DNA-based Data Storage

Devin Leake
VP of R&D

Gen9
There is a growing need for storage

- By 2020, 40 zettabytes will be generated
- Predominant form of storage is tape

Content created by DOMO, “Data Never Sleeps”

IDC (International Data Corporation)

https://mozy.com/infographics/where-is-the-worlds-data-stored
Facebook Facility in Fort Worth

750,000 sq.ft.
$1,000,000,000
10-15 year life span
1 Exabyte Storage
DNA-based Data Storage

• DNA has been recovered and analyzed from:
  o Wooly mammoth (20,000 years ago)
  o Neanderthal (40,000 years ago)
  o Bison (60,000 years ago)
  o Pollen samples (~500,000 years ago)

• DNA Storage Process:
  o Data IN: Synthesize DNA that contain user information
  o Data OUT: Sequencing is performed via high-throughput sequencing technologies (e.g. Illumina, PacBio, Oxford Oxford Nanopore Technologies)

Bornholt, et.al. ASPLOS 2016
Proof-of-concept for DNA-based Storage

- Simple encoding scheme (substitution cipher) used for initial experiments
- Microdot storage similar to prior ‘espionage’ formats
- Indexing and access primers flank DNA message

Clelland, Risca and Bancroft Nature 1999
Evolution of Encoding Architecture

Goldman, et.al. Nature 2013
Evolution of Encoding Architecture

Encoding architecture maximizes content density and redundancy while avoiding biophysical features (e.g. homopolymers) that impair storage performance.

Bornholt, et.al. ASPLOS 2016
Volumetric Density of DNA

<table>
<thead>
<tr>
<th></th>
<th>Volumetric Density (Bits/cm^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard Disk</td>
<td>$10^{13}$</td>
</tr>
<tr>
<td>Flash Memory</td>
<td>$10^{16}$</td>
</tr>
<tr>
<td>DRAM</td>
<td>$10^{13}$</td>
</tr>
<tr>
<td>Cellular DNA</td>
<td>$10^{19}$</td>
</tr>
</tbody>
</table>

DNA provides several orders of magnitude higher density than traditional storage media

Determining Optimal Length of DNA

With longer DNA, sequences for addressing and indexing ‘payload’ are a smaller percentage of overall length.

Bornholt, et.al. ASPLOS 2016
Types of Data Storage Based on Need

**Active**
- Uninterrupted availability
- High performance read/write access

**Backup**
- High media capacity
- High performance read/write access
- Low storage cost per GB

**Archive**
- Data authenticity – data is protected from modification
- Extended media longevity/ durability
- **High performance random read access**
- Extremely low storage cost per GB
Coding for Random Access Scheme

6 words = 6 × 21 bits = 126 bits

binary to integer

\[ x = 123 \ldots 666 \]

Encode_{a_1, 80}(x)

GATAGT...CGTCAT ... GACAGT...TGTAAT

80 bases 80 bases

12 sub-blocks = 12 × 80 bases = 960 bases

20 + 960 + 20 = 1,000 bases
Active Archive Model

- Method of tiered storage using multiple systems
  - Multiple media employed for optimal storage (Return on Byte)
  - Metadata used to organize stored data
- Model facilitates necessary deployment of support systems, such as organizational AI

Reducing Cost of DNA-based Storage

- Synthesis is currently the primary driver for price (2 orders of magnitude more than ‘reading’)
- Improving synthesis efficiency regain ~1 order of magnitude

Goldman, et.al. Nature 2013

Bornholt, et.al. ASPLOS 2016
The Synthetic Biology Solution

Using **semiconductor** technology, Gen9 developed a high-throughput process to build **DNA** at a very high quality and low cost.
Thank You!