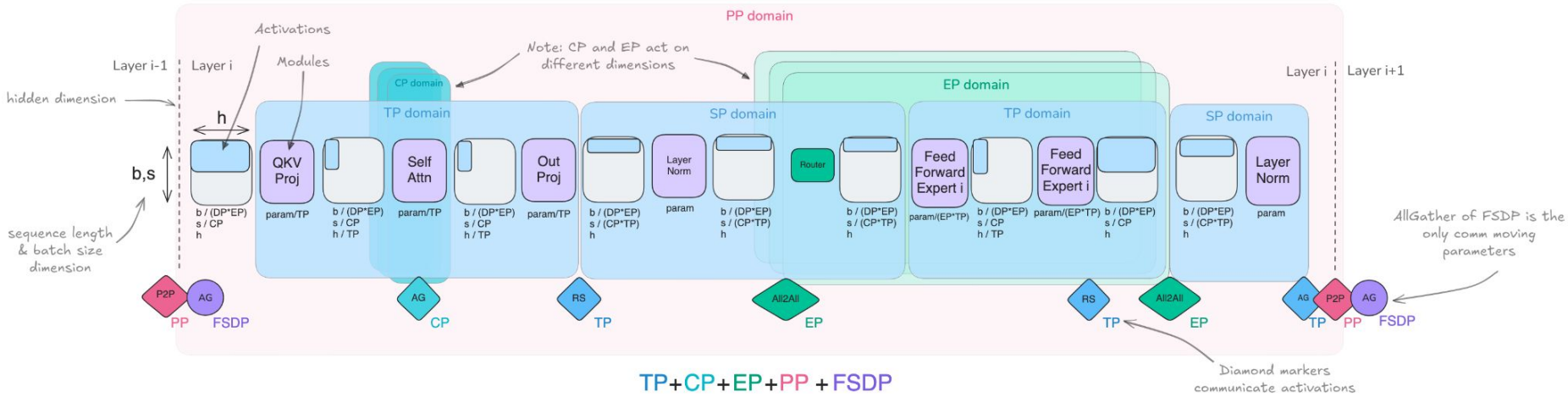
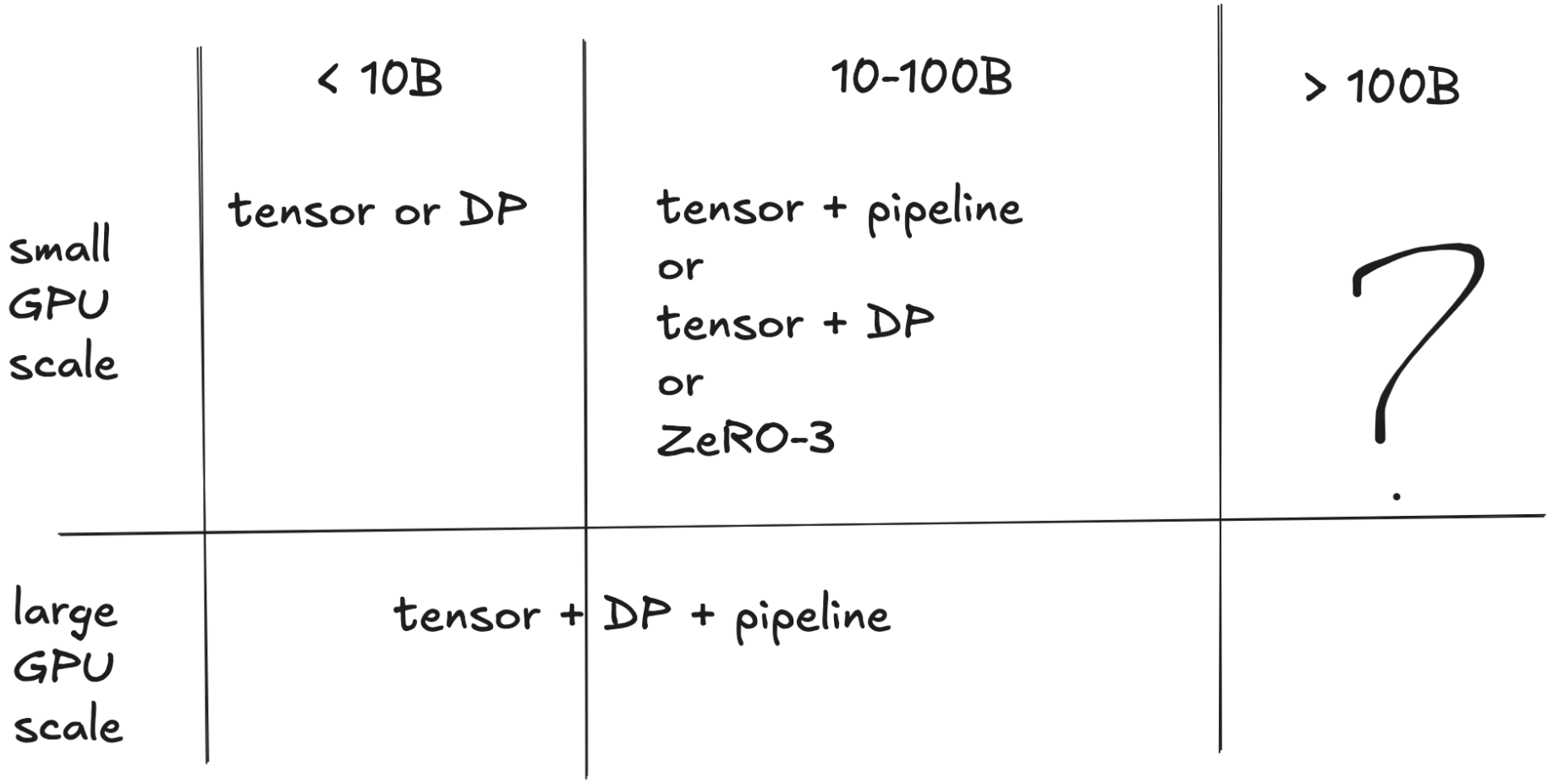
Finding the Best Training Configuration

