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

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

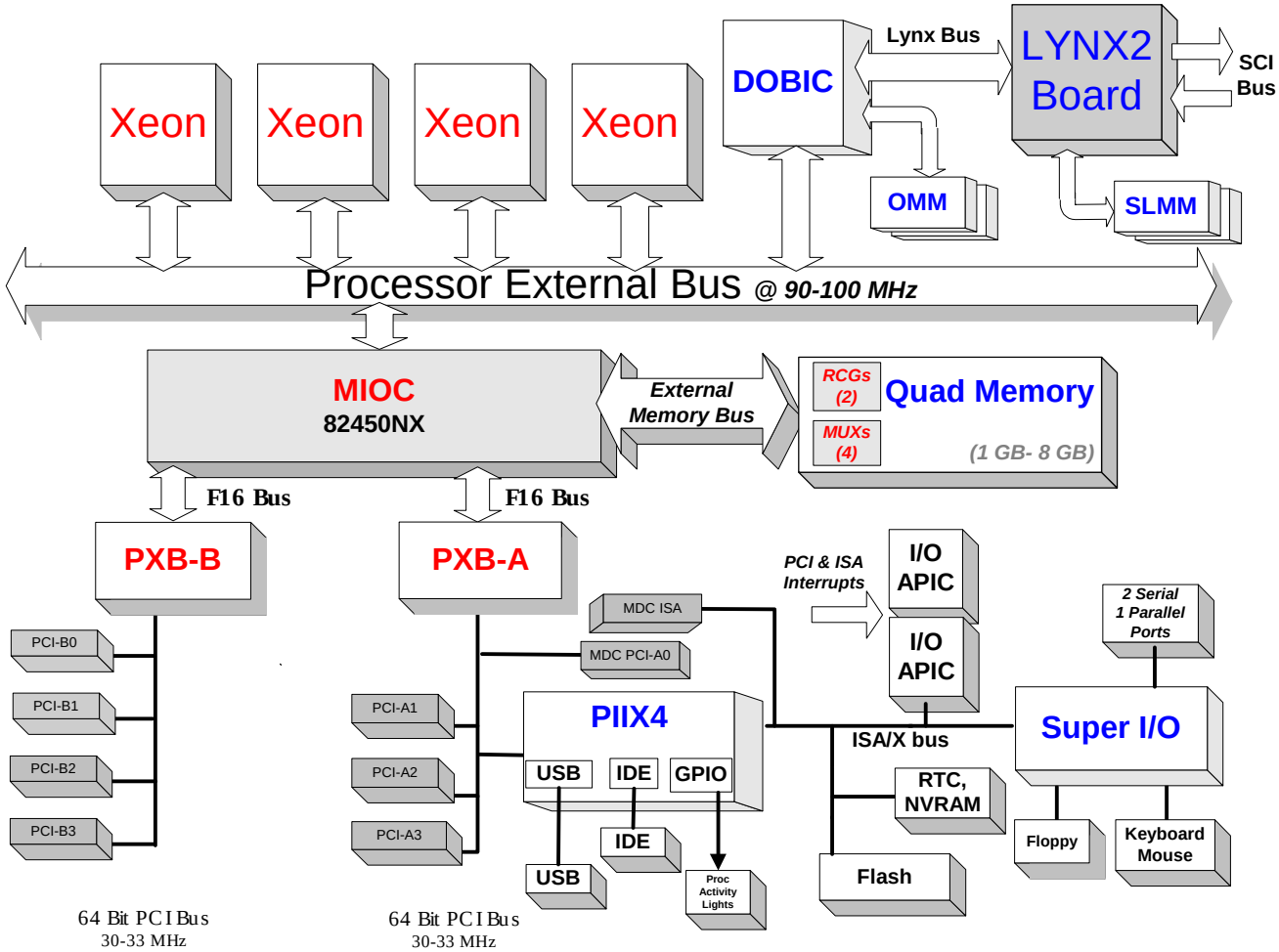
Overview

- TPC-H
  - System configuration
  - Data
- TPC-C
  - System configuration
  - Data
- Synthetic workload
- Conclusions

