελούν πίνακες που διαθέτουν καταχωρήσεις και συσχετίζονται: με πληροφορίες αντιστοίχησης, πληροφορίες επεξεργασίας και πληροφορίες εντολών. Ωστόσο, τα πλαίσια, οι καταχωρήσεις και συνεπώς η ευρύτερη δομή των Open Flow Tables ενδέχεται να έχουν αρκετές παραλλαγές, αναλόγως με τις υπηρεσίες τις οποίες έχει κληθεί να εξυπηρετήσει ο εκάστοτε μεταγωγέας [44]. Γενικότερα, επικρατεί μία ρευστότητα στο συγκεκριμένο τμήμα του πρωτοκόλλου, διότι η αρχιτεκτονική SDN και το πρωτόκολλο Open Flow βρίσκονται ακόμα σε ανάπτυξη και εξέλιξη.



Εικόνα 4. 5 - Αναπαράσταση του πρωτοκόλλου Open Flow

4.3.2 Στοιχεία του πρωτοκόλλου Open Flow

Στην παραπάνω ενότητα, αναφέραμε πως ο Controller μίας άκρης (**edge**) στέλνει δεδομένα στο Open Flow Switch του ιδίου edge. Αυτά τα δεδομένα αποτελούν πληροφορίες, που αξιοποιούνται για την επεξεργασία των πακέτων, που ενδέχεται να προωθηθούν από το edge και αποθηκεύονται στα Flow Tables [49]. Όπως φαίνεται και από την εικόνα που ακολουθεί, στην γενική τους δομή, τα Flow Tables στην πραγματικότητα συνιστούν πίνακες, που αποτελούνται από τρείς στήλες:

Πεδία αντιστοίχησης	Μετρητές	Οδηγίες Προώθησης

Εικόνα 4. 6 - Αναπαράσταση ενός πίνακα ροής (flow table)

Στην πρώτη στήλη ενός Open Flow Table καταχωρούνται πληροφορίες αντιστοίχησης (**matching fields**). Τα συγκεκριμένα δεδομένα χρησιμοποιούνται από τα Open Flow Switches για την αναγνώριση της προέλευσης και του προορισμού ενός πακέτου προς προώθηση. Στη συνέχεια, ακολουθεί η στήλη καταχώρησης μετρητών (counters), οι οποίοι χρησιμοποιούνται για να διαπιστωθεί αν ένα πακέτο, ενός αποστολέα που είναι καταχωρημένος στο ανάλογο πεδίο αντιστοίχησης, είναι απαρχαιωμένο (outdated) ή όχι. Επιπλέον, όπως και με τα παραδοσιακά πρωτόκολλα TCP/IP, οι καταχωρήσεις του πεδίου μετρητών συνήθως χρησιμοποιούνται για την εξακρίβωση διπλών πακέτων και πακέτων εκτός σειράς. Στην τρίτη και τελευταία στήλη ενός Open Flow Table αποθηκεύεται ένα σύνολο από εντολές, που στην πραγματικότητα αποτελούν οδηγίες προώθησης (forwarding instructions). Η εν λόγω στήλη είναι το πιο σημαντικό τμήμα σε όλο τον πίνακα, καθότι σε αυτό το πεδίο του πίνακα αποθηκεύονται οι οδηγίες του Controller, που συσχετίζονται με τη διαχείριση των πακέτων και της ευρύτερης ροής που συνιστούν τα πακέτα ενός προορισμού. Αυτές οι οδηγίες αποτελούν προγραμματιστικές εντολές που καθορίζονται από το πρωτόκολλο Open Flow και καθορίζουν τις απαιτήσεις της εκάστοτε ροής από πακέτα. Συνεπώς, οι εντολές που εισάγονται από τους διαχειριστές του δικτύου στον Controller, μέσω της βοήθειας του λειτουργικού συστήματος δικτύου, μεταφέρονται μέσω του ασφαλούς καναλιού στο Open Flow Switch και αποθηκεύονται σε αυτό το πεδίο.

Όταν ένα πακέτο καταφθάσει για πρώτη φορά σε ένα Open Flow Switch, τότε προσκολλάται στο μπροστινό του σημείο μία νέα κεφαλίδα (**header**). Η εικόνα που ακολουθεί, παρουσιάζει την ευρύτερη δομή ενός SDN header:

Εικόνα 4. 7 - Αφαιρετική αναπαράσταση ενός header SDN

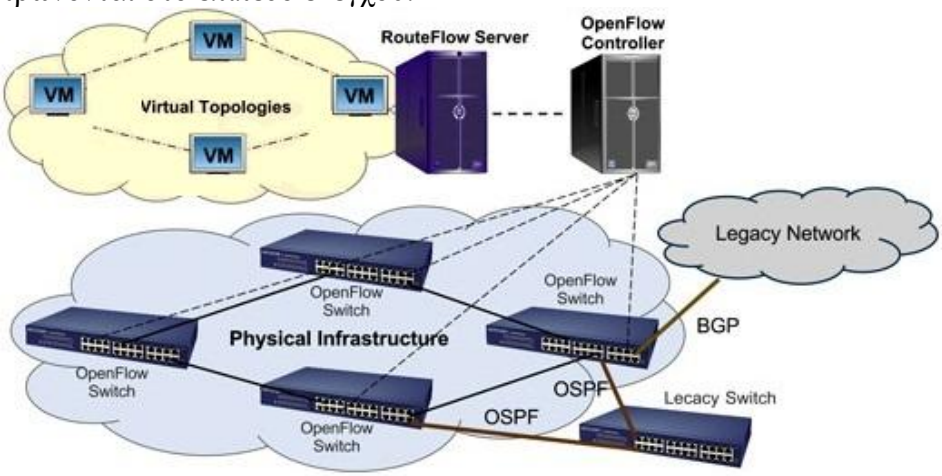
Όπως θα εξεταστεί και στην συνέχεια, η δομή της εν λόγω κεφαλίδας συσχετίζεται άμεσα με την δομή των Open Flow Tables και συνεπώς η κεφαλίδα ενδέχεται αρκετές παραλλαγές, ανάλογα με την υλικοτεχνική υποδομή του δικτύου και τις υπηρεσίες που παρέχει. Ωστόσο, στη γενική του δομή ένας header SDN διαθέτει ένα πεδίο για κάθε πρωτόκολλο ή ειδική τεχνολογία που αξιοποιείται από την υποδομή του δικτύου, τα οποία καταγράφουν πληροφορίες όπως τη διεύθυνση προορισμού και τη διεύθυνση αφετηρίας. Ενδεικτικά, πεδία μπορεί να υπάρχουν για τα πρωτόκολλα IP και TCP, καθώς και για τις τεχνολογίες VLAN και Ethernet.

4.3.3 Λειτουργία του πρωτοκόλλου Open Flow

Έχοντας πλέον αναλύσει τις γενικότερες προδιαγραφές της αρχιτεκτονικής SDN και του μοντέλου Open Flow, καθώς και ποία είναι τα απαραίτητα πράγματα για την υλοποίηση τους, σε αυτήν την ενότητα θα αναλυθεί περαιτέρω η λειτουργία ενός δικτύου SDN/Open Flow. Όπως φαίνεται από την εικόνα που ακολουθεί, η ευρύτερη δομή ενός δικτύου SDN/Open Flow αποτελείται: από ένα σύνολο εικονικών μηχανημάτων ή VM που συνδέονται με ένα εικονικό τοπικό δίκτυο ή VLAN, ένα Open Flow / Route Flow Controller, ένα σύνολο από Open Flow

Switches και τέλος ένα δίκτυο που αξιοποιεί παραδοσιακές τεχνολογίες για τη λειτουργία του, όπως για παράδειγμα τεχνολογίες IP.

Το δίκτυο, που απαρτίζεται από τα εικονικά μηχανήματα σε συνδυασμό με τον Route Flow Server, αποτελεί το επίπεδο ελέγχου του ευρύτερου δικτύου SDN [50]. Σκοπός των εικονικών μηχανημάτων είναι η υποστήριξη του λειτουργικού συστήματος δικτύου (network operating system), το οποίο προσφέρει ένα προγραμματιστικό περιβάλλον στους διαχειριστές για τη διαχείριση των πόρων του. Επίσης, στο επίπεδο ελέγχου συμπεριλαμβάνεται ο Route Flow Server, που σκοπός του είναι η συλλογή και αποθήκευση πληροφοριών που συσχετίζονται με τις απαιτήσεις της κάθε ροής που βρίσκεται μέσα στο δίκτυο. Αυτό σημαίνει ότι στα δίκτυα SDN/Open Flow όλες τις απαραίτητες πληροφορίες για την καθιέρωση όλων των διαφορετικών ροών συγκεντρώνονται στο επίπεδο ελέγχου.



