Computing Architectures exploiting Optical Interconnect and Optical Memory Technologies

DOCTORAL THESIS

Pavlos Maniotis

Submitted to the School of Informatics of Aristotle University of Thessaloniki

Thessaloniki, November 2017

Advisory Committee: N. Pleros, Assistant Professor, A.U.TH. (Supervisor)
A. Miliou, Associate Professor, A.U.TH.
K. Tsichlas, Lecturer, A.U.TH.
Examination Committee:  
A. Gounaris, Assistant Professor, A.U.TH.  
A. Miliou, Associate Professor, A.U.TH.  
G. Papadimitriou, Professor, A.U.TH.  
A. Papadopoulos, Associate Professor, A.U.TH.  
N. Pleros, Assistant Professor, A.U.TH.  
K. Tsichlas, Lecturer, A.U.TH.  
K. VyrsoKinos, Lecturer, A.U.TH.
Η έγκρισή της παρούσας Διδακτορικής διατριβής από το τμήμα Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέως. (Ν. 5343/1932, σχ. 202, παρ. 2)
Ευχαριστίες

Η παρούσα διδακτορική διατριβή αποτελεί τον καρπό των ερευνητικών εργασιών μου που πραγματοποιήθηκαν κατά τη διάρκεια της συμμετοχής μου στην ερευνητική ομάδα των Φωτονικών Συστημάτων και Δικτύων (Photonic Systems and Networks – PhoS-net) του τμήματος Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης. Στο σημείο αυτό, έχοντας πλέον ολοκληρώσει τη συγγραφή του τεχνικού κειμένου που ακολουθεί, θέλω να εκφράσω τις ειλικρινείς μου ευχαριστίες προς τα πρόσωπα των ανθρώπων που με την καθοδήγησή τους, τις συμβουλές τους και τη στήριξή τους συνέβαλαν στην πραγματοποίηση της παρούσας διατριβής.

Πρώτα από όλους, θέλω να ευχαριστήσω ολόψυχα τον επιβλέποντα της διδακτορικής μου κ. Νίκο Πλέρος, Επίκογυ Καθηγητή του τμήματος Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης. Η οικτετής και πλέον γνωμοδοτική ομάδα μας ξεκίνησε το 2009, όταν και ανάλογα να εκπληρώσει την πυρηνική μου εργασία ως προπτυχιακός φοιτητής υπό την επίβλεψή του. Ειλικρινά, ποτέ μου δεν είχα φανταστεί κατά την πρώτη μας συνάντηση πως θα έφτανε κάποτε η στιγμή που θα ολοκλήρωσα τη διδακτορική μου διατριβή υπό την επίβλεψή του, μία στιγμή που κουβαλά μαζί της χαρά και ευτυχία πως θα έφτανε κάποτε η στιγμή που θα ολοκλήρωσα τη διδακτορική μου διατριβή. Ήδη κατά τη διάρκεια της εργατικής μας συνεργασίας, η καθοδήγησή του έλαβε μια προσωπική σημασία για μας, κι αυτή η σχέση διέρρευσε κατά τη διάρκεια της εργαστικής μας συνεργασίας από τη διπλωματική μέχρι την ενημερωτική συνεργασία.

Πέρα από τους καθηγητές μου, θέλω επίσης να ευχαριστήσω τους ανθρώπους με τους οποίους συνέργασα ως συστηματικά όλα αυτά τα χρόνια. Ευχαριστώ θερμά τους κ. Κωνσταντίνο Βυρσωκινό, Λέκτορα του τμήματος Φυσικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, για τη βοήθεια που μου προσέφερε κατά τη διάρκεια εκμάθησης του λογισμικού προσομοίωσης VPI Photonics κατά τα πρώτα στάδια της
διατριβής μου. Επίσης, ευχαριστώ θερμά τον κ. Γεώργιο Κανέλλο, Λέκτορα του πανεπιστημίου του Bristol, τόσο για την καθοδήγησή του που έδωσε σχετικά με τη σχεδίαση των κυκλωμάτων οπτικής μνήμης όσο και για όλες τις εποικοδομητικές συζητήσεις που έχουμε κάνει κατά τη διάρκεια της διατριβής μου. Ευχαριστώ επίπλεον το μεταδιδακτορικό ερευνητή Δρ. Σάββα Γκιτζένη για τις συζητήσεις μας όσο και για τις συμβουλές του πάνω σε διάφορα ζητήματα που αφορούν τις αρχιτεκτονικές των σύγχρονων επεξεργασιακών συστημάτων. Θέλω επίσης να ευχαριστήσω το μεταδιδακτορικό ερευνητή Δρ. Δημήτρης Παπαγιάννη και Θεότητης Παπαδαβίδη για τις συζητήσεις μας κατά τη διάρκεια της διατριβής μου.

Δημήτρης Παπαγιάννη και Δημήτρης Παπαγιάννη, τους οποίους και ευχαριστώ για τη συνεργασία μας κατά τα χρόνια της συνύπαρξης μας στην εργασία του PhoS-net. Ευχαριστώ επίπλεον για τις συζητήσεις μας κατά τη διάρκεια της διατριβής μου.

Από την άψογη συνεργασία μας κατά τη διάρκεια της διατριβής μου, θέλω επίσης να ευχαριστήσω θερμά τον σκληρό μου συνεργάτη Νίκο Σερζένη για τις συζητήσεις μας κατά τη διάρκεια της διατριβής μου.

Με τιμή,
Πάυλος Μανιώτης

vi
Abstract

The unprecedented evolution of modern High Performance Computing Systems along with the predictions and expectations for exascale computing by near future have introduced the need for revising the internal interconnect and compute technologies and architectures towards meeting the demands of future computing systems. In this framework, optical interconnects and photonic integration technology fields emerge as promising alternative solutions for future computing environments by replacing conventional interconnects and circuitry with alternative high-bandwidth optical solutions. The research efforts of this thesis build upon the emerging optical interconnect technology aiming at resolving the long-standing “Bandwidth” and “Memory-Wall” problems and enabling exascale processing powers within the available energy boundaries. This thesis proposes new optical systems that exploit Optical Interconnect and Optical Memory Technologies and synergizes them with processors in innovative Computing Architectures for increasing bandwidth and reducing energy consumption. The novel optically-enabled concepts proposed in this thesis are spanning from complete High Performance Computing network environments down to chip-scale multi-core computing architectures.

At first, OptoHPC-Sim simulation platform is demonstrated which supports the system-scale utilization of novel electro-optical boards and routing technologies in complete and fully operational High Performance Computing network architectures. OptoHPC-Sim can effectively simulate optically enabled High Performance Computing network configurations and reliably compare them with conventional ones while at the same time incorporates a representative set of predefined modules that include all the recent advances around the Electro-Optical Printed Circuit Boards technology. OptoHPC-Sim forms an open tool targeting the adoption of Optical Technologies in future High Performance Computing configurations and can effectively support the architectural and technology decision making on the way to the exascale era. By using OptoHPC-Sim, an optical board-based High Performance Computing network architecture is proposed with the respective comparative simulation analysis demonstrating up to 190% mean throughput improvement and 83% mean packet delay reduction compared to world’s #3 Titan CRAY XK7 High Performance Computer network configuration.

Extending the use of optics from complete High Performance Computing to chip-level environments, this thesis demonstrates via physical layer simulations a complete optical
Abstract

cache memory layout that operates at speeds up to 16 Gb/s, which is significantly faster compared to any of the conventional technologies. The proposed cache memory layout offers a significant leap forward for optical memory technologies that by so far have been restricted to single-bit capacity layouts. Besides the memory elements, all the layouts for the necessary subsystems are being presented, introducing at the same time Wavelength Division Multiplexing in the memory domain. These subsystems include (a) an optical interface module that is used for translating the wavelength-formatted data and address fields between the processor and the cache memory unit, (b) a Read/Write Selector for ensuring correct access to the stored data, (c) a completely passive row decoder for addressing the different storing locations of the cache memory, (d) a 2D optical Random Access Memory bank for storing the actual data and (e) a tag-comparison circuit used for determining whether the desired data are stored in the cache memory unit. This pioneering design of an optical memory layout is demonstrated to perform successfully in both Write and Read functionalities at 16 Gb/s via optical physical layer simulations using the commercially available VPI software tool.

Going a step further towards highlighting the benefits of the proposed optical cache memory architecture in addressing the long-lasting “Memory Wall” problem in the computing industry, this thesis demonstrates a Chip-Multiprocessor architecture that uses the optical cache unit as a shared single-level Level-1 cache, discarding the complex cache hierarchy and offering significant speed and energy advantages to multi-core architectures. The Chip-Multiprocessor architecture is validated via Gem5 simulation engine, demonstrating that the optical cache-enabled multi-core architecture can significantly improve system performance, increase memory bandwidth and discard the need for complex coherency protocols. The simulation results suggest either an execution speed-up of 19.4% or a cache capacity requirements reduction of ~63% on average for the 12 benchmarks of PARSEC suite. The proposed Chip-Multiprocessor architecture offers also the possibility for using the optical L1 cache as off-die module relieving in this way valuable chip area.

Finally, in order to transfer the optical memory technology benefits from Chip Multiprocessor layouts also in novel all-optical routing table architectures, this thesis presents the first design of an all-optical Ternary-Content Addressable Memory (T-CAM) cell and a complete optical T-CAM row architecture for use in routing look-up table implementations. The Optical T-CAM row follows a novel Wavelength Division Multiplexing encoding matchline design, providing in this way successful comparison
operation for complete optical words. The proposed scheme allows for the essential subnet-masked operation that is needed in modern router applications, while its evaluation by means of physical-layer simulations reveals successful Search and Write operation at speeds of up to 20 Gb/s, i.e. >4x faster operation than the respective electronic counterparts.
Keywords

Περίληψη

Η πρωτοφανής εξέλιξη των σύγχρονων υπερυπολογιστικών συστημάτων σε συνδυασμό με τις προβλέψεις και προσδοκίες για exascale υπολογιστικές μηχανές στο μέλλον έχουν δημιουργήσει την ανάγκη για αναθεώρηση των τεχνολογιών που αφορούν τις εσωτερικές διασυνδέσεις, τα υπολογιστικά κυκλώματα αλλά και τις αρχιτεκτονικές οργάνωσης των μελλοντικών υπολογιστικών μηχανών. Σε αυτό το πλαίσιο, οι τεχνολογίες των οπτικών διασυνδέσεων και της φωτονικής ολοκλήρωσης αναδεικνύουνται ως πολλά υποσχόμενες εναλλακτικές για την αντικατάσταση των συμβατικών εσωτερικών διασυνδέσεων και κυκλωμάτων στα μελλοντικά υπολογιστικά περιβάλλοντα χάρη στο πολύ υψηλότερο εύρος ζώνης που προσφέρουν. Η έρευνα που έχει πραγματοποιηθεί στο πλαίσιο της παρούσης διατριβής έχει επικεντρωθεί γύρω από την επίλυση των μακροχρόνων προβλημάτων "Τείχους Μνήμης" και "Εύρους Ζώνης" και έχει ως στόχο την παρουσίαση προτάσεων που θα ενσαχύσουν τις προσπάθειες για την επίτευξη exascale υπολογιστικών μηχανών στο προσεχές μέλλον. Στην παρούσα διατριβή παρουσιάζονται καινοτόμα οπτικά συστήματα που κάνουν χρήση οπτικών τεχνολογιών διασύνδεσης και μνήμης τα οποία στη συνέχεια χρησιμοποιούνται για την παρουσίαση καινοτόμων υπολογιστικών αρχιτεκτονικών που έχουν ως στόχο τη συνωστική σύξηση της απόδοσης αλλά και τη μείωση της κατανάλωσης ενέργειας.

Η παρούσα διατριβή ξεκινάει με την παρουσίαση της πλατφόρμας προσομοίωσης OptoHPC-Sim η οποία έχει ως στόχο τη μελέτη της χρήσης καινοτόμων οπτικών τεχνολογιών σε συστήματα σε μελλοντικά συστήματα υπερυπολογιστών. Πέρα από τη μελέτη πλήρων αρχιτεκτονικών δικτύων υπερυπολογιστικών συστημάτων που κάνουν χρήση καινοτόμων οπτικών τεχνολογιών, ο OptoHPC-Sim προσομοιωτής μπορεί επίσης να χρησιμοποιηθεί και για την παρουσίαση συμβατικών αρχιτεκτονικών των σύγχρονων υπερυπολογιστών. Ο OptoHPC-Sim συμπεριλαμβάνει ένα αντιπροσωπευτικό σύνολο από μοντέλα προσομοίωσης τα οποία αντικατοπτρίζουν τις πρόσφατες τεχνολογικές εξελίξεις γύρω από τις οπτοηλεκτρονικές μητρικές κάρτες αλλά και τις οπτοηλεκτρονικές τεχνολογίες δρομολόγησης δεδομένων. Ο OptoHPC-Sim στοχεύει στην υιοθέτηση αυτών και περισσότερων οπτικών τεχνολογιών από τα μελλοντικά υπερυπολογιστικά συστήματα και μπορεί να χρησιμοποιηθεί καθώς διανύουμε το δρόμο προς την exascale εποχή. Κάνοντας χρήση του συγκεκριμένου προσομοιωτή προτείνεται μία καινοτόμα αρχιτεκτονική δικτύου για μελλοντικά υπερυπολογιστικά συστήματα η οποία κάνει χρήση σύγχρονων οπτικών τεχνολογιών. Σας αποτελέσματα της
συγκεκριμένης ανάλυσης μεταξύ της προτεινόμενης αρχιτεκτονικής και της αρχιτεκτονικής που χρησιμοποιείται στον Titan CRAY XK7 υπομονοπλακτή αναδιανύουν σημαντικά οφέλη
tόσο ως προς τη ρυθμοαπόδοση, η οποία παρουσιάζεται αυξημένη κατά 190% κατά μέσο όρο για 8 benchmarks, όσο και ως προς τη μέση καθυστέρηση πακέτου, η οποία παρουσιάζεται κατά 83% βελτιωμένη κατά μέσο όρο για τα ίδια benchmarks.

Επεκτείνοντας την χρήση των οπτικών τεχνολογιών από τις αρχιτεκτονικές δικτύων υπερυπολογιστών σε περιβάλλοντα επεξεργαστή, η παρούσα διατριβή παρουσιάζει μία καινοτόμα αρχιτεκτονική οπτικής κρυφής μνήμης η οποία επεκτείνει τις μέχρι τώρα διαδικασίες που περιορίζονται στη χωρητικότητα ενός και μόνο bit. Η ανάλυση απόδοσης της προτεινόμενης κρυφής κρυφής μνήμης μέσω προσομοίωσης φυσικού επιπέδου αναδεικνύει ταχύτητες λειτουργίας έως και 16 GHz, σημαντικά υψηλότερες συγκριτικά με οποιαδήποτε συμβατική τεχνολογία, διατηρώντας με αυτόν τον τρόπο το συνολικό διαθέσιμο εύρος ζώνης μνήμης. Πέρα από τις διαδικασίες αποθήκευσης των δεδομένων, παρουσιάζονται ένα πλήρες σύνολο απαραίτητων περιφερειακών υποσυστημάτων, οι οποίοι εισάγουν την έννοια της εκμετάλλευσης κύματος σε καμία μάθηση. Ιδίως, οι κύματα συμπληρώνουν τη σειρά μεταφοράς, κάνοντας έτσι δυνατή την πρόσβαση στην κρυφή μνήμη.

Πηγαίνοντας ένα βήμα παραπάνω προς την ανάδειξη των πλεονεκτημάτων της προτεινόμενης αρχιτεκτονικής κρυφής μνήμης ως προς την επίλυση του προβλήματος του “Σείχους Μνήμης”, η παρούσα διατριβή παρουσιάζει μία αρχιτεκτονική συστήματος πολυεπεξεργαστή όπου ορίζεται η χρήση της προτεινόμενης αρχιτεκτονικής κρυφής μνήμης ως διαμορφωθέντος κρυφής μνήμης κατά μέσο όρο.

Πηγαίνοντας ένα βήμα παραπάνω προς την ανάδειξη των πλεονεκτημάτων της προτεινόμενης αρχιτεκτονικής κρυφής μνήμης ως προς την επίλυση του προβλήματος του "Σείχους Μνήμης", η παρούσα διατριβή παρουσιάζει μία αρχιτεκτονική συστήματος πολυεπεξεργαστή που χρησιμοποιεί την προτεινόμενη αρχιτεκτονική κρυφής μνήμης ως διαμορφωθέντος κρυφής μνήμης κατά μέσο όρο. Η επιλογή της προτεινόμενης αρχιτεκτονικής κρυφής μνήμης ως διαμορφωθέντος κρυφής μνήμης κατά μέσο όρο εκμηδενίζει την ανάγκη για πολύπλοκες και πολυεπεξεργαστής επεξεργασίες συμβατικών κρυφών μνημών και κατά συνέπεια καταφέρνει να εκμηδενίζει και την ανάγκη για πολύπλοκα πρωτοκόλλα συνάφειας των δεδομένων. Η αξιολόγηση της απόδοσης της προτεινόμενης αρχιτεκτονικής πολυεπεξεργαστής γίνεται με τη χρήση της ευφέδως διαδεδομένης πλατφόρμας προσομοίωσης
Περίληψη

Gem5. Τα αποτελέσματα της συγκριτικής μελέτης μεταξύ της προτεινόμενης αρχιτεκτονικής και μίας συμβατικής αρχιτεκτονικής επέξεργαστή που κάνει χρήση μίας διαδικαστικής εισαγωγής χρωμάτων μνημών αναδεικνύουν σημαντική βελτίωση στη συνολική απόδοση του συστήματος.

Πιο συγκεκριμένα αυτό μπορεί να μεταφραστεί είτε σε 19.4% βελτίωση του χρόνου εκτέλεσης ή σε 63% μείωση της απαιτούμενης χωρητικότητας σε χρωμή μνήμη κατά μέσο όρο για 12 benchmarks από την PARSEC συστή. Επιπλέον, η προτεινόμενη αρχιτεκτονική πολυπεξεργαστή προσφέρει επίσης τη δυνατότητα τοποθέτησης της συμβατικής χρωμής μνήμης σε εξωραϊστό chip δίπλα στο chip του επεξεργαστή, απελευθερώνοντας με αυτό τον τρόπο σημαντικό χώρο προς όφελος των πυρήνων επεξεργασίας.

Τέλος, μεταφέροντας τα πλεονεκτήματα της τεχνολογίας της συμβατικής μνήμης από το επίπεδο των chip των επεξεργαστών στο επίπεδο των καινοτόμων αρχιτεκτονικών πινάκων δρομολόγησης για chip δρομολογητών, η παρούσα διατριβή παρουσιάζει το πρώτο πλήρως οπτικό και συμβατικό σύστημα συσχετιστικής μνήμης που υποστηρίζει τη χρήση μασκών υποδικτύου, μία λειτουργία απαραίτητη για όλους τους σύγχρονους δρομολογητές δεδομένων.

Χρησιμοποιώντας ως βασικό δομικό στοιχείο τη συγκεκριμένη κατεχόμενη το συγκεκριμένο κελί γίνεται επιπλέον η παρουσίαση μίας αρχιτεκτονικής χαρακτήρας συσχετιστικής μνήμης η οποία υλοποιεί ένα καινοτόμο μηχανισμό αναζήτησης των δεδομένων που βασίζεται στην αρχή της πολυπλεξίας κύματος. Ο συγκεκριμένος μηχανισμός εφαρμόζεται στα τελικά σήματα εξόδου όλων των κελιών της προτεινόμενης αρχιτεκτονικής χαρακτήρας και με αυτό τον τρόπο επιτυγχάνεται η ορθή λειτουργία σύγκρισης σε πλήρεις υποδικτές λέξεις δεδομένων.

Τα αποτελέσματα των προσομοιώσεων σε φυσικό επίπεδο παρουσιάζουν επιτυχή λειτουργία της προτεινόμενης αρχιτεκτονικής χαρακτήρας σε ταχύτητες μέχρι και 20 Gb/s, ταχύτητες οι οποίες είναι κατά πολύ υψηλότερες συγκριτικά με οποιαδήποτε αντίστοιχη συμβατική τεχνολογία.
# Table of Contents

ABSTRACT .................................................................................................................. VII
KEYWORDS ................................................................................................................ XI
LIST OF FIGURES ........................................................................................................ XIX
LIST OF TABLES .......................................................................................................... XXIII

CHAPTER 1 .................................................................................................................. 25
  1.1 INTERCONNECTION CHALLENGES FACED BY MODERN HIGH PERFORMANCE COMPUTING SYSTEMS .............................................................. 25
  1.2 EMERGING OPTICAL TECHNOLOGIES INTEGRATION IN MODERN HIGH PERFORMANCE COMPUTING SYSTEMS .................................................. 26
  1.3 SIMULATION ENGINES FOR EXPLORING OPTICAL TECHNOLOGIES INTEGRATION IN HIGH PERFORMANCE COMPUTING-LEVEL ........................................ 27
  1.4 DATA ACCESS CHALLENGES FACED BY MODERN CHIP-MULTIPROCESSOR AND ROUTER CHIPS ................................................................. 28
  1.5 THE OPTICAL SRAM TECHNOLOGY EVOLUTION .............................................. 31
  1.6 CONTRIBUTION AND STRUCTURE OF THIS THESIS ...................................... 33
  1.7 LIST OF REFERENCES ...................................................................................... 38

CHAPTER 2 .................................................................................................................. 49
  2.1 INTRODUCTION ............................................................................................. 49
  2.2 THE OPTICAL BOARD TECHNOLOGY PLATFORM CONSIDERED BY OPTO-HPC-SIM .................................................................................. 50
  2.2.1 HIGH END ROUTING PLATFORM USING OPTICAL INTERCONNECTS .................................................................................................................. 51
  2.2.2 MULTI-MODE ELECTRO-OPTICAL PCB TECHNOLOGY ................................ 55
  2.2.3 PASSIVE OPTICAL CONNECTOR AND POLYMER COUPLING INTERFACES .................................................................................................... 58
  2.2.4 FIBER AND POLYMER WAVEGUIDE FLEXPLANE TECHNOLOGIES ............ 60
  2.3 THE OPTO-HPC-SIM SIMULATION ENGINE ....................................................... 61
  2.4 EOPCB-BASED HPC NETWORK PERFORMANCE ANALYSIS AND COMPARISON WITH CRAY XK7 HPC NETWORK CONFIGURATION ........................................... 68
  2.5 CONCLUSION ............................................................................................... 76
  2.6 LIST OF REFERENCES ...................................................................................... 76

CHAPTER 3 .................................................................................................................. 81
  3.1 INTRODUCTION ............................................................................................. 81
  3.2 OPTICAL CACHE MEMORY ARCHITECTURE FOR DIRECT CACHE MAPPING ...................................................................................................... 82
  3.2.1 INTERFACES TO CPU AND MAIN RANDOM ACCESS MEMORY .................... 86
  3.2.2 READ/WRITE SELECTION STAGE ................................................................ 88
  3.2.3 ROW ADDRESS SELECTION STAGE & ROW ACCESS GATES ......................... 88
  3.2.4 2D OPTICAL RAM BANK: COLUMN ADDRESS SELECTOR (CAS) AND OPTICAL FLIP-FLOPS ................................................................. 90
  3.2.5 TAG COMPARATOR STAGE ......................................................................... 95
  3.3 SIMULATION RESULTS .................................................................................... 100
  3.4 OPTICAL CACHE SCALABILITY ANALYSIS ...................................................... 104
  3.5 EXPANDING THE OPTICAL CACHE MEMORY ARCHITECTURE FROM DIRECT TO 2-WAY ASSOCIATIVE CACHE MAPPING SCHEME .................................................. 109
  3.6 CONCLUSION ............................................................................................... 111
  3.7 LIST OF REFERENCES ...................................................................................... 112

CHAPTER 4 .................................................................................................................. 117
  4.1 INTRODUCTION ............................................................................................. 117

xvii
Table of Contents

4.2 OPTICAL-BUS-BASED CHIP-MULTIPROCESSOR ARCHITECTURE WITH OPTICAL CACHE MEMORIES .......... 118
4.2.1 OPTICAL INTERFACES .................................................................................. 121
4.3 SYSTEM-LEVEL SIMULATION RESULTS ............................................................ 124
4.3.1 CMP CACHE HIERARCHIES ........................................................................ 126
4.3.2 SIMULATION PARAMETERS ....................................................................... 128
4.3.3 BODYTRACK MISS RATES FOR VARYING NUMBER OF CORES .................. 130
4.3.4 MISS RATES FOR ALL PROGRAMS .............................................................. 132
4.3.5 EXECUTION TIMES FOR ALL PROGRAMS ..................................................... 135
4.4 DISCUSSION .................................................................................................. 139
4.5 CONCLUSION ............................................................................................... 142
4.6 LIST OF REFERENCES .................................................................................. 143

CHAPTER 5 ....................................................................................................... 147

AN ALL-OPTICAL TERNARY-CONTENT ADDRESSEE MEMORY (T-CAM) ROW ARCHITECTURE FOR ADDRESS LOOKUP AT 20 Gb/s .................................................................................................................. 147
5.1 INTRODUCTION .................................................................................................. 147
5.2 OPTICAL T-CAM CELL AND ROW ARCHITECTURES ..................................... 148
5.3 SIMULATION RESULTS .................................................................................. 151
5.4 CONCLUSION .................................................................................................. 153
5.5 LIST OF REFERENCES .................................................................................. 153

CHAPTER 6 ....................................................................................................... 155

CONCLUSIONS AND FUTURE WORK .................................................................... 155
6.1 INTRODUCTION .................................................................................................. 155
6.2 THESIS CONCLUSIONS .................................................................................. 155
6.3 FUTURE WORK .................................................................................................. 158

ANNEX A: PUBLICATIONS .................................................................................. 161

LIST OF PUBLICATIONS .................................................................................... 161
A.1 JOURNAL PUBLICATIONS .............................................................................. 161
A.2 CONFERENCE PUBLICATIONS ...................................................................... 162

ANNEX B: ABBREVIATIONS ............................................................................... 165

LIST OF ABBREVIATIONS .................................................................................... 165
List of Figures

Fig. 1.1: Data is growing at a 40% compound annual rate, reaching nearly 45 zettabytes by 2020. [1.2] .................................................................25

Fig. 1.2: (left) SPARC T5 Processor specifications (~40% cache area) [1.46], (right) Nvidia Fermi GPU Processor specifications (~30% cache area) [1.49] .........................................................29

Fig. 1.3: Processor – DRAM Memory performance gap. [1.53] .................................................................30

Fig. 2.1: OptoHPC-Sim’s main GUI frame demonstrating an example HPC model incorporating 4 racks. Each rack consists of 24 OPCBs being grouped in 3 chassis of 8 OPCBs each. An EOPCB design with 2 router interfaces is demonstrated at the bottom .........................................................................50

Fig. 2.2: Schematic outline of the hybrid optical interconnect showing chip and optical packaging (top) and the actual chip with the CMOS die and assembled VCSEL and PD matrices (bottom, A) and the packaged chip (bottom, B & C) with a cutout hole in the package for optical coupling [2.7] ............................................................................53

Fig. 2.3: Optical coupling of the on-chip optical interconnect (A) with a 2D fiber bundle assembled on the PCB (B); coupling to a double-layered embedded waveguide array in a PCB using microlens and prisms (C). [2.7] ............................................................................54

Fig. 2.4: PRBS31 eye diagrams from a 12×14 VCSEL matrix at 8Gb/s line-rate. [2.4] .................................................................54

Fig. 2.5: Optical/Electrical PCB demonstrator with 16 copper and 1 optical embedded layer. Black rectangles at Router areas represent the optical I/Os that couple to the underlying embedded waveguide layer following the concept of Fig. 2.3 (c). MT#1 and MT#2 areas are targeted for vertical out-of-board connections via MT Termination Push-On sites for fiber-to-waveguide connections. Optical waveguide tracing layout is shown in blue. .........................................................55

Fig. 2.6: (a) Schematic of chip-to-waveguide coupling concept (b) folding element comprising of beam splitter and microlens array (MLA), (c) cut out section in OEPCB with waveguide illuminated. ........................................................................56

Fig. 2.7: (top) EOPCB with embedded MM polymer waveguide layer fabricated by TTM Technologies, (bottom) Cross-section of the EOPCB showing 14+14 waveguides .................................................................57

Fig. 2.8: a) Electro-optical backplane with embedded waveguides, b) out-of-plane receptacle connected to an MT ferrule, c) Out-of-plane receptacle passively aligned onto optical waveguide interface, d) polymer waveguide test board with generic waveguide interfaces and waveguide illuminated with 650 nm light from out-of-plane test cable .........................................................59

Fig. 2.9: Optical fiber flexplanes deployed for an optically enabled data storage and switch test platform for data centers [2.23]: a) Photo of electro-optical midplane with MT terminated flexplane, b) Schematic view of Prizm MT terminated flexplane ........................................................................61
List of Figures

Fig. 2.10: A single rack 3D-Torus topology example where a total number of 24 PCBs are organized in groups of 8 PCBs where each of the group forms a chassis. Each PCB incorporates 4 computing nodes.................................................................64

Fig. 2.11: OptoHPC-Sim’s PCB-layer view where two router modules are connected together using the link module where each of them is directly connected to two node modules by using again another instance of the link module.................................................................64

Fig. 2.12: 1x4x1 torus layout for a) a CRAY XK7 blade and b) a Dual Layer EOPCB ..............................................69

Fig. 2.13: DOR vs MOV R algorithm in an HPC system with Conventional Router configuration and for both Uniform Random and Nearest Neighbor traffic patterns.................................................................70

Fig. 2.14: Throughput simulation results for 8 synthetic traffic profiles .................................................................72

Fig. 2.15: Mean packet delay simulation results for 8 synthetic traffic profiles.................................................................73

Fig. 3.1: Direct cache mapping .......................................................................................................................82

Fig. 3.2: (a) Optical cache memory architecture (b) Memory Address and Data Interfaces .................83

Fig. 3.3: Flowchart of proposed optical Cache operation..................................................................................86

Fig. 3.4: (a) Row Address Selector (RAS) architecture (b) combined spectrum of four ring resonator-based filters used in the RAS (c) RAS truth table (d) Ring resonator-based Filter design parameters.................................................................89

Fig. 3.5: Coupled-SOA-MZI-based optical Flip-Flop .......................................................................................91

Fig. 3.6: (a) Write mode operation showing the complete path of a Data bit and its inverted value through the different cache memory stages, Time scale: 62.5ps/div. (b) Timing diagram for the Write mode operation .................................................................92

Fig. 3.7: Spectrum of: (a) Line and input signals at the RAS input (b) Line and input signals at the RAS output (c) 5-bit Tag and 5-bit at the input of the Row “00” AG1 (d) Row “00” AG1 output .........................................................................................................................95

Fig. 3.8: (a) Tag Comparator (b) SOA-MZI-based XOR gate .................................................................96

Fig. 3.9: Read mode operation example showing the complete set of XOR Tag Comparisons and the path of a stored Data bit to the Data-to-Read Interface. Time scale: 62.5ps/div for the pulse traces - 20ps/div for the eye diagram .................................................................98

