

Petascale I/O and Adaptive Performance Control

Dan Reed

reed@cs.uiuc.edu

University of Illinois

Thanks To ...

- Ruth Aydt (adaptive control)
- Mario Medina (catastrophe theory)
- James Oly (hidden Markov models)
- Huseyin Simitci (adaptive striping)
- Nancy Tran (time series forecasting)
- Dan Wells (measurement)
- Ying Zhang (measurement)

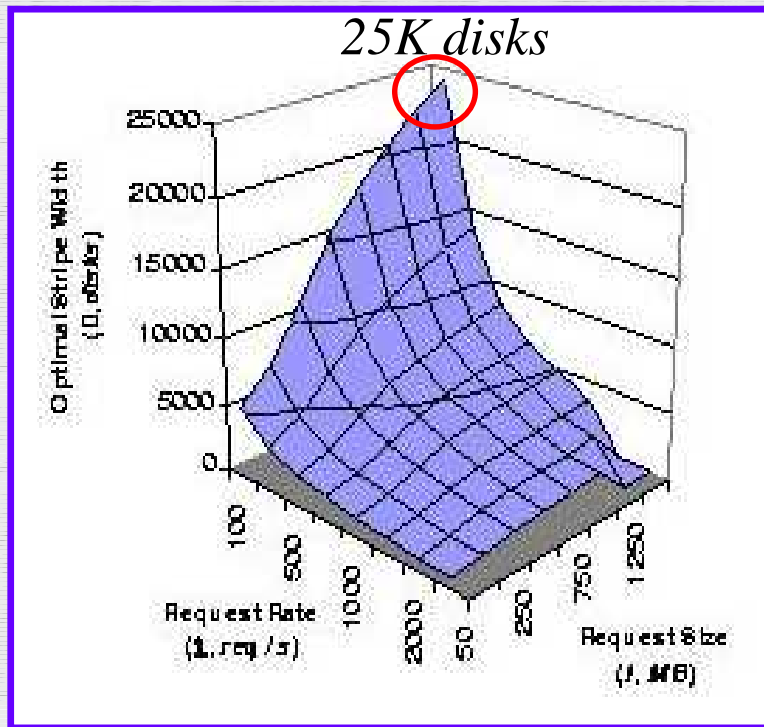
Presentation Outline

- Petascale I/O implications
 - technology trends and parallelism
- ASCI code characterizations
 - complexity motivation
- Adaptive I/O control
 - classification and prediction
 - distributed control and adaptation

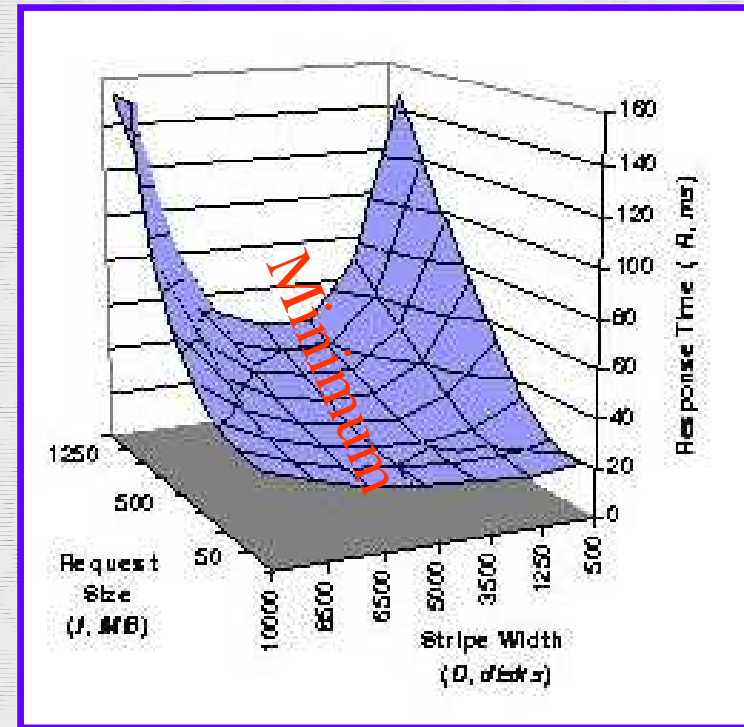
Storage Device Trends

- Commodity disks (apologies to Hans for errors)
 - 100 percent annual increase in areal density
 - *8 percent annual decrease in access times*
 - 40 percent annual increase in media transfer rates
- Implications
 - **bandwidth/capacity ratio approach zero!**
 - massive parallelism for petascale systems

Petascale Striping (100K Disks)



Optimal Striping Factor



Request Size Variation

Based on a model of parallel I/O performance (*stay tuned*)

Petascale I/O Systems

- Massive parallelism
 - $O(10K-100K)$ disks
 - distributed data coordination and placement
 - flexible data striping
- Design implications
 - accept complexity
 - trade capacity for bandwidth
 - adaptively distribute data

Scale and Adaptivity

- Large-scale system lessons
 - unexpected complexity and interactions
- Examples
 - command and control systems
 - international financial system
 - electric power grid
- *Embrace adaptation at large scale*
- *Think distributed visualization and analysis*

