STORAGE DEVELOPER CONFERENCE

SDC 22 | Fremont, CA
September 12-15, 2022

BY Developers FOR Developers

A SNIA Event

DNA Data Storage

DNA Data Storage Alliance - Rosetta Stone Initiative

Presented by

- Joel Christner, Director, Distinguished Engineer, Dell Technologies
- Alessia Marelli, CTO, DNAalgo
- Mark Wilcox, CEO and CTO, 21e8

DNA DATA STORAGE ALLIANCE

A SNIA Technology Affiliate

Helpful Links

- Preserving our Digital Legacy – an Introduction to DNA Data Storage

Agenda

- Differences: DNA vs Traditional Media
- Overview of the Rosetta Stone
- How to Participate
- Summary

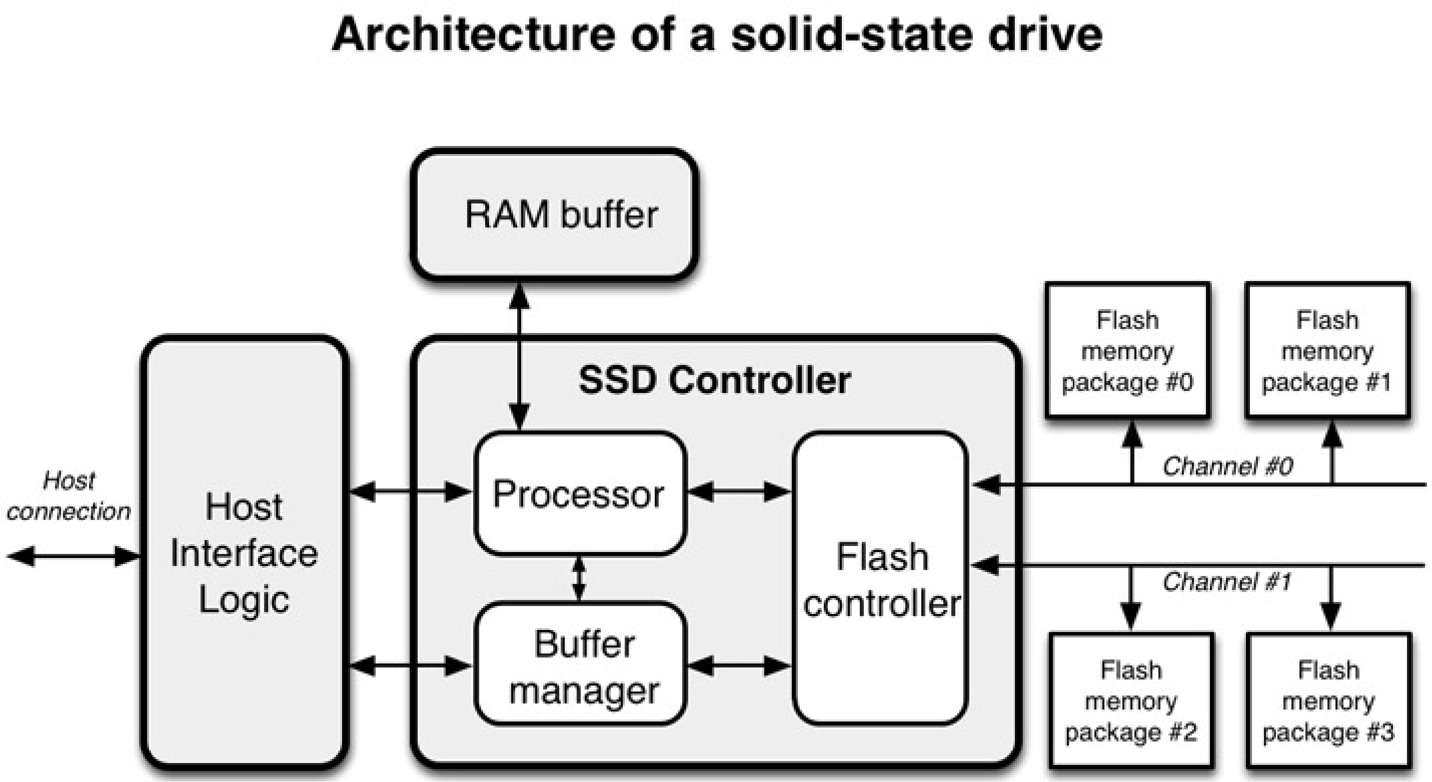
DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