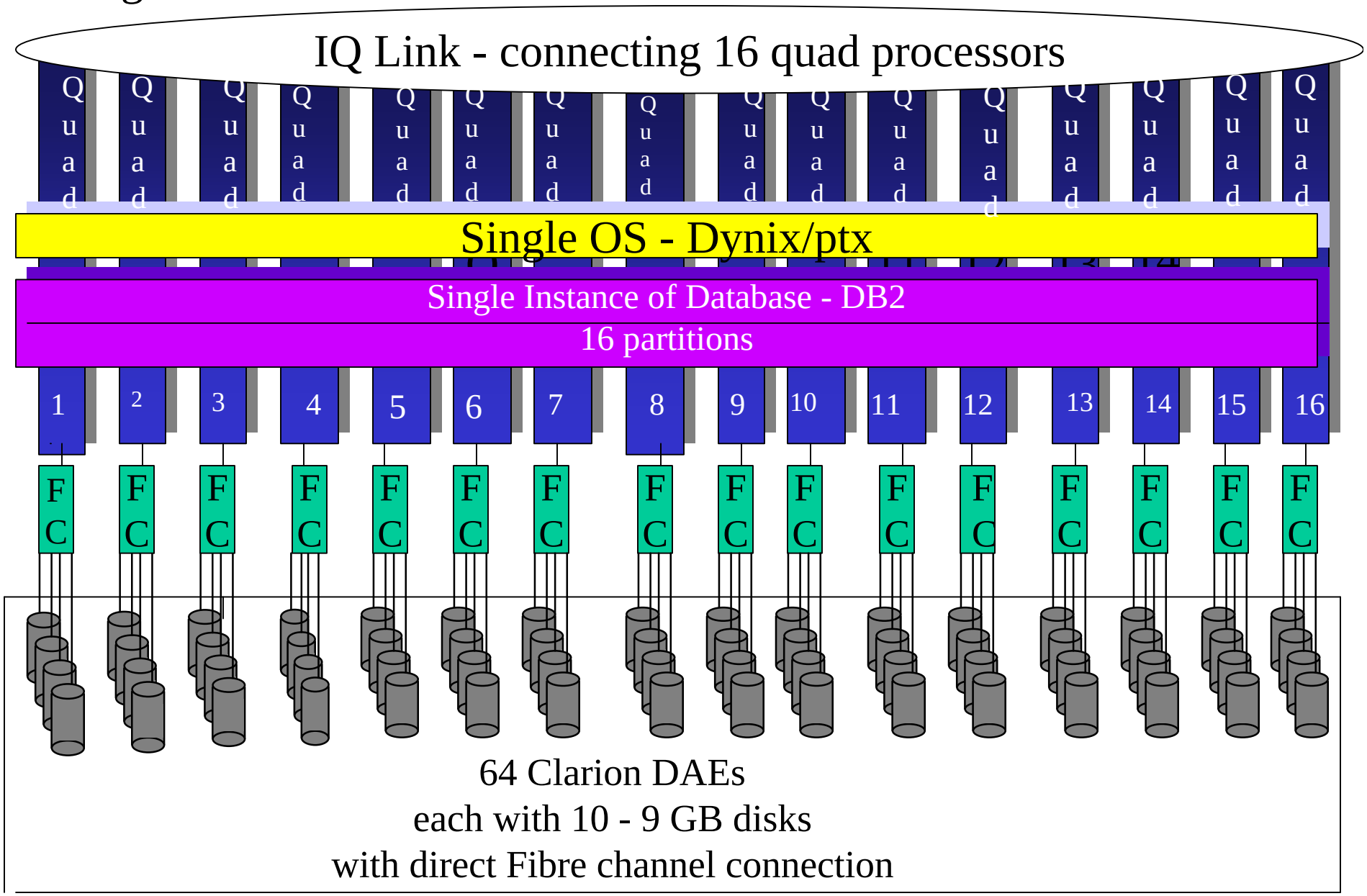
TPC-H System Configuration

- 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



Configuration Details

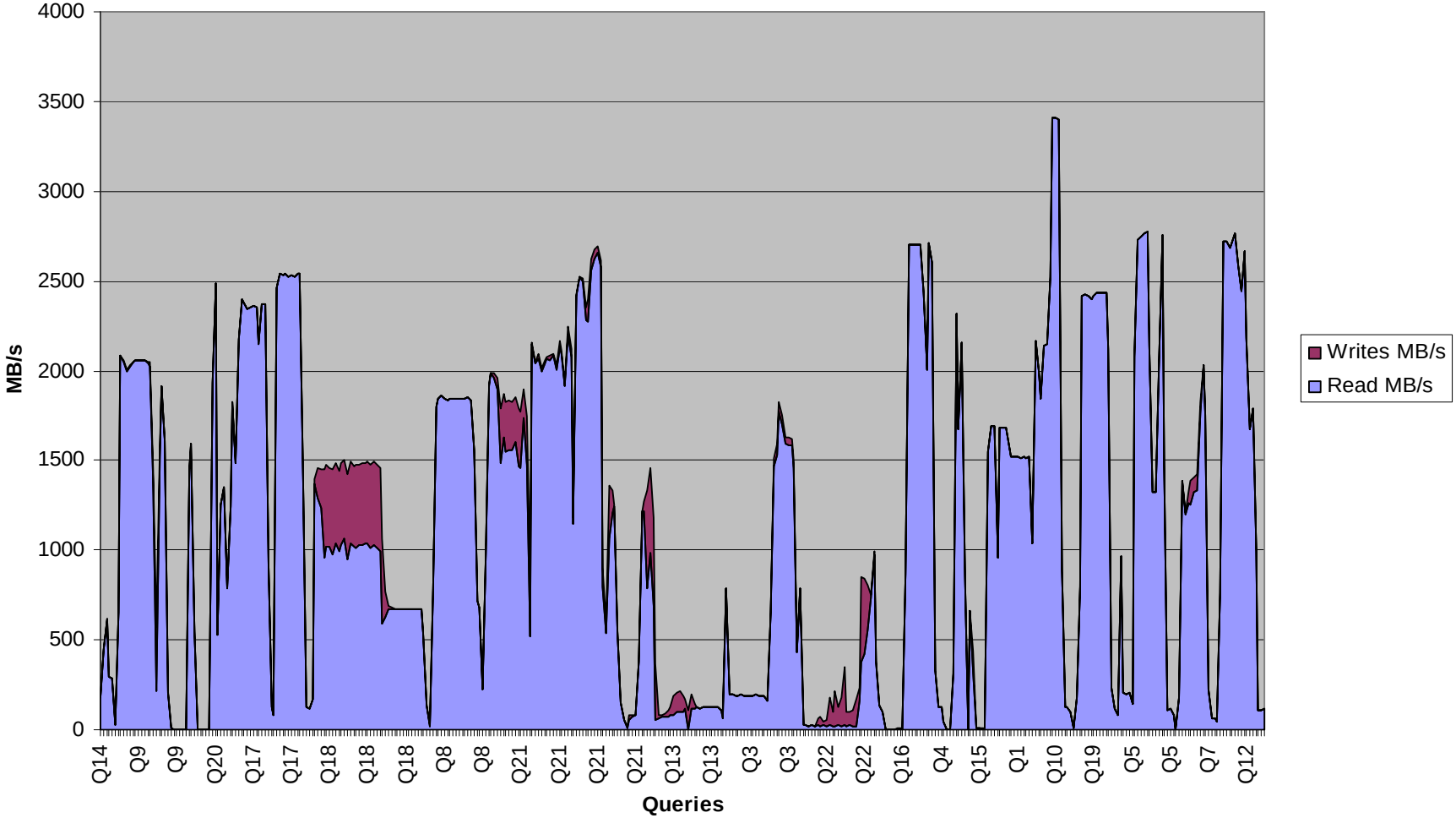


TPC-H I/O Characteristics

