Life of a Storage Packet (Walk)

J Metz, SNIA Board of Directors, Cisco
Chad Hintz, SNIA-ESF Board, Cisco
November 19, 2015

SNIA™
Ethernet Storage Forum

Who We Are

J Metz SNIA
Board of Directors
Cisco

Chad Hintz
SNIA-ESF Board
Cisco

SNIA Legal Notice

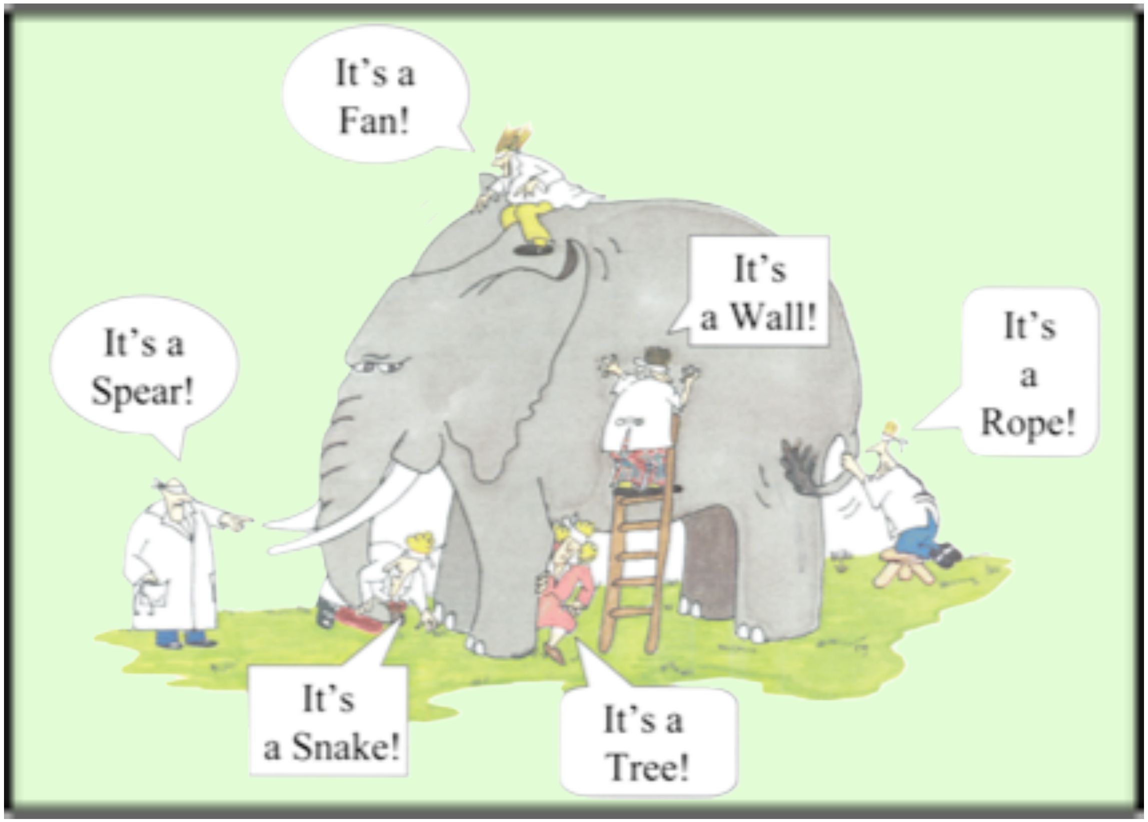
- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
- Any slide or slides used must be reproduced in their entirety without modification
- The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Why a Packet Walk?

- It's not a question of smarts, it's a question of scope
- A lot of focus on the details, but not enough on the relationships between details

What Is It?

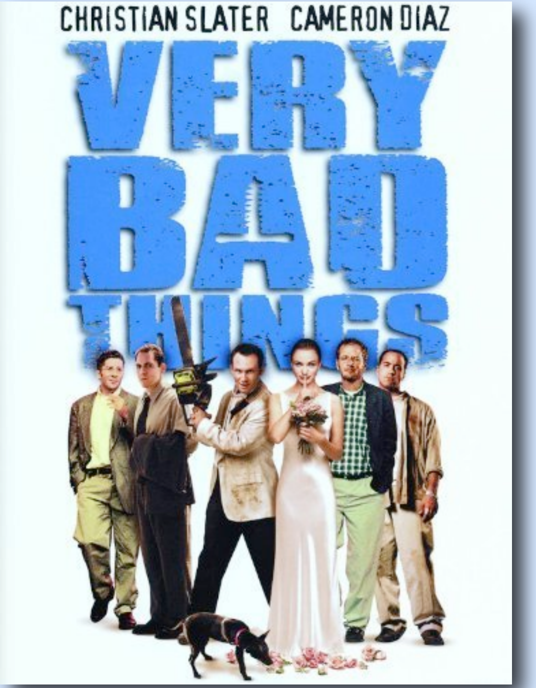


- Putting many little pieces together may not give you the right big picture
- Who is this for?
  - People who want an introduction to storage systems (i.e., beginners)
  - Experts in one field, but not all
  - People who want to know more about the basics (but were afraid to ask)

Avoiding Bad Things

SNIA Ethernet Storage Forum

- Ignorance in storage can be a Very Bad Thing
- Can cause "religious" differences based upon what people are comfortable with, rather than technological merits
- Can lead to incompatible solutions and unintended consequences



Agenda

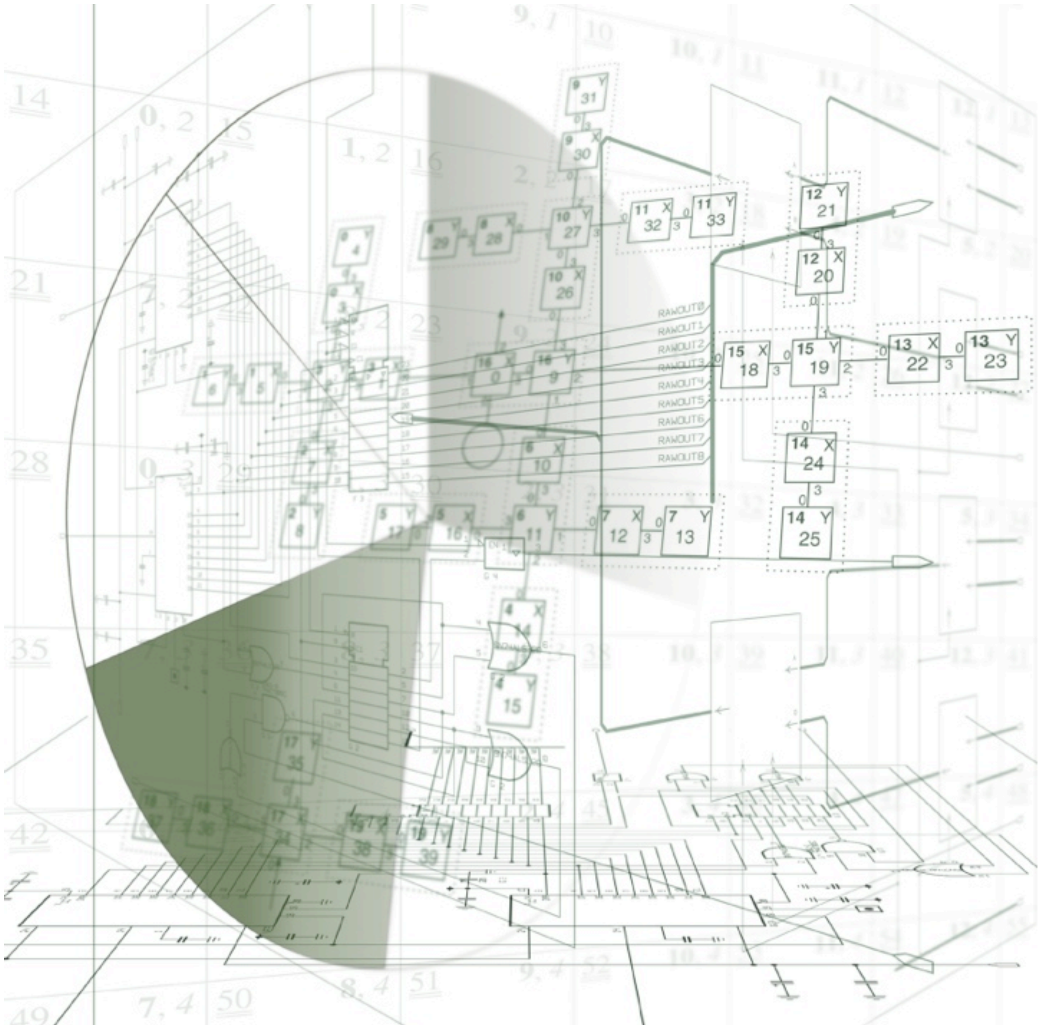
- Introduction/What this Presentation Is
- Understanding the Parts (pieces of the puzzle)
- Understanding What the Application Sees
- Understanding What the Storage Sees
- Understanding What the Network Sees
- Putting It All Together
- Additional Resources
- Conclusion

What This Presentation Is

- Focus on the holistic storage problem
- Emphasis on the relationships between storage elements
- Visualizing the concepts in a different way
- Keeping it Simple and Sane
- Keeping a level head and a balanced view

IT'S ALL ABOUT PERSPECTIVE.

What This Presentation Is NOT

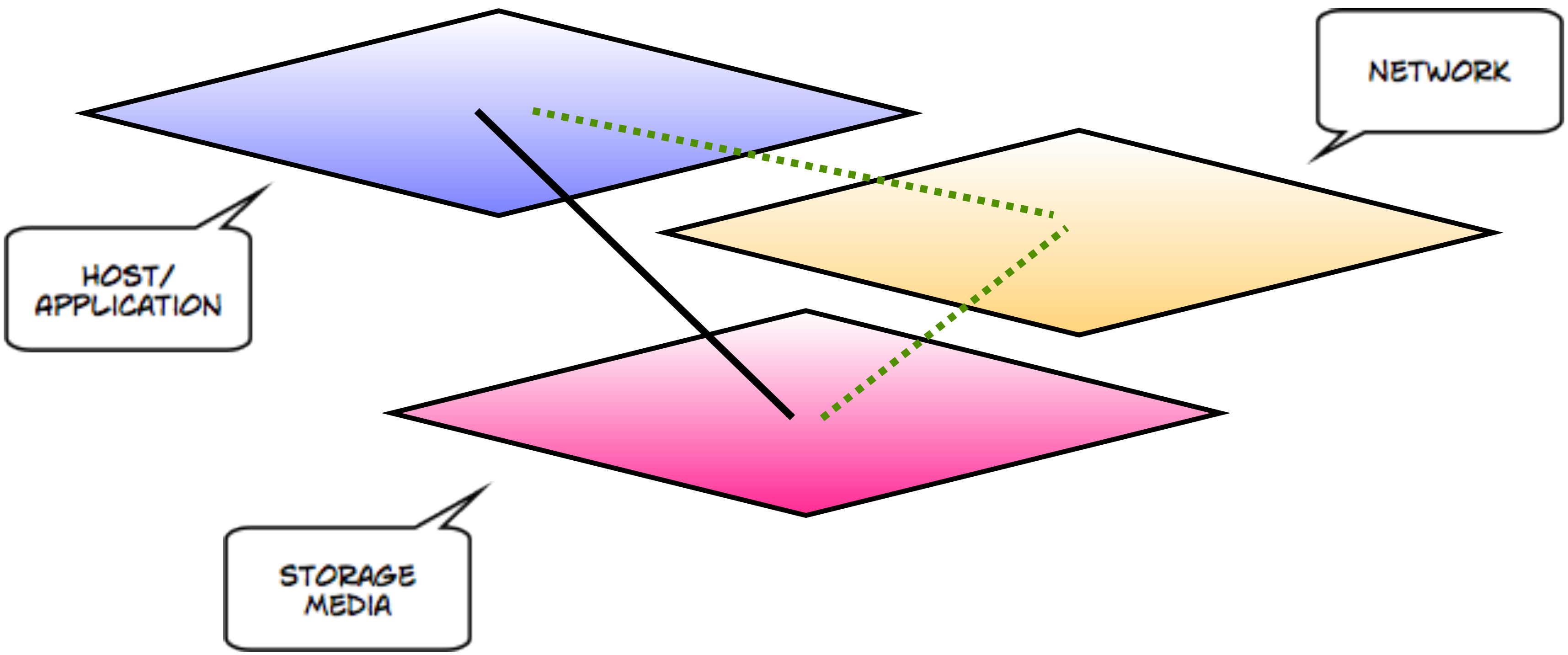


- Virtualized Storage
- Software Defined Storage
- BC/DR/Storage-over-Distance
- Security
- Comprehensive
  - Many of the nuances will be missing!

The Common Parts

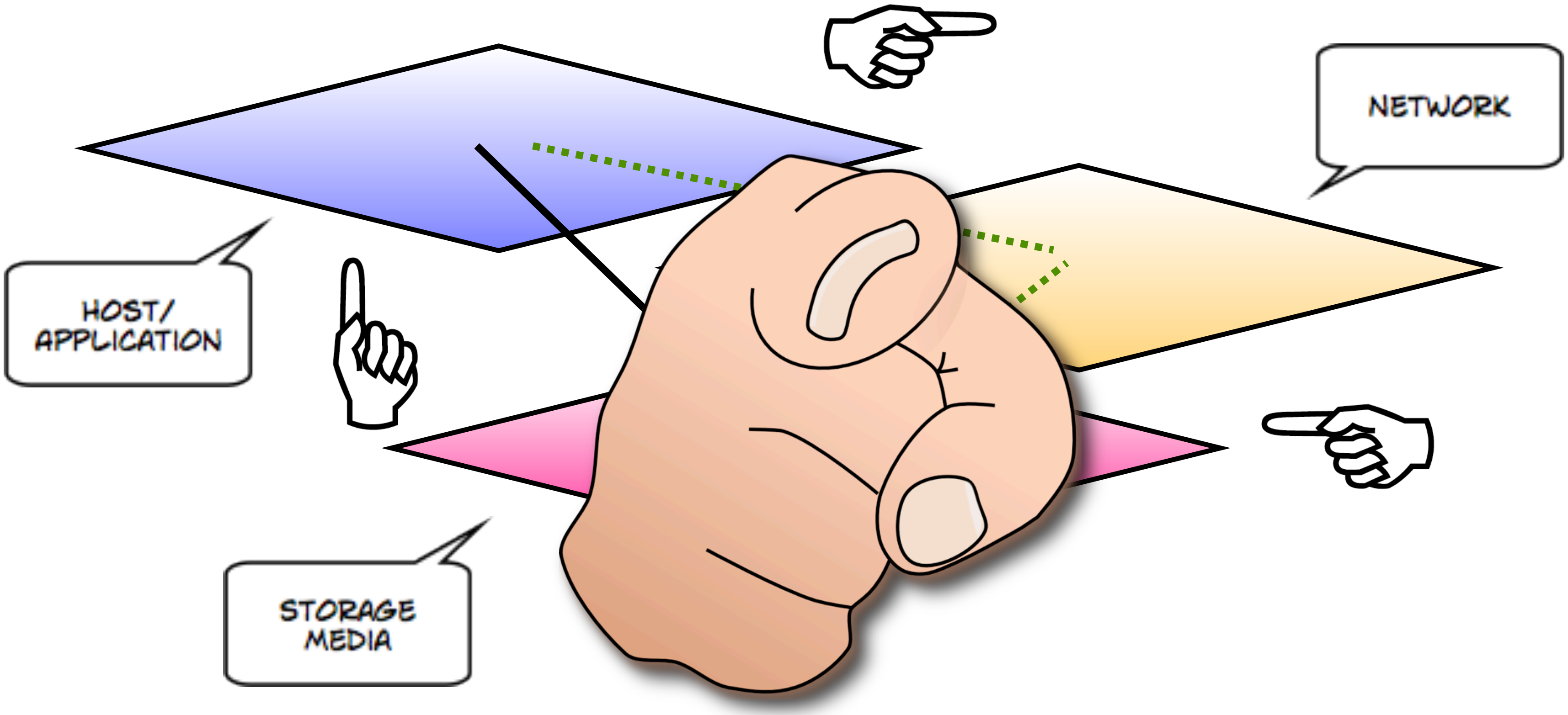
SNIA™
Ethernet Storage Forum

Bigger Picture



- Three main conceptual areas: Host/Application, Storage Media, and Storage Network

Bigger Picture



- Three main conceptual areas: Application, Storage Media, and Storage Network

Storage Has One Job

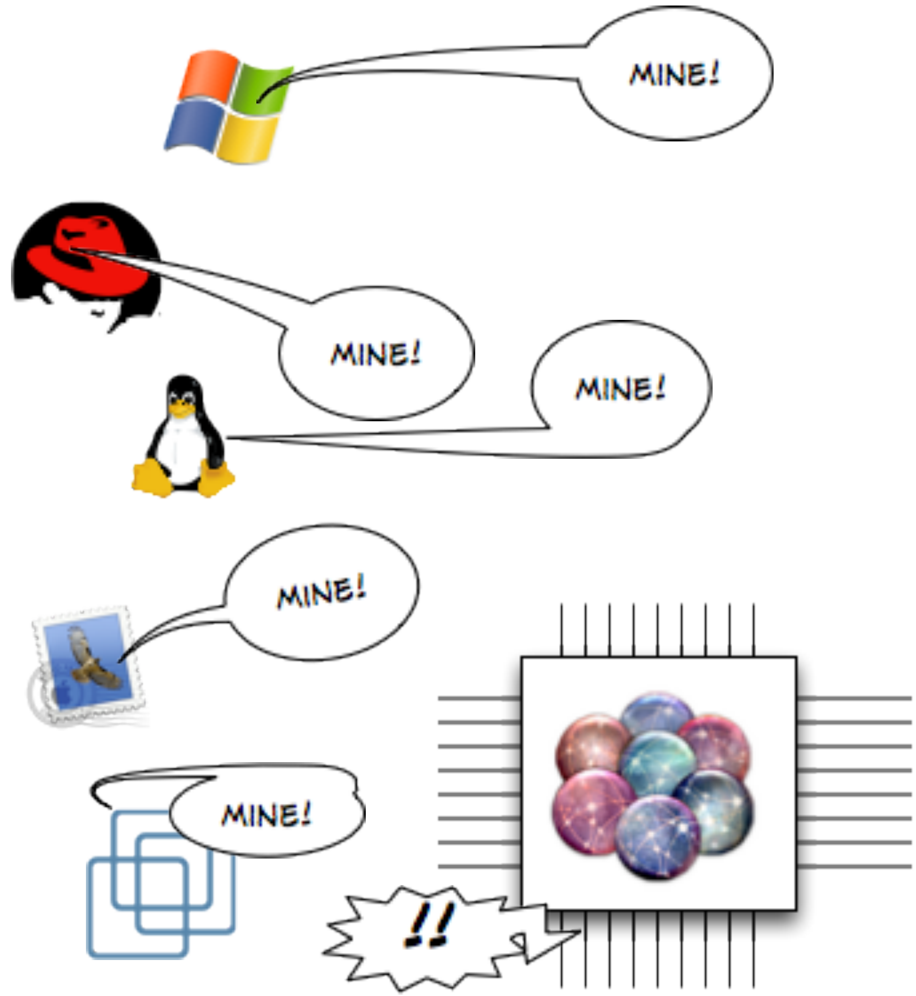
- "Here is a bit of data. Hold onto it. Give that same bit back to me when I ask for it."

