Mass Storage System Performance Prediction Using a Trace Driven Simulator

Bill Anderson
National Center for Atmospheric Research
Boulder, CO

Acknowledgements

- Gene Harano
- Mark Love
- John Merrill
- Craig Ruff
- Erich Thanhardt
- George Williams

Copyright© 2005 University Corporation for Atmospheric Research

Introduction

- NCAR has a 2 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© 2005 University Corporation for Atmospheric Research

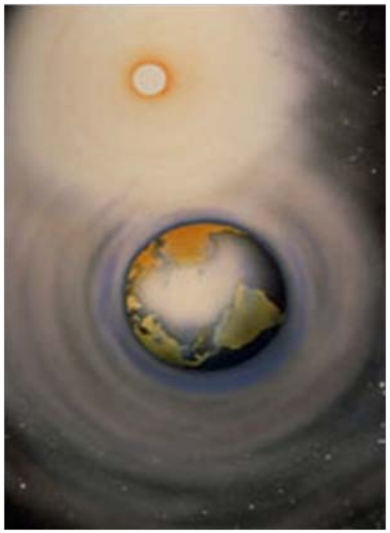
Overview of Talk

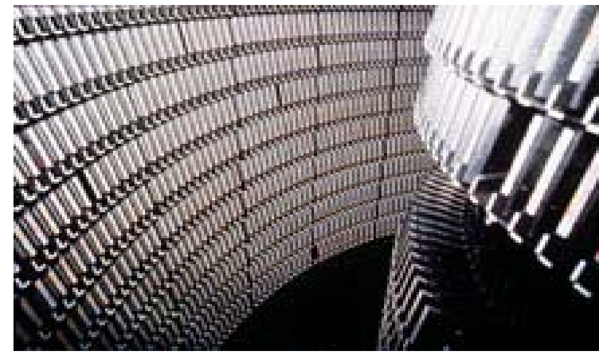
- NCAR & the NCAR MSS
- Simulator Functionality
- Simulator Results

Copyright© 2005 University Corporation for Atmospheric Research

NCAR

National Center for Atmospheric Research





Copyright© 2005 University Corporation for Atmospheric Research

The NCAR MSS

- 2.2 PBs total, 1.3 PBs of unique data (as of 1 Feb 2005)
- 26 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

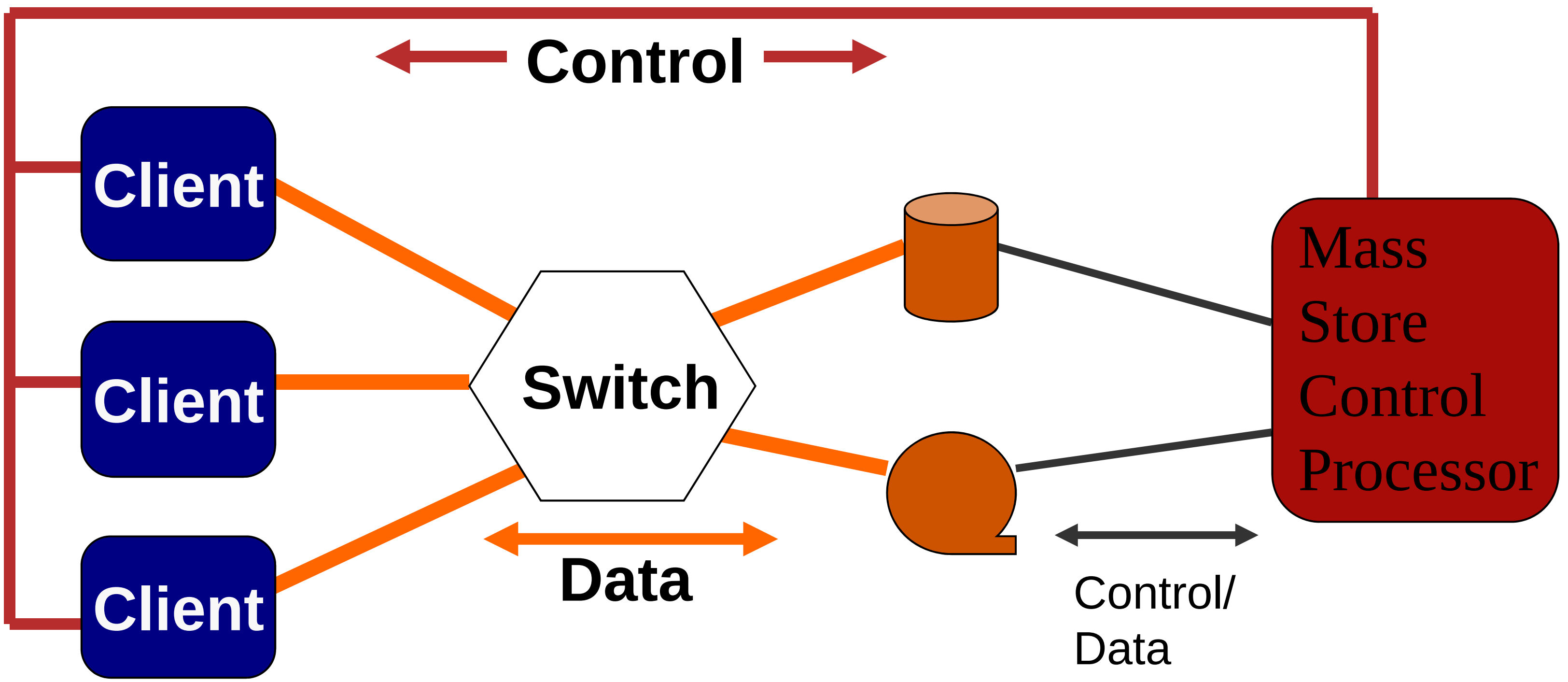
Copyright© 2005 University Corporation for Atmospheric Research

