

z/TPF I/O Performance Study Results

Robert Blackburn, Allan Feldman,
Lee LaFrese and Leslie Sutton

Systems & Technology Group
March 27, 2008

Notices, Disclaimer and Trademarks

Copyright © 2008 by International Business Machines Corporation.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation. Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This information may include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or programs(s) at any time without notice. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT.

IBM shall have no responsibility to update this information. IBM products are warranted according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) Under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein. The performance data contained herein was obtained in a controlled, isolated environment. Actual results that may be obtained in other operating environments may vary significantly. While IBM has reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Statements regarding IBM's future direction and intent are subject to change or withdraw without notice, and represent goals and objectives only. The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

IBM, Enterprise Storage Server, ESCON, FICON, FlashCopy, System Storage, System p, z/OS, z9 and DS8000 are trademarks of International Business Machines Corporation in the United States, other countries, or both. Other company, products or service names may be trademarks or service marks of others.

z/TPF I/O Performance Study Results

Abstract

This paper documents the results of an I/O performance study that was done at IBM Poughkeepsie in 2007. The purpose of the study was to document the performance of z/TPF using an IBM z9 host processor along with IBM System Storage DS8300 disk storage. A cross brand team from IBM's TPF development, System z performance and Storage performance organizations collaborated on the project. Key findings of the study are included as well as discussion of how these results were achieved and some recommendations on how users may apply this information to their own z/TPF environments.

Overview

Transaction Processing Facility (TPF) software is IBM's leading software offering for industries that require high performance, high throughput, and high availability solutions. Most of the world's major airlines and Global Distribution Systems (GDS) are TPF customers, as well as many railroads, hotels, banks, and major credit card systems. z/TPF is the newest version of TPF that provides customers with a highly scalable solution for their online transaction processing needs. To achieve these results, z/TPF exploits the performance capabilities of the latest System z host processors and disk storage systems. The objective of this study was to evaluate the performance of an end-to-end IBM solution using z/TPF, z9 host and DS8300 hardware.

z/TPF's extremely efficient MP workload capabilities and I/O subsystem enable it to drive I/O rates well beyond any other platform. z/TPF's largest customers are able to achieve throughput exceeding 1.4M I/O operations per second (IO/s) and transaction rates well over 25K tps in production. In addition to throughput, z/TPF systems have extremely high availability and fast system restart characteristics and also have the ability to connect to very large networks.

The System z9 Enterprise Class processor (formerly z9-109) provides a combination of previous mainframe strengths with new functions designed around scalability, including flexible granularity solutions, virtualization, availability, and security. It can scale to offer a large capacity mainframe with up to 54 CPs and 512 GB of memory in a single footprint. The z9 EC, like earlier IBM mainframes, includes dedicated I/O processors (the System Assist Processor or SAP) to manage high rates of concurrent data accesses. It also provides more available channel bandwidth and FICON channels than previous zSeries® processors.

The DS8300 is IBM's industry leading enterprise disk storage offering. It provides front end cache hit throughput up to 4.9 million IO/s and end-to end sequential bandwidth up to 3.9 Gigabytes per second (GB/s). The DS8300 has a rich set of advanced replication functions such as Metro Mirror for synchronous remote copy, Global Mirror for asynchronous remote copy and FlashCopy for local point in time copies. Another important feature of the DS8300 is the ability to logically partition a system into two independent system facility images (SFIs) each with their own set of dedicated hardware resources and running their own licensed internal code. This dual SFI feature was enabled on the DS8300 used in this study.

Configuration

z/TPF

Source Level:

The latest available z/TPF system at PUT 3 was used for the test. In addition, a key PUT4 APAR (PJ32163) to correct CPU loop task dispatching and SWISC load balancing was loaded to the system.

z/TPF I/O Performance Study Results

Storage & VFA:

The z9 processor used was configured with 18GB of memory. Of this memory, 6GB was allocated to the system and 12GB was allocated to Virtual File Access (VFA) storage. VFA is a TPF unique read cache for I/O buffers. By adjusting the total memory size, and thus the amount of VFA, the test achieved a reads per write ratio of 1.13.

Tools

Several z/TPF tools were used to monitor and assess system performance:

- **Data Collection/Data Reduction (DC/DR):** The purpose of this package (DC/DR) is to observe by judicious sampling the operation of the TPF system environment and report on its performance parameters and their statistical characteristics. The package is comprised of an online component referred to as data collection (DC) and an offline (z/OS) reporting component called data reduction (DR).
- **Continuous Data Collection (CDC):** CDC is an application that collects real-time z/TPF system performance information. The performance data is processed, formatted, and displayed by a component of CDC called the CDC client.
- **Software profiler:** The z/TPF software profiler is a performance measurement tool that collects and analyzes data for external interrupt (EI) analysis, ECB entry analysis (EA), macro analysis (MA), page range (PR) analysis, and resource consumption (RC).

