Improving RAID-Based Storage Systems with Flash Memory

Rosie Wacha
Scott Brandt (UCSC Advisor)
John Bent, Gary Grider, James Nunez (LANL)

First Annual ISSDM/SRL Research Symposium
October 20-21, 2009

Data Centers

- Lots of data
- Must be able to keep up with read and write requests
- Power is a huge concern
- Data must be stored reliably, which impacts performance

Alternatives to Hard Drives

- Storage class memory
  - Solid state, persistent storage
  - E.g. flash, magnetic memory, phase change memory
- Flash is an attractive alternative to magnetic disks
  - 5-10x lower power
  - 2x throughput
  - 10x faster random access to data
  - 3x-10x more expensive than disks

Flash SSDs Replacing Disks

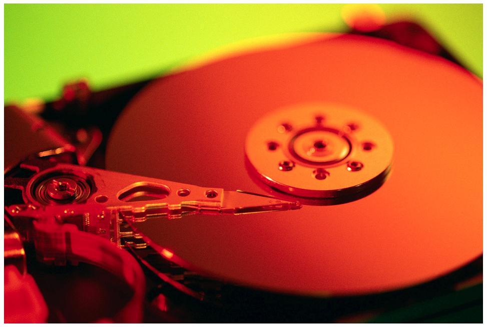
- Laptops
- Sensor networks
- Virtual memory
- Satellites
- No clear solution in data centers (EuroSys ‘09)
  - Not cost-effective to replace
  - Caching tier only cost-effective for 10% of workloads

Phase Change Memory in Storage Systems

- RAID 6 + PCM (MASCOTS ‘09)
  - Storing parities on PCM >=doubles reliability
  - No performance benefit assumed
- BPFS – byte-addressable persistent FS (SOSP ‘09)
  - Replace all disks with PCM
  - Use atomicity guarantees to do fewer writes
  - Faster than NTFS
- Neither look at power or cost of PCM

Our Solution: Replace Some Disks with Flash

- Flash Solid State Drives (SSDs) are available
  - Future work generalizing to other technologies
- RAID 4 + SSD = RAID 4S
  - Reduces load on remaining drives by up to 50%



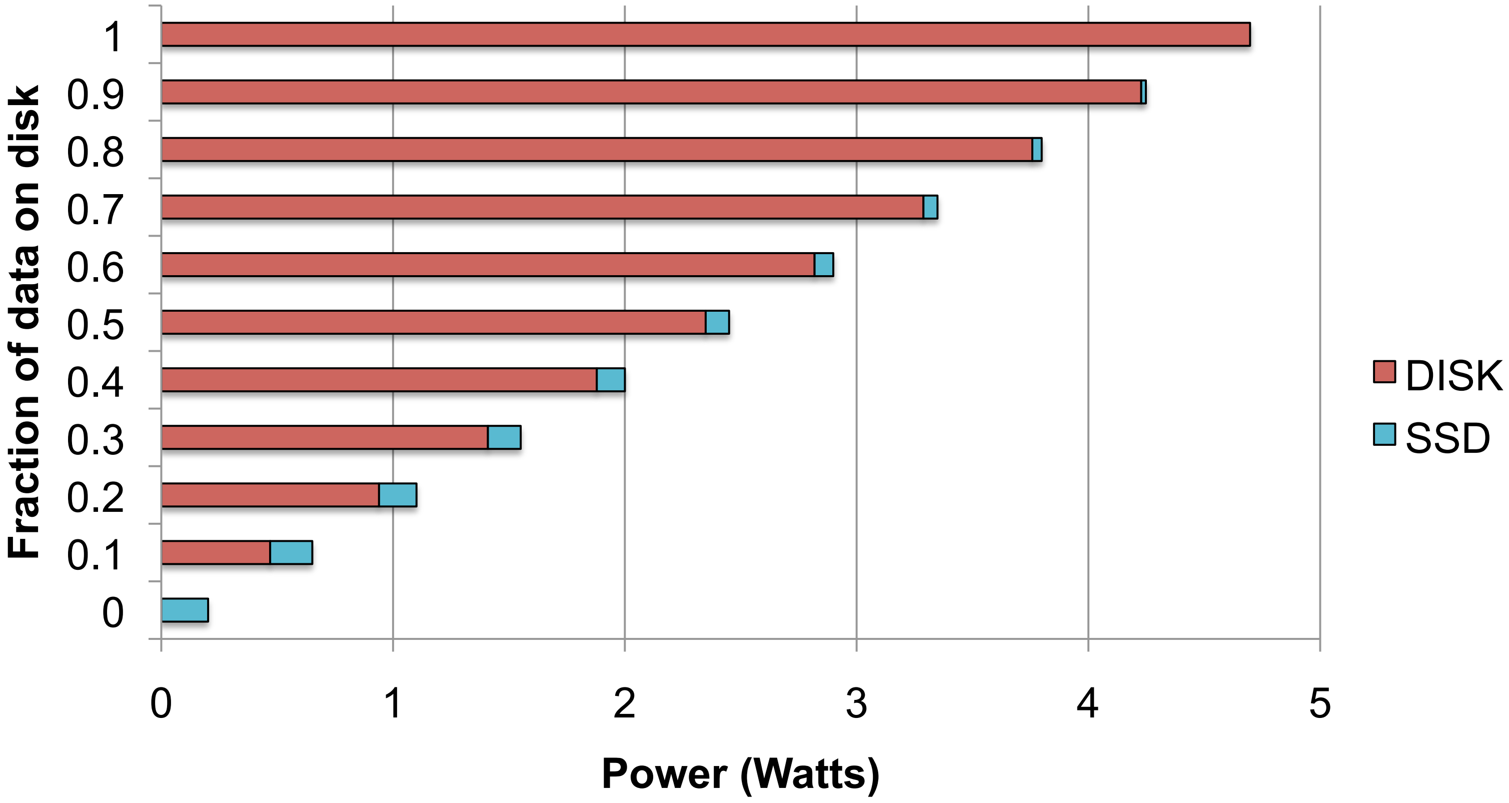
Solid State Drives (SSDs)

