Flash Memory Summit

DNAssim: A Full System Simulator for DNA Storage

Alessia Marelli[1], Thomas Chiozzi[1] , Lorenzo Zuolo[1],Nicholas Battistini[1], Giacomo Lanzoni[2], Piero Olivo[3], Cristian Zambelli[3], Rino Micheloni[1,3]

[1] DNAalgo
[2] Consorzio Futuro in Ricerca Ferrara
[3] Università degli studi di Ferrara

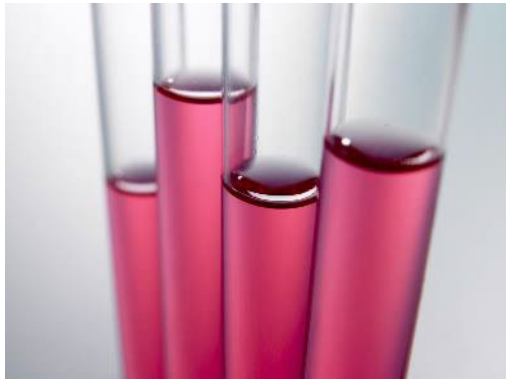
Outline

- The need of new storage media
- What is DNA storage
- Error sources
- Why DNAssim
- Encoding & decoding
- SW/HW co-simulation
- Conclusions

Why DNA?

- More and more applications are data hungry
  - Earth is covered with data centers
- DNA storage enables

Longevity



Low power

Capacity

DNA issues

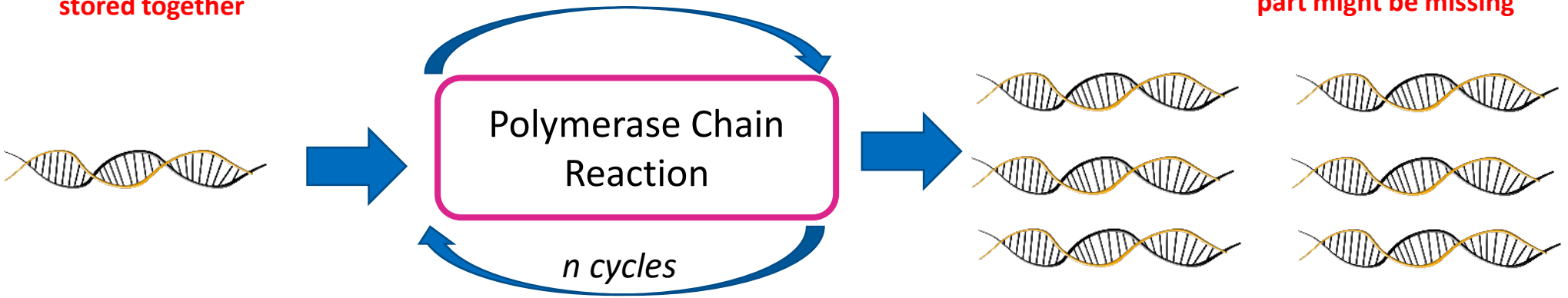
- Nothing comes for free, so the main DNA storage issues are

