Improving performance

- Reduce number of context switches
- Reduce data copying
- Reduce interrupts by using large transfers, smart controllers, polling
- Use DMA
- Balance CPU, memory, bus, and I/O performance for highest throughput

Disk scheduling

- Schedule disk accesses to gain performance
  - FCFS - first come first service
  - SSTF - shorted seek time first
    - starvation
  - SCAN
    - Elevator algorithm
  - CSCAN
    - Restarts from the beginning after each cycle
  - LOOK
    - Look till end of direction
  - CLOOK
- Disk scheduling harder with smart disks that can rearrange bad sectors

Disk attachment

- Host-attached storage
  - SCSI, Fibre-Channel
- Network attached storage (NAS)
  - Device implements a complete file system
- Storage-Area Networks
  - High speed interconnect
  - Can dynamically reassign disks to other servers
- iSCSI
  - SCSI using IP protocols
  - Security, congestion etc. are issues
- Direct Access File System (DAFS)
  - Emerging standard leveraging Remote Direct Memory Access infrastructure
  - http://www.dafscollaborative.org/

RAID

- Reliability vs redundancy
- Performance via parallelism

- Raid 0: striping w/o redundancy
  - No redundancy
  - Good I/O performance
- Raid 1: Mirrored disks
  - Highly redundant
  - Twice read rate, same write performance
- Raid 2: Hamming code ECC
  - Separate disks for data and error correction code
  - Commercially not viable

Raid levels

- Raid 3: bit-interleaved parity organization
  - Data with separate parity disks
- Raid 4: block-interleaved parity
  - Separate parity disk
- Raid 5: Block-interleaved distributed parity
  - Parity data is distributed across all disks
  - Complex implementation on the controller
- Raid 6: Independent Data disks with two independent distributed parity schemes
- Raid 10 (striped array whose segments are RAID 1 arrays), 50, 0+1, 53, …

RAID animations

- http://www.acnc.com/04_01_00_flash.html

Snapmirror

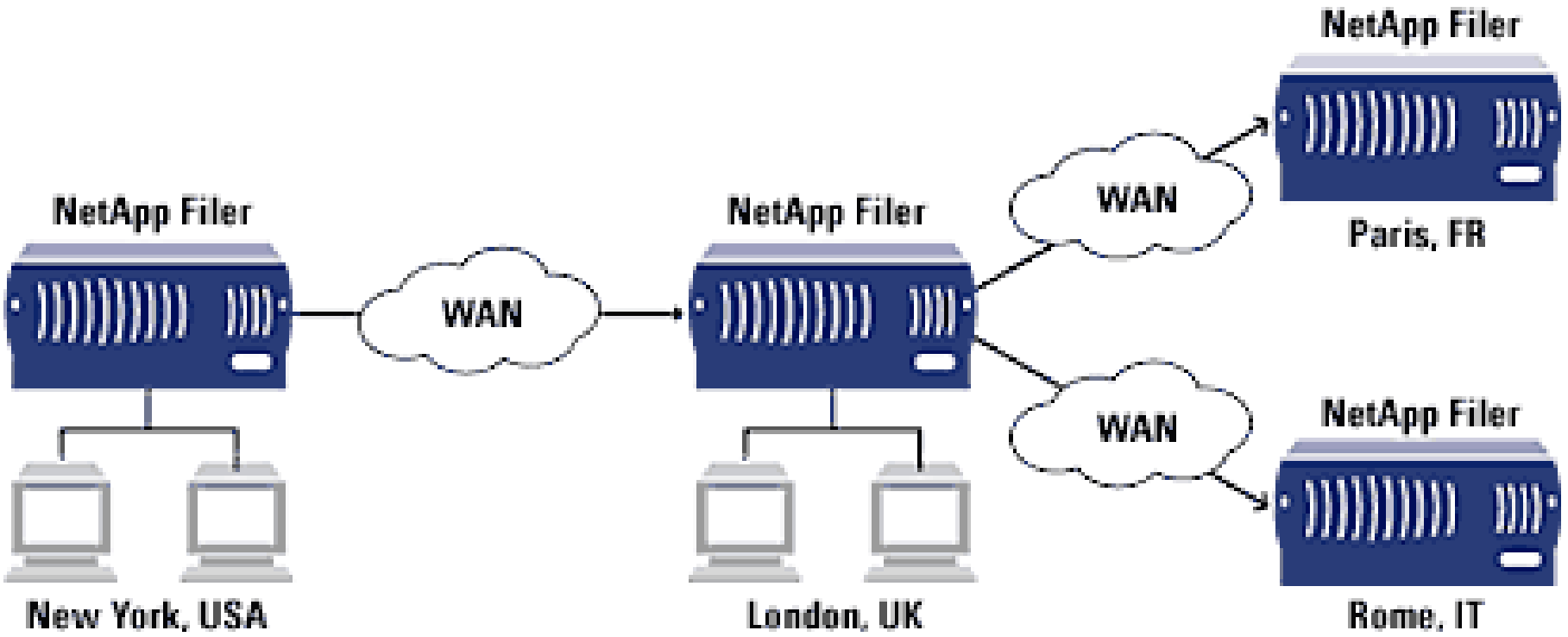
- SnapMirror: File System based asynchronous Mirroring for disaster recovery
- Trends:
  - Persistent and reliable data is crucial for businesses
  - Disks are getting cheaper and bigger, backup technologies are not keeping up
  - RAID to guard against disk failures
    - Hybrid levels (level 50) can provide redundancy and performance
  - Disaster recovery
    - Create off-site online backups to guard against disasters

