Energy-Aware Design Space Exploration for Network-on-Chip Architectures

Antonios Psathakis

Thesis submitted in partial fulfillment of the requirements for the

Masters' of Science degree in Computer Science

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes University Campus, Heraklion, Crete, GR-70013, Greece

Thesis Advisor: Prof. *Manolis Katevenis*

This work has been supported and performed at **Computer Architecture and VLSI Systems (CARV) Laboratory, Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH)**

University of Crete
Computer Science Department

Energy-Aware Design Space Exploration for Network-on-Chip Architectures

Thesis submitted by
Antonis Psathakis
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science

THESIS APPROVAL

Author: _____
Antonis Psathakis

Committee approvals: _____
Manolis Katevenis
Professor, University of Crete
Thesis Supervisor

_____
Angelos Bilas
Professor, University of Crete
Committee Member

_____
Dionisios Pnevmatikatos
Professor, Technical University of Crete
Committee Member

Departmental approval: _____
Angelos Bilas
Professor, Director of Graduate Studies

Heraklion, October 2013

Abstract

As semiconductor technology advances, more area is available to inject memory and logic onto a single chip. This area can be either partitioned into few complex processor elements (PEs) or many simple PEs. The current trend is to inject as many simple PEs as possible. In order to connect many PEs, Network-on-Chip (NoC) solutions are becoming more favourable due to their modular characteristics. However, NoC's utilization is affected by the increase in PE count, decreasing the network's and system's energy efficiency.

Various proposals have appeared for tackling energy-utilization inefficiencies at the network architecture level. These fall within three architectural parameters: network partitioning (P), concentration (C) and express physical links (X). However, these efforts assume a small design space among the P,C and X parameters, or use unscalable schemes for express physical links, or unfairly distribute buffer space and bisection bandwidth along the network architectures. As a consequence, researchers end up with different conclusions in terms of how a network's energy efficiency is affected by the P,C and X architectural parameters.

This work evaluates the area, performance and energy of the network architectures that have been derived by applying each of the three architectural parameters (P, C and X), either separately or combinatorially, to a (baseline) single 2D mesh network. The P parameter considers homogeneous partitioning and two types of heterogeneous partitioning, the C parameter explores one concentration degree of 4 PEs per network node, and finally, the X parameter assumes two express intervals (2-hop or 4-hop). This has resulted in a design space of 20 and 24 network configurations for 64 and 256 PEs respectively. All of the network instances were simulated using various traffic patterns that exhibit diverse communication behaviors and a varying range of control packets per data packet ratios. To enforce strong fairness, we kept each network's buffer space allocation and bisection bandwidth almost equal to the baseline, by properly adjusting the respective router micro-architecture without degrading performance. In some cases, the router micro-architecture adjustments improved performance.

Drawing on insights from our analysis, we observe that in future NoCs of hundreds of PEs, the exclusive use of express physical links utilizing express interval equal to 2, without concentration or network partitioning, is the best approach in terms of energy-area savings and energy-area efficiency. Furthermore, we demonstrate that network partitioning under a fair buffer space and bisection bandwidth allocation decreases the area efficiency rather than increases, regardless of partitioning scheme. Energy consumed and energy efficiency declines too, except for a particular type of heterogeneous scheme that gives slight improvements. However this happens only in cases where a specific range of control packets per data packet ratios are injected. Finally, through this work, one can determine the best suited network architecture for each different use-case and PE count.

Περίληψη

Καθώς εξελίσσεται η τεχνολογία των ημιαγωγών, περισσότερος χώρος διατίθεται στους σχεδιαστές, για να τοποθετήσουν λογική και μνήμη σε ένα τσιπ. Αυτός ο χώρος, μπορεί να αξιοποιηθεί βάζοντας λίγα πολύπλοκα ή πολλά απλά στοιχεία επεξεργασίας (PEs). Οι τρέχουσες όμως τάσεις, υποδεικνύουν περισσότερα απλά PEs. Για την διασύνδεση του ολοένα και αυξανόμενου πλήθους των PEs, οι Δίκτυο-σε-Τσιπ (NoC) λύσεις φαίνονται οι πιο προσιτές, εξαιτίας του δομοστοιχειωτού τρόπου διασύνδεσης τους. Δυστυχώς, ο ρυθμός αξιοποίησης (utilization) των NoCs μειώνεται σε σχέση με το αυξανόμενο πλήθος των PEs, επηρεάζοντας τόσο την ενεργειακή απόδοση του δικτύου όσο και του συστήματος συνολικά.

Διάφορες λύσεις έχουν προταθεί σε αρχιτεκτονικό επίπεδο, για να αντιμετωπιστεί το πρόβλημα της ενεργειακής απόδοσης. Αυτές υπόκεινται σε τρεις παραμέτρους: διχοτόμηση δικτύου (Δ), συγκεντροποίηση (Σ) και εισαγωγή φυσικών καναλιών παράκαμψης (Π). Οι προτάσεις αυτές θεωρούν κάποιες περιπτώσεις των Δ,Σ και Π παραμέτρων, ή χρησιμοποιούν μη κλιμακώσιμες τεχνικές για φυσικά κανάλια παράκαμψης, ή δεν κατανέμουν δίκαια την συνολική μνήμη και το εύρος ζώνης της διατομής. Ως εκ τούτου, δεν υπάρχουν ξεκάθαρα συμπεράσματα για το πως επηρεάζεται η ενεργειακή απόδοση ενός δικτύου σε σχέση με τις Δ,Σ και Π παραμέτρους.

Στην παρούσα εργασία, γίνεται εκτίμηση του εμβαδού, της επίδοσης και της ενέργειας της κάθε αρχιτεκτονικής, η οποία προκύπτει όταν εφαρμόσουμε ξεχωριστά ή συνδυαστικά καθεμία από τις Δ, Σ και Π παραμέτρους, σε ένα μονό δίκτυο αναφοράς. Η Δ παράμετρος περιλαμβάνει μία ομογενής και 2 ετερογενής διχοτομήσεις, η Σ υποθέτει 4 PEs ανά δικτυακό κόμβο, και η Π, θεωρεί 2 ή 4 γειτονικές ζεύξεις παράκαμψης. Ο σχεδιαστικός χώρος τελικά που διαμορφώνεται, φτάνει στα 20 και 24 δίκτυα, για τα 64 και 256 PEs αντίστοιχα που μελετούνται. Οι μετρήσεις πραγματοποιήθηκαν με διαφορετικά είδη δικτυακών κινήσεων και ποικίλων λόγων πακέτων ελέγχου ανά πακέτο δεδομένου. Για να υπάρξει δίκαιη σύγκριση, διατηρήσαμε τη συνολική μνήμη και το εύρος ζώνης της διατομής του κάθε δικτύου σχεδόν ίσο με το δίκτυο αναφοράς, αλλάζοντας προσεκτικά την μίκρο-αρχιτεκτονική του αντίστοιχου δικτυακού κόμβου, χωρίς απώλεια επιδόσεων. Σε κάποιες περιπτώσεις, η μετατροπή αυτή, βελτίωνε την επίδοση.

Με βάση στοιχεία που προέκυψαν από την ανάλυση μας, παρατηρούμε ότι στα μελλοντικά NoCs της τάξης των εκατοντάδων PEs, η αποκλειστική χρήση φυσικών καναλιών παράκαμψης 2 γειτονικών ζεύξεων, χωρίς συγκεντροποίηση ή διχοτόμηση, φαίνεται να είναι η καλύτερη επιλογή όσον αναφορά την εξοικονόμηση και απόδοση ενέργειας-εμβαδού. Επίσης, επιδεικνύουμε ότι η διχοτόμηση ενός δικτύου κάτω από ίσες συνθήκες συνολικής μνήμης και εύρους ζώνης διατομής, μειώνει αντί να αυξάνει τη χωρική απόδοση, ανεξάρτητα από το είδος της διχοτόμησης. Η ενέργεια και η ενεργειακή απόδοση επίσης μειώνεται, εκτός από ένα συγκεκριμένο τρόπο διχοτόμησης όπου δίνει ελάχιστες βελτιώσεις. Αυτό συμβαίνει μονάχα στις περιπτώσεις που ο λόγος του πλήθους των πακέτων ελέγχου ανά πακέτο δεδομένου, βρίσκεται σε σε ένα ορισμένο εύρος τιμών. Τέλος, μέσα από τη παρούσα εργασία, μπορεί κανείς να διαλέξει τη καλύτερη δυνατή αρχιτεκτονική, ανάλογα με τη περίπτωση χρήσης.

Acknowledgements

This work was done at the Computer Architecture and VLSI Systems (CARV) laboratory of the Institute of Computer Science (ICS) of the Foundation for Research and Technology Hellas (FORTH), and was financially supported by a FORTH ICS scholarship. First and foremost, I would like to thank my academic advisor Prof. Manolis GH. Katevenis for supervising and guiding my thesis work. Next, I would like to thank Prof. Dionisios Pnevmatikatos, Prof. Panagiota Fatourou, Dr. Iakovos Mavroidis and Dr. Polyvios Pratikakis, for their support during the weekly meetings and Prof. Angelos Bilas for his constructive questions in the defense of my thesis. I would also like to thank Vassilis Papaefstathiou who helped me handle many technical and architectural problems, George Kalokerinos for his constructive comments and encouraging advices, Kostas Harteros for reviewing and a last-minute issue that he found in the evaluation results, Dr. Sypros Lyberis for discussing various interconnection trade-offs, Michael Lygerakis for the technical support, Vagelis Vasilakis for reviewing, commenting and encouraging this work, Apostolos Glenis, Dimitris Poulios, Maria Chalkiadaki, Tasos Papagiannis and Antonis Papaioannou for helping me a lot with the latex and writing, Tryfon Farmakakis and George Constantinidis for reviewing the abstract, Nikos Papakonstantinou for his repeated questions about my thesis date, and finally, Dr. Manolis Marazakis, Giannis Sfakianakis, Stelios Mavridis, Spyros Papagewrgiou, Markos Foundoulakis, Foivos Zakkak, Yannis Manousakis, Chrysti Simeonidou and Dimitris Yannopoulos.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Trends in on-chip communications

In Chip Multiprocessor Systems (CMPs), using many simple PEs instead of fewer complex ones, seems to be the best power efficient practice [7, 8]. CMPs can take full advantage of technology scaling as they can deploy more PEs and have the ability to power-gate[1] idle PEs while others perform computations [9]. Apart from PEs, CMPs also require other sub-systems, like caches and memory controllers. Each sub-system connects to an on-chip router through a network interface in order to communicate with other sub-systems.

As the number of PEs increases, area limitations are introduced, affecting design complexity, productivity and power density [10, 11]. In order to tackle such limitations, Network-on-Chip (NoC) solutions are becoming more preferable over shared buses and ad-hoc wiring. This happens due to NoC's modular capabilities and scalability characteristics [12]. Most existing NoCs utilize a 2D mesh topology [13, 14, 15, 16, 3], as meshes have lower design complexity and map well to the 2D chip substrate. Although 2D meshes target the above limitations, their utilization decreases as network node increases [17]. This under-utilization affects either the network's energy efficiency and potentially, the whole system efficiency as other parts of the system are stalled waiting for communication.

1.2 NoC Power Dissipation

It has been relatively recent that Intel's Tera-Scale Computing Research Program, produced the teraflop processor [16]. The teraflop (block diagram of teraflop's components is depicted in Fig. 1.1) chip consist of 80 simple cores.

Along with 80 cores, the chip also contains 80 routers. A core is comprised of a single-precision floating-point core, an instruction memory, a data memory slice, one computation block and a dedicated on-chip router. The on-chip router

[1] power-gating is a technique to reduce power consumption by shutting off the current to blocks of the circuit that are not in use.

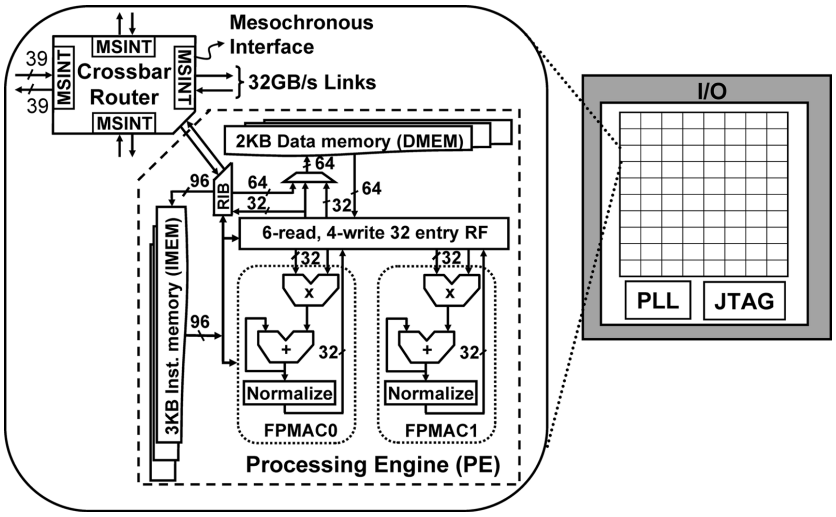
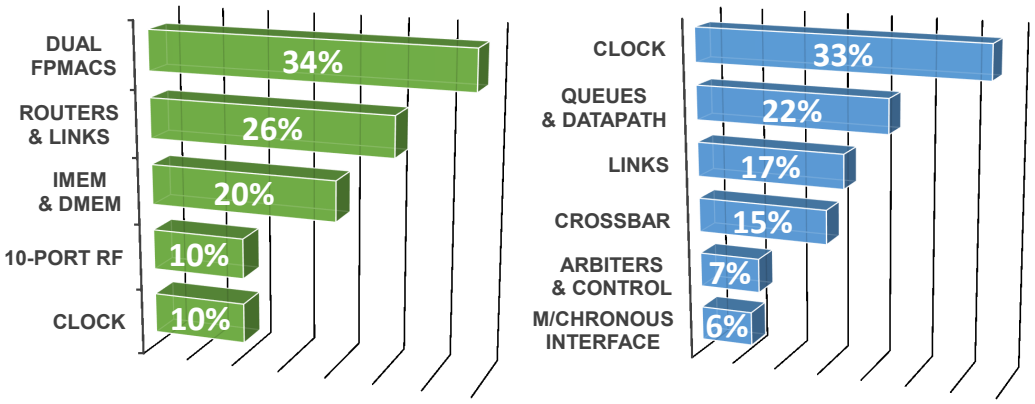


Figure 1.1: Block diagram of teraflop's tile architecture. (Source: Intel Teraflop).

is responsible for the communication of that core with all other cores and components of the processor. All routers, are interconnected by a single 2D mesh network clocked at 5GHz. As illustrated in Fig. 1.2i, the on-chip network consumes 26% of each tile's power. Other works show similar conclusions too, by reporting 36% NoC power dissipation of total chip power. The MIT Raw on-chip network which connects 16 tiles of PEs [18] is such an example. This magnitude of power dissipation has introduced another confirmation of the importance of reducing the energy consumption of on-chip networks, especially when considering such large-scale systems. Fig. 1.2ii illustrates power breakdown of the teraflop NoC. Among the NoC components, total router power is higher relative to the interconnection links. Except for clock, buffer and crossbar power are significant factors to consider too.

1.3 Related Work

Router and links (Sec. 1.2) play the second most significant role in total chip power dissipation. Three on-chip interconnect components seem to contribute most in NoC power dissipation. Buffers, links and crossbars. Numerous techniques have been proposed to encounter scalability, power dissipation, energy consumption and area inefficiency. Some are focused on router micro-architecture level (buffer read bypass, link encoding, segmented crossbar, etc.), others at higher a level (multiple networks, concentration and express physical links). We first discuss about the former (router micro-architecture level) and next, the latter (NoC architecture).



(i) Chip power breakdown. **(ii)** NoC power breakdown.

Figure 1.2: Chip power breakdown and NoC power breakdown for Intel teraplop (Source: Intel).

1.3.1 Focusing on Router Component Power

In the context of router micro-architecture, we distinguish three power hungry components. Buffer, crossbar and links. As previously illustrated (Fig. 1.2ii), the buffer power strongly affects the NoC power dissipation. Thus, initially we begin by discussing about buffer power techniques.

Buffers

In terms of buffer power, buffer reduction or elimination techniques have been investigated. We outline some of these below.

Using buffer-read bypassing [19], a flit[2] of packet can leave straight away if a free output exists at the time it enters the input port. This saves read energy in cases where no contention exists, but write energy still exists. However area savings become zero, and maybe below zero, as extra circuity is needed to implement bypassing.

Circuit switched networks [20] or a combination of circuit and packet switched networks [21, 22]. Circuit-switched networks prevent contention by preallocating bandwidth instead of buffering blocked packets, causing router buffers to be excluded. However, circuits paths need to be established before packet transmission

[2] In wormhole switching (a technique which dictates when the packet moves forward from a router), network packets are broken into small pieces called flits. Buffer and channel bandwidth allocation happens in a flit level (instead of a packet level as in Virtual Cut-through). To avoid head-of-line blocking and improve latency and bandwidth, more than one virtual channels per message type can be used. A virtual channel holds the state needed to coordinate the handling of the flits of a packet over a channel. At a minimum, this state identifies the output channel of the current node for the next hop of the route and the state of the virtual channel (idle, waiting for resources, or active). Since a packet is transmitted flit by flit, it may occupy several flit buffers along its path, creating a worm-like image. This, however, can be confusing since cut-through routing does the same thing.

imposing large latency and power overheads. Circuit-Packet switching came into play to fix the above circuit switching issue, but as it was shown [23], by dividing traffic into best effort (BE) (packet switched network) and guaranteed service (GS) (circuit switched network) BE network doesn't offer low latency in a loaded NoC, performance:cost ratio of BE network is much worse than GS, and finally, the use of BE service breaks the composability among applications.

Two more techniques are presented, that NoC community tried to eliminate Router Buffer: Buffer-less networks [24, 25] and Elastic Buffer (EB) flow control based networks [26]. The former eliminates router buffers by miss-routing packets if the output port is busy, or drop them if no-port is available at that time. This technique comes at the price of lacking support for QoS, traffic class services, fault tolerance, congestion awareness and energy management. A promising effort for reducing router buffer power is the latter, where actual buffering is migrating from the router to the pipelined channels, where flip-flops are controlled through Elastic Buffer (EB) flow control [26].

Crossbars

Crossbars are the basic component that "turn" the packet to the correct non-busy output port. Total power dissipation is mostly due to input-output ports and less in control lines that make the decision of packet switching. But this is not always true. As crossbar becomes narrower (datapath width decreases) or/and smaller (input-output ports decrease), control:input/output power ratio becomes larger. For common crossbars (64bit datapath, 5 input-output ports) control[3] power is around 10% to 15% of the total crossbar's power. We can see power reduction efforts in either Matrix-Crossbars or Tree Multiplex-based crossbars. Below we outline some of them:

Matrix crossbars use cross-points that connect input ports to output ports using transmission gates or tristate drivers. For a NxN Matrix crossbar, cross-points grow quadratically, and area complexity grows $O(N^2W^2)$ [27]. Segmented crossbar [19] is a modification of a matrix crossbar which uses segmented bus scheme [28] to save power but not area.

On the other hand, tree multiplex-based crossbars have area complexity of $O(N^2W)$ [29, 30] and are much more power efficient [30]. This happens due to lower dynamic power complexity $O(N^2W\sqrt{W})$ (or $O(N^2W^{1.5})$) [30] compared to matrix crossbars $O(N^2W^2)$ [27].

Links

Capacitively Driven Low-Swing Interconnection (CDLSI) links and Link Encoding are some techniques for low power dissipation.

Using CDLSI, binary logic is encoded using a lower voltage (V_{swing}) smaller than V_{dd}. Using differential signaling, the sender converts full-swing signal into

[3]Measurements are obtained from ORION [27] power-area simulator.

two signals of opposite polarity bounded by V_{swing}. The receiver, uses a differential sense amplifier that restores the signal swing to its full-swing level. Due to the quadratic relation between voltage and power, power dissipation decreases quadratically. However, this comes at a price of of extra power supply distributed to each sender-receiver and double wiring.

Link-encoding schemes attempt to reduce switching activity in links through intelligent coding [31, 32]. Bus-invert coding [31] uses a simple scheme, where a control signal is asserted and propagated along with inverted data every time hamming distance between transferred data and data to be transferred is larger than half of total bits per cycle transfer (i.e. flit width in bits). SiLENT encoding [32] is another link encoding technique effort.

1.3.2 Targeting the Whole NoC's Power

Various proposals have appeared for tackling energy-utilization inefficiencies at the network architecture level. We indicate these efforts to fall within P,C and X parameters and discuss them below in chronological order. These parameters that can be introduced to a 2D mesh baseline[4], are: network partitioning (P), concentration (C) and express physical[5] links (X).

Concentration have been used by researchers [35, 36, 37] and designers [3, 38] to minimize network node count and design complexity. Balfour and Dally [37] proposed that homogeneously partitioned[6] and concentrated networks that use express physical links on the peripheral network nodes (Cmesh[7] for non partitioned network, CmeshX2 for partitioned network), are the best in terms of energy-area efficiency.

Boris et. al [35] proposed a new architecture, called Multidrop Express Channels (MECS). MECS uses concentrated network nodes and additional router ports for express physical links that scale better than flattened butterfly [36]. They conclude that a homogeneously partitioned concentrated 2D mesh has more energy savings than a single concentrated 2D mesh and a homogeneously partitioned MECS had also better energy savings compared with single MECS.

In Chen et. al's research [34], between homogeneously partitioned meshes and meshes that use express physical links that bypass 4 local hops per express hop, they end up that express architectures are more power-area efficient.

Yoon et. al [39] searched single 2D mesh topologies and heterogeneously[8] par-

[4] A baseline network always refers to a single physical 2D mesh network.

[5] Express virtual channels [33] have appeared in bibliography too. However, this work investigates the physical ones, as they show better power efficiency [34].

[6] A homogeneously partitioned network (HOM) is sliced into many subnetworks (this work investigates two subnetworks). Each subnetwork has identical router-micro-architecture. Thus, each subnetwork has equal number of traffic classes, virtual channels per traffic class and flit depth. Details are found in Sec. 2.1.1 on Pg. 13

[7] Cmesh refers to Balfour and Dally's 64 PE network. It uses concentration equal to 4 and express physical links on peripheral network nodes. It should not be confused by a concentrated network.

[8] A heterogeneously partitioned network is sliced at maximum number of subnetworks as the

titioned ones. They showed that heterogeneously partitioned networks, where each traffic class has its own subnetwork, gives better power-area savings than a single network with virtual traffic classes.

Finally, Volos et. al [40] used 2D mesh networks either partitioned homogeneously or heterogeneously. For homogeneous case, they used Balfour and Dally's scheme. For heterogeneous case, they used two schemes. The first one, was the proposed one, which was similar to Yoon et. al. The difference was that instead of one traffic class per subnetwork, they assigned two traffic classes at one subnetwork, and one traffic class to the other. The other scheme, was similar to Balfour and Dally, but not properly defined. The baseline was partitioned by two subnetworks, where each subnetwork had equal number of traffic classes and varying datapath width. They concluded that heterogeneously partitioned networks are more power-area and energy-efficient than homogeneously partitioned networks.

1.4 Motivation

Our work was motivated by two main observations. The first results in existing efforts while the other in P,C and X network architecture parameters. We begin discussing about the former, and then the latter.

1.4.1 Need for a Unified Evaluation

According to the previous efforts (Sec. 1.3.2 on Pg. 5), there is not a common conclusion. For example, Balfour and Dally and Boris et. al both agree that concentrated homogeneously partitioned networks give better energy savings, but Volos et. al., argue that heterogeneous (using HET1 scheme) architectures (see 2.1.1 for detailed information about HET1 scheme) are better than homogeneous ones in terms of energy efficiency. Chen et. al on the other hand, shows that non partitioned networks with express physical links are better than homogeneous ones. However, they considered homogeneously partitioned networks without express links. Another research of Yoon et. al, shows better power-area savings for heterogeneously partitioned networks (using HET2 scheme) than baseline ones.

Considering the above conclusions, there is not a clear "picture" which architecture can be more energy-friendly. We believe that the main reasons that this happens, are the following:

- Previous design space exploration was insufficient. Network configuration count which arises when all three architectural (P, C and X) parameters are used separately or combinatorially, is much more.

number of total message classes. Our work classifies heterogeneous architectures into 2 types. When the total number of traffic classes equals 3, the first type (HET1) assigns one traffic class to its one subnetwork, and the remaining two, to the second subnetwork. The HET2 type, assigns one traffic class per subnetwork. In both types, each subnetwork has different router-micro-architecture. More details are found in Sec. 2.1.1 on Pg. 13

- Buffer space and bisection bandwidth distribution along the networks are not always distributed fairly. We prove (Sec. 2 on Pg. 11) that even though one keeps the baseline's bisection bandwidth equal to a converted network N (in which N's network P,C and X parameters were modified), the total N network's buffer space is strongly affected.

- Express physical link schemes are not always scalable (i.e., flattened butterfly) relative to network node count. Introducing express physical links, port count increase per router, (See Sec. 2.3.1 on Pg. 2.3.1). plays a critical role in crossbar's area/power scalability. (see Sec. 1.3.1 on Pg. 4).

- Matrix crossbars vs. tree-based ones. Some researches assume matrix crossbars for crossbar components. As discussed in Sec. 1.3.1 on Pg. 4, Matrix crossbar's area-power decreases quadratically as crossbar width decreases due to $O(W^2)$ area and power complexity. Thus, when network partitioning is assumed (P parameter), the power-area savings are appeared to be huge. However, this is a pitfall. Tree-based crossbars offers more scalability and power-area efficiency than matrix-based ones[30, 27].

Work	PEs	Non P'ed				Partitioned												Bis. Alc.	Mem. Alc	X. Scal.
		SPN				HOM				HET1				HET2						
		CX	CX	CX	CX	CX	CX	CX	CX	CX	CX	CX	CX	CX	CX	CX	CX			
balfour'06	64								•											
grot'09	64,256								•											
chen'09	256		•																	
yoon'10	16													•						
volos'12	16									•										
this	64,256																			

Figure 1.3: Design space exploration of previous efforts relative to network partitioning (P), concentration (C) and insertion of express physical links (X). Green(Gray): the respective parameter was(was not) investigated. Red: the respective parameter was not applied. Yellow: the respective parameter was not properly defined. Black dot: the respective NoC architecture that gives better energy/area savings. SPN: Single Physical Network (not partitioned). HOM: Homogeneously partitioned network. HET1/HET2: Heterogeneously partitioned network using HET1/HET2 scheme.

In Fig. 1.3 we depict the design space that each research [37, 35, 34, 39, 40] evaluated, in the context of the P,C and X parameters and some characteristics. These characteristics include: a) bisection bandwidth allocation, b) buffer space allocation and c) NoC's scalability relative to the respective express physical link scheme (if exists). As we observe, bisection bandwidth and buffer space allocation are not always distributed equally along the networks. Scalability imposed from express physical links, is also not maintained in some network configurations

(Detailed analysis why scalability is not maintained in some existing network configuration efforts, is found in Sec. 2 on Pg. 11). Even though some efforts (Chen et. al. and Yoon et. al) fairly distribute buffer space and bisection bandwidth, the considered design space is not large enough to reach to a safe conclusion. This work generates network configurations that try to cover the design space that P,C and X parameters create, and simultaneously keeps the bisection bandwidth and buffer space allocation equally distributed along the network architectures.

1.4.2 NoC Architectural Trade-offs

In this subsection, we outline the trade-offs among the P,C and X architectural parameters. For each parameter, we present the benefits/drawbacks that a single parameter could impose to a baseline network.

Network Partitioning (P)

A flit can choose the appropriate subnetwork (if heterogeneous partitioning is chosen) to leave straight away to its destination. If homogeneous partitioning is chosen, a flit could either leave through the first or the second subnetwork, balancing the load.

- Prioritization delay will be less. Final strict message type ordering, will take place at the destination node (heterogeneous partitioning).
- Each subnetwork's router datapath would be narrower (because bisection wire count is kept constant), imposing larger serialization latency.
- Due to the fact that dynamic power of a (tree-based) crossbar is super-linear relative to width, replacing a wide crossbar by narrow ones, will give less overall power than a wide one (for more details see Sec. 2.1 on Pg. 11).

Concentration (C)

Less router traversals and total routers versus more ports per router and wider crossbar datapath (crossbar needs to become wider in order to keep bisection bandwidth fixed between the baseline and the concentrated network) has the following impacts:

- Less head latency (as long as router's pipe-depth is greater than one cycle) and less serialization latency (because wider datapath is used).
- Crossbar traversal energy increases (because larger/wider crossbar dissipate quadratically/super-linearly more power).
- Less buffer control logic for the same memory budget. This happens as we have few large buffers instead of many small ones.

- Allocator area increases. This happens because request width of each arbiter increases linearly with respect to the number of ports. Total number of arbiters is proportional to the number of ports. Thus, allocator area can increase quadratically with the number of ports.
- In terms of neighbour communication a single router traversal costs more when a packet travels through a larger router.
- Less but longer links per dimension may impose link pipelining. Thus, larger router buffer depth may is needed in order to cover credit return latencies from extra link traversals.

Express Physical Links (X)

Less router traversals versus extra input ports per router have the following impacts:

- Flit traverses less routers. Thus, the cost we have to pay for buffering, arbitration and XB traversal is less.
- Total number of router pipe-stage traversals is decreasing, improving latency. This is has huge effect when average latency is mostly due to head latency rather than serialization latency (i.e. small packets).
- Crossbar complexity grows quadratically ($O(N^2)$) with number of ports (N).
- Bisection wire count becomes greater. Thus, narrower router datapaths are needed and serialization latency becomes larger as a consequence.
- Potential insertion of link pipelining and buffer depth increment in order to cover credit return latencies.

