High Performance Computing

13M37098 Yuki Takasaki

Review Paper

“DASH: a Recipe for a Flash-based Data Intensive Supercomputr”

[SC’12 Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis]

Jiahua He, Arun Jagatheesan, Sandeep Gupta, Jeffrey Bennett, and Allan Snavely

San Diego Supercomputer Center (SDSC)

University of California, San Diego

Outline

1. Introduction
2. System Overview
3. I/O System Design and Tuning
4. Performance of Real-World Data-Intensive Applications
5. Related Work
6. Conclusions

- Comment

1.Introduction

- Data intensive computing can be defined as computation involving large dataset and complicated I/O patterns.
  - Data mining application : a large amount of raw data on disk and complex data that make parallelization difficult
  - Predictive science application : a large amount of generated intermediate data
- Data intensive computing is challenging
  - There is a five-order-of-magnitude latency gap between main memory and spinning hard disk.

1.Introduction

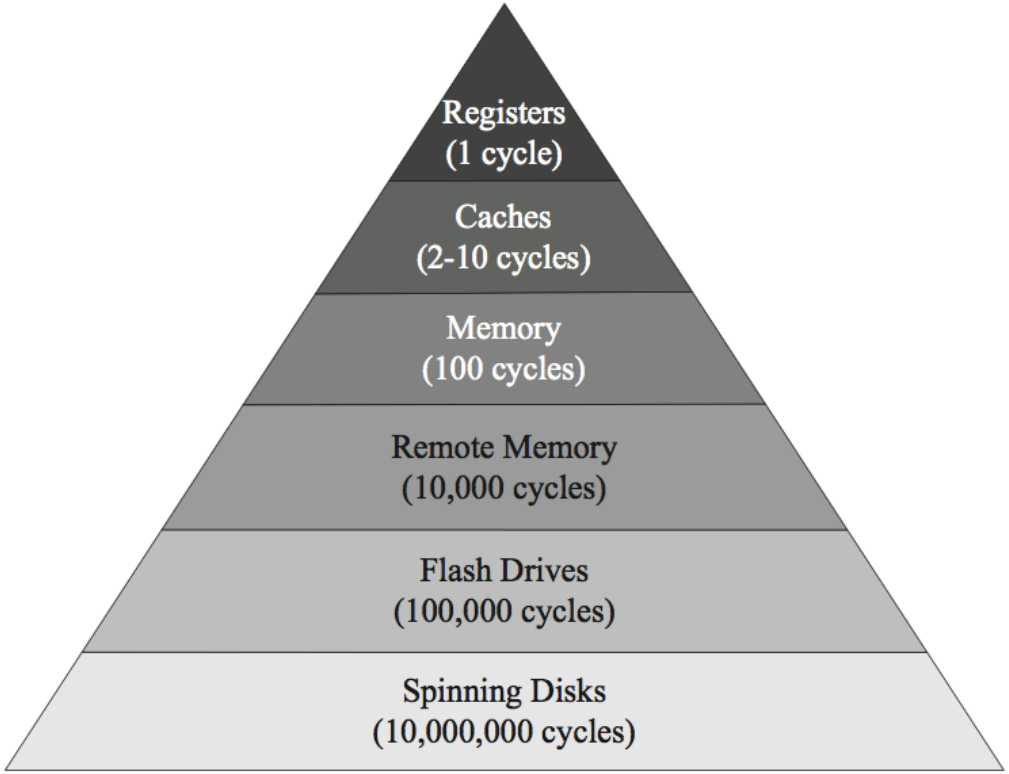

Figure 1. The memory hierarchy. *Each level shows the typical access latency in processor cycles. Note the five-orders-of-magnitude gap between main memory and spinning disks.*

1.Introduction

- They designed and built a prototype data intensive supercomputer named DASH
  - flash-based Solid State Drive (SSD) technology
  - virtually aggregated DRAM to fill the latency gap
  - Use commodity parts including Intel X25-E flash drives and distributed shared memory(DSM) software called vSMP from ScaleMP