Presentation Outline

- Petascale I/O implications
 - technology trends and parallelism
- *ASCI code characterizations*
 - *complexity motivation*
- Adaptive I/O control
 - classification and prediction
 - distributed control and adaptation

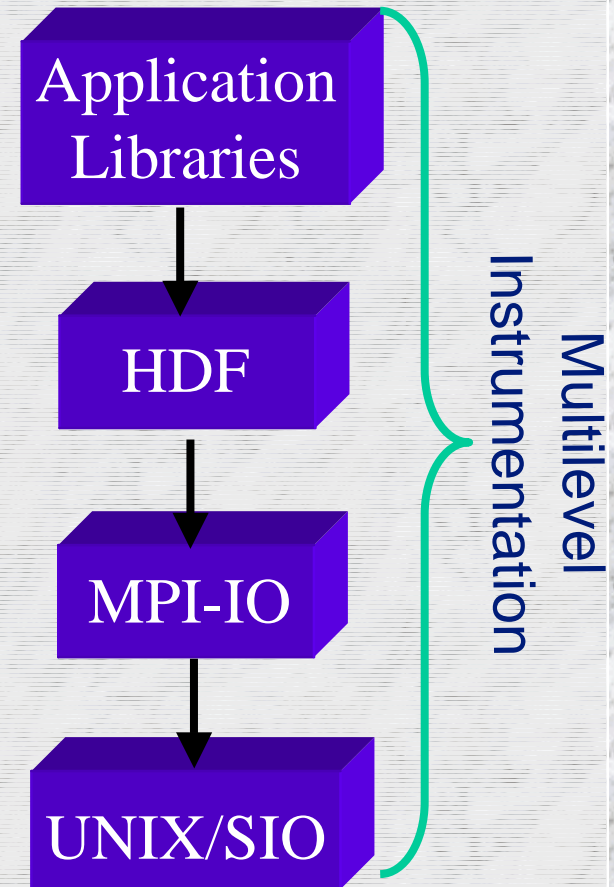
I/O Characterization Approach

■ Multilevel analysis

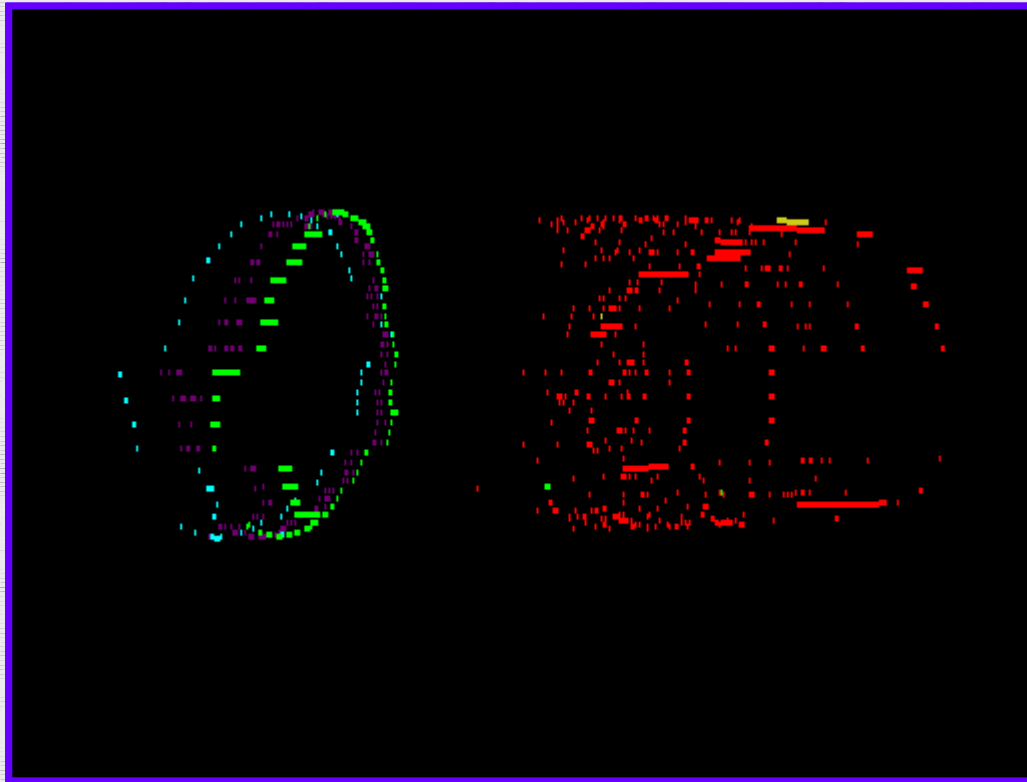
- correlation across levels
- overhead/policy assessment
- longitudinal studies
- Pablo I/O measurements

■ Rationale

- performance sensitivity
- library interface data sharing
- optimization guide



LLNL ARES I/O Activity



- ALE hydro code
- parallel domain decomposition
- explicit MPI
- 64 processors
- ASCI Blue Pacific