Putting it All Together

When examining each parameter separately, as we previously saw, there are many trade-offs. Although trade-offs are well defined for each parameter there still the the question, what happens if the parameters are combined under the same bisection and buffer space allocation. This is another motivation that caused us to evaluate a large design space of NoC architectures.

1.5 Contribution

This thesis makes the following contributions:

1. Proves that maximum energy-area efficiency and energy savings for tens to hundreds of PEs, are possible as long as we use exclusively express physical links without concentration or/and network partitioning. Energy savings for

tens of PEs can reach 41% and for hundreds of PEs 62%. The optimal express interval that optimizes energy-area efficiency is found to be equal to 2.

2. Shows which is the best NoC architecture for every use-case, such as area, performance, power and energy throughout a large number of evaluated NoC architectures (20 and 24 NoC architectures for 64 and 256 PEs respectively).

3. Explains why there have been different conclusions in the literature, in terms of how a network's energy efficiency is affected by the P,C and X architectural parameters.

4. Shows how control:data packet ratio can affect a NoC architecture's performance and energy.

1.6 Thesis Structure

The rest of this thesis is structured as follows: Chapter 2 makes a mathematical analysis of how each of the P,C and X parameters (applied separately or combinatorially to a baseline network) can affect the buffer space, bisection bandwidth and scalability (for X parameter). Then illustrates how router micro-architecture adjustments should be done, in order to fairly distribute wires an buffer space without degrading the performance. Chapter 3 presents the whole experimental infrastructure and simulation methodology, along with the detailed NoC architecture configurations. Chapter 4 presents the evaluation results for 64 and 256 PEs in terms of area, performance and energy. Finally, Chapter 5 concludes and provides future research directions.

Chapter 2

Quantitative Analysis of NoC Architectures

This Chapter analyses modifications that could be done on a single 2D mesh network to optimize either a subset or the overall efficiency of a NoC. We categorize these optimizations into 3 parameters. Network Partitioning, Concentration and Express Physical Link insertion. For each parameter that we could apply seperately, firstly are discussed some potential advantages and drawbacks, secondly it is shown how buffer allocation is affected, then it is presented a mathematical framework on how router micro-architecture should be adjusted for overall fair buffer allocation for each changing parameter, and then a unified framework when aplying a combination of those 3 parameters.

2.1 Network Partitioning

Fig. 2.1 depicts a 2D baseline mesh network partitioned into 2 subnetworks. The partitioned network has equal bisection wire count as baseline. This means that router datapath of subnetwork #1 plus router datapath of subnetwork #2 is equal to baseline's router datapath ($Datapath_{base} = Datapath_{partioned} = Datapath_1 + Datapath_2$). In general, we should always keep in mind that for every already converted 2D mesh (the partitioned network here), bisection wire count is equal to baseline 2D mesh (the single 2D mesh network). Below we show some motivations how energy-power efficiency could potentially be achieved by network partitioning.

- A packet could leave its source without requesting arbitration among other message types for each hop traversal, thus saving energy due to the elimination of message priority arbitration and prioritization delay [17]. The final strict message type ordering would be done at the destination node. This could impose less head latency, less circuit logic for router pipelining (pipeline depth would be narrower, as less pipe-stages would be necessary for a head flit to initiate a path), and an injection node could potentially inject

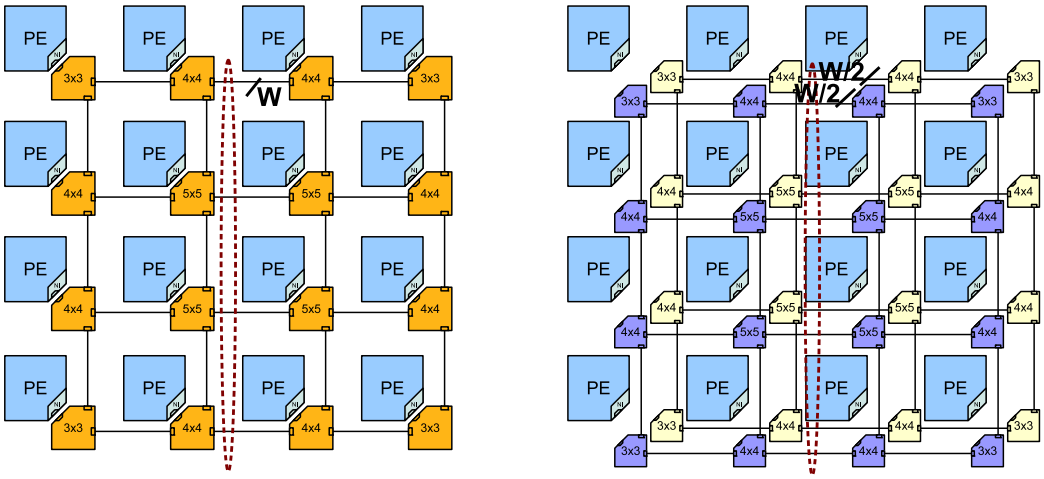


Figure 2.1: A 2D Mesh (baseline) on the left, a partitioned 2D Mesh network on the right. The ellipse shows bisection wire count.

at each cycle, many or all types of messages (packets) having some kind of parallelization. On the other hand, serialization latency would become larger because the partitioned network would have narrower router datapaths. Narrow datapaths are resulted, as previously mentioned, because bisection wire count is equal to both baseline 2D mesh and partitioned network.

- Less overall power dissipation for crossbars. Dynamic power of a crossbar is super-linear relative to width of a crossbar due to super-linear growth of wires [41] interconnecting multiplexors. Any partitioning could give less switching capacitance. Although Passas et. al [41] shows $O(W^{1.5})$ dynamic power complexity in high radix crossbars, for low-radix crossbars, datapath width complexity could be as low as $O(W^{1.1})$. This was confirmed by Post-PnR simulations done in a 65nm [1] technology library of various crossbar configurations for various radixes and datapath widths. Some of the results are shown in table 2.1. In table 2.1 we can observe that partitioning a 4x4 128bit crossbar into 2x64bit crossbars, can give us 9% power savings, For a crossbar with the above width configurations but radix 9x9, power savings could reach 11%. Another scenario is to slice a wide 192bit crossbar into 3 64bit crossbars. For 5x5 radix configuration, power savings could reach 12%, whereas for a 9x9 radix configuration could reach 24%. We can conclude that we can have power savings at around 10% to 20% when slicing a crossbar. The reason of this super-linearity is due to the super-linear growth of wires. Fig. 2.2ii, 2.2iv proves this relationship betwwen total length growth and crossbar width.

Crossbar Dyn. Power (65nm [1]) *pattern:uniform, load:50%, sw. probability:50%*				
Width \ **Radix**	**4x4**	**5x5**	**8x8**	**9x9**
16	.224	.276	.912	1.01
32	.485	.584	1.92	2.13
48	.739	.899	2.94	3.21
64	.987	1.20	4	4.43
96	1.55	1.82	6.36	7.11
128	2.17	2.49	9.04	10.2
192	3.51	4.11	15.2	17.6
256	4.93	5.88	23	26.5

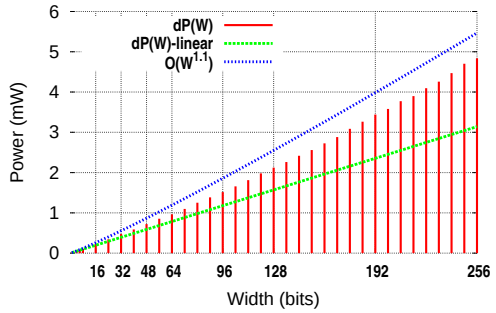
Table 2.1: Post-PnR crossbar dynamic power (65nm [1]) for 4x4, 5x5, 6x6, 7x7, 8x8, 9x9 radixes, and 16, 32, 48, 64, 96, 128, 192, 256 widths.

2.1.1 Homogeneous and Heterogeneous Partitioned Networks

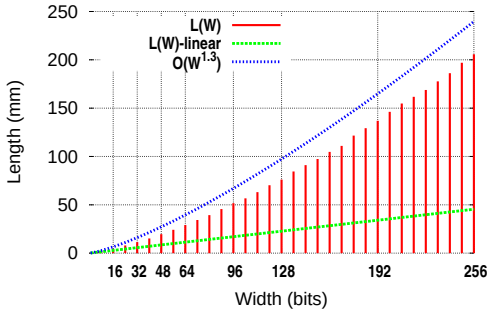
Balfour and Dally [37] first introduced Network Partitioning for 2D meshes. For their baseline network they assume that traffic consists of two classes. (Fig. 2.3a illustrates a 2D mesh with three traffic classes). They assume that each traffic class is based on packet size rather than packet type. So, short packets form one traffic class, and long packets form the other traffic class. Network partitioning was implemented either homogeneously or heterogeneously.

Homogeneously partitioned (HOM): A baseline single 2D mesh network is partitioned into two subnetworks. Each subnetwork implements two traffic classes and both subnetworks have equal bisection wire count. Heterogeneously partitioned (HET): baseline Network is partitioned into two subnetworks (as many as traffic classes)and each subnetwork implements one traffic class.

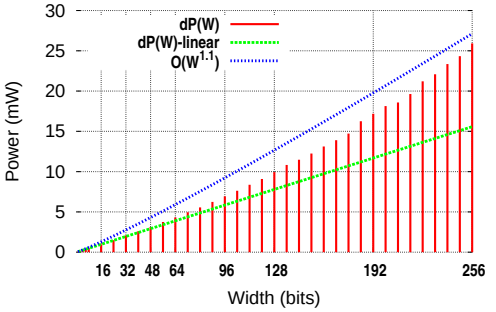
In terms of bisection wire count fairness, we extend Balfour and Dally's scheme for partitioned networks, so that a single 2D mesh network and a resulted 2D mesh partitioned network (Homogeneous or Heterogeneous) have the same bisection wire count. This means that: Each HOM's subnetwork (HOM has two subnetworks) should have half the bisection wire count of baseline 2D mesh. In other words, the router datapath of each subnetwork should be halved. Fig. 2.3b shows such a HOM partitioned network with three traffic classes per subnetwork. In the case of a HET partitioned mesh network, bisection wire count of all HET's subnetworks should be the same as baseline, but each subnetwork's bisection wire count does need to be equal with each other. Bisection wire count of one subnetwork (and hence, router datapath) can be larger than the other. For example, if each traffic class represents packet size, as Balfour and Dally's scheme, short packets could travel through a narrow datapath (small bisection bandwidth subnetwork), and long packets through wide datapath (larger bisection bandwidth subnetwork). Fig.



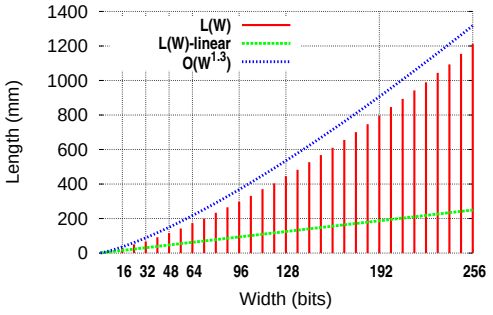
(i) Dynamic power vs. width (4x4 crossbar)



(ii) Link length vs. width (4x4 crossbar)



(iii) Dynamic power vs. width (9x9 crossbar)



(iv) Link length vs. width (9x9 crossbar)

Figure 2.2: Dynamic power and total wire length post-PnR measurements (65nm [1]) of 4x4 and 9x9 crossbars, each configured at 1 to 256 bits width. Red line depicts the absolute numbers of dynamic power in mWs for left figures or total link length in right figures, green line shows how dynamic power (left figures) or total link length (right figures) would have scaled if a linear relationship had been presented between power and width or link length and width respectively. The blue line shows the complexity that was estimated empirically.

2.3d illustrates a HET-partitioned network comprised of three subnetworks.

Boris et. al, Chen et. al [35, 34] used Balfour and Dally HOM scheme. They used 3 traffic classes. Each traffic class was based on packet type instead of packet size. Generally, 2 traffic classes are necessary for message-dependent deadlock avoidance, but if protocol supports forwarding (intervention messages), deadlock will be avoided using at least three message classes [42]. For example, the first traffic class is for response traffic class (high priority), second traffic class is for forward (or intervention) traffic (medium priority) and last traffic class, for request traffic (low priority).

Yoon et. al, Volos et. al [39, 40] used Balfour and Dally HET scheme keeping bisection wire count equal on both the baseline network (the single baseline network that has virtual traffic classes) and the partitioned one (HET). Yoon et. al used four traffic classes to avoid message-dependent deadlock, each having its

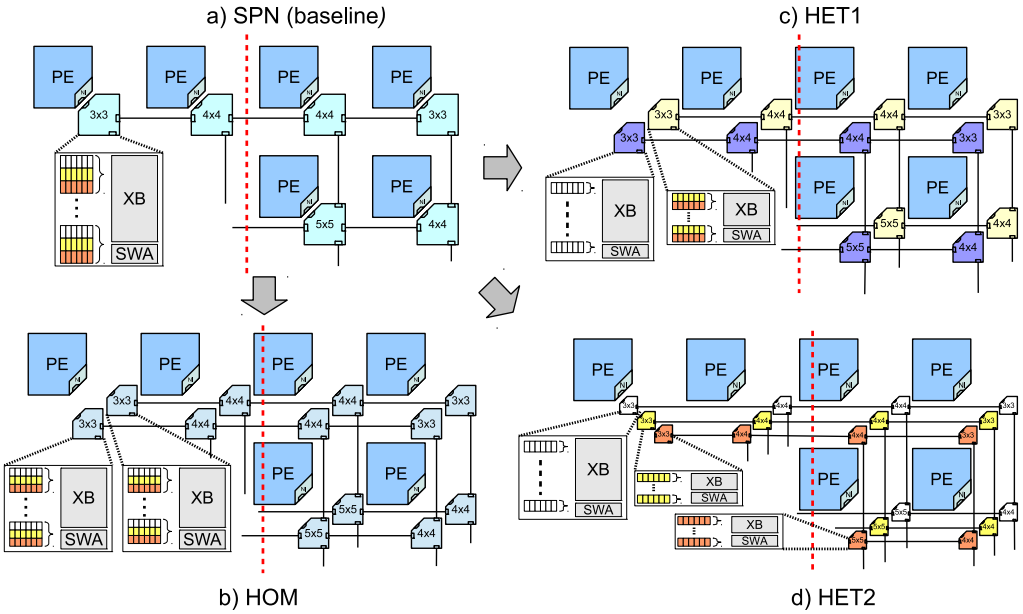


Figure 2.3: a): a single physical 2D mesh network (SPN), unpartitioned and unconcentrated (baseline) using three traffic classes. b): a homogeneously partitioned 2D mesh network with three traffic classes per subnetwork (HOM). c): a heterogeneously partitioned 2D mesh network with variable traffic classes per subnetwork (HET1). d): a heterogeneously partitioned 2D mesh network with one traffic class per subnetwork (HET2).

subnetwork. Volos et. al among the 3 traffic classes that they used, one traffic class was devoted to its own subnetwork and remaining two to the other subnetwork. Fig. 2.3c illustrates such effort.

HET partitioned networks can thus form two types. One of Yoon et. al and the other of Volos et. al. To distinguish between the two heterogeneous architectures we will refer to Volos et. al. as HET1 and as HET2 for Yoon et. al. For all NoCs we assume 3 traffic classes similar to [42].

2.1.2 Bisection Wire Count and Buffer Allocation

Except for Balfour and Dally's [37] effort, bisection wire count was adjusted by [35, 34, 39, 40] to be equal in both baseline and partitioned networks, for fair comparison among NoC architectures

In terms of Buffer allocation, by converting a single baseline 2D mesh into a HOM network while keeping the total HOM's bisection wire count equal to baseline, total HOM's buffer allocation is equal to baseline. However, this is not true for heterogeneous partitioning. As we prove in Sec. 2.1.3 on Pg. 18 and Sec. 2.1.3 on Pg. 21, the resulted HET network will end up having with less buffer resources. Yoon et. al [39] observed this unfairness between a baseline 2D mesh and a heterogeneous partitioned network and carefully fixed total baseline network buffer

allocation to be equal to the partitioned network increasing the depth of input buffers.

A similar observation was made by Chen et. al [34]. Although Chen et. al used HOM networks for evaluation, they adjusted total baseline network buffer allocation for the express physical networks (increasing buffer depth). As we will prove in Sec. 2.3.3 on Pg. 33, by converting a single 2D baseline mesh network into an express physical network, converted network has less buffer allocation in proportion to baseline up to 67%.

2.1.3 Resource Adjustment

In order to adjust buffer allocation, one can do this with two ways. Either adjusting the number of virtual channels (VCs) per message class or adjusting buffer depth. Buffer Width could be adjusted too, but buffer width is restricted by bisection wire count.

For example, let a baseline 2D mesh network of 64bit datapath width (buffer width equal to 64 bit), a HOM partitioned network should have 2 subnetworks of 32 bits each. Thus, each router buffer would have half the baseline's buffer width. HOM partitioned networks have the advantage of not needing any buffer allocation adjustment, so, there is no need to modify any router micro-architecture parameter, such as number of VCs, or buffer depth. HOM partitioned networks have the same buffer allocation as baseline 2D-mesh. This is due to the fact that each HOM subnetwork keeps the baseline's number of virtual traffic classes and buffer depth. On the other hand, HET partitioned networks don't keep buffer allocation equal to baseline. To do so, someone needs to increase buffers in some way.

2D mesh baseline NoC Buffer allocation can be computed as follows:

$$BUF_{base} = RTs \times avgP \times MSG \times VC \times FD \times FW$$

Where:

RTs: is the total number of routers per (sub)network,

avgP: is number of ports per router on average,

MSG: is numbers of message classes (same as traffic class) per port per (sub)network,

VC: is virtual channels per message class (or traffic class),

FD: is flit depth, which meanns the number of buffers per virtual channel and

FW: is flit width in bits per buffer.

The next paragraphs will show how much of total buffer is allocated when converting a baseline network into a HOM, HET1 or HET2 partitioned network under equal bisection wire count among baseline and partitioned networks.

HOM Partitioned Network

Total buffer allocation of a HOM partitioned network is equal to:

$$BUF_{HOM} = \sum_{net=1}^{2} RTs_{net} \times avgP_{net} \times MSG_{net} \times VC_{net} \times FD_{net} \times FW_{net} \tag{2.1}$$

In HOM partitioned networks each subnetwork has half the bisection wire count of the baseline network (router datapaths are equally halved, hence, $FW_1 = FW_2 = FW/2$), and router micro-architecture is kept equal at both subnetworks, so MSG, VC, FD have the same value as baseline. Moreover average ports per router ($avgP$) and the total number of routers (RTs) of each subnetwork are equal to the baseline 2D mesh. So, the above equation, Eq. (2.1) becomes:

$$\begin{aligned} BUF_{HOM} &= \sum_{net=1}^{2} RTs \times avgP \times MSG \times VC \times FD \times \frac{1}{2} \times FW \\ &= RTs \times avgP \times MSG \times VC \times FD \times \frac{1}{2} \times FW \\ &\quad + RTs \times avgP \times MSG \times VC \times FD \times \frac{1}{2} \times FW \\ &= \frac{1}{2} \times BUF_{base} + \frac{1}{2} \times BUF_{base} \\ &= BUF_{base} \end{aligned}$$

Generally, $BUF_{HOM} = BUF_{base}$ is true for a variable bisection wire count per subnetwork and bisection wire count between baseline and partitioned network ($FW = FW_1 + FW_2$). For example if $FW_1 = \mu \times FW$, $FW_2 = (1-\mu) \times FW$ and $\mu = (0, 1)$, we will have that:

$$\begin{aligned} BUF_{HOM} &= RTs \times avgP \times MSG \times VC \times FD \times \mu \times FW \\ &\quad + RTs \times avgP \times MSG \times VC \times FD \times (1-\mu) \times FW \\ &= \mu \times BUF_{base} + (1-\mu) \times BUF_{base} \\ &= BUF_{base} \end{aligned}$$

We have just proved that for fair buffer allocation between a baseline 2D-mesh and HOM partitioned network, under the same bisection wire count at both networks (baseline and HOM), there is no need to adjust any micro-architectural parameter such as VC or FD. $BUF_{base} = BUF_{HOM}$ will always be true. Next, we will see what happens with Heterogeneous Network Partitioning.

HET1 Partitioned Network

For HET1 (among 3 traffic classes, one is implemented on the first subnetwork and the other two on the second subnetwork, Sec. 2.1.1 on Pg. 15), total buffer storage is computed as follows:

$$BUF_{HET1} = \sum_{net=1}^{2} RTs_{net} \times avgP_{net} \times MSG_{net} \times VC_{net} \times FD_{net} \times FW_{net}$$

Average ports per router and total router count of each subnetwork of HET1 is equal to baseline 2D mesh. So BUF_{HET1} becomes:

$$\begin{aligned} BUF_{HET1} = {} & RTs \times avgP \times MSG_1 \times VC_1 \times FD_1 \times FW_1 \\ & + RTs \times avgP \times MSG_2 \times VC_2 \times FD_2 \times FW_2 \end{aligned} \tag{2.2}$$

Supposing that we use the same flit dept and virtual channel per message class on both subnetworks ($FD_1 = FD_2 = FD_{HET1}$, $VC_1 = VC_2 = VC_{HET1}$) of HET1 partitioned networks, equation Eq. (2.2) becomes:

$$\begin{aligned} BUF_{HET1} = {} & (RTs \times avgP \times VC_{HET1} \times FD_{HET1}) \\ & \times (MSG_1 \times FW_1 + MSG_2 \times FW_2) \end{aligned}$$

To explore a baseline's 2D mesh Network's buffer allocation in proportion to a HET1's partitioned network's buffer allocation, we can take ratio α be:

$$\begin{aligned} \alpha &= \frac{BUF_{HET1}}{BUF_{base}} \\ &= \frac{(RTs \times avgP \times VC_{HET1} \times FD_{HET1}) \times (MSG_1 \times FW_1 + MSG_2 \times FW_2)}{RTs \times avgP \times MSG \times VC \times FD \times FW} \\ &= \frac{(VC_{HET1} \times FD_{HET1}) \times (MSG_1 \times FW_1 + MSG_2 \times FW_2)}{MSG \times VC \times FD \times FW} \end{aligned}$$

When converting a baseline mesh into a HET1 network, we can keep virtual channels (VC) per message class and flit depth (FD) of baseline same as the HET1 partitioned network ($VC_{HET1} = VC_{base}$ and $FD_{HET1} = FD_{base}$) as Volos et. al [40]. In other words, we keep architectural parameters (VC_{HET1}, FD_{HET1}) of the partitioned network unadjusted and check if total buffer allocation between baseline and partitioned network are equal, using the α ratio. If α ratio becomes less than one, total buffer allocation in the partitioned network will be less than the baseline's, otherwise, greater if $\alpha > 1$ or equal if $\alpha = 1$. We have:

$$\alpha = \frac{MSG_1 \times FW_1 + MSG_2 \times FW_2}{MSG \times FW}$$

Due to the equal bisection wire count of both baseline and HET1 partitioned network ($FW = FW_1 + FW_2$) and the fact that one traffic class is implemented

for subnetwork 1 and two traffic classes the other subnetwork ($MSG = MSG_1 + MSG_2$, $MSG = 3$,$MSG_1 = 1$,$MSG_2 = 2$) α becomes as follows:

$$\alpha = \frac{MSG_1 \times FW_1 + MSG_2 \times FW_2}{(MSG_1 + MSG_2) \times (FW_1 + FW_2)} \tag{2.3}$$

$$= \frac{(MSG_1 \times FW_1 + MSG_2 \times FW_2)}{(MSG_1 \times FW_1 + MSG_2 \times FW_2) + (MSG_1 \times FW_2 + MSG_2 \times FW_1)} \tag{2.4}$$

for κ and λ as follows:

$$\kappa = MSG_1 \times FW_1 + MSG_2 \times FW_2$$
$$\lambda = MSG_1 \times FW_2 + MSG_2 \times FW_1$$

Eq. (2.4) becomes:

$$\alpha = \frac{\kappa}{\kappa + \lambda}$$

We have proven that always $\alpha < 1$. This means that we will always have less buffer allocation in HET1 partitioned networks if we keep bisection wire count and other architectural parameters such as VC per message class (VC) and flit dept (FD) equal in both networks (baseline and HET1). In order to adjust HET1's buffer allocation to be the same as the baseline ($a = 1$), we could either adjust flit depth (FD) or number of VC per message class (VC). Below there is an example, that we adjust HET1 partitioned network, by modifying flit depth (FD).

Example: For buffer allocation equality adjustment, equation $BUF_{base} = BUF_{HET1}$ needs to be true. This results in,

$$MSG \times VC \times FD \times FW = \sum_{net=1}^{2} MSG_{net} \times VC_{net} \times FD_{net} \times FW_{net} \tag{2.5}$$

solving for FD, we have:

$$FD = \sum_{net=1}^{2} \frac{MSG_{net} \times VC_{net} \times FW_{net}}{MSG \times VC \times FW} \times FD_{net}$$
$$= (\frac{MSG_1 \times VC_1 \times FW_1}{MSG \times VC \times FW}) \times FD_1 + (\frac{MSG_2 \times VC_2 \times FW_2}{MSG \times VC \times FW}) \times FD_2$$

In each HET1's subnetwork we can use the same flit depth, ($FD_1 = FD_2 = FD_{HET1}$), making the above equation:

$$FD = \frac{MSG_1 \times VC_1 \times FW_1 + MSG_2 \times VC_2 \times FW_2}{MSG \times VC \times FW} \times FD_{HET1} \tag{2.6}$$

The baseline network is assumed to have 3 traffic classes, so MSG is equal

to 3 (Sec. 2.1.1 on Pg. 15). HET1 subnetwork 1 has one traffic class for response traffic and subnetwork 2 has two traffic classes (MSG_1 equal to 1 and MSG_2 equal to 2). One for intervention traffic and the another for request traffic (Fig. 2.3c). Moreover, let baseline has one VC per message class ($VC = 1$) and in HET1 network has one VC per message class too ($VC_1 = VC_2 = 1$). Suppose baseline's datapath is $FW = 64bit$ and HET1 partitioned networks' datapats are: $FW_1 = 32bit$, $FW_2 = 32bit$

We have chosen $FW_1 = 32bit, FW_2 = 32bit$ instead of an asymmetric partition like $FW_1 = 48bit, FW_2 = 16bit$ or $FW_1 = 40bit, FW_2 = 24bit$ and so forth, to show that Heterogeneous partitioning doesn't distinguish exclusively from datapath cut. A Heterogeneous architecture HET1 could possibly have two subnetworks with equal data-path each. We proceed leaving flit depth as the only free parameter, and apply the above numbers in equation 2.6:

$$FD = \frac{1 \times 1 \times 32 + 2 \times 1 \times 32}{3 \times 1 \times 64} \times FD_{HET1}$$
$$FD = \frac{FD_{HET1}}{2}$$

In order to keep bisection wire count and buffer allocation equal at both networks with equally halved datapaths, flit depth (FD) of HET1 should be 2x the baseline's 2D mesh network flit depth.

VC per message class (VC) and Flit depth (FD) parameters in a baseline network, are proportional to the total network's buffer allocation. The same holds for HET1 Partitioned Network (Eq. (2.2)). Thus, solving for FD, we can see how much of the total buffer is allocated. In the current case, HET1 buffer allocation is 50% of the baseline's.

If we had chosen HET1 partitioned network to have $FW_1 = 48bit, FW_2 = 16bit$ (we could do this if we would like to pass the most traffic through subnetwork 1, comprising of mostly response traffic data packets, and through narrow datapath $FW_2 = 16bit$ request traffic, comprising of short request packets) FD should be equal to:

$$FD = \frac{1 \times 1 \times 48 + 2 \times 1 \times 16}{3 \times 1 \times 64} \times FD_{HET1}$$
$$= 0.42 \times FD_{HET1}$$

This means, that the HET1 buffer allocation would become 42% of baseline, even less than with $FW_1 = 32bit, FW_2 = 32bit$ configuration. Also, for equal buffer allocation adjustment, HET1 flit depth FD should be 2.4x times more than baseline's flit depth.