The NCAR MSS

- 5 StorageTek 9310 Tape Silos plus an offline (manual mount) archive
- 40 9940B FC tape drives, 38 9940A Escon tape drives and 14 9840A Escon tape drives
- 10 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 near future, all data transfers will occur over IP networks to FC devices.

Copyright© 2005 University Corporation for Atmospheric Research

Data Flow in the NCAR MSS



Copyright© 2005 University Corporation for Atmospheric Research

Configuring the MSS

- System complexity and component interdependence can make it hard to estimate the effects of system changes.
- We can measure current disk cache hit ratio, but would a larger cache lead to a commensurate improvement in hit ratio?
- Would more tape drives improve user response time significantly?

Copyright© 2005 University Corporation for Atmospheric Research

Configuring the MSS

- Can be difficult and expensive to experiment with a production MSS.
- Analytic approaches (e.g., queueing theory) have lower accuracy than we wanted.
- Simulations can be more accurate than analytic models and easier than experimenting with a production MSS.

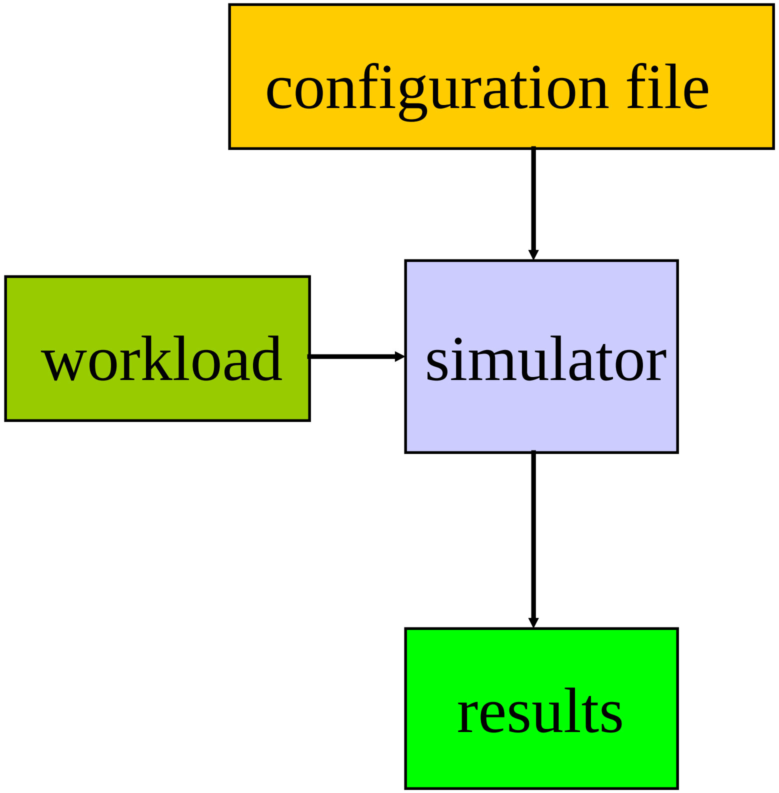
Copyright© 2005 University Corporation for Atmospheric Research

