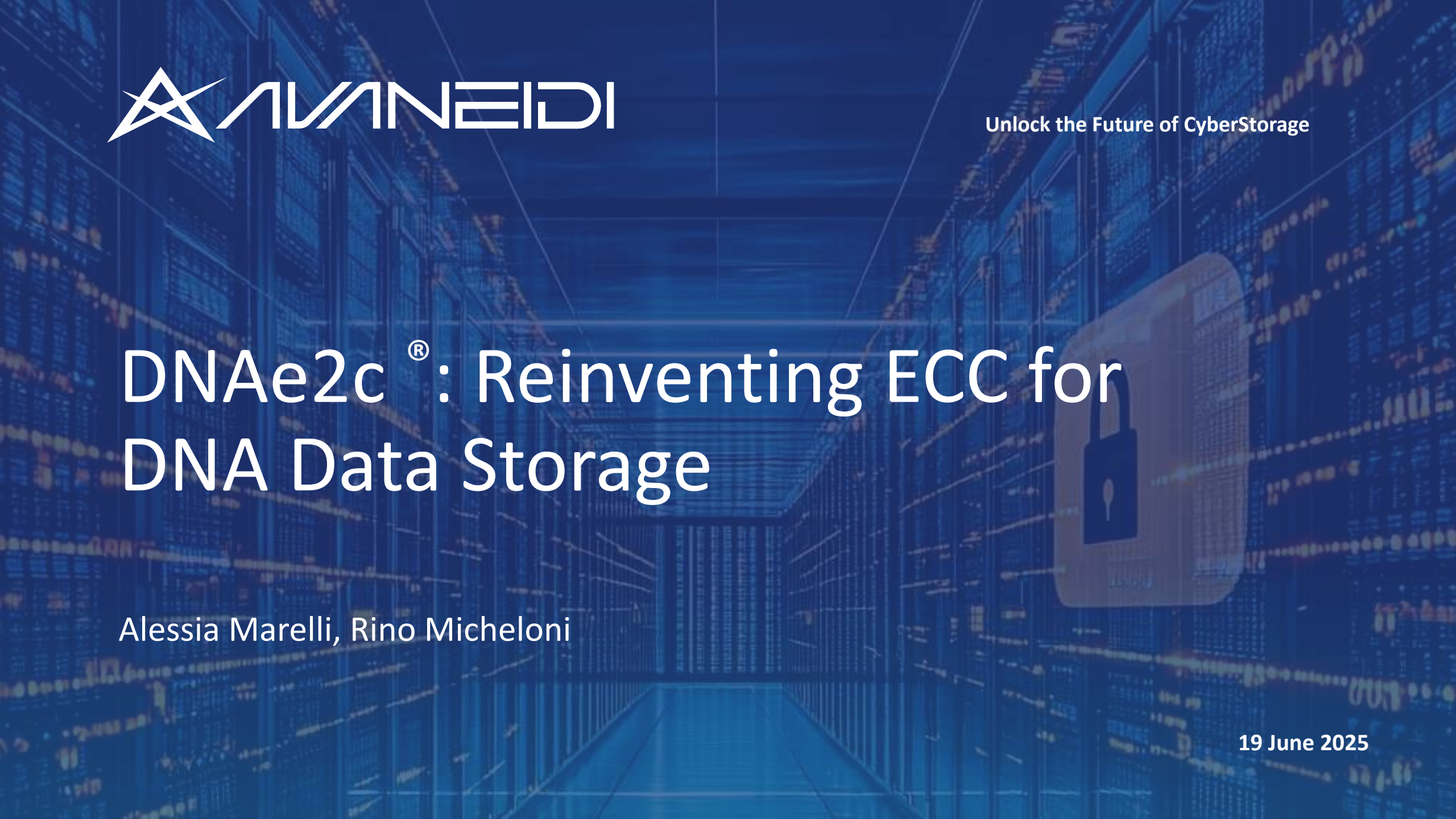


AVANEIDI

Unlock the Future of CyberStorage

DNAe2c ®: Reinventing ECC for DNA Data Storage

Alessia Marelli, Rino Micheloni

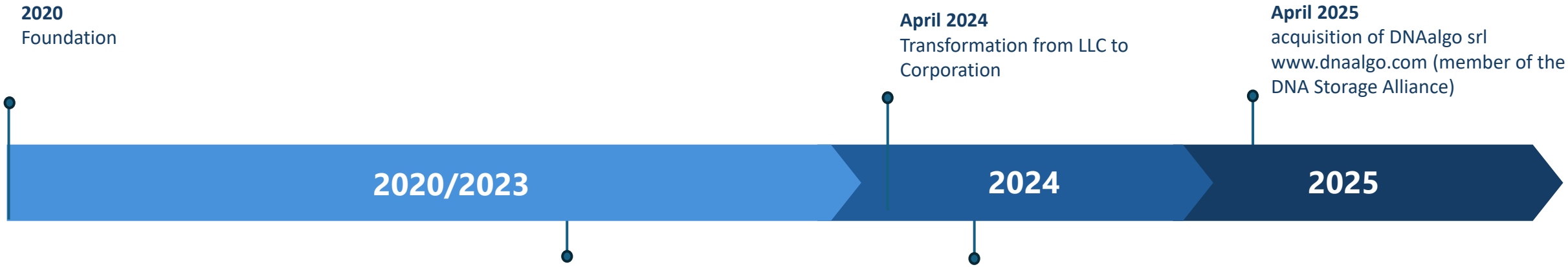
19 June 2025

AGENDA

- Who we are
- Why DNA storage
- Synthetic DNA pipeline
- Why ECC?
- Tailored DNAe2c
- Conclusions

AVANEIDI

WHO WE ARE



- First CyberStorage Company in Europe

WHO WE ARE

Data Repatriation - Main driving forces

Safe-AI

4C-model

Data Security

geopolitics

On-

Edge Data Center

Cloud issues: 4C model (Gartner)

- Cost
- Compliance
- Complexity
- Lack of Control

“Safe-AI refers to the development and deployment of AI systems that are resilient to cyber threats, adversarial attacks, and data breaches.
It ensures that AI models operate securely, protect sensitive data, and prevent malicious manipulation”.

