Protección de Datos y Backup

José M. Peña

<jmpena@fi.upm.es>

Contenidos

1. Definiciones y términos
2. Requisitos de usuario:
   - Plazos de recuperación
   - Planificación de la organización
3. Granularidad de la copia.
4. Topologías de sistemas de backup
5. Dispositivos físicos
6. Fabricantes (hardware/software)

Definiciones

- Un **backup** es una **copia adicional** de la información que puede utilizarse con fines de **recuperación y restauración** ante fallos.
  - Su utilización se hace cuando la copia original está inutilizada o corrupta.
  - La copia puede ser:
    - Copias de los ficheros en instantes de tiempo determinados.
    - Copias especulares de los datos originales completamente sincronizados.

Tipologías de Backups

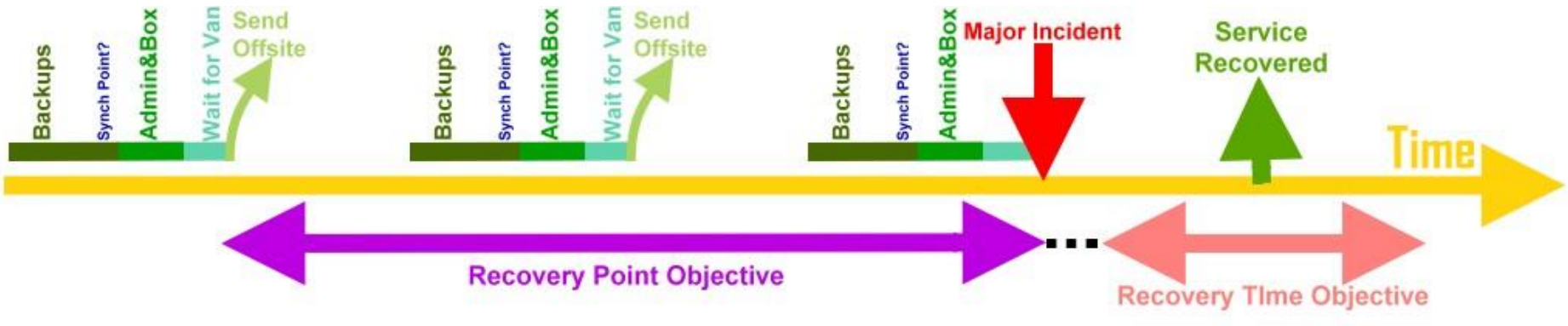
- Según necesidades:
  - **Copias para recuperación ante desastres**: El objeto es disponer de una copia que subsane la perdida potencial de datos valiosos para el usuario.
  - **Copias operacionales**: Se hacen para disponer de una instantánea de los datos del sistema en un momento determinado, con la intención de poder regresar a esa situación (sin necesidad de que haya un desastre):
    - E.g., versiones de un repositorio software.
  - **Copias reguladas**: Se realizan para cumplir con normativas legales que exigen el almacenado de datos históricos durante un periodo de tiempo. (LOPD en España).

Requisitos de Usuario

- Plazos de recuperación
- Instalaciones:
  - Original
  - De recuperación
- Elementos a recuperar:
  - Ficheros con poca variación.
  - Ficheros con mucha variación.
- Temporizaciones:
  - Cuándo se hacen los backups.
  - Cuánto tiempo dura la operación de copia.
  - Durante cuánto tiempo se guarda copias.

Plazos de Recuperación

- **Recovery Point Objective (RPO):** Periodo máximo de tiempo en el cual se han podido ver afectados datos antes de un incidente.
- **Recovery Time Objective (RTO):** Periodo máximo de tiempo en el que es asumible tener los sistemas de información parados después de un incidente.



© Wikipedia

Planificación de la Organización

- La empresa debe incluir en sus procedimientos internos diferentes documentos de reglamentación:
  - Plan de continuidad del negocio (*business continuity plan*): Que indica la exposición de la organización a amenazas internas y externas y las contramedidas para prevención y recuperación. Incluye:
    - Análisis de impacto en el negocio (*business impact analysis* – BIA): Se diferencian sistemas críticos de no críticos y donde se definen, por ejemplo RTP y RPO.
    - Análisis de amenazas y riesgos (*threat and risk analysis* – TRA): Se identifican los tipos de amenazas.

Business Continuity Planning, FEMA, Retrieved: June 16, 2012

http://www.ready.gov/business/implementation/continuity

Planificación de la Organización

- Plan de recuperación ante desastres (*disaster recovery plan*): Determina los pasos a realizar para realizar las acciones de recuperación ante un incidente.
  - Incluye las prioridades de esas tareas, el entrenamiento de los grupos participantes y los canales de comunicación.
  - Debe realizarse un ensayo de recuperación de forma periódica para verificar la integridad de datos y la agilidad de los procedimientos.
  - Lleva asociado acciones relativas al inventario sistemático de equipos, las pólizas de seguro y garantías de los mismos y un listado de números de emergencias y similares.

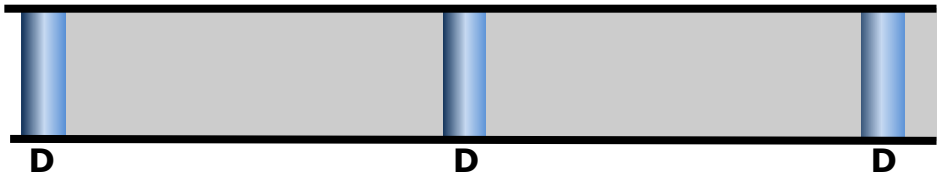
Disaster Recovery Planning Process. Geoffrey H. Wold. Disaster Recovery Journal. Adapted from Vol. 5 #1. Disaster Recovery World© 1997
http://www.drj.com/new2dr/w2_002.htm

Tipologías de Backups

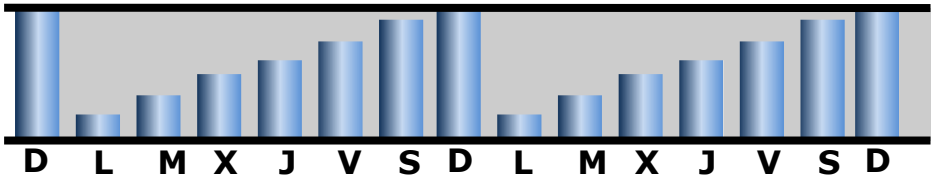
- Por Granularidad:
  - "*Cuándo y de qué se hace copia*".
  - Se determinan diferentes tipos de backups de acuerdo a cuáles son los ficheros copiados.
  - Los diferentes tipos de backups se hacen en "*ciclos de backup*"
- Por Operatividad del sistema:
  - "*En qué estado está el sistema cuando se realiza la copia*".
  - Se determina si es necesario detener la operativa del sistema (dejar de proporcionar servicio) para hacer el backup.

