DNAe^{2}c® ECC for DNA Data Storage: 10x Improvement over RS Codes

M. Montana, A. Marelli, R. Micheloni, V. DeCian, C. Spolaore, C. Tocalli

Presented by Mario Montana

Agenda

- About DNAalgo
- Why DNA storage and ECC
- Error Sources in the DNA Channel
- CODECs in literature
- The DNAalgo CODEC: DNAe^{2}c ®
- Conclusion

DNAalgo Inc

DNAalgo: Company Profile

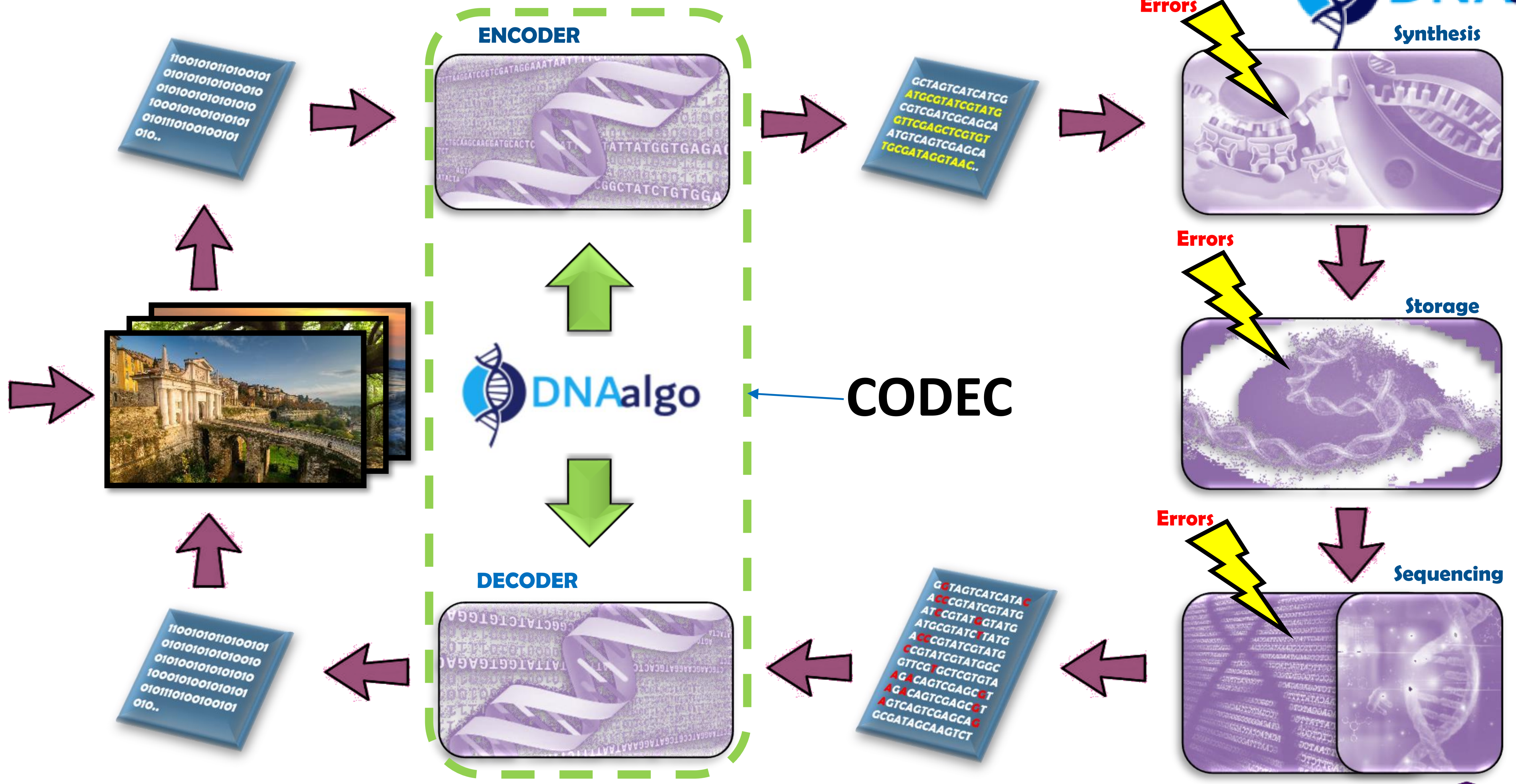
- Executive Team
  - Sabrina Barbato (biologist), CEO
  - Rino Micheloni (engineer), COO
  - Alessia Marelli (mathematician), CTO
  - Mario Montana (senior executive), Chief Strategy and Alliance Officer – Board Member
- Located in Italy
- Privately held
- www.dnaalgo.com

DNAalgo: Mission

to leverage "Information Theory" for a *fast* and *reliable* DNA storage

At DNAalgo we believe that data "manipulation" is the only way for making DNA storage reliable and fast enough for the storage industry; without reliability and speed, DNA storage won't go too far from Today's proof-of-concept stage.

DNAalgo's role inside the DNA Storage Pipeline



What we offer: 3 pillars

Noise Modeling

We build stochastic models of the storage errors associated with any Synthesis/Sequencing technology; these models can be used to run software simulations instead of expensive and long Synthesis/Sequencing experiments.

Encoding and Decoding IPs

Using synthetic DNA for data storage implies two steps of digital data processing: Encoding and Decoding. We combine a full set of error stochastic models with a proprietary simulator (DNAssim) to develop the most efficient IPs for both Encoding and Decoding.

DNAssim

Because of the intrinsic statistical behavior of the storage errors, a simulator is required for figuring out the impact of Encoding/Decoding algorithms. We have built from scratch a full-system simulator for DNA storage which enables a complete design exploration of Encoding/Decoding: DNAssim.

Why DNA storage and Why Error Correction?

Why DNA Storage?

- Massive amounts of data are being generated every day
  - A new archival storage layer is needed beyond tape
- DNA storage enables..

Longevity

Low power

Capacity

DNA Storage Creates Challenges wrt Errors

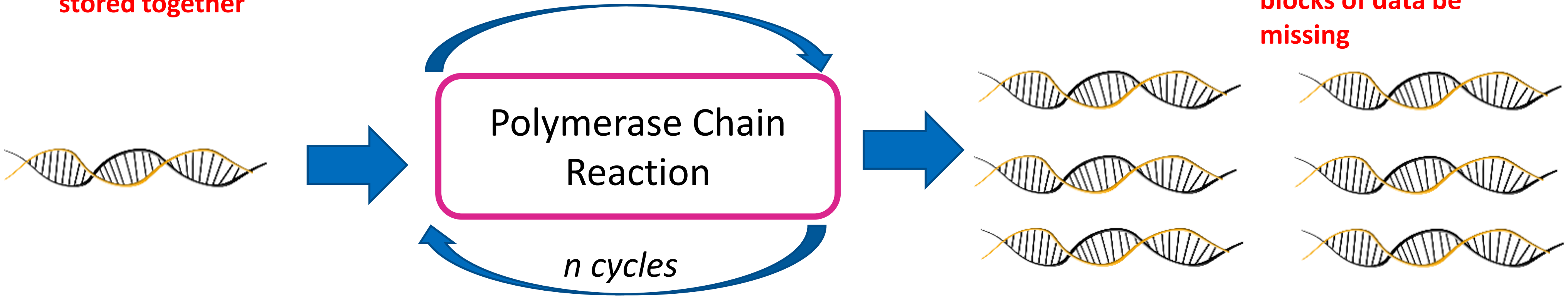
- Nothing comes for free, so the main DNA storage issues are

