Characterization of I/O for TPC-C and TPC-H workloads

Don DeSota
IBM
NUMA-Q
desotad@us.ibm.com
Overview

• TPC-H
  – System configuration
  – Data

• TPC-C
  – System configuration
  – Data

• Synthetic workload

• Conclusions
TPC-H System Configuration

• Based on ~6 month old measurements
• 300 GB TPC-H
• 16 Quads
  – 700 MHz PIII Xeon with 2MB L2
  – 64 Clarion DAEs
    • 640 9 GB disks
    • 64 FC PCI controllers
  – 64 GB Memory
  – Dynix/Ptx V4.5.1
  – Ptx/SVM 2.2.1
  – DB2 UDB V7.1 - 128MB disk blocks
Quad Block Diagram

Processor External Bus @ 90-100 MHz

Xeon Xeon Xeon Xeon

DOBIC

LYNX2 Board

OMM SLMM

SCI Bus

External Memory Bus

Quad Memory

(1 GB-8 GB)

MIOC

82450NX

64 Bit PCI Bus
30-33 MHz

PXB-B

PCI-B0
PCI-B1
PCI-B2
PCI-B3

PCI & ISA Interrupts

PCI & ISA

RCGs (2)
MUXs (4)

MDC ISA

External Memory Bus

64 Bit PCI Bus
30-33 MHz

PXB-A

PCI-A0
PCI-A1
PCI-A2
PCI-A3

PIIX4

USB IDE GPIO

I/O APIC

Super I/O

PIIIx

USB IDE

I/O APIC

RTC, NVRAM

Flash

Floppy Keyboard Mouse

2 Serial 1 Parallel Ports

64 Bit PCI Bus
30-33 MHz

SCI Bus

30-33 MHz

Xeon Xeon Xeon Xeon

64 Bit PCI Bus
30-33 MHz
Configuration Details

IQ Link - connecting 16 quad processors

Single OS - Dynix/ptx

Single Instance of Database - DB2

16 partitions

64 Clarion DAEs

each with 10 - 9 GB disks

with direct Fibre channel connection
TPC-H I/O Characteristics

- Mostly reads
- 128MB blocks
- Sequential I/O
- Partitionable
TPC-H I/O Throughput
NUMA-Q 200/DB2 300GB TPC-H Power Test
IO Throughput (MB/s)

Queries

1/21/01 Don DeSota IBM
TPC-H IOs

TPC-H I/Os

IO/s

Write

Read

1/21/01 Don DeSota IBM

Q9 Q9 Q9 Q6 Q17 Q17 Q18 Q18 Q18 Q8 Q8 Q21 Q21 Q21 Q21 Q21 Q13 Q13 Q3 Q3 Q22 Q22 Q16 Q15 Q1 Q10 Q10 Q19 Q5 Q7 Q7 Q12
Query 10

Query 10 & SHY: Returned Item Reporting

TPC-H Business Question: The Returned Item Reporting Query finds the top 20 customers, in terms of their effect on lost revenue for a given quarter, who have returned parts. The query considers only parts that were ordered in the specified quarter. The query lists the customer’s name, address, nation, phone number, account balance, comment information and revenue lost. The customers are listed in descending order of lost revenue. Revenue lost is defined as \( \text{sum}(l_{\text{extendedprice}}*(1-l_{\text{discount}})) \) for all qualifying lineitems.

```
select
c_custkey, c_name, sum(l_extendedprice * (1 - l_discount)) as revenue, c_acctbal, n_name, c_address, c_phone, c_comment
from customer, orders, lineitem, nation
where c_custkey = o_custkey
  and l_orderkey = o_orderkey
  and o_orderdate >= date '[DATE]' 
  and o_orderdate < date '[DATE]' + interval '3' month
  and l_returnflag = 'R'
  and c_nationkey = n_nationkey
group by
c_custkey, c_name, c_acctbal, c_phone, n_name, c_address, c_comment
order by revenue desc;
```

Functionality: Query 10 is a 4-table join of three large tables and one look-up table. Customer detail is returned by this query, alongside of only one column of summarized data. 1/28 of the Order rows qualify based on date, while 1 in 4 Lineitems match the criteria of having a return flag of 'R'. Millions of rows are returned from this query, but the final sort determines which 20 rows are to be displayed in the answer set.
Query 10 Breakdown

Query 10 I/O Throughput

![Chart showing I/O throughput for Query 10 with MB/s values and Write vs Read comparison.]

1/21/01 Don DeSota                IBM 10
TPC-C System Configuration

• Based on ~18 month old measurements

• TPC-C V.3

• 1 - 4 Quads
  – 495 MHz PIII Xeon with 2MB L2
  – Clarion DAEs
  – 64 GB Memory
  – FC Switched Fabric
  – Dynix/Ptx
  – Ptx/SVM
  – 2K disk blocks
Configuration Details

IQ Link connecting 4 quad processors

Single OS - Dynix/ptx

Single DB

FC Switch connecting 4 quad and DAEs

Clarion DAEs
with switched Fibre channel connection and multipath I/O
Multipath I/O

• With multipath I/O, the all DMA transfers are to/from local memory addresses.
TPC-C I/O Characteristics

• Mostly reads
• 2K blocks
• Random I/O
TPC-C I/O Throughput

TPC-C V.3 I/O Throughput

1/21/01  Don DeSota  IBM
TPC-C I/Os

TPC-C V.3 I/Os

Quads

1 2 3 4

Write

Read

1/21/01  Don DeSota  IBM  16
Database I/O Synthetic Tests

- Run on the TPC-H system
- Large Sequential Reads at 128K blocksize, simulating Full Table Scans
- Random Reads at 2K blocksize, simulating Index Range Scans
- 80% Random Reads, 20% Random Writes at 2K blocksize, Simulating Insert and Update DML.

<table>
<thead>
<tr>
<th></th>
<th>IO/sec</th>
<th>MB/sec</th>
<th>% Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>128K Sequential Reads</td>
<td>30,080</td>
<td>3760</td>
<td>14</td>
</tr>
<tr>
<td>2K Random Reads</td>
<td>88,345</td>
<td>176</td>
<td>13</td>
</tr>
<tr>
<td>2K 80% Reads, 20% Writes</td>
<td>83,651</td>
<td>167</td>
<td>12</td>
</tr>
</tbody>
</table>
Conclusions

- TPC-C has lower bandwidth but higher I/Os requirement.
- TPC-H can push I/O throughput close to the system maximum.
- An I/O fabric and multipath I/O is required in a multinode system for TPC-C to achieve maximum throughput.
- TPC-H I/O is highly variable throughout the run.