A Scalable Multi-Dimensional NoC Simulation Model for Diverse Spatio-temporal Traffic Patterns

Awet Yemane Weldezion[*], Matt Grange[†], Dinesh Pamunuwa[‡], Axel Jantsch[*] and Hannu Tenhunen[*]

[*]KTH Royal Institute of Technology, School of ICT, Electronic Systems, Forum 120, Kista SE 16440, Sweden
Email: {aywe, axel, hannu}@kth.se
[†]Mentor Graphics, Calibre Division, Wilsonville, Oregon, USA
Email: matthew_grange@mentor.com
[‡]University of Bristol, Dept. Electrical & Electronic Eng., Woodland Road, Bristol, BS8 1UB, UK
Email: dinesh.pamunuwa@bristol.ac.uk

Abstract—This paper describes a powerful simulation platform that enables accurate simulations of numerous network configurations under realistic traffic patterns to predict the performance and power needs of a 3-D integrated system early in the design flow. The simulation platform can model virtually any sized 2-D or 3-D network configuration, providing low-cost and fast trade-off evaluations of various systems architectures. The network simulator uses scalable RTL-level models that can be used for accurate power and timing analyses. We demonstrate the capability of our simulation model by analyzing the performance of various network topologies under spatio-temporal traffic patterns to show how the network topology can be adjusted to meet the performance requirements of a design before it is manufactured. The simulation results can be used to optimize the placement of cores and communication buses early in the flow. By using the model, standard applications such as mobile application processor, femto-cell base-stations on-chip and wide-IO TSV memory stacking can be simulated.

I. INTRODUCTION

Networks-on-Chip (NoC) are a well-researched solution for the communication infrastructure in many-core and heterogeneous designs. The on-chip network infrastructure is used as a backbone, where all IP cores are physically linked with a regular bus structure to on-chip and off-chip modules. The emergence of complex 2-D, 2.5-D, and 3-D SoC ICs pushes the boundaries of such conventional communication technology. By considering recent scalability trends, in the next ten years many-core processors will have 1000s of cores in a single networked chip. As the number of cores increases, and the level of dissimilar technologies rises, designs must use more advanced and scalable bus systems. Nevertheless, such progress is not without limitations and challenges.

Firstly, the design complexity increases with the degree of integration of the dissimilar technologies. This is also worsened by the lack of integrated design tools and standards. Secondly, the high-cost of 3-D chip prototyping makes design implementation limited as far as running simulators and analyzing simulation results. Finally, even with current 2-D mesh networks with hundreds of cores, extracting results by running software and RTL simulators by itself is tedious and time-consuming process. The problem worsens when the number of variables to be extracted increases, and when the complexity

level of integration increases, for example from 2-D mesh to 3-D cube network. Simulation environments reported in [1]- [4] proposed NoC simulation models for specific applications or configurations. However, the design approaches reported are not comprehensive and scalable to accommodate the emerging integration of dissimilar technologies and heterogenous systems.
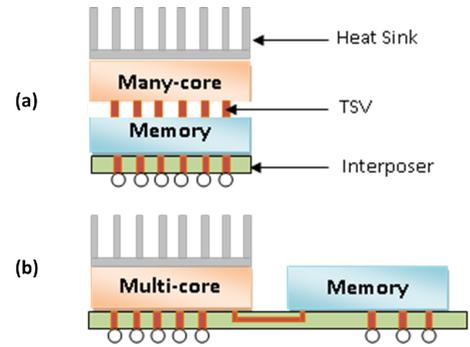

Fig. 1. Processor-memory configurations with (a) 3D-IC and (b) 2.5D-IC with interposer technology

The conceptual 2.5D-IC and 3D-IC chip models shown in Figure 1 are relevant examples of vertical integration of dissimilar technologies. With existing process technology, it is possible to stack standard CMOS technology, distributed DRAM layers, and NAND Flash memory technologies together in a single chip and increase system performance in terms of memory access time and bandwidth [5] [6]. The key driver to scalable vertical integration is the through-silicon-via (TSV), which has emerged as state-of-the-art technology [7]. However, prototyping such systems requires a multi-disciplinary design team with new design tools that significantly increase research cost and overall time-to-market. Thus, heavy use of simulation based analysis is imperative.

