Disk Array Architectures

Walt Burkhard

burkhard@cs.ucsd.edu

http://www-cse.ucsd.edu/~burkhard

http://www-cse.ucsd.edu/groups/gemini

Gemini Storage Systems Laboratory
Computer Science and Engineering Department
University of California, San Diego
9500 Gilman Drive
La Jolla, California 92093-0114

- RAID Level 5: Left Symmetric Data Layout

disk 0	disk 1	disk 2	disk 3	disk 4
A.0	A.1	A.2	A.3	B.0
B.1	B.2	B.3	C.0	C.1
C.2	C.3	D.0	D.1	D.2
D.3	E.0	E.1	E.2	E.3
pE	pD	pC	pB	pA

- Stripes Reliability
- Parallel reads.
- Writes.
- Degraded Performance (single disk failure)

  Stripe-unit accesses – reads or writes

  Both increase the workload.

- Improvements

  Performance, Availability / Reliability, Cost

Disk Array Architectures II

Workload: *R reads* **and** *W writes* n disks

- **Small Operation Fault Free Performance:**

  workload: $R + 4W$ operations for small accesses.
  workload is evenly divided among the disks.

Degraded Mode designates single disk failure

- **Small Operation Degraded Mode Performance:**

  Read workload: $2\,R(n\text{-}1)/n$ operations.

  Write workload: $4W + W(n\text{-}7)/n$ operations.

- **Performance Consequences:**

  Degraded mode is inevitable. To meet the needs of our clients during the degraded mode, we must limit the utility of the system during fault free operation.

Disk Array Architectures III

Declustering improves the degraded mode performance.

- **Disk Array Declustering**

  The client data is spread over additional disks; n disks with stripe width k, typically $n > k$.

Example: $n = 6$, $k = 5$.

disk 0	disk 1	disk 2	disk 3	disk 4	disk 5
A.0	A.1	A.2	A.3	B.0	B.1
B.2	B.3	C.0	C.1	C.2	C.3
E.1	D.0	D.1	D.2	D.3	E.0
F.3	E.2	E.3	F.0	F.1	F.2
pC	pF	pB	pE	pA	pD

Disk Array Architectures IV

- ## Data Layout Goals

Error Correction	stripe units of each stripe mapped to distinct disks.
Distributed Parity	each disk contains the same number of parity stripe units.
Distributed Reconstruction	the incremental workload is evenly distributed over the surviving disks.
Efficient Mapping	run-time efficient requiring little additional memory to represent the data layout.
Parallel Reads	contiguous client data mapped to distinct disks as much as possible.
Large Writes	contiguous client data mapped to single stripes as much as possible.

Parity Declustering I

- **Balanced Incomplete Block Designs (BIBD)**

- (b, n, r, k, λ)

1. there are b subsets of $\{0, 1, \ldots, n\text{-}1\}$ of size k,
2. each element in $\{0, 1, \ldots, n\text{-}1\}$ appears in r subsets, and
3. each pair of elements of $\{0, 1, \ldots, n\text{-}1\}$ is in λ subsets.

conditions:

$$n \cdot r = b \cdot k$$

$$r\ (k\text{-}1) = \lambda (n - 1)$$

Data Layout I

disk 0	disk 1	disk 2	disk 3
A.0	A.1	pA	pB
B.0	B.1	C.1	pC
C.0	D.0	D.1	pD
E.0	pE	E.1	F.1
F.0	pF	pG	G.1
G.0	H.0	pH	H.1
pI	I.0	I.1	J.1
pJ	J.0	K.0	K.1
pK	pL	L.0	L.1

0	1	2
0	1	3
0	2	3
1	2	3

- Our Goals

  Simple (easy to specify) data layouts.

  Run-time efficient access calculations.

Parity Declustering Simulation I

- RAIDframe simulator

Workload

Access sizes:	8,24,48,72,120,133,168,192,240,288,336KB
Concurrency:	1,2,4,8,19,15,20,25 clients
Alignment:	stripe unit boundaries

random accesses uniformly distributed over address space

Disks HP 2247

Capacity:	1.03GB
Geometry:	1981 cylinders, 13 heads, 8 zones
Average seek time:	10 ms
Rotational speed:	5400 rpm
Queueing discipline:	shortest seek time first

Data Layout

Stripe unit size:	8KB
Number of disks, n:	13
Stripe width, k:	4 stripe units; RAID 5 is 13

Parity Declustering II

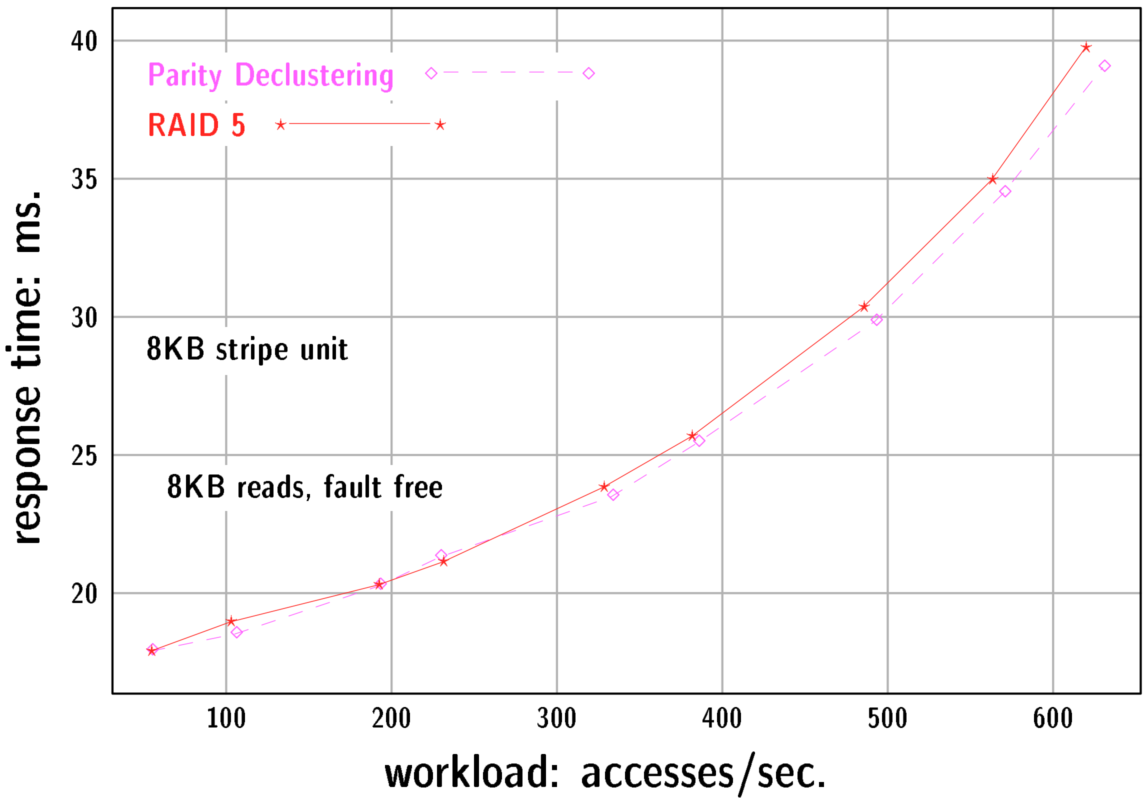
- **data layout schema** 13 disks, stripe width 4.

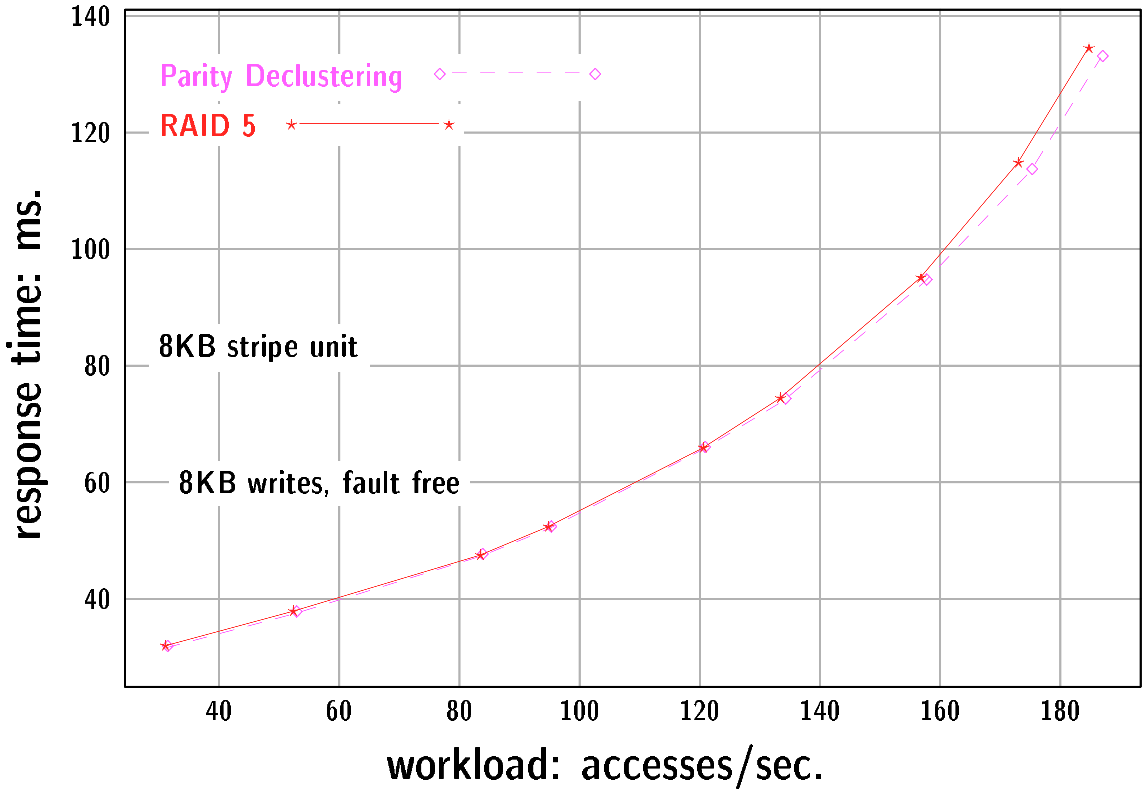
0	1	3	9
1	2	4	10
2	3	5	11
3	4	6	12
0	4	5	7
1	5	6	8
2	6	7	9
3	7	8	10
4	8	9	11
5	9	10	12
0	6	10	11
1	7	11	12
0	2	8	12