A huge amount of data are stored together

Data are read without order

Channel IDS

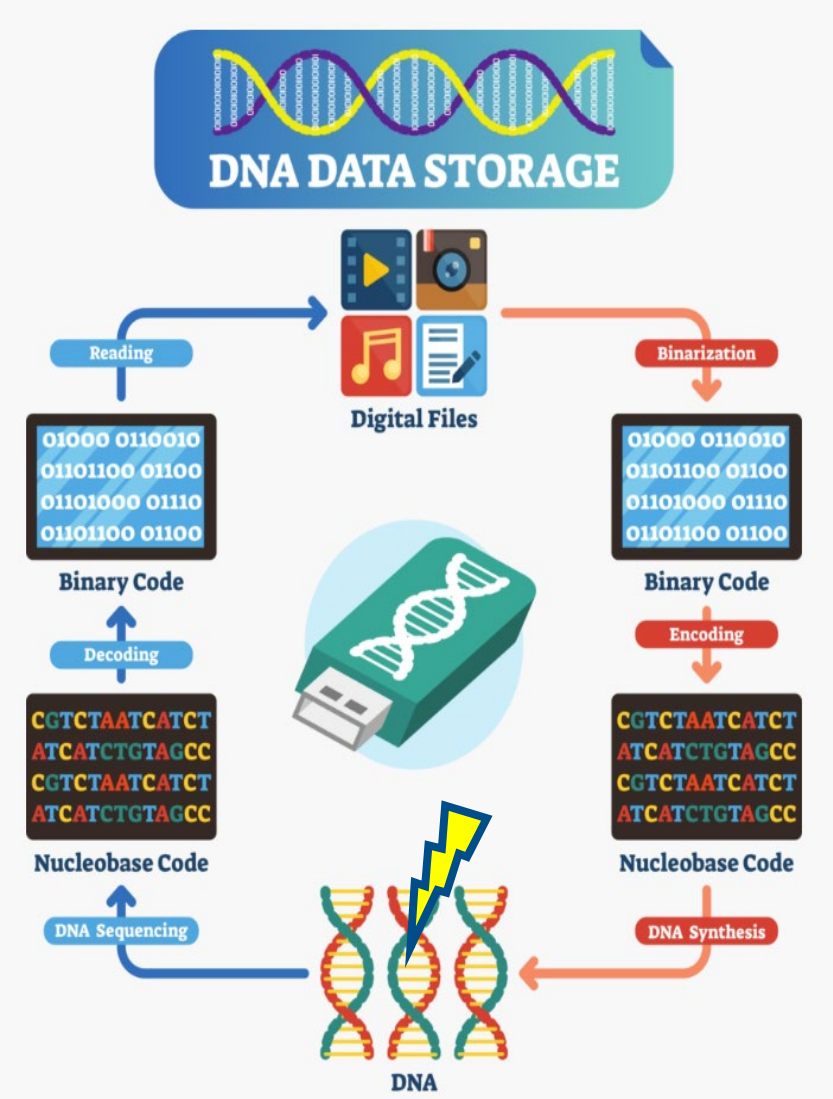
PCR replicas: some part might be missing



