STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

Update on Standards for Consuming DNA Data Storage

Daniel Chadash

Co-Founder, DNA Data Storage Alliance

Joel Christener

Director and Distinguished Engineer, Dell

Helpful Links

- Preserving our Digital Legacy – an Introduction to DNA Data Storage
- Ballot result for sector zero, sector one specification proposals

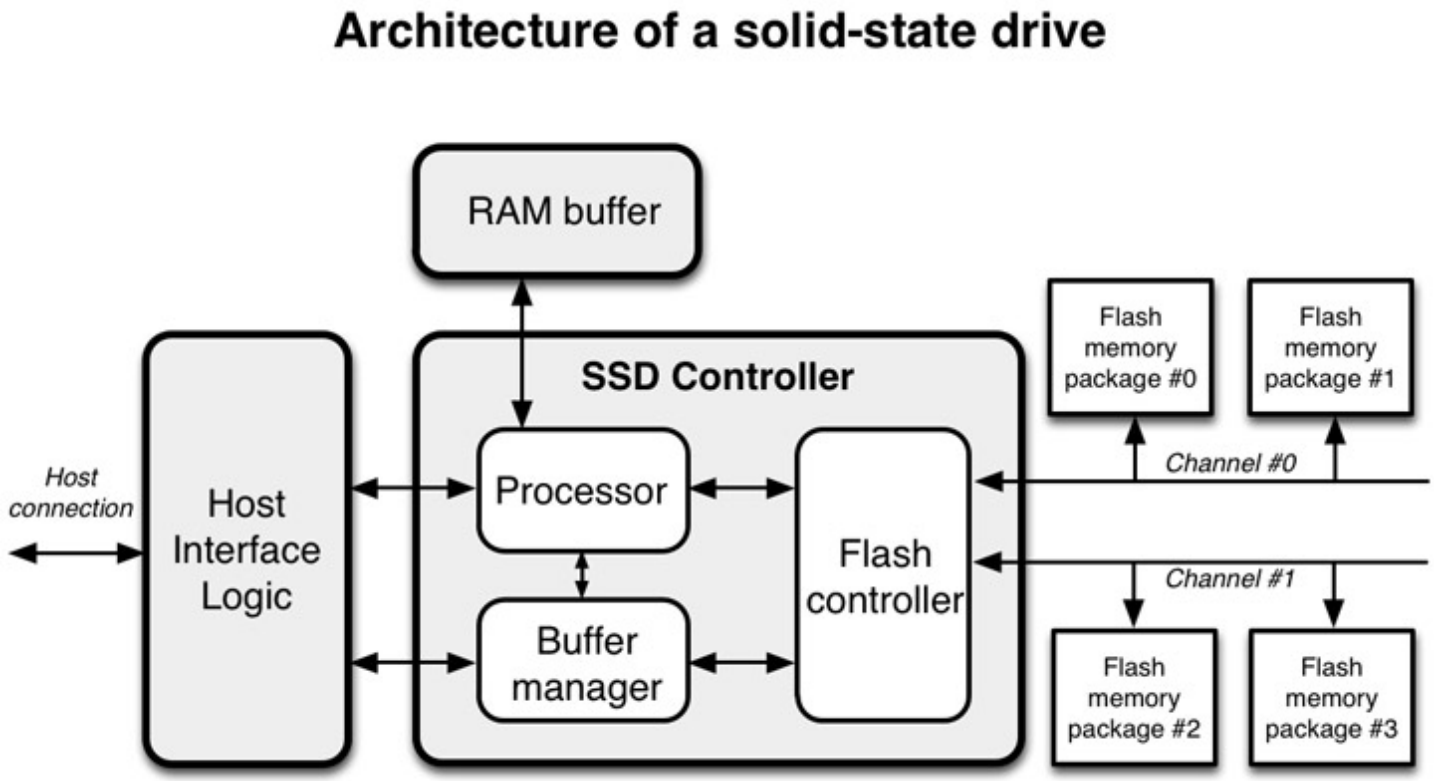
Agenda

- Differences: DNA vs Traditional Media
- Overview of the **D**NA **A**rchive **R**osetta **S**tone (DARS)
- Status, Details, and Standardization
- Summary

