Exploiting Disk Intelligence for Decision Support Databases

Kimberly Keeton  David A. Patterson
Hewlett-Packard Laboratories  University of California at Berkeley
kkeeton@hpl.hp.com  Patterson@cs.berkeley.edu

Third Workshop on Computer Architecture Evaluation using Commercial Workloads (CAECW ’00)
January 9, 2000

Motivation: Increasing I/O & Compute Needs

* Greg’s Law: Greg Papadopoulos, CTO, Sun Microsystems
  - DSS database I/O demand growth: 2X / 6-12 months
  - Storage capacity and associated processing

* Contributing factors:
  - Collect richer data (more detailed)
    - “Just-in-time” inventory
  - Keep longer historical record
  - Increased data access via network
  - Business consolidation

* Winter VLDB Survey (1997):
  - Telecomm., retail & financial DBs ~doubled from 1996 to 1997
Motivation: Architectural Trends

- More sophisticated & modularized disk drives
  - Increased disk-resident memory, processing
  - Fast serial lines replacing busses
  - By 2001, Seagate estimates 100-200 MIPS, <= 64 MB memory

- Communication trends
  - Switched networks overtake bus-based networks
  - Serial communication advances: Gbps serial I/O lines

- Processor trends
  - Emergence of low cost, low power embedded processors
  - Embedded integer performance: ~ 1/2 desktop performance
  - Integrated logic and DRAM on same chip

Motivation: Intelligent Disks

- Intelligent disk (IDISK):
  - Low cost, low power processor
  - Memory
  - Scalable, switch-based interconnect

- Longer-term (5 to 10 years):
  - Sufficient processing, memory for no front-end host?
Motivation: Performance Feasibility

- How well does IDISK perform for DSS workloads?
- How does IDISK performance compare with that of other popular server architectures?
- What’s the limiting factor(s) for performance?
  - Disk bandwidth?
  - Processor speed?
  - Memory capacity?
  - Network bandwidth?

Outline

- Motivation
- Methodology
  - TPC-D measurements
  - Scaled hardware configurations
  - Analytic models
- Case studies
  - Selection
  - Hash join
- Conclusions
Approach

- Analytic models of DSS queries
- Calibrate models using measurements from full-scale (100 GB) TPC-D DSS system
- Compare several DSS server architectures:
  - IDISK: thin-node cluster
  - Cluster of quad SMPs
  - Single large SMP
- Scaled up hardware and data sets

Estimated Instruction Counts per I/O

<table>
<thead>
<tr>
<th>Database Operation</th>
<th>Read vs. Write</th>
<th>Sequential vs. Random (I/O size)</th>
<th>Est. Inst. per I/O</th>
<th>Used in analysis:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scan+select+project+aggregate (simple)</td>
<td>Read</td>
<td>Sequential (64 KB)</td>
<td>800,000</td>
<td>Selection</td>
</tr>
<tr>
<td>Scan+select+project+aggregate (complex)</td>
<td>Read</td>
<td>Sequential (64 KB)</td>
<td>4,000,000</td>
<td>Selection</td>
</tr>
<tr>
<td>Scan+select+project+hash join (one-pass)</td>
<td>Read</td>
<td>Sequential (64 KB)</td>
<td>1,200,000</td>
<td>Hash join</td>
</tr>
<tr>
<td>Index scan + nested loops join</td>
<td>Read</td>
<td>Random (4 KB)</td>
<td>280,000</td>
<td>Index nested loops join</td>
</tr>
<tr>
<td>Write int. results to disk</td>
<td>Write</td>
<td>Random (8 KB)</td>
<td>400,000</td>
<td>Hash join</td>
</tr>
</tbody>
</table>

- Based on measurements of 100 GB TPC-D queries (single-stream)
  - 4-processor Pentium Pro-based server running Informix/NT 4.0
  - Simple scan (Q6), complex scan (Q1), simple hash join (Q4), complex hash join (Q5, Q8), simple index NL join (Q11)
Base Systems for Performance Study

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>NCR WorldMark 5200 w/ Teradata</th>
<th>HP 9000 V2500 Enterprise Server w/ Oracle8i</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processors per node</td>
<td>4 * 450 MHz</td>
<td>32 * 440 MHz</td>
</tr>
<tr>
<td>Mem. capacity per node</td>
<td>2 GB</td>
<td>32 GB</td>
</tr>
<tr>
<td>Disk capacity per node</td>
<td>40 * 9 GB</td>
<td>680 * 9.1 GB</td>
</tr>
<tr>
<td>Proc. interconnect B/W</td>
<td>120 MB/s</td>
<td>N/A</td>
</tr>
<tr>
<td>I/O interconnect B/W</td>
<td>264 MB/s (1 64b 33 MHz PCI)</td>
<td>2112 MB/s (8 64b 33 MHz PCI)</td>
</tr>
<tr>
<td>Nodes</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>Total processors</td>
<td>128</td>
<td>32</td>
</tr>
<tr>
<td>Total mem. capacity</td>
<td>64 GB</td>
<td>32 GB</td>
</tr>
<tr>
<td>Total disks</td>
<td>1280</td>
<td>680</td>
</tr>
</tbody>
</table>

