INFORMATION STORAGE AND MANAGEMENT

Hassan I. Eissa

HASSAN3ISSA@GMAIL.COM

http://fb.com/hassan3issa

Contents

❖ Part I Storage System

- Chapter 1. Introduction to Information Storage and Management
  - Information Storage
  - Evolution of Storage Architecture
  - Data Center Infrastructure
  - Virtualization and Cloud Computing
- Chapter 2. Data Center Environment
  - Application
  - Database Management System (DBMS)Host (Compute)
  - Connectivity
  - Storage
  - Disk Drive Components
  - Disk Drive Performance
  - Host Access to Data
  - Direct-Attached Storage
  - Storage Design Based on Application Requirements and Disk Performance
  - Disk Native Command Queuing
  - Introduction to Flash Drives
  - Concept in Practice: VMware ESXi
- Chapter 3. Data Protection: RAID
  - RAID Implementation Methods
  - RAID Array Components
  - RAID Techniques
  - RAID Levels
  - RAID Impact on Disk Performance
- Chapter 4. Intelligent Storage Systems
  - Components of an Intelligent Storage System
  - Storage Provisioning
  - Types of Intelligent Storage Systems
  - Concepts in Practice: EMC Symmetrix and VNX

❖ Part II Storage Networking Technoligies

Chapter 5: Fibre Channel Storage Area Networks

- Fibre Channel: Overview
- The SAN and Its Evolution
- Components of FC SAN
- FC Connectivity
- Switched Fabric Ports
- Fibre Channel Architecture
- Fabric Services
- Switched Fabric Login Types
- Zoning
- FC SAN Topologies
- Virtualization in SAN
- **Error! Reference source not found.**

Chapter 6. IP SAN and FCoE

- iSCSI
- FCIP
- FCoE

Chapter 7. Network-Attached Storage

- Benefits of NAS
- File Systems and Network File Sharing
- Components of NAS
- NAS I/O Operation
- NAS Implementations
- NAS File-Sharing Protocols
- Factors Affecting NAS Performance
- File-Level Virtualization

Chapter 8. Object-Based and Unified Storage

- Object-Based Storage Devices
- Content-Addressed Storage
- CAS Use Cases
- Unified Storage

Part I

Storage System

Chapter 1. Introduction to Information Storage and Management

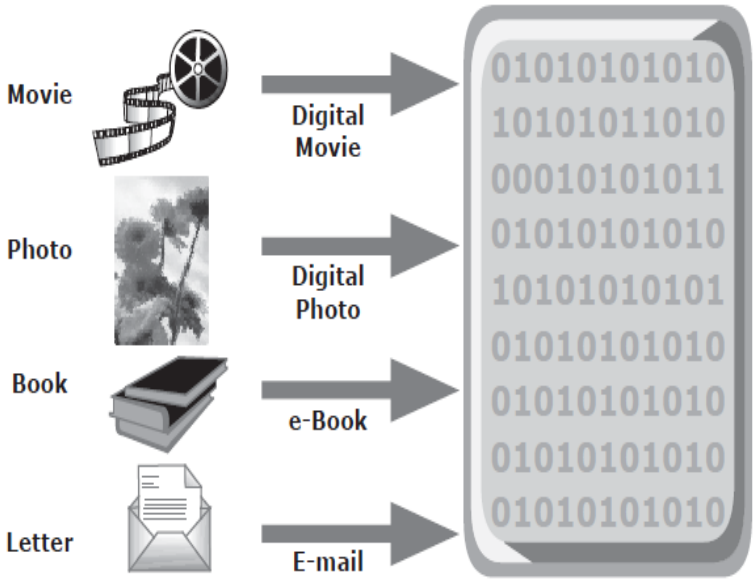
Information Storage

هو النظام الذى يحافظ على الوصول إلى البيانات لمعالجة المعلومات. او بعبارة أخرى وحدة تخزين المعلومات في النظام إما على محرك أقراص ثابت أو خادم يحتوي عادة على قاعدة بيانات.

- Data

تعريف البيانات هي مجموعة من الحروف أوالكلمات أو الأرقام أو الرموز أو الصور المتعلقة بموضوع معين مثال على ذلك: بيانات الموظفين (الأسماء - الأرقام الوظيفية - المهن - الصور) بدون ترتيب ، يتم معالجتها ببرامج حاسوبية خاصة لكي نحصل على معلومات (Information).

هذه البيانات يمكن انشائها باستخدام جهاز كمبيوتر وتخزينها على هيئة مجموعه من الارقام الثنائية (Binary) وتسمى بال Digital Data, يمكن استحدامها بعد ذلك من قبل المستخدمين بعد ان يقوم الكمبيوتر بمعالجتها.

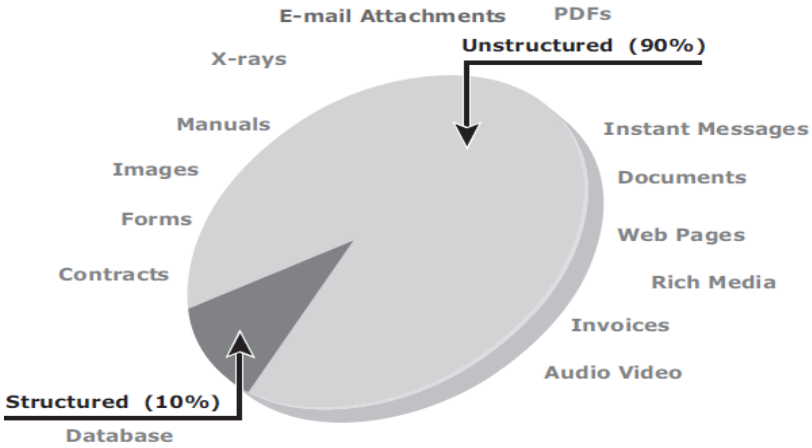


- Types of Data

يمكن تصنيف الداتا الى بيانات مهيكله وغير مهيكله Structured و unstructured بناءا على كيفية تخزينها واداراتها.

حيث انه يتم تنظيم ال Structured Data على هيئة صفوف واعمدة في جداول او قواعد بيانات بشكل محدد مما يسهل على اي تطبيق معالجتها بسهولة ويتم تخزينها عن طريق مايسمى بال database management system (DBMS).

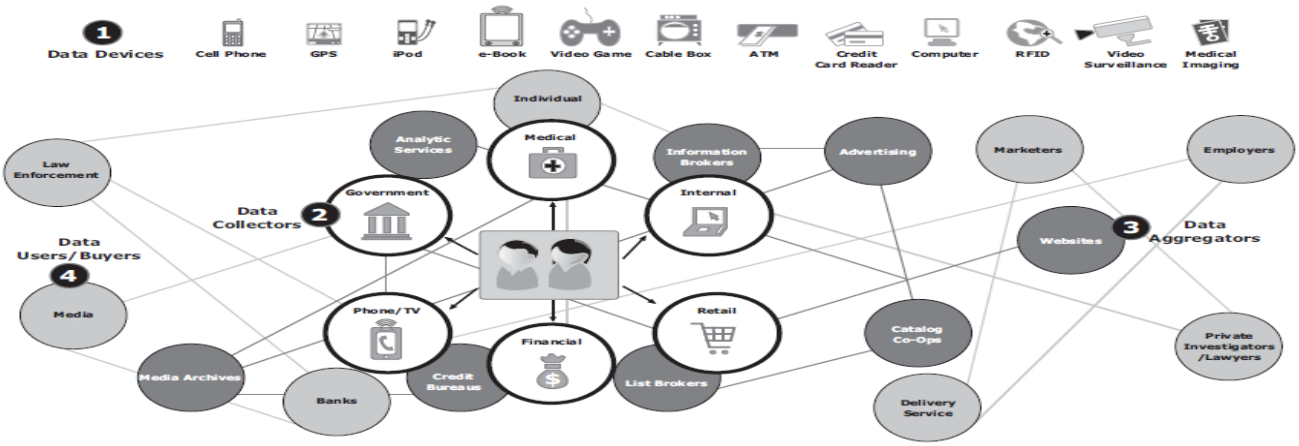
اما ال Unstructured Data فهي تشكل النسبة الاكبر للبيانات ولا يمكن تخزينها فى صفوف واعمدة مما يصعب على التطبيقات معالجتها بسهولة الا من خلال تطبيقات محددة, على سبيل المثال البيانات التى يتم تخزينها من المستخدمين على هيئة sticky notes, e-mail messages, business cards, or even digital format files, such as .doc, .txt, and .pdf



Big Data -

مجموعة من البيانات الكبيرة جدا بحجم يفوق قدرة أدوات قواعد البيانات التقليدية من تخزين، إدارة و تحليل تلك البيانات فى فترة زمنية مقبولة.

وتتضمن البيانات المهيكلة والغير مهيكلة (structured and unstructured) التى تنشأ عن طريق مجموعه متنوعه من المصادر والتطبيقات المختلفة مثل business application transactions, web pages, videos, images, e-mails, social media



- **Information**

مفهوم المعلومات هي نتائج معالجة البيانات، فالمعلومات عبارة عن البيانات التي تمت معالجتها بتصنيفها وتنظيمها وتحليلها، وأصبح لها معنى لتحقق هدف معين وتستعمل لغرض معين حتى توفر ما يسمى بالمعرفة.

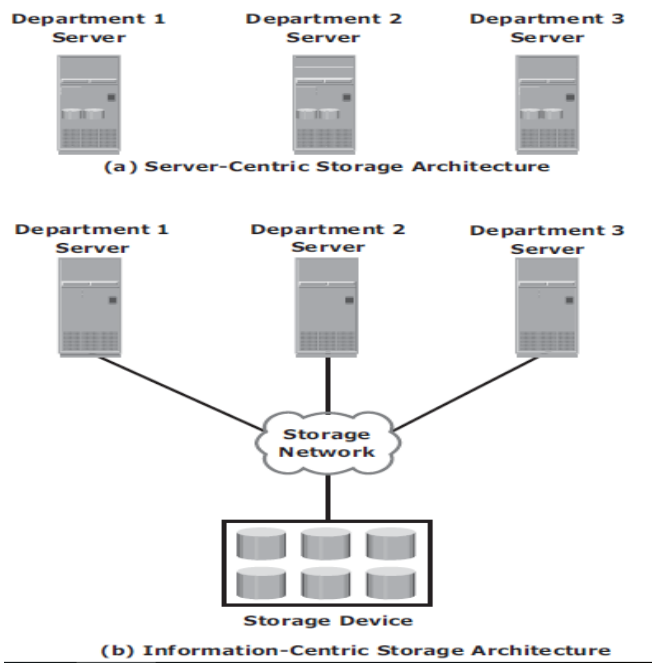
- **Storage**

يجب تخزين البيانات والمعلومات المنشأة من قبل الأفراد او الشركات بحيث يمكن سهولة الوصول اليها ومعالجتها. وتسمى الاجهزة الخاصة بتخزين البيانات بال Storage وتختلف انواع الStorage على حسب اختلاف انواع الداتا والطريق المنشأة بها مثل كروت الذاكرة للموبيل والكاميرات او الDVD لمشغلات الاقراص او الDisk Drive فى اجهزة الكمبيوتر والسيرفرات.

Evolution of Storage Architecture

فى البداية كان يتم استخدام اجهزة كمبيوتر مركزية تسمى بال mainfarms بها اجهزة تخزين للبيانات tape reels and disk packs ولكن مع تطور الانظمه اصبح بمقدور الشركات ان يكون لها اجهزة التخزين والسيرفرات الخاصه بها ولكن عملية التخزين كانت تطبق داخليا على اجهزة السيرفر حيث لايمكن مشاركة البيانات مع اي اجهزة اخرى ويكون لكل سيرفر مساحه محدوده من التخزين وأي مهام إدارية، مثل صيانة الخادم أو زيادة سعة التخزين قد تؤدي إلى عدم توافر المعلومات وبالتالى كانت غير محمية او غير مدارة بالشكل الامثل

وللتغلب على هذه التحديات تطورت اجهزة التخزين حيث اصبحت تدار مركزيا وبشكل مستقل عن اجهزة السيرفر حيث يتم مشاركة اجهزة التخزين مع العديد من السيرفرات ويمكن زيادة مساحة التخزين ايضا دون التاثير على البيانات مما يجعل عملية ادارة المعلومات اكثر سهولة وبتكلفة اقل والتى تمكن من حماية وتحسين الاستفادة من المعلومات.



Data Center Infrastructure

تقوم الشركات بتوفير مراكزالبيانات داخل المؤسسة لتقديم قدرات مركزية لادارة و معالجة البيانات.

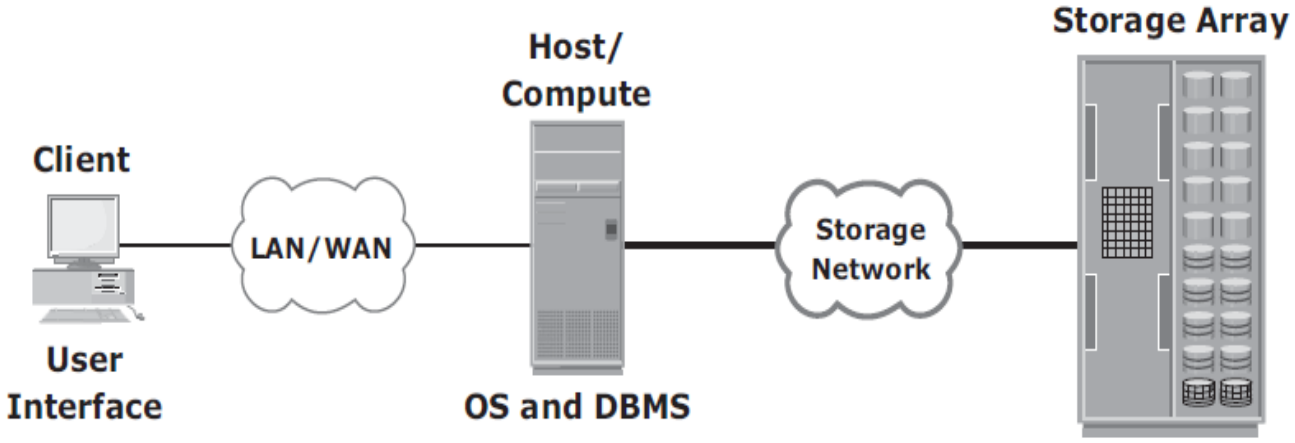
ومراكز البيانات هي عباره عن مركز ضخم مكون من العديد من الخوادم الضخمة ومزودات الطاقة مثل أجهزة الكمبيوتر وأنظمة التخزين وأجهزة الشبكة والنسخ الاحتياطي للطاقة, ومكونات البرمجيات، مثل التطبيقات وأنظمة التشغيل وبرامج الإدارة.

ومن الممكن ان تحتفظ المنظمات الكبيرة بأكثر من مركز بيانات واحد لتوزيع أعباء العمل على معالجة البيانات وتوفير النسخ الاحتياطي في حالة حدوث كارثة.

- Core Elements of a Data Center

كما ذكرنا العناصر الاساسية فى الداتا سنتر

- **Application**: تطبيقات الكمبيوتر لتقديم عمليات التشغيل المختلفة.
- **Database management system (DBMS)**: تقدم طريقة تخزين وتنظيم البيانات داخل قواعد البيانات.
- **Host or compute**: التى تقوم بتشغيل البرامج والتطبيقات المختلفة مثل (Hardware, Software, Firmware).
- **Network**: تحدد مسار البيانات داخل الداتا سنتر عن طريق الاجهزة المختلفة المتصلة بالشبكة.
- **Storage**: الاجهزة التى تقوم بتخزين البيانات داخلها.



- Modern IT Data Center Layers

يمكن تقسيم مراكز البيانات الحديثة التى تتناسب مع الCloud Computing الى 8 Layers منها 5 Logical Layers و 3 cross-layer functions.

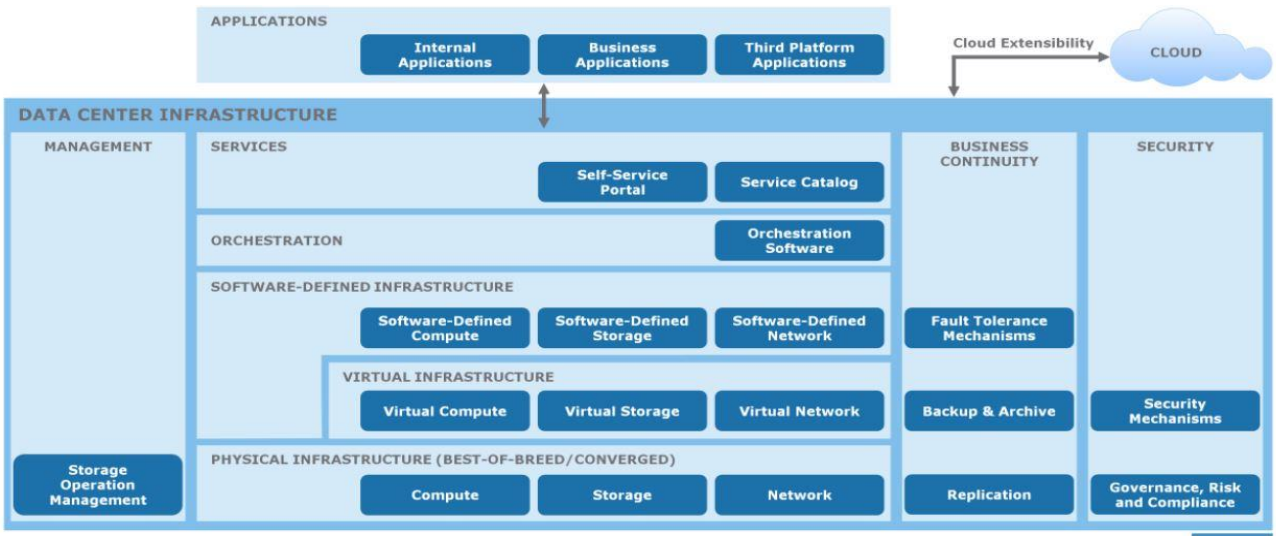
:Five Logical Layers

- **Physical infrastructure**: التى تحتوى على الResources مثل اجهزة الهاردوير والسيرفرات
- **Virtual infrastructure**: تقوم بعملية الabstracting لل Physical Resources وتحويلها الى Virtual Resources
- **Software-defined infrastructure**: تقوم بتقديم ال ITaaS (IT as Service) حيث يستطيع المستخدم من خلالها الوصول الى التطبيقات المختلفة.
- **Orchestration Layer**: تحتوي على Orchestration software يقوم بعمل تنسيق وادارة المهام للRequests المختلفة للمستخدمين من خلال عمل workflow.
- **Services Layer**: تشمل كل الخدمات والمعلومات الموجوده حول الIT Resources والتى تقوم بتقديمها الى الend users.

:Three cross-layer functions

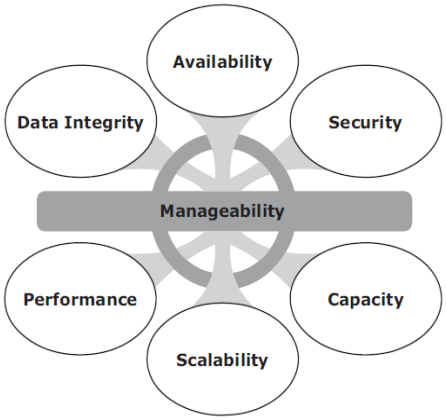
- **Business continuity Layer:** تمكنك من تخفيف أثر التوقف عن العمل (Down Time) بسبب الانقطاعات المخطط لها والغير مخطط لها من خلال عدة عوامل مثل
  - **Proactive Measure** متضمنه: business impact analysis, risk assessment, backup, archiving, and replication.
  - **Reactive Measures** متضمنه: disaster recovery and disaster restart.

- **Security Layer:** تدعم هذه الطبقة جميع الطبقات السابقة لتقديم خدمة آمنة للمستخدمين اعتمادا على administrative mechanisms and technical mechanisms
- **Management Layer:** تقوم بدعم جميع الطبقات لتنفيذ مهام ادارة ومراقبة وعمل التقارير المتعلقه بالداتا سنتر



- Key Characteristics of a Data Center

اهم الخصائص الرئيسية لمراكز البيانات:



- Availability
- Security
- Performance
- Data integrity
- Capacity
- Manageability

- Managing a Data Center

- Monitoring
- Reporting
- Provisioning

Virtualization and Cloud Computing

اولا يجب ان نتعرف على تكنولوجيا ال**Virtualization** وهي عملية انشاء نسخ وهمية Virtual من اجهزة حقيقية Physical, مثل Compute, Storage & Network تحويلها الى Logical Resources.

على سبيل المثال فان جهاز الكمبيوتر او السيرفر يتكون من عدة طبقات:

- Software Layer وهي السوفتوير والتى تحتوي على النظام التشغيلي OS والتطبيقات المختلفة
- Virtualization Layer وهي تفصل بين الاولى والثالثة وهنا تطبق التكنولوجية التخيليه عن طريق ما يسمى بالhypervisor
- Physical Layer وهي الهاردوير والتى تحتوي على ... CPU, Memory, Hard Desk,

بالنسبة للHypervisor هو نظام تشغيلي يكون وسيط بين الHardware و الOS وينشا منه Virtual machine.

اما ال**Cloud Computing** هو مصطلح يشير الى الخدمات او الموارد التى تقدم الى الشركة من مزود الخدمة عن طريق الانترنت بغرض الاستفادة من الامكانيات لديها, بمعنى انه يمكنك نقل الداتا سنتر الخاصه بك الى داتا سنتر موجوده عند مزودى خدمة الcloud وتستطيع الوصول اليها عن طريق سحابة الانترنت.

Chapter 2. Data Center Environment

الان اصبح وجود الداتا سنتر جزء اساسي فى اي شركة سواء كانت صغيرة, متوسطة او كبيرة.

ومن اهم عناصر الداتا سنتر الرئيسية كما ذكرنا هي ال host, storage, Network, applications, and DBM

هذه العناصر تعمل على معالجة وتخزين البيانات. وظهر تطور جديد للداتا سنتر مع ظهور الVirtualization اصبح قادر على تحويل الphysical resources الى Logical resources. هذا التطور ادى الى قلة استخدام الPhysical resources مما ادى الى تحسين الInfrastructure وتقليل التكلفة الاجمالية لامتلاكها.

Application

هو تطبيق يمكن الوصول اليه من خلال الكمبيوتر و السيرفر, يقوم ببعض العمليات الحاسوبية المحدده التى يحتاجها المستخدم حيث يقوم بعملية القراءة والكتابة للبيانات (Database) باستخدام خدمات نظام التشغيل لأداء عمليات R / W على أجهزة التخزين (Storage).

يمكن تصنيف التطبيقات الى business applications - infrastructure management applications - data protection applications - security applications.

ومنها على سبيل المثال Email, enterprise resource planning (ERP), decision support system (DSS), resource management, backup, authentication and antivirus applications وغيرها.

Note: Application virtualization breaks the dependency between the application and the underlying platform (OS and hardware). Application virtualization encapsulates the application and the required OS resources within a virtualized container.

Database Management System (DBMS)

نظام ادارة قواعد البيانات وهوعبارة عن مجموعة من البرامج التى تتحكم في تنظيم وتخزين وإدارة وسحب البيانات من قواعد البيانات. و يمكنها ادارة العديد من قواعد البيانات كما يمكن العديد من المستخدمين من الوصول إلى هذه القواعد في الوقت نفسه.

Host (Compute)

يقوم المستخدمين بتخزين واسترجاع البيانات من خلال التطبيقات المختلفة, وتعمل هذه التطبيقات على اجهزة الكمبيوتر التى تسمى بال Host ومن الممكن ان يكون الHost اما Physical او Logical. وامثلة على الPhysical Host مثل اجهزة الكمبيوتر والسيرفرات واجهزة الجوال والتى تحتوى على CPU, memory, I/O devices وتحتوى ايضا على مجموعة من الSoftware لتنفيذ عمليات الحوسبة مثل معالجة المدخلات المخرجات الى الHost

العناصر الاساسية لل (Compute) Host:

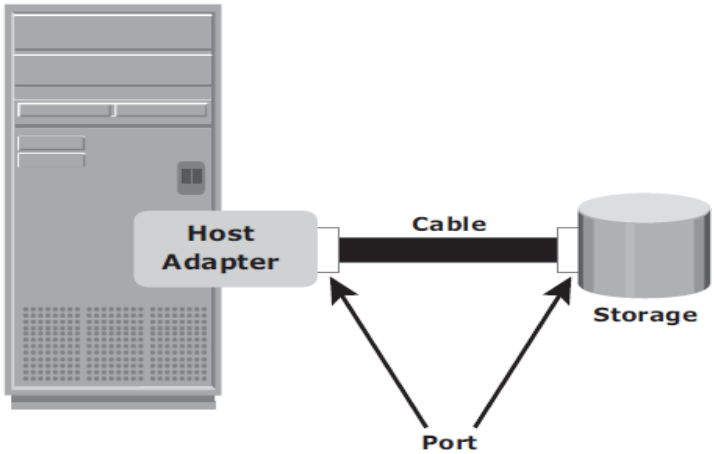
- Operating System
- Device Driver
- Volume Manger
- File System
- Compute Virtualization

Connectivity

تشير الى كيفية الربط بين الHost والاجهزة المختلفة مثل السيرفرات واجهزة الStorage وسنقوم بالتركيز فى هذا الجزء على التوصيل بال Storage Devices حيث يوجد هناك نوعين من الاتصال:

- Physical Components of Connectivity

المقصود بال Physical connectivity لل Host هو كروت الشبكة ويوجد منها انواع حيث كارت الشبكة المتصل باجهزة الStorage يكون نوعه host bus adapter (HBA) اما المتصل باجهزة الشبكة يكون NIC .



- Interface Protocols

مجموعة من البروتوكولات التى تمكن من التوصيل بين الhost وال Storage ومنها انواع مختلفة مثل

- **IDE/ATA and Serial ATA**: يكون الStorage موصولة مباشرة بالHost بطريقة تسمى DAS
- **SCSI and Serial SCSI**: تقنية نقل بيانات من النوع Parallel تستطيع نقل البيانات بسرعة 16Mbps
- **Fiber Channel**: تستخدم الكابلات الضوئية فى الاتصال وتصل سرعة نقل البيانات الى 10Gbps
- **Internet Protocol (IP)**: بروتوكول يستخدم فى host-to-host traffic

Storage

- اولا مفهوم ال **Storage**

بداية كان هناك Device يستخدم لتخزين البيانات والمعلومات وهناك انواع منه تسمى بالهارد ديسك HDD واهم مايميز الهارد ديسك هوه قدرة التخزين وسرعة الاداء لذلك كان هناك انواع منها ATA, PATA, SATA, SCSI, SAS & Fiber Channel

- ثم ظهر مؤخرا تقنية ثانية عبارة عن ذاكرة متنقلة (Flash Memory) تسمى بالSolid-State Drive (SSD) وتعتبر أكثر كفاءة من الاولى لانها لا تعتمد فى قراءة البيانات على اى اجزاء متحركة كالHDD فهو عبارة عن مجموعه من الاقراص المتحركة يتم قرائتها بواسطة راس متحرك (ابره) اما الSSD عبارة عن عدة شرائح NAND يتم التنسيق فيما بينهما بواسطة متحكم Controller .

مع زيادة حجم البيانات وحفاظا على البيانات والمعلومات داخل الهارد ديسك كان لابد من وجود عملية Redundantly وهنا ظهرت تقنية الRAID لتحسين اداء تخزين البيانات ويوجد انواع من الRaid منها RAID-0, RAID-1, RAID-5, RAID-10

- ومع ظهور الانترنت وزيادة حجم البيانات المراد تخزينها فظهر الFile Server وهو عباره عن Server يضم العديد من الهارديسك تخزن عليه البيانات عن طريق الشبكة ولكن كان هذا التصميم غير كافي ايضا

- **Direct-Attached Storage (DAS)**

ومعناها ان الهارد ديسك واصل مباشرة بالكمبيوتر او السيرفر عن طريق IDE كابل ATA or SATA ثم تتم عملية الRead & Write للبيانات مباشرة على حسب الOS سواء كانت NTFS or FAT32 باستخدام بروتوكول ال SCSI

- **Network attached storage (NAS)**

هناك طريقة اخرى لتوصيل ال Storage عن طريق الشبكة العادية LAN تسمى NAS نفس طريقة ال File Server باختلاف ان الStorage تكون داخل Arrays تقرا البيانات من عليها عن طريق برتوكولات NFS, CIFS & SMB ومن عيوبها انها File-Level Storage System ولاتستطيع ان تعمل منها Boot .

- **A tape drives**

نوع من انواع الStorage عبارة عن شريط ممغنط يتم الكتابة والقراءه عليه عن طريق Tape Library

يتميز بسرعة نقل البيانات مثال الLinear Tape-Open 7(LTO7) تصل سرعته الى 300 Mbps

Disk Drive Components

يتكون الهارد ديسك للStorage من مجموعه من المكونات الرئيسية مثل:

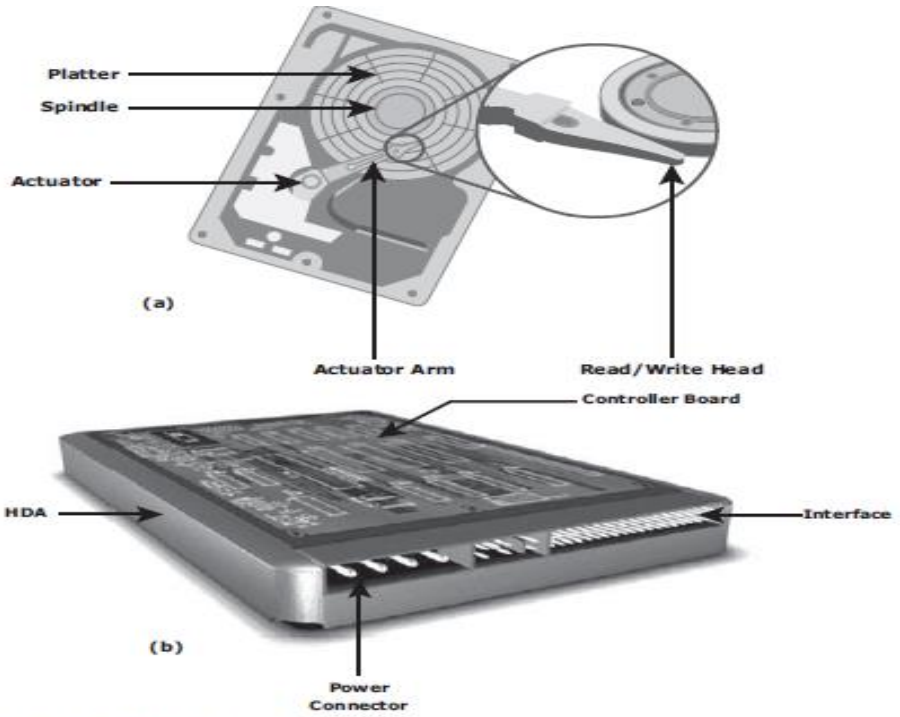
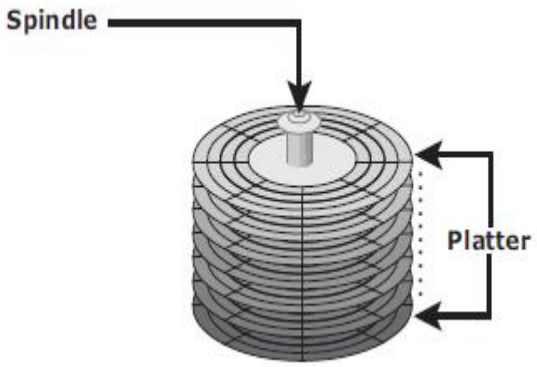


Figure 2-5: Disk drive components

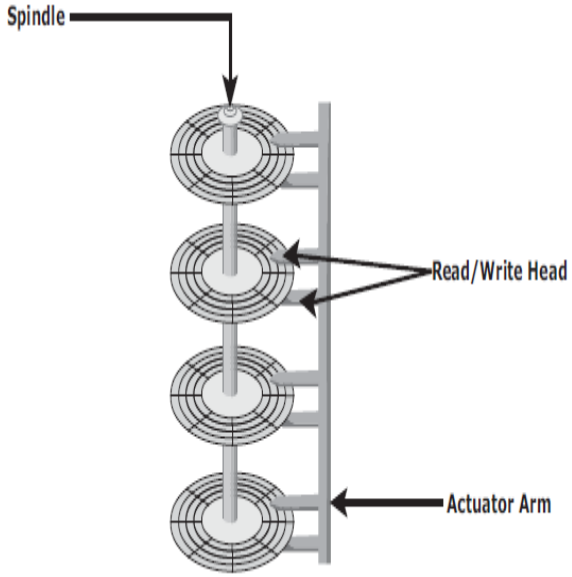
- **Platter**

هي واحد او مجموعه من الاقراص الدائرية التى تخزن عليها البيانات علي هيئة binary codes.



- **Spindle**

هو الذى يربط جميع الاقراص ببعضها وترتبط بمحرك لتحريكه بسرعه ثابتة حيث ان الاقراص تدور الاف المرات فى الدقيقة الواحدة.



- **Read / Write Head**

تقوم بعملية القراءة والكتابة على الاقراص حيث يحتوي كل قرص على رأسين واحد لكل وجه للقرص يتم تحريكهما عن طريق محرك للقيام بعملية القراءة والكتابة.

- **Actuator Arm Assembly**

هو المحرك الذي يقوم بتحريك الرؤوس على الاقراص للقيام بعملية القراءة والكتابة.

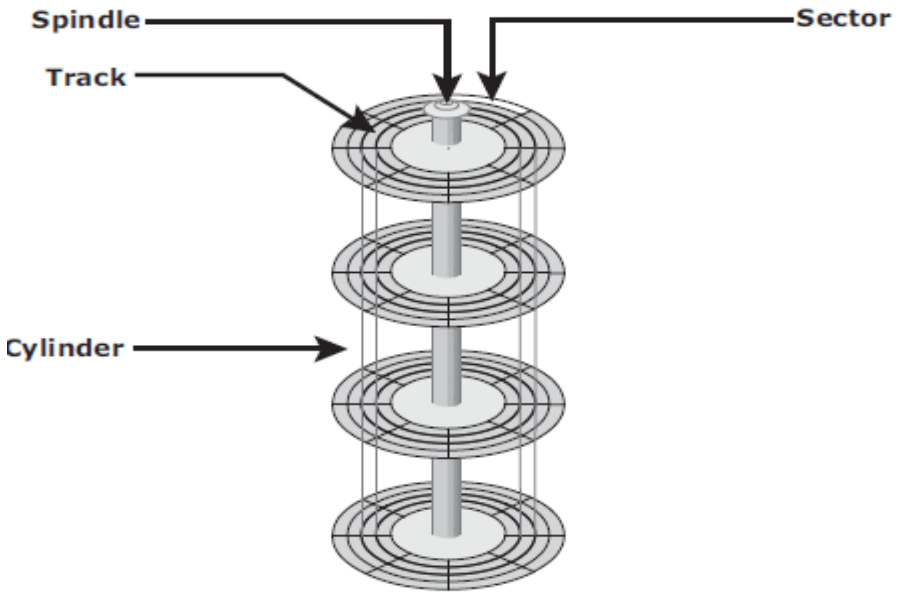
- Drive Controller Board

هي لوحة الدوائر الموجوده فى الهارد ديسك وتحتوى على مجموعه من الmicroprocessor و internal memory and firmware's.

