Energy Efficient NoC design through Supervised Machine Learning

D. Shanthi¹, Ashwani Kumar²

¹,² Associate Professor, Sreyas Institute of Engineering and Technology, Hyderabad
dshanthi01@gmail.com, ashwani.kumarcse@gmail.com

Abstract. 3D NetworkOnChip (NoC) will provide high multi-core data processing with the minimal amount of energy consumption. It is necessary to consider the architectural limits of three-dimensional NoCs. A deep reinforced framework is proposed that uses router-less NoC for evaluation in the case study. The framework uses prior approaches that may be unreliable because of the difficulties with the design and search process, and the inflexibility that results because of space restrictions. The framework has better loop placements for networking chips with various design constraints. A Monte Carlo search tree is used to develop a Deep Neural Network that searches in parallel for a NoC design.

1. Introduction
The 3D chips are capable of providing good performance, packaging density and function at low costs, which are characteristics of 3D chips. Nocc's are integrated circuits which enable the integration of a number of cores on a chip. The building shapes comply the benefits of paradigms for unprecedented performance. While the architectures of NoC were previously prohibitive because of the wiring in planar IC’s, these were then out excelled by 3D variants of this methodology. This 2D NoC usually follows straight forward extensions of 2D NoC architectures, which does not generates all of the benefits from NoC 3D architecture technology. We've selected a topology where connections are such as vertical and horizontal. This is often difficult to get optimum performance. To maximize this objective, functions like machine learning and communication cost are used to explore a 3D Sw element architecture of NoC. Thus it reduces both latency and energy consumption. New Architecture for Cloud Computing performs the best among the other NoC platforms in terms of the state of the art of NoC architectures.

Many conventional three-dimensional meshes[1] have been used in NoC applications. One of the disadvantages of meshed based architectures is the delay and more energy consumption. The Nocc-bus hybrid architecture is a way to meet the needs of the lower classes of the 3D Interconnect, which uses Dynamic Time Division Multiple Access (DTDMA) for lower latency. DimDe is used in keeping energy consumption to a minimum. 3D NoC is improved by decreasing the number of port of input. The buses of all the types are of Z-dimension. As the number of nodes increases the latency of the system increase since the high traffic injection increases will increase. The recent developments like 3D printing bear the manufacturing defects and wearing out needs [9]. This resulted in development of vertical [9]. Both NoC architecture with vertical links and fault tolerant router are proposed to avoid performance loss in case of load shedding or other disasters. In addition, the results in greater capital expense, Application-specific NoC architectures are investigated in [12]. The network spanners used here uses ILP based algorithm in order to develop long link deployment network. The whole project aims to stabilize energy consumption in Singapore. This paper offers novel deep reinforcement
learning framework for design space exploration, and presents a research implementation example for routerless NoC. The Monte-Carlo Tree Search (MCTS) algorithm produces some data to a neural network, which, in turn, guides the search. Loop positions within constraints of designed design rules are learned automatically by the framework.

2. NoC Architecture
A ring of nodes manages one node. ring-directed injection Single-ring NoC follows: Single-ring designs are simple to build, but they restrict the amount of data they can process. The network network overload occurs quickly when nodes are added. Single-ring designs scale to a relatively small number of processors[2]. Router-based NoCs have the potential to change the future. Network routers monitor resource availability before packets are transferred between nodes via these routers[2]. Most network architectures use mesh architecture.

Table 1: Difference between Bus-based chips and NoC

| Factor           | Bus                  | NoC                  |
|------------------|----------------------|----------------------|
| High Frequency   | 250MHz               | >750MHz              |
| A ring of nodes  | directed injection   | ring NoC             |
| manages one      | Single               | follows: Single      |
| node. ring       |                      |                      |
| Low Latency of   | 6                    | 6                    |
| Cluster          | Cycles@250MHz        | Cycles@250MHz        |
| Low Delay of     | 14-18 cycles         | 12 cycles            |
| inter Latency    | @250MHZ              | @250MHZ              |
| Dynamic Power    | High                 | Low                  |
| Static Power     | High                 | Low                  |
| Gate Count       | 400k                 | 210k                 |

Figure 1(b) illustrates a basic 2D grid of nodes, with a router at every node. This solution incurs 11% of the total chip area [12] and can see a maximum of 28% of total chip power overhead [5][7][15][19][30].

3. Optimization of 3D-NoC
The design problem was first described, and then the high level overview of machine learning proposed in this paper was presented. Then key details of NoC optimization are provided. A computer that needs to communicate in a high-performance, low-power way needs to use low-power hardware and software to send and receive data. For this scenario, NoC networks which are based on CSP designs are good. SW calculations are done with values inserted to a regular grid. In order to achieve better performance, power, and other resources, using power law SW connectivity is better than regular multi hop mesh networks. NoC structures are similar to those found in our home. A NoC 3D configuration must support the complete design which is a long range shortcut network. There should be proper placement of long range links that are along the vertical dimension. Thus, the objective is to design the best interconnection architecture that accomplishes the link selection criteria by the complete interconnection architecture.

