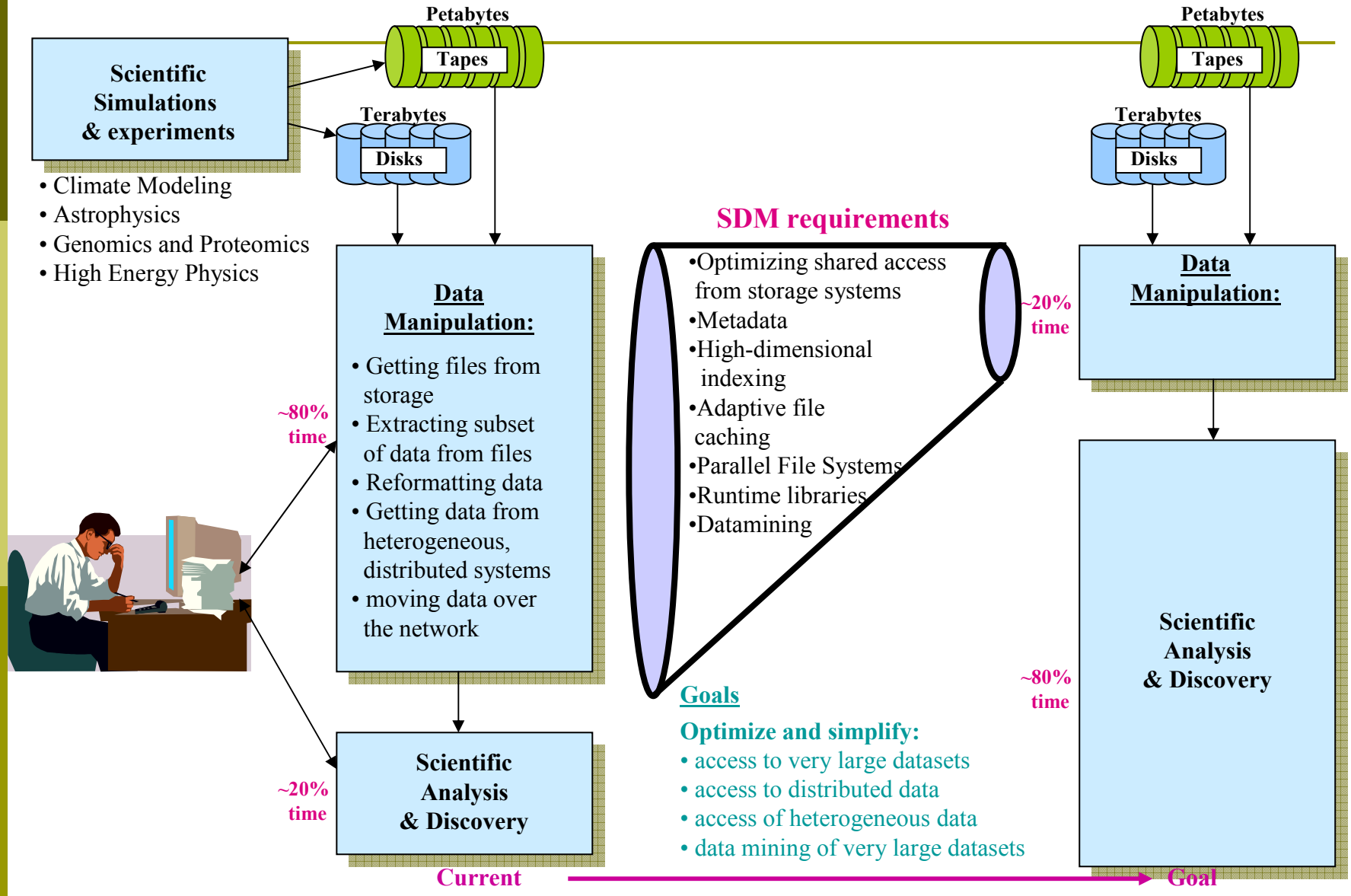
Large-Scale Scientific Knowledge Discovery: Problems and Potential Approach

