Mokey

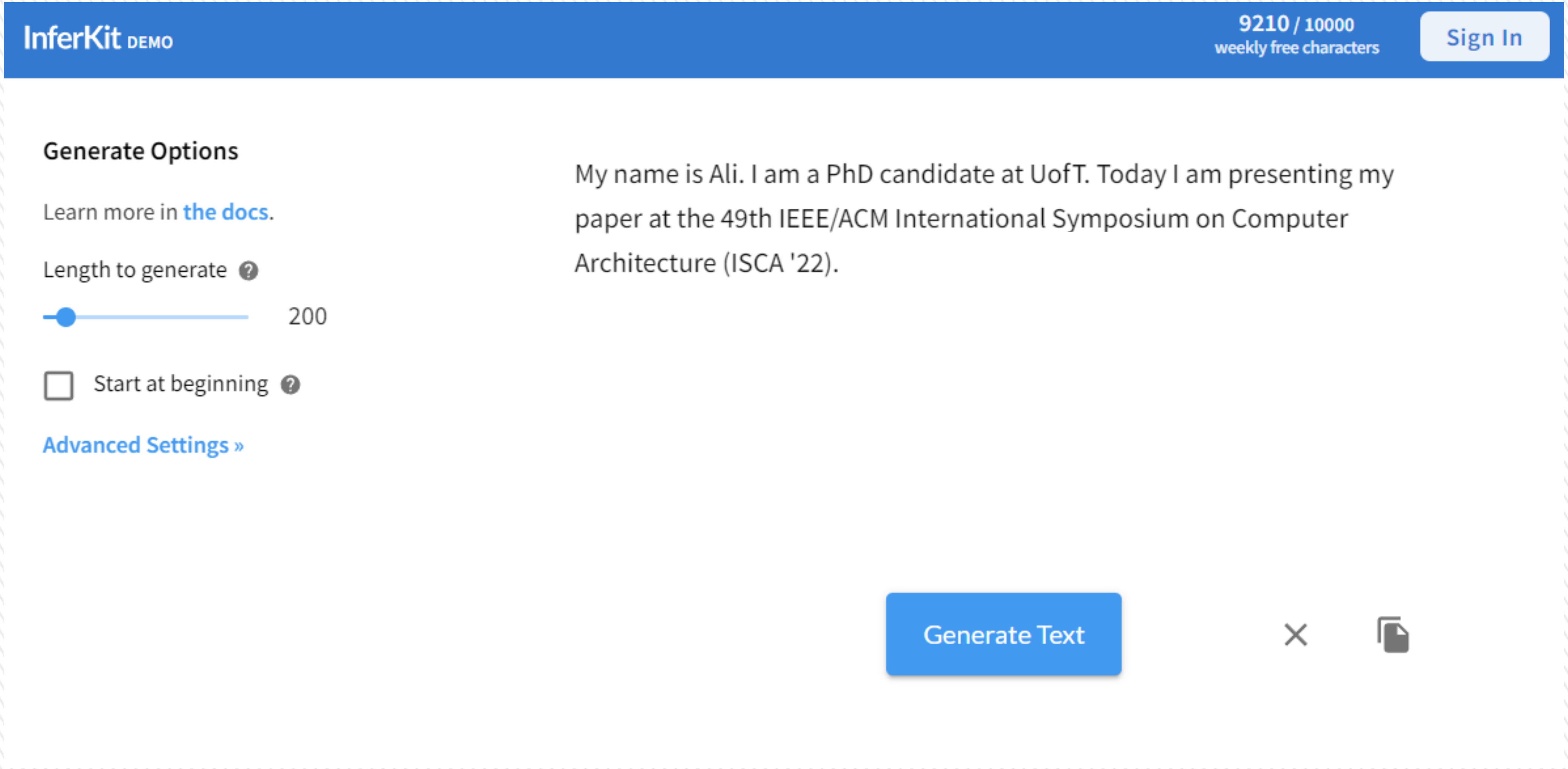
Enabling Narrow Fixed-Point Inference

for Out-of-the-Box Floating-Point Transformer Models

49th IEEE/ACM International Symposium on Computer Architecture
(ISCA '22)

Ali Hadi Zadeh[1,2], Mostafa Mahmoud[1], Ameer Abdelhadi[1], and Andreas Moshovos[1,2]
[1] University of Toronto , [2] Vector Institute

Transformers for Text Generation



Generate Options

Learn more in the docs.

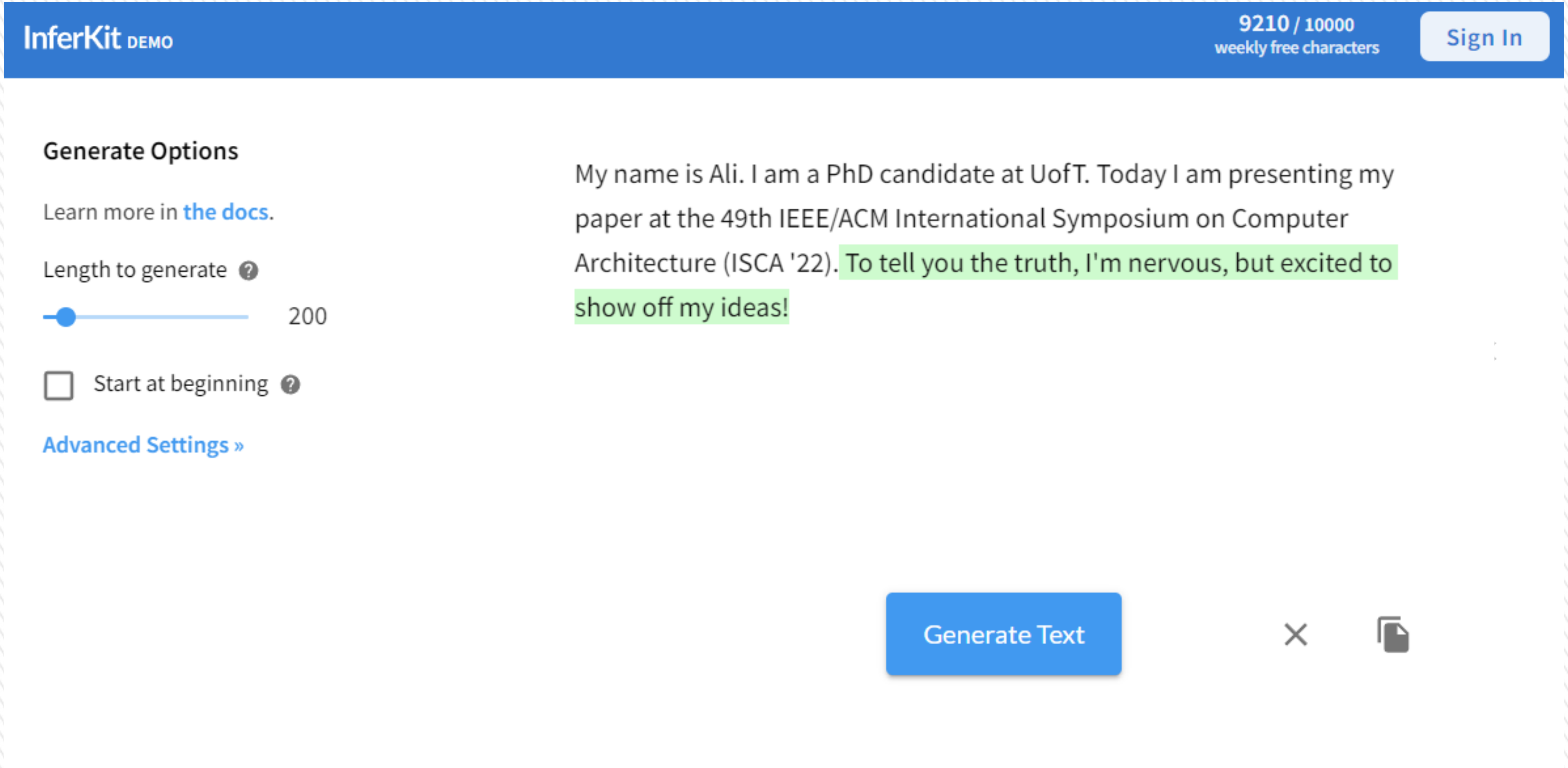
Length to generate: 200

☐ Start at beginning

My name is Ali. I am a PhD candidate at UofT. Today I am presenting my paper at the 49th IEEE/ACM International Symposium on Computer Architecture (ISCA '22).

https://app.inferkit.com/demo

Transformers for Text Generation



Generate Options

Learn more in the docs.

Length to generate: 200

☐ Start at beginning

My name is Ali. I am a PhD candidate at UofT. Today I am presenting my paper at the 49th IEEE/ACM International Symposium on Computer Architecture (ISCA '22). To tell you the truth, I'm nervous, but excited to show off my ideas!

https://app.inferkit.com/demo

Challenges

Weights

2018

1.2GB

2021

2TB

Activations

Galileo Galilei
@TheREALGalileo

Follow

Solar eclipses never get old. #SolarEclipse2017

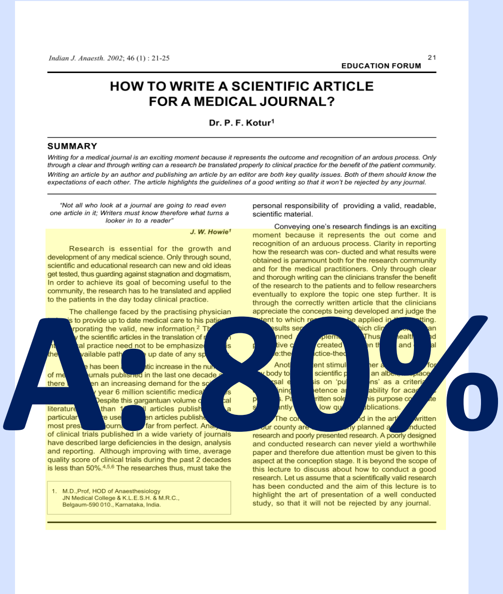
12:15 [illegible] 17

54,442 Retweets 158,426 Likes

28K 54K 158K

created @ tweeterino.com

A: 5%



Indian J. Anaesth. 2002; 46 (1) : 21-25 21

EDUCATION FORUM

HOW TO WRITE A SCIENTIFIC ARTICLE FOR A MEDICAL JOURNAL?

Dr. P. F. Kotur[1]

SUMMARY

Writing for a medical journal is an exciting moment because it represents the outcome and recognition of an ardous process. Only through a clear and through writing can a research be translated properly to clinical practice for the benefit of the patient community.

Writing an article by an author and publishing an article by an editor are both key quality issues. Both of them should know the expectations of each other. The article highlights the guidelines of a good writing so that it won't be rejected by any journal.

> *"Not all who look at a journal are going to read even one article in it; Writers must know therefore what turns a looker in to a reader"*
>
> ***J. W. Howie[1]***

Research is essential for the growth and development of any medical science. Only through sound, scientific and educational research can new and old ideas get tested, thus guarding against stagnation and dogmatism, In order to achieve its goal of becoming useful to the community, the research has to he translated and applied to the patients in the day today clinical practice.

The challenge faced by the practising physician [illegible] is to provide up to date medical care to his patie[illegible] [illegible]rporating the valid, new information.[2] Th[illegible] [illegible] the scientific articles in the translation of [illegible] [illegible]al practice need not to be emphasized [illegible] is the [illegible]vailable path[illegible] up date of any sp[illegible]

[illegible] has been [illegible]tic increase in the nu[illegible] of me[illegible]urnals published in the last one decade [illegible] there [illegible]en an increasing demand for the sc[illegible] [illegible] year 6 million scientific medica[illegible] [illegible] Despite this gargantuan volume o[illegible]cal literature [illegible] than [illegible]l articles publish[illegible] a particular [illegible] use[illegible]en articles publishe[illegible] most pres[illegible] journa[illegible] far from perfect. Ana[illegible] of clinical trials published in a wide variety of journals have described large deficiencies in the design, analysis and reporting. Although improving with time, average quality score of clinical trials during the past 2 decades is less than 50%.[4,5,6] The researches thus, must take the personal responsibility of providing a valid, readable, scientific material.

Conveying one's research findings is an exciting moment because it represents the out come and recognition of an arduous process. Clarity in reporting how the research was con- ducted and what results were obtained is paramount both for the research community and for the medical practitioners. Only through clear and thorough writing can the clinicians transfer the benefit of the research to the patients and to fellow researchers eventually to explore the topic one step further. It is through the correctly written article that the clinicians appreciate the concepts being developed and judge the extent to which re[illegible] be applied in th[illegible]tting. [illegible]esults ser[illegible]hich cli[illegible] [illegible]an [illegible]nned [illegible]pleme[illegible] Thus [illegible]ealth [illegible]d [illegible]tive c[illegible] created [illegible]en th[illegible] and [illegible]al [illegible]e:the[illegible]ctice-theo[illegible]

Anot[illegible]ent stimul[illegible]her a[illegible] for [illegible] body to [illegible] scientific p[illegible] an alb[illegible] [illegible]rsal e[illegible]sis on 'pu[illegible]ns' as a criteria [illegible]ining [illegible]etence a[illegible]ability for acad[illegible] p[illegible]s. Pa[illegible]ritten sole[illegible] this purpose co[illegible]te s[illegible]antly [illegible] low qu[illegible]ublications.

The co[illegible]d in the ar[illegible] written [illegible]ur county are [illegible]y planned a[illegible]nducted research and poorly presented research. A poorly designed and conducted research can never yield a worthwhile paper and therefore due attention must be given to this aspect at the conception stage. It is beyond the scope of this lecture to discuss about how to conduct a good research. Let us assume that a scientifically valid research has been conducted and the aim of this lecture is to highlight the art of presentation of a well conducted study, so that it will not be rejected by any journal.