Fig. 3.10: Timing diagram for the Read mode operation ..................................................................................99

Fig. 3.11: Write mode operation showing the storing of an 8-bit Data word in the 2D Optical RAM bank. Time scale: 62.5ps/div .................................................................101
List of Figures

Fig. 3.12: Read mode operation example showing the reading of an 8-bit Data word from the 2D Optical RAM bank and its transmission to the Data-to-Read Interface. Time scale: 62.5ps/div. ..................................................................................................................103

Fig. 3.13: (a) Schematic representation of the scaled optical cache architecture (b) 8-bit Column Address Selector based on PhC-based AG and filtering elements (c) 8-bit Row Address Selector based on PhC-based filtering elements (d) Tag Comparator employing PhC nanocavity-based XOR gates..................................................................................105

Fig. 3.14: 2-Way set associative cache mapping scheme ..........................................................................................................................109

Fig. 3.15: (a) Optical cache memory architecture. (b) Flowchart of the proposed optical cache operation..................................................................................................................................................110

Fig. 4.1: (a) Conventional CMP architecture with on-chip Cache Memories and Electrical Bus for CPU-MM communication (b) The proposed CMP architecture with off-chip optical Cache Memories between CPU-MM and Optical Busses between them (c) The CPU-L1d bus’ communication layers (d) The CPU-L1d bus’ 64-bit Data-Read Layer comprising 8 waveguides ..........................................................................................................................119

Fig. 4.2: Optical interfaces utilized in CPU cores and MM: (a) WDM Optical Transmitter Interface (b) WDM Optical Receiver Interface. Both Transmitter and Receiver interfaces are followed by operation examples presenting the TDM-based bus/cache access scheme. The examples assume 4 processing cores and a 4x optical system operation speed (compared to electronic cores) ..................................................................................................................................................122

Fig. 4.3: CMP Architectures: Conventional L1+L2 cache (up) and Optical L1 cache (down). The Conventional L1 (not shown) is the same with the Conventional L1+L2 but without the L2 module. ..................................................................................................................................................125

Fig. 4.4: The miss rates of L1i and L1d caches for the Conventional L1, Conventional L1+L2 and Optical L1 architecture for varying number of cores N. ..................................................................................................................................................131

Fig. 4.5: Miss rates simulation results for 12 PARSEC programs in an 8-core CMP..................................................................................133

Fig. 4.6: Execution time for the 12 programs of PARSEC and N = 8 cores ...............................................................................................136

Fig. 5.1: (a) T-CAM-based routing table along with highlighted operation example, (b) T-CAM row architecture that comprises an indicative number of 4 T-CAM cells. ..................................................................................................................................................148

Fig. 5.2: all-optical T-CAM cell architecture with 2 FFs (TCFF & XFF) and a XOR gate and T-CAM row’s AWG multiplexer for 4 indicative T-CAM cells. ..................................................................................................................................................150

Fig. 5.3: 20 Gb/s simulation results for the T-CAM row architecture of Fig. 1 (b). Time scale: 50ps for traces and 25ps/div for eye diagrams ........................................................................................................................................152
Fig. 6.1: Possible extension for the optical bus-based CMP architecture of Fig. 4.1 (b) where multiple L1 cache units are being interconnected with an optical crossbar or an AGWR .................. 159
List of Tables

Table 2.1: Router Configurations’ IO Capacities ............................................................................. 69
Table 2.2: Simulation Parameters ...................................................................................................... 70
Table 3.1: Wavelength assignment on Data and Tag bits .................................................................. 87
Table 3.2: Operating Conditions for the cache subsystems ................................................................. 104
Table 3.3: Energy consumption and footprint for different electrical and optical cache memory* technologies .......................................................................................................................... 107
Table 4.1: Simulation Parameters ...................................................................................................... 129
Table 4.2: Classification of the Programs of PARSEC suite ............................................................... 130
Table 4.3: L2 Miss rates for the all the 12 PARSEC benchmarks.......................................................... 137
Table 4.4: Performance gain for the Optical architecture compared to the Conventional L1+L2 .... 137
Chapter 1
Introduction: Background and Contribution of this thesis

1.1 Interconnection Challenges Faced by Modern High Performance Computing Systems

The predictions and expectations for exaflop High Performance Computing Systems (HPCs) by 2020 \[1.1\] rely mainly on the aggregation of vast numbers of Chip-Multiprocessors (CMPs) within the HPC platforms, constantly pushing the performance envelope at all three critical factors: (1) bandwidth, (2) latency and (3) energy efficiency. The capacity and timing limitations imposed by conventional interconnect technology come in contrast to the increasing number of computing nodes inside modern HPC platforms, forming in this way one of the main sources of bottleneck in data exchange across all the network hierarchy communication levels: (a) rack-to-rack, (b) backplane, (c) chip-to-chip and (d) even on-chip. At the same time, the extended parallelization of modern applications associated with the “data deluge” problem, has introduced unprecedented requirements with respect to the amounts of traffic that have to be exchanged within modern HPC environments, essentially transforming computing problems into communication problems. Only in 2012 for example, an estimated 2.5 zettabytes of business data were generated and current trends

![Data in zettabytes (ZB)](image)

Fig. 1.1: Data is growing at a 40% compound annual rate, reaching nearly 45 zettabytes by 2020. \[1.2\]
and estimations indicate that the volume of generated data will continue growing significantly every year with a 40% compound annual rate [1.2], reaching nearly 45 zettabytes by 2020 as can be seen in Fig. 1.1. Taking into account that poor data management can cost up to 35% of a business’s operating revenue [1.2] and since conventional interconnect systems are still comprising a major bottleneck, optical interconnect and photonic integration technologies are being promoted as highly promising interconnect solutions with the aim to translate their proven high-speed, low-latency and energy-efficient data transfer advantages into respective benefits at system-level.

1.2 Emerging Optical Technologies Integration in Modern High Performance Computing Systems

Optics are rapidly replacing electrical interconnects with Active Optical Cables (AOCs) forming already a well-established technology in rack-to-rack communications. At the same time, mid-board optical subassemblies and compact board-level flexible modules, like FlexPlane [1.1], have recently entered the market targeting the replacement of conventional on-board interconnects for chip-to-chip communication purposes while at the same time emerging optical technologies are continuously penetrating at deeper hierarchy levels.

In [1.3] a mid-board optical transceiver with 12 Coolbit Optical Engines [1.4] is presented, performing at ~25.8 Gb/s and offering a total bandwidth of 300 Gb/s per square inch. The Coolbit Engine development process begins with the semiconductor fabrication of the Vertical Surface Emitting Laser (VCSEL) and Photodiode Integrated Circuits (ICs), moves to the automated wafer assembly of the VCSEL, photodiode and other ICs and ends with the operational testing of the wafer. Moreover, INTEL has introduced the hybrid silicon 100G Parallel Single Mode 4-channel (PSM4) Quad Small Form-factor Pluggable-28 (QSFP-28) Optical Transceiver [1.5], targeted for use in optical interconnects for data communications applications. Luxtera from the other side has successfully entered the Silicon Photonics market [1.6] by presenting the LUX62608 OptoPHY and LUX42604 QSFP optical transceiver modules [1.7] that offer a small form-factor, high speed, and low power consumption solution for Cloud Data Interconnect, Local Area Network (LAN) and HPC applications.

Going a step further, Optical Printed Circuit Board (OPCB) layouts can offer high-density, energy efficient and low-loss Tb/s on-board data transmission forming a
promising solution for completely replacing the copper printed wires and their associated low bandwidth and distance- and speed-dependent energy dissipation problems. OPCBs have successfully revealed various optical wiring solutions like (a) optical fiber to the board [1.8], (b) optical polymer waveguides to the board [1.9]-[1.13], (b) embedded optical polymer waveguides [1.14], [1.15] and (c) embedded glass optical waveguides [1.16], [1.17], [1.8], while at the same time very high density parallel interfaces have been presented [1.18], [1.19]. Single-layered and single-mode arrays of optical waveguides in OPCBs have been recently presented to offer as low as 0.6 dB/cm propagation losses at 1310 nm and a total density of 50 wires/cm [1.20]. Bringing multiple optical layers hybridly integrated in Electro-Optical Printed Circuit Board (EOPCB) layouts with several electrical interconnect layers comprises the next big goal towards increasing the number of wiring and routing paths, with recent works reporting already on successful implementations of multi-layer embedded optical [1.17] and polymer [1.21], [1.22] waveguides. Although extensive hardware reliability studies have yet to be done [1.23] and low-cost mass-manufacturing processes have to be deployed for enabling their market adoption, the EOPCB technology holds much promise for eliminating the low bandwidth and distance- and speed-dependent energy dissipation problems originating from copper printed wires.

This roadmap, combined with the rapid progress on mid-board optical transceiver chips [1.10], [1.24]-[1.26] has also triggered expectations for on-board optoelectronic routing schemes either via optically interfaced electronic router Application-specific integrated circuits (ASICs) [1.27], or via silicon photonic switching platforms [1.28]. After the successful examples of circuit-switched optical solutions in Data Center environments [1.29], [1.30], the approach of on-board optically enabled routing seems to gain momentum as the line-rates of ASIC I/O ports reached already 25Gb/s [1.31]–[1.32]. Bringing optics as close as possible to the ASIC I/Os can yield significant power benefits at board-level signal routing, mimicking the case of the board-to-board connectivity where the recent release of fiber-coupled router ASIC from Compass EOS allows for just 10pJ/bit consuming optical Input/Output (IO) ports [1.27].

1.3 Simulation Engines for Exploring Optical Technologies Integration in High Performance Computing-Level

However, the rapid progress witnessed in the fields of board-level optical interconnects and optoelectronic routing technologies has still not been proven neither
Chapter 1

tailored nor reflected in system-scale benefits in HPC environments. Although advantages at link-level are being thoroughly addressed, the EOPCB layout and the performance of a complete HPC engine that exploits EOPCBs and performs with workload applications is usually still an unknown parameter. One main reason for the disassociation between hardware technology development and HPC-scale performance lies also in the lack of a corresponding system-scale simulation engine that would allow for optimally exploiting the new technology toolkit through performance evaluation at HPC level. Although photonics have already emerged in chip-scale simulation platforms like PhoeniXSim [1.33] suggesting optimal technology and network architecture design rules through system-scale performance [1.34], state-of-the-art sophisticated HPC simulators still cannot efficiently support the use of advanced electro-optic router and interconnect solutions at board-level. Among the few HPC open-source simulators that are free of charge and available to the research community, none of them is focused on or can even efficiently explore the adoption of optical technology advancements in the HPC field. The Extreme-scale Simulator [1.35] implements a parallel discrete event HPC simulator but is mainly targeting the investigation of parallel applications’ performance at extreme-scale Message Passing Interface (MPI) environments. SST+gem5 [1.36] is a scalable simulation infrastructure for HPCs and comes as the result of the integration of the highly detailed gem5 performance simulator into the parallel Structural Simulation Toolkit (SST). SST is a system of disparate hardware simulation component entities integrated via a simulator core, which provides essential services for interfacing, executing, synchronizing and monitoring the various components with gem5 [1.37] being integrated as one of them. However, gem5 gives emphasis in simulating detailed Central Processing Unit (CPU)-cores and computer memory hierarchies, yielding high simulation times due to its highly-detailed CMP hardware models.

1.4 Data Access Challenges Faced by Modern Chip-Multiprocessor and Router Chips

Going down through the HPC hierarchy and focusing now on the processor chip level, the unprecedented high processing speeds released by modern CMP configurations [1.38]-[1.40] have put a tremendous pressure on the shoulders of memory units and the respective interconnect facilities. Solutions such as deployment of large on chip cache memories, the widening of the CPU-Main Memory (CPU-MM) buses and prefetching from the MM have been devised to ease the limited off-chip bandwidth and the high MM’s response latency imposed by the constraints of the electronic technology [1.41].
Higher spatial multiplexing degrees through wider buses allow for more efficient and simultaneous multi-bit data transfer within a single cycle. On the other hand, trading bandwidth for reduced average delay and buffering data close to CPU in anticipation of future requests through prefetching reduced on average the access delay stalling processing. However, the high degree of parallelism introduced by modern CMP configurations has further aggravated the bottleneck between the CPU and MM and led to larger two- or even three-level cache memory hierarchies that take up almost 40% of the total chip energy consumption [1.42] and more that 40% of chip real estate [1.43]-[1.45]. This trend can be seen for example in Fig. 1.2 where SPARC T5 processor [1.38], [1.46] and Nvidia Fermi Graphics Processing Unit (GPU) [1.47]-[1.49] chips devote ~40% and ~30% of their chip real estate respectively in cache memory circuitry. Equipping the CMP module with high amount of hierarchical caches that struggle for a position closer to the CPU is currently the main line for coping with the well-known and continuously increasing processor-memory performance gap [1.50], which was already 20 years ago identified as the main bottleneck in further system-level power advances, commonly referred to as the “Memory Wall” [1.50]. Fig. 1.3 presents the “Memory Wall” problem and as can be seen the Processor – DRAM memory performance gap is growing with a ~50% rate every year.

The origin of the bottleneck lies mainly in the electronic interconnect performance constraints associated with low bandwidth densities and speed- and distance-dependent

Fig. 1.2: (left) SPARC T5 Processor specifications (~40% cache area) [1.46], (right) Nvidia Fermi GPU Processor specifications (~30% cache area) [1.49]
energy dissipation, as well as in the limited memory bandwidth and long access times provided by current Electronic Random Access Memory (RAM). The most important steps forward for coping with these limitations were the introduction of memory hierarchy and the widening of CPU-memory buses. Wider buses allowed for speed enhancement through the exploitation of spatial multiplexing, using multiple wires for simultaneous multi-bit transfer. At the same time, memory hierarchy led to cache memories based on Static RAMs (SRAMs) and closely located to the CPU [1.51]-[1.52]. In this way, a small amount of data that is most probable to be accessed by the CPU is stored in the first level of proximity so as to reduce access times. Only when a piece of data is unavailable in this memory level, it will be sought in the next slower-performing level of hierarchy. For example, Intel 486 appeared in 1989 and was the first Intel processor with a Level-1 on-chip cache that tried to cope with the “Memory Wall” problem by incorporating both processing and buffering elements on the same chip while Intel Itanium 2 appeared in 2003 was the first Intel processor that incorporated a 3-Level cache hierarchy on the processor die [1.53].

However, the divergence between CPU and memory speeds has not still been fully undone, remaining still a daunting issue and calling for interventions also to the basic hardware apart from the architectural landscape. Efforts towards the deployment of faster electronic SRAM cell layouts are still on-going, including 6T [1.54]-[1.56], 8T [1.57] and 10T cells [1.58], while the exploitation of Fin Field Effect Transistor (FinFET) technology for tri-gate architectures allowed for superior levels of scalability [1.56]. Moreover, hybrid cache architectures have emerged with their intension being to replace SRAM cache with future memory technology [1.59].
Moving from CMPs to modern Router chip configurations we observe that the always increasing need for faster buffering technologies is also present. The tremendous increase in Internet Protocol (IP) network traffic is continuously pushing for faster Address Look-up (AL) speeds in modern router applications [1.60]. On top of that, the massive growth of the devices connected to the Internet dictates the utilization of large Routing Tables [1.61] that turns fast AL operation even more challenging. Hence, this has led to the need for hardware based AL solutions in order to cope with the need for faster routing operations [1.62]. Towards achieving low-latency routing operation Content Addressable Memories (CAMs) have entered the game; CAMs are a special type of SRAM memory targeted to latency-sensitive search applications that allow the realization of search operations within the time frame of a single clock-cycle [1.63]. In order to accomplish this and assuming a 2-Dimension (2D) array of \( n \) CAM-cells, each CAM-cell embodies also a XOR gate that allows for parallel search throughout the complete 2D array [1.63].

However, the AL operation speed is still limited by the relatively low operation speeds imposed by electronic logic circuits and by the electronic interconnects, which are employed for connecting the different memory cells and the 2D CAM-cell array with the forwarding table that usually exploits RAMs. Despite the impressive optical packet payload data-rate increase that took place during the last years [1.64], modern electronic CAMs can barely offer AL speeds of few GHz [1.65].

1.5 The Optical SRAM Technology Evolution

Facing these problems, the rapidly progressing fields of optical interconnects and photonic integration appear as potentially promising alternative technologies for future computing architecture environments [1.66]. The replacement of electrical cabling with optically connected CPU-Dynamic RAM (DRAM) technologies has already outlined the benefits in several experimental demonstrations so far [1.67]-[1.71] where the main effort has shifted to the replacement of the electronic busses with optical wires. With the current technology, fetching 256 bits of operand from the MM module consumes more than 16 nJ [1.72] and requires four stages of transmission (assuming a 64-bit wide bus at an operation speed just above 1GHz). Bringing photonics into the game by replacing the electrical busses with optical wiring solutions, either over a Silicon-on-Insulator (SOI) platform or over an OPCB, is expected to (a) reduce energy consumption down to 1mW/Gb/s [1.73], (b) raise operation speed to several tens of GHz, and at the same
time, (c) dispense with the traditional issue of Resistance-Capacitance (RC)-induced delay of the electrical wiring. In addition, the maturing silicon photonics platform was recently utilized for introducing a whole new class of optical interfaces for DRAM integration [1.67], fully equipped with modulator and photodiode elements [1.68], [1.69], so as to expedite optically connected CPU-memory solutions. Inter-core interconnects through silicon photonic waveguides have been also proposed, improving throughput by almost an order of magnitude [1.70] and leading to micro-processor architectures relying on Silicon-photonic multi-bus Network-on-Chip (NoC) solutions [1.74]. However, all these enhancements cannot mitigate the need for memory caching: CMP dies will continue to struggle to find an optimum balance in the processor, cache and interconnect circuitry considerations.

Research in optics has been active also in offering high-speed memory cells as an approach to avoid the long access times associated with electronic RAM cells. Although the lack of electric charge places photons at disadvantage when coming to storage, the combined utilization of photonic integration and of novel circuit architectures has led to novel optical memory technologies with significant speed and energy benefits [1.75]-[1.80]. A variety of all-optical Flip-Flops (FFs) has been presented so far, including layouts based on coupled Semiconductor Optical Amplifiers (SOAs) [1.75], III-V-on-Silicon On Insulator (SOI) microdisk lasers [1.76], polarization bistable VCSELs [1.77] as well as coupled SOA-based Mach-Zehnder Interferometers (SOA-MZIs) [1.78]. Photonic crystal (PhC) nanocavities have also recently reached remarkable standards with respect to speed, energy and footprint [1.79], already demonstrating multi-cell integrated memory modules with more than 100 bit storage capacity [1.80] However, no true random access read/write capabilities as would be required in functional cache storage were reported in all of the above cases and functionality was limited to FF-based storage operation.

The demonstration of the first optical SRAM cell performing successfully up to 5 Gb/s [1.81] was accomplished by exploiting two SOAs as Access Gates and a coupled SOA-MZI-based FF, with its performance analysis suggesting Read/Write speed capabilities up to 40Gb/s [1.82]. Subsequent optical RAM cell developments reduced the number of active elements through the introduction of wavelength diversity in the incoming signals [1.83]. More recently, the successful adoption of Wavelength Division Multiplexing (WDM) principles into optical RAM architectures [1.84] has highlighted the advantages that can be gained also in RAM peripheral circuits like Column Decoding.
(CD) [1.85] and Row Decoding (RD) [1.86], which can simplify the RAM row architectures when using WDM-formatted word and address fields. On top of that, the first effort towards breaking the speed barrier associated with conventional CAMs has been recently presented in [1.87] by experimentally demonstrating the first optical Binary-CAM (B-CAM) cell utilizing a monolithically integrated SOA-MZI-switch-based FF. Despite being the first prototype, this optical B-CAM cell implementation offers already an operation speed of up to 10 Gb/s, i.e. 1.5x times faster compared to the fastest conventional CAM cell [1.87]. However, for practical routing AL tables where subnet-masked routing has to be accomplished, optical CAMs are required to support also the “X” or “Care/Don’t Care” operation besides the simple comparison offered by B-CAMs [1.87].

All the advantages offered by optical memory units and RAM devices presented above could be in principle translated also to respective speed and energy benefits of complete optical cache memories. This requires, however, appropriate optical cache architectures that will be capable of incorporating the unique characteristics of optical storage. So far, no optical cache memory layouts have been attempted, with the vast experience of electronics [1.88] suggesting that a number of functional subsystems are additionally needed to get implemented in the optical domain and yet properly interconnected to allow for successful optical cache operation. A Read/Write selection module has to ensure whether Read or Write functionality will be performed from or to the cache content, respectively. In addition, memory address fields have to get translated into the proper cache row and column combination where the data words will be read from or written to, necessitating the use of Row and Column address selection circuits. Given the smaller size of cache memory compared to the MM module, a tag comparison circuit is also required to decide whether the set of data requested by the processor resides in the cache module and to determine if we have a cache hit or miss [1.88]. Finally, all these subsystems have to perform both in the case of CPU-cache and in the case of cache-MM communication, the latter being realized through the employment of an additional Memory Controller (MC) unit at the electronic Main RAM Memory interface.

1.6 Contribution and Structure of this Thesis

The research efforts of this thesis have been shaped around resolving the long-standing “Bandwidth” and “Memory Wall” problems towards enabling exascale
processing powers within the available energy boundaries. This thesis proposes new optical systems and synergizes them with processors in innovative Computing Architectures for increasing overall system performance in terms of application execution time, memory and interconnection bandwidth, latency and energy consumption. The remaining part of the thesis consists of 5 chapters, whereas the contents and the contributions of each chapter are described in detailed as follows:

- **Chapter 2:** Towards addressing the lack of proper system-scale simulation engines that would allow for exploring the integration of novel optical interconnect technologies into future HPC systems as it has been described in section 1.3, chapter 2 introduces the OptoHCP-Sim simulation platform that supports the use of novel electro-optical boards and routing technologies in complete and fully operational HPC network architectures. As such, OptoHPC-Sim can effectively simulate optically enabled HPC network configurations and reliably compare them with conventional ones. The development of OptoHCP-Sim has been dedicated in offering a valuable contribution to the research community by effectively supporting the architectural and technology decision making on the way to the exascale era. In its first version OptoHPC-Sim incorporates a representative set of predefined modules that include all the recent advances around the EOPCB technology which are being presented in section 2.2.

  OptoHPC-Sim forms a powerful, modular and light-weight solution and has been implemented on top of the Omnet++ discrete event simulation framework [1.89]. It relies on a careful balance between the model detail and the simulation execution time, employing a queue-based HPC model and including only the absolutely necessary details for reliably evaluating an optically enabled HPC system. OptoHPC-Sim is a fully reconfigurable and extendable platform and supports both mesh and torus network topologies, two different routing algorithms, virtual channels for deadlock-avoidance, Store-and-Forward (SF) and Virtual-Cut-Though (VCT) [1.90] flow control methods, and a user-friendly Graphical User Interface (GUI) that allows the detailed exploration of the complete HPC topologies and can successfully be used for both demonstration and education purposes. It incorporates 8 synthetic traffic profiles [1.90] and can also be fed with custom trace files or user-defined statistical distribution traffic profiles.

  Besides describing the simulator features, chapter 2 proceeds to a comparative performance analysis between a system following the Titan Cray XK7 network
specifications (Titan Cray XK7 has been ranked as the world’s no. 3 Supercomputer as of June 2016 [1.91]) and a respective HPC architecture where electronic CRAY blades have been replaced by a completely optical blade design that makes use of novel optical technologies. The results obtained with OptoHPC-Sim suggest that the employment of board-level optics in appropriate layouts can lead to optically enabled HPC network system configurations that can significantly outperform top-class HPC machines, on average offering throughput improvements higher than ~190% as well as mean packet delay reduction of ~83% for the 8 synthetic traffic profiles that come with OptoHPC-Sim.

- **Chapter 3**: Towards addressing the data access challenges associated with modern CMPs, chapter 3 demonstrates for the first time a complete and fully functional optical cache memory physical layer architecture that is capable of performing Write and Read functionalities directly in the optical domain. The proposed optical cache has been designed for 64-bit storage capacity organized in 4 cache lines and is following a direct mapped cache associativity scheme with a 5-bit tag field. Its performance has been evaluated through physical layer simulations using the commercially available VPI Photonics simulation platform [1.92], showing successful Read and Write operation at speeds up to 16 GHz.

The optical cache memory architecture introduces WDM in the memory domain while five discrete subsystems that are designed relying on already experimentally demonstrated optical building blocks are presented: (i) an optical interface module that is used for translating the wavelength-formatted data and address fields between the processor and the cache memory unit, (ii) a Read/Write Selector for ensuring correct access to the stored data, (iii) a completely passive row decoder for addressing the different storing locations of the cache memory, (iv) a 2D optical RAM bank for storing the actual data and (v) a tag-comparison circuit used for determining whether the desired data are stored in the cache memory unit. Subsequently, a scalability analysis of the proposed architecture is demonstrated where various state-of-the-art optical technologies have been taken into account. The results present comparable footprint and energy consumption measurements with the respective conventional technologies. Finally, chapter 3 demonstrates that minor modifications to the proposed cache architecture can allow for supporting the 2-way set associative mapping scheme, extending in this way the flexibility of
the proposed architecture.

- **Chapter 4**: In order to highlight the transformative character of the optical cache architecture when employed in CMP architectures as well as its credentials to address the “Memory Wall” problem, chapter 4 proceeds to the demonstration of an optically-enabled CMP architecture that makes use of an off-chip shared single-level (L1) optical cache memory that negates the need for complex cache hierarchy schemes. In the proposed scheme the all-optical cache memory is connected to CPU and MM via optical buses based on waveguides that make use of both Wavelength Division and Spatial multiplexing. For the optical buses design implementation, various novel optical interconnect technologies have been taken into account and can offer significantly lower energy consumption as well as significantly better performance. In contrast to the conventional hierarchy paradigm, the proposed approach relies on the basis that the optical cache memory can operate significantly faster than the electrical cores; the cycle of the optical cache memory is a fraction of the cycle of the electronic cores, making thus possible to serve multiple concurrent core requests without stalling the execution. The proposed optical cache CMP architecture is, hence, a single-level shared L1, which sidesteps the issue of data consistency in the conventional paradigm of dedicated per core L1 caches, since only one copy of the data exists in the shared L1. Moreover, the shared L1 cache of the proposed scheme is placed off-chip, sparing in this way precious chip area from the die in favor of the processing elements.

Chapter 4 uses the proposed optical cache-based CMP architecture and proceeds to its comparison with a typical CMP hierarchy of dedicated L1 and shared L2 electronic caches. The system-scale performance is addressed for 12 parallel workloads, using the PARSEC benchmark suite [1.93] on top of the Gem5 simulator [1.37]. The simulation findings suggest that the shared optical cache architecture can improve substantially the Level-1 miss rate, and either speed-up execution by 19.4% or slash the required cache capacity by ~63% on average for the 12 PARSEC benchmarks. As such, the optical cache-enabled CMP architecture can significantly improve system performance, increase memory bandwidth and discard the need for complex coherency protocols, offering also the possibility for using the optical L1 cache as off-die module relieving valuable CPU area.
• **Chapter 5:** Towards transferring the optical memory technology benefits from CMP layouts also in novel all-optical routing table architectures, chapter 5 extends the so far work on optical CAM technology, presented in section 1.5, by proposing the first all-optical Ternary-CAM (T-CAM) cell and its interconnection in an optical T-CAM row architecture, where 4 T-CAM cells and a novel WDM-encoded matchline design can provide comparison operation for a complete 4-bit optical word. The optical T-CAM cell extends the B-CAM cell operation by allowing the storage of a third state “X”, enabling in this way the essential subnet-masked operation needed in modern router applications.

The proposed optical T-CAM cell architecture comprises two optical FFs and an optical XOR gate; the 1\textsuperscript{st} optical FF is used for storing the actual T-CAM cell contents, the 2\textsuperscript{nd} FF for implementing the “X” state support while the XOR gate for enabling the T-CAM cell search capability. The 4-cell T-CAM-row architecture follows a proper wavelength encoding scheme by using an Arrayed Waveguide Grating (AWG)-multiplexer; the multi-wavelength output signal produced at the final row output determines whether a success comparison result is achieved throughout the complete T-CAM row. The performance evaluation of the 4-cell T-CAM row architecture has been carried out using the VPI Photonics simulation suite and by employing experimentally verified SOA-based building blocks. The simulation results demonstrate successful T-CAM row operation at 20 Gbps for both Search and Write functionalities, i.e. more than 2x the speed that can be offered by state-of-the-art electronic CAM look-up table designs [1.87]. The proposed T-CAM row architecture can be easily scaled to form complete optical T-CAM tables required in AL, while the recent experimental developments in high-speed and ultra-low-power integrated photonic crystal InP-on-Si FF devices [1.94] could potentially allow for its experimental implementation in low-footprint and low-energy prototypes.

• **Chapter 6:** This chapter concludes the thesis by providing a summary of the major results and contributions along with some potential lines for future investigation.
1.7 List of References


1.35. S. Bohm and C. Engelmann, "xSim: The extreme-scale simulator," International Conference in High Performance Computing and Simulation (HPCS), pp. 280-286, 4-8 July 2011


1.44. L. Zhao, R. Iyer, S. Makineni, J. Moses, R. Illikkal, and D. Newell, Performance, area and bandwidth implications on large-scale CMP cache design, In Proceedings of the Workshop on Chip Multiprocessor Memory Systems and Interconnects, 2007, Phoenix, AZ, USA


1.46. Die shot of the Sparc T5 processor from Oracle [http://www.theregister.co.uk/2012/09/04/oracle_sparc_t5_processor](http://www.theregister.co.uk/2012/09/04/oracle_sparc_t5_processor)


1.55. A. Veloso et al., “Demonstration of scaled 0.099 µm² FinFET 6T-SRAM cell using full-field EUV lithography for (Sub-) 22nm node single-patterning technology,” in Proc. IEEE International Electron Devices Meeting (IEDM) 2009, 7-9 Dec. 2009, Baltimore, MD, USA


1.64. http://www.ethernetalliance.org/roadmap/


1.69. K. Lee et al., “10Gb/s silicon modulator based on bulk-silicon platform for DRAM optical interface,” Optical Fiber Communication Conference (OFC) 2011, JThA33, 6-10 March 2011, Los Angeles, CA, USA.