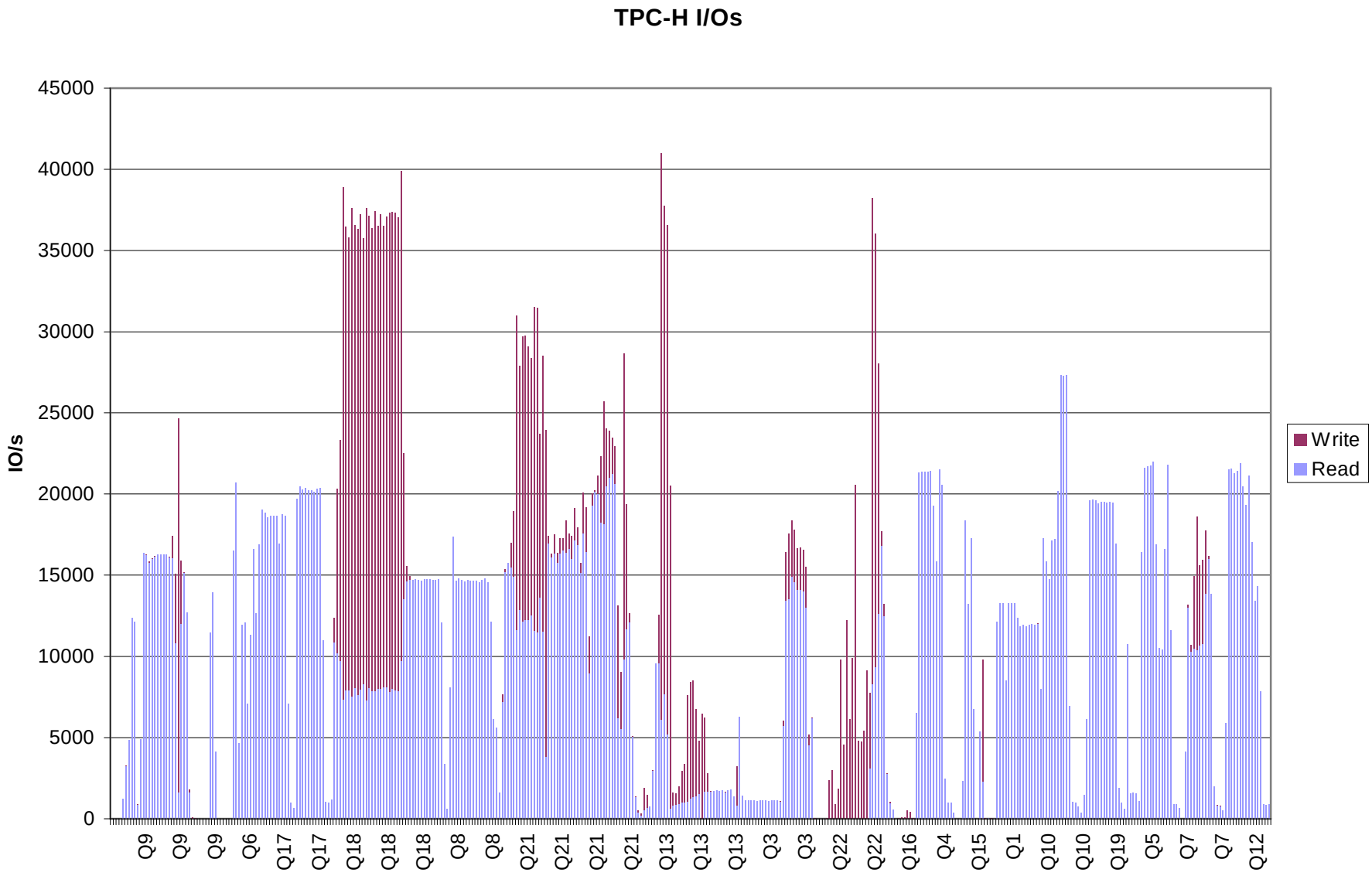
- Mostly reads
- 128MB blocks
- Sequential I/O
- Partitionable

TPC-H I/O Throughput

NUMA-Q 200/DB2 300GB TPC-H Power Test
IO Throughput (MB/s)



TPC-H IOs


TPC-H I/Os
IO/s
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
Q9
Q9
Q9
Q6
Q17
Q17
Q18
Q18
Q18
Q8
Q8
Q21
Q21
Q21
Q21
Q13
Q13
Q13
Q3
Q3
Q22
Q22
Q16
Q4
Q15
Q1
Q10
Q10
Q19
Q5
Q7
Q7
Q12
Write
Read