- **Data Layout Goals**

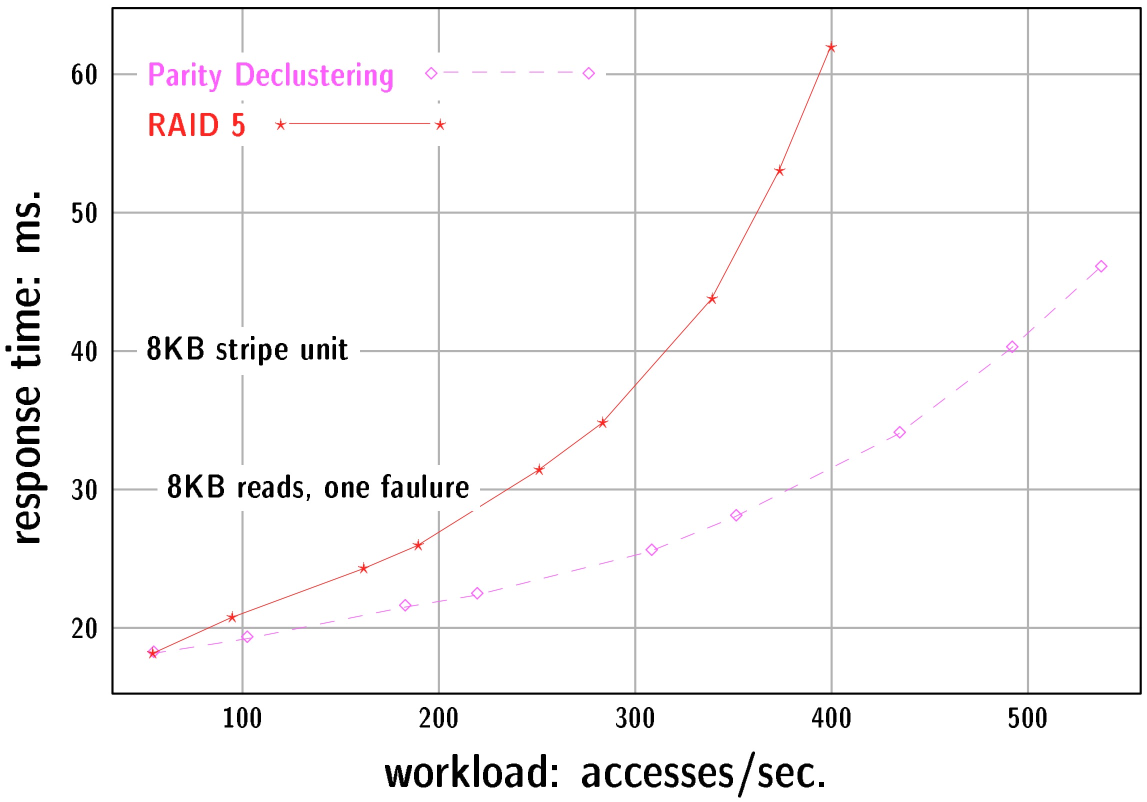
  all satisfied except parallel reads!

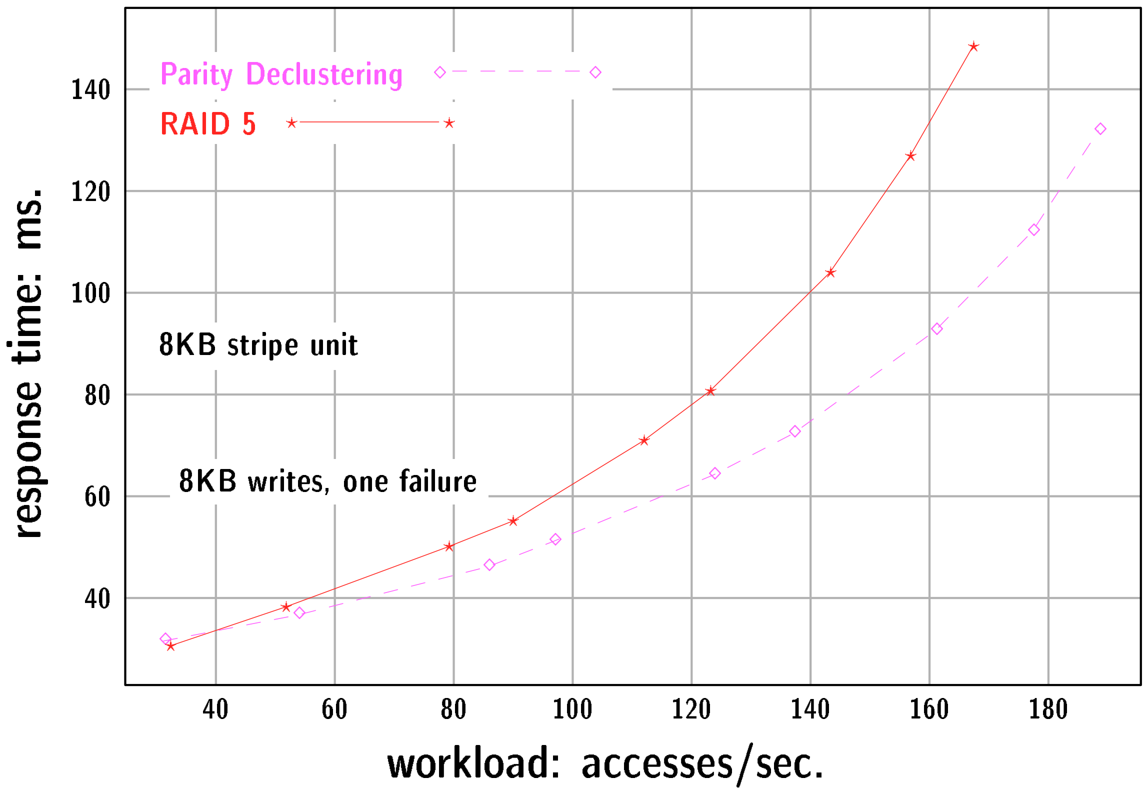
Parity Declustering III





Parity Declustering IV





Disk Array Architectures V

- **Generic Workloads**

Datum:	Guillermo Alvarez Walt Burkhard Flaviu Cristian
Prime:	Guillermo Alvarez Walt Burkhard Larry Stockmeyer Flaviu Cristian
Permutation Development:	Thomas Schwarz Walt Burkhard Jesse Steinberg

- **Video Workloads**

Segmented Information Dispersal:	Ariel Cohen Walt Burkhard

- ISCA'97

- Complete block designs

- Binomial number system for $0 \le s < \binom{n}{k}$

$$s = \binom{d_0}{1} + \binom{d_1}{2} + \cdots + \binom{d_{k-1}}{k}$$

with $0 \le d_0 < d_1 < d_2 < \cdots < d_{k-1} < n$.

For $0 \le s < \binom{n}{k}$ and $0 \le i \le d_i < d_{i+1} < n$,
the binomial number system mapping

$$s \mapsto \{ d_0, d_1, \cdots , d_{k-1} \}$$

is a bijection.

Datum II

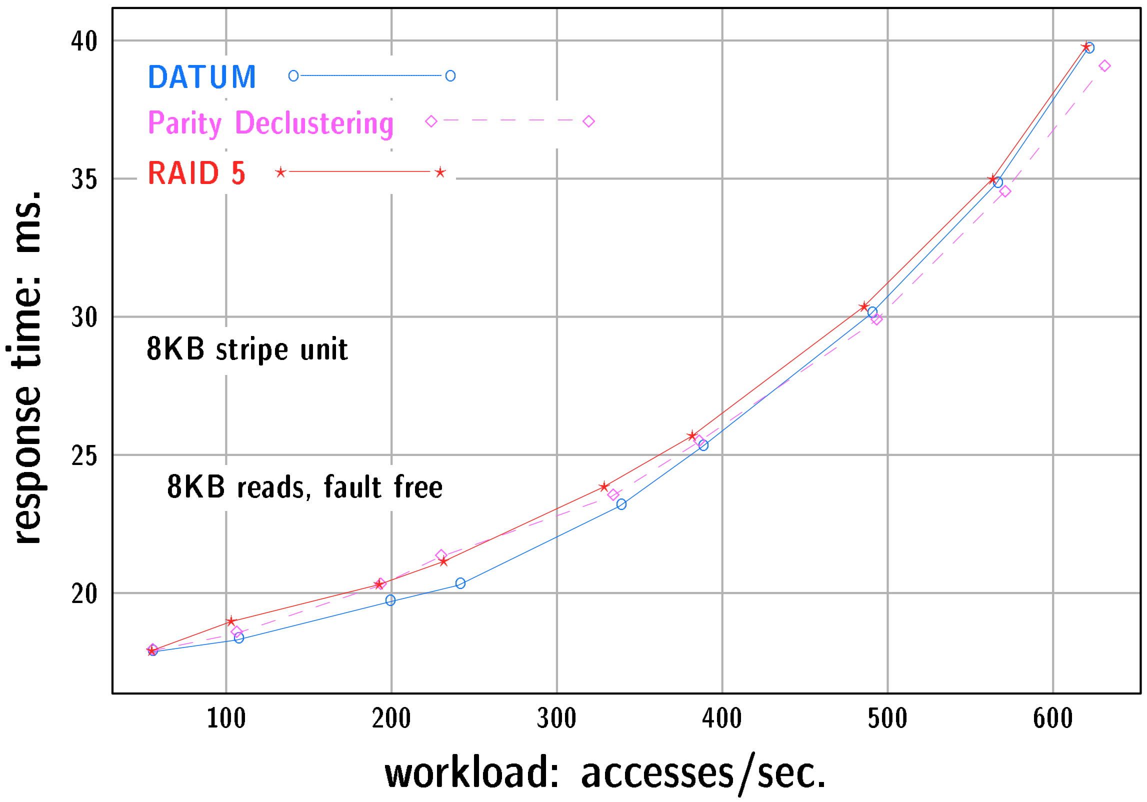
- **data layout schema** 13 disks, stripe width 4.

