

# Using 128 TB Capacity per X-Blade with EMC Celerra

*Best Practices Planning*

---

## **Abstract**

This white paper provides an overview of considerations when utilizing the supported 128 TB storage limit per X-Blade. It includes discussions on existing best practices, approaches to scaling, and design implications of existing hard/soft limits within feature sets of EMC<sup>®</sup> Celerra<sup>®</sup> Network Server.

March 2009

---

---

Copyright © 2009 EMC Corporation. All rights reserved.

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” EMC CORPORATION MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com

All other trademarks used herein are the property of their respective owners.

Part Number h6154

---

Using 128 TB Capacity per X-Blade  
with EMC Celerra  
Best Practices Planning

---

## Table of Contents

<b>Executive summary .....</b>	<b>4</b>
<b>Introduction .....</b>	<b>4</b>
Audience .....	4
Terminology .....	4
<b>Memory usage.....</b>	<b>5</b>
System memory usage .....	5
X-Blade cache.....	5
<b>File count and sizes.....</b>	<b>6</b>
<b>File system planning .....</b>	<b>6</b>
Planning phase tip: Expected IOPS calculation.....	8
<b>SnapSure checkpoints .....</b>	<b>8</b>
<b>Celerra Replicator .....</b>	<b>9</b>
<b>Backup and restore considerations .....</b>	<b>10</b>
<b>Planning and design workflow .....</b>	<b>11</b>
<b>Conclusion .....</b>	<b>12</b>
<b>References .....</b>	<b>12</b>

---

## Executive summary

EMC® Celerra® Network Server version 5.6 supports the utilization of 128 TB of usable storage per X-Blade on certain platforms. This increase in supported storage per X-Blade requires discussions on and understanding of existing limitations and best practices because it is now possible to reach these limits more frequently.

The ability to utilize 128 TB per X-Blade allows for more capacity per X-Blade. However, the performance characteristics of the system do not change. Therefore, it is vital to understand that this increase in usable capacity does not translate into a greater number of I/O operations per second (IOPS). The performance characteristics of the solution are still dependent on the workload, number of available spindles per file system, and system resources within the X-Blade.

Before Celerra version 5.6, the usable capacity per X-Blade was 32 TB for Fibre Channel (FC) storage and 64 TB for SATA storage. The SATA limit was set higher than the FC limit because SATA solutions were often used for archiving and backup solutions and even the other use cases for SATA drives were focused on capacity over performance. Over the past year, the usable capacity for each X-Blade has dramatically increased for the EMC Celerra platform. This increase creates the need for better planning with special considerations for file system sizing, replication sessions, SnapSure™ checkpoint scheduling, and backup/restore planning. It is important to note that the underlying limits — including those for maximum file system size, number of files per directory, number of checkpoints per file system, and number of replication sessions — have not changed. The guidelines provided within this paper currently pertain only to the EMC Celerra NS80, NSX, NS-960, and NS-G8 platforms that run Celerra version 5.6 or later. The other available platforms and Celerra versions currently do not support 128 TB of usable capacity per X-Blade. It is important to note that although the limit has been increased for FC, SATA, and enterprise Flash drive storage, the general driver for solutions that utilize 128 TB should be storage capacity and not performance. If a solution is being developed for a performance-sensitive use case or application, then utilizing the full amount of available capacity is not recommended. The limits provided here were the limits at the time this paper was published. Please check the *NAS Support Matrix* for the latest limits.

The general approach to solution design with the increased capacity per X-Blade should focus on performance requirements first and then on utilizing the increased limit for scaling capacity. When positioning this enhancement to capacity, you must be certain to set expectations properly by using the correct value drivers within the discussion. When conducting the planning and design phases, you must model the forecasted capacity utilization effectively and create the initial file systems by using appropriate sizes to balance the solution performance, capacity, and scaling requirements.

## Introduction

The allowance of additional capacity impacts existing NAS solution design and scaling. This white paper provides considerations for certain features and areas because of this impact. The features and areas covered include memory usage, file system design, SnapSure checkpoints, Celerra Replicator™, and backup/restore operations. Numerous existing best practices and design recommendations are applicable to these features and they are discussed in detail in this white paper.

