2. ΕΝΑ ΒΗΜΑ ΑΚΟΜΑ ΣΤΟ ΔΙΑΔΙΚΤΥΟ

2.1 Πρωτόκολλα Δρομολόγησης IP

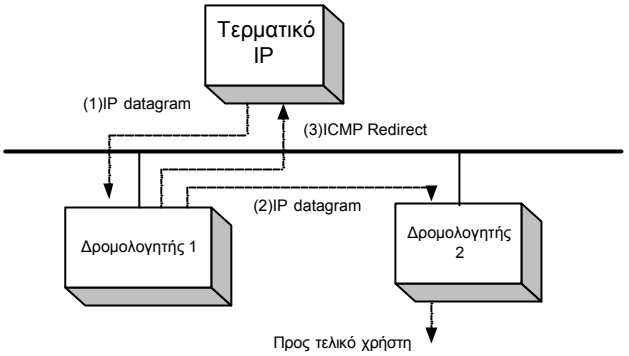
Τα πρωτόκολλα δρομολόγησης IP χωρίζονται σε δυο κατηγορίες: στα Πρωτόκολλα Εσωτερικής Δρομολόγησης (Interior Gateway Protocols, IGPs) και στα Πρωτόκολλα Εξωτερικής Δρομολόγησης (Exterior Gateway Protocols, EGPs). Τα IGPs εκτελούν δρομολόγηση σε δίκτυα που έχουν κοινή δικτυακή διαχείριση. Παραδείγματα IGPs είναι τα Routing Information Protocol (RIP), Open Shortest Path First (OSPF), Interior Gateway Routing Protocol (IGRP) κλπ. Αντίθετα προς τα IGPs, τα EGPs χρησιμοποιούνται για την ανταλλαγή πληροφορίας δρομολόγησης μεταξύ δικτύων που δεν μοιράζονται την ίδια διαχειριστική αρχή (τα ονομαζόμενα Αυτόνομα Συστήματα, AS). Το πιο χαρακτηριστικό πρωτόκολλο είναι το Border Gateway Protocol (BGP). Επιπλέον τα πρωτόκολλα δρομολόγησης διακρίνονται σε στατικής και δυναμικής δρομολόγησης.

2.1.1 Πρωτόκολλα Στατικής Δρομολόγησης

Τα πρωτόκολλα του τύπου αυτού δεν ανταλλάσσουν πληροφορία σχετική με την κατάσταση του δικτύου με άλλους δρομολογητές. Ο πίνακας δρομολόγησης δημιουργείται κάθε φορά που ενεργοποιείται μια διασύνδεση από τον διαχειριστή του κάθε σταθμού και παραμένει αναλλοίωτη, εκτός εξαιρέσεων που παρουσιάζονται στη συνέχεια, μέχρι την επόμενη ενημέρωση από τον διαχειριστή του συστήματος. Ο πίνακας δρομολόγησης δημιουργείται χειροκίνητα και κατόπιν αποθηκεύεται στα μέσα αποθήκευσης, ώστε να δημιουργείται αυτόματα πλέον σε κάθε επανεκκίνηση του συστήματος.

‘Οπως αναφέρθηκε παραπάνω, είναι δυνατόν, σε εξαιρετικές περιπτώσεις, η μορφή του πίνακα να αλλάξει, χωρίς την παρέμβαση του διαχειριστή. Αυτές οι περιπτώσεις χρησιμοποιούν το πρωτόκολλο ICMP, το οποίο είναι πρωτόκολλο στρώματος δικτύου.

Σε όλες τις περιπτώσεις μέχρι στιγμής έχουμε θεωρήσει ότι ο τελικός παραλήπτης θα λάβει το πακέτο του αποστολέα είτε μέσω ανεύρεσης της τελικής διεύθυνσης μέσα στον πίνακα δρομολόγησης είτε μέσω της εγγραφής default. Ας εξετάσουμε όμως την περίπτωση που δεν υπάρχει η εγγραφή default και καμία από τις εγγραφές μέσα στον πίνακα δεν είναι κατάλληλη για να προωθήσει το πακέτο. Σε αυτήν την περίπτωση, ο δρομολογητής θα στείλει μήνυμα ICMP Host Unreachable, πληροφορώντας το τερματικό αποστολής ότι δεν υπάρχει μονοπάτι προς τον τελικό παραλήπτη. Το πρωτόκολλο ICMP εκτός από την ενημέρωση για την μη ύπαρξη μονοπατιού προς κάποιον παραλήπτη έχει και άλλη χρησιμότητα. Ας δούμε την πρώτη περίπτωση χρησιμοποιώντας το Σχήμα 2.1.



Σχήμα 2.1
Η περίπτωση ICMP Redirect

Υποθέτουμε ότι το τερματικό στέλνει ένα πακέτο στον δρομολογητή 1, ο οποίος εμφανίζεται ως η εγγραφή default στον πίνακα δρομολόγησης του τερματικού. Ο δρομολογητής 1 μόλις λάβει το πακέτο θα ψάξει στον πίνακα δρομολόγησης του για κάποια εγγραφή που να ταιριάζει με τη διεύθυνση προορισμού IP. Κατά την αναζήτηση ανακαλύπτει ότι το μονοπάτι που οδηγεί στον τελικό παραλήπτη περνάει μέσα από τον δρομολογητή 2, οπότε προωθεί το πακέτο στην αντίστοιχη ζεύξη. Την στιγμή αυτή ανακαλύπτει ότι η ζεύξη στην οποία προωθείται το πακέτο είναι η ίδια με αυτήν από την οποία προήλθε. 'Ετσι μετά την αποστολή θα στείλει και ένα μήνυμα ICMP Redirect στο τερματικό IP ενημερώνοντάς το ότι μελλοντικά πακέτα της συγκεκριμένης διεύθυνσης παραλήπτη πρέπει να προωθούνται στον δρομολογητή 2. Με τον τρόπο αυτόν είναι δυνατόν το τερματικό να εκκινήσει έχοντας ως μόνη εγγραφή την εγγραφή default και με την πάροδο του χρόνου να δημιουργήσει έναν πίνακα δρομολόγησης, χρησιμοποιώντας το πρωτόκολλο ICMP καθώς και πληροφορία από γειτονικούς δρομολογητές, οι οποίοι θα πρέπει να περιέχουν κάποια πληροφορία σχετική με την τοπολογία του δικτύου. 'Ενα πρόβλημα με την μέθοδο αυτή είναι το ότι για κάθε προορισμό δημιουργείται μια εγγραφή στον πίνακα δρομολόγησης του τερματικού. Τα μονοπάτια που εγγράφονται στον πίνακα αφορούν μόνο προορισμούς τερματικών και όχι δικτύων. 'Ετσι ακόμη και αν δυο τερματικά βρίσκονται στο ίδιο δίκτυο, δεν είναι δυνατόν να γίνει μια μόνο εγγραφή στον πίνακα δρομολόγησης. Με την χρήση του ICMP είναι λοιπόν δυνατόν να δημιουργηθούν πίνακες δρομολόγησης σε ένα σύστημα, χωρίς παρέμβαση του διαχειριστή, ακόμη και αν αυτοί αρχικά περιέχουν μόνο την εγγραφή default. 'Οπως είδαμε οι πίνακες που δημιουργούνται είναι μεγάλοι λόγω της μη υποστήριξης διευθύνσεων δικτύου από το πρωτόκολλο. Το αποτέλεσμα είναι ότι οδηγούμαστε σε άσκοπη κατανάλωση μνήμης και αύξηση του χρόνου αναζήτησης μιας διεύθυνσης μέσα στον πίνακα. Τέλος το θέμα της χειροκίνητης ενημέρωσης του πίνακα δεν εξαλείφεται τελείως, αλλά απλά μετατίθεται. Ενώ δηλαδή ο διαχειριστής του συστήματος δεν απαιτείται να δημιουργήσει πίνακα στα τερματικά δεν ισχύει το ίδιο και για

τους δρομολογητές. Η στατική δρομολόγηση μπορεί ναι είναι αποτελεσματική στην περίπτωση των μικρών δικτύων, δεν ισχύει όμως το ίδιο όταν το δίκτυο μεγαλώσει.

2.1.2 Πρωτόκολλα Δυναμικής Δρομολόγησης

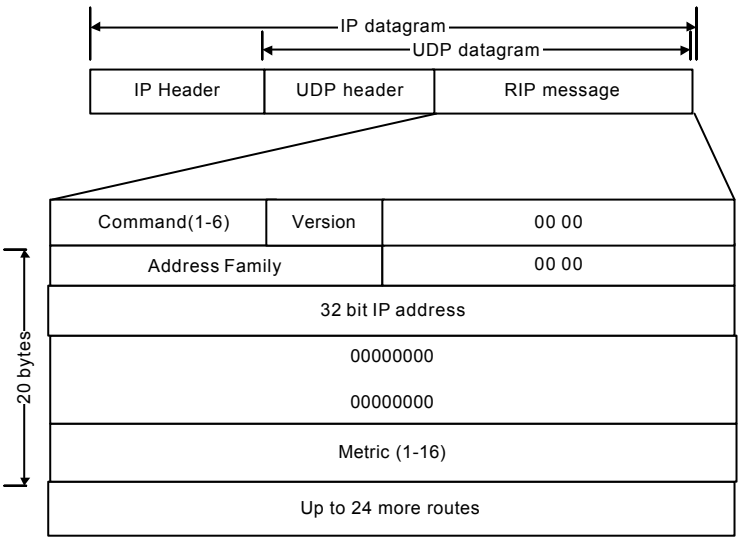
Με τον όρο δυναμική δρομολόγηση εννοούμε την ανταλλαγή πληροφοριών, μεταξύ γειτονικών δρομολογητών, σχετικών με τις τοπολογίες των δικτύων στα οποία συνδέονται οι δρομολογητές . Η διαδικασία, η οποία χρησιμοποιείται για την επικοινωνία μεταξύ των γειτονικών δρομολογητών ονομάζεται δαίμονας δρομολόγησης. Ο δαίμονας δρομολόγησης αναλαμβάνει την ενημέρωση των πινάκων βασισμένος στην πληροφορία που λαμβάνει από πίνακες γειτονικών δρομολογητών. Ο μηχανισμός δρομολόγησης δεν αλλάζει σε σχέση με τα πρωτόκολλα στατικής δρομολόγησης. Αυτό που αλλάζει είναι η πληροφορία που εγγράφεται στον πίνακα. Με άλλα λόγια ο δαίμονας καθορίζει την «πολιτική» δρομολόγησης αποφασίζοντας ποιο είναι το καλύτερο μονοπάτι προς έναν προορισμό (στην περίπτωση ύπαρξης περισσότερων του ενός μονοπατιού) και εγγράφοντας το στον πίνακα.

Τα πρωτόκολλα αυτά διακρίνονται σε εσωτερικής και εξωτερικής δρομολόγησης. Το διαδίκτυο είναι οργανωμένο σε αυτόνομα συστήματα. Το κάθε αυτόνομο σύστημα έχει τον δικό του διαχειριστή, ο οποίος καθορίζει ποιο πρωτόκολλο θα χρησιμοποιηθεί για την επικοινωνία μεταξύ των δρομολογητών του αυτόνομου συστήματος. Αυτά ονομάζονται Πρωτόκολλα Εσωτερικής Δρομολόγησης (IGP, Interior Gateway protocols). Για την επικοινωνία μεταξύ δρομολογητών διαφορετικών αυτόνομων συστημάτων χρησιμοποιούνται τα Πρωτόκολλα Εξωτερικής Δρομολόγησης (EGP, Exterior Gateway Protocols), ένα από τα οποία παρουσιάζεται στο επόμενο κεφάλαιο. Κατά καιρούς έχουν παρουσιαστεί διάφορα πρωτόκολλα που εμπίπτουν στην κατηγορία Π Εσωτερικής Δρομολόγησης. Οι διαφορές τους, άλλοτε σημαντικές και άλλοτε όχι, βρίσκονται στον τρόπο επιλογής του βέλτιστου μονοπατιού προς έναν προορισμό, στην παροχή πληροφορίας διευθύνσεων δικτύου κλπ.

2.1.2.1 Routing Information Protocol (RIP)

Μέχρι πριν μερικά χρόνια το πιο δημοφιλές πρωτόκολλο της κατηγορίας αυτής ήταν το Routing Information Protocol, RIP. Τα μηνύματα του RIP μεταδίδονται με χρήση της πόρτας 520 του UDP. Στο Σχήμα 2.2 παρουσιάζεται τόσο η μορφή του πακέτου UDP που μεταφέρει το μήνυμα RIP, όσο και η μορφή του ίδιου του μηνύματος. Η πληροφορία για κάθε μονοπάτι έχει μέγεθος 20 bytes και σε κάθε μήνυμα μπορεί να μεταφερθεί πληροφορία για 1 έως 25 μονοπάτια. Το όριο των 25 μονοπατιών τέθηκε ώστε το συνολικό μέγεθος του μηνύματος να είναι μικρότερο από 512 bytes. Με τον περιορισμό αυτόν είναι πολύ συχνό το φαινόμενο να χρειάζονται περισσότερα του ενός μηνύματα για την αποστολή ολόκληρου του πίνακα δρομολόγησης. Η επικεφαλίδα του μηνύματος περιέχει τα

πεδία command και version. Το πεδίο command μπορεί να έχει τις τιμές 1(αίτηση παροχής πληροφορίας) ή 2 (απάντηση). Μπορεί ακόμη να περιέχει τις τιμές 5 και 6, δηλαδή αίτηση παροχής πληροφορίας για μέρος του πίνακα και παροχή της πληροφορίας αυτής αντίστοιχα. Το πεδίο version περιγράφει την έκδοση του πρωτοκόλλου RIP. Μετά την επικεφαλίδα ακολουθεί η πληροφορία των διαφόρων μονοπατιών, η οποία για κάθε μονοπάτι έχει μέγεθος 20 bytes. Το πεδίο metric παρέχει τον αριθμό των διαδοχικών δρομολογητών (hop count) από τους οποίους θα περάσει ένα μήνυμα μέχρι τον τελικό παραλήπτη, η διεύθυνση του οποίου περιέχεται στο πεδίο “32 bit IP address”.



