Flash Memory and PRAM: Sleeping with the Enemy

- Accelerating In-Page Logging with PRAM -

Apr. 19, 2011

Sang-Won Lee

Motivation??

NVRAMOS 2010

Flash is Coming

- The age of flash-based DBMSs is coming
  - Oracle's TPC-C BM result @ 2010 using Exadata
    - ✓ Oracle + Sun Flash Storage
    - ✓ Total cost: 49M $
      - Storage: 23M $
        - Sun Flash Array: 22M $
        - 720 2TB 7.2K HDD: 0.7M$
  - IBM proposed SSD Buffer (VLDB 10)
  - And MS SQL Server @ Jim Gray Lab ..

ORACLE	SPARC SuperCluster with T3-4 Servers	TPC-C 5.11.0 TPC-Pricing 1.5.0 Report Date December 2, 2010

Total System Cost	TPC-C Throughput	Price/Performance	Availability Date
$30,528,863USD	30,249,688 tpmC	$1.01USD/tpmC	June 1, 2011

Database Server Processors/Cores/Threads	Database Manager	Operating System	Other Software	Number of Users
SPARC T3 1.65GHz 108 / 1,728 / 13,824	Oracle Database 11g Release 2 Enterprise Ed. With Oracle Real Application Clusters and Partitioning	Oracle Solaris 10 09/10	Tuxedo CFS-R Tier 1 Oracle iPlanet Web Server	24,300,000

Clients

81 Sun Fire X4170M2
2.93GHz Intel
Xeon X5670 HC
48GB Memory
2 146GB SAS disk

Database Nodes

27 Sun SPARC T3-4 Servers
4 1.65GHz SPARC T3
512GB Memory
3 300GB 10K RPM SAS
4 8Gb/s FC HBA, 2 port
10GbE SFP+
5RU High

Storage

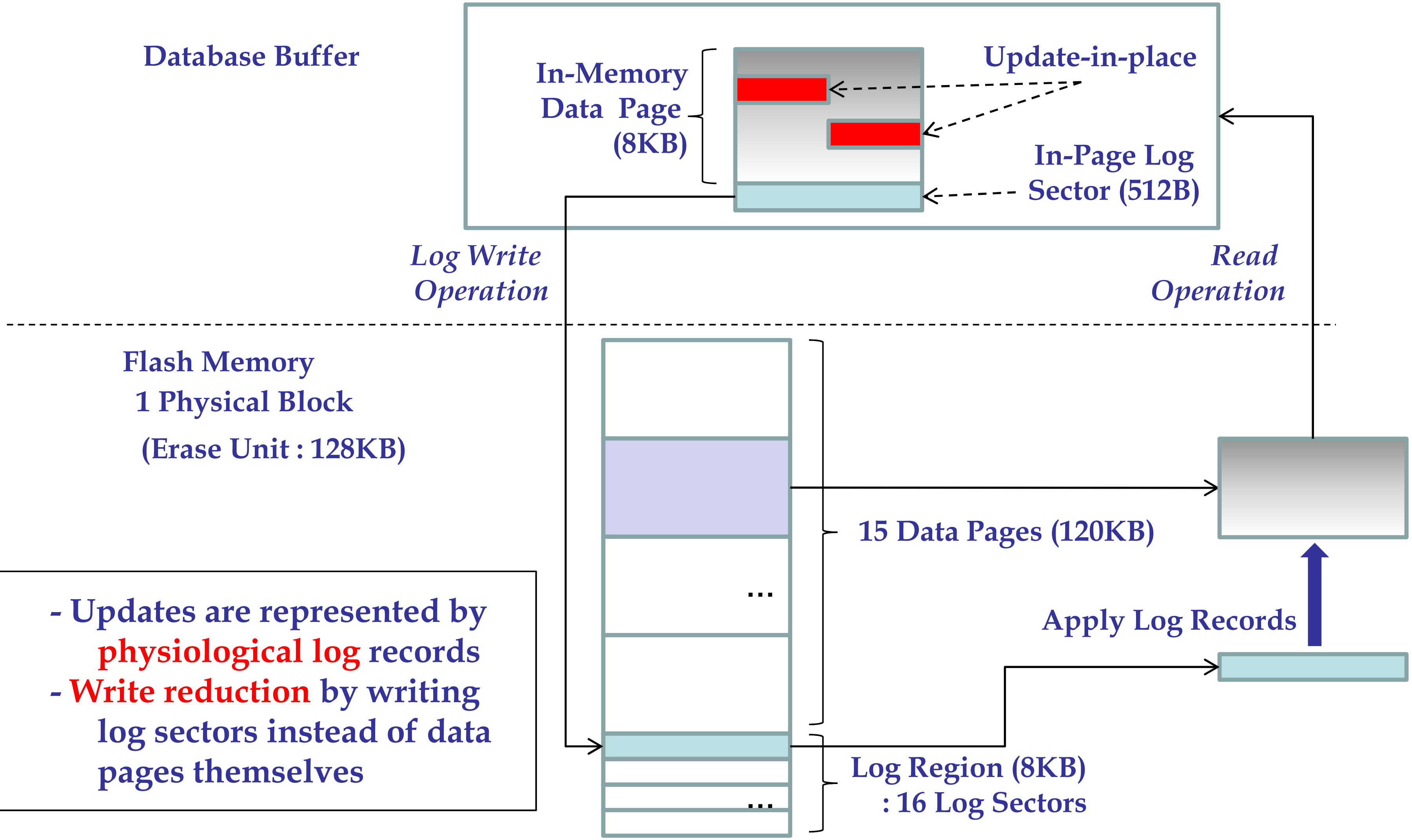
67 X4270M2 DATA COMSTAR
6 2TB 7.2K RPM SAS
2 Sun F5100 Flash Arrays