1.76. L. Liu et al., “An ultra-small, low-power, all-optical flip-flop memory on a


1.86. G.T. Kanellos, et. al, “WDM-enabled optical RAM architectures for ultra-fast,
low-power optical cache memories,” in Proc. 15th International Conference on Transparent Optical Networks (ICTON) 2013, 23-27 Jun. 2013, Cartagena, Spain


1.89. A.Varga, "The OMNeT++ discrete event simulation system," in Proc. of the European Simulation Multiconference (ESM’), Prague, Czech Republic, 2001


1.91. Top 500 Supercomputers’ list of June 2016 (http://www.top500.org)


Chapter 2

Bringing Optical Boards to HPC-scale environments: The OptoHPC-Sim simulation engine

2.1 Introduction

The increased communication bandwidth demands of HPC-systems calling at the same time for reduced latency and increased power efficiency have designated optical interconnects as the key technology in order to achieve the target of exascale performance. In this realm, technology advances have to be accompanied by corresponding simulation tools that support end-to-end system modeling in order to evaluate the performance benefits offered by optical components at the system-scale environment. This chapter presents the OptoHPC-Sim simulator that supports the utilization of optical interconnect and electro-optical routing technologies at system-scale offering complete end-to-end simulation of HPC-systems and allowing for reliable comparison with existing HPC platforms. OptoHPC-sim’s simulation engine has been developed using the Omnet++ platform and has been designed to offer the optimum balance between the model detail and the simulation execution time. Before proceeding to OptoHPC-Sim presentation, the chapter introduces an application-oriented technology development roadmap along with all the recent advances around the EOPCB technology that have been taken into account during the OptoHPC-Sim development. These advancements include both high-density and multi-layered EOPCBs as well as all the necessary building blocks that are necessary for enabling true optical blade technology, including multi-mode polymer-based single- and dual-layer EOPCBs, a board-compatible optically-interfaced router chip and passive board-level connectors. Taking into account all these technology advancements, a complete optical blade design for use in future HPC systems is presented. This optical blade design is later used for the performance comparison analysis realization between an HPC system that employs state-of-the-art optoelectronic routers and optical interconnects and the Cray XK7 system platform which has been ranked as the 3rd fastest supercomputer in 2016 [2.1].

Chapter 2 is organized as follows: Section 2.2 describes the optical blade design layout
along with all the technological advancements around electro-optical boards that have been taken into account during the OptoHPC-Sim development. Section 2.3 presents the OptoHPC-Sim platform, while Section 2.3 proceeds with a performance evaluation analysis by comparing an HPC network system employing state-of-the-art optoelectronic routers and optical interconnects with a system employing a purely electrical board layout as is being used in Titan CRAY XK7. Section 2.4 summarizes the findings of the performance analysis and concludes the chapter.

2.2 The Optical Board Technology Platform considered by OptoHPC-Sim

The application-oriented technology roadmap used for the OptoHPC-Sim development is illustrated in Fig. 2.1. It presents an example HPC network of 4 racks, as it appears at the GUI interface of the OptoHPC-Sim simulator. The internal rack architecture hierarchy follows the architecture of the Titan CRAY XK7 supercomputer, where 8 CRAY XK7 Blades are grouped together forming a chassis and three chassis are

Fig. 2.1: OptoHPC-Sim's main GUI frame demonstrating an example HPC model incorporating 4 racks. Each rack consists of 24 OPCBs being grouped in 3 chassis of 8 OPCBs each. An EOPCB design with 2 router interfaces is demonstrated at the bottom.
grouped together forming an HPC rack. At the right of Fig. 2.1, a cluster of 8 OPCBs forming a chassis is highlighted and illustrated as inset in more detail. Each OPCB includes proper sockets for hosting 4 transceiver optochips and 2 optoelectronic router chips along with the proper pin connections between them. The OPCB’s optical bandwidth allocation has been realized using the Automatic Topology Design Tool (ATDT) [2.2], whose role is to provide the optimum OPCB bandwidth allocation for a given layout strategy. Transceiver optochips serve as the interface between the CPU chips and the board-level optical waveguides, while the optoelectronic router chips connect the CPU chips together as well as with the outer world off-board devices. The inset at the bottom of Fig. 2.1 presents the EOPCB prototype design that is currently being fabricated by TTM Technologies [2.3] in order to validate the basic blade functionality that is necessary for the operation of an optical blade-based HPC network. This EOPCB prototype is capable of hosting two Compass EOS optoelectronic router chip modules [2.4] that allow for both chip-to-chip as well as off-board communication by optical means. The critical technology blocks required for enabling the correct operation of this EOPCB prototype include a) the EOPCB, b) the board-adaptable electro-optic router ASIC together with the appropriate chip-to-board interfaces, and c) the board-level connectors and coupling interfaces. The following sections describe in more detail the progress that has been carried out so far regarding all these individual technology blocks towards realizing an optical blade capable to serve the needs of an optical blade-based HPC network as the one shown in Fig. 2.1.

2.2.1 High end routing platform using optical interconnects

This section briefly reviews the optoelectronic router chip developed by Compass Networks [2.4]. Its board-adaptable version has been considered for serving as the on-board routing machine within the frame of OptoHPC-Sim simulator. The router ASIC developed by Compass Networks constitutes a traffic manager that operates as a queuing machine, in which the conventional electrical serial I/O have been replaced with a pure optical I/O interface. In CMOS design, high frequency chip I/O ports can be located only on the perimeter of the package since an areal array of I/Os is more form factor efficient than a linear array at the same linear pitch. More details on optical and electrical escape bandwidth can be found in [2.5]. The I/O ports number is thus limited by the package size and the ability to connect these serial I/Os to the metal traces on the PCB. An optical chip interface allows overcoming this radix limitation via the
assembly of dense, two-dimensional arrays of lasers and photodiodes directly on the ASIC. The first operating demonstration presented by Compass Networks incorporated a matching 2D fiber array that is assembled above the optoelectronic matrices allowing for direct coupling of light to and from the fibers. This technology allows for very high data densities and increases the bandwidth-distance product of the device while minimizing at the same time the power required for chip I/O.

Using this approach, several line cards can be connected directly by using fibers, thereby eliminating the need for a backplane in such scenarios. In a typical router application, traffic from the packet processing unit is routed to the traffic manager ASIC with its on-chip parallel optical interconnect connected via parallel fiber arrays to several traffic managers on different line cards and racks. According to Compass Networks, the high bandwidth available from the large optical matrices allows building large, non-blocking Clos networks since the router nodes have enough optical bandwidth to connect directly without the need for an intermediate switch fabric. This approach has already been followed in [2.4], where various line cards are connected with fiber optics thereby eliminating the package constraints and greatly simplifying the line card architecture: the traffic from the packet processing unit is routed to a traffic manager/queuing machine ASIC with an on-chip parallel optical interconnect which is linked via parallel fiber arrays to several traffic managers on different line cards with minimal queuing constraints.

A cross-sectional view of the parallel optical interconnect assembled on the traffic manager chip is shown in Fig. 2.2. This is a mixed signal chip with both digital and analog functionalities. Two dimensional matrices of Vertical-Cavity Surface-Emitting Lasers (VCSELs) and Photodiodes (PDs) are directly attached to their analog circuits in the chip. Each VCSEL is located directly above a Tx cell containing the laser driver and serializer. Similarly, each PD is located above an Rx cell containing the TIA, limiting amplifier, equalizer, de-serializer and clock data recovery circuit (CDR). This is a localized design with each optoelectronic VCSEL/PD module being electrically isolated from all other VCSEL/PD modules. As has been reported by the company, the transmission length from the analog circuit to each VCSEL/PD module is in the 100µm range thereby minimizing the effect of parasitics on the link.
This design consumes significantly lower power with respect to conventional chip I/O. In standard design, most of the power dissipation is in the serializer/deserializer (SerDes) arrays. However, unlike the typical chip package case where the SerDes drives a lossy (~25dB) copper trace, in this case it drives a low loss (<0.2dB) fiber optic link. This allows packaging all of the analog components densely in a 250×250µm area. With an 8Gb/s clock, the power for an optical link is 10pJ/bit, including the SerDes arrays. For a 168 element device like the optoelectronic router developed by Compass Networks, the total power consumption of the optical interconnect and its analog circuits is about 12W—significantly lower than any comparable copper I/O technology, offering at the same time 1.34Tb/s full duplex bandwidth. According to the designers of the optoelectronic router, future generation of this technology based on 16nm CMOS will have a power efficiency of 3-4 pJ/bit including the SerDes arrays [2.6].

Optical coupling of the on-chip interconnect can be carried either to a 2D fiber optic array mounted on the PCB as can be seen in Fig. 2.3 (B), or to an embedded waveguide array as can been seen in Fig. 2.3 (C). In either case, a 2-lens relay is used for light coupling. A cutout hole is required thus also in the PCB. In the fiber coupling case, the Tx and Rx fiber arrays are mounted vertically above the VCSEL and PD matrices. The fiber bundles are glued to the PCB surface and this solution has passed extensive reliability tests and is already being applied in real field applications of the optically-interconnected router chip. In the case of waveguide coupling solution, which is still in the development phase, a proper angle prism has to be used with the microlens array to facilitate in-plane light coupling. This approach can also be used for fiber coupling in future generations.
Chapter 2

The high bandwidth of the optical interconnect is obtained by using large matrices in the transceiver. The device has 168 optical channels in a 12×14 layout and uses 8Gb/s optoelectronic chips, leading to an aggregate bandwidth of 1.34Tb/s with a data density of 64Gb/s/mm². This chip is currently in the process of serving as the board-adaptable router chip in the Optical Blade Design presented in the next sections; however the recent progress towards 336-element optical I/O matrix size [2.7] raises expectation for future on-board router chips with record high aggregate capacity values. Fig. 2.4 shows the eye diagrams from a 168 element VCSEL matrix performing at 8 Gb/s line-rates and producing a $2^{31}-1$ Pseudorandom Binary Sequence (PRBS) test pattern [2.4]. All 168 eyes exhibit BER < $10^{-12}$ at the center of the eye and are clearly open with an extinction ratio of about 5dB and high Signal-to-Noise Ratio (SNR) values. The optical crosstalk between neighboring cells was in the -32dB range indicating good optical and electrical isolation in the matrix. Sensitivity measurements with a 2m, 200m and 300m multimode OM3 fiber reported a sensitivity level of about -10dBm at a BER of $10^{-12}$ for all fiber.

Fig. 2.3: Optical coupling of the on-chip optical interconnect (A) with a 2D fiber bundle assembled on the PCB (B); coupling to a double-layered embedded waveguide array in a PCB using microlens and prisms (C). [2.7]

Fig. 2.4: PRBS31 eye diagrams from a 12×14 VCSEL matrix at 8Gb/s line-rate. [2.4]
length. This is indicative that the receiver Clock and Data Recovery (CDR) design is sensitive enough to overcome the distortions resulting from modal dispersion. With an average VCSEL power of \( \sim 2 \text{dBm} \), this result indicates a dynamic range of about 10dB [2.4].

2.2.2 Multi-mode Electro-Optical PCB Technology

Fig. 2.5 depicts the mask layout for an EOPCB prototype that can host two optoelectronic router chips (presented in the previous section) and follows the EOPCB design illustrated in the inset of Fig. 2.1. This EOPCB prototype constitutes one of the base technologies incorporated in OptoHPC-Sim simulator. This prototype layout is currently being fabricated by TTM Technologies [2.3] and aims at all-optical chip-to-chip connectivity using multimode polymeric waveguide arrays embedded in conventional multilayer PCB card with up to 16 electrical layers. The two optoelectronic chips are located at a distance of 15cm and have their optical I/O matrix facing the PCB, so that the VCSEL transmitter matrix of the first chip can connect to the PD receiver matrix of the second chip via a 14-element multimode polymer waveguide array. The waveguide loop has been implemented in the EOPCB prototype in order to allow for a step-wise experimental testing approach for the chip-to-chip connectivity. This loop allows the transmission of data from a router chip back to the same chip, so that the experimental testing can be carried out even when assembling the first chip prior having both router chips deployed on-board.

Waveguides were designed to have rectangular core cross-section with nominal size of 50 \( \times \) 50 \( \mu \text{m} \), and channel-to-channel spacing of 250\( \mu \text{m} \). Dow Corning liquid silicone materials were used to fabricate the waveguides. Core material (WG-1010) with refractive
index $n=1.519$ were surrounded by lower index cladding material (OE-4141) with $n=1.501$ to result $\text{NA}=0.24$ channels. Clad layer thicknesses were 50µm and 25µm for bottom and top cladding, respectively.

Fig. 2.6 (a) presents a schematic of chip-to-waveguide coupling concept. The beam folding element to couple light from/to TX/RX to waveguide comprises of beam splitter with microlens array (MLA) as it is shown in Fig. 2.6 (b) and glued on one of the flat surfaces. After EOPCB fabrication, the folding element is assembled into cut out hole in the PCB (size 5.13×11.5mm) as it is shown in Fig. 2.6 (c).

Previous OPCB link demonstrators show optical waveguides as separate entity from PCB [2.8], as layer built on board surface [2.9]-[2.12] or embedded part of PCBs with varying complexity and a limited number of up to 4 electrical layers [2.13]-[2.18]. In this attempt by TTM Technologies, the optical waveguide array has been embedded for the first time inside a high layer count PCB product with 16 to 20 copper layers including one or two optical layers in the stack for compliance with product form factors, constructions and board materials [2.14]. The board contains all required electrical layers
Bringing Optical Boards to HPC-scale environments: The OptoHPC-Sim simulation engine

and via structures (Plated-Through Holes (PTHs), n-PTH, stacked and buried microvias) built around optical cores, following certain process and design strategies during the development for: (a) rerouting of all signals to avoid areas with optical waveguides, (b) managing processing of sub-cores with different copper thickness (17µm for signal (S), 35µm for power (P) and 70µm for ground (G) layers), (c) providing three microvia layers as part of the EOPCB, (d) controlling registration and material movement during lamination of dissimilar materials and (e) providing a process flow with minimal thermal load to waveguides.

Besides chip-to-chip connectivity via embedded polymer waveguides, the EOPCB prototype hosts two mid-board Multifiber Termination Push-On (MTP) sites for fiber-to-waveguide connections. These MTP sites provide out-of-plane waveguide connection with embedded micro-mirrors, which were embedded directly into the waveguide substrate as part of the PCB fabrication progress and connected to lensed MT ferrules assembled in a slot perpendicular to the mirrors. Except from the two chip-to-board interfaces presented in Fig. 2.5, two mid-board MTP fiber-to-WG test connectors can be seen at the right side of the board. In addition, Fig. 2.7 shows an overview of the fabricated board as well as a cross-section across the stack detailing the electrical layers and the embedded waveguides. The outline of the fabricated board has dimensions of 190 mm x 420 mm and comprises 16 electrical layers for signal and power line interconnects and one optical waveguide layer stacked between copper layers 8 (L8) and 9 (L9). In more detail, the construction of optical/electrical build is 8 electrical + 1...
Optical + 8 electrical and the optical loop consists of 14 parallel waveguides.

Focusing now to 16”x20” standard production panels and taking advantage of the established fabrication processes [2.19], TTM Technologies reported for the first time that the developed process has been scaled up to support EOPCBs with two optical layers. Fig. 2.3 (C) presents the targeted coupling scheme that is based on microlenses and prisms, which is in principle transferrable also to the fiber-connector scheme and could allow for reducing the form factor of the 2D fiber bundle approach that was presented in Fig. 2.3 (B). However, using this coupling scheme in the dual-layer embedded waveguide board layout, as shown in Fig. 2.3(c), would allow the use of only a small percentage of the actual optoelectronic router chip interconnect-size developed by Compass Technologies, which has a 12×14 layout. Especially, this layout can allow for the utilization of the two outer rows of the router’s 12 x 14 I/O optical matrix where the first outer-row 48 peripheral I/O pins connect to the first PCB waveguide layer and the second-periphery row 40 pins connect to the second waveguide layer, so that finally only 88 out of the 168 optical pins can be accommodated. In order to fully exploit the whole 12x14 optical I/O matrix of the router at on-board setups, the electro-optical PCB should be replaced by FlexPlane technology (which is later presented in section 2.2.4), since multi-layer EOPCB deployments are still facing severe difficulties towards accommodating more than 2 waveguide layers.

2.2.3 Passive Optical Connector and Polymer Coupling Interfaces

In order to fully utilize the number of available channels and exploit the off-board interconnect capabilities of the optoelectronic routing chips presented in section 2.2.1, appropriate passive coupling interfaces and pluggable connectors are needed for the interconnection part. MT ferrules are by far the most common parallel optical connector interface available. Passive parallel optical interfaces based on the MT standard can accommodate up to 6 rows of 12 optical channels per connector ferrule, whereby adjacent channels will have a center-to-center separation of 0.25 mm. MT ferrules are designed to house arrays of multimode or single mode optical fibers. In order to ensure that each connecting fiber pair in the connecting ferrules can make full physical contact with each other even when the connecting MT facets are not completely parallel, the fibers are arranged to protrude slightly out of the MT ferrule facet. The MT ferrule of Fig. 2.8 has been manufactured by USConec [2.20] and the MLAs which can be connected onto the MT ferrule interface have been manufactured by the Japanese
company Enplas [2.21]. The MT ferrules have been manufactured using a durable thermoplastic composite, Poly-phenylene Sulfide (PPS). The termination process involves multiple parallel fibers being aligned and secured to the ferrules with an optical connector grade thermal cure epoxy and polished with a variety of commercially available batch connector polishing machines in order to achieve the interface requirements set out in the IEC standard 61754-5. The fibers are arranged and polished such that the tip of each fiber should protrude by between 0.5 µm and 2.5 µm from the surface of the MT ferrule facet, depending on the quality required. A new generation of parallel optical connector was developed by USConec in 2013 in collaboration with Intel and Facebook as part of the Open Compute project [2.22] in order to address the problem of scaling such connectors into future mega Data Centers. The expanded beam PrizmMT™ ferrules incorporate microlens arrays into the fiber holding structure to ensure that, at the exposed connecting interfaces, the optical beam width is actually increased to about 3.5 times the size of the multimode fiber aperture, thus making it far less susceptible to contamination. The MXC connector, which formed a key part of the publicity drive surrounding the Open Compute project houses a Prizm MT ferrule in a plastic shell and
clip and is designed for host side access.

Moving to polymer coupling interfaces, a suite of receptacles to allow coupling of MT fiber interfaces to PCB embedded multimode polymer waveguides has been developed. Fig. 2.8 (a) shows two waveguide coupling interfaces on an electro-optical PCB with embedded multimode polymer waveguides. One type of receptacle allows in-plane fiber-to-waveguide coupling, whereby the optical axis of the connecting fiber will be co-linear with the axis of the embedded waveguide. The other receptacle types allow out-of-plane fiber-to-waveguide coupling, whereby the axis of the connecting fiber will be orthogonal to the waveguide axis. The receptacle of Fig. 2.8 (b) includes a discrete micro-mirror system. This will allow MT ferrule-based connectors to plug to the top of the PCB and launch or receive light to and from the embedded waveguides. The receptacles are passively aligned and attached to the polymer waveguide interface using a proprietary assembly method which is shown in Fig. 2.8 (c). Fig. 2.8 (d) shows a test board with generic waveguide coupling interfaces, designed to accommodate either in-plane or out-of-plane receptacles. An MTP fiber optic cable is attached to an out-of-plane receptacle and illuminates an embedded multimode polymer waveguide with visible 650 nm light.

2.2.4 Fiber and Polymer Waveguide Flexplane Technologies

Following a realistic scenario that combines a dual-layer embedded polymer waveguide PCB with the Compass EOS router chip, only the two outer rows of the router’s 12x14 I/O optical matrix can be used. In this arrangement, the first outer-row 48 peripheral I/O pins connect to the first PCB waveguide layer and the second-periphery row 40 pins connect to the second waveguide layer. In order to fully exploit the whole 12x14 optical I/O matrix of the router without migrating to still immature deployments of multi-layer OPCBs with more than 2 waveguide layers, the electro-optical PCB should be replaced by flexplane technology. Fiber flexplanes are laminated fiber-optic circuits, in which optical fibers are pressed and glued into place on a substrate. These structures benefit from the reliability of conventional optical fiber technology and are currently promoted as the mature and low-cost alternative to EOPCBs, which have still to reach certain reliability and cost-efficiency metrics prior entering the market. However, unlike embedded optical waveguides, these circuits cannot accommodate waveguide crossings in the same layer i.e. fibers must cross over each other and cannot cross through each other. Moreover, each additional fiber layer necessitates typically the addition of backing substrates in order to hold the fibers in place, thus significantly increasing the thickness
Bringing Optical Boards to HPC-scale environments: The OptoHPC-Sim simulation engine

Fig. 2.9: Optical fiber flexplanes deployed for an optically enabled data storage and switch test platform for data centers [2.23]: a) Photo of electro-optical midplane with MT terminated flexplane, b) Schematic view of Prizm MT terminated flexplane of the circuit. As such, flexplanes can be attached at best as separate entities onto the surface of a conventional PCB.

Fig. 2.9 (a) shows a 196 fiber flexplane with MT ferrule terminations in an optically enabled data storage and switch test platform for data centers [2.23]. The average insertion loss of the flexplane alone is ~0.32 dB and it has been measured using an 850 nm VCSEL source from an Intel XFP transmitter. Fig. 2.9 (b) depicts the design of a more complex 196 fiber flexplane with Prizm MT terminations, which is more suitable for forced air environments in Data Centers.

2.3 The OptoHPC-Sim Simulation Engine

The deployment of on-board optical technologies even with brilliant physical layer performance characteristics cannot ensure on its own an excellent performance at HPC-scale environments. Since both network topology and bandwidth allocation between the
nodes in an HPC comprise significant performance factors on top of the underlying technologies, proper simulation tools are needed in order to deal with the performance evaluation of future HPC network architectures that make use of optical technologies. However state-of-the-art sophisticated HPC simulators still don’t support the use of advanced electro-optic router and interconnect solutions at board-level. Among the few HPC open-source simulators that are free of charge and available to the research community, none of them is focused on or can even efficiently explore the adoption of optical technology advancements in the HPC field. The Extreme-scale Simulator (xSim) [2.24] implements a parallel discrete event HPC simulator but is mainly targeting the investigation of parallel applications’ performance at extreme-scale MPI environments. SST+gem5 [2.25] is a scalable simulation infrastructure for HPCs and comes as the result of the integration of the highly detailed gem5 performance simulator into the parallel Structural Simulation Toolkit (SST). SST is a system of disparate hardware simulation component entities integrated via a simulator core, which provides essential services for interfacing, executing, synchronizing and monitoring the various components with gem5 [2.26] being integrated as one of them. However, gem5 gives emphasis in simulating detailed CPU-cores and computer memory hierarchies, yielding high simulation times due to its highly-detailed CMP hardware models.

At this point, OptoHPC-Sim simulation platform emerges as a new simulation engine targeted towards supporting the use of electro-optical boards and routing technologies in complete and fully operational HPC network architectures. OptoHPC-Sim can evaluate throughput and latency of complete HPC networks based on EOPCBs and for a wide range of traffic profiles typically used for benchmarking in HPCs. OptoHPC-Sim can yield valuable feedback on the technology development towards conforming to application-driven performance requirements, facilitating critical decisions such as the number of optical links finally required and the number of optoelectronic chips that need to be hosted on an EOPCB. OptoHPC-Sim forms a powerful, modular and light-weight solution being implemented on top of the Omnet++ discrete event simulation framework [2.27]. It relies on a careful balance between the model detail and the simulation execution time, employing a queue-based HPC model and including only the absolutely necessary details for reliably evaluating an optically enabled HPC system. OptoHPC-Sim offers a user-friendly GUI that allows the detailed exploration of complete HPC topologies and can successfully be used for both demonstration and education purposes.
OptoHPC-Sim’s GUI is presented in Fig. 2.1, where an example HPC network of 4 racks along with the internal rack architecture hierarchy is demonstrated. The same rack architecture is also employed in Titan CRAY XK7 supercomputer [2.28], where 8 CRAY XK7 Blades are grouped together forming a chassis and three chassis are grouped together forming an HPC rack. Depending on the size of network determined as the number of computing nodes, the number of racks may vary between 1-3 racks up to 49-320 racks. Building for example a class0 network of 96-288 computing nodes would require 1-3 racks organized in a single rack-row, while a class2 network of 1632-4608 nodes would require 17-48 racks organized in two rack-rows [2.29].

At the top of OptoHPC-Sim’s GUI in Fig.1, the main menu’s toolbar allows the management of the simulation process providing the options for a step-by-step, fast and express simulation mode. Along with the main menu’s simulation toolbar, kernel-statistics are reported including the simulation clock-time and the number of scheduled/executed events. At the left side of OptoHPC-Sim’s GUI, the parameters explorer allows the exploration of the configurations regarding the current simulation setup. At the bottom of GUI, the event-list section informs the user for the executed events. Last but not least, the network explorer appears in the middle of GUI allowing the top-down exploration of the simulation model hierarchy by double-clicking to the individual modules.

OptoHPC-Sim currently supports both Mesh and Torus network topologies in up to 3-dimensional arrangements, as being widely used in many of the industry’s HPC systems [2.28]. Fig. 2.10 presents an example topology of a single rack 3D torus architecture where a total number of 24 PCBs are organized in groups of 8 PCBs, where each of the groups forms a chassis. Using the OptoHPC-Sim’s GUI and moving down through the HPC hierarchy concludes to the PCB-layer view demonstrated in Fig. 2.11. In this example, two router modules are connected together using a specifically configured instance of the link module, with each router being directly connected to two node modules by using again a specifically configured instance of the same link module.

Router model represents the router chips used in the HPC network and is responsible for all the routing decisions which are taken on a hop-by-hop basis. Router model comes with support for Dimension Order Routing (DOR) and Minimal Oblivious Valiant Routing (MOVR) algorithms that ensure deadlock free operation by eliminating any cyclic dependencies that could arise through the individual routing decisions [2.30]. During the OptoHPC-Sim’s initialization stage, the router model is responsible for
generating the routing-table structures that are necessary for taking the routing decisions. Routing tables are organized in rows where the number of rows is equal to the total number of routers in the network minus one since traffic should never be routed to the source router again. Each routing table row is organized in two columns, where the first
column contains a unique router address and the second column contains a set of one or more possible output gates that should be followed in order to route any data destined to the router of the first column. The routing table generation is based on the Dijkstra’s shortest paths algorithm ensuring minimal routing operation for both DOR and MOVR routing algorithms [2.30].

Router model comes with a set of three predefined configurations, where all the router network-level characteristics have been taken into account, such as the input and output port organization as well as their specific bandwidth specifications. The first configuration has been derived by considering the Gemini router’s specifications, which is currently used in Titan Cray’s XK7 blades. The other two configurations have been derived by considering the specifications of the Optoelectronic Router presented in section 2.2.1. Regarding the first OE Router configuration, named OE-Router-88ch, a total number of 88 bi-directional I/O links is considered, where every link operates at 8Gbps. In this case, a realistic scenario is followed where only the two outer rows of the router’s 12x14 I/O optical matrix are used over a dual-layer embedded polymer waveguide PCB. In this arrangement, the first outer-row 48 peripheral I/O pins connect to the first PCB waveguide layer and the second-periphery row 40 pins connect to the second waveguide layer. In order to fully exploit the whole 12x14 optical I/O matrix of the router without migrating to still immature deployments of multi-layer OPCBs with more than 2 waveguide layers, the case where all 168 optical I/Os are utilized is also considered, by using a fiber-optic Flexplane technology (presented in section 2.2.4) for realizing the on-board interconnections. This OE Router configuration is named OE-Router-168ch.

Router model incorporates also the buffer, ResourcesManager and switchFabric models that are necessary for the internal router organization but are not depicted in Fig. 2.11. Buffer model implements a basic First-In-First-Out (FIFO) policy and supports Virtual Channel (VC) organization, which ensures deadlock-free operation with regard to the wrap-around links existing in Torus networks. VC organization is also essential for MOVR routing algorithm in order to eliminate any cyclic dependences arising by the individual routing decisions [2.30]. The Buffer model can be used for modeling either an input- or an input-output-buffer router organization. ResourcesManager implements a FIFO arbitration policy with respect to the router’s input buffers, while at the same time orchestrates the output ports resource allocation. ResourcesManager module is also
responsible for driving the *switchFabric* module that forwards the input buffers transmitted data to the proper output ports.

Link model incorporates all the physical-layer relevant parameters, such as the link bandwidth, link length/propagation delay and Bit-Error-Rate (BER). The link module is utilized in all HPC network connections and not only at on-board level, as shown in the example of Fig. 2.11, using the corresponding parameters for every hierarchy level.

Node model simulates the HPC’s computing nodes and is responsible for the per node traffic generation according to the applications running on the HPC and described later along with *trafficPatternsManager*. Node also sinks any incoming data updating at the same time the per node simulation statistics (global statistics management described later along with *statisticsManager*). Node model incorporates both the buffer and *trafficGenerator* models that are necessary for the internal node organization.

Buffer model is the same with the one incorporated in the router model, where in the case of node it is capable of simulating an infinite depth queue which separates the packet source (*trafficGenerator*) from the simulated network. It is important to note here that the traffic injection process is operated in lock-step with the rest of the network simulation, achieving in this way a bounded memory footprint even for network saturation conditions \[2.30\].

*TrafficGenerator* manages the actual traffic generation by generating and forwarding proper messages to the node’s infinite buffer. Due to the fact that messages may be arbitrarily long, they are further divided into one or more packets that have a predefined maximum length. Each packet carries a segment of the message’s payload and a packet header is always preceding. Considering the SF flow control mechanism, both the header and payload data are packed together into a single group of bits and are transmitted to node’s buffer. When the VCT flow control mechanism is followed, the packet payload is further divided into zero or more body flits that are followed by a tail flit. In this case all the header, body flits and tail flit are individually transmitted to the node’s buffer and subsequently to the entire network.

Three more auxiliary modules, namely *networkAddressesManager*, *trafficPatternsManager* and *statisticsManager*, have been incorporated to support the successful network initialization setup and the correct simulation operation process. All these three modules can be seen in the OptoHPC-Sim’s GUI network explorer of Fig. 2.1 and are accessible directly below the four racks of the HPC network example.
NetworkAddressesManager is responsible for the network’s addresses allocation along both the computing nodes and the routers. Two automatic address allocation schemes are supported with the first one following a sequential address allocation policy like in the case of Titan CRAY XK7 [2.28] and the second one following a random-uniform address allocation policy. If desired, custom address-allocation schemes can be fed to the simulator in the form of input text files. For all the cases each node is assigned both a decimal address and a location identifier in the form of X.Y.Z coordinates with regard to its absolute position in the Torus/Mesh grid. Taking as an example the second node of Fig. 2.11, its decimal address equals to 4 where its location identifier equals to 0.1.0. All addresses are unique and start counting from zero up to the number of nodes minus one. The same address allocation scheme is also applied to the router nodes. Finally, networkAddressesManager is responsible for defining the dateline routers, which are essential for ensuring deadlock free operation in the Torus topologies [2.30]. Considering the example of Fig. 2.11, the first router serves as dateline in all three X, Y and Z dimensions, while the second router serves as dateline only in X and Z dimensions.

TrafficPatternsManager's main responsibility is to define and manage the applications executing in the simulated system by means of traffic pattern distributions. OptoHPC-Sim currently supports 8 most well-known synthetic traffic patterns in the literature [2.30]: 1) Random Uniform, 2) Bit Complement, 3) Bit Reverse, 4) Bit Rotation, 5) Shuffle, 6) Transpose, 7) Tornado, and 8) Nearest Neighbor. Two more configuration options are additionally offered, where the simulator can be fed with either real-world packet traces or files describing the traffic pattern distribution among the computing nodes. On top of that, the user can choose between constant and exponential message inter-arrival times as well as constant and variable message size distributions.