Query 10

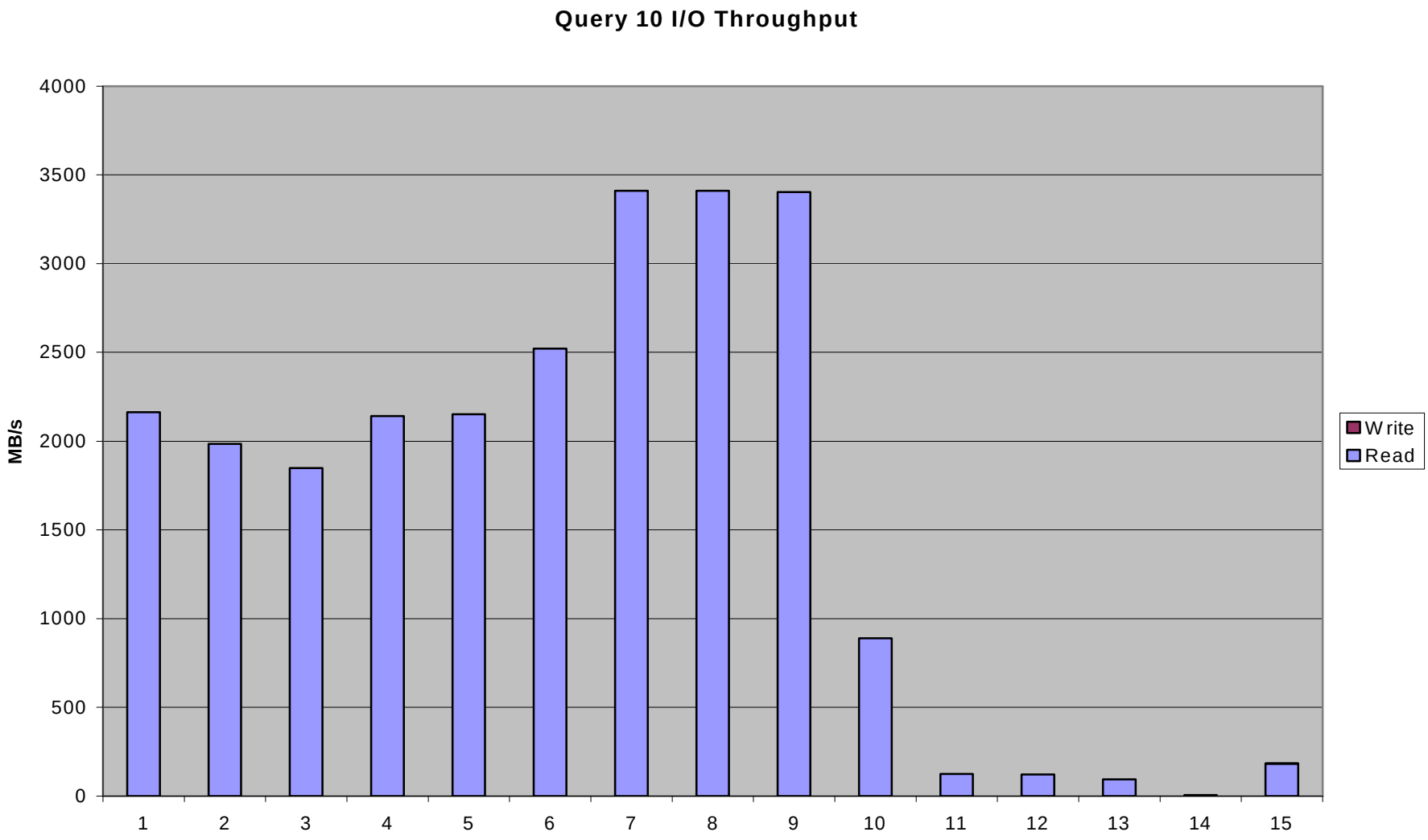
Query 10 &SHY; Returned Item Reporting

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 sum(l_extendedprice*(1-l_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.