Σχήμα 2.2
Μορφή μηνύματος RIP

Κατά την εκκίνηση ενός συστήματος που χρησιμοποιεί το RIP, το σύστημα αναζητά τις ενεργές διασυνδέσεις, με τις οποίες είναι συνδεδεμένο. Κατόπιν στέλνει πακέτα RIP ζητώντας τους πλήρεις πίνακες δρομολόγησης των γειτονικών δρομολογητών. Για την αίτηση παροχής του πλήρους πίνακα από γειτονικούς δρομολογητές τα πεδία command, address family και metric της επικεφαλίδας του μηνύματος τίθενται στις τιμές 1, 0 και 16 αντίστοιχα. Μετά την άφιξη της αίτησης αποστέλλεται μέσω ενός ή διαδοχικών μηνυμάτων ο πλήρης πίνακας του δρομολογητή αν η αίτηση είναι για την περίπτωση του πλήρους πίνακα. Αν όχι, τότε για κάθε εγγραφή της αίτησης εξετάζεται το πεδίο metric και τίθεται στην αντίστοιχη τιμή αλλιώς τίθεται στην τιμή 16. Η τιμή 16 στο πεδίο αυτό σημαίνει ότι δεν υπάρχει μέσα στον πίνακα μονοπάτι προς τον συγκεκριμένο προορισμό. Στη συνέχεια λαμβάνεται η απάντηση και το σύστημα που ζήτησε την πληροφορία ενημερώνει τον πίνακα δρομολόγησης του. Στην περιγραφή του πρωτοκόλλου καθορίζεται ότι για

κάθε σύστημα είναι απαραίτητες οι μεταδόσεις μέρους ή συνόλου του πίνακα του στους γειτονικούς δρομολογητές. Η πληροφορία που περιέχεται σε κάθε πίνακα έχει περιορισμένη διάρκεια ζωής. Αν κάποιο σύστημα διαπιστώσει ότι για κάποιο συγκεκριμένο μονοπάτι δεν έχει έρθει ενημέρωση για διάστημα πέραν των τριών λεπτών, το πεδίο metric του συγκεκριμένου μονοπατιού τίθεται στην τιμή 16. Αν περάσει ακόμη ένα λεπτό και η πληροφορία που αφορά το συγκεκριμένο μονοπάτι δεν ενημερωθεί, τότε διαγράφεται ολόκληρη η γραμμή από τον πίνακα. Το [RFC 1058] καθορίζει επίσης ότι εκτός από την ενημέρωση κατά τακτά χρονικά διαστήματα υπάρχει και η εξαναγκασμένη ενημέρωση που συμβαίνει όποτε αλλάζει το πεδίο metric για κάποιο μονοπάτι. Σε αυτήν την περίπτωση η εγγραφή που αφορά το μονοπάτι αυτό πρέπει να διαδοθεί στο δίκτυο μέσω του RIP. Το πεδίο metric που χρησιμοποιείται από το RIP μας παρέχει πληροφορία σχετική με την καθυστέρηση άφιξης του μηνύματος στον τελικό παραλήπτη. Το πεδίο metric έχει την τιμή 1 για όλα τα δίκτυα τα οποία συνδέονται απ’ ευθείας σε έναν δρομολογητή.

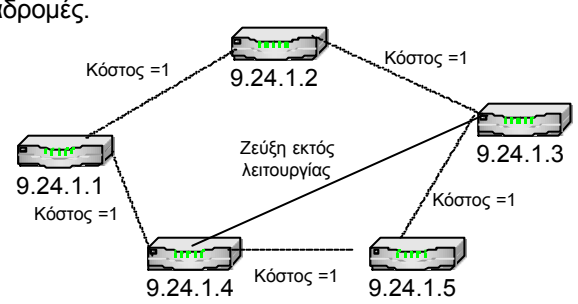
Το RIP παρόλο που είναι πολύ απλό πρωτόκολλο έχει σημαντικά μειονεκτήματα. Κατά αρχήν ο μέγιστος αριθμός δρομολογητών μέσω των οποίων μπορεί να διέλθει ένα μήνυμα είναι 15 και έτσι δεν είναι δυνατή η χρήση του πρωτοκόλλου σε μεγάλα δίκτυα. Επίσης δεν περιέχει πληροφορία διευθυνσιοδότησης υποδικτύων. ‘Ετσι, δεν είναι ικανό να διακρίνει αν μια διεύθυνση αντιστοιχεί σε υποδίκτυο ή σε τερματικό. Για την αντιμετώπιση των προβλημάτων αυτών παρουσιάστηκε μια νέα έκδοση του πρωτοκόλλου, το RIP Version2. Η νέα έκδοση δεν αλλάζει το πρωτόκολλο απλά παρέχει περισσότερη πληροφορία χρησιμοποιώντας τα πεδία που εμφανίζονται στο Σχήμα 2.2 που έχουν την τιμή 0. Το RIP-2 παρέχει πληροφορία για διευθυνσιοδότηση υποδικτύου καθώς και ένα πεδίο route Tag το οποίο διευκολύνει την επικοινωνία με πρωτόκολλα εξωτερικής δρομολόγησης. Παρόλο που τα παραπάνω προβλήματα αντιμετωπίστηκαν στην νέα έκδοση, ο αλγόριθμος που υλοποιείται στα RIP τα έχει κάνει να θεωρούνται πλέον ξεπερασμένα. Πιο συγκεκριμένα, σημαντικό μειονέκτημα είναι ο μεγάλος χρόνος που απαιτείται μέχρι να ισορροπήσει μετά την αστοχία ή την απενεργοποίηση μιας ζεύξης. Ακόμη, η χρήση μόνο του αριθμού των διαδοχικών δρομολογητών μέσω των οποίων διέρχεται το πακέτο, ως κριτήριο δρομολόγησης, δεν έχει αποδειχθεί σαν το αποτελεσματικότερο. Είναι δυνατόν, διαδρομή με λίγους κόμβους να αποτελείται από ζεύξεις χαμηλού ρυθμού διέλευσης και τα πακέτα να καθυστερούν περισσότερο απ’ ότι αν ακολουθήσουν διαδρομή με περισσότερους κόμβους. Τέλος, το πρωτόκολλο δεν προβλέπει εξισορρόπηση φορτίου (load balancing) σε περιπτώσεις ισοδύναμων διαδρομών αφού στον πίνακα υπάρχει μόνο μια εγγραφή προς κάθε προορισμό.

2.1.2.2 Open Shortest Path First (OSPF)

‘Οπως έχει ήδη αναφερθεί στην προηγούμενη παράγραφο τα πρωτόκολλα στατικής δρομολόγησης χρησιμοποιούνται κυρίως σε μικρά δίκτυα. Με την αύξηση του μεγέθους ενός δικτύου, μεγαλώνει και ο πίνακας

δρομολόγησης σε σημείο που κάποια στιγμή η δρομολόγηση σε πραγματικό χρόνο να καθίσταται αδύνατη. Για τον λόγο αυτό, στα μεγάλα δίκτυα δε χρησιμοποιούνται τέτοιου είδους πρωτόκολλα. 'Ενα πρωτόκολλο εσωτερικής δρομολόγησης για μεγάλα δίκτυα είναι το OSPF. Πρόκειται για πρωτόκολλο δυναμικής δρομολόγησης πράγμα που σημαίνει ότι ο πίνακας ενός δρομολογητή ενημερώνεται σε τακτά χρονικά διαστήματα, χωρίς παρέμβαση από τον διαχειριστή, μέσω πληροφοριών που καταφθάνουν από γειτονικούς δρομολογητές.

Ας δούμε την λειτουργία του πρωτοκόλλου με την χρήση του παρακάτω παραδείγματος στο Σχήμα 2.3. Το δίκτυο αποτελείται από πέντε δρομολογητές. Σε κάθε ζεύξη μεταξύ δυο καταχωρητών ανατίθεται κάποιο κόστος, το οποίο μπορεί να βασίζεται σε παράγοντες όπως η ικανότητα ρυθμού διέλευσης της ζεύξης, η αξιοπιστία κλπ. Τα κόστη κάθε ζεύξης γίνονται γνωστά μέσω του δικτύου σε όλους τους δρομολογητές. Αρχικά όλες οι ζεύξεις είναι ενεργές και έχουν κόστος 1 (ακόμη και η ζεύξη μεταξύ 9.24.1.3 και 9.24.1.4). 'Ετσι για το δεδομένο δίκτυο και τη δεδομένη χρονική στιγμή όλοι οι δρομολογητές έχουν την βάση δεδομένων (Σχήμα 2.3). Τα πακέτα δεδομένων που διέρχονται από τον 9.24.1.2 και πηγαίνουν στον 9.24.1.4, πηγαίνουν είτε μέσω του .1.3 είτε μέσω του .1.1 αφού και οι δύο διαδρομές έχουν το ίδιο κόστος. Στην πραγματικότητα, επειδή το πρωτόκολλο έχει και πρόβλεψη για εξισορρόπηση φορτίου (load balancing), το φορτίο θα διανεμηθεί ομοιόμορφα (κατά το δυνατόν) και στις δυο διαδρομές.



Σχήμα 2.3
Δρομολόγηση σε δίκτυο OSPF

Πίνακας 2.1
Κόστη Ζεύξεων

Δρομολογητής	9.94.1.1	9.94.1.2	9.94.1.3	9.94.1.4	9.94.1.5
9.94.1.1		1	2	1	2
9.94.1.2	1		1	2	2
9.94.1.3	2	1		1	1
9.94.1.4	1	2	1		1
9.94.1.5	2	2	1	1	

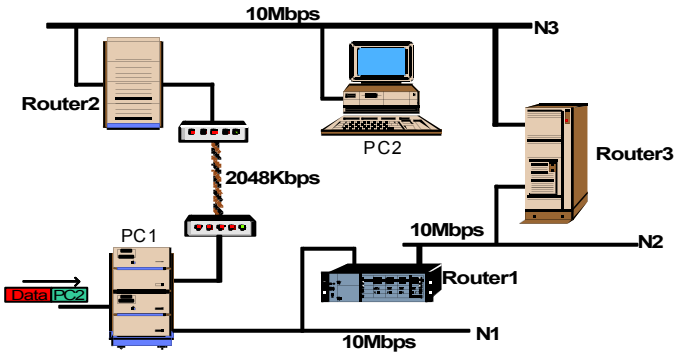
Στο OSPF κάθε δρομολογητής στέλνει προς τους γειτονικούς του το μήνυμα HELLO για να επισημάνει ότι είναι ενεργός. Επίσης μεταδίδει

πληροφορία στο δίκτυο σχετική με την κατάσταση των ζεύξεων που χρησιμοποιεί ανά 30 δευτερόλεπτα. ‘Εστω ότι το δίκτυο για κάποιο χρονικό διάστημα λειτουργεί κανονικά και κάποια χρονική στιγμή η ζεύξη μεταξύ των 9.24.91.3 και 9.24.1.4 τίθεται εκτός λειτουργίας. Οι δρομολογητές 9.24.91.3 και 9.24.1.4 αντιλαμβάνονται την δυσλειτουργία και μεταδίδουν στο δίκτυο την πληροφορία ότι η ζεύξη μεταξύ τους δεν είναι πλέον ενεργή. Οι υπόλοιποι δρομολογητές του δικτύου ενημερώνουν τους πίνακές τους μόλις λάβουν την πληροφορία και κατασκευάζεται ένας πίνακας αρκετά διαφορετικός από τον πίνακα 1. Θα υπάρχουν 2 διαδρομές από τον .1.2 στον .1.4 εκ των οποίων η μία θα έχει κόστος 2 (μεσώ του .1.1) και η άλλη κόστος 3 (μέσω των .1.3 και .1.5). Τώρα πλέον η ροή της πληροφορίας θα πραγματοποιείται μέσω της διαδρομής που εμφανίζει το χαμηλότερο κόστος.

Ο πιο συνηθισμένος τρόπος υπολογισμός κόστους είναι βασισμένος στο ρυθμό διέλευσης μιας ζεύξης. ‘Ενας τύπος υπολογισμού του κόστους είναι ο παρακάτω:

Κόστος=100.000.000/ρυθμος διέλευσης σε bits/sec.

Για παράδειγμα, το κόστος σε μία ζεύξη Ethernet είναι 10EXP(8)/10EXP(7)=10 ενώ σε μια γραμμή E1 είναι 10EXP(8)/2.048.000≅49. Με βάση το Σχήμα 2.4 μπορούμε να καταλάβουμε την διαφορά μεταξύ των πρωτοκόλλων RIP και OSPF όσον αφορά την λειτουργία τους.



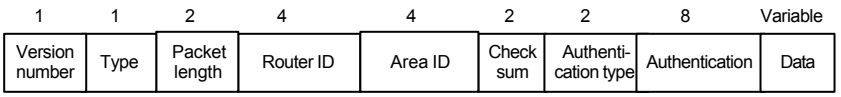
Σχήμα 2.4
Τοπολογία

‘Εστω ότι έρχεται στο PC1 πακέτο με τελικό προορισμό το PC2. Ας εξετάσουμε πως θα προωθούσαν το πακέτο αυτό τα δύο πρωτόκολλα εσωτερικής δρομολόγησης που έχουμε εξετάσει έως τώρα. Υπάρχουν δυο διαφορετικοί δρόμοι μέχρι τον τελικό προορισμό. Ο ένας περνάει μέσω των modems στον δρομολογητή Router 2 και προωθείται στο δίκτυο N3 απ’ όπου και φτάνει στον τελικό προορισμό. Ο δεύτερος δρόμος περνάει από το δίκτυο N1, τον δρομολογητή Router 1 τον Router 3 και φτάνει στο προωθείται στο δίκτυο N3. Στην περίπτωση του RIP η δρομολόγηση θα γινόταν μέσω του πρώτου δρόμου, αφού ως ενδιάμεσο σταθμό

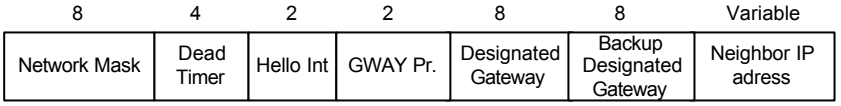
περιλαμβάνει μόνο τον δρομολογητή Router 2. Αντίθετα το OSPF θα επιλέξει τον δεύτερο δρόμο μέσω των δρομολογητών 1 και 3. Αυτό γιατί το συνολικό κόστος της πρώτης διαδρομής είναι 49+10=59, ενώ της δεύτερης είναι 10+10+10=30 σύμφωνα με παραπάνω τρόπο υπολογισμό του κόστους.

Ας δούμε όμως σε ποια περίπτωση το πακέτο φτάνει πιο γρήγορα στον τελικό προορισμό. Το τελικό μονοπάτι είναι το ίδιο και στις δυο περιπτώσεις αφού και στις δυο προωθείται μέσω του δικτύου N3. Θα υπολογίσουμε λοιπόν τον χρόνο που κάνει να φτάσει το πακέτο στο δίκτυο N3 και στις δυο περιπτώσεις. Αν χρησιμοποιηθεί το RIP, ο χρόνος μέχρι να φτάσει το πακέτο στον δίκτυο N3 είναι t_1=8*1.000/2.048.000 secs=3.90625msec. Στην περίπτωση του OSPF ο αντίστοιχος χρόνος είναι t_2=8*1.000/10.000.000secs=0.8msec για κάθε ζεύξη (εφ' όσον έχουν τον ίδιο ρυθμό διέλευσης), δηλαδή σύνολο 1.6 msec. Στο παράδειγμα έχουμε παραλείψει τον χρόνο επεξεργασίας και αναζήτησης του μονοπατιού δρομολόγησης στον κάθε δρομολογητή, ο οποίος είναι πολύ μικρότερος από την διαφορά t_2-t_1, οπότε δεν οδηγούμαστε σε λάθος συμπέρασμα.