This paper presents a scalable multi-dimensional NoC based simulation model for fast and seamless integration of dissimilar technologies and heterogeneous systems. On-chip resources share and use a NoC as a backbone to communicate with each other as well as with off-chip resources. The model exploits the characteristics of horizontal and vertical interconnects and

addresses such issues through the development of various standardized interfacing circuit blocks, communication protocols, and architectural designs while meeting the requirements of high-performance at low-cost and low-power consumption. By using the simulation model, chip design issues such as system performance, power consumption, form-factor, and high-level issues such as programmability, operating systems, and memory consistency can be significantly addressed early in the design exploration phase. By combining the approach of system-on-a-chip design with the new features of 3-D integration technology, the simulation model can serve as a low cost integration and evaluation medium for dissimilar technologies.

II. SIMULATION ENGINE

For any topology, the placement of IP cores varies depending on design requirements. Heterogeneous IP-Cores run applications of all sorts. Some applications require real-time responses, while others require more bandwidth and memory space. An interesting aspect in a 3-D architecture is the flexibility to choose and place cores within a die as well as within different layers in the stack. In practice, since all these IP-cores are integrated in one platform, it is expected that any simulation model must run all applications equally according to their requirements. Thus, the simulation engine must be a general purpose type that accommodates such demands. Each IP-core can then be easily configured and evaluated based on a set of parameters. The engine is composed of several key components with scalable features. The engine is also extensible and new components can be added when required.

A. Router Micro-architecture

Any NoC is comprised of routers located in general at the corners, edges, surface and center of the physical structure. For example a 2-D router has four links to other routers in 4 directions, namely, North, South, East and West. A 3-D router has two additional, Up and Down links giving a total of 6 bi-directional links. There is also an additional bi-directional port from router to resource in each case. For a given NoC structure, not all of the links can be used for routing packets. For example, only half of the router-to-router links in the corner of the network are connected, halving the available router bandwidth due to the unconnected edge ports.

In our simulation model, a single router template with reconfigurable ports is used. By detecting its network position at run-time, a router auto-enables its ports and makes connections with neighboring routers. The length of the connection links can be specified based on the physical dimension of the individual resources. This feature is particularly important in heterogeneous systems with each IP-core delivered by a third party with fixed physical sizes.

The NoC protocol employed in the model is a hot-potato implementation with router architecture as described in [8]. The routing strategy is based on non-minimal and load dependent deflection type packet switching, with adaptive per hop routing. A relative addressing scheme is implemented which

simplifies the duplication of identical routers when network structures of varying sizes are designed.

The routers instantiated in all of the network configurations are bufferless. A packet cannot be stored in a router, and thus in each cycle the packets must be moved to the next router or deflected back to a resource. The implemented router uses separate input and output links for each router-to-router and resource-to-router connection.
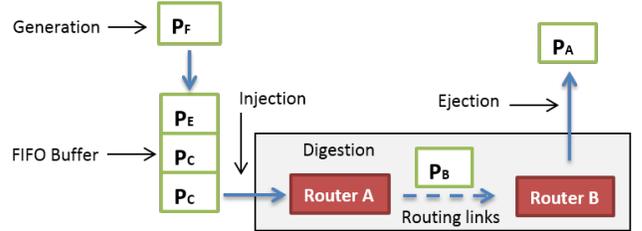

Fig. 2. A simple model showing the packetization process

B. Traffic Generator

The simulation engine has a number of built-in temporal and spatial traffic patterns. These traffic patterns can be used to stress-test a design under realistic scenarios so that the network configuration and placement strategies can be defined to produce the maximum performance and efficiency.

The temporal component of the traffic specifies the number of packets to be generated per cycle within a pre-defined time period. In our simulator, a Bursty model (B-Model) is used to emulate bursty traffic. The B-model uses the property of *self-similarity* exhibited by real traffic [9] to define temporal traffic patterns.

