Accelerating Attention Based Models via HW-SW Co-Design using Fine-Grained Sparsification.

Abhimanyu Bambhaniya[1], Amir Yazdanbakhsh[2], Suvinay Subramanian[3], and Tushar Krishna[1]

[1]Georgia Institute of Technology [2]Google [3]Google Deepmind

Email: abambhaniya3@gatech.edu

Google Research

Georgia Tech

Outline

- Motivation
- Fine-Grained Sparsification (FiGS)
- FIGS Engine
- Evaluation

Outline

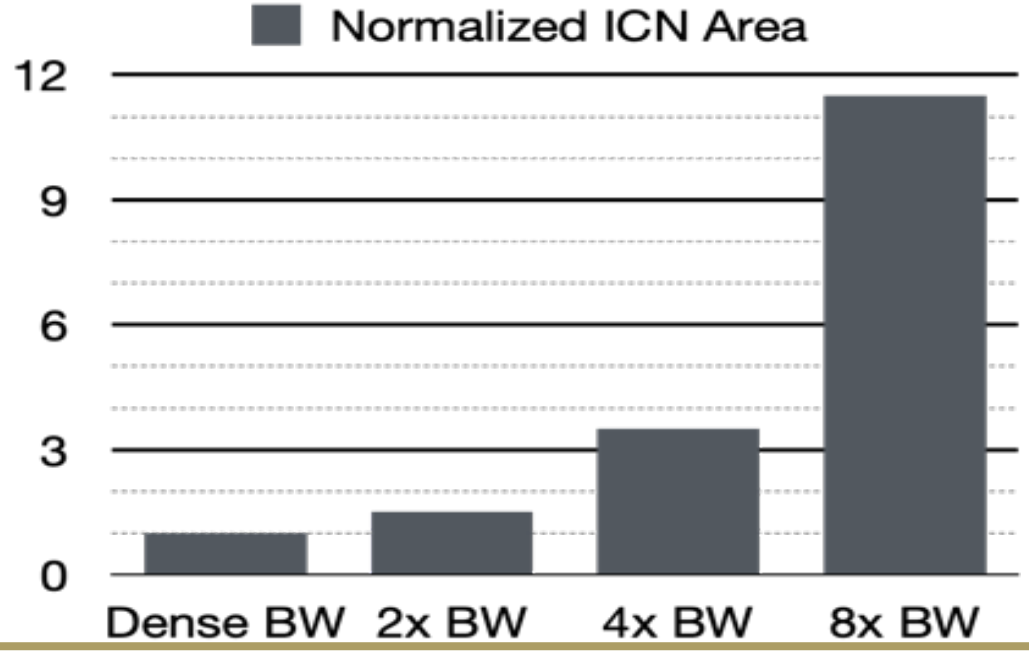
- Motivation
- Fine-Grained Sparsification (FiGS)
- FIGS Engine
- Evaluation

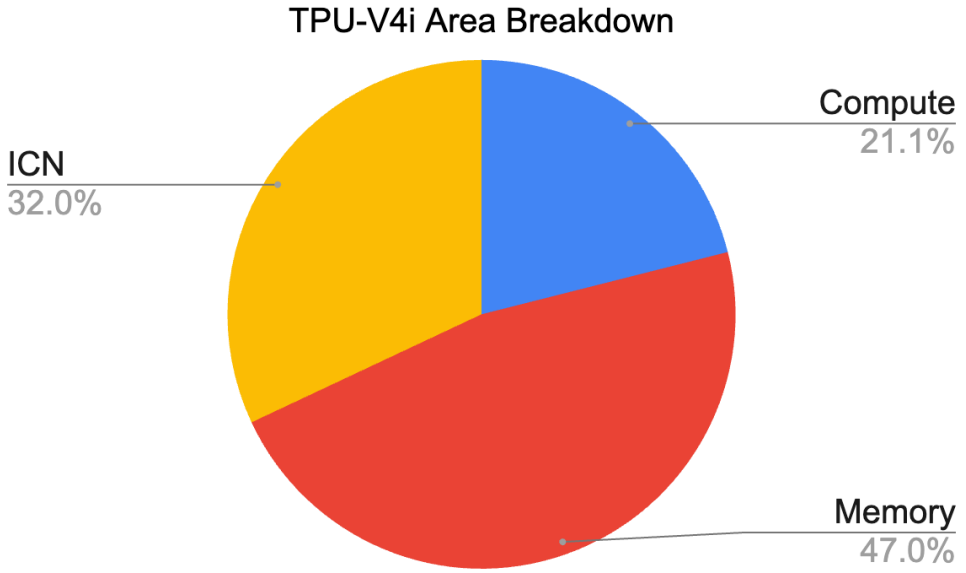
Motivation

- Attention based models are becoming a significant portion of real-world workloads.
- Size of these models is growing at alarming pace.
- We need solution that can help run these models effectively.
- Sparsity is realistic solution that has shown significant promise.
- Nvidia came out with GPU's supporting 2:4 structured sparse workloads, giving path to N:M structured sparse matrices.
- Finding an optimal training recipe to train model with N:M is still an open research question and many people are working on it.

