Striped Tape Arrays

Ann L. Drapeau

Computer Science Department
University of California
571 Evans Hall
Berkeley, CA 94720

alc@cs.berkeley.edu

Striped Tape Arrays

Ann L. Drapeau
alc@cs.berkeley.edu

Motivation

- Applications require **high throughput** (100 MB/sec), **massive storage** (Terabytes, Petabytes)
- Technology Trends
  - Magnetic tape: high capacity, low bandwidth
  - Robots: automatic loading of tape cartridges
- **Striping: a technique for increasing throughput**
- Issues in striping effectively
- Tape array reliability

Outline

- Introduction to Striping
- Applications
- Tape Technologies
- Robots
- Access Times
  - Drive and Robot Measurements
- Striping Options and Issues
- Reliability Issues
- Summary

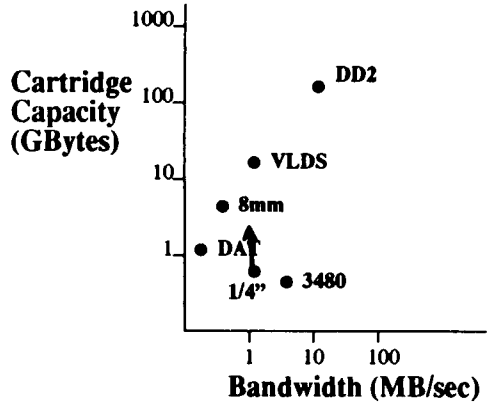
Data Striping

- **Spread data from individual files across several devices**
- Advantages:
  - **Increase bandwidth to a single file**
  - **Reduce latency of large accesses**
  - Allows independent "smaller" accesses
  - Easy to incorporate error correction
- Problems:
  - Increase latency of some accesses
  - Synchronization

Do Applications Need Striped Tape?

- **Large scientific archives (NASA EOS)**
  - High sustained bandwidth (100 MB/s)
  - Total storage very large (Petabytes)
  - Would benefit from striping **throughput**
- **Interactive access to large data sets (Sequoia)**
  - Researchers across California
  - Want reasonable response time over network
  - Total storage large (Terabytes)
  - Striping would **reduce large access latency**

Tape Technologies



Tape Technologies

Technology	Capacity per Cartridge	Cost	Bandwidth
1/4"	2 GB	$ 1000	3 MB/sec
1/2" 3480	480 MB	$ 20000	6 MB/sec
4mm DAT	1.3 GB	$ 1000	183 KB/sec
8mm Exabyte	5 GB	$ 3000	500 KB/sec
1/2" Metrum VLDS	14.5 GB	$ 40000	2 MB/sec
Ampex DD2	165 GB	$150000	15 MB/sec

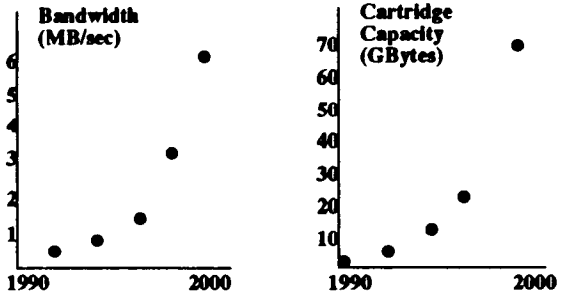
Linear Recording: 1/4" cartridge, 1/2" 3480

Helical Scan: DAT 4mm, 8mm, 1/2" VLDS, 19mm D2

Tape Tradeoffs: No "Perfect" Drive

- **Inexpensive** helical scan drives have **low bandwidth** (DAT, 8mm)
- **Inexpensive** serpentine drives have **moderate bandwidth** (1/4")
- **High capacity** drives have **long access times** (helical scan, 1/4")
- Drives with **short access times** are **low capacity** (1/2" 3480)
  - Moderate price and bandwidth
- **High bandwidth** drives **very expensive** (DD2)
  - **Bandwidth not high enough**
  - **Very high capacity**

Future Tape Drives (8mm)



- Source: Harry C. Hinz, Exabyte Corp.
- **Changes:** increase track density, decrease track width & pitch, reduce tape thickness, increase rotor speed

Robots

- **Large Libraries:**
  - many cartridges, several drives
  - expensive
  - one or more robot arms
- **Carousels**
  - around 50 cartridges, one or two drives
  - moderate cost
- **Stackers**
  - around 10 cartridges, one drive
  - inexpensive

Robots

	Metrum RSS-600 (1/2" VLDS)	Spectra Logic STL-8000H Carousel (8mm)	Exabyte EXB-10 Stacker (8mm)
# Drives	up to 5	1 or 2	1
# Cartridges	600	45	10
Total Capacity (GBytes)	>6000	225	50
Cost	$540,000 (2 drives)	$27,500 (1 drive)	$7000
Avg. Robot Access Time (sec)	8	10	<20

Tape Access Time (Cartridge Switch)

- Access time =

  rewind time +
  eject time +
  robot unload +
  robot load +
  device load +
  fast search +
  transfer time