NEW CHALLENGES

A Holistic Approach to safeguarding Data

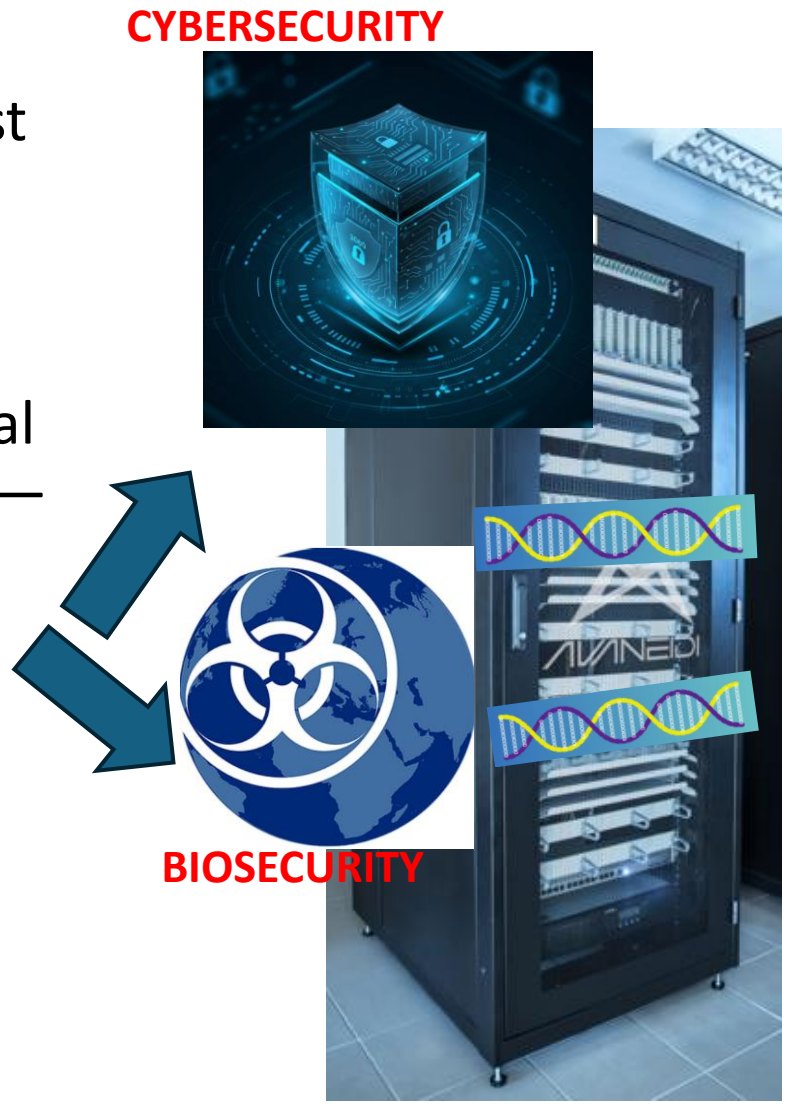
CyberStorage by Avaneidi is a scalable & reliable storage infrastructure that seamlessly combines a tailored Secure Hardware, your-own Data Governance and an advanced CyberShield into one unified platform

- **Advanced CyberShield**
- **Your-Own Data Governance**
- **Tailored Secure Hardware**

All **Designed** & **Manufactured** by Avanedi in **Italy**

CYBERSTORAGE BY AVANEIDI

- Today, CyberStorage systems already provide robust data security through advanced electronic technologies and sophisticated software/firmware.
- DNA storage extends conventional data security paradigms by integrating electronic and biochemical processes—such as DNA synthesis and sequencing—necessitating a dual-layered cybersecurity framework.
- This includes protecting digital encoding and retrieval systems, as well as securing biological materials against unauthorized synthesis, sequencing, or tampering.



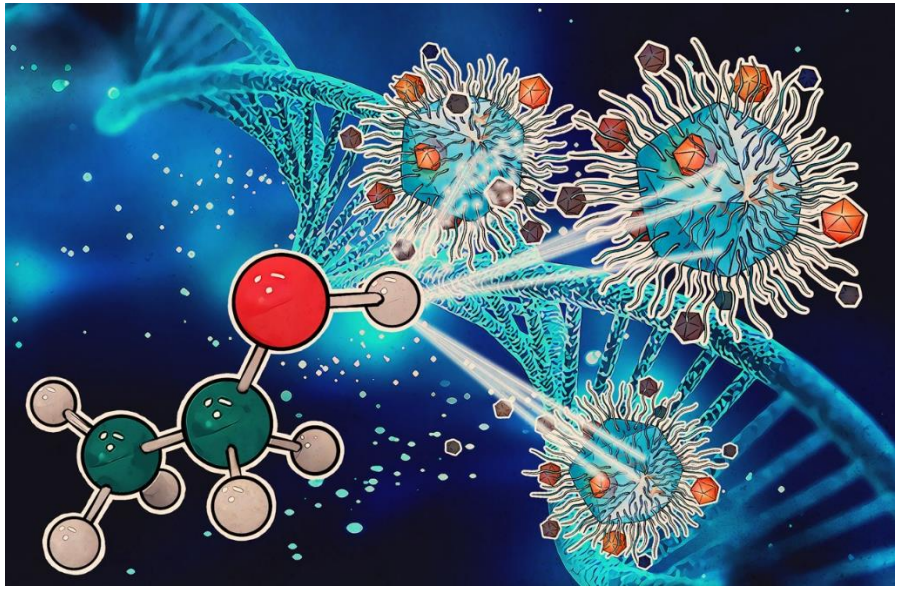
DNA

WHY DNA DATA STORAGE

1. **Extreme Storage Density**: DNA can store enormous amounts of data in extremely small spaces. One gram of DNA can theoretically hold 215 petabytes (215 million GB).
2. **Long-Term Durability**: If stored properly (in cool, dry conditions), DNA can last thousands of years, unlike hard drives, magnetic tapes, or SSDs, which degrade within decades.
3. **Sustainability and Low Environmental Impact**: Once synthesized, DNA requires no energy to be preserved. It has a potentially lower ecological footprint compared to traditional data centers.
4. **Long-Term Stability**: DNA is resistant to corruption from radiation, magnetic fields, or power fluctuations, ensuring the data remains intact over time.
5. **Universality and Future-Proof Format**: DNA is the universal language of life. Technologies to read it will likely remain available in the future, making DNA a future-proof data format.