2.System Overview

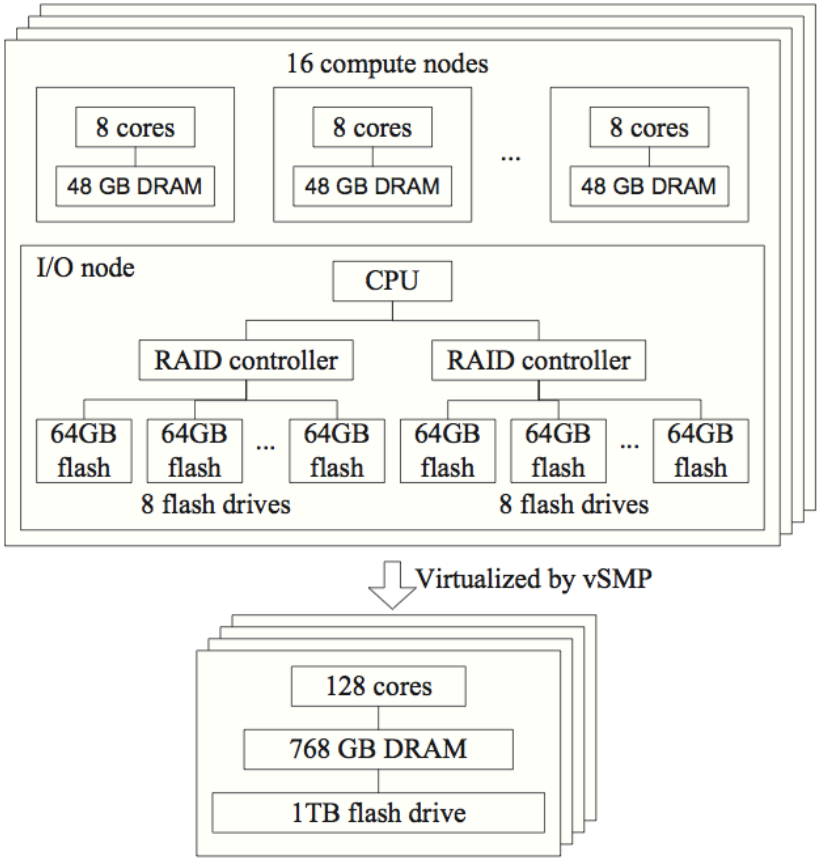

Figure 2. Physical and virtual structure of DASH supernodes. *DASH has in total 4 supernodes IB interconnected of the type shown in the figure.*

2.System Overview

- A. Storage hierarchy
  - SLC(Single-Level Cell) drive of 1 TB/supernode
    - Longer lifetime, lower bit error rate, and lower latency than MLC(Multi-Level Cell) drive.
  - local DDR3 DRAM memory of 48GB/computenode
  - Distributed shared memory of 768GB
    - Use vSMP software to aggregate distributed into a single address space.
    - Can use all that memory as a RAM disk for fast I/O

2.System Overview

- B. Cost efficiency

TABLE 1. COST EFFICIENCY COMPARISON BETWEEN DASH AND COMMERCIAL PRODUCTS.

	Generic HDD (SATA)	DASH-I/O node	DASH Super node	Fusion –IO	Sun – F5100
GB	2048	1024	768	160	480
MB/s/$	~0.4	0.16	0.49	0.12	0.07
$/GB	~0.15	19.43	112.63	41.06	90.62
IOPS/$	0.4-1.0	28	52	18	9
IOPS/GB	0.05-0.1	549	5853	725	828

2.System Overview

- C. Power efficiency

TABLE 2. COMPARISON OF POWER METRICS BETWEEN SSD AND HDD.

	DRAM 7x2 GB Dimms (14 GB)	Flash SSD 64GB	HDD 2TB
Active Power	70 W	2.4 W	11 W
Idle Power	35 W	0.1 W	7 W
IOPS per Watt	307	712	35

3.I/O System Design and tuning

- To evaluate the performance of storage systems, bandwidth and IOPS are both important metrics.
  - Bandwidth measures sequential performance.
  - IOPS shows the throughput of random accesses.
- They biased towards achieving high IOPS
  - Their target applications are characterized as intensive random accesses
- To pursue and measure the peak I/O performance of the system, they adopted RAID 0.
- They used IOR and XDD
  - The most accurate, reliable, and well-known I/O benchmarks in their experience.
  - They verify each other and their result were always similar in our tests.