Challenges

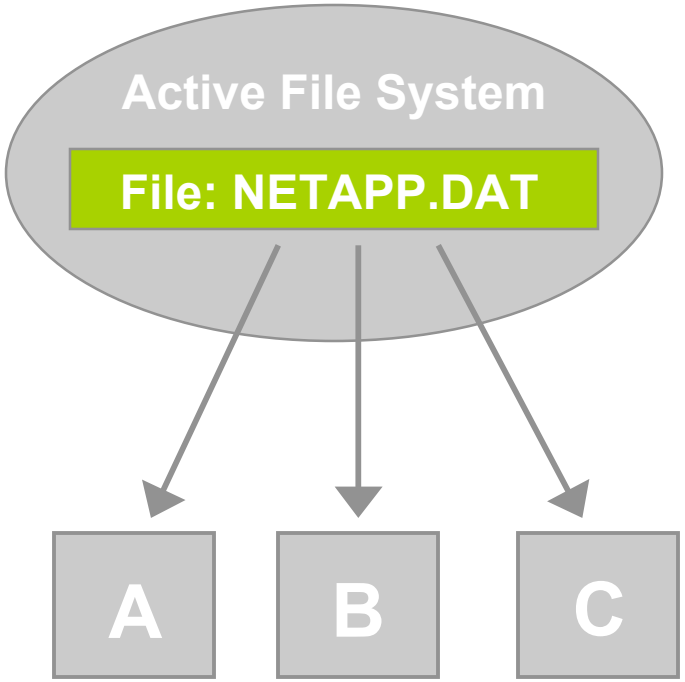
- Backup restore from tapes are cheap but slow
  - Tapes can achieve around 60 GB/hour for restore
  - Terabyte data stores can take a long time to restore
- Online remote backup
  - Expensive (network bandwidth requirements)
  - Performance slow because transaction cannot complete till WAN update finishes
- Asynchronous backups
  - Backup at regular intervals
  - If backup goes to multiple devices, then the event ordering can create inconsistent backups
- We want cheaper, faster restore capable mechanism