Host Processor

Hardware: z9 processor, model 2094-S18, 2 books, 18-way MP

- 128 GB memory
- 32 FICON Express2 2GB channels
- 4 SAPs
- measured at 1-16 I-Streams

Disk Storage

Hardware: DS8300 Turbo, 2107- 9B2 dual frame, dual SFI

- 64 GB total cache memory (32 GB per SFI)
- 2 GB write cache total (1 GB per SFI)
- 360 active disk drives (146 GB 15K RPM) and 24 hot spares
- 48 RAID-10 arrays (12 configured as 3+3 and 36 configured as 4+4)
- 16 x 2 Gb Longwave host adapters, two ports per adapter used for host connectivity
- 2880 TPF volumes (3390-9) = 26 TeraBytes usable storage
 - 720 volumes used for performance testing, remainder for copies

Volume Layout:

2880 volumes of 3390-9 size (10017 cylinders) were created, which consumed all of the available capacity of the DS8300. The total capacity of these volumes was 24.5 Terabytes (TB). Of these volumes, 720 were attached to a VM test partition for debug and workload development (TEST volumes). Another 720 volumes were used for the performance testing (PERF volumes). The remaining volumes were reserved for copy services testing or spares. The PERF volumes were divided into 36 Logical Subsystems (LSS) each containing 20 volumes (10 primes and 10 duplicates). Three LSS were spread across groups of four RAID-10 arrays (also called ranks). Each group of four RAID-10 arrays contained one 3+3 array and three 4+4 arrays. This provided for excellent workload balance across the disk drives. The volume layout across 8 of the 48 RAID-10 arrays is shown in Figure 1 below. This basic configuration was repeated for the 40 other arrays and 30 other LSS.

z/TPF I/O Performance Study Results

Server 0

Rank 0	Sets	LSS0 (P,D)	LSS2 (P,D)	LSS4 (P,D)
R-10 3+3	TEST	4 (2,2)	4 (2,2)	4 (2,2)
48 x 3390-9	PERF	4 (2,2)	4 (2,2)	4 (2,2)
	FC TGT	2 (2,0)	2 (2,0)	2 (2,0)
	PPRC SEC	4 (2,2)	4 (2,2)	4 (2,2)
	SPARE	2	2	2
	TOTAL	16	16	16
		48		

Server 1

Rank 1	Sets	LSS1 (P,D)	LSS3 (P,D)	LSS5 (P,D)
R-10 3+3	TEST	4 (2,2)	4 (2,2)	4 (2,2)
48 x 3390-9	PERF	4 (2,2)	4 (2,2)	4 (2,2)
	FC TGT	2 (2,0)	2 (2,0)	2 (2,0)
	PPRC SEC	4 (2,2)	4 (2,2)	4 (2,2)
	SPARE	2	2	2
	TOTAL	16	16	16
		48		

Rank 2	Sets	LSS0 (P,D)	LSS2 (P,D)	LSS4 (P,D)
R-10 4+4	TEST	6 (3,3)	5 (3,2)	5 (2,3)
64 x 3390-9	PERF	5 (2,3)	6 (3,3)	5 (3,2)
	FC TGT	3 (3,0)	3 (3,0)	2 (2,0)
	PPRC SEC	5 (3,2)	5 (2,3)	6 (3,3)
	SPARE	2	3	3
	TOTAL	21	22	21
		64		

Rank 3	Sets	LSS1 (P,D)	LSS3 (P,D)	LSS5 (P,D)
R-10 4+4	TEST	6 (3,3)	5 (3,2)	5 (2,3)
64 x 3390-9	PERF	5 (2,3)	6 (3,3)	5 (3,2)
	FC TGT	3 (3,0)	3 (3,0)	2 (2,0)
	PPRC SEC	5 (3,2)	5 (2,3)	6 (3,3)
	SPARE	2	3	3
	TOTAL	21	22	21
		64		

Rank 4	Sets	LSS0 (P,D)	LSS2 (P,D)	LSS4 (P,D)
R-10 4+4	TEST	5 (2,3)	6 (3,3)	5 (3,2)
64 x 3390-9	PERF	5 (3,2)	5 (2,3)	6 (3,3)
	FC TGT	2 (2,0)	3 (3,0)	3 (3,0)
	PPRC SEC	6 (3,3)	5 (3,2)	5 (2,3)
	SPARE	3	2	3
	TOTAL	21	21	22
		64		

Rank 5	Sets	LSS1 (P,D)	LSS3 (P,D)	LSS5 (P,D)
R-10 4+4	TEST	5 (2,3)	6 (3,3)	5 (3,2)
64 x 3390-9	PERF	5 (3,2)	5 (2,3)	6 (3,3)
	FC TGT	2 (2,0)	3 (3,0)	3 (3,0)
	PPRC SEC	6 (3,3)	5 (3,2)	5 (2,3)
	SPARE	3	2	3
	TOTAL	21	21	22
		64		

Rank 6	Sets	LSS0 (P,D)	LSS2 (P,D)	LSS4 (P,D)
R-10 4+4	TEST	5 (3,2)	5 (2,3)	6 (3,3)
64 x 3390-9	PERF	6 (3,3)	5 (3,2)	5 (2,3)
	FC TGT	3 (3,0)	2 (2,0)	3 (3,0)
	PPRC SEC	5 (2,3)	6 (3,3)	5 (3,2)
	SPARE	3	3	2
	TOTAL	22	21	21
		64		

