Multivariate Resource Performance Prediction in the NWS

Martin Swany
UCSB

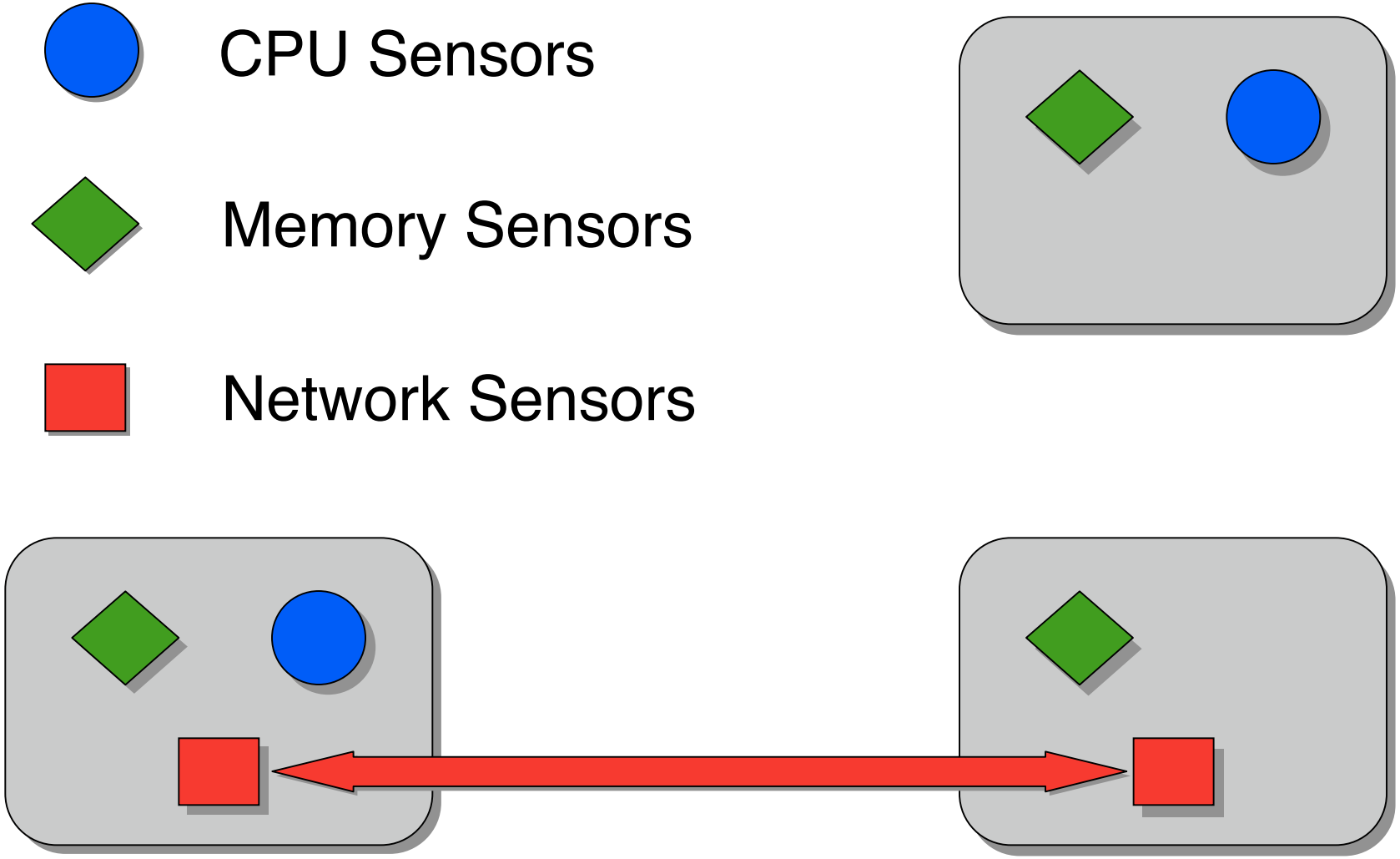
UCSB

NWS Architecture

CPU Sensors

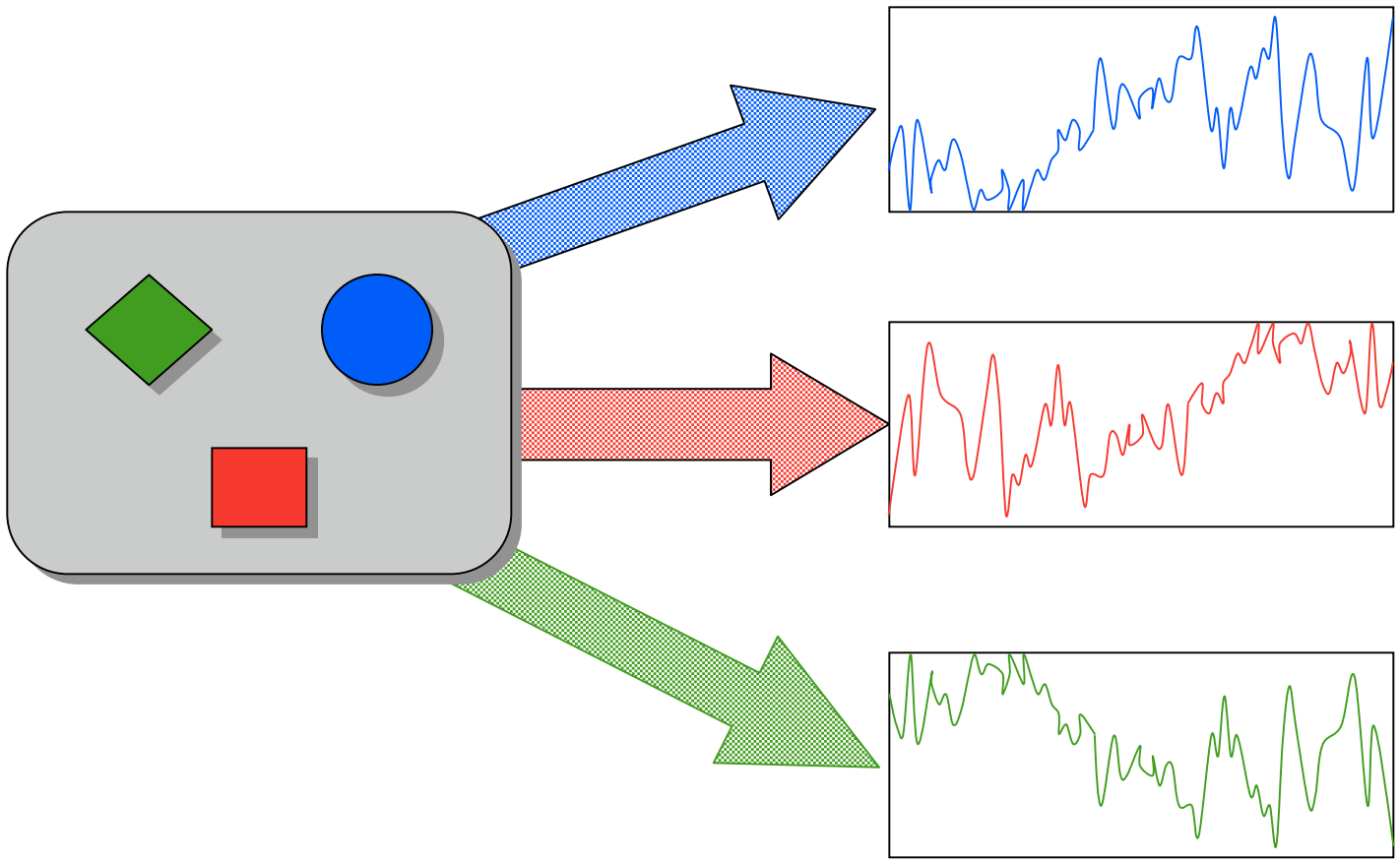
Memory Sensors

Network Sensors