Εικόνα 4. 8 - Αναπαράσταση ευρύτερης δομής δικτύου SDN/Open Flow

Οι πληροφορίες των ροών διαμοιράζονται από τον Route Flow Server προς τα τερματικά των διαχειριστών με χρήση αλγορίθμων πλημμύρας (**flooding algorithms**). Κατά αυτόν τον τρόπο, οι διαχειριστές του δικτύου μπορούν μέσω των τερματικών εικονικών μηχανημάτων να έχουν γνώση όλων των διαφορετικών ροών, που διέρχονται από το δίκτυο, με αποτέλεσμα να μπορούν να θεσπίσουν ξεχωριστές πολιτικές δρομολόγησης για τα πακέτα της εκάστοτε ροής του δικτύου. Όταν οι διαχειριστές αποφασίσουν ποία είναι η βέλτιστη πολιτική διαχείρισης της δικτυακής υποδομής, μπορούν να χρησιμοποιήσουν το προγραμματιστικό περιβάλλον που παρέχεται από το λειτουργικό σύστημα δικτύου των τερματικών τους, με σκοπό να προγραμματίσουν τα Open Flow Switch για το πώς θα πρέπει να προωθήσουν τα πακέτα που εισέρχονται από το δίκτυο IP. Ο προγραμματισμός των Open Flow Switches επιτυγχάνεται με την χρήση των εντολών του πρωτοκόλλου Open Flow, για τις οποίες και έγινε λόγος στην ενότητα 4.3.1. Οι εντολές Open Flow αποτελούν ένα σύνολο από συγκεκριμένες ενέργειες, που πρέπει να εκτελέσουν τα Open Flow Switches στα πακέτα, που ανήκουν σε κάποια συγκεκριμένη ροή. Έτσι, λοιπόν, για να αναγνωριστεί σε ποια ροή ανήκει ένα πακέτο εξετάζεται κυρίως η διεύθυνση αφετηρίας του και η διεύθυνση προορισμού του και η δομή των εντολών Open Flow. Έτσι λοιπόν, δεδομένου ότι οι διαχειριστές γνωρίζουν από τον Open Flow Server τις διευθύνσεις πηγής και προορισμού των πακέτων μίας ροής, στέλνουν τις κατάλληλες εντολές Open Flow στα edges του δικτύου, δηλαδή στους Controllers και τα Open Flow Switches, κάνοντας και πάλι χρήση αλγορίθμων πλημμύρας.

Όταν οι εντολές του επιπέδου ελέγχου έχουν μεταφερθεί στα Open Flow Switches, θεωρείται πως πλέον το επίπεδο δεδομένων μπορεί να τεθεί σε λειτουργία. Όπως φαίνεται από την

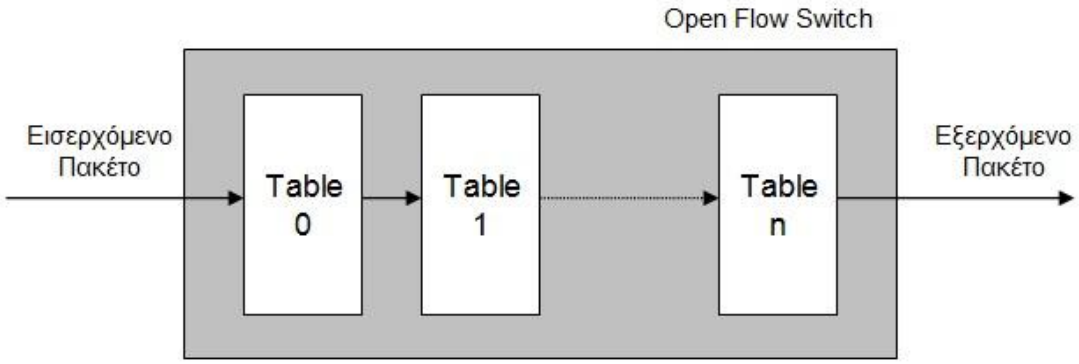
προηγούμενη εικόνα απεικόνισης μίας ευρύτερης δομής ενός δικτύου SDN/Open Flow, το επίπεδο δεδομένων αποτελείται από τα Open Flow Switch και γενικότερα την υλικοτεχνική υποδομή και τα πρωτόκολλα IP που αξιοποιούνται για την προώθηση των πακέτων. Οι εντολές Open Flow του επίπεδου ελέγχου αποθηκεύονται στα Flow Tables των Open Flow Switches, όπως αναφέρθηκε και στην προηγούμενη ενότητα. Παρατηρώντας την δομή των Flow Table στην αντίστοιχη εικόνα 4.6, κατά αντιστοιχία με τη δομή των εντολών Open Flow είναι εύκολα κατανοητή η διαδικασία με την οποία αποθηκεύονται οι εντολές στον πίνακα Open Flow. Συγκεκριμένα, σε κάθε γραμμή του πίνακα Open Flow καταχωρούνται οι πληροφορίες για την αναγνώριση μίας ροής, καθώς και οι ενέργειες που αντιστοιχούν στην εκάστοτε ροή στο πεδίο οδηγιών. Κατά αυτόν τον τρόπο, τα Open Flow Switches μπορούν να καταλάβουν σε ποια ροή ανήκουν τα εισερχόμενα πακέτα, καθώς και ποιες είναι ενέργειες που αντιστοιχούν στην εκάστοτε ροή [46].

Καθώς τα πακέτα ταξιδεύουν στην υποδομή του δικτύου διέρχονται από δρομολογητές, κόμβους IP και Open Flow Switches. Όταν το πρώτο πακέτο μίας νέας ροής καταφτάσει για πρώτη φορά σε ένα edge του δικτύου, δηλαδή έναν Open Flow Switch που συνδέεται με ένα ασφαλές κανάλι με έναν Controller, το Open Flow Switch διαπιστώνει ότι πρόκειται για μία νέα ροή, διότι πρώτον το πακέτο δεν διαθέτει έναν header SDN και δεύτερον δεν διαθέτει καμία εγγραφή στο Flow Table, που να συσχετίζεται με αυτό. Έτσι, λοιπόν, σε αυτήν την περίπτωση το Open Flow Switch προωθεί το πακέτο προς τον Controller, ο οποίος στη συνέχεια γνωστοποιεί στους διαχειριστές την ύπαρξη ενός νέου πακέτου, που ενδεχομένως να σηματοδοτεί την αφετηρία μίας νέας ροής. Όταν θεσπιστεί από τους διαχειριστές μια κατάλληλη πολιτική δρομολόγησης για τα πακέτα της συγκεκριμένης ροής, οι διαχειριστές επεμβαίνουν δυναμικά και προγραμματίζουν τα Open Flow Switches, καταχωρώντας στα Open Flow Tables των Open Flow Switches τις κατάλληλες πληροφορίες για τη δρομολόγηση των πακέτων της επικείμενης ροής. Κατά αυτόν τον τρόπο, τα επόμενα πακέτα της ροής που θα διέλθουν από τα Switches, θα μπορέσουν να εξυπηρετηθούν, όπως έχουν ορίσει οι διαχειριστές του δικτύου.

Όταν ένα πακέτο μίας ροής καταφτάνει σε ένα Open Flow Switch, για το οποίο τα Switches είναι ενημερωμένα από τους διαχειριστές, τότε εκτελούνται οι εξής ενέργειες: Αν το πακέτο διέρχεται για πρώτη φορά από ένα Open Flow Switch, άλλα το Switch είναι ενήμερο για τη ροή στην οποία ανήκει, το Open Flow Switch προσκολλά στο πακέτο τον header SDN, ο οποίος περιγράφεται στην εικόνα 4.7 και στην συνέχεια προωθείται προς την κατάλληλη διεπαφή, σύμφωνα πάντα με τις πληροφορίες που είναι καταχωρημένες στο Open Flow Table [49]. Ωστόσο, η παραπάνω περίπτωση θα συμβαίνει μόνο στο πρώτο Switch από το οποίο θα διέρθει το πακέτο, άρα στην πλειοψηφία των περιπτώσεων όταν ένα πακέτο καταφτάσει σε ένα Switch, θα διαθέτει ένα header SDN. Προφανώς, οι εν λόγω headers υπάρχουν για την ταχύτερη επεξεργασία των πακέτων, έτσι ώστε οι μετέπειτα Switches να μην χρειαστεί να διαβάσουν όλο το πακέτο ή ένα μεγάλο μέρος για την εύρεση πληροφοριών αναγνώρισης.

Στην εικόνα που ακολουθεί, διαφαίνεται λεπτομερώς η διαδικασία με την οποία τα Open Flow Switches επεξεργάζονται τα πακέτα τα οποία διαθέτουν SDN headers. Συγκεκριμένα, όταν καταφτάσει ένα τέτοιο πακέτο, το Open Switches επεξεργάζεται το πακέτο κάνοντας χρήση ενός συνόλου από πίνακες, μέσω μία διαδικασία που ονομάζεται διοχέτευση (**pipeline**). Κατά αυτήν την διαδικασία, το πακέτο εξετάζεται σειριακά από ένα σύνολο από πίνακες ροής, οι οποίοι εκτελούν μία σειρά από ενέργειες, με την κάθε ενέργεια να εκτελείται η μία μετά την άλλη. Προφανώς, αν ένας πίνακας στην αλυσίδα που σχηματίζεται δεν διαθέτει κάποια εγγραφή για μία ροή, τότε τα πακέτα αυτής της ροής απλά τον αγνοούν. Μέχρι στιγμής, όταν γινόταν αναφορά στον Open Flow Table, στην πραγματικότητα γινόταν λόγος για τον πίνακα 0 της

εικόνας 4.9 καθότι αυτός είναι υπεύθυνος για τις βασικές διεργασίες επεξεργασίας των πακέτων. Ωστόσο, ανάλογα με τις ανάγκες και τα πρωτόκολλα που αξιοποιούνται από την εκάστοτε ροή, ένα πακέτο της μπορεί να χρειαστεί να επεξεργαστεί από έναν πίνακα ή και περισσότερους.