Rank 7	Sets	LSS1 (P,D)	LSS3 (P,D)	LSS5 (P,D)
R-10 4+4	TEST	5 (3,2)	5 (2,3)	6 (3,3)
64 x 3390-9	PERF	6 (3,3)	5 (3,2)	5 (2,3)
	FC TGT	3 (3,0)	2 (2,0)	3 (3,0)
	PPRC SEC	5 (2,3)	6 (3,3)	5 (3,2)
	SPARE	3	3	2
	TOTAL	22	21	21
		64		

Figure 1 - Volume Layout on Eight RAID-10 Ranks

The DS8300 used in the testing was a dual SFI 2107-9B2. Each SFI contained half of the configured volumes. Additionally, each SFI consists of two server clusters. In order to build a resilient configuration, the volumes were laid out so that each LSS would have prime and duplicate volumes. However, each prime volume would have its duplicate on the opposite server cluster of the other SFI. Thus the configuration would survive the loss of two server clusters at the same time even if they were both on the same SFI. Figure 2 below shows this concept graphically.

z/TPF I/O Performance Study Results

<p>SFI 0, Server 0</p> <table style="width: 100%; border-collapse: collapse;"> <tr><td>Prime 001</td><td>Prime 061</td><td>Prime 121</td></tr> <tr><td>Dup 181</td><td>Dup 241</td><td>Dup 301</td></tr> <tr><td>Prime 003</td><td>Prime 063</td><td>Prime 123</td></tr> <tr><td>Dup 183</td><td>Dup 243</td><td>Dup 303</td></tr> <tr><td>... etc.</td><td>--- etc.</td><td>--- etc.</td></tr> <tr><td>Prime 057</td><td>Prime 117</td><td>Prime 177</td></tr> <tr><td>Dup 237</td><td>Dup 297</td><td>Dup 357</td></tr> <tr><td>Prime 059</td><td>Prime 119</td><td>Prime 179</td></tr> <tr><td>Dup 239</td><td>Dup 299</td><td>Dup 359</td></tr> </table>	Prime 001	Prime 061	Prime 121	Dup 181	Dup 241	Dup 301	Prime 003	Prime 063	Prime 123	Dup 183	Dup 243	Dup 303	... etc.	--- etc.	--- etc.	Prime 057	Prime 117	Prime 177	Dup 237	Dup 297	Dup 357	Prime 059	Prime 119	Prime 179	Dup 239	Dup 299	Dup 359	<p>SFI 0, Server 1</p> <table style="width: 100%; border-collapse: collapse;"> <tr><td>Prime 002</td><td>Prime 062</td><td>Prime 302</td></tr> <tr><td>Dup 182</td><td>Dup 242</td><td>Dup 122</td></tr> <tr><td>Prime 004</td><td>Prime 064</td><td>Prime 304</td></tr> <tr><td>Dup 184</td><td>Dup 064</td><td>Dup 124</td></tr> <tr><td>... etc.</td><td>--- etc.</td><td>--- etc.</td></tr> <tr><td>Prime 058</td><td>Prime 298</td><td>Prime 358</td></tr> <tr><td>Dup 238</td><td>Dup 118</td><td>Dup 178</td></tr> <tr><td>Prime 060</td><td>Prime 300</td><td>Prime 360</td></tr> <tr><td>Dup 240</td><td>Dup 120</td><td>Dup 180</td></tr> </table>	Prime 002	Prime 062	Prime 302	Dup 182	Dup 242	Dup 122	Prime 004	Prime 064	Prime 304	Dup 184	Dup 064	Dup 124	... etc.	--- etc.	--- etc.	Prime 058	Prime 298	Prime 358	Dup 238	Dup 118	Dup 178	Prime 060	Prime 300	Prime 360	Dup 240	Dup 120	Dup 180
Prime 001	Prime 061	Prime 121																																																					
Dup 181	Dup 241	Dup 301																																																					
Prime 003	Prime 063	Prime 123																																																					
Dup 183	Dup 243	Dup 303																																																					
... etc.	--- etc.	--- etc.																																																					
Prime 057	Prime 117	Prime 177																																																					
Dup 237	Dup 297	Dup 357																																																					
Prime 059	Prime 119	Prime 179																																																					
Dup 239	Dup 299	Dup 359																																																					
Prime 002	Prime 062	Prime 302																																																					
Dup 182	Dup 242	Dup 122																																																					
Prime 004	Prime 064	Prime 304																																																					