	Samsung Flash SSD PB22-J (MLC)	**WD VelociRaptor 10,000 RPM**
Cost	$799.31	$229.99
Capacity	256 GB	300 GB
$/GB	$3.12	$0.77
Read / Write Throughput	220 / 200 MB/s	120 / 120 MB/s
Latency	0.1 ms	3 ms
Power	≤ 1.5 W	≤ 6 W

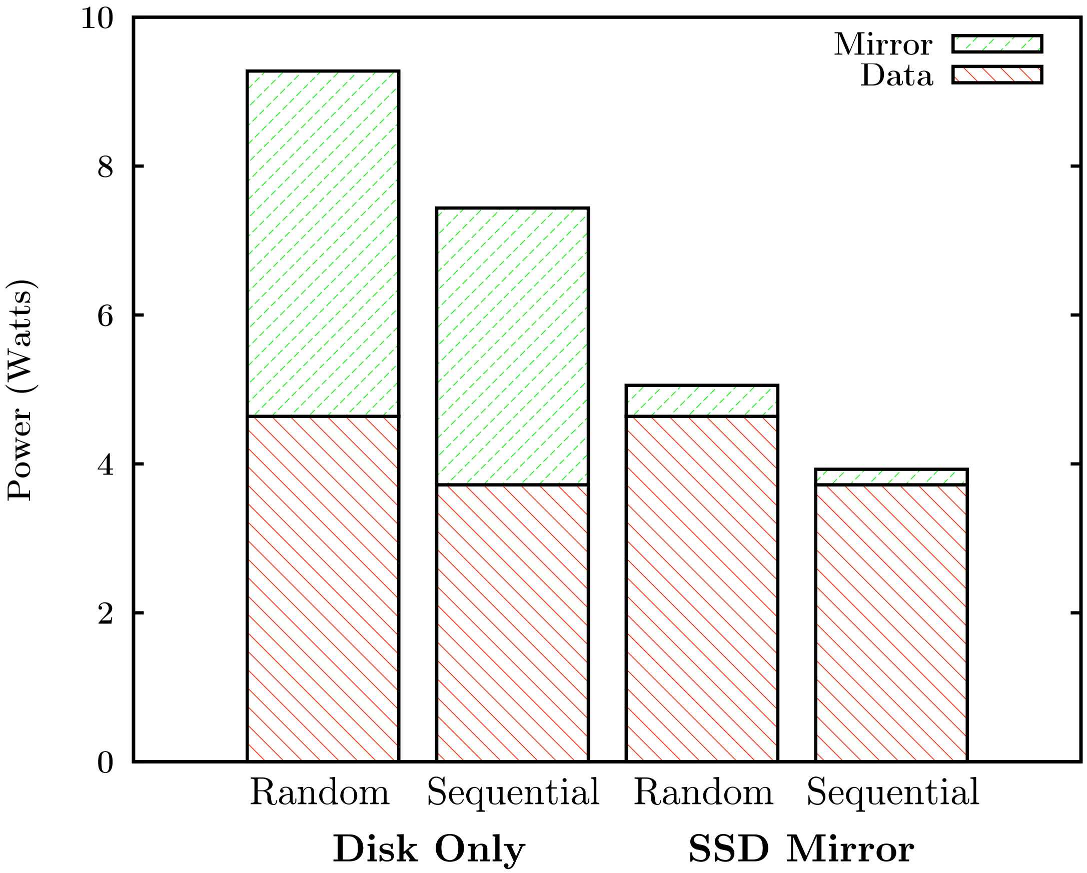
Power Simulation

- 1.5 MB/s synthetic random read workload
  - 64 KB request size
