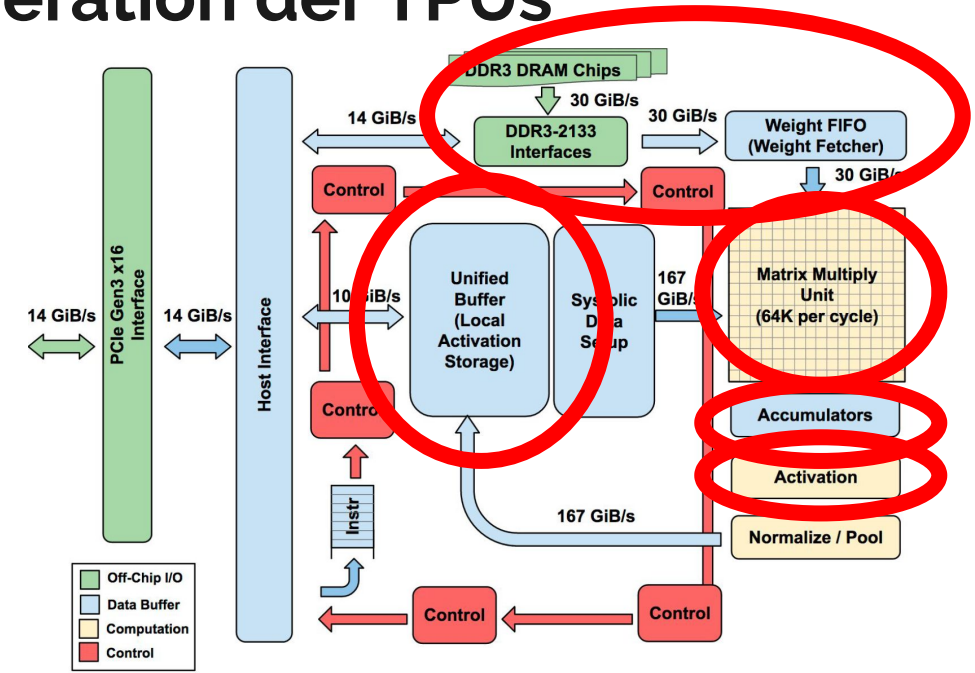
Implementierung einer Tensor Processing Unit

mit dem Fokus auf Embedded Systems und das Internet of Things

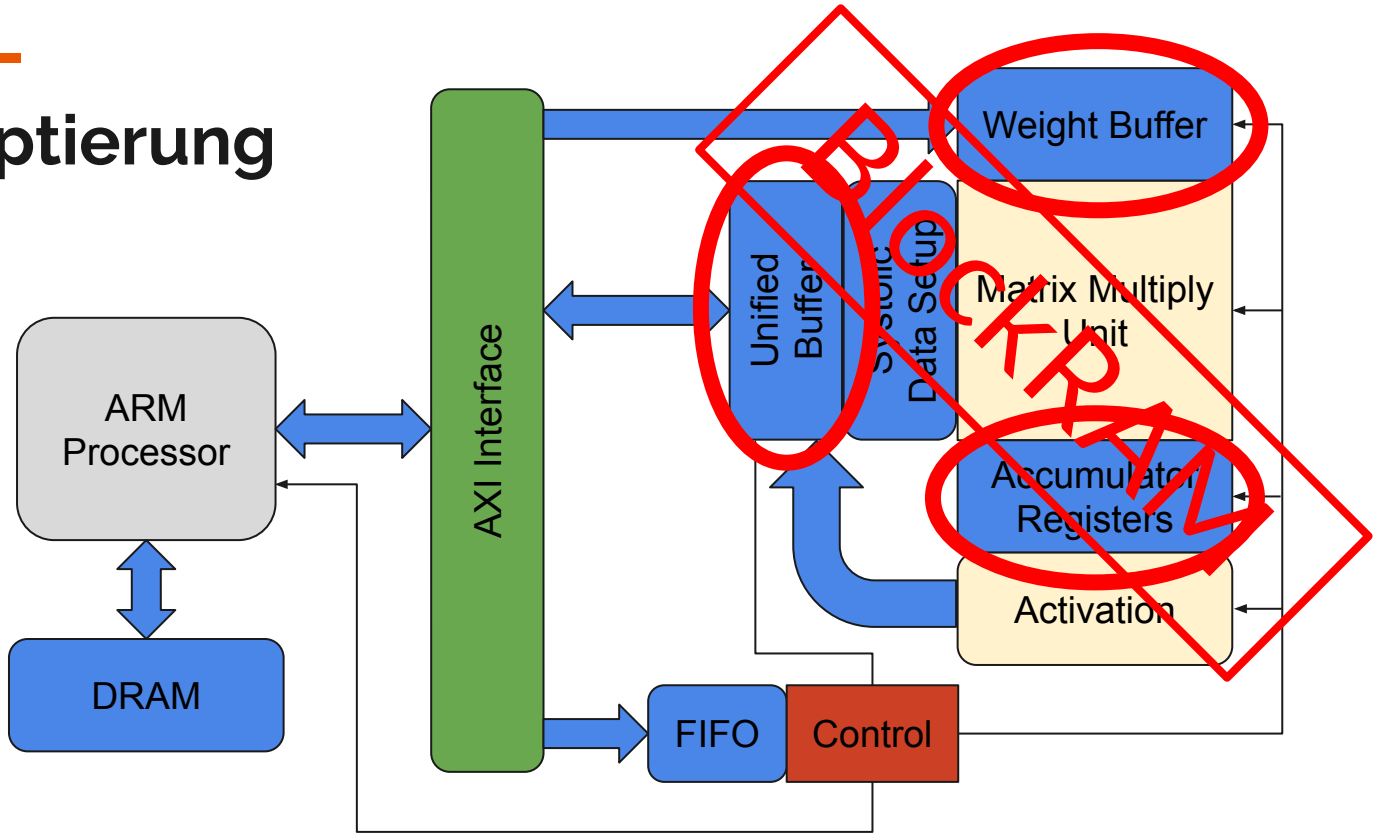
Jonas Fuhrmann
HAW Hamburg
Robert Bosch GmbH