Understanding What the Application Sees

SNIA
Ethernet Storage Forum

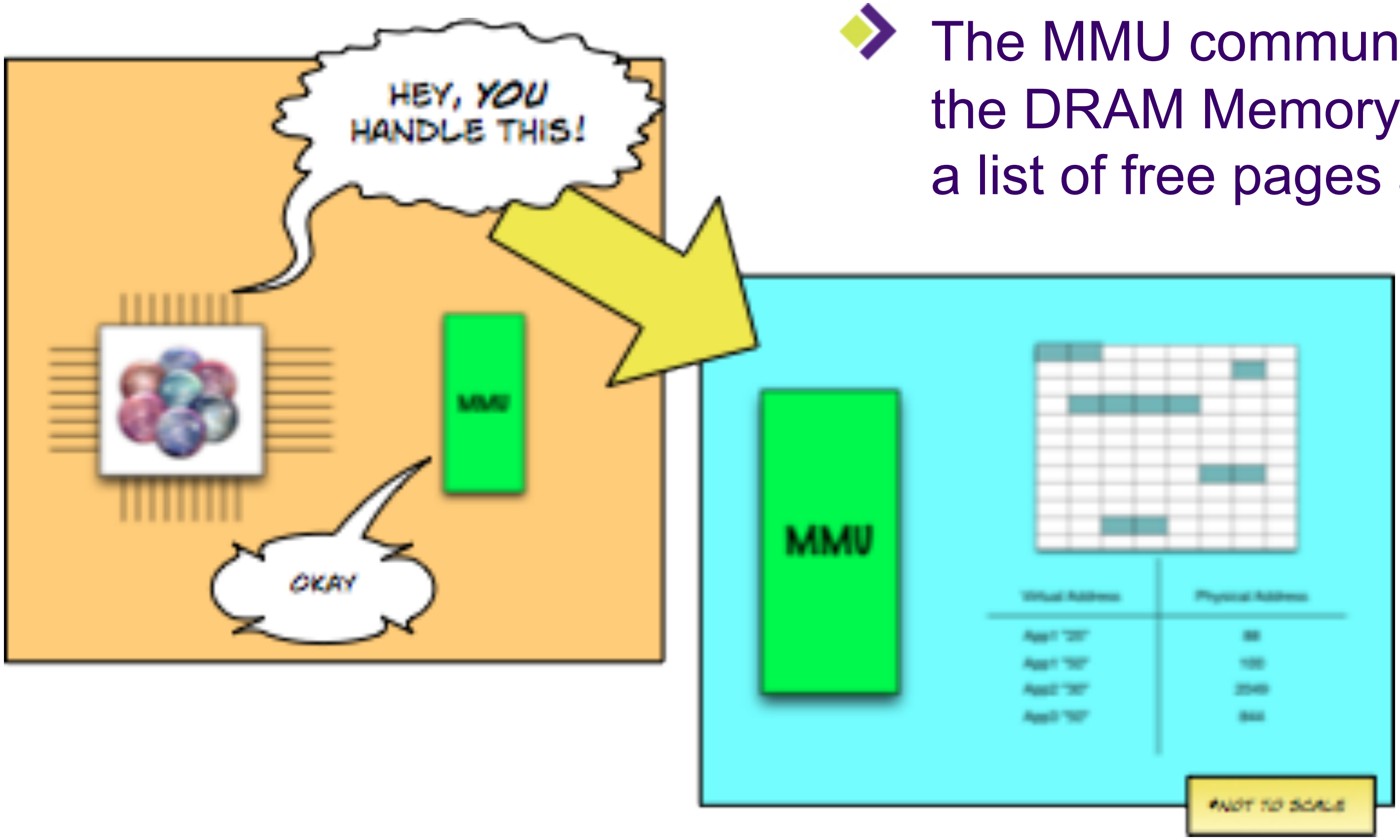
Applications

- Each system has a CPU with many applications running
- These applications think they have all the available memory

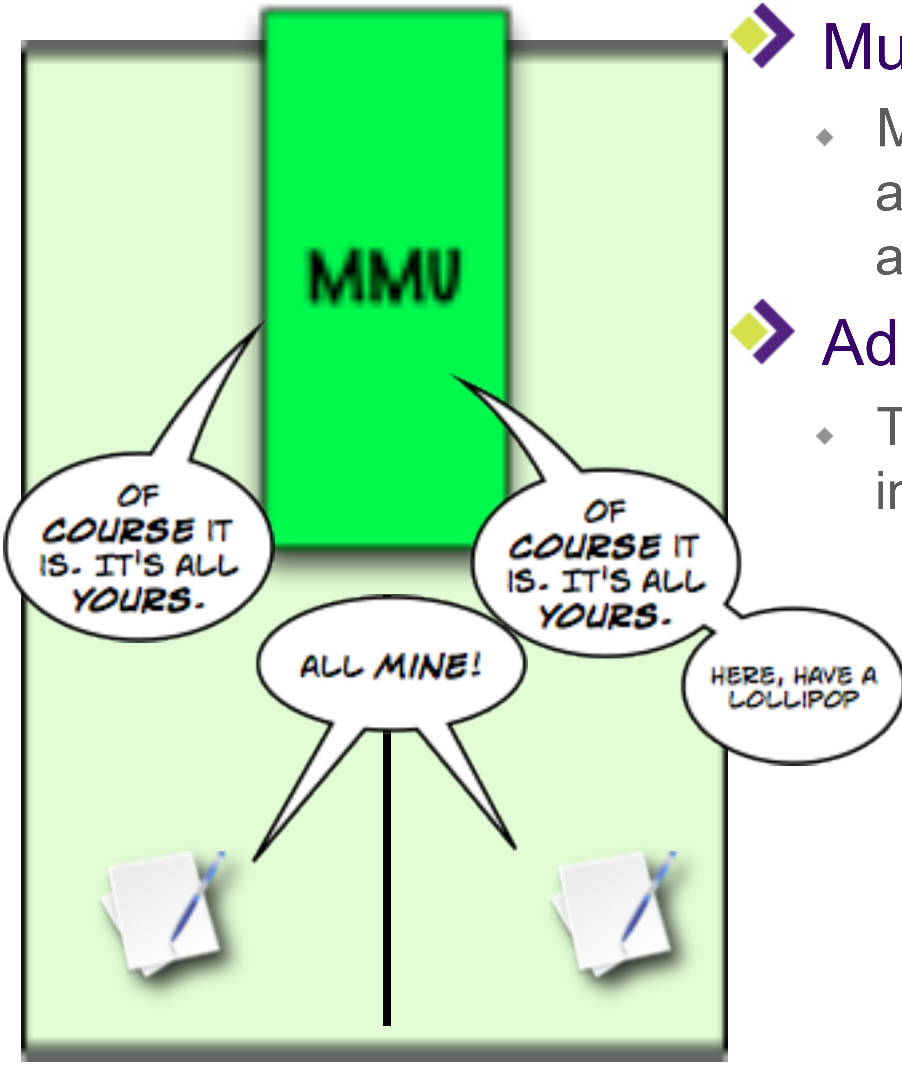


CPU -> MMU

- Compute systems have a Memory Management Unit (MMU)
- The MMU communicates with the DRAM Memory directly, gets a list of free pages available



Memory Management Module



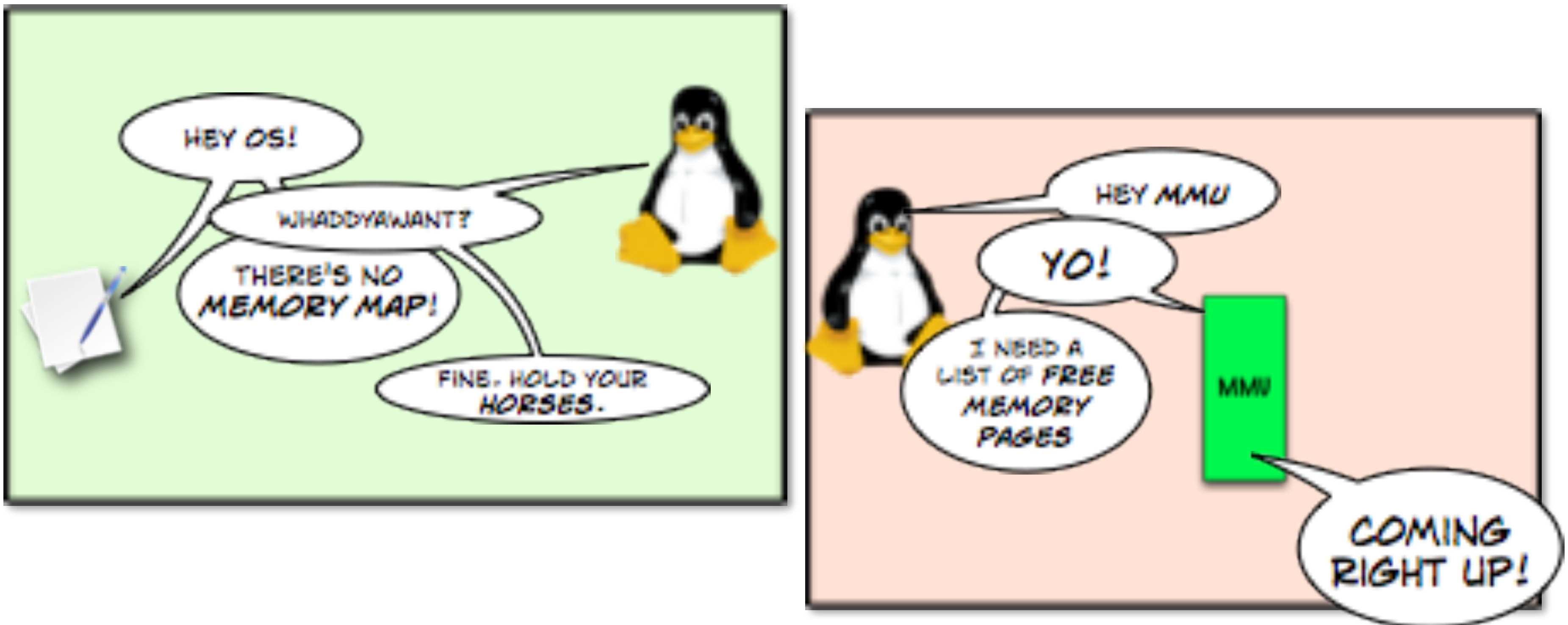
Multi-tenancy

- Memory owned by one process (e.g., application) can't be overwritten by another process

Addressing

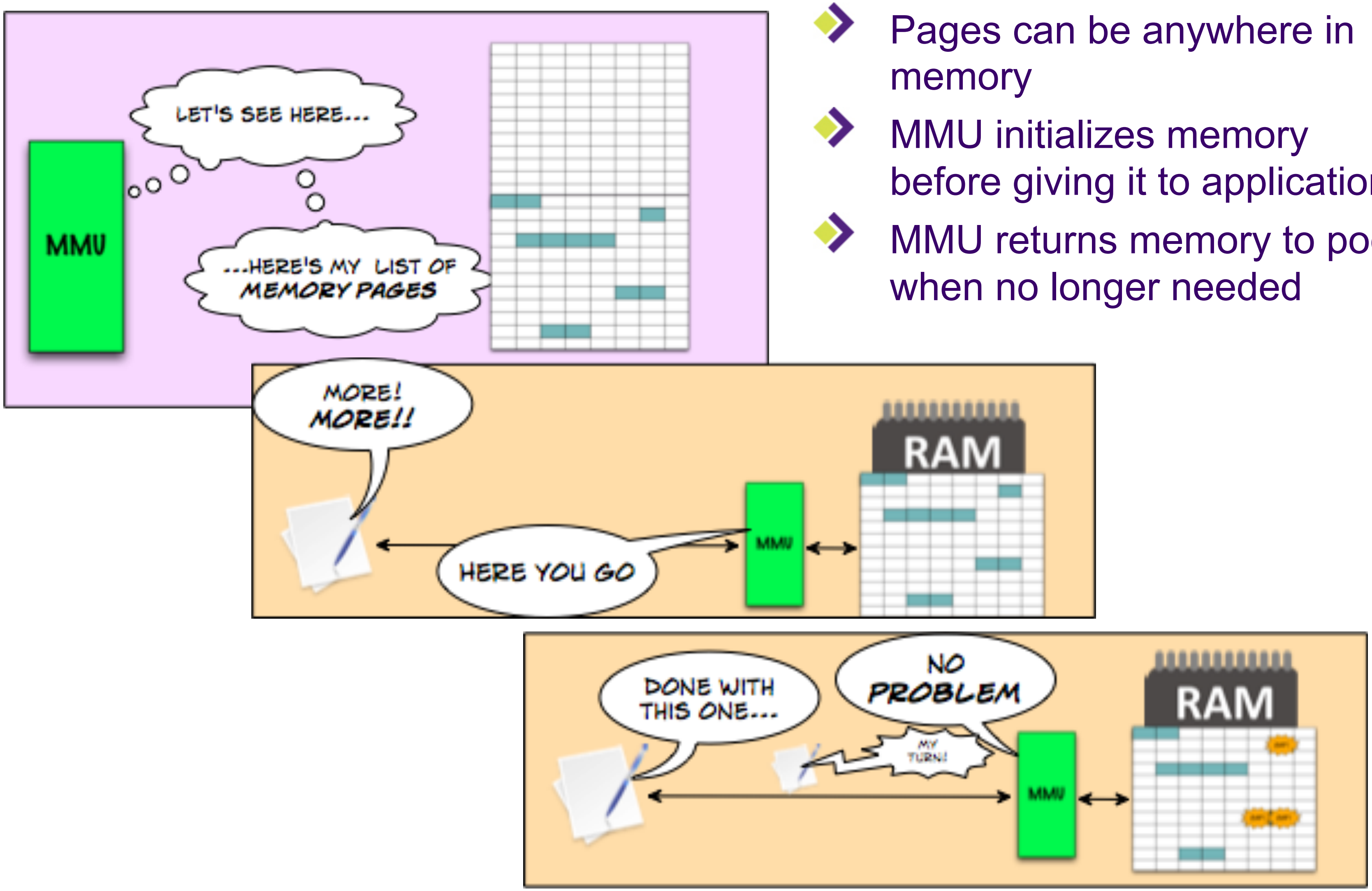
- Translating CPU's physical address into particular DRAM or row of DRAMs

Accessing Memory



- Applications get memory when they try to access it

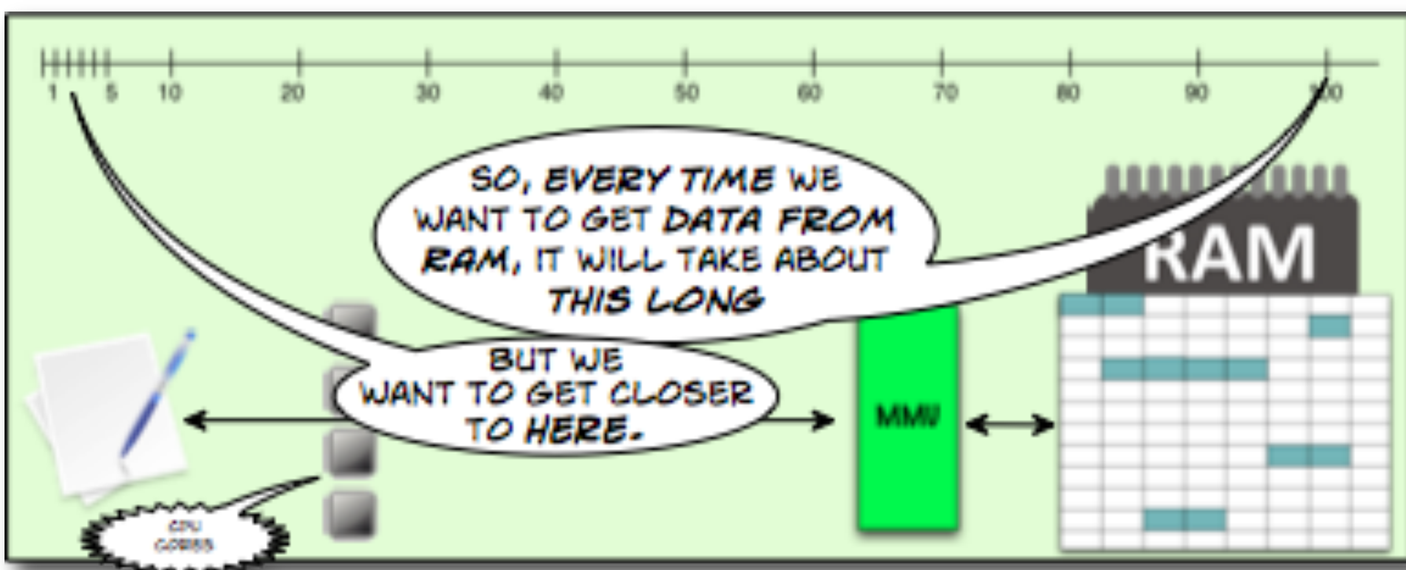
Assigning Memory



- Pages can be anywhere in memory
- MMU initializes memory before giving it to application
- MMU returns memory to pool when no longer needed

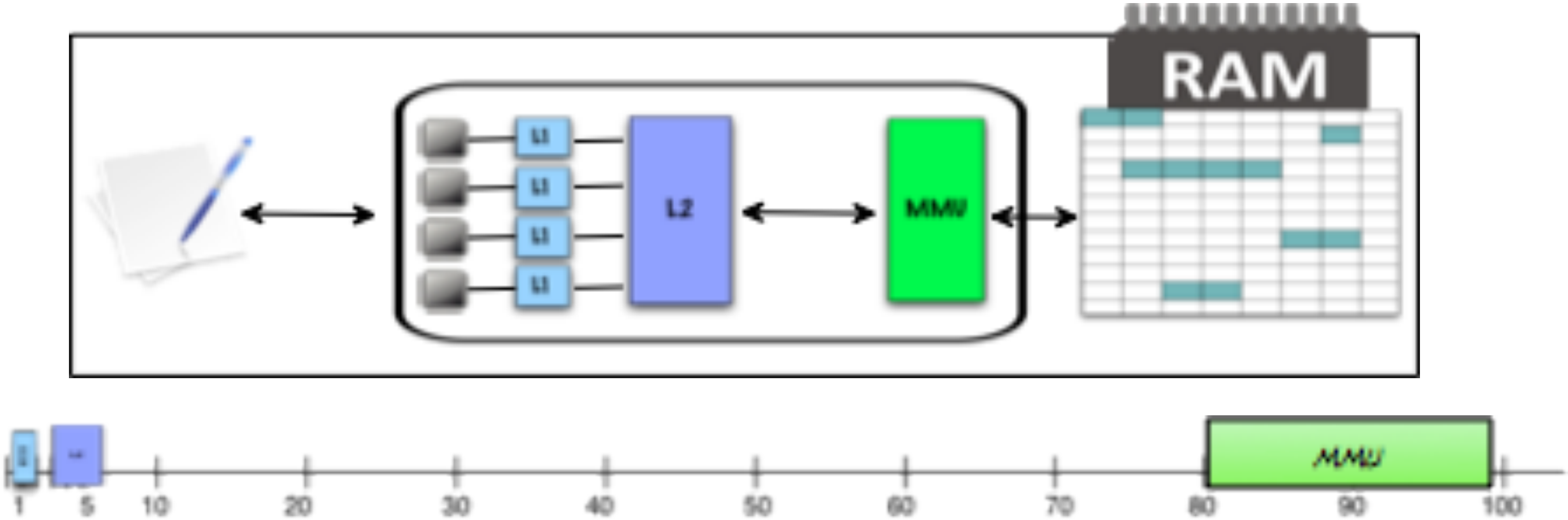
Making It Faster

- Rule of Thumb: Always put storage/ memory as close to the CPU as possible
- Improving time constraints will be a constant theme in storage
- Accessing DRAM takes anywhere from 60-100ns
- Need to get closer to “zero”