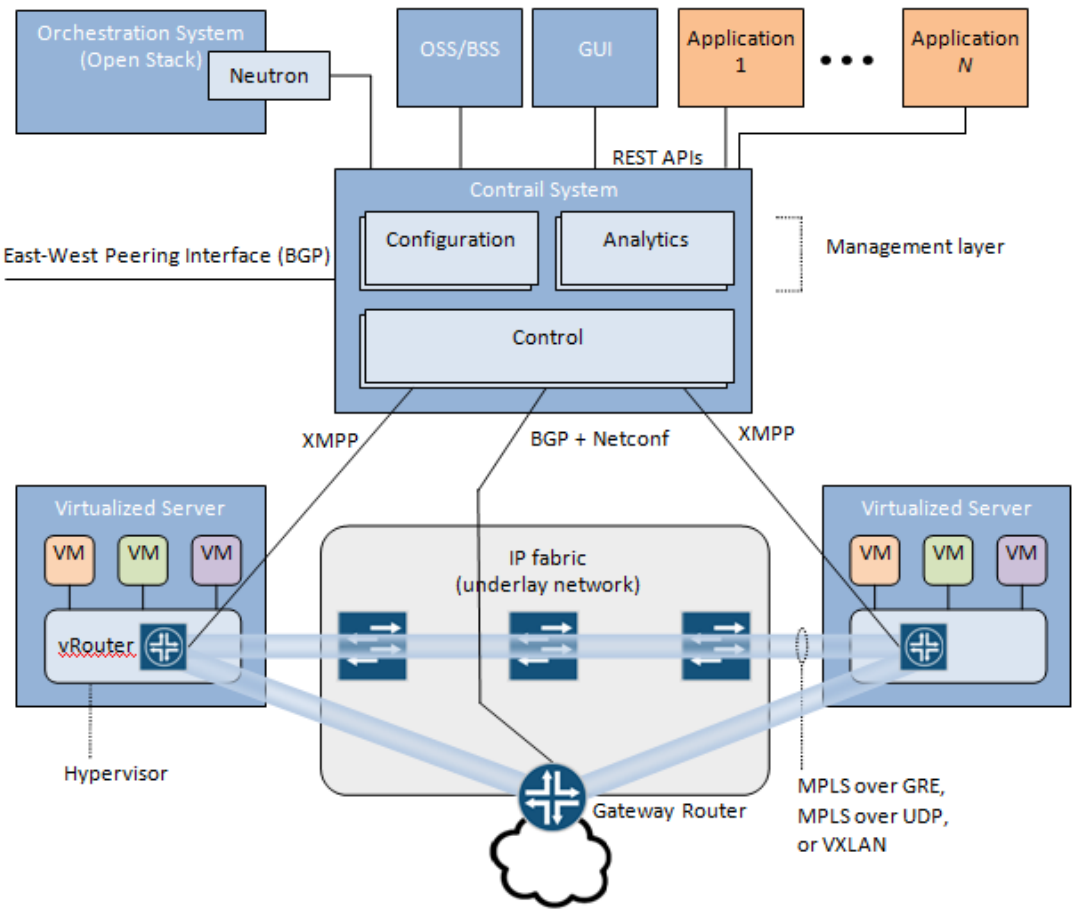
Εικόνα 4. 9 - Αναπαράσταση διαδικασίας διοχέτευσης πακέτων σε ένα Open Flow Switch

4.4 OpenContrail και OpenDaylight πρωτόκολλα

Στην ενότητα αυτή θα γίνει μία αναφορά σε κάποια πρωτόκολλα τα οποία προορίζονται για software defined networks. Μέχρι στιγμής το πιο διαδεδομένο πρωτόκολλο που χρησιμοποιείται από τους SDN controllers είναι το Open Flow το οποίο αναλύθηκε εκτενώς στην προηγούμενη ενότητα 4.3, αλλά ο SDN controller μπορεί να βασιστεί και σε διαφορετικά πρωτόκολλα. Να σημειωθεί ότι μερικά από τα πρωτόκολλα αυτά βρίσκονται ακόμα σε πειραματικό στάδιο.

OpenContrail

Το OpenContrail σύστημα χαρακτηρίζεται από τους δημιουργούς του ως μία επεκτάσιμη πλατφόρμα για SDN υποδομές [51]. Αποτελείται από 2 κύριες συνιστώσες: τον OpenContrail Controller και το OpenContrail vRouter. Ο OpenContrail Controller είναι ένας λογικά κεντρικοποιημένος αλλά φυσικά κατανεμημένος SDN controller που είναι υπεύθυνος για την παροχή της διαχείρισης, του ελέγχου και των αναλυτικών λειτουργιών ενός δικτύου. Το OpenContrail vRouter είναι ένα πεδίο προώθησης το οποίο εκτελείται στον hypervisor ενός διακομιστή. Επεκτείνει το δίκτυο από τους φυσικούς δρομολογητές και τους μεταγωγείς σε ένα κέντρο δεδομένων, σε ένα εικονικό δίκτυο επικάλυψης (virtual overlay network). Το OpenContrail vRouter είναι λειτουργικά παρόμοιο με τα υπάρχοντα εμπορικά vSwitches, όπως για παράδειγμα το Open vSwitch αλλά παρέχει επίσης εντολές δρομολόγησης και υπηρεσίες υψηλότερου επιπέδου. Ο OpenContrail ελεγκτής παρέχει το λογικά κεντρικοποιημένο πεδίο ελέγχου και το πεδίο διαχείρισης του συστήματος και καθορίζει την λειτουργία των vRouters. Στην επόμενη εικόνα που ακολουθεί απεικονίζεται η αρχιτεκτονική του OpenContrail:



Εικόνα 4. 10 - Αναπαράσταση αρχιτεκτονικής OpenContrail

Το OpenContrail σύστημα αποτελείται από δύο κύρια μέρη τα οποία επαναλαμβάνουμε διευκρινιστικά: τον λογικά κεντρικοποιημένο αλλά φυσικά κατανεμημένο controller και ένα σύνολο από vRouters που λειτουργούν ως στοιχεία προώθησης [51].

Ο controller αποτελείται από 3 κύριες οντότητες:

- Κόμβους διαμόρφωσης, που είναι αρμόδιοι για τη μετάφραση του υψηλού επιπέδου μοντέλου δεδομένων σε μία μορφή χαμηλότερου επιπέδου κατάλληλη για την αλληλεπίδραση με τα στοιχεία του δικτύου.
- Κόμβους ελέγχου, οι οποίοι είναι υπεύθυνοι για τη διάδοση αυτής της μορφής χαμηλού επιπέδου από και προς τα στοιχεία του δικτύου.
- Κόμβους ανάλυσης, οι οποίοι είναι υπεύθυνοι για την καταγραφή δεδομένων σε πραγματικό χρόνο από τα στοιχεία του δικτύου, απομακρύνοντας τα από το δίκτυο και τελικά για την παρουσίαση τους σε μια μορφή κατάλληλη για εφαρμογές- applications.

Οι vRouters πρέπει να θεωρηθούν ως στοιχεία δικτύου υλοποιημένα εξ ολοκλήρου σε λογισμικό. Αυτοί είναι υπεύθυνοι για την προώθηση πακέτων από μια εικονική μηχανή σε άλλες εικονικές μηχανές μέσα από μια σειρά μονοπατιών από server σε server. Τα μονοπάτια αυτά αποτελούν ένα δίκτυο επικάλυψης που κάθεται στην κορυφή ενός φυσικού δικτύου IP-over-

Ethernet. Κάθε vRouter αποτελείται από δύο μέρη: ένα χώρο του χρήστη που υλοποιεί το επίπεδο ελέγχου και ένα module του πυρήνα που υλοποιεί τον μηχανισμό προώθησης. Περισσότερες λεπτομέρειες για τη λειτουργία του OpenContrail μπορούν να βρεθούν στην ηλεκτρονική του διεύθυνση [51].

OpenDaylight

Το OpenDaylight είναι ένα έργο ανοικτού πηγαίου κώδικα της LINUX για να προωθήσει την τεχνολογία του SDN [52]. Το OpenDaylight επικεντρώνεται στην οικοδόμηση μιας ανοικτής πλατφόρμας, που βασίζονται σε πρότυπα SDN πλατφόρμα ελεγκτή, που είναι κατάλληλο για την ανάπτυξη σε διάφορα περιβάλλοντα δικτύων παραγωγής. Το OpenDaylight αναμένεται να περιλαμβάνει υποστήριξη για μια σειρά από είτε καθιερωμένα ή ανερχόμενα πρωτόκαλλα SDN, υπηρεσίες δικτύου όπως εικονικοποίηση (virtualization) αλλά και στοιχεία του επιπέδου δεδομένων στις SDN εφαρμογές. Επίσης το OpenDaylight θα περιλαμβάνει υποστήριξη και για το πρωτόκολλο OpenFlow αλλά και πιθανώς για άλλες ανοιχτές SDN εφαρμογές και υλοποιήσεις. Να τονιστεί σε αυτό το σημείο ότι το OpenDaylight βρίσκεται σε πειραματικό στάδιο με τον κώδικά του να αναβαθμίζεται αλλά αναμένεται να είναι ένα πολλά υποσχόμενο πρωτόκολλο για SDN αρχιτεκτονικές [52].