- Physical Disk Structure

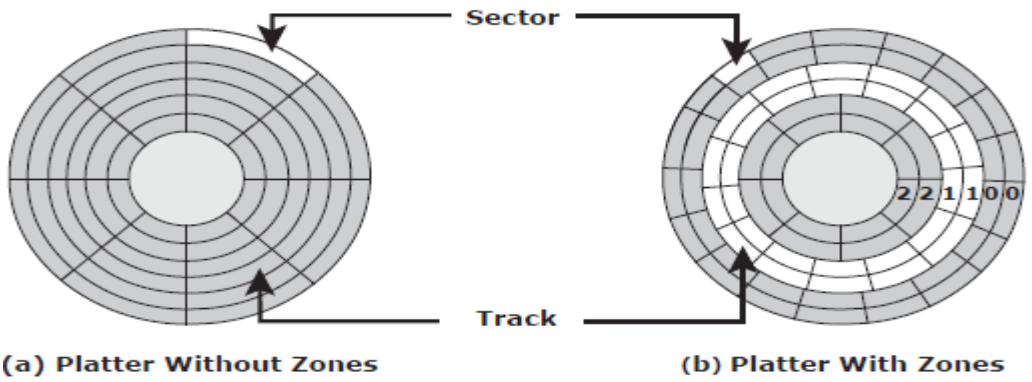
البيانات تخزن فى القرص على مايسمى بال tracks تكون عباره عن مجموعه من الدوائر او الحلقات حول الspindle كما فى الصورة ويتم ترقيمها ابتداءا من الصفر بداية من الحلقة الخارجية The number of tracks per inch (TPI)

كل track ينقسم الى مجموعة من الSectors كل sector يكون قابل لتخزين البيانات بشكل فردي ويختلف عدد الtracks & sectors على الاقراص على حسب اختلاف نوع الهارد ديسك.



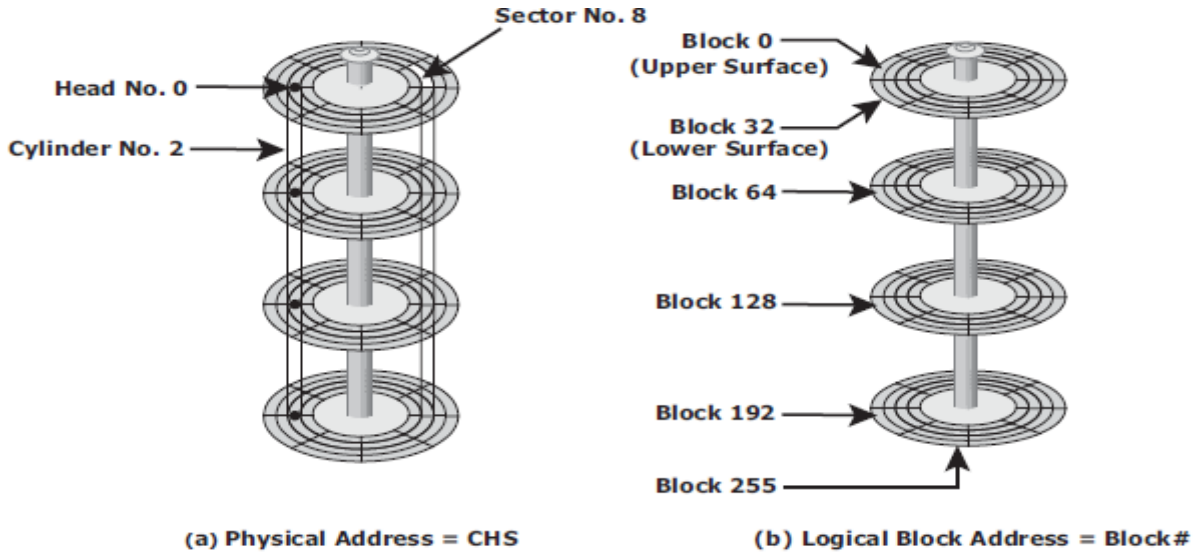
- Zoned Bit Recording

يتم تقسيم الtracks داخل القرص الى zones بناءا على المسافه الى منتصف القرص حيث ان الtracks الخارجيه على طرف القرص تكون مساحتها اكبر من الtracks الداخلية القريبه من المنتصف حيث كل track لا يحتوى على نفس حجم البيانات لذلك تم عمل zone bit recording لتتساوى حجم الtracks وتم ترقيمها حيث ان الzone على الtrack الخارجية تبدا من صفر كما فى الصورة.



- **Logical Block Addressing**

الهارد ديسك يستخدم physical address يتكون من ال cylinder, head and sector (CHS) هذا الرقم يشير الى العنوان المحدد فى الديسك ويجب ان يكون نظام التشغيل على علم بهذه العناوين باستخدام Logical block addressing (LBA) حيث يقوم الdisk controller بتحويل الLBA الى CHS حيث ان نظام التشغيل يريد فقط معرفة حجم الهارددیسك بناءا على عدد الBlocks حيث يتم مقارنتها بال physical sectors على اساس 1:1



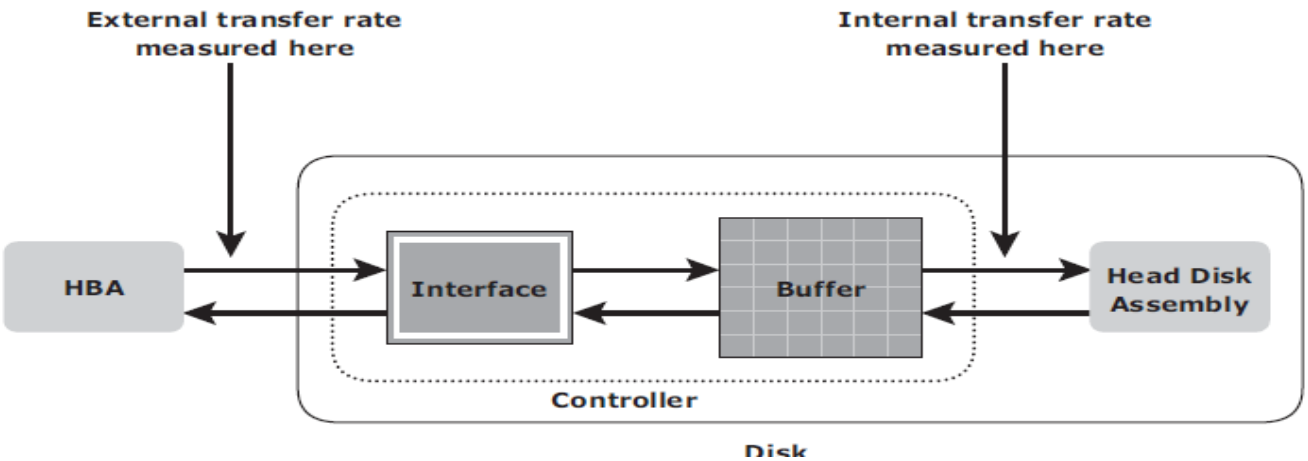
Disk Drive Performance

هناك عدة عوامل تؤثر على اداء محركات الاقراص منها:

- **Disk Service Time**

هو الوقت الذى يستغرقه الهارد ديسك لاكمال مهامه I/O Requests ومن العوامل التى تساهم فى ذلك:

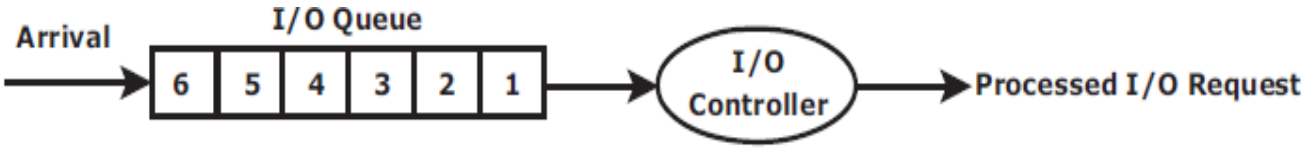
- **Seek Time**: وهو الوقت المستغرق فى تحريك رأس القراءه والكتابه على الplatter ليصل الى الtrack المطلوب, وهناك مواصفات محدده لل Disk Service Time:
  - Full Stroke
  - Average
  - Track-to-Track
- **Rotational Latency**: الوقت المستغرق للقرص للقيام بالدوران وظبط موقع الSector المطلوب تحت رأس القراءة والكتابة.
- **Data Transfer Rate**: متوسط كمية البيانات على الوحده الزمنية المستغرقة لوصول الداتا الى كارت الHBA.



- **Disk I/O Controller Utilization**

استخدام جزء كبير من الهارد ديسك له تاثيرعلى سرعة الاستجابة ويعتمد هذا على عنصرين اساسين :

- **Queue**: موقع انتظار الـI/O Requests قبل المعالجة
- **Disk I/O Controller**: معالجة الـI/O Requests المنتظره فى الـqueue واحده بواحده.

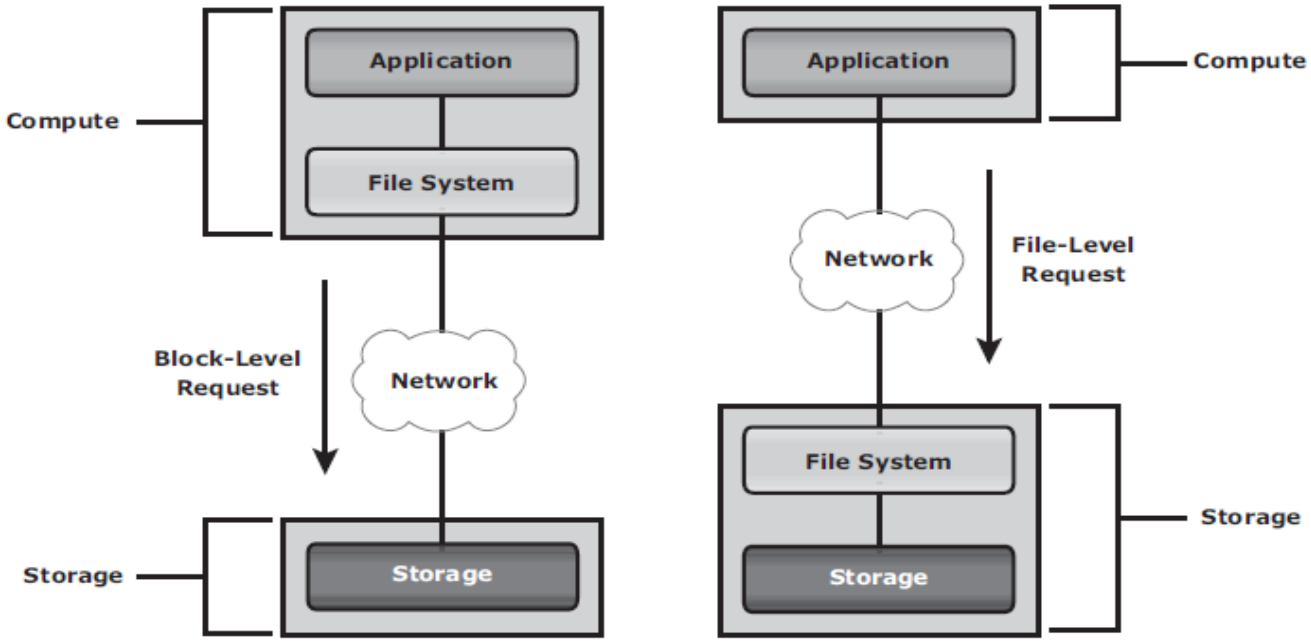


Host Access to Data

- طرق التخزين والوصول الى البيانات تختلف على حسب اختلاف تصميم البنية التحتية للشبكة حيث تعتمد على طريقة توصيل الـHost بالـStorage, فمن الممكن ان تكون الـStorage موصوله داخليا او خارجيا ويتم وصول الـHost للـStorage باستخدام بعض البروتوكولات المختلفة مثل IDE/ATA SCSI or Fibre Channel.
- FC & iSCSI تستخدم للوصول الى البيانات على الـStorage الموصوله خارجيا, اما اذا كانت الـStorage موصوله مباشرة بالـHost فتستخدم ما يسمى بال direct-attached storage (DAS) للوصول الى البيانات.
- يمكن الوصول الى البيانات فى الشبكة بعدة طرق تسمى block level, file level & object level.

حيث انه فى الـFile System يقوم بتقسيم الـfile attributes الى logical block address (LBA) وارسالها لطلب البيانات من الـStorage ثم يقوم الـstorage بتحويل الـLBA الى CHS ثم يقوم بجلب البيانات.

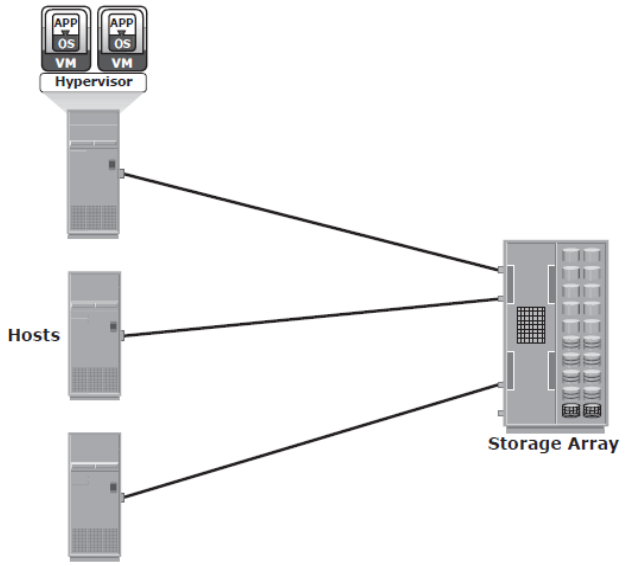
- اما فى الـBlock Level يتم انشاء الـFile System داخل الـHost ويتم الوصول للبيانات من خلالها كما موضح بالصورة.



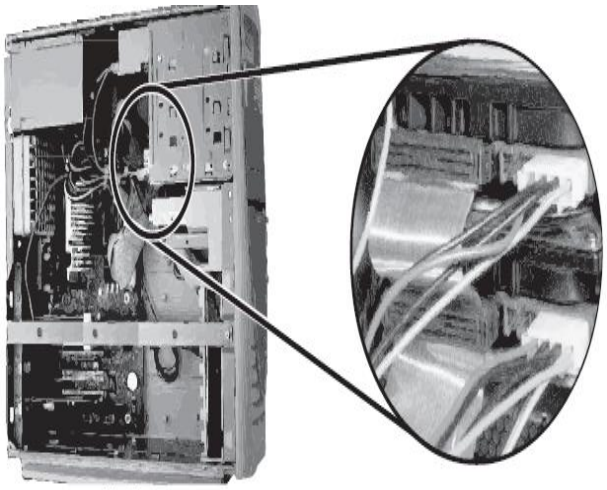
(a) Block-Level Access (b) File-Level Access

Direct-Attached Storage

ومعناها ان الهارد ديسك واصل مباشرة بالكمبيوتر او السيرفر عن طريق IDE كابل ATA or SATA ثم تتم عملية الRead & Write للبيانات مباشرة على حسب الOS سواء كانت NTFS or FAT32 ببروتوكول SCSI.



(b) External DAS



(a) Internal DAS

- DAS Benefits and Limitations

يتطلب تكلفة اقل فى بناء الشبكة بالنسبة للStorage وكذلك الاعدادات تكون بسيطه وسريعة مما يجعل التحكم فى الStorage ابسط واسهل.

Storage Design Based on Application Requirements and Disk Performance

- يسهل تحديد حجم الStorage المطلوبه لاي Application حيث تقدر بحجم وارقام كل من ال file system ومكونات الdatabase المستخدمة فى الApplication نفسه.
- حجم وعدد الI/O Requests وخصائصه التى تم انشائها عن طريق الApplication تؤثر على سرعة استجابة الStorage وبالتالى يؤثر على ال Disk Performance.
- حجم الI/O block يعتمد على ال file system و الdatabase التى تم انشائها عن طريق الApplication
- من المفترض ان ال(Ts) disk service time لاي I/O Request هو مقياس لل Disk Performance وكما ذكرنا سابقا ان ال Ts هي مجموع ال(T) seek time و(L) rotational latency و(X) internal transfer time
- اذن **(Ts = T + L + X)** فعلى سبيل المثال....

لو افترضنا ان معدل ال (T) seek time هو 5ms

وان سرعة دوران القرص هي 15000 revolutions Per minuteاو250 Per Second اي ان L=0.5/250 rps

وان حجم البيانات المراد نقلها 40Mb وعلى سبيل المثال ان حجم ال 32kb=I/O block اذن L=32kb/40mb

اذن الوقت المستغرق لل I/O Controller لنقل block size 32kb هو

(TS) = 5 ms + (0.5/250) + 32 KB/40 MB = 7.8 ms

وبالتالي

the maximum number of I/Os serviced per second or IOPS is **(1/ Ts)** = 1/(7.8 × 10*3) = 128 IOPS

وهنا جدول لتوضيح الحد الاقصى لل IOPS لخدمة الBlock Size المختلفة.

BLOCK SIZE	$T_S = T + L + X$	$IOPS = 1/T_S$
4 KB	5 ms + (0.5/250 rps) + 4 K/40 MB = 5 + 2 + 0.1 = 7.1	140
8 KB	5 ms + (0.5/250 rps) + 8 K/40 MB = 5 + 2 + 0.2 = 7.2	139
16 KB	5 ms + (0.5/250 rps) + 16 K/40 MB = 5 + 2 + 0.4 = 7.4	135
32 KB	5 ms + (0.5/250 rps) + 32 K/40 MB = 5 + 2 + 0.8 = 7.8	128
64 KB	5 ms + (0.5/250 rps) + 64 K/40 MB = 5 + 2 + 1.6 = 8.6	116

نجد هنا ان معدل الIOPS يتراوح من 140 – 116 ومن الممكن ان يصل الى الحد الاقصى للاستخدام وكما ذكرنا من قبل انه معدل الاستخدام يؤثر على سرعة الاستجابة Response Time (R) على سبيل المثال لو ان معدل الاستخدام (U) ل Block size حجمه 32kb هو 96% اذن سرعة الاستجابة (R) هي:

R = TS/(1 – U) = 7.8/(1 – 0.96) = 195 ms

هنا نجد ان سرعة الاستجابة تختلف على حسب معدل الاستخدام

وبالتالى تحديد الStorage المطلوبه لاي Application تعتمد على ال Capacity & IOPS لو افترضنا ان الApplication يتطلب 200GB فمن الممكن استخدام Single Disk. لكن ماذا لو ان معدل الIOPS لل application نفسه عالى فان سرعة استجابة ال Single Disk تكون قليلة على عملية I/O Requests.

اذن عدد الDisks المطلوبة (D_R) تساوي

$$D_R = \text{Max}\ (D_C, D_I)$$

D_C = the number of disks required to meet the capacity

D_I = number of disks required to meet the application IOPS

على سبيل المثال:

لو افترضنا ان المساحة المطلوبه ل Application ما, هى 1.46 TB وان معدل الI/O المنشأه من خلال هذا الApplication هي 9000 IOPS, وعلى حسب مواصفات الVendors بان مساحة الديسك 146GB بمعدل دوران 15000 rbm قادرة على القيام ب 180 IOPS حد اقصى

اذن من المساحة المطلوبه نجد ان عدد الديسك المطلوبه = D_C **10 disks** = 1.46 TB/146 GB

ايضا عدد الديسك المطلوبه لمقابلة الIOPS application = 9000 / 180= 50 Disks

لكن لو افترضنا ان هناك معدل استخدام 70% للديسك فان معدل الIOPS يصبح 126 IOPS =0.7 / 180

وبالتالى يكون عدد الديسك المطلوبه لمقابلة الIOPS application = D_I **72 Disks** = 126 / 9000

اذن عدد الStorage المطلوبه لل Application = **Max (10, 72) = 72 disks**

Disk Native Command Queuing

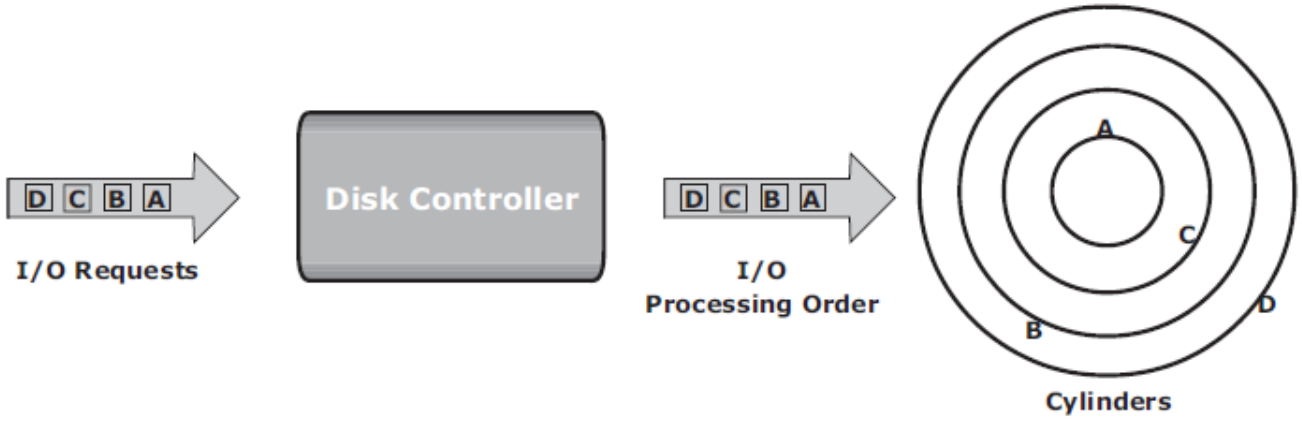
الCommand Queuing هي تقنية جديدة تقوم بتحديد اوامر التنفيذ لل I/O المستقبلة وازالة حركة رؤوس القراءة الكتابه الغير ضرورية منها لتحسين مستوى الديسك.

عندما يستقبل الديسك الI/O للتنفيذ فان command queuing algorithms يقوم بوضع tags لتحديد تسلسل الاوامر التى يجب تنفيذها اولا.

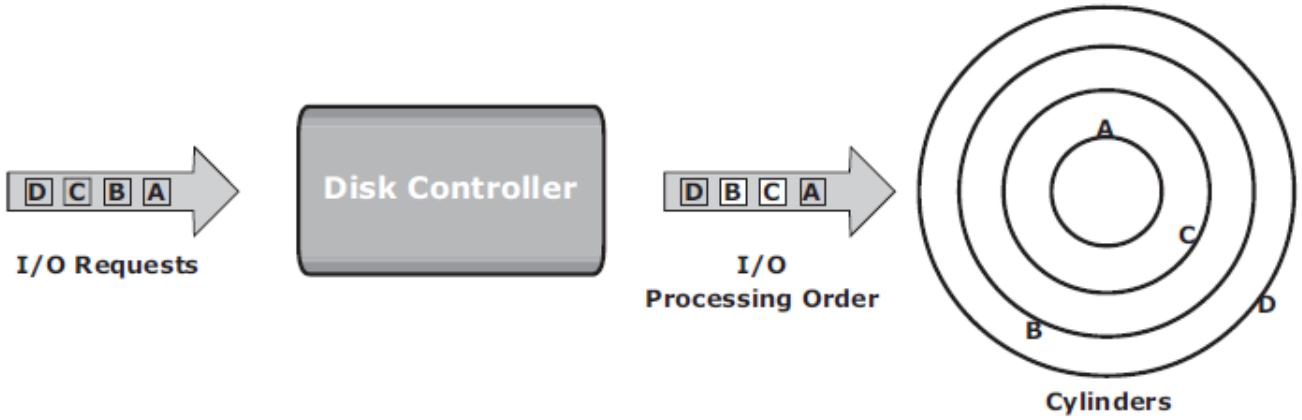
فى الCommand Queuing تعتمد الCommands على تنظيم الداتا داخل الديسك بغض النظر عن الاوامر المستقبله فى اي command وهنا تستخدم الSeek Optimization

تعتمد الCommands ايضا على تحسين القراءة / الكتابة لحركات الرأس، مما قد يؤدي إلى إعادة ترتيب الأوامر

حيث انه بدون الseek optimization يتم تنفيذ الاوامر بالترتيب كما هي . كما بالصورة التالية



(a) Without Optimization



(b) With Seek Time Optimization

Introduction to Flash Drives

- ال Flash Drives هو جيل جديد من الديسكات تسمى ايضا ب solid state drives (SSDs) وهي اكثر كفائه واعلى فى الاداء عن الHDD.
- لانها لا تعتمد فى قراءة البيانات على اى اجزاء متحركة. ولانها تستخدم مايسمى بال semiconductor-based solid state memory (flash memory) لحفظ واسترداد البيانات
- لانها ايضا لا تحتوي على seek time او rotational latencies
- الFlash Drives ايضا تستلم عدد كبير من الIOPS فى وقت استجابه (Respond Time) قليل جدا.
- ايضا الFlash Drives يستهلك طاقة اقل مقارنة بال HDD.
- الFlash Drives مناسب خاصة لل Block Size الصغيرة والتى تتطلب اقل من 1ms للRespond Time
- هناك بعض الApplications التى تتطلب سرعه فى معالجة الداتا مثل currency exchange, electronic trading systems والتى تجد فائدة فى استخدام الFlash Drives.

- Components and Architecture of Flash Drives

- الFlash Drives تشبه فى شكلها الفيزيائى الPhysical Disks الاخرى مما يسهل عملية استبدال الديسكات داخل الStorage Array.

المكونات الرئيسية لل Flash Drives هي :

- **Controller:** لادارة اداة الDrive
- **I/O Interface:** يوفر الطاقة والوصول للبيانات
- **mass storage (collection of memory chips):** هي مجموعه من الmemory chips تسمى nonvolatile NAND (negated AND) تستخدم لتخزين الداتا.
- **Cashe:** يستخدم للتخزين الاحتياطي للداتا للقيام بعملية الBuffering.

- الFlash Drives تستخدم العديد من الI/O Channels الموصوله من الController الى memory chips للوصول الى الداتا مما يؤدى الى ارتفاع الBandwidth الداخلية وكذلك الPerformance
- وتستخدم عادة من 8 الى 24 Channels.
- يتم تقسيم الMemory chips الى blocks & pages
- الpages هي اصغر شئ يمكن القراءة والكتابة عليه فى الflash drives ويتم تقسيمه الى مجموعات داخل الblocks, ليس لها مساحه معينه ممكن ان تبدا من 4 KB, 8 KB – 16 KB
- يمكن ان يحتوى الblocks على 32, 63 or 128 Pages.

- Features of Enterprise Flash Drives

- **NAND flash memory technology**

هي تكنولوجيا مناسبة للقراءة العشوائية للبيانات لانها تستخدم مايسمى بال bad block tracking and error-correcting code (ECC)

- **Single-Level Cell (SLC)-based flash**

تقنية ال NAND تستخدم شكلين مختلفين من الCell Designs

- A multi-level cell (MLC): يمكنها من تخزين اكثر من pit per cell.
- single-level cell (SLC): التى يمكنها تخزين 1 pit فقط

الSLC هي التكنولوجيا المفضله لدى الشركات بسبب ادائها الاعلى وسرعة القراءه والكتابه لديها عن الMLC

- **Write leveling technique**

عنصر مهم فى تحقيق اقصى استفاده من عمر الFlash drives حيث كتابة البيانات يتم تحديثها بشكل متكرر لتجنب اعادة الكتابة على نفس الخلايا. واستخدام اصغر block عند عملية الكتابة الجديدة.

Concept in Practice: VMware ESXi

VMWare هي الشركة الرائدة فى تقنية الServer virtualization حيث تقدم نظام تشغيلي يسمى بالHypervisor

- ال hypervisor هو عباره عن نظام تشغيلي يتعامل مباشرة مع الهاردوير للسيرفر حيث انه يتم تنزيله على الx86 hardware ويقوم بتخصيص نسبة من الresource الخاصه به مثل ال Hard drive, ram, CBU لانشاء virtual machines داخله تعمل على Physical machine واحده.
- الvirtual machine يتم تخزينها داخل ملفات منفصله يمكن نقلها او نسخها على هيئه template
- تخزن الداتا داخل الVM فى single directory داخل ال file system cluster وتسمى بال virtual machine file system (VMFS).
- الESXi هو الvirtual machine host ل VMWare يستخدم الphysical resource لانشاء الVM
- ال ESXi يحتوى على مكونين رئيسين VMkernel and Virtual Machine Monitor
  - VMkernal الكروت التى تستخدم للربط بين الVM والhardware.
  - Virtual Machine Monitor هو المسؤول عن تنفيذ الاوامر داخل الCBUs وعن عملية ال binary translation.

Chapter 3. Data Protection: RAID

مراكز البيانات اصبحت الان تحتوي على العديد من وسائط التخزين Disk Drives داخل ال Storage Infrastructure و مع زيادة حجم البيانات وحفاظا على البيانات والمعلومات داخل الهارد ديسك كان لابد من وجود عملية Redundantly & data protection وهنا ظهرت تقنية الRAID لتحسين اداء تخزين البيانات.
ويوجد انواع من الRaid منها RAID-0, RAID-1, RAID-5, RAID-10.

وهي إختصار لـ redundant array of independent disks وتعد هذه التقنية من أحد أهم التقنيات المستخدمة في السيرفرات وأجهزة حفظ البيانات فهي تقوم بعملية دمج عدة وحدات تخزين مع بعضها البعض ليس بشكل فيزيائي أو عن طريق الكابلات ولكن بطريقة منطقية مقسمة إلى عددة مستويات مثل RAID 0 و RAID 1 و RAID 5 ولكل واحد منها خصائصه فمثلا RAID 0 يقوم بزيادة سرعة النقل والأداء وذلك بتقسيم حزمة البيانات بين عدة وحدات تخزين حيث أن كل وحدة تأخذ جزء من البيانات لتحفظها وهذا يقلص من وقت الحفظ كون وجود عدة أقراص تشتغل في عملية الحفظ بدلا من واحد . أما RAID 1 فهي تعمل عند تواجد وحدتيين تخزين أو أكثر فهي تقوم بأخذ نسخة إحتياطية من وحدة التخزين ونسخها للأقراص الأخرى بشكل كامل بهدف الحفاظ على البيانات من الضياع . وهنالك خصائص أخرى أكثر تعقيدا في المستويات الأخرى حيث يمكننا دمج عدة مستويات مع بعض للحفاظ على البيانات من الضياع بشكل أفضل ولتسريع أداء النقل أيضا.

RAID Implementation Methods

- **Software RAID**

يستخدم host-based software لتقديم وظائف الRAID حيث يكون فى طبقة نظام التشغيل ولا يستخدم hardware خارجي لادارة الRAID Array. ومن مميزاته:

- Performance
- Supported features
- Operating system compatibility

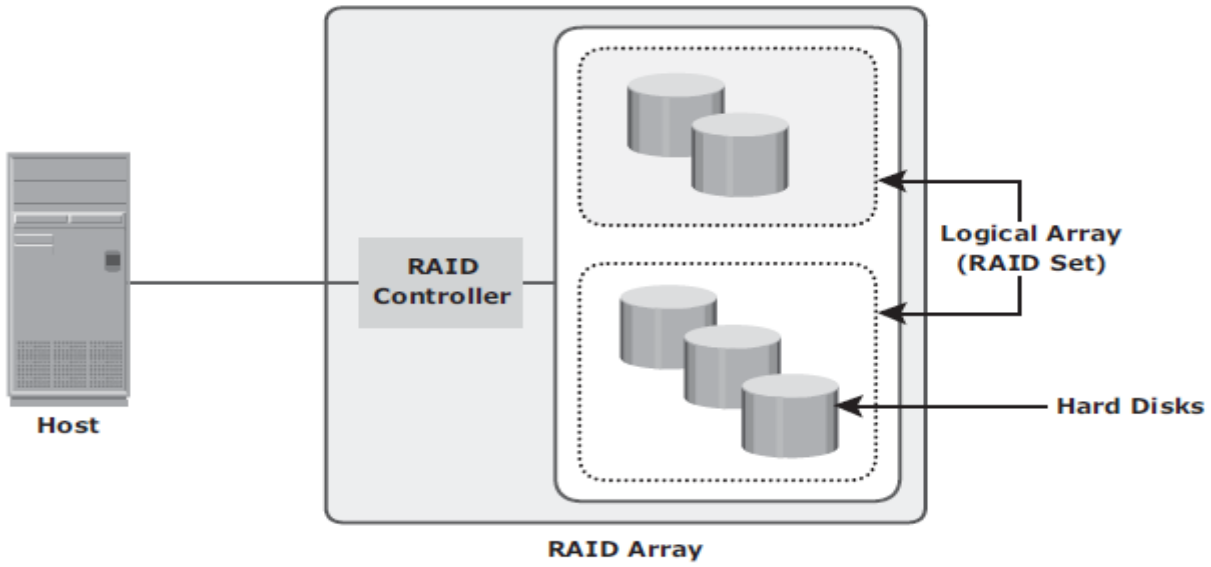
- **Hardware RAID**

يستخدم hardware controller للقيام بوظائف الRAID يكون اما على الhost ويسمى Controller card RAID او على ال array ويسمى The external RAID controller. ومن مميزاته:

- Management and control of disk aggregations
- Translation of I/O requests between logical disks and physical disks
- Data regeneration in the event of disk failures

RAID Array Components

ال RAID array عباره عن enclosure يحتوى على عدد من الديسكات التى تدعم ال RAID من الممكن تجميعهم فى one logical disk وتسمى بالlogical array او RAID group.

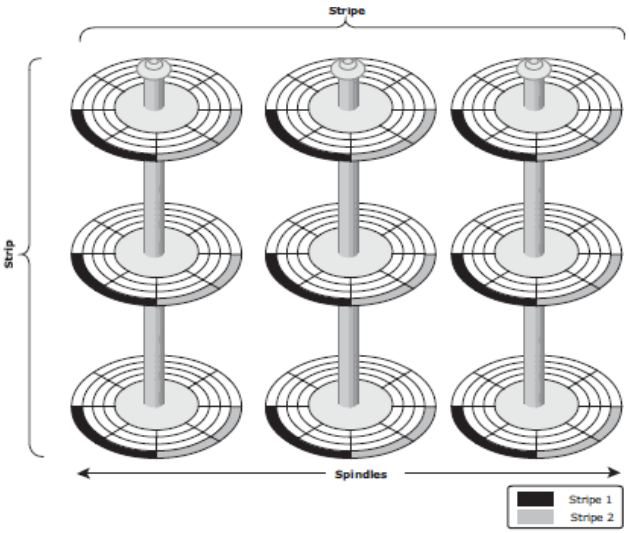


RAID Techniques

يوجد ثلاث تقنيات للRAID وهي:

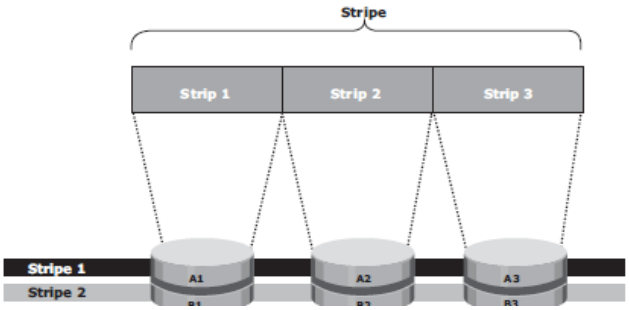
- **Striping**

تعني تقسيم الداتا الى احجام صغيره وتوزيعها على الDrives (more than one) بالتوازي. كل منهم لهم نفس الحجم وتعمل رؤوس القراءه والكتابه فى وقت واحد مما يعمل على معالجة البيانات فى وقت اقصر وبالتالى تزيد سرعة الاداء.



داخل كل ديسك مجموعه من ال RAID ويحتوى كل RAID على عدد من الblock size يتم تحديدها وتعرف بال strips.

حجم الStrip هو عدد الblocks وهو ايضا عباره عن الحد الاقصى للبيانات التى يتم قرائتها وكتابتها على الديسك الواحد.