Ganularidad

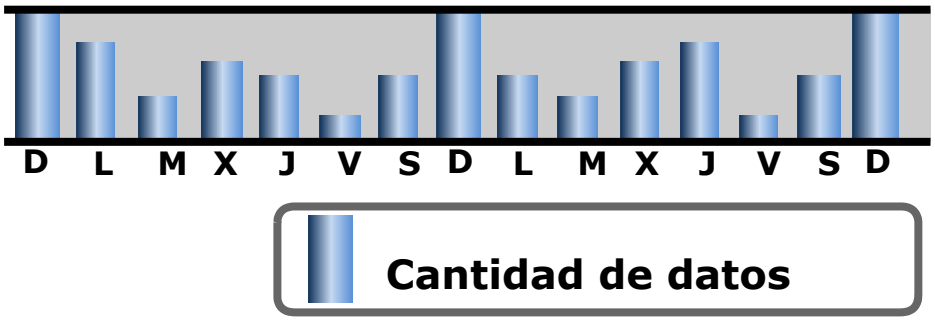
Backup completo: Se realiza una copia integral de los datos, copiando todos los contenidos de los sistemas a mantener.



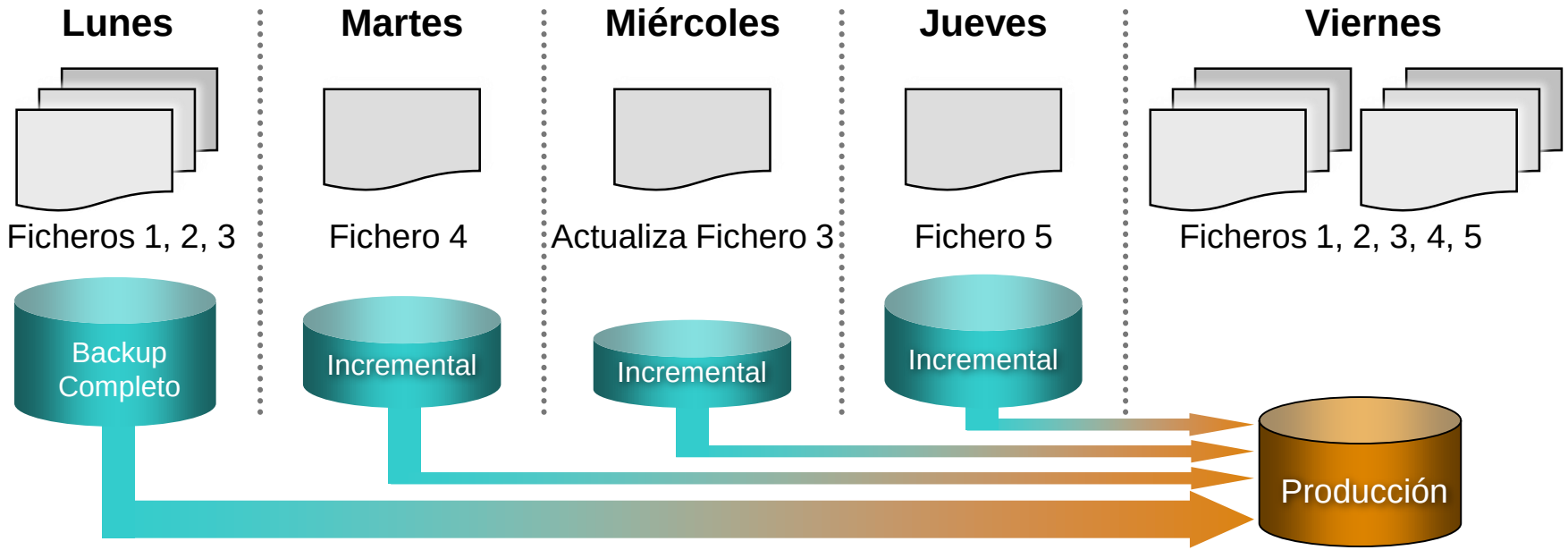
Backup diferencial: Partiendo de una copia de backup completa, se realiza una copia de todos los datos modificados desde que se hizo ese backup completo.



Backup incremental: Partiendo de una copia de backup completa, se realiza una copia sólo de los datos modificados desde el último backup (sea completo o incremental).



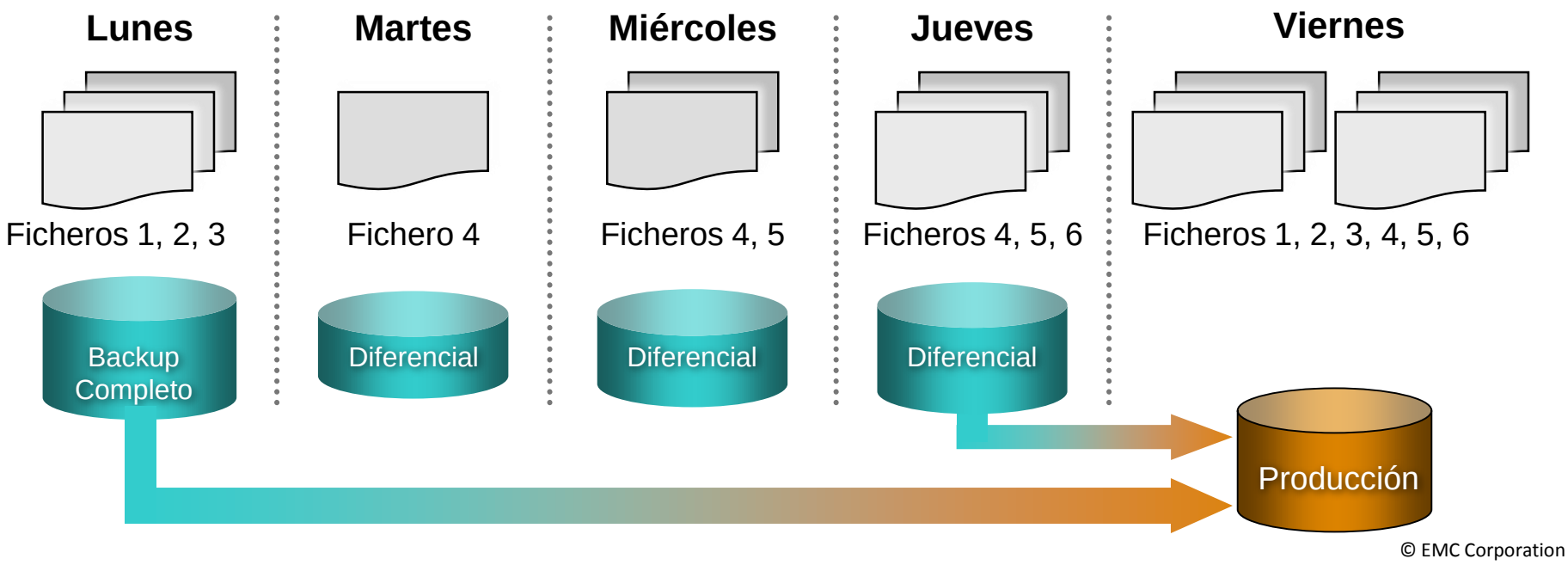
Recuperación de un Backup Incremental



© EMC Corporation

- Aspectos clave:
  - Los ficheros que se han modificado después del último backup se guardan.
  - Se realizan un número menor de copias de ficheros, que requieren una menor capacidad de almacenamiento y backups más rápidos.
  - Mayor tiempo de recuperación porque resulta necesario deshacer el último backup completo y todos los incrementales.

