Πλέγματα Μεταγωγής Benes με Εσωτερικό Backpressure Πολυπλοκότητας $O(N)$

Γεώργιος Σαπουντζής

Μεταπτυχιακή Εργασία

Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Περίληψη

Τα πλέγματα μεταγωγής με πολλαπλά στάδια και εσωτερικούς ενταμιευτές είναι η πιο αποδοτική μέθοδος για την κλιμάκωση των μεταγωγέων πακέτων σε πολύ μέγαλο αριθμό από πόρτες. Το δίκτυο Benes είναι το χαμηλότερου κόστους πλέγμα μεταγωγής που επιτρέπει λειτουργία χωρίς εσωτερική φραγή (internal blocking). Η αναδραστική πίεση (backpressure) εσωτερικά στο πλέγμα μεταγωγής μπορεί να περιορίσει την χρήση ακριβών off-chip ενταμιευτών μνήμης μόνο στις εικονικές ουρές εξόδου (VOQ) μπροστά από το στάδιο εισόδου. Η παρούσα εργασία επεκτείνει τις γνωστές αρχιτεκτονικές backpressure στο δίκτυο Benes. Για να επιτευχθεί αυτό, έπρεπε να συνδυάσουμε επιτυχώς το backpressure ανά-ροή, τη δρομολόγηση μέσω πολλαπλών μονοπατιών (αντίστροφη πολυπλεξία), και την επαναδιάταξη κελιών (cells). Επίσης, παρουσιάζουμε ένα σχήμα συγχώνευσης ροών το οποίο χρειάζεται για τη μείωση του κόστος του backpressure σε $O(N)$ ανά στοιχείο μεταγωγής. Αποδεικνύουμε ανυπαρξία αδιεξόδων για μια μεγάλη κλάσση από αλγορίθμους δρομολόγησης μέσω πολλαπλών μονοπατιών. Με τη χρήση προσομοιώσεων με ακρίβεια cell time, αξιολογήσαμε διάφορες μεθόδους για την κατανομή και επαναδιάταξη των cells, βρήκαμε ότι η καθυστέρηση για εκρηκτική κίνηση είναι μόνο 25 εώς 50 τοις εκατό υψηλότερη από το ιδανικό σύστημα με ουρές στις εξόδους, και δείξαμε ότι η καθυστέρηση των μη-συμφορημένων ροων παραμένει ανεπηρέαστη από την παρουσία συμφορημένης κίνησης προς ορισμένες πόρτες εξόδου. Με τη χρήση απλοποιημένων μοντέλων για το πλέγμα Benes, δείχνουμε ότι η κατανομή των cells δεν δημιουργεί περιορισμούς στην διαπερατότητα, και εντοπίζουμε τα σημεία του πλέγματος όπου επιλύονται οι συγκρούσεις μεταξύ των cells.

Επόπτης Εργασίας:
Μανόλης Κατεβαίνης
Καθηγητής, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Ευχαριστίες

Καταρχήν, θα ήθελα να ευχαριστήσω τον επόπτη καθηγητή μου Μανόλη Κατεβαίνη για την υπομονή του και την καθοδήγηση καθόλη τη διάρκεια αυτής της εργασίας. Επίσης, θα ήθελα να ευχαριστήσω τα μέλη της ομάδας αρχιτεκτονικής μεταγωγέων πακέτων, Ν. Χρυσό, Β. Σύρρη και Π. Φραγκοπούλου για τις συζητήσεις και τα πολύτιμα σχόλια τους καθώς και το ΙΤΕ για την υλικοτεχνική υποδομή και την οικονομική υποστήριξη. Τέλός, θα ήθελα να ευχαριστήσω τους φίλους και την οικογένεια μου για την υποστήριξη τους.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Οι μεταγωγείς, και οι δρομολογητές που τους χρησιμοποιούν, είναι τα βασικά δομικά στοιχεία για την κατασκευή δικτύων υψηλής ταχύτητας που χρησιμοποιούν συνδέσμους από-σημείο-σε-σημείο (point-to-point). Καθώς οι απαιτήσεις για τη διαπερατότητα του δικτύου μεγαλώνουν, χρειαζόμαστε μεταγωγείς με περισσότερες και ταχύτερες πόρτες. Αυτή η εργασία αφορά την *κλιμάκωση του μεταγωγέα*, όταν *αυξάνει ο αριθμός από πόρτες*. Για μικρό και μέτριο αριθμό από πόρτες –εώς 64 περίπου– το crossbar είναι η τοπολογία μεταγωγής που επιλέγουμε, λόγω της απλότητας του και της ικανότητας του για λειτουργία χωρίς φραγή (non-blocking). Ωστόσο, το κόστος του αυξάνει με το N^2, όπου N είναι ο αριθμός από πόρτες, που το κάνει πολύ ακριβό για μεγάλα N. Επιπλέον, ο χρονοπρογραμματισμός του crossbar είναι δύσκολο πρόβλημα, και γίνεται δυσκολότερο όσο μεγαλώνει το N.

Για μεταγωγείς με εκατοντάδες ή χιλιάδες από πόρτες, χρειάζονται αρχιτεκτονικές βασισμένες σε *πλέγματα μεταγωγής με πολλαπλά στάδια*, το κόστος των οποίων μεγαλώνει με ρυθμό μικρότερο από τετραγωνικό. Οι ερευνητές έχουν μελετήσει τέτοιες κλιμακώσιμες τοπολογίες ήδη από τις μέρες της ηλεκτρομηχανικής τηλεφωνίας [1]. Το δίκτυο banyan [2] χαρακτηρίζεται από χαμηλό κόστος, $N \cdot logN$ και μεγάλο αριθμό από μονοπάτια. Αν και μπορεί να υποστηρίξει πλήρη απασχόληση των συνδέσμων εξόδου για κίνηση με ομοιόμορφη κατανομή προορισμών, οπώς και για ορισμένα άλλα συγκεκριμένα μοντέλα κίνησης, υποφέρει από *εσωτερική φραγή*: δεν είναι δυνατό να δρομολογηθούν μέσω του δικτύου banyan όλα τα σύνολα *εφικτών* ρυθμών $\lambda_{i,j}$. Το $N \times N$ δίκτυο με το χαμηλότερο κόστος που δεν υποφέρει από εσωτερική φραγή είναι το δίκτυο *Benes* [3], το κόστος του οποίου είναι $N \cdot 2logN$. Το δίκτυο Benes είναι χωρίς εσωτερική φραγή αλλά μετά από *αναδιευθετήσεις* (rearrangeably non-blocking), δηλαδή, όταν κάθε σύνδεση δρομολογείται μέσω ενός και μόνο μονοπατιού, η εγκατάσταση νέων συνδέσεων μπορεί να απαιτήσει την επανα-δρομολόγηση ήδη υπαρχόντων συνδέσεων. Ωστόσο, με τη χρήση δρομολόγησης μέσω πολλαπλών μονοπατιών, αυτό το μειονέκτημα μπορεί να εξαλειφθεί: βλέπε ενότητα 2.1.1. Αυτή η εργασία αφορά το δίκτυο Benes.

Αν ένα πλέγμα μεταγωγής με πολλαπλά σταδια δεν περιέχει ενταμιευτές για αποθήκευση, πρέπει να υπάρχει ένας μηχανισμός που να χειρίζεται τις συγκρούσεις που παρουσιάζονται μεταξύ των cells *(α)* στα εσωτερικά μονοπάτια λόγω του αλγορίθμου δρομολόγησης, *(β)* εξαιτίας των συγκρούσεων στις εξόδους. Οι πρώ-

τες συγκρούσεις μπορούν να αντιμετωπιστούν με κατανεμημένο τρόπο ("πλέγματα με αυτο-δρομολόγηση") με τη χρήση δικτύων ταξινόμησης Batcher [4]. Οι δεύτερες συγκρούσεις –cells που προορίζονται για την ίδια έξοδο την ίδια χρονική στιγμή– πρέπει να εξαλειφθούν στις εισόδους ή να αντιμετωπιστούν μέσα στο πλέγμα. Η εξάλειψη στις εισόδους είναι ισοδύναμη με τον χρονοπρογραμματισμό του crossbar και απαιτεί καθολική συνεργασία, επομένως είναι ανεδαφική για μεγάλα πλέγματα. Για την αντιμετώπιση των συγκρούσεων στις εξόδους μέσα στο πλέγμα, οι σχεδιαστές έχουν χρησιμοποιήσει επανα-κυκλοφορία των cells [5] ή πολλαπλά μονοπάτια προς κάθε έναν από τους ενταμιευτές εξόδου [6]. Όλοι αυτοί οι μηχανισμοί κοστίζουν πολύ όσον αφορά των αριθμό των σταδίων και τα μονοπάτια ανά στάδιο στο πλέγμα μεταγωγής: το κόστος του πλέγματος είναι $O(N \cdot log^2 N)$, και η σταθερά μπροστά από το πραγματικό κόστος είναι σημαντική. Στην ουσία, αυτές οι τεχνικές ξοδεύουν (ακριβούς) επικοινωνιακούς πόρους για να κάνουν οικονομία σε (φθηνούς) αποθηκευτικούς πόρους, το οποίο είναι το λάθος tradeoff στην σύγχρονη τεχνολογία VLSI.

Είναι προτιμότερο για το πλέγμα μεταγωγής να περιέχει εσωτερικούς ενταμιευτές προκειμένου να αποθηκεύει τα προσωρινά συγκρουόμενα cells εως ότου απαλειφθεί η σύγκρουση. Αυτός ο αποθηκευτικός χώρος μπορεί να είναι "μικρός" ώστε να χωράει στα chips των στοιχείων μεταγωγής, ή μπορεί να είναι αρκετά "μεγάλος" προκειμένου να αντικαθιστά όλο τον αποθηκευτικό χώρο που συνήθως υπάρχει στις κάρτες εισόδου, –συνήθως εκατοντάδες MBytes– με αποτέλεσμα να απαιτείται off-chip DRAM. Στην πρώτη περίπτωση, χρησιμοποιείται *backpressure* για να αποτρέψει την υπερχείλιση των μικρών ενταμιευτών, με αποτέλεσμα η πλειοψηφία των αποθηκευμένων cells να απωθούνται στις κάρτες εισόδου, σε virtual-output queues (VOQ). Δεδομένου ότι οι κάρτες εισόδου είναι πολύ λιγότερες από του συνδέσμους μέσα στο πλέγμα, αυτή η αρχιτεκτονική έχει ως αποτέλεσμα τη σημαντική μείωση του κόστους σε σύγκριση με την αρχιτεκτονική που χρησιμοποιεί off-chip DRAM για τους ενταμιευτές μέσα στο πλέγμα, όπως έδειξε και η αξιολόγηση του μεταγωγέα ATLAS I [7]. Η εργασία αυτή αφορά την εφαρμογή αυτής της επωφελούς αρχιτεκτονικής με *εσωτερικό backpressure* στο δίκτυο Benes –το χαμηλότερου κόστους, κλιμακώσιμο πλέγμα μεταγωγής.

Στην εργασία αυτή επεκτείνουμε τις αρχιτεκτονικές backpressure από πλέγματα με ένα μονοπάτι ανά είσοδο-έξοδο (όπως τα δίκτυα banyan) σε τοπολογίες με πολλαπλά μονοπάτια ανά είσοδο-έξοδο, και ειδικά στο δίκτυο Benes. Η επέκταση αυτή δεν είναι τετριμένη. Προκειμένου να μην υπάρχει εσωτερική φραγή στο πλέγμα Benes, τα cells κάθε ροής πρέπει να δρομολογηθούν μέσω πολλαπλών μονοπατιών, και μετά να επαναδιαταχθούν κατάλληλα, όπως εξηγείται στην ενότητα 2.1.1. Προκειμένου να μην υπάρχουν προβλήματα head-of-line-blocking στην λειτουργία του backpressure, πρέπει το backpressure να μπορεί να χειρίζεται κάθε ροή χωριστά, όπως εξηγείται στην ενότητα 2.1.2. Αν αυτές οι δύο απαιτήσεις – δρομολόγηση μέσω πολλαπλών μονοπατιών και δυνατότητα του backpressure να χειρίζεται κάθε ροή χωριστά – συνδυάζονταν με απλοϊκό τρόπο, θα είχαμε ως αποτέλεσμα πολυπλοκότητα της τάξης $O(N^2)$ για τα στοιχεία μεταγωγής στο μεσαίο στάδιο του πλέγματος Benes. Δείχνουμε πώς να μειωθεί αυτή η πολυπλοκότητα σε $O(N)$, με τη χρήση κατάλληλων τεχνικών συγχώνευσης των ροών με ελάχιστη επίπτωση στην απόδοση του συστήματος: βλέπε ενότητα 2.2.1. Η πολυπλοκότητα $O(N)$ που προκύπτει είναι ρεαλιστική για τη σημερινή τεχνολογία VLSI αφού πλέγματα με μέγεθος N της τάξης των μερικών χιλιάδων απαιτούν on-chip αποθηκευτικό χώρο της τάξης των μερικών χιλιάδων cells (μερικά Mbits), το οποίο είναι εφικτό.

Η κατανομή των cells σε πολλαπλά μονοπάτια αλληλεπιδρά με τη συγχώνευση των ροών, και τα δύο αυτά αλληλεπιδρούν με την οργάνωση και την τοποθέτηση των ενταμιευτών, δείχνουμε ποια οργάνωση είναι προτιμητέα, και αποδεικνύουμε την ανυπαρξία αδιεξόδων για αυτή (ενότητα 3.3). Η ενότητα 4 παρουσιάζει τα αποτελέσματα των προσομοιώσεων, δείχνοντας ότι *(α)* επιτυγχάνεται λειτουργία χωρίς εσωτερική φραγή με πλήρη απασχόληση των εξόδων, *(β)* τα γραφήματα της καθηστέρησης συναρτήσει του φόρτου για αυτό το πλέγμα μεταγωγής και για εκρικτική κίνηση είναι συγκρίσιμα με έναν παράγοντα 1.5 με εκείνα για το ιδανικό σύστημα με ουρές στις εξόδους, *(γ)* η καθυστέρηση σε προορισμούς χωρίς συμφόρηση επηρρεάζεται ελάχιστα από την παρουσία συμφόρησης σε άλλα σημεία του δικτύου, και *(δ)* όποια μέθοδο κατανομής των cells σε πολλαπλά μονοπάτια και αν επιλέξουμε μέσα από την κλάσση των μεθόδων που θεωρούμε δεν επηρρεάζει πολύ την καθυστέρηση των cells. Τέλος, με τη χρήση απλοποιημένων μοντέλων για το πλέγμα Benes, δείχνουμε ότι η κατανομή των cells δεν περιορίζει την διαπερατότητα του συστήματος, και εντοπίζουμε τα σημεία του πλέγματος όπου επιλύονται οι συγκρούσεις.

Από όσο γνωρίζουμε, αυτή είναι η πρώτη φορα που μελετάται η εφαρμογή backpressure ανά-ροή στο πλέγμα μεταγωγής Benes. Επιπλέον, δεν γνωρίζουμε άλλες μελέτες του backpressure σε συνδυασμό με δρομολόγηση cells μέσω πολλαπλών μονοπατιών. Η δρομολόγηση cells μέσω πολλαπλών μονοπατιών έχει μελετηθεί και αλλού, π.χ. [8] [9] [10], αλλά όχι σε συνδυασμό με backpressure.

Κεφάλαιο 2

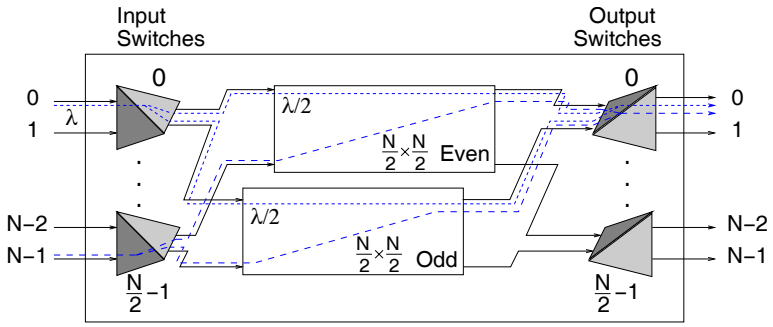
Γενική Αρχιτεκτονική

2.1 Το Πλέγμα Benes

Η ενότητα αυτή είναι μια ανασκόπηση των δύο βασικών ιδεών της σχεδίασης, που είναι το πλέγμα Benes, και το εσωτερικό backpressure σε μεταγωγείς.

