The Premier Advanced Recording Technology Forum

THIC Inc.

NCAR

Simulating the NCAR Mass Storage System to Improve Performance

Bill Anderson

National Center for Atmospheric Research

1850 Table Mesa Dr., Boulder, CO, 80305-5602

Phone:+1-303-497-1243 FAX: +1-303-497-1848

E-mail: andersnb@ucar.edu

Presented at the THIC Meeting at the National Center for Atmospheric Research, 1850 Table Mesa Drive, Boulder CO 80305-5602

June 29-30, 2004

Introduction

- NCAR has a 1.9 PB Mass Storage System.
- The size and complexity of the hardware and software can make it difficult to estimate the effects of system configuration changes on performance.
- A trace-driven performance simulator was built to aid us in ranking design and configuration alternatives.

Copyright© 2004 University Corporation for Atmospheric Research

Overview

- NCAR & the NCAR MSS
- High Level View of the Simulator
- Lower Level View of the Simulator
- Validation
- Limitations
- Example of Simulator Use
- Future Uses

Copyright© 2004 University Corporation for Atmospheric Research

National Center for Atmospheric Research

Copyright© 2004 University Corporation for Atmospheric Research

The NCAR MSS

- 1.9 PBs total, 1.1 PBs of unique data (as of 1 June 2004)
- 23 M files
- Average net growth rate ~1.7 TB per day
- Average 16,000 reads per day and 22,000 writes per day
- Average 4500 tape mounts per day (mostly robotic)

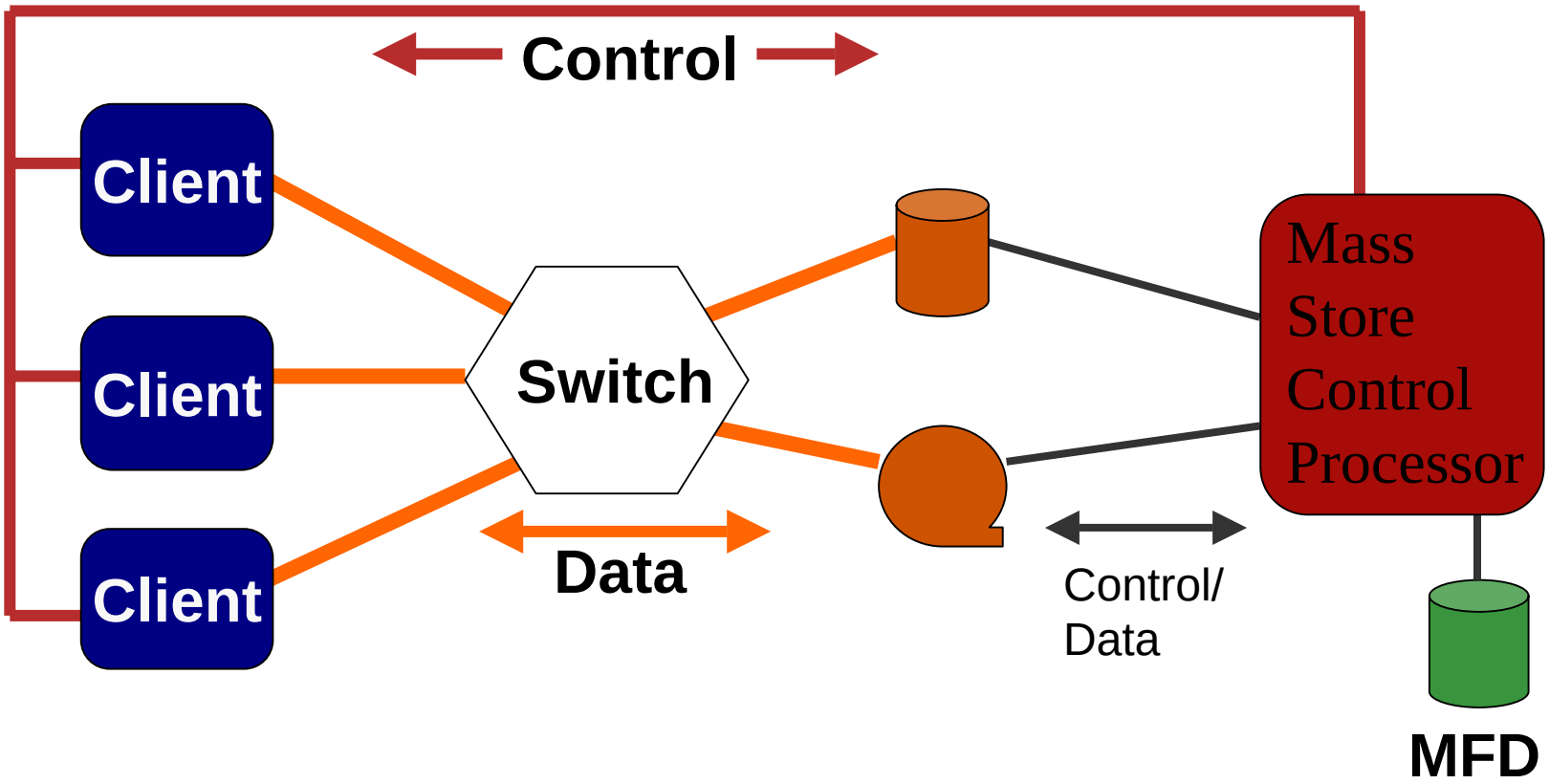
Copyright© 2004 University Corporation for Atmospheric Research