Spatial traffic specifies the destination address in the network that is set according to a synthetic traffic pattern in use. Patterns such as uniform random traffic (URT), local traffic, bit-reverse, bit-complement, bit-rotation, bit-shuffle, and hotspot are used. Each pattern differs from the other by the traffic probability of sending a packet from a source to every destination in the network. In URT, every destination gets packets with equal probability, whereas in local traffic, the destinations placed close to the source get packets with higher probability than those placed farther.

C. Network Interface (NI)

The Network Interface (NI) is a block, connecting a resource to a router, where the packetization and queuing of packets in first-in, first-out (FIFO) buffers takes place before injection. In our simulation model, packetization is done by creating an envelope to encapsulate data ready to be delivered to a destination through the network. The model in Figure 2 shows the abstraction of flow of packets in a network. Each packet specifies its source, destination, packet id, and router-to-router hop counter. To extract more network information
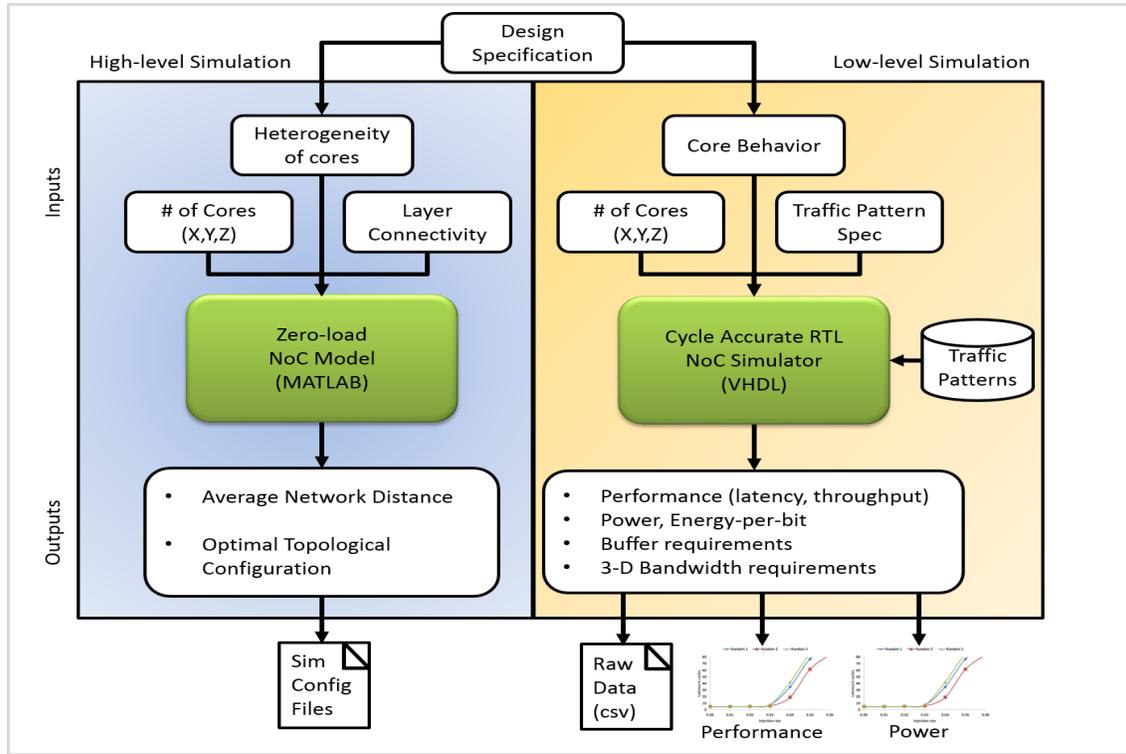
Fig. 3. The Simulation Environment with Inputs and Outputs

additional parameters such as FIFO delay time, individual horizontal (X, Y) and vertical (Z) hop counter can be set during simulation configuration. Each resource has a FIFO buffer to temporarily store packets that cannot be injected to the network due to congestion. The FIFO size is variable and unlimited. When a fresh packet is generated and ready for injection, the FIFO is checked for any queued packets. Packets that are already queued are re-injected into the network with higher priority than newly generated packets. The network does not drop packets. The packet headers generated by the resource contain final destination addresses and the routers make routing decisions on the fly based on this information.

