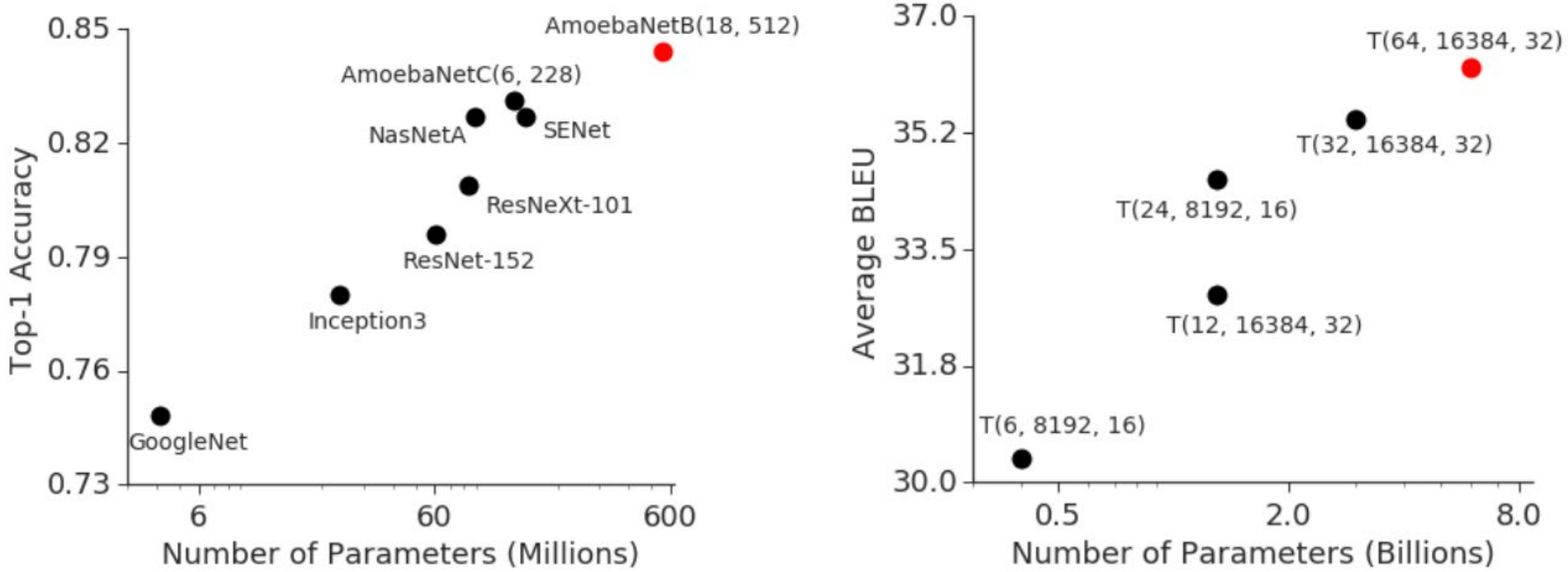
DISTRIBUTED MACHINE LEARNING SYSTEMS EFFICIENCY

PEDRAM ROSTAMI

University of Tehran – School of Electrical and Computer Engineering – DMLS (Dr.Dousti) – 2023 Jan

TABLE OF CONTENTS

LARGE SCALE MODELS ARE BETTER



Gpipe: Efficient training of giant neural networks using pipeline parallelism

LARGE SCALE MODELS 2012 VS 2022

Distbeleif (2012)

