Verification of Resilient Communication Models for the Simulation of a Highly Adaptive Energy-Efficient Computer

[Extended Abstract]

Mario Bielert
Technische Universität Dresden, ZIH
01069 Dresden, Germany
mario.bielert@tu-dresden.de

Kim Feldhoff
Technische Universität Dresden, ZIH
01069 Dresden, Germany
kim.feldhoff@tu-dresden.de

Florina M. Ciorba
University of Basel
Department of Mathematics and Computer Science
4003 Basel, Switzerland
florina.ciorba@unibas.ch

Stefan Pfennig
Technische Universität Dresden, Faculty of Computer Science
01069 Dresden, Germany
stefan.pfennig@tu-dresden.de

Elke Franz
Technische Universität Dresden, Faculty of Computer Science
01069 Dresden, Germany
elke.franz@tu-dresden.de

Thomas Ilsche
Technische Universität Dresden, ZIH
01069 Dresden, Germany
thomas.ilsche@tu-dresden.de

Wolfgang E. Nagel
Technische Universität Dresden, ZIH
01069 Dresden, Germany
wolfgang.nagel@tu-dresden.de

Motivation. Delivering high performance while operating in an energy-efficient manner is of great importance in conducting scientific research and in the use of daily life technology. From a computing perspective, a novel concept, namely the HAEC Box [3], utilizes innovative ideas of optical and wireless chip-to-chip communication to allow a new level of runtime adaptivity for future computers. To achieve high performance and energy efficiency, the HAEC Box creates a platform for flexibly adapting to the needs of the computational problem.

HAEC-SIM [1] is an integrated simulation environment designed for the study of the performance and energy costs of the HAEC Box running energy-aware applications. The design of the HAEC Box is based on individual abstraction models of hardware (e.g., CPUs, links), architecture (e.g., computing nodes, network), and software (e.g., runtime system, code generation).

Given the characteristics of the HAEC Box [3][1], any simulation of the execution of communication intensive benchmarks on this platform needs to account for packet loss due to errors, failures, or malicious attacks. Therefore, HAEC-SIM provides three resilient communication models: resilient dimension order routing (denoted DOR_r), resilient practical network coding (denoted PNC_r), and resilient dynamic network coding (denoted NCD_r).

Goal. Verification is an integral activity in the modeling and simulation process. The goal of this work, therefore, is the verification of the implementation of the resilient communication models [5] in HAEC-SIM.

Approach. The following steps were conducted for verification:

1 Selected two communication intensive applications (LU and BT) from the NAS Parallel Benchmarks Suite.
2 Traced of the execution of the class D version of the benchmarks using 4096 MPI processes on 256 16-way nodes of an HPC system at TU Dresden [6]. The resulting execution traces form the input to HAEC-SIM.
3 Specified the simulated HPC platform characteristics: a 3-D torus, with $16 \times 16 \times 16$ computing nodes. Given that, herein, only the communication models are verified, the computational power of the target platform is assumed equivalent to that of the execution platform, and that the commu-

nication links have characteristics similar to Infiniband: 700 ns latency and 54,54 Gbit/s bandwidth.
4 For each application, launched HAEC-SIM with the three resilient communication models.
5 Used a verification tool to compare the transfer time for each point-to-point message of the application obtained from simulation and from an independent implementation of the models based on polynomial approximation.
6 Assessed the acceptability of the verification comparative results.

Applications and traces. , while BT.D.4096 spends approximately 48 %. From the traces of the two communication intensive benchmarks (denoted as LU.D.4096 and BT.D.4096) executed on Taurus [6], we observed that they spend ≈68 % and ≈48 % respectively of the execution time in MPI functions. In both cases, processes communicate primarily with their neighbors regarding the application topology.

Simulated HPC platform. For the simulation scenarios, we consider a 3D torus topology, with $16 \times 16 \times 16$ computing nodes and Infiniband links. Each link is assumed to have a 0.01 packet loss probability. The MPI processes of the benchmarks are mapped to the simulated compute nodes in an xyz order.

Communication models. Each point-to-point message exchanged between MPI processes of the application is split into multiple packets of 1500 bytes each. Sender nodes compute a digital signature for every packet, to prevent their unrecognized modifications. Digital signatures are verified by the intermediate and receiver nodes. Packets are sent according to three communication models:

DOR$_r$: Denotes the common store-and forward approach of sending packets. Similar to TCP, packets are organized into windows. The receiver confirms the successful receipt of each packet by means of acknowledgments. Upon sending one window, the sender starts to send packets of the next window.

PNC$_r$: Refers to Practical Network Coding [2]. The sender organizes packets in matrices called generations (of 5 packets) and computes random linear combinations out of the packets of one generation. The receiver decodes upon receiving sufficient linear independent combined packets. It sends acknowledgments to confirm the current rank of the matrix of received packets. The sender proceeds to the next generation upon completing the sending of the current one.

NCD$_r$: Similar to PNC$_r$. However, the sender estimates the delivery probability by means of the receiver acknowledgments. Thus, the sender can estimate the number of remaining packets to be sent such that the receiver will have full rank [4].

Simulation and verification. A total of six parallel HAEC-SIM simulations have been conducted for each benchmark with each communication model. Due to the high complexity of the simulation process, high performance computing was necessary to deliver meaningful verification results. The simulator writes the transfer time of each message and the number of hops traveled in the simulated output trace. This data is extracted by the verification tool and compared against the polynomial approximation of the data obtained from the implementation of the models in Sage [5]. The verification reveals the fact that the data from simulation and polynomial approximation is identical.

Conclusions and Future Work. This work represents the first step towards accurate HAEC-SIM-based simulations for studying the behavior of communication intensive applications running on the HAEC Box using three resilient communication models. Future work directions include developing communication models that consider link congestion and collective communication (or multicast messages).

1. REFERENCES

[1] M. Bielert, F. M. Ciorba, K. Feldhoff, T. Ilsche, and W. E. Nagel. HAEC-SIM: A Simulation Framework for Highly Adaptive Energy-Efficient Computing Platforms. In *Proc. of the Conference on Simulation Tools and Techniques*, 2015.

[2] P. A. Chou, Y. Wu, and K. Jain. Practical Network Coding. In *Proc. Annual Allerton Conf. on Comm., Control, and Computing*, 2003.

[3] G. Fettweis, W. Nagel, and W. Lehner. Pathways to servers of the future. In *Proc. of the Design, Automation Test in Europe Conference Exhibition*, Mar 2012.

[4] S. Pfennig and E. Franz. Adjustable Redundancy for Secure Network Coding in a Unicast Scenario. In *Proc. of International Symposium on Network Coding*, 2014.

[5] S. Pfennig, E. Franz, F. M. Ciorba, T. Ilsche, and W. E. Nagel. Modeling communication delays for network coding and routing for error-prone transmission. In *Proc. of the 3rd Intl. Conf. on Future Gen. Comm. Techn.*, pages 19–24. IEEE, Aug 2014.

[6] Technische Universität Dresden, ZIH. HPC System Taurus. `https://doc.zih.tu-dresden.de/hpc-wiki/bin/view/Compendium/HardwareTaurus`.