

Note: late submissions accepted without penalty until 5 May.

In class we spent much of the semester showing how recent generation GPUs can be used for matrix/matrix multiplication, though understanding that they were designed for a larger class of computations. In this assignment we will look at some accelerators designed specifically for machine learning workloads that are matrix/matrix multiplication. These include Google's TPU chips (which are used in production) and an idea proposed in the research literature, *Scale-Out Systolic Arrays (SOSA)*. Both designs use systolic arrays, of course. As mentioned in class, the computation of matrix/matrix multiplies using systolic arrays requires far less register reads and writes than the same computation performed on a device that reads operands for each FMADD instruction from a register file.

For this assignment, and to prepare for the final exam read the following papers. First, read the papers describing Google's TPU accelerators, all of which use large systolic arrays. The first paper, Jouppi 17 [3], describes Google's first TPU, now called TPUv1. Though TPUv1 is primitive compared to later designs, the paper does provide more detail, and so will be the basis of some questions in this assignment. The TPUv1 systolic array was limited to integer arithmetic, limiting its use to inference. Google's next GPUs TPUv2 and TPUv3 were designed to handle both inference and training. To support training these chips use systolic arrays that operate on 16-bit floating-point data, the BF16 (brain-float) format. These are described in several papers, see [2, 4]. TPUv4i, designed just for inference, is described in a fourth paper [1].

Also read the SOSA paper, [5], in which a design using smaller systolic arrays is proposed. The paper shows how the lower energy efficiency of the smaller systolic arrays can be overcome by better utilization when the systolic arrays are appropriately interconnected and a more meaningful metric, according to the paper, is used.

You should be able to get copies of all of these papers for free on campus. Off campus you might be asked to pay. Please E-mail me if you have any problems getting a free copy.

Problem 1: Answer the following questions about Google's TPUs and described by Norman Jouppi and his many collaborators. Several of the questions below are based on the TPUv1 paper [3]. This paper shows data bandwidth in units of GiB/s. Be aware that $1 \text{ GiB} = 2^{30} \text{ B}$. Some of the questions below ask about the number of clock cycles or to reason about clock cycles with other information provided in GiB/s. To answer such questions one needs to use the clock frequency, which for TPUv1 is $\phi = 0.7 \text{ GHz}$.

(a) In TPUv1 how many clock cycles would it take to load one set of weights (one tile) from the off-chip memory (DDR3 DRAM Chips) into the Weight FIFO? Helpful information is in Figure 1 in [3].

(b) The paper explains on page 3 that it takes 256 cycles to load the weight matrix. (That's not the answer to the subproblem above.) Suppose the array were 128×512 . How long would it take to load the weight matrix from the FIFO?

(c) Figure 1 shows a bandwidth of 167 GiB/s between the Unified Buffer and the MXU, but off-chip bandwidth for activations is only 10 GiB/s. Why the imbalance? That is, how is it possible that enough data can be provided to the MXU from off chip to keep it fed at 256 bytes per cycle?

(d) To compute the large layer on Homework we need to multiply a 1536×512 matrix by a 512×2970 matrix. Roughly how long would that take on TPUv3 (not TPUv1 this time)? Explain any assumptions.

Problem 2: Yüzügüler *et al* describe an accelerator in which *pods* containing a 32×32 systolic array are interconnected by carefully chosen networks.

(a) The paper notes that the utilization of the systolic arrays depends on both the bisection bandwidth and the latency of the network connecting the pods to storage for weights, activations, and partial sums. Suppose that each SM in an Nvidia H100 had a pod with a 32×32 systolic array, along with the skewing and de-skewing hardware. The systolic arrays would read and write data from the H100 L2 cache, and perhaps make use of the L1. What would the utilization of the systolic arrays in this setup be based on the

