Design and Implementation of 3D Network on Chip Architectur

Sangeetha. T[1],Dr. Latha. R[2], Sathish. B [3], Dhineshvalavan. S[4], Sakthivelan. R[5] , Selvendhran. A[6]

[1]Asst. Prof , Department of ECE, K.S.K college of engineering and Technology,
Darasuram, , Thanjavur, Tamil Nadu, India.

[2]Prof, Department of ECE, K.S.K college of engineering and Technology ,
Darasuram, Thanjavur, Tamil Nadu, India

[3, 4,5] student, Department of ECE, K.S.K college of engineering and Technology,
Darasuram, Thanjavur, Tamil Nadu

***Abstract*- Design of 3D network on chip is evaluated here. Conventional NoCs are designed predominantly for unicast data exchanges and the multicast traffic is generally handled by converting each multicast message to multiple unicast transmissions .Hence, applications dominated by multicast traffic experience high queuing latencies and significant performance penalties when running on systems designed with unicast based NoC architectures. In the existing design of a wireless NoC (WiNoC) architecture incorporating necessary multicast support .By integrating congestion-aware multicast routing with network coding, the WiNoC is able to efficiently handle heavy multicast injections. For applications running with a broadcast heavy Hammer cache coherence protocol, the proposed multicast aware WiNoC achieves an average of 47% reduction in messagelatency compared with the XY-tree-based multicast-aware mesh NoC .This network level improvement translates into a 26% saving in full-system energy delay product. In the proposed system, an efficient 3D network on chip is evaluated in which the internal SOC platform is splitted up in a 3D Fashion. The Node to Node communication can happen even at nodes present in different layers too .Node to node communication delay and distance will be reduced here.**

INTRODUCTION

ThisthesisaddressestheanalysisanddesignofalgorithmsandprotocolsforNetwork on Chip (NoC) interconnection systems. We propose to investigate the following research issues: Architecture analysis: we intend to evaluate the main NoC architectures proposed in Literature such as 2D Matrix, Ring, Crossbar and the novel SpidergonNoC to understand their main characteristic and the cases where they constitute the best choice. protocol design: investigation and proposal of novel transport and routing algorithms for the SpidergonNetwork on Chip.Networkdesign: investigate the major issues in the actual NoC implementation and interconnection. The analysis and the characterization of protocols and architectures for NoC systems that we propose throughout this Thesis have been obtained through computer-based simulation. According to the International Technology Roadmap for Semiconductors projections, by the end of this decade complex systems, called Multi Processor System-on- Chip (MP SoC), will contain billions of transistors running at a frequency of many GHz.As depicted in Figure1.1 the technology of semiconductors keeps on scalingdown allowing more and more components to be installed within the same area of a chip. As a consequence complex systems that once required many microchip for being

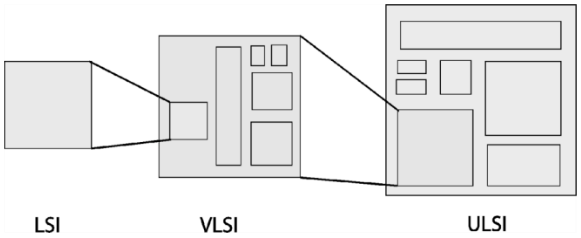


Figure 1.1: Improvement in the semiconductor technology leads a scale down of the components on achip.

Nowcanbeinstalledonasinglemicrochipcontainingallthelogicofthesystemandtheinterconnectionchannelsconnectingthem.

Examples of the secapabilities are there centeightcores IBM's Cell processor installed on the Sony's Play station III and them ore futuristic eighty cores Intel's Tera flop processors.

Acentral and key element in future complex MP So Cis the global On-Chip Communication Architecture (OCCA) or On Chip Interconnect(OCIN) Theinfrastructure that interconnects the components of a MPSoC and provides themeans necessary for distributed computation among different processingelements.

The natural evolution of the bus-based solution reported in Figure1.2(a) and the poorly scalablepointtopointnetworksseeninFigure1.2(b)arethenewgenerationarchitecturescalledNetworkon-Chip(NoC)represented.
Examples of innovative NoC architectures include the Lip6 Spin, the

Raw, the Vtt(and various Universities) Eclipse and Nos- trum, Philips's Ethereal NoC , Stanford/Bologna Universities' Netchi and ST Microelectronics Spiderg on No C.No Care packet-switched communication networks derived from the parallel computing domain.

