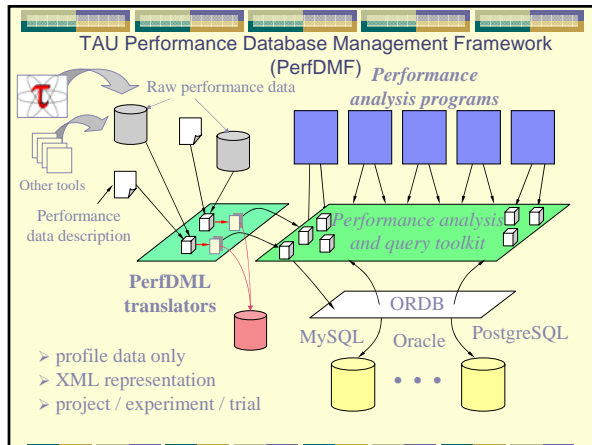


# Performance Databases and Multivariate Statistical Analysis

Shirley Moore  
shirley@cs.utk.edu

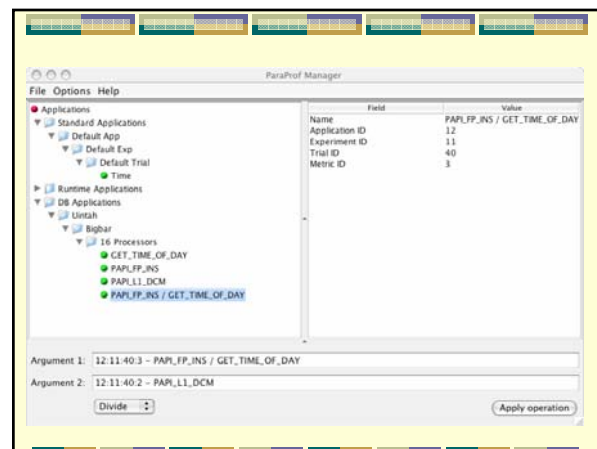
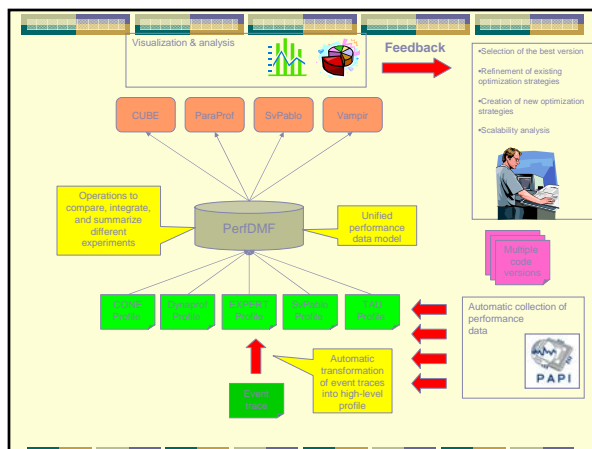
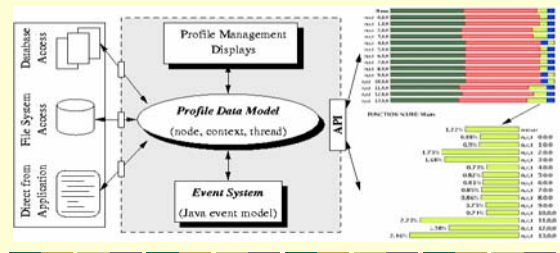
## Scalability and Interoperability Issues