Erste Generation der TPUs

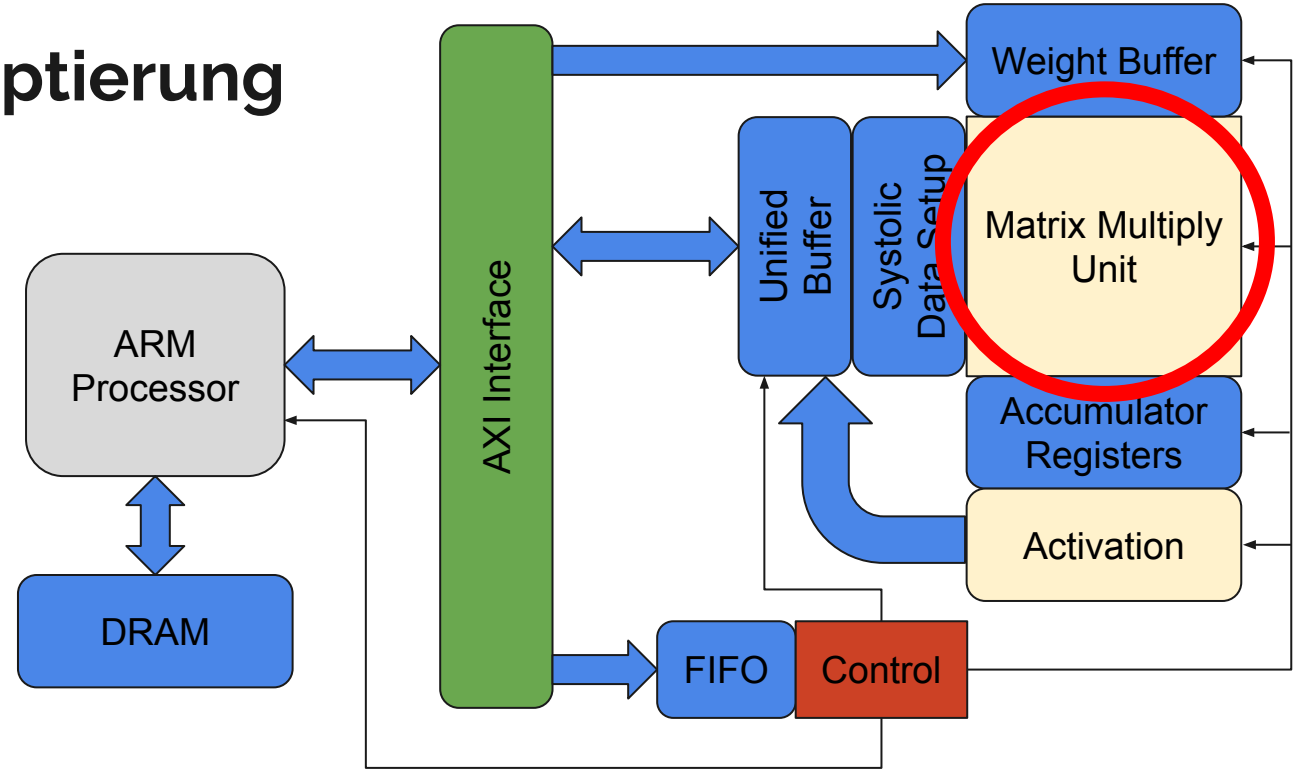


Quelle: Jouppi, Norman P. ; Young, Cliff u. a.: In-Datacenter Performance Analysis of a Tensor Processing Unit / Google, Inc. 2017.

