MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

A Novel Adaptive Quantization Methodology for 8-bit Floating-Point Training

Norbert Wehn

Carl Zeiss Stiftung

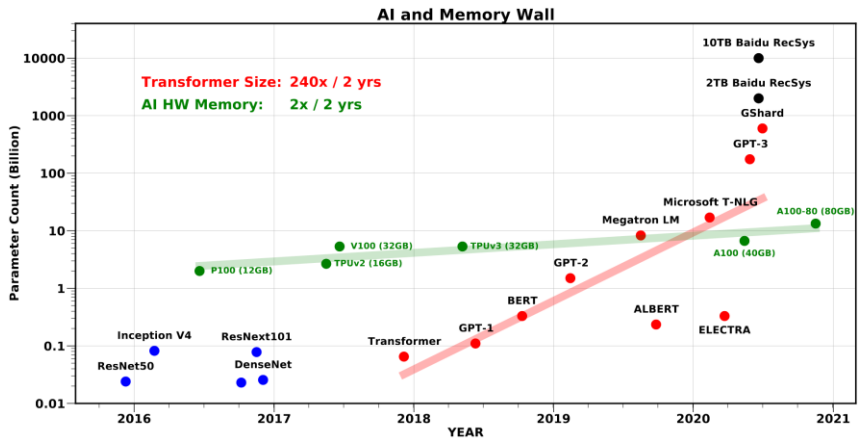
Open6GHub

SustAInML

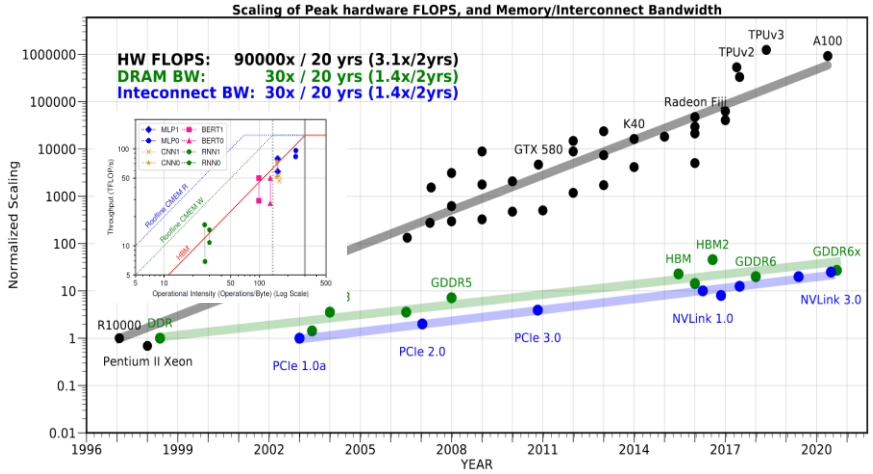
RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