Complexity is the norm! **Caveat: old Silo I/O library**

Intelligent I/O Libraries

■ I/O characterization experience

- policies must match application stimuli
- policy configuration dependent on many factors
 - hardware capabilities
 - access pattern attributes (*consider DVCs*)
 - resource contention (other users)

■ Implication

- static optimization alone is not sufficient
- adaptive optimization and tuning

Presentation Outline

- Petascale I/O implications
 - technology trends and parallelism
- ASCI code characterizations
 - complexity motivation
- *Adaptive I/O control and optimization*
 - *classification and prediction*
 - *distributed control and adaptation*

Closed Loop Adaptive Control

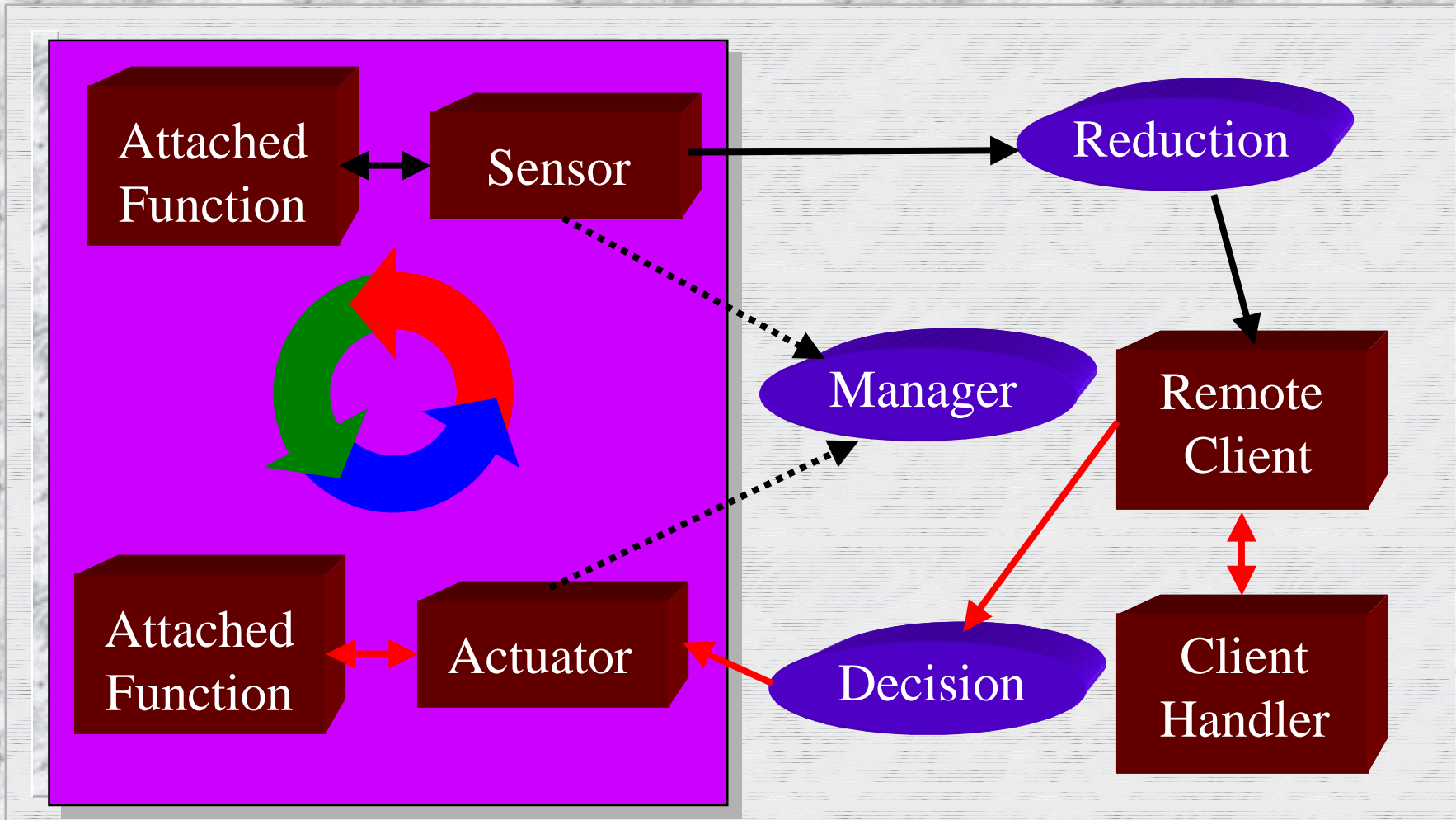
■ Requirements

- monitor resource demands and responses
- select policies based on observed behavior
- implement policy changes locally and globally

■ Observations

- dynamic attachment to remote software
- *control theory meets performance analysis*