- 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

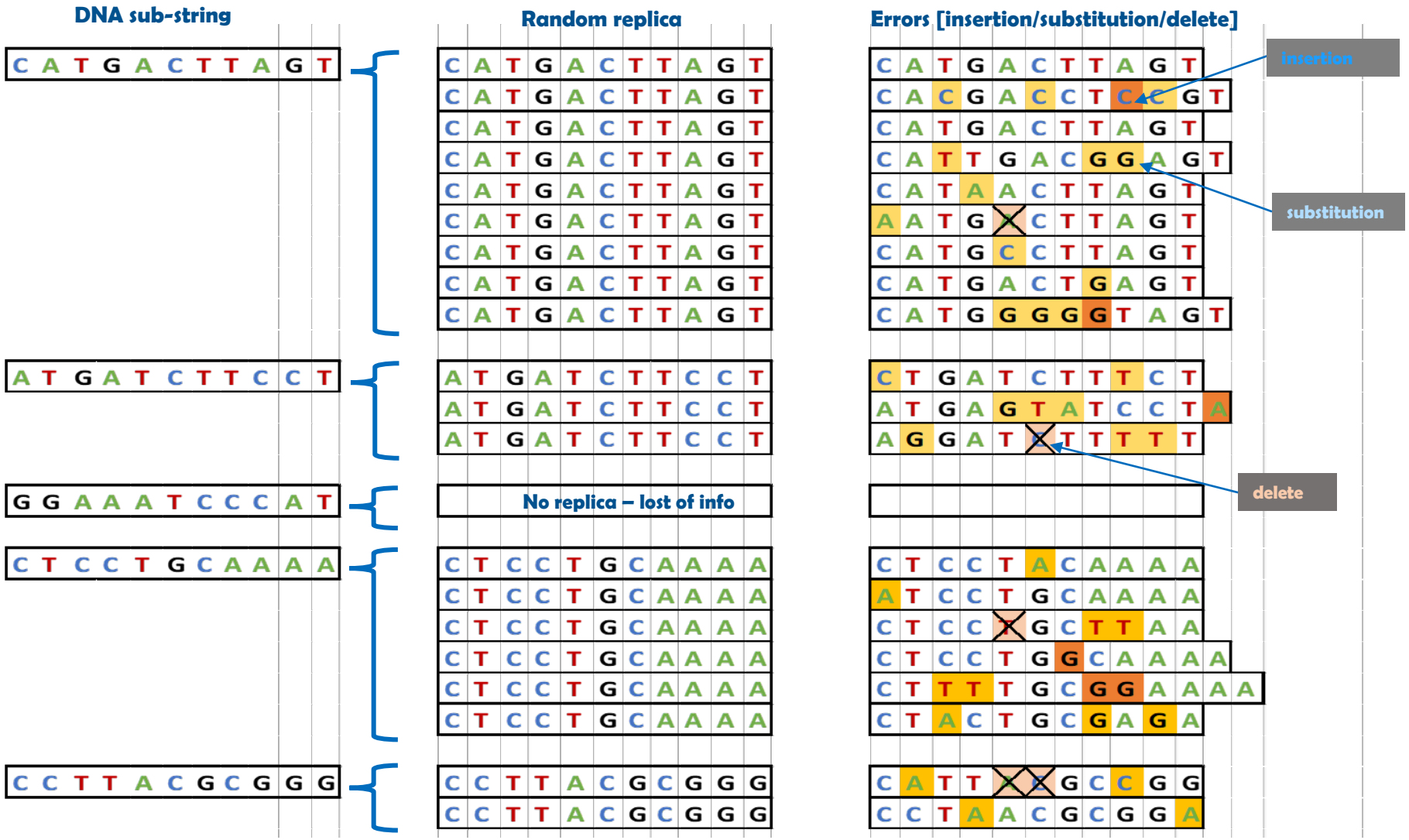
DNA storage

- During synthesis, sequencing and storing some errors might occur.
- Errors can be insertion, deletion and substitution
- In addition to that, in order to sequence the information PCR is applied so that each strand is read a variable number of times (also 0 times)



