Can We Store the Whole World's Data in DNA Storage?

HotStorage'20

Bingzhe Li, Nae Young Song, Li Ou, and David H.C. Du
University of Minnesota, Twin Cities

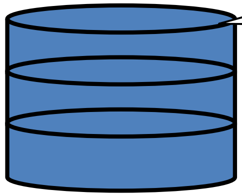
Center for Research in Intelligent Storage

University of Minnesota

Outlines

- Motivation
- DNA background
- Contributions
  - Trade-offs in DNA storage
  - DNA storage modeling
  - How many tubes to store the whole world’s data?
- Indexing scheme
- Conclusion

Big Data Era



Data is doubled almost every **2 years**
44 Zettabytes in 2020
175 Zettabytes in 2025

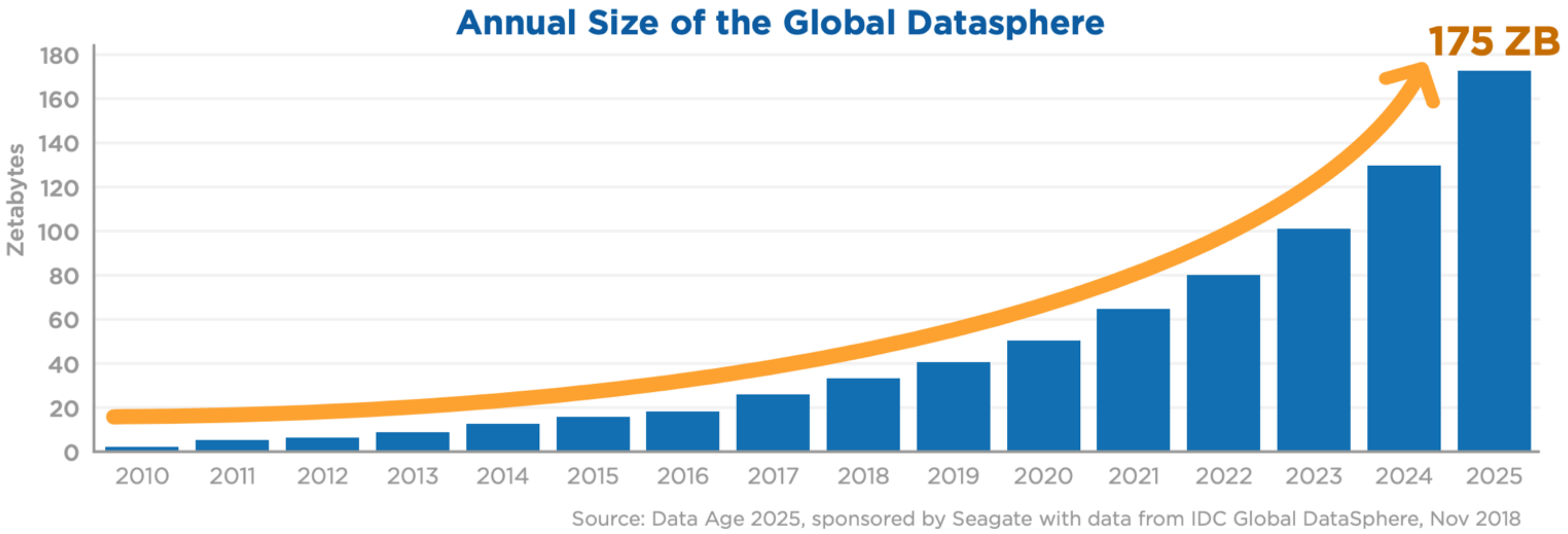
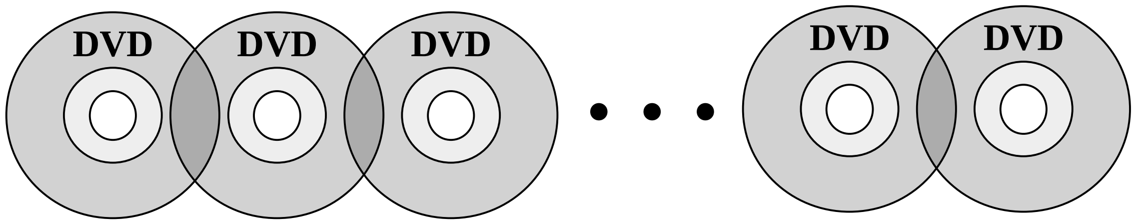


Image from: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

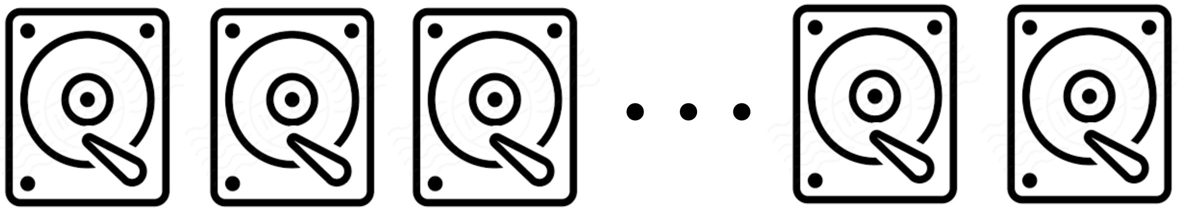
How to Store these Data?