2 X4270M2 DATA COMSTAR
5 2TB 7.2K RPM SAS
2 Sun F5100 Flash Arrays

28 X4270M2 REDO COMSTAR
11 2TB 7.2K RPM SAS

System Component	Each Server Node		Each Client	
Processors/Cores/Threads and cache	4/64/512	SPARC T3 1.65GHz 6 MB L2 Cache	2/12/24	Intel Xeon X5670 12MB Smart Cache
Memory		512GB (13.5TB Total)		48GB
Disk Controllers	4	8Gb/s FC HBA 2 Port	1	8 port Internal SAS
OS Disks (each system)	3	300GB 10K RPM SAS	2	146GB 10K RPM SAS
External Storage (Equally visible to all T3-4 Server nodes)	11,040 720	24GB SSD Flash Modules 2TB 7.2K RPM SAS		
Total Storage		1.76PB		

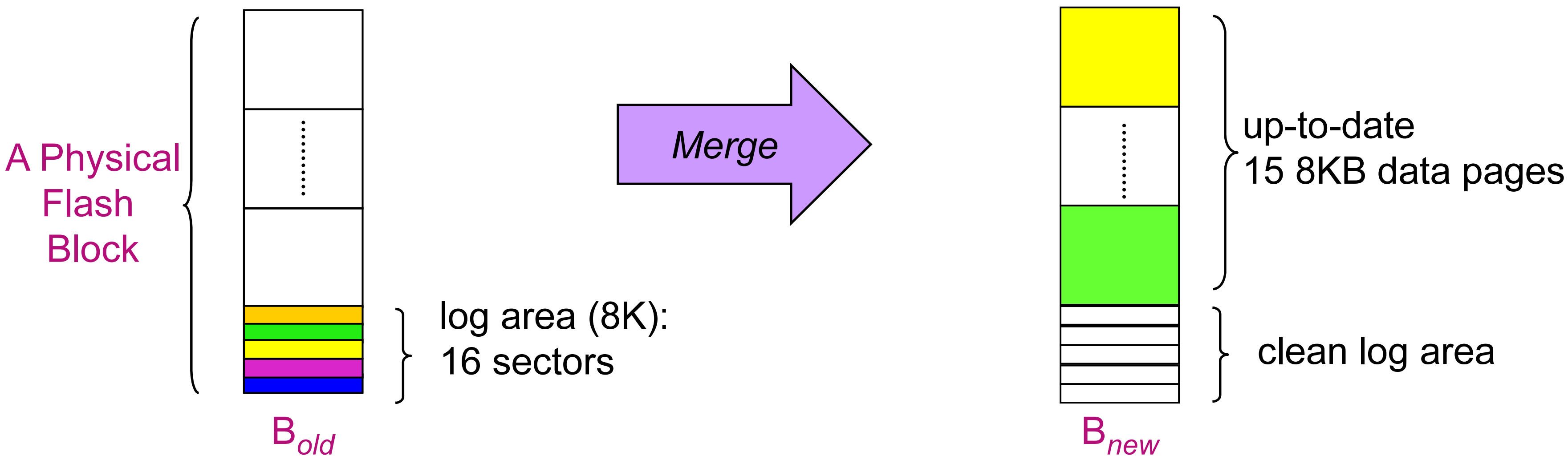
In-Page Logging (IPL) @ SIGMOD 2007



- Updates are represented by physiological log records
- Write reduction by writing log sectors instead of data pages themselves