A huge amount of data are stored together

Data are read without order

Channel IDS

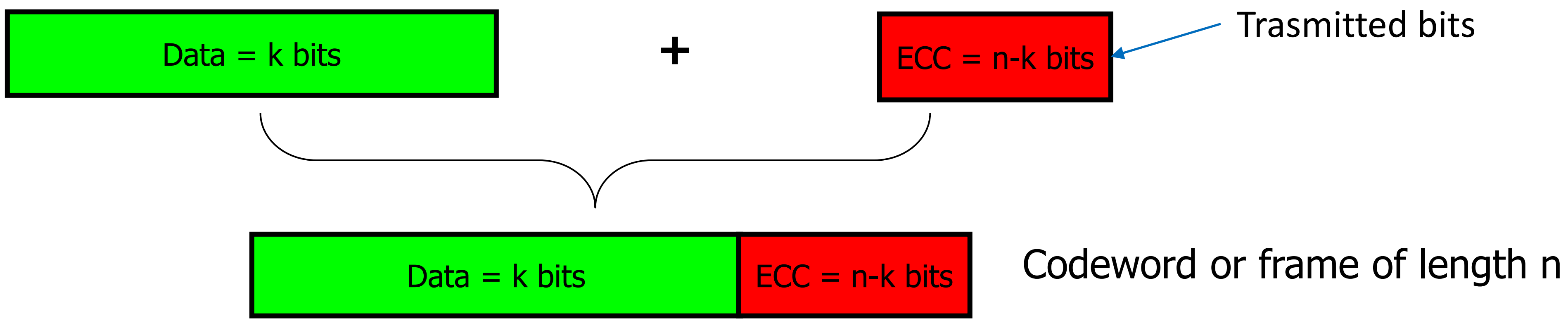
Erasure/PCR replicas: blocks of data be missing



- At DNAalgo we believe that data "manipulation" is the only way for making DNA storage reliable and fast enough for the storage industry; without reliability and speed, DNA storage won't go too far from Today's proof-of-concept stage

Error Correction Codes (ECC)

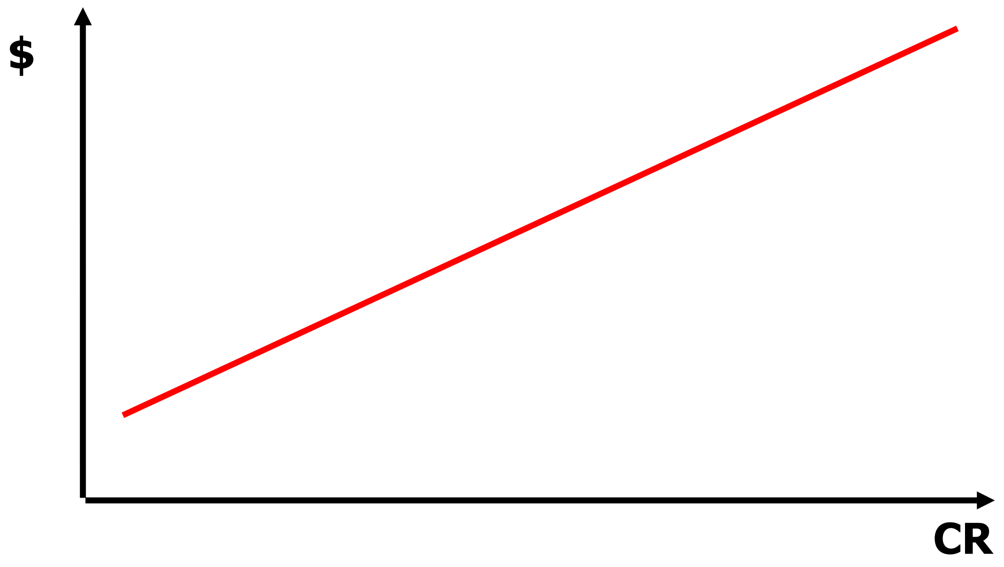
- Error Correction was born as a part of information theory for telecomunications



- The main purpose is to correct the errors that occur during transmission over a medium

Code Rates & ECC

- Code rate CR is defined as the ratio k/n



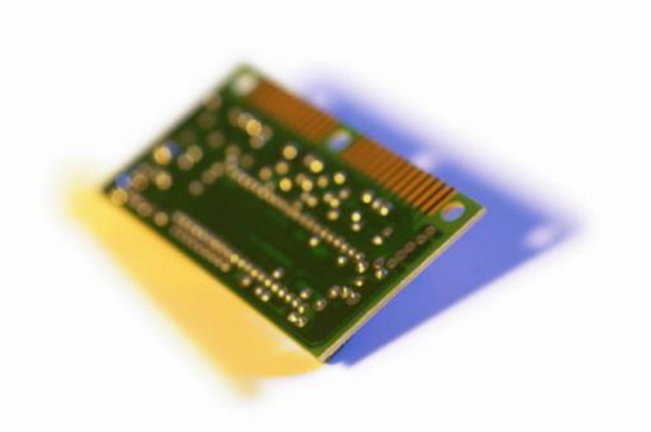
A high code rate guarantees less overhead and so a monetary gain
A low code rate guarantees high correctability

Example: ECC in Communications
Phone wires and the internet

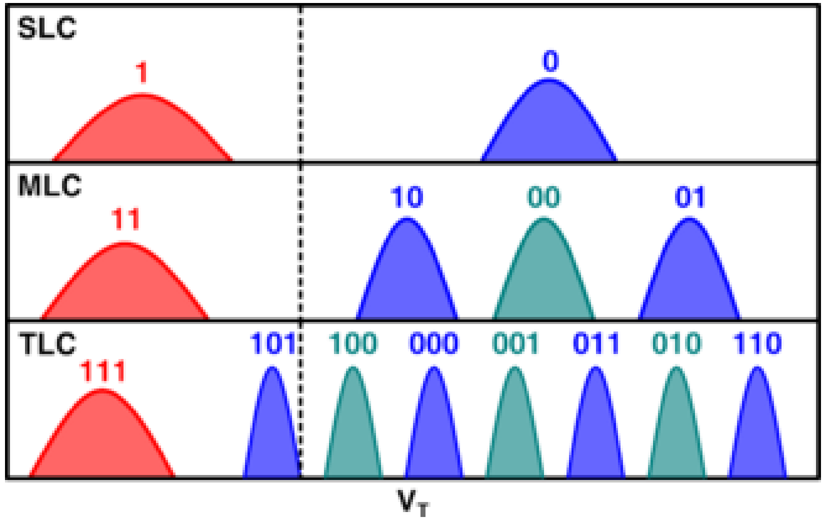
- In the late 80s home phones were a commodity on quite all the houses.
- At that time, internet was becaming popular in houses.
- How was it possible to send digital data in all the houses without changing the infrastructure?

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

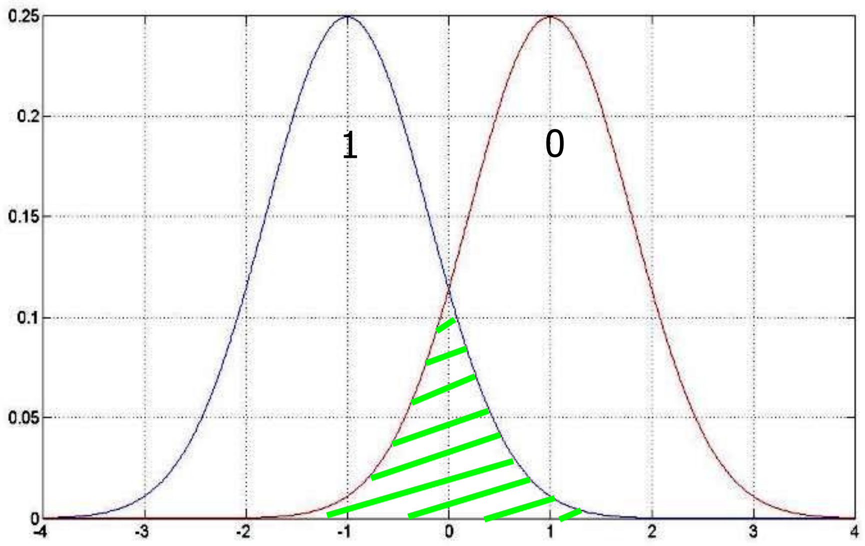
Application Example: Flash Storage (SSD)



- In early 2000s the Flash market was dominated by NOR memories. At that time the NAND Flash arrived. While NOR Flash was reliable, NAND was not due to their intrinsic structure
- But NAND is fast and very scalable.
- In the same voltage space were it was possible to discriminate between 2 digital values, in few years it was passible to have 4, 8, 16, ... Distributions
- NAND Flash is low cost and is used in a lot of applications such as SSDs.