44 ZB

50 trillion DVD movies



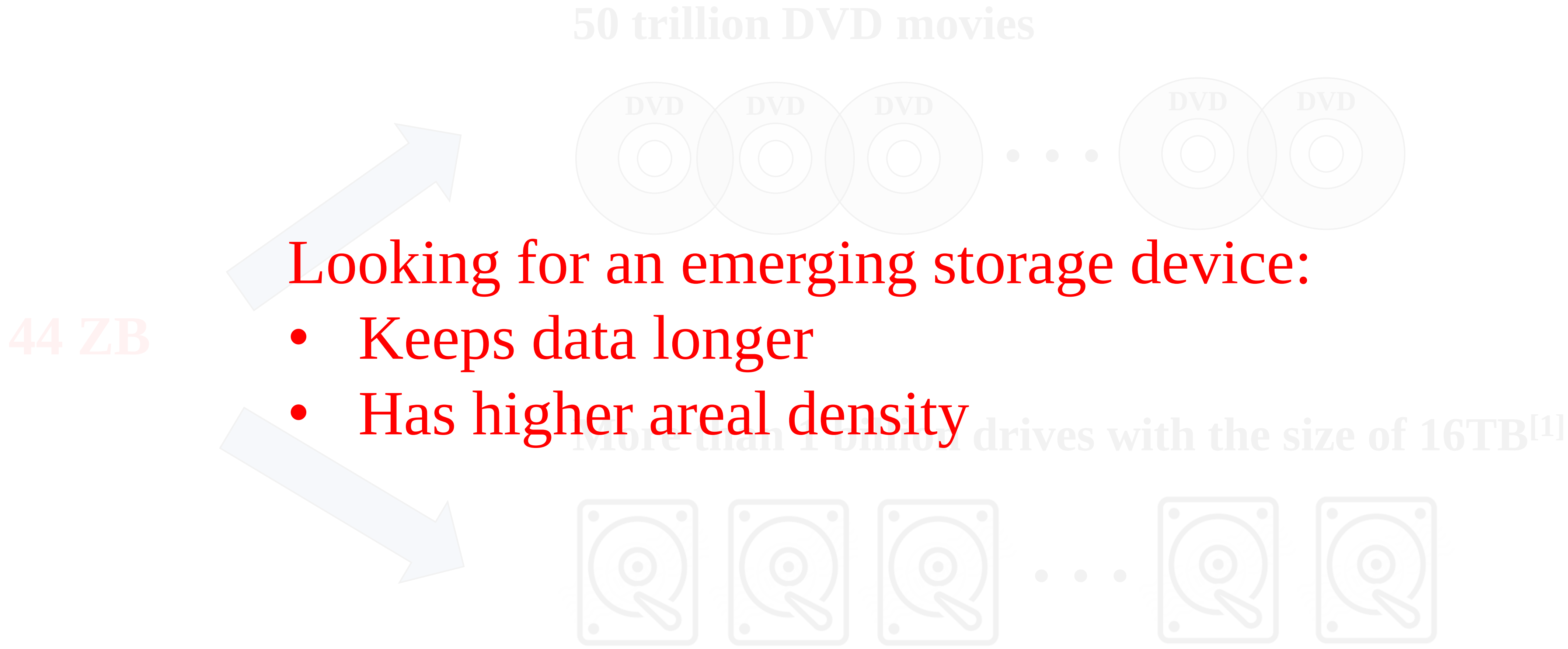
More than 1 billion drives with the size of 16TB[1]



5 years or 10 years warranty

[1] https://www.seagate.com/enterprise-storage/

How to Store these Data?



Looking for an emerging storage device:

- Keeps data longer
- Has higher areal density

[1] *https://www.seagate.com/enterprise-storage/*

DNA Storage

- High spatial density
  - A theoretical density of 455 EB/g [1]

- Long persistency
  - Several centuries [2][3]

[1] Raja Appuswamy, Kevin Le Brigand, Pascal Barbry, Marc Antonini, Olivier Madderson, Paul Freemont, James McDonald, and Thomas Heinis. Oligoarchive: Using dna in the dbms storage hierarchy. In CIDR, 2019.
[2] Morten E Allentoft, Matthew Collins, David Harker, James Haile, Charlotte L Oskam, Marie L Hale, Paula F Campos, Jose A Samaniego, M Thomas P Gilbert, Eske Willerslev, et al. The half-life of dna in bone: measuring decay kinetics in 158 dated fossils. Proceedings of the Royal Society B: Biological Sciences, 279(1748):4724–4733, 2012.
[3] Robert N Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, and Wendelin J Stark. Robust chemical preservation of digital information on dna in silica with error-correcting codes. Angewandte Chemie International Edition, 54(8):2552–2555, 2015.