Dup 184	Dup 064	Dup 124																																																					
... etc.	--- etc.	--- etc.																																																					
Prime 058	Prime 298	Prime 358																																																					
Dup 238	Dup 118	Dup 178																																																					
Prime 060	Prime 300	Prime 360																																																					
Dup 240	Dup 120	Dup 180																																																					
<p>SFI 1, Server 0</p> <table style="width: 100%; border-collapse: collapse;"> <tr><td>Prime 182</td><td>Prime 242</td><td>Prime 302</td></tr> <tr><td>Dup 002</td><td>Dup 062</td><td>Dup 122</td></tr> <tr><td>Prime 184</td><td>Prime 244</td><td>Prime 304</td></tr> <tr><td>Dup 004</td><td>Dup 064</td><td>Dup 124</td></tr> <tr><td>... etc.</td><td>--- etc.</td><td>--- etc.</td></tr> <tr><td>Prime 238</td><td>Prime 298</td><td>Prime 358</td></tr> <tr><td>Dup 058</td><td>Dup 118</td><td>Dup 178</td></tr> <tr><td>Prime 240</td><td>Prime 300</td><td>Prime 360</td></tr> <tr><td>Dup 060</td><td>Dup 120</td><td>Dup 180</td></tr> </table>	Prime 182	Prime 242	Prime 302	Dup 002	Dup 062	Dup 122	Prime 184	Prime 244	Prime 304	Dup 004	Dup 064	Dup 124	... etc.	--- etc.	--- etc.	Prime 238	Prime 298	Prime 358	Dup 058	Dup 118	Dup 178	Prime 240	Prime 300	Prime 360	Dup 060	Dup 120	Dup 180	<p>SFI 1, Server 1</p> <table style="width: 100%; border-collapse: collapse;"> <tr><td>Prime 181</td><td>Prime 241</td><td>Prime 301</td></tr> <tr><td>Dup 001</td><td>Dup 061</td><td>Dup 121</td></tr> <tr><td>Prime 183</td><td>Prime 243</td><td>Prime 303</td></tr> <tr><td>Dup 003</td><td>Dup 063</td><td>Dup 123</td></tr> <tr><td>... etc.</td><td>--- etc.</td><td>--- etc.</td></tr> <tr><td>Prime 237</td><td>Prime 297</td><td>Prime 357</td></tr> <tr><td>Dup 057</td><td>Dup 117</td><td>Dup 177</td></tr> <tr><td>Prime 239</td><td>Prime 299</td><td>Prime 359</td></tr> <tr><td>Dup 059</td><td>Dup 119</td><td>Dup 179</td></tr> </table>	Prime 181	Prime 241	Prime 301	Dup 001	Dup 061	Dup 121	Prime 183	Prime 243	Prime 303	Dup 003	Dup 063	Dup 123	... etc.	--- etc.	--- etc.	Prime 237	Prime 297	Prime 357	Dup 057	Dup 117	Dup 177	Prime 239	Prime 299	Prime 359	Dup 059	Dup 119	Dup 179
Prime 182	Prime 242	Prime 302																																																					
Dup 002	Dup 062	Dup 122																																																					
Prime 184	Prime 244	Prime 304																																																					
Dup 004	Dup 064	Dup 124																																																					
... etc.	--- etc.	--- etc.																																																					
Prime 238	Prime 298	Prime 358																																																					
Dup 058	Dup 118	Dup 178																																																					
Prime 240	Prime 300	Prime 360																																																					
Dup 060	Dup 120	Dup 180																																																					
Prime 181	Prime 241	Prime 301																																																					
Dup 001	Dup 061	Dup 121																																																					
Prime 183	Prime 243	Prime 303																																																					
Dup 003	Dup 063	Dup 123																																																					
... etc.	--- etc.	--- etc.																																																					
Prime 237	Prime 297	Prime 357																																																					
Dup 057	Dup 117	Dup 177																																																					
Prime 239	Prime 299	Prime 359																																																					
Dup 059	Dup 119	Dup 179																																																					

