近三年AI系统类论文参考列表：

[Shaoqi Wang](https://dblp.uni-trier.de/pid/156/4566.html), [Oscar J. Gonzalez](https://dblp.uni-trier.de/pid/24/356.html), [Xiaobo Zhou](https://dblp.uni-trier.de/pid/13/6395-2.html), [Thomas Williams](https://dblp.uni-trier.de/pid/21/4101.html), [Brian D. Friedman](https://dblp.uni-trier.de/pid/31/9986.html), [Martin Havemann](https://dblp.uni-trier.de/pid/69/1532.html), [Thomas Y. C. Woo](https://dblp.uni-trier.de/pid/45/3217.html):  
**An efficient and non-intrusive GPU scheduling framework for deep learning training systems.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#WangG0WFHW20): 90

[Cong Guo](https://dblp.uni-trier.de/pid/117/1754-3.html), [Bo Yang Hsueh](https://dblp.uni-trier.de/pid/222/1570.html), [Jingwen Leng](https://dblp.uni-trier.de/pid/131/5131.html), [Yuxian Qiu](https://dblp.uni-trier.de/pid/233/8648.html), [Yue Guan](https://dblp.uni-trier.de/pid/54/7820.html), [Zehuan Wang](https://dblp.uni-trier.de/pid/274/1546.html), [Xiaoying Jia](https://dblp.uni-trier.de/pid/73/9555-1.html), [Xipeng Li](https://dblp.uni-trier.de/pid/274/1507.html), [Minyi Guo](https://dblp.uni-trier.de/pid/99/6797.html), [Yuhao Zhu](https://dblp.uni-trier.de/pid/72/8314-1.html):  
**Accelerating sparse DNN models without hardware-support via tile-wise sparsity.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#0003HLQGW0LG020): 16

[Trevor Gale](https://dblp.uni-trier.de/pid/190/7144.html), [Matei Zaharia](https://dblp.uni-trier.de/pid/36/2133.html), [Cliff Young](https://dblp.uni-trier.de/pid/64/1248.html), [Erich Elsen](https://dblp.uni-trier.de/pid/94/3606.html):  
**Sparse GPU kernels for deep learning.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#GaleZYE20): 17

[Mohamed Wahib](https://dblp.uni-trier.de/pid/10/6150.html), [Haoyu Zhang](https://dblp.uni-trier.de/pid/168/0332.html), [Truong Thao Nguyen](https://dblp.uni-trier.de/pid/233/1462.html), [Aleksandr Drozd](https://dblp.uni-trier.de/pid/06/10957.html), [Jens Domke](https://dblp.uni-trier.de/pid/11/10083.html), [Lingqi Zhang](https://dblp.uni-trier.de/pid/183/6577-1.html), [Ryousei Takano](https://dblp.uni-trier.de/pid/42/679.html), [Satoshi Matsuoka](https://dblp.uni-trier.de/pid/57/4464.html):  
**Scaling distributed deep learning workloads beyond the memory capacity with KARMA.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#WahibZNDD0TM20): 19

[Samyam Rajbhandari](https://dblp.uni-trier.de/pid/115/9021.html), [Jeff Rasley](https://dblp.uni-trier.de/pid/90/8732.html), [Olatunji Ruwase](https://dblp.uni-trier.de/pid/72/2838.html), [Yuxiong He](https://dblp.uni-trier.de/pid/50/174.html):  
**ZeRO: memory optimizations toward training trillion parameter models.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#RajbhandariRRH20): 20

[Minhui Xie](https://dblp.uni-trier.de/pid/06/5769.html), [Kai Ren](https://dblp.uni-trier.de/pid/01/7592.html), [Youyou Lu](https://dblp.uni-trier.de/pid/96/8005.html), [Guangxu Yang](https://dblp.uni-trier.de/pid/266/9100.html), [Qingxing Xu](https://dblp.uni-trier.de/pid/284/4647.html), [Bihai Wu](https://dblp.uni-trier.de/pid/284/5134.html), [Jiazhen Lin](https://dblp.uni-trier.de/pid/240/9420.html), [Hongbo Ao](https://dblp.uni-trier.de/pid/166/6635.html), [Wanhong Xu](https://dblp.uni-trier.de/pid/81/5760.html), [Jiwu Shu](https://dblp.uni-trier.de/pid/60/3690.html):  
**Kraken: memory-efficient continual learning for large-scale real-time recommendations.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#XieRLYXWLAXS20): 21

