
Researchers at The George Washington University are developing an innovative method to tackle the major challenges in on-chip communication: performance scaling, energy-efficiency, and reliability. The novel network-on-chip technology invented at GW will significantly improve performance, power, and reliability simultaneously. Considering that the innovation can be applied to any current and future parallel computing systems and the high demand for multicore parallel computing, we believe that in a couple of years this technology can be worth tens of billions of dollars.

The market for parallel computing is growing rapidly, and the global parallel computing market had been valued at USD 31.20 billion in the year 2017, which is expected to reach USD 50.50 billion by 2023 growing at 8.35% CAGR. Parallel computing works on a network or a series of processing units that are itself bound by hundreds and thousands of computational cores. These cores are capable of executing a number of high-performance software and programs. The first consumer multicore processors hit the market in 2005, and super computers were using multicore processors as early as 2001. As technology scales, the number of cores on a single chip is increasing: GPUs and TPUs now can have multiple thousands of individual cores, while a supercomputer can carry up to over 100,000 cores. While the parallel computing market is pursuing higher performance of multicore systems, much of the perceived speed increases for those systems in the last 5–10 years has to do with improvements related to coordinating multiple cores on a multicore processor, not with individual processors becoming faster. For example, in fact, the 4th generation (2014) Intel core i7 has the same base processor frequency (3.60 GHz) as the 9th generation (2018) Intel core i7 — but the new version has twice as many cores, 8 instead of 4. Due to that fact, the communication among the cores is becoming the key to achieve higher performance for multicore systems: multiprocessing, parallel processing, and data parallelism all rely on network-on-chips for communication.

Over the past years, network-on-chips (NoCs) have emerged as the standard interconnect fabric solutions for connecting multiple cores, caches, memory controllers, and other hardware components on the chip. Unfortunately, NoC architectures are facing several urgent challenges including increased power consumption, waning reliability, and increased latency. There has been a significant amount of work recently devoted to solving these problems individually, yet simultaneously addressing these problems is proving to be difficult due to the explosion of the design space and the complexity of handling many trade-offs: different techniques can conflict and offset each other's desired goals.

Professor Louri has invented a method of simultaneously tackling the three urgent limitations of NoCs. With the novel NoC architecture design, it allows the router to proactively and intelligently switch among several different operation modes at runtime. Each operation mode has different trade-offs among fault-tolerant, transmission traffic, network latency and power consumption. A machine learning based controller is implemented in each router. The ML-based controller observes the NoC attributes at runtime and automatically evolves optimal per-router control policy, which selects the optimal operation modes at any given time. This technology is a
powerful innovation in multicore parallel computing and implicitly in several important hot areas, ranging from mobile processors, chip multiprocessors (CMPs) to supercomputers and exascale systems, with an enormous commercial potential in a growing number of environments.

Applications:

- On-chip communication for parallel computing systems (multi-core CPU, GPU, FPGA, TPU, etc.)


Advantages:

- Simultaneously tackles three urgent limitations of conventional computer architecture designs: performance, energy efficiency, and reliability.

- Experiment shows the innovation can reduce network latency by 55%, improve energy efficiency by 64%, and reduce retransmission caused by faults by 48% comparing to the conventional NoC design.

- The key innovations are scalable and can be implemented to nearly any current and future parallel computing world including mobile computing, IoT, multicore processors, embedded systems, servers, cloud computing, data centers, machine learning, etc.

Inventors

Professor Ahmed Louri

Ke Wang