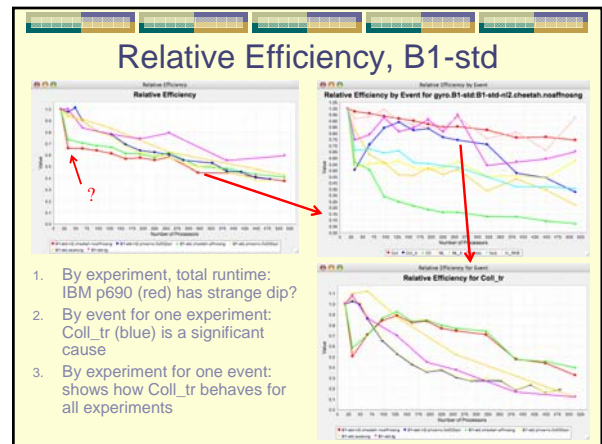
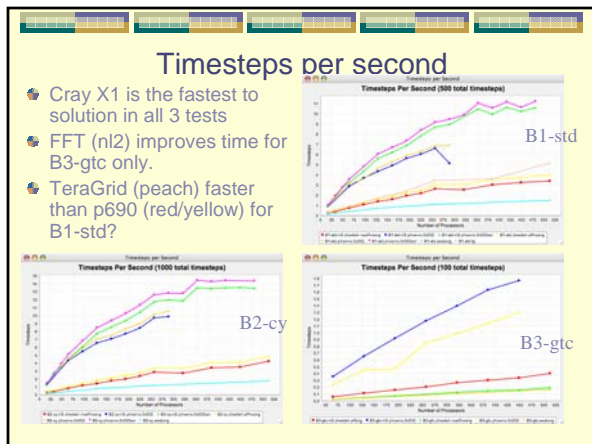
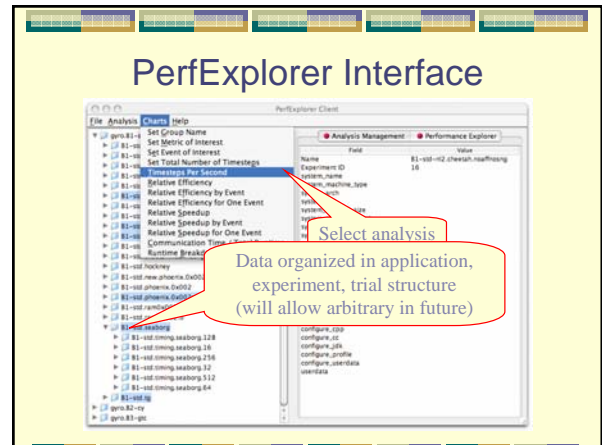
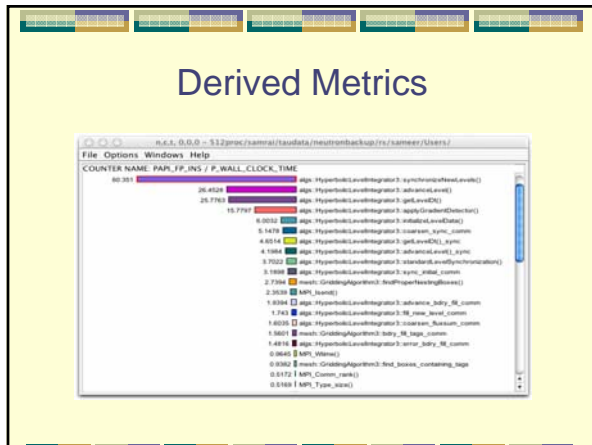
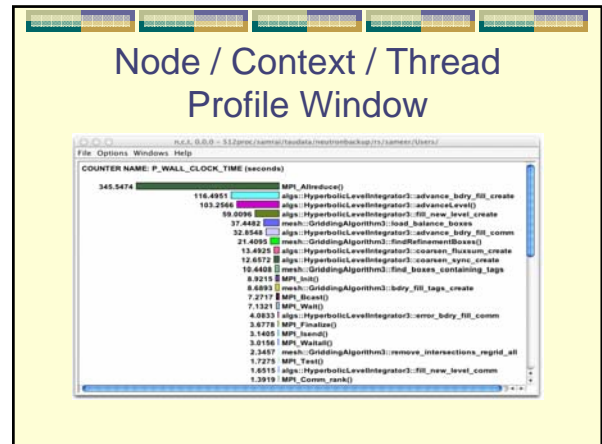
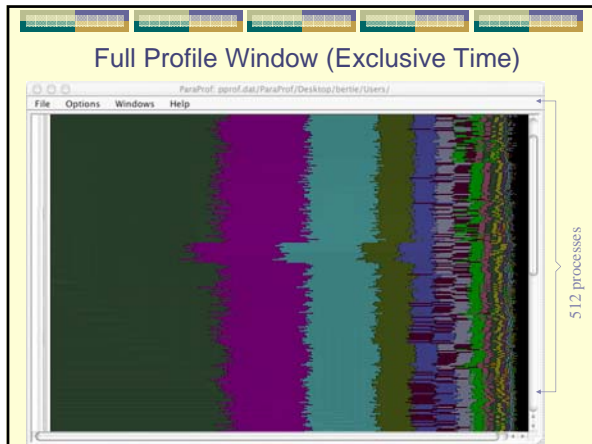
- Need to organize and archive large amounts of performance data
- Need to preserve “metadata” – e.g., machine and compiler configuration, test case, application info
  - Important for repeatability
- Need for automated analysis and correlation of large amounts of data for different metrics and from different sources and tools
- Need for multi-experiment analyses