3.I/O System Design and tuning

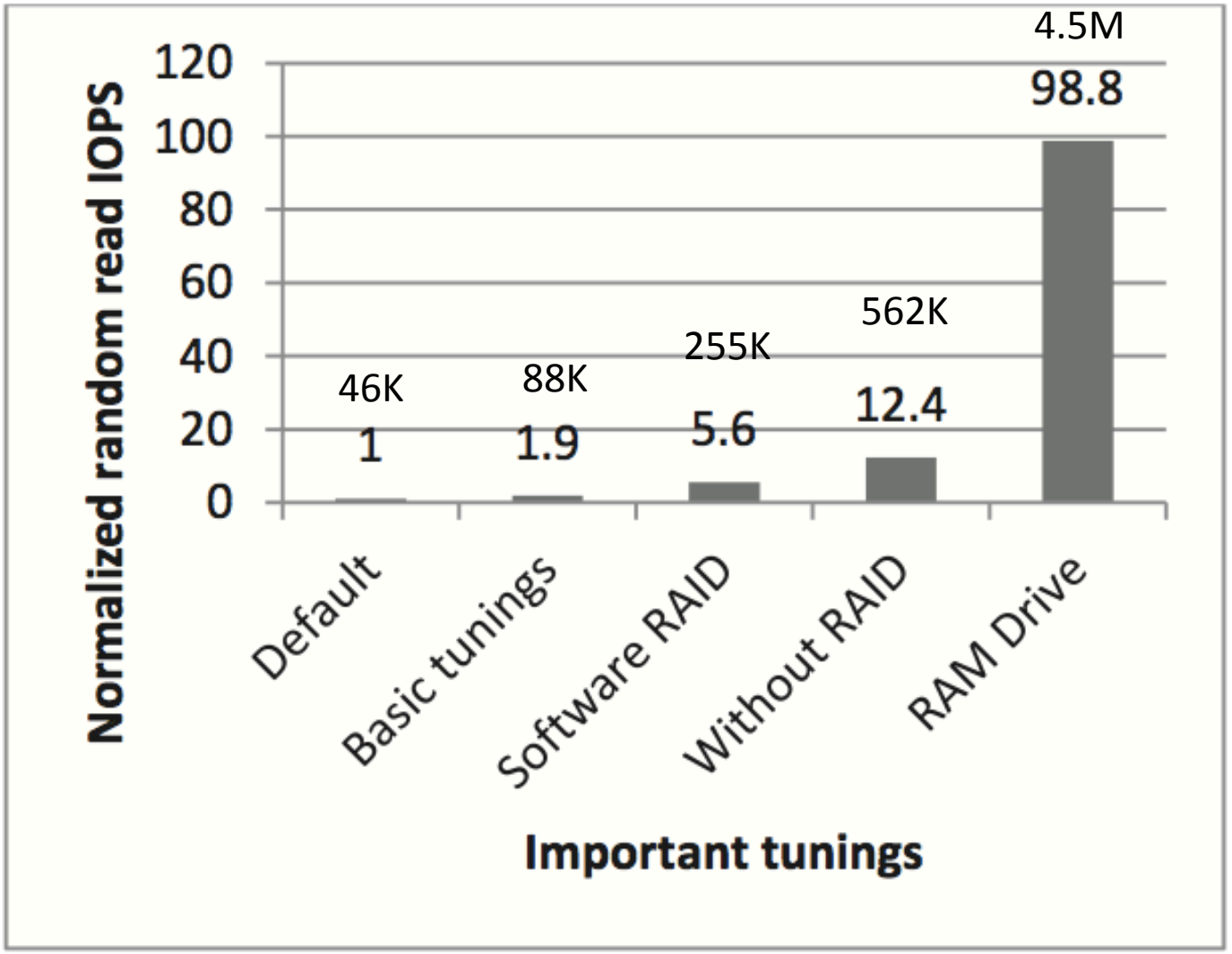

Figure 3. Random read performance improvements with important tunings.

