Redbooks
ibm.com/redbooks

Best Practices Guide for Databases on IBM FlashSystem

Jagadeesh Papaiah

Storage

IBM
Redpaper

Best Practices Guide for Databases on IBM FlashSystem

The purpose of this IBM® Redpaper® document is to provide best practice guidelines to design and implement IBM FlashSystem® storage for database workloads. The recommended settings and values are based on lab testing, proof of concept (PoC) and experience drawn from customer implementations. Suggestions that are presented in this document are applicable to most production database environments to increase performance of I/O and availability. However, more considerations might be required while designing, configuring, and implementing storage for extreme transactional, analytical, and database cluster environments.

Customers are migrating database storage to IBM FlashSystem largely due to low latency performance of the IBM FlashSystem family of Storage. Using IBM FlashSystem, IBM customers are able to achieve low latency for queries and transactions from milliseconds to microseconds, realize a multi-fold increase in application level transactions per second, increase CPU efficiency and reduce database licensing costs.

Recent additions of data reduction technologies to IBM FlashSystem further increase overall TCO benefits. All IBM FlashSystem models now offer compression, which can reduce database storage by 40 - 80% depending on database software.

In addition to best practices that are described in this document, the IBM FlashSystem Worldwide Solutions Engineering Team can further assist customers with performing analysis of current database workloads for IBM FlashSystem benefits, perform PoCs at our labs, and help with implementation.

Oracle

Oracle is one of the most popular databases on IBM FlashSystem in terms of size of the databases and the resulting low latency requirements. For Oracle Databases, IBM FlashSystem average latency ranges 300 microseconds - 1 millisecond depending on the IBM FlashSystem model and database workload type variations.

Table 1 shows the recommended `init.ora` values, ASM, and log file considerations to achieve a balanced performance, also refer to OS consideration sections and storage layout for settings corresponding to OS and IBM FlashSystem model.

Table 1 The database for Oracle systems settings

Parameter	Default setting	Recommendations	Description
`FILESYSTEMIO_OPTIONS`	Varies by database and OS	`SETALL`	`SETALL` enables both direct and asynch I/O.
Block size	8 KB (Range 2 KB - 32 KB)	Not modifiable after DB creation, 8 KB optimal for most DBs	Large block size for LOBs and set at the table space level.
`DB_FILE_MULTIBLOCK_READ_COUNT`	Default value corresponds to the maximum I/O size and is platform-dependent	32 optimal for IBM FlashSystem, at 32 MBR and Blksize 8 KB, average read scan size that is issued to 256 KB sequential for table scans, see table below for testing results	Specifies the maximum number of blocks that are read in one I/O operation during a sequential scan.
Redo log file block size	512 Bytes	4 KB blksize, set `"disk_sector_size_override"=TRUE` to add log file with 4 KB blksize	On IBM FlashSystem, 4 KB block size is optimal and reduces '`log file synch`' waits.
ASM Disk Redundancy	Created, with the following options: ► External = 1x copy ► Normal = 2x copies ► High = 3x copies	External	For high availability, create disk group as `Normal` and mirror disks across two arrays.
ASM Allocation Unit - AU size	1 MB	4 MB for OLTP DB and 8 MB for VLDBs	4 MB is optimal for most databases and larger value 16-32 might provide performance benefits for data warehouse type applications.

On databases with more DSS/analytic type workloads, significant number of table scans are issued by Oracle database, which results in large block sequential reads. Sequential reads to IBM FlashSystem might be further tuned for optimization. The following table illustrates query response time differences with varying multi-block read count. Based on our lab testing that uses HammerDB TPCH schema tables, a combination of 8 KB blksize and 32 multiblock read (MBR) count achieved the lowest response time.

Table 2 on page 3 shows the query response time differences based on SQL queries against HammerDB TPCH schema tables. Negative query response time % differences are shown for varying block size MBR combination.