[Di Zhang](https://dblp.uni-trier.de/pid/80/3482.html), [Dong Dai](https://dblp.uni-trier.de/pid/55/6260-1.html), [Youbiao He](https://dblp.uni-trier.de/pid/191/4569.html), [Forrest Sheng Bao](https://dblp.uni-trier.de/pid/98/5980.html), [Bing Xie](https://dblp.uni-trier.de/pid/94/2438.html):  
**RLScheduler: an automated HPC batch job scheduler using reinforcement learning.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#ZhangDHBX20): 31

[Dhiraj D. Kalamkar](https://dblp.uni-trier.de/pid/95/4155.html), [Evangelos Georganas](https://dblp.uni-trier.de/pid/121/2450.html), [Sudarshan Srinivasan](https://dblp.uni-trier.de/pid/76/9547.html), [Jianping Chen](https://dblp.uni-trier.de/pid/61/3598.html), [Mikhail Shiryaev](https://dblp.uni-trier.de/pid/205/2909.html), [Alexander Heinecke](https://dblp.uni-trier.de/pid/84/6756.html):  
**Optimizing deep learning recommender systems training on CPU cluster architectures.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#KalamkarGSCSH20): 43

[Indu Thangakrishnan](https://dblp.uni-trier.de/pid/284/4779.html), [Derya Cavdar](https://dblp.uni-trier.de/pid/49/9873.html), [Can Karakus](https://dblp.uni-trier.de/pid/81/11533.html), [Piyush Ghai](https://dblp.uni-trier.de/pid/189/9993.html), [Yauheni Selivonchyk](https://dblp.uni-trier.de/pid/284/4687.html), [Cory Pruce](https://dblp.uni-trier.de/pid/284/5048.html):  
**Herring: rethinking the parameter server at scale for the cloud.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#ThangakrishnanC20): 44

[Arpan Jain](https://dblp.uni-trier.de/pid/154/8472.html), [Ammar Ahmad Awan](https://dblp.uni-trier.de/pid/75/11190.html), [Asmaa M. Aljuhani](https://dblp.uni-trier.de/pid/284/4650.html), [Jahanzeb Maqbool Hashmi](https://dblp.uni-trier.de/pid/194/1394.html), [Quentin G. Anthony](https://dblp.uni-trier.de/pid/252/7024.html), [Hari Subramoni](https://dblp.uni-trier.de/pid/65/1958.html), [Dhabaleswar K. Panda](https://dblp.uni-trier.de/pid/p/DhabaleswarKPanda.html), [Raghu Machiraju](https://dblp.uni-trier.de/pid/65/1768.html), [Anil Parwani](https://dblp.uni-trier.de/pid/132/4688.html):  
**GEMS: GPU-enabled memory-aware model-parallelism system for distributed DNN training.** [SC 2020](https://dblp.uni-trier.de/db/conf/sc/sc2020.html#JainAAHASPMP20): 45

[Nikoli Dryden](https://dblp.uni-trier.de/pid/148/1273.html), [Naoya Maruyama](https://dblp.uni-trier.de/pid/99/5529.html), [Tim Moon](https://dblp.uni-trier.de/pid/183/6612.html), [Tom Benson](https://dblp.uni-trier.de/pid/238/0178.html), [Marc Snir](https://dblp.uni-trier.de/pid/40/3047.html), [Brian Van Essen](https://dblp.uni-trier.de/pid/57/3302.html):  
**Channel and filter parallelism for large-scale CNN training.** [SC 2019](https://dblp.uni-trier.de/db/conf/sc/sc2019.html#DrydenMMBSE19): 10:1-10:20

[Cédric Renggli](https://dblp.uni-trier.de/pid/215/5117.html), [Saleh Ashkboos](https://dblp.uni-trier.de/pid/195/5539.html), [Mehdi Aghagolzadeh](https://dblp.uni-trier.de/pid/19/5184.html), [Dan Alistarh](https://dblp.uni-trier.de/pid/36/3251.html), [Torsten Hoefler](https://dblp.uni-trier.de/pid/16/3869.html):  
**SparCML: high-performance sparse communication for machine learning.** [SC 2019](https://dblp.uni-trier.de/db/conf/sc/sc2019.html#RenggliAAAH19): 11:1-11:15