AI Memory Challenge

Model Size versus AI accelerator memory capacity



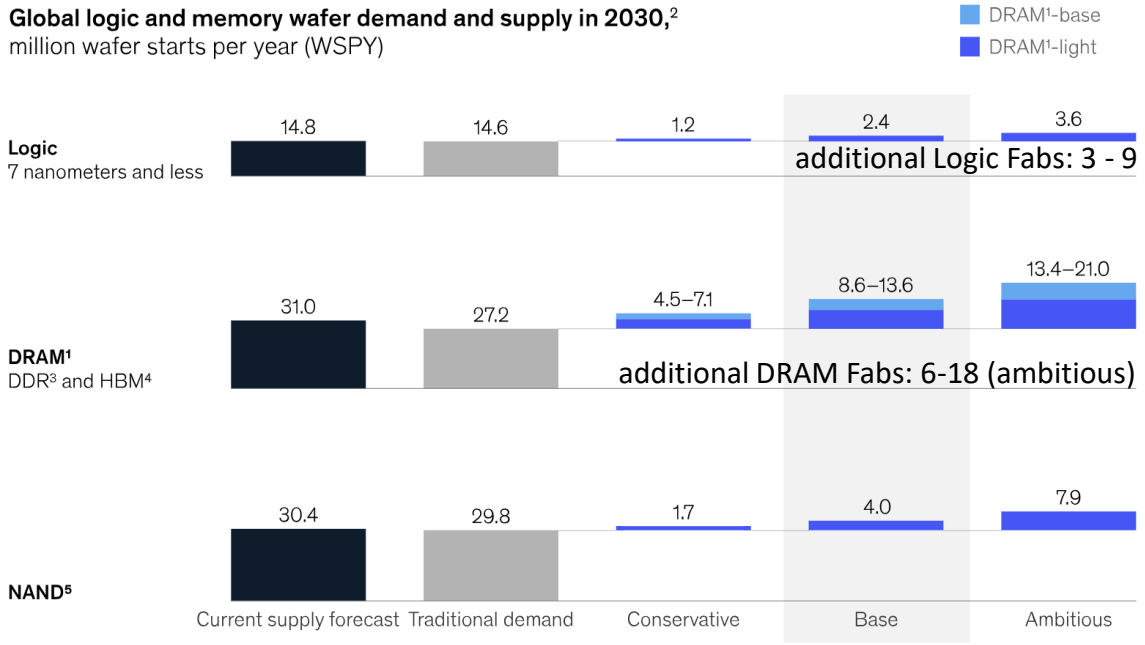
Memory Bandwidth



Source: AI and Memory Wall/Medium Post

By 2030, generative AI will increase demand for wafers significantly.

Global logic and memory wafer demand and supply in 2030,[2]
million wafer starts per year (WSPY)

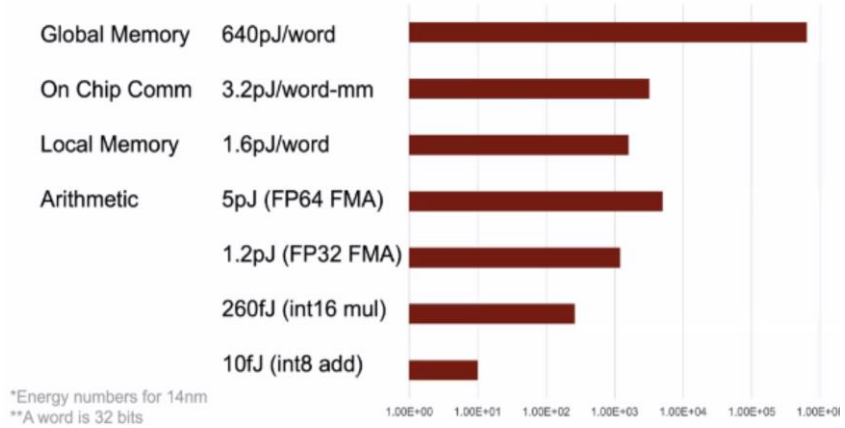


[1]Dynamic random-access memory.
[2]Two DRAM demand scenarios are considered. DRAM base case: AI accelerators and CPU+GPU systems have same memory content per server. DRAM light case: CPU+AI accelerator systems have lower memory content per server (50% less) than CPU+GPU systems.
[3]Double data rate memory.
[4]High-bandwidth memory.
[5]NAND = "not-and," a type of memory.
Source: *World fab forecast*, SEMI, December 12, 2023; McKinsey analysis