The NCAR MSS

- 5 StorageTek 9310 Tape Silos plus an offline (manual mount) archive
- 20 9940B FC tape drives, 38 9940A Escon tape drives and 24 9840A Escon tape drives
- 7 TB of disk cache
- rcp model – moves entire files to/from host at user request
- IP and Hippi are used for the data paths.
- In the future, all data transfers will occur over IP networks through custom software to FC devices.

Copyright© 2004 University Corporation for Atmospheric Research

Data Flow in the NCAR MSS



Copyright© 2004 University Corporation for Atmospheric Research

Configuring the MSS

- System complexity and component interdependence can make it hard to estimate the effects of system changes.
- Can be difficult and expensive to experiment with a production MSS.
- Analytic approaches (e.g., queueing theory) often have low accuracy.
- Simulations are generally more accurate than analytic approaches and easier than experimenting with a production MSS.

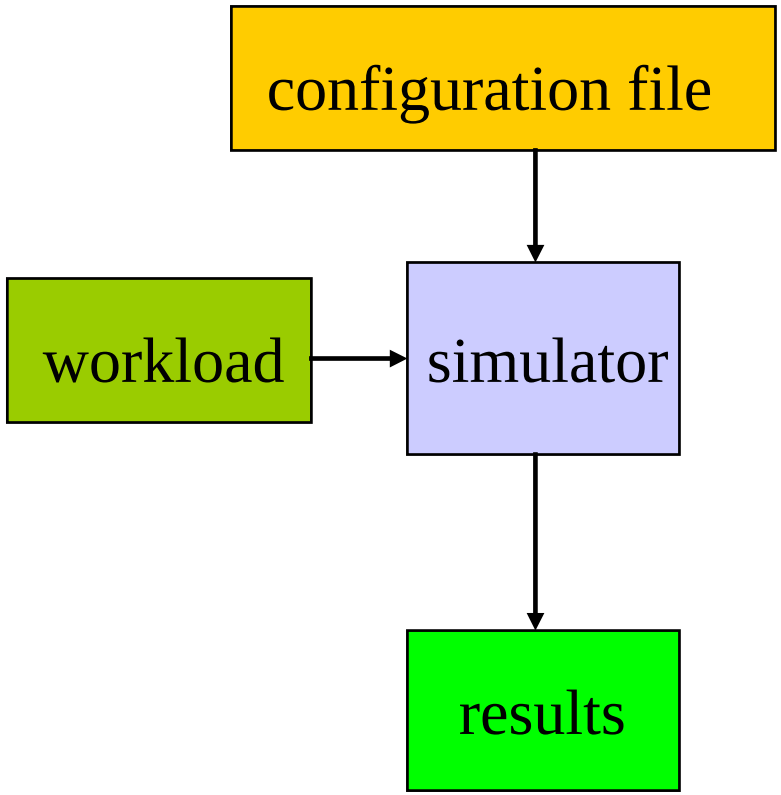
Copyright© 2004 University Corporation for Atmospheric Research