Background of DNA storage

- **Nucleotides**: molecules form the building blocks of DNA.
  - Adenine (A) ←→ Thymine (T)
  - Cytosine (C) ←→ Guanine (G)

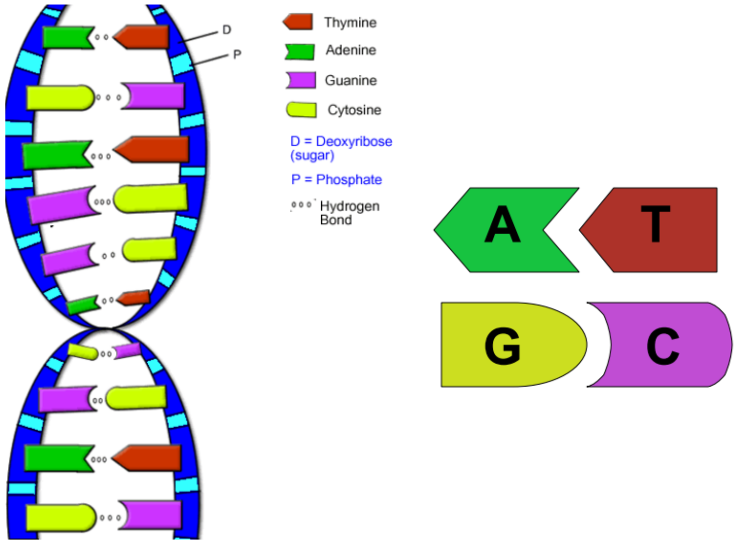


Figure 1

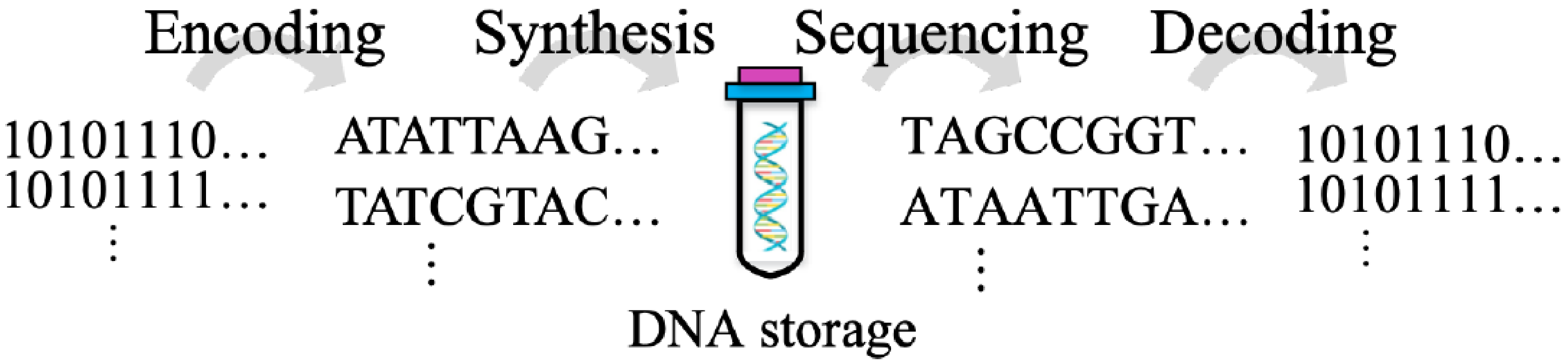
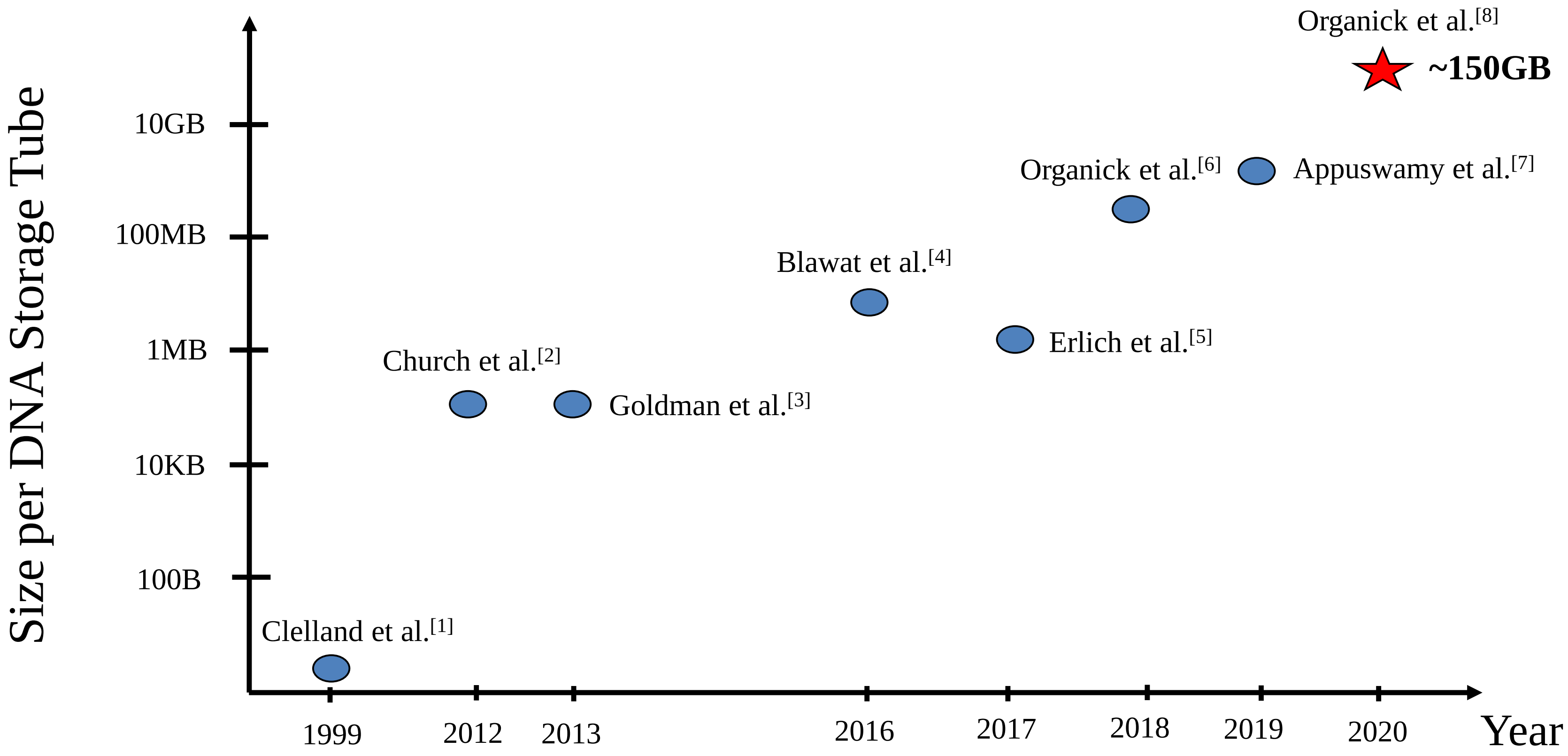