https://dnaalgo.com/

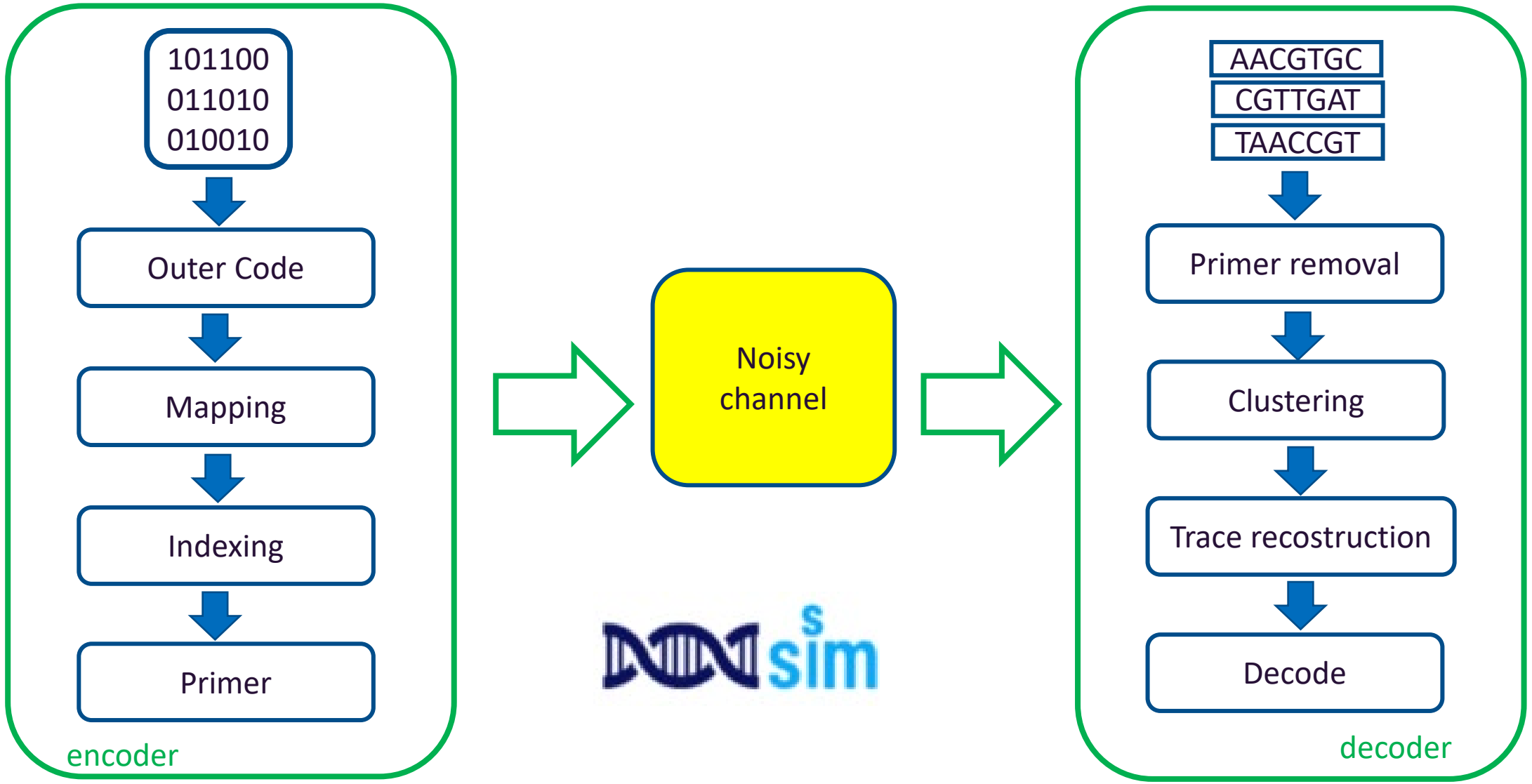
Information Channel example



Why a simulator?

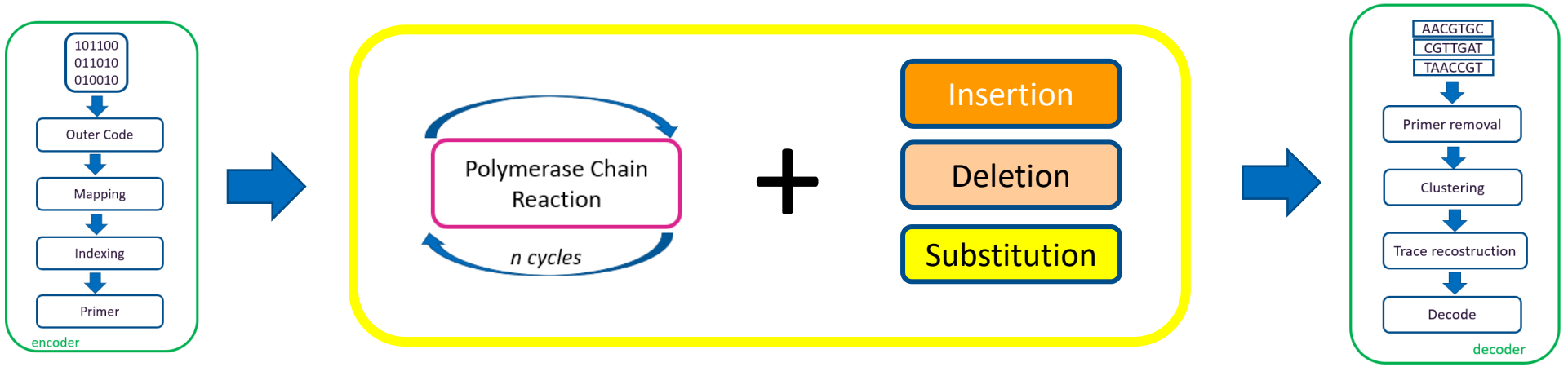
- While encoding and decoding can be described by a set of equations, errors are not deterministic and must be modeled.
- Encoding and Decoding can be optimized if tailored to a specific noise model.
- Because of the intrinsic statistical behavior of the noise, a simulator is required for figuring out the impact of encoding/decoding algorithms.

