PAT-Noxim: A Precise Power & Thermal Cycle-Accurate NoC Simulator

Amin Norollah[1], Danesh Derafshi[1], Hakem Beitollahi[1], and Ahmad Patooghy[2]

[1]Department of Computer Engineering, Iran University of Science & Technology
beitollahi@iust.ac.ir
{a_norollah, derafshi_danesh}@comp.iust.ac.ir
[2]Department of Computer Science, University of Central Arkansas, USA
apatooghy1@uca.edu

***Abstract*—Ever increasing number of on-chip cores magnifies the need for efficient Network on Chip (NoC) architecture designs. To have a wide design space exploration, accurate simulators play a key role to estimate power consumption, area and temperature profile of NoCs. Access-Noxim as one of the well-known NoC simulators is used by several researchers to perform cycle-accurate simulations for different NoC architectures. However, this simulator lacks power and thermal accuracy due to ignoring the temperature/power cross impact which is also known as Temperature Effect Inversion. This paper presents PAT-Noxim, to address this source of inaccuracy in the Access-Noxim simulator. PAT-Noxim uses architecture-level tools to calculate power consumption and area more accurately with considering the power/temperature cross impact. PAT-Noxim simulator now supports several architectural alternatives including pipelined and non-pipelined routers, and arbitrary number of virtual channels per each port. In addition, the PAT-Noxim renders results for a wide ranges of VLSI fabrication technologies from 90 nm down to 22 nm. Obtained results by the PAT-Noxim simulator indicate 5.45% and 11.6% more accuracy in estimation of power consumption and temperature parameters respectively.**

***Index Terms*—Energy Efficiency, Temperature Effect Inversion, Design Space Exploration, Network-on-Chip.**

I. Introduction

Network-on-Chip (NoC) is nowadays accepted by most of researchers and hardware designers as an efficient communicating architecture for multi-core systems. However, power consumption and temperature are two limiting factors against the performance of NoC enabled architecture [1]. The power consumption of the on-chip network is the summation of dynamic power consumption and static power consumption. In technologies above 90 nm, the dynamic power consumption dominates total power consumption; while, today the role of static power consumption is more significant in technologies below 90 nm [2]. Lakshminarayanan et al. [3] show that the power consumption has direct relation with the temperature of the chip such that increasing rises the other. In general, power and temperature have significant effects on reliability, lifespan of the chip, performance and cooling cost. Therefor, it is very important to have a good estimation or measurement about the power consumption and temperature of the chip. Such information might help designers to do a wider design space exploration and find the better design solutions. In this regard, several research teams developed NoC simulators in high level programming languages such as C/C++ [4], [5], [6] or hardware description languages [7], [8], [9].

Noxim is a NoC chip simulator tool [7] written in C++ language, based on the SystemC library. The implemented routing strategy can be modified easily by writing and testing code using C++ programming language. Noxim uses a runtime engine to support runtime modifications in some parts of the network architecture. For instance, the buffer depth, packet size, traffic distribution and routing algorithm can be changed at runtime without any recompilation [10]. The Access-Noxim simulator [11] is the upgraded version of the Noxim and improves the power calculation, along with obtaining the temperature of routers and processing cores at specified time intervals using the Hotspot temperature model.

The ATLAS simulator [8] written in Java, is presented by the GAPH Group to automate the design flow of some networks on chips (e.g. Hermes, Mercury). VHDL and SystemC are used to implement hardware descriptions and test-benches, respectively. Booksim-2 [5], a cycle accurate simulator for interconnection networks, is presented by the Concurrent VLSI Architecture team at Stanford University. This simulator is implemented by C++ and has a flexible modular architecture. DARSIM [6] is a cycle accurate simulator for NoCs that is written in C++ language. It uses the architecture of ingress-queued wormhole router for simulation. This simulator supports 2D and 3D mesh networks.