---

1. M.D.,Prof, HOD of Anaesthesiology
JN Medical College & K.L.E.S.H. & M.R.C.,
Belgaum-590 010., Karnataka, India.

A: 80%

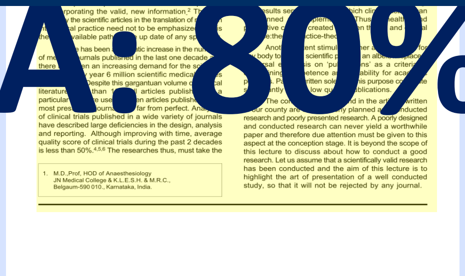
Challenges

Weights

2021

2TB

Activations



of clinical trials published in a wide variety of journals have described large deficiencies in the design, analysis and reporting. Although improving with time, average quality score of clinical trials during the past 2 decades is less than 50%.[4,5,6] The researches thus, must take the

research and poorly presented research. A poorly designed and conducted research can never yield a worthwhile paper and therefore due attention must be given to this aspect at the conception stage. It is beyond the scope of this lecture to discuss about how to conduct a good research. Let us assume that a scientifically valid research has been conducted and the aim of this lecture is to highlight the art of presentation of a well conducted study, so that it will not be rejected by any journal.

1. M.D.,Prof, HOD of Anaesthesiology
JN Medical College & K.L.E.S.H. & M.R.C.,
Belgaum-590 010., Karnataka, India.

A:80%

Memory:

Performance & Energy Bottleneck

Challenges

Weights

2018

1.2GB

2021

2TB

Activations

Galileo Galilei
@TheREALGalileo

Solar eclipses never get old. #SolarEclipse2017

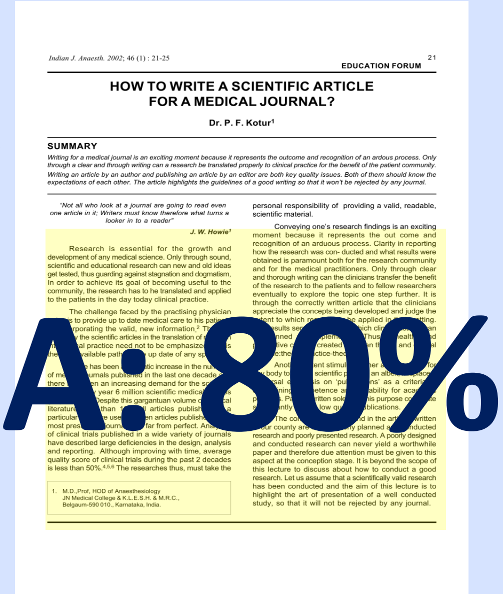
12:15 [illegible]17

54,442 Retweets 158,426 Likes

28K 54K 158K

created @ tweeterino.com

A: 5%



Indian J. Anaesth. 2002; 46 (1) : 21-25 21

EDUCATION FORUM

HOW TO WRITE A SCIENTIFIC ARTICLE
FOR A MEDICAL JOURNAL?

Dr. P. F. Kotur[1]

SUMMARY

Writing for a medical journal is an exciting moment because it represents the outcome and recognition of an ardous process. Only through a clear and through writing can a research be translated properly to clinical practice for the benefit of the patient community. Writing an article by an author and publishing an article by an editor are both key quality issues. Both of them should know the expectations of each other. The article highlights the guidelines of a good writing so that it won't be rejected by any journal.

"Not all who look at a journal are going to read even one article in it; Writers must know therefore what turns a looker in to a reader"

J. W. Howie[1]

Research is essential for the growth and development of any medical science. Only through sound, scientific and educational research can new and old ideas get tested, thus guarding against stagnation and dogmatism, In order to achieve its goal of becoming useful to the community, the research has to he translated and applied to the patients in the day today clinical practice.

The challenge faced by the practising physician [illegible] is to provide up to date medical care to his patie[illegible] [illegible]rporating the valid, new information.[2] Th[illegible] [illegible]y the scientific articles in the translation of [illegible] [illegible]al practice need not to be emphasize[illegible] is the [illegible]vailable path[illegible] up date of any sp[illegible]

[illegible]e has been [illegible]atic increase in the nu[illegible] of me[illegible]urnals published in the last one decade [illegible] there [illegible]en an increasing demand for the sc[illegible] [illegible]y year 6 million scientific medica[illegible] [illegible] Despite this gargantuan volume o[illegible]cal literature [illegible] than 1[illegible]ll articles publish[illegible] a particular [illegible]re use[illegible]en articles publishe[illegible] most pres[illegible] journa[illegible] far from perfect. Ana[illegible] of clinical trials published in a wide variety of journals have described large deficiencies in the design, analysis and reporting. Although improving with time, average quality score of clinical trials during the past 2 decades is less than 50%.[4,5,6] The researches thus, must take the personal responsibility of providing a valid, readable, scientific material.

Conveying one's research findings is an exciting moment because it represents the out come and recognition of an arduous process. Clarity in reporting how the research was con- ducted and what results were obtained is paramount both for the research community and for the medical practitioners. Only through clear and thorough writing can the clinicians transfer the benefit of the research to the patients and to fellow researchers eventually to explore the topic one step further. It is through the correctly written article that the clinicians appreciate the concepts being developed and judge the [illegible]tent to which re[illegible] be applied in th[illegible]tting. [illegible]esults se[illegible]hich cli[illegible]an [illegible]nned [illegible]pleme[illegible] Thus [illegible]ealth[illegible]nd p[illegible]tive c[illegible] created [illegible]en th[illegible] and [illegible]al [illegible]e:the[illegible]actice-theo[illegible]

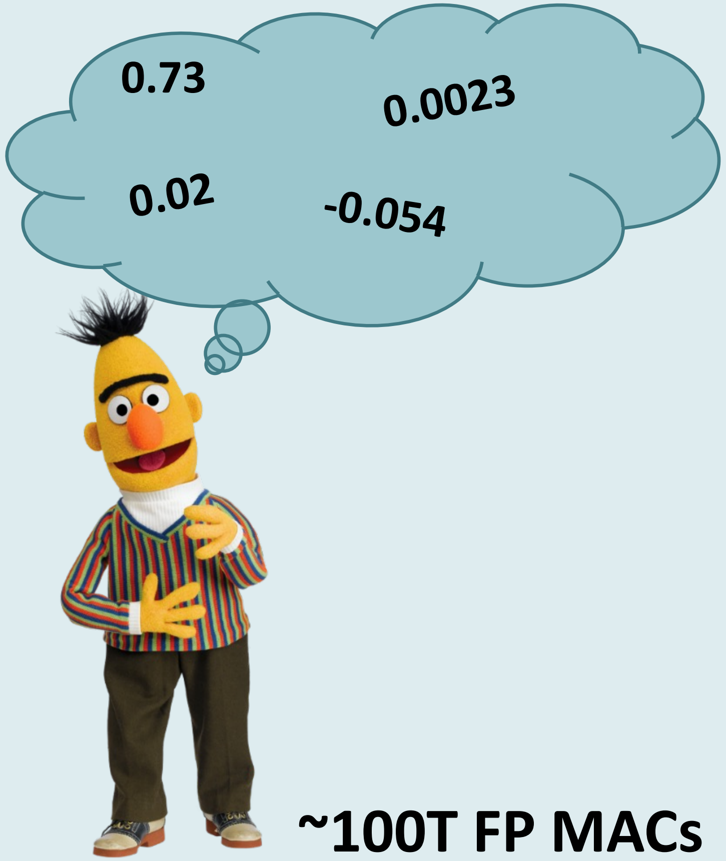
Anot[illegible]ent stimul[illegible]her a[illegible] for [illegible]y body to [illegible] scientific p[illegible] an albe[illegible]place [illegible]rsal e[illegible]sis on 'pu[illegible]ns' as a criteri[illegible] [illegible]ining [illegible]etence a[illegible]ability for aca[illegible] p[illegible]s. Pa[illegible]ritten sole[illegible]his purpose co[illegible] s[illegible]antly [illegible] low qu[illegible]ublications.

The co[illegible]nd in the art[illegible] written [illegible]our county are [illegible]ly planned a[illegible]onducted research and poorly presented research. A poorly designed and conducted research can never yield a worthwhile paper and therefore due attention must be given to this aspect at the conception stage. It is beyond the scope of this lecture to discuss about how to conduct a good research. Let us assume that a scientifically valid research has been conducted and the aim of this lecture is to highlight the art of presentation of a well conducted study, so that it will not be rejected by any journal.

[1] M.D.,Prof, HOD of Anaesthesiology
JN Medical College & K.L.E.S.H. & M.R.C.,
Belgaum-590 010., Karnataka, India.

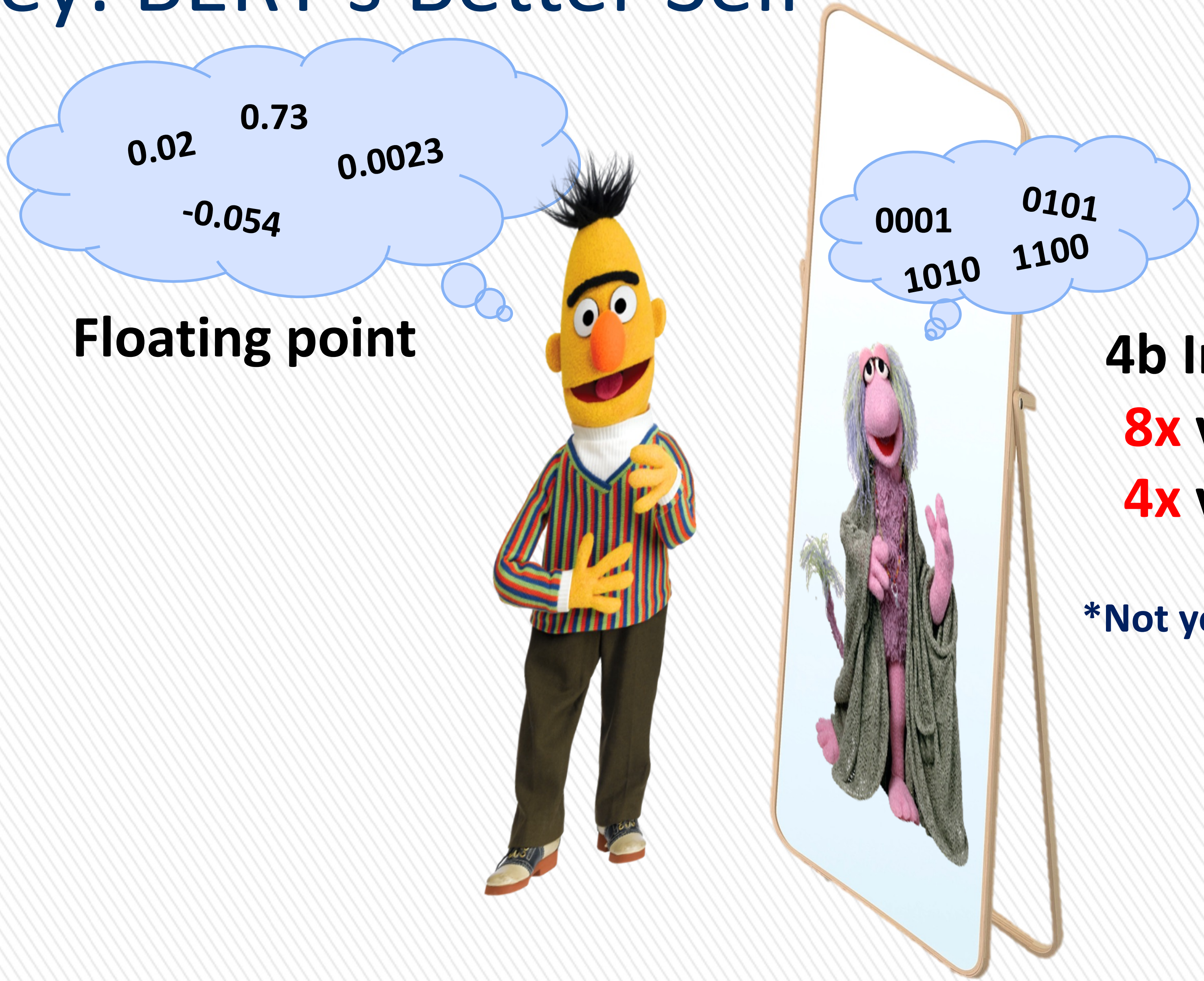
A:80%

FP Compute



~100T FP MACs

Mokey: BERT’s Better Self



Floating point

4b Int index

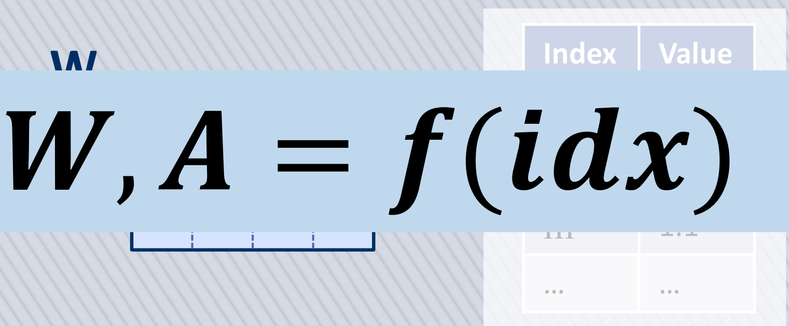
8x vs FP32

4x vs FP16

***Not your typical 4b quantization ;)**

Mokey

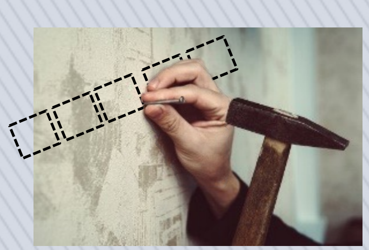
4-bit Quantization: W+A



$$W, A = f(idx)$$

Post-training

+= A x W. ➔ Count idx



Fixed-point compute

Mokey HW Accelerator

Vs. Tensor Cores: ***15x*** Faster + ***100x*** Energy Efficient

Mokey Memory Compression

For Tensor Cores:

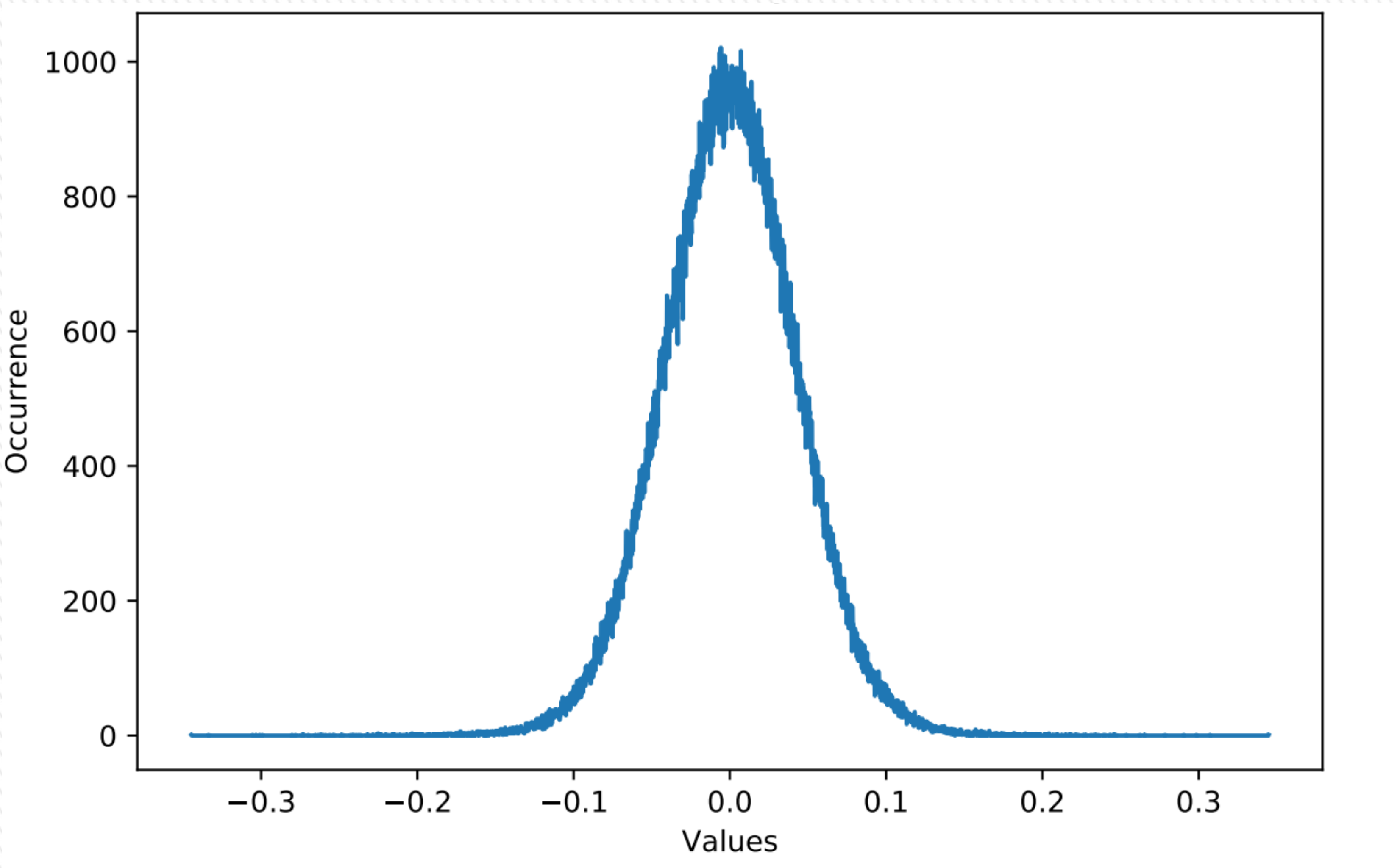
Off-chip Only: ***4x*** Faster + ***8x*** Energy Efficient

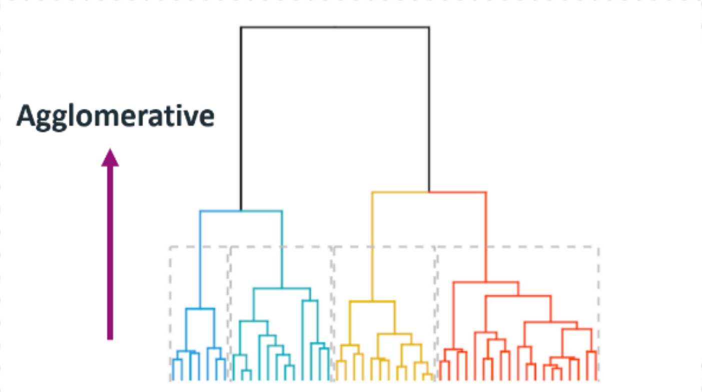
Off- and on-chip*: ***10x*** Faster + ***50x*** Energy Efficient

Roadmap

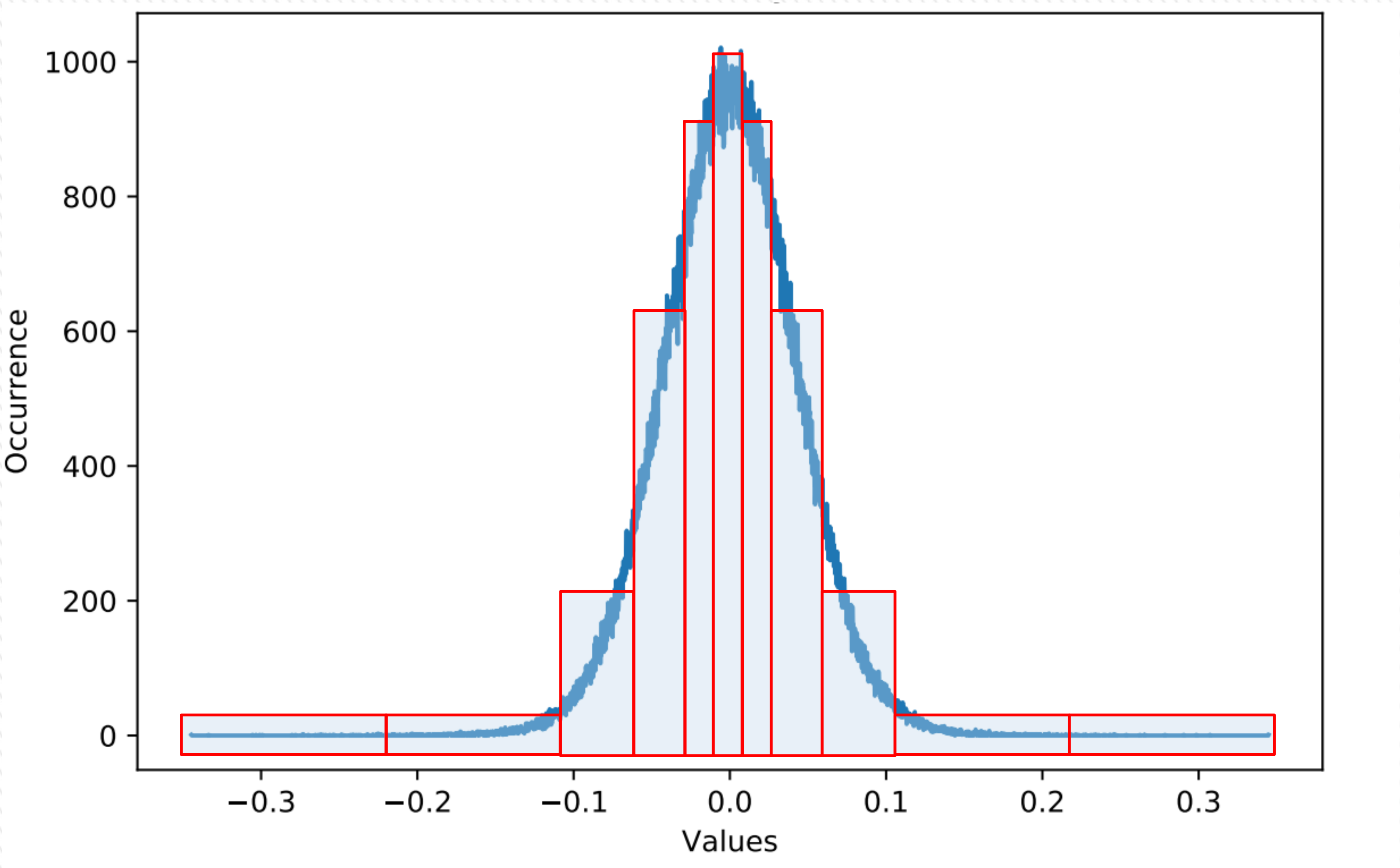
- Mokey Quantization
- Computation on indices
- Models' Accuracy
- Hardware Evaluation
- Conclusion

Dictionary-Based Quantization





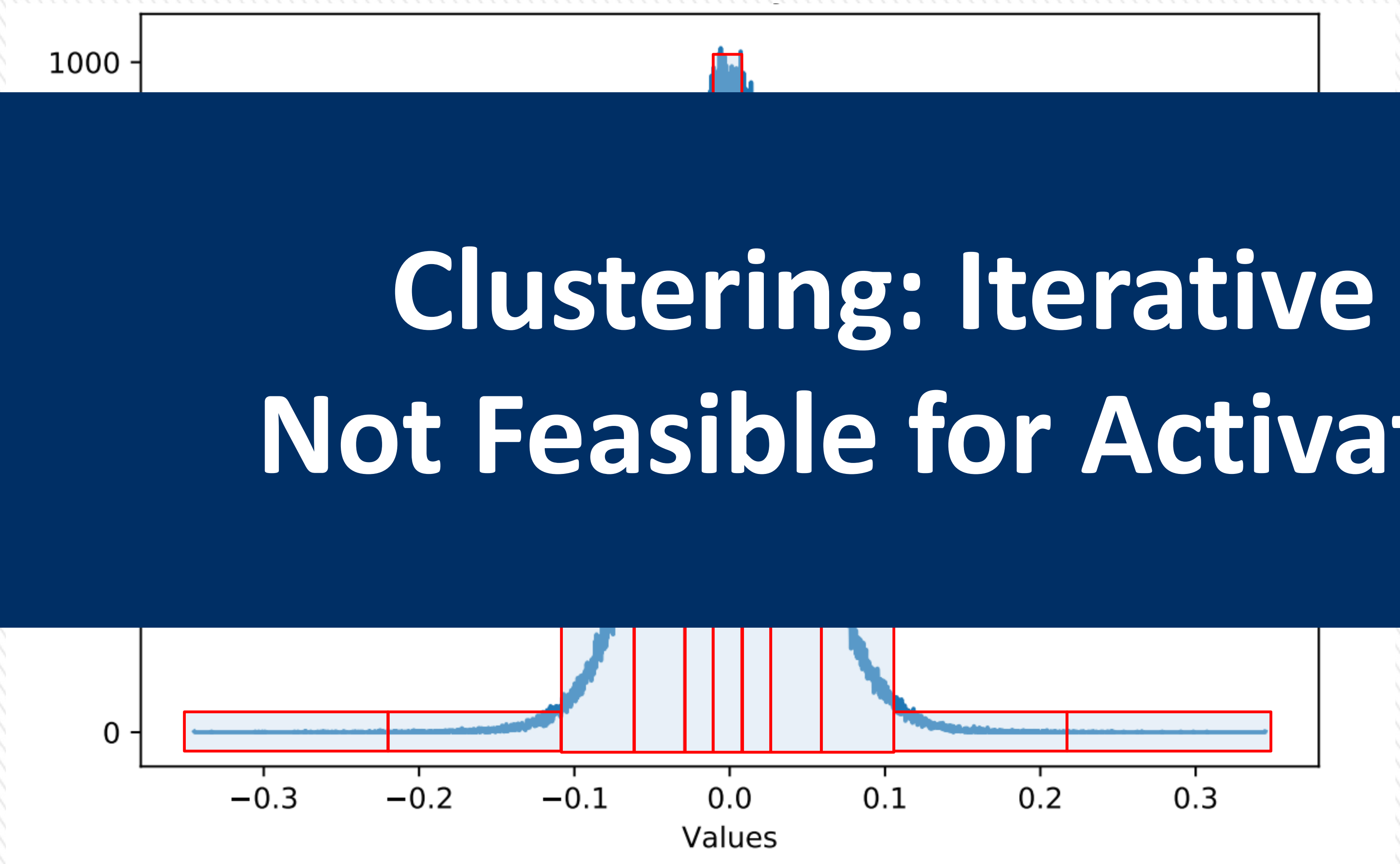
Dictionary-Based Quantization



Quantization Dict.