NAND and SSD: Also a challenging medium



- Error region is on each overlapping region between distributions.
- Distributions overlap due to the usage and retention
- In a space with 16 or 32 distributions, they overlap also in a fresh device. The estimated error probability is 10^{-2}
- How can we such an error prone media for enterprise grade applications that need closer to error rates of 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**

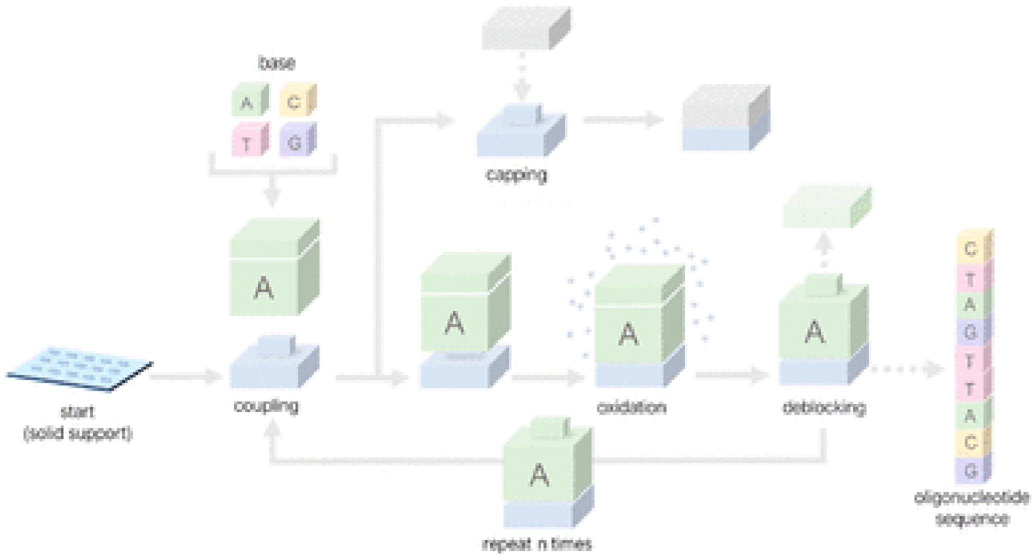
The mission of ECC in DNA storage..

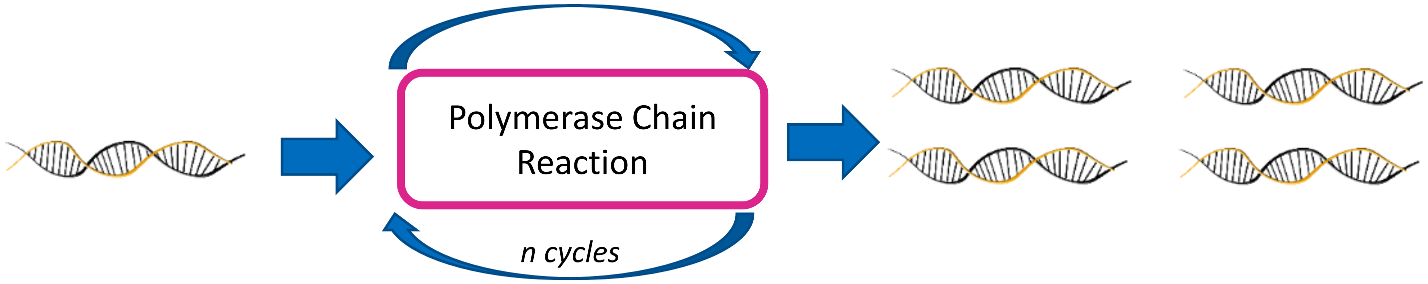
- When DNA is created for a storage use, it is not necessarily by itself a bad medium – it is actually quite stable resulting in strong data retention.
- The problem lies in the synthesis (writing) sequencing (reading) processes which generate errors.
- In order to reduce the «noise» in the process, processes employ expensive and time consuming techniques to write and read the information
- A storage system more resilient to errors could potentially tolerate a more approximate synthesis or sequencing process hence, decreasing time and cost of the biological methods

