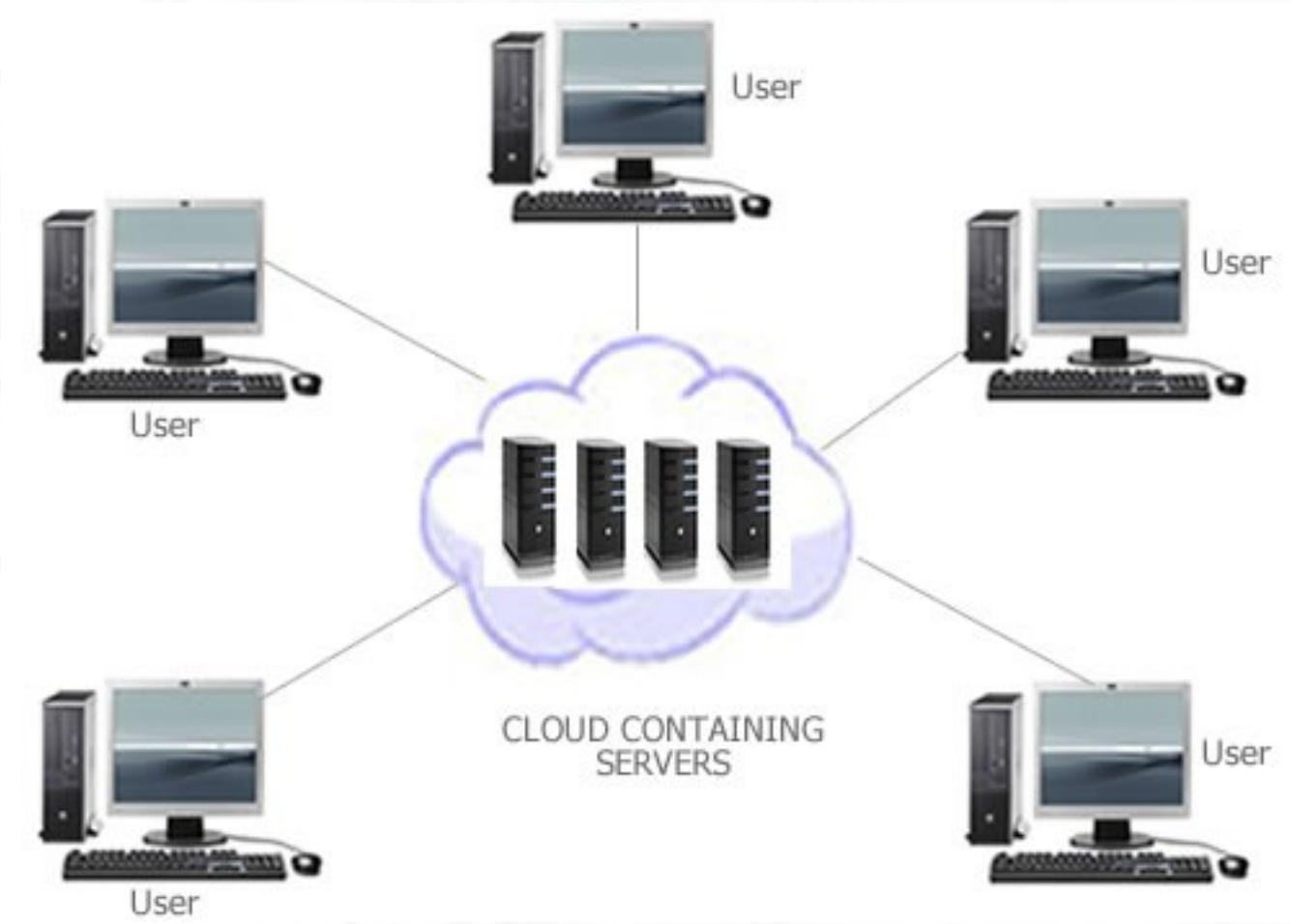
Overview of Cloud Computing large-scale processing technologies

What is Cloud Computing



How does Cloud Computing work

Large scale processing technologies

- Supercomputers
- Transaction Processing Systems
- Grid Systems
- Data Intensive Scalable Computing

MapReduce

- programming model to process large data sets
- high level abstraction
- 2 functions:

map (k1, v1) → list (k2, v2)
reduce (k2, list(v2)) → list (v2)