STORAGE DEVELOPER CONFERENCE
SDC 22

Differences: DNA vs Traditional Media

Differences: DNA vs Traditional Media

1. Exposing a device to the system

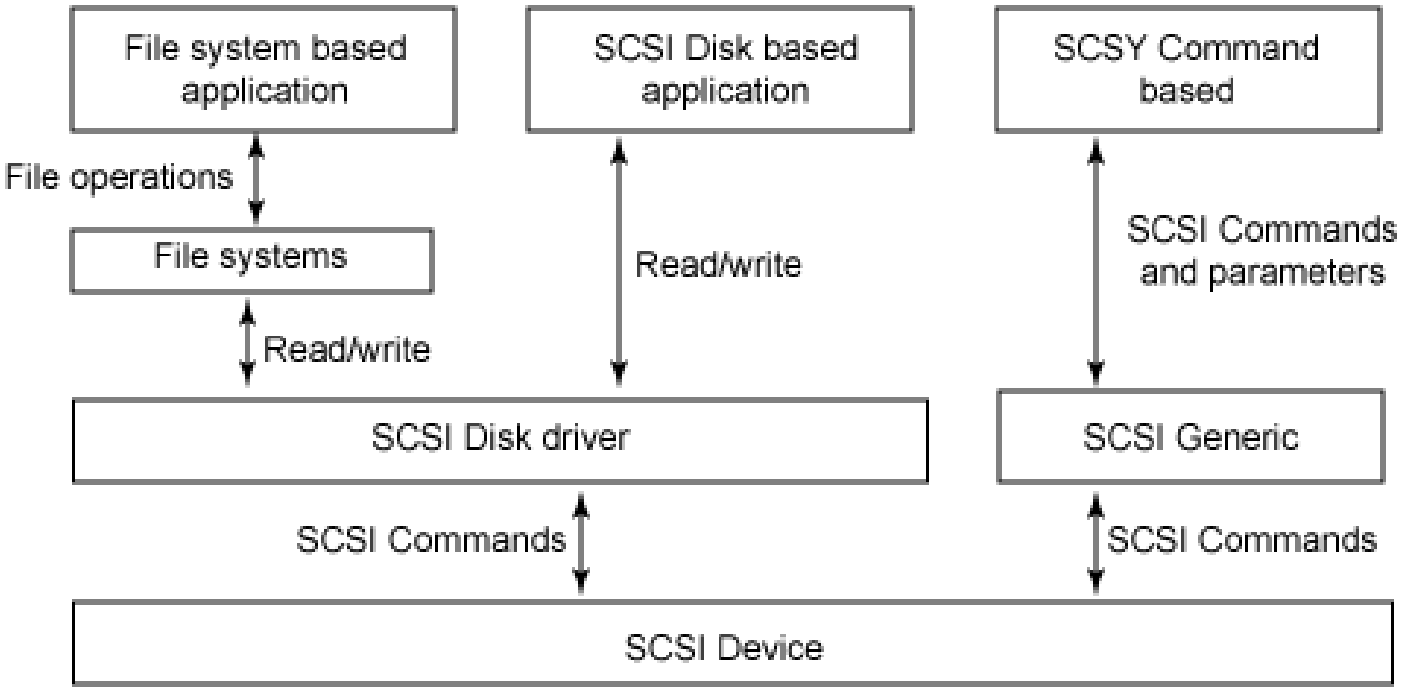


DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

STORAGE DEVELOPER CONFERENCE
SDC 22

Differences: DNA vs Traditional Media

2. Organizing abstractions to create filesystem storage



Differences: DNA vs Traditional Media

3. Media without integrated controller

Barcode with volume serial number, generation, and type of cartridge

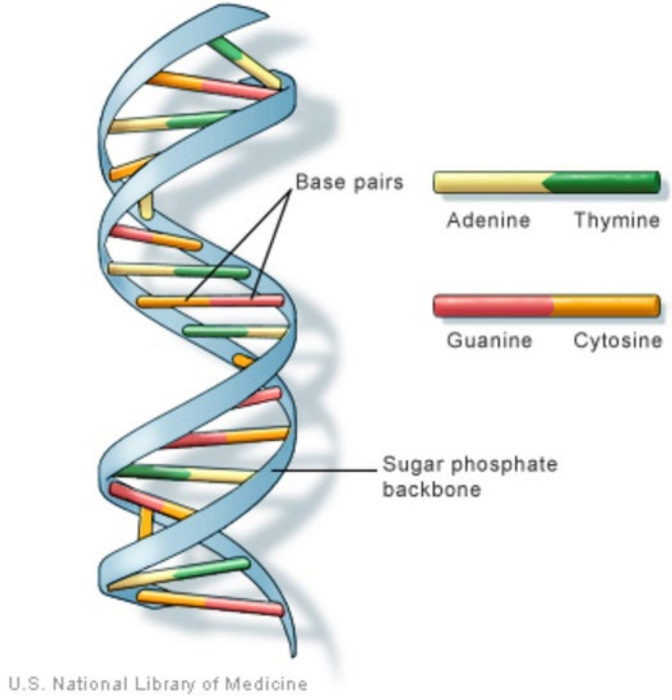
Read LTFS from beginning of the tape

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

