IBM CICS Transaction Server V4.2

A Comparison of CICS QR and OTE Performance

March 2012

IBM Hardware Acceleration Lab

Nicholas C. Matsakis
Wei K. Liu
Greg Dyck
Terry Borden

Overview

In this paper we compare the performance of a 3270-based workload running in CICS® QR (Quasi-Reentrant) and CICS OTE (Open Transaction Environment – threadsafe) modes. We used RTW, one of our traditional CICS®/DB2® workloads running on configurations with different numbers of dedicated CPs (8 to 32) on an IBM zEnterprise 196 (z196). We used RMF™ measurement data to compare the number of transactions per second (ETR), response time achieved, and total LPAR CPU busy percentage between QR and OTE modes. Other than the transaction mode, the number of CPs, and matching the appropriate number of AORs for the number of CPs, there were no other configuration differences between the measurement environments.

Result Highlights

This paper may be useful in understanding the potential scale improvement and or cost reductions that might be achieved by using the CICS Open Services in OTE mode.

The measurements showed the following points:

- The transaction rate (ETR) scaled nearly linearly as additional CPs were added in both CICS QR and OTE modes.
- CICS OTE mode achieved significantly higher transaction rates (ETR) than CICS QR at every measurement point.
- CICS OTE had better response time, achieving up to an 80% improvement in most of the measurements.

Comparison of CICS QR and CICS OTE modes

Prior to OTE, all application code ran under the main CICS® TCB called the Quasi-Reentrant (QR) TCB/Task. The CICS dispatcher sub-dispatched the use of the QR TCB between the CICS transactions. Each transaction voluntarily gives up control when it issues a CICS service. There is only one CICS transaction active at any time on the QR TCB.

OTE introduced a new class of TCB called an open TCB, which can be used by threadsafe applications. An open TCB is characterized by the fact it is assigned to a CICS transaction for the life of the transaction and multiple OTE TCBs may run concurrently in the CICS region. A threadsafe application is defined as a program which uses appropriate serialization techniques, such as compare and swap or enqueue, when accessing any shared resource(s). It must be capable of running concurrently on multiple TCBs and must not rely on quasi-reentrancy to serialize access to shared resources and storage.

There is no sub-dispatching of other CICS transactions under the open TCB. An application executing under an open TCB can issue non CICS API requests which may involve the TCB being blocked. Blocking is allowed because only this TCB is halted, and not the whole of CICS, which is what happens if a blocking request is issued under the QR TCB. Examples of non CICS APIs would be z/OS services such as GETMAIN and z/OS UNIX® System Services functions. In CICS TS 2.2 support was added to enable CICS/DB2 applications to run in an OTE. Previous to this they had to switch TCB when issuing DB2 requests. With OTE they benefited from

reduced TCB switching which improved their performance. Existing or new CICS DB2 applications written in any language which access DB2 now had the opportunity to gain the performance benefits provided by the OTE technology.

3270 CICS/DB2 workload (RTW)

RTW is a standard workload used by the CICS Hursley Performance team to assess changes in performance characteristics within new releases of CICS code when running DB2 applications. In these applications the presentation logic is separated from the business logic by an EXEC CICS LINK.

The workload has the following characteristics

- All COBOL programs
- 7 unique transactions
- 20 Database Tables
- Average of 200 DB2 calls per transaction
- 54% Select, 1% insert, 1% update, 1% delete, 8% open cursor, 27% fetch cursor, 8% close cursor
- Terminal and application processing is included in the same region rather than separate TORs and AORs. The CICS MVS High Performance Option (HPO) was enabled.

The front-end presentation logic is very simple. It receives data from the terminal, passes it to the back-end business logic, and sends a response to the terminal when the logic returns control. The workload was designed to be threadsafe. When the workload was run in CICS OTE mode there were no non-threadsafe transactions running.

Since the workload was used to assess changes in performance characteristics within new releases of CICS code it was necessary to make changes to the workload to perform scale measurements. In this comparison the workload has been scaled up (primarily by increasing the size of the database tables) so that we can run on a z196 LPAR with 32 CPs to compare the throughput performance of the workload running in traditional QR and in OTE modes.

Network simulation

In the 3270 version of the RTW the network is simulated by TPNS LU2 terminals. TPNS runs on a separate z/OS® system to avoid effecting the CPU usage of the system under test. The transaction rate is changed dynamically by altering the millisecond 'user think time' (UTI). For this evaluation we have used different "user think time" values (i.e., creating different transaction arrival rates) to achieve approximately 90% processor utilization in order to compare the transaction rates of the two modes.

