Venti:
a new approach to archival data storage

Sean Quinlan

Sean Dorward

Bell Labs

Lucent Technologies

why archival storage?

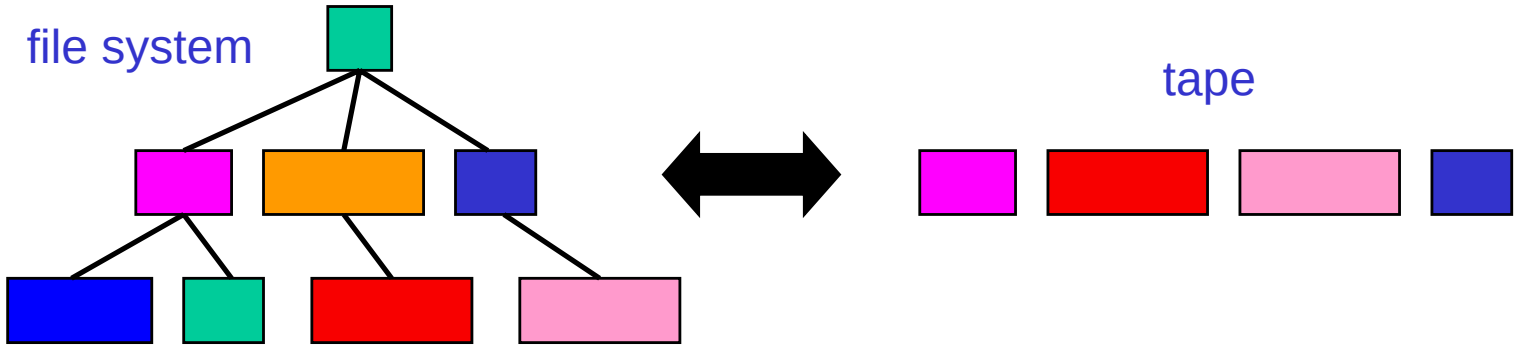
Lucent Technologies
Bell Labs Innovations

- disaster recovery
  - necessary but rarely used
- history of changes
  - many uses if available
  - experience with Plan 9 file system
- storage is plentiful
  - infinite if capacity increases faster than consumption
  - why delete anything?

tape backup

Lucent Technologies
Bell Labs Innovations

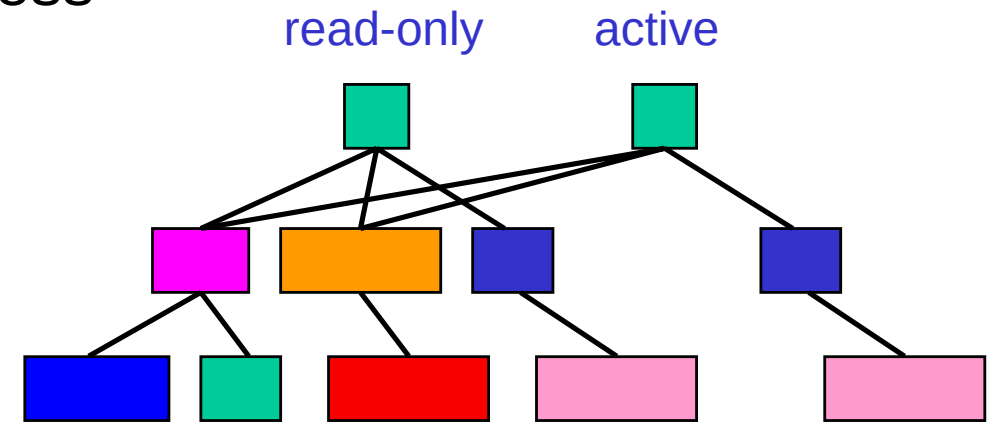
- backup
  - copy data from file system to tape
- restore
  - copy from tape to file system
  - often painful
- tapes are streaming devices
  - tension between full and incremental backup



file system snapshots

Lucent Technologies
Bell Labs Innovations

- a consistent read-only view of the file system
- access with standard tools
  - ls, cat, cp, grep, diff
  - retain file system permissions
- looks like a full backup
- implementation resembles an incremental back-up
- fast random access
  - use in place
  - share blocks



