Η Υλοποίηση του Τηλέγραφου: μια Αρχιτεκτονική για Επικοινωνία Υψηλής Ταχύτητας

Παναγιώτα Βατσολάκη, Γιώργος Καλοκαιρινός, Μανόλης Στρατάκης,
Χαρά Ξανθάκη, Μιχάλης Λυγεράκης, Γιώργος Κορνάρος,
Απόστολος Δόλλας [1], Γιώργος Παπαδουράκης [2], Μανόλης Κατεβαίνης [3]

Ινστιτούτο Πληροφορικής (ΙΠ)
Ιδρυμα Τεχνολογίας και Ερευνας (ΙΤΕ)
Επιστημονικό και Τεχνολογικό Πάρκο, Τ.Θ. 1385 Ηράκλειο, Κρήτη, GR-711 10
E-mail: vatsola@ics.forth.gr; τηλ.: +30 (81) 391665; fax: +30 (81) 391661.

ΠΕΡΙΛΗΨΗ

Παρουσιάζουμε την υλοποίηση του *Τ*ηλέγραφου, μιας αρχιτεκτονικής για επικοινωνία υψηλής ταχύτητας παράλληλων και κατανεμημένων συστημάτων. Ο Τηλέγραφος χρησιμοποιεί διασυνδέσεις μεταξύ επεξεργαστών και δικτύου που βασίζονται στην πράξη της μακρινής εγγραφής για να υποστηρίξει επικοινωνία με χαμηλό κόστος και μικρή καθυστέρηση. Το δίκτυο του Τηλέγραφου κατασκευάζεται από μεταγωγείς που μπορούν να συνδεθούν μεταξύ τους σε οποιαδήποτε τοπολογία εξασφαλίζοντας υψηλό εύρος ζώνης. Η πρωτότυπη υλοποίηση αυτής της αρχιτεκτονικής που διασυνδέει σταθμούς εργασίας DEC Alpha βασισμένη σε τεχνολογία χαμηλής ολοκλήρωσης, FPGAs και μνήμες RAM, έχει κατασκευαστεί και ελεγχθεί επιτυχώς στο εργαστήριο μας. Η δεύτερη υλοποίηση, με την μορφή ενός ολοκληρωμένου κυκλώματος έχει ήδη σταλεί για κατασκευή. Η Αρχιτεκτονική του Τηλέγραφου, απλή αλλά ταυτόχρονα ισχυρή, δίνει την δυνατότητα κατασκευής πολυεπεξεργαστών με χαμηλό κόστος.

*1. Ε*ισαγωγή

Στην εργασία αυτή παρουσιάζουμε την πρώτη υλοποίηση του *Τ*ηλέγραφου, μιας αρχιτεκτονικής για επικοινωνία υψηλής ταχύτητας. Στα πλαίσια του Τηλέγραφου, σχεδιάσαμε και υλοποιήσαμε μεταγωγείς δικτύου (switches) καθώς και διασυνδέσεις επεξεργαστή–δικτύου (network – processor interfaces) που επιτρέπουν την παράλληλη επεξεργασία σε σταθμούς εργασίας συνδεδεμένους μέσω ενός υπερταχέος δικτύου. Η αρχική μας πρωτότυπη υλοποίηση, ο Τηλέγραφος Ι, αποτελείται από τυπωμένα κυκλώματα διασύνδεσης με το δίκτυο που μπαίνουν στον δίαυλο εισόδου-εξόδου Turbochannel των σταθμών εργασίας DEC Alpha, και τυπωμένα κυκλώματα μεταγωγής που συνδέονται μέσω καλωδιοταινιών μεταξύ τους αλλά και με τις διασυνδέσεις δικτύου για να σχηματίσουν ένα δίκτυο υψηλής ταχύτητας. Ο Τηλέγραφος Ι έχει ήδη κατασκευαστεί χρησιμοποιώντας τεχνολογία χαμηλής ολοκλήρωσης (FPGAs and SRAMs) με

1 Ο Α. Δόλλας είναι και με το Τμήμα Ηλεκτρονικής και Μηχανικών Η/Υ του Πολυτεχνείου Κρήτης.
2 Γ. Παπαδουράκης είναι και με το Γενικό Τμήμα, Σχολή Τεχνολογικών Εφαρμογών, ΤΕΙ Ηρακλείου.
3 Ο Μ. Κατεβαίνης είναι και με το Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης.

σκοπό την γρήγορη υλοποίηση (rapid prototyping). Ο μεταγωγέας του Τηλέγραφου ΙΙ έχει επίσης υλοποιηθεί σε ένα ολοκληρωμένο κύκλωμα τεχνολογίας ASIC, πετυχαίνοντας καλύτερη απόδοση.

Το όνομα της αρχιτεκτονικής είναι "Τηλέγραφος" – από τις λέξεις "τηλέ" και "γράφω "– γιατί η πράξη της μακρινής εγγραφής *(remote write)* είναι ένα βασικό στοιχείο πάνω στο οποίο χτίζεται ένα γρήγορο σύστημα επικοινωνίας, όπως περιγράφεται στο τμήμα 2. Η Διασύνδεση Δικτύου του Τηλέγραφου η οποία υλοποιεί την πράξη της μακρινής εγγραφής όπως και άλλες πράξεις σε μακρινή μνήμη, παρουσιάζεται στο μέρος 3.

Το δίκτυο που συνδέει τους σταθμούς εργασίας κατασκευάζεται από μεταγωγείς που χρησιμοποιούν έλεγχο ροής βασισμένο σε "εισιτήρια" (credit-based). Μέσα στους μεταγωγείς, η αποθήκευση των πακέτων έχει οργανωθεί σαν ενιαία κεντρική μνήμη (shared buffering) η οποία δίνει βέλτιστη χρήση του διαθέσιμου εύρους ζώνης των συνδέσμων. Τα δύο αυτά χαρακτηριστικά εξασφαλίζουν την υψηλή και σταθερή απόδοση του μεταγωγέα ακόμα και σε περιπτώσεις υπερβολικού φόρτου στο δίκτυο. Η οργάνωση του μεταγωγέα παρουσιάζεται στο μέρος 4, ακολουθούμενο από την περιγραφή της υλοποίησης των μεταγωγέων και των διασυνδέσεων δικτύου του Τηλέγραφου στο τμήμα 5.