Recuperación de un Backup Diferencial



© EMC Corporation

- Aspectos clave
  - Se copian más ficheros, por lo tanto el backup lleva más tiempo y usa más espacio de almacenamiento.
  - Las recuperaciones son mucho más rápidas porque sólo conllevan recuperar el backup completo y el último de los diferenciales.

Operatividad del Sistema Durante Copia

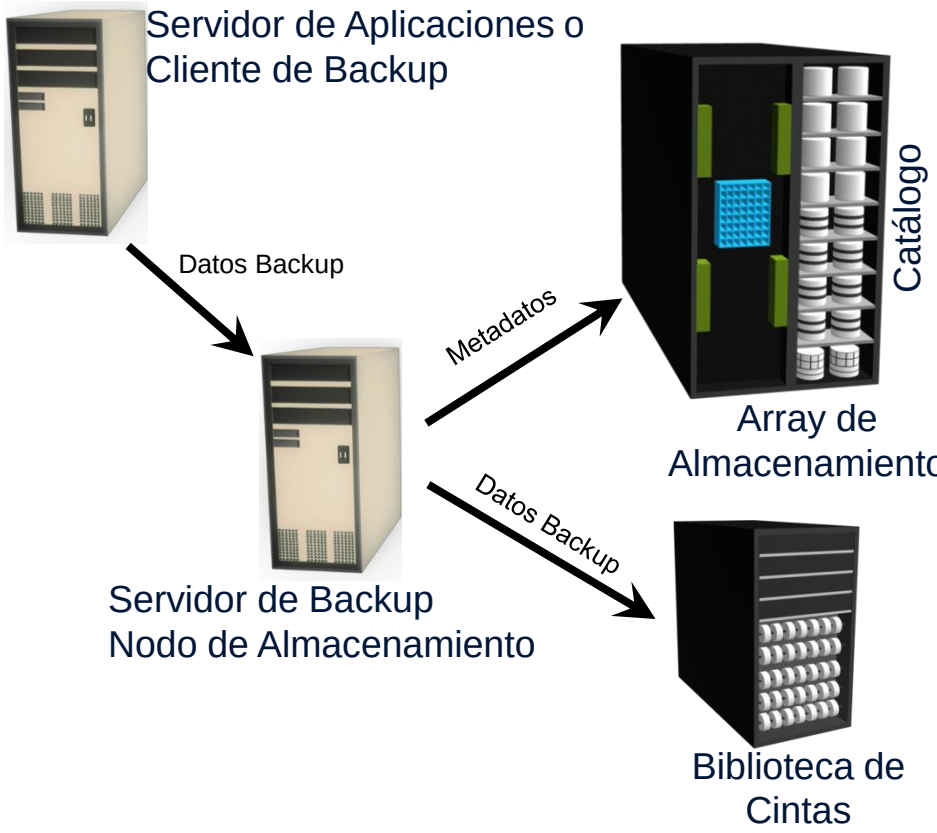
- **Backup frio (*cold*) u *off-line***: La operativa del sistema se detiene.
  - Entre el comienzo de la fase de copia y el final de la misma no se hace ninguna operación sobre los datos.
  - Requiere ventanas de tiempo para realizar esas copias que deben ser programadas y validadas.
  - No válido para sistemas 24x7 (e.g., un comercio on-line).
- **Backup caliente (*hot*) u *on-line***: La operativa del sistema no se detiene y la copia se hace con el sistema en producción.
  - Requiere fijar el instante de tiempo de referencia.
  - Gestionar no sólo los datos estables sino las modificaciones (log de operaciones) entre ese instante y el final de la copia.

Hot Backup

- Muy utilizado típicamente en bases de datos (pero también aplicable a sistemas de ficheros).
  - Se configura el sistema en modo *hot backup* (a veces llamado *point-in-time recovery*)
  - Se crea un log de operaciones (*redo log*) donde se almacenan todas las modificaciones que se piden sobre los datos al comenzar la copia:
    - Eso implica que los datos estables no se modifican por esas operaciones.
  - Al finalizar la copia el redo log se ejecuta y se aplican todos los cambios.
  - Durante la operación de copia el sistema funciona en modo degradado (peores prestaciones).

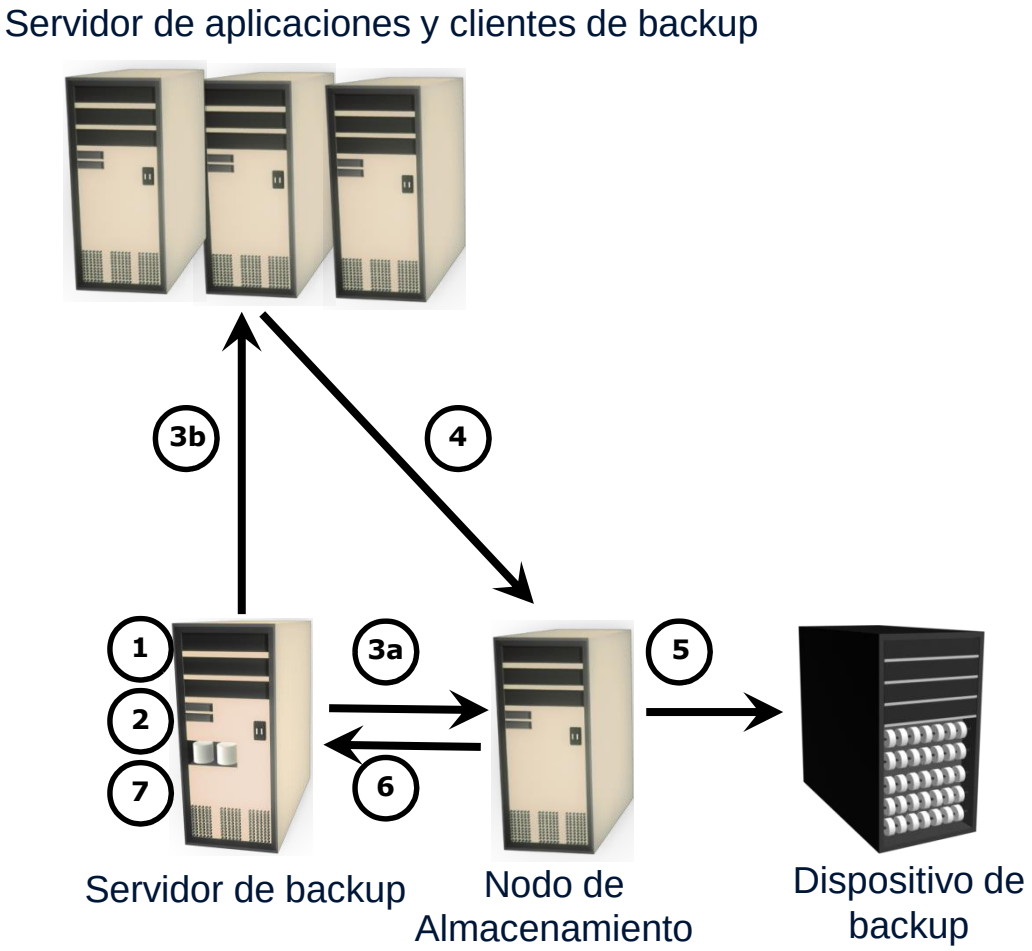
Arquitectura de un Sistema de Backup

- Cliente de backup
  - Manda los datos a copial al servidor de backup
- Servidor de backup
  - Puede ser uno de los nodos de almacenamiento en sistemas con varios de estos nodos.
  - Gestiona las operaciones de copia y mantiene un catálogo con los metadatos de la copia.
  - Si es uno de los nodos de almacenamiento se comunica con el dispositivo.
- Bibliotecas de cintas
  - Armario con almacenamiento secundario.