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 20% 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© 2005 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© 2005 University Corporation for Atmospheric Research

Components in the Simulator

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

Copyright© 2005 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.

Copyright© 2005 University Corporation for Atmospheric Research

Simulating a Component

- Next, implement the model in software.
- Test
- Refine and add more detail as necessary to achieve desired accuracy

Copyright© 2005 University Corporation for Atmospheric Research

Estimating Delay Parameters

- Examples include time to mount a tape, time to load a tape and I/O transfer rates.
- Deterministic approach used for some parameters (e.g., tape load time).
- Some delays may be too difficult to determine accurately at a given point in time.
- In those cases, a probabilistic approach can be useful.

Copyright© 2005 University Corporation for Atmospheric Research

Estimating Delay Parameters

- For example, tape positioning times for reads can vary a lot and be relatively large.
- If file positions on tape were modeled, could be calculated deterministically.
- We use a probability distribution to model that delay.
- Parameter values (deterministic or probabilistic) were obtained from multiple sources and are configurable.

Copyright© 2005 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© 2005 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 host.
- It takes about 24 wall clock hours and 7 GB of memory for a 6 month simulation.

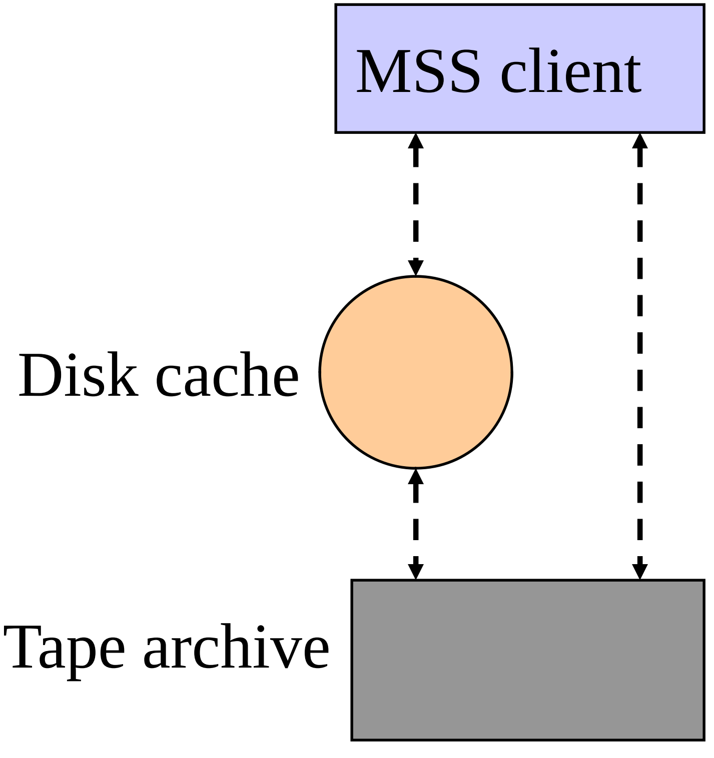
Copyright© 2005 University Corporation for Atmospheric Research

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 20%.

Copyright© 2005 University Corporation for Atmospheric Research

Simulator Results: Disk Cache Study



- 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.

