z/TPF V1.1

TPF Users Group Fall 2007
Title: z/TPF I/O Performance Study

Name: Robert Blackburn, Allan Feldman,
Lee LaFrese and Leslie Sutton

Venue: System Control Program

AIM Enterprise Platform Software
IBM z/Transaction Processing Facility Enterprise Edition 1.1.0

Any reference to future plans are for planning purposes only. IBM reserves the right to change those plans at its discretion. Any reliance on such a disclosure is solely at your own risk. IBM makes no commitment to provide additional information in the future.

Agenda

- **Objectives**
- **IBM Team**
- **Hardware and Configuration**
- **Key Results**
  - Workload Development
  - Throughput Scaling
  - Adaptive Record Caching
  - FlashCopy Performance
  - Failover Test Results
- **Future Work**
- **Summary and Conclusions**

Objectives

- **Use z/TPF as a high intensity I/O driver for performance testing of large scale system including storage, processor and software**
- **Demonstrate the outstanding performance, throughput and scalability of a z/TPF system**
- **Validate the resilience and performance of z/TPF in an I/O failover scenario**
- **Incorporate the knowledge gained to improve the I/O performance of a z/TPF complex**

IBM Team

- **This project was a collaborative effort across the following IBM organizations**
  - TPF Development (Poughkeepsie)
  - Enterprise Disk Performance (Tucson)
  - zSeries Performance (Poughkeepsie)
- **Cross team skills used to analyze the complete hardware/software stack**

z/TPF Configuration

- **z/TPF (PUT 03)**
- **AIR1 driver**
  - Random 4 KB I/O evenly distributed across:
    - 5M 4 KB records – FINDC
      - Configurable up to 15M records
    - 50K 4 KB records – FILEC
      - Configurable up to 2M records
    - 3 Device types – A (147K), B(473K), and C(4.43M)
  - Other macros mixed in – storage access, CREMC, etc
  - Simulated SNA network
    - Project underway to upgrade AIR1 to use TCP/IP
- **TPF Operations Server 1.2.04**

Hardware Configuration

- **z9 processor, model 2094-S18, 2 books, 18-way**
  - 128 Gb memory
  - 32 FICON Express2 2Gb channels
  - 1-16 I-Streams
  - VTS TS7740 and 3592-E05 TS1120
  - 4 SAPs
- **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)
  - 384 x 146 GB 15K RPM disks (48 RAID
  - 16 x 2 Gb LW host adapters
  - 4 x 720 TPF volumes (3390-9) = 26 TeraBytes usable storage
    - Note: 720 volumes used for performance testing, remainder for copies
  - Configured for resilience to rank, server cluster or SFI failure

z/TPF Workload Characteristics

- **Total storage – 18 GB**
- **VFA storage – 12 GB**
- **1-16 way MP**
- **170K I/O per second max**
- **20K Messages per second**
- **Destage percentage – 10-13%**
- **Read Hit percentage – 96%**
- **Reads per writes – 1.13**