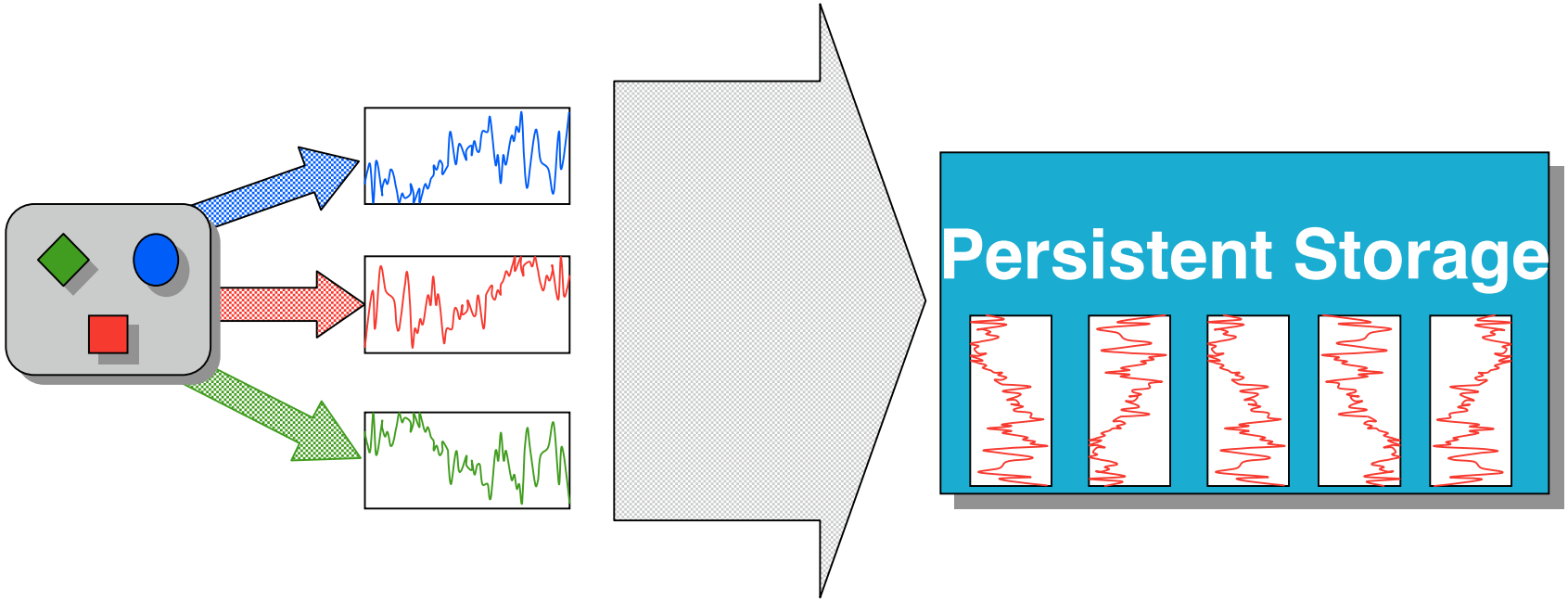
UCSB

Sensor Elements Produce Time Series



UCSB

Time Series are Stored



In the nws_memory

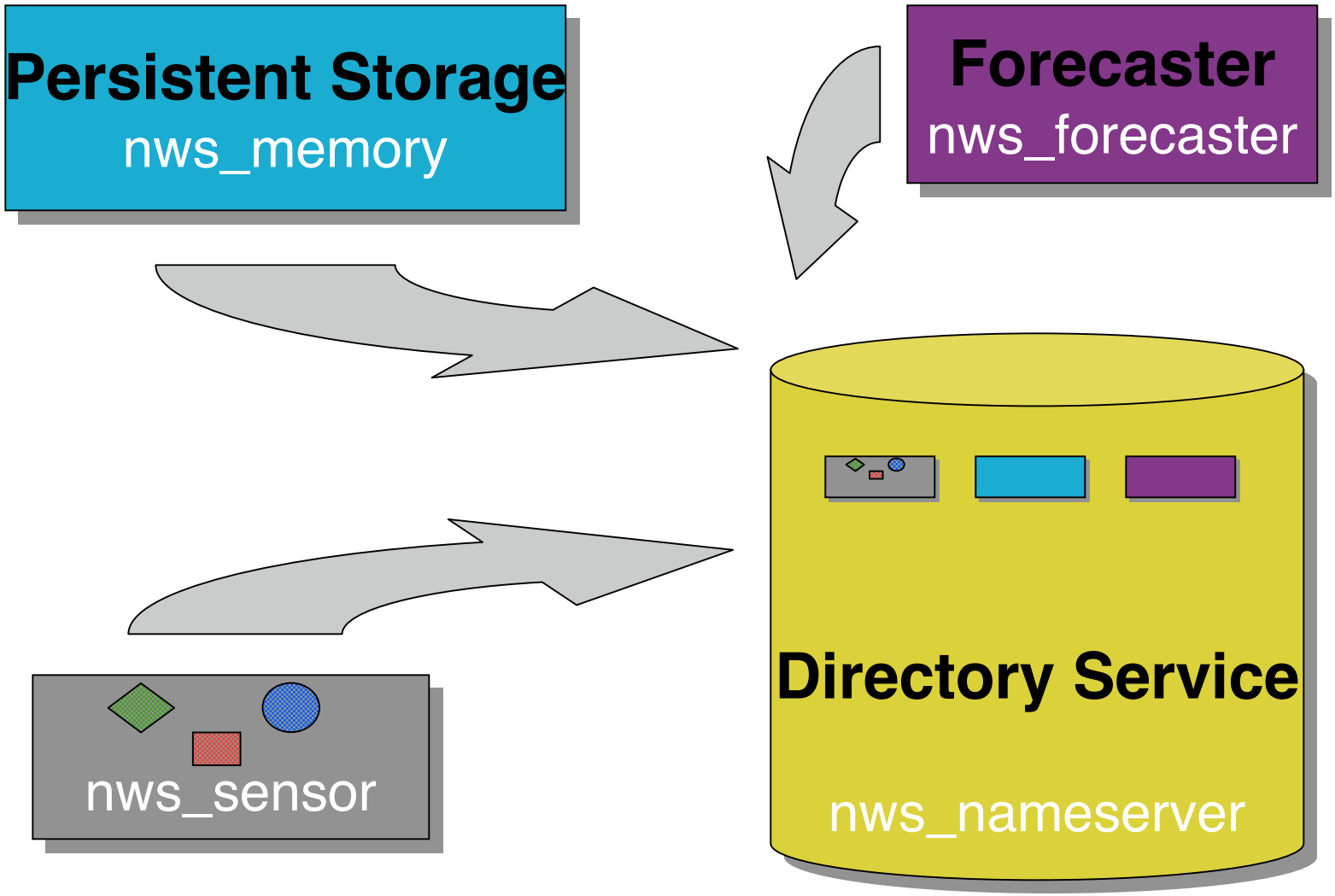
UCSB

Time Series are Passed through Forecasters

