Machine Learning Reading Group

Speeding Up Distributed Machine Learning Using Codes

Kangwook Lee, Maximilian Lam, Ramtin Pedarsani, Dimitris Papailiopoulos, and Kannan Ramchandran, *Fellow*, IEEE, March 2018

University “Transilvania”
Faculty of Mathematics and Informatics
Brasov
07.Nov.2018

Kerestély Árpád
k_arpi2004@yahoo.co.uk

Distributed Machine Learning

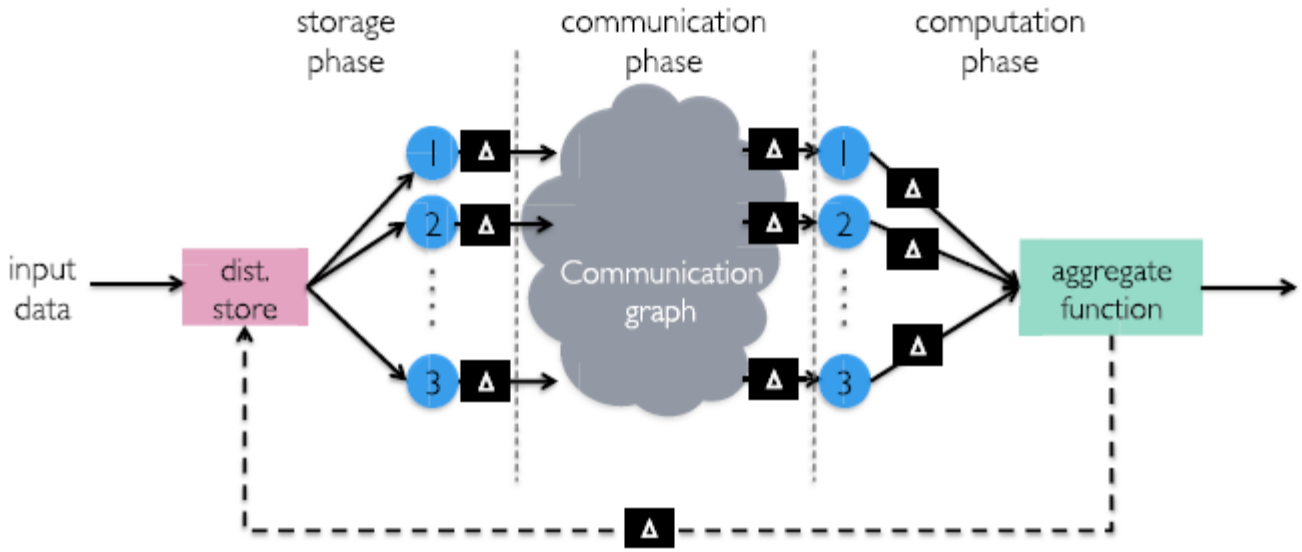
- largescale machine learning and data analytics has shifted towards massively large distributed systems
- more commonly the domain that uses “massively large distributed systems” is called Cloud Computing

What is Codes?

- Codes refers to coding (modifying, transforming) some messages
- The two usages of message coding are:
  - Coded Computation
  - Coded Shuffling