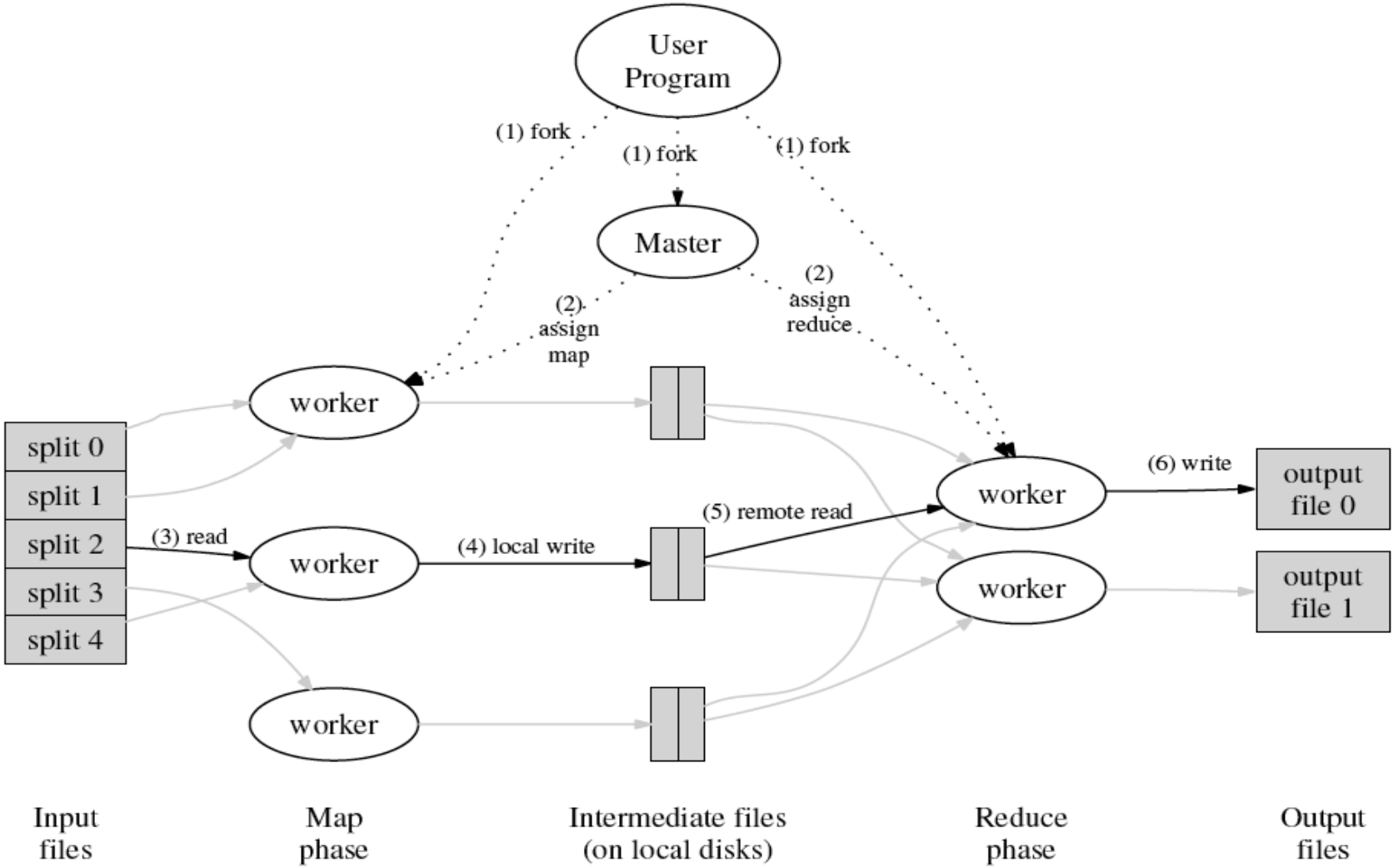
MapReduce

Example:

Map (document name, document contents)
↳ emit (word,”1”)

Reduce (word, list of counts)
↳ emit(word_count)

