Energy-Efficient and Reliability-Aware Data Management in Mobile Storage Systems

Tao Xie

Supported by NSF under grant CNS-0834466

Department of Computer Science
San Diego State University

Content

- Introduction and Motivation
- Research Tasks and Preliminary Results
- Education-Related Activities
- Future Research Directions

Introduction

- This project develops a hybrid disk array architecture, which integrates small capacity flash disks with hard disk drives to form a robust and energy-efficient storage system for mobile data-intensive applications.
- An array of new data management techniques for data-intensive mobile applications will be developed.
- A prototype and a simulation toolkit will be implemented.
- It will also promote teaching, learning, and training by exposing students to technological and scientific underpinnings in the field of energy-efficient storage systems.

Stationary Data Centers

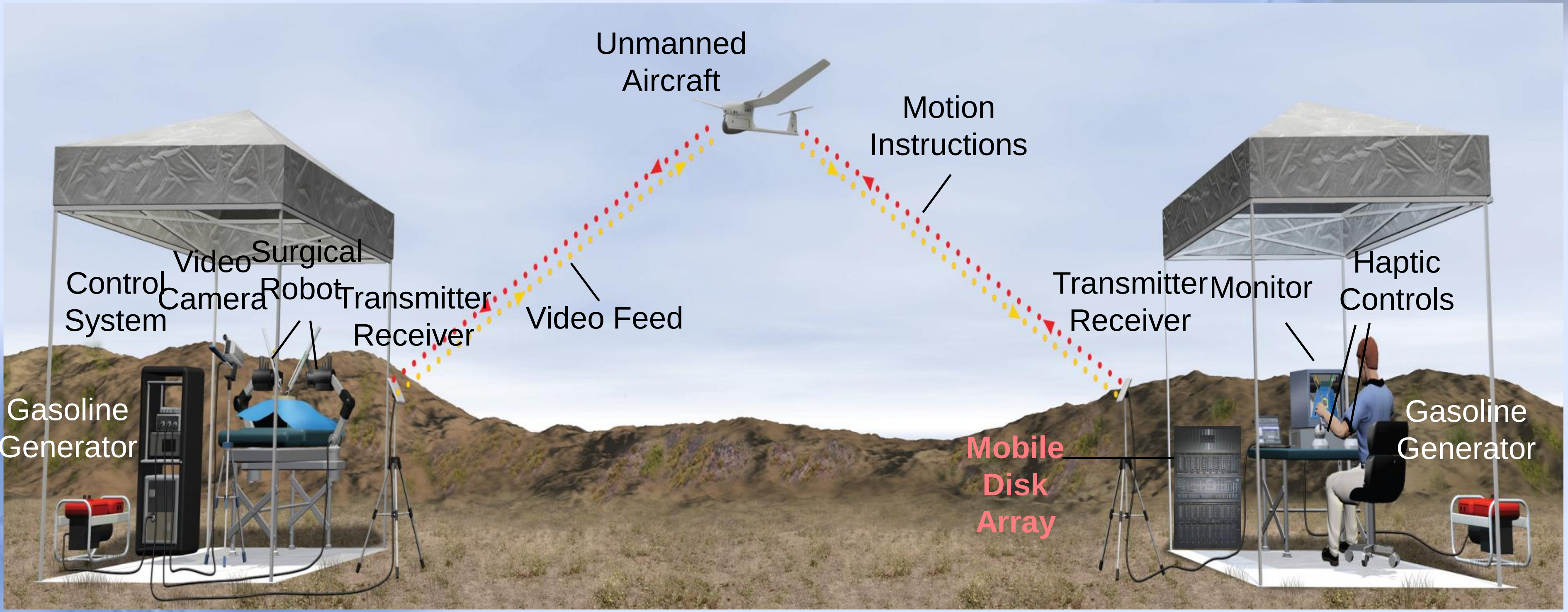
Mobile Disk Arrays

- Existing mobile disk array consists of an array of independent small form factor hard disks connected to a host by a storage interface in a mobile computing environment.

- Hard disks have some intrinsic limitations such as long access latencies, high annual disk replacement rates, fragile physical characteristics, and energy-inefficiency.

- Due to their severe application environments, mobile disk arrays must be energy-efficient, extremely reliable, highly fault-tolerant and physically robust.

Application One: Remote Surgery



Application Two: Mobile Data Center

New Challenges of Mobile Disk Arrays

- Very limited power supply
- Stringent reliability requirement
- High demands on fault-tolerance
- Robust physical characteristics

File Open Issues That Will Be Addressed

1. The lack of a high-performance, highly reliable, and energy-efficient storage architecture
2. New energy-saving data management schemes for mobile data-intensive applications
3. Understanding of the relationship between disk energy saving techniques and disk reliability
4. The absence of an energy-ware fault-tolerant mechanism
5. A prototype and a simulation toolkit

Five Research Tasks

1. Develop a hybrid disk storage architecture
2. Develop a reliability model
3. Establish an energy conservation infrastructure
4. Develop an energy-aware fault-tolerant mechanism
5. Implement a mobile disk array prototype and a simulation toolkit

Education Objectives

- To train 1 Ph.D. student and 2 undergraduate students
- To conduct a training workshop
- To develop one senior-level undergraduate course on energy-aware storage systems

Task 1: Developing a flash-hard hybrid disk storage architecture

1. We are implementing a novel *f*lash d*i*sk s*t*orage architecture (FIT) for high performance, energy conservation and highly reliable mobile disk arrays.

2. The basic idea of the FIT architecture is to construct mobile disk arrays by using both non-volatile NAND flash memory based SSD (solid state disk) and small-factor hard disk drives.

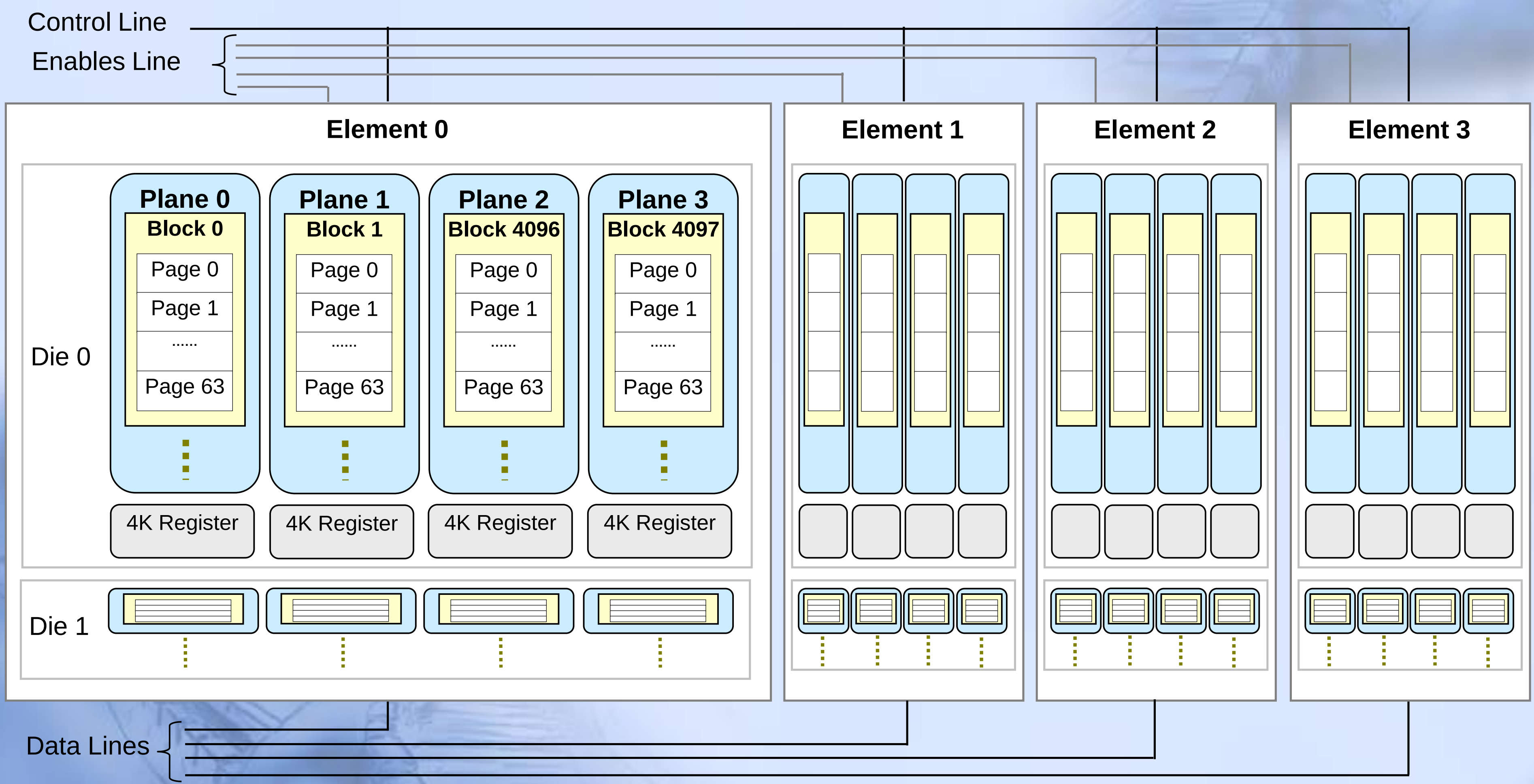
Flash SSD

1. They are physically robust with high vibration-tolerance and shock-resistance.
2. They inherently consume much less energy than mechanical mechanism based hard disks.
3. They offer much fast read access times.
4. Very recent breakthrough largely relaxes the three constraints on existing flash disks: small capacity, low throughput, and limited erasure cycles.

