Modeling of DNA Data Storage Architectures

1 Contact Information and Supervisory Team

Main Contact	Jakob Luettgau (SRP, Inria) jakob.luettgau@inria.fr https://jakobluettgau.com/
Laboratory	Institut national de recherche en sciences et technologies du numérique (Inria) Inria Centre at Rennes University, France
Research Group	KerData (Scalable Storage for Clouds and Beyond; leader: Gabriel Antoniu) https://www.inria.fr/en/kerdata https://team.inria.fr/kerdata/
Supervisory Team	Jakob Luettgau, PhD (Inria, France), KERDATA Dominique Lavenier, DR (CNRS , France), GENSCALE François Tessier, PhD (Inria, France), KERDATA

2 Context and Overview

Data volumes are rapidly increasing and are widely expected to surpass 175 zettabytes by 2025 [2, 3]. Current technologies are struggling to both store and keep this data accessible. The reason are unfavorable data volume to access bandwidth ratios and rapidly increasing costs both for hardware and energy. In fact, data and compute infrastructure today have a notable environmental footprint: Data centers are currently accounting for about 3% of of the worlds total electricity usage [6] and are projected to double by 2026 [7]. In addition, even though both total data capacity and total data access bandwidth are continuing to increase, we are slowly loosing "surface area" to also access all this data. With no way to access all this data, it may become ultimately pointless to store unless new technologies would allow much more fine-granular random and parallel access to data. For long-term data storage, DNA-based systems [11] could overcome many existing limitations, offering very high data densities while allowing extreme levels of parallel access.

For a long time, both the cost to "read" (sequence) and "write" (synthesize) DNA have been cost and time prohibitive. But array-based nanopore sequencing devices have significantly reduced the cost of reading DNA sequencing to 0.01 cents per megabase [10, 1], with multiple techniques being well established and available as commercial products. Similarly, more cost-efficient synthesis techniques have been demonstrated, including array-based synthesis on a chip [14, 12].

Research on DNA-based data storage is concerned with the developing of parallel handling of DNA, as well as suitable encoding and management strategies that bring together benefits from digital technologies and pair them with biochemical technologies to provide data read and write functionality with DNA molecules as the storage medium. Various candidates for suitable building blocks [13, 14, 10] exist but research for highly-scalable architectures that cater to data center needs are still in their infancy.

3 Internship objectives and approach

This project aims to investigate DNA-storage architectures for HPC and cloud data centers.

1. As a first step, the student shall become familiar with the constraints of DNA-based storage systems such as imposed by sequencing and synthesis technologies
2. The student will then experiment with the simulation of different architectures for DNA storage and analyze their performance characteristics with respect to data capacity, latency, and bandwidth of the system

The student will learn state-of-the-art practices, tools, and standards, such as:

- System simulation and modeling with Wrench [15] and SimGrid [8]
- Applications of next-generation DNA sequencing and synthesis approaches and fundamental DNA encoding schemes [5, 4]
- Self-Describing Data Formats (HDF5[9], NetCDF[16])

We will build upon previous work and active research of the members in the supervisory team and international collaborators in the USA and Germany.

Required skills

- Good knowledge/understanding of computer architectures and storage systems
- Strong programming skills (Python, C/C++)
- Bonus: Familiarity with DNA fundamentals from classes on biology or bio-informatics

Languages: Proficiency in written English is required; fluency in spoken English is required.
Relational skills: The candidate will work in a research team, where regular meetings will be set up. The candidate has to be able to present the progress of their work in a clear and detailed manner.
Other values appreciated: Open-mindedness, strong integration skills, and team spirit.

Most importantly, we seek highly motivated candidates.

References

[1] DNA Sequencing Costs: Data. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

[2] Domo Resource - Data Never Sleeps 11.0. https://www.domo.com/learn/infographic/data-never-sleeps-11.

[3] Worldwide IDC Global DataSphere Forecast, 2023-2027: It's a Distributed, Diverse, and Dynamic (3D) DataSphere. https://www.idc.com/getdoc.jsp?containerId=US50554523.

[4] DNA Data Storage Sector One v1.0, November 2023.

[5] DNA Data Storage Sector Zero v1.0, November 2023.

[6] DOE: Data Centers and Servers. https://www.energy.gov/eere/buildings/data-centers-and-servers, 2024.

[7] Electricity 2024 - Analysis and Forecast to 2026. Technical report, 2024.

[8] H. Casanova. Simgrid: A toolkit for the simulation of application scheduling. In *Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid*, pages 430–437, May 2001.

[9] HDF Group. HDF5: Hierarchical Data Format, 2019.

[10] John J. Kasianowicz, Eric Brandin, Daniel Branton, and David W. Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, November 1996.

[11] Dave Landsman and Karin Strauss. The DNA Data Storage Model. *Computer*, 56(7):78–85, July 2023.

[12] Dominique Lavenier. DNA Storage: Synthesis and Sequencing Semiconductor Technologies. In *IEDM 2022 - 68th Annual IEEE International Electron Devices Meeting*, pages 1–4, San Francisco, United States, December 2022. IEEE.

[13] Frieder Mugele and Jean-Christophe Baret. Electrowetting: From basics to applications. *Journal of Physics: Condensed Matter*, 17(28):R705, July 2005.

[14] Bichlien H. Nguyen, Christopher N. Takahashi, Gagan Gupta, Jake A. Smith, Richard Rouse, Paul Berndt, Sergey Yekhanin, David P. Ward, Siena D. Ang, Patrick Garvan, Hsing-Yeh Parker, Rob Carlson, Douglas Carmean, Luis Ceze, and Karin Strauss. Scaling DNA data storage with nanoscale electrode wells. *Science Advances*, 7(48):eabi6714, November 2021.

[15] Ryan Tanaka, Rafael Ferreira Da Silva, and Henri Casanova. Teaching Parallel and Distributed Computing Concepts in Simulation with WRENCH. In *2019 IEEE/ACM Workshop on Education for High-Performance Computing (EduHPC)*, pages 1–9, Denver, CO, USA, November 2019. IEEE.

[16] Unidata. NetCDF: Network Common Data Form, 2019.