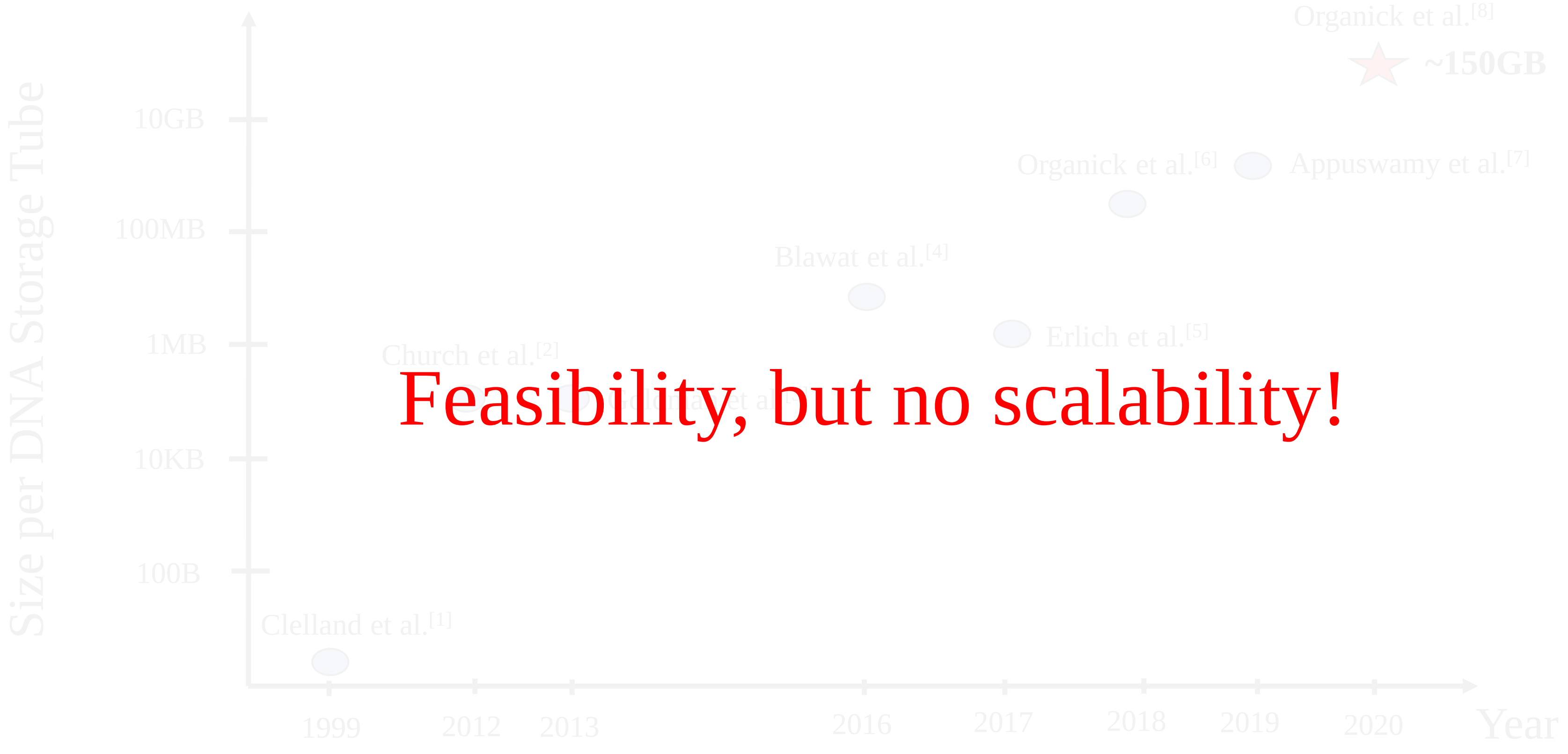
Figure 1 from https://www.genome.gov/Pages/Education/Modules/BasicsPresentation.pdf