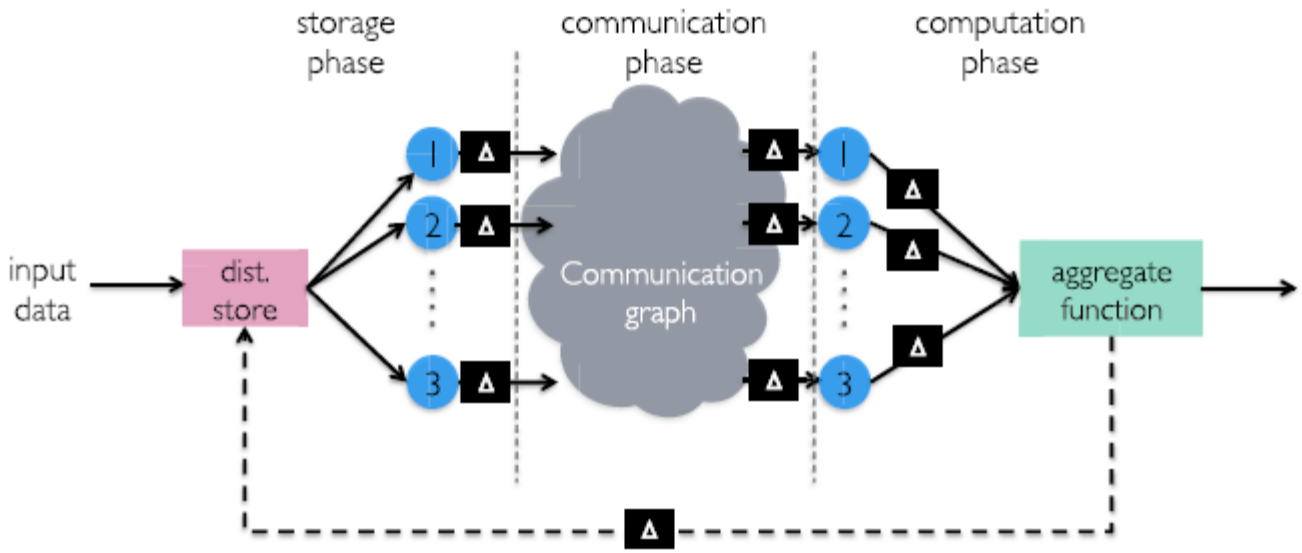
Cloud Computing

- comprising individually small and unreliable computational nodes (low-end, commodity hardware)
- enable the execution of production-scale tasks on data sizes of the order of petabytes (1 PB = 1000 TB)
- modern distributed systems for ML: Apache Spark (based on Hadoop)
- computational primitive: MapReduce



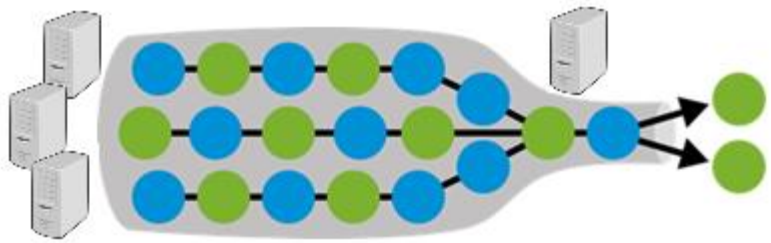
Cloud Computing

- the workflow of distributed machine learning algorithms in a large-scale system can be decomposed into three functional phases: a storage, a communication, and a computation phase
- for many Machine Learning algorithms:
  - Communication: data shuffling
  - Computation: matrix multiplication



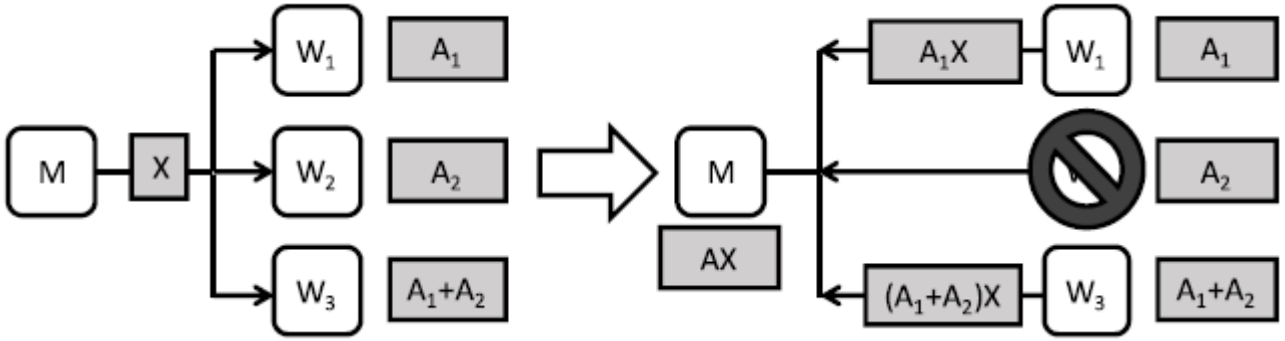
Cloud Computing Bottlenecks

- performance of a modern distributed system is significantly affected by anomalous system behavior and bottlenecks
  - straggler nodes
  - system failures
  - communication bottlenecks



