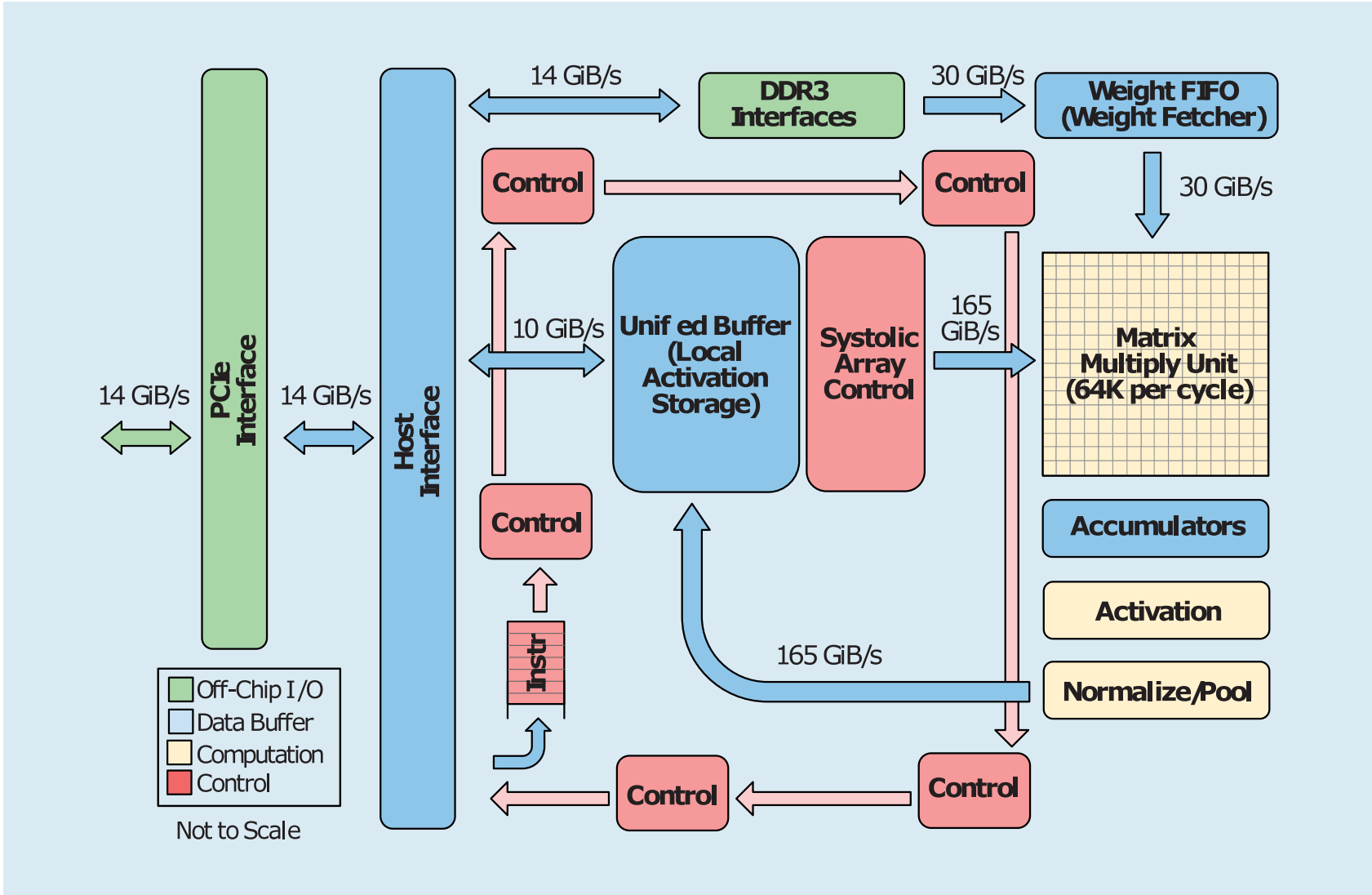
Arhitectura Calculatoarelor

Opriţoiu Flavius
flavius.opritoiu@cs.upt.ro

30 Septembrie, 2020

Studiu de caz: Tensor Processing Unit (TPU)



1

[1] J. Hennessy, D. Patterson 2019: "A new golden age for computer architecture", [HePa19]

Studiu de caz: TPU (contin.)

Arhitectura Google a TPU:

- descrisă într-un articol prezentat în cadrul International Symposium on Computer Architecture din 2017
  - una din cele mai anticipate prezentări
  - arhitectura TPU a ajuns, între timp, la a III-a iterație

Originea dispozitivului [JYPP17]:

- în 2006, Google analizeaza bugetul de calcul necesar pentru rularea aplicațiilor speciale, precum Deep Neural Networks (DNNs)
  - la acel moment, acestea pot fi executate, gratis, utilizând HW-ul existent
- în 2013, o proiecție de utilizare a serviciilor Google: 3 minute, zilnic de căutare vocală
  - utilizează aplicații DNNs pentru recunoaștere vocală
  - un astfel de scenariu revendică dublarea necesarului computațional al corporației
    - costisitor dacă se folosesc Central Processing Unit (CPU)-uri

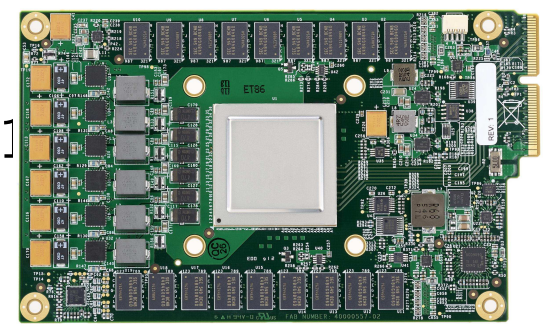
Studiu de caz: TPU (contin.)