Existing Work



[1] Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. Nature 399, 533–534 (1999).
[2] Church, G. M., Gao, Y. & Kosuri, S. Next- generation digital information storage in DNA. Science ***337****, 1628–1628 (2012)*
[3] Goldman, N. et al. Towards practical, high- capacity,low- maintenance information storage in synthesized DNA. Nature 494, 77–80 (2013).
[4] Blawat, M. et al. Forward error correction for DNA data storage. Procedia Comput. Sci. 80, 1011–1022 (2016)
[5] Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. Science 355, 950–954 (2017).
[6] Organick, L. et al. Random access in large- scale DNA data storage. Nat. Biotechnol. 36, 242–248 (2018).
[7] Appuswamy, Raja, et al. "OligoArchive: Using DNA in the DBMS storage hierarchy." CIDR. 2019.
[8] Organick, Lee, et al. "Probing the physical limits of reliable DNA data retrieval." Nature communications 11.1 (2020): 1-7.

Existing Work



Feasibility, but no scalability!

[1] Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. Nature 399, 533–534 (1999).
[2] Church, G. M., Gao, Y. & Kosuri, S. Next- generation digital information storage in DNA. Science 337, 1628–1628 (2012)
[3] Goldman, N. et al. Towards practical, high- capacity,low- maintenance information storage in synthesized DNA. Nature 494, 77–80 (2013).
[4] Blawat, M. et al. Forward error correction for DNA data storage. Procedia Comput. Sci. 80, 1011–1022 (2016)
[5] Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. Science 355, 950–954 (2017).
[6] Organick, L. et al. Random access in large- scale DNA data storage. Nat. Biotechnol. 36, 242–248 (2018).
[7] Appuswamy, Raja, et al. "OligoArchive: Using DNA in the DBMS storage hierarchy." CIDR. 2019.
[8] Organick, Lee, et al. "Probing the physical limits of reliable DNA data retrieval." Nature communications 11.1 (2020): 1-7.