Differences: DNA vs Traditional Media

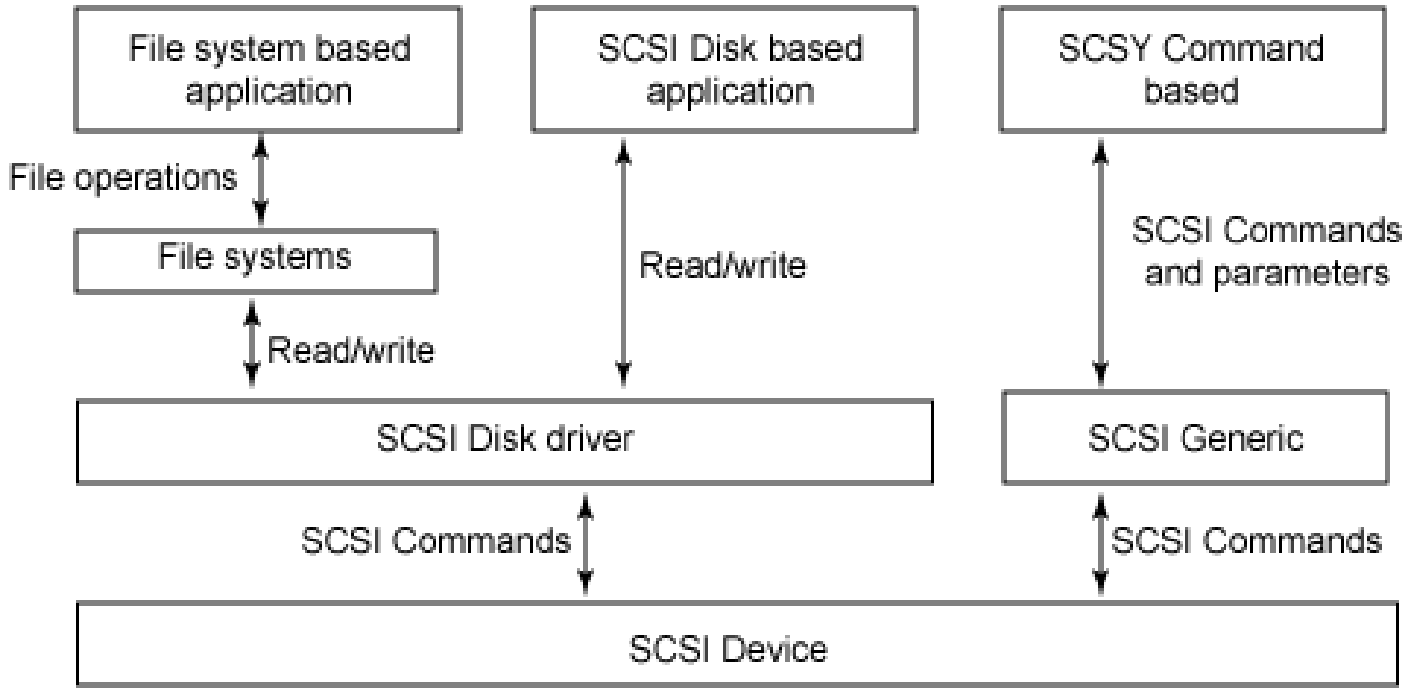
Differences: DNA vs Traditional Media

1. Exposing a device to the system



Differences: DNA vs Traditional Media

2. Organizing abstractions to create filesystem storage



Differences: DNA vs Traditional Media

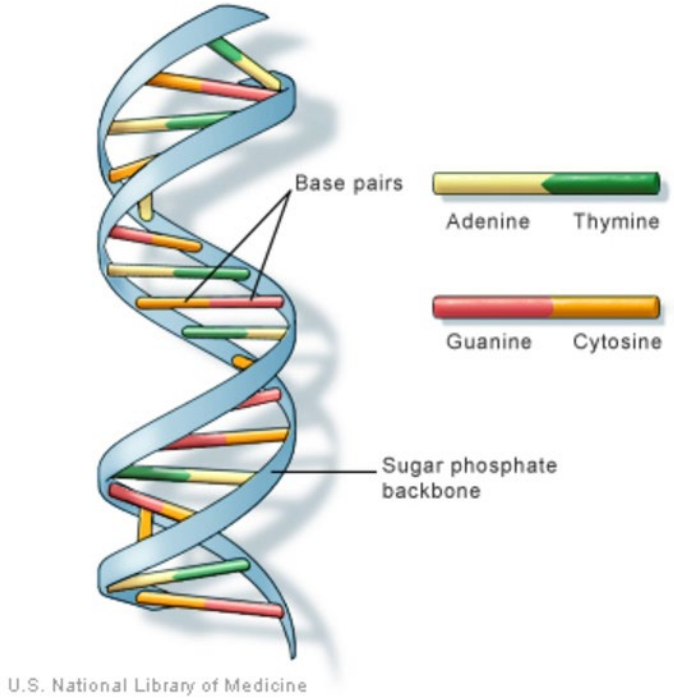
3. Media without integrated controller

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

Read LTFS from beginning of the tape

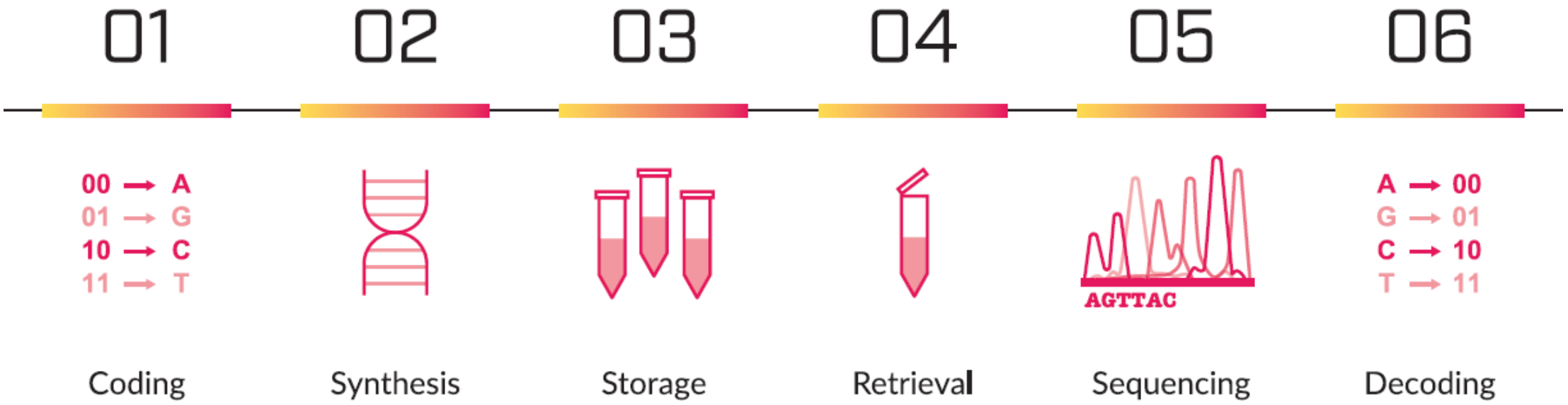
A “Primer” on DNA Data Storage Media

- The fundamental unit of storage in DNA is an oligonucleotide (also called ‘oligo’)
  - Definition: polymer containing a small number of subunits
  - Short, single strand of synthetic DNA or RNA
  - Sugar phosphate backbone
  - Base compounds ***A***denine, ***C***ytosine, ***T***hymine, ***G***uanine
  - Bases attach to the strand and to a mate on other strand
  - Adenine bonds w/ Thymine, Guanine bonds w/ Cytosine
- A double-stranded DNA molecule is a pair of single-stranded DNA molecules (oligos), tightly wound around one another, held together by the bonds between the bases