Τέλος να σημειωθεί ότι υπάρχουν πολλές εφαρμογές σε εξέλιξη των οποίων σκοπός είναι να δημιουργήσουν αξιόπιστα πρωτόκολλα για τις διάφορες SDN εφαρμογές και υλοποιήσεις. Άξια αναφοράς είναι τα ForCES [53], PARHAM [54] και το I2RS [55] (Να σημειωθεί διευκρινιστικά για τα παραπάνω ότι ενώ το OpenFlow είναι πρωτόκολλο διεπαφής μεταξύ του SDN controller και του μεταγωγέα, τα OpenDaylight, OpenContrail και PARHAM είναι SDN network controller (σχήμα 4.4).

4.5 Μελλοντικές ερευνητικές προκλήσεις στα κέντρα δεδομένων

Καθώς το μοντέλο που καθορίζει το SDN οικειοποιείται από ολοένα και περισσότερα κέντρα δεδομένων και καθιερώνονται πρωτόκολλα όπως το Open Flow νέες προκλήσεις και λύσεις σε αυτές προκύπτουν καθημερινώς. Στην ενότητα αυτή της εργασίας γίνεται αναφορά σε διάφορες προκλήσεις οι οποίες τίθενται από το SDN, καθώς επίσης και τις μελλοντικές ερευνητικές κατευθύνσεις στις οποίες πρέπει να στραφούνε τα κέντρα δεδομένων.

1) Σχεδιασμός controller και μεταγωγέα

Στην ενότητα αυτή θα αναφερθούμε σε ερευνητικές προσπάθειες οι οποίες αφορούν την επεκτασιμότητα, την απόδοση και την ασφάλεια, μίας SDN αρχιτεκτονικής, στο επίπεδο ελέγχου και στον μεταγωγέα.

Υπάρχουν ερευνητικές μελέτες στις οποίες οι εισερχόμενες ροές δεδομένων προωθούνται απευθείας στους μεταγωγείς με απώτερο σκοπό τη μείωση των αιτήσεων στον controller, έτσι ώστε να περιοριστεί το contention σε controllers που δέχονται μεγάλο αριθμό από applications (DIFANE) [56]. Μία άλλη πρόταση είναι το μοίρασμα των ροών δεδομένων (Devoflow) [56].

Πιο συγκεκριμένα, οι σύντομες ροές δεδομένων θα αναλαμβάνονται από τους μεταγωγείς ενώ οι μεγάλες ροές θα αντιμετωπίζονται από τον controller έτσι ώστε να μετριάζεται η

καθυστέρηση εγκατάστασης της ροής και το επίβαρο του ελεγκτή (controller overhead). Το δίκτυο FLARE [56] αποτελεί ένα νέο δικτυακό μοντέλο το οποίο εστιάζει σε πλήρως προγραμματίσιμα δίκτυα, τα οποία παρέχουν ικανότητα προγραμματισμού για το επίπεδο δεδομένων, το επίπεδο ελέγχου, αλλά και την ενδιάμεση διεπαφή τους.

Στον τομέα της επεκτασιμότητας και της απόδοσης, μελέτες έχουνε δείξει ότι ένας μοναδικός controller μπορεί να χειριστεί μέχρι και 6 εκατομμύρια ροές ανά δευτερόλεπτο χωρίς προβλήματα. Πιο σύγχρονες μελέτες έχουνε δείξει ότι ένας controller μπορεί τελικά να διαχειριστεί πολύ περισσότερες ροές της τάξεως των 12,8 εκατομμύριων ανά δευτερόλεπτο σε 12-πύρηνο επεξεργαστή με μέση λανθάνουσα καθυστέρηση της τάξεως του 24.7 us για κάθε ροή (Beacon controller) [56]. Ωστόσο, για αυξημένη επεκτασιμότητα και ειδικά για σκοπούς αξιοπιστίας και στιβαρότητας έχει αναγνωρισθεί ότι ο λογικά κεντρικοποιημένος controller πρέπει να είναι φυσικά κατανεμημένος. Αυτή η προσέγγιση, με σκοπό την εύκολη επεκτασιμότητα του πεδίου ελέγχου χρησιμοποιείται από διάφορες ερευνητικές ομάδες όπως οι Onix [56], και HyperFlow[56]. Επιπλέον ερευνητικές μελέτες γίνονται για το εάν περισσότεροι controllers απλοποιούν το σύστημα ή αν το περιπλέκουν καθώς επίσης και το μέρος στο δίκτυο όπου θα είναι εγκατεστημένοι αυτοί. Σε πιο πρόσφατες μελέτες για κατανεμημένο έλεγχο, οι οποίες πρόεκυψαν από την ανάγκη για τη δυναμική ανάθεση των μεταγωγέων στους controllers, προτείνεται ένας αλγόριθμος για την αύξηση ή μείωση του συνόλου των controllers ανάλογα με τις εκτιμήσεις του φορτίου των ελεγκτών. Προτείνεται επίσης ένας μηχανισμός για την δυναμική μεταπομπή των μεταγωγέων από τον έναν ελεγκτή στον άλλο ανάλογα με τις ανάγκες του δικτύου.

Παρότι ο έλεγχος και η μέτρηση είναι δύο σημαντικά στοιχεία της διαχείρισης του δικτύου, δεν έχουν γίνει πολλά όσο αφορά το σχεδιασμό APIs για τη μέτρηση στα κέντρα δεδομένων. Μπορούμε να αναφερθούμε μονάχα σε μια SDN αρχιτεκτονική μέτρησης της κίνησης, η οποία διαχωρίζει το πεδίο των δεδομένων μέτρησης από το πεδίο ελέγχου.

2) Έρευνες σχετικά με νέα πρωτόκολλα και καλύτερο Internet

Οι απαιτήσεις για επεκτασιμότητα και καλύτερη απόδοση οι οποίες δημιουργούνται από ολοένα και αυξανόμενες πολύπλοκες εφαρμογές έχουν θέσει μια ποικιλία προκλήσεων οι οποίες δύσκολα θα αντιμετωπιστούν με τη σημερινή κατάσταση και αρχιτεκτονική του Internet. Πολλοί ειδικοί αναφέρουν ότι είναι απαραίτητη η αλλαγή ή η αναβάθμιση και εξέλιξη των πρωτοκόλλων που το διέπουν καθώς και της φυσικής υποδομής. Γίνεται λόγος για το ότι απαιτείται ένα **Software-Defined Internet** [56]. Ένα αξιοσημείωτο παράδειγμα είναι η ανάπτυξη του IPv6 [56].

Σε αυτόν τον τομέα οι λίγες έρευνες που έχουν γίνει περιλαμβάνουν την πρόταση μίας software-defined Internet αρχιτεκτονικής η οποία δανείζεται από MPLS [56] τη διάκριση μεταξύ των άκρων του δικτύου (edges) και του πυρήνα για να διαχωρίσει τα έργα ανάμεσα σε ‘’εντός τομέα’’ (intra- domain) και ‘’μεταξύ τομέων’’ (inter-domain). Δεδομένου ότι μόνο οι συνοριακοί δρομολογητές και ο αντίστοιχος controller τους σε κάθε τομέα συμμετέχουν σε inter-domain εφαρμογές, οι αλλαγές στα inter-domain μοντέλα υπηρεσιών θα περιοριζόντουσαν σε τροποποιήσεις του λογισμικού στους inter-domain controllers παρά σε ολόκληρη την υποδομή. Παραδείγματα του πως αυτή η αρχιτεκτονική θα μπορούσε να χρησιμοποιηθεί για να παρέχει νέες υπηρεσίες internet βρίσκονται ακόμα υπό εξέλιξη σε πειραματικά στάδια.

3) Αλληλεπίδραση μεταξύ Controller-Service

Ενώ η αλληλεπίδραση μεταξύ controller και μεταγωγέων (southbound) καθορίζεται ικανοποιητικά από πρωτόκολλα όπως το OpenFlow, δεν υπάρχει κανένα πρότυπο για τις αλληλεπιδράσεις μεταξύ των ελεγκτών και των δικτυακών υπηρεσιών ή εφαρμογών (northbound). Μια πιθανή εξήγηση είναι ότι η northbound διασύνδεση καθορίζεται εξ ολοκλήρου στον τομέα του λογισμικού, ενώ οι αλληλεπιδράσεις ελεγκτή-μεταγωγέα απαιτούν υλοποίηση με hardware [56].