Our Contributions

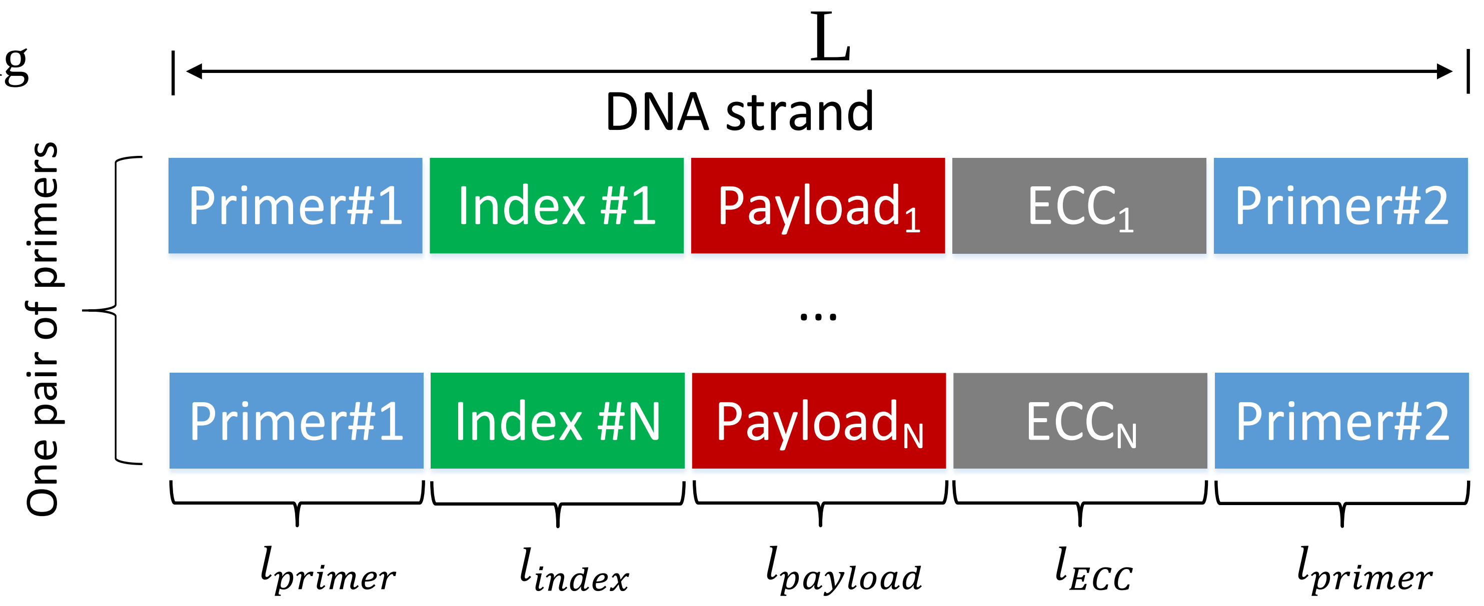
- Investigate the effect of different factors on the capacity of DNA storage (in-house simulator)
- Analyze trade-offs between different factors and scalability of DNA storage
- How to index the whole world's data in DNA storage

Factors and Modeling of DNA Storage

- **Primer**: is used to read data out (sequencing process based on Polymerase Chain Reaction (PCR))
- **Index**: distinguishes DNA strands in the same primer pair
- **Payload**: useful information
- **ECC**: corrects errors from synthesis and sequencing processes

- **PF (primer factor)**: **N** DNA strands attached to the same primer pair
- **Coding density**: useful information (bit)
  $Info = coding\ density\ * l_{payload}$

- **Solubility**
- **Droplet volume**
- **...**



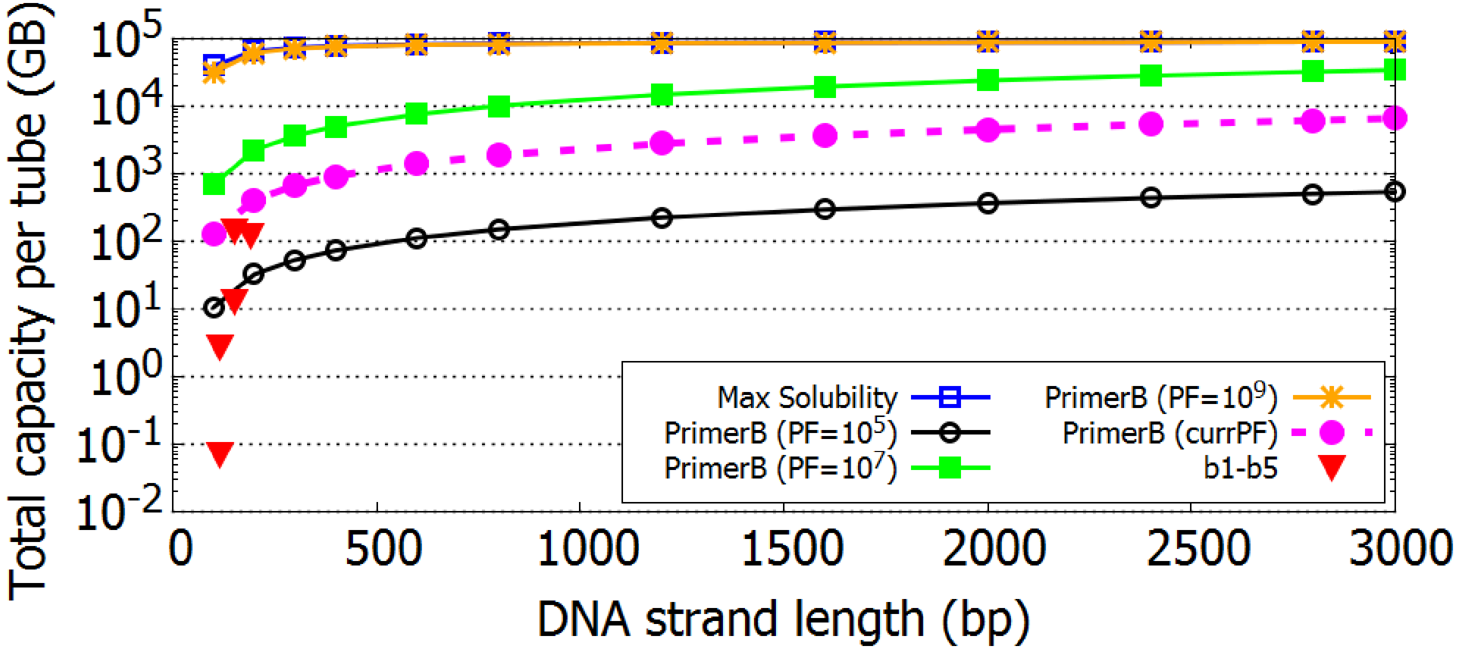
$$L = l_{primer} * 2 + l_{index} + l_{payload} + l_{ECC}$$