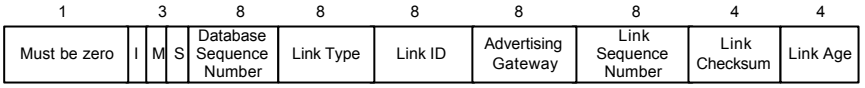
Ακόμη δεν έχουμε αναφερθεί στον τρόπο με τον οποίο η πληροφορία σχετικά με τα κόστη των διαφόρων ζεύξεων γίνεται γνωστή στους δρομολογητές ενός δικτύου OSPF. Αυτό επιτυγχάνεται μέσω της χρήσης του πρωτοκόλλου HELLO. Το HELLO εκτός από την μετάδοση πληροφορίας σχετικής με τις ζεύξεις συγχρονίζει και τα ρολόγια των συστημάτων που συμμετέχουν σε ένα δίκτυο OSPF. Πριν την αποστολή ενός πακέτου, ο δρομολογητής εισάγει σε ένα πεδίο την τρέχουσα ώρα αντιγράφοντας την από το τοπικό ρολόι. Το σύστημα που λαμβάνει το πακέτο αφαιρεί την τιμή αυτή από την τιμή που υπολογίζει ότι έχει ο αποστολέας την χρονική στιγμή της λήψης. Με τον τρόπο αυτό μπορεί να υπολογίσει την διάρκεια για την αποστολή του πακέτου μέσω της ζεύξης και συνεπώς τον ρυθμό διέλευσης. Μπορούμε να παρατηρήσουμε ότι το μήνυμα περιέχει πεδίο για ανίχνευση λαθών ενώ το RIP δεν περιείχε. Στην περίπτωση του RIP, τα μηνύματα χρησιμοποιούν το UDP, το οποίο περιέχει τέτοιο κώδικα για την πληροφορία που μεταφέρει. Το OSPF όμως χρησιμοποιεί απ' ευθείας το IP (με protocol identifier 89), το οποίο περιέχει τέτοιο κώδικα μόνο για την επικεφαλίδα του, ενώ η πληροφορία που μεταφέρει δεν είναι προστατευμένη από ενδεχόμενα λάθη.



Σχήμα 2.5
Επικεφαλίδα μηνύματος OSPF



Σχήμα 2.6
Μήνυμα HELLO



Σχήμα 2.7
Μήνυμα Database Description

Στο Σχήμα 2.5 παρουσιάζεται το γενικό μήνυμα OSPF. Τον πεδίο version περιέχει την τρέχουσα έκδοση του πρωτοκόλλου. Το type χρησιμοποιείται για την αναγνώριση του τύπου του μηνύματος που ακολουθεί. Υπάρχουν οι τύποι HELLO, Database Description, Link Status Request, Link Status Update και Link Status Acknowledgement. Το τελευταίο χρησιμοποιείται σαν αναγνώριση για τα μηνύματα Link Status Update. Το πεδίο message length δίνει το μήκος του μηνύματος ενώ στο router ID περιέχεται η διεύθυνση του αποστολέα του μηνύματος. Ακόμη στο area id περιέχεται ο αριθμός της περιοχής (περιγράφεται στην επόμενη παράγραφο). Το checksum παρέχει τρόπο να ελέγξουμε αν το μήνυμα έχει λάθη ενώ, τέλος, το authentication type και το authentication μας παρέχει τρόπο να διαχωρίσουμε τις εκπομπές των πιστοποιημένων δρομολογητών από ενδεχόμενες εκπομπές κάποιων κακόβουλων που θα προκαλούσαν προβλήματα στη διαδικασία δρομολόγησης.

- Το μήνυμα HELLO: Το μήνυμα αυτό αποστέλλεται περιοδικά για να διαπιστωθεί αν υπάρχει επικοινωνία μεταξύ των δρομολογητών σε ένα δίκτυο OSPF. Στο πεδίο type εισάγεται η τιμή 1 και έχει την μορφή του που παρουσιάζεται στο Σχήμα 2.6 Το πεδίο Network Mask περιέχει την μάσκα του δικτύου από το οποίο εστάλη το μήνυμα. Αν το χρονικό διάστημα που περιέχεται στο Dead Timer παρέλθει χωρίς να απαντήσει ο δρομολογητής στον οποίον απευθύνεται, τότε ο δρομολογητής αυτός δε θα ληφθεί υπό όψιν για την διαδικασία δρομολόγησης. Το επόμενο πεδίο δίνει το διάστημα που μεσολαβεί μεταξύ της αποστολής διαδοχικών μηνυμάτων του τύπου αυτού. Τα πεδία Designated Gateway και Backup Designated Gateway περιέχουν τις διευθύνσεις των προεπιλεγμένων δρομολογητών μιας περιοχής. Τέλος τα Neighbor1-n IP Address δίνουν τις διευθύνσεις όλων των δρομολογητών από τους οποίους έχει λάβει μήνυμα HELLO ο αποστολέας.
- Το μήνυμα Database Description: Χρησιμοποιείται για την ανταλλαγή πληροφοριών που περιέχονται στις βάσεις δεδομένων των δρομολογητών. Το περιεχόμενο των μηνυμάτων αυτών βοηθάει κάποιον δρομολογητή να καταλάβει την τοπολογία του δικτύου στο

οποίο βρίσκεται. Κατά την έναρξη της διαδικασίας ο ένας δρομολογητής (master) απαιτεί από κάποιον άλλον (slave) να του παράσχει πληροφορίες από την βάση δεδομένων του. Ο slave απαντάει με μήνυμα που φαίνεται στο Σχήμα 2.7. Επειδή η βάση που απαιτεί να του αποσταλεί μπορεί να είναι μεγάλη και να μην χωράει σε ένα μόνο μήνυμα χρησιμοποιούνται τα bits I, το οποίο έχει την τιμή 1 μόνο στο πρώτο μήνυμα και M, το οποίο είναι 1 αν ακολουθούν και άλλα μηνύματα που αφορούν την ίδια αίτηση. Το bit S διαχωρίζει τα μηνύματα του slave από του master. Τα πεδία από Link Type έως Link Age επαναλαμβάνονται για κάθε ζεύξη. Παρέχουν πληροφορία σχετικά με τον τύπο της ζεύξης (προς άλλη περιοχή, άλλο δίκτυο κ.λ.π.), τον χρόνο που έχει παρέλθει από την στιγμή που ενεργοποιήθηκε η ζεύξη, την ταυτότητα του δρομολογητή που παρέχει την πληροφορία αυτή όπως επίσης χρησιμεύουν για την ανίχνευση λαθών στο πακέτο.

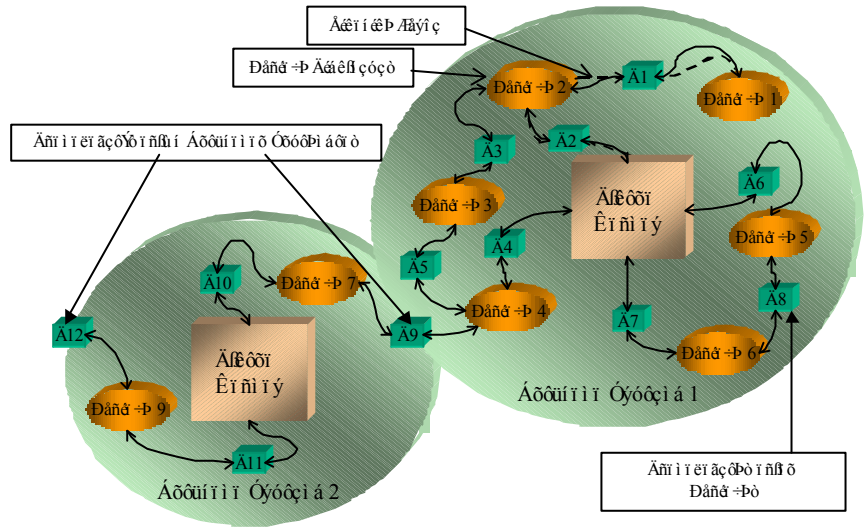
- Το μήνυμα Link Status Request: Μετά την απόκτηση πληροφορίας σχετικής με την τοπολογία ενός δικτύου, κάποιος δρομολογητής είναι πιθανόν να ανακαλύψει ότι μέρος της βάσης που αφορά κάποιες ζεύξεις του δεν έχει ενημερωθεί για μεγάλο διάστημα. Χρησιμοποιεί το μήνυμα αυτό για να απαιτήσει, από γειτονικό δρομολογητή, πληροφορία σχετική με τις ζεύξεις αυτές.
- Το μήνυμα Link Status Update: Σε περιοδικά χρονικά διαστήματα ή όταν η κατάσταση μιας ζεύξης αλλάξει οι δρομολογητές χρησιμοποιούν το μήνυμα αυτό για να ενημερώσουν τους γειτονικούς τους. Το μήνυμα είναι μεταβλητού μήκους και στην επικεφαλίδα του περιέχει το πλήθος των εγγραφών που περιέχονται σε αυτό. Η μορφή της κάθε εγγραφής είναι ίδια με την περίπτωση του μηνύματος Database Description και περιέχει πεδία που περιγράφουν αν η ζεύξη είναι προς άλλο δίκτυο, άλλη περιοχή κλπ.

Οργάνωση ενός δικτύου OSPF

Σε ένα δίκτυο που αποτελείται από K δρομολογητές το συνολικό πλήθος των μηνυμάτων HELLO που μεταδίδονται ανά τακτό χρονικό διάστημα θα είναι K^2 , εφ' όσον ο καθένας από αυτούς θα μεταδώσει μήνυμα προς όλους. Το γεγονός αυτό είναι δυνατόν να μειώσει το αξιοποιήσιμο για μετάδοση δεδομένων εύρος ζώνης σε χαμηλά επίπεδα. Η περίπτωση αυτή έχει προβλεφθεί από το OSPF και για τον λόγο αυτό σε κάθε φυσικό δίκτυο υπάρχει ένας προεπιλεγμένος δρομολογητής, στον οποίο όλοι οι άλλοι στέλνουν τα μηνύματά τους. Αυτός είναι και ο υπεύθυνος για τη μετάδοση των μηνυμάτων προς όλους τους άλλους δρομολογητές "εκπροσωπώντας" τους. Με τον τρόπο αυτό το συνολικό πλήθος των μεταδιδομένων μηνυμάτων μειώνεται στο 2K. 'Οπως έχει ήδη αναφερθεί, στο πρωτόκολλο αυτό ορίζεται ότι οι βάσεις μεταξύ δρομολογητών που ανταλλάσσουν πληροφορία κατάστασης ζεύξεων πρέπει να είναι συγχρονισμένες, δηλαδή να περιέχουν τα ίδια δεδομένα. 'Ετσι στην περίπτωση του OSPF όπου οι δρομολογητές ανταλλάσσουν πληροφορία μόνο με τον επιλεγμένο δρομολογητή, οι βάσεις των δρομολογητών

απαιτείται να είναι συγχρονισμένες μόνο με αυτόν και όχι και μεταξύ όλων των δρομολογητών του δικτύου. Επειδή υπάρχει η πιθανότητα δυσλειτουργίας του ή αστοχίας των ζεύξεων που τον συνδέουν με το υπόλοιπο δίκτυο υπάρχει και ο εφεδρικός προεπιλεγμένος δρομολογητής ο οποίος αναλαμβάνει τον ρόλο του μόλις διαπιστωθεί ότι η επικοινωνία με την πρωτεύοντα δεν είναι δυνατή. Αυτό γίνεται μέσω των περιοδικών μεταδόσεων των μηνυμάτων HELLO, όπως έχει ήδη αναφερθεί.

Ακόμη όμως και με την μείωση της διακινούμενης πληροφορίας δρομολόγησης είναι δυνατόν αν το δίκτυο είναι πολύ μεγάλο, να οδηγηθούμε σε τεράστιες βάσεις δεδομένων. Για να αντιμετωπιστεί το πρόβλημα της αύξησης στην κατανάλωση πόρων (μνήμη, επεξεργαστική ισχύς) με την αύξηση του μεγέθους ενός δικτύου, το πρωτόκολλο προβλέπει την διαίρεση ενός δικτύου σε περιοχές. Με τον τρόπο αυτό επιτυγχάνεται καλύτερη διαχείριση πόρων και αποδοτικότερη δρομολόγηση. Μια περιοχή είναι ένα σύνολο δικτύων μέσα σε ένα αυτόνομο σύστημα. Κάθε δρομολογητής διατηρεί μια βάση που περιέχει πληροφορίες σχετικές με την κατάσταση των ζεύξεων μεταξύ των δρομολογητών καθώς και την τοπολογία του δικτύου. Σε μια περιοχή όλοι οι δρομολογητές έχουν κοινή βάση δεδομένων (όχι κοινούς πίνακες δρομολόγησης).



Σχήμα 2.8
Οργάνωση Δικτύου OSPF

Με τον τρόπο αυτόν είναι ανά πάσα στιγμή γνωστή η κατάσταση όλων των ζεύξεων μέσα στην περιοχή και η λειτουργία δρομολόγησης μέσα στην περιοχή καθίσταται εύκολη. ‘Ολες οι περιοχές έχουν άμεση ή έμμεση πρόσβαση στο δίκτυο κορμού (backbone), το οποίο είναι υπεύθυνο για την μετάδοση της πληροφορίας δρομολόγησης σε όλες τις περιοχές του

αυτόνομου συστήματος. Κανονικά όλες οι περιοχές πρέπει να είναι συνδεδεμένες με το δίκτυο κορμού. Πολλές φορές όμως, λόγω της τοπολογίας του δικτύου αυτό δεν είναι δυνατό. Μια τέτοια περίπτωση είναι η περίπτωση της περιοχής 1 του Σχήματος 2.8 η οποία συνδέεται με το δίκτυο κορμού μέσω της περιοχής 2. Η σύνδεση η οποία πραγματοποιεί η περιοχή 1 με το δίκτυο κορμού ονομάζεται εικονική σύνδεση αφού διέρχεται μέσω άλλης περιοχής (περιοχή διέλευσης) και εμφανίζεται στο Σχήμα 2.8 με την διακεκομμένη γραμμή.

Στο πρωτόκολλο ορίζονται διάφοροι τύποι περιοχών ανάλογα με τα χαρακτηριστικά τους. Χρησιμοποιώντας το παράδειγμα του Σχήματος 2.8 μπορούμε να κατανοήσουμε τους διαφορετικούς τύπους περιοχών που ορίζει το πρωτόκολλο. Η περιοχή 1 ονομάζεται απομονωμένη περιοχή επειδή συνδέεται με τις υπόλοιπες μέσω ενός μόνου σημείου εξόδου. Οι περιοχές αυτές ενώ μπορούν να εγκαταστήσουν εικονικές ζεύξεις, δεν είναι δυνατόν να διέρχονται από αυτές εικονικές ζεύξεις άλλων περιοχών. Με λίγα λόγια μια απομονωμένη περιοχή δεν είναι ποτέ δυνατόν να είναι και περιοχή διέλευσης.