Σε αυτό το τομέα οι έρευνες διεξάγονται με βάση το γεγονός ότι οι εφαρμογές θα έπρεπε να μπορούσαν να αποκτούν πρόσβαση στους μεταγωγείς, να συνυπάρχουν και να αλληλεπιδρούν με άλλες εφαρμογές και να κάνουν χρήση των υπηρεσιών του δικτύου όπως πχ η προώθηση, χωρίς να απαιτείται ο developer της εφαρμογής να γνωρίζει τις λεπτομέρειες εφαρμογής του ελεγκτή.

Μερικές προτάσεις υποστηρίζουν τη χρήση μιας προγραμματιστικής γλώσσας για δικτυακή διάρθρωση έτσι ώστε να εκφράζουν τις δικτυακές πολιτικές (Procera, Frenetic, FML, Nettle) [56]. Για παράδειγμα η Procera [56] παρασκευάζει ένα επίπεδο πολιτικής πάνω από τους υπάρχοντες controllers, για να επικοινωνεί με τα αρχεία διαμόρφωσης και τα GUIs. Το προτεινόμενο επίπεδο πολιτικής είναι υπεύθυνο για τη μετατροπή των υψηλών επιπέδων πολιτικής σε εντολές ροής που θα χρησιμοποιηθούν από τον controller. Σε άλλες πειραματικές μελέτες προτείνονται διαρθρώσεις του δικτύου και μηχανισμοί διαχείρισης, οι οποίοι εστιάζουν στη δραστικοποίηση των αλλαγών της κατάστασης του δικτύου, στην υποστήριξη της τήρησης της πολιτικής, και στην παροχή πλήρης ενημέρωσης και ελέγχου των εφαρμογών, για τη διάγνωση και την αντιμετώπιση προβλημάτων δικτύου.

Επιπλέον το northbound API θα έπρεπε να επιτρέπει σε εφαρμογές να ασκούν διαφορετική πολιτική για την ίδια ροή. Υπάρχουν ερευνητικές ομάδες οι οποίες εστιάζουν στο να διασφαλίσουν ότι κανόνες οι οποίοι σχεδιάστηκαν για νε εκτελέσουν ένα έργο ή να διαμορφώσουν μία εφαρμογή δεν θα παραμερίζουν τους υπόλοιπους κανόνες [56].

4) Εικονικοποίηση και υπηρεσίες Cloud

Η ζήτηση για εικονικοποίηση και υπηρεσίες cloud αυξάνεται με ταχείς ρυθμούς και προσελκύει μεγάλο ενδιαφέρον από τη βιομηχανία και τους τομείς ανώτατης εκπαίδευσης. Οι προκλήσεις που παρουσιάζονται περιλαμβάνουν, γρήγορη παροχή, αποτελεσματική διαχείριση των πόρων και επεκτασιμότητα, οι οποίες μπορεί να αντιμετωπιστούν με τη χρήση του μοντέλου SDN όπως και έχει προαναφερθεί στις ενότητες 4.1 και 4.2.

Για παράδειγμα οι FlowVisor [56] και AutoSlice [56] εστιάζουν στη δημιουργία διαφορετικών τομέων από τους πόρους ενός δικτύου (π.χ., το εύρος ζώνης, τοπολογία, CPU, πίνακα προώθησης) και στην ανάθεση τους σε διαφορετικούς controllers αλλά και στην αναγκαστική απομόνωση μεταξύ των τομέων. Άλλοι SDN ελεγκτές μπορούν να χρησιμοποιηθούν ως ένα σύστημα υποστήριξης δικτύου για την υλοποίηση της εικονικοποίησης σε λειτουργικά συστήματα cloud όπως είναι το Floodlight [56] για το OpenStack [56] και το NOX [56] για το Mirage[56]. Το FlowN [56] έχει ως στόχο να προσφέρει μια επεκτάσιμη λύση για το virtualization του δικτύου παρέχοντας μια αποτελεσματική χαρτογράφηση μεταξύ των φυσικών και εικονικών δικτύων και με την χρήση επεκτάσιμων συστημάτων για τις βάσεις δεδομένων.

Μία ακόμα ερευνητική ομάδα πρότεινε έναν αλγόριθμο για αποτελεσματική μετακίνηση ή μεταφορά του bandwidth χρησιμοποιώντας OpenFlow. Το LIME [56] είναι μια SDN-based λύση για μετακίνηση των εικονικών μηχανών, η οποία χειρίζεται την κατάσταση του δικτύου κατά τη διάρκεια της μετακίνησης και αυτόματα διαρθρώνει τις συσκευές του δικτύου σε νέες τοποθεσίες. Το NetGraph [56] παρέχει ένα σύνολο από APIs έτσι ώστε οι υπεύθυνοι του δικτύου να έχουν πρόσβαση στις εικονικές λειτουργίες του δικτύου.

5) Information-Centric Networking

Η κεντρική δικτύωση πληροφοριών- **Information-Centric Networking (ICN)** [56] αποτελεί ένα νέο πρότυπο που προτείνονται για τη μελλοντική αρχιτεκτονική του Διαδικτύου, η οποία στοχεύει στην αύξηση της απόδοσης της παράδοσης του περιεχομένου και της διαθεσιμότητας αυτού. Η νέα αυτή αντίληψη έχει διαδοθεί πρόσφατα από έναν αριθμό προτάσεων αρχιτεκτονικής δικτύων, όπως το Content- Centric Networking (**CCN**) [56]. Το κίνητρο για την ανάπτυξη του ICN είναι ότι η σημερινή κατάσταση του διαδικτύου είναι βασισμένη γύρω από το χτίσιμο πληροφοριών (information driven) ενώ η τεχνολογία δικτύων εξακολουθεί να επικεντρώνεται στην ιδέα του location – based και host-to-host επικοινωνία. Προτείνοντας μια αρχιτεκτονική η οποία θα απευθύνεται στα επώνυμα δεδομένα (named data) παρά στους hosts, η διανομή του περιεχομένου υλοποιείται απευθείας στο δικτυοδόμημα,αντί να στηρίζεται στην περίπλοκη χαρτογράφηση, διαθεσιμότητα και μηχανισμούς ασφαλείας που χρησιμοποιούνται σήμερα για τη χαρτογράφηση του περιεχόμενου σε μια συγκεκριμένη τοποθεσία.

Ο διαχωρισμός μεταξύ της επεξεργασίας των πληροφοριών και την προώθηση αυτών στην ICN είναι αντίστοιχο με τον διαχωρισμό των πεδίων δεδομένων και πεδίο ελέγχου στο SDN. Το ερώτημα γίνεται στη συνέχεια πώς θα συνδυαστούν το ICN με το SDN για την δημιουργία του "**Software-Defined Information-Centric Networks**" [56]. Ένας σχετικά εντυπωσιακός αριθμός από ερευνητικές ομάδες έχουν προτείνει τη χρήση εννοιών SDN για την εφαρμογή των ICN. Καθώς το OpenFlow επεκτείνεται για να υποστηρίξει εξατομικευμένες κεφαλίδες, το SDN μπορεί να χρησιμοποιηθεί ως μια καθοριστική τεχνολογία κλειδί για την ανάπτυξη του ICN [56].

Κεφάλαιο 5
Σύνοψη και Συμπεράσματα

Ως πόρισμα της συγκεκριμένης διπλωματικής εργασίας, προκύπτει από όλα τα προαναφερθέντα, ότι με τους ολοένα και αυξανόμενους ρυθμούς κίνησης και κυκλοφορίας στα κέντρα δεδομένων αλλά και με το virtualization και την ανάγκη για εξυπηρέτηση των cloud εφαρμογών, τα κέντρα δεδομένων πρέπει να εξελιχθούν για να καλύψουν τις απαιτήσεις αυτές, καθώς η σημερινή τεχνολογία δεν θα επαρκεί για να καλύψει τις μελλοντικές τους ανάγκες όπως επεξηγήθηκε διεξοδικά στην ενότητα 4.2. Αυτο θα πραγματοποιηθεί με την οικειοποίηση οπτικών στοιχείων στις αρχιτεκτονικές δομές των κέντρων δεδομένων. Σήμερα τα κέντρα δεδομένων κάνουν χρήση οπτικής τεχνολογίας αλλά αποκλειστικά με την μορφή point to point ζεύξεων υψηλής χωρητικότητας, ενώ το έργο της μεταγωγής αναλαμβάνουν ηλεκτρονικές διατάξεις, μετά από οπτο-ηλεκτρονική (O-E) μετατροπή του οπτικού σήματος. Το συμπέρασμα και η θέση της παρούσας διπλωματικής είναι ότι η παρούσα τεχνολογία που βασίζεται σε point to point οπτικές ζεύξεις και ηλεκτρονικούς switches για την μεταγωγή των πακέτων, δεν μπορεί να ακολουθήσει τους ρυθμούς αύξησης της κίνησης με τρόπο αποδοτικό και εφικτό οικονομικά στο μέλλον σύμφωνα με τις αναμενόμενες ανάγκες.