Figure 2 - Layout of Prime and Duplicate Volumes

Workload Development

Based on experience with data from many customers, the following I/O performance characteristics are considered typical of production TPF systems. They represent goals that governed the setup and configuration of the system:

- Reads per Writes = between 1 and 3 with an optimal goal of 2.
- Read Hit ratio = between 85% and 97% with an optimal goal of 95%.
- Destage Rate = between 8% and 20% with an optimal goal of 10%.

The first step in the study was to develop a large scale TPF workload meeting the customer profile goals outlined above. This workload would serve as the I/O driver for performance testing. Although smaller scale workloads were available, they were not appropriate for performance testing with the highly capable DS8300. The approach used was to scale up a TPF in-house lab test driver called AIR1 for the performance testing. AIR1 is a stand alone driver that simulates message traffic and has been used by the TPF lab for many years to measure system performance. The driver executes an instruction mix that invokes TPF system services emulating traditional TPF customer environments. A significant portion of the services invoked result in I/O intensive activity. These were obviously of most interest for testing I/O.

The AIR1 driver uses a z/TPF database that consists of 4K duplicated pool records (4DP). The 4DP records were divided into two different record IDs – TSIM4 for reads and TSIM5 for writes - so that the amount of read and write I/O's executed could be carefully managed and correctly mapped to the customer profile described above. The tuned workload had a read to write ratio of 1.13 which is roughly 56.5% reads. Using knowledge of the record access distribution function combined with empirical methods we were able to project and verify the cache hit ratio of .95 and the destage rate at .10.

z/TPF I/O Performance Study Results

The z/TPF system was defined with 3 device types – A, B, and C. Of the 720 fully duplicated modules, only 20 were defined as Device A, 64 were Device B, and the bulk of the database, 636 modules, were defined as Device C.

TSIM4 records

TSIM 4 records are used by the AIR1 driver for reading data. 15 Million 4K records were defined but only 5 Million total records were used in order to achieve a read cache hit ratio close to 95%. The data was spread over three device types as follows to ensure a smooth distribution of I/O activity among all the devices:

- Device A – 145,833 records used out of total 416,667 defined
- Device B – 466,667 records used out of total 1,333,333 defined
- Device C – 4,387,500 records used out of total 13,250,000 defined

TSIM5 records

TSIM 5 records are used by the AIR1 driver for writes. 2 Million 4K records were defined but only 50,000 records were used to achieve a 10% de-stage rate. The data was spread over three device types as follows to ensure a smooth distribution of I/O activity among all the devices:

- Device A – 1,389 records used out of a total 55,556 defined
- Device B – 4,444 records used out of a total 177,778 defined
- Device C – 44,167 records used out of a total 1,766,666 defined

Results and Discussion

Baseline Performance Testing

After developing a representative customer workload, the next step was to complete some baseline tests designed to measure the I/O performance. The workload was run at increasing I/O loads until further increases could not be sustained. The average disk response time and service time were measured using TPF data collection. Figure 3 shows a plot of the I/O rate vs. response time and service time – a typical “knee shaped curve” commonly seen in performance studies. The maximum throughput measured was approximately 163K IO/s at just over 2 milliseconds (ms) response time. This is an impressive level of throughput and service but was easily achieved via read and write caching. All of the writes were buffered by non-volatile storage so they proceeded at electronic speeds. With a 95% read hit ratio and 56.5% reads, less than 3% of I/O operations experienced any disk latency.

z/TPF I/O Performance Study Results

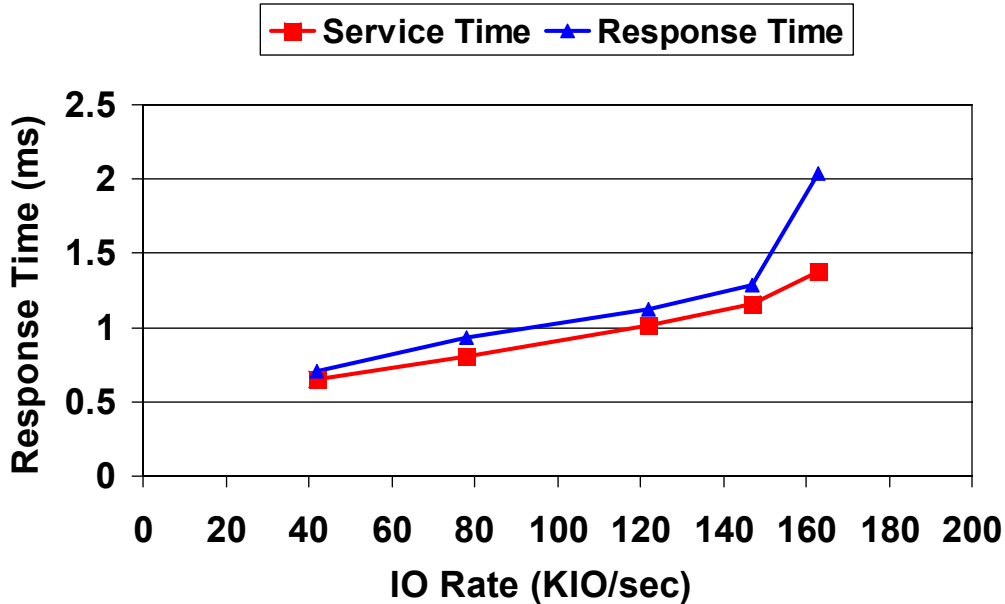


Figure 3 - I/O Service and Response Time Curve

It is fairly straightforward to estimate the number of disk operations per second that the workload generated. As mentioned earlier, about 3% of the 163K IO/s were read misses. This works out to 4,890 read misses per second spread across 360 disk drives. Thus read misses only account for about 13.5 IO/s/disk. Although the writes were buffered, they eventually must be destaged to disk. Since the destage rate was 10%, about 16,300 destages per second went to disk. The disks were configured with RAID-10 so each destage is mirrored to two physical disk drives. Thus 32,600 disk operations per second were generated by destaging or about 95.5 IO/sec/disk. Adding in the read misses gives us a total of 109 disk operations/s/disk. An enterprise class 15K RPM disk drive can sustain up to approximately 240 disk ops/s. Although the disk drives were busy, they were running at less than 50% of maximum utilization. For this workload profile, the throughput limit must have been due to a different bottleneck.