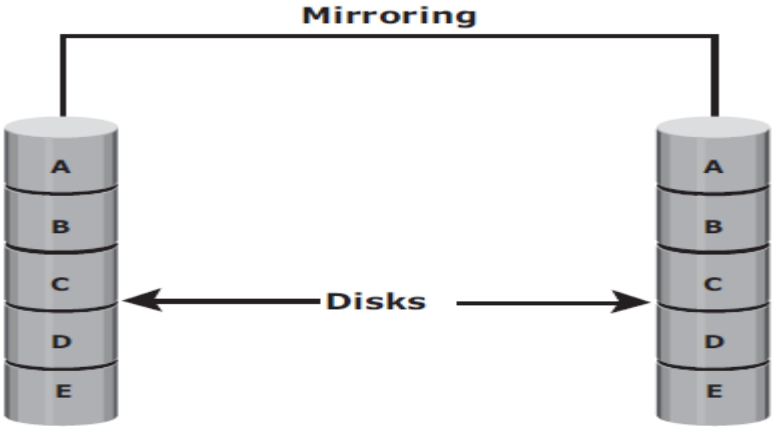


- Mirroring

وهي تعنى تخزين نفس الداتا على اتنين ديسك مختلفين اي انه يتم عمل ال Back-up بشكل فوري ومستمر وذلك بانه يتم كتابة البيانات بنفس الوقت على الtwo disks.

يمكن إضافة أو إزالة اي من الdisks الموجوده داخل الRAID وذلك يضمن استرجاع البيانات فى حالة تعطل اي ديسك دون الحاجه لوجود backup. وذلك يوفر ال Data redundancy & Data Protection.

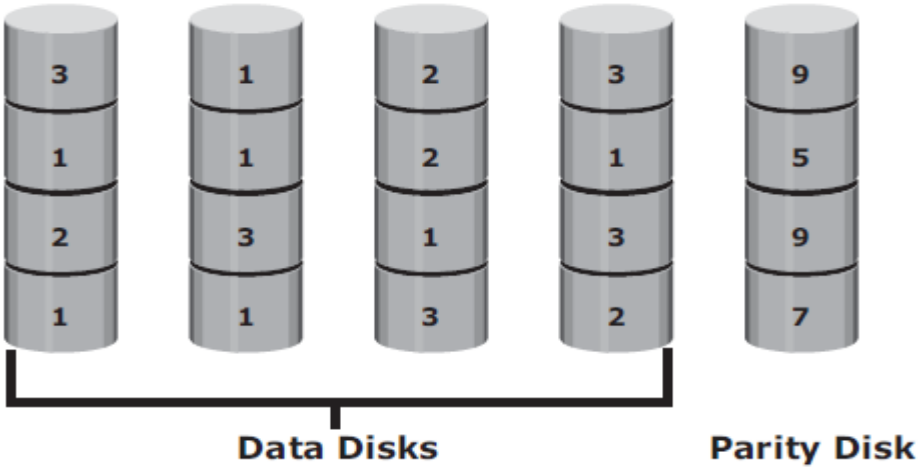
ولكن من عيوبها بطئ عملية كتابة البيانات حيث انه يقوم بالكتابه على two disks وايضا ارتفاع التكلفة المالية.



- Parity

هى طريقه اخرى لحماية البيانات دون زيادة التكلفة كما يحدث فى الmirroring, حيث يتم اضافة ديسكات احتياطيه لضمان عملية الParity حيث ان هذه الديسكات الاضافية تحتوى على معلومات من كل ديسك يمكنها من اعادة انشاء البيانات فى حالة حدوث عطل بها. هذه التقنية ايضا توفر الredundancy & protection دون عمل duplicate data. بمعنى اخر فانه يتم تقسيم البيانات الى عدد من الأجزاء (Stripes) بعدد الديسكات ما عدا ديسك واحد يتم تخزين به جزء اخر يطلق عليه جزء الاحتياط (parity) ويتم توليده من الأجزاء الاخرى,وانشاء الParity يتم عن طريق مقارنة باقي الأجزاء عن طريق (XOR) هذه الدالة تقوم باجراء عملية حسابية على أجزاء البيانات الاخرى لكتابة واسترجاع البيانات.

ال Parity information يمكن تخزينها فى separate, dedicated disk drives او distributed disk موزعه على كل الديسك داخل الRAID كما موضحه بالصورة



RAID Levels

يوجد العديد من انواع الRAID يمكن استخدامها, كل منها تستخدم تقنية معينه سواء كانت Striping, Mirroring or Parity كما موضح بالجدول التالى:

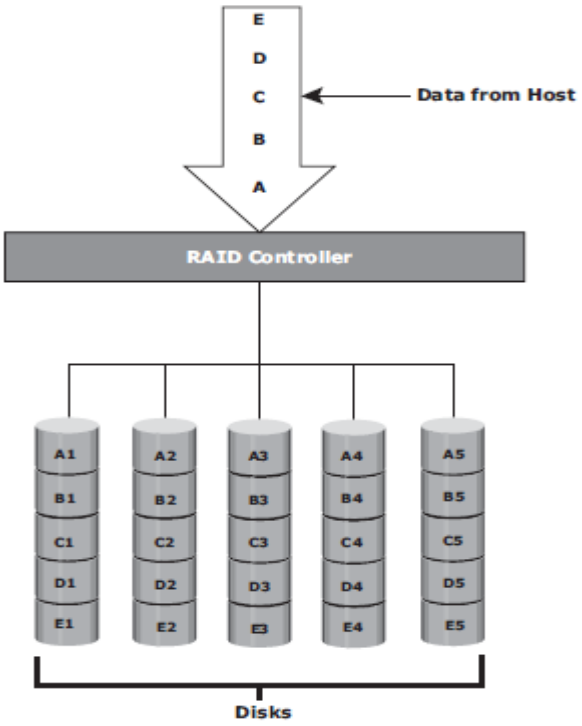
LEVELS	BRIEF DESCRIPTION
RAID 0	Striped set with no fault tolerance
RAID 1	Disk mirroring
Nested	Combinations of RAID levels. Example: RAID 1 + RAID 0
RAID 3	Striped set with parallel access and a dedicated parity disk
RAID 4	Striped set with independent disk access and a dedicated parity disk
RAID 5	Striped set with independent disk access and distributed parity
RAID 6	Striped set with independent disk access and dual distributed parity

- RAID 0

يستخدم تقنية الstriping حيث يقوم الRAID controller بتوزيع البيانات على جميع الديسكات الموجوده بالتساوي(strip size) وبالتالى القراءه والكتابه تعمل بشكل متزامن مما يؤدى الى زيادة فى الاداء (Performance)

مساحة الRAID = جميع مساحات الdisks الموجوده داخل الRAID.

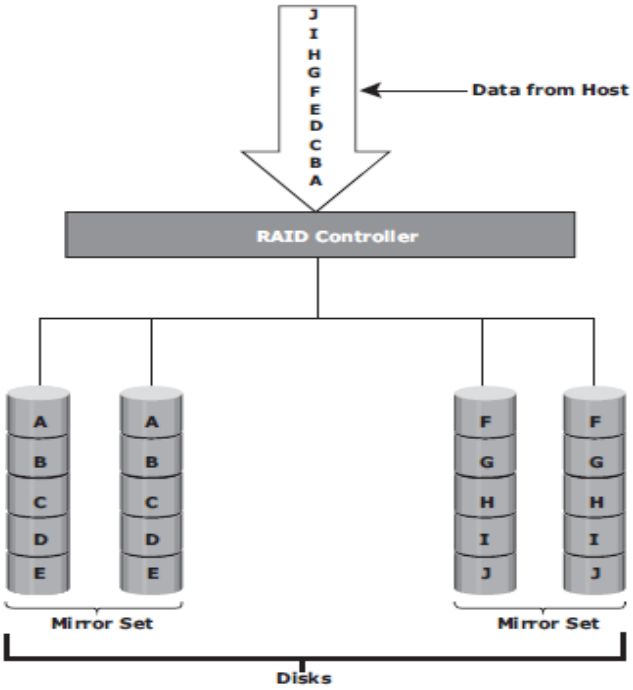
من عيوبه عند تلف اي ديسك داخل الRAID تضيع كل البيانات المخزنه. يكون استخدامه جيد مع الApplication التى تتطلب high I/O throughput.



RAID 1 -

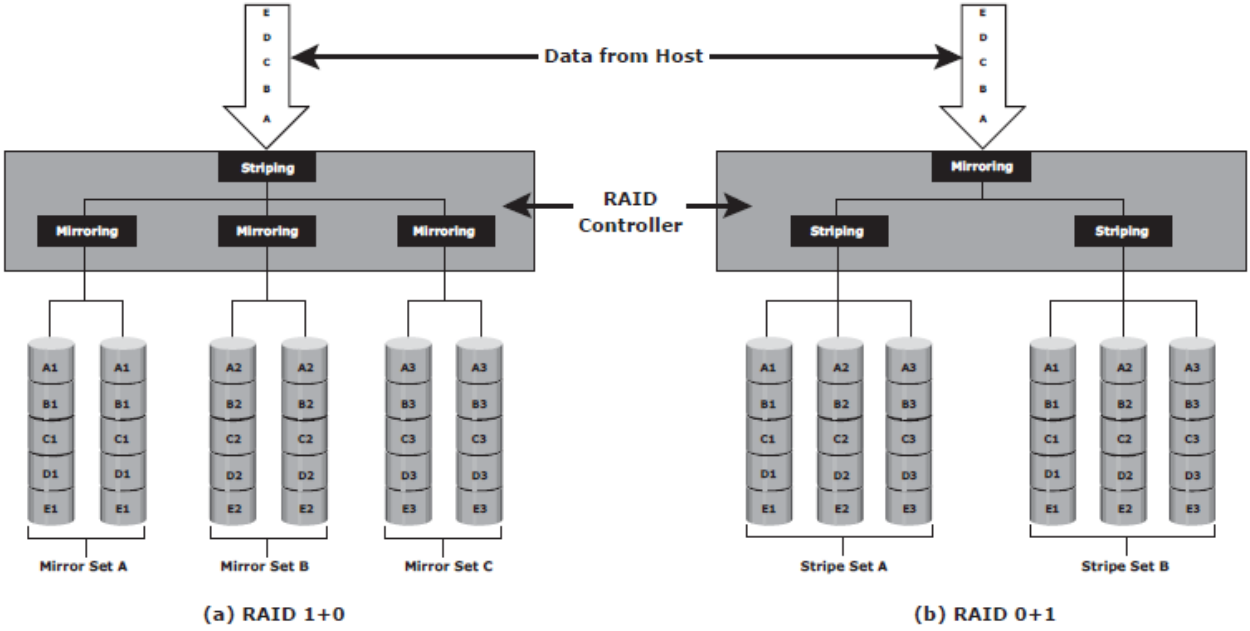
تستخدم تقنية الmirroring حيث يقوم بالحفاظ على البيانات من خلال نسخها على ديسك اخر كـ backup يكون الديسك الاخر نفس الصوره للاول (Redundancy).

يعتبر هو الافضل فى عملية الRecovery وبذلك يعتبر افضل للتطبيقات التى تحتاج High Availability لكن يوجد به بعض العيوب حيث يحتاج الى واحدات تخزين اكثر لانه يقوم بتكرار البيانات على اكثر من ديسك وبالتالى يكون ابطئ فى عملية الكتابة للبيانات.



Nested RAID -

هي خليط من RAID 0 و RAID 1 بحيث تحقق الاداء المميز (Performance) من RAID 0 والحماية المطلوبه (Redundancy) من RAID 1. تعرف باسم RAID 1+0 (RAID 10) or RAID 0+1 (RAID 01)



لكي نفهم طريقة عمل ال Nested RAID على سبيل المثال من الصورة السابقة

RAID 1+0 نجد ان هناك 6 disks تم دمجهم ليعملو ك (RAID 1 first and then RAID 0) RAID1+0

حيث RAID 1 طبقت الmirroring على كل اتنين disk كما يلي:

Drives 1+2 = RAID 1 (Mirror Set A)

Drives 3+4 = RAID 1 (Mirror Set B)

Drives 5+6 = RAID 1 (Mirror Set C)

ثم RAID 0 طبقت الstriping على كل ال Sets A,B & C

اما فى **RAID 0+1** نجد انه تم تطبيق ال striping اولا لكل من:

Drives 1 + 2 + 3 = RAID 0 (Stripe Set A)

Drives 4 + 5 + 6 = RAID 0 (Stripe Set B)

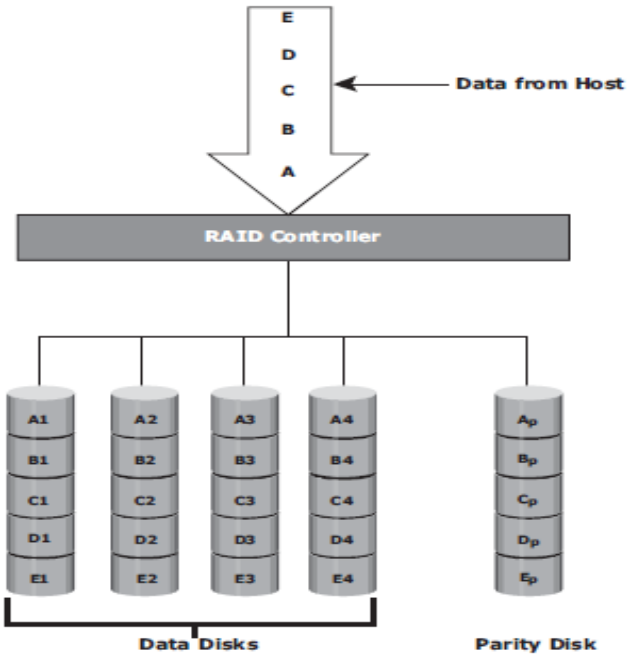
ثم تم تطبيق ال mirroring بين set A & B.

- RAID 3,4

تستخدم تقنية الParity حيث يقوم بعمل stripes للداتا من اجل زيادة الاداء وكذلك يقوم بعمل Parity من اجل fault tolerance.

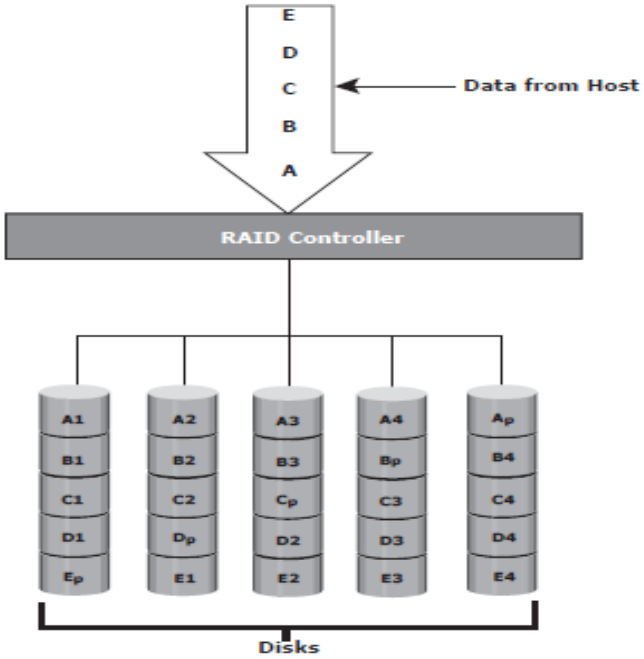
حيث ان الParity يقوم بتخزين معلومات عن كل الDrive الموجوده بحيث لو حدث اي خطا فى اي منهم يقوم باعادة بناء البيانات.

ولكن يختلف فى RAID 4 انه يقوم بالوصول الى البيانات على الDisks المطبق عليها الstriping بشكل مستقل دون الحاجه الى قراءة البيانات كامله على الstriping.



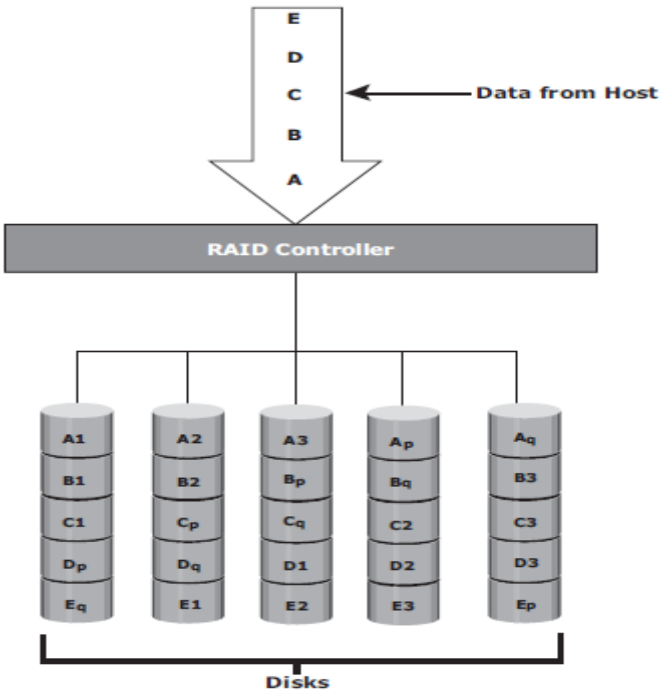
RAID 5 -

يتشابه مع RAID 4 من حيث عمل الStriping والوصول الى strips drivers بشكل مستقل مع تطبيق ال Parity ايضا مع اختلاف موقع الParity حيث انه فى RAID 4 يكون ديسكات محدده لكن فى RAID 5 تكون موزعه على كل الديسكات كما موضح بالصورة لتجنب المشاكل التى تحدث عند الكتابه على dedicated parity disk.



RAID 6 -

يعمل بنفس طريقة ال RAID 5 مع اختلاف وجود Parity اخر فى RAID 6 وذلك لتجنب ضياع البيانات اذا حدث خطأ فى اتنين ديسك مع بعض داخل الRAID ولذلك تكون الكتابة ابطئ به عن RAID 5 نظرا لوجود Parity اخر وايضا اعادة انشاء البيانات تكون ابطئ.

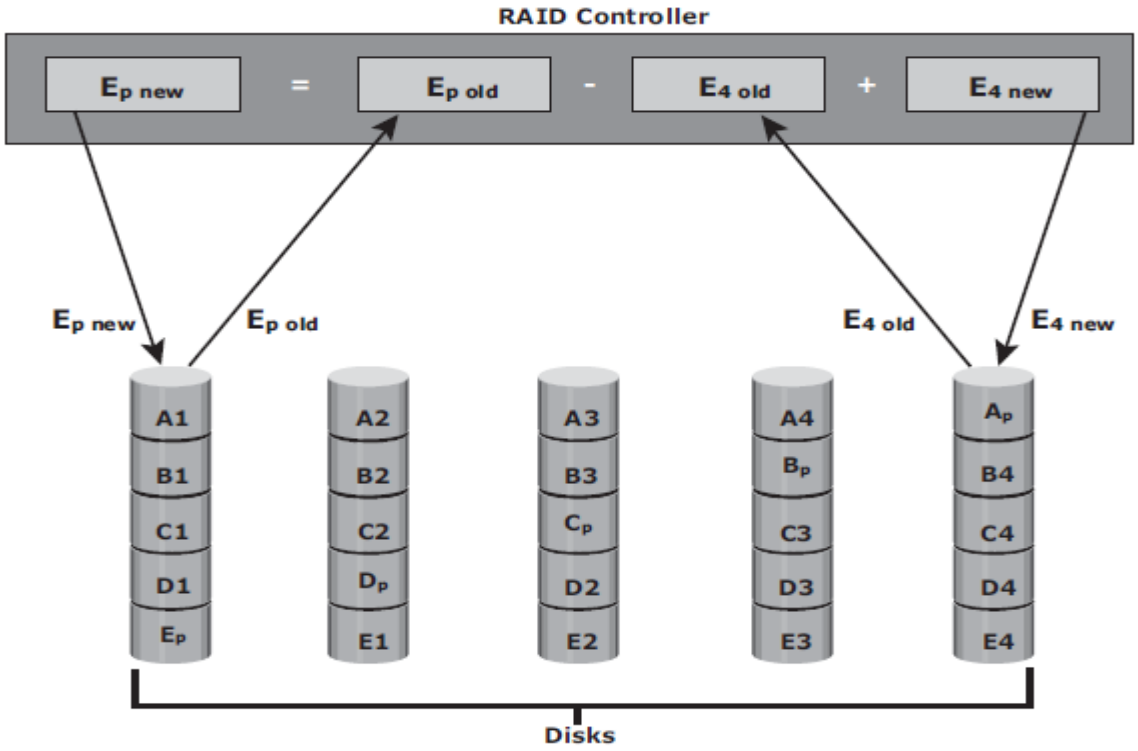


RAID Impact on Disk Performance

عند اختيار نوع الRAID الذي سيتم تطبيقه يجب النظر اولا الى مدى تاثيره على اداء الDisks والتطبيقات (IOPS).

على سبيل المثال فى نظامي ال mirroring & Parity يكون هناك العديد من المدخلات والمخرجات I/O مثل قراءة وكتابة البيانات.

فنجد مثلا فى RAID 1 كل عملية كتابة يجب ان تتم على اتنين ديسك مختلفين الذين يستخدمون نظام الmirroring, كذلك فى RAID 5 عملية الكتابة من المحتمل ان تمر على اربعة عمليات مختلفه حيث ان الRAID controller يقوم بقراءة وكتابة واعادة تهيئة البيانات وكتابتها على الParity لكل داتا يتتم قرائتها على الRAID. كما نجد فى الصورة التالية



تم حساب الParity (P) على اساس المعادلة XOR حيث

$$E_p = E_1 + E_2 + E_3 + E_4 \text{ (XOR operations)}$$

وعندما يتم قراءة بيانات جديدة عن طريق الRAID controller يتم حساب الParity عن طريق قراءة الParity القديم (Ep old) بالاضافة الى قراءة البيانات القديمة (E4 old)وبعد ذلك يقوم الController بكتابة البيانات الجديدة على الديسك وعلى الParity (Ep new) وهذا معناه انه يوجد عمليتين قراءة وعمليتين كتابة لكل عملية كتابة جديدة للبيانات داخل الRAID.

$$E_{p\ new} = E_{p\ old} - E_{4\ old} + E_{4\ new} \text{ (XOR operations)}$$

وكذلك فى RAID 6 نجد انه يتطلب 3 عمليات قراءة (2 Parity & 1 Data) وكذلك ايضا القراءة. اي 6 عمليات I/O على الديسكات.

- Application IOPS and RAID Configurations

عند تحديد عدد الديسكات المطلوبه لاي تطبيق يجب الاخذ فى الاعتبار مدى تاثير الRAID بناءا على عدد الIOPS المنشأه من خلال الapplication. حيث ان الDisk load يحسب بناءا على الRAID configuration وعمليات القراءة والكتابه التى تتم من الhost.

على سبيل المثال لو ان هناك application ينشا 5200 IOPS, حوالى 60% منهم يتم قرائتهم

اذن ال disk load يحسب كالاتى:

The disk load in RAID 5 is calculated as follows:

RAID 5 disk load (reads + writes) = 0.6 x 5,200 + 4 x (0.4 x 5,200) [because the write penalty for RAID 5 is 4]

= 3,120 + 4 x 2,080

= 3,120 + 8,320

= 11,440 IOPS

The disk load in RAID 1 is calculated as follows:

RAID 1 disk load = 0.6 x 5,200 + 2 x (0.4 x 5,200) [because every write manifests as two writes to the disks]

= 3,120 + 2 x 2,080

= 3,120 + 4,160

= 7,280 IOPS

بالمقارنة RAID 1 مع RRAID 5 نجد ان عدد الديسكات المطلوبه للعمل مع الworkload مع اعتبار ان الحد الاقصى للديسك الواحد حسب مواصفاته هي 180 IOPS

RAID 5: 11,440/180 = 64 disks

RAID 1: 7,280/180 = 42 disks (approximated to the nearest even number)

RAID Comparison -

Table 3-2: Comparison of Common RAID Types

RAID	MIN. DISKS	STORAGE EFFICIENCY %	COST	READ PERFORMANCE	WRITE PERFORMANCE	WRITE PENALTY	PROTECTION
0	2	100	Low	Good for both random and sequential reads	Good	No	No protection
1	2	50	High	Better than single disk	Slower than single disk because every write must be committed to all disks	Moderate	Mirror protection
3	3	[(n-1)/n] × 100 where n= number of disks	Moderate	Fair for random reads and good for sequential reads	Poor to fair for small random writes and fair for large, sequential writes	High	Parity protection for single disk failure
4	3	[(n-1)/n] × 100 where n= number of disks	Moderate	Good for random and sequential reads	Fair for random and sequential writes	High	Parity protection for single disk failure
5	3	[(n-1)/n] × 100 where n= number of disks	Moderate	Good for random and sequential reads	Fair for random and sequential writes	High	Parity protection for single disk failure
6	4	[(n-2)/n] × 100 where n= number of disks	Moderate but more than RAID 5.	Good for random and sequential reads	Poor to fair for random writes and fair for sequential writes	Very High	Parity protection for two disk failures
1+0 and 0+1	4	50	High	Good	Good	Moderate	Mirror protection

Hot Spare -

يقصد بها ديسكات احتياطيه داخل الRAID array تستبدل مؤقتا مع الديسكات التى يحدث بها مشكلة لتقوم بعملية الData Recovery حسب نظام ال RAID المطبق. على سبيل المثال

اذا كان Parity RAID هو المستخدم فيتم اعادة انشاء البيانات داخل الhot spare من خلال الParity Disk داخل الRAID set.

اما اذا كان Mirroring RAID فيتم نسخ البيانات من الmirror disk الاخر.

وعندما يتم اضافة ديسكات جديدة الى RAID array يتم نسخ البيانات من الhot spare الى الديسك الجديد

ثم يعود الhot spare الى وضع الاستعداد ليحل مكان اي ديسك اخر عند حدوث اي مشكله به.

Chapter 4. Intelligent Storage Systems

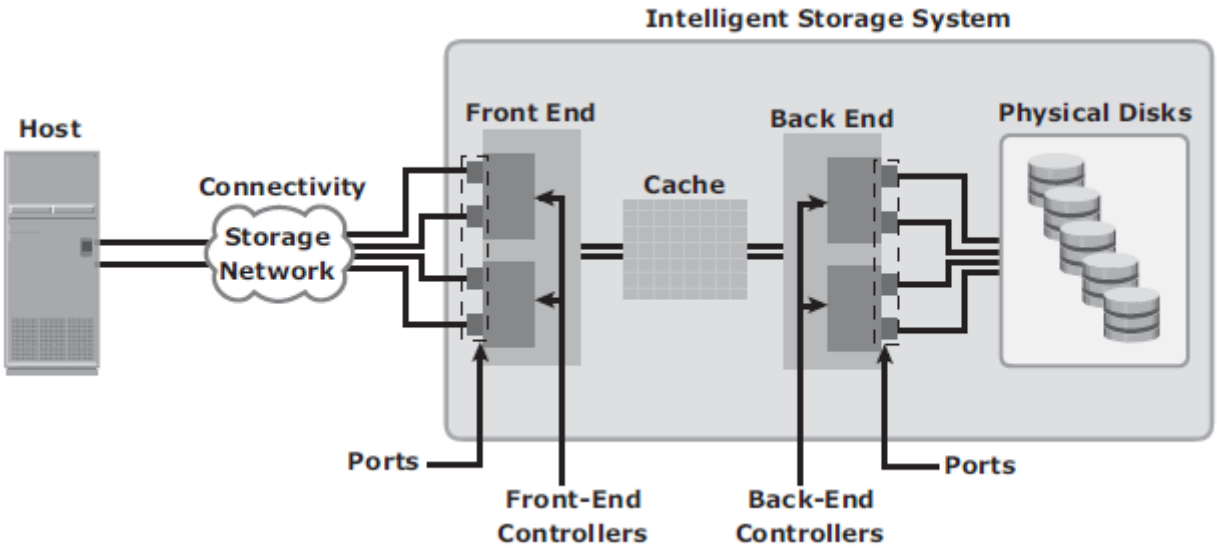
تكنولوجيا الRAID تساهم بشكل كبير فى تحسين اداء الStorage System لكنها قد لاتكون كافية بالنسبة لبعض التطبيقات الحديثه.

ومع تقدم التكنولوجيا ظهرت حلول اخرى للStorage تسمى ب intelligent storage systems هي تطور لتكنولوجيا الRAID حيث يمكنها القيام باداء افضل لعمليات ال I/O بقدرات عالية. حيث يتم تهيئتها بعدد كبير من الmemory تسمى (cashe) ايضا عدد كبير من الI/O Paths عن طريق algorithm متطورة لتلبية متطلبات التطبيقات الحديثة والتى تحتاج الى اداء عالى.

Components of an Intelligent Storage System

تحتوى الIntelligent Storage Systems على اربع مكونات رئيسية هم front end, cache, back end, & physical disks. بالاضافة الى االربط الداخلى بينهم كما موضح بالصورة

حيث يتم استقبال الI/O Request من الHost عن طريق الFront-end ثم يتم معالجتها عن طريق الcashe و الBack-end حتى تمكن من تخزينها واسترجاعها من الPhysical-Disks.



Front End -

هو الInterface الموصول بين الStorage system وال Host ويحتوى على جزئين هما ال Front-end controller & front-end ports ويكون هناك اكثر منController من اجل عملية الRedundancy يحتوى كل منهم على اكثر من Ports لكي يقدر على توصيل العديد من الhost على الIntelligent storage.

ويستخدم الfront-end controller العديد من البروتوكولات لعملية نقل البيانات مثل Fibre Channel, iSCSI, FICON, or FCoE, حيث يقوم بتوجيه البيانات المستقبله وارسال Acknowledgement الى الHost.

- Cashe

هي ذاكرة تخزين احتياطي يتم وضع البيانات بها مؤقتا لتقليل الوقت المطلوب لمعالجة الI/O requests المستقبله من الHost.

الCashe يقوم بتحسين اداء الStorage System عن طريق عزل الHost عن الHDD

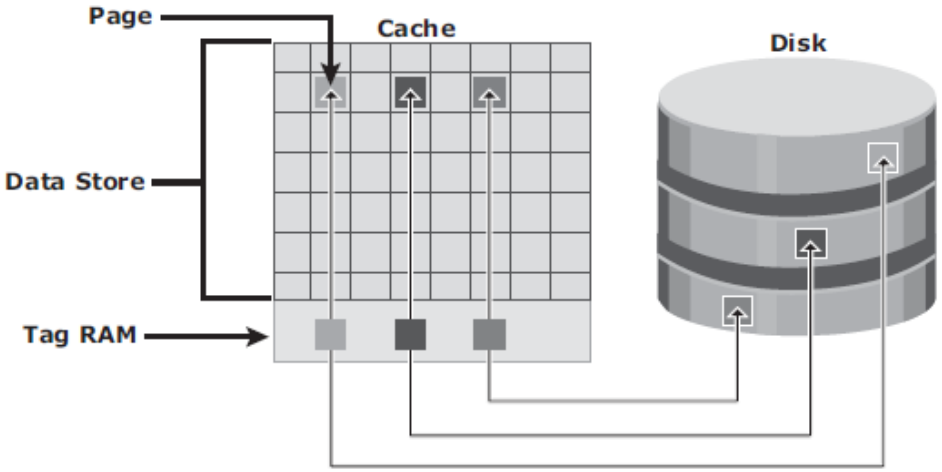
الديسكات بشكلها الفيزيائى تعتبر دائرية والحركة الدائرية لها تعتبر هي الجزء الابطئ داخل نظام الintelligent storage حيث ان الوصول للداتا خلال دوران الديسك ياخذ بعض الوقت milliseconds بسبب الseek time & rational latency بعكس الوصول الى الداتا الموجوده على الcashe تاخذ اقل من millisecond داخل الintelligent arrays.

- **Structure of Cache**

يتم تخزين الcashe داخل صفحات (Pages) وهي اصغر وحده تخزين داخل الcashe

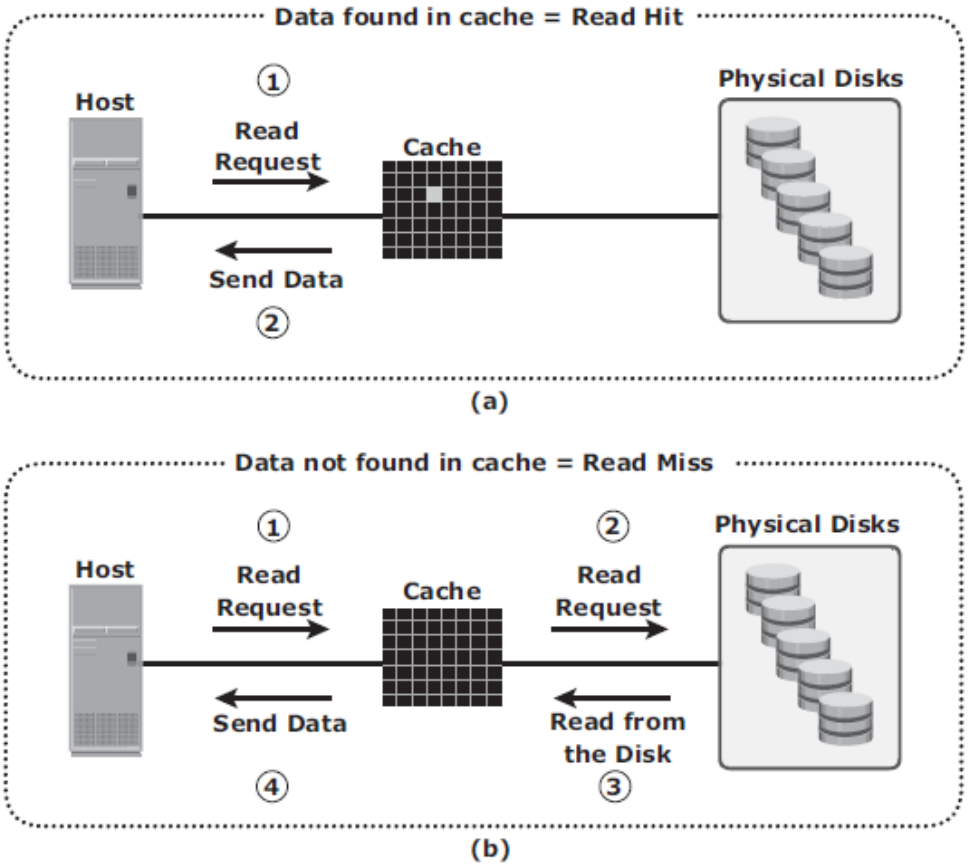
حجم كل page يحدد بناءا على I/O size الخاص بالapplication.

والcashe يتكون من data store & Tag RAM حيث الdata store تحمل البيانات بينما الtag RAM تتتبع مواقع البيانات داخل الdata store.



- **Read Operation with Cache**

عندما يطلب ال host قراءة اي بيانات يقوم الStorage controller بقراءة الtag RAM لكي يحدد هل هذه البيانات موجوده فى الcashe ام لا, فاذا كانت موجوده ترسل مباشرة الى الhost بدون اي اجراء للديسك وتسمي هذه القراءة بال read cashe hit, اما اذا كانت الداتا غير موجوده بال cashe تسمى بال read cashe miss وتتطلب قراءة البيانات من الديسك حيث يقوم الback-end باسترجاع الداتا المطلوبه.



- **Write Operation with Cache**

الكتابة على الcashe تتضمن مزايا اكثر من ناحية الاداء حيث انها تاخذ وقت اقل بكثير من الكتابة مباشرة على الديسك. وايضا من خلال تجميع الكتابات الصغيره من الcashe ودمجها الى كتابات كبيرة وتحويلها الى الديسك.