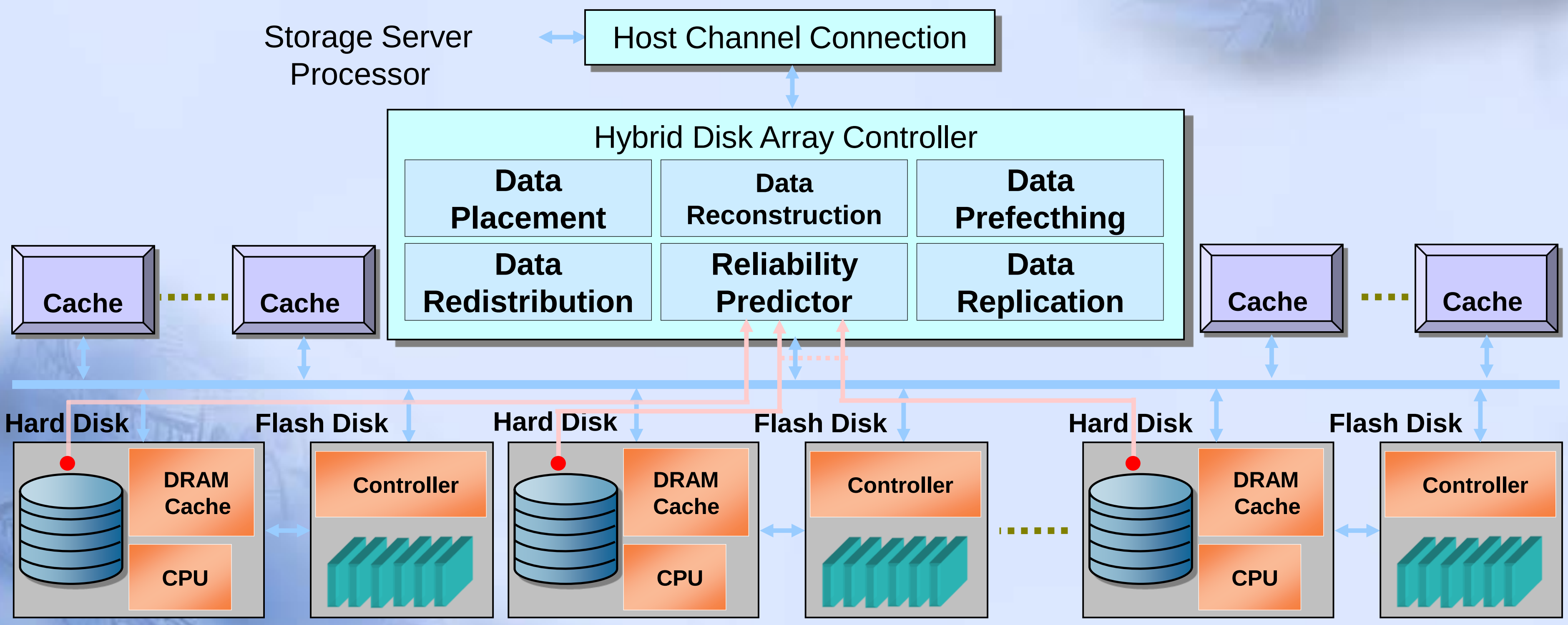
Flash SSD vs. HDD

2.5" SATA 3.0Gbps SSD		2.5" SATA 3.0Gbps HDD
Solid NAND flash based	**Mechanism type**	Magnetic rotating platters
64GB	**Density**	80GB
73g	**Weight**	365g
Read: 100MB/s, Write :80MB/s	**Performance**	Read: 59MB/s, Write: 60MB/s
1W	**Active Power consumption**	3.86W
20G (10~2000Hz)	**Operating Vibration**	0.5G (22~350Hz)
1,500G for 0.5ms	**Shock resistance**	170G for 0.5ms
0°C to 70°C	**Operating temperature**	5°C to 55°C
None	**Acoustic Noise**	0.3 dB
MTBF >2M hours	**Endurance**	MTBF < 0.7M hours

Internal Structure of a SSD with Four Elements



The FIT Architecture



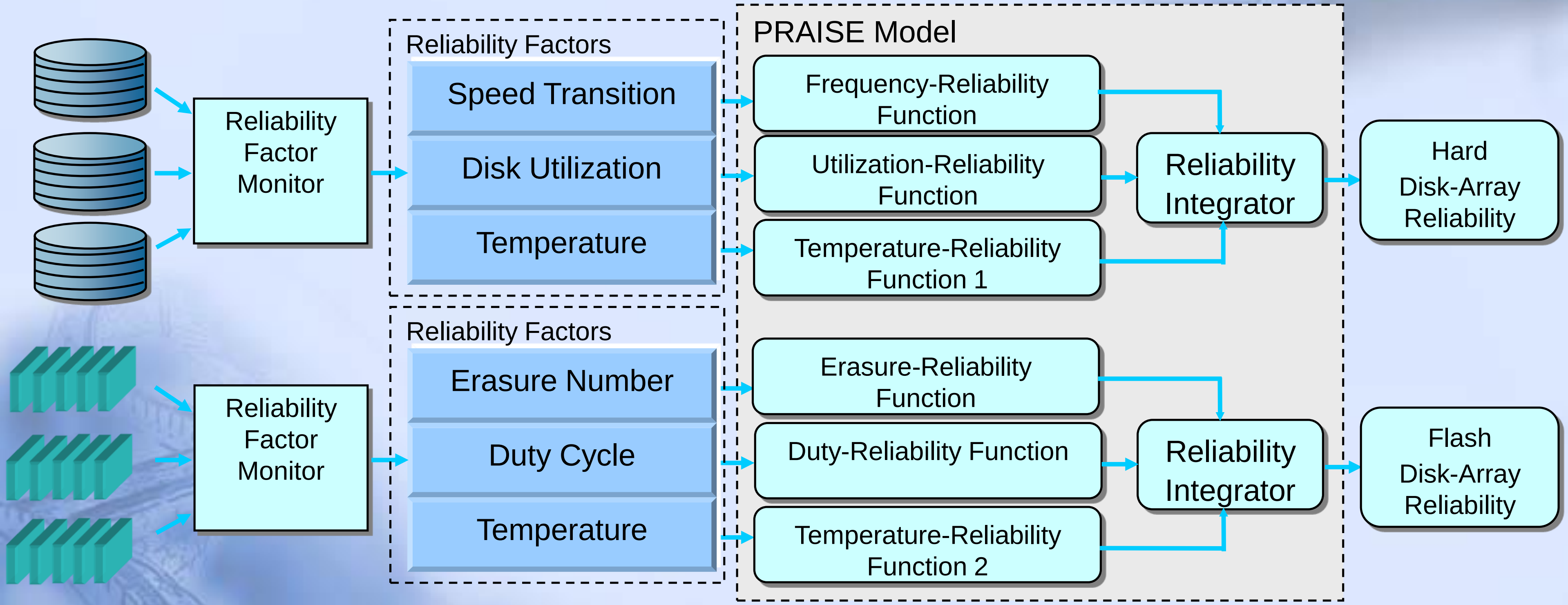
Task 2: Developing a disk reliability predictor

1. Developing a reliability predictor that is capable of estimating failure rate for both hard disks and flash disks is challenging.

2. A deep understanding of the relationship between energy saving techniques and disk reliability is an open question.

3. The a *re*liability *p*redictor (REP) will be built.

The Reliability Predictor (REP)



The PRAISE Model

- PRAISE: *p*redictor of *r*eliability for fl*a*sh-assisted d*i*sk *s*torag*e*
- The top three reliability functions are dedicated for hard disk reliability estimation, whereas the bottom three reliability functions are reserved for flash disk reliability prediction.
- We plan to use data mining approaches to discover the relationship between reliability-affecting factors and the reliability level.