StatisticsManager's role is to handle the global result collection during the simulation process and to record the results into proper output files when the simulation comes to an end. One of its most significant features is that it can detect whether a steady state has been reached through continuously monitoring the global network’s performance metrics, informing the simulation kernel via a special termination signal that denotes that a steady state has been reached.

OptoHPC-Sim’s configuration procedure can be easily handled by only a single configuration file, which specifies the network configuration parameters that must be taken into account.
2.4 EOPCB-Based HPC Network Performance Analysis and Comparison with CRAY XK7 HPC Network Configuration

In this section OptoHPC-Sim is used in order to evaluate and compare the performance of an HPC network that employs three different types of on-board routing: a) the OE-Router-88ch, b) the OE-Router-168ch and c) a Conventional Router model that complies with the Gemini router’s specifications along with a purely electrical board layout, as is being used in CRAY XK7 HPC which is the world’s 3rd fastest supercomputer [2.1]. For the OE-Router models, router channel allocation has been realized in both OE-Router-88ch and OE-Router-168ch cases with the Automatic Topology Design Tool (ATDT) in order to offer optimum saturation throughput for the case of Uniform Random traffic pattern when considering optimal routing conditions. ATDT has been developed by University of Patras [2.2] and is targeted to the generation of the optimal on-board OPCB topology and bandwidth allocation within a specific set of topology families like meshes, tori and fully connected networks. ATDT (a) satisfies any given physical- and packaging-related parameters as well as performance requirements and (b) it also takes into account that OPCBs are parts of a larger system/topology. Since only Uniform Random traffic pattern is considered by ATDT, it cannot provide any information about the network performance when different traffic profiles are employed which indicates the need for use of HPC network simulation engines like OptoHPC-Sim.

In this performance analysis and in order to allow for a direct comparison between an HPC network architecture relying on optical blade technology with the CRAY XK7 blade employed in the 3rd world’s fastest supercomputer, both topology type and size and for both the whole network and for the individual boards were kept constant and equivalent to CRAY XK7 systems. Fig. 2.12 (a) shows the detailed layout of a single CRAY XK7 PCB while Fig. 2.12 (b) shows the detailed layout of a double layered OPCB that corresponds to the OE-Router-88ch configuration. As can be seen, the EOPCB includes 4 sockets for hosting the transceiver optochips and 2 optoelectronic router chips along with the proper pin connections between them. Transceiver optochips serve as the interface between the CPU traffic generating modules, called computing nodes, and the board-level optical waveguides, while the optoelectronic router chip version is here shown to support 88 out of 168 multi-mode optical I/Os as it has been described in detail in sections 2.2.2 and 2.3.
Bringing Optical Boards to HPC-scale environments: The OptoHPC-Sim simulation engine

Table 2.1 summarizes the I/O link capacities per dimension for the 2 OE-Router and the Conventional Router configurations, as well as the maximum router capacity for all the three cases. In this performance analysis, a 4x12x8 3D Torus HPC network is assumed which can be classified as a class1 network incorporating a total number of 384 computing nodes [2.29]. The computing nodes are organized in a single rack-row of 4 racks, where each rack incorporates 3 chassis of 8 PCB Blades. Each PCB Blade incorporates 2 directly connected router modules, where each router module is directly connected to 2 computing nodes. A sequential address allocation policy is followed and all the eight synthetic traffic patterns presented in section 2.3 are used. Regarding the Conventional Router configuration, the VCT flow control mechanism has been utilized complying with the respective mechanism of the Gemini router that is used in the Titan CRAY XK7 supercomputer [2.28]. In both cases of OE-Router-configurations, both SF and VCT flow control methods have been evaluated. The rest of the simulation parameters employed is being summarized in Table 2.2.

Before proceeding with the comparison analysis between the OE Router and Conventional Router model-based HPC systems, the Conventional Router model-based

Table 2.1: Router Configurations’ I/O Capacities

<table>
<thead>
<tr>
<th>Router Port Type</th>
<th>Conventional Router</th>
<th>OE-Router-88ch</th>
<th>OE-Router-168ch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node-Router (Gbps)</td>
<td>83.2</td>
<td>64</td>
<td>120</td>
</tr>
<tr>
<td>X dimension* (Gbps)</td>
<td>75</td>
<td>64</td>
<td>120</td>
</tr>
<tr>
<td>Y dimension* (Gbps)</td>
<td>75 (Mezzanine)</td>
<td>37.5 (Cable)</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>37.5 (Cable)</td>
<td>96 (Mezzanine)</td>
<td>192</td>
</tr>
<tr>
<td>Z dimension* (Gbps)</td>
<td>120 (Backplane)</td>
<td>75 (Cable)</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>120 (Cable)</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>Max Capacity (Tbps)</td>
<td>0.706</td>
<td>0.704</td>
<td>1.344</td>
</tr>
</tbody>
</table>

*per direction
network is being evaluated for both MOVIR and the DOR routing algorithms so as to identify the more suitable routing paradigm. Fig. 2.13 illustrates the respective simulation results. Fig. 2.13(a) shows the mean node throughput versus the mean node offered load for both Random Uniform and Nearest Neighbor traffic patterns, while Fig. 2.13 (b) depicts the respective mean message delay versus mean node offered load considering all the messages exchanged among the computing nodes. For both traffic patterns, mean node throughput increases proportionally to the offered load until it reaches the corresponding saturation point. Beyond saturation point and in both Nearest Neighbor DOR and MOVIR cases, a constant network throughput is observed where the DOR algorithm offers ~16% better performance compared to the MOVIR. Regarding the Uniform Random pattern, DOR algorithm offers ~13% better performance around the saturation point, with further load increases leading to sharp throughput decreases for DOR resulting in ~12% better performance for MOVIR. For all cases, mean message delay comes in agreement with the respective mean node throughput observations yielding a constant mean message delay increase until becoming unbounded at the saturation point. Moreover, a slightly better performance is obtained for the case of Nearest Neighbor DOR,

### Table 2.2: Simulation Parameters

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Size</td>
<td>4 x 12 x 8</td>
</tr>
<tr>
<td>Message generation distribution</td>
<td>Exponential</td>
</tr>
<tr>
<td>Header Size (Bytes)</td>
<td>64</td>
</tr>
<tr>
<td>Packet Size* (Bytes)</td>
<td>1536</td>
</tr>
<tr>
<td>Router Buffer Size (KBytes)</td>
<td>250</td>
</tr>
</tbody>
</table>

*Message Size was set equal to Packet Size*
being in agreement with the mean node throughput observations. As such, DOR algorithm is being chosen to be employed as the routing algorithm for the rest of the performance analysis since it is shown to outperform MOVR algorithm in the Conventional Router-based network topology and for both Uniform Random and Nearest Neighbor synthetic traffic patterns.

Fig. 2.14 and Fig. 2.15 illustrate the simulation comparison results among all the three OE-Router-88ch, OE-Router-168ch and Conventional Router (termed as CRAY in the figure) cases and for all the 8 synthetic profiles supported by OptoHPC-sim. Fig. 2.14 presents the mean node throughput versus mean node offered load while Fig. 2.15 presents the respective mean message delay versus mean node offered load considering all the messages exchanged among the computing nodes. As expected, for all throughput measurements and for both OE-Router cases, no variations between the SF and VCT flow control methods are observed irrespective of the traffic pattern applied.

Fig. 2.14 reveals that the use of Uniform Random pattern leads to the highest saturation throughput among all 8 traffic patterns for both OE-Router cases. This comes in agreement with the channel allocation and design strategy that were followed by the ATDT tool towards ensuring maximum throughput for Uniform Random patterns. However, given that ATDT considers optimal routing conditions that are certainly not met by realistic routing algorithm implementations like DOR, both OE-Router-based cases saturate below the 100% offered load that should be theoretically expected. On the other hand, although the maximum capacity of the Conventional Router is slightly higher compared to the OE-Router-88ch, the Conventional Router CRAY system throughput saturates much earlier at ~14.5 Gb/s, being ~3.3 times lower compared to the 48 Gb/s saturation point of the OE-Router-88ch. This particular observation reveals the important role of total router’s bandwidth channel allocation strategy. The throughput performance is significantly improved in the case of the OE-Router-168ch compared to the OE-Router-88ch due to the 1.9x higher router capacity offered in this case.

Beyond the corresponding saturation points, a slight throughput drop for all the three router configurations is observed. This behavior stems from the channel arbitration unfairness introduced by the network routers with respect to the individual packet flows of Uniform Random pattern. OptoHPC-Sim employs a per router First-In-First-Out (FIFO) arbitration policy with respect to the desired output router port. Packets are grouped together according to the desired output port and are prioritized according to the absolute arrival time at the input ports of each individual router. This would
Fig. 2.14: Throughput simulation results for 8 synthetic traffic profiles
Bringing Optical Boards to HPC-scale environments: The OptoHPC-Sim simulation engine

Fig. 2.15: Mean packet delay simulation results for 8 synthetic traffic profiles
eventually allow packets that require fewer hops and therefore fewer resource arbitrations to get a higher proportion of the available bandwidth, since no global routing criteria are taken into consideration. Hence, some flows may become starved and their throughput can drop dramatically as the load increases beyond saturation. Solutions like (a) the adoption of age-based arbitration criteria (e.g. # of hops) or (b) the implementation of non-interfering networks with one virtual channel per destination (unrealistic for big networks) are well-known in the literature for offering network stabilization beyond saturation point [2.30]. However, the implementation and analysis of such advanced solutions falls out of the scope of this analysis.

Proceeding to the remaining traffic patterns shown in Fig. 2.14, mean node throughput increases proportionally to the offered load until reaching the corresponding saturation points for all three router configurations, similarly to the case of Uniform Random. In the cases of Tornado’s CRAY and OE-Router-88ch and of Bit Complement’s CRAY, saturation throughput is reached even from the first measurement at an offered load of 10 Gb/s. The OE-Router-88ch configuration outperforms the CRAY system for all traffic patterns, with the only exception being in the case of the Nearest Neighbor traffic profile. In the case of the Nearest Neighbor pattern, the CRAY-based network saturates at ~36 Gb/s, offering ~14.6% better performance compared to the OE-Router-88ch and confirming in this way that the Titan CRAY XK7 design favors the use of this specific traffic pattern.

For the rest of the patterns, the comparative analysis yields almost similar behavior as for the Uniform Random; although the total maximum capacity of the CRAY Conventional Router is slightly higher compared to the OE-Router-88ch, the system throughput saturates much earlier resulting in significantly worse performance. In the case of the OE-Router-168ch-based layout, the network throughput outperforms both the OE-Router-88ch and the CRAY cases for all traffic patterns including the Nearest Neighbor, taking advantage of the highest router capacity employed in this network configuration.

For the Nearest Neighbor and Bit Rotation cases, the network continues to deliver the peak throughput even after reaching the saturation point, designating the behavior of a stable network. For the Tornado and Bit Complement traffic patterns, the throughput drops beyond the corresponding saturation points, following a similar behavior as in the case of the Uniform Random pattern. Again, this stems from the channel arbitration unfairness introduced by the network routers with respect to the individual packet flows.
of each pattern. The significantly sharper drops experienced in these two patterns compared to the Uniform Random pattern indicate that the unfairness related to these patterns is much more severe than for the Uniform Random case.

For the Shuffle, Transpose and Bit Reverse traffic profiles, the mean node throughput continues to increase even beyond the respective saturation points but at a significantly lower rate. This can be explained by the use of different link capacities in the different dimensions of the network. In multi-dimensional networks with different link capacities per dimension (see Table 2.1), there may be some dimensions that get saturated earlier depending on the applied traffic pattern. As such, a portion of the traffic gets favored as this has only to travel through unsaturated areas of the network, resulting in a lower-rate throughput increase even beyond the saturation point.

Proceeding to the mean message delay measurements shown in Fig. 2.15, mean message delay for the Uniform Random case increases constantly until becoming unbounded at the saturation point. However, before reaching the saturation point, the VCT flow control method performs better offering lower mean message delay values compared to the SF for every individual OE-Router configuration, being fully in agreement with respective theoretical expectations [2.30]. In both the VCT and SF flow control methods, the OE-Router-168ch system outperforms the respective OE-Router-88ch case taking advantage of its 1.9x higher capacity value. Finally, all OE-Router cases outperform the respective CRAY system, which leads to unbounded delay values even from the second measurement at a 20 Gb/s offered load.

Similar behavior is witnessed for the mean message delay performance of the network for all traffic patterns shown in Fig. 2.15, with the Nearest Neighbor forming again the sole exception as the CRAY system offers in this case again lower delay values compared to the OE-Router-88ch system. Table 2.3 provides a summary of the results for both the throughput and delay values and for all available traffic patterns and router configurations. Performance of the CRAY system is illustrated in 2 columns; one presenting the mean node throughput in Gb/s, and the other presenting the mean packet delay in us. The corresponding columns for the OE-Router-88ch and OE-Router-168ch systems include, apart from their individual throughput and delay metrics, the difference as percentage compared with the respective CRAY performance. For all three configurations the reference for throughput metrics is considered the saturation point of the CRAY system. Regarding the delay metrics, performance of the CRAY system just before the saturation point is considered as the reference except for Tornado, Transpose
and Bit Reverse patterns where the CRAY system becomes saturated before the measurement of 10 Gb/s injection bandwidth and consequently this first point is considered as reference. Important to note is that the OE-Router-88ch system provides on average a 50% higher throughput value and a 54% lower delay value compared to CRAY despite the router module has a slightly lower capacity than the Gemini router employed in the CRAY XK7 configuration. The OE-Router-168ch system, when compared to the CRAY system provides even more significant performance improvements, yielding almost 190% higher throughput and 83% lower delay which comes from the combination of the larger bandwidth offered by the optical technology and the optimization of the router channel allocation realized with the use of the ATDT tool.

2.5 Conclusion

This chapter presented OptoHPC-Sim network simulator that supports the utilization of optical interconnect and electro-optical routing technologies at system-scale offering complete end-to-end simulation of HPC systems and allowing for reliable comparison with existing HPC platforms. OptoHPC-Sim can yield valuable feedback on the technology development for future HPC systems by allowing for the performance evaluation of optical blade-enabled HPC network architectures. The chapter presented also the recent technological advances towards implementing high-density and multi-layered EOPCBs that have been taken into account during the OptoHPC-Sim development. OptoHPC-Sim can support the design and utilization of optical interconnect and electro-optical routing technologies at system-scale, offering at the same time complete end-to-end simulation of HPC-systems and allowing for reliable comparison with existing HPC platforms. The comparison analysis between an HPC network system employing state-of-the-art optoelectronic routers and optical interconnects with a system following the Cray XK7 system platform specifications reveals the benefits that can be gained by incorporating these technology advancements to future HPC networks in terms of both throughput and mean message delay.

2.6 List of References


2.3. http://www.ttmtech.com/


“Development of electro-optical PCBs with embedded waveguides for data center and high performance computing applications,” in Proc. of SPIE vol. 8991, 2014


2.21. http://www.enplas.co.jp


2.27. A. Varga, "The OMNeT++ discrete event simulation system," in Proc. of the European Simulation Multiconference (ESM’), Prague, Czech Republic, 2001

2.28. M. Ezell, “Understanding the impact of interconnect failures on system operation,” in Proc. of Cray User Group Conference, May 2013

Chapter 3

Optical Cache Memory Architectures for future Chip-Multiprocessors

3.1 Introduction

The growing disparity between processor and Main Memory (MM) speeds, the well-known “Memory Wall”, has been stalling computer performance increase for the past two decades. High access latency imposed by electronic RAM and limited off-chip bandwidth imposed by electronic technology are the main barriers against the “Memory Wall” elimination. This has led to a large number of on-chip cache memories involved in modern CMPs, which often occupy more than 40% of their chip real estate [3.1], [3.2]. An alternative solution for future computing platforms could potentially emerge from the optical interconnects and photonic integration technology fields. Replacing the CPU-MM buses with optical waveguides has already proven beneficial [3.3], while research in optical memories has reported proof-of-concept demonstrations of all-optical Static-RAM (SRAM) architectures [3.4], operating in principle at significant higher speeds than their electronic counterparts [3.5]. These achievements could in principle lead to high-speed optical cache memories, replacing the entire CPU-MM circuitry with waveguides and off-chip optical cache memories, improving CPU chip area utilization in favour of processing elements. This chapter presents a complete and fully functional physical layer optical cache memory architecture that is capable of performing Write and Read functionalities directly in the optical domain. The proposed architecture complies with the direct cache mapping scheme and is organized in four cache lines with every line being capable of storing two bytes of optical data. WDM formatting of both the memory address and the optical data words is exploited, while the proposed design relies on the interconnection of subsystems that comprise experimentally proven optical building blocks. The optical cache memory architecture comprises five discrete subsystems that are necessary for offering correct cache memory operation: (a) the optical interface module that is used for translating the wavelength-formatted data and address fields between the processor and the cache memory unit, (b) the Read/Write Selector for ensuring correct access to the stored data, (c) the completely passive row decoder for addressing the different storing locations of the cache memory, (d) the 2D optical RAM
bank for storing the actual data, and (e) the tag-comparison circuit used for determining whether the desired data are stored in the cache memory unit. Moreover, minor modifications to the cache architecture can allow for supporting the 2-way set associative mapping scheme as it is being presented later in this chapter. The performance of the proposed optical cache memory architecture is evaluated via physical layer simulations using the VPI Photonics Simulation suite [3.6]. The simulation results present successful functionality during both Read and Write operations and for speeds up to 16 GHz. Given that all necessary subsystems for Read/Write selection, Row/Column Address Selection, Tag Comparison, Access Gate mechanism and optical memory cell function employed in the proposed architecture are designed relying on already experimentally demonstrated optical building blocks [3.4], [3.7]-[3.9], the proposed optical cache memory architecture could potentially form the basic layout of future optical cache systems when extending its size to higher capacity metrics.

Chapter 3 is organized as follows: section 3.2 describes the architectural layout of the proposed optical cache architecture and performs an in-depth analysis of its subcomponents. Section 3.3 provides the performance evaluation of the optical cache architecture through bit-level simulation runs. Section 3.4 elaborates on the scalability perspectives of optical cache concepts while section 3.5 concludes the chapter.

### 3.2 Optical Cache Memory Architecture for Direct Cache Mapping

The optical cache architecture has been designed to comply with the direct mapping access scheme, where every block of bytes in the MM can be stored only at a certain cache line. The principle of direct mapping is schematically represented in Fig. 3.1, showing how every MM block is mapped onto a constant location in the cache for an indicative cache size of four lines. The size of the MM block equals the size of the cache line, so that the entire block of bytes gets transferred even when only a single byte within

![Fig. 3.1: Direct cache mapping](image-url)
this block is requested from the MM, following the principles of the workload spatial locality [3.10]. Cache line selection is predefined for all Memory blocks and is fully determined by the so called Line field within the memory address. For a number of k MM blocks that have to be associated with 2^r cache lines, with r being the number of bits comprising the Line field, the (i modulo 2^r) number corresponds to the cache line that is associated with the i-th MM block. In this way, it is obvious that more than one MM blocks can be stored on the same cache line when k>2^r, as is usually the case, and the information about which MM block resides at a given time in the cache line is provided by the field of tag bits [3.10]. Whenever the CPU requests the content of a certain memory address, the decision about whether this content resides in the cache is determined by comparing the tag bits of the memory address field and the tag bits that are stored in the cache line to be accessed. In the case of identical tag fields, the content can be directly retrieved from the cache without the intervention of the MM, corresponding to a cache hit, while in the case of different tag fields a cache miss is obtained and the requested data have to be transferred again from the MM to the cache.

Fig. 3.2 (a) shows the layout of the proposed optical cache architecture. It is organized in four cache lines with a per line two byte storage capacity, assuming a MM block size of two bytes and direct mapped associativity. The use of five tag bits indicates that the 8-byte cache can be associated with a 2^5x8=256 byte MM module.

![Fig. 3.2: (a) Optical cache memory architecture (b) Memory Address and Data Interfaces](image)
The optical cache comprises five discrete subsystems: i) the optical interface module that is responsible for receiving the memory address and the data words from the CPU Memory Address Register (MAR) and Memory Buffer Register (MBR), respectively, and generating the respective optical memory and word signals during Write operation, as well as for providing the data word content to the CPU registers during Read operation, ii) the Read/Write selection stage, that comprises three optical Write Access Gates (WAGs), iii) the Row Address Selection (RAS) circuit that is responsible for activating only one out of the four cache lines based on the information included in the Line field of the memory address, iv) the Tag Comparison circuit, whose role is to compare the tag bits contained in the selected cache line and the tag bits of the memory address included in the CPU MAR to determine whether a cache hit or cache miss takes place, and finally the v) 2D optical RAM bank that consists of four identical cache lines, where every line can be broken down into three sections: The Tag storage area formed by a Row Access Gate (AG) and followed by an Arrayed Waveguide Grating (AWG-based) CD element prior reaching the five optical FFs where the five tag bits are getting stored, the 1st byte storage and the 2nd byte storage sections, where the 8-bit word#1 and word#2 of the block are stored, respectively. Both 1st and 2nd byte storage sections comprise a RAM Row Access gate (AG) followed by an AWG-based CD unit that terminates to 8 optical memory cells where an optical byte is stored.

Read or Write mode operation is determined by the RW optical signal. When RW equals a logical “0”, write operation is enforced by activating the WAG gates in the Read/Write selection stage and allowing the optical data words and the tag bits of the memory address to enter the 2D optical RAM bank. At the same time, the RAG gates involved in the Tag comparison circuit are blocked prohibiting read operation. Once the RAS subsystem has determined which cache line has to be activated based on the line field of the memory address, all AG elements employed in the non-selected cache lines are blocked impeding communication of their row content with the outer world. Only the three AG elements encompassed in the selected cache line remains open, allowing the tag bits and two optical data words to enter and then get distributed to their corresponding FF cells through the AWG-based CD modules, completing successfully the cache write functionality.

When Read operation is required, the RW value is a logical “1” blocking the WAGs so that the cache content gets protected from altering their content. The RAS circuit selects again the cache line that has to be read and the tag and two byte word information
contained in the selected cache line gets the permission to exit through the respective AG modules, while the tag and word bits of the non-selected cache lines find their respective AG units blocked. The cache line tag content emerging then at the AG output enters then the Tag Comparison circuit in order to participate in the cache hit/miss decision process. The tag bits are spatially discriminated through an AWG and every tag bit gets logically XORed with the respective tag bit contained in the memory address field originating from the CPU MAR register. The outputs of the five logical XOR operations are then multiplexed into a common line through an AWG to form a common “Compare” (COMP) signal output, whose logical value is “0” only when all five XORs provide a 0 at their output implying a cache hit. At the same time, each cache line word is launched as input to respective Read AG modules (RAGs). The RAGs are driven by the inverted RW signal implying that they are open when Read operation is attempted and blocked when Write functionality is performed. In addition, RAGs are also controlled by the Tag Comparison COMP outcome signal. Whenever a cache hit is present and COMP equals “0”, RAGs are activated and the incoming optical words are allowed to transit them and reach the “Data To Read” optical interface, corresponding to successful Read operation. In the case of a cache miss, the COMP signal has a “1” logical value blocking the RAG circuits. Both in the cache hit and cache miss cases, the COMP signal is immediately communicating also the Memory Controller (MC) of the MM module so that the MM gets prepared to send the requested data words to the cache if finally a cache miss is determined. It should be noted that the MC unit receives also the Line and Tag field information of the Memory Address interface almost instantly without having to wait for the final cache hit or miss decision, so as to reduce the number of clock cycles required to fetch the data from the MM in case a miss occurs. This immediate communication of the Tag and Line fields to the MC unit expedites also the Write operation process in case the cache coherency protocol enforces the constant update of the MM about the newly written data in the cache [3.11]-[3.13]. A flow-chart of the cache memory operation concept is depicted in Fig. 3.3.

A more detailed analysis for each optical cache subsystem follows in the rest of this section, providing also details about their implementation in the complete optical cache physical layer simulations carried out via the VPI Photonic simulation software package.
3.2.1 Interfaces to CPU and Main Random Access Memory

Fig. 3.2 (b) provides a more detailed illustration of the Memory Address and Data Interfaces, considering a four-bit size for every possible field (Line, Tag, Data-To-Write and Data-To-Read) for simplicity purposes instead of the field sizes used in the cache design of Fig. 3.2 (a) (5-bit Tag, 2-bit Line and 8-bit Data words). The Memory Address Interface receives the Line and Tag bits from the CPU MAR register and forwards them to the next block of the cache architecture, as well as to the MM MC. The latter ensures that this information is made immediately available to the Main RAM memory so as to reduce latency in case a cache miss takes place during Read operation, since the data words have to be then fetched from the MM into the cache unit again. Considering that optically connected solutions for the CPU-cache and for the cache-Main RAM communication can be available [3.3], [3.14]-[3.17] we can assume that both the CPU and the MM can make Line and Tag bits as well as their inverted values $Line$ and $Tag$ available to the cache. Moreover, we employ WDM-formatting for both Line and Tag fields as well as for their complementary values, assuming their generation by directly modulated VCSEL arrays, where each VCSEL emits at a different wavelength. Similar approaches have been widely utilized in optochip deployments so far [3.18]-[3.20], reporting on modulation speeds up to 25 Gb/s and currently moving towards 40Gb/s data rates. The Line and Tag VCSEL array outputs are then multiplexed into a Line and Tag optical bus, respectively, via two corresponding optical Arrayed Waveguide Grating (AWG) multiplexers.
The Data Interface consists of two discrete parts, the Data-To-Write and the Data-To-Read units that are used when Write or Read operation are performed, respectively. The Data-To-Write interface receives the data words from the CPU MBR register and forwards them in the optical cache when Write operation is required, again assuming that also the word information is available by the CPU MBR unit. Similar to the address field bits, WDM-formatting is used also for the data words, assuming again the employment of a WDM VCSEL array. The VCSEL array output for every data word gets again multiplexed into a common Data word optical bus through an AWG module. During Read operation, a respective AWG demultiplexer and a photodiode (PD) array are used at the Data-To-Receive interface in order to welcome the WDM formatted data words fetched from the cache module and to convert them into their electrical form prior allowing their usage from the CPU MBR register. PD array solutions have been already proven efficient for a number of up to 168 photodiodes [3.18] and for speeds up to 25 Gb/s [3.19] per PD element.

It should be noted that the WDM formatting of the memory address and data word information does not imply the use of a continuously increasing number of wavelengths as the cache design scales to higher capacities. The proposed design supports spatially discriminated optical buses for the Line, Tag, word#1 and word#2 fields, so that the same number of wavelengths can be reused in these four information fields. It should also be mentioned that instead of the VCSEL-based WDM signal generation solution, the proposed cache architecture could alternatively also rely on already demonstrated silicon-based solutions for the CPU-DRAM optical interfaces [3.14]. In order to minimize four-wave-mixing phenomena imposed by the presence of multiple wavelength signals in the subsequent optical cache building blocks, the wavelengths of the VCSEL emitter array and all respective AWG elements have been selected according to the algorithm presented in [3.21].

The list of assigned wavelengths to the Tag and Data bits can be seen in Table 3.1. As can be easily seen, the spatial multiplexing that is employed allows for re-using the

<table>
<thead>
<tr>
<th>WORD1</th>
<th>WORD2</th>
<th>TAG</th>
<th>( \lambda_{D_1}, \lambda_{T_1} )</th>
<th>( \lambda_{D_2}, \lambda_{T_2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>D9</td>
<td>T1</td>
<td>1.564,19</td>
<td>1.559,31</td>
</tr>
<tr>
<td>D2</td>
<td>D10</td>
<td>T2</td>
<td>1.562,56</td>
<td>1.557,36</td>
</tr>
<tr>
<td>D3</td>
<td>D11</td>
<td>T3</td>
<td>1.554,78</td>
<td>1.548,35</td>
</tr>
<tr>
<td>D4</td>
<td>D12</td>
<td>T4</td>
<td>1.551,88</td>
<td>1.546,12</td>
</tr>
<tr>
<td>D5</td>
<td>D13</td>
<td>T5</td>
<td>1.542,30</td>
<td>1.538,19</td>
</tr>
<tr>
<td>D6</td>
<td>D14</td>
<td>-</td>
<td>1.576,69</td>
<td>1.572,72</td>
</tr>
<tr>
<td>D7</td>
<td>D15</td>
<td>-</td>
<td>1.574,37</td>
<td>1.571,07</td>
</tr>
<tr>
<td>D8</td>
<td>D16</td>
<td>-</td>
<td>1.568,44</td>
<td>1.566,80</td>
</tr>
</tbody>
</table>

Table 3.1: Wavelength assignment on Data and Tag bits
assigned wavelengths in different bit words. More specifically, having Word1 comprising \( D_i \) bits and their respective wavelength pairs, the same wavelength pairs are re-used for carrying the bits of Word2 and Tag streams.

### 3.2.2 Read/Write Selection Stage

The Read/Write selection mechanism is carried out by three identical WAG gates, all of them being controlled by the Read/Write (RW) optical signal. One WAG gate uses the optical tag bits as its input signal, while each of the remaining two WAGs exploits one of the incoming optical words (word#1 and word#2) as input. Each WAG is implemented by a SOA-based MZI (SOA-MZI) configured as a differentially-biased switch with bi-directional control signal injection, following the layout described in [3.22]. In the absence of any control signal, i.e. RW=0, Write operation has to be executed and the incoming tag and word data bits transit the respective WAGs and get forwarded to the Row AG stage of the 2D optical RAM bank subsystem. When Read operation is enforced, RW equals a logical “1” blocking the WAG, impeding in this way the tag and data word bits to reach the subsequent optical cache subsystems.

At the same time, the Read/Write Selection mechanism determines whether the cache content will reach the Data-To-Read interface through controlling the RAG switch employed in the Tag Comparator stage. The RAG switch device is identical to the WAG, being controlled by the logical sum of the \( RW \) and COMP signals, the latter originating from the XOR outputs involved in the Tag Comparator. While in Write mode, the \( RW \) equals a logical “1” and prohibits the cache word content to propagate through the RAG and reach the Data-to-Read interface module irrespective of the COMP signal value. In the case of Read functionality, \( RW \) is “0” allowing the cache words to exit the RAG and get successfully received by the Data-To-Read interface, yielding successful Read operation.

### 3.2.3 Row Address Selection stage & Row Access Gates

A detailed schematic representation of the RAS stage followed by the Row AGs is shown in Fig. 3.4 (a). RAS is responsible for activating only one out of the four possible cache rows, decoding the information that is encompassed in the incoming Line bits. Taking advantage of the WDM-formatted Line and \( Line \) bit streams, RAS is performed
via an all-passive optical configuration initially proposed in [3.9] and already demonstrated in a proof-of-concept experiment in [3.23].