Index	Value
I	0.02
II	0.07
III	0.12
IV	0.25
…	…

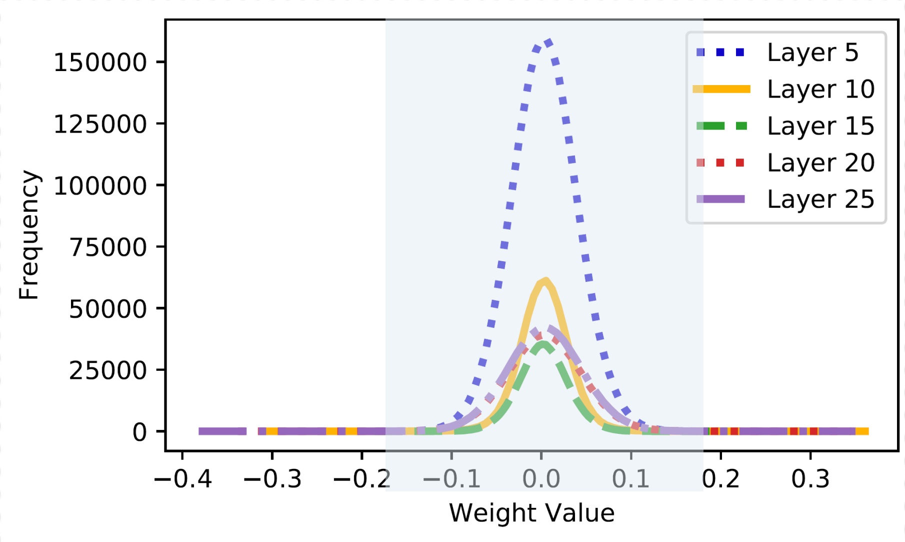
Dictionary-Based Quantization



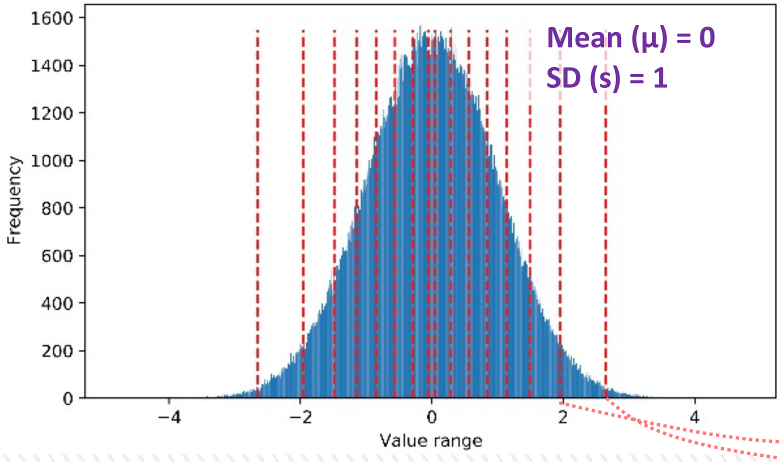
Clustering: Iterative ☹

Not Feasible for Activations

A Dictionary for All Layers



Reference Distribution



Golden Dict. (GD)

Index	Value
I	-2.7
II	-1.98
…	…
VI	1.98
XVI	2.7

Scale and **Shift** is All You Need!

A Dictionary for All Layers

Reference Distribution

Weights & Embeddings: Offline
Activation: Profiling

Scale and **Shift** is All You Need!

Inference Computation

Original

A = 0.2

W = 0.7

A×W += 0.2 × 0.7 = 0.14

Dictionary Quant.

A = I

W= II

Index	Value
I	0.2
II	0.7
III	1.1
IV	1.4
…	…

A × W += I × II

A × W += 0.2 × 0.7 = 0.14

Mokey Quant.

A = I

W= II

How?!

A × W += I × II = 0.14

How to Use Indices for Computation

Original

Dictionary Quant.

Mokey Quant.

Let’s See in Practice!

A × W += I × II

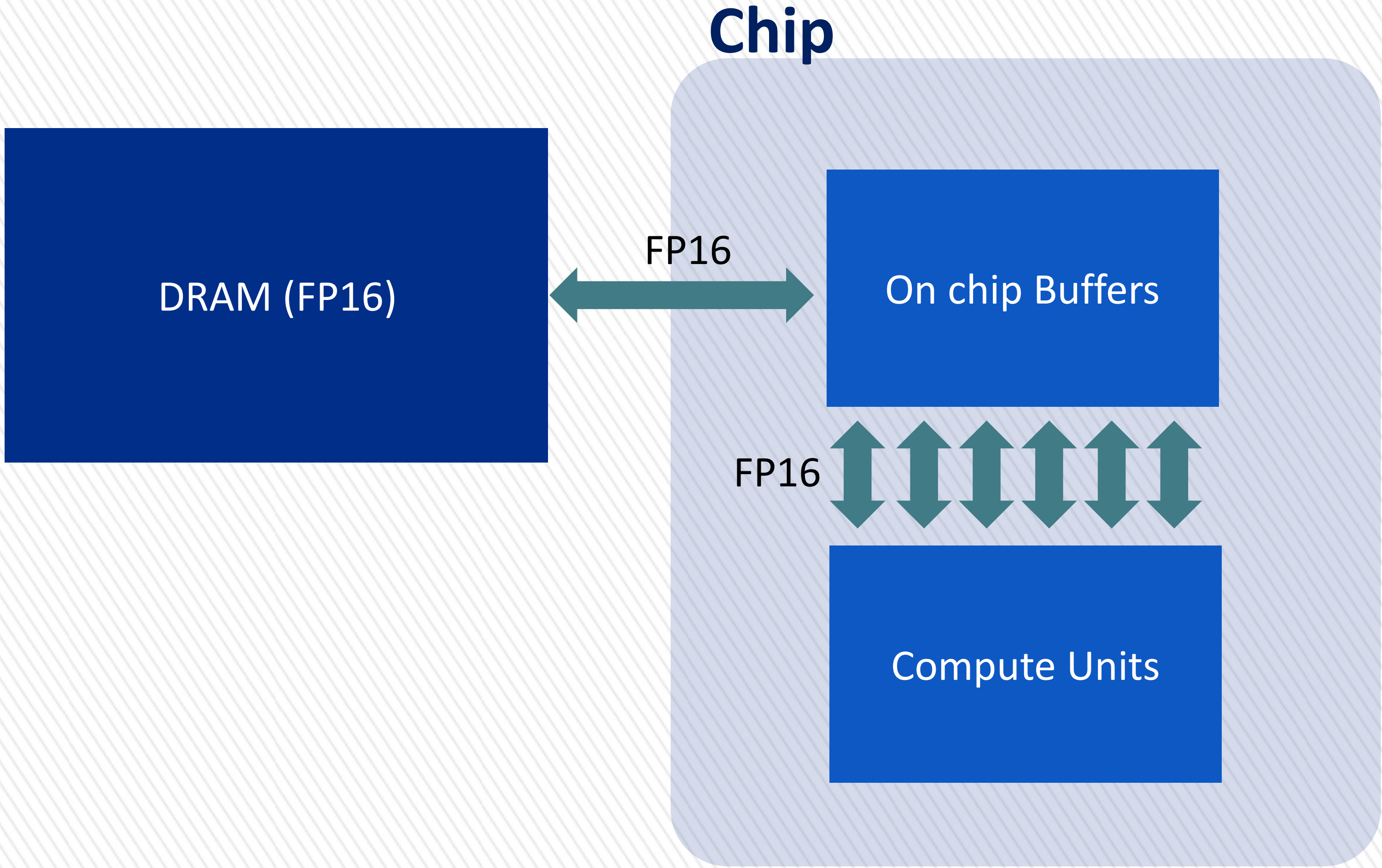
A × W += 0.2 × 0.7 = 0.14

FP16 Baseline

Original

A = 0.2 W = 0.7

A×W += 0.2 × 0.7 = 0.14



Dictionary Quantization

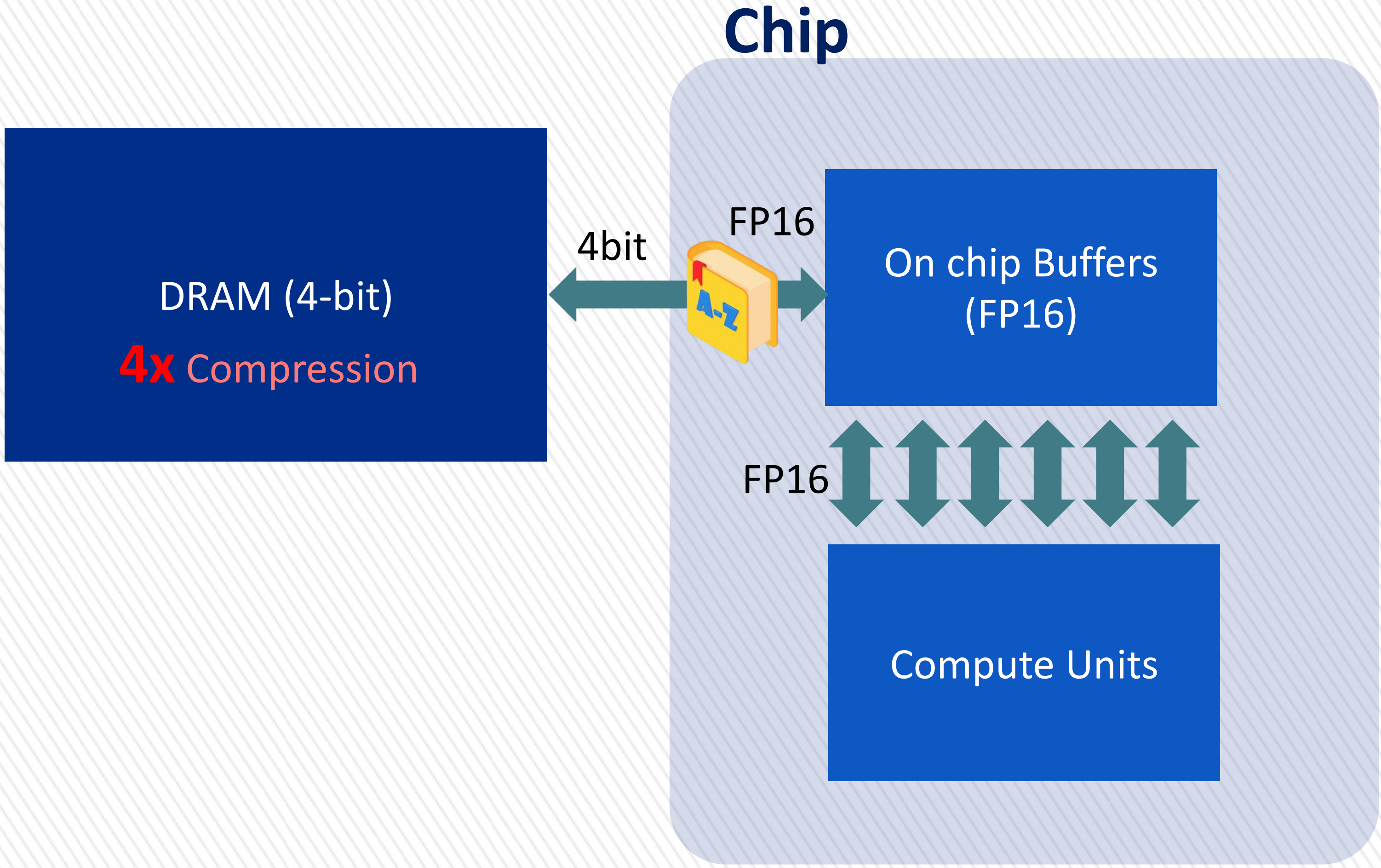
Dictionary Quant.

A = I W= II

Index	Value
I	0.2
II	0.7
III	1.1
IV	1.4
…	…

A × W += I × II

A × W += 0.2 × 0.7 = 0.14



Chip

DRAM (4-bit)

4x Compression

4bit

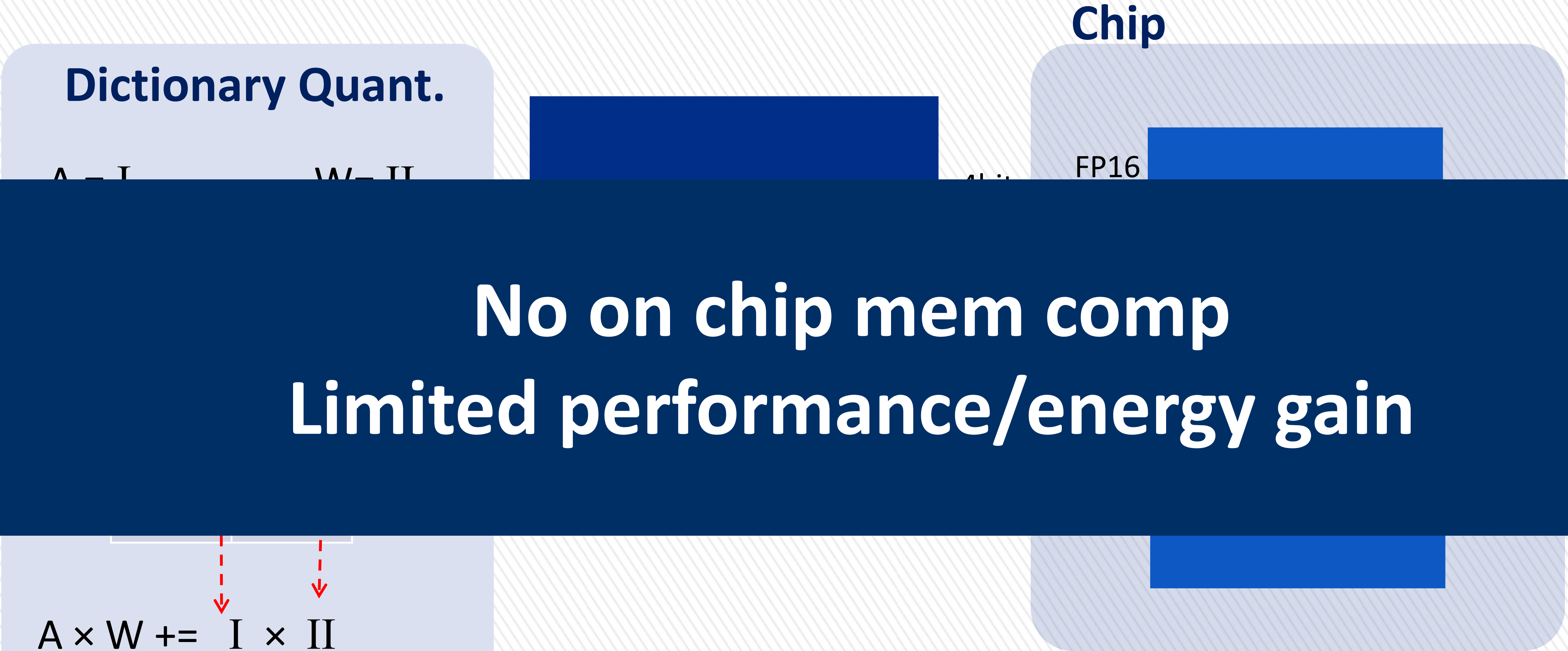
FP16

On chip Buffers (FP16)