TPU:

- design Application Specific Integrated Circuit (ASIC) specializat
  - 256 x 256, 8-bit, Matrix Multiply Unit
  - 28MiB memorie interna
  - 92 Tera Operations per Second (TOpS)
  - de 15/30 ori mai rapid decât CPU/Graphics Processing Unit (GPU)
  - de 30/80 ori mai eficient (TOpS/Watt) decât CPU/GPU

Accelerarea operațiilor unui DNN:

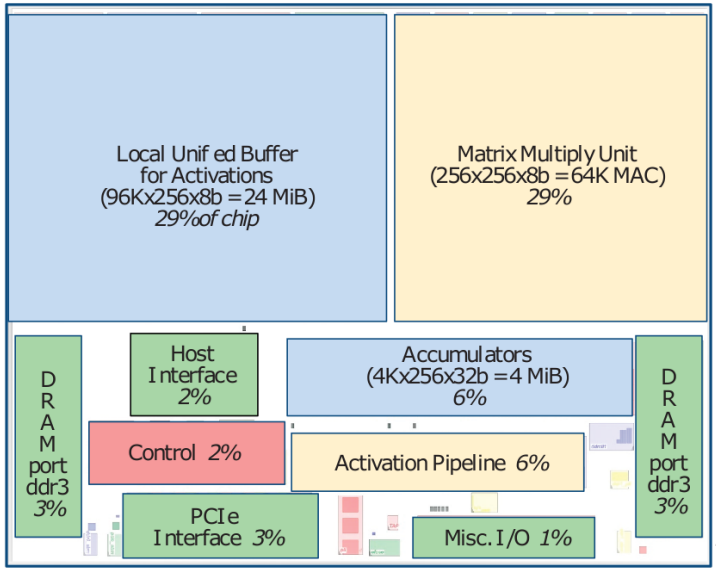
- cuantizare: transformarea numerelor de virgulă flotantă în întregi (tipic, pe 8 biți)
- înmulțirea - întregi 8-bit versus virgulă flotantă 16-bit
  - arie de Si de 6 ori mai mică
  - energie consumată de 6 ori mai mică
- adunarea - întregi 8-bit versus virgulă flotantă 1
  - arie de Si de 38 mai mică
  - energie consumată de 13 ori mai mică



2

[2] Jouppi et al.: "In-Datacenter Performance Analysis of a Tensor Processing Unit" [JYPP17]

Studiu de caz: TPU (contin.)



[3]

Elemente de arhitectură:

- Matrix Multiply Unit
  - elementul central al unității
  - rețea de 65'536, unități de tip înmulțire cu acumulare, pe 8-bit
  - architecturi sistolice pentru accelerarea înmulțirii matricilor
- Adunarea și înmulțirea întregilor și a numerelor de virgulă flotantă sunt în centrul mecanismelor de accelerare ale TPU

[3] Jouppi et al.: "In-Datacenter Performance Analysis of a Tensor Processing Unit" [JYPP17]

Cap. 1 Analiza funcțională și sinteza dispozitivelor de adunare și scădere, binară și zecimală

1.1 - Sumatoare paralele bazate pe propagarea serială a transportului

Ripple Carry Adder (RCA): utilizează celule dedicate pentru fiecare rang

- celula Full Adder Cell (FAC)
- propagarea carry-ului: de la un rang către rangul mai semnificativ

Arhitectură RCA pe n biți:

FAC:

- tabel de adevăr
- ecuații ale ieșirilor

Sinteza FAC:
A. porți de tip EXOR, AND, OR

B. porți de tip EXOR, NAND

C. multiplexoare

D. porți de tip NAND

Dacă $c_0 = 0 \Rightarrow$ cea mai puțin semnificativă FAC se simplifică:

- ecuații ale ieșirilor

Sinteza Half Adder Cell (HAC):
A. porți de tip EXOR, AND

B. porți de tip NOR

C. porți de tip NAND

Calea critică pentru un RCA pe 2 biți

Ipoteze simplificatoare:

- toate porțile primitive au latență de *1d*
  - indiferent de numărul de intrări
- inversoarele nu au întârzieri (*0d*)
- porțile EXOR au latență de *2d*
- toți operanzii sunt disponibili la momentul *0d*

Întârizerea unui segment RCA pe n biți:
$D_{RCA}^{c_{out}} = 2nd$
$D_{RCA}^{z} = 2nd$

Referințe

[HePa19] J. L. Hennessy and D. A. Patterson, "A new golden age for computer architecture," *Commun. ACM*, vol. 62, no. 2, pp. 48–60, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3282307

[JYPP17] N. P. Jouppi, C. Young, N. Patil, D. Patterson *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture*, ser. ISCA '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1–12.