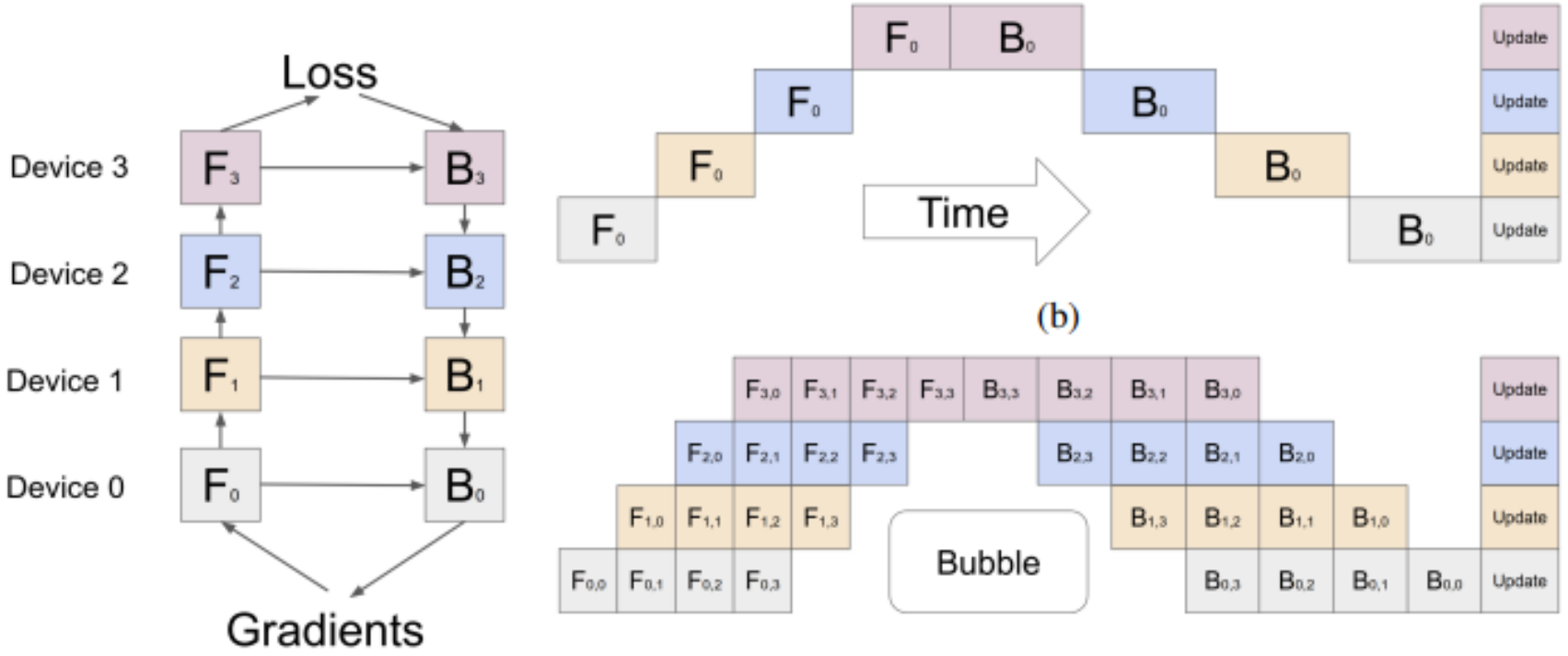
- 1.7B parameters
- Trained on tens of thousands CPU cores

BLOOM (2022)

- 176B parameters
- Trained on 384 A100 80GB GPU (48 nodes)
- GPU memory: 640GB per node
- CPU memory: 512GB per node

ONLY BIG TECH COMPANIES

GPIPE



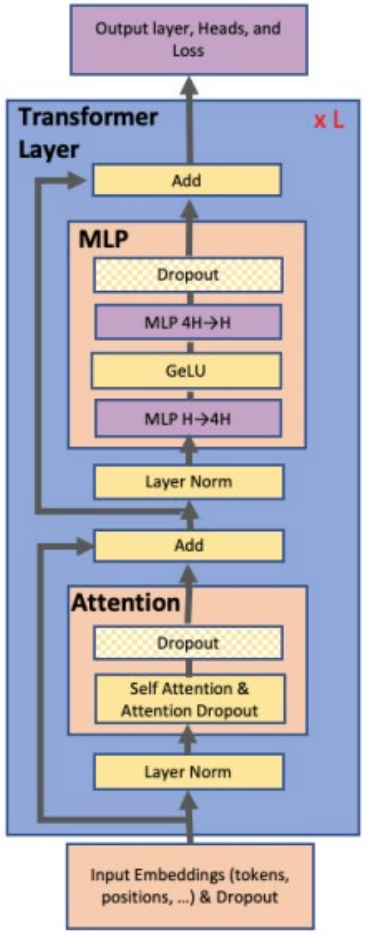
Gpipe: Efficient training of giant neural networks using pipeline parallelism