## Audience

This white paper is intended for EMC field personnel, partners, and customer IP storage managers. This paper requires an intermediary understanding of IP storage technologies and concepts.

## Terminology

Checkpoint — Point-in-time, logical image of a production file system. A checkpoint is a file system and is also referred to as a checkpoint file system or a SnapSure file system.

File system — Method of cataloging and managing the files and directories on a storage system.

---

Inode — “On-disk” data structure that holds information about files in a file system. This information identifies the file type as being a file including Celerra FileMover stub files, a directory, or a symbolic link.

Production file system (PFS) — Production file system on a Celerra Network Server. A PFS is built on Symmetrix® volumes or CLARiiON® LUNs and mounted on a Data Mover in the Celerra Network Server.

Redundant Array of Independent Disks (RAID) — Method for storing information where the data is stored on multiple disk drives to increase performance and storage capacities and to provide redundancy and fault tolerance.

## Memory usage

### System memory usage

The EMC Celerra X-Blade has a finite amount of system memory, which is allocated and managed by the DART operating system. The available memory is divided into two main areas for the system. A portion of the available memory is reserved for system functions such as storing the active configuration for the X-Blade, and mapping user IDs (secmap cache). X-Blade uses the rest of the available memory for caching. The increase in supported capacity per X-Blade has shown no significant impact on the portion of memory resources dedicated to system functionality. Therefore, the existing method and amount of “reserved” memory do not require any modification because of the additional capacity.

### X-Blade cache

Table 1 and Table 2 show that the only significant impact of increased usable memory is that although the same percentage of memory is available for use, the percentage of operations that might be cached per file system changes based on the number and size of the mounted user file systems. You can consider the amount of memory used by each file system as a function of the workload and size of the file system.

**Table 1. System cache distribution with 64 TB per X-Blade**

	<b>X-Blade</b>
<b>Total capacity of mounted file systems (TB)</b>	64
<b>Cache usage (GB)</b>	3
<b>Percentage of cached data per X-Blade</b>	0.0046%
<b>Percentage of cached data per FS (two 16 TB file systems)</b>	0.0046%

**Table 2. System cache distribution with 128 TB per X-Blade**

	<b>X-Blade</b>
<b>Total capacity of mounted file systems (TB)</b>	128
<b>Cache usage (GB)</b>	3
<b>Percentage of cached data per X-Blade</b>	0.0023%
<b>Percentage of cached data per FS (eight 16 TB file systems)</b>	0.0023%

The total capacity allocated per X-Blade has increased from 64 TB to 128 TB, but each file system on an X-Blade cannot exceed the existing maximum size limit of 16 TB. The performance characteristics of Celerra version 5.6 and the specific EMC Celerra platform do not change because of this increase in capacity. Therefore, it is still recommended to design solutions with both the capacity and performance objectives in mind. The following example illustrates this concept.

Your organization has recently purchased a brand new EMC Celerra NS80 (with three X-Blades) and plans to use it to host two projects that are ready for implementation. The SPECsfs@97\_R1 results are achieved by using specific configurations and workloads that do not reflect any conceivable customer workload. This

example uses the SPECsfs@97 results for illustration. The actual performance of the Celerra may differ based on the workloads and configuration at your site. The NS80 platform has a published SPECsfs@97\_R1 benchmark result of achieving 43,318 IOPS per X-Blade. The two projects that are being considered for IP storage have differing requirements for performance and capacity. The performance requirements are based solely on transaction count, and throughput/I/O sizes are not factored for simplicity. One project requires 18,000 IOPS and 80 TB of storage, and the other requires 30,000 IOPS and 20 TB of storage. Even though the total capacity required can be accommodated on a single X-Blade (20 TB + 80 TB = 100 TB), the required number of IOPS (30K + 18K = 48K) cannot be accommodated. Therefore, you must use two X-Blades because the sizing factor is performance and not capacity. The planning and design phase should always account for performance requirements first and then factor in capacity requirements. There is a common misconception that a lesser number of drives are required for a solution if the drive density increases. However, this is true only for solutions that must deliver only capacity. In most cases, the drive count cannot change for a solution due to the number of spindles necessary to achieve performance requirements.