FP16

Compute Units

Dictionary Quantization



No on chip mem comp
Limited performance/energy gain

$A \times W \mathrel{+}= \mathrm{I} \times \mathrm{II}$

$A \times W \mathrel{+}= 0.2 \times 0.7 = 0.14$

Dictionary Quantization

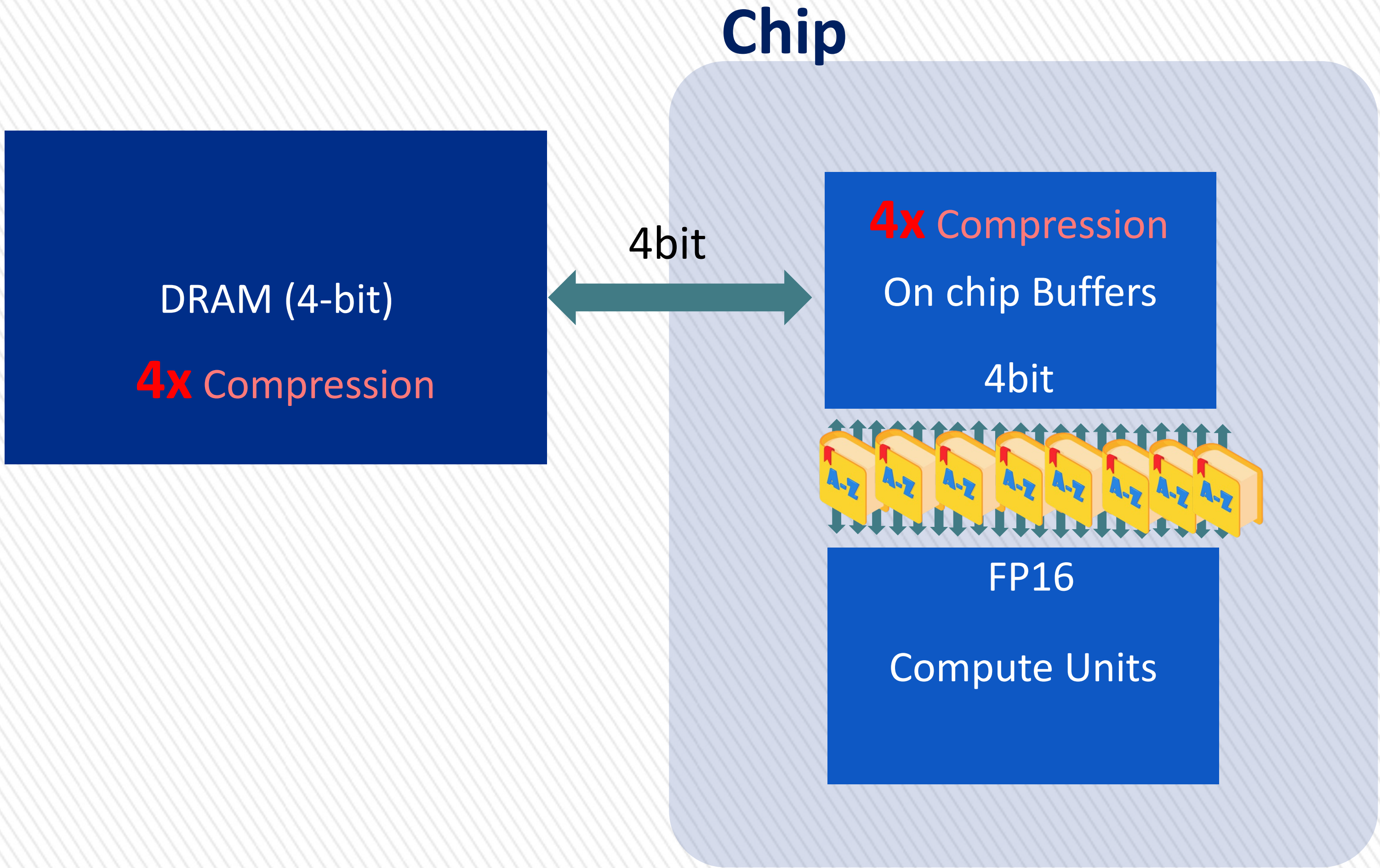
Dictionary Quant.

$A = \mathrm{I}$ $W = \mathrm{II}$

Index	Value
I	0.2
II	0.7
III	1.1
IV	1.4
…	…

$A \times W \mathrel{+}= \mathrm{I} \times \mathrm{II}$

$A \times W \mathrel{+}= 0.2 \times 0.7 = 0.14$

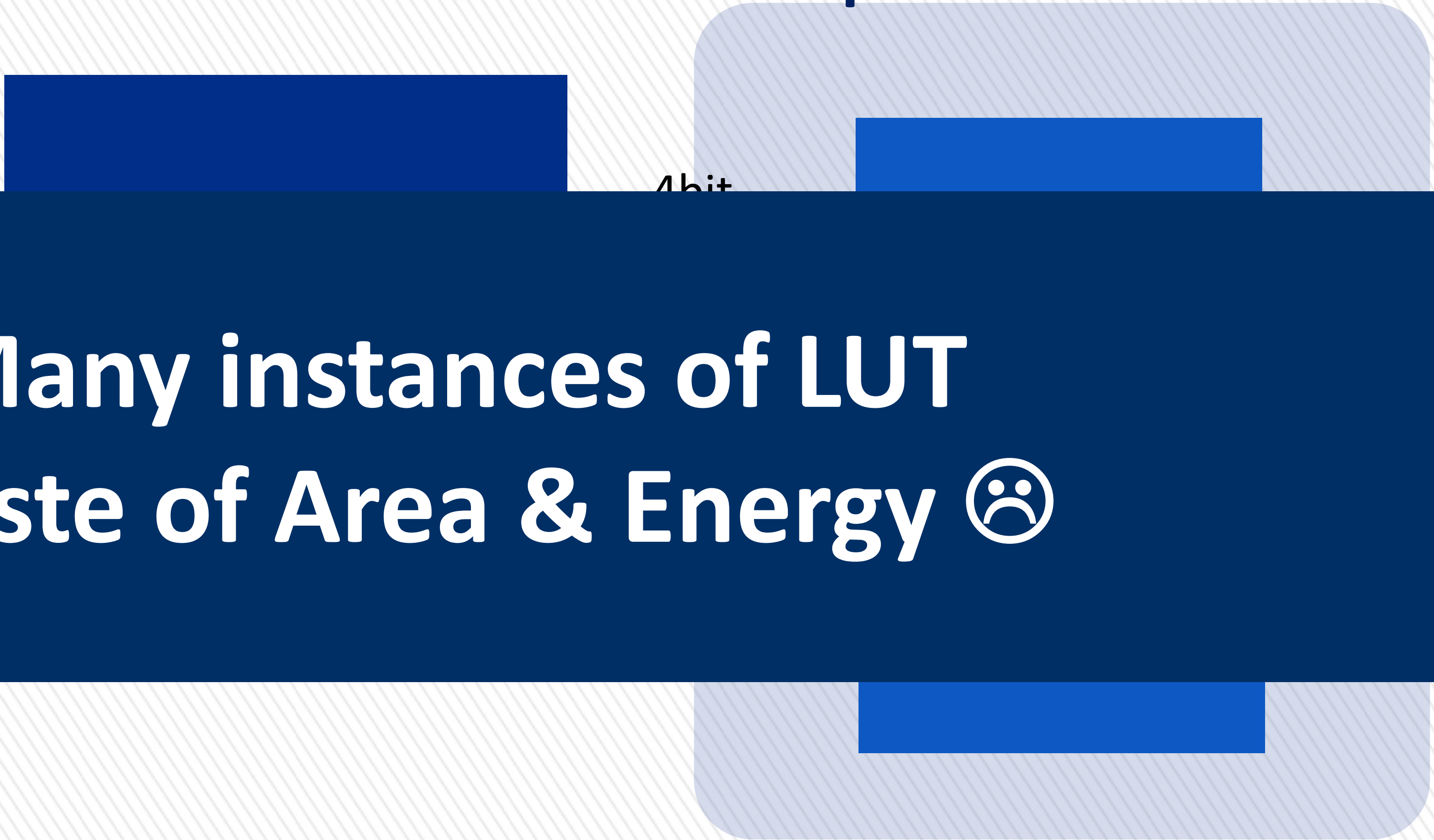


Dictionary Quantization

Dictionary Quant.