2.1.1 Λειτουργία χωρίς Εσωτερική Φραγή

Το δίκτυο Benes [3] μπορεί να κατασκευαστεί αναδρομικά, με τη χρήση *αντίστροφης πολυπλεξίας*, όπως φαίνεται στο σχήμα 2.1. Το $N \times N$ δίκτυο Benes αποτελείται από δύο $\frac{N}{2} \times \frac{N}{2}$ υποδίκτυα Benes, $\frac{N}{2}$ μεταγωγείς μεγέθους 2×2 που συνδέονται στις εισόδους των δύο υποδικτύων, και τα ονομάζουμε μεταγωγείς εισόδου, και $\frac{N}{2}$ μεταγωγείς μεγέθους 2×2 που συνδέονται στις εξόδους των δύο υποδικτύων, και τα ονομάζουμε μεταγωγείς εξόδου.

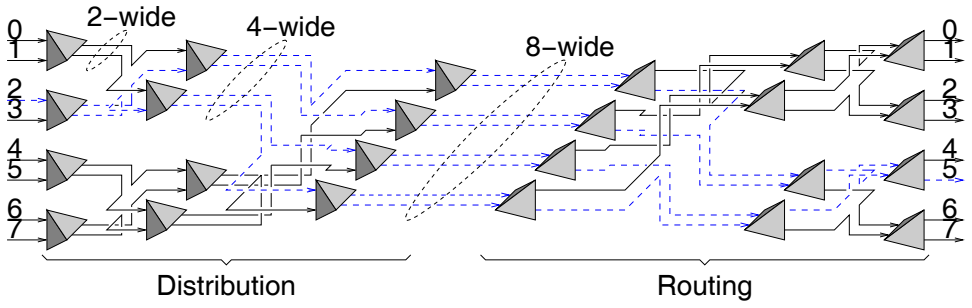


Σχήμα 2.1: *Αναδρομική κατασκευή ενός δικτύου Benes μεγέθους $N \times N$.*

Έστω $\lambda_{i,j}$ η κίνηση που εισέρχεται στο δίκτυο από την είσοδο i και προορίζεται για την έξοδο j. Προκειμένου το $N \times N$ δίκτυο να μην έχει εσωτερική φραγή ο 2×2 μεταγωγέας που συνδέεται στην είσοδο i πρέπει να κατανέμει το $\lambda_{i,j}$ εξίσου στις δύο εξόδους του. Ο μεταγωγέας εξόδου που τροφοδοτεί την έξοδο j λαμβάνει $\frac{1}{2}\lambda_{i,j}$ σε κάθε μία από τις εισόδους του, ανακατασκευάζει το $\lambda_{i,j}$ και το δρομολογεί στην κα-

τάλληλη έξοδο. Η ανυπαρξία εσωτερικής φραγής προκύπτει ως εξής. Για οποιοδήποτε σύνολο από *εφικτούς* ρυθμούς $\lambda_{i,j}$ που εισέρχονται στο $N \times N$ δίκτυο (δηλ. $\sum_{j=0}^{N-1} \lambda_{i,j} \leq 1, \ \forall i$) και εξέρχονται από το $N \times N$ δίκτυο (δηλ. $\sum_{i=0}^{N-1} \lambda_{i,j} \leq 1, \ \forall j$), οι ρυθμοί που εισέρχονται σε και εξέρχονται από κάθε $\frac{N}{2} \times \frac{N}{2}$ υποδίκτυο θα είναι επίσης εφικτοί. Συγκεκριμένα, η είσοδος k είτε του ενός είτε του άλλου υποδικτύου θα δέχεται $\sum_{j=0}^{N-1} \frac{1}{2}\lambda_{2k,j} + \sum_{j=0}^{N-1} \frac{1}{2}\lambda_{2k+1,j}$ το οποίο είναι $\leq \frac{1}{2} + \frac{1}{2} = 1$ λόγω του ότι η συνολική κίνηση είναι εφικτή. Συμμετρικά, ο φόρτος στην έξοδο m είτε του ενός είτε του άλλου υποδικτύου θα είναι $\sum_{i=0}^{N-1} \frac{1}{2}\lambda_{i,2m} + \sum_{i=0}^{N-1} \frac{1}{2}\lambda_{i,2m+1} \leq \frac{1}{2} + \frac{1}{2} = 1$. Υποθέτοντας τώρα ότι κάθε υποδίκτυο δεν έχει εσωτερική φραγή, δηλ. μπορεί να δρομολογήσει οποιοδήποτε εφικτή κίνηση, αναδρομικά προκύπτει ότι το συνολικό $N \times N$ δίκτυο επίσης δεν έχει εσωτερική φραγή.

Ξεδιπλώνοντας την αναδρομή στο σχήμα 2.1, και για N = 8, προκύπτει η τοπολογία που φαίνεται στο σχήμα 2.2. Η κίνηση $\lambda_{i,j}$ περνάει από $\log N$ στάδια κατανομής και $\log N$ αντίστοιχα στάδια ανακατασκευής. Το σχήμα επίσης δείχνει ότι το $N \times N$ δίκτυο Benes μπορεί να κατασκευαστεί τοποθετώντας δύο δίκτυα banyan το ένα δίπλα στο άλλο. Τα δίκτυα banyan ονομάζονται και δίκτυα *κατανομής* και *δρομολόγησης*, αντίστοιχα [11], αφού το πρώτο δίκτυο κατανέμει την εισερχόμενη κίνηση στους N συνδέσμους στη καρδιά του δικτύου Benes – οι N σύνδεσμοι είναι ένας "φαρδύς" εικονικός σύνδεσμος με διαπερατότητα N – και το δεύτερο δίκτυο δρομολογεί τα cells στο κατάλληλο σύνδεσμο εξόδου.



Σχήμα 2.2: *Δίκτυο Benes μεγέθους* 8×8 *που δείχνει την κατανομή και ανακατασκευή της κίνησης* $\lambda_{2,5}$.

Η ανυπαρξία εσωτερικής φραγής προλύπτει από (επαναλαμβανόμενη) *αντίστροφη πολυπλεξία* ή *κατανομή φορτίου* με ισομερή τρόπο. Μία απλοϊκή μέθοδος για την κατανομή φορτίου είναι η αποστολή όλων των πακέτων που ανήκουν σε "μισές" από τις μικροροές μέσω ενός μονοπατιού, και όλων των πακέτων που ανήκουν στις άλλες μισές ροές μέσω του άλλου μονοπατιού, π.χ. χρησιμοποιώντας μια ψευδοτυχαία συνάρτηση που, για να αποφασίσει το μονοπάτι μιας μικροροής, κάνει hashing με βάση τις διευθύνσεις IP πηγής και προορισμού. Αυτή η μέθοδος εγγυάται ότι όλα τα πακέτα μιας δεδομένης μικροροής ακολουθούν την ίδια διαδρομή και επομένως φτάνουν στη σειρά. Το μειονέκτημα αυτής της μεθόδου είναι ότι η κατανομή φορτίου μπορεί να μην είναι ισομερής μακροπρόθεσμα, και ακόμα χειρότερα σε βραχυπρόθεσμη βάση, ειδικά όταν ο αριθμός των μικροροών είναι περιορισμένος. Ανισομερής κατανομή φορτίου προκαλεί εσωτερική φραγή στο πλέγμα Benes και επομένως δεν χρησιμοποιούμε την μέθοδο αυτή. Στην άλλη άκρη των δυνατών μέθοδων είναι μία μέθοδος για επακριβή κατανομή φορτίου που μοιάζει με τους *bit-sliced* επεξεργαστές της δεκαετίας του 70. Κάθε cell διαιρείται σε δύο μονάδες, κάθε μία με μέγεθος το μισό του

αρχικού cell, και κάθε μονάδα στέλνεται σε μία από τις δύο κατευθύνσεις. Η μέθοδος αυτή χρησιμοποιείται σε αρκετά εμπορικά chip sets, αλλά μόνο με διαιρέσεις μέχρι και σε 8 μέρη και με προσεκτικό σχεδιασμό των μονοπατιών ώστε να έχουν ίσες καθυστερήσεις [12]. Η μέθοδος αυτή δεν είναι κλιμακώσιμη, λόγω της πάγιας επιβάρυνσης για την επεξεργασία των επικεφαλίδων του cell και της κάθε μονάδας στις οποίες διαιρείται το cell, και επομένως δεν τη χρησιμοποιούμε.

Για την επίτευξη ισομερούς κατανομής φορτίου μακροπρόθεσμα –ίσως, όμως, χωρίς την επίτευξη της βραχυπρόθεσμα– και λειτουργία του συστήματος σε επίπεδο cells, έχουν προταθεί διάφορες μέθοδοι: τυχαιοκρατικές (randomized) [13], με προσαρμοστικότητα (adaptive) [8], ανα-ροή, εκ-περιτροπής κατανομή των cells (per-flow round-robin cell distribution) [14]. Σε όλες τις παραπάνω μεθόδους, τα cells μιας δεδομένης μικροροής δρομολογούνται μέσω είτε του ενός είτε του άλλου μονοπατιού, και επομένως μπορούν να φτάσουν στην έξοδο εκτός σειράς. Προκειμένου το πλέγμα μεταγωγής να διατηρεί τη σειρά των cells που ανήκουν όλα στην ίδια μικροροή, πρέπει να γίνεται επαναδιάταξη (resequencing) των cells στα σημεία που ξανασυγκλίνουν τα μονοπάτια [15] [9]. Η επαναδιάταξη είναι ένα σημαντικό θέμα για το σύστημα μας, με το οποίο ασχολούμαστε στις ενότητες 2.2.1 και 3.3.

2.1.2 Πρωτόκολλα για Εσωτερικό Backpressure

Οι μεταγωγείς με ενταμιευτές σε πολλαπλά στάδια ως επι το πλείστον χρησιμοποιούν αναδραστικό έλεγχο μεταξύ των σταδίων βασισμένο σε *backpressure*, *(α)* για να αποφύγουν υπερχείλιση των ενταμιευτών στα κατάντη σταδια, και *(β)* για τον έλεγχο του ρυθμού κάθε μίας ροής χωριστά όταν πολλές ροές ανταγωνίζονται για πόρους που έχουν γίνει oversubscribed, και με αυτό τον τρόπο παρέχουν εγγυήσεις ποιότητας υπηρεσίας (QoS).

Το απλούστερο πρωτόκολλο backpressure είναι το *stop-and-go*: το ανάντη σταδιο διατηρεί ένα bit κατάστασης (συνολικά ή ανα-ροή), που καθορίζει εάν επιτρέπεται ή όχι η μετάδοση cells προς το κατάντη στάδιο. Ένα πιο εξεζητημένο πρωτόκολλο, που έχει μικρότερες (λιγότερες από τις μισές) απαιτήσεις σε αποθηκευτικό χώρο, είναι αυτό που χρησιμοποιεί *credits*: το ανάντη στάδιο διατηρεί έναν μετρητή για τα credits (συνολικά ή ανα-ροή), που καθορίζει πόσα cells επιτρέπεται να μεταδωθούνε στην κατάντη κατευθυνση πριν φτάσουν καινούργια credits μέσω αναδραστικών σημάτων backpressure. Ο απαιτούμενος χώρος για ενταμιευτές είναι $\lambda \times RTT$ (συνολικά ή ανα-ροή), όπου λ είναι ο μέγιστος ρυθμός και RTT είναι ο χρόνος μετ'επιστροφή. Η εργασία αυτή χρησιμοποιεί credit-based backpressure.

Τα σήματα backpressure μπορεί να αναφέρονται σε κάθε μία (μικρο)ροή χωριστά ή σε συνενώσεις ροών, ή αδιάκριτα σε όλη τη κίνηση που περνάει από έναν σύνδεσμο. Η εφαρμογή backpressure αδιάκριτα έχει ως αποτέλεσμα πολύ κακή QoS, γιατί μία και μόνο συμφορημένη ροή μπορεί να διακόψει την εξυπηρέτηση όλων των ροών με τις οποίες διαμοιράζεται έναν σύνδεσμο ή έναν ενταμιευτή (αυτό είναι ανάλογο με το πρόβλημα head-of-line (HOL) blocking). Επομένως, χρειάζεται backpressure *ανα-ροή* ή με *εικονικά κανάλια* (virtual-channel) ή με *πολλαπλές λωρίδες* (multilane). Ο ορισμός και ο αριθμός των "ροών" είναι μια κρίσιμη παράμετρος και επηρρεάζει το κόστος –το συνολικό μέγεθος της κατάστασης και το αντικείμενο στο οποίο αναφέρεται (granularity) η πληροφορία ανάδρασης– και το QoS –βαθμός απομόνωσης μεταξύ αντα-

γωνιζόμενων ροών. Όταν είναι ακριβό το granularity του backpressure να είναι η κάθε ροή, μπορούμε να χρησιμοποιήσουμε μια "συμβιβαστική" λύση ή κατάλληλη συγχώνευση ροών. Τα συμβιβαστικά πρωτόκολλα backpressure πετυχαίνουν καλή απόδοση στη συνήθη περίπτωση αλλά πολύ κακή επίδοση στις χείριστες περιπτώσεις, τέτοια πρωτόκολλα είναι τα παρακάτω: το wormhole virtual channels [16], μια πρόταση από την DEC [17], το Quantum Flow Control [18], και το multilane backpressure του ATLAS I [19].

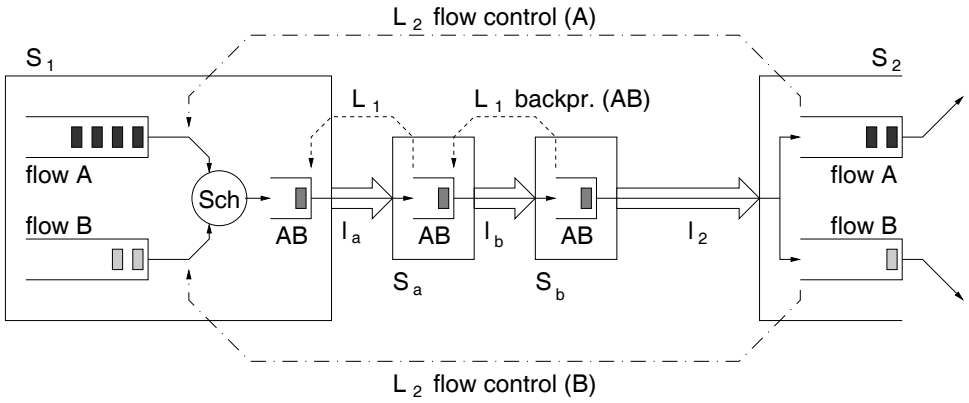
Αυτή η εργασία ασχολείται με backpressure ανα-ροή, το οποίο εγγυάται ότι ακόμη και αν όλες οι πόρτες εξόδου εκτός από μια είναι oversubscribed, η κίνηση που κατευθύνεται στη μία μη-συμφορημένη έξοδο θα υφίσταται καθυστερήσεις συγκρίσιμες με αυτές σε ένα ιδανικό σύστημα με ουρές στις εξόδους. Πετυχαίνουμε ισχυρές εγγυήσεις QoS με κόστος όχι μεγαλύτερο από $O(N)$ ανά στοιχείο μεταγωγής, το οποίο κόστος είναι ρεαλιστικό στη σύγχρονη τεχνολογία VLSI.

Τα βασικά εργαλεία αυτής της προσπάθειας είναι η *συγχώνευση ροών με κοινό προορισμό* και το *ιεραρχικό backpressure*. Όταν πολλές ροές που ανήκουν στο ίδιο επίπεδο προτεραιότητας ακολουθούν το ίδιο μονοπάτι προς ένα κοινό προορισμό, τότε μπορούμε να τις θεωρήσουμε σαν μία ροή όσον αφορά την δέσμευση των ενταμιευτών και το granularity του backpressure μετά από συγχώνευση κατά μήκος του κοινού μονοπατιού. Ο λόγος είναι ότι τα cells μιας ροής δεν θα χρειαστεί ποτέ να προσπεράσουν τα cells μιας άλλης μετα το σημείο της συγχώνευσης. Ένα (μικρό) μειονέκτημα της συγχώνευσης αυτού του είδους είναι η μεταβατική συμπεριφορά του συστήματος όταν μία από τις ροές μεταβαίνει από την ανενεργή στην ενεργή κατάσταση: η "pipeline" μπροστά από το σημείο συγχώνευσης έχει ήδη γεμίσει με cells άλλων ροών. Για συστήματα χρονοπρογραμματισμού τύπου weighted round robin (WRR) υπάρχει ο κίνδυνος αυτή η pipeline να αδειάζει με ρυθμό που αντιστοιχεί στα βάρη των παλιών ροών, ενώ η ροή που μόλις έχει ενεργοποίηθει έχει πολύ μεγαλύτερο βάρος.