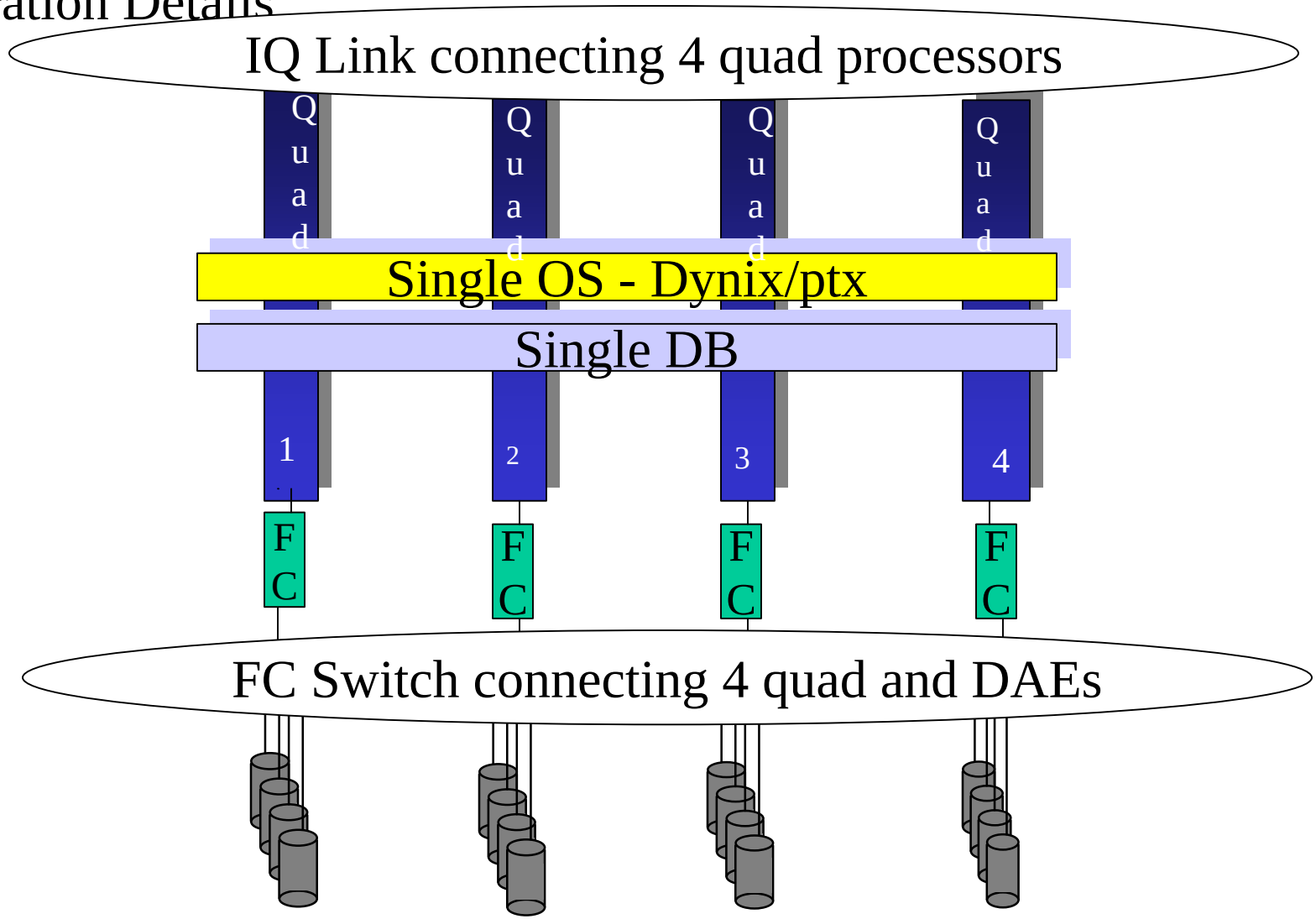
Query 10 Breakdown



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

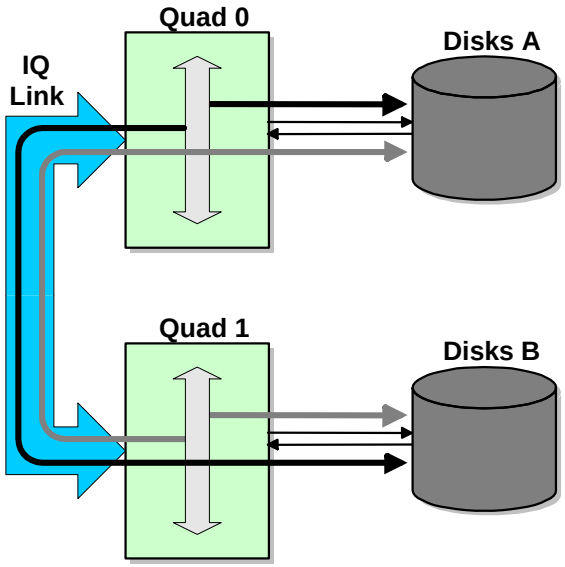
Configuration Details



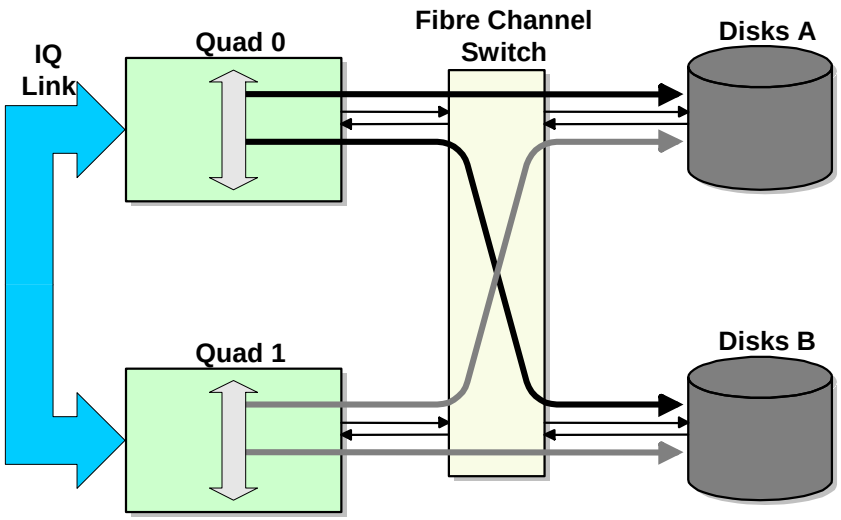
Clarion DAEs
with switched Fibre channel connection and multipath I/O