Implemented as a library or a daemon - nws_forecaster

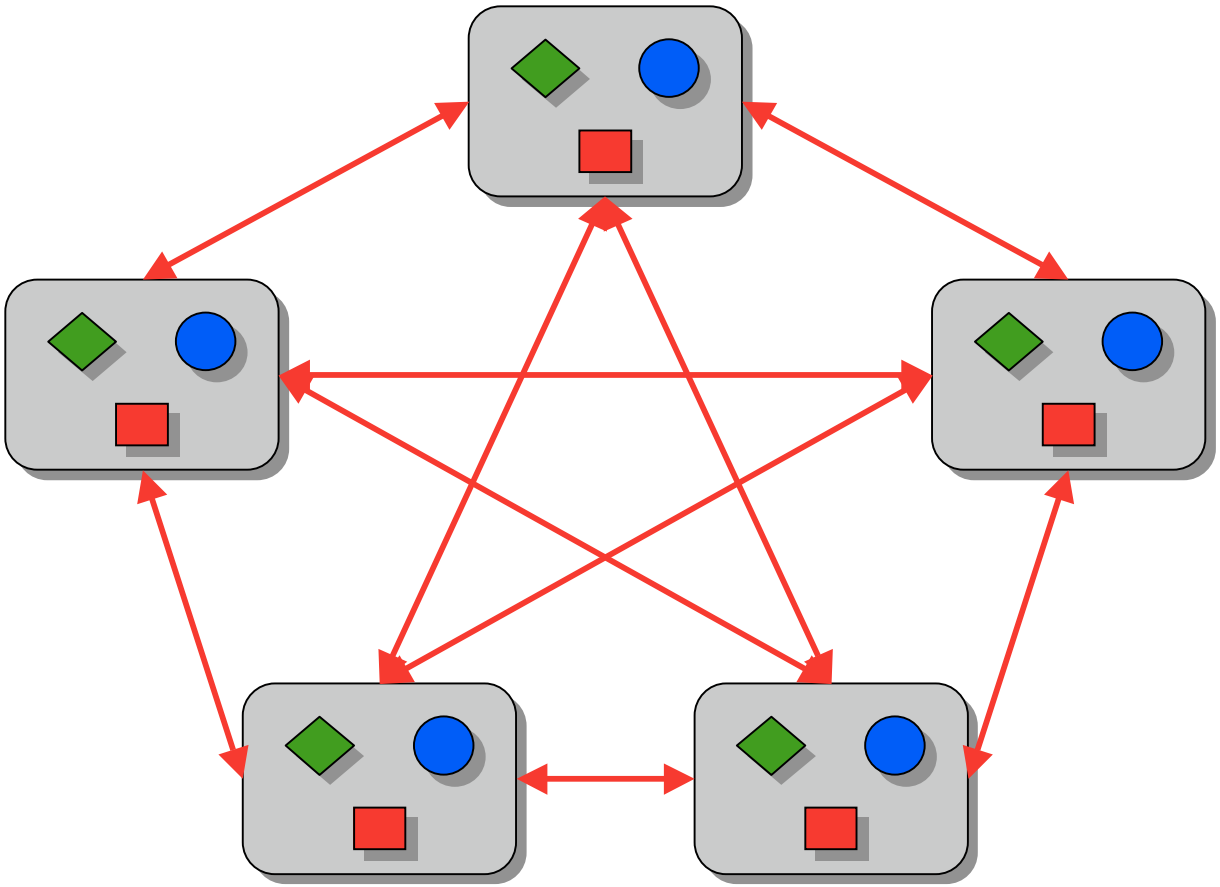
UCSB

Elements Register with the Directory Service



UCSB

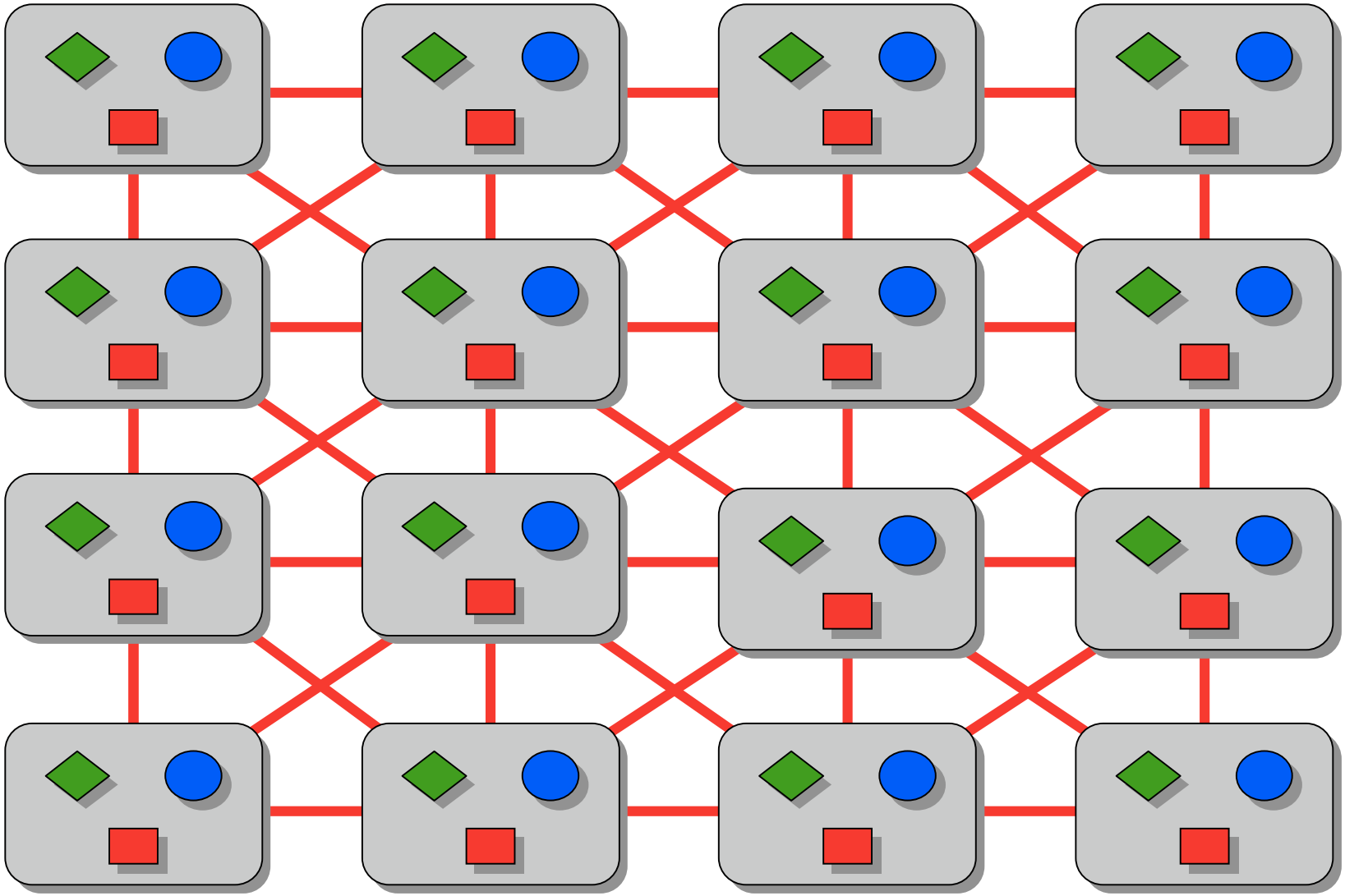
Network Sensors Form Peer Groups



Called “cliques”