2. *Η Α*ρχιτεκτονική του *Τ*ηλέγραφου βασισμένη στην πράξη της μακρινής εγγραφής

Παρότι τα σημερινά δίκτυα παρέχουν ολοένα και μεγαλύτερο εύρος ζώνης, η καθυστέρηση των κλασικών αρχιτεκτονικών επικοινωνίας μέσω μηνυμάτων δεν μειώνεται με αντίστοιχο ρυθμό. Αυτό οφείλεται στο γεγονός ότι το μεγαλύτερο ποσοστό χρόνου για την μετάδοση ενός μηνύματος δαπανάται μέσα στο λειτουργικό σύστημα και μέσα σε πρωτόκολα επικοινωνίας. Ο Τηλέγραφος, αντίθετα από τις κλασικές αρχιτεκτονικές επικοινωνίας, δίνει ένα σύστημα διασύνδεσης με το δίκτυο με χαμηλή καθυστέρηση και χαμηλό επιπλέον κόστος. Με την πράξη της μακρινής εγγραφής ο χρήστης μπορεί να στείλει μηνύματα κατευθείαν στο δίκτυο, και να πάρει μηνύματα από το δίκτυο, χωρίς να χρειάζεται παρέμβαση του λειτουργικού συστήματος.

Η αρχιτεκτονική του Τηλέγραφου, που περιγράφεται αναλυτικά στο [Kate94], υποστηρίζει ένα ενιαίο χώρο φυσικών διευθύνσεων μεταξύ όλων των επεξεργαστών (υπολογιστών) του δικτύου. Η μετάφραση ιδεατών διευθύνσεων σε φυσικές χρησιμοποιείται για την αντιστοίχιση τοπικών ιδεατών διευθύνσεων ενός επεξεργαστή σε φυσικές διευθύνσεις του δικτύου. Η τελική φυσική διεύθυνση μπορεί να βρίσκεται είτε στην τοπική μνήμη του επεξεργαστή είτε σε μια μακρινή μνήμη ενός άλλου επεξεργαστή. Αυτό προσδιορίζεται με τον εξής τρόπο: τα περισσότερο σημαντικά bits (MS) κάθε αριθμού φυσικής σελίδας δηλώνουν την ταυτότητα του επεξεργαστή, ενώ τα λιγότερο σημαντικά (LS) bits δηλώνουν τον αριθμό σελίδας στην μνήμη αυτού του επεξεργαστή. Η υλοποίηση των παραπάνω γίνεται ως εξής: όταν η ιδεατή διεύθυνση (virtual address) μιας εντολής αποθήκευσης *(store)* που εκτελείται από τον επεξεργαστή μεταφραστεί σε μια φυσική διεύθυνση του χώρου διευθύνσεων της συσκευής εισόδου-εξόδου του Τηλέγραφου, η Διασύνδεση Δικτύου του Τηλέγραφου αναλαμβάνει να κατασκευάσει ένα πακέτο με τη διεύθυνση και τα δεδομένα και να το στείλει στον κόμβο προορισμού (που καθορίζεται από τα MS bits της φυσικής σελίδας). Η Διασύνδεση Δικτύου στον κόμβο προορισμού εκτελεί την (τοπική εκεί) εγγραφή στην μνήμη στον χώρο διευθύνσεων του δέκτη χρήστη.

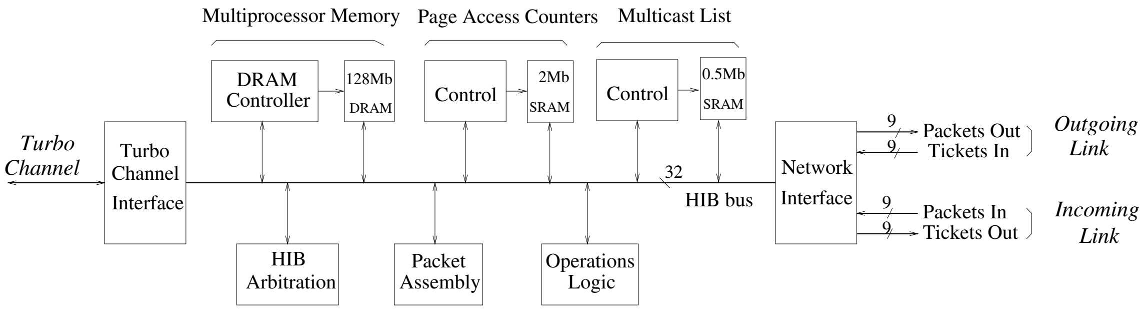
Με τον παραπάνω τρόπο, μπορούν να στέλνονται μηνύματα σαν μια ακολουθία από μακρινές εγγραφές, δηλαδή, σαν μια ακολουθία εντολών αποθήκευσης σε κατάλληλες ιδεατές διευθύνσεις. Η προστασία των μηνύματων βασίζεται στον ήδη υπάρχοντα μηχανισμό προστασίας μέσω της μετάφρασης διευθύνσεων. Κατά συνέπεια, η ανταλλαγή μηνυμάτων μπορεί να γίνει σε επίπεδο χρήστη, χωρίς να χρειάζεται καμία χρονοβόρα κλήση του λειτουργικού συστήματος. Επιπλέον, με αυτό το σχήμα αποκλείονται οι υπερχειλίσεις του ενταμιευτή του δέκτη, και παρέχεται αυτόματη ανασύνθεση του μηνύματος από τα πολλαπλά πακέτα του. Επειδή δεν χρειάζεται αντιγραφή του μηνύματος μέσα στον ίδιο κόμβο, επιταχύνεται η διαδικασία ανταλλαγής μηνυμάτων και ταυτόχρονα απλοποιείται η υλοποίηση της διασύνδεσης με το δίκτυο.

Πέρα από την γρήγορη ανταλλαγή μηνυμάτων, η Διασύνδεση Δίκτυου του Τηλέγραφου υποστηρίζει κατανεμημένη κοινόχρηστη μνήμη με χαμηλό κόστος. Παρέχει υποστήριξη σε hardware για αυτόματη ενημέρωση των αντιγράφων μια σελίδας, που μπορεί να υπάρχουν στο δίκτυο, όταν γίνει μια εγγραφή στην αρχική σελίδα (eager sharing). Επίσης, εκτός από μακρινές εγγραφές, παρέχει μακρινές ατομικές πράξεις και μακρινές αναγνώσεις που μπλοκάρουν ή όχι τον επεξεργαστή. Ετσι, για κάθε σελίδα, ο μεταφραστής ή το λειτουργικό σύστημα μπορούν να επιλέξουν τοπική αντιγραφή ή μακρινή πρόσβαση. Σ' αυτή την επιλογή βοηθούν μετρητές τοπικών και μακρινών προσβάσεων που περιλαμβάνονται στην Διασύνδεση Δικτύου.

