# DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters

Jeff Rasley
Microsoft

Samyam Rajbhandari
Microsoft

Olatunji Ruwase
Microsoft

Yuxiong He
Microsoft

## ABSTRACT

Explore new techniques in Microsoft's open source library called DeepSpeed, which advances large model training by improving scale, speed, cost, and usability, unlocking the ability to train 100-billion-parameter models. DeepSpeed is compatible with PyTorch. One piece of our library, called ZeRO, is a new parallelized optimizer that greatly reduces the resources needed for model and data parallelism while massively increasing the number of parameters that can be trained. Researchers have used these breakthroughs to create Turing Natural Language Generation (Turing-NLG), which at the time of its release was the largest publicly known language model at 17 billion parameters. In addition we will also go over our latest transformer kernel advancements that led the DeepSpeed team to achieve the world fastest BERT pretraining record.

The Zero Redundancy Optimizer (ZeRO) is a novel memory optimization technology for large-scale distributed deep learning. ZeRO can train deep learning models with over 100 billion parameters on the current generation of GPU clusters at three to five times the throughput of the current best system. It also presents a clear path to training models with trillions of parameters, demonstrating an unprecedented leap in deep learning system technology.

DeepSpeed brings state-of-the-art training techniques, such as ZeRO, optimized kernels, distributed training, mixed precision, and checkpointing, through lightweight APIs compatible with PyTorch. With just a few lines of code changes to your PyTorch model, you can leverage DeepSpeed to address underlying performance challenges and boost the speed and scale of your training.

## KEYWORDS

Distributed deep learning, machine learning

**ACM Reference Format:**
Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3394486.3406703

Corresponding author email: jeff.rasley@microsoft.com.


*KDD '20, August 23–27, 2020, Virtual Event, CA, USA*

ACM ISBN 978-1-4503-7998-4/20/08.
https://doi.org/10.1145/3394486.3406703

## 1 TUTORIAL OUTLINE

The latest trend in AI is that larger natural language models provide better accuracy; however, larger models are difficult to train because of cost, time, and ease of code integration. With the goal of advancing large model training by improving scale, speed, cost, and usability for model developers across the world, Microsoft made the DeepSpeed library open source in February of 2020. Since then, the DeepSpeed team has been hard at work extending the library to continue pushing the boundaries of scale and speed of deep learning training.

In this tutorial, the DeepSpeed team will discuss what DeepSpeed is (https://www.deepspeed.ai), how to use it with your existing PyTorch [1] models, and advancements in the ZeRO optimizer that are central to supporting training of 100-200 billion parameter models and higher. The team will present deep-dive results on how they were able to obtain the world record for fastest BERT training. In addition, other features of DeepSpeed will be discussed such as mixed precision [2] support to take full advantage of NVIDIA's Volta architecture [3].

DeepSpeed can efficiently train models with 100-200 billion parameters up to 10 times faster than the state-of-the-art via the use of a memory optimization system that we called ZeRO (Zero Redundancy Optimizer) [4]. ZeRO is a parallelized optimizer that greatly reduces the resources needed for model and data parallelism while massively increasing the number of parameters that can be trained. Researchers used these breakthroughs to create the Turing Natural Language Generation (Turing-NLG), one of the largest publicly known language models at 17 billion parameters.

DeepSpeed recently obtained the fastest BERT training record of 44 minutes on 1024 NVIDIA V100 GPUs [5]. This is a 34% improvement over the best published result of 67 minutes [6] in end-to-end training time to achieve the same accuracy on the same number and generation of GPUs. This improvement does not come at the cost of excessive hardware resources but comes from a result of improved software efficiency. For example, DeepSpeed can attain 64 teraflops of single GPU performance on a NVIDIA V100 GPU, which is over 50% of the hardware peak.

## 2 TUTORS AND BIOGRAPHIES

***Jeff Rasley, Ph.D.:*** Jeff Rasley is a researcher at Microsoft with a background in distributed systems and networking. His research is focused broadly on improving the performance and usability of deep learning training. Jeff's research has been used in multiple Microsoft systems and projects such as Bing, Ads, AzureML, and HyperDrive. He is one of the founding developers of the DeepSpeed training library that has resulted in order(s)-of-magnitude

speed/scale improvements for DL training. Jeff received his PhD in Computer Science from Brown University.

***Samyam Rajbhandari, Ph.D.:*** Samyam Rajbhandari is a researcher at Microsoft with a background in High Performance Computing. He works on developing high performance infrastructures for accelerating large scale deep learning training and inference on parallel and distributed systems. His research results are used in multiple Microsoft systems and products, such as Bing, Ads, AzureML to improve performance and capacity. He developed the core technology in DeepCPU and DeepSpeed resulting in order(s)-of-magnitude improvement on speed and scale for DL inference and training. His recent work on memory optimization has enabled training of very large models including the 17.2B Turing-NLG model from Microsoft. Samyam received his PhD in Computer Science from Ohio State University.

***Olatunji Ruwase, Ph.D.:*** Olatunji Ruwase is a researcher at Microsoft with a systems background spanning compilers, operating systems, and hardware accelerators. His research results on training, inference, and hyperparameter search for deep learning models are used in multiple Microsoft systems and products, such as Bing, Ads, HyperDrive, and Catapault. He is recently working on systems support and convergence techniques for accelerating large-scale training of deep learning models on distributed GPUs. Tunji received his PhD in Computer Science from Carnegie Mellon University.

***Yuxiong He, Ph.D.:*** Yuxiong He is a research manager at Microsoft. She works on performance optimization of parallel and distributed systems such as machine learning systems, information retrieval systems, data management systems, and large-scale cloud infrastructure. Her work was selected among the best papers for SIGIR, ICDE, WSDM and Middleware. Her research results have been utilized by various Microsoft systems and products, such as Bing, Ads, AzureML and AzureSQL, boosting system performance and capacity. Her recent work focuses on optimizing deep learning systems. She leads DeepCPU and DeepSpeed projects, which strive for order(s)-of-magnitude improvement on speed and scale for DL inference and training. Yuxiong received her PhD in Computer Science from Singapore-MIT Alliance.

## REFERENCES

[1] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[2] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training, 2017.

[3] NVIDIA Tesla V100 GPU architecture. http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf, 2017. [Online, accessed 22-April-2020].

[4] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. *arXiv:1910.02054*, 2019.

[5] Microsoft DeepSpeed achieves the fastest BERT training time. https://www.deepspeed.ai/news/2020/05/27/fastest-bert-training.html, 2020.

[6] Shar Narasimhan. NVIDIA Clocks World's Fastest BERT Training Time ... https://devblogs.nvidia.com/training-bert-with-gpus/, 2019. [Online; accessed 25-September-2019].