A = I

W= II



Many instances of LUT

Waste of Area & Energy ☹

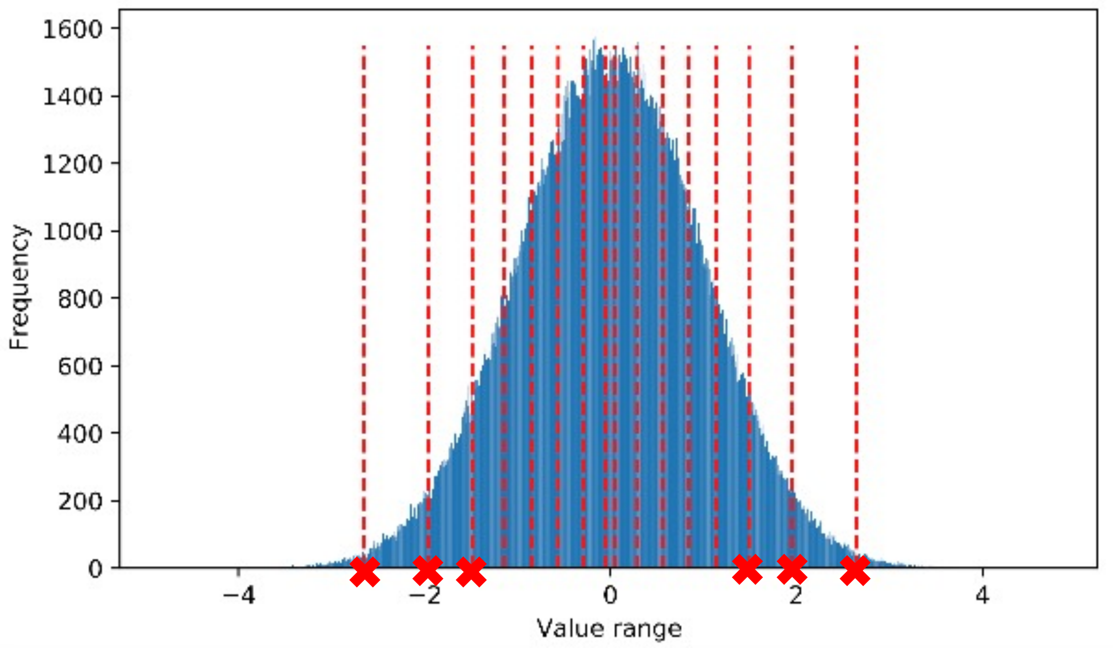
A × W += I × II

A × W += 0.2 × 0.7 = 0.14

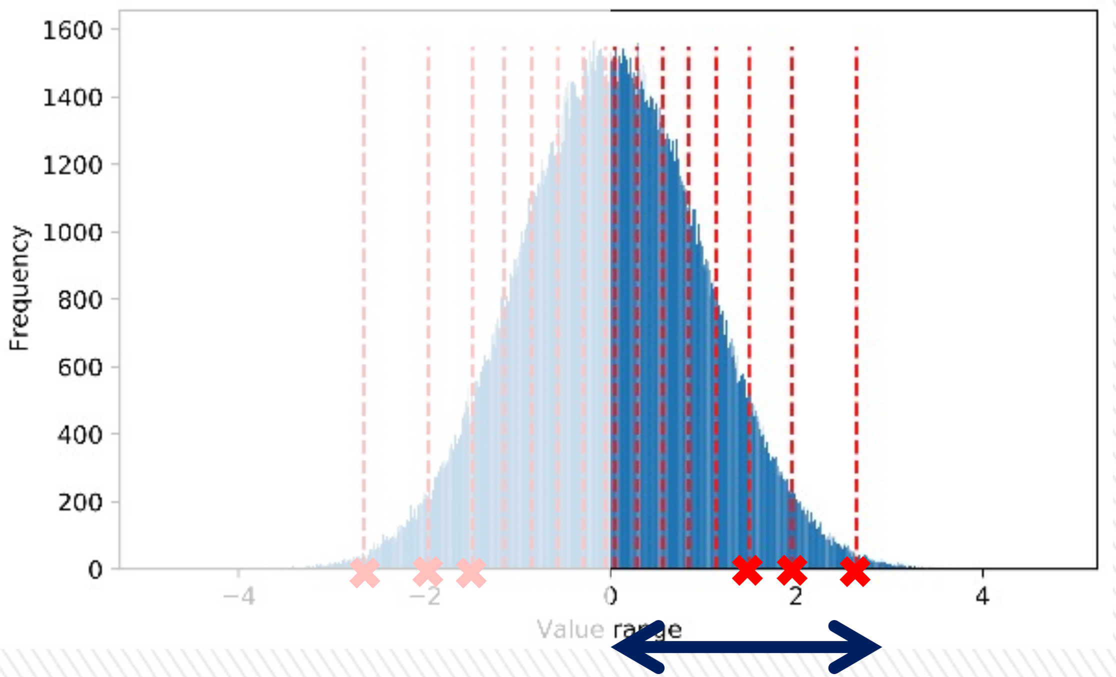
Mokey Quantization

“Simple” Relationship between Index and Value

Mokey Quantization



Symmetrical Dictionary



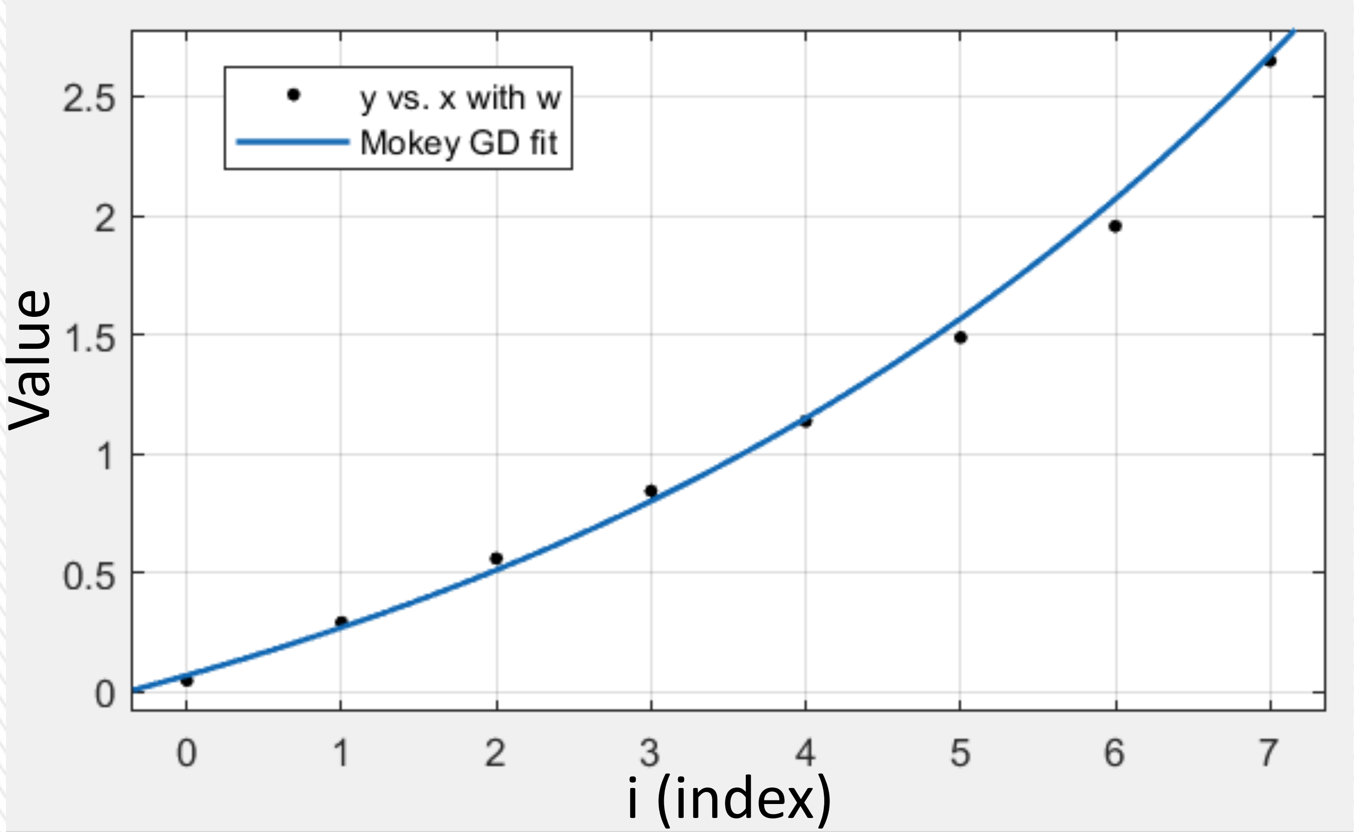
Index	Value
I	0.05
II	0.35
…	…
VI	1.97
VII	2.61

Golden Dict. (GD)

Exponential Function

Index	Value
I	0.05
II	0.35
…	…
VI	1.97
VII	2.6

Golden Dict. (GD)

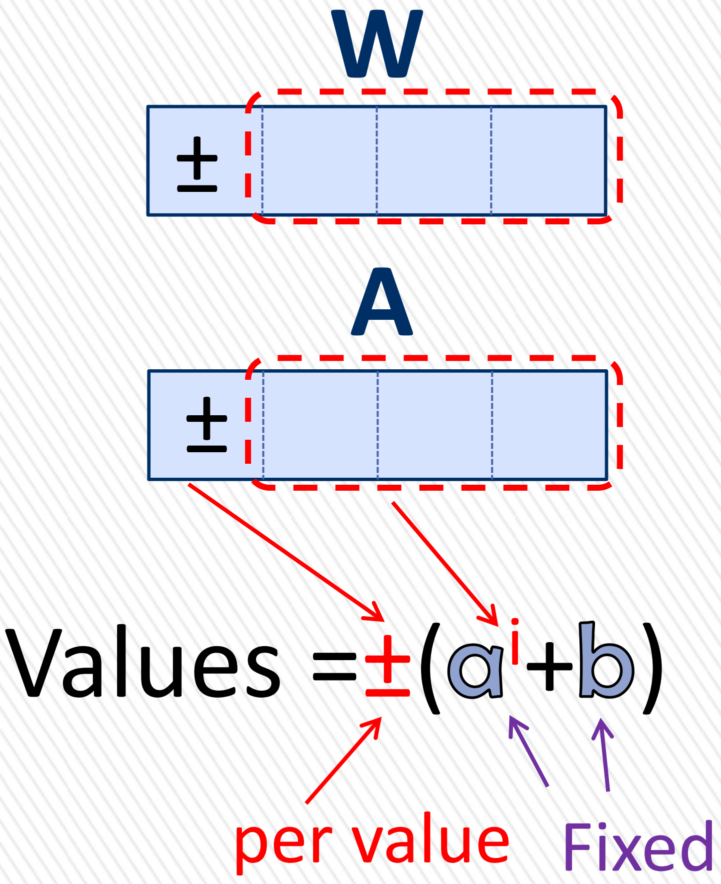


$GD=a^i+b$

Values Format

Index	Value
I	0.05
II	0.35
…	…
VI	1.97
VII	2.6

Golden Dict. (GD)

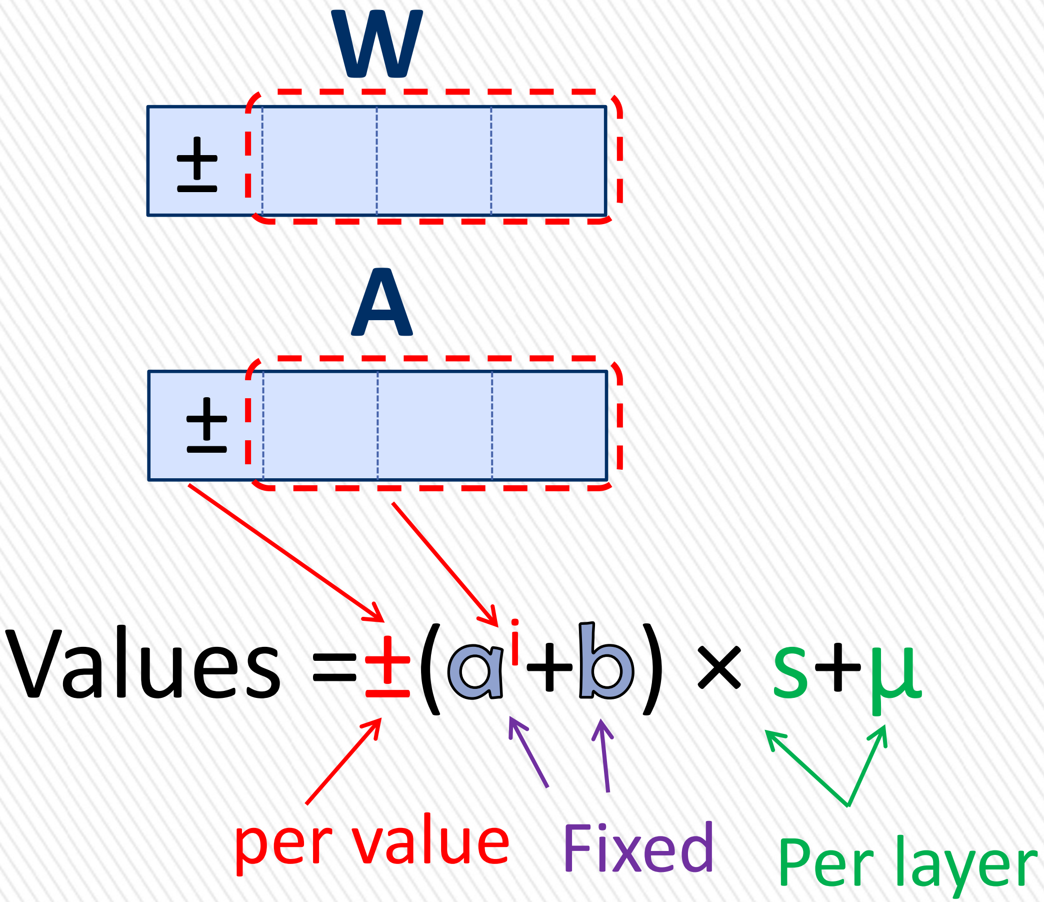


$$\text{Values} = \pm(a^i + b)$$

Values Format

Index	Value
I	0.05
II	0.35
…	…
VI	1.97
VII	2.6

Golden Dict. (GD)



$$\text{Values} = \pm(a^i + b) \times s + \mu$$

per value

Fixed

Per layer

Revisiting Computation

$A_0 = a^1 + b$ $\quad$ $W_0 = a^2 + b$

$A_1 = a^4 + b$ $\quad$ $W_1 = a^5 + b$

… $\quad$ …

Revisiting Computation

$A_0 = a^1 + b \quad W_0 = a^2 + b$

$A_1 = a^4 + b \quad W_1 = a^5 + b$

$\dots \quad \dots$

$$\sum_N AW = A_0W_0 + A_1W_1 + \dots$$

$$= a^{1+2} + ba^1 + ba^2 + b^2 +$$

$$a^{4+5} + ba^4 + ba^5 + b^2 +$$

Revisiting Computation

$$A_0 = a^1 + b \qquad W_0 = a^2 + b$$

$$A_1 = a^4 + b \qquad W_1 = a^5 + b$$

$$\ldots \qquad \ldots$$

$$\begin{aligned}\textstyle\sum_N AW &= A_0W_0 + A_1W_1 + \ldots \\ &= a^{1+2} + ba^1 + ba^2 + b^2 + \\ &\quad\; a^{4+5} + ba^4 + ba^5 + b^2 + \\ &\ldots\end{aligned}$$

Range [0,7]

$$= (a^3 + a^9 + \ldots) + b(a^1 + a^4 + \ldots) + \underbrace{b(a^2 + a^5 + \ldots) + Nb^2}_{\textit{Pre-computed}}$$

Revisiting Computation

$A_0 = a^1 + b \quad W_0 = a^2 + b$

$A_1 = a^4 + b \quad W_1 = a^5 + b$

$\dots \quad \dots$

$$\begin{aligned}\sum_N AW &= A_0W_0 + A_1W_1 + \dots \\ &= a^{1+2} + ba^1 + ba^2 + b^2 + \\ &\quad\ a^{4+5} + ba^4 + ba^5 + b^2 + \\ &\dots\end{aligned}$$

Range [0,7]

$$= (a^3+a^9+\dots) + b\underbrace{(a^1+a^4+\dots)}_{\text{Computed during last layer's quantization}} + b\underbrace{(a^2+a^5+\dots)}_{\text{Pre-computed}} + Nb^2$$

Revisiting Computation

$$A_0 = a^1 + b \qquad W_0 = a^2 + b$$

$$A_1 = a^4 + b \qquad W_1 = a^5 + b$$

$$\dots \qquad \dots$$

$$\textstyle\sum_N AW = A_0W_0 + A_1W_1 + \dots$$

$$= a^{1+2} + ba^1 + ba^2 + b^2 +$$

$$a^{4+5} + ba^4 + ba^5 + b^2 +$$

$$\dots$$

Range [0,7]+[0,7]=[0,14]

Range [0,7]

$$= \underbrace{(a^3+a^9+\dots)}_{\text{(1) Histogram (2) Weighted reduction}}+b\underbrace{(a^1+a^4+\dots)}_{\text{Computed during last layer's quantization}}+ b\underbrace{(a^2+a^5+\dots)+Nb^2}_{\text{Pre-computed}}$$

