deepspeed

Microsoft

DeepSpeed-MoE: Advancing MoE inference & training to power next generation AI scale

ICML 2022

Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, Yuxiong He.
Microsoft DeepSpeed Team.

AI Scale is limited by Compute

- Compute is the primary challenge of training massive models
- Ambitious Model Scale and Time to Train

Model	Model Size	Hardware	Days to Train
Megatron-LM GPT-2	8.3B	512 V100 GPU	9.2 days
OPT	175B	992 A100 GPU	56 days
MT-NLG	530B	2200 A100 GPU	60 days
PaLM	540B	6144 TPU v4	57 days

- Next jump in scale:
  - next generation of hardware
  - significant investment in GPUs

Next AI Scale on current hardware

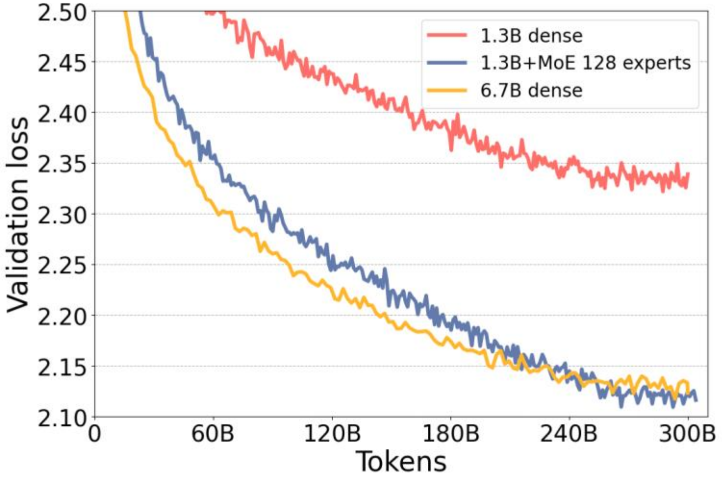
- Can we achieve next generation model quality on current generation of hardware?

- From a training perspective MoE provides a promising path
  - Scale at sub-linear cost

- MoE is *promising* but is it *practical*?
  - **Limited Scope:** Does it work for NLG or NLR or other models?
  - **Massive Memory Requirements:** 8-10x in size compared to quality equivalent dense
  - **Limited Inference Performance:** Massive model size == slow and expensive inference?

Cheaper NLG Model Training with MoE

- 1.3B+MoE with 128 experts, compared to 1.3B and 6.7B dense (GPT-3 like)
- **5x** lower training cost to same accuracy using MoE
- **8x** more parameters to same accuracy using MoE