Source: McKinsey, March 2024

- DRAM increase due to generative AI larger than logic increase

AI Memory Challenges - Energy

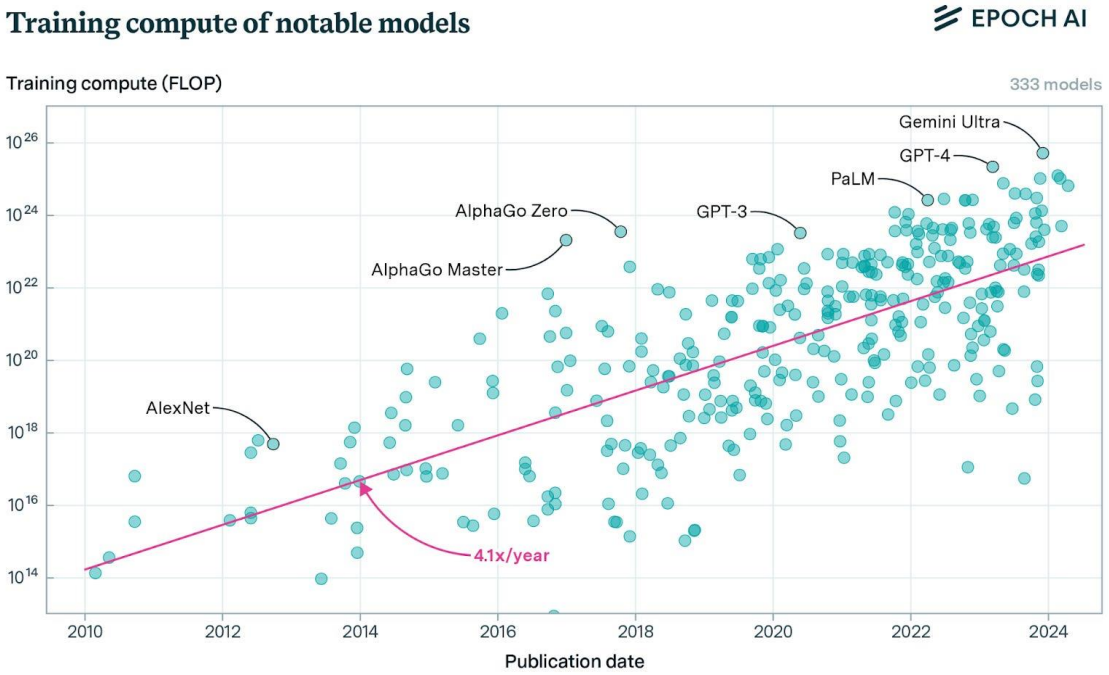


Source: TSMC

Operation		Picojoules per Operation		
		45 nm	7 nm	45 / 7
+	Int 8	0.03	0.007	4.3
	Int 32	0.1	0.03	3.3
	BFloat 16	--	0.11	--
	IEEE FP 16	0.4	0.16	2.5
	IEEE FP 32	0.9	0.38	2.4
×	Int 8	0.2	0.07	2.9
	Int 32	3.1	1.48	2.1
	BFloat 16	--	0.21	--
	IEEE FP 16	1.1	0.34	3.2
	IEEE FP 32	3.7	1.31	2.8
SRAM	8 KB SRAM	10	7.5	1.3
	32 KB SRAM	20	8.5	2.4
	1 MB SRAM[1]	100	14	7.1
GeoMean[1]		--	--	2.6
DRAM		Circa 45 nm	Circa 7 nm	
	DDR3/4	1300[2]	1300[2]	1.0
	HBM2	--	250-450[2]	--
	GDDR6	--	350-480[2]	--

Table 2. Energy per Operation: 45 nm [16] vs 7 nm. Memory is pJ per 64-bit access.

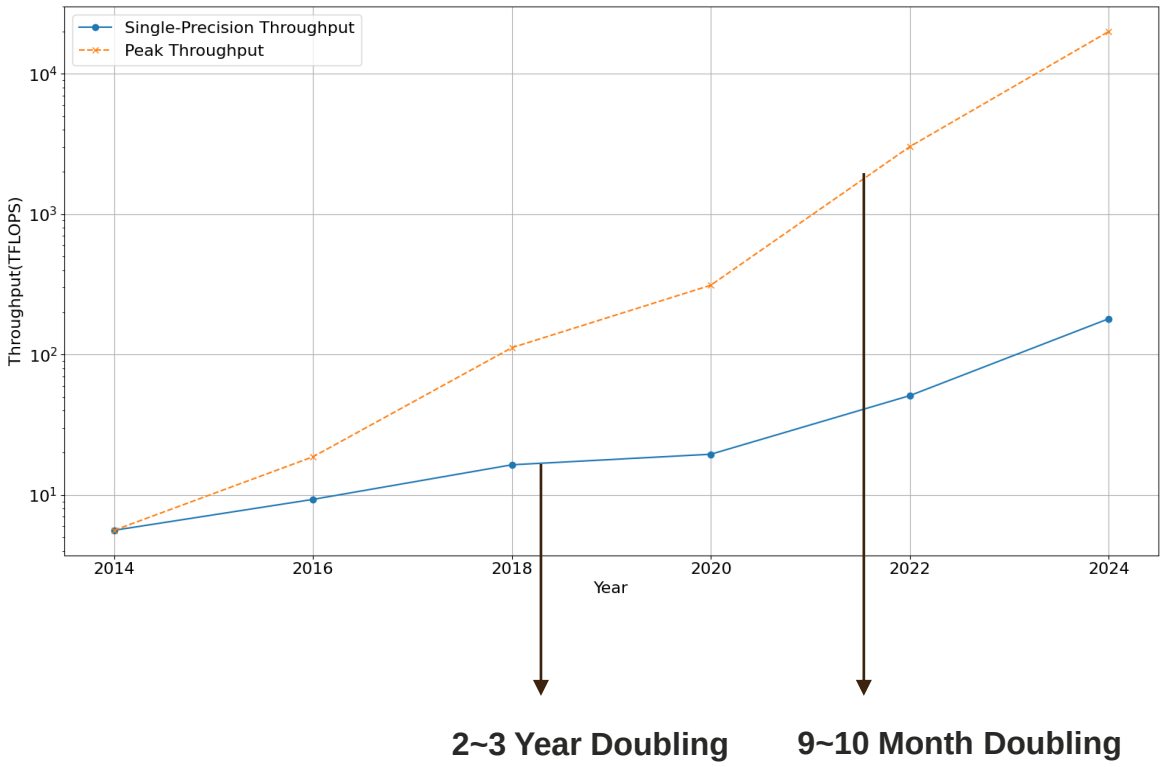
Jouppi, et al. "Ten lessons from three generations shaped google's tpuv4i: Industrial product." In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pp. 1-14. IEEE, 2021.



