Measuring Tape Drive Performance

Theodore Johnson
AT&T Labs - Research
johnsont@research.att.com

Publications

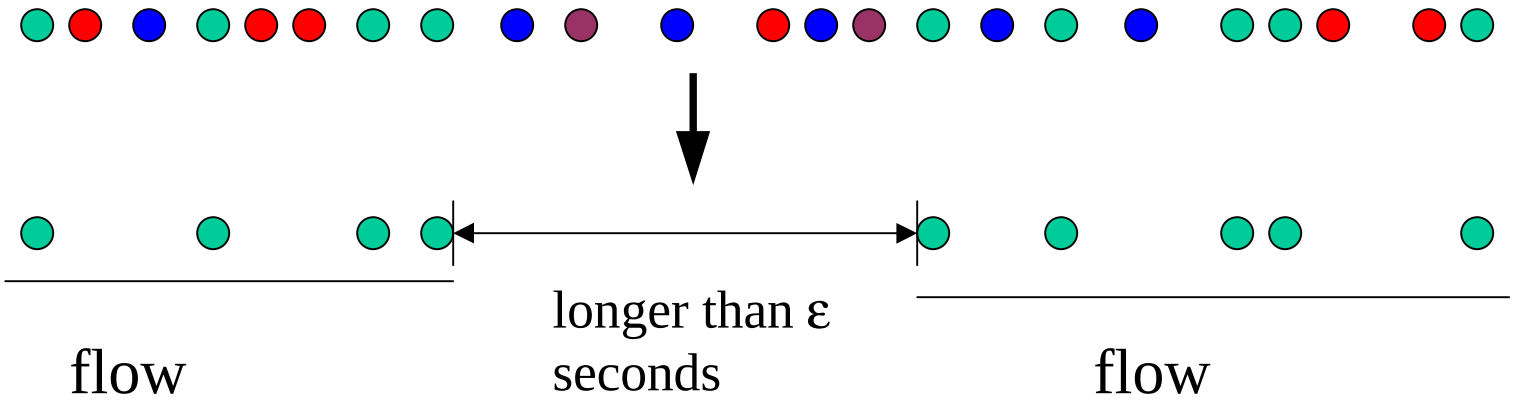
- *Benchmarking Tape System Performance*, T. Johnson , E. Miller, Joint IEEE Symposium on Mass Storage Systems / NASA GSFC Conf. On Mass Storage Systems and Technologies, pg. 95 - 112, 1998
- *Performance Measurements of Tertiary Storage Devices*, T. Johnson, E. Miller, Proc. 24th Intl. Conf. on Very Large Data Bases, pg. 50-61, 1998

Background

- My research at AT&T is on very large data warehouses.
  - Currently, IP network traffic.
- My research area is databases and performance modeling.
- Databases on tape
  - Multimedia
  - Scientific databases
  - Data warehouses
  - Write-once, read-never
- Current project
  - Store terabytes of network traffic data
  - Use it for business and scientific analyses
  - technical challenges
    - complex aggregation queries
    - use inexpensive tertiary storage

Complex Aggregation

- Complex aggregation
  - More interesting than sum, average, min, max, count
- Observation
  - Most queries can be computed with sequential scans
    - EMF query language
  - Sequential aggregates common in network analysis
  - Example : computing “IP flows”



Motivation

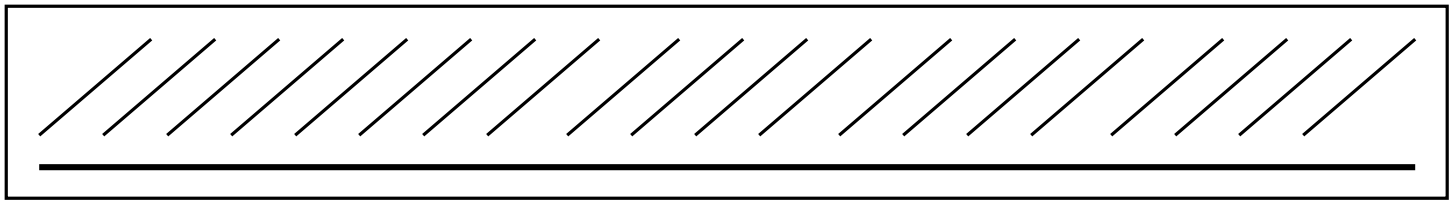
- Performance Modeling
  - Develop analytical or simulation models of tertiary storage system behavior.
- Data Warehouse Architecture Design
  - Is my approach feasible?
  - How should I design the system to obtain the best performance?
- Data Storage Optimization
  - What data do I put on tape?
  - How should I partition it?
  - Where should I put the partitions?

Why Measure Tape Performance?

- Many uses in database systems
  - Backup, scientific databases, multimedia, data warehouses
- The performance of tape drives is unusual
  - Inherently sequential
  - Many performance quirks.
- The performance of tape drives (and robotic storage libraries) has not been systematically studied.
- The lack of basic information inhibits my work.

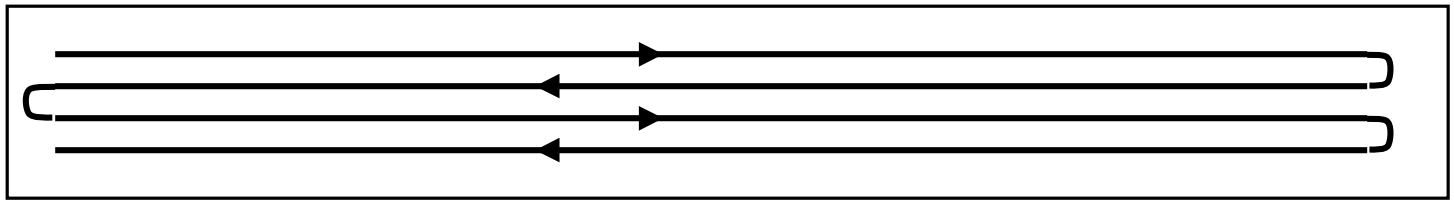
Tape drive characteristics

Serpentine layout



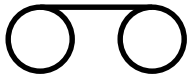
The data tracks are written diagonally using a rotating head.
A directory track can be used to speed up seeks.

Serpentine (linear) layout



Data tracks are written linearly, in alternating directions.
Fast seeks are accomplished by moving the tape head to
a new track.

Tape drive characteristics (cont.)

- Tape package
  - cassette 
    - might not need to rewind before eject
  - cartridge 
    - must rewind before eject.
- Compression
- Block sizes
  - fixed or variable
- Cost, size, reliability, etc.

Measurements

- Measure all aspects of a data request response time.
  - Robot fetch, mount time, seek time, transfer rate, rewind time, rewind time, unmount, media return time.
- Measure special characteristics
  - Short seeks, seek positioning hints, unmount without rewind, compression rates
- Measure factors affecting data transfer rates
  - block size, compression, delays in I/O requests.
- Be consistent
  - Use `mtio` calls.
  - Consistency not always possible (esp. when measuring robot performance and mount/unmount times).