D. Memory Model

A 3-D integration technology enables processor-memory stacking in many ways for different applications [5]. Our simulation model includes a complete memory system to address the needs of the different applications. For simple embedded many-core applications, the basic components that define the complete memory levels are L1/L2 cache, DRAM and NAND flash blocks.

E. TSV Model

TSVs are short and fast vertical interconnects as compared to the long and thin planar (horizontal) wires. In order to exploits such unique TSV features, the simulator provides TSV models which can be optimized and reused for different applications allowing high-speed inter-layer communication.

For example, double data rate (DDR) configuration of inter-layer communication of the simulation model uses a clock pumping technique to vertically deliver two bits per cycle through a single TSV [10].

An accurate TSV model is particularly important for 3-D chip design as it is relatively expensive and complex to prototype. Each TSV model can be specified as a bundle where a number of TSVs are set and the clocking mode is configured. It has sender and receiver components in its internal configuration, physically connected with TSV bundles to the immediate cores in both ends.

III. NETWORK MONITORING SERVICE

Network monitors are used to check runtime activities of a router. The information can be used by the same or other routers to make future routing decision. It can also be extracted and used by designers to analyze the network behaviour of a simulated system and make design decisions. In our model, the network monitoring service will be used to manage network traffic congestion, system faults, thermal variations and other challenges that may arise with the 3-D integration.

A. Fault Monitors

As 3-D integration technology advances, the complexity and the challenges rise. Increasing TSV aspect ratios, thinning of wafers, and increasing heterogeneity of the integration bring about new challenges or increase the complexity of the existing challenges. Reliability becomes a burning issue. Our simulator, fault monitors can be configured in a router to detect transient

and permanent faults. Several fault tolerant schemes can be simulated. A permanent fault can be introduced to a router and the extracted information can be used to make comparative analysis with other network configurations.

B. Energy Meter

The energy meter measures router-specific power consumption per cycle. The dynamic power of a specific router can be measured based on the rate of packet flow. The power consumption of a network can be extracted by synthesizing for specific technology node and frequency. By employing the values extracted from the energy meter and from the synthesis report, energy per bit can be calculated for the simulated system.

C. Link Monitor

The link monitor reads the per cycle link activity in a router. The information is mainly extracted as an output from the simulator for further analysis of link utilization of the simulated network configuration.

D. Thermal Monitor

Based on the traffic activity, the heat level can be monitored in hotspot regions of the network. This is important for 3-D systems as thermal heating is one of the major challenges. The monitoring feedback can also be used to adjust heat level (for example by reducing clock speed)

E. Buffer Monitor

Buffer monitors measure the packets contained in a FIFO buffer. Buffer sizing is one of the challenges in NoC design. In order to capture all packets without dropping or without limiting resource activity, the buffer size must be optimized. Any extra buffer will cost more area and power. Any less could lead to packet dropping. Information about buffer size in use in each router can be extracted and analyzed from the simulator model.

F. Congestion Meter

As the number of cores increases, relieving network traffic congestion becomes the major challenge. Packets are deflected to a different path whenever there is no free link in the intended direction. In order to avoid unnecessary hopping towards a congested part of the network, it requires continuous monitoring for any congestion in that part. The congestion meter measures router congestion levels in its recent past history (previous 16 cycles). Based on the reading, neighbouring routers will make dynamic decisions whether or not to send a packet towards that direction. The reading can also be captured and extracted from the output for further analysis.

IV. THE SIMULATION FLOW

Figure 3 shows the basic blocks and flow of the simulation model. There are two major parts of the flow: Zero-load model and RTL simulator. Preparation for model configuration starts at the entry point of the flow. Design specification regarding network size, IP cores' dimensions, TSV-specific

settings such as clocking, traffic patterns Based on the specification, Spatio-temporal traffic with varying injection rates are generated. The zero-load model produces information that doesn't require RTL simulations such as average network distance and optimal topological configuration based on the network size and IP core dimensions that can be used for zero-load network analysis. All outputs from this simulator are simulation configuration files that in turn can be used as input to the RTL simulator.

The RTL simulator top-level file integrates all the NoC components. A single test-bench reads all inputs including the network size in terms of X,Y,Z, injection rates, spatio-temporal traffic patterns and other relevant information. Data samples are extracted from the output files following the warm-up phase of the network and preceding the cool-down phase to ensure reliable results.