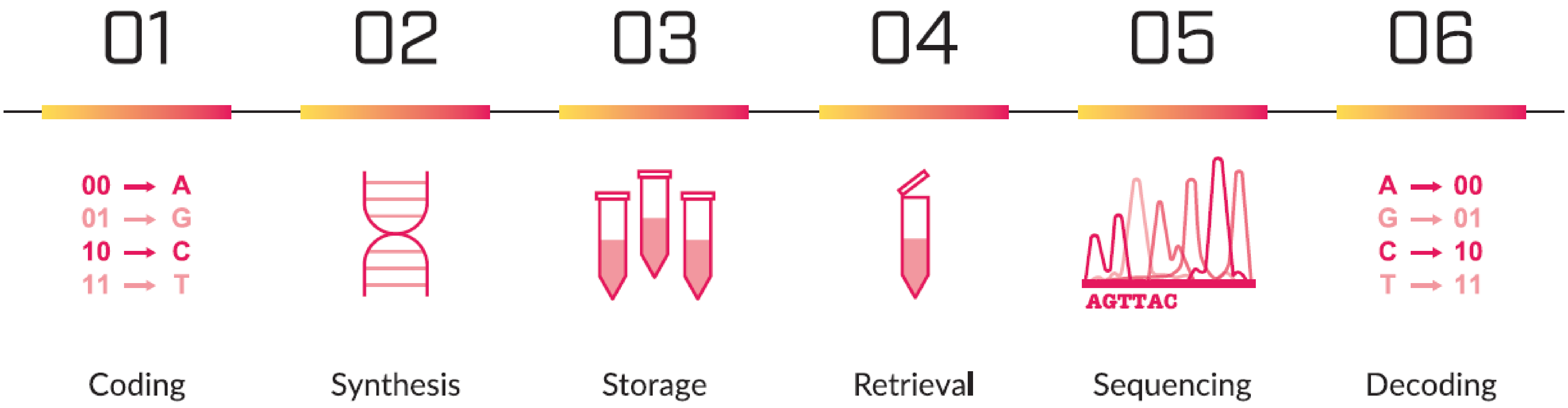
STORAGE DEVELOPER CONFERENCE
SDC 22

A "Primer" on DNA Data Storage Media

- The fundamental unit of storage in DNA is an oligonucleotide (also called 'oligo')
  - Short, single strand of synthetic DNA or RNA
  - Often a sugar phosphate backbone
  - Base compounds **_A_**denine, **_C_**ytosine, **_T_**hymine, **_G_**uanine
  - Base compounds attach to the strand …
  - … and to a mate on the opposing strand
  - Adenine bonds w/ Thymine, Guanine bonds w/ Cytosine
- A DNA molecule is a pair of strands (oligos), tightly wound around one another, held together by the bonds between the bases