    - **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**

Error Sources in the DNA Channel

IDS errors & Erasures

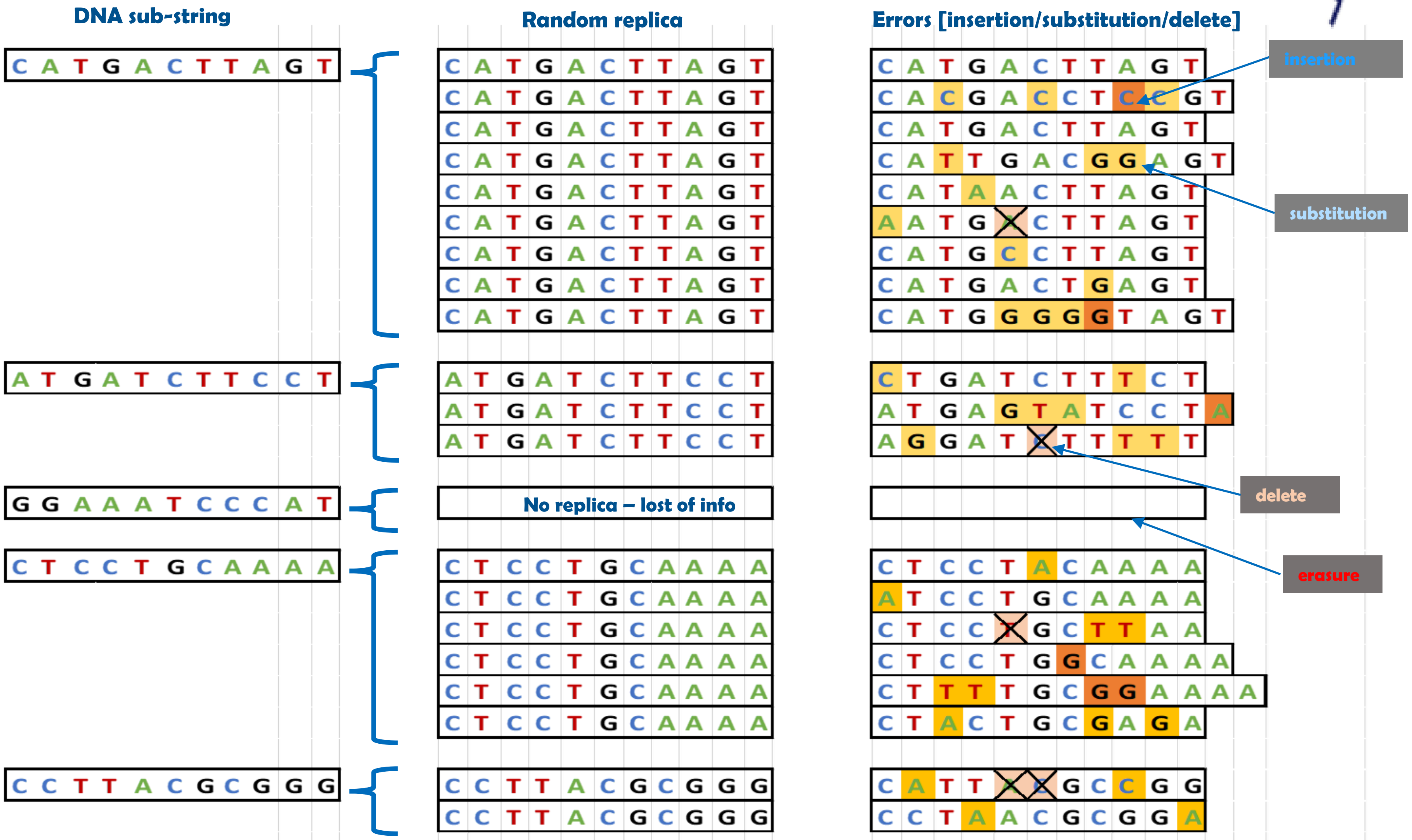




During synthesis, errors arise from ***incomplete capping*** and ***DNA damage*** during oxidation and deblocking steps These errors can be an ***insertion, a deletion or a substitution (IDS)*** at a nucleotide level

During sequencing, ***PCR is applied*** so that each strand is read a variable number of times (also 0 times) creating possible ***erasures*** of entire strands

Information Channel example



Current approaches for CODECs in DNA storage and How to Evaluate

State-of-the-art ECCs/CODECs for DNA Storage

- Error Correction Codes are used mainly for substitution errors, or erasure errors
- Known codes used so far in the industry are:
  - **Reed Solomon** (Organick, Lee, et al. "Random access in large-scale DNA data storage." Nature biotechnology 36.3 (2018): 242-248.)
  - **LDPC** (Chandak, Shubham, et al. "Improved read/write cost tradeoff in DNA-based data storage using LDPC codes." 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2019.)
  - **Fountain** (Erlich, Yaniv, and Dina Zielinski. "DNA Fountain enables a robust and efficient storage architecture." science 355.6328 (2017): 950-954.)
  - **Hedges** (Press, William H., et al. "HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints." Proceedings of the National Academy of Sciences 117.31 (2020): 18489-18496.)

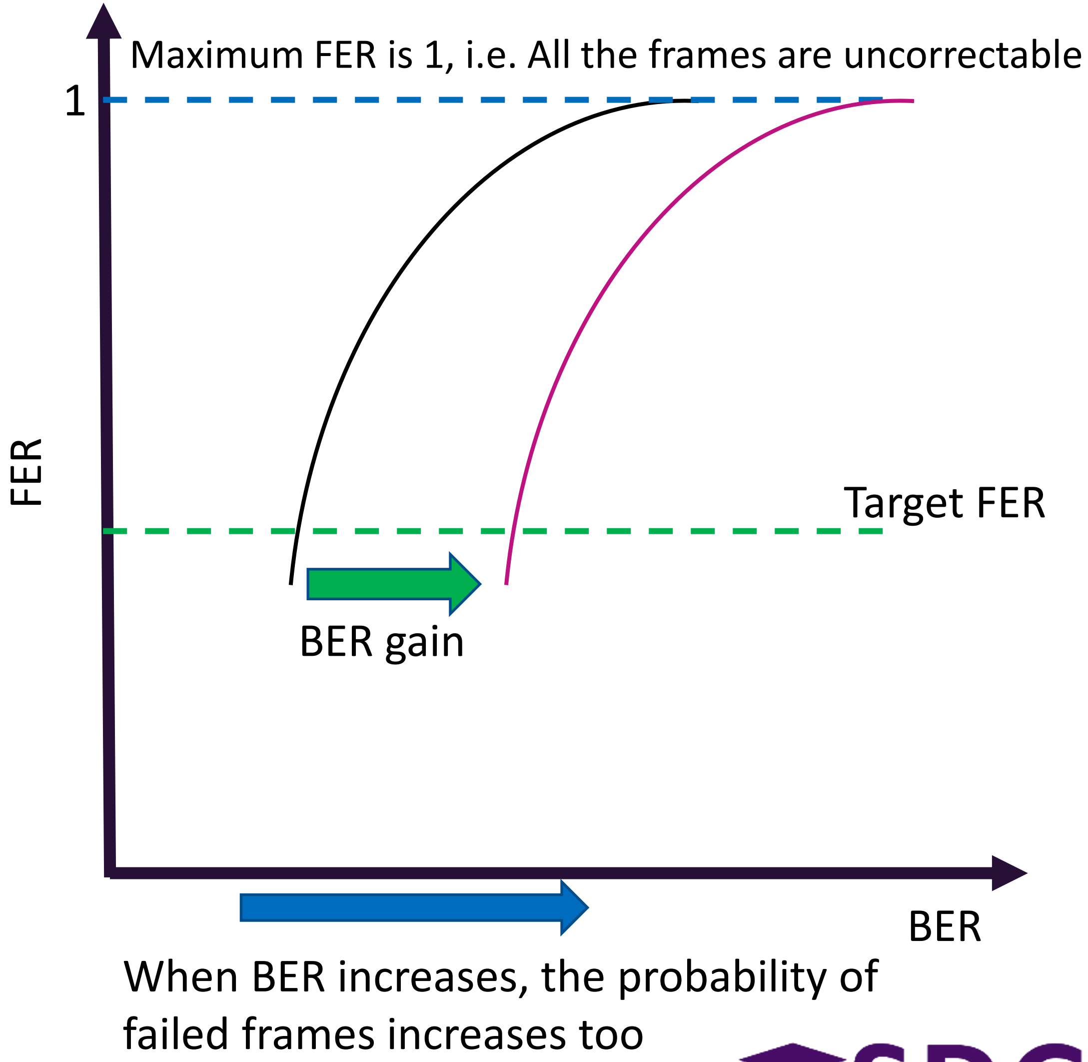
How we do compare ECCs?

- Reed Solomon and LDPC are standard ECC described in a substitution channel or erasure channel
- Fountain codes are very powerful but mainly for erasure
- Hedges code has a very low code rate and are very computationally intensive and are also used for erasure

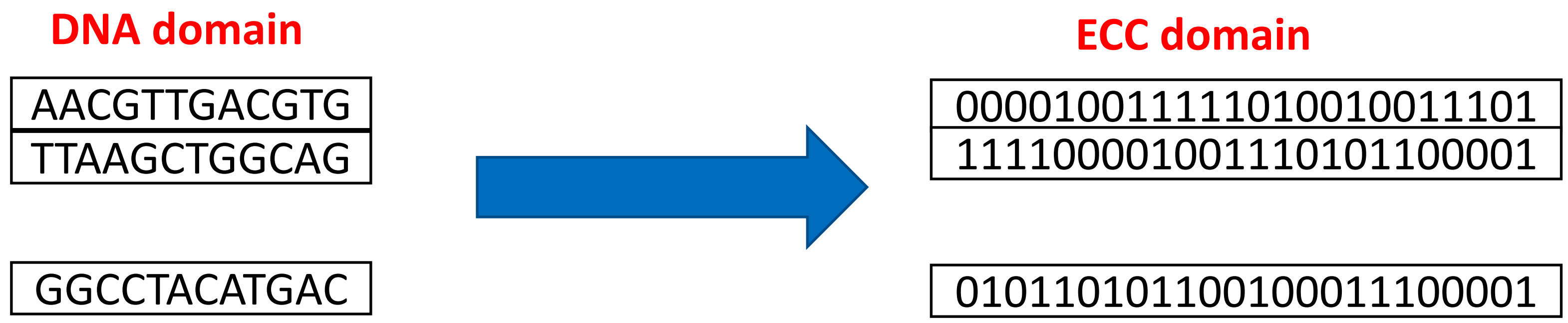
Code	Erasure	Insertion	Deletion	Substitution
No codec	NO	partially	partially	Partially
Reed Solomon	partially	NO	NO	partially
LDPC	NO	YES	YES	YES
Fountain	YES	NO	NO	NO
Hedges	NO	YES	YES	YES
DNAalgo DNAe^{2}c ®	YES	YES	YES	YES

FER vs BER?