Πολλές από τις λειτουργίες του Τηλέγραφου συναντώνται σε παλιότερα συστήματα, όπως πράξεις σε μακρινή μνήμη (π.χ. [Cray93]), eager sharing (π.χ. [WiMa89], [Mapl92], [Serp90], [WLT93]), ή πολύ γρήγορη επεξεργασία μηνυμάτων (π.χ. [ECGS92], [HeJo92], [NWD93]). Ειδικά σχετικά με την γρήγορη ανταλλαγή μηνυμάτων, πολλές πρόσφατες εργασίες εστιάζουν σε διασυνδέσεις υπολογιστών (και ήδη υπαρχόντων σταθμών εργασίας) με το δίκτυο και σε ανταλλαγή μηνυμάτων σε επίπεδο χρήστη, (π.χ. το [Blum94], και το [MiBH95]) αλλά και σε μεταγωγείς χαμηλής καθυστέρησης σαν δομικά στοιχεία του δικτύου, όπως το [Myri95]. Εκτός από τον Τηλέγραφο, κι άλλες αρχιτεκτονικές υποστηρίζουν γρήγορη ανταλλαγή μηνυμάτων πάνω από ένα σύστημα κατανεμημένης κοινόχρηστης μνήμης, (όπως για παράδειγμα το [Leno92], και το [HeGDG94]). Ομως, άλλες από αυτές βασίζονται σε πρωτόκολα συνάφειας μεταξύ των κρυφών μνημών (cache coherence), και άλλες (π.χ. [Agar91]) χρησιμοποιούν έναν αφιερωμένο επεξεργαστή για την διασύνδεση με το δίκτυο. Και στις δύο περιπτώσεις η πολυπλοκότητα του hardware είναι πολύ μεγαλύτερη, και υψηλότερου κόστους από την αντίστοιχη του Τηλέγραφου για παρόμοια απόδοση. Σε σχέση με αυτά, ο Τηλέγραφος παρέχει αυτά τα χαρακτηριστικά με χαμηλότερο κόστος, ενώνοντας ήδη υπάρχοντα εξοπλισμό μέσω ενός υπερταχέος τοπικού δικτύου για να πετύχει την επίδοση ενός πολυεπεξεργαστή.

*3. Τ*ο κύκλωμα Διασύνδεσης με το Δίκτυο

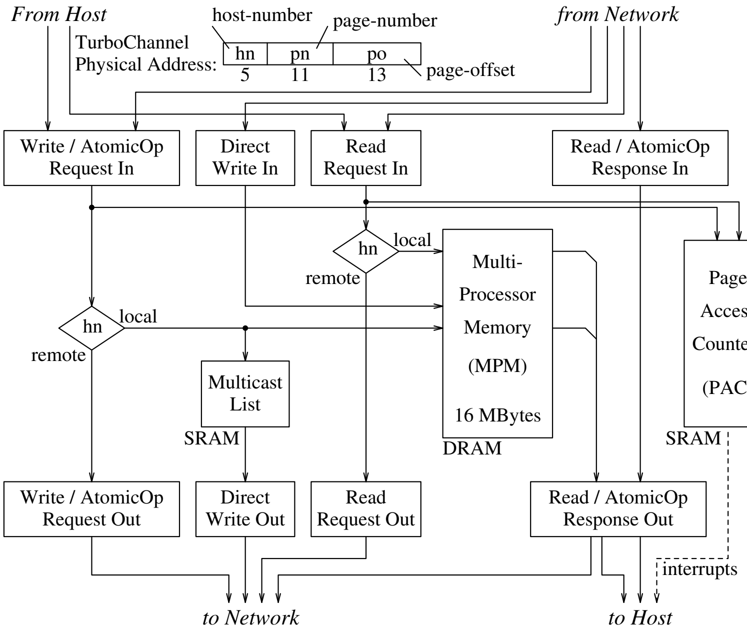
Το κύκλωμα Διασύνδεσης των σταθμών εργασίας με το Δίκτυο του Τηλέγραφου Ι (που θα το ονομάζουμε Host Interface Board – HIB) , όπως παρουσιάζεται στην εικόνα 1 περιέχει την τοπική κοινόχρηστη μνήμη (που ονομάζουμε Multiprocessor Memory - MPM), την μνήμη SRAM με την λίστα multicast , την μνήμη SRAM με τους μετρητές πρόσβασης στις σελίδες της μνήμης, τα κυκλώματα που ενεργοποιούν τις λειτουργίες δικτύου σε απάντηση συγκεκριμένων προσβάσεων από την CPU, τα κυκλώματα που χειρίζονται αιτήσεις πράξεων που έρχονται από το δίκτυο ή από τον τοπικό επεξεργαστή, και τις διασυνδέσεις με το Turbochannel και με το δίκτυο.



*Ε*ικόνα *1:* Διάγραμμα της Διασύνδεσης Δικτύου του *Τ*ηλέγραφου *Ι*

Ο τοπικός επεξεργαστής επικοινωνεί με το σύστημα του Τηλέγραφου μέσω του Turbochannel το οποίο διαθέτει 29 bits για φυσικές διευθύνσεις. Χρησιμοποιώντας αυτά τα 29 bits, κάθε επεξεργαστής μπορεί να προσπελάσει την μνήμη του πολυεπεξεργαστή (MPM) του συνολικού συστήματος του Τηλέγραφου Ι. Μέσα σ' αυτό το χώρο 2^{29} διευθύνσεων περιλαμβάνεται και ο χώρος με τους καταχωρητές και τους πίνακες του Τηλέγραφου Ι που χρησιμοποιείται αποκλειστικά από το Λειτουργικό Σύστημα για έλεγχο και διαχείριση. Κάθε HIB έχει 16Mbytes μνήμης MPM. Τα 24 από τα 29 bits της φυσικής διεύθυνσης του Turbochannel χρησιμοποιούνται για την προσπέλαση της τοπικής μνήμης του κάθε HIB και υπόλοιπα 5 ορίζουν σε ποιο HIB αναφέρεται η συγκεκριμένη διεύθυνση. Στην εικόνα 2 παρουσιάζεται η χρήση των bits της φυσικής διεύθυνσης στον Τηλέγραφο Ι: τα 5 περισσότερο σημαντικά bits, *hn*, δίνουν τον αριθμό του σταθμού εργασίας, τα επόμενα 11, *pn*, περιγράφουν τον αριθμό της σελίδας, και τα 11 λιγότερο σημαντικά, *po*, την θέση μέσα στην σελίδα. Η αντιστοίχιση αυτή έγινε με δεδομένο το μέγεθος σελίδας του επεξεργαστή Alpha που είναι 8 Kbytes.

Ο περιορισμένος αριθμός των bits διεύθυνσης που απομένουν για την ταυτότητα σταθμού (5 bits), μας επιτρέπει να έχουμε μόνο 32 σταθμούς εργασίας στον Τηλέγραφο Ι. Σε ένα εμπορικό σύστημα αυτό δεν είναι αποδεκτό, και μπορεί να αποφευχθεί με τους δύο παρακάτω τρόπους : *(i)* ο Τηλέγραφος συνδέεται με σταθμούς εργασίας που παρέχουν διαύλους με περισσότερες φυσικές διευθύνσεις από αυτές που δίνει το Turbochannel, ή *(ii)* ένας πίνακας μετάφρασης προστίθεται στην Διασύνδεση Δικτύου. Ο Τηλέγραφος ΙΙ που σχεδιάζουμε υλοποιεί την δεύτερη αυτή λύση. Τα 10 περισσότερο σημαντικά ψηφία της φυσικής διεύθυνσης χρησιμοποιούνται σαν δείκτης σε μια στατική μνήμη (που ελέγχεται από το λειτουργικό σύστημα) η οποία τα αντιστοιχεί σε 23 bits. Αυτά τα 23 bits περιέχουν 10 bits με την ταυτότητα του σταθμού προορισμού, και 13 bits που είναι το περισσότερο σημαντικό μέρος μιας 32 bit φυσικής διεύθυνσης της κύριας μνήμης του σταθμού προορισμού.