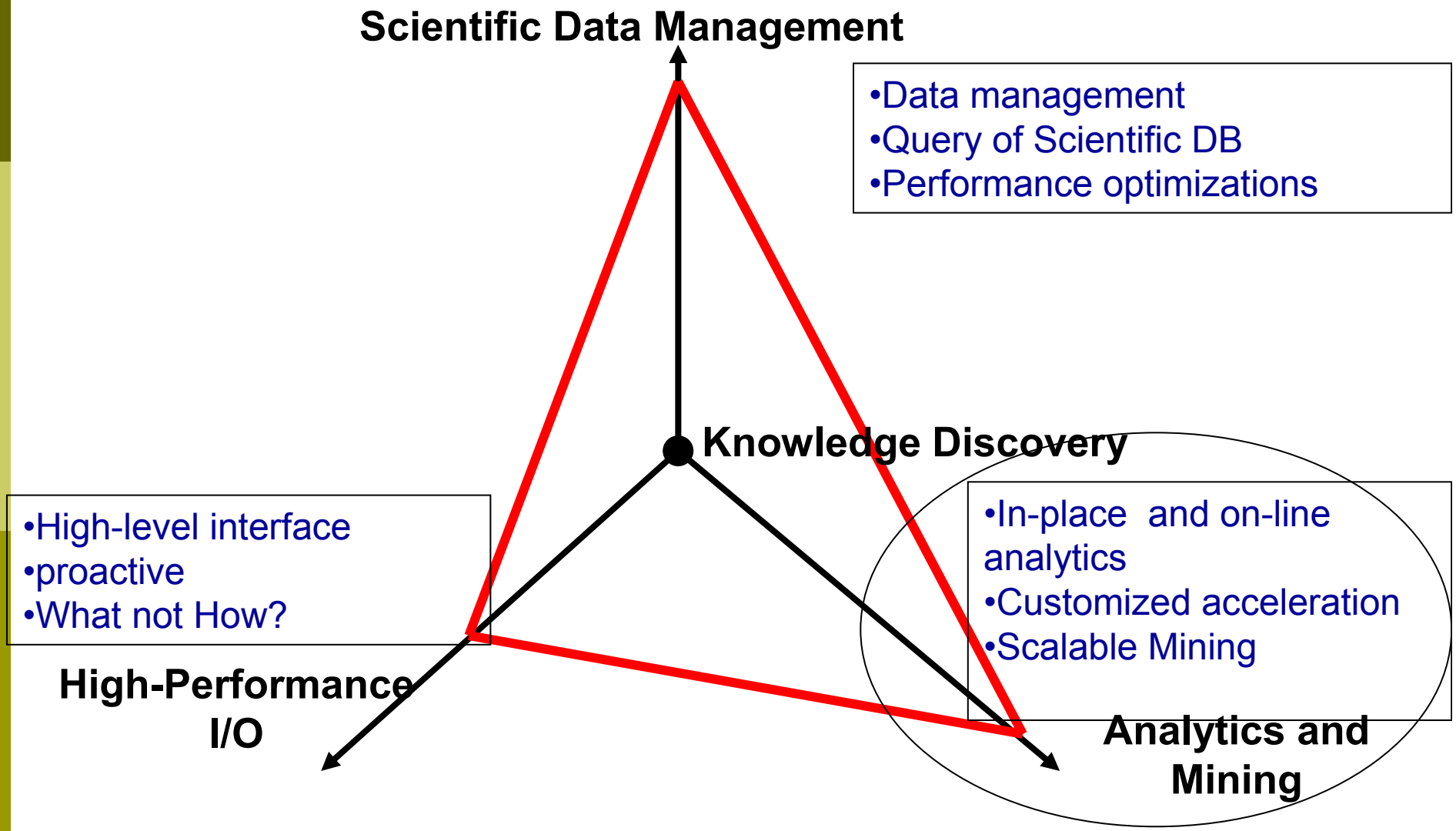
Alok Choudhary, Professor
Director: Center for Ultra-Scale Computing and Security
Dept. of Electrical Engineering and Computer Science
And Kellogg School of Management
Northwestern University
choudhar@ece.northwestern.edu

Acknowledgements:
DOE (SCIDAC)
NSF: (HECURA, CRI, Fellowships)
Students: Kenin Coloma, Avery Ching, Ramanathan, Berkin, Jianwei Li (Now at Wallstreet), Ying Liu (now faculty at Chinese Academy of Sciences), Joe Zambreno (now faculty at Iowa State), Wei-Keng Liao (Research prof at NWU), G. Memik (Asst prof at NWU)

Scientific Data Management and Analysis: Productivity and Performance



Challenges in Scientific Knowledge Discovery

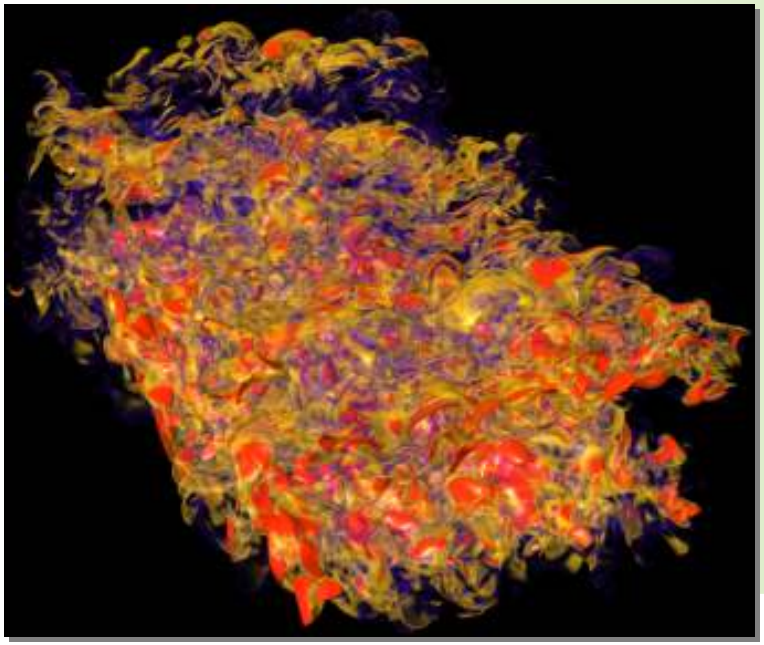


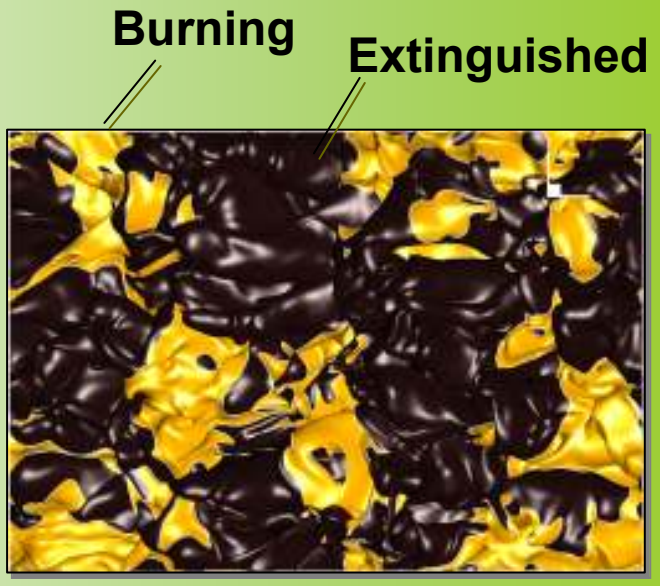
Combustion Application using DNS: Extinction and reignition in a CO/H2 jet flame

Understanding extinction/reignition in non-premixed combustion is key to flame stability and emission control in aircraft and power producing gas-turbines

Discovered dominant reignition mode is due to engulfment of product gases, not flame propagation

