classmate
Date 7.4.2020
Page 1

COMPUTER ARCHITECTURE OF WAREHOUSE SCALE COMPTS.

the total storage capacity

DOMAIN SPECIFIC ARCHITECTURE

① "tensorflow a system for large scale machine arch." usenix nov 2016.

② norman jappy, nishant patil, davaid patterson. motivation for evalution of 1st tensor processing unit.

③ a domain specific arch. for ML apps.

④ jeff , cliff young.
a golden age: empowering the ML revolution.

⑤ norman p jouppy.
anly data center perf. analysis of a TPU.
44th international symposium, jun 2017.

4 DSAs discussed in book
- TPU - catapult - crust by intel.

- pixel visual core intel. 7.4 -- 7.7.

GUIDELINES

- use dedicated memories to minimize distance over which data is moved.

data movement is managed by SW.

- invest in resources

- use easiest form of ||lism that matches the domain.

- reduce the data size and type to simplest needed for the domain.

- use domain specific prog. lang. to port the code to DSA.

5 main instructions. TPU

- read host memory → unified buffer.

- reads weights from weights memory into weight fifo for matrix unit.

- matrix multiply : stroke convolve. multiply. dot. convolution.

- activate
  non-linear functions for NN, relu, sigmoid, tanh.

- write host memory

TPU, 28 nm, 700 MHz,
1/3rd space = unified buffers.
1/4 space = matrix multiply unit.