Environment

- z196 with up to 32 dedicated CPs
- TPNS on separate system
- DS8800 DASD
- z/OS V1R13
- DB2 10 for z/OS
- CICS TS V4.2

Performance data collection

We used RMF to measure three important metrics for this comparison:

- Transaction rate (ETR) and response time reported by assigning the CICS APPLID in a unique WLM reporting group in the CICS subtype.
- The average LPAR CPU busy percentage reported in the CPU Activity Report.

Tuning and Configuration Used

- **CICS**
  - MAXOPENTCBS(8)
- **DB2CONN**:
  - THREADLIMIT(8) PRIORITY(EQUAL) TCBLIMIT(8) REUSELIMIT(10000)
- **DB2ENTRY**:
  - PROTECTNUM(8) THREADLIMIT(8)
- **DB2**
  - CTHREAD set to ensure that the threash hold was never reached.
  - Log datasets striped over 2 dedicated volumes on separate DS8800 control units
  - Buffer pools tuned with best practice guidelines

The QR and L8 TCBs were defined with an EQUAL dispatch priority. Some configurations showed slight benefit with HIGH priority for the L8 TCB, but overall an EQUAL dispatch priority provided the best results for the RTW workload.

Switching to OTE

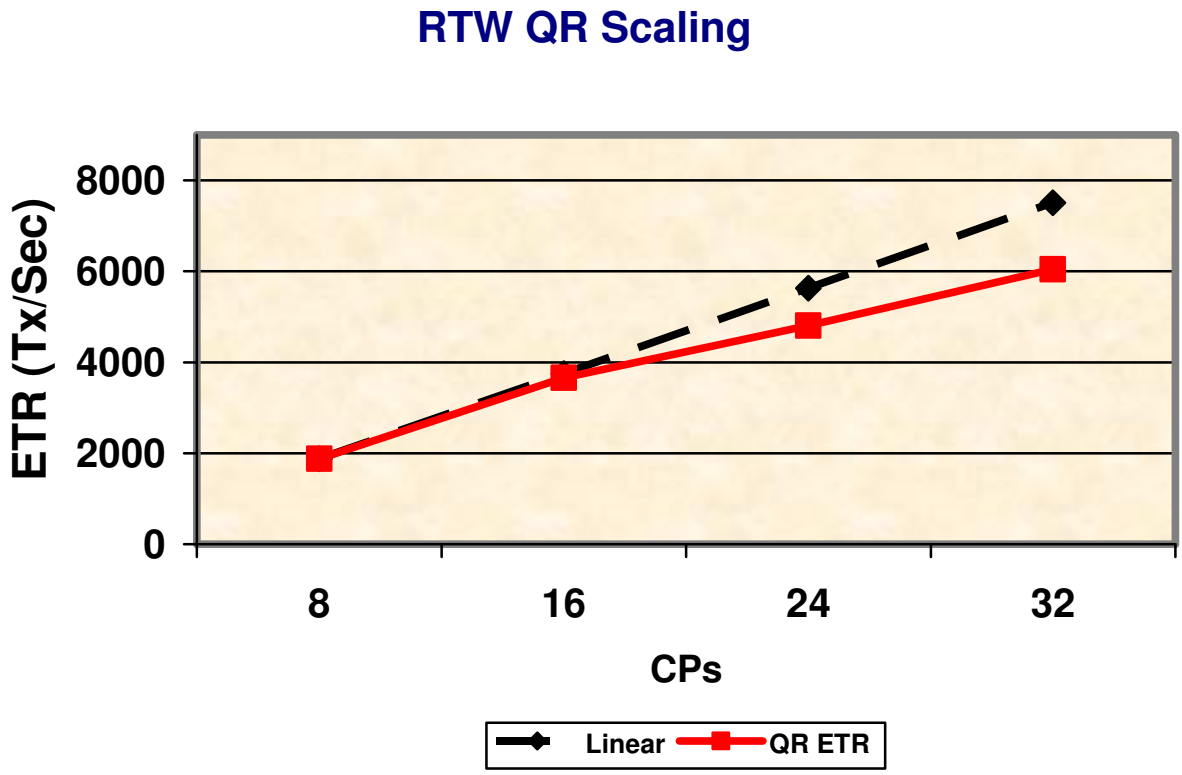
The workload was designed to be threadsafe. The only change necessary to run in OTE mode was to update the CICS program resource definitions.