A “Primer” on DNA Data Storage Media



DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

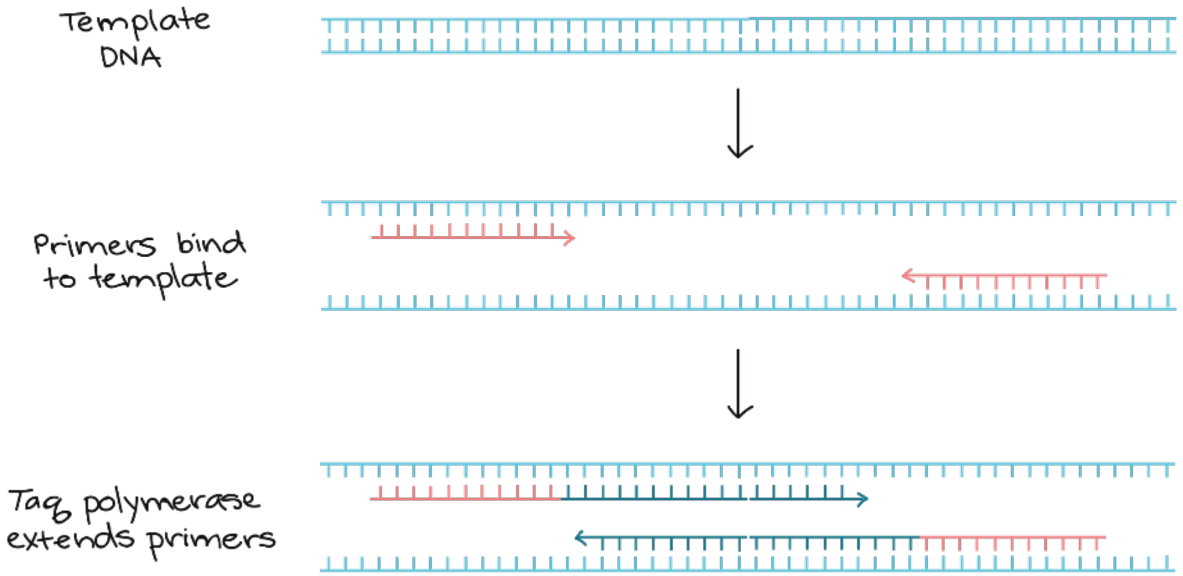
STORAGE DEVELOPER CONFERENCE

A “Primer” on DNA Data Storage Media

- The process of storing binary data into DNA data storage media involves
  - Coding – conversion of binary numbers to ATGC base pairs (“bits to bases”)
  - Synthesis – creation of the strands and chemical bonds between the base compounds
  - Storage – placing constructed strands into sealed medium until contents are needed
- The process of retrieving binary data from DNA data storage media involves
  - Retrieval – accessing the sealed medium containing the required strands
  - Sequencing – discerning the bases found in a segment of DNA
  - Decoding – converting ATGC pairs into binary (“bases to bits”)
- DNA has neither addressable sectors (disk) or relative position (tape)
  - Locations and addresses must be encoded into the material itself

A “Primer” on DNA Data Storage Media

- A ***primer*** is a short stretch of DNA targeting a unique sequence, generally to identify the sequence for ***amplification***
- ***Polymerase chain reaction (PCR)*** is the process used to create one or many copies of the amplified DNA



DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

STORAGE DEVELOPER CONFERENCE
SDC 22

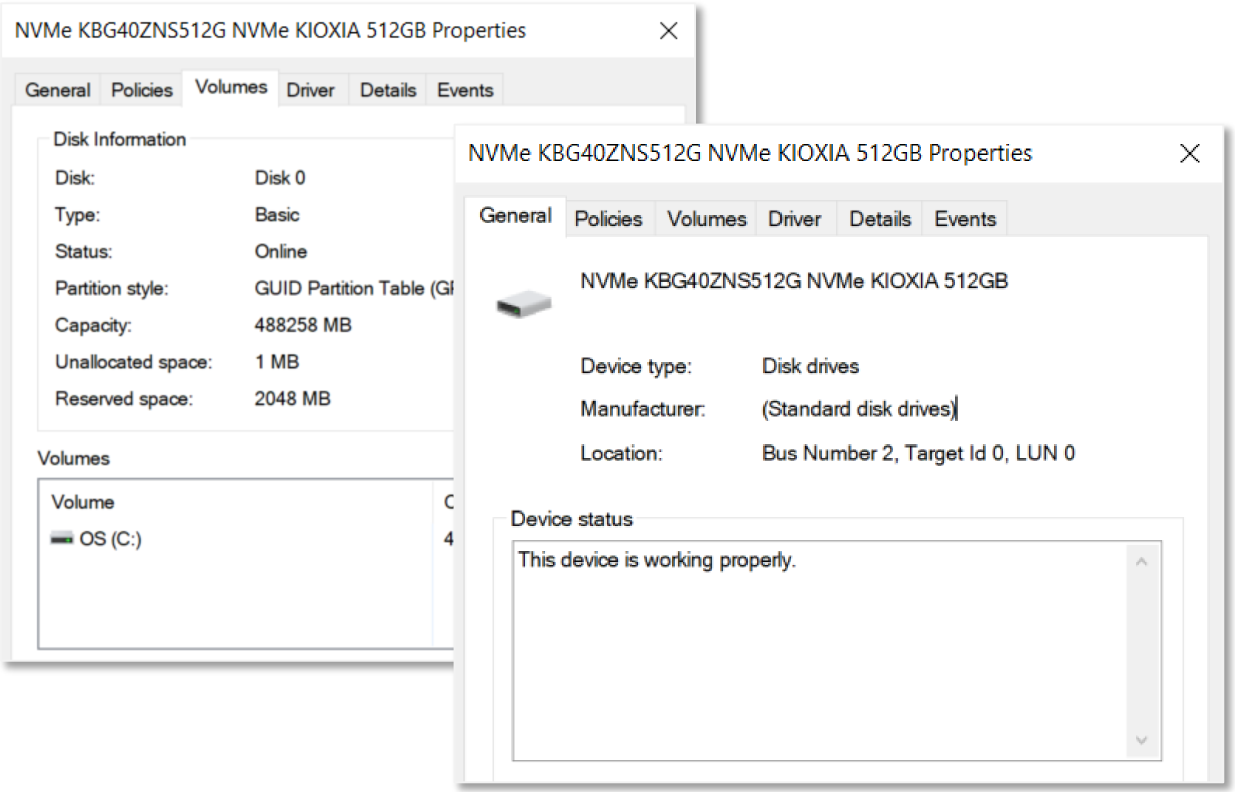
Overview of the DNA Data Storage Rosetta Stone project (DARS)