2.1.3 Ιεραρχικό Backpressure

Το ιεραρχικό backpressure [20, ενότητα III-D], που απεικονίζεται και στο σχήμα 2.3, μπορεί να χρησιμοποιηθεί όταν πολλές ροές (που ανήκουν στο ίδιο επίπεδο προτεραιότητας κατά προτίμηση) μοιράζονται ένα κοινό τμήμα από τα μονοπάτια τους. Αυτή η περίπτωση διαφέρει από την προηγούμενη: τώρα επιτρέπεται στις ροές να αποκλίνουν μετά το κοινό τους μονοπάτι. Μπορούμε να θεωρήσουμε τις ροές σαν μία, μετά από συγχώνευση, κατά μήκος του κοινού μονοπατιού, υπό την προυπόθεση ότι υπάρχει ένα υψηλότερο επίπεδο ελέγχου ροής από το σημείο της απόκλισης (στο τέλος του κοινού μονοπατιού) προς το σημείο της συγχώνευσης. Τότε το κοινό μονοπάτι συμπεριφέρεται σαν ένας ενιαίος, "εικονικός" σύνδεσμος, και το δεύτερο επίπεδο ελέγχου ροής αφορά την πρόσβαση στον εικονικό σύνδεσμο.

Στο σχήμα 2.3, οι ροές A και B έχουν ένα κοινό μονοπάτι, που εκτείνεται από το μεταγωγέα (ή στάδιο) S_1 στο S_2, διαμέσου των μεταγωγέων (ή σταδίων) S_a και S_b. Κατά μήκος του κοινού μονοπατιού (σύνδεσμοι l_a, l_b, και l_2) μπορούμε να θεωρήσουμε τις ροές σαν μία ροή 'AB" μετά από συγχώνευση. Στο σημείο συγχώνευσης, ένας χρονοπρογραμματιστής "Sch" αποφασίζει από ποια ροή θα προέλθει το επόμενο cell τύπου AB, ο οποίος βασίζει τις αποφάσεις του σε αναδραστική πληροφορία ελέγχου ροής επιπέδου 2 (L_2) την οποία λαμβάνει από το σημείο απόκλισης, S_2. Το granularity του backpressure επιπέδου 1 (L_1)



Σχήμα 2.3: *Ιεραρχικός έλεγχος ροής.*

είναι η ροή μετά από συγχώνευση AB. Το backpressure τύπου L_1 μεταδίδεται από το S_b στο S_a μέσω του l_b, και από το S_a στο S_1 μέσω του l_a. Είναι σημαντικό να παρατηρήσουμε ότι δεν χρειάζεται έλεγχος ροής τύπου L_1 για τον σύνδεσμο l_2: η ροή προς τους ενταμιευτές A και B του μεταγωγέα S_2 διευθετείται μέσω ελέγχου ροής τύπου L_2, και όχι τύπου L_1.

Μία σημαντική εφαρμογή του ιεραρχικού ελέγχου ροής, όπως φαίνεται στο σχήμα 2.3, είναι σε μεγάλους μεταγωγείς που αποτελούνται από πλέγματα μεταγωγής με πολλαπλά σταδια και κάρτες συνδέμένες στους συνδέσμους εισόδου που περιέχουν ενταμιευτές με εικονικές ουρές εξόδου (VOQ). Στο σχήμα 2.3, θεωρείστε ότι S_1 είναι η κάρτα εισόδου ενός (μεγάλου) μεταγωγέα S_1, και S_2 είναι η κάρτα εισόδου του επόμενου κατάντη (μεγάλου) μεταγωγέα S_2; S_a και S_b είναι τα στάδια του πλέγματος μεταγωγής S_1. Οι σύνδεσμοι l_a και l_b είναι εσωτερικοί στο πλέγμα μεταγωγής, ενώ ο σύνδεσμος l_2 είναι σύνδεσμος του δικτύου (ενός WAN). Παρατηρείστε ότι στο παραπάνω σενάριο το backpressure τύπου L_1 είναι αμιγώς εσωτερικό στο πλέγμα μεταγωγής, ενώ ο έλεγχος ροής τύπου L_2 είναι στο επίπεδο του δικτύου (οποιουδήποτε τύπου έλεγχος ροής: hop-by-hop ή end-to-end, credit- ή rate-based). Δεν υπάρχει backpressure τύπου L_1 στο σύνδεσμο l_2 του δικτύου.

Αυτό είναι το μοντέλο που υποθέτουμε στην παρούσα εργασία: ασχολούμαστε αποκλειστικά με τον έλεγχο ροής τύπου "L_1" *εσωτερικά* στο πλέγμα Benes. Υποθέτουμε ότι αυτός είναι τύπου credit-based backpressure, ανεξάρτητα από τον τύπο του ελέγχου ροής που εφαρμόζεται στο συνολικό δίκτυο. Προσέξτε στο σχήμα 2.3 ότι το backpressure L_1 λειτουργεί πάνω σε συννενώσεις ροών που αποτελούνται από όλες τις μικροροές του δικτύου που μοιράζονται το ίδιο μονοπάτι (και το ίδιο επίπεδο προτεραιότητας) μέσα στο πλέγμα μεταγωγής. Επομένως, στο υπόλοιπο αυτής της εργασίας, και για ένα $N \times N$ πλέγμα με pl επίπεδα προτεραιότητας, θεωρούμε μόνο τις $N^2 \times (pl)$ ροές που ορίζονται η κάθε μία από μία συγκεκριμένη πόρτα εισόδου, i, μία συγκεκριμένη πόρτα εξόδου, j, και ένα συγκεκριμένο επίπεδο προτεραιότητας. Η συγχώνευση πολλών εξωτερικών ροών σε μία από τις παραπάνω εσωτερικές ροές πραγματοποιείται από το χρονοπρογραμματιστή "Sch" του σχήματος 2.3 στην κάρτα εισόδου του μεγάλου μεταγωγέα, πρίν από την είσοδο στο πλέγμα μεταγωγής Benes, και δεν αποτελεί αντικείμενο μελέτης της παρούσας εργασίας.

2.2 Οργάνωση των Στοιχείων Μεταγωγής

Στην ενότητα αυτή, παρουσιάζουμε σχήματα συγχώνευσης ροών τα οποία μειώνουν το κόστος του back-pressure (ανά στοιχείο μεταγωγής) από $O(N^2)$ σε $O(N)$. Ακολούθως, περιγράφουμε τις ουρές και την λειτουργικότητα εσωτερικά στα στοιχεία μεταγωγής των δικτύων banyan κατανομής και δρομολόγησης.

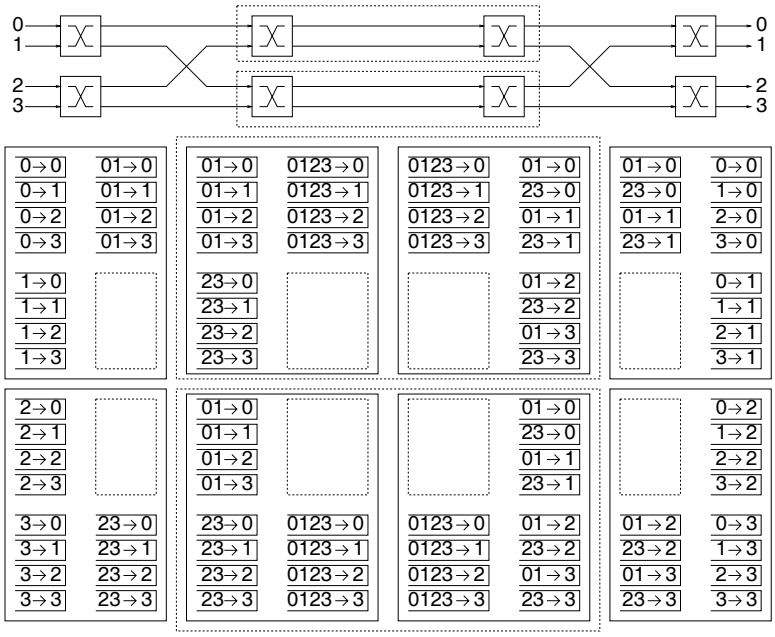
2.2.1 Ομάδες Ροών

Όπως επισημάναμε στις ενότητες 2.1.2 και 2.1.3, το granularity του backpressure πρέπει να είναι οι N^2 ροές (ανά επίπεδο προτεραιότητας) που ορίζονται από όλα τα ζεύγη εισόδου-εξόδου. Στα πλέγματα banyan, αν και ο συνολικός αριθμός των ροών είναι N^2, μόνο N ροές διέρχονται από κάθε σύνδεσμο του πλέγματος. Στο πλέγμα Benes, ωστόσο, η κίνηση κάθε ροής κατανέμεται και στα δύο, "ζυγό" και "περιττό", υποδίκτυα του σχήματος 2.1, συνεπώς, από όλα τα υποδίκτυα, ανεξάρτητα του πόσο μικρά είναι, μέχρι και από τα στοιχεία μεταγωγής στον πυρήνα του πλέγματος, διέρχονται N^2 ροές (ανά επίπεδο προτεραιότητας). Ο αριθμός αυτός πρέπει να μειωθεί, με τη χρήση τεχνικών συγχώνευσης ροών οι οποίες παρουσιάζονται στις ενότητες 2.1.2 και 2.1.3.

Πρώτα εξετάζουμε τη συγχώνευση *ανα-έξοδο* των ροών που κατευθύνονται στην ίδια πόρτα εξόδου του πλεγμάτος. Στο σχήμα 2.4 φαίνονται οι *ομάδες ροών* πάνω στις οποίες δουλεύει το backpressure, το "$01 \rightarrow 0$" υποδηλώνει τη συγχώνευση των ροών $0 \rightarrow 0$ και $1 \rightarrow 0$, και το '$0123 \rightarrow 0$" τη συγχώνευση των ομάδων ροών $01 \rightarrow 0$ και $23 \rightarrow 0$. Το συγκεκριμένο παράδειγμα χρησιμοποιεί στοιχεία μεταγωγής μεγέθους 2×2. Κάθε στοιχείο μεταγωγής του δικτύου κατανομής (το αριστερό μισό του πλέγματος Benes) συγχωνεύει, μία-προς-μία, τις N ομάδες ροών που εισέρχονται από μια από τις εισόδους του με τις N ομάδες ροών που εισέρχονται από την άλλη, και παράγει N ομάδες ροών μετά τη συγχώνευση, ο παράγοντας συγχώνευση είναι δύο-προς-ένα. Αυτά τα στοιχεία μεταγωγής επιπλέον κατανέμουν τα cells και στις δύο εξόδους, οπότε οι N ομάδες ροών μετά τη συγχώνευση εμφανίζονται σε κάθε μία από τις εξόδους, το σχήμα 2.4 δείχνει μία από τις εξόδους λεπτομερώς, και χρησιμοποιεί ένα άδειο κουτί για την άλλη. Επομένως, από όλους τους συνδέσμους διέρχονται ακριβώς N ομάδες ροών. (Τα δύο κεντρικά στάδια του πλέγματος φαίνοται χωριστά μόνο για εννοιολογικούς λόγους, στην πραγματικότητα, υλοποιούνται σαν ένα στάδιο.)

Στο δίκτυο δρομολόγησης (το δεξιό μισό του πλέγματος Benes), τα cells που έχουν κατανεμηθεί στο ζυγό και στο περιττό υποδίκτυο πρέπει να επαναδιαταχθούν. Η επαναδιάταξη, στους μεταγωγείς εξόδου, πρέπει να γίνει χωριστά για κάθε ροή που ανήκει σε μία συγχωνευμένη ομάδα ροών. Ο λόγος είναι ότι οι συγχωνευμένες ομάδες ροών περιέχουν cells που έχουν κατανεμηθεί σε διαφορέτικούς μεταγωγείς εισόδου, ανεξάρτητα το ένα από το άλλο, και πρίν το σημείο συγχώνευσης. Επομένως, οι συγχωνευμένες ομάδες ροών από διαφορετικές εισόδους στην ίδια έξοδο πρέπει να διαχωριστούν προκειμένου η επαναδιάταξη να δουλέψει σωστά.

Ο διαχωρισμός των ομάδων ροών και η επαναδιάταξη των cells μπορούν πραγματοποιηθούν βαθμιαία, ανα-στάδιο, ή συσσωρευτικά, στο τελευταίο στάδιο του πλέγματος. Στη δεύτερη περίπτωση, δεν χρειάζεται να διαχωρίσουμε τις ροές στο banyan δρομολόγησης, οπότε, θα υπήρχαν $\frac{N}{2}, \ldots, 2, 1$ ροές οι οποίες διέρχο-



Σχήμα 2.4: *Ένα πλέγμα μεγέθους* 4×4 *και οι ομάδες ροών στις εισόδους και στις εξόδους των στοιχείων μεταγωγής για την περίπτωση που κάνουμε συγχώνευση ροών ανα-έξοδο.*

νται από τα στοιχεία μεταγωγής στα $\log_2 N$ στάδια του banyan δρομολόγησης, αντιστοίχως. Ωστόσο, κάθε συσκευή επαναδιάταξης στις πόρτες εξόδου του πλέγματος σε αυτή τη περίπτωση απαιτεί N ενταμιευτές για την επαναδιάταξη, έναν για κάθε μία από τις N (ανα-είσοδο) ροές που οδηγούνε στη συγκεκριμένη έξοδο, και κάθε ενταμιευτής πρέπει να έχει μέγεθος $O(N)$. Δεν υπάρχει λόγος να συσσωρεύσουμε τόση πολυπλοκότητα στο τελευταίο στάδιο του πλέγματος, επομένως προτιμάμε την πρώτη λύση –βαθμιαίος διαχωρισμός των ομάδων ροών και επαναδιάταξη των cells.

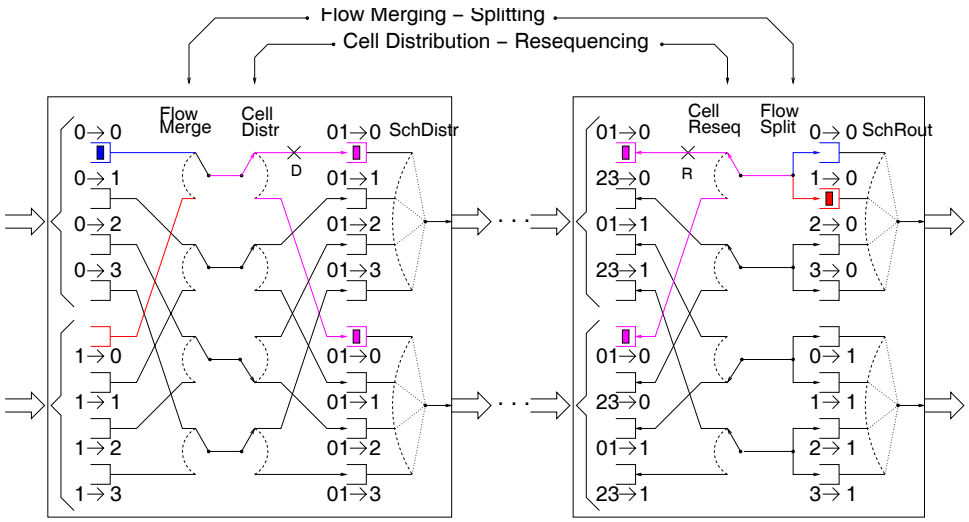
Μία εναλλακτική μέθοδος για τη μείωση του αριθμού των ροών στο κέντρο του πλέγματος είναι η *ιεραρχική συγχώνευση ροών*, η οποία ακολουθεί την αναδρομική κατασκευή του δικτύου Benes. Συγκεκριμένα, τα στοιχεία μεταγωγής στην περιφέρεια ενός $K \times K$ υποδικτύου Benes (βλέπε σχήμα 2.1) διατηρούν πληροφορία κατάστασης για τις K ροές που κατευθύνονται σε κάθε μία από τις εξόδους ή προέρχονται από κάθε μία από τις εισόδους του συγκεκριμένου υποδικτύου. Αυτή η μέθοδος μειώνει τον αριθμό των ροών ακόμη περισσότερο από την συγχώνευση ανά-έξοδο, μέχρι και σε $N/2^k$ στο στάδιο k. Ωστόσο, συγχωνεύονται ροές που κατευθύνονται σε διαφορετικούς *τελικούς* προορισμούς, οπότε χρειάζεται *ιεραρχικό backpressure* (ενότητα 2.1.3), το οποίο είναι πιο πολύπλοκο να υλοποιηθεί από το απλό backpressure. Αυτό το σχήμα χρειάζεται συνολικά $\log N$ επίπεδα backpressure. Αν και δεν χρειάζεται πλέον backpressure από-κόμβο-σε-κόμβο (hop-by-hop) στο δίκτυο δρομολόγησης, η καθυστέρηση της ανάδρασης για τον έλεγχο ροής είναι μέχρι και $2 \log N$ για το υψηλότερο επίπεδο ελέγχου ροής, επομένως ενταμιευτές μεγέθους μέχρι και $2 \log N$ χρειάζονται για πλήρη αποσχόληση των εξόδων, το οποίο είναι ανεπιθύμητο.