Training CHAT-GPT3 (175B parameters)

- 10,000 Nvidia V100 Cores (Microsoft Datacenter)
- 15 days training
- 1287 MWh: 552t CO_2
- Equivalent ~ 3 jet place CO_2 round trips San Francisco/New York
- Large source of energy consumption is off-chip memory access

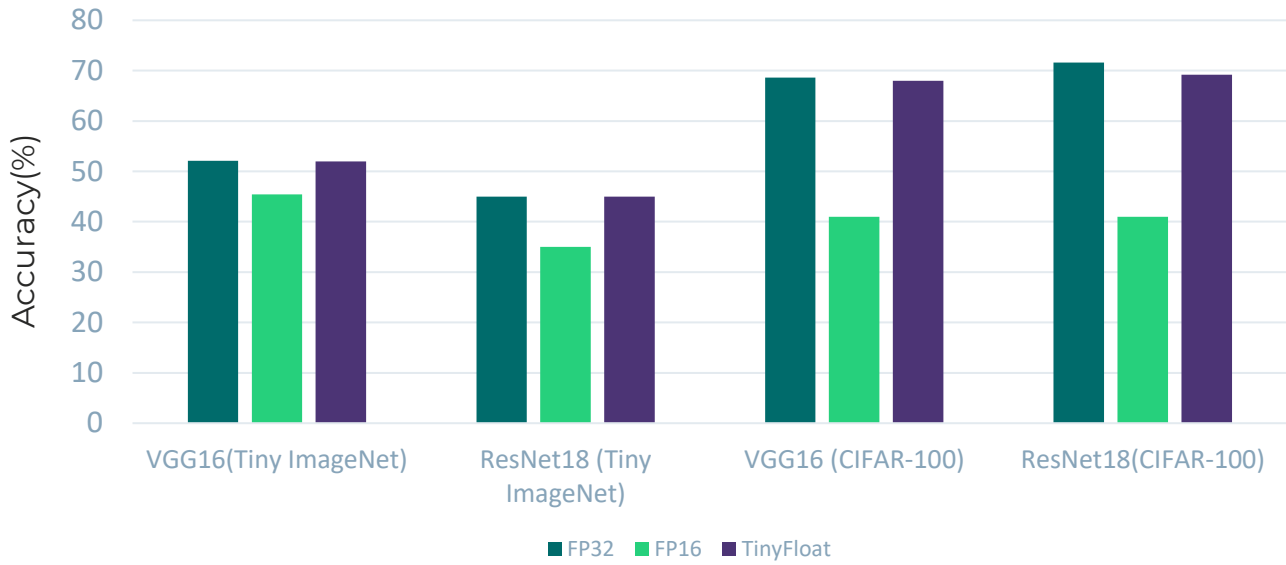
Nvidia GPUs Throughput



- Large portion of GPU's throughput increase by utilizing low bit-width compute cores
- GPUs equipped with cores for different data formats (int8, int4, FP4, FP8, FP16, FP32, FP64) → large area
- Inference: low precision data formats well investigated, but not for training