- **Write-back cashe:** تخزن البيانات على الcashe ويتم ارسال acknowledgment الى الhost مباشرة ثم مؤخرا بعد عدة كتابات يتم حفظه على الديسك
- **Write-through cashe:** تخزن البيانات على الcashe ويتم ارسال acknowledgment الى الhost ويتم تخزينها على الديسك مباشرة.

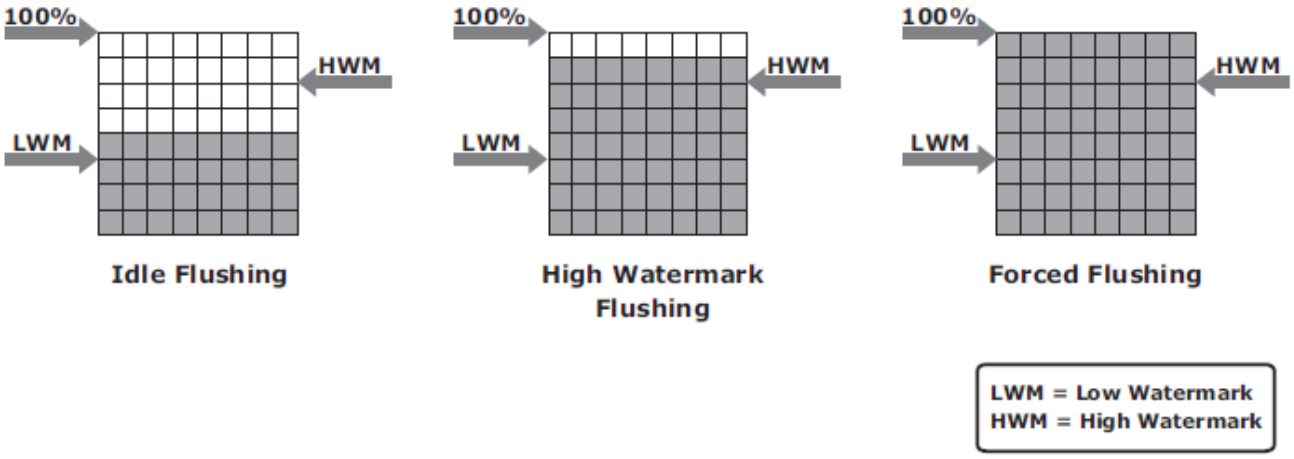
- **Cache Implementation**

يتم تطبيق الcashe اما dedicated cache او global cashe, حيث فى الdedicated cashe يتم تخصيص جزء منفصل من الmemory للقراءة والكتابة, اما فى الglobal cashe تتم القراءه والكتابه على اي جزء من الmemory المتاحه.

- **Cache Management**

تاتى الIntelligent storage مع عدد كبير من الcashe (ذاكرة التخزين المؤقته) والتى يجب من ادارتها بشكل جيد حيث انه عند ملئ الصفحات الخاصه بال cashe هناك بعض الصفحات الاخرى يجب تحريرها لاستيعاب البيانات الجديدة وتتم هذه عن طريق استخدام بعض الalgorithms مثل:

- Least Recently Used (LRU)
- Most Recently Used (MRU)
- Idle flushing
- High watermark flushing
- Forced flushing



- **Cache Data Protection**

الcashe تعتبر ذاكره مؤقته واي فقدان للداتا بها تتسبب فى ضياع البيانات التى لم تخزن الى الديسك ولذلك لابد من حمايتها لتخفيف خطر فقدان تلك البيانات وذلك باستخدام

- **Cashe Mirroring**

حيث كل كتابة على الcashe تتم على اتنين memory loacotin مختلفين ايضا على اتنين memory cards منفصلين.

- **Cashe vaulting:**

لتقليل خطر ضياع البيانات الناتجه عن فصل الطاقه عن طريق استخدام battery مع الذاكره بحيث لو يتم استعادة الطاقه ليتم كتابة البيانات على الcashe ولكن اذا حدث انقطاع لمدة كبيره فان استخدام البطاريات لا يكون قابل للتطبيق هذا لان فى الintelligent storage تستخدم كميات كبيره من البيانات التى تحتاج الى وقت كبير لحفظها على الديسكات والتى لاتوفرها عمل البطاريات, لذلك يتم استخدام physical disks لتفريغ محتويات الcashe بها خلال انقطاع الطاقه وهذا مايسمى بالcashe vaulting وهذه الديسكات تسمىvault drive وعندما تعود الطاقه يتم كتابة البيانات مره اخرى فى الcashe ثم تخزن الى الديسكات.

- **Back End**

هو الInterface الوسيط بين الCashe وال Physical disks ويحتوى على جزئين back-end ports and back-end controllers. حيث يقوم الcontroller بنقل البيانات من الcashe الى الPhysical disks الموصوله بالback-end ports.

- **Physical Disk**

التى تخزن عليها البيانات وتكون موصوله بال back-end controller حيث ان ال modern intelligent storage تدعم مجموعه متنوعه من الديسكات بسرعات مختلفه مثل الFC, SATA,SAS, and flash drives

Storage Provisioning

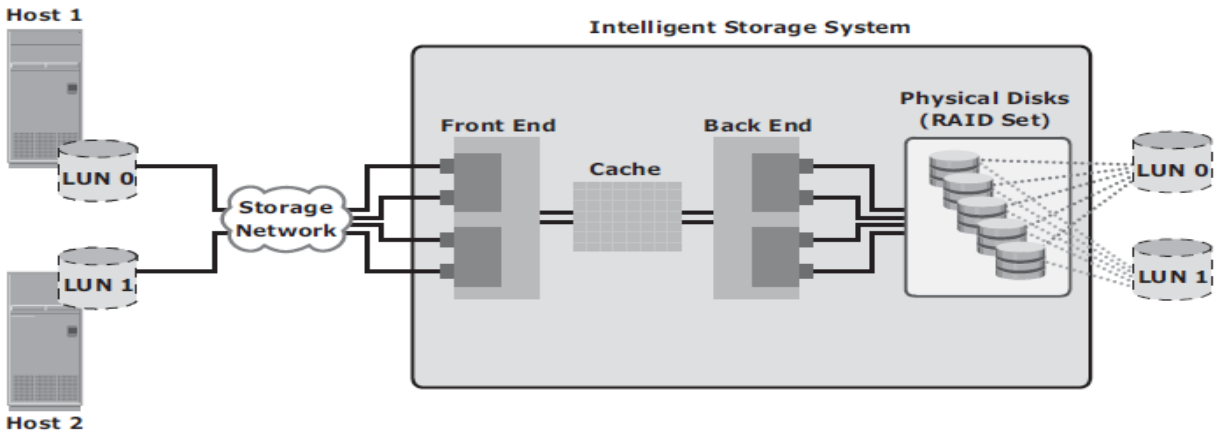
هي عملية تخصيص الStorage resources الى الHosts بناءا على السعة المطلوبه والاداء المطلوب من الapplication المشغل على الhost ويمكن تنفيذها بطريقتين:

- **Traditional Storage Provisioning**

فى هذه الطريقه يتم تجميع الPhysical Disks داخل الRAID المطلوب تطبيقه وبناءا عليه يتم تخصيص الAvailability, Capacity & Performance المتاحه خلال الRAID set ومن المفترض ان يكون الRAID set مطبق من خلال ديسكات لها نفس النوع والسرعه والحجم لتحقيق الاستفادة القصوى من الاداء.

ال RAID sets تحتوى دائما على مساحات عالية لانها تقوم بدمج مجموعه من الديسكات الفردية وتقوم بانشاء ما يسمى بال logical units عن طريق تقسيم المساحه المتوفره الى units صغيره هذه ال units يتم تخصيصها الى الhosts بناءا على المساحه المطلوبه منها.

كل logical unit تم انشائها من خلال ال RAID يكون لها رقم خاص بها Unique ID يسمى بال Logical unit number (LUN) هذه الLUN تقوم بتنظيم وتكوين الRAID وتخصيصها الى الhosts وهي تمثل طريقة الTraditional Storage Provisioning



عندما يتم تخصيص ال LUN الى non-virtualized host يظهر كانه raw-disk داخل نظام التشغيل ويتم تهيئته واستخدامه من خلال نظام التشغيل, اما اذا كان ال LUN مخصص الى virtualized host فيتم تخصيصه الى ال Hyper-visor والذي يقوم باعادة هيكلته الى raw disks لاستخدامه داخل ال virtual machines من خلال ال virtual disks.

يمكن ايضا للVirtual machines الوصول مباشرة الى الLUN, حيث فى هذه الحاله يتم تخصيص ال LUN الى single virtual machine ويتم استخدامها عندما يكون الapplication يعمل على الvirtual machine.

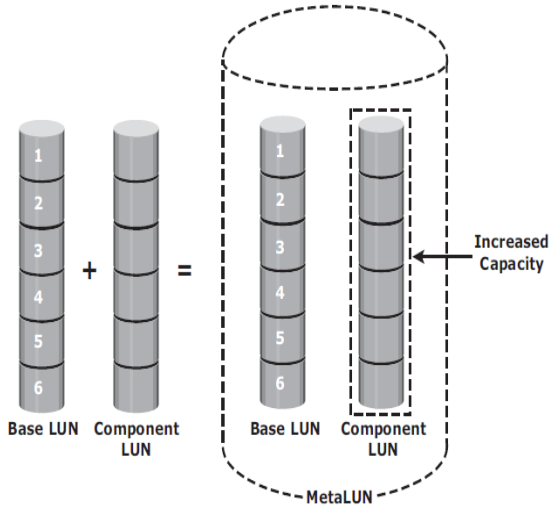
- **LUN Expansion: MetaLUN**

هي طريقه تستخدم لتمديد ال LUN التى تحتاج الى مساحه اضافية او اداء اعلى عن طريق دمج اتنين LUN او اكثر ومن الممكن ان تكون concatenated or Striped.

- **Concatenated expansion**

فى هذه الحاله ليس بالضروره ان تكون الLUN المضافه نفس حجم الLUN الاساسية وكل الLUN بداخلها يجب ان تكون اما محميه عن طريق الmirrored or parity او غير محمية باستخدام RAID 0.

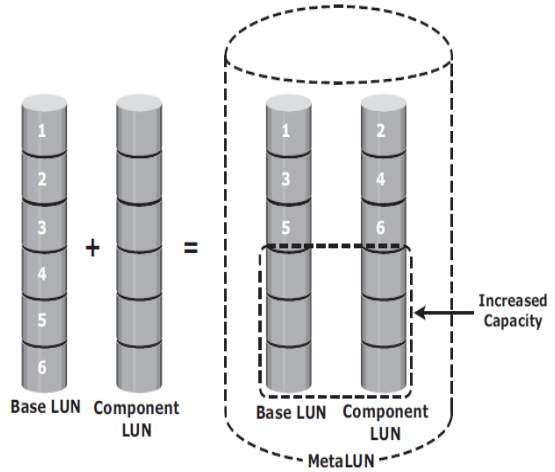
يمكن دمج الLUN فى انواع الRAID المختلفه على سبيل المثال RAID 10 LUN يمكن دمجها (Concatended) مع RAID 5 LUN, لكن فى RAID 0 LUN يمكن دمجها فقط مع RAID 0 LUN.



- **Striped expansion**

تعيد تقسيم ال LUN الاساسية مع الLUN الاضافية وفى هذه الحاله يجب ان تكون جميعها نفس الحجم وحالة الRAID

تحسن من مستوى الاداء من خلال زيادة عدد الديسكات التى تكون Striped



جميع ال LUNs فى ال concatended and striped يجب ان يكون لها نفس نوع الديسكات سواء جميعهم Fibre Channel او جميعهم ATA.

- Virtual Storage Provisioning

تمكن من انشاء LUN مع Capacity اعلى من التى يتم تخصيصها Physically. وهنا تعرف الLUN باسم thin

الLUN هنا لا تحتاج الى Physical Storage لكي يتم تخصيصها بشكل كامل الى الhosts, حيث يتم

التخصيص حسب الطلب من خلال shared pool على الPhysical capacity

ال Shared pool تتكون من physical disks وتكون مشابهه للRAID group حيث انها عباره عن مجموعه

من الDrives التى تنشأ من خلالها الLUNs.

ال Shared Poolايضا تدعم الsingle RAID و تحتوى على عدد كبير من الDrives ممكن ان تكون متجانسه

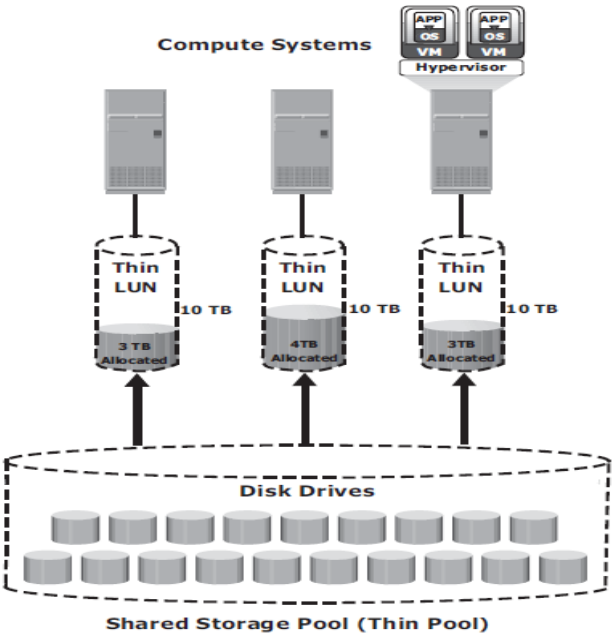
اي لها نفس النوع او غير متجانسة مثل الFlash, FC, SAS & SATA.

الvirtual storage تخصص Storage اكثر كفاءة الى الhosts لانها توفر capacity اعلى من الذى يوفره

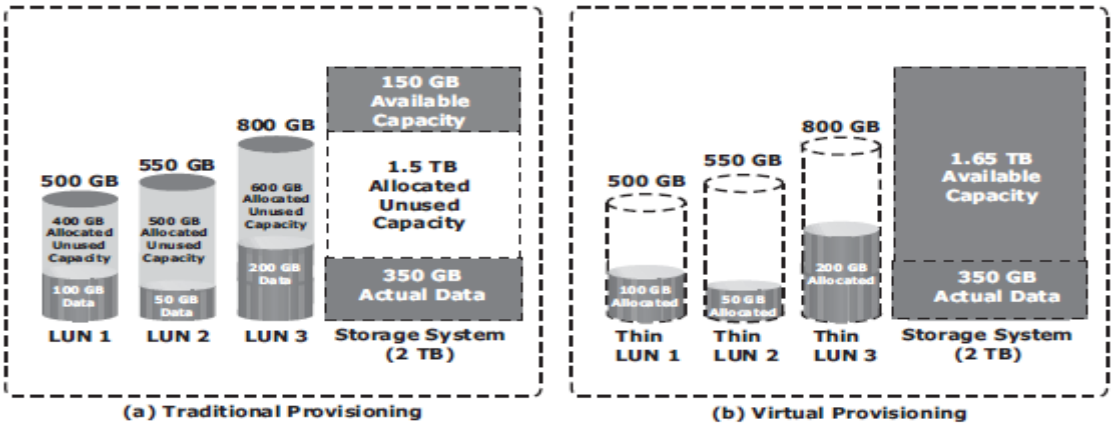
ال storage array العادى.

من الممكن انشاء العديد من الShared Pool داخل الStorage array, ايضا تحتوي الShared pool على

العديد من الthin LUNs.



- **Comparison between Virtual and Traditional Storage Provisioning**



نلاحظ هنا فى الTraditional Provisioning ان هناك 3 LUNs تم انشائها لواحد او اكثر من host وان المساحه الاجمالية لل Storage Systemهي 2TB تم تخصيص 500 GB الى LUN1 بينما المستخدم منها فعليا هو 100 GB والباقى غير مستخدم كذلك فى LUN3 تم تخصيص 550 GB بينما المستخدم هو 50 GB, ايضا فى LUN 3 تم تخصيص 800 GB والمستخدم 200 GB. اي انه فى المساحه الاجماليه للStorage System نجد ان 350 GB هي المستخدمه فقط و 1.5 TB مخصصه للLUNs لكن غير مستخدمه والمساحه المتاحه هي 150 GB فقط اما فى الVirtual Provisioning نجد ان هنا 3 Thin LUNs لا يوجد بهم مساحه غير مستخدمه من المساحه الاجمالية حيث اجمالى المستخدم هو 350 GB واجمالى المتاح هو 1.65 TB.

- **Use Cases for Thin and Traditional LUNs**

الThin LUN يعتبر اكثر كفاءة فى عملية توفير مساحة التخزين والذي يكون مناسب اكثر لبعض الapplication, ايضا مناسب للشركات من حيث توفير كمية الطاقه المستهلكه وسهولة الادارة.

اما الTraditional LUNs مناسبة اكثر للتطبيقات التى تحتاج اداء افضل حيث تكون اكثر تحكم ودقة فى البيانات المستخدمه, كذلك يمكنها انشاء العديد من الLUNs داخل RAID group مختلفه والذي يمنع حدوث workload. ومناسب اكثر للشركات الغير مهتمه بمساحة التخزين.

- LUN Masking

هي خاصية للتحكم فى الوصول الى الداتا حيث تخصص LUNs لكل Host على الStorage Array حيث تتحكم بهم بشكل مناسب لمنع استخدام الLUN من الhost الغير مصرح بها داخل الshared environment.

على سبيل المثال لو انه يوجد داتا لقسم معين كالمبيعات داخل الشركة بدون LUN masking فانه من السهل الوصول الى هذه البيانات من قبل اي قسم اخر.

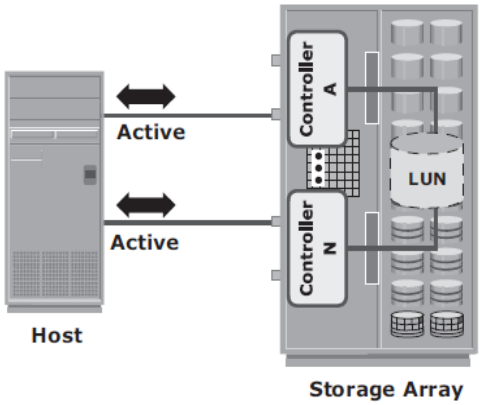
Types of Intelligent Storage Systems

ال Intelligent Storage عامة تندرج تحت نوعين رئيسين اما High-end storage systems او Midrange storage systems

- High-end storage systems

الHigh-end Storage يعرف ب active-active arrays, وغالبا يهدف الى enterprise application الكبيرة خاصة وانه مصمم بعدد كبير من الcontrollers وال cashe memory.

الactive-active تعنى ان الhost بامكانه ارسال اي I/O request الى الLUNs خلال اي controller متاح.



اهم مميزات ال High-end Storage:

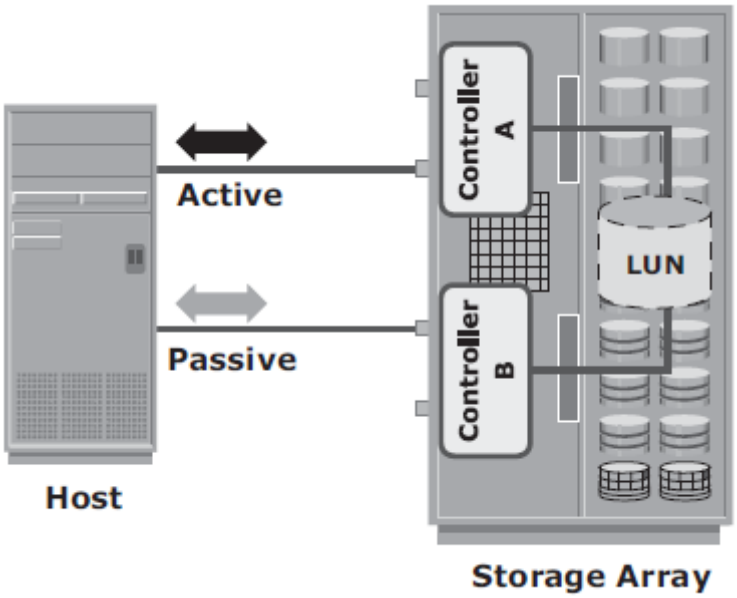
- Large storage capacity
- Large amounts of cache to service host I/Os optimally
- Fault tolerance architecture to improve data availability
- Connectivity to mainframe computers and open systems hosts
- Availability of multiple front-end ports and interface protocols to serve many of hosts
- Availability of multiple back-end controllers to manage disk processing
- Scalability to support increased connectivity, performance, and storage capacity requirements
- Ability to handle large amounts of concurrent I/Os from many hosts and applications
- Support for array-based local and remote data replication

- Midrange Storage Systems

الMidrange-end Storage تعرف ب active-passive arrays, مناسبة اكثر للenterprise applications الصغيره او المتوسطة, حيث انها توفر الStorage solution بتكاليف اقل.

ال active-passive تمكن الhost من الوصول الى ال LUN فقط من خلال الcontroller الذي يحتوى على هذه الLUN كما موضح بالصورة.

ال Midrange Storage مصمم ب اتنين controller كل منهما يحتوى على host interfaces, cache, RAID controllers و Interface الى الDisk drive.



Concepts in Practice: EMC Symmetrix and VNX

ال EMC Symmetrix هو active-active storage array وهو مناسب جدا للشركات التى تتطلب مستوى عالى من الاداء والخدمه بالنسبه الى الstorage system حيث يمكنه دعم عدد ضخم من ال application workloads الغير متوقعه.

اما ال EMC VNX Storage هي active-passive array تعتبر storage platform موحده حيث توفر تخزين الداتا لكل من block, file & object-based فى نفس الarray, تعتبر مناسبة للapplication workloads المتوقعه.

- EMC Symmetrix Storage Array

يوفر اعلى مستوى من الاداء والمساحه التخزينية بالنسبة للـ Storage solution حيث يوفر قدر عالى من الscalability لمواجهة متطلبات الI/O workloads الغير متوقعه, يحتوى ايضا على Symmetrix Virtual Matrix (VMAX) series.

The EMC VMAX هو Storage Platform صمم لمواجهة نمو وتطور مستقبل الstorage بالنسبة لمجال الVirtualization. ومن اهم مميزاته:

- Incrementally scalable to 2,400 disks
- Supports up to 8 VMAX engines (Each VMAX engine contains a pair of directors.
- Supports flash drives, fully automated storage tiering (FAST), virtual provisioning, and Cloud computing
- Supports up to 1 TB of global cache memory
- Supports FC, iSCSI, GigE, and FICON for host connectivity
- Supports RAID levels 1, 1+0, 5, and 6
- Supports storage-based replication through EMC TimeFinder and EMC SRDF
- Highly fault-tolerant design that allows nondisruptive upgrades and full component-level redundancy with hot-swappable replacements

- EMC Symmetrix VMAX Component

يحتوى على VMAX على منطقة واحده للSystem اما الStorage تصل الى 10 مناطق كل منها تدعم حتى 16 Drives array enclosures (DAEs) وكل enclosure من الممكن ان يحتوى الى 15 drives.

اما منطقة الSystem تحتوى على مكونات النظام مثل VMAX Engines, Matrix Interface Board Enclosure (MIBE), standby power supply (SPS) modules, and service processor.

- Symmetrix VMAX Architecture

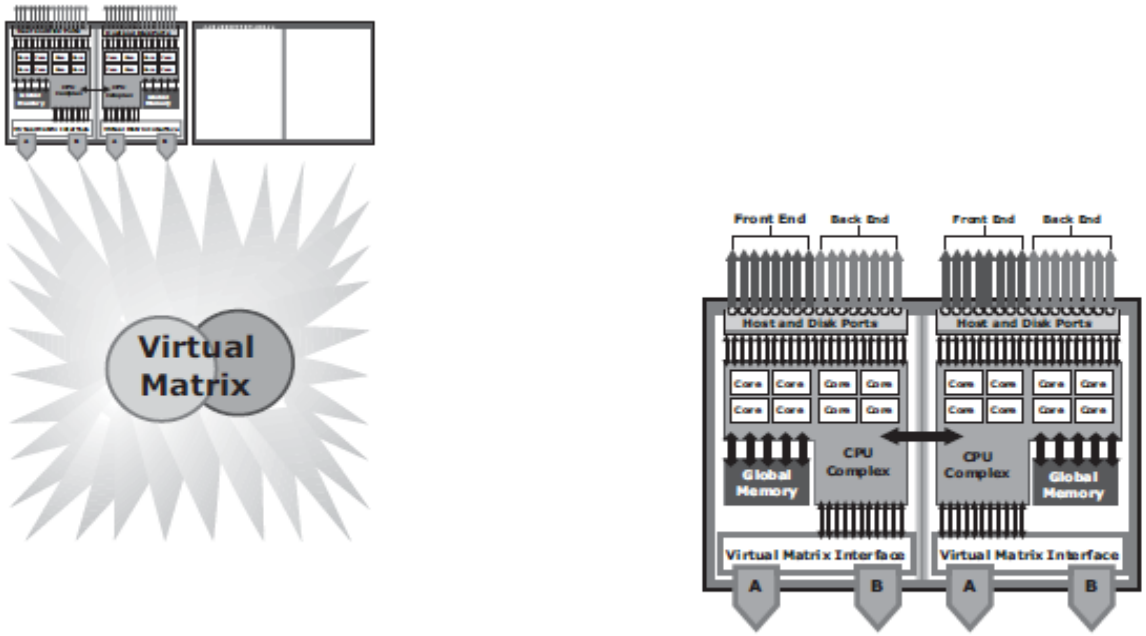


Figure 4-13: VMAX architecture

Part II

Storage Networking Technologies

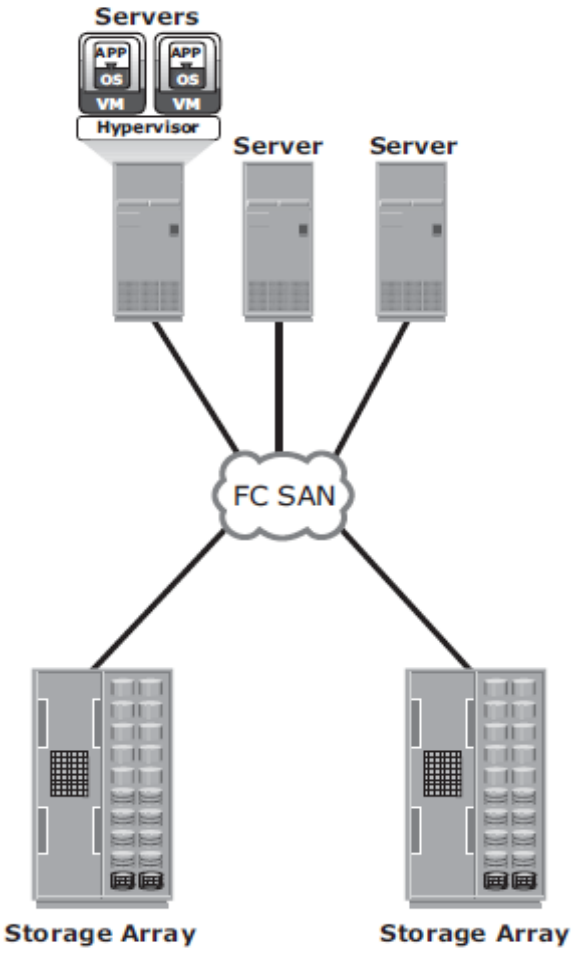
Chapter 5: Fibre Channel Storage Area Networks

Fibre Channel: Overview

تقنية جديدة طورت لتتناسب مع زيادة سرعات نقل البيانات بين الServers المختلفة وال Storage Systems حيث يمكنها من ربط وحدات التخزين مع السيرفرات بإستخدام كابلات الفايبر او الكابلات الضوئية والتي تسمح لنا بنقل البيانات بسرعة عالية جداً تصل إلى 16 Gbps. كما أنها تمكنك من نقل البيانات للمسافات الطويلة.

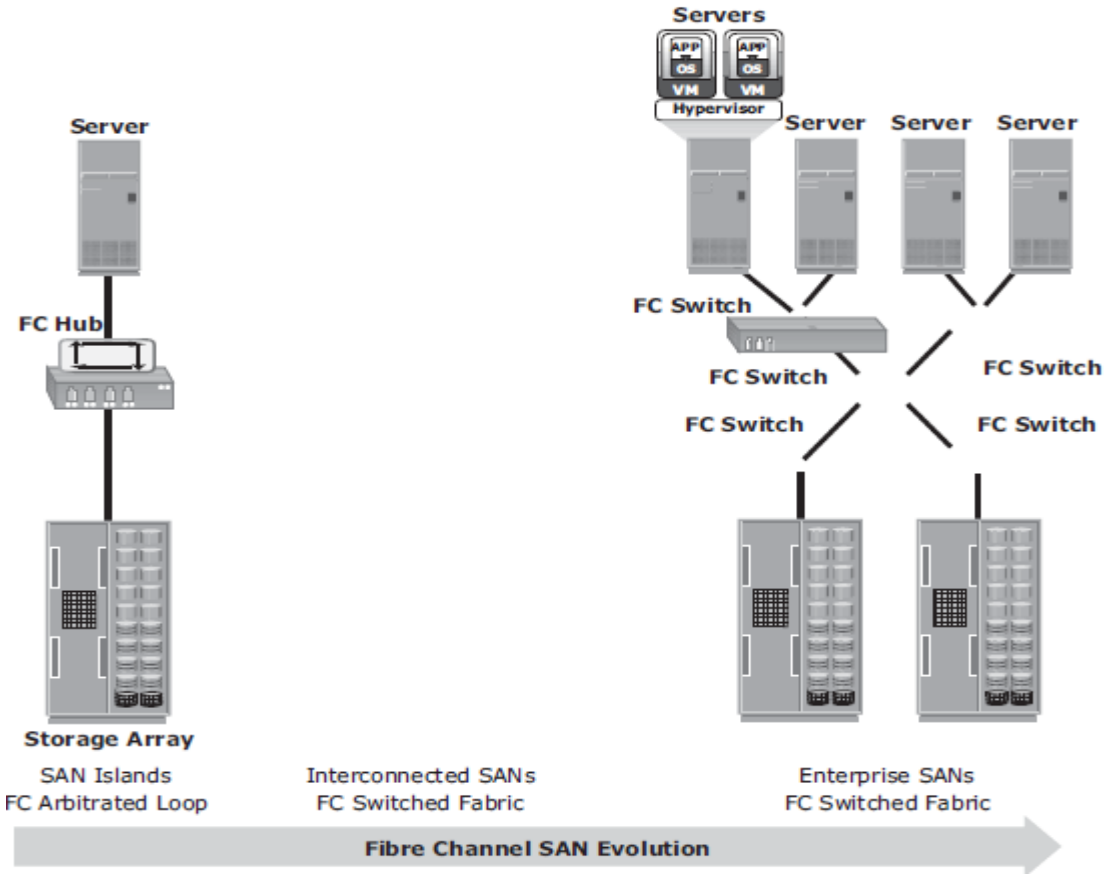
The SAN and Its Evolution

الSAN هي شبكة يتم بها نقل البيانات بين ال Host وال Storage باستخدام بروتوكول الFibre Channel فيمكنه نقل البيانات بسرعات عاليه ويمكنه من دمج ومركزية الStorage لكي تكون متاحه الى العديد من Servers. وتوفر ايضا تحسين استغلال موارد الStorage وسهولة ادارتها مقارنة بالDAS, حيث تعمل بنظام Block-Level Storage System.



فى الصورة السابقة نجد مثال بسيط لل FC SAN حيث يحتوى على مجموعه من الhosts موصوله على Storage devices باستخدام FC hub هذه التكوين يعرف بـ Fibre Channel Arbitrated Loop (FC-AL) لكن استخدام الhub فى شبكة الSAN يكون له limited connectivity & bandwidth.

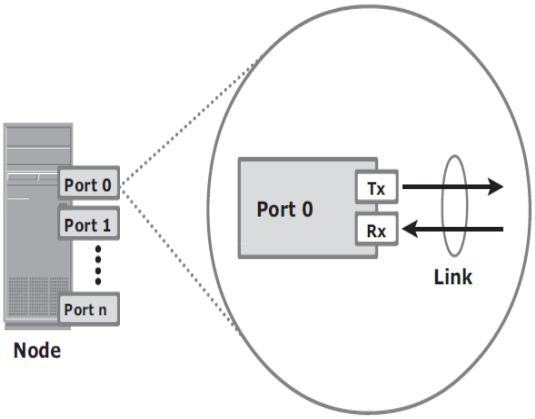
ولكن جاء تطور شبكات الSAN باستخدام الFC Switches تعرف ايضا بالMDS switch حيث قامت بتحسين وتطويرسرعة واداء نقل البيانات داخل الشبكة.



Components of FC SAN

- Node Ports

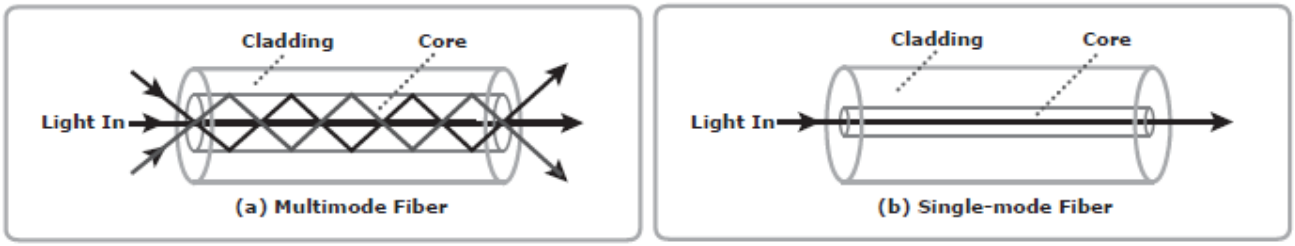
تحتوى الSAN Network على العديد من الPorts والتى تساعد فى نقل البيانات بين الHosts والStorage داخل ال Fibre Channel وتسمى الاجهزه المتصله بالStorage System بال Node, كل node تعتبر source او destination للمعلومات وتحتوى على port او اكثر يمثل ال physical interface لتبادل البيانات مع ال nodes الاخرى. يعتبر هذا البورت جزء من الhost adapter الذى يسمى بال HBA.



- Cables and Connectors

يستخدم فى شبكات الSAN الكابلات النحاسية copper cable للمسافات القصيره for MAX 20 m, او الالياف الضوئية Fiber Optics للمسافات البعيدة ويوجد منه نوعين تعتمد عليها الSFP الموصله عليها:

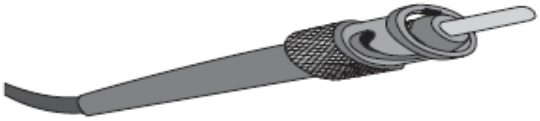
- Multi-mode Fiber (MMF): يستخدم للمسافات القصيرة up to 500 m
- Single-Mode Fiber (SMF): يستخدم للمسافات البعيدة up to 10 km



اما الConnectors الموصوله فى نهاية كل cable لتمكنه من الاتصال بالبورت لها انواع ايضا بالنسبه لل Fiber cables مثل Standard connector (SC) & Lucent connector (LC) تستخدم فى الFiber Optic cables وايضا Straight Tip (ST) تستخدم فى fiber patch panels.

(a) Standard Connector (SC)

(b) Lucent Connector (LC)



(c) Straight Tip Connector (ST)

- Interconnect Devices

تتكون من الFC Hubs, Switches & Directors

- **Hub:** تستخدم فى الاتصال داخل Fibre Channel Arbitrated Loop (FC-AL) حيث تقوم بتوصيل الnodes داخل physical star topology, هذا معناه ان كل الnodes تشارك المعلومات مع بعضها داخل loop لان الداتا تمر على كل الconnections point.
- **Switches:** اكثر كفاءة من الhub حيث يقوم بتوجيه البيانات فقط الىPhysical port الاخر بدون مشاركه للbandwidth بين الnodes حيث كل منها لها مسار اتصال مخصص.