Oligo Example: ATTCGAGCGTTTTCGCGGTATAAGGAT

A “Primer” on DNA Data Storage Media



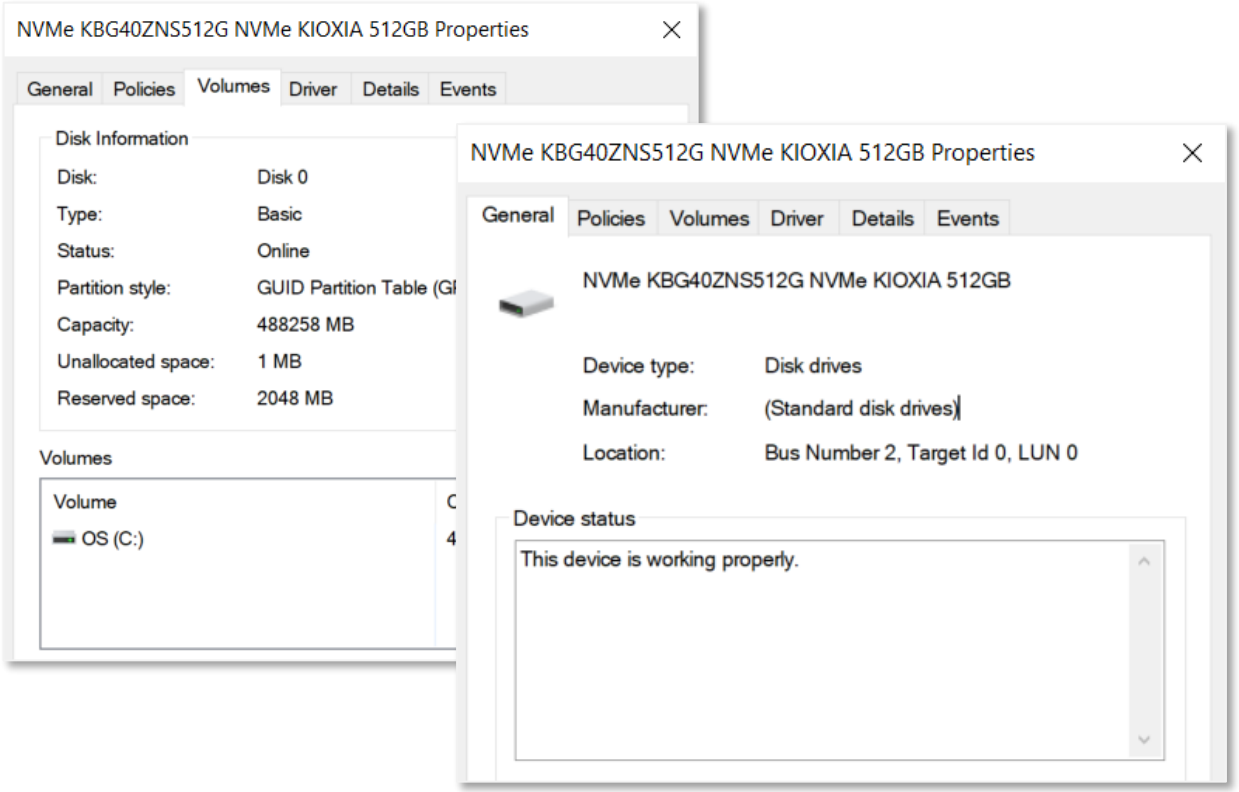
The problem

- DNA media does not share properties found in other storage media types
  - No built-in controller, or linear addressing of physical storage regions
  - Not built on a fixed substrate; not addressable memory, media is built as data is written
  - Addresses (sectors) need to be encoded into the oligos 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
  - Codecs are currently proprietary, as they are a competitive advantage unlike LTFS
- 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 happens if companies that wrote the DNA are gone by the time the data is accessed?

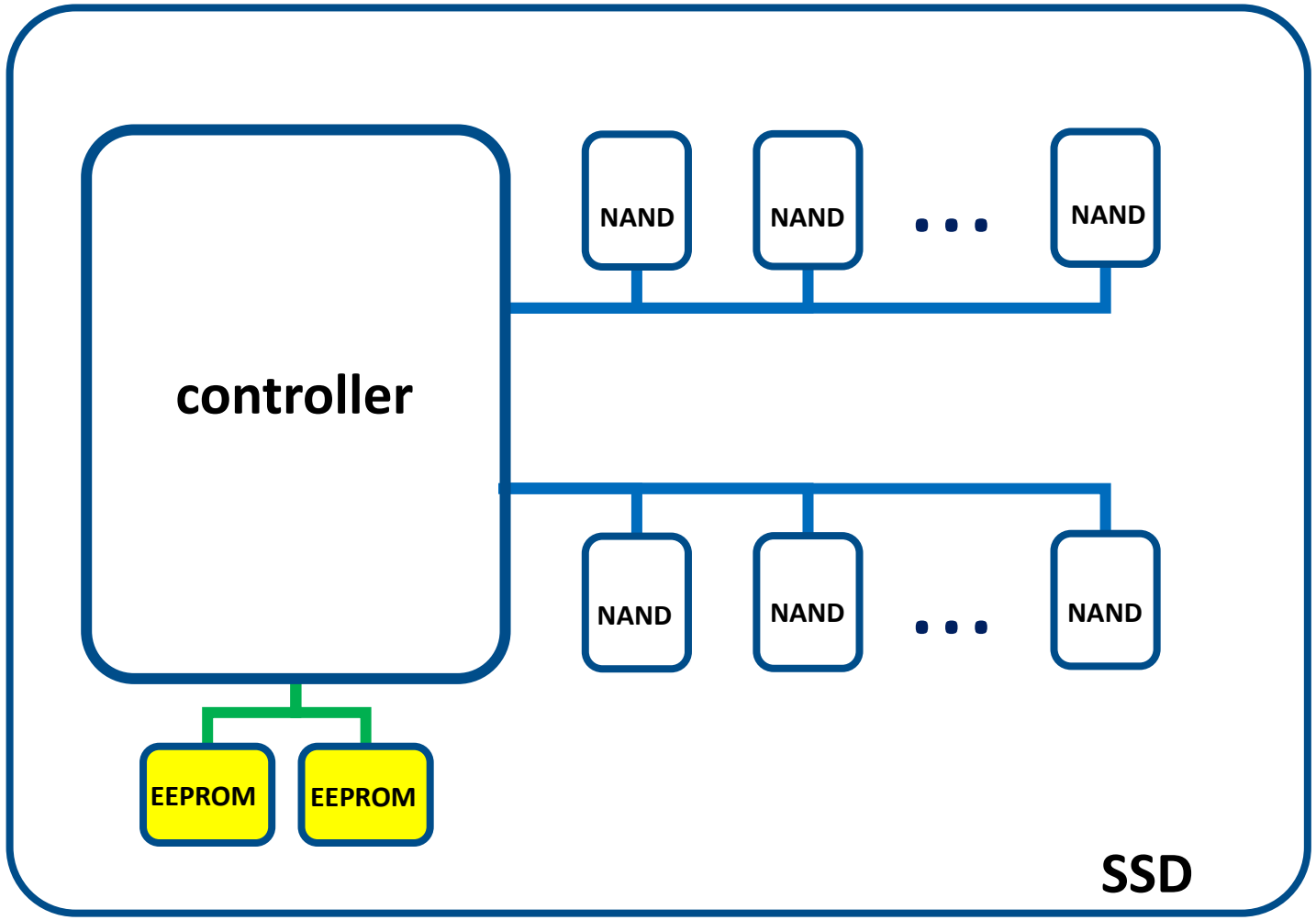
Overview of the DNA Archive Rosetta Stone (DARS)