Low Precision Formats for DNN Training

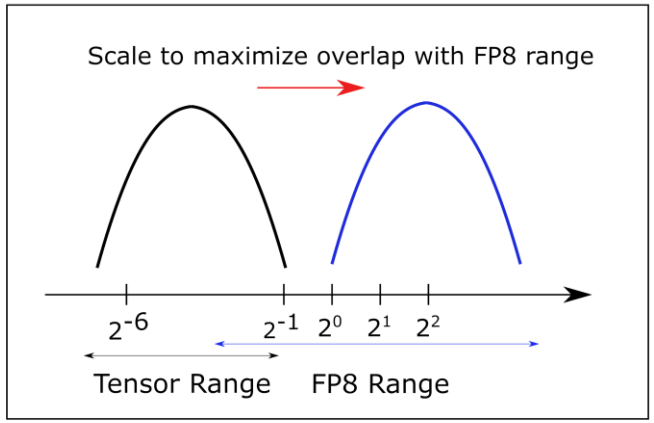
	Exp-Size	Mantissa Size	Min	Max
FP32	8	23	1.40e-45	3.0e38
FP16	5	10	5.96e-8	6.5e4
BFloat16	8	7	1.0e-38	1.0e38
TinyFloat12	7	4	1.35e-20	8.64e18
FP8	5	2	3.09e-5	6.0e4



- Mantissa width has minimal impact on accuracy (FP32, TinyFlaot)
- Significant accuracy drop with 5-bit Exp (FP16)
- Exponent width can not be reduced due to the dynamic range demands of DNN training

8-bit Floating Point Quantization (introduced by IBM)

FP8 range : 3.09e-5 ~ 6.0e4



- MobileNet backward pass activations range: 1.0e-40 ~5.0e-1
- FP8 range cannot cover the range requirement of DNN training
- Possible solution: scaling data to FP8 range

Challenges

- FP8 state-of-the-art relies on scaling data to the FP8 representable region using offline experiments with the trained model
- NVIDIA presents online scaling in the latest GPUs, but the details of the methodology is not publicly available
- Scaling operation requires multiplication and division