Το RFC 2328 ορίζει επίσης και διαφορετικούς τύπους δρομολογητών. Σε ένα OSPF δίκτυο υπάρχουν δρομολογητές που είναι υπεύθυνοι για την δρομολόγηση μόνο εντός της περιοχής, δρομολογητές που ενώνουν δυο περιοχές και ονομάζονται δρομολογητές ορίου περιοχής, και οι δρομολογητές ορίου αυτόνομου συστήματος που ενώνουν το αυτόνομο σύστημα με κάποιο άλλα. Ο διαχωρισμός στους τύπους αυτούς έγινε για να μειωθεί ο όγκος της διακινούμενης πληροφορίας καθώς και των πινάκων που διατηρούν οι δρομολογητές σε κάθε περιοχή. Πιο συγκεκριμένα ο πρώτος τύπος δρομολογητή διατηρεί στα αποθηκευτικά του μέσα πληροφορία που είναι σχετική μόνο με την περιοχή στην οποία ανήκει. Οι δρομολογητές ορίου περιοχής συνδέουν δυο περιοχές μεταξύ τους και διατηρούν βάση που περιέχει την τοπολογία όλων των περιοχών στις οποίες συνδέονται ανταλλάσσοντας πληροφορία με τους δρομολογητές όλων των περιοχών αυτών. Τέλος, οι δρομολογητές ορίων αυτόνομου συστήματος, χρησιμοποιούν πρωτόκολλα εξωτερικής δρομολόγησης για να επικοινωνούν με δρομολογητές άλλων αυτόνομων συστημάτων. Είναι υπεύθυνοι να μεταδίδουν στην περιοχή στην οποία ανήκουν πληροφορία σχετική με εξωτερικές συνδέσεις, αλλά και για την μετάδοση σε εξωτερικά δίκτυα πληροφορίας σχετικής με την κατάσταση των ζεύξεων του αυτόνομου συστήματος τους.

Τέλος πρέπει να σημειωθεί ότι το πρωτόκολλο υποστηρίζει και την δρομολόγηση διαφορετικών τύπων υπηρεσίας σε διαφορετικά μονοπάτια.

2.1.3 Πρωτόκολλα Εξωτερικής Δρομολόγησης

Τα πρωτόκολλα εξωτερικής δρομολόγησης επιτρέπουν την επικοινωνία μεταξύ δρομολογητών που βρίσκονται σε διαφορετικά αυτόνομα συστήματα. Στο κεφάλαιο αυτό παρουσιάζεται το πρωτόκολλο BGP το οποίο είναι και το πλέον διαδεδομένο.

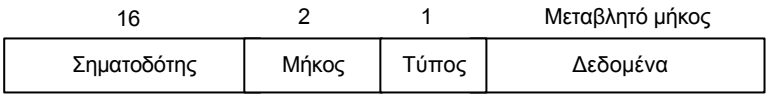
2.1.3.1 Border Gateway Protocol (BGP)

Το BGP επιτρέπει την ανταλλαγή πληροφορίας μεταξύ δρομολογητών που βρίσκονται στο ίδιο ή σε διαφορετικά αυτόνομα συστήματα. Υπάρχουν οι εξής περιπτώσεις ανταλλαγής πληροφορίας:

- Δρομολόγηση μεταξύ διαφορετικών αυτόνομων συστημάτων: Συμβαίνει μεταξύ δρομολογητών που βρίσκονται σε διαφορετικά αυτόνομα συστήματα. ‘Ολοι οι δρομολογητές που συμμετέχουν σε ανταλλαγή πληροφοριών δρομολόγησης του τύπου αυτού πρέπει να βρίσκονται στο ίδιο φυσικό δίκτυο.
- Δρομολόγηση εντός ενός αυτόνομου συστήματος: Συμβαίνει μεταξύ δρομολογητών οι οποίοι βρίσκονται στο ίδιο αυτόνομο σύστημα. Σκοπός είναι η ενημέρωση όλων των δρομολογητών σχετικά με την τοπολογία του δικτύου αλλά και ο καθορισμός κάποιου δρομολογητή ως συνδετικό σημείο προς άλλα αυτόνομα συστήματα.
- Δρομολόγηση Διέλευσης (Pass Through): Συμβαίνει κατά την ανταλλαγή πληροφορίας μεταξύ δρομολογητών που συνδέονται με το πρωτόκολλο BGP μέσω ενός άλλου αυτόνομου συστήματος το οποίο δε συμμετέχει στο πρωτόκολλο αυτό.

Η πρωταρχική αρμοδιότητα ενός συστήματος BGP είναι η ανταλλαγή πληροφορίας σχετικής με την τοπολογία και την ύπαρξη μονοπατιών προς διάφορα δίκτυα. Κάθε δρομολογητής διατηρεί έναν πίνακα στην μνήμη του, που περιέχει όλα τα πιθανά μονοπάτια προς ένα συγκεκριμένο προορισμό. Σε αντίθεση με τα πρωτόκολλα εσωτερικής δρομολόγησης δεν έχουμε περιοδική ενημέρωση όλου του πίνακα αλλά μόνο ενημέρωση σε περιπτώσεις μεταβολής της τοπολογίας του δικτύου. Κάθε δρομολογητής διατηρεί στην μνήμη του την πληροφορία για όλα τα μονοπάτια έως ότου κάποια πληροφορία μεταβολής για ένα τουλάχιστον από αυτά έρθει από κάποιον άλλο δρομολογητή. Η μόνη περίπτωση ανταλλαγής πληροφορίας ολόκληρου πίνακα είναι κατά την είσοδο κάποιου δρομολογητή στο σύστημα. Σε όλες τις άλλες περιπτώσεις η μόνη πληροφορία που ανταλλάσσεται είναι πληροφορία μεταβολής των μονοπατιών.

Οι δρομολογητές δεν στέλνουν πληροφορία σε τακτά χρονικά διαστήματα παρά μόνο όταν σημειωθεί κάποια μεταβολή. Ο λόγος για τον οποίο γίνεται αυτό είναι η μείωση του όγκου δεδομένων δρομολόγησης μέσα στο δίκτυο. Επίσης η πληροφορία που ανταλλάσσεται δεν είναι λεπτομερής. Κάθε δρομολογητής μεταδίδει μόνο το βέλτιστο μονοπάτι προς κάποιον προορισμό χρησιμοποιώντας ένα μόνο μέτρο για την αξιολόγηση του από τους ομότιμους δρομολογητές και τον βαθμό επιλογής της κάθε ζεύξης. Ο βαθμός αυτός εξαρτάται από μία πληθώρα παραμέτρων όπως ο αριθμός των αυτόνομων συστημάτων μέσω των οποίων διέρχεται το μονοπάτι, η σταθερότητα της ζεύξης, η ταχύτητα και η καθυστέρηση μετάδοσης. Οι τιμές των διαφόρων παραμέτρων τίθενται από τον διαχειριστή του συστήματος.



Σχήμα 2.9
Η επικεφαλίδα του BGP

Οι δρομολογητές πριν από την οποιαδήποτε ανταλλαγή πληροφοριών εγκαθιστούν μεταξύ τους σύνδεση χρησιμοποιώντας το πρωτόκολλο TCP. Ο λόγος χρήσης του TCP είναι η αύξηση της αξιοπιστίας των μεταδόσεων εφ' όσον οι πληροφορίες διανύουν μεγάλες αποστάσεις και διέρχονται από πολλά αυτόνομα συστήματα. 'Ολα τα πακέτα που μεταδίδονται χρησιμοποιούν το βασικό τύπο πακέτου το οποίο αποτελείται από την επικεφαλίδα (όπως εμφανίζεται στο Σχήμα 2.9) και το πεδίο δεδομένων. Η επικεφαλίδα αποτελείται από τα εξής τέσσερα πεδία

- Σηματοδότης: Περιέχει μια τιμή πιστοποίησης την οποία μπορεί να υπολογίσει ο δέκτης για να πιστοποιήσει την αυθεντικότητα και την ορθότητα του πακέτου. Επίσης χρησιμοποιείται για την ανίχνευση της απώλειας συγχρονισμού μεταξύ δρομολογητών που συμμετέχουν στο σύστημα. Στην περίπτωση μηνύματος OPEN το πεδίο αυτό δεν έχει χρησιμότητα και όλα τα bits τίθενται σε '1'.
- Μήκος: καθορίζει το μήκος του πακέτου σε bytes περιλαμβάνοντας την επικεφαλίδα. Η τιμή του πρέπεί να είναι τουλάχιστον 19 και όχι μεγαλύτερη από 4096.
- Τύπος : καθορίζει έναν από τους εξής τέσσερις διαφορετικούς τύπους πακέτων:
- Open: Εγκατάσταση σύνδεσης μεταξύ δρομολογητών ή την απόσυρση ενός ή περισσότερων.
- Update: Χρησιμοποιείται για την εγκατάσταση ενός μονοπατιού ή την απόσυρση ενός ή περισσότερων.
- Notification: Αποστέλλεται αμέσως μετά την ανίχνευση λάθους από κάποιον δρομολογητή. Η σύνδεση στην οποία αναφέρεται κλείνει αμέσως μετά την αποστολή του μηνύματος αυτού.
- Keep Alive: Ανταλλάσσονται μεταξύ δρομολογητών για να επιβεβαιώσουν ότι η σύνδεση μεταξύ τους είναι υπαρκτή.
- Δεδομένα: Περιέχει πληροφορία δρομολόγησης.

Βάσεις πληροφορίας δρομολόγησης (RIBs)
Κάθε δρομολογητής διατηρεί τρεις βάσεις δεδομένων στα αποθηκευτικά του μέσα: την Adj-RIB-In, την Loc-RIB και την Adj-RIB-Out. Στην Adj-RIB-In περιέχεται πληροφορία που καταφθάνει από άλλους δρομολογητές του συστήματος και προέρχεται από μηνύματα τύπου UPDATE. Η πληροφορία αυτή θα ληφθεί υπ' όψιν από τον αλγόριθμο επιλογής του βέλτιστου μονοπατιού για εισαγωγή στον πίνακα δρομολόγησης. Στην Loc-RIB αποθηκεύεται πληροφορία που θα χρησιμοποιηθεί μόνο τοπικά από τον ίδιο τον δρομολογητή. Τα δεδομένα που εισάγονται σε αυτήν την βάση καθορίζονται από την πολιτική δρομολόγησης που ακολουθεί ο

δρομολογητής και προέρχονται από επεξεργασία της πληροφορίας που περιέχεται στην βάση Adj-RIB-In.

Τέλος, στην Adj-RIB-Out βρίσκεται η πληροφορία την οποία έχει επιλέξει ο δρομολογητής για μετάδοση, προκειμένου να ενημερωθούν και άλλοι δρομολογητές του συστήματος για την τοπολογία του συστήματος. Μεταφέρεται με την αποστολή μηνυμάτων τύπου UPDATE.

Τύποι μηνυμάτων

Στο BGP υπάρχουν τέσσερις τύποι μηνυμάτων όπως έχει ήδη αναφερθεί. 'Ολα χρησιμοποιούν την ίδια επικεφαλίδα που παρουσιάστηκε στο Σχήμα 2.9 με μόνη διαφορά στο πεδίο τύπος μηνύματος. Ενώ λοιπόν η επικεφαλίδα παραμένει η ίδια αυτό που αλλάζει είναι το πεδίο δεδομένων. Στις εικόνες που ακολουθούν παρουσιάζεται μόνο το πεδίο δεδομένων του κάθε μηνύματος. Το πλήρες πακέτο κάθε τύπου, αποτελείται από την επικεφαλίδα που παρουσιάστηκε παραπάνω και το αντίστοιχο πεδίο δεδομένων.

Α. Το μήνυμα OPEN

Μετά την επιτυχή εγκατάσταση συνδέσεως TCP το πρώτο μήνυμα που μεταδίδεται είναι το μήνυμα OPEN. Αν το μήνυμα αυτό επιβεβαιωθεί, μπορούν να ακολουθήσουν και οι άλλοι τύποι μηνύματος. Σε περίπτωσης λάθους, η σύνδεση εγκαταλείπεται και γίνονται καινούργιες προσπάθειες εγκατάστασης της. Το Σχήμα 2.10 δείχνει τα πεδία από τα οποία αποτελείται το μήνυμα.

Σχήμα 2.10
Το μήνυμα OPEN

Η έκδοση δείχνει την τρέχουσα έκδοση του πρωτοκόλλου, το αυτόνομο σύστημα δείχνει τον αριθμό του αυτόνομου συστήματος του αποστολέα και ο χρόνος συγκράτησης καθορίζει το μέγιστο χρονικό διάστημα μεταξύ διαδοχικών μηνυμάτων KEEP ALIVE ή UPDATE σε δευτερόλεπτα. Η τιμή του πρέπει να είναι 0 ή μεγαλύτερη από 3. Στην περίπτωση ενός δικτύου που δεν παρουσιάζει μεταβολές είναι πιθανόν να περάσει μεγάλο χρονικό διάστημα χωρίς την αποστολή δεδομένων δρομολόγησης. Για τον λόγο αυτό σε τακτά χρονικά διαστήματα οι δρομολογητές, που έχουν εγκαταστήσει σύνδεση μεταξύ τους, αποστέλλουν μήνυμα τύπου KEEP ALIVE για την επιβεβαίωση της ύπαρξης της σύνδεσης μεταξύ τους. Η ταυτότητα BGP χρησιμοποιείται για την αναγνώριση του αποστολέα. Επειδή σε ένα αυτόνομο σύστημα είναι πιθανόν να υπάρχουν περισσότεροι του ενός δρομολογητές που χρησιμοποιούν το BGP, το πεδίο αυτόνομο σύστημα δεν είναι επαρκές για την αναγνώριση του αποστολέα. Στο πεδίο αυτό εισάγεται η IP διεύθυνση του αποστολέα. Το μήκος προαιρετικών παραμέτρων δείχνει το μήκος του πεδίου που ακολουθεί σε bytes. Είναι δυνατό το πεδίο αυτό να έχει την τιμή 0 και να μην υπάρχει

επόμενο πεδίο. Τέλος, το πεδίο προαιρετικών παραμέτρων περιέχει μια λίστα παραμέτρων που αποτελείται από την τριπλέτα Τύπος, Μήκος και Τιμή.

Β. Το μήνυμα UPDATE
Το μήνυμα UPDATE χρησιμοποιείται για την μετάδοση πληροφορίας δρομολόγησης μεταξύ δρομολογητών. Η πληροφορία που περιέχεται μπορεί να χρησιμοποιηθεί για την κατασκευή του γράφου του συστήματος αποφεύγοντας έτσι την περίπτωση δρομολόγησης σε βρόχους (loop routing). Μπορεί να μεταφέρει πληροφορία για την εισαγωγή ενός μόνο καινούργιου μονοπατιού ή την απόσυρση περισσότερων του ενός μονοπατιών. Είναι ακόμη δυνατός ο συνδυασμός των δυο παραπάνω περιπτώσεων. Στο Σχήμα 2.11 παρουσιάζεται το πακέτο αυτού του τύπου.