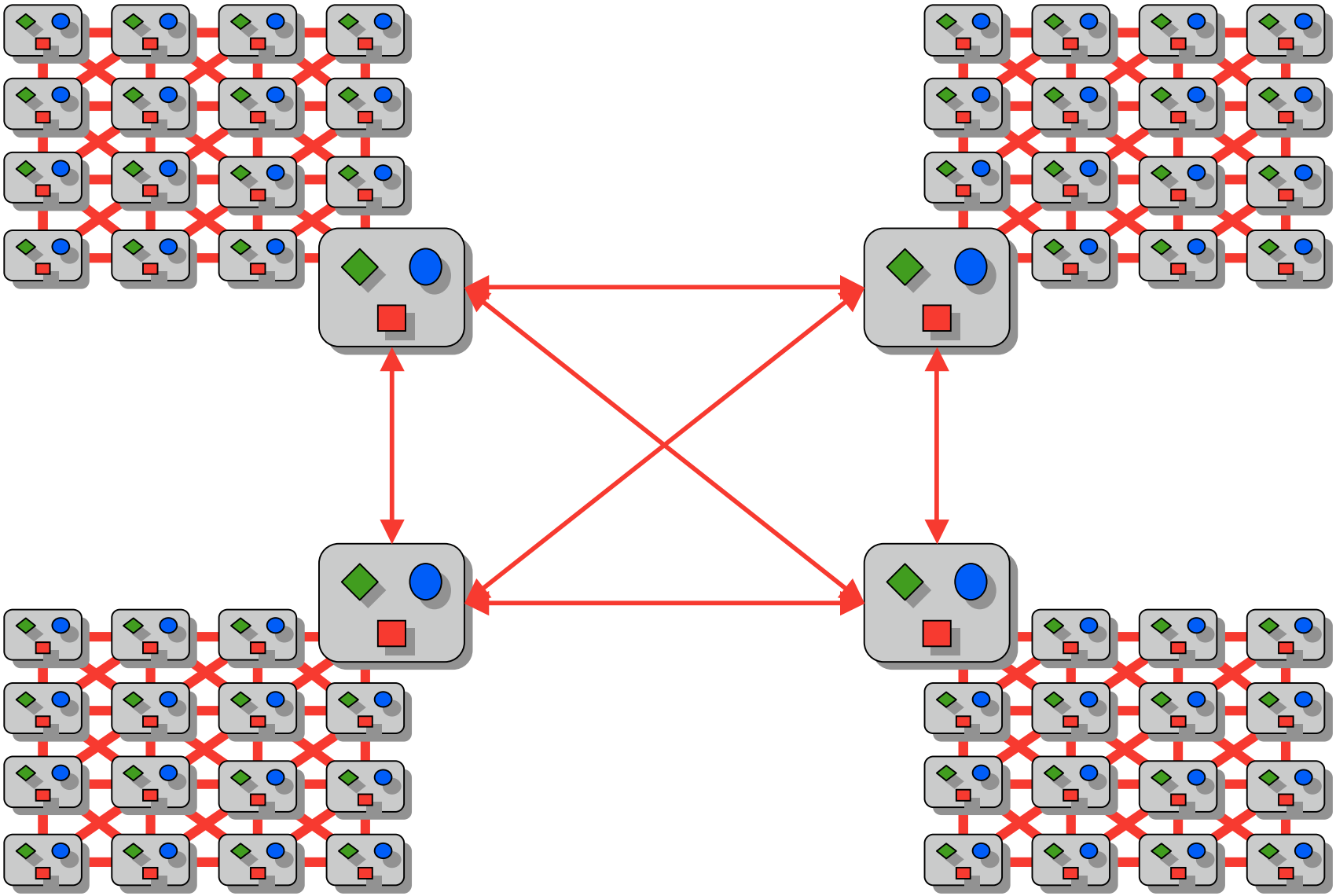
UCSB

A Fully-Connected Cluster or Site



UCSB

A Hierarchy of Cliques for scalability



UCSB

Measurement Issues

- Active measurements are intrusive
- More accuracy often means more intrusiveness
- The NWS defaults to lightweight probes that have questionable usefulness
  - 64KB transfers
  - Clearly not a measure of nominal bandwidth!
    - Especially as bw *delay grows
- Nominal bandwidth isn't the idea here