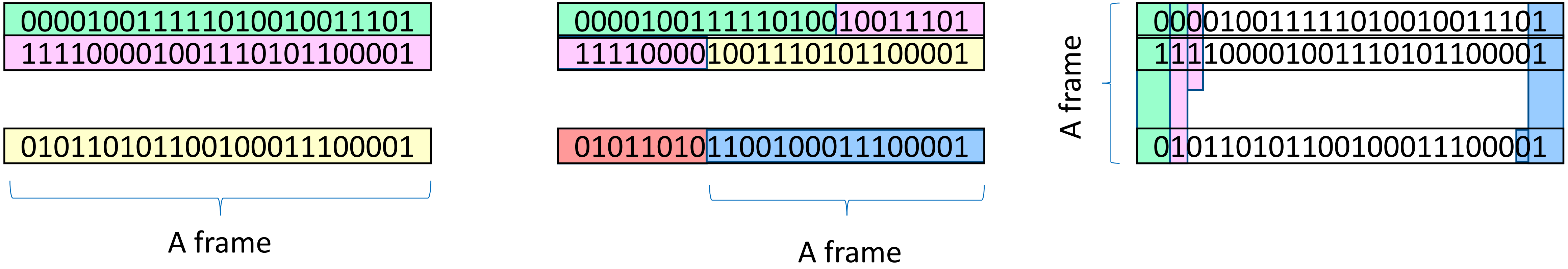
- Generally codes are evaluated as Bit Error Rate (BER) against Frame Error Rate (FER) which represents the probability of having un uncorrectable frame given a specific BER.
- Given a target FER, the better ECC is the one that can reach the target with the highest BER possible. In other words the most right we are, the better ECC we have
- We can compare two codes by computing the percentage of BER we can gain to reach the same target



FER in DNA data storage



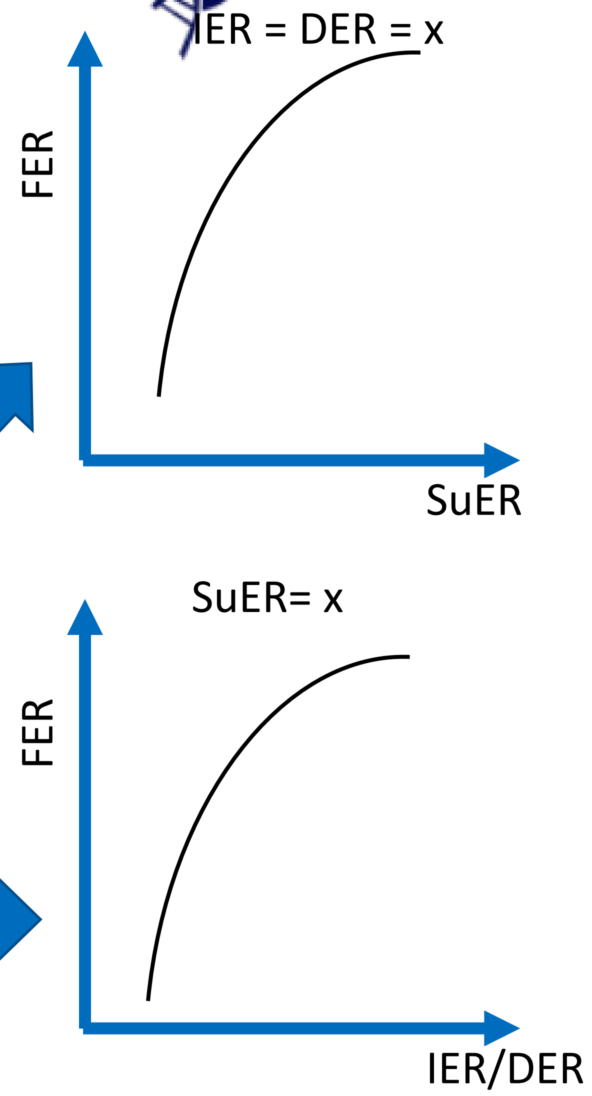
In ECC domain strands have a longer length than in DNA domain, the lenght is 1.5-2 times bigger depending on the mapping used



A frame can coincide with the strand length, but it can also go across strands in different ways depending on the coding strategy used

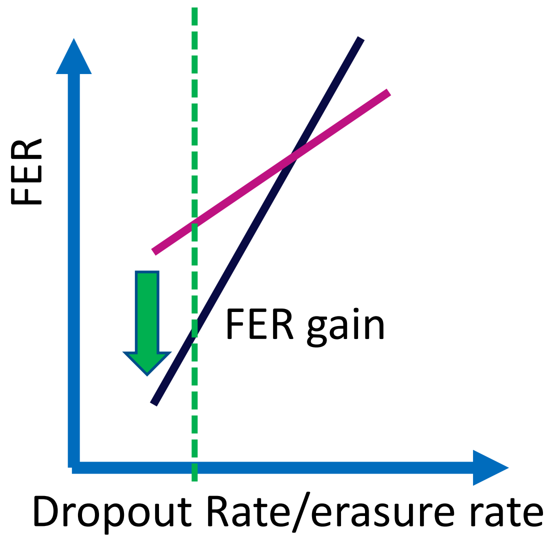
FER vs SuER/IER/DER

- In DNA channel, we do not have BER because the errors can be created by insertion, deletion and substitution and all those probabilities are independent
- The analysis is split in 2 graphs:
  - FER vs SuER (substitution error rate) where IER (Insertion Error Rate) and DER (Deletion Error Rate) are fixed and equal
  - FER vs DER/IER where SuER is fixed



FER vs erasures

- In DNA channel, some strands can be lost, this is what we call erasures. In any case it is possible to scramble the erased nucleotides among all the frames.
- In addition to that we may add erasures in trace reconstruction for example, if we found out that a strand hasn't the correct length and we decide to erase the whole strands.
- In order to evaluate performance against erasures we provide a graph of FER vs Dropout Rate/erasure rate
  - In this graph, by fixing a dropout rate, the better code is the one that shows a smaller FER



DNAe^{2}c ®

The Goal

DNAalgo

- Error sources can change a lot depending on the synthesizing or sequencing machines ➡ we want an ECC with some «tuning» parameters so that it can be targeted to a specific error channel
- In order to understand performances we need to perform simulations with known codes with a lot of different parameters (SuER, erasure, PCR distribution, etc.) ➡ we need to implement different ECC on DNAssim and perform many simulations against our solution. i.e SW/HW co-simulations
- We want to keep the CR as high as possble in order to avoid «writing too much» and to keep computational complexity and power consumption as low as possible ➡ the solution must be implemented in HW

