Using Tensor Processing Units for Neural Networks

Jonas Engicht

Department of Psychological Methodology,
Friedrich-Schiller University Jena

ai.uni-jena.de

SUMMER SCHOOL 2023

INTRODUCTION

The Tensor Processing Unit (TPU) is a custom hardware by Google that has been developed in response to the growing demand for more computational power to run large neural networks efficiently.

This poster takes an in-depth look at the cutting-edge world of TPUs and their transformative impact on AI technologies. Is also presents different methods to access TPUs.

TPU VERSION 1

GENERAL SPECIFICATIONS
- Inference only
- Connection via SATA slot with host system

COMPUTATIONAL RESOURCES
- Matrix Multiply Unit (MXU)
- Weight FIFO
- Activation Unit

QUANTIZATION
- 8-bit Int for matrix multiplication
- Approximation of arbitrary values between minimum and maximum

INSTRUCTION SET
- CISC

PARALLEL PROCESSING
- 4 stages of instructional execution
- Matrix multiplication can be performed in parallel with other tasks

PERFORMANCE
- Up to 92 TIOPS

LIMITATIONS
- Energy proportionality
- Slow data transfer rates
- Thermal design power high
- Inference only

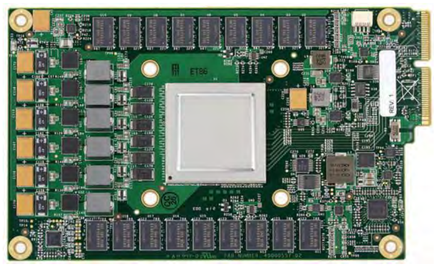

Figure 1 *TPUv1 Printed Circuit Board. [5]*

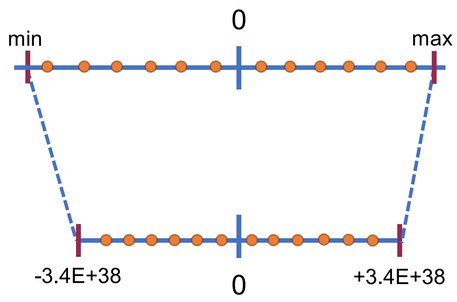

Figure 2 *Representation of quantization with float32 values. Adapted from [6].*

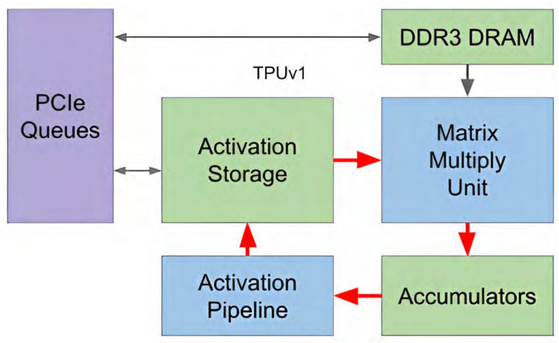

Figure 3 *TPU Block Diagram. [4]*

SYSTOLIC ARRAYS

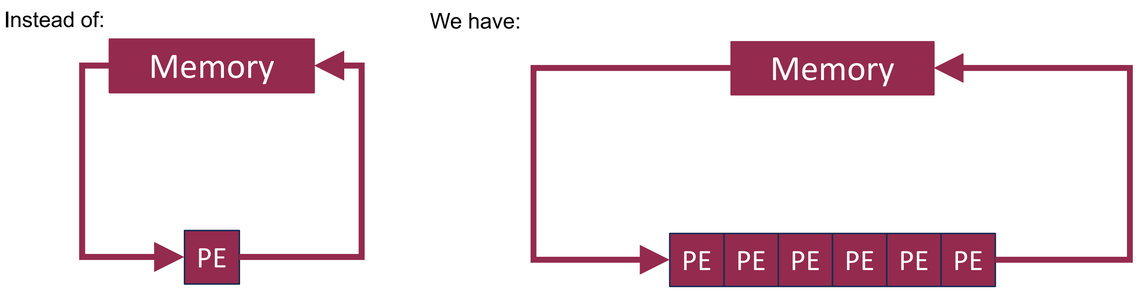

Figure 4 *Illustration of a systolic array in comparison to a non-systolic architecture. Adapted from [3].*

- Data is read from memory once, used multiple times
- Possible Designs:
  - Two-dimensional
  - Triangular
  - Rectangular
  - Etc.
- Data moves with fixed rhythm
- Significantly higher throughput

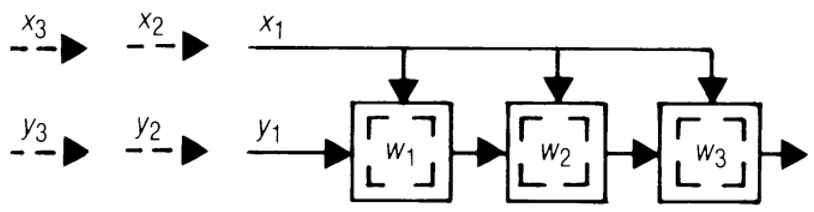

Figure 5 *Systolic array with cells. Input (x_i) is broadcasted to the cells. Weights (w_i) stay. Results (y_i) move systolically. [3]*