UCSB

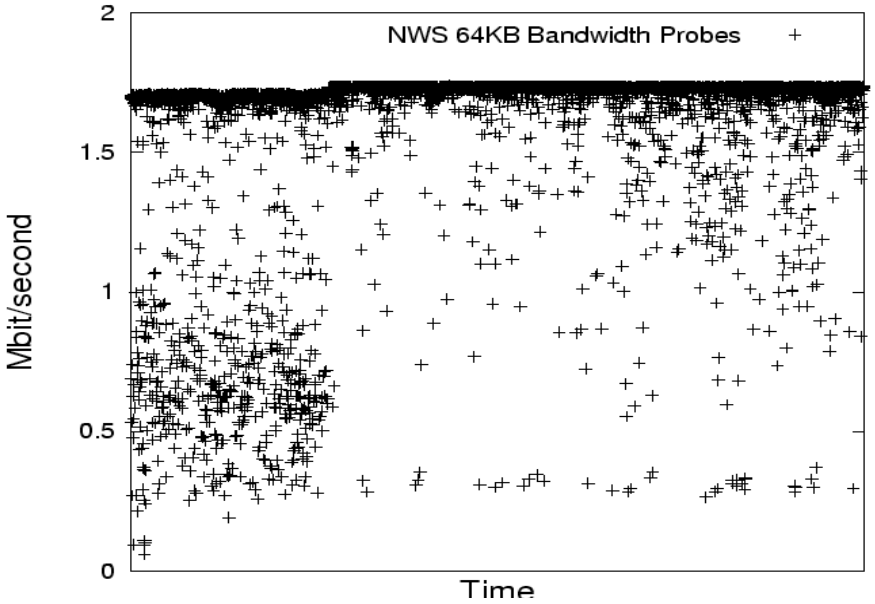
Some philosophy

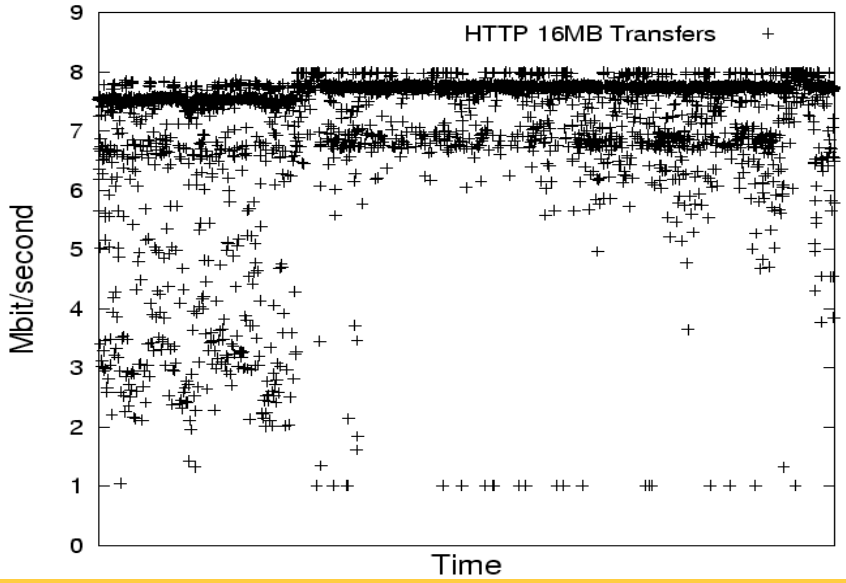
- Resource performance is extremely dynamic
- Presenting up to date performance information to distributed systems is essential for their effective operation
  - Particularly the case for Grid environments
- Most measurement in this context is used as a prediction of future performance
- No privelaged access
- There is no mind / body dualism

UCSB

How much Information?

- Intuitively, there seems to be a relationship between shorter and longer measurements
  - Go ahead, squint
- But, they are significantly different





UCSB

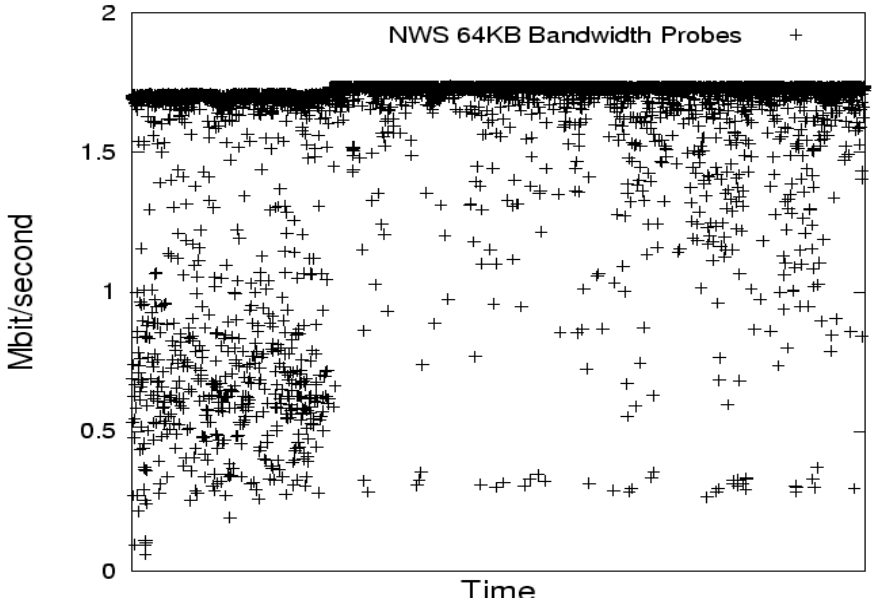
Experimental Methodology

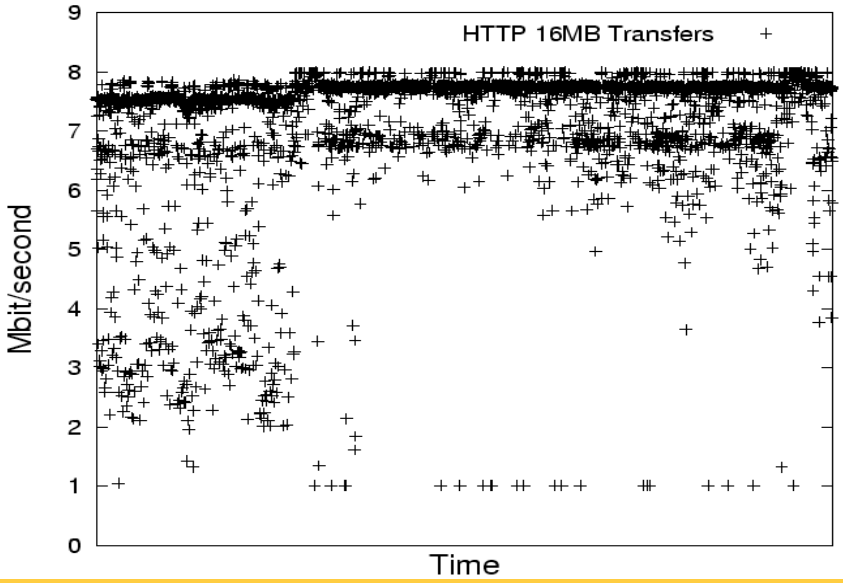
- Collect 64KB NWS measurements every 10 seconds
- Time 16MB HTTP transfers every 60 seconds
  - Using wget
  - The file to be transferred comes from the filesystem
  - Heavily used, general purpose system
    - It can't all be in the buffer cache

UCSB

Recall this picture

- Factor of 4 (ish) difference in maximum throughput
- Figure out the difference and use a simple linear scaling function?





UCSB

The relationship isn't that easy

- The relationship isn't necessarily linear
  - This makes the regression more difficult
- The relationship might change over time
- The problem of data matching is tricky, too
- Perhaps different amounts of information
  - Particularly if application instrumentation data is used.

UCSB

Recent Work

- Experiments of Network Throughput Measurement and Forecasting Using the Network Weather Service, P. Primet, R. Harakaly, F. Bonnassieux (INRIA, ENS-Lyon), CCGrid 02
- Attempts to compute the relationship between NWS data and Iperf data with a magic scaling factor
  - We tried this too, but weren't happy with the results

Recent Work

- Predicting Sporadic Grid Data Transfers, S. Vazhkudai, J. Schopf, to appear HPDC 11
- Focuses on using NWS data to predict GridFTP transfers
- Explores a variety of matching techniques for regression

UCSB

Multivariate Forecasters

- We want to take a suspected predictor **X**, and use it to make forecasts of a target **Y**