Table 2 Query response time:

ORA Block size KB	MBR Count	Query Response Time Differences	Read Scan Block Size KB
8	128	-25%	1024
8	64	- 2%	512
8	32		256
8	16	- 4%	128
8	8	- 8%	64
8	4	-23%	32

Oracle I/O calibration

Consider using the Oracle-provided stored procedure **DBMS_RESOURCE_MANAGER.CALIBRATE_IO** for I/O calibration. This action is optional, and it is not required for IBM FlashSystem implementation.

The I/O calibration feature of Oracle Database enables you to assess the performance of the storage subsystem, and determine whether I/O performance problems are caused by the database or the storage subsystem. Unlike other external I/O calibration tools that issue I/Os sequentially, the I/O calibration feature of Oracle Database issues I/Os randomly uses Oracle data files to access the storage media, producing results that more closely match the actual performance of the database.

This procedure issues an I/O intensive read-only workload, made up of 1 MB of random I/Os, to the database files to determine the maximum I/O operations per second (IOPS) and megabytes of I/O per second (MBPS) that can be sustained by the storage subsystem.

The I/O calibration occurs in two steps:

1. In the first step of I/O calibration with the **DBMS_RESOUCE_MANGER.CALIBRATE_IO** procedure, the procedure issues random database-block-sized reads (by default, 8 KB) to all data files from all database instances. This step provides the maximum IOPS, in the output parameter **MAX_IOPS**, that the database can sustain. The value **MAX_IOPS** is an important metric for OLTP databases. The output parameter **MAX_LATENCY** provides the average latency for this workload. When you need a specific target latency, you can specify the target latency with the input parameter **MAX_LATENCY** specifies the maximum tolerable latency in milliseconds for database-block-sized I/O requests.
2. The second step of calibration that uses the **DBMS_RESOUCE_MANGER.CALIBRATE_IO** procedure issues random, 1 MB reads to all data files from all database instances. The second step yields the output parameter **MAX_MBPS**, which specifies the maximum MBPS of I/O that the database can sustain. This step provides an important metric for data warehouses.

The calibration runs more efficiently if the user provides the number of physical disks input parameter, which specifies the approximate number of physical disks in the database storage system.

Run the **DBMS_RESOURCE_MANAGER.CALIBRATE_IO (<*DISKS*>, <*MAX_LATENCY*>, iops, mbps, lat);** procedure.

The input values for IBM FlashSystem `DISKS` = number of back-end IBM Tivoli® Storage IBM FlashCopy® Manager modules or NVMe drives, and `MAX_LATENCY` = `10`.

Thee outputs are `maxiops`, `maxmbps`, and `latency`.

Caution: Due to the resources required to run the I/O workload, I/O calibration should be performed only when the database is idle, or during off-peak hours, to minimize the impact of the I/O workload on the normal database workload.

Microsoft SQL Server

SQL Server implementations on IBM FlashSystem are mostly on VMware virtual machines, and recommendations apply to both bare metal and virtual machines. More considerations are listed under VMware considerations. Table 3 shows the parameters for SQL server.

Table 3 Parameters for Microsoft SQL Server

Parameter	Default setting	Recommendations	Description
Page size	8 KB	Not modifiable	Disk I/O operations are performed at the page level.
Extent Size	64 KB	Not modifiable	Extent is eight physically contiguous pages, and the databases have 16 extents per megabyte.
Log files	One log file	Use separate drive for logs and use dedicated volumes for log files	I/Os to log file are primarily writes.
TempDB	One data file	Multiple, 1 datafile/cpu or core, and pre-size	On databases with significant sorts, multiple files and dedicated LUNs improves performance.
Data files	One data file	Multiple data files (per filegroup) for each CPU on the host server, and pre-size	For large databases creating multiple volumes 8-32 for data files improves performance.
Backup: **BUFFERCOUNT**	Varies	32, commands in queue	Option can be set at SQL command level or at tools level.
Backup: **MAXTRANSFERSIZE**	1 MB	2 MB - 4 MB	Option can be set at SQL command level or at backup tools level.