Unlike our workload, it is possible to have a large mix of QR and threadsafe programs running in the same AOR.

Results

3270 based RTW Benchmark results for QR

The following graph shows the measurement results for the RTW workload running in CICS QR mode.

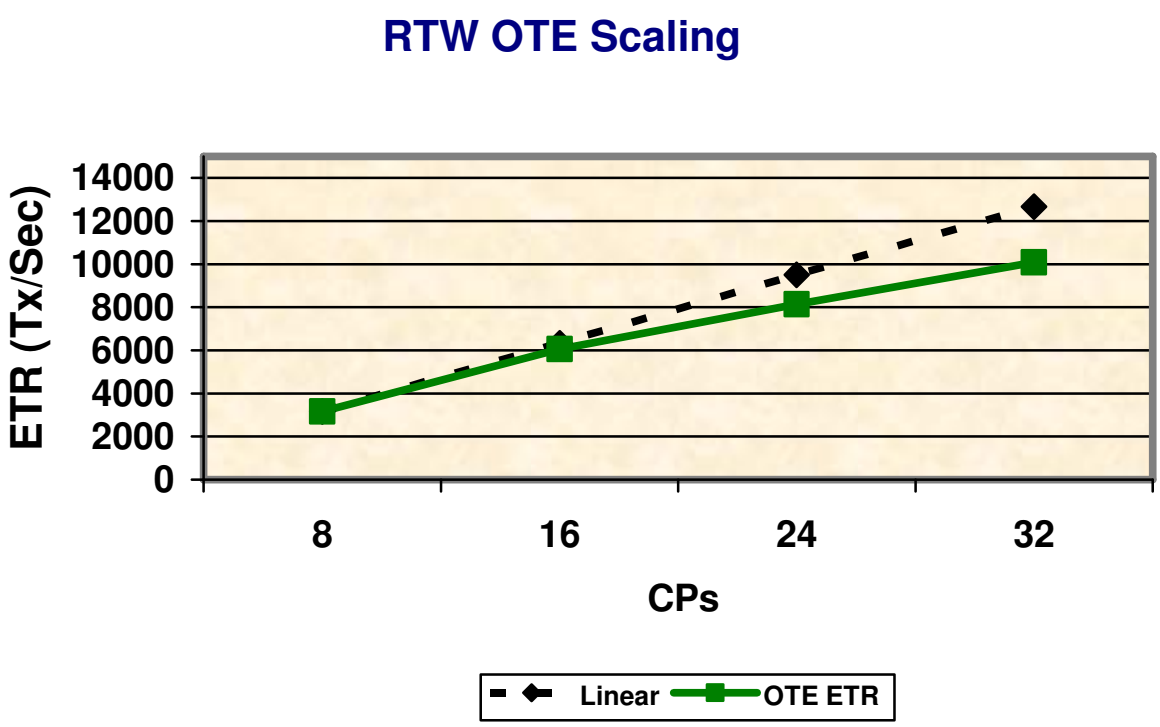


The following table shows the number of Regions, Terminals, and think time (UTI) used for each measurement point and the resulting ETR, Response time per transaction in milliseconds, and LPAR CPU busy percentage.

CPs	Regions	Terminals	UTI	ETR	Response	LPAR CPU %
8	4	8000	208	1877	105	95.7
16	8	16000	218	3660	14	90.3
24	12	24000	244	4807	87	87.5
32	16	32000	252	6046	255	95.7

3270 based RTW Benchmark results for Threadsafe OTE

The following graph shows the measurement results for the RTW workload running in CICS OTE mode.