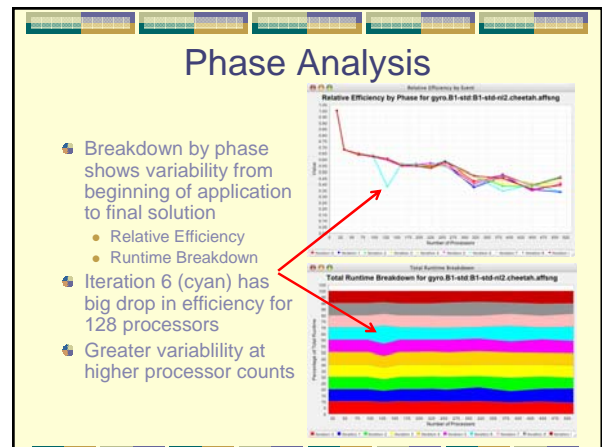
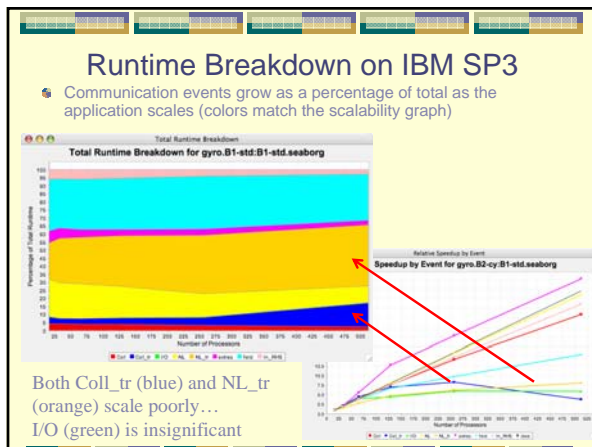
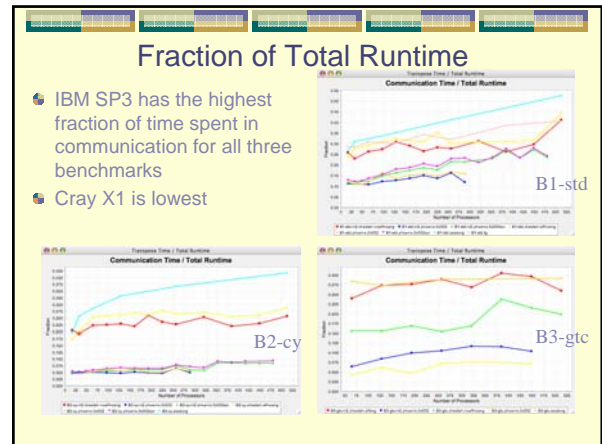
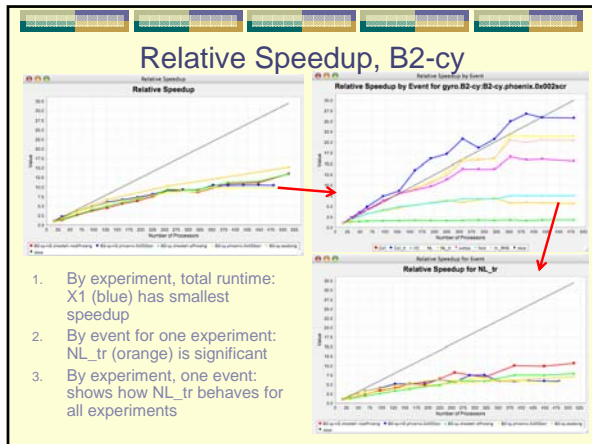


## ParaProf Architecture

- Portable, extensible, scalable tool for profile analysis
- Upload performance data from file system, database, and/or at runtime







### Multivariate Statistical Analysis

- Focus attention on important metrics and show the distribution of those metrics across parallel tasks and code regions
- Statistical reductions can be used as filters to reduce the total amount of performance data, either at the time the data are generated or at analysis time.

### MATLAB Statistical Toolbox

- Multivariate Statistics
  - Principal Components Analysis (PCA)
  - Factor Analysis
  - Multivariate Analysis of Variance (MANOVA)
  - Cluster Analysis
  - Multidimensional Scaling

## Principal Components Analysis

- More than one variable may be measuring the same driving principle.
- Often there are only a few driving forces for dozens of system variables.
- Principal Components Analysis (PCA) generates a new set of variables, called *principal components*.
  - Each principal component is a linear combination of the original variables.
  - Principal components are orthogonal so that there is no redundant information.

## Cluster Analysis

- Creates groups of objects, called *clusters*, or *equivalence classes*, such that objects in the same cluster are similar and objects in different clusters are distinct.
- Hierarchical (e.g., dendrogram) and non-hierarchical (e.g., K-means clustering) methods

## Combined PCA and Cluster Analysis of GYRO Profile Data

ps96.steps:

```
>> R = buildMatrix( data, 11 );  
>> [counters, newdataset, percent_explained] = findPrincipalCounters( R , 3 );  
>> [indices, q] = findClusters( newdataset , 2 );  
>> plotClusters( newdataset , indices )  
>> xlabel('FPUD produced a result');  
>> ylabel('Processor cycles');  
>> zlabel('FPU executed PDIV instruction');
```