Motivation

- Growing on Nvidia's approach, current SOTA N:M accelerators support high degree of sparsity.
- But supporting these N:M structured sparse matrices, requires higher BW.
- This results in higher area of interconnect, which makes up a significant area of modern chips.





Contributions

- Here are the contributions we make in this work:
  1. Introduce a new training recipe, FiGS, that helps train models to gain better accuracy compared to current SOTA recipes.
  2. A sparse architecture to accelerate variable N:M networks at lower input bandwidth requirement compared to current SOTA sparse architectures.

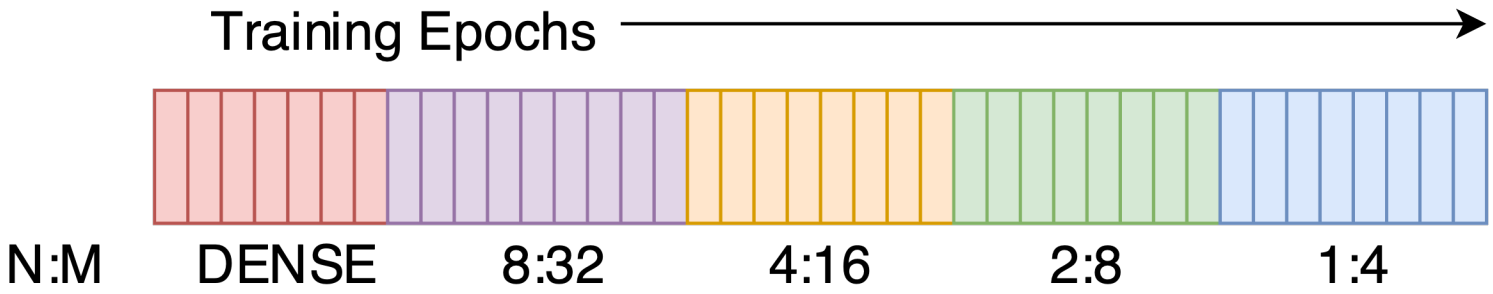
Outline

- Motivation
- Fine-Grained Sparsification (FiGS)
- FIGS Engine
- Evaluation

Fine-Grained Sparsification (FiGS)

Training Recipe:

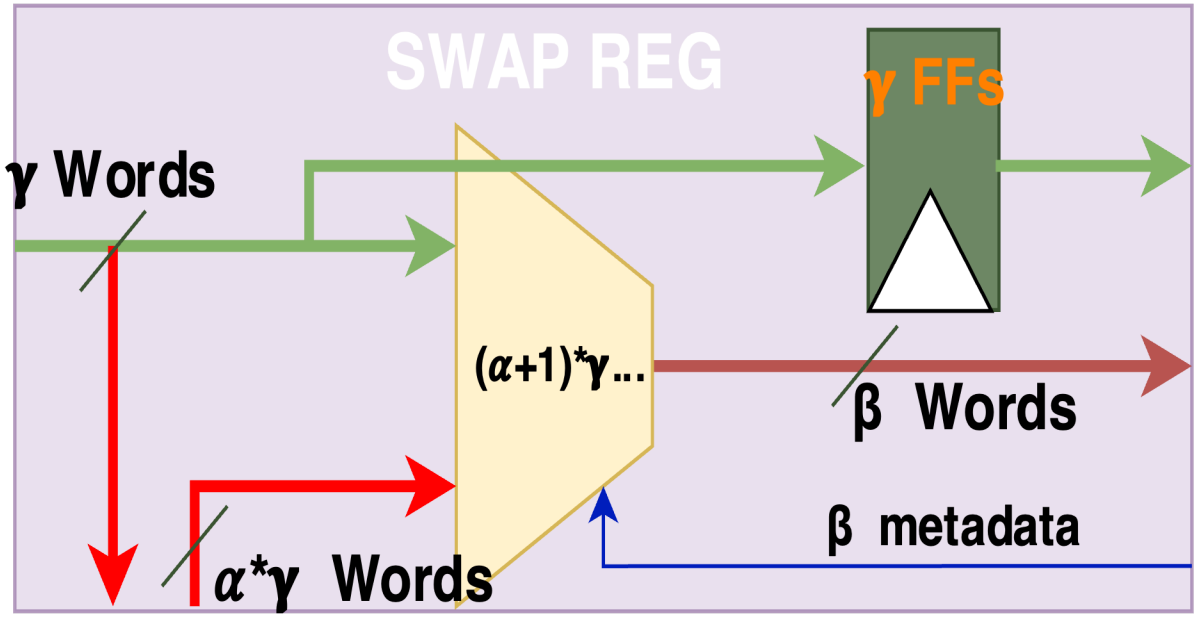
- Derivative of base N:M sparsity
- Intuition: Same # of non-0s in a row.
- Helps the gradients gather locally before pruning which helps achieve better accuracy.
- Fig shows the training schedule for reaching 1:4. Each step would be a subset of the previous sparsification N:M, i.e.1:4 ∈ 2:8 ∈ 4:16 ∈ 8:32.