Multipath I/O

Point-to-Point



Multipath

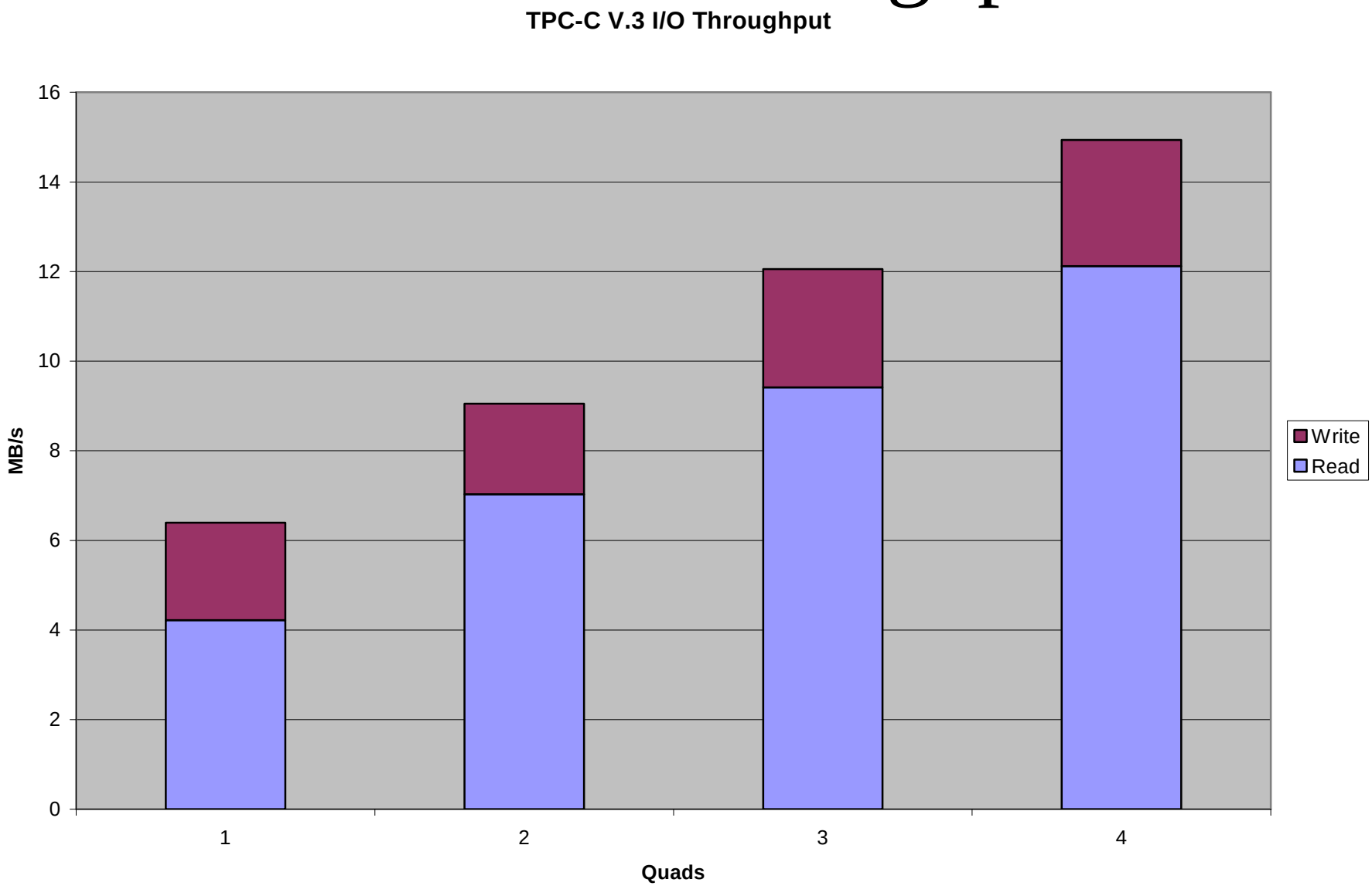


- With multipath I/O, the all DMA transfers are to/from local memory addresses.