H100 latency and throughputs shown below collected using the microbenchmarks described in a previous homework. Be sure to state assumptions made in computing your answer.

```
GPU 0: NVIDIA H100 PCIe @ 1.75 GHz WITH 81089 MiB GLOBAL MEM
GPU 0: L2: 51200 kiB MEM<->L2: 2039.0 GB/s
GPU 0: CC: 9.0 SM: 114 SP-FP32/SM: 128 DP-FP64/SM: 64 TH/BL: 1024
GPU 0: SHARED: 49152 B/BL 233472 B/SM CONST: 65536 B # REGS: 65536
GPU 0: PEAK: 25609 SP GFLOPS 12804 DP GFLOPS COMP/COMM: 50.2 SP 50.2 DP
```

```

          --Insn--  -----L1 Per SM-----  -L2 GB/s-- DRAM
Lv Blk V SPL      s /add % SW TW  BXW B/cyc  GB/s   /SM /GPU GB/s
ME 114 1  s0     280 2.9 5   4 1  0.0 10.9 17.8 17.8 2025 2023
L2 114 1  s0     503 2.8 18  4 1  0.0 48.6 66.9 66.9 7629  78
L1 114 2  s0     464 1.9 46  8 2  0.0 133.1 225.9  0.1   8   8
L1 114 1  s0     467 2.5 61  4 1  0.0 131.9 224.7  0.1   8   8
L2 114 1  s0     507 2.8 18  4 1  0.0 48.2 66.5 66.5 7581  77
L2  57 1  s0     369 2.8 13  4 1  0.0 54.5 91.3 91.3 5206  99
L2   2 1  s0     561 2.8  0  4 1  0.0 34.7 60.1 60.1  120  60
L2   1 1  s0     513 2.8  0  4 1  0.0 37.6 65.7 65.7   66  66

```

Kernel mb_g:

```

---Data Touched---  -----Latency-----
nbl  iter  Block      Total  ns  cyc  !<-----500 ns----->!
512  20480  512 kiB    256 MiB  394  692 *****
512  10240  256 kiB    128 MiB  396  695 *****
256  10240  256 kiB     64 MiB  391  686 *****
128  10240  256 kiB     32 MiB  147  257 *****
 64  10240  256 kiB     16 MiB  139  244 *****
 32  10240  256 kiB      8 MiB  139  244 *****
 16  10240  256 kiB      4 MiB  144  252 *****
  8  10240  256 kiB      2 MiB  144  252 *****
  4  10240  256 kiB      1 MiB  143  251 *****
  2  10240  256 kiB    512 kiB  150  263 *****
  1  10240  256 kiB    256 kiB  150  262 *****
  1  10000  128 kiB    128 kiB   37   65 **
  1  10000   64 kiB    64 kiB   28   50 **
  1  10000   32 kiB    32 kiB   24   42 *
  1  10000   16 kiB    16 kiB   22   38 *
  1  10000    8 kiB     8 kiB   21   36 *

```

References:

- [1] Jouppi, N. P., Hyun Yoon, D., Ashcraft, M., Gottscho, M., Jablin, T. B., Kurian, G., Laudon, J., Li, S., Ma, P., Ma, X., Norrie, T., Patil, N., Prasad, S., Young, C., Zhou, Z., and Patterson, D. Ten lessons from three generations shaped googles TPUv4i : Industrial product. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)* (2021), pp. 1–14. <http://dx.doi.org/10.1109/ISCA52012.2021.00010>.
- [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., Ghaemmghami, 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] Norrie, T., Patil, N., Yoon, D. H., Kurian, G., Li, S., Laudon, J., Young, C., Jouppi, N., and Patterson, D. The design process for google’s training chips: TPUv2 and TPUv3. *IEEE Micro* 41, 2 (2021), 56–63. <http://dx.doi.org/10.1109/MM.2021.3058217>.
- [5] Yüzügüler, A. C., Sönmez, C., Drumond, M., Oh, Y., Falsafi, B., and Frossard, P. Scale-out systolic arrays. *ACM Trans. Archit. Code Optim.* 20, 2 (mar 2023). <https://doi.org/10.1145/3572917>.