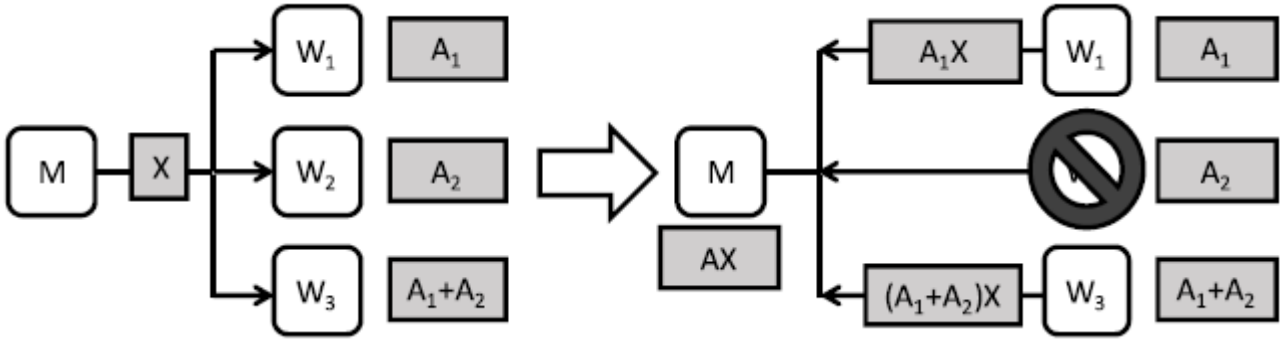
Coded Computation example

- a system with three worker nodes, noted $\boldsymbol{W_i}$ and one master, noted $\boldsymbol{M}$
- the goal is to compute matrix multiplication $\boldsymbol{AX}$ for data matrix $\boldsymbol{A} \in \mathbb{R}^{q \times r}$ and input matrix $\boldsymbol{X} \in \mathbb{R}^{r \times s}$
- the data matrix is divided into two submatrices $\boldsymbol{A_1} \in \mathbb{R}^{q/2 \times s}$ and $\boldsymbol{A_2} \in \mathbb{R}^{q/2 \times s}$ stored in node 1 and 2; the sum of the two submatrices is stored in node 3
- after the master node transmits $\boldsymbol{X}$ the workers compute the multiplication and send the result back to the master
- the master can compute $\boldsymbol{AX}$ as soon as it receives **any** two computation results



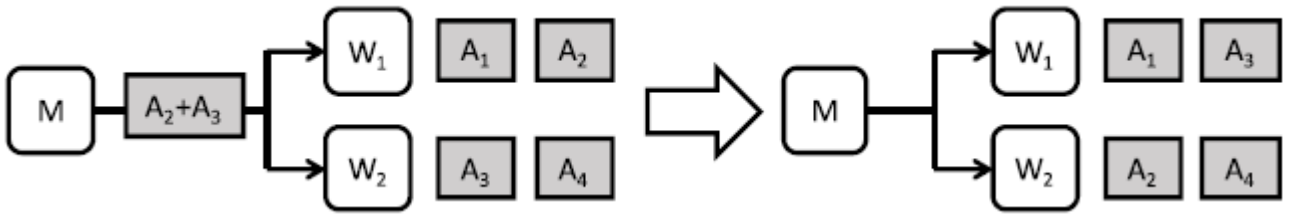
Coded Computation

- *Coded Computation* designs parallel tasks for a **linear operation** using **erasure codes** such that its runtime is not affected by up to a certain number of stragglers (may be node failures)
- Matrix multiplication is one of the most basic linear operations and is the workhorse of a host of machine learning and data analytics algorithms, e.g., gradient descent based algorithm for regression problems, power-iteration like algorithms for spectral analysis and graph ranking applications



Coded Shuffling example

- a system with two worker nodes, noted $\boldsymbol{W_i}$ and one master, noted $\boldsymbol{M}$
- the data set consists of 4 batches noted $\boldsymbol{A_i}$ which are stored as shown below
- the objective of the master is to transmit $\boldsymbol{A_3}$ to the first worker and $\boldsymbol{A_4}$ to the second
- the solution is that the master sends a multicast message $\boldsymbol{A_2 + A_3}$ to the worker nodes, since the workers can decode the desired batches using the stored batches
- compared to the naïve shuffling scheme in which the master node transmits $\boldsymbol{A_2}$ and $\boldsymbol{A_3}$ separately, this shuffling scheme can save 50% of the communication cost, speeding up the overall machine learning algorithm (considering that the cost of sending a multicast message is the same as the cost for sending a unicast message)



The core idea of optimizations

- Trade computation for communication by introducing redundancy into subtasks of a distributed algorithm such that the original task's result can be decoded from a subset of the subtask results
- Possible by using erasure codes

	Computation (sec)	Communication (sec)	Total (sec)
TeraSort	15.53	945.72	961.25

trade computation for communication

Coded Computing, Salman Avestimehr, et.al
http://www-bcf.usc.edu/~avestime/papers/CodedComputingWeb2018.pdf

Erasure code

- An erasure code is a method of introducing redundancy to information for robustness to noise.
- A method that encodes a message of *k* symbols into a longer message of *n* coded symbols such that the original *k* message symbols can be recovered by decoding a subset of coded symbols.
- The article proposes the usage of *MDS (Maximum Distance Separable)* codes

Conclusions

- *Coded matrix multiplication* significantly outperforms the one with the uncoded matrix multiplication in the *gradient descent* algorithm for linear regression; the average runtime is reduced by 31.3% to 35.7%.
- Using *Coded shuffling,* the communication overhead for data-shuffling is reduced by more than 81% compared to uncoded shuffling. Thus, at a very low storage overhead for caching, the algorithm can be significantly accelerated.