Magic of Caching

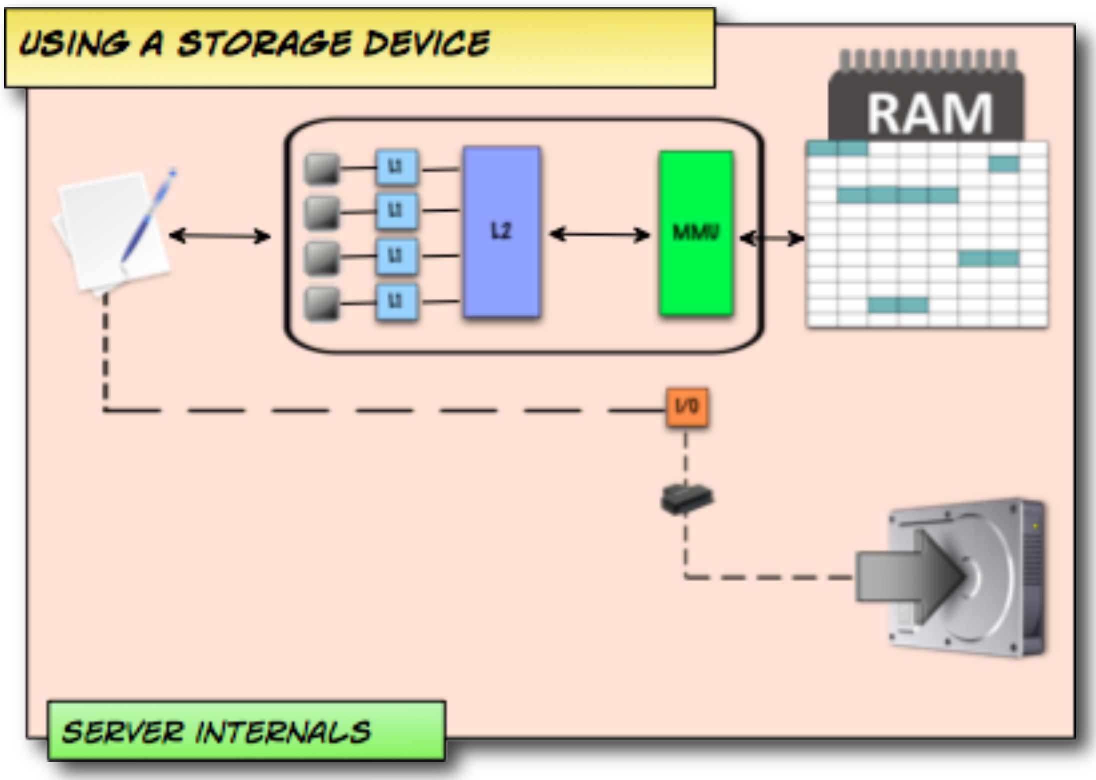
- A Level 1 (L1) cache directly connects to a CPU core, taking ~1ns
- Level 2 (L2) cache takes about 3-6 ns



No More Room = Accessing Disk

- When you have no more room at the DRAM, you need to go to more permanent storage (e.g.,) disk
- Going to disk is expensive (time-wise)
- Do you want to drive:
  - 1km for your pizza?
  - 5 km for your pizza?
  - 100 km for your pizza?
  - 8 million km for your pizza?
    - (i.e., more than 10 round trips to the moon!)

Fork in the Road

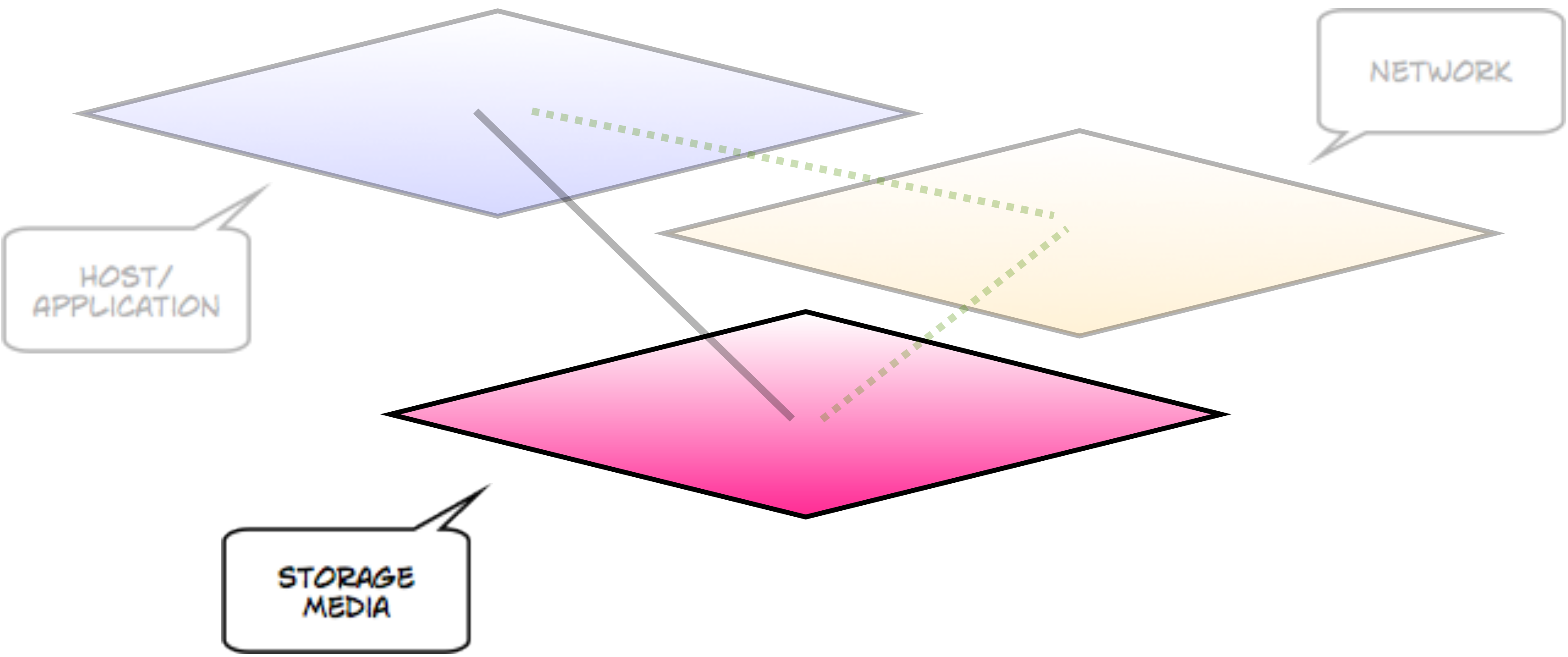


- Best Practice:
  - Always keep storage as close to the application as possible
- The storage drive is a different part of the process, and has some additional pieces
- Unlike RAM, storage devices don't "speak CPU" natively
  - Need some additional parts to get them to talk

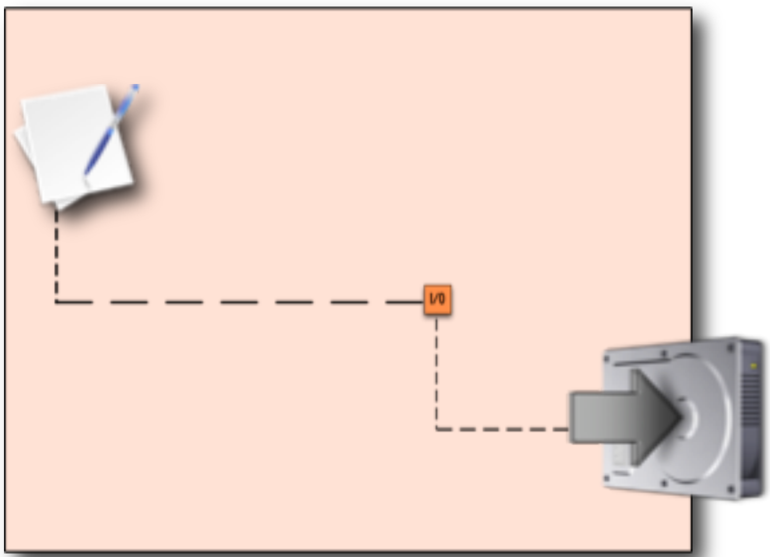
Understanding What the Storage Sees

SNIA
Ethernet Storage Forum

Storage Media

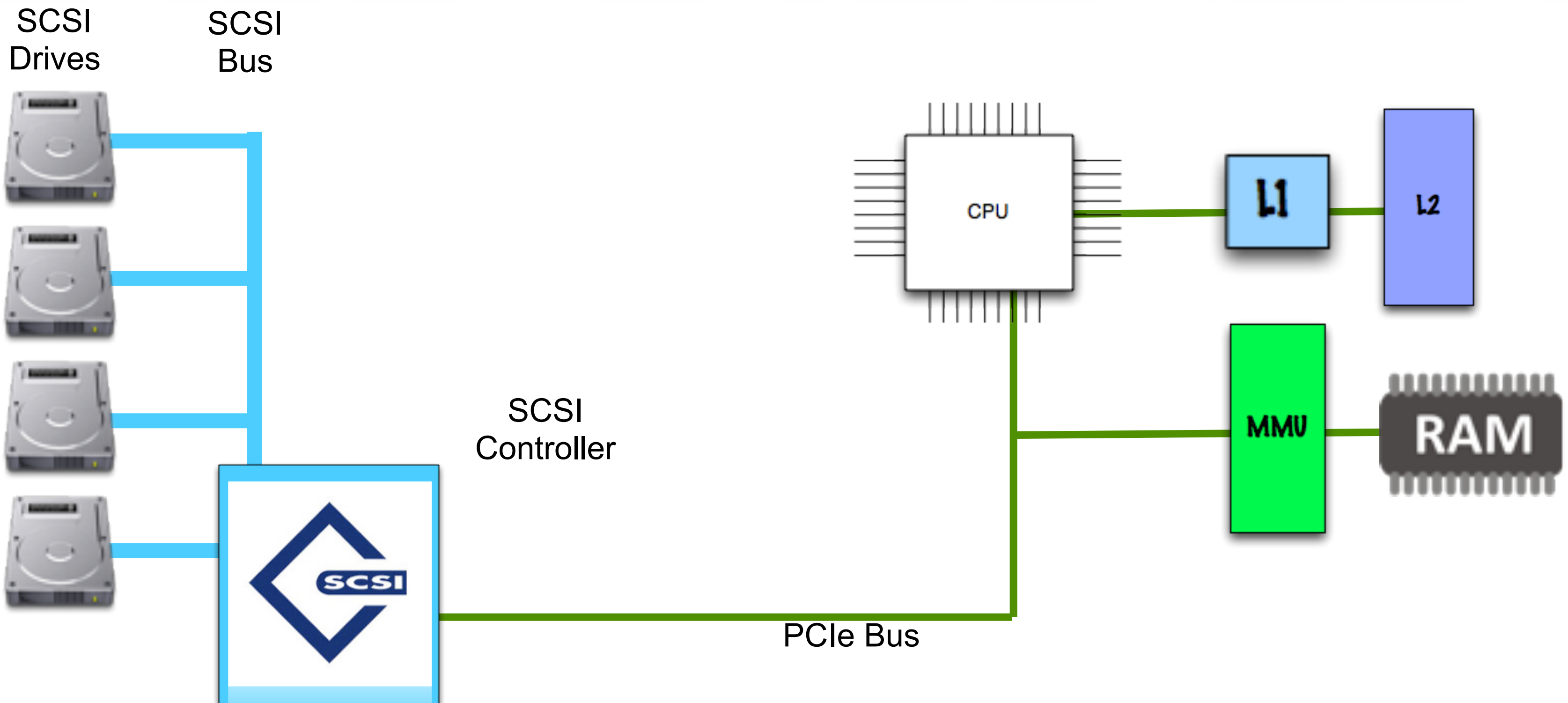


Block I/O Stream



- CPU and memory are connected to storage via the PCI bus (usually PCIe nowadays)
- Applications ask for a location and a length (range) of data storage
- The system needs to translate both the address and the location of where that range of data exists
- The system needs help with that translation in the form of a protocol

Go for a Bus Ride



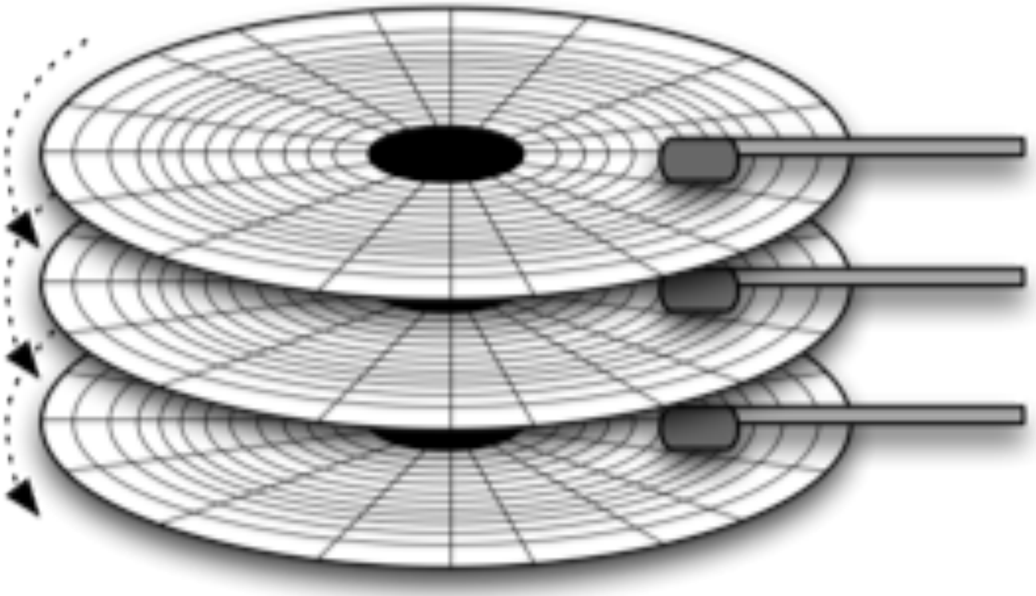
- Commands from the CPU need to be adapted/translated to speak to storage devices (e.g., SCSI, IDE/ATA)

A Little Bit About SCSI

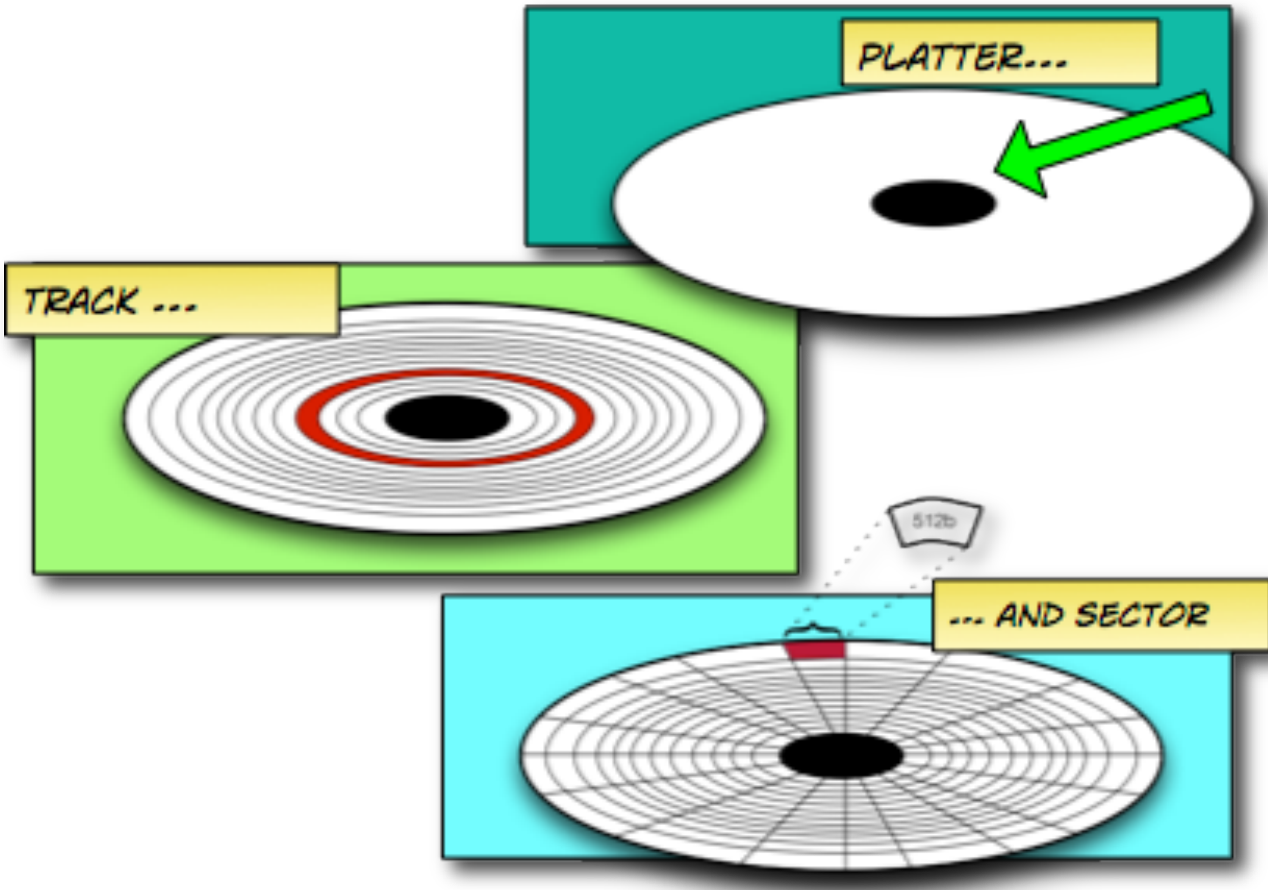
- SCSI is ubiquitous
  - Hard Drives, SSDs, Tape Drives, etc.
- Backwards Compatible
- Common SCSI Components
  - Initiator - Issues requests for service by SCSI devices, can be on-board or part of an adapter
  - Target - physical storage device, can be single disk or array
  - Service Delivery Subsystem - Communication mechanisms (usually over a wire) between initiators and targets

More About Blocks

- Blocks are logical and physical units that are located on storage media
- It is the smallest unit writable by a disk or file system
- All storage - including file storage and object storage, eventually winds up talking to blocks



