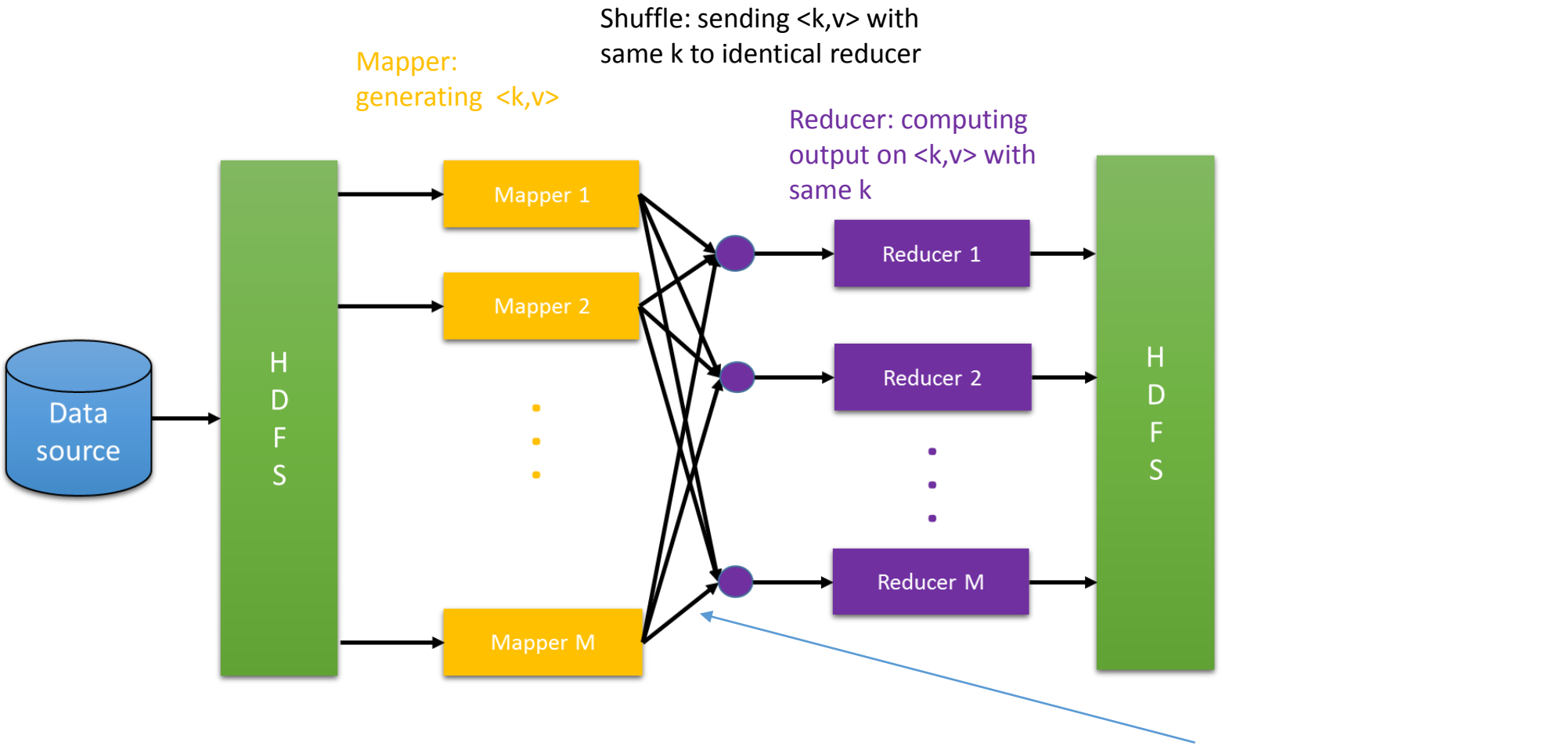
Efficient computation of aggregate functions in large scale data processing frameworks

