LSU EE 7722 **Homework 3** **Due: 19 April 2021**

For the week of 19 April 2021 read Yang ASPLOS 20 (Interstellar) [5] and Jouppi CACM 20 (TPU v2 and v3) [2]. These papers are freely available when accessed from on campus. E-mail me if you have any problems obtaining the papers.

There are no questions to be answered for this assignment. Questions will be assigned in an upcoming assignment, which possibly may be called Homework 3 too.

Yang ASPLOS 20 [5] describes an exploration of possible hardware designs to implement convolutional and fully connected neural network layers. In class we talked about an earlier design study, Eyeress [1], which also reported on a design-space exploration, but one which focused on what they called the row-stationary dataflow. Chen, with Parashar as the lead author report on a broader exploration in their Timeloop paper, Parashar ISCA 19 [4]. For this assignment concentrate on the Yang paper and look at the other papers for background.

Jouppi 20 CACM, [2] describes Google's TPU v2 and v3. As mentioned in class, the TPU uses one or two large systolic arrays. Google's first TPU, described in an earlier paper [3], was intended only for inference and since it was something of a first effort, avoided complex and unimportant features. TPU v2 and v3 took advantage of this experience, but also was designed for training workloads, and so required true FP MAC units and more flexible programmability.

References:

[1] Chen, Y.-H., Emer, J., and Sze, V. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. In *Proceedings of the 43rd International Symposium on Computer Architecture* (Piscataway, NJ, USA, 2016), ISCA '16, IEEE Press, pp. 367–379. https://doi.org/10.1109/ISCA.2016.40.

[2] Jouppi, N. P., Yoon, D. H., Kurian, G., Li, S., Patil, N., Laudon, J., Young, C., and Patterson, D. A domain-specific supercomputer for training deep neural networks. *Commun. ACM 63*, 7 (June 2020), 6778. https://doi.org/10.1145/3360307.

[3] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., Boyle, R., Cantin, P.-l., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., Ghaemmaghami, T. V., Gottipati, R., Gulland, W., Hagmann, R., Ho, C. R., Hogberg, D., Hu, J., Hundt, R., Hurt, D., Ibarz, J., Jaffey, A., Jaworski, A., Kaplan, A., Khaitan, H., Killebrew, D., Koch, A., Kumar, N., Lacy, S., Laudon, J., Law, J., Le, D., Leary, C., Liu, Z., Lucke, K., Lundin, A., MacKean, G., Maggiore, A., Mahony, M., Miller, K., Nagarajan, R., Narayanaswami, R., Ni, R., Nix, K., Norrie, T., Omernick, M., Penukonda, N., Phelps, A., Ross, J., Ross, M., Salek, A., Samadiani, E., Severn, C., Sizikov, G., Snelham, M., Souter, J., Steinberg, D., Swing, A., Tan, M., Thorson, G., Tian, B., Toma, H., Tuttle, E., Vasudevan, V., Walter, R., Wang, W., Wilcox, E., and Yoon, D. H. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture* (New York, NY, USA, 2017), ISCA '17, ACM, pp. 1–12. http://doi.acm.org/10.1145/3079856.3080246.

[4] Parashar, A., Raina, P., Shao, Y. S., Chen, Y., Ying, V. A., Mukkara, A., Venkatesan, R., Khailany, B., Keckler, S. W., and Emer, J. Timeloop: A systematic approach to DNN accelerator evaluation. In *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (2019), pp. 304–315. https://ieeexplore.ieee.org/abstract/document/8695666.

[5] Yang, X., Gao, M., Liu, Q., Setter, J., Pu, J., Nayak, A., Bell, S., Cao, K., Ha, H., Raina, P., Kozyrakis, C., and Horowitz, M. Interstellar: Using Halide's scheduling language to analyze DNN accelerators. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems* (New York, NY, USA, 2020), ASPLOS 20, Association for Computing Machinery, p. 369383. https://doi.org/10.1145/3373376.3378514.