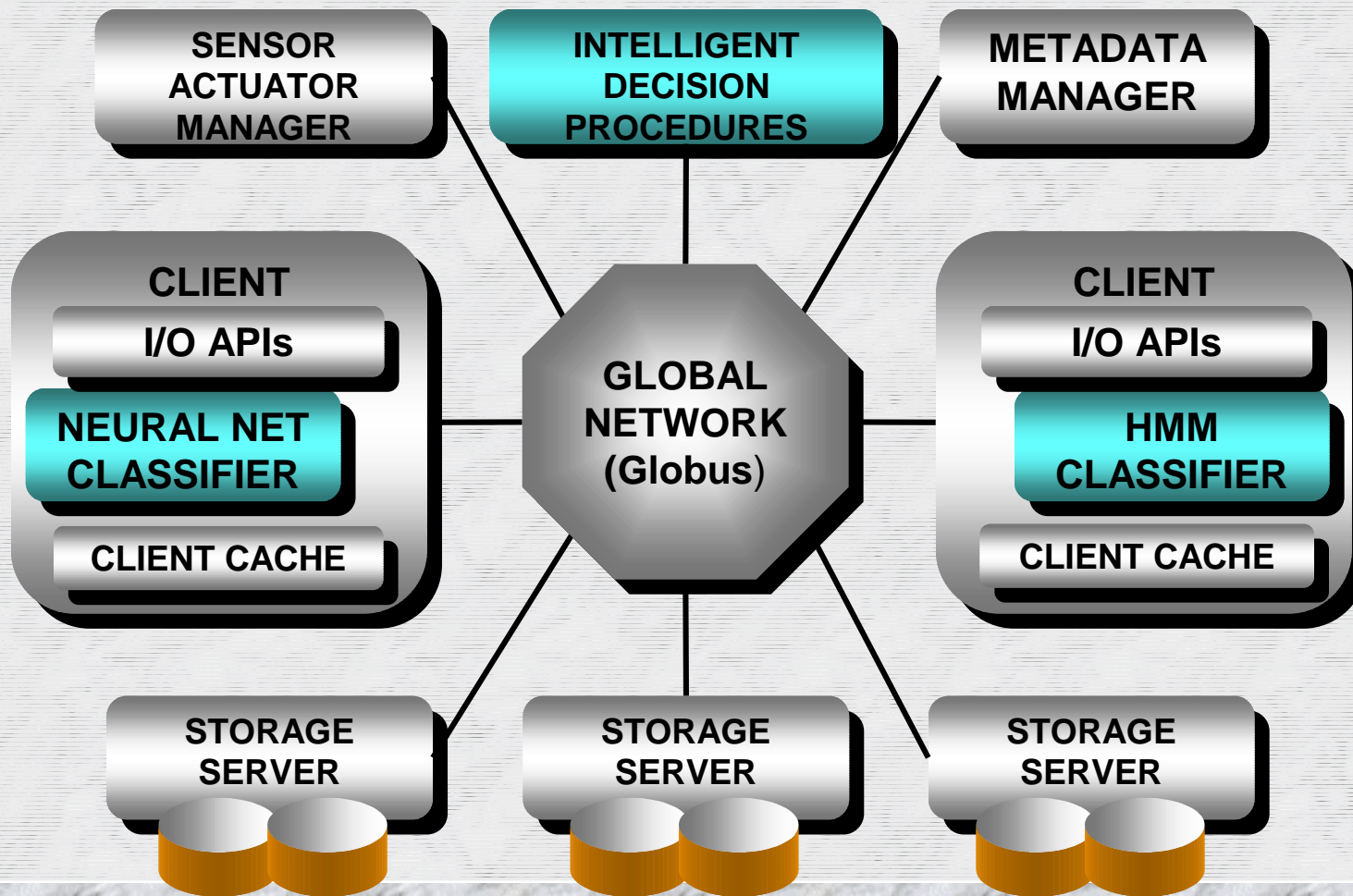
Autopilot Adaptive Control



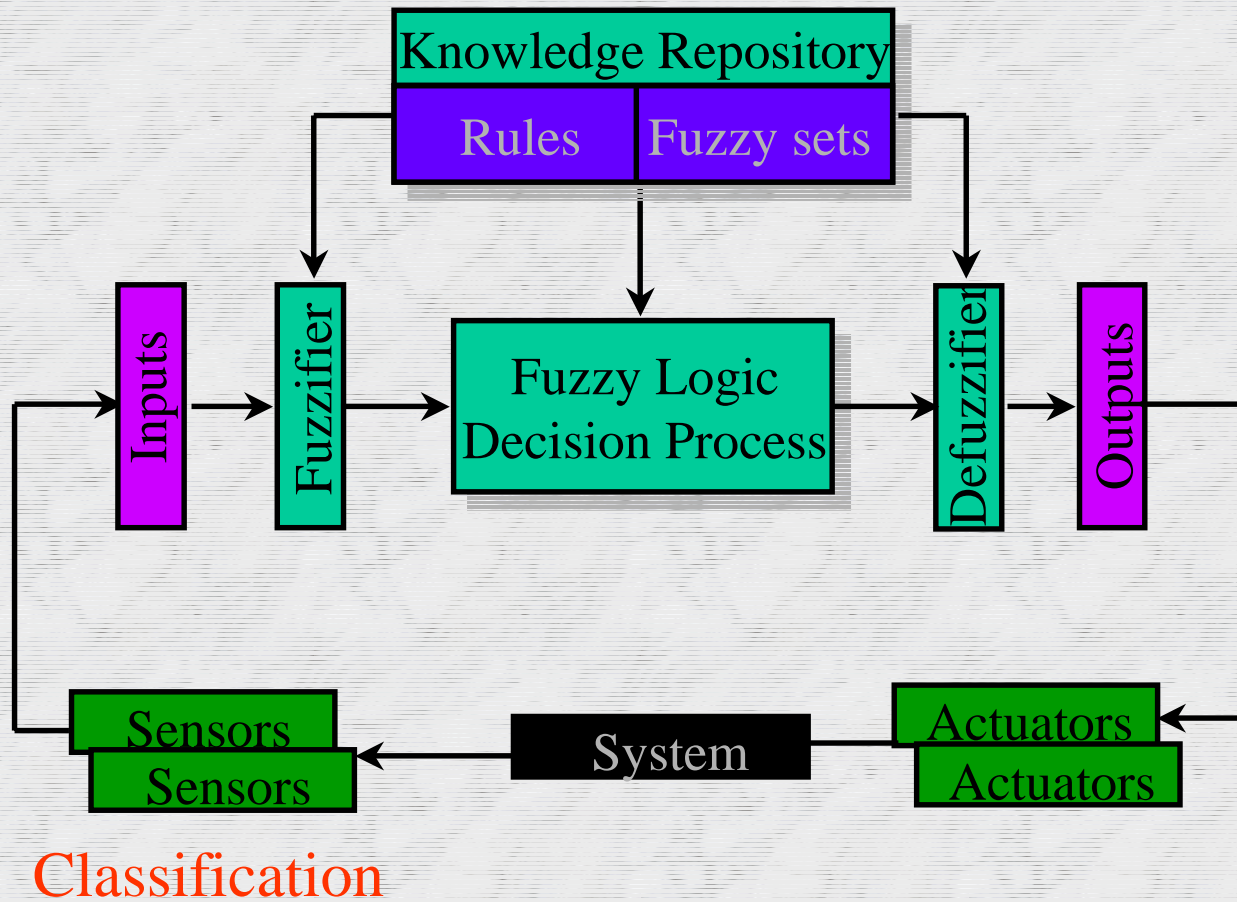
PPFS II Adaptive I/O Library

- Automatic access pattern classification
 - artificial neural networks
 - hidden Markov models
 - ARIMA time series forecasting
- Dynamic policy selection
 - adjustable prefetching, caching, and striping
 - redundant storage
 - fuzzy logic control