Although there are some available NoC simulators developed by researchers in the literature, to the best of our knowledge, none of them considers the power/temperature cross impact. In this paper, we enhance the last improved version of the Access-Noxim simulator to measure power consumption, temperature of the chip and area of the circuit more efficiently and accurately. Our tool is called PAT-Noxim[1]. In PAT-Noxim, several parameters are added to the tool and new methods have been proposed for measuring the mentioned metrics. Simulation results prove our claim that PAT-Noxim measures power-consumption and area more accurately in

[1]PAT-Noxim is an open-source code and available in https://github.com/Amin-Norollah/PAT-Noxim

978-1-5386-1491-4/18/$31.00 ©2018 IEEE

compare to the last version of Access-Noxim.

The reminder of the paper is organized as follows. Section II introduces the used power, temperature, and area model in the PAT-Noxim. Section III examines the effect of temperature on the leakage current and introduces a new way to calculate the leakage current through a linear equation. In Section IV, the architecture of PAT-Noxim is presented. Section V evaluates the PAT-Noxim simulator and finally Section VI concludes the paper.

II. Power, Temperature and Area Models

We use a set of recently proposed and accurate models to measure different design parameters in the PAT-Noxim simulator. The models are added as some additional classes to the PAT-Noxim in some cases (e.g., Orion 3). In other cases (e.g., Hotspot), some constant values are calculated using the model, then the constant values are used in the code of PAT-Noxim. In particular we use the following models.

Orion 3 [12], [13], [14]: As the network bandwidth increases, the power consumption of the network rises as well. One way to approximate the power consumption is to use the architectural power model [12]. This model is an accurate and high-speed solution to measure the power consumption of the router components for different manufacturing technologies. The Orion 3 model is built on the architectural power model and supports the majority of router architectures for different technologies. The third version of the simulator adds power and area models for clock and links, along with consideration of virtual channels [13], [14].

The system-level simulator *McPAT* [15]: Thanks to its ability to include the interconnections between cores, McPAT is capable of modeling area, power consumption, and timing in multi-core systems for technologies between 90 to 22 nm [15]. The simulator takes parameters such as number of cores, type of interconnections, amount of required memory, and also core related parameters such as in-order or out-of-order nature of the core, number of bits per instruction, data, and the number of processing components for integer, decimal and complex numbers, to calculate the static and dynamic power consumption. It also measures the time parameters and the occupied area for each component, using a system-level model. However, one needs to run a benchmark on the proposed processor in order to use all options of McPAT.

Hotspot [16]: This model estimates the temperature of the chip using s network of resistor/capacitor model for every components of the target system. To achieve this goal, the tool uses the occupied area and the power consumption of each component which are extracted by Orion 3 and fed into the Hotspot for temperature modeling.

III. The Leakage Current Measurement

The leakage current is directly dependent on the temperature of the chip so that the increased leakage current translates to a hike in the static power and temperature of the chip. This rise continues exponentially, until one of two values (temperature or leakage current) becomes stable. This phenomenon is known as TEI (Temperature Effect Inversion).

Therefore, considering the temperature generated by network elements not only increases the power consumption and the static to dynamic power ratio, but also increases the temperature of the chip and as a result, will improve the accuracy of simulation. Different models are presented to calculate the leakage current [17], [18], [19]. Given that Access-Noxim uses constant values of power for different components based on the constant temperature of the room ($25°C$); we need a model for obtaining the leakage current in the presence of temperature, regarding its initial value.

The leakage current in sub-90 nm technologies depends largely on sub-threshold and gate leakage currents. The sub-threshold leakage is the current between source and drain of a transistor in the weak inversion region and is directly related to temperature, threshold and gate-source voltages [20]. The leakage current of the gate is independent of temperature; hence, we ignore it in our calculations [21]. Thereby, we realize that the leakage model proposed by Yongpan Liu et at. [2] is appropriate for our purpose. They have proposed the actual leakage current behavior that can be used for blocks of the circuit at the architecture level. For different temperatures, one can calculate the leakage current, based on its value for the reference temperature ($25°C$ or $298.15K$). The leakage current of each transistor in this model is as follows:

$$I_{leakage}(T) = I_{subthreshold}(T) + I_{gate} \quad (1)$$

$$I_{subthreshold}(T) = I_{subthreshold}(T_{ref}) + \frac{I_{subthreshold}(T_{ref})}{T_{ref}^2} \times (2T_{ref} - \frac{q(V_{GS} - V_{th})}{nk})(T - T_{ref}) \quad (2)$$