Devices measured

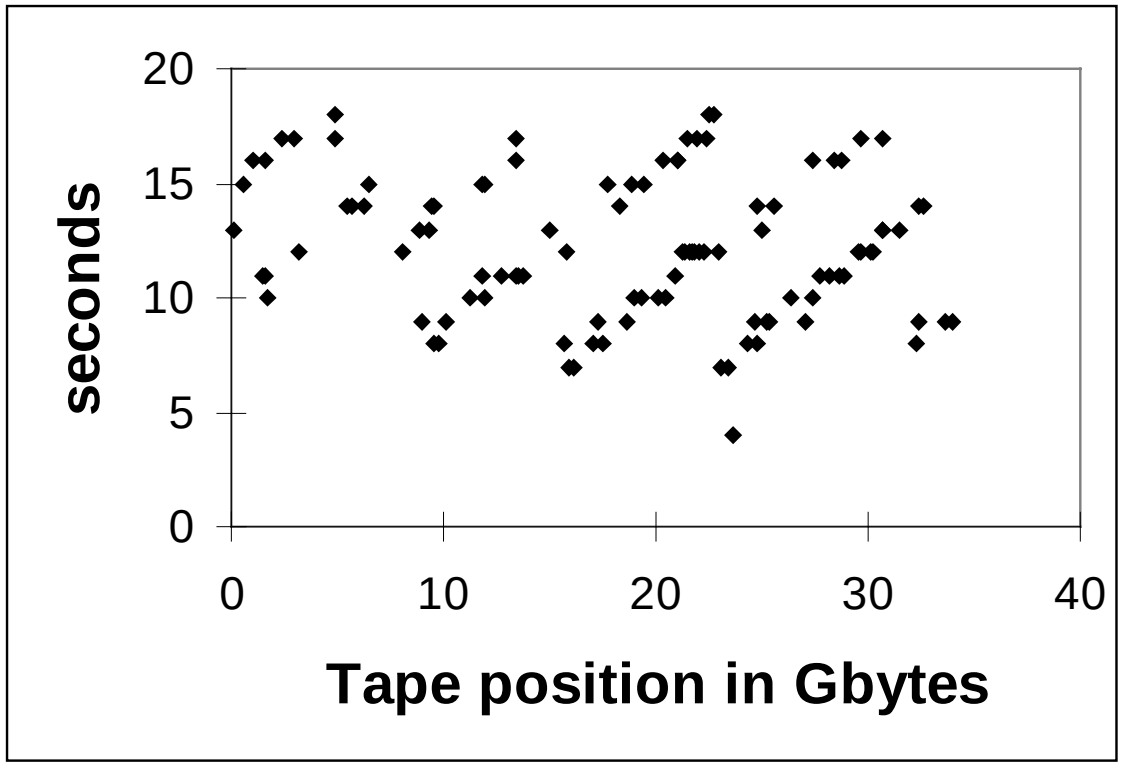
- Robotic storage libraries
  - GRAU ABBA/2 (6000 tapes)
  - Storagetek 9710 (404 tape, expandable to 588)
  - Ampex 810 (256 tape)
  - Sony DMS-B9 (9 tape)
- Tape drives
  - 4mm (helical scan cassette, low cost, low performance)
  - Ampex DST 310 (helical scan cassette, high cost, high performance)
  - Sony DTF (helical scan cassette, high cost, high performance)
  - DLT 4000 (serpentine cartridge, low cost, low performance)
  - DLT 7000 (serpentine cartridge, medium cost, medium performance)
  - IBM 3590 (serpentine cartridge, high cost, high performance)
    - Thanks to Ethan Miller.

Robotic storage library performance

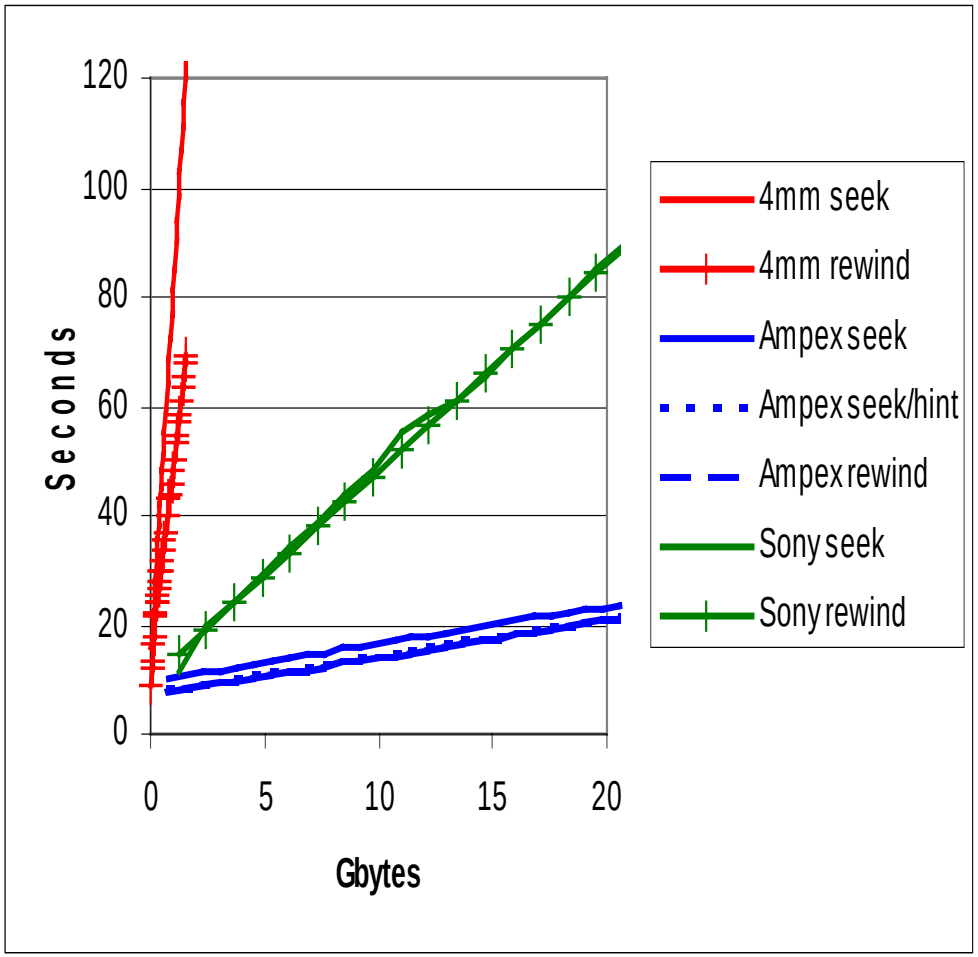
Robot name	Avg. fetch time	Std. Dev. Fetch time	Avg. return time	Std. Dev. Return
GRAU ABBA (6000 tapes)	19.7 Sec.	.1 sec	16.4 sec	.2
Storagetek 9710 (404 tapes)	Approx. 9	small	Approx. 9	small
Ampex 810 (256 tapes)	2.9	.3	3.7	.5
Sony DMS-B9 (9 tapes)	12.8	2.1	17.2	1.9