Εν κατακλείδει, η συγχώνευση ροών ανα-έξοδο με επαναδιάταξη ανα-στάδιο είναι πολύ απλούστερα στην υλοποίηση και έχουν ομοιόμορφο κόστος υλοποίησης $O(N)$ ανά στοιχείο μεταγωγής, επομένως χρησιμο-

ποιούμε αυτή την αρχιτεκτονική στο υπόλοιπο της εργασίας.

2.2.2 Λογική Οργάνωση των Ενταμιευτών

Στο σχήμα 2.5 φαίνεται η λογική οργάνωση των ενταμιευτών που έχει επιλεγεί για τα στοιχεία μεταγωγής των δικτύων κατανομής και δρομολόγησης, καθώς και τα απαιτούμενα ενεργά συστατικά μέρη. Στο σχήμα ακολουθείται η αρχιτεκτονική συγχώνευσης ροών και επαναδιάταξης των cells που επιλέξαμε παραπάνω. Επίσης φαίνονται οι ροές από τις εισόδους 0 και 1 σε τέσσερες διαφορετικές εξόδους του πλέγματος στο αριστερό (κατανομής) στοιχείο μεταγωγής, καθώς και οι ροές προς τις εξόδους 0 και 1 από τέσσερες διαφορετικές εισόδους του πλέγματος στο δεξί (δρομολόγησης) στοιχείο μεταγωγής. Οι ουρές FIFO που φαίνονται είναι *λογικές* ουρές, που περιέχουν *αναφορές* σε cells, τα πραγματικά cells δεν μετακινούνται μέσα στο στοιχείο μεταγωγής.



Σχήμα 2.5: *Λογική οργάνωση των ενταμιευτών ενός στοιχείου μεταγωγής στο δίκτυο κατανομής και ενός αντίστοιχου στοιχείου μεταγωγής στο δίκτυο δρομολόγησης.*

Τα στοιχεία μεταγωγής στο δίκτυο κατανομής πρέπει να κάνουν συγχώνευση ροών και κατανομή των cells, μπορούν να κάνουν αυτές τις εργασίες με οποιαδήποτε σειρά. Τα στοιχεία μεταγωγής στο δίκτυο δρομολόγησης πρέπει να κάνουν επαναδιάταξη των cells και διαχωρισμό των ροών στη σωστή, αντίστοιχη σειρά. Η κατανομή των cells μπορεί να πραγματοποιηθεί με διάφορους τρόπους, όπως εξηγούμε στην ενότητα 3.3, οι αποφάσεις της στηρίζονται σε κατάσταση που διατηρείται ανα-ροή και σκοπεύει στην βελτιστοποιήση βάσει κριτηρίων που τίθενται επίσης ανα-ροή. Η συγχώνευση των ροών πρίν από την κατανομή των cells μειώνει των αριθμό των ροών που εκτίθενται στην κατανομή των cells. Όσο μικρότερος είναι ο αριθμός των ροών, τόσο ευκολότερο είναι ο συντονισμός των (τοπικών) ανα-ροή αποφάσεων έτσι ώστε να βελτιστοποιούνε και βάσει καθολικών κριτηρίων. Επίσης, μειώνεται ο απαιτούμενος χώρος για ενταμιέυτες, όπως εξηγούμε παρακάτω σε αυτή την ενότητα. Επομένως, επιλέγουμε αυτή τη διαρρύθμιση, όπως φαίνεται και στο σχήμα 2.5.

Στις εισόδους των στοιχείων μεταγωγής, χρειάζονται ενταμιευτές ανά πόρτα εισόδου και ανά ομάδα ροών, διότι πρέπει να σταλούν στον ανάντη γείτονα credits για τους συγκεκριμένους ενταμιευτές και στο συγκεκριμένο granularity. Εκτός από αυτούς τους ενταμιευτές στις εισόδους, είναι επωφελές ή απαραίτητο να έχουμε και ενταμιευτές στις εξόδους, όπως φαίνεται στο σχήμα 2.5. Η επιλεχθείσα διαρρύθμιση απαιτεί $2 \times P \times N$ ουρές FIFO ανά στοιχείο μεταγωγής του δικτύου κατανομής. Αν τα στοιχεία μεταγωγής πραγματοποιούσαν κατανομή των cells πριν από την συγχώνευση των ροών, τότε, κάθε ένα από αυτά θα χρειαζόταν $P^2 \times N$ ουρές FIFO.

Στα στοιχεία μεταγωγής του δικτύου κατανομής (το αριστερό μισό του δικτύου), είναι επωφελές να έχουμε ενταμιευτές στις εξόδους *(α)* προκειμένου οι χρονοπρογραμματιστές στις εξόδους να λειτουργούν ανεξάρτητα ο ένας από τον άλλο, και *(β)* για αποδοτικότητα σε ορισμένες περιπτώσεις που εμφανίζονται κατά την κατανομή, όπως εξήγείται παρακάτω. Υποθέστε ότι δεν υπάρχουν ενταμιευτές στις εξόδους. Έστω, καταρχάς, ότι ο ενταμιευτής εισόδου $0 \rightarrow 0$ περιέχει ένα cell ενώ ο ενταμιευτής εισόδου $1 \rightarrow 0$ είναι άδειος (όπως στο σχήμα 2.5), και ότι ο αλγόριθμος κατανομής των cells επιτρέπει στο cell να αναχωρήσει προς οποιαδήποτε κατεύθνση. Τότε, μέχρι ένας αλλά όχι και οι δύο χρονοπρογραμματιστές του συγκεκριμένου στοιχείου μεταγωγής θα ήταν επιτρεπτό να διαλέξει την ομάδα ροών $01 \rightarrow 0$ προς εξυπηρέτηση, επομένως οι δύο χρονοπρογραμματιστές δεν θα μπορούσαν να λειτουργήσουν παράλληλα. Έστω, τώρα, ότι και οι δύο ενταμιευτές εισόδου $0 \rightarrow 0$ και $1 \rightarrow 0$ περιέχουν cells, και υποθέστε ότι ο αλγόριθμος κατανομής των cells υπαγορεύει ότι το επόμενο στη σειρά cell της ομάδας ροών $01 \rightarrow 0$ πρέπει να αναχωρήσει μέσω του επάνω χρονοπρογραμματιστή του στοιχείου μεταγωγής, Μέχρι ο επάνω χρονοπρογραμματιστής εξυπηρετήσει την ομάδα ροών $01 \rightarrow 0$, είναι δύσκολο για τον κάτω χρονοπρογραμματιστή να εξυπηρετήσει την ομάδα ροών $01 \rightarrow 0$, αν και υπάρχουν δύο cells που ανήκουν στη συγκεκριμένη ομάδα ροών, επειδή δεν γνωρίζουμε επακριβώς ποιο είναι το δεύτερο στη σειρά cell αυτής της ομάδας ροών.

Στα στοιχεία μεταγωγής του δικτύου δρομολόγησης (το δεξί μισό του δικτύου), χρειάζονται ενταμιευτές στις εισόδους για τον ίδιο λόγο όπως και στα στοιχεία μεταγωγής του δικτύου κατανομής, εκτός και αν γνωρίζουμε από που να περιμένουμε το επόμενο cell στην οποία περίπτωση χρειαζόμαστε μόνο μία θέση ενταμιευτή και το credit για αυτή τη θέση στέλνεται στον ανάντη κόμβο από τον οποίο περιμένουμε να φτάσει το επόμενο cell. Κάθε ενταμιευτής εξόδου, μαζί με τον ομόλογο του στην είσοδο του γειτονικού κατάντη μεταγωγέα, αποτελούν έναν εικονικό ενταμιευτή διπλάσιου βάθους ο οποίος χρειάζεται για την ανυπαρξία αδιεξόδων στην επαναδιάταξη των cells και για τις προτιμητέες μεθόδους κατανομής των cells, όπως θα δείξουμε στην ενότητα 3.3. Ωστόσο, οι ενταμιευτές στις εξόδους δεν είναι απαραίτητοι, μπορούν να παραλειφθούν αυξάνοντας παράλληλα το βάθος των ενταμιευτών στις εισόδους.

Κεφάλαιο 3

Κατανομή και Επαναδιάταξη των Cells

3.1 Εισαγωγή

Στο πλέγμα Benes υπάρχουν N εσωτερικά μονοπάτια για κάθε ζεύγος εισόδου-εξόδου, ένα μονοπάτι μέσα από κάθε ένα από τους N μεσαίους συνδέσμους. Η μέθοδος κατανομής των cells είναι υπεύθυνη για την επιλογή ενός κατάλληλου μονοπατιού διαμέσου του πλέγματος για κάθε εισερχόμενο cell – δηλ. η κατανομή των cells είναι μια περίπτωση του προβλήματος δυναμικής δρομολόγησης στο Benes. Ενδειαφερόμαστε για κλιμακώσιμες, κατάλληλες για hardware μεθόδους. Οπότε, μελετούμε μεθόδους της παρακάτω μορφής: κάθε μεταγωγέας εισόδου (βλέπε σχήμα 2.1) εξετάζει τα εισερχόμενα cells και ανάλογα με τη ροή που ανήκουν (δηλ. πηγή και προορισμός), και πληροφορία που διατηρείται τοπικά, καθορίζει το υποδίκτυο Benes μέσω του οποίου προωθείται το cell.

Στο υπόλοιπο αυτού του κεφαλαίου, παρουσιάζουμε καταρχάς την κλάσση των μεθόδων κατανομής των cells που θεωρούμε, το σκεπτικό πίσω από αυτή τη κλάσση, και την ανα-ροή εκ-περιτροπής (per-flow, round-robin) μέθοδο κατανομής των cells που ανήκει σε αυτή τη κλάσση, μετά αποδεικνύουμε ανυπαρξία αδιεξόδων για την παραπάνω κλάσση μεθόδων κατανομής.

3.2 Κατανομή των Cells με Μέγιστη Ανισομέρεια ανα Ροή ίση με 1

Ο αντικειμενικός σκοπός της μεθόδου κατανομής των cells είναι η εξισορρόπηση του φόρτου της κίνησης που οδηγείται στα δύο υποδίκτυα Benes. Η μέθοδος που προτείνεται στο Wide Links [21] είναι η προώθηση των cells κάθε ροής *εναλλασσόμενα* στα δύο υποδίκτυα Benes. Η παραπάνω μέθοδος μπορεί να υλοποιηθεί με κατανεμημένο τρόπο και είναι αρκετά απλή για υλοποίηση σε hardware και σε υψηλή ταχύτητα. Υπό την απουσία απωλειών των cells, η επαναδιάταξη των cells ανάγεται στην παραλαβή των cells κάθε ροής εναλλασσόμενα από τα δύο υποδίκτυα Benes. Για την παράδοση των cells εν σειρά, αρκεί να αρχίσουμε να παραδίδουμε τα cells μιας ροής από το ίδιο υποδίκτυο μέσω του οποίου το στοιχείο μεταγωγής του δικτύου

κατανομής είχε αρχίσει να προωθεί τα cells της συγκεκριμένης ροής. Η παραπάνω μέθοδος έχει επίσης προταθεί στα πλαίσια της αρχιτεκτονικής parallel packet switch (PPS) [14] προκειμένου να απαληφθεί η απαίτηση για εσωτερικό speedup.

Θεωρείστε τη μικροροή $i \rightarrow j$, το στοιχείο μεταγωγής του δικτύου κατανομής που συνδέεται στην είσοδο i και το στοιχείο μεταγωγής του δικτύου δρομολόγησης που συνδέεται στην έξοδο j. Έστω C_{ij}^e το σύνολο των cells της μικροροής $i \rightarrow j$ τα οποία βρίσκονται σε ένα μονοπάτι του πλέγματος που αρχίζει από το στοιχείο μεταγωγής του δικτύου κατανομής, ακριβώς μετά το σημείο κατανομής των cells και προς το ζυγό υποδίκτυο Benes (π.χ. το σημείο "D" του σχήματος 2.5) και τελειώνει στο στοιχείο μεταγωγής του δικτύου δρομολόγησης ακριβώς πριν από το σημείο επαναδιάταξης των cells και από το ζυγό υποδίκτυο Benes (π.χ. το σημείο "R" του σχήματος 2.5). Έστω C_{ij}^o το ομόλογο του C_{ij}^e για το περιττό υποδίκτυο Benes. Σημειώστε ότι τα σύνολα C_{ij}^e και C_{ij}^o δεν έχουν κοινά στοιχεία αφού τα αντίστοιχα μονοπάτια δεν έχουν κοινά μέρη.

Θεώρημα 1 *Για την ανα-ροή εκ-περιτροπής κατανομή των cells ισχύει:*

$$| \, |C_{ij}^e| - |C_{ij}^o| \, | \leq 1, \; \forall \, (i,j) \tag{3.1}$$

Απόδειξη: Βλέπε [22, ενότητα 3.2] ◁

Η εξίσωση 3.1 σημαίνει ότι η ανα-ροή εκ-περιτροπής κατανομή των cells ισοκατανέμει επακριβώς την κίνηση ανα-ροή μεταξύ των δύο υποδικτύων Benes. Ωστόσο, η εξίσωση 3.1 επίσης συνεπάγεται ότι:

$$| \sum_j |C_{ij}^e| - \sum_j |C_{ij}^o| \, | \leq N, \quad \forall \, input \; i \tag{3.2}$$

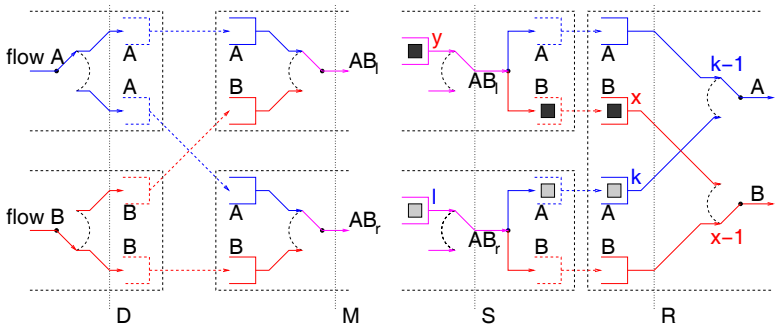
$$| \sum_i |C_{ij}^e| - \sum_i |C_{ij}^o| \, | \leq N, \quad \forall \, output \; j \tag{3.3}$$

Οι εξισώσεις 3.2 και 3.3 σημαίνουν ότι η ανα-ροή εκ-περιτροπής κατανομή των cells δεν ισοκατανέμει επακριβώς στα δύο υποδίκτυα το συνολικό όγκο της κίνησης που μπαίνει από κάθε είσοδο ή κατευθύνεται σε κάθε έξοδο του δικτύου Benes. Η απόδοση της ανα-ροή εκ-περιτροπής κατανομής των cells μελετάται αναλυτικά στο κεφάλαιο 5.

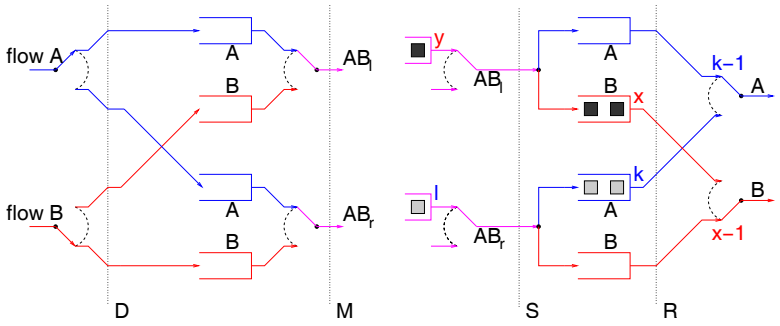
Η ανα-ροή εκ-περιτροπής κατανομή των cells δεν είναι η μοναδική μέθοδος που μελετάμε στην εργασία αυτή. Εξετάζουμε την κλάσση των μεθόδων κατανομής των cells με μέγιστη ανισομέρεια ανα ροή ίση με 1: κάθε στιγμή, οι συνολικοί αριθμοί των cells μια συγκεκριμένης ροής που έχουν προωθηθεί μέσω δύο οποιονδήποτε μονοπατιών που είναι διαθέσιμα για τη συγκεκριμένη ροή διαφέρουν το πολύ κατά 1. Στο άλλο άκρο των δύο μονοπατιών, λόγω της επαναδιάταξης τα cells "καταναλώνονται" στη σειρά, η συσκευή επαναδιάταξης στην ουσία ακολουθεί πιστά τη συσκευή κατανομής με κάποια καθυστέρηση. Προκύπτει ότι, για αυτές τις μεθόδους κατανομής, ο αριθμός των cells που είναι αποθηκευμένα προσωρινά κατά μήκος δύο μονοπατιών διαφέρουν το πολύ κατά 2. Βλέπουμε δηλ. ότι οι παραπάνω μέθοδοι κατανομής τείνουν να εξισορροπούν το φόρτο σε δύο οποιαδήποτε μονοπάτια. Όπως αποδείχθηκε παραπάνω, η ανα-ροή εκ-περιτροπής κατανομή των cells είναι μια τέτοια μέθοδος όπου ο αριθμός των cells σε δύο οποιαδήποτε μονοπάτια μπορεί να διαφέρει το πολύ κατά 1.

