EEL6871 Autonomic Computing Fall 2013 Homework 3

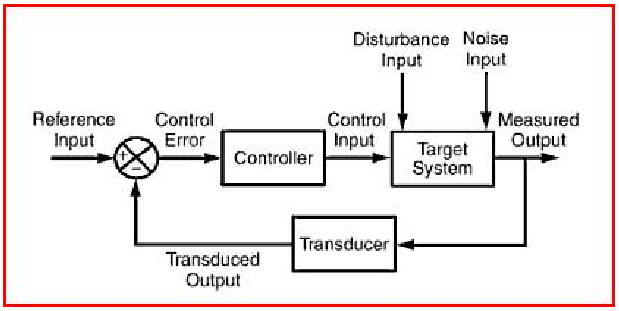
Assigned: Tuesday, Sep-17-2013
Due: Thursday, Sep-26-2013 5pm

Overview

The purpose of this homework is to collect data points from a Hadoop cluster build using the same procedure of the previous homework and to construct a model of this Hadoop system. Using this model you will, in a later assignment, develop a controller for the system.

Context

The following figure is a block diagram of a feedback control system, also colloquially referred to as a closed-loop system or MAPE (Monitor, Analyze, Predict and Execute) loop. In this homework we are concerned with developing a model of the Target System.



Specifically, the Target System is your choice of Hadoop application and its supporting Hadoop cluster. The Control Input is the number of Mappers allocated to your application and the Measured Output is the program execution rate. Note that both Disturbance Input and Noise Input are not being considered here.

Procedure

The development of system models requires experimental data. You will collect a set of data points (# of mappers, job execution rate). To control how many mappers are allocated to a job, you will start a Hadoop cluster with a fair scheduler plugged in. Navigate to /usr/local/hadoop/conf/fair-scheduler.xml on the master and edit the following code snippet.

```
<allocations>
  <pool name="root">
    <minMaps>1</minMaps>
    <minReduces>1</minReduces>
    <maxMaps>81</maxMaps>
    <maxReduces>81</maxReduces>
    <minSharePreemptionTimeout>300</minSharePreemptionTimeout>
  </pool>
</allocations>
```

The maxMaps and maxReduces properties determine the maximum number of mappers and reducers allocated to the root user, as you will usually login to the Hadoop cluster as root. If you run only one job at a time, then these tags can be used to control the maximum # of mappers and reducers allocated to the current running job. By default, each node in a cluster can run two mappers and two reducers at most. That is to say if your cluster has n nodes, then the maximum number of mappers or reducers is *2n.* There is no use in setting these properties bigger than *2n.* After you modify these properties, you can go to the Hadoop scheduler UI (https://master:50030/scheduler)to confirm the changes.

The job execution rate is derived from the job completion percentage devided by the time elapsed so far in hours. For example, if your job started at 15:51:46 and the current time is 15:58:46 the mapers has finished 36%, then job execution rate is 36/(7/60) = 308.57

Start a Hadoop cluster with n nodes and vary your mappers from *1* to *2n.* Collect the corresponding data points. You are to determine a linear difference equation, in the form of Equation 2.5 in the testbook, to serve as a model of the Target System. Read Section 2.4 of the text for the details on how this is to be done. Note that you will develop a single model from the *N=1, 2, ..., 2n* data collected.Use N=n as the operating point. Be careful that you job lasts long enough for you to have enough time to vary the number of mappers from *1* to *2n.* It's better to have a script to change these tags at a certain interval and just measure the job completion percentage in the Hadoop output.

Deliverables

To demonstrate you have completed the assignment successfully you will submit the following items:

1. In a readme.txt file include the target system related information:
   - Command line to run your chosen application.
   - Cluster size
   - Application running time
   - Collected data points
2. The linear difference equation to be used as a model of the target system based on the data you collected.
3. A plot like Fig. 2.11 showing your data points compared to the model you determined. Have clearly labeled axes with **units** and an appropriate title.
4. The RMSE ,R^2 and Correlation Coefficient calculation showing the accuracy of your model.

5. The MATLAB (or whatever language) code you used to produce the model, plot, RMSE, R^2 and CC.

Submission Policy

- Follow strictly to the format specified in every homework. Incorrect submission formats will lead to a grade reduction.
- All submissions are expected by the deadline specified in the homework assignment. Grade is automatically reduced by 25% for every late day.
- Make sure you test your submitted code using the tar file you submitted before submission. If untar, make, compile or any other command needed to execute your program do not work, your homework grade will be zero.