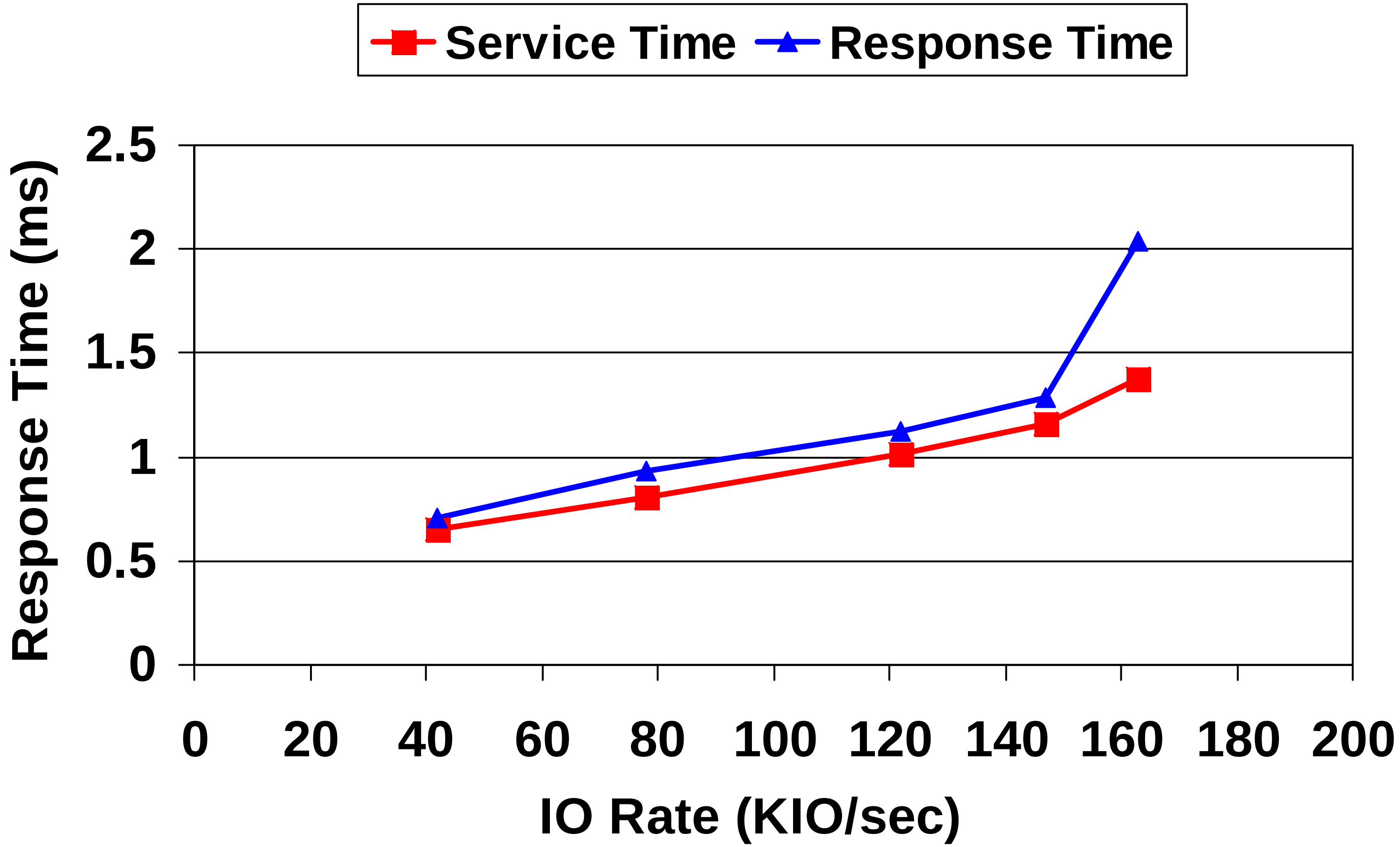
Processor testing

- **Validated processor hardware architecture efficiency for z/TPF workload**
  - Processor Instrumentation
  - Complete instruction traces
  - Processor cache modeling
  - Data Collection
  - Continuous Data Collection
  - Software profiler
- **Verified predicted MP ratio**
  - 1 – 16 way MP