Μήκος του επόμενου πεδίου	Αποσυρόμενα μονοπάτια	Μήκος του επόμενου πεδίου	Ιδιότητες Μονοπατιών	Προσεγγισιμότητα δικτύου

Σχήμα 2.11
Το μήνυμα UPDATE

Το μήκος του πεδίου Αποσυρόμενα μονοπάτια έχει μήκος 2 bytes και δίνει το μήκος του επόμενου πεδίου. Αν έχει τιμή 0 τότε όλα τα μονοπάτια είναι ακόμη ενεργά. Το πεδίο Αποσυρόμενα μονοπάτια έχει μεταβλητό μήκος και περιέχει την λίστα με τις IP διευθύνσεις όλων των μονοπατιών που δεν είναι πλέον ενεργά. Το Μήκος του πεδίου 'ιδιότητες μονοπατιών' έχει μήκος 2 bytes και δίνει το μήκος του επόμενου πεδίου σε bytes. Το πεδίο Ιδιότητες μονοπατιών είναι μεταβλητού μήκους και περιγράφει την ιδιότητα κάθε μονοπατιού. Αποτελείται από την τριπλέτα τύπος, μήκος, τιμή. Υπάρχουν διάφοροι τύποι ιδιοτήτων άλλες υποχρεωτικές σύμφωνα με το πρωτόκολλο και άλλες προαιρετικές που δεν είναι απαραίτητο να υποστηρίζονται από όλες τις εκδόσεις του πρωτοκόλλου. Το πρώτο bit του πεδίου τύπος καθορίζει αν η ιδιότητα που ακολουθεί είναι υποχρεωτική ή προαιρετική. Στην περίπτωση των προαιρετικών παραμέτρων ο δρομολογητής που λαμβάνει το μήνυμα είτε την αγνοεί (αν δεν την υποστηρίζει) στην περίπτωση που προορίζεται γι' αυτόν είτε την προωθεί χωρίς να επέμβει πάνω της αν το πακέτο προορίζεται για άλλον δρομολογητή. Ιδιότητες που καθορίζει το πρωτόκολλο και οι οποίες είναι απαραίτητες για την κατανόηση του είναι η εξής:

- ORIGIN: καθορίζει την προέλευση του μηνύματος. Αν δηλαδή αυτό έφτασε από κάποιον δρομολογητή που βρίσκεται εντός του αυτόνομου συστήματος ή από κάποιον εξωτερικό.
- AS_PATH: αποτελείται από μια ακολουθία εγγραφών και μας πληροφορεί από ποια αυτόνομα συστήματα έχει περάσει το μήνυμα. Κάθε δρομολογητής που λαμβάνει το μήνυμα αυτό πραγματοποιεί τις εξής λειτουργίες. Αν το μήνυμα προορίζεται για κάποιον δρομολογητή που βρίσκεται στο ίδιο αυτόνομο σύστημα τότε το πεδίο αυτό μεταδίδεται ως έχει χωρίς καμία μεταβολή. Στην περίπτωση όμως που προορίζεται για δρομολογητή που βρίσκεται σε άλλο αυτόνομο

σύστημα τότε ο δρομολογητής εισάγει τον αριθμό του αυτόνομου συστήματος του ως τελευταία εγγραφή στο πεδίο. Με τον τρόπο αυτό ο δρομολογητής που λαμβάνει το μήνυμα μπορεί να γνωρίζει όλη την διαδρομή την οποία ακολούθησε το πακέτο αυτό. Στην περίπτωση που ένας δρομολογητής δημιουργεί αυτό το μήνυμα προς αποστολή, τότε στο πεδίο αυτό εισάγεται το νούμερο του αυτόνομου συστήματος μόνο αν προορίζεται για δρομολογητή που βρίσκεται εκτός του αυτόνομου συστήματος.

- NEXT_HOP: περιέχει τη διεύθυνση IP του δρομολογητή που πρέπει να χρησιμοποιηθεί ως επόμενος σταθμός μέχρι τον τελικό προορισμό.
- MULTI_EXIT_DISC: χρησιμοποιείται στην περίπτωση ύπαρξης περισσότερων της μια συνδέσεων προς ένα συγκεκριμένο εξωτερικό αυτόνομο σύστημα και καθορίζει το ποιος δρομολογητής θα χρησιμοποιείται για την προώθηση πακέτων προς το σύστημα αυτό. Η τιμή που περιέχεται στο πεδίο αποτελεί το μέτρο (κόστος) για την σύνδεση αυτή. Ο δρομολογητής που έχει την μικρότερη τιμή στο πεδίο αυτό αποτελεί και την έξοδο του συστήματος προς το εξωτερικό αυτόνομο σύστημα.
- LOCAL_PREF: χρησιμοποιείται κατά την επικοινωνία με άλλους δρομολογητές του ίδιου αυτόνομου συστήματος. Μεταδίδει τον βαθμό προτίμησης ενός μονοπατιού προς εξωτερικά δίκτυα.

Γ. Το μήνυμα KEEP ALIVE

Το μήνυμα KEEP ALIVE χρησιμοποιεί μόνο τον βασικό τύπο πακέτου όπως εμφανίζεται στο Σχήμα 2.8, χωρίς πρόσθετες παραμέτρους. Χρησιμοποιείται για να επιβεβαιώσει ότι η σύνδεση μεταξύ δυο δρομολογητών είναι ενεργή σε περιπτώσεις όπου αυτοί δεν ανταλλάσσουν μηνύματα τύπου UPDATE. Η συχνότητα μετάδοσης του μηνύματος τίθεται στο 1/3 της τιμής του πεδίου χρόνου συγκράτησης, ώστε ακόμη και αν χαθεί ένα πακέτο του τύπου να υπάρχει χρόνος για την μετάδοση άλλου χωρίς να πέσει η σύνδεση. Το πρωτόκολλο καθορίζει ότι το ελάχιστο χρονικό διάστημα μεταξύ μετάδοσης διαδοχικών μηνυμάτων του τύπου αυτού είναι 1 δευτερόλεπτο. Για τον λόγο αυτό η τιμή του πεδίου Χρόνος συγκράτησης πρέπει πάντοτε να είναι μεγαλύτερη από 3.

Δ. Το μήνυμα NOTIFICATION

Το μήνυμα αυτό αποστέλλεται στην περίπτωση ανίχνευσης λάθους από κάποιον δρομολογητή. Περιέχει τον τύπο του λάθους ακολουθούμενο από περιγραφή του λάθους που ανιχνεύθηκε. Πιθανοί τύποι λαθών είναι λάθος στο μήκος του μηνύματος (υπέρβαση μήκους ή μήκος μικρότερο από 19 bytes), μη υποστηριζόμενη έκδοση του πρωτοκόλλου, λάθος πιστοποίησης, λήξη του χρόνου συγκράτησης κλπ. Αμέσως μετά την αποστολή του μηνύματος, ο δρομολογητής που ανίχνευσε το λάθος κλείνει την σύνδεση με τον δρομολογητή από τον οποίο προήλθε το λάθος.

Λειτουργία του BGP

Η περιγραφή της λειτουργίας του πρωτοκόλλου BGP μπορεί να παρασταθεί με την ακολουθία λειτουργιών μιας μηχανής πεπερασμένων καταστάσεων. Η κατάσταση αυτή αναφέρεται σε κάθε σύνδεση, όμως ο δρομολογητής για κάθε σύνδεση με κάποιον ομότιμό του 'τρέχει' και από ένα στιγμιότυπο της μηχανής αυτής. Η αρχική κατάσταση είναι η ανενεργή κατάσταση, όπου δεν υπάρχει σύνδεση με έναν συγκεκριμένο δρομολογητή. Με την λήψη ενός μηνύματος εγκατάστασης σύνδεσης ο δρομολογητής ξεκινάει έναν χρονιστή και προχωράει την κατάσταση 'αναμονή σύνδεσης'. Σε περίπτωση ανίχνευσης λάθους ή αν παρέλθει το διάστημα που καθορίζεται από τον χρονιστή τότε η προσπάθεια σύνδεσης διακόπτεται και η μηχανή γυρνάει στην ανενεργό κατάσταση. Για τον λόγο αυτό η τιμή που εισάγεται στον χρονιστή πρέπει να είναι επαρκής, ώστε να είναι δυνατή η εγκατάσταση σύνδεσης TCP, πριν την ανταλλαγή μηνυμάτων του πρωτοκόλλου. Αν η σύνδεση αρχικοποιηθεί με επιτυχία, ο δρομολογητής στέλνει μήνυμα OPEN και προχωράει στην κατάσταση 'αποστολή OPEN', περιμένοντας αντίστοιχο μήνυμα από τον άλλο δρομολογητή. Αν το μήνυμα που θα ληφθεί είναι σωστό τότε ο δρομολογητής στέλνει μήνυμα KEEP ALIVE, ως επιβεβαίωση σωστής λήψης. Στην αντίθετη περίπτωση στέλνει μήνυμα NOTIFICATION και η σύνδεση κλείνει. Είναι δυνατόν να σταλεί μήνυμα του τύπου αυτού ακόμη και αν δεν υπάρχει λάθος. Η περίπτωση αυτή είναι η ταυτόχρονη εγκατάσταση σύνδεσης (κατάσταση σύγκρουσης) και από του δύο δρομολογητές. Αν ανιχνευθεί η κατάσταση σύγκρουσης τότε η μία από τις 2 συνδέσεις πρέπει να κλείσει, οπότε αποστέλλεται μήνυμα NOTIFICATION.

Μετά την επιτυχή εγκατάσταση της σύνδεσης αποστέλλεται ολόκληρος ο πίνακας δρομολόγησης και οι δυο δρομολογητές σε τακτά χρονικά διαστήματα αποστέλλουν μηνύματα KEEP ALIVE με σκοπό τη διατήρηση της σύνδεσης.

Κατά διαστήματα καταφθάνουν μηνύματα τύπου UPDATE με πληροφορία μεταβολής της κατάστασης ορισμένων μονοπατιών. Η πληροφορία αυτή εισάγεται στην βάση Adj-RIB-In και εκτελείται η διαδικασία ενημέρωσης του τοπικού πίνακα δρομολόγησης. Στην περίπτωση απόσυρσης κάποιου μονοπατιού αυτό αφαιρείται από την βάση και ενημερώνεται ο πίνακας. Αν φτάσει πληροφορία προσθήκης νέου μονοπατιού, τότε αν το κόστος προσεγγισιμότητας του δικτύου είναι το ίδιο με αυτό άλλου μονοπατιού που υπάρχει ήδη στη βάση, τότε το νέο μονοπάτι αντικαθιστά το παλιό. Υπάρχει επίσης η περίπτωση η πληροφορία που αφορά το νέο μονοπάτι να μην είναι πληροφορία προς συγκεκριμένο δρομολογητή ή τερματικό, αλλά να περιέχει πληροφορία προσεγγισιμότητας υποδικτύου. Σε αυτήν την περίπτωση αν δυο εγγραφές, μία προς συγκεκριμένο σύστημα και μια προς το υποδίκτυο που περιλαμβάνει το σύστημα έχουν το ίδιο κόστος τότε θα προτιμηθεί η πληροφορία υποδικτύου. Με τον τρόπο αυτό ο δρομολογητής μπορεί να περιέχει πληροφορία δρομολόγησης γα περισσότερα του ενός τερματικά έχοντας μόνο μια εγγραφή στον πίνακά του.

Α. Διαδικασία αξιολόγησης μονοπατιών
Η διαδικασία αυτή επιλέγει το ποια πληροφορία θα μεταδοθεί στο δίκτυο εφαρμόζοντας την πολιτική δρομολόγησης του BGP πάνω στις εγγραφές που περιέχονται στην βάση Adj-RIB-In. Τα αποτελέσματα που προκύπτουν από την διαδικασία αποθηκεύονται στην βάση Adj-RIB-Out. Πραγματοποιείται σε τρία διαδοχικά βήματα:

1. Στην πρώτη φάση, υπολογίζεται ο βαθμός προτίμησης κάθε μονοπατιού. Εκτελείται κάθε φορά που λαμβάνεται μήνυμα τύπου UPDATE το οποίο περιέχει πληροφορία σχετική με νέο μονοπάτι. Κατά την φάση αυτή δεσμεύεται η βάση Adj-RIB-In και δεν επιτρέπεται η εισαγωγή νέων εγγραφών μέχρι να ολοκληρωθεί. Για κάθε νέο μονοπάτι ή μεταβολή για κάποιο υπάρχον, ο δρομολογητής υπολογίζει το βαθμό προτίμησης του.
2. Στην δεύτερη φάση, γίνεται η επιλογή του καλύτερου μονοπατιού προς συγκεκριμένο προορισμό και εκτελείται μόνο αν δεν τρέχει την συγκεκριμένη στιγμή η τρίτη φάση κάποιας προηγούμενης ενημέρωσης. Κατά την φάση αυτή επιλέγεται το μονοπάτι που έχει τον μεγαλύτερο βαθμό προτίμησης (μικρότερο κόστος) προς έναν συγκεκριμένο προορισμό σε περίπτωση ύπαρξης περισσότερων του ενός μονοπατιού. Αν υπάρχει ένα μόνο μονοπάτι τότε αυτό επιλέγεται κατευθείαν. Ο δρομολογητής κατόπιν εισάγει το επιλεχθέν μονοπάτι στην βάση Local-RIB. Σε περίπτωση που υπάρχουν πολλά μονοπάτια τα οποία όλα περνάνε από δρομολογητή του ίδιου αυτόνομου συστήματος, επιλέγεται αυτό με την μικρότερη τιμή MULTI_EXIT_DISC.
3. Η τρίτη φάση, που εκτελείται μετά το πέρας της δεύτερης, είναι υπεύθυνη για την ενημέρωση των άλλων δρομολογητών του συστήματος και ενεργοποιείται όταν έχουν αλλάξει οι πληροφορίες δρομολόγησης ή όταν ένας νέος δρομολογητής εισέλθει στο δίκτυο. Απαραίτητη προϋπόθεση για να ενεργοποιηθεί είναι το να μην τρέχει την συγκεκριμένη χρονική στιγμή η δεύτερη φάση κάποιας αξιολόγησης. Μετά το τέλος της φάσης αυτής ο δρομολογητής ενημερώνει τους γειτονικούς του σε περίπτωση ύπαρξης μεταβολών. Αν η πληροφορία που οδήγησε στην μεταβολή αυτή προήλθε από δρομολογητή του ίδιου αυτόνομου συστήματος τότε η μεταβολή δεν διαδίδεται στο δίκτυο. Στην αντίθετη περίπτωση μεταδίδεται.