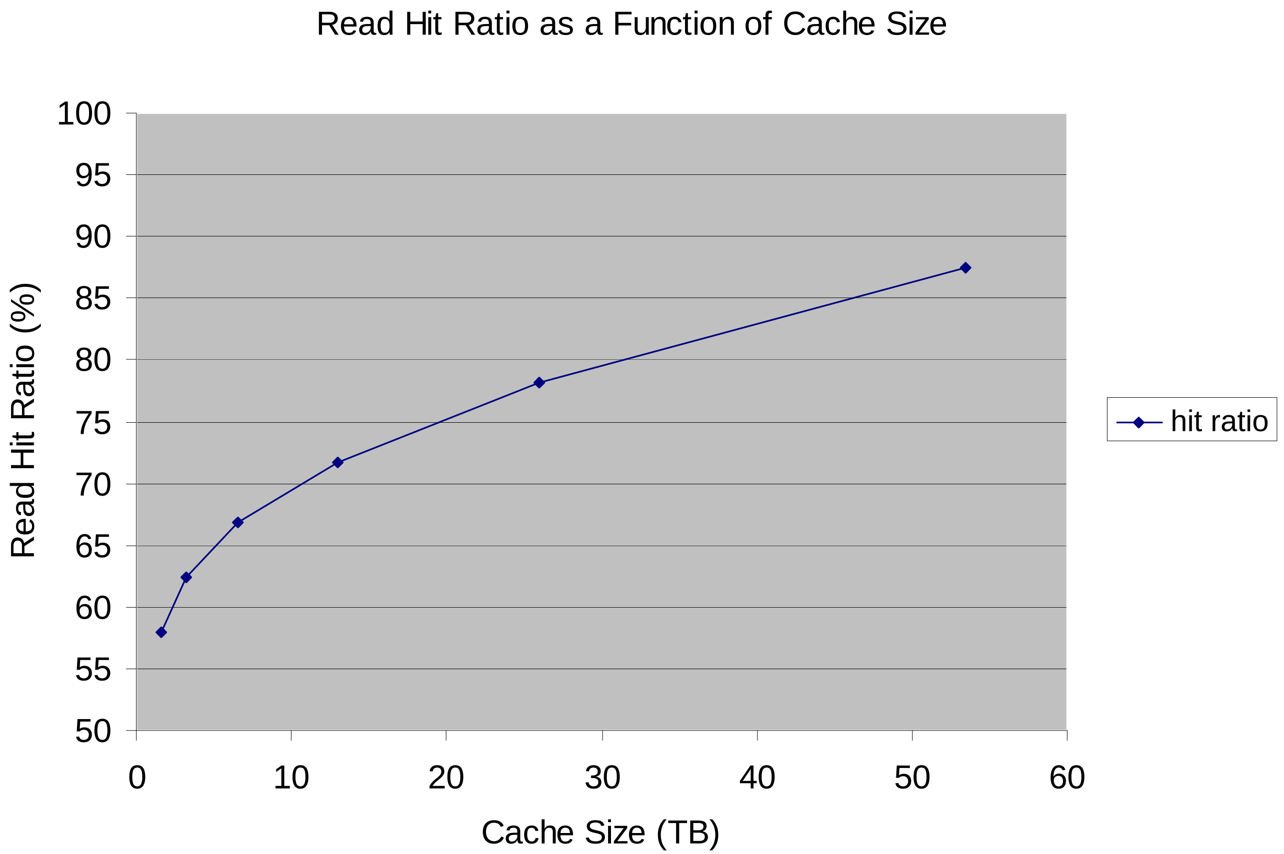
Copyright© 2005 University Corporation for Atmospheric Research

Disk Cache Study

- Simulation runs showed that there was a large 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%.