The DS8000 is powered by two IBM p570 servers running the AIX operating system. Each server in the DS8300 used for the testing was a 4-way Symmetric Multiprocessor (SMP) using Simultaneous Multi-Threading (SMT). SMT allows a 4-way SMP to perform comparably to an 8-way without SMT. An IBM internal tool called command history was used to estimate the processor complex utilization while running the TPF I/O driver at various throughput levels. Figure 4 below shows the results of the command history measurements. Note that there are some AIX operations that command history can't capture. In practical use, a command history utilization of 85% corresponds to a fully utilized SMP complex. At an operating level of 163K IO/s, the command history result was 73%. This indicates that the processor utilization was approaching its limit. It is likely that some additional throughput above 163K IO/s could be driven but the granularity of the workload driver was such that an absolute limit could not be found precisely. Further tuning of the I/O driver would likely have resulted in a maximum throughput in the range of 180 – 190K IO/s for this hardware configuration.

z/TPF I/O Performance Study Results

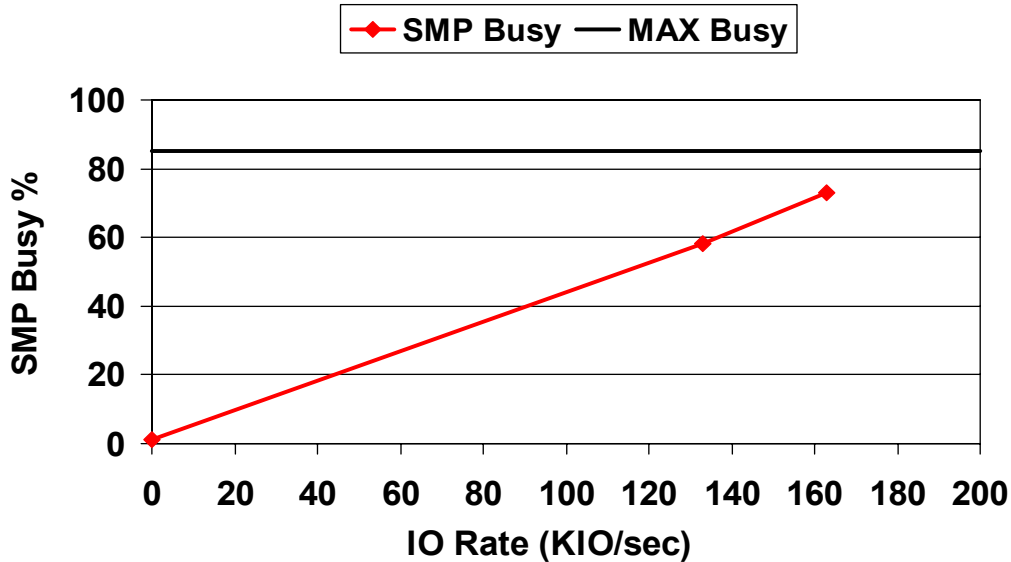


Figure 4 - SMP Busy Estimates from Command History

Adaptive Record Caching

TPF data records typically show good temporal locality of reference but poor spatial locality. What this means is that when a particular record is read by TPF, there is a fairly good chance that the same record will be read again within a short time (temporal locality). However, it is not very likely that other records on the same disk track will be referenced any time soon (spatial locality). This is because the TPF database is scattered uniformly across the disk volumes. For this reason, it is usually most efficient to manage TPF disk caches on a record basis.

Other operating systems such as z/OS typically have data access patterns that display both temporal and spatial locality of reference. A technique called track caching is used to take advantage of this. When a record is read from a disk on a system using track caching, the entire track (or in some cases a portion of the track) is staged from the disk and stored in cache memory. Thus subsequent reads to nearby records on the same track may be cache hits. For TPF, track caching is usually not as efficient as record caching.

The DS8000 features an adaptive record caching algorithm which dynamically adjusts between track and record modes at a fine granularity based on past experience. This algorithm helps the DS8000 cache perform efficiently across a wide range of applications.

A test was done to demonstrate the performance value of adaptive record caching compared to traditional track caching. A TPF test was run with adaptive record caching disabled and the results were compared to a similar run with adaptive record caching enabled. Figure 5 below shows the read response time of the physical hard disk drive (HDD) for the two runs. Adaptive record caching reduced the HDD response time by 19% compared to track caching and no reduction in the read hit percentage was seen. The size of data staged per disk read operation decreased from 57 KB to 21 KB. This test validates that adaptive record caching continues to be a more efficient approach to disk caching than track caching in a TPF environment. The key benefits are lower response time for read misses and lower overall HDD utilization at the same I/O load.

