PHOENIX
A real-time fault-tolerant network-attached storage device

Ashish Raniwala, Srikant Sharma
Anindya Neogi, Tzi-cker Chiueh

Experimental Comp Systems Lab
Department of CS
SUNY@Stony Brook

Introduction

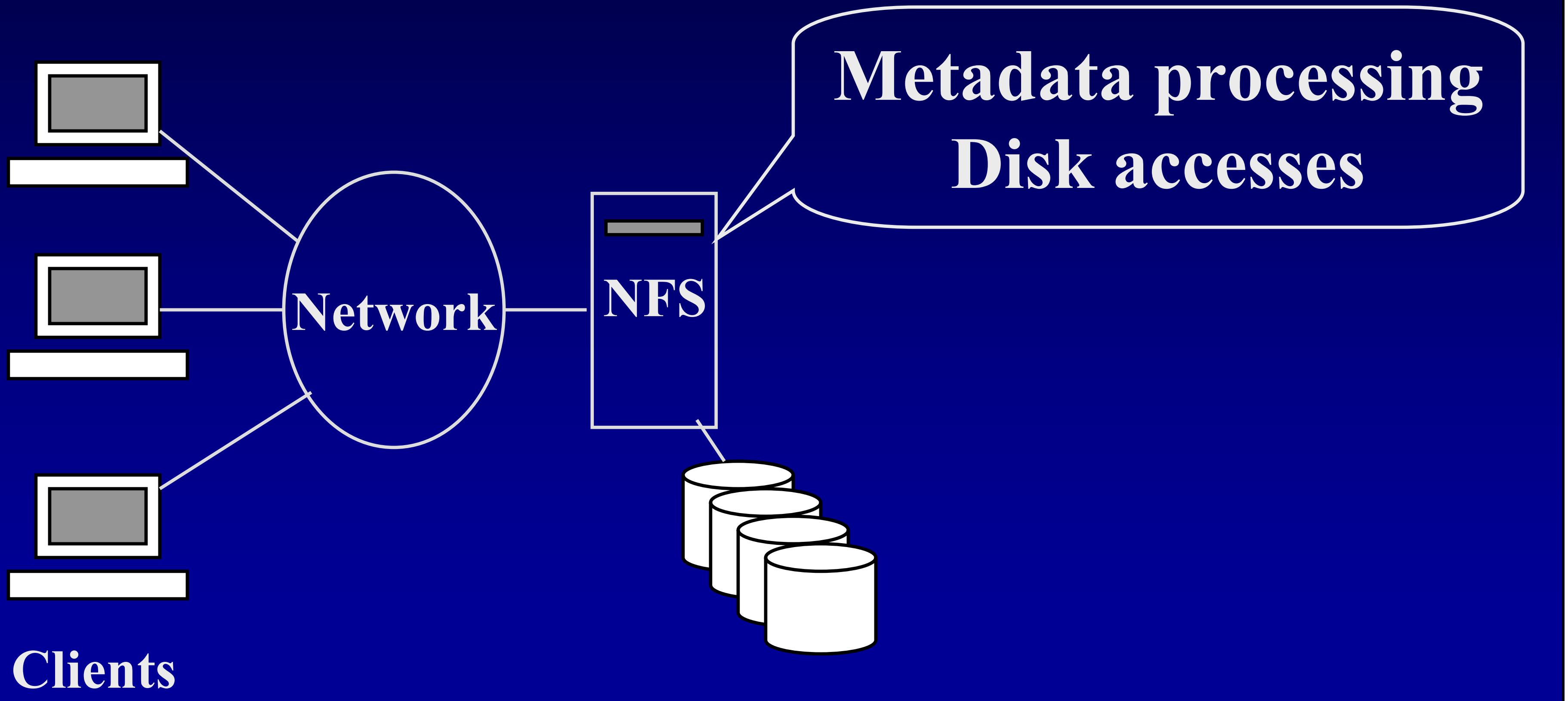
NFS

- File-system sitting on the network
- Provides a file-system interface

NASD

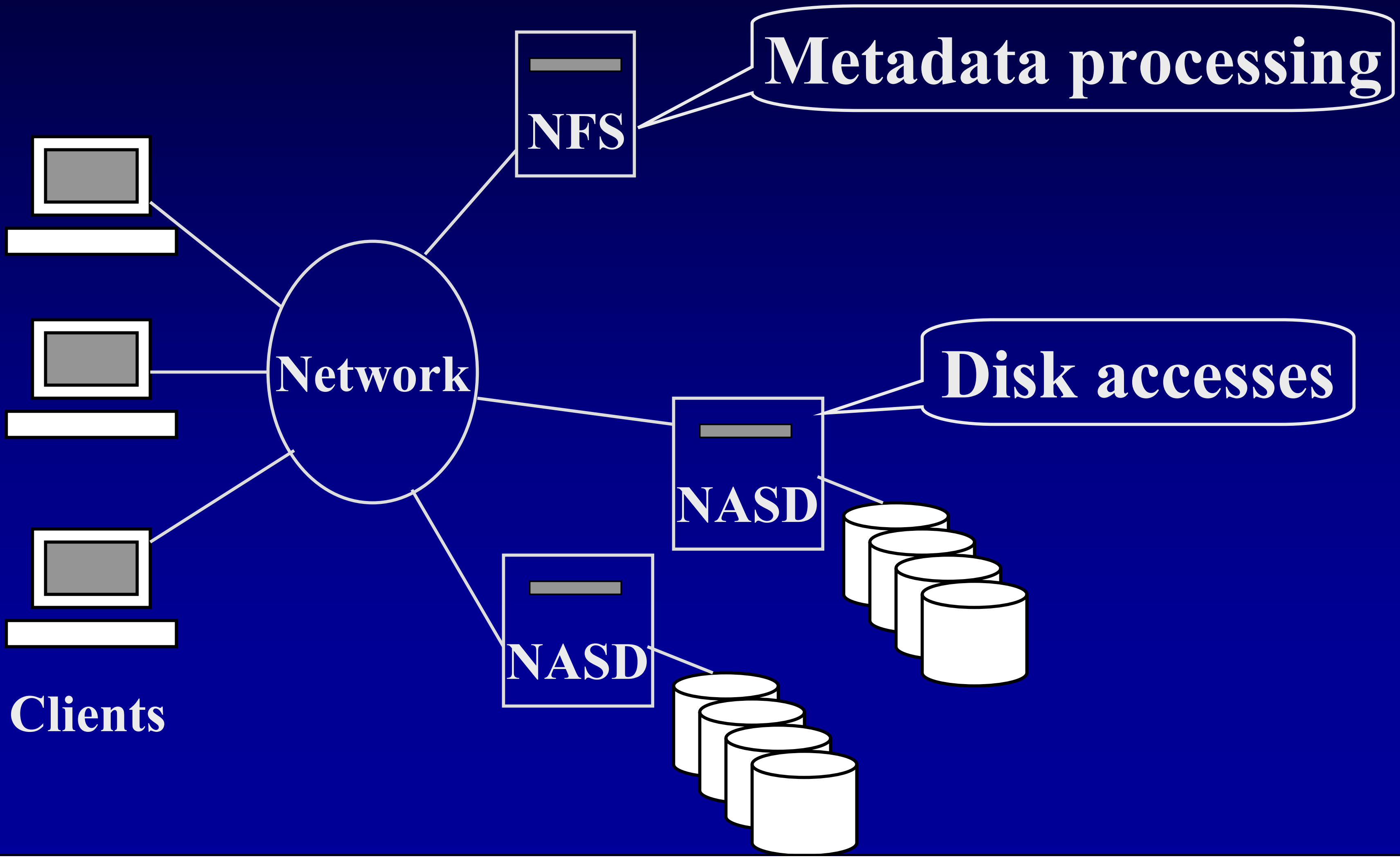
- Storage-device sitting on the network
- Provides a disk-like interface