Block Merge in In-Page Logging

- Merge: new internal operation in IPL



Transactional IPL (*TIPL*) @ ICDE 2011

Traditional In-Place Update

↓ *No in-place update*

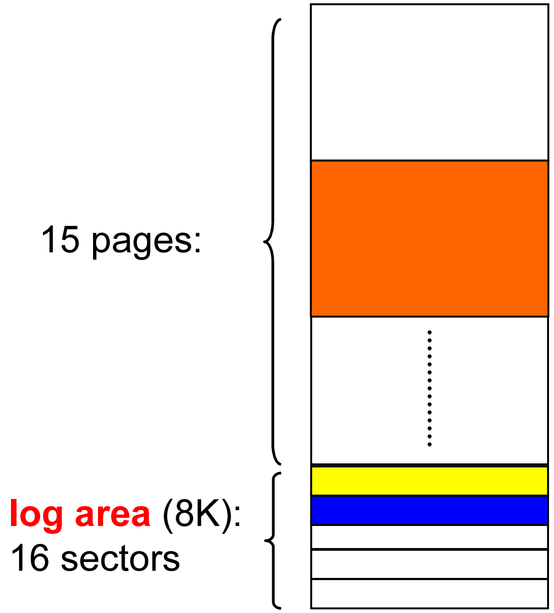
Log-Structured Approach

↓ *No mechanical latency*
Fast Read Speed

In-Page Logging Approach

↓ ***Page-oriented Redo Log***

New Recovery & Multiversion Store



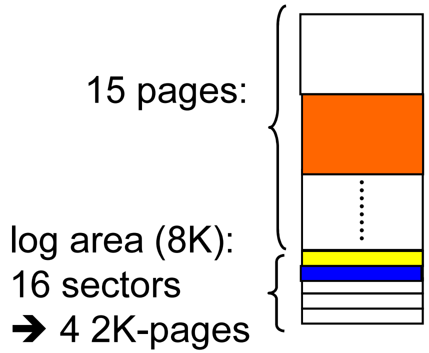
- Dual uses of IPL log

1. Better write performance
2. Transactional support

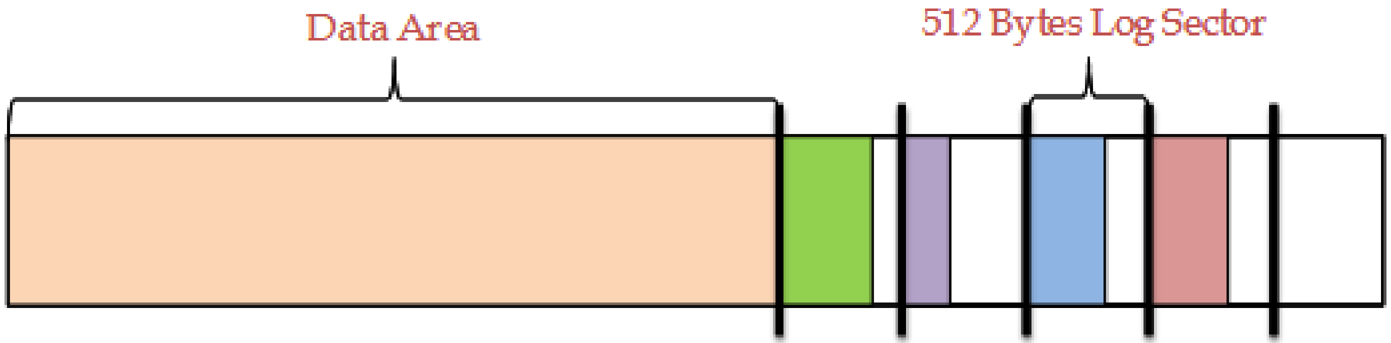
   (with nominal overhead)

IPL: Threats and a Reliefer

- IPL key point
  - Write reduction by capturing minimal change (or delta)
- Threats
  - The smallest unit of write is expected to increase: 512B → 2KB
    - ✓ The benefit of IPL can reduce
  - Read overhead
- PRAM