made with ❤️ for “Little ML book club”



TP+CP+EP+PP +FSDP


GPU-RICH &
EMOTIONALLY
POOR


< 10B
10-100B
> 100B
small GPU scale
tensor or DP
tensor + pipeline
or
tensor + DP
or
ZeRO-3
?
large GPU scale
tensor + DP + pipeline


GPU-POOR &
GPU-POOR & STILL EMOTIONALLY POOR

- Enable recomputation
- Gradient accumulation
- Stop buying coffee out and save for more GPUs

Achieving the target global batch size

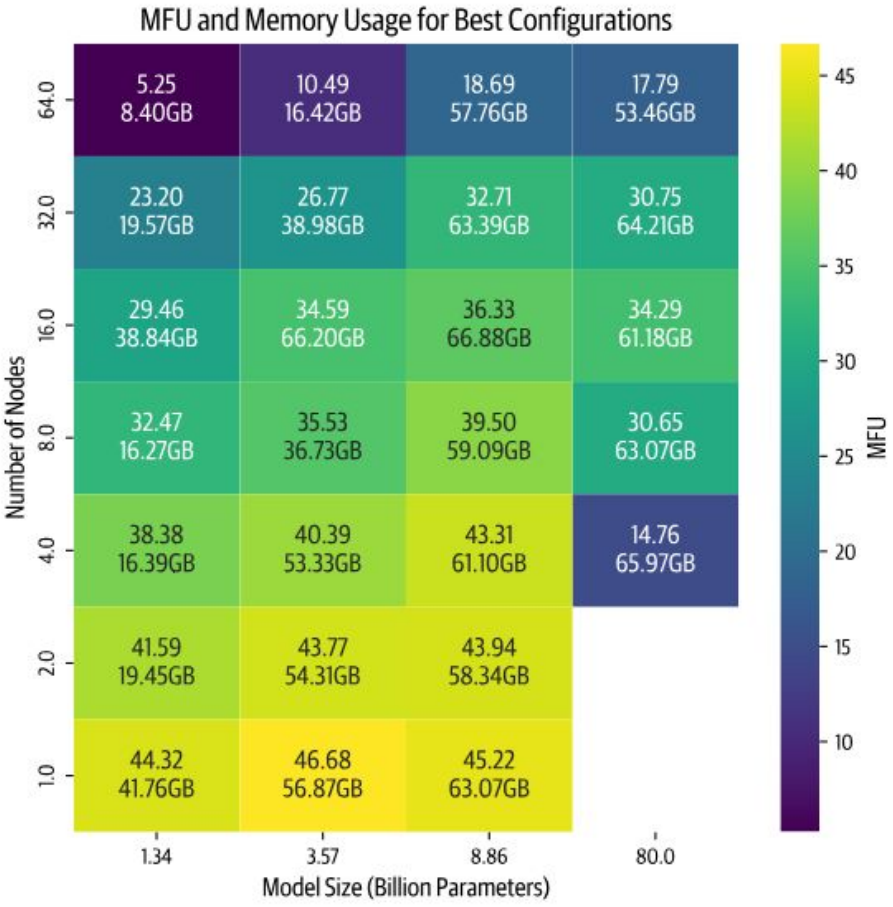
Surprise-surprise

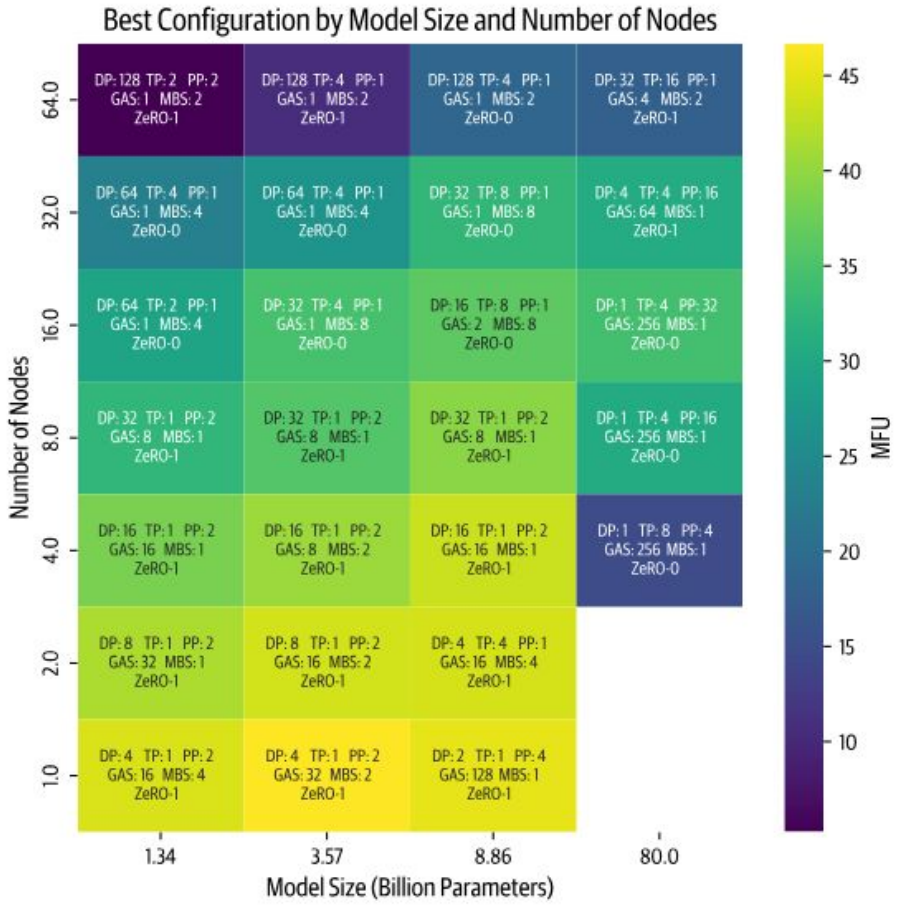
Achieving the target global batch size

Surprise-surprise
DP + gradient accumulation + context

Optimizing training throughput

1. Tensor
2. ZeRO-3
3. DP -> PP
4. Scale one technique at a time
5. Experiment with micro-batch sizes


MFU and Memory Usage for Best Configurations
5.25
8.40GB
10.49
16.42GB
18.69
57.76GB
17.79
53.46GB
23.20
19.57GB
26.77
38.98GB
32.71
63.39GB
30.75
64.21GB
29.46
38.84GB
34.59
66.20GB
36.33
66.88GB
34.29
61.18GB
32.47
16.27GB
35.53
36.73GB
39.50
59.09GB
30.65
63.07GB
38.38
16.39GB
40.39
53.33GB
43.31
61.10GB
14.76
65.97GB
41.59
19.45GB
43.77
54.31GB
43.94
58.34GB
44.32
41.76GB
46.68
56.87GB
45.22
63.07GB
Number of Nodes
64.0
32.0
16.0
8.0
4.0
2.0
1.0
1.34
3.57
8.86
80.0
Model Size (Billion Parameters)
MFU
45
40
35
30
25
20
15
10