Thorough planning and design of an IP storage deployment can provide numerous benefits and increase the likelihood of reaching a desired performance level. When planning for performance-based solutions, ensure that you address the following three parameters:

- Drive count (the number of spindles available to a file system can directly affect its performance)
- Proper distribution of data across RAID groups (to minimize hot spots, file systems should be created across LUNs with similar characteristics and storage capacity)
- Capacity

## File count and sizes

The number of files and their size also must be factored in when designing a high capacity solution. The current limits are provided in Table 3.

**Table 3. Current file size and count limits (as of February 2009)**

	<b>Limit</b>
<b>Maximum file size</b>	16 TB
<b>Maximum number of subdirectories per parent directory</b>	65,533
<b>Maximum path name length</b>	1,024 bytes

The size of each individual file system cannot exceed 16 TB. However, the total usable capacity per X-Blade is now 128 TB, which indicates that in a typical environment, you can now mount more file systems on a single X-Blade than in the past. It is also important to note that when creating file systems you should always try to build “right-sized” file systems and balance the performance and capacity requirements both at a file system and at an X-blade level. The concept of “right-sizing” is discussed in the next section.

You must take into account the number of files expected to reside within each file system during the design phase. Celerra allows 257 million inodes (roughly equal to the number of files that can be supported) by default, but this number can be increased to a maximum value of approximately 4 billion. If the solution you are deploying requires more than 257 million objects per file system, consider using multiple file systems and combining them into a namespace by using Nested Mount File (NMFS) systems. The number of files per file system can also affect your ability to meet backup/restore requirements. The greater the number of files in a given file system, the longer it takes to back up/restore the file system.

## File system planning

The current limitations regarding file system size and the number of file systems per X-Blade are provided in Table 4 on page 7.

---

**Table 4. File system size and count limits (as of February 2009)**

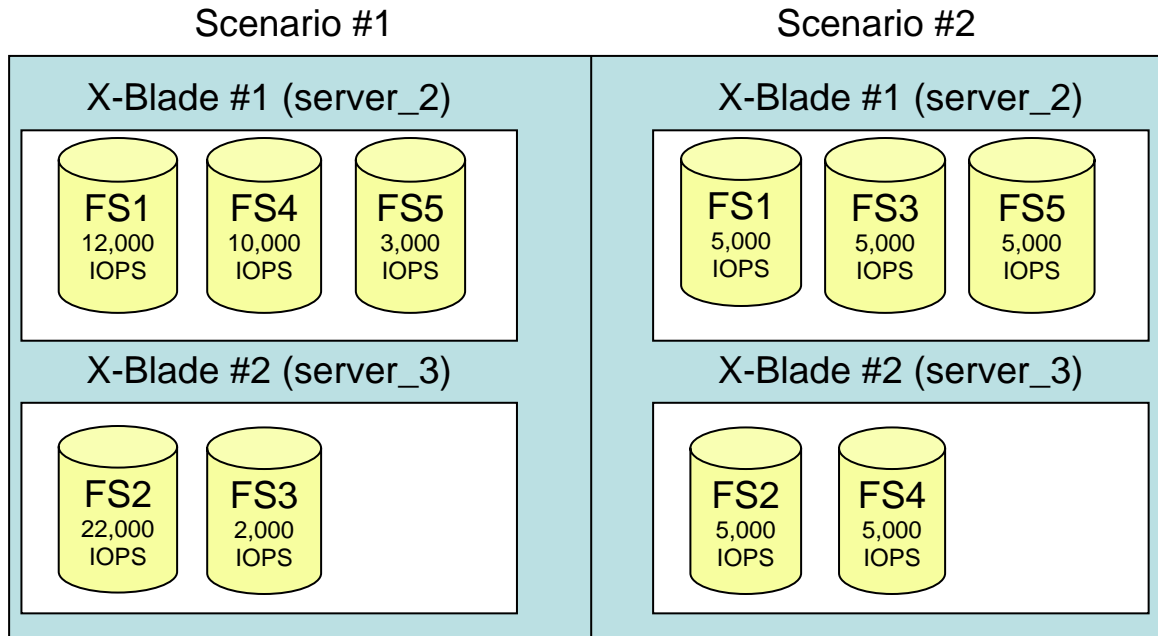
	<b>Limit</b>
<b>Maximum file system size</b>	16 TB
<b>Maximum number of FS per X-Blade (including VDMs and checkpoints)</b>	2048
<b>Maximum number of FS per Celerra (including VDMs and checkpoints)</b>	4096