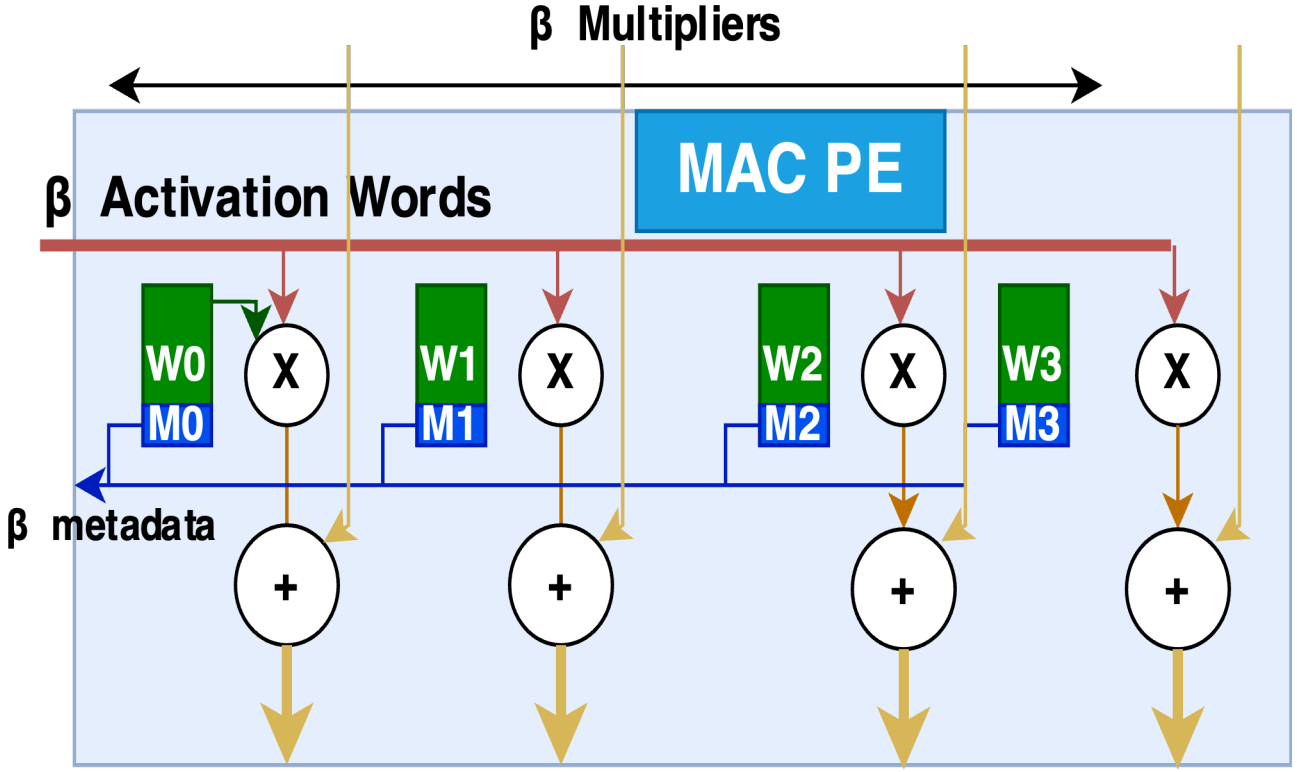


Outline

- Motivation
- Fine-Grained Sparsification (FiGS)
- FIGS Engine
- Evaluation

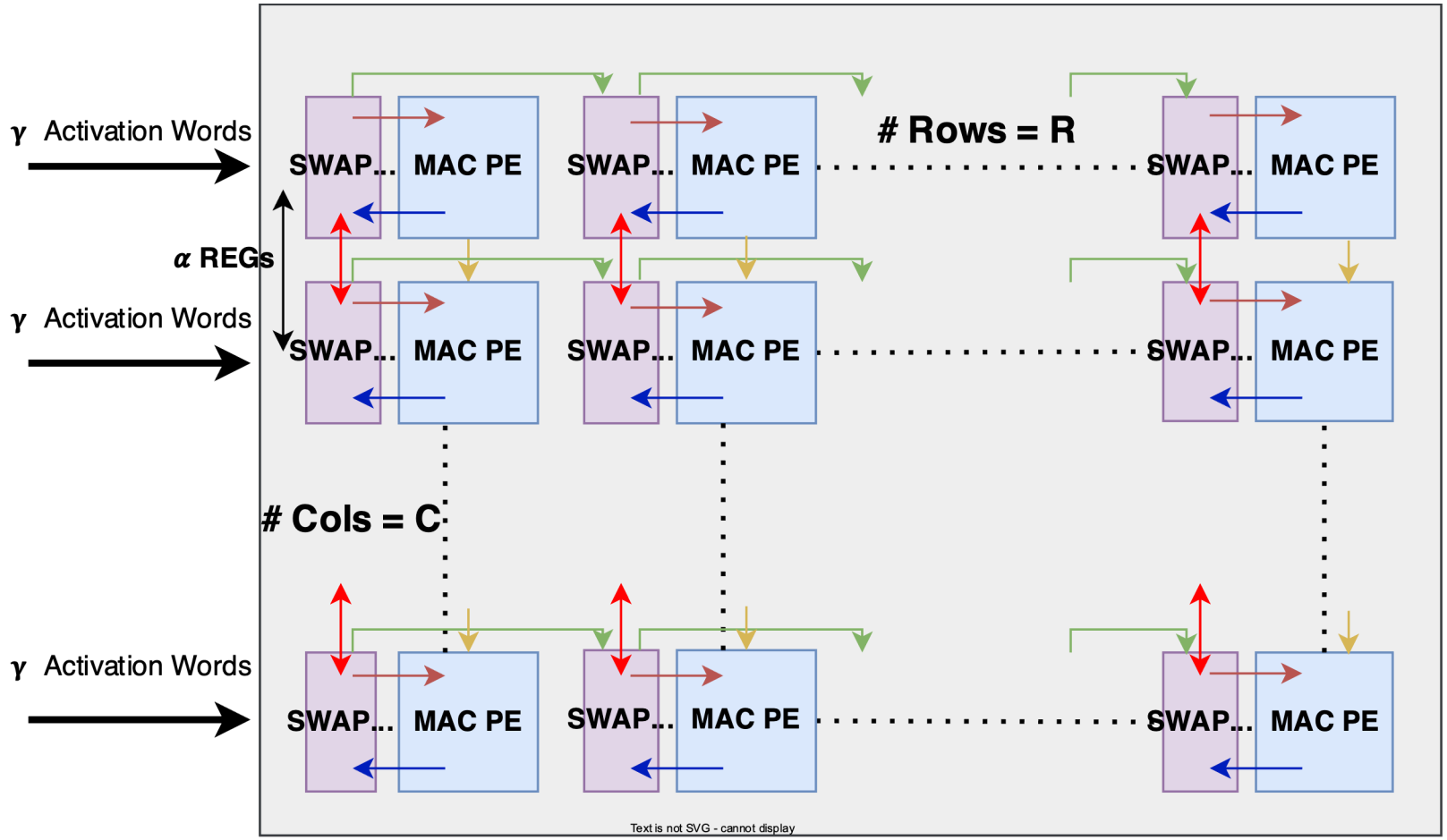
FIGS Architecture - PEs