0	1	2	3
0	1	2	4
0	1	3	4
	⋮		
9	10	11	12

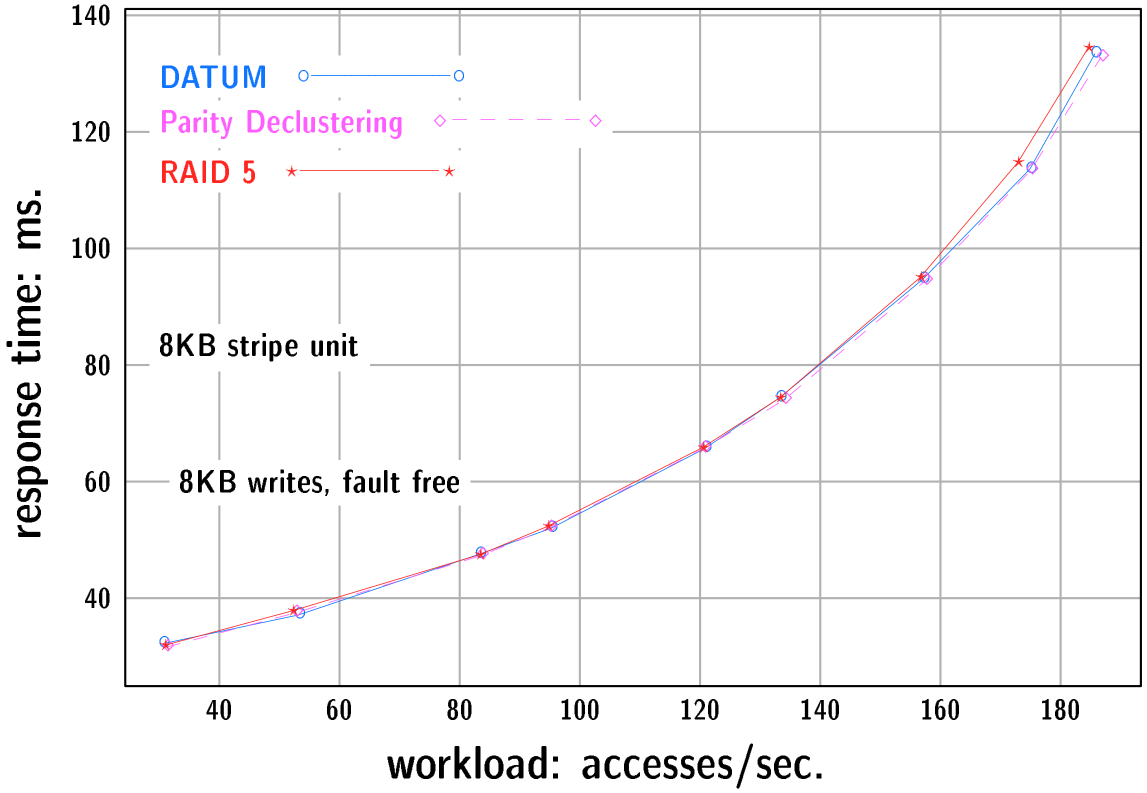
- **Data Layout Goals**

  all satisfied except parallel reads!

Datum III

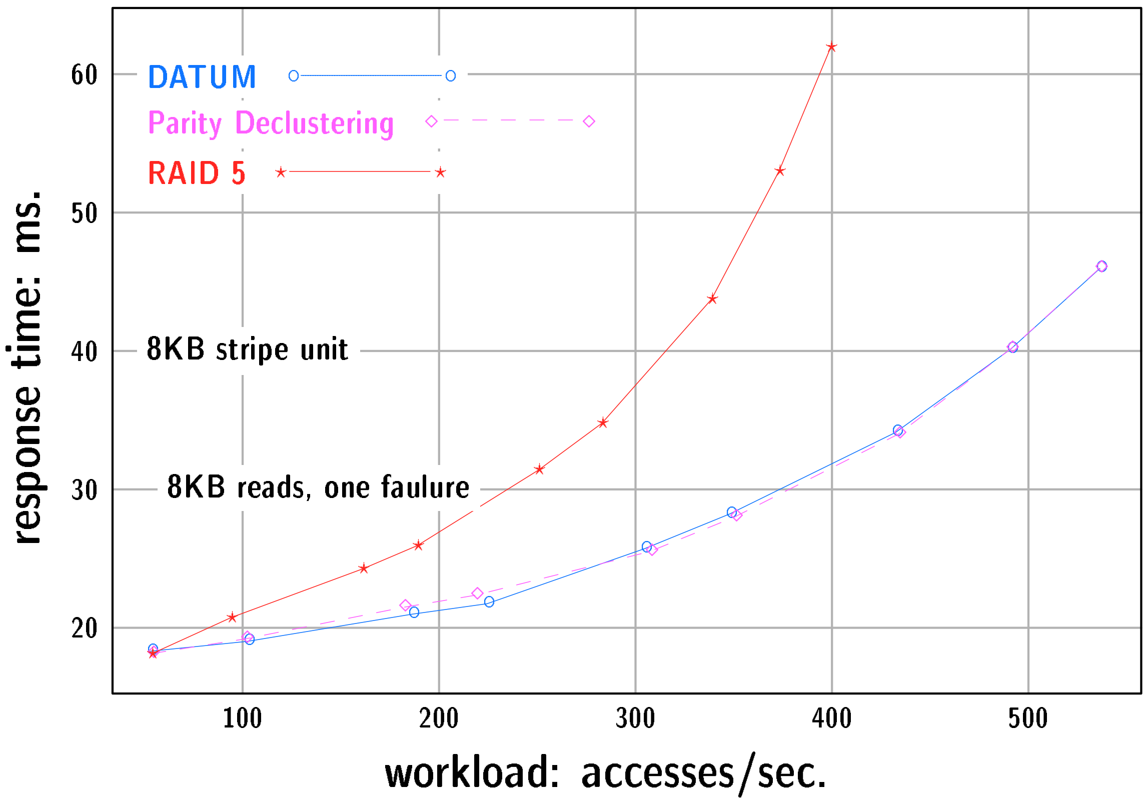


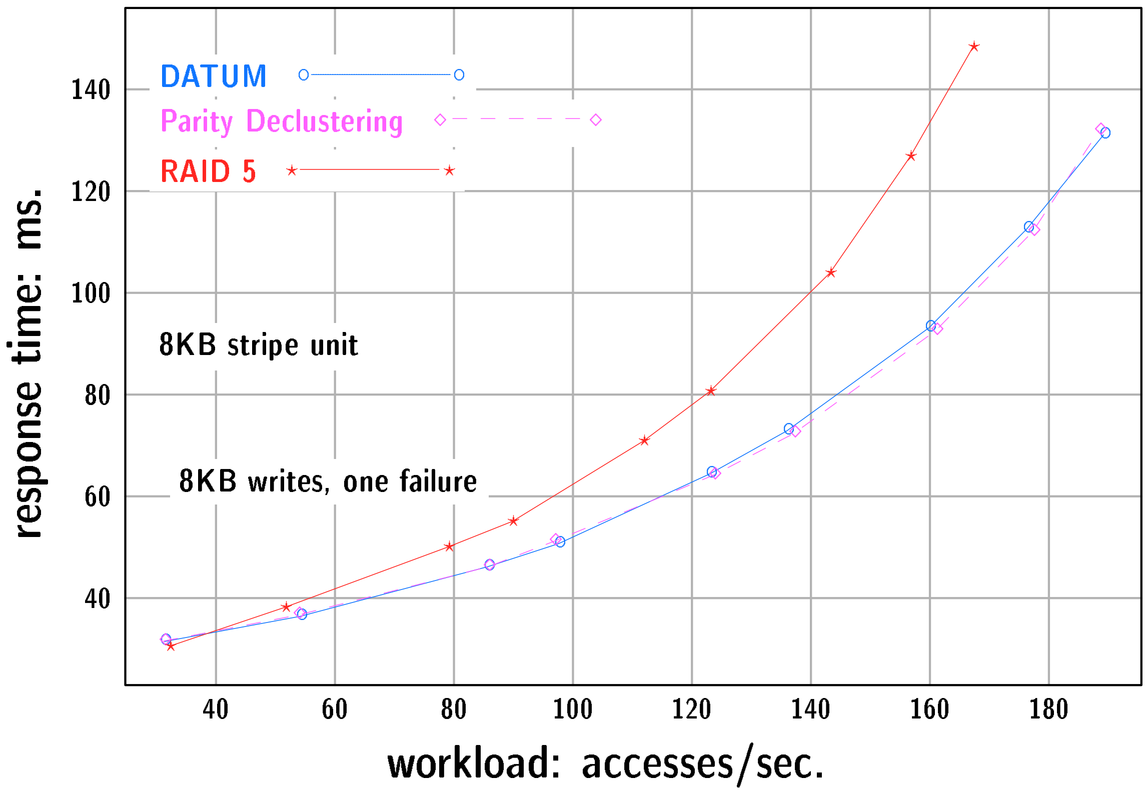
DATUM
Parity Declustering
RAID 5
8KB stripe unit
8KB reads, fault free
response time: ms.
workload: accesses/sec.
20
25
30
35
40
100
200
300
400
500
600



DATUM
Parity Declustering
RAID 5
8KB stripe unit
8KB writes, fault free
response time: ms.
workload: accesses/sec.
40
60
80
100
120
140
40
60
80
100
120
140
160
180

Datum IV


DATUM
Parity Declustering
RAID 5
8KB stripe unit
8KB reads, one faulure
response time: ms.
workload: accesses/sec.
20
30
40
50
60
100
200
300
400
500


DATUM
Parity Declustering
RAID 5
8KB stripe unit
8KB writes, one failure
response time: ms.
workload: accesses/sec.
40
60
80
100
120
140
40
60
80
100
120
140
160
180

Disk Array Architecture VI

- Existential Result

n disks k stripe width

r stripe units per disk

Theorem: There is a data layout satisfying the data layout goals with parameters n, k, and r if and only if

a. k divides r,

b. $(n\text{-}1)$ divides $r(k\text{-}1)$,

and one of the following conditions holds –

i) $k = n$,

ii) $k = n\text{-}1$,

iii) $k = 2$,

iv) $k = 3$ & $n = 5$ or $n = 7$,

v) $k = 4$ & $n = 7$.

Prime I

- ISCA'98

disk 0	disk 1	disk 2	disk 3	disk 4
A.0	A.1	A.2	B.0	B.1
B.2	C.0	C.1	C.2	D.0
D.1	D.2	E.0	E.1	E.2
pE	pB	pD	pA	pC
F.0	G.0	F.1	G.1	F.2
G.2	H.2	H.0	I.0	H.1
I.1	J.1	I.2	J.2	J.0
pJ	pF	pG	pH	pI
		...		