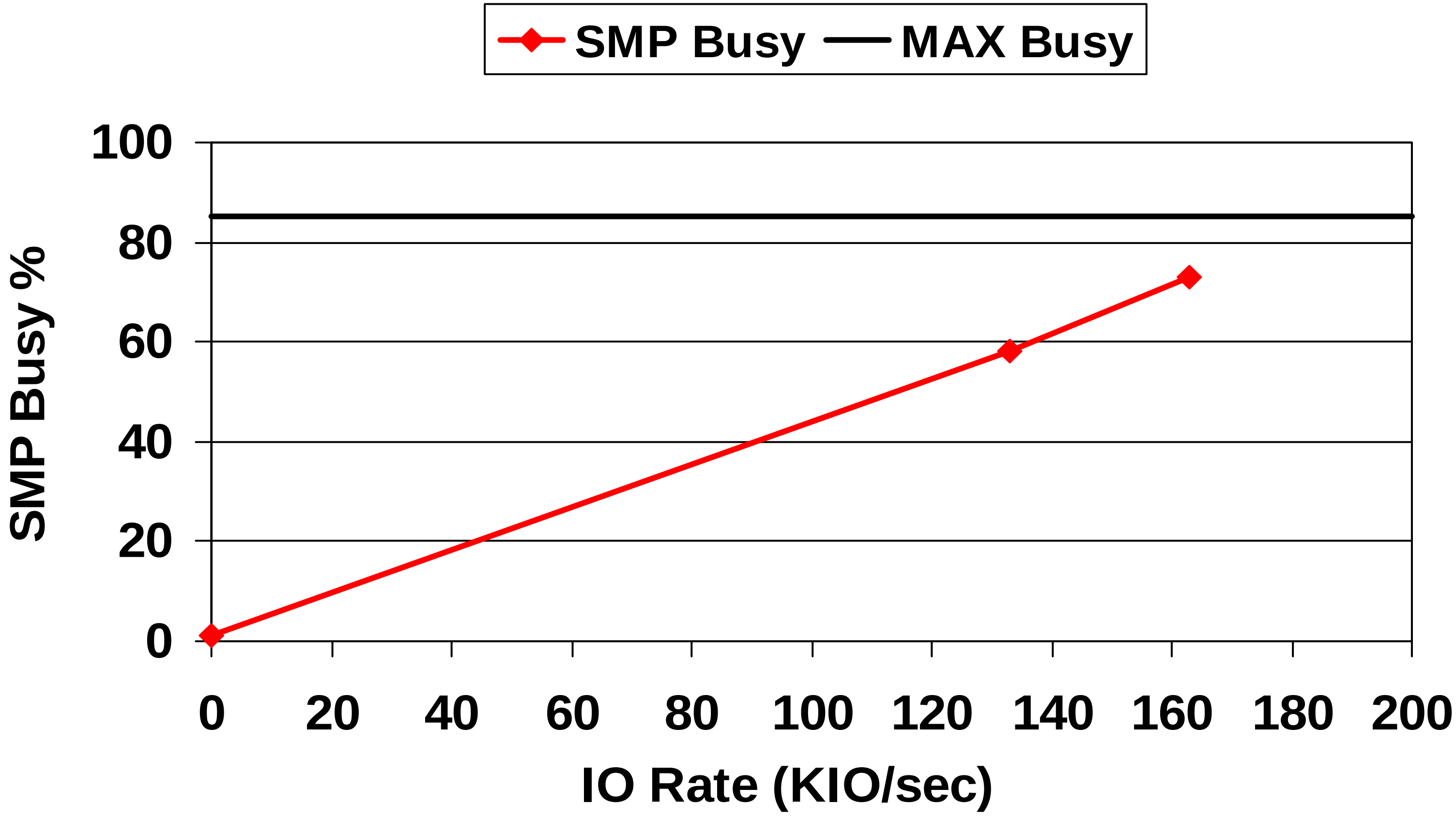
DS8300 Volume Configuration Overview

- **A total of 2880 x 3390-9 volumes were configured**
  - 720 connected to TPF Test system (Test set)
  - 720 connected to z9 host for Performance testing (Perf set)
  - The remainder were allocated for Flash Copy targets, Metro Mirror secondary volumes and spares
- **A total of 36 LCUs were created on 48 RAID ranks**
  - Three LCUs striped across groups of four RAID ranks
  - 20 volumes per LCU in performance set
- **For the performance test volume set**
  - Prime and dup. volumes were placed on separate RAID ranks, SFIs and server clusters for resilience in case of a hardware failure