[Prasanna Balaprakash](https://dblp.uni-trier.de/pid/51/3696.html), [Romain Egele](https://dblp.uni-trier.de/pid/248/7760.html), [Misha Salim](https://dblp.uni-trier.de/pid/248/7934.html), [Stefan M. Wild](https://dblp.uni-trier.de/pid/05/1044.html), [Venkatram Vishwanath](https://dblp.uni-trier.de/pid/30/6369.html), [Fangfang Xia](https://dblp.uni-trier.de/pid/90/6711.html), [Tom Brettin](https://dblp.uni-trier.de/pid/81/5091.html), [Rick Stevens](https://dblp.uni-trier.de/pid/s/RickLStevens.html):  
**Scalable reinforcement-learning-based neural architecture search for cancer deep learning research.** [SC 2019](https://dblp.uni-trier.de/db/conf/sc/sc2019.html#BalaprakashESWV19): 37:1-37:33

[Linnan Wang](https://dblp.uni-trier.de/pid/169/9748.html), [Jinmian Ye](https://dblp.uni-trier.de/pid/190/7150.html), [Yiyang Zhao](https://dblp.uni-trier.de/pid/33/3791.html), [Wei Wu](https://dblp.uni-trier.de/pid/95/6985-16.html), [Ang Li](https://dblp.uni-trier.de/pid/33/2805.html), [Shuaiwen Leon Song](https://dblp.uni-trier.de/pid/23/7512.html), [Zenglin Xu](https://dblp.uni-trier.de/pid/68/1538.html), [Tim Kraska](https://dblp.uni-trier.de/pid/26/6037.html):  
**Superneurons: dynamic GPU memory management for training deep neural networks.** [PPOPP 2018](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2018.html#WangYZWLSXK18): 41-53

[hen Jia](https://dblp.uni-trier.de/pid/06/180-1.html), [Aleksandar Zlateski](https://dblp.uni-trier.de/pid/163/2100.html), [Frédo Durand](https://dblp.uni-trier.de/pid/87/2617.html), [Kai Li](https://dblp.uni-trier.de/pid/181/2853.html):  
**Optimizing N-dimensional, winograd-based convolution for manycore CPUs.** [PPOPP 2018](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2018.html#JiaZDL18): 109-123

[Yue Zhao](https://dblp.uni-trier.de/pid/48/76-11.html), [Jiajia Li](https://dblp.uni-trier.de/pid/89/9032.html), [Chunhua Liao](https://dblp.uni-trier.de/pid/44/6478.html), [Xipeng Shen](https://dblp.uni-trier.de/pid/36/4172.html):  
**Bridging the gap between deep learning and sparse matrix format selection.** [PPOPP 2018](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2018.html#ZhaoLLS18): 94-108

[Joel Hestness](https://dblp.uni-trier.de/pid/60/3063.html), [Newsha Ardalani](https://dblp.uni-trier.de/pid/53/7913.html), [Gregory F. Diamos](https://dblp.uni-trier.de/pid/06/1767.html):  
**Beyond human-level accuracy: computational challenges in deep learning.** [PPoPP 2019](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2019.html#HestnessAD19): 1-14

[Martin Winter](https://dblp.uni-trier.de/pid/55/6309.html), [Daniel Mlakar](https://dblp.uni-trier.de/pid/218/5849.html), [Rhaleb Zayer](https://dblp.uni-trier.de/pid/96/4208.html), [Hans-Peter Seidel](https://dblp.uni-trier.de/pid/s/HansPeterSeidel.html), [Markus Steinberger](https://dblp.uni-trier.de/pid/89/9395.html):  
**Adaptive sparse matrix-matrix multiplication on the GPU.** [PPoPP 2019](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2019.html#WinterMZSS19): 68-81

[Da Yan](https://dblp.uni-trier.de/pid/81/9436-2.html), [Wei Wang](https://dblp.uni-trier.de/pid/35/7092-30.html), [Xiaowen Chu](https://dblp.uni-trier.de/pid/24/2536.html):  
**Optimizing batched Winograd convolution on GPUs.** [PPoPP 2020](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2020.html#YanWC20): 32-44