  تصميم الswitch يكون اما fixed port count او بشكل modular حيث فى الmodular ممكن زيادة عدد الports بتركيب port card اضافى
- **Directors:** هي تعتبر high-end switches حيث يوجد بها عدد اكثر من ال ports اكثر بقدرات اكثر كفاءة. اما التصميم فجميعها modular يمكن زيادة الports بتركيب port cards او blades لل chassis

- **SAN Management Software**

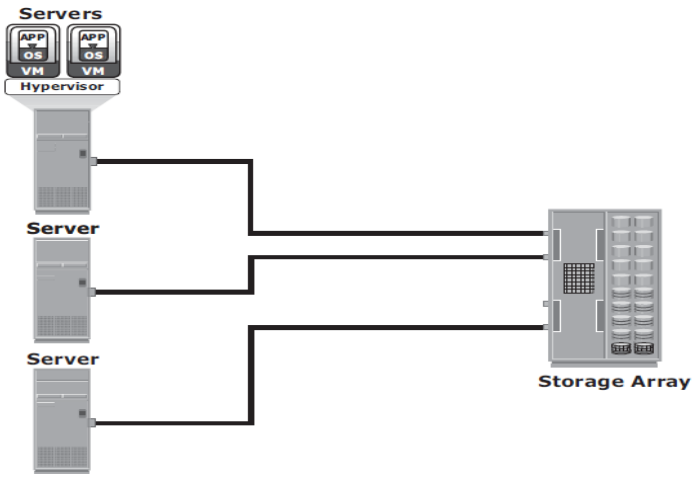
هو سوفتوير يستخدم لادارة الInterfaces الموصوله بين الHosts والStorage Arrays, حيث يقوم بادارة موارد شبكة الSAN عن طريق وحدة تحكم مركزية مثل ادارة الوظائف المختلفة وطرق التوصيل المختلفة بين ال storage system والservers ايضا مراقبة ومتابعة التنبيهات للاجهزه الجديدة وال Zoning.

FC Connectivity

تمثل التوصيلات داخل الFC SAN Architecture مثل:

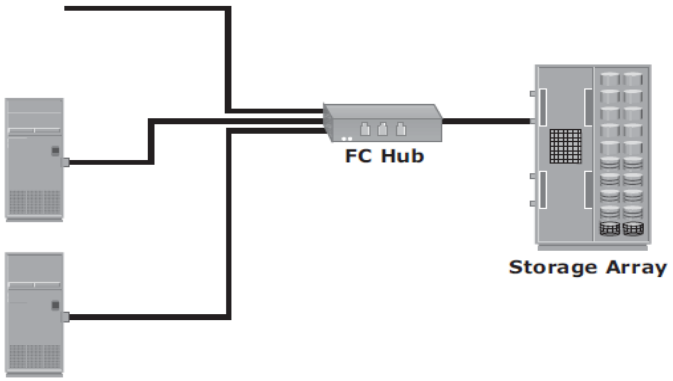
- **Point-to-Point**

ابسط مثال لطريقة التوصيل داخل شبكة الSAN حيث يتم توصيل الhost بالstorage مباشرةً, يسمح بنقل البيانات بين الnodes لكن يكون له عدد اتصال محدود وهذا يعتمد على عدد المنافذ الموجودة في Fiber Channel الخاصة بـ Storage Unit.



- **Fibre Channel Arbitrated Loop**

فى هذا التصميم تكون الاجهزه مرتبطه بـShared Loop لانه يكون مصمم للعمل فى token ring toplogy او star topology وهنا يتم استخدام FC-Hub حيث كل الاجهزه تشارك الاجهزه الاخرى فى الI/O operation داخل الloop ويجب على جهاز واحد فقط التحكم فى الLoop لتنفيذ اوامر الI/O operation داخله.



- Fibre Channel Switched Fabric

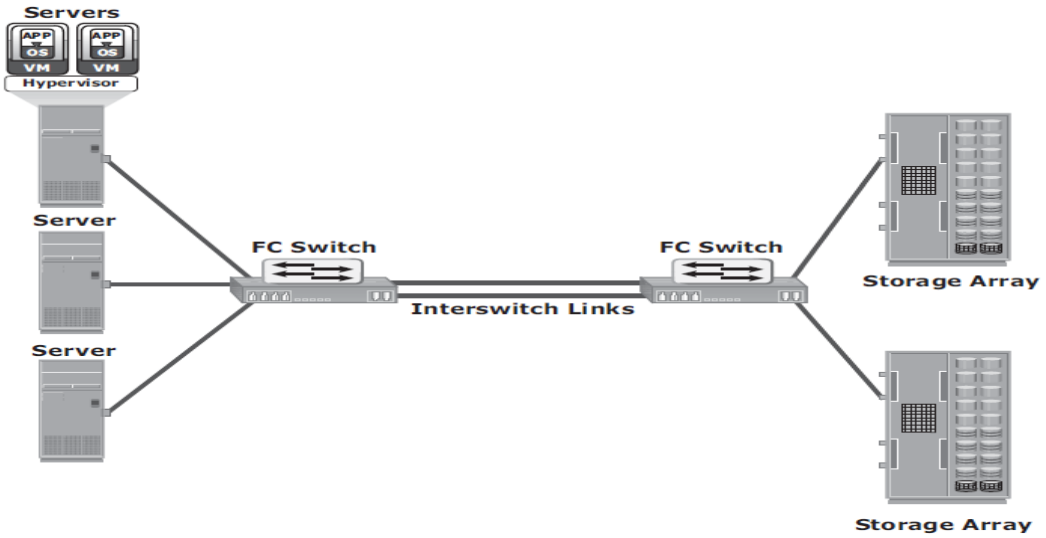
هنا يتم استخدام FC-Switch يعرف ايضا بالFabric connect حيث يقوم بربط جميع الnodes بالstorage داخل الشبكة.

كل سويتش داخل الFabric يكون له unique domain identifier والذى هو جزء من ال fabric addressing scheme.

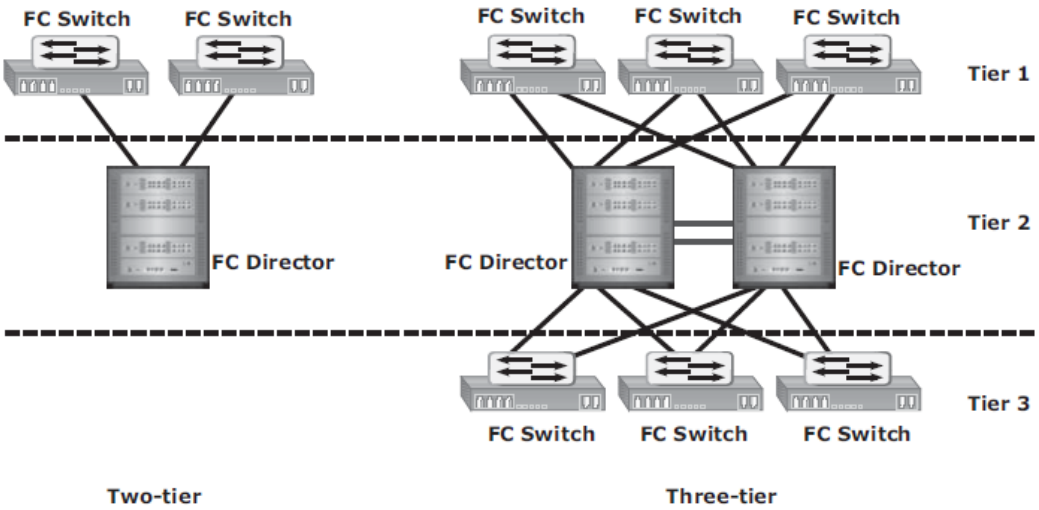
ايضا الnodes لا تشارك فى الloop حيث يتم نقل البيانات داخل مسار خاص بين الnodes.

كل port داخل الfabric يكون له 24-bit Fibre channel address.

الربط بين اتنين سويتش داخل الswitched fabric تسمى بال Interswitch Link (ISL) وتستخدم فى نقل البيانات بين الhost والstorage وادارة الترافيك بين الswitches وتستخدم ايضا لتوسيع الشبكة لتوصيل العديد من الnodes.



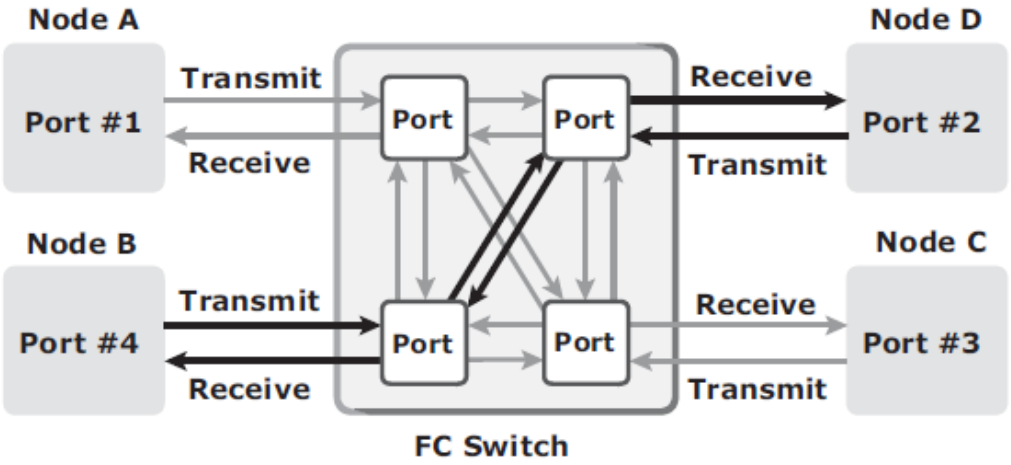
يمكن وصف الFabric بانه مجموعه من الطبقات عدد هذه الطبقات تحدد بعدد السويتش التى يتم مرور البيانات عليها بين نقطتين ومنها تحدد كيف يتم التوصيل بين الHost وال Storage, فانه عند زيادة عدد الطبيقات بينهم داخل الfabric فالبتالى تزيد المسافه فيزيد الوقت المستغرق فى وصول البيانات



- **FC-SW Transmission**

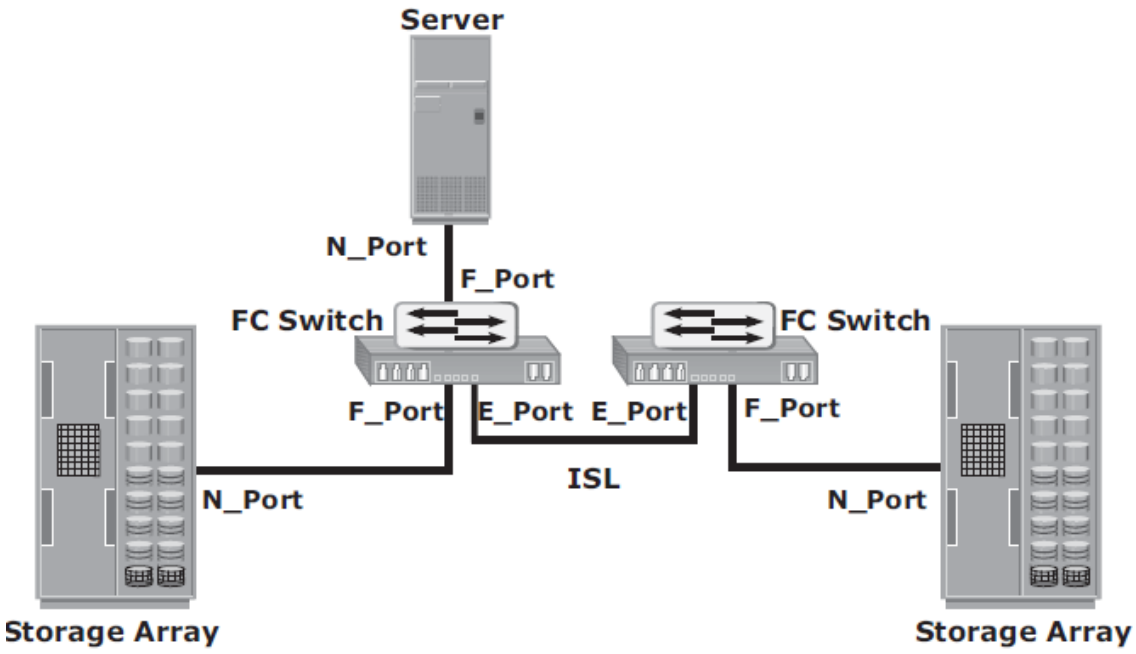
الFC-Switches تستخدم الswitches التى تقوم بتحويل الداتا بين الnodes مباشرة من خلال الPorts.

حيث تقوم بتحويل الFrames بين الsource وال Destination بواسطة الfabric.



Switched Fabric Ports

- **N_Port:** يعرف بالNode Port وهو البورت على ال Host (HBA) او الStorage array موصول بالسويتش على الFabric Port.
- **E_Port:** يعرف بال Expansion port وهو البورت على الFC switch الموصول بFC Switch اخرعن طريق ال ISL
- **F_Port:** يعرف بالFabric port وهو البورت على الFC Switch وموصول بال N_Port.
- **G_Port:** هو Generic Port على السويتش يعمل ك E_Port او F_Port حيث تحدد وظيفته تلقائيا اثنا ال initialization.



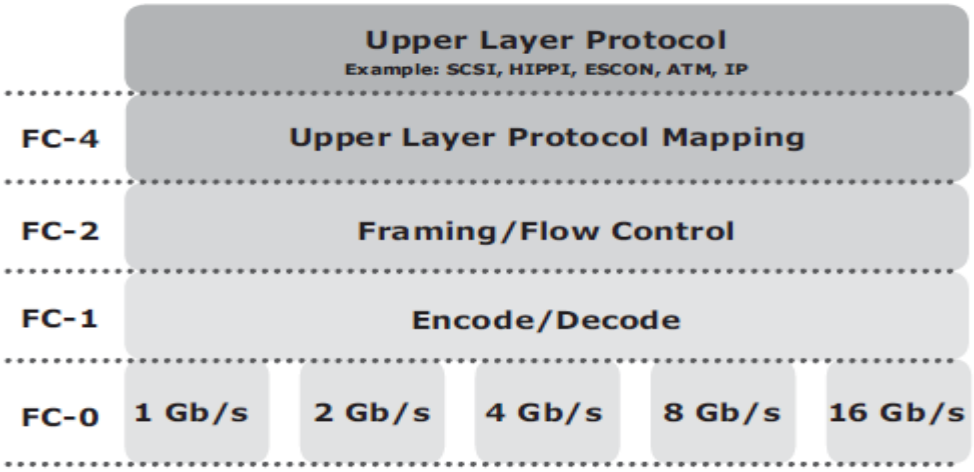
Fibre Channel Architecture

كما ذكرنا ان الFibre Channel تقنية جديدة ظهرت لنقل البيانات بسرعات عالية بين الHost والStorage من خلال شبكات الSAN باستخدام Fibre Channel Protocol (FCP).

ولكن هناك مراحل او طبقات تمر بها الداتا فى شبكات ال SAN بين Host والStorage . مثل الOSI Layers فى شبكات الLAN بين الSource وال Destination.

- Fibre Channel Protocol Stack

الFCP يقوم بتوصيل البيانات داخل الشبكة خلال 5 مراحل او طبقات



- **FC-4 Layer**

هذه الطبقة تحدد الApplication Interface والطريقه التى يتم بها التوصيل بين الطبقه العليا (ULPs) الى الطبقات السفلى باستخدام البروتوكولات المختلفة مثل SCCI, HIPPI, ESCON, ATM & IP.

- **FC-2 Layer**

توضح حجم او شكل الداتا داخل الفريم الذى يتم ارساله داخل ال Fibre Channel حيث انها تحتوى الداتا على Control Frame & Data Frame حيث فى الData Frame يتم تحديد حجم البيانات من خلال الExchange & sequence, اما فى الControl Frame يتم تحديد حجم البيانات من خلال عمل negotiate لل Classes و Flow Control لل Fibre Channel.

- **FC-1 Layer**

لنقل البيانات بسرعات عالية يتم تشفير الداتا قبل ارسالها وفك الشفرة عند الاستلام وتتم هذه المرحله بطريقة معينه بحيث يتم تحديد عدد الbits التي سيتم نقلها وتحديد بداية ال Frame (SOF) ونهاية الFrame (EOF) . ايضا طريقة التحكم فى الداتا كتحديد بداية الHeader ونهاية الHeader

- **FC-0 Layer**

الطبقة التى تتم فيها Transmission للBits عن طريق الكابلات النحاسية copper cable للمسافات القصيره for MAX 20 m, او الالياف الضوئية Fiber Optics للمسافات البعيدة

- Fibre Channel Addressing

الFC Address يتم تحديده تلقائيا عندما يسجل الN_Port دخوله على الFabric ويكون له هيئه معينه حيث يحتوى على 24-bit.

الجزء الاول منها على الDomain ID وهو unique number لكل switch وهناك 239 عنوان فقط مخصص للDomain ID لان هناك بعض العناوين محجوزه للfabric management service على سبيل المثال FFFFFC محجوز للname server ايضا FFFFFE مخصص لل fabric login service.

اما الجزء الثانى Area ID يحدد مجموعه من الswitch port تستخدم للتوصيل الى الnodes.

اما الجزء الاخير Port ID يحدد عدد الports داخل المجموعه.

وبالتالى اقصى عدد يمكن توصيله لل nodes الى الswitched fabric يحدد كالتالى

239 domains x 256 areas x 256 ports = 15,663,104

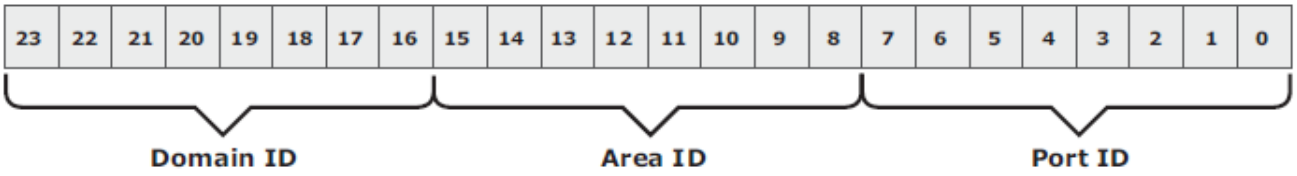


Figure 5-13: 24-bit FC address of N_Port

- World Wide Names

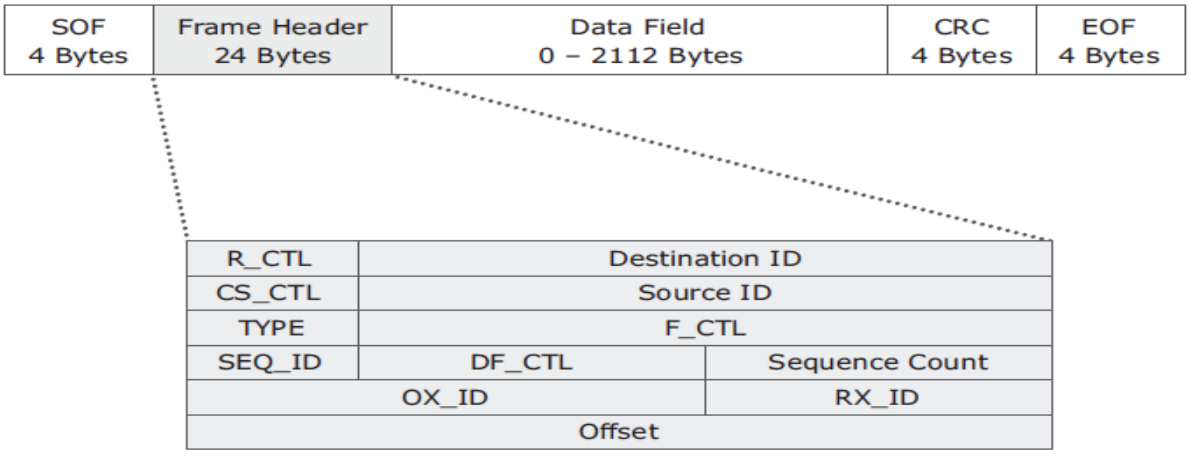
كل جهاز داخل الFC environment يحتوى على عنوان يسمى بال WWN شبيه للMAC address ويوجد منه نوعين (World Wide Node Name (WWNN) & World Wide Port Name (WWPN حيث ان الWWNN الذي يكون على الarray وال WWPN يكون موجود على الHPA للسيرفر او الجهاز.

World Wide Name - Array															
5	0	0	6	0	1	6	0	0	0	6	0	0	1	B	2
0101	0000	0000	0110	0000	0001	0110	0000	0000	0000	0110	0000	0000	0001	1011	0010
Format Type	Company ID 24 bits						Port	Model Seed 32 bits							

World Wide Name - HBA															
1	0	0	0	0	0	0	0	c	9	2	0	d	c	4	0
Format Type	Reserved 12 bits			Company ID 24 bits						Company Specific 24 bits					

- FC Frame

حجم ال Frame فى الFibre Channel لا يتعدى 2112 Bytes مضافا اليه الHeader تصبح 2148 Bytes ويتكون من 5 اجزاء هي start of frame (SOF), frame header, data field, cyclic redundancy check (CRC), and end of frame (EOF).



ال

SOF & EOF فقط لتحديد بداية ونهاية الفريم, اما الFrame Header حجمه 24-Bytes ويحتوى على كل العناوين والمعلومات للفريم مثل ال(S_ID) Source ID والـ(D_ID) Destination ID و Sequence ID (SEQ_ID) و Sequence count (SEQ_CNT) و(OX_ID) Originating Exchange ID و Responder Exchange ID (RX_ID) بالاضافه الى بعض control field مثل:

Routing Control (R_CTL), Class Specific Control (CS_CTL), TYPE, Data Field Control (DF_CTL) & Frame Control (F_CTL).

اما الData Field الحد الاقصى للبيانات على الفريم وحجمها 2112-Bytes, والCRC تستخدم لعمل CHECK على الداتا لتسهيل اكتشاف الاخطاء.

- Structure and Organization of FC Data

عند نقل البيانات داخل الFC network فانها يكون لها شكل وحجم معين داخل الفريم ويتم تحديد حجم الداتا عن طريق مراحل بين الhost والarray.

- **Exchange:** عملية تبادل تتم بين الnodes لتحديد وادارة المعلومات مثل نوع البروتوكول المستخدم لنقل البيانات هذه المعلومات تسمى بالinformation units حيث كل unit تحتوى على مجموعه من ال Sequence بترتيب معين ولا يتم ارسال الExchange الى الReceiver حتى تكتمل الSequence.
- **Sequence:** هي عباره عن مجموعه من الFrames المرسله من port الى اخر.
- **Frame:** هي الجزء الاساسي للبيانات والحد الاقصى لحجم الفريم هي 2112 Bytes.

- Flow Control

التحكم فى نقل البيانات يضمن وصول البيانات كامله وبشكل سليم وهناك نوعان للتحكم فى نقل البيانات وهي End-to-End Credit (EE_Credit) & Buffer-to-Buffer Credit (BB_Credit).

وتتم بعدة خطوات :

1- يرسل الSender لل Receiver طلب Request عباره عن Write or Read.

2- الReceiver يرسل Credit الى الSender والCredit هي الfree Buffer او الCapacity للReceiver لاستلام الداتا على سبيل المثال Credit 3.

3- الSender ينشأ اتنين Counters اولا وهم Credit (CR)=3 والثاني Receiver Ready (R_Ready)=0

4- الSender سيبدأ بارسال جزء من الداتا ويخصم 1 من الCredit CR=2

5- الReceiver يرسل Acknowledgment بالاستقبال R_Ready=1

6- الى ان تكتمل الترافيك عندما تكون الCR=0 وال R_Ready=3

- Classes of Service

الFC Standard تحدد انواع مختلفه من الclasses وال Services للعمل مع التطبيقات المختلفة.

Table 5-1: FC Class of Services

	CLASS 1	CLASS 2	CLASS 3
Communication type	Dedicated connection	Nondedicated connection	Nondedicated connection
Flow control	End-to-end credit	End-to-end credit B-to-B credit	B-to-B credit
Frame delivery	In order delivery	Order not guaranteed	Order not guaranteed
Frame acknowledgment	Acknowledged	Acknowledged	Not acknowledged
Multiplexing	No	Yes	Yes
Bandwidth utilization	Poor	Moderate	High

Fabric Services

جميع ال FC Switches تقدم خدمات موحده بغض النظر عن الشركة المصنعه كما هو محدد فى Fibre Channel Standard , على سبيل المثال خدمات مثل Fabric Login Server, Fabric Controller, Name Server, and Management Server

Fabric Login Server محدد بعنوان FFFFFE ويستخدم اثناء عملية تسجيل دخول الnodes الى الfabric

Name Server محدده ايضا بعنوان ثابت FFFFFC وهي المسؤوله عن عملية تسجيل اسم الnode ports

Fabric Controller ومعرفة بعنوان FFFFFD وهو مسؤول عن ادارة وتوزيع ال Registered State Change Notifications (RSCNs) الى الnode port المسجلة.

Management Server معرفه بعنوان FFFFFA يستخدم لتفعيل FC SAN management software لادارة واسترداد المعلومات.

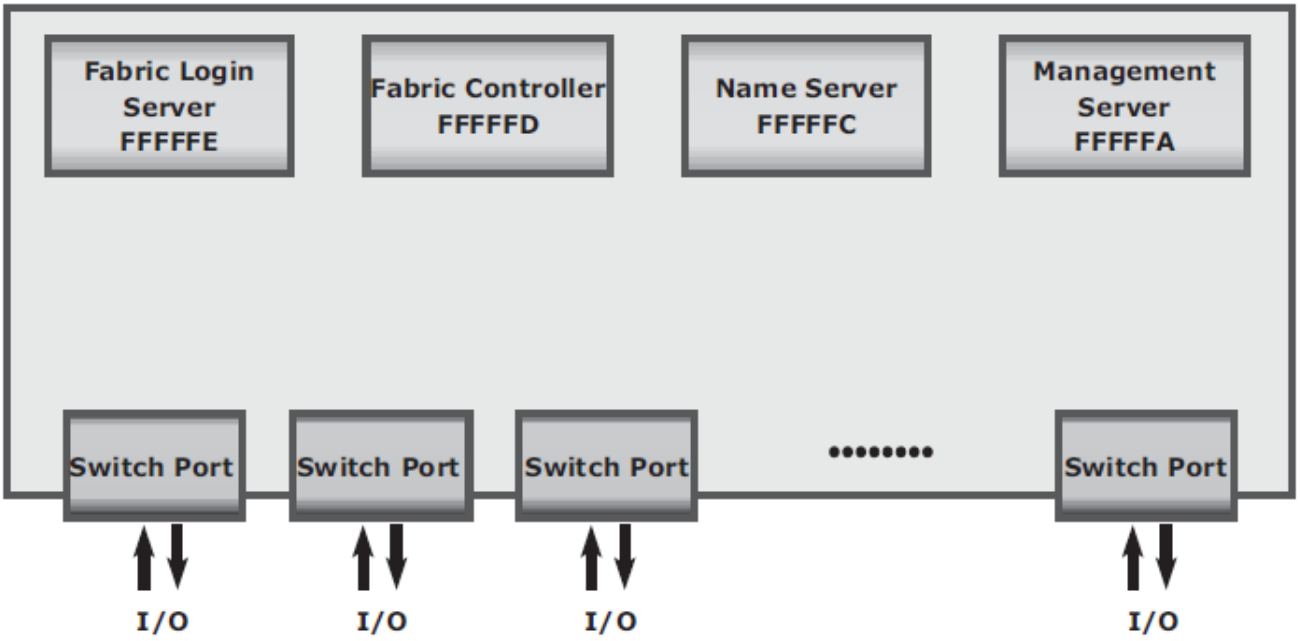


Figure 5-16: Fabric services provided by FC switches

Switched Fabric Login Types

يوجد مراحل عند تسجيل دخول الPort فى الFabric Service

- **Fabric Login (FLOGI)**

تتم بين الN_Port و ال F_Port حيث يرسل الnode طلب يحتوى على الWWNN & WWPN الى الlogin server service. عندها ال switch fabric login يقوم بالموافقه Accept (ELS Reply) على الطلب وينشأ FCID خاص بالN_Port ويكون associated على نفس البورت.

- **Port login (PLOGI)**

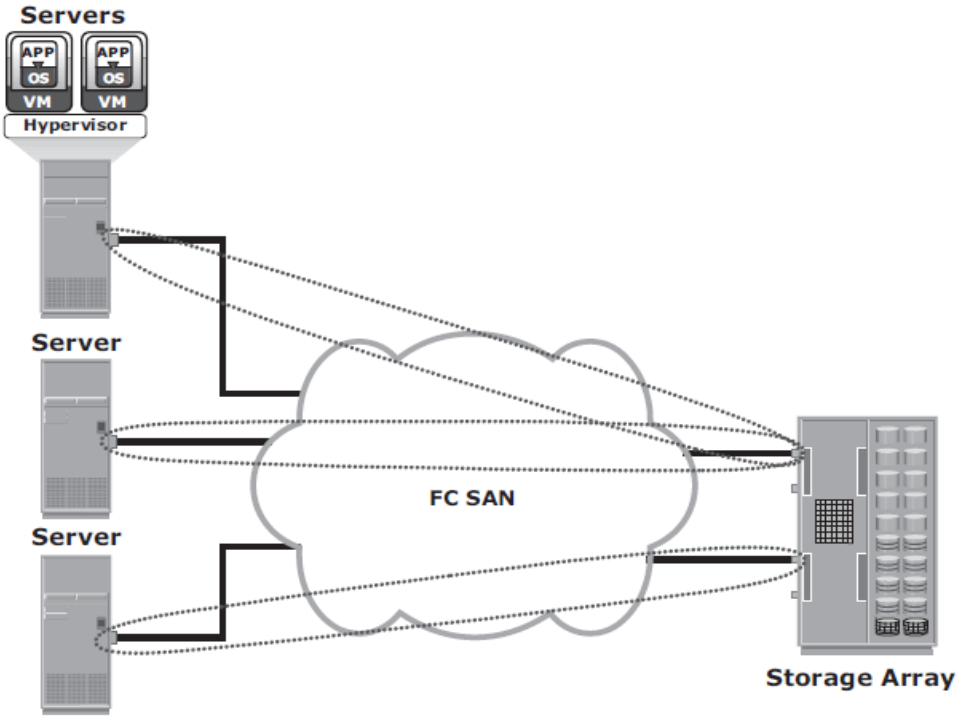
الInitiator يقوم بفتح session مع الname server وهي service على السويتش يسجل عليها الداتابيز الخاصه بالPort ID ومن الممكن ان يكون نفس ال switch fabric login او مختلف.

- **Process login (PRLI)**

تتم بين اتنين N_Ports وهذه المرحله لها علاقه بالطبقة FC-4 ULPs مثل فاذا كان الULP عباره عن SCSI فيتم التبادل للخدمات التى لها علاقه بال SCSI.

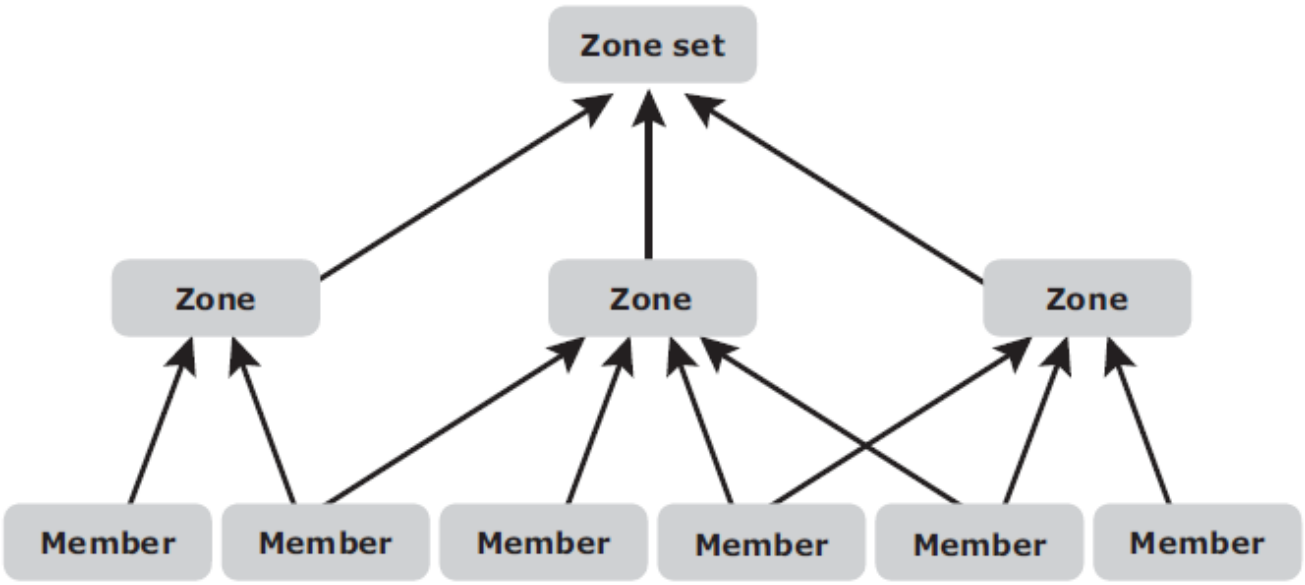
Zoning

يتم تقسيم الFC Switch الى Logical group تحتوى على node ports للاتصال مع بعضهما البعض داخل المجموعه ويسمى هذا التقسيم بال Zoning.



نلاحظ عند حدوث اي تغيير داخل الname server database يقوم ال fabric controller بارسال Registered State Change Notification (RSCN) الى جميع الnodes داخل الfabric اضافة الىnodes الغير متعلقه بنتائج هذا التغيير وهذا يؤدى الى زيادة الfabric-management traffic الذى يؤثر بشكل كبير على عملية الprocessing داخل الfabric.

هنا الZoning يساعد على تقليل عدد ال RSCNs داخل الFabric.



- Types of Zoning

- **Port Zoning**

يتم على مستوى الPort, تخصيص كل بورت الى zone محدده. لكن من عيوبه انه عند تغيير البورت لابد من تغيير الconfiguration.

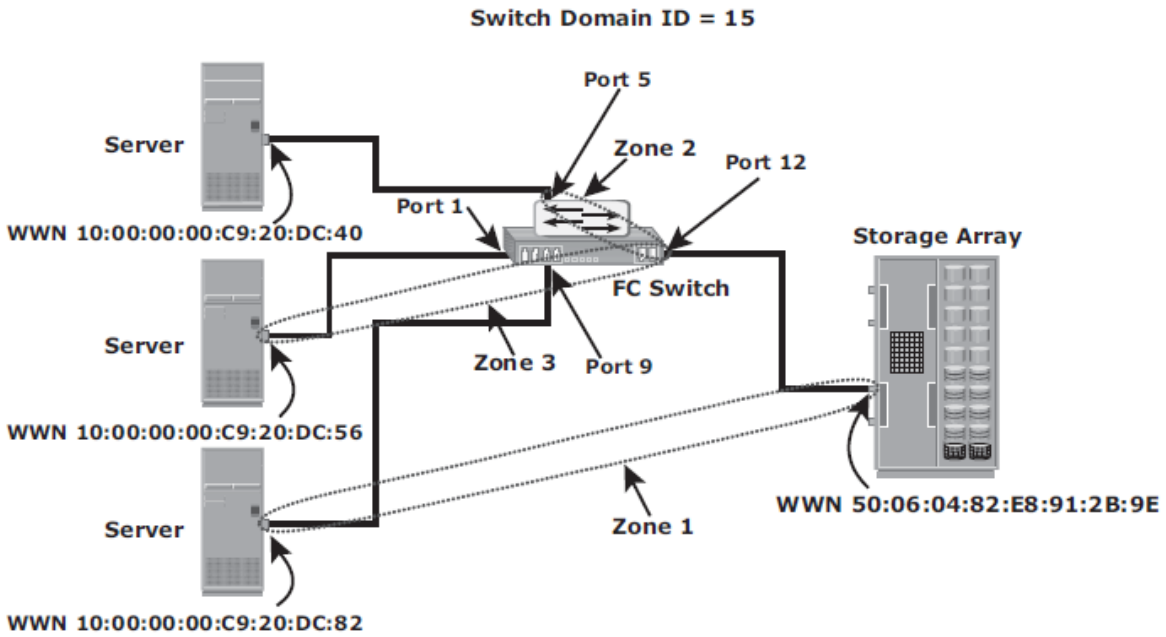
- **WWN zoning**

تتم على مستوى علىWWN, تخصيص كل WWN للinitiator الى Zone.