Preliminary Results of Task 2

- We developed an empirical reliability model, called PRESS (Predictor of Reliability for Energy Saving Schemes) [Xie and Sun, IPDPS'08]

- Fed by operating temperature, disk utilization, disk speed transition frequency, three energy-saving-related reliability affecting factors, PRESS estimates the reliability of entire hard disk array

Operating Temperature

- High temperature was discovered as a major culprit for a number of disk reliability problems.

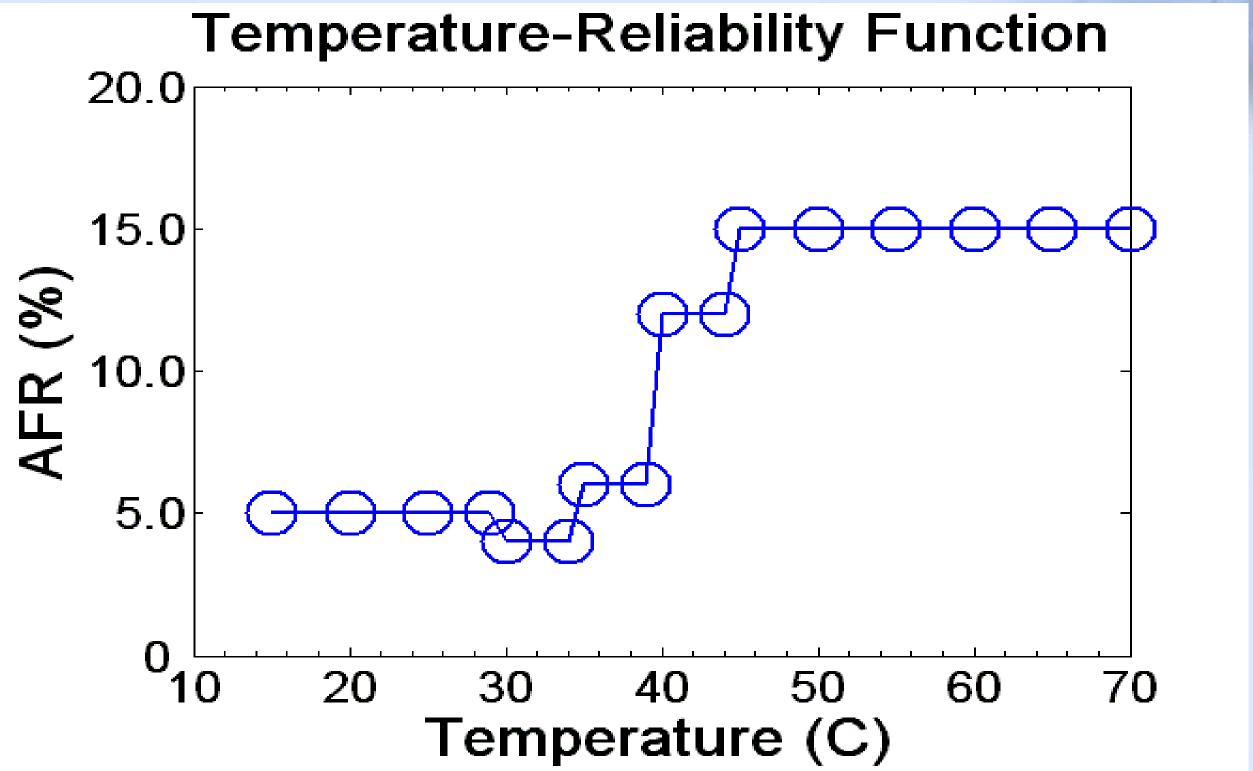
- Results from both camps indicate that disk operating temperature generally has observable effects on disk reliability.

A Two-Speed Disk

- Assume that the low speed mode is 3,600 RPM (revolutions per minute) and the high speed mode is 10,000 RPM.

- Based on related work, we derive the temperatures of two-speed disks as “[45, 50] C for the high speed mode” and “[35, 40] C for the low speed mode”.

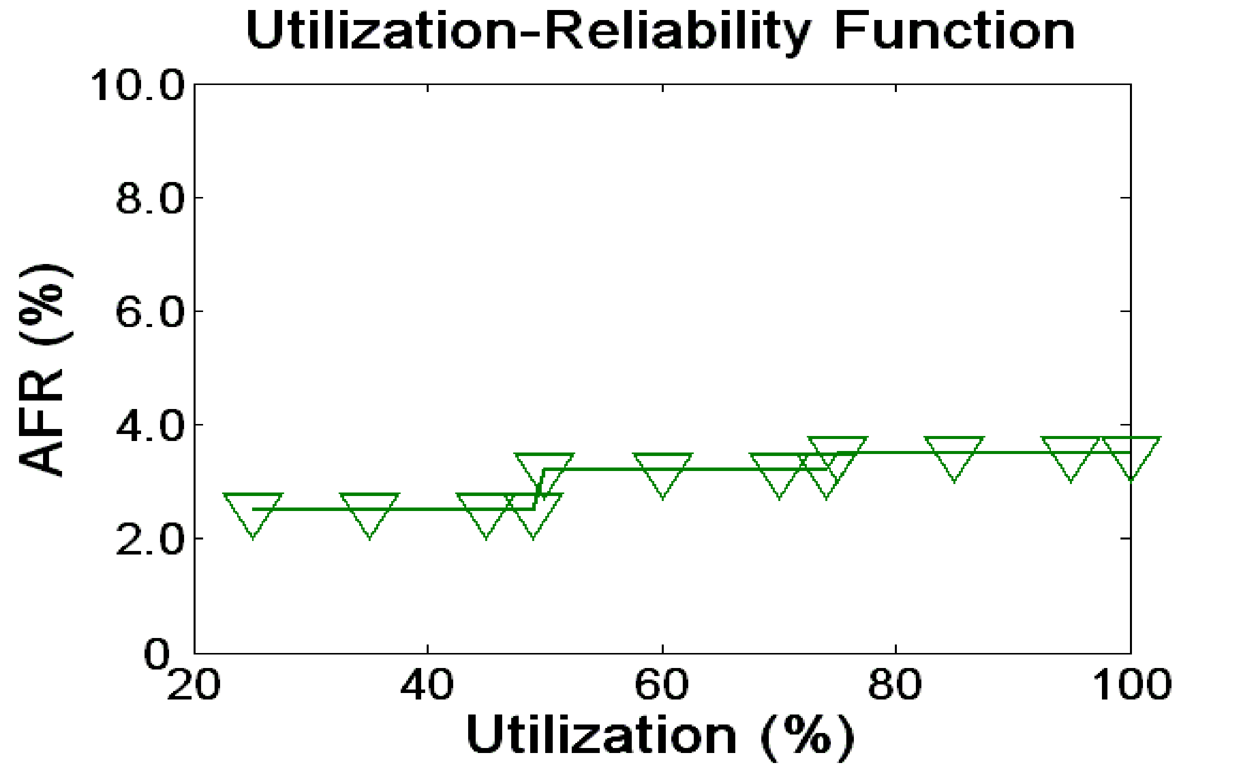
From Google Data [Pinheiro et al. 2007]



Disk Utilization

- Disk utilization is defined as the fraction of active time of a drive out of its total power-on-time.
- A conclusion that higher utilizations in most cases affect disk reliability negatively has been generally confirmed by two widely recognized studies ([Cole 2000] and [Pinheiro et al. 2007]).

From Google Data [Pinheiro et al. 2007]