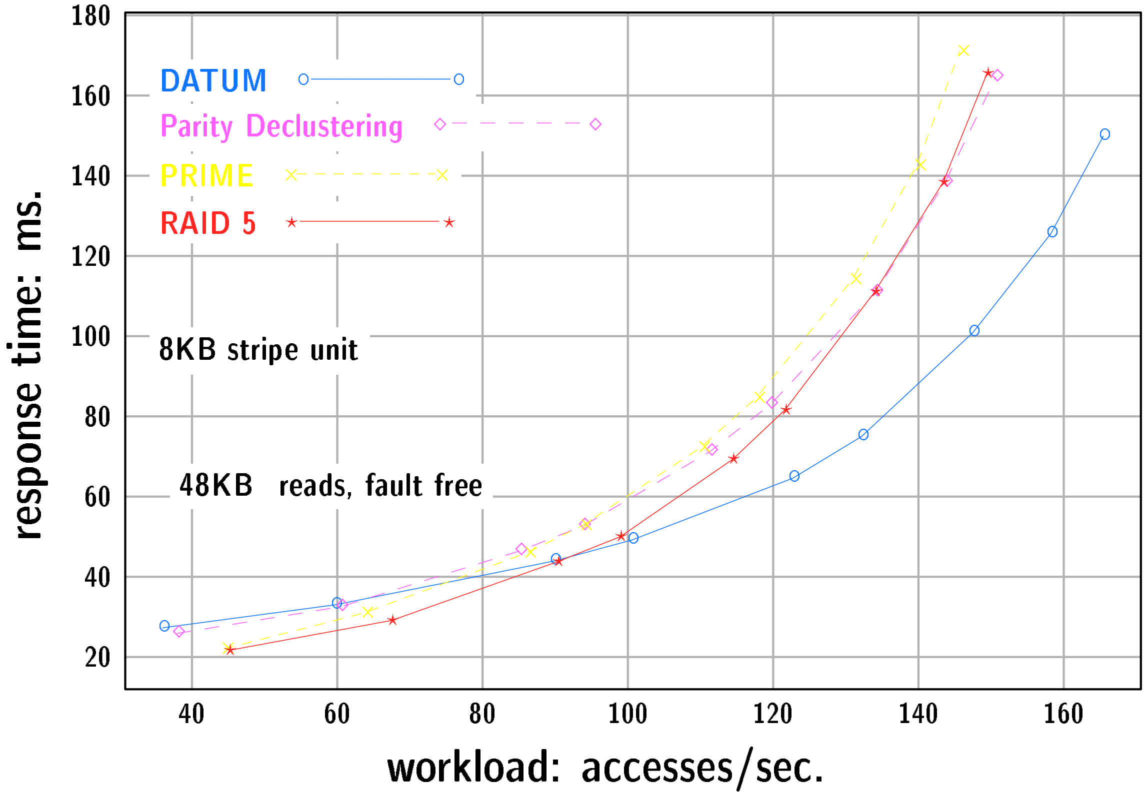
- Data Layout Goals

  Almost optimal read of contiguous user stripe units.

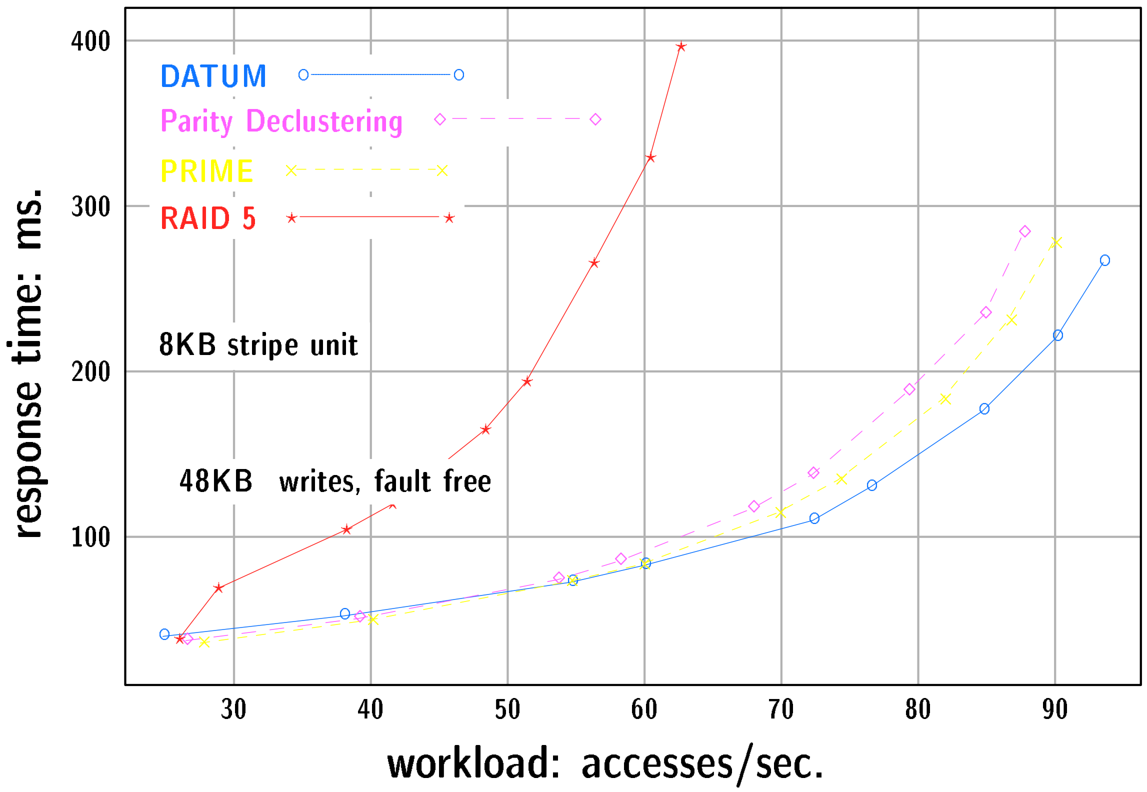
  Large write optimization.

  Evenly distributed parity and reconstruction workload.

Prime II

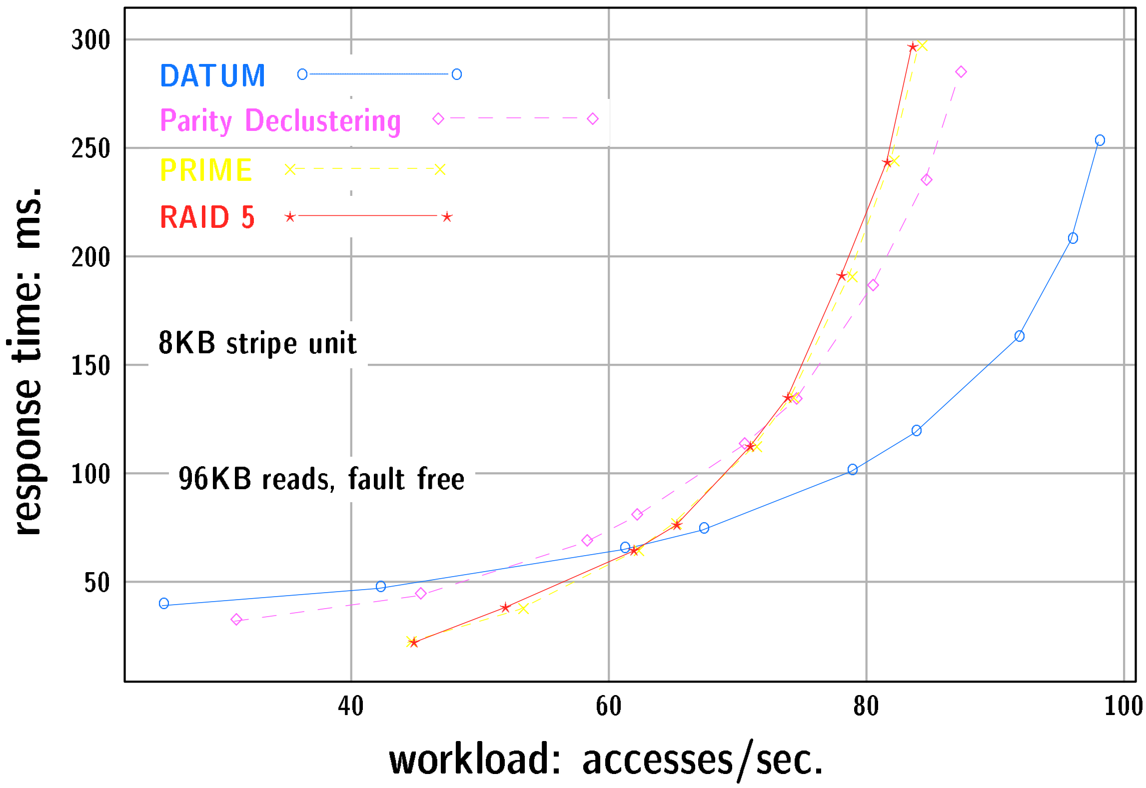


DATUM
Parity Declustering
PRIME
RAID 5
8KB stripe unit
48KB reads, fault free
response time: ms.
workload: accesses/sec.
180
160
140
120
100
80
60
40
20
40
60
80
100
120
140
160