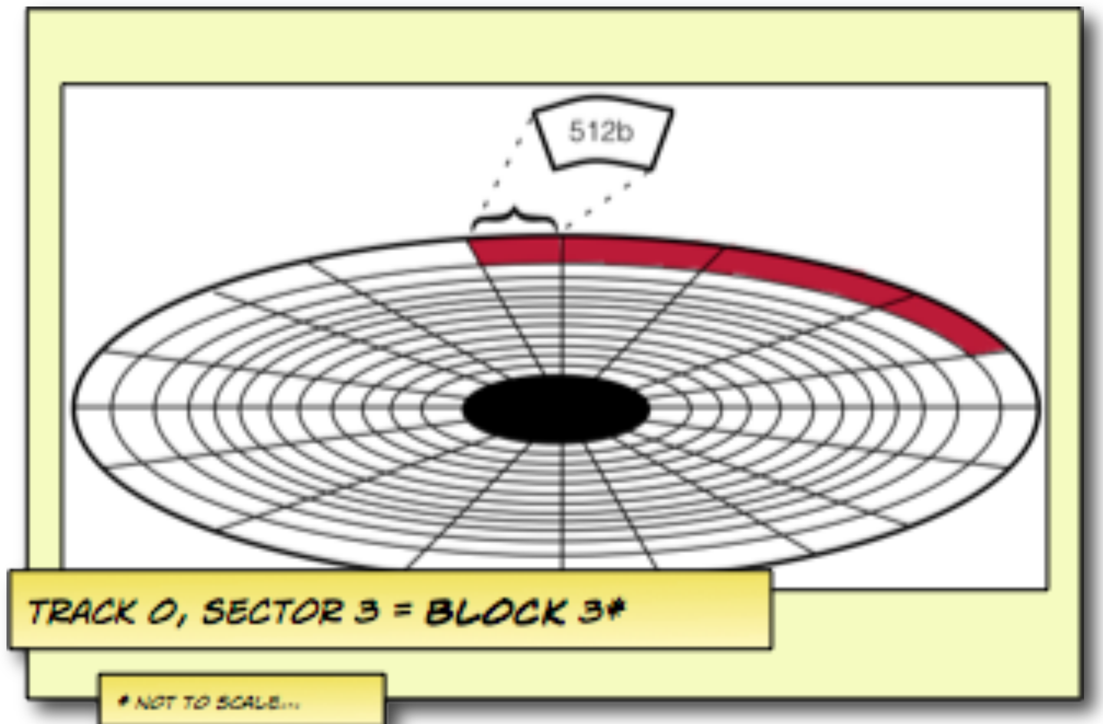
Anatomy of a Disk Drive



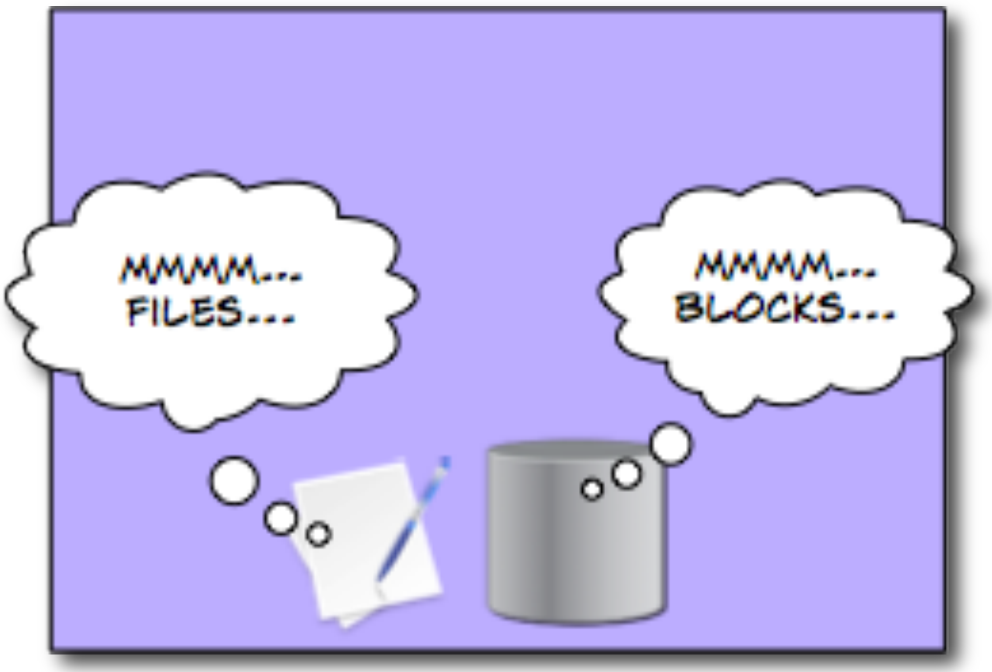
- Components:
  - Platter
  - Track
  - Sector
- Location of data greatly affects performance

Blocks and Sectors

- Blocks are made up of sectors on a drive
- Each block is given a unique number
- Everything that a file system does with storage media is composed of operations on blocks



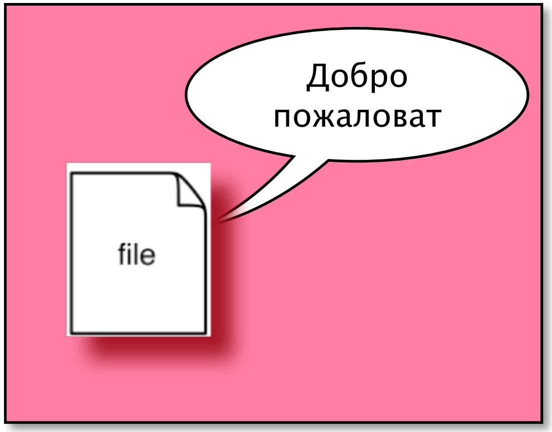
Files and Blocks



- Applications (including OS) think in terms of files
- Storage thinks in terms of blocks
- Need to match these up somehow

Understanding Files and Blocks

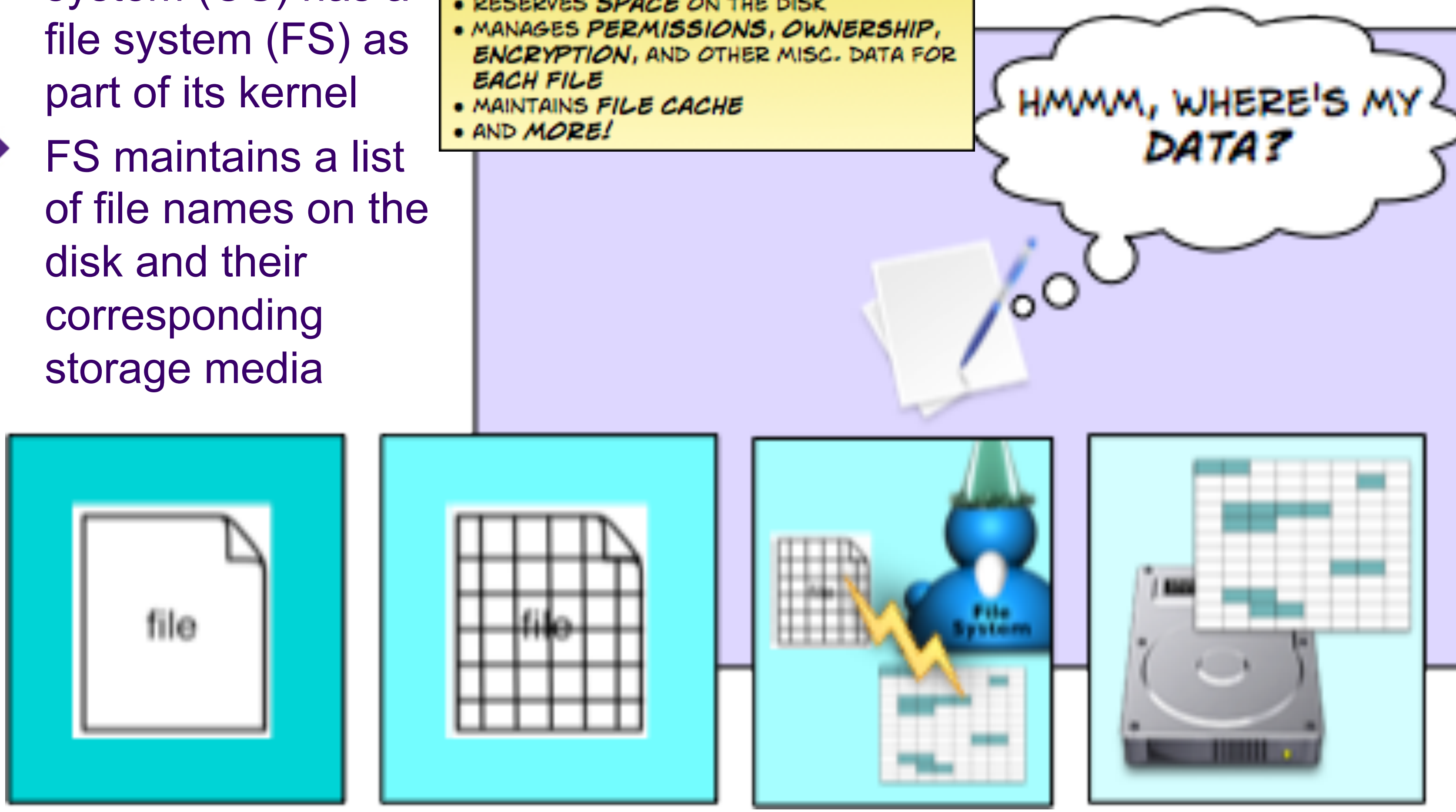
SNIA
Ethernet Storage Forum



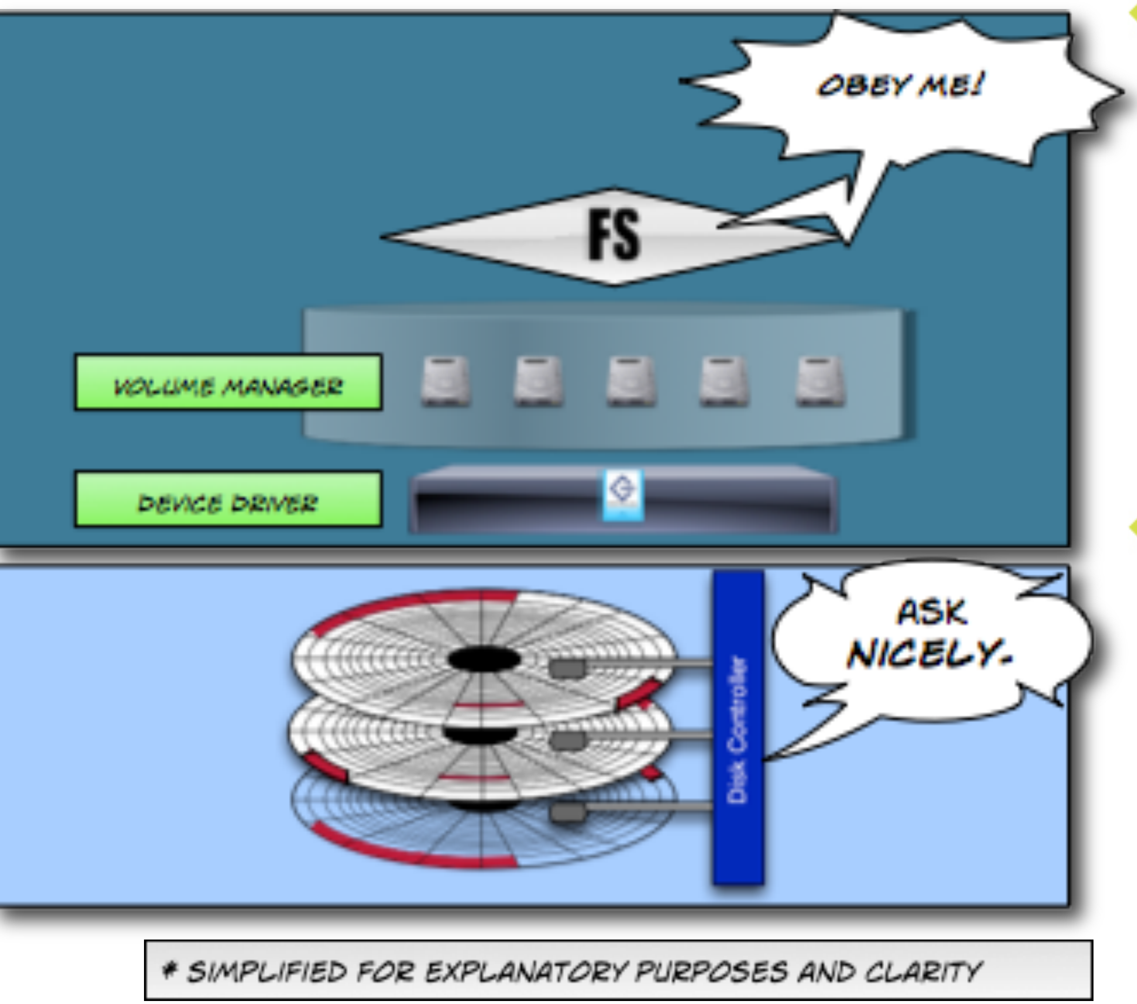
File Systems

- Every operating system (OS) has a file system (FS) as part of its kernel
- FS maintains a list of file names on the disk and their corresponding storage media

- Manages the **relationship** of the file (and the corresponding **metadata**) to their **storage locations**
- Reserves **space** on the disk
- Manages **permissions, ownership, encryption**, and other misc. data for **each file**
- Maintains **file cache**
- And **more!**



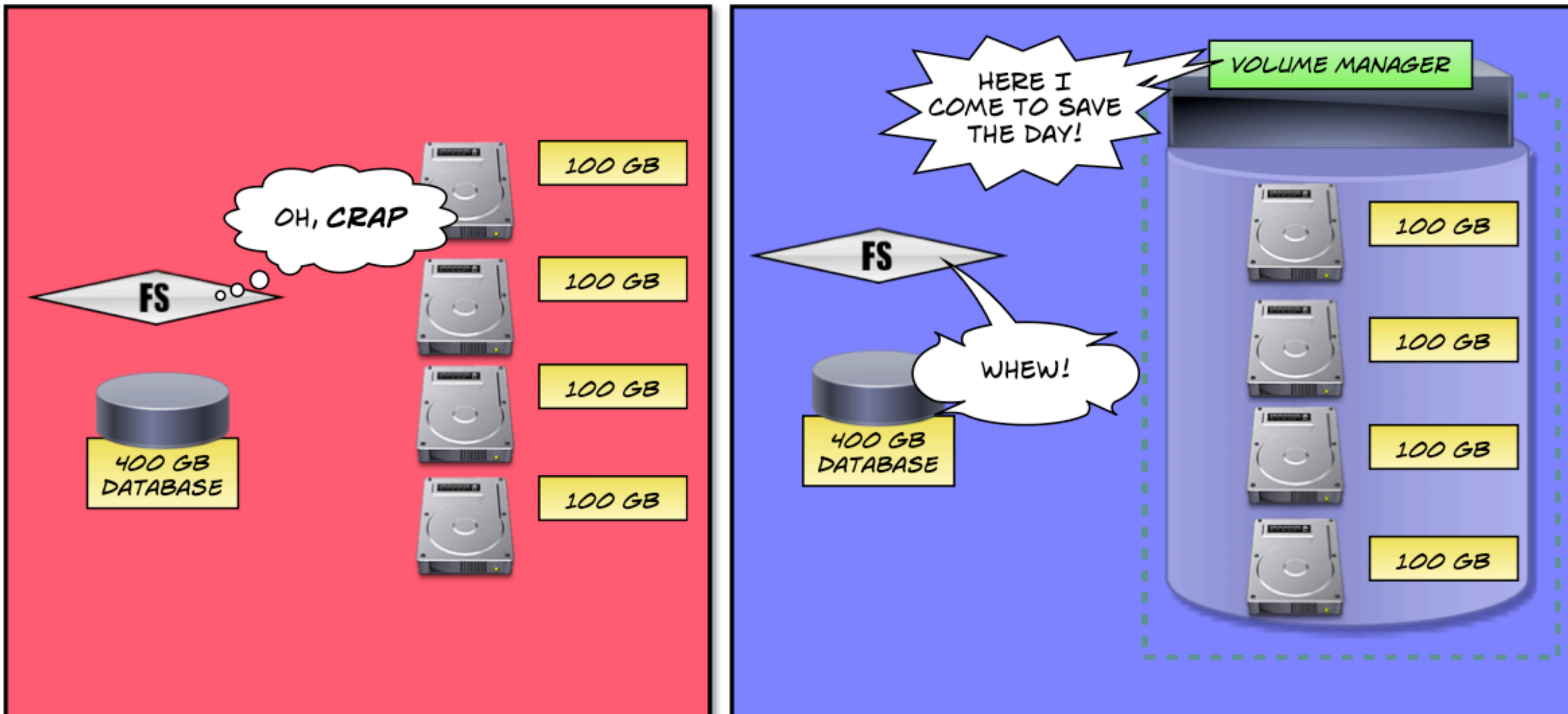
Enter the File System



- Drives are managed by a drive controller
  - Takes I/O commands from the file system
  - Done through I/O module using a protocol (such as SCSI)
- In-between the file system and the drive controller is a Volume Manager*
  - Aggregates and creates “fake disks” that the File System uses

Volume Manager

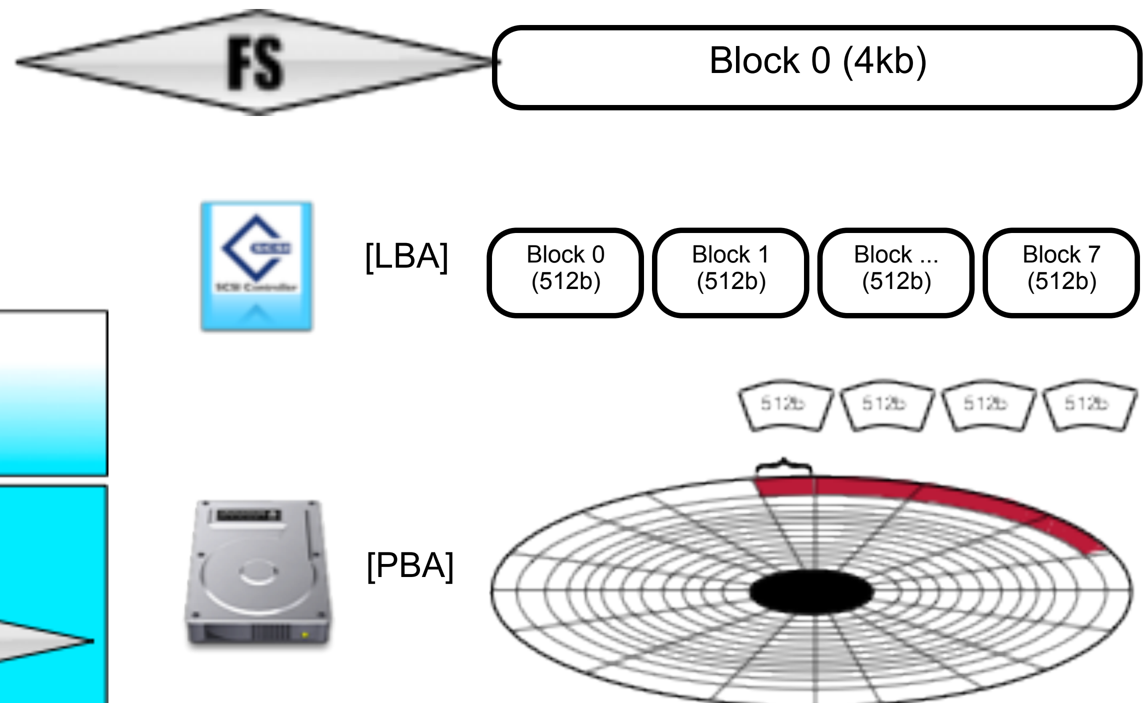
- Creates a virtual volume layer

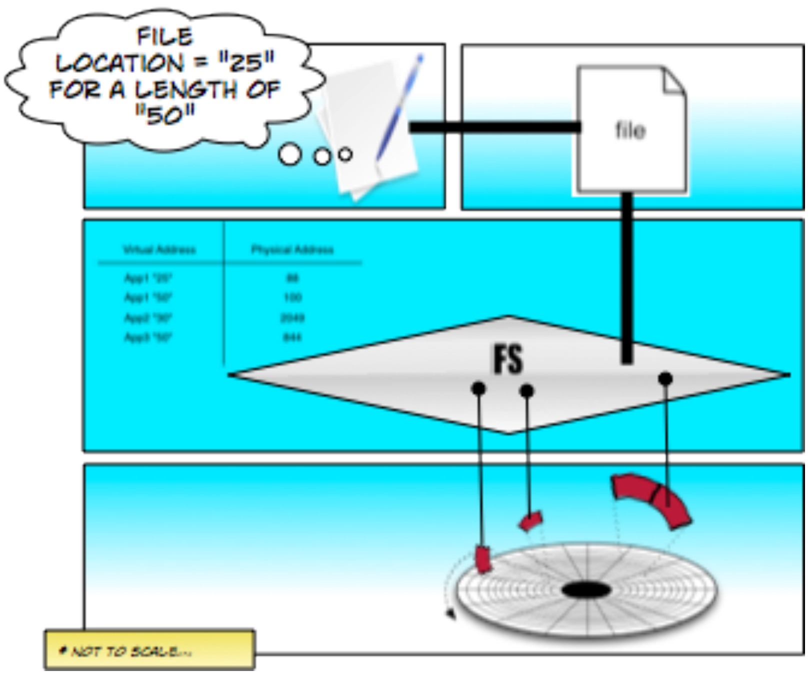


* SIMPLIFIED FOR EXPLANATORY PURPOSES AND CLARITY

File System and Drives

- Translates and manages the virtual addresses that the application sees to the block address on a drive
- Not all blocks are created equal