BENEFIT

1. **High Costs:** Currently, DNA synthesis and sequencing are expensive. For example, storing one gigabyte of data in DNA can cost around $3,500, while the same on a traditional hard drive costs less than $0.03.
2. **Slow Read and Write Speeds:** Encoding and decoding data in DNA is significantly slower compared to traditional methods, making this technology more suitable for long-term archival storage than for quick data access.
3. **Error Rates:** DNA synthesis and sequencing can introduce errors such as substitutions, insertions, or deletions. While error-correcting codes have been developed, handling these errors remains a technical challenge.
4. **Limited Accessibility and Infrastructure:** Reading and writing data to DNA requires specialized equipment and advanced technical expertise, limiting accessibility and increasing operational complexity.
5. **Ethical and Regulatory Concerns:** The use of synthetic DNA for data storage raises ethical issues, particularly around privacy and the potential misuse of genetic information. Additionally, it is subject to complex legal and regulatory frameworks.

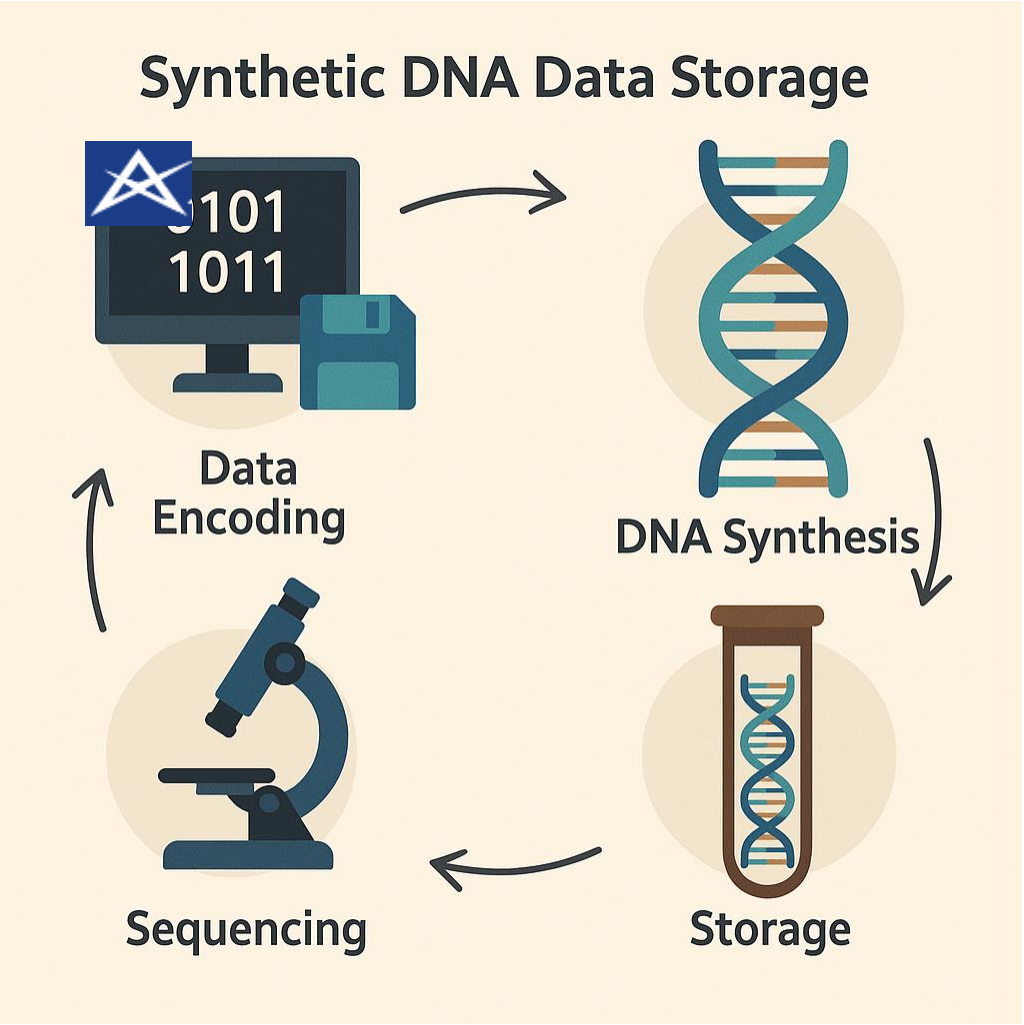


CHALLENGES

DNA DATA STORAGE PIPELINE

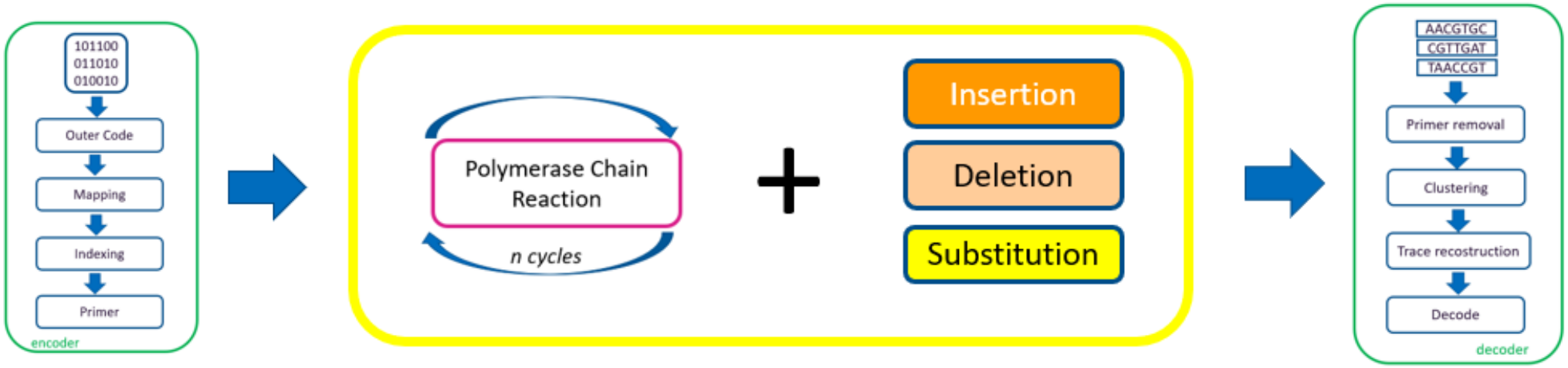
Avaneidi IPs works on the storage side so on the encoding and decoding process. In order to build that, we need these 3 pillars:

- **1. Noise Modeling:** We create mathematical models to simulate the types of errors that occur during DNA synthesis and sequencing.
  ✅ **Why it matters:** It saves time and cost by replacing long lab experiments with software simulations.
- **2. Encoding & Decoding IPs:** We develop smart algorithms to **encode digital data into DNA** and then **decode it back** accurately.
  ✅ We use advanced simulations (DNAssim®) to optimize these steps for speed, reliability, and efficiency.
- **3. DNA Simulation Platform – DNAssim® :** We built DNAssim, a **custom simulator** for DNA-based storage systems.
  ✅ It helps test how encoding and decoding perform under real-world error conditions, enabling better system design.



WHAT IS SYNTHETIC DNA

- While encoding and decoding processes can be defined using equations, the occurrence of errors is inherently non-deterministic and requires modeling.
- To achieve optimal performance, encoding and decoding should be customized based on a specific noise model.
- Due to the statistical nature of noise, a simulator is essential to evaluate the effectiveness of different encoding and decoding strategies.



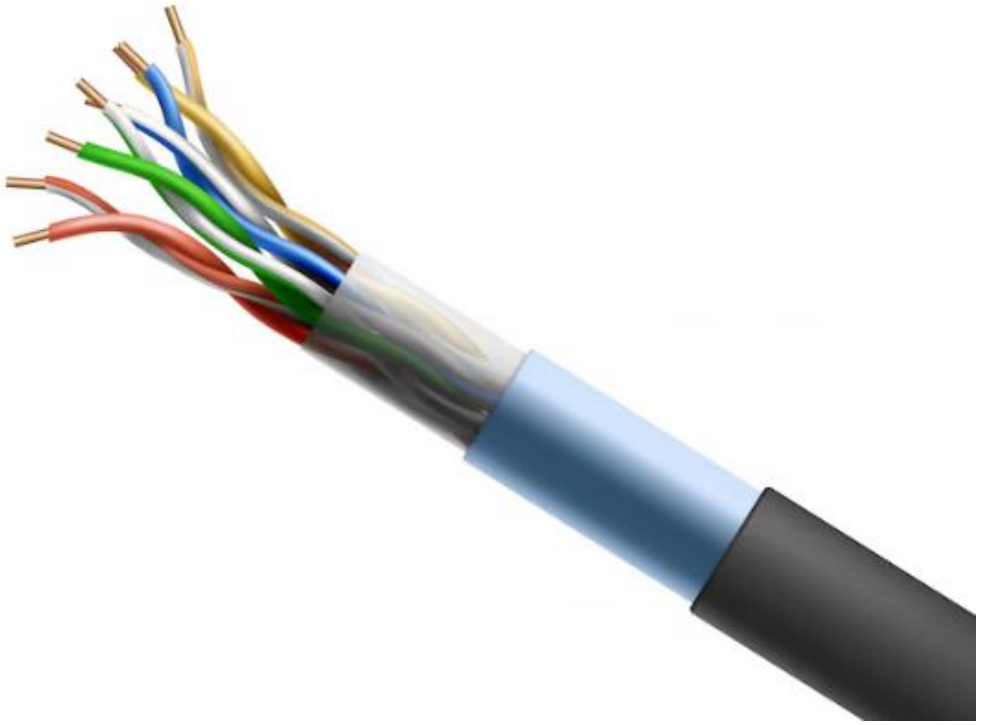
DNASSIM

WHY ECC?

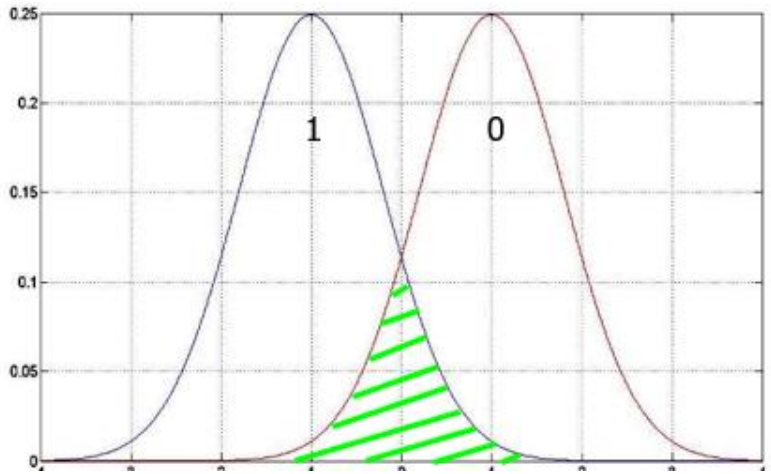
Internet & phone cables

- In the late '80s, home phones were widespread, and the internet was starting to enter households.
  How could digital data be delivered without changing the existing infrastructure?

- **Through the use of ECC, high speed communications over a medium not created originally for this purpose, was made possible.**



ENABLING A TECHNOLOGY



NAND Flash

- In the early 2000s, NOR Flash dominated the market for its reliability, but NAND Flash emerged as a faster, more scalable alternative despite its structural limitations.
  NAND enabled storing multiple levels (4, 8, 16...) within the same voltage range, driving down costs and enabling widespread use in devices like SSDs.
- Errors occur in the overlapping regions between distributions, which worsen over time due to usage and retention effects.
  In devices with 16 or 32 levels, overlaps exist even when new, with error rates around 10^{-2}.
  So how can such an error-prone medium be used in enterprise-grade applications that demand error rates near 10^{-14}?
- **Another example of where, through the use of ECC, a very poor media is able to be used for an application requiring higher performance than the media can provide on its own**

ENABLING A TECHNOLOGY

The mission of ECC in DNA storage..