The concept of “right-sizing” is that you plan for acceptable performance levels within your IP storage and size the file system so that file system extensions are needed less frequently. This technique maximizes the benefits of using automatic file system extension by minimizing the frequency at which the extensions are triggered. If a proposed use case requires 100 TB of storage capacity over the next five years by using 300 GB drives, and the growth rate is expected to be 25 percent annually, then we should size the initial file system large enough to store at least 25 percent of the forecasted capacity requirement initially. When an automated file system extension is triggered in this scenario, the resulting extension size is 10 percent of the current size, which increases the file system capacity by 2.5 TB (Initial File System Size: 100 TB \* 25% = 25 TB; Automated Extension Size: 25 \* 10% = 2.5 TB). We can utilize a large portion of each RAID group due to the large extension size (AVM attempts to use four RAID groups for each file system/extension). This limits the amount of available space left for other file systems on the RAID group, which minimizes spindle contingency within the solution. This approach balances the amount of storage that you must purchase initially, the risk of changing forecasts/growth rates, and the performance. The initial size of the file system is not influenced by the growth rate, but by the size of the extension of each file system (each extension should be large enough to utilize 35 percent to 50 percent of each RAID group). You must use this approach for sizing file systems on an individual basis rather than setting a pre-determined, generic, “minimum” size for each new file system. If the data hosted on the new file system is being migrated from another server, then the new file system should be large enough to handle the initial data migration without requiring a file system extension. Review the size and utilization of the current source file system during the migration-planning phase and use this information for sizing the initial target file system.

You should also attempt to balance file systems across X-Blades to maximize the utilization of the physical resources available within the system. This gives rise to two scenarios:

- The file systems that are expected to generate the majority of the I/O requests should be mounted on separate X-Blades if possible.
- If all file systems are expected to have a similar workload, then dividing the total number of file systems across both X-Blades should suffice.

As mentioned previously, when planning for performance-based solutions, ensure that the following three parameters are addressed: drive count (the number of spindles available to a file system can directly affect its performance), proper distribution of data across RAID groups (to minimize hot spots, file systems should be created across LUNs with similar characteristics and storage capacity), and capacity. Always address the performance requirements first and the capacity requirements later.



**Figure 1. Scenarios for balancing file systems**

Large file systems are generally used either for environments that are not performance sensitive or for environments that illustrate characteristics of having a large, sequential I/O based workload. The combination of multiple sequential I/O streams to the same RAID group displays the same performance characteristics as a random workload because each sequential request interrupts another. Therefore, as the number of sequential I/Os to the same RAID group increases, the data pattern on the storage array tends to display characteristics that are more common for a random workload.

### ***Planning phase tip: Expected IOPS calculation***

The following method can be used to determine the theoretical number of expected IOPS for a given file system:

#### **Front-end IOPS:**

$$\frac{\text{Number of Clients} * \text{Number of Processes} * \text{Number of Threads per Process} * \text{Average Application I/O Size (in KB)}}{\text{Protocol Transfer Size:}}$$

CIFS: 56K for writes and 16K for reads | NFS: Use the client read/write buffer size

#### **Storage array IOPS:**

$$\frac{\text{Front-End IOPS}}{8K \text{ (the predominant I/O size to storage array)}}$$

## **SnapSure checkpoints**

Celerra version 5.6 supports 96 NAS checkpoints per primary file system. The number of checkpoints counts against the total number of file systems allowed per X-Blade/Celerra, and against the allocated capacity per X-Blade. The total available capacity per X-Blade is 128 TB and the Checkpoint Save Volumes that are associated with the mounted checkpoints on an X-Blade count against this limit. Figure 2

illustrates that while the amount of storage used by primary file systems increases, the storage remaining for Checkpoint Save Volume decreases.