© EMC Corporation

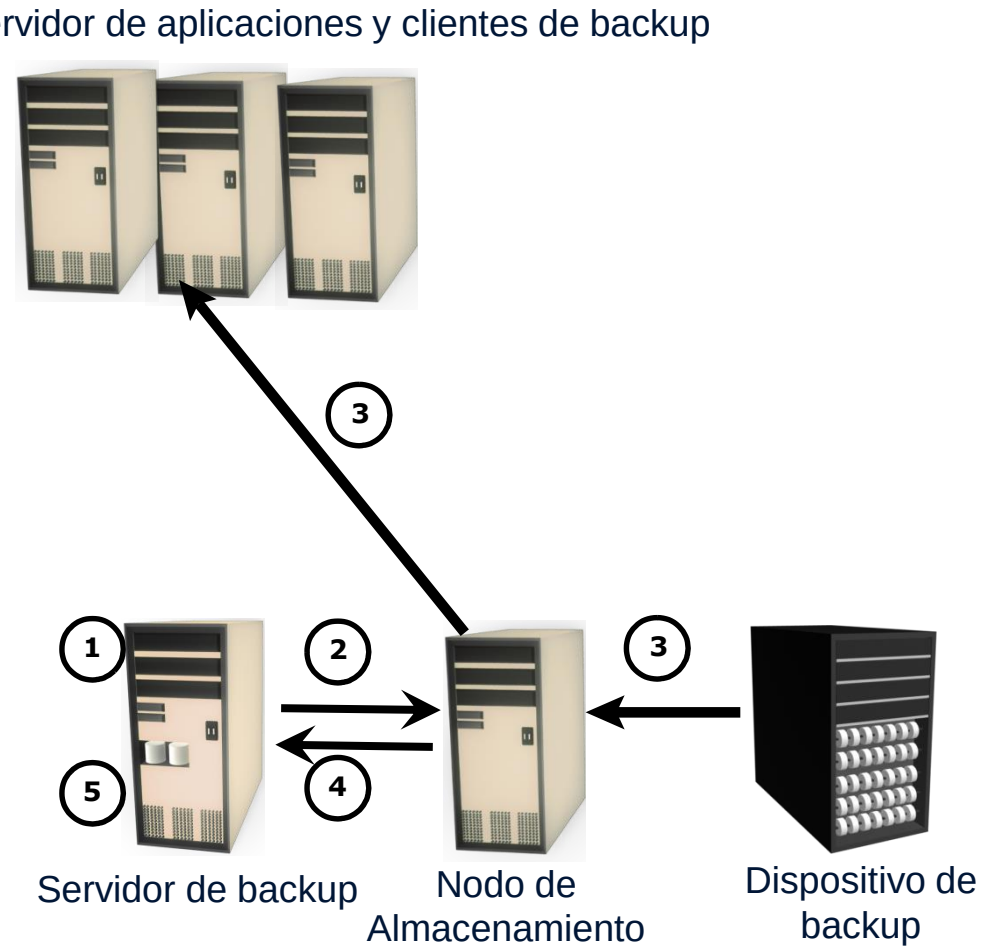
Operation de Copia



1. **Comienza un proceso de backup planificado**
2. **El servidor de backup recupera del catálogo la información relativa a la copia**
3. **3a El servidor le pide al nodo de almacenamiento que carge la cinta en el dispositivo de backup**
4. **3b El servidor da la orden a los clientes que le manden los metadatos al servidor y los datos al servidor de almacenamiento**
5. **Clientes manda datos al servidor de almacenamiento**
6. **El nodo de almacenamiento le manda los datos Al dispositivo de backup**
7. **El nodo de almacenamiento le remite la información sobre el número de cinta al nodo de backup**
8. **El servidor de back lo registra en el catálogo y actualiza el valor de estado del backup**

© EMC Corporation

Operación de Recuperación



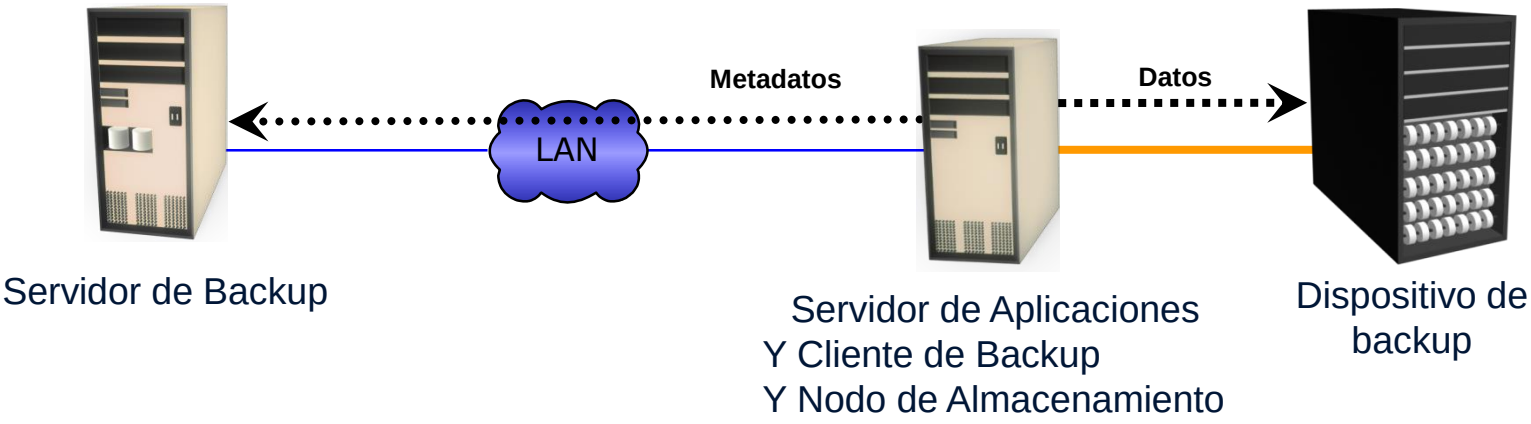
1. **El servidor recorre el catálogo de para identificar los datos a recuperar y el cliente que debe recibir los datos**
2. **El servidor da la orden al nodo de almacena-miento para que carge la cinta en el dispostivo**
3. **Se leen los datos y se mandan al cliente correspondiente**
4. **El nodo de almacenamiento le transmite al servidor los metadatos de backup recuperado**
5. **El servidor de backup actualiza el catálogo**