- Principles applicable to all resource policies

PPFS II Architecture

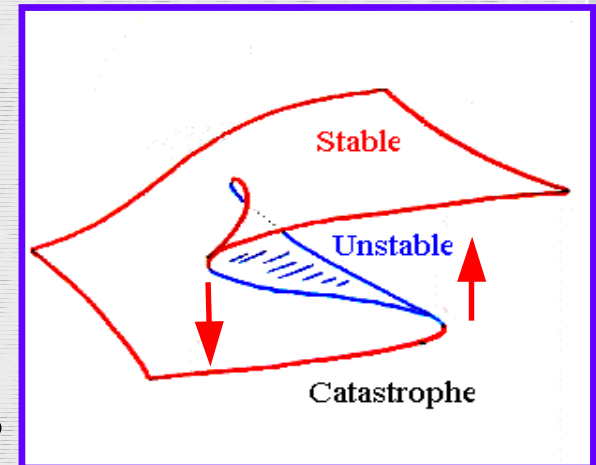


Fuzzy Logic Decision Process



Catastrophe Theory

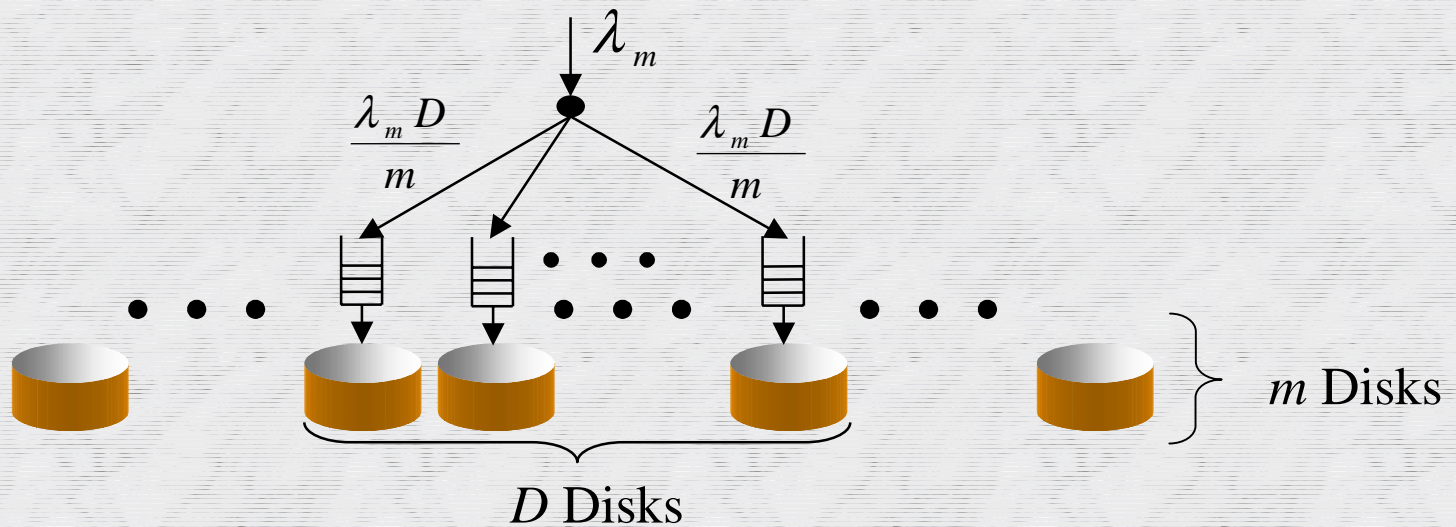
- A model for reasoning about complexity
 - stable/unstable regions
 - transitions and hysteresis
 - optimal/near-optimal control
- Operate at the “ragged edge”
 - maximize performance
 - exploit dynamic adaptation