UCSB

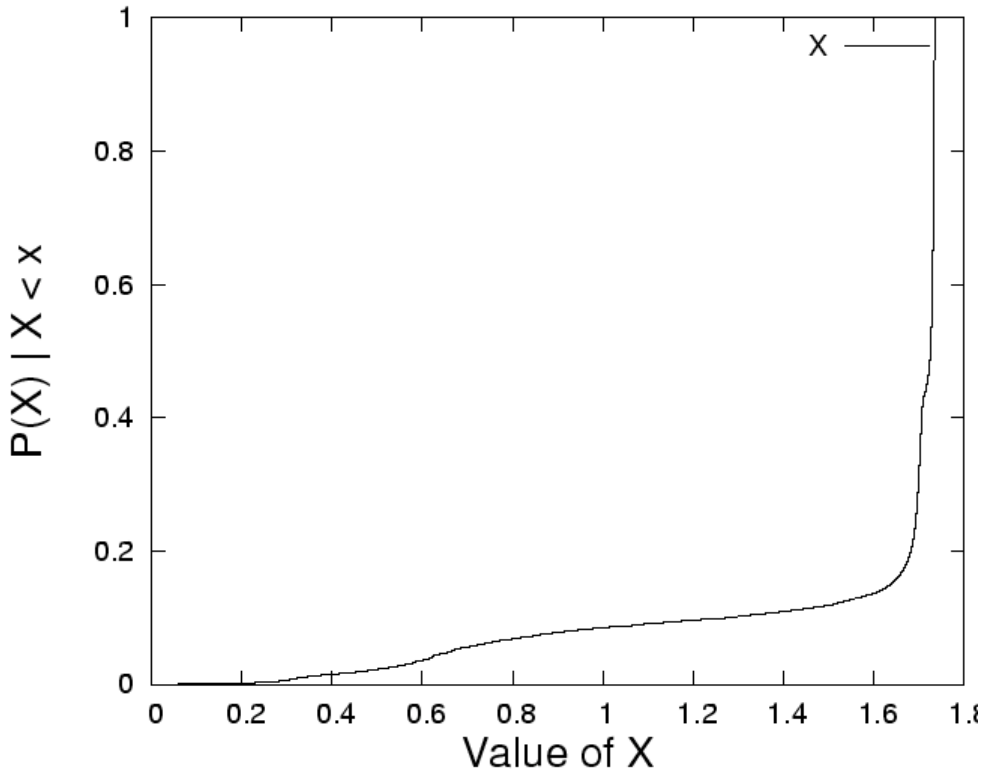
Correlation Mechanism

- Most correlation mechanisms assume normal distributions (as they deal with standard deviation, etc.)
- Network traffic does not enjoy a normal distribution
  - (see, well, lots of stuff)
- Distribution-free correlation mechanisms such as Spearman Rank correlation assume datasets of the same size