Microsoft recommendations for SQL Server files and filegroups:

- Most databases work well with a single data file and a single transaction log file.
- If you use multiple data files, create a second filegroup for the additional file and make that filegroup the default filegroup. In this way, the primary file contains only system tables and objects.
- To maximize performance, create files or filegroups on different available disks as possible. Put objects that compete heavily for space in different filegroups.
- Use filegroups to enable placement of objects on specific physical disks.
- Put different tables used in the same join queries in different filegroups. This step improves performance because of parallel disk I/O searching for joined data.
- Put heavily accessed tables and the nonclustered indexes that belong to those tables on different filegroups. Using different filegroups improves performance because of parallel I/O if the files are on different physical disks.
- Do not put the transaction log files on the same physical disk that has the other files and filegroups.
- If you need to extend a volume or partition on which database files reside by using tools like Diskpart, you should back up all system and user databases and stop SQL Server services first. Also, after disk volumes are extended successfully, you should consider running DBCC CHECKDB command to ensure the physical integrity of all databases residing on the volume.

IBM Db2

Table 4 shows the recommendations that apply to IBM Db2® Linux, UNIX, Windows (LUW) and does not apply to Db2 on IBM z/OS®.

Table 4 Db2 Linux, UNIX, and Windows recommendations

Parameter	Default setting	Recommendations	Description
Page Size	4 KB - 32 KB parameter contains the value that was used as the default page size when the database was created.	4 KB for default.	4 KB optimal for OLTP and 16 KB - 32 KB for analytics and LOB set at table space level.
dft_extent_sz	Thirty-two pages.	Use default size for all table spaces.	Default subject to change by config advisor.
dft_prefetch_sz	Automatic.	Default good enough for most table spaces.	Default subject to change by config advisor.
Table space management	Automatic if managed by clause is not specified or specified as 'Automatic' during table space creation.	Automatic.	Use Automatic and avoid using SMS and DMS table space as they will be deprecated in future versions.

Parameter	Default setting	Recommendations	Description
`OVERHEAD`, `DEVICE READ RATE`	6.725 ms, 100 MBps.	Default.	Defaults are OK for most databases, and consider changing `OVERHEAD` to 1 ms for high OLTP. If the Db2 database was upgraded from versions earlier than 10.1, then the existing table spaces retain the `OVERHEAD` and `DEVICE READ RATE` attributes for that storage group, which is set to undefined.

Operating system considerations

This section shows the operating system settings that are optimal for IBM FlashSystem based testing, customer implementations, and corresponding vendor recommendations.

Linux

Use deadline for I/O scheduler and consider using the `tuned-profiles-oracle` package for RHEL and other database specific tuned profiles available based on Linux distribution and releases. To ensure path failover policy and that timeouts are set to appropriate values, see the IBM recommended device mapper multipath and udev rules for the corresponding Linux distribution and specific release.

AIX

Consider migrating to AIXPCM and refer to multipath configuration and best practices on IBM Documentation.

VMware considerations for SQL Server

Table 5 shows the VMware parameters for SQL Server.

Table 5 VMware parameters for SQL server

Description	Recommendations
VMFS	Place SQL Server data (system and user), transaction log, and backup files into separate VMDKs (if not using RDMs). The SQL Server binary files are usually installed in the OS VMDK. Separating SQL Server installation files from data and transaction logs also provides better flexibility for backup, management, and troubleshooting.

Description	Recommendations
Data store versus RDMs	Performance differences are not high enough except for high OLTP databases based on VMware testing. However, RDMs are required for SQL Server `AlwaysON` FCI.
Storage I/O Control	Consider Storage I/O Control setting for mixed VM environment.
ESX HBA queue depth	Default 32 - 64, set to 128.
ESX Disk.DiskMaxIOSize	Default 32767 KB, set to 4 MB.
ESX PSP policy	Round robin.
ESX PSP IOPS limit	Default 1000, set 1 - 10.

Disk layout for IBM FlashSystem

Figure 1 shows the disk layout for designing and mapping database volumes to IBM FlashSystem volumes. Balanced performance and lower latency can be achieved by using multiple VDisks for database data volumes.