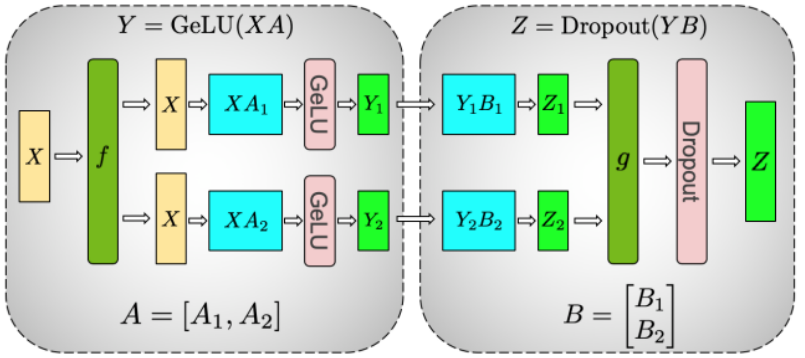
GPIPE

NVIDIA GPUs (8GB each)	Naive-1	Pipeline-1	Pipeline-2	Pipeline-4	Pipeline-8
AmoebaNet-D (L, D)	(18, 208)	(18, 416)	(18, 544)	(36, 544)	(72, 512)
# of Model Parameters	82M	318M	542M	1.05B	1.8B
Total Model Parameter Memory	1.05GB	3.8GB	6.45GB	12.53GB	24.62GB
Peak Activation Memory	6.26GB	3.46GB	8.11GB	15.21GB	26.24GB
Cloud TPUv3 (16GB each)	**Naive-1**	**Pipeline-1**	**Pipeline-8**	**Pipeline-32**	**Pipeline-128**
Transformer-L	3	13	103	415	1663
# of Model Parameters	282.2M	785.8M	5.3B	21.0B	83.9B
Total Model Parameter Memory	11.7G	8.8G	59.5G	235.1G	937.9G
Peak Activation Memory	3.15G	6.4G	50.9G	199.9G	796.1G

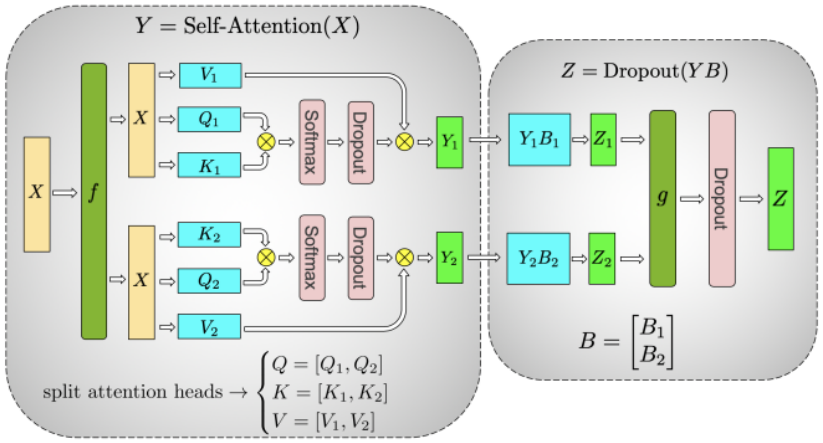
TPU	AmoebaNet			Transformer		
$K =$	2	4	8	2	4	8
$M = 1$	1	1.13	1.38	1	1.07	1.3
$M = 4$	1.07	1.26	1.72	1.7	3.2	4.8
$M = 32$	1.21	1.84	3.48	1.8	3.4	6.3