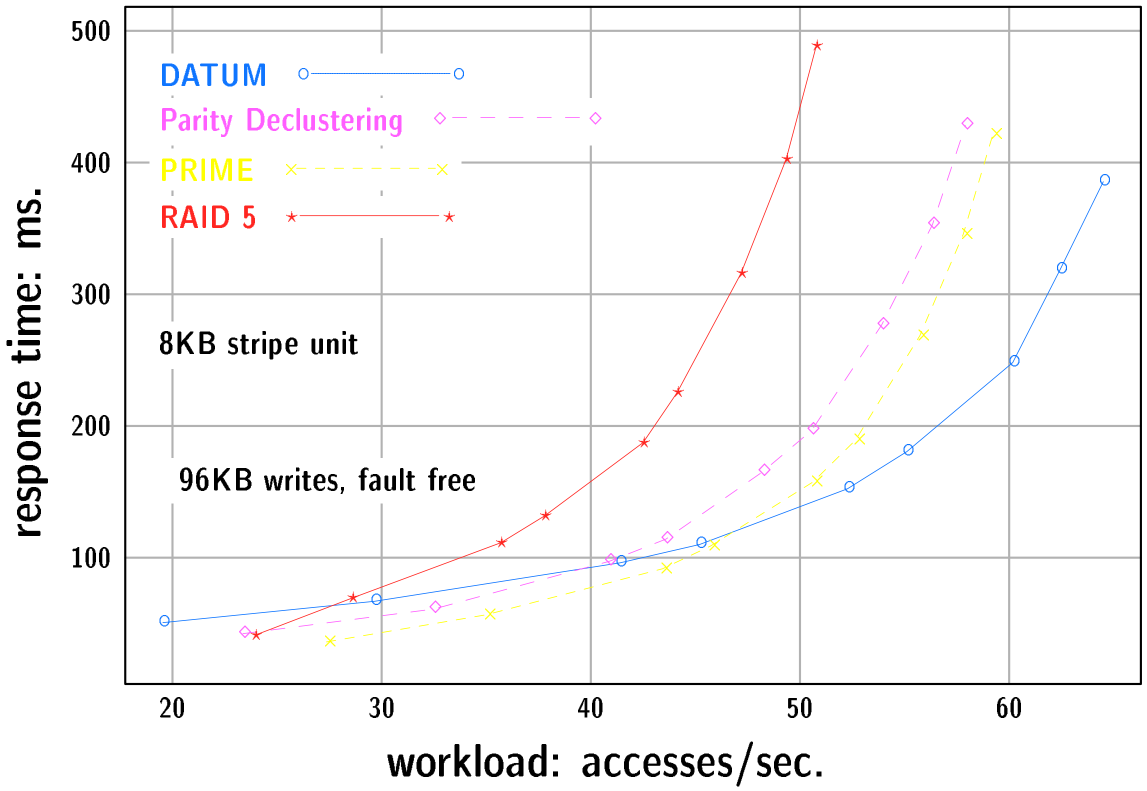


DATUM
Parity Declustering
PRIME
RAID 5
8KB stripe unit
48KB writes, fault free
response time: ms.
workload: accesses/sec.
400
300
200
100
30
40
50
60
70
80
90

Prime III

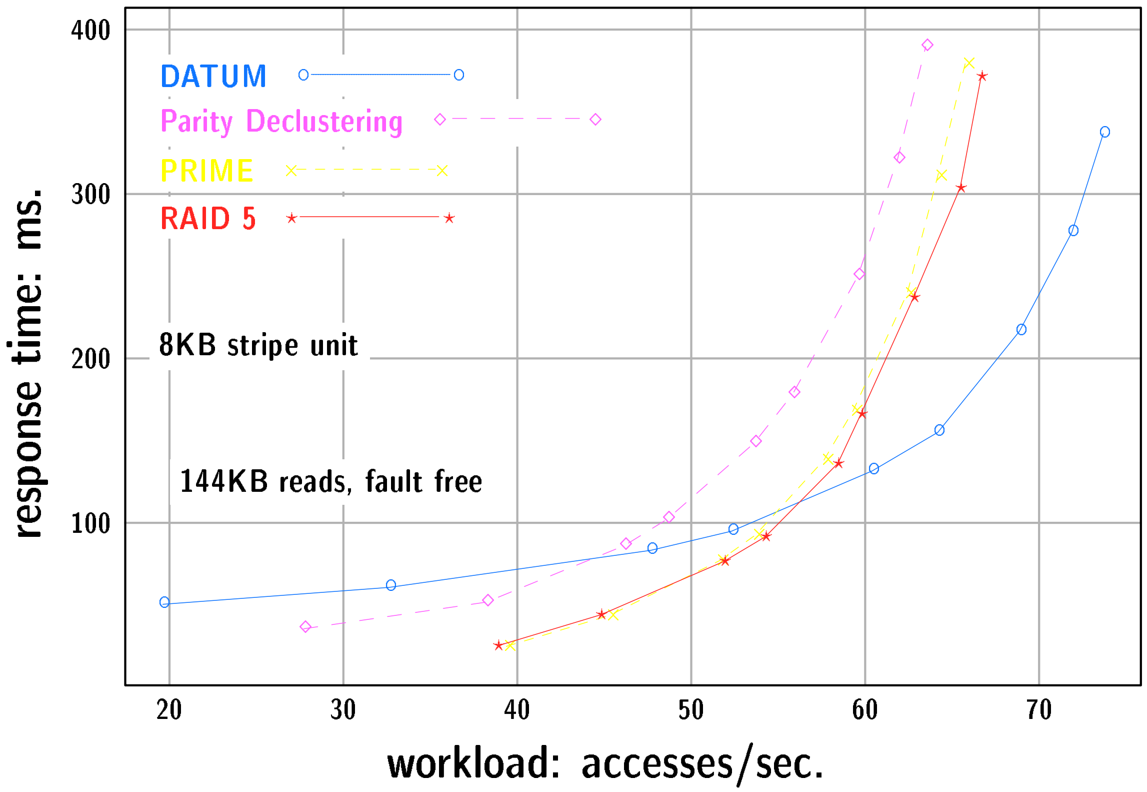


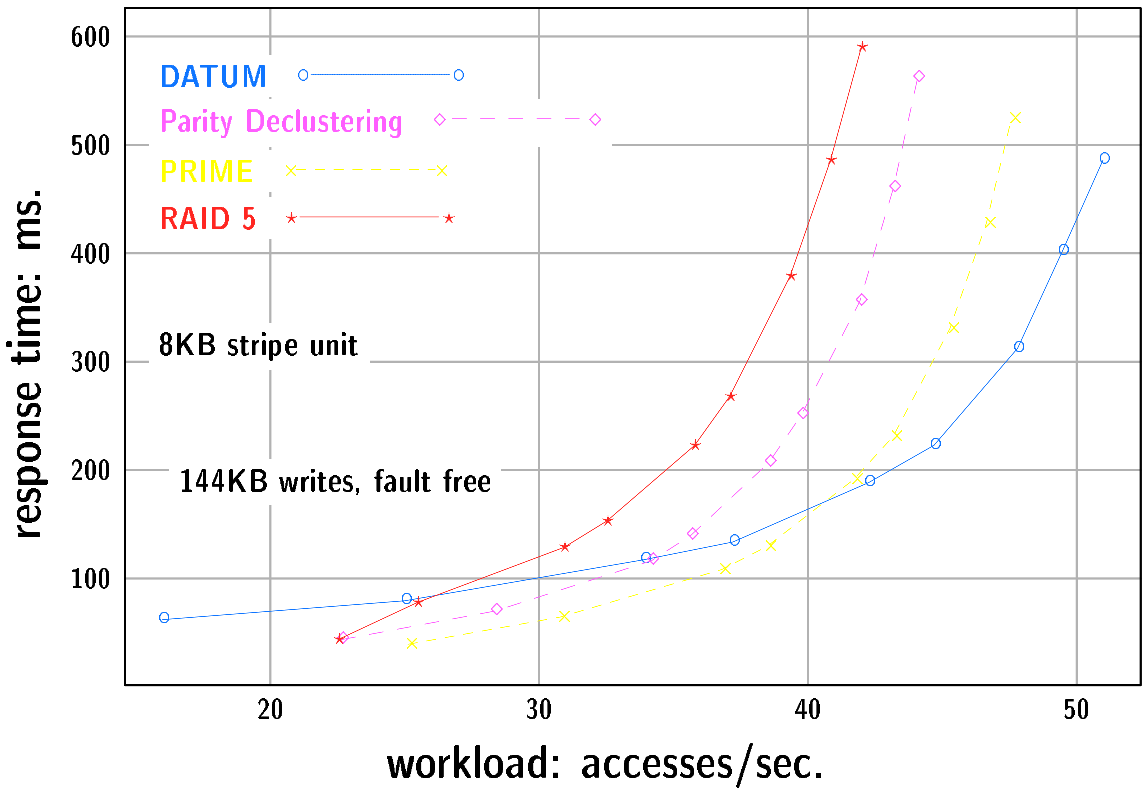
DATUM
Parity Declustering
PRIME
RAID 5
8KB stripe unit
96KB reads, fault free
response time: ms.
workload: accesses/sec.
300
250
200
150
100
50
40
60
80
100



DATUM
Parity Declustering
PRIME
RAID 5
8KB stripe unit
96KB writes, fault free
response time: ms.
workload: accesses/sec.
500
400
300
200
100
20
30
40
50
60

Prime IV


DATUM
Parity Declustering
PRIME
RAID 5
8KB stripe unit
144KB reads, fault free
response time: ms.
workload: accesses/sec.
400
300
200
100
20
30
40
50
60
70


DATUM
Parity Declustering
PRIME
RAID 5
8KB stripe unit
144KB writes, fault free
response time: ms.
workload: accesses/sec.
600
500
400
300
200
100
20
30
40
50

Permutation Development (PDDL) I

- Economical Sparing

disk 0	disk 1	disk 2	disk 3	disk 4
S	A.0	B.0	B.1	A.1
C.1	S	C.0	D.0	D.1
F.1	E.1	S	E.0	F.0
H.0	H.1	G.1	S	G.0
I.0	J.0	J.1	I.1	S