3.3 Ανυπαρξία Αδιεξόδων

Το πλέγμα Benes με πεπερασμένους ενταμιευτές, εσωτερικό backpressure, συγχώνευση ροών ανα-έξοδο, κατανομή και επαναδιάταξη των cells είναι ένα κατανεμημένο σύστημα με πεπερασμένους πόρους και διαμοιρασμό πόρων. Σε ένα τέτοιο σύστημα, πρέπει να σιγουρευτούμε ότι καταστάσεις αδιεξόδων είτε δε συμβαίνουν καθόλου, είτε αν συμβαίνουν, το σύστημα τις ανιχνεύει και τις επιλύει. Στην ενότητα αυτή, δείχνουμε ότι για την κλάσση των μεθόδων κατανομής με μέγιστη ανισομέρεια ανα ροή ίση με 1, καταστάσεις αδιεξόδων δεν μπορούν να συμβούν.

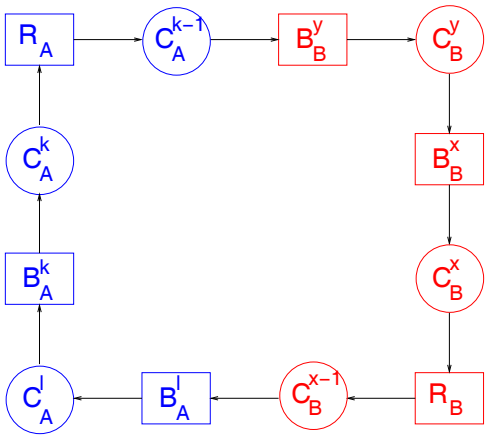


Σχήμα 3.1: *Η κατάσταση αδιεξόδου όταν η συγχώνευση των ροών προηγείται της κατανομής των cells. Τα στοιχεία μεταγωγής που συμμετέχουν φαίνονται με σιακεκομένες γραμμές και οι ροές που συμμετέχουν υποδηλώνονται ως A, B και AB. Οι αριθμοί δίπλα στους ενταμιευτές FIFO δηλώνουν τους αύξοντες αριθμούς των cells στην κεφαλή των ενταμιευτών.*



Σχήμα 3.2: *Απλοποιημένη αλλά ισοδύναμη όψη της κατάστασης αδιεξόδου που φαίνεται στο σχήμα 3.1.*

Στο σχήμα 3.1 φαίνεται ο τρόπος που μπορεί να εμφανιστεί αδιέξοδο στα στοιχεία μεταγωγής του συστήματος μας. Το σχήμα 3.2 δείχνει μια απλοποιημένη αλλά ισοδύναμη όψη του αδιεξόδου: οι συνεχόμενοι ενταμιευτές FIFO που είναι αφιερομένοι στην ίδια ροή έχουν συγχωνευτεί σε έναν ενταμιευτή FIFO με βάθος ίσο με το άθροισμα των βαθών των ξεχωριστών ενταμιευτών FIFO. Έστω c_f^s το cell της ροής "f" με αύξων αριθμό "s", B_f^s η θέση ενταμίευσης όπου βρίσκεται το cell c_f^s, και R_f η συσκευή επαναδιάταξης για τη ροή "f". Η κατάσταση αδιεξόδου είναι η παρακάτω (βλέπε σχήμα 3.3 για το γράφο δέσμευσης πόρων): *(α)* η συσκευή επαναδιάταξης για τη ροή A περιμένει το cell c_A^{k-1}, *(β)* το cell c_A^{k-1} βρίσκεται μέσα στο



Σχήμα 3.3: *Ο γράφος δέσμευσης πόρων για την κατάσταση αδιεξόδου. Οι κύκλοι αναπαριστούν cells, ενώ τα τετράγωνα αναπαριστούν πόρους που μπορεί να είναι είτε θέσεις ενταμιευτών είτε συσκευές επαναδιάταξής.*

πλέγμα κάπου πίσω από το cell c_B^y και χρειάζεται κάποιον ενταμιευτή (B_B^y) ή συσκευή επαναδιάταξης που είναι δεσμευμένη από το cell c_B^y προκειμένου να προχωρήσει – παρατηρείστε ότι το cell c_A^{k-1} δεν είναι δυνατό να βρίσκεται στο ίδιο μονοπάτι με το cell c_A^k αφού ανήκουν στην ίδια ροή, οπότε, το πλέγμα Benes δεν θα άλλαζε τη σειρά ανάμεσα στα δύο cells, *(γ)* το cell c_B^y χρειάζεται έναν ενταμιευτή που είναι δεσμευμένος από το cell c_B^x για να προχωρήσει, *(δ)* το cell c_B^x περιμένει για επαναδιάταξη από το R_B, κ.ο.κ. μέχρι που ο κύκλος κλείνει στο R_A.

Τα cells διέρχονται από το banyan κατανομής αντιμετωπίζουν μόνο καθυστερήσεις λόγω ανταγωνισμού. Στην περίπτωση που χρησιμοποιούσαμε αναδιάταξη στις τελικές εξόδους, τα cells θα διέρχονταν από το banyan δρομολόγησης αντιμετωπίζοντας μόνο καθυστερήσεις λόγω ανταγωνισμού στις εξόδους, αλλά δεν θα αντιμετώπιζαν καθυστερήσεις λόγω αναδιάταξης μέσα στο πλέγμα – σε αυτή την περίπτωση θα έπρεπε να βρούμε ικανές συνθήκες για την ανυπαρξία αδιεξόδων στις συσκευές αναδιάταξης στις τελικές εξόδους. Στην περίπτωση που χρησιμοποιούμε αναδιάταξη ανά στάδιο, τα cells αντιμετωπίζουν τις καθυστερήσεις λόγω αναδιάταξης μέσα στο πλέγμα – παρακάτω στην ουσία βρίσκουμε ικανές συνθήκες για την ανυπαρξία αδιεξόδων στις συσκευές αναδιάταξης του κάθε σταδίου.

Έστω b_D το μέγεθος των ενταμιευτών FIFO στο δίκτυο κατανομής, και b_R το μέγεθος των ενταμιευτών FIFO στο δίκτυο δρομολόγησης όπως αυτοί φαίνονται στο σχήμα 3.2.

Θεώρημα 2 *Υπό τις προυποθέσεις ότι δεν χάνονται cells, $b_D = 1$, $b_R = 2$ και ότι η προώθηση των cells γίνεται βάσει ελέγχου ροής hop-by-hop credit-based, δεν εμφανίζονται αδιέξοδα για οποιαδήποτε μέθοδος κατανομής των cells με μέγιστη ανισομέρεια ανα ροή ίση με 1.*

Απόδειξη: Βλέπε [22, ενότητα 3.3.1] ◁

Να σημειώσουμε ότι η παραπάνω απόδειξη δεν υποθέτει κάποιες ιδιότητες για τη μέθοδο χρονοπρογραμματισμού στο σημείο συγχώνευσης των ροών. Ωστόσο, υποθέτει ότι δεν χάνονται cells λόγω ηλεκτρικού θορύβου μέσα στο πλέγμα. Αυτό προφανώς είναι μια μη ρεαλιστική υπόθεση και ένα πραγματικό σύστημα

πρέπει να χρησιμοποιεί πρωτόκολλα ανεκτικά σε λάθη όπως είναι αυτά που περιγράφονται στο [9]. Η απόδειξη επεκτείνεται εύκολα και για την περίπτωση όπου συμμετέχουν περισσότερες από δύο ροές. Όσον αφορά την ανα-ροή εκ-περιτροπής κατανομή των cells, δεν εμφανίζονται αδιέξοδα ακόμα και για $b_R = 1$. Η απόδειξη συνεχίζει να ισχύει και για τις γενικές περιπτώσεις όπου *(α)* οι ενταμιευτές FIFO στο δίκτυο κατανομής, έχουν μέγεθος b_D, όπου $b_D > 1$ και *(β)* τα στοιχεία μεταγωγής είναι μεγέθους $P \times P$, όπου $P > 2$. Για την εξήγηση βλέπε [22, ενότητα 3.3.2].

Για να συνοψίσουμε, για δεδομένη μέθοδο κατανομής των cells και δεδομένο μέγεθος b_D των ενταμιευτών το δίκτυο κατανομής, μπορούμε να επιλέξουμε το b_R έτσι ώστε να μην εμφανίζοντα αδιέξοδα στο πλέγμα Benes. Στην ειδική περίπτωση όπου η μέγιστη ανισομέρεια ανα ροή είναι 1 και $b_D = 2$, το απαιτούμενο b_R είναι 3. Όσον αφορά την ανα-ροή εκ-περιτροπής κατανομή των cells, το απαιτούμενο b_R είναι ίσο με τό b_D για οποιαδήποτε τιμή του b_D.

Κεφάλαιο 4

Αποτελέσματα Προσομοιώσεων

4.1 Εισαγωγή

Αναπτύξαμε ένα μοντέλο για προσομοίωση που λειτουργεί με ακρίβεια ενός cell time προκειμένου να επαληθεύσουμε την ορθότητα της σχεδίασης και να αποτιμήσουμε την απόδοση του κάτω από διάφορα σενάρια για την κίνηση και το μέγεθος του μεταγωγέα, και προκειμένου να συγκρίνουμε διάφορες μεθόδους για την κατανομή και επαναδιάταξη των cells. Στο μοντέλο για προσομοίωση, ο χρόνος μετ'επιστροφή για ένα cell και το αντίστοιχο credit είναι 1 cell time, και κάθε ενταμιευτής που φαίνεται στο σχήμα 2.5 έχει μέγεθος 1 cell, εκτός από τους ενταμιευτές στις εισόδους των στοιχείων μεταγωγής του δικτύου δρομολόγησης που έχουν μέγεθος 2 για μια συγκεκριμένη μέθοδο κατανομής των cells.

Προσομοιώσαμε τον μεταγωγέα με ομαλή κίνηση, εκρηκτική κίνηση και κίνηση hotspot. Η ομαλή κίνηση είναι αφίξεις Bernoulli με ομοιόμορφα κατανεμημένους προορισμούς. Στην εκρηκτική κίνηση κάθε πηγή εναλασσόμενα παράγει ένα burst από cells (όλα με τον ίδιο προορισμό) που ακολουθείται από μια ανενεργή περίοδο, τα bursts και οι ανενεργές περίοδοι περιέχουν αριθμούς από cells οι οποίοι ακολουθούν την *γεωμετρική* κατανομή. Τα αποτελέσματα που παρασιάζουμε είναι για κίνηση bursty/12, το οποίο σημαίνει ότι το μέσο μέγεθος του burst είναι 12 cells, το οποίο αντιστοιχεί σε ένα από τα δημοφιλή μεγέθη της κατανομής της κίνησης IP (θεωρώντας ότι το payload των cells είναι 48 bytes). Στην κίνηση hotspot, κάθε προορισμός που ανήκει σε ένα προεπιλεγμένο σύνολο από "hot spots" λαμβάνει κίνηση (ομαλή ή εκρηκτική) η οποία συγκεντρωτικά έχει φόρτο 100%, και όλες οι πηγές συμμετέχουν εξίσου, οι υπόλοιποι προορισμοί λαμβάνουν είτε ομαλή είτε εκρηκτική κίνηση όπως έχει οριστεί παραπάνω. Τα αποτελέσματα που παρασιάζουμε είναι για κίνηση hotspot/4, δηλ. υπάρχουν 4 hot spots, συγκεκριμένα οι πόρτες 0, 1, 2, and 3.

Η καθηστέρηση που φαίνεται στα γραφήματα είναι η μέση καθυστέρηση για όλα τα cells από τη στιγμή που γενιούνται μέχρι τη στιγμή που φτάνουν στην έξοδο από όπου *έχει αφαιρεθεί το μήκος του πλέγματος* (αριθμός των σταδίων). Για παράδειγμα, τα περισσότερα από τα αποτελέσματα που παρασιάζουμε είναι για πλέγμα μεγέθους 64×64 κατασκευασμένο από στοιχεία μεταγωγής μεγέθους 4×4, οπότε το πλέγμα έχει $2 \cdot \log_4 64 = 2 \cdot 3 = 6$ στάδια, και ο αριθμός που έχει αφαιρεθεί είναι 6 cell times. Στο μοντέλο μας για τις προσομοιώσεις, ένα cell χρειάζεται 1 cell time για να διασχίσει 1 στάδιο του πλέγματος, σε ένα σύ-

στημα που κατά τα άλλα είναι άδειο. Επομένως, αφαιρώντας το μήκος του πλέγματος από την πραγματική καθυστέρηση, αναφέρουμε το άθροισμα των καθυστερήσεων στις ουρές. Σε όλα τα αποτελέσματα που αναφέρουμε, η διάρκεια των προσομοιώσεων είναι 200,000 cell times και η συλλογή στατιστικών μετρήσεων ξεκινάει μετά τα πρώτα 40,000 cell times.

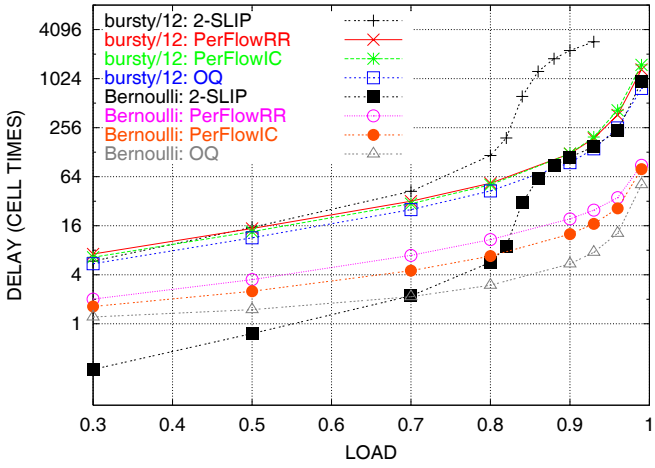
Ως μία ένδειξη όσον αφορά τη λειτουργία χωρίς εσωτερική φραγή, προσομοιώσαμε ένα πλέγμα μεγέθους 64×64 με είσοδο τη παρακάτω τεχνητή κίνηση. Σε όλα και σε καθένα cell time, μία πλήρη αντιμετάθεση που είχε επιλεγεί τυχαία εμφανιζόταν στην είσοδο του μεταγωγέα, δηλ. ο φόρτος στις εισόδους συνεχώς ήταν 100%, ενώ η συνολική κίνηση στο πλέγμα ήταν *εφικτή*, με την έννοια που έχουμε από την ενότητα 2.1.1, κατά τη διάρκεια όλων και καθενός cell time. Μετά από ένα εκατομμύριο cell times, κατ'ουσίαν δεν υπήρχαν cells στις ουρές στις εισόδους: η πλειοψηφία των VOQ's ήταν άδειες, ενώ λιγοστές περιείχαν 1 ή 2 cells η κάθε μία.

4.2 Μέθοδοι Κατανομής των Cells και Σύγκριση με OQ και iSLIP

Πειραματιστήκαμε με δύο μεθόδους κατανομής των cells, που ονομάζουμε *PerFlowRR* και *PerFlowIC*, σε ένα πλέγμα Benes μεγέθους 64×64 κατασκευασμένο από στοιχεία μεταγωγής μεγέθους 4×4. Το PerFlowRR είναι η ανα-ροή εκ-περιτροπής κατανομή των cells, όπου οι δείκτες που δείχνουν σε ποιο υποδίκτυο πρέπει να στείλουμε ένα cell αρχικοποιούνται με τυχαίο τρόπο. Το PerFlowIC (από το per-flow imbalance count) διαλέγει την πόρτα για την προώθηση ενός cell όπως εξηγείται παρακάτω: μέσα από το σύνολο από πόρτες που έχουν λάβει τα λιγότερα cells τις προκειμένης ροής μέχρι τώρα, διαλέγει την πόρτα που έχει τον μικρότερο αριθμό από *έτοιμα cells*. *Έτοιμα cells* είναι τα cells (από οποιαδήποτε ομάδα ροών) που περιμένουν σε κάποια ουρά της συγκεκριμένης πόρτας και έχουν διαθέσιμο credit από τον κατάντη κόμβο. Και οι δύο μέθοδοι έχουν μέγιστη ανισομέρεια ανα ροή ίση με 1, και, μακροπρόθεσμα, στέλνουν τον ίδιο αριθμό από cells σε κάθε διαθέσιμο μονοπάτι. Η μέθοδος PerFlowIC, ωστόσο, είναι πιο ευέλικτη κάθε φορά που ο μετρητής ανισομέρειας γίνεται 0. Επίσης κάναμε προσομοιώσεις με μεγαλύτερου μεγέθους ενταμιευτές, μέχρι και 4, και βρήκαμε ότι η απόδοση δεν είναι ευαίσθητη όσον αφορά αυτή τη παράμετρο. Τα αποτελέσματα φαίνονται στο σχήμα 4.1, για κίνηση με ομοιόμορφα κατανεμημένους προορισμούς, και στο σχήμα 4.2, για κίνηση με παρουσία hot spots.