The subthreshold leakage current for reference temperature is calculated as below:

$$I_{subthreshold}(T_{ref}) = \frac{W}{L}.(\frac{K.T_{ref}}{q})^2.e^{\frac{q(V_{GS}-V_{th})}{nKT_{ref}}} \quad (3)$$

where in equations 1 to 3:

- V_{th} is the threshold voltage,
- L and W are the device effective channel length and width,
- V_{GS} is the gate-to-source voltage,
- n is the sub-threshold swing coefficient for the transistor,
- V_{DS} is the drain-to-source voltage,
- q is the magnitude of the electrical charge on the electron with a value 1.6021761019 c, and
- k the Boltzmann constant is 1.380648x1023 eV/K.

Equation (2) is a linear approximation of the leakage current in presence of temperature [19].

IV. The PAT-Noxim Simulator

This section introduces our main expansions to the Access-Noxim simulator which are done in three domains: Adding virtual channel to pipeline routers, providing new models to calculate the metrics (power consumption, area and temperature) and adding manufacturing technologies to the tool. Next, we explain the simulation procedure using the PAT-Noxim simulator.

A. Major Expansions

Virtual Channel Pipelined Router: Virtual channels were first developed to solve the deadlock conundrum in NoCs, but were further used to improve the network latency and throughput. Today, it is common practice to use virtual channels in pipelined routers. PAT-Noxim supports virtual channels in its pipeline structure, leading to a precise measurement of the power consumption and occupied area of network components.

Power Consumption, Area and Temperature Modeling: To improve the power consumption model, we implement the Orion 3 and McPAT tools in the PAT-Noxim simulator. Moreover, Orion 3 and McPAT are used to measure the area of routers and processing elements, respectively. Orion 3 calculates the power consumption and the area occupied by transistors and logical gates such as NOT, NOR, NAND, MUX and D-FF for different router components. It can also consider the power consumption and the area of links and clock. We have modified Orion 3 to improve its coordination with the PAT-Noxim simulator which leads to a boost in calculation of the power consumption and the leakage current under temperature influence.

We incorporated the McPAT tool to calculate static power, dynamic power and occupied area of processing elements of several reference processors. Hotspot is used to obtain the temperature of each tile in an NoC. Access-Noxim employs the Hotspot tool to measure the temperature of routers, but due to the lack of an area model, it uses a constant area for components of each tile. But, in our model, the area of each component is dynamically calculated based on its manufacturing technology. The area value is passed to the Hotspot tool in order to measure the temperature of the chip. The temperature calculated by our tool is very accurate and close to the actual value as we dynamically provide the area model for hotspot.

Manufacturing Technologies: To better investigate the effect of TEI, we should use smaller manufacturing technologies such as 22 nm. In these technologies, the contribution of static power consumption in total power consumption is highlighted and thereby, the effect of TEI on temperature of the chip becomes significant [22], [23]. So, to have an accurate power consumption, the manufacturing technologies of 90, 65, 45, 32 and 22 nm are added to PAT-Noxim.

B. Architecture of PAT-Noxim Simulator

The C++ language and the systemC library are used to implement PAT-Noxim. Figure 1 demonstrates the overall

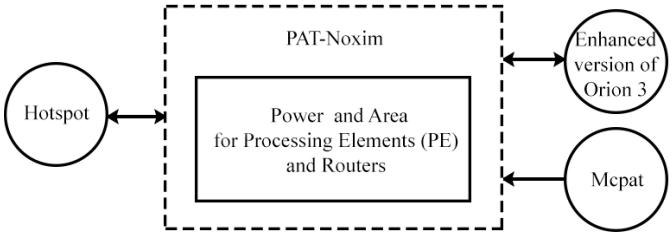

Fig. 1. Architecture of the PAT-Noxim simulator

schematic of PAT-Noxim. At the beginning of simulation, PAT-Noxim extracts and buffers the power consumption values of different components of routers, by examining the input parameters through the Orion 3 tool. The amount of power consumed by the processing elements and their required memory is obtained from a library stored in the project source. PAT-Noxim uses the library to cross-reference the processing element with its respective area and power consumption values.