Internal Fragmentation



- Reduced write buffering
- Frequent merges
- Wear leveling

PRAM Researches in DB Communities

- Query processing using PRAM @ CIDR 2010

- PRAM as Log Device @ ICDE 2011

Flash Memory vs. PRAM

- The performance of PRAM is far lagging behind its promise

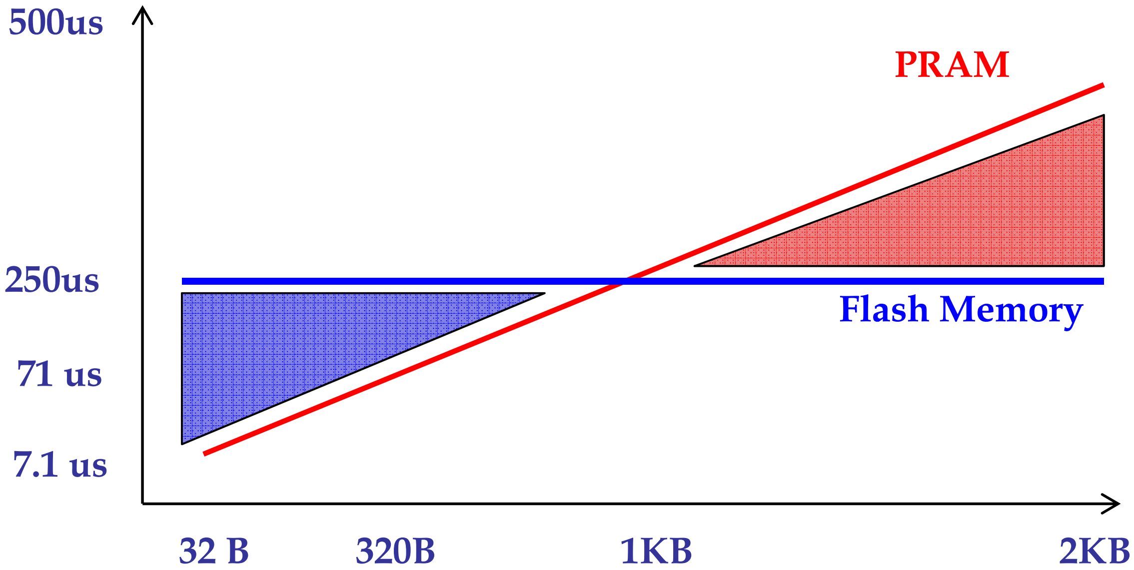
Media	Access time		
	Read	Write	Erase
Magnetic Disk†	12.7 ms (2KB)	13.7 ms (2KB)	N/A
NAND Flash‡	75 μs (2KB)	250 μs (2KB)	1.5 ms (128KB)
PCRAM¶	206 ns (32B)	7.1 μs (32B)	N/A
DRAM§	70 ns (32B)	70 ns (32B)	N/A

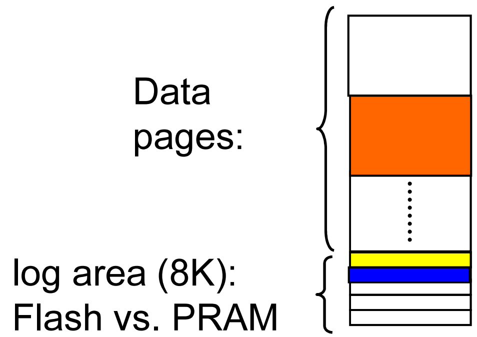
†Disk: Seagate Barracuda 7200.7 ST380011A;
‡NAND Flash: Samsung K9F8G08U0M 16Gbits SLC NAND [15];
¶PCRAM: Samsung 90nm 512Mb PRAM [8];
§DRAM: Samsung K4B4G0446A 4Gb DDR3 SDRAM [16]

Table 1: Access Speed: Magnetic disk vs. NAND Flash vs. PCRAM vs. DRAM

Flash vs. PRAM

- Write performance of PRAM





- Key difference b/w Flash and PRAM: (from IPL viewpoint)
  - Faster read/write latency for small size data
  - Byte-addressability for read and write