3.I/O System Design and tuning

- A. Single drive tuning

TABLE 3. IMPORTANT TUNING PARAMETERS FOR FLASH DRIVES.

Parameters	Descriptions	DASH setting
Write Caching	Write through or write back in the drive ram-cache	Write back
Read Ahead	Read the data into the drive ram-cache before they are requested according to the access pattern.	On
AHCI	Advanced Host Controller Interface, API for SATA host bus adapters.	On

3.I/O System Design and tuning

- A. Single drive tuning

TABLE 4. I/O TEST RESULTS OF A SINGLE FLASH DRIVE.

	Sequential Write (MB/s)	Sequential Read (MB/s)	Random Write (4KB IOPS)	Random Read (4KB IOPS)
Measured	203	261	10724	39756
Spec	170	250	3300	35000

3.I/O System Design and tuning

- B. Basic RAID tuning

TABLE 5. IMPORTANT TUNING PARAMETERS FOR THE DASH I/O SYSTEM.

Components	Parameters	Descriptions	Final DASH setting
I/O Benchmarks	Cache Policy	Cached or direct I/O, use the OS buffer cache or not.	Direct I/O
	API	I/O APIs to access drives such as POSIX, MPIIO, HDF5 and netCDF.	POSIX
	Chunk Size	The data size of each request. I/O benchmarks usually generate fixed-sized requests.	4MB for sequential tests, 4KB for random tests
	Queue Depth	The number of outstanding I/O requests.	1 for sequential tests and 128 for random tests
Operating System	I/O Scheduler	Schedule and optimize I/O accesses. There are 4 algorithms in the 2.6 Linux kernel: CFQ (default), Deadline, Anticipatory, and No-op.	No-op
	Read Ahead	Read the data into cache before they are requested according to the previous access pattern.	Off
Hardware RAID	Cache Policy	Cached or direct I/O, use the RAID controller cache or not.	Direct I/O
	Write Policy	Write through or write back.	Write through
	Read Ahead	RAID-level read ahead.	Off
	Stripe Size	The block size in which RAID spread data out to drives.	64KB

3.I/O System Design and tuning

- B. Basic RAID tuning

TABLE 6. I/O TEST RESULTS WITH 2 DIFFERENT STRIPE SIZES.

Stripe Size (KB)	Sequential Write (MB/s)	Sequential Read (MB/s)	Random Write (4KB IOPS)	Random Read (4KB IOPS)
64	1179	2199	3749	87563
128	1275	2056	3121	79639

3.I/O System Design and tuning

- C. Advanced tuning
  - They suspected that the bottleneck might be the RAID controller.
    - Replace the RAID controller with the state-of-the-art RAID controller(Intel RS2BL080)
    - Use simple Host Bus Adapter(HBA) without embedded processors
      - Share the power from the host CPU
    - They connected only 6 flash drives to compose a software RAID and achieved 153,578 4KB IOPS
      - Almost 2x of the hardware RAID performance
  - Their motherboard has on-board HBA, which is rated higher than 150K 4KB IOPS by the vendor.
    - Each HBA can connect 4 flash drives and their motherboard can hold 4 HBAs.