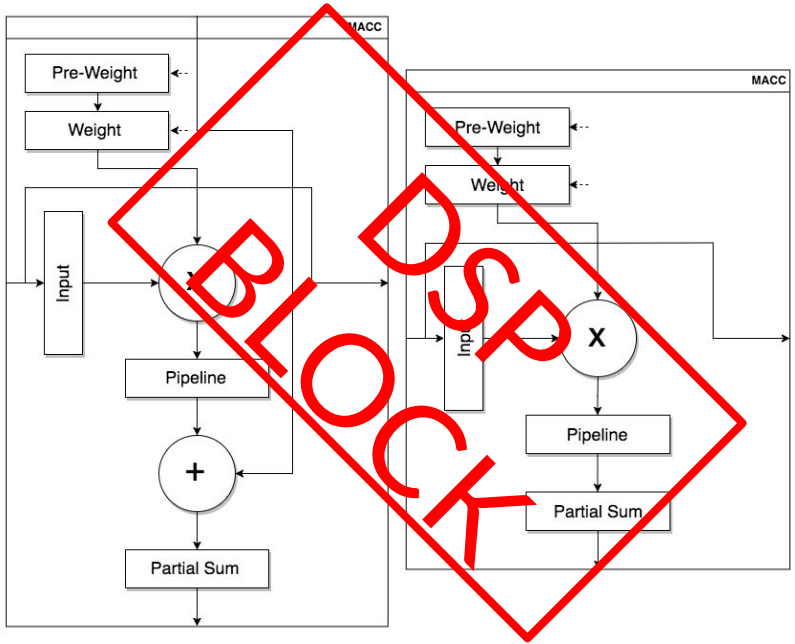
Adaptierung

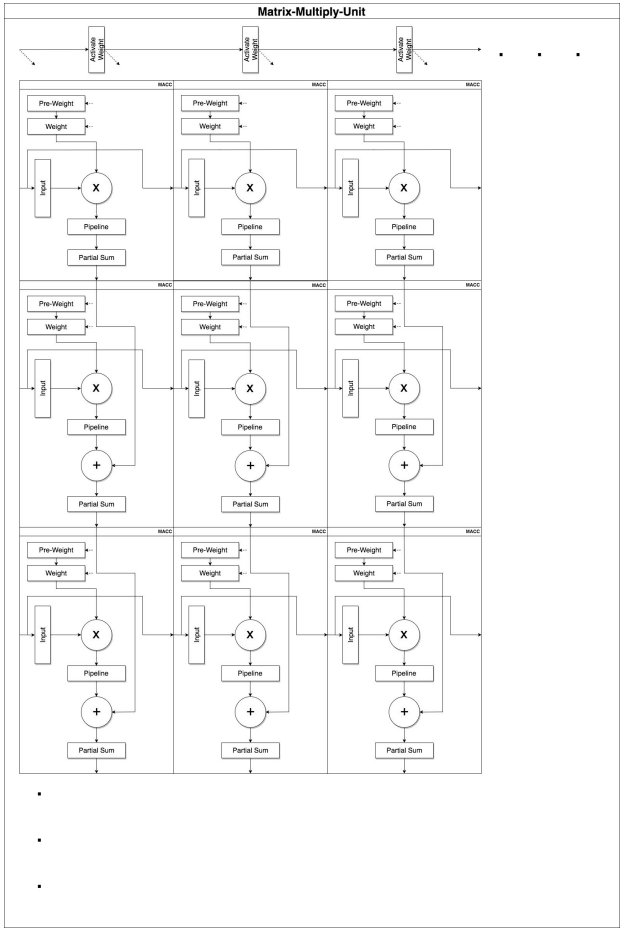


Adaptierung

