1. [**Cross-Platform Remote File Systems**](http://syrcose.ispras.ru/2020/submissions/SYRCoSE_2020_paper_26_14.pdf)**:**

The ability to access files from anywhere, regardless of where they are kept, is made possible by remote file systems, which are a crucial component of modern computing. Sharing files between platforms, however, can be challenging due to the use of several file systems by different operating systems. By allowing users of various operating systems to access the same file system, cross-platform remote file systems help solve this problem.

According to the article, cross-platform remote file systems operate by translating file system requests across various operating systems. For instance, a request for a file in the Windows-specific NTFS storage system can be converted into a request for an identical file in the Linux-specific EXT4 file system. As a result, file sharing between various platforms is made possible.

The ease with which files can be shared between various operating systems is one of the key benefits of cross-platform remote file systems. Users may access files from any device, which can be very helpful in businesses that employ a variety of operating systems. Additionally, by enabling users to access files via secure connections, cross-platform remote file systems can enhance security.

The essay also covers some of the difficulties in putting cross-platform remote file systems into practice. It may be challenging to deliver consistent performance across platforms, for instance, if different file systems have different performance traits. Additionally, it's possible that some file systems won't support specific features like file permissions, which can make security maintenance challenging.

Overall, the paper offers a thorough analysis of cross-platform remote file systems, including advantages and drawbacks. Cross-platform remote file systems are undoubtedly a crucial piece of technology for businesses that employ a variety of operating systems since they make file sharing easier and boost security. It is also obvious that putting in place cross-platform remote file systems calls for thorough consideration of the various file systems involved as well as the organization's performance and security needs.

In the current digital era, cross-platform remote file systems have grown in significance due to the development of several operating systems and the demand for seamless file sharing and collaboration between them. Samba, NFS, and WebDAV are a few of the most widely used cross-platform remote file systems.

Samba is a well-liked open-source version of the SMB/CIFS protocol that enables file and printer sharing between Windows and Linux devices. Contrarily, NFS, a popular network file system, allows for file sharing across Unix-based systems. Users can read and edit files online using the WebDAV protocol, which stands for Web Distributed Authoring and Versioning.

Assuring compatibility between various file systems is one of the difficulties in establishing cross-platform remote file systems. For instance, features like support for file permissions and encryption are not present in Linux's EXT4 file system, but NTFS, the file system used by Windows, offers these characteristics. NTFS does not provide features like extended attributes or support for symbolic links, but EXT4 does. Cross-platform remote file systems must therefore be able to convert these various file system capabilities into an operating system-compatible format.

Performance presents another obstacle. Performance aspects including read and write rates, file size restrictions, and support for concurrent access may vary amongst different file systems. To guarantee that users can access their files consistently and reliably regardless of the platform they are using, cross-platform remote file systems must be able to handle these performance variations.

The final issue with cross-platform remote file systems is security. Sharing files online exposes sensitive information to potential security threats like viruses, hacking, and unauthorized access. Cross-platform remote file systems often employ encryption and authentication protocols to make sure that data is delivered securely and that only authorized users have access to it, thereby reducing these dangers.

In conclusion, cross-platform remote file systems are a crucial piece of technology for contemporary businesses that depend on cross-platform information sharing and collaboration. Although they present some implementation difficulties, the advantages they provide in terms of performance, security, and simplicity make them a useful tool for any organization that needs to share files across various platforms.

1. [Data Integrity Checking (HADOOP)](https://www.jcscm.net/fp/115.pdf):

In big data contexts where the amount of data can be overwhelming, data integrity is a crucial component of data management and storage. A well-liked method for storing and handling massive volumes of data in a distributed setting is the Hadoop Distributed File System (HDFS). This article describes Hadoop's data integrity checking, a crucial component for ensuring that data is neither lost or corrupted.

Data retrieval and storage integrity are both checked as part of the data integrity checking process in Hadoop. Hadoop employs a number of methods, such as checksums, replication, and block verification, to guarantee data integrity.

A mathematical procedure known as a checksum is used to confirm the accuracy of data as it is being stored and retrieved. Checksums are used by Hadoop to identify any modifications to the data that might have been made during transmission or storage. A file is split into blocks and given a checksum value before being saved in Hadoop. Hadoop computes the checksum value for each block in the file and compares it to the original checksum value to make sure the data has not been corrupted when the file is later retrieved.

Hadoop also employs replication to guarantee data integrity. A file that is saved in Hadoop is duplicated across several data nodes to make sure it is always accessible in the event of a node failure. By default, Hadoop replicates data three times, so each block is present on three separate data nodes. This guarantees that access to and usage of the data are not affected even if one node fails.

Thirdly, Hadoop uses block verification to guarantee data integrity. Hadoop periodically checks each block stored in the system to make sure it hasn't been corrupted or lost, and it ensures that it is still intact. Hadoop can automatically recover the data by replicating it if any damage or loss is found.