Β. Περιορισμός του όγκου πληροφορίας δρομολόγησης
Το BGP περιορίζει την διακινούμενη πληροφορία δρομολόγησης στο δίκτυο για να αυξηθεί το διαθέσιμο για μετάδοσης δεδομένων εύρος ζώνης. Για τον λόγο αυτό καθορίζει ένα ελάχιστο χρονικό διάστημα μεταξύ διαδοχικών ενημερώσεων για κάθε σύνδεση BGP. Σε αυτόν τον περιορισμό υπάρχουν δύο εξαιρέσεις. Επειδή απαιτείται ταχεία σύγκλιση του πρωτοκόλλου σε ένα αυτόνομο σύστημα, πληροφορίες που προέρχονται από δρομολογητές του ίδιου αυτόνομου συστήματος επιτρέπεται να υπερβούν τον περιορισμό αυτό. Επίσης στην περίπτωση απόσυρσης ενός μονοπατιού μπορούν να υπερβούν τον περιορισμό αυτό, για να μην προωθούνται πακέτα που δεν πρόκειται να φτάσουν τον

προορισμό τους. Ο περιορισμός αυτός δεν επεκτείνεται και στην επιλογή των βέλτιστων μονοπατιών. Αυτό σημαίνει ότι στο χρονικό διάστημα μεταξύ δυο διαδοχικών μεταδόσεων, ο δρομολογητής μπορεί να αλλάξει το βέλτιστο μονοπάτι όσες φορές θέλει βασιζόμενος σε πληροφορίες που καταφθάνουν από άλλους και να μεταδώσει την εγγραφή που έχει κατά την χρονική στιγμή της παρόδου του διαστήματος αυτού.

Εκτός από την συχνότητα μετάδοσης μηνυμάτων το πρωτόκολλο υποστηρίζει και ένα είδος συμπίεσης των δεδομένων που διακινούνται. Είναι δυνατόν στην περίπτωση διέλευσης μηνύματος UPDATE από ένα αυτόνομο σύστημα να μην προστίθεται η IP διεύθυνση όλων των δρομολογητών από τους οποίους διέρχεται το μήνυμα παρά μόνο ο αριθμός του αυτόνομου συστήματος. 'Ετσι κατά την διαδικασία δρομολόγησης των δεδομένων, το μήνυμα προωθείται σε ένα αυτόνομο σύστημα χωρίς να χρειαστεί να περάσει από συγκεκριμένη ακολουθία δρομολογητών, και το σύστημα αυτό αναλαμβάνει την δρομολόγηση του όσο το μήνυμα βρίσκεται εντός του συστήματος.

2.2 Από άκρο σε άκρο αποφυγή συμφόρησης (End to end Congestion avoidance)

2.2.1 Εισαγωγή

Το Πρωτόκολλο Ελέγχου Μετάδοσης (Transmission Control Protocol, TCP) είναι το πρωτόκολλο στρώματος μεταφοράς (transport layer) στο διαδίκτυο που παρέχει αξιόπιστη μεταφορά δεδομένων. Η αξιοπιστία εξασφαλίζεται χάρη σε κάποια χαρακτηριστικά που το διαφοροποιούν από το UDP, το οποίο δεν παρέχει εγγυήσεις για την αξιόπιστη μεταφορά των δεδομένων. Τα χαρακτηριστικά αυτά είναι:

- Τα δεδομένα προς μεταφορά αντιμετωπίζονται ως ένα ρεύμα byte. Η διεργασία – παραλήπτης λαμβάνει τα δεδομένα με την ίδια σειρά που τα μεταδίδει η διεργασία αποστολέας. Το TCP αριθμεί τα δεδομένα που μεταδίδονται χρησιμοποιώντας το πεδίο Sequence Number της επικεφαλίδας του TCP. Η παράδοση των δεδομένων στη διεργασία – παραλήπτη γίνεται κατά αύξοντα Sequence Number ακόμα και αν τα δεδομένα δεν μεταφέρονται στο δίκτυο με αυτή τη σειρά.
- Χρήση επιβεβαιώσεων (acknowledgements) για τα δεδομένα που μεταδίδονται. Ο αποδέκτης των δεδομένων στέλνει στον αποστολέα πακέτα (segments) που επιβεβαιώνουν την ορθή λήψη των δεδομένων. Δεδομένα για τα οποία δεν έχει ληφθεί από τον αποστολέα επιβεβαίωση αναμεταδίδονται. 'Ετσι το TCP εγγυάται την ορθή μεταφορά των δεδομένων.
- 'Ελεγχος Ροής. Με τη χρήση των επιβεβαιώσεων και ενός μηχανισμού κυλιόμενου παραθύρου ο παραλήπτης των δεδομένων είναι σε θέση να ελέγξει το ροή των δεδομένων. Συγκεκριμένα ο παραλήπτης ορίζει το μέγεθος του παραθύρου (σε bytes) το οποίο χρησιμοποιείται από

τον αποστολέα ως άνω όριο για τα δεδομένα που μπορεί να στείλει χωρίς να λάβει επιβεβαίωση. Κατά τον τρόπο αυτό εξασφαλίζεται ότι ο αποστολέας δεν στέλνει δεδομένα ταχύτερα από όσο μπορεί να τα απορροφήσει ο παραλήπτης και ότι δεν θα εξαντληθεί ο χώρος ενταμίευσης.

- Μεταφορά δεδομένων με σύνδεση. Πριν ξεκινήσει η μεταφορά των δεδομένων, ο αποστολέας και ο παραλήπτης εγκαθιστούν μια σύνδεση (TCP connection) η οποία καταργείται με την ολοκλήρωση της μεταφοράς (ή αν η επικοινωνία καθίσταται αδύνατη). Η εγκατάσταση της σύνδεσης αποτελεί προϋπόθεση για την υλοποίηση των παραπάνω μηχανισμών.
- Ο ‘Ελεγχος Συμφόρησης (Congestion Control) είναι έννοια παρεμφερής με τον ‘Ελεγχο Ροής και στηρίζεται στον ίδιο μηχανισμό κυλιόμενου παραθύρου. Υπάρχει ωστόσο μια θεμελιώδης διαφορά: ο έλεγχος ροής αποσκοπεί στην προσαρμογή της ροής των δεδομένων στο ρυθμό που αυτά μπορούν να απορροφηθούν από τη διεργασία – παραλήπτη. Ο έλεγχος συμφόρησης από την άλλη πλευρά αποβλέπει στην ομαλή ροή των δεδομένων στο δίκτυο και την αποφυγή καταστάσεων συμφόρησης του δικτύου.

Στη συνέχεια αυτής της ενότητας παρουσιάζονται οι μηχανισμοί που οφείλουν οι υλοποιήσεις του TCP να χρησιμοποιούν για τον έλεγχο συμφόρησης, όπως ορίζονται στο RFC 2581.

2.2.2 Μηχανισμός Συρρόμενου Παραθύρου

Το TCP χρησιμοποιεί την τεχνική του συρρόμενου παραθύρου (sliding window technique) για να προσφέρει αξιόπιστες υπηρεσίες. Δηλαδή επιτρέπει σε κάθε χρήστη του δικτύου να μεταδώσει ένα αριθμό πακέτων ίσο με το μέγεθος του παραθύρου (window size) πριν αρχίσει να περιμένει για τις επιβεβαιώσεις των πακέτων που έστειλε. Με αυτήν την τεχνική, ελέγχεται αν τα πακέτα έφθασαν στο σωστό προορισμό, αν έφθασαν σωστά και αν δημιουργήθηκαν προβλήματα στα πακέτα που μεταδίδονται. Επίσης, γίνεται σωστή χρήση του δικτύου αντίθετα με την περίπτωση που ένα πακέτο μεταδιδόταν και στην συνέχεια έπρεπε να περιμένει να έρθει η επιβεβαίωση πριν μεταδοθεί το επόμενο.

Πέρα όμως από το πρόβλημα της σωστής μεταφοράς δεδομένων και της ικανοποιητικής χρησιμοποίησης του δικτύου, το TCP με το μηχανισμό του συρρόμενου παραθύρου λύνει άλλο ένα πρόβλημα: τον από άκρη σε άκρη έλεγχο ροής της πληροφορίας (end to end flow control). Με αυτό τον τρόπο δίνει την δυνατότητα στον προορισμό να περιορίσει τον ρυθμό μετάδοσης μέχρι να αδειάσουν οι ενταμιευτές του. Ο μηχανισμός αυτός του TCP είναι κάπως πιο περίπλοκος. Κατά αρχήν αναφέρουμε ότι ο δέκτης των δεδομένων διατηρεί ένα ίδιο παράθυρο προκειμένου να τοποθετεί τα δεδομένα με την ίδια σειρά και να τα παραδίδει μόλις ολοκληρωθεί η λήψη τους στην κατάλληλη εφαρμογή. Επίσης σε κάθε πακέτο επιβεβαίωσης ο δέκτης βάζει μια ένδειξη διαθέσιμου χώρου που υπάρχει στους

καταχωρητές του. Με αυτόν τον τρόπο ο πομπός ενημερώνεται και έτσι μπορεί να μεταβάλει το μέγεθος του δικού του παραθύρου.

Το πλεονέκτημα του μεταβλητού μήκους παραθύρου είναι αυτό που επιτυγχάνει τον σωστό έλεγχο ροής από άκρη σε άκρη. Στέλνοντας ένα μηδενικό διαθέσιμο χώρο ο δέκτης μπορεί να διακόψει όλες τις μεταδόσεις έτσι ώστε το δίκτυο ή η σύνδεση να μπορέσει να επανέλθει μετά από μια άσχημη κατάσταση.

Το TCP εγγυάται λοιπόν μια αξιόπιστη μεταφορά και επαναμεταδίδει κάθε πακέτο εάν μια επιβεβαίωση (ACK) δεν έχει ληφθεί σε μια συγκεκριμένη χρονική περίοδο. Το TCP θέτει αυτή τη χρονική περίοδο (timeout) σαν μια συνάρτηση του round-trip-time (RTT) που περιμένει μεταξύ των δυο άκρων της σύνδεσης. Δυστυχώς, δεδομένου του εύρους των πιθανών RTT μεταξύ οποιουδήποτε ζεύγους hosts στο Διαδίκτυο, καθώς και στην μεταβολή στον χρόνο του RTT μεταξύ των ίδιων hosts, η επιλογή μιας κατάλληλης τιμής για την χρονική περίοδο (timeout) δεν είναι τόσο εύκολη.

Για να διευθετήσει αυτό το πρόβλημα, το TCP χρησιμοποιεί έναν προσαρμοζόμενο μηχανισμό επαναμετάδοσης (adaptive retransmission mechanism). Κάθε φορά που το TCP έχει μια καινούρια μέτρηση για το χρόνο καθυστέρησης (sample round trip time), υπολογίζει μια νέα εκτιμώμενη τιμή για το RTT με βάση την νέα τιμή και την παλιά εκτίμηση. Στη συνέχεια ο χρόνος αναμονής (timeout) υπολογίζεται σαν μια συνάρτηση του χρόνου καθυστέρησης. Προβλήματα εμφανίζονται ως προς το τι θεωρείται ακριβής μέτρηση του RTT.

Το πρόβλημα λύθηκε με τον αλγόριθμο του Karn, σύμφωνα με τον οποίο όταν υπολογίζεται μια εκτίμηση του RTT αρκούν τα πακέτα που δεν επαναμεταδόθηκαν. Με τη μέθοδο αυτή η αρχική εκτίμηση του RTT επαναπροσδιορίζεται μέχρι να επιτευχθεί η μετάδοση του πακέτου. Μόλις ένα πακέτο μεταδοθεί σωστά με την πρώτη προσπάθεια, η εκτίμηση του RTT γίνεται ξανά από την αρχή. Για την αποφυγή καταστάσεων συμφόρησης, το TCP προτείνει τέσσερις μηχανισμούς που πλαισιώνουν τον μηχανισμό του συρρόμενου παραθύρου: Αργή ‘Εναρξη (Slow Start), Αποφυγή Συμφόρησης (Congestion avoidance), Ταχεία Επανεκπομπή (Fast Retransmit) και Ταχεία Ανάρρωση (Fast Recovery).

Πριν εξηγηθούν αναλυτικά οι μηχανισμοί, θα δώσουμε μερικούς βασικούς ορισμούς εννοιών σχετικά με το μηχανισμούς:

Segment: Οποιοδήποτε πακέτο IP μεταφέρει δεδομένα TCP ή / και επιβεβαιώσεις.

Sender Maximum Segment Size (SMSS): Μέγιστο μήκος ενός segment (σε bytes, χωρίς να περιλαμβάνει τις επικεφαλίδες TCP/IP) που μπορεί να μεταδώσει ο αποστολέας. Συνήθως βασίζεται στο MTU του δικτύου στο οποίο συνδέεται ο αποστολέας.

Receiver Maximum Segment Size (RMSS): Μέγιστο μήκος ενός segment που ο παραλήπτης μπορεί να δεχτεί. Το στέλνει ο παραλήπτης μέσω του MSS TCP option κατά την εγκατάσταση της σύνδεσης TCP. Αν δε σταλεί τιμή MSS χρησιμοποιείται η default τιμή των 536 bytes. Η τιμή του RMSS χρησιμοποιείται ως άνω φράγμα για το SMSS.

Full-sized Segment: ‘Ενα segment μήκους SMSS bytes.

Receiver Window (rwnd): Το πιο πρόσφατο μήκος παραθύρου που ανακοίνωσε ο παραλήπτης στον αποστολέα.

Congestion Window (cwnd): Το μήκος του παραθύρου που χρησιμοποιεί ο αποστολέας για τη μετάδοση των δεδομένων. Πρέπει πάντα cwnd <= rwnd.

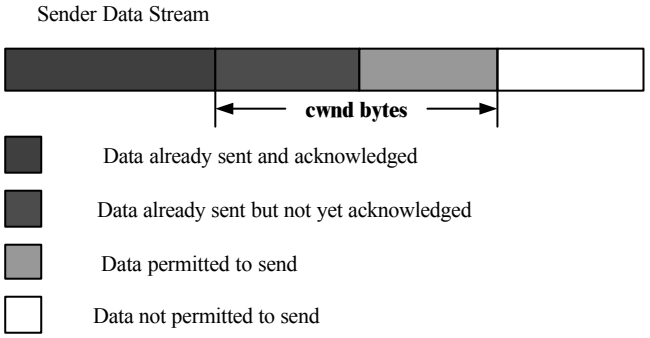
Initial Window (IW): Η αρχική τιμή του cwnd (αμέσως μετά την εγκατάσταση της σύνδεσης).

Loss Window (LW): Το cwnd γίνεται ίσο με LW όταν ο αποστολέας αντιληφθεί απώλεια πακέτου με τη λήξη του retransmission timer.

Restart Window (RW): Το cwnd γίνεται ίσο με RW όταν ο αποστολέας ξεκινά τη μετάδοση μετά από περίοδο σιωπής.

Flight Size: Το πλήθος των δεδομένων (bytes) που έχουν μεταδοθεί αλλά δεν έχει (ακόμα) επιβεβαιωθεί η λήψη τους.