The Goal: Produce an Archive Boot Sector

- 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

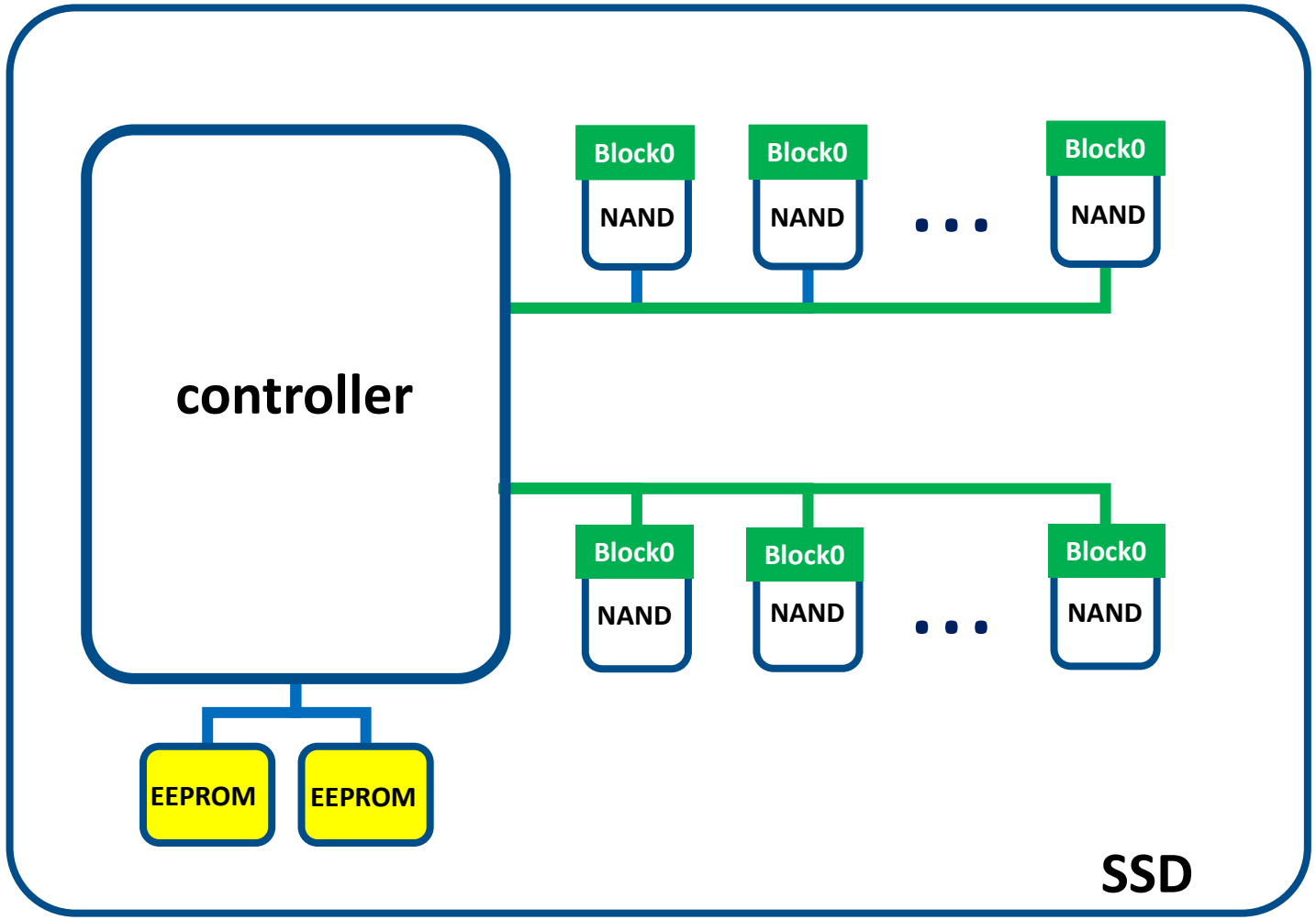


Current State – Initializing an SSD



- 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

Challenge: 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

DARS (1/2)

