RETROSPECTIVE: DNA-based Molecular Architecture with Spatially Localized Components

Luis Ceze*, Georg Seelig*, Karin Strauss†
*University of Washington
†Microsoft Research

I. Context and Motivation: Bring Architecture Principles to Molecular Computing

Performing computation inside living cells offers life-changing applications, from improved medical diagnostics to better cancer therapy to intelligent drugs. Due to its biocompatibility and ease of engineering, one promising approach for performing in-vivo computation is DNA strand displacement.

The key contribution of this paper was practical spatial isolation of components, leading to more easily designed DNA-based circuits. The concept is the molecular equivalent of a breadboard or a gate array, where individual components forming logic gates can be arbitrarily "plugged in" and connected into logical circuits on a 2D structure. Prior approaches relied mostly on stochastic interaction of freely diffusing components – which poses challenges to composability, performance, and efficient use of resources.

Apart from the actual key innovation of spatial organization in DNA strand displacement, our goals in submitting this paper to ISCA were to (1) expose our computer architecture community to molecular computing, a concept that had been, until then, mostly constrained to the DNA nanotechnology community; and (2) to show that many computer architecture principles apply quite nicely to molecular computing. In the paper, we covered several analogies between electronics and molecular circuits as examples. In fact, even the spatially localized components idea drew from electronics concepts such as the layout of circuits on a surface, reducing the problem from all-to-all possible interactions (allowed in free-flow "3D" solution, where components are diffused in a fluid and move freely) to a constrained set of allowed interactions prescribed by the geometry of a pre-defined structure (leveraging its 2D nature).

II. Evolution: From molecular computing to storage to storage+computing

While this work focused on DNA nanotechnology, composable DNA computation, in-vivo therapies and in-vitro diagnostics, very impactful areas about which we were very excited, it did not generate the flurry of activity we expected in the computer architecture community.

However, this work primed us to think differently and brought us awareness of accelerating advances in DNA reading and writing capabilities, which pointed towards the possibility of storing large amounts of digital data in DNA with viable throughput. Ultimately, this led us down a path we had initially not expected, even if its application was more aligned with information technology: DNA data storage, or storing digital data in synthetic DNA, an attractive medium for archival storage purposes [1].

Interestingly, one of the key differences between DNA data storage and other types of storage is the ability to dissociate the written "bits" from the physical structure of the substrate used to write them. Again, the 2D vs. 3D contrast is interesting: dissociating the bits from the substrate structure (2D) allows molecules representing digital data to be packed tightly (in 3D), independent from fabrication limitations of the substrate. This results in much denser data storage than if DNA molecules remained attached to the 2D substrate.

This high density, along with great durability under easy-to-maintain environmental conditions, the unwavering relevance of DNA as media (due to its importance in life sciences), and the prospect of better environmental sustainability made it a promising approach for long-term data storage [2]. By working with amazing collaborators across several disciplines, we approached it from an end-to-end perspective, resulting in a series of contributions all the way from theoretical computer science, with new encoding, decoding, and error correction algorithms [3], [4], to mechanical and electrical engineering for process automation [5], [6], [7], physical storage and random access [8], and DNA synthesis [9], [10] (the process that implements writes in DNA data storage), to molecular biology and chemistry for chemical random access [11], [12], [13] and preservation techniques [14], [15], [16].

The DNA data storage work has brought us back full circle to performing limited forms of computations with data in DNA, directly in molecular form, when we showed how to implement image similarity search in DNA [17]. This approach does not use DNA gates like those in DNA strand displacement, but shares their underlying concept of hybridization – bases in opposite sides of the DNA double helix attract each other (A pairs with T, C with G). We developed the idea of mapping feature vectors to DNA sequences such that the probability of DNA strands sticking to each other is proportional to the similarity of the feature vectors.

Ironically, the key aspect of making similarity search work well in molecular form is to not have any spatial localization at all (departing from a 2D substrate) and in fact, quite the opposite, to rely on diffusion of the molecules in a fluid to increase molecular interactions needed to perform the search. Here, the computer architecture analogy is from in-memory

processing to at-molecule processing. Diffusion creates space between the molecules representing data items such that "queries" may approach and interact with them – if they bind with a query, which is attached to a magnetic nanoparticle, the data item can then be retrieved using a magnet.

This is an attractive architectural aspect because one can adjust the access bandwidth to capacity ratio dynamically. Unlike electronic memories, where the access bandwidth to capacity ratio is fixed at fabrication time – e.g., wires in solid-state memories, or optical/magnetic heads in disks. Hence the throughput of this search access operation implemented in DNA is proportional to its volume. This is true processing in memory, where the processing diffuses through the physical embodiment of the memory elements.

III. The Future

After we started working on DNA data storage, molecular data storage research [18] has been receiving significant attention from governments and industry: Multiple large projects were funded by government agencies and industrial consortia. Programs in which we participated in the U.S. include those by DARPA, I-ARPA, and SRC, but we are aware of targeted programs in other countries as well, such as UK, France, Germany, and Brazil.

On the commercial side, an ecosystems of companies formed by a large set of major IT and biotechnology companies, as well as a number of startups, is actively working on DNA data storage. Finally, a few years ago, the DNA Data Storage Alliance was formed to support this ecosystem by socializing the technology, exploring usage scenarios, and ensuring interoperability. Although data storage is a challenging market where incumbent technologies tend to have an upper hand due to the magnitude of sunken investment, we believe that DNA data storage offers an attractive point in the digital data storage space and are hopeful that it will reach commercial viability.