Open I/O Research Problems

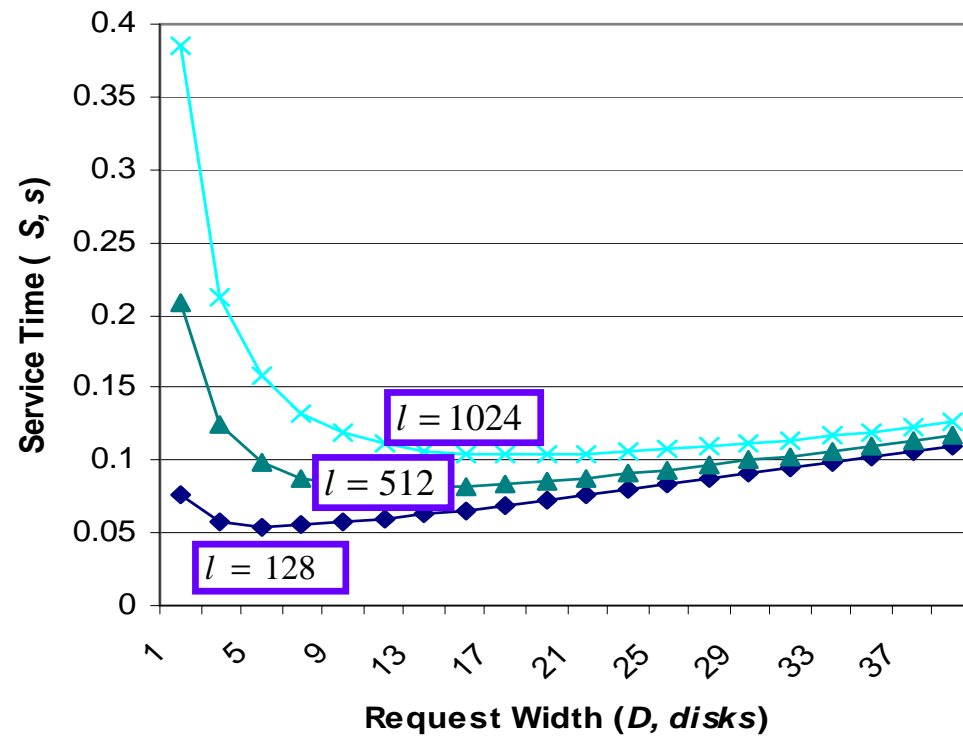
- Analytical models of disk striping
 - to study the performance of very large systems
- Adaptive selection of disk striping parameters
 - smoothly adapting policies
 - fuzzy logic with configurable rule bases
- Space-time tradeoffs
 - redundant storage formats (multiple file copies)
 - off-line and/or online transformations
 - request scheduling for multiple copies

