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

$M_{n-1} \rightarrow M_n$ $k$ times
2-way Pipeline Merge Sort

\[ S_i \text{'s input} = \begin{array}{cccccc} 8 & 2 & 1 & 3 & 7 & 6 & 5 & 4 \end{array} \]
\[ S_i \text{'s output} = \begin{array}{cccccc} 2 & 8 & 1 & 3 & 6 & 7 & 4 & 5 \end{array} \]
\[ S_2 \text{'s output} = \begin{array}{cccccc} 1 & 2 & 3 & 8 & 4 & 5 & 6 & 7 \end{array} \]
\[ S_3 \text{'s output} = \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \]

number/rectangle indicate record/string respectively

Number of Merge Way (k)

<table>
<thead>
<tr>
<th>merge ways</th>
<th># of comparators / chip</th>
<th># of chips / sort system</th>
<th># of DRAMs / sort system</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-way</td>
<td>1</td>
<td>24</td>
<td>200</td>
</tr>
<tr>
<td>4-way</td>
<td>3</td>
<td>12</td>
<td>120</td>
</tr>
<tr>
<td>8-way</td>
<td>7</td>
<td>8</td>
<td>96</td>
</tr>
<tr>
<td>16-way</td>
<td>15</td>
<td>6</td>
<td>84</td>
</tr>
</tbody>
</table>

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.

---

![Graph showing robustness for record length with and without SLT](graph.png)
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.
Physical Specifications

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LSI Technology</td>
<td>0.35µm CMOS, 2 metal layers</td>
</tr>
<tr>
<td>Gate Counts</td>
<td>91K gates + 41K-bit RAM</td>
</tr>
<tr>
<td>Package</td>
<td>320-pin BGA (ball grid array)</td>
</tr>
<tr>
<td>Frequency</td>
<td>66MHz</td>
</tr>
<tr>
<td>Voltage</td>
<td>3.3V</td>
</tr>
<tr>
<td>Volume Production</td>
<td>May, 1998</td>
</tr>
</tbody>
</table>

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
Hardware Sorter Board

---

**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 UltraSPARCis (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 SCISIs, 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

<table>
<thead>
<tr>
<th></th>
<th>Disk read</th>
<th>Sorter write</th>
<th>Sorter read</th>
<th>Disk write</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>0.86sec</td>
<td>0.62sec</td>
<td>0.86sec</td>
<td>0.62sec</td>
</tr>
</tbody>
</table>

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