- Calculate transfer, seek, and idle times
  - Disk
  - SSD
- Vary amount of disk vs. SSD performing the workload
  - Calculate power based on workload

Disk vs. SSD Power Consumption



RAID 1 Power Reduction



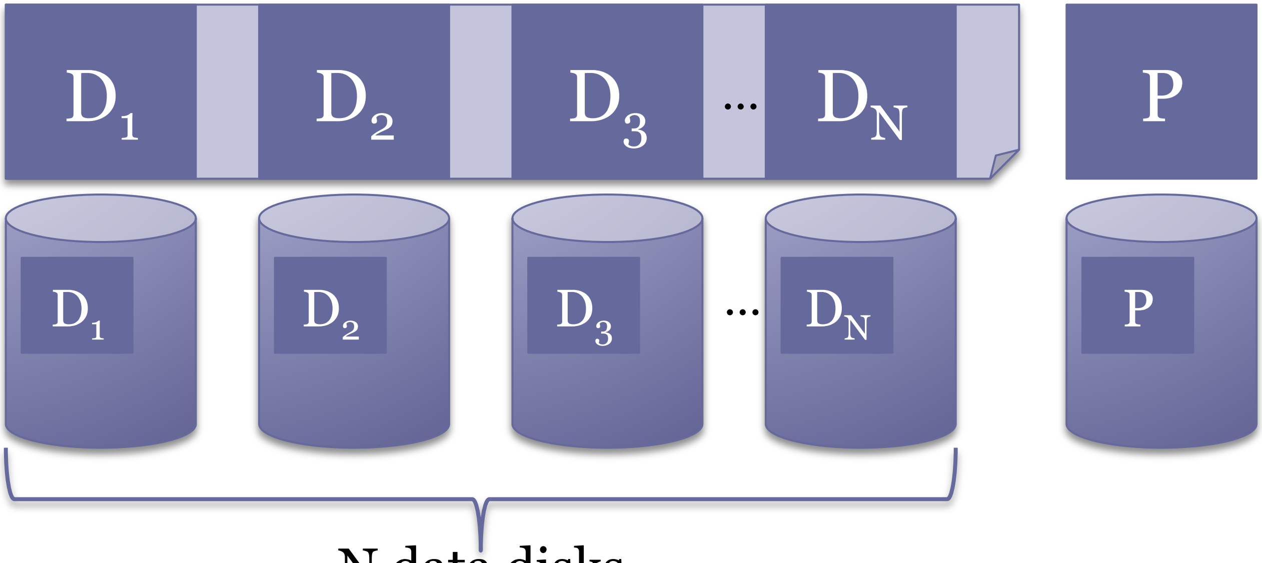
- Simulated workload
  - 1.5 MB/s read
  - 64 KB requests
- Samsung SLC SSD
- Western Digital WD20EADS