Remote mirrors for disaster recovery

- Courtesy: NetApp

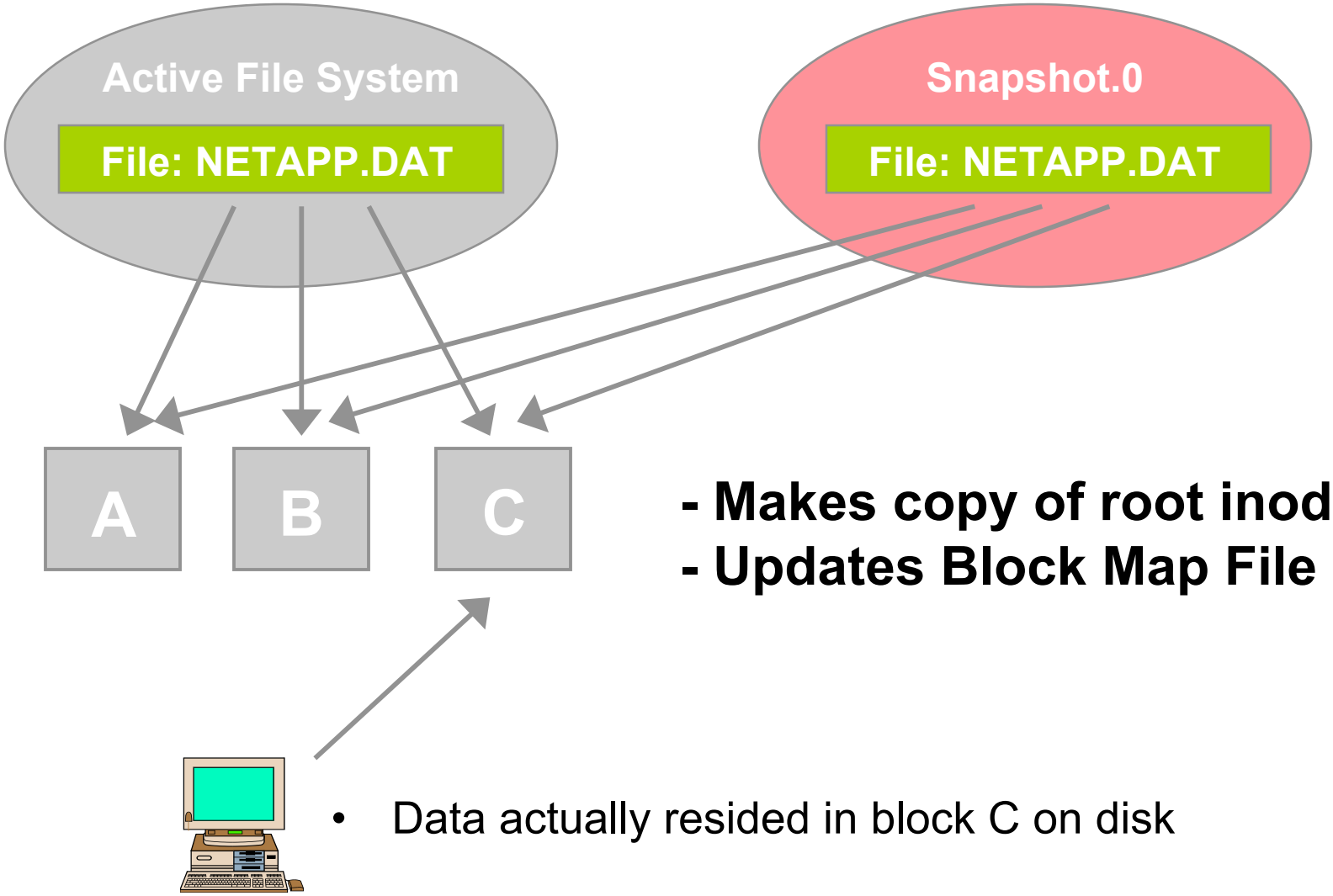


Snapshot Internals (1)



- Data actually resided in block C on disk

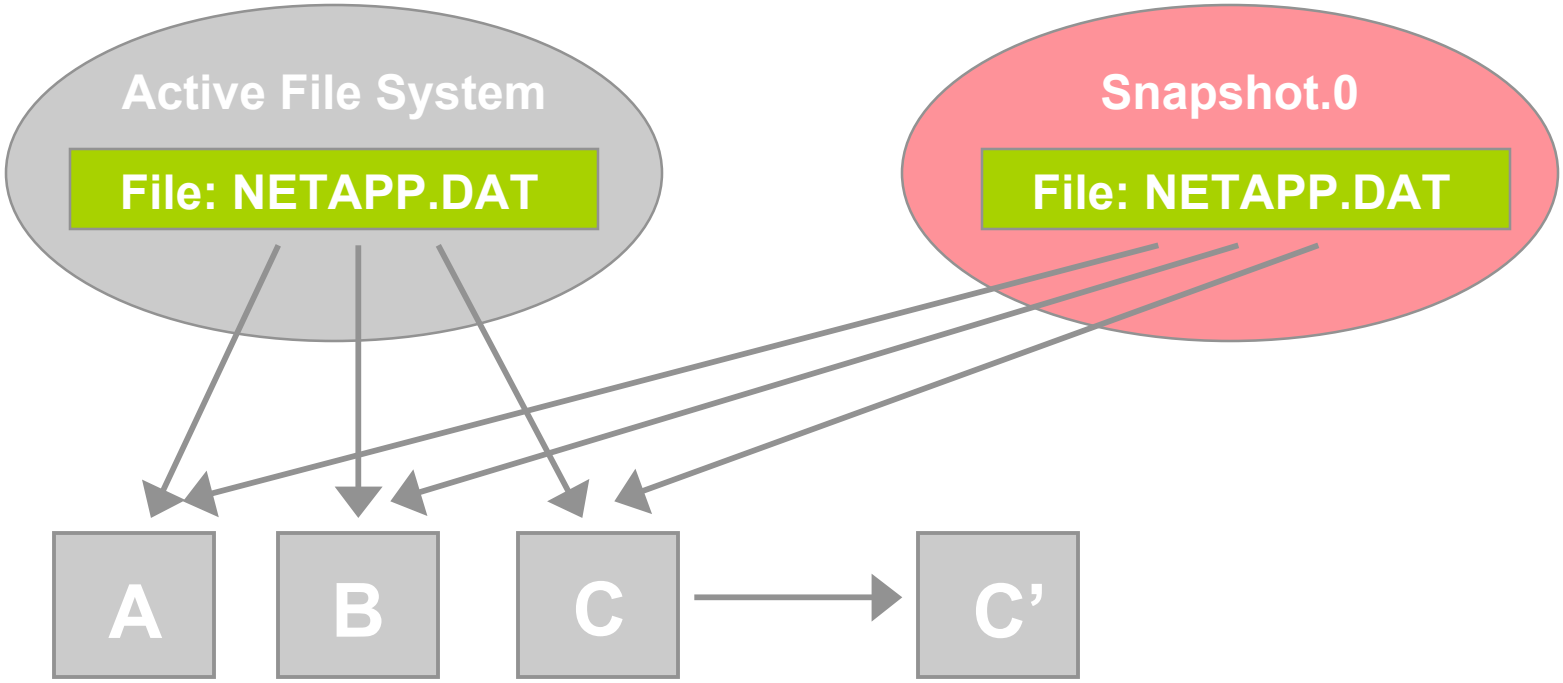
Snapshot Internals (2)