Chao Zhang, Farouk Toumani and Emmanuel Gangler

University Clermont Auvergne, France

MapReduce Paradigm

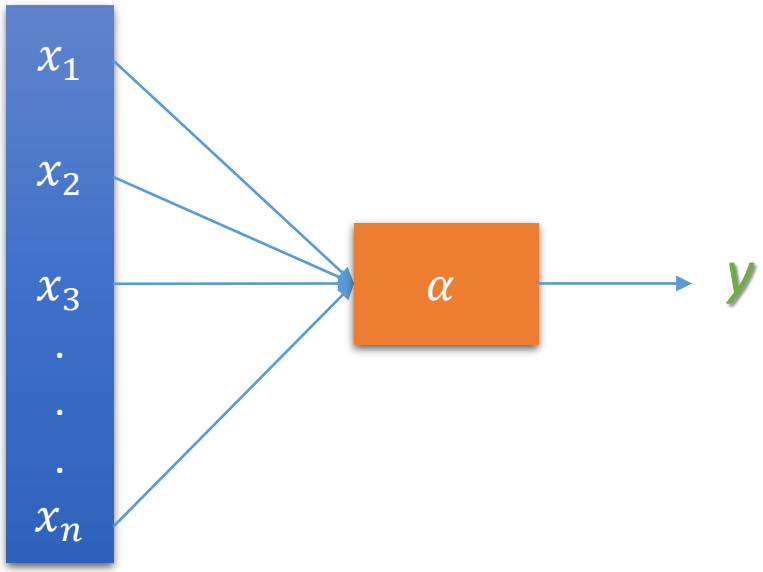


Communication cost： a bottleneck of MapReduce

Aggregation

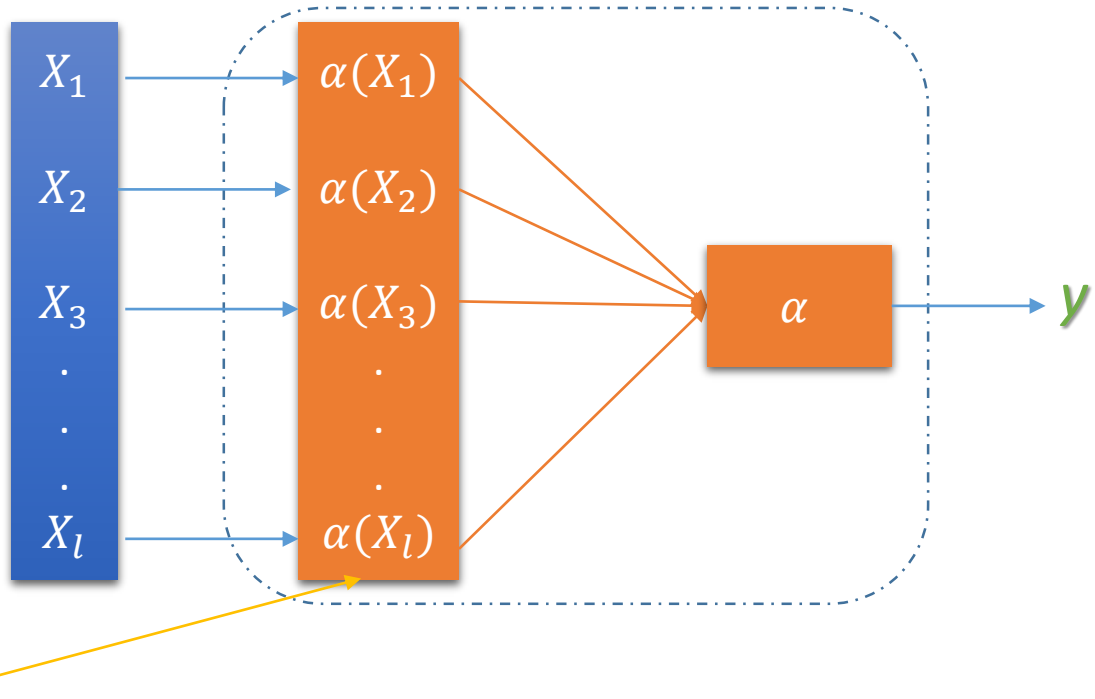
The inherent property of aggregation : taking all values as input and returning only one value as output.

Input : $X = \{x_1, x_2, x_3, \dots, x_n\}$
Output : $y = \alpha(X)$



The property to break down the dependency : **associativity**.

Input : $X = X_1 \cup X_2 \dots \cup X_l$
Output : $y = \alpha(X)$



Partial aggregation： reducing communication cost

Research agenda

Q: given an arbitrary aggregation α, when it is decomposable and how to decompose it?

- Classifying aggregations into symmetric (commutative) and asymmetric families.
- **Generic frameworks** to decompose aggregation.
- Properties for the frameworks to be efficient:
  - *Sufficient and necessary* condition for symmetric aggregation;
  - *Sufficient* condition for a class of asymmetric aggregation.

An overview of solutions