Messages are created in various traffic patterns in each processing element, and after being divided into packets, and then into flits, they are ready to be sent. During sending each flit to its corresponding router in a tile, the process of sending is executed in the processing element. By running this process, the dynamic power required to send a flit from this router is measured and added to the total power consumption of the processing element. The same procedure is done for the receiving router as we measure the amount of power consumed to send the flit on its way to next hop and add it to the total power consumption of the router (relative to its architecture).

Once a packet is sent from the source processing element to a destination router, it has the following power dissipation properties:

- Creation of a packet and splitting it into flits in the processing element,
- Forwarding flits from the source processing element to its router by internal links,
- Power consumption due to input (output) buffer for each flit (for read and write),
- Virtual channel allocation stage for header flit,
- Switch allocation (SA) and crossbar stages for each flit,
- Power consumption due to the links between two routers for each flit.

After passing a specific number of cycles required to calculate the temperature, the simulator estimates the static power consumption and adds it to the total power consumption of each component. Then, the simulator sends dynamic and static power consumption values of a tile (router, processing element and their corresponding memory) to the Hotspot tool. Hotspot models each tile with an RC model to measure and send the temperature of all tiles back to PAT-Noxim. Then the received values are sent as feedback to Orion 3 where the static power consumption of each router is calculated for the second time in the presence of temperature. Static power is dependent on leakage current and is called TEI. Through the

formula (2) provided in the last section.

The Orion 3 simulator has been modified to calculate the static power in the presence of temperature based on formula (2) so that it requires the temperature of router to obtain the leakage current. Then the static power value is sent back to PAT-Noxim.

In Access-Noxim, a user must manually change the characterization of latency, power consumption and area when the architecture of a router or a processing element is changed. This issue may lead to occurrence of error in the system and moreover, it is a time-consuming task. In PAT-Noxim, a user can modify the architecture of routers and processing elements dynamically which leads to reduction of error occurrence and higher accuracy in computation.

V. Evaluation of the PAT-Noxim Simulator

In this section, we setup for power, performance and area experiments using the PAT-Noxim simulator in various working conditions and technology sizes. In addition, we compare the PAT-Noxim results with those of Access-Noxim to show the amount of accuracy improvement of PAT-Noxim simulator.

A. Platform Description

Figure 2.a demonstrates the arrangement of 64 tiles in a $4\times4\times4$ 3D mesh architecture. Each node contains a processing element, an L2 cache, and a router. In this architecture, a processing element needs to communicate with other existing processing elements in the network during the execution of its tasks. It creates a packet for its message, splits it up into flits and hands it over to the corresponding router via physical lines.

Figure 2.b shows the architecture of a pipelined router with p input and p output ports including virtual channels. Each input port includes v virtual channels (VC1, VC2, VC3, ... VCv). The control unit of each router contains following components: 1) routing computation (RC) unit which determines the appropriate outgoing port for every packet, 2) virtual channel allocation (VA) unit which finds an idle output VC from the output port determined by RC unit to forward packet, and 3) switch allocation (SA) unit which allocates output ports of the crossbar switch to requesting packets. The input and output routers are interconnected using a crossbar which is controlled by the SA unit.

B. Power, Area & Thermal Consumption

Figure 3 and 4 shows the measured area and power consumption, respectively for each router in each cycle for technologies of 22, 32, 45, 65 and 90 nm with the same voltage and frequency. By shrinking the technology, the ratio of static power to dynamic power increases, which clearly states that the leakage current plays an important role in total power consumption of the chip. Since the power consumption depends on temperature, we can conclude that by shrinking the manufacturing technology, temperature will have a higher impact on power consumption of the whole chip.

