



CONFERENCE

—NORTH AMERICA 2026—

Scaling Frontier Model Training using DeepSpeed

Zipeng Wang
Google, co-lead of DeepSpeed project

What is DeepSpeed?

DeepSpeed is an OSS training optimization library that makes distributed training easy, efficient, and effective.

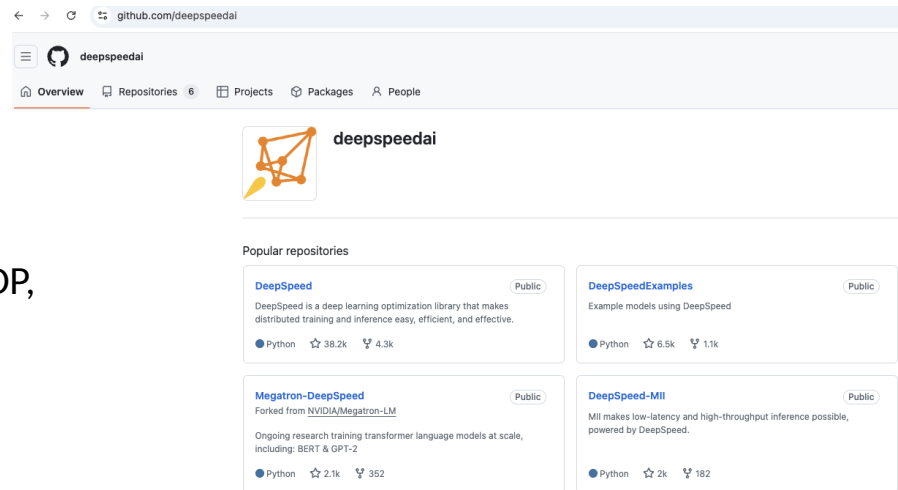
Efficiency: high-performance on 1 to 100K GPUs

Ease-of-use: minimal effort to run out of the box

Democratization: low-budget access to SOTA AI

Inspired other training optimization frameworks - e.g. FSDP, Megatron-LM etc.

DeepSpeed joins PyTorch Foundation since 2025



DeepSpeed: Tackling the GPU Memory Wall

ZeRO: Memory Optimizations Toward Training Trillion Parameter Models

Samyam Rajbhandari*, Jeff Rasley*, Olatunji Ruwase, Yuxiong He
{samyamr, jerasley, olruwase, yuxhe}@microsoft.com

ZeRO-Offload: Democratizing Billion-Scale Model Training

Jie Ren <i>UC Merced</i>	Samyam Rajbhandari <i>Microsoft</i>	Reza Yazdani Aminabadi <i>Microsoft</i>	Olatunji Ruwase <i>Microsoft</i>
Shuangyan Yang <i>UC Merced</i>	Minjia Zhang <i>Microsoft</i>	Dong Li <i>UC Merced</i>	Yuxiong He <i>Microsoft</i>