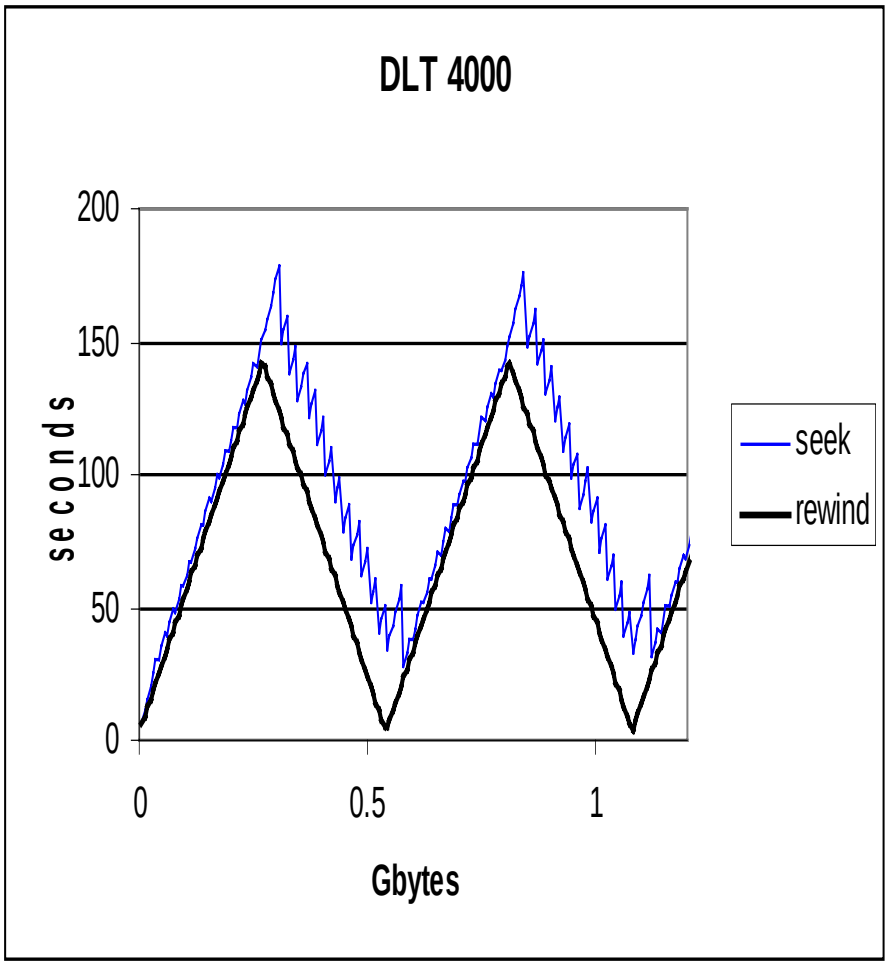
Mount and unmount time

- Mount : 10 to 50 seconds, low variance.
- Unmount : 4 to 20 seconds, low variance.
- Ampex can unmount from mid tape:

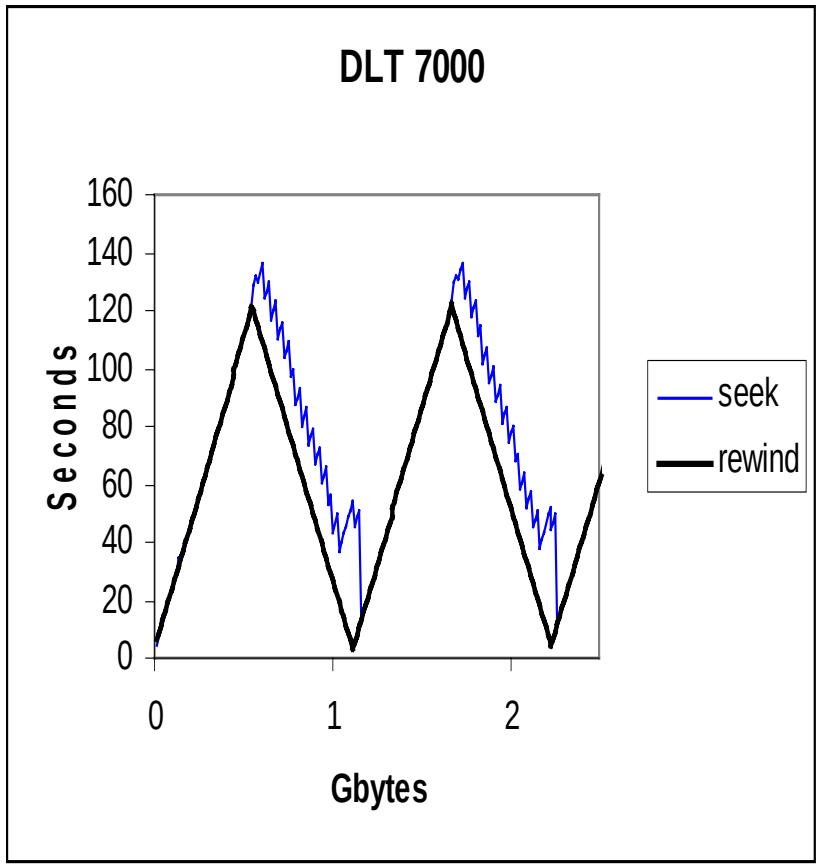


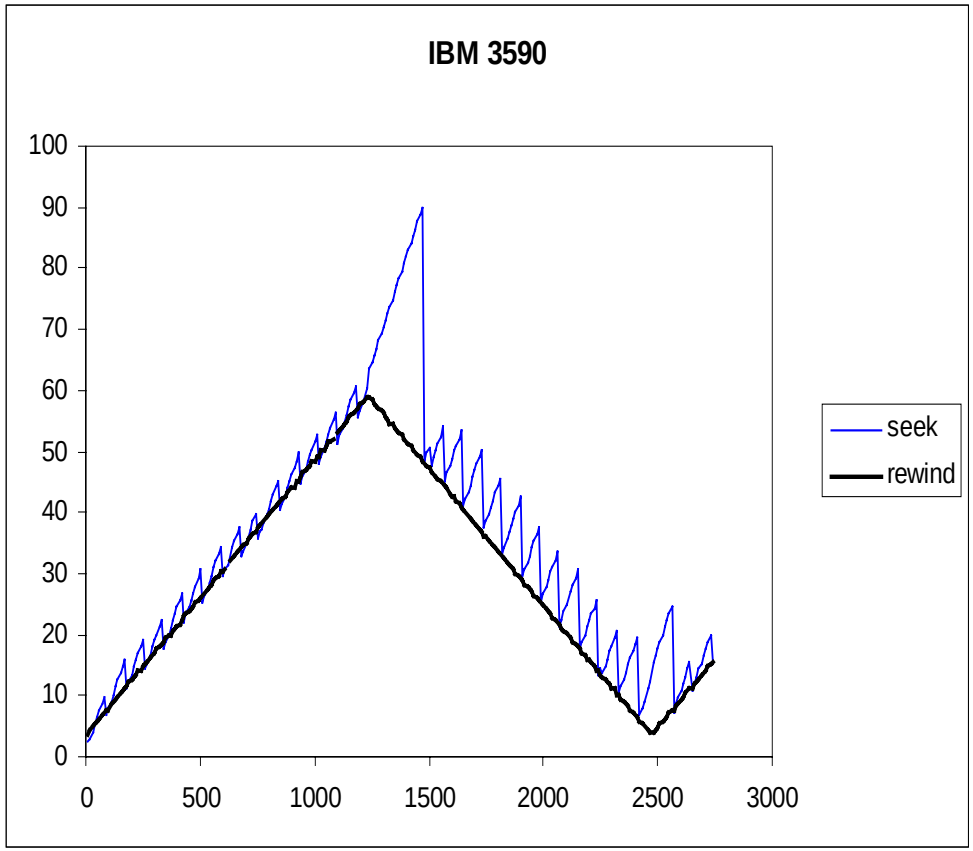
Seek time from BOT





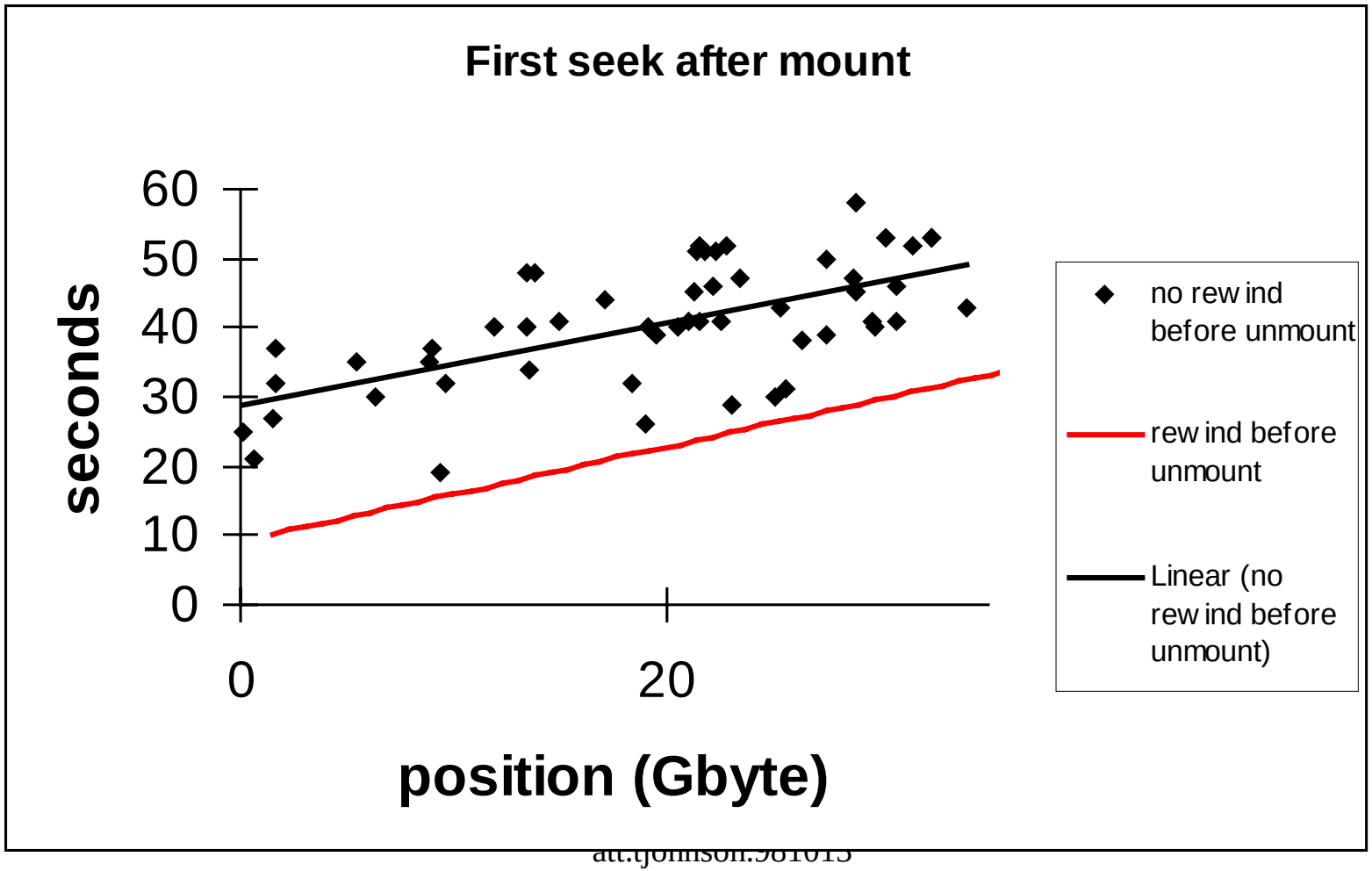
Seek time from BOT (cont.)