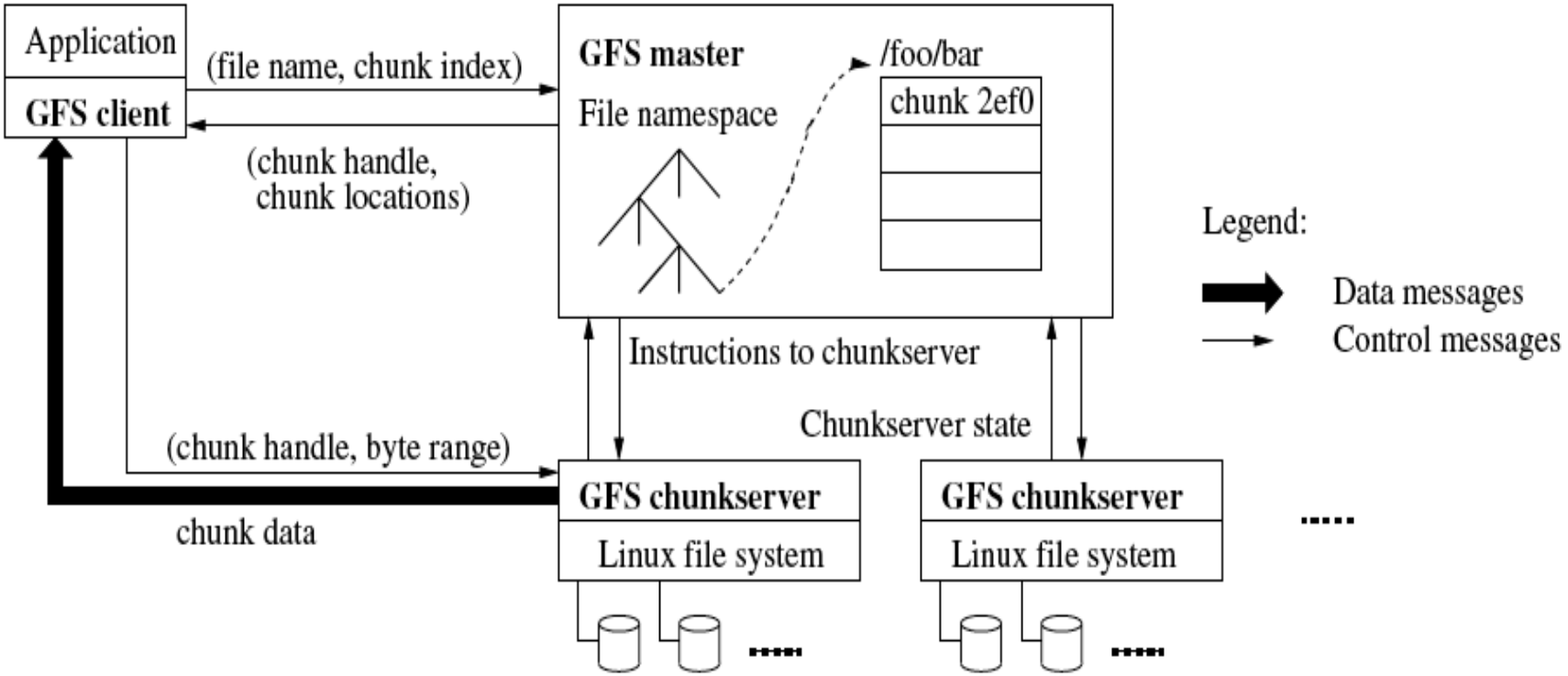
MapReduce


User Program
(1) fork
(1) fork
(1) fork
Master
(2) assign map
(2) assign reduce
worker
worker
worker
worker
worker
split 0
split 1
split 2
split 3
split 4
(3) read
(4) local write
(5) remote read
(6) write
output file 0
output file 1
Input files
Map phase
Intermediate files (on local disks)
Reduce phase
Output files

Google File System (GFS)

- Distributed file system
- Used by MapReduce and other Google technologies
- Similar tools as in normal system: *open, close, create, delete, read, write*
- Fault management, replication, monitoring

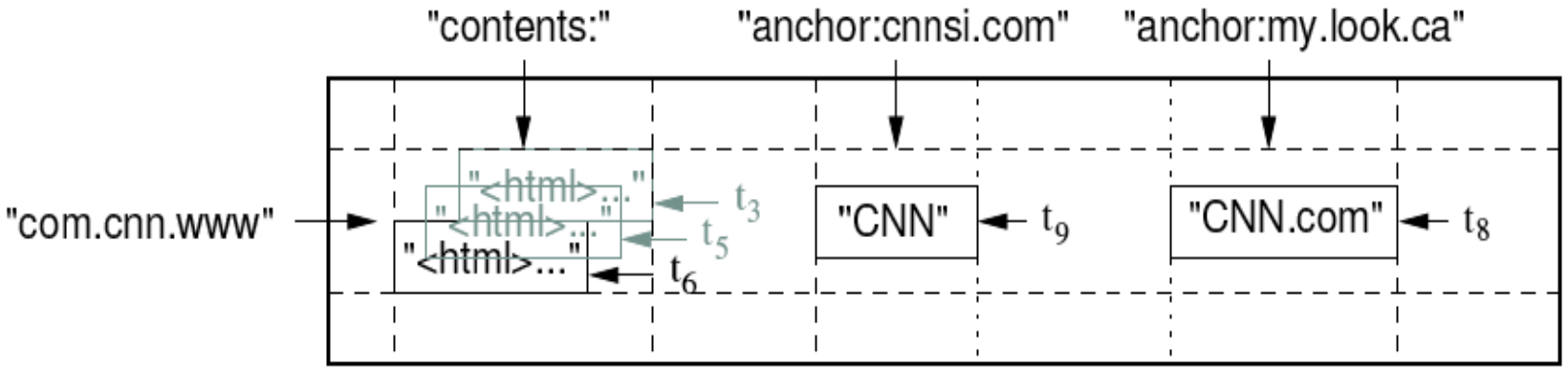
Google File System (GFS)



Bigtable

- Large scale database
- Based on GFS
- Non relational
- Only simple read/write operations

Bigtable



(row:string, column:string, time:int64) → string

Implementations

- Google implementations
- Open-source projects
- Commercial implementations
- Research and proof of concept implementations

Use cases and products

and many others...

Literature

- MapReduce, Bigtable, GFS publications
- Open-source projects documentation
- Interest groups and community websites
- Product vendors websites

Questions....