The key driver for SnapSure Checkpoint Save Volume capacity utilization is the daily rate of change on the primary file system that is being protected. A best practice is to plan for approximately 10 percent of the primary file system size being required for the Checkpoint Save Volume. You must have at least 10 GB of available capacity to create the first checkpoint for a primary file system. The number of checkpoints for a primary file system also impacts SnapSure Checkpoint Save Volume utilization. The greater the number of checkpoints is, the greater the probability of a changed block having to be stored in the SnapSure Checkpoint Save Volume.

The capability to create writeable checkpoints was introduced in Celerra version 5.6 and the impact of this new functionality on SnapSure Checkpoint Save Volume utilization should be taken into account. When data is being written or deleted from a writeable checkpoint, the original data is stored in the SnapSure Checkpoint Save Volume. Therefore, a large number of changes to files can dramatically increase the amount of storage that is required for checkpoints.

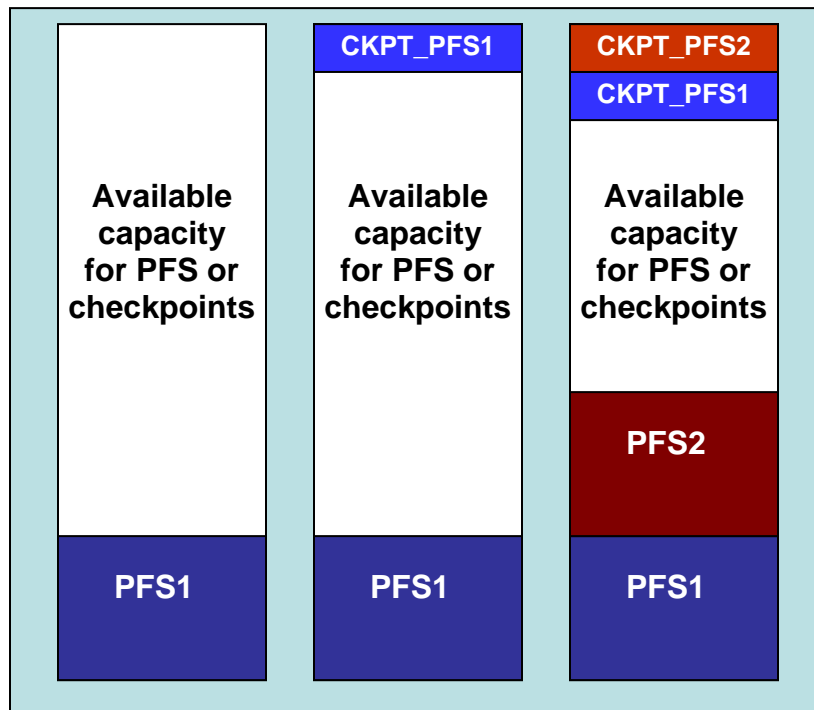


Figure 2. Comparison of storage used by primary file systems and storage available for the Checkpoint Save Volume

## Celerra Replicator

Table 5. Current limits for Celerra Replicator (V2) (as of February 2009)

	Limit
Maximum number of configured replication sessions per X-Blade	1024
Maximum number of replication sessions actively in initial transfer per X-Blade	16
Maximum number of replication sessions actively in transfer per X-Blade	256

The same considerations that apply to the SnapSure checkpoint feature also apply to Celerra Replicator because Celerra Replicator utilizes internal checkpoints for tracking changes and delta-sets between the

---

source and destination file systems. Celerra Replicator V2 creates two internal checkpoints per replication session for system use. The total number of file systems, VDMs, and checkpoints mounted on a single X-Blade cannot exceed 2,038 and these internal checkpoints count against this limit.

The Celerra Replicator feature requires the creation of interconnects, which are specified by either IP address or interface names, between the source and target X-Blades. When the capacity of an I/O load on the system increases, interconnect network utilization should be monitored closely by using either the `server_stats` command from CLI or Celerra Monitor from Celerra Manager Advanced Edition user interfaces. This must be done to ensure that throughput requirements for new shares and exports can be sufficiently handled while Celerra Replicator is active.