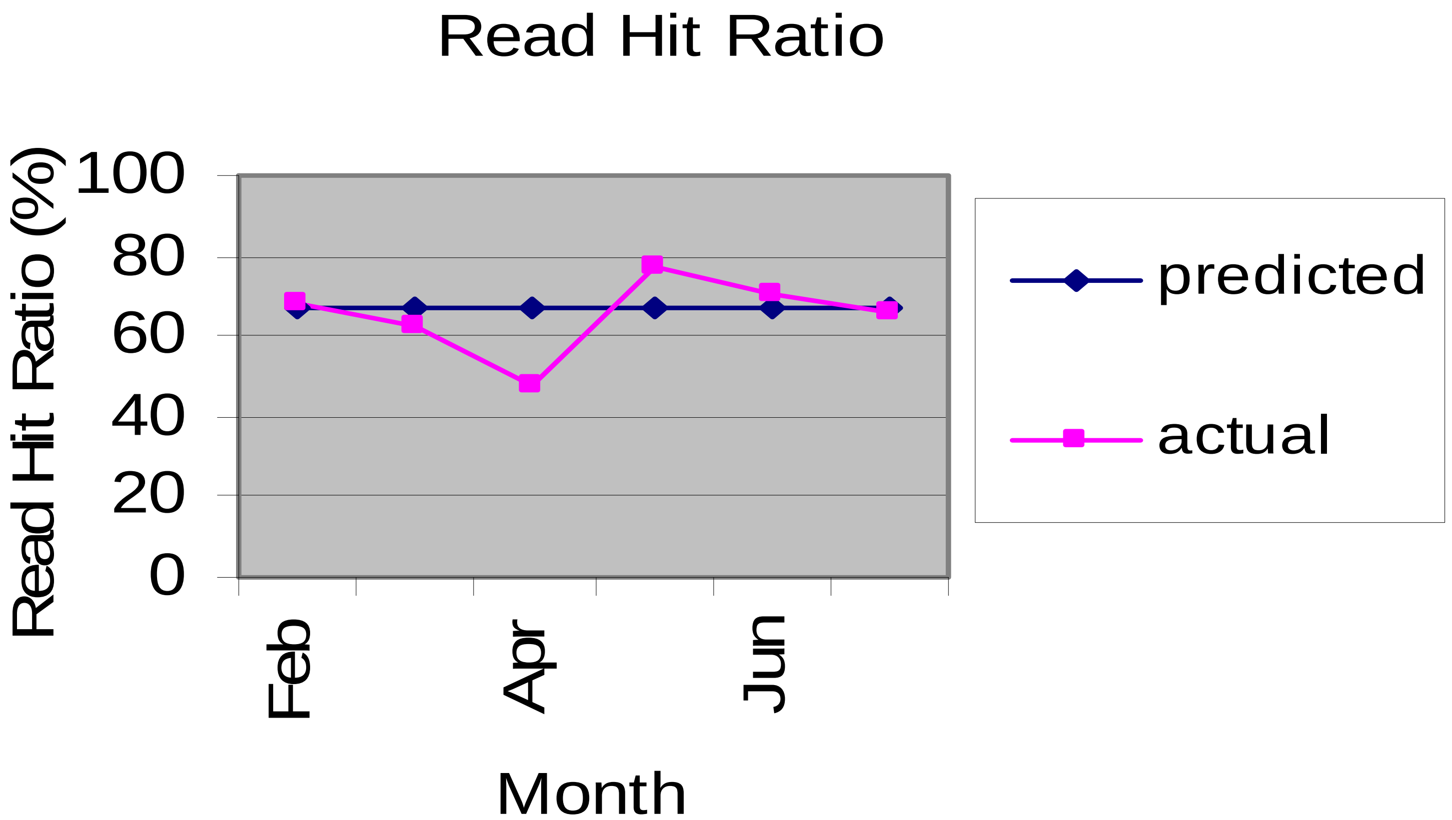
Copyright© 2005 University Corporation for Atmospheric Research

Predicted Read Hit Ratio for Files <= 50 MB



Copyright© 2005 University Corporation for Atmospheric Research

Actual Read Hit Ratio



Copyright© 2005 University Corporation for Atmospheric Research

Simulator Results: Disk and Tape Drive Study

- Purpose of this study was to estimate user response time as a function of both disk cache configuration and the number of STK 9940-B tape drives.
- Metric was user response time.
- Simulator was configured similar to our actual system. Only disk cache configuration and number of 9940-B drives were varied.