For a four-line cache unit, a 2-bit Line field is required as input. Considering that the Line and Line bits are both encoded onto a total number of four different wavelengths, i.e. $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, the multiplexed WDM signal is broadcasted into the four RAS rows through optical couplers. Every row utilizes a cascade of two band-rejecting filters centered at two of the incoming wavelengths and configured to support all possible permutations of the four incoming wavelengths, forming in this way an all-passive $\lambda$-selective matrix. The $\lambda$-selective matrix was implemented in the cache design by means of cascaded first-order Silicon-on-Insulator Add/Drop microring resonators acting as band-rejecting filters, as shown in Fig. 3.4 (a). An indicative combined spectrum of all four band rejecting filters used in the RAS can be seen in Fig. 3.4 (b), with every resonance having an average -3dB bandwidth of 0.4 nm and extinction ratio higher than 30 dB. In this way, a “0” optical output is obtained only at the row that has to be

Fig. 3.4: (a) Row Address Selector (RAS) architecture (b) combined spectrum of four ring resonator-based filters used in the RAS (c) RAS truth table (d) Ring resonator-based Filter design parameters
activated, while all non-selected rows will provide optical power at their output in the form of either one or two wavelengths, as shown in the Truth Table of Fig. 3.4 (c). The dimensions and design characteristics of the filters employed can be seen in the Table of Fig. 3.4 (d), following parameters that have been already reported in the literature with respect to successfully fabricated SOI microring structures [3.24].

After exiting the \( \lambda \)-selective matrix, the RAS output signals are used for driving the Row AGs employed at the front-end of the 2D optical RAM bank. More specifically, every of the four possible RAS outputs drives three AG units that control the access to a complete optical RAM row. As can be seen in Fig. 3.4 (a), AG1 is responsible for the access of the Tag and \( \text{Tag} \) bits, while AG2 and AG3 are each responsible for one out of the two incoming or outgoing 8-bit optical words and their respective complementary \( \text{Word} \) values. All AG units have been again implemented following the differentially-biased bidirectional SOA-MZI switch layout [3.22], similar to the WAG and RAG arrangements already described in the Read/Write selection stage. It should be noted that SOA-MZIs have been already demonstrated experimentally to operate when using multi-wavelength input signals [3.25] as well as when having both their input and control signals in multi-wavelength format [3.23].

In the case of an activated RAS line, the RAS output reaching the three AGs as their control signal will equal “0”, so that the propagation of the Tag and Data word signals as well as of their complementary values is allowed through the AG. In the case of a non-selected RAS row, the optical power of the non-rejected Line and \( \text{Line} \) wavelengths reaching the AGs will change the operational state of the AGs prohibiting the Tag and Data signals to transit the AG elements and continue to the optical RAM row.

### 3.2.4 2D optical RAM Bank: Column Address Selector (CAS) and optical Flip-Flops

The Tag and Data word signals are inserted into the Column Address Selector (CAS) unit after being granted access through the Row AG#1, AG#2 and AG#3 to a specific cache line while operating in Write mode. The CAS module is a completely passive configuration relying on three discrete AWG components, with every individual AWG located directly at the output of the respective Row AG module, as has been already demonstrated in [3.7], [3.25]. Taking advantage of the WDM-formatted Tag bits, the AWG1 following the Row AG1 is responsible for spatially separating the incoming Tag
and Tag bits into the respective AWG output ports. In the same way, the WDM-formatted word#1 and word#2 data bits and their complementary values entering AWG2 and AWG3, respectively, which are located at the output of the corresponding Row AG#2 and AG#3, get demultiplexed into the discrete AWG output ports. Any two AWG outputs that produce the pair of wavelengths corresponding to an incoming bit and its respective inverted value are considered to form a specific column within the 2D optical RAM bank. Consequently, each wavelength pair emerging at a column will drive an individual optical FF, providing the necessary Set and Reset signals for the proper FF operation. Given the 5-bit size of the Tag field and the associated 10 wavelengths required for representing the Tag and Tag information, AWG1 is a 1x10 structure that drives five wavelength pairs to the correct optical FFs. Following the same principle, AWG2 and AWG3 are 1x16 structures that drive 16 wavelength pairs to the correct FFs to store the 8-bit optical word#1 and word#2, respectively.

During Read mode operation, the contents of the optical FF row propagate through the AWG and emerge at the AWG common port getting multiplexed into a common WDM-formatted data signal before being fed into the Row AG unit, this time in an opposite direction compared to the case of Write functionality.

The optical FF implementation that has been used in the proposed optical cache design for storing the Tag and Data bits relies on the coupled SOA-MZI-based optical FF architecture [3.26], which has been already demonstrated experimentally even in complete optical RAM cell functionality [3.4] and holds the potential to operate at speeds up to 40Gb/s [3.5]. It should be noted, however, that our proposed optical cache architecture can operate with any optical FF configuration that utilizes wavelength diversity in the input and output signals, like for example the III-V-on-SOI microdisk lasers [3.27]. Fig. 3.5 depicts the optical FF layout that has been used for the purposes of our simulation, closely following the experimental configuration reported in [3.26]. It comprises two coupled SOA-MZIs (MZI1 and MZI2) powered by two external continuous-wave (CW) input signals at wavelengths that are identical to the incoming Set
and Reset signal wavelength pair, i.e. $\lambda_a$ and $\lambda_b$ in Fig. 3.5. A detailed description of the principle of operation of this specific optical FF arrangement can be found in [3.4], [3.26].

Having described so far all subsystems being necessary during Write mode operation, Fig. 3.6 (a) shows an indicative example of the whole Write procedure depicting the optical signals at the various cache stages starting from the optical interfaces up to getting stored in the optical FF. All Tag and Data signals employed follow a Non-Return-to-Zero (NRZ) $2^7$-1 PRBS sequence format at 16 Gb/s. Fig. 3.6 (a)(i) illustrates the RW toggling signal, which is injected into the WAG Selection gate as control signal. Fig. 3.6 (a)(ii) and Fig. 3.6 (a)(iii) show the Line bit streams encoded on $\lambda_1$ and $\lambda_2$ after exiting the RAS row that corresponds to a “00” RAM row and as such rejects only the inverted Line bits carried by $\lambda_1$ and $\lambda_2$ wavelengths. These two RAS output streams are
being injected as a single optical control signal into the Row AGs (AG1, AG2 and AG3), implying that access of the Tag and Data bits to the “00” row of the complete 2D optical RAM bank will be granted only when both λ1 and λ2 signals are simultaneously “0”. A pair of the complementary Data streams denoted as Bit and Bit̅ is shown in Fig. 3.6 (a)(iv) and 3.6 (a)(v), respectively, prior entering the Read/Write WAG stage. Figures 3.6 (a)(vi) to 3.6 (a)(vii) show the corresponding pair of Data signals after having sequentially propagated through the R/W Selection WAG gate, the “00” RAM row AG module and the AWG-based CAS stage, forming the Set and Reset signals prior entering the optical FF. As can be seen, whenever the R/W signal has a logical “1” value (Read mode) the Data streams are prohibited to reach the optical FF resulting to “0” logical values for both the Set and Reset signals. Moreover, even if the R/W signal has a logical value of “0” suggesting Write mode operation, the Data streams are again not allowed to reach the optical FF in case at least one of the λ1 and λ2 Line signals equals a logical “1”.

To this end, Data bit and Bit̅ streams will transit the WAG, the Row AG and the subsequent AWG-based CAS module only when all three λ1 and λ2 Line and R/W bits are ‘0’ simultaneously. In this case, the Set signal will carry the respective information of the Data Bit sequence while the Reset signal will be identical to the Data Bit̅ value. In the example depicted in Fig. 3.6 (a), the Reset signal has a wavelength of λ1=1528.77 nm, while the Set signal has a wavelength of λ2=1524.42 nm. The Set signal will enter MZI-2, which is powered by a CW signal at λ2 in this case, while the Reset signal will enter MZI-1, which is powered by a CW signal at λ1. Fig. 3.6 (a)(viii) and 3.6 (a)(ix) illustrate the corresponding FF content as it is imprinted at λ2 and λ1 at the S2 and S1 MZI output ports, respectively. When all three R/W and λ1 and λ2 Line signals equal simultaneously “0”, the FF content is modified according to the respective Set and Reset bits. In the presence of a Set optical pulse, a logical “1” is stored and monitored as FF OUT signal obtained at the S2 port of MZI-2. Whenever a Reset optical pulse enters the FF, the FF OUT switches back to zero output level retaining this state until the arrival of the next Set pulse. In all cases, FF OUT obtained at the S1 port of MZI-1 continuously holds the complementary value of FF OUT.

It should be noted that the pulse traces depicted in Fig 3.6 (a) do not incorporate the latency induced by every cache subsystem in order to allow for a clearer visualization of the obtained results and to verify the successful Write functionality. The timing diagram and the associated latency during Write cycle are shown in detail in Fig. 3.6 (b). Fig. 3.6 (b)(i) through Fig. 3.6 (b)(v) depict the pulse traces for the RW, λ1, λ2, BIT and
signals, which are perfectly synchronized when leaving the CPU interface towards the cache memory. Fig. 3.6 (b)(vi) and 3.6 (b)(vii) depict the Set and Reset pulse traces, respectively, after exiting the AWG-based CAS stage within a cache line and prior entering the optical FF. Fig. 3.6 (b)(viii) and 3.6 (b)(ix) illustrate the FF content and its complementary value as recorded directly at the FF output. As can be noticed, the focus in this timing diagram will be within two distinct Write cycles where the FF state is modified, commencing at t1 and t4 time instants. Once the RW signal gets “0” and enters the R/W selector’s WAG at t1 and t4, respectively, the optical words to be written in the cache will appear at the AG array input ports after a certain time duration $\Delta t_1$, which involves the time-of-flight in the Data-to-Write Interface AWG and in the R/W Selector WAG. Assuming AWG dimensions as reported in [3.28], the time of flight in the AWG turns out to be 4.65 ps, while in the case of the WAG gate the time-of-flight equals 9.1 psec considering a SOA length of 1.1mm [3.29]. As such, $\Delta t_1$ equals 13.75 psec if passive connections of negligible length are assumed between the AWG, the WAG and the AG units. The $\lambda_1$ and $\lambda_2$ Line bits can get easily delayed by means of passive waveguide lines in the RAS stage so as to enter the AG array as control signals again after $\Delta t_1$, in order to allow for synchronized operation with the optical words within the AGs. The logical AND operation between the Line bits and the incoming optical words within the AG as well as the subsequent propagation of the optical word bits through the CAS-included AWG result to an additional $\Delta t_2$ delay until the Set and Reset pulses appear at the FF input ports. This $\Delta t_2$ incorporates again the time-of-flight values for propagating through the AG module and the following CAS-included AWG, so that $\Delta t_2 = \Delta t_1$. As such, the Set and Reset pulses emerge at the FF inputs at t2 and t5, with $t_2-t_1 = t_5-t_4 = \Delta t_1 + \Delta t_2 = 27.5$ psec. The time needed for the FF to respond upon Set/Reset pulse injection equals 12.5 psec, considering that the 10-90% rise/fall times of the FF logical levels are in the order of 25psec [3.29] as can be also confirmed in Fig. 3.6 (b)(viii) and 3.6 (b)(ix). To this end, the originally incoming Bit values at t1 and t4 emerge as FF content at t3 and t6, respectively, with $t_3-t_1 = t_6-t_4$ denoting the entire latency during a single Write cycle. Given that $t_3-t_1 = t_2 + 12.5 \text{ psec} = 40 \text{ psec}$, this indicates that there are 22.5 ps left within a single Write cycle of 62.5 ps at 16 GHz, which could be used for any potential additional propagation paths and still have a total latency lower than a clock cycle. Assuming silicon-based passive waveguide circuitry, the 22.5 psec correspond to approximately 2.7mm of total path length required for having all necessary elements interconnected and still retain latency below one clock cycle.
Fig. 3.7: Spectrum of: (a) Line and Tag input signals at the RAS input (b) Line and Tag input signals at the RAS output (c) 5-bit Tag and 5-bit at the input of the Row “00” AG1 (d) Row “00” AG1 output

Fig. 3.7 illustrates the respective spectra of the Line and Tag field signals during Write operation at various stages of the setup. Fig. 3.7 (a) shows the four Line and Tag input wavelengths at the input of the RAS stage, while Fig. 3.7 (b) depicts the respective spectra obtained at the output of the “00” RAS row and prior entering the Row AG1, where only $\lambda_1$ and $\lambda_2$ are obtained since $\bar{\lambda}_1$ and $\bar{\lambda}_2$ have been rejected by the RAS row. Fig. 3.7 (c) represents the 5-bit Tag and 5-bit Tag field spectrum at the input of the Row “00” AG1, while finally Fig 3.7 (d) shows the output of the Row “00” AG1 before reaching the AWG-based CD unit, where all ten Tag and Tag wavelengths as well as the two $\lambda_1$ and $\lambda_2$ Line wavelengths are present. The extra wavelength observed at 1531.89 nm is used as an external Continuous Wave (CW) signal inserted into the Row AG in order to provide the necessary SOA-MZI differential biasing as reported in [3.22].

3.2.5 Tag Comparator Stage

The Tag Comparator module is utilized during Read mode to compare the Tag bits emerging from the optical cache line with the Tag bits of the CPU-requested memory address. In every Read cycle, a set of data is requested by the CPU from a specific cache line, which is fully determined by the Line field of the memory address. As such, the Line bits will activate the appropriate Row AG elements allowing the Tag and word contents of the selected cache line to get forwarded through the corresponding AWG-based CAS
stage to the Row AGs. The Tag and word bit streams are then forwarded to the Tag Comparison stage: the two optical data words will enter two respective RAG modules as input signals, while the Tag bits are demultiplexed in an AWG device in order to be compared with the CPU-requested Tag field. The comparison is conducted through a logical XOR operation implemented by means of a SOA-MZI dual-rail control logic switch, as has been already demonstrated in [3.30]. A total number of five parallel SOA-MZI-based XOR gates is utilized, as shown in Fig. 3.8 (a), with every gate comparing the corresponding cache-included and CPU-requested Tag bits.

Every SOA-MZI XOR gate exploits a CW input signal and a dual-rail control logic with the cache-included and the CPU-requested Tag bits acting as the two control signals that enter the two MZI branches, respectively, as shown in Fig. 3.8 (b). Following the SOA-MZI-based XOR principle described in more detail in [3.30], the XOR output signal will have a logical “1” value imprinted on the CW input wavelength only when the two Tag bits are different, denoting a cache miss. In case of identical Tag bit values, the XOR will produce a logical “0” at its output. By selecting five different wavelengths for the five CW signals powering the 5-element XOR array and using an AWG module at the XOR array output, the five XOR outputs are multiplexed into the common AWG port performing a logical WDM-enabled OR operation between the five individual XOR outcomes. The 5-input OR result comprises the COMP output signal and indicates a successful cache hit only when having a logical “0” value, i.e. all five XORs have attained
perfect matching between the cache-included and the CPU-requested Tag bits. In all other cases, the COMP signal will be in general a multi-level WDM-formatted signal designating a cache miss. It is important to note that the COMP signal generated by the Tag Comparison stage is immediately communicated also to the MC of the Main RAM Memory, so as to immediately initiate all necessary MC functions for transferring the correct data from the MM to the cache unit in the case of a cache miss. The COMP signal is then used as the control signal into the RAG modules to decide upon the transmission allowance of the cache-line word content. The detailed implementation and operation of the RAG modules have been described in Section 3.2.2.

Fig. 3.9 illustrates an indicative Read mode operation example describing the path followed by the Bit and Bit content of a single optical FF until arriving at the Data-To-Read interface module, again ignoring the timing delays induced by the different cache sub-systems. Fig. 3.9 (i) shows the RW signal, while Fig. 3.9 (ii) and 3.9 (iii) depict the Line bit streams encoded onto the λ1 and λ2 wavelengths, respectively, and launched as the control signal into the AG elements of row “00”. Fig. 3.9 (iv) shows the Tag1 bit information stored in the first of the five optical FFs used in the optical cache line for Tag field storage, while Fig. 3.9 (v) represents the corresponding Tag1 information that enters the XOR1 comparison stage as the control signal. As can be noted, the Tag1 FF content will emerge at the XOR1 control input only when the two Line bits equal simultaneously “0”, implying successful “00” cache line selection. Fig. 3.9 (vi) shows the corresponding Tag1 information requested by the CPU and entering XOR1 as the second control signal to be compared with Tag1 originating from the cache line FF. The logical XOR result is depicted in Fig. 3.9 (vii), confirming successful XOR operation since optical pulses are obtained only when the two XOR1 control inputs are of different values.

In the same way, the set of Figures 3.9 (viii) until 3.9 (xxiii) can be divided in groups of four figure rows each, with every group illustrating the corresponding Tag#i information stored in the cache line FF, the Tag#i stream stemming from the FF and reaching XOR#1 as the control signal, the Tag#i sequence requested by the CPU and the final XOR#i output beam, with i=2,3,4,5. Fig. 3.9 (xxiv) depicts the word bit information D1 that is contained in the optical FF of the respective cache line, which has to be read by the CPU whenever requested. The Read output result emerging at the RAG device
Fig. 3.9: Read mode operation example showing the complete set of XOR Tag Comparisons and the path of a stored Data bit to the Data-to-Read Interface. Time scale: 62.5ps/div for the pulse traces - 20ps/div for the eye diagram
output and reaching the Data-to-Read interface module is shown in Fig. 3.9 (xxv). As can be easily noticed, the D1 cache content is successfully read at the cache output only when the following conditions apply simultaneously: RW=1, denoting Read mode operation, the λ1 and λ2 Line bits equal both “0” implying successful “00” cache line selection, and all five XOR output signals equal “0” suggesting identical cache-included and CPU-requested Tag fields and corresponding to a cache hit. Fig. 3.9 (xxvi) depicts the eye diagram of the D1 cache content when emerging at the Data-To-Read optical interface module, showing a clearly open eye with an extinction ratio higher than 12 dB, an average amplitude modulation of 1.2 dB and a Quality (Q-)factor equal to 8.8, which corresponds to a Bit-Error Rate (BER) better than $10^{-12}$.

The delays induced by every subsystem within a single Read cycle and the respective timing diagram are illustrated in Fig. 3.10. Fig. 3.10 (i) depicts the pulse trace for the RW signal when leaving the CPU interface towards the cache memory’s R/W selector. The time periods between the dash-dotted lines $t_1$-$t_5$, $t_5$-$t_6$, $t_6$-$t_10$ and $t_10$-$t_13$ represent the corresponding clock cycles for the RW signal. Fig. 3.10 (ii) and 3.10 (iii) depict $\lambda_1$ and $\lambda_2$ pulse traces, respectively, after exiting the Row Address Selector’s row “00” and prior entering the 2D optical RAM bank’s AGs. Once the RW signal gets a value of logical “1” and enters the R/W selector’s WAG at $t_1$ and $t_10$, respectively, the $\lambda_1$ and $\lambda_2$ Line bits will appear at the AG array input ports after a certain time duration of $\Delta t_1 = t_2 - t_1 = 13.75$ psec. This corresponds to the delay time induced by the passive waveguide lines in the RAS stage as already described in the timing diagram for Write mode operation.

![Fig. 3.10: Timing diagram for the Read mode operation](image-url)
Fig. 3.10 (iv) and 3.10 (v) illustrate the D1 and Tag4 FF contents, respectively, as they are recorded prior entering the AGs of row “00”. Fig 3.10 (vi) depicts Tag4 FF content prior entering the XOR4 gate, i.e. after having been logically ANDed with the λ1 and λ2 Line bits in the AG of row “00” and after having propagated through the Tag Comparator’s AWG. The total delay for these two stages equals to Δt2=t8-t7=13.75ps, including the time-of-flight in the AG’s SOAs, which equals to 9.1ps considering a SOA length of 1.1mm [3.29] and time-of-flight in Tag comparator’s AWG, which equals to 4.65ps assuming the AWG dimensions as reported in [3.28]. Fig. 3.10 (vii) depicts the XOR4 input as requested by the CPU, which is completely synchronised by means of passive waveguide delay lines with the XOR4 input coming from the Tag4 FF. Fig 3.10 (viii) depicts the XOR4 result as it is recorded at the XOR4 output after a time delay of Δt3=t4-t3=t9-t8=9.13ps, which corresponds to the time-of-flight in the XOR4’s SOAs. Fig. 3.10 (ix) depicts the final output emerging at the Data-to-Read interface, which carries the D1 information depicted in Fig 3.10 (iv). After the XOR output, the final output is delayed by Δt4=t12-t11=18.43ps, a value which includes the time-of-flight of the XOR4 output in the AWGs prior to- and after the RAG and the RAG itself. The D1 contents are delayed accordingly using passive waveguide delay lines, in order to get synchronous AND operation at the RAG. It should be noted that t11 is the time instance where the XOR product emerges at the XOR4 output, as is the case for t4 and t9. Finally, by adding all the Δt1, Δt2, Δt3 and Δt4 delay values induced by cache subsystems during the Read mode operation we get a total latency lower than a single clock cycle and equal to 55.12ps. This can also be derived from Fig. 3.10, where the final Read output of the Read-operation clock cycle starting at t10 is obtained at the Data-to-Read interface at t12, before the end of the clock cycle at t13.

3.3 Simulation Results

Fig. 3.11 illustrates an example of the Write procedure for storing an 8-bit Data word. Fig. 3.11 (i) depicts the RW toggling signal, which is injected into the WAG Selection gate as control signal, while Fig. 3.11 (ii) and Fig. 3.11 (iii) show the Line bit streams encoded on λ1 and λ2 after exiting the RAS row. As this example illustrates the writing of an 8-bit word to the “00” RAM row, only the inverted Line bits carried by λ1 and λ2 wavelengths are rejected. These two RAS output streams are being injected as a single optical control signal into the Row AGs and access of the Tag and Data bits to the “00” row is granted only when both λ1 and λ2 signals are simultaneously “0”. The set of
Figures 3.11 (iv) until 3.11 (xix) can be divided in figure row pairs, with every pair illustrating the Bit\#i Data stream prior entering the Read/Write WAG stage and the corresponding D\#i OUT signal finally stored as the content of the respective FF, with i=1,2…8. As already described for the case of writing to the “00” cache line, Data bit and Bit streams will transit the WAG, the Row AG and the subsequent AWG-based CAS module only when all three λ1 and λ2 Line and R/W bits are ‘0’ simultaneously. These cases are highlighted as gray-shaded areas in Fig. 3.11. Consequently, the FF content is altered according to the rules described in section 3.2.4 and the 8-bit Data word is stored successfully. The FF-output signals had an average extinction ratio value
of 6.2 dB while the average amplitude modulation was 1.9 dB, a performance similar to the respective experimental values reported in literature [3.4], [3.26].

Fig. 3.12 illustrates an example of the Read mode operation describing the path followed by an 8-bit Data word from its FF storage location until arriving at the Data-To-Read interface module. Fig. 3.12 (i) shows the RW signal, while Fig. 3.12 (ii) and 3.12 (iii) depict the Line bit streams encoded onto the λ1 and λ2 wavelengths, respectively, and launched as the control signal into the AG elements of row “00”. Only when the RW signal equals “1” (signaling Read operation) and the two Line bits equal “0” simultaneously (implying successful “00” cache line selection), the Tag FF content will emerge at the XOR gate to serve as control signal. As already described in section 3.2.5, the Tag#i stored in one of the five optical FFs used in the optical cache line will be driven into the Tag Comparator and will be compared through a XOR operation to the Tag#i information requested by the CPU, with i=1,2,…5. The five logical XOR results are depicted in Fig. 3.12 (iv) until Fig. 3.12 (viii), where optical pulses are obtained only when the two XOR#i control inputs are of different logical values.

The set of Figures 3.12 (ix) to 3.12 (xxxii) can be divided in groups of three figure rows each, with every group illustrating the corresponding D#i information stored in the cache line FF, the D#i Final Out stream reaching the Data-to-Read interface module and its complementary value $\overline{D}#i$ Final Out, with i=1,2,…8. The gray-shaded areas in Fig. 3.12 designate the bit slots where the 8-bit Data word can be successfully retrieved from the cache memory, corresponding to the case of having RW=1, both λ1 and λ2 Line bits and all five XOR output signals being “0”. The Line, Tag and Data input signals had an extinction ratio value of 20 dB, while all final Read output signals had extinction ratio values higher than 12 dB and average amplitude modulation of 1.2 dB, indicating a significantly improved signal quality compared to the originally stored FF content.

All Data signals followed a Non-Return-to-Zero (NRZ) $2^7$-1 PRBS sequence format at 16 Gb/s. The SOA model used in the VPI simulation environment and utilized in all SOA-based subsystems was identical to the model presented by Kehayas et al in [3.31], having a small signal gain of 27dB and a gain recovery time of 25psec when driven at 300mA. The optical power levels used at all SOA-MZI modules employed as cache subsystems were very close to respective experimental observations [3.4], [3.22], [3.30] and are summarized in Table 3.2. For the optical FF layout depicted in Fig. 3.5, the CW input power for both MZIs was equal to -11.5 dBm, while the peak powers of the Set
Fig. 3.12: Read mode operation example showing the reading of an 8-bit Data word from the 2D Optical RAM bank and its transmission to the Data-to-Read Interface. Time scale: 62.5ps/div.
and Reset optical pulses were around 3.2 dBm. The peak power of the pulse sequence emerging at the FF output was ranging between -9.3 and -8.1 dBm, depending on the specific values of signal wavelengths employed. With respect to the passive components used in the proposed cache layout, loss values of 0.2dB per coupler, 1dB per ring resonator-based filter and 1.5 dB per AWG were assumed. Without involving any amplification stages, the optical peak power levels emitted by the Data-To-Write and Tag VCSEL arrays were around to ~ -7 dBm and ~5 dBm, respectively. The respective optical peak power level emitted by the Line VCSEL array was 19 dBm per channel, as the total propagation losses until the Access Gates were ~13 dB. As VCSEL sources cannot emit such high power values, an SOA amplification stage would have to be added after each RAS row output in an integrated version of the proposed optical cache layout.

### 3.4 Optical Cache Scalability Analysis

The physical layer optical cache architecture demonstrated as demonstrated in section 3.2 has a capacity of only 64 bits organized in four cache lines with all of its active subsystems relying on SOA-based implementations. The latter was promoted for two main reasons: a) due to the enhanced maturity and reliability of the SOA simulation model compared to alternative integrated switch technologies being available in the VPI software suite, and b) in order to follow respective experimental layouts for the different cache subsystems like the optical RAS/CAS stages including the multi-wavelength AGs [3.7], the optical SOA-MZI-based FFs [3.4], the optical XOR gates [3.30] and the

<table>
<thead>
<tr>
<th>Cache Subsystem</th>
<th>Input signal (peak power in dBm)</th>
<th>Control signal (peak power in dBm)</th>
<th>External CW (avg. power in dBm)</th>
<th>Output signal (peak power in dBm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAG</td>
<td>-9.2 to -8.2*</td>
<td>6.00 (RW)</td>
<td>-0.7</td>
<td>4 to 5.3*</td>
</tr>
<tr>
<td>AG 1-3</td>
<td>-9.2 to -8.2*</td>
<td>6.00 (per Line ch.)</td>
<td>-0.7</td>
<td>4 to 5.3*</td>
</tr>
<tr>
<td>RAG</td>
<td>-9.2 to -8.2*</td>
<td>6.00 (per COMP ch.)</td>
<td>-0.7</td>
<td>4 to 5.3*</td>
</tr>
<tr>
<td>XOR 1-5</td>
<td>-3 (CW avg. Power)</td>
<td>1.00 (FF &amp; CPU TAG ch.)</td>
<td>'-'</td>
<td>6.00 (per XOR gate)</td>
</tr>
</tbody>
</table>

*per channel
differentially-biased SOA-MZI \([3.22]\) WAG/RAG units. However, scaling the optical cache to higher capacity values can probably not rely on SOA-based implementations, since they would typically result to high-cost, large-size and high-power cache layouts, given the well-known size- and energy-hungry characteristics of SOA technology. Attempting higher capacity SOA-based cache demonstrations at least via physical layer simulations was also not effective, mainly due to software execution time constraints associated with the VPI simulation platform.

Despite small capacity cache modules can be of practical use in certain application areas like optical packet routing with small-size optical packet traffic, higher capacity values are certainly needed when targeting CMP environments. Although the SOA-based technology approach seems not to support this perspective, the proposed optical cache architecture is in principle fully compatible with alternative active and passive photonic technology blocks that hold all necessary credentials for low-power, low-size and eventually future low-cost deployment. Examples include the low-power optical microdisk laser FF \([3.27]\) and optical switch \([3.32]\) technologies relying on the III-V-on-SOI material platform, or even the more disruptive photonic crystal (PhC) nanocavity technology platform that has made remarkable progress in low-power and ultra-compact all-optical RAM \([3.33]\), switch \([3.34]\) and multiplexer \([3.35]\) technology.

Towards addressing possible high-capacity future optical cache developments, Fig. 3.13 illustrates a possible 8kB cache architecture when adopting the PhC nanocavity
technology platform, relying of course on high-level assumptions regarding the interconnectivity of different PhC-based building blocks and the actual chip-scale topology layout. The 8kB size was selected so as to comply with state-of-the-art L1 capacity standards involved in current CMPs, like in the case of the NIAGARA CMP that incorporates a 8kB 4-way L1 data cache organized in four different cache sets with each set comprising 128 rows and a line size of 16 bytes [3.36]. This corresponds to a total number of 4 sets x 16 byte/block x 8 = 512 bits stored per row, or alternatively to 64-Byte columns.

The AG and XOR gate implementation in the PhC-based cache design of Fig. 3.13 relies on the all-optical Fano resonance switch based on coupled PhC nanocavities [3.34]. A total of 8448 switches are employed as AGs, taking into account 64 AG devices per cache row (one AG per column). Assuming a 16-bit Tag field, the XOR gate array will require 16 Fano resonant elements, yielding a total number of 8464 Fano resonance PhC nanocavity switches. The optical FFs can be realized by means of the ultralow-power all-optical nanocavity-based memory cells presented in [3.33], with the 8KB cache memory with 16-bit tag field necessitating a total number of 67584 FF elements. Apart from the active subsystems, the passive photonic AWG-based multiplexing and ring-resonator-based filtering circuitry can be also replaced by PhC resonant structures [3.35] so as to reduce physical size dimensions. These devices exhibit very small footprint in the order of 7x7 μm² per filtering element and can be configured in cascaded stages to function as multiplexers/demultiplexers.