RECENT TPU VERSIONS

TPU v2

- Neural network training
- Bfloat16 for matrix multiplications
- Using VLIWs (Multiple operations invoked and executed at the same time)
- High Bandwidth Memory (HBM)
- Tensor Core: Scalar Unit + Vector Unit + MXU

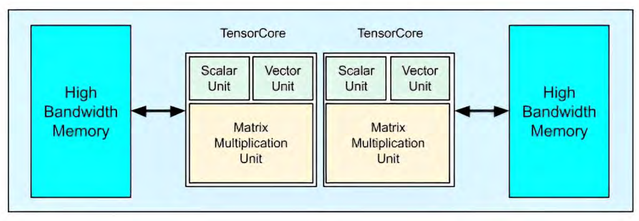

Figure 6 *Illustration of a TPUv2 chip. [2]*

TPU v3

- 2 MXUs per Tensor Core
- 2× more HBM Capacity
- Higher memory bandwidth
- Water cooling

TPU v4

- 4 MXUs per Tensor Core
- Bfloat16 and int8 mode
- Higher memory bandwidth
- Compiler compatible with previous generations

USING TPUS

- TPU v2, v3, v4 available through Google Cloud Platform (Cloud TPUs)
- Usable frameworks: i.a. TensorFlow, PyTorch, JAC

TPU VM Architecture

- SSH connection to TPU Hosts
- Direct access to compiler, logs, and errors
- Execution of arbitrary code
- *libtpu*: compiler, drivers, and software for execution

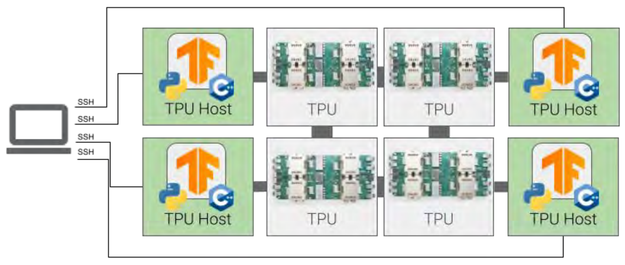

Figure 7 *Illustration of the TPU VM architecture. TPU Hosts are accessed directly via SSH. [2]*

TPU Node Architecture

- Direct access to User VM
- TPU Hosts running TensorFlow Server
- Communication via gRPC connection
- Conversion of framework-specific code is needed

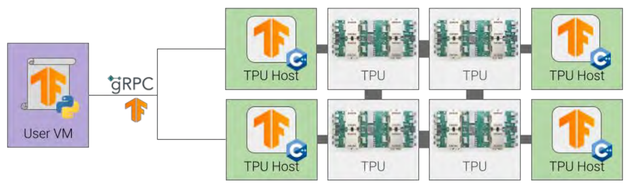

Figure 8 *Illustration of the TPU Node architecture. No direct access to the TPU Hosts is possible. [2]*

CONCLUSION

- TPUs are ASICs developed by Google for accelerating machine learning workloads.
- The TPUv1 was optimized for inference.
  - Uses quantization, a high-level instruction set and several computational resources that were specifically designed for matrix-multiplications.
- The TPU v2 can train neural networks by using bfloat16 floating-point numbers, redesigned hardware components and VLIWs
- The third and the fourth generation of TPUs are improved versions of the second TPU generation.
- Only TPU generations two, three, four are publicly available and can be accessed via Google Cloud Platform through either the TPU VM architecture or the TPU Node architecture.

REFERENCES

[1] Google Cloud, “The bfloat16 numerical format | Cloud TPU,” Google Cloud, Nov. 18, 2022. https://cloud.google.com/tpu/docs/bfloat16 (accessed Nov. 25, 2022).

[2] Google Cloud, “System Architecture | Cloud TPU | Google Cloud,” Jan. 13, 2023. https://cloud.google.com/tpu/docs/system-architecture-tpu-vm (accessed Jan. 15, 2023).

[3] H.-T. Kung, “Why systolic architectures?,” Computer, vol. 15, no. 01, pp. 37–46, 1982.

[4] N. P. Jouppi et al., “Ten Lessons From Three Generations Shaped Google’s TPUv4i : Industrial Product,” 2021, pp. 1–14. doi: 10.1109/ISCA52012.2021.00010.

[5] N. P. Jouppi et al., “In-datacenter performance analysis of a tensor processing unit,” in 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), 2017, pp. 1–12. doi: 10.1145/3079856.3080246.

[6] M. Vandersteegen, “Aspects and best practices of quantization aware training for custom network accelerators,” Apr. 2020. [Online]. Available: https://iiw.kuleuven.be/onderzoek/eavise/starttodeeplearn/aspects-and-best-practices-of-quantization-aware-t.pdf

CONTACT

Email: jonas.engicht@uni-.jena.de