- 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)

DARS (2/2)

- 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

Status, Details, and Standardization

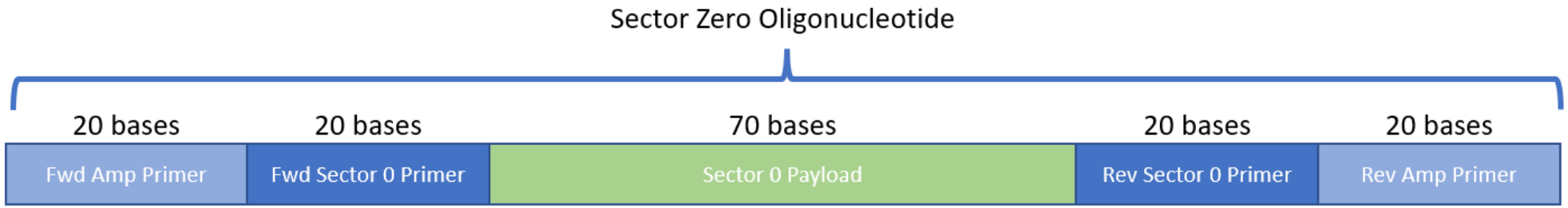
Sector Zero vs Sector One

"We can solve any problem by introducing an extra level of indirection"
~Wheeler

- The problem space: go from zero understanding of the archive to an understanding of how to consume the archive contents
- Subdivide it into two steps:
- Step one: create a mechanism wherein a small amount of data can be reliably retrieved and well-understood (sector zero, e.g. discern how to access the archive logical structure and metadata)
- Step two: create a mechanism wherein a larger amount of metadata can be reliably retrieved and consumed (e.g. the logical structure and metadata)

Sector Zero (1/3)

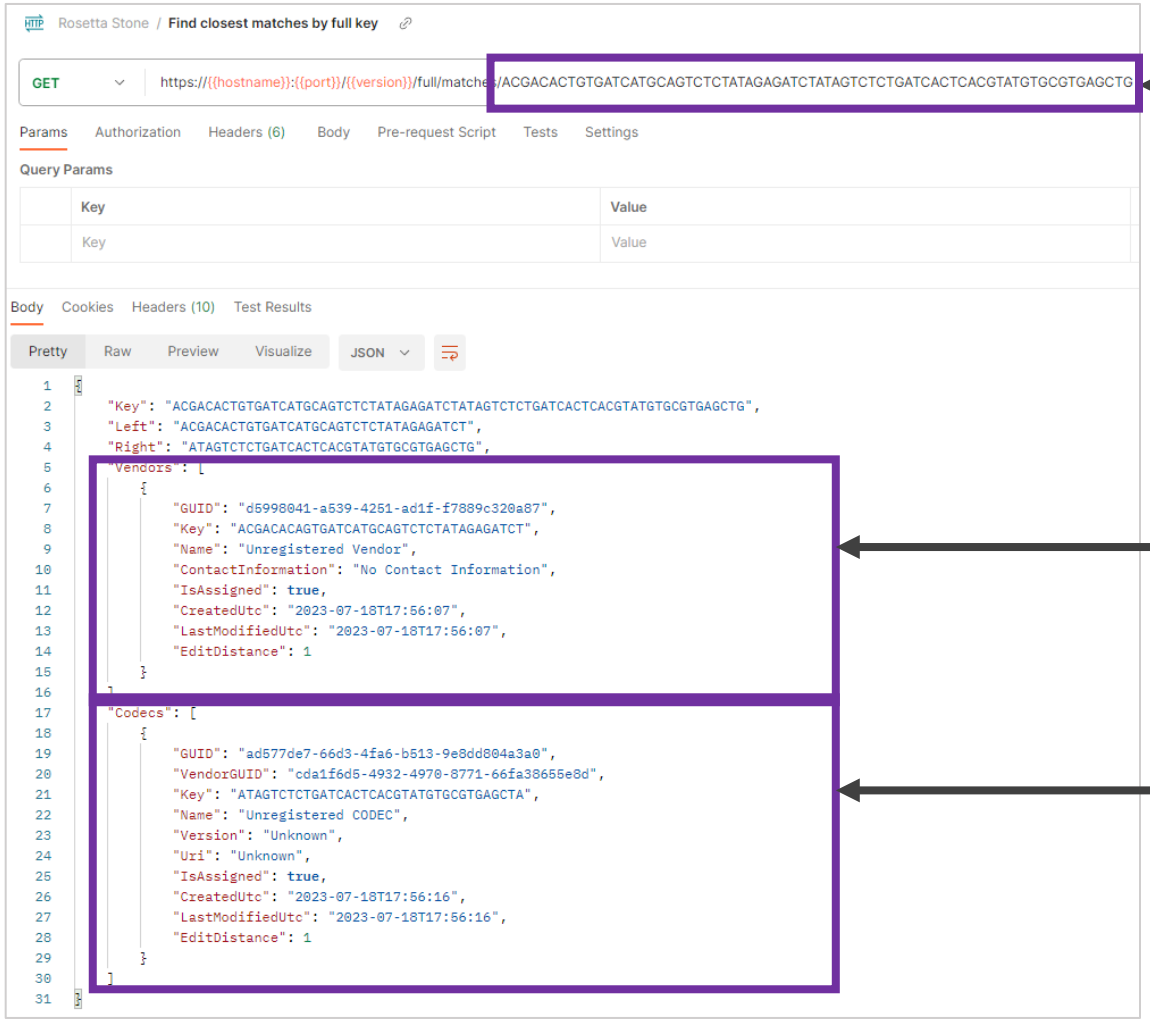
- The goal and intent of sector zero is to enable those with no external metadata about the archive to:
  - Retrieve a key identifying the archive writer
  - Retrieve a key identifying the codec used to write sector one
- Sector zero fits into a single oligonucleotide and can be amplified from an archive using an alliance-defined set of primers
- Sector 0 is not “encoded” = no codec is needed to read it