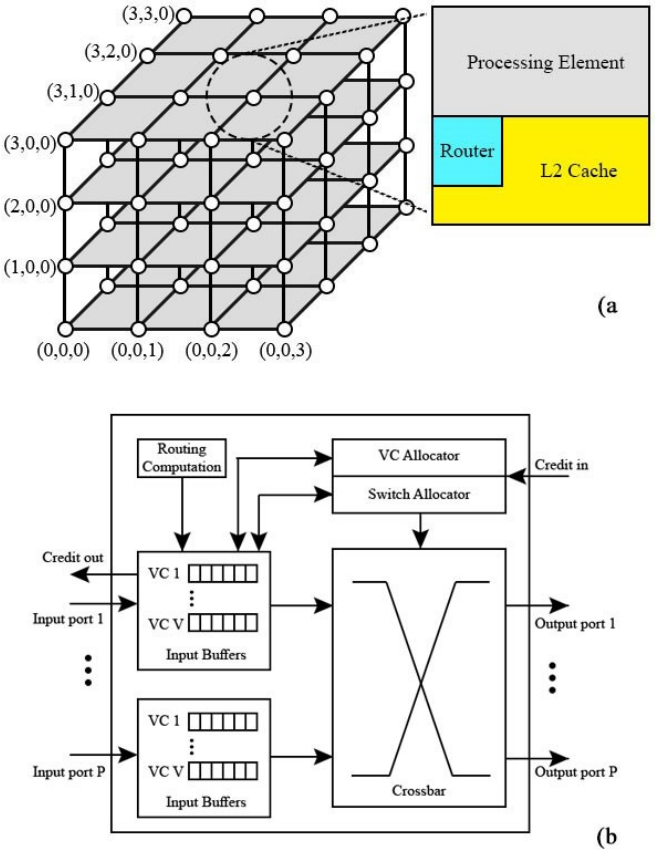

Fig. 2. a) 64 tiles in a 3D mesh NoC, b) The architecture of a pipelined router

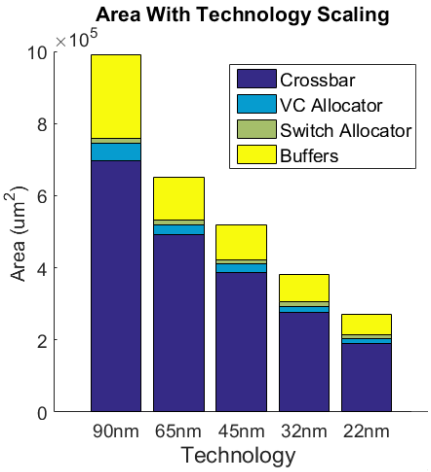

Fig. 3. Area report of router components for 22, 32, 45, 65 and 90 nm technologies

Figure 5 compares the calculated power consumption in Access-Noxim with PAT-Noxim in 65nm technology. For the normal simulation, PAT-Noxim is configured to resemble the network applied in Access-Noxim. In the pipeline simulation, a three-stage pipelined routing architecture is considered with no virtual channel implementation. The pipelined-4VC simula-

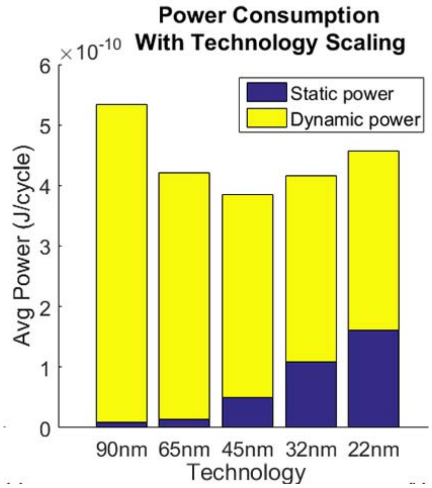

Fig. 4. Power consumption of router components under 22, 32, 45, 65 and 90 nm technologies

port there are 4 virtual channels. This figure indicates that the PAT-Noxim simulator calculates the power consumption, 5.45% more accurately compared to Access-Noxim in normal simulation test.