- **Mixed zoning**

تتم على مستوى الWWN & Port او ممكن استخدم الFCID

يوجد صعوبة فى عملية الZoning بالنسبه للWWN وFCID اذا كان هناك اكتر منinitiator كل منها يحتوى على اكثر من بورت فيكون العدد كبير وتم حل هذه المشكلة بعمل مايسمى بال Device Aliases اما على مستوى الarray يتم تقسيمها الى LUN Masking لتوفير ايضا الSecurity وكل Initiator يستطيع الوصول لل LUN الخاصه به عن طريق عمل mapping بين الWWN له والLUN ID



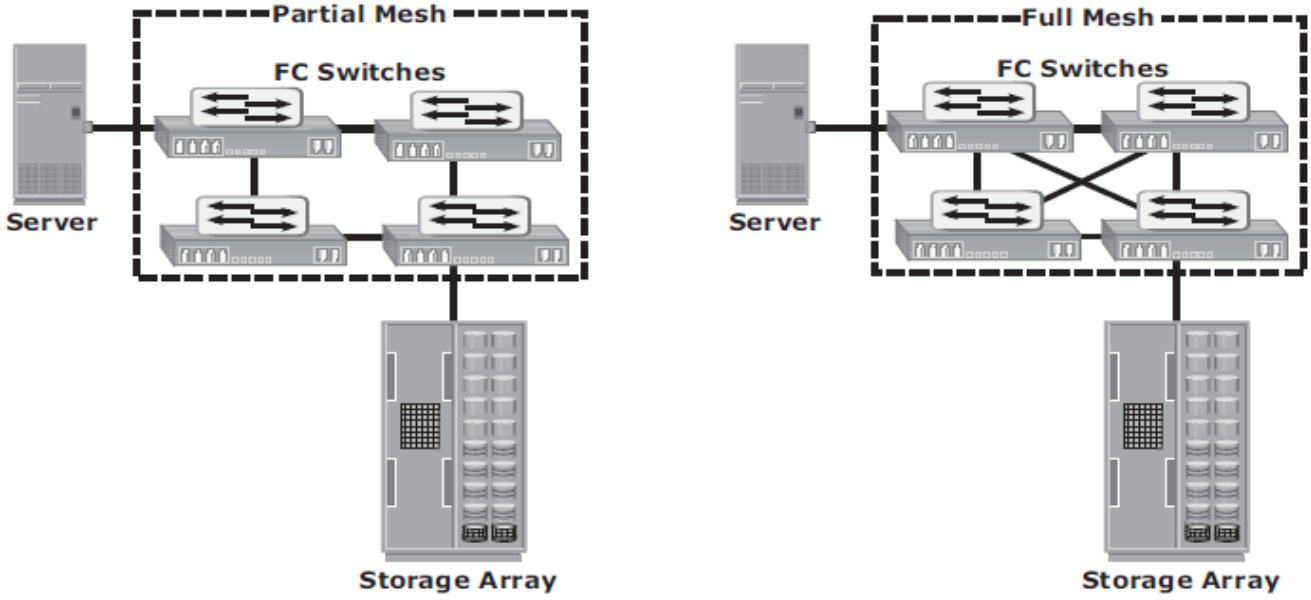
FC SAN Topologies

هناك انواع مختلفة لطرق تصميم شبكات الFC SAN منها:

- Mesh Topology

يوجد منه نوعين هما الFull Mesh & Partial Mesh

- Full Mesh: يكون كل switch مرتبط بجميع الswitches الاخرى داخل الشبكة, ويكون مناسب عندما تكون الشبكة صغيره حيث يحتوى الى حد اقصى 4 switches or directors, ايضا يحتوى على حد اقصى one ISL تستخدم لنقل الترافيك بين الhost-to-storage. ومع زيادة عدد السويتشات داخلها تقل عدد الports المتاحه للnodes connectivity.
- Partial Mesh: يكون هناك العديد من ال ISL بين الhost-to-storage لنقل الترافيك.



- Core-Edge Fabric

هذا التصميم يحتوى على نوعين من الSwitches هما الedge tier & core tier

حيث فى الedge tier تتكون من مجموعه من السويتشات التى تسمح باضافة وتوصيل العديد من الhosts وكل سويتش داخل الedge tier يكون متصل بسويتش اخر داخل الcore tier من خلال ISL.

اما الcore tier يحتوى غالبا على الDirectors التى تحتوى على high fabric availability وذلك لان جميع الترافيك لابد ان تمر او تصل الى هذه الطبقة, ايضا جميع الStorage تكون متصله بهذه الطبقة عن طريق ISL بالاضافه الى الhost التى تحتاج اداء عالى فى سرعة نقل البيانات.

- ***Benefits and Limitations of Core-Edge Fabric***

الcore-edge لديه نقطة اتصال واحده الى كل الStorage devices داخل الfabric هذا لان الترافيك تمر على نمط محدد من الedge-to-core مما يسهل عملية حساب الtraffic load, ايضا لان كل بورت على الswitch مخصص لstorage or host مما يسهل عملية تحديد الnetwork resources

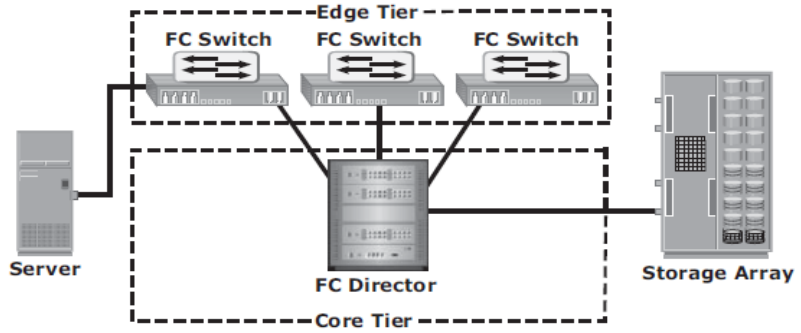


Figure 5-21: Single-core topology

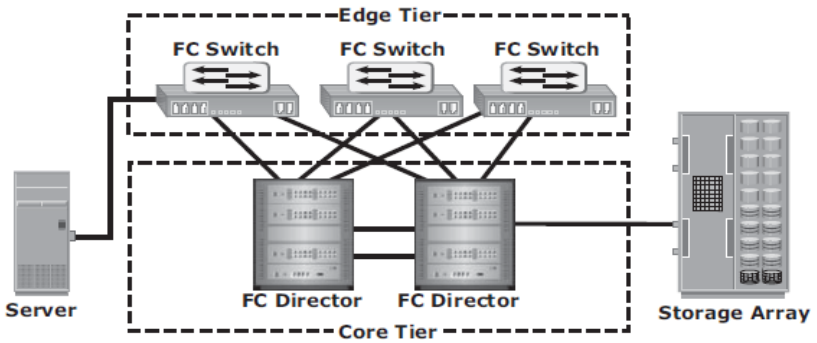


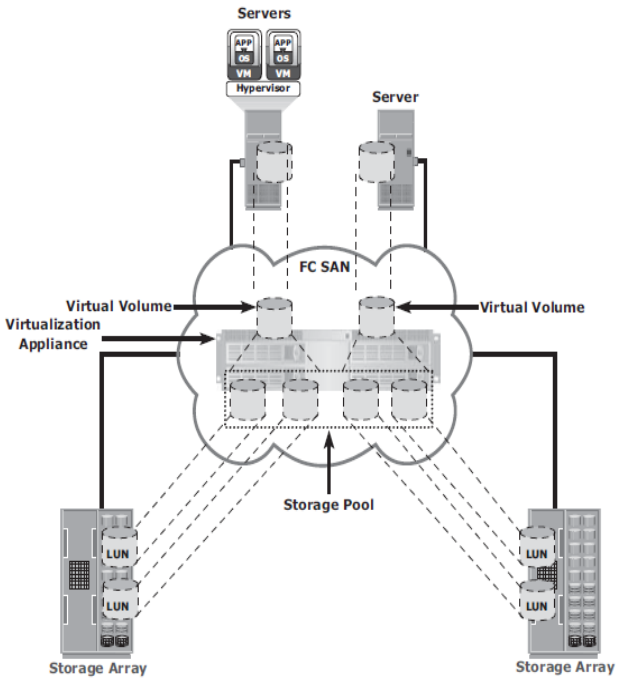
Figure 5-22: Dual-core topology

Virtualization in SAN

سنتحدث هنا عن تقنية ال Virtualization داخل شبكات الSAN باستخدام Block-Level Storage Virtualization & VSAN

- Block-level Storage Virtualization

كما ذكرنا ان شبكات الFC-SAN تستخدم نظام الblock-level system حيث يتم تقسيم الstorage array الى LUNs وتخصيصها الى الHosts.



هنا الblock level يستخدم طبقة الvirtualization عن طريق انشاء virtual storage volume مستقله عن الphysical storage حيث تقوم بعمل abstracts الى الphysical storage وعمل storage pool من جميع اجهزة الstorage الموجودة وانشاء الvirtual volume عليها يتم تخصيصها الى الhosts بدلا من ربطها مباشرة الى الLUNs, اذن تكون هذه الhosts موجوده فى الvirtualization layer حيث تقوم هذه الطبقه بعمل mapping لل virtual volume الى الLUNs

فى المثال السابق كان الblock-level storage virtualization يدعم الdata migration فقط داخل داتا سنتر واحده, هناك جيل جديد من block-level storage virtualization يدعم الdata migration بين العديد من الداتا سنتر والتى يمكن ربطها وادارتها من خلال virtualization layer واحده تربط جميع الداتا سنتر.

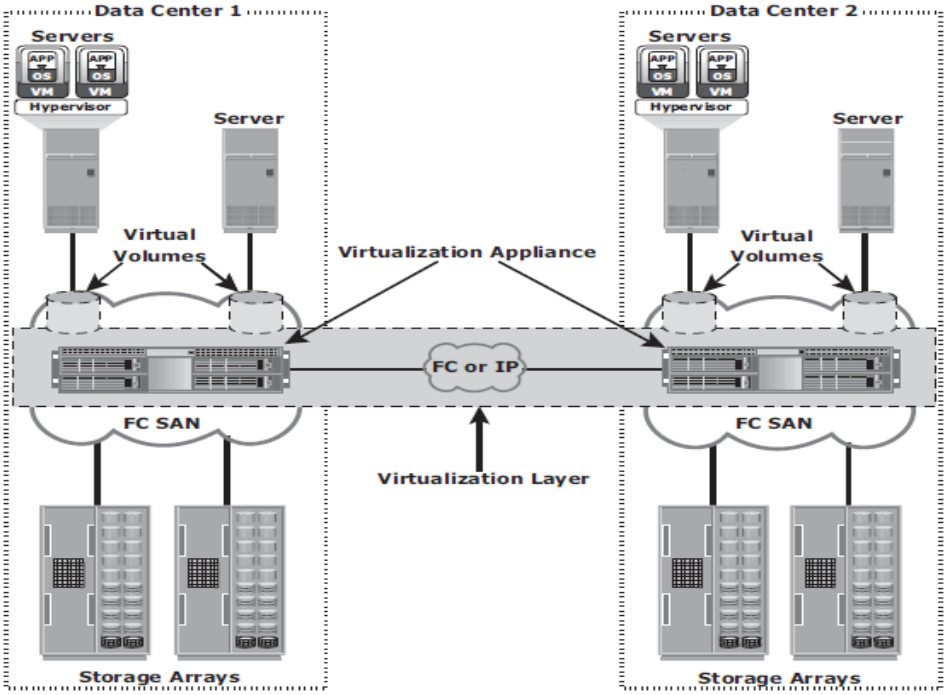


Figure 5-25: Federation of block storage across data centers

- Virtual SAN (VSAN)

تسمى ايضا بال virtual fabric حيث تخصص مجموعه من الhosts or storage ports(nodes) للتواصل فيما بينهم باستخدام virtual topology داخل الPhysical SAN بناءا على ال physical location لهم داخل الfabric.

من الممكن انشاء العديد من الVSANs داخل الPhysical SAN, كل VSAN تكون مستقله لها الservice الخاص بها مثل name server, and zoning

الVSAN توفر الsecurity scalability, availability, and manageability عن طريق عزل البيانات الخاص بكل VSAN وتقييد الوصول اليها لغير المصرح لهم داخل الVSAN membership.

- **Dynamic Port VSAN Membership**

بشكل افتراضى كل بورت ينتمى الى Default VSAN ويوجد طريقتين لعملassign للبورت داخل الVSAN اما manually على كل بورت access VSAN1

اما dynamically عن طريق تخصيص الVSAN Membership بناءا على الWWN وهذه الطريقه تسمى Dynamic Port VSAN Membership (DPVM) والDPVN يحتوى على كل الداتابيز الخاصة بالpWWN والnWWN ويختار الVSAN عن طريق عمل mapping

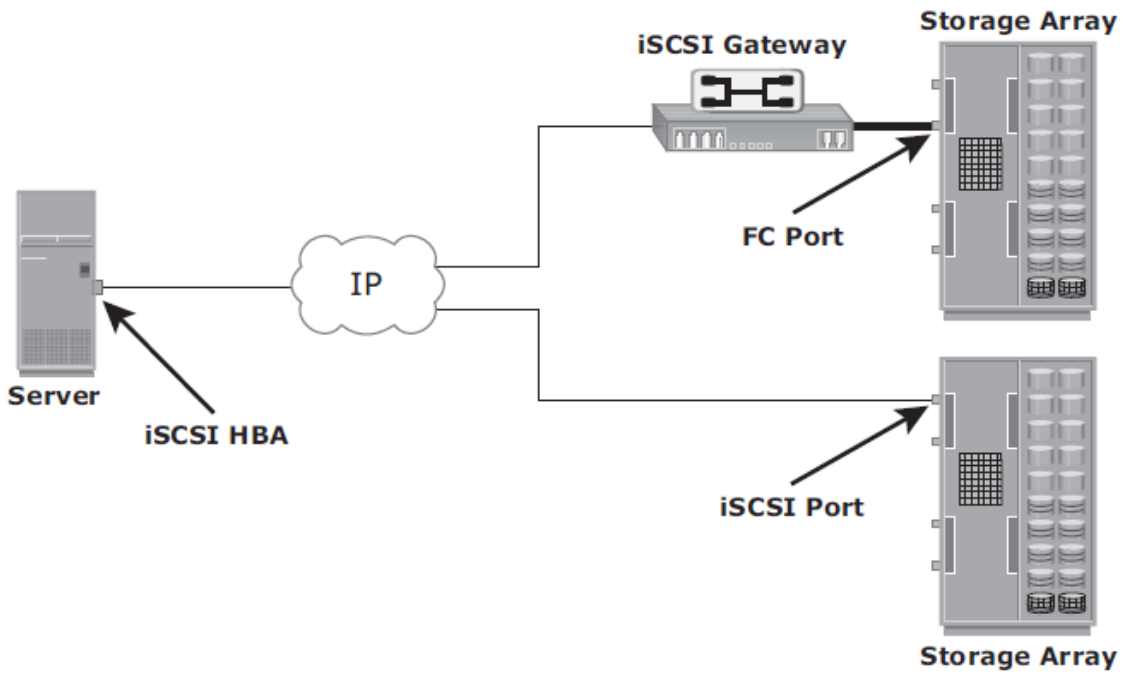
Chapter 6. IP SAN and FCoE

iSCSI

نفس مفهوم الSCSI بروتوكول يقوم بنقل البيانات بين الHost والStorage لكن باستخدام الIP protocol

حيث يتم وضع ال SCSI commands والداتا داخل IP Packet ويتم ارسالها عن طريق الTCP/IP.

يعتمد عليه بشكل واسع لانه قليل التكلفه وسهل التطبيق لكن من عيوبه محدودية الcapacity وغير مناسب للتطبيقات الكبيرة.



- **Components of iSCSI**
  - Initiator (Host)
  - Target (Storage or ISCSI gateway)
  - IP based network

- **iSCSI Host Connectivity**
  - Standard NIC with software ISCSI initiator
  - TCP offload engine (TOE)
  - ISCSI HBA

- iSCSI Topologies

- **Native iSCSI Connectivity**

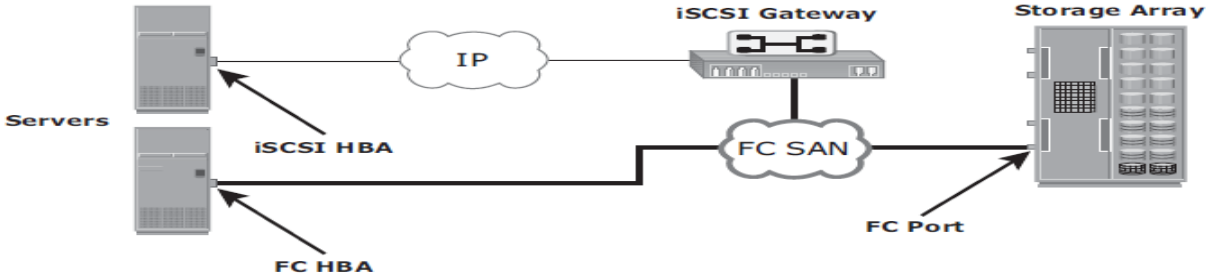
هنا مكونات الFC غير مطلوبه فمن الممكن ان يكون ال initiator متصل مباشرة بال Target او عن طريق الIP network. كما فى الصورة التالية الArray يحتوى على iSCSI port وله IP Address متصل بالشبكة من خلال Standard ethernet switch وعندما يتصل ال initiator بالشبكه يمكنه الوصول الى ال LUNs المتاحه على الArray.

(a) Native iSCSI Connectivity

- **Bridged iSCSI Connectivity**

تحتوى على مكونات الFC حيث ان ال Array لا يحتوى على iSCSI port ولكنه متصل بجهاز خارجى يسمى iSCSI gateway يستخدم للاتصال بين الiSCSI host & FC Storage حيث يقوم بتحويل الIP Packets الى FC Frames

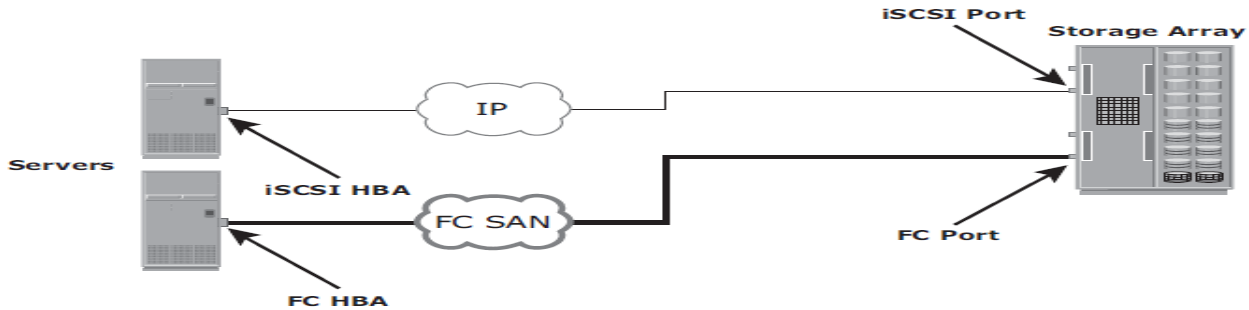
يتم ضبط اعداد ال iSCSI initiator بال IP Address لل gateway على اساس انه الTarget وبالعكس بالنسبه للArray يتم اعداده على انه الFC Initiator.



(b) Bridged iSCSI Connectivity

- **Combining FC and Native iSCSI Connectivity**

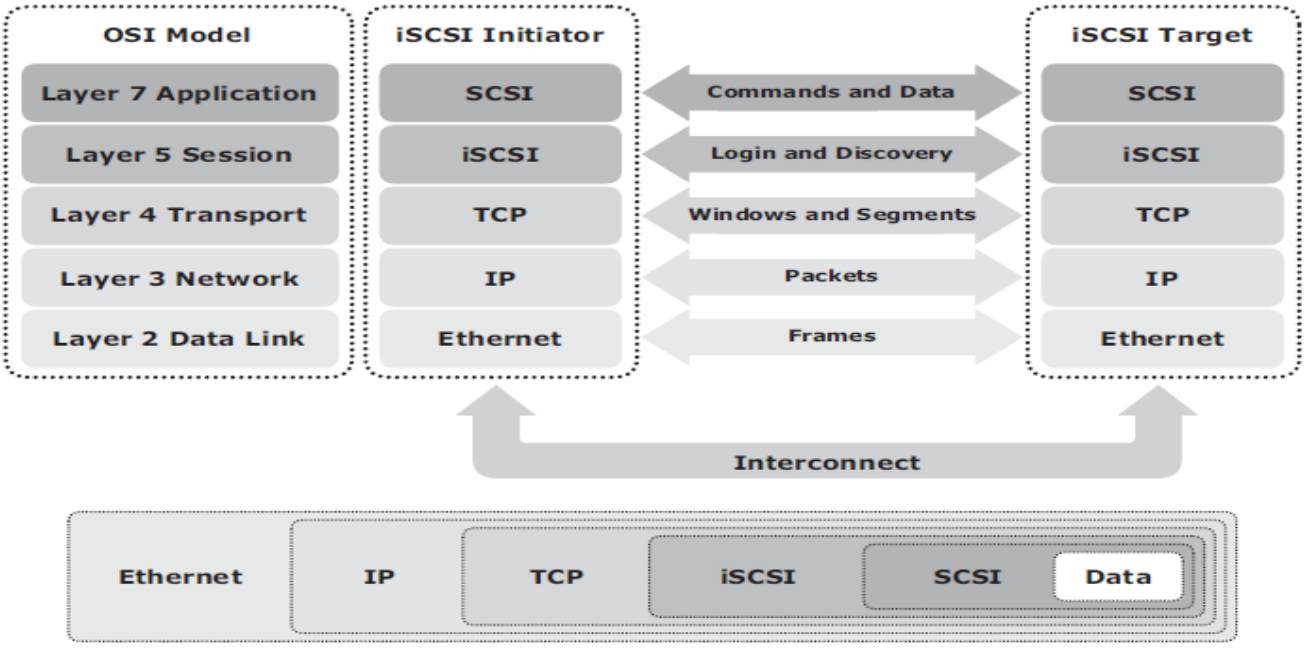
حيث تحتوى على النوعين معا, حيث يوجد بها iSCSI Port & FC Port.



(c) Combining FC and Native iSCSI Connectivity

- iSCSI Protocol Stack

الSCSI هو command protocol يعمل فى الapplication layer حيث يتم عمل Encapsulate لل SCSI Request سواء كان Read or Write داخل TCP/IP Header ويرسل من خلال NIC فى الEthernet العادية.



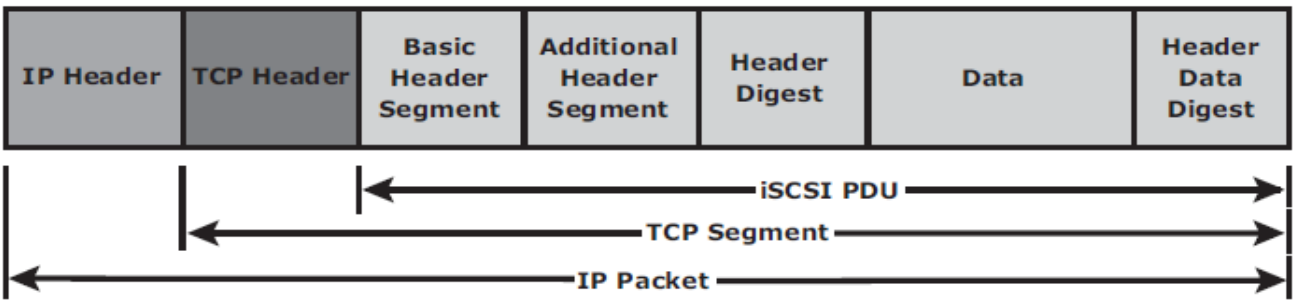
- iSCSI PDU

يتم التواصل بين الiSCSI initiators بينهما البعض باستخدام iSCSI PDU, تحتوى هذه الاتصالات على iSCSI connections, iSCSI sessions, iSCSI discovery, sending SCSI commands and data

الiSCSI PDU عباره عن واحده او اكثر من header segments اضافة الى ال Data Segment يتم عمل Encapsulate له داخل IP Packet لتسهيل عملية النقل.

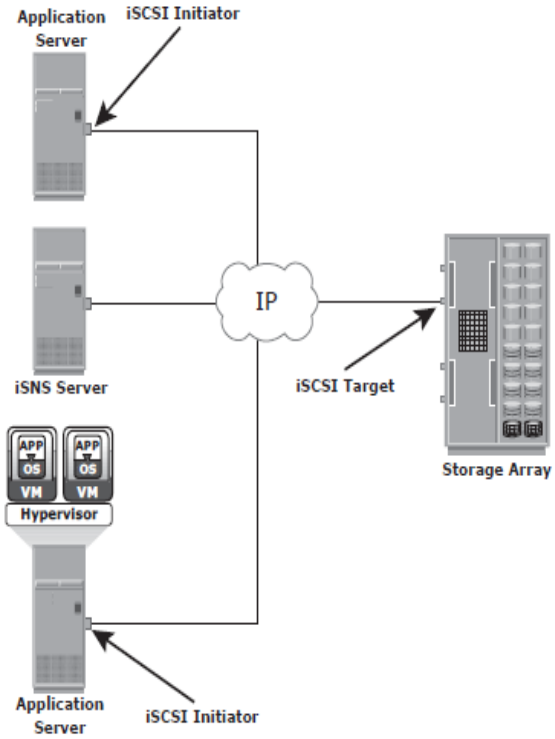
يحتوى الPDU للIP Packet على IP Header يقوم بعملية الrouting information لنقل البيانات داخل الشبكة, ويحتوى ايضا على TCP Header الذي يحتوى على المعلومات المطلوبه لضمان وصول الباكيت الى الtarget.

اضافة الى الiSCSI Header الذى يقوم باستخراج الSCSI commands والبيانات وتوصيلها الى ال target ويحتوى ايضا على CRC او digest الذي يقوم بالتاكد من سلامة البيانات.



- iSCSI Discovery

الinitiator لابد ان يستكشف مواقع واسماء الtargets قبل البدء فى عملية التواصل وتتم عملية الاستكشاف بطريقتين اما Send Targets discovery او (iSNS) internet Storage Name Service.



- **Send Targets discovery** يتم اعداد الinitiator يدويا داخل شبكة الtargets لبدء عملية الاستكشاف حيث يقوم بارسال Send Target commands ويقوم الtarget network portal بالاستجابة وارسال اسماء الtargets والIP Addresses المتاحه.
- **iSNS** يقوم باستكشاف اجهزة الiSCSI بشكل automatic حيث يتم اعداد الinitiators & targets للتسجيل داخل الiSNS server اوتوماتكيا حيث يحتوى هذا السيرفر على list بكل الtargets داخل الشبكة ايضا يتم استخدام service location protocol (SLP) فى عملية الاستكشاف.

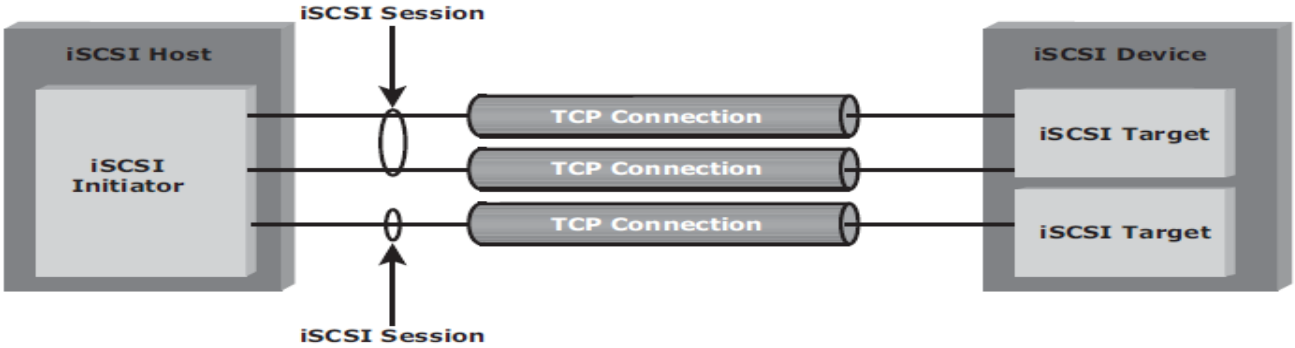
- iSCSI Names

تستخدم لتعريف الinitiators & target داخل الiSCSI network لتسهيل عملية التواصل ويوجد نوعين من اسماء الiSCSI.

- iSCSI Qualified Name (IQN)
- Extended Unique Identifier (EUI)

- iSCSI Sessions

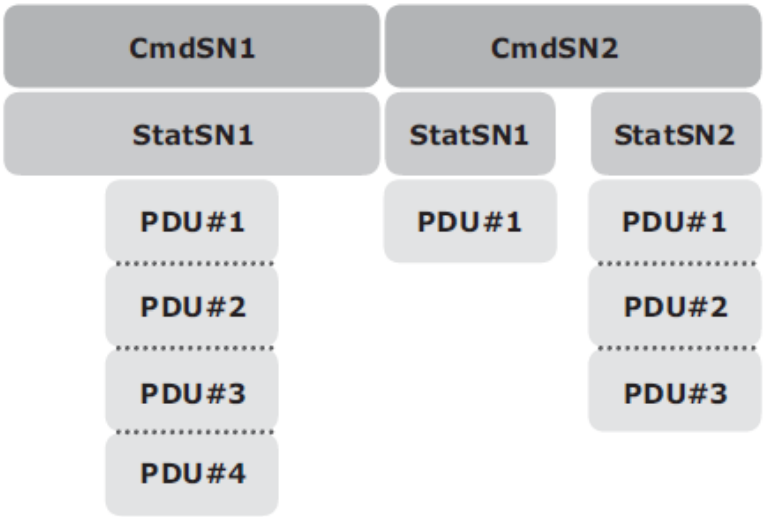
هي الsession التى تبدا بين الinitiator & target وتعرف بـsession ID (SSID), ويتم انشائها فى عدة حالات مثل مرحلة Discovery عملية استكشاف الinitiator الى الtarget المتاحة او مرحلة نقل الداتا العادية بينهما.



- **iSCSI Command Sequencing**

عملية التواصل بين الtarget & initiators تعتمد على بعض الcommand sequence مثل ال – request respond هذه الsequences من الممكن ان تنشأ العديد من الPDUs لذلك يتم ترقيمها وتعرف باسم command sequence number (CmdSN) ليتم تعريف الcommand PDUs الخاصه بكل session.

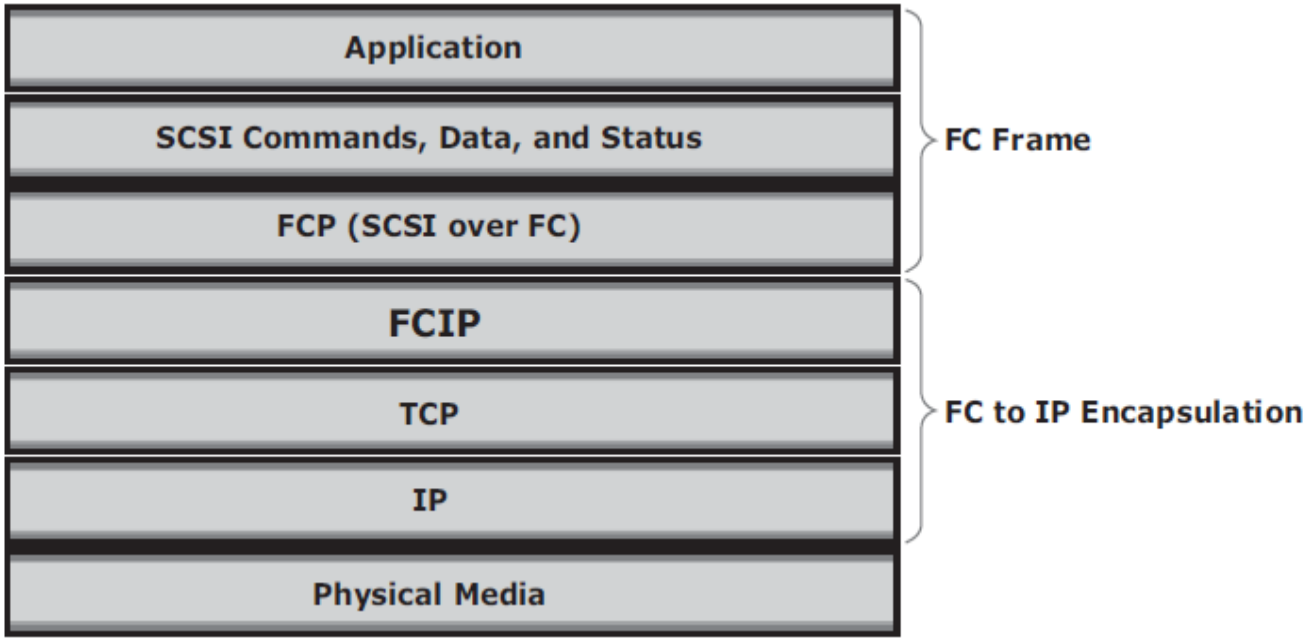
ايضا هناك الstatus sequence number (StatSN) تستخدم لمعرفة حالة استجابة الsequence number



FCIP

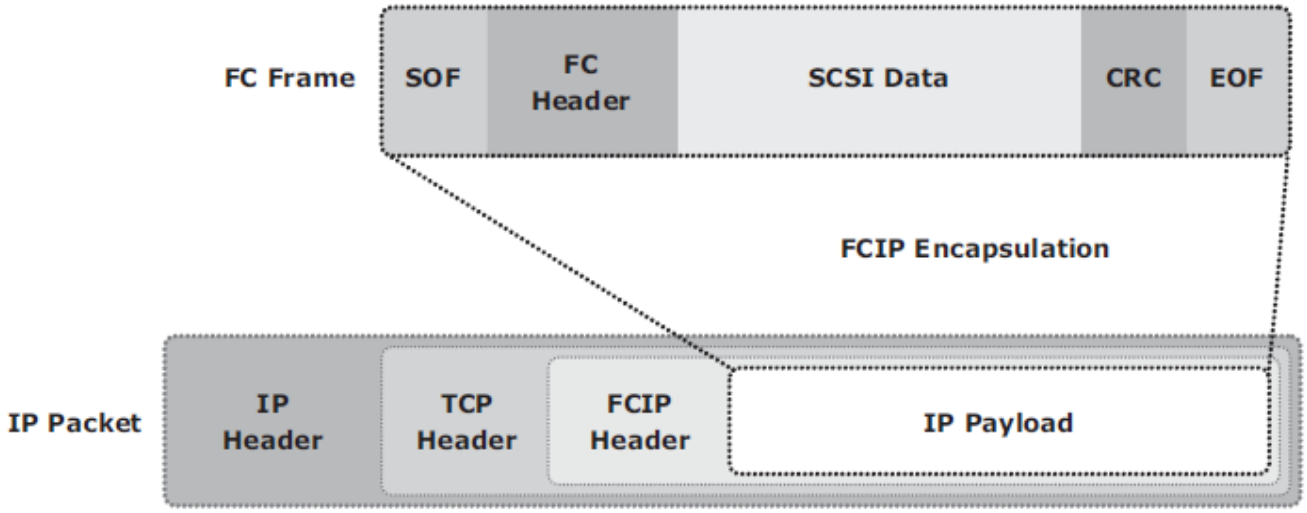
هو tunneling protocol يستخدم لنقل البيانات داخل ال FC SAN من خلال ال IP based network عن طريق عمل Stacking لل Fibre Channel Frame داخل IP Packet.