## Backup and restore considerations

The ability to utilize 128 TB per X-Blade affects the backup and restore functionality and the planning tasks associated with it most heavily. This section helps you identify the amount of storage you can utilize per X-Blade to ensure that you can meet the backup and restore time expectations. This is because the amount of file systems that were mounted on two X-Blades, before Celerra 5.6, to achieve the desired capacity objectives can now be mounted on a single X-Blade. It is important to note that there are several backup/restore options available and that each option provides different benefits. Although this section focuses primarily on NDMP backup/restore, some of the other backup/restore options are:

- **SnapSure checkpoints:** SnapSure checkpoints can be utilized to recover from logical errors by using an online point-in-time copy of the data set.
- **Celerra Replicator:** You can use this feature to create a secondary copy of the file system on either the local or remote EMC Celerra system. This method utilizes IP communication to transfer periodic updates to the secondary file system after an initial baseline copy. Celerra Replicator supports multi-protocol file systems.
- **Network-based backup:** This method does not have an explicit limit on the number of active backup/restore jobs. However, it supports only single-protocol file systems and provides very slow backup/restore.
- **NDMP:** This option provides separate communication paths for the data payload and control data during a backup/restore session. There is an active limit of four concurrent sessions with this backup/restore method and it supports multi-protocol file systems.
- **MPFS:** This option provides benefits similar to NDMP in terms of the data path. However, its limitations are similar to network-based backups (single protocol). There is no explicit limit on the number of active backup/restore jobs.

Celerra allows four active NDMP backup/restore sessions concurrently on a single X-Blade. Therefore, it is recommended that at least four tape drives be allocated per X-Blade to achieve maximum session count. Due to the large capacity of usable storage, a full backup may not be feasible regularly. Therefore, you should prefer vendors that support the “incremental forever” backup type. The initial backup may be equivalent to a full backup, but after the initial backup, back up only files that have changed. A read-only copy of the file system can be created by utilizing the automated checkpoint capability that was introduced with Celerra version 5.5. This feature allows supported backup applications to create a checkpoint of the file system that is being backed up automatically. Although you should create file systems as large as possible for performance gains, sizing a file system too large can affect backup and restore times. Usually, a restore operation takes longer to complete than a backup operation. The average file size and quantity of files also affect backup performance. Usually, a file system comprised of files lesser than 8K in size takes longer to back up because more metadata calls have to be made and transmitted to the backup server during the course of the backup operation.

You can use several techniques to ensure that the backup and restore operations are conducted as efficiently as deemed by the solution requirements:

- Use NDMP file filtering to exclude unnecessary files from being backed up. This reduces the backup size.

- Use Volume Based Backup (VBB) when backing up a file system with predominantly small files. The speed of conducting a destructive restore operation for such a file system is significantly faster than a traditional tar- or dump-based restore.
- Consider using faster backup mediums, for example, backup to disk, if the source file system is large (over 2 TB).
- Celerra Replicator V2 can also be used to provide a backup copy of the current file system. This approach to file system backup may incur higher setup costs when compared to tape, but it provides additional benefits such as faster restores and online single file restores.

Disk read performance most frequently dominates backup performance. Because most file systems are much bigger than the memory caches provided by the Celerra Data Movers and the attached storage arrays, only a small percentage of the file system can be cache resident. Consequently, the read performance is limited by the performance of the physical disk spindles. Optimal backup/restore performance requires that each concurrent backup/restore operation targets file systems that reside on separate RAID groups/disks to achieve maximum throughput.

## Planning and design workflow

This section provides a workflow that includes the topics that must be addressed for a successful NAS capacity allocation plan. It is recommended that you read the specific section in detail when preparing to deploy NAS in a production environment.

