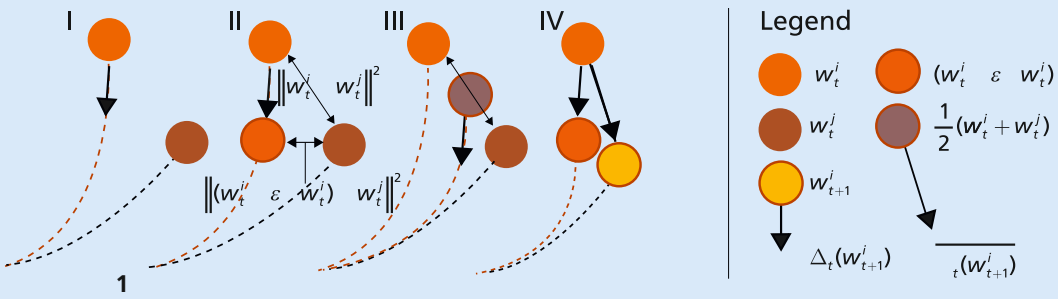
Fraunhofer
ITWM

FRAUNHOFER INSTITUTE FOR INDUSTRIAL MATHEMATICS ITWM



1

1 *Graph based visualization oft he ASGD update scheme. Individual processes send updates completely uninformed of the recipients. On the recipient side, available updates are included in the local computation as available. In this scheme, no process ever waits for any communication to be send or received.*

FAST PARALLEL ALGORITHMS FOR LARGE SCALE MACHINE LEARNING APPLICATIONS

Asynchronous Stochastic Gradient Descent (ASGD)

ASGD – Asynchronous Stochastic Gradient Descent is a fast parallel optimization method for Machine Learning on HPC cluster and HTC cloud applications.

Stochastic Gradient Descent (SGD) is the standard numerical method used to solve the core optimization problem for the vast majority of machine learning algorithms. In the context of large scale learning, as utilized by many **Big Data** applications, the efficient parallelization of SGD on distributed systems is a key performance factor.

We introduce **Asynchronous Stochastic Gradient Descent,** outperforming current, mostly MapReduce based, parallel SGD algorithms in solving the optimization task for large scale machine learning problems in distributed memory environments. We are able to show [1], that ASGD is faster, has better convergence and scaling properties and leads to better error rates than other state of the art methods.

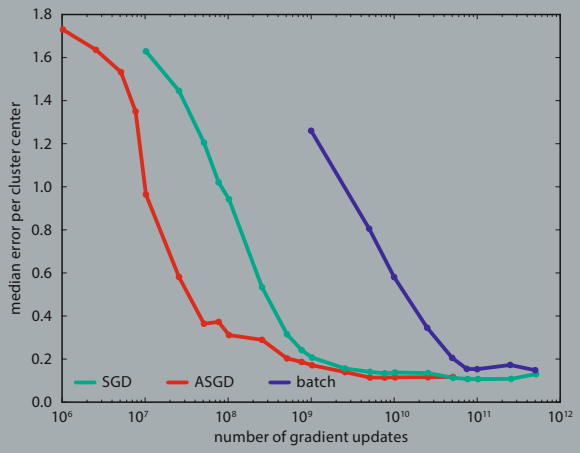
- Optimizing Machine Learning on HPC and HTC platforms
- Asynchronous communication on shared memory systems
- Strong linear scaling
- Superior convergence
- Better error rates

Fraunhofer-Institut für Techno- und Wirtschaftsmathematik ITWM

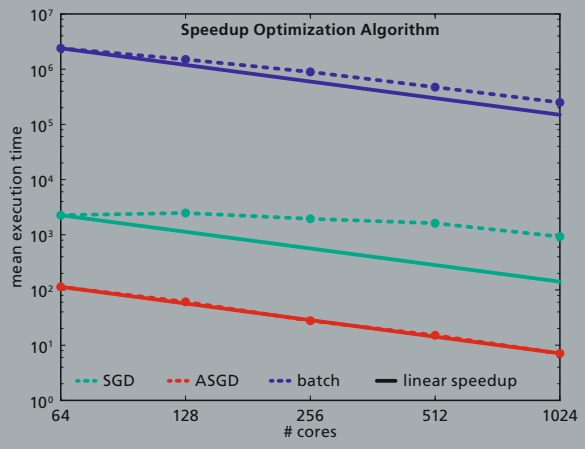
Fraunhofer-Platz 1
67663 Kaiserslautern
Germany

Contact
Dr.-Ing. Janis Keuper
phone +49 631 31600-47 15
janis.keuper@itwm.fraunhofer.de

www.itwm.fraunhofer.de



2



3

2 *Convergence properties of our ASGD algorithm compared to parallel SGD and MapReduce BATCH optimization applied to K-Means clustering with k=100, d=100 and ~1TB of data samples*

3 *Scaling properties of ASGD applied to K-Means clustering with k=10, d=10 and ~1TB of data samples*

Asynchronous Communication outperforms MapReduce

Figures 2 and 3 show the performance of an unsupervised learning with the K-Means clustering algorithm. Using ASGD optimization lower error rates with less iterations than MapReduce based SGD or BATCH methods (see [1] for detailed analysis). Also, ASGD provides linear strong scaling in the number of cores and is considerably faster than MapReduce based SGD or BATCH methods.

Asynchronous Communication: applying the PGAS scheme to machine learning applications

ASGD is implemented using the PGAS programming model: we apply the asynchronous, single-sided communication scheme provided by the GPI 2.0 [3] API of the GASPI [2] protocol. Individual processes send mini-BATCH updates completely uninformed of the recipients status whenever they are ready to do so. On the recipient side, available updates are included in the local computation as available. In this scheme, no process ever waits for any communication to be send or received. Hence, communication is literally "free" (in terms of latency).

2 *J. Keuper, F.-J. Pfreundt*
Asynchronous parallel stochastic gradient descent – a numeric core for scalable distributed machine learning algorithms
In arxiv.org/abs/1505.04956, 2015

3 *D. Grünewald, C. Simmendinger*
The gaspi api specification and its implementation
gpi 2.0. In 7th International Conference on PGAS Programming Models, volume 243, 2013

4 *GPI 2.0 is our open source implementation of the GASPI standard. Downloads available at: www.gpi-site.com/gpi2/*