Best Configuration by Model Size and Number of Nodes
DP: 128 TP: 2 PP: 2
GAS: 1 MBS: 2
ZeRO-1
DP: 128 TP: 4 PP: 1
GAS: 1 MBS: 2
ZeRO-1
DP: 128 TP: 4 PP: 1
GAS: 1 MBS: 2
ZeRO-0
DP: 32 TP: 16 PP: 1
GAS: 4 MBS: 2
ZeRO-1
DP: 64 TP: 4 PP: 1
GAS: 1 MBS: 4
ZeRO-0
DP: 64 TP: 4 PP: 1
GAS: 1 MBS: 4
ZeRO-0
DP: 32 TP: 8 PP: 1
GAS: 1 MBS: 8
ZeRO-0
DP: 4 TP: 4 PP: 16
GAS: 64 MBS: 1
ZeRO-1
DP: 64 TP: 2 PP: 1
GAS: 1 MBS: 4
ZeRO-0
DP: 32 TP: 4 PP: 1
GAS: 1 MBS: 8
ZeRO-0
DP: 16 TP: 8 PP: 1
GAS: 2 MBS: 8
ZeRO-0
DP: 1 TP: 4 PP: 32
GAS: 256 MBS: 1
ZeRO-0
DP: 32 TP: 1 PP: 2
GAS: 8 MBS: 1
ZeRO-1
DP: 32 TP: 1 PP: 2
GAS: 8 MBS: 1
ZeRO-1
DP: 32 TP: 1 PP: 2
GAS: 8 MBS: 1
ZeRO-1
DP: 1 TP: 4 PP: 16
GAS: 256 MBS: 1
ZeRO-0
DP: 16 TP: 1 PP: 2
GAS: 16 MBS: 1
ZeRO-1
DP: 16 TP: 1 PP: 2
GAS: 8 MBS: 2
ZeRO-1
DP: 16 TP: 1 PP: 2
GAS: 16 MBS: 1
ZeRO-1
DP: 1 TP: 8 PP: 4
GAS: 256 MBS: 1
ZeRO-0
DP: 8 TP: 1 PP: 2
GAS: 32 MBS: 1
ZeRO-1
DP: 8 TP: 1 PP: 2
GAS: 16 MBS: 2
ZeRO-1
DP: 4 TP: 4 PP: 1
GAS: 16 MBS: 4
ZeRO-1
DP: 4 TP: 1 PP: 2
GAS: 16 MBS: 4
ZeRO-1
DP: 4 TP: 1 PP: 2
GAS: 32 MBS: 2
ZeRO-1
DP: 2 TP: 1 PP: 4
GAS: 128 MBS: 1
ZeRO-1
Number of Nodes
64.0
32.0
16.0
8.0
4.0
2.0
1.0
1.34
3.57
8.86
80.0
Model Size (Billion Parameters)
MFU
45
40
35
30
25
20
15
10

1. Meta LLaMA Family

Model	Date	Parameters	Hardware	TP	PP	DP/ FSDP	CP	EP	Key Innovations
LLaMA 1	Feb 2023	7B–65B	2,048 A100 80GB	—	—	DP	—	—	Basic data parallelism; RSC cluster
LLaMA 2	Jul 2023	7B–70B	RSC + prod clusters	—	—	FSDP	—	—	Introduced FSDP; GQA for 70B; 4K context; 1.73M GPU-hours for 70B
LLaMA 3	Apr 2024	8B–70B	16,384 H100	8	16	FSDP (128)	1	—	4D parallelism; 8K context; 126 layers (not 128) for balanced PP
LLaMA 3.1	Jul 2024	8B–405B	16,384 H100	8	16	FSDP (128)	1–16	—	128K context via CP=16; all-gather CP (not ring attention); 38-43% MFU

GPUs	TP	CP	PP	DP	Seq. Len.	Batch size/DP	Tokens/Batch	TFLOPs/GPU	BF16 MFU
8,192	8	1	16	64	8,192	32	16M	430	43%
16,384	8	1	16	128	8,192	16	16M	400	41%
16,384	8	16	16	8	131,072	16	16M	380	38%

Table 4 Scaling configurations and MFU for each stage of Llama 3 405B pre-training. See text and Figure 5 for descriptions of each type of parallelism.

2. Google PaLM/Gemini Family

Model	Date	Parameters	Hardware	TP	PP	DP	CP	EP	Key Innovations
PaLM	Apr 2022	540B	6,144 TPU v4	12	**None**	256 (2D FSDP)	—	—	Pipeline-free; 57.8% HW utilization; Pathways system
PaLM 2	May 2023	Undisclosed	TPU v4	✓	—	✓	—	✓ (sparse)	MoE architecture; improved compute-optimal scaling
Gemini 1.0 Ultra	Dec 2023	Undisclosed	Multi-DC TPU v4/ v5e	✓	—	✓	—	✓	**Multi-datacenter training**; 97% goodput; optical circuit switching
Gemini 1.5 Pro	Feb 2024	Undisclosed (MoE)	TPU v5+	✓	—	✓	✓	✓	Sparse MoE; up to 1M context; long-context specialization

3. DeepSeek Family

Model	Date	Total Params	Active Params	Hardware	TP	PP	DP	EP	Key Innovations
DeepSeek 67B	Jan 2024	67B	67B (dense)	H800 cluster	✓	✓	ZeRO	—	Baseline dense model
DeepSeek-V2	May 2024	236B	21B	H800 cluster	—	16 (ZeroBubble)	ZeRO-1	8	MLA attention; DeepSeekMoE; 42.5% cost reduction vs 67B
DeepSeek-V3	Dec 2024	671B	37B	2,048 H800	**None**	16 (DualPipe)	ZeRO-1	64	Aux-loss-free balancing; FP8; **$5.6M total cost**; 180K GPU-hr/T tokens

see you next time