The problem

- DNA media does not share properties found in other storage media types
  - No built-in controller, or linear addressing of physical storage regions
  - No structured media topology
  - No built-in facilities for addressing specific parts of the media
  - Addresses (sectors) need to be encoded for later reading
- Multiple mechanisms (CODECs) exist for encoding data into DNA
  - CODEC must be discernable from within the media itself in a standard way
- With >100 year lifespan, we must anticipate technology evolution
  - Categories of innovation expected within DNA media and the value chain?
  - What is considered a safe assumption today that may not be one tomorrow?

What is an archive "boot record"?

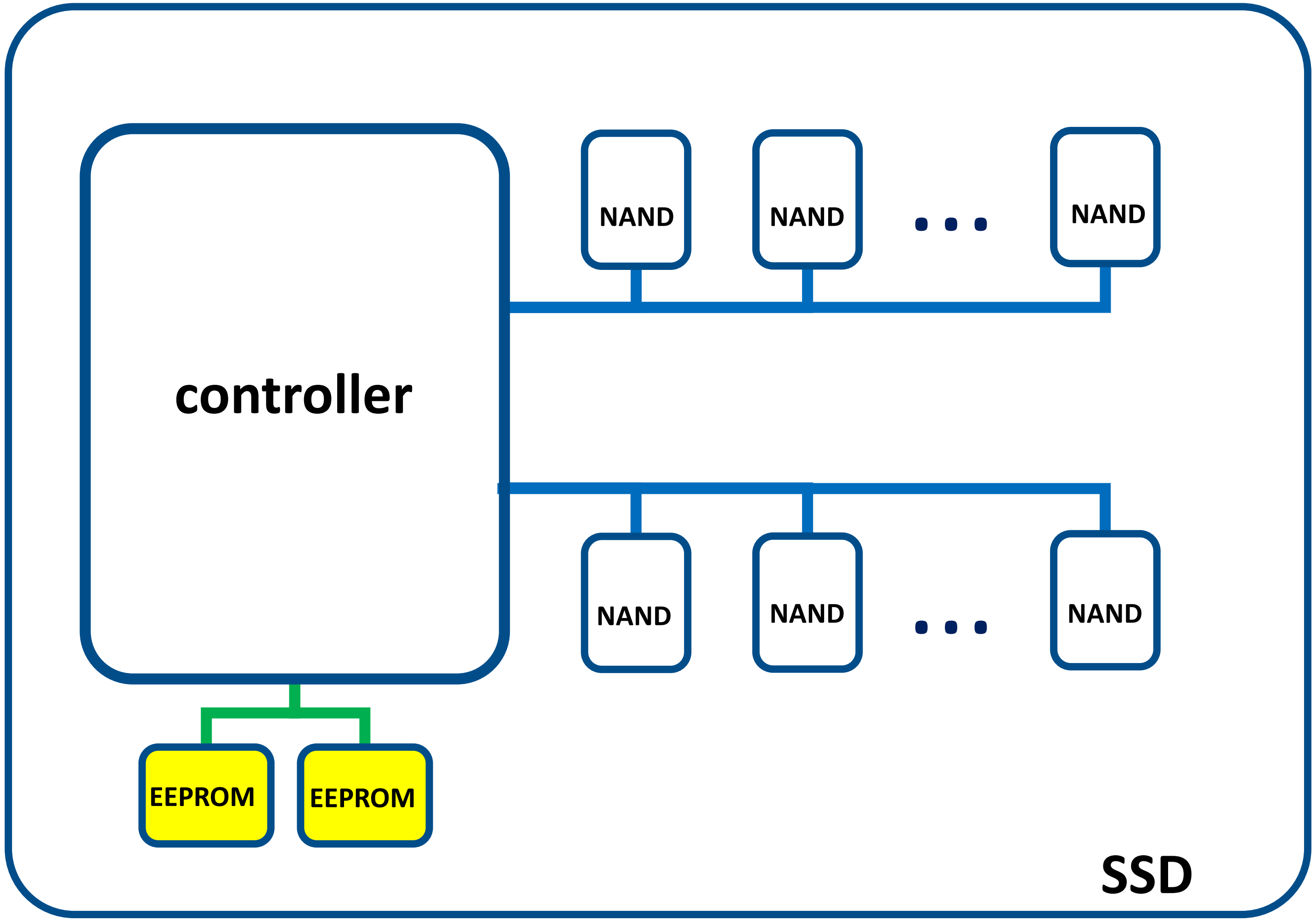
- With traditional media, controller knows where sector zero resides, packages device metadata for the consumer
  - Operating system connects to and initializes device for consumption
  - Manages translation of upper layer APIs (e.g. POSIX) into lower layer protocol primitives (e.g. SCSI)
  - Generally governed by an intermediary (e.g. filesystem)
- No controller within DNA media, no linear addressing within the media, and no file system



DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

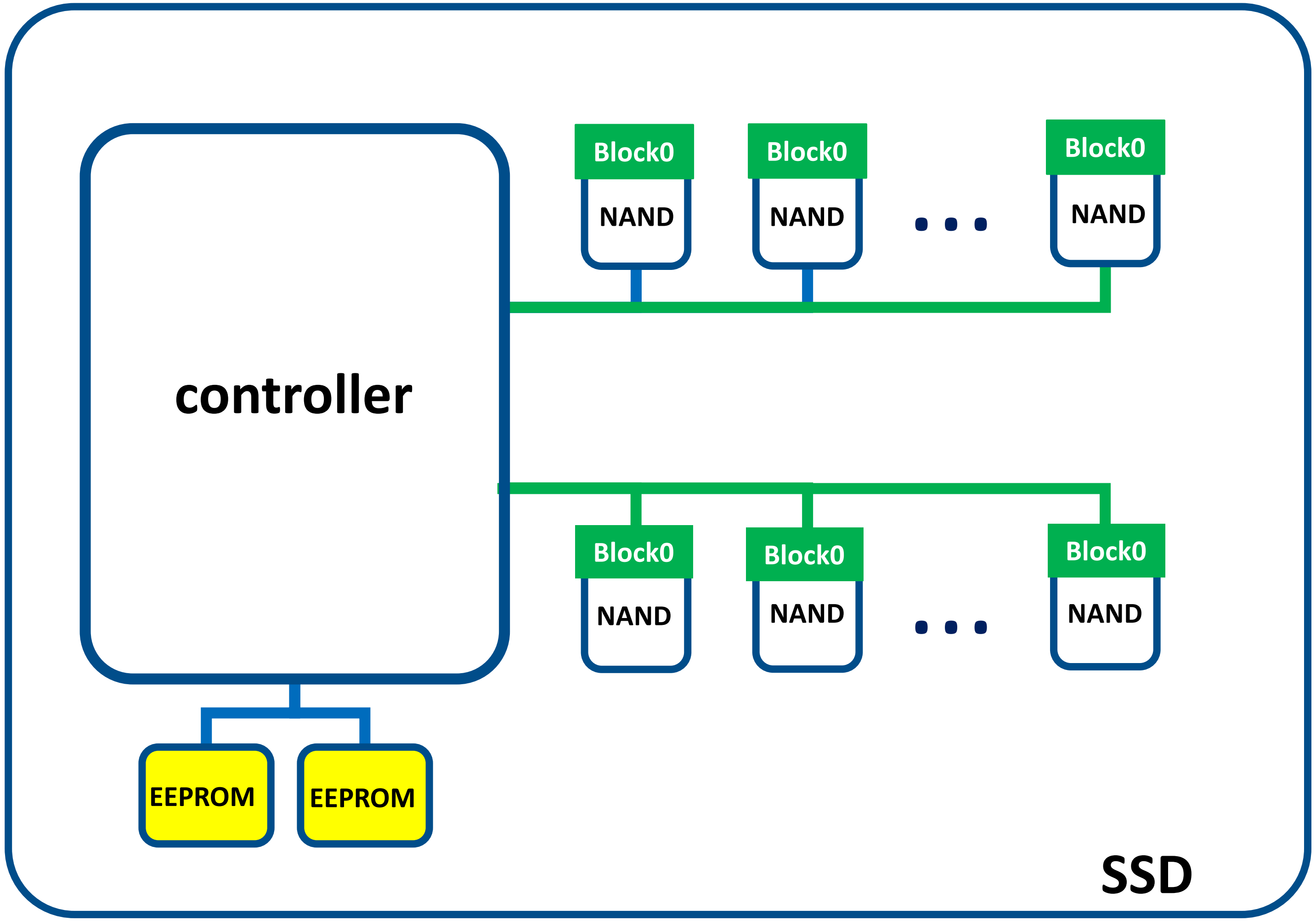
STORAGE DEVELOPER CONFERENCE
SDC 22

Current State – Initializing an SSD



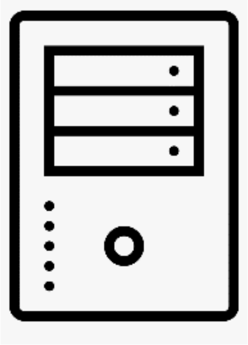
- Controller first reads information on E2PROM about HW configuration (type of NAND, timings, vendor ID, channel addressing, type of ECC used to load FW)
- Data read from E2PROM is protected by ECC to ensure reliability