0	1	4	2	3
1	2	0	3	4
2	3	1	4	0
3	4	2	0	1
4	0	3	1	2

- Optimal reconstruction workload possible

- Very small layout description

Workload Calculation

0	1	2	8	3	5	7	4	6	9
	0	1	7	0	2	4	0	2	5
	9	0	6	8	0	2	8	0	3
	3	4	0	6	8	0	5	7	0

1	2	3	4	5	6	7	8	9
1	3	2	2	2	2	2	3	1

0	1	2	4	3	7	8	5	6	9
	0	1	3	0	4	5	0	1	4
	9	0	2	6	0	1	9	0	3
	7	8	0	5	9	0	6	7	0

1	2	3	4	5	6	7	8	9
3	1	2	2	2	2	2	1	3

Nearly Resolvable Designs

Permutation Development (PDDL) III

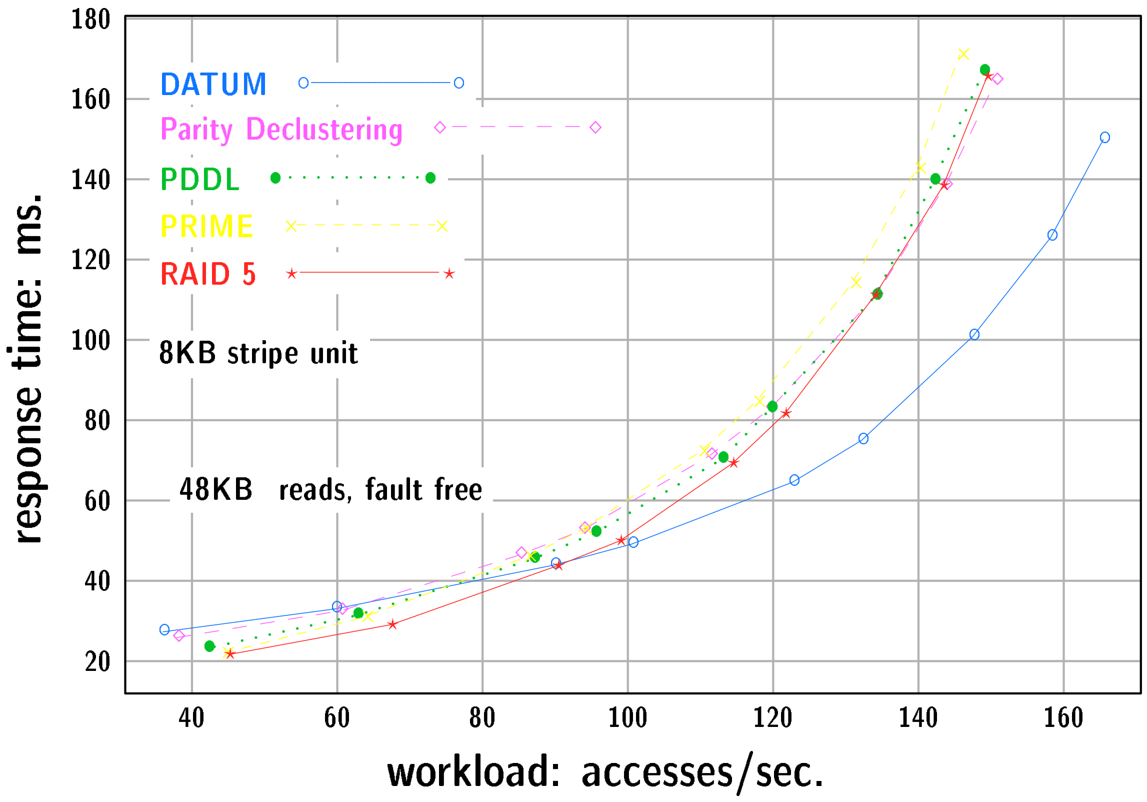
- **data layout scheme** 13 disks, stripe width 4.

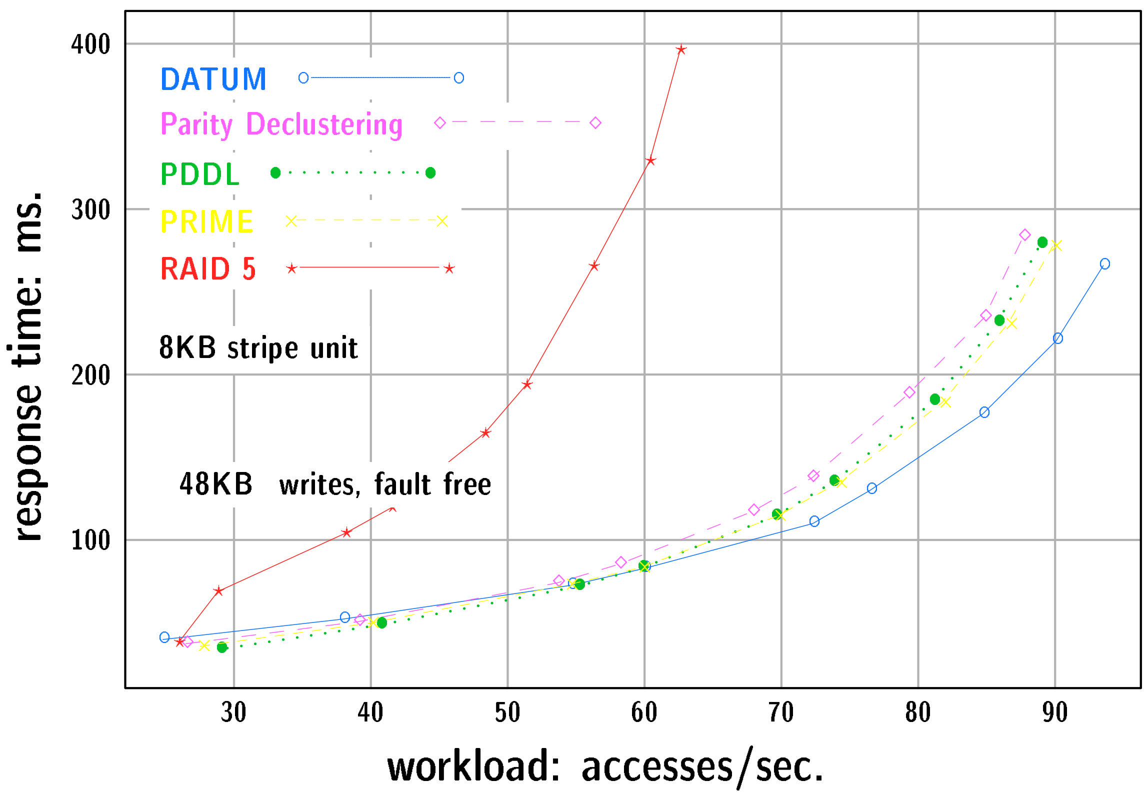
  0   1 5 8 12   2 3 10 11   4 6 7 9

- **Data Layout Goals**

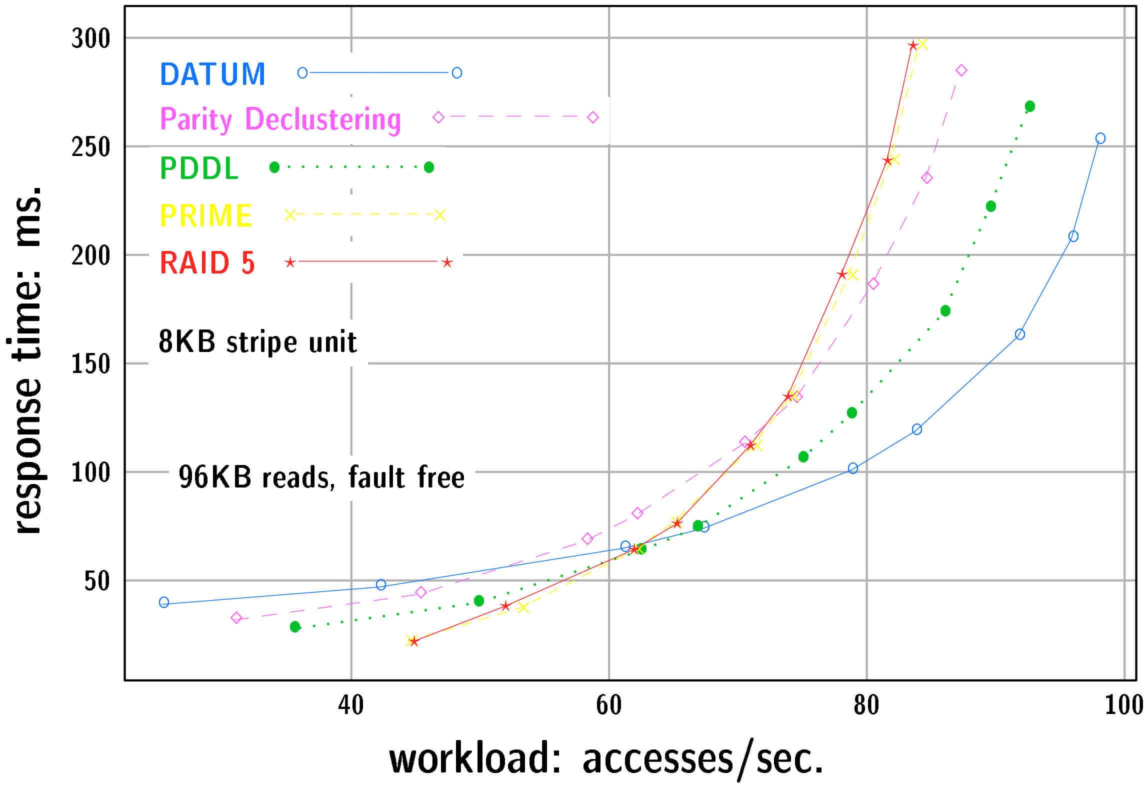
  all satisfied except parallel reads.