Striping Queueing Model



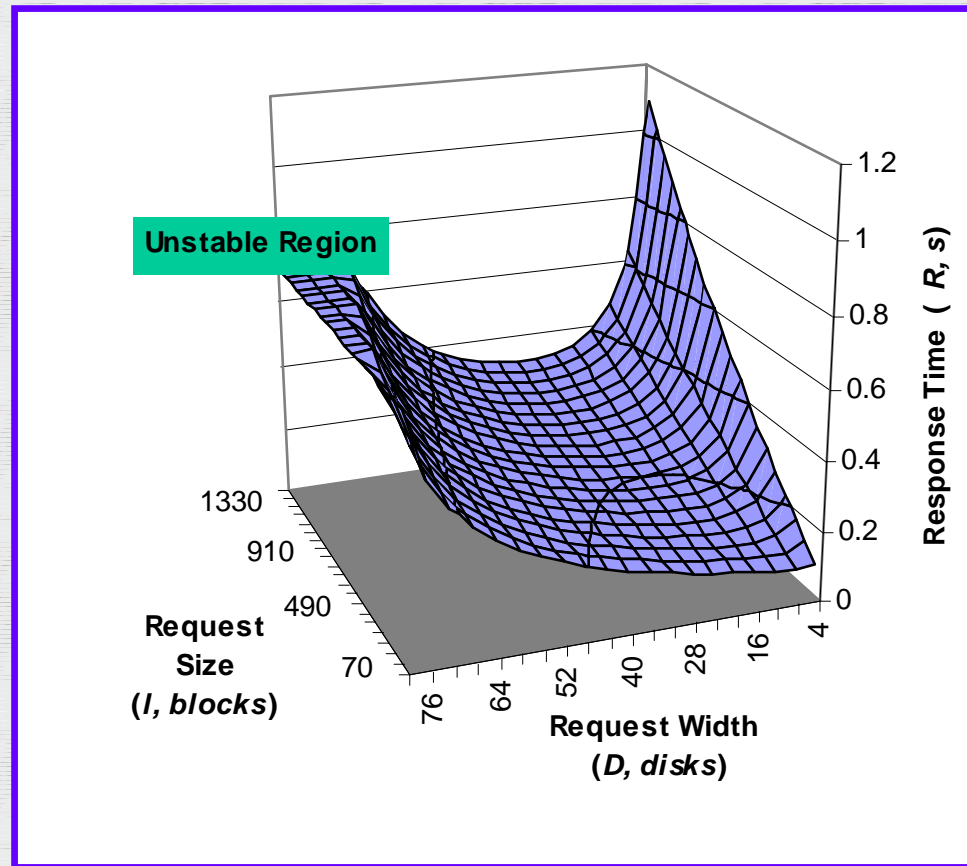
Parallelism decreases service time, but only to a point!

Optimal Striping Example



Optimal number of disks for each request size

Striped Request Response Time



Striping Rule Base

```
if ( RequestRate == INFREQUENT &&  
    RequestSize == TINY )  
    {RequestWidth = SMALL;}
```

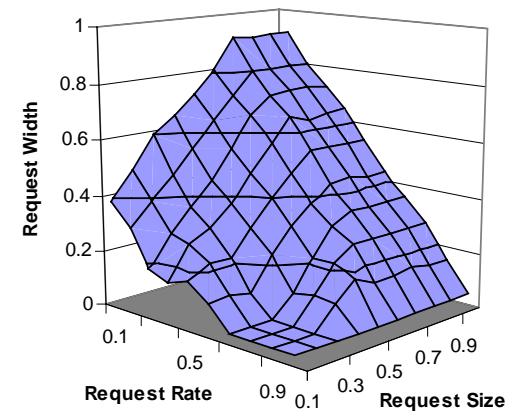
... ..

```
if ( RequestRate == CONTINUOUS &&  
    RequestSize == LARGE )  
    {RequestWidth = TINY;}
```

... ..

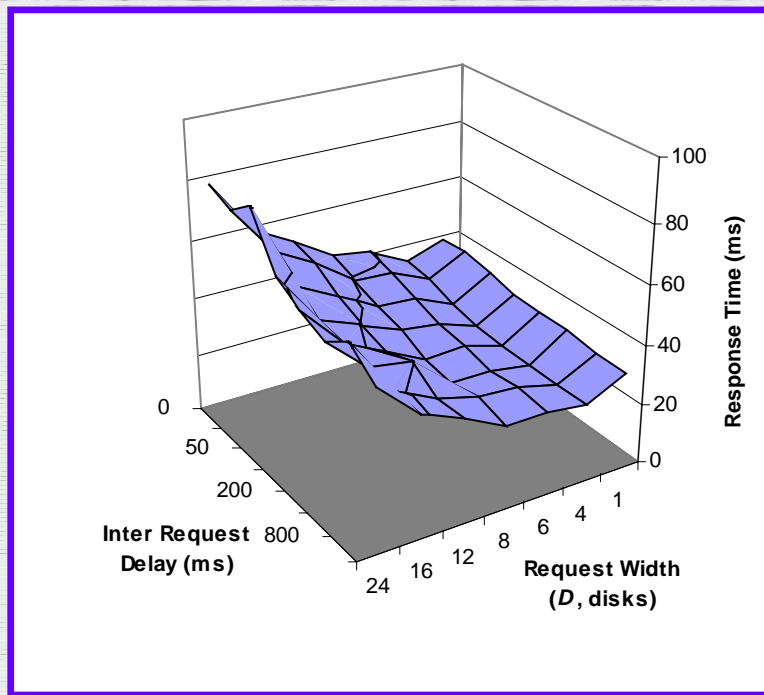
```
if (RequestRate == FREQUENT)  
    {FileReplicationTime = OFFLINE;}
```

... ..