3.I/O System Design and tuning

- C. Advanced tuning

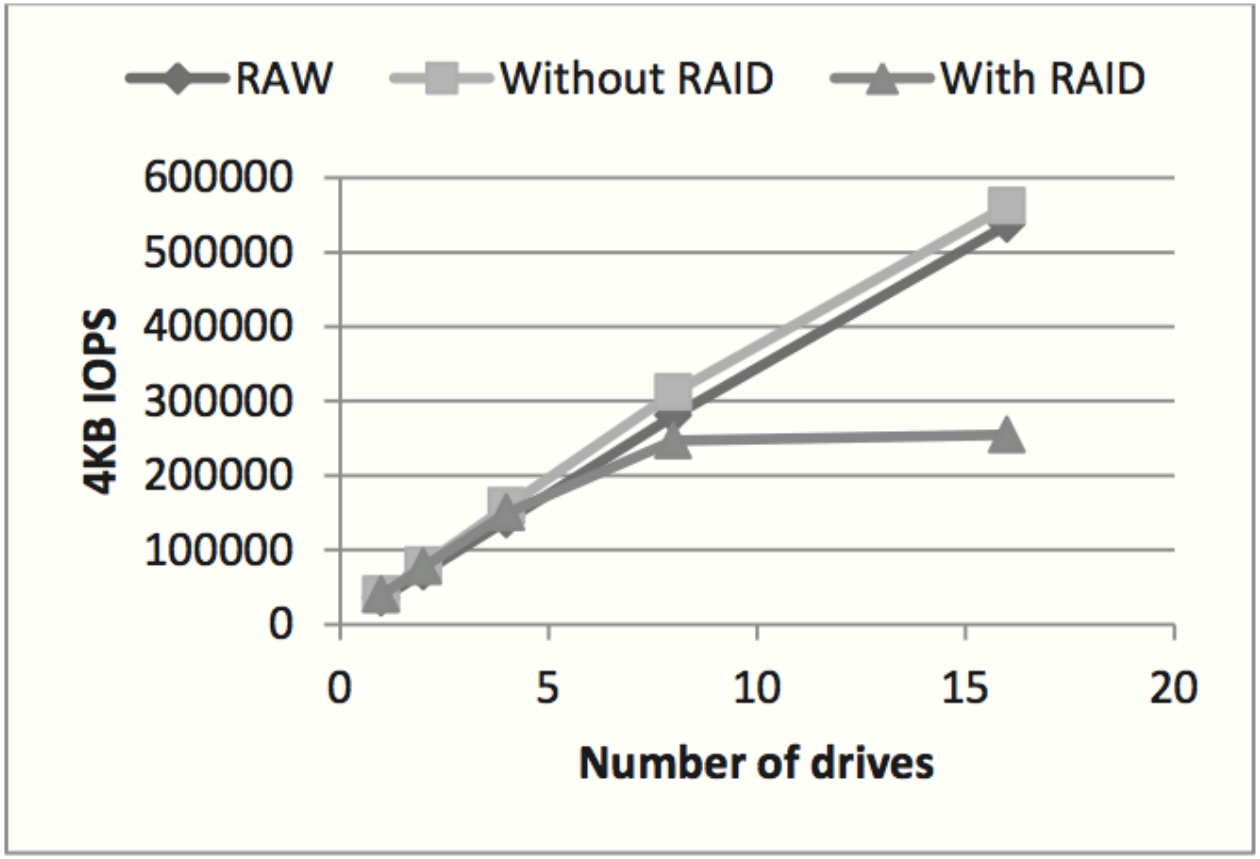

Figure 4. Random read performance with and without RAID. *The configuration with RAID only scales up to 8 drives while the one without RAID can scale linearly up to 16 drives. We also ran tests with raw block devices.*

3.I/O System Design and tuning

- C. Advanced tuning

TABLE 7. I/O TEST RESULTS WITH AND WITHOUT RAID.

	Sequential Write (MB/s)	Sequential Read (MB/s)	Random Write (4KB IOPS)	Random Read (4KB IOPS)
With RAID	1395	2119	19784	254808
Without RAID	2958	3225	143649	562365

3.I/O System Design and tuning

- D. RAM drive

TABLE 8. I/O TEST RESULTS OF THE RAM DRIVE.