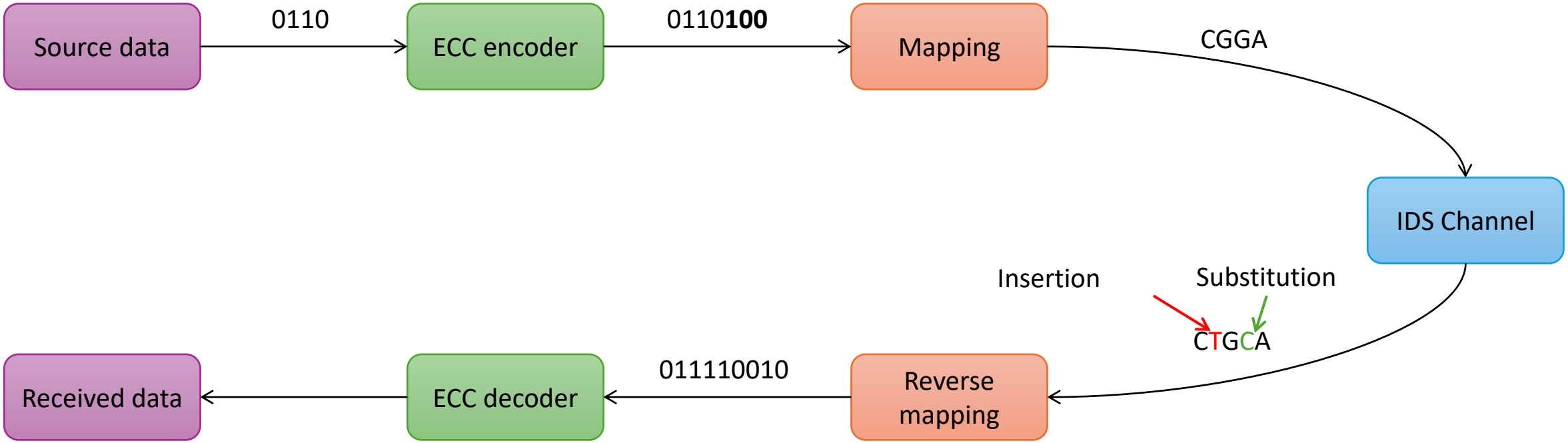
- DNA is a stable medium with strong data retention, making it suitable for storage.
- The main challenge lies in the synthesis (writing) and sequencing (reading) steps, which introduce errors.
- To minimize this noise, current methods rely on costly and time-consuming processes.
- A more error-tolerant storage system could allow for less precise synthesis/sequencing, reducing both time and cost.

- **Maybe... Through the use of a strong ECC approach, poor but lower cost, and faster writing and reading processes can be used for Enterprise Grade storage applications**

ENABLING A TECHNOLOGY

DNAe2c®

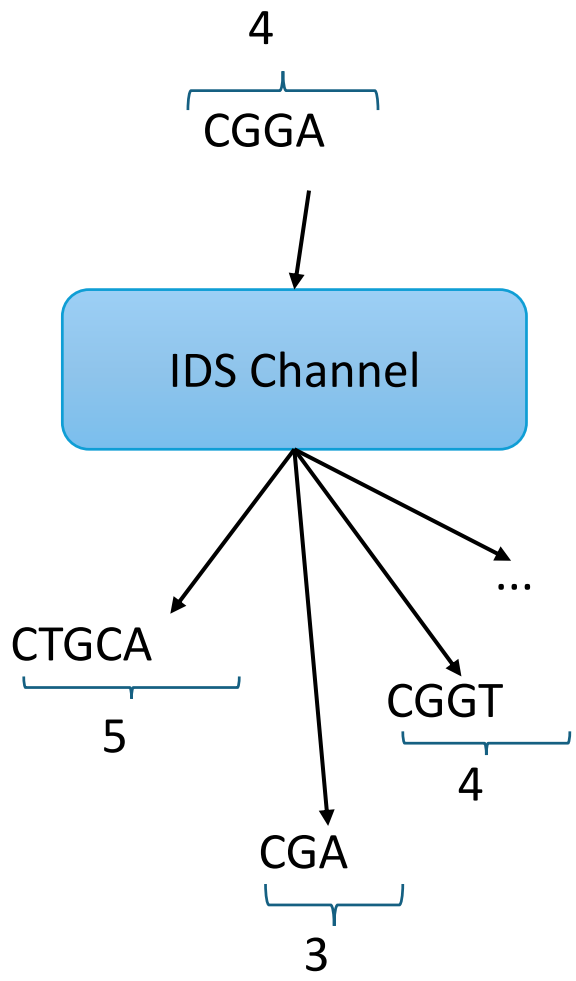
Basic DNA storage system with Hamming (7, 4) code



We cannot calculate the syndrome, because the length of the codeword (9) is not equal to the number of columns of the parity-check matrix (7), so matrix multiplication is not possible!

BLOCK CODES

- A key issue with IDS channels and block codes is that they can alter the length of DNA strands.
- When a strand has the wrong length, errors can still be detected, but correction isn't possible.
- As a result, these strands must be discarded, leading to data loss.
- While block codes are powerful, the scheme shown in the previous slide needs additional steps to reach acceptable performance levels
- In addition to that block codes are very powerful with Hamming distance while in the DNA channel we have to deal with Levenshtein distance