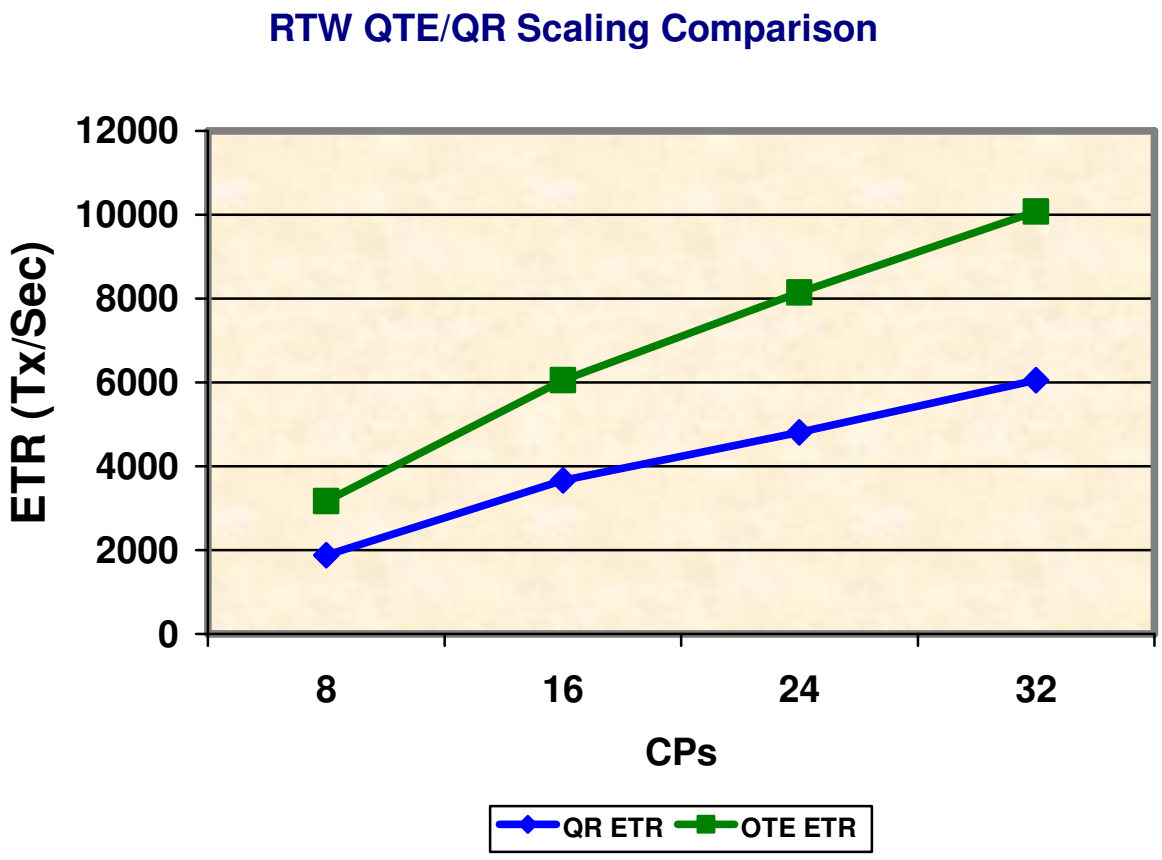


The following table shows the number of Regions, Terminals and think time used for each measurement point and the resulting ETR and Response time per transaction in milliseconds.

CPs	Regions	Terminals	UTI	ETR	Response	LPAR CPU%
8	4	8000	128	3169	7	91.9
16	8	16000	146	6053	6	90.6
24	12	24000	148	8145	9	91.7
32	16	32000	156	10079	43	92

Conclusion

The following graph compares the ETR achieved by each mode (QR and OTE) as the number of CPs increases from 4 to 32.



As shown in the above graph, the RTW workload running in OTE mode was able to achieve about **1.65** times the throughput (transactions per second) over QR mode. These results indicates that for this workload running it in CICS OTE mode offers advantages over running it in CICS QR mode.

1. Greater scale can be achieved on a given number of CPs
2. A desired transaction rate can be achieved with fewer CPs

It is clear that moving from QR to OTE can provide a considerable reduction in CPU time and substantial improvement in response time. This is largely because in QR most of the time has to be spent on the single QR TCB which results in less parallelism, worse hardware cache benefits, and substantial region TCB chatter between the QR TCB and the L8 or DB2 TCBs.

Given these savings and industry trends showing a declining growth rate in individual processor thread speeds which is being offset by increasing the number of concurrent processor threads it is recommended that consideration be given for migrating QR based applications to the Threadsafe OTE environment.

More information about what is involved in this migration can be found in the IBM book "Threadsafe Considerations for CICS".

Reference Material

- Threadsafe Considerations for CICS, SG24-6351
- IBM CICS TS 4.2 Application Programming Guide, SC34-7158
- IBM CICS TS 4.2 Application Programming Reference, SC34-7159