Sequential Write (MB/s)	Sequential Read (MB/s)	Random Write (4KB IOPS)	Random Read (4KB IOPS)
11,264	42,139	2,719,635	4,495,592

4.Performance of Real-World Data-Intensive Applications

- A. External memory BFS
  - External memory BFS is a common component in several predictive science graph-based applications.
  - They used the external memory package 0.39 implemented by Deepak Ajwani et al. in their experiments.
    - They use one of the algorithms, MR-BFS.
    - They ran a range of tests on a dataset size of 200GB and compared the performance of three different storage media with similar and comparable configurations.
      - RAM drive, flash drives, and spinning disks

4.Performance of Real-World Data-Intensive Applications

- A. External memory BFS

TABLE 9. AVERAGE MR-BFS RESULTS ON THE DASH SUPERNODE FROM DIFFERENT STORAGE MEDIA

	RAM Drive	Flash Drives	Spinning Disks
Total I/O Time (sec)	854 (5.2x)	1862 (2.4x)	4444
Total Run Time (sec)	1917 (3.0x)	3130 (1.8x)	5752

4.Performance of Real-World Data-Intensive Applications

- B. Palomar Transient Factory
  - Palomar Transient Factory is a data base application used to discover time-variable phenomena in astronomy data.
  - The response times of the forward query and the backward query are crucial for PTF.
  - They measured these query response times on DASH

4.Performance of Real-World Data-Intensive Applications

- B. Palomar Transient Factory

TABLE 10. COMPARISON OF PTF QUERY RESPONSE TIMES ON DASH AND PTF PRODUCTION DATABASE WITH SPINNING DISKS.

Query type	Forward Query	Backward Query
DASH-IO (SDSC)	11ms (124x)	100s (78x)
Existing DB	1361ms	7785s

4.Performance of Real-World Data-Intensive Applications

- C. Biological pathways analysis
  - Biological pathway analysis are an integrated data-mining of heterogeneous biological data framework.
  - BiologicalNetworks is a Systems Biology software platform for analysis and visualization of biological pathways, gene regulation and protein interaction network.
  - They ran some popular queries of BiologicalNetworks on three different media on SDSC DASH including hard disks, SSDs and memory(using vSMP)

4.Performance of Real-World Data-Intensive Applications

- C. Biological pathways analysis

TABLE 11: QUERY RESPONSE TIMES OF POPULAR QUERIES IN BIOLOGICAL NETWORKS ON DIFFERENT STORAGE MEDIA (HARD DISK, SSD AND MEMORY) AND THEIR SPEED-UP IN COMPARISON TO HARD DISK.

Query	Q2C	Q3D	Q5F	Q6G	Q7H
RAMFS (vSMP)	11338ms (1.42x)	62850ms (3.60x)	3ms (186x)	17957ms (1.54x)	211ms (5.64x)
SSD	11120ms (1.45x)	176873ms (1.28x)	11ms (50.73x)	24879ms (1.11x)	495ms (2.41s)
HDD	16090ms	226023ms	558ms	27661ms	1191ms

5.Related Work

- A.ccNUMA machines
  - ccNUMA machines have single shared memory space by special hardware.
    - SGI Altix 4000 series, HP Superdome, and Bull NovaScale 5000 series
  - With these machines, people can program across all the nodes in shared-memory model.
  - However, these products usually adopt proprietary technology based on customized hardware, and need a long development period, which makes their ratios of performance to price pretty low.
  - vSMP is a software implementation of ccNUMA and is much more cost efficient.

5.Related Work

- B.Distributed Shared Memory (DSM)
  - People try to achieve ccNUMA's function with a software implementation called DSM, such as vSMP.
    - Data intensive applications are becoming dominant and the requirement for large shared memory is becoming stronger.
    - Most of the new system exploit the virtual machine technology and implement the DSM layer under the operating system and right above the hardware.
    - It provides a single system image to the operating system and eases the management burden.

6.Conclusion

- They designed and built a new prototype system called DASH, exploiting flash drives and remote memory to fill the gap.
- They tuned the system and achieved ~560K 4KB IOPS with 16 flash drives and ~4.5M 4KB IOPS with 650GB RAM drive.
- With 3 real applications from graph theory, biology, and astronomy, we attained up to two-order-of-magnitude speedup with RAM drives compared with traditional spinning disks.

Comment

- Strong point
  - Experiment environment is suitable.
- Weak point
  - I don’t know which benchmark do they use in section 3