Large, Sequential Writes in RAID 4

- N write requests → N+1 writes to disk
  - N data writes and 1 parity write
- average per write request → 1+1/N writes to disk
- RAID 5 performance is same for this workload

Large, Sequential Writes in RAID 4

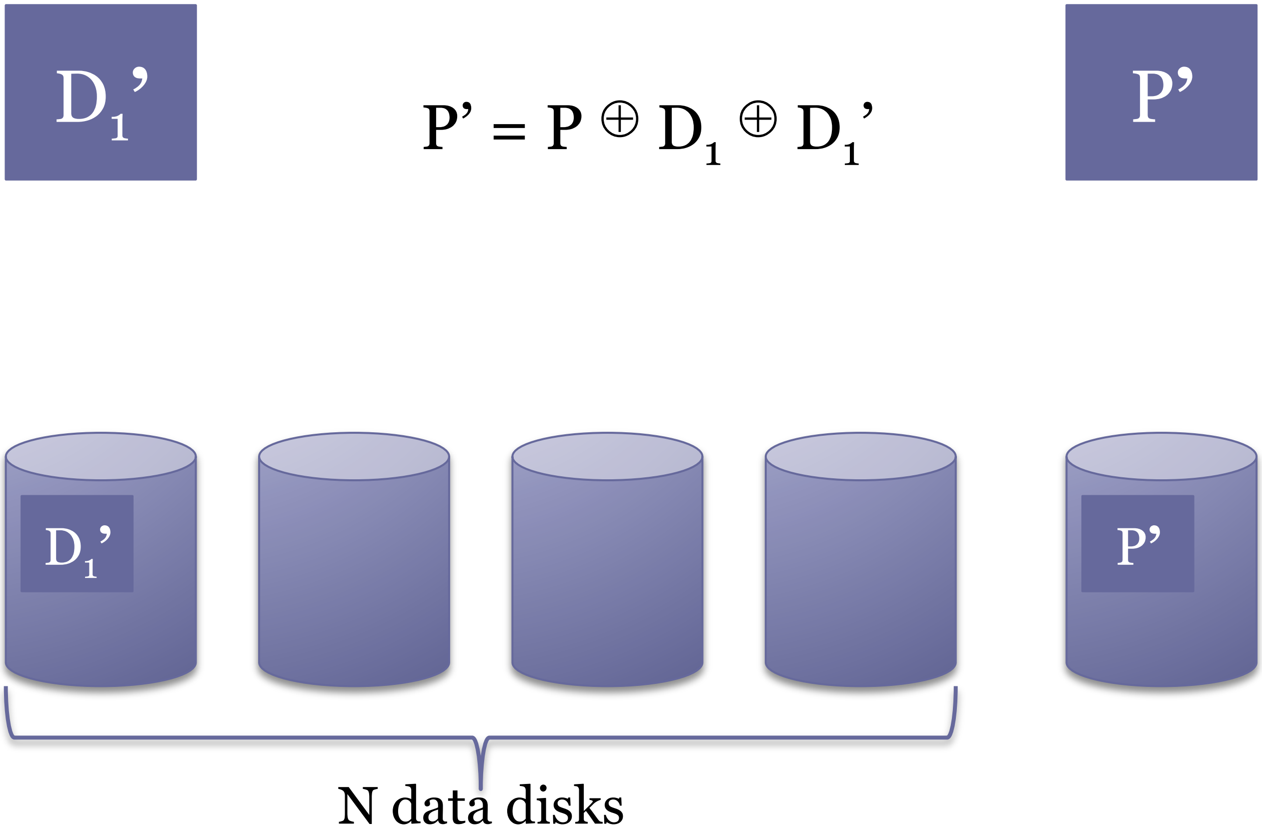
- N write requests → N+1 writes to disk
  - N data writes and 1 parity write
- average per write request → 1+1/N writes to disk
- RAID 5 performance is same for this workload