- **FCIP Protocol Stack**



- Application Layer تقوم بانشاء الSCSI commands & Data.
- اما الSCSI Layer يحتوى على تطبيق SCSI driver الذي يقوم بتنفيذ اوامر read write commands
- الFCP Layer هو الFibre Channel frame الذي يحمل الSCSI يكون اعلى الtransport layer مما يسهل مروره خلال الSAN fabric environment بالاضافه انه يتم عمل encapsulation لل Fibre Channel Frame داخل IP Packet
- FCIP Layer هي الطبقة التى يتم فيها عمل encapsulate لل Fibre Channel Frame داخل IP Packet وتمريرها الى الTCP Layer .
- IP Layer تستخدم لنقل البيانات حيث تقوم بعمل encapsulated data داخل الethernet.

عندما يتم عمل encapsulate للبيانات الى IP Packet من الممكن للData link ان لاتدعم حجم الMTU لل IP Packet مما يتسبب فى عمل fragmentation لها والذي يتطلب وضع headers على كل fragments ويقوم الTCP بعملية استقبال واعادة هيكلة البيانات resequencing وتمريرها الى ال FC processing للجهاز.

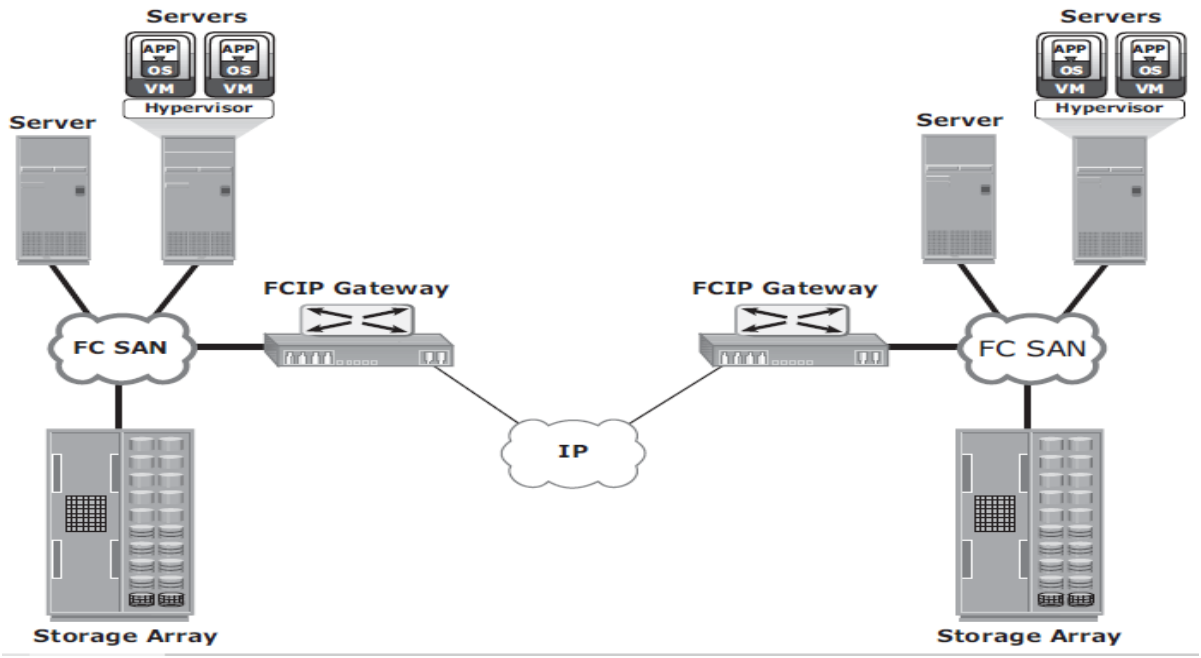


- FCIP Performance and Security

يجب الاخذ بالاعتبار وجود عملية الPerformance, reliability, and security عند تطبيق الFCIP ومن ضمن هذه الاعتبارات, وجود اكثر من مسار بين الFCIP gateways وهذا لتجنب single point of failure وزيادة الbandwidth. ايضا الFCIP يستخدم الunified fabric والتى يتم فيها الدمج بين الLAN Network وال SAN Network للاستفادة من الخدمات الموجوده فى الشبكة الى اقصى حد لتحقيق اداء افضل لها وانخفاض تكاليف التشغيل وزيادة اداء التطبيقات فى الشبكة.

- **FCIP Topology**

داخل الFCIP environment يوجد FCIP gateway متصل بكل Fabric SAN يقوم واحد منهم متصل بال IP network بعمل encapsulate لل FC Frame داخل IP Packet ويقوم الاخر باستقبال الPacket وعمل ازالة لل IP وارسال الFC data داخل الfabric



FCoE

هو بروتوكول يستخدم لنقل الFibre Channel Traffic من خلال ال Ethernet. اي انه يقوم بنقل الداتا الخاصه بالLAN & SAN حيث يتم عمل Encapsulate لل Fibre Channel frame التى تخرج من ال CNA الى TCP/IP Header داخل ال Ethernet.

- **Consolidation Using FCoE**

من اهم فوائد الFCoE هي عملية الدمج بين الشبكات, حيث انه قبل الFCoE كان يستخدم كروت NIC للوصول الى IP network ويتم استخدام كروت الHBA للوصول الى Storage resources فكان كل سيرفر يحتوى على الاقل على 2 NIC & 2 HBA من اجل عملية الredundancy فاذا كان هناك العديد من السيرفرات يتطلب هذا العديد من الadapters, cables & switches فتصبح العملية معقده داخل الenvironment.

لكن جاء الFCoE وقام بعملية الدمج وذلك باستخدام FCoE switches & Converged Network Adapters (CNAs) يتم استبدالها مع الNIC & HBA داخل السيرفر لدمج IP & FC traffic لتقليل عدد الadapters, cables & switches داخل الشبكة وبالتالى تقليل التكلفة وسهولة الادارة.

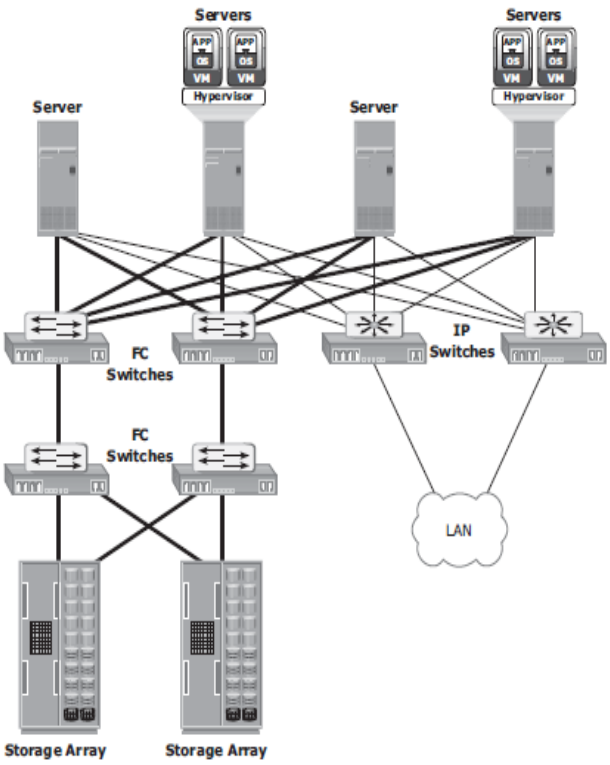


Figure 6-12: Infrastructure before using FCoE

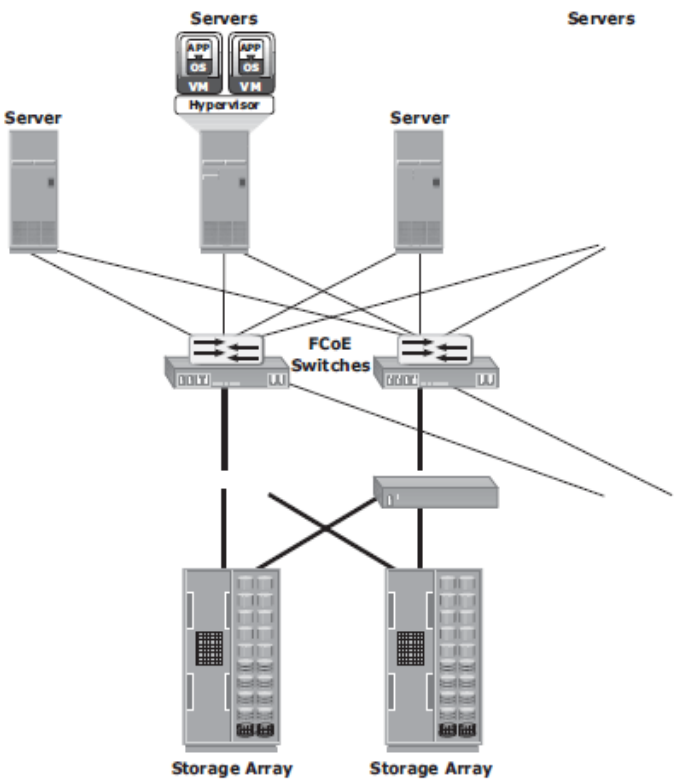
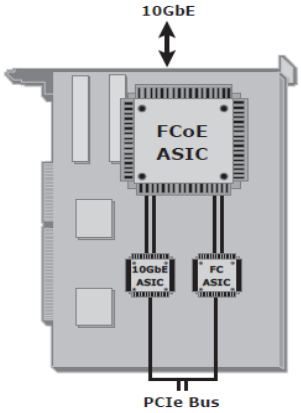


Figure 6-13: Infrastructure after using FCoE

- Compenants of an FCoE Network

- **Converged Network Adapter**



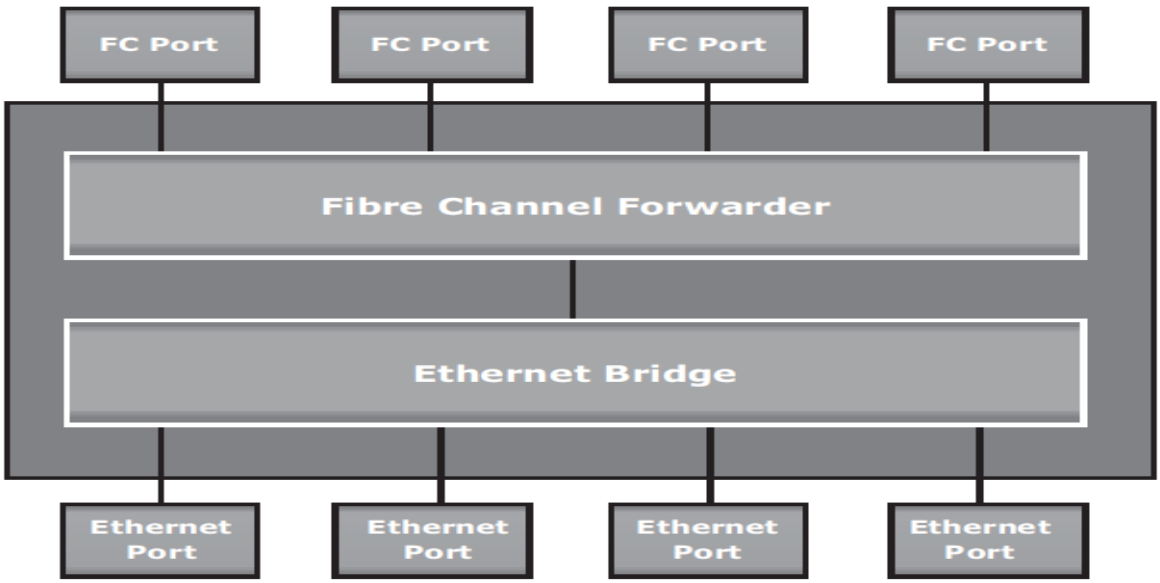
وهي كروت تدعم الEthernet & Fibre Channel بحيث انه يرسل SCSI Request يحتوى على FCoE Header الى السويتش. وكذلك مع الEthernet يتم عملEncapsulate لل الTCP/IP Header داخل FCoE Header مع ال SCSi Request.

- **Cables**

هناك اختيارين من الكابلات المستخدمه اما Copper based Twinax للمسافات القصيره حد اقصى 10 meter او standard fi ber optical بسرعات 10Gbps يتم توصيلهم بال Small Form Factor Pluggable Plus (SFP+).

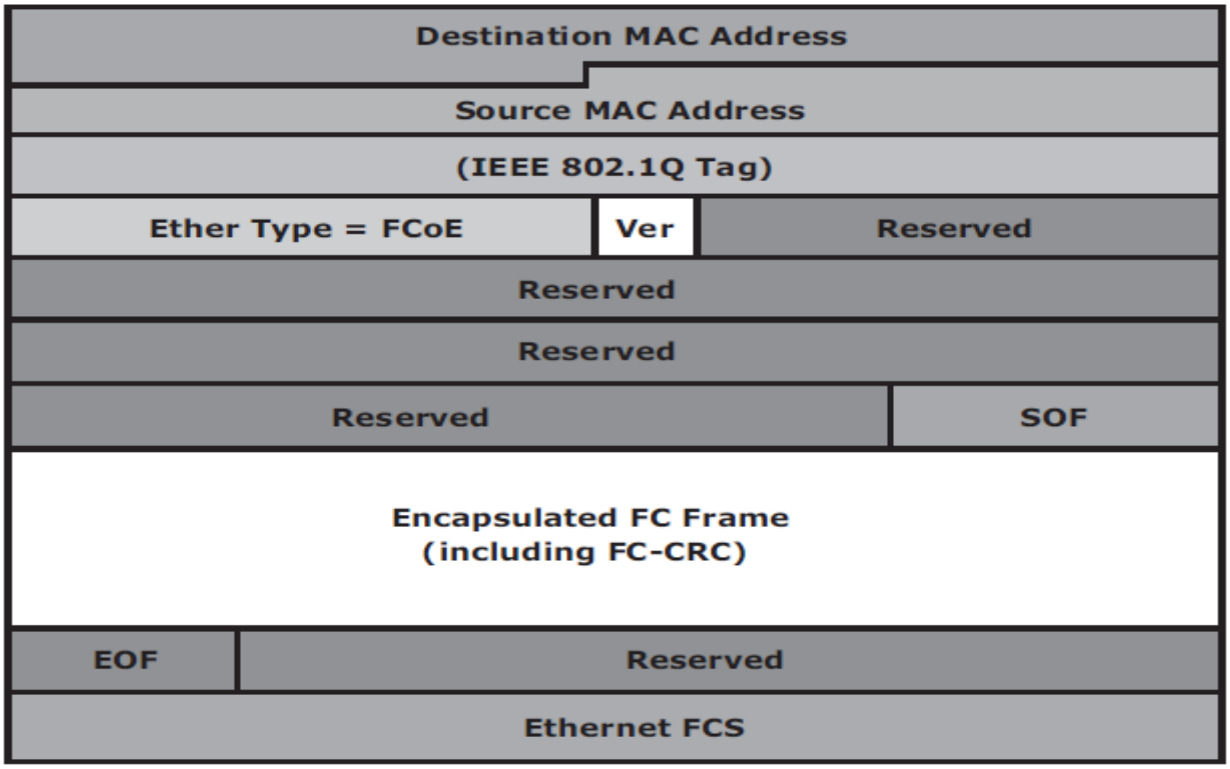
- **FCoE Switches**

تقوم بعمل وظائف النوعين ethernet switch & fibre channel switch حيث تحتوى على Fibre Channel Forwarder (FCF), Ethernet Bridg, Ethernet ports and optional FC ports حيث يقوم الFCF بعمل encapsulate لل FC Frames المستقبله من الFC Ports ايضا عمل de-encapsulate لل FCoE frames المستقبله من الethernet bridge



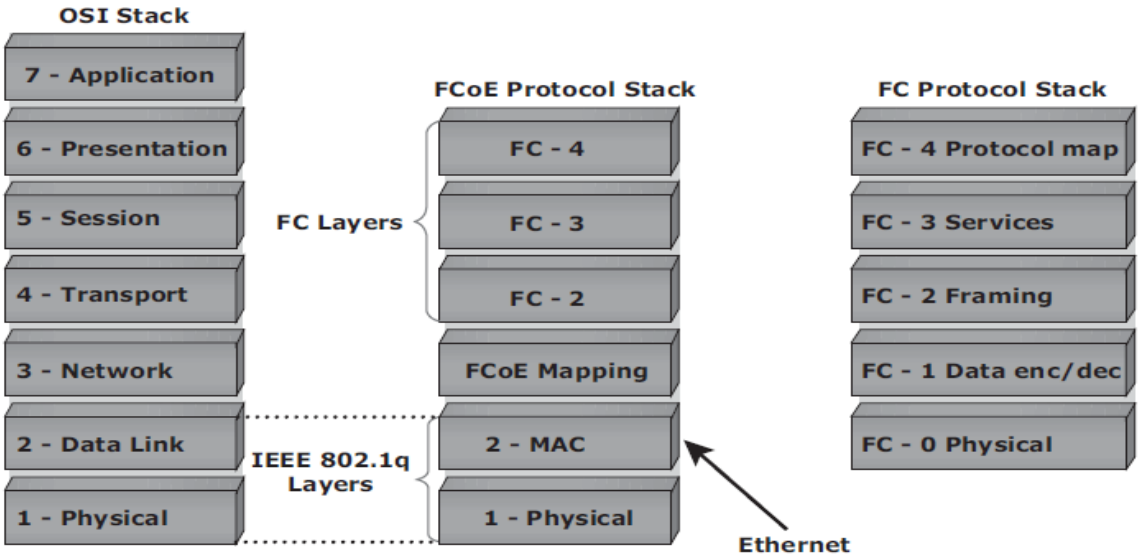
- FCoE Frame Structure

الFCoE Frame هو عباره عن Ethernet Frame تحتوى على FCoE protocol data unit كما موضح بالصورة. اول 48-bits تستخدم لتحديد الDestenatin MAC Address وال 48-bits الثانية لتحديد الSource MAC Address, 32-bits التالية IEEE 802.1Q tag المسؤوله عن انشاء الVLANs,ثم16-bits خاصة بالFCoE معها 4-bit للversion, 100-bits محجوزه معها 8-bit لراس الفريم (SOF) ثم ياتى الفريم نفسه حيث حجم ال Frame فى الFibre Channel لا يتعدى 2112 Bytes مضافا اليه الHeader تصبح 2148 Bytes, ثم 8-bit اخرى لنهاية الفريم متبوعه ب 24-bits محجوزة, وفى النهاية 32-bits للFrame Check Sequence (FCS) مسؤوله عن التحقق من الفريم واستكشاف الاخطاء.



- **FCoE Frame Mapping**

عملية الencapsulation للFCoE Frame تحدث من خلال عمل mapping لل FC frame داخل الethernet كما موضح بالصورة حيث لكل منهم طبقات محددة تمر من خلالها الداتا لكل طبقه وظائف معينه, بالنسبه للFCoE يقوم باستبدال طبقة ال FC-0 & FC-1 الخاصه بالFibre Channel بطبقة الethernet مما يوفر القدرة على نقل بيانات FC2 – FC4 داخل طبقة الethernet.



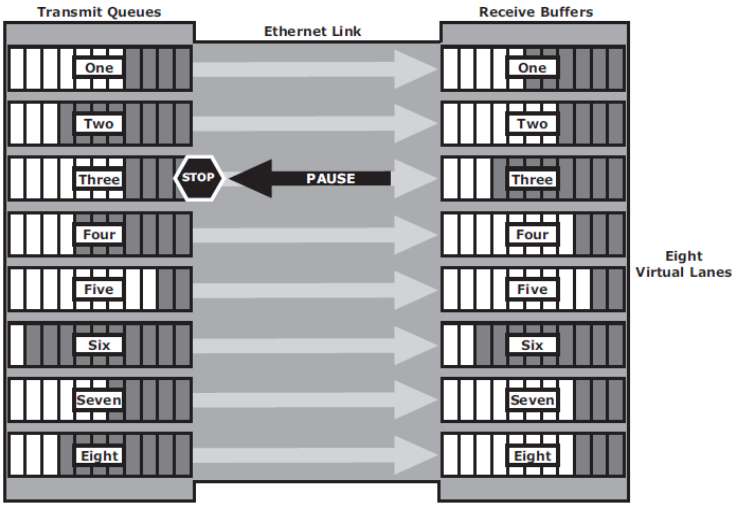
- **FCoE Enabling Technologies**

الEthernet على مستوى Layer 2 يوجد لديه مشكلتين اساسيتين اولا يتطلب القدره على توفير lossless traffic delivry للتاكد من وصول الترافيك كامله, ايضا مشكلة الPause وتحدث عند وجود congestion فى الشبكة فيقوم بارسال pause request لوقف ارسال الداتا سواء كانت خاصة بال LAN or SAN. ولذلك لابد من تحسين اداء الEthernet لانه يستخدم فى نقل البيانات الخاصة بالSAN عن طريق بروتوكول الFCoE ولايجب وجود فقدان للداتا ووجود حل لمشكلة الCongestion الموجوده فى الشبكه وذلك باستخدام DCB Standards وهي:

Feature	Benefit	IEEE Standard
Priority-based Flow Control (PFC)	Enables lossless Ethernet—manages I/O between initiator and target on a multi-protocol Ethernet link	802.1Qbb
Enhanced Transmission Selection (ETS)	Allocate bandwidth to IP, iSCSI and FCoE traffic—managed with OneCommand 5.0	802.1Qaz
Congestion Notification (CN)	Transmit congestion information end-to-end per traffic flow	802.1Qau
Data Center Bridging Capability Exchange (DCBX)	Extends DCB network by exchanging Ethernet parameters between DCB switches	802.1ab

- **Priority Flow Control (PFC)**

هي ميزه تستخدم لحل مشكلة Losseless traffic delivry عن طريق تقسيم اللينك المستخدم لنقل الترافيك الى 8 Virtual Links وتخصيص Bandwith محدده لكل Virtual link ومن خلال عمل QoS Configuration ممكن تقسيم الترافيك الى Classes وتخصيص كل class لكل virtual link



- **Enhanced Transmission Selection (ETS)**

هي ميزه اخرى لتحسين الEthernet من خلال تحسين اداء الPFC وذلك عن طريق انه اذا حدث Congesstion فى احدى ال Virtual link المستخدمه لانها استهلكت الBandwith المخصصه لها وكان هناك virtual link اخرى غير مستخدمه فتقوم بعمل reallocation للترافيك بحيث تخرج من الFree virtual link وتقوم بارجاعها مره اخرى اذا جاءت ترافيك لهذه اللينك.

- **Data Center Bridging Exchange (DCBX)**

هو بروتوكول شبيه لل LLDP او الCDP بحيث انه يقوم بعمل Negotiation & Exchange لل FCoE بروتوكول بين ال initiotor & target ليقوم بعمل Discovering لكل ال DCB-capable Switches الموجوده داخل الFibre Channel Netowr وكل المعلومات الخاصه بها.

- **Congestion Notification (QCN)**

اذا كان هناك اكثر من سويتش على الشبكه وموصل ليهم اكثر من سيرفر كلهم يقوموا بارسال ترافيك الى ان تصل الى السويتش الاخير الموصول بال Arrays فمن الممكن حدوث Congession عليه فيرسل Pause Message فيتم ايقاف جميع الترافيك على ال هنا ياتى دور الCN حيث انه باستخدام الDCBX ومعرفة ال exchange capable والtraffic flows وال traffic classes لكل الSwitches يقوم كل سويتش بعمل Decapsulate للترافيك الى ان يصل الى السويتش صاحب الoverload traffic ويقوم بعمل pause traffic له

Chapter 7. Network-Attached Storage

NAS هي طريقه للوصول الى ال Storage من خلال الethernet network باستخدام بروتوكولات مثل NFS, CIFS & SMB وتعمل بنظام file-level storage system.

Benefits of NAS

Comprehensive access, Improved effi ciency, Improved fl exibility, Centralized storage, Simplified management, Scalability, High availability, Security, Low cost & Ease of deployment

File Systems and Network File Sharing

- **Accessing a File System**

الFile System يجب ان يكون mounted وهي طريقة في نظام التشغيل تستخدم للتحكم في كيفية تخزين وإسترجاع وتنظيم وإدارة ملفات الحاسوب والبيانات التي تحتوى عليها تلك الملفات لتسهيل إيجادها وإستخدامها، والfile system يستعمل جهاز تخزين للبيانات مثل القرص الصلب او NAS وعملية الmounting هذه تقوم بعمل اتصال يربط بين الfile system على الNAS وبين انظمة التشغيل.

- **Network File Sharing**

تشير الى عملية التخزين والوصول الى البيانات من خلال الشبكة. وتسمى بال file sharing وهنا يقوم الuser بانشاء الملفات ويحدد طريقة الوصول اليه مثل القراء والكتابة والتنفيذ وازالة الملفات التى يمنحها للمستخدمين الاخرين.

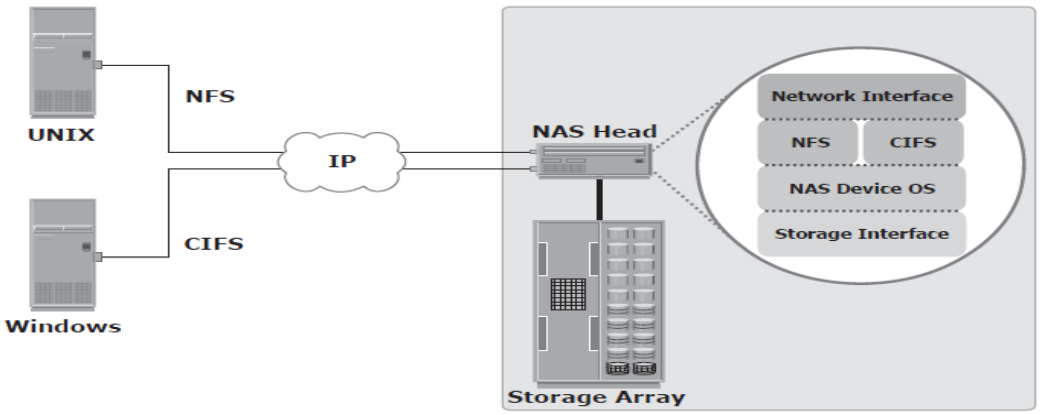
ومن الامثلة لعملية الfile sharing هي file transfer protocol (FTP), Distributed File System (DFS), ايضا الclient-server models التى تستخدم بروتوكولات الfile sharing مثل NFS & CIFS.

Components of NAS

تحتوى الNAS على مكونين رئيسين هما الHead & Storage, بالنسبة للStorage ممكن ان تكون خارجية على اجهزة NAS وتكون Shared على الhosts, اما بالنسبة للHead يحتوى على المكونات التالية

- CPU and memory
- One or more network interface cards (NICs)
- optimized operating system for managing the NAS functionality
- NFS, CIFS, and other protocols for fi le sharing
- Industry-standard storage protocols and ports

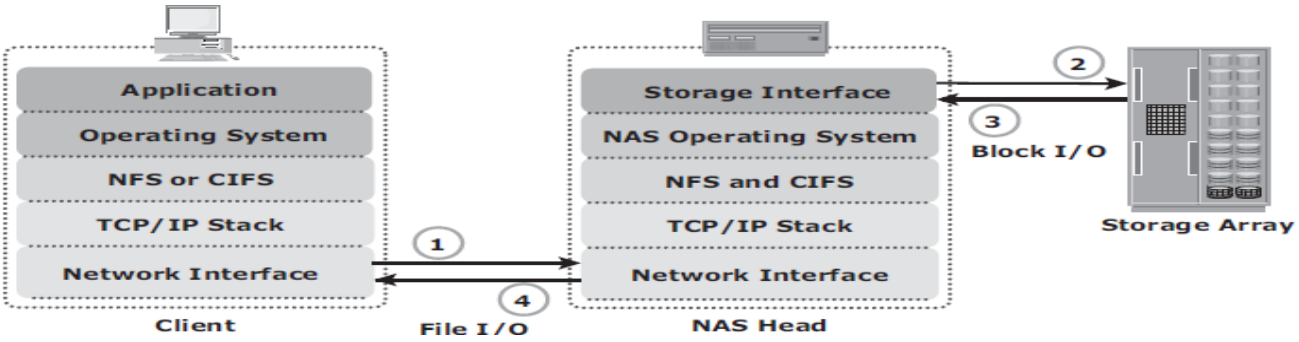
شبكة الNAS تحتوى ايضا على clients تقوم بالوصول الى اجهزة الNAS عن طريق IP Network باستخدام file-sharing protocols.



NAS I/O Operation

الNAS يعمل بنظام الfile-level لوصول ال client الى البيانات فعملية الوصول الى البيانات من الclient الى الstorage تمر بعدة مراحل:

1- ال client يقوم بارسال طلب file I/O داخل TCP/IP package وارسالها خلال الشبكة
2- يقوم الNAS device باستقبال طلب الfile I/O وتحويلها الى block-level I/O لارسالها الى الstorage
3- عندما تعود الداتا من الstorage تقوم الNAS device بعمل repackages لها باستخدام البروتوكول المناسب مثل NFS, CIFS, SMB من خلال TCP/IP Packages وتحويلها الى الclient.



NAS Implementations

هناك ثلاث عوامل مشتركة عند تطبيق الNAS وهي Unified, Gateway & Scale-out.

حيث ان الunified يقوم بتوحيد الوصول الى البيانات فى الNAS-based & SAN-based ضمن storage موحده وتوفير interface موحد ايضا لادارة كل منهما.

اما فى الgateway اجهزة الNAS تستخدم storage خارجية لتخزين واسترجاع البيانات وتختلف عن ال unifed storage بان هناك مهام ادارية منفصله لاجهزة الNAS والstorage.

اما الscale-out يقوم بتجميع العديد من الnodes داخل cluster.

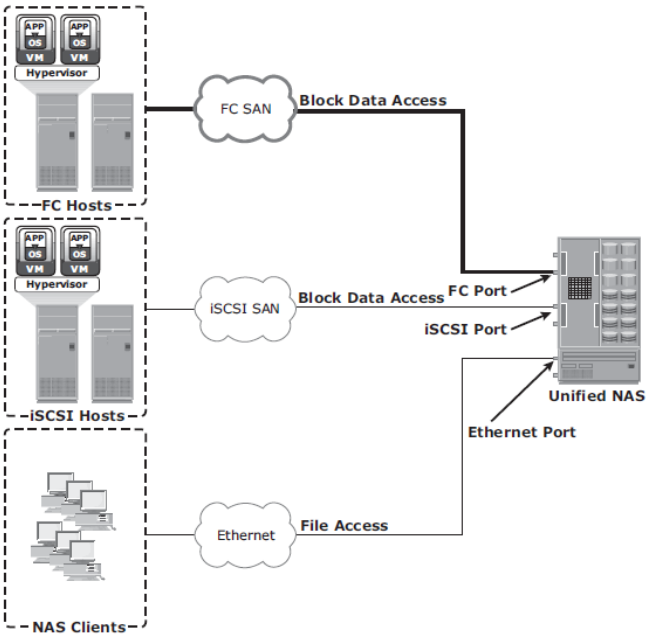
- Unified NAS

يقوم بخدمة تخزين واسترجاع البيانات مع امكانية الوصول الى block-level data, حيث تدعم بروتوكولات CIFS & NFS بالنسبة للfile-level access ايضا تدعم بروتوكولات iSCSI & FC بالنسبة للblock-level access . والتى تؤدى الى دمج الNAS-based & SAN-based داخل storage platform واحد. وبالتالى تساعد على تقليل تكلفة الinfrastructure & management للمنظمه.

وتحتوى على واحد او اكثر من NAS heads & storage داخل السيستم الواحد.

- Unified NAS Connectivity

كل NAS head يعتبر unified NAS حيث يحتوى على front-end Ethernet ports موصوله بال IP network لتوفير وصول الclients الى البيانات واستقبال الI/O requests. ايضا يحتوى على back-end ports موصوله بالstorage controller حيث تدعم بروتوكولات iSCSI & FC.



- Gateway NAS

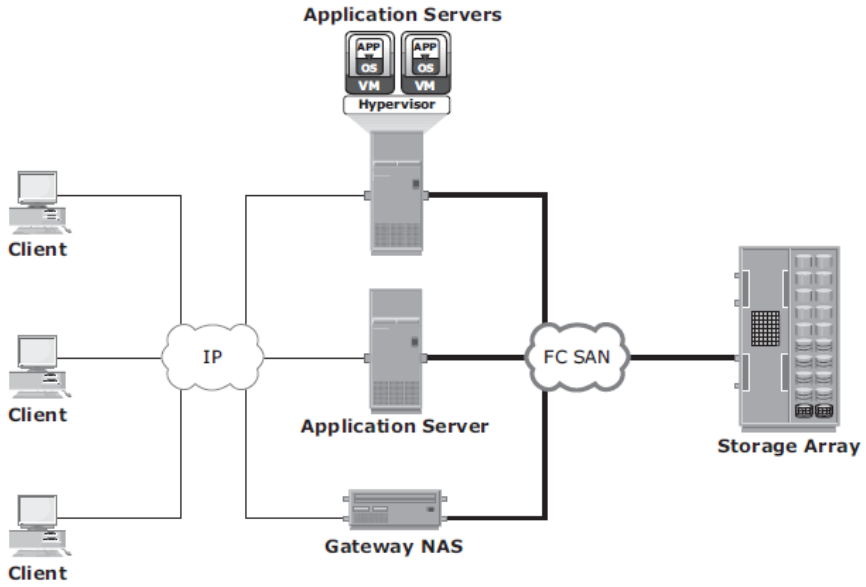
تحتوى على واحد او اكثر من الNAS heads بالاضافه الى storage الخارجية التى تدار بشكل مستقل.

تشبه الunified storage من حيث المشاركه فى التطبيقات المختلفة التى تستخدم الblock-level I/O وتختلف عنها حيث لها مهام ادارية منفصله لاجهزة الNAS والstorage.

يعتبر اكثر قابلية للتطوير مقارنة بالunified storage حيث من الممكن توسيع او تطوير ال NAS heads & storage اذا تتطلب الامر, على سبيل المثال من الممكن اضافة NAS heads لتحسين اداء ال NAS devices, ايضا يمكن اضافة capacity اذا وصل التخزين للحد الاقصى مثل الunified NAS ايضا يمكنها تحسين استخدام تخزين الstorage.

- Gateway NAS Connectivity

الfront-end connectivity تشبه الunified storage connectivity, للاتصال بين الNAS gateway وال Storage يكون من خلال traditional FC SAN لدعم الحلول المختلفة مثل multiple paths for data, redundant fabrics, and load distribution.



- Scale-Out NAS

كما ذكرنا ان كل من الNAS unified & gateway له القابلية للتوسيع والتطوير والتى تؤدى الى زيادة سعة البيانات وتحسين الاداء من خلال اضافة بعض الresources مثل الCPUs, memory & storage حيث تكون الstorage محدده بسعه معينه.