The Cumulative Distribution Function

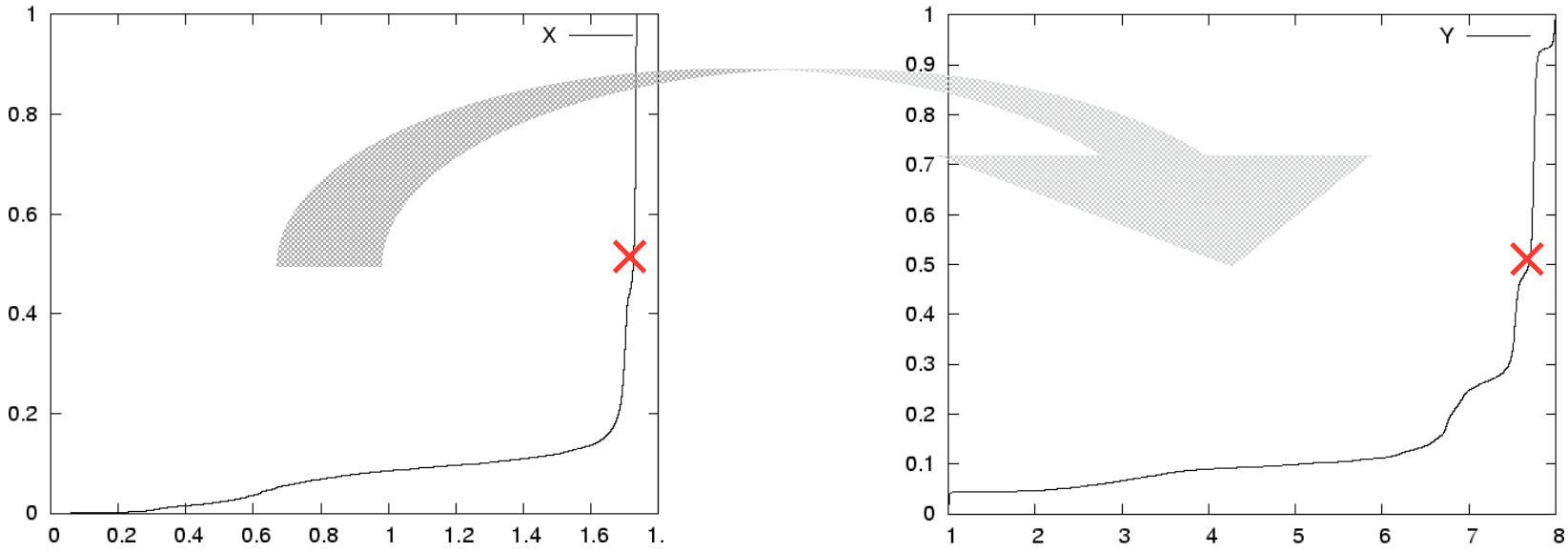
- The empirical CDF is defined as

$$CDF_X(x) = \sum_{i=1}^{position_x} \frac{1}{comnt_x}$$



UCSB

So, our CDF correlator uses the CDF to translate X into Y



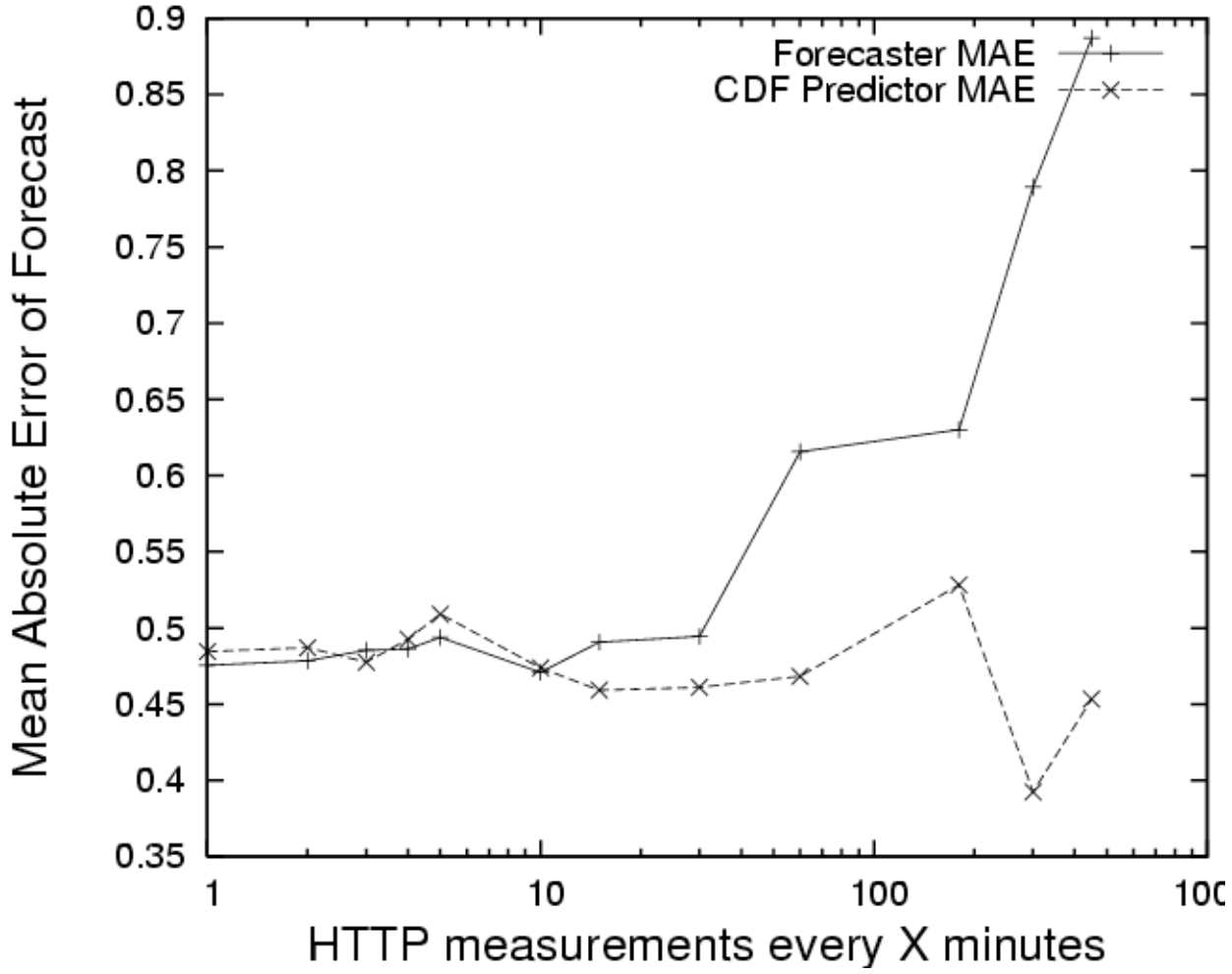
UCSB

Some Terminology

- MAE – Mean Average Error
- MSE – Mean Square Error
- MNEP – Moving Normalized Error Percentage

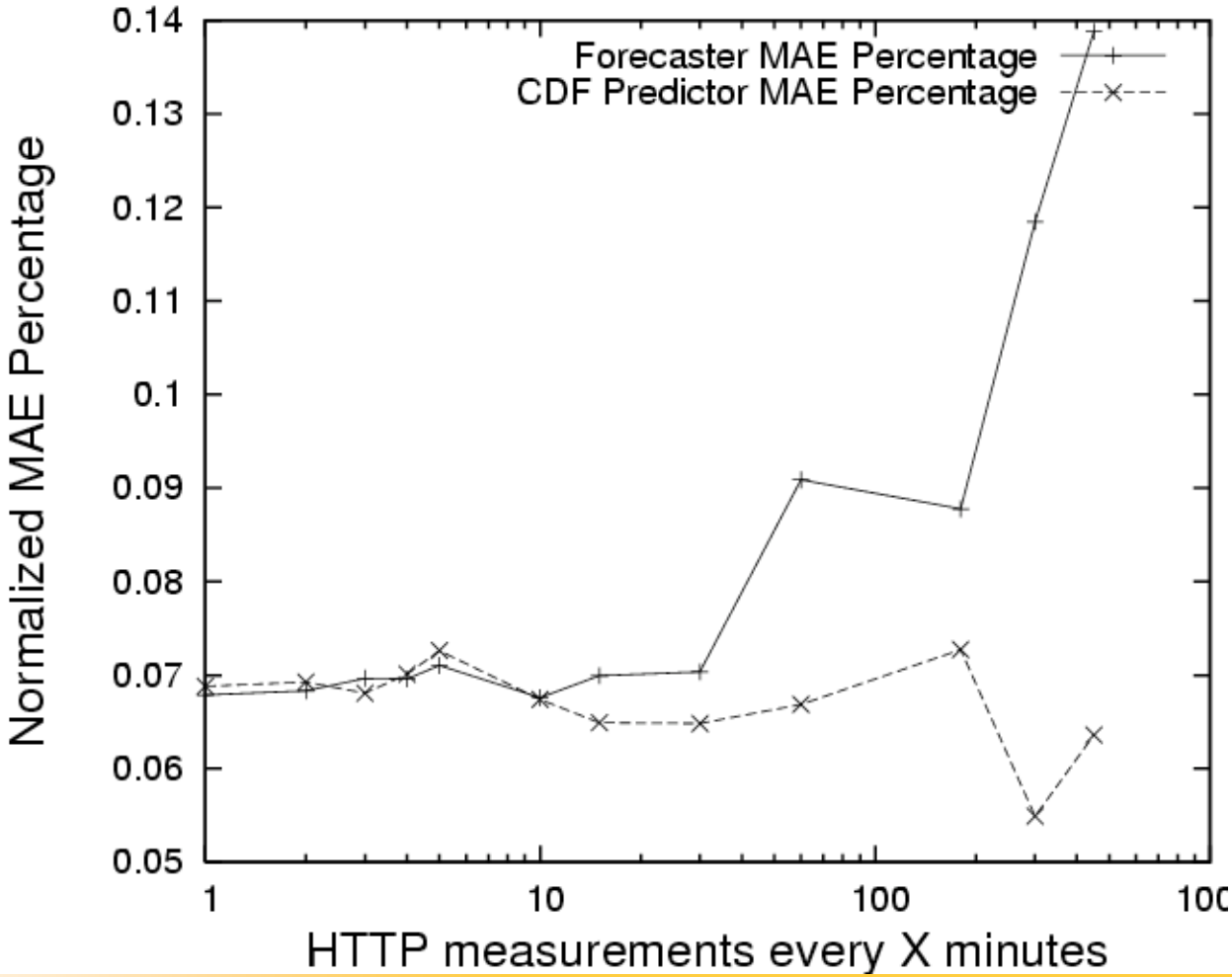
UCSB

Comparison of Mean Absolute Error (MAE) between univariate and multivariate forecasts for different frequencies of HTTP measurements