Goal of Simulator

- main goal: estimate the ***average*** response time and other metrics for user file transfers over time periods of a month or more to within about 15% percent of the true values.
- Of the set of performance metrics, users are probably most aware of response time.
- Simulator only provides information about performance related metrics.

Copyright© 2004 University Corporation for Atmospheric Research

High Level View of Simulator



- Trace driven using logs
- discrete-event simulator
- Hardware & software components are simulated
- Simulator written in Java
- ASCII configuration file used
- Calculates metrics such as response time, # of tape mounts, device utilization

Copyright© 2004 University Corporation for Atmospheric Research

Components in the Simulator

- Tape drives
- Silos
- Disk subsystems
- Numerous software components

Copyright© 2004 University Corporation for Atmospheric Research

Simulating a Component

- First, build a conceptual model of a component such as a tape drive.
- Information was obtained by talking with other group members and vendors, by reading documentation and source code and by running tests.
- Components are modeled with the least amount of detail consistent with the desired accuracy.
- Many details can be ignored.
- Next, implement the model in software.

Copyright© 2004 University Corporation for Atmospheric Research

Simulating a Component

- Test
- Refine and add more detail as necessary to achieve desired accuracy
- Each component that takes time to perform operations is modeled using a separate thread.
- Example…

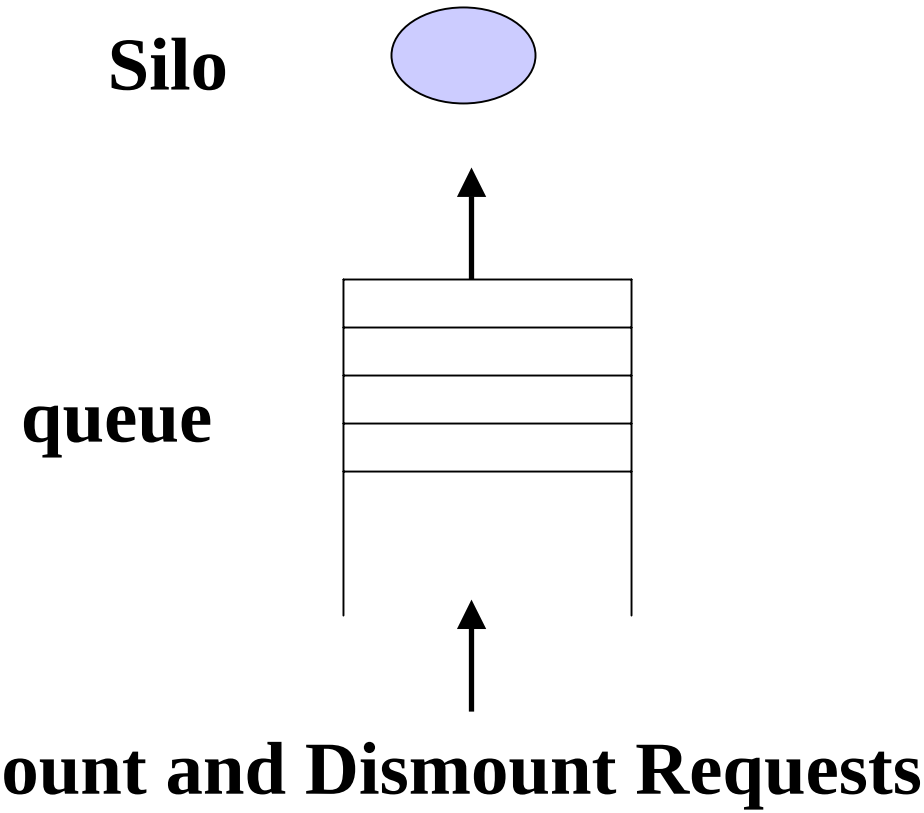
Copyright© 2004 University Corporation for Atmospheric Research