SDC 23

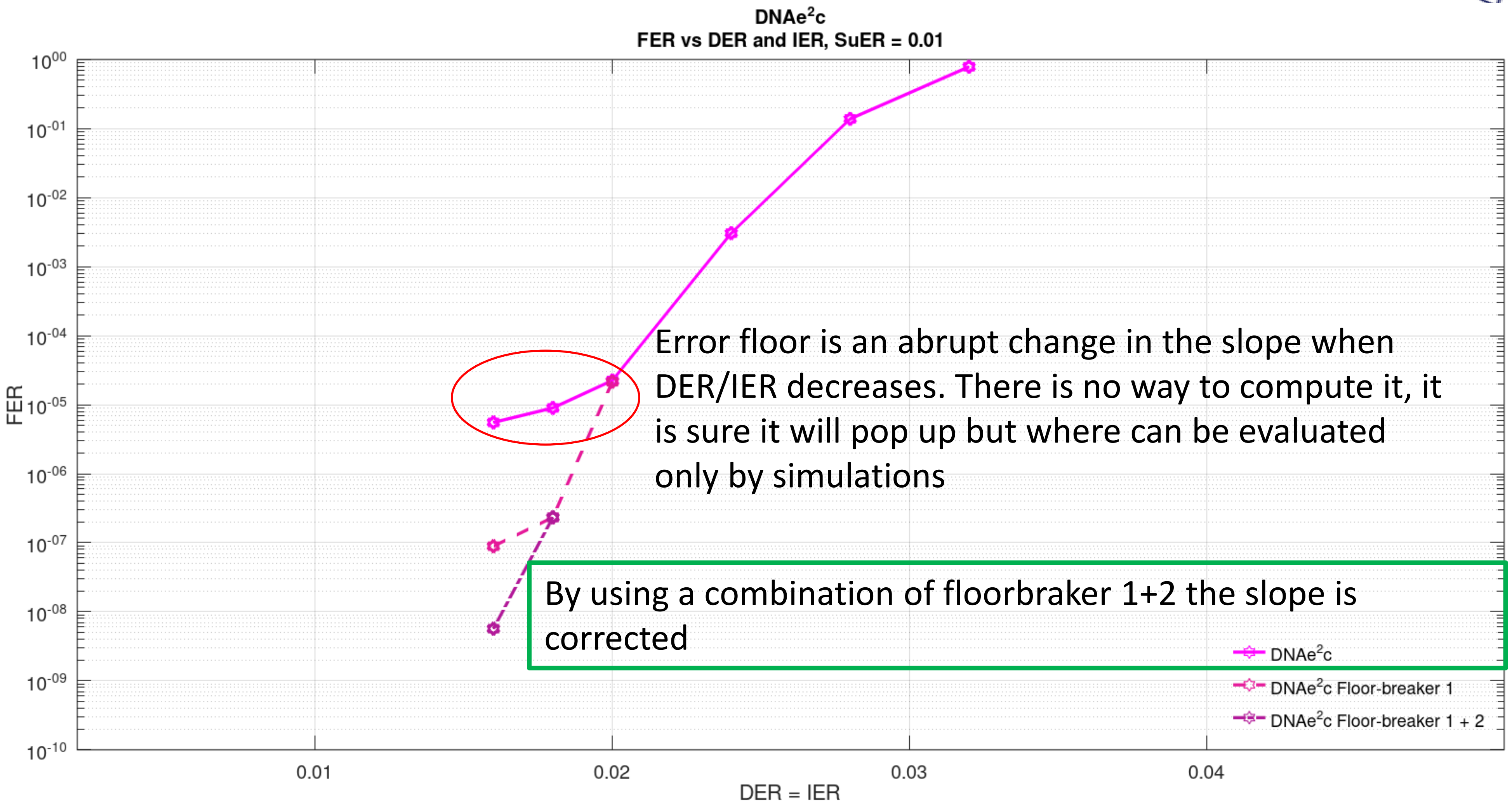
DNAe^{2}c ®

DNAe^{2}c ®

NA | oise | ware | rrors&erasures | leaner

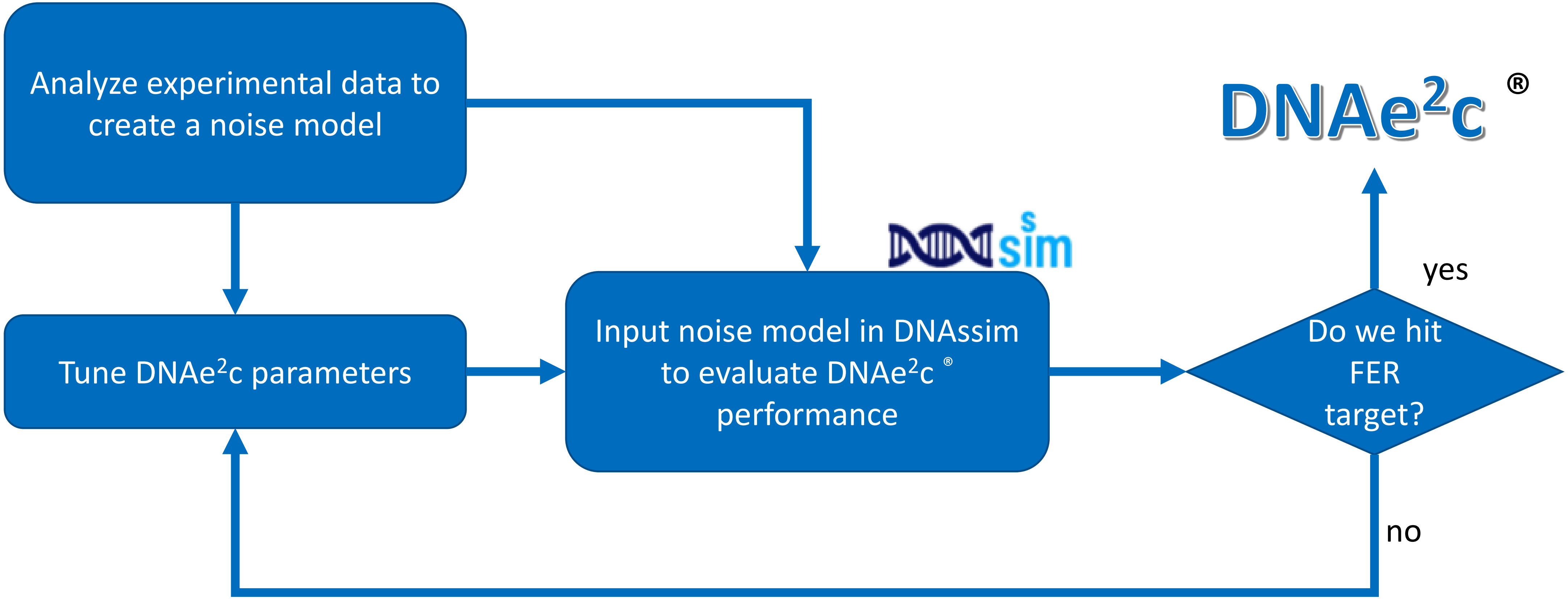
- Solution is based on a proprietary code which must be
  - Iterative so that latency can be changed by changing the number if iterations
  - Flexible code rate so that we can change the number of parity bits that must be written according to the synthesizing machine in use
  - Tricks (Recovery Mechanisms)– 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