Revisiting Computation

Histogram: Most of the computation => 3-bit INT Add

Reduction: Post processing => 16-bit Fixed-point

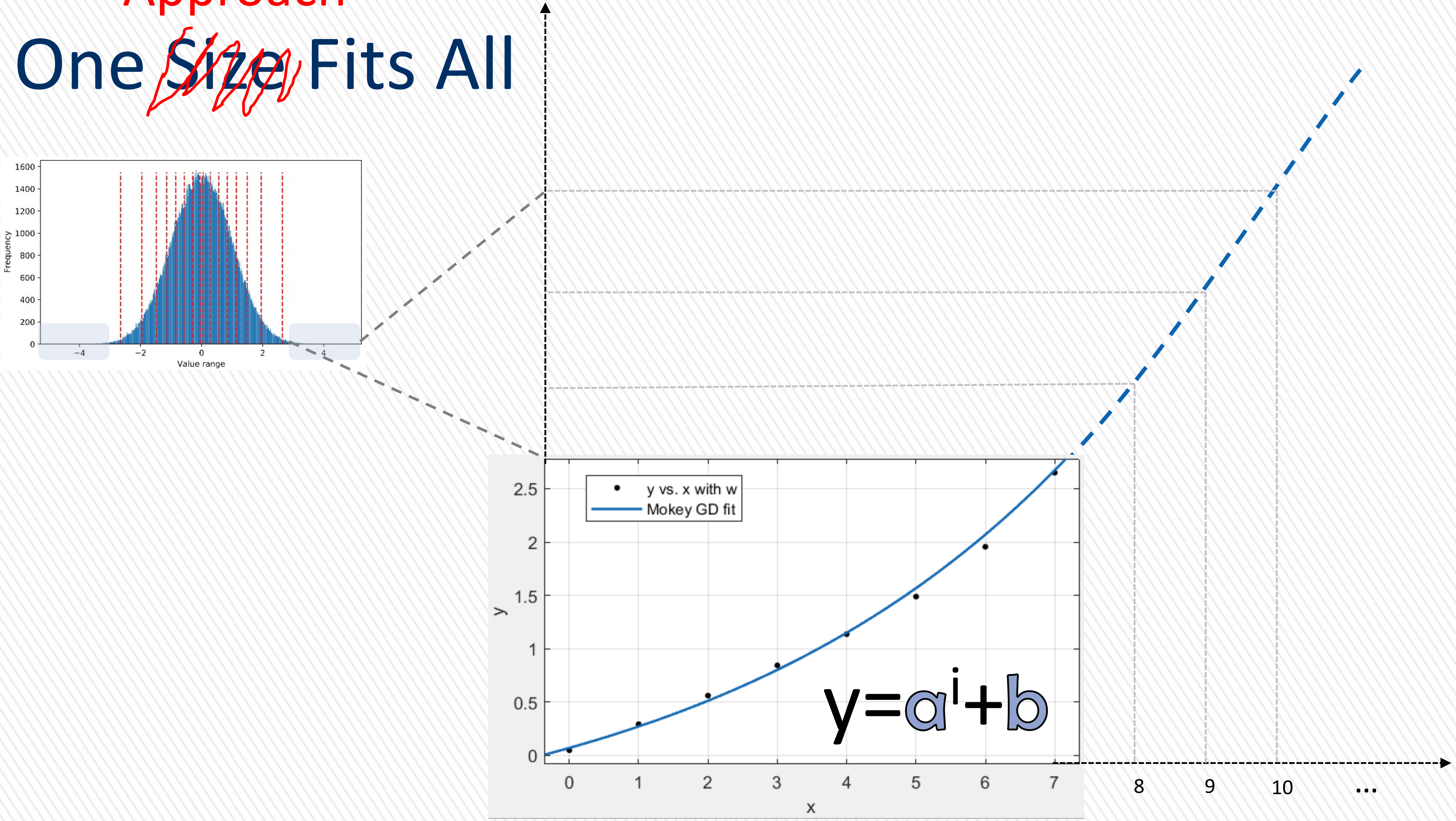
Range [0,7]+[0,7]=[0,14] •••

Range [0,7]

$$= (\underbrace{a^3+a^9+\ldots}_{\text{(1) Histogram, (2) Weighted reduction}})+b(\underbrace{a^1+a^4+\ldots}_{\text{Computed during last layer's quantization}})+ b(\underbrace{a^2+a^5+\ldots}_{\text{Pre-computed}})+Nb^2$$

Approach

One ~~Size~~ Fits All



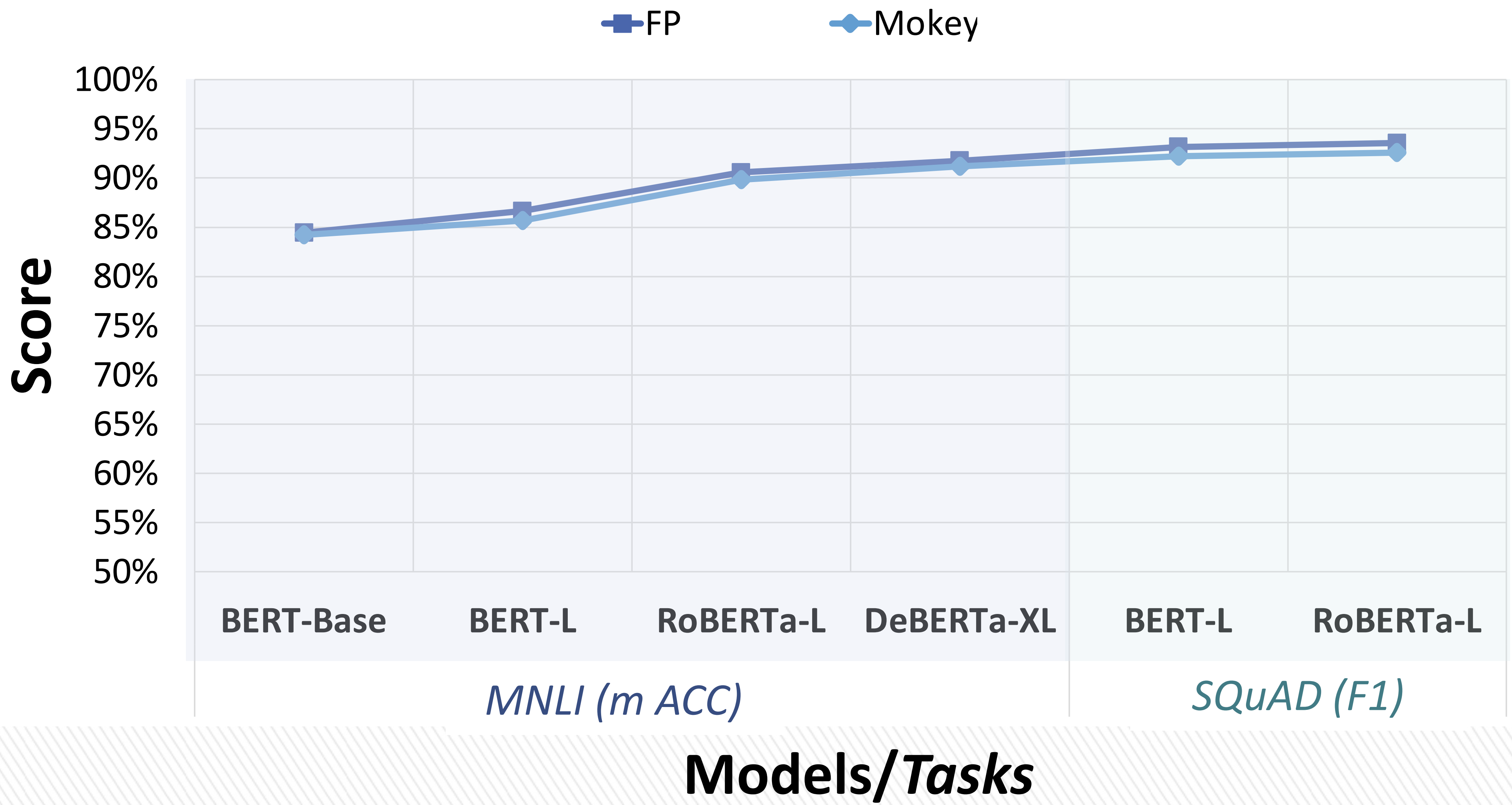
$y=a^i+b$

Evaluation

- FP16 Tensor Cores baseline
- Wide range of on-chip buffers
- 110M - 750M parameter models

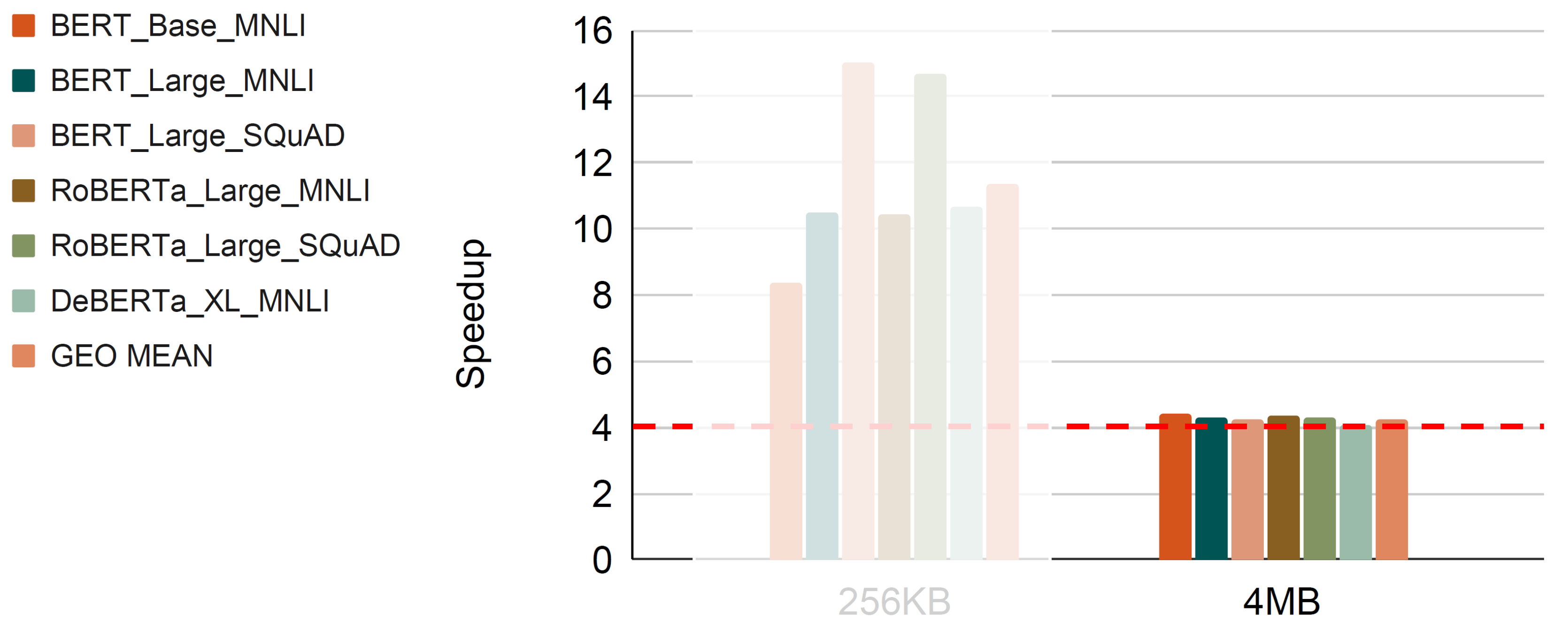
- Custom cycle accurate simulator.
  - DRAMsim3: Dual Channel DDR4-3200
- On-chip Memory: CACTI
- Synthesis: Synopsis Design Compiler
  - 65nm TSMC – 1Ghz
- Layout: Cadence Innovus
- Signal Activity: Modelsim
- Power Estimation: Cadence Innovus