IV. Acknowledgements

The work that happened downstream from this effort benefited from significant effort of over 50 researchers and engineers, as well as immense institutional support. We thank Microsoft Research, DARPA, I-ARPA, the National Science Foundation, Oxford Nanopore Technologies and Twist Bioscience for their support. We especially thank Anne Fischer and David Markowitz for their program leadership. We thank Ed Lazowska and Craig Mundie for earmarked donations for the construction of the Molecular Information Systems Lab, which has sparked many collaborative efforts in the intersection of biotechnology and computing. We also thank Doug Carmean for his support, as well as many strategic and intellectual contributions. Finally, we thank the ISCA community for welcoming this kind of more adventurous work that takes decades to show impact.

References

[1] J. Bornholt, R. Lopez, K. Strauss, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," in *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, April 2016, selected for IEEE Micro Top Picks 2016.

[2] B. Nguyen, J. Sinistore, J. Smith, P. S. Arshi, L. M. Johnson, T. Kidman, T. J. diCaprio, D. Carmean, and K. Strauss, "Architecting datacenters for sustainablity: Greener data storage using synthetic DNA," in *Electronic Goes Green*, September 2020.

[3] C. Rashtchian, K. Makarychev, M. Racz, S. Ang, D. Jevdjic, S. Yekhanin, L. Ceze, and K. Strauss, "Clustering billions of reads for DNA data storage," in *NeurIPS*, November 2017.

[4] R. Lopez, Y.-J. Chen, S. D. Ang, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Seelig, K. Strauss, and L. Ceze, "DNA assembly for nanopore data storage readout," *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[5] M. Willsey, A. P. Stephenson, C. Takahashi, P. Vaid, B. H. Nguyen, M. Piszczek, C. Betts, S. Newman, S. Joshi, K. Strauss, and L. Ceze, "Puddle: A dynamic, error-correcting, full-stack microfluidics platform," in *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2019.

[6] S. Newman, A. P. Stephenson, M. Willsey, B. H. Nguyen, C. N. Takahashi, K. Strauss, and L. Ceze, "High density DNA data storage library via dehydration with digital microfluidic retrieval," *Nature Communications*, vol. 10, no. 1, p. 1706, 2019.

[7] C. N. Takahashi, B. H. Nguyen, K. Strauss, and L. Ceze, "Demonstration of end-to-end automation of DNA data storage," *Scientific Reports*, vol. 9, no. 1, p. 4998, 2019.

[8] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. N. Takahashi, S. Newman, H.-Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze, and K. Strauss, "Random access in large-scale DNA data storage," *Nature Biotechnology*, 2018. [Online]. Available: http://www.nature.com/doifinder/10.1038/nbt.4079

[9] B. H. Nguyen, C. N. Takahashi, G. Gupta, J. A. Smith, R. Rouse, P. Berndt, S. Yekhanin, D. P. Ward, S. D. Ang, P. Garvan, H.-Y. Parker, R. Carlson, D. Carmean, L. Ceze, and K. Strauss, "Scaling DNA data storage with nanoscale electrode wells," *Science Advances*, vol. 7, 2021.

[10] J. Smith, B. Nguyen, R. Carlson, J. G. Bertram, S. Palluk, D. H. Arlow, and K. Strauss, "Spatially selective electrochemical cleavage of a polymerase-nucleotide conjugate," *ACS Synthetic Biology*, May 2023.

[11] L. Organick, Y.-J. Chen, S. D. Ang, R. Lopez, X. Liu, K. Strauss, and Ceze, "Probing the physical limits of reliable DNA data retrieval," *Nature Communications*, vol. 11, no. 1, pp. 1–7, 2020.

[12] C. Winston, L. Organick, D. Ward, L. Ceze, K. Strauss, and Y.-J. Chen, "A combinatorial PCR method for efficient, selective oligo retrieval from complex oligo pools," *ACS Synthetic Biology*, February 2022.

[13] B. W. A. Bögels, B. Nguyen, D. Ward, L. Gascoigne, D. P. Schrijver, A.-M. M. Pistikou, A. Joesaar, S. Yan, I. K. Voets, W. J. M. Mulder, A. Phillips, S. Mann, G. Seelig, K. Strauss, Y.-J. Chen, and T. F. A. deGreef, "DNA storage in thermoresponsibe microcapsules for repeated random multiplexed data access," *Nature Nanotechnology*, May 2023.

[14] W. D. Chen, A. X. Kohll, B. Nguyen, J. Koch, R. Heckel, W. J. Stark, L. Ceze, K. Strauss, and R. N. Grass, "Combining data longevity with high storage capacity – layer-by-layer DNA encapsulated in magnetic nanoparticles," *Advanced Functional Materials*, May 2019.

[15] A. X. Kohll, P. L. Antkowiak, W. D. Chen, B. Nguyen, W. J. Stark, L. Ceze, K. Strauss, and R. N. Grass, "Stabilizing synthetic DNA for long-term data storage with earth alkaline salts," *Chemical Communications*, February 2020.

[16] L. Organick, B. Nguyen, R. McAmis, W. D. Chen, A. X. Kohll, S. D. Ang, R. N. Grass, L. Ceze, and K. Strauss, "An empirical comparison of preservation methods for synthetic DNA data storage," *Small Methods*, January 2021.

[17] C. Bee, Y.-J. Chen, M. Queen, D. Ward, X. Liu, L. Organick, G. Seelig, K. Strauss, and L. Ceze, "Molecular-level similarity search brings computing to DNA data storage," *Nature Communications*, vol. 12, pp. 1–9, 2021.

[18] L. Ceze, J. Nivala, and K. Strauss, "Molecular digital data storage using DNA," *Nature Reviews Genetics*, p. 1, 2019.