ΠΕΡΙΛΗΨΗ

Σύμφωνα με τον Νόμο του Moore (1965), η πυκνότητα των τρανζίστορ σε ένα ολοκληρωμένο κύκλωμα διπλασιάζεται κάθε δεκαοκτώ μήνες. Με ολοένα μεγαλύτερο αριθμό τρανζίστορ σε ένα ολοκληρωμένο κύκλωμα, αυξάνεται τόσο ο αριθμός των ενσωματωμένων ψηφιακών λειτουργικών μονάδων όσο και η ποσότητα της μνήμης, και επακόλουθος η αποδοτικότητα των αρχιτεκτονικών Η/Υ. Έτσι έχουμε φτάσει στο σημείο σήμερα όπου είναι δυνατή η ενσωμάτωση δεκάδων ή ακόμα εκατοντάδων συνιστώμενων ψηφιακών ηλεκτρονικών μερών όπως επεξεργαστικές μονάδες (ΕΜ), μνήμες, και δρομολογητές σε ένα ενιαίο ολοκληρωμένο κύκλωμα που περιέχει μια πλειάδα πλακιδίων, όπου το καθένα περιέχει επεξεργαστή, μνήμη, κ.ο.κ. Η συστοιχία αυτών τον πολλαπλών πλακιδίων σε ένα και μόνο ολοκληρωμένο κύκλωμα έχει επιφέρει την καινοτόμα ιδέα του πολύ-πύρηνου επεξεργαστή.

Ένα επακόλουθο αυτού είναι το γεγονός ότι λόγω των αυξανόμενων αναγκών επικοινωνίας ανάμεσα στα πλακίδια ενός πολύ-πύρινου επεξεργαστή, ότι τα ολοκληρωμένα ενδό-συνδεόμενα δίκτυα (Networks-on-Chip, NoC), τα οποία προσφέρουν γρήγορη και αποδοτική επικοινωνία, άρχισαν να αντικαταστούν άλλους παλαιότερους και μη αποδοτικούς τρόπους επικοινωνίας όπως οι αφοσιωμένες καλωδιώσεις και τα διαζεύγματα. Πέραν από την ψηλή τους απόδοση, τα ολοκληρωμένα ενδό-συνδεόμενα δίκτυα προσφέρουν και άλλα πλεονεκτήματα όπως τη διαχείριση κατανάλωσης ενέργειας, και την μεγαλύτερη ανοχή σε σφάλματα που προκύπτουν κατά τη διάρκεια της λειτουργίας ενός πολύ-πύρηνου επεξεργαστή.

Για να αποσταλούν και να παραδοθούν μηνύματα, που είναι σε μορφή ψηφιακών πακέτων, από ένα πλακίδιο σε άλλο σε έναν πολύ-πύρηνο επεξεργαστή, γίνεται χρήση αλγορίθμου δρομολόγησης. Ο αλγόριθμος δρομολόγησης φέρει άμεση επίπτωση στην αποδοτικότητα ενός ολοκληρωμένου δικτύου, εφόσον επηρεάζει άμεσα την καθυστέρηση στην αποστολή των μηνυμάτων και στην επίτευξη εκτενούς ζωνικού εύρους και διεκπεραιωτικότητας. Καθόλα αυτά, τα περισσότερα ολοκληρωμένα ενδό-συνδεόμενα δίκτυα τείνουν να χρησιμοποιούν απλούς αλγόριθμους για την δρομολόγηση των δεδομένων οι οποίοι ναι μεν εγγυώνται ότι όλα τα πακέτα δεδομένων θα φτάσουν στον προορισμό τους αλλά δεν λαμβάνουν υπόψη τους τη συμφόρηση που υπάρχει στο δίκτυο, και γενικά την κατάσταση στην

οποία βρίσκεται το δίκτυο, με αποτέλεσμα να περιορίζεται σημαντικά η αποδοτικότητα του δικτύου, και μετέπειτα η απόδοση ολόκληρου του πολύ-επεξεργαστή. Στη χειρότερη περίπτωση, αλγόριθμοι δρομολόγησης οι οποίοι είναι «αδιάφοροι» ως προς τη συμφόρηση δικτύου, δεν αποσκοπούν στο να εξισορροπήσουν το φορτίο που μεταφέρεται χρησιμοποιώντας εναλλακτικές διαδρομές, και έτσι παρατείνεται η συμφόρηση στο δίκτυο. Άλλοι μεν, λαμβάνουν υπόψη τους την τοπική κατάσταση στο δίκτυο, και προσπαθούν να προσαρμόσουν τις αποφάσεις τους για τις διαδρομές που θα ακολουθήσουν τα πακέτα, αλλά η έλλειψη μη βέλτιστης όψης της κατάστασης του δικτύου, περιορίζει και σε αυτούς το επιταγμένο ύψος απόδοσης τους.