Scalar dissipation rate



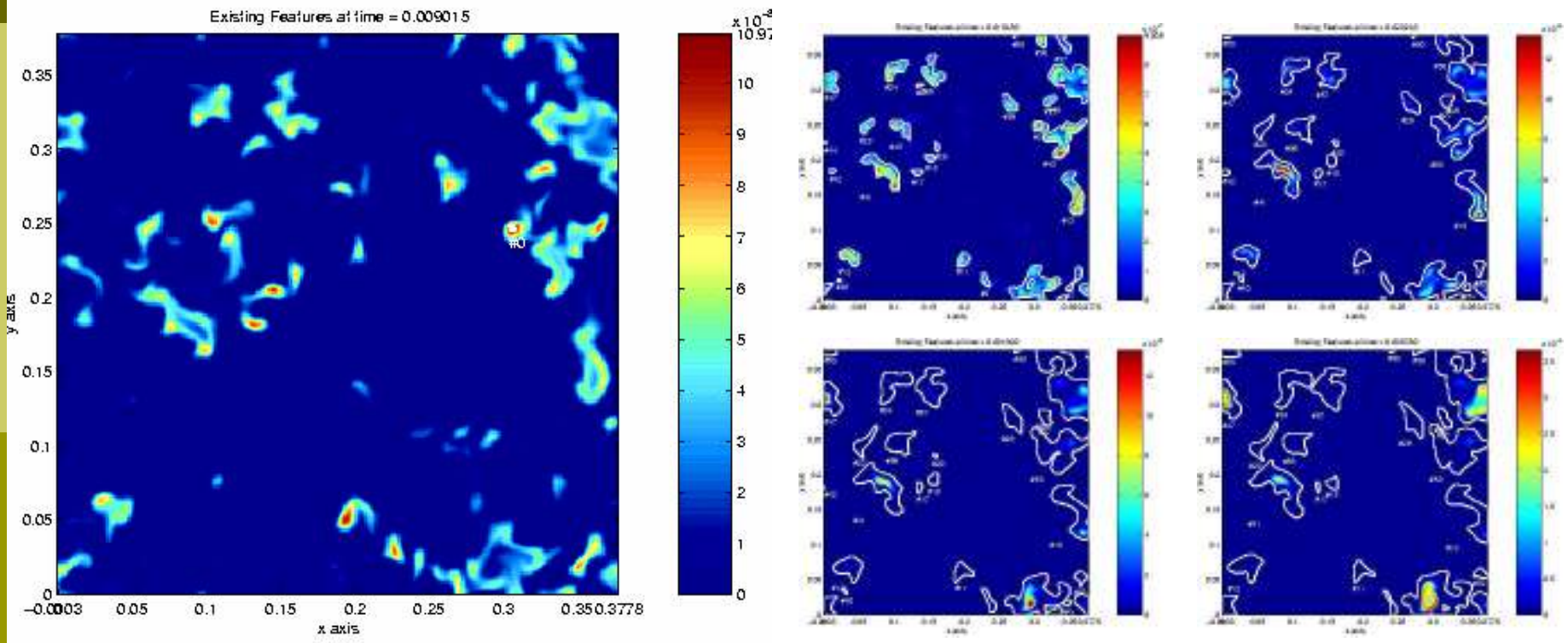


The ***largest ever simulations of combustion*** have been performed to advance this goal:

- 500 million grid points
- 11 species and 21 reactions
- 16 DOF per grid point
- 512 Cray X1E processors
- 30 TB raw data
- 2.5M hours on IBM SP NERSC (INCITE); 400K hours on Cray X1E (ORNL)

Combustion understanding and modeling: Detection and tracking of autoignition features on-line

Direct simulation of a 3D turbulent flame with detailed chemistry (200 million grids, 12 species, 5 TB raw data, 5 TB derived data

