Venti

Construction and Maintenance of a Centralized Hash Table

Russ Cox

PDOS Group Meeting
November 15, 2005

`http://swtch.com/~rsc/talks/`

History

Cached WORM file server (Quinlan and Thompson):

- active file system on magnetic disk acts as worm cache
- mark all disk blocks copy-on-write at 5am to take snapshot
- slowly dribble snapshot to worm
- maintain forward linked list of snapshots
- present snapshot tree to users
- became integral part of our computing environment

```
% ls -lp /n/dump/*/*/386/bin/8c | uniq
--rwxrwxr-x presotto sys 243549 Jan 21  1997 8c
...
--rwxrwxr-x presotto sys 298289 Dec 14 18:55 8c
%

% yesterday -D authsrv.c
diff -n /n/dump/2003/0106/sys/src/cmd/auth/authsrv.c authsrv.c
/n/dump/2003/0106/sys/src/cmd/auth/authsrv.c:100 c authsrv.c:100
<           break;
---
>           exits(0);
%
```

Quinlan, ‘‘A Cached WORM File System’’, SP&E December 1991.
`http://plan9.bell-labs.com/~seanq/cw.pdf`

History, ii

WORM was right choice in 1990

- one jukebox is infinite: capacity grows faster than our storage needs
- no head crashes
- plausible random access times
- magnetic disks too small, tape too slow
- bootes (1990): 100MB mem, 1GB disk, 300GB juke box
- emelie (1997): 350MB mem, 54GB disk, 1.2TB juke box

What about 1999?

- disks cheap and big, getting cheaper and bigger
- disks cheaper and bigger than optical disk
- disks much faster than optical disk
- disks have head crashes
- build a better base out of magnetic disk?

Venti

Archival block store (Quinlan and Dorward):

- SHA1-addressed
- blocks never reclaimed
- omit duplicate blocks
- compress

Implementation:

- log of all blocks ever written
- log broken into fixed-size (say, 500MB) chunks called *arenas*
- arenas copied to other media (tape, DVD, etc.) as they fill
- index on the side makes lookups efficient

Initial system:

- iolaire (1999): 2GB mem, 36GB index, 480GB hw raid arenas

Quinlan and Dorward, ‘‘Venti: a new approach to archival storage’’, FAST 2002.
`http://plan9.bell-labs.com/sys/doc/venti.pdf`

PDOS Backups

Store FFS disk images into venti

- Merkle hash tree of blocks
- backup program parses only FFS block-in-use bitmap
- very simple, reliable

Separate user-level NFS server presents backups

- parses full FFS
- ‘okay’ to be buggy (data still in venti)
- mounted on `/dump` on amsterdam

Depends on venti

- prototype adequate but slow
- intended to run on RAID or other ‘reliable’ storage

PDOS Venti Servers

amsterdam (2002)

- two 120GB IDE drives that Chuck wasn't using.
- original venti software

venti.csail (2004)

- machine hand-built by me and Frank
- four 300GB IDE drives packed in tight
- rewritten venti, stabilized in spring 2005
- disks failed in September 2005 (some fans were unplugged)

backup.pdos (2005)

- Coraid EtherDrive storage server, at least nominally
- rackmount PC running Plan 9 with 15 hot-swap SATA disk slots
- Coraid software serves various RAID configurations
- access via ATA-over-Ethernet (AoE)
- twenty 320GB SATA disks w/ five-year warranty; three have failed
- replaced Coraid software with new venti

Photos

Outline

History & Background

(you are here)

Sad state of modern disks

Venti architecture

Reliability measures

Modern ATA/SATA Disks

Good

- very cheap
- very large
- very fast at sequential access

Bad

- unreliable
- very slow at random access (seeks)

Ugly

- commercial pressures (Dell) will make the Good parts better
- but not fix the bad parts (blame Windows)
- (SCSI disks are more reliable, but 4x the price)

Modern Disks — Speed (MB/s)

read		write		
seq	rand	seq	rand	
23	0.64	25	0.78	IBM DTLA 30GB IDE
25	0.64	26	1.2	IBM/Hitachi Deskstar GXP 80GB IDE
12	0.47			IBM/Hitachi Deskstar GXP 120GB IDE
45	0.51	36	1.9	Maxtor 300GB IDE
49	0.60	49	1.0	Western Digital 320GB SATA
5.5	0.45	0.65	0.40	Seagate 4GB SCSI
48	1.1	1.34	1.1	Maxtor 300GB SCSI
17	0.46	14	2.4	Raidweb.com IDE RAID w/ SCSI interface
1.9	0.39			Canon 32MB Compact Flash / USB
4.9	3.7			PNY 256MB Compact Flash / USB
3.7	3.1			Toshiba 16MB SD Card / USB
6.4	4.1			Kingmax 512MB SD Card / USB
6.4	2.3			Patriot 1GB SD Card / USB
2.0	2.0			PQI 32MB Flash IDE
3.1	2.7			Transcend 512MB Flash IDE

Modern Disks — Reliability

No one will give out hard numbers

- we bought a 20-pack of WD3200JD SATA disks
- started using 15 of them five weeks ago
- 4 have died - one can't spin up, one has many bad sectors, two unprocessed

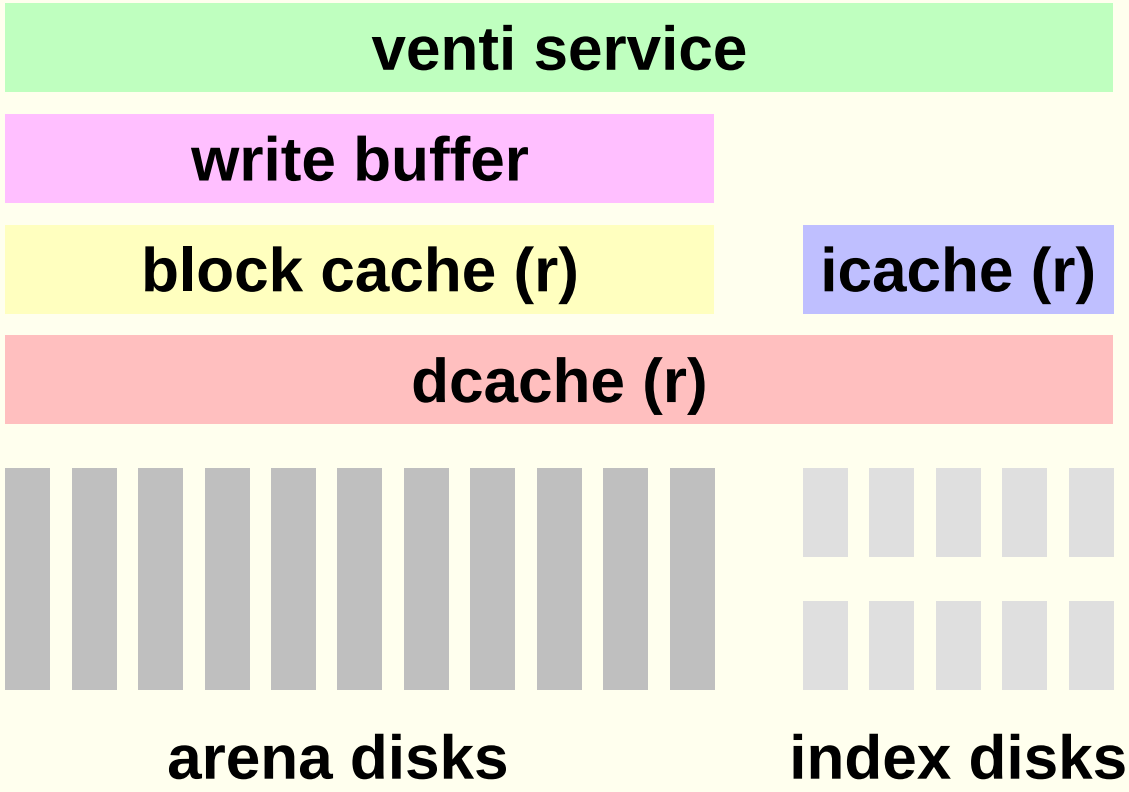
Warranty reflects quality (?)

- October 1, 2002: Seagate, Maxtor, and Western Digital all switched from 3Y to 1Y standard warranties
- seem to be back to 3Y on most drives now
- can still find 5Y warranties if you look hard
- our SATA have 5Y warranties

Prices reflect quality (?)

- longer warranty means higher prices
- SCSI means higher prices (4x)

Original venti architecture



- all caches read-only (write-through)

Optimization Plan

Arena updates

- write buffer, batching

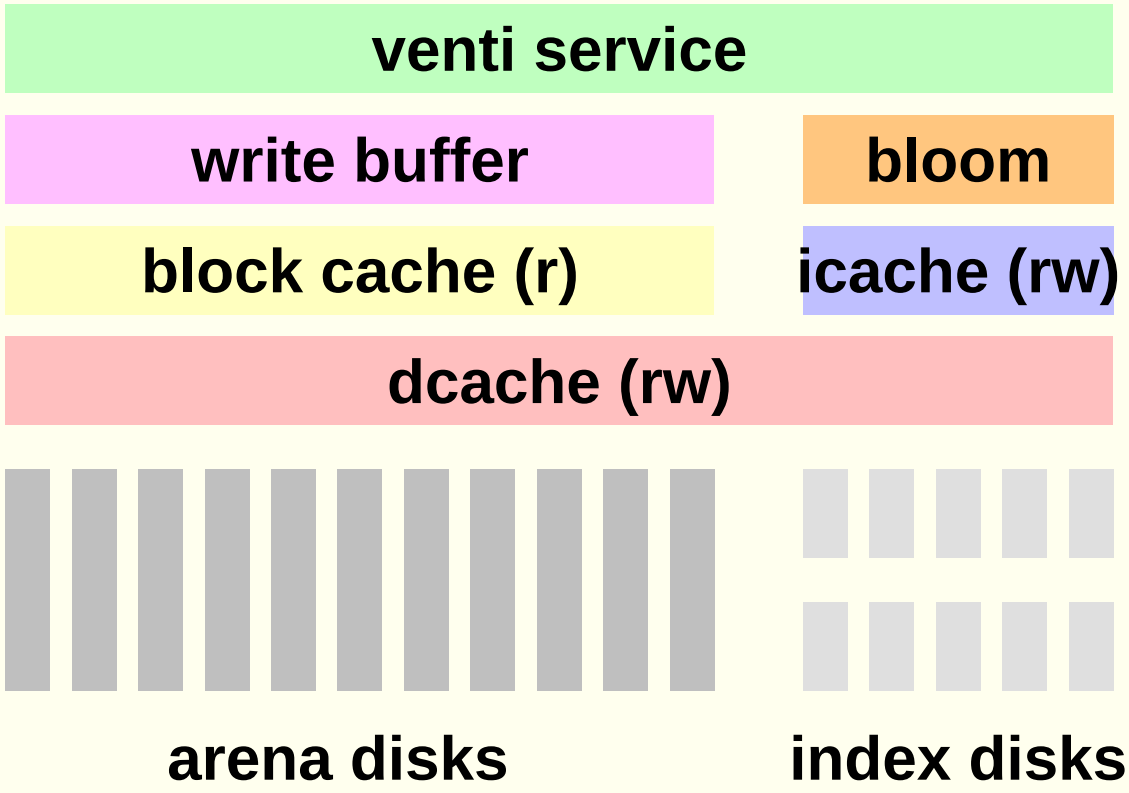
Index updates

- write buffer, batching

Common cases

- sequential reads of sequentially written data
- writes of fresh data

New venti architecture



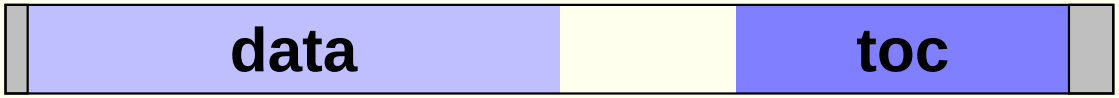
Arena updates

Delay and batch writes, queuing in memory

- added `sync` rpc
- modest buffer size (200MB)
- write in big sequential bursts

Write ordering

- arena data + table of contents first (committed)
- then index entries
- then tail stats



arena section (~1GB)

Index updates

Delay and batch writes, queueing in memory

- enormous buffer size – 1GB, 26M entries, 219GB in 8K blocks
- buffer holds individual entries to be added (40 bytes each), not disk blocks
- flush buffer by sequential updates passes over entire index, 4MB chunks at a time)
- (can update each index disk in parallel)

index section (~20GB)

Sequential reads

Notice sequential index lookups for blocks in same arena

- in response, load entire arena table of contents into index cache
- should make more fine-grained

Fresh writes

Bloom filter of all SHA1 hashes stored

- potentially large – 512MB
  32 hash bits/block, 0.7 overlap factor, 200M 16kB blocks, 1.5TB
  20 hash bits/block: 2.2TB
- might be too big, okay for now

Most fresh writes don't need to check the index

- chance of false positive 0.7^{32} when optimally full
- 1 in 100,000
- even lower when filter is mostly empty

Performance

Microbenchmarks

- writing fresh data, 25% random - 13 MB/s
- reading same w/ prefetching - 13 MB/s
- reading same w/o prefetching - 7 MB/s

Would have more numbers, but a disk failed last night

Bigger difference with fewer index disks

Performance

Fragmentation breaks sequential read hypothesis

- reading SHA1 hash list to start backup takes too long
- need to measure fragmentation + effects

  plenty of real data — need to process it

Performance

What about the backups? Two examples.

/disk/am2 - 22GB, 8kB blocks

- backup on 12/18/2003 - 53 kB/s avg venti rpc speed
- backup on 11/13/2005 - 282 kB/s avg venti rpc speed
  349 kB/s write speed (venti write rpc) 310 kB/s read speed (venti read rpc)
- could still be faster

Backup time (all of amsterdam, all disks)

- old system — median time 6.5 hours
  25% 4.5 hours, 75% 9 hours
- new system — consistently 2.25 hours
  with 2x data to back up
  dominated by time to read disks
  could be faster: run disks in parallel

Reliability

End-to-end SHA1 checks

- clients can (and do) compute SHA1 of blocks on read/write
  detected bad memory installed in machine

- Venti '`fsck`' can reconstruct arenas as much as possible
  SHA1s say whether we're right
  cleaned up aftermath of bad memory very easily

- 'seal' arena by recording SHA1 hash over entire arena
  provides simple future check that arena has not changed

Disk failures

- mirrored pairs of arena disks
- mirror knows to copy only new parts of log
- not old pieces that changed due to corruption

- no mirror for index - can rebuild (40 minutes yesterday)

Summary

Disks are out to get you.

Seeks hurt a lot.

Prefetching speeds positive lookups

Bloom filter speeds negative lookups

End-to-end data checks are good