They are based on a well-defined protocol stack similar to the ISO/OSI seen in the network on computers. A layered-stack approach to the design of the on-chip inter-core communications can be defined accordingly with the communication-based methodology that will

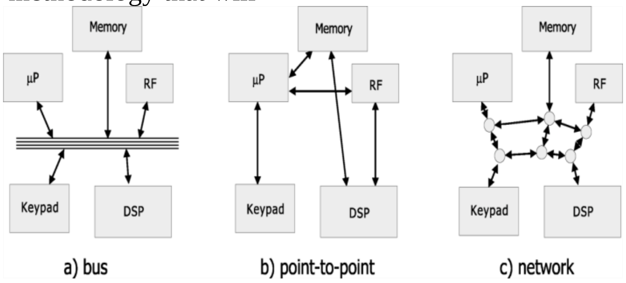


Figure 1.2: Examples of communication structures in Systems-on-Chip

Be conceived for the system. Exploiting the lesson learned by telecommunication community, the global on-chip communication is decomposed into layers similar to the ISO-OSI Reference Model (see Figure 1.3).

The protocol stack enables different services, providing to the programmer an abstraction of the communication frame- work. Layers interact through well-defined interfaces and they hide the low level details.

1. Chip Multi Processor

(CMP) are networks designed to support any kind of traffic. At the network's design time no knowledge about the traffic is available. CMP hence are built to offer best effort services. Quality of ser- vices (QoSs) capabilities can be granted by differentiating the traffic into classes of priorities each on eassigned to a specific virtual channel (VC) or by providing guaranteed service level son dedicated connection.

CMP systems are mainly composed by similarcomponents grouped such as sets of processors and memories. This allows the use of regular topologies often borrowed from the parallel computing world such as 2D Mesh, Torus, Ring etc. . .;

2. SystemOnChip

(SoC)aresystemsintegratingheterogeneouscomponents, often developed by third party companies at the purpose of building an appli- cation specific systems. In SoC often the traffic patterns are known since the design time hence the interconnection networks can be build to exactlymatch.

The application requirement. In SoC for embedded applications designers use standard industrial CAD-tool flows for the synthesis of a platform-specific NoC and must cope with an increasing number of timing-closure exceptions due the differences in size across its heterogeneous processing cores. This problem becomes particularly hard when using nanometer technology processes as the impact of global inter- connect wires raises exponentially the number of wire exceptions, i.e. timing-closure violations due to the delay of a global wire exceeding the target clock period (the clock at which the system is deigned to run)T_{clk}.The research results discussed in this Thesis will be mainly focused on System on Chip architectures.

PROPOSED SYSTEM

In the proposed system, an efficient 3D network on chip is evaluated in which the internal SOC platform is splitted up in a 3D Fashion. The Node to Node communication can happen even at nodes present in different layers too. Node to node communication delay and distance will be reduced here.

EXISTING SYSTEM

In the existing design of a wireless NoC (WiNoC) architecture incorporating necessary multicast support.By integrating congestion-aware multicast routing with network coding, the WiNoC is able to efficiently handle heavy multicast injections. For applications running with a broadcast heavy Hammer cache coherence protocol, the proposed multicast aware WiNoC achieves an average of 47% reduction in message latency compared with the XY-tree-based multicast-aware mesh NoC. This network level improvement translates into a 26% saving in full-system energy delay product.

3.1 Existing System Overview

3.1.1 Xy-Tree Multicast Mechanism for Mesh Nocs

The multicast message is first forwarded from the source node to all the intermediate nodes lying in the same row.

The message is then replicated at these intermediate nodes and a copy of the original message is forwarded to all the destinations.

3.1.2 Cache-Coherence-Induced Traffic Patterns

Each multicast message is associated with a set of acknowledgement (ACK) messages that are transmitted from each multicast destination back to the source node.

3.1.3 Mesh Noc

The network latency is usually high due to the inherent multihop nature of the system. High network latencies cause undesired delays in forwarding the multicast messages as well as in collecting the ACKs, leading to stalled processor cycles

PROTOCOLSDESIGN

6.1 Routing

Routing packets along an interconnection network is a well known problem of the parallel computing systems. Literature offers many different algorithms suited for almost any kind of architecture proposed so far .

Routing algorithms can be subdivided into source or distributed routing func- tions. In the first case a source node computes the whole path of a packet through the NoC. In the latter the source node just forwards the packet to the connected router that will compute the first step and leaving to the following routers theburden of computing the followingpath. A routing function can be either minimal or not. In the first case each hop performed by a packet takes it always closer to its final destination. In the case.

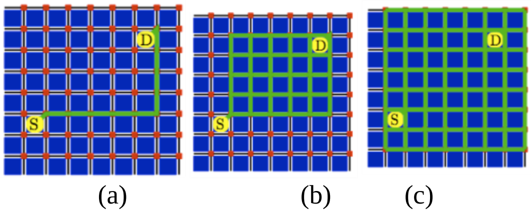


Figure 6.1: Example of (a) minimal deterministic, (b) minimal adaptive, (c) non minimal adaptive routing.