Motivation



Metadata processing
Disk accesses

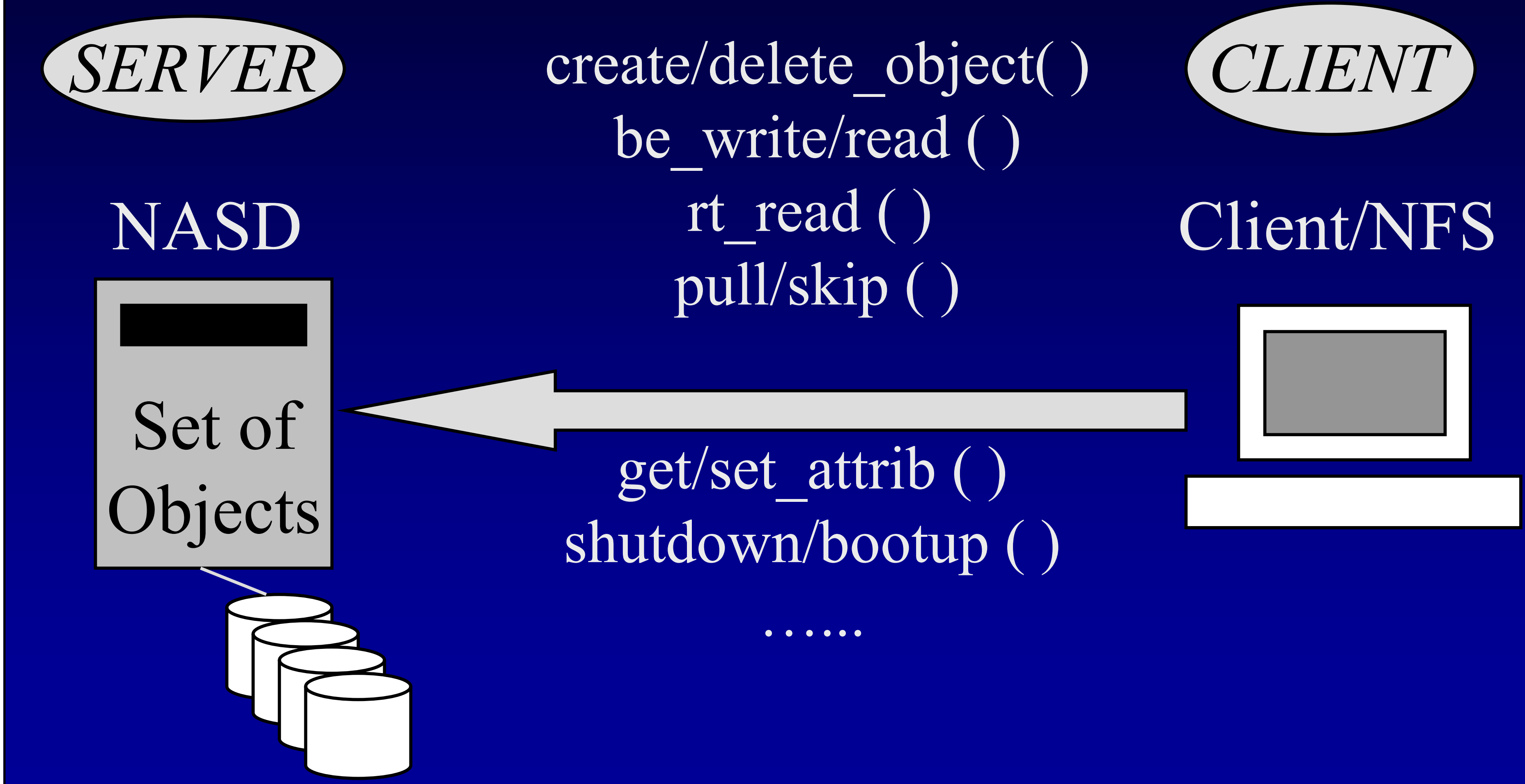
Motivation (Scalable NFS!)



Functionalities

- Object-level interface (on top of disk i/f)
- QoS-guaranteed disk accesses
- QoS valid across single disk failures
  - Service availability
  - Data availability
- Dynamic utilization of unused space
- Low power for increased reliability

Interface



SERVER

NASD

Set of Objects

create/delete_object( )

be_write/read ( )

rt_read ( )

pull/skip ( )

get/set_attrib ( )

shutdown/bootup ( )

……

CLIENT

Client/NFS

Road Map

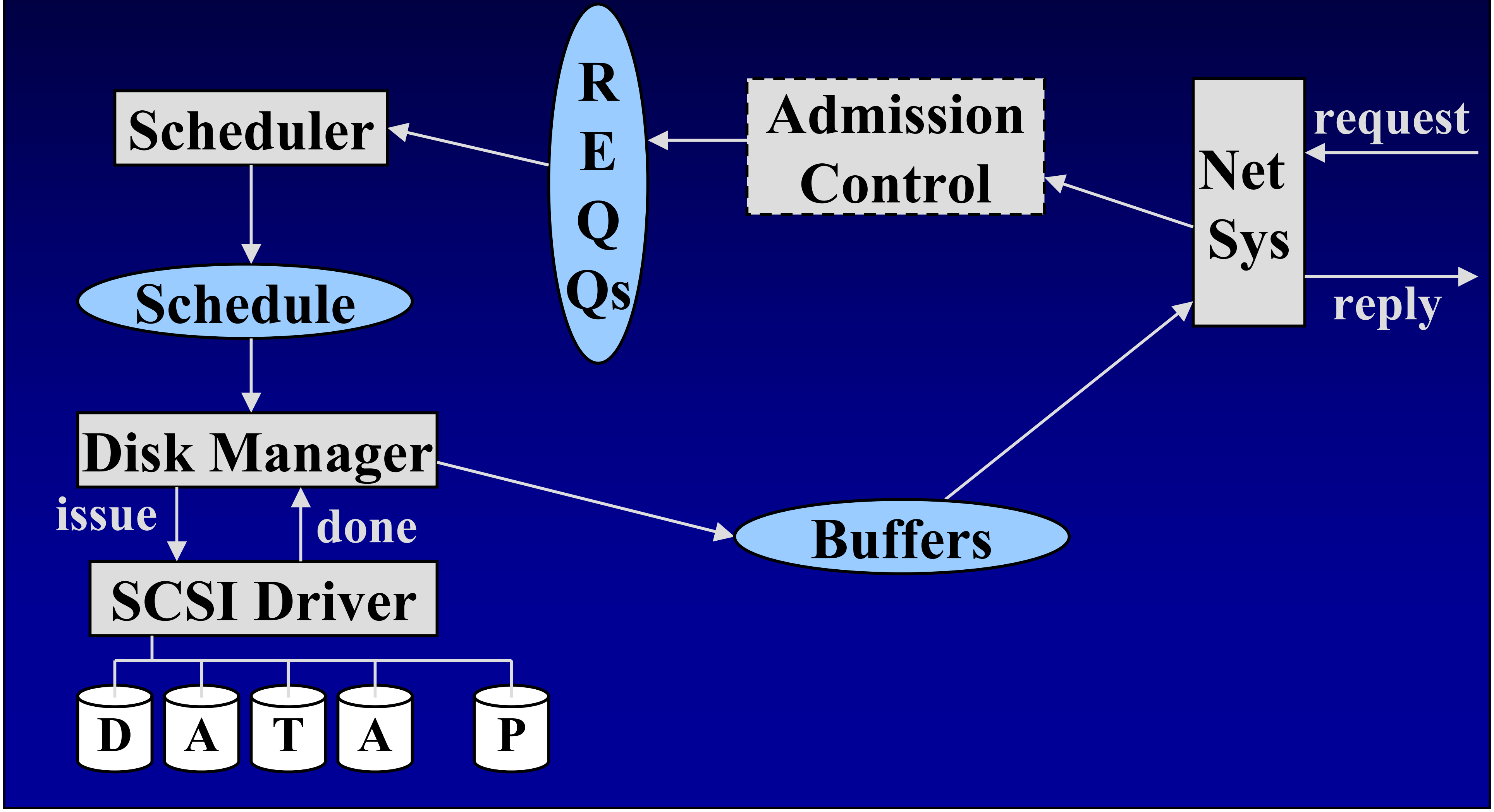
- Introduction
- Motivation
- Functionalities
- Interface