MEGATRON-LM





(a) MLP



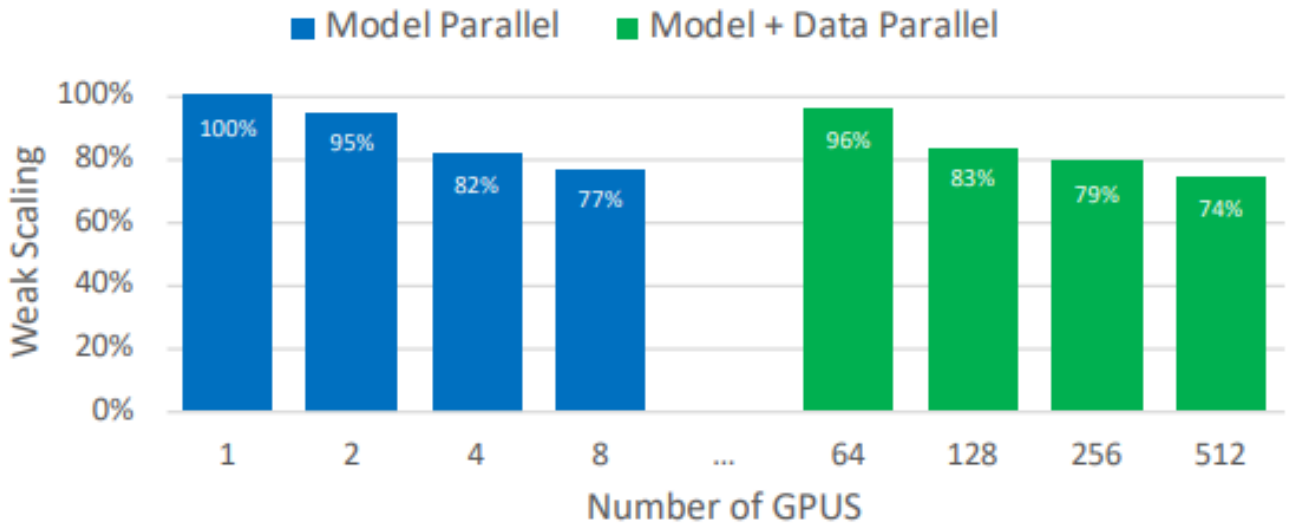
(b) Self-Attention

Megatron-lm: Training multi-billion parameter language models using model parallelism

MEGATRON-LM

- Training GPT-2 models
- All GPUs are V100 32GB
- 64-way data parallelism

Hidden Size	Attention heads	Number of layers	Number of parameters (billions)	Model parallel GPUs	Model +data parallel GPUs
1536	16	40	1.2	1	64
1920	20	54	2.5	2	128
2304	24	64	4.2	4	256
3072	32	72	8.3	8	512



Megatron-lm: Training multi-billion parameter language models using model parallelism

DEEPSPEED (ZERO)