Εικόνα 2: Διάγραμμα ροής των πράξεων της Διασύνδεσης Δικτύου του *Τ*ηλέγραφου *I*

Οπως φαίνεται και στο σχήμα 2, κάθε MPM πρόσβαση που φτάνει στο HIB από το τοπικό επεξεργαστή ελέγχεται για να εξακριβωθεί αν αναφέρεται σε τοπική ή μακρινή σελίδα. Αιτήσεις που έρχονται από το δίκτυο αναφέρονται πάντοτε σε τοπική σελίδα. Οσον αφορά τις τοπικές αιτήσεις, η ταυτότητα του σταθμού προορισμού διευκρινίζεται με την σύγκριση του πεδίου που περιέχει την ταυτότητα του σταθμού, *hn*, με τον τοπικό αριθμό ταυτότητας ο οποίος ορίζεται από την τιμή διακοπτών DIP που περιέχονται σε κάθε HIB. Οι αιτήσεις τοπικών αναγνώσεων και ατομικών πράξεων παίρνουν πάντοτε άμεση απάντηση. Για να απλουστευθεί το HIB, δεδομένου ότι το Alpha/Turbochannel δεν υποστηρίζει πολλαπλές αναγνώσεις μνήμης σε αναμονή, επιτρέπουμε να υπάρχει μόνο μία ανάγνωση ή αίτηση ατομικής πράξης από τον τοπικό επεξεργαστή σε αναμονή. Για τις τοπικές εγγραφές, ελέγχεται η λίστα multicast της συγκεκριμένης σελίδας, και για κάθε καταχώρηση της στέλνεται στον αντίστοιχο σταθμό προορισμού ένα πακέτο που περιέχει μια αίτηση άμεσης εγγραφής *(direct write).* Η αίτηση αυτή άμεσης εγγραφής είναι όμοια με μια κοινή αίτηση εγγραφής, με την διαφορά ότι δεν γίνεται έλεγχος στην λίστα multicast του σταθμού προορισμού, και κατά συνέπεια δεν καταλήγει σε multicasting.

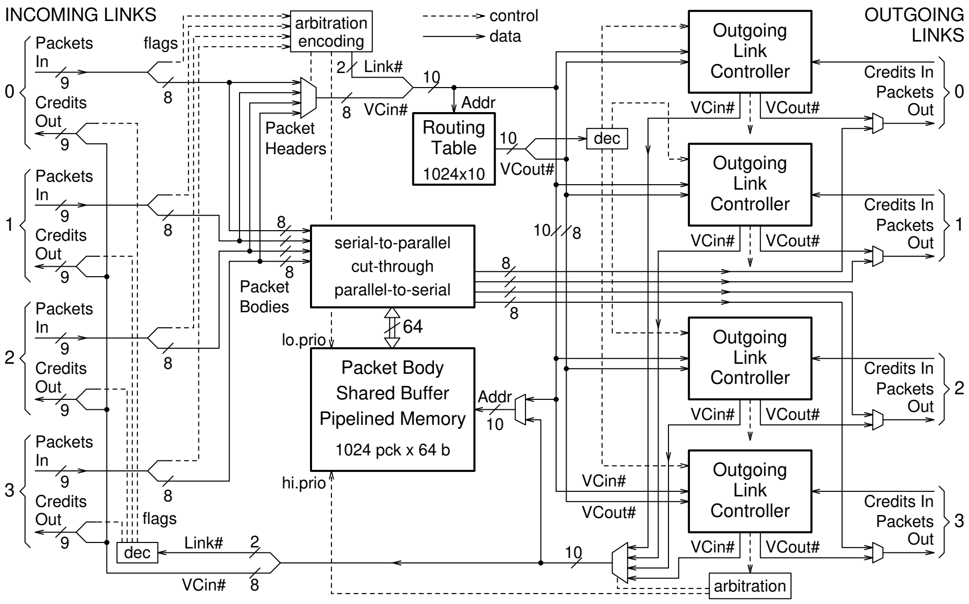
Ο Τηλέγραφος, εκτός από τις πράξεις εγγραφής και ανάγνωσης στην MPM μνήμη, υποστηρίζει επίσης την πράξη της μακρινής αντιγραφής (remote copy), ένα σύνολο από ατομικές πράξεις (fetch-and store, fetch-and-add, compare-and-swap), και εντολή διακοπής σε μακρινό επεξεργαστή (remote interrupt). Η έκδοση των παραπάνω πολύπλοκων πράξεων, γίνεται μέσω μιας ακολουθίας δύο ώς τριών εντολών αποθήκευσης/φόρτωσης, οι οποίες περνούν στο HIB τον τύπο της πράξης (opcode) και τα ορίσματα (διευθύνσεις, δεδομένα), κάνοντας ταυτόχρονα μετάφραση και έλεγχο των δικαιωμάτων προσπέλασης των διευθύνσεων. Κατά την εκτέλεση των παραπάνω πράξεων, οι εντολές φόρτωσης και αποθήκευσης που απευθύνονται στον Τηλέγραφο, εμφανίζονται μέσω του Turbo Channel στο HIB bus, τον κύριο διάδρομο επικοινωνίας του HIB. Εκεί, ανάλογα με τον τύπο της πράξης ενεργοποιείται κάποιο κύκλωμα του HIB που αναλαμβάνει είτε να εκτελέσει την πράξη (πραγματοποιώντας τις απαιτούμενες αριθμητικές και λογικές πράξεις, και προσπελάσεις στην μνήμη) αν είναι τοπική, ή να κατασκευάσει ένα πακέτο και να το προωθήσει στο δίκτυο, αν είναι μακρινή.

*4. Τ*ο διάγραμμα του *Μ*εταγωγέα Πακέτων

Για την σύνδεση κατανεμημένων συστημάτων, οι μεταγωγείς είναι τα ιδεώδη στοιχεία για υψηλές ταχύτητες και με δυνατότητα επέκτασης (σε αντίθεση με τις συνδέσεις τύπου διαύλου), [Kung92]. Το δίκτυο του Τηλέγραφου I δομείται από μεταγωγείς που υλοποιούν προληπτικό έλεγχο ροής από σύνδεσμο σε σύνδεσμο. Αυτός ο μηχανισμός, που βασίζεται σε εισιτήρια, αποκλείει την υπερχείλιση των ενταμιευτών μέσα στους μεταγωγείς, και την απώλεια πακέτων δεδομένων. Παρακάτω, θα μελετήσουμε σε λεπτομέρεια τα κυκλώματα του μεταγωγέα.