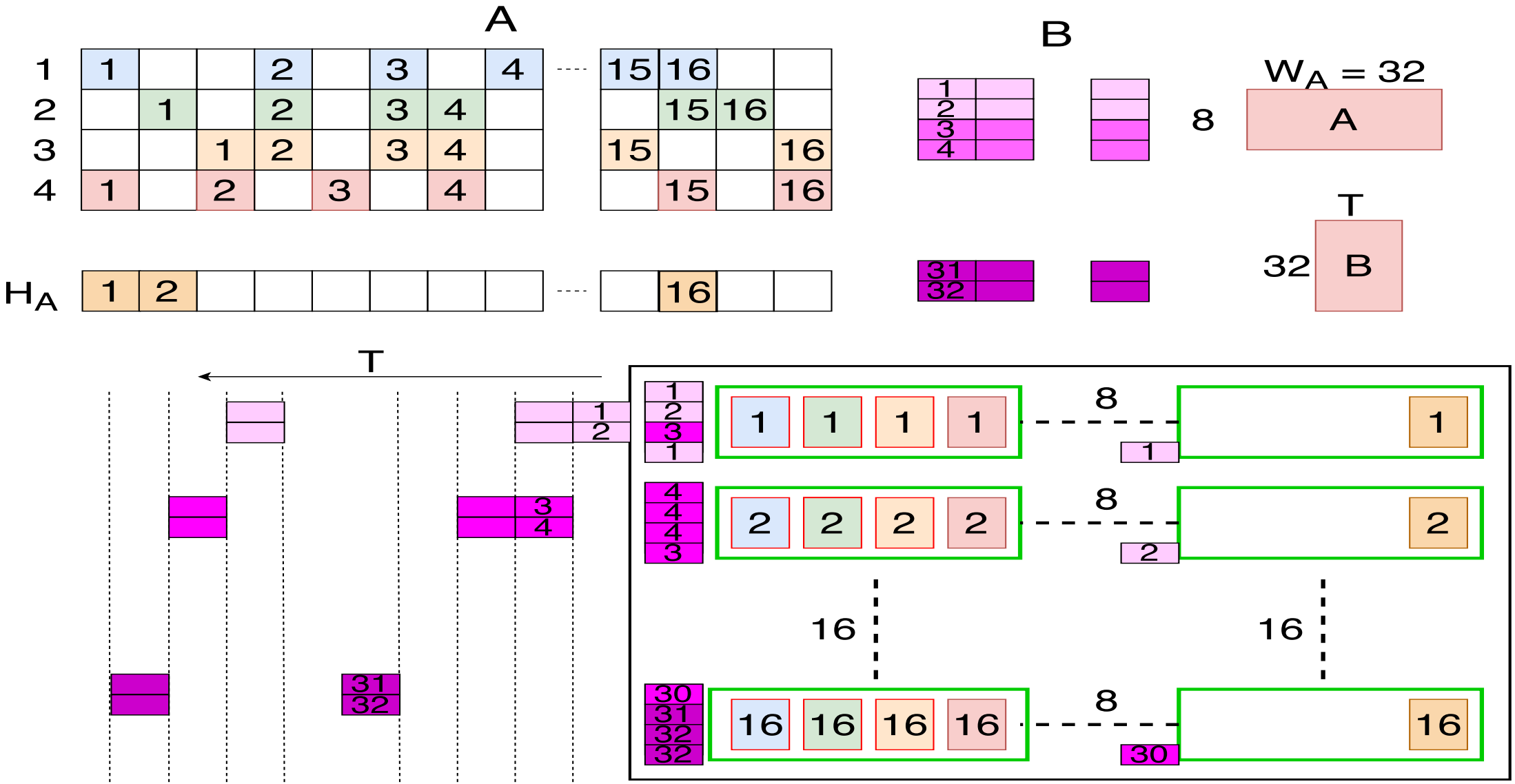


- A **MAC PE** is composed of β MAC units that share the input swap reg. It generates β separate partial sums.
- A **Swap Reg** handles input values coming in from east. They register these inputs and send them to next swap reg.
  - Each swap reg can be connected to α more swap regs in the vertical direction.
  - They can exchange the data among the swap regs for getting correct activation value.