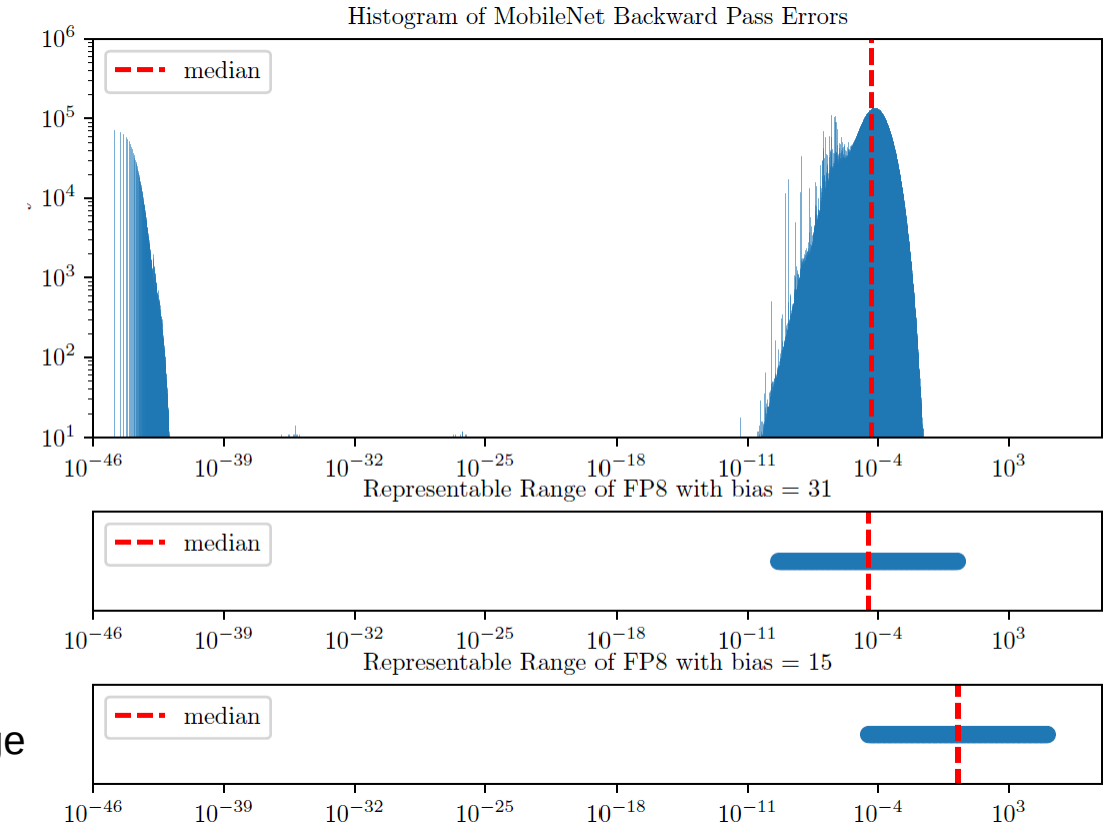
Adjusting FP8 data format

$$FP\ Value = (-1^{Sign}) * 1.Mantissa * 2^{(Exponent-Bias)}$$

$$Bias = 2^{(ExponentSize-1)} - 1$$

Idea: consider *Bias* as a variable instead of constant=15

- Instead of scaling we shift the data
- No multiplication/division necessary

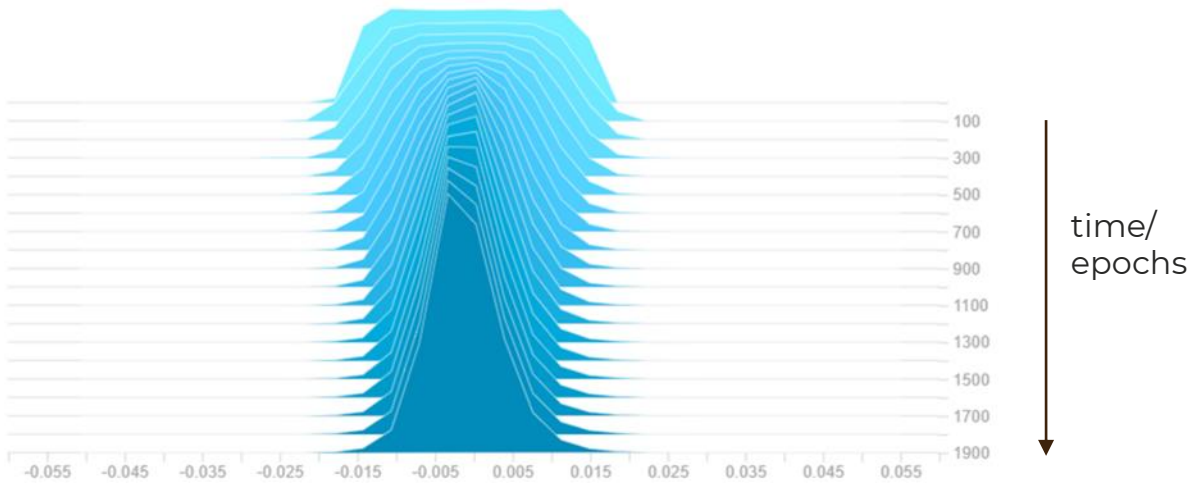


Bias = 31 → match of dynamic range

Bias = 15 → mismatch of dynamic range

How to calculate the *bias* online?

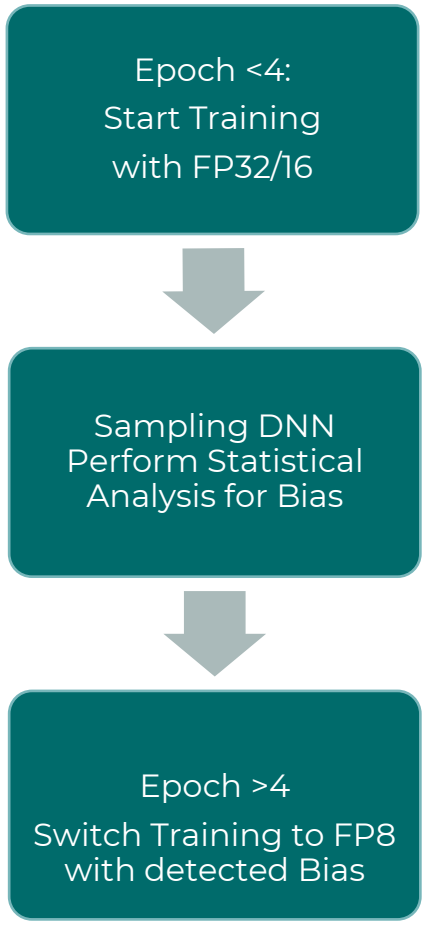
Online Bias Calculation



- Data range remains quite stable -> initial epochs are sufficient for bias calculations
- Median is used for bias calculations since it has robustness to outliers