© EMC Corporation

Topologías de Backup

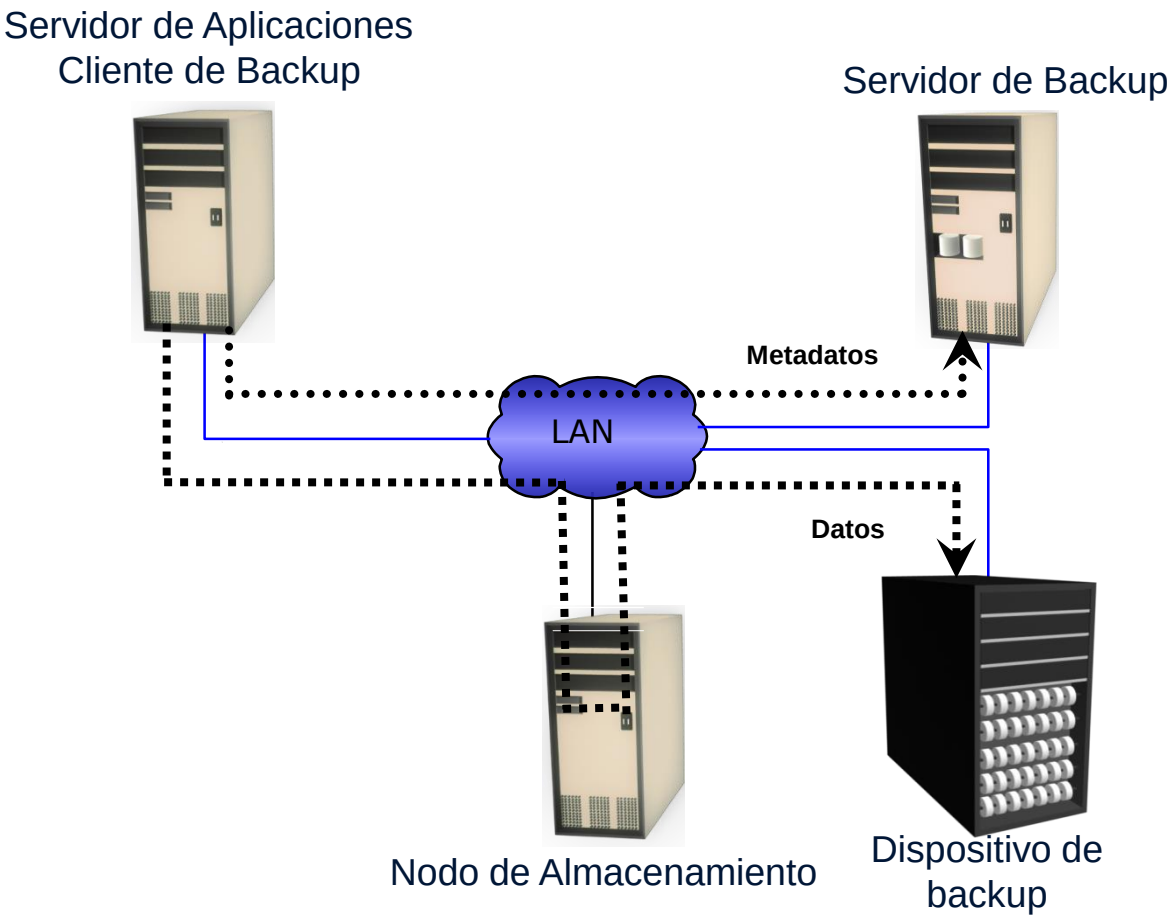
- Determinan la configuración de conexión de los sistemas operacionales de almacenamiento con el sistema de copias de backup.
  - Son muy dependientes de la topología de conexión de los sistemas de almacenamiento.
  - Tipos:
    - Backup de conexión directa.
    - Backup vía LAN (*Local Area Network*)
    - Backup vía SAN (*Storage Area Network*)
    - Backup mixto
    - Backup vía NAS (*Network-Attached Storage*): Con/sin servidor