Για ομαλή (Bernoulli) κίνηση, η συγκεκριμένη μέθοδος κατανομής των cells έχει σημασία: η μέθοδος PerFlowIC παράγει καθυστερήσεις οι οποίες είναι από 30% ως 60% χαμηλότερες σε σύγκριση με τη μέθοδο PerFlowRR. Η διαφορά είναι περισσότερο εμφανής για μέτριο φόρτο, και λιγότερο εμφανής για χαμηλό και υψηλό φόρτο. Η παρουσία ή η απούσία κίνησης hot-spot δεν επηρεάζει αυτή την πλευρά των αποτελεσμάτων. Για *εκρηκτική* κίνηση, ωστόσο, η συγκεκριμένη μέθοδος κατανομής των cells επί της ουσίας *δεν κάνει διαφορά*. Αυτό πρέπει να οφείλεται στο ότι υπάρχει μεγάλος αριθμός από cells της ίδιας ροής που καταφτάνουν συνεχόμενα: σε αυτή τη περίπτωση, η μέθοδος PerFlowIC γίνεται παρόμοια με τη μέθοδο PerFlowRR εκτός από μακροπρόθεσμα και βραχυπρόθεσμα.

Συγκρίνοντας τις καθυστερήσεις από το σχήμα 4.2 με εκείνες του σχήματος 4.1, παρατηρούμε ότι είναι



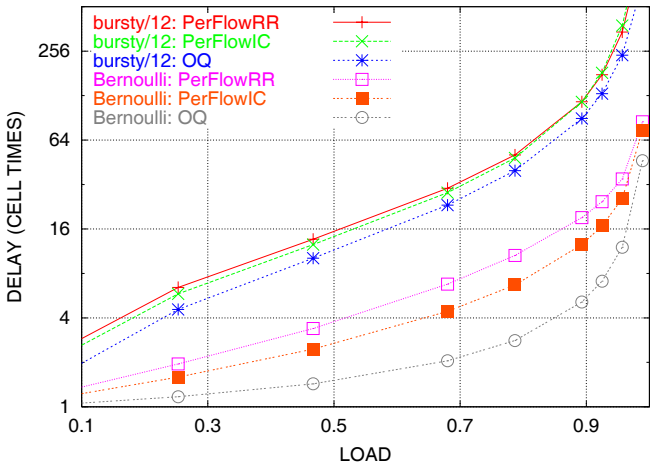
Σχήμα 4.1: *Καθυστέρηση συναρτήσει του φόρτου για ομοιόμορφους προορισμούς. Πλέγμα μεγέθους* 64×64 *κατασκευασμένο από στοιχεία μεταγωγής μεγέθους* 4×4*, πάνω καμπύλες: κίνηση bursty/12, κάτω καμπύλες: κίνηση Bernoulli. Φαίνεται επίσης το ιδανικό σύστημα με ουρές στις εξόδους (OQ) για λόγους σύγκρισης.*

σχεδόν πανομοιότυπες, το οποίο δείχνει ότι η κίνηση non-hotspot επί της ουσίας παραμένει *ανεπηρέαστη* από την παρουσία hot spots στο δίκτυο, και αποδεικνύει τις *άριστες ιδιότητες QoS* του μεταγωγέα. Στα γραφήματα δεν φαίνεται η διαπερατότητα (απασχόληση) των προορισμών hotspot (να θυμήσουμε ότι ο φόρτος σε αυτούς είναι 100%). Για ομαλή κίνηση η απασχόληση αυτών των εξόδων ήταν σε όλες τις περιπτώσεις πάνω από 99%, για εκρηκτική κίνηση, η απασχόληση κυμαινόταν από 92% εώς 98%.

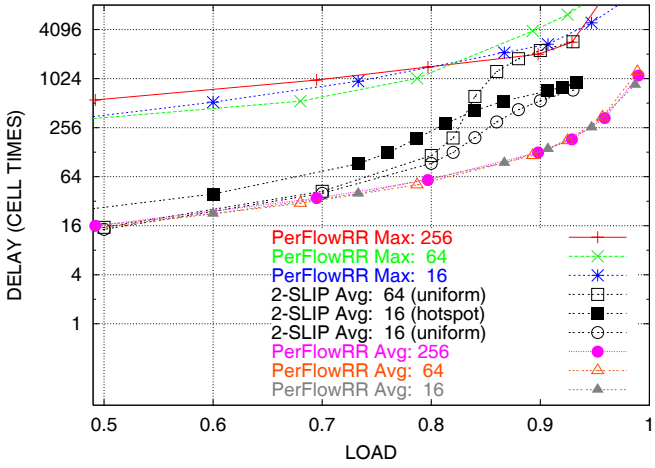
Τα σχήματα 4.2 και 4.1 επίσης δείχνουν, για λόγους σύγκρισης, την καθυστέρηση στον ιδανικό μεταγωγέα με ουρές στις εξόδους (OQ) και με κάθε ένα από τα μοντέλα κίνησης, σε κάθε τριάδα από καμπύλες, το OQ είναι η χαμηλότερη από τις τρείς. Βλέπουμε ότι, για εκρηκτική κίνηση, το πλέγμα Benes έχει χειρότερη καθυστέρηση μόνο κατά 25% εώς 50% σε σύγκριση με το OQ. Για ομαλή κίνηση, η καθυστέρηση για το πλέγμα μεταγωγής είναι μεγαλύτερη κατά έναν παράγοντα από 1.6 εώς 4, η διαφορά είναι λιγότερο εμφανής για χαμηλό φόρτο και περισσότερο εμφανής για φόρτο γύρω στο 80%.

Τέλος, συγκρίνουμε τις επιδόσεις του πλέγματος Benes με εκείνες ενός συστήματος με crossbar, VOQ's και το αλγόριθμο χρονοπρογραμματισμού crossbar 2-SLIP [23] [1]. Παρατηρούμε ότι, για φόρτο κάτω από 70%, η καθυστέρηση για τον 2-SLIP είναι μικρή, συγκρίσιμη με εκείνη για το πλέγμα Benes. Σε περίπτωση ομαλής κίνηση, η καθυστέρηση γίνεται μικρότερη και από 1, αυτό ωφείλεται στο ότι το crossbar δεν έχει εσωτερικούς ενταμιευτές, οπότε τα cells δεν χρειάζεται να περιμένουν για 1 cell time μέσα στο crossbar. Όσο ο φόρτος μεγαλώνει, γύρω στο 80%, η καθυστέρηση για τον 2-SLIP αυξάνεται σημαντικά, και για εκρηκτική κίνηση γίνεται 14 εώς 18 φορές χειρότερη από την καθυστέρηση για το πλέγμα Benes.

[1] Για τις προσομοιώσεις του αλγορίθμου 2-SLIP, χρησιμοποιήσαμε τον προσομοιωτή SIM από το πανεπιστήμιο Stanford. Το μοντέλο για την εκρηκτική κίνηση δεν υποστηρίζει φόρτο μεγαλύτερο από $\frac{b}{b+1}$, όπου b είναι το μέσο μέγεθος burst, οπότε, παρουσιάζουμε αποτελέσματα για φόρτο μέχρι 0.923%.



Σχήμα 4.2: *Καθυστέρηση για τους προορισμούς non-hotspot παρουσία κίνησης hotspot/4. Ο οριζόντιος άξονας είναι ο φόρτος προς τις εξόδους non-hotspot, οι υπόλοιπες παράμετροι είναι όπως στο σχήμα 4.1.*



Σχήμα 4.3: *Απόδοση για διάφορα μεγέθη του πλέγματος, από* 16×16 *εώς* 256×256*: μέση και μέγιστη καθυστέρηση συναρτήσει του φόρτου, για εκρηκτική κίνηση παρουσία hot spots. Τα αποτελέσματα είναι για τη μέθοδο κατανομής cells PerFlowRR.*

4.3 Εξάρτηση της Απόδοσης από το Μέγεθος του Πλέγματος

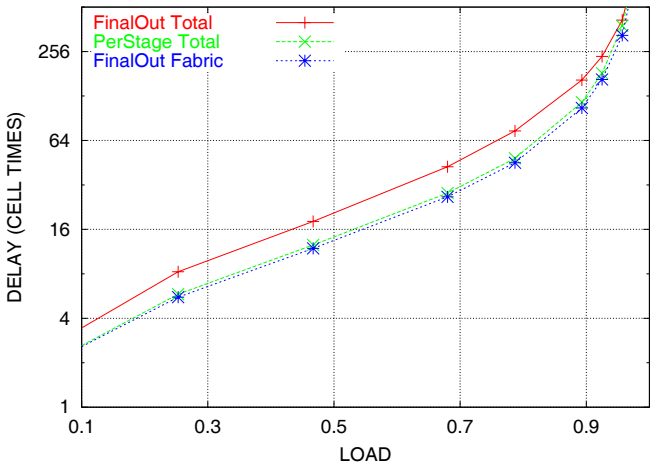
Ένα από τα πλεονεκτήμα της προτεινόμενης αρχιτεκτονικής είναι η δυνατότητα της για κλιμάκωση σε πολύ μεγάλα μεγέθη. Είναι λοιπόν σημαντικό να μην υποβαθμίζεται η απόδοση του πλέγματος όσο αυξάνει το μέγεθος του. Πειραματιστήκαμε με πλέγματα μεγέθους εώς και 256 πόρτες. Ως κίνηση στις εισόδους χρησιμοποιήσαμε το πιο "ενδειαφέρον" από τα παραπάνω μοντέλα, αφίξεις bursty/12 με προορισμούς hotspot/4.

Τα αποτελέσματα φαίνονται στο σχήμα 4.3, και δείχνουν ότι η μέγιστη καθυστέρηση γενικά αυξάνει όσο αυξάνει το μέγεθος του πλέγματος, κατα προσέγγιση από 25% εώς 75% όταν το μέγεθος του πλέγματος τετραπλασιάζεται. Ωστόσο, η *μέση* καθυστέρηση επί της ουσίας παραμένει *ανεπηρέαστη* από το μέγεθος του πλέγματος.

Επίσης, παρουσιάζουμε αποτελέσματα για μεταγωγείς που χρησιμοποιούν τον αλγόριθμο χρονοπρογραμματισμού crossbar 2-SLIP. Βλέπουμε ότι στην περίπτωση αυτή, για φόρτο κάτω από 70%, η καθυστέρηση παραμένει ανεπηρέαστη όσο μεγαλώνει το μέγεθος του πλέγματος, αλλά για υψηλότερο φόρτο γίνεται περίπου 4 φορές χειρότερη όσο μεγαλώνει το μέγεθος του πλέγματος.

4.4 Εναλακτικές Μέθοδοι Επαναδιάταξης των Cells

Όπως αναφέρθηκε στην ενότητα 2.2.1, η επαναδιάταξη των cells μπορεί να γίνει βαθμιαία, "PerStage", ή συσσωρευτικά, στο τελευταίο στάδιο του πλέγματος, "FinalOut". Από την οπτική γωνία της υλοποίησης, η ανα-στάδιο επαναδιάταξη είναι απλούστερη και φτηνότερη από την επαναδιάταξη στο τελευταίο στάδιο, αλλά η ερώτηση όσον αφορά την απόδοση παρέμενε: φαίνεται ότι η μέθοδος FinalOut επιτρέπει στα cells να προχωρήσουν ταχύτερα μέσα στο δίκτυο δρομολόγησης, και επομένως μπορεί να οδηγεί σε χαμηλότερες καθυστερήσεις. Στην πραγματικότητα, τα πράγματα είναι αντίτροφα!



Σχήμα 4.4: *Μέση καθυστέρηση για διάφορες μεθόδους επαναδιάταξής, εκρηκτική κίνηση παρουσία hot spots. Τα αποτελέσματα είναι για τη μέθοδο κατανομής cells PerFlowIC.*

Στο σχήμα 4.4 φαίνεται η μέση καθυστέρηση για τις δύο μεθόδους επαναδιάταξης, η κίνηση στις εισόδους είναι bursty/12 και hotspot/4, όπως και στην ενότητα 4.3. Για τη μέθοδο "FinalOut", δείχνουμε ξεχωριστά την καθυστέρηση των cells να περάσουν μέσα από το πλέγμα, χωρίς να έχουν ακόμη επαναδιαταχθεί ("FinalOut Fabric"), και χωριστά τη συνολική τους καθυστέρηση, η οποία περιέχει και τη διαδικασία επαναδιάταξης στο τελευταίο στάδιο του πλέγματος ("FinalOut Total"). Τα αποτελέσματα έχουν πολύ ενδειαφέρον: αν και τα cells περνούν μέσα από το πλέγμα λίγο ταχύτερα, σε σύγκριση με την περίπτωση όπου υφίστανται καθυστερήσεις στο δίκτυο δρομολόγησης λόγω της επαναδιάταξης ανα-στάδιο, η συνολική καθυστέρηση για τη μέθοδο FinalOut είναι χειρότερη.

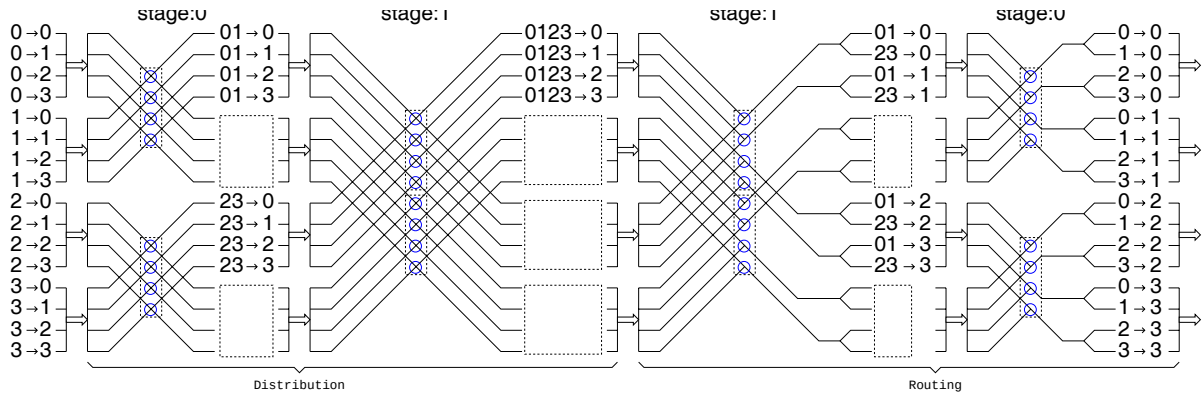
Βλέπουμε ότι το να αφήνουμε ορισμένα cells να περάσουν γρήγορα μέσα από το πλέγμα, πιο μπροστά από τη σειρά τους, χωρίς επαναδιάταξη ανα-στάδιο, φαίνεται να καταναλώνει τέτοιους πόρους του πλέγματος που, συνολικά, βλάπτει τα υπόλοιπα cells περισσότερο από ότι ωφελεί τα cells που εξέρχονται νωρίς. Συμπεραίνουμε ότι η *η επαναδιάταξη ανα-στάδιο είναι γνησίως καλύτερη* από την συσσωρευτική επαναδιάταξη στο τελευταίο στάδιο του πλέγματος, και από την οπτική γωνία του κόστους υλοποίησης και της πολυπλοκότητας όπως επίσης και από την οπτική γωνία της απόδοσης.

Κεφάλαιο 5

Αναλυτική Μελέτη της Απόδοσης

5.1 Εισαγωγή

Οι συνέπειες της ανισομέρειας που περιγράφεται από τις εξισώσεις 3.2 και 3.3 είναι εσωτερικές καθυστερήσεις οι οποίες δεν θα παρουσιάζονταν στο μοντέλο κίνησης με ρευστά ή σε μια πιο περίπλοκη μέθοδο κατανομής των cells. Συγκεκριμένα, ανεπάρκειες της μεθόδου κατανομής των cells σε ένα στάδιο του banyan κατανομής έχουν ως συνέπεια καθυστερήσεις στην πρόσβαση *(α)* στους συδέσμους εξόδου των στοιχείων μεταγωγής που ανήκουν στο συγκεκριμένο στάδιο του δικτύου κατανομής, *(β)* στους συδέσμους εισόδου των στοιχείων μεταγωγής που ανήκουν στο αντίστοιχο στάδιο του δικτύου δρομολόγησης. Ωστόσο, όπως παρατηρείται στο [24], οι παραπάνω καυθστερήσεις είναι αδύνατο να εξαλειφθούν όταν τα στοιχεία μεταγωγής του δικτύου κατανομής δουλεύουν ανεξάρτητα, όπως συμβαίνει σε ένα κατανεμημένο αλγόριθμο.



