Geographic Arithmetic Routing for Exascale Interconnects

Caroline Concatto, Jose A. Pascual and Javier Navaridas
{caroline.concatto,jose.pascual,javier.navaridas}@manchester.ac.uk
School of Computer Science
The University of Manchester
Manchester, United Kingdom

ABSTRACT

Reaching Exascale computer ability poses tremendous challenges to the computer architecture community. In ExaNeSt we are developing a novel computing architecture putting together the latest, most promising emerging technologies: liquid cooling, optical interconnects and non-volatile memories. These technologies are being leveraged to produce a small-scale, 2-cabinet prototype. Here we present an overview of our FPGA-based custom-made top-of-rack switch that envisions a paradigm shift by avoiding the costly and inefficient, both in terms of area and energy, routing tables enrooted in most current networking technologies used in HPC systems.

KEYWORDS

Exascale systems, Interconnects, Routing, Top of Rack Switch

ACM Reference format:
Caroline Concatto, Jose A. Pascual and Javier Navaridas. 2017. Geographic Arithmetic Routing for Exascale Interconnects. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 1 pages.
https://doi.org/0000001.0000001

1 TABLE-FREE ROUTING FOR EXASCALE

We are currently designing an advanced computing rack, consisting of densely-packed 16nm Xilinx UltraScale+ FPGA SoCs with low-power 64-bit ARM processors, in-node distributed NVM storage and immersion liquid cooling [2]. This platform provides a huge raw computing power and aims at accelerating a diverse set of applications relevant to exascale – scientific, engineering, big data, and cloud computing. Due to the system scale and peculiarities, the interconnection network is essential to *minimize* I/O, communication and synchronization delays as well as to *maximize* the sustained computing throughput achieved by the applications. Also, the system scale commends a multi-tier interconnect: AXI in the lowest-level, APENet+ within the cabinets and an FPGA-based, custom-made interconnect among cabinets. We focus here on the design of the Top-tier switch architecture, see Fig. 1, whose main novelty is to avoid using *routing tables* by devising an arithmetic routing scheme based around geographical addressing.

This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 671553 (ExaNeSt).

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/0000001.0000001

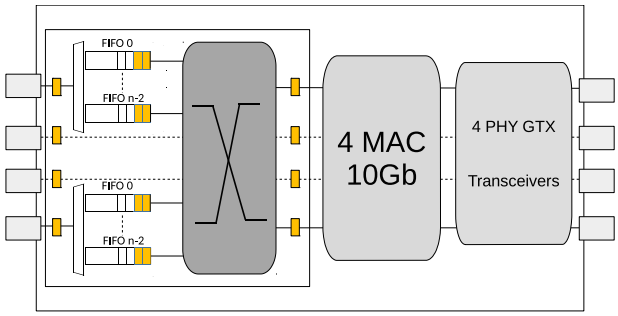

Figure 1: 4-port Router Architecture

This contrasts with the typical technologies used in large-scale HPC systems (see Top500 list) and datacentres – e.g, Ethernet, Aries, Omnipath or Infiniband – where we can see that routing tables are an integral part of the design [1]. These tables are the biggest consumers of chip resources (i.e. *area* and *power*) in the switching elements and, indeed, they are usually much larger than the router logic (i.e. buffering, flow control mechanisms plus the crossbar) as they need thousands of entries to sustain large number of endpoints without harming the performance. This is because the table-updating protocols, invoked when there is a change in the network or when an address is not in the routing table, pose huge overheads in terms of extra traffic. Furthermore, inconsistent states can be produced when some routing tables are updated but others are not which, in turn, can produce pathological behavior such as misrouting or livelock. Larger networks require more, larger tables which exacerbates these challenges.

Our TOR Switch design, re-purposes the area saved by not using routing tables so to integrate Virtual Output Queues (VOQ) which will improve significantly the performance of the network by reducing the likelihood of congestion being formed. Instead of routing tables it uses an arithmetic routing policy based on the positions of the router in the Network. This routing policy only uses a register with the local address and a comparator to select the output port. Fig. 1 shows the router architecture: input registers, arithmetic routing engine followed by the VOQs and a non-blocking switch allocator currently using a round robin policy. Traversing the router takes 3 clock cycles (with no contention) and uses a custom-made packet format devised within ExaNeSt. At present, it is interconnected using 10Gbps transceivers, but faster links will be considered in the future.

REFERENCES

[1] Top 500. 2016 (accessed March 22, 2017). *Top500 Supercomputer ranking.* http://www.top500.org
[2] M. Katevenis et al. 2016. The ExaNeSt Project: Interconnects, Storage, and Packaging for Exascale Systems. *Conf. on Digital System Design (DSD)* 00 (2016).