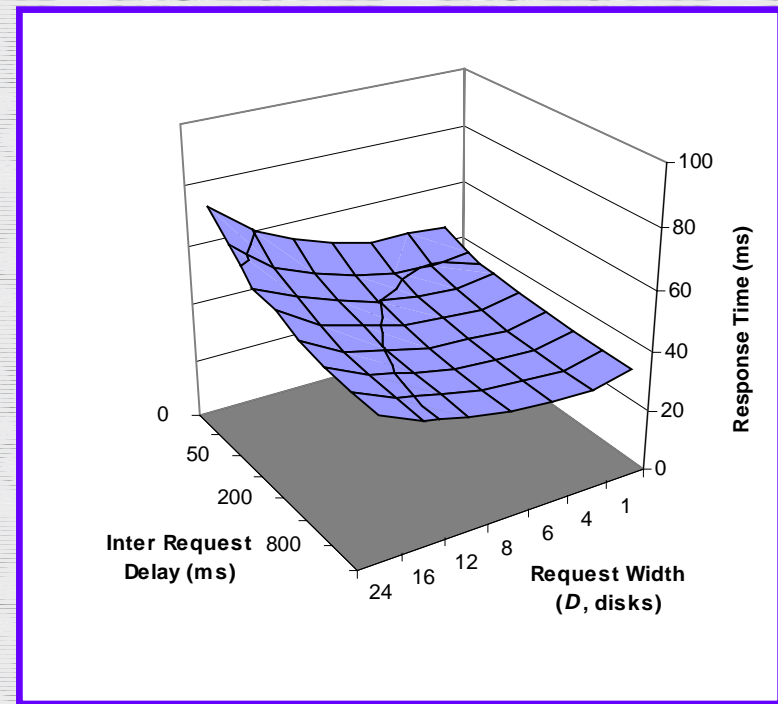


Model Verification (Linux Cluster)

Cluster Response Time



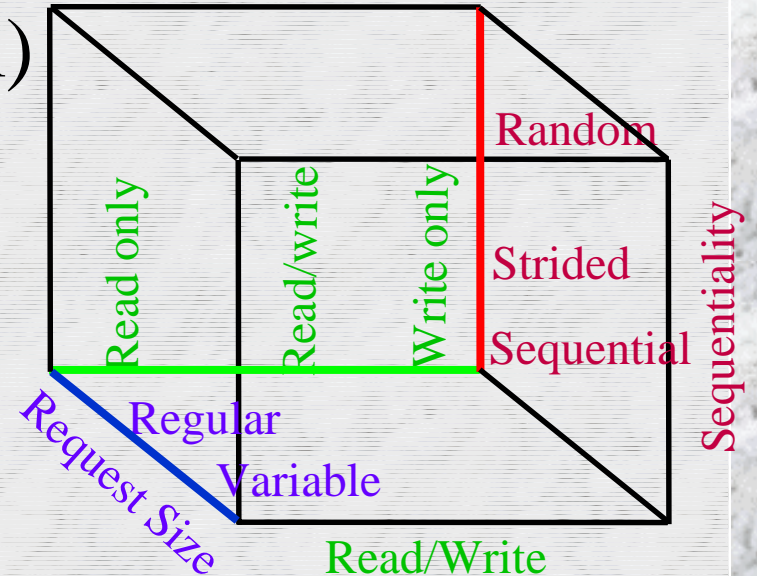
Queueing Model



64 KB random reads, 11% difference on average

Access Pattern Exploitation

- Local (per parallel thread)
 - artificial neural networks (ANNs)
 - hidden Markov models (HMMs)
- Global (parallel program)
 - temporal algebra
- Universal
 - across job mix



Neural Net Classification

■ I/O access abstraction

- file byte offset
- request size
- operation type (read/write)

■ Neural network classification

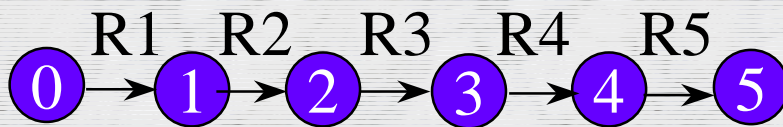
- operates in real-time
- access abstraction (input)
- qualitative classification (output)

Hidden Markov Classification

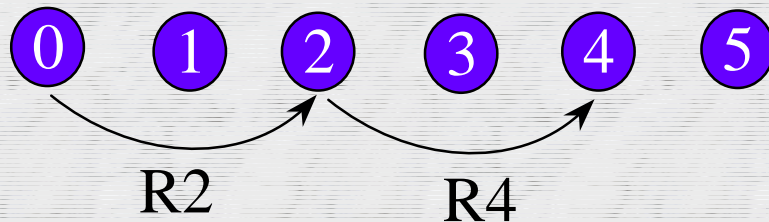
■ Modeling I/O patterns

- states correspond to file blocks or records
- transitions are reads or writes

■ Examples



Read only sequential



Read only strided

ARIMA Forecasting

- Three basic steps
 - model identification
 - parameter estimation
 - forecasting (prediction)
- ACT (autocorrelation function)
 - linear relation between observation pairs
- PACF (partial autocorrelation function)
 - conditional correlation
 - intervening observations removed

Concluding Thoughts

- Think adaptive
 - complex patterns
 - intelligent management
 - distributed control

- Think massively parallel
 - thousands of disks
 - adaptive control