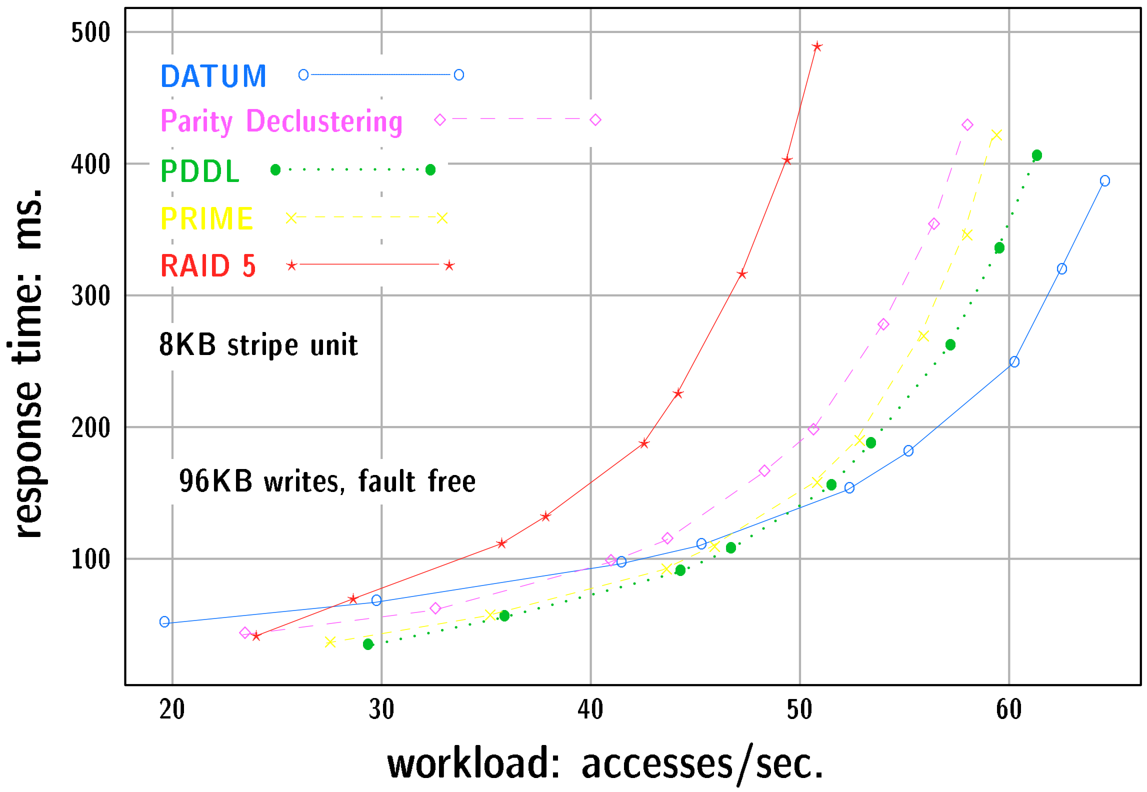
Permutation Development (PDDL) IV



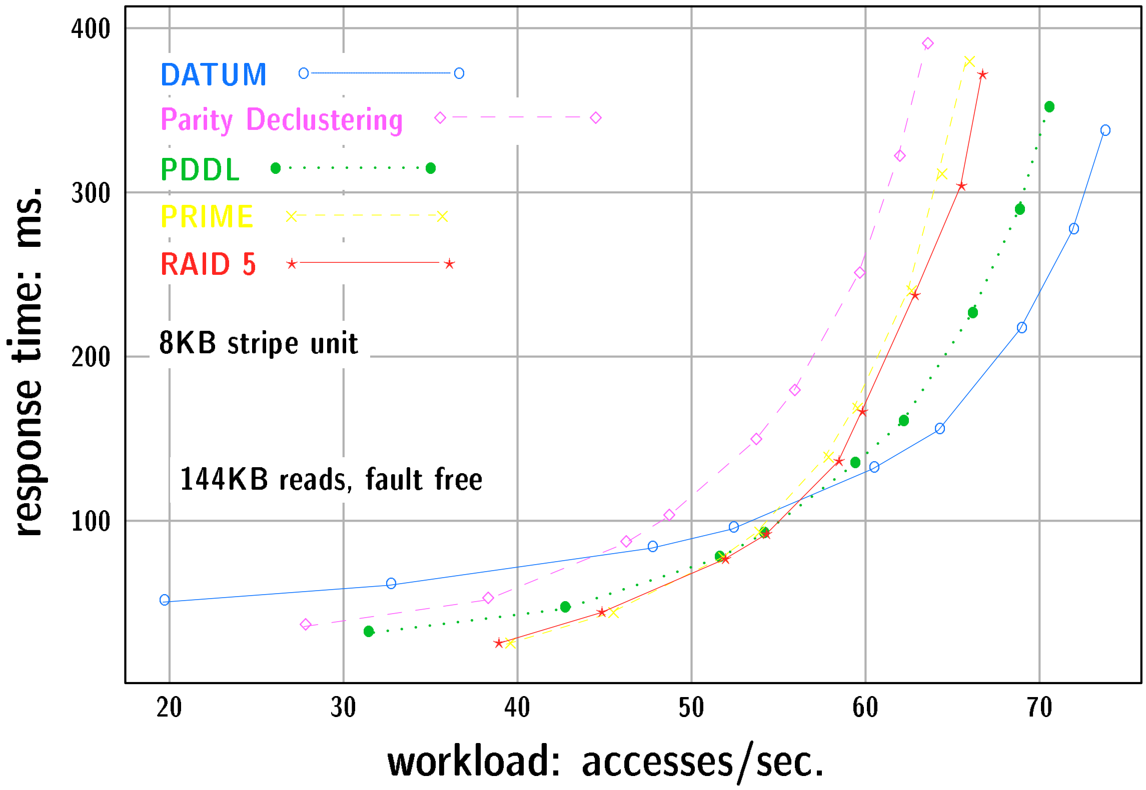


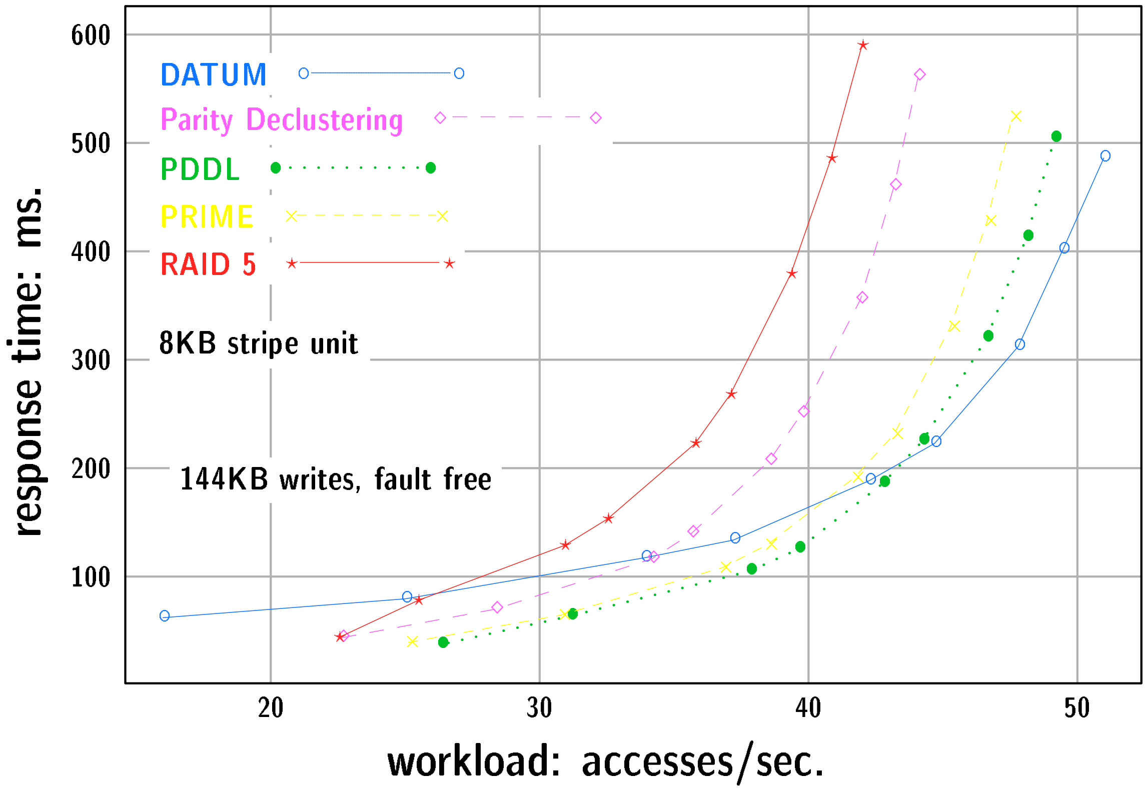
Permutation Development (PDDL) V



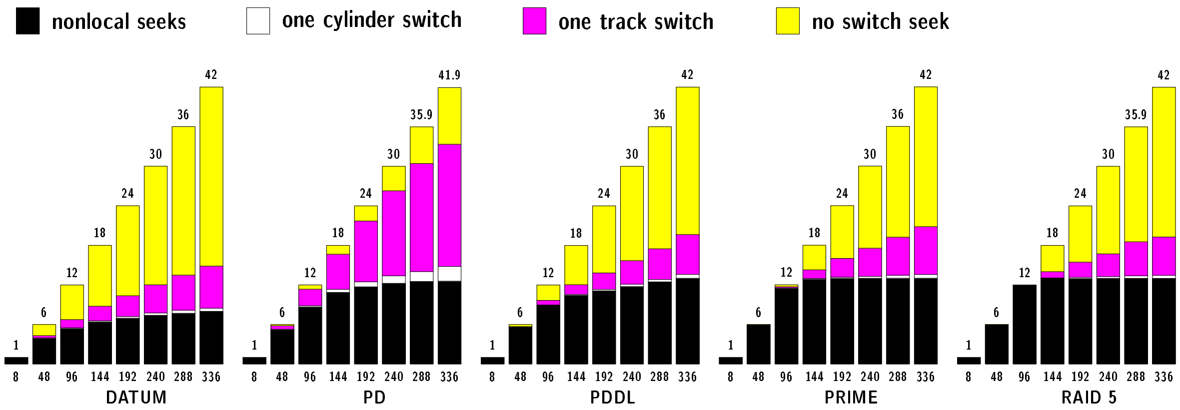


Permutation Development (PDDL) VI

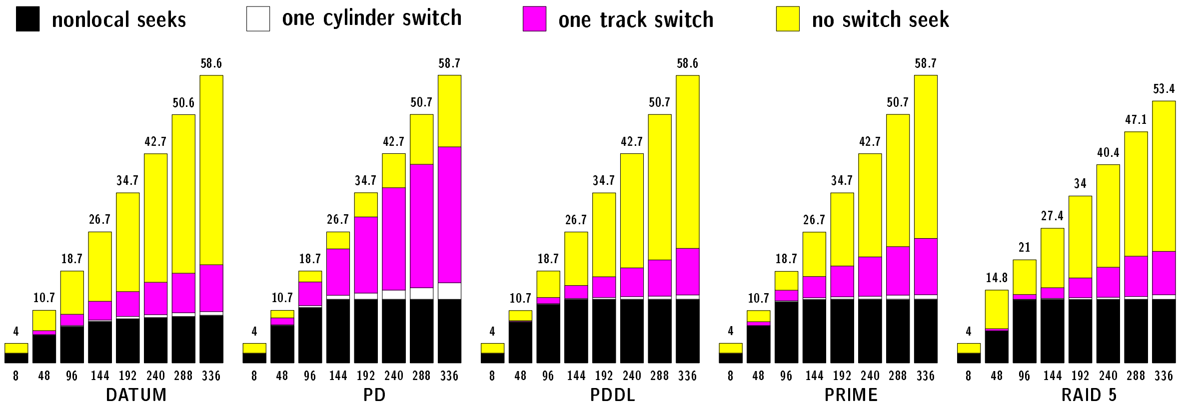




Permutation Development (PDDL) VII



Reads fault free



Writes fault free

Segmented Information Dispersal I

- ACM MM'96

  very large accesses $\approx$ 300KB

disk 0	disk 1	disk 2	disk 3	disk 4
A.0	B.0	C.0	D.0	E.0
A.1	B.1	C.1	D.1	E.1
pA	pB	pC	pD	pE

