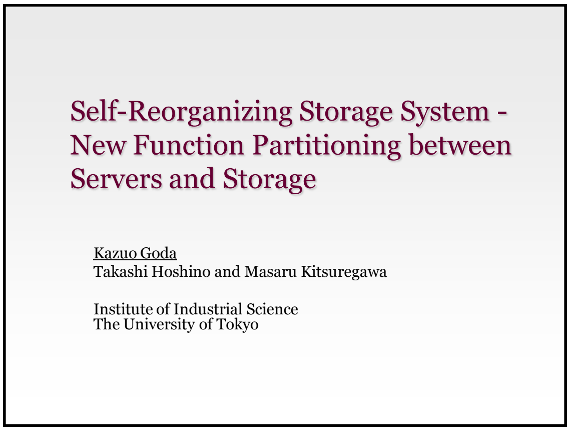
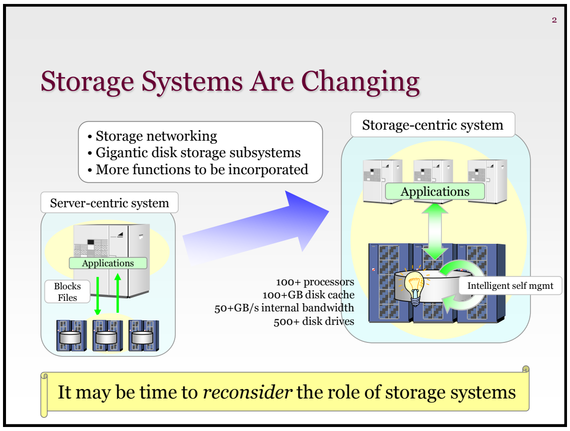


Self-Reorganizing Storage System - New Function Partitioning between Servers and Storage

Kazuo Goda
Takashi Hoshino and Masaru Kitsuregawa

Institute of Industrial Science
The University of Tokyo

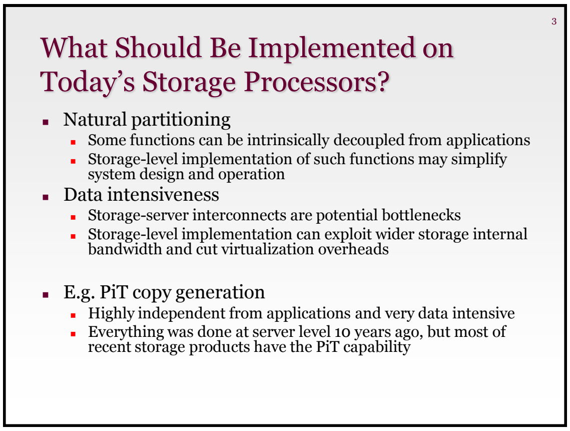


Storage Systems Are Changing

- Storage networking
- Gigantic disk storage subsystems
- More functions to be incorporated

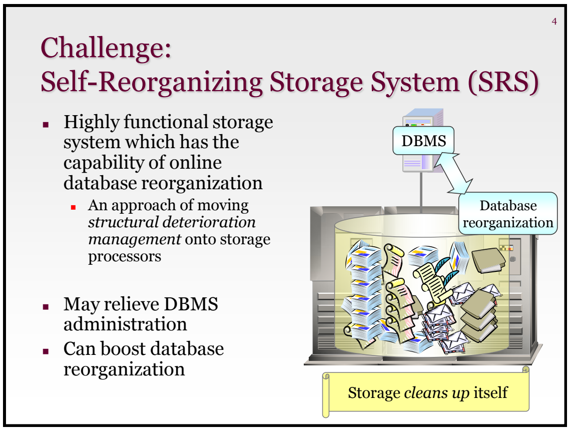
100+ processors
100+GB disk cache
50+GB/s internal bandwidth
500+ disk drives

It may be time to *reconsider* the role of storage systems



What Should Be Implemented on Today's Storage Processors?

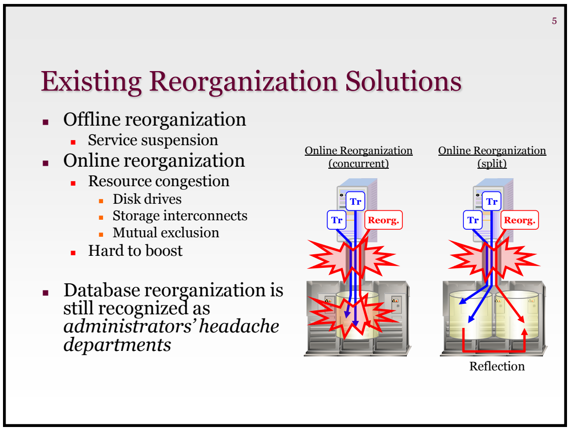
- Natural partitioning
  - Some functions can be intrinsically decoupled from applications
  - Storage-level implementation of such functions may simplify system design and operation
- Data intensiveness
  - Storage-server interconnects are potential bottlenecks
  - Storage-level implementation can exploit wider storage internal bandwidth and cut virtualization overheads
- E.g. PiT copy generation
  - Highly independent from applications and very data intensive
  - Everything was done at server level 10 years ago, but most of recent storage products have the PiT capability



Challenge: Self-Reorganizing Storage System (SRS)

- Highly functional storage system which has the capability of online database reorganization
  - An approach of moving *structural deterioration management* onto storage processors
- May relieve DBMS administration
- Can boost database reorganization

Storage *cleans up* itself



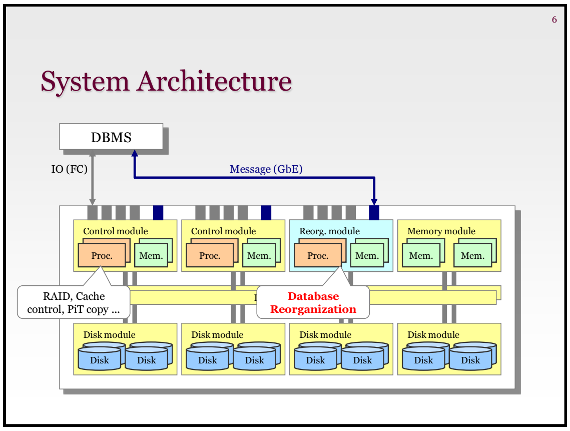
Existing Reorganization Solutions

- Offline reorganization
  - Service suspension
- Online reorganization
  - Resource congestion
    - Disk drives
    - Storage interconnects
    - Mutual exclusion
  - Hard to boost
- Database reorganization is still recognized as *administrators' headache departments*

Online Reorganization (concurrent)

Online Reorganization (split)

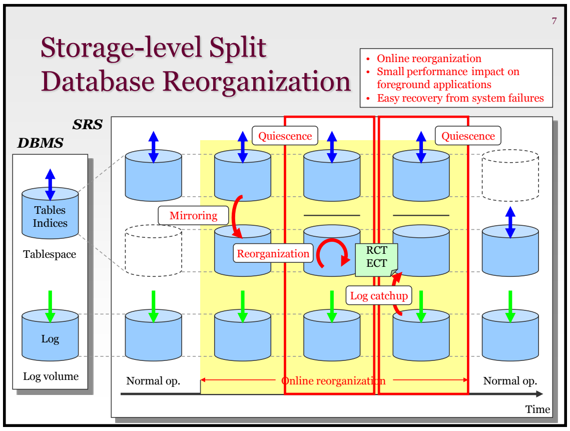
Reflection



System Architecture

RAID, Cache control, PiT copy ...

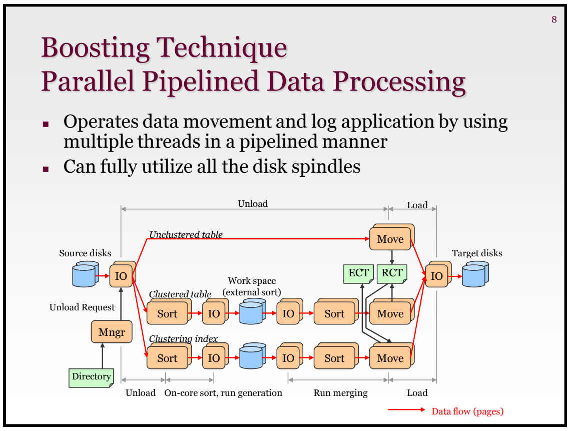
Database Reorganization



7

Storage-level Split Database Reorganization

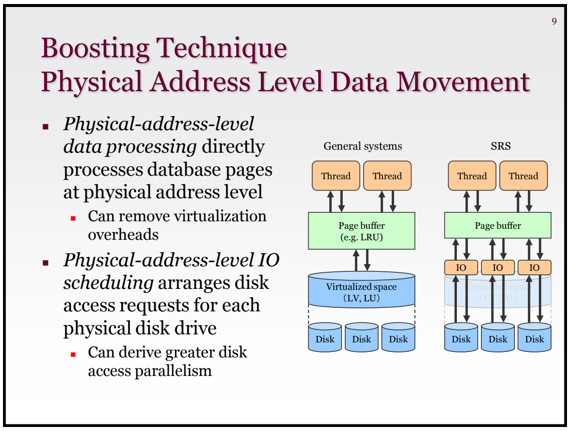
- Online reorganization
- Small performance impact on foreground applications
- Easy recovery from system failures



8

Boosting Technique Parallel Pipelined Data Processing

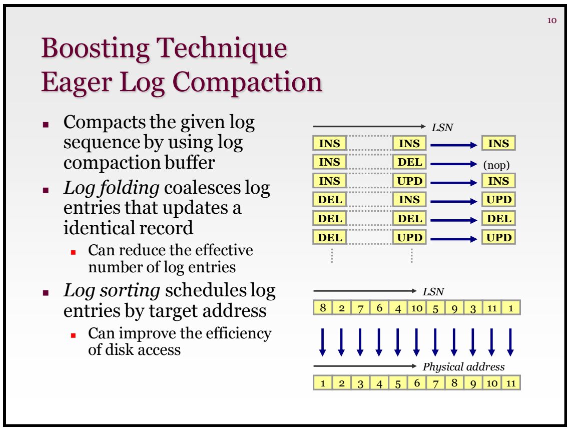
- Operates data movement and log application by using multiple threads in a pipelined manner
- Can fully utilize all the disk spindles



9

Boosting Technique Physical Address Level Data Movement

- *Physical-address-level data processing* directly processes database pages at physical address level
  - Can remove virtualization overheads
- *Physical-address-level IO scheduling* arranges disk access requests for each physical disk drive
  - Can derive greater disk access parallelism



10

Boosting Technique Eager Log Compaction

- Compacts the given log sequence by using log compaction buffer
- *Log folding* coalesces log entries that updates a identical record
  - Can reduce the effective number of log entries
- *Log sorting* schedules log entries by target address
  - Can improve the efficiency of disk access

LSN

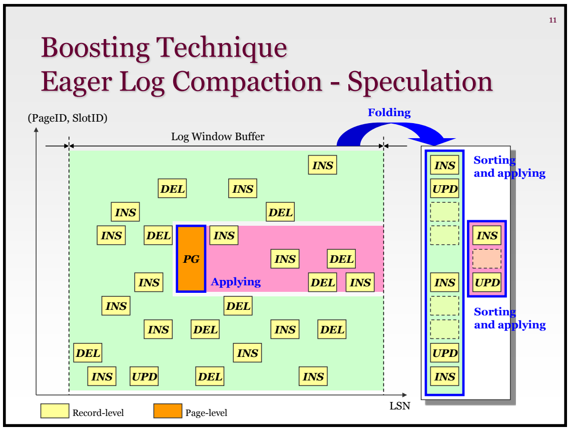
INS	INS	INS
INS	DEL	(nop)
INS	UPD	INS
DEL	INS	UPD
DEL	DEL	DEL
DEL	UPD	UPD

LSN

8	2	7	6	4	10	5	9	3	11	1

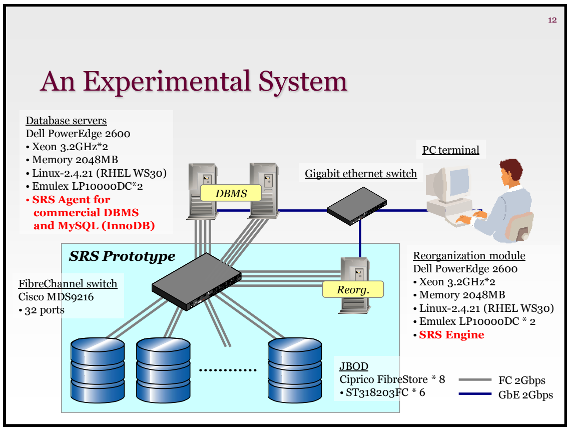
Physical address

1	2	3	4	5	6	7	8	9	10	11



11

Boosting Technique Eager Log Compaction - Speculation



12

An Experimental System

Database servers
Dell PowerEdge 2600
- Xeon 3.2GHz*2
- Memory 2048MB
- Linux-2.4.21 (RHEL WS30)
- Emulex LP10000DC*2
- **SRS Agent for commercial DBMS and MySQL (InnoDB)**

PC terminal

Gigabit ethernet switch

Reorganization module
Dell PowerEdge 2600
- Xeon 3.2GHz*2
- Memory 2048MB
- Linux-2.4.21 (RHEL WS30)
- Emulex LP10000DC * 2
- **SRS Engine**

FibreChannel switch
Cisco MDS9216
- 32 ports

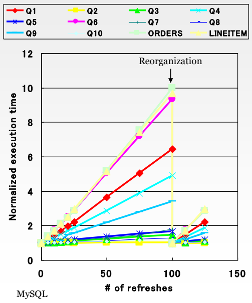
JBOD
Ciprico FibreStore * 8
- ST318203FC * 6

Reorganization Case Study #1
TPC-H Queries and Refresh Functions

- TPC-H benchmark
  - Queries (Q.1-10) Scans (orders, lineitem)
  - Refresh functions (RF1, RF2)
- Reorganization trigger

$$\frac{\text{Query Response Time}}{\text{Initial Query Response Time}} > 1000\%$$

Performance degradation could be kept within the specified range

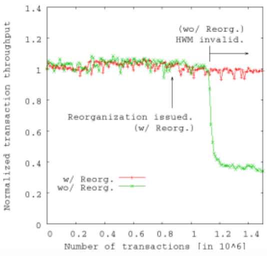


Reorganization Case Study #2
TPC-C with Data Migration

- TPC-C benchmark revised
  - Five types of regular transactions
  - Plus two procedures of data migration
    - Delete too old sales records
- Reorganization trigger

$$\frac{\text{Free space size reserved for insertion}}{\text{Table space size}} < 5\%$$

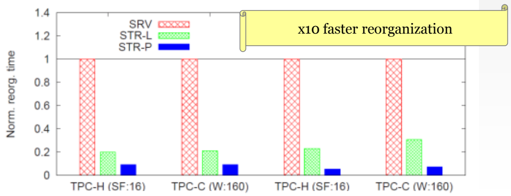
Sudden performance degradation could be avoided



Commercial DBMS

Reorganization Boosting

SRV
Reorganization utility packaged in DBMS: server-level reorganization
STR-L
SRS prototype: storage-level reorganization without physical address awareness
STR-P
SRS prototype: storage-level reorganization with physical address awareness

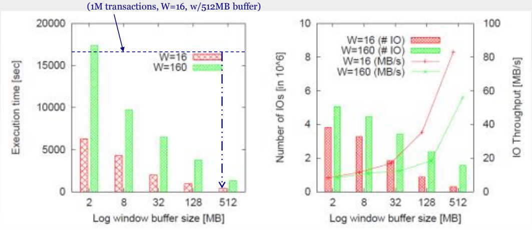


x10 faster reorganization

Commercial DBMS MySQL

Log Catchup Boosting

Transaction processing time
(1M transactions, W=16, w/512MB buffer)



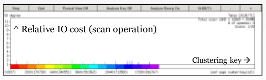
x40 faster than transaction processing

TPC-C, 1M Transactions
Commercial DBMS

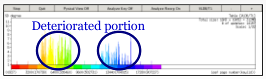
Online Monitoring of Database Structural Deterioration

- Core technique for autonomic database reorganization management
- SDMon: a prototyped tool for MySQL
  - Captures structural deterioration information from InnoDB storage engine on line, and
  - Visualizes the captured information at runtime

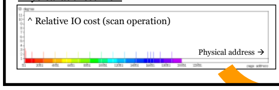
Clustering key view



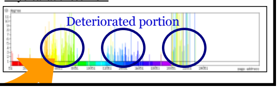
Clustering key view



Physical address view



Physical address view



Self-Reorganizing Storage System

- An approach of running database reorganization on storage processors
- Storage-level split database reorganization
- Boosting techniques to exploit storage hardware potential
  - Parallel pipelined data processing
  - Physical address level data movement
  - Eager log compaction
- Key findings
  - SRS can work for TPC-H and TPC-C on two different DBMS products
  - x10 faster reorganization than server-level reorganization tools
  - x40 faster log catchup than transaction processing
- SDMon: an online monitoring tool of database structural deterioration