Inodes, Files, and Directories

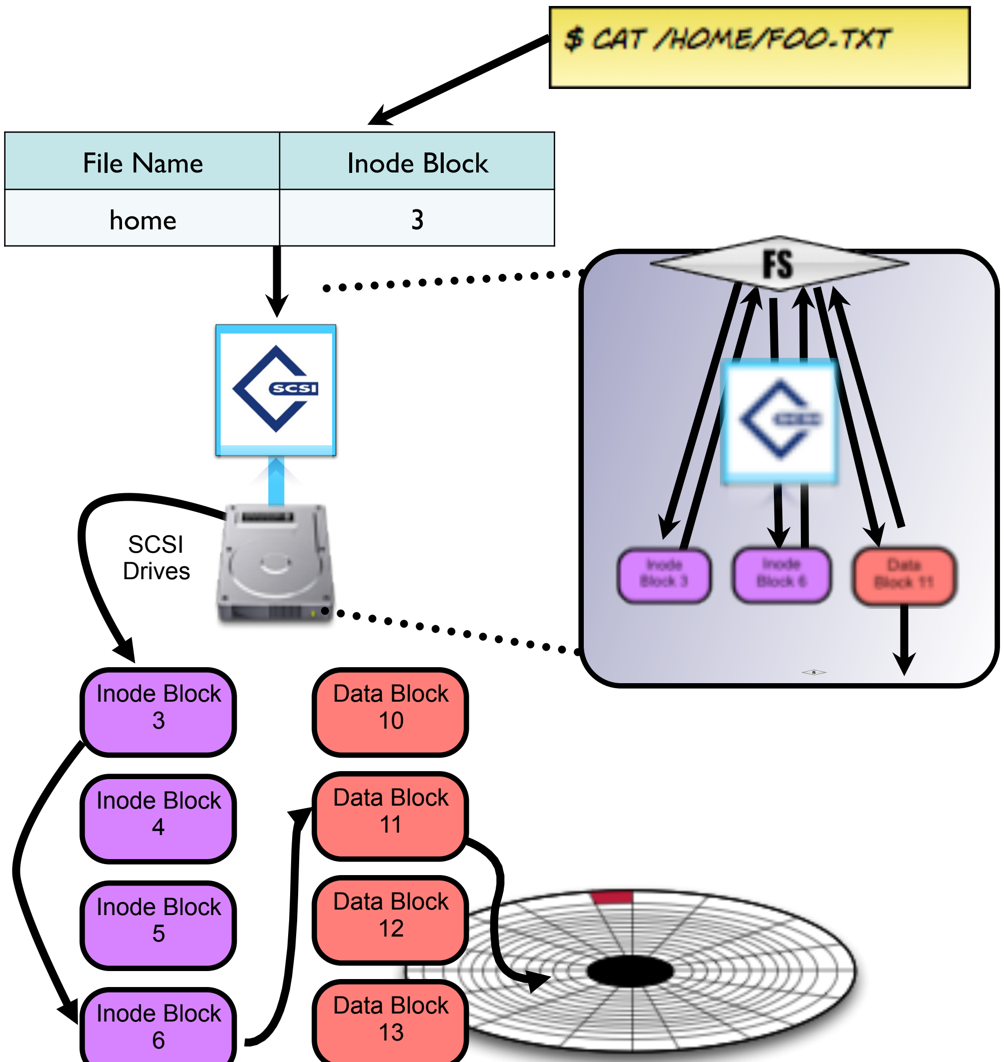
Inode Block

Owner
Mode (permission/directory/file)
Timestamp
Address
Data Block (e.g., file data, inode pointer)
[some other stuff*]

* SIMPLIFIED FOR EXPLANATORY PURPOSES AND CLARITY

- Inodes are metadata
- Mapping of files to blocks is handled through Inodes
- Each Inode describes one file
  - Every file or folder will have a corresponding Inode
- Each Inode contains a list of the disk block numbers in the file it describes
- Names of files live in Directory Structures
  - Directory Structure maps names to Inode numbers
- Directory Inodes
  - Will have Data Blocks that contain file names and inodes of the files
- File Inodes
  - Will have data blocks that contain the actual data of the file

File System Inode Process*



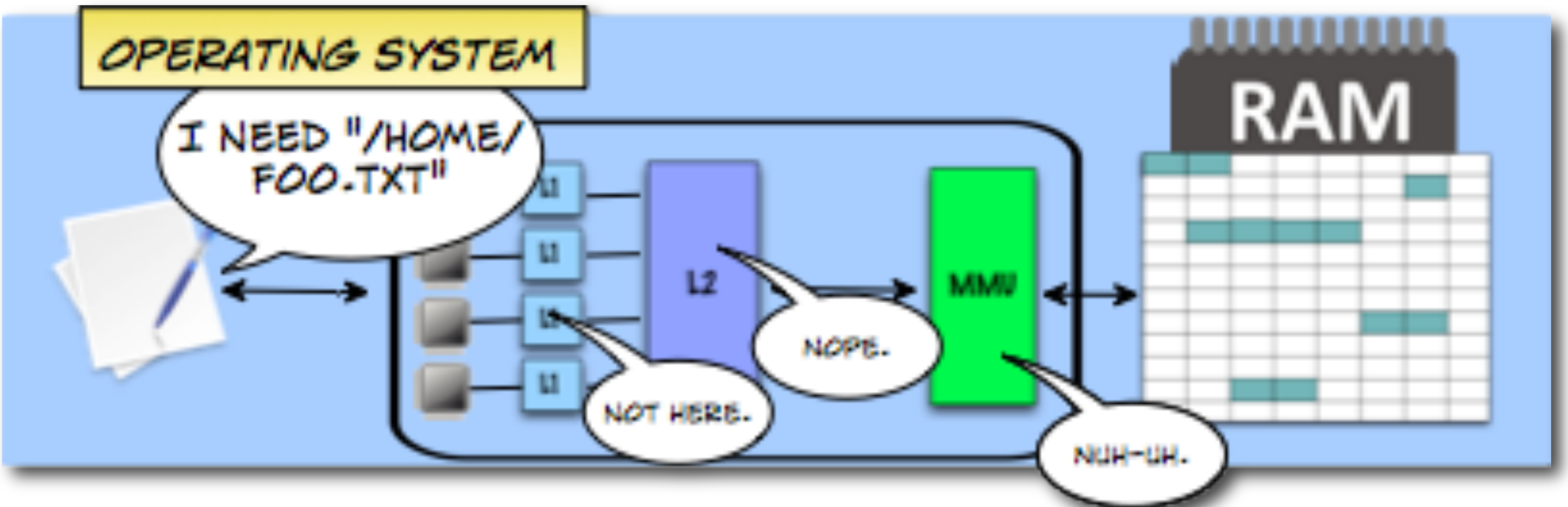
- Kernel starts at root of the file system: "/"
- Looks for directory called "home"
- Goes to that Inode, sees its a directory structure, looks up entry for "foo.txt"
- Goes to that Inode for "foo.txt" and holds list of disk block numbers
  - Step that converts from FS into real disk block numbers
- Goes to SCSI controller, which puts those block numbers into a SCSI command (e.g., "read")

* SIMPLIFIED FOR EXPLANATORY PURPOSES AND CLARITY

Storage Packet Walk

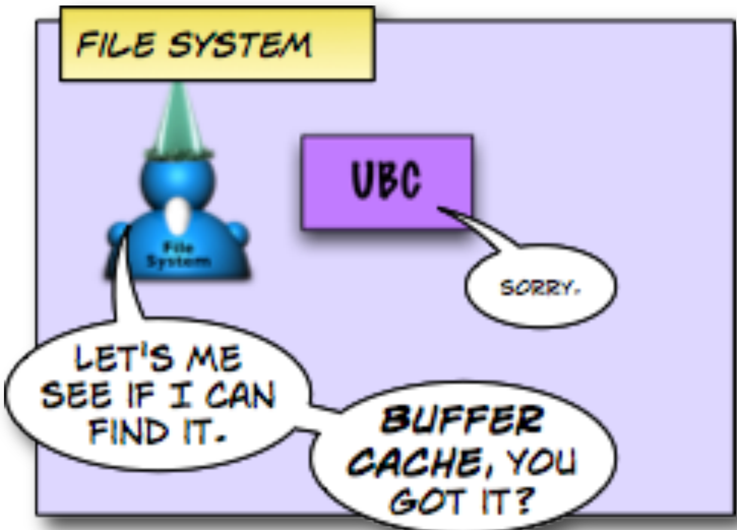
SNIA™
Ethernet Storage Forum

Packet Walk

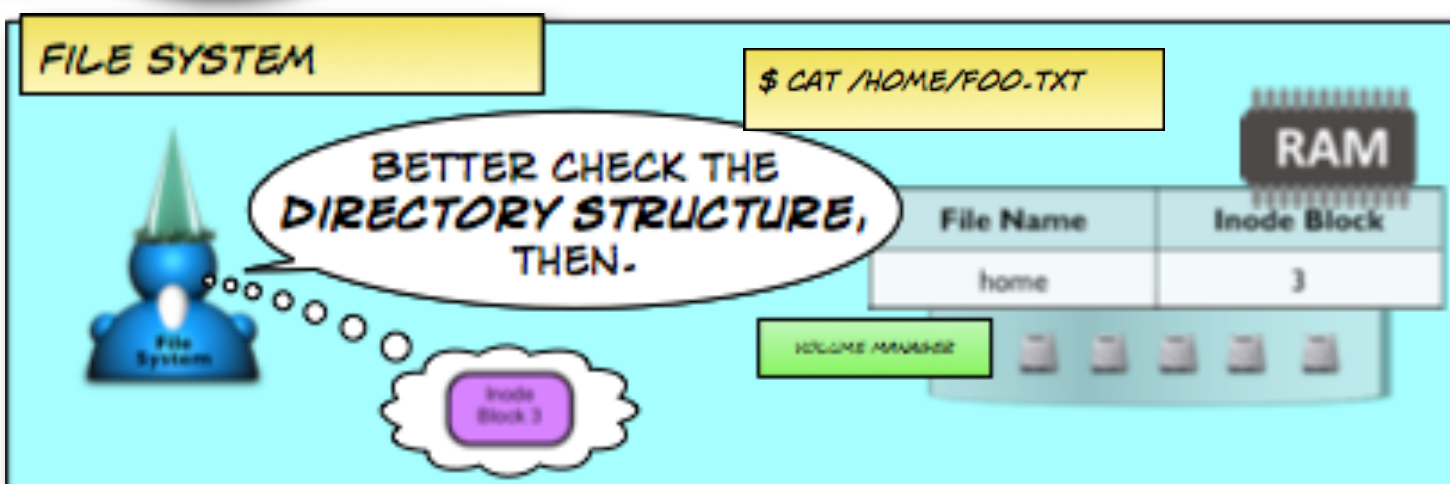


- Operating System checks main memory (and associated caches) first
- If not there, it needs to use the file system's ability to retrieve from disk

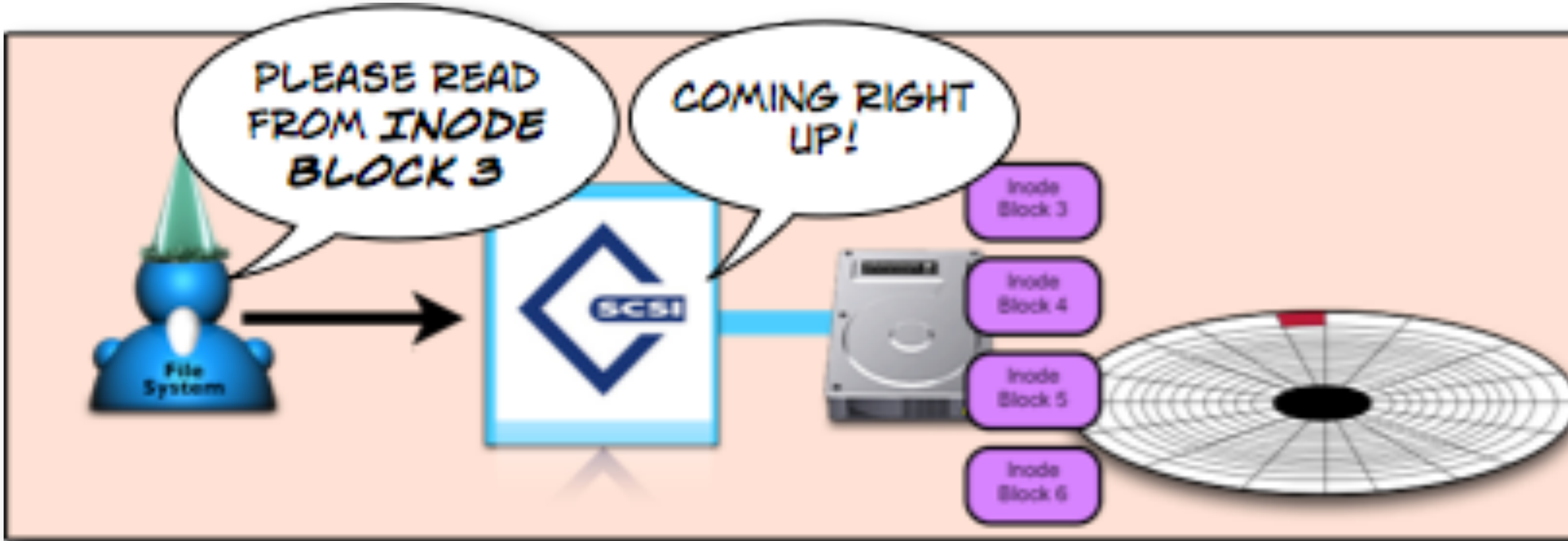
Packet Walk

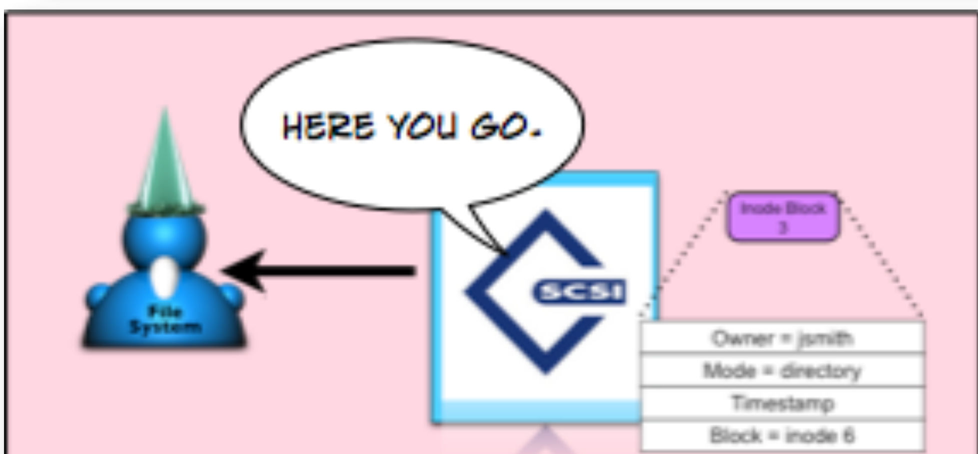


- File system checks Unified Buffer Cache (UBC) to see if file has been previously accessed
- If not, then it needs to check its directory structure to see how the file name is associated with a disk block (i.e., Inode)



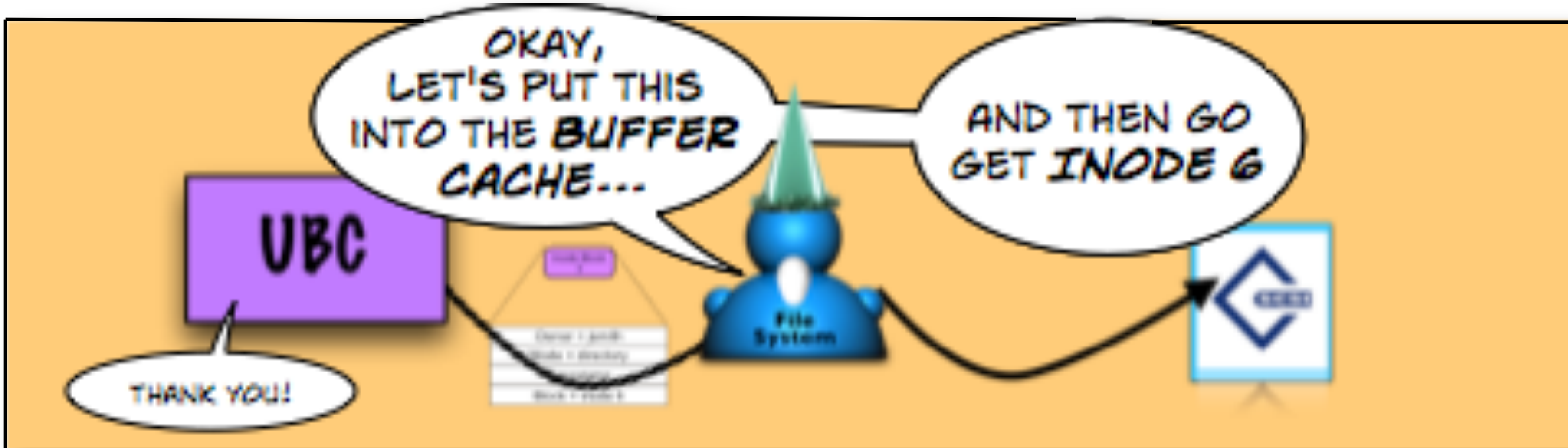
Packet Walk





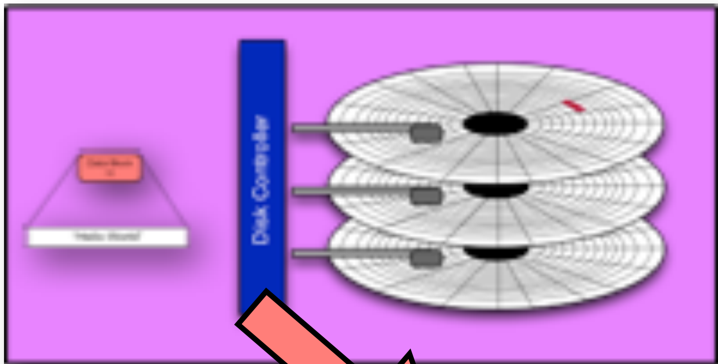
- SCSI controller returns the contents of the Inode
- The response is a directory ("home"), which has its own Inode pointer