Quantization Accuracy



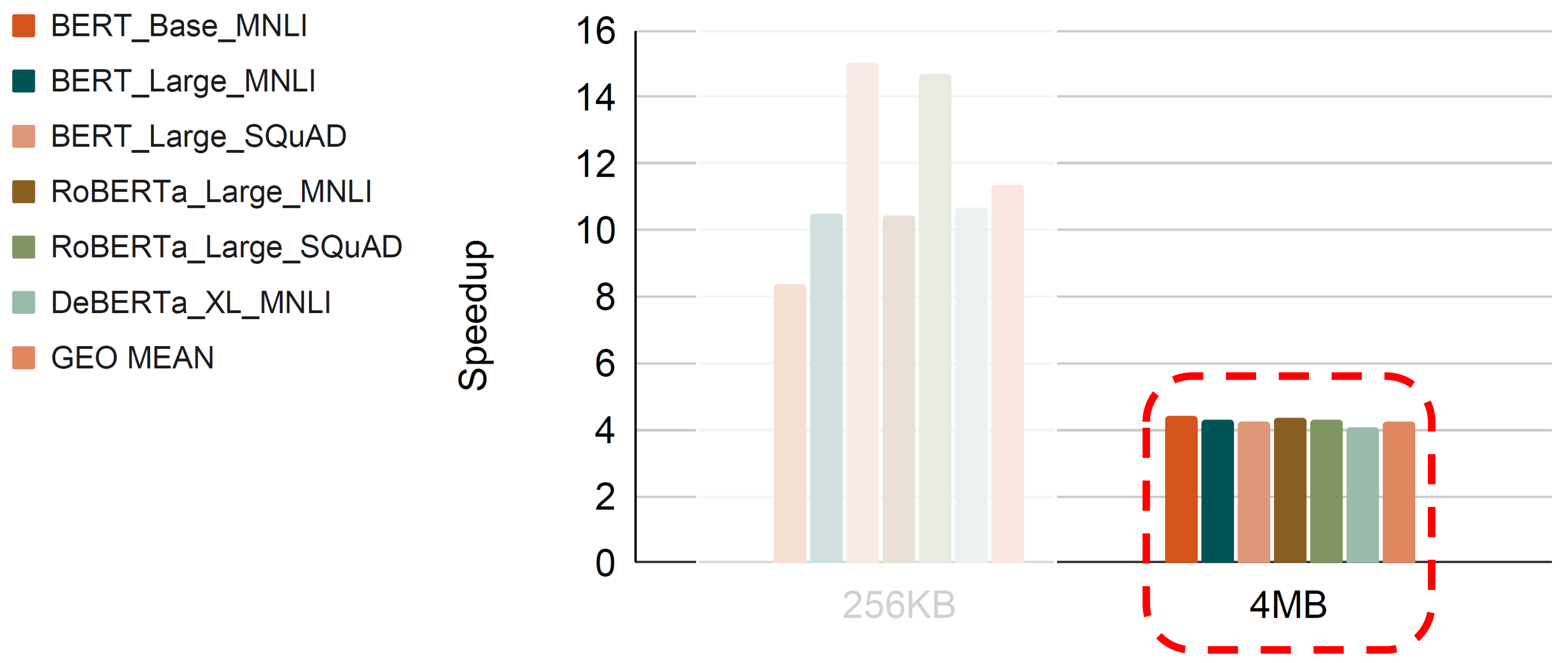
Accelerator Performance

Compression: 16-bit to 4-bit => 4x



Accelerator Performance

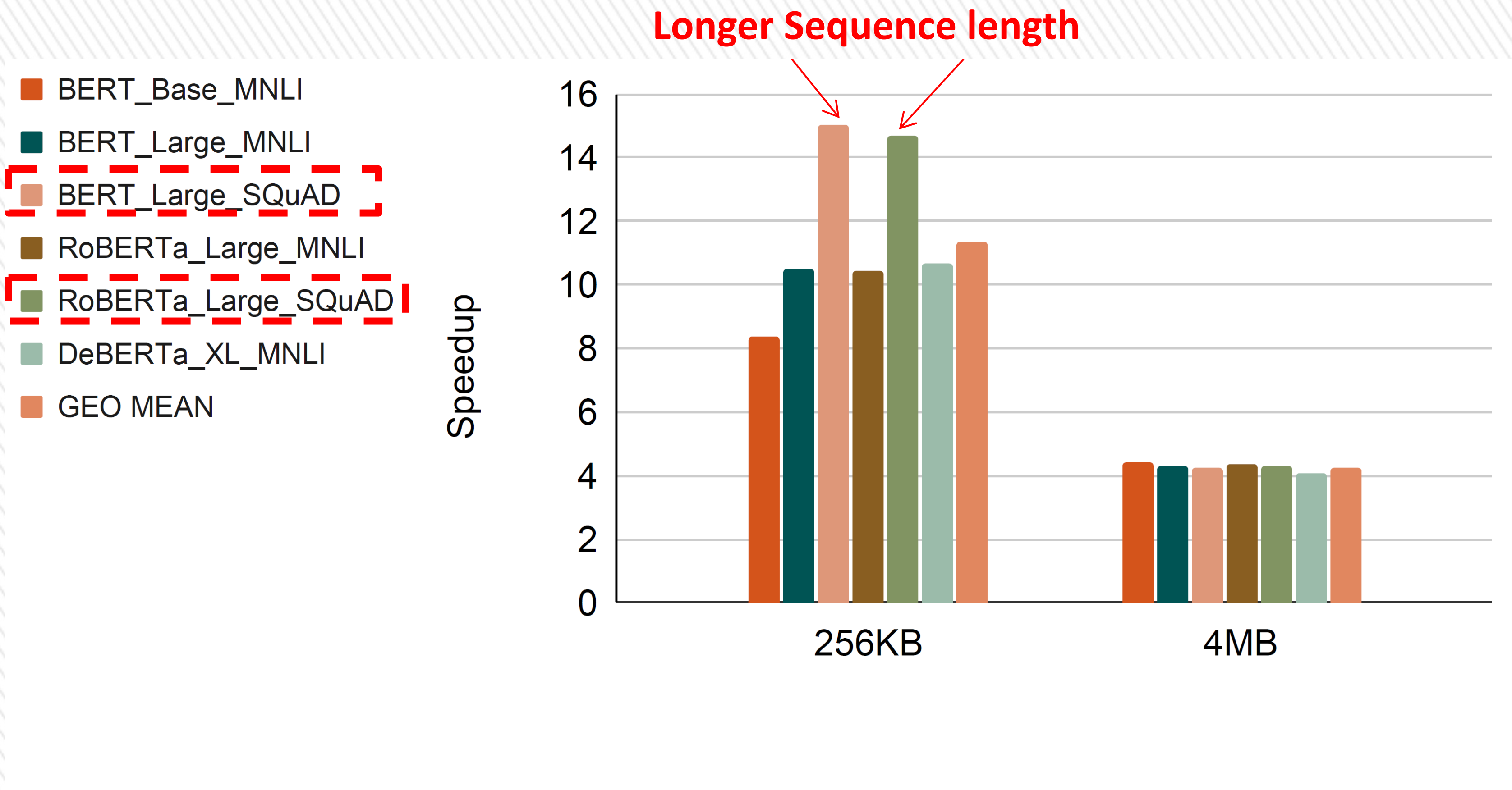
Compression: 16-bit to 4-bit => 4x



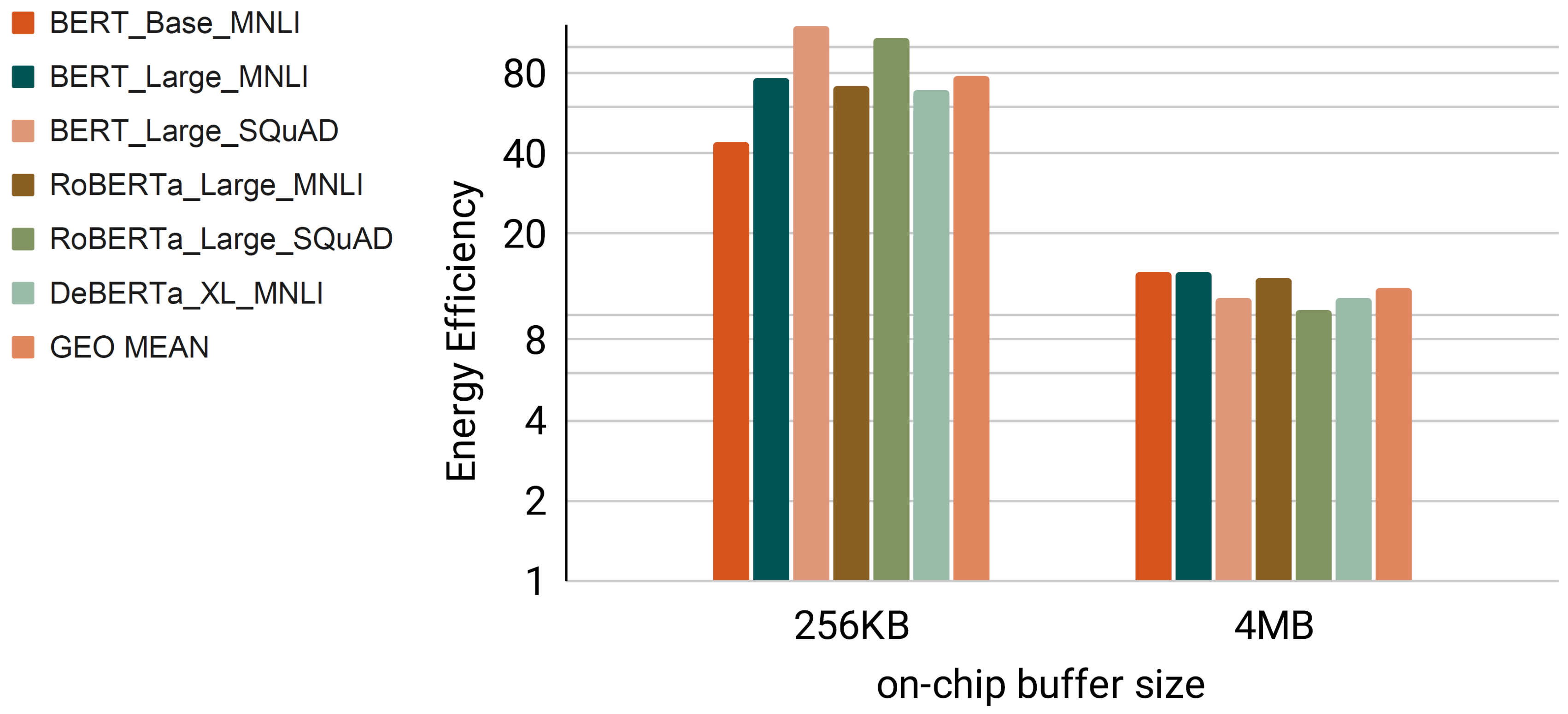
Baseline can almost fit activations on-chip

Accelerator Performance

Compression: 16-bit to 4-bit => 4x

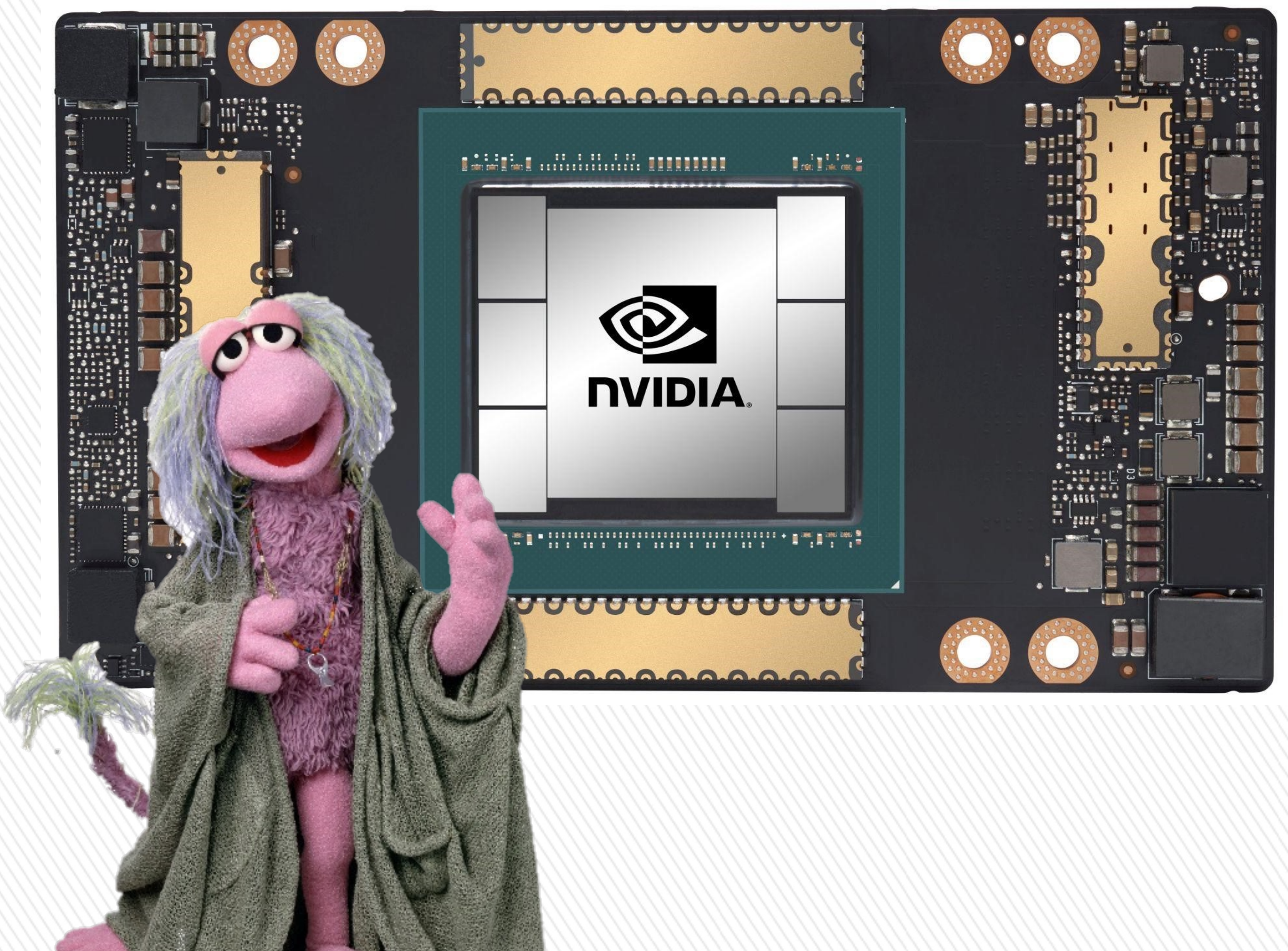


Accelerator Energy Efficiency



Memory Compression and more in paper ☺

Memory Compression



Conclusion

- Mokey Quantization:
  - 4-bit Dictionary-based
  - Focus on a subspace => closed-form representation
  - Fixed-point compute
  - No fine-tuning

- Mokey HW Accelerator:
  - Compute directly on indices
  - 1.6x Smaller tiles vs. Tensor Cores
  - 15x Faster and 100x Energy efficient
  - Can be adapted to other accelerators

Ali Hadi Zadeh
hadizade@ece.utoronto.ca