The probability of a direct link is a function of the number of links, and as such changes exponentially with respect to link length. The connectivity coefficient ($\alpha$) decides the connectivity nature, i.e., for examples consider a larger $\alpha$ means which is locally connected with few or no long-range links. The number of connections per unit area between the neurons of the network is assumed to be independent of the distance between the neurons, even if their actual distance is very large. An intermediate network is in between two totally connected networks. There is a topological randomness. SW graphs contain average paths that are the total number of hops between all the nodes. The total number of nodes in the graph is used in the SW graph. This graph feature can be used to communicate with very little resources.

![Figure 2. High-Level view of optimization algorithm.](image)

If communication costs are optimized, then the cost per hop is lowered, as well as the consumption of energy. Also, the latency between nodes is lowered. The space of SW graph which involves a combinatorial design "D" based on clusters is a complex space. The goal is to find the minimal X that would minimize O. Hill climbing and simulated annealing can be used to select the community. We use machine learning which significantly improves the efficiency of the many algorithms for exploring design space. The optimization process is undertaken before implementation of the NoC.

### 4. Optimization using machine learning

We employ a recent methodology called STAGE [1]. Online random-restart combinatorial optimization algorithms have been developed to significantly improve performance via local search (for e.g., hill climbing). STAGE aims to offer additional features $\sigma(d) \in R$ where m is the total number of features of the optimization problem to find an improvised function of evaluation, known as a "starting point" in local search process. A. E selects states that help A make progress. locally optimal sites appear to be convex and globally optimal in the middle (2). Compared to other popular algorithms, Stage learns the structure and design of the space. In this method, design space is extended through the use of available space. This is the first time a NoC application is using STAGE.
4.1 Instantiation of complex no-cubes dynamics:
Here, all the information pertaining to the STAGE 3D NoC strategies is outlined.

4.2 The Design Sphere:
The design space of network depends mainly on the resources of a particular network, which are used as the input to an optimization algorithm. A communication cost arising from the use of a 3D NoC that is the product of frequency of communication, hop count, link length.

4.3 Network Constraints:
Enforcement of constraints on to the placement of router configuration and vertical links is performed to explore feasible 3dimensional NoC designs.

5. Instructional Leadership
Reinforcement learning is one of several forms of machine learning where researchers explore multiple actions in an environment in order to maximize cumulative rewards. A central issue in this paper is the nature of the environment where software agents act. The environment is represented by a single-hop routed NoC design [11].

Learning approximations to optimal looping mechanisms is difficult due to a high need for training data. In AlphaGo[3][6], the policy function selected one move at each of several hundred steps, but required over 30 million data samples. To address the challenge mentioned earlier, you must utilize data efficiency and threading. State must include all information for the template to determine optimal loop placement, and must be compatible with Angular’s input/output structure. The hopping needs to be known in order to average the number of hops. Information quality also influences learning efficiency because incorrect information necessitates inferential work. The decision tree structure is also fixed throughout the design process. Returns and their impact on NoC performance must be disclosed to all involved.

Figure 3. Multi-agent system for NoC design.

Figure 4. Proposed deep learning framework for router less NoC designs reinforcement.
The key metrics of embedded machine learning are throughput/latency, accuracy, and cost and energy consumption. Machine learning algorithm accuracy is calculated for a larger dataset. There are large datasets in big numbers which are publicly available for anyone like Image Net. It is important to update the weights in applications or environment using programmability. In deep neural networks, the processor should support various networks which have variable number of layers, channels, filter sizes and filters.
6. Conclusions

By combining machine learning and SW networks, a design optimization method is proposed that Improvises the energy efficiency to explore design space. It is shown that the 3D SW graph of NoC architecture performs better than the regular 3D NoC. The optimized 3D NoC which has SW architecture gets a 35% EDP reduction compared to 3D mesh. The proposed methodology performs better than fully connected 3D mesh in terms of robustness and efficiency with a minimal number of vertical links. In the case of reduction of vertical links up
to 50%, we achieved a 25% EDP reduction compared to 3D mesh. Fixed connectivity of 1.3% is converged to 9 epochs with the help of network training and also with the accuracy of 96.6% in classification, for the test set given in fig 7. There is 1.6% below the performance of the baseline of 98.2% which is obtained by a fully connected network which could not be trained on to a chip. To compare another network on chip is trained in which the connections of the network are initialized from the similar or same distribution which is identified by DEEP R bud was not re-wired the connections in the period of training. The classification accuracy of the network reached only to 81.3%. Thus, the comparison projects that the rewiring during on-line is very important for a good performance of a network. On 4 core setup (4PA case) 1.3% of connectivity uses very less memory, but it fully utilizing the memory in the available four processors. This way, it could be up to 9.3% of connectivity and also achieves 97.7% of accuracy on the assigned test set.

