Ripunjay Narula (19BCE0470)

Check Yourself 2

Approximate Relative Access Time

Semiconductor memory also has much faster access times than other types of data storage; a byte of data can be written to or read from semiconductor memory within a few nanoseconds, while access time for rotating storage such as hard disks is in the range of milliseconds(9-15).

Volatility

DRAM is volatile whereas Disk Memory and Flash Storage are not.

Cost

Flash memory has a higher cost per bit than hard drives whereas DRAM is cheaper than them.

MIPS:

Stands for "Million Instructions Per Second." It is a method of measuring the raw speed of a computer's processor. Since the MIPS measurement doesn't take into account other factors such as the computer's I/O speed or processor architecture, it isn't always a fair way to measure the performance of a computer. For example, a computer rated at 100 MIPS may be able to computer certain functions faster than another computer rated at 120 MIPS.
MIPS is a RISC-type, Load/Store instruction set. The early implementations like MIPS 1 and MIPS 2 were 32 bits while MIPS 3, 4 and 5 are 64 bits.

MFLOPS:

MFLOP. Short for mega floating-point operations per second, **MFLOPs** are a common measure of the speed of computers used to perform floating-point calculations. Another common measure of computer speed and power is MIPS (million instructions per second), which indicates integer performance.
A floating-point operation is an addition, subtraction, multiplication, or division operation applied to a number in a single or double precision floatingpoint representation. Such data items are heavily used in scientific calculations and are specified in programming languages using key words like float, real, double, or double precision.

NUMERICALS ON CPU EXECUTION TIME

Q1. CPU clock rate is 1 MHz. Program takes 5 million cycles to execute. What is the CPU time?

=> 5,000,000 * (1 / 1,000,000) = 5 seconds

Q2. CPU clock rate is 500 MHz. Program takes 45 million cycles to execute. What is the CPU time?

=> 45,000,000 * (1 / 500,000,000) = 0.09 seconds

Q3. **Let assume that a benchmark has 100 instructions:**
35 instructions are loads/stores (each take 2 cycles)
75 instructions are adds (each takes 1 cycle)
20 instructions are square root (each takes 50 cycles)
What is the CPI for this benchmark?

=> CPI = ((0.35 * 2) + (0.75 * 1) + (0.20 * 50)) = 11.45

INTERNAL ARCHITECTURE OF GOOGLE TENSOR PROCESSING UNIT

In a Google data center, TPU devices are available in the following configurations for both TPU v2 and TPU v3:

- Single device TPUs, which are individual TPU devices that are not connected to each other over a dedicated high-speed network. You cannot combine multiple single device TPU types to collaborate on a single workload.
- TPU Pods, which are clusters of TPU devices that are connected to each other over dedicated high-speed networks.

Single-device TPUs

A single-device TPU configuration in a Google data center is one TPU device with no dedicated high-speed network connections to other TPU devices. Your TPU node connects only to this single device.