Silo Example

- Consider a single silo.
- It primarily receives requests to mount or dismount tapes and can queue up requests.
- If a request arrives while a silo is currently mounting or dismounting a tape, the request waits in a queue.
- The time to actually service a mount request is a function of drive location, cell location and robot speed.
- Average service times can be measured and obtained from vendor.

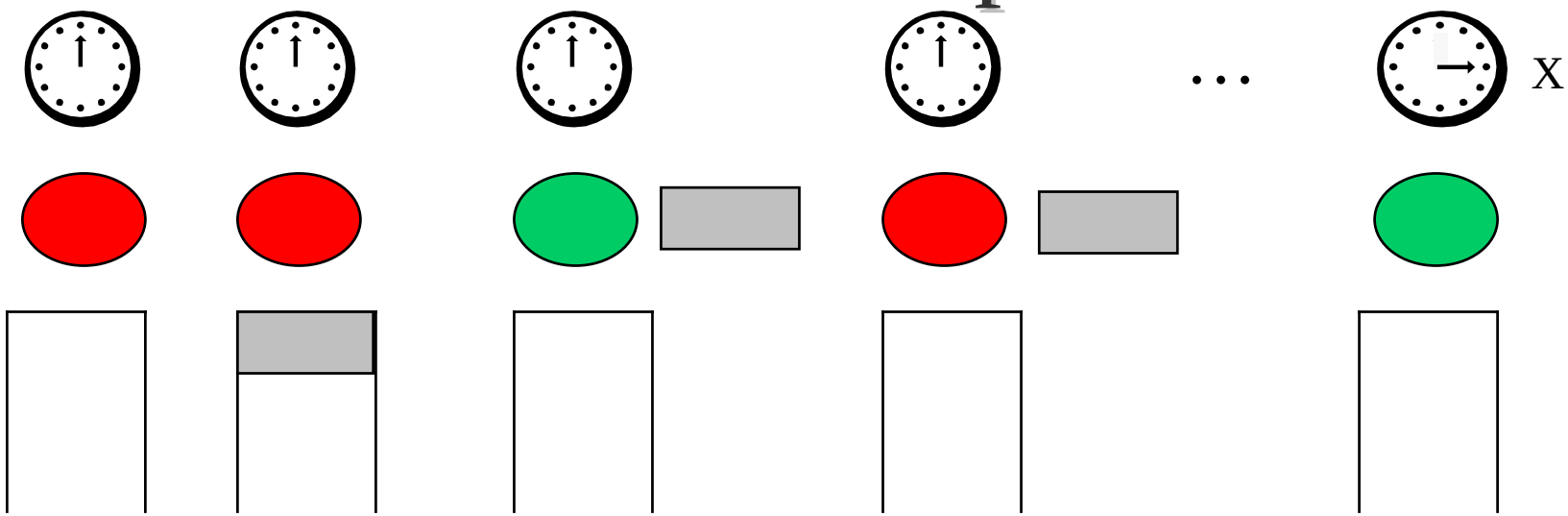
Copyright© 2004 University Corporation for Atmospheric Research

Silo Example



Copyright© 2004 University Corporation for Atmospheric Research

Silo Example



IDLE

REQUEST ARRIVES

THREAD IS RESUMED

THREAD IS SUSPENDED UNTIL TIME X

THREAD IS RESUMED AND RETURNS SUCCESS

Copyright© 2004 University Corporation for Atmospheric Research

Silo Example

- Many details of a real silo were ignored.
- As in real system, if request arrival rate exceeds rate at which requests can be processed, queue length grows.
- Example is a typical server with a queue.
- Example is representative of how other servers are simulated. Each one is a thread and typically has a queue.