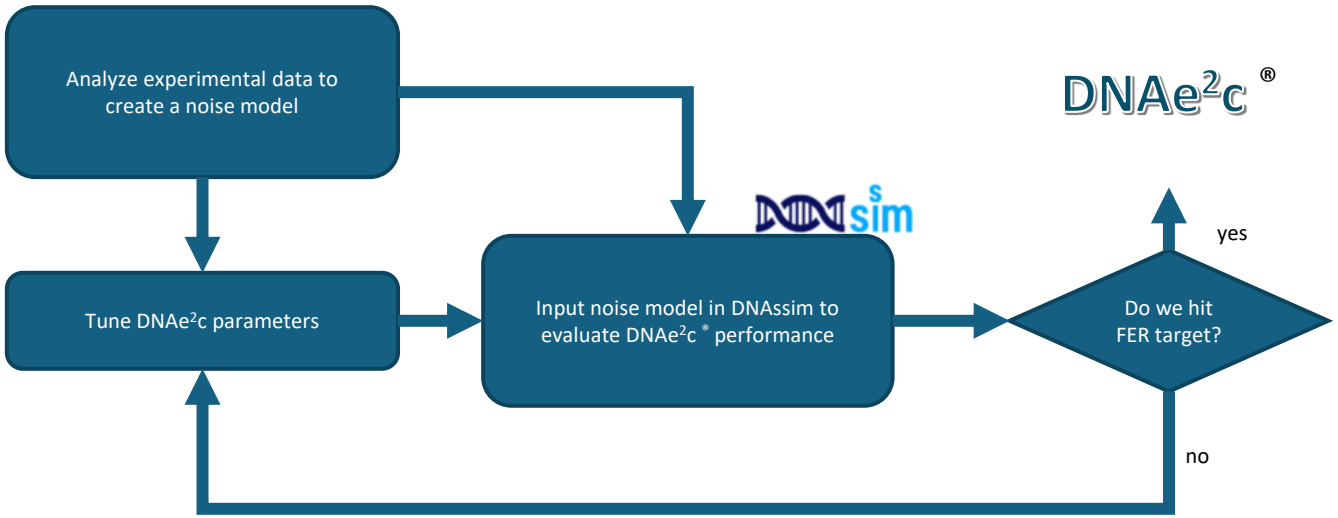
BLOCK CODES AND IDS CHANNEL

- **Error sources vary significantly based on the synthesizing and sequencing platforms**
  → We need an ECC scheme with tunable parameters to adapt to specific error characteristics of each channel.
- **To evaluate performance, extensive simulations are required**
  → We must test known codes across a wide range of conditions (e.g., Substitution Error Rate, erasures, PCR noise, etc.).
  → This involves implementing various ECCs in DNAssim® and running large-scale simulations, including SW/HW co-simulations.
- **Maintain a high code rate (CR) to optimize efficiency**
  → Minimizing written data helps reduce both computational complexity and power consumption.
- **Final solution must be hardware-implementable**
  → The chosen ECC and decoding architecture should be suitable for efficient hardware deployment.

DNAe²c®

NA noise aware errors & erasures leaner

NEW CODE

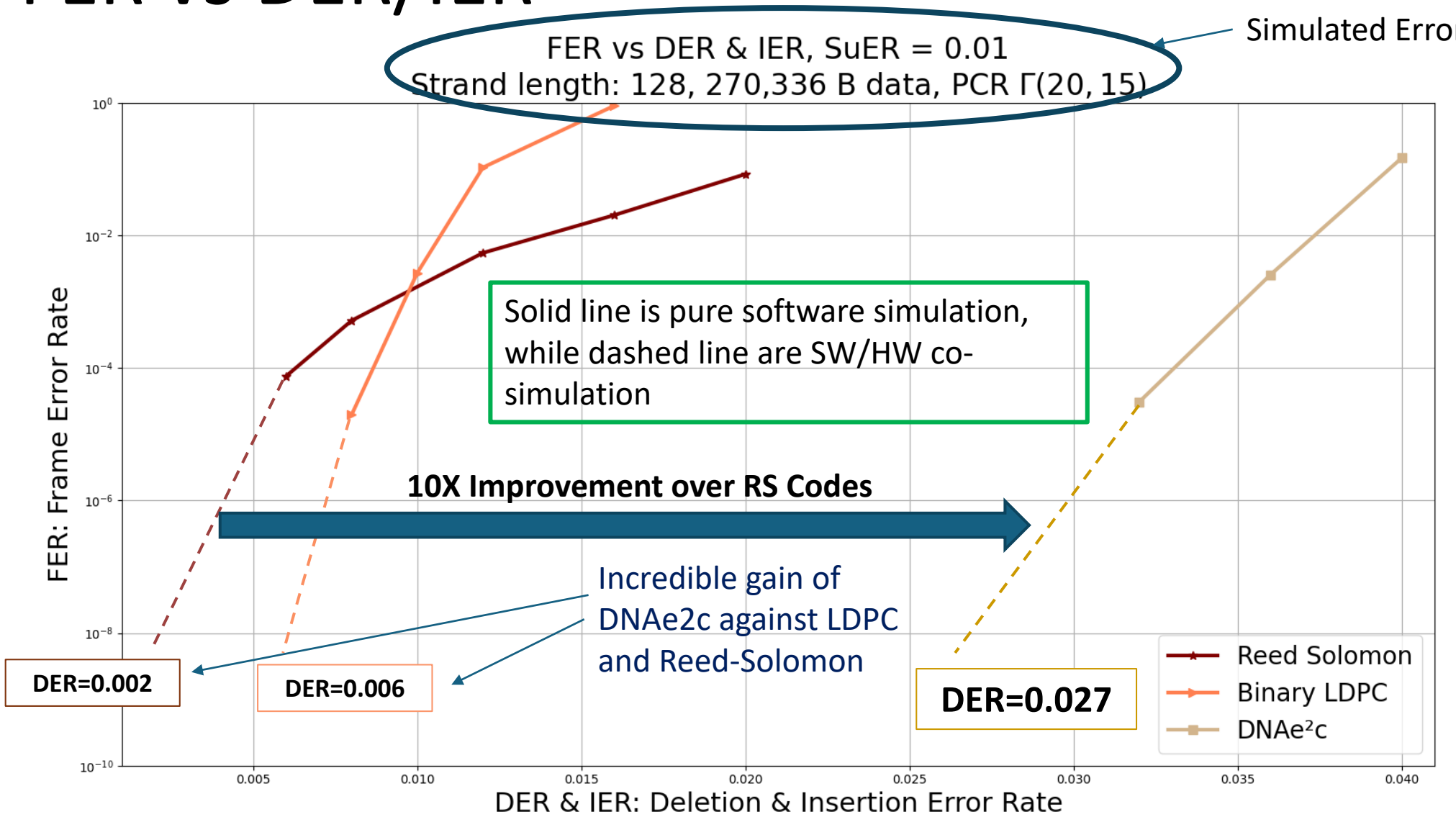


Solution is based on a proprietary code which must be

- Iterative so that latency can be changed by changing the number if iterations
- Flexible code rate so that we can change the number of parity bits that must be written according to the synthesizing machine in use
- Tricks (Recovery Mechanisms)– such that we can enable/disable different tricks depending on the error conditions
- HW implementable so that it will be easier to deploy it in data center solutions