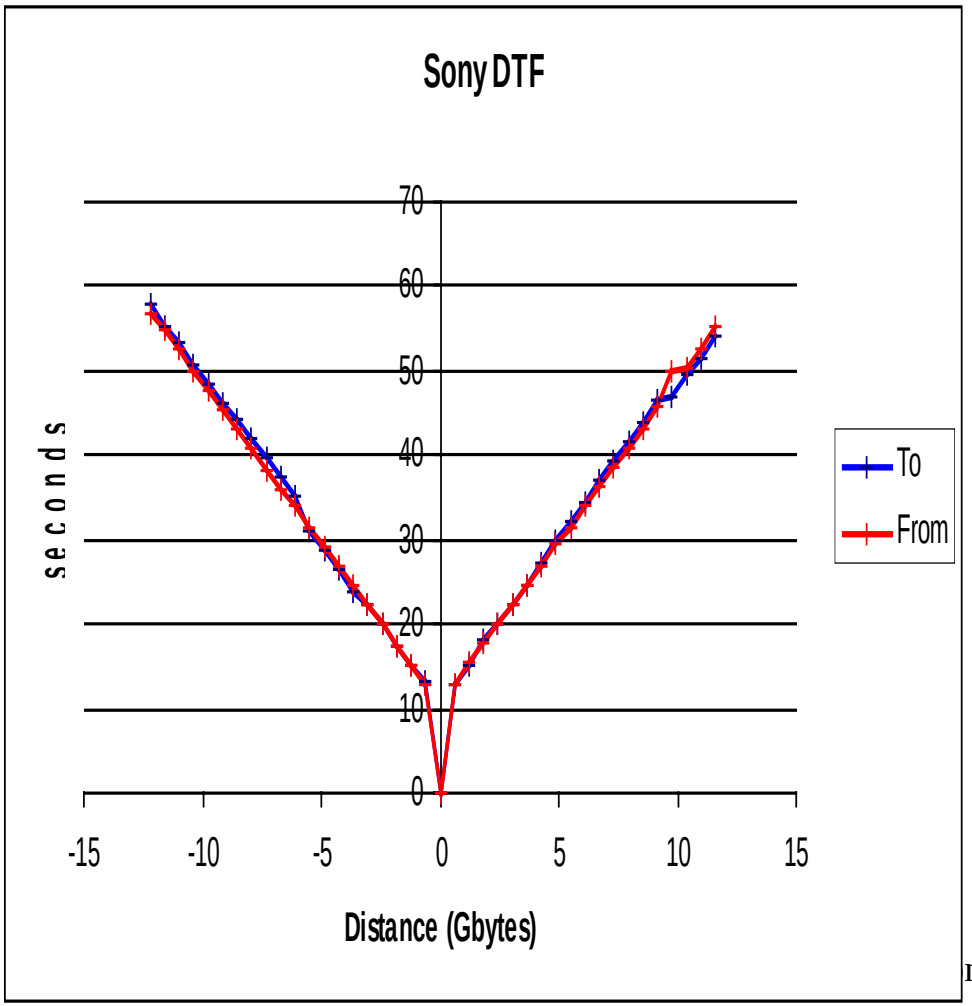


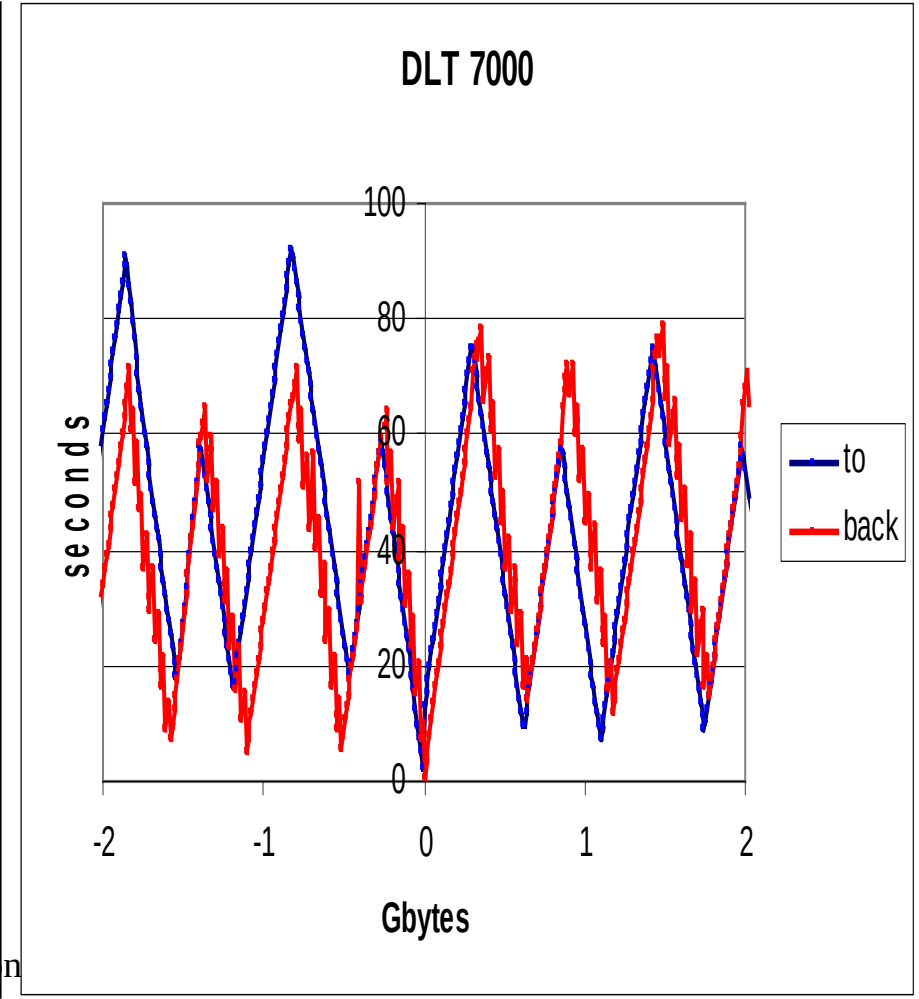
Unmount without rewind

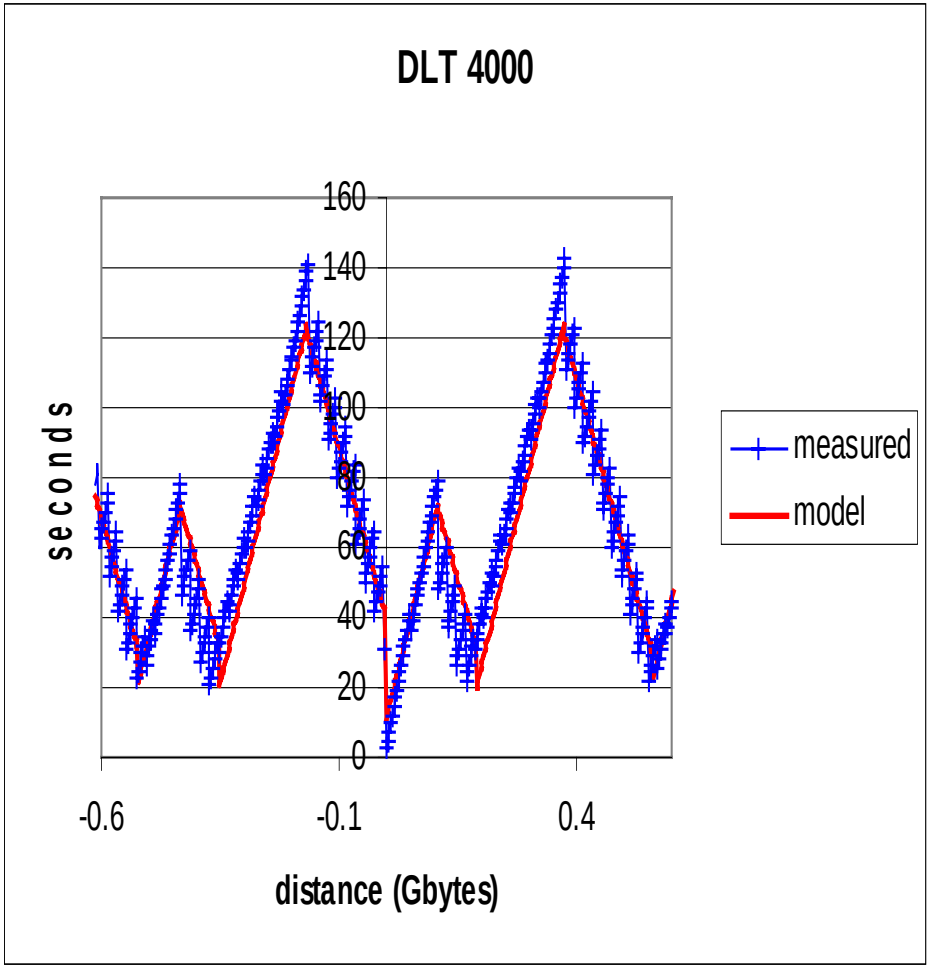
- Theory : faster seeks to data in mid-tape.
- Practice : slower seeks to data in mid-tape.