References

[1] V. Pavlidis and E. Friedman, Three-Dimensional Integrated Circuit Design, 2009.
[2] B. S. Feero, P.P. Pande “Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation”, IEEE Trans. on Computers, Vol 53, pp. 32-45, Aug. 2008.
[3] P. Jacob et al., “Predicting the Performance of a 3D Processor-Memory Stack,” IEEE Design and Test of Computers, vol. 22, no. 6, pp. 540-547, Nov./Dec. 2005.
[4] H.G. Lee et al., “On-Chip Communication Architecture Exploration: A Quantitative Evaluation of Point-to-Point, Bus, and Network-on-Chip Approaches,” ACM Trans. Design Auto. of Electronic Sys., vol. 12, no. 3, pp. 1-20, Aug. 2007.
[5] F. Li et al., “Design and Management of 3D Chip Multi-processors Using Network-in-Memory,” Proc. of ISCA, 2006 pp. 130-141, doi.10.1109/ISCA.2006.18.
[6] J. Kim et al., “A novel dimensionally-decomposed router for on-chip communication in 3D architectures,” in Proc. of Annual Intl. Symposium on Comp. Arch., June 2007, pp. 138–149.
[7] A. M. Rahmani, K. R. Vaddina, K. Latif, L. P. Plosila, H. Tenhunen, “HighPerformance and Fault-Tolerant 3D NoC-Bus Hybrid Architecture Using ARB-NET Based Adaptive Monitoring Platform”, IEEE Transaction on Computers, vol 63, pp. 734-747, Mar 2014.
[8] L. Jiang, F. Ye, Q. Xu, K. Chakrabarty, B. Eklow, “On effective and efficient in-field TSV repair for stacked 3D ICs”, Proc. of DAC. 2013.
[9] Kumar, A., S.P. Ghrera, and V. Tyagi, Implementation of wavelet based modified buyer-seller watermarking protocol (BSWP), 2014.
[10] M. Bahmani, A. Sheibanyrad, F. Petrot, F. Dubois, P. Durante, “A 3D-NoC Router Implementation Exploiting Vertically-Partially-Connected Topologies,” Annual Symp. on VLSI, Aug. 2012, vol., no., pp.9-14.
[11] I. Loi, F. Angiolini, S. Mitra, L. Benini, “Characterization and Implementation of Fault-Tolerant Vertical Links for 3-D Networks-on Chip”, IEEE Trans. Computer Aided Design of Integrated Circuit and Sys., vol 30, pp. 124-134, Jan. 2011.
[12] D. Shanthi, R. K. Mohanty, Dr. G. Narsimha, “Application of Intelligent Techniques to predict software reliability”, National Conference on Information Systems and Knowledge Sharing vol.18, pp.318-321, 2013.

[13] Kumar, A., S. P. Ghrera, and V. Tyagi, An ID-based Secure and Flexible Buyer-seller Watermarking Protocol for Copyright Protection. 2017.

[14] D. Shanthi, Dr. R. K. Mohanty, Dr. G. Narsimha “Application of Machine Learning Techniques for Statistical Analysis of Software Reliability Data Sets,” Journal of Computers and Electrical Engineering Special Section on Soft Computing Approaches for Sustainable Systems”, 2018.

[15] D. Shanthi, Dr. R. K. Mohanty, Dr. G. Narsimha,” APPLICATION OF PARTICAL SWARM INTELLIGENCE,”, International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE Proceedings, 2017.

[16] Kumar, A. and S. Srivastava, Object detection system based on convolution neural networks using single shot multi-box detector. Procedia Computer Science, 2020. 171: p. 2610-2617.

[17] Kumar, A., Z. J. Zhang, and H. Lyu, Object detection in real time based on improved single shot multi-box detector algorithm. EURASIP Journal on Wireless Communications and Networking, 2020. 2020(1): p. 1-18.

[18] Fatima, S. A., et al. Object recognition and detection in remote sensing images: a comparative study. in 2020 International Conference on Artificial Intelligence and Signal Processing (AISP). 2020. IEEE.

[19] Shailendra Mishra and Mayank Singh, “An improved Energy Efficient Communication Protocol (IEECP) for Wireless Sensor Networks,” International Journal of Engineering and Advanced Technology (IJEAT), vol. 8, no. 4c, April 2019.