- Measured three tape drives, one robot:
  Accurate access time models for simulation

Drive Measurements

Drive Load and Eject Times

	4mm DAT	8mm Exabyte	Metrum VLDS
Mean Load Time (sec)	16	35.4	28.3
Mean Eject Time (sec)	17.3	16.5	3.8

Data Transfer Rates

	4mm DAT	8mm Exabyte	Metrum VLDS
Read Rate (MB/sec)	0.17	0.47	1.2
Write Rate (MB/sec)	0.17	0.48	1.2

Rewind and Search Behavior

	4mm DAT	8mm Exabyte	Metrum VLDS
Rewind Startup (sec)	15.5	23	15
Rewind Rate (MB/sec)	23.1	42.0	350
Search Startup (sec)	8	12.5	28
Search Rate (MB/sec)	23.7	36.2	115

- Constant startup
- Approximately linear search/rewind

Tape Access Time Example
(Exabyte EXB8500 Drive, EXB-120 Robot)

- Average Access time =

rewind time	(1/2 tape)	(75 sec)	+
eject time		(17 sec)	+
robot unload		(21 sec)	+
robot load		(22 sec)	+
device load		(35 sec)	+
fast search	(1/2 tape)	(84 sec)	+
transfer time			

- **Not including data transfer: 4 minutes!**

Options for Striped Tape

- Within a robot
  - + cartridges in stripe kept together
  - - few readers, robot arms
  - - single point of failure
- Between robots
  - + several robot arms used in access
  - - harder to keep cartridges together
- Between small robots (stackers)
  - + highest proportion arms to readers and cartridges

Striping Issues

- Configuration depends on workload
- **Interleave factor crucial:**
  - Too small: cartridge switches increase latency (Long access times -- big penalty)
  - Too big: lose potential parallelism
- **Workloads that will benefit from striping**
  - Large archives
  - Interactive systems with large avg. request size
- **Striping will hurt** performance of some accesses
  - Interleave much smaller than average request
  - High load/scarce readers

More Striping Issues

- Striping with improved devices/robots
  - **Higher bandwidth drives**
    - Bandwidth, aerial density may increase 30X by end of decade
    - Less need for striping?
    - **Still get throughput benefits**
  - **Faster access times (drives and robots)**
    - faster load, eject, search, rewind, robot arms
    - no rewind before eject
    - cartridge switch penalties reduced
    - **striping more effective**

Synchronization Issues

- Drives retry after failed writes
  - Bad tape would retry indefinitely
  - Pat Savage (Shell Oil): after write error, retry on all tapes in stripe
- If "RAID-5" (large interleaving)
  - Single cassettes may satisfy smaller requests independently
  - Large requests spanning several tapes may be out of synchronization by minutes
  - Buffer space required to hold stripe units while request completes

Reliability Issues: Tape Media

- **High rates of raw bit errors**
  - before internal ECC
  - one in 10^5 bits
- **Dropouts**
  - Debris
    - Slicing of tape
    - Particles in atmosphere
    - Start/stop wear
  - Nonhomogeneous Tape Coating

Uncorrectable Bit Error Rates

Drive	Bit Error Rate
1/4"	10^{-14}
4mm DAT	10^{-15}
Exabyte 8mm	10^{-13}
Metrum VLDS	10^{-13}
Ampex DD2	10^{-12}

- **Error rates after ECC**
- Terabyte approximately 10^{13} bits
- **MSS will contain uncorrectable errors!**

- **Need Error Correction**
  - Easy to implement in striped systems
  - **How much?**
  - How reliable are error rates?
  - How will ECC affect performance?
- Error Rates Increase with Wear
- **Tapes last around 2000 passes**
- Severe wear: tape unreadable
- **If tapes are rewritten often, need to copy tapes periodically**

Reliability: Tape Heads

- Drive design includes tape/head wear
- Accumulate debris
  - tape debris
  - atmosphere
  - tape coating (friction, humidity)
- Wear with tape medium helps clean heads
- **Heads last around 2000 hours of tape contact**
- Algorithms for
  - Periodic head cleaning
  - Fast replacement on failure

More Reliability Issues

- Other drive problems

Megatape 1991 Repair Statistics (8mm)

Repair type	%
Replace heads	44
Tape mechanism (reel motors, tape tension, etc.)	21
Card failure	17
Other (firmware, power supply, etc.)	14
No defect found	4

- Robot reliability
- Support hardware

Summary

- **Applications want high sustained throughput**
- **Technology Trends:**
  - Tape drives increasing in capacity, bandwidth (currently inadequate)
  - Robots allow automatic handling of cartridges
- **Striping:**
  - Increased throughput
  - Reduced latency of large requests
- **Striping configurations:**
  - Within or between robots
  - Tradeoffs: ratio of readers, robot arms, tapes

- **Striping issues:**
  - Interleave factor for best performance
  - Effect of improved drives, robots
  - Synchronization problems
- **Reliability Issues:**
  - Media Wear
  - Head Wear
  - Other drive failures
  - Robot failures
  - Error correction needed: how much?