High-Performance Sort Chip

Shinsuke Azuma, Takao Sakuma, Takashi Nakano,
Takaaki Ando, Kenji Shirai

azuma@icc.melco.co.jp

Mitsubishi Electric Corporation

Overview

- Background
- Algorithm
- Functional Features
- Architecture of Sort Chip
- Performance Evaluation
- Summary

Background

- Sorting
  - One of the most fundamental operations in databases
  - Key operation in integrating data warehouses from legacy databases
    Almost all the processing consists of sorting
    (by Dr. Ralph Kimball)
- Problem
  - Data movements occupy most of the execution time

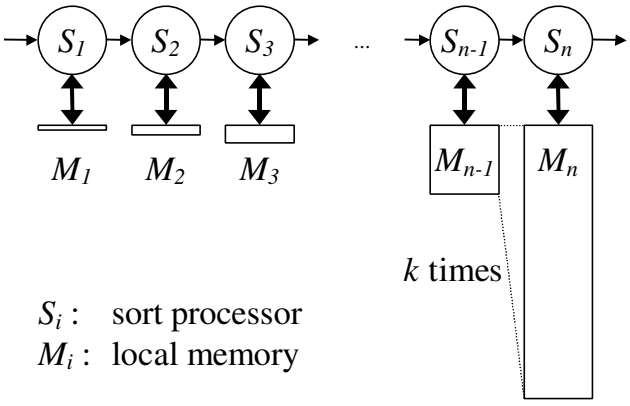
Our Approach

- Off-load such memory intensive operation onto a special hardware
- Employ multi-bank memory organization which improves memory throughput
- By installing the hardware sorter DIAPRISM/SS in a PCI slot, high-speed sorting can be achieved on a commodity PC

Algorithm

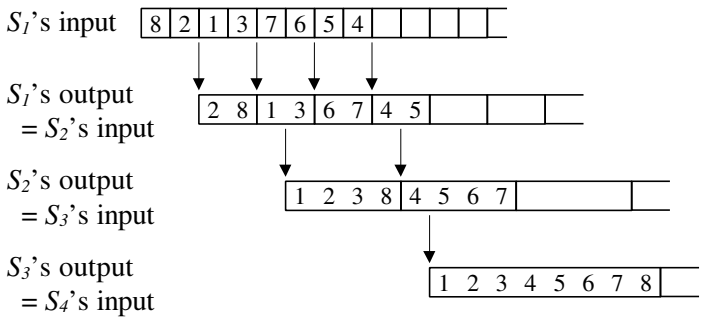
- "Pipeline merge sort algorithm" can sort data in a time proportional to the data size
- Multiple sort processors are connected linearly
- Each sort processor merges *k* strings (series of sorted records) and outputs a string *k* times as long as input
- Each sort processor has local memory to save the first *k*-1 strings

k-way Hardware Sorter



S_i : sort processor
M_i : local memory

2-way Pipeline Merge Sort



number/rectangle indicate record/string respectively

Number of Merge Way (k)

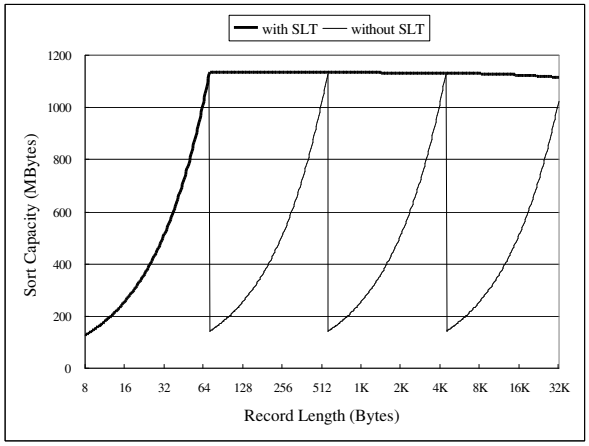
merge ways	# of comparators / chip	# of chips / sort system	# of DRAMs / sort system
2-way	1	24	200
4-way	3	12	120
8-way	7	8	96
16-way	15	6	84

The above system can handle 1GB data or 16 million records

Robustness for Record Length

- If the incoming record length is different from the one presumed in hardware, the memory utilization efficiency of the naive merge sort algorithm deteriorates.
- By String Length Tuning (SLT) algorithm, each sort chip dynamically chooses the number of merge ways from one to eight in order to fully utilize its memory.
- With SLT, the sort system can sort a constant amount of data regardless of the record length.

Robustness for Record Length



I/O Interface

- Three groups of I/O interfaces
  - Incoming / outgoing interfaces consist of 32-bit data lines, parity bits, and control signals.
  - Memory interface consists of 64-bit data lines, parity bits, and control signals.
  - Memory interface is connected directly to EDO or synchronous DRAMs.
- Data transfer rate
  - 4-byte data at each clock cycle between adjacent chips
  - 1-page (256-byte) data in 32-clock cycles between a chip and its local memory

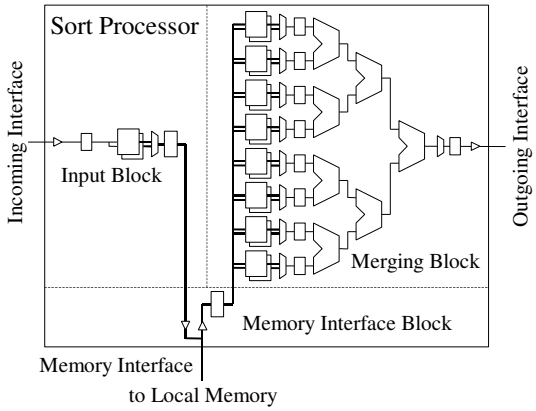
Architecture of Sort Chip

- Three major blocks
  - Input block has two 256-byte page buffers.
  - Input data is sent to the local memory through memory interface block.
  - Merging block has 8-way entries.
  - Each way has two 256-byte page buffers, and reads the corresponding string from the local memory.
  - Merging block merges the eight strings and outputs the result string.

Architecture of Sort Chip

- Merging process
  - Merging block has seven comparators in a tournament tree style.
  - Eight records are put into the comparator tree where each comparator selects the smaller one, and consequently the smallest record is output.
  - Repeating this process produces a sorted string eight times as long as an input string.
- Comparator capabilities
  - 4-byte-wide binary comparator
  - Selects one out of eight in each clock cycle
  - Maximum record length of 32K bytes, including sort key

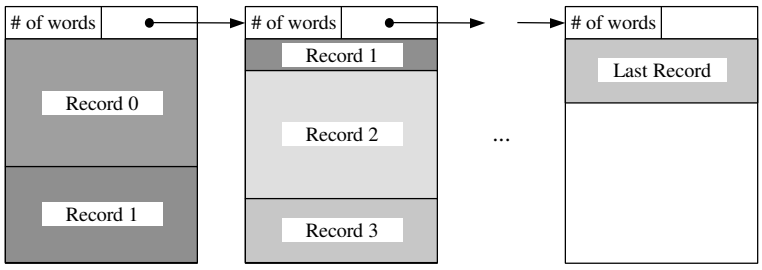
Data Path Diagram



High-Performance Memory Management

- Sort chip adopts page mode access to its local memory.
- Page size is 256 bytes.
- If the size of a string is larger than the page size, the string is stored in multiple pages.
- The pages are connected as a list and each page contains a pointer to the next page and the number of data words in the page, in addition to the data itself.

High-Performance Memory Management

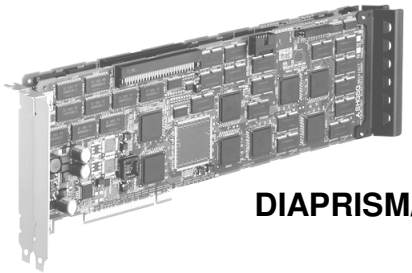


Physical Specifications

LSI Technology	0.35μm CMOS, 2 metal layers
Gate Counts	91K gates + 41K-bit RAM
Package	320-pin BGA (ball grid array)
Frequency	66MHz
Voltage	3.3V
Volume Production	May, 1998

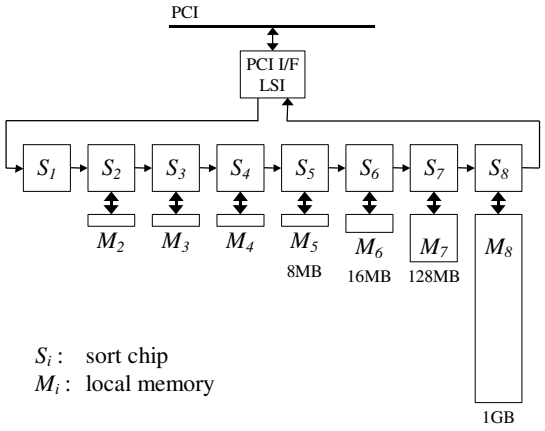
Hardware Sorter Board

- Eight sort chips and one PCI interface LSI
- 1GB memory is connected to the eighth sort chip
- Sorter can handle up to 16 million records or 1.1GB data



DIAPRISM/SS

Hardware Sorter Board



	SDRAMs
M_1	none
M_2	16Mb x 5
M_3	16Mb x 5
M_4	16Mb x 5
M_5	16Mb x 5
M_6	16Mb x 9
M_7	128Mb x 9
M_8	128Mb x 72

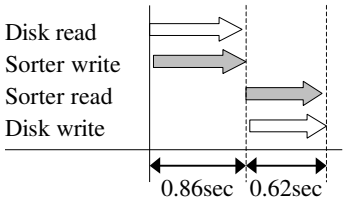
including parity

Sort Benchmark

- Datamation Benchmark
  - Elapse time to sort a million 100-byte records
  - Each record has a 10-byte key
  - Input records are in random order and output records must be in ascending order
  - Includes the time to read the input from disk and to write the output to disk
- World Record
  - 2.41 seconds by a 32-node cluster of UltraSPARCs (1998)
  - 1.18 seconds by a 16-node cluster of PentiumII PCs (1999)

Performance Evaluation

- System Configuration
  - Single-node PC with one PentiumII Xeon, 128MB main memory
  - Four Ultra2 SCSIs, four disks on each SCSI for data, one disk on another SCSI for OS
  - Hardware sorter in a PCI slot
  - Windows NT Server 4.0
- Benchmark Result
  - 1.48 seconds



Summary

- 8-way merge sort chip to achieve high-speed sorting for database processing or data warehousing
- Constant sort capacity regardless of the record length by proposed algorithm (SLT)
- 1GB of data can be sorted at one time by only eight chips
- Benchmark result proves that the performance is comparable to the world record with much less cost
- Marketing
  - the primary version of DIAPRISM/SS in July 1998
  - the enhanced version in October 1999