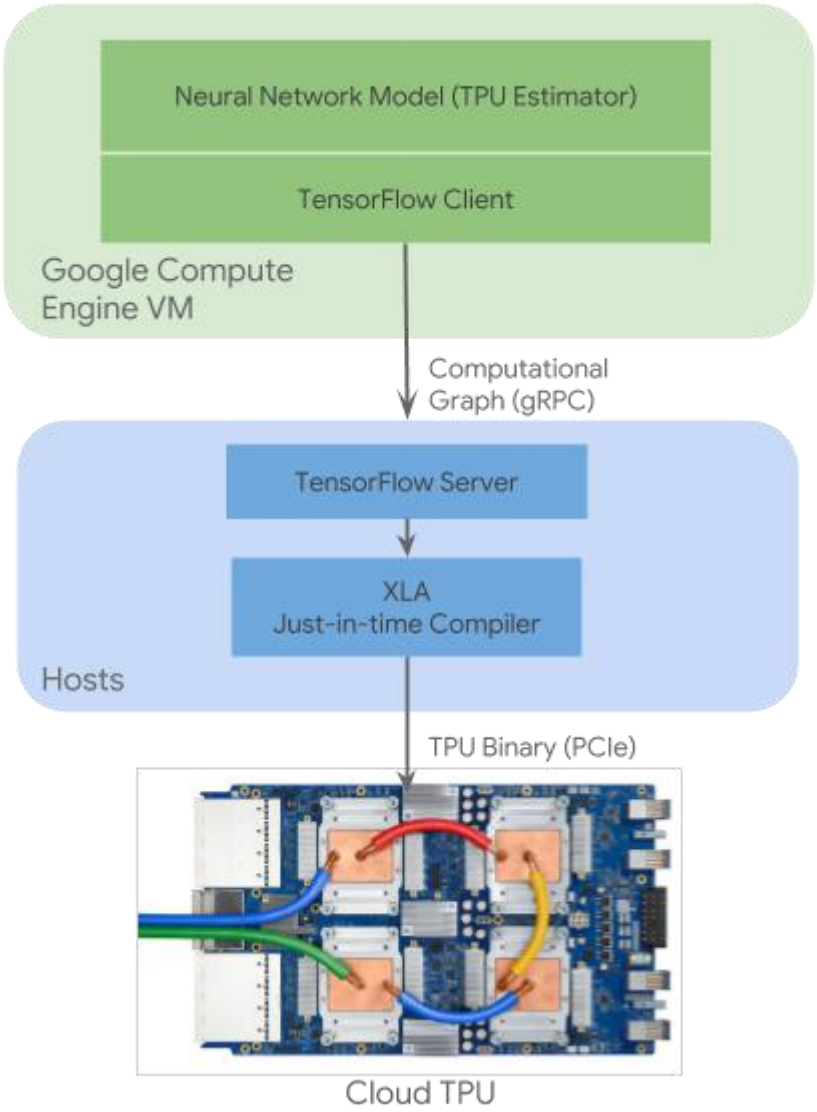
For single-device TPUs the chips are interconnected on the device so that communication between chips does not require host CPU or host networking resources.

TPU Pods

A TPU pod configuration in a Google data center has multiple TPU devices connected to each other over a dedicated high-speed network connection. The hosts in your TPU node distribute your machine learning workloads across all of the TPU devices.

In a TPU Pod, the TPU chips are interconnected on the device so that communication between chips does not require host CPU or host networking resources. Additionally, each of the TPU devices in a TPU Pod are connected to each other over dedicated high-speed networks that also do not require host CPU or host networking resources.

Software architecture



TPU estimator

TPU Estimators are a set of high-level APIs that build upon Estimators which simplify building models for Cloud TPU and which extract maximum TPU performance. When writing a neural network model that uses Cloud TPU, you should use the TPU Estimator APIs.

TensorFlow client

TPU Estimators translate your programs into TensorFlow operations, which are then converted into a computational graph by a TensorFlow client. A TensorFlow client communicates the computational graph to a TensorFlow server.

TensorFlow server

A TensorFlow server runs on a Cloud TPU server. When the server receives a computational graph from the TensorFlow client, the server performs the following actions:

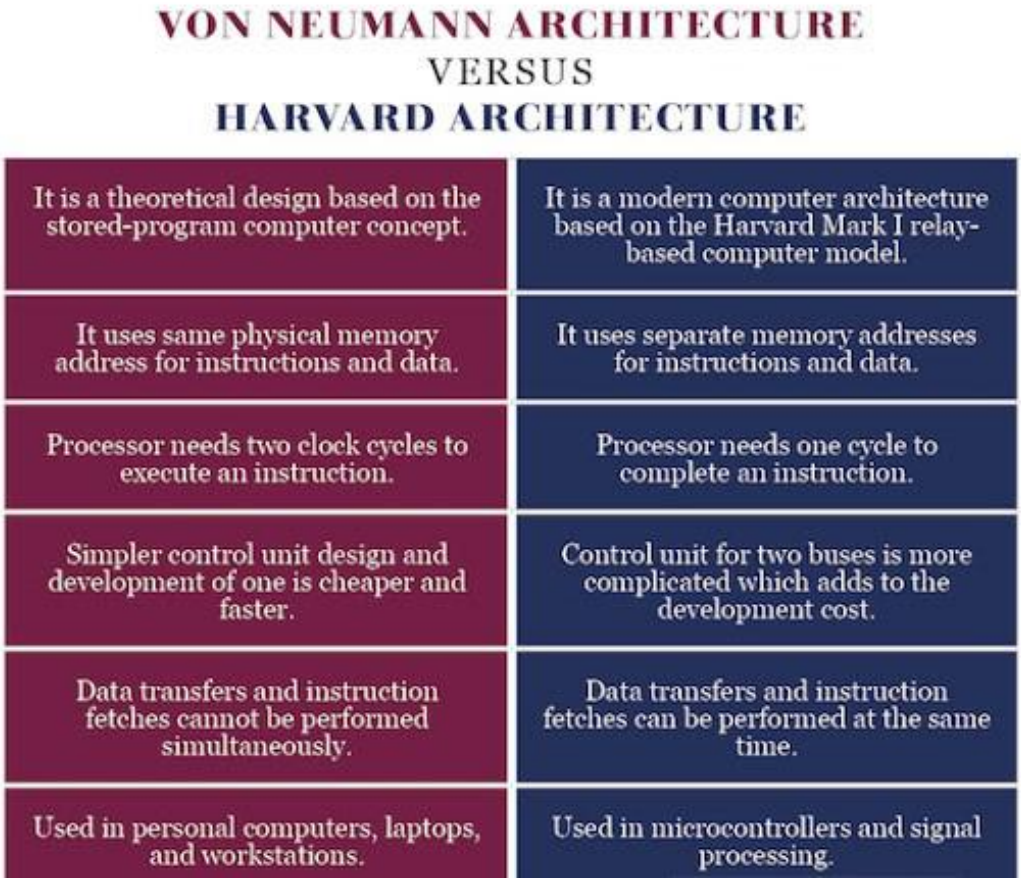
1. Load inputs from Cloud Storage
2. Partition the graph into portions that can run on a Cloud TPU and those that must run on a CPU
3. Generate XLA operations corresponding to the sub-graph that is to run on Cloud TPU
4. Invoke the XLA compiler

XLA compiler

XLA is a just-in-time compiler that takes as input High Level Optimizer (HLO) operations that are produced by the TensorFlow server. XLA generates binary code to be run on Cloud TPU, including orchestration of data from on-chip memory to hardware execution units and inter-chip communication. The generated binary is loaded onto Cloud TPU using PCIe connectivity between the Cloud TPU server and the Cloud TPU and is then launched for execution.

Von Neumann vs Harvard Architecture

There is no relations between Instruction Set (RISC and CISC) with architecture of the processor (Harvard Architecture and Von Neumann Architecture). Both instruction set can be used with any of the architecture.



CISC vs RISC


RISC VERSUS CISC
RISC
CISC
An instruction set architecture that is designed to perform a smaller number of computer instructions so that it can operate at a higher speed
A full set of computer instructions that intends to provide the necessary capabilities in an efficient way
Stands for Reduced Instruction Set Computer
Stands for Complex Instruction Set Computer
Utilizes a small, highly optimized set of instructions
Utilizes a large, specialized and a complex set of instructions
More machine oriented
More programmer oriented
Simple and requires one clock cycle to execute instructions
Complex and requires multiple clock cycles to execute an instruction
More registers
Fewer registers
Instructions have simple, fixed formats with few addressing modes
Instructions have variable formats with several complex addressing modes
Has simple instructions - the program length is long
Has complex instructions - the program length is short
Requires more RAM
Requires a minimum amount of RAM
Used in Hardwired Control Unit; used in applications such as mobile phones and tablets
Used in Microprogrammed Control Unit; used in applications such as desktop computer and laptops
Visit www.PEDIAA.com