Seminar Arbeit
Eingebettete Systeme und Robotik
im Sommersemester 2025

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Potenzial und Herausforderung der DNA-Speicherung

Philipp Jonathan Mohorko
Rheinland-Pfälzische Technische Universität Kaiserslautern, Fachbereich Informatik

Note: *This report is a compilation of publications related to some topic as a result of a student seminar. It does not claim to introduce original work and all sources should be properly cited.*

Abstract *Die stetig wachsende Datenmenge stellt bestehende Rechenzentren zunehmend vor große Herausforderungen in Bezug auf Speicherkapazität, Energieverbrauch und Nachhaltigkeit. Konventionelle Speichertechnologien wie HDDs, SSDs oder Magnetbänder stoßen hierbei an physikalische und ökonomische Grenzen. Ein innovativer Ansatz zur Lösung dieser Probleme ist die Speicherung digitaler Informationen in DNA-Molekülen. DNA bietet eine extrem hohe Speicherdichte, eine lange Haltbarkeit und benötigt keine permanente Energiezufuhr, wodurch sie sich insbesondere für die Archivierung sogenannter* kalter Daten *eignet. Diese Arbeit untersucht die theoretischen Grundlagen der DNA-Speicherung, vergleicht sie mit bestehenden Speichertechnologien und diskutiert mögliche Architekturen für Rechenzentren, die DNA-Speicher integrieren. Abschließend werden die Leistungsfähigkeit, die aktuellen Einschränkungen sowie die langfristigen Potenziale dieser Technologie bewertet.*

1 Einleitung

Die Digitalisierung schreitet in nahezu allen Lebensbereichen rasant voran. Anwendungen wie Cloud-Dienste, Soziale Medien, Datenarchivierungen, bei denen Informationen über viele Generationen hinweg erhalten bleiben müssen, oder autonomes Fahren erzeugen täglich riesige Datenmengen. Rechenzentren bilden dabei das Rückgrat moderner Informationssysteme, stoßen mit klassischen Speicherlösungen jedoch zunehmend an ihre Grenzen.

Neben der steigenden Nachfrage nach Speicherkapazität erhalten auch Energieeffizienz, Langlebigkeit und Datensicherheit eine immer größere Bedeutung. Dadurch wächst das Interesse an alternativen Technologien, die neue Möglichkeiten für die Speicherung bieten könnten.

Einer dieser vielversprechenden Alternativen ist die DNA-Speicherung. Als Trägermolekül genetischer Informationen bietet die DNA Eigenschaften, die sie potenziell zu einem neuartigen Speichermedium machen könnte. Im Rahmen dieser Arbeit werden die Grundlagen dieser Technologie betrachtet und ihre Chancen sowie Herausforderungen im Kontext modernener Rechenzentren diskutiert.

Zunächst werden in Kapitel 2 die theoretischen Grundlagen gelegt, indem wesentliche biologische und informationstechnische Konzepte erläutert werden. Im Anschluss folgt in Kapitel 3 eine Analyse bestehender Rechenzentren, die sowohl konventionelle Speichertechnologien als auch Netzwerkarchitekturen und ihre zentralen Herausforderungen umfasst. Darauf aufbauend rückt Kapitel 4 die DNA-Speicherung in den Fokus: Es beschreibt das Grundprinzip, zentrale Eigenschaften sowie aktuelle Einschränkungen dieser Technologie. Ein möglicher Integrationsansatz in Form hybrider Rechenzentrumsstrukturen wird in Kapitel 5 vorgestellt und hinsichtlich Chancen und Grenzen bewertet. Die Leistungsfähigkeit der DNA-Speicherung wird in Kapitel 6 durch Kosten- und Effizienzvergleiche untersucht. Den Abschluss bildet Kapitel 7, das die Ergebnisse zusammenführt und ein Fazit zieht.

2 Theoretische Grundlagen

Um die Potenziale und Grenzen der DNA-Speicherung zu verstehen, ist es zunächst notwendig, sich die Grundlagen dieser Technologie vor Augen zu führen:

2.1 DNA:

Die Desoxyribonukleinsäure (DNS), ist ein organisches Molekül, welches das Erbgut aller Lebewesen trägt. Das Molekül besteht aus langen Ketten von Nukleotiden, bestehend aus den vier Basen **Adenin (A), Thymin (T), Guanin (G)** und **Cytosin (C)**. Diese Sequenzen ermöglichen eine extrem hohe Informationsdichte: Theoretisch kann ein Gramm DNA bis zu 455 Exabyte (455 * 10^{18} Bytes) an Daten speichern, somit ist die DNA herkömmlichen Speichermedien um mehrere Größenordnungen überlegen [4, 6].