HET2 Partitioned Network

We continue similarly, to see what is HET2 partitioned network's total buffer allocation in proportion to 2D mesh baseline. Ending up that ($a < 1$), buffer allocation for HET2 is less than baseline 2D Mesh. So we have:

$$BUF_{HET2} = \sum_{net=1}^{3} RTs_{net} \times avgP_{net} \times MSG_{net} \times VC_{net} \times FD_{net} \times FW_{net} \tag{2.7}$$

As before, average ports per router and total routers of each subnetwork of HET2 is equal to baseline 2D mesh. So BUF_{HET2} becomes:

$$\begin{aligned} BUF_{HET2} = {} & RTs \times avgP \times MSG_1 \times VC_1 \times FD_1 \times FW_1 \\ & + RTs \times avgP \times MSG_2 \times VC_2 \times FD_2 \times FW_2 \\ & + RTs \times avgP \times MSG_3 \times VC_3 \times FD_3 \times FW_3 \end{aligned}$$

Supposing that we use the same flit depth and virtual channel per message class at each HET2's subnetwork ($FD_1 = FD_2 = FD_3 = FD_{HET2}$, $VC_1 = VC_2 = VC_3 = VC_{HET2}$), the above equation is transformed:

$$\begin{aligned} BUF_{HET2} = {} & (RTs \times avgP \times VC_{HET2} \times FD_{HET2}) \\ & \times (MSG_1 \times FW_1 + MSG_2 \times FW_2 + MSG_3 \times FW_3) \end{aligned} \tag{2.8}$$

We evaluate the α ratio using the same methodology as previously for HET1:

$$\alpha = \frac{BUF_{HET2}}{BUF_{base}}$$

and we have:

$$\begin{aligned} \alpha = \frac{(MSG_1 \times FW_1 + MSG_2 \times FW_2 + MSG_3 \times FW_3)}{(MSG_1 \times FW_1 + MSG_2 \times FW_2 + MSG_3 \times FW_3)} \\ +(MSG_1 \times FW_2 + MSG_1 \times FW_3 + MSG_2 \times FW_1 \\ + MSG_2 \times FW_3 + MSG_3 \times FW_1 + MSG_3 \times FW_2) \end{aligned} \tag{2.9}$$

for κ and λ:

$$\begin{aligned} \kappa &= MSG_1 \times FW_1 + MSG_2 \times FW_2 + MSG_3 \times FW_3, \\ \lambda &= MSG_1 \times FW_2 + MSG_1 \times FW_3 + MSG_2 \times FW_1 \\ &\quad + MSG_2 \times FW_3 + MSG_3 \times FW_1 + MSG_3 \times FW_2 \end{aligned}$$

(2.9) becomes:

$$\alpha = \frac{\kappa}{\kappa + \lambda} \tag{2.10}$$

We have just proven that $\alpha < 1$ for HET2 partitioned networks too. This means that we will always have less buffer allocation in HET2 partitioned networks, when we keep bisection wire count and other architectural parameters, such as VC per message class (VC) and flit dept (FD), equal in both networks (baseline and HET2). If we would like to adjust HET2's buffer allocation to be same as baseline ($a = 1$), we could either adjust flit depth (FD) or the number of VC per message class (VC). Below there is an an example, that we can adjust a HET2 network, by modifying the number of virtual channels (VC) per message class.

Example: For buffer allocation equality adjustment, equation $BUF_{base} = BUF_{HET2}$ needs to be true. This results in,

$$MSG \times VC \times FD \times FW = \sum_{net=1}^{3} MSG_{net} \times VC_{net} \times FD_{net} \times FW_{net} \quad (2.11)$$

solving for VC, we have:

$$\begin{aligned} VC &= \sum_{net=1}^{3} \frac{MSG_{net} \times FD_{net} \times FW_{net}}{MSG \times VC \times FW} \times VC_{net} \\ &= (\frac{MSG_1 \times FD_1 \times FW_1}{MSG \times VC \times FW}) \times VC_1 \\ &\quad + (\frac{MSG_2 \times FD_2 \times FW_2}{MSG \times FD \times FW}) \times VC_2 \\ &\quad + (\frac{MSG_3 \times FD_3 \times FW_3}{MSG \times FD \times FW}) \times VC_3 \end{aligned}$$

In each HET2's subnetwork we can use the same number of virtual channels per message class ($VC_1 = VC_2 = VC_3 = VC_{HET2}$) and the above equation that gives VC, becomes:

$$VC = \frac{\begin{array}{c} MSG_1 \times FD_1 \times FW_1 \\ +MSG_2 \times FD_2 \times FW_2 \\ + MSG_3 \times FD_3 \times FW_3 \end{array}}{MSG \times FD \times FW} \times VC_{HET2} \quad (2.12)$$

baseline 2D-mesh Network has 3 traffic classes ($MSG = 3$), (Sec. 2.1.1 on Pg. 15). HET2 subnetwork 1 has one traffic class for response traffic, HET2 subnetwork 2 has one traffic class for intervention traffic, and HET2 subnetwork 3 has one traffic class for responce traffic ($MSG_1 = MSG_2 = MSG_3 = 1$, Fig. 2.3d). Moreover, suppose baseline's flit depth is equal to 6 ($FD = 6$) and HET2's flit depth is equal to 6 for each subnetwork ($FD_1 = FD_2 = FD_3 = 6$). Finally, let baseline's datapath be $FW = 64bit$ and HET2's datapaths be: $FW_1 = 32bit$, $FW_2 = FW_3 = 16bit$. We proceed leaving VC as the only free parameter and

apply above numbers in equation 2.12:

$$VC = \frac{1 \times 6 \times 32 + 1 \times 6 \times 16 + 1 \times 6 \times 16}{3 \times 6 \times 64} \times VC_{HET2}$$
$$VC = \frac{VC_{HET2}}{3}$$

In order to have bisection wire count and buffer allocation equal to both baseline and HET2 network, HET2's virtual channels (VC_{HET2}) should be 3x of the baseline's virtual channels VC.

Any of the baseline's parameters such as VC per message class (VC) and flit depth (FD), are proportional to the buffer allocation. The same holds for the HET2 (Eq. (2.8)). Thus, solving for VC, we can see how much of the total buffer is allocated. In the current case, HET2 buffer allocation is 33.3% of the baseline's.

2.2 Concentration

A baseline mesh network, converted into a concentrated mesh is depicted in Fig. (2.4). The resulted concentrated mesh network has equal bisection wire count to the baseline. This means that the concentrated mesh network's router datapath should be multiplied by a factor of σ. In current case $\sigma = 2$. For example EPN's router datapath should be:

$$\begin{aligned} Datapath_{EPN} &= \sigma \times Datapath_{base} \\ &= 2 \times Datapath_{base} \end{aligned}$$

Below we show some intuitions of what would potentially happen when concen-

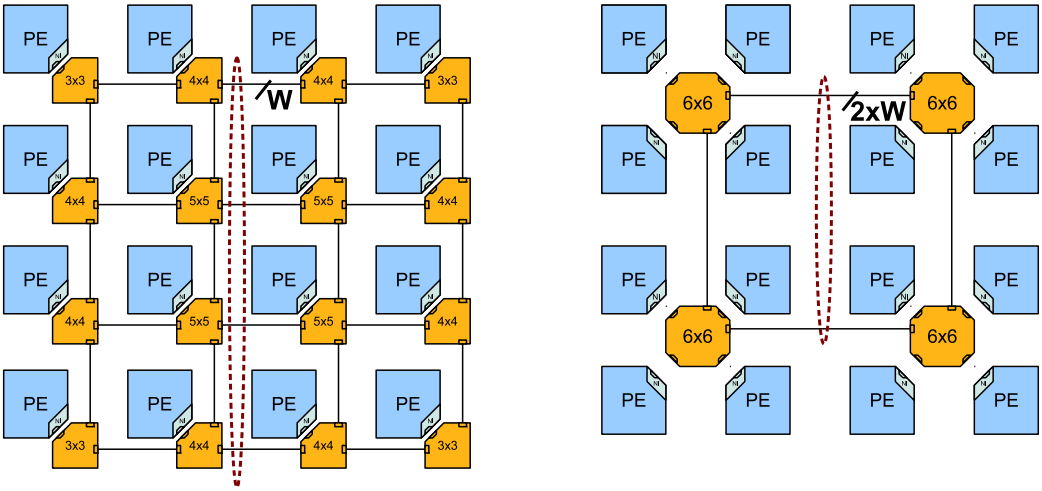


Figure 2.4: A baseline 2D mesh on the left, a 2D mesh concentrated network on the right. Bisection wire count is illustrated inside the ellipse.

trating a 2D baseline mesh network. Benefits and drawbacks are discussed too.

- Router traversals will be definitely less. A packet has now less routers to pass through, regardless of traffic pattern. A packet could reach its destination without crossing many router pipelines and so head latency can potentially become smaller (as long as router pipeline depth is more than one cycle). Serialization latency may become smaller, because in order to keep bisection wire count the same as the baseline, wider router datapath is required. This comes at a price of the additional energy overhead that may be introduced, as wider datapaths dissipate super-linearly more crossbar dynamic power (Fig. 2.2 Table 2.1).

- Total router count is less than baseline's. This may affect total size of control logic that manipulates buffers. Few large buffers instead of many small buffers may introduce less total buffer control logic. Although a concentrated network consists of less routers, more ports per router are needed to connect additional PEs. Thus, each crossbar traversal will cost approximately $(\frac{P_{in}+C}{P_{in}})^2$ times more energy, due to quadratic dependence between crossbar dynamic power and ports. Total allocator's area is affected too, as it increases quadratically with the number of ports. When the request width of each arbiter increases linearly with respect to the number of ports, and the total number of arbiters is proportional to the number of ports, allocator area can increase quadratically with the number of ports.

- Communication among neighbors connected to the same router results in a single router traversal, but each traversal costs more. In addition, temporary neighbor communication may under-utilize non-neighbor nodes, and the whole network performance will degrade. For example, if many local requests content a single router, remote requests that need to pass through the already contented router, will be under-utilized.

- Larger router-to-router links may be introduced. An X dimension distance for example, will be sliced by less number of routers, with potential insertion of link pipelining too. Link pipelining can introduce additional buffer requirements in terms of buffer depth in order to cover flow control's round trip time (RTT). In addition, if large degree of concentration has been applied, neighbor requests may need to travel long wire distances.

2.2.1 Buffer and Bisection Wire Count Allocation

Among concentrated network designs [37, 35], bisection wire count in concentrated network schemes were carefully adjusted by Boris et. al [35] to be equal to the baseline 2D Mesh for fair comparison. However, in terms of buffer allocation in concentrated networks, neither of [37, 35] made such adjustment. We will see in the following section (Sec. 2.2.2 on Pg. 26) that when a 2D mesh is converted to a concentrated mesh, buffer allocation ranges from 30% up to 21%.

2.2.2 Resource Adjustment

As previously described, a baseline NoC's buffer allocation is:

$$BUF_{base} = RTs \times avgP \times MSG \times VC \times FD \times FW \tag{2.13}$$

In addition, in Sec. 2.2 on Pg. 23, we saw σ and when concentration degree is 4 (four PEs per router), σ needs to become 2 because this is the only way for concentrated network's bisection wire count to be equal to the baseline's. Thus, σ depends on concentration degree. For example, if a 6x6 2D mesh network is concentrated by a degree of 9 (4 clusters, each cluster comprised of 3x3 mesh) σ should be equal to 3. Which means that the converted concentrated network should have 3× wider router datapaths relative to the baseline. In general, if concentration degree is C and each cluster has $\sqrt{C} \times \sqrt{C}$ routers, σ should simply be: $\sigma = \sqrt{C}$. When there is no concentration, C is equal to 1 ($C = 1$).Total buffer allocation for concentrated network is:

$$BUF_{con} = RTs_{con} \times avgP_{con} \times MSG_{con} \times VC_{con} \times FD_{con} \times (\sigma_{con} \times FW) \tag{2.14}$$

Total number of routers (RTs) could also be written as: $RTs = \frac{PEs}{C}$ thus, equations (2.13),(2.14) become:

$$BUF_{base} = \frac{PEs}{C} \times avgP \times MSG \times VC \times FD \times FW$$

$$BUF_{con} = \frac{PEs}{C_{con}} \times avgP_{con} \times MSG_{con} \times VC_{con} \times FD_{con} \times (\sigma_{con} \times FW) \tag{2.15}$$

Supposing that we use the same flit dept and virtual channel per message class on both the baseline and concentrated network ($FD = FD_{con}$, $VC = VC_{con}$) and traffic classes are the same in both networks ($MSG = MSG_{con}$) as well, Eq. (2.15) becomes:

$$BUF_{con} = \frac{PEs}{C_{con}} \times avgP_{con} \times MSG \times VC \times FD \times (\sigma_{con} \times FW)$$

To explore baseline's buffer allocation in proportion to concentrated network's

buffer allocation, we take ratio α to be:

$$\begin{aligned}\alpha &= \frac{BUF_{con}}{BUF_{base}} \\ &= \frac{\frac{PEs}{C_{con}} \times avgP_{con} \times MSG \times VC \times FD \times (\sigma_{con} \times FW)}{\frac{PEs}{C} \times avgP \times MSG \times VC \times FD \times FW} \\ &= \frac{C \times avgP_{con} \times \sigma_{con}}{C_{con} \times avgP}\end{aligned}$$

concentration degree of the baseline is always 1 ($C = 1$) and $\sigma_{con} = \sqrt{C_{con}}$, thus:

$$\alpha = \frac{avgP_{con}}{avgP} \times \frac{1}{\sqrt{C_{con}}} \tag{2.16}$$

It isn't clear if $\alpha < 1$ or $\alpha = 1$ or $\alpha < 1$. But we can compute α for different PEs values and concentration degrees. Equation (2.16) that gives α, has 3 variables.

- average ports per router ($avgP$) for 2D mesh baseline (where $C = 1$),
- $avgP_{con}$ for the concentrated network and
- concentration degree C_{con} of the concentrated network.

The concentration degree of each network is known, hence, before we compute α we need to compute $avgP$ for each network. Table 2.2 below, shows $avgP$ values. Knowing the average port per router parameter ($avgP$) ratio α is easy to compute.

Average ports per router (avgP)

C \ PEs	**16**	**36**	**64**	**81**	**144**	**256**
1	4	4.34	4.5	4.56	4.67	4.75
4	6	6.67	7	×	7.34	7.5
9	×	11	×	11.67	12	×
16	×	×	18	×	18.67	19

Table 2.2: Average ports per router ($avgP$) for various concentration degrees (1, 4, 9 and 16) and number of PEs (16, 32, 64, 81, 144 and 256).

Below are the α ratios resulting from $avgP$ parameters of table 2.2. Table 2.3 tells us that the α ratio is less than than one for 16,36,64,81,144,256 PEs and concentration degrees of 4 and 9, and α is equal to one for concentration degree equal to 16. We can conclude that, for a concentrated network of $C = 4$, buffer allocation will be up to 21% less than a baseline network, when keeping bisection wire count and other architectural parameters (VC and FD) equal to both networks .

Ratio α

C \ PEs	**16**	**36**	**64**	**81**	**144**	**256**
4	0.75	0.77	0.78	×	0.79	0.79
9	×	0.84	×	0.85	0.86	×
16	×	×	1	×	1	1

Table 2.3: Ratio α for various concentration degrees (1, 4, 9, 16) and number of PEs (16, 32, 64, 81, 144, 256).

When transforming a baseline network into a concentrated one, although the concentrated network's router datapaths are increased (multiplied by σ) to keep bisection wire count equal, buffer allocation is still less than the baseline. Evaluation of networks need to consider this buffer allocation impact for fair comparison. To keep concentrated network's buffer allocation the same as the baseline's ($a = 1$), we could either adjust flit depth (FD) or number of VC per message class (VC) of the concentrated network. Below, an example is illustrated where adjustment to the concentrated network is made, by modifying only flit depth (FD).

Example: For buffer allocation equality adjustment, equation $BUF_{base} = BUF_{con}$ needs to be true. This results in,

$$\frac{1}{C} \times avgP \times FD = \frac{1}{C_{con}} \times avgP_{con} \times FD_{con} \times \sigma_{con}$$

solving for FD_{con}, we have:

$$FD_{con} = (\frac{C_{con}}{C} \times \frac{avgP}{avgP_{con}} \times \frac{1}{\sigma_{con}}) \times FD$$

Assume we have a 64-node 2D mesh baseline network, with each PE connected to a single router. We convert the baseline to a concentrated network of $C = 4$. The resulted concentrated network will have 16 nodes, where each 4-PE cluster is connected to a single router as shown in Fig. 2.5. Furthermore, let the baseline flit depth be 7. Then the concentrated network's flit depth should be:

$$\begin{aligned} FD_{con} &= (4 \times \frac{4.5}{7} \times \frac{1}{2}) \times 7 \\ &= 1.29 \times 7 \\ &= 9 \end{aligned}$$

2.3 Express Physical Links

Considering large-scale 2D mesh Networks, router traversals are scaled proportionally to the network diameter. One way to reduce average hop count on large

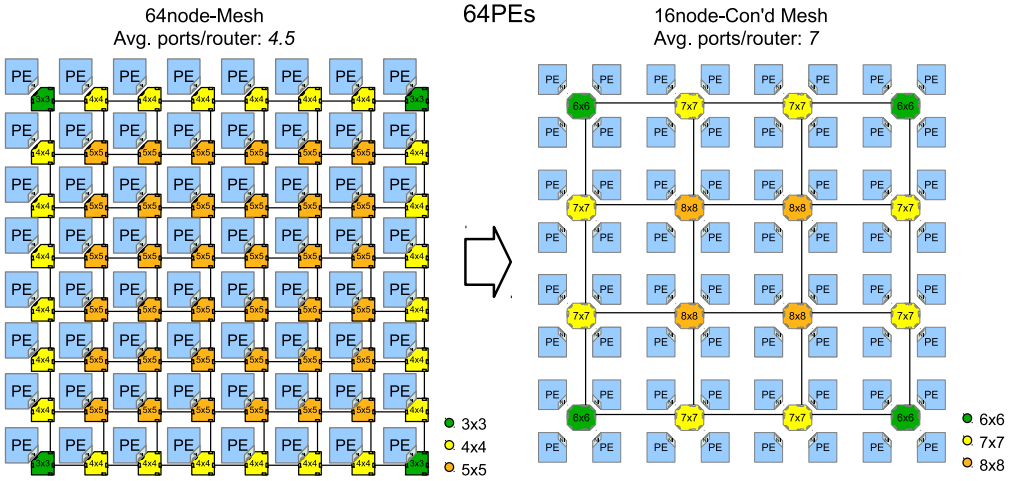


Figure 2.5: A 64-node 2D mesh converted into a 16-node concentrated mesh network.

scale mesh networks is by inserting express physical links. A 2D baseline mesh network, converted into an express physical network using Chen et. al scheme [34] is depicted in Fig. 2.6. The resulted express physical network (EPN) has equal bi-section wire count as the baseline. This means that EPN's router datapath should be divided by a factor of ρ. Where ρ is the bisection wire reduction factor. In the current case $\rho = 2$. Thus:

$$\begin{aligned} Datapath_{EPN} &= \frac{1}{\rho} \times Datapath_{base} \\ &= \frac{1}{2} \times Datapath_{base} \end{aligned}$$

Below we show some intuitions how energy-power efficiency could potentially be achieved inserting express physical links on a baseline 2D mesh network. Except from potential benefits, we show potential drawbacks too.

- A packet could have an alternative path to leave as long as an express path exists and dimension distance is greater or equal to express interval (express interval is equivalent to local hops. For example an express interval of 2 means two local hops per express hop). This imposes less average hop count per packet with less router traversals. Assuming a constant destination for a packet, it would be more energy efficient if the packet could exclusively use only wire paths. Router traversal adds extra cost because we have to pay for buffering, arbitration and crossbar traversal.

- A head flit will not traverse the whole router pipeline. Less router traversals means less total pipe-stage traversals improving head latency (less overall head latency in proportion to baseline). Reducing head latency is very im-

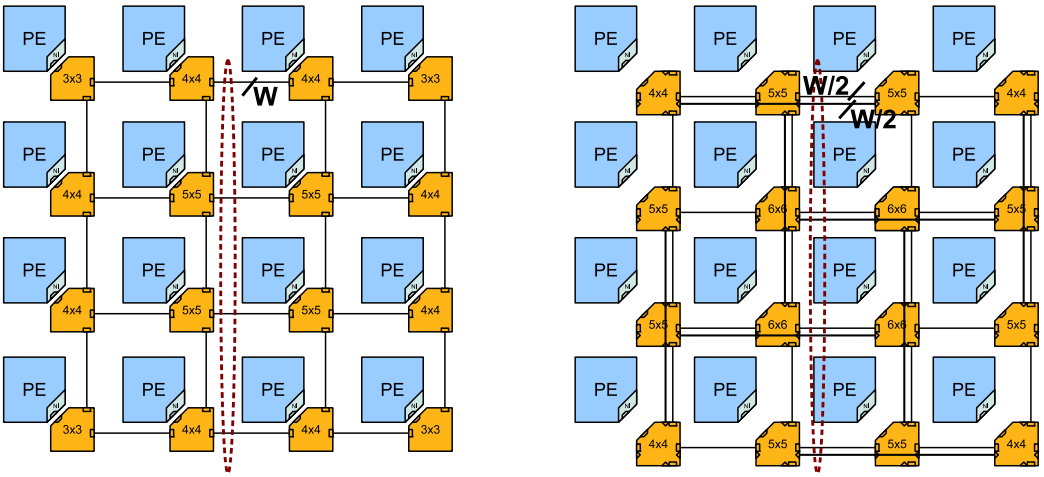


Figure 2.6: A baseline 2D mesh on the left, a 2D mesh express physical network (EPN) on the right. Bisection wire count is illustrated inside the ellipse. The network on the right should have the half datapath width of baseline's, so that the wires inside the ellipse be equal to both networks.

portant, especially when it has stronger effect than serialization latency. This is true for for short packets and packets that need to travel far away from the source (many router traversals).

On the other hand, depending on the scheme of express links (how topologically express links could interconnect routers), extra input ports would have to be added to the routers affecting all of the following parameters directly or indirectly:

- Crossbar complexity grows quadratically ($O(N^2)$) with number of input ports(N).
- For each express link crossbar power increases by approximately $(\frac{p_{in}+1}{p_{in}})^2$ times.
- Bisection wire count becomes greater as more links are crossing from one half side of the network to the other.
- Potential insertion of link pipelining introduces extra pipe stages to existing wire, if clock frequency is wire-limited
- Total Buffer allocation per router could be increased due to two reasons. Firstly, because of input port growth $\frac{P_{in}+1}{P_{in}}$ and secondly due to the buffer depth increment that is possibly required in order to cover Round Trip Time (RTT) in case link pipelining has imposed extra link traversal delay.
- Serialization latency becomes larger as datapath is shrinking, to accommodate with the baseline's bisection bandwidth.

2.3.1 Express Physical Link Schemes and Scalability

This section discusses about schemes that have been used to utilize a 2D mesh network by express physical links. In Fig. 2.7 we see four schemes, for 16 network nodes: Cmesh [37], flatten butterfly (Fbatfly) [36], MECS [35] and Chen et. al's EPN [34].

Dally and Balfour's "Cmesh" is a concentrated mesh with express physical links on peripheral nodes, (it should not be confused with concentrated mesh of thesis's work. A concentrated mesh has not express physical links). Fbatfly is a butterfly network mapped to a 2D mesh like topology, MECS is a Multidrop Express Channel network similar to Fbatfly, but each express physical link is shared as local link too, EPN is a 2D mesh topology with extra physical links interconnected in a special manner (Fig. 2.7). As is easily perceived, each scheme differs in interconnectivity and router micro-architecture. Schemes are described below in same order as in Fig. 2.7.

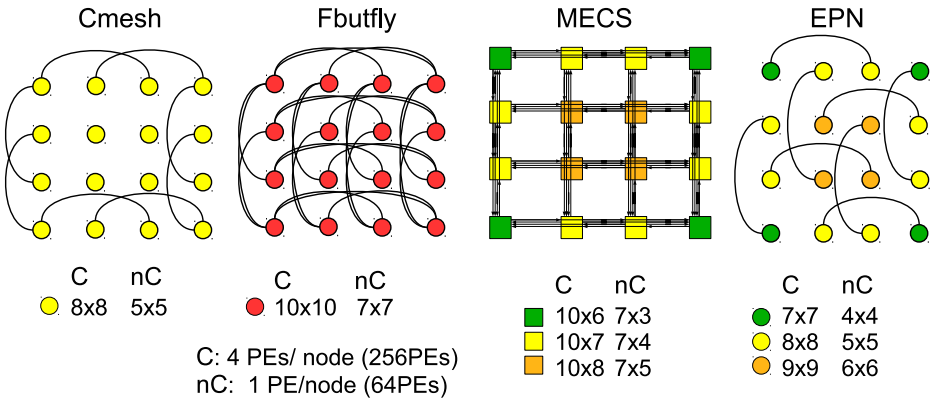


Figure 2.7: Schemes for express physical links interconnecting 16 (4x4) network nodes. Local links are not shown for clarity (except for MECS) for clarity.

Cmesh [37] was introduced as a concentrated 2D mesh topology that interconnects 16 nodes (each node is connected to 4 PEs) with express physical links on peripheral network nodes. By introducing concentrating, a baseline 64-node 2D mesh network diameter could be reduced from 14 to 6. Network diameter could be reduced even more (from 6 to 4) if express physical links were inserted in peripheral nodes. Apart from average hop count reduction, total bisection wire count of 16-node Cmesh topology is the same as in a 64-node 2D mesh, there is no need to increase 16-node Cmesh's router datapath (See section 2.2 on Pg. 23 for bisection wire adjustment).

However, when PE count grows beyond 64 (beyond 16 network nodes), non-peripheral nodes does not have express channels to utilize network. This is easily perceived in Fig. 2.8. Only packets traveling exclusively at peripheral links are utilized though express links. The rest of links are local links making network diameter large. Also, total bisection wire count of 64-node (256PEs) Cmesh has

not the same bisection wire count as a 256-node(256PEs) 2D baseline mesh. Extra peripheral express links of 64-node Cmesh are not enough to equalize the total bisection wire count of a respective 256-node 2D baseline mesh. Router datapath increase should be considered for 64node Cmesh.

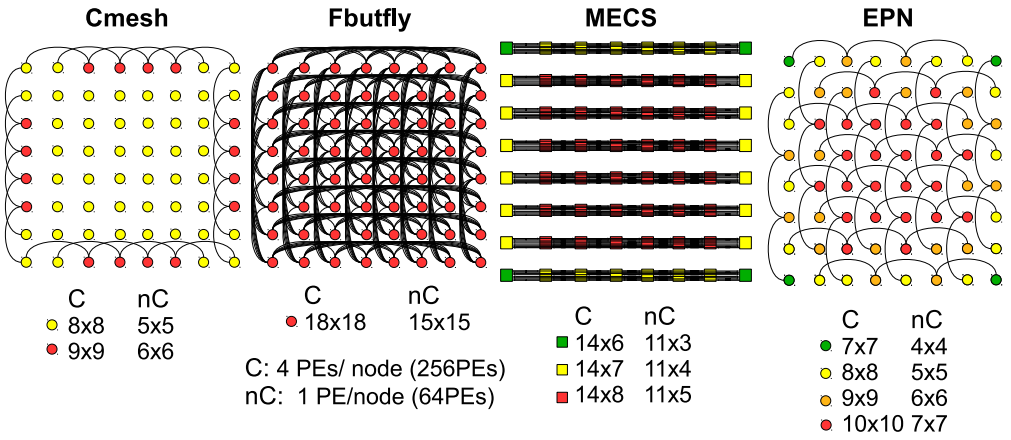


Figure 2.8: Cmesh, FBatfly, MECS and EPN use express physical links interconnecting 64 (8x8) network nodes (64 PEs or 256 PEs if concentration is applied). Local links and MECS's column links are not shown for clarity.

Kim et. al. introduced flattened butterfly [36]. Regardless of network node count ($k \times k$), network diameter of such architecture is always 2. This happens because each network node is directly connected to every other node at the same row/column. Although network diameter is reduced to minimum, as network node count increases, bisection wire count increases by $O(k^2)$ complexity, and crossbar input-output ports by $O(k)$. This has serious effect in scalability. For example, a 8x8 flattened butterfly network, should have additionally $(\frac{1}{4} \times 8^2) - 1 = 15$ bisection bidirectional wires per row compared with a 8x8 2D mesh network, and $2 \times (k-1) - 4 = 10$ more input-output ports than a 2D mesh baseline's 5x5 router.

MECS [35] was an alternative of the flattened butterfly to reduce the complexity of bisection wire count to $O(k)$, crossbar input port growth to $O(k)$ and crossbar output port growth to $O(1)$. For example, a 8x8 MECS network, should have additionally $(\frac{1}{2} \times 8) - 1 = 3$ more bisection bidirectional wires per row than a 8x8 2D mesh network, $2 \times (k-1) - 4 = 10$ more input ports than a 5x5 router and equal number of output ports to a 5x5 router.

Chen et. al [34] uses express physical links with 4-hop express interval (4 local hops per express hop) for 256 nodes, similar to Fig. 2.9. This structure's express physical links can be generalized for an express interval of 2, too (Fig. 2.7,2.8). This kind of express physical network scheme (EPN) has a bisection wire growth complexity independent of network count (or independent of k). It only depends on express interval.

For example, a 4x4 mesh network, when converted to EPN scheme of express

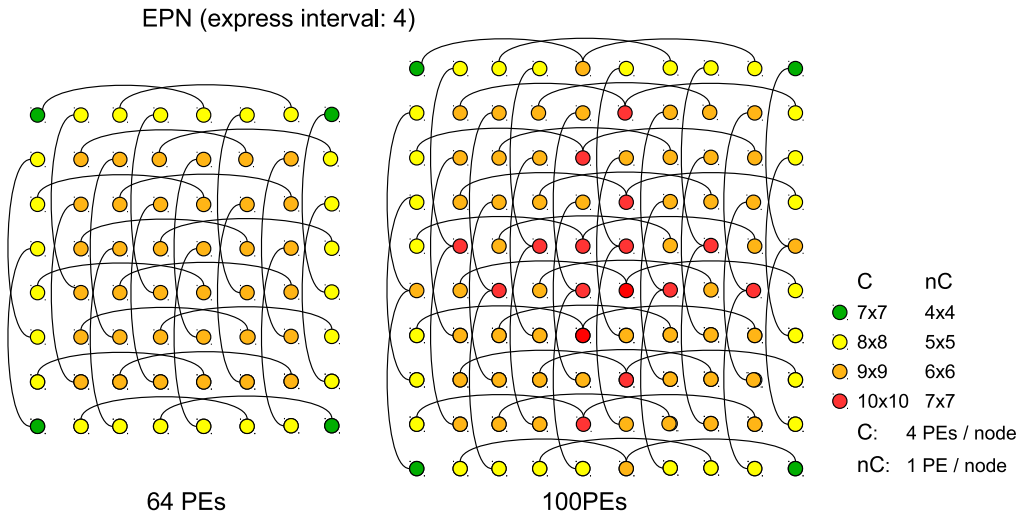


Figure 2.9: Express physical links for 64-PEs and 100-PEs networks with express interval equal to 4.

interval equal to 2, has one extra bisection bidirectional wire per row. This is true for 16x16, 64x64, 100x100 and so forth. In Fig. 2.10 there are three EPNs, of express interval equal to 2, for 36PEs, 64PEs and 100PEs. For each case bisection wire count has been incremented only by one. Total crossbar inputs-outputs per router remain the same as well, regardless of the PE count.

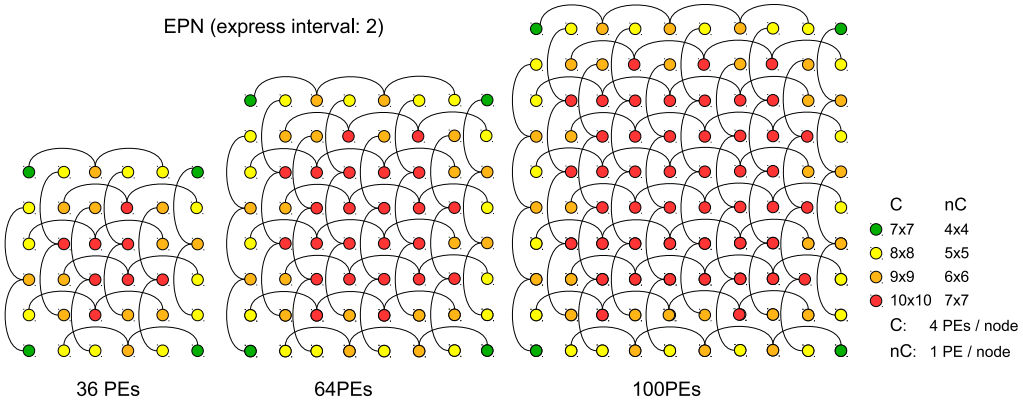


Figure 2.10: Express physical links for 36-PEs, 64-PEs and 100-PEs with express interval equal to 2.

When express interval becomes 4, the maximum additional bisection bidirectional wires per row is 2. Fig. 2.9 shows two networks (64 PEs and 100PEs) with express interval equal to 4. Again, as for express interval equal to 2, for each case of total network nodes, bisection wire count has been incremented by same number (here two instead of one). Total crossbar inputs-outputs per router remain the

same too, regardless of PE count and express interval (See Fig. 2.10, 2.9).

So far, the most scalable scheme for express physical links seems to be EPN. A mathematical proof will be shown later in following subsections.

2.3.2 Bisection Wire Count and Buffer Allocation

Except for Balfour and Dally [37], Boris et. al and Chen et. al [35, 34] carefully adjusted the bisection wire count to be equal in both baseline 2D Mesh and networks that use express physical links for fair comparison.

For example, Boris et. al [35] reduced router datapath by a factor of 2 for MECS 16-node network, and by a factor of 4 for MECS 64-node network, because a 16-node MECS network has 2 times more bisection wires than 64-node baseline (or a 16-node concentrated 2D mesh) and a 64-node MECS network has 4× bisection wires than 256-node baseline (or a 64-node concentrated 2D mesh). Chen et. al. [34] reduced router datapath too by a factor of 3 for 256-node EPN, because a 256-node EPN has 3× bisection wires than a 256-node baseline 2D mesh (or a 64-node concentrated 2D mesh).

On the other hand, Balfour and Dally's (16-node only) Cmesh [37], did not have equal bisection wire count to a (64-node) 2D mesh baseline, despite the fact that the additional peripheral express links of Cmesh were enough to equalize total bisection wire count to the baseline network, Cmesh's bisection wire count was bigger than at the baseline due to wider router datapaths that have been used.

In terms of Buffer allocation only Chen et. al [34] adjusted total baseline network buffer allocation to be equal to the express physical network (EPN), and this done by increasing buffer depth. When a baseline 2D mesh network is being converted into an express-physical network (EPN) forcing express physical network (EPN) to have the same bisection wire count as a 2D mesh baseline, overall buffer allocation of EPN ends up less than then baseline up to 60%. This is explained later in section 2.3.3.

2.3.3 Resource Adjustment

In this section buffer allocation adjustment between a baseline and an express physical network using Chen et. al [34] scheme for express physical links is analyzed. The 2D mesh baseline NoC buffer allocation is:

$$BUF_{base} = RTs \times avgP \times MSG \times VC \times FD \times FW$$

In Sec. 2.3 on Pg. 27, ρ parameter was introduced as the bisection wire reduction factor for express physical networks, when they are converted from a 2D mesh baseline. Furthermore, we saw in Sec. 2.3.1 on Pg. 31, that when converting a 2D mesh into a EPN, additional bisection bidirectional links per row depend only on express interval. If express interval is 2 we have one additional bisection bidirectional link per row, if express interval is 4 we have two additional bisection

bidirectional links per row, regardless of PE count. Let express interval be exp, then ρ could be modeled as:

$$\rho = \frac{1}{\frac{exp}{2} + 1} \tag{2.17}$$

When $exp = 2$, ρ becomes $\rho = \frac{1}{2}$, and when $exp = 4$, $\rho = \frac{1}{3}$. At the first case, router datapath should be $\frac{1}{2}$ times narrower, and at the other, router datapath should be $\frac{1}{3}$ times narrower for fair bisection wire allocation. The equation (2.17) is true for non express networks too. For a 2D mesh baseline, for instance, exp is zero and ρ becomes $\rho = 1$. Total buffer allocation for EPN now becomes:

$$BUF_X = RTs_X \times avgP_X \times MSG_X \times VC_X \times FD_X \times (\frac{1}{\rho} \times FW) \tag{2.18}$$

Supposing that we use the same flit dept and virtual channel per message class on both baseline and EPN ($FD = FD_{con}$, $VC = VC_{con}$) and also $MSG = MSG_X$ $RTs = RTs_X$, as traffic classes and router count is the same in both networks equation Eq. (2.18) becomes:

$$BUF_X = RTs \times avgP_X \times MSG \times VC \times FD \times (\frac{1}{\rho} \times FW)$$

To explore baseline's 2D mesh Network's buffer allocation in proportion to EPN's buffer allocation, we compute the ratio α as:

$$\begin{aligned}
\alpha &= \frac{BUF_X}{BUF_{base}} \\
&= \frac{RTs \times avgP_X \times MSG \times VC \times FD \times (\frac{1}{\rho} \times FW)}{RTs \times avgP \times MSG \times VC \times FD \times FW} \\
&= \frac{avgP_X}{avgP} \times \frac{1}{\rho}
\end{aligned}$$

Chen et. al's EPN scheme, as we saw previously, has the best scalability characteristics. The maximum number of additional input ports per router, regardless of PE count and express interval, is 2. Thus, the worst (not real) case scenario of α could be when EPN's average ports per router is baseline's $avgP$ plus 2 ($avgP+2$):

$$\alpha = \frac{avgP + 2}{avgP} \times \frac{1}{\rho}$$

The only possibility for ratio α to be greater than or equal to one ($\alpha \leq 1$), for $exp = 2$ and $exp = 4$, is when $avgP \leq 2$ and $avgP \leq 1$ respectively. This could never happen, because the smallest possible 2D mesh has always $avgP \geq 3$.

We have proven that, when using Chen et. al's EPN scheme, α will always be less than one. This means than, when converting a 2D mesh baseline into a EPN,

total buffer allocation will always be less than the baseline's, despite the fact that total ports of EPN are more.

A 64-node 2D mesh has $avgP = 4.5$. The respective EPN, for $exp = 2$, has $avgP = 6$ and, for EPN of $exp = 4$, $avgP = 5.5$. Thus, for each case α ratio becomes as follows:

$$\alpha_{exp=2} = \frac{6}{4.5} \times \frac{1}{2} = 0.67 \qquad \alpha_{exp=4} = \frac{5.5}{4.5} \times \frac{1}{3} = 0.41$$

It is easily perceived that total buffer allocation for 64PEs can be from 35% ($exp = 2$) up to 60% less that buffer allocation of 2D mesh baseline. Evaluation of networks need to consider this buffer allocation impact for fair comparison.

For EPN's buffer allocation to be the same as the baseline ($a = 1$), we could adjust flit depth (FD) or number of VC per message class (VC) of EPN. Below there is an example, where a 64-node EPN of $exp = 4$ is adjusted, by modifying virtual channels per message class (VC).

Example: Equation $BUF_{base} = BUF_X$ needs to be true. Solving for VC_X and substituting baseline's VC with 2, results in:

$$\begin{aligned} VC_X &= \frac{avgP \times \rho}{avgP_X} \times VC \\ &= \frac{4.5 \times 3}{5.5} \times 2 \approx 5 \end{aligned}$$

Thus, result shows that a 64-node EPN-4 ($exp = 4$) VC count needs to be 2.5× times more than the baseline's VC count to equalize its buffer allocation.

2.4 Combined NoC Architectures

So far, we saw each adjustment technique separately, where a 2D mesh baseline network was converted into a partitioned network, a concentrated network and an express physical network (EPN). Also, it was proved that when forcing bisection wire count to be equal at both the baseline and the converted network, keeping the rest of router micro-architecture parameters unmodified, buffer allocation could be less than baseline's up to 67% for HET networks, 30% for concentrated networks and 60% for EPNs). This observation was later used to adjust router micro-architectures by increasing VCs and/or FD for fair buffer allocation and performance boost. This memory adjustment could assist the system to: a) Avoid head of line blocking and increase saturation throughput. b) Fix potential turn-around credit delays that express links or/and concentration could introduce and increase latency [43]. This section discusses all possible NoC instances when all three parameters are combined and shows how buffer allocation is affected. Fig. 2.11 depicts this design space exploration.

Before proceeding to buffer allocation analysis, let partitioning parameter be P which can be SPN, HOM, HET1 and HET2, where SPN means that partitioning

has not been applied (Single Physical Network), concentration parameter be C, and express physical link parameter be X, with $X2$ meaning that network has been converted with express physical links of express interval equal to 2 and $X4$ with express interval equal to 4. To refer a combination, firstly place concentration symbol, then express physical symbol and finally partitioning symbol. Thus, we have the following (table 2.4) combinations of converted networks plus baseline (SPN): 24 total cases have been arisen including baseline 2D mesh network. Three

NoC architecture instances

		Not P'ed	**Partitioned**		
Not C'ed	$exp = 0$	SPN	HOM	HET1	HET2
	$exp = 2$	X2-SPN	X2-HOM	X2-HET1	X2-HET2
	$exp = 4$	X4-SPN	X4-HOM	X4-HET1	X4-HET2
C'ed	$exp = 0$	C-SPN	C-HOM	C-HET1	C-HET2
	$exp = 2$	CX2-SPN	CX2-HOM	CX2-HET1	CX2-HET2
	$exp = 4$	CX4-SPN	CX4-HOM	CX4-HET1	CX4-HET2

Table 2.4: NoC architecture instances for combined architectural parameters.

of them are explained below:

- X2-HOM: the baseline network has been converted into HOM partitioned network with express physical links of $exp = 2$
- C-HET2: the baseline has been converted into a HET2 concentrated network
- CX4-HET1: the baseline has been converted into a HET1 concentrated network with express physical links of $exp = 4$

2.4.1 Bisection Wire Count and Buffer Allocation

A unified equation that gives buffer allocation for every NoC case (24 instances) is:

$$BUF = \sum_{net=1}^{N} RTs_{net} \times avgP_{net} \times MSG_{net} \times VC_{net} \times FD_{net} \times (\frac{\sigma}{\rho} \times FW_{net})$$

Below, there is an example that shows how much buffer is allocated when combining all three architectural parameters. We take as an example a CX2-HET1 NoC, (table 2.5 shows the baseline and CX2-HET1's topology and router architecture parameters):

$$\begin{aligned} BUF_{base} &= 64 \times 4.5 \times 3 \times 1 \times 6 \times (\frac{1}{1} \times 96) \\ &= 497664 bits \end{aligned}$$

RT/NoC parameters \ NoC type	SPN	CX2-HET1	
Net	#1	#1	#2
avgP	4.5	8	8
MSG	3	1	2
VC	1	1	1
FD	6	6	6
σ	1	2	2
ρ	1	2	2
FW	96	64	32

Table 2.5: NoC configuration for a baseline SPN network and a CX2-HET1 network without buffer allocation adjustment.

$$\begin{aligned} BUF_{CX2-HET1} &= 16 \times 8 \times 1 \times 1 \times 6 \times (\frac{2}{2} \times 64) \\ &\quad + 16 \times 8 \times 2 \times 1 \times 6 \times (\frac{2}{2} \times 32) \\ &= 98304 bits \end{aligned}$$

Keeping router micro-architecture of CX2-HET1 unmodified, buffer allocation becomes 80% less than the baseline's!

2.4.2 Resource Adjustment

But, CX2-HET1 has express physical links that may need pipelining. Moreover, due to concentration, local links will increase, too, because less routers are sharing the same dimension distance. For a $144mm^2$ ($12mm \times 12mm$) available die space for a NoC, each baseline's router link could ideally be $\frac{12}{8} = 1.5mm$ (8 routers per dimension). For a 45nm LSTP (Low STandby Power) technology and 30% buffer insertion overhead, wire delay could be 220 picoseconds per mm [44]. For a 1.5mm link, wire delay could be $1.5mm \times 220\frac{ps}{mm} = 330ps$. This means that a 2GHz frequency would be enough for the circuit to operate normally. On the other hand CX2-HET1's local links could ideally be $\frac{12}{4} = 3mm$ (4 router per dimension). Thus 3mm link wire delay could be $3mm \times 220\frac{ps}{mm} = 660ps$ which means that circuit could not operate at higher frequencies than 1.5GHz. Thus, one extra pipe stage is needed for CX2-HET1's local links. Express links, respectively, will be approximately 2× times more than CX2-HET1's local links (express interval is 2) which means wire delay could be $2 \times 3mm \times \frac{ps}{mm} \approx 1.3ns$. This means that circuit could not operate at higher frequencies than 0.7GHz. Therefore, wire links should be partitioned to at least 3 pipe stages, to accommodate 2GHz operating frequency. So, 2 more pipe registers per wire are needed. By increasing router link traversal latency from 1 to 3, flit depth FD buffer should be increased to cover

the RTT delay. In current case CX2-HET1's FD should be increased from 6 to 10 $(6 + 2 \times 2)$.

Thus, CX2-HET1 buffer allocation now becomes:

$$\begin{aligned} BUF_{CX2-HET1} &= 16 \times 8 \times 1 \times 1 \times 10 \times (\frac{2}{2} \times 64) \\ &\quad + 16 \times 8 \times 2 \times 1 \times 10 \times (\frac{2}{2} \times 32) \\ &= 163840 bits \end{aligned}$$

Which is still less than baseline (67% of baseline less). This amount of memory can be filled by increasing CX2-HET2's VC by 3. Increasing VC will boost the performance of CX2-HET2 increasing saturation throughput. Hence, final CX2-HET2's buffer allocation is:

$$\begin{aligned} BUF_{CX2-HET1} &= 16 \times 8 \times 1 \times 3 \times 10 \times (\frac{2}{2} \times 64) \\ &\quad + 16 \times 8 \times 2 \times 3 \times 10 \times (\frac{2}{2} \times 32) \\ &= 491520 bits \end{aligned}$$

Result shows that the CX2-HET1's buffer allocation is almost equal to the baseline's (497664 bits) and a fair evaluation can be taken.

NoC type / **RT/NoC parameters**	**SPN**	**CX2-HET1**	
Net	#1	#1	#2
VC	1	**3**	**3**
FD	6	**10**	**10**
FW	96	64	32
Bis.	1546	1546	
B.Alc	497664	491520	

Table 2.6: CX2-HET1 configuration after buffer allocation and bisection wire count adjustment.

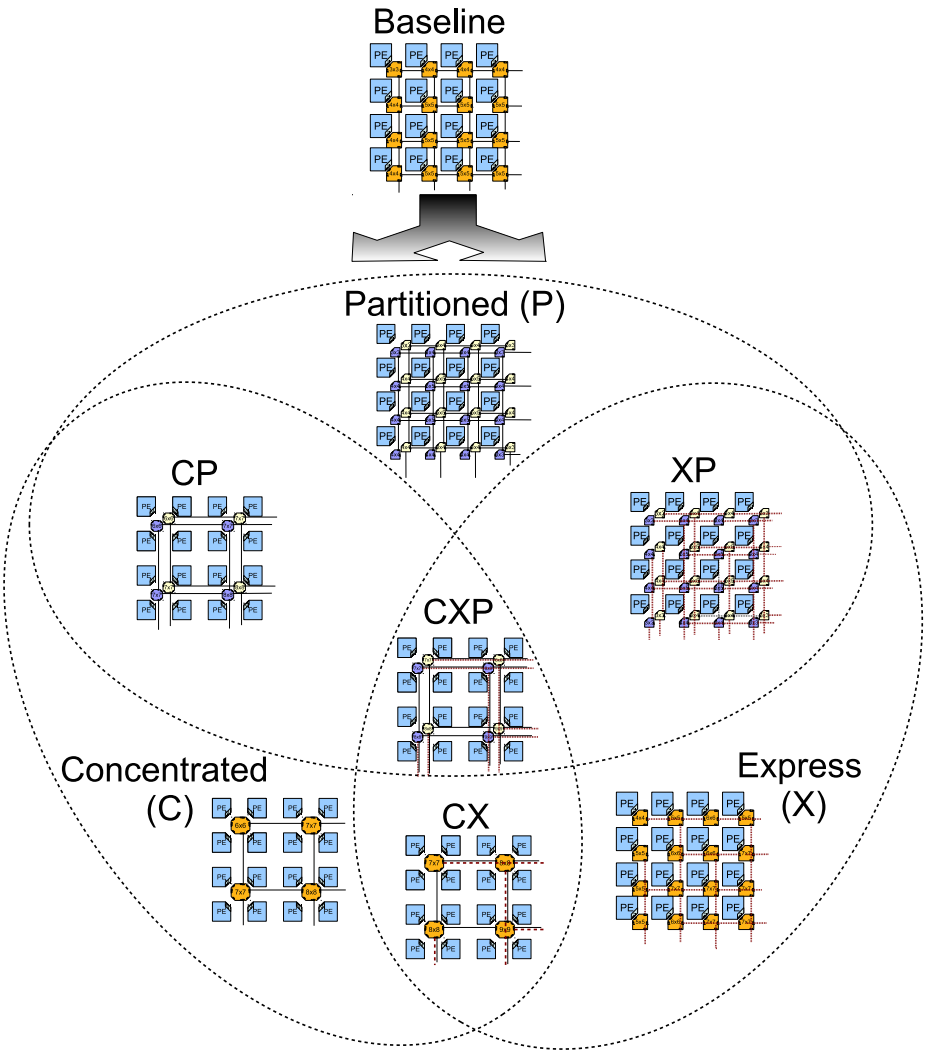


Figure 2.11: A high level view of NoC architecture instances when combining architectural parameters.

Chapter 3

NoC Model and Experimental Infrastructure

This chapter demonstrates the process for the evaluation of 20 and 24 network instances of 64 and 256 PEs respectively. All of the network instances (44 in total) are adjusted to have fixed buffer allocation and bisection bandwidth (bisection wire count). The NoC model is presented first and later the experimental infrastructure.

3.1 Model

In the following subsections, we present: a) the assumed NoC architecture in terms of system and device level (router and links) and b) the area-power-energy equations used together with power-performance simulators. The system level is presented in Sec. 3.1.1, the device level in Sec. 3.1.2 and area, power and energy mathematical expressions in Sec. 3.1.3.

3.1.1 System

A single chip multiprocessor (CMP) is assumed, comprising of either 64PEs or 256PEs. PEs, routers and links are supposed to fit on $150mm^2$ die area. A PE consists of computing logic and memory slices. Each memory slice could be a level of cache hierarchy (i.e. L1/L2 cache). In our case, a LLC (last level cache) or MCU (memory controller unit) and is connected to a router's injection/ejection port through a PE's NI.

In order to facilitate communication for parallel workloads, task migration or other functions, threads need to communicate either via explicit message passing [45] or via shared memory address space [46]. In either case, at least three traffic classes are needed [47, 42] to avoid message-dependent network deadlocks. Accordingly, this work considers the following three traffic classes: A high priority traffic class for response messages, a medium priority traffic class for intervention messages and a low priority traffic class for request messages.

Packet sizes are considered to be fixed, consisting of two sizes. Short and long. A short packet is assumed to be 128 bits, which is enough to fit a memory address, protocol-oriented commands (read request, write reply e.t.c.) and network specific data (flow-control, routing, buffer info e.t.c). Long packets are assumed to be 640 bits, which is enough to fit memory data (one cache line), protocol commands and network-specific data.

Furthermore, packets can be classified as either control or data packets. Control packets are read/intervention requests and write replies (acks). Data packets can be read replies and write requests. Read/intervention requests and write replies are short while write requests and read replies are long packets.

For fixed-size packets, shared memory address space mechanisms give higher $\frac{C}{D}$ ratio of control packets per data packets ranging from ≈2 to ≈4. On the other hand explicit messaging mechanisms give ratios ranging from ≈0.1 to ≈1 [48, 40]. This work explores $\frac{C}{D}$ packet ratios ranging from 0.1 to 3.

3.1.2 Router and Links

This subsection explains how and why the router model was selected and how link assignment was modeled along with the pipelining.

Router

A 3-stage input-buffered router is modeled similar to [2, 3]. These three pipe-stages include (See Fig. 3.1) a buffer access (BA) and route computation (RC) stage, a switch allocation (SA) combined with a virtual channel (VA) assignment stage and a final crossbar traversal (XB) stage.

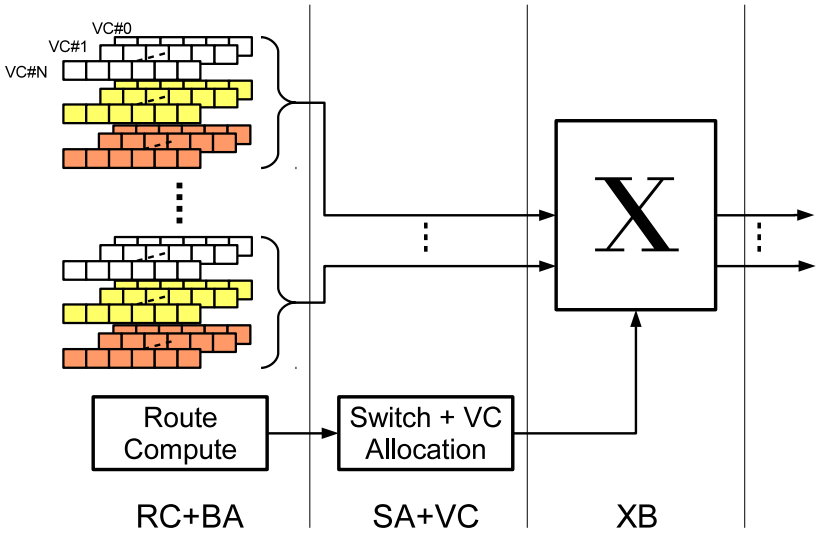


Figure 3.1: A three-stage router pipeline model similar to [2, 3]. Each color of buffer queues represent a message class.

1st stage: During buffer access (BA), read/write is performed in parallel with route computation (RC). Buffer's logic is modeled as two-port register-FIFOs and routing computation consists of deterministic XY ordered routing.

2nd stage: When head flit enters the second pipeline-stage, switch allocation (SA) and virtual channel assignment (VA) is performed in parallel. At the beginning of the cycle, the flit looks for a free VC for every output port from a pool of free VC's. If VA and SA have successfully granted a virtual channel and an output port the flit is ready to proceed to crossbar traversal (XB), and at the same time a credit is sent to the upstream router (credit-based scheme for flow control). If SA+VA phase fails, the flit is moved to the back of the SA queue and retries for both virtual and physical channel (SA+VA is repeated). Routers without virtual channels, have only the SA phase.

3rd stage: When virtual channel and output port are successfully granted, control signals steer the flit to the right output through crossbar traversal phase (XB). The crossbar is modeled as a tree-multiplexer due to the fact that tree-multiplexers appear to have better area, power and scalability characteristics than matrix implementations [30] (see Sec. 1.3.1).

The combined SA+VA router differ from the textbook's [17] virtual-channel router, because it doesn't have separate arbiters for VA. VA is performed through a FIFO memory simultaneously with the SA phase. It also differs from the router described in [43] which uses speculation to enable both allocators to operate concurrently. Combined (SA+VA), selects a VC from a queue of pointers of free VCs, after a successful SA, reducing the complexity of VC allocation's delay, area and power. This is because the number of arbiters and the arbiter's request width increase linearly with respect to the number of VCs. Thus, leakage power and area increases quadratically with the number of virtual channels. On the other hand dynamic power increases linearly with the number of VCs, because the utilization rate of each arbiter is inversely proportional to their number [27].

Links

Links are router-to-router wire segments that bi-directionally connect two routers or a NI to a router. Router-to-router links can be local links, express physical links of $exp = 2$ and express physical links of $exp = 4$. Fig. 3.2 depicts those links for 64 and 256 PEs (concentrated or not). Due to the fact that the power dissipation and delay of the buffered wires are proportional to their length [44], wire length is carefully considered in this work.

We assume all tiles and routers are squares, placed on a square of $X \times X$ area. For Link length computation we need to know X, router's side (r), concentration degree (C), when concentration is examined, and express interval (exp), if express

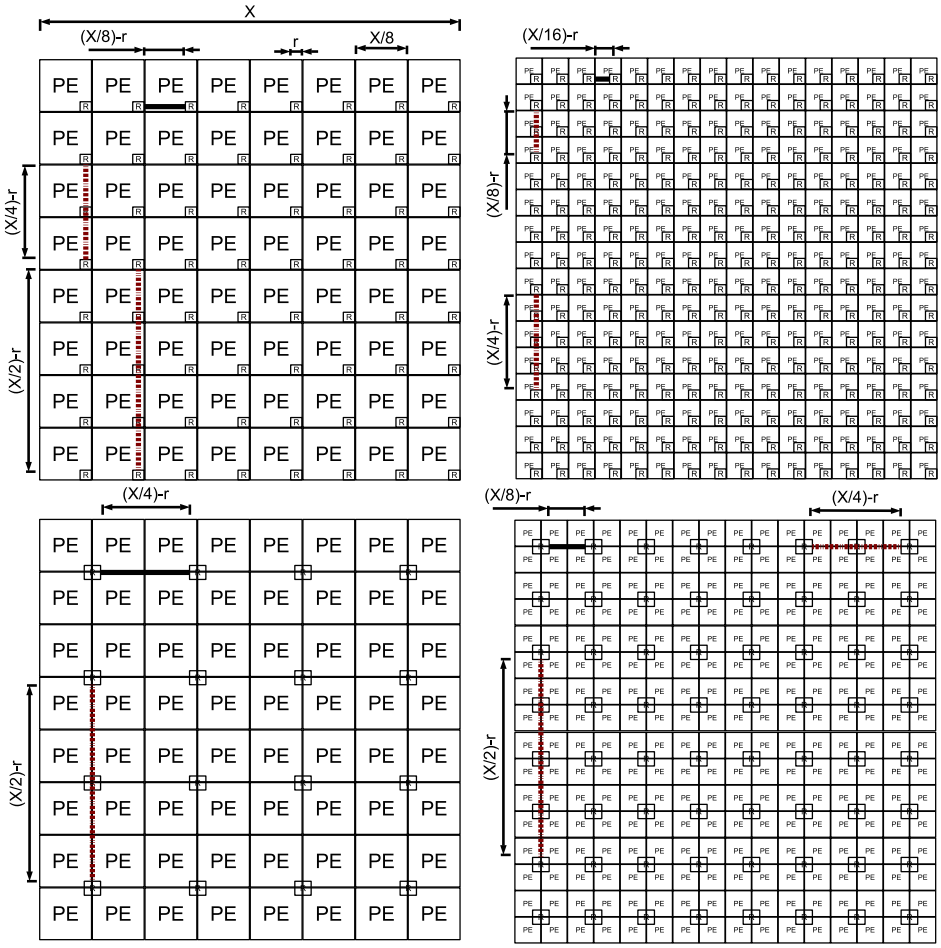


Figure 3.2: Four identical (same area) floor-plans for 64 (left) and 256 PEs (right). Top floor-plan are non-concentrated. Bottom are concentrated. Red links are express channels, black are local channels.

physical links are included. Finally, link length can be calculated by:

$$l = \frac{X \times exp}{\sqrt{\frac{PEs}{C}}} - r \tag{3.1}$$

(Note: for non-express networks, here, *exp* should not substituted to zero as 2.17, but substituted to one).

As it can be seen from Eq. 3.1, the router's area plays a role in the length of the links too. Some networks with more area-hungry routers can have smaller link length than networks with less area-hungry ones. Also, concentration plays a significant role in terms of link length. Link power trade-offs can be grasped, when considering all (four) of the above parameters (X,r,C,exp)

The next step for modeling NoC's wire length is link pipelining. Depending

on the network, link length can vary greatly, affecting wire delay and possibly clock frequency. To enforce pipelining, a length threshold was assumed. After that threshold, a pipe register is inserted. For example, if length threshold is 2 mm, for a 3 mm link, one pipeline stage is inserted. Later, the assumed threshold will be discussed in more detail in Sec. 3.2.2 and Page 50.

3.1.3 Power, Energy and Area

So far, system and device level level has been described. This subsection encompasses area-power-energy mathematical expressions along with the above architectural levels (system and device). First, equations are shown for power, then energy, and finally, area.

Power: NoC power ($P = P_d + P_s$, dynamic: $P_d = \frac{1}{2}\alpha f C V^2$ and static: $P_s = I_{static} \times V$) is broken down to that of routers and of links. Dynamic power of a single router and link is proportional to the switching capacitance, the square of voltage source, the operating frequency and the activity factor. When no DVFS (Dynamic Voltage and Frequency Scaling) or power gating techniques are assumed, as traffic injection rate increases, only the activity factor (α) increases, while other parameters like frequency (f) and voltage (V) remain intact. Injection rate is the utilization rate of router's injection port, whereas, activity factor is the resulting bit-switching utilization caused by input ports activity (injection port and routing ports). Router power estimators use the router's input activity factors as activity factors [49, 27].

This work measures input activity factors when the network begins to saturate at the time that zero-load[1] latency doubles. The reason that saturation point was chosen close to zero-load was due to the latency intensive nature of CMPs.

If N is the total number of networks, RTs is the number of routers per network, $\bar{P}_{RT}$ is the average power per router, Ls is the total number of local links, XLs is total number of express physical links (if they exist), $\bar{P}_L$ is the average power of each link and $\bar{P}_{XL}$ is the average power of each express physical link (if exists), then, average NoC power (P_{NoC}) of a network instance, is given by:

$$P_{NoC} = \sum_{net=1}^{N} (RTs \times \bar{P}_{RT_{net}} + Ls \times \bar{P}_{L_{net}} + XLs \times \bar{P}_{XL_{net}}) \tag{3.2}$$

where $\bar{P}_{RT}$, $\bar{P}_L$ and $\bar{P}_{XL}$ are further decomposed as:

$$\bar{P}_{RT} = \frac{\sum_{i=1}^{RTs} P_{RT_i}}{RTs}, \quad \bar{P}_L = \frac{\sum_{i=1}^{Ls} P_{L_i}}{Ls}, \quad \bar{P}_{XL} = \frac{\sum_{i=1}^{XLs} P_{XL_i}}{XLs}$$

and each of P_{RT}, P_L and P_{XL}, depending on type of wire (piped or not piped),

[1]The zero-load latency is the head flit latency plus the serialization latency that a packet encounters on average, when almost no load exists on the network.

comprising of:

$$P_{RT} = P_{BUF} + P_{ALC} + P_{XB}$$
$$P_L = 2 \times l_L \times \frac{P_{wire}}{mm} \times FW \quad or \quad P_{pL} = 2 \times l_L \times \frac{P_{pwire}}{mm} \times FW$$
$$P_{XL} = 2 \times l_{XL} \times \frac{P_{wire}}{mm} \times FW \quad or \quad P_{pXL} = 2 \times l_{XL} \times \frac{P_{pwire}}{mm} \times FW$$

Energy: Energy is consumed as a flit traverses through router and link components. From the first cycle of its generation, at the time that NI injects it to the source router injection port, until it reaches its final destination router (flit ejects destination router's port), each flit traverses on average $\bar{H}+1$ routers, $\xi \times \bar{H}$ local links and $(1-\xi) \times \bar{H}$ express physical links (if they exist). Thus, the average flit energy is:

$$\bar{E}_{flit} = \sum_{net=1}^{N} (\bar{H}+1) \times \bar{E}_{RT_{net}} + (\xi \times \bar{H}) \times \bar{E}_{L_{net}} + ((1-\xi) \times \bar{H}) \times \bar{E}_{XL_{net}}$$

where ξ is local hops:total hops ratio ($\xi = \frac{H_L}{H_L+H_{XP}}$), $1-\xi$ is express hops:total hops ratio, $\bar{E}_{RT}$ is average energy per router per active input port, $\bar{E}_L$ is average energy per active local link, either pipelined or not, $\bar{E}_{XL}$ is average energy per active express link, again, either pipelined or not.

$$\bar{E}_{RT} = \frac{1}{2}\alpha_{RT} C_{RT} V^2$$
$$\bar{E}_L = \frac{1}{2}\alpha_L (l_L \times \frac{C_{wire}}{mm} \times FW) V^2 \quad or \quad \bar{E}_{pL} = \frac{1}{2}\alpha_L (l_L \times \frac{C_{pwire}}{mm} \times FW) V^2$$
$$\bar{E}_{XL} = \frac{1}{2}\alpha_{XL} (l_{XL} \times \frac{C_{wire}}{mm} \times FW) V^2 \quad or \quad \bar{E}_{pXL} = \frac{1}{2}\alpha_{XL} (l_{XL} \times \frac{C_{pwire}}{mm} \times FW) V^2$$

For a network with a saturation bandwidth of $BW_{sat} = \chi$, injected data per cycle is $\chi\ bits/cycle$. The total number of cycles per injected data (D), is $D/BW_{sat} = D/\chi$ cycles. Thus, energy per injected data (D), for a saturated network at bandwidth BW_{sat}, is:

$$\begin{aligned} E_{data} &= E_{flit} \times cycles \\ &= E_{flit} \times \frac{D}{BW_{sat}} \end{aligned}$$

For example, assume two networks. SPN1 and SPN2 (single 2D mesh physical network 1 and 2). Let SPN1's saturation bandwidth be $BW_{SPN1} = 10 bits/cycle$ and SPN2's $BW_{SPN2} = 14 bits/cycle$. 1 KB data, will need $\frac{8192 bits}{10 bits/cycle} \approx 819$ cycles in SPN1 and $\frac{8192 bits}{14 bits/cycle} \approx 585$ cycles for SPN2 to be injected. The average energy per 1KB of transmitted data for SPN1, SPN2 is respectively $E_{SPN1_{1KB}} = E_{SPN1_{flit}} \times 819$, $E_{SPN2_{1KB}} = E_{SPN2_{flit}} \times 585$.

Area: In this paragraph we discuss the overall measured area. NoC area is comprised of router area plus link area (pipelined or not). An equation for the total NoC area is shown,

$$A_{NoC} = \sum_{net=1}^{N} RTs \times \bar{A}_{RT_{net}} + Ls \times \bar{A}_{L_{net}} + XLs \times \bar{A}_{XL_{net}}$$

where $\bar{A}_{RT}$ is the average area per router, $\bar{A}_L$ is the average area per local link, and $\bar{A}_{XL}$ is average area per express physical link. $\bar{A}_{RT}$, $\bar{A}_L$ and $\bar{A}_{XL}$ are further decomposed as:

$$\bar{A}_{RT} = \frac{\sum_{i=1}^{RTs} A_{RT_i}}{RTs}, \quad \bar{A}_L = \frac{\sum_{i=1}^{Ls} A_{L_i}}{Ls}, \quad \bar{A}_{XL} = \frac{\sum_{i=1}^{XLs} A_{XL_i}}{XLs}$$

and each of A_{RT}, A_L and A_{XL}, which are depending on type of wire (pipelined or not), are comprised of:

$$A_{RT} = A_{BUF} + A_{ALC} + A_{XB}$$
$$A_L = 2 \times l_L \times \frac{A_{wire}}{mm} \times FW \quad or \quad A_{pL} = 2 \times l_L \times \frac{A_{pwire}}{mm} \times FW$$
$$A_{XL} = 2 \times l_{XL} \times \frac{A_{wire}}{mm} \times FW \quad or \quad A_{pXL} = 2 \times l_{XL} \times \frac{A_{pwire}}{mm} \times FW$$
$$A_{wire} = A_{buf} + A_{rout} \quad or \quad A_{pwire} = A_{buf} + A_{rout} + A_{pipe}$$

3.2 Experimental Infrastructure

This section describes the simulation and configuration procedure. In the beginning we present a) the experimental methodology in high abstract level (Sec. 3.2.1 on Pg. 47), then b) the simulation methodology and parameters that were used, such as simulator tools, process technology, packet distribution e.t.c. (Sec. 3.2.2 on Pg. 49) and finally c) the detailed configuration setup (Sec. 3.2.3 on Pg. 51) for 44 network instances (20 networks for 64PEs and 24 networks for 256 PEs).

3.2.1 High-Level Abstraction Methodology

Fig. 3.3 shows a high level view of the experimental methodology. Firstly, the individual topologies are generated, using as a reference point a 2D mesh baseline network with 3 traffic classes, 6 flit buffer depth, 64 bit router datapath, 1 virtual channel per traffic class, 1 cycle latency of local hops with a resulted credit return latency at 4 cycles in total.

As the topologies have been generated, the router micro-architecture is configured. This configuration adjusts routers so as to conform with

1. network architecture specifications (type of network partitioning, presence of express physical links and concentration),

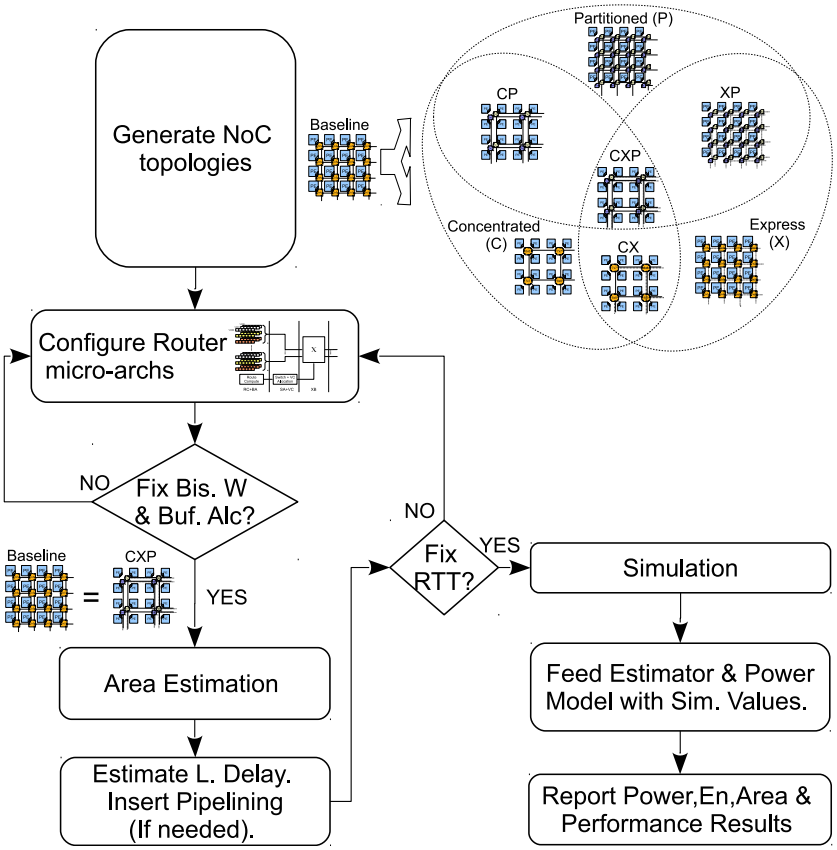


Figure 3.3: A high level view of the experimental methodology.

2. bisection wire count (to be the same as the baseline's) and

3. buffer allocation (to be the same as the baseline's).

After the router micro-architecture configuration, area estimation takes place followed by the link length computation using the procedure described in Sec. 3.1.2 43. Then, link delay specification is checked, and extra pipeline stages are inserted if needed. Depending on the number of the pipe-stages inserted, buffer depth is checked if it is conformed with the credit return latency. If it does not, the procedure goes back to router configuration, to adjust the router's buffer depth according to the new credit return latency.

Next, the simulation starts, and each network is being tested. As the simulation finishes, the values obtained from the network simulator (such as activity factors, average local/express hops, saturation point, throughput e.t.c.) are used by the power estimator together with the power-energy-area mathematical expressions described in Sec. 3.1.3 on Pg. 45 and results are then reported.

3.2.2 Simulation Methodology and Tools

A modified version of Booksim [50] was used for cycle accurate micro-architecture level network simulation to *a*) evaluate network performance and *b*) get statistical values (switching-activity factors, average local/express hops, saturation throughput e.t.c.) that was necessary for power and energy estimation. Additional features were implemented in the Booksim simulator to further support heterogeneous networks of the HET1 and HET2 schemes, as described in Sec. 2.1.1 on Pg. 13. Also, a generalized scalable express physical channel scheme was extended into Booksim, which can include additional express physical links into a network, with a variable express hop interval (in this work, $exp = 2$ and $exp = 4$) similar to the Chen et. al. [34] scheme.

Simulation values of Booksim were later added to the power and energy NoC models described in Sec. 3.1.3 on Pg. 45, together with combined orion2 [27] and orion3 [51, 29] router-link power estimator tools. Orion estimators were further modified to incorporate a custom model for link-pipelining. This was necessary as there was no link-pipelining model supported by the tools.

Orion3 [51, 29] was used for router's buffer power/area estimations, whereas orion2 [27] was used similarly for power/area estimation for the rest of the router components (combined switch/virtual channel allocator and crossbar) and links. Combination of both tools was used for two reasons:

1. Orion3 [51, 29] has less average error in buffer power/area estimation, and takes into account buffer manipulation logic. Buffer manipulation logic includes control signals (like FIFO select and buffer enable control) and housekeeping logic (like number of free buffers available per virtual channel and VC identification tag per buffer) [29]. Careful estimation for buffer power/area was a significant factor to consider, as this work's evaluated networks intentionally have almost the same total buffer allocation among them, causing some networks to have more but smaller memory instances (and hence, more control logic per memory logic), or less but bigger memory instances (and hence less control logic per memory logic).

2. Orion2, in terms of crossbar power dissipation, better captures the super linear characteristics of dynamic power relative to datapath width (which happens because orion3 model's regression post-PnR crossbar datapath values are ranging from 16 to 64 bits. As we saw in Sec. 2.1 on Pg. 11, the superlinearity nature between crossbar dynamic power and the crossbar width, is mostly observable beyond 64 bits.) This is especially useful, as many of evaluated networks have taken into account network partitioning. When a crossbar datapath is partitioned into more than one slices, overall power dissipation can potentially be less than using a single crossbar. Crossbar power relationship with datapath width is described in Sec.2.1 on Pg. 11 and depicted in Fig. 2.2.

Power-area estimations used ITRS predictions for a 45nm, and LSTP (Low

Standby Power) process technology with operating network's frequency at 2GHz. Due to the fact LSTP's leakage power isn't as significant as dynamic power [44], only dynamic power is fed into our model.

In terms of link pipeline insertion, delay estimation of non-pipedlined wires was done by cacti [44], to figure out if pipelining was needed. When link pipelining is needed, extra pipeline area is estimated by orion2, and extra power of pipelining is estimated by a power pipeline overhead factor ($pipe_{ovhd}$). As shown below,

$$pipe_{ovhd} = \gamma \times \frac{pipes}{mm} \tag{3.3}$$

this pipeline overhead factor relates non-pipelined wire with pipelined one. This was done using post-PnR simulations for 65,90,130,180nm [1, 4, 5, 6] and then projecting γ for 45nm (see Fig. 3.4). Depending on the maximum clock cycle that

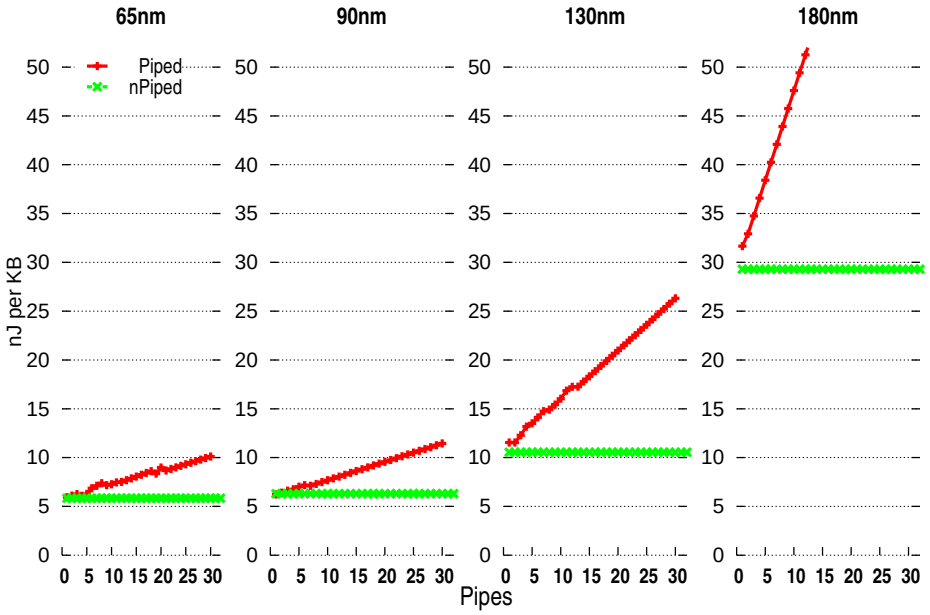


Figure 3.4: Energy scaling between non-pipelined and pipelined wires. Values are obtained from Post PnR simulations, for 65, 90, 130 and 180nm [1, 4, 5, 6].

we can afford, $\frac{pipes}{mm}$ of Eq. 3.3 is adjusted accordingly. For a 45nm LSTP technology and a 30% buffer insertion overhead, wire delay can be 220 picoseconds per mm [44]. For a clock frequency equal to $2GHz$, clock cycle should be at least $500ps$. In this work we assume that extra pipe is inserted when a non-piped wire reaches $\approx 330ps$ delay. This means that for each $\frac{330ps}{220ps/mm} \approx 1.5mm$ one pipe is inserted.

The synthetic workloads used in this evaluation, consist of four traffic patterns that exhibit diverse communication behaviors: a) uniform, b) neighbor, c) bit complement and d) transpose. As mentioned earlier, in Sec. 3.1.1 on Pg. 42,

traffic comprises of either control or data packets. Also, $\frac{C}{D}$ ratio has been shown to be larger for communication models such as shared memory address space, and smaller for explicit messaging. This ratio has been shown to span from 2 to 4 for shared memory address space mechanisms, and, ≈ 0 to 1 for explicit messaging communication models [48, 40]. This work takes into account for each of the traffic pattern (uniform, neighbor, bit complement and transpose) the $\frac{C}{D}$ ratio by injecting traffic at 3,2,1,0.5,0.33 and 0.1 $\frac{C}{D}$ ratios, to investigate potential in networks, and especially the heterogeneous ones.

3.2.3 Detailed NoC Architecture Instances

Combining P, C and X parameters 20 networks arise for 64 PEs, and 24 networks for 256 PEs. For 64 PEs, concentrated networks with express physical links of $exp = 4$ do not exist, because a 4x4 mesh cannot have an express link interval equal to 4. Table 3.1 shows valid topologies that result for 64 PEs and 256 PEs.

		NoC architecture instances				**64PEs**	**256PEs**
		Not P'ed	**Partitioned**			**Nodes per Net**	
	nX	SPN	HOM	HET1	HET2	64	256
nC	$X2$	X2-SPN	X2-HOM	X2-HET1	X2-HET2	64	256
	$X4$	X4-SPN	X4-HOM	X4-HET1	X4-HET2	64	256
	nX	C-SPN	C-HOM	C-HET1	C-HET2	16	64
C	$X2$	CX2-SPN	CX2-HOM	CX2-HET1	CX2-HET2	16	64
	$X4$	CX4-SPN	CX4-HOM	CX4-HET1	CX4-HET2	×	64

Table 3.1: NoC topologies for 64 and 256 PEs. Concentration or no concentration is depicted with "C/nC". Express physical links with express interval equal to 4/2/0 is depicted with "X4/X2/nX".

The experimental methodology described in Sec 3.2.1, depicted in Fig. 3.3, was followed for each generated NoC to adjust buffer and bisection wire resources to be as similar as possible for all NoCs. Flit depth was carefully adjusted in cases where pipelining was inserted, to cover RTT latency, and, extra virtual channels were added to boost performance in NoCs that had available memory. We present the NoC configurations of 44 2D mesh networks for either 64PEs (20 networks) or 256 PEs (24 networks) in the following two tables. For 64 PEs, table 3.2.3, and for 256 PEs, table 3.2.3.

A brief explanation of some column variables is shortly given below:

Ps is the total number of ports per NoC.

D_L is the delay in cycles of a local link. If $D_L > 1$ then link pipelining has been inserted with D_L number of pipe-stages.

D_{XL} is similar with D_L, but for express physical links.

Bis.W is the total bisection wire count of a NoC.

$\frac{\textbf{Bits}}{\textbf{port}}$ are the average bits per input port per router.

$\textbf{BUF}_{\textbf{KB}}$ gives the overall buffer allocation in KBs per NoC.

$\textbf{MSG}_{\textbf{2}}$ is the number of traffic classes (message classes) per input port for the second subnetwork.

NoC	Nets	RTs	Ps	avgP	D_L	D_{XL}	VC	FD	Bis.W	MSG_1	FW_1	MSG_2	FW_2	MSG_3	FW_3	$\frac{Bits}{port}$	BUF_{KB}
SPN	1	64	288	4.5	1	0	1	6	512	3	64	0	0	0	0	1152	40.5
HOM	2	128	576	4.5	1	0	1	6	512	3	32	3	32	0	0	576	40.5
HET1	2	128	576	4.5	1	0	2	6	512	1	32	2	32	0	0	576	40.5
HET2	3	192	864	4.5	1	0	3	6	528	1	22	1	22	1	22	396	41.8
X2-SPN	1	64	384	6	1	2	1	9	512	3	32	0	0	0	0	864	40.5
X2-HOM	2	128	768	6	1	2	1	9	512	3	16	3	16	0	0	432	40.5
X2-HET1	2	128	768	6	1	2	2	9	512	1	16	2	16	0	0	432	40.5
X2-HET2	3	192	1152	6	1	2	4	7	480	1	10	1	10	1	10	280	39.4
X4-SPN	1	64	352	5.5	1	4	1	14	528	3	22	0	0	0	0	924	39.7
X4-HOM	2	128	704	5.5	1	4	1	13	576	3	12	3	12	0	0	468	40.2
X4-HET1	2	128	704	5.5	1	4	3	10	480	1	10	2	10	0	0	450	38.7
X4-HET2	3	192	1056	5.5	1	4	3	13	576	1	8	1	8	1	8	312	40.2
C-SPN	1	16	112	7	2	0	1	8	512	3	128	0	0	0	0	3072	42
C-HOM	2	32	224	7	2	0	1	8	512	3	64	3	64	0	0	1536	42
C-HET1	2	32	224	7	2	0	2	8	512	1	64	2	64	0	0	1536	42
C-HET2	3	48	336	7	2	0	3	8	504	1	42	1	42	1	42	1008	41.3
CX2-SPN	1	16	128	8	2	4	1	14	512	3	64	0	0	0	0	2688	42
CX2-HOM	2	32	256	8	2	4	1	14	512	3	32	3	32	0	0	1344	42
CX2-HET1	2	32	256	8	2	4	2	14	512	1	32	2	32	0	0	1344	42
CX2-HET2	3	48	384	8	2	4	3	13	528	1	22	1	22	1	22	858	40.2

Table 3.2: NoC Architecture instances (20 in total) for 64 PEs.

NoC	Nets	RTs	Ps	avgP	D_L	D_{XL}	VC	FD	Bis.W	MSG_1	FW_1	MSG_2	FW_2	MSG_3	FW_3	$\frac{Bits}{port}$	BUF_{KB}
SPN	1	256	1216	4.75	1	0	1	6	1024	3	64	0	0	0	0	1152	171
HOM	2	512	2432	4.75	1	0	1	6	1024	3	32	3	32	0	0	576	171
HET1	2	512	2432	4.75	1	0	2	6	1024	1	32	2	32	0	0	576	171
HET2	3	768	3648	4.75	1	0	3	6	1056	1	22	1	22	1	22	396	176.3
X2-SPN	1	256	1664	6.5	1	1	1	9	1024	3	32	0	0	0	0	864	175.5
X2-HOM	2	512	3328	6.5	1	1	1	9	1024	3	16	3	16	0	0	432	175.5
X2-HET1	2	512	3328	6.5	1	1	3	6	1024	1	16	2	16	0	0	432	175.5
X2-HET2	3	768	4992	6.5	1	1	5	6	960	1	10	1	10	1	10	300	182.8
X4-SPN	1	256	1600	6.25	1	2	2	7	1056	3	22	0	0	0	0	924	180.5
X4-HOM	2	512	3200	6.25	1	2	2	6	1152	3	12	3	12	0	0	432	168.8
X4-HET1	2	512	3200	6.25	1	2	4	7	960	1	10	2	10	0	0	420	164.1
X4-HET2	3	768	4800	6.25	1	2	5	7	1152	1	8	1	8	1	8	280	164.1
C-SPN	1	64	480	7.5	1	0	1	8	1024	3	128	0	0	0	0	3072	180
C-HOM	2	128	960	7.5	1	0	1	8	1024	3	64	3	64	0	0	1536	180
C-HET1	2	128	960	7.5	1	0	3	5	1024	1	64	2	64	0	0	1440	168.8
C-HET2	3	192	1440	7.5	1	0	4	6	1008	1	42	1	42	1	42	1008	177.2
CX2-SPN	1	64	576	9	1	2	2	6	1024	3	64	0	0	0	0	2304	162
CX2-HOM	2	128	1152	9	1	2	2	6	1024	3	32	3	32	0	0	1152	162
CX2-HET1	2	128	1152	9	1	2	3	8	1024	1	32	2	32	0	0	1152	162
CX2-HET2	3	192	1728	9	1	2	5	7	1056	1	22	1	22	1	22	770	162.4
CX4-SPN	1	64	544	8.5	1	4	2	10	1008	3	42	0	0	0	0	2520	167.3
CX4-HOM	2	128	1088	8.5	1	4	2	10	1056	3	22	3	22	0	0	1320	175.3
CX4-HET1	2	128	1088	8.5	1	4	3	13	1056	1	22	2	22	0	0	1287	170.9
CX4-HET2	3	192	1632	8.5	1	4	5	12	1008	1	14	1	14	1	14	840	167.3

Table 3.3: NoC Architecture Instances for 256 PEs. 24 NoC architectures in total.

Chapter 4

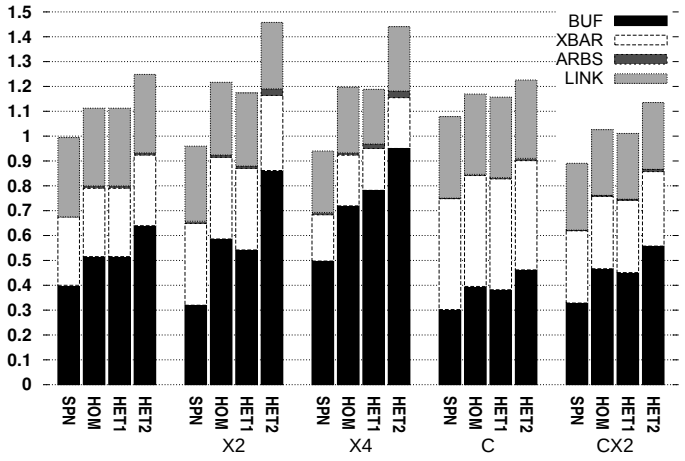
Evaluation

This chapter presents the evaluation results of the 44 NoC architectures (20 for the 64 PEs and 24 for the 256 PEs) that were generated when the three design parameters were combined together: a) Network partitioning or not (partitioned: HOM, HET1 and HET2, not partitioned: SPN). b) Network Concentration or not. c) Express physical links or not (with 2-hop/4-hop express interval). Furthermore, router-microarchitecture adjustment was introduced to restrict bisection wire count and buffer space allocation to be as close as possible the same for all of the NoCs instances.

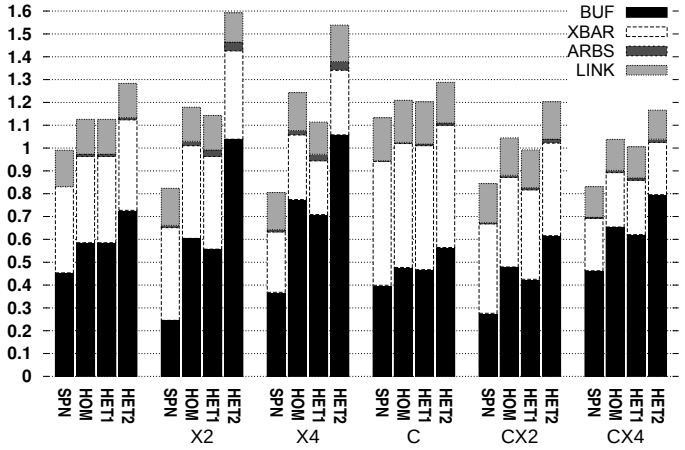
The respective specification parameters are in Table 3.2.3 and Table 3.2.3. Y-axis values are normalized relative to the baseline 2D mesh network (SPN). X-axis depicts the NoC's architecture keyword-names in same order as in first column of configuration tables. The measurements are presented into 3 sections: area (Sec. 4.1 on Pg. 4.1), performance (Sec. 4.2 on Pg. 4.2) and power/energy (Sec. 4.3 on Pg. 4.3). For convenience when we refer to a group of NoCs, either partitioned or not, we use a single prefix-word. For example, if we want to discuss about NoCs that use express physical links with express interval equally to 4 ($exp = 4$), we do this by referring as ''X4 networks'' meaning that X4 networks can be non-partitioned (X4-SPN), homogeneously partitioned (X4-HOM), heterogeneous partitioned with HET1 scheme (X4-HET1) or heterogeneous partitioned with HET2 scheme (X4-HET2).

4.1 Area Analysis

Area breakdown is depicted in Fig. 4.1i for 64 PEs and Fig. 4.1ii for 256 PEs. Area is comprised of buffer (BUF), crossbar (XBAR), allocator (ARBS), links (LINK). Link area includes link buffering, link pipelining and routing. Respective (normilized to baseline) values are shown in Table 4.1.



(i) 64PEs.



(ii) 256PEs.

Figure 4.1: Area for 64 and 256 PEs.

4.1.1 Area Savings

The maximum possible area savings that we can get, is 11% for 64 PEs and 20% for 256 PEs. The 11% is obtained from the CX2-SPN network and the 20% from the X4-SPN network. As we observe from Table 4.1, area is saved when: a) concentration is combined with express physical links, b) express physical links are used exclusively and c) network partitioning is not applied. Case *a* is mostly suitable for 64 PEs and case *b* for 256 PEs. Case *c* is not suitable for 64 and 256 PEs.

NoC	mm²	norm
CX2-SPN	6.419	0.89
X4-SPN	6.750	0.94
X2-SPN	6.841	0.95
SPN	7.204	1.00
CX2-HET1	7.294	1.01
CX2-HOM	7.400	1.03
C-SPN	7.777	1.08
HET1	8.029	1.11
HOM	8.042	1.12
CX2-HET2	8.182	1.14
C-HET1	8.334	1.16
C-HOM	8.424	1.17
X2-HET1	8.453	1.17
X4-HET1	8.572	1.19
X4-HOM	8.628	1.20
X2-HOM	8.753	1.22
C-HET2	8.819	1.22
HET2	9.073	1.26
X4-HET2	10.279	1.43
X2-HET2	10.512	1.46

(i) 64PEs.

NoC	mm²	norm
X4-SPN	21.655	0.80
X2-SPN	22.341	0.82
CX4-SPN	22.492	0.83
CX2-SPN	22.884	0.84
CX2-HET1	26.907	0.99
SPN	27.105	1.00
CX4-HET1	27.274	1.01
CX4-HOM	28.163	1.04
CX2-HOM	28.346	1.05
X4-HET1	30.101	1.11
HOM	30.579	1.13
HET1	30.592	1.13
C-SPN	30.691	1.13
X2-HET1	30.858	1.14
CX4-HET2	31.605	1.17
X2-HOM	32.020	1.18
CX2-HET2	32.557	1.20
C-HET1	32.600	1.20
C-HOM	32.809	1.21
X4-HOM	33.730	1.24
HET2	34.841	1.29
C-HET2	34.858	1.29
X4-HET2	41.775	1.54
X2-HET2	43.053	1.59

(ii) 256 PEs.

Table 4.1: NoC area for 64 and 256 PEs.

4.1.2 Impacts Imposed by P,C and X Parameters

Network Partitioning

When network partitioning is applied to each of the non-partitioned NoC (which means: SPN, X2-SPN, X4-SPN, CX2-SPN, CX4-SPN) total area is increasing. This increase, as seen in Fig. 4.1i, 4.1ii is mostly due to buffer area. There are two reasons for this: a) Buffer area increases because partitioned networks need more, but smaller memory instances, increasing control logic for the same memory budget. Depending on partitioning scheme, HET1 seems slightly better than HOM networks, and HET2 is the worst. This happens as buffer control logic is replicated three times (equal to the number of subnetworks) for the same memory budget, and simultaneously, HET2 has more virtual channels assigned relative to other networks and that means additional manipulation logic for VCs. b) As seen from figures Fig. 4.1i, 4.1ii crossbar area remains almost the same as network partitioning is applied for each of SPN, X2-SPN, X4-SPN, CX2-SPN and CX4-SPN networks. This is due to the fact that crossbars are comprised of tree-based multiplexers rather than matrix crossbars. Area complexity of tree-based crossbars is

$O(W)$ whereas matrix-based crossbars is $O(W^2)$. (See crossbar's complexity Sec. 1.3.1 on Pg. 4) Allocator's area is affected too. As number of subnetworks increase, we pay the same arbitration logic for each subnetwork, increasing the overall arbitration logic. Depending on partitioning scheme, HET1 seems slightly better than HOM networks, and HET2 the worst. In terms of wiring, there are no important differences. This is due to the fact that bisection wire count was kept as much as possible the same for all NoCs.

Concentration and Express Physical Links

As we saw in Sec. 2.2 on Pg. 23 about concentration, and in Sec. 2.3 on Pg. 27 about express physical links,

a) when concentration is introduced, router datapath becomes $\sigma = \sqrt{C}$ times wider and input ports are increased by $\sigma - 1$ input ports per router but, the number of routers is decreased by C times.

b) when express physical links are introduced, router datapath becomes ρ times narrower ($\rho = \frac{1}{\frac{exp}{2}+1}$) and router ports are increased by less than 2 input ports on average (see Sec. 2.3.3 on Pg.33) per router.

It seems that in case *a*, although the number of routers are kept the same, and router input ports are increased by 2 input ports per router, the datapath reduction helps a lot for area efficiency even for 256 PEs. In case *b*, although router input ports are increased by 3 input ports per router on average and the datapath becomes wider, the reduction of the number of routers due to concentration does not helps further the area efficiency. However, when concentration and express physical links are combined together, they are giving very good area savings ranging from 11% for 64 PEs, to 17% for 256 PEs. But, when the express physical links are used exclusively, as PE count increases to 256 PEs, they are good enough to outperform CX-SPN networks, giving better area savings up to 20%.

4.2 Performance Analysis

Performance capabilities of each NoC are reported. The major metrics included are the average zero load latency and average transfer time per injected amount of data. The injection rate used in this evaluation, as already mentioned in Sec. 3.1.3 on Pg. 45, is at saturation point where average latency becomes 2× times the zero-load latency, as CMPs, are mostly latency intensive. We measure the relationship between average hops and average utilized link length too in order to investigate how much of link length is switching on average, relative to the number of average hop count.

For each of all the above metrics, as mentioned in Sec. 3.2.2 on Pg. 50, four synthetic workloads (uniform, neighbour, bit complement and transpose) were used that exhibit diverse communication behaviours, and a range of $\frac{C}{D}$ ratios

(3,2,1,0.5,0.33 and 0.1) for each simulation run. This was done, firstly to explore how explicit communication models which impose less control packets per data packets, affect NoC's efficiency and how shared address space models similarly affect NoC's power, energy and performance. Secondly, to examine how heterogeneous partitioned networks are affected into this message distribution. Due to the fact that heterogeneous networks assign traffic classes onto a physical subnetwork, $\frac{C}{D}$ ratio could under/over-utilize subnetworks, imposing imbalance among subnetworks, and thus degrading the overall network performance and potentially energy efficiency. For clarity we show two cases of packet ratios, 3 and 0.33 $\frac{C}{D}$. Conclusions equally stand for the rest of the ratios (2, 1, 0.5 and 0.1), as the trend is kept unchanged.

4.2.1 Latency

Each packet latency comprises of head latency and serialization latency. Head latency depends on router pipeline delay and link traversals (hop counts), whereas serialization depends on flit count of a packet [17]. Average zero-load latency is depicted in Fig. 4.2i for 64 PEs, and Fig. 4.2ii for 256 PEs. Respective values are shown in table Table 4.2. For either 64 PEs or 256 PEs, there are two $\frac{C}{D}$ cases (0.33 and 3). $\frac{C}{D}$ ratios slightly affect latency (1% to 2%) as average zero-load is mostly affected due to flit count per packet and head latency. Thus, we report latency numbers that are between those two $\frac{C}{D}$ cases.

Latency Improvement

With 64 PEs, the lowest achieved latency is achieved by C-SPN networks, with 38% less latency than the baseline. On the other hand, when PE count is increased to 256 PEs, CX2-SPN and C-SPN architectures are almost equal in terms of latency, with 45-48% savings. CX2-SPN are slightly better than C-SPN networks when $\frac{C}{D}$ becomes larger than one.

Network Partitioning Impact

When a network is partitioned, router datapaths are divided to keep overall bisection wire count constant. This introduces larger serialization latencies as packets need more cycles to be transmitted. Thus, partitioned networks, cannot offer better zero-load latency than SPN networks. When PE count is 64 PEs, the lowest latency by partitioned networks is from the C-HOM and C-HET1 networks with 25% latency savings, as PE count increases to 256 PEs, again, C-HOM and C-HET1 partitioned networks can give less latency at around 36% than the baseline.

Concentration and Express Physical Links Impact

This observation shows that CX networks are a favorable choice as PE count increases, but not the best when PE count is around 64 PEs. Although C-SPN

64 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
C-SPN	0.62
CX2-SPN	0.71
C-HOM	0.76
C-HET1	0.76
C-HET2	0.92
CX2-HOM	0.97
CX2-HET1	0.98
SPN	1.00
X2-SPN	1.10
CX2-HET2	1.25
HET1	1.26
HOM	1.26
X4-SPN	1.38
HET2	1.53
X2-HOM	1.63
X2-HET1	1.65
X4-HOM	1.99
X2-HET2	2.26
X4-HET1	2.32
X4-HET2	2.70

64 PEs, $\frac{C}{D} = 3$	
NoC	norm
C-SPN	0.63
CX2-SPN	0.69
C-HOM	0.74
C-HET1	0.75
C-HET2	0.86
CX2-HOM	0.90
CX2-HET1	0.91
SPN	1.00
X2-SPN	1.05
CX2-HET2	1.12
HOM	1.20
HET1	1.22
X4-SPN	1.28
HET2	1.43
X2-HOM	1.47
X2-HET1	1.49
X4-HOM	1.77
X2-HET2	1.98
X4-HET1	2.06
X4-HET2	2.32

256 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
CX2-SPN	0.54
C-SPN	0.54
C-HET1	0.63
C-HOM	0.63
CX4-SPN	0.65
CX2-HET1	0.71
C-HET2	0.74
CX2-HOM	0.74
X2-SPN	0.87
CX4-HET1	0.89
CX2-HET2	0.90
CX4-HOM	0.91
SPN	1.00
X4-SPN	1.01
HET1	1.17
HOM	1.18
CX4-HET2	1.19
X2-HET1	1.22
X2-HOM	1.23
HET2	1.36
X4-HOM	1.48
X4-HET1	1.60
X2-HET2	1.65
X4-HET2	1.89

256 PEs, $\frac{C}{D} = 3$	
NoC	norm
CX2-SPN	0.52
C-SPN	0.55
CX4-SPN	0.60
C-HET1	0.61
C-HOM	0.62
CX2-HET1	0.65
CX2-HOM	0.68
C-HET2	0.69
CX4-HET1	0.79
CX2-HET2	0.80
CX4-HOM	0.81
X2-SPN	0.82
X4-SPN	0.91
SPN	1.00
CX4-HET2	1.02
X2-HET1	1.09
X2-HOM	1.10
HET1	1.14
HOM	1.14
HET2	1.28
X4-HOM	1.30
X4-HET1	1.40
X2-HET2	1.43
X4-HET2	1.61

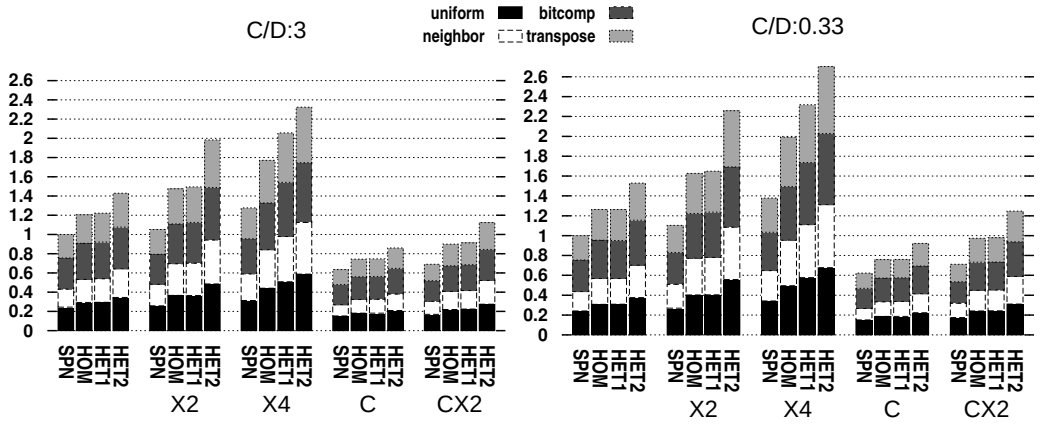
Table 4.2: Zero-load latency for 64 and 256 PEs and $\frac{C}{D} \in \{0.33, 3\}$.

has wider datapath than CX2-SPN, CX2-SPN's express physical links give lower head latencies and slightly outperform the C-SPN networks at 256 PEs. However, at 64 PEs, head latency is not as important as in 256 PEs, and overall latency can not further be improved by CX2-SPN networks. Finally an other interesting observation point is that, between X2 and X4 networks, X2 networks give better latency values, for every PE count. This happens as X4 networks have narrower router datapaths than X2 networks, imposing more serialization latency.

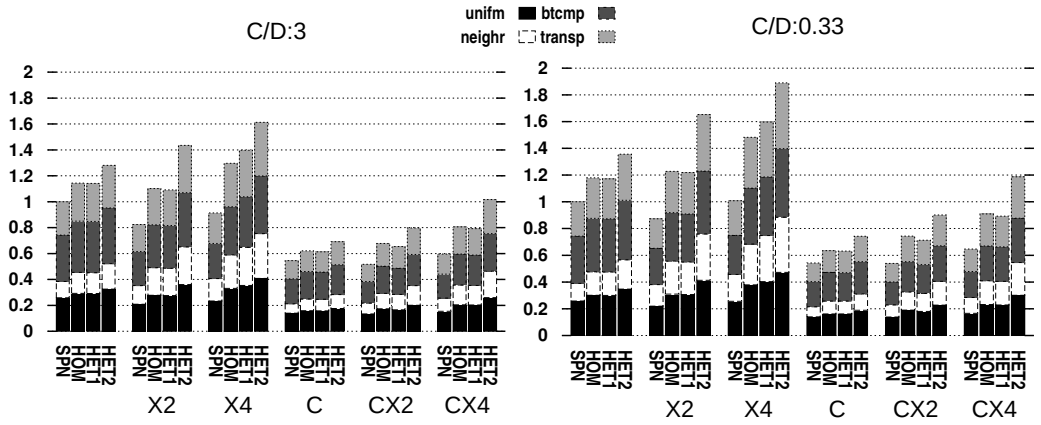
4.2.2 Transfer Time

Average transfer time per injected amount of data (1 KB) is measured in cycles and reported in normalized numbers. As mentioned earlier in Sec. 3.1.3 on Pg. 45, the saturation injection rate assumed is at the point where average latency becomes 2× times the zero-load latency.

When saturation injection rate per traffic pattern is measured, the respective saturation bandwidth is computed. Then, average transfer cycles per injected amount of data is computed and aggregate transfer times are then reported for all of traffic patterns. Depending on application, saturation bandwidth of a network affects the completion time of a source-to-destination transaction. The most com-



(i) 64PEs.



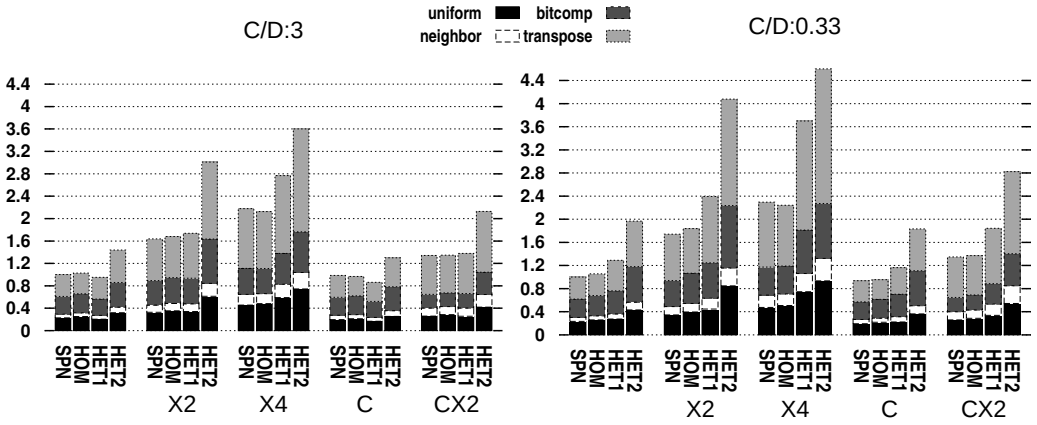
(ii) 256PEs.

Figure 4.2: Zero-load latency for 64 and 256 PEs and $\frac{C}{D} \in \{0.33, 3\}$.

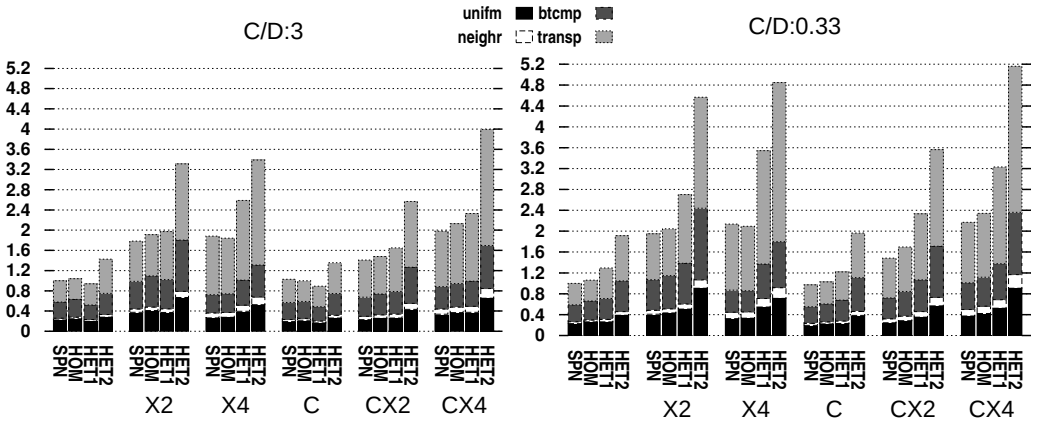
mon factors that affect saturation point for single physical networks, are how often a packet is blocked waiting for resources. In input-queuing, wormhole switching, a flit can be blocked more "easily" when it waits for a message type arbitration, free virtual channel assignment, free physical channel allocation and credits from the upstream router. For partitioned networks with multiple subnetworks, saturation point can be affected by load imbalance issues when heterogeneous networks are used. Measurements are depicted in Fig. 4.3i for 64 PEs, and Fig. 4.3ii for 256 PEs. Respective normalized values are shown in Table 4.3.

Transfer Time Savings

When few control packets ($\frac{C}{D} = 0.33$) are injected per data packet (as in explicit messaging models), the best we can take is 6% less cycles for 64 PEs, and 3% less



(i) 64PEs.



(ii) 256PEs.

Figure 4.3: Transfer time per 1KB injected data for 64 and 256 PEs and $\frac{C}{D} \in \{0.33, 3\}$.

cycles for 256 PEs using in both cases, C-SPN networks. When more ($\frac{C}{D} = 3$) control packets are used (as in shared memory models), C-HET1 NoCs outperform the baseline by 14% and 11% for 64 and 256 PEs respectively. As $\frac{C}{D}$ ratio becomes 1, C-SPN networks are better than the baseline for 64 PEs (only 4% improvement), and for 256 PEs, the baseline network is the most sufficient. If the ratio is reduced close to zero ($\frac{C}{D} = 0.33$), C-SPN seems a good option for both 64 PEs (6% improvement) and 256 PEs (3% improvement).

Express Physical Links and Concentration Impact

In terms of number of cycles per injected data, concentration helps the most. Especially, C-SPN,C-HOM and C-HET1 networks are good enough to optimize the completion of transactions in network communications. If $\frac{C}{D} = 3$ ratios are

64 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
C-SPN	0.94
C-HOM	0.95
SPN	1.00
HOM	1.06
C-HET1	1.16
HET1	1.29
CX2-SPN	1.35
CX2-HOM	1.37
X2-SPN	1.74
C-HET2	1.83
X2-HOM	1.84
CX2-HET1	1.84
HET2	1.97
X4-HOM	2.24
X4-SPN	2.30
X2-HET1	2.39
CX2-HET2	2.83
X4-HET1	3.70
X2-HET2	4.08
X4-HET2	4.60

64 PEs, $\frac{C}{D} = 3$	
NoC	norm
C-HET1	0.86
HET1	0.95
C-HOM	0.97
C-SPN	0.98
SPN	1.00
HOM	1.03
C-HET2	1.30
CX2-SPN	1.34
CX2-HOM	1.35
CX2-HET1	1.38
HET2	1.44
X2-SPN	1.64
X2-HOM	1.68
X2-HET1	1.74
X4-HOM	2.13
CX2-HET2	2.13
X4-SPN	2.18
X4-HET1	2.77
X2-HET2	3.01
X4-HET2	3.60

256 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
C-SPN	0.97
SPN	1.00
C-HOM	1.03
HOM	1.06
C-HET1	1.22
HET1	1.29
CX2-SPN	1.48
CX2-HOM	1.70
HET2	1.91
X2-SPN	1.95
C-HET2	1.96
X2-HOM	2.04
X4-HOM	2.09
X4-SPN	2.14
CX4-SPN	2.17
CX2-HET1	2.33
CX4-HOM	2.34
X2-HET1	2.70
CX4-HET1	3.23
X4-HET1	3.54
CX2-HET2	3.56
X2-HET2	4.57
X4-HET2	4.85
CX4-HET2	5.16

256 PEs, $\frac{C}{D} = 3$	
NoC	norm
C-HET1	0.89
HET1	0.94
C-HOM	1.00
SPN	1.00
C-SPN	1.03
HOM	1.04
C-HET2	1.36
CX2-SPN	1.41
HET2	1.43
CX2-HOM	1.48
CX2-HET1	1.65
X2-SPN	1.78
X4-HOM	1.84
X4-SPN	1.88
X2-HOM	1.91
X2-HET1	1.97
CX4-SPN	1.98
CX4-HOM	2.13
CX4-HET1	2.33
CX2-HET2	2.57
X4-HET1	2.59
X2-HET2	3.31
X4-HET2	3.39
CX4-HET2	4.00

Table 4.3: Average transfer time per KB for 64 and 256 PEs when $\frac{C}{D} \in \{0.33, 3\}$.

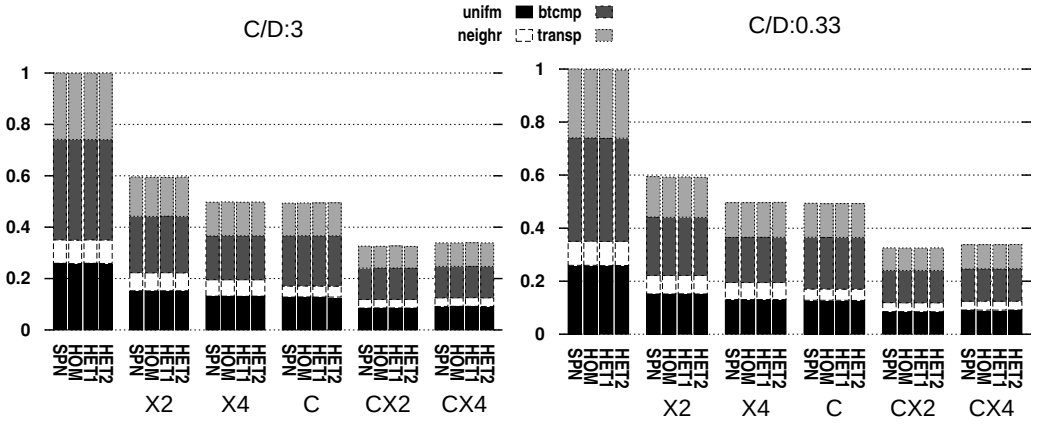
used, C-HET1 seems the best choice. If other $\frac{C}{D}$ ratios are used, then C-SPN and C-HOM networks are the best choices.

Packet Type Distribution Impact

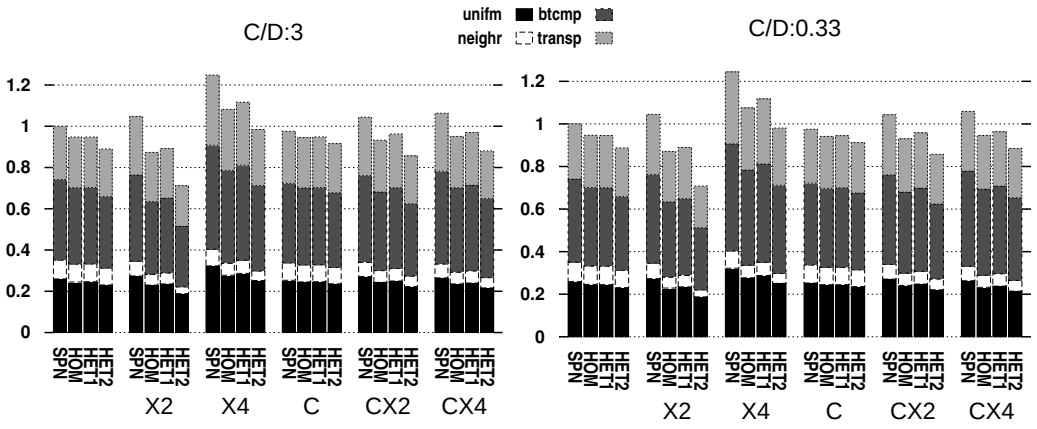
We can conclude that C-HET1 networks can be advantageous in terms of saturation bandwidth, as long as control packets are more than data packets. However, in cases that control packets are equal or less than data packets, C-SPN networks are good enough for both 64 PEs and 256 PEs. As it is explained in Sec. 4.3.1 on Pg.66 for example, when larger/smaller volume of control packets (for i.e. read requests) are injected relative to data volume (for i.e. read replies), response/request subnetwork can be under/over-utilized (for i.e. request subnetwork could reach its saturation throughput before response does, over-utilizing request subnetwork and under-utilizing response subnetwork, or response subnetwork could first reach in saturation state, before request subnetwork, and thus, over-utilizing response subnetwork and under-utilizing request subnetwork) imposing load imbalance between subnetworks and overall throughput degradation.

4.2.3 Utilized Link Length

Although average hop count of a NoC may be less than another, average switching link length may be equal or greater. Thus, a comparison is investigated here between average hop count, and average link length utilization to clarify in which case (which network architecture) we have the longer wire length utilization.



(i) Average hop count.



(ii) Average link length utilization.

Figure 4.4: Average hop count and link length utilization for 256 PEs when $\frac{C}{D} \in \{0.33, 3\}$.

Concentration and Express Physical Links Impact

In Fig. 4.4i we see the effect of concentration to the average hop count. Average hop count becomes smaller as concentration is applied. The wire length utilization on the other hand, slightly increases as concentration or express link are applied. Furthermore, between express links of express interval equal to 2 and 4, X2 networks have less average hop count than X4 networks and X4 networks have slightly

more wire length utilized than X2 networks. This show us that in X4 networks, there are not as many possibilities as in X2 networks for a packet, to take an express physical channel. On average, when a packet in X4 networks is going to travel through a path of less than 4 hops, it will use local links.

Network Partitioning Impact

As network partitioning is applied, partitioned networks are having less switching link length utilized than non partitioned ones. This happens because partitioned networks occupy more area, and thus less link routing is needed to connect bigger router blocks.

4.3 Power and Energy Analysis

This section reports: a) the overall average power and b) the average remote transfer energy per one KB of injected data with its respective energy efficiency in EDP (Energy Delay Product). Router and link's activity factors (obtained from the simulator) were captured when the average packet latency had reached 2× times the zero-load packet latency (similar to Wang et. al. [52]). Further details of the methodology are found in Sec. 3.2 on Pg. 47, and depicted in high level abstraction in Fig. 3.3.

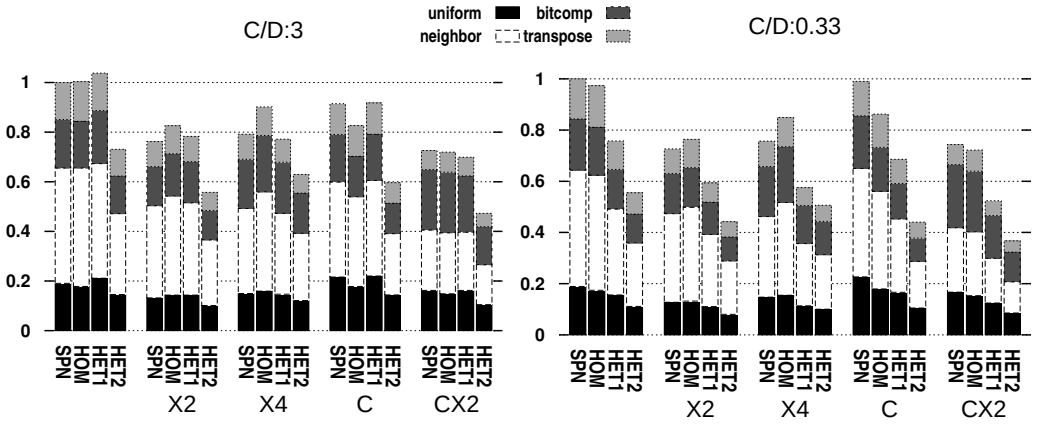
Four synthetic workloads (uniform, neighbor, bit complement and transpose) are used in power-energy simulations, similarly as in performance simulations, and a range of $\frac{C}{D}$ ratios are explored. $\frac{C}{D}$ ratios were explored to see how vulnerable heterogeneous architectures could become. Again, as in performance evaluation, two cases of control packet to data packet ratios are shown for clarity, 3 and 0.33. Each NoC's total power dissipation, is comprised of router and piped/non-piped link power (detailed power modeling is described in Sec. 3.1.3 on Pg. 45). In addition, as described in Sec.3.2.2 on Pg. 49, LSTP (Low Standby Power) process was used for the evaluation. In following paragraphs, total power is reported.

4.3.1 Power Dissipation

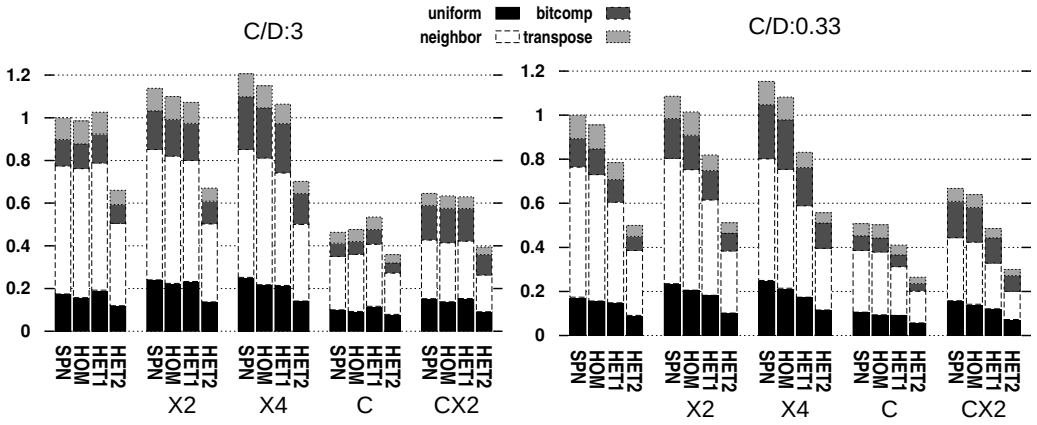
Normalized values are shown in table 4.4, and depicted in Fig. 4.5i for 64 PEs and in Fig. 4.6i for 256 PEs. For 64 and 256 PEs, CX2-HET2 networks for either $\frac{C}{D} = 3$ or $\frac{C}{D} = 0.33$ cases, we have the least total NoC power dissipation. 64 PE configuration at $\frac{C}{D} = 0.33$ and $\frac{C}{D} = 3$, shows 63% and 53% less power dissipation respectively, and, for 256 PEs, at $\frac{C}{D} = 0.33$ and $\frac{C}{D} = 3$, 68% and 61% less power is dissipated respectively.

Utilization

To explore whether flit rate utilization is the main contributor to power dissipation, we show average injected flit rates for each network (Fig. 4.5ii, 4.6ii). We can



(i) Total power dissipation.



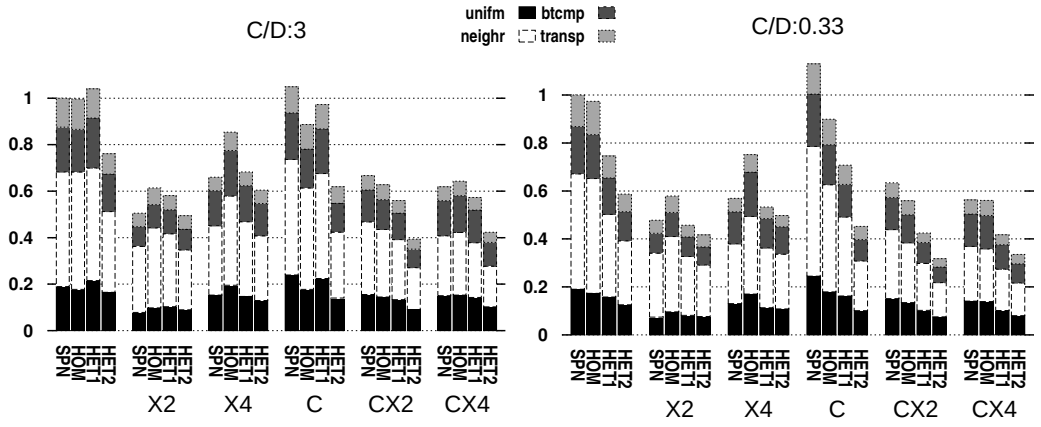
(ii) Average utilization.

Figure 4.5: Total power and average utilization for 64 and $\frac{C}{D} \in \{0.33, 3\}$.

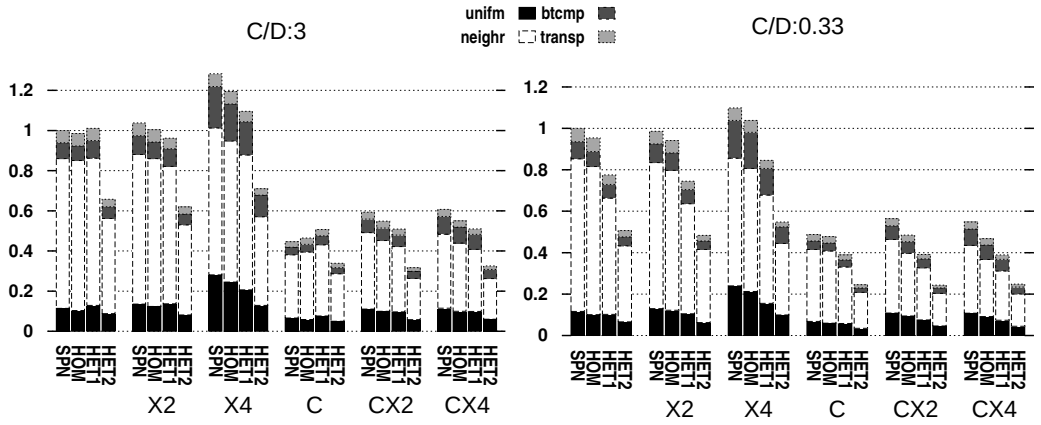
safely say that reduced NoC power dissipation of CX2-HET2 NoCs, happens due to under-utilization of HET2 partitioning. heterogeneous networks in general, as seen in Fig. 4.5ii, 4.6ii, are having more utilization degradation as $\frac{C}{D}$ ratio tends to zero. Respective power is further degraded as seen in Fig. 4.5i, 4.6i. This packet distribution effect is further explained in the following section.

Packet Distribution Impact on Heterogeneous Networks

To understand why Power dissipation of heterogeneous partitioning is greatly influenced by the $\frac{C}{D}$ ratio, we illustrate a simple example. Let a partitioned NoC, sliced into 2 subnetworks, each one with the same flit-width. In subnetwork A, only data packets are allowed (i.e. read replies) and in subnetwork B, only control packets (read requests). Saturation flit rate of subnetwork A is 25%, and subnet-



(i) Total power dissipation.



(ii) Average utilization.

Figure 4.6: Total power and average utilization for 256 PEs when $\frac{C}{D} \in \{0.33, 3\}$.

work B's is 25% as well. Each data packet size is twice the control packet size and the communication model imposes more control packets per data packet, which is on average 4 control packets per data packet. Both network's flit width is 64bit. Data packet is equal to 128bits and control packet equal to 64bits. From the above assumptions, we have that subnetwork's A and B's saturation bandwidth is 16 *bits/cycle*. Let's suppose that one captures network's state at saturation point, due to the fact that (in this example) data packets are generated only if control packets are generated first, dependence is created between data-subnetwork and control-subnetwork. Subnetwork B will start injecting control packets and subnetwork A will inject respective data packets. Let us see which subnetwork will reach its saturation point first. For every cycle, we need 4 control packets per data packets. This means, twice control volume on average per data volume. Thus, when NoC does not saturated yet, suppose, that 8 *bits/cycle* are injected into subnet-

64 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
CX2-HET2	0.37
C-HET2	0.44
X2-HET2	0.44
X4-HET2	0.51
CX2-HET1	0.52
HET2	0.56
X4-HET1	0.58
X2-HET1	0.59
C-HET1	0.69
CX2-HOM	0.72
X2-SPN	0.73
CX2-SPN	0.74
X4-SPN	0.76
HET1	0.76
X2-HOM	0.76
X4-HOM	0.85
C-HOM	0.86
HOM	0.97
C-SPN	0.99
SPN	1.00

64 PEs, $\frac{C}{D} = 3$	
NoC	norm
CX2-HET2	0.47
X2-HET2	0.56
C-HET2	0.60
X4-HET2	0.63
CX2-HET1	0.70
CX2-HOM	0.72
CX2-SPN	0.73
HET2	0.73
X2-SPN	0.76
X4-HET1	0.77
X2-HET1	0.78
X4-SPN	0.79
X2-HOM	0.83
C-HOM	0.83
X4-HOM	0.90
C-SPN	0.92
C-HET1	0.92
SPN	1.00
HOM	1.00
HET1	1.04

256 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
CX2-HET2	0.32
CX4-HET2	0.34
X2-HET2	0.42
CX4-HET1	0.42
CX2-HET1	0.42
C-HET2	0.45
X2-HET1	0.46
X2-SPN	0.48
X4-HET2	0.50
X4-HET1	0.53
CX2-HOM	0.56
CX4-HOM	0.56
CX4-SPN	0.56
X4-SPN	0.57
X2-HOM	0.58
HET2	0.59
CX2-SPN	0.63
C-HET1	0.71
HET1	0.75
X4-HOM	0.75
C-HOM	0.90
HOM	0.97
SPN	1.00
C-SPN	1.13

256 PEs, $\frac{C}{D} = 3$	
NoC	norm
CX2-HET2	0.39
CX4-HET2	0.42
X2-HET2	0.49
X2-SPN	0.51
CX2-HET1	0.56
CX4-HET1	0.57
X2-HET1	0.58
X4-HET2	0.60
X2-HOM	0.61
CX4-SPN	0.62
C-HET2	0.62
CX2-HOM	0.63
CX4-HOM	0.64
X4-SPN	0.66
CX2-SPN	0.67
X4-HET1	0.68
HET2	0.76
X4-HOM	0.85
C-HOM	0.89
C-HET1	0.97
HOM	1.00
SPN	1.00
HET1	1.04
C-SPN	1.05

Table 4.4: Total power dissipation for 64 and 256 PEs when $\frac{C}{D} \in \{0.33, 3\}$.

work B on average, then, 4 *bits/cycle* are injected on average, in subnetwork A. As control packet rate is increasing, there will be a time period that 16 *bits/cycle* (saturation area) are injected on average in subnetwork B. At that time, subnetwork A, will be injecting at 8 *bits/cycle* on average. So, aggregate *bits/cycle*, would be 24 *bits/cycle*. However, saturation point of both subnetworks is 16 *bits/cycle*, and thus, overall saturation bandwidth will only be 24 *bits/cycle*. Thus, subnetwork A (data network) has been under-utilized whereas subnetwork B over-utilized.

Homogeneous versus Heterogeneous Networks

In terms of network partitioning, we conclude that: a) heterogeneous NoC's utilization is always smaller than homogeneous (see Fig. 4.5ii, 4.6ii), however, there is an exception where C-HET1 NoCs, are better utilized than C-HOM NoCs. b) HET NoCs that have less control packets than data packets, are under-utilized. On the other hand, HOM NoCs are not affected from $\frac{C}{D}$ ratios. The most utilized NoCs for 64 PEs and 256 PEs are X4-SPN NoCs. They give 15% (for $\frac{C}{D} = 0.33$) and 20%(for $\frac{C}{D} = 3$) more utilization than baseline for 64 PEs, 10%($\frac{C}{D} = 0.33$) and 24% ($\frac{C}{D} = 3$) more utilization for 256 PEs.

4.3.2 Energy Consumed

Energy Savings

Normalized values are shown in table 4.5, and depicted in Fig. 4.7i for 64 PEs and in Fig. 4.7ii for 256 PEs. The X2-SPN networks are having the least average remote transfer energy per KB for either 64 and 256 PEs. 64 PE configuration (at $\frac{C}{D} = 0.33$ and $\frac{C}{D} = 3$) shows 41% less consumed energy, and, for 256 PEs (at $\frac{C}{D} = 0.33$ and $\frac{C}{D} = 3$) 62% less energy is consumed too.

64 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
X2-SPN	0.59
X2-HET1	0.69
X2-HOM	0.72
X4-SPN	0.89
HET1	0.97
SPN	1.00
X2-HET2	1.01
X4-HOM	1.03
HOM	1.04
HET2	1.12
CX2-HET1	1.12
CX2-SPN	1.14
X4-HET1	1.16
CX2-HOM	1.16
CX2-HET2	1.23
X4-HET2	1.34
C-HET1	1.41
C-HET2	1.43
C-HOM	1.45
C-SPN	1.61

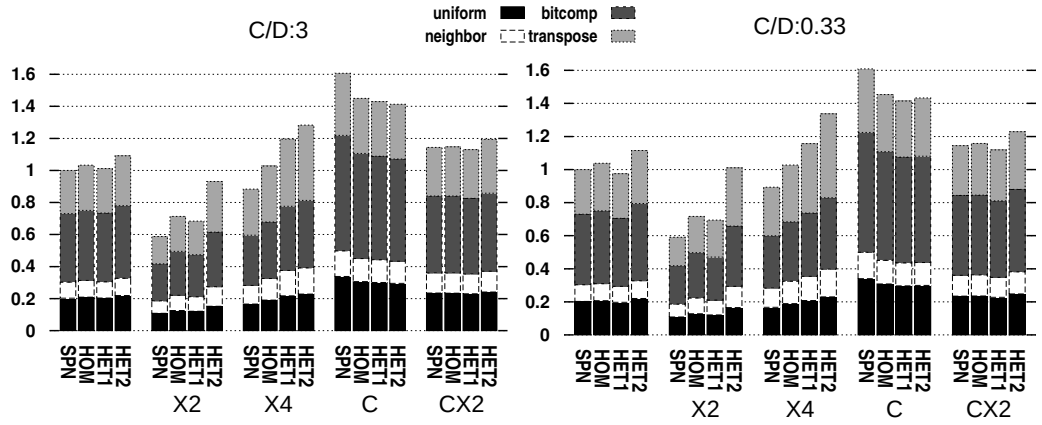
64 PEs, $\frac{C}{D} = 3$	
NoC	norm
X2-SPN	0.59
X2-HET1	0.68
X2-HOM	0.71
X4-SPN	0.88
X2-HET2	0.93
SPN	1.00
HET1	1.01
X4-HOM	1.03
HOM	1.03
HET2	1.09
CX2-HET1	1.13
CX2-SPN	1.14
CX2-HOM	1.15
CX2-HET2	1.19
X4-HET1	1.20
X4-HET2	1.28
C-HET2	1.41
C-HET1	1.43
C-HOM	1.45
C-SPN	1.61

256 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
X2-SPN	0.38
X4-SPN	0.43
X2-HOM	0.55
X2-HET1	0.56
X4-HOM	0.62
X4-HET1	0.71
CX2-SPN	0.74
CX2-HET1	0.75
CX2-HOM	0.77
CX2-HET2	0.93
HET1	0.96
SPN	1.00
X4-HET2	1.00
CX4-SPN	1.01
X2-HET2	1.03
HOM	1.04
CX4-HET1	1.07
CX4-HOM	1.14
HET2	1.16
C-HET1	1.17
C-HOM	1.23
C-HET2	1.24
C-SPN	1.48
CX4-HET2	1.54

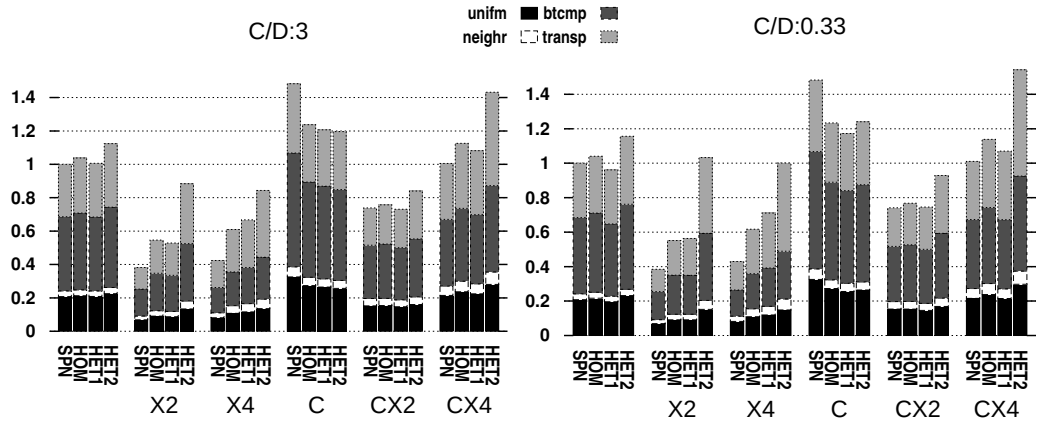
256 PEs, $\frac{C}{D} = 3$	
NoC	norm
X2-SPN	0.38
X4-SPN	0.42
X2-HET1	0.53
X2-HOM	0.55
X4-HOM	0.61
X4-HET1	0.67
CX2-HET1	0.73
CX2-SPN	0.74
CX2-HOM	0.76
CX2-HET2	0.84
X4-HET2	0.84
X2-HET2	0.88
SPN	1.00
CX4-SPN	1.01
HET1	1.01
HOM	1.04
CX4-HET1	1.08
HET2	1.12
CX4-HOM	1.13
C-HET2	1.20
C-HET1	1.21
C-HOM	1.24
CX4-HET2	1.43
C-SPN	1.48

Table 4.5: Remote transfer energy per 1KB of injected data for 64 and 256 PEs when $\frac{C}{D} \in \{0.33, 3\}$.

Network Partitioning Impact In terms of network partitioning, for 64 PEs, X2-HET1 networks can give up to 31% savings (less than X2-SPN's 41%) slightly more than X2-HOM networks (28%). For 256 PEs, X2-HET1 and X-HOM have similar savings (around 45%). If only network partitioning is applied to the baseline, the only possible configurations that give (almost negligible) energy savings, are HET1 partitioned networks, giving 3% less energy for 64 PEs and 4% less energy for 256 PEs. However this happens as long as $\frac{C}{D} = 0.33$. When $\frac{C}{D} = 3$, we have 1% more energy consumed for 64 and 256 PEs.



(i) 64 PEs.

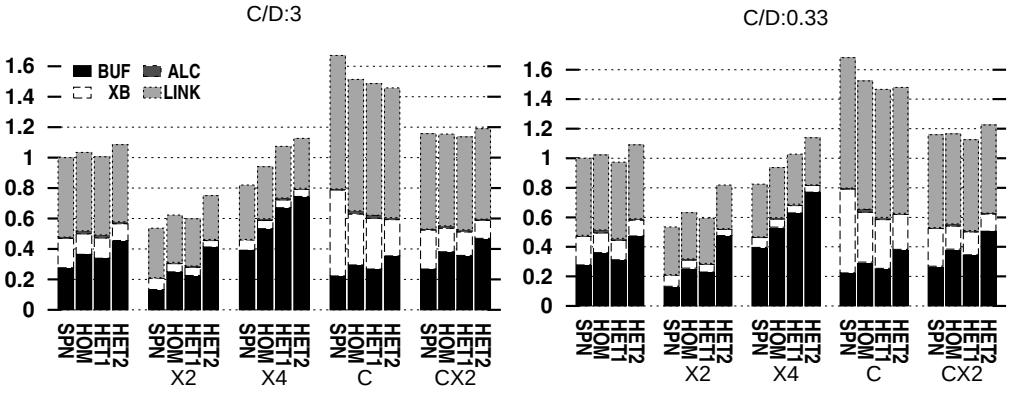


(ii) 256 PEs.

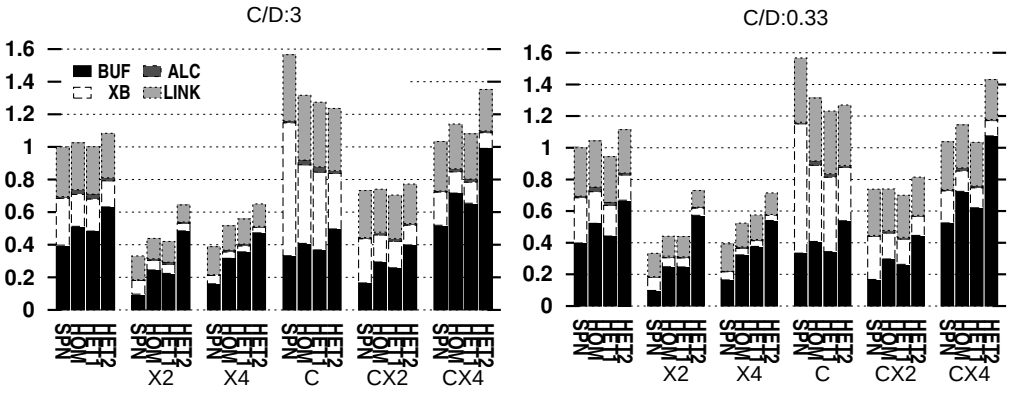
Figure 4.7: Remote transfer energy per 1KB injected data for 64 and 256 PEs when $\frac{C}{D} \in \{0.33, 3\}$.

Concentration and Express Physical Links Impact Energy savings obtained from X2-SPN networks show that router by-passing saves enormous amounts of energy. Combined concentration with express physical links saves energy too, but using express physical links exclusively is better. Another observation is that when concentration-only applied, does not give any energy savings. Neither for 64 PEs nor 256 PEs. This happens because crossbar and link energy increases. Router breakdown is shown in Fig. 4.8i,4.8ii

Packet Distribution Impact For 64, 256 PE count cases, energy consumed, remain almost constant regardless of $\frac{C}{D}$ ratio. This is true for non heterogeneous networks, as average utilization per network is not affected too much from $\frac{C}{D}$ ratio variations. But in heterogeneous architectures, average utilization per network is



(i) 64 PEs.



(ii) 256 PEs.

Figure 4.8: Router energy (uniform case) for 64 and 256 PEs when $\frac{C}{D} \in \{0.33, 3\}$.

affected from $\frac{C}{D}$ ratio (as we saw in Sec. 4.3.1 on Pg. 65) but energy consumed is almost the same regardless of $\frac{C}{D}$ ratio. This happens because utilization is inversely proportional to the number of cycles per injected amount of data and proportional to power dissipation (average energy per cycle). When utilization is decreased average energy per cycle is decreased, but the total number of cycles that are needed per amount of injected data, is increased, and thus energy ($\frac{energy}{cycle} \times cycles$) remains almost the same for each $\frac{C}{D}$ ratio.

Preferred Network Architecture It seems that X2-SPN networks are the best in terms of consumed energy, even from X4 networks for every PE count. For 256 PEs, energy savings can be up to 62%. The main reason comes from the fact that average hop count is decreasing (Sec. 4.2.3 on Pg. 64 and Fig. 4.4i) and that utilization is kept constant relative to baseline (Sec. 4.3.1 on Pg. 65). If concentration is exclusively used, there are not any energy savings. For 256 PEs,

concentration combined with X2 physical links, can give energy savings, but up to 25%. These savings are far less than X2-SPN networks, which are 62%.

4.3.3 Energy Efficiency

As we observe from EDP results (table 4.6, Fig. 4.9i for 64 PEs and Fig. 4.9ii for 256 PEs) for 64 PEs, when $\frac{C}{D} = 0.33$, baseline and X2-SPN perform equally in terms of EDP. When $\frac{C}{D} = 3$, EDP savings are ≈7% for X2-SPN and ≈4% for HET1 networks. For 256 PEs, X2-SPN networks have 26% less EDP for $\frac{C}{D} = 0.33$ case, 32% less EDP for $\frac{C}{D} = 3$ case.

64 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
X2-SPN	1.00
SPN	1.00
HOM	1.11
HET1	1.26
X2-HOM	1.29
CX2-SPN	1.40
C-HOM	1.40
CX2-HOM	1.45
C-SPN	1.50
C-HET1	1.65
X2-HET1	1.66
CX2-HET1	1.91
X4-SPN	2.01
HET2	2.22
X4-HOM	2.25
C-HET2	2.64
CX2-HET2	3.25
X2-HET2	4.19
X4-HET1	4.39
X4-HET2	6.41

64 PEs, $\frac{C}{D} = 3$	
NoC	norm
X2-SPN	0.93
HET1	0.96
SPN	1.00
HOM	1.07
X2-HET1	1.17
X2-HOM	1.17
C-HET1	1.23
CX2-SPN	1.39
CX2-HOM	1.40
C-HOM	1.41
CX2-HET1	1.42
HET2	1.57
C-SPN	1.57
C-HET2	1.83
X4-SPN	1.88
X4-HOM	2.14
CX2-HET2	2.35
X2-HET2	2.81
X4-HET1	3.31
X4-HET2	4.73

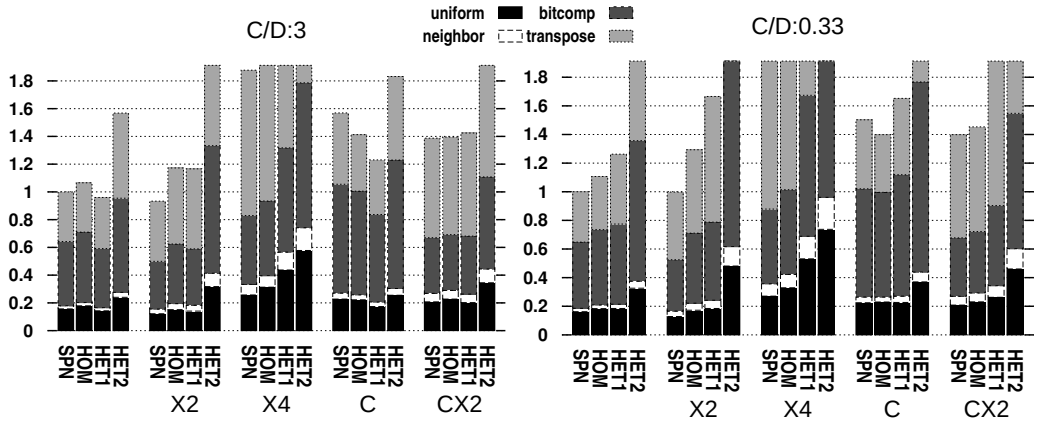
256 PEs, $\frac{C}{D} = 0.33$	
NoC	norm
X2-SPN	0.74
X4-SPN	0.94
SPN	1.00
CX2-SPN	1.06
HOM	1.11
X2-HOM	1.14
C-HOM	1.25
HET1	1.26
CX2-HOM	1.27
X4-HOM	1.38
C-HET1	1.41
C-SPN	1.42
X2-HET1	1.58
CX2-HET1	1.73
CX4-SPN	2.13
HET2	2.25
C-HET2	2.43
CX4-HOM	2.61
X4-HET1	2.84
CX2-HET2	3.40
CX4-HET1	3.53
X2-HET2	5.02
X4-HET2	5.98
CX4-HET2	8.31

256 PEs, $\frac{C}{D} = 3$	
NoC	norm
X2-SPN	0.68
X4-SPN	0.82
HET1	0.95
SPN	1.00
CX2-SPN	1.00
X2-HOM	1.05
C-HET1	1.06
X2-HET1	1.06
HOM	1.08
CX2-HOM	1.09
CX2-HET1	1.18
X4-HOM	1.20
C-HOM	1.22
C-SPN	1.50
C-HET2	1.60
HET2	1.63
X4-HET1	1.89
CX4-SPN	1.93
CX2-HET2	2.16
CX4-HOM	2.35
CX4-HET1	2.51
X2-HET2	3.05
X4-HET2	3.31
CX4-HET2	5.97

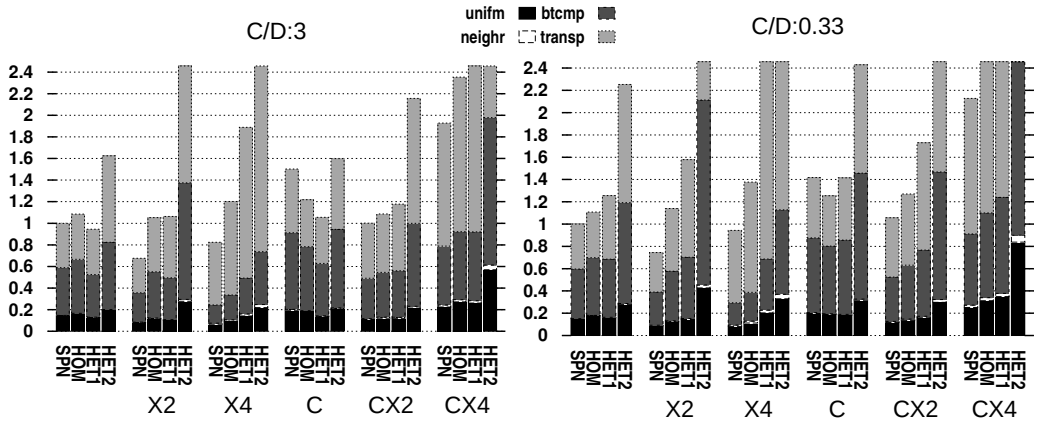
Table 4.6: Energy delay product (EDP) for 64 and 256 PEs when $\frac{C}{D} \in \{0.33, 3\}$.

Network Partitioning Impact

Partitioned networks seems to offer savings only if HET1 type of network architecture is used, and $\frac{C}{D}$ ratio is more than two to three. Otherwise, partitioned networks cannot overcome energy inefficiencies.



(i) 64 PEs.



(ii) 256 PEs

Figure 4.9: Energy delay product (EDP) for 64 and 256 PEs when $\frac{C}{D} \in \{0.33, 3\}$.

Concentration and Express Physical Links Impact

Energy efficiency is obtained by X2-SPN networks for 64 and 256 PEs. In 64 PEs, we observe that X2-SPN and the baseline network are sufficient enough for energy efficiency for $\frac{C}{D} = 0.33$ ratio, but, as $\frac{C}{D}$ increases, X2-SPN are better. In general, for every $\frac{C}{D}$ ratio, X2-SPN are better than baseline. As PE count increases, when express physical links are used exclusively with express interval equal to 2, X2-SPN networks are the best in terms of energy efficiency.

Packet Distribution Impact

Here packet distribution of control and data packets play significant role mostly in HET1 or HET2 partitioned networks. The worst EDP values are obtained when less control packets per data packets are transmitted and heterogeneous architectures

are applied. For example in 64 PEs, the worst EDP (6.4× times the baseline) is by X4-HET2 network at $\frac{C}{D} = 0.33$ and for 256 PEs (8.3× times worst than the baseline), is by CX4-HET2 at $\frac{C}{D} = 0.33$.

Preferred Network Architecture

For both 64 and 256 PEs, the most preferred architecture in terms of EDP, is the X2-SPN. EDP savings are the best for every PE count and $\frac{C}{D}$ ratios.

4.4 Summary

This section summarizes evaluation results into a unified table (4.7 and 4.8 for 64 PEs and 256 PEs). The first column presents the the network architecture, and the rest columns are as follows:

1. total area (**AREA**),
2. zero load latency (**LAT**),
3. transfer time per KB (**TM**),
4. total power (**PW**),
5. energy consumed per KB (**EN**),
6. energy delay product (**EDP**) and
7. energy delay area product (**EDAP**).

For convenience, the network architectures are ordered according to *EDAP* value in descending manner. This means that network architectures residing on top of the table, have the best EDAP. *EDAP* was chosen, similarly used as in [53], because it includes both the operational cost (energy) as well as a capital cost (area). Below we present the evaluated results for 64 PEs and 256 PEs:

Area:
64 PEs: Area savings do exist only for the networks that are not partitioned and utilize express physical links of express interval equal to 2 or 4. These networks are CX2-SPN(11%), X2-SPN(5%), X4-SPN(6%).

256 PEs: Significantly area savings do exist but only for those that are not partitioned and use express physical links either express interval of 2 or 4 (CX2-HET1 gives 1% savings). These networks are X4-SPN (20%), X2-SPN (18%), CX4-SPN (17%), CX2-SPN (16%).

Zero-Load Latency:
64 PEs: Zero-Load latency savings are mostly obtained from concentrated networks. These networks are: C-{SPN, HOM, HET1, HET2}, CX2-{SPN, HOM, HET1}. When $\frac{C}{D} \in \{0.33, 3\}$ C-SPN gives 38% savings.

256 PEs: Zero-Load latency savings are obtained mostly from C, CX2 and CX4 networks, such as {C,CX2}-{SPN, HOM, HET1, HET2} and CX4-{SPN, HOM, HET1}. And X2-SPN networks for $\frac{C}{D} \in \{0.33, 3\}$ and X4-SPN when $\frac{C}{D} = 0.33$ Most savings come from CX2-SPN networks by 46% (when $\frac{C}{D} = 0.33$) and 48% (when $\frac{C}{D} = 3$).

Transfer Time:
64 PEs: In network's saturation mode, the least total cycles needed per injected amount of data is given by C-{SPN,HOM} networks as long as few control packets per data packet are injected, and by C-{SPN, HOM, HET1}, HET1 when control packets per data packet ratio is increasing. The best average transfer time for $\frac{C}{D} = 0.33$ is achieved by C-SPN networks (6%), and by C-HET1 (14%) partitioned networks when $\frac{C}{D} = 3$.

256 PEs: When network's state is in saturation mode, the least total cycles needed per injected amount of data is by C-SPN (at 3%) networks when few control packets per data packet are injected, and by C-HET1 (at 11%) and HET1 (at 4%) when control packets per data packet ratio is increasing.

Power:
64 PEs: We have observed that power savings exist almost for every network architecture that has been resulted from the baseline (except the case that we have many control packets per data packet and network architecture is HET1 partitioned network). Most power savings are gained from CX2-HET2 (63% for $\frac{C}{D} = 0.33$ and 53% for $\frac{C}{D} = 3$) architectures due to their high under-utilization relative to the other networks.

256 PEs: We observe that power savings exist almost for every network architecture that has been resulted from baseline (except the case for $\frac{C}{D} = 0.33$ and C-SPN, and $\frac{C}{D} = 3$ and HOM, HET1 and C-SPN.) The most power savings are coming from CX2-HET2 architectures due to their high under-utilization relative to other networks.

Energy:
64 PEs: Energy savings are mostly obtained from X2-{SPN, HOM, HET1} and X4-SPN networks for $\frac{C}{D} \in \{0.33, 3\}$ ratios, and additionally from X2-HET2 (7% savings) partitioned networks when $\frac{C}{D} = 3$. The best energy savings are from X2-SPN networks at around 41%.

256 PEs: Energy savings are mostly obtained from X2-{SPN,HOM,HET1}, X4-{SPN,HOM,HET1}, CX2-{SPN,HOM,HET1} for $\frac{C}{D} \in \{0.33, 3\}$ ratios, and additionally from HET1 partitioned networks when $\frac{C}{D} = 0.33$ (slightly savings:4%) and from X2-HET2, X4-HET2 and CX2-HET2 when $\frac{C}{D} = 3$. The enormous energy savings are from X2-SPN networks by 62%.

EDP:
64 PEs: When few control packets per data packet are injected ($\frac{C}{D} = 0.33$) X2-SPN and baseline is sufficient enough keeping the best EDP relative to the rest of network architectures. As control packets per data packet is increasing, X2-SPN network has 7% better EDP than the baseline.

256 PEs: When few control packets per data packet are injected ($\frac{C}{D} = 0.33$) the only energy efficient network architectures are X2-SPN and X4-SPN giving 26% and 4% savings respectively. As control packets per data packet is increasing, X2-SPN network have still the best EDP than baseline with 32% savings and X4-SPN networks with 18%.

EDAP:
64 PEs: Our results shows that EDAP is able to decrease regardless of $\frac{C}{D}$ ratio as long as X2-SPN networks are used. The numbers are: 5% ($\frac{C}{D} = 0.33$) to 11% ($\frac{C}{D} = 3$) better EDAP for 64 PEs, and 41% ($\frac{C}{D} = 0.33$) to 46% ($\frac{C}{D} = 0.33$) for 256 PEs.

256 PEs: As PE count increases {X2,X4,CX2}-SPN networks have the best EDAP. X2-SPN networks show 41% better EDAP than baseline when $\frac{C}{D} = 0.33$ and 46% better EDAP when $\frac{C}{D} = 3$. The rest of network architectures give better EDAP as follows: X4-SPN: up to 28% when $\frac{C}{D} = 0.33$, and 37% when $\frac{C}{D} = 3$. CX2-SPN: up to 14% when $\frac{C}{D} = 0.33$, and 19% when $\frac{C}{D} = 3$.

NOC	AREA	LAT	TM	PW	EN	EDP	EDAP
X2-SPN	0.95	1.10	1.74	0.73	**0.59**	**1.00**	**0.95**
SPN	1.00	1.00	1.00	1.00	1.00	1.00	**1.00**
CX2-SPN	**0.89**	0.71	1.35	0.74	1.14	1.40	1.28
HOM	1.12	1.26	1.06	0.97	1.04	1.11	1.29
HET1	1.11	1.26	1.29	0.76	0.97	1.26	1.46
CX2-HOM	1.03	0.97	1.37	0.72	1.16	1.45	1.60
C-SPN	1.08	**0.62**	**0.94**	0.99	1.61	1.50	1.65
X2-HOM	1.22	1.63	1.84	0.76	0.72	1.29	1.70
C-HOM	1.17	0.76	0.95	0.86	1.45	1.40	1.72
C-HET1	1.16	0.76	1.16	0.69	1.41	1.65	2.00
X4-SPN	0.94	1.38	2.30	0.76	0.89	2.01	2.00
CX2-HET1	1.01	0.98	1.84	0.52	1.12	1.91	2.06
X2-HET1	1.17	1.65	2.39	0.59	0.69	1.66	2.10
X4-HOM	1.20	1.99	2.24	0.85	1.03	2.25	2.98
HET2	1.26	1.53	1.97	0.56	1.12	2.22	3.00
C-HET2	1.22	0.92	1.83	0.44	1.43	2.64	3.45
CX2-HET2	1.14	1.25	2.83	**0.37**	1.23	3.25	4.02
X4-HET1	1.19	2.32	3.70	0.58	1.16	4.39	5.96
X2-HET2	1.46	2.26	4.08	0.44	1.01	4.19	6.99
X4-HET2	1.43	2.70	4.60	0.51	1.34	6.41	10.48

(i) $\frac{C}{D} = 0.33$

NOC	AREA	LAT	TM	PW	EN	EDP	EDAP
X2-SPN	0.95	1.05	1.64	0.76	**0.59**	**0.93**	**0.89**
SPN	1.00	1.00	1.00	1.00	1.00	1.00	1.00
HET1	1.11	1.22	0.95	1.04	1.01	0.96	1.11
HOM	1.12	1.20	1.03	1.00	1.03	1.07	1.24
CX2-SPN	**0.89**	0.69	1.34	0.73	1.14	1.39	1.26
X2-HET1	1.17	1.49	1.74	0.78	0.68	1.17	1.47
C-HET1	1.16	0.75	**0.86**	0.92	1.43	1.23	1.49
CX2-HOM	1.03	0.90	1.35	0.72	1.15	1.40	1.53
CX2-HET1	1.01	0.91	1.38	0.70	1.13	1.42	1.54
X2-HOM	1.22	1.47	1.68	0.83	0.71	1.17	1.54
C-HOM	1.17	0.74	0.97	0.83	1.45	1.41	1.73
C-SPN	1.08	**0.63**	0.98	0.92	1.61	1.57	1.73
X4-SPN	0.94	1.28	2.18	0.79	0.88	1.88	1.86
HET2	1.26	1.43	1.44	0.73	1.09	1.57	2.12
C-HET2	1.22	0.86	1.30	0.60	1.41	1.83	2.40
X4-HOM	1.20	1.77	2.13	0.90	1.03	2.15	2.84
CX2-HET2	1.14	1.12	2.13	**0.47**	1.19	2.35	2.91
X4-HET1	1.19	2.06	2.77	0.77	1.20	3.31	4.49
X2-HET2	1.46	1.98	3.01	0.56	0.93	2.81	4.69
X4-HET2	1.43	2.32	3.60	0.63	1.28	4.73	7.73

(ii) $\frac{C}{D} = 3$

Table 4.7: Evaluation results for $\frac{C}{D} \in \{0.33, 3\}$ (64PEs). Bold numbers shows the least value per column.

NOC	AREA	LAT	TM	PW	EN	EDP	EDAP
X2-SPN	0.82	0.87	1.95	0.48	**0.38**	**0.74**	**0.59**
X4-SPN	**0.80**	1.01	2.14	0.57	0.43	0.94	0.72
CX2-SPN	0.84	**0.54**	1.48	0.63	0.74	1.06	0.86
SPN	1.00	1.00	1.00	1.00	1.00	1.00	1.00
HOM	1.13	1.18	1.06	0.97	1.04	1.11	1.28
CX2-HOM	1.05	0.74	1.70	0.56	0.77	1.27	1.33
X2-HOM	1.18	1.23	2.04	0.58	0.55	1.14	1.38
HET1	1.13	1.17	1.29	0.75	0.96	1.26	1.45
C-HOM	1.21	0.63	1.03	0.90	1.23	1.25	1.53
C-SPN	1.13	**0.54**	**0.97**	1.13	1.48	1.42	1.60
C-HET1	1.20	0.63	1.22	0.71	1.17	1.41	1.71
CX2-HET1	0.99	0.71	2.33	0.42	0.75	1.73	1.71
X4-HOM	1.24	1.48	2.09	0.75	0.62	1.38	1.75
CX4-SPN	0.83	0.65	2.17	0.56	1.01	2.13	1.76
X2-HET1	1.14	1.22	2.70	0.46	0.56	1.58	1.84
CX4-HOM	1.04	0.91	2.34	0.56	1.14	2.61	2.78
HET2	1.29	1.36	1.91	0.59	1.16	2.25	3.01
C-HET2	1.29	0.74	1.96	0.45	1.24	2.43	3.18
X4-HET1	1.11	1.60	3.54	0.53	0.71	2.84	3.24
CX4-HET1	1.01	0.89	3.23	0.42	1.07	3.53	3.63
CX2-HET2	1.20	0.90	3.56	**0.32**	0.93	3.40	4.17
X2-HET2	1.59	1.65	4.57	0.42	1.03	5.02	8.53
X4-HET2	1.54	1.89	4.85	0.50	1.00	5.98	9.68
CX4-HET2	1.17	1.19	5.16	0.34	1.54	8.31	10.14

(i) $\frac{C}{D} = 0.33$

NOC	AREA	LAT	TM	PW	EN	EDP	EDAP
X2-SPN	0.82	0.82	1.78	0.51	**0.38**	**0.68**	**0.54**
X4-SPN	**0.80**	0.91	1.88	0.66	0.42	0.82	0.63
CX2-SPN	0.84	**0.52**	1.41	0.67	0.74	1.00	0.81
SPN	1.00	1.00	1.00	1.00	1.00	1.00	1.00
HET1	1.13	1.14	0.94	1.04	1.01	0.95	1.09
CX2-HOM	1.05	0.68	1.48	0.63	0.76	1.09	1.14
CX2-HET1	0.99	0.65	1.65	0.56	0.73	1.18	1.16
X2-HET1	1.14	1.09	1.97	0.58	0.53	1.06	1.24
HOM	1.13	1.14	1.04	1.00	1.04	1.08	1.25
C-HET1	1.20	0.61	**0.89**	0.97	1.21	1.06	1.28
X2-HOM	1.18	1.10	1.91	0.61	0.55	1.05	1.28
C-HOM	1.21	0.62	1.00	0.89	1.24	1.22	1.48
X4-HOM	1.24	1.30	1.84	0.85	0.61	1.20	1.53
CX4-SPN	0.83	0.60	1.98	0.62	1.01	1.93	1.60
C-SPN	1.13	0.55	1.03	1.05	1.48	1.50	1.69
C-HET2	1.29	0.69	1.36	0.62	1.20	1.60	2.09
X4-HET1	1.11	1.40	2.59	0.68	0.67	1.89	2.16
HET2	1.29	1.28	1.43	0.76	1.12	1.63	2.17
CX4-HOM	1.04	0.81	2.13	0.64	1.13	2.35	2.51
CX4-HET1	1.01	0.79	2.33	0.57	1.08	2.51	2.58
CX2-HET2	1.20	0.80	2.57	0.39	0.84	2.16	2.64
X2-HET2	1.59	1.43	3.31	0.49	0.88	3.05	5.18
X4-HET2	1.54	1.61	3.39	0.60	0.84	3.31	5.36
CX4-HET2	1.17	1.02	4.00	**0.42**	1.43	5.97	7.28

(ii) $\frac{C}{D} = 3$

Table 4.8: Evaluation results for $\frac{C}{D} \in \{0.33, 3\}$ (256PEs). Bold numbers shows the least value per column.

Chapter 5

Conclusions

Interconnecting many elements onto a single die, the interconnection network is heavily under-utilized. In order to tackle these limitations, the existing proposals intervene at the network architectural level, without taking into account at least one of the following criteria:

1. The appropriate design space that P, C and X parameters can introduce.
2. Use of scalable schemes for express physical links.
3. Take care of the router and network resources (ie., bisection bandwidth and buffer space allocation).

Hence, researchers end up with different conclusions in terms of how a network's energy efficiency is affected by the P,C and X architectural parameters.

In this work we carefully generate network architectures derived by applying each of the three architectural parameters (P, C and X), either separately or combinatorially, to a baseline (single 2D mesh) network. Then, we show how buffer space allocation is affected for each case (20 cases for 64 PEs and 24 cases for 256 PEs) when changing the architectural parameters. For each network architecture configuration, we prove that buffer space can only decrease or remain constant. In order to keep the buffer space and bisection bandwidth equal to the baseline, we properly adjust the respective router micro-architecture without degrading performance. We do this by adding more virtual channels per message class or/and increase the buffer depth[1] in cases we have more than one link traversals per router-to-router connection. Drawing on insights from our analysis, we observe that:

It is more energy efficient when a router is bypassed. However, scalability issues arise when extra links are introduced. Extra links means that extra ports per router are needed. Thus, router port count must be constant as PE count increases, otherwise, energy of each crossbar traversal will increase as PE count increases.

[1]Buffer depth increase is needed in order to cover flow control's round trip time (RTT). If RTT round time is not covered, then throughput degrades due to credit return latencies [43]

In this work, we observe that the exclusive use of express physical links of express interval equal to two[2], without concentration or network partitioning, is the best approach in terms of energy-area savings and energy-area efficiency.

When network partitioning is applied, we need more total router ports per network, and thus we pay more control logic for the same memory budget. On the other hand, when concentration is applied, although we have less total router ports per network, and thus, we don't pay more memory logic for the same memory budget, we pay more energy for the crossbar traversals and links. Even though the average crossbar traversal count of packet in a concentrated network is less than a non-concentrated, this happens because a) crossbar energy per cycle increases quadratically with the number of ports and super linearly with datapath width and b) link energy increases due to pipelining and c) utilization decreases as more PEs are injecting traffic per router.

The best zero-load latency for network architectures of tenths of PEs and hundreds of PEs, are the concentrated ones without express physical links and the concentrated ones with express physical links of express interval equal to 2, respectively. In both cases network partitioning does not offer latency improvements. In terms of throughput, regardless of PE count, the exclusive use of concentration seems to be the best. Network partitioning can further improve the throughput only in case we have heterogeneous partitioning and control to data packet ratio is bigger than one. Thus, network partitioning is preferable only in cases that partitioning is combined with concentration, and, control to data packet ratio is bigger than one.

[2]Express physical links are similar to Chen et. al's scheme. Chen et. el. however used express interval equal to 4.

Bibliography

[1] Taiwan Semiconductor Manufacturing Company Limited, *65nm Standar Cell Process*, 2010.

[2] A. Kumary, P. Kunduz, A. Singhx, L.-S. Pehy, and N. Jhay, ``A 4.6 tbits/s 3.6 ghz single-cycle noc router with a novel switch allocator in 65nm cmos,'' in *Computer Design, 2007. ICCD 2007. 25th International Conference on*. IEEE, 2007, pp. 63--70.

[3] J. Howard, S. Dighe, Y. Hoskote, S. Vangal, D. Finan, G. Ruhl, D. Jenkins, H. Wilson, N. Borkar, G. Schrom *et al.*, ``A 48-core ia-32 message-passing processor with dvfs in 45nm cmos,'' in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*. IEEE, 2010, pp. 108--109.

[4] FARADAY Technology Corporation, *UMC's 90nm Standard Cell Process*, 2006.

[5] Virtual Silicon, *UMC's 130nm Standar Cell Process*, 2003.

[6] Virtual Silicon Technology, *UMC's 180nm Standar Cell Process*, 2001.

[7] D. G. Chinnery, ``Low power design automation,'' Ph.D. dissertation, Citeseer, 2006.

[8] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, ``Low-power cmos digital design,'' *IEICE Transactions on Electronics*, vol. 75, no. 4, pp. 371--382, 1992.

[9] Y. Wang, J. Xu, Y. Xu, W. Liu, and H. Yang, ``Power gating aware task scheduling in mpsoc,'' *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 10, pp. 1801--1812, 2011.

[10] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, ``Interconnect-power dissipation in a microprocessor,'' in *Proceedings of the 2004 international workshop on System level interconnect prediction*. ACM, 2004, pp. 7--13.

[11] J. D. Owens, W. J. Dally, R. Ho, D. Jayasimha, S. W. Keckler, and L.-S. Peh, ``Research challenges for on-chip interconnection networks,'' *Micro, IEEE*, vol. 27, no. 5, pp. 96--108, 2007.

[12] W. J. Dally and B. Towles, ``Route packets, not wires: On-chip interconnection networks,'' in *Design Automation Conference, 2001. Proceedings.* IEEE, 2001, pp. 684--689.

[13] P. Gratz, C. Kim, R. McDonald, S. W. Keckler, and D. Burger, ``Implementation and evaluation of on-chip network architectures,'' in *Computer Design, 2006. ICCD 2006. International Conference on.* IEEE, 2006, pp. 477--484.

[14] J. Held, J. Bautista, and S. Koehl, ``From a few cores to many: A tera-scale computing research overview,'' *White Paper, at http://www. intel. com/technology/magazine/rese arch/tera-scale-1006. htm*, 2006.

[15] D. Wentzlaff, P. Griffin, H. Hoffmann, L. Bao, B. Edwards, C. Ramey, M. Mattina, C.-C. Miao, J. F. Brown, and A. Agarwal, ``On-chip interconnection architecture of the tile processor,'' *Micro, IEEE*, vol. 27, no. 5, pp. 15--31, 2007.

[16] S. R. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain *et al.*, ``An 80-tile sub-100-w teraflops processor in 65-nm cmos,'' *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 1, pp. 29--41, 2008.

[17] W. J. Dally and B. P. Towles, *Principles and practices of interconnection networks.* Access Online via Elsevier, 2004.

[18] J. S. Kim, M. B. Taylor, J. Miller, and D. Wentzlaff, ``Energy characterization of a tiled architecture processor with on-chip networks,'' in *Proceedings of the 2003 international symposium on Low power electronics and design.* ACM, 2003, pp. 424--427.

[19] H. Wang, L.-S. Peh, and S. Malik, ``Power-driven design of router microarchitectures in on-chip networks,'' in *Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture.* IEEE Computer Society, 2003, p. 105.

[20] P. T. Wolkotte, G. J. Smit, G. K. Rauwerda, and L. T. Smit, ``An energy-efficient reconfigurable circuit-switched network-on-chip,'' in *Parallel and Distributed Processing Symposium, 2005. Proceedings. 19th IEEE International.* IEEE, 2005, pp. 155a--155a.

[21] N. D. E. Jerger, L.-S. Peh, and M. H. Lipasti, ``Circuit-switched coherence,'' in *Proceedings of the Second ACM/IEEE International Symposium on Networks-on-Chip.* IEEE Computer Society, 2008, pp. 193--202.

[22] K. Goossens, J. Dielissen, and A. Radulescu, ``Æthereal network on chip: concepts, architectures, and implementations,'' *Design & Test of Computers, IEEE*, vol. 22, no. 5, pp. 414--421, 2005.

[23] K. Goossens and A. Hansson, ``The aethereal network on chip after ten years: Goals, evolution, lessons, and future,'' in *Design Automation Conference (DAC), 2010 47th ACM/IEEE*. IEEE, 2010, pp. 306--311.

[24] T. Moscibroda and O. Mutlu, ``A case for bufferless routing in on-chip networks,'' in *ACM SIGARCH Computer Architecture News*, vol. 37, no. 3. ACM, 2009, pp. 196--207.

[25] G. Dimitrakopoulos and K. Galanopoulos, ``Switch allocator for bufferless network-on-chip routers,'' in *Proceedings of the Fifth International Workshop on Interconnection Network Architecture: On-Chip, Multi-Chip*. ACM, 2011, pp. 19--22.

[26] G. Michelogiannakis, J. Balfour, and W. J. Dally, ``Elastic-buffer flow control for on-chip networks,'' in *High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th International Symposium on*. IEEE, 2009, pp. 151--162.

[27] A. B. Kahng, B. Li, L.-S. Peh, and K. Samadi, ``Orion 2.0: A power-area simulator for interconnection networks,'' *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 20, no. 1, pp. 191--196, 2012.

[28] J. Chen, W.-B. Jone, J.-S. Wang, H.-I. Lu, and T.-F. Chen, ``Segmented bus design for low-power systems,'' *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 7, no. 1, pp. 25--29, 1999.

[29] A. B. Kahng, B. Lin, and S. Nath, ``Explicit modeling of control and data for improved noc router estimation,'' in *Proceedings of the 49th Annual Design Automation Conference*. ACM, 2012, pp. 392--397.

[30] G. Passas, M. Katevenis, and D. Pnevmatikatos, ``Crossbar nocs are scalable beyond 100 nodes,'' *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 31, no. 4, pp. 573--585, 2012.

[31] M. R. Stan and W. P. Burleson, ``Bus-invert coding for low-power i/o,'' *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 3, no. 1, pp. 49--58, 1995.

[32] K. Lee, S.-J. Lee, and H.-J. Yoo, ``Silent: serialized low energy transmission coding for on-chip interconnection networks,'' in *Proceedings of the 2004 IEEE/ACM International conference on Computer-aided design*. IEEE Computer Society, 2004, pp. 448--451.

[33] A. Kumar, L.-S. Peh, P. Kundu, and N. K. Jha, ``Express virtual channels: towards the ideal interconnection fabric,'' in *ACM SIGARCH Computer Architecture News*, vol. 35, no. 2. ACM, 2007, pp. 150--161.

[34] C.-H. O. Chen, N. Agarwal, T. Krishna, K.-H. Koo, L.-S. Peh, and K. C. Saraswat, ``Comparison of physical and virtual express topologies for future many-core on-chip networks.''

[35] B. Grot, J. Hestness, S. W. Keckler, and O. Mutlu, ``Express cube topologies for on-chip interconnects,'' in *High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th International Symposium on.* IEEE, 2009, pp. 163--174.

[36] J. Kim, J. Balfour, and W. Dally, ``Flattened butterfly topology for on-chip networks,'' in *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture.* IEEE Computer Society, 2007, pp. 172--182.

[37] J. Balfour and W. J. Dally, ``Design tradeoffs for tiled cmp on-chip networks,'' in *Proceedings of the 20th annual international conference on Supercomputing.* ACM, 2006, pp. 187--198.

[38] S. Lyberis, G. Kalokerinos, M. Lygerakis, V. Papaefstathiou, D. Tsaliagkos, M. Katevenis, D. Pnevmatikatos, and D. Nikolopoulos, ``Formic: Cost-efficient and scalable prototyping of manycore architectures,'' in *Field-Programmable Custom Computing Machines (FCCM), 2012 IEEE 20th Annual International Symposium on.* IEEE, 2012, pp. 61--64.

[39] Y. J. Yoon, N. Concer, M. Petracca, and L. Carloni, ``Virtual channels vs. multiple physical networks: A comparative analysis,'' in *Design Automation Conference (DAC), 2010 47th ACM/IEEE.* IEEE, 2010, pp. 162--165.

[40] S. Volos, C. Seiculescu, B. Grot, N. K. Pour, B. Falsafi, and G. De Micheli, ``Ccnoc: Specializing on-chip interconnects for energy efficiency in cache-coherent servers,'' in *Networks on Chip (NoCS), 2012 Sixth IEEE/ACM International Symposium on.* IEEE, 2012, pp. 67--74.

[41] G. Passas, M. Katevenis, and D. Pnevmatikatos, ``A 128 x 128 x 24gb/s crossbar interconnecting 128 tiles in a single hop and occupying 6% of their area,'' in *Proceedings of the 2010 Fourth ACM/IEEE International Symposium on Networks-on-Chip.* IEEE Computer Society, 2010, pp. 87--95.

[42] L.-S. Peh and N. E. Jerger, ``On-chip networks (synthesis lectures on computer architecture),'' *Morgan and Claypool, San Rafael*, 2009.

[43] L.-S. Peh and W. J. Dally, ``A delay model and speculative architecture for pipelined routers,'' in *High-Performance Computer Architecture, 2001.*

HPCA. The Seventh International Symposium on. IEEE, 2001, pp. 255--266.

[44] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi, ``Cacti 6.0: A tool to model large caches,'' *HP Laboratories*, 2009.

[45] M. Snir, *Mpi the Complete Reference: The Mpi Core.* MIT Press, 1998, vol. 1.

[46] E. H. Jensen, G. W. Hagensen, and J. M. Broughton, ``A new approach to exclusive data access in shared memory multiprocessors,'' Technical Report UCRL-97663, Lawrence Livermore National Laboratory, Tech. Rep., 1987.

[47] D. E. Culler, J. P. Singh, and A. Gupta, *Parallel computer architecture: a hardware/software approach.* Gulf Professional Publishing, 1999.

[48] M. G. Katevenis, V. Papaefstathiou, S. Kavadias, D. Pnevmatikatos, D. S. Nikolopoulos, and F. Silla, ``Explicit communication and synchronization in sarc,'' *Micro, IEEE*, vol. 30, no. 5, pp. 30--41, 2010.

[49] N. Agarwal, T. Krishna, L.-S. Peh, and N. K. Jha, ``Garnet: A detailed on-chip network model inside a full-system simulator,'' in *Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on.* IEEE, 2009, pp. 33--42.

[50] N. Jiang, J. Balfour, D. U. Becker, B. Towles, W. J. Dally, G. Michelogiannakis, and J. Kim, ``A detailed and flexible cycle-accurate network-on-chip simulator,'' in *Performance Analysis of Systems and Software (ISPASS), 2013 IEEE International Symposium on.* IEEE, 2013, pp. 86--96.

[51] A. B. Kahng, B. Lin, and K. Samadi, ``Improved on-chip router analytical power and area modeling,'' in *Proceedings of the 2010 Asia and South Pacific Design Automation Conference.* IEEE Press, 2010, pp. 241--246.

[52] H.-S. Wang, X. Zhu, L.-S. Peh, and S. Malik, ``Orion: a power-performance simulator for interconnection networks,'' in *Microarchitecture, 2002.(MICRO-35). Proceedings. 35th Annual IEEE/ACM International Symposium on.* IEEE, 2002, pp. 294--305.

[53] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, ``Mcpat: an integrated power, area, and timing modeling framework for multicore and manycore architectures,'' in *Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on.* IEEE, 2009, pp. 469--480.