Current State – Initializing an SSD

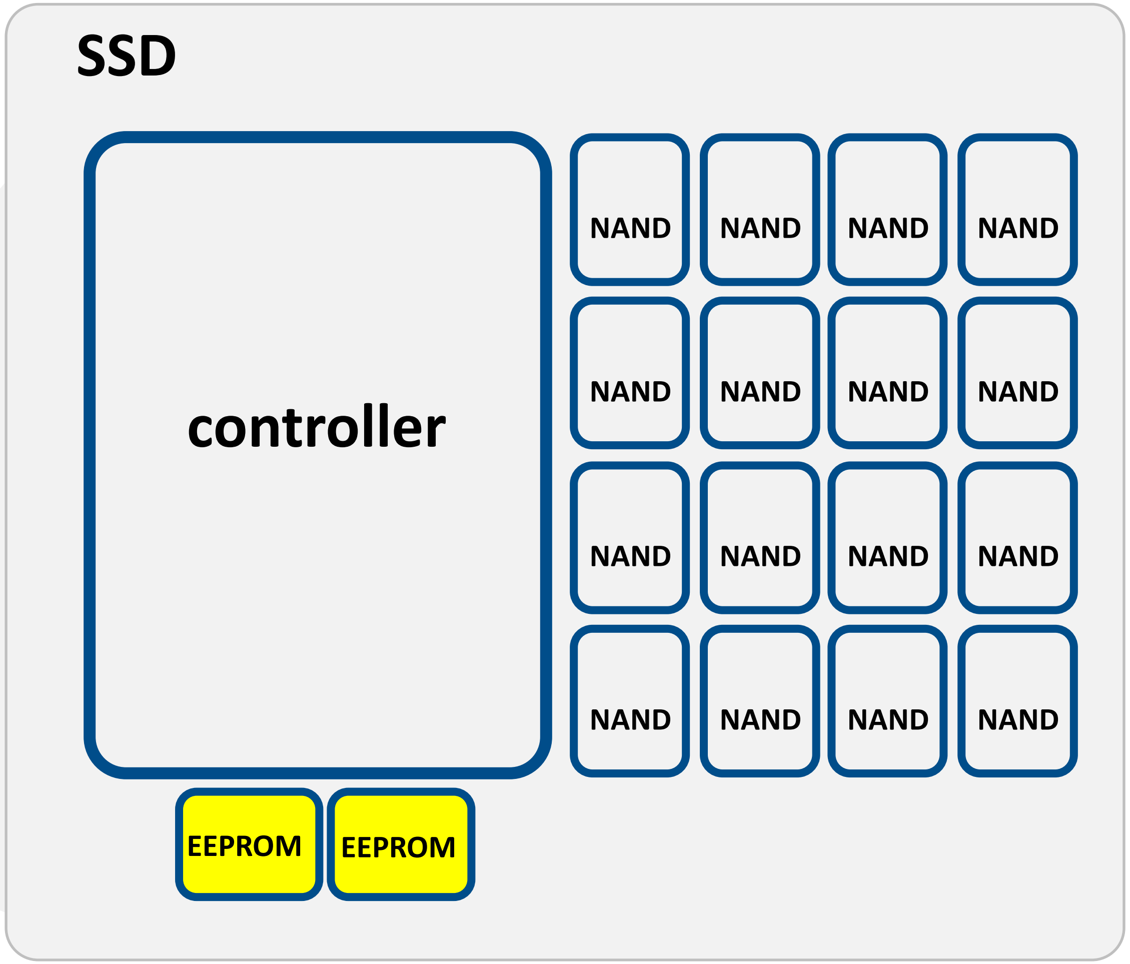


- Using previously read information, controller is able to read NANDs
- By reading block0 of NAND devices, controller loads the firmware
- Block0 is guaranteed good by NAND vendors for this purpose

Booting a Machine from an SSD



File System
Sector 0 (MBR)
Sector 1
…
Sector n



Booting a DNA Data Storage Archive

- Without a controller how can we read the archive?
- Where can we discover metadata such as vendor ID, CODEC used in the archive?
- This metadata is contained in the archive itself, but we need a way to discriminate it from other data

Rosetta Stone Project (1/3)

- Part of DNA Data Storage Alliance
- Goals:
  - Agree on a common identifier format for universally bootstrapping any DNA Archive
  - Enable identification of the CODEC used to encode an archive, from within the archive
  - Enable innovation in DNA CODECs for the main archive by enabling a standard for discovering the CODEC that was used
  - Provide fast access to archive metadata

DNA Archive Rosetta Stone (DARS)

Rosetta Stone Project (2/3)

- Working Assumptions
  - A generally-available specification document is accessible
  - Archive boot record is built using natural DNA bases (ACTG)…
  - …but the archive may contain non-natural DNA bases
  - Standard means of identifying the CODEC used within the archive is needed
  - We assume a reader will have some form of Internet connectivity
  - DNA will primarily be used as a write-once archival medium

Rosetta Stone Project (3/3)

- Decisions to Make
  - Agreement upon length of oligonucleotides
  - …error recovery metrics and mechanisms
  - …how many “sectors” are required
- Progress to Date
  - Initial proposals drafted and discussed
  - Covering sector zero implementation, identification
  - Outlining payload contents and their meaning
  - Discussions and tests around error modeling and recoverability
- Roadmap
  - Reviewing future proposals
  - Creation of and maintenance of a specification of a standard
  - Build policy and procedure documentation
  - External registry of CODECs

Future State