Οπως έχουμε ήδη δεί, ο Τηλέγραφος I, που είναι το αρχικό μας πρωτότυπο, σχεδιάστηκε σαν μοντέλο γρήγορης υλοποίησης (rapid prototyping), και αποτελείται κυρίως από FPGAs και στατικές μνήμες. Οι διάφορες επιλογές παραμέτρων του Τηλέγραφου I οφείλονται σε περιορισμούς από την τεχνολογία που χρησιμοποιήθηκε και δεν είναι εγγενή χαρακτηριστικά της αρχιτεκτονικής του. Στον Τηλέγραφο-I οι σύνδεσμοι μεταφέρουν παράλληλα 8 bits δεδομένων (+ 1 flag bit), το μέγεθος πακέτου είναι 72 bits, άρα χρειάζονται 9 κύκλοι ρολογιού για να περάσει ένα πακέτο μέσα από ένα σύνδεσμο. Η επικεφαλίδα του πακέτου (8 bits) είναι ο αριθμός του virtual circuit VC, και τα υπόλοιπα 64 bits αποτελούν το περιεχόμενο του πακέτου. Με 8 bits επικεφαλίδα υποστηρίζονται 256 VC's σε κάθε σύνδεσμο. Για λόγους απλοποίησης και επιτάχυνσης της πρωτότυπης σχεδίασης, τα εισιτήρια ταξιδεύουν μέσα από διαφορετικούς συνδέσμους. Σε ένα μεταγωγέα υλοποιημένο σε ένα ολοκληρωμένο ASIC, θα γινόταν οικονομία πολυπλέκοντας πάνω στους ίδιους συνδέσμους τα πακέτα και τα εισιτήρια – έτσι ακριβώς γίνεται στον Τηλέγραφο II. Ο μεταγωγέας του Τηλέγραφου-I έχει 4 συνδέσμους πακέτων εισόδου και 4 εξόδου (και ίδιο αριθμό συνδέσμων εισιτηρίων).

Στην εικόνα 3 παρουσιάζεται το διάγραμμα του μεταγωγέα του Τηλέγραφου I, παραλείποντας τις λεπτομέρειες χρονισμού και καταχωρητών pipeline. Παρακάτω, θα περιγράψουμε την ροή των πακέτων μέσα από τον μεταγωγέα. Οι επικεφαλίδες των εισερχομένων πακέτων στέλνονται στον Πίνακα Δρομολόγησης (Routing Table) (γνωστός και σαν πίνακας μετάφρασης σε ορολογία δικτύων ATM), όπου εξυπηρετούνται σε ένα κύκλο ρολογιού η καθεμία. Ενα κύκλωμα διαιτησίας σειριοποιεί πιθανές ταυτόχρονες αφίξεις επικεφαλίδων. Μέσα σε 9 διαδοχικούς κύκλους ρολογιού, το πολύ 4 επικεφαλίδες πακέτων μπορούν να εμφανιστούν στις εισόδους του μεταγωγέα.



Εικόνα 3: Διάγραμμα του *Μ*εταγωγέα του *Τ*ηλέγραφου *Ι*

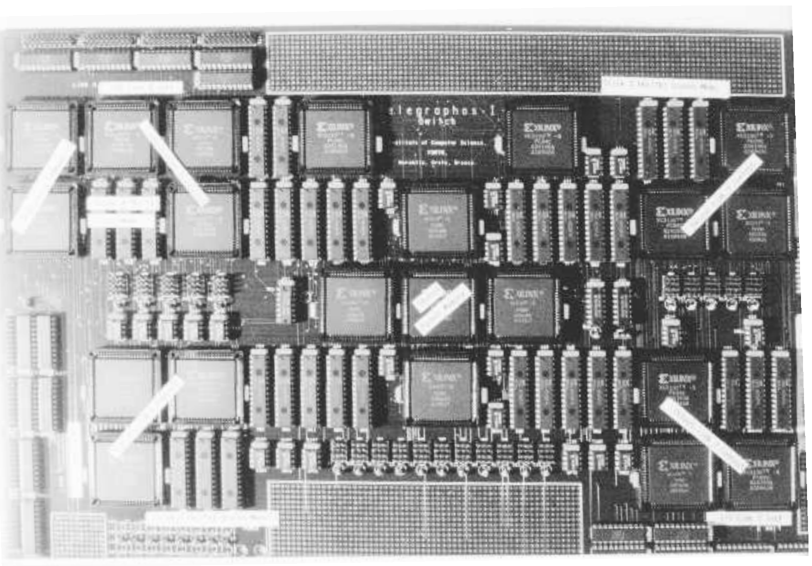
Τα σώματα των εισερχομένων πακέτων αποθηκεύονται σε μία κεντρική μνήμη που την μοιράζονται όλοι οι σύνδεσμοι. Αυτή η μνήμη είναι οργανωμένη σαν *shared buffer* γιατί αυτή η οργάνωση δίνει την μικρότερη καθυστέρηση μέσα στον μεταγωγέα [HlKa88], και βέλτιστη χρήση της διαθέσιμης ποσότητας μνήμης. Ενώ η αρχιτεκτονική shared buffer είναι η καλύτερη από άποψη απόδοσης, πολλοί σχεδιαστές την θεωρούν πολύ ακριβή γιατί απαιτεί από την μνήμη μεγάλο ρυθμό μεταφοράς δεδομένων. Στον Τηλέγραφο ακολουθούμε μια νέα οργάνωση μιας τέτοιας μνήμης και των κυκλώματων εισόδου, εξόδου και υλικού μεταγωγής της, που ονομάζουμε Pipelined Memory, και που παρουσιάζεται λεπτομερώς στο [KaVE94]. Ανακεφαλαιώνοντας, αυτή η μνήμη δέχεται ή στέλνει ένα ολόκληρο πακέτο (64 bits) σε κάθε κύκλο ρολογιού. Αρα, έχει αρκετό ρυθμό μεταφοράς δεδομένων ώστε στη διάρκεια οποιονδήποτε 9 διαδοχικών κύκλων, να γράφει μέχρι 4 εισερχόμενα πακέτα και να διαβάζει μέχρι 4 εξερχόμενα πακέτα. Η είσοδος/έξοδος των δεδομένων σε αυτή τη μνήμη γίνεται με το σειριακό/παράλληλο/cut-through κύκλωμα που φαίνεται στο μέσο της εικόνας 3, και του οποίου η οργάνωση περιγράφεται στο [KaSC91, εικόνα 11]. Το μέγεθος της μνήμης για τα σώματα των πακέτων είναι 1024 πακέτα, και η διεύθυνση όπου αποθηκεύεται το κάθε πακέτο σχηματίζεται συρράπτοντας την ταυτότητα (ID) του εισερχομένου συνδέσμου και τον αριθμό του VC.

