**Paper 1. On chip memory**

Y. Chen *et al*., "DaDianNao: A Machine-Learning Supercomputer," *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*, 2014, pp. 609-622, doi: 10.1109/MICRO.2014.58.

CNN’s and DNN’s require a large amount of memory, but only modestly large.

By utilizing on chip memory for multi-cihp systems (L2 cache I assume) and maybe the TLB. This is called the DinNao accelerator. The architecture is made up of buffers for caching latent variables in CNN or DNN and a Neural Functional Unit which is a pipelined version of typical back-propogation algorithms for CNN and DNN. This achieves a 450x speedup on GPU (on average).

**Research number 2. META**

See: <https://ai.facebook.com/blog/ai-rsc/>

Meta has introduced a new AI HPC supercomputer at enterprise scale. The need for the supercomputer arose in the need to train large models at scale. 760 NVIDIA DGX A100 used for compute and 6,080 GPUs — with each A100 GPU being more powerful than the V100 used in the system. 600 Gb/s InfiniBand two-level Clos fabric that has no oversubscription was used along with 175 petabytes of flash memory.

**Research number 3. SMFLOW**

SMFLOW implements optimizations around which mitigate bottlenecks on a central server allowing for better usage of parallelization specific to the TaihuLight supercomputer. This is shown to better support scaling up to ~1024 parallel processes. These optimizations result in a 10.42X speedup in CNN convolutional layers (multiply add’s) the convolutional layers. This is accomplished by usage of elastic stochastic gradient descent to reduce memory transfers and increase cache utilization. Further a Convolutional layer offload is implemented along with large batch model traning.

Reference:

<https://www.sciencedirect.com/science/article/abs/pii/S0020025520312457>

Mingfan Li, Han Lin, Junshi Chen, Jose Monsalve Diaz, Qian Xiao, Rongfen Lin, Fei Wang, Guang R. Gao, Hong An,

swFLOW: A large-scale distributed framework for deep learning on Sunway TaihuLight supercomputer,

Information Sciences,

Volume 570,

2021,

Pages 831-847,

ISSN 0020-0255,

https://doi.org/10.1016/j.ins.2020.12.079.

(https://www.sciencedirect.com/science/article/pii/S0020025520312457)