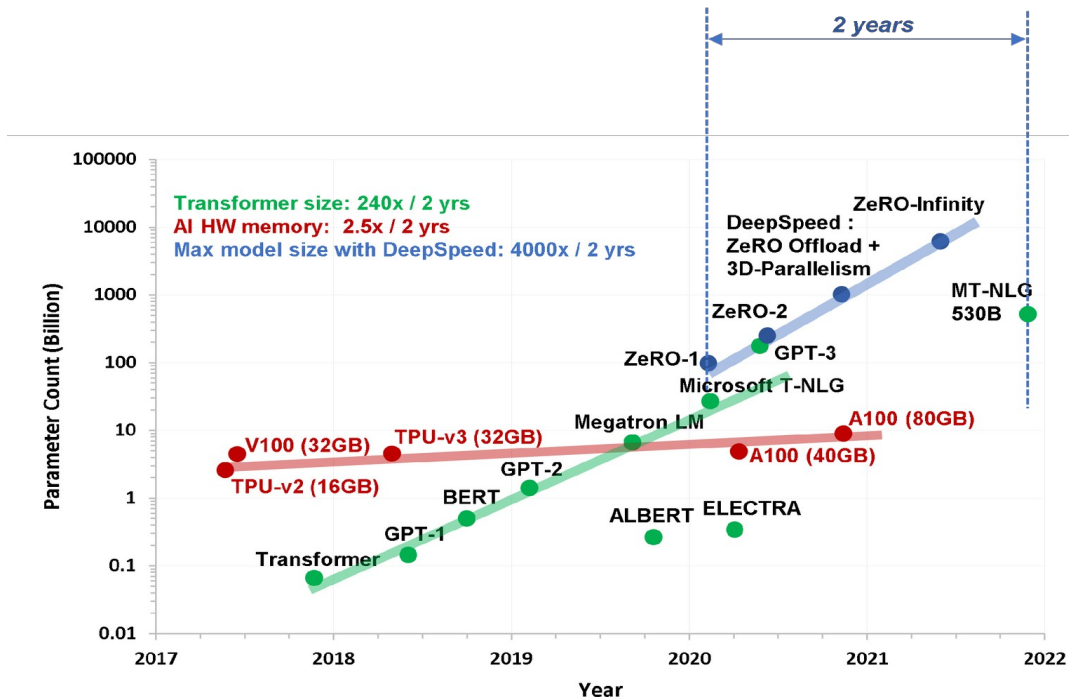
ZeRO-Infinity: Breaking the GPU Memory Wall for Extreme Scale Deep Learning

Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, Yuxiong He
{samyamr, olruwase, jerasley, shsmi, yuxhe}@microsoft.com

Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model

Shaden Smith^{§1}, Mostofa Patwary^{§1}, Brandon Norick¹, Patrick LeGresley¹, Samyam Rajbhandari¹, Jared Casper¹, Zhun Liu¹, Shrimai Prabhumoye¹, George Zerveas^{*1}, Vijay Korthikanti¹, Elton Zhang¹, Rewon Child¹, Reza Yazdani Aminabadi¹, Julie Bernauer¹, Xia Song¹, Mohammad Shoeybi¹, Yuxiong He¹, Michael Houston¹, Saurabh Tiwary¹, and Bryan Catanzaro²

[§]equal contribution
¹Microsoft
²NVIDIA



Impact on GPU Memory Wall

Trillion-parameter support available in ~2021

Inspired other scaling frameworks: FSDP, Colossal-AI, Megatron-LM, etc.

DeepSpeed Updates

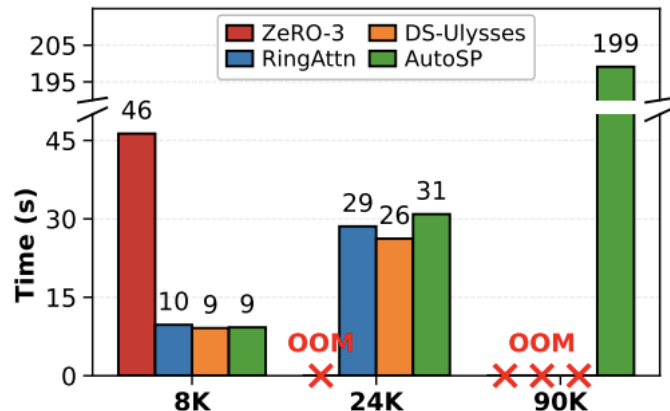
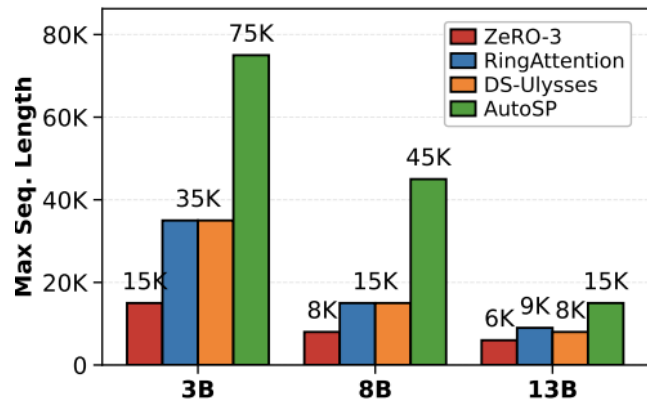
AutoSP