Packet Walk

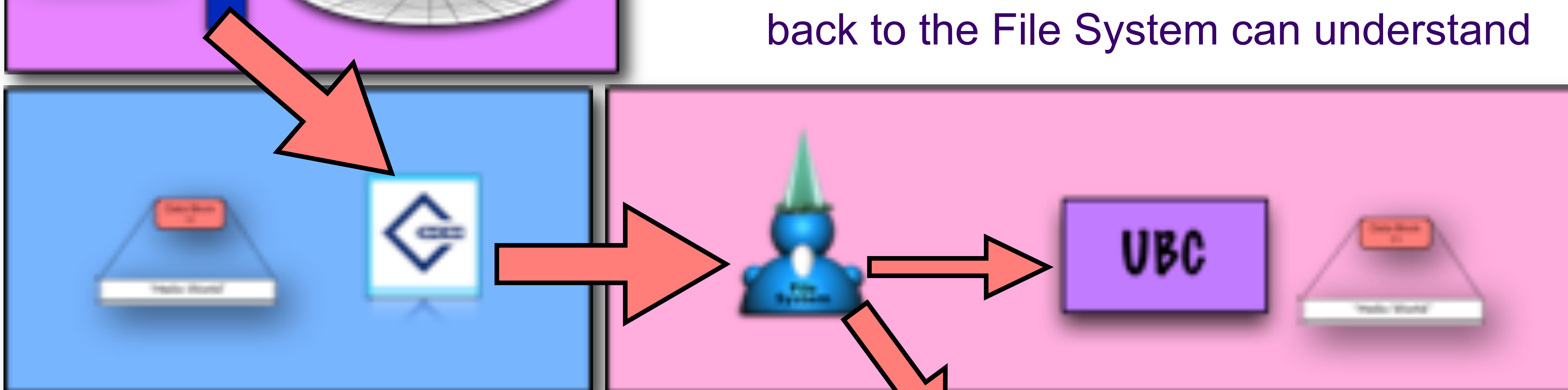


- The data is put into the Unified Buffer Cache (i.e., specifically an Inode cache)
- Process continues until we get to actual data blocks

Packet Walk



- The contents of the data block is collected by the Drive controller, and sent to the SCSI controller
- SCSI controller translates the format back to the File System can understand

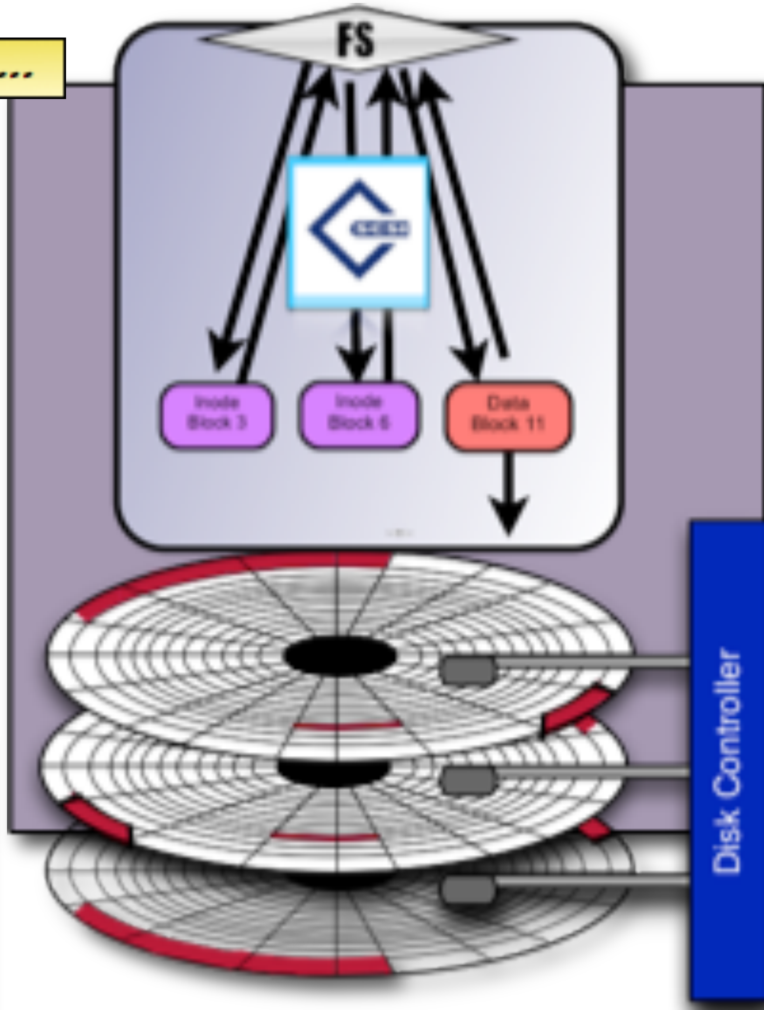


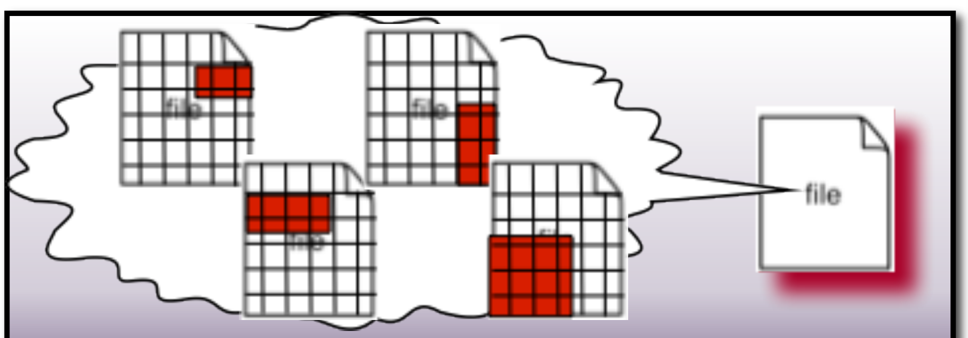
- File system copies data to its own buffer cache
- Sends data back to application, storing copies in cache in main memory

Packet Walk

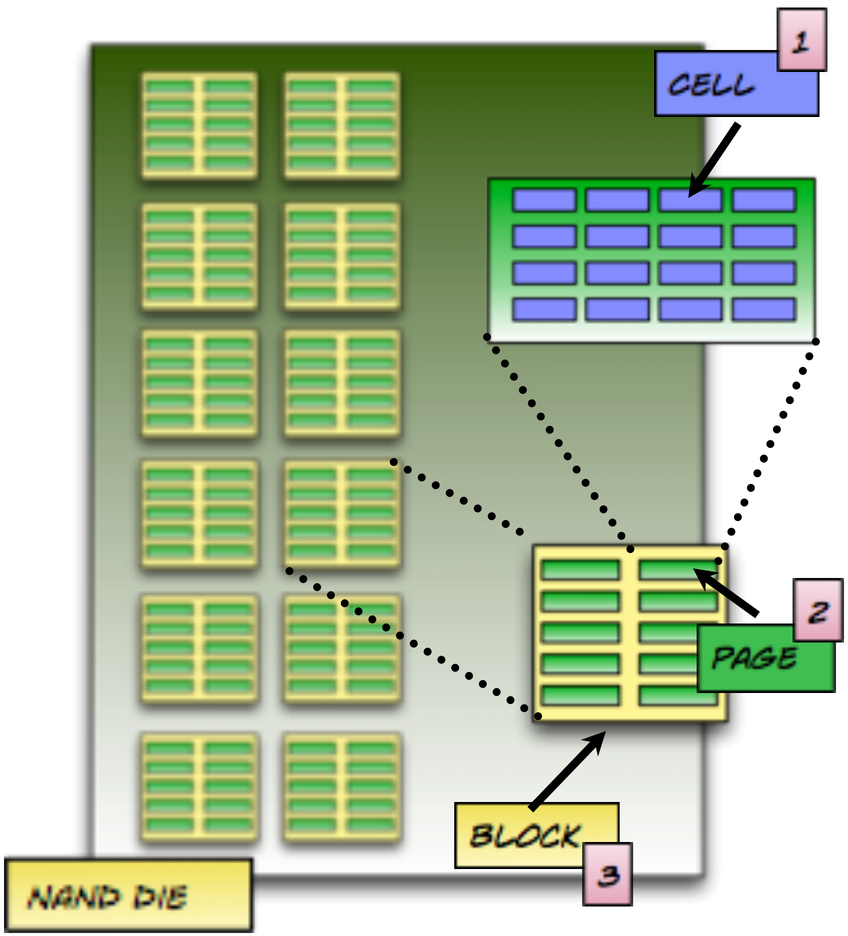
SOMETHING TO CONSIDER...

- Each time we access blocks we have to wait for the platters to move to the correct address for the read/write head to access the contents
- If the blocks are not stored on the drive in sequence, we have to wait even longer

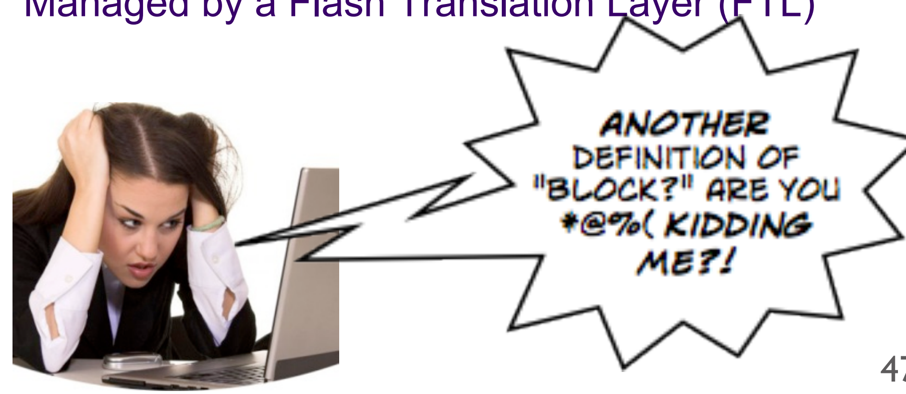




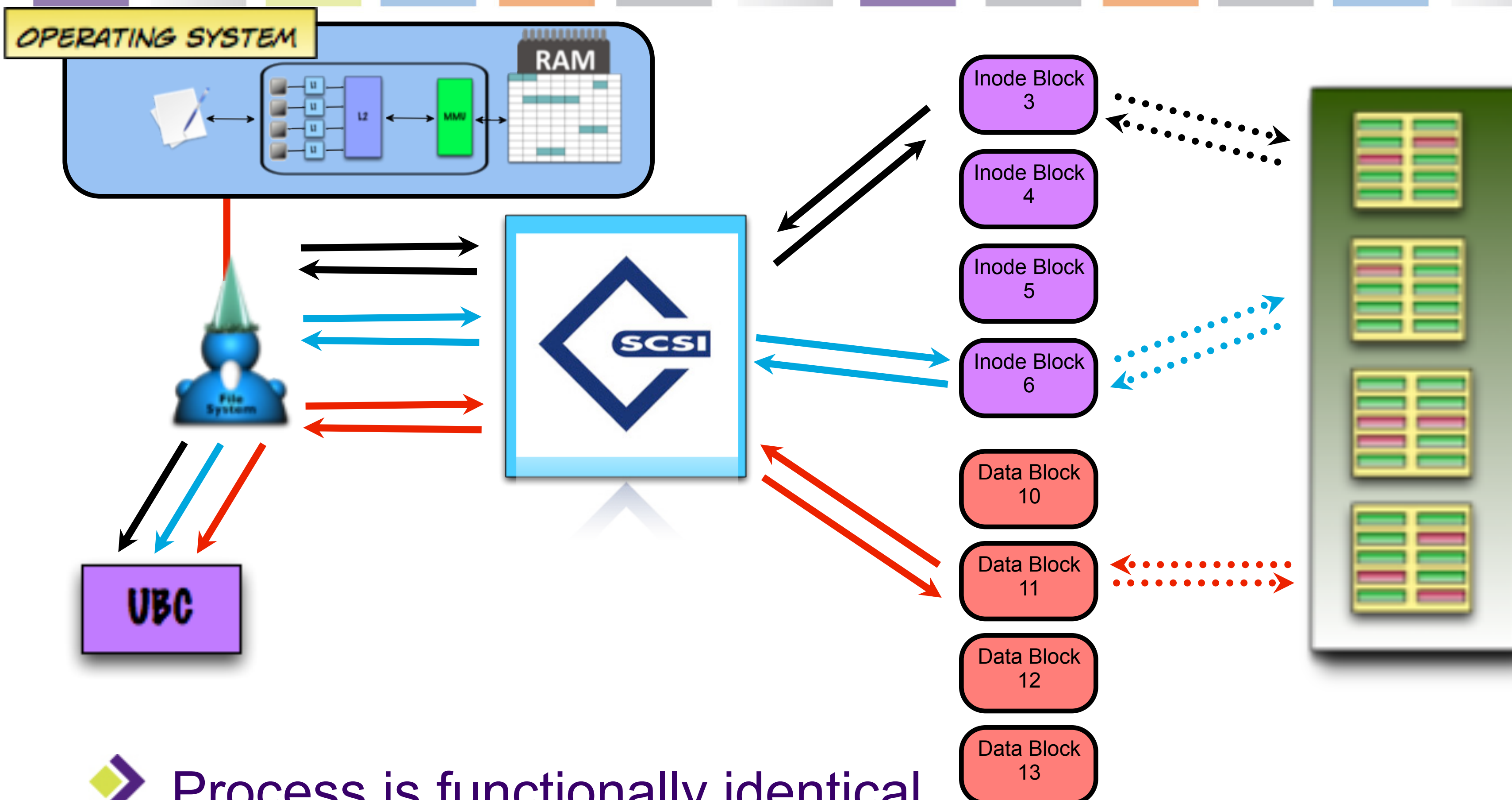
Anatomy of a Flash Drive



- No spinning media
  - All data is randomly accessed
- 1 Cells contain bits
  - SLC - 1 bit per cell
  - MLC - 2 bits per cell
  - TLC - 3 bits per cell
- 2 Pages are the smallest unit that can be programmed
  - Made up of cells
  - Can come in 2k, 4k, 8k, 16k
- 3 Blocks are the smallest unit that can be erased
  - Made up of pages
  - Most blocks have 128 or 256 pages
- Managed by a Flash Translation Layer (FTL)