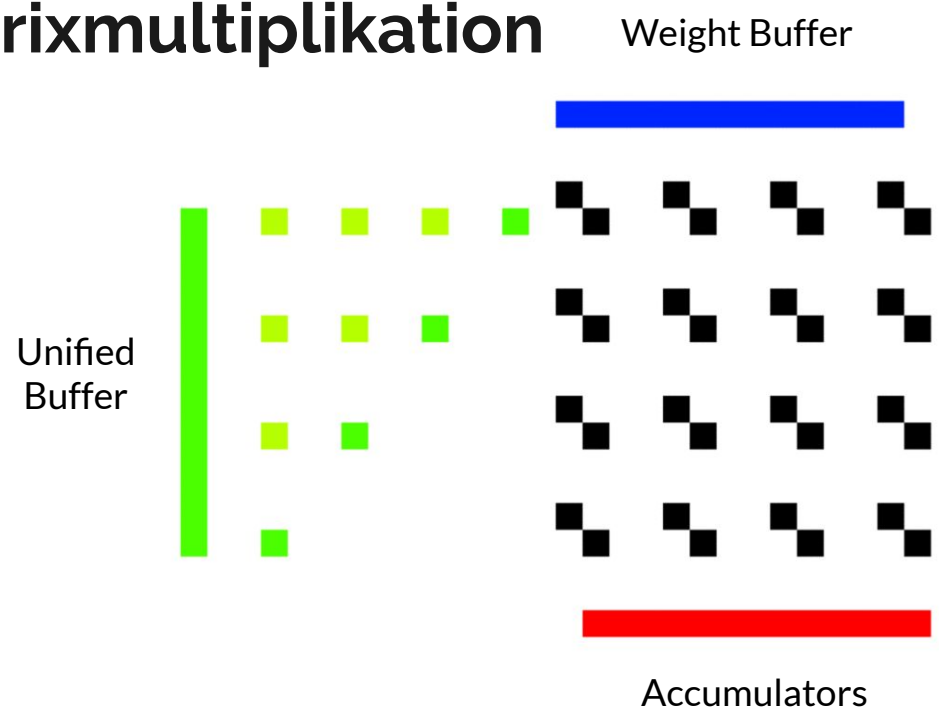


Matrixmultiplikationseinheit

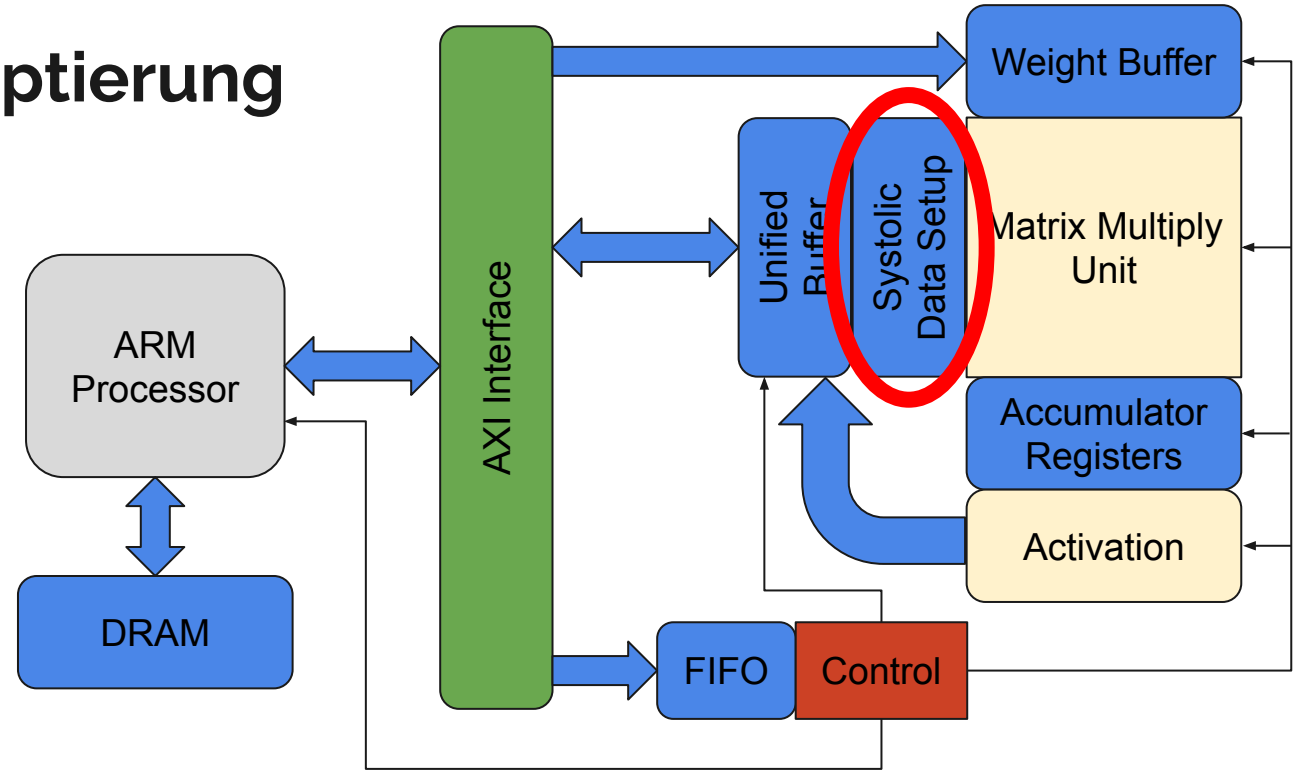