Copyright© 2004 University Corporation for Atmospheric Research

Estimating Delay Parameters

- Once code was implemented, service time parameters had to be estimated.
- Examples include time to mount a tape, time to load a tape and I/O transfer rates.
- Some parameters are constants and others vary.
- Estimates obtained from multiple sources.
- Parameters are configurable.

Copyright© 2004 University Corporation for Atmospheric Research

Validation

- Checking that the predicted results closely match what would be observed in a real system.
- All components were individually validated and the simulator as a whole was also validated.
- Dozens of validation runs were performed for simplified cases where the exact answers are known and cases where simulator was configured like the real MSS.
- Validation is an on-going process.

Copyright© 2004 University Corporation for Atmospheric Research

Running the Simulator

- Simulator is configured with a primitive ASCII configuration file.
- It is currently run on an IBM Power4 system.
- It takes about 24 wall clock hours and 7 GB of memory for a 6 month simulation.

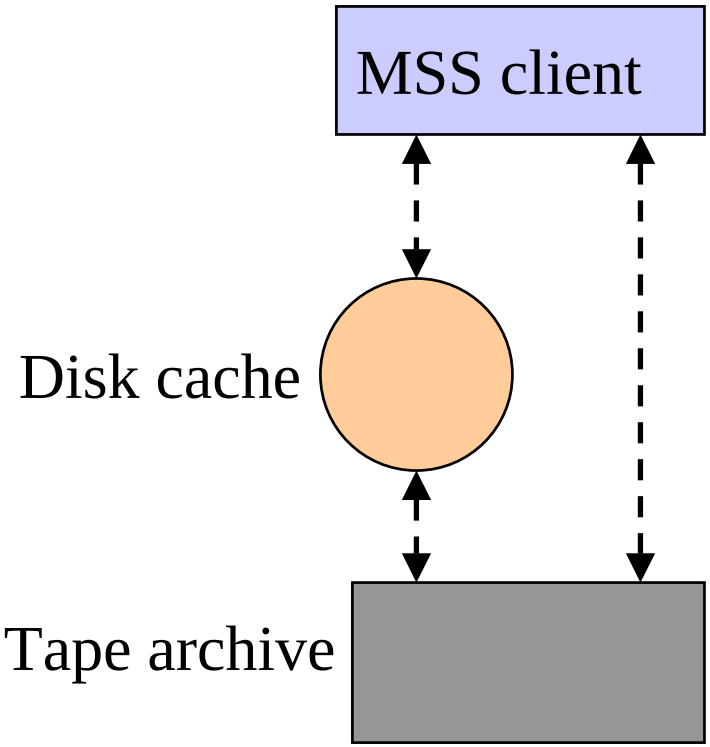
Copyright© 2004 University Corporation for Atmospheric Research

Limitations

- Simulator cannot predict the workload.
- All approaches to configuring the MSS are limited in the same way.
- Fortunately, workload is fairly well behaved.
- Not all components of the MSS currently taken into account.
- Metrics are averages and have error bounds of about 15%.

Copyright© 2004 University Corporation for Atmospheric Research

Example of Simulator Use



- Simulator was used to study benefits of expanding the disk cache.
- Read hit ratio was primary metric.
- Cache sizes and migration and staging policies were varied.