Cascaded placement of the PhC resonant structures can lead to the implementation of the all passive Row decoder, as can be noticed in Fig. 3.13 (e). Moreover, PhC resonant structures can be combined with Fano resonance PhC nanocavity switches to form the Column Decoder and Tag Comparator structures, as seen in Fig. 3.13 (b) and 3.13 (d), respectively. Taking also into account the area occupied by the bending radii and optical interconnection of the devices, the scaled optical cache architecture could reach an approximate footprint value of 7.89 mm² for a storage capacity of 8 KB. Moreover, as state-of-the-art PhC nanocavity components require almost 30nW bias power [3.33], the total energy consumption should not exceed approximately 142.6 fJ/bit period for storing 8 KB.
In order to perform a direct comparison between the 8 kB optical cache and existing electronic cache solutions, the Energy/bit period and footprint requirements of an 8 KB 2GHz electronic cache for 32, 45 and 68 nm processing technology have been calculated by using the CACTI 6.5 cache power estimation tool [3.37]. For the scaled 8 KB 16GHz optical cache, two different cases have been taken into account regarding the technology used for the active elements. In the first case, microdisk lasers were used for all switches, Access Gates [3.32] and optical FFs [3.27]. In the second case, optical Fano resonance PhC nanocavities [3.34] were used for the switches and Access Gates, while the all-optical nanocavity-based memory presented in [3.33] was used for the optical FFs. Nanocavity-based filtering structures [3.35] were used in both cases for the Row and Column decoding peripherals. However, due to their passive nature, they were only taken into consideration for footprint calculations. The obtained results are summarized in Table 3.3. As can be seen, significant reduction in the required Energy/bit period can be achieved through the utilization of the PhC nanocavity technology. It should be noted that in the case of optical cache technology the power consumption does not increase with clock speed. So, the energy can be reduced by a factor of 2.5 if, for example, the all-optical nanocavity-based memory operates for data of 40 Gb/s as reported in [3.33]. The footprint required for the integrated implementation of both cases of optical cache is still higher than the respective electronic cache footprint but still within acceptable limits towards getting attached close to the processor chips. Moreover, this footprint does not come any more at the cost of processor chip real-estate, since the optical cache can be located off-processor-chip and get interconnected via optical board waveguide technology to the core as it is being presented in the next Chapter (Chapter 4).

Considering an optical cache chip topology as shown in Fig. 3.13, the routes that the input and output signals originating from the CPU during Write mode operation and from the optical FFs during Read mode operation, respectively, have been drawn so as to allow for corresponding latency time calculations. The Line, Tag and Data interfaces

<table>
<thead>
<tr>
<th>Technology</th>
<th>2GHz Electronic</th>
<th>16GHz Optical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>68nm</td>
<td>45nm</td>
</tr>
<tr>
<td>Footprint (mm²)</td>
<td>0.13</td>
<td>0.057</td>
</tr>
</tbody>
</table>

*For a cache size of 8 KB*
together with the Tag Comparator and Row Decoders have been placed in the center of the circuit, so as to reduce the maximum data traveling distance corresponding to the most distant optical FFs. Thus, X stands for the horizontal distance between the 1st and the 33rd column, also equal with the distance between the 34th and 66th column. Also, Y is denoted as the vertical distance between the 1st and 64th row being identical to the distance between the 65th and 128th row. To calculate the circuit latency, the longest-path scenario corresponding to data being stored at the four circuit edges has been selected. Line bits have to travel a distance of Y to reach Row 128 and unlock it and then travel another X, so as to reach the optical FFs of the first column. By placing the DATA1 interface vertically confined with the optical FFs of Row1, the input data will have to travel a distance of Y to be stored in (Row 128, Column 1). Thus, the worst case scenario latency for Write operation accounts for the time that the Line bits need to travel a distance of X+Y. For Read operation, Line bits will again have to travel a distance of X+Y to unlock the furthest data Word, as in the case of Write operation. Consequently, the Data bits will need to travel another Y to reach the corresponding Data-to-Read interface and Read-Selector gate. In total, a latency corresponding to a traveling distance of 2xY+X will be imposed for the Data bits to be unlocked and led to the Memory Controller. At the same time, Line bits will only need to travel a distance of Y to unlock the necessary row. Then, Tag bits will emerge from the optical FFs and travel a distance of Y to reach the Tag Comparator. After the comparison is made, its product will have to travel a distance of X to reach the corresponding Data-to-Read interface and Read-Selector gate, controlling propagation of the Data signal to the Memory Controller. Thus, the latency imposed from the Tag bit propagation also corresponds to a traveling distance of 2xY+X. So, the maximum latency for Read operation corresponds to a traveling distance of 2xY+X. Taking into account the approximate dimensions of the discrete subsystems depicted in Fig. 3.13 as well as the area required for the optical interconnection and bending radii, X reaches a value of 1504.8 μm while Y is approximately 1280 μm. Considering that Si waveguides have a refractive index of around 2.5 [3.38], the latency for Write and Read operation accounts for ~23.2 ps and ~33.7 ps, respectively. Even if we take into account the 44ps writing speed of the PhC nanocavity memory [3.33], this corresponds to less than two memory cycles at a 16 GHz clock operation and is significantly smaller than electronic cache latency values [3.39].

The optical bandwidth of the proposed cache layout is calculated as the aggregate data rate that can be transferred within a single clock cycle to or from the cache unit.
Considering a 64-bit wide data bus, as usually employed in conventional CMP configurations, and WDM-formatted optical words, a total data rate of 8x16=128Gb/s is propagating through every bus wire suggesting a total throughput of 64x128=8.192 Tb/s for the entire data bus. Assuming the same data bus width in respective electronic cache configurations operating at 2GHz, the bandwidth of the electrical cache does not exceed 128Gb/s indicating a 64x improvement in the case of optical cache units that originates both from its high data-rate carrying capabilities as well as by its WDM-enabling characteristics. The low latency of the scaled optical cache architecture, combined with its low power consumption and high optical bandwidth, renders the proposed optical cache scheme a promising solution for employment in future CMP architectures.

3.5 Expanding the Optical Cache Memory Architecture from Direct to 2-Way Associative Cache Mapping Scheme

In contrast to the direct cache mapping scheme, the 2-way approach relies on dividing the cache memory module into a number of identical sets, each one comprising two discrete buffering locations termed as WAY#1 and WAY#2. Fig. 3.14 shows how each MM block is mapped to a certain set for an indicative 4-set cache memory module. As it is the case with the proposed optical cache that follows the direct cache mapping scheme, in the optical implementation of the 2-way optical cache memory setup, high-speed CPU-cache memory interconnection via optical waveguide solutions can be achieved through the implementation of proper interfaces at the CPU and cache memory.

Fig. 3.14: 2-Way set associative cache mapping scheme
Fig. 3.15: (a) Optical cache memory architecture. (b) Flowchart of the proposed optical cache operation

Asides as it has been presented in detail in section 3.2.1.

Fig. 3.15 (a) shows the layout for the proposed cache memory architecture that follows the 2-way cache mapping scheme. As can be seen it is organized in four sets (each cache row defines a separate set) of two buffering locations (WAY#1 and WAY#2) and in contrast to the direct mapping scheme it also employs the Way Selector (WS) which is used in order to control access between the two possible buffering locations. Similarly to the direct cache mapping approach, the Read/Write Selector (RWS) comprises three identical Write Access Gates (WAGs), all of them being controlled by the Read/Write (RW) signal. WAG1 receives as input the Tag signal while the other two WAGs receive as input the two words transmitted from the Data-to-Write (DTW) interface. The presence of a RW=0 signal suggests Write operation and the incoming signals exit the WAGs and get forwarded to the WS. If RW=1, Read operation is performed and both Tag and Data signals are blocked from reaching the WS. The WS comprises three identical switches controlled by the Select Way (SW) signal. In the presence of a SW=0 signal, both the Tag and Data signals are routed to WAY#1 in the Row Address Selector- (RAS-) activated set, otherwise they are routed to WAY#2. The RAS receives as input two Line bits, as required for a four-line cache memory, and is responsible for activating one of the four possible output rows. Each RAS row utilizes two cascaded band-rejection filters centered at two of the four incoming wavelengths $\lambda_1, \lambda_2, \bar{\lambda}_1$ and $\bar{\lambda}_2$, forming a $\lambda$-selective matrix as it has already described in section 3.2.3. The RAS output signals are used to drive the row AGs, namely AG1, AG2 and AG3. In the case of an activated set, the incoming RAS signals will equal “0”, allowing the propagation of Tag
and Data signals to the corresponding 2D optical RAM bank’s (2DRB’s) FFs (FFs). In the case of a non-activated set, the non-rejected RAS signals will block access to the corresponding FFs. The 2DRB comprises four identical sets organized in four lines. Each line can be split into two WAYs of three discrete storage areas: One 5-bit area for the Tag information and two 8-bit areas that form together a block buffering area. Each storage area comprises an AG followed by an AWG-based Column Decoder (CD) leading to the all-optical FFs. Finally, the Tag Comparator (TC) is utilized in Read operation to compare the Tag bits emerging from the RAS-activated set with the Tag bits of the processor’s requested memory address. For each WAY, both streams are demultiplexed through two AWGs and the comparison is performed through 5 cascaded XOR gates. The XOR results are multiplexed through an AWG per WAY. These WDM-formatted results form the COMP#1 and COMP#2 signals which are fed into the TC’s RAGs along with the \( \overline{\text{rw}} \) signal. The presence of a logical value “0” in either of two COMP signals indicates that the requested data are buffered in the cache memory and if \( \overline{\text{rw}} = 0 \), suggesting Read operation, a cache hit takes place and the data words exit the RAGs. In all other cases, the COMP signals will form two multi-level WDM-formatted signals which will block the Data signals from reaching the Data-to-Read (DTR) interface, also informing the Memory Controller for the miss in order to fetch the data from the MM. The principle of operation behind the proposed 2-way associative optical cache approach is summarized in the flowchart of Fig. 3.15 (b). Similar simulation scenarios to the ones presented in section 3.3 revealed that in comparison to the direct cache mapping approach the addition of the WS subsystem does not affect the cache memory operation and results of similar quality were obtained for all possible data and address combinations indicating successful functionality in both WAYs and at all the four cache memory rows.

3.6 Conclusion

This chapter demonstrated a 16 GHz physical layer optical cache memory architecture for direct and 2-way cache mapping associativity. The proposed optical cache architecture takes advantage of WDM principles by using WDM-formatted words and address fields, thus significantly relaxing the requirements in terms of active components and power consumption. The chapter presents all the necessary subsystem layouts that are needed for correct cache memory operation: (a) an optical interface module that is used for translating the wavelength-formatted data and address fields between the
processor and the cache memory unit, (b) a Read/Write Selector for ensuring correct access to the stored data, (c) a completely passive row decoder for addressing the different storing locations of the cache memory, (d) a 2D optical RAM bank for storing the actual data and (e) a tag-comparison circuit used for determining whether the desired data are stored in the cache memory unit. Moreover, all functional subsystems comprise experimentally proven optical building blocks adding to the feasibility of the setup. Successful Read and Write functionality at 16 Gb/s is demonstrated via physical layer simulations. The proposed optical cache architecture provides a promising system solution for bridging the gap between optically connected CPU-cache schemes and high-speed optical RAM.

3.7 List of References


3.16. K. Lee et al., “10Gb/s silicon modulator based on bulk-silicon platform for DRAM optical interface,” Optical Fiber Communication Conference (OFC) 2011, JThA33, 6-10 March 2011, Los Angeles, CA, USA.


Chapter 4

A Chip-Multiprocessor Architecture Optically Connected to a High-Speed Optical Cache Memory: Performance Analysis Using the Gem5 Simulator

4.1 Introduction

The performance gap between the CPU and Main Memory, commonly referred to as “Memory Wall”, has been identified since twenty years ago as the main barrier against increases in the computer performance. With the introduction of CMPs, the need for memory bandwidth further increased leading to the widening of the CPU-MM buses and the use of hierarchical on-chip cache memories that often occupy more than 40% of the chip area [4.1]-[4.3]. As the “Memory Wall” is still present, new solutions are needed to bridge the gap. Optical interconnects and photonic integration technology fields emerge as such, thanks to their proven high-speed data transfer abilities. The benefits of replacing electrical CPU-MM buses with optical waveguides have been outlined in several [4.4] while the 16 GHz optical cache memory physical layer architecture presented in the previous chapter, advocates the replacement of the entire CPU-MM electronic circuitry with optical buses and optical caches, freeing up valuable space in favor of the processing elements and thus improved CPU chip area utilization while offering ultra-high speeds at the memory layer.

This chapter presents an optical bus-based CMP architecture where the processing cores share an optical single-level cache. Modern CMPs typically employ Level-1 (L1) caches dedicated to each core, and Level-2 (L2) and sometimes Level-3 (L3) caches shared among the cores. In contrast to this paradigm, this chapter presents a shared-L1 CMP architecture on the basis that the optical cache memory can operate significantly faster than the electrical cores; it thus can theoretically be shared among the cores without becoming a bottleneck, serving concurrently multiple requests. The proposed optical cache CMP architecture is, hence, a single-level shared L1, which sidesteps the issue of data consistency in the conventional paradigm of dedicated per core L1 caches [4.5]. The optical cache is implemented in a separate chip next to the CPU die.
The interconnection system is realized through WDM optical interfaces connecting the shared cache with the cores and the Main-Memory via spatial-multiplexed waveguides. Evaluating the proposed approach, system-level simulations of a wide range parallel workloads using Gem5 simulator are realized. The proposed Optical cache-enabled CMP architecture is compared against the conventional one that uses dedicated on-chip L1 electronic caches and a shared L2 cache. The simulation results show significant L1 miss rate reduction of up to 96% for certain cases; on average, a performance speed-up of 19.4% or cache capacity requirements reduction of ~63% is attained. Combined with high-bandwidth CPU-DRAM bus solutions based on optical interconnects, the proposed design is a promising architecture bridging the gap between high-speed optically connected CPU-DRAM schemes and high-speed optical memory technologies.

Chapter 4 is organized as follows: Section 4.2 describes the detailed physical layer architecture of the bus-based CMP architecture, section 4.3 presents the system-scale simulation results, section 4.4 recapitulates the findings of this work and finally section 4.5 concludes the chapter.

### 4.2 Optical-bus-based Chip-Multiprocessor Architecture with Optical Cache Memories

Fig. 4.1 (a) presents a typical example of a modern CMP with multi-level electronic caches and an indicative number of eight processing cores. Specifically, the standard approach is to put dedicated L1 data (L1d) and L1 instruction (L1i) caches at each core that run at the same speed with the core in order to maintain stall-free core operation assuming cache hits. L1d and L1i caches independently buffer the instruction and data fetch and store operations towards doubling the cache bandwidth and reducing interference between the instruction and data streams. Reducing the interference between the instruction and data streams can improve the overall system’s performance and is the standard approach followed by most current CMP systems [4.5]. Going down through the memory hierarchy, a second unified Level-2 (L2) cache stores both instructions and data and depending on the number of cores and the target application, Level-3 (L3) caches may be eventually also employed and shared among the processing cores. Last, the MM connects to the CPU chip with a spatially multiplexed electrical bus. Although L2 and L3 are slower than L1, they are much faster to access than MM and, typically much larger in size than L1, diminishing thus the penalty of an L1 miss.
In contrast to the conventional paradigm of Fig. 4.1 (a), this chapter examines the impact of using optical cache memory technology as single-level shared cache between the processing cores and the MM. The optical cache memory technology has relied on the hardware design presented in the previous Chapter (Chapter 3) and being capable of operating at 16GHz line-rate speeds. In the proposed CMP architecture, the shared L1 cache is an optical cache memory technology, connected to CPU and MM via spatial multiplexed optical waveguides. The direct sharing of the cache among the cores does not necessarily stall the core operation as the optical cache operates at significantly higher speeds compared to the electronic cores, managing to serve multiple concurrent requests from many cores during each electronic core cycle.

Fig. 4.1 (b) depicts the layout of the proposed optical-bus-based CMP architecture comprising three discrete subsystems: (i) the cache-free CMP chip (8 cores are shown as in Fig. 4.1 (a)), (ii) the optical cache chip with separate L1i and L1d caches lying next to the CMP chip, and (iii) the MM module. The interconnection system between the three subsystems consists of three optical buses with proper WDM optical interfaces at the edge of the CPU cores and the MM. Note that optical to electronic conversion is not required at the cache memory’s sides as the optical cache memory operates completely in the optical domain. Section 4.2.1 presents all the details regarding the WDM optical interface technologies.

The short access time of the optical cache memory layer (originating from its higher clock speed) combined with a proper Time-Division-Multiplex (TDM)-based access
scheme (explained in detail in section 4.2.1 along with the WDM optical interface technologies) can sidestep any bottleneck phenomena arising from the aggregation of the multiple memory requests from the different cores to the single shared cache. At the same time, the shared buffering approach manages to eliminate any coherency issues faced by the multiple discrete caches extensively used in conventional CMPs (Fig 4.1 (a)), as data is cached uniquely in a common shared structure (Fig. 4.1 (b)).

In the proposed architecture of Fig. 4.1 (b) there are three optical buses where the first two connect the CPU cores with the L1i and L1d caches and the third one connects the cache chip with the MM. It should be noted that in contrast to current CMPs based on electronic caches, in the proposed design there not any caches residing in the CPU chip. Placing the cache memories at a separate chip can be advantageous allowing for better utilization of the area of the CPU chip in favor of the actual processing elements. This would translate into either more cores on the same die or smaller processor dies.

Each of the three optical buses consists of multiple communication layers where the number of layers depends on the operations carried out by the attached subsystems. The CPU-L1d bus is used whenever a CPU core wants to access data in L1d cache and consists of the three discrete layers depicted in Fig. 4.1(c): the Address, the Data-Write and the Data-Read Layer. The Address Layer is always used along with one of the other two layers and especially with the Data-Write Layer in the case of a write operation or the Data-Read Layer in the case of a read operation. In both cases, the Address Layer forwards the address of the data from the CPU core to L1d. In the case of a write operation the Data-Write Layer forwards the transmitted data from the core to L1d, while in the case of a read operation, the Data-Read layer forwards the data from L1d to the proper CPU core. In contrast to the CPU-L1d bus, the CPU-L1i bus is used to load instructions from the L1i cache to the core. As there is no need to write in the L1i cache (instructions are always requested by CPU cores and never written back to the cache [4.5]), only the Address and the Data-Read layers are implemented in the CPU-L1i bus and operate similarly to the CPU-L1d bus. Finally, the L1-MM bus consists of all the three layers, similar to the CPU-L1d bus, in order to allow for both read and write operations between the optical chip’s caches and the MM.

Considering a 64-bit system architecture, all the Data-Write and Data-Read Layers must be implemented 64-bit wide. The width of the Address Layer depends on the desirable address memory space. For example, a system with 32-bit Address Layer width can successfully address $2^{32}$ memory locations and assuming 1 byte per memory location
results in 4 GByte addressable memory space. In contrast to conventional electronic buses where \( n \) parallel wire lanes are used to implement \( n \)-bit wide buses, wavelength multiplexing permits to send multiple bits over a single waveguide. The proposed architecture employs 16 wavelengths per waveguide able to transfer 8 bits of information, given that each bit requires a pair of wavelengths for encoding the actual bit value on one wavelength and its complement on a second wavelength, following the physical layer cache architecture specifications presented in Chapter 3. To this end, the transmission of these 16 wavelengths in every optical waveguide suggests that every waveguide can carry now 1 byte instead of only 1 bit, as is the case in electronic buses where no wavelength or frequency multiplexing capabilities are supported. As such, a 64-bit wide bus can be here implemented by employing 8 optical waveguides for the Data-Read and Data-Write Layers, while a 32-bit wide Address Layer bus requires only 4 waveguides. It should be noted that the same 16 wavelengths can be reused in each parallel waveguide.

Fig. 4.1(d) illustrates in more detail the different lanes within the Data-Read Layer of the CPU-L1d bus, where 8 waveguides (with each one carrying 1 byte of information encoded on 16 wavelengths) form the desired 64-bit width. A simple connection scheme between the bus waveguides and every individual core is considered by incorporating an optical coupler that drops part of the data power level to its attached core and lets the remaining power continue its propagation in the bus. High-density optical waveguides can be easily deployed by conventional Silicon-on-Insulator (SOI) technology or by single mode polymer waveguides that have been recently presented to offer as low as 0.6 dB/cm propagation losses and a total density of 50 wires / cm [4.6].

### 4.2.1 Optical Interfaces

One of the most promising approaches for implementing on-chip optical interfaces has been presented in [4.4] where external laser sources combined with SOI waveguides and ring-resonator-based filters and modulators operating at 10 Gb/s carry out the necessary opto-electronic and electro-optical conversions between the CPU cores and the optical interconnection system.
Fig. 4.2: Optical interfaces utilized in CPU cores and MM: (a) WDM Optical Transmitter Interface (b) WDM Optical Receiver Interface. Both Transmitter and Receiver interfaces are followed by operation examples presenting the TDM-based bus/cache access scheme. The examples assume 4 processing cores and a 4x optical system operation speed (compared to electronic cores).

Fig. 4.2 (a) illustrates the optical interface’s WDM transmitter module, which is placed on the same die with the processing cores. External laser sources provide a stream of Continuous Wave (CW) signals at 16 different wavelengths, which is coupled into the on-chip waveguide and routed through a cascade of resonant ring modulators driven by proper drivers connected to the CPU core’s registers. Each ring modulator is tuned to a different wavelength and is capable of modulating the CW signal at that specific wavelength. Every bit-register’s driver drives two ring modulators in order to encode both the register content as well as its complementary value on two discrete wavelengths coming in accordance with the physical layer architecture specifications presented in Chapter 3. In order to enable a basic Time Division Multiplexing (TDM)-based bus access scheme among the cores, a proper Clock Generator (CG) circuit distributes periodical Return-to-Zero clock pulses with 1:N duty cycle at 2 GHz line-rate to all register drivers. In this way, the duration of every clock pulse equals the duration of a cache clock cycle and performs a gating operation to the register drivers defining the
time slot where the electro-optical conversion of the register content will take place. Each individual core is here shown to be equipped with a separate CG circuit that is synchronized with a different cache cycle time-slot, but also a single common CG element could be eventually used distributing its clock pulses among the different core interface subsystems.

After having all 16 wavelengths modulated, the WDM signal travels through the waveguide to the optical cache memory module. A total number of eight identical transmitter modules is required at every core interface to implement the complete 64-bit interface of the optical buses, given that each waveguide carries 8 bits of information. An operation example of the TDM-based bus access scheme is presented below the WDM Transmitter Interface. This example assumes a CMP configuration of 4 processing cores combined with an optical bus and an optical cache subsystem that both operate at 4X higher clock line-rate speeds when compared to the electronic cores. As can be seen in the example, 1 CPU core cycle corresponds to 4 optical-cache cycles where the 1st CPU core transmits its data during the 1st cache cycle, the 2nd CPU core during the 2nd cache cycle, the 3rd CPU core during the 3rd cache cycle and the 4th CPU core during the 4th cache cycle. When one core completes a single transmission of its data, a WDM signal is formed occupying a time-slot that equals the period of the cache clock cycle. When all cores complete a data transmission, the resulting signal comprises four time-multiplexed WDM signals that extend along an entire CPU clock cycle. Fig. 4.2 illustrates the time traces only for the first two registers of the core WDM interface for facilitating readiness of the picture.

Fig. 4.2 (b) illustrates the optical interface for the respective WDM receiver module. The WDM-formatted signal coming from the optical cache memory module enters a series of Add/Drop electro-optical ring resonator filters, with each filter being tuned to and dropping a different wavelength. The dropped wavelength enters a photo-detector module where it gets converted into electric current and then gets stored to the corresponding bit-register. As in the case of the WDM transmitter interface module and in order to enable a basic TDM-based bus access scheme among the cores, a proper CG circuit distributes periodical clock pulses to all ring filters indicating whether they should read the data transmitted in the bus or not. It should be noted here that the same CG circuit can be used for both the WDM transmitter and receiver interfaces per individual core as long as they are both synchronized with the time-slot allocated for this specific core. Similarly to the transmitter module, eight identical 8-bit receiver modules are
necessary at every core’s interface to implement the 64-bit interface. Fig. 4.2(b) depicts below the WDM receiver interface an operation example of the TDM-based bus access scheme. The operation example assumes the same configuration of 4 cores preserving also the same bus access order as was used in the respective transmission example, allowing in this way for contention free cache-to-core communication. Again, only the traces corresponding to the first two registers of every core are presented, while the initial content of the registers has been assumed to equal a logical ‘0’.

The complete path between the CPU core and the MM transmitter/receiver optical interfaces relies exclusively on the use of optical subsystems for all necessary transmission, processing and storage functions without incorporating any electronics.

An alternative solution for the optical interfaces has already been presented in the previous Chapter (Chapter 3) along with the presentation of the optical cache physical layer architecture. According to that solution VCSEL-array sources and photodiode arrays are placed on the same die with the CPU cores. Regarding the optical interface’s transmitter module, each VCSEL emits at a different wavelength and is driven by its corresponding bit-register. The VCSEL array outputs are then multiplexed into the waveguide using an Arrayed Waveguide Grading (AWG) multiplexer. Similar approaches have been presented in various demonstrations [4.7]-[4.9], offering modulation speeds of up to 25 Gb/s (currently moving towards 40 Gb/s). Regarding the optical interface’s receiver module, an AWG demultiplexer is used to demultiplex the WDM-formatted signal coming from the optical cache, where the individual wavelengths are finally fed into a photodiode array in order to get converted to electrical form. Photodiode array solutions have already proved sufficient for a number of up to 128 photodiodes [4.7] and for speeds of up to 25 Gb/s per photodiode element [4.8].

4.3 System-level Simulation Results

To assess the performance of the optical cache module presented in the previous Chapter (Chapter 3), this section proceeds with the study of a CMP system that integrates such an optical off-chip cache connected to the CPU and MM via optical buses.

The standard CMP design is based on electronic memory elements and deploys a multi-level cache hierarchy. The most common approach is to use one dedicated L1 cache per core (partitioned in separate L1i and L1d caches) and a common L2 cache shared among all cores (through the dedicated L1 units). In high-end systems, L3 cache
is added between L2 and the MM in order to further reduce the performance penalty of the L2 misses propagating towards the MM. In this study, a two-level cache hierarchy CMP system is considered as depicted at the top of Fig. 4.3 and is termed Conventional L1+L2. A second all electronic cache system, Conventional L1, is also considered where it is a system with individual per core L1 caches, but no L2. Although Conventional L1 does not reflect a realistic CMP, it is used to facilitate the one-to-one direct comparison between dedicated L1 caches per core and the proposed shared L1 optical cache architecture.

The proposed Optical L1 system to be compared against Conventional L1+L2 and Conventional L1 deploys just a single L1 optical cache shared among all the cores and is depicted at the bottom of Fig. 4.3. Such a flat hierarchy is possible thanks to the high speed of the optical memory when compared to the electronic cores. The high clock speed of the optical memory combined with the TDM-based bus/cache access scheme allows for contention-free core-to-cache and cache-to-core communication without stalling their execution when only cache hits are assumed. This is in sharp contrast to the dedicated L1 caches in current CMPs where they cannot efficiently used in shared CMP.

Fig. 4.3: CMP Architectures: Conventional L1+L2 cache (up) and Optical L1 cache (down). The Conventional L1 (not shown) is the same with the Conventional L1+L2 but without the L2 module.
architectures serving multiple cores’ requests without stalling their execution. At the other end, the shared electronic L2 has a considerable latency in the order of tens of cycles and at top of that it can easily become a bottleneck if multiple L1 caches miss concurrently, an event not infrequent in demanding workloads that stress the front line dedicated L1 units. Compared to the slow electronic cores, the optical cache is possible to share efficiently among the cores, eliminating thus the need for dedicated units and the complexity of multi-level organization. What is more, the flat hierarchy manages to get rid of any data consistency issues that arise in the conventional multi-level cache architectures where the same piece of data is cached in different cache units and gets updated. As such, the proposed CMP architecture eliminates the need for complex coherency protocols reducing at the same time the core-to-cache and cache-to-core communication to the minimum number of stages, since only one copy of the data resides now in the single shared cache and there is no need for updating any invalidated copies distributed among the memory hierarchy in a multi-level cache hierarchical scheme.

### 4.3.1 CMP Cache Hierarchies

As shown in Fig. 4.3 and for both Conventional and Optical L1 architectures L1 is divided in discrete instruction (L1i) and data (L1d) caches. Each of them is separately connected via an electronic link or optical bus to the cores, doubling in this way the bandwidth between the core and L1 and permitting the core to simultaneously access an instruction and a data word [4.5]. In the simple case of a single core (N = 1) and read operation, when the requested data are found in L1, an L1 cache hit occurs and the core is served immediately without delay. Otherwise, the request propagates to the next level of the memory hierarchy, i.e.,

(i) the unified (stores both instructions and data) and shared L2 cache via an electronic bus in the Conventional L1+L2 cache architecture, or

(ii) the MM via an optical or electronic link in the optical cache and Conventional L1 cache architectures, respectively.

In the Conventional L1+L2 cache, if the data is found in L2 (L2 hit), L2 responds with some delay, which is significantly lower than the latency of accessing MM through the electronic bus.

The multi-core case (N > 1) is more complicated, as the multiple cache units (dedicated L1s and shared L2) in the Conventional systems provide the opportunity for a
referenced address to be stored in multiple places within the cache hierarchy. If not handled properly, some stale copies of the referenced data could possibly result to erroneous computations. As such, a coherency protocol that manages the sharing of the data in the different caches is required.

Cache coherency mechanisms can be classified into Directory Based and Snooping protocols [4.5]. This work focuses on the later which is the typical choice in small to medium scale CMPs. In early CMPs, L1 caches were write-through, hence any memory write had to propagate to the (typically write-back) L2 via the L1-L2 bus. In snooping, each dedicated L1 unit monitors the address bits broadcasted in the L1-L2 address bus and invalidates its cache line if it gets updated at a different core.

Going back to the case of an L1 cache miss, snooping permits the implementation of an additional enhancement towards reducing the L2 access delay. If the missed line can be found in a different L1, the latter L1 can respond the request placed at the L1-L2 bus on behalf of the slower L2. In such designs, L1 units become two ported, accepting and servicing requests from both the respective core and the L1-L2 bus. As expected, the L1-L1 communication over the broadcast bus takes some cycles to complete (and is subject to contention if the bus is busy with other requests). Hence the communication over the L1-L2 bus is slower than the communication of the core to its dedicated L1 unit. However, it is still (and typically much) faster than the access to L2 and can be particularly beneficial in parallel workloads that share and exchange a significant amount of data between their threads.