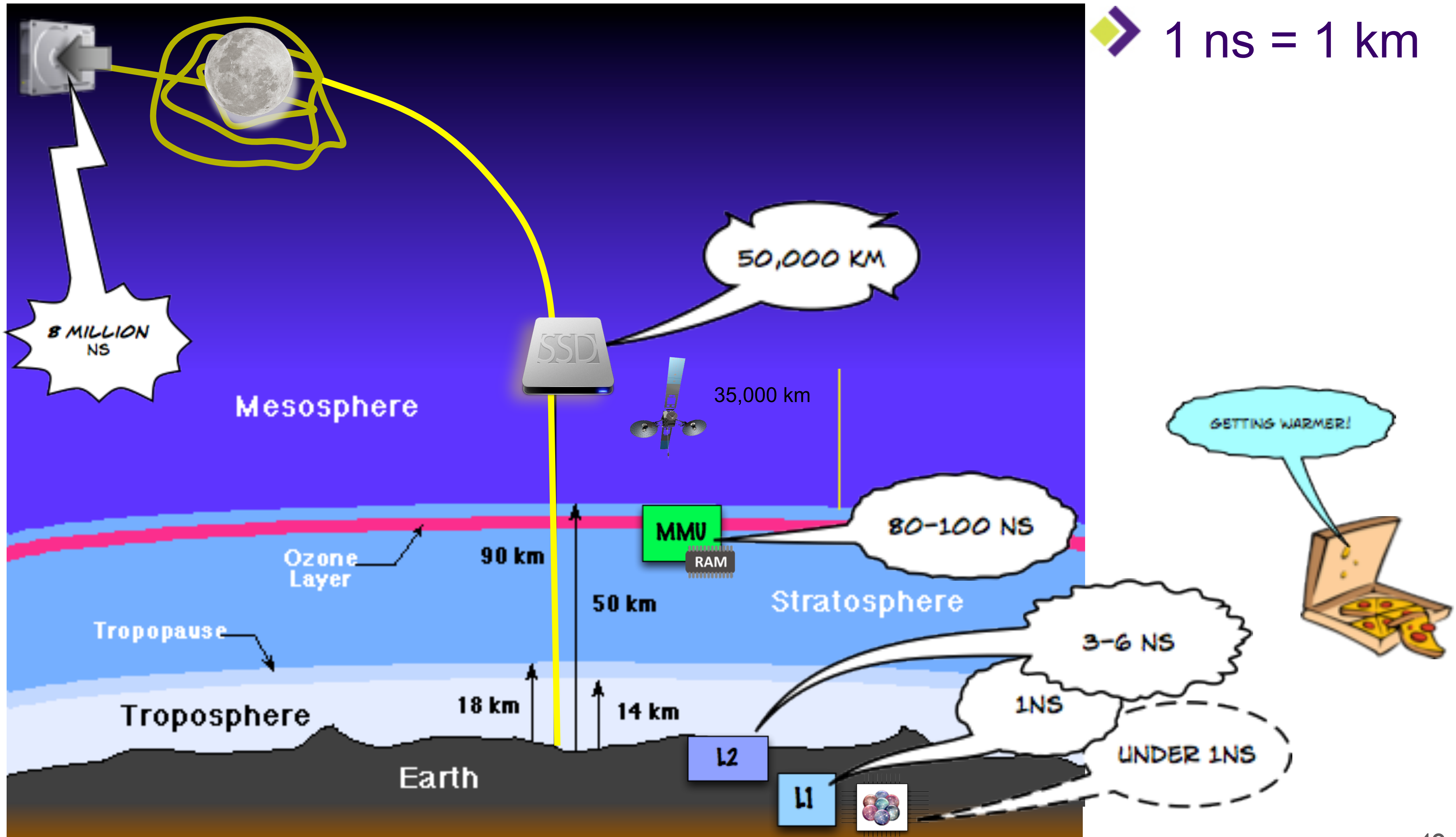
Once More, With Flash



- Process is functionally identical
- No moving parts; reduced time

Back to the Pizza Delivery

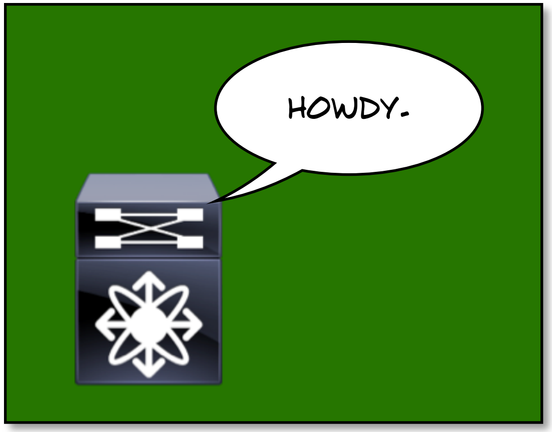
Image source: NASA



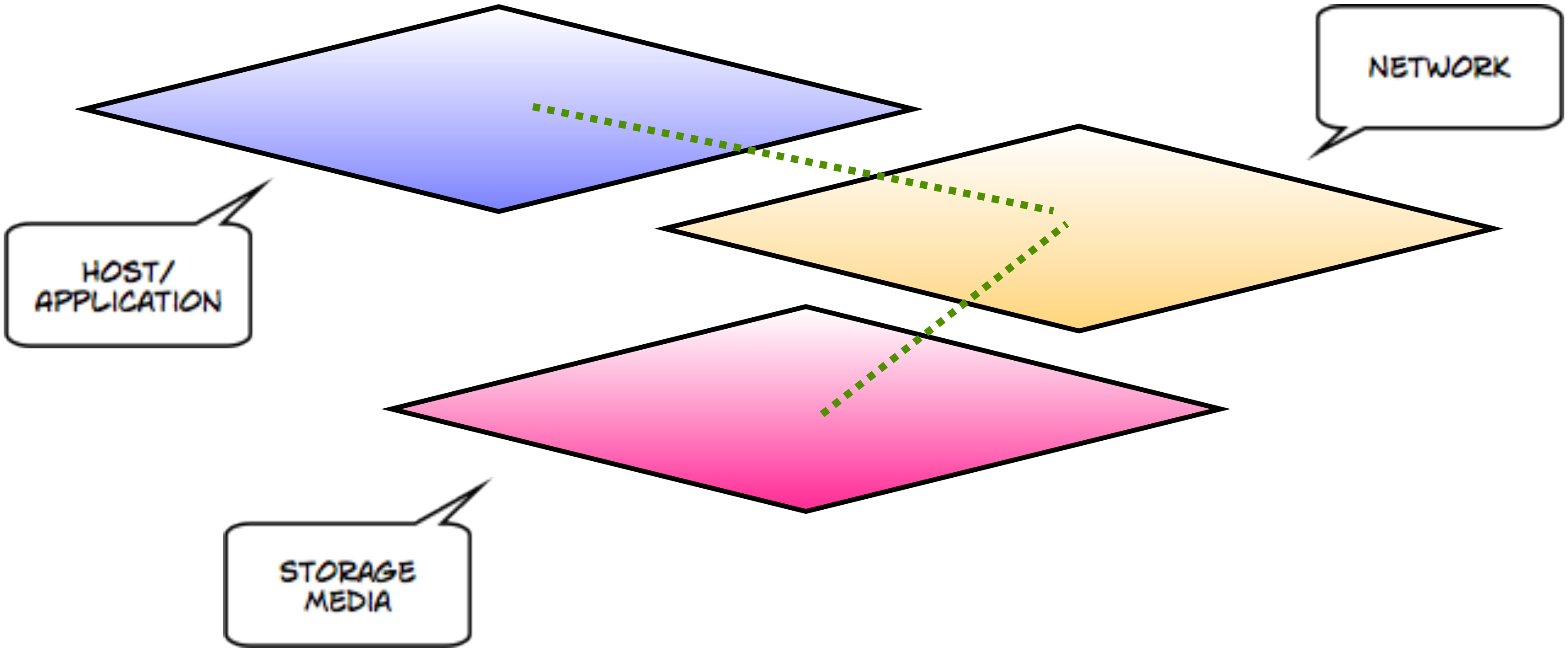
- 1 ns = 1 km

Understanding What the Network Sees

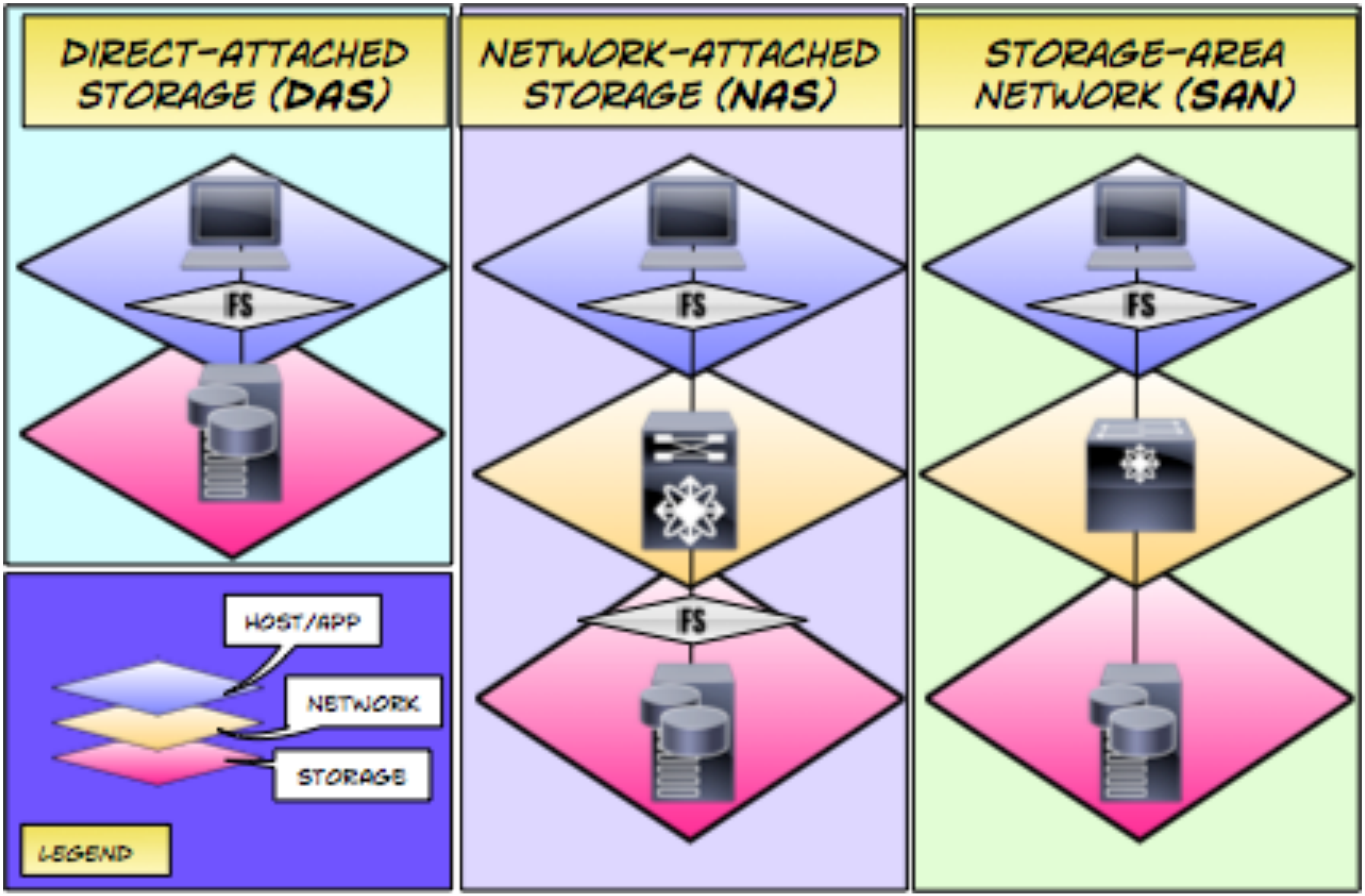
SNIA™ Ethernet Storage Forum



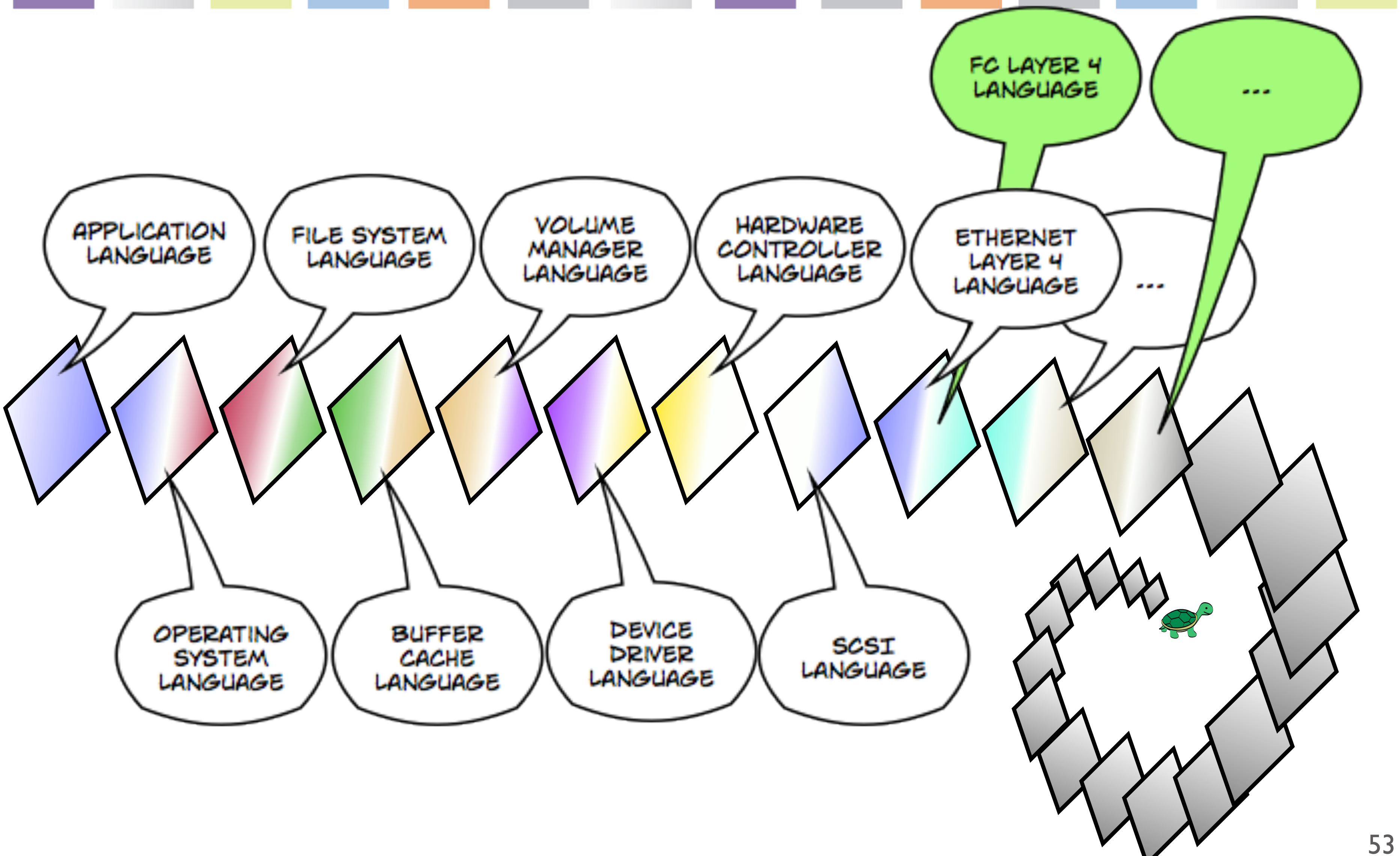
Networks



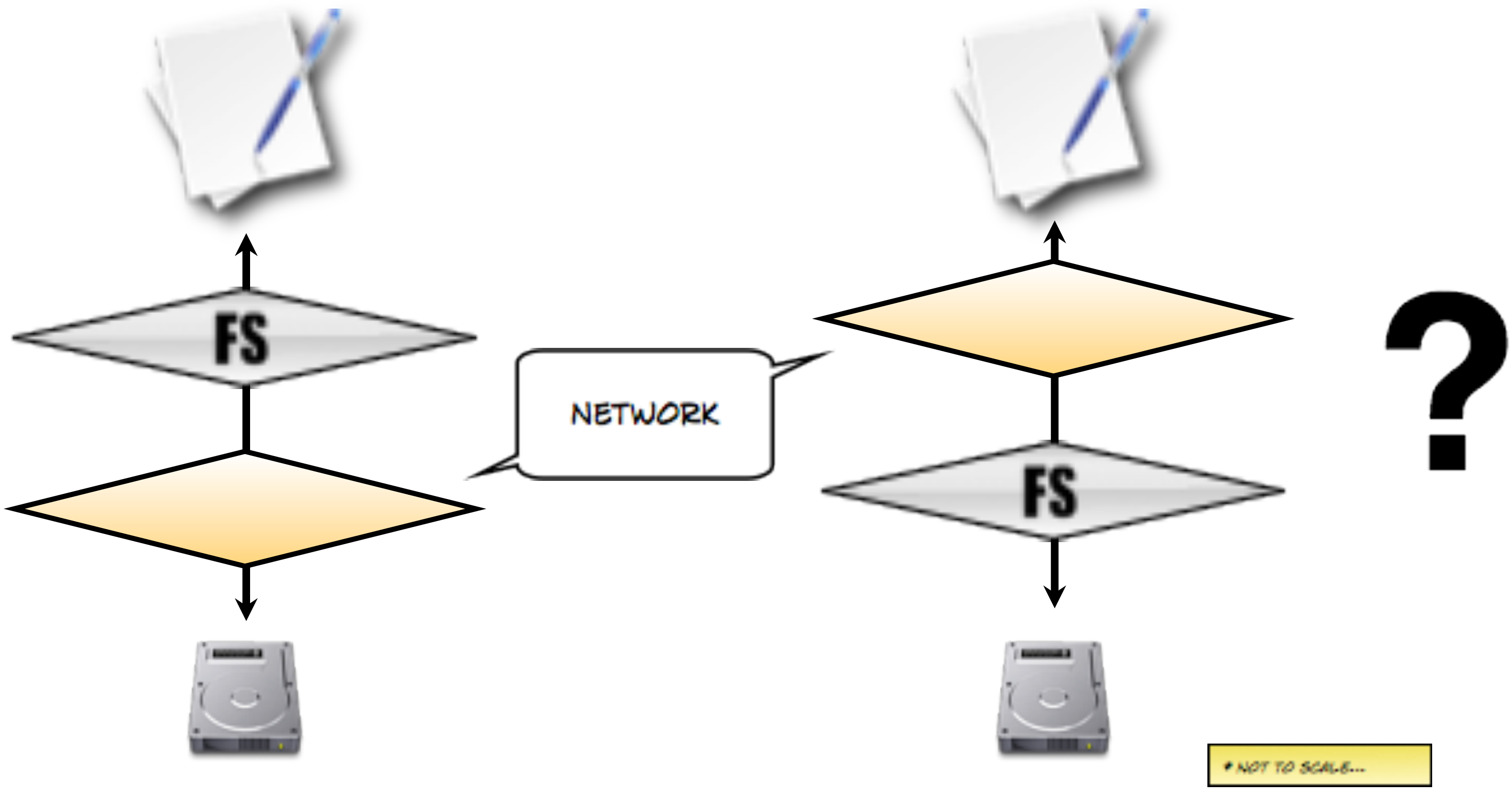
Impact of the Network



Translation Layer Extravaganza

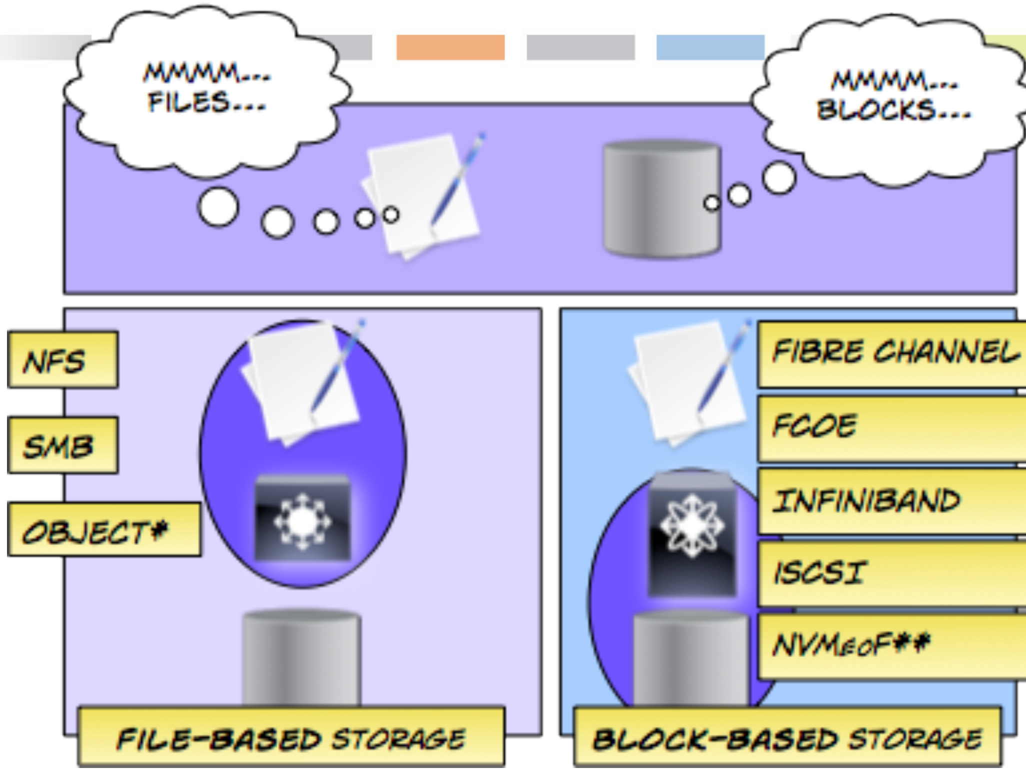


Where does the network go?