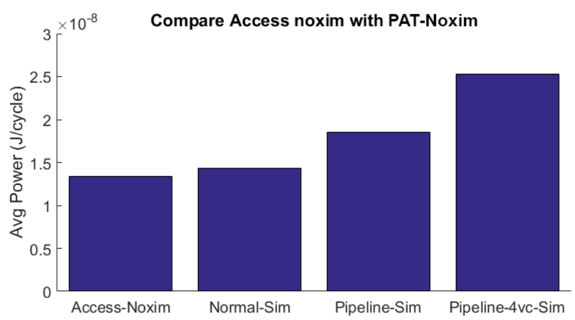

Fig. 5. Comparing the measured power consumption of Access-Noxim with PAT-Noxim

Figure 6 indicates the effect of TEI as described in Section III. Temperature and power consumption levels will increase simultaneously till a cooling system alleviates the temperature and settles it down to a constant value. The Hotspot tool offers a built-in cooling model in which the user can modify the chip cooling parameters to his/her preferences. As shown in Figure 6, after 30 million cycles past the start point of the network, the temperature and static power consumption have reached a mutual equilibrium. To see the effect of TEI, the total static power consumption and the average total temperature of the network are calculated with and without TEI effect. Our results show that the accuracy of the mentioned metrics is enhanced 12.3% and 11.4% more accurately than the Access-Noxim simulator, respectively, when TEI is considered. In these simulations, we improved the accuracy of temperature measurement using the precise area of each element.

Figure 7 shows the temperature of different layers of tiles in a network on chip after 40 million clock cycles, taking into account the effect of temperature inversion. We also assume that

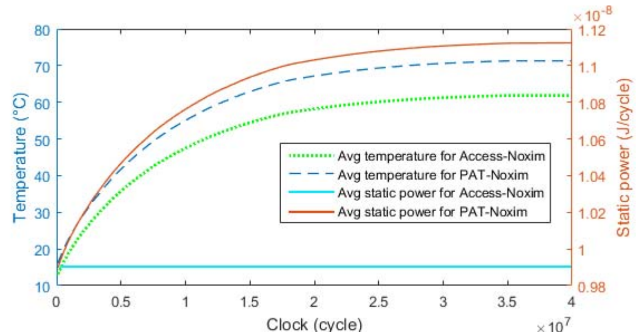

Fig. 6. Comparing Access-Noxim and PAT-Noxim in the presence of TEI effect on temperature and static power consumption

processing cores consume power only for sending messages to another core. In the temperature model, we use the default configuration of heat sink, heat spreader and heat transfer path provided by Hotspot. As we traverse further layers from the heat sink, the temperature of the upcoming layers increases as well, but according to Figure 6, the temperature stabilizes after 30 million cycles.