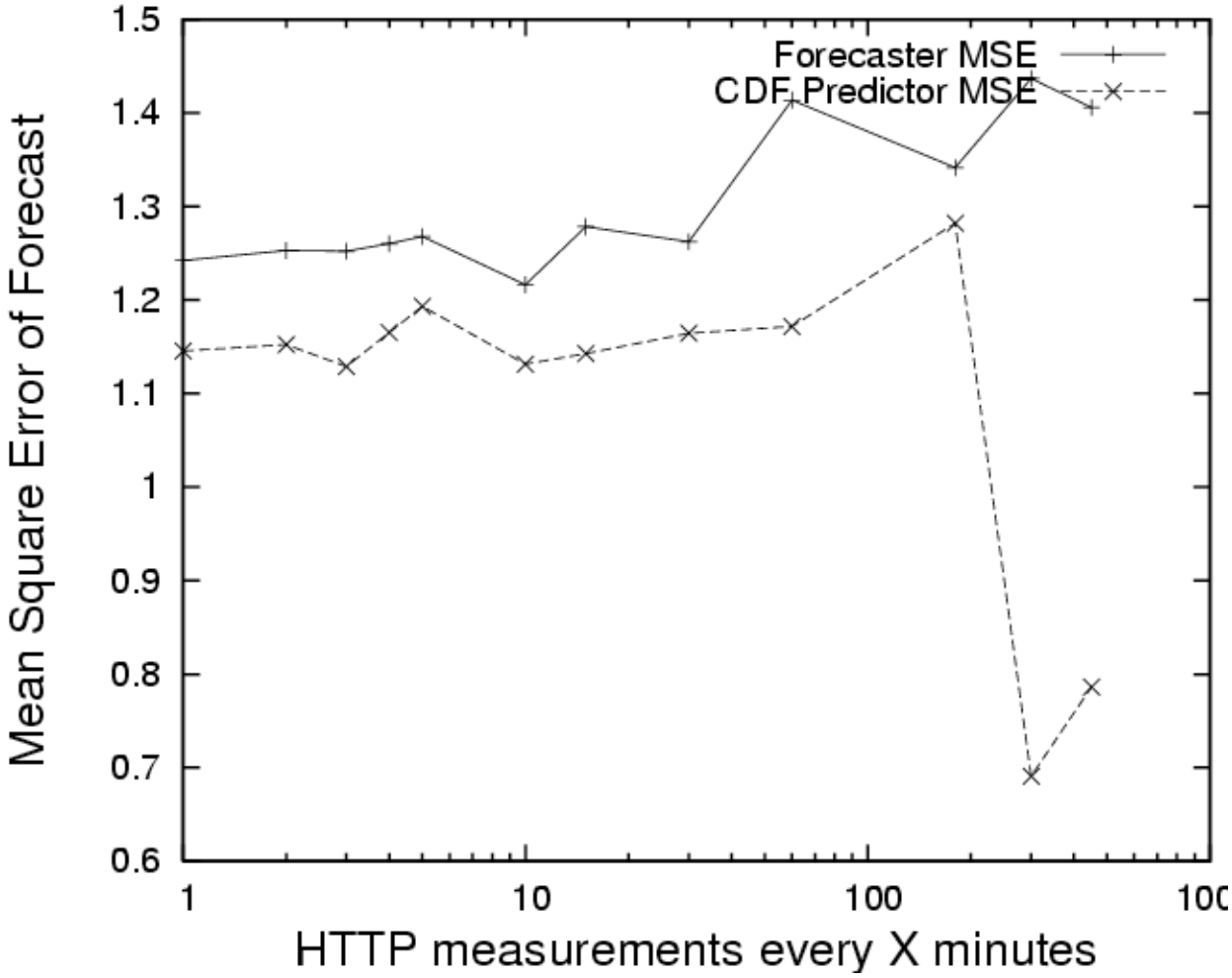
UCSB

Comparison of Moving Normalized Error Percent (MNEP) of the Mean Absolute Error (MAE) between univariate and multivariate forecasts for different frequencies of HTTP



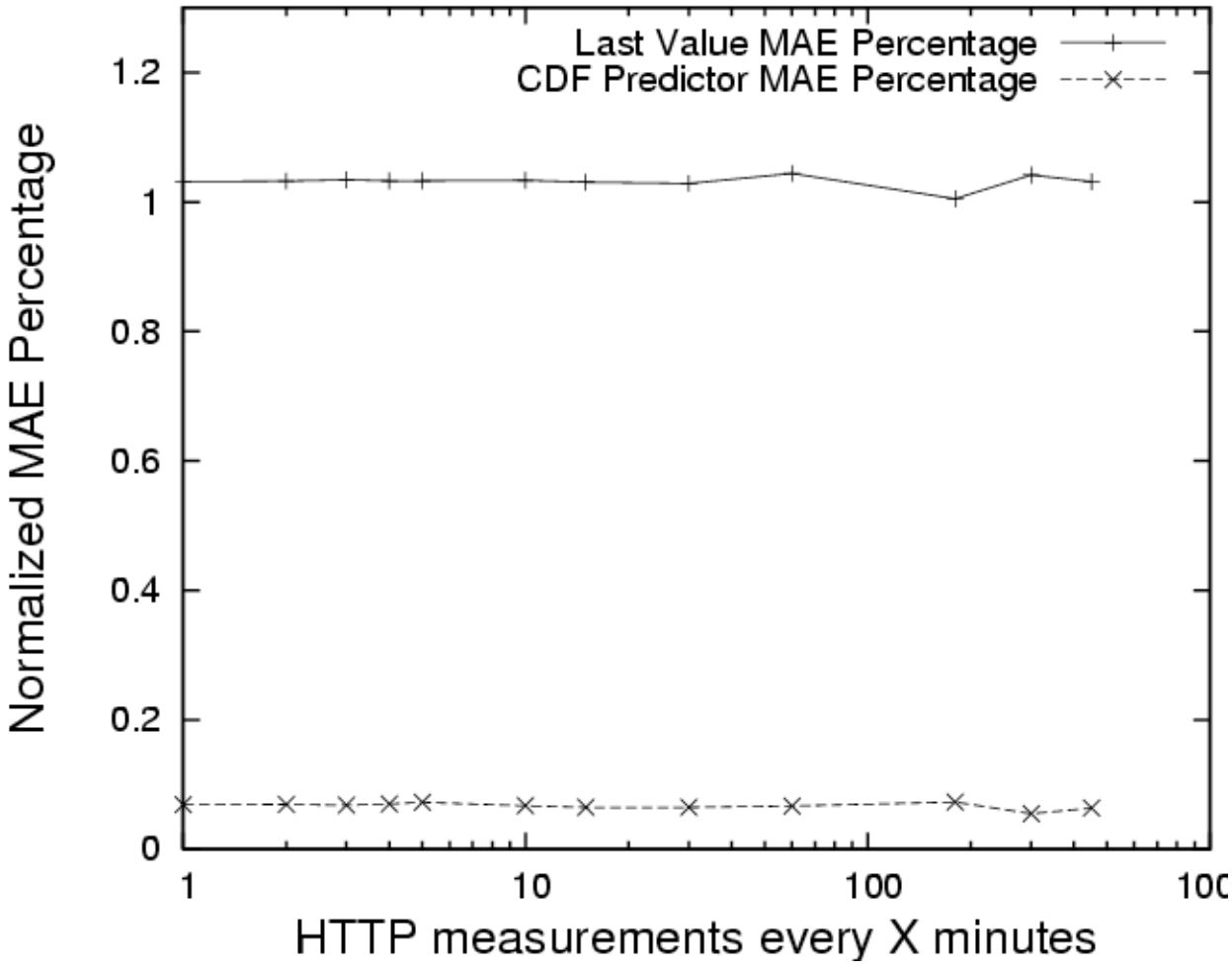
UCSB

Comparison of the square root of the Mean Square Error (MSE) between univariate and multivariate forecasts for different frequencies of HTTP measurements.



UCSB

Comparison of Moving Normalized Error Percent (MNEP) of the Mean Absolute Error (MAE) between ``Last Value'' and multivariate forecasts.



UCSB

Conclusions

- We have developed a novel multivariate prediction technique
- Much yet to be done, although the CDF is proving useful in situations where a distribution is required

UCSB