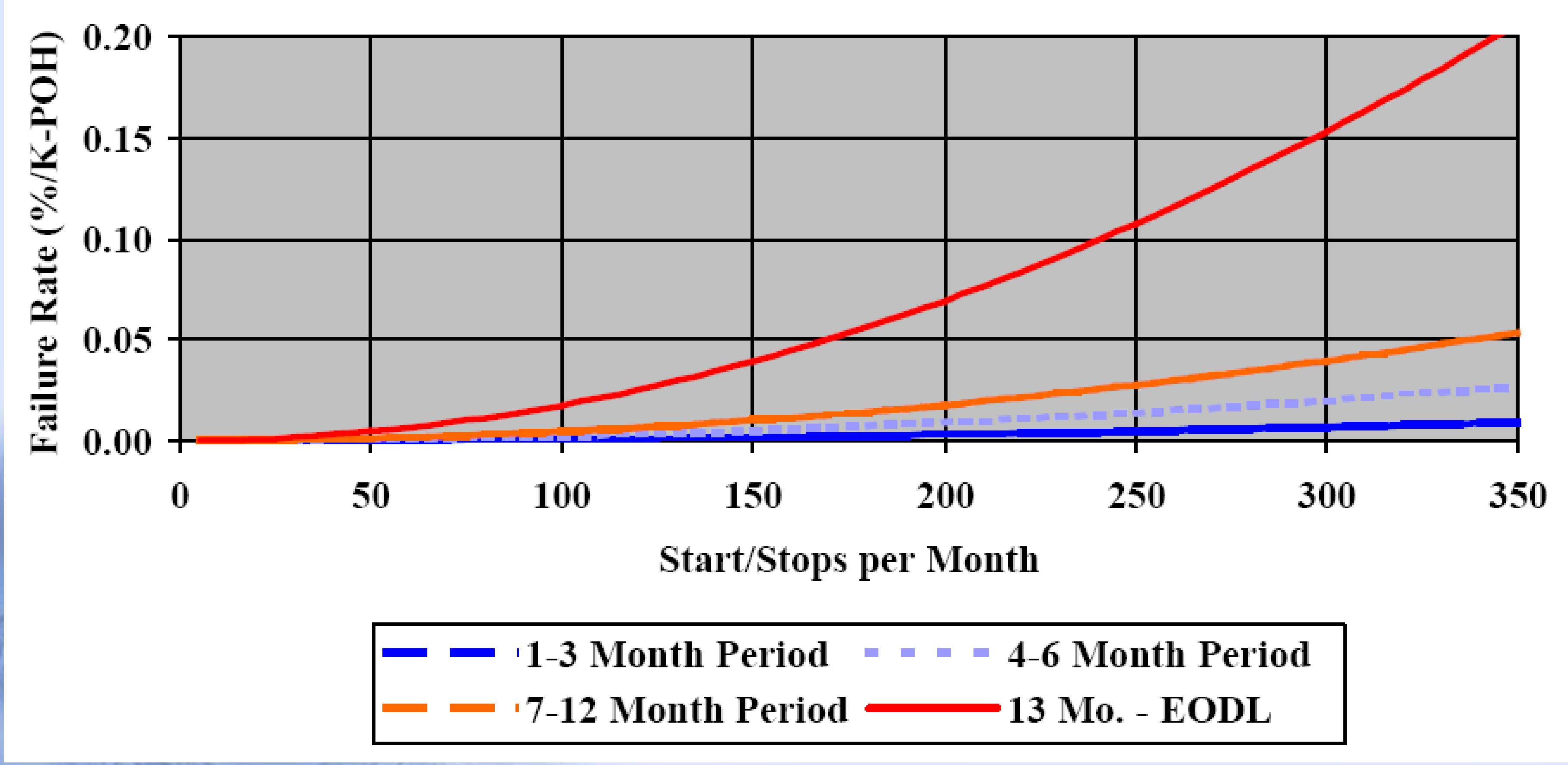


Disk Speed Transition Frequency

- The disk speed transition frequency (hereafter called frequency) is defined as the number of disk speed transitions in one day.

- The frequency-reliability function is built on a combination of the spindle start/stop failure rate adder suggested by IDEMA and the modified Coffin-Manson model.

Spindle Start/Stop Failure Rate Adder (IDEMA)



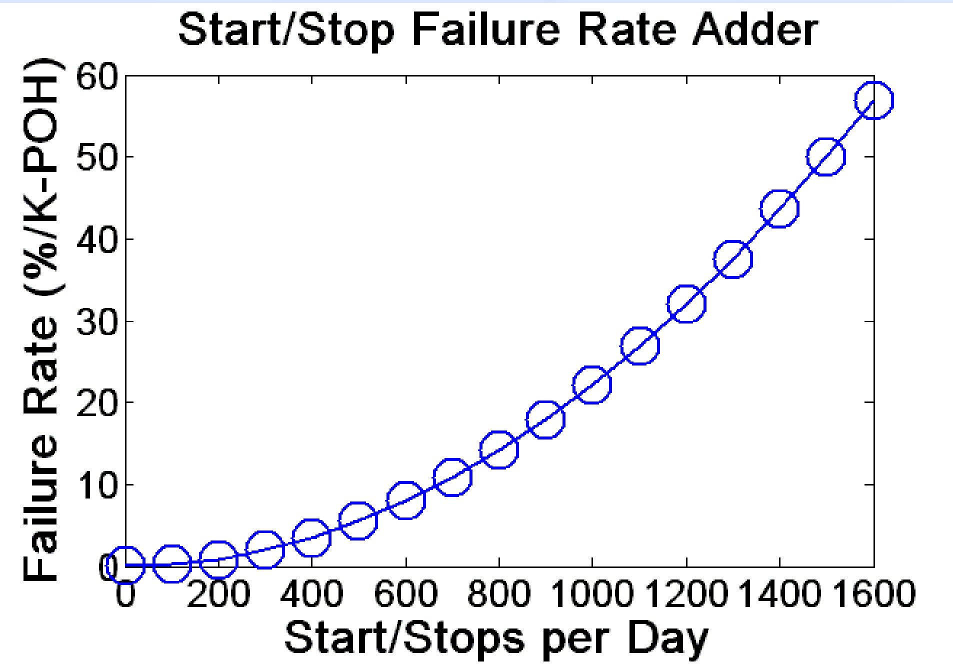
Modified Coffin-Manson Model

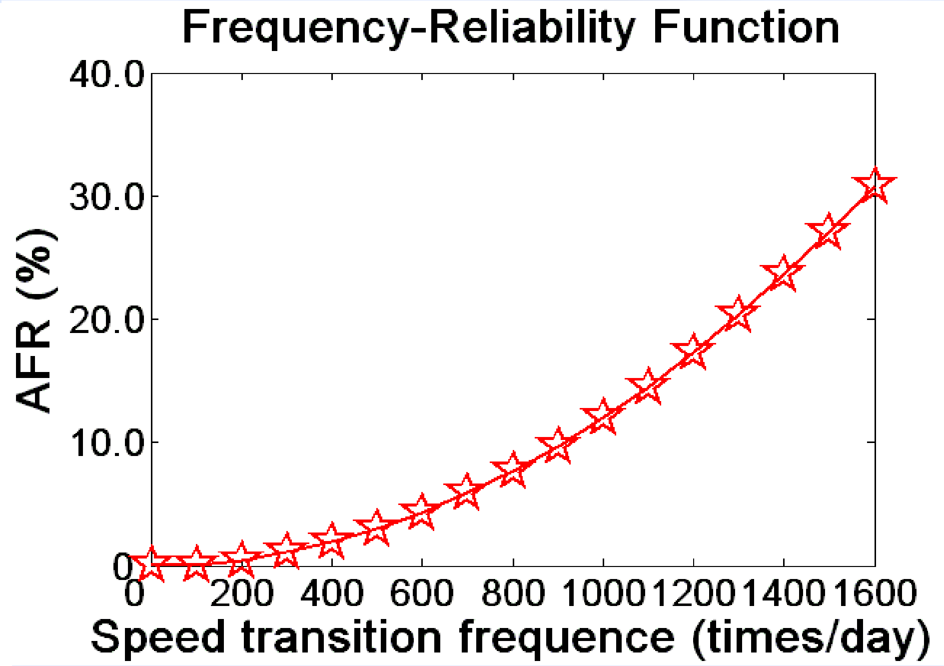
$$N_f = A_0 f^{-\alpha} \Delta T^{-\beta} G(T_{\max})$$

N_f is the number of cycles to failure, A_0 is a material constant, f is the cycling frequency, ΔT is the temperature range during a cycle, and $G(T_{\max})$ is an Arrhenius term evaluated at the maximum temperature reached in each cycle.