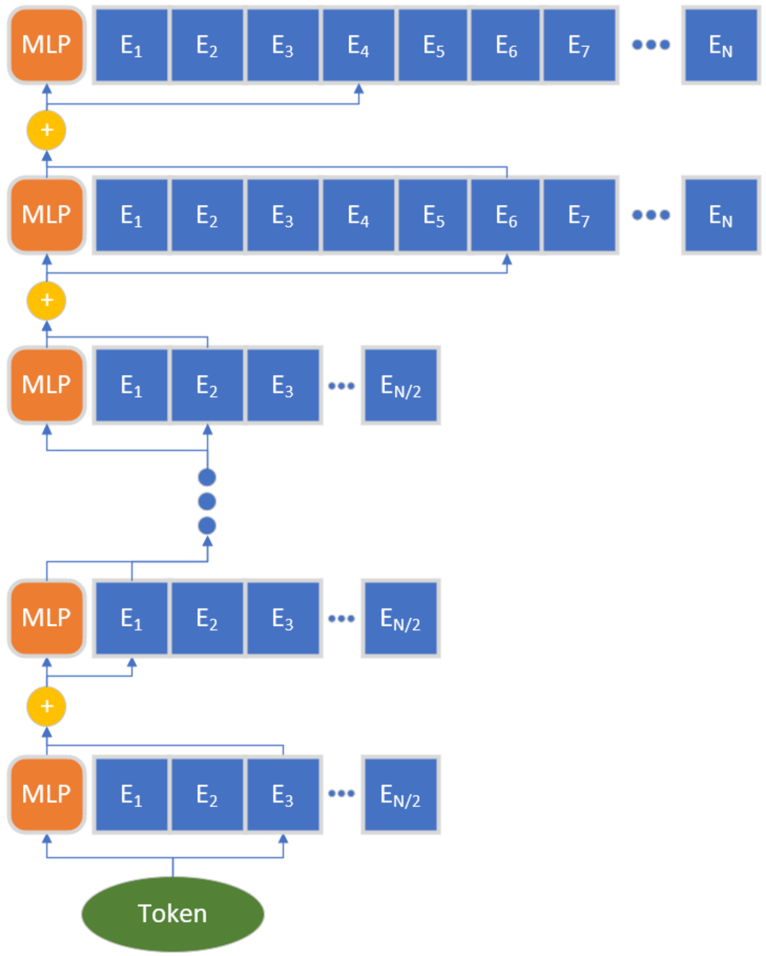


Case	Model size	LAMBADA: completion prediction	PIQA: commonsense reasoning	BoolQ: reading comprehension	RACE-h: reading comprehension	TriviaQA: question answering	WebQs: question answering
Dense NLG:							
(1) 350M	350M	52.03	69.31	53.64	31.77	3.21	1.57
(2) 1.3B	1.3B	63.65	73.39	63.39	35.60	10.05	3.25
(3) 6.7B	6.7B	**71.94**	**76.71**	**67.03**	37.42	23.47	5.12
Standard MoE NLG:							
(4) 350M+MoE-128	13B	62.70	74.59	60.46	35.60	16.58	5.17
(5) 1.3B+MoE-128	52B	69.84	**76.71**	64.92	**38.09**	**31.29**	**7.19**

	Training samples per sec	Throughput gain/ Cost Reduction
6.7B dense	70	1x
1.3B+MoE-128	372	5x

PR-MoE: a parameter efficient MoE model design

- New architecture: Pyramid-Residual MoE (PR-MoE)
  - Pyramid MoE: 2x experts in last two layers
  - Residual MoE: a fixed MLP plus a chosen expert per layer per token
- Mixture-of-Student (layer reduced version of PR-MoE)
  - First MoE-to-MoE distillation work
  - A novel staged knowledge distillation algorithm



Standard MoE vs. PR-MoE + MoS

- PR-MoE: model size reduction from 1.7x to 3.2x ; no performance degradation
- PR-MoE + MoS: model size reduction from 1.9x to 3.7x; maintaining >99% performance

Case	Model size (Reduction)	LAMBADA: completion prediction	PIQA: commonsense reasoning	BoolQ: reading comprehension	RACE-h: reading comprehension	TriviaQA: question answering	WebQs: question answering
MoE NLG with 350M base model:							
(1) MoE	13B (1x)	62.70	**74.59**	**60.46**	35.60	**16.58**	5.17
(2) PR-MoE	4.0B (3.2x)	**63.65**	73.99	59.88	**35.69**	16.30	4.73
(3) PR-MoE + MoS	**3.5B (3.7x)**	63.46	73.34	58.07	34.83	13.69	**5.22**
MoE NLG with 1.3B base model:							
(4) MoE	52B (1x)	69.84	76.71	64.92	**38.09**	**31.29**	7.19
(5) PR-MoE	31B (1.7x)	**70.60**	**77.75**	**67.16**	**38.09**	28.86	7.73
(6) PR-MoE + MoS	**27B (1.9x)**	70.17	77.69	65.66	36.94	29.05	**8.22**

Designing a highly scalable MoE Inference System

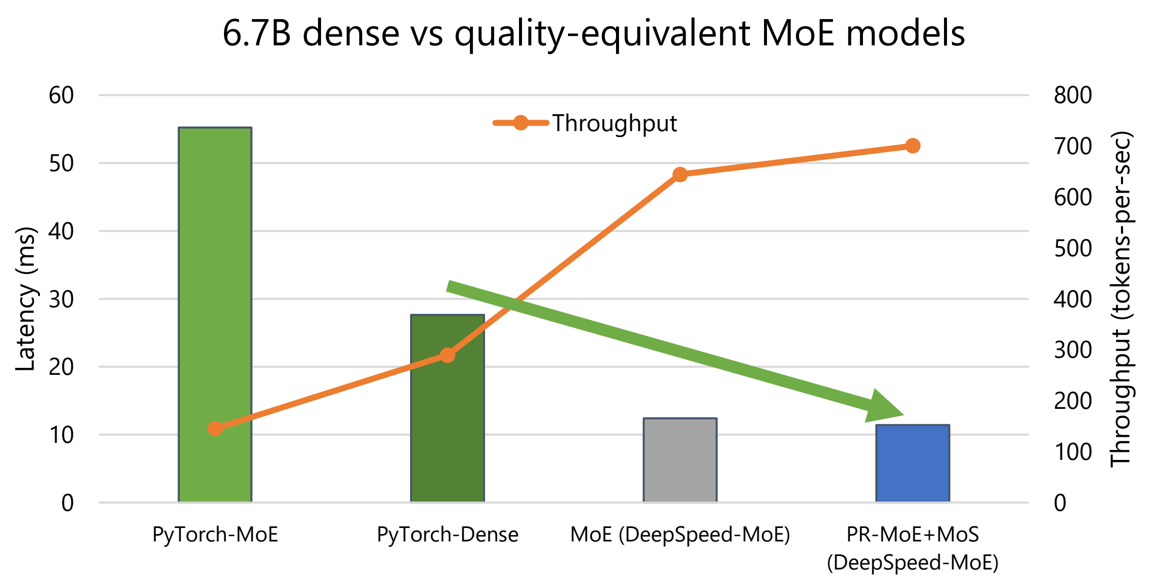
- Key Challenge:
  - 4x larger MoE model size than quality-equivalent-dense models (QEDM)
    - Requires 4x higher bandwidth/parallelism/scalability for latency parity