TPF Disk Ramp Test Results

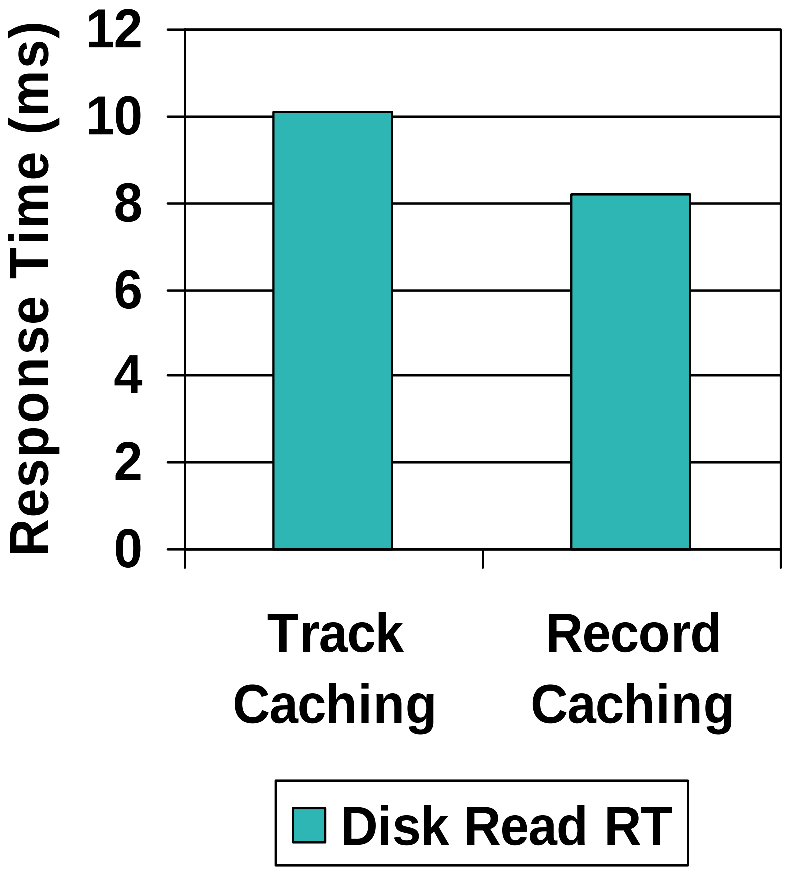


DS8300 Command History SMP Utilization Results



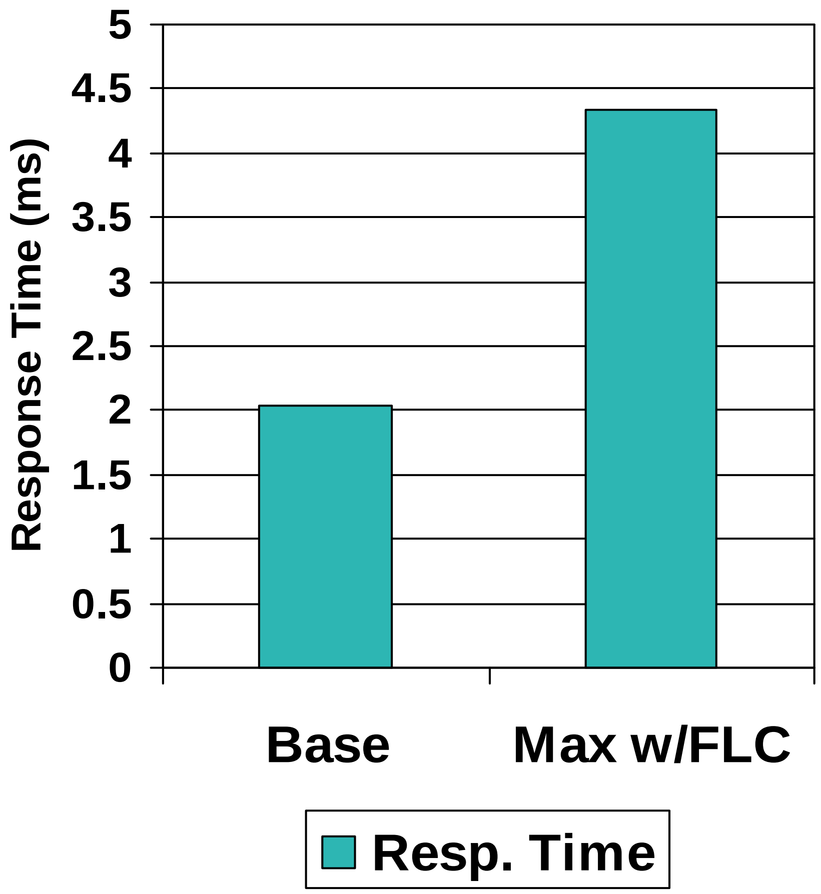
Note: Practical maximum for SMP Busy from Command History is about **85%** because internal mail dispatching times are not measured