FIGS Engine Architecture



FIGS Engine Dataflow



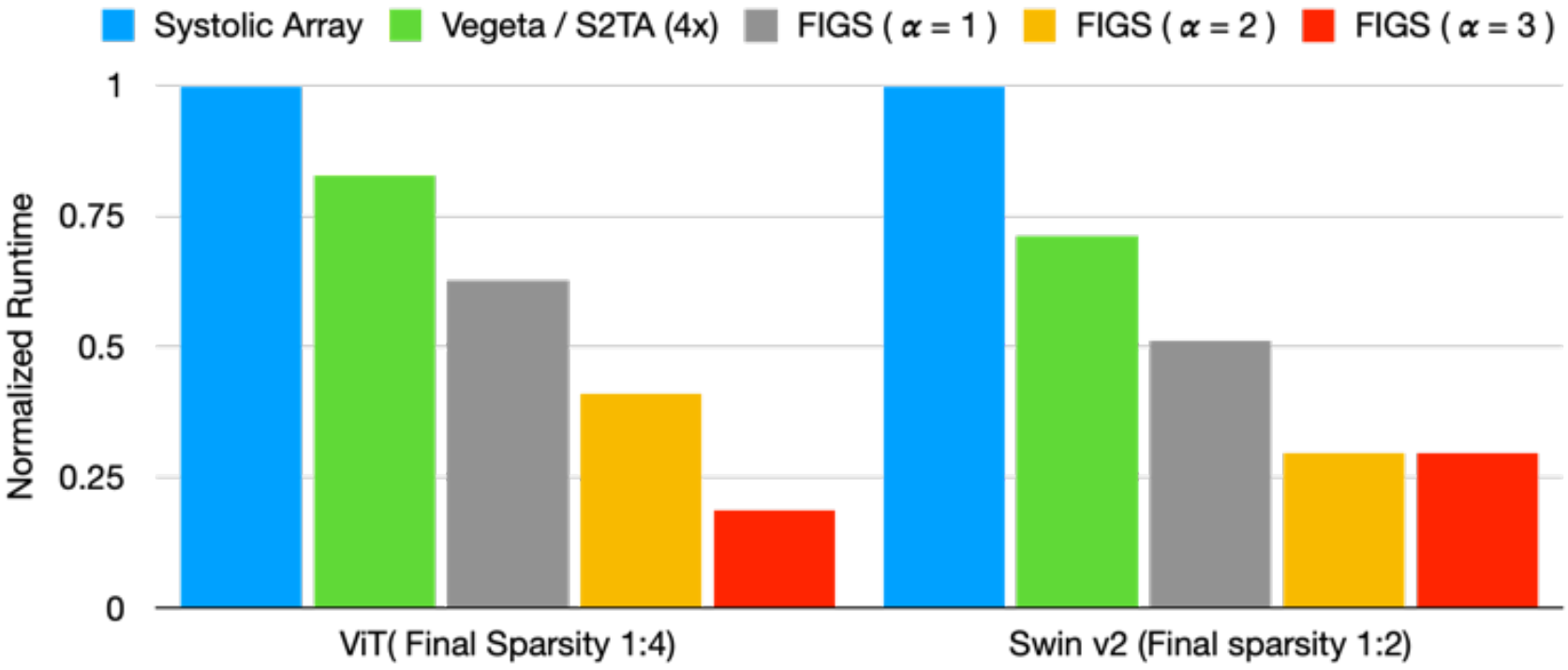
Outline

- Motivation
- Fine-Grained Sparsification (FiGS)
- FIGS Engine
- Evaluation

Fine-Grained Sparsification (FiGS) Evalution

Technique	Target Sparsity	ViT	Swin-V2	T5X-Base
Dense	-	**76.39**	83.45	86.2
SR-STE	1:8 (FF)	77.87	81.44	-
FIGS		**78.17**	**81.47**	-
SR-STE	1:16 (FF)	75.64	80.15	-
FIGS		76.87	**80.27**	-
SR-STE	1:32 (FF)	73.06	78.97	**79.4**
FIGS		**74.91**	**79.28**	79.3
SR-STE	1:8 (FF+QK)	75.53	81.22	75.8
FIGS		**76.33**	**81.44**	**76.8**

FiGS Engine Performance Analysis



- Overall, compared to the state-of-the- art (SOTA) sparse matrix engines, a FIGS engine can provide upto 2.4x and 4.4x speedup when the network are trained to sparsity of 1:2 and 1:4 respectively.

Future Works

- Investigate the efficacy of FIGS-Train in other models.
- Try to extend the uarch to support more kind of sparsity techniques.
- Get performance, area, energy, analysis of FIGS uarch understand its feasibility in realistic implementations.

Conclusion

- We introduce a new training recipe, FiGS that help us train model to better accuracy compared to SOTA training recipe for N:M structured sparse.

- We introduce a new flexible engine that support various N:M ratios at lower BW compared to SOTA N:M sparse engines.

Question? Please send me an email:
abambhaniya3@gatech.edu

Thank you for listening!

If you are interested,
please check out our paper here ☺