An Approximate Communication Framework for Network-on-chips

Prof. Louri and his research team at the George Washington University have developed an approximate communication framework for network-on-chips (NoCs), which significantly reduces the latency and power consumption of on-chip data movement. The invented framework leverages the fact that big data applications (e.g., recognition, mining, and synthesis) can tolerate modest error and transfers data with the necessary accuracy, thereby improving the energy-efficiency and performance of multi-core processors.

Analyzing big data is critical for modern society. Retailers like Amazon, Costco, and Safeway analyze a large amount of user data, such as shopping and browsing history, to optimize their advertising and business strategy. As these companies expand their businesses each year, the need for processing big data is also growing (5—10%). Therefore, big data service providers, such as Google and Microsoft and chip manufactures such as Intel, AMD, and Nvidia have invested a large amount of human resources to enhance the efficiency and performance of the processing platform. One effective solution to ensure efficient big data processing is using approximation techniques. Considering that big data applications are error-resilient in nature, many approximation techniques are being developed to trade quality for better system performance and efficiency. For example, Google applies approximation techniques to their tensor processing unit (TPU). By reducing the precision of computation, TPU achieves significant gains on power and execution time. However, recent research on approximation technologies focuses mainly on the computation and neglects that current multi-/many-core systems spend large amounts of time and power to transmit data across on-chip interconnects. Recent studies show that NoC power consumption can reach up to 40% of the overall chip power. This problem is aggravated when a big-data application is executed on multi-/many-core systems, since processing a large amount of data incurs intense on-chip communication.

To this end, we propose an approximate communication framework to trade data quality for more efficient on-chip communication. The proposed framework decreases the size of transmitted data to reduce the consumption of time and power by on-chip communication. The proposed framework incorporates a hardware-software co-design to decrease network power consumption and latency while ensuring a necessary result quality. On the software side, the framework automatically identifies error-resilient variables in the application and calculates the error tolerance for the variables. On the hardware side, we augment the conventional network interface with lossy compression/decompression modules to reduce packet size. As a result, the proposed framework significantly reduces the network latency and dynamic power consumption compared to conventional NoC designs while meeting the application’s requirements of result quality.

As user data grows exponentially in the future, this groundbreaking invention can improve communication efficiency for a wide variety of parallel computation platforms, ranging from the IoT device, chip multiprocessor (CMP), to data centers with enormous commercial potential.
Applications:

1. Parallel Computing Systems (e.g., multi-core CPU, GPU, TPU, etc.)


Advantages:

1. Approximate communication is a novel technique that trades communication quality for better network performance.

2. Approximate communication framework eliminates unnecessary data communication while meeting the application’s requirements of result quality.

3. The performance evaluation shows that our proposed framework reduces network latency and dynamic power consumption by 62% and 43%, respectively, when compared to the conventional NoC design.

Inventors

Prof. Ahmed Louri

Yuechen Chen