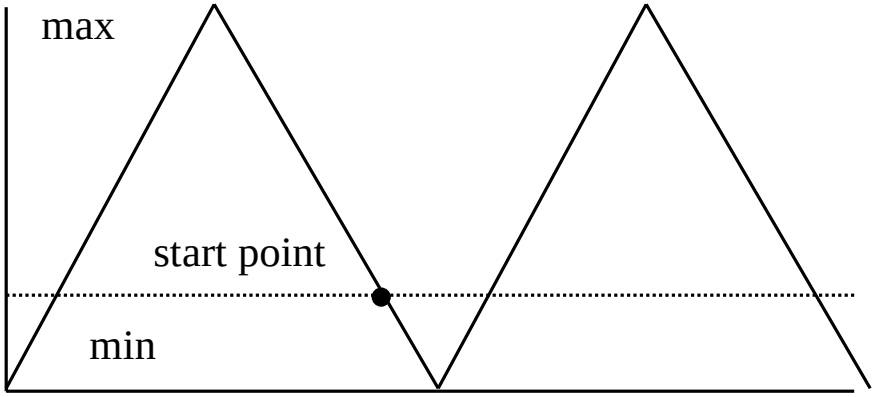
Seeks from mid tape



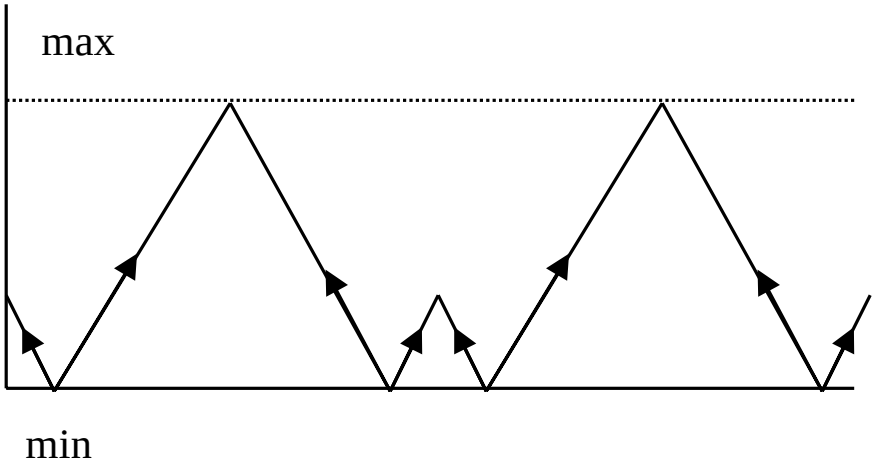




Direction on track

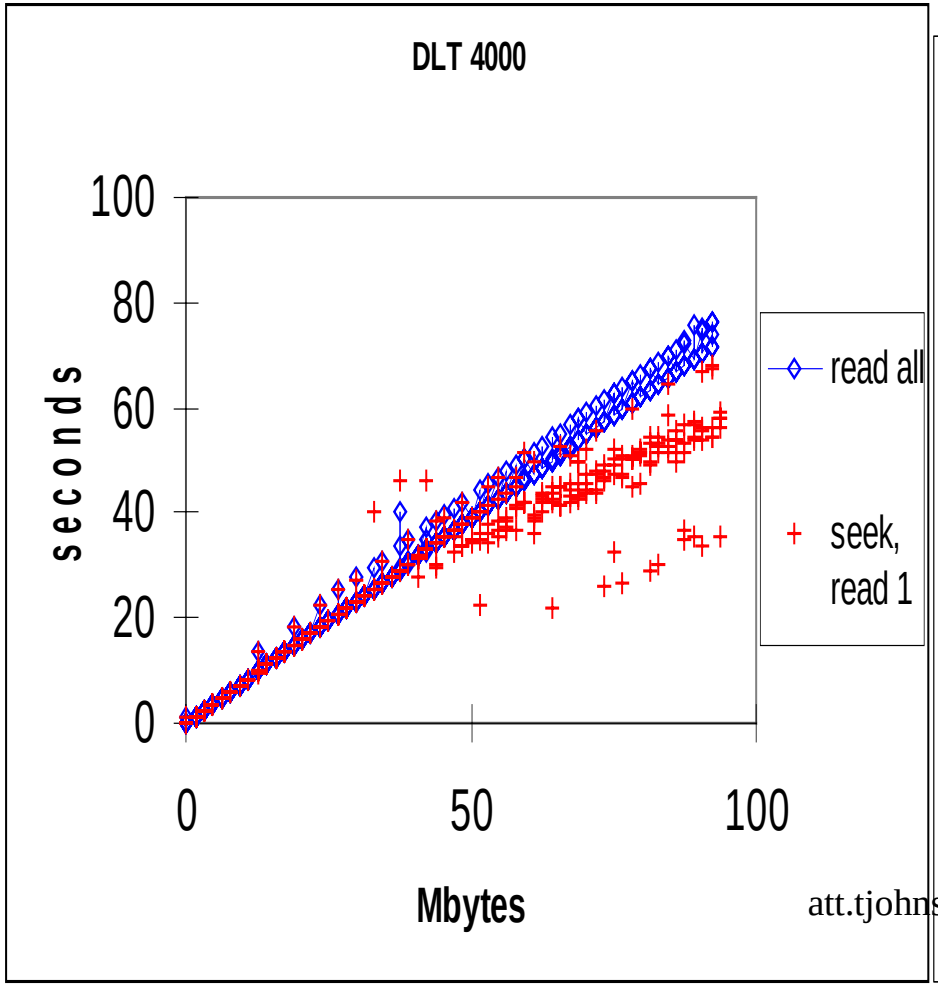


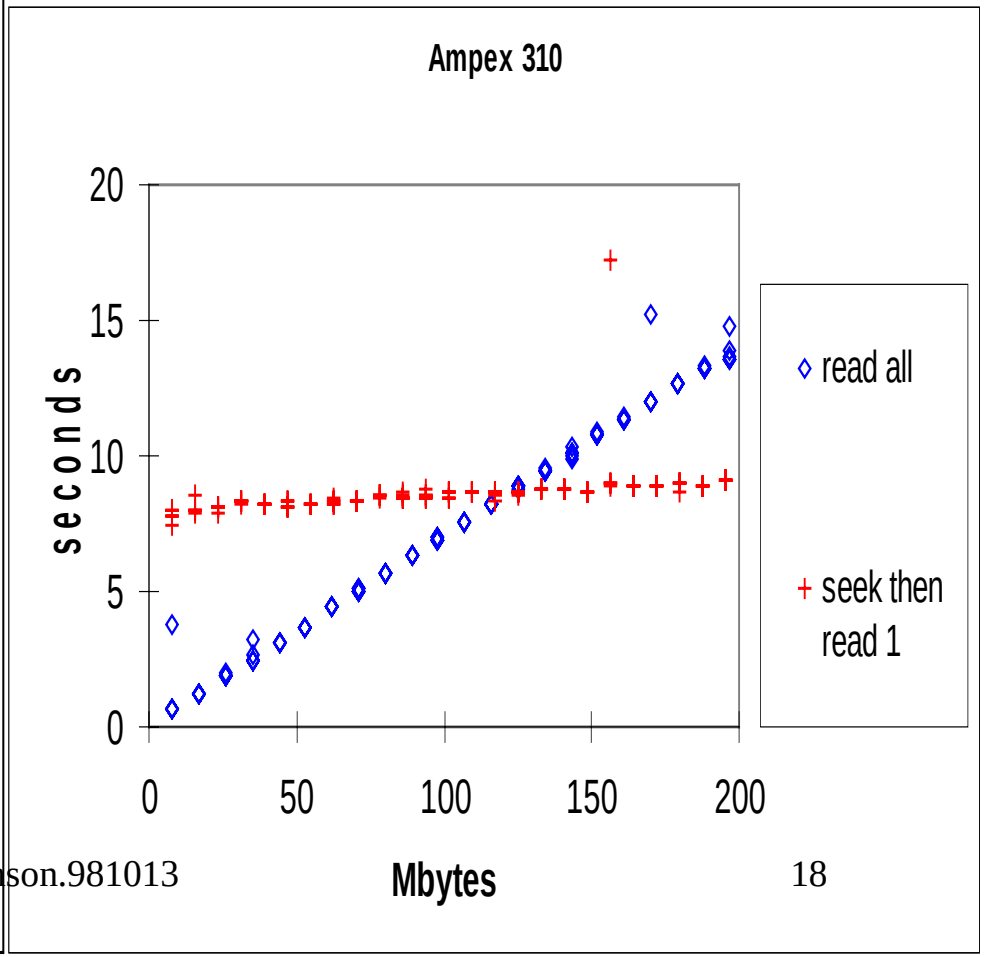