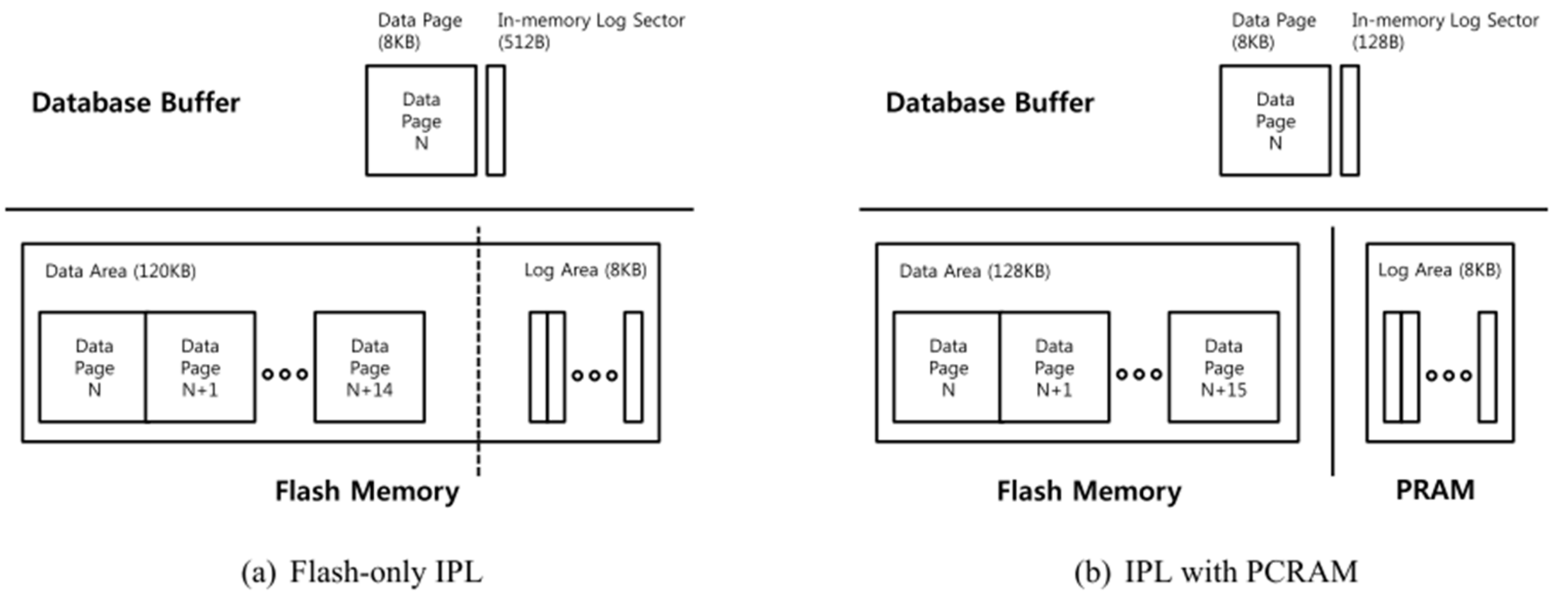
A Personal Prediction on Flash and PRAM

- Although some advocates of non-volatile memory predict that flash memory will give way to non-volatile memory soon(e.g. by 2012),

- We believe that they will co-exist, complementing each other, for a while until the hurdles in its manufacturing process are lifted and non-volatile memory becomes commercially competitive in both capacity and price.

- Vendors did not find any killer application for PRAM.
  - Chicken and egg dilemma!

IPL-P: IPL with PRAM



- Advantages
  - Fast write latency for small log data
  - Delay merge operation (e.g. 4 writes → 80 writes)
  - Reduce (or almost hide) the read overhead of IPL
  - Can use commercial Flash SSDs (even MLC-based SSD)

IPL–P: Performance (Simulation)