Τη στιγμή άφιξης του πακέτου, πέρα από την αποθήκευση του εισερχομένου πακέτου στην κεντρική μνήμη πακέτων, και τη μετάφραση του εισερχομένου αριθμού VC (VCin) στον εξερχόμενο αριθμό VC (VCout), ειδοποιείται επίσης ο αρμόδιος ελεγκτής συνδέσμου εξόδου (Outgoing Link Controller) για την άφιξη αυτού του πακέτου. Ενας τέτοιος ελεγκτής υπάρχει για κάθε σύνδεσμο εξόδου. Η λειτουργία του είναι να κρατάει λογαριασμό των πακέτων που περιμένουν να φύγουν από τον συγκεκριμένο σύνδεσμο, καθώς και των εισιτηρίων για θέσεις ενταμίευσης στον επόμενο μεταγωγέα. Επίσης, είναι υπεύθυνος για να προγραμματίσει την αναχώρηση πακέτων από τον αντίστοιχο σύνδεσμο και να αποστείλει το επόμενο έτοιμο πακέτο όταν ο σύνδεσμος είναι ανενεργός. Οταν ο ελεγκτής αποφασίσει ότι είναι ώρα για ένα συγκεκριμένο πακέτο να μεταδοθεί από τον σύνδεσμό του, ζητάει προσπέλαση στην μνήμη σώματος πακέτου από τα κυκλώματα διαιτησίας, και μόλις την εξασφαλίσει στέλνει *(i)* τον αριθμό VCout# στον εξερχόμενο σύνδεσμο σαν νέα επικεφαλίδα *(ii)* τον αριθμό VCin# στην μνήμη σώματος πακέτου σαν την διεύθυνση για το διάβασμα του σώματος του πακέτου και *(iii)* τον αριθμό VCin# στον κατάλληλο σύνδεσμο εισιτηρίου προς τον προηγούμενο γειτονικό μεταγωγέα, ώστε να τον ειδοποιήσει ότι υπάρχει χώρος αποθήκευσης για ένα ακόμη πακέτο με αυτό τον αριθμό VC.

Σε ένα ελαφρά φορτωμένο δίκτυο, η καθυστέρηση cut-through μέσα από τον μεταγωγέα (η καθυστέρηση της επικεφαλίδας δια μέσου του chip) διαρκεί 3 κύκλους ρολογιού. Για ένα ζεύγος από γειτονικούς μεταγωγείς με χρόνο καθυστέρησης πάνω στους συνδέσμους 1 κύκλο, η καθυστέρηση round-trip πακέτου-εισιτηρίου είναι 9 κύκλοι ρολογιού. Αρα, είναι αρκετό να έχουμε χώρο αποθήκευσης για ένα πακέτο ανά VC για να πετύχουμε πλήρη χρήση του συνδέσμου ακόμη και από ένα μοναδικό ενεργό VC.

5. Η υλοποίηση των *Μ*εταγωγέων και των Διασυνδέσεων με το Δίκτυο

Ο Τηλέγραφος Ι έχει κατασκευαστεί και αυτή την στιγμή βρίσκεται στα τελικά στάδια του ελέγχου, με τα περισσότερα μέρη του να έχουν ήδη επιβεβαιωθεί ότι δουλεύουν σωστά. Οσον αφορά στον μεταγωγέα, έχουμε ελέγξει και επιβεβαιώσει όλες τις βασικές λειτουργίες του: τα κυκλώματα γύρω από τις πόρτες εισόδου και εξόδου, την ορθότητα ελέγχου και προσπελάσεων στον κεντρικό ενταμιευτή πακέτων, τις εντολές διαχείρισης και επίβλεψης. Εχουμε επιτυχώς περάσει πακέτα μέσα από τον μεταγωγέα προς μία πόρτα εξόδου και έχουμε επιβεβαιώσει ότι ενεργοποιείται η λειτουργία cut-through όταν αυτό είναι δυνατό.



Εικόνα 4: Φωτογραφία του τυπωμένου κυκλώματος του μεταγωγέα του *Τ*ηλέγραφου *Ι*

Στην εικόνα 4 βλέπουμε μια φωτογραφία του 4×4 μεταγωγέα του Τηλέγραφου Ι. Ο Τηλέγραφος Ι με μέγεθος πακέτου 72 bits, χρόνο πακέτου 9 κύκλους ρολογιού, και κύκλο 75 ns, έχει εύρος ζώνης ανά σύνδεσμο 103 Mbits/s σε κάθε κατεύθυνση. Ο υπερβολικά αργός κύκλος οφείλεται στην τεχνολογία FPGA που χρησιμοποιήσαμε για να διατηρηθεί χαμηλό το κόστος ανάπτυξης: τα FPGAs της σειράς Xilinx 3000 (κι όχι της σειράς 4000 που είναι σημαντικά πιο γρήγορα και με μεγαλύτερη χωρητικότητα). Επίσης, είχαμε πολύ μεγάλες και αναπόφευκτες καθυστερήσεις στις περιπτώσεις που η λογική χρειαζόταν να μοιραστεί μεταξύ πολλών FPGA's.

Ο μεταγωγέας του Τηλέγραφου Ι είναι ένα τυπωμένο κύκλωμα 6-επιπέδων (με 4 επίπεδα για σύνδεση σήματων και 2 επίπεδα για τροφοδοσία), μεγέθους $42 \times 30\ cm^2$, με 4000 pins, 1000 δίκτυα σημάτων, και γραμμές πάχους $0.2\ mm$. Περιέχει 110 Kbits μονόπορτης μνήμης SRAM και το αντίστοιχο περίπου 14,500 πυλών τυχαίας λογικής που περιέχονται σε 21 συνολικά FPGA's. Στον Πίνακα 1 παρουσιάζεται με λεπτομέρεια η υλοποίηση κάθε υποκυκλώματος του μεταγωγέα όπως υλοποιήθηκε σε μια Xilinx 3100 FPGA. Για κάθε υποκύκλωμα, ο πίνακας δίνει των αριθμό των ''πρωτογενών κυττάρων'' της FPGA (CLB) που καταλαμβάνει, τον αριθμό των flip-flops τύπου-D (DFF) που χρησιμοποιεί, το ισοδύναμο του προσεγγιστικού αριθμού πυλών τυχαίας λογικής μέσα στην FPGA, καθώς και την ποσότητα εξωτερικής μνήμης που χρησιμοποιεί. Οπως βλέπουμε, ο κεντρικός ενταμιευτής πακέτων καταλαμβάνει 5,900 πύλες και 64 Kbits μνήμης. Ο έλεγχος του, που περιλαμβάνει την διαιτησία πρόσβασης μεταξύ εισερχομένων και εξερχομένων συνδέσμων, και την δημιουργία των σημάτων ελέγχου για την πρώτη βαθμίδα της pipeline, χωράει σε μια Xilinx 3130 FPGA. Το datapath εισόδου και εξόδου του κεντρικού ενταμιευτή υλοποιήθηκε σαν τέσσερεις φέτες των 2-bit. Οι ελεγκτές για τους τέσσερεις εξερχόμενους συνδέσμους καταναλώνουν 6,500 πύλες και 28 Kbits μνήμης, και χωράνε σε τρεις FPGA's για κάθε εξερχόμενο σύνδεσμο.