Sector Zero (2/3)

- The 70-base payload is then split into two 35-base strings; the left-most representing the vendor and the right-most representing the codec used for sector one
- These values can then be passed into an API service provided by SNIA and the DNA Data Storage Alliance to determine:
  - Which vendor wrote the archive
  - Which codec was used to write sector one
- In the case of errors (insert/delete/replace) the nearest records by key can be retrieved, along with their edit distance (Levenshtein)

Sector Zero (3/3)



Sector zero payload

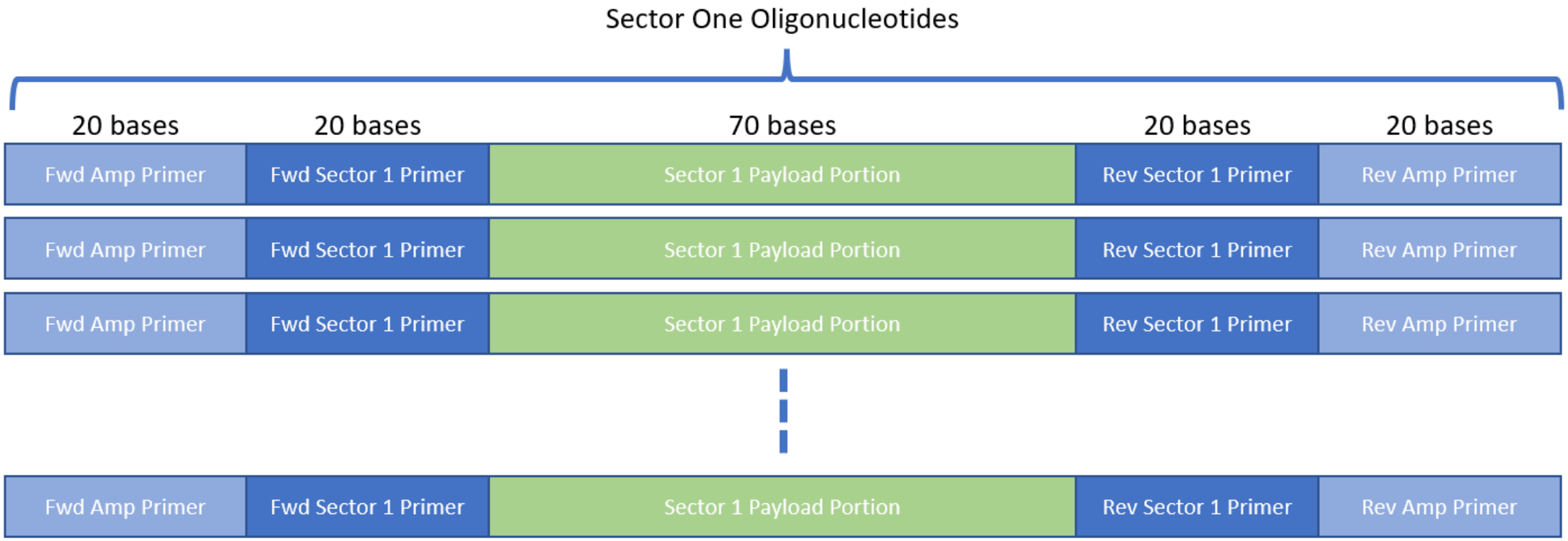
Closest vendor match(es)