- Makes copy of root inode
- Updates Block Map File

- Data actually resided in block C on disk

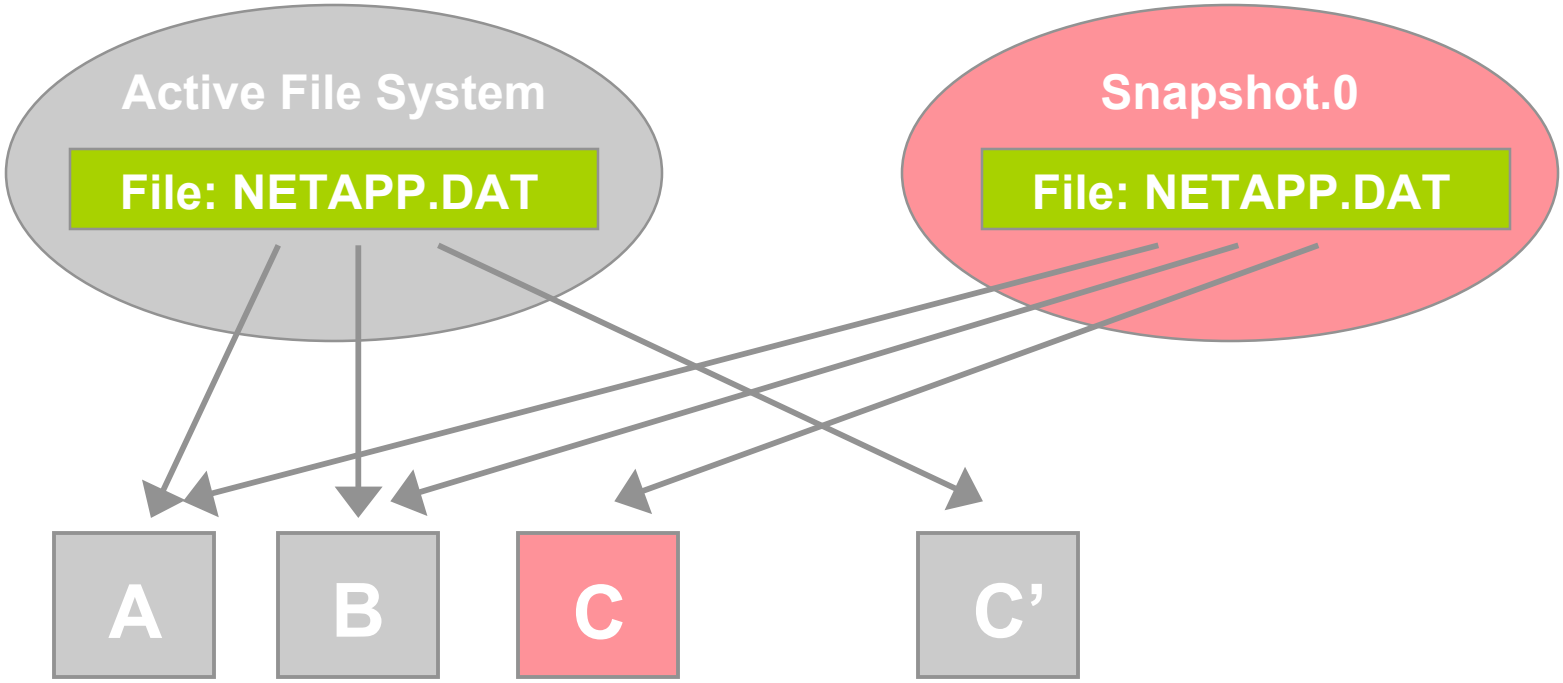
Snapshot Internals (3)