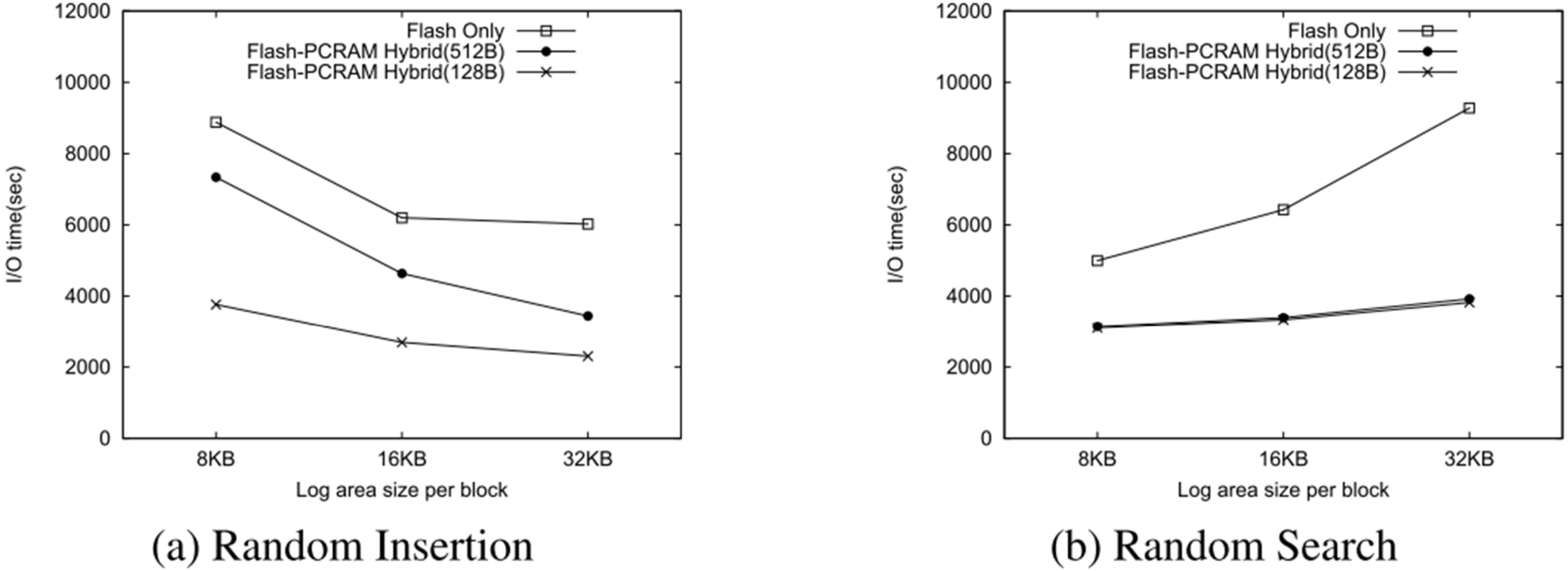
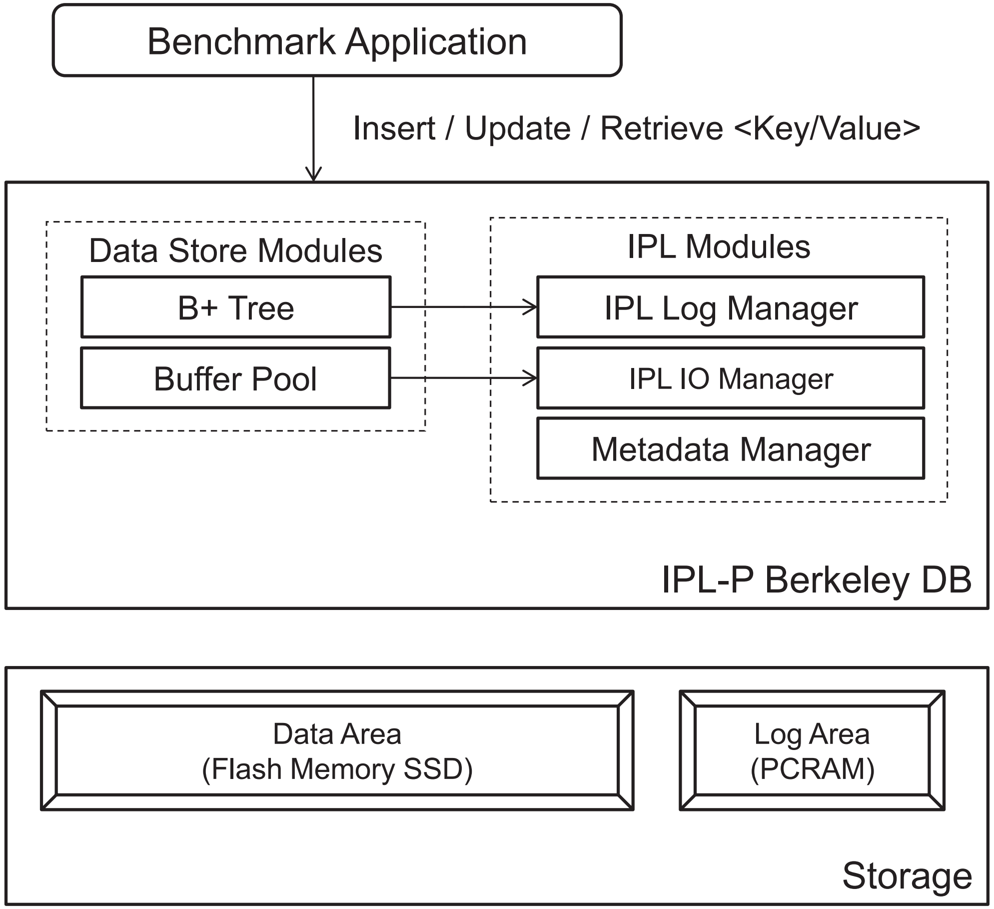


Figure 2: IPL Performance: Flash-only vs. Hybrid (5M records)