Όπως αναλύθηκε και διεξοδικά στην εργασία υπάρχει ανάγκη για αναβάθμιση των data centers σε επόμενο επίπεδο ώστε να μπορούν να επεκταθούν περαιτέρω για να ακολουθήσουν τις απαιτήσεις. Η ραγδαία διάδοση των υπηρεσιών cloud οδηγεί σε αντίστοιχη αύξηση του μεγέθους των datacenters, εισάγοντας το παράδειγμα των λεγόμενων mega-detacenters με έκταση εκατομμύρια τετραγωνικά μέτρα, που στεγάζουν εκατοντάδες χιλιάδες εξυπηρετητές. Μια από τις μεγαλύτερες προκλήσεις για τη δημιουργία τόσο μεγάλων δομών είναι η κατανάλωση ενέργειας, καθώς δεν είναι εφικτό να κατασκευαστεί datacenter με ενεργειακές απαιτήσεις περισσότερες των 20 MW περίπου, λόγω πρακτικών δυσκολιών στην παροχή ενέργειας. Συνεπώς είναι ύψιστη προτεραιότητα η βελτίωση της ενεργειακής απόδοσης κάθε υποσυστήματος του datacenter, συμπεριλαμβανομένου φυσικά του δικτύου. Η οπτική τεχνολογία είναι πολλά υποσχόμενη καθώς μπορεί να προσφέρει μεταγωγή με χαμηλότερη κατανάλωση ενέγειας σε σχέση με τα αποκλειστικά ηλεκτρικά κυκλώματα μεταγωγής, διαφάνεια (transparency) σε τύπο και ρυθμό δεδομένων, καλύτερη διαχείριση των πόρων του συστήματος (λόγω δυναμικής αναδιαμόρφωσης του δικτύου) αλλά ακόμη και μεγαλύτερο εύρος ζώνης. Η οπτική τεχνολογία όμως έχει μια σειρά από ιδιαιτερότητες που πρεπει να λαμβάνονται υπόψιν στο σχεδιασμό του δικτύου (πχ χρόνος μεταγωγής, έλλειψη buffer, περιορισμοί στις διαστάσεις στοιχείων – scalability) και έτσι μία πλήρης αντικατάσταση της υλικής υποδομής ενός data center από ηλεκτρικά κυκλώματα σε οπτικά θα είναι εξαιρετικά δύσκολη έως και αδύνατη χωρίς να αλλάξει ολόκληρη η αρχιτεκτονική του συστήματος. Υπάρχουν μια σειρά από αρχιτεκτονικές που μπορούν να ξεπεράσουν τα προβλήματα αυτά που περιγράφηκαν. Ιδιαίτερα όσον αφορά το πρόβλημα του χρόνου μεταγωγής προτείνονται τα υβριδικά δίκτυα τα οποία χρησιμοποιούν σε πρώτη πρόσβαση ένα ηλεκτρικό κύκλωμα για μεταγωγή και όταν το κύκλωμα αυτό έχει κορεστεί τότε γίνονται χρήσεις οπτικών κυκλωμάτων για να εξυπηρετήσουν τις πιο μεγάλες ροές, που μεταβάλλονται πιο αργά. Η χρήση και των δύο τεχνολογιών είναι ο ομαλότερος τρόπος μετάβασης των κέντρων δεδομένων σε πρώτη φάση σήμερα. Μελλοντικά ανάλογα και με τις ανάγκες του κάθε κέντρου ίσως να απαιτούνται μόνο υλοποιήσεις με οπτική τεχνολογία. Τέλος, οι ιδιαιτερότητες των δικτύων αυτών επεκτείνονται και στον έλεγχό τους.

Προτείνεται ο έλεγχος στα πλαίσια της τεχνολογιας SDN που περιγράφηκε στο κεφάλαιο 4 και επισημάνθηκαν καίρια σημεία για περαιτέρω μελλοντική ανάπτυξη και έρευνα.

Στην εργασία παρουσιάστηκαν και εξεταστήκαν 4 κατηγορίες αρχιτεκτονικών δομών, με βάση τα ενεργά στοιχεία που χρησιμοποιούν για τη δρομολόγηση, τα οποία και καθορίζουν τα βασικά χαρακτηριστικά της εκάστοτε αρχιτεκτονικής: Plexxi, AWG-based, 3D MEMS και WSS-based. Όλες οι μορφές αυτές κάνουν χρήση οπτικής τεχνολογίας η οποία προσφέρει σημαντικά πλεονεκτήματα σε σύγριση με τα ηλεκτρικά κυκλώματα πακέτων. Το δίκτυο Plexxi όπως παρουσιάστηκε στην ενότητα 3.1 έχει ως φυσική τοπολογία δακτύλιο αλλά η λογική του τοπολογία λειτουργεί ως mesh. Το Plexxi κάνει χρήση της WDM τεχνολογίας οπτικής διασύνδεσης με δύο lightrail καλώδια ένα προς κάθε κατεύθυνση μεταφέροντας μέχρι και 240 Gbps. Η αρχιτεκτονική αυτή κάνει χρήση 10 GbE συνδέσεων και έχει ως προδιαγραφές 2,56 Tb/s χωρητικότητα μεταγωγής, 480 nanosecond latency και 400 W ώς μέγιστη κατανάλωση ενέργειας σε κάθε μεταγωγέα. Χαρακτηρίζεται από γραμμική επεκτασιμότητα όπου για την προσθήκη νέου μεταγωγέα στο δακτύλιο, αυτός εισέρχεται ανάμεσα σε δυο ήδη υπάρχοντες μεταγωγείς, διαδικασία που άλλες φορές είναι εύκολη και άλλες όχι καθώς εξαρτάται από πολλούς παράγοντες αλλά αυτό καθιστά προβλέψιμα τα κόστη επεκτασιμότητας. Το σημαντικό με αυτή την αρχικτεκτονική είναι ότι είναι εμπορικά διαθέσιμη και εφαρμόσιμη σε κέντρα δεδομένων και κυρίως υλοποιήσιμη με τη σημερινή υποδομή της οπτικής τεχνολογίας. Δημοφιλές είναι το υβριδικό μοντέλο Plexxi-Calient το οποίο παρέχει καλύτερο scaling και επαναδιαρθρωσιμότητα στο κέντρο δεδομένων. Η αρχιτεκτονική Plexxi επίσης αποτελεί μία ομαλή μετάβαση του κέντρου δεδομένων στα δίκτυα της επόμενης γενιάς όπως αναφέρθηκε και προηγουμένως και θα εξυπηρετεί τις ανάγκες των κέντρων δεδομένων για αρκετά χρόνια.

Η ομαλή μετάβαση των κέντρων δεδομένων σε δίκτυα επόμενης γενιάς είναι επίσης εύκολη και με την τεχνολογία των 3D MEMS τα οποία αποτελούν υβριδικές αρχιτεκτονικές κάνοντας χρήση και οπτικού αλλά και ηλεκτρικού κυκλώματος. Αυτο συμβαίνει καθώς εάν αναλογιστούμε ένα σημερινό κέντρο δεδομένων όπως παρουσιάστηκε στην ενότητα 2.2, αποτελείται κυρίως από ηλεκτρονικούς μεταγωγείς και έτσι μια υβριδική προσέγγιση θα ήταν ομαλή μετάβαση για το προσεχές μέλλον. Τα οπτικά κυκλώματα των MEMS είναι NxN συμπλέγματα από κάτοπτρα που περιστρέφονται και αντανακλούν την οπτική δέσμη. Καθώς λειτουργούν σε οπτικό επίπεδο δεν απαιτούν αποκωδικοποιητές πακέτων και μετατροπείς από ηλεκτρικά σε οπτικά και το ανάστροφο, πράγμα το οποίο συμβάλλει καθοριστικά στην λιγόστευση του κόστους του δικτύου. Η περιστροφή των κατόπτρων εισάγει μια μικρή καθυστέρηση μεταγωγής η οποία είναι της τάξης των 25 ms – 50 ms. Ενώ αυτό για ένα ηλεκτρικό κύκλωμα μεταγωγής θα ήταν μη αποδεκτό στις υβριδικές υλοποιήσεις όπου το οπτικό κομμάτι του δικτύου διαχειρίζεται την αργά μεταβαλλόμενη κίνηση, οι χρόνοι πληροφορίας και μεταγωγής είναι μεγάλοι (μερικές φορές είναι και λεπτά) η καθυστέρηση αυτή είναι αδιάφορη. Τόσο τα 3D MEMS όσο και το δίκτυο Plexxi είναι πιο άμεσα υλοποιήσιμα για τα κέντρα δεδομένων καθώς συνήθως το κύριο κομμάτι της κίνησης εξυπηρετείται από τα ηλεκτρικά κυκλώματα μεταγωγής (ή το κύριο δακτύλιο στη περίπτωση του Plexxi) και όταν παρουσιαστεί μεγάλο traffic bulk ή πληροφορία προτεραιότητας ή όταν το ηλεκτρικό κύκλωμα δεν θα μπορεί να ανταπεξέλθει τότε γίνεται χρήση του οπτικού κυκλώματος. Αυτές οι δύο τεχνολογίες προτείνονται για το άμεσο μέλλον από την εργασία αυτή καθώς είναι πιο εύκολα υλοποιήσιμες σύμφωνα με την εικόνα των κέντρων δεδομένων σήμερα. Δυστυχώς όμως η επεκτασιμότητα των 3D-MEMS είναι περιορισμένη σήμερα με το κάθε OCS να αποτελείται απο συγκεκριμένο μέγιστο αριθμό θυρών. Ενδειχτικές τιμές αριθμού θυρών είναι 320 ενώ έρευνητικές διεργασίες διεξάγονται για επέκταση αυτών σε 1000 θύρες.