(l_{primer} is about 18 – 25 bp)

DNA Storage Trade-offs: varying DNA length (L)

L ↑ ⟹ l_{ECC} ↑ $l_{payload}$? l_{index} →

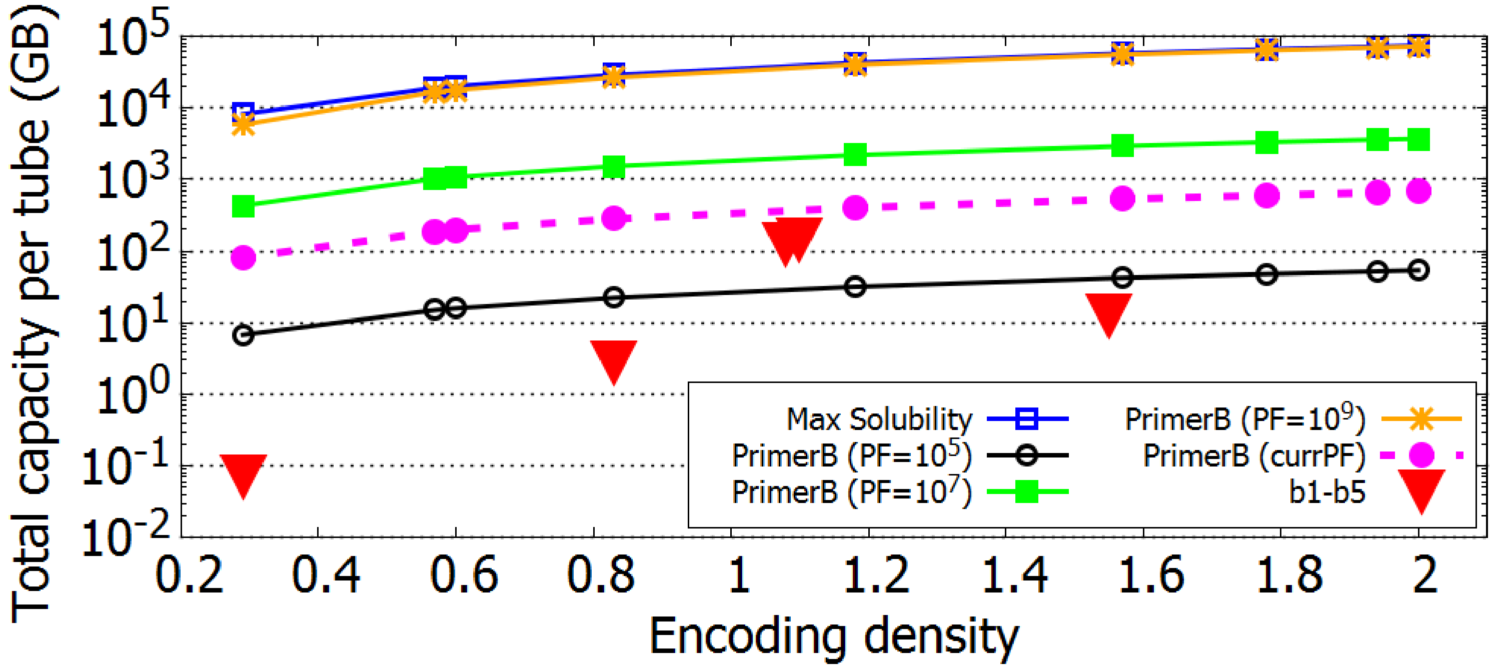
L = 100 – 3000 bp



DNA Storage Trade-offs: varying coding density

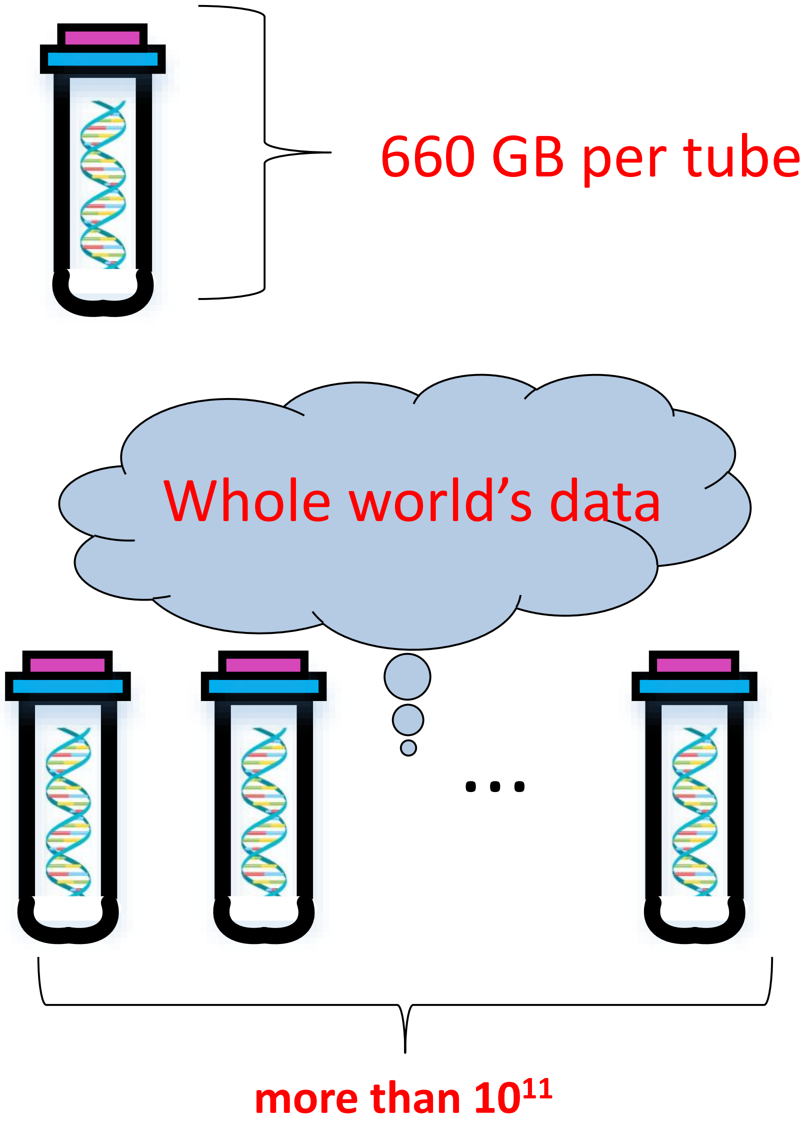
Coding density ↑ ⇒ l_{ECC} ↑ $l_{payload}$ ↓ l_{index} →

Coding = 0.29 - 2



Store the Whole World's Data based on Today's Technology

Factor	Value
Whole world's data (ZB)	44
DNA Strand Length(bp)	300
Primer length (bp)	20
Coding density	1
ECC	15%
Tube size (mL)	1.7
Max DNA solubility in liquid (mg/mL)	500
Droplet size (mL)	0.001
PF	1.55*10E6



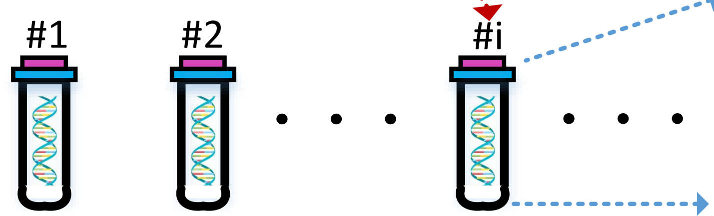
Center for Research in Intelligent Storage

University of Minnesota

DNA Storage Indexing

Block-based storage device:

Request OFFSET mod $capacity_{tube}$