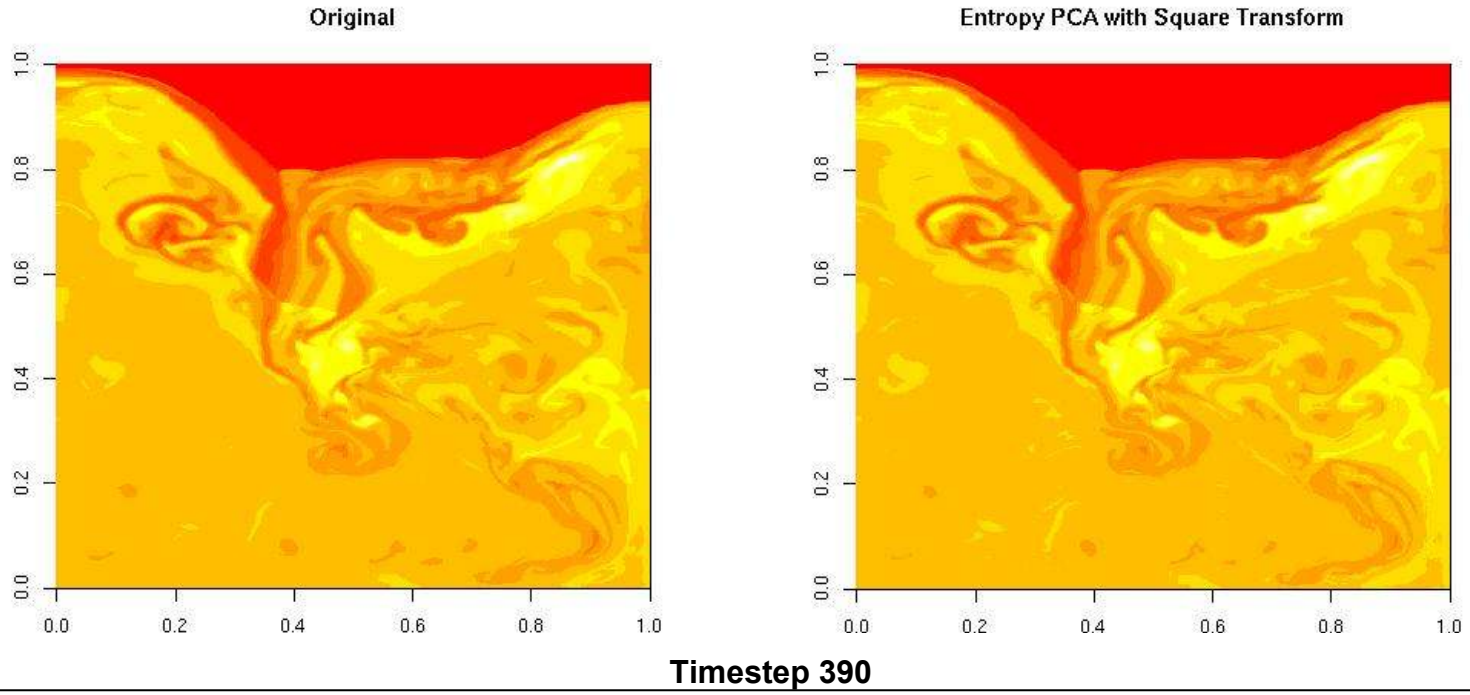


ACK: Jackeline Chen, SNL

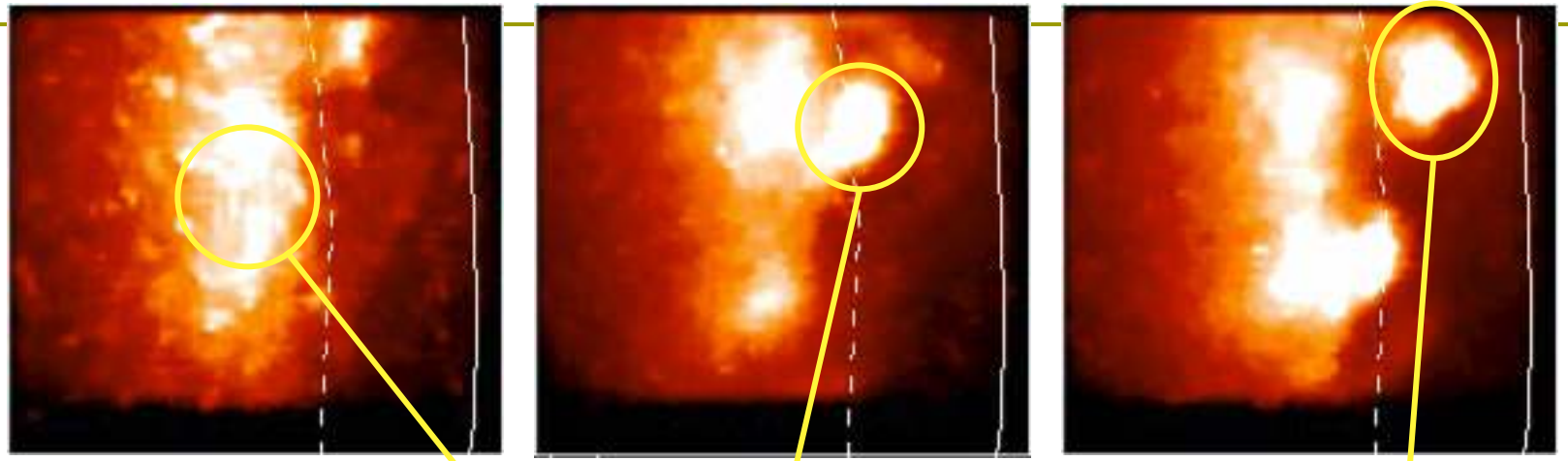
Example - Mining-based Data Reduction for Multigrid Simulation

- **Based on PCA of contiguous field blocks**
- **Astrophysics supernova simulation:**
  - **16 to 200 times reduction per time step**

Ack: Nagiza Samatova ORNL



Fusion: Using image processing/mining to analyze blob formation



Second, track blobs back to their source in the "sea of turbulence"

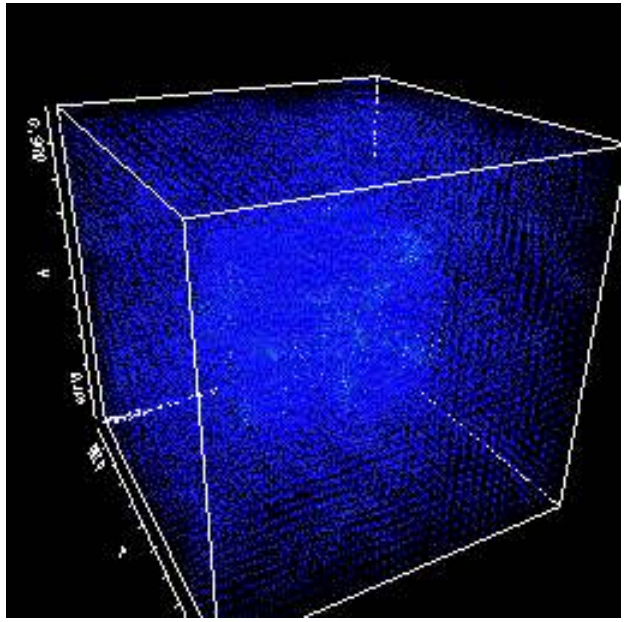
First, identify well-defined blobs using image analysis.

Fundamental question: Why does turbulence produce coherent structures such as blobs?