Οι αρχιτεκτονικές βασισμένες στην μονάδα AWG και η λειτουργία της αναλύθηκαν στην ενότητα 3.2. Τα AWG προσφέρουν την πιθανότητα επίτευξης υψηλής ταχύτητας μεταγωγής λόγω της ύπαρξης tunable lasers με την ικανότητα γρήγορης επαναρύθμισης του μήκους κύματος (πχ DS-DBR lasers). Υπάρχει μία μεγάλη ποικιλία εφαρμογών με AWG όπου η κάθε αρχιτεκτονική προσφέρει διαφορετικά οφέλη κάθε φορά και προορίζεται για να καλύψει συγκεκριμένες ανάγκες. Το μέσο latency στις εφαρμογές αυτές που προκύπτει από μετρήσεις είναι 2,5 ms το οποίο είναι αρκετά μικρό για τα δεδομένα των οπτικών τεχνολογιών, ενώ πειραματικές υλοποιήσεις tunable-laser με γρήγορο χρόνο ρύθμισης μήκους κύματος έχουν δείξει τη δυνατότητα για χρόνους μεταγωγής της τάξης των μερικών ns. Δυστυχώς ένα μεγάλο μειονέκτημα των αρχιτεκτονικών αυτών είναι η επεκτασιμότητα καθώς γενικά εξαρτάται από τη διάσταση του AWG, καθώς και το κόστος καθώς απαιτείται η διασύνδεση με tunable lasers που γενικά είναι πιο ακριβά από αντίστοιχα lasers που εκπέμπουν σε σταθερό μήκος κύματος. Επίσης η επεκτασιμότητα γίνεται ακόμα πιο δύσκολη αν αναλογιστούμε τον αριθμό των καλωδίων με τον οποίο συνδέεται η κάθε AWG μονάδα με το υπόλοιπο δίκτυο. Παρά τη δύσκολη επεκτασιμότητα, με σωστό σχεδιασμό μιας αρχιτεκτονικής βασισμένης στα AWG έχουν προταθεί διάφορες αρχιτεκτονικές που επιτυγχάνουν τη δημιουργία ενός μεγάλου δικτύου, ενώ με κατάλληλο έλεγχο (controlplane) μπορεί να περιοριστεί σημαντικά μέχρι και να μηδενιστεί το contention των πακέτων στις εξόδους. Ένα μεγάλο μέρος του latency στις AWG εφαρμογές εξαρτάται από το μέγεθος και το πλήθος των πακέτων στις εισόδους των tunable lasers και αυτό ειναι το κύριο πρόβλημα που μπορεί να παρουσιαστεί εάν το δίκτυο δεν μπορεί να διαχειριστεί ένα μεγάλο πλήθος πακέτων. Έξυπνος σχεδιασμός της αρχιτεκτονικής του δικτύου όμως σε συνδυασμό με αλγορίθμους, όπως για παράδειγμα αποτελεί η αρχιτεκτονική DOS στην ενότητα 3.2.5 με την ανακύκλωση των πακέτων και την επαναεισαγωγή τους στο δίκτυο μπορεί μέχρι και να μηδενίσει το contention. Έτσι τα AWG based δίκτυα μπορεί να αντιμετωπίσουν αρκετές εφαρμογές ευαίσθητες στο latency.

Επιπλέον, στην εργασία παρουσιάστηκαν και αρχιτεκτονικές που βασίζονται σε στοιχεία WSS. Το σημαντικό και κύριο χαραακτηριστικό των αρχιτεκτονικών αυτών είναι η μεγάλη ευελιξία κατανομής των πόρων του συστήματος καθώς και οι εξαιρετικά μικροί χρόνοι μεταγωγής λόγω της τεχνολογίας WSS που χρησιμοποιήθηκε που μεταφράζεται σε αρκετά μικρό latency. Στο δίκτυο Mordia που αποτελεί και την πιο ολοκληρωμένη εφαρμογή το switching time ήταν μόλις 10-11,5 μs πράγμα ασυναγώνιστο με όλες τις προηγούμενες εφαρμογές και υλοποιήσεις. Ακόμη οι WSS-based εφαρμογές παρουσίαζαν αρκετά υψηλές τιμές του throughput, με αυτό να είναι ίσο με το 87,9 % - 95,4% του συνολικού bandwidth. Επίσης στις WSS εφαρμογές η επεκτασιμότητα δεν ήταν ιδιαίτερα δύσκολη, ίδιας κλίμακας και επιπέδου με του Plexxi μιας και κάνουμε λόγο για δακτύλιο αλλά οι εφαρμογές είχαν ένα ανώτατο όριο θυρών στις συσκευές. Καμία υλοποίηση δεν ξεπερνούσε τις 704 θύρες και γενικά τα 22 με 23 stations χωρίς να δημιουργεί προβλήματα στο δίκτυο. Ένας βασικός περιορισμός των αρχιτεκτονικών αυτών είναι το κόστος, καθώς τα στοιχεία αυτά κατασκευάζονται για τις ανάγκες τηλεπικοινωνιακών δικτύων και είναι αρκετά ακριβά. Οι WSS-based αρχιτεκτονικές είναι πολλά υποσχόμενες αλλά απαιτούν περαιτέρω έρευνα για την εφαρμογή τους σε κέντρα δεδομένων μεγάλης κλίμακας και σίγουρα θα απασχολήσουν τους σχεδιαστές δικτύων αρκετά στο μέλλον.

Τέλος, καθώς το Internet γίνεται όλο και πιο δημοφιλές και οι απαιτήσεις των διαδικτυακών υπηρεσιών αυξάνονται, χρειάστηκε να δημιουργηθούν νέα πρωτόκολλα καθώς σωστή και αποδοτική λειτουργία μιας αρχιτετονικής χρειάζεται κατάλληλο έλεγχο για την ορθή λειτουργία της. Οι ειδικοί δικτύων παρατήρησαν ότι η κυρίαρχη αιτία, που οδήγησε στην αύξηση της

πολυπλοκότητας, ήταν το γεγονός ότι πολλές οντότητες ενός δικτύου εκτελούσαν ταυτόχρονα ελεγκτικές διαδικασίες και ενέργειες που σχετίζονται με την προώθηση ενός πακέτου. Έτσι λοιπόν, προτάθηκε η αρχιτεκτονική **SDN**, κατά την οποία η οργάνωση και διαχείριση των δικτυακών πόρων διαχωρίζεται στο επίπεδο ελέγχου (control plane) και στο επίπεδο δεδομένων (data plane). Περιληπτικά, η αρχιτεκτονική SDN καθορίζει πως το επίπεδο ελέγχου μπορεί να θέσει τις παραμέτρους των ροών του δικτύου κάνοντας χρήση του πρωτοκόλλου Open Flow. Αντίστοιχα, το επίπεδο δεδομένων συνιστάται από δίκτυα IP, που όμως περιλαμβάνουν ένα σύνολο από μεταγωγείς που αποκαλούνται Open Flow Switches. Τα Open Flow Switches ορίζουν την βέλτιστη προώθηση των πακέτων του δικτύου, δεδομένου ότι μπορούν να προγραμματιστούν δυναμικά μέσω του πρωτοκόλλου Open Flow από τους διαχειριστές του δικτύου. Η επέκταση του SDN ώστε να περιλαμβάνει τον έλεγχο οπτικών πομποδεκτών αποτελεί ανοικτό θέμα έρευνας των τελευταίων ετών, ενώ η περαιτέρω επέκτασή του για την υποστήριξη οπτικών στοιχείων δρομολόγησης είναι ένα εξαιρετικά ενδιαφέρον πεδίο που αναμένεται να απασχολήσει την ερευνητική κοινότητα αλλά και να μεταφραστεί σύντομα σε έντονη επιχειρηματική δραστηριότητα στον τομέα των δικτύων δεδομένων.

Βιβλιογραφία

[1] Cisco, 'Cisco Global Cloud Index: Forecast And Methodology 2013–2018 White Paper' white paper

[2] Inc, S., (2002). Networking Complete. Third Edition. San Francisco: Sybex

[3] Bicsi, B., (2002). Network Design Basics for Cabling Professionals. City: McGraw-Hill Professional

[4] Cisco,. 'What Is A Network Switch Vs. A Router?' (http://www.cisco.com/cisco/web/solutions/small_business/resource_center/articles/connect_employees_and_offices/what_is_a_network_switch/index.html)

[5] Peter Newman ‘’Fast Packet Switching for Integrated Services Chapter 4Multi-Stage Interconnection Networks’’ (http://pnewman.com/papers/thesis/chapter4.pdf),Thesis March 1989

[6] Douglas Comer, ‘’Computer Networks and Internets’’, page 99 ff, Prentice Hall 2008

[7] Aroca, Jordi Arjona, and Antonio Fernández Anta, ‘’Bisection (band) width of product networks with application to data centers, Theory and Applications of Models of Computation’’, Springer Berlin Heidelberg, 2012. 461-472.