Adaptive Record Caching



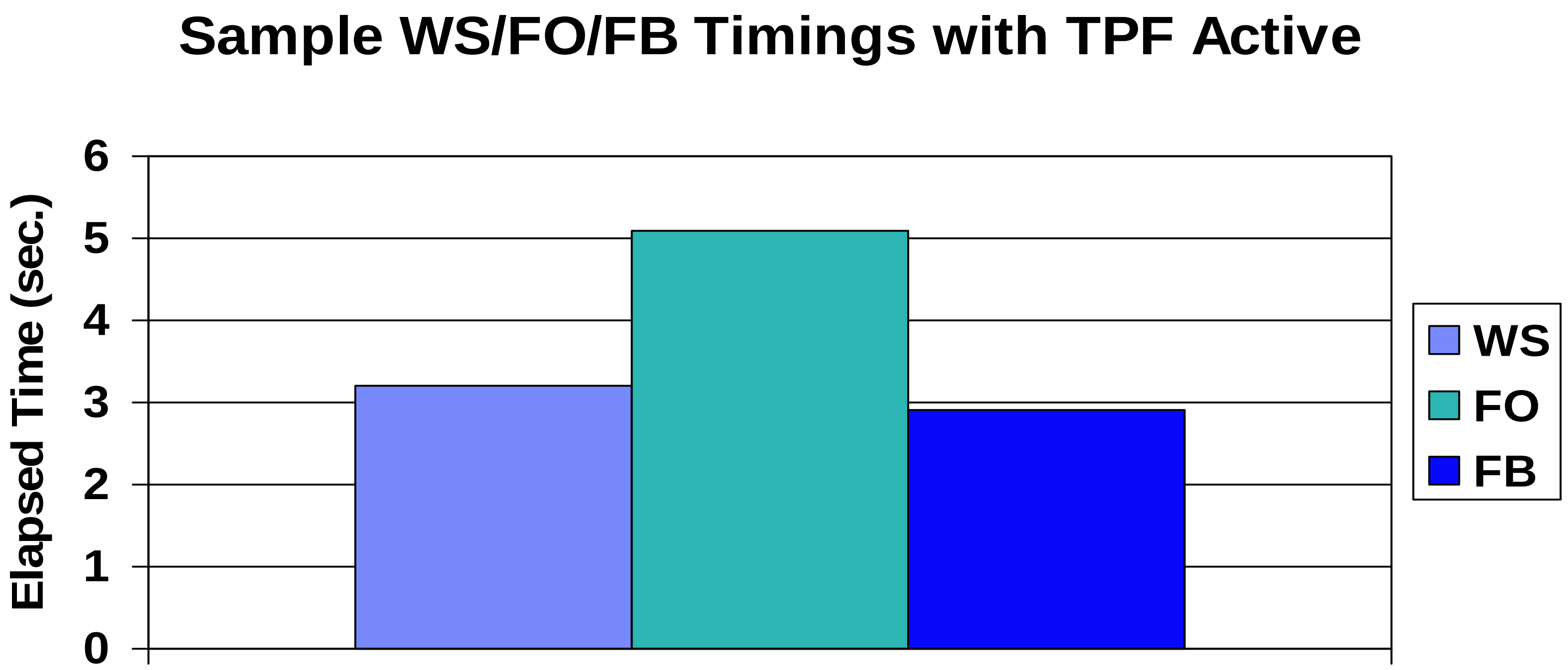
- **Average stage size reduced from 57 KB (full track) to 21 KB when using adaptive record caching.**
- **Adaptive record caching is the default setting for DS8000 and ESS**
- **19% RT improvement with record caching**
  - Drive more disk IOPS
  - Better read miss performance