N data disks

Small, Random Writes in RAID 4

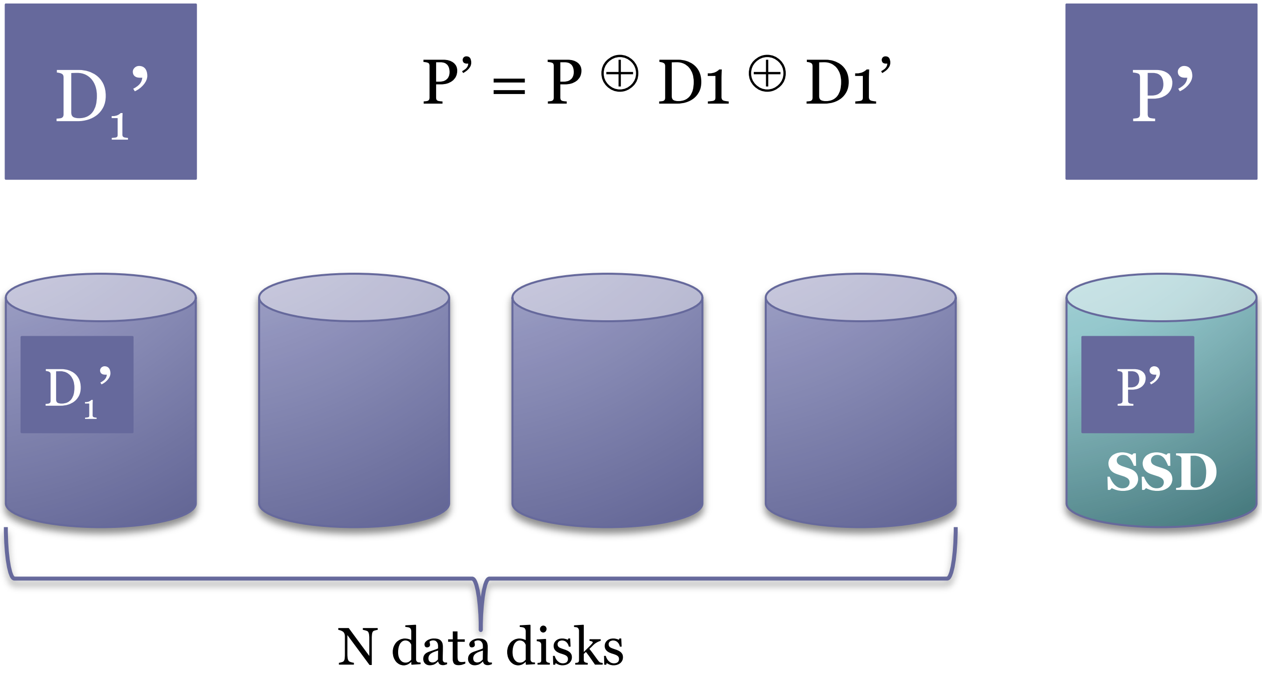
- 1 write request → 2 disk reads + 2 disk writes



$$P' = P \oplus D_1 \oplus D_1'$$

RAID 4S – SSD Parity

- 1 write request → 1 disk read + 1 disk write +
  1 SSD read + 1 SSD write
  → 1 disk read + 1 disk write