Backup de Conexión Directa



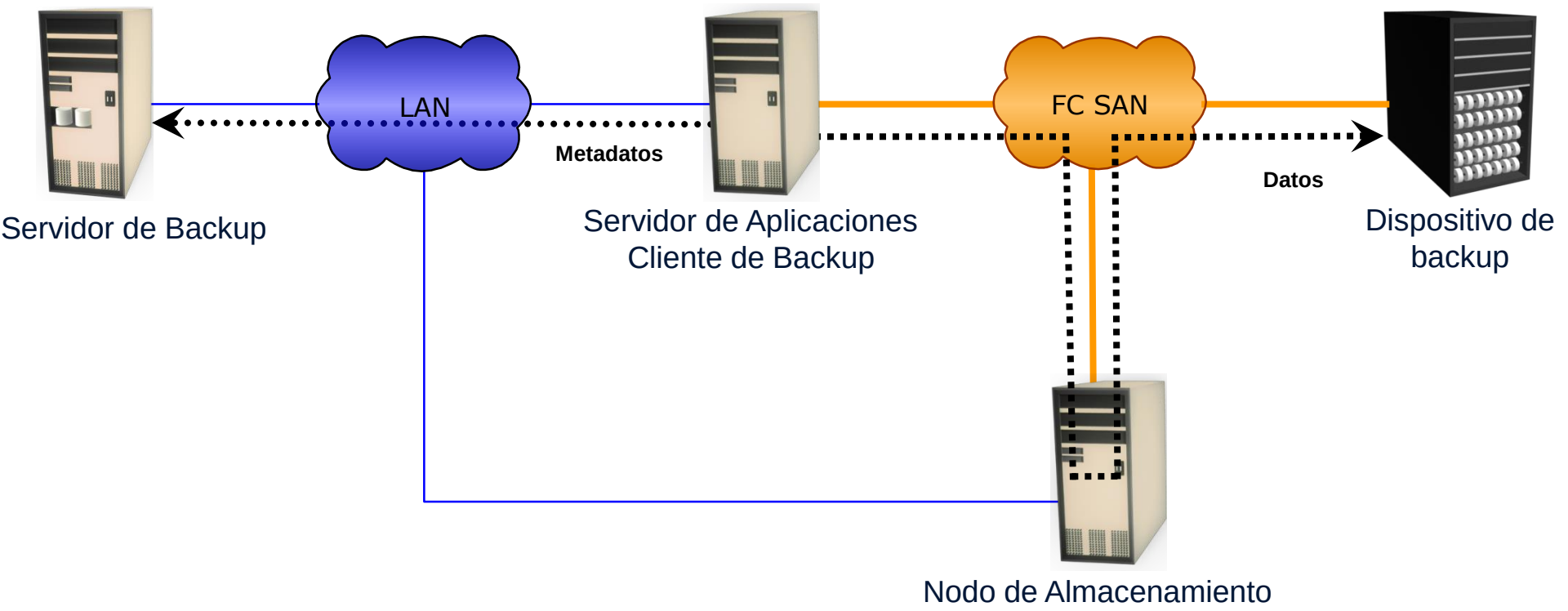
© EMC Corporation

Backup Vía LAN



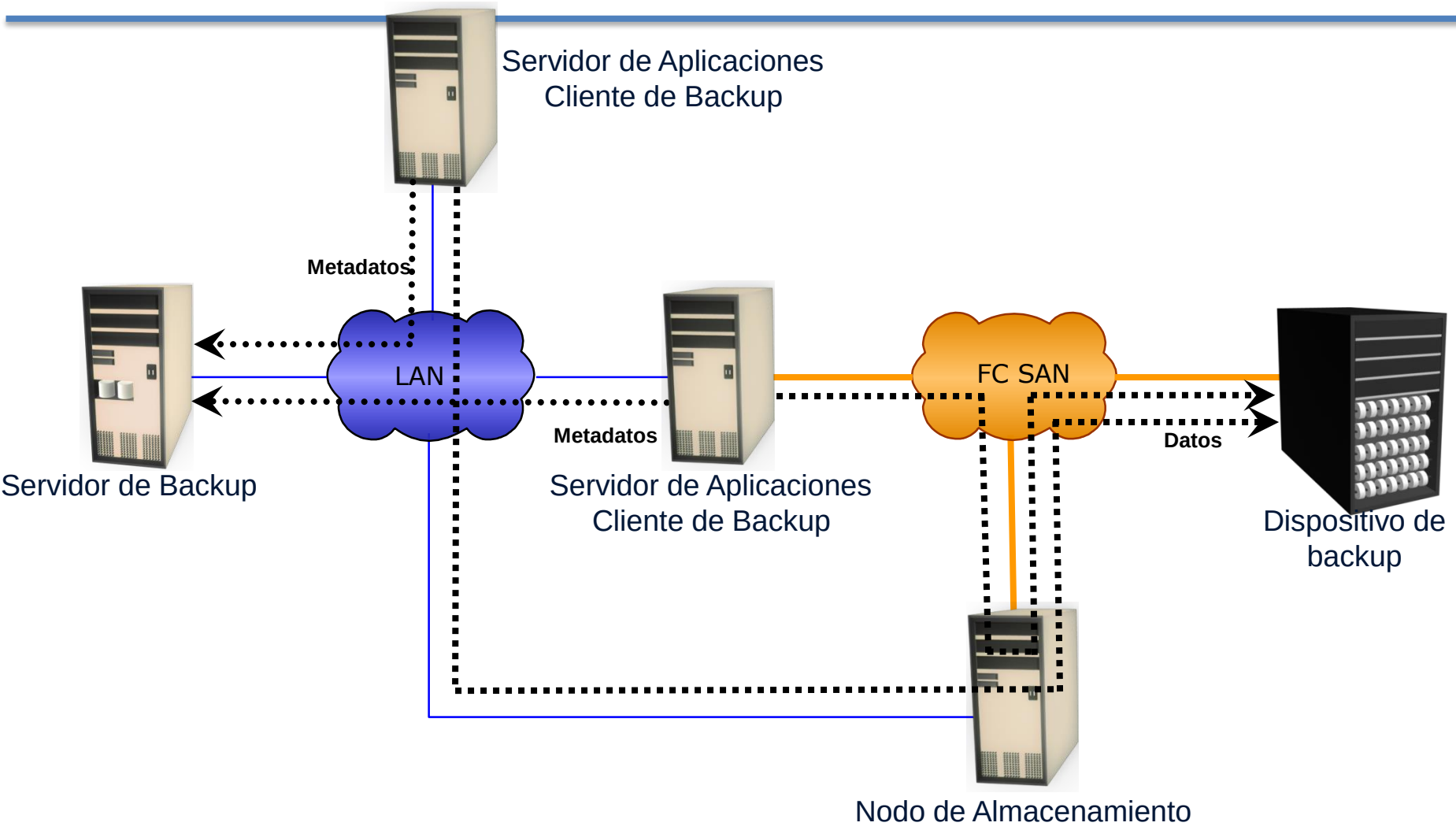
© EMC Corporation

Backup Vía SAN (Sin LAN)