Error Floor

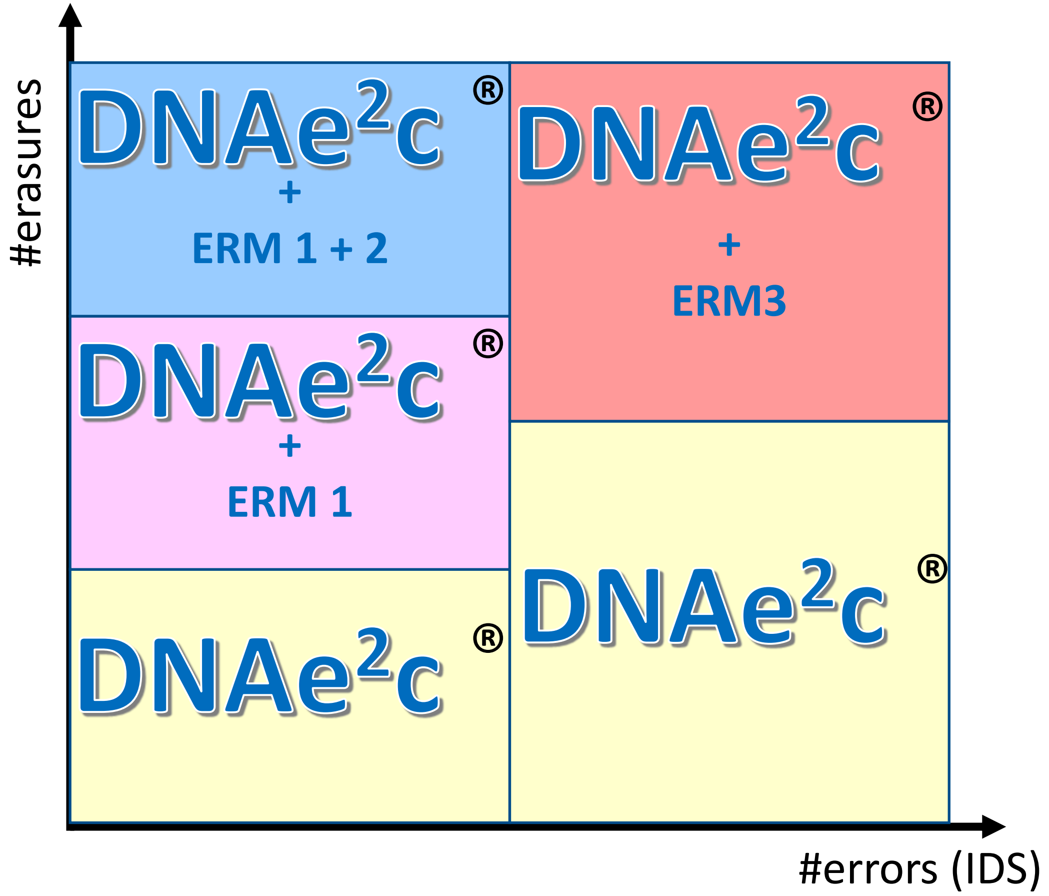


- Iterative decoders exhibits error floor
- In order to avoid it we studied and verified different floor breaker strategies that can be enabled standalone or in combination

How DNAe^{2}c ® tuning works

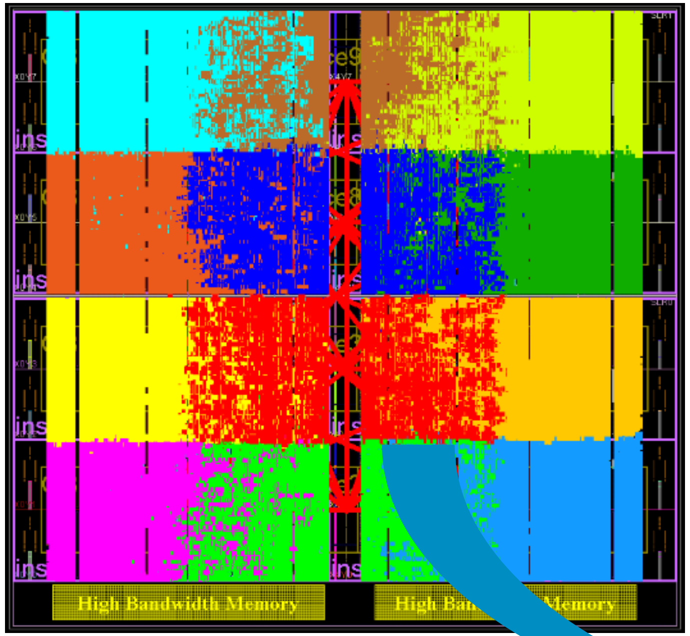


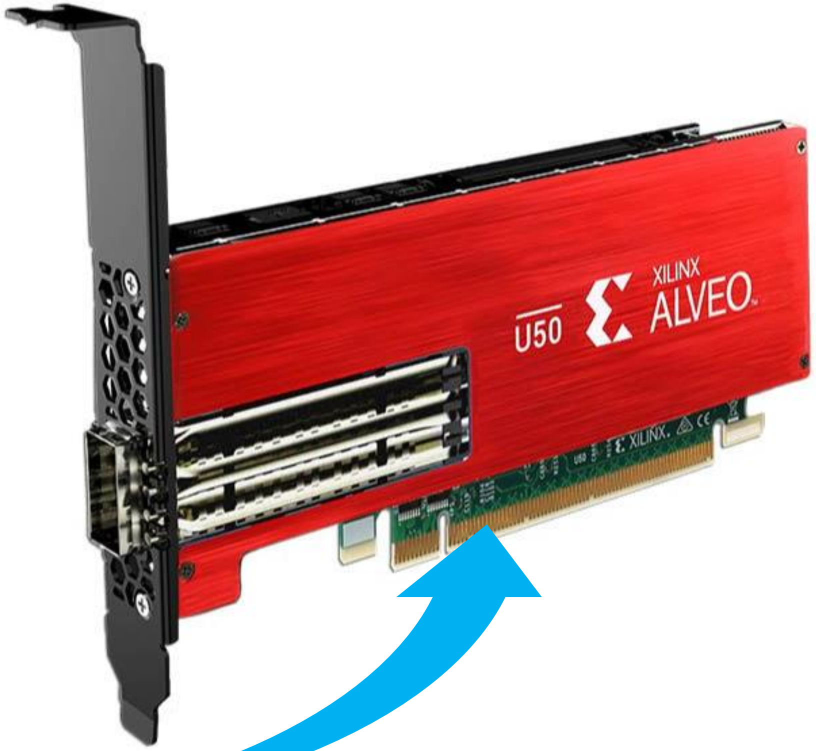
Enabling & Disabling Recovery Mechanisms



- DNAe^{2}c ® is a complete set of solutions based on the error condition
- If the number of erasure increases we can add ERM1 (Erasure Recovery Mechanism) or a combination of two different tricks
- If the number of both IDS errors and erasures dramatically increases we can add other ERM3