Seek distance



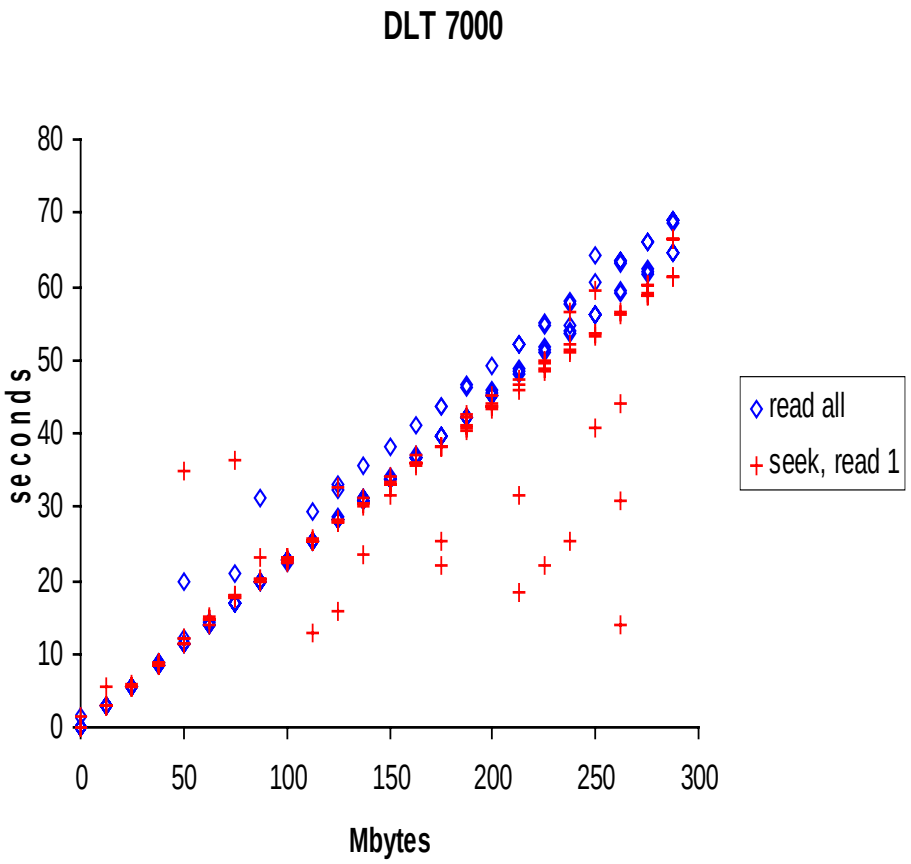
Short seeks

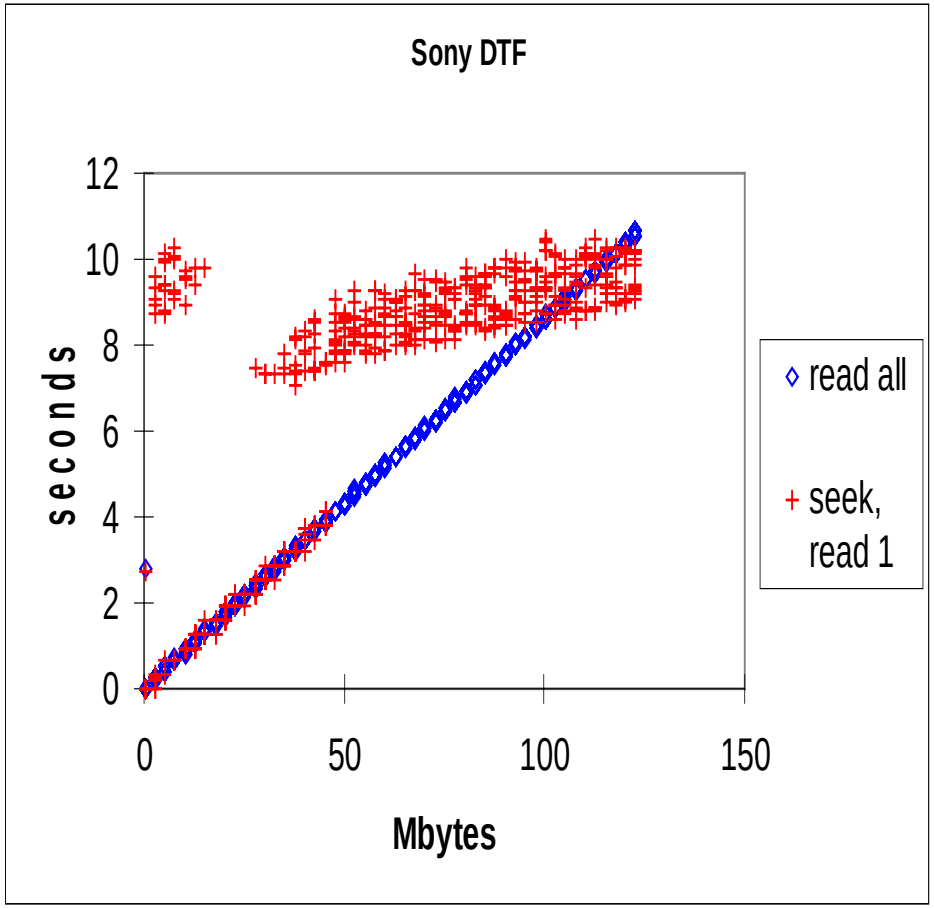
- Is it better to seek or to sequentially read?





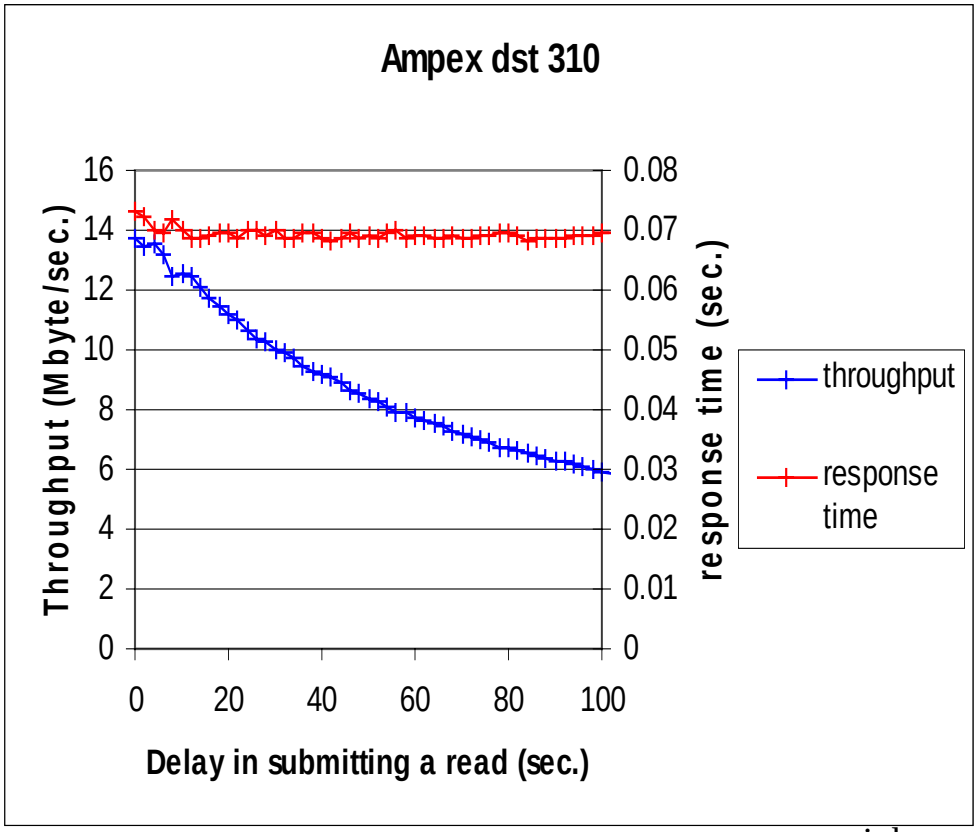
Short seeks (cont.)