- Rosetta Stone sets the stage for controllers, drivers, and ecosystem
  - Agreeing on decoding standard enables vendors to work on consumers of sector zero
  - Standard controller functions for management (e.g. SMART) may come from our error models
- Address space governance
  - CODEC ID / address issuance may work similar to IP addresses
  - DNA Data Storage Alliance could operate similar to ICANN
- Technology will always evolve
  - CODECs, address space will form part of the Alliance's industry roadmap
  - Working assumptions based on current technology
    - Advances may lead to review of assumptions, error model, number of codecs etc.
  - Synergy with other SNIA storage technologies i.e. computational
    - Exposing novel CODEC capabilities enabled by the DNA medium

How to Participate

How to Participate

- Standards only succeed when they consider and support the needs of a broad base of constituents with an eye toward the future
- Our working group is growing and diverse, and looking to
  - Increase representation from both public and private sector
  - Increase representation from a variety of markets and domains

Subscribe to our newsletter on our website
https://dnastoragealliance.org

Follow us on Twitter
@DnaDataStorage

Follow us on LinkedIn
@dna-data-storage-alliance

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

STORAGE DEVELOPER CONFERENCE SDC 22

Summary

Summary

- DNA data storage media provides the promise of density, durability, and cost effectiveness to meet the challenges of data growth, retention, compliance, and climate change
- Writing data to DNA involves coding, synthesis, and storage, and conversely, reading data from DNA involves retrieval, sequencing, and decoding
- DNA as a storage media does not share properties found in other storage media types, e.g. no built-in controller, or linear addressing of physical storage regions
- Rosetta Stone aims to ensure that a DNA archive can be consumed in a consistent manner by making discoverable the structure and encoding of the archive

Thank You

Please take a moment to rate this session.

Your feedback is important to us.

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate