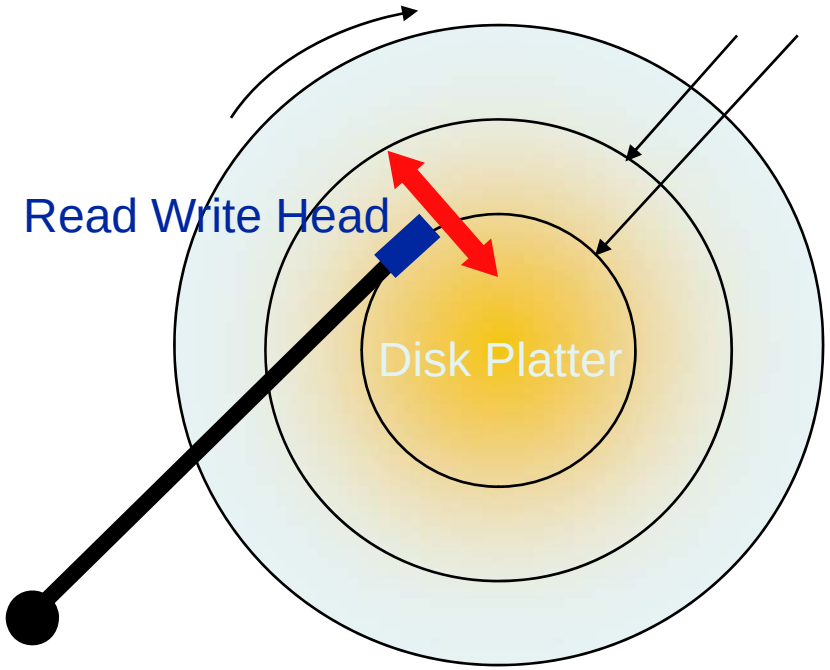
Comparison of Drive Technologies for High Transaction Databases

Wade Tuma

Founder, and CEO, Solid Data Systems

Delays in Rotating Disks.



- Seek Latency
  - Time Required To Move Heads To The Correct Track
- Rotational Latency
  - Time It Takes Data To Rotate Under The Head
- Typically 8 To 10 Msec Total On Random Accesses
- Combined, These Delays Are Glacial Compared To Speeds Of Modern Computers

Rotating Disk Random Reads

Seagate 15K HDD
Random Benchmarks

RANDOM READ BENCHMARK			
Block Size	Read IO/s	Read MB/s	Avg Service Time - ms
512B	185	0.09	10.4
1K	185	0.18	10.5
2K	182	0.37	10.5
4K	175	0.70	10.8
8K	176	1.41	10.9
16K	172	2.75	11.0
32K	170	5.44	11.0
64K	152	9.76	11.0
128K	132	16.96	11.2

- Disks work poorly on small blocks
- Work better on big blocks
- Rotating disks look much the same on writes as on reads

Understanding the scale of Database IO

- 4K Block At 100MB/Sec Takes 40 Micro Sec.
- 8K Block At 100MB/Sec Takes 80 Micro Sec.

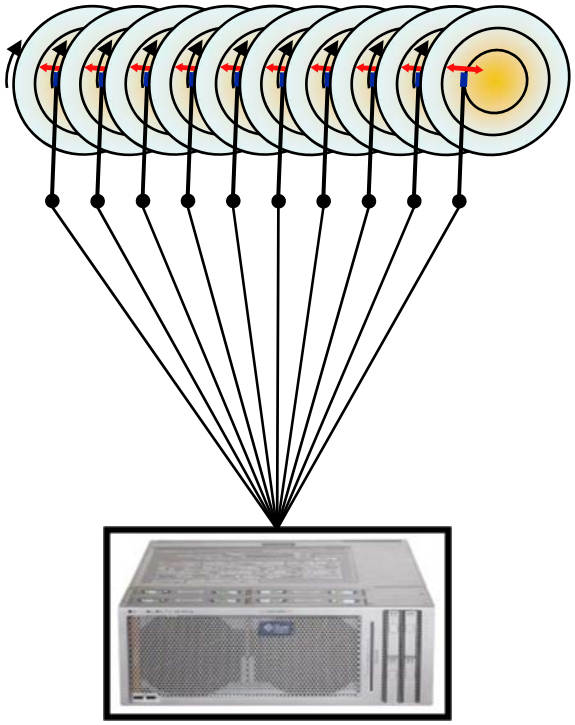
- So It Takes ~ 8-10 Millisecond. To Get 80 Micro Sec. Of Data When Accesses Are Purely Random.

- Drive Is Delivering Data About 1% Of The Time.
- Often More Data Is Brought In Than Was Requested. (Read Ahead)

Implications of Disk Latency

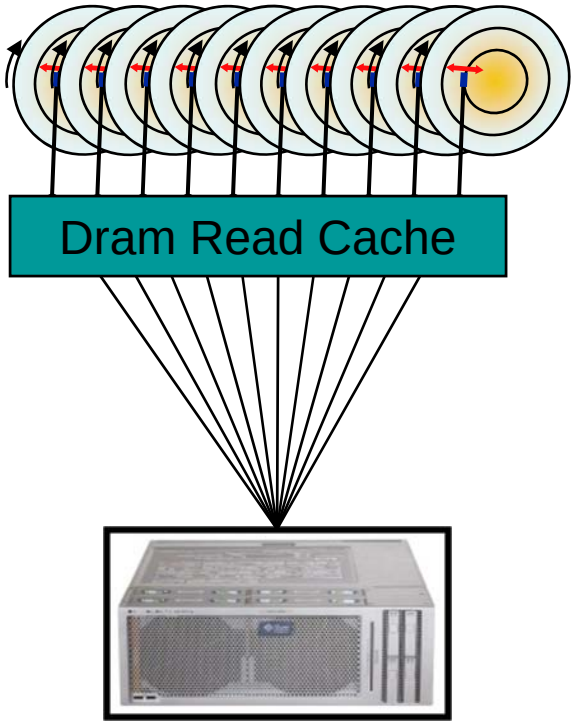
- Unless The Desired Data Is Stored Elsewhere, Server Is Going To Wait ~ 10 Milliseconds To Get The Data.
  - Caches Are A "Store Elsewhere" Option
    - In The Disk Array
    - In The Server
- Paralleling Disks Multiplies The Potential Number Of Access, (Increases IOPS) But *Does Not Improve Latency*
  - ***Also Requires An Application And Database Capable Of Reading And Writing In Parallel***

Parallel IOs



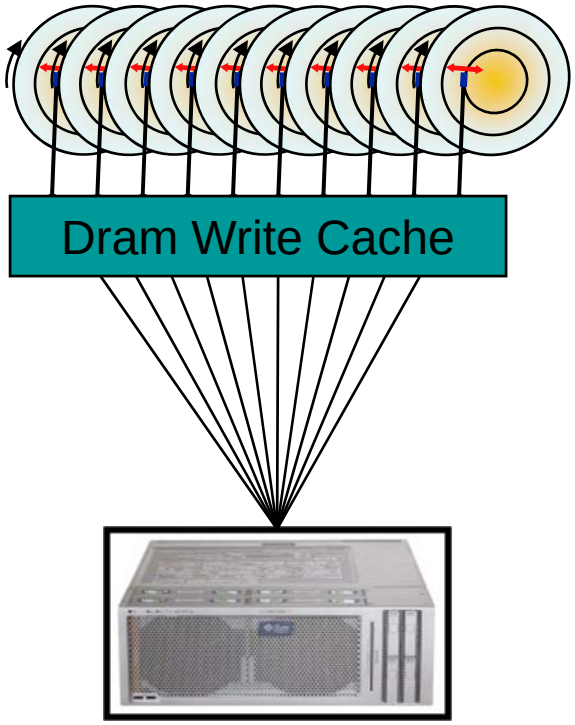
- 10 Drives, 10 Times The IOPS
- No Improvement In Latency
- IO Requests Remain Active In The Server For The Full Latency Period
- Can The App And Database Really Issue 10 Independent IO Requests?

Read Cache



- Read Cache Is Dram Memory
- Small Percentage Of Total Disk Capacity.
- Data Is Read From Cache To Avoid Disk Latency
- Read Cache Requires Mapping Table And High Speed Processor
- Odds Of Data Being In Cache Is App (And Everything Else) Dependant
- Fails To Help When Data Is Rarely In Cache

Write Cache



- Write Cache Is Dram Memory
- Small Percentage Of Total Disk Capacity.
- Data Is Written To Cache To Avoid Waiting For Disk Latency
- Write Cache Requires Mapping Table And High Speed Processor
- Fails To Help Once Cache Capacity Is Exceeded

Other end of the Scale

Solid Data Model SD3000 Dram SSD Random Benchmarks

RANDOM READ BENCHMARK			
Block Size	Read IO/s	Read MB/s	Avg Service Time - ms
512B	23584	11.79	0.036
1K	22068	22.07	0.033
2K	17066	34.13	0.044
4K	12641	50.57	0.064
8K	8299	66.40	0.119
16K	3980	63.69	0.414
32K	2689	86.08	0.724
64K	1288	82.45	2.526
128K	685	87.76	7.137

- Capable Of Thousands Of Accesses Per Sec.
- Uses A Significant Percentage Of Buss Bandwidth.
- Server Latency Becomes Significant
- Don't Need Caches
- *Don't Need To Operate In Parallel*

A New Option, Flash SSD

M-Systems Model FDD 3.5" Flash SSD Random Benchmarks

RANDOM READ BENCHMARK			
Block Size	Read IO/s	Read MB/s	Avg Service Time - ms
512B	1315	0.66	1.4
1K	1217	1.22	1.5
2K	1206	2.41	1.5
4K	1075	4.30	1.7
8K	906	7.26	2.0
16K	666	10.66	2.8
32K	447	14.33	4.2
64K	322	20.62	5.9
128K	204	26.16	9.5

- Read Performance Between Dram SSD and Rotating Disks

Read and Write Performance

Solid Data Model SD3000 Dram SSD
Random Benchmarks

RANDOM READ BENCHMARK

Block Size	Read IO/s	Read MB/s	Avg Service Time - ms
512B	23584	11.79	0.036
1K	22068	22.07	0.033
2K	17066	34.13	0.044
4K	12641	50.57	0.064
8K	8299	66.40	0.119
16K	3980	63.69	0.414
32K	2689	86.08	0.724
64K	1288	82.45	2.526
128K	685	87.76	7.137

RANDOM WRITE BENCHMARK

Block Size	Write IO/s	Write MB/s	Avg Service Time - ms
512B	20768	10.38	0.046
1K	17391	17.39	0.053
2K	14988	29.98	0.057
4K	11365	45.46	0.079
8K	7400	59.20	0.150
16K	3582	57.32	0.511
32K	2371	75.87	0.932
64K	1140	72.96	3.225
128K	604	77.32	9.187

M-Systems Model FDD 3.5" Flash SSD
Random Benchmarks

RANDOM READ BENCHMARK

Block Size	Read IO/s	Read MB/s	Avg Service Time - ms
512B	1315	0.66	1.4
1K	1217	1.22	1.5
2K	1206	2.41	1.5
4K	1075	4.30	1.7
8K	906	7.26	2.0
16K	666	10.66	2.8
32K	447	14.33	4.2
64K	322	20.62	5.9
128K	204	26.16	9.5

RANDOM WRITE BENCHMARK

Block Size	Write IO/s	Write MB/s	Avg Service Time - ms
512B	22	0.01	92.5
1K	22	0.02	91.7
2K	21	0.04	92.3
4K	21	0.09	94.5
8K	21	0.17	92.5
16K	21	0.34	93.7
32K	21	0.68	102.1
64K	19	1.23	106.7
128K	18	2.37	113.2

Seagate 15K HDD
Random Benchmarks

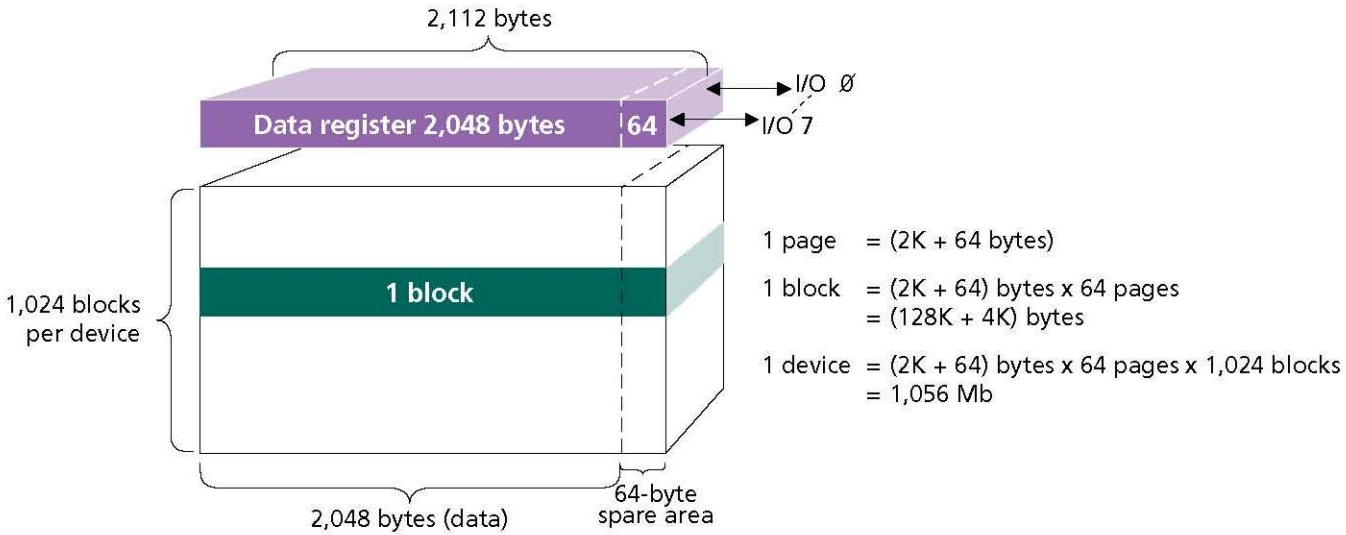
RANDOM READ BENCHMARK

Block Size	Read IO/s	Read MB/s	Avg Service Time - ms
512B	185	0.09	10.4
1K	185	0.18	10.5
2K	182	0.37	10.5
4K	175	0.70	10.8
8K	176	1.41	10.9
16K	172	2.75	11.0
32K	170	5.44	11.0
64K	152	9.76	11.0
128K	132	16.96	11.2

RANDOM WRITE BENCHMARK

Block Size	Write IO/s	Write MB/s	Avg Service Time - ms
512B	290	0.14	6.7
1K	290	0.29	6.5
2K	283	0.57	6.9
4K	280	1.12	6.3
8K	284	2.27	6.2
16K	264	4.23	6.3
32K	237	7.58	6.6
64K	211	13.51	6,5
128K	183	23.48	8.0

Structure of Flash Memory



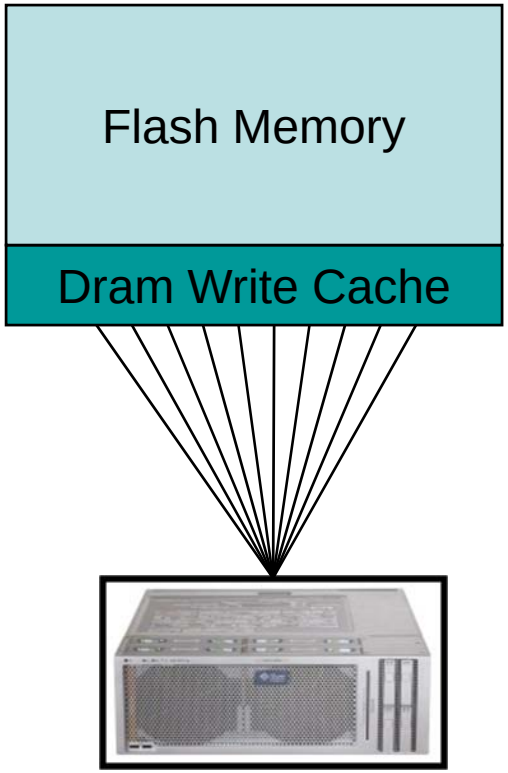
- To Write To This Memory We Need To Read The Entire Block (132KBytes) To A Scratch Area.
- Erase The Entire Block (2 Msec.)
- Modify Our 4k Or 8k Data In The Scratch Area Then Rewrite All 132KBytes Back To Flash Block.
- Effectively, Regardless Of The Block Size We're Really Reading Then Writing An Entire 132kbyte Block

Why So Slow on Writes?

RANDOM WRITE BENCHMARK			
Block Size	Write IO/s	Write MB/s	Avg Service Time - ms
512B	22	0.01	92.5
1K	22	0.02	91.7
2K	21	0.04	92.3
4K	21	0.09	94.5
8K	21	0.17	92.5
16K	21	0.34	93.7
32K	21	0.68	102.1
64K	19	1.23	106.7
128K	18	2.37	113.2

- Data is read from pages
- Data is written in blocks consisting of many pages
- Before data can be written, old data must be erased
- Erasing takes several milliseconds.

Improving the Write Performance of Flash



- Write Cache Is Dram Memory
- Small Percentage Of Total Flash Capacity.
- Data Is Written To Cache To Avoid Waiting For Flash Latency
- Write Cache Requires Mapping Table And High Speed Processor
- Fails To Help Once Cache Capacity Is Exceeded

Use for Reads Only?

50/50 RANDOM READ/WRITE BENCHMARK			
Block Size	R/W IO/s	R/W MB/s	Avg Service Time - ms
512B	58	0.03	33.9
1K	54	0.05	35.5
2K	55	0.11	38.8
4K	55	0.22	34.1
8K	54	0.44	38.7
16K	49	0.80	45.7
32K	45	1.45	45.8
64K	45	2.89	46.6
128K	41	5.28	52.7

- If You Can Guarantee Read Only Access Then You Get Full Flash Read Performance
- Most Operating Systems Write Even On Reads (Last Access Time For Example)

- Even If The Write Percentage Was only 5% The Total Number Of 4K IOPs Would Fall From 1075 In Pure Reads To About 300