* 1999 TPC-D 300 GB SF performance-leading configurations
* Assumed Seagate Cheetah 9LP characteristics: 28.9 MB/s

Back of the Envelope Benchmarks

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>IDISK04</th>
<th>“NCR04”</th>
<th>“HP04”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processors per node</td>
<td>1 * 2500 MHz</td>
<td>4 * 4500 MHz</td>
<td>32 * 4400 MHz</td>
</tr>
<tr>
<td>Mem. capacity per node</td>
<td>32 – 512 MB</td>
<td>20 GB</td>
<td>320 GB</td>
</tr>
<tr>
<td>Disks per node</td>
<td>1</td>
<td>21</td>
<td>672</td>
</tr>
<tr>
<td>Proc. interconnect B/W</td>
<td>600 MB/s</td>
<td>600 MB/s</td>
<td>N/A</td>
</tr>
<tr>
<td>I/O interconnect B/W</td>
<td>N/A</td>
<td>800 MB/s (1 64b 100 MHz PCI)</td>
<td>6400 MB/s (8 64b 100 MHz PCI)</td>
</tr>
<tr>
<td>Nodes</td>
<td>672</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>Total processors</td>
<td>672</td>
<td>128</td>
<td>32</td>
</tr>
<tr>
<td>Total mem. capacity</td>
<td>21.5 –344 GB</td>
<td>640 GB</td>
<td>320 GB</td>
</tr>
<tr>
<td>Total disks</td>
<td>672</td>
<td>672</td>
<td>672</td>
</tr>
</tbody>
</table>

* Projected 2004 systems based on today’s configurations
* All configurations have 672 disks:
  - Per disk: 95.4 GB, 154.6 MB/s
* IDISK processor speed ~ 1/2 central processor speed
* IDISK memory varied (128 - 256 MB typical)
Case Study 1: Selection

- Scaled up data sets
  - 1000 GB scale factor data set
- Query based on TPC-D Q1, Q6
  - Scan 6 billion 145 B rows
- Assume sequential table scan used (no materialized views)
- Computation per I/O
  - Simple: $0.8 \times 10^6$ inst (Q6)
  - Complex: $4.0 \times 10^6$ inst (Q1)

4 rows * 69 B/row

SORT BY

GROUP BY

SCAN lineitem

5.94 billion rows * 145 B/row

Selection

- Embarrassingly parallel task
  - Simple: I/O-limited
  - Complex: compute-limited
- What about faster interconnect?
  - Assume 10X the scaled speed
- IDISK04 Simple Speedup (10X):
  - NCR04: 2.4X, HP04: 9.6X
  - Now also compute-limited
- IDISK04 Complex Speedup (10X):
  - NCR04: 2.9X, HP04: 11.9X
  - (Same: compute-limited)

Scan/selection is best-case scenario for IDISK
- Embarrassingly parallel
- Streaming data access
Case Study 2: Hash Join

Query based on TPC-D Q12
- Hybrid hash join
- Order: 910M rows x 135 B
- Lineitem: 3.5B rows x 145 B
- Memory-sensitive algorithm
  - Build hash table from first relation
  - Probe hash table from second relation
  - 1.2 M inst. per I/O (1-pass)
- Assume network comm. for order table

2 rows * 25 B/row

SORT BY

GROUP BY

HASH JOIN

SCAN order ("build")

0.91 bill. rows *
135 B/row

SCAN lineitem ("probe")

3.49 bill. rows *
145 B/row

Hash-Join

IDISK04 256 MB:
- Computation dominates
- NCR04, HP04: PCI-limited
  - NCR04: 3.7X
  - HP04: 11.2X
- 10X PCI: Compute-limited
- IDISK04 Speedups (10X PCI):
  - NCR04 10X: 3.5X
  - HP04 10X: 10.6X
- What if IDISK memory isn’t big enough?
**IDISK Memory Sensitivity for Hash Join**

- Hash-join is memory-sensitive algorithm
  - “One-pass” if data fits in memory
  - “Two-pass” if data too big to fit into memory
- Crossover point: ~200 MB

**IDISK04 256 MB:**
- Computation dominates

**IDISK04 128 MB:**
- Temp. I/O costs dominate
- Performance within 15% of NCR04

**Hash Join Two-Pass Crossover Points**

- How much memory required per node for our hash join query to be one pass?
  - Assume 8 KB comm. buffers
- Small datasets (up to 30 GB SF)
  - Limited by size of communication buffers
- Larger datasets (100 GB and above)
  - Limited by size of build relation
Conclusions

- DSS database workloads present challenging I/O demands
- Analytic modeling based on measurements of full-scale DSS system
- IDISK system achieves high-performance and scalability for variety of DSS operations
  - Outperforms cluster and SMP systems with faster processors and higher aggregate memory capacity by 2X to 12X
  - Due to increased I/O parallelism & larger aggregate computation
- IDISK can trade off disk I/O B/W for memory capacity
  - Two-pass hash join: ~15% slowdown over cluster system