In this experimentation, Gem5 simulator [4.10] is used, which implements a MOESI [4.11] snooping coherency protocol (a variation of the standard MESI protocol [4.5]) that allows the direct L1 to L1 communication. Gem5 is an open source tool that can simulate accurately a complete CMP. It takes into account the timing at the cores and the memory system, including the cache hierarchy and interconnects. However, it should be noted that Gem5 uses the notion of “express snoops” to propagate instantly any cache miss without contention in the bus; this is an inaccuracy in the timing of the snooping protocol that falls on the optimistic side with respect to the performance of a real system (the Gem5 designers acknowledge it and judge it as having minor effect in comparison with real systems). Last, it should be noted that Gem5
models a non-inclusive cache hierarchy with write-back L1 units, allowing the efficient use of the total cache capacity.\(^1\)

### 4.3.2 Simulation Parameters

In the simulation runs, a CMP comprising \(N\) cores is considered, with \(N\) either equal to 8, or varying from 2 to 16 being always a power of 2, depending on the experiment. Each core runs at 2 GHz, while the MM is set equal to 512 MB and runs at 1 GHz. Regarding the electronic cache, the speed of both L1 and L2 is set to 2 GHz (same as the cores), and the electronic bus from L2 to MM runs at 1 GHz. As for the optical cache, the speed of the Optical L1 is set equal to \(2 \times N\) GHz, where \(N\) is the number of cores; likewise, the buses cores-L1 and L1-MM run at the same speed with the optical cache clock, which should be possible via silicon photonics or single-mode polymer-based optical waveguides. These configurations comply with the assumption that all cores are concurrently served during a single CPU cycle when considering only L1 cache hits (see TDM-based access scheme in section 4.2.1). Although the physical-layer scalability analysis presented in section 3.4 indicates a hardware operation speed of up to 16 GHz, this analysis probes up to 32 GHz in order to allow for covering also the case of \(N = 16\) cores in one of the simulation scenarios as well as to point out the extra benefits that may arise by further boosting the operational speed capabilities of the optical cache hardware.

The architectures are compared against each other with respect to:

(i) their miss rates on the basis of an equal L1 cache size in sections 4.3.3 and 4.3.4, and

(ii) on their execution times on the basis of an equal total cache size (including L2 too in the Conventional L1+L2 case) in section 4.3.5.

All the simulation runs iterate over the powers of 2 for the L1i size with the minimum being equal to 2 KB per core (which results in \(2 \times N\) KB total capacity for the shared optical cache) and the maximum equal to 512 KB in total for all the cores. Hence, for \(N = 8\) cores, the minimum total L1i is equal to 16 KB. As for L1d, it was set equal to twice the size of L1i. Last, in Conventional L1+L2, L2 was set equal to 8 times of L1d, as it is the case with the SPARC T5 processor [4.12].

---

\(^1\) Early CMP systems enforced an inclusive property to facilitate memory writes and coherency. In inclusive cache hierarchies, all L1 cache lines should be presented in L2, hence L2 should be (much) larger than the total L1 to be efficient.
All simulation configuration values along with the related physical layer metrics derived from the optical cache architecture presented in chapter 3 were incorporated in the Gem5 simulator \[4.10] and are listed in Table 4.1. For the needs of this study, a Reduced Instruction Set Computer (RISC) Instruction Set Architecture (ISA) is chosen, and specifically, the Alpha ISA, and the Timing Simple model about the CPU. The Timing Simple CPU is a model of in-order execution that does not set up a pipeline, but “uses timing memory accesses, stalling the CPU on cache accesses and waits for the memory system to respond prior to proceeding” \[4.13\]. Its peak Instructions Per Cycle (IPC) is equal to 1 if the memory system does not impose any delay, i.e., the cache responds on requests within the 500 ps clock of the CPU cycle (where CPU cores run at 2 GHz) and no misses occur. Thanks to its simplicity, it is particularly useful to focus on the stalls in execution and assess the performance of the memory system (cache hierarchy and MM), which sophisticated out-of-order execution could blur. Finally, according to the scalability analysis of section 3.4 the optical cache’s hit latency has been considered to equal 1/N nsec suggesting a latency value of 125 psec for the case of N=8 cores, without having this rounded to the CPU clock cycle.

PARSEC benchmark suite \[4.14\] is chosen to get simulated within the Gem5 simulation framework, in order to assess the performance of the optical cache on top of Linux. PARSEC is an open source suite intended for research in CMPs and consists of 13 multi-threaded programs (listed in Table 4.2) that cover diverse workloads. Table 4.2 shows their characteristics, as described by the designers about: (i) the programming

<table>
<thead>
<tr>
<th>Table 4.1: Simulation Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main Memory</strong></td>
</tr>
<tr>
<td>Size</td>
</tr>
<tr>
<td>Speed</td>
</tr>
<tr>
<td>tCAS, tRCD, tRP</td>
</tr>
<tr>
<td>tRAS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Memory Bus</strong></th>
<th><strong>Cache L1-L2 Bus</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed (GHz)</td>
<td>1 (2xN)</td>
</tr>
<tr>
<td>Width</td>
<td>64 bits</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Cache Parameters</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cache Type</td>
</tr>
<tr>
<td>Total Size (KB)</td>
</tr>
<tr>
<td>Speed (GHz)</td>
</tr>
<tr>
<td>Block Size</td>
</tr>
<tr>
<td>Associativity</td>
</tr>
<tr>
<td>Hit Latency (nsec)</td>
</tr>
<tr>
<td>Replacement Policy</td>
</tr>
</tbody>
</table>

The simulation parameters used in both architectures, conventional and optical cache. **Bold** entries in parenthesis show the values of the optical cache when different from the conventional cache.
model used to parallelize the task, (ii) the size of the working set, (iii) the sharing and (iv) the exchange data properties among the different threads of the parallel program. These properties provide useful intuition about the program behavior and performance. In the simulations that follow, 12 out of 13 PARSEC benchmarks are used (freqmine has been excluded as it has not been possible to simulate and get performance metrics).

### 4.3.3 Bodytrack Miss Rates for Varying Number of Cores

Focusing first on the bodytrack benchmark, a program with high data sharing and medium data exchange properties, Fig. 4.4 presents the miss rates for the L1d and L1i caches of the two Conventional (L1 and L1+L2) and Optical L1 architectures for \( N = 2, 4, 8 \) and 16 cores. Regarding the Conventional cases, the average of the \( N \) dedicated L1 caches’ miss rates is presented which means that the sum of the misses regarding all the dedicated cache units would appear even higher.

The plots reveal clearly that the miss rates of the shared Optical L1 architecture are lower than in the Conventional architectures. This benefit is more proclaimed in the L1d cache than in the L1i cache, and increases with the number of cores \( N \). The two Conventional architectures have similar miss rates, which is intuitively expected as the existence or absence of L2 does not directly affect the L1 misses (it is however important in minimizing the penalty of L1 misses, which is studied in section 4.3.5). Some minor differences in the L1d miss rates between the Conventional cases without and with L2 for the 8-core and 16-core scenario owe probably to the full system mode execution of the simulations, where benchmarks were executed on top of the Linux Operating System (OS) which was running within the Gem5 simulator framework.
What is important to note here is that both the L1d and L1i miss rates of the two Conventional architectures increase with the number of cores $N$, which demonstrates the parallelization saturation: the problem is broken into more pieces as the computing threads increase, raising thus the communication requirements among the cores. In contrast, the Optical L1 shared architecture is only mildly affected given the fact that the shared cache topology allows for efficient inter-thread data sharing and exchange by providing a common cache unit for all threads’ data and instructions.

Comparing more closely the two Conventional against the Optical L1 architecture for $N = 2$ cores, the Optical L1 architecture has always better L1d miss rates. For the first case of 8 KB total L1d, the Optical L1 architecture performs 14% better than the Conventional ones; increasing the cache size leads progressively to 40% and 55% improvement for 128 and 256/512 KB L1d size respectively. In the 1024 KB L1d case, all the three architectures’ miss rates are almost equal to zero. Regarding the L1i miss rate, in the first measurement of 4 KB L1i, the Optical L1 architecture has one fourth of the miss rate of the Conventional ones, while in all the other cases, all miss rates are almost zero.

Examining bigger CMPs of $N = 4$ cores and above, the shared optical L1d miss rate is always better than the other two, except from the run at the lowest cache capacity (L1i equal to 2xN KB). In these cases the Optical L1 shared architecture suffers from high data miss rates due to the cores’ competition on the limited number of cache lines that falls short of the application’s working set. In the cases of $N = 4$ and 8 cores and for L1d sizes of 32-64 and 64 KB, respectively, the optical L1d miss rate is about the two-thirds of the Conventional ones. In the 128 KB L1d case and CMP sizes of $N = 4$, 8 and 16 cores the Optical L1 architecture performs about half the miss rate compared to the Conventional

![Fig. 4.4: The miss rates of L1i and L1d caches for the Conventional L1, Conventional L1+L2 and Optical L1 architecture for varying number of cores N.](image)
ones. Doubling the L1d size to 256 KB leads to miss rate reduction to around the one fourth of the Conventional cases. The performance advantages are further magnified to one tenth and above when considering the two highest CMP sizes (8 and 16 cores) and L1d capacities of 512 and 1024 KB.

Regarding the L1i cache and for small capacities the Conventional architectures suffer from discernible instruction miss rates due to the limited capacity of the dedicated L1i caches. In contrast, in the Optical L1 architecture the instructions used by all the cores manage to fit in the common cache. As shown in the plots, the Optical L1 architecture always has lower miss rates than the Conventional ones. In particular, Optical L1 manages to almost completely eliminate the cache miss ratio.

4.3.4 Miss Rates for All Programs

Although Table 4.2 is useful to intuitively analyze the benchmark behavior in a CMP system, it can only provide an approximate characterization. It can be observed that there do not exist programs that share exactly the same characteristics (except from ferret and dedup). As explained in the extended analysis and comparison of PARSEC suite with SPLASH-2 suite [4.15], each application has been designed and programmed so as to be unique. As such, PARSEC forms a representative suite for a wide range of application domains.

Fixing the number of cores to \( N = 8 \) the miss rates for the 12 benchmarks are plotted in Fig. 4.5. Comparing the miss rates in the two Conventional against the Optical L1 architecture reveals that a class of programs behave similar to bodytrack (canneal, dedup, facesim, ferret, streamcluster, vips and on the borderline x264) and have L1 miss rates equal to a fraction of the respective values in the Conventional architectures (Optical L1 architecture has a proclaimed advantage in the facesim’s L1i misses). In a second class of programs (blackholes, fluidanimate, rtview and swaptions), the Optical L1 architecture has worse L1d miss rate performance for small cache sizes and small or no improvement for higher cache sizes where the miss rates are close to zero in all the three compared architectures.
Focusing on the benchmarks of the first class, *dedup*, *ferret* and *x264* are based on the pipelined parallelization model and all the three share the common high data sharing and high data exchange properties. Although *dedup* and *ferret* exhibit identical characteristics in Table 4.2, for both *Conventional* architectures *ferret* has higher miss rates than *dedup* which translates to increased gains for the shared optical topology. Regarding *x264*, it differs from the other two in terms of working set and parallelization granularity, featuring a medium working set and coarse parallelization granularity (vs. unbounded and medium respectively). Comparing the *Optical L1* architecture against the *Conventional* ones, *ferret* presents a 48-66% L1d miss rate improvement within the tested cache size range, while the respective miss rate improvement for *dedup* and *x264* starts from 11% and scales up to 37% and 54%, respectively. Regarding the L1i miss rates and for all the three benchmarks, *Optical L1* architecture outperforms both *Conventional* schemes within the complete tested cache size range.
Canneal is the only benchmark of PARSEC that is based on the Unstructured parallelization model and has both high data sharing and high data exchange properties. For small L1d sizes Optical L1 architecture achieves an about 30% lower miss rate compared to the two Conventional architectures; in the high values of cache capacity, the miss rate advantages are in the range of 20-25%. Taking into account the unbounded working set and the Unstructured parallelization model of canneal can intuitively explain why, in all architectures, canneal has the highest L1d miss rates among all the benchmarks. Regarding the L1i miss rates the Optical L1 shared architecture manages to fit the instruction set used by all the cores and even for small cache sizes, outperforming both Conventional architectures.

Bodytrack, streamcluster, vips and facesim follow the Data-Parallel parallelization model and complete the class of programs that perform better miss rates for the Optical L1 architecture. Bodytrack is a program with high data sharing and medium data exchange properties and the benefits gained by utilizing the Optical L1 shared architecture have been analyzed in the previous section. All three streamcluster, vips and facesim benchmarks combine low data sharing with medium data exchange properties but differ in terms of working set and parallelization granularity. Streamcluster has medium parallelization granularity and medium working set, and performs 18-96% better in the L1d miss rate in the Optical L1 architecture when compared to both Conventional ones. Vips has coarse parallelization granularity and a medium working set, and shows an almost constant gain for the Optical L1 architecture of 20-27% for all the tested L1d sizes. Facesim’s coarse parallelization granularity combined with large working set results in benefits of 22-31% for the Optical L1 architecture in the range of 64-256 KB. Regarding the L1i miss rates and for all the three streamcluster, vips and facesim benchmarks Optical L1 architecture manages to fit well the instruction set utilized by all the cores, outperforming thus even for small cache sizes both Conventional architectures.

Blackscholes, fluidanimate, rtview and swaptions are based also on the Data-Parallel parallelization model and have low data usage exchange and low data usage sharing requirements (except fluidanimate’s medium data exchange and rtview’s high data sharing). In small L1d sizes, Optical L1 architecture has worse L1d miss rate than both Conventional architectures. Hence, it seems that these applications of low communication requirements benefit from dedicated L1 cache topologies that keep isolated the threads’ individual data streams: a shared cache of limited capacity induces a high inter-thread competition for the insufficient number of cache lines. However, increasing the cache
size mitigates the inter-thread competition in the Optical L1 architecture, leading to miss rates close to zero for all the three architectures with no variations observed among them. As for the L1i miss rate performance, both Conventional architectures suffer from high miss rates at low cache capacities, as the small and dedicated cache fragments cannot efficiently fit each thread’s instruction set. Similarly to all benchmarks reported so far, Optical L1 architecture avoids such problems as the pooling of the dedicated fragments capacity to a shared cache unit suffices to hold the instruction set used by all cores.

Summing up the above, it is observed that the single shared optical cache operates better in terms of L1d miss rates in about two thirds of the benchmarks, and almost the same (or worse for small cache sizes) in the other one third. Not surprisingly, the programs of the first class share the common characterization of high or medium data usage exchange in Table 4.2 (in a fifty-fifty ratio). Additionally, more than half of these programs (five out of eight) share the high usage data sharing property. The latter two agree with the performance advantages observed for the shared cache topology. Parallel programs of high data sharing and exchange needs among their threads benefit from the shared cache architecture; the high volumes of data exchange increase the traffic and consequently the miss rate among the dedicated L1d caches. In contrast, programs of generally low data sharing and exchange among their threads fill the dedicated L1d cache with each thread’s working data and have little to gain from the cache pooling. Finally, regarding the L1i miss rates both class of programs benefit from the shared cache architecture as even for small cache sizes it always performs close to zero miss rates, exceeding in performance both Conventional architectures.

### 4.3.5 Execution Times for All Programs

Moving to the performance metric of execution time, Fig. 4.6 presents the simulation results for the 12 PARSEC benchmarks and for \( N = 8 \) cores. In contrast to the previous analysis where the Optical L1 was compared to the Conventional architectures on the basis of equal L1 sizes (ignoring the size of L2 for the Conventional L1+L2 case), the comparison here is carried out on the basis of the system’s total cache size. For both Conventional architectures and in order to approach the realistic systems’ setups two system configurations are tested:

1. **System A** corresponds to a system with 16 KB L1d per core, as it is the case with the Sparc T5 processor [4.12] and
(2) *System B* corresponds to a system with the 32 KB L1d per core, double of *System A*, as it is the case with IBM’s Power7 processor [4.16].

Thus, two points are plotted for each of the *Conventional* architectures (*L1* and *L1+L2*) with the left corresponding to *System A* and the right to *System B*. For the *Optical L1* architecture all the power of 2 L1i sizes are tested within the range of 16-512 KB.

The first to note from Fig. 4.6 is that the *Optical L1* architecture in small cache sizes results in high execution times for all the applications, which agrees with the miss rates analysis presented above. Focusing on the applications from the first class of the classification of section 4.3.4, it is observed that for equal cache sizes, *Optical L1* shared architecture achieves better execution times than both *Conventional L1* configurations (except for *canneal* in *System B*). Moreover, by varying the total cache capacity from small to high values, the execution time is significantly reduced. Comparing the *Optical L1* against the *Conventional L1+L2* architectures, the former has better system performance (except from *canneal* in *System A*). As for the second class of programs that exhibited similar miss rates across the three architectures (presented in section 4.3.4), the execution time for the *Optical L1* architecture is in most cases similar to the *Conventional* ones except for *fuidanimate* in both *Conventional L1* systems and *Conventional L1+L2 System A* and in *rtview*, the latter exhibits a highly irregular behavior in the *Optical L1* architecture as it performs strong variations in its execution time as cache size increases.
Table 4.3: L2 Miss rates for the all the 12 PARSEC benchmarks

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>System A (%)</th>
<th>System B (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>blackscholes</td>
<td>0.06</td>
<td>0.19</td>
</tr>
<tr>
<td>swaptions</td>
<td>0.18</td>
<td>0.26</td>
</tr>
<tr>
<td>streamcluster</td>
<td>0.32</td>
<td>0.10</td>
</tr>
<tr>
<td>bodytrack</td>
<td>1.88</td>
<td>1.10</td>
</tr>
<tr>
<td>x264</td>
<td>2.45</td>
<td>2.88</td>
</tr>
<tr>
<td>ferret</td>
<td>9.16</td>
<td>7.70</td>
</tr>
<tr>
<td>rtview</td>
<td>10.51</td>
<td>20.96</td>
</tr>
<tr>
<td>dedup</td>
<td>11.89</td>
<td>6.77</td>
</tr>
<tr>
<td>fluidanimate</td>
<td>12.48</td>
<td>35.35</td>
</tr>
<tr>
<td>facesim</td>
<td>20.93</td>
<td>29.17</td>
</tr>
<tr>
<td>vips</td>
<td>29.82</td>
<td>32.65</td>
</tr>
<tr>
<td>canneal</td>
<td>47.71</td>
<td>46.36</td>
</tr>
</tbody>
</table>

Table 4.4: Performance gain for the Optical architecture compared to the Conventional L1+L2

<table>
<thead>
<tr>
<th>Program</th>
<th>Capacity Reduction (%) over System A *</th>
<th>Performance Speed-up (%) over System A *</th>
<th>Capacity Reduction (%) over System B **</th>
<th>Performance Speed-up (%) over System B **</th>
</tr>
</thead>
<tbody>
<tr>
<td>blackscholes</td>
<td>68.4</td>
<td>7.4</td>
<td>84.2</td>
<td>-2</td>
</tr>
<tr>
<td>bodytrack</td>
<td>68.4</td>
<td>17.0</td>
<td>68.4</td>
<td>10.4</td>
</tr>
<tr>
<td>canneal</td>
<td>N/A</td>
<td>-12.0</td>
<td>92.1</td>
<td>5.9</td>
</tr>
<tr>
<td>dedup</td>
<td>36.8</td>
<td>14.3</td>
<td>36.8</td>
<td>9.3</td>
</tr>
<tr>
<td>facesim</td>
<td>84.2</td>
<td>23.8</td>
<td>92.1</td>
<td>16.1</td>
</tr>
<tr>
<td>ferret</td>
<td>36.8</td>
<td>14.0</td>
<td>36.8</td>
<td>4.4</td>
</tr>
<tr>
<td>fluidanimate</td>
<td>68.4</td>
<td>20.5</td>
<td>36.8</td>
<td>-2.8</td>
</tr>
<tr>
<td>rtview</td>
<td>92.1</td>
<td>29.0</td>
<td>N/A</td>
<td>-36.1</td>
</tr>
<tr>
<td>streamcluster</td>
<td>68.4</td>
<td>34.2</td>
<td>68.4</td>
<td>24.3</td>
</tr>
<tr>
<td>swaptions</td>
<td>36.8</td>
<td>-11.0</td>
<td>68.4</td>
<td>-17.8</td>
</tr>
<tr>
<td>vips</td>
<td>68.4</td>
<td>22.6</td>
<td>36.8</td>
<td>8.8</td>
</tr>
<tr>
<td>x264</td>
<td>68.4</td>
<td>11.3</td>
<td>68.4</td>
<td>6.4</td>
</tr>
<tr>
<td>Mean***</td>
<td>63.37</td>
<td>19.41</td>
<td>62.65</td>
<td>10.7</td>
</tr>
</tbody>
</table>

* System A: Conventional L1+L2 with 16 KB L1d / core
** System B: Conventional L1+L2 with 32 KB L1d / core
*** Negative values of speed-up were not included in the mean calculation

Table 4.3 presents the L2 miss rates for all the benchmarks and for both Systems A and B. The miss rate values refer to the percentage of both L1i and L1d misses that could not be processed by the L2 cache and were forwarded to MM. The results come in accordance with the multi-level caching approach followed by the conventional systems so far due to the fact that for most benchmarks most of L1 misses does not have to wait for the significantly slower MM to respond and are directly processed by the second level of caching.

Table 4.4 summarizes the performance improvements that are possible from the Optical L1 architecture over both Conventional L1+L2 System A and B configurations. The
“Optical L1 vs Conventional L1” comparison is not included in the table, as the Conventional L1 system (without L2) is not representative of current CMPs. Specifically the first and the third columns present the reduction in capacity requirements for same performance, and the second and the fourth columns show the performance gain in execution time over the two Conventional L1+L2 configurations, respectively.

The capacity requirements reduction has been individually computed for each benchmark by (i) finding the two reference Optical L1 systems with the smallest cache sizes whose execution time is almost equal or less than System A and B, respectively and (ii) interpolating to compute the capacity requirements reduction. As shown in the first and third columns of Table 4.4, a high reduction in the required cache capacity can be obtained for all the programs (except from canneal in System A and rtview in System B), regardless whether they belong in the first or second class of programs presented in section 4.3.4. Moreover, for both System A and B almost the same or better execution times can be obtained combining at the same time a mean capacity reduction of ~63%.

Second and fourth columns of Table 4.4 present the performance speed-up over both Conventional L1+L2 configurations by taking as reference point the Optical L1 system with total 1536 KB L1 cache size (1024 + 512 KB of L1d and L1i respectively). Clearly, the Optical L1 system achieves an average speed-up of 19.41% for a total of 10 out of 12 benchmarks when compared against the System A. From the first class of programs, only canneal has a worse execution time in the Optical L1 architecture. Canneal, similar to rtview, exhibits irregular behavior as cache size increases. Regarding the second class of programs three out of four applications improve their execution time and only one, swaptions, performs worse in the Optical L1 architecture. Last, comparing the optical system against System B we observe an average performance speed-up of 10.7% for a total of 8 out of 12 benchmarks combined also with a total capacity requirements reduction of 37%. As expected, the rest of the four applications that not benefit by the shared architecture belong to the second class of programs presented in section 4.3.4.

The above results show that in the first class of programs, the advantage in the miss rates of the Optical L1 architecture against the two Conventional ones carries over to the execution time but to a lower degree. This is a result of the advanced snooping protocol modeled in Gem5 which manages to reduce most of the penalty of an L1 miss to a handful of cycles through the L1 to L1 communication. In conventional systems, where an L1 miss would necessarily be served by L2 at a cost of tens of cycles, the Optical L1 architecture advantage in execution time would be more proclaimed. Thus, the higher
miss rate of the previous analysis results in a modest (and sometimes mild) increase in the total execution time for the Conventional architecture, as the total number of cycles is often dominated by the number of instructions to execute rather than the stalls in the memory system.

One important exception to the above observation is canneal. Although it showed lower miss rates in the Optical L1 in comparison to the Conventional, its execution time is higher or almost equal to the Conventional cases. It should be stressed that canneal exhibits a highly irregular behavior, increasing its execution time in Conventional L1+L2 (only bodytrack in Conventional L1 behaves similarly) and performing strong variations in Optical L1 as cache size increases (rtview in Optical L1 behaves similarly). Although this experimentation does not analyze the source code of canneal, this behavior could be attributed to the combination of the unstructured model with the unbounded working set that creates the possibility for a core to pollute the shared optical cache with unimportant data (with regard to the other cores) and create starvation for cache lines to the other cores hampering the overall parallel program performance. In a dedicated L1 configuration, a thread cannot interfere with the other threads’ L1 lines.

4.4 Discussion

The previous section’s simulation results reveal the potential miss-rate reduction benefits that could be gained by adopting the shared high-speed optical cache topology in multi-core chip configurations. These translated to either significant capacity requirements reduction, or important speed-up in the overall system’s performance.

The transition from traditional computing on CPUs to massively parallel general purpose computing on GPUs has identified the cache sharing concept as one of the dominant factors towards maximizing the computing resources utilization. It has not been many years ago since NVIDIA adopted the cache sharing concept in General Purpose computing GPUs (GPGPUs) by releasing the G80 architecture where 16 KB of cache memory were shared among the 8 cores of each Streaming Multiprocessor (SM) on the GPU chip [4.17], [4.18]. Taking advantage of the GPU massive parallelization and low-energy characteristics in conjunction with the growing popularity of general purpose computing on GPU-based systems allowed NVIDIA to play the key-role in GPU-based HPC systems by introducing the Fermi and Kepler GPU architectures [4.19], [4.20], which appeared to power several of the top-10 machines in the Top500 list of November 2013 [4.21]. The cache sharing concept allowed, from the application’s threads
perspective, to improve the inter-thread communication by accelerating the data and result sharing in parallel applications [4.19], [4.20].

The above benefits have been corroborated in the simulation analysis of the previous section and in particular for workloads characterized by high parallelization and high data sharing and exchange needs (as in bodytrack). Adopting a shared cache architecture that uses the high-speed all-optical cache memory technology can theoretically strengthen the existing advantages of the cache sharing concept by allowing off-chip ultra-fast caching and lowering total cache capacity requirements, effectively replacing in many cases the Conventional L1+L2 hierarchy with a single shared high-speed cache.

In addition to the advantages coming from the shared all-optical cache memory architecture, the proposed design has also much to offer in the crucial problem of efficiently moving data between the basic system’s sub-modules, i.e., the processors, the cache and the MM. Increasing the maximum offered bandwidth by cache memory modules to processing cores in future parallel GPU-based HPC systems requires novel solutions capable of tackling the new trend of transforming high-parallel application workloads into high-throughput problems. Assuming an electronic 64-bit wide bus operating at 2 GHz, typical in conventional on-chip bus configurations, the total theoretical maximum throughput does not exceed 128 Gb/s. Considering then a 16 GHz optical bus of the same width combined with WDM-formatted optical words, a total data rate of \( M \times 16 \times 64 \text{ Gb/s} \) propagates through the entire bus, with \( M \) representing the number of multiplexed channels per optical wire. Using an 8x WDM format results in a total theoretical throughput of 8.192 Tb/s, which corresponds to a 64x improvement for the proposed architecture that originates from both optical bus’s high-data rate capabilities as well as by its WDM characteristics.

The number of cores that can be served by a single shared optical cache certainly depends on the high clock frequency of the optical cache in order to enhance the degree of Time-Division-Multiplexing between the cores being served. Although this work relies on 16GHz optical building blocks for the cache module, state-of-the-art memory and switch technology have shown that operational speeds up to 100GHz [4.22] should be in principle possible, significantly extending the limit of cores that can be served compared to conventional electronic approaches. However, even after reaching the cache-clock limits and in order to support higher number of cores, the optical nature of the proposed cache and interconnect infrastructure allows for a broad portfolio of additional scaling factors: a combination of TDM, spatial multiplexing and WDM schemes offers a three-
fold increase compared to the electronics, where only spatial multiplexing is available as a bandwidth scaling factor. To this end, a higher number of cores can be supported by adopting the well-known cache multi-porting technique \[4.23\] that would allow for true concurrent accesses from multiple cores to multiple cache lines, increasing in this way the total available cache bandwidth. The next step could be the utilization of the multiport optical cache unit in conjunction with a multiport WDM switch like the Arrayed Waveguide Router (AWGR) \[4.24\] that would allow the sharing of the cache resources to even higher number. Such capabilities are impossible for the electronic technology. As a result, the proposed system’s scale factor would be proportional to both the cache-clock speed and the number of total network ports.

The total energy consumption requirements of the proposed approach can be separated into the energy consumed by the optical cache and by the optical busses employed for the CPU-cache and cache-DRAM links. Regarding optical cache, its energy consumption has been extensively covered in the cache scalability analysis of section 3.4, concluding that the PhC nanocavities technology could in principle lead to lower power consumption memory modules when compared to the electronic approach; the energy consumption of an 8 KB electronic cache equals 1.1 pJ / bit period while the respective energy consumption of an equal-capacity optical cache module equals ~143 fJ / bit period. All the detailed findings of scalability analysis can be found in Table 3.3 of section 3.4. As for the optical links, the proposed scheme is again advantageous compared to the state-of-the-art electronic interconnects. According to \[4.25\], data movement is currently the most energy consuming factor in electronic CPU and GPU chips. Moving 256 bits of data with an electronic bus costs around 1pJ/bit for a link distance of ~5 mm, around 4pJ/bit when link distance increases to 20mm and reaches even 30nJ/bit when off-chip connections to the DRAM are pursued. The respective value for the proposed optical bus is independent of the link distance and depends solely on the modulator’s and on the photo-receiver’s energy consumption at the transmitter and receiver’s site, respectively; in the case of CPU-cache communication, however, only one of these modules is utilized depending on the communication direction (CPU-to-cache or cache-to-CPU), given that the optical cache unit operates exclusively in the optical domain and as such does not require any additional o/e/o modules. For the CPU-to-cache path silicon micro-ring modulators can achieve as low as 471 fJ / bit energy consumption at 25 Gb/s or ~29 fJ / bit at 20 Gb/s \[4.26\]. For the cache-to-CPU communication, complete optical receiver modules that can achieve a total energy
consumption of only 1.27 pJ/bit [4.27] have been already reported. Even more important, these optical solutions can be applied in the proposed scheme also for the cache-to-DRAM communication, offering tremendous energy savings as the 30nJ/bit requirements of current electronics will reduce to pJ/bit scale.

Given that the optical cache and the optical transceiver technology rely on resonant optical structures like the ring modulators and the master-slave FF modules included in the RAM cells, thermal stabilization could be probably a significant issue to take into account. Thermally stabilized ring modulators can be controlled within an energy consumption of only 100fJ/bit at 25 Gb/s [4.28] (even if we have to spend 2mW in resonance tuning) and could be a possible option in the proposed platform. Another promising solution could be the utilization of athermal silicon modulator devices, which have been recently presented to consume only 2.77 fJ/bit energy consumption at speeds of up to 25 Gb/s [4.29]. Regarding the master-slave FF modules, the main problem arising by possible thermal drifts would originate from the thermally induced link length variation at the coupling section between the two coupled switches. However, the analysis reported in [4.30] has revealed that a coupling length of even 0.2 cm would allow for operation speeds of up to 30 Gb/s for a SOA-MZI-based FF module. In a realistic scheme, the coupling length between the master-slave modules would be limited to only a few μm and SOA-MZIs should be replaced by ultra-small PhC switches, which even with the temperature variations would certainly allow for successful operation speeds much higher than the 16 GHz used in the current work. As such, provided that the coupling distance can be kept small, which is indeed feasible using state-of-the-art Silicon waveguides, temperature variations are not expected to cause any problems in the performance of the optical RAM cells. Finally, placing the shared cache on a separate chip next to the processor die is expected to eliminate the temperature variations caused by the processing cores.