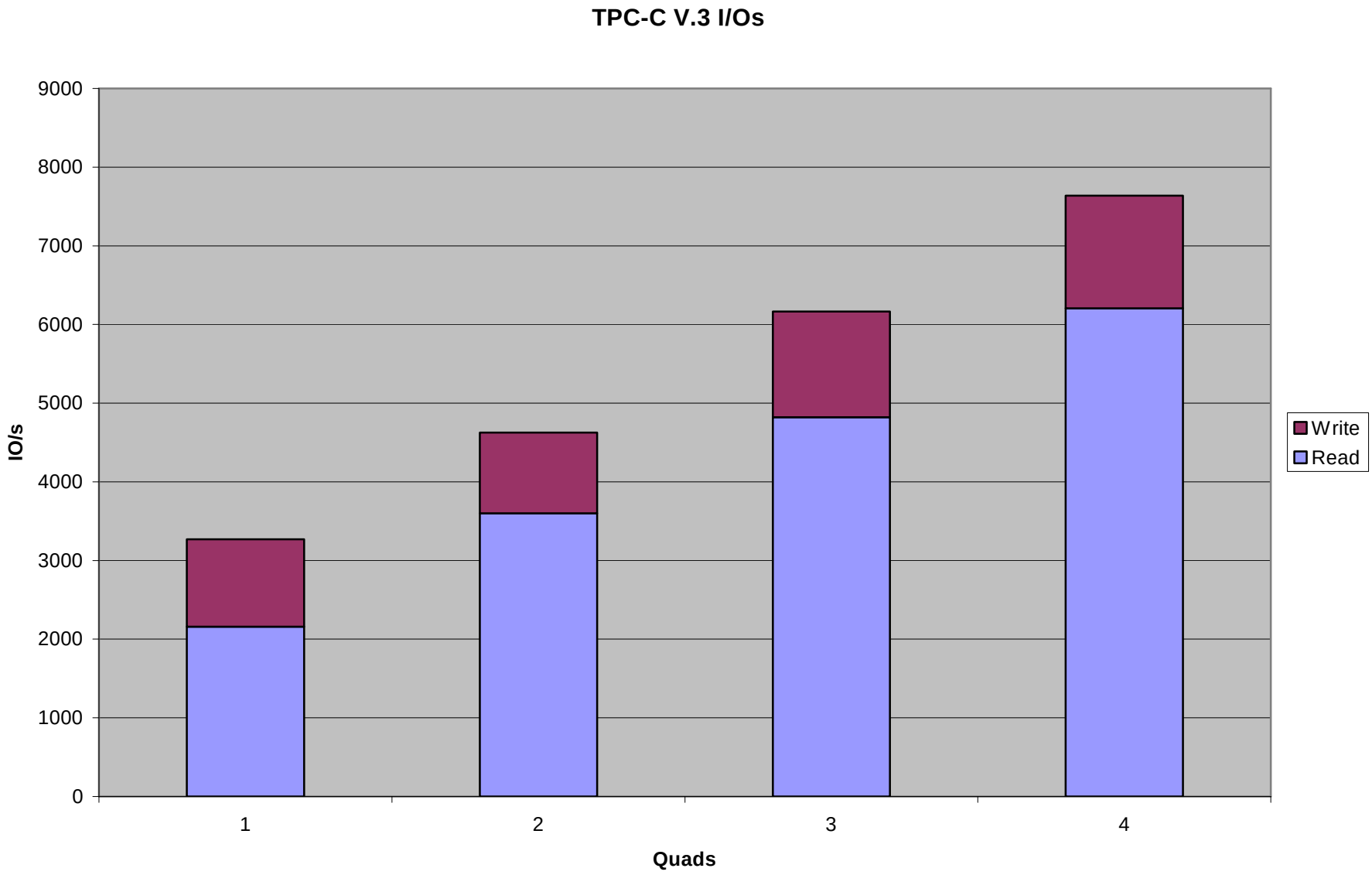
TPC-C I/O Characteristics

- Mostly reads
- 2K blocks
- Random I/O

TPC-C I/O Throughput



TPC-C I/Os


TPC-C V.3 I/Os
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
IO/s
1
2
3
4
Quads
Write
Read

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.

	IO/sec	**MB/sec**	**%Kernel**
128K Sequential Reads	30,080	3760	14
2K Random Reads	88,345	176	13
2K 80% Reads, 20% Writes	83,651	167	12

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.