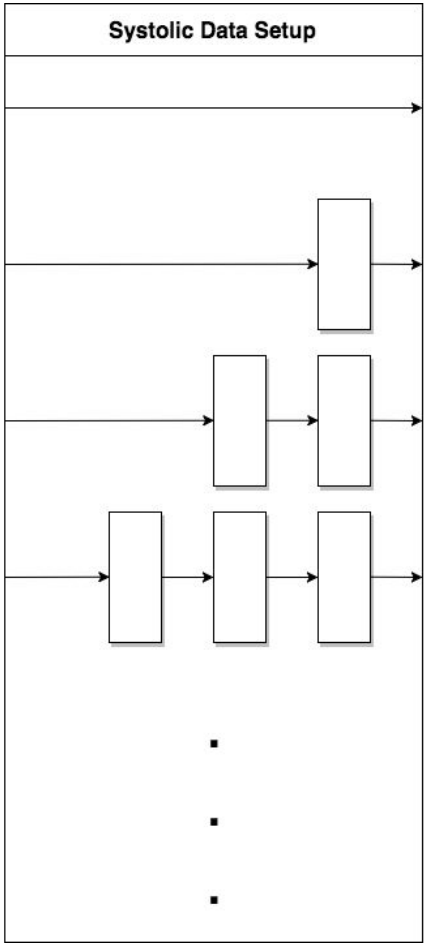
Matrixmultiplikation



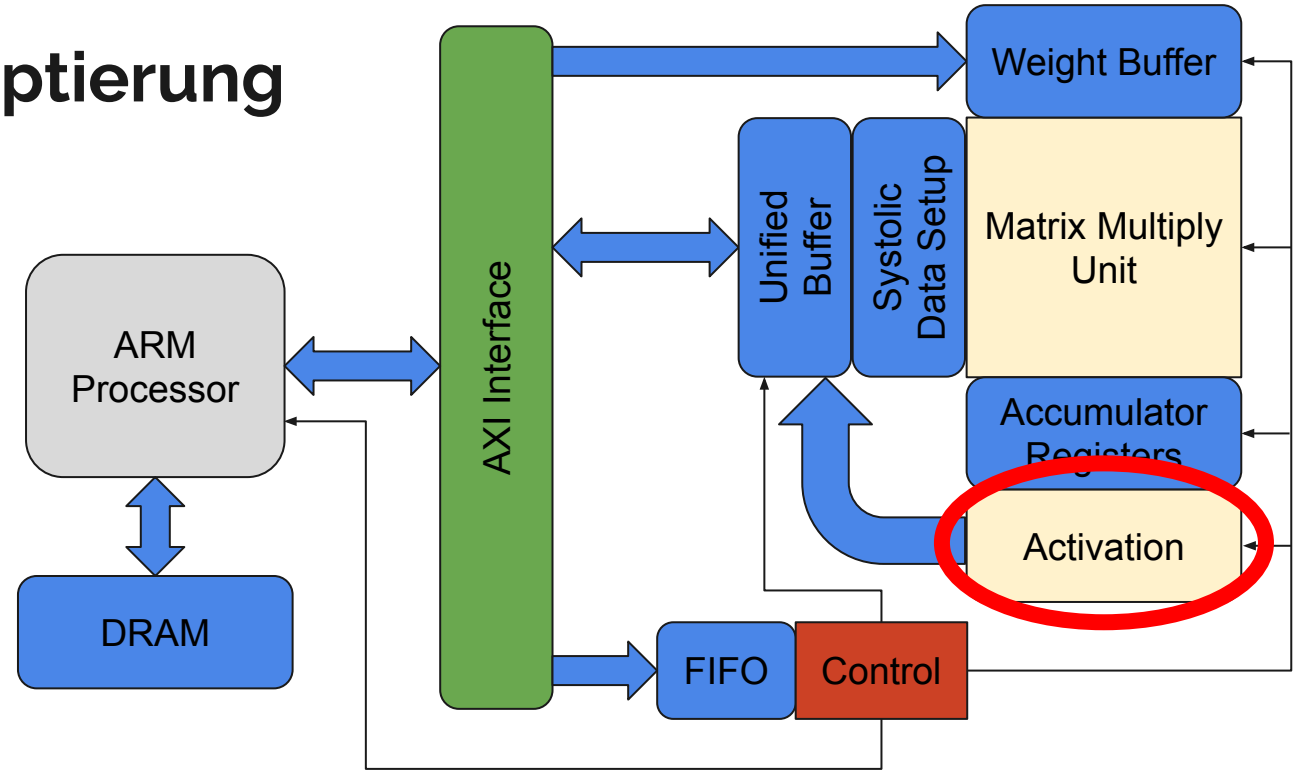
Adaptierung