4.5 Conclusion

This chapter demonstrated an optical bus-based CMP architecture where the all-optical cache memory architecture of Chapter 3 is used as shared cache among the processing cores suggesting a totally flat cache hierarchy. Placing the shared cache on a separate chip next to the processor die allows for better chip area utilization in favor of the processing elements. All the CPU-DRAM communication is realized completely in the optical domain by utilizing proper WDM optical interfaces combined with spatial-multiplexed
optical waveguides that connect both processing cores and MM with the single cache unit. The proposed architecture leads to significant L1 miss rate reduction of up to 96% for certain cases that can be translated to an average performance speed-up of 19.4% or an average cache capacity requirements reduction of ~63%. Thus, the proposed architecture forms a promising system solution for bridging the gap between the optically connected CPU-DRAM schemes and high speed optical memory technologies.

4.6 List of References


4.21. Top 500 Supercomputers’ list of November 2013 (http://www.top500.org)


4.29. E. Timurdogan, C. M. Sorace-Agaskar, J. Sun, E. S. Hosseini, A.
Biberman, and M. R. Watts, An ultralow power athermal silicon modulator, Nature Communications 5, Article number: 4008, June 2014, doi:10.1038/ncomms5008

Chapter 5

An All-Optical Ternary-Content Addressable Memory (T-CAM) row Architecture for Address Lookup at 20 Gb/s

5.1 Introduction

Content Addressable Memories (CAMs) form a popular design choice for routing table implementations thanks to their fast searching capabilities. However, high speed Address Look-up (AL) operation is still challenging due to the speed limitations imposed by conventional electronic technologies. An optical CAM-based routing table would in principle allow for keeping up with the increasing data-rates of optical packet payload. The first experimental all-optical Binary-CAM (B-CAM) cell architecture has been recently demonstrated showing for the first time 10 Gb/s error-free Search and Write operation [5.1]. This chapter demonstrates the first all-optical Ternary-CAM (T-CAM) cell and its interconnection in an optical T-CAM row architecture, where 4 T-CAM cells and a novel WDM-encoded matchline design can provide comparison operation for a complete 4-bit optical word. The optical T-CAM cell extends the B-CAM cell operation by allowing the storage of a third state “X”, enabling in this way the essential subnet-masked operation needed in modern router applications. The 4-cell T-CAM-row architecture follows a proper wavelength encoding scheme by using an Arrayed Waveguide Grating (AWG)-multiplexer; the multi-wavelength output signal produced at the final row output determines whether a success comparison result is achieved throughout the complete T-CAM row. The performance evaluation of the 4-cell T-CAM row architecture has been carried out using the VPI Photonics simulation suite and by employing experimentally verified SOA-based building blocks. The proposed T-CAM row architecture can be easily scaled to form complete optical T-CAM tables required in AL, while the recent experimental developments in high-speed and ultra-low-power integrated photonic crystal InP-on-Si FF devices [5.2] could potentially allow for its experimental implementation in low-footprint and low-energy prototypes.

Chapter 5 is organized as follows: section 5.2 presents the all-optical T-CAM cell architecture as an extension to the all-optical Binary-CAM cell by using an additional
optical FF and proceeds also with the presentation of the complete T-CAM-row architecture that consists of an indicative number of 4 cells. Section 5.3 presents the corresponding simulation results that demonstrate successful T-CAM row operation at 20 Gb/s for both Search and Write functionalities. Finally, section 5.4 concludes the chapter.

5.2 Optical T-CAM cell and row Architectures

A CAM-based routing table usually consists of a T-CAM table interlinked to a RAM table. The T-CAM table stores all possible destination addresses following typically a layout where every CAM row includes a single address, while the RAM table employs the addresses of the proper router output ports. In this way, the T-CAM table is responsible for identifying the desired destination address of the incoming packet so as to activate the corresponding RAM table entry that will in turn activate the appropriate router output port. An indicative T-CAM-based routing scheme is presented in the example of Fig. 5.1 (a). The destination address of an incoming packet is fed as the search-input into the T-CAM table, while the proper router output port that should be used for forwarding the incoming packet to the desired destination emerges as the search-output signal at the RAM table output. The T-CAM AL operation is realized in a single step since the
An All-Optical T-CAM row Architecture for Address Lookup at 20 Gb/s

destination address of the incoming packet is broadcasted to all T-CAM rows having its constituent bits compared in parallel with the content of every T-CAM row. In case the search-input is identical to a word stored in a T-CAM row, a proper matchline signal identifier emerges at the corresponding row output. A proper encoding and decoding circuit is being used in between the interlinked T-CAM and RAM tables in order to associate the matched T-CAM row matchline signal with the correct RAM table row that stores the proper router output port. In the example of Fig. 5.1 (a), the search-input matches successfully only to the word “X011” stored in the 2nd T-CAM row, with the “X” state denoting that this bit can be successfully matched with an input search value of either 1 or 0. The matchline signal generated at the 2nd T-CAM row is then translated via the encoding and decoding circuitry into the address “01” of the RAM table, which designates that “port B” should be activated at the router output in order to allow the incoming packet to safely propagate to its desired next-hop.

The proposed optical T-CAM row architecture is demonstrated in Fig. 5.1 (b). The row incorporates an indicative number of 4 T-CAM cells where each of them consists of 2 Flip-Flop (FF) modules and 1 XOR gate. The XOR gates are necessary for realizing the comparison operation between the search-input bits and the values stored in the T-CAM cells. The left side FF of each cell is named XFF and is necessary for implementing the third state “X”, enabling in this way the subnet-masked operation which is widely used in modern router applications. On the other hand, the right side FF of each cell is named T-CAM Content FF (TCFF) and stores the actual T-CAM cell content that can be either a logical 0 or 1. When subnet-masked operation is desired, the XFF’s content equals 0 implying that the TCFF respective content has to be ignored. As such, the respective XOR operation does not take into account the TCFF content and the comparison result equals to a logical 0 independently of the value of the search-input bit. On the contrary, in case the TCFF value has to be taken into account, then the XFF content equals to a logical 1 and the XOR output depends upon the comparison between the TCFF value and the respective search-input bit. By assigning a different wavelength for carrying the optical XOR output at every individual T-CAM cell within a row, all the 4 T-CAM cell outputs can be combined at the row output by using an Arrayed-Waveguide Grating (AWG) multiplexer unit (Fig. 5.1 (b)). This leads to a WDM-encoding scheme that produces the corresponding matchline signal at the final row output. In this way, a matchline signal of a logical value 0 indicates a completely matched comparison result since all the individual XOR outputs will equal to a logical 0. On the contrary, a non-zero
optical power level obtained at the encoder input indicates that at least one individual XOR output produces a comparison miss, denoting a non-completely matched row.

Fig. 5.2 presents the proposed all-optical T-CAM cell architecture comprising the TCFF and XFF that subsequently feed an optical XOR gate. The optical XOR gate consists of a single SOA-MZI switch, while both the TCFF and XFF units comprise two interlinked SOA-MZI switches that form a well-known optical master-slave FF configuration [5.3]-[5.6]. For both FFs, a proper Set/Reset pulse mechanism is used in order to switch between the two possible logical states [5.4], [5.5]. Both XFF and TCFF are powered by 2 Continuous Wave (CW) laser beams: $\lambda_e$ is used as input signal at the right-side switches of both XFF and TCFF, while a CW signal at $\lambda_a$ and $\lambda_f$ is launched as the input signal at the left-side SOA-MZI switches of the XFF and TCFF, respectively. As such, the content of the XFF and the TCFF gets encoded on $\lambda_a$ and $\lambda_f$ wavelengths as the FF output signals, respectively. The XFF output signal is then fed as the input signal at the XOR gate, after being filtered in an Optical Bandpass Filter (OBF). On the other side, the TCFF output at $\lambda_f$ enters the XOR gate as the control signal of the upper-branch SOA. The lower branch SOA of the XOR gate is being fed with the input search bit that acts as the second control signal. In this way, the TCFF output and search bit values get logically XORed and the comparison result gets imprinted on the XFF output signal at $\lambda_a$ that is used as the XOR input. Whenever the T-CAM cell is in the “X” state, the XFF output equals to a logical 0 resulting to a logical “0” at the final XOR output irrespective of the TCFF and search bit values. On the contrary, the final XOR output depends indeed on the comparison result between the TCFF output and search bit values when the XFF output equals a logical “1”: when both the TCFF output and search bit signals have the same value the XOR output is 0, while in the opposite case that they have different values the XOR output equals “1” and is imprinted on the XFF output at $\lambda_a$. 

Fig. 5.2: all-optical T-CAM cell architecture with 2 FFs (TCFF & XFF) and a XOR gate and T-CAM row’s AWG multiplexer for 4 indicative T-CAM cells.
Assigning different wavelengths to all the T-CAM cell outputs within a single row allows for the realization of a simple wavelength encoding scheme by using an AWG multiplexer, as presented at the right side of Fig. 2; λa through λd are used for the different cell outputs, while λe, λf and the wavelengths used for the Set/Reset signals are employed in all T-CAM cells. The complete absence of optical power at the final multiplexed T-CAM row output indicates a perfect match between the search-input bits and the row’s contents.

5.3 Simulation Results

This section presents the physical layer simulation-based performance analysis of the T-CAM row architecture for both Search and Write operations at a line-rate of 20 Gb/s. The simulation models have been developed using the VPI Photonics suite and both XOR gate and FF models are based on experimentally verified building blocks. More specifically, XOR gates follow the SOA-MZI-based dual-rail logic model, as this has been presented by [5.7], while the FF models follow closely the FF model configuration used in the extended simulation analysis of the all-optical cache memory architectures for CMPs [5.5]. The only difference has been the use of co-propagating instead of counter-propagating Set/Reset signals that are used for switching between the two possible FF logical states. More details around the principle of operation of the all-optical FF technology can be found in [5.4], [5.6]. Regarding the SOA model used in both XOR gates and FFs, this is identical with the one presented and experimentally validated by [5.8]. The wavelengths used in the 4-cell arrangement of Fig. 5.2 are equal to: λa: 1564.19 nm, λb: 1562.56 nm, λc: 1559.31 nm, λd: 1557.36 nm, λe: 1554.78 nm, λf: 1546.12 nm, Set: 1548.35 nm and Reset: 1551.88 nm.

Fig. 5.3 presents the simulation results for all the 4 T-CAM cells employed in a single row. Fig. 5.3 (i) and (ii) illustrate the Set/Reset pulse traces that are fed into the XFFs and determine whether the XFF has to define a “X” state for the T-CAM cell or not. Fig. 5.3 (iii) and (iv) illustrate the Set/Reset pulse traces that are fed into the TCFFs of the 4 cells dictating the logical content of every TCFF. Fig. 5.3 (v) depicts the XFF output signal, while Fig. 5.3 (vi) illustrates the TCFF content transitions for every T-CAM cell. As can be seen in both Fig. 5.3 (v) and (vi), successful bit storage operation is achieved according to the respective Set/Reset pulse traces; the presence of a Set pulse leads to a FF content transition to the 0 logical state, while the presence of a Reset pulse leads to a FF content transition to the logical state of 1. Fig. 5.3 (vii) presents the search-bit
pulse traces that are fed into the XOR gates of the 4 T-CAM cells as parallel streams in order to get compared with the respective T-CAM cell contents. The search-bit pulse traces are Non-Return-to-Zero (NRZ) $2^7-1$ Pseudorandom Binary Sequences (PRBS) at a line-rate of 20Gb/s. Fig. 5.3 (viii) shows the XOR output signals that form also the T-CAM cell outputs and Fig. 5.3 (ix) illustrates the power level of the final matchline signal that is produced at the row output and just after the AWG multiplexer. As can be noticed, this is a multilevel signal with every different power level corresponding to a different number of bit-level search misses. When all T-CAM cells match the 4 bits of the incoming search-input signal, no optical power is recorded at the AWG output.

Successful matchline operation of the complete T-CAM row can be verified for the entire pulse traces used as the 4 parallel search bit sequences. Three representative examples at the timeslots #1, #9 and #27 have been highlighted in order to facilitate understanding of the T-CAM row performance in different situations. In the example of timeslot #1, all 4 T-CAM cells are in the “X” state since all respective XFF outputs are equal to 0, which finally results to XOR output of 0 regardless of the TCFF and search-bit values. As expected, the final matchline signal at timeslot #1 is also equal to 0, corresponding to a complete match between the T-CAM row and the search-input contents. Within timeslot #9, none of the T-CAM cells is in the “X” state since all XFF outputs are equal to a logical 1. For T-CAM cell #1, the XOR output equals to a logical 0 since the TCFF and the search-bit content are equal. However, for the remaining three T-CAM cells the respective XOR outputs equal to a logical 1 denoting the different content between the corresponding TCFF and search-bit signals. The presence of three
optical pulses at different wavelengths but within the same timeslot #9 designates that the optical power obtained at the AWG output will equal the sum of the power levels of the three individual pulses, obviously leading to a matchline signal with non-zero power that indicates a non-perfectly matched search operation. In the example of timeslot #27, T-CAM cells #2 and #4 are in the “X” state since their XFF content equals a logical 0. As such, the respective XOR outputs are also equal to 0 and this happens even in the case of T-CAM cell #4 where the TCFF content and its respective search-bit are different. Regarding cell #3, the XOR output equals 0 because both the TCFF content and the search-bit are equal. However, cell #1 has a XOR output of 1 since TCFF content and the respective search-bit have different values. This single optical pulse obtained as the result of the comparison along the entire T-CAM row is then also translated into a non-zero power level at the AWG output, suggesting again a non-matched row. Fig. 5.3 (x) presents clearly open eyes for all 4 T-CAM cells with an average extinction ratio of 9.1 dB.

5.4 Conclusion

An all-optical T-CAM cell and row architecture for high-speed AL operation for future network router applications has been presented. Physical layer simulation results demonstrate successful Search and Write operation at a line-rate of 20 Gbps for a complete 4-bit optical word. Although the proposed T-CAM cell and row architectures employ SOA-based switch and FF modules, they are compatible with alternative optical switch and FF technologies [5.2] that can in principle yield higher integration levels, lower-footprint and lower-energy T-CAM-based routing table implementations.

5.5 List of References


Chapter 6
Conclusions and Future Work

6.1 Introduction

This chapter completes the dissertation by summarizing the main contributions, while also providing some potential research lines for future investigation. In particular, section 6.2 contains the most significant concluding remarks of the research realized in this thesis, while section 6.3 outlines some of the open research issues related to it.

6.2 Thesis Conclusions

The main research contribution presented in this thesis revolves around the long-standing “Bandwidth” and “Memory Wall” problems and has been focused towards enabling exascale processing powers in future HPC computing systems. Innovative optical-enabled Computing Architectures are proposed for increasing overall system performance in terms of application execution time, memory and interconnection bandwidth, latency and energy consumption.

Chapter 2 presented OptoHPC-Sim simulation platform, which supports the utilization of optical interconnect and electro-optical routing technologies at system-scale offering complete end-to-end simulation of HPC-systems and allowing for reliable comparison with existing HPC platforms. OptoHPC-Sim takes into account a complete set of recent technological advances towards implementing high-density and multi-layered EOPCBs that can get benefit of novel on-board optoelectronic routing approaches. As such, OptoHPC-Sim can be effectively used in the design of future EOPCB and HPC-interconnect hardware architectures towards optimizing the overall system performance.

Besides describing the simulator features, chapter 2 proceeded to a comparative performance analysis between a system following the Titan Cray XK7 network specifications, which has been ranked as the world’s no. 3 Supercomputer as of June 2016, and a respective HPC architecture where electronic CRAY blades have been replaced by a completely optical blade design that makes use of novel optical
technologies. The results obtained with OptoHPC-Sim suggest that the employment of board-level optics in appropriate layouts can lead to optically enabled HPC network system configurations that can significantly outperform conventional HPC machines, on average offering throughput improvements higher than ~190% as well as mean packet delay reduction of ~83% for the 8 synthetic traffic profiles that are currently supported by OptoHPC-Sim.

Going down through the HPC hierarchy and focusing on the processor chip level chapter 3 demonstrated a 16 GHz physical layer optical cache memory architecture that can support both direct and N-way cache mapping associativity. The proposed optical cache architecture takes advantage of WDM principles by using WDM-formatted words and address fields, thus significantly relaxing the requirements in terms of active components and power consumption. The proposed cache architecture incorporates five discrete subsystems that have been designed relying on already experimentally demonstrated optical building blocks. The subsystems include (i) an optical interface module that is used for translating the wavelength-formatted data and address fields between the processor and the cache memory unit, (ii) a Read/Write Selector for ensuring correct access to the stored data, (iii) a completely passive row decoder for addressing the different storing locations of the cache memory, (iv) a 2D optical RAM bank for storing the actual data and (v) a tag-comparison circuit used for determining whether the desired data are stored in the cache memory unit. Moreover, chapter 3 demonstrates a scalability analysis of the proposed architecture where various state-of-the-art optical technologies are taken into account. The results present comparable footprint and energy consumption measurements with the respective conventional technologies. The performance evaluation of the proposed optical cache memory architecture has been realized with the commercially available VPI Photonics simulation platform, showing successful Read and Write operation at operation speeds of up to 16 GHz.

Towards highlighting the transformative character of the proposed optical cache architecture when employed in CMP architectures as well as its credentials in addressing the “Memory Wall” problem, chapter 4 demonstrated an optical bus-based CMP architecture where an all-optical cache memory is shared among the processing cores suggesting a totally flat cache hierarchy. The suggested flat hierarchy negates the need for complex cache hierarchy schemes while placing the shared cache on a separate chip next to the processor die allows for better chip area utilization in favor of the processing
elements. As such, the proposed CMP architecture is a single-level shared L1, which sidesteps the issue of data consistency since only one copy of the data exists in the shared L1. All the CPU-DRAM communication is realized completely in the optical domain by utilizing proper WDM optical interfaces combined with spatial-multiplexed optical waveguides that connect both processing cores and MM with the single cache unit. The optical buses design implementation takes into consideration various novel optical interconnect technologies that can offer significantly lower energy consumption as well as significantly better performance. A comparison analysis between the proposed optical cache-based CMP architecture and a typical CMP hierarchy of dedicated L1 and shared L2 electronic caches has been realized by using 12 parallel workloads of the PARSEC benchmark suite on top of the Gem5 simulator. The simulation results presented significant L1 miss rate reduction that can be translated to an average performance speed-up of 19.4% or an average cache capacity requirements reduction of ~63%. Thus, the proposed architecture forms a promising system solution for bridging the gap between the optically connected CPU-DRAM schemes and high speed optical memory technologies.

The final chapter of this thesis (Chapter 5) has been dedicated in transferring the optical memory technology benefits from CMP layouts also in novel all-optical routing table architectures. More specifically the first all-optical T-CAM cell and row architecture for high-speed Address Lookup operation for future network router applications has been presented where a novel WDM-encoded matchline scheme, by using an Arrayed Waveguide Grating (AWG)-multiplexer, can provide comparison operation for a complete 4-bit optical word; the multi-wavelength output signal produced at the final row output determines whether a success comparison result is achieved throughout the complete T-CAM row. The proposed optical T-CAM cell extends the first all-optical B-CAM cell operation by allowing the storage of a third state “X”, enabling in this way the essential subnet-masked operation needed in modern router applications. The physical layer simulation results presented in chapter 5 have been extracted by using the VPI Photonics simulation suite and by employing experimentally verified SOA-based building blocks and demonstrate successful Search and Write operation at a line-rate of 20 Gb/s for a complete 4-bit optical word, i.e. more than 2x the speed that can be offered by state-of-the-art electronic CAM look-up table designs. Taking into account that the proposed T-CAM cell and row architectures are compatible with alternative high-speed and ultra-low-power integrated photonic crystal InP-on-Si FF and switch devices, this
can in principle yield higher integration levels, lower-footprint and lower-energy T-CAM-based routing table implementations.

6.3 Future Work

The main contributions presented in this dissertation can be followed by several new research lines for future investigation. Some of the open topics for future work are:

- The proposed OptoHPC-Sim simulation engine has all the credentials for being enriched with new capabilities that can extend its usability regarding the architectural and technology decision making on the way to the exascale era. Supporting for example the energy consumption performance analysis of the network’s subsystems, forms an extension of significant importance since it can be particularly useful both in system design as well as in the comparison analysis among different HPC network configurations. As such, the simulation analysis of section 2.4 could be further extended by giving a better insight about the energy characteristics of the studied HPC network architectures and technologies.

- Extending the supported set of hardware modules (including both the electrical and optical ones) can be proven beneficial since it will allow the exploration of a broader range of HPC network topologies and architectures. Future versions of OptoHPC-Sim can be enriched with other popular network topologies like: (a) Butterfly, (b) Dragonfly, (c) Hypercube, (d) Fat Tree, (e) 5D-Torus and (f) 6D-Torus (a.k.a. Tofu). Towards this direction, new simulation models have to be developed in order to extend the current simulation library for reflecting all the necessary hardware modules that are needed for building such topologies. On top of that, the current set of routing algorithms must be extended, since DOR and MOVVR implementations that come with OptoHPC-Sim can only support up to 3-dimensional torus and mesh arrangements. Numerous algorithms have been proposed in the literature representing all the three (a) Deterministic, (b) Oblivious and (c) Adaptive routing algorithm classification categories.

- Another topic of significant interest for extending OptoHPC-Sim would be the capture and analysis of network traffic extracted from applications being executed in real HPC systems. The captured message sequences can be effectively used for
either creating real-world packet traces or generating proper statistical data-sets describing the traffic pattern distribution among the computing nodes. Feeding OptoHPC-Sim with such data will offer the significant advantage of more accurate and realistic simulation capabilities that will be focused around the desired applications.

- Moving now to the scalability perspective of CMP architectures that make use of optical cache memories, Fig. 6.1 presents a possible first step towards extending the optical bus-based CMP architecture of Fig. 4.1 (b). According to the TDM-based access scheme presented in section 4.2.1, the number of cores that can be efficiently served in a bus-based shared-cache approach certainly depends on the clock frequency of the optical cache that is shared among the cores. After reaching the optical cache clock limits, the next step can be the utilization of an optical crossbar or a multiport WDM switch like the Arrayed Waveguide Router (AWGR) in order to allow the sharing of the cache memory resources to even higher number of processing cores. As can be seen in the concept of Fig. 6.1, an optical crossbar or an AWGR can be used for interconnecting multiple L1 optical cache units, mimicking the approach of SPARC T5 processor (see Fig. 1.2) where an electronic crossbar is used for interconnecting the L3 caches. Validating such an approach would require (a) the design and physical-layer performance evaluation of all the necessary optical hardware components that are needed for implementing.

Fig. 6.1: Possible extension for the optical bus-based CMP architecture of Fig. 4.1 (b) where multiple L1 cache units are being interconnected with an optical crossbar or an AWGR
the interconnect scheme among the caches and (b) the system-level evaluation of the complete system in order to identify the possible performance benefits that can be gained compared to conventional approaches. The physical-layer performance evaluation can be realized again with the use of VPI Photonics suite while the system-level evaluation can be realized with the use of Gem5 in conjunction with a number of parallel benchmarks like the PARSEC benchmark suite.

- Regarding the expansion of the T-CAM row architecture of chapter 5 to a complete all-optical T-CAM-based routing table scheme, the next research steps have to deal with both the design of new all-optical subsystems as well as with their performance evaluation by means of physical-layer simulations. In more detail, both the optical designs for (a) the encoding/decoding circuit between the T-CAM and RAM tables and (b) the optical RAM table of Fig. 5.1 (b) have to be implemented and subsequently evaluated in order to demonstrate complete routing operation for realistic scenarios. The desired solution must effectively distinguish the matching T-CAM table row and associate it with the correct RAM table row in order to identify the proper router output port. Again, VPI Photonics suite can be effectively used for the physical-layer performance evaluation of all these subsystems.
Annex A: Publications

List of Publications

A.1  Journal Publications


Annex A: Publications

2016


A.2 Conference Publications


Annex A: Publications

mmwave massive MIMO and MT-MAC protocols,” will appear in Proceedings of SPIE Photonics West, San Francisco, California, USA, 27 January - 1 February 2018


Fukuoka, Japan, 25-28 October 2015


# Annex B: Abbreviations

## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2DRB</td>
<td>2D Random Access Memory Bank</td>
</tr>
<tr>
<td>AG</td>
<td>Access Gate</td>
</tr>
<tr>
<td>AL</td>
<td>Address Look-up</td>
</tr>
<tr>
<td>AOC</td>
<td>Active Optical Cable</td>
</tr>
<tr>
<td>ASIC</td>
<td>Application-Specific Integrated Circuit</td>
</tr>
<tr>
<td>ATDT</td>
<td>Automatic Topology Design Tool</td>
</tr>
<tr>
<td>AWG</td>
<td>Arrayed Waveguide Grating</td>
</tr>
<tr>
<td>B-CAM</td>
<td>Binary-Content Addressable Memory</td>
</tr>
<tr>
<td>BER</td>
<td>Bit Error Rate</td>
</tr>
<tr>
<td>CAM</td>
<td>Content Addressable Memory</td>
</tr>
<tr>
<td>CAS</td>
<td>Column Address Selector</td>
</tr>
<tr>
<td>CD</td>
<td>Column Decoder</td>
</tr>
<tr>
<td>CDR</td>
<td>Clock and Data Recovery</td>
</tr>
<tr>
<td>CG</td>
<td>Clock Generator</td>
</tr>
<tr>
<td>CMP</td>
<td>Chip Multiprocessor</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit</td>
</tr>
<tr>
<td>CW</td>
<td>Continuous Wave</td>
</tr>
<tr>
<td>DOR</td>
<td>Dimension Order Routing</td>
</tr>
<tr>
<td>DRAM</td>
<td>Dynamic Random Access Memory</td>
</tr>
<tr>
<td>DTR</td>
<td>Data to Read</td>
</tr>
<tr>
<td>DTW</td>
<td>Data to Write</td>
</tr>
<tr>
<td>EOPCB</td>
<td>Electro-Optical PCB</td>
</tr>
<tr>
<td>FF</td>
<td>Flip-Flop</td>
</tr>
<tr>
<td>FIFO</td>
<td>First In First Out</td>
</tr>
<tr>
<td>FinFET</td>
<td>Fin Field Effect Transistor</td>
</tr>
<tr>
<td>GPGPU</td>
<td>General Purpose Graphics Processing</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphics Processing Unit</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>HPC</td>
<td>High Performance Computing</td>
</tr>
<tr>
<td>I/O</td>
<td>Input/Output</td>
</tr>
<tr>
<td>IC</td>
<td>Integrated Circuit</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>IPC</td>
<td>Instructions Per Cycle</td>
</tr>
<tr>
<td>ISA</td>
<td>Instruction Set Architecture</td>
</tr>
<tr>
<td>L1</td>
<td>Level-1</td>
</tr>
<tr>
<td>L1d</td>
<td>Level-1 data</td>
</tr>
<tr>
<td>L1i</td>
<td>Level-1 instruction</td>
</tr>
<tr>
<td>L2</td>
<td>Level-2</td>
</tr>
<tr>
<td>L3</td>
<td>Level-3</td>
</tr>
<tr>
<td>LAN</td>
<td>Local Area Network</td>
</tr>
<tr>
<td>LRU</td>
<td>Least Recently Used</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MAR</td>
<td>Memory Address Register</td>
</tr>
<tr>
<td>MBR</td>
<td>Memory Buffer Register</td>
</tr>
<tr>
<td>MC</td>
<td>Memory Controller</td>
</tr>
<tr>
<td>MLA</td>
<td>Microlens Array</td>
</tr>
<tr>
<td>MM</td>
<td>Main Memory</td>
</tr>
<tr>
<td>MOVVR</td>
<td>Minimal Oblivious Valiant Routing</td>
</tr>
<tr>
<td>MPI</td>
<td>Message Passing Interface</td>
</tr>
<tr>
<td>MT</td>
<td>Multifiber Termination</td>
</tr>
<tr>
<td>MTP</td>
<td>Multifiber Termination Push-On</td>
</tr>
<tr>
<td>MZI</td>
<td>Mach-Zehnder interferometer</td>
</tr>
<tr>
<td>NoC</td>
<td>Network-on-Chip</td>
</tr>
<tr>
<td>NRZ</td>
<td>Non Return to Zero</td>
</tr>
<tr>
<td>OPCB</td>
<td>Optical Printed Circuit Board</td>
</tr>
<tr>
<td>OS</td>
<td>Operating System</td>
</tr>
<tr>
<td>PCB</td>
<td>Printed Circuit Board</td>
</tr>
<tr>
<td>PD</td>
<td>Photodiode</td>
</tr>
<tr>
<td>PhC</td>
<td>Photonic Crystal</td>
</tr>
<tr>
<td>PPS</td>
<td>Poly-phenylene Sulfide</td>
</tr>
<tr>
<td>PRBS</td>
<td>Pseudorandom Binary Sequence</td>
</tr>
<tr>
<td>PSM4</td>
<td>Parallel Single Mode 4-channel</td>
</tr>
<tr>
<td>PTH</td>
<td>Plated-Through Hole</td>
</tr>
<tr>
<td>QSFP</td>
<td>Quad Small Form-factor Pluggable</td>
</tr>
<tr>
<td>RAG</td>
<td>Read Access Gate</td>
</tr>
<tr>
<td>RAM</td>
<td>Random Access Memory</td>
</tr>
<tr>
<td>RAS</td>
<td>Row Address Selector</td>
</tr>
<tr>
<td>RD</td>
<td>Row Decoder</td>
</tr>
<tr>
<td>RISC</td>
<td>Reduced Instruction Set Computer</td>
</tr>
<tr>
<td>RWS</td>
<td>Read/Write Selector</td>
</tr>
<tr>
<td>SerDes</td>
<td>Serializer/Deserializer</td>
</tr>
<tr>
<td>SF</td>
<td>Store and Forward</td>
</tr>
<tr>
<td>SM</td>
<td>Streaming Multiprocessor</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>SOA</td>
<td>Semiconductor Optical Amplifier</td>
</tr>
<tr>
<td>SOI</td>
<td>Silicon on Insulator</td>
</tr>
<tr>
<td>SRAM</td>
<td>Static Random Access Memory</td>
</tr>
<tr>
<td>SST</td>
<td>Structural Simulation Toolkit</td>
</tr>
<tr>
<td>TC</td>
<td>Tag Comparator</td>
</tr>
<tr>
<td>T-CAM</td>
<td>Ternary-Content Addressable Memory</td>
</tr>
<tr>
<td>TDM</td>
<td>Time Division Multiplexing</td>
</tr>
<tr>
<td>VC</td>
<td>Virtual Channel</td>
</tr>
<tr>
<td>VCSEL</td>
<td>Vertical Surface Emitting Laser</td>
</tr>
<tr>
<td>VCT</td>
<td>Virtual Cut Through</td>
</tr>
<tr>
<td>WAG</td>
<td>Write Access Gate</td>
</tr>
<tr>
<td>WDM</td>
<td>Wavelength Division Multiplexing</td>
</tr>
<tr>
<td>WS</td>
<td>Way Selector</td>
</tr>
<tr>
<td>XGM</td>
<td>Cross Gain Modulation</td>
</tr>
<tr>
<td>XPM</td>
<td>Cross Phase Modulation</td>
</tr>
</tbody>
</table>