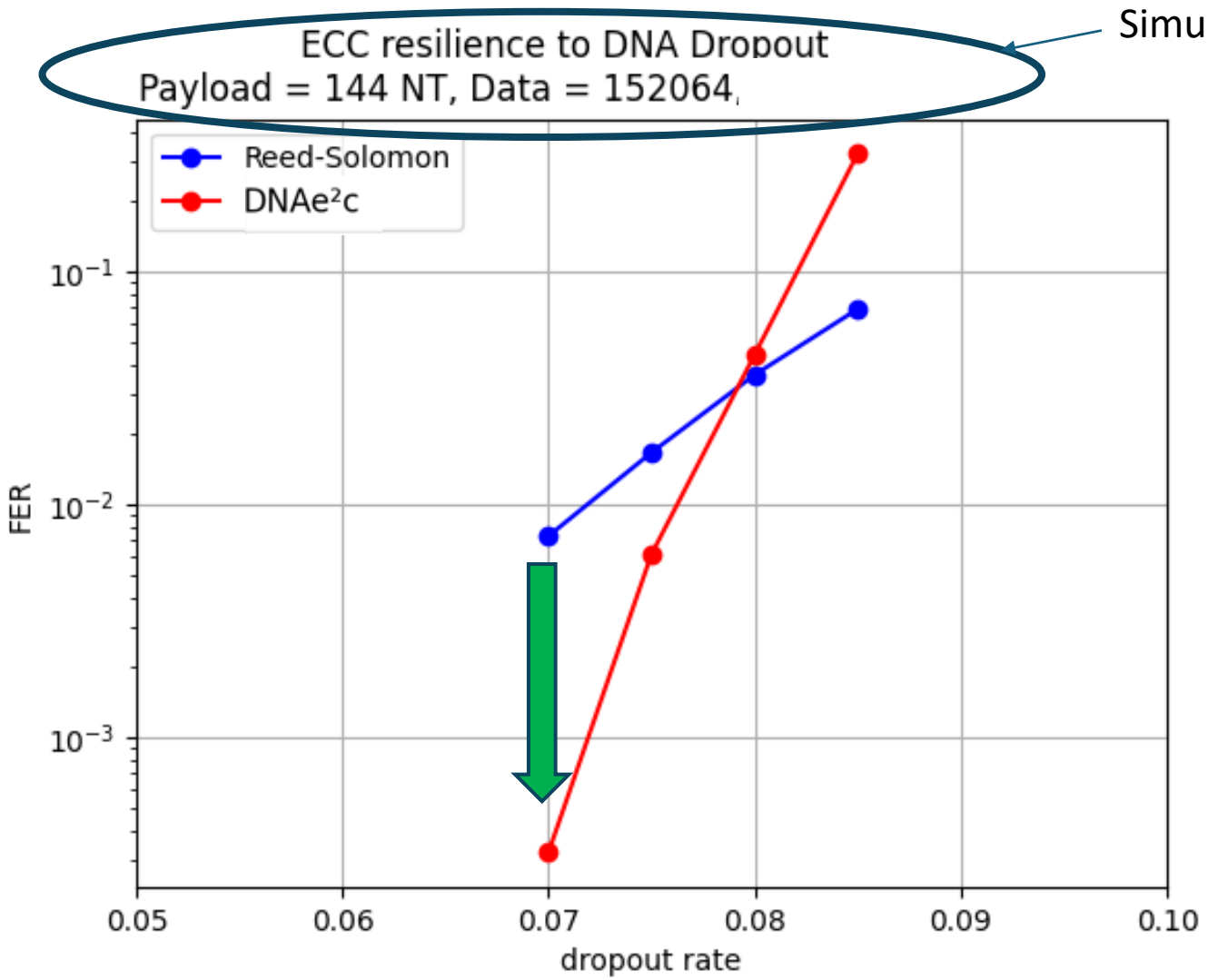
DNAe²c ®

FER vs DER/IER



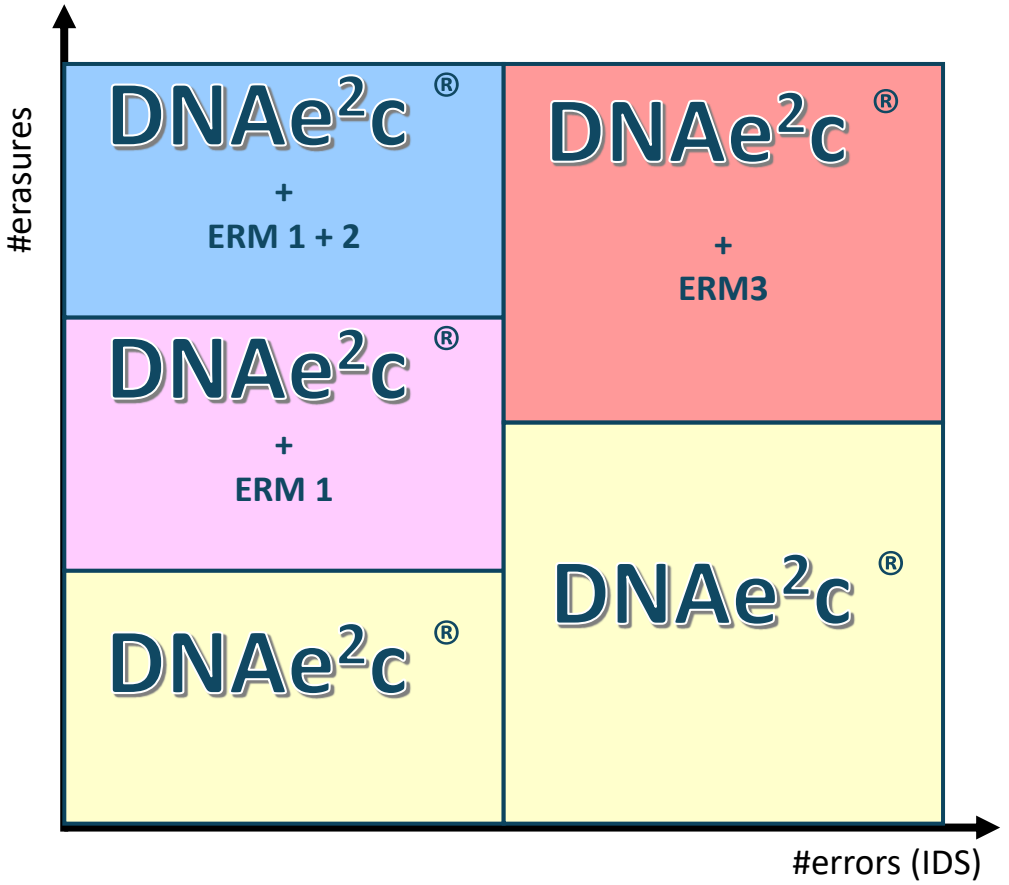
If we fix a target FER of 10^{-9} (0 is not an option!), in other words we accept of losing one frame over a billion, DNAe²c ® is able to have a DER/IER 10x bigger in comparison to Reed-Solomon codes!