Introducing DNAsim (DNAssim)



Noise Model

- Noise can be modeled as PCR (Polymerase Chain Reaction) + IDS (**I**nsertion **D**eletion **S**ubstitution) Channel
- PCR is represented by a variable number of strand replicas
  - Tunable multiplicity
- IDS channel translates into a statistical number of apply insertion, deletion and substitution for each strand
  - Tunable substitution/insertion/deletion probabilities



Simulation tool

- DNAssim is managed by a Graphical User Interface (GUI), where all the different parameters and options can be chosen
- When simulation is completed a bunch of graphs and texts are output in order to analyze results.

Simulation tool

https://dnaalgo.com/

Comparing results

Outer Code: A
Trick on code: A,C
Mapping: natural
Randomizer: on
Indexing: A
Cluster: A
Trace Rec: A

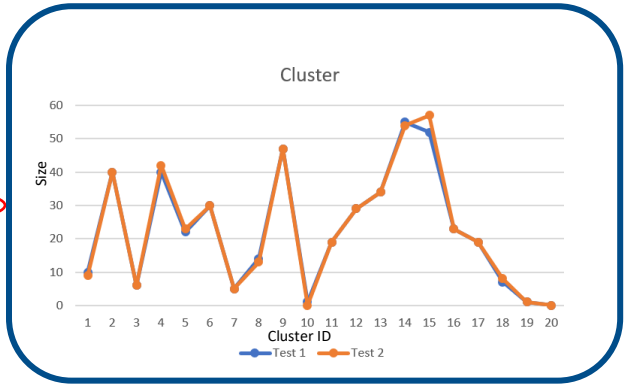
Test 1

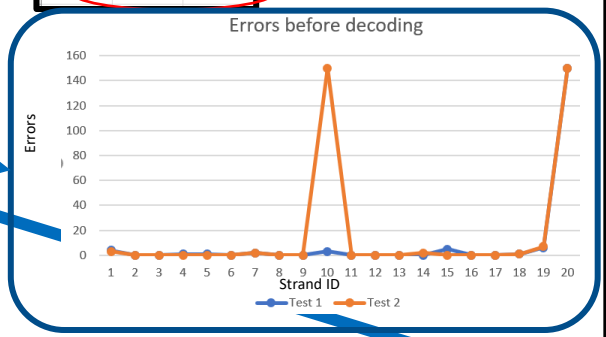
Outer Code: B
Trick on code: B,C
Mapping: natural
Randomizer: on
Indexing: A
Cluster: D
Trace Rec: A

Test 2

Percentage of G/C : 45%
Number of initial strands: 20
Number of sequenced strands: 450

1	10	9
2	40	40
3	6	6
4	40	42
5	22	23
6	30	30
7	5	5
8	14	13
9	47	47
10	1	0
11	19	19
12	29	29
13	34	34
14	55	54
15	52	57
16	23	23
17	19	19
18	7	8
19	1	1
20	0	0