Finally routing algorithms can be classified in deterministic, adaptive and oblivious . In the NoC domain deterministic algorithms are usually preferred because they are simple to implement and their behavior is easily predictable: given a source and a destination node, a deterministic routing algorithm computes through a look-up table or a mathematical equation the exact path towards the packets destination node.

An example of minimal deterministic routing algorithm: given a source S and a destination D the routing function returns only one single path connecting them.the result of a minimal adaptive routing algorithm.

Adaptive routing algorithms compute the path of a given message considering the source and destination nodes address and also some information relative to the status of the network. Because of their nature these kind of algorithms can easily fall in a deadlock state but can greatly improve the performance of the system as they manage to evenly distribute the traffic along all the channels interconnecting two peers.

These algorithms are difficult to predict and the greater degree of freedom can easily degenerate in deadlocks. Non minimal algorithms can be used in complex NoC systems featuring a certain degree off aulttolerance.

Network on Chip are a special case of parallel computing systems characterized by the tight constraints such as resource availability, area and power consumption and cost of the NoC architecture. Many of the currently adopted architectures and protocols derive directly from the distributed computing research area from which NoC are a special case. Never- theless new and NoC-specific solutions are currently being published.

EXPERIMENTAL RESULTS

A. Experimental Setup

In this paper, we evaluate the performance of the proposed WiNoC architecture against three other multicast-aware NoC architectures, XY-tree mesh NoC, path multicast mesh NoC, and the wireless mesh (WiMesh). These three architectures are shown in Fig. 14. We have already explained the salient features of the XY-tree mesh NoC and the path multicast meshin Section II. The WiMesh NoC is constructed b placing WIson the traditional wireline mesh architecture. In mesh wireline networks, it is not possible to create the two sets of orthogonallinks as in our WiNoC, and hence, it is not possible to employ the congestion-aware MALASH routing described earlier. Thisis because to create two sets of orthogonal links, at least

$2(N-1)$ wireline links are needed in a region with N nodes. However, in the mesh NoC, only $2(N-\sqrt{N})$ intra-regionwireline links are available. Hence, the wireline distributions in WiMesh follow simple XY-tree routing [9], [21].We use the same five applications (CNL, FLD, FFT, LU,and RAD) discussed in Section III for the following analyses.

We consider a 64-core ALPHA core system operating with afrequency of 2.5 GHz and having a die size of 20×20 mm2.The memory system is composed of private 64-kbyteL1 instruction and data caches and one shared 16-MbyteL2 cache (256-kbyte distributed L2 per core). All the processingcores and memory system are interconnected using anNoC fabric with 64 routers. For all the NoC architecturesconsidered, we employ a generic three-stage router architecturemodifying [38]. This architecture has three functional

stages, namely, input arbitration, routing/switch traversal, andoutput arbitration with link traversal. Following routing, amulticast flit is replicated to all necessary output VirtualChannel buffers (VCs) simultaneously. The flits at output VCsare then handled by the output arbiter.

The additional cyclesrequired for congestion-aware routing and MAC protocol

processing are accounted for while determining the networklatency. Energy dissipation of the network routers, includingthe routing, MAC, and NC blocks, was obtained from thesynthesized netlist using a 28-nm commercial Fully DepletedSilicon On Insulator technology by running Synopsys PrimePower.

The total area overhead required in a WiNoC routerto incorporate the WI, routing logic, NC, and MAC protocolunits is 0.2514 mm2 (4.02% area overhead for an NoC tilesized 2.5 mm× 2.5 mm). The router ports are provided witha buffer depth of two flits. The width of all wireline links isthe same as the considered flit width (32 bits).1 Each wirelineis designed with the optimum number of uniformly placedand sized repeaters in the 28-nm technology node. The energydissipation of the wireline links was obtained through CadenceSPECTRE simulations.We use GEM5, a full-system simulator, to obtain processorand network-level information [39]. We keep track of eachmessage injected through the network interface module ofGEM5 to extract the multicast traffic traces associated withreal applications [40].

We use a modified GARNET interconnectionnetwork along with GEM5 to model the multicast supported NoC s in full-system simulations performed to obtainexecution times. The processor-level statistics generated bythe GEM5 simulations are incorporated into multicore power,area, and timing to determine processor power values [41].