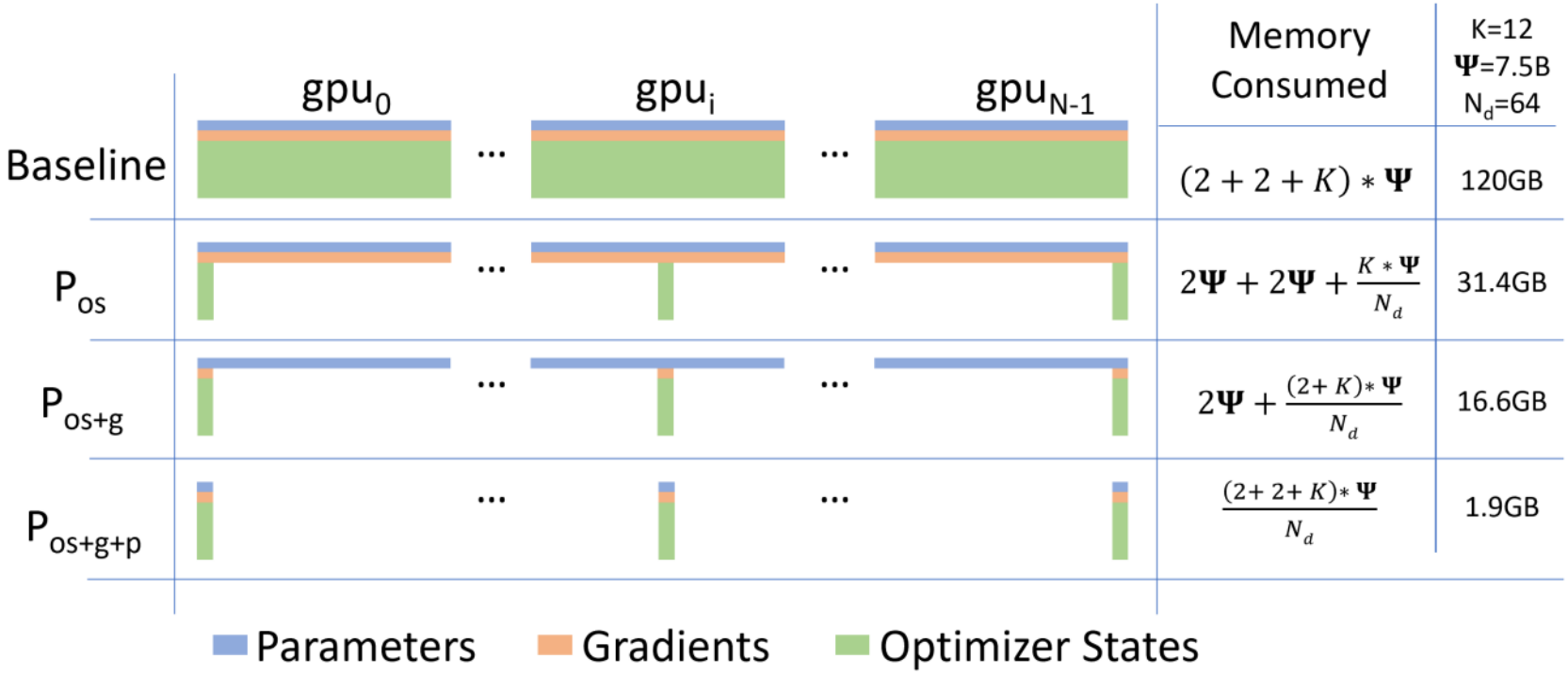
ZeRO-DP

- Optimizer State partitioning
- Gradient partitioning
- Parameter partitioning

ZeRO-R

- Optimize activation memory
- Reduce temporary buffers
- Memory management for preventing memory fragmentation

DEEPSPEED (ZERO)



Zero: Memory optimizations toward training trillion parameter models

DEEPSPEED (ZERO)

MP	GPUs	Max Theoretical Model Size				Measured Model Size	
		Baseline	P_{os}	P_{os+g}	P_{os+g+p}	Baseline	*ZeRO*-DP (P_{os})
1	64	2B	**7.6B**	14.4B	128B	1.3B	**6.2B**
2	128	4B	**15.2B**	28.8B	256B	2.5B	**12.5B**
4	256	8B	**30.4B**	57.6B	0.5T	5B	**25B**
8	512	16B	**60.8B**	115.2B	1T	*10B*	**50B**
16	1024	32B	**121.6B**	230.4B	*2T*	20B	**100B**

Zero: Memory optimizations toward training trillion parameter models

OTHERS

https://huggingface.co/docs/accelerate

https://github.com/facebookresearch/fairscale

COMPARISON

	Released Year	Released Co.	Platform	Community
Gpipe	2019	Google	Tensorflow	Inactive
Megatron-LM	2019	Nvidia	PyTorch	Large
DeepSpeed	2020	Microsoft	PyTorch	Large and Active
fairScale	2022	Meta (Facebook AI)	PyTorch	Small
Accelerate	2022	Hugging Face	Transformers	Small

CONCLUSION

Introduction

- Large Scale Models
- Large Scale DMLSs

Systems

- GPipe
- Megatron-LM
- DeepSpeed
- Others

Comparison

- Compared Systems

REFERENCES

- Dean, Jeffrey, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato et al. "Large scale distributed deep networks." *Advances in neural information processing systems* 25 (2012).
- Huang, Yanping, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, and Yonghui Wu. "Gpipe: Efficient training of giant neural networks using pipeline parallelism." *Advances in neural information processing systems* 32 (2019).
- Shoeybi, Mohammad, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. "Megatron-lm: Training multi-billion parameter language models using model parallelism." *arXiv preprint arXiv:1909.08053* (2019).
- Rajbhandari, Samyam, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. "Zero: Memory optimizations toward training trillion parameter models." In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1-16. IEEE, 2020.
- https://github.com/facebookresearch/fairscale
- https://huggingface.co/docs/accelerate

Thanks for your attention and time!