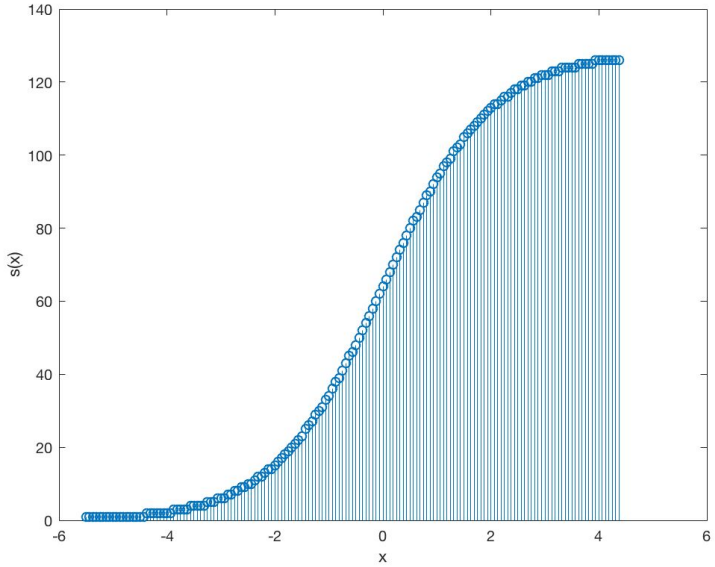
Systolic Data Setup

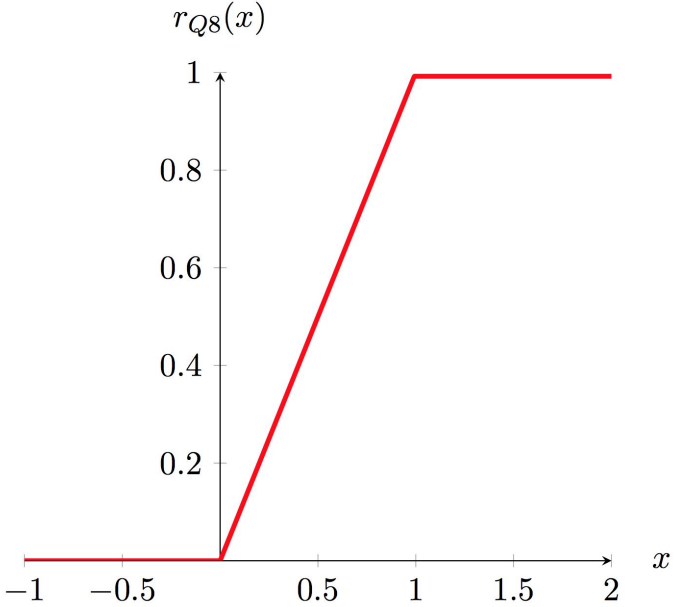


Adaptierung