1	4	3
2	0	0
3	0	0
4	1	0
5	1	0
6	0	0
7	2	2
8	0	0
9	0	0
10	3	150
11	0	0
12	0	0
13	0	0
14	0	2
15	5	0
16	0	0
17	0	0
18	1	1
19	6	7
20	150	150

Missing strands: 1
Errors before decoding: 23
Number of recovered strand: 19

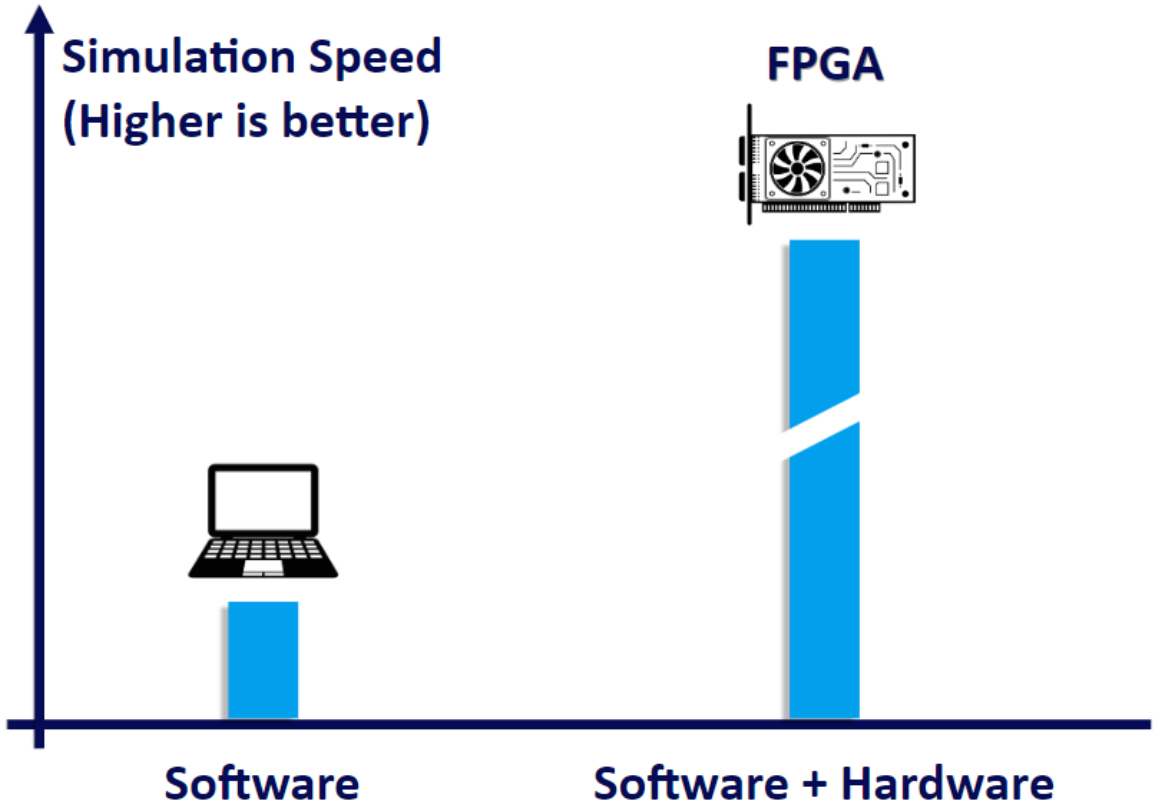
Test 1

Missing strands: 2
Errors before decoding: 15
Number of recovered strand: **20**

Test 2

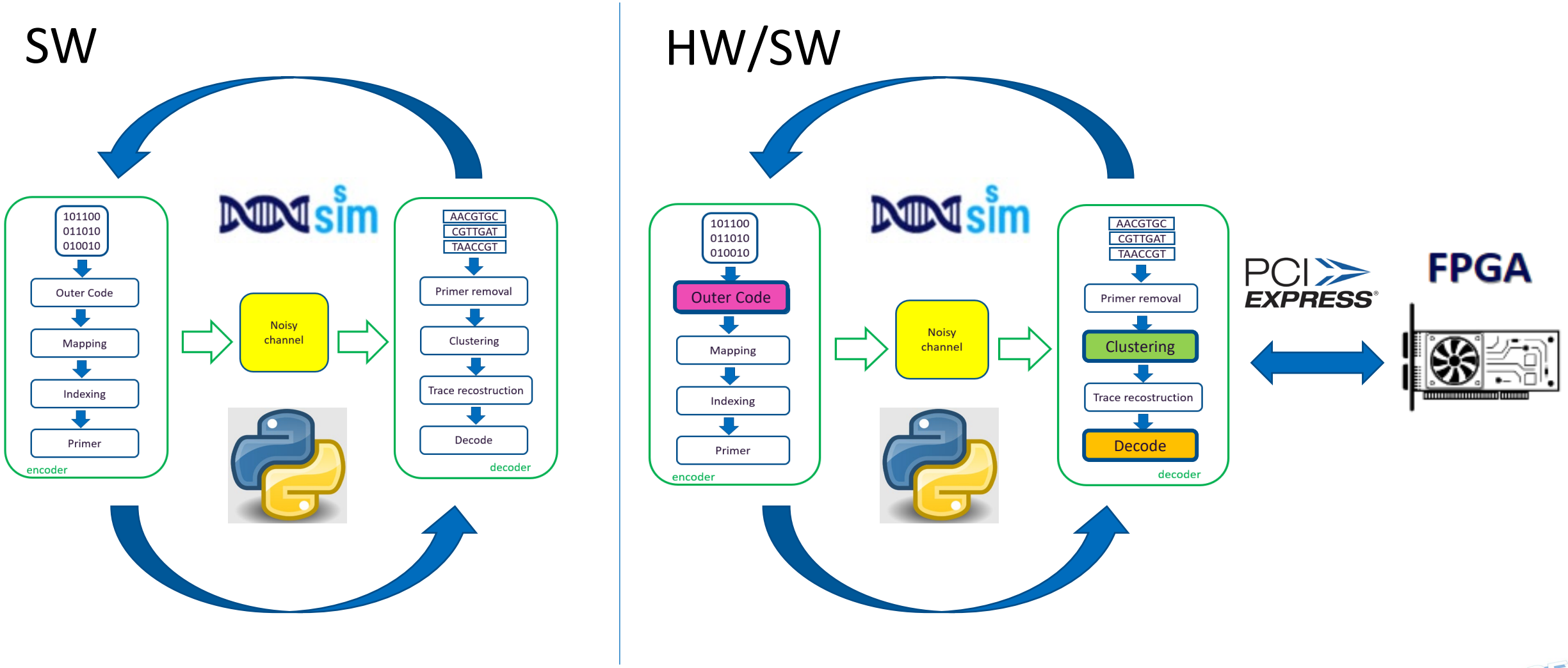
HW/SW co-simulation

Because of the number and complexity of the steps involved in the DNA storing process, the number of simulations is huge and a “pure software” simulator can easily run out of gas. To overcome this limitation, at DNAalgo we developed a custom co-simulation (i.e. mix of hardware and software) platform



https://dnaalgo.com/

HW/SW co-simulation



A photograph of the test rig used to assess the performance of the DNAssim framework. The Graphical User Interface (GUI) of the software engine and the FPGA-based hardware accelerator attached to the host motherboard are highlighted

Conclusions

- A new media is needed to store all data produced every day
- DNA storage is a promising candidate
- Encoding and Decoding involve multiple functions -> much more complicated w.r.t. Flash or HDD
- Noise channel can be modeled as a combination of PCR + IDS channel
- DNAssim is used to find the best encoding and decoding combinations tailored to a specific error model
- DNA simulations are accelerated by a combination of HW/SW (co-simulation)

THANK YOU!