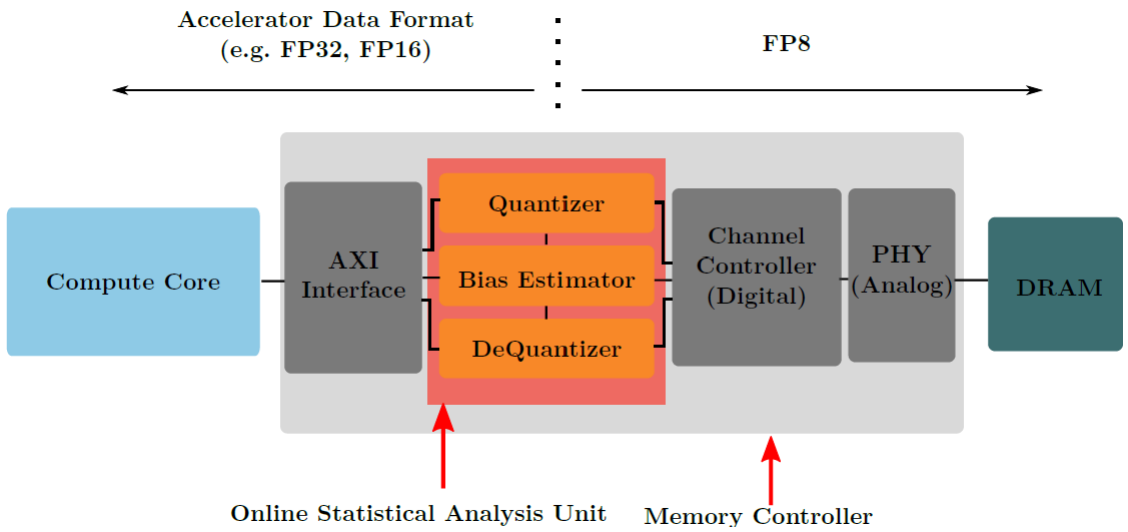
Median Value	Bias	Median Value	Bias	Median Value	Bias
64	10	2	15	0.0625	20
32	11	1	16	.	.
16	12	0.5	17	.	.
8	13	0.25	18	.	.
4	14	0.125	19	0.0000305175	31

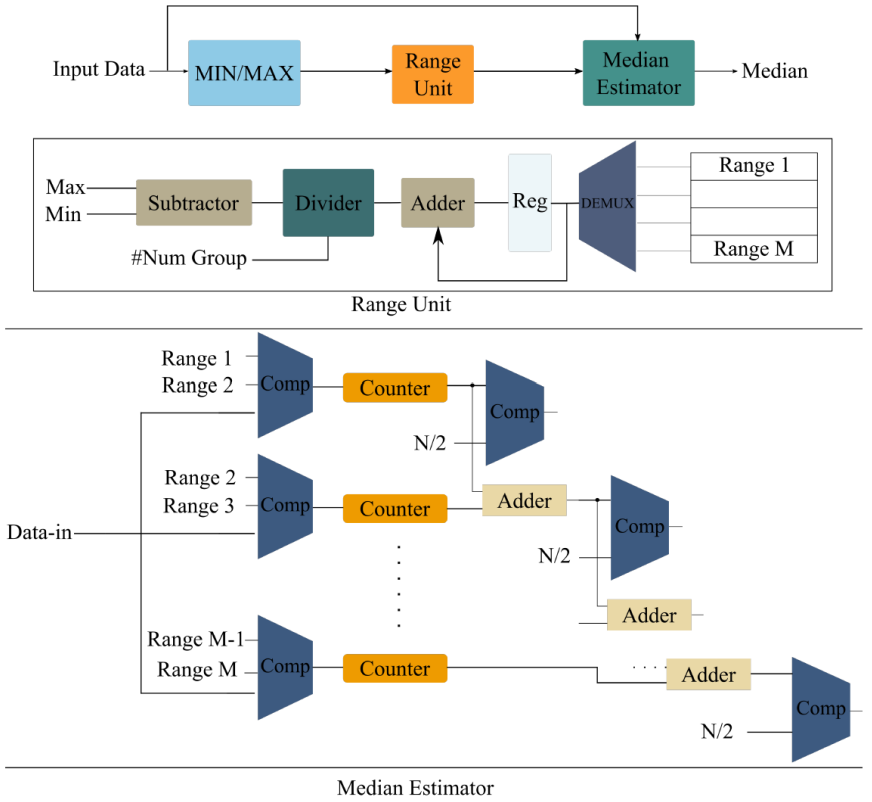
- Warm-Up phase with FP32 or FP16 to calculate biases



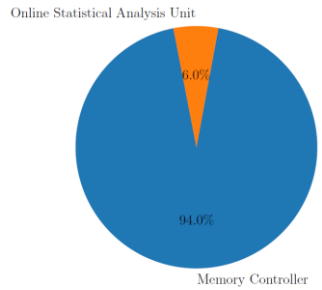
Implementation and Compatibility with HW Platforms

- Online monitoring and compression/decompression integrated into the DRAM memory controller
- Low complexity median calculation unit
- No modification of compute core necessary





Block	OP	Area	Freq	Power
Quantizer	FP32toFP8	66 μm2	1 GHz	0.0033 mW
DeQuantizer	FP8toFP32	31 μm2	1 GHz	0.003 mW
Bias Estimator	Median Calc	18996 μm2	1 GHz	10.8 mW



22nm FDSOI technology

Results/Accuracy

DNN training accuracy results comparison.