Πίνακας *1:* Στοιχεία Περιπλοκότητας του *Μ*εταγωγέα του *Τ*ηλέγραφου *Ι*

Υποκύκλωμα	*CLBs*	*DFFs*	Λογική (ισοδ. πύλες)	Ποσ.	Συνολική λογική (ισοδ. πύλες)	*SRAM* (Kbits)
Datapath Ενταμιευτή Δεδομένων 2-bit slice	138	199	1347	4	5390	8*2=16
Ελεγχος Ενταμιευτή Δεδομένων	69	59	510	1	510	
Συνολικά *Ε*νταμιευτής Δεδομένων					*5900*	*64*
Κατάσταση Πακέτων, Εισητηρίων	53	28	325	4	1300	4*3.5=14
Ουρά Ετοιμων Πακέτων	95	70	660	8	5280	4*4.5=14
Συνολικά Σύνδεσμοι *Ε*ξόδου					*6580*	*28*
Εντολές Διαχείρισης (για setup)	59	58	470	1	468	
Εντολές Διαχείρισης (για επίβλεψη)	76	49	500	1	500	
Συνολικά *Ε*ντολές Διαχείρισης					*970*	
Διασυνδέσεις *Ε*ισερχόμενων Συνδέσμων	70	65	540	*2*	*1080*	
Συνολικά Μεταγωγέας					**14530**	**108**

Παράλληλα με τον έλεγχο του Τηλέγραφου Ι, δουλεύουμε και στον μεταγωγέα του Τηλέγραφου ΙΙ – μια ολοκληρωμένη σε ένα chip εκδοχή του μεταγωγέα του Τηλέγραφου Ι. Το ASIC του μεταγωγέα του Τηλέγραφου ΙΙ σχεδιάσθηκε βάσει της 0.7 μm CMOS τεχνολογίας standard-cell της European Silicon Structures Inc. (ES2), έχει σταλεί για κατασκευή, και περιμένουμε τα πρωτότυπα chips. Ο μεταγωγέας του Τηλέγραφου ΙΙ είναι 4×4, με μέγεθος πακέτου 144 bits, και σύμφωνα με τις προσομοιώσεις μας δουλεύει στα 25 MHz, και στα 400 Mbps ανά σύνδεσμο σε κάθε κατεύθυνση (κάθε σύνδεσμος έχει πλάτος 16-bit μέσα στο chip, ενώ οι σύνδεσμοι εκτός του chip είναι διπλοχρονιζόμενοι και μεταφέρουν 8 bits κάθε 20 ns).

Πίνακας *2:* Στοιχεία της Υλοποίησης του *HIB* του *Τ*ηλέγραφου *I*

Υποκύκλωμα	Λογική (πύλες)	*SRAM* (Kbits)	Σημειώσεις*:*
Κεντρικός έλεγχος	1000	0.5	
Διασύνδεση με το Turbochannel	550		300 πύλες + 64 bits καταχωρητών
Διασύνδεση με Εισερχόμενο σύνδεσμο	1000	2.	2+2 Kb σύγχρονης (2-πορτης) FIFO's
Διασύνδεση με Εξερχόμενο σύνδεσμο	750	2.	
Μερικό Σύνολο – σχετικό με ανταλλαγή μηνυμάτων	3300	4.5	
Ατομικές πράξεις	1500		
Multicast (eager sharing)	400	512	16 K θέσεις multicast λίστας x 32 bits
Μετρητές Πρόσβασης σε Σελίδες	800	2048	64 K σελίδες x (16+16) bits
Μνήμη Πολυεπεξεργαστή (MPM)			16 MBytes = 128 Mbits DRAM
Μερικό Σύνολο – σχετικό με κοινόχρηστη μνήμη	2700	2500	

Η Διασύνδεση Δικτύου του Τηλέγραφου Ι υλοποιείται σε ένα τυπωμένο κύκλωμα μεγέθους μεγέθους $42 \times 30\ cm^2$, με 3000 pins, και 750 δίκτυα σημάτων. Περιέχει 90 chips, από τα οποία 15 είναι FPGAs, 15 μνήμες SRAM, 32 μνήμες DRAM, 10 FIFOs και 20 TTL chip καταχωρητών που δουλεύουν στα 12.5 MHz που δίνει το TurboChannel. Σχετικά με το τυπωμένο κύκλωμα της Διασύνδεσης Δικτύου, έχει επιβεβαιωθεί η σωστή λειτουργία όλων των κυκλωμάτων που αφορούν στην αποστολή και λήψη πακέτων, και στην ορθή εκτέλεση ατομικών πράξεων. Τα μέρη αυτά είναι: η διασύνδεση με το TurboChannel, τα κυκλώματα που διαχειρίζονται την τοπική δυναμική μνήμη MPM, τα κυκλώματα γύρω από τον σύνδεσμο εξόδου, και τα κυκλώματα ατομικών πράξεων. Στον πίνακα 2 απαριθμούμε την αριθμό πυλών αντίστοιχης πολυπλοκότητας και τα μεγέθη μνήμης για τα διάφορα μέρη του HIB του Τηλέγραφου Ι. Οπως φαίνεται, το μέρος της Διασύνδεσης Δικτύου που είναι απαραίτητο για να υποστηρίξουμε ένα σύστημα ανταλλαγής μηνυμάτων είναι πολύ μικρό: 3300 πύλες και λίγα Kbits μνήμης. Το μεγαλύτερο μέρος καταλαμβάνεται από τα κυκλώματα γύρω από τους συνδέσμους που όμως, ούτως ή άλλως υπάρχουν σε οποιαδήποτε μορφή δικτύου. Το υπόλοιπο κομμάτι που είναι υπεύθυνο και για τις μακρινές εγγραφές περιέχει περίπου 1500 πύλες και 512 bits SRAM, δηλαδή είναι πολύ μικρό και απλό. Το κυκλώματα της Διασύνδεσης Δικτύου για την υποστήριξη κοινόχρηστης μνήμης είναι επίσης απλά (2700 πύλες), αλλά χρησιμοποιεί σημαντική ποσότητα μνήμης (2.5 megabits SRAM) η οποία θα ήταν εξωτερική, στην περίπτωση υλοποίησης της Διασύνδεσης Δικτύου σε ένα ολοκληρωμένο.