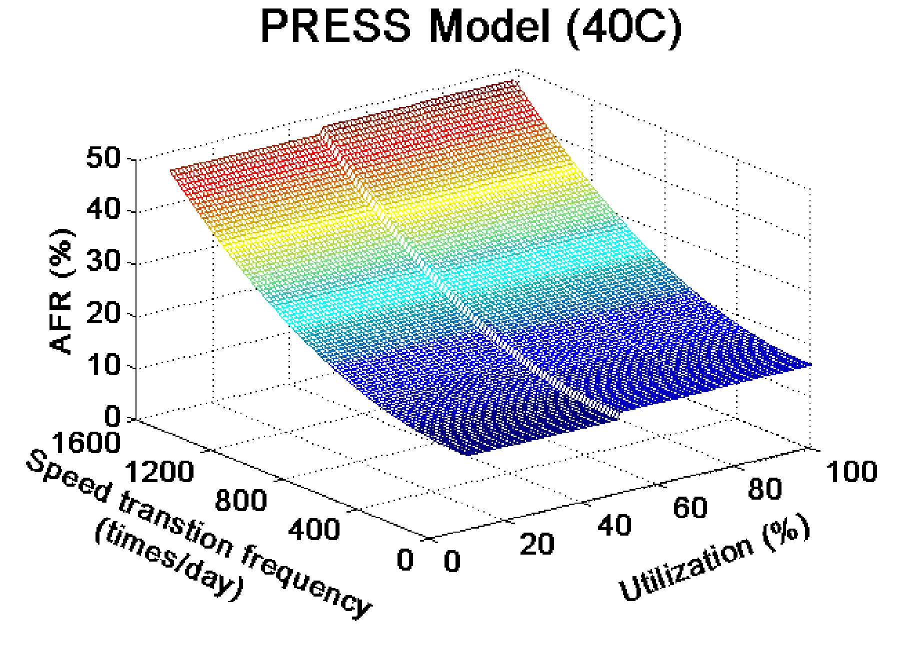
Frequency-reliability Function

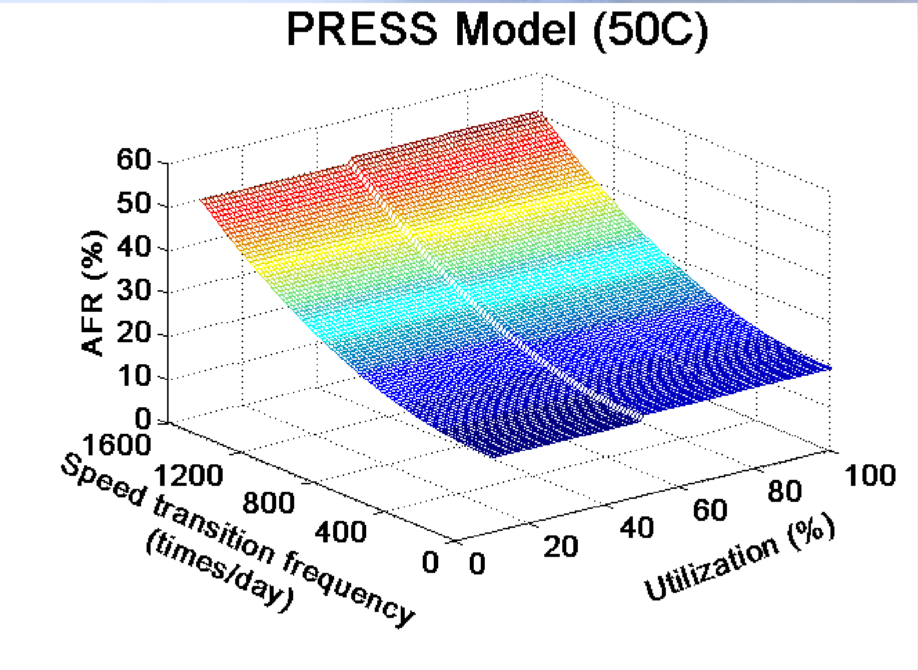
$$R(f) = 1.51e^{-5} f^2 - 1.09e^{-4} f + 1.39e^{-4}, f \in [0,1600]$$





PRESS It Altogether

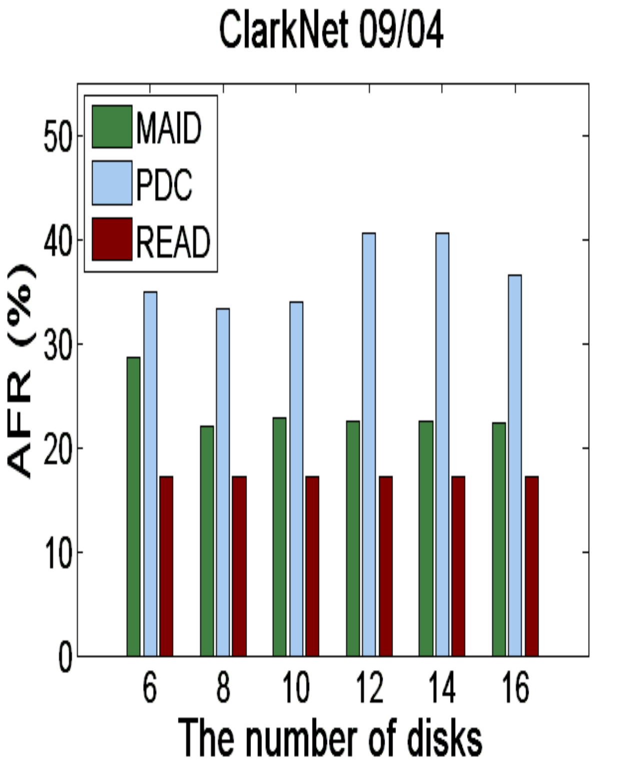


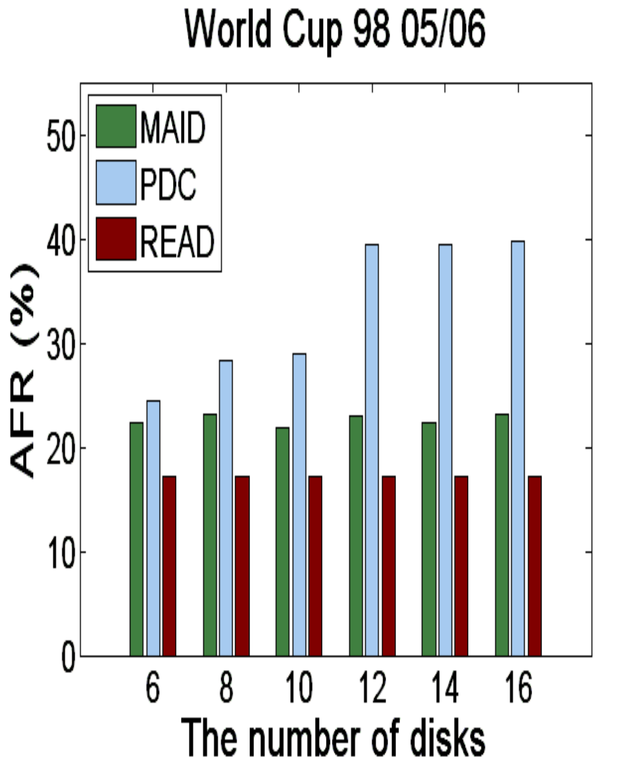


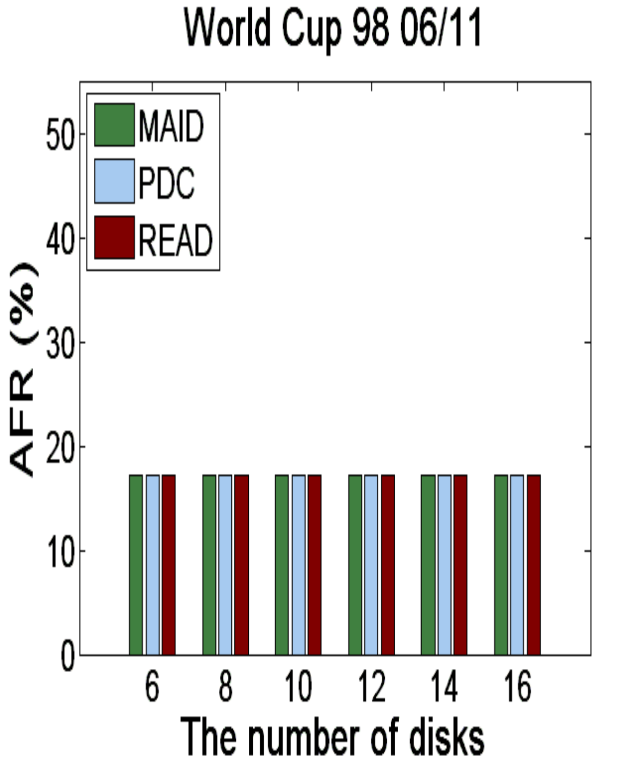
The Idea of READ

- The general idea of READ is to control disk speed transition frequency based on the statistics of the workload.
- It employs a dynamic file redistribution scheme to periodically redistribute files across a disk array in an even manner.
- A low disk speed transition frequency and an even distribution of disk utilizations imply a lower AFR based on our PRESS model.