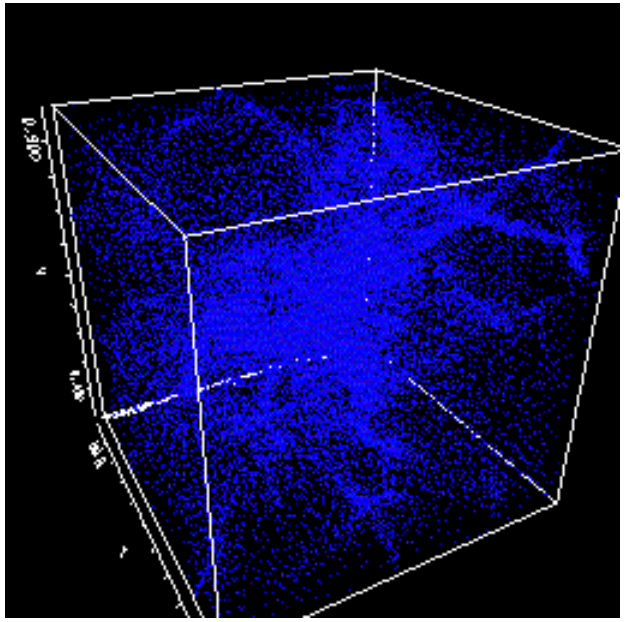
Ack: Scott Klasky

Cosmology

ENZO: simulates the formation of galaxies from the beginning of the universe to the present day



Data set 1

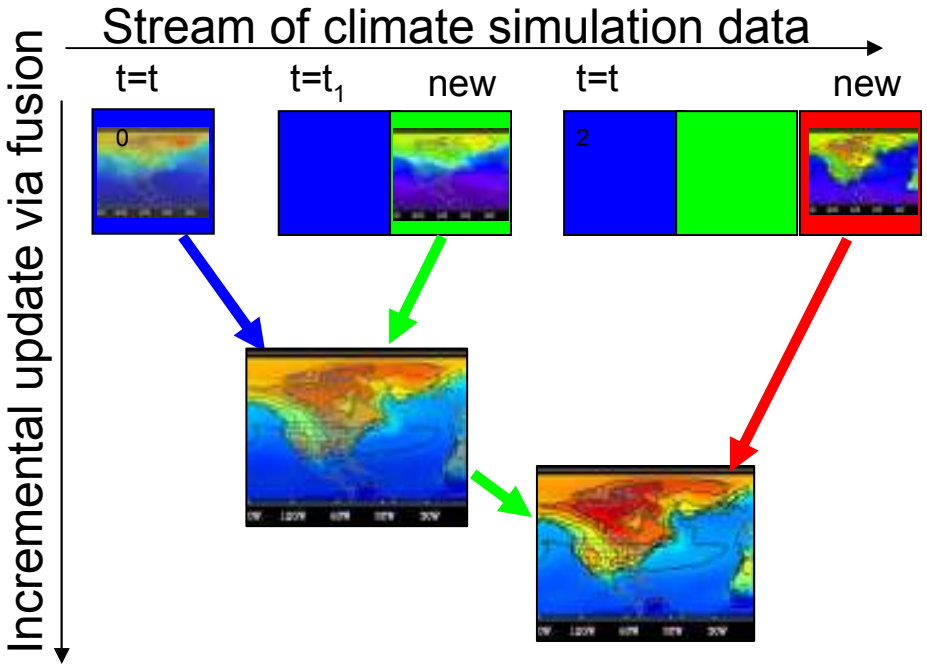


Data set 2

Each data set contains 491520 particles

Simulation Data Sets Dynamically Change

- Scientific simulations (e.g., climate modeling and supernova explosion) typically run for days to month and produce data sets in the order of one to ten terabytes per simulation.
- Effectively and efficiently analyzing these streams of data is a challenge:
  - Static analysis techniques are not sufficient. Any changes require complete re-computation.



Computations MUST be able to efficiently analyze streams of data while they are being produced, rather than wait until they are produced

Complexity of Scientific Simulation Produced Data Sets requires mining

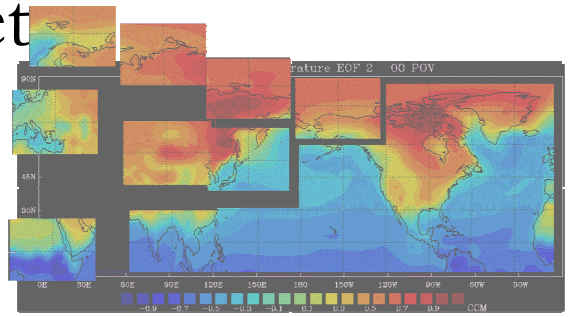
Challenge: Develop effective & efficient methods for mining scientific data set

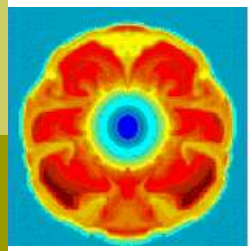
Tera&Petabytes

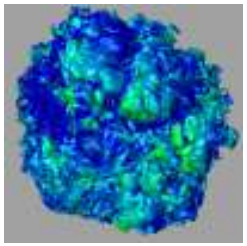
Existing methods do not scale in terms of time and storage

Distributed

Existing methods work on single centralized dataset. Data transfer is prohibitive







Supernova Explosion:
1-D simulation: 2GB
2-D simulation: 1TB
3-D simulation: 50TB