- WAFL (write anywhere file layout) writes modified data block to new location on disk (C')

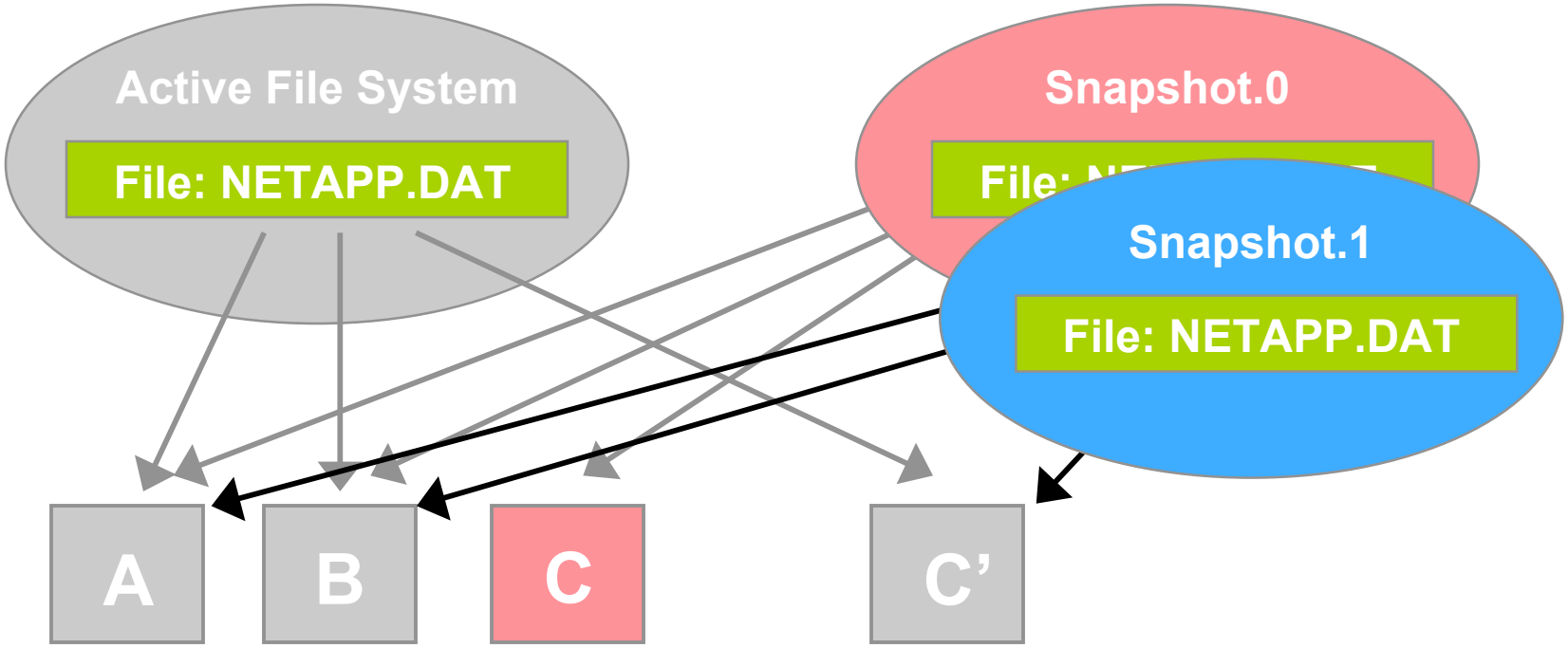
- Client modifies data at end of file
- Data actually resided in block C on disk

Snapshot Internals (4)



- Active file system version of NETAPP.DAT is now composed of disk blocks A, B & C'.
  - Snapshot.0 file system version of NETAPP.DAT is still composed of blocks A, B & C

Snapshot Internals (5)



- Snapshot.1 file system version of NETAPP.DAT is composed of blocks A, B & C'

SnapMirror

- Can use this mechanism to mirror data across WAN
- Can reduce data storage requirements by not backing up deleted/updated data
- Identifying dirty blocks are easier than logical, file system aware mechanisms