Closest CODEC match(es)

Sector One (1/3)

- The goal and intent of sector one is to enable the archive reader to
  - Understand the general logical structure
  - Get a clue about the content
  - Understand the parameters needed to read the archive's contents
- Sector one contains a significant amount of metadata and uses JSON as its representation
- Due to its size (potentially thousands of bytes) sector one spans multiple oligos and requires a codec (the codec addresses identification of oligos)
- Contents may be accessible outside the archive (e.g. barcode, QR code, NFC) to mitigate the need to sequence

Sector One (2/3)



Sector One (3/3)

- Once read, the data is decoded to a UTF8 string and deserialized from JSON into an object or dictionary
- The object contains:
  - Description of archive contents
  - Hashing and non-repudiation
  - Details for the sequencer
  - Details of CODEC used for data
  - Timestamp
  - Details about the archive writer
  - Optional fields
  - Additional parameters

Current Status

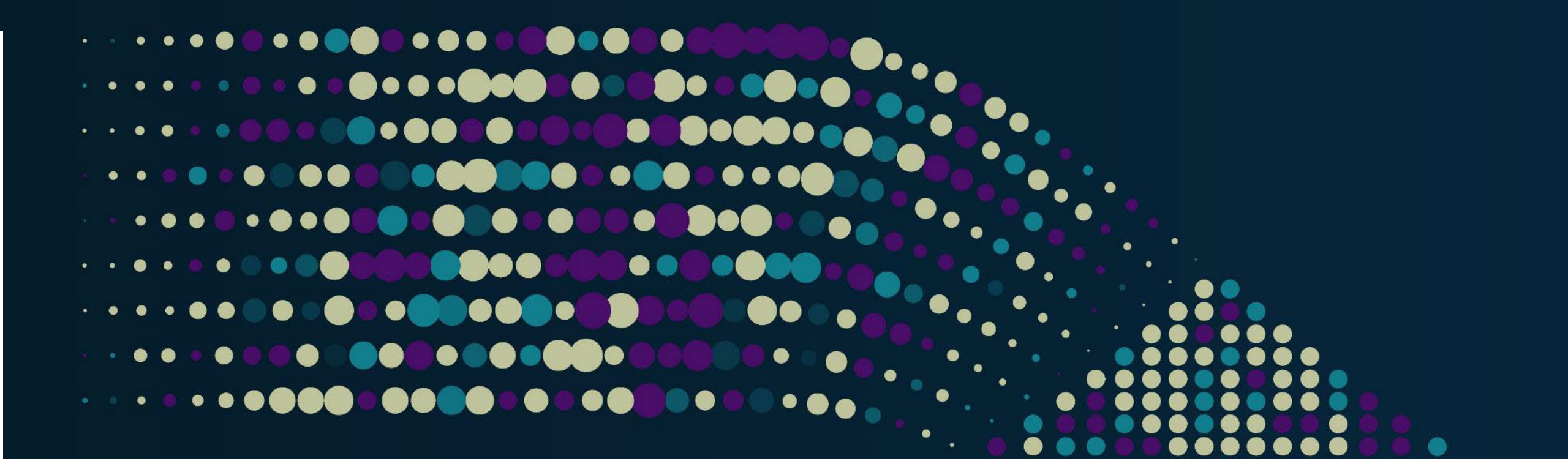
- Both sector zero and sector one have been approved by the DNA TA Governing Board and technical working group
- We are currently in the IP Review stage and about to publish the specs by EOY

Summary

Summary

- DNA as a storage media presents unique challenges that have not yet been faced in other storage media types
- The goal and intent of the **D**NA **A**rchive **R**osetta **S**tone initiative is to enable an archive reader to go from zero understanding of an archive to a logical understanding of how to consume the archive's contents
- Sector zero is responsible for identifying who wrote the archive and what codec was used for writing sector one
- Sector one is responsible for providing codec, sequencing, and other details necessary to consume the data within the archive
- Both proposed specifications have been approved for ratification by the DNA technical working group and we are awaiting the rest of the standardization process

Thank you!



Please take a moment to rate this session.

Your feedback is important to us.