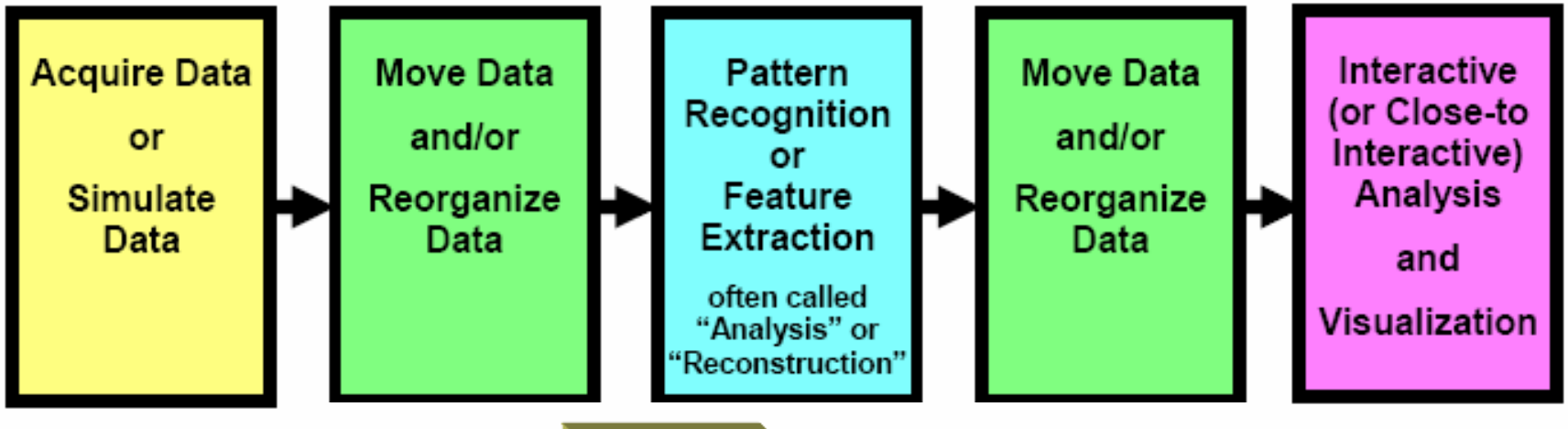
High-dimensional

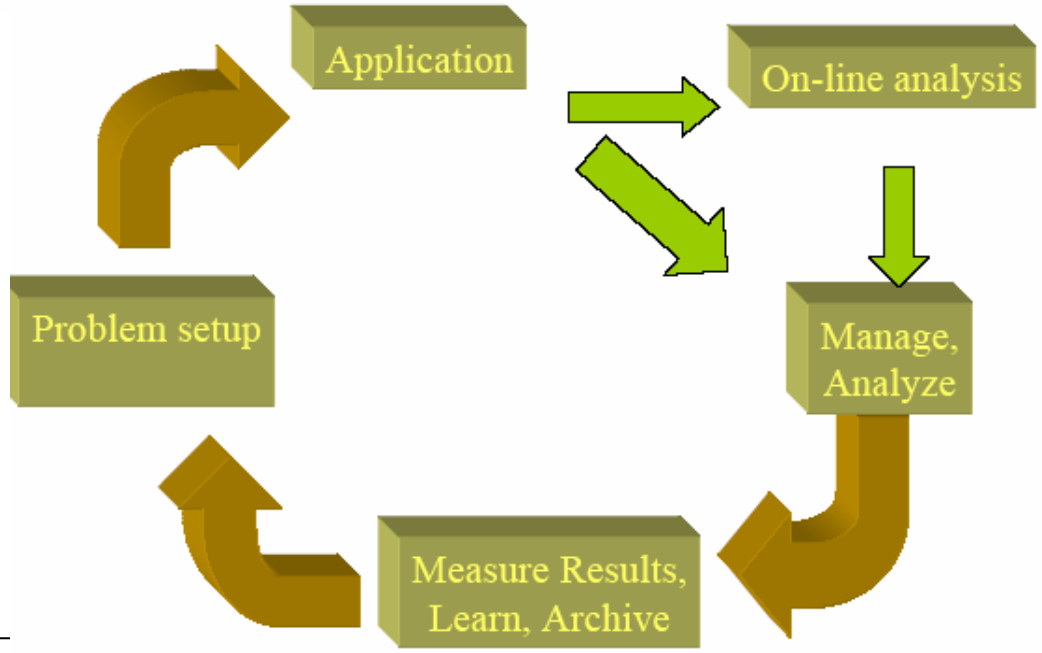
Existing methods do not scale up with the number of dimensions

Dynamic

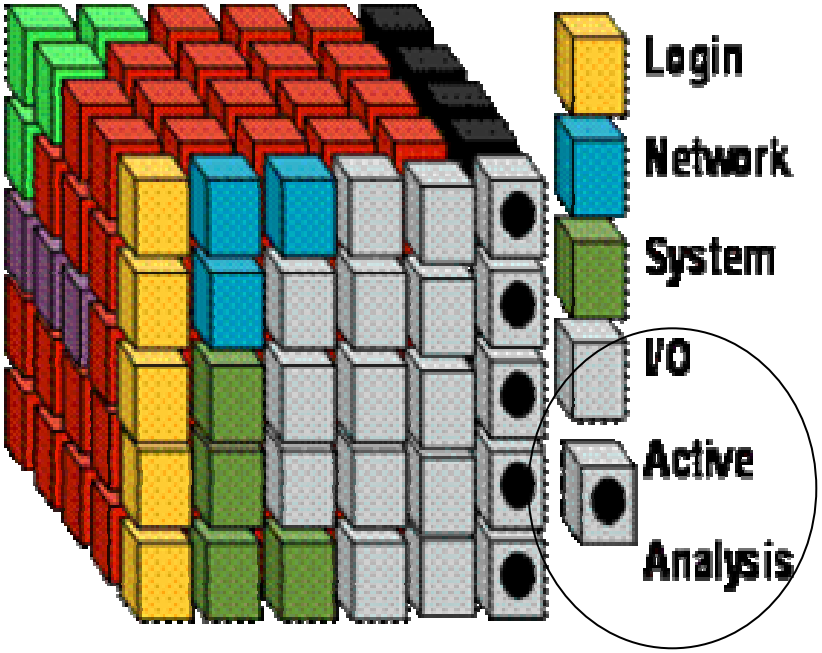
Existing methods work w/ static data. Changes lead to complete re-computation