Copyright© 2005 University Corporation for Atmospheric Research

Disk and Tape Drive Study

- Three different disk cache configurations were tried:
  1. 7.8 TB cache with a max cache file size of 50 MB
  2. 36.1 TB cache with a max cache file size of 500 MB
  3. 62.6 TB cache with an unlimited cache file size
- Writes of files that met the cache criteria (file size, etc.) were written directly to the cache; otherwise they were written directly to tape.
- Reads of files that resided in the cache were read from there; otherwise they were read from tape.
- Files were aged off of the cache based on LRU

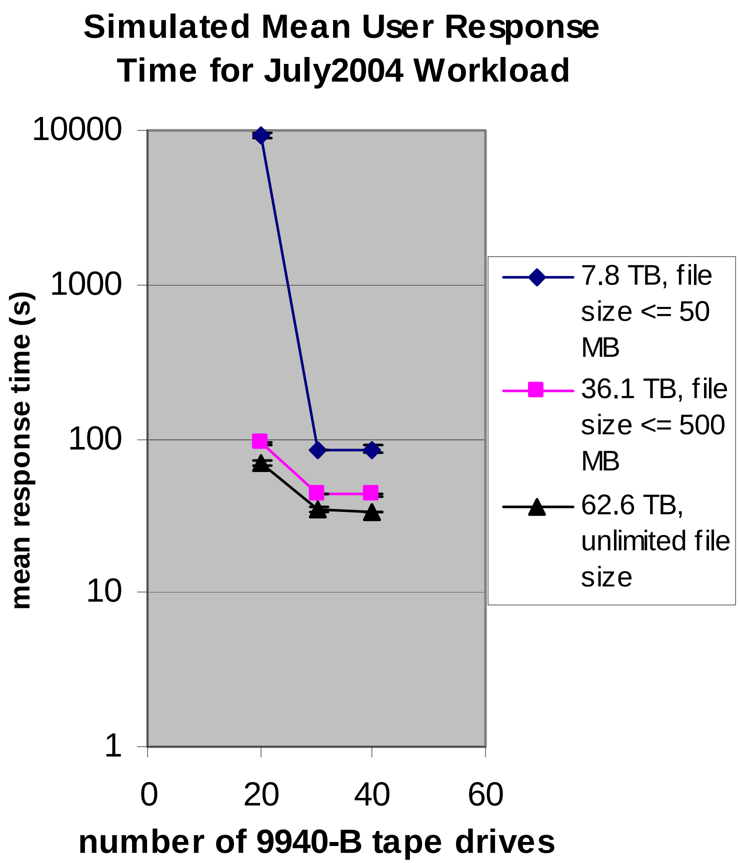
Copyright© 2005 University Corporation for Atmospheric Research

Disk and Tape Drive Study

- The number of 9940-B drives tried were 20, 30, and 40.
- All new writes to tape went to 9940-B drives; reads were serviced by 9940-B, 9940-A or 9840-A
- Some parameter information:
  - Cache had an aggregate b/w of 180 MB/sec
  - 9940-B load/unload time was 18 seconds
  - 9940-B position times based on prob. distributions and constants
  - 9940-B max transfer rate 30 MB/sec

Copyright© 2005 University Corporation for Atmospheric Research

Results



- Caching files beyond 500 MB may not be worth it
- A balance of tape drives and disk cache seems better than an extreme of either
- Plot illustrates non-linear behavior of response time

Copyright© 2005 University Corporation for Atmospheric Research

Limitations of Study

- We did not investigate the effects of workload changes that might be induced by system configuration changes.
- We did not study the economic costs of the various scenarios.
- Would also be interesting to experiment with different parameter values.

Copyright© 2005 University Corporation for Atmospheric Research

Conclusions

- A trace-driven performance simulator has been developed to aid us in ranking design and configuration alternatives.
- We found that a modest sized cache could provide a reasonable read hit ratio.
- We also used it to estimate the number of 9940-B tape drives that we would initially need.

Copyright© 2005 University Corporation for Atmospheric Research

Questions?

Copyright© 2005 University Corporation for Atmospheric Research