

---

## Discussion Session 4:

# Programming Assignment 2

Department of Electrical Engineering  
Stanford University

<http://eeclass.stanford.edu/ee282>

# Today's Agenda

---

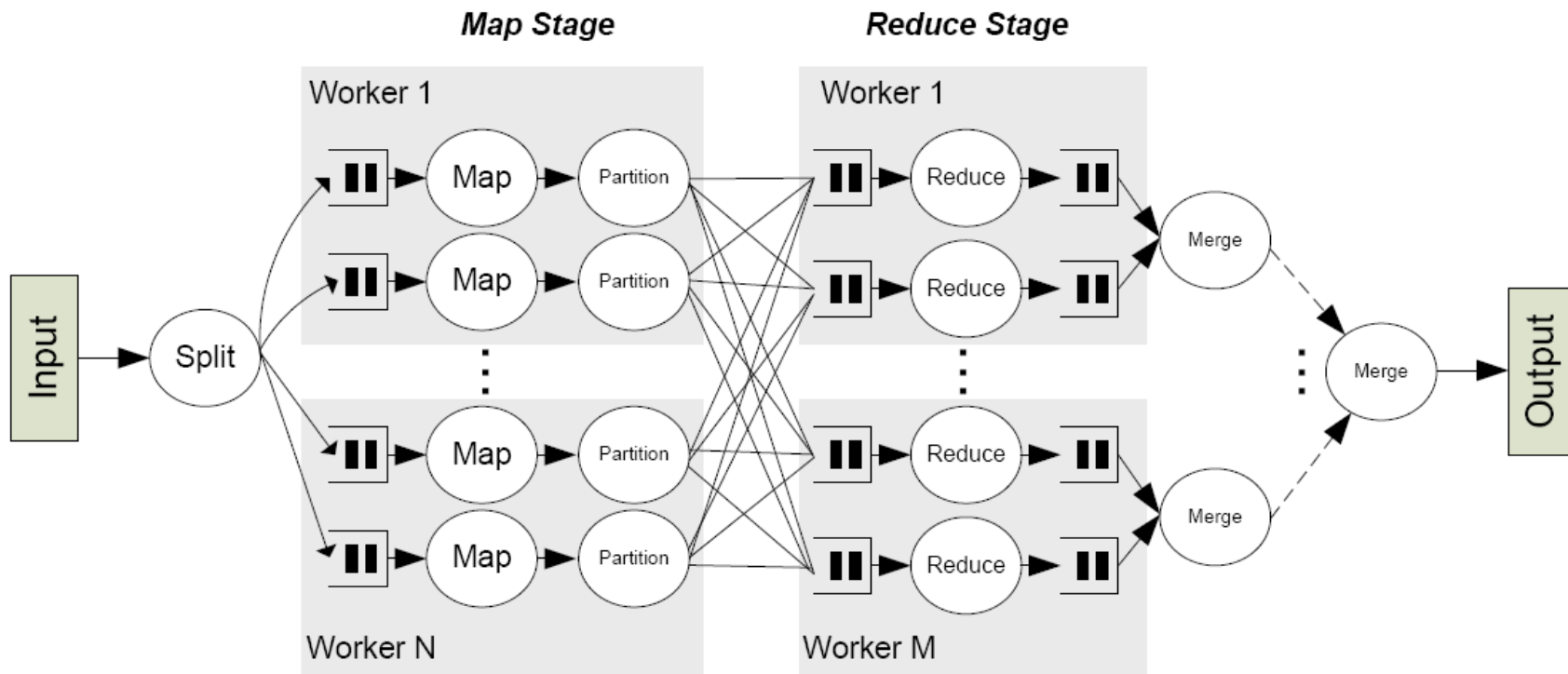
- Programming Assignment 2
- Map-Reduce example: WordCount

# Programming Assignment 2

---

- Handout available on the website
  - Contains much of what we'll go over today
- Write 3 Map-Reduce Applications
  - DigramCount
  - WordCorrelate
  - FragmentFinder
- Perform basic performance experiments
- Report (answer questions posed by handout)
- Due December 3<sup>rd</sup> @ 5PM PST

# Map-Reduce



# Example: Word Count

---

wget/untar pa2.tar.gz

source setup.sh

look at tiny.txt

look at WordCount.java

make tiny wordCount (inspect output)

make check (inspect out/part-00000)

<http://kastri:50030/>

mapTasks -> 16

# Application 1: DigramCount

---

- Definition
  - Digram: pair of letters
  - E.g. “science” has the digrams  
sc ci ie en nc ce
- Count all digrams in the input text
- Output example:
  - th 10030
  - to 3299
  - ...

## Application 2: Word Correlate

---

- Goal: Find words that appear on the same line as each other

- Example

- input: “how about that”

- output:

- how:about            1

- how:that             1

- about:how            1

- about:that           1

- that:how             1

- that:about           1

# Application 3: Fragment Finder

---

- Find most frequent 5-word sentence fragment that starts with a word
  - i.e. What is the most frequent fragment that starts with 'and'?
    - “and at the same time”
  - What word? Every word.
- \*2\* Map-Reduce passes
  - Pass 1: Find and count all fragments
  - Pass 2: Find most frequent fragment that starts with a word

# Report

---

1. Correctness results
  - Turn in the output of 'make check' for each application
2. Sensitivity analysis
  - Vary the dataset size
  - Vary the number of map tasks
  - Evaluate the efficacy of Combiner classes