Σχήμα 5.1: *Το πλέγμα Benes μεγέθους* 4×4 *και όλες οι ροές που το διασχίζουν. Οι μπλε κύκλοι στο banyan κατανομής υποδηλώνουν τη λειτουργικότητα για συγχώνευση των ροών και κατανομή των cells, ενώ οι μπλε κύκλοι στο banyan δρομολόγησης υποδηλώνουν τη λειτουργικότητα για επαναδιάταξη των cells και διαχωρισμό των ροών. Οι μπλε κύκλοι που έχουν ομαδοποιηθεί ανήκουν σε ένα και μόνο στοιχείο μεταγωγής.*

Στην ενότητα αυτή δείχνουμε με αναλυτικές μεθόδους ότι αυτές οι εσωτερικές καθυστερήσεις δε συνεπά-

γονται περιορισμούς στην διαπερατότητα του συστήματος. Ωστόσο, για να καταλήξουμε στο αποτέλεσμα αυτό, θεωρούμε ένα απλοποιημένο μοντέλο του πλέγματος Benes με *(α)* συγχώνευση ροών ανα-ροή, *(β)* ανα-ροή εκ-περιτροπής κατανομή των cells *(γ)* ενταμιευτές με άπειρο μεγέθος στο μεσαίο στάδιο του πλέγματος, και *(δ)* καθόλου backpressure.

Καταρχήν, παρατηρείστε ότι στο μοντέλο με ρευστά και ενταμιευτές με άπειρο μεγέθος, καθόλου ανάδραση, και μία μέθοδο κατανομής των cells η οποία κατανέμει επακριβώς την ανα-ροή κίνηση στις εξόδους κάθε στοιχείου μεταγωγής του δικτύου κατανομής, οποιοσδήποτε συνδυασμός κίνησης στις εισόδους περνάει μέσα από το banyan κατανομής στο μεσαίο στάδιο του πλέγματος χωρίς καθυστερήσεις στις ουρές. Σε ένα σύστημα πακέτων, ωστόσο, τα εισερχόμενα cells υφίστανται καθυστερήσεις στις ουρές μέσα στο δίκτυο κατανομής οι οποίες είναι πεπερασμένες. Επίσης, παρατηρείστε ότι ένα μοντέλο με ρευστά όπως το παραπάνω όταν χρησιμοποιεί και συγχώνευση ροών ανα-έξοδο δεν έχει τη δυνατότητα να επιβάλλει την επιθυμητή αναλογία ανάμεσα στους ρυθμούς εξυπηρέτησης των ροών που κατευθύνονται στην ίδια έξοδο. Ο λόγος είναι ότι η εισερχόμενη κίνηση φθάνει χωρίς καθυστερήσεις στο μεσαίο στάδιο του πλέγματος, και όταν φθάσει εκεί έχει ήδη σχηματιστεί μία και μοναδική ομάδα ροών ανα έξοδο. Ωστόσο, το παραπάνω σύστημα έχει τη δυνατότητα να απομονώσει μεταξύ τους ομάδες ροών που κατευθύνονται σε διαφορετικές εξόδους: αρκεί οι χρονοπρογραμματιστές που τροφοδοτούν τους συνδέσμους στην καρδιά του πλέγματος να κατανέμουν την χωρητικότητα τους *στατικά* και *εξίσου* μεταξύ των N ομάδων ροών που εξυπηρετούν. Σε μια τέτοια περίπτωση, η κίνηση ανά τελική έξοδο που εισέρχεται στο banyan δρομολόγησης είναι το πολύ ίση με 1 cell ανά cell time σε οποιαδήποτε μονάδα του χρόνου και αν μετρηθεί, και η κίνηση περνάει μέσα από το banyan δρομολόγησης χωρίς συγκρούσεις.

Για τους παραπάνω λόγους, στην ανάλυση η οποία μελετάει επιπτώσεις από την κατανανομή των cells, χαρακτηρίζουμε την απόδοση του πλέγματος Benes μεγέθους $N \times N$ σε σχέση με την απόδοση του ιδανικού μεταγωγέα μεγέθους $N \times N$ με ουρές FIFO στις εξόδους. Η προσέγγιση που χρησιμοποιούμε και οι τεχνικές που εφαρμόζουμε κατά την ανάλυση είναι επηρεασμένες από αυτές που παραουσιάζονται στο [25] και οι οποίες είναι μια εφαορμογή της θεωρίας της άλγεβρας δικτύων (network calculus).

$A(t)$	Συσσωρευτικός αριθμός cells που έχουν *φτάσει* στο διάστημα $[0, t)$
$B(t)$	Συσσωρευτικός αριθμός cells που έχουν *αναχωρήσει* στο διάστημα $[0, t)$
$C(t)$	Συσσωρευτικός αριθμός cells *επισκέψεων* του χρονοπρογραμματιστή στην ουρά $[0, t)$
$q(t)$	Backlog της ουράς τη χρονική στιγμή t

Πίνακας 5.1: *Σημειολογία για τις μεταβλητές που περιγράφουν την κατάσταση μίας ουράς FIFO.*

Παρακάτω εξηγούμε τη βασική εξίσωση της άλγεβρας δικτύων. Θεωρείστε μια ουρά FIFO που εξυπηρετείται από έναν χρονοπρογραμματιστή τύπου work-conserving με χωρητικότητα που μπορεί να μεταβάλλεται με το χρόνο. Η κατάσταση του παραπάνω συστήματος μπορεί να περιγραφεί από ένα σύνολο μεταβλητών που φαίνονται στον πίνακα 5.1. Για μία ουρά FIFO αυτού του είδους ισχύει:

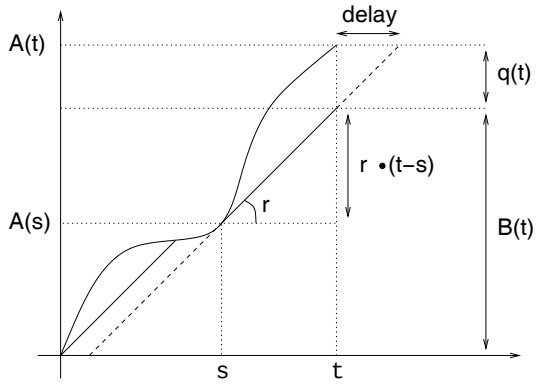
$$q(t) \quad = \quad \max_{0 \le s \le t} [A(t) - A(s) - (C(t) - C(s))]$$

$$B(t) = \min_{0 \le s \le t} [A(s) + C(t) - C(s)]$$

Όταν ο χρονοπρογραμματιστής εξυπηρετεί την ουρά FIFO με σταθερό ρυθμό r, είναι $C(t) = r \cdot t$, οποτε:

$$\begin{aligned} q(t) &= \max_{0 \le s \le t} [A(t) - A(s) - r \cdot (t - s)] \\ B(t) &= \min_{0 \le s \le t} [A(s) + r \cdot (t - s)] \end{aligned}$$

Η γραφική αναπαράσταση των εξισώσεων για την περίπτωση με σταθερό ρυθμό φαίνονται στο σχήμα 5.2.



Σχήμα 5.2: *Γραγική αναπαράσταση του backlog $q(t)$ μίας ουράς FIFO που εξυπηρετείται με σταθερό ρυθμό r.*

5.2 Banyan Κατανομής

Καταρχήν εξετάζουμε τις καθυστερήσεις στους μεταγωγείς εισόδου. Το αποτέλεσμα από αυτή την ενότητα είναι παρόμοιο με το αποτέλεσμα από το [14, Lemma 3]. Ωστόσο, το επαναλαμβάνουμε εδώ στα πλαίσια του πλέγματος Benes προκειμένουμε η αναφορά να είναι αυτοτελής. Χρησιμοποιούμε την σειμιογραφία που φαίνεται στον πίνακα 5.2.

$A_{i,j}$	Αριθμός αφίξεων στην είσοδο i του πλέγματος που προορίζονται για την έξοδο j του πλέγματος
A_j	$\sum_{i=0}^{N-1} A_{i,j}$
$A^{D(k)}_{i_k,p_k,j}$	Θεωρείστε το στοιχείο μεταγωγής με δείκτη i_k στο στάδιο $k-1$ του banyan κατανομής. Αριθμός των αφίξεων στην ουρά για την τελική έξοδο j στην πόρτα εξόδου p_k του παραπάνω στοιχείου μεταγωγής.
$A^{D(k)}_{i_k,p_k}$	$\sum_{j=0}^{N-1} A^{D(k)}_{i_k,p_k,j}$
$d^{D(k)}$	Μέγιστη καθυστέρηση ενός cell στον ενταμιευτή εξόδου ενός στοιχείου μεταγωγής στο στάδιο $k-1$ του banyan κατανομής.

Πίνακας 5.2: *Σημειολογία για τις μεταβλητές που περιγράφουν τις αφίξεις στις ουρές εξόδου των στοιχείων μεταγωγής του δικτύου κατανομής. Η σημειολογία για τις υπόλοιπες μεταβλητές (B, C, b) είναι ανάλογη.*

Χρησιμοποιούμε jitter control ή εξομοίωση καθυστερήσεων παρομοίως με το [25]. Δηλ., τα cells κρατιούνται στις ουρές εισόδου των στοιχείων μεταγωγής του δικτύου κατανομής πριν προωηθηούνε στις ουρές εξόδου, έτσι ώστε όλα τα cells να υφίστανται την ίδια καθυστέρηση $d^{D(k)}$ από τη στιγμή που εισέρχονται στην ουρά εξόδου του σταδίου $k-1$ μέχρι τη στιγμή που εισέρχονται στην ουρά εξόδου του επόμενου σταδίου του banyan κατανομής. Επί της ουσίας, το jitter control αποτρέπει τα cells που φτάνουν στο πλέγμα σε ένα δεδομένο cell time και κατευθύνονται σε μία δεδομένη έξοδο του πλέγματος να προσπεράσουν cells που κατευθύνονται στην ίδια έξοδο του πλέγματος αλλά έφτασαν σε προηγούμενα cell times. Δηλ., το jitter control αποτρέπει την αλλαγή σειράς μεταξύ των cells που κατευθύνονται στην ίδια έξοδο.

Θεώρημα 3 *Έστω ότι οι χρονοπρογραμματιστές στις εξόδους των στοιχείων μεταγωγής του δικτύου κατανομής (δηλ. τα* SchDistr*) χρησιμοποιούν τη μέθοδο εξυπηρέτησης FIFO. Τότε, για* $(1 \leq k \leq L = \log_P N)$ *ισχύει:*

$$q_{i_k,p_k}^{D(k)}(t) < k \cdot N \tag{5.1}$$

$$B_{i_k,p_k}^{D(k)}(t + k \cdot N) \geq A_{i_k,p_k}^{D(k)}(t) \tag{5.2}$$

$$d^{D(k)} < k \cdot N \tag{5.3}$$

Απόδειξη: Βλέπε [22, παράρτημα A.1] ◁

Το θεώρημα 3 παρέχει ένα άνω όριο για τη καθυστέρηση χειρότερης περίπτωσης μέσα στο banyan κατανομής και για μέγιστο απαιτούμενο χώρο για ενταμιευτές στις εξόδους των στοιχείων μεταγωγής του δικτύου κατανομής. Επιπλέον, μας λέει ότι το όριο αυτό αυξάνει γραμμικά, και όχι εκθετικά, με τον αριθμό k του σταδίου.

Έστω $L = \log_P N$, τότε το άνω όριο για τη καθυστέρηση χειρότερης περίπτωσης μέσα στο banyan κατανομής είναι της τάξης του $N \cdot L^2$ για την περίπτωση με jitter control. Το άνω όριο χωρίς jitter control είναι της τάξης του $N \cdot L$, αλλά, όπως εξηγείται στο [26], η εγκατάλειψη του jitter control μπορεί να προκαλέσει καθυστερήσεις στο πλέγμα Benes που είναι μεγαλύτερες από το πρότυπο μεταγωγέα FIFO.

5.3 Μεσαίο Στάδιο

Το μεσαίο στάδιο αναφέρεται στο τελευταίο στάδιο του banyan κατανομής το οποίο τροφοδοτεί τους N συνδέσμους στη καρδιά του πλέγματος, βλέπε σχήμα 5.1. Όλα τα στοιχεία κατανομής του δικτύου κατανομής που αποτελούνε το μεσαίο στάδιο διατηρούν το ίδιο σύνολο από ουρές εξόδου: μία ουρά για κάθε ένα από τις N τελικές εξόδους του πλέγματος.

Έστω:

$$q_j(t) = \max_{0 \leq s \leq t} [A_j(t) - A_j(s) - 1 \cdot (t - s)]$$

Τότε, το $q_j(t)$ είναι το backlog τη χρονική στιγμή t στην έξοδο j του μεταγωγέα μεγέθους $N \times N$ με ουρές FIFO στις εξόδους και στον οποίο η κίνηση που εισέρχεται είναι πανομοιότυπη με αυτήν που εισέρχεται στο πλέγμα Benes.

Θεώρημα 4 *Έστω ότι οι χρονοπρογραμματιστές που τροφοδοτούνε τους συνδέσμους στην καρδιά του πλέγματος κατανέμουν την χωρητικότητα τους στατικά και εξίσου μεταξύ των N ουρών που εξυπηρετούν. Επίσης, έστω $L = \log_P N$, $d^D = \sum_{k=1}^{L-1} d^{D(k)}$ και $d^M = (L+2) \cdot N$. Τότε, ισχύει:*

$$q^{D(L)}_{i_L,p_L,j}(t) < \frac{q_j(t-d^D)}{N} + L + 1 \quad (5.4)$$

$$B^{D(L)}_{i_L,p_L,j}(t + q_j(t-d^D) + d^M) \geq A^{D(L)}_{i_L,p_L,j}(t) \quad (5.5)$$

Απόδειξη: Βλέπε [22, παράρτημα A.2] ◁

Συνδυάζοντας τις εξισώσεις 5.2 και 5.5, παίρνουμε τον πίνακα 5.3. Από τον πίνακα 5.3 προκύπτει

Είσοδος του πλέγματος	t_0
Ενταμιευτής στο μεσαίο στάδιο	$t_0 + d^D$
Σύνδεσμος στο μεσαίο στάδιο	$(t_0 + d^D) + q_j((t_0 + d^D) - d^D) + d^M$ $= t_0 + q_j(t_0) + d^D + d^M$

Πίνακας 5.3: *Καθυστερήσεις που υφίσταται ένα cell από την είσοδο του πλέγματος μέχρι που φτάνει σε ένα μεσαίο σύνδεσμο.*

ότι ένα cell που φτάνει στην είσοδο του πλέγματος τη χρονική στιγμή t_0 θα φτάσει σε κάποιον ενταμιευτή του μεσαίου σταδίου *ακριβώς* τη χρονική στιγμή $t_0 + d^D$ – αυτή η χρονική στιγμή είναι ακριβής λόγω του jitter control –, και θα περάσει από κάποιον σύνδεσμο του μεσαίου σταδίου όχι αργότερα από την χρονική στιγμή $t_0 + q_j(t_0) + d^D + d^M$. Παρατηρείστε ότι η καθυστέρηση του ίδιου cell από την είσοδο στην έξοδο του πρότυπου μεταγωγέα FIFO είναι μεταξύ $t_0 + q_j(t_0) - (N-1)$ και $t_0 + q_j(t_0)$ εξ'ορισμού της μεθόδου ρονοπρογραμματισμού FIFO. Οπότε, η καθυστέρηση κάθε cell από την είσοδο του πλέγματος σε κάποιο σύνδεσμο του μεσαίου σταδίου είναι μεγαλύτερη από την καθυστέρηση του ίδιου cell μέσα από τον πρότυπο μεταγωγέα με ουρές FIFO στις εξόδους κατά έναν όρο το πολύ ίσο με $d^D + d^M + N = O(N \cdot L^2)$. Να σημειώσουμε ότι ο αντίστοιχος της καθυστέρησης στο [25] είναι της τάξης του $O(N^2)$. Ο λόγος για τη μειωμένη επιβάρυνση στην καθυστέρηση είναι ότι το πλέγμα Benes επιτρέπει πιο ευέλικτα χρονοδιαγράμματα για τα cells. Τέτοια χρονοδιαγράμματα είναι εφικτά λόγω των ενταμιευτών που υπάρχουν εσωτερικά στα στοιχεία μεταγωγής και υπολογίζονται επίσης με κατανεμημένο τρόπο.

5.4 Banyan Δρομολόγησης

Στην ενότητα αυτή, αναλύουμε την καθυστέρηση που υφίστανται τα cells στα $\log_P N$ στάδια του banyan δρομολόγησης. Χρησιμοποιούμε τη σημειολογία που φαίνεται στον πίνακα 5.4.