Στο Σχήμα 2.12 περιγράφεται ο μηχανισμός του Κυλιόμενου Παραθύρου (Congestion Window - cwnd) στο TCP. Ο αποστολέας μεταδίδει segments και παράλληλα περιμένει επιβεβαιώσεις για τα segments που έχει ήδη μεταδώσει. Το μήκος του παραθύρου ορίζεται ως το πλήθος των bytes που επιτρέπεται να μεταδώσει ο αποστολέας μετά το τελευταίο byte για το οποίο έλαβε επιβεβαίωση.



Σχήμα 2.12
TCP Congestion Window

Το αριστερό όριο του παραθύρου είναι ο αριθμός (Sequence Number) του τελευταίου byte για το οποίο έχει ληφθεί επιβεβαίωση, ενώ το δεξί όριο του παραθύρου ορίζει το μέγιστο Sequence Number που μπορεί να μεταδώσει ο αποστολέας πριν λάβει νέα επιβεβαίωση. Λαμβάνοντας μια επιβεβαίωση τα όρια του παραθύρου μετατοπίζονται δεξιά κατά το πλήθος των bytes που επιβεβαιώθηκαν. Επιπλέον κάθε επιβεβαίωση που λαμβάνεται (όπως θα περιγραφεί αναλυτικά παρακάτω) προκαλεί αύξηση του μήκους του παραθύρου. Το παράθυρο συρρικνώνεται όταν το ζητήσει ο παραλήπτης (θέτοντας το rwnd σε κάποια μικρή τιμή, μικρότερη από το τρέχον cwnd, π.χ. στο 0) ή όταν ο αποστολέας αντιληφθεί απώλεια πακέτων.

Ο αποστολέας δικαιούται να μεταδώσει ένα segment αν το Sequence Number του είναι μικρότερο από το δεξί άκρο του cwnd. Αν το SN του επόμενου προς μετάδοση segment έχει φτάσει στο όριο του παραθύρου, ο αποστολέας θα πρέπει να περιμένει μέχρι να λάβει νέα επιβεβαίωση για να το μεταδώσει.

2.2.2.1 Αργή 'Εναρξη και Αποφυγή Συμφόρησης

Οι αλγόριθμοι Αργής 'Εναρξης (Slow Start) και Αποφυγής Συμφόρησης (Congestion Avoidance) πρέπει να χρησιμοποιούνται από τον αποστολέα TCP ώστε να ελέγχουν τον όγκο των δεδομένων που εισάγονται στο δίκτυο. Για την υλοποίηση των δυο αυτών αλγορίθμων χρειάζεται επιπλέον το παράθυρο συμφόρησης του αποστολέα (rwnd), που καθορίζει την ποσότητα των δεδομένων που ο αποστολέας μπορεί να μεταδώσει στο δίκτυο χωρίς επιβεβαίωση σωστής λήψης (ACK) και το παράθυρο του δέκτη, που θέτει ένα όριο για την ποσότητα των δεδομένων που περιμένουν χωρίς να έχουν επιβεβαιωθεί. Για την τελική αποστολή των δεδομένων επιλέγεται το μικρότερο από τα δυο παράθυρα.

Μια επιπλέον μεταβλητή που χρειάζεται είναι ένα κατώφλι αργής έναρξης, το ssthresh (Slow Start Threshold), που χρησιμοποιείται για να καθορίσει το εάν θα χρησιμοποιηθεί ο αλγόριθμος Slow Start ή ο Congestion Avoidance.

Η έναρξη της μετάδοσης δεδομένων σε ένα δίκτυο με άγνωστες συνθήκες απαιτεί από το TCP να εξετάσει προσεκτικά το δίκτυο ώστε να καθορίσει τη διαθέσιμη χωρητικότητά του, με σκοπό την αποφυγή συμφόρησης. Για αυτό το σκοπό ο αλγόριθμος Slow Start χρησιμοποιείται όταν ο όγκος δεδομένων αυξάνεται μετά από μια κατάσταση συμφόρησης.

Ο αποστολέας συγκρίνει την τρέχουσα τιμή του cwnd με τη μεταβλητή ssthresh (Slow Start Threshold) και με βάση το αποτέλεσμα της σύγκρισης επιλέγει μεταξύ Slow Start και Congestion Avoidance. Αν cwnd < ssthresh τότε επιλέγεται ο Slow Start ενώ αν cwnd > ssthresh επιλέγεται ο Congestion Avoidance (σε περίπτωση ισότητας μπορεί να επιλεγεί οποιοσδήποτε από τους 2). Η αρχική τιμή του ssthresh μπορεί να είναι οποιαδήποτε (συνήθως αρχικοποιείται στο rwnd). Η αρχική τιμή του cwnd είναι φυσικά IW. Στο RFC 2581 ορίζεται ότι το IW πρέπει να είναι το πολύ ίσο με 2*SMSS bytes (και σε κάθε περίπτωση όχι πάνω από 2 segments).

Κατά τη διάρκεια του Slow Start ο αποστολέας αυξάνει το cwnd κατά SMSS bytes (το πολύ) για κάθε νέα επιβεβαίωση (ACK segment) που λαμβάνει. Αυτός είναι και ο ταχύτερος ρυθμός αύξησης του cwnd. Αν δεν παρατηρηθεί απώλεια πακέτου το cwnd διπλασιάζεται σε κάθε RTT (Round Trip Time).

Κατά τη διάρκεια του Congestion Avoidance το cwnd αυξάνεται κατά SMSS ανά RTT. Για να επιτευχθεί αυτός ο ρυθμός το cwnd αυξάνεται κατά SMSS*SMSS/cwnd για κάθε νέο ACK segment που λαμβάνει ο αποστολέας. Εναλλακτικά ο αποστολέας μπορεί να μετράει το σύνολο των bytes που επιβεβαιώνονται από τα ACKs που λαμβάνει και να αυξήσει το cwnd κατά SMSS όταν το σύνολο αυτό γίνει ίσο με cwnd.

Με τους αλγορίθμους αυτούς επιδιώκεται η σταδιακή διεκδίκηση του διαθέσιμου εύρους ζώνης κατά την έναρξη της σύνδεσης. ‘Οταν το διαθέσιμο εύρος ζώνης έχει προσεγγιστεί τότε ο ρυθμός αύξησης του cwnd μειώνεται ώστε να αποφευχθεί πιθανή συμφόρηση στο δίκτυο.

2.2.2.2 Ταχεία Επανεκπομπή και Ταχεία Ανάρρωση

Η τιμή του cwnd ελαττώνεται όταν ο αποστολέας αντιληφθεί ότι στο δίκτυο υπάρχει συμφόρηση ή όταν το ζητήσει ο παραλήπτης ανακοινώνοντας νέο rwnd (μικρότερο από το τρέχον cwnd). Ο αποστολέας θεωρεί ότι επικρατεί συμφόρηση στο δίκτυο όταν αντιληφθεί απώλεια πακέτων. Η απώλεια πακέτου γίνεται αντιληπτή αν δεν ληφθεί επιβεβαίωση για το συγκεκριμένο πακέτο μέσα σε ορισμένο χρονικό διάστημα (ίσο με RTO – Retransmission Time Out), ή αν ληφθούν περισσότερα του ενός όμοια ACK segments.

Η περίπτωση του timeout υποδηλώνει σοβαρό πρόβλημα συμφόρησης καθώς έχει σταματήσει η ροή των segments. Η περίπτωση των πολλαπλών ομοίων επιβεβαιώσεων υποδηλώνει ελαφρύτερη μορφή συμφόρησης: έχει χαθεί κάποιο segment όμως άλλα segments (μεταγενέστερα του «χαμένου») φτάνουν στον παραλήπτη και προκαλούν τα πολλαπλά όμοια ACKs. Για το λόγο αυτό ο αποστολέας χειρίζεται τις 2 αυτές περιπτώσεις διαφορετικά. Στο RFC 2581 ορίζεται ότι αν συμβεί Retransmission Timeout, το cwnd γίνεται ίσο με LW (Loss Window, ίσο με SMSS). Επιπλέον ο αποστολέας πρέπει να θέσει το ssthresh σε τιμή όχι μεγαλύτερη από:

$$\text{ssthresh} = \max\ (\frac{\text{FlightSize}}{2}, 2 \cdot \text{SMSS})$$

‘Οταν ο αποστολέας λάβει 4 όμοια ACK segments (χωρίς να μεσολαβεί μεταξύ τους άλλο segment) θέτει σε λειτουργία τον αλγόριθμο Ταχείας Επανεκπομπής (Fast Retransmit). Ο αλγόριθμος αυτός προβλέπει ότι με τη λήψη του 4^{ου} όμοιου ACK ο αποστολέας θα αναμεταδώσει το ζητούμενο segment χωρίς να περιμένει τη λήξη του Retransmission Timer. Μετά την αναμετάδοση του «χαμένου» segment ο αποστολέας ενεργοποιεί τον αλγόριθμο Ταχείας Ανάρρωσης (Fast Recovery) μέχρις ότου λάβει ένα νέο (διαφορετικό) ACK segment. Ο συνδυασμός των 2 αυτών αλγορίθμων αποτελείται από τα παρακάτω βήματα:

1. Με τη λήψη του 4^{ου} όμοιου ACK το ssthresh τίθεται σε τιμή όχι μεγαλύτερη από όσο ορίζει η εξίσωση (1).
2. Αναμεταδίδεται το «χαμένο» segment και το cwnd γίνεται ίσο με ssthresh + 3*SMSS. Τα 3 επιπλέον segments στη νέα τιμή του παραθύρου αντιστοιχούν στα 3 «διπλά» ACKs που προκάλεσαν την αναμετάδοση.
3. Για κάθε «διπλό» ACK που λαμβάνεται το cwnd αυξάνεται κατά SMSS.
4. Αν η τιμή του cwnd το επιτρέπει, μεταδίδεται ένα segment.

5. ‘Οταν ληφθεί νέα επιβεβαίωση (στην πράξη η επιβεβαίωση για το «χαμένο» segment) το cwnd γίνεται ίσο με ssthresh και η φάση του Fast Recovery τελειώνει.

Ο συνδυασμός των 4 αλγορίθμων (Slow Start, Congestion Avoidance, Fast Retransmit και Fast Recovery) αποτελεί το minimum των απαιτήσεων που οφείλει μια υλοποίηση του TCP να υποστηρίζει σχετικά με τον έλεγχο συμφόρησης και είναι γνωστός ως TCP Reno.

2.2.2.3 Παραγωγή Επιβεβαιώσεων

Η συμμετοχή του παραλήπτη στη διαδικασία του ελέγχου συμφόρησης περιορίζεται στη μετάδοση των επιβεβαιώσεων, η λήψη (ή η έλλειψη) των οποίων χρησιμοποιείται από τον αποστολέα για τη ρύθμιση του cwnd.

Για την παραγωγή των επιβεβαιώσεων χρησιμοποιείται συνήθως (και συνιστάται στο RFC 2581) ο αλγόριθμος Καθυστερημένης Επιβεβαίωσης (Delayed ACK). Σύμφωνα με αυτόν ο παραλήπτης δεν παράγει 1 επιβεβαίωση για κάθε segment που λαμβάνει, αλλά καθυστερεί την παραγωγή τους με σκοπό να συνδυάσει πιθανά data και ACK στο ίδιο segment (τεχνική γνωστή ως piggyback) και γενικότερα να μειώσει τη χρησιμοποίηση του δικτύου. Η ακριβής λειτουργία του αλγορίθμου έχει ως εξής:

- Επιβεβαιώσεις πρέπει να παράγονται τουλάχιστον για κάθε $2^{ο}$ segment πλήρους μεγέθους (full sized) που λαμβάνεται και μέσα σε 500ms από τη λήψη του πρώτου segment που δεν έχει επιβεβαιωθεί. (Ο παραλήπτης μπορεί να θεωρήσει ως μέγεθος ενός full sized segment είτε το RMSS είτε το default των 536 bytes).
- Εναλλακτικά ο παραλήπτης μπορεί να επιβεβαιώνει κάθε δεύτερο segment, ανεξάρτητα από το μέγεθός του.
- Segments εκτός σειράς (δηλαδή με Sequence Number μεγαλύτερο από το αναμενόμενο) θα πρέπει να επιβεβαιώνονται αμέσως (ώστε να επιταχύνεται η έναρξη του Fast Retransmit). Φυσικά σε αυτή την περίπτωση η επιβεβαίωση που παράγεται είναι αντίγραφο της προηγούμενης που στάλθηκε.
- Segments που γεμίζουν ένα κενό στη ροή πακέτων που λαμβάνει ο παραλήπτης πρέπει να επιβεβαιώνονται αμέσως.

Ο παραλήπτης δεν πρέπει να παράγει περισσότερες από 1 επιβεβαιώσεις για το ίδιο segment εκτός από την περίπτωση που ανακοινώνει νέα τιμή για το rwnd.

2.2.3 Τροποποιήσεις και Επεκτάσεις στο TCP

Στην ενότητα αυτή παρουσιάζεται μια σειρά από τροποποιήσεις που έχουν προταθεί για τη βελτίωση της απόδοσης του TCP. Η υποστήριξη των μηχανισμών αυτών (αν και κάποιοι αποτελούν IETF standards) δεν αποτελεί υποχρέωση για μια υλοποίηση του TCP. Σε κάθε περίπτωση πάντως η χρήση μηχανισμών διαφορετικών από αυτούς που περιγράφονται στο RFC 2581 θα πρέπει να γίνεται με κριτήριο το ρυθμό

που μεταδίδει ο αποστολέας. Πιο συγκεκριμένα, αν ένα τροποποιημένο TCP πρόκειται να χρησιμοποιηθεί ευρέως στο Internet ο αποστολέας θα πρέπει να αυξάνει το cwnd χρησιμοποιώντας υποχρεωτικά παραλλαγές των αλγορίθμων Slow Start και Congestion Avoidance, που δεν θα αυξάνουν το παράθυρο ταχύτερα από όσο ορίζει το RFC 2581. Ως προς τους αλγορίθμους αποκατάστασης απώλειας πακέτων υπάρχει μεγαλύτερη ελευθερία, αλλά και πάλι το παράθυρο κατά την αναμετάδοση δεν πρέπει να ξεπερνά τα όρια του RFC 2581.

Εκτός από την τήρηση των περιορισμών του RFC 2581 τροποποιώντας κανείς το TCP θα πρέπει να λαμβάνει υπόψη του και την έκταση των απαιτούμενων αλλαγών. Κάποιοι μηχανισμοί υλοποιούνται με τροποποίηση μόνο του αποστολέα ή του παραλήπτη και επομένως είναι πιο απλή η αξιοποίησή τους. ‘Αλλοι μηχανισμοί προϋποθέτουν αλλαγές και στα 2 άκρα και συνεπώς η υιοθέτησή τους δεν εξασφαλίζει το προσδοκώμενο κέρδος σε κάθε περίπτωση. Ωστόσο σημαντικοί τέτοιου είδους μηχανισμοί (όπως αυτός των Επιλεκτικών Επιβεβαιώσεων, Selective Acknowledgements) αν και δεν αποτελούν υποχρεωτικό συστατικό μιας υλοποίησης TCP, έχουν υιοθετηθεί ως πρότυπα από την IETF και υποστηρίζονται από τα περισσότερα σύγχρονα λειτουργικά συστήματα.