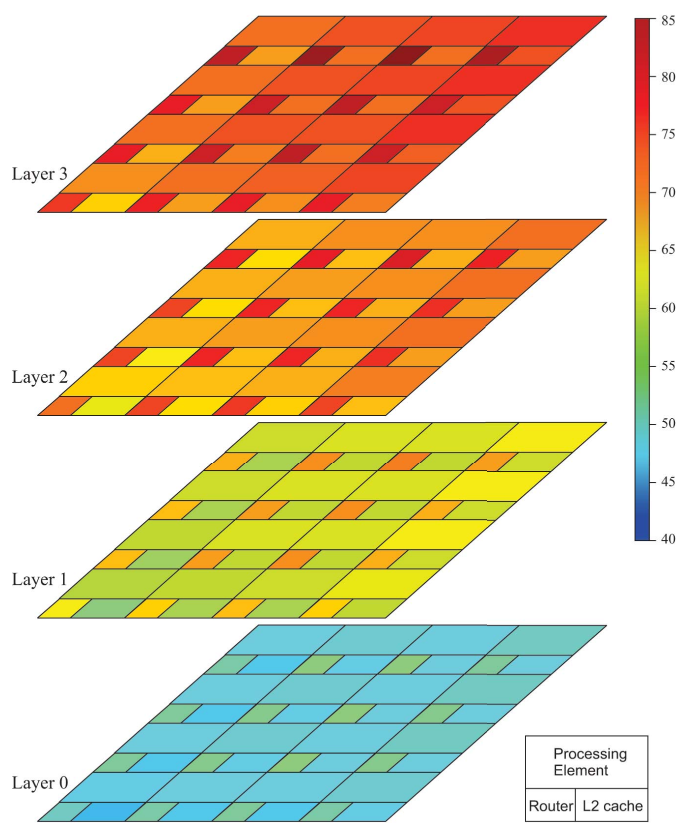

Fig. 7. The temperature of different layers of tiles after 40 million clock cycles

C. Throughput

The throughput of a NoC is defined as a rate in which message traffic can traverse across a network. An increase in the number of virtual channels is translated to an increase in the number of completed messages. Nonetheless, the throughput

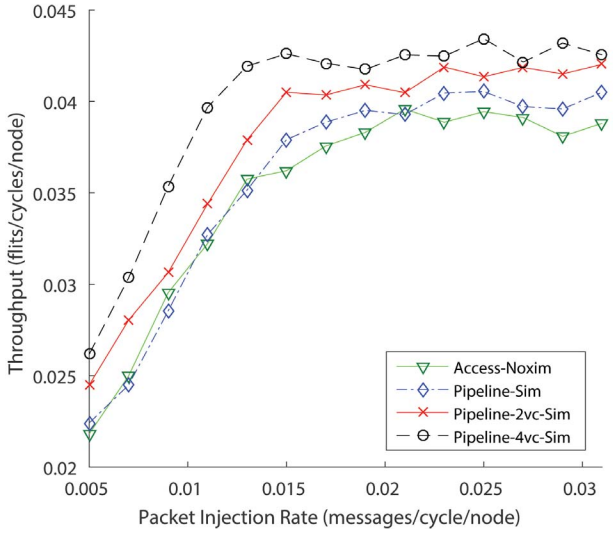

Fig. 8. Network throughput under different router architectures

is saturated when the number of virtual channels is increased beyond four [24]; hence, one, two and four virtual channels are used for simulations. Figure 8 depicts the amount of throughput versus packet injection rate for different architectures in both PAT-Noxim and Access-Noxim.

As can be seen the amount of throughput is increased proportionally with the number of virtual channels. The sporadic reaction of each individual graph to the increase in packet injection rate is originated from random traffic distribution of packets.

VI. Conclusions

This paper enhances the Access-Noxim simulator to measure power consumption, temperature of the chip and the area occupied by the components more accurately. Unlike the previous NoC simulators, we use the area of both routers and processing elements in order to calculate the chip temperature. PAT-Noxim captures temperature effect inversion phenomenon to estimate the power consumption of the chip more realistically. The PAT-Noxim simulator offers 5.45% and 11.6% better accuracy in calculating power consumption and temperature, respectively compared to the Access-Noxim simulator. Finally, the tool leverages the simulation flexibility via supporting base-line router architecture, pipelined architectures, and arbitrary number of virtual channels.

References

[1] G. N. Hardavellas, M. Ferdman, and B. Falsafi, "Toward Dark Silicon in Servers," *IEEE Micro*, vol. 31, no. 4, pp. 6–15, 2011.

[2] Y. Liu and et al., "Accurate Temperature-Dependent Integrated Circuit Leakage Power Estimation is Easy," in *Proc. Design, Automation and Test in Europe Conf*, March 2007, pp. 204–209.

[3] V. Lakshminarayanan and N. Sriraam, "The effect of temperature on the reliability of electronic components," in *IEEE International Conference on Electronics, Computing and Communication Technologies (IEEE CONECCT)*, July 2014.

[4] A. Nayebi, S. Meraji, A. Shamaei, and H. Sarbazi-Azad, "Xmulator: A listener-based integrated simulation platform for interconnection networks," in *First Asia International Conference on Modelling Simulation (AMS'07)*, March 2007, pp. 128–132.