Έστω $* \rightarrow j$ το σύνολο από ροές που αποτελείται από όλες τις ροές που κατευθύνονται στην τελική έξοδο j και περνάνε μέσα από ένα συγκεχριμένο στοιχείο μεταγωγης. Στην ανάλυση, υποθέτουμε ότι οι χρονοπρογραμματιστές στις εξόδους των στοιχείων μεταγωγής του δικτύου δρομολόγησης κατανέμουν την χωρητικότητα τους *στατικά* και *εξίσου* μεταξύ των συνόλων από ροές που παριστάνονται ως $* \rightarrow j$.

	Θεωρείστε το στοιχείο μεταγωγής με δείκτη i_k στο στάδιο $k-1$ του banyan δρομολόγησης.
$A^{R(k),rsq}_{i_k,q_k,j}$	Αριθμός των αφίξεων στην ουρά επαναδιάταξης για την τελική έξοδο j στην πόρτα εισόδου q_k του παραπάνω στοιχείου μεταγωγής.
$A^{R(k)}_{i_k,p_k,j}$	Αριθμός των αφίξεων για την τελική έξοδο j στην πόρτα εξόδου p_k του παραπάνω στοιχείου μεταγωγής.
$d^{R(k)}$	Μέγιστη καθυστέρηση ενός cell στον ενταμιευτή εξόδου ενός στοιχείου μεταγωγής στο στάδιο $k-1$ του banyan δρομολόγησης.

Πίνακας 5.4: *Σημειολογία για τις μεταβλητές που περιγράφουν τις αφίξεις στις ουρές των στοιχείων μεταγωγής του δικτύου δρομολόγησης. Η σημειολογία για τις υπόλοιπες μεταβλητές (B, C, b) είναι ανάλογη.*

Σημειώστε ότι οι είσοδοι των στοιχείων μεταγωγής του δικτύου δρομολόγησης πρέπει να χειρίζονται τις διαθέσιμες ροές κάθε μία χωριστά προκειμένου να επαναδιατάσουν τα cells.

Θεώρημα 5 *Έστω $d^{mid} = d^D + d^M + N$. Έστω ότι οι χρονοπρογραμματιστές στις εξόδους των στοιχείων μεταγωγής του δικτύου δρομολόγησης (i.e.* SchRout*) κατανέμουν την χωρητικότητα τους στατικά και εξίσου μεταξύ των συνόλων από ροές $* \to j$. Τότε, για $(L = \log_P N \geq k \geq 1)$ ισχύει:*

$$\bigcup_{i_L,q_L} B^{R(L),rsq}_{i_L,q_L,j}(t' + d^{mid}) \quad \supseteq \quad B_j(t') \tag{5.6}$$

$$d^{R(k)} \quad = \quad 2 \cdot k \cdot N + P^k \tag{5.7}$$

$$\bigcup_{i_k,p_k} B^{R(k)}_{i_k,p_k,j}(t' + d^{mid} + \sum_{s=L}^{k} d^{R(s)}) \quad \supseteq \quad B_j(t') \tag{5.8}$$

Στην παραπάνω εξίσωση, το i_L, q_L διατρέχει όλες τις πόρτες εισόδου στο στάδιο $L-1$ του banyan δρομολόγησης, ενώ το i_k, p_k διατρέχει όλες τις πόρτες εξόδου στο στάδιο $L-1$ του banyan δρομολόγησης.

Απόδειξη: Βλέπε [22, παράρτημα A.3] ◁

Από το θεώρημα 5 προκύπτει ότι ένα cell διέρχεται μέσα από τους ενταμιευτές για επαναδίαταξη στο στάδιο $L-1$ του banyan δρομολόγησης με επιπλέον καθυστέρηση το πολύ ίση με d^{mid} σε σχέση με την καθυστέρηση μέσα από τον πρότυπο μεταγωγέα FIFO, και διέρχεται μέσα από το banyan δρομολόγησης με επιπλέον καθυστέρηση το πολύ ίση με d^R, όπου

$$d^R = \sum_{s=L}^{1} d^{R(s)} = \sum_{s=L}^{1} (2 \cdot s \cdot N + P^s) = 2 \cdot N \cdot \sum_{s=L}^{1} s + \sum_{s=L}^{1} P^s = O(N \cdot \log_P^2 N) + O(N)$$

Οπότε, η επιπλέον καθυστέρηση κάθε cell διαμέσου του πλέγματος Benes σε σχέση με την καθυστέρηση διαμέσου του πρότυπου μεταγωγέα FIFO είναι:

$$d^D + d^M + d^R + N \quad < \quad \frac{3}{2} \cdot N \cdot L^2 + \frac{3}{2} \cdot N \cdot L + 5 \cdot N$$

Κεφάλαιο 6

Συμπεράσματα και Μελλοντική Εργασία

Δείξαμε έναν αποδοτικό τρόπο για την κλιμάκωση των μεταγωγέων πακέτων σε πολύ μεγάλο αριθμό από πόρτες, διατηρώντας παράλληλο τη λειτουργία χωρίς εσωτερική φραγή και την υψηλή ποιότητα υπηρεσίας. Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας το δίκτυο Benes, την χαμηλότερου κόστους τοπολογία που δεν έχει εσωτερική φραγή. Μνήμες με τους μεγάλους ενταμιευτές χρειάζονται μόνο στις εισόδους του συστήματος, για την υλοποίηση εικονικών ουρών εξόδου (VOQ), ο αριθμός τους κλιμακώνεται γραμμικά συναρτήσει του μεγέθους του συστήματος, ο αριθμός των ουρών που πρέπει να διατηρούνται σε κάθε μνήμη επίσης κλιμακώνεται γραμμικά, ενώ η διαπερατότητα κάθε μνήμης παραμένει σταθερή. Στο πλέγμα Benes χρησιμοποιείται εσωτερικό backpressure, προκειμένου να έχουμε: *(α)* χαμηλού κόστους στοιχεία μεταγωγής, αφού χρειάζονται μόνο on-chip μνήμη για ενταμιευτές, *(β)* μηδενική απώλεια cells μέσα στο πλέγμα μεταγωγής, αν και οι μνήμες για ενταμιευτές είναι μικρές, *(γ)* χαμηλού κόστους σύστημα, αφού το πλέγμα δεν χρειάζεται εσωτερικό speedup, *(δ)* χαμηλού κόστους σύστημα, αφού το πλέγμα δεν χρειάζεται περίσσεια μονοπάτια προκειμένου να χειριστεί συγχρούσεις μεταξύ των cells με τη μέθοδο του deflection routing, *(ε)* χαμηλού κόστους σύστημα, αφού δεν χρειάζεται κεντρικοποιημένος χρονοπρογραμματιστής, και όλος ο χρονοπρογραμματισμός και ο συντονισμός γίνεται με κατανεμημένο τρόπο, και *(στ)* υψηλή απόδοση του συστήματος και υψηλή ποιότητα υπηρεσίας, ακόμα και με χαμηλό κόστος για το σύστημα όπως εξηγείται λεπτομερώς παραπάνω.

Για την επίτευξη όλων των παραπάνω, έπρεπε να επεκτείνουμε την γνωστή αρχιτεκτονική με backpressure ανα-ροή έτσι ώστε να μπορεί να εφαρμοστεί δρομολόγηση μέσω πολλαπλών μονοπατιών (αντίστροφη πολυπλεξία) και επαναδιάταξη των cells, διατηρώντας παράλληλα το κόστος του συστήματος σε αποδεκτά επίπεδα. Αυτό το πετύχαμε χρησιμοποιώντας ένα κατάλληλο σχήμα συγχώνευσης ροών που διατηρεί το κόστος του backpressure ίσα με $O(N)$ ανά στοιχείο μεταγωγής. Αποδείξαμε ανυπαρξία αδιεξόδων για μία ευρεία κλάσση από μεθόδους για την κατανομή των cells σε πολλαπλά μονοπάτια. Τέλος, χρησιμοποιώντας προσομοιώσεις με ακρίβεια cell time, *(α)* δείξαμε ότι η επαναδιάταξη ανα-στάδιο είναι προτιμότερημ *(β)*

βρήκαμε ότι η κατανομή των cells βάσει μετρητών ανισομέρειας πετυχαίνει μικρότερης καθυστερήσεις από την εκ-περιτροπής κατανομή, αλλά για εκρηκτική κίνηση η διαφορά γίνεται αμελητέα, *(γ)* παρατηρήσαμε ότι η καθυστέρηση για εκρηκτική κίνηση είναι μόνο 25 με 50 % υψηλότερη σε σχέση με το ιδανικό σύστημα με ουρές στις εξόδους, και *(δ)* δείξαμε ότι η καθυστέρηση των μη-συμφορημένων ροών παραμένει ανεπηρέαστη από την παρουσία συμφορημένης κίνησης σε πόρτες εξόδου που είναι oversubscribed, το οποίο αποδεικνύει τις άριστες ιδιότητες ποιότητας υπηρεσίας του συστήματος.

Η παρούσα εργασία περιγράφει την αρχιτεκτονική του συστήματος, το σκεπτικό της και τον βασικό τρόπο λειτουργίας της. Σκοπεύουμε να επεκτείνουμε αυτή τη δουλειά σε δύο περιοχές: θέματα που αφορούν πρακτικές υλοποιήσεις και μοντελοποίηση και μελέτη με αναλυτικές μεθόδους. Όσον αφορά τα πρακτικά θέματα, σκπεύουμε να ασχοληθούμε με *(α)* βελτιστοποιήσεις για πλέγματα 3 σταδίων, και *(β)* απλοποίηση της λογικής οργάνωσης των ενταμιευτών στα στοιχεία μεταγωγής. Παρόλο που έχουμε κάνει δουλειά στον χαρακτηρισμό με αναλυτικές μεθόδους της επίδρασης που έχει η κατανομή των cells, αυτή η δουλειά πρέπει να επεκταθεί, οπότε σκοπεύουμε: *(α)* να πραγματοποιήσουμε μία ανάλυση βάσει μεθόδων από τη θεωρία network calculus που λαμβάνουν υπόψη τους το έλεγχο ροής [27], και *(β)* να μελετήσουμε μεθόδους κατανομής των cells που προέρχονται από θεωρητικά αποτελέσματα [28] [29]. Τέλος, σημαντικά θέματα μελλοντικής εργασίας αποτελούν η αρχιτεκτινική χρονοπρογραμματισμού και η υποστήριξη για multicasting.

Βιβλιογραφία

[1] M. Marcus, "The Theory of Connecting Networks and their Complexity: a Review," *IEEE Proceedings*, vol. 65, no. 9, pp. 1263–1271, Sept. 1977.

[2] C.-L. Wu and T.-Y. Feng, "On a Class of Multistage Interconnection Networks," *IEEE Trans. on Computers*, vol. 29, no. 8, pp. 694–702, Aug. 1980.

[3] V. Benes, "Optimal Rearrangeable Multistage Connecting Networks," *Bell Systems Technical Journal*, vol. 43, no. 7, pp. 1641–1656, July 1964.

[4] K. Batcher, "Sorting Networks and their Applications," in *AFIPS Proc. 1968 Spring Joint Computer Conf.*, 1968, vol. 32, pp. 307–314.

[5] A. Huang and S. Knauer, "Starlite: A Wideband Digital Switch," in *Proc. IEEE GLOBECOM '84 Conf., Atlanta GA USA*, Dec. 1984, pp. 121–125.

[6] J. Giacopelli, J. Hickey, W. Marcus, W. Sincoskie, and M. Littlewood, "Sunshine: a High Performance Self-Routing Broadband Packet Switch Architecture," *IEEE-JSAC*, vol. 9, no. 8, pp. 1289–1298, Oct. 1991.

[7] G. Kornaros, D. Pnevmatikatos, P. Vatsolaki, G. Kalokerinos, C. Xanthaki, D. Mavroidis, D. Serpanos, and M. Katevenis, "ATLAS I: Implementing a Single-Chip ATM Switch with Backpressure," *IEEE Micro*, vol. 19, no. 1, pp. 30–41, Jan. 1999, http://archvlsi.ics.forth.gr/atlasI/hoti98/.

[8] F. Chiussi, D. Khotimsky, and S. Krishnan, "Generalized Inverse Multiplexing for Switched ATM Connections," in *Proc. IEEE GLOBECOM Conf., Australia*, Nov. 1998, pp. 3134–3140, http://www.bell-labs.com/org/113480/Papers/fabio-globecom98B.ps.

[9] D. Khotimsky, "A Packet Resequencing Protocol for Fault-tolerant Multipath Transmission with Non-Uniform Traffic Splitting," in *Proc. IEEE GLOBECOM Conf., Brasil*, Dec. 1999, pp. 1283–1289, http://www.bell-labs.com/org/113480/Papers/dkh-globecom99.ps.

[10] Sundar Iyer, Amr A. Awadallah, and Nick McKeown, "Analysis of a Packet Switch with Memories Running Slower than the Line-Rate," in *IEEE INFOCOM*, Mar. 2000, http://klamath.stanford.edu/~sundaes/Papers/infocom2000.pdf.

[11] J. Turner, "Design of a Broadcast Packet Switching Network," *IEEE Transactions on Communications*, vol. 36, no. 6, pp. 734–743, June 1988.

[12] "IBM PowerPRS Q-64G Packet Routing Switch Datasheet," Dec. 2001, http://www.ibm.com/-chips/techlib/techlib.nsf/products/PowerPRS_Q-64G_Packet_Routing_Switch.

[13] L. G. Valiant and G. J. Brebner, "Universal Schemes for Parallel Communication," in *ACM STOC*, 1981, pp. 263–277.

[14] Sundar Iyer and Nick McKeown, "Making Parallel Packet Switches Practical," in *IEEE INFOCOM*, Mar. 2001, http://klamath.stanford.edu/~sundaes/Papers/infocom2001.pdf.

[15] J. Duncanson, "Inverse Multiplexing," *IEEE Communications Magazine*, vol. 32, no. 4, pp. 34–41, Apr. 1994.

[16] William J. Dally, "Virtual Channel Flow Control," *IEEE Transactions on Parallel and Distributed Systems*, vol. 3, no. 2, pp. 194–205, Mar. 1992.

[17] C. Ozveren, R. Simcoe, and G. Varghese, "Reliable and Efficient Hop-by-Hop Flow Control," *IEEE Journal on Selected Areas in Communication*, vol. 13, no. 4, pp. 642–650, May 1995.

[18] Quantum Flow Control Alliance, "Quantum Flow Control: A cell-relay protocol supporting an Available Bit Rate Service," July 1995, version 2.0.

[19] M. Katevenis, D. Serpanos, and E. Spyridakis, "Credit-Flow-Controlled ATM for MP Interconnection: the ATLAS I Single-Chip ATM Switch," in *HPCA*, Feb. 1998, pp. 47–56, http://archvlsi.ics.forth.gr/atlasI/atlasI_hpca98.ps.gz.

[20] M. Katevenis, "Fast Switching and Fair Control of Congested Flow in Broad-Band Networks," *IEEE Journal on Selected Areas in Communication*, vol. 5, no. 8, pp. 1315–1326, Oct. 1987.

[21] Manolis Katevenis, "Wide Links: $2 \cdot logN$-Stage Non-Blocking Networks with In-Order Packet Delivery," Internal Note, June 1995.

[22] Georgios Sapountzis, "Benes Switching Fabrics with $O(N)$-Complexity Internal Backpressure," M.S. thesis, University of Crete, Sept. 2002, English version.

[23] Nick McKeown, "iSLIP: A Scheduling Algorithm for Input-Queued Switches," *IEEE/ACM Transactions on Networking*, vol. 7, no. 2, Apr. 1999, http://tiny-tera.stanford.edu/~nickm/papers/-ToN_April_99.pdf.

[24] D. A. Khotimsky and S. Krishnan, "Stability Analysis of a Parallel Packet Switch with Bufferless Input Demultiplexors," in *ICC 2001*, June 2001.

[25] Cheng-Shang Chang, Duan-Shin Lee, and Yi-Shean Jou, "Load Balanced Birkhoff-von Neumann Switches, Part II: Multi-stage Buffering," *Computer Communications, special issue on "Current Issues in Terabit Switching"*, 2001.

[26] D. A. Khotimsky and S. Krishnan, "Towards the Recognition of Parallel Packet Switches," in *GBN Workshop*, Apr. 2001, http://www.cs.columbia.edu/~sk/research/papers/pps-gbn01.html.

[27] R. Agrawal, R. L. Cruz, C. Okino, and R. Rajan, "Performance Bounds for Flow Control Protocols," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 310–323, June 1999.

[28] Allan Borodin, Jon Kleinberg, Prabhakar Raghavan, Madhu Sudan, and David P. Williamson, "Adversarial Queueing Theory," in *Proc.* 28^{th} *ACM STOC*, May 1996, pp. 376–385.

[29] A. Richa M. Mitzenmacher and R. Sitaraman, "The Power of Two Random Choices: A survey of the Techniques and Results," in *Handbook of Randomized Computing, P. Pardalos, S. Rajasekaran, and J. Rolim, Eds.* Kluwer, 2000.