[8] Performance Testing Basics – What is Throughput (http://www.joecolantonio.com/2011/07/05/performance-testing-what-is-throughput)

[9] W. J. Dally, B. Towles ‘’Principles and Practices of Interconnection Networks’’ Morgan Kaufmann, 2004

[10] Oppenheimer, P. 2011 Top-Down Network Design. Indianapolis:Cisco Press

[11] B. Parhami, ‘Exact Formulas for the Average Internode Distance in Mesh and Binary Tree Networks, Computer Science and Information Technology’, Vol. 1, No. 2, pp. 165-168, 2013

[12] Al-Fares, Mohammad, Alexander Loukissas, and Amin Vahdat. 'A Scalable, Commodity Data Center Network Architecture'. *SIGCOMM Comput. Commun. Rev.* 38.4 (2008):63. Web.

[13] Han, Sangjin et al. 'Network Support For Resource Disaggregation In Next-Generation Datacenters'. *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks - HotNets-XII* (2013):n. pag.

[14] Hedlund, Brad. 'Top Of Rack Vs End Of Row Data Center Designs'. *Brad Hedlund*. N.p., 2009.(http://bradhedlund.com/2009/04/05/top-of-rack-vs-end-of-row-data-center-designs/).

[15] Liao, James. 'Inside A TOR Switch'. *Pronto Systems*. (https://prontosystems.wordpress.com/2011/02/04/inside-a-tor-switch/)

[16] Cisco, 'Data Center Top-Of-Rack Architecture Design'. white paper (http://www.cisco.com/c/en/us/products/collateral/switches/nexus-5000-series-switches/white_paper_c11-522337.html)

[17] Calient, The Software Defined Hybrid Packet Optical Datacenter Network (http://comdate.com.au/media/dnload/TheHybridPacketOpticalDatacenterNetwork.pdf) white paper

[18] Plexxi Switch 1x Affinity Networking™ Switch (http://www.plexxi.com/wp-content/uploads/2014/06/Plexxi_Switch_1_Datasheet.pdf) Datasheet

[19] Plexxi Switch 2x Affinity Networking™ Switch (http://www.plexxi.com/wp-content/uploads/2013/11/DS_PLX_Switch2_2013_vfinal1.pdf) Datasheet

[20] Plexxi paths and topologies (http://www.plexxi.com/2013/09/plexxi-paths-and-topologies-part-1-let-there-be-light/)

[21] Affinities in Action Plexxi and Calient (http://www.plexxi.com/wp-content/uploads/2013/03/Plexxi-and-CALIENT-Solution-Brief-March-2013.pdf)

[22] Plexxi In Depth Switch 2 (http://www.plexxi.com/wp-content/uploads/2013/11/Switch-2-Product-Brief.pdf) Datasheet

[23] Amersfoort, M., 'Arrayed Waveguide Grating', in Application note. 1998, C2V:Enschede, Netherlands. (http://web.archive.org/web/20070927081252/http://www.c2v.nl/products/software/support/files/A1998003B.pdf) 15 June 1998,white paper.

[24] D. Siracusa, G. Maier, V. Linzalata, A. Pattavina 'Scalability of Optical Interconnections based on the Arrayed Waveguide Grating in High Capacity Routers' (2011), Optical Network Design and Modeling (ONDM), 2011 15th International Conference

[25] Rastegarfar, Houman et al. 'A High-Performance Network Architecture For Scalable Optical Datacenters'. *IEEE Photonic Society 24th Annual Meeting* (2011).

[26] Neel, Brian et al. 'SPRINT:Scalable Photonic Switching Fabric For High-Performance Computing (HPC)'. *J. Opt. Commun. Netw*. 4.9 (2012):A38.

[27] Ye, Xiaohui et al. 'DOS'. *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems - ANCS '10* (2010).

[28] Proietti, Roberto et al. 'Scalable Optical Interconnect Architecture Using AWGR-Based TONAK LION Switch With Limited Number Of Wavelengths'. *J. Lightwave Technol.* 31.24 (2013):4087-4097..

[29] Mems-exchange.org,. 'What Is MEMS Technology?'. (https://www.mems-exchange.org/MEMS/what-is.html)

[30] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat. Helios:A hybrid electrical/optical switch architecture for modular data centers. In ACM SIGCOMM, 2010.

[31] Singla, Ankit et al. 'Proteus:A Topology Malleable Data Center Network'. *Proceedings of the Ninth ACM SIGCOMM Workshop on Hot Topics in Networks - Hotnets '10* (2010):n. pag..

[32] He Liu, et al. 'Reactor:A Reconfigurable Packet And Circuit Tor Switch'. *2013 IEEE Photonics Society Summer Topical Meeting Series* (2013):n. pag..

[33] Yuzo Ishii, Naoki Ooba, Akio Sahara, and Koichi Hadama 'WSS Module Technology for Advanced ROADM' NTT Technical Review, Vol. 12 No. 1 Jan. 2014

[34] Mapyourtech.com,. 'What Is Wavelength Selective Switch–WSS?'. 2013. (http://www.mapyourtech.com/entries/general/what-is-wavelength-selective-switch%E2%80%93wss)

[35] Robert Anderson,US Patent 6.542,657:'Binary Switch for an Optical Wavelength Router", April 1, 2003 (http://www.google.com/patents/WO2002050588A1?cl=en), Network Photonics Inc

[36] Wall, Pierre et al. 'WSS Switching Engine Technologies'. *OFC/NFOEC 2008 - 2008 Conference on Optical Fiber Communication/National Fiber Optic Engineers Conference* (2008).

[37] G. Baxter et al., "Highly Programmable Wavelength Selective Switch Based on Liquid Crystal on Silicon Switching Elements," in Optical Fiber Communication Conference, 2006 and the 2006 National Fiber Optic Engineers Conference. OFC 2006

[38] N. Farrington, A. Forencich, Pang-Chen Sun, S. Fainman, J. Ford, A. Vahdat, G. Porter, and G.Papen 'A 10 ms Hybrid Optical-Circuit/Electrical-Packet Network for Datacenters', Communications Magazine, IEEE (Volume:51, Issue:9)

[39] Porter, George et al. 'Integrating Microsecond Circuit Switching Into The Data Center'. *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM - SIGCOMM '13* (2013).

[40] Farrington, Nathan et al. 'Hunting Mice With Microsecond Circuit Switches'. *Proceedings of the 11th ACM Workshop on Hot Topics in Networks - HotNets-XI* (2012).

[41] Xu, Lin et al. 'A Hybrid Optical Packet And Wavelength Selective Switching Platform For High-Performance Data Center Networks'. *Opt. Express* 19.24 (2011).

[42] Hassan, Qusay. Demystifying Cloud Computing [J], The Journal of Defense Software Engineering (CrossTalk) (2011).

[43] 'Why Cloud Computing Technology is the New Revolution' (http://www.fonebell.in/cloud-computing-technology-new-revolution/)

[44] Das,S. et al (2011) 'Application-Aware Aggregation and Traffic Engineering in a Converged Packet-Circuit Network'. OFC/NFOEC

[45] 'SDN Architecture Overview' (https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/SDN-architecture-overview-1.0.pdf) white paper.

[46] 'Software-Defined Networking:The New Norm for Networks' (https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf) white paper.

[47] Jim Theodoras 'Software-Defined Networking for Transport Networks', ADVA Optical Networking, application white paper

[48] Software-Defined Networking (SDN) Definition (https://www.opennetworking.org/sdn-resources/sdn-definition) white paper.

[49] McKeown,N. et alia (2008) ''OpenFlow:Enabling Innovation in Campus Networks''. ACM SIGCOMM Computer Communication Review March 14, 2008

[50] Nascimento,M.R. et alia (2011) 'Virtual Routers as a Service:The RouteFlow Approach Leveraging Software-Defined Networks'.

[51] OpenContrail Architecture Documentation (http://www.opencontrail.org/opencontrail-architecture-documentation/)

[52] OpenDaylight the Project (http://www.opendaylight.org/project/about)

[53] A. Doria, Ed. et al. 'Forwarding and Control Element Separation (ForCES) Protocol Specification' (http://tools.ietf.org/html/rfc5810), Internet Engineering Task Force (IETF), March 2010

[54] S. Azodolmolky, M. N. Peterson, A. Manolova Fagertun, P. Wieder, S. R. Ruepp, R. Yahyapour, “SONEP:A Software-Defined Optical Network Emulation Platform,” ONDM 2014, 19-22 May 2014, Stockholm, Sweden.

[55] Hares, Susan, and Russ White. 'Software-Defined Networks And The Interface To The Routing System (I2RS)'. *IEEE Internet Comput.* 17.4 (2013):84-88.

[56] Nunes, Bruno Astuto A. et al. 'A Survey Of Software-Defined Networking:Past, Present, And Future Of Programmable Networks'. *IEEE Commun. Surv. Tutorials* 16.3 (2014):1617-1634.