TPF Disk Performance with FlashCopy



- Workload was running at 164 KIO/sec on the DS8300
- Base is without FlashCopy
- FlashCopy w/ background copy started on 360 prime volumes
- Max FLC is the maximum observed response time after the FlashCopy was initiated (Max IOB count)
- This is a worst case scenario. Normally FlashCopy would be initiated when the host workload was more moderate

TPF I/O Hardware Error Handling Behavior



WS - Warmstart, reset memory in both server clusters

FO - Failover, one server cluster takes over for the other

FB – Failback, the failed server cluster is brought back online

Useful information for tuning TPF shutdown parameters

Future Work Plans

- **Further MP testing beyond 16-way**
- **Loosely Coupled**
- **Remote Copy**
- **Multi-Path Reconnect**
- **Study different TPF workload variations**
- **Study/Improve Error Handling Behavior timings and TPF resiliency**

Summary and Conclusions

- **Very satisfactory results from this full scale z/TPF I/O performance test**
  - Excellent throughput and scalability
  - Performance and resilience meet expectations during simulated I/O hardware failures
- **This type of cross-brand collaboration will continue**
  - Results of studies like this feed into future product designs

Backup

Materials

DS8300 Volume Configuration (repeated six times)

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						

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 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 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						

Server 1

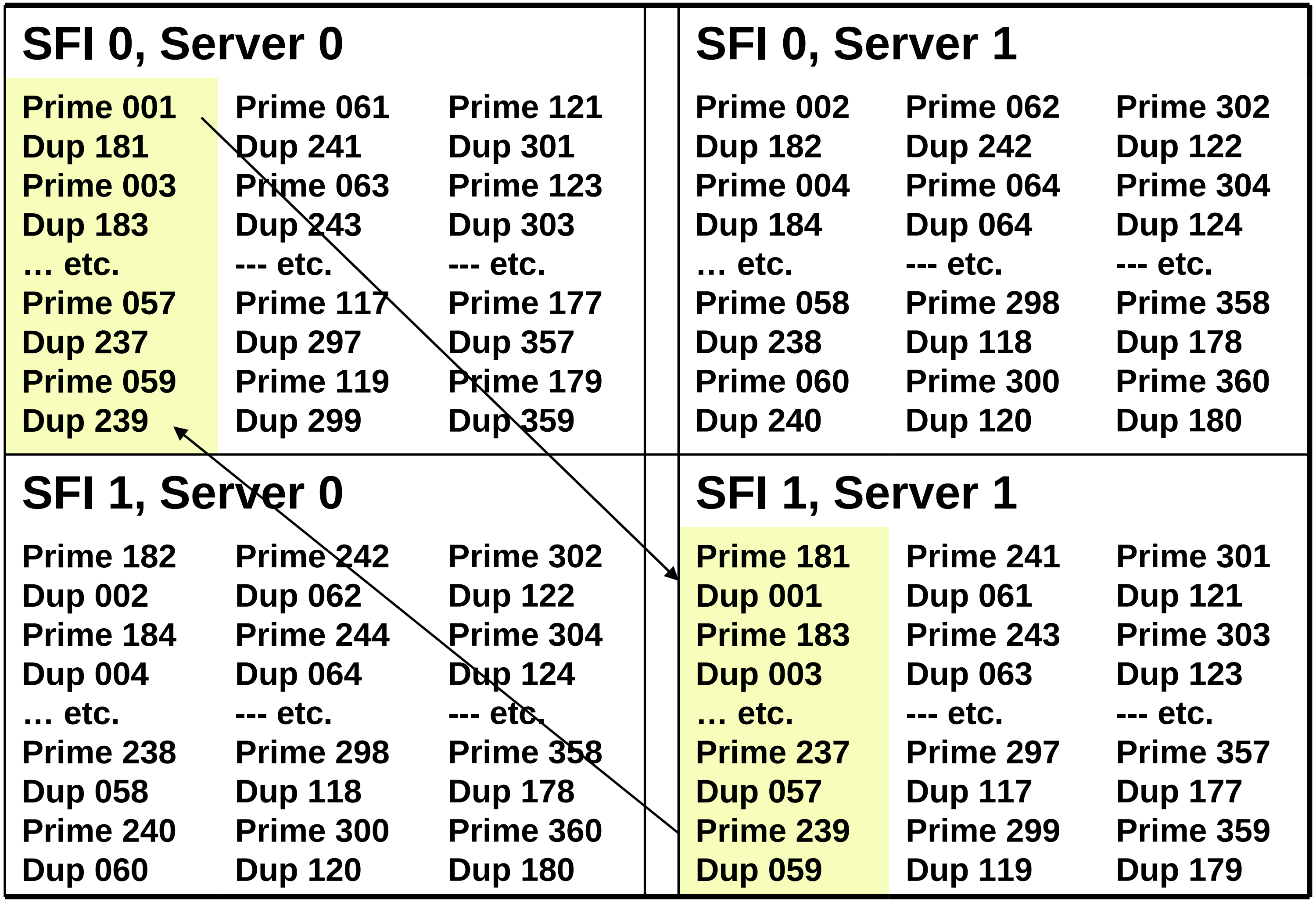
Rank 1	Sets	**LSS1**	(P,D)	**LSS3**	(P,D)	**LSS5**	(P,D)
R-10 3+3	TEST	4	(2,2)	4	(2,2)	4	(2,2)
48 x 3390-9	PERF	4	(2,2)	4	(2,2)	4	(2,2)
	FC TGT	2	(2,0)	2	(2,0)	2	(2,0)
	PPRC SEC	4	(2,2)	4	(2,2)	4	(2,2)
	SPARE	2		2		2	
	TOTAL	**16**		**16**		**16**	
	48						