V. APPLICATION EXAMPLES

The simulation model can be used to evaluate several applications running in regular, irregular or heterogenous system configurations. Most standard simulators analyze regular networks with identical routers and resource sizes. However, in practice, NoC routers are used in many different application chips integrated with heterogenous IP cores and with dissimilar technologies. As an example, figure 4 shows the top-view model of a 3-D integrated mobile application processor. It has processing elements of different sizes, off-chip interfacing blocks, L2 and L3 caches, as well as IO interface positioned at the center of the chip. The wide-IO communicates to DRAM stack placed on top of the processor [11].

Our simulation model is able to configure such heterogenous systems with irregular networks by taking the abstraction of the system. The processing elements, interfacing units and
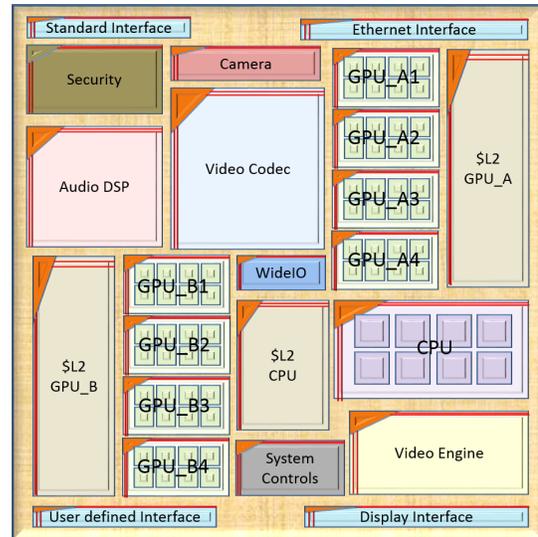

Fig. 4. An Application Example of Heterogenous Integration of Mobile Application Processor

memory cores are considered as resources each connected to its own router. The frequency of communication between

Throughput	RawLatency	Hopcount	Zero − load
0.010028	19.151449	3.787862	3.772826
0.020028	19.413949	3.853487	3.819785
0.029997	19.354099	3.838525	3.804563
0.040075	19.343731	3.835933	3.778540
0.050053	19.483549	3.870887	3.803709
0.060084	19.595569	3.898892	3.817340
0.069947	19.574320	3.893580	3.791449
0.079966	19.687991	3.921998	3.800617
0.089938	19.746630	3.936657	3.799479
0.099978	19.841965	3.960491	3.804270
0.199944	20.775593	4.193898	3.808556
0.300047	22.089913	4.522478	3.806187
0.400069	24.446072	5.111518	3.802214
0.500000	30.063269	6.515817	3.808944
0.599953	392.389098	97.097275	3.804381
0.699981	1028.814062	256.203515	3.802026
0.800047	1668.793235	416.198309	3.804718
0.900019	2312.714280	577.178570	3.805594
1.000000	2948.796994	736.199248	3.805841

TABLE I

AN EXAMPLE OF DATA EXTRACTED FROM THE SIMULATION OUTPUT

resources both on- and off-chip is the traffic probability that defines where and when to send data. The physical dimension of each resource defines the link length between two neighbouring routers.

Table 1 shows an output from RTL level simulation for performance evaluation of a 64 core network arranged in a $4 \times 4 \times 4$ configuration. Basic metrics such as throughput, latency and additional information from network hop-count and delay at zero-load are presented as an example.

In every cycle, packets arrive at all destinations at a rate based on the injection rate and the congestion level of the network. The throughput per resource per cycle, λ, is defined in (1), where P_{Total} is the total number of packets received over the simulated range, N is the number of resources in the network and C is the number of cycles in the sampling region.

$$\lambda = \frac{P_{Total}}{N \times C} \quad (1)$$

The raw latency, T_{Raw}, is the distance traveled by a packet from the source to the destination address in terms of hop counts, denoted by H_C. The parameters C_{Final} and C_{Init} represent the final and initial clock cycles respectively. When the network is at zero-load, the raw latency is equivalent to the minimum distance.

$$T_{Raw} = \frac{(C_{Final} - C_{Init})}{H_C} \quad (2)$$

The model allows easy integration with basic development tools such as ModelSim and Matlab through which input variables can be set for measurement and results can be evaluated. Metrics of performance and power such as average network latency, processor-memory access latency, network throughput: ejection per core, hotspot bandwidth, power consumption of individual layers and hot spots, and core energy per bit can be measured. In addition, vertical and horizontal data traffic distribution and local and global cache utilization status can be rapidly extracted and analyzed.