Another Way To Think About It

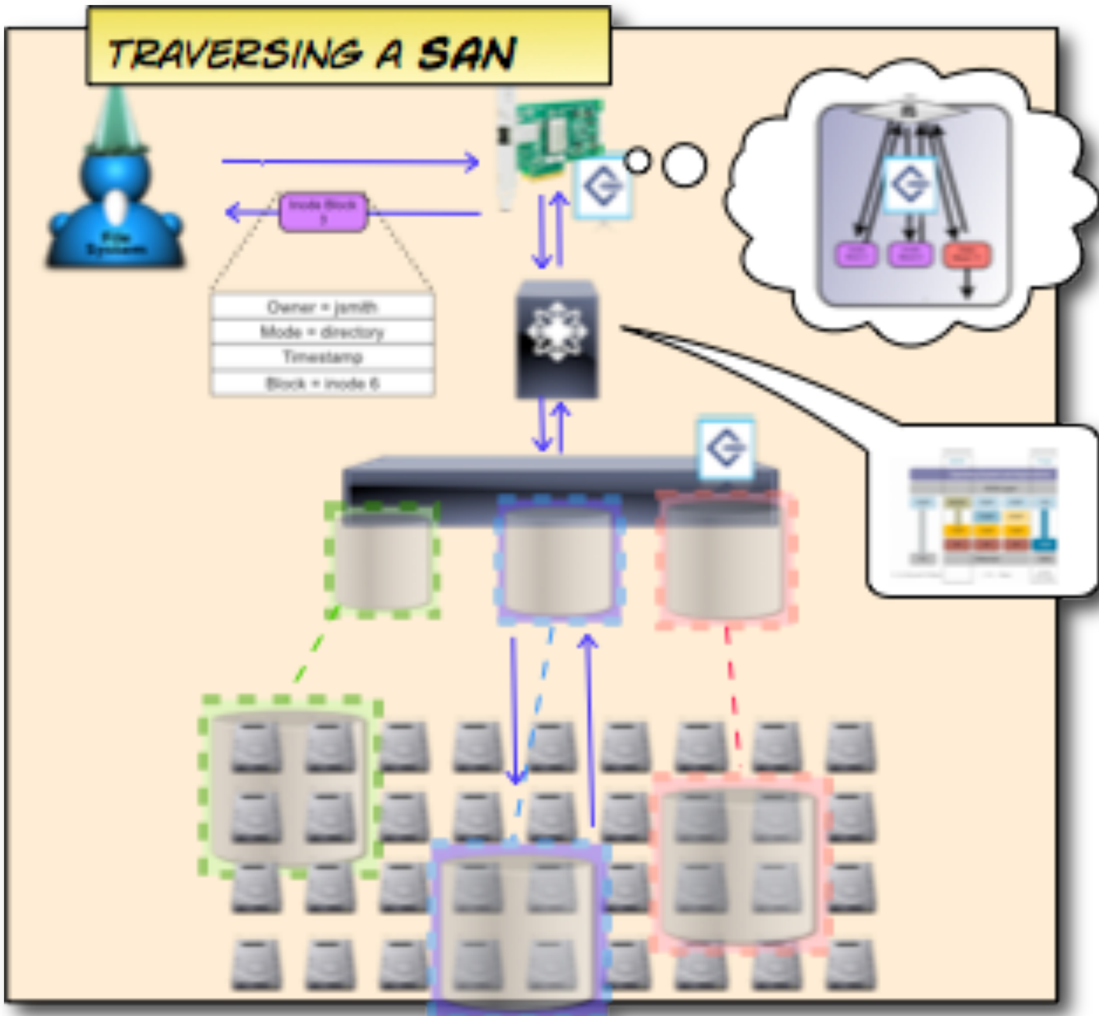
- Do you want a network to understand the applications better?
- Do you want a network to understand the storage better?
- Typical Trade-Off: Flexibility versus Performance



* OBJECT-BASED STORAGE CAN BE SEEN AS A SPECIAL CLASS OF FILE-BASED STORAGE, BUT OUTSIDE THE SCOPE OF THIS PRESENTATION BEYOND WHAT IS SHOWN HERE.

** NVMEOF IS A BLOCK-BASED PROTOCOL BASED ON NON-VOLATILE MEMORY EXPRESS, NOT SCSI, BUT ALSO FALLS OUTSIDE THE SCOPE OF THIS PRESENTATION

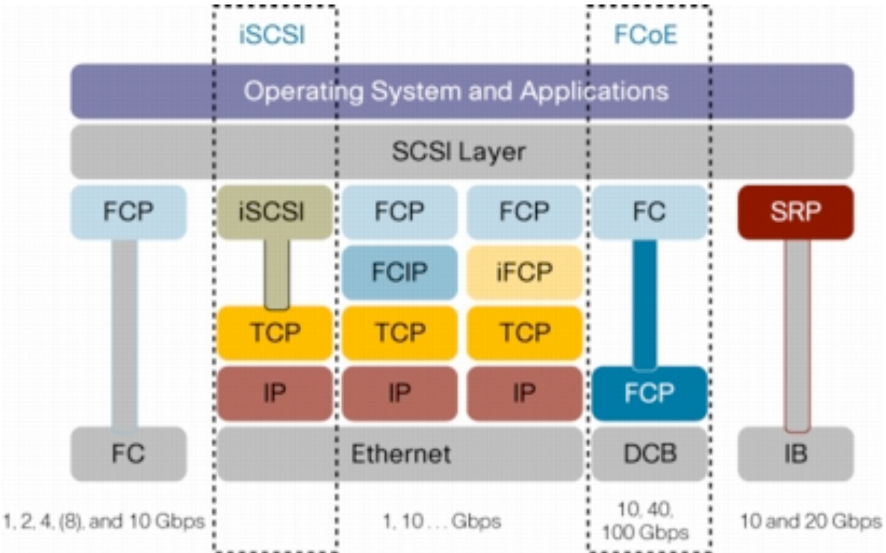
File System Inode Process to a SAN*



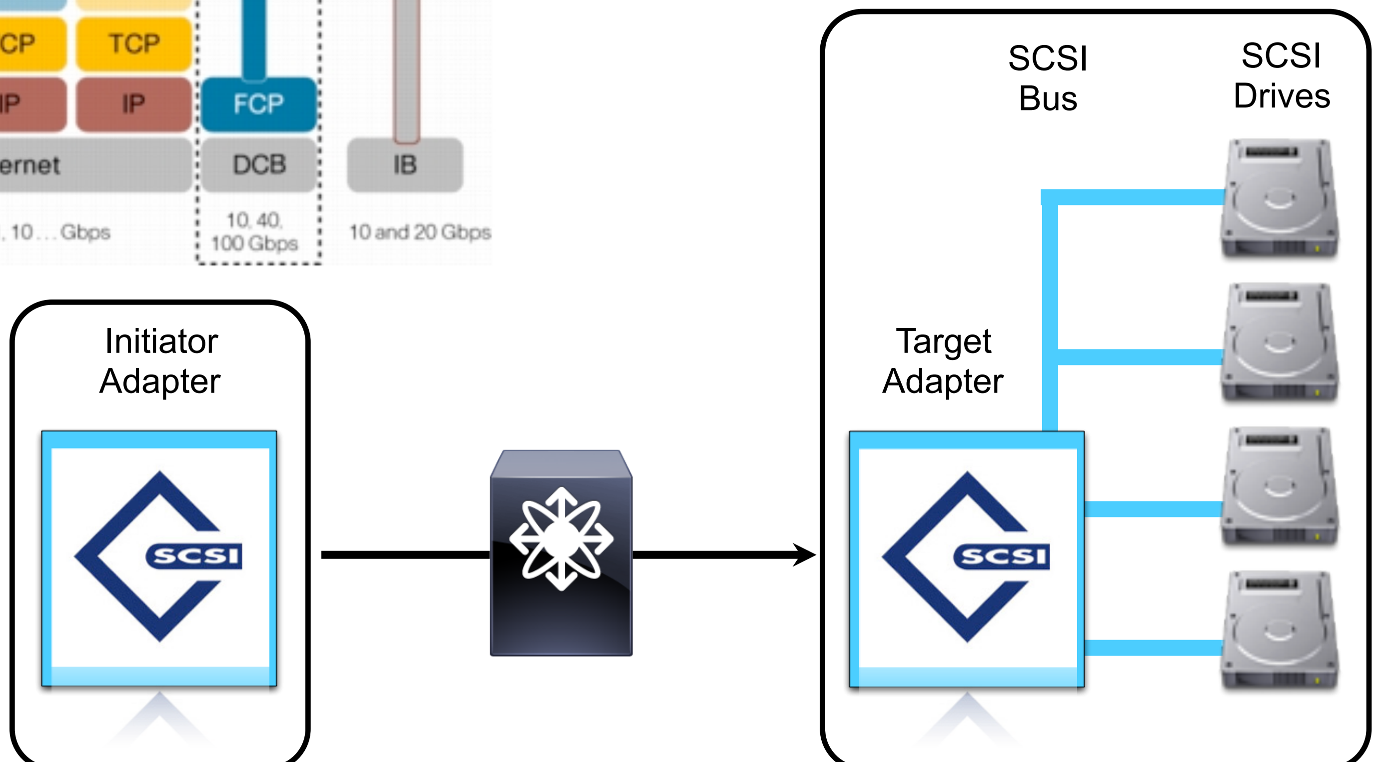
- SANs insert SCSI communication networks
- Permit consolidation of storage for multiple hosts over a storage network
- Each host controls (i.e., “owns” its assigned storage)
  - “Blue” owns blue logical drive, “Green” owns green logical drive, etc.

* SIMPLIFIED FOR EXPLANATORY PURPOSES AND CLARITY

SCSI Network Communication Options

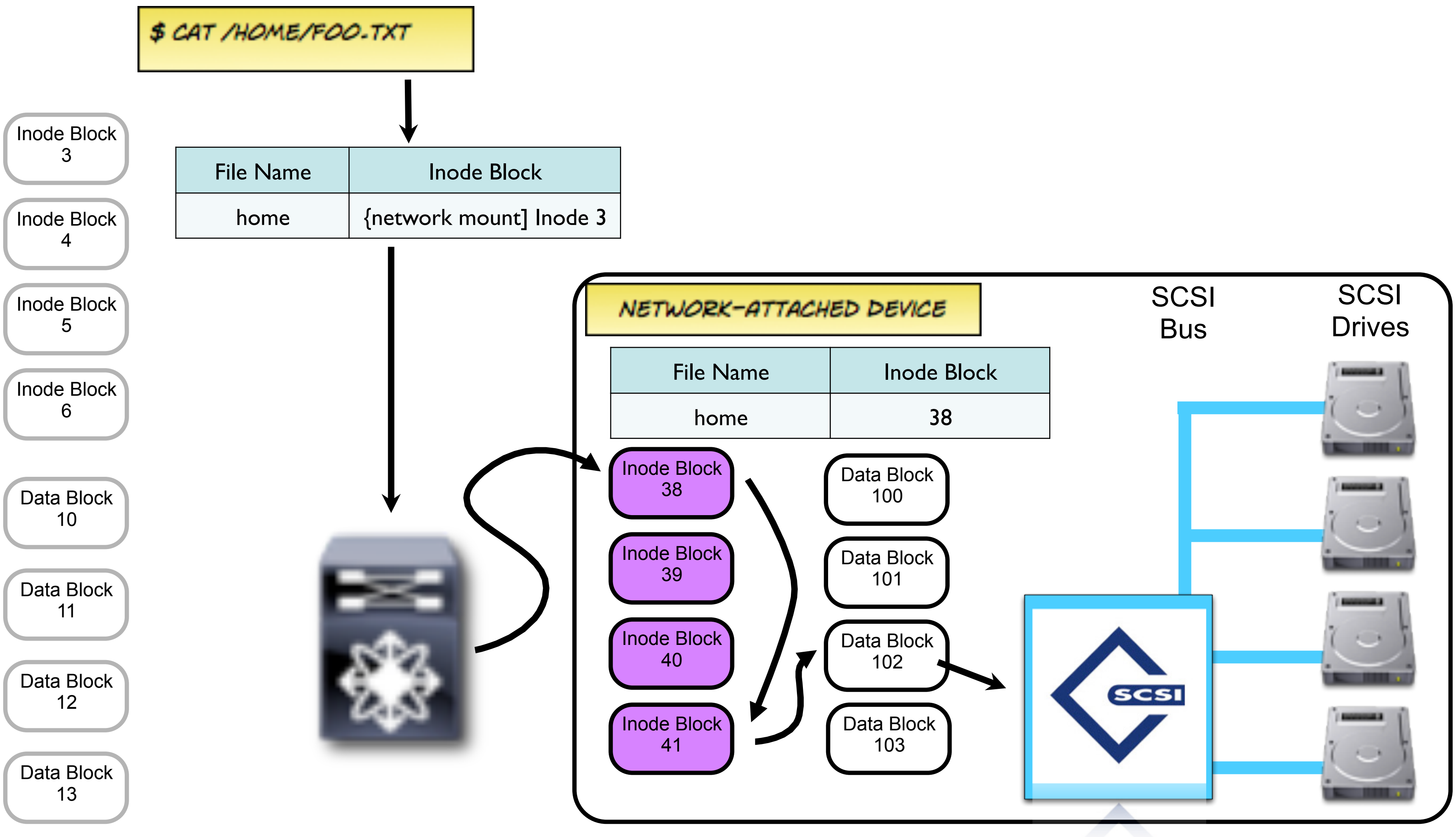


- Many different ways to have SCSI communicate over a network



File System Inode Process To Network-Attached Storage

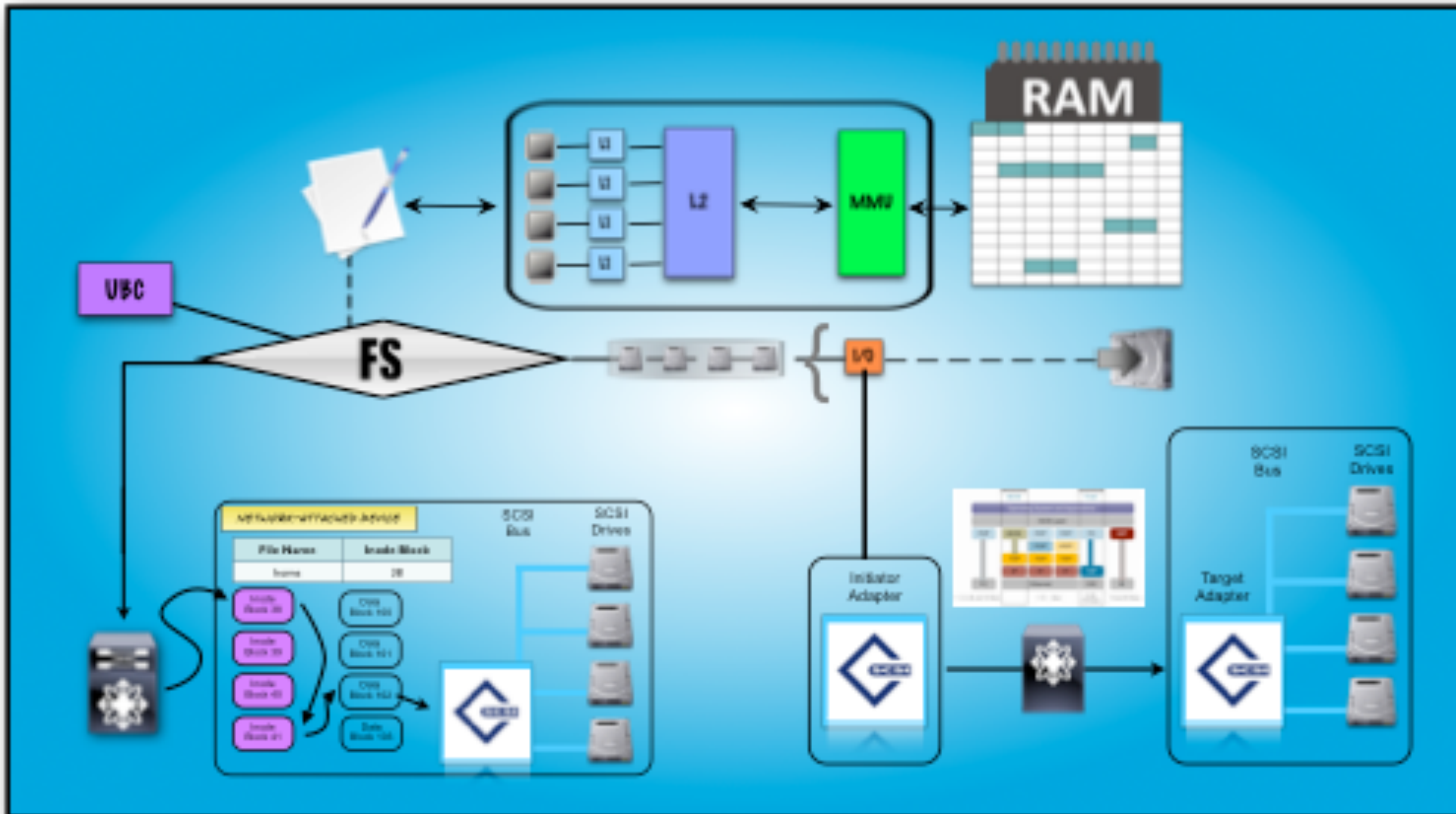
SNIA Ethernet Storage Forum



Summary

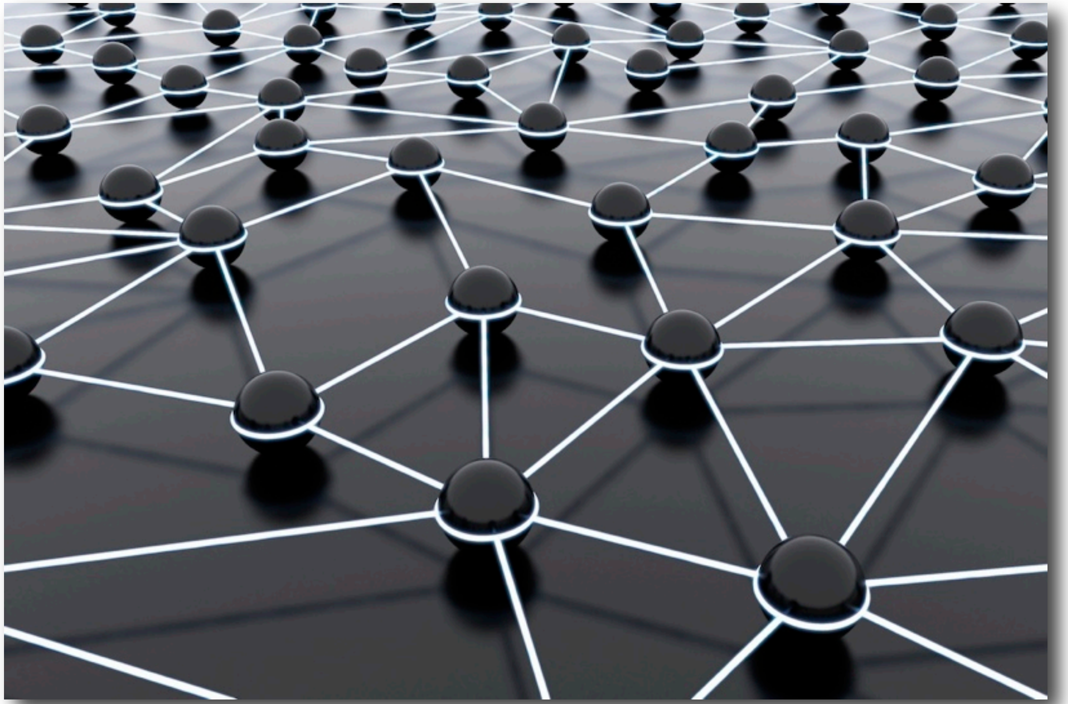
Ethernet Storage Forum

The Big Picture



Summary Points

- These were just the basics, the bare-bones minimum
- The storage process is more involved than simply having a favorite storage medium, network, or file system
- Understand the process, understand the trade-offs
- Watch more SNIA-ESF webinars to fill in the details!



Additional Info

SNIA Ethernet Storage Forum

For Your Reference

- "Anatomy of a File System" by Benno Joy
  - https://youtu.be/0Yf-W7Ps6u4
  - Excellent video on SCSI, drives, and file systems
- Operating Systems Course by John Bell, U of Illinois (Chicago)
  - Great overview of memory and I/O for OSes
  - https://www2.cs.uic.edu/~jbell/CourseNotes/OperatingSystems/
- Schulz, Greg. Cloud and Virtual Data Storage Networking. CRC Press. 2012.
  - Excellent foundation book for storage as a holistic concept (not just the networking piece).
  - http://storageio.com
- SNIA's website
  - http://snia.org

SPECIAL THANKS...

ALEX MCDONALD, NETAPP
GREG SCHULZ, STORAGEIO
STEVE CHALMERS, HP
JOE PELISSIER, CISCO
FRED KNIGHT, NETAPP

After This Webcast

- This webcast and a PDF of the slides will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand
- http://www.snia.org/forums/esf/knowledge/webcasts

- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-ESF blog
- http://sniaesfblog.org/

- Follow us on Twitter @ SNIAESF

THANK YOU!

SNIA™
Ethernet Storage Forum