2.2.3.1 Βελτιώσεις στην Αργή ‘Εναρξη

Σε TCP συνδέσεις που μεταφέρουν λίγα δεδομένα σε σχέση με το γινόμενο Bandwidth*Delay ο αλγόριθμος Αργής ‘Εναρξης (Slow Start) αποτελεί περιοριστικό της απόδοσης παράγοντα. Αιτία γι’ αυτό είναι το γεγονός ότι το cwnd πρέπει να φτάσει σε υψηλή τιμή και μέχρι να γίνει αυτό (με το Slow Start) απαιτείται χρόνος κάποιων RTT. Αν τα δεδομένα προς μεταφορά είναι λίγα τότε ο χρόνος αυτός είναι σημαντικό ποσοστό της συνολικής διάρκειας της σύνδεσης. ‘Ετσι αντί η σύνδεση να αξιοποιήσει το εύρος ζώνης που της αναλογεί, αναλώνεται στη διαδικασία να το αναζητήσει.

Μεγαλύτερο αρχικό παράθυρο (IW)

Η χρήση μεγαλύτερης τιμής αρχικού παραθύρου από όσο ορίζεται στο RFC 2581 (το πολύ 2*SMSS) επιταχύνει την αύξηση του παραθύρου κατά την Αργή Εναρξη καθώς ελαττώνει τη διάρκειά του κατά αρκετά RTT’s. Για πολύ σύντομες συνδέσεις είναι δυνατό όλα τα δεδομένα να μεταφερθούν μέσα στο 1^{ο} παράθυρο (σε 1 RTT).

Βέβαια τιμή πάνω από 2*SMSS δεν επιτρέπεται για το IW από το RFC 2581, και ο λόγος γι’ αυτό είναι ότι το μεγάλο IW προκαλεί εκρηκτική μετάδοση δεδομένων (bursty transmission) που είναι δυνατό να οδηγήσει το δίκτυο σε κατάσταση συμφόρησης (ιδιαίτερα αν μεταδίδουν έτσι πολλοί κόμβοι). Στο RFC 2414 προτείνεται πειραματικά η χρήση αρχικού παραθύρου μέχρι 4 segments. Η ακριβής τιμή φαίνεται στην πιο κάτω εξίσωση.

$$IW = \min(4*SMSS, \max(2*SMSS, 4380bytes))$$

Χρήση των Καθυστερημένων Επιβεβαιώσεων μετά την Αργή 'Εναρξη
Η χρήση των delayed ACKs (παραγωγή επιβεβαιώσεων για κάθε 2^o segment ή μέσα σε 500ms) έχει ως αποτέλεσμα ο αποστολέας να λαμβάνει λιγότερα ACKs από τον αριθμό των segments που στέλνει. 'Ετσι και η αύξηση του παραθύρου γίνεται με χαμηλότερο ρυθμό γιατί βασίζεται στον αριθμό των ACK segments που λαμβάνει ο αποστολέας. Η φάση της Αργής 'Εναρξης θα μπορούσε να επιταχυνθεί αν ο παραλήπτης δεν χρησιμοποιούσε delayed ACKs στη φάση του Slow Start. Ωστόσο ο παραλήπτης δεν γνωρίζει αν ο αποστολέας βρίσκεται σε Αργή 'Εναρξη ή όχι. Αυτό θα μπορούσε να το μάθει αν ο αποστολέας υποδηλώνει τη χρήση ή όχι του Slow Start σε κάποιο πεδίο της επικεφαλίδας του TCP ή με κάποια επιλογή TCP (option).

Μέτρηση Bytes
'Οπως προαναφέρθηκε η χρήση των Καθυστερημένων Επιβεβαιώσεων (delayed ACKs) έχει ως αποτέλεσμα τη λήψη λιγότερων επιβεβαιώσεων από τον αποστολέα και την πιο αργή αύξηση του παραθύρου κατά τη διάρκεια της Αργής 'Εναρξης. Προκειμένου να επιταχυνθεί η αύξηση του παραθύρου κατά τη φάση του Slow Start ο αποστολέας αντί να μετράει το πλήθος των επιβεβαιώσεων που λαμβάνει (και να αυξάνει το cwnd κατά SMSS ανά ACK) θα μπορούσε να μετρά τα bytes που επιβεβαιώνονται από κάθε ACK και να αυξάνει κατά αυτή την ποσότητα το cwnd. Η μέθοδος αυτή αναφέρεται ως Απεριόριστη Μέτρηση Bytes (Unlimited Byte Counting, UBC) και είναι δυνατό να οδηγήσει σε ιδιαίτερα εκρηκτική μετάδοση (ειδικά κατά τη φάση αποκατάστασης απώλειας πακέτων όπου ένα ACK segment είναι δυνατό να επιβεβαιώνει πολύ περισσότερα από 1-2 segments). Μια παραλλαγή της μεθόδου αναφέρεται ως Περιορισμένη Μέτρηση Bytes (Limited Byte Counting) και βασίζει την αύξηση του παραθύρου στα bytes που επιβεβαιώνει το κάθε ACK, θέτοντας όμως ένα άνω φράγμα ίσο με 2*SMSS. Κατά τον τρόπο αυτό περιορίζονται οι εκρήξεις του UBC.

Εκτίμηση Ρυθμού Μετάδοσης
Η χρήση της Αργής 'Εναρξης για την επανεκκίνηση της μετάδοσης σε περιπτώσεις που αυτή σταματά για ένα διάστημα κατά τη διάρκεια μιας TCP σύνδεσης οδηγεί σε μειωμένη απόδοση. Αντί της χρήσης της Αργής 'Εναρξης κάποιες υλοποιήσεις προτιμούν να συνεχίσουν τη μετάδοση με χρήση της τιμής του cwnd τη στιγμή που σταμάτησε η μετάδοση. Ωστόσο η επιλογή αυτή ενέχει τον κίνδυνο η τιμή αυτή του cwnd να μην αντικατοπτρίζει πια το διαθέσιμο εύρος ζώνης στο δίκτυο και επομένως η επανεκκίνηση της μετάδοσης με αυτή την τιμή να οδηγήσει σε συμφόρηση.

Η μέθοδος της Εκτίμησης Ρυθμού Μετάδοσης (Rate Based Pacing) βασίζεται στη χρήση μιας εκτίμησης του διαθέσιμου εύρους ζώνης για την τρέχουσα σύνδεση. Μεταδίδει segments με ένα σταθερό ρυθμό (όχι back-to-back) και με τιμή cwnd μεγαλύτερη από αυτή του RW. Ο ρυθμός και το cwnd που χρησιμοποιούνται υπολογίζονται με βάση την εκτίμηση του εύρους ζώνης.

2.2.3.2 Βελτιώσεις στην αποκατάσταση απώλειας πακέτου

Η χρήση των αλγορίθμων Ταχείας Επανεκπομπής και Ανάρρωσης για την αποκατάσταση απώλειας πακέτου δίνουν τη δυνατότητα στον αποστολέα να αναμεταδώσει κατά μέγιστο 1 χαμένο segment ανά RTT. Δεδομένου ότι σε χρόνο ίσο με 1 RTT είναι δυνατό να μεταδοθούν κατά μέγιστο segments ενός παραθύρου, η απώλεια περισσότερων του ενός segments στο ίδιο παράθυρο καθιστά τους αλγορίθμους αναποτελεσματικούς. Ο λόγος είναι ότι η αναμετάδοση n segments απαιτεί χρόνο n RTTs και επιπλέον κάθε RTT αποτελεί ξεχωριστή εκτέλεση του κύκλου Fast Retransmit & Fast Recovery με όλο και μικρότερο ssthresh.

Σε δίκτυα με υψηλό γινόμενο Bandwidth*Delay, η αξιοποίηση του διαθέσιμου εύρους ζώνης προϋποθέτει τη χρήση μεγάλων παραθύρων. Με δεδομένο το ρυθμό σφαλμάτων και απωλειών, η χρήση μεγαλύτερων τιμών παραθύρου αυξάνει την πιθανότητα περισσότερων της μιας απωλειών πακέτου στο ίδιο παράθυρο.

Παρουσιάζουμε στη συνέχεια δύο μεθόδους για τη βελτίωση του TCP ως προς την αποκατάσταση απώλειας πακέτου παρουσία περισσότερων της μιας απωλειών στο ίδιο παράθυρο.

New-Reno TCP

Το New-Reno TCP είναι μια παραλλαγή του Reno TCP που διαφοροποιείται μόνο στον τρόπο που τερματίζεται η φάση της Ταχείας Ανάρρωσης. Συγκεκριμένα εισάγεται η έννοια της Μερικής Επιβεβαίωσης (Partial ACK). Partial ACK ονομάζεται κάθε ACK που επιβεβαιώνει μέρος μόνο από τα δεδομένα που είχαν μεταδοθεί όταν ξεκίνησε το η Ταχεία Επανεκπομπής.

Στο Reno TCP η λήψη ενός partial ACK από τον αποστολέα τερματίζει την Ταχεία Ανάρρωσης. Στο New-Reno TCP το partial ACK εκλαμβάνεται ως ένδειξη ότι το επόμενο από αυτό που επιβεβαιώνεται segment επίσης χάθηκε και επομένως δεν λήγει η φάση της Ταχείας Ανάρρωσης. Με τη λήψη του partial ACK ο αποστολέας θα αναμεταδώσει αμέσως το ζητούμενο segment. Η Ταχεία Ανάρρωση θα ολοκληρωθεί μόνο όταν επιβεβαιωθούν όλα τα δεδομένα που είχαν ήδη μεταδοθεί όταν ξεκίνησε η Ταχεία Επανεκπομπή.

Το New Reno αναμεταδίδει 1 χαμένο segment ανά RTT όπως και το Reno TCP. Ωστόσο αποφεύγει τη διαδοχική μείωση στο μισό του ssthresh για κάθε ένα από τα χαμένα segments του ίδιου παραθύρου. Αυτό δεν μπορεί να το αποφύγει το Reno TCP, το οποίο για κάθε χαμένο segment στο ίδιο παράθυρο θα εκτελέσει ξεχωριστό κύκλο Ταχείας Επανεκπομπής και Ανάρρωσης.

Επιλεκτικές Επιβεβαιώσεις

Οι επιβεβαιώσεις στο TCP, όπως τουλάχιστον τις έχουμε δει μέχρι το σημείο αυτό, έχουν μια συσσωρευτική (cumulative) φύση. Δεν επιβεβαιώνουν τη λήψη συγκεκριμένων segments αλλά ολόκληρου του

data stream μέχρι ενός σημείου. Αν υπάρξει ένα κενό στο stream του παραλήπτη ο αποστολέας δεν θα πληροφορηθεί από τις επιβεβαιώσεις ότι segments «δεξιότερα» του κενού έχουν παραληφθεί.

Αυτή ακριβώς την πληροφορία επιχειρούν να μεταφέρουν οι Επιλεκτικές Επιβεβαιώσεις (Selective Acknowledgements, SACK). Τα SACKs δεν αντικαθιστούν τα cumulative ACKs αλλά δρουν συμπληρωματικά ως προς αυτά. Υλοποιούνται με τη χρήση 2 TCP options. Τα options αυτά είναι το SACK-permitted και το SACK. Το πρώτο ανταλλάσσεται κατά την εκκίνηση της σύνδεσης και χρησιμοποιείται για τη διαπραγμάτευση μεταξύ αποστολέα και παραλήπτη για τη χρήση ή όχι των SACKs. Το SACK option περιλαμβάνεται σε ACK segments και μεταφέρει την πληροφορία του Selective Acknowledgement. Πιο συγκεκριμένα ένα SACK option περιγράφει τμήματα του data stream που έχει λάβει ο παραλήπτης και που βρίσκονται «δεξιότερα» από το μεγαλύτερο επιβεβαιωμένο (με cumulative ACK) sequence number.

Η πληροφορία αυτή δίνει τη δυνατότητα στον αποστολέα να αναμεταδώσει χαμένα segments χωρίς να χρειάζεται να περιμένει 1 RTT για το καθένα από αυτά όπως συμβαίνει στο Reno και στο New-Reno TCP. ‘Ετσι η χρήση των SACKs είναι δυνατό να επιταχύνει σημαντικά τη διαδικασία αποκατάστασης απώλειας πακέτων όταν έχουν χαθεί περισσότερα του ενός segments από το ίδιο παράθυρο.

2.2.3.3 ‘Αλλες τροποποιήσεις

Στην ενότητα αυτή παρουσιάζονται μερικές ακόμα τροποποιήσεις - βελτιώσεις που δεν εμπίπτουν στις κατηγορίες που ήδη παρουσιάστηκαν.

- TCP for Transactions (T/TCP): Πρόκειται για παραλλαγή του TCP κατά την οποία μετά την 1η TCP σύνδεση μεταξύ 2 κόμβων παρακάμπτεται η διαδικασία του “3-Way Handshake” για την εγκατάσταση της σύνδεσης και η μετάδοση των δεδομένων ξεκινά από το 1ο segment. Με το T/TCP επιχειρείται η επιτάχυνση των σύντομων TCP συνδέσεων.
- Header Compression: Η χρήση compression στους IP και TCP headers έχει προφανές όφελος. Το όφελος αυτό αποκτά ιδιαίτερη σημασία στην αποσυμφόρηση του αργού reverse link. Εκεί (θεωρώντας ως πλειοψηφία της κίνησης στο αργό link τα TCP ACKs) το κέρδος από τη συμπίεση μεγιστοποιείται.
- Υπολογισμός του ssthresh: Η χρήση ενός μηχανισμού εκτίμησης του διαθέσιμου για την TCP σύνδεση είναι δυνατό να οδηγήσει στον υπολογισμό μιας καταλληλότερης τιμής του ssthresh και κατά συνέπεια σε αποδοτικότερη λειτουργία του συνδυασμού Slow Start και Congestion Avoidance.
- Corruption Detection – Explicit Congestion Notification: Αγνοώντας το TCP αν η απώλεια πακέτου οφείλεται σε λάθος μετάδοσης (corruption) ή σε συμφόρηση στο δίκτυο (congestion) επιλέγει να την ερμηνεύσει ως συμφόρηση. Ωστόσο η γνώση ότι η απώλεια οφείλεται σε σφάλμα μπορεί να αξιοποιηθεί ώστε η αναμετάδοση του χαμένου segment να μην συνοδευτεί από μείωση του cwnd. Η χρήση μηχανισμών

Corruption detection ή Explicit Congestion Notification μπορεί να παρέχει τη γνώση αυτή και να βελτιώσει την απόδοση του TCP.