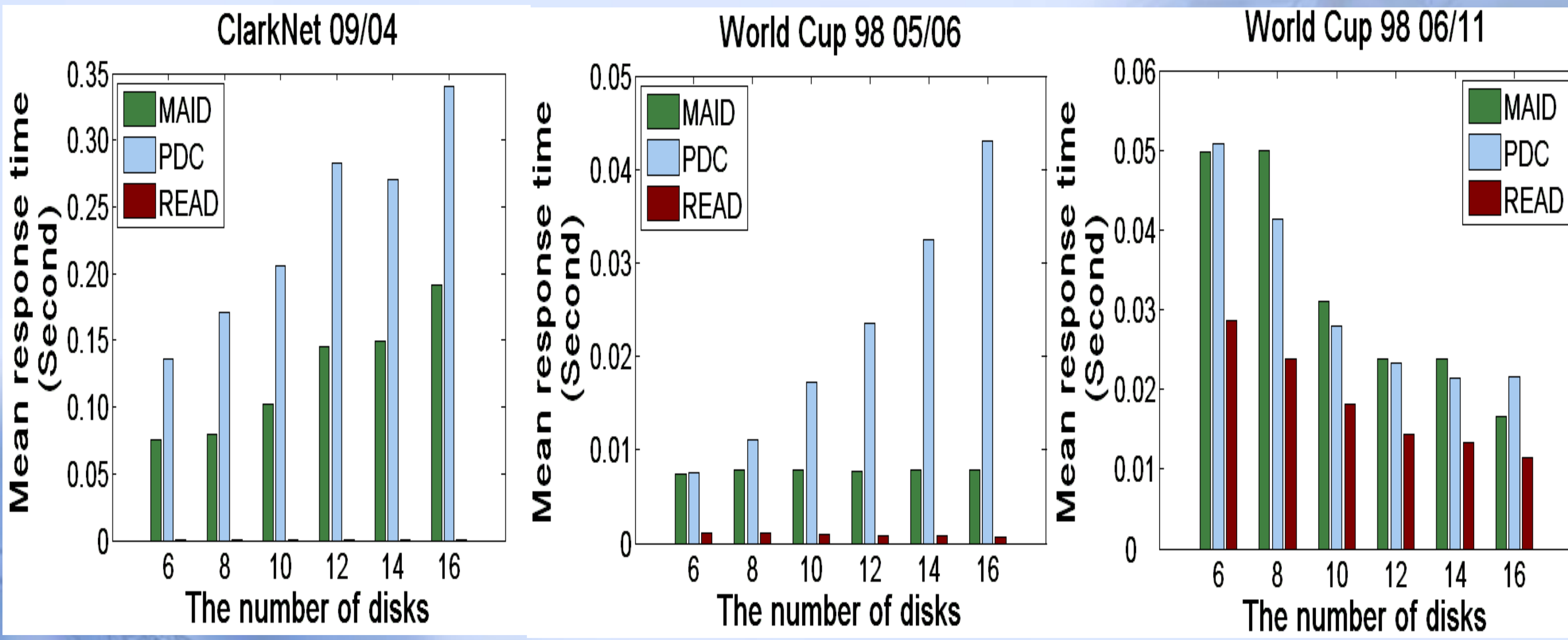
Reliability



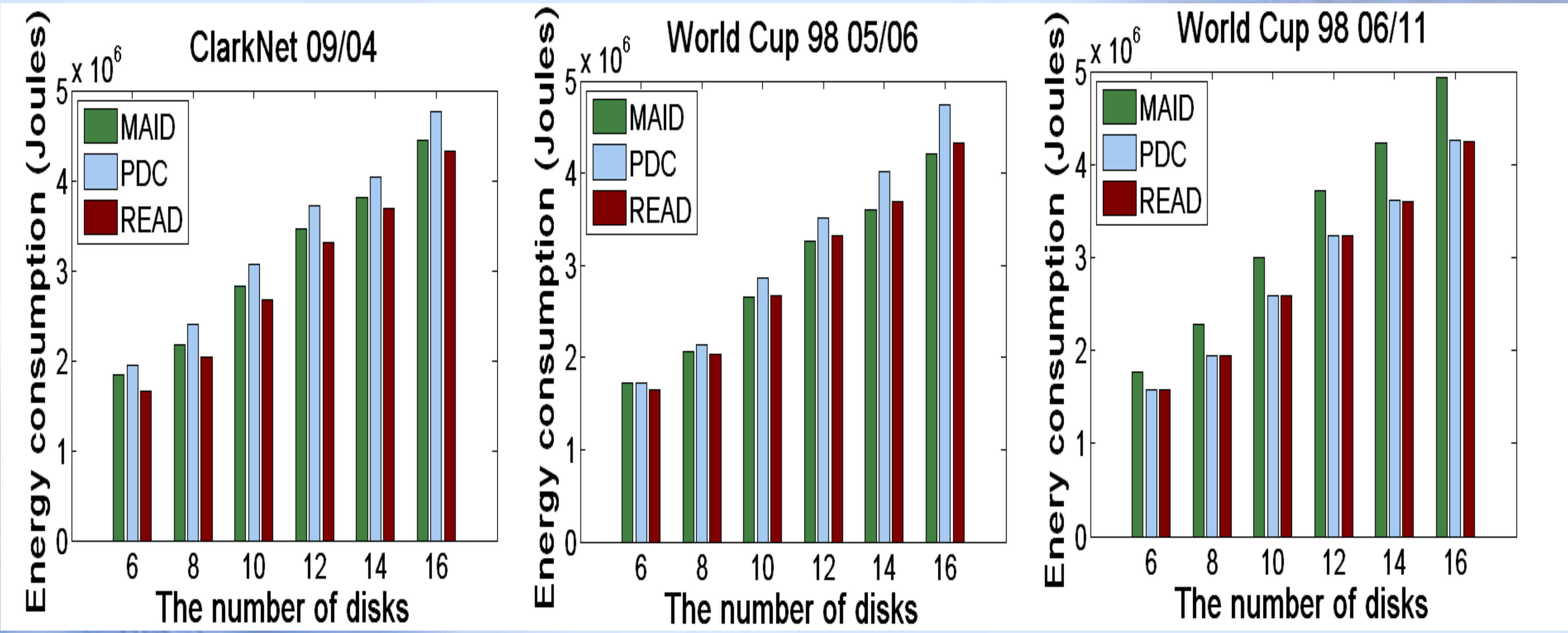




Performance



Energy Conservation



Task 3: Developing data management schemes

- Traditional data management schemes like data placement algorithms only concentrated on improving system performance data reliability, while normally overlooked energy efficiency.

- An array of energy-aware data management software modules: data placement algorithms, data redistribution strategies, data replication policies, and data prefetching schemes will be developed.

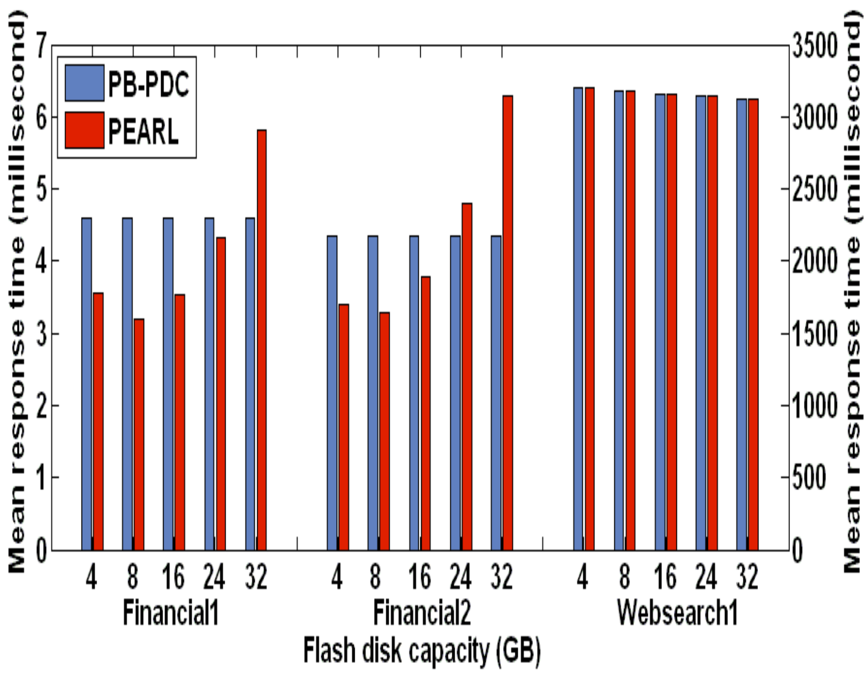
Modules in Task 3

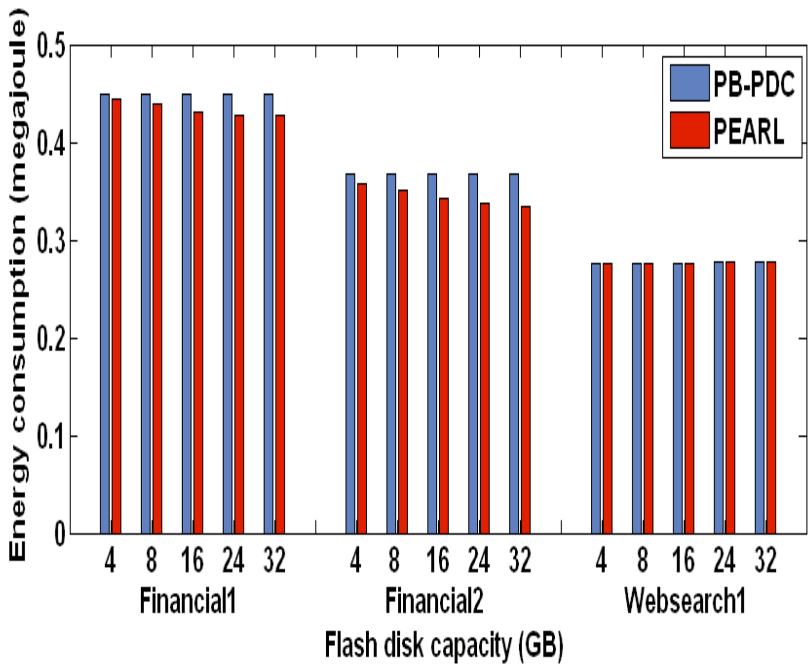
- Energy-Efficient Data Placement Algorithms
- Self-Adaptive and Reliability-Aware Data Redistribution Strategies
- Self-Triggered Data Replication Policies
- Automatic Data Prefetching Schemes