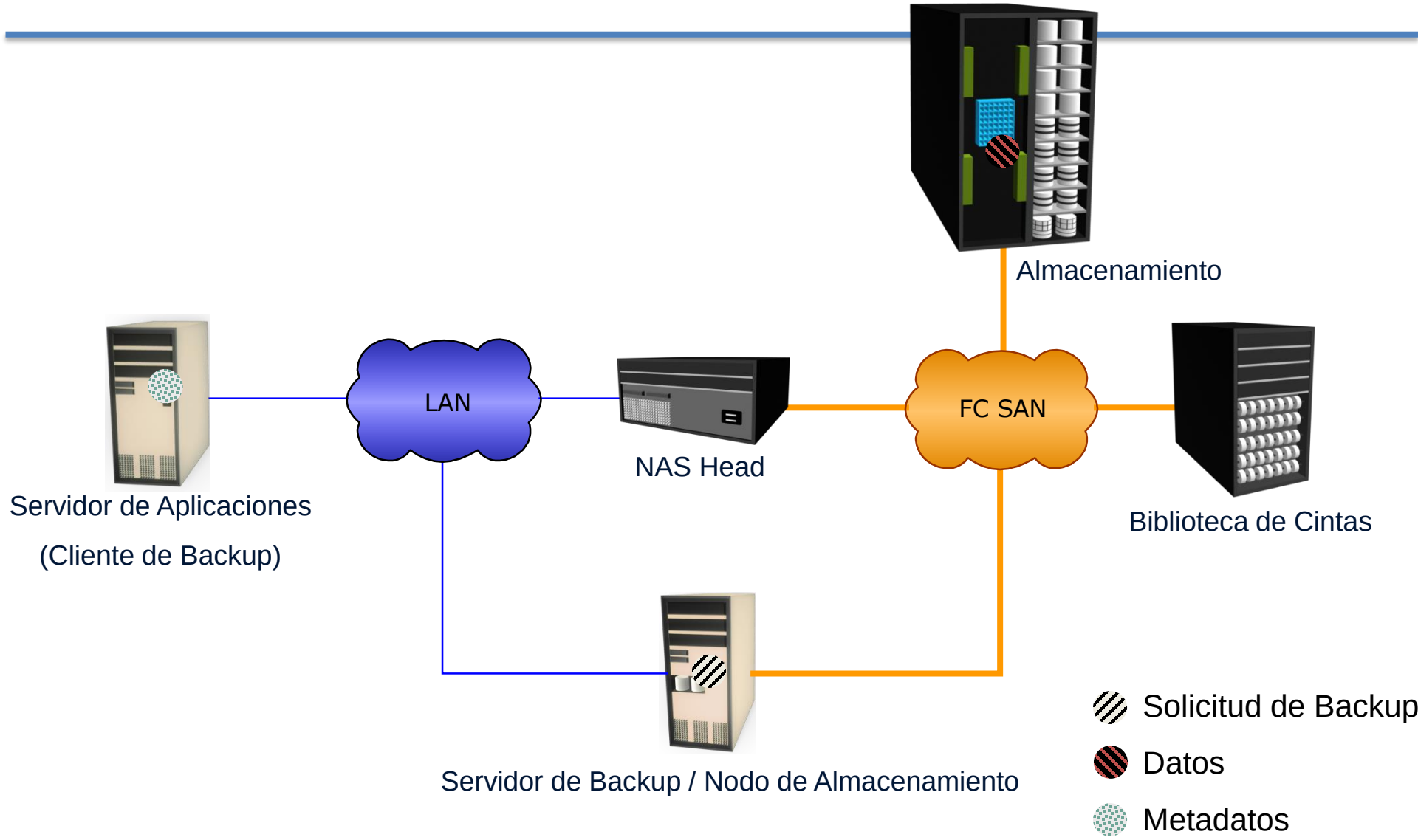
© EMC Corporation

Backup Mixto



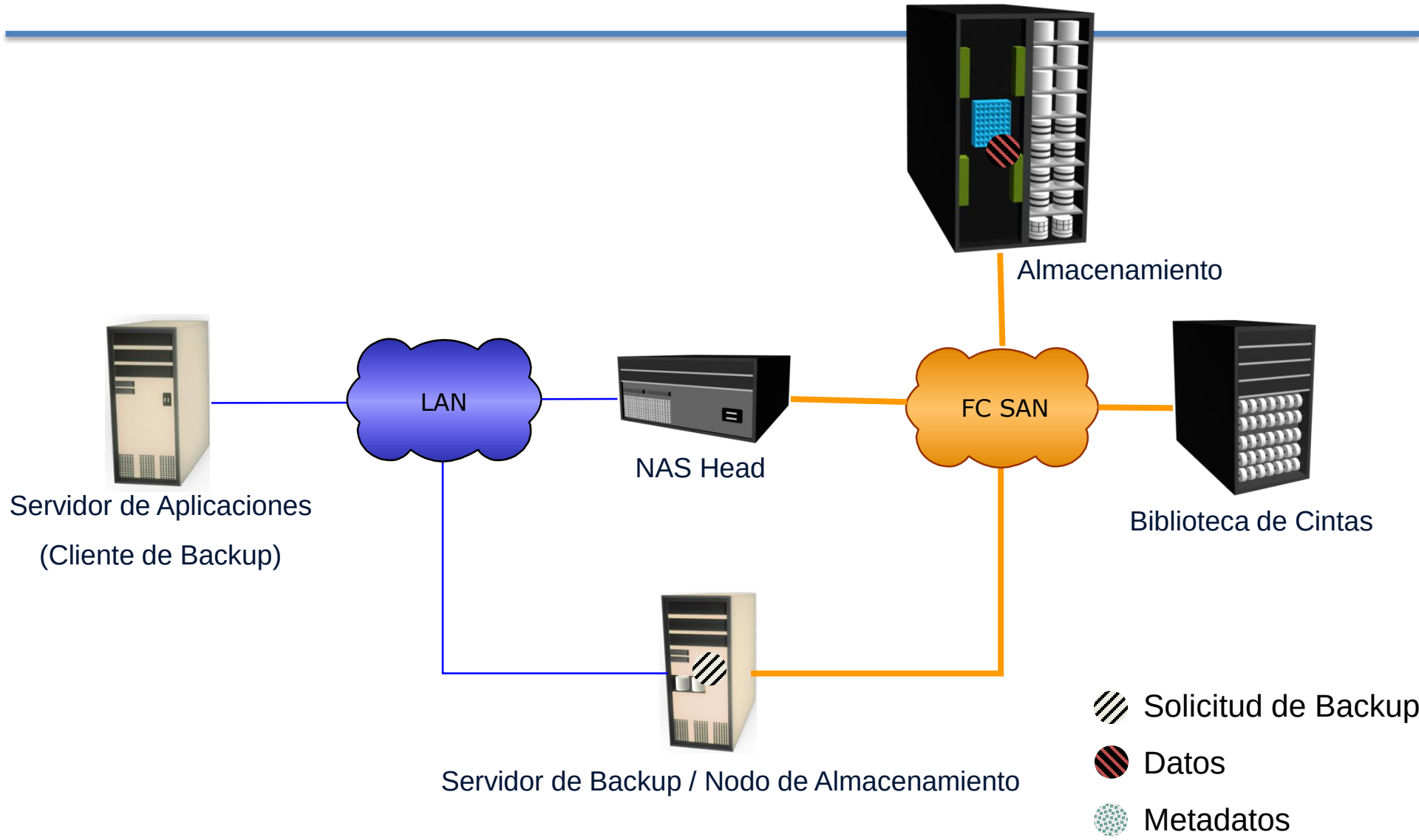
© EMC Corporation

Backup Vía NAS – Con Servidor



© EMC Corporation

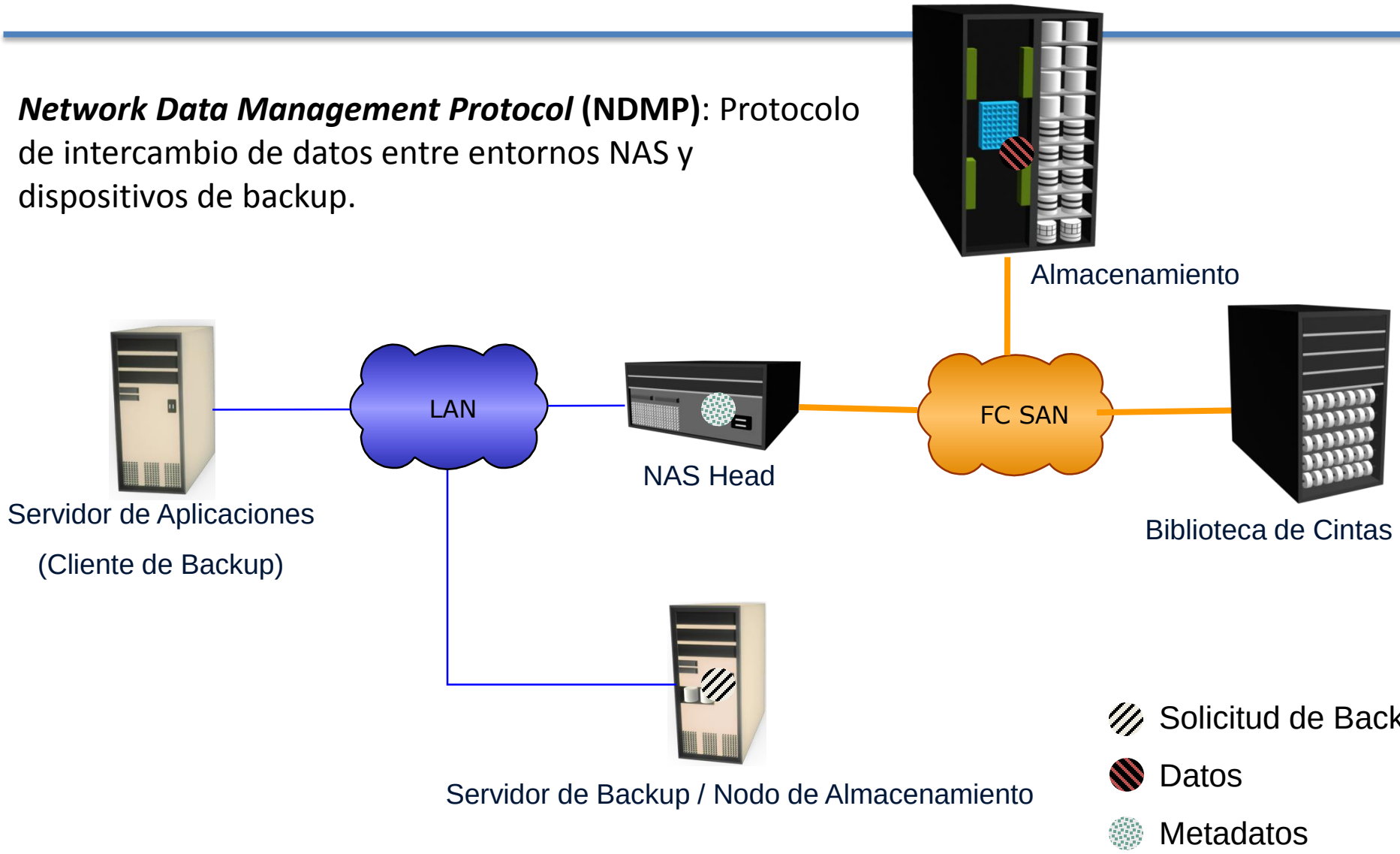
Backup Vía NAS – Sin Servidor



© EMC Corporation

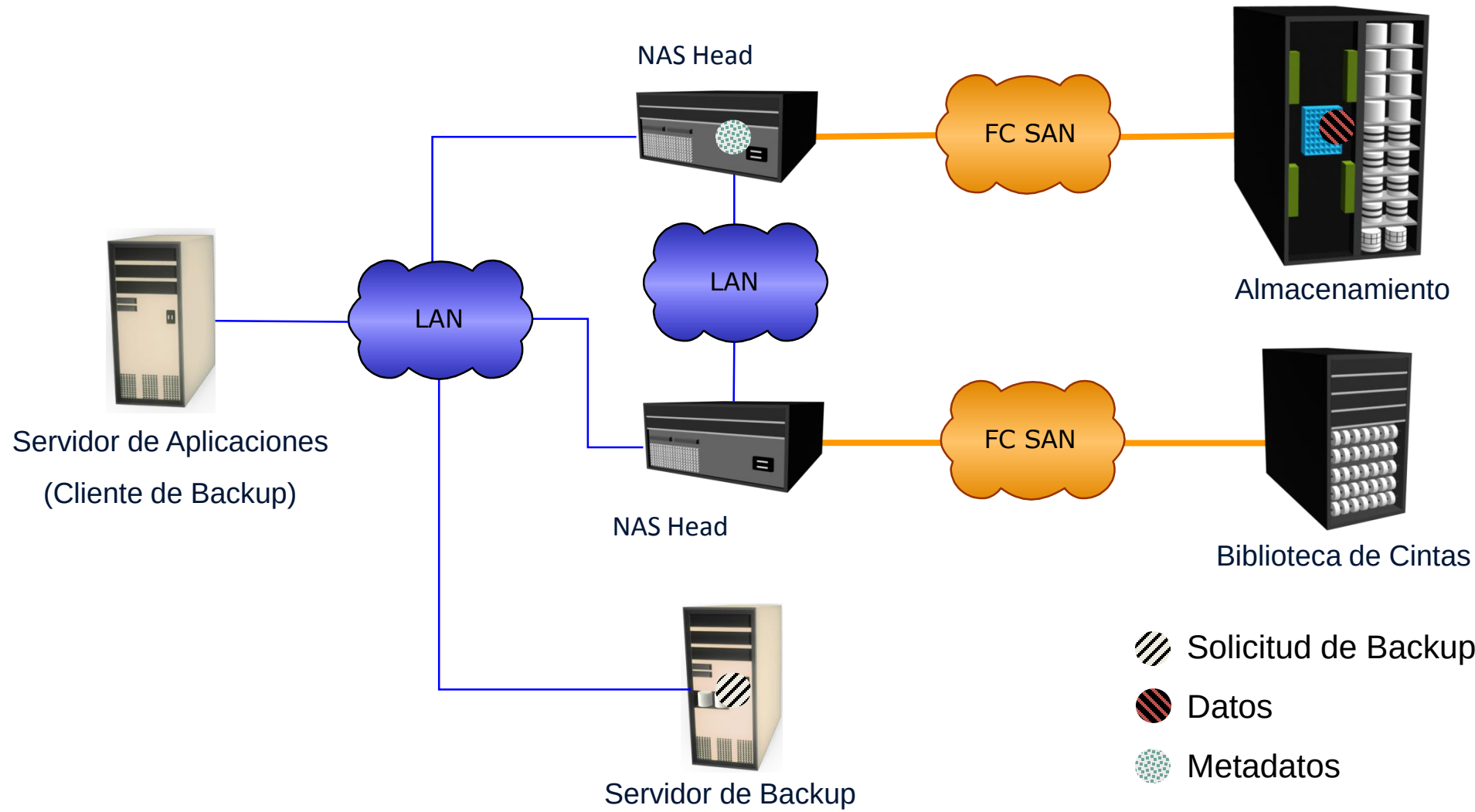
Backup Vía NAS – NDMP de 2 Vías

***Network Data Management Protocol* (NDMP)**: Protocolo de intercambio de datos entre entornos NAS y dispositivos de backup.



© EMC Corporation

Sistema de Backup como Dispositivo NAS NDMP de 3 Vías



© EMC Corporation

Dispositivos Físicos de Almacenamiento

- Tienen que considerarse aspectos de eficiencia, capacidad y durabilidad del medio.
- Su localización:
  - *On-line*: Directamente accesible, por lo general en disco.
  - *Near-line*: Accesible pero con una latencia mayor, por lo general cinta en un robot o biblioteca de backup.
  - *Off-line*: No está accesible sin intervención humana, requiere transportar el medio de almacenamiento desde otra localización.
  - Centro de recuperación de desastres: Instalación que dispone de una copia de datos sincronizada con alta frecuencia (o especular) y que puede estar operativa en un intervalo de tiempo mínimo.

Dispositivos Físicos de Almacenamiento

- Tipo de dispositivo:
  - Cinta:
    - Bajo coste, acceso lento.
  - Disco:
    - Alto coste, acceso rápido.
    - Por lo general se usan extensiones RAID para dar mayor fiabilidad.
  - Cintas virtuales:
    - Usualmente proporcionada por dispositivos de backup de gama alta.
    - Compuesta por discos que “cachean” contenidos de una biblioteca de cintas asociada.
    - Recuperación con prestaciones simulares (salvo cacheado), pero backup mucho más eficiente (disco→ disco → cinta).

Fabricantes y Software

- IBM:
  - Tivoli
- EMC:
  - EMC Networker / RecoverPoint
- Hitachi (HDS):
  - True Copy
- HP:
  - HP Data protector

- Software de propósito general:
  - Libre:
    - AMANDA
    - BACULA
  - Propietario:
    - NovaBACKUP
    - Acronis
    - Symantec/Veritas