[Mathias Parger](https://dblp.uni-trier.de/pid/230/8215.html), [Martin Winter](https://dblp.uni-trier.de/pid/55/6309.html), [Daniel Mlakar](https://dblp.uni-trier.de/pid/218/5849.html), [Markus Steinberger](https://dblp.uni-trier.de/pid/89/9395.html):  
**spECK: accelerating GPU sparse matrix-matrix multiplication through lightweight analysis.** [PPoPP 2020](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2020.html#PargerWMS20): 362-375

[Peng Jiang](https://dblp.uni-trier.de/pid/92/1104-4.html), [Changwan Hong](https://dblp.uni-trier.de/pid/176/9177.html), [Gagan Agrawal](https://dblp.uni-trier.de/pid/a/GAgrawal.html):  
**A novel data transformation and execution strategy for accelerating sparse matrix multiplication on GPUs.** [PPoPP 2020](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2020.html#JiangHA20): 376-388

[Kezhao Huang](https://dblp.uni-trier.de/pid/264/1773.html), [Jidong Zhai](https://dblp.uni-trier.de/pid/72/7413.html), [Zhen Zheng](https://dblp.uni-trier.de/pid/115/5528.html), [Youngmin Yi](https://dblp.uni-trier.de/pid/45/4951.html), [Xipeng Shen](https://dblp.uni-trier.de/pid/36/4172.html):  
**Understanding and bridging the gaps in current GNN performance optimizations.** [PPoPP 2021](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2021.html#HuangZZYS21): 119-132

[Xiaoyang Zhang](https://dblp.uni-trier.de/pid/61/10371.html), [Junmin Xiao](https://dblp.uni-trier.de/pid/135/9326.html), [Guangming Tan](https://dblp.uni-trier.de/pid/80/2272.html):  
**I/O lower bounds for auto-tuning of convolutions in CNNs.** [PPoPP 2021](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2021.html#ZhangXT21): 247-261

[Boyuan Feng](https://dblp.uni-trier.de/pid/227/2946.html), [Yuke Wang](https://dblp.uni-trier.de/pid/85/4549.html), [Guoyang Chen](https://dblp.uni-trier.de/pid/151/5469.html), [Weifeng Zhang](https://dblp.uni-trier.de/pid/19/949.html), [Yuan Xie](https://dblp.uni-trier.de/pid/x/YuanXie.html), [Yufei Ding](https://dblp.uni-trier.de/pid/127/9591.html):  
**EGEMM-TC: accelerating scientific computing on tensor cores with extended precision.** [PPoPP 2021](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2021.html#FengWCZ0D21): 278-291

[Jiarui Fang](https://dblp.uni-trier.de/pid/147/0860.html), [Yang Yu](https://dblp.uni-trier.de/pid/46/2181.html), [Chengduo Zhao](https://dblp.uni-trier.de/pid/214/1505.html), [Jie Zhou](https://dblp.uni-trier.de/pid/00/5012.html):  
**TurboTransformers: an efficient GPU serving system for transformer models.** [PPoPP 2021](https://dblp.uni-trier.de/db/conf/ppopp/ppopp2021.html#FangYZZ21): 389-402

提交时间6月25日23：59

International Conference for High Performance Computing, Networking, Storage and Analysis (SC)

<https://dblp.org/db/conf/sc/index.html>

ACM SIGPLAN Symposium on Principles & Practice of Parallel Programming (PPoPP)

<https://dblp.org/db/conf/ppopp/index.html>

IEEE International Conference on Cluster Computing (CLUSTER)

<https://dblp.org/db/conf/cluster/index.html>

USENIX Conference on File and Storage Technologies (FAST)

<https://dblp.org/db/conf/fast/index.html>

USENIX Annual Technical Conference (USENIX)

<https://dblp.org/db/conf/usenix/index.html>

Symposium on Networked Systems Design and Implementation (NSDI)

<https://dblp.org/db/conf/nsdi/index.html>

European Conference on Computer Systems (EuroSys)

<https://dblp.org/db/conf/eurosys/index.html>

International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)

<https://dblp.org/db/conf/asplos/index.html>

International Symposium on Computer Architecture (ISCA)

<https://dblp.org/db/conf/isca/index.html>

International Symposium on High-Performance Computer Architecture (HPCA)

<https://dblp.org/db/conf/hpca/index.html>

Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)

<https://dblp.org/db/conf/micro/index.html>