Preliminary Results of Task 3

- "Dynamic Data Reallocation in Hybrid Disk Arrays" [Xie and Sun, IEEE TPDS]
- "PEARL: Performance, Energy, and Reliability Balanced Dynamic Data Redistribution for Next Generation Disk Arrays" [Xie and Sun, MASCOTS'08]
- "SAIL: Self-Adaptive File Reallocation on Hybrid Disk Arrays" [Xie and Madathil, HiPC'08]

PEARL


Mean response time (millisecond)
PB-PDC
PEARL
Financial1
Financial2
Websearch1
Flash disk capacity (GB)


Energy consumption (megajoule)
PB-PDC
PEARL
Financial1
Financial2
Websearch1
Flash disk capacity (GB)

Task 4: Developing an energy-aware fault-tolerant mechanism for FIT

- Data reconstruction algorithms, which are executed in the presence of disk failure, for mobile storage systems must be reliability-aware, performance-driven and energy-efficient.

- We developed two novel reconstruction strategies that can be applied to mobile storage systems to noticeably save energy while providing shorter reconstruction times and user response times.

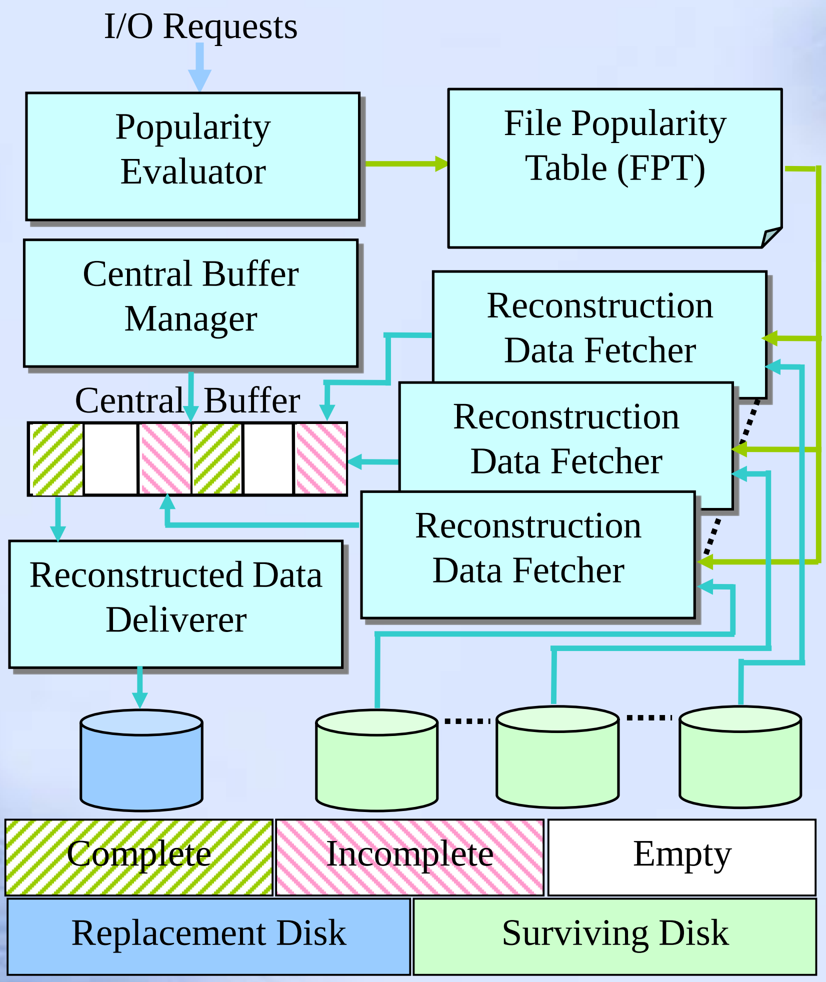
Preliminary Results of Task 4

- "MICRO: A Multi-level Caching-based Reconstruction Optimization for Mobile Storage Systems" [Xie and Wang, IEEE Transactions on Computers, October 2008]
- "Collaboration-Oriented Data Recovery for Mobile Disk Arrays" [Xie and Sharma, ICDCS'09]

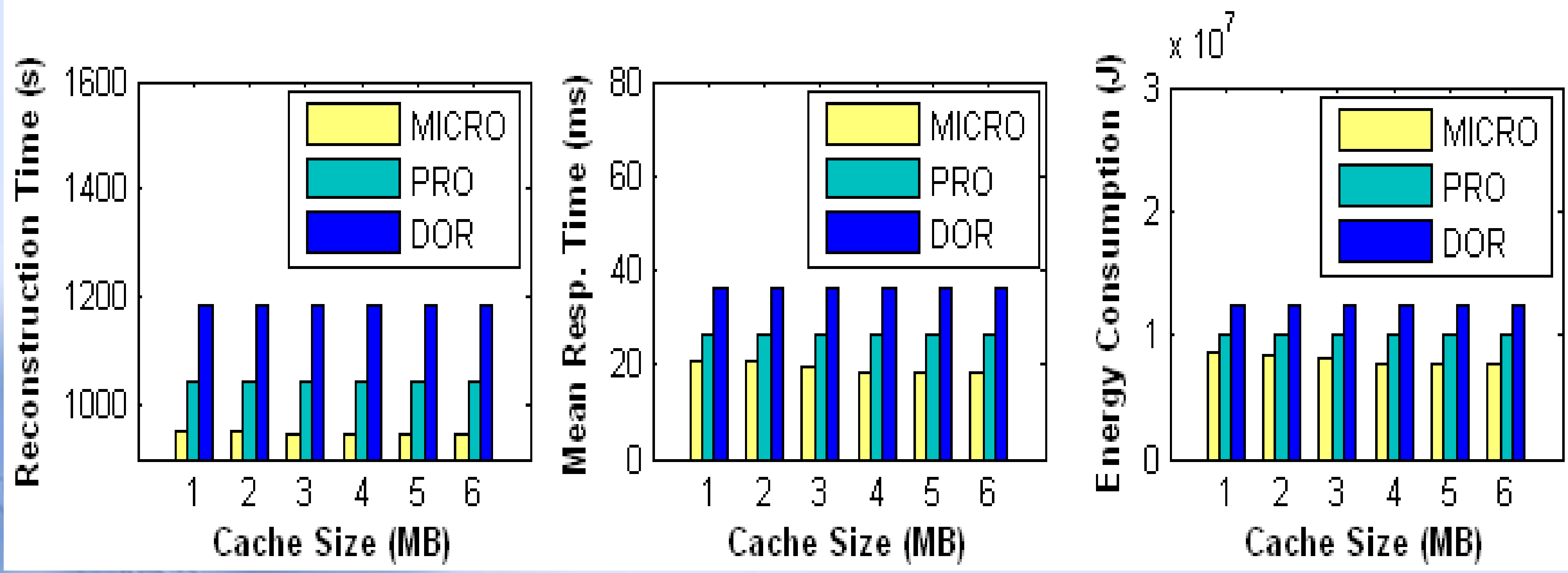
MICRO

- MICRO collaboratively utilizes storage cache and disk array controller cache to diminish the number of physical disk accesses caused by reconstruction.