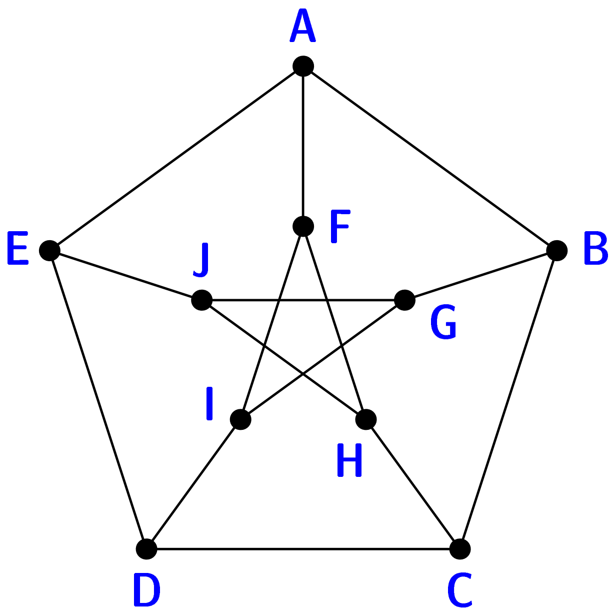
$$pA = B.0 \oplus E.1$$
$$pB = C.0 \oplus A.1$$
$$pC = D.0 \oplus B.1$$
$$pD = E.0 \oplus C.1$$
$$pE = A.0 \oplus D.1$$

Disk 2 fails

$$C.0 = pB \oplus A.1$$
$$C.1 = pD \oplus E.0$$

- Performance Consequences

Segmented Information Dispersal II



$$pA = B.0 \oplus F.0 \oplus E.0 \qquad pF = A.2 \oplus H.1 \oplus I.1$$
$$pB = A.0 \oplus G.0 \oplus C.0 \qquad pG = B.2 \oplus I.2 \oplus J.1$$
$$pC = B.1 \oplus H.0 \oplus D.0 \qquad pH = C.2 \oplus J.2 \oplus F.1$$
$$pD = C.1 \oplus I.0 \oplus E.1 \qquad pI = D.2 \oplus G.1 \oplus F.2$$
$$pE = D.1 \oplus J.0 \oplus A.1 \qquad pJ = E.2 \oplus G.2 \oplus H.2$$

- Disk F fails:

$$F.0 = pA \oplus B.0 \oplus E.0$$
$$F.1 = pH \oplus C.2 \oplus J.2$$
$$F.2 = pI \oplus D.2 \oplus G.1$$

Segmented Information Dispersal III

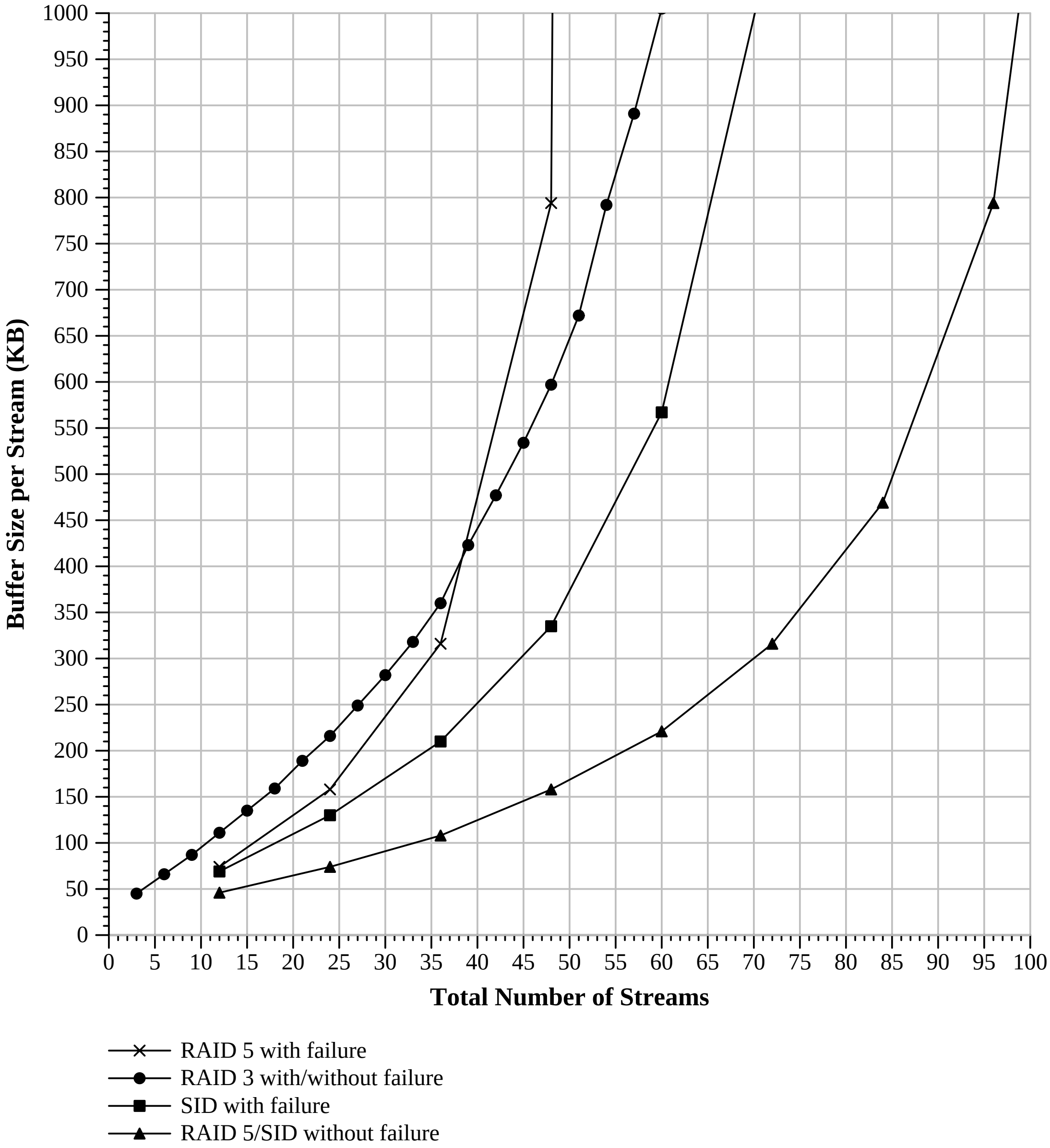
A separated difference set, SDS(n,q), is a set $C = \{ c_0, c_1, \ldots, c_{q-1} \}$ of positive integers less than n satisfying the following condition: let D be the set of differences between all ordered pairs of elements of C,

$$D = \{ (c_i - c_j) \bmod n \mid 0 \leq i,j < q \ \&\ i \neq j \}$$

then $C \cup D$ has cardinality q^2.

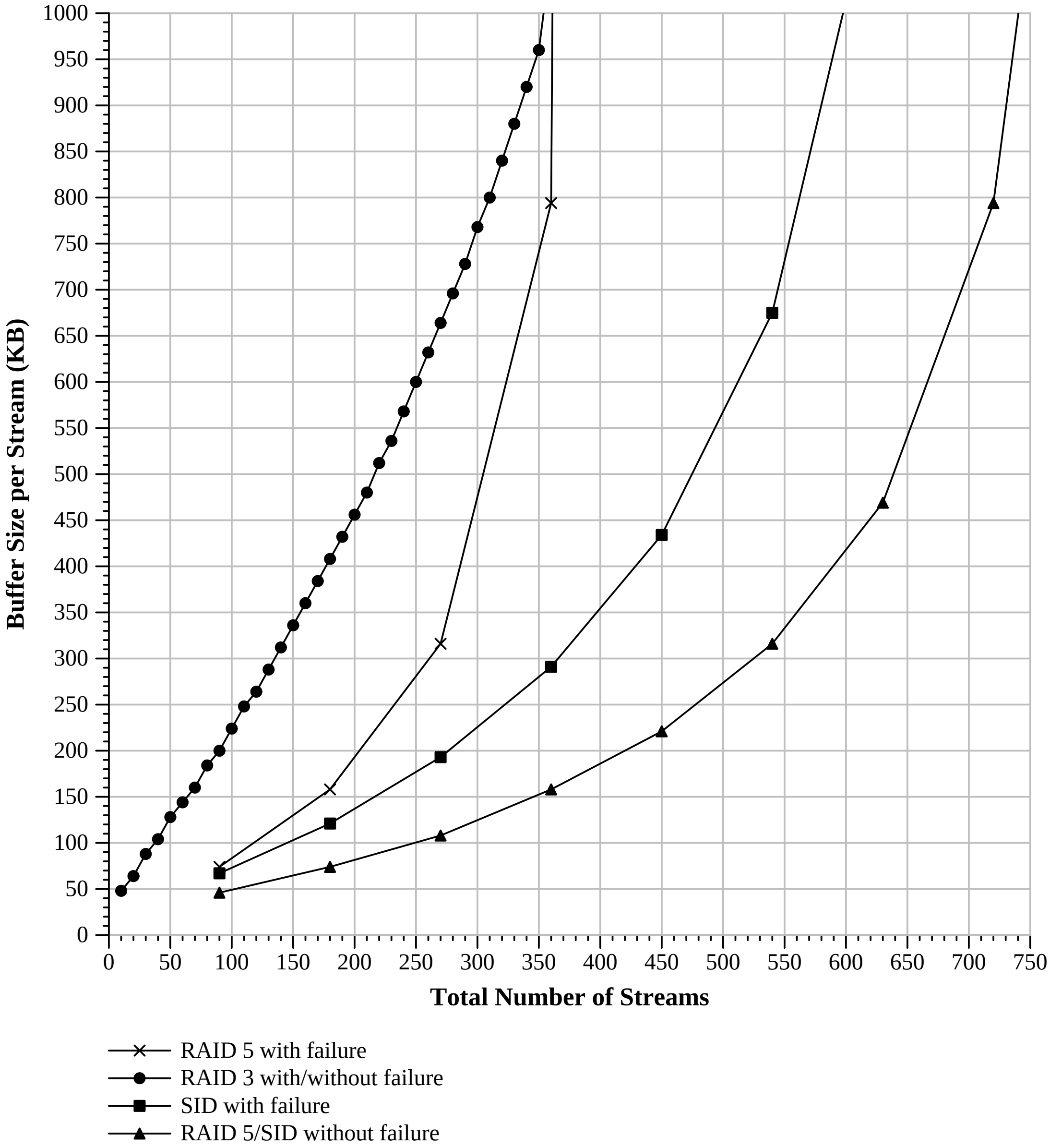
- SDS(5,2): $C = \{1, 4\}$ $D = \{2, 3\}$
- SDS(n,3): $C = \{1, 4, n\text{-}1\}$
  $D = \{2, 3, 5, n\text{-}5, n\text{-}3, n\text{-}2\}$
  n must be at least 11.

Segmented Information Dispersal IV



SID vs RAID Level 3 & 5 $n = 12 \quad q = 3$

Segmented Information Dispersal V



SID vs RAID Level 3 & 5 $n = 90 \quad q = 8$