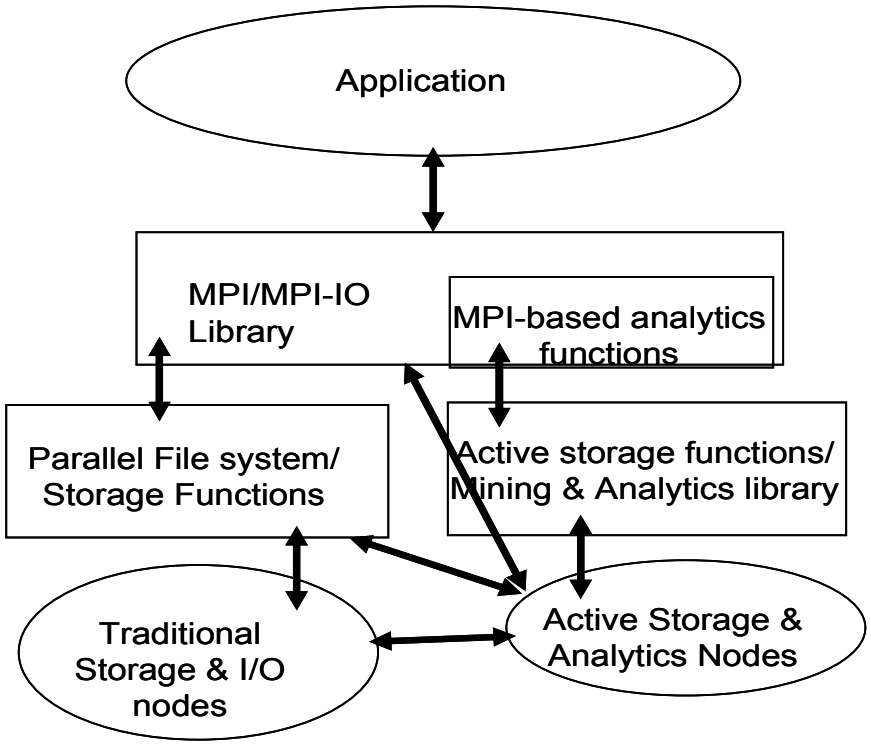
Scientific Work-Flow



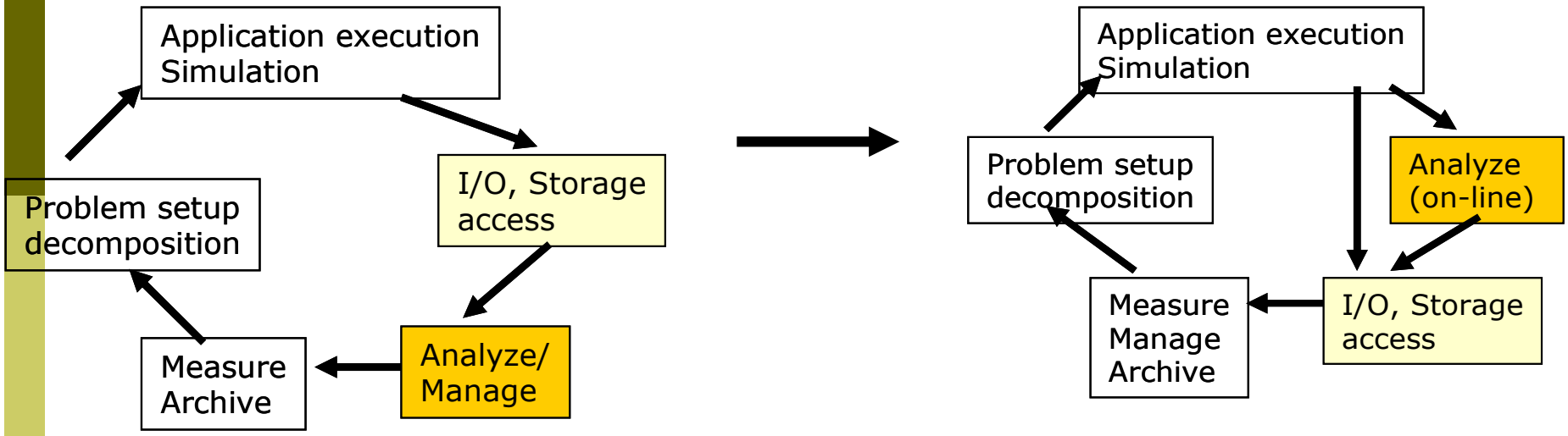


In-Place On-Line Scalable Mining

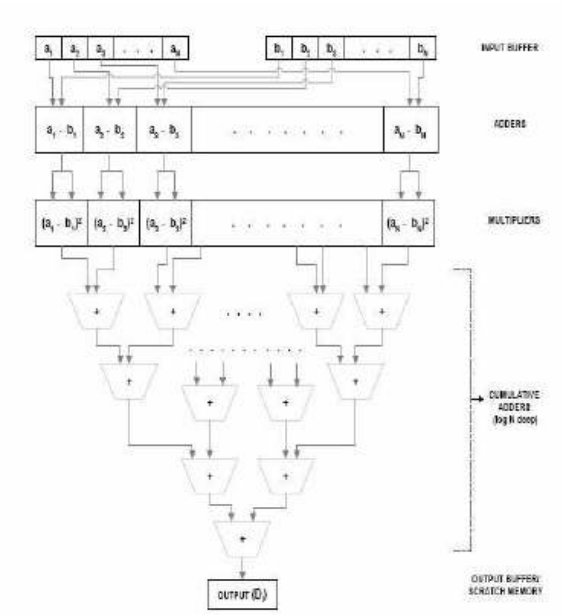


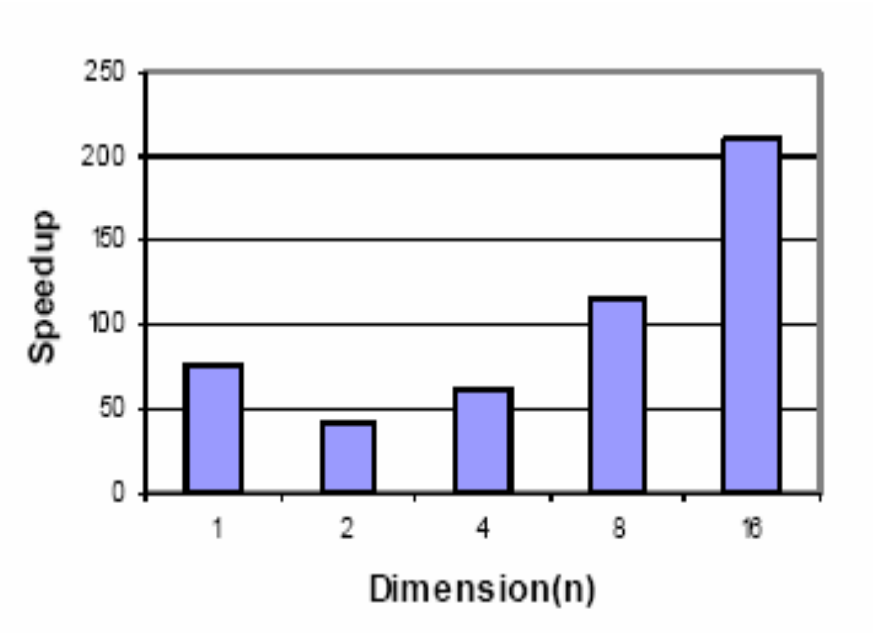


Accelerating and Computing in the Storage

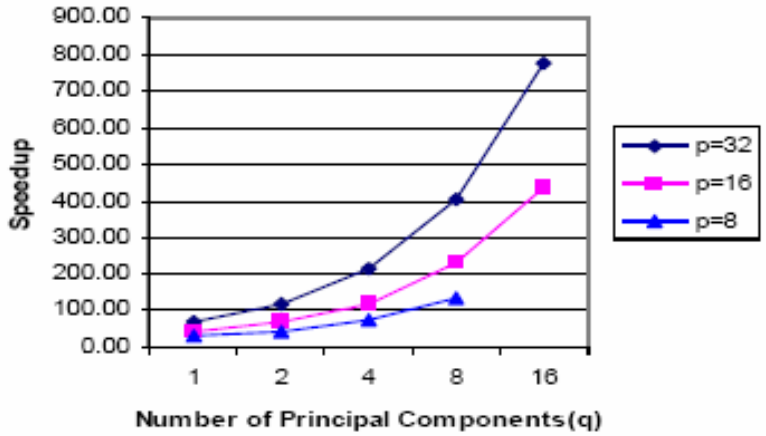


Active Storage System





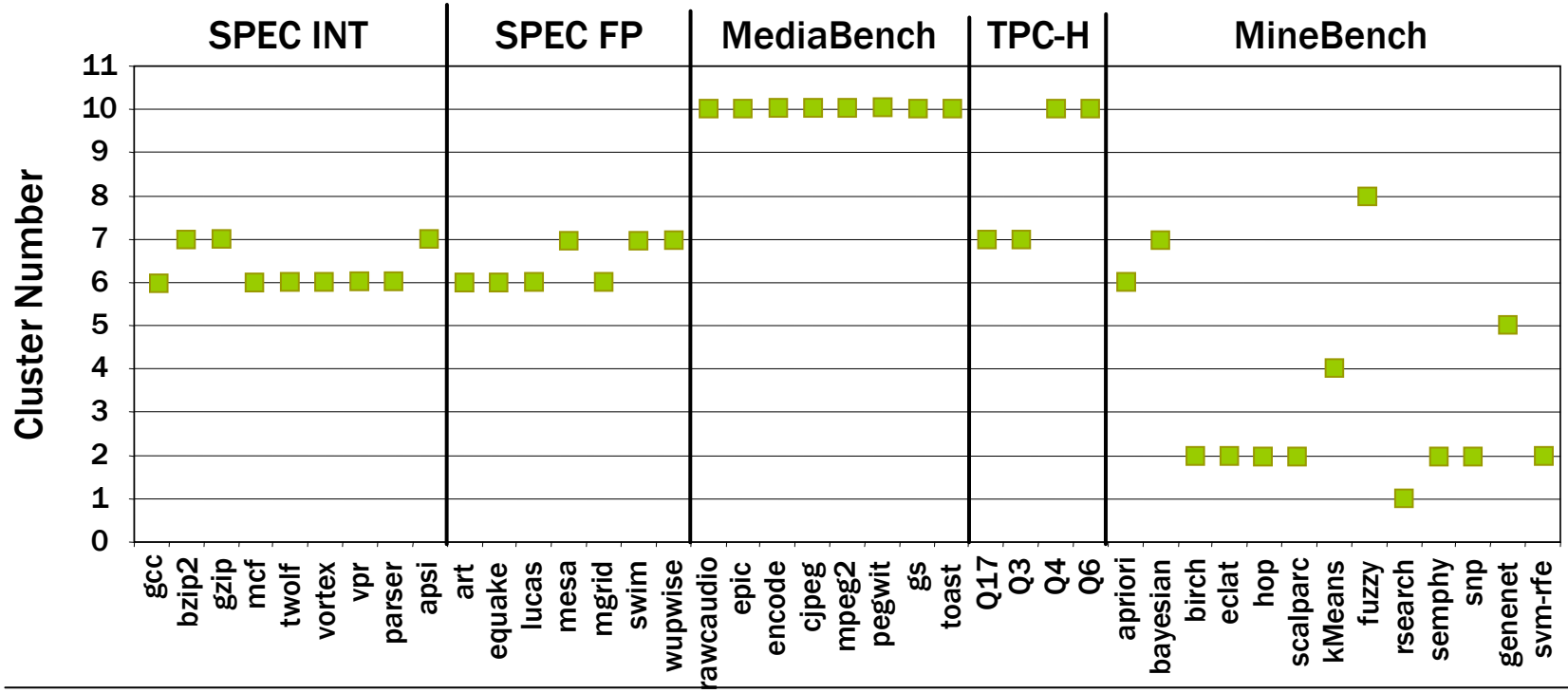
Distance kernel in Clustering data mining: Speedup over a 2.4GHz AMD Opteron



PCA

Data Mining – Is it different from other application domains?

- 25 dimensional performance and characterization data. Mining used to cluster
- NU MINEBENCH
- http://cucis.ece.northwestern.edu/projects/DMS/MineBench.html



Community Resource: MineBench Project Homepage

http://cucis.ece.northwestern.edu/projects/DMS