VI. CONCLUSIONS

We propose a scalable simulation model that addresses the design challenges of 2-D and 3-D systems characterized by a high level of complexity and heterogeneity early in the design. The unique properties of TSVs and high-speed features of inter-layer communication are accurately modeled. A spatio-temporal traffic generator is included and several network monitors are added.

Moreover, the model can be easily configured to simulate specific applications such as wide-IO many-core processor memory stacking and mobile application processor development, or heterogeneous integration of dissimilar technologies such as implementation of femto-cell base-stations on-chip integrated with RF modules, baseband processor, and memory units. In the future, the model will be extended to include deterministic routing and wormhole packet switching schemes.

REFERENCES

[1] Yang Hu, Shouyi Yin, Leibo Liu and Shaojun Wei. "A Mixed-Level Modeling for Network on Chip Infrastructure in SoC Design," IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Kuala Lumpur - Malaysia, 2010

[2] Menwang Xie, Duoli Zhang and Yao Li, "Meshim: A high-Level Performance Simulation Platform for ThreeDimensional Network-on-Chip," IEEE 9th International Conference on ASIC (ASICON), pp. 349-352, Xiamen - China, October 2011

[3] Young Jin Yoon, Concer, N. and Carloni, L. "VENTTI: a Vertically Integrated Framework for Simulation and Optimization of Networks-On-Chip" IEEE International SOC Conference (SOCC), pp. 171-176, Niagara Falls, NY, 2012

[4] Gottschling, P., Haoyuan Ying and Hofmann, K. "GSNOC UI - A Comfortable Graphical User Interface for Advanced Design and Evaluation of 3-Dimensional Scalable Networks-on-Chip", International Conference on High Performance Computing and Simulation (HPCS), Madrid - Spain, 2012

[5] Awet Yemane Weldezion, Z. Lu, R. Weerasekera, and H. Tenhunen, "3-D Memory Organization and Performance Analysis for Multi-processor Network-On-Chip Architecture". In Proceedings of IEEE International Conference on 3D System Integration (3DIC 2009), San Francisco USA, September 2009.

[6] M. Grange, A. Jantsch, A. R. Weerasekera, and D. Pamunuwa, Modelling the computational efficiency of 2-D and 3-D silicon processors for early chip planning," in Proc. Int. Conf. Comp.-Aided Design (ICCAD), 2013, pp. 310-17.

[7] Awet Yemane Weldezion, R. Weerasekera, D. Pamunuwa ,L. Zheng and H. Tenhunen, Bandwidth Optimization for Through Silicon Via(TSV) bundles in 3D Integrated Circuits. In 3D Integration Workshop, The Design, Automation, and Test in Europe (DATE) conference, Nice France, April 2009.

[8] E. Nilsson. "Design and Implementation of a hot-potato Switch in a Network on Chip". Mémoire, Departement of Microelectronics and Information Technology, Royal Institute of Technology, 2002.

[9] Park, K., W. Willinger, "Self-Similar Network Traffic: An Overview, in Self-Similar Network Traffic and Performance Evaluation," edited by K. Park and W. Willinger. Wiley-Interscience, 2000.

[10] Awet Yemane Weldezion, R. Weerasekara, H. Tenhunen, Design Space Exploration of Clock pumping Techniques to Reduce Through Silicon Via TSV Manufacturing Cost In 3D Integration. in Proceedings of the 14th IEEE Electronics Packaging Technology Conference (EPTC 2012), Singapore, December 2012.

[11] Wide I/O Single Data Rate (WIDE I/O SDR), JEDEC standard specification, Document number JESD229, December 2011 http://www.jedec.org/news/pressreleases/jedec-publishes-breakthrough-standard-wide-io-mobile-dram.