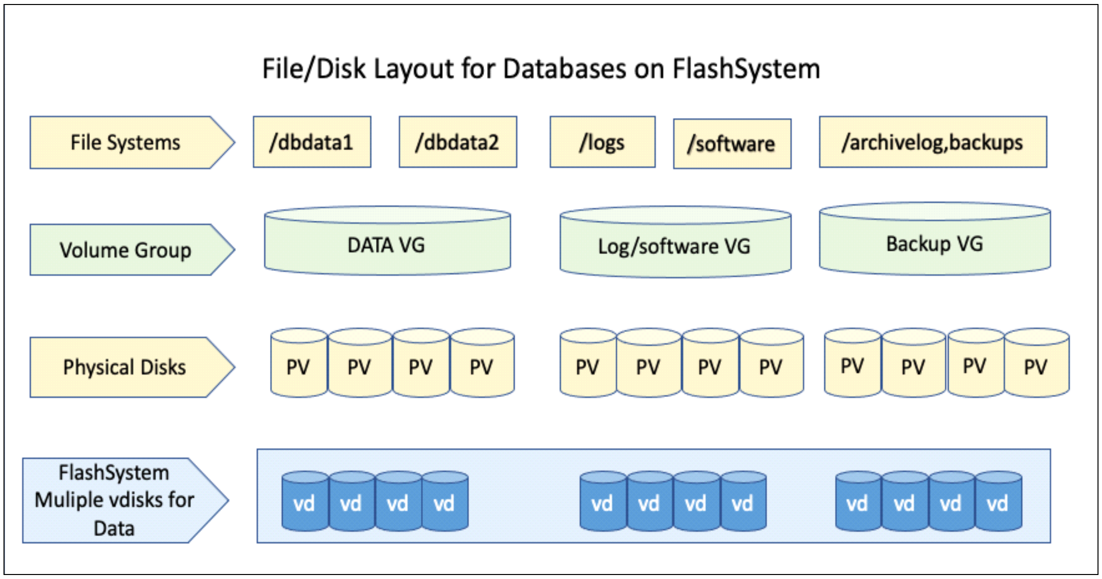

Figure 1 File/disk layout for databases on IBM FlashSystem

Considerations for IBM FlashSystem

Consider the following practices when configuring IBM FlashSystem devices:

- Separate data by creating separate VDisks for data, logs, archive logs, backups, and software installation binary files.
- Multiple VDisks are recommended for database data (16 - 32) for High OLTP.

- Volume level compression is not recommended database redo or transaction logs.
- Volume level compression is not recommended if compression is turned on at the database or table level.

References

These websites are also relevant as further information sources:

- https://docs.oracle.com/en/database/oracle/oracle-database/21/tgdba/IO-configuration-and-design.html
- http://docs.oracle.com/cd/B19306_01/server.102/b14211/iodesign.htm#i19636
- https://docs.microsoft.com/en-us/sql/relational-databases/pages-and-extents-architecture-guide
- https://www.ibm.com/docs/en
- https://technet.microsoft.com/en-us/library/cc966534.aspx
- https://access.redhat.com/solutions
- https://www.ibm.com/docs/en/flashsystem-9x00/8.2.1?topic=system-settings-linux-hosts

Author

Jagadeesh Papaiah is a Corporate Solutions Architect. As a member of the IBM Worldwide IBM FlashSystem Solutions Engineering team, he works with customers, IBM Business Partners, and IBM employees worldwide on consulting, designing, and implementing infrastructure solutions. He has over 25 years of experience in Information Management, integration architecture, infrastructure services, IT strategy and architecture, and solution design.

Now you can become a published author, too

Here’s an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time. Join an IBM Redbooks® residency project and help write a book in your area of expertise, while honing your experience by using leading-edge technologies. Your efforts help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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 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.

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

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

Db2®
FlashCopy®
IBM®
IBM FlashSystem®
Redbooks®
Redbooks (logo) ®
Tivoli®
z/OS®

The following terms are trademarks of other companies:

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

VMware, and the VMware logo are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

Other company, product, or service names may be trademarks or service marks of others.

REDP-5520-01

ISBN 0738459828

Printed in U.S.A.