Rank 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 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 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						

Placement of Primes and Dups for Resilience



SFI 0, Server 0

Prime 001
Dup 181
Prime 003
Dup 183
… etc.
Prime 057
Dup 237
Prime 059
Dup 239

Prime 061
Dup 241
Prime 063
Dup 243
--- etc.
Prime 117
Dup 297
Prime 119
Dup 299

Prime 121
Dup 301
Prime 123
Dup 303
--- etc.
Prime 177
Dup 357
Prime 179
Dup 359

SFI 0, Server 1

Prime 002
Dup 182
Prime 004
Dup 184
… etc.
Prime 058
Dup 238
Prime 060
Dup 240

Prime 062
Dup 242
Prime 064
Dup 064
--- etc.
Prime 298
Dup 118
Prime 300
Dup 120

Prime 302
Dup 122
Prime 304
Dup 124
--- etc.
Prime 358
Dup 178
Prime 360
Dup 180

SFI 1, Server 0

Prime 182
Dup 002
Prime 184
Dup 004
… etc.
Prime 238
Dup 058
Prime 240
Dup 060

Prime 242
Dup 062
Prime 244
Dup 064
--- etc.
Prime 298
Dup 118
Prime 300
Dup 120

Prime 302
Dup 122
Prime 304
Dup 124
--- etc.
Prime 358
Dup 178
Prime 360
Dup 180

SFI 1, Server 1

Prime 181
Dup 001
Prime 183
Dup 003
… etc.
Prime 237
Dup 057
Prime 239
Dup 059

Prime 241
Dup 061
Prime 243
Dup 063
--- etc.
Prime 297
Dup 117
Prime 299
Dup 119

Prime 301
Dup 121
Prime 303
Dup 123
--- etc.
Prime 357
Dup 177
Prime 359
Dup 179

Disclaimers

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 could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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. Customer experiences described herein are based upon information and opinions provided by the customer. The same results may not be obtained by every user.

Reference 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. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead. It is the user's responsibility to evaluate and verify the operation on any non-IBM product, program or service.

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 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.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

The providing 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
USA

Trademarks

- **IBM, z9, zSeries, FICON and DS8000 are trademarks of International Business Machines Corporation in the United States, other countries, or both. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml**
- **Other company, product, or service names may be trademarks or service marks of others.**

- **Notes**
- **Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.**
- **All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.**
- **This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.**

- **All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.**

- **Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.**

- **Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.**

- **This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.**