Data integrity checking in Hadoop has many advantages, one of which is that it makes sure that data is always accurate and available, despite hardware or network problems. This is especially crucial in big data contexts because losing data can have negative effects. Data integrity checking can further enhance Hadoop's speed by lowering the risk of data corruption and guaranteeing that data is always accessible when needed.

It is important to note that Hadoop offers a number of choices for customizing the checksum technique used to test data integrity in order to further elaborate on data integrity checking in Hadoop. The CRC32 checksum technique, which is quick and effective at finding data problems, is what Hadoop employs by default. Although they have a higher computational overhead, other algorithms supported by Hadoop include the Adler-32 and MD5 checksum algorithms, which have a stronger error-detecting capacity.

Block verification is a crucial step in Hadoop's procedure for verifying data integrity. Every block that is stored in the system is routinely inspected for integrity by Hadoop, which may automatically recover any lost or corrupted data by replicating it from a different node or by utilizing checksum values to locate and fix mistakes. Block verification can, however, be a computationally demanding procedure, especially in massive Hadoop clusters with numerous data nodes. Hadoop offers a variety of configuration options for adjusting the frequency and severity of block verification in order to minimize the negative effects on system performance.

Data scrubbing, a tool offered by Hadoop in addition to these mechanisms, is intended to proactively identify and fix data errors in the system. In order to find any discrepancies or errors, data cleaning periodically scans the data kept in Hadoop and compares it to the checksum values. Data scrubbing can automatically fix any problems by reproducing the offending blocks from other cluster nodes if any are found.

Overall, any distributed file system must include a vital component for data integrity checking, and Hadoop offers strong mechanisms for guaranteeing the reliability and availability of data in large data situations. Checksums, replication, block verification, and data scrubbing are some of the techniques used by Hadoop to identify and fix data mistakes, guarantee that data is always accessible when required, and boost system performance by lowering the probability of data corruption. The demand for efficient data integrity verification in Hadoop and other distributed file systems will only increase as the amount and complexity of big data continue to rise.

**Part 2: Answer the following questions:**

1. (10 Points) Differentiate between HDD and SSD devices in terms of speed/performance, form factors, capacity. 1-2 pages, double-spaced.

Ans: Solid-state drives (SSDs) and hard disk drives (HDDs) are two categories of storage devices that vary in terms of capacity, performance, form factors, and underlying technology.

The performance and speed of HDDs and SSDs are two of their main distinctions. Since HDDs read and write data on spinning disks, the speed at which the disks can spin determines how fast HDDs can operate. In comparison, SSDs are more faster and more dependable than HDDs because they don't have any moving components and instead use flash memory to store and retrieve data. HDDs typically have a seek time of several milliseconds, which can cause slower performance and longer load times, but SSDs can access data nearly instantly.

HDDs are often heavier and bulkier than SSDs in terms of form factor, which makes them less portable and unsuitable for use in laptops, tablets, and other mobile devices. As opposed to SSDs, HDDs can be found in larger capacities, with some models offering up to 16 terabytes of storage. But larger SSD capacities are also becoming more widely available, with some models providing up to 4 terabytes of storage.

The dependability and durability of HDDs and SSDs should also be taken into account when comparing them. HDDs use spinning disks, making them more prone to mechanical failure and data loss from shock or vibration. SSDs, on the other side, don't have any moving parts, so they're more durable and less likely to encounter problems.

The price of storage per unit of storage is one thing to take into account when contrasting HDDs vs SSDs. HDDs have traditionally cost less than SSDs, making them the preferred option for those that require high storage capacities at a reduced price. However, as the price of SSDs has dropped recently, they have grown in popularity as a more affordable option to HDDs for specific applications.

The durability and life expectancy of HDDs and SSDs are additional factors to take into account. SSDs can only handle a certain number of write cycles before they start to degrade, which over time may lead to data loss or decreased performance. The development of SSDs with longer lifespans and increased endurance, however, as a result of advancements in flash memory technology, makes them appropriate for usage in high-traffic areas and mission-critical applications.

System performance and user experience can be significantly impacted by the speed and performance differences between HDDs and SSDs. For instance, SSDs can speed up computer startup, application launch, and file loading, which can increase output and efficiency. In high-performance computing applications like gaming, video editing, and 3D rendering, where quick data access and processing are essential, SSDs are also perfect.

HDDs have a cost-effective way of storing massive volumes of data, which gives them an edge over SSDs. HDDs are the best choice for customers that need to store huge multimedia assets, such as images, videos, and music, as they come in capacities up to 16 terabytes. However, SSDs' appeal to users who require a lot of storage space may be constrained by the fact that they are still relatively expensive at higher storage capacities.

The decision between an HDD and an SSD is ultimately influenced by a number of variables, including performance, cost, durability, and storage capacity. SSDs are the best option for high-performance computer tasks like gaming and video editing because of their quick read and write speeds. HDDs continue to be a popular option for consumers that need to store huge volumes of data at a reasonable cost. Both HDDs and SSDs are likely to keep getting faster, more affordable, and more capable as technology advances, giving consumers a variety of options to fulfill their storage demands.