Venti

Lucent Technologies
Bell Labs Innovations

- block-level network storage system
  - back end storage for multiple clients
- write-once
  - once data is stored, can not be deleted
  - simplifies administration and security
- data stored on magnetic disks
  - impressive technology curve
  - high performance random access
- blocks are identified by a hash of their contents
  - write(data) *not* write(block, data)
  - read(H(data)) $\rightarrow$ data *not* read(block) $\rightarrow$ data

interesting properties

- no way to overwrite a block
  - different blocks have different hashes
- blocks can be shared
  - multiple writes of the same data will be coalesced
- multiple clients can share a server
  - the hash function is a universal name space
- a secure hash authenticates data
  - server can not lie
- simple to replicate/cache/load balance

sha1: secure hash algorithm 1

Lucent Technologies
Bell Labs Innovations

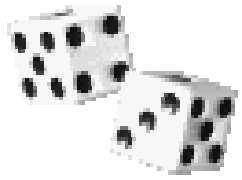
- proposed by NIST
  - US National Institute for Standards and Technology
- hash value is 160 bits → 20 bytes → venti
- no known collisions
- believed to be secure
  - difficult to generate data with a given hash value

- reasonably fast software implementations
  - ~60 Mbytes/sec on 700Mhz Pentium 3

is 160 bits enough?

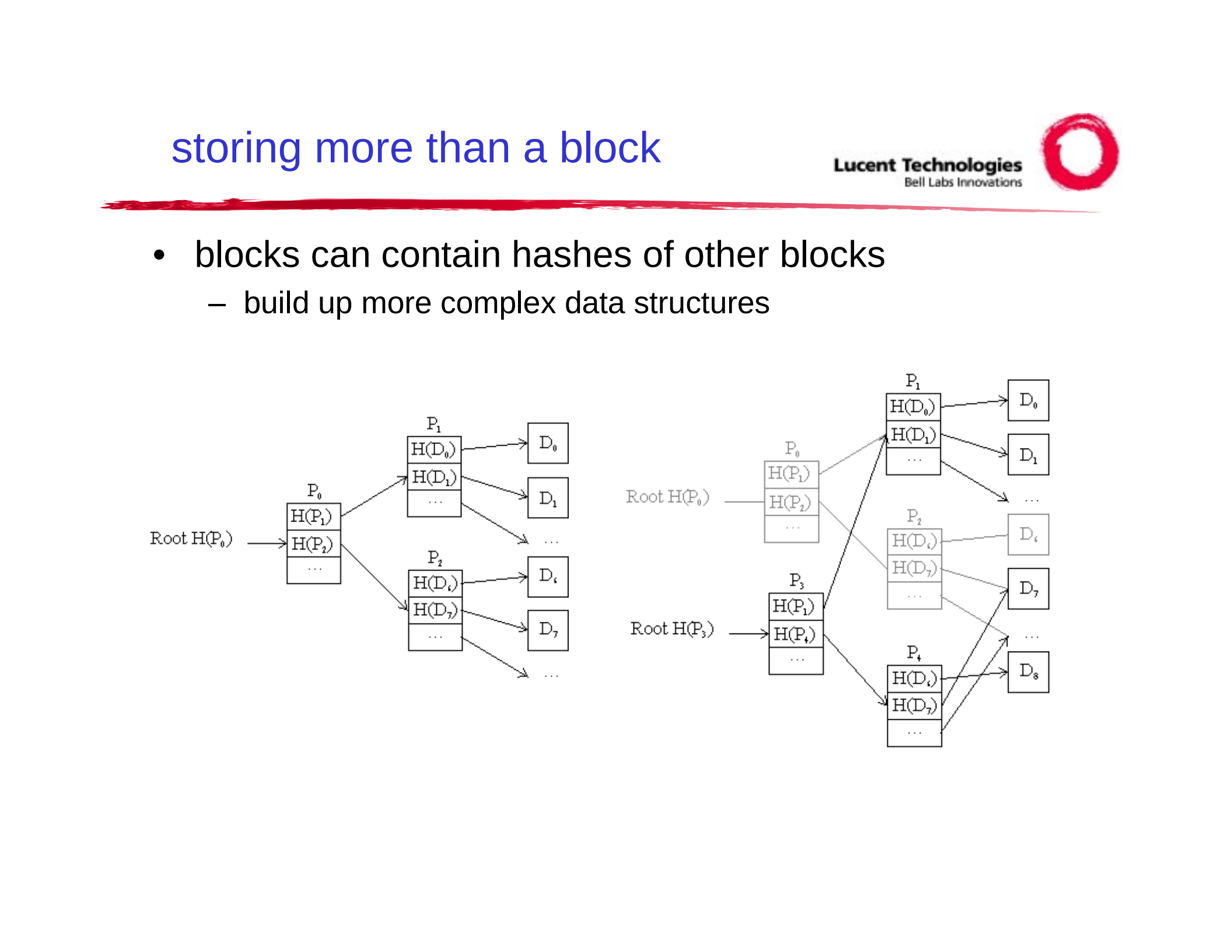
Lucent Technologies
Bell Labs Innovations

- $2^{160} \sim 10^{48}$ hash values
- suppose 10^{14} 8KB block (~ 1 exabyte)
- less than 1 in 10^{20} chance of collision
  - assert no collisions
  - although we do check
- possible to move to sha256 in the future



storing more than a block

Lucent Technologies
Bell Labs Innovations



- blocks can contain hashes of other blocks
  - build up more complex data structures

vac: a zip like application

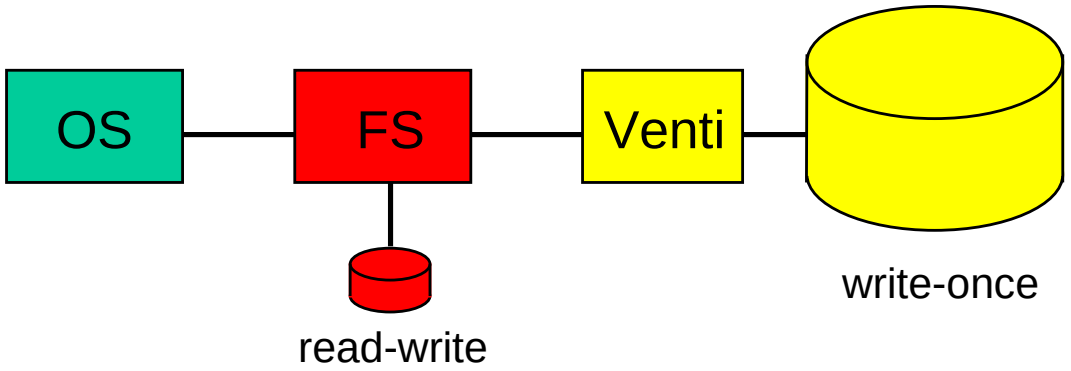
- vac files ...
  - produces a tree of blocks corresponding to specified files
  - similar to zip
  - output is a single hash
- compress any amount of data to 20 bytes!
- Venti will coalesce multiple copies of data
  - using vac multiple times will not consume extra storage
  - using vac on slightly changed data should only consume storage proportional to the delta (assuming block alignment)
- unvac
- vacfs

block-level (physical) backup

- copy raw disk blocks
  - avoid interpreting and walking file system
  - potentially much higher throughput
- block-level backup to Venti
  - coalesces duplicate blocks based on data
  - space advantages of incremental backup
- random access
  - directly mount
  - lazy restores

new plan 9 file system

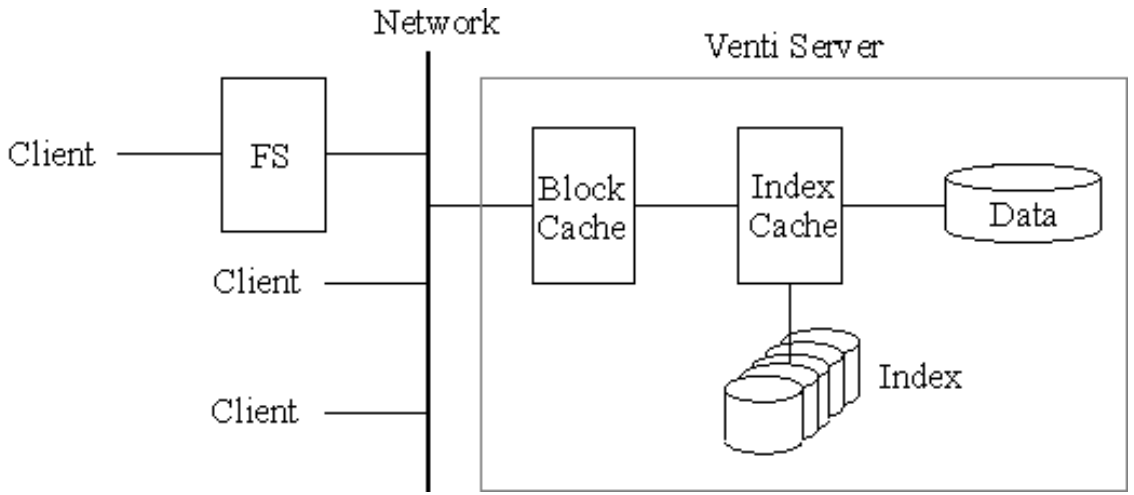
- build a file system directly on top of Venti
  - primary location for active data
- use a small amount of read/write storage
  - smaller than active file system
  - accumulates changes to the file system
  - snapshot flushes changes to Venti
- permanently retain all snapshots



Venti implementation

Lucent Technologies
Bell Labs Innovations

- network service accessed via a simple protocol
  - supports multiple clients
  - variable sized blocks
- combines a data log, index, and caches



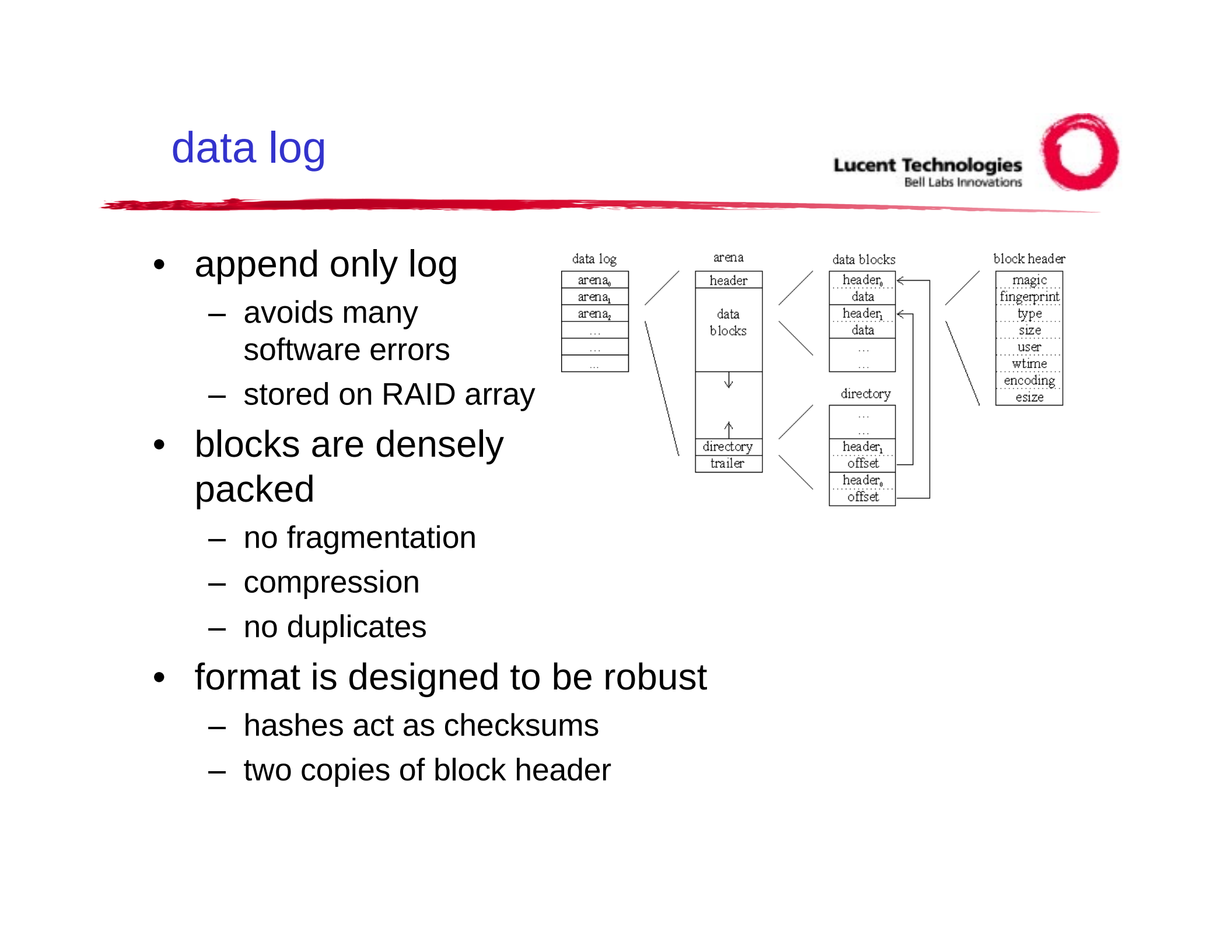
hardware

Lucent Technologies
Bell Labs Innovations

- server
  - 2 processor x86 box with 2GB of memory
- index
  - 8x 10,000rpm 9Gb scsi disks
- data
  - raid array with 8 7200rpm 75GB ide disks
  - total of 500GB using RAID 5
- cost ~$14K in 4Q 2000

data log

Lucent Technologies
Bell Labs Innovations

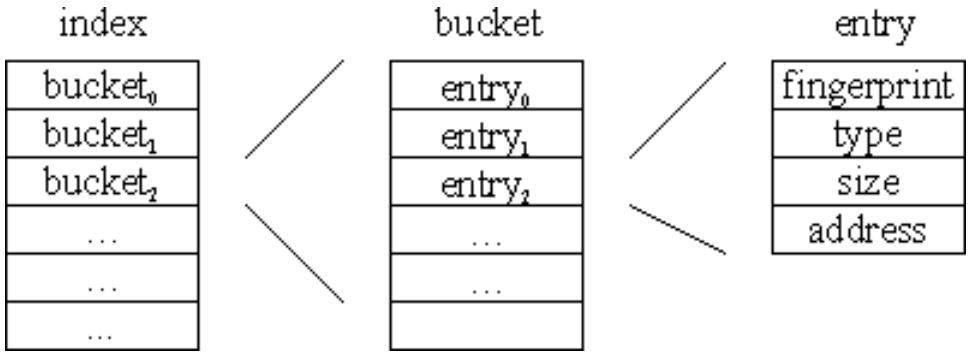


- append only log
  - avoids many software errors
  - stored on RAID array
- blocks are densely packed
  - no fragmentation
  - compression
  - no duplicates
- format is designed to be robust
  - hashes act as checksums
  - two copies of block header

index

Lucent Technologies
Bell Labs Innovations

- maps 160 bit hash to location in log
- implemented as a disk-resident hash table
  - hash the hash
  - one disk access
- index can be rebuilt from log
- multiple index disks provided improved throughput



memory caches

- block cache
  - avoid any disk I/O
  - ~100,000 entries in 0.5 Gbyte
- index cache
  - cache hash $\rightarrow$ log mapping
  - avoids I/O to index disks
  - ~10,000,000 entries in 0.5 Gbyte
- caches improve
  - reads
  - duplicate writes
- caches do not improve
  - virgin writes

read & write performance

Lucent Technologies
Bell Labs Innovations

8Kbyte blocks in Mbyte/sec

	Sequential Reads	Random Reads	Virgin Writes	Duplicate Writes
Uncached	0.9	0.4	3.7	5.6
Index Cache	4.2	0.7	-	6.2
Block Cache	6.8	-	-	6.5
Raw Raid	14.8	1.0	12.4	12.4