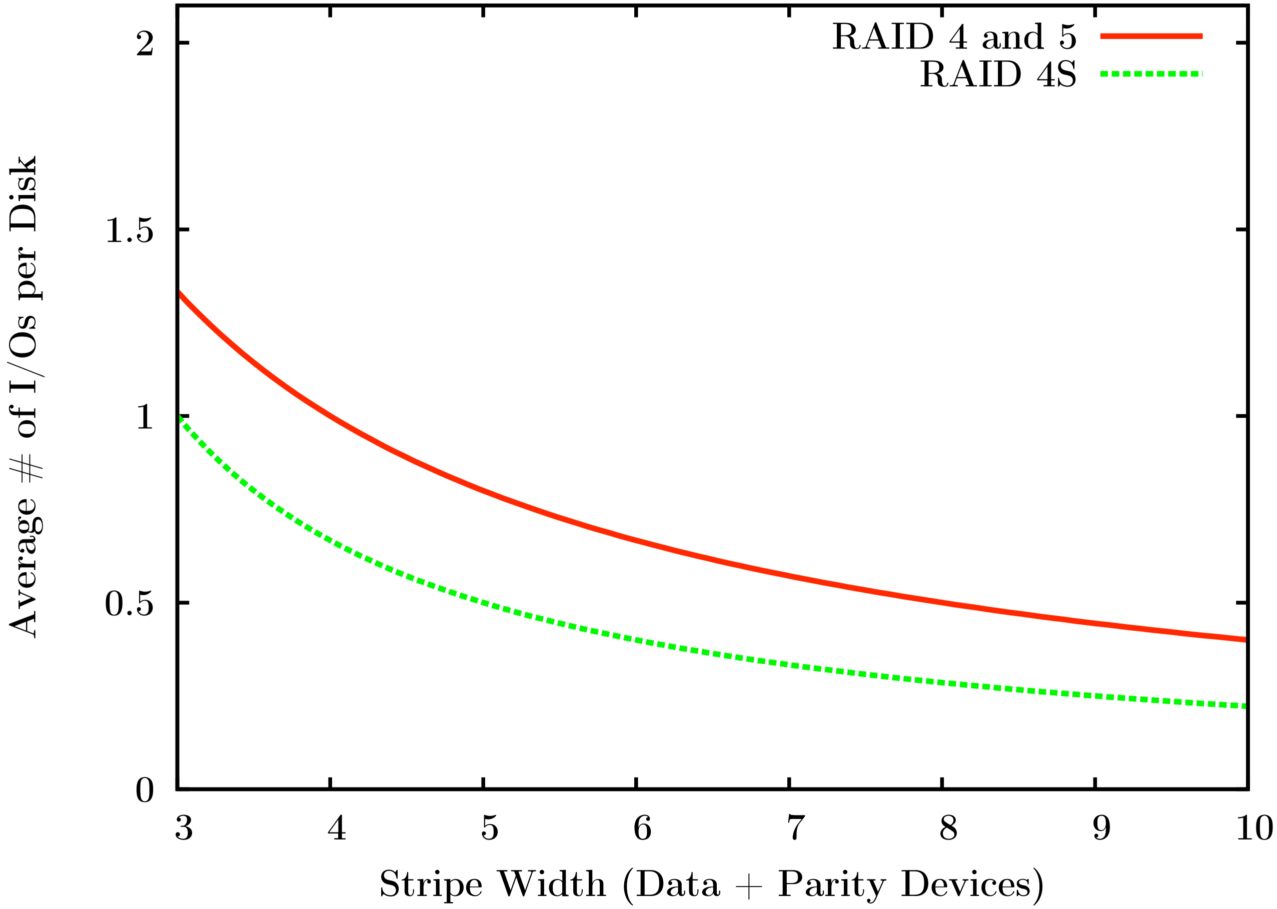


$P' = P \oplus D1 \oplus D1'$

RAID 4S: Use SSD for RAID 4 Parity

- RAID 5 small writes
  - 1 write → 2 disk reads + 2 disk writes
  - k writes → 2k reads + 2k writes
  - Avg. # I/Os per disk in stripe size N → **4/(N+1)**
- RAID 4S small writes
  - 1 write → 1 disk read + 1 disk write
  - k writes → k disk reads + k disk writes
  - Avg. # I/Os per disk in stripe size N → **2/N**

Small Write Performance of RAID 4S



- Theoretical experiment
- RAID 5:

  4/(N+1)
- RAID 4S:

  2/N

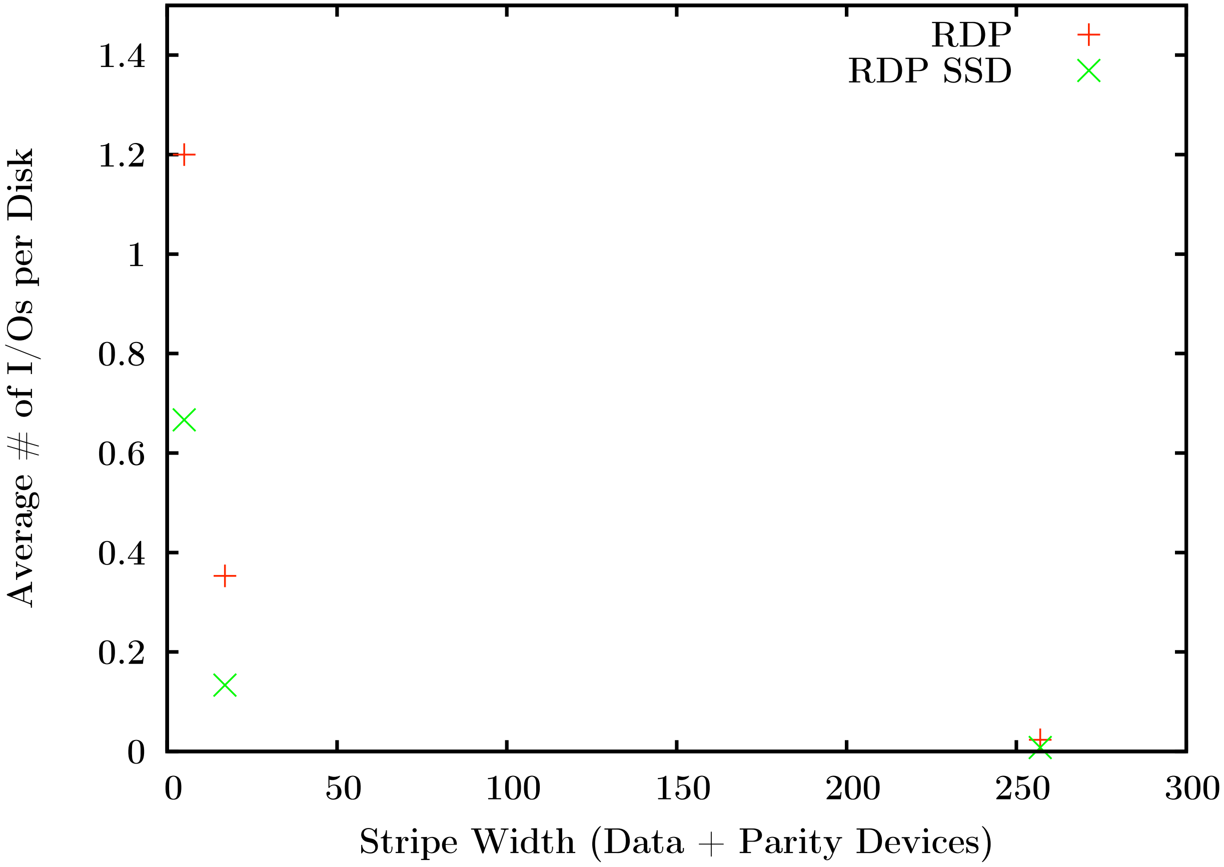
Reduction in Average # of I/Os per Disk

Stripe width	RAID 5	RAID 4S	% reduction
3	1.33	1	25
4	1	0.67	33
5	0.8	0.5	38
10	0.4	0.22	44
50	0.08	0.041	49
100	0.04	0.020	49.5

Row-Diagonal Parity with SSD

- Replace both parity drives with SSDs
- RDP small writes
  - 1 write → 3 disk reads + 3 disk writes
  - Avg. # I/Os per disk in stripe size N → **6/(N+2)**
- RDP SSD small writes
  - Offload both parities to SSDs
  - 1 write → 1 disk read + 1 disk write
  - Avg. # I/Os per disk in stripe size N → **2/N**

Row-Diagonal Parity with SSD



- Theoretical experiment
- RDP:

  6/(N+2)
- RDP SSD:

  2/N

Degraded Mode and Reconstruction

- Degraded mode
  - Reads and writes access all disks in stripe
  - Disks are more fully utilized
  - Parity SSD is more idle
    - Lower small write overhead than all-disk array
- Rebuild onto spare SSD
  - Read all data
  - Compute lost data
  - Faster than spare disk
  - Small writes don’t overwhelm the parity SSD

Conclusions and Future Work

- Incorporating a small number of SSDs improves RAID
  - Lower power
  - Better performance
  - Feasible higher reliability
- Performance analysis with real workloads
- Cost / benefit analysis of adding flash
- Implement RAID 4S prototype
- Degraded mode / reconstruction