Λαμβάνοντας υπόψη τους πιο πάνω περιορισμούς, σε αυτή τη διπλωματική εργασία προτείνουμε έναν προσαρμοστικό αλγόριθμο δρομολόγησης με σφαιρική εξισορρόπηση ροής φορτίου, όπου λαμβάνει υπόψη του την κατάσταση συμφόρησης της ροής δεδομένων καθόλη την επιφάνεια της τοπολογίας δικτύου, με γνώμονα τη βέλτιστη καθοδήγηση της ροής δεδομένων έτσι που να περιορίζεται η συμφόρηση και να αυξάνεται η διεκπεραιωτικότητα του δικτύου, και έτσι η συνολική απόδοσή του. Προτείνουμε τρεις διαφορετικές εκδοχές του σφαιρικού αλγορίθμου, όπου η κάθε εκδοχή παρουσιάζει διαφορική επίτευξη αποδοτικότητας του ολοκληρωμένου δικτύου. Τα αποτελέσματα των πειραμάτων μας με χρήση εξομοιωτή δικτύου, έχουν καταδείξει μέχρι 17.5% αύξηση στην απόδοση σε σχέση με μη προσαρμοστικό αλγόριθμο δρομολόγησης.

Λέξεις κλειδιά: Ολοκληρωμένα ενδό-συνδεόμενα δίκτυα, προσαρμοστικοί αλγόριθμοι δρομολόγησης, εξισορρόπηση φορτίου.

ABSTRACT

According to Moore's law (1965), the density of transistors in an integrated circuit doubles every eighteen months. With integrated circuits containing an increasing amount of transistors, the number of on-chip digital modules and memory units that can fit onto a chip also increases, and as a consequence the performance of computer architectures has been growing at an exponential rate. This trend has led us to the point today where it is possible to pack tens or even hundreds of digital electronic modules onto a single silicon die, comprising multiple tiles, where every tile contains a processing element, a graphic engine, multi-level cache memory, a router, etc. The array of these multiple tiles in a single integrated circuit has given rise to the multi-core processor paradigm.

With multi-core processors there is increasing demand for communication between its tiles. In such a communication-centric system, dedicated wires and limited-connectivity crossbars are no longer adequate in handling the exponentially increasing communication demands, which come in the form of packetized messages exchanges. Alternatively, Networks-on-Chips (NoCs), miniature-scale counterparts of large-scale off-chip networks found in computer clusters, servers and supercomputers, instead provide fast and efficient communication among the various tiles. Besides their high-throughput capabilities, NoCs offer further desirable advantages, such as those of power consumption management, greater tolerance to faults that arise during the operational life of a multi-core processor, modularity, and scalability.

Packetized message delivery among the various tiles of a multi-core processing chip requires the use of a reliable and efficient routing algorithm. The routing algorithm directly impacts the attainable performance of a NoC, since it affects the network's latency observed when sending messages, and also the network's bandwidth and effective throughput. Most NoCs tend to use simplified routing algorithms to route packets among tiles, and although they do ensure that the routed data will eventually arrive at their destinations, however, they do not take into consideration the network's congestion levels and statuses that are currently present. This behavior, unfortunately, limits the network's effective throughput, and

subsequently constraints the entire performance of the multi-core processor system. In the worst-case scenario, routing algorithms that are oblivious to the current contention state of a network do not aim toward balancing the traffic workload using alternative topology paths, and thus, the congestion in the network is left unmonitored and unconstrained. Next, semi-adaptive routing algorithms, though, do take into consideration the localized, i.e., in their immediate vicinity, network state and try to adapt their routing decisions to the network's state, but they lack a global and hence optimal view of the network which consequently limits their performance too.

Taking the above shortcomings into consideration, in this thesis, we propose an adaptive routing algorithm with global load-balancing capabilities that takes into consideration the congestion level and state of the traffic's data flow across the entire network topology. Our goal is to provide the best guidance in routing the data flow across alternative topology paths so that network congestion is restricted, and the network's throughput and overall performance are maximized. We propose three different versions of this global congestion-aware load-balancing routing algorithm, each with a different achievable performance level. Our experimental evaluation and results, using a cycle-accurate network simulator, show up to 17.5% improvement in the network's effective throughput when compared to a non-congestion-adaptive routing algorithm.

KEYWORDS: Networks-on-Chip, Adaptive routing algorithms, Load balancing