[5] N. Jiang, D. U. Becker, G. Michelogiannakis, J.Balfour, B. Towles, J. Kim, and W. J. Dally., "A Detailed and Flexible Cycle-Accurate Network-On-Chip Simulator," in *IEEE International Symposium on Performance Analysis of Systems and Software,*, 2013.

[6] M. Lis, K. S. Shim, M. H. Cho, P. Ren, O. Khan, and S. Devadas, "DARSIM: A parallel cycle-level NOC simulator," in *6th Annual Workshop on Modeling, Benchmarking and Simulation*, 2010.

[7] V. Catania and et al., "Noxim: An open extensible and cycle-accurate network on chip simulator," in *IEEE 26th International Conference onApplication-specific Systems Architectures and Processors (ASAP)*, March 2015, pp. 162–163.

[8] A. Mello, N. Calazans, and F. Moraes, "ATLAS-an environment for NoC generation and evaluation," in *Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2011.

[9] 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 *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, April 2013, pp. 86–96.

[10] P. P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Performance Evaluation and design Trade-offs for Network-On-Chip Interconnect Architectures," *IEEE Transactions On Computer*, vol. 54, no. 8, pp. 1025 – 1040, 2005.

[11] *Access Noxim*, http://access.ee.ntu.edu.tw/noxim/index.html.

[12] N. Eisley and L.-S. Peh, "High-level power analysis for on-chip networks," in *Proceedings of the 2004 International Conference on Compilers, Architecture, and Synthesis for Embedded Systems*, ser. CASES '04. New York, NY, USA: ACM, 2004, pp. 104–115.

[13] A. B. Kahng, B. Li, L.-S. Peh, and K. Samadi, "ORION 2.0: A Power-Area Simulator for Interconnection Networks," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 1, pp. 191 – 196, March 2011.

[14] J. Fong, S. Nath, A. B. Kahng, and B. Lin. (2017) [21] orion3.0. [Online]. Available: http://vlsicad.ucsd.edu/ORION3/

[15] 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 *42nd Annual IEEE/ACM International Symposium on Microarchitecture*, New York, NY, USA, Jan 2010.

[16] R. Zhang, M. R. Stan, and K. Skadron, *HotSpot 6.0: Validation, Acceleration and Extension*, 2017.

[17] E. Cai and D. Marculescu, "Temperature Effect Inversion-Aware Power-Performance Optimization for FinFET-Based Multi-Core Systems," *IEEE Transactions on CAD of Integrated Circuits and Systems*, vol. 36, no. 11, 2017.

[18] S. N. Mozaffari and A. Afzali-Kusha, "Statistical model for subthreshold current considering process variations," in *2nd Asia Symposium on Quality Electronic Design (ASQED)*, Penang, Malaysia, August 2010.

[19] E. P. Vanndamme, P. Janson, and L. Deferen, "Modelling the subthreshold swing in MOSFETs," *IEEE Electron Device Letters*, vol. 18, pp. 369–371, 1997.

[20] M. Pedram and S. Nazarian, "Thermal Modeling, Analysis, and Management in VLSI Circuits: Principles and Methods," *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1487 – 1501, August 2006.

[21] W. P. Liao, L. He, and K. M. Lepak, "Temperature and supply voltage aware performance and power modeling at micro-architecture level," *IEEE TonCAD*, vol. 24, no. 7, pp. . 1042–1053, July 2005.

[22] E. Cai and D. Marculescu, "Temperature Effect Inversion-Aware Power-Performance Optimization for FinFET-Based Multicore Systems," *IEEE Transactions on Computer*, vol. 36, no. 11, 2017.

[23] Z.Abbas and M.Olivieri, "Impact of technology scaling on leakage power in nano-scale bulk CMOS digital standard cells," *Microelectronics Journal*, vol. 45, no. 2, 2014.

[24] P. P. Pande, C. Grecu, M. Jones, A. Lvanov, and R. Saleh, "Performance Evaluation and Design Trade Offs for Network on Chip Interconnect Architectures," *IEEE Transaction on Computers*, vol. 54, no. 8, pp. 1025 – 1040, August 2005.