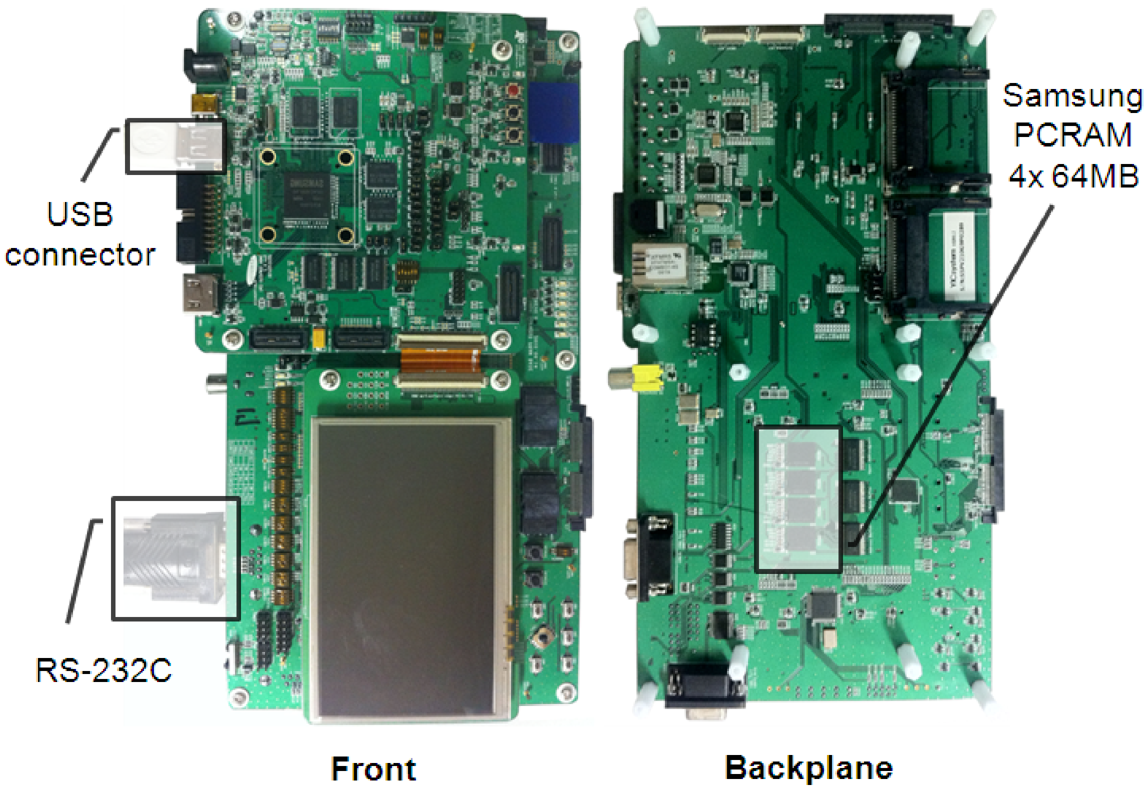
IPL –P: Performance on Real Board

- System Architecture



IPL –P: Performance on Real Board (2)

- Hardware

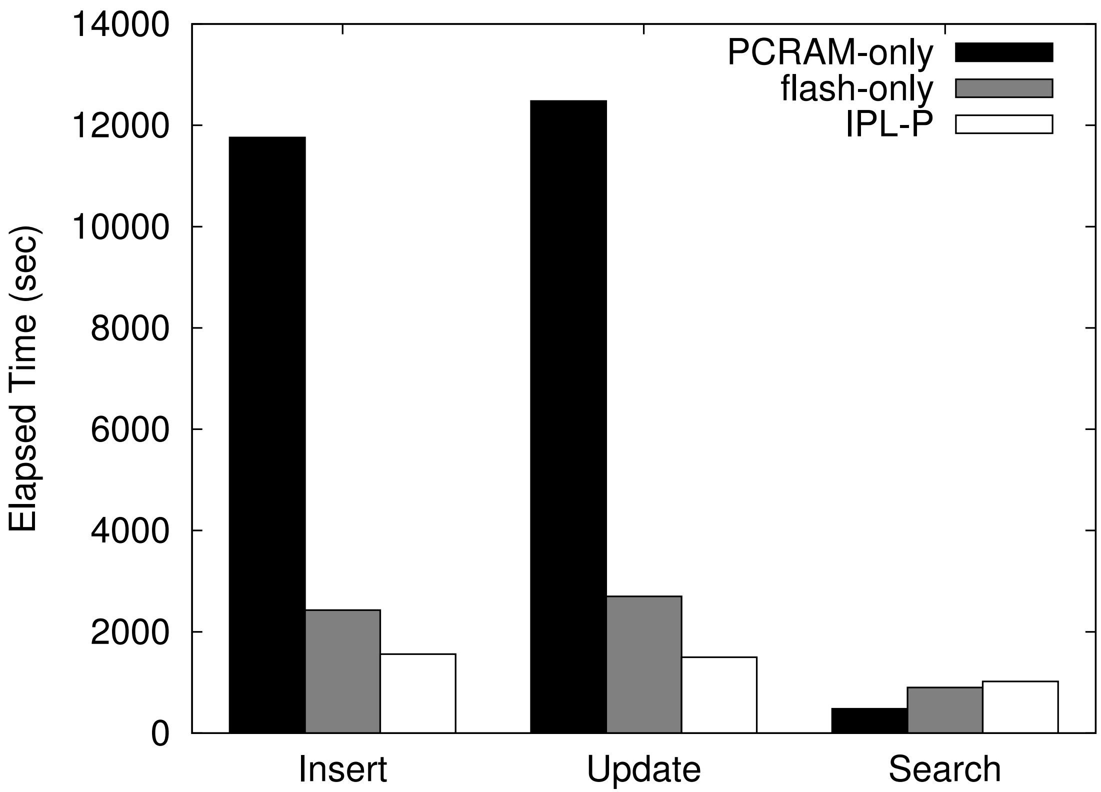


IPL –P: Performance on Real Board (3)

- Flash vs. PRAM: On-Board Performance

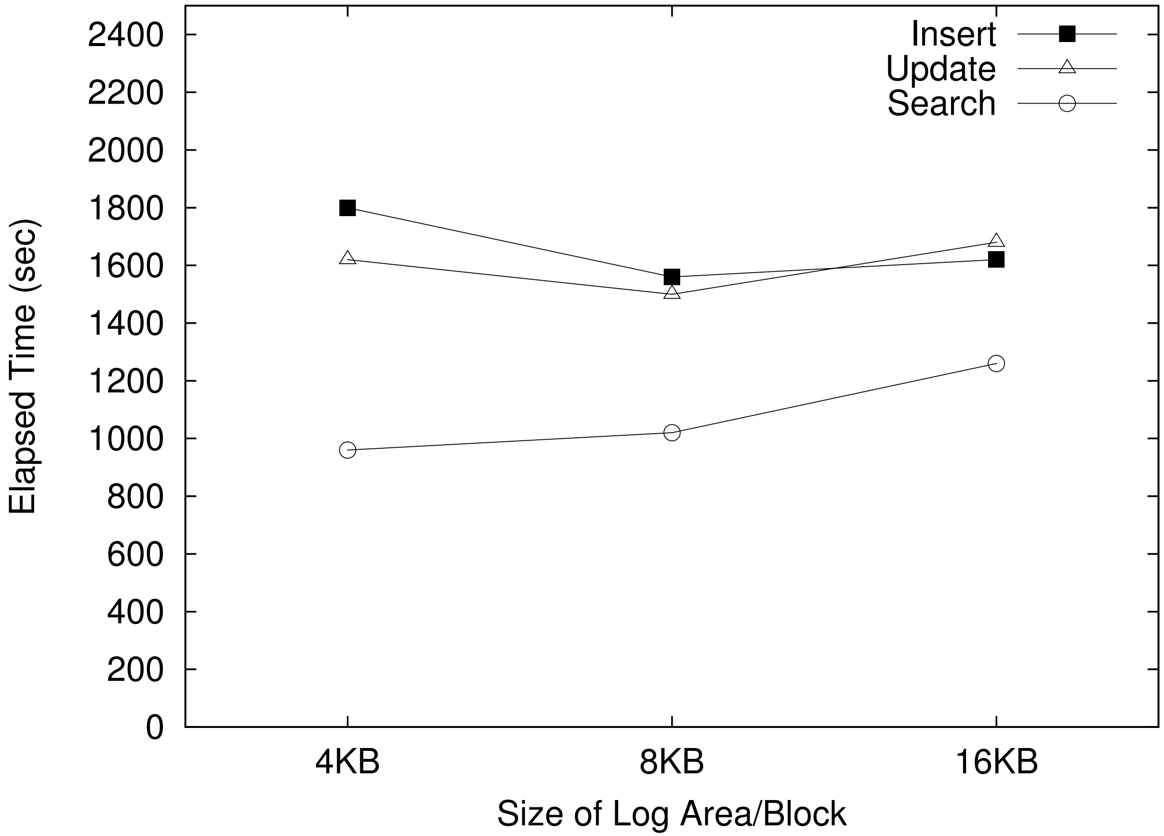
IPL –P: Performance on Real Board (4)

- With 8K log area



IPL –P: Performance on Real Board (5)

- By varying log area size



Conclusion and Future Works

- Flash memory and PRAM will complement each other ..
- As a model case of hybrid storage design based on flash memory and PRAM, we proposed IPL-P

- Future works
  - DIMM module?
  - Implement TIPL-P using MySQL inno DB storage engine