Application	DataSet	DNN Model	Accuracy Measured in	Number of Training Epochs	FP32	FP16	BFloat16	FP8 (This Work)
Natural Language Processing	Multi30k	Transformer-base	BLEU	250	33.2	29.6	32.8	30.8
	IMDB review	LSTM	%	15	88.20	85.10	88	87.31
	PennTreeBank	LSTM	PPL	35	104.4	107.7	104.6	109.2
Image Classification	Cifar100	DenseNet	%	65	80.20	32	80	79.10
		ResNet18	%	65	71.60	41	71.50	71
		ResNet101	%	65	79.78	68.10	79.70	78.80
		VGG16	%	65	68.60	24	67.70	67
		GoogleNet	%	65	78.10	75	78	77.03
	Cifar10	ResNet18	%	100	93.02	91.80	93	93
		VGG16	%	100	93.64	92.10	93.60	93.25
		GoogleNet	%	100	95	91	95	94.60
		ResNet101	%	100	95.50	89.40	95.50	94.80
	TinyImageNet	MobileNet-V2	%	50	52.30	37.20	52	51.10
		VGG16	%	50	52.10	45.40	52	52
		ResNet18	%	50	45	35	45	44.60
		GoogleNet	%	50	50.70	12	50.40	50

- Accuracy comparable with FP32 data format

Results/Energy

SCALE-Sim Config	Data Width	Type	DDR3-DIMM DQ = 64, DataRate=1866		DDR4-DIMM DQ = 64, DataRate=2133	
			Time(ms)	Energy(mJ)	Time(ms)	Energy(mJ)
TPU-like #MACs:128×128 on-chip:36MB	8-bit	F	1.16	2.14	0.68	1.53
		B	1.18	2.17	0.69	1.58
		G	1.03	1.95	0.63	1.23
	32-bit	F	4.13	7.32	2.49	5.35
		B	4.13	7.35	2.51	5.47
		G	4.44	8.96	2.83	5.38
Qualcomm-like #MACs:64×64 on-chip:9MB	8-bit	F	1.01	1.93	0.6	1.44
		B	1	1.94	0.61	1.47
		G	1.07	2.02	0.63	1.31
	32-bit	F	4.16	5.13	2.39	7.11
		B	4.15	5.21	2.4	7.09
		G	4.35	4.77	2.68	8.49

- DRAM energy saving 3x compared to FP32

RTU MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Results/Energy

SCALE-Sim Config	Data Width	Type	DDR3-DIMM DQ = 64, DataRate=1866		DDR4-DIMM DQ = 64, DataRate=2133	
			Time(ms)	Energy(mJ)	Time(ms)	Energy(mJ)
TPU-like #MACs:128×128 on-chip:36MB	8-bit	F	1.16	2.14	0.68	1.53
		B	1.18	2.17	0.69	1.58
		G	1.03	1.95	0.63	1.23
	32-bit	F	4.13	7.32	2.49	5.35
		B	4.13	7.35	2.51	5.47
		G	4.44	8.96	2.83	5.38
Qualcomm-like #MACs:64×64 on-chip:9MB	8-bit	F	1.01	1.93	0.6	1.44
		B	1	1.94	0.61	1.47
		G	1.07	2.02	0.63	1.31
	32-bit	F	4.16	5.13	2.39	7.11
		B	4.15	5.21	2.4	7.09
		G	4.35	4.77	2.68	8.49

- DRAM energy saving 3x compared to FP32

RTU MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Special Thanks to the members of my AI group:

Mohammad Hassani Sadi, Chirag Sudarshan

For more information

Novel Adaptive Quantization Methodology for 8-bit Floating-Point DNN Training

M.H. Sadi, C. Sudarshan, N. Wehn

Springer Journal on Design Automation for Embedded Systems, 2024

https://eit.rptu.de/fgs/ems/start