z/TPF I/O Performance Study Results

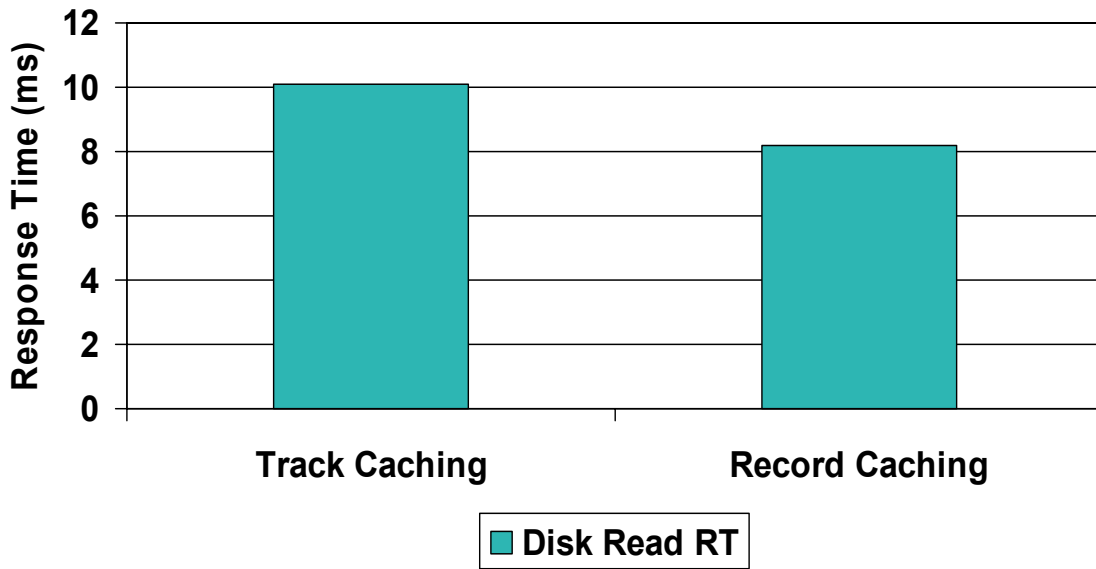


Figure 5 – Physical Disk Read Response Time Lower with Adaptive Record Caching

Performance with FlashCopy

IBM FlashCopy is a method for making a point in time copy of data. When FlashCopy is used, a near instantaneous logical copy of a source volume is created and available for use. A FlashCopy may be made with a background copy or without one (NoCopy). If a background copy is requested, data is asynchronously moved from the source to target volume. As changes are made to tracks on the source, the original copy of that track is moved to the target volume before it is overwritten via destaging the changed track to the source. This process is referred to as "copy source to target" or CST. If the track had previously been copied to the target by the background copy operation or an earlier CST, then no further data movement is needed. With the NoCopy option, all data movement to the target is handled by CST.

FlashCopy can be a valuable method for making copies in a TPF environment. Copies may be made for online backup, testing, or simply dumping data to tape. From a performance perspective, FlashCopy makes additional use of disk system resources for moving data so some level of performance degradation is to be expected. To evaluate how much of an effect FlashCopy has in a TPF environment, a test was run using the Air1 driver and FlashCopy with background copy. Since TPF used software duplexing, it was only necessary to FlashCopy 360 of the 720 total volumes in the configuration. The Air1 driver was ramped up to 163K IO/sec - close to the maximum I/O throughput for the disk system. Just prior to starting the FlashCopy the I/O response time was 2 ms. During the FlashCopy, the average response time more than doubled. Figure 6 below shows the before and after FlashCopy response time results. Considering the heavy I/O load that the storage system was servicing, this result is quite good. Typically a FlashCopy is run when the I/O demand is more modest. Even at this heavy load, the overall response time was still less than 5 ms.

