Distributed Systems Project, Spring 2015

Jussi Kangasharju
Course Outline

3 exercises to look at distributed systems in practice

Exercises mostly programming

2 individual exercises, 1 group exercise

Groups of up to 3 people allowed

Group work not mandatory, but recommended
Course Schedule

13.1. Start of first exercise (individual)
15.1. Q&A for first exercise

20.1. Deadline for first exercise

20.1. Start of second exercise (individual)
22.1., 27.1., and 29.1., Q&A for second exercise

3.2. Deadline for first and second exercise

3.2. Start of third exercise (group)
5.2., 10.2., 12.2., 17.2., 19.2., 24.2., and 26.2. Q&A for third exercise

8.3. Deadline for third exercise
People

Jussi Kangasharju
   Office hour: Tue 13-14 or ask for appointment by email

Liang Wang
   Office hour: During meetings or ask appointment by email

Twitter: #UnivHelsinkiCS_DSP15 (also visible on course page)
Assignments

Distributed algorithms

   Individual assignments about algorithms

Hadoop/Spark

   Use Hadoop/Spark to analyze a data set

Overlay networks

   Design, analyze, and implement an overlay network

Details for assignments 2 and 3 presented later
Grading

Each assignment graded on scale 1-5
Must get at least 1 in every assignment

Same grade for all members of group

Overall grade is weighted average of assignment grades

Assignments 1 and 2: Weight 1
Assignment 3: Weight 2
Assignment 1: Algorithms

Link to assignment will be posted to course website
Individual Assignments on Distributed Algorithms

1. Lamport clocks
2. Vector clocks
3. Bully election algorithm
4. Gossiping

Simple programs communicating over the network
Select assignment: (student ID % 4) + 1
General Idea

Multiple programs on different machines

Everybody knows everybody

Programs communicate to implement a given algorithm

Key points: Network communication, correct algorithm
House Rules

Configuration file for nodes and ports
Format:

<ID> <IP/HOST> <PORT>

Command line argument indicates what is client’s ID
File has an arbitrary number of lines

Must conform to specified output format
Deviation results in a reduced grade
Programs must be runnable on Ukko cluster
Assignment 2: Hadoop/Spark

Link to assignment will be posted to course website
Working with Large Data Sets

Recall MapReduce and Spark from Distributed Systems course

Hadoop = Open source implementation of MapReduce (and several other things)
Spark = Implementation of Spark

Assignment goal: Get familiar with Hadoop/Spark and MapReduce

Task: Work with a large data set
Practical Matters

Provide your user ID to us
Either today or by email to Liang

We create work directories for everyone
Next steps

Q&A session on 15.1.

Deadline for returning January 20th at 10:00

Details for next exercise announced on 20.1.

Return to Liang.Wang@cs.helsinki.fi

See assignment sheet for instructions