B. Efficiency of the Wireless Links

In this section, we demonstrate that the mm-wave wirelesslinks are more efficient than the traditional metal wires to1We have used 32-bit flits since Directory protocol only requires smallerflit widths [43]. However, when an NoC is specifically designed for densebroadcasts (such as Hammer protocol broadcasts), employing higher flitwidths [22], [42] can enhance the bandwidths at the network edges.Fig. 15. Per bit link EDPs for varying communication lengths.

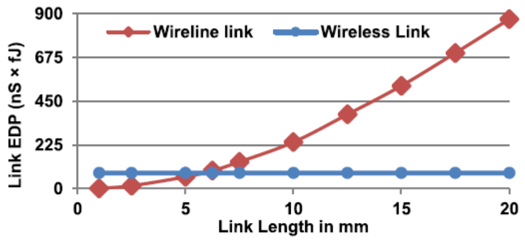
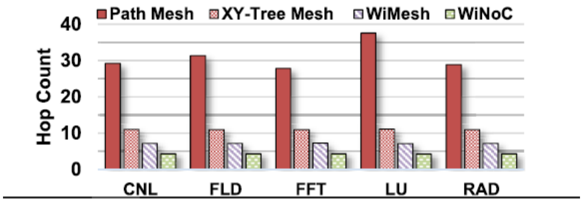


Fig. 15. Per bit link EDPs for varying communication lengths.



CONCLUSION

With an ever-expanding application pool, today's manycorearchitectures are expected to handle a high diversityof on-chip traffic patterns. The Hammer cache coherenceprotocol is one such system-on-chip application that stresseson chip networks with heavy multicast injections. In thispaper, we presented a multicast-aware WiNoC architecturethat can efficiently handle multicast-heavy cache coherencecommunications. Incorporated with a congestion-aware multicast routing and NC.

WiNoC eliminates the initial andintermediate queuing latencies seen in conventional wirelinemesh NoCs. Moreover, using wireless shortcuts, the WiNoCachieves significant reductions in network latencies leadingto improved system performances. Compared with a manycoresystem using the multicast-aware XY-tree mesh NoC, the WiNoC incorporated many-core platform achieves an averageof 26% full system EDP improvement in the presence ofHammer protocol.

REFERENCES

[1] A. Karkar, N. Dahir, R. Al-Dujaily, K. Tong, T. Mak, and A. Yakovlev, “Hybrid wire-surface wave architecture for one-to-many communication in networks-on-chip,” in Proc. IEEE Design Autom. Test Eur. Conf.Exhibit. (DATE), Mar. 2014, pp. 1–4.

[2] D. Vainbrand and R. Ginosar, “Network-on-chip architectures for neuralnetworks,” in Proc. 4thACM/IEEE Int. Symp. Netw.-Chip (NOCS),May 2010, pp. 135–144.

[3] J.-Y. Kim, J. Park, S. Lee, M. Kim, J. Oh, and H.J. Yoo, “A 118.4 GB/smulti-casting network-on-chip with hierarchical star-ring combinedtopology for real-time object recognition,” IEEE J. Solid-State Circuits,vol. 45, no. 7, pp. 1399–1409, Jul. 2010.

[4] Y. Xue and P. Bogdan, “User cooperation network coding approachfor NoC performance improvement,” in Proc. 9th ACM Int. Symp.Netw.-Chips, 2015, Art. no. 17.

[5] Neuromorphic Computing: From Materials to Systems Architecture,Report of a Roundtable Convened to Consider Neuromorphic ComputingBasic Research Needs, U.S. Dept. Energy, Gaithersburg, MD, USA,Oct. 2015.

[6] N. E. Jerger, L.-S. Peh, and M. Lipasti, “Virtual circuit tree multicasting:A case for on-chip hardware multicast support,” in Proc. 35th Int. Symp.Comput. Archit. (ISCA), Jun. 2008, pp. 229–240.

[7] A. Ros, M. E. Acacio, and J. M. García, “Dealing with traffic-area tradeoffin direct coherence protocols for many-core CMPs,” in *Proc. Int.Workshop Adv. Parallel Process. Technol.*. Berlin, Germany: Springer,Aug. 2009, pp. 11–27.

[8] P. Conway and B. Hughes, “The AMD opteron northbridge architecture,”*IEEE Micro*, vol. 27, no. 2, pp. 10–21, Mar. 2007.

[9] M. Lodde, J. Flich, and M. E. Acacio, “Heterogeneous NoC design forefficient broadcast-based coherence protocol support,” in *Proc. NOCS*, May 2012, pp. 59–66.

[10] S. Deb, A. Ganguly, P. P. Pande, B. Belzer, and D. Heo, “WirelessNoC as interconnection backbone for multicore chips: Promises andchallenges,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 2