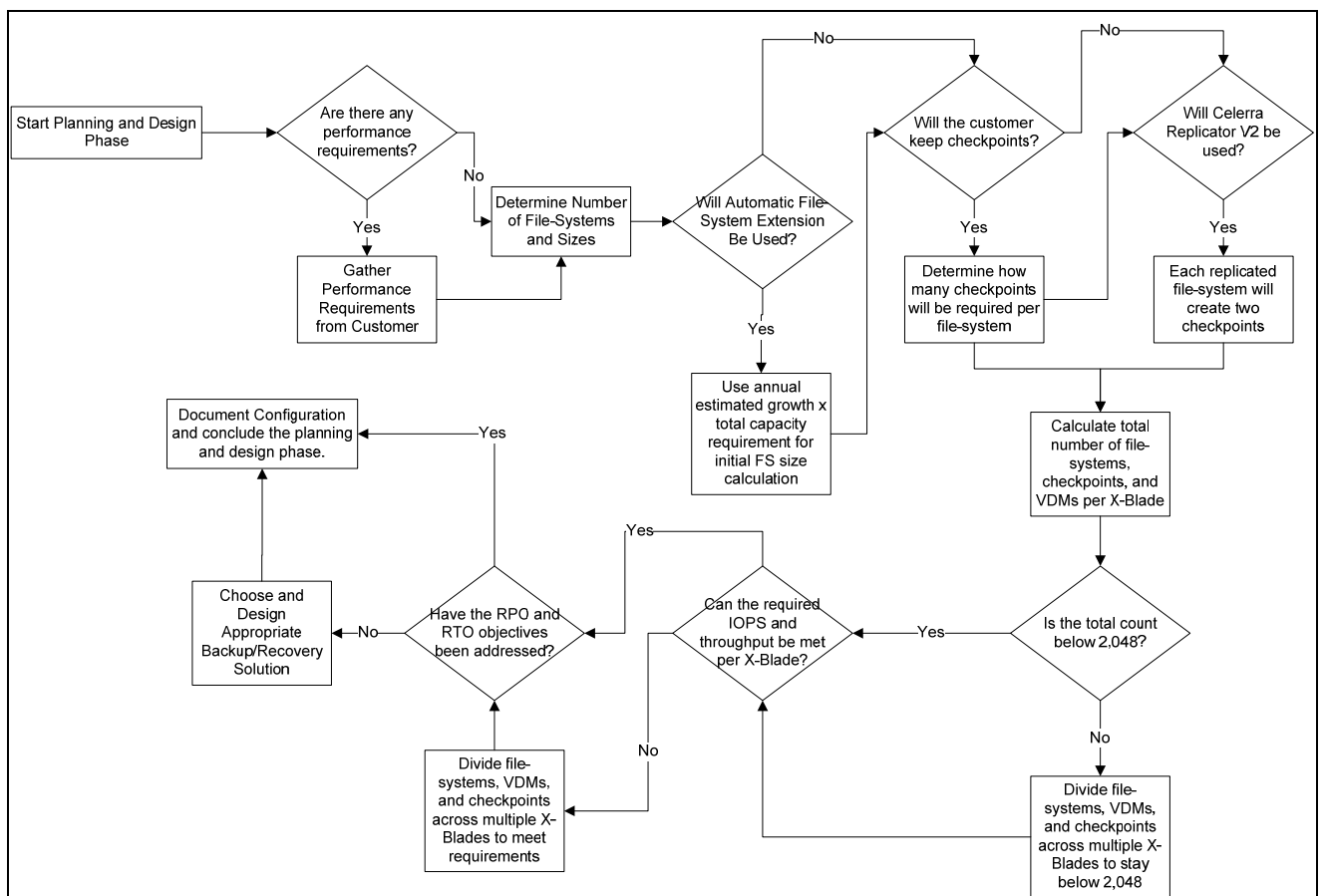


Figure 3. Workflow for a successful NAS capacity allocation plan

---

## Conclusion

The underlying limits of Celerra 5.6 such as file system size, number of checkpoints per file system, and number of file systems have not been modified with the introduction of 128 TB capacity support. However, the probability of encountering these limits in a solution has increased because solutions that once required two X-Blades can now utilize one. When planning a NAS deployment, performance requirements should drive the design and capacity requirements should be accounted for as well. This approach allows rational decisions on capacity utilization for your specific environment.

## References

Name: *Celerra Network Server 5.5 Best Practices for Performance - Best Practices Planning*

Type: White paper

URL: <http://powerlink.emc.com>

Audience: Field pre-sales (restricted to EMC employees only)

Technical Depth: Medium

Name: *E-Lab Interoperability Navigator*

Type: Web-based application

URL: <http://powerlink.emc.com>

Audience: Field pre-sales, technical customer

Technical Depth: Medium

Name: *NAS Support Matrix*

Type: Technical publication

URL: <http://powerlink.emc.com>

Audience: Field pre-sales, technical customer

Technical Depth: Medium