Muon Optimizer

New Accelerators Enablement

AutoEP + AutoTP

AutoSP: Automatic Sequence Parallelism



AUTO SP: UNLOCKING LONG-CONTEXT LLM TRAINING VIA COMPILER-BASED SEQUENCE PARALLELISM

Ahan Gupta*, Zhihao Wang*, Neel Dani
SSAIL Lab, University of Illinois Urbana-Champaign
Champaign, IL, USA
{ag82, zhihaow6, neeld2}@illinois.edu

Masahiro Tanaka
Anyscale
San Francisco, CA, USA
mtanaka@anyscale.com

Olatunji Ruwase
Snowflake
San Mateo, CA, USA
tunji.ruwase@snowflake.com

Minjia Zhang
SSAIL Lab, University of Illinois Urbana-Champaign
Champaign, IL, USA
minjiaz@illinois.edu



DeepSpeed

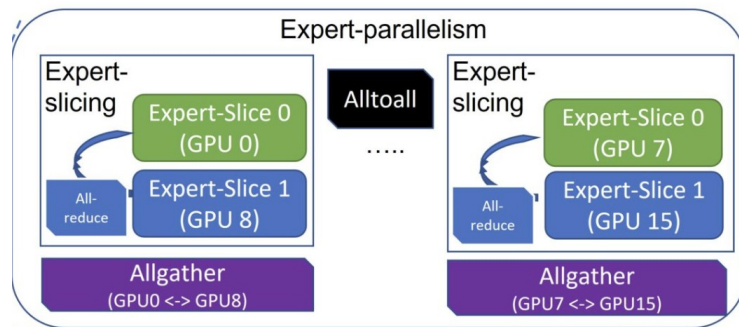
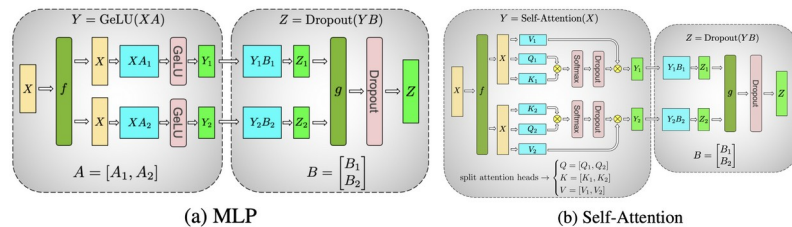
Muon Optimizer

- **Muon** (MomentUm Orthogonalized by Newton-Schulz) is an optimizer for 2D parameters of neural network hidden layers
- **~1.33x training speed up** compares to adam optimizer for several OSS frontier Models
- Muon is the optimizer used by Kimi K3 models
- DeepSpeed integrated Muon optimizer to work with **ZeRO stage 1/2/3** through **tensor unflattening**
- Support Muon in CPU offloading
- In DeepSpeed, you can change optimizer type to Muon by a line of change in config file

<https://pytorch.org/blog/using-muon-optimizer-with-deepspeed/>

AutoTP + AutoEP training for HuggingFace models

- **AutoTP + AutoEP -- Automatic Tensor parallel and Expert Parallelism folding**
 - Analyze model config of a HF model and shard model attention heads, MLP, etc. among accelerators
 - Auto inject communication into model computation
 - No modeling change, no preprocessing of model checkpoint required
- DeepSpeed now brings AutoTP + AutoEP to training
 - Turn on with simple configuration change
 - Training checkpoint directly usable
 - Free up more memory for batch scaling



DeepSpeed

New Accelerators Enablement

- DeepSpeed aims on democratizing Large Models training across multiple HW platforms and accelerators
- We have successfully enabled support for multiple Accelerator types ranging from Nvidia GPU, AMD GPU, Apple Silicon, NPU, XPU etc.
- We are also working on supporting TPU with TorchTPU, as well as other HW vendors to enable and optimize DeepSpeed performance on scaling large model training