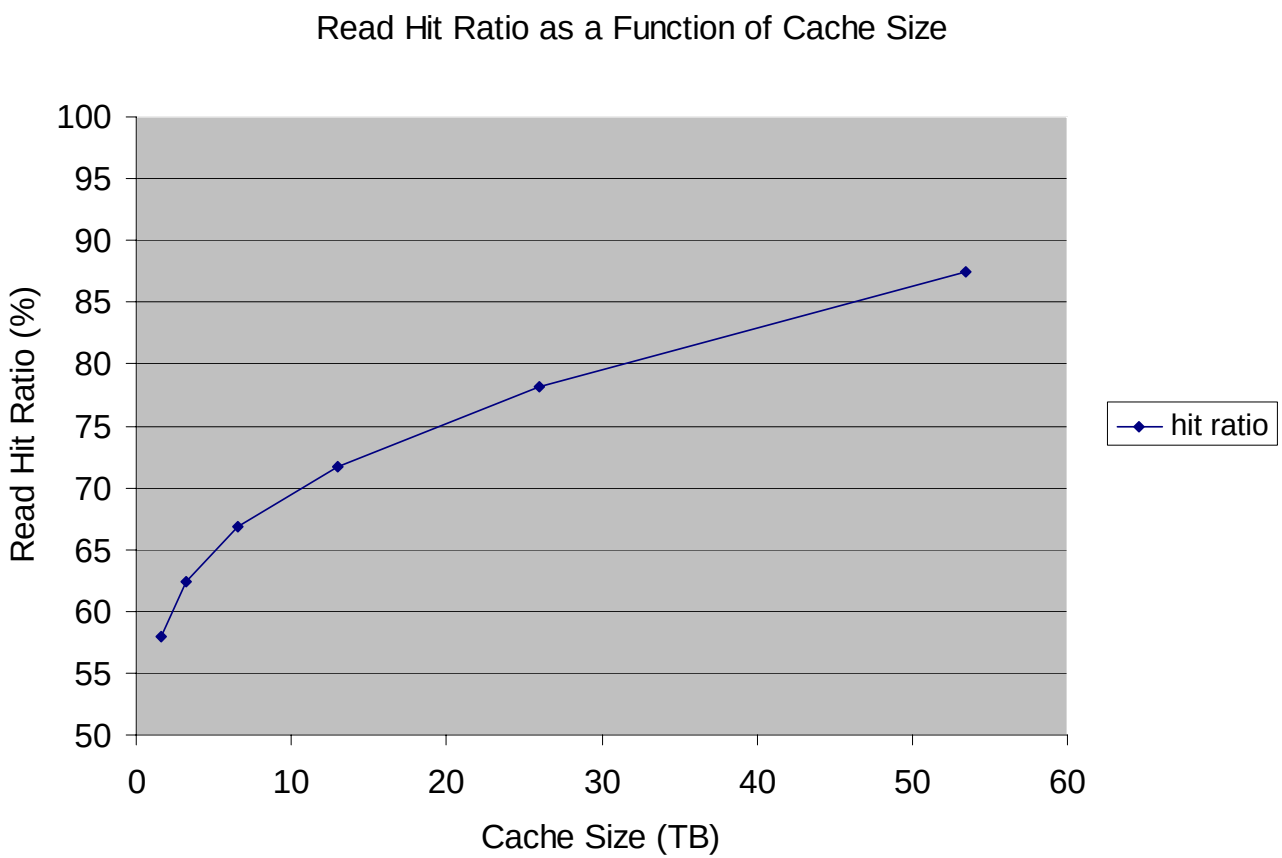
Copyright© 2004 University Corporation for Atmospheric Research

Example of Simulator Use

- Simulation runs showed that there was a fair number of files that were read back within about 30 days of being written.
- Simulator was then used to help size a disk cache to offload reads from the tapes and provide faster response time to users.
- Estimated read hit ratios were around 60%.