Scale-out NAS يساعد على هذا التوسيع من خلال اضافة nodes الى الclustered NAS architecture.

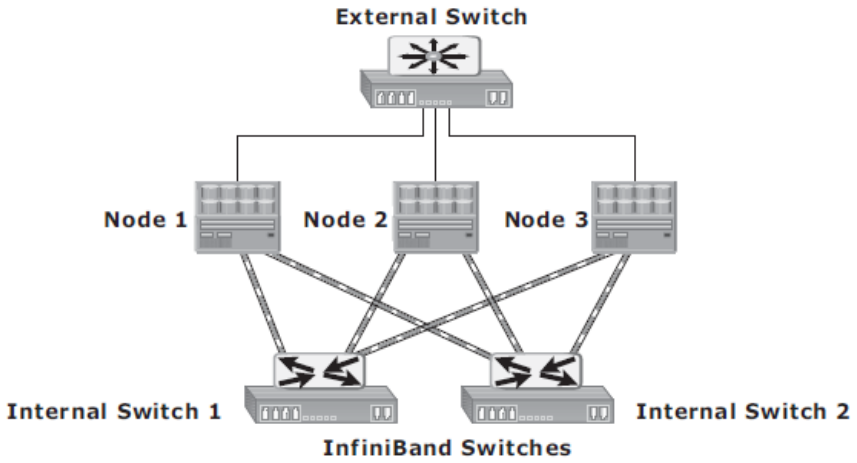
الcluster يعمل ك NAS device ويدار مركزيا, ويمكن اضافة الnodes الى الcluster عند الحاجه الى زيادة السعه والاداء بدون اي downtime.

الscale-out يقوم بانشاء file system يعمل على جميع الnodes داخل الcluster وتكون المعلومات داخلها shared لكل الnodes و يكون الfile system متاح لاي client موصول بالnodes.

يقوم ايضا بعمل stripes للبيانات داخل الnodes باستخدم mirror or parity protection, حيث تكون البيانات المرسله من ال client الى ال cluster مقسمه ومخصصه الى الnodes المختلفة, وعندما يرسل الclient طلب بقراءة البيانات يقوم ال scale-out NAS باعادة blocks المناسبة من العديد من الnodes الموجودة واعادة توجيها داخل file الى الclient.

تعتبر الscale-out مناسبة لاستخدام الBig Data حيث لها القدرة على ادارة وتخزين مساحة ضخمه من البيانات داخل نظام واحد مع المحافظه على الاداء المناسب.

- **Scale-Out NAS Connectivity**



NAS File-Sharing Protocols

اجهزة الNAS تمكن المستخدمين من مشاركة البيانات عبر الانظمة المختلفة وذلك باستخدام بعض البروتوكولات

- **NFS**

هو client-server protocol يستحدم لمشاركة الملفات ويستخدم عادة على انظمة الUNIX. يعتمد فى الاصل على الconnectionless باستخدام الUDP Protocol, يقوم باستخدام اجهزة منفصلة لعرض الuser data

يستخدم ايضا طريقة Remote Procedure Call (RPC) لعملية التواصل بين جهازين كمبيوتر

NFS يقوم بانشاء اتصال بين الclient وال remote system لنقل البيانات

- NFSv2: هو stateless protocol الذي يعنى انه لايحتفظ باي معلومات تخص الملفات
- NFSv3: مبنى على stateless protocol ايضا يقوم باستحدام الTCP or UDP
- NFSv4: يستخدم الTCP حيث انه statefull protocol والذي يوفر الenhanced security.

- **CIFS**

هو client-server application protocol حيث يمكن الclient programs من مشاركة الملفات والخدمات على الremote computers من خلال الTCP/IP protocol

يقوم باستخدام تامين خاص على الملفات والسجلات المخزنه من قبل المستخدم لمنع حدوث overwriting من قبل مستخدمين اخرين.

تدعم ايضا الfault tolerance حيث تقوم باعادة الاتصال اوتوماتيكيا واعادة فتح الملفات التى تاثرت عند حدوث الانقطاع

هو statefull protocol حيث ان الCIFS server تحافظ على معلومات الاتصال الخاصه بكل client.

Factors Affecting NAS Performance

الNAS يقوم باستحدام الIP network لمشاركة الملفات لذلك تتاثر الPerformence بشكل كبير حيث ترتبط بالbandwidth & latency, حيث تقوم الNAS بتجنب حدوث Network congestion باتباع مايلي:

1- Number of hops: حيث زيادة عدد الhops يؤدى الى زيادة سرعة الاستجابة
2- Authentication with a directory service such as Active Directory or NIS حيث يجب ان تتوفر الموارد المتاحة لاجهزة السيرفر الخاصه بعملية الauthentecation لسرعة اداء هذه العملية.
3- Retransmission: حيث من الممكن حدوث Link errors and buffer overfl ows
4- Over utilized routers and switches: حيث يكون وقت الاستجابة اكبر من optimally utilized
5- File system lookup and metadata requests
6- Over utilized NAS devices: محاولة الوصول للملفات من قبل العديد من المستخدمين تؤدى زيادة معدل الاستحدام والتى يمكن تحديدها من خلال عرض احصاءات استخدام ال memory, CPU, or disk subsystem utilization
7- Over utilized clients: ايضا الclient التى تحاول الوصول الى CIFS or NFS من الممكن ان تؤدى الى over utilized والتى تؤدى الى زيادة وقت الprocessing وسرعة الاستجابة.

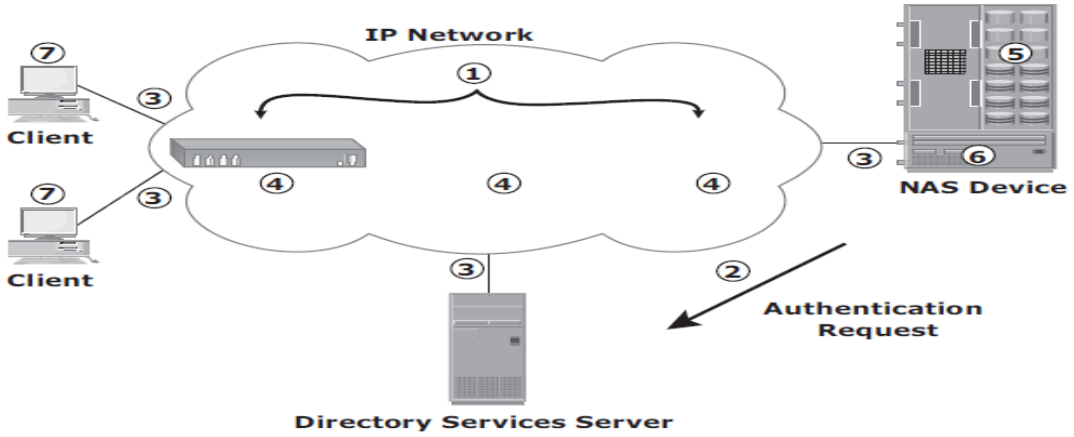
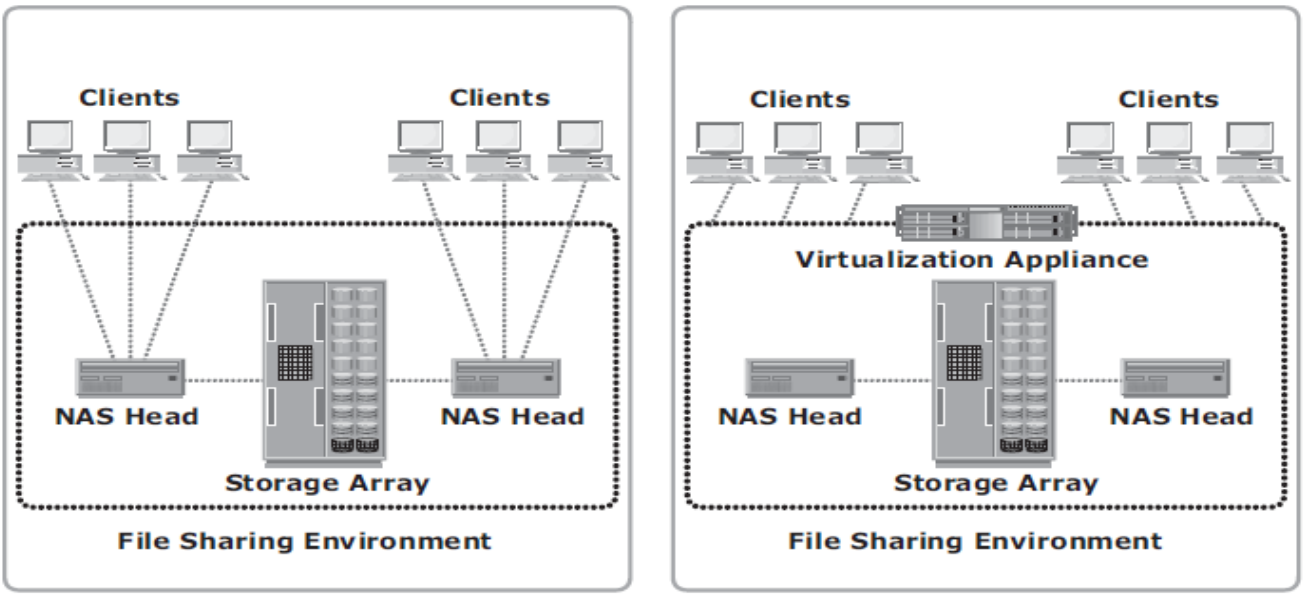


Figure 7-8: Causes of latency

File-Level Virtualization

الFile-Level virtualization يقوم بالغاء تبعية الوصول الى البيانات داخل الfile level والموقع الذى يخزن عليه البيانات بمعنى انه اولا كان كل جهاز يعلم الموقع الذي يخزن عليه البيانات والذى يؤدى الى مشكلة فى حجم التخزين غير المستغل بشكل كامل حيث كل جهاز ملزم بالتخزين فى اجهزة الNAS او الfile sever بالاضافة الى مشكلة فى الperformence حيث عند ملئ الfile server فليس بالسهولة نقل البيانات او اعادة تهيئة التطبيقات للتخزين فى سيرفر اخر.

File-Level viritualizaion يقوم بتبسيط عملية تخزين ونقل البيانات حيث يوفر تطبيق مستقل للموقع الذى يخزن عليه البيانات ويقوم بانشاء logical pool of storage تمكن المسخدمين من استخدام logical path للوصول الى البيانات كما انها تحل مشكلة الوصول الى البيانات من خلال اجهزة الNAS عند تغيير الfile server.



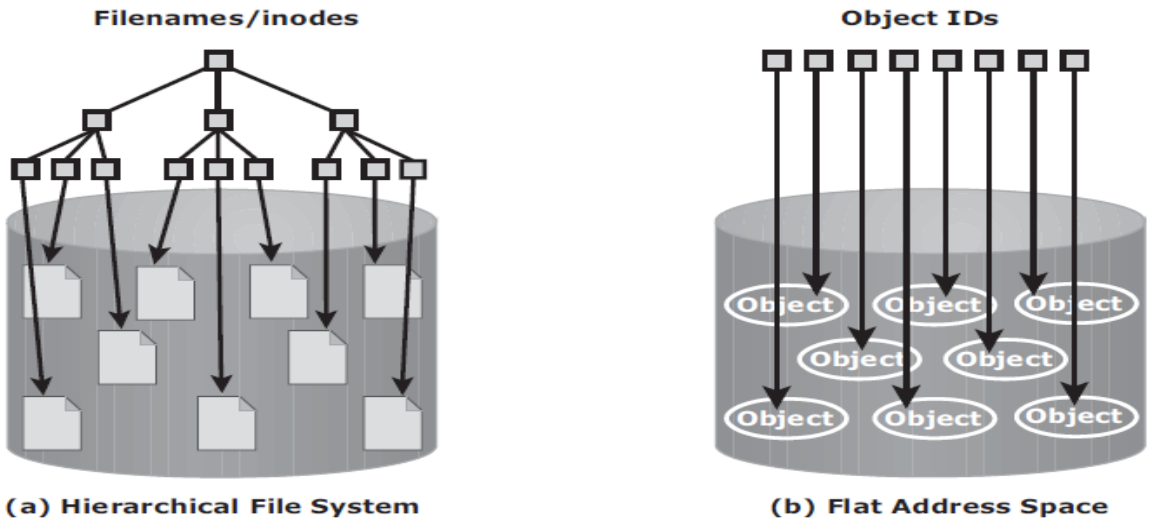
(a) Before File-Level Virtualization

(b) After File-Level Virtualization

Chapter 8. Object-Based and Unified Storage

Object-Based Storage Devices

هي اجهزة تقوم بتنظيم وتخزين الUnstructured Data مثل الgraphics & office documents, movies كـ objects. وتستخدم مساحة تخزين متساوية للتخزين وبالتالى لايوجد تسلسل هرمي فى تنظيم وتخزين البيانات.

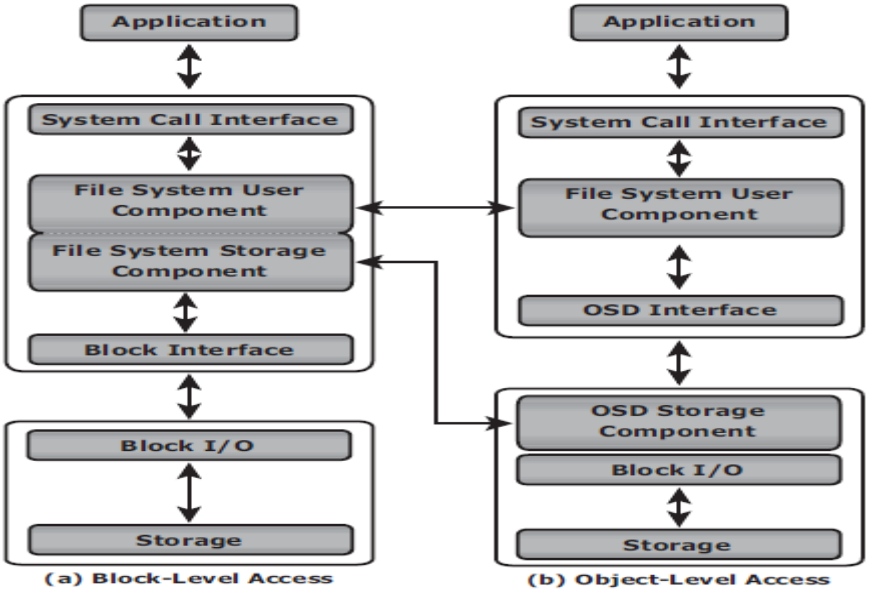


كل opject يكون له رقم محدد يسمى بال opject ID ويحتوي ايضا على user data و metdata مثل (size, date, ownership,..) بالاضافة الى بعض الattributes مثل (retention, access pattern,...).

- Object-Based Storage Architecture

اوامر الI/O داخل الBlock access التقليدية تمر بعدة مراحل حيث يتم انشاء الI/O من خلال الapplication وتمر من خلال الfile system وصولا الى الdisk drive.

عندما يستقبل الfile system الI/O من الapplication يقوم بعمل مقارنة مع الI/O القادمه الى ال disk blocks, ويتم استخدام ال block interface لارسال الI/O الى الstorage device خلال الشبكة.

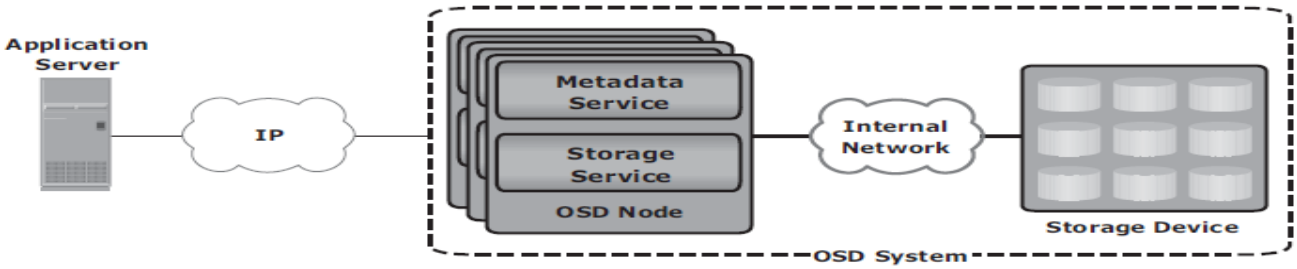


يتكون الfile system من مكونين هما ال user component والذى يقوم بتنفيذ الوظائف المختلفه مثل hierarchy management, naming, and user access control ويتكون ايضا منstorage component والذي يقوم بتوصيل الfiles الى الphysical location للdisk drive.

عندما يقوم الapplication بمحاولة الوصول الى بيانات مخزنة على OSD يقوم بارسال طلب الى ال user compenent للfile system والتى تقوم بالتواصل مع الstorage device من خلال الOSD interface والتى تحتوى على ال storage compenet والتى تكون مسؤوله عن عملية تنظيم الوصول الopject الى الstorage device

- Components of OSD

تتكون من ثلاث مكونات هي الnodes, private network & storage



ال OSD system يحتوى على node او اكثر والnode عباره عن سيرفر الذي يقوم بالعمليات التشغيلية للOSD وتقديم خدمات لتخزين واستعادة البيانات. حيث يقوم الOSD node بخدمتين رئيستين هما:

الmetdata service وهي المسؤوله عن انشاء الopject ID وتحتوى ايضا على الattributes للداتا وتقوم بالمقارنه بين الopject ID وال file system namespace.

يقوم ايضا بالStorage service والذي يقوم بادارة مجموعة الديسكات التى تخزن عليها البيانات.

الOSD node يتصل بالstorage عن طريق شبكة داخلية من خلال node-to-node connectivity.

- Object Storage and Retrieval in OSD

عملية التخزين للopjects داخل الOSD تتم على مراحل كما يلي:

1- يقوم الapplication server بارسال الfile للتخزين فى الOSD node.

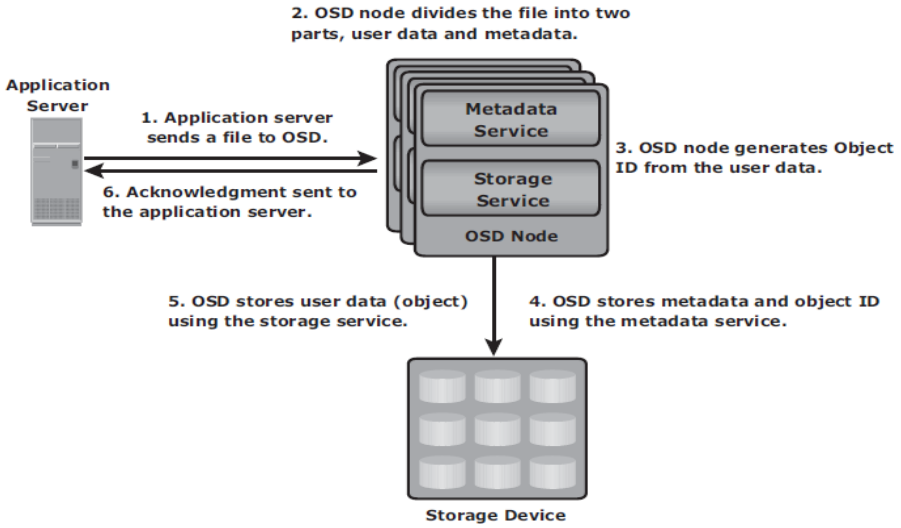
2- يقوم الOSD node بتقسيم الfile الى جزئين metdata & userdata.

3- يقوم الOSD node بانشاء الopject ID باستخدام algrothim مخصص.

4- الOSD node يقوم بتخزين الmetdata & opject ID لاستخدامها مستقبلا باستخدام الmetdata service.

5- الOSD node يقوم بتخزين الuser data (opject) داخل الstorage device باستخدام الstorage service.

6- يتم ارسال acknowledgment الى الapplication server بان الopject تم تخزينه.



بعد تخزين الopject بنجاح يكون جاهز للاسترجاع ويتم الاستعادة على مراحل ايضا كما يلي:

1- يقوم الapplication server بارسال read request الى الOSD system.

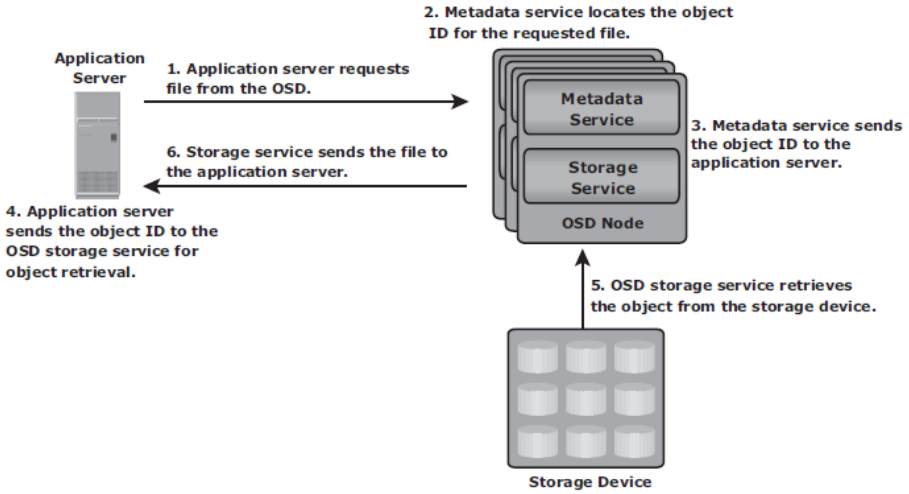
2- الmetdata service تقوم باستعادة ال opject ID للملف المطلوب.

3- تقوم الmetdata service بارسال الopject ID الى الapplication server.

4- يقوم الapplication server بارسال الopject ID الى الOSD storage لاستعادة الملفات.

5- الOSD storage service يقوم باستعادة الopject من الstorage device.

6- يقوم الstorage device بارسال الملفات الى الapplication server.



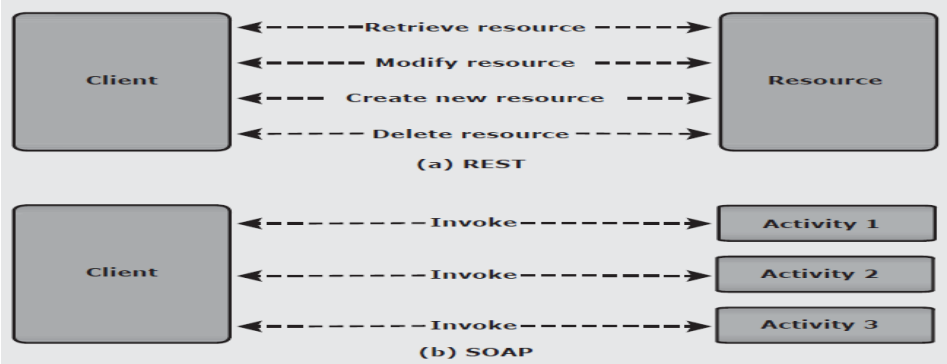
- Benefits of Object-Based Storage

تقدم اجهزة الOSD العديد من المزايا بالنسبة للunstructured data من خلال انظمة التخزين التقليدية.
حيث ان نظام التخزين المثالى يجب ان يوفر performance, scalability, security, and data sharing across multiple platforms. ولكن انظمة التخزين التقليدية مثل الSAN & NAS لاتوفر جميع هذه المزايا فى نظام واحد. حيث تقوم بالعديد من المزايا مثل

- Security and reliability
- Platform independence
- Scalability
- Manageability

- Common Use Cases for Object-Based Storage

من اهم الاسباب الشائعه لاستخدام اجهزة الOSD هي الdata archival solution. حيث ان سلامة وحماية البيانات هي اهم جزء مطلوب داخل الdata archiving solution.
حيث يقوم الOSD بتخزين البيانات داخل opject form وربطهم ب unique opject ID والذي يضمن امان عالى للبيانات بالاضافة بانه يوفر الscability & data protection, هذه القدرات تجعل الOSD الخيار الافضل للتطبيق لانظمة الdata archiving وتخزين الfixed content.
Content addressed storage (CAS) هو نوع خاص لاجهزة ال OSD صمم لتخزين الfixed content.
ايضا استخدام اخر للOSD وهي cloud-based storage حيث يقوم باستخدام web interface للوصول الى الstorage resources اضافة الى security, scalability and automated data management يجعله الخيار الافضل بالنسبة للcloud-based storsge.
ال OSD يدعم ايضا الweb service access من خلال representational state transfer (REST) و ايضا الsimple object access protocol (SOAP).
Note: الRESET هو بروتوكول طور للاستخدام فى الweb application حيث يقوم بعدة عمليات عن طريق الويب (web service) مثل retrieving, modifying, creating, and deleting resources.
اما الSAOP هو XML-based protocol يمكن من ربط قواعد البيانات والتواصل بين ال web application المستعمله على الانظمه التشغيليه المختلفه عن طريق عمل encoding لل HTTP & XML



Content-Addressed Storage

هو opject-based storage device صمم خصيصا لتخزين واستعادة المحتوى الثابت fixed content (البيانات الغير متوقع تحديثها) حيث انها تقوم بتخزين البيانات بناءا على محتواها وليس موقع التخزين.

الCAS يقوم بتحزين بيانات المستخدمين والattributes الخاصه بهم كـ opject ويخصص لها unique address يعرف بال content address (CA) وهذ يوفر ادارة مركزية افضل للتخزين.

طرق الوصول للبيانات فى الCAS تختلف عن الOSD حيث ان الapplication server يصل الى ال CAS devices عن طريق CAS API يعمل على السيرفر, ولكن يتشابها معا فى طريقة التخزين.

من اهم مزايا الCAS فى تخزين المحتوى الثابت fixed content:

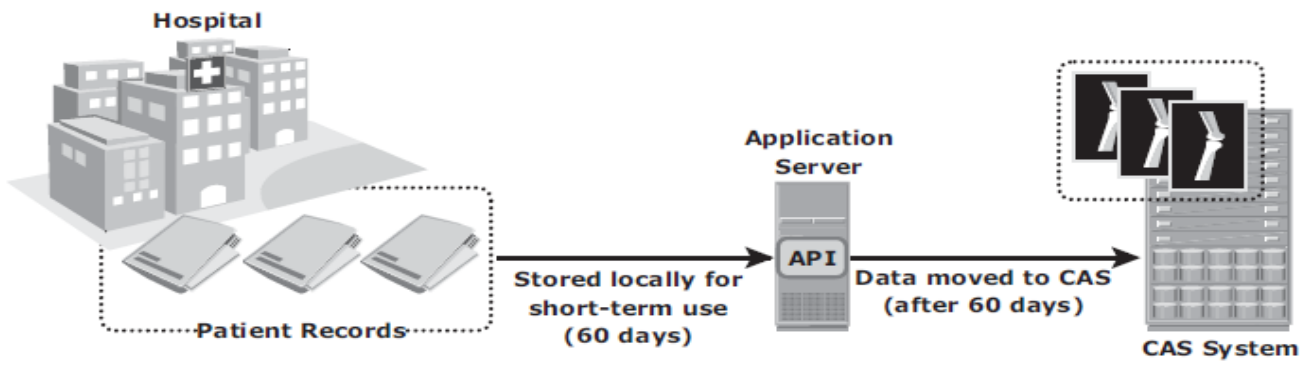
- Content authenticity
- Content integrity
- Location independence
- Single-instance storage (SIS)
- Retention enforcement
- Data protection
- Fast record retrieval
- Load balancing
- Scalability
- Event notifi cation
- Self diagnosis and repair
- Audit trails

CAS Use Cases

الكثير من المنظمات تستخدم ال CAS Solutionلمواجهة تحديات الاعمال ومن الامثله على ذلك:

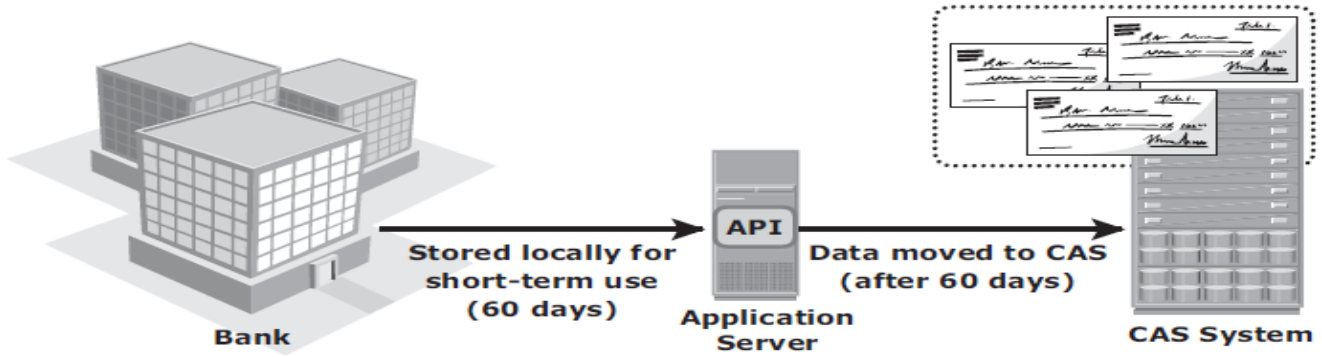
- **Healthcare Solution: Storing Patient Studies**

المراكز الصحية الكبرى تستقبل مئات المرضى يوميا وتقوم بانشاء العديد من التقارير الصحية التى تحتوى على بيانات المريض وصور التقارير ويتم تخزين هذه المعلومات online لمدة من الوقت للاستعمال المباشر, ثم يتم تخزين هذه السجلات والاحتفاظ بها للعديد من السنين حتى لو كانت غير مطلوبه بالنسبة للمريض.



- **Finance Solution: Storing Financial Records**

ايضا فى البنوك المعاملات اليوميه والتقارير المحاسبية يتم انشاء العديد منها يوميا ثم ترسل لل archive services من خلال IP network ثم يتم تخزينها بعد 60 يوم ايضا على الCAS System اذا تم الحاجه لها فيما بعد.



Unified Storage

الunified storage تقوم بتوحيد عمليات الblock, file & opject access داخل storage solution واحد.

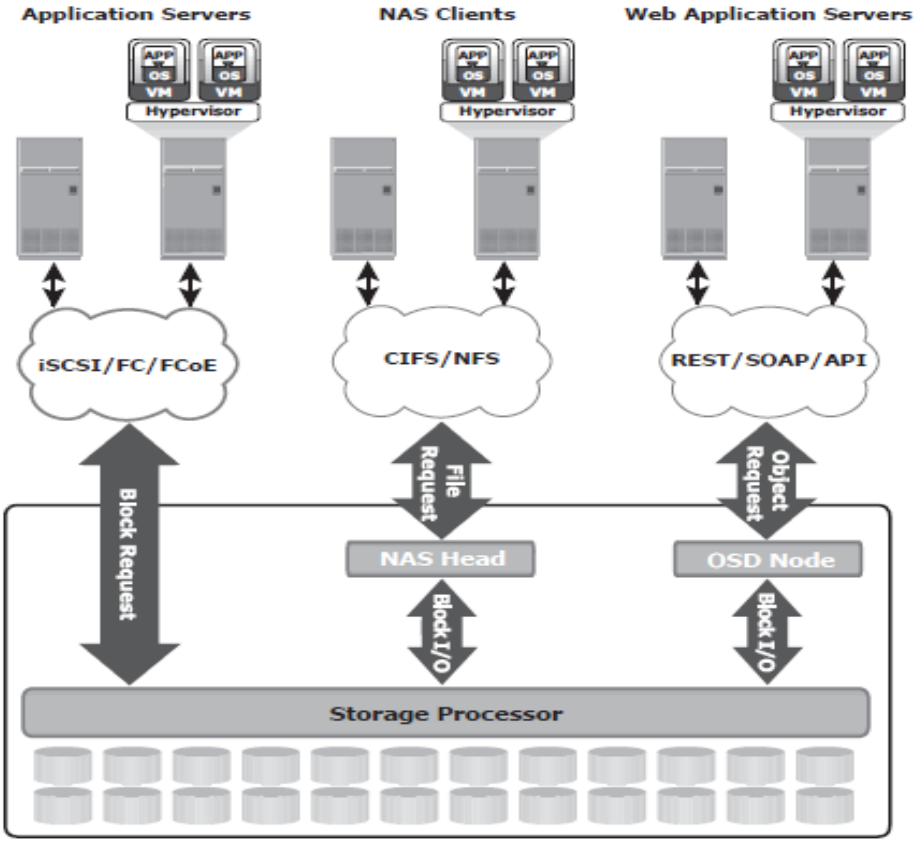
وتدعم العديد من البروتوكولات مثل CIFS, NFS, iSCSI, FC, FCoE, REST & SOAP.

- **Components of Unified Storage**

- يتكون من مجموعه من المكونات الرئيسية مثل storage controller, NAS head, OSD node & storage
- ال Storage controller يدعم الblock-level access للapplication server من خلال استخدام بروتوكولات مثل iSCSI, FC, or FCoE وهو المسؤول ايضا عن ادارة الback-end storage pool داخل الstorage system.

ايضا الStorage Controller يقوم بانشاء LUNs وعرضها على ال application servers, NAS heads & OSD node حيث تظهر وكانها local physical disks ويتم تكوين الfile system داخل هذه الLUNs بحيث تكون متاحه الى الapplication لتخزين البيانات.

- الNAS head هو dedicated file server يوفر الوصول للfiles من خلال الNAS client ويكون موصول على الstorage من خلال الstorage controller باستخدام بروتوكول ال FC or FCoE ومن الممكن وجود اكثر من NAS head من اجل عملية الRedundancy, وتظهر الLUN's ايضا على الNAS head كانها physical disks يقوم الNAS head بتكوين الfile systems داخل هذه الLUNs وانشاء NFS, CIFS or mixed share وتصديرها الى الNAS clients.
- الOSD nodes موصوله بالstorage عن طريق الstorge controller باستخدام الFC or FCoE ايضا والLUNs المخصصه لها تظهر كـphysical disk يتم اعدادها من خلال الOSD nodes لتتمكن من تخزين البيانات عن طريق الweb application server باستخدام بروتوكولات الREST, SOAP or API.



- Data Access from Unified Storage

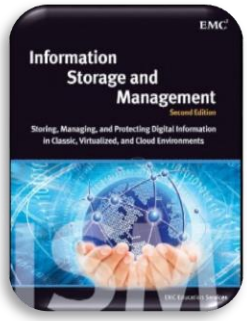
فى الunified storage عمليات الblock, file, and object requests تمر على مسارات I/O مختلفة للوصول الى الstorage

- **Block I/O request:** يقوم السيرفر بارسال block request الى الstorage باستخدام بروتوكول FC, FCoE or Iscsi حيث يكون متصل مباشرة بالstorage controller
- File I/O request: يتم ارسال file request من خلال الNAS client الى NAS head باستخدام NFS or CIFS ثم يقو الNAS head بتحويلها الى block request الى storage controller والعكس.
- Object I/O request: يتم ارسال opject request من خلال الweb application server باستخدام REST or SOAP الى الOSD node والذي يقوم بتحويلها الى block request الى storage controller والعكس.

End of Part 1 ☺

➢ مصادر الكتاب

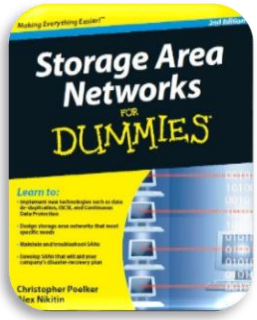
❶



كتاب Information Storage and Management second edetion

من شركة **EMC**

❷



كتاب Storage Area Networks for Dummies 2nd edetion

تحياتي

حسن عيسى

22-11-2017