- initial results
- uncached sequential reads need work
  - limited by latency of index disks
- uncached writes benefit from multiple index disks

Plan 9 historical data

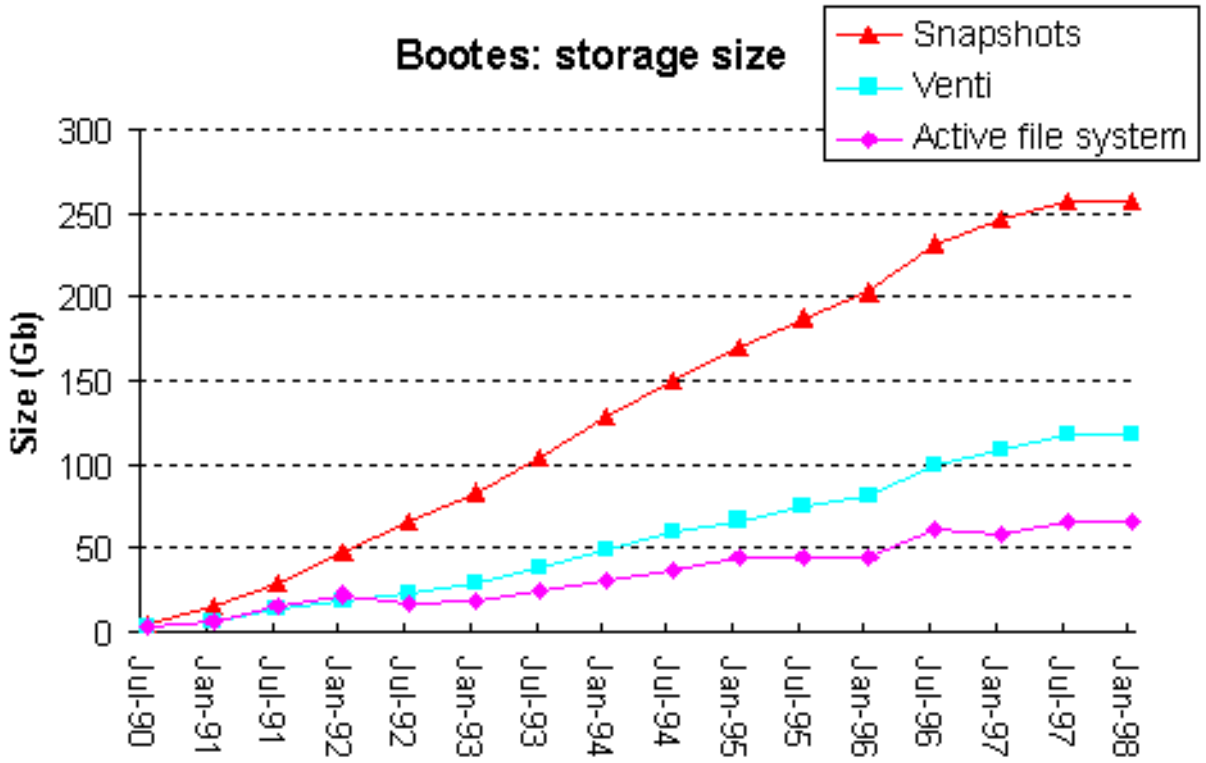
Lucent Technologies
Bell Labs Innovations

- two Plan 9 file servers spanning 1990 - 2001
- daily snapshots
  - stored on optical WORM
  - ~ 650 Gbytes of data
- 522 user accounts
  - 50 - 100 active users at any time
- many software development projects
- several large data sets
- traces are available on the web