HW/SW Implementation of CODEC Function





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

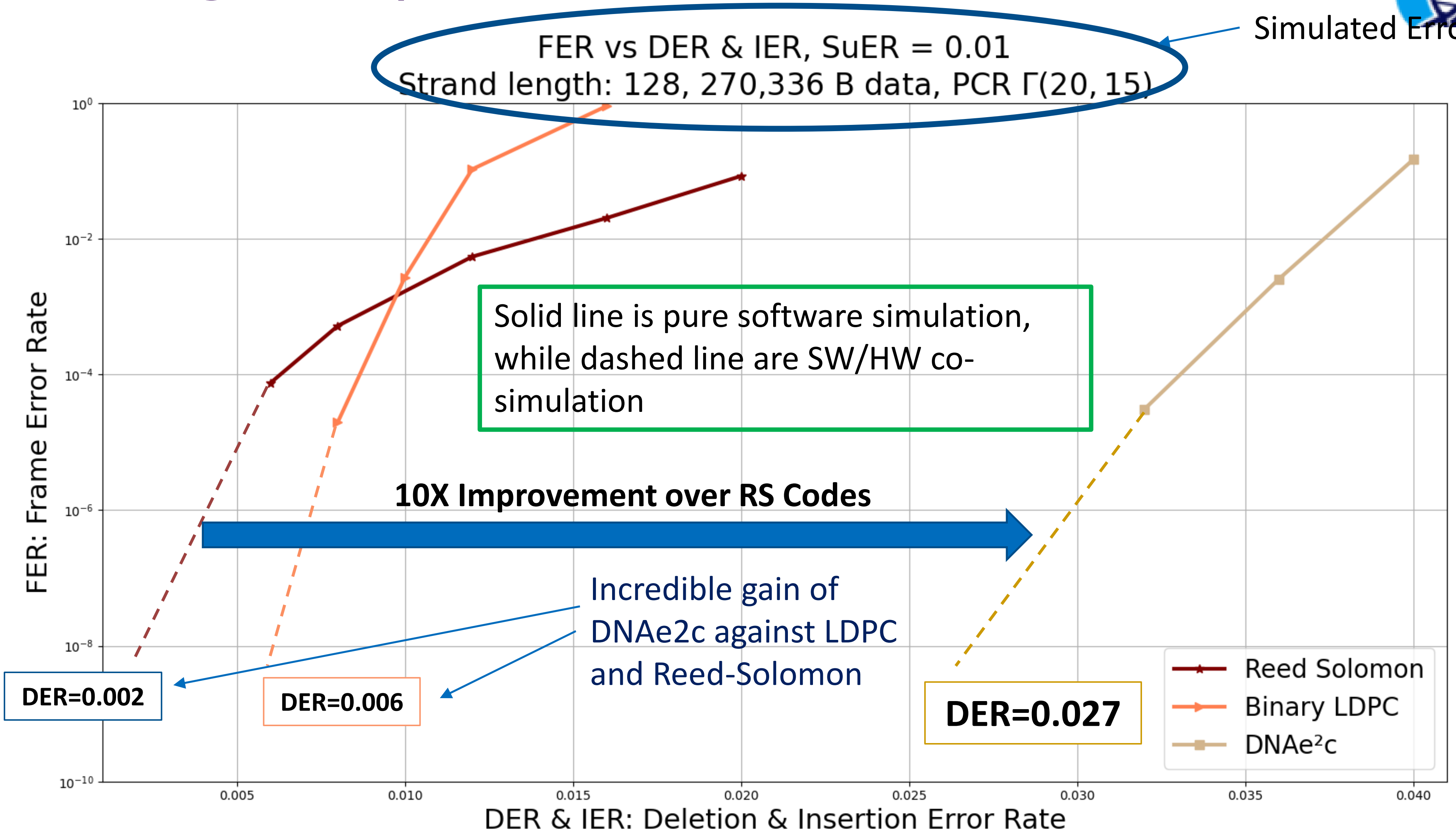
DNAe^{2}c ® comparison

Section Subtitle

Graph comparison of ECCs/CODECS

- In the following we will see some comparison of different graph in different error conditions
- Error conditions can be determined by SuER, IER, DER and Dropout Rate
- Error conditions can vary depending on the algorithms used in the pipeline (e.g. Trace reconstruction)
- DNAe^{2}c ® is a set composed by a propretary ECC + different ERMs enabled or disabled by analysizing the set of errors in a particular environment
- In order to have a curve, many simulations are performed by DNAssim in pure SW of HW/SW co-simulations

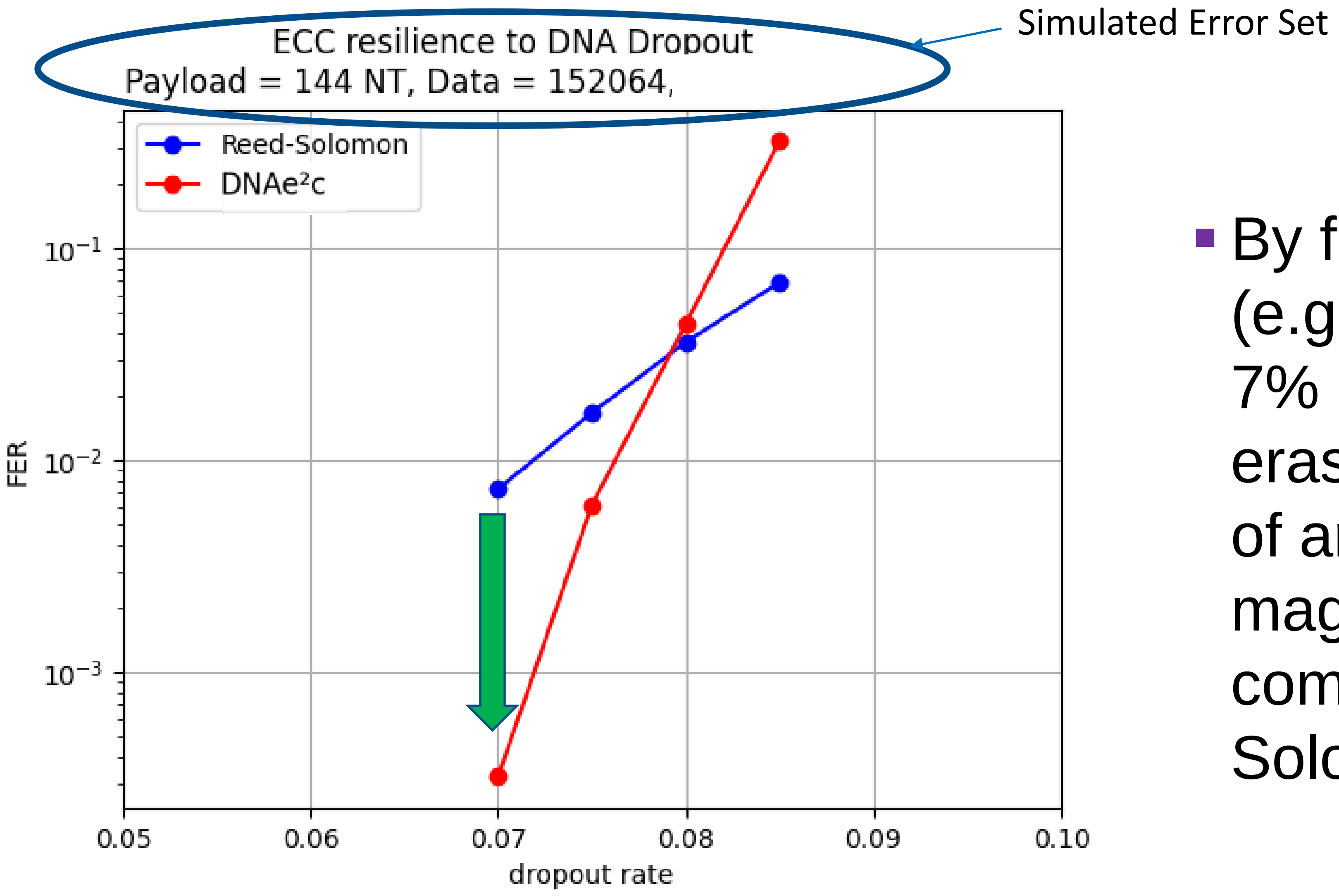
FER vs DER/IER



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 DNAe^{2}c ® compared to Reed-Solomon

Conclusion

Section Subtitle

Conclusion

- ✓ Error Correction Codes can enable the use of poor media for high performance system applications
- ✓ DNA data storage is a different channel in comparison to standard storage or telcommunications channels
- ✓ We need a way to compare different coding strategies in a DNA data storage environment
- ✓ DNAe^{2}c ® is a complete set of solutions that can be tuned based on a specific noise set. By using the knowledge of the noise condition the codec can be tuned in order to reach the target FER
- ✓ DNAe^{2}c ® has been implemented in HW and SW on a standalone accelerator card
- ✓ Do you want to challenge DNAe^{2}c ® on your particular error conditions (synthesizing machine + sequencing machine)? Reach us!

Please take a moment to rate this session.

Your feedback is important to us.