z/TPF I/O Performance Study Results

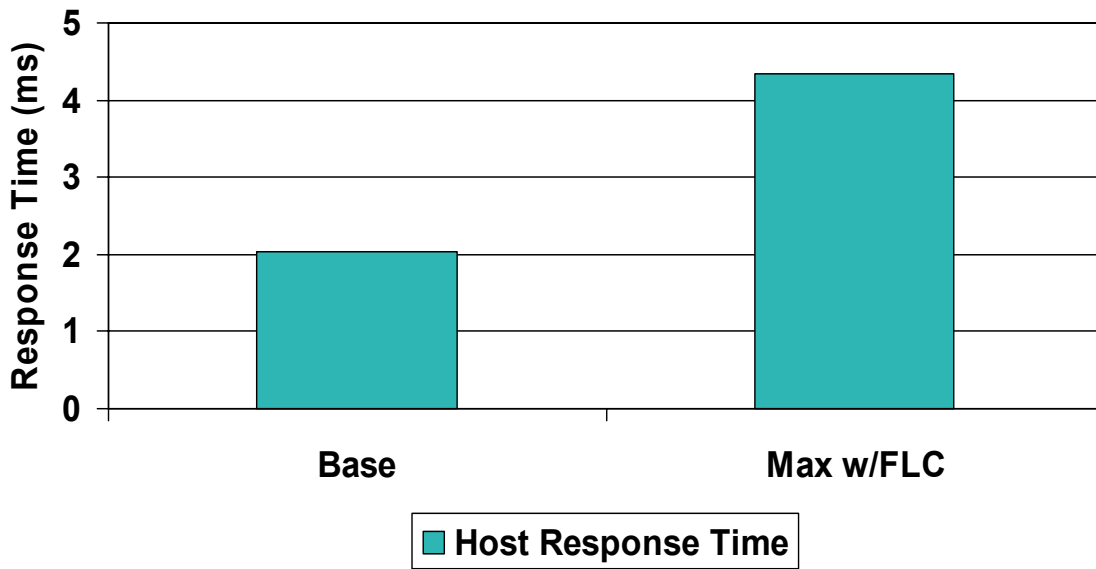


Figure 6 - Host Response Time With and Without FlashCopy Active

Error Handling Behavior Performance

As mentioned earlier, the DS8300 is powered by two dual active p570 servers running licensed internal code that manages most of the storage functionality of the system. Each server contains a 4-way Power 5+ SMP complex. The DS8300 is designed with redundancy such that the storage unit will continue to operate after a hardware failure. For example, if one of the servers were to fail, the surviving server can take over and manage the entire system. This process is called a "failover". Although a failover occurs quickly, there is a small amount of time required to transition to single server mode. During this time, host I/O will be delayed. Likewise, after repairs are made, there will be a small amount of time required to "failback" to dual server mode. Finally there are some soft errors that can be corrected via a simple "warmstart" of one of the server clusters. Again this results in a small amount of time when host I/O may be delayed. These types of scenarios are generally called error handling behavior. To properly tune TPF shutdown parameters, it is useful to have some idea about the typical duration of these kinds of events.

To protect TPF from catastrophic failure during a short term loss of disk access such as a DS8000 failover event, critical TPF block shutdowns (e.g. ECBs, Frames, etc) should be set to roughly 70%. This ensures that while TPF is in Input List Shutdown, the in flight transactions will not exhaust critical resources.

In fact, TPF can protect itself from a catastrophic IPL for any length of time due to an event that causes loss of disk access. Of course this does not mean the time duration is unimportant because the system can be unresponsive during this period and thus appear to end users as an outage. As we shrink the length of time of the failover event, we drive the impact of the TPF system failover to zero. Hence, maintaining short durations for failover events is critically important.

Figure 7 below shows timings for warmstart failover and failback while running TPF at a moderate I/O load using the scaled up Air1 driver. In all cases the timings are very good, never exceeding six seconds. These results are within design expectations for the DS8300 and

z/TPF I/O Performance Study Results

demonstrate the system's excellent performance and resilience under simulated hardware failures.

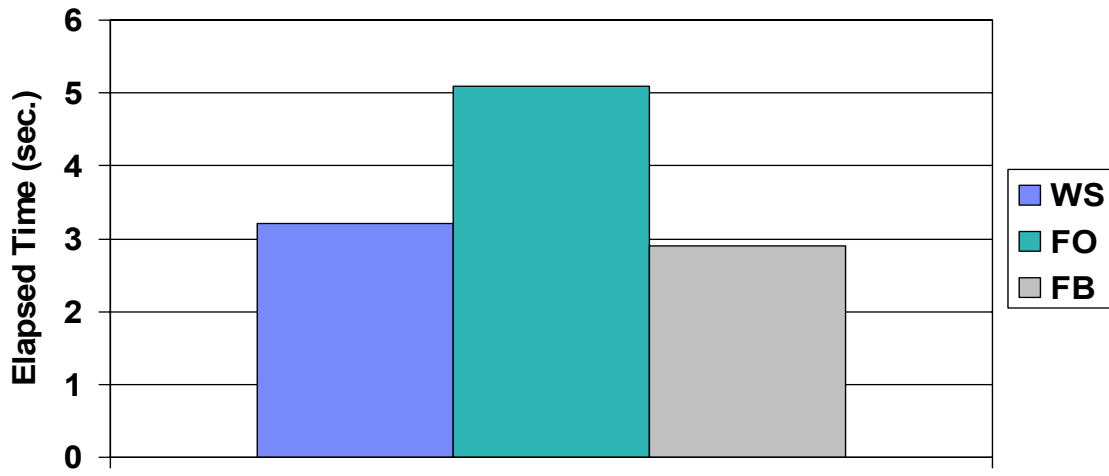


Figure 7 – Sample DS8000 Warmstart/Failover/Failback Timings with TPF Active

Conclusions and Recommendations

The testing detailed in this paper provides an indication of the outstanding performance and throughput that can be achieved using a complete IBM hardware and software solution for TPF. Developing a large scale I/O driver based on TPF was the key to obtaining these results. In the future this driver and variations upon it can be used to test new hardware and software versions as well as enhanced algorithms designed with TPF in mind.

The following expectations may be extrapolated based on the test results and analysis:

- A single DS8300 may achieve over 160K IO/sec in a TPF environment if the cache hit ratio is high.
- IBM adaptive record caching is an effective algorithm for limiting disk read service times and reducing disk and cache utilization in a TPF environment.
- IBM FlashCopy may be used in a TPF environment without excessive impact to host I/O response times.
- The performance of the error handling capabilities of the DS8000 is more than adequate for a TPF requirement. Measured I/O delays were less than 6 seconds, well within the range that a properly tuned TPF system can handle non-disruptively.