http://www.cs.bell-labs.com/~seanq/p9trace.html

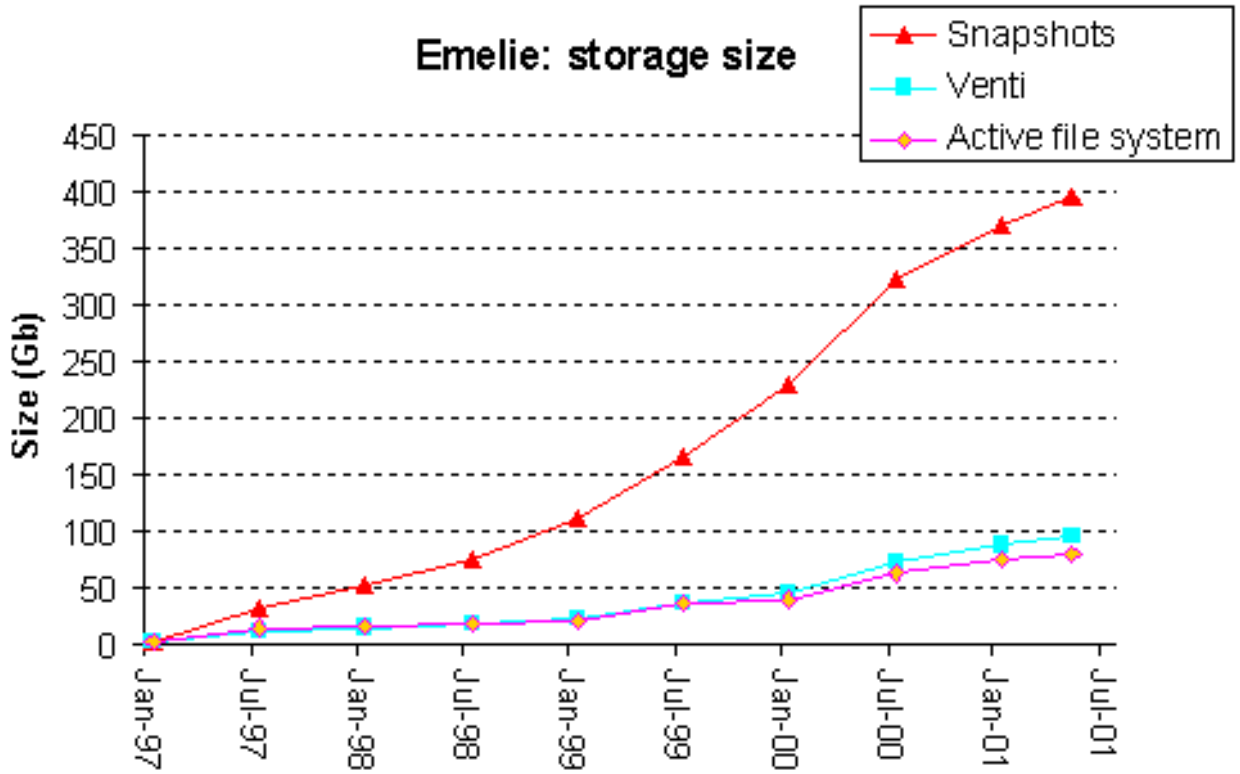
bootes: 1990 - 1997

Lucent Technologies
Bell Labs Innovations



emelie: 1997 - 2001

Lucent Technologies
Bell Labs Innovations



sources of compression

Lucent Technologies
Bell Labs Innovations

	bootes	emelie
Elimination of duplicates	27.8%	31.3%
Elimination of fragments	10.2%	25.4%
Data Compression	33.8%	54.1%
Total Reduction	59.7%	76.5%

reliability & recovery

Lucent Technologies
Bell Labs Innovations

- tools that run on the server
  - check index
  - check log
  - rebuild index
  - copy section of log to removable media
- RAID 5 provides some protection for log
- would like offsite mirror
  - simple to implement
- would like write-once disks
  - protection against a buggy or compromised server
- currently backup log to tape
  - append only structure of the log makes this easier

more work

Lucent Technologies
Bell Labs Innovations

- load balancing
  - divide work based on hash
  - add proxies to hide from client
  - scalable performance
- replication/caching
  - background exchange of write operations
  - forward failed read operations
  - no coherency problems!
- access control
  - currently authenticate user
  - hash is a weak form of capability

conclusions

- hashes as block addresses
  - simple model with attractive properties
- write-once
  - easy to share data
  - simplifies implementation
  - simplifies administration
  - improves security
  - practicable
- magnetic disks for archival storage
  - amazing technology curve
  - high performance random access