len

External index:

$offset_1$	Primer pair #1
…	…
$offset_i$	Primer pair #i
…	…
$offset_M$	Primer pair #M

len * # of entries * # of tube ~77 TB

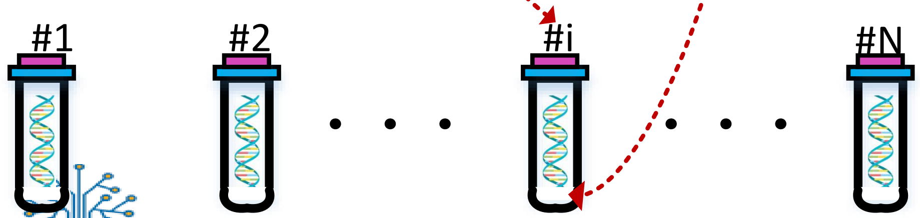
Object-based storage device:

len

External index:

ID_1	Tube #1	Primer pair #1	$Index_{start1}$	$Index_{end1}$
…	…	…	…	…
ID_i	Tube #i	Primer pair #j	$Index_{starti}$	$Index_{endi}$
…	…	…	…	…
ID_M	Tube #N	Primer pair #M	$Index_{startM}$	$Index_{endM}$

Request Global ID/Key



len * # of IDs ~7.17*10^5 TB

Indexing for the object-based storage device is more challenging

Conclusion

- Modeling of DNA storage based on different factors
- Investigate the trade-offs between different factors
- Scalability of DNA storage
- Introduce simple schemes to index the whole world’s data in DNA storage

Thanks!

lixx1743@umn.edu