You are here!

- Basic Working
- Performance Optimizations
- Performance Measurements
- Current Research
- Related Work

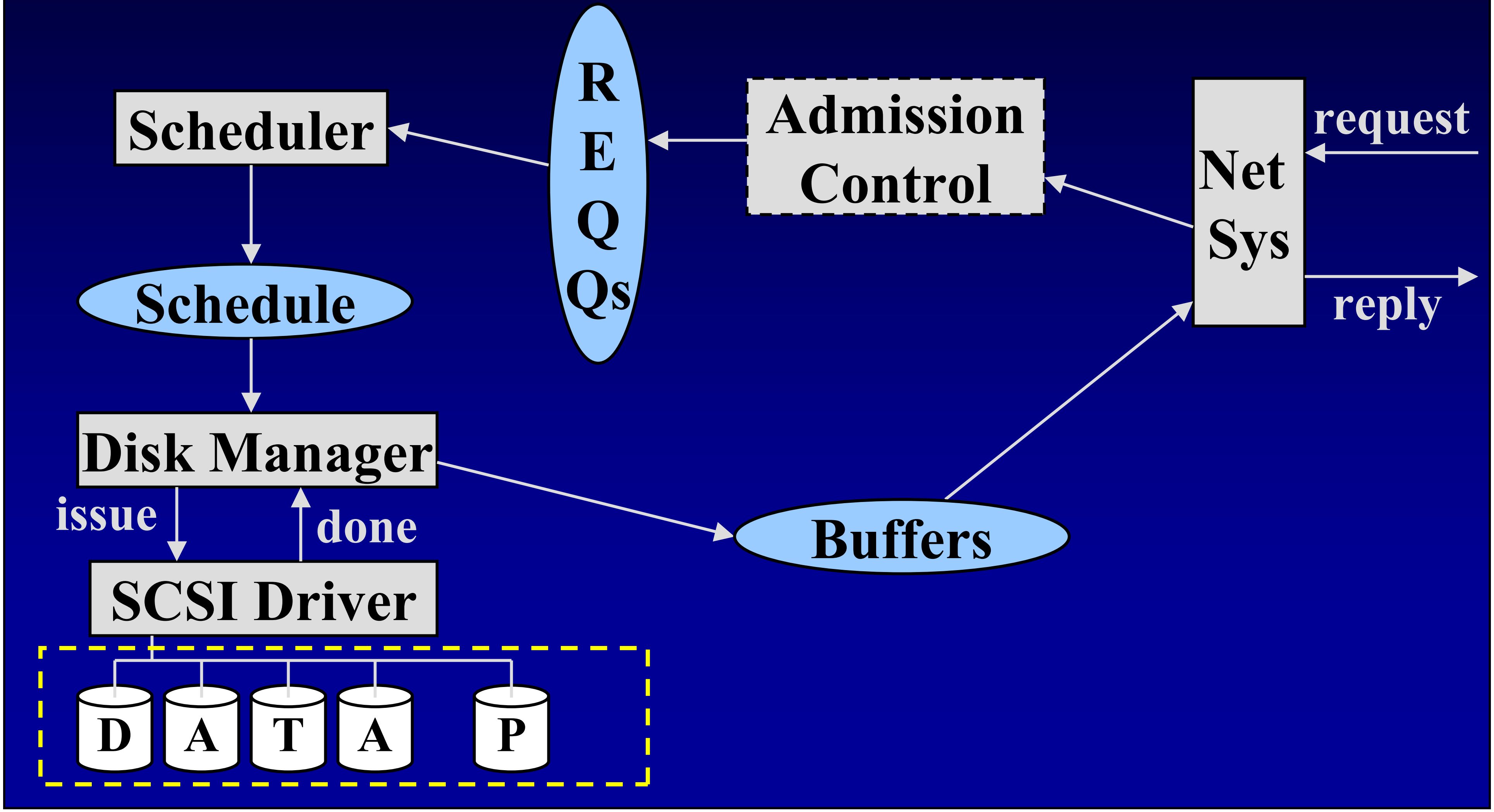
Basic Working (Normal Case)

Software Architecture



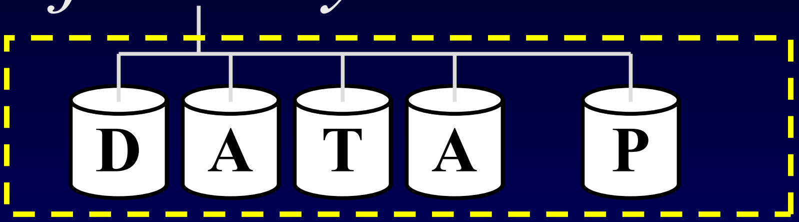
Basic Working (Normal Case)

Software Architecture



Basic Working (Normal Case)