Der Aufbau der DNA ist eine Doppelhelix, in der die Basen Adenin und Thymin sowie Guanin und Cytosin jeweils durch Wasserstoffbrücken miteinander verbunden sind. Die Struktur sorgt für Stabilität und Langlebigkeit, weshalb die DNA ideal für langfristige Datenspeicherung ist. Im Vergleich zu elektronischen Speichertechnologien ist die DNA resistenter gegen Umweltfaktoren wie extreme Temperaturen, Strahlung und Feuchtigkeit. Zudem benötigt DNA keine Energie zum Erhalt der gespeicherten Daten, was die Betriebskosten deutlich reduziert.

Ein wichtiger Schritt bei der Nutzung von DNA zur Datenspeicherung ist die sogenannte **DNA-Synthese**, der zentrale Schreibprozess, bei dem digitale Informationen in die Basenabfolge von DNA übersetzt werden.
Die Synthese ist aktuell der zeit- und kostenintensivste Schritt der DNA-Speicherung, bildet jedoch die Voraussetzung dafür, digitale Informationen überhaupt dauerhaft in molekularer Form ablegen zu können. Dieses Thema und dessen Folgen werden in **Kapitel 4.3** erneut aufgegriffen und mit den anderen Verfahren des DNA-Speicherungsprozesses verknüpft [3, 6].

Ergänzend zur DNA-Synthese, die den Schreibvorgang der Daten darstellt, ist die
DNA-Sequenzierung der zentrale Leseprozess der DNA-Speicherung. Dabei handelt es sich

um ein Verfahren zur Bestimmung der Basenabfolge der DNA-Stränge, mit deren Hilfe die gespeicherten Daten wieder in digitale Form überführt werden können. Dieser Prozess ist essenziell, um die gespeicherten Daten wieder zugänglich zu machen und bildet die Grundlage für das Dekodieren der in DNA gespeicherten Informationen [3, 6].

2.2 Shuffle-Stambling:

Beim Begriff „Shuffle-Stambling“ handelt es sich um ein Modell aus dem Bereich der DNA-Datenspeicherung, das die zufällige Neuordnung (Shuffling) der DNA-Stränge beschreibt, wenn diese im Speichersystem abgelegt werden. Anders als bei klassischen Speichermedien, bei denen die Reihenfolge der Daten fest vorgegeben ist, sind bei der DNA-Speicherung die einzelnen Stränge, weder beim Lesen noch beim Schreiben der Daten nicht in einer bestimmten Reihenfolge fest vorgegeben. Dadurch entsteht eine sogenannte „ungeordnete“ Anordnung, bei der die ursprüngliche Reihenfolge verloren gehen kann.

Um die ursprüngliche Reihenfolge wiederherzustellen, werden Indexe oder Barcodes verwendet. Somit ist das Shuffle-Stambling-Modell wichtig, um effiziente Algorithmen und Fehlerkorrekturverfahren zu entwickeln, die trotz Zufallslayouts eine zuverlässige Datenwiederherstellung ermöglichen [9].

2.3 Fat-tree-Architektur:

Die „Fat-tree“-Architektur ist eine hierarchische Netzwerkstruktur, die in modernen Rechenzentren verwendet wird, um hohe Bandbreiten und Skalierbarkeit, also die Fähigkeit, das System bei steigenden Datenmengen ohne grundlegende Strukturänderungen effizient erweitern zu können, zu gewährleisten. Dabei sind die Verbindungen in den oberen Ebenen dicker, um den Datenverkehr effizient zu verteilen und Engpässe zu vermeiden. Diese Topologie unterstützt den gleichmäßigen Datenfluss, hohe Ausfallsicherheit und eine einfache Erweiterbarkeit der Netzwerkstruktur.
Für DNA gestützte Speicherlösungen in großen Rechenzentren kann die „Fat-tree“-Architektur eine effiziente Netzwerkverbindung zwischen biologischen und netzwerkbasierten Komponenten sicherstellen [12].

2.4 Reed-Solomon-Code:

Das Reed-Solomon-Code-Verfahren ist ein Fehlerkorrekturverfahren, welches die Daten als Polynome [*Mathematischer Ausdruck, der aus Variablen (z.B. x) und deren Potenzen besteht, die mit Zahlen multipliziert und addiert werden, z.B.* $2x^2 + 3x + 1$] über endliche Körper [*Eine Menge mit endlich vielen Elementen, in der die vier Grundrechenarten (Addition, Subtraktion, Multiplikation und Division außer durch Null) definiert sind und die