1. Differentiate between NAS and SAN and how they implement RAID. 1-2 pages, double-spaced.

Ans:

For storing and managing massive volumes of data, two popular technologies are Storage Area Networks (SAN) and Network Attached Storage (NAS). While they both serve the same purpose of giving users access to shared storage resources, their architecture, connectivity, and RAID implementation are different.

A network-based storage system called NAS offers a centralized location for hosting and distributing data files. Users can access the contents stored on NAS devices using common file-sharing protocols like SMB, NFS, or FTP. NAS systems are frequently Ethernet-connected to the network. NAS devices are often simple to set up and administer, and they can be used as standalone units or as part of a broader network architecture.

On the other side, SAN is a block-level storage solution that offers dedicated networks high-speed access to shared storage resources. Enterprise environments where significant amounts of data need to be accessed rapidly and reliably frequently employ SANs. Block-level access to storage devices is made possible by SANs using specific storage protocols like Fibre Channel or iSCSI. SANs can be difficult to set up and maintain, and they frequently need for specialized training and experience.

Both NAS and SAN systems can apply RAID in a number of different ways to increase system stability and provide data redundancy. In order to increase speed and reduce costs, a technology called RAID (Redundant Array of Inexpensive Disks) integrates many hard drives into a single logical unit.

Depending on the level of performance and data redundancy necessary, NAS devices can use RAID 0, RAID 1, RAID 5, RAID 6, or RAID 10. While RAID 1 offers data redundancy by mirroring data across multiple disks, RAID 0 is used to increase performance by striping data across multiple disks. By striping data and parity information across numerous disks, RAID 5 and RAID 6 offer a compromise between performance and data redundancy. By striping data over mirrored pairs of drives, RAID 10, commonly referred to as RAID 1+0, combines the performance advantages of RAID 0 with the data redundancy advantages of RAID 1.

The same RAID levels that are used by NAS devices can also be used by SANs, but these often employ more sophisticated RAID configurations, like RAID 50 or RAID 60, which offer a better level of data redundancy and performance. RAID 50 and RAID 60 offer high levels of performance and data redundancy by combining multiple RAID 5 or RAID 6 arrays into a single logical unit.

There are a few significant aspects to take into account while implementing RAID in NAS and SAN systems. First, the organization's unique needs in terms of performance, data redundancy, and cost will determine which RAID level is selected. RAID 10 is frequently used in SANs due to its high degree of performance and data redundancy, while RAID 5 and RAID 6 are popular choices for NAS systems because they offer a decent balance between performance and data redundancy.

The quantity of disks needed to implement the selected RAID level is another thing to think about. The number of disks in the RAID array could need to be increased depending on the level of data redundancy and performance needed. For instance, RAID 5 requires a minimum of three disks, whereas RAID 6 demands a minimum of four drives. It could be required to deploy additional disks in a SAN environment when high performance and high reliability are crucial in order to achieve the needed degree of performance and data redundancy.

In conclusion, the two storage systems known as NAS and SAN are utilized to give users access to shared storage resources. SANs often use more complicated RAID configurations than NAS devices due to their high-performance and high-reliability needs. Both can implement a range of RAID configurations to offer data redundancy and enhance system reliability.

1. (28 Points) 2 Points Each - Define the following terms in 2-3 sentences from the eBook.
   1. Connecting External HDD using eSATA vs USB
   2. Current costs per MB of flash/SSD, SRAM, DRAM, and HDD
   3. Data Integrity: What is CRC? What is ECC? Checksum? What is Parity for storage in RAID?

Ans:

a.

eSATA (External Serial Advanced Technology Attachment), a high-speed data transmission technology, offers faster data transfer rates than USB when connecting an external hard drive. For huge file transfers or for programs that need a lot of bandwidth, such video editing, eSATA is a better choice. On the other hand, USB is a more generally accepted standard that is appropriate for daily usage and offers compatibility with the majority of devices.

b.

For flash/SSD, SRAM, DRAM, and HDD, the current price per megabyte is: Depending on elements like manufacturing volume, market demand, and technological improvements, the cost per MB (megabyte) of various memory/storage systems can vary significantly. By 2021, the price per MB of flash/SSD (solid-state drive) will often be lower than that of SRAM and DRAM, but more than that of HDD (hard disk drive).

c.

Data accuracy and consistency over the course of a piece of data's lifecycle are referred to as data integrity. Data integrity is checked using the CRC (Cyclic Redundancy Check) error detection technique by comparing a calculated checksum with a stored checksum. The error correction technique known as ECC (Error Correction Code) employs additional data to find and fix faults in stored data. A checksum is a number that is computed from the data to identify any data changes. In RAID systems, parity is a form of redundancy that offers fault tolerance by enabling the reconstruction of lost data in the event of a disk failure.