Architecture of MICRO



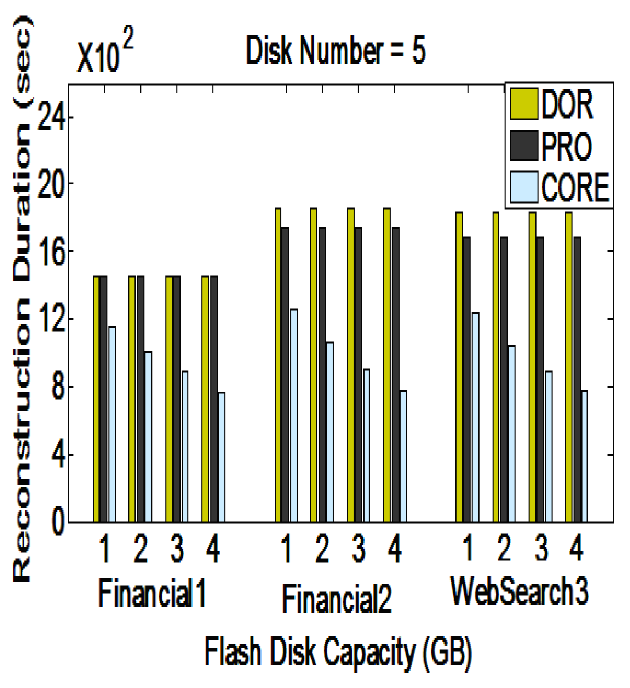
Experimental Results of MICRO

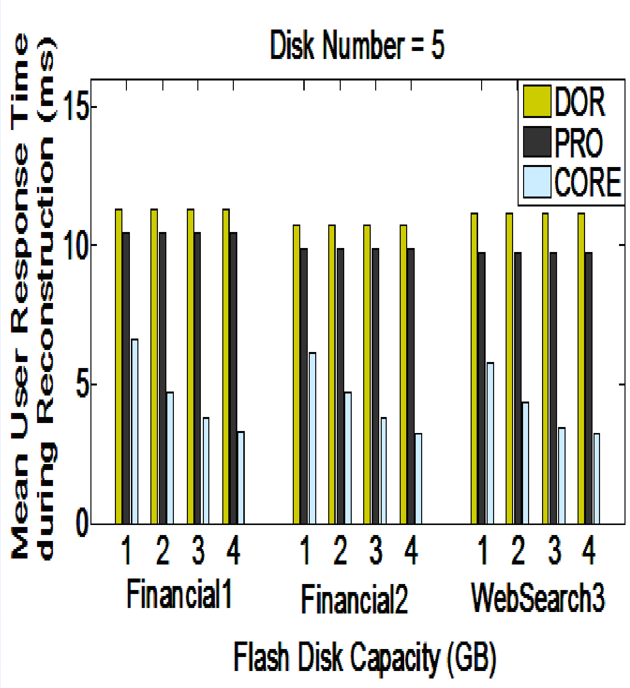


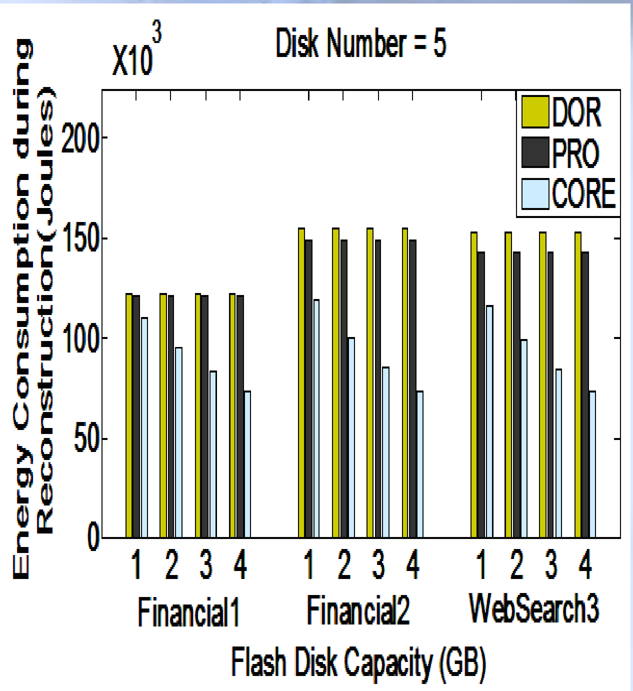
CORE

- We developed a flash assisted data reconstruction strategy called CORE (collaboration-oriented reconstruction) on top of a hybrid disk array architecture.

Experimental Results of CORE







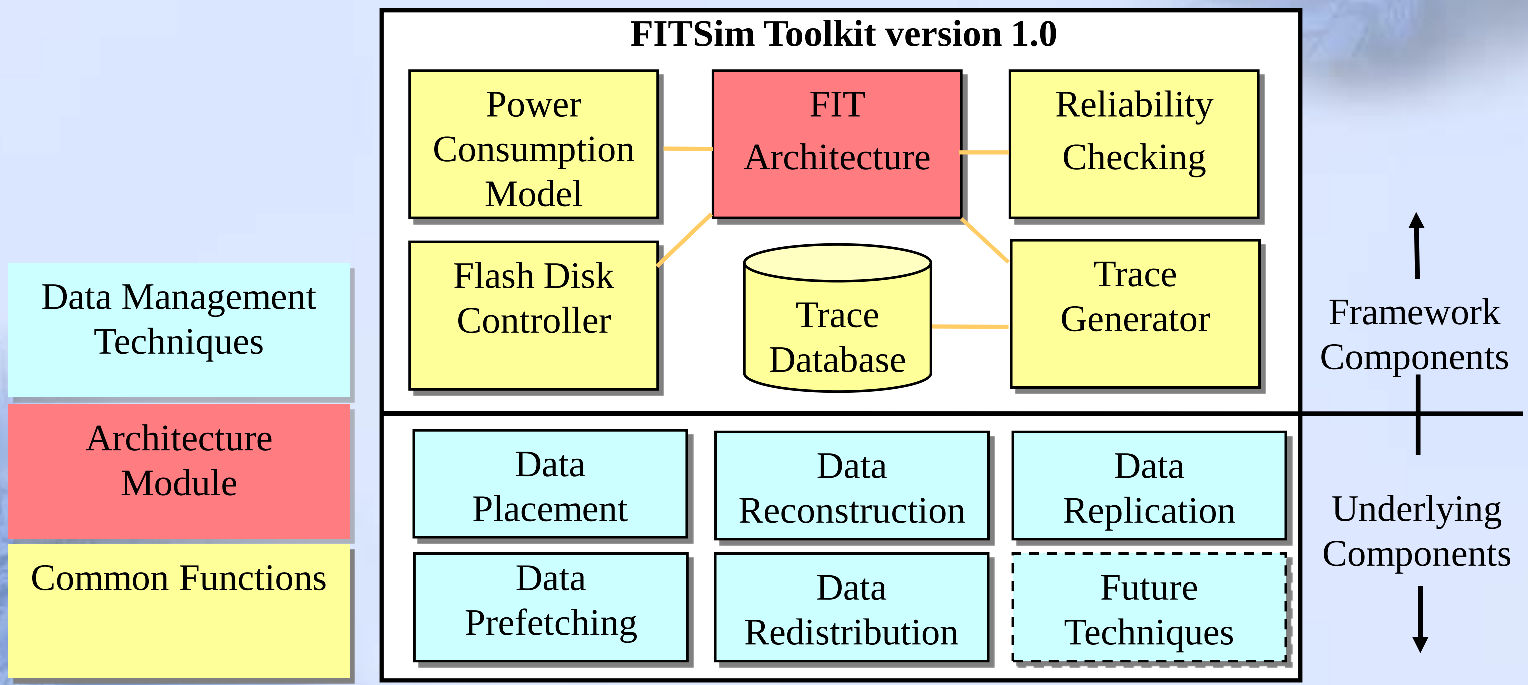
Research Task 5: a prototype and a simulation toolkit

- We will implement a FIT prototype, which will be the first of its kind.
- Although existing simulation tools can model disk arrays, they are inadequate for the modeling of a combination of energy-aware data management techniques for mobile disk arrays without adversely affecting system reliability.
- We will implement a simulation software toolkit called FITSim Toolkit, which consists of the FIT architecture, a disk reliability model, a power consumption model, a trace generator, and an array of energy-aware data management techniques.

The FIT Prototype

- Implement a FIT prototype to work on Linux
- Evaluate the FIT architecture and the new data management schemes’ performance and tradeoffs
- Address the issues that will arise during the realization of the FIT architecture

FITSim


FITSim Toolkit version 1.0
Power Consumption Model
FIT Architecture
Reliability Checking
Flash Disk Controller
Trace Database
Trace Generator
Framework Components
Data Placement
Data Reconstruction
Data Replication
Data Prefetching
Data Redistribution
Future Techniques
Underlying Components
Data Management Techniques
Architecture Module
Common Functions

Funded Projects

1. A Device-Array Based Flash Storage System for Emerging Data-Intensive and Mission-Critical Mobile Applications: from Architecture Redesign to New File System (PI, NSF CNS-1320738, $440,727, 10/2013 ~ 09/2016)
2. CAREER: Architectural Support for Integrating NAND Flash Solid State Disks into Enterprise-Class Storage Systems (PI, NSF CNS-0845105, $436,000, 09/2009~ 08/2014)
3. Energy-Efficient and Reliability-Aware Data Management in Mobile Storage Systems (PI, NSF CNS-0834466, $160,000, 09/2008 ~ 08/2010)

- "BUD: A Buffer-Disk Architecture for Energy Conservation in Parallel Disk Systems" (Co-PI, NSF CCF-0742187, among the $311,999 grant, $90,244 flows to SDSU, 05/2007 ~ 04/2010).

Questions?