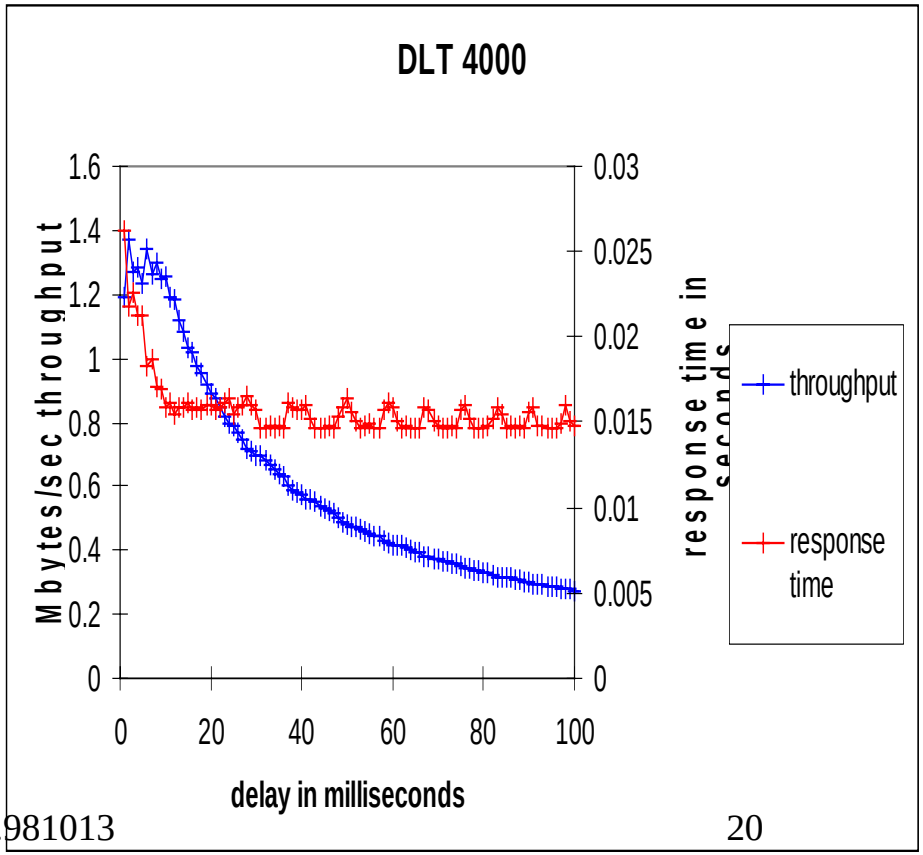




Delays in submitting read requests

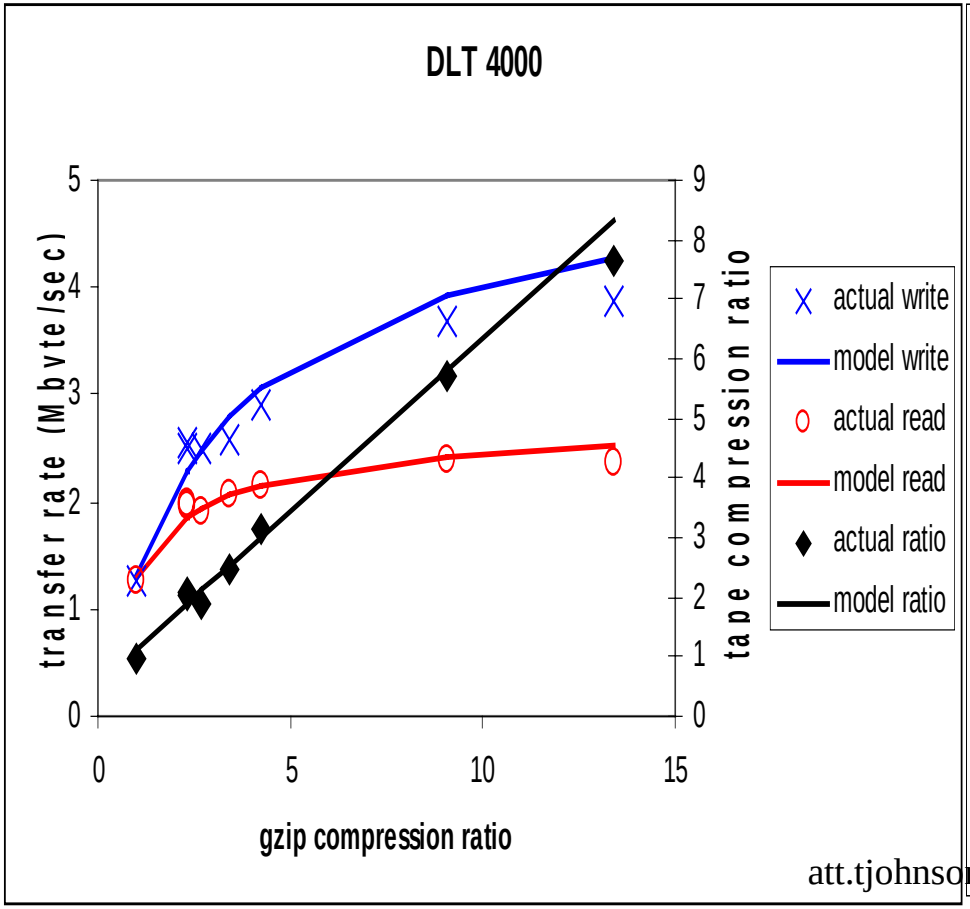
- Little effect for small reads.
- Startup penalty for helical scan with large reads, large delays
  - Ampex DST 310 : 7 second penalty for 70Mbyte reads, 32 sec. delays

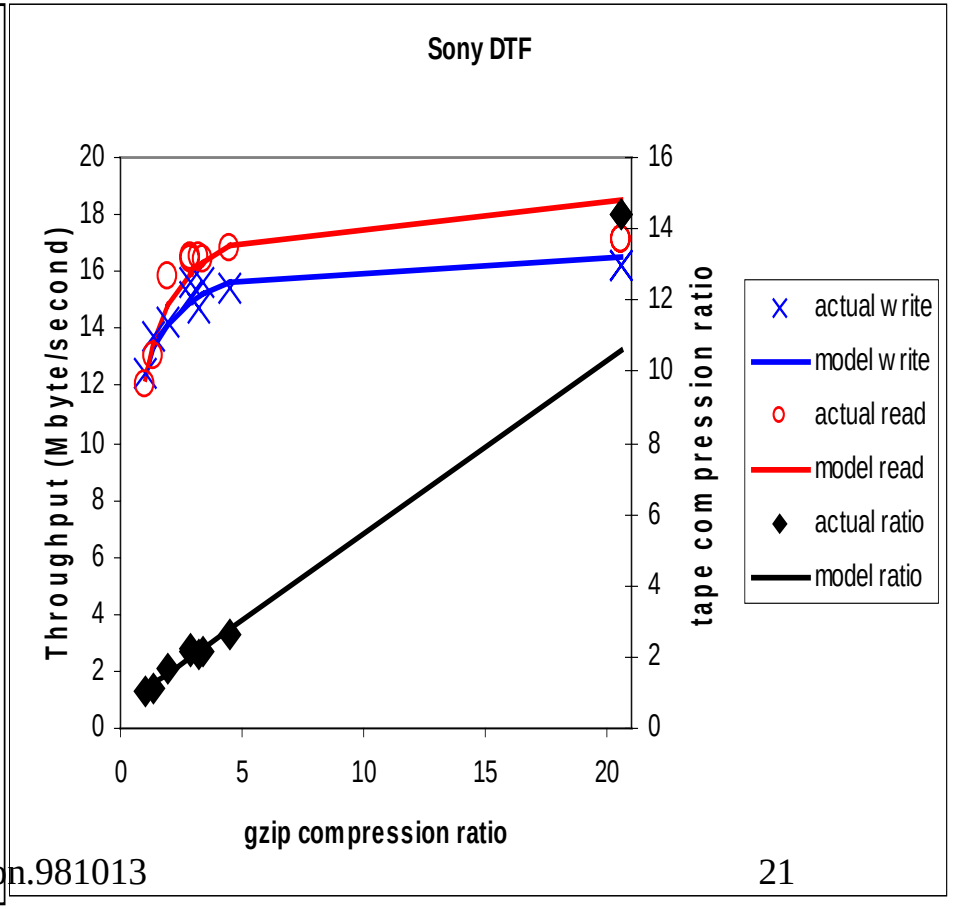




Compression

- $C_{tape} = a + b * C_{gzip}$
- Throughput = $1/(a + b/C_{gzip})$
  - Different constants for reading, writing.





Conclusions

- Performance measurements are necessary for planning.
  - Simple example: How big should the file sizes be?
  - Cutoff : the minimum file size to get 50% of the tape drive throughput.

Drive	file location	transfer rate	min file size	% of tape
4mm	128 sec.	.325 MB/sec.	42 MB	2.6 %
Ampex DST 310	35.7	14.2	507	1.2
Sony DTF	143	12.0	1720	4.2
DLT 4000	145	1.27	184	1.0
DLT 7000	121	4.3	520	1.5
IBM 3590	61	8.9	543	5.3

Understanding Tertiary Storage Performance

- Understanding tape-based tertiary storage is too complex.
  - Evaluating new equipment is a major chore.
- Things would be easier if someone systematically evaluated equipment.
- How complete are my measurements?
  - They work for me.
  - But perhaps not for others.
    - Arie Shoshani et al. of LLNL: building a scientific database using HPSS. Can only access data in units of a file. They consistently find a significant delay in the time to move from the end of a file to the start of the next file.

Providing Performance Information

- Why can’t the tape device tell the user about the performance they can expect?
  - I.e., for the current tape, given the current state of the drive.
- Precedent : predicting tape failures.
- Examples:
  - A topology map would allow optimized seeks on serpentine tape.
  - Seek time estimator?
  - Is the data rate too high? Too low?
  - Estimated remaining space?

Thanks to

- NASA Goddard Space Flight Center
  - EOSDIS / Mass Storage Testing Laboratory
  - Jeanne Behnke, P.C. Hariharan, Joel Williams
- Dept. of Defense
  - Ethan Miller
  - Jim Ohler, Jim Finlayson
- AT&T Labs - Research
  - Gary Sagendorf