Thu, Aug 13, 2020 at 1:56 PM

To: tgeng@bu.edu

Hi, Tong,

I am a researcher from XXXX. I really appreciate your work "AWB-GCN: A Graph Convolutional Network Accelerator with Runtime Workload Rebalancing", which achieves superior performance. The method you use is also interesting. May I ask a small question about this paper?

In the Table 3: The execution time for Cora is 2.3E-3 ms. What is included in this execution time? If I assume the X1 and A are 32-bit floating-point with the CSC format, and I assume the external memory bandwidth is 77GB/s as your paper says, even loading A and X1 from the external memory will have more time than 2.3E-3 ms.

The paper says the execution time is end-to-end including the loading data from DDR. I am a little confused about it. Could you mind giving me a tip about how you measure the performance.

Thanks,

XXXX

---

Tong Geng <tgeng@bu.edu> Mon, Aug 17, 2020 at 12:26 PM

To:

Dear XXXX,

Thank you for your email and interest in our work.

As mentioned in our paper, we implemented scratchpad memories on the device which hold a part of the input matrices on-chip to further reduce the bandwidth requirements. These data can be preloaded on the FPGA, and therefore our measurement does not include the preload overhead. The FPGA has hundreds of Mb on-chip memory.

My collaborators are revising the paper. I will let them know your concern and make this more clear.

Best regards,

Tony

---

Mon, Aug 17, 2020 at 1:12 PM

To: Tong Geng <tgeng@bu.edu>

Hi, Tong,

Thanks for your reply. Wish you good health and recover soon!

Thanks for providing the detailed information about the performance measurement. I am clear about it now.

Best,

XXXX