FER vs dropout rate/erasure rate



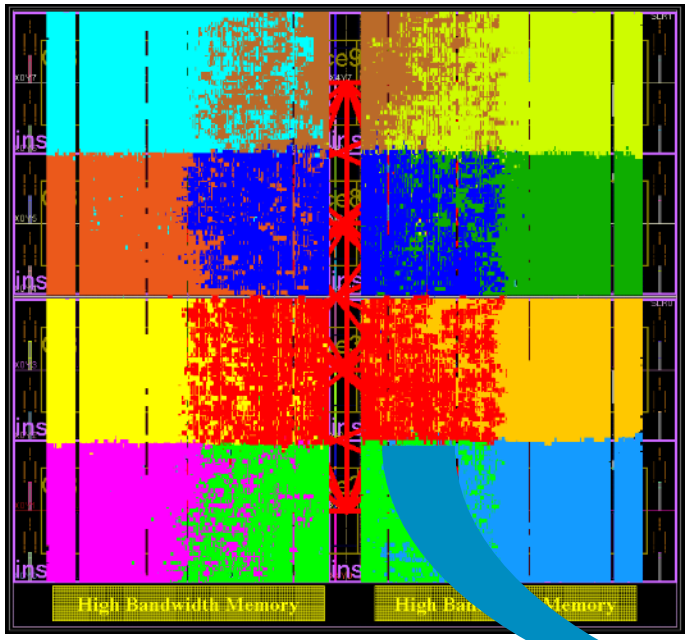
- By fixing a droput rate (e.g. 0.07) that means that 7% of the strands are in erasure, we see the gain of around two orders of magnitude of DNAe²c ® compared to Reed-Solomon

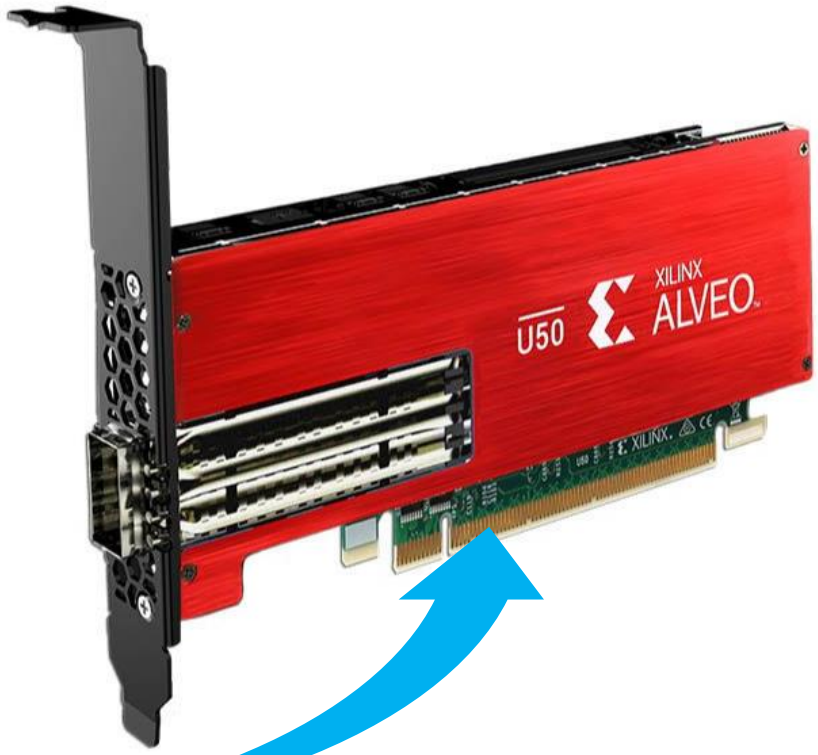
- DNAe2c ® offers a complete solution suite tailored to specific error conditions.
- If erasures increase, we can enhance reliability by adding ERM1 (Erasure Recovery Mechanism) or combining it with other correction techniques.
- If both IDS errors and erasures increase significantly, we can further strengthen the system by introducing ERM3, an advanced recovery mechanism.



ENABLING/DISABLING RECOVERY MECHANISM

- In order to evaluate power consumption and computational effort we implemented DNAe^{2}c ® on a FPGA (Xilinx Alveo U50)





HW/SW IMPLEMENTATION OF CODEC

CONCLUSIONS

- A new storage medium is essential to handle the ever-growing volume of daily data.
- DNA storage is a strong candidate, and Error Correction Codes (ECC) make it viable even when the medium is inherently unreliable.
- Encoding and decoding for DNA are complex, involving multiple stages—significantly more intricate than in Flash or HDD systems.
- The noise channel is modeled as a mix of PCR and IDS errors, capturing realistic error behavior.
- DNAssim® is used to explore and optimize encoding/decoding strategies tailored to specific noise profiles.
- DNAe2c® offers a flexible and complete solution, allowing tuning based on known noise characteristics to achieve a target Frame Error Rate (FER).

CONCLUSIONS

avaneidi.com

marketing@avaneidi.com

Avaneidi Spa

THANK YOU