Object Layout on Disks



Stripe

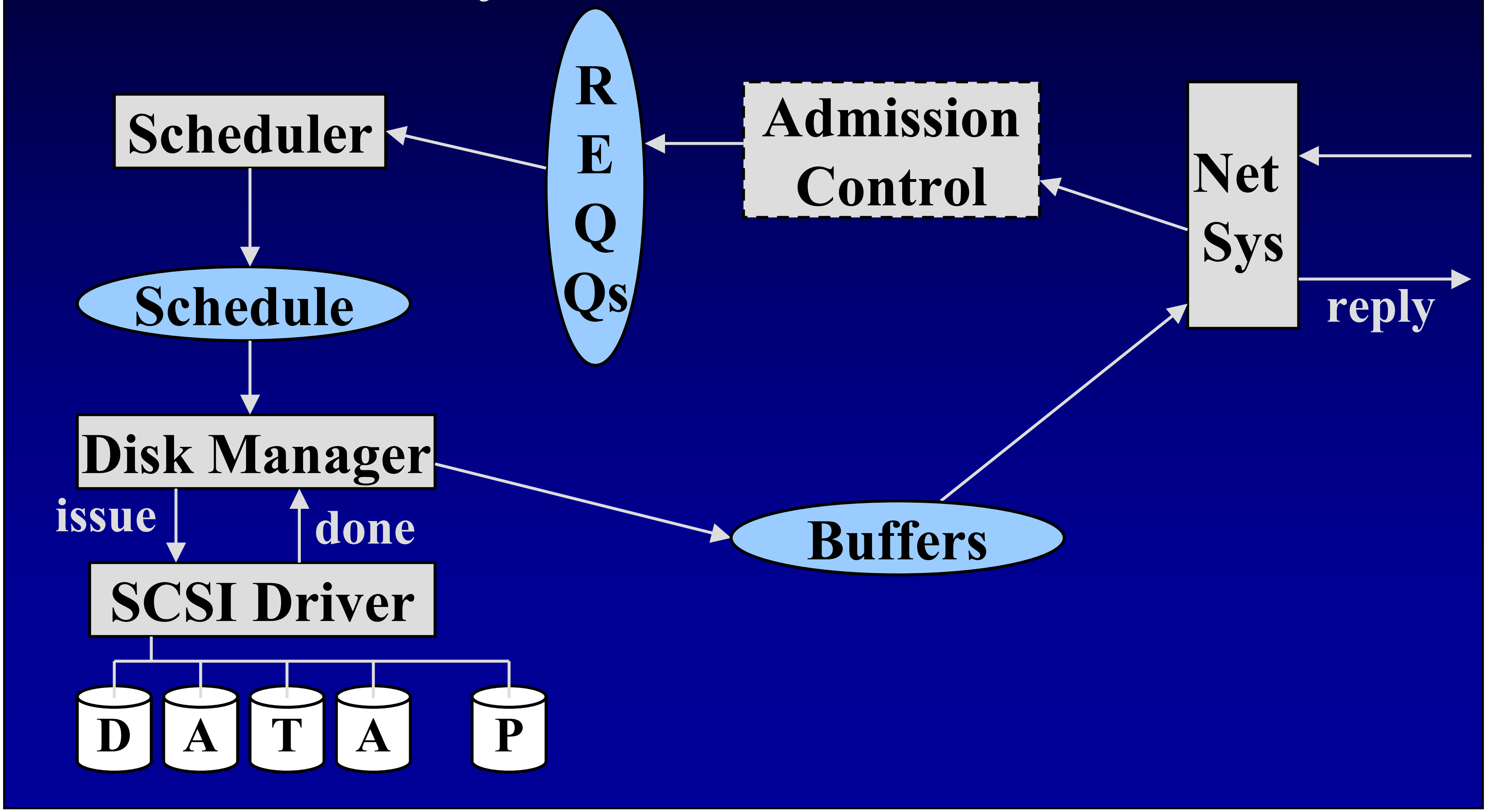
1	2	3	4	P1
5	6	7	8	P2
9	10	11	12	P3
⋮	⋮	⋮	⋮	⋮
N-3	N-2	N-1	N	P N/4

Object

Parity

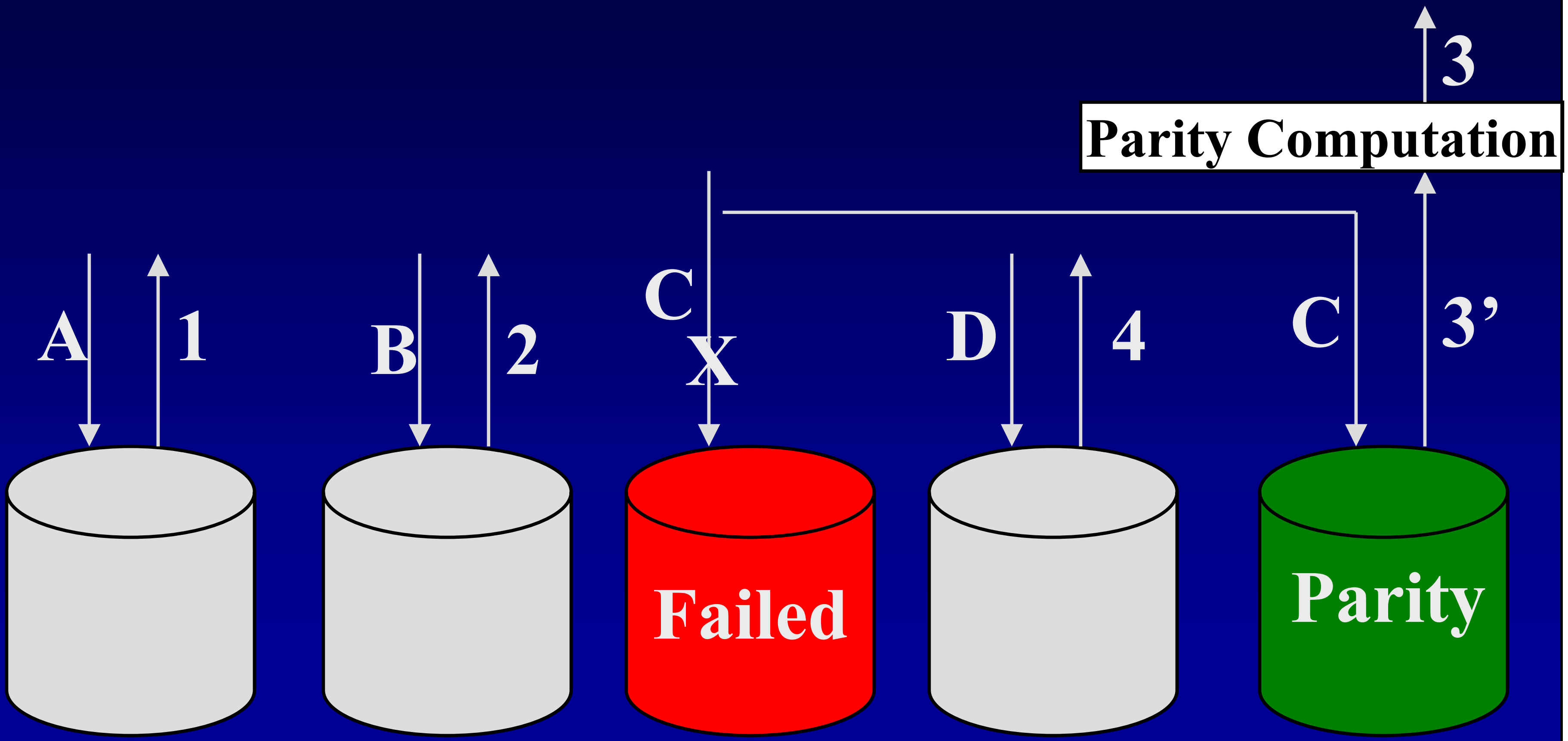
Basic Working (Normal Case)

Software Architecture



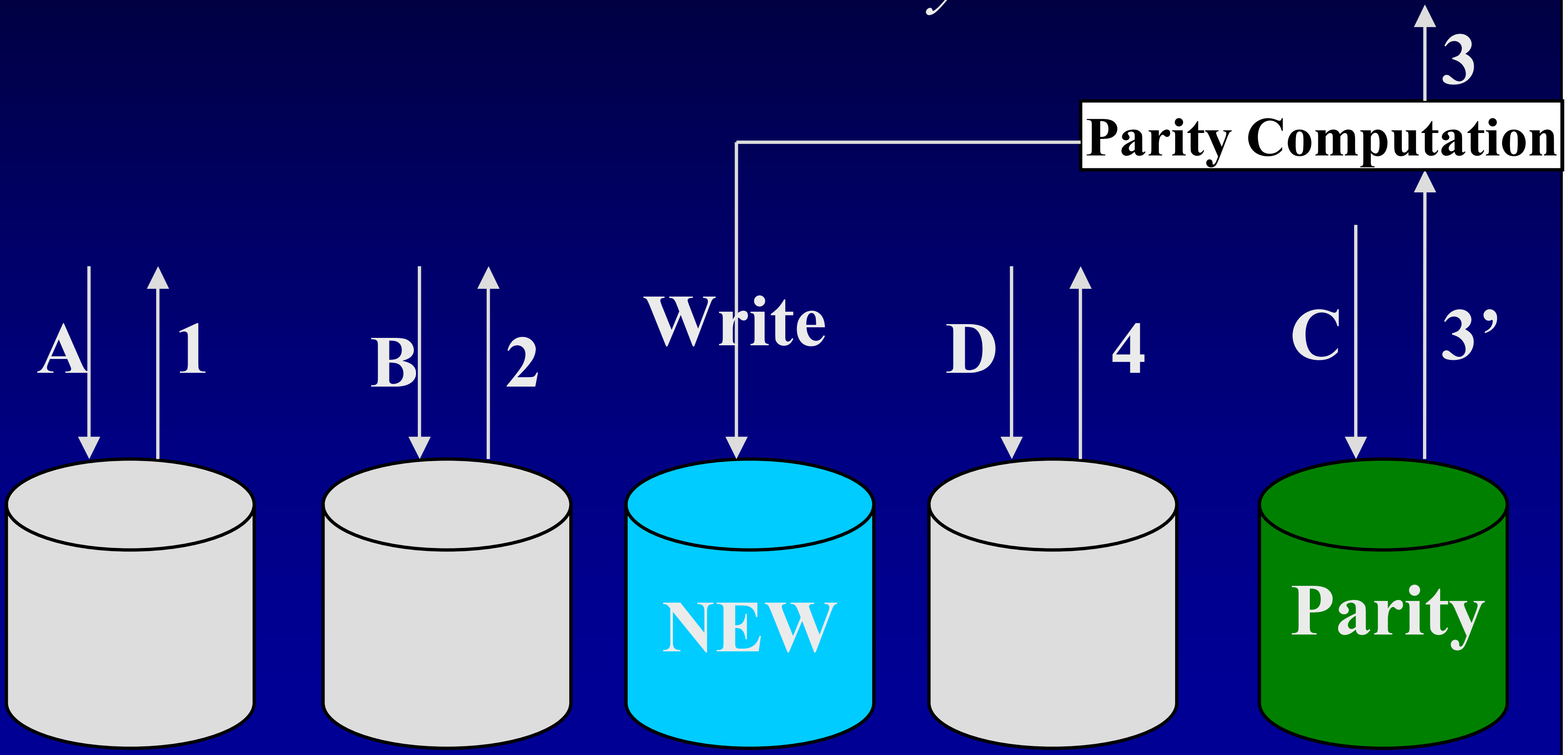
Basic Working (Failure Case)

Service Availability



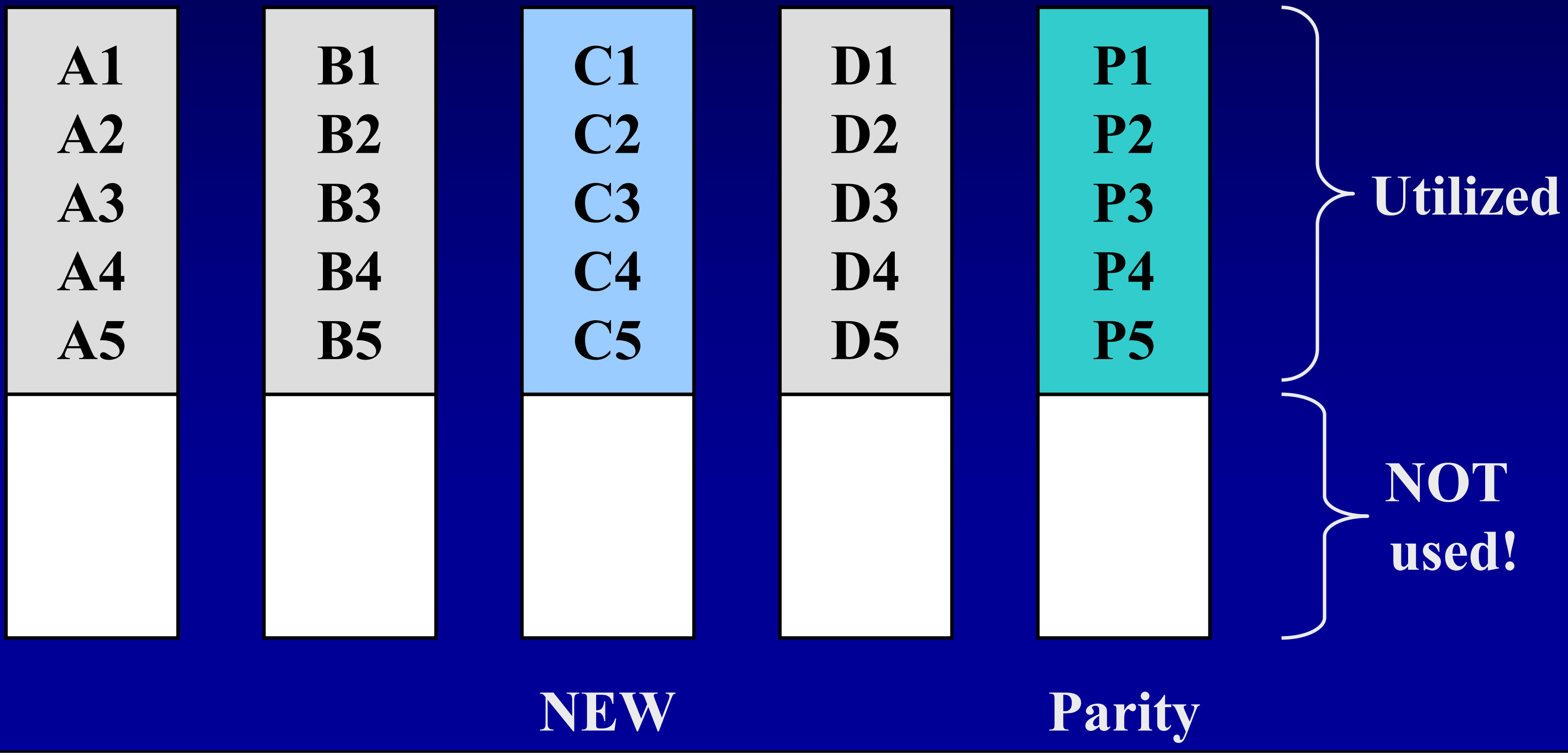
Basic Working (Failure Case)

Data Recovery



Performance Optimization-1

Observation



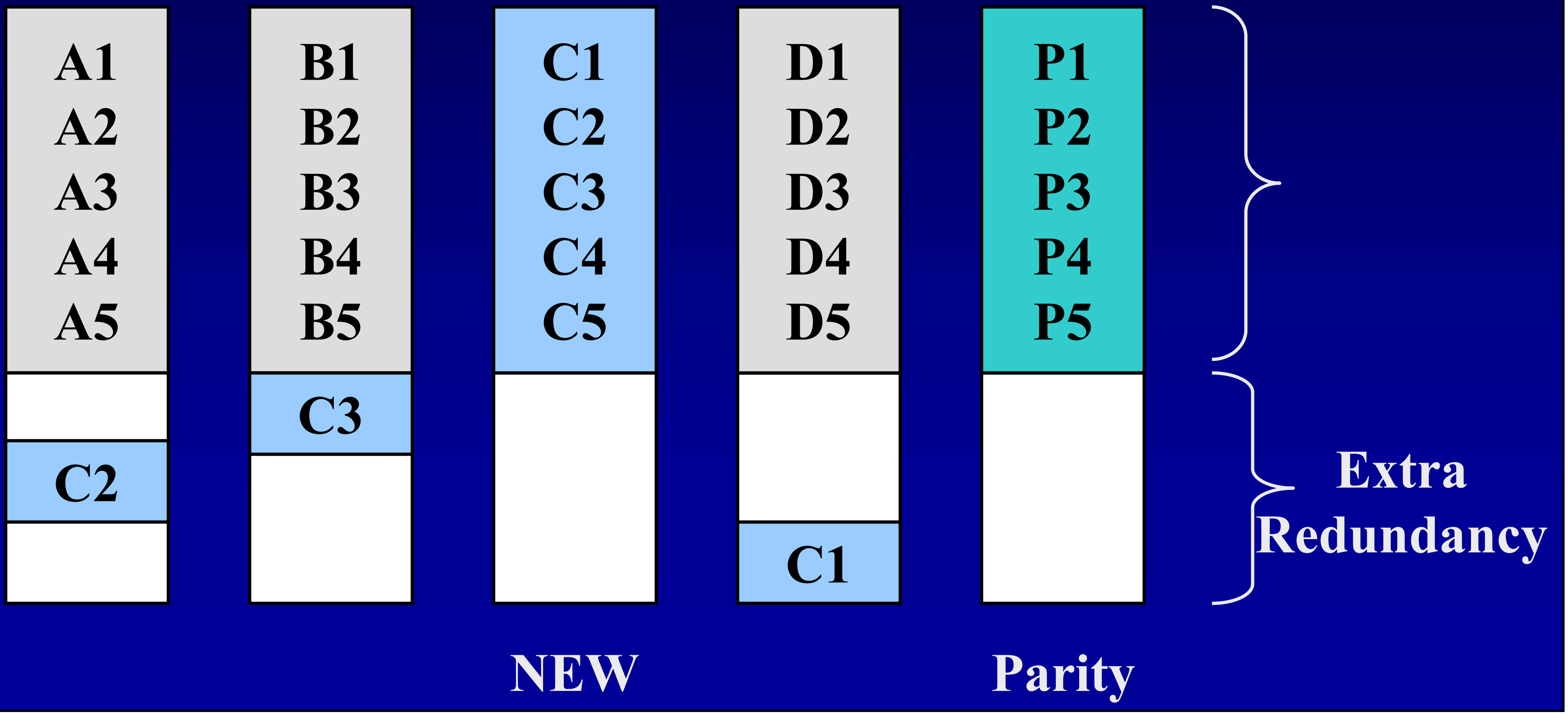
A1	B1	C1	D1	P1
A2	B2	C2	D2	P2
A3	B3	C3	D3	P3
A4	B4	C4	D4	P4
A5	B5	C5	D5	P5
		NEW		Parity

Utilized

NOT used!

Performance Optimization-1

Utilizing unused space
for faster data reconstruction



A1 A2 A3 A4 A5 C2

B1 B2 B3 B4 B5 C3

C1 C2 C3 C4 C5

NEW

D1 D2 D3 D4 D5 C1

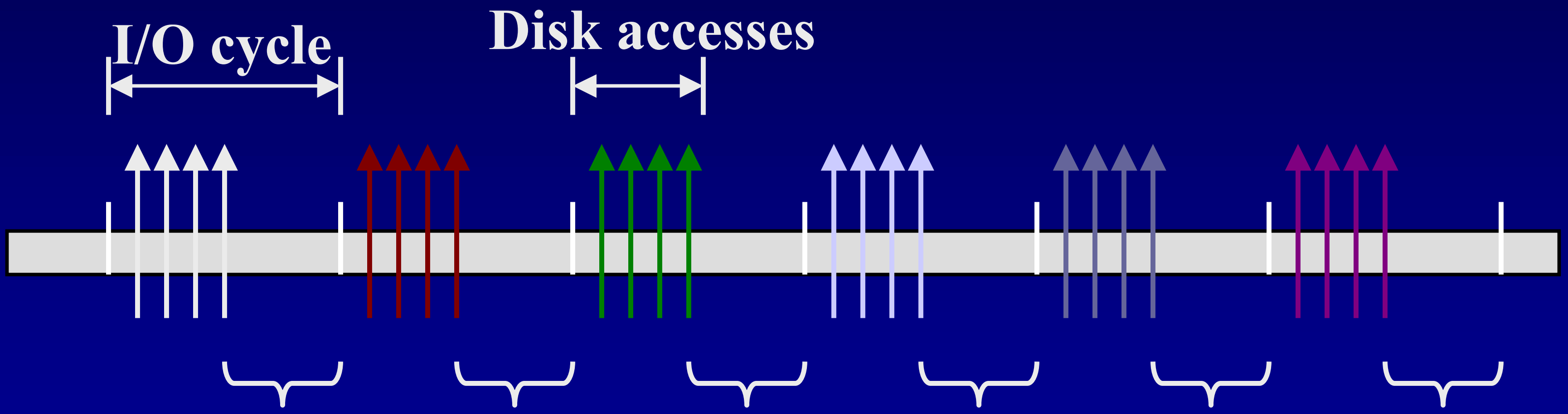
P1 P2 P3 P4 P5

Parity

Extra Redundancy

Performance Optimization-2

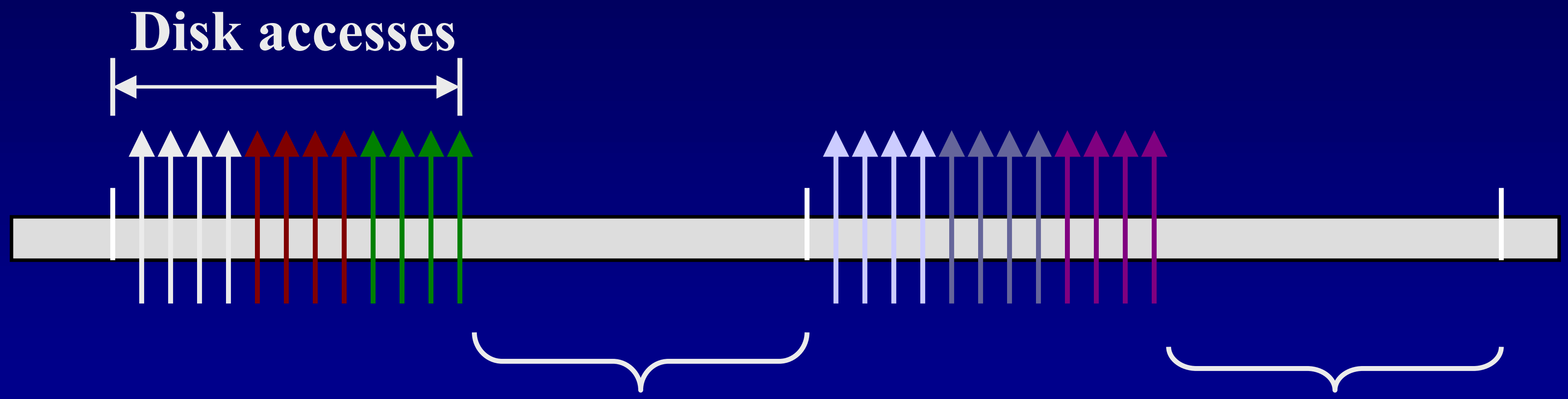
Observation



Unused disk cycles,
But disks are consuming power!

Performance Optimization-2

Lowered Power Consumption
for Increased Disks Reliability



Disks switched to
"low-power" mode

Performance Measurements

Relevant Hardware Configuration

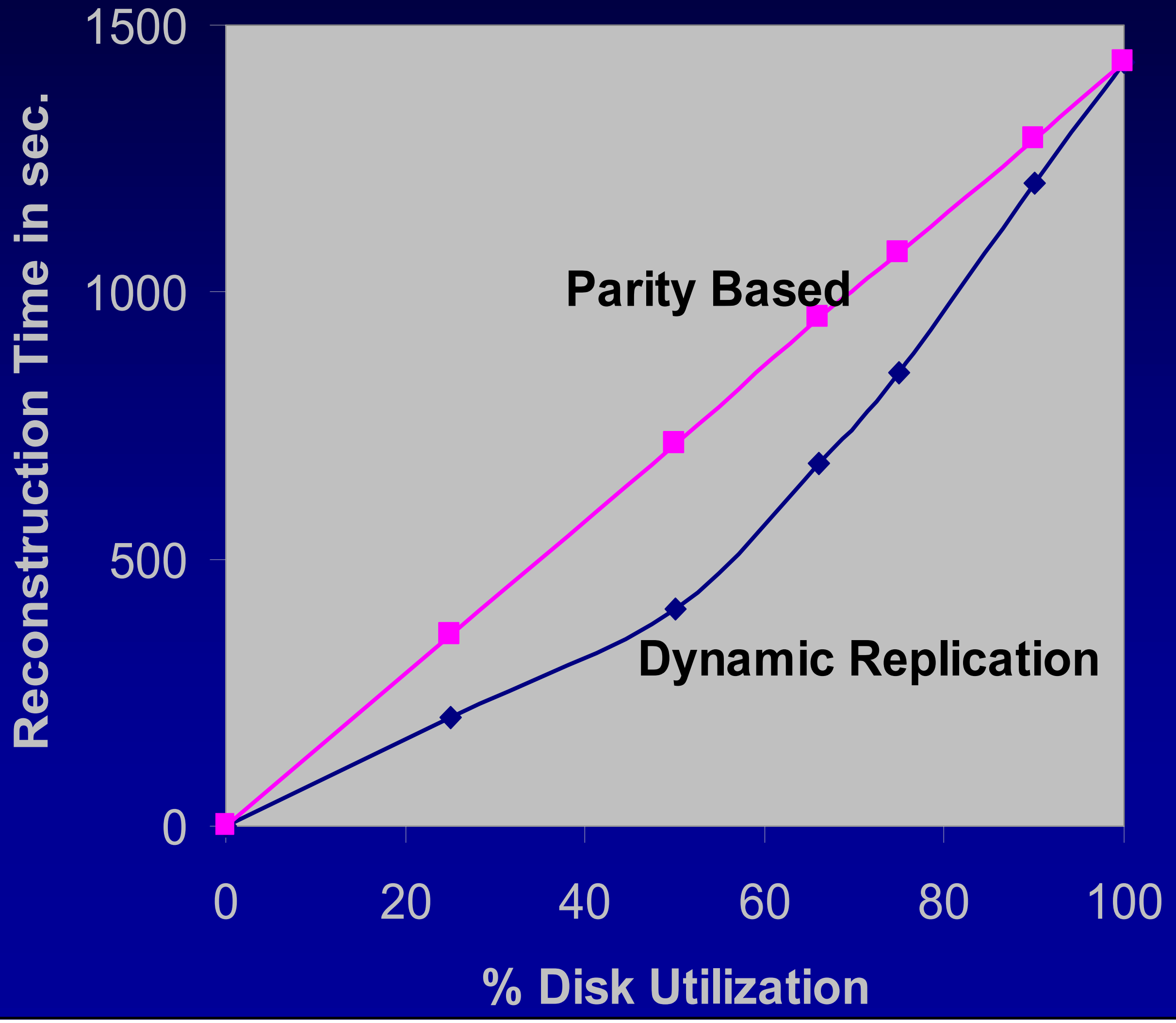
PentiumPro 200MHz PC, 128 MB RAM,
Array of five 4-GB Ultra Wide SCSI disks,
Ultra-Wide SCSI adapter sitting on a 33MHz PCI bus

Throughput (MPEG-1 streams)

Normal	52
Failure	42
Recovery	36

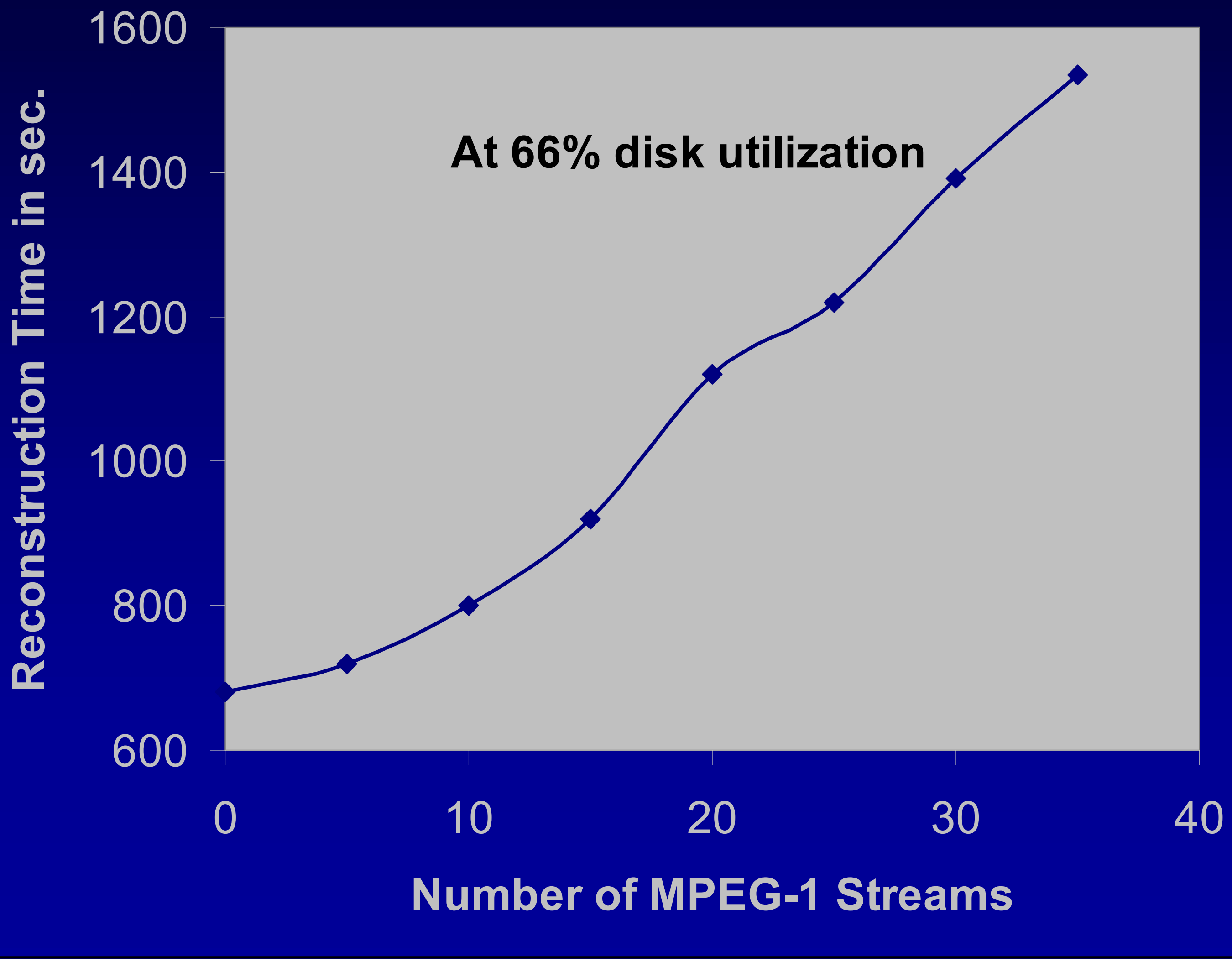
Performance Measurements

Raw Reconstruction



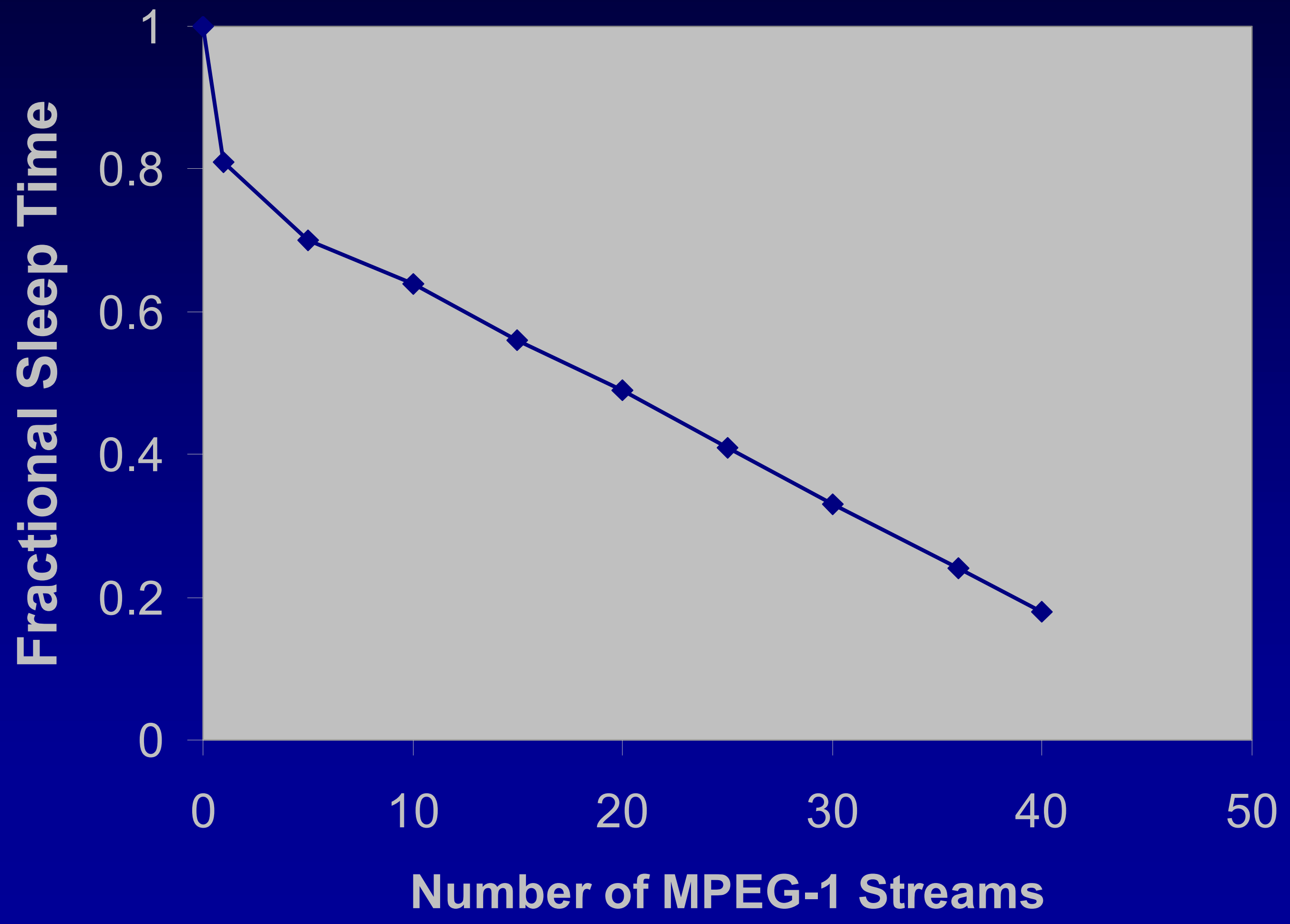
Performance Measurements

Reconstruction with client-streams



Performance Measurements

Lowered Power Consumption



Current Research

QoS-guaranteed virtual disks

A client can specify (storage, bandwidth, reliability) requirements independent of each other!

Self-management!

Related Work

- NASD project at CMU
- RAID systems
- Active Disks, HP AutoRAID
- Petal, Frangipani
- Global File System
- SB Video Server, Microsoft Tiger Server
- Power management for mobile computers

Road Map

- Introduction
- Motivation
- Functionalities
- Interface
- Basic Working
- Performance Optimizations
- Performance Measurements
- Current Research
- Related Work