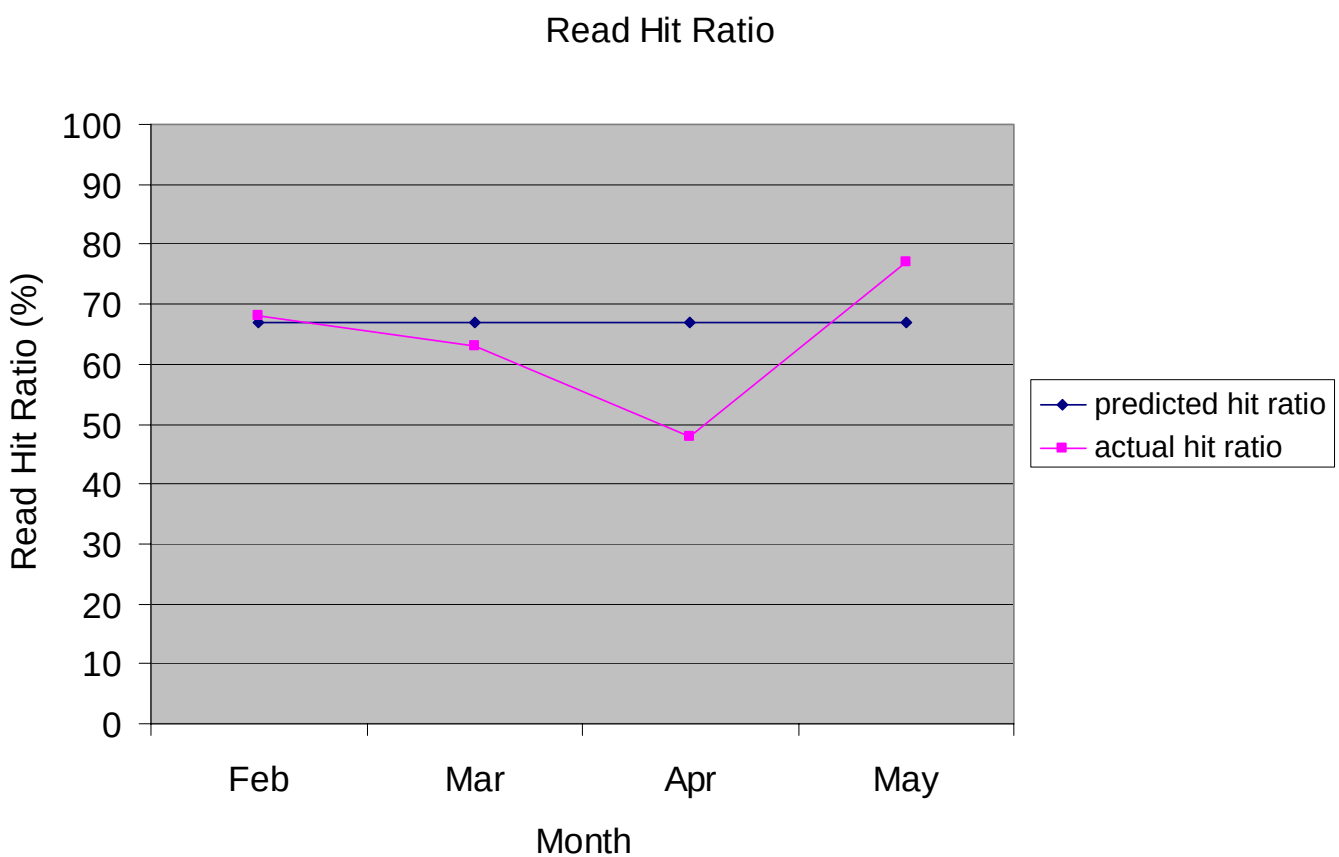
Copyright© 2004 University Corporation for Atmospheric Research

Predicted Read Hit Ratio



Copyright© 2004 University Corporation for Atmospheric Research

Actual Read Hit Ratio



Copyright© 2004 University Corporation for Atmospheric Research

Future Uses of Simulator

- Study the effect of tape throughput density (ratio of transfer rate to capacity) on MSS performance
- Experiment with synthetic workloads
- Study what the most cost effective way to improve MSS performance might be

Copyright© 2004 University Corporation for Atmospheric Research

Questions?

Copyright© 2004 University Corporation for Atmospheric Research