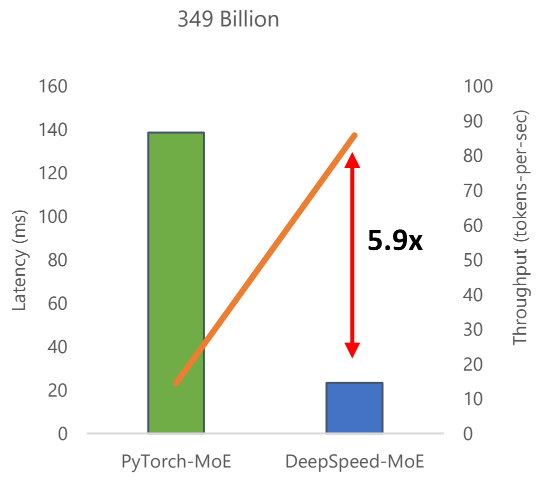
- Goal:
  - Achieve aggregate memory bandwidth across **<u>hundreds of devices</u>**

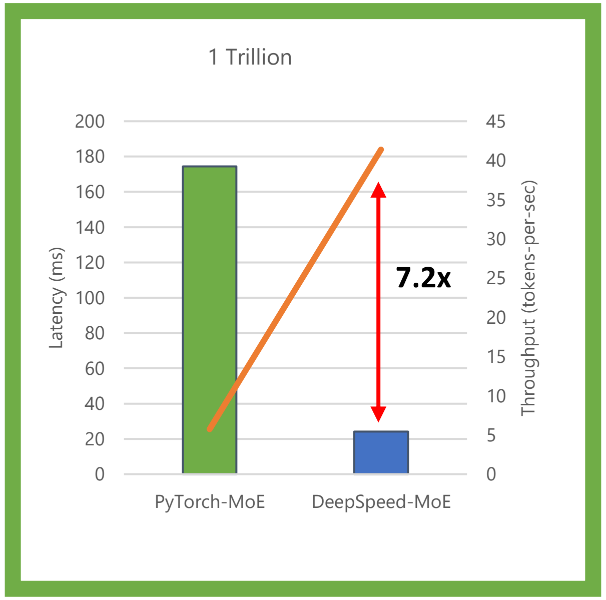
- Three main area of optimizations for maximizing aggregate bandwidth
  - A symphony of parallelism
    - Careful orchestration of tensor, data and expert parallelism
  - Parallelism coordinated Communication Optimization Strategies
    - Minimize communication overhead
  - Kernel Optimizations
    - Maximize bandwidth utilization per device

Lower-latency & Higher-throughput at Unprecedented Scale

- 7.2x faster inference
  - 25ms for serving a 1T model

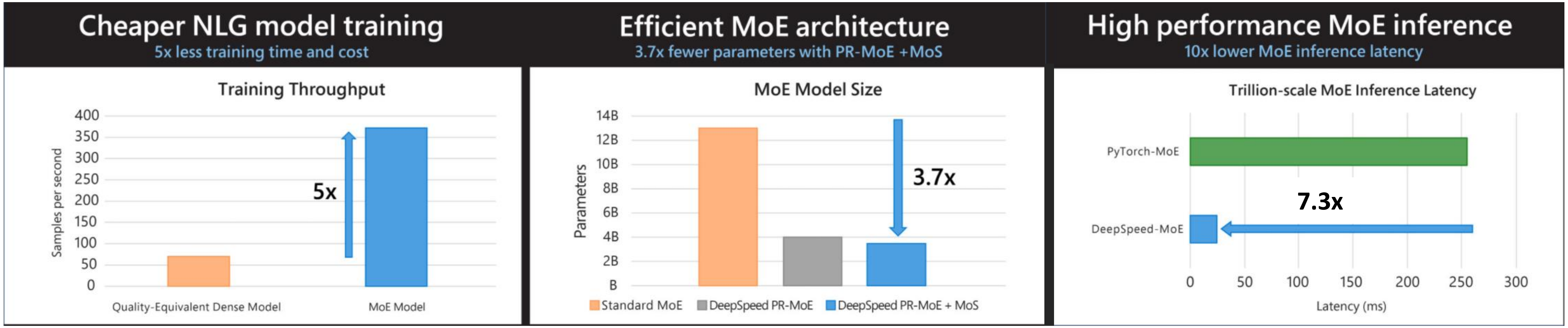






Faster than dense model inference with DeepSpeed-MoE

DeepSpeed-MoE: Powering the next generation of AI Scale



Thank you!

To Learn more: www.deepspeed.ai