Μια επόμενη εκδοχή της Διασύνδεσης με το Δίκτυο, ολοκληρωμένη μέσα σε ένα chip ASIC σε τεχνολογία (ES2 0.7μm std cell CMOS) σχεδιάζεται για να χρησιμοποιηθεί στον Τηλέγραφο II. Σ' αυτή την υλοποίηση, η μνήμη του πολυεπεξεργαστή δεν θα είναι ξεχωριστή από την κεντρική μνήμη του μηχανήματος, και ο δίαυλος όπου συνδέεται είναι ο δίαυλος PCI.

6. Συμπεράσματα

Συστήματα που χρησιμοποιούν ήδη υπάρχοντες σταθμούς εργασίας μέσω της ομαδοποίησης τους για παράλληλες εφαρμογές κερδίζουν μεγάλο ενδιαφέρον τον τελευταίο καιρό. Ενα τέτοιο σύστημα είναι ο Τηλέγραφος, ο οποίος διασυνδέει υπολογιστές, μέσω ενός επεκτάσιμου δικτύου που δομείται απο υπερταχείς μεταγωγείς και που παρέχει τις βασικές λειτουργίες για γρήγορη επικοινωνία. Η υλοποίηση του Τηλέγραφου με απλή τεχνολογία χαμηλής ολοκλήρωσης επιβεβαιώνει την αρχιτεκτονική, και πετυχαίνει απόδοση μεγάλης κλίμακας κατανεμημένου πολυεπεξεργαστή με πολύ χαμηλό κόστος.

Βιβλιογραφία

[Agar91] A. Agarwal, D. Chaiken, G. D'Souza, K. Johnson, D. Kranz, J. Kubiatowicz, K. Kurihara, et al.: ''The MIT Alewife: A large-Scale Distributed Memory Multiprocessor'', Proceeding of Workshop on Scalable Shared Memory Multiprocessors, Kluwer Academic Publishers, 1991.

[Blum94] M. Blum, K. Li, R. Alpert. C. Dubnicki, E. Felten: ''Virtual Memory Mapped Network Interface for the SHRIMP Multicomputer'' Proceedings of the 21th International Symposium on Computer Architecture, April 18-21, 1994, pp. 142-153.

[Cray93] Cray Research Inc.: ''Cray T3D System Architecture Overview'' System Architecture Overview, 1993, Manual HR-04033.

[ECGS92] T. Von Eicken, D. Culler, S. GoldStein, K. Schauser: ''Active Messages: a Mechanism dor integrated Communication and Computation'', Proceedings of the 19th International Symposium on Computer Architecture, pp. 256-266, May 1992.

[HeGDG94] J. Heinlein, K. Gharachorloo, S. Dresser, A. Gupta: ''Integration of Message Passing and Shared Memory in the Stanford FLASH Multiprocessor'', Proceedings of the 6th Intnl. Conference on Arch. Support for Progr. Languages and Oper. Sys (ASPLOS), October 1994

[HeJo92] D. Henry, C. Joerg: ''A Tightly-Coupled Processor-Network Interface'', ASPLOS V Conference Proceedings, ACM SIGARCH/SIGOPS/SIGPLAN, October 1992, Ma. USA, pp. 111-122.

[HlKa88] M. Hluchyj, M. Karol: ''Queueing in High-Performance Packet Switching'', IEEE Journal on Sel. Areas in Communications, Vol. 6, No. 9, Dec. 1988, pp. 1587-1597.

[KaSC91] M. Katevenis, S. Sidiropoulos, C. Courcoubetis: ''Weighted Round-Robin Cell Multiplexing in a General-Purpose ATM Switch Chip'', IEEE Journal on Selected Areas in Communications, Vol. 9, No. 8, October 1991, pp. 1265-1279.

[Kate94] M. Katevenis: ''Telegraphos: High-Speed Communication Architecture for Parallel and Distributed Computer Systems'', Technical Report FORTH-ICS/TR-123, Institute of Computer Science, FORTH, Heraklio, Crete, Greece, May 1994; available on-line from ftp.ics.forth.gr, directory tech-reports/1994, file 94.TR123.Telegraphos.ps.Z

[KaVE94] M. Katevenis, P. Vatsolaki, A. Efthymiou: ‘‘Pipelined Memory Organization for High Performance Switching and Buffering’’, Technical Report FORTH-ICS/TR-127, Institute of Computer Science, FORTH, Heraklio, Crete, Greece, December 1994; available on-line from ftp.ics.forth.gr, directory tech-reports/1994, file 94.TR127.PipelinedMemory.ps.Z

[Kung92] H. T. Kung: ‘‘Gigabit Local Area Networks: A Systems Perspective’’, IEEE Communications Magazine, Vol. 30, No. 4, April 1992, pp. 79-89.

[Leno92] D. Lenoski, J. Laudon, K. Gharachorloo, W-D. Weber, A. Gupta, J. Hennessy, M. Horowitz, M. Lam: ‘‘The Stanford Dash Multiprocessor’’, IEEE Computer Magazine, March 1992, pp. 63-79.

[Mapl92] C. Maples: ‘‘A High-Performance Memory-Based Interconnection System for Multicomputer Environments’’, Proceedings of the Supercomputing Conference, 1992, pp. 295-304.

[MiBH95] R. Minnich, D. Burns. F. Hady: ‘‘The Memory-Integrated Network Interface’’, IEEE Micro, February 1995, pp. 11-20

[Myri95] N. Boden, D. Cohen, R. Felderman, A. Kulawik, Charles Seitz, J. Seizovic, W.K. Su: ‘‘Myrinet: A Gigabit-per-Second Local Area Network’’, IEEE Micro, February 1995, pp. 29-36

[NWD93] M. Noakes, D. Wallace, William J. Dally, ‘‘The J-Machine Multicomputer: An Architectural Evaluation’’ Proceedings of the 20th International Symposium on Computer Architecture, pp. 224-235, May 16-19, 1993.

[Serp90] D. Serpanos: ‘‘Scalable Shared-Memory Interconnections’’, Ph.D. Thesis, Technical Report CS-TR-227-90, Princeton University, Dept. of Computer Science, Oct. 1990.

[WiMa89] L. Wittie, C. Maples: ‘‘Merlin: Massively Parallel Heterogeneous Computing’’, Proc. 1989 Int. Conf. on Parallel Processing, St. Charles, Illinois, Aug. 1989, pp. 142-150.

[WLT93] A. Wilson Jr., R. LaRowe Jr., M. Teller: ‘‘Hardware Assist for Distributed Shared Memory’’, Proceedings of the 13th International Conference on Distributed Computing Systems, pages 246-255, May 1993.