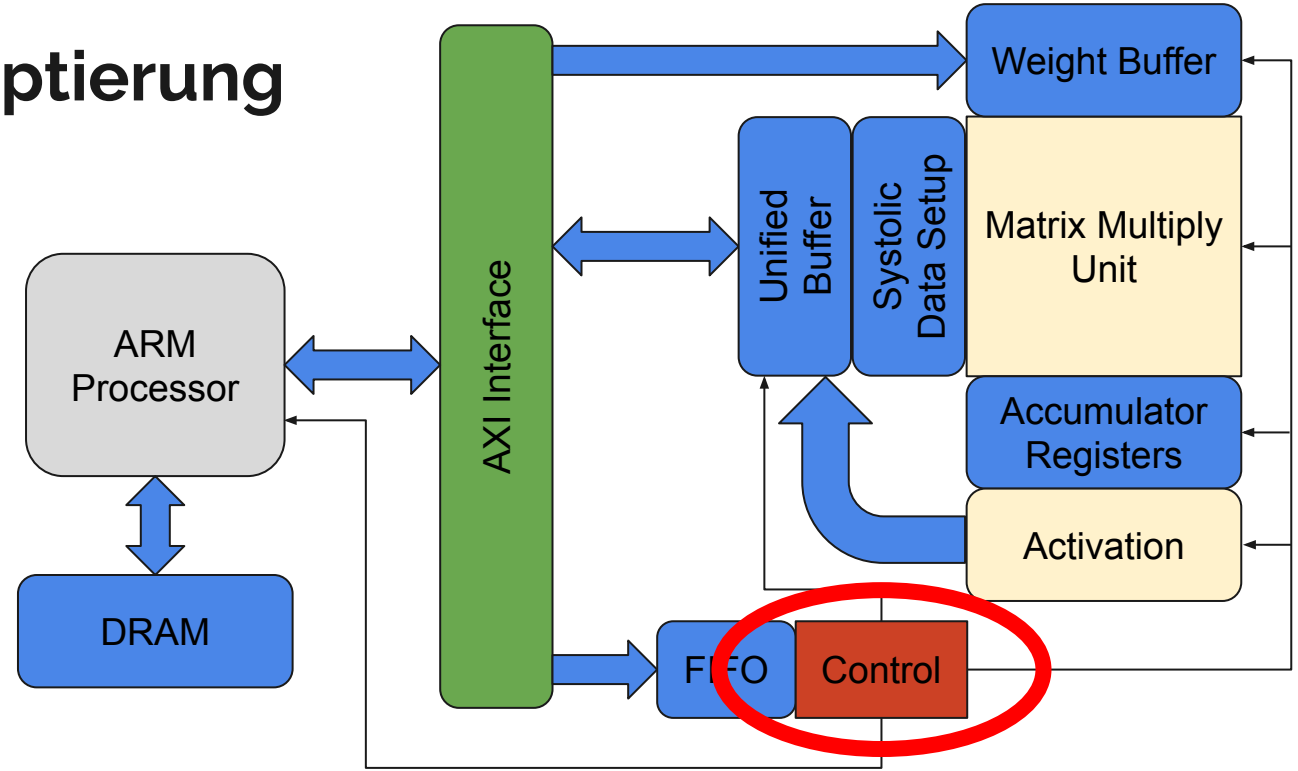


Aktivierungseinheit





Adaptierung



Instruktionssatz

Standard-Instruktionstyp:

- Arithmetische Operationen
  - matrix_multiply
  - activate
- Allgemeine Operationen
  - nop
  - halt
  - synchronize

Weight-Instruktionstyp:

- Laden von Gewichten
  - read_weights

Auswertung

Evaluation und Vergleiche

- Ressourcennutzung
- Leistungsaufnahme
- Performance
- Genauigkeit
- Berechnungsfehler

Ressourcennutzung und Leistungsaufnahme

mit 177,77MHz

Größe	6	8	10	12	14
LUT	802	873	936	970	3985
LUTRAM	93	109	125	141	239
FF	2112	2536	3031	3519	7144
BRAM	49	68	86.5	105.5	139
DSP	47	77	115	161	218
BUFG	1	1	1	1	1

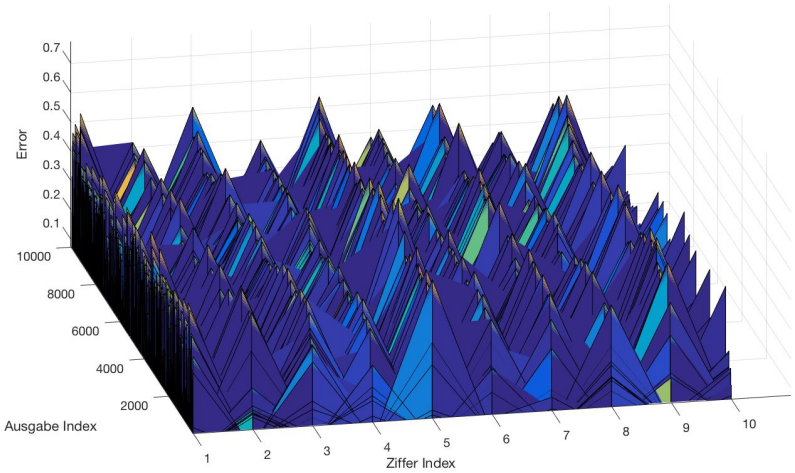
Größe	6	8	10	12	14
Leistung des Prozessor in W	1.493				
Dynamische Leistung der TPU in W	0.04	0.051	0.061	0.071	0.186
Statische Leistung der TPU in W	0.146	0.148	0.15	0.152	0.159
Temperatur in $C°$	44.4	44.5	44.7	44.8	46.2

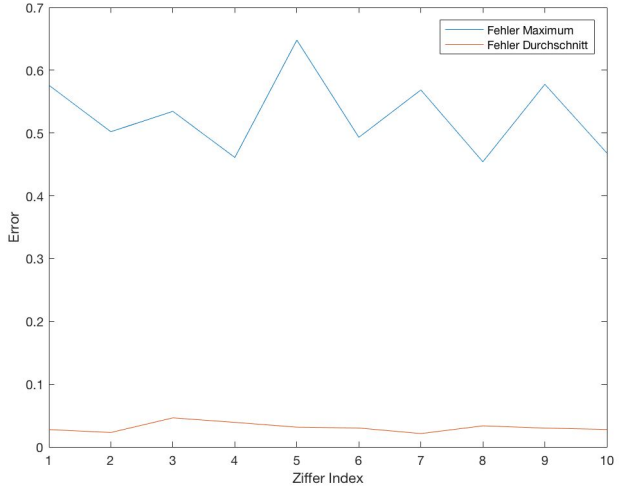
Performance mit dem MNIST Datensatz

Tensor Processing Unit mit $177,\overline{77}MHz$						**Intel**	**BCM2837**
Matrix Größe	**6**	**8**	**10**	**12**	**14**	**Core i5**	**4x ARM**
Anzahl der						**-5287U**	**Cortex-A53**
Instruktionen	431	326	261	216	186	mit	mit
Dauer in μs	383,746	289,014	234,277	194,222	165,528	$2,9GHz$	$1.2GHz$
Dauer pro							
Vektor in μs	63,958	36,127	23,428	16,185	11,823	62	763

Genauigkeit und Fehler mit dem MNIST Datensatz

- **TensorFlow: 97,86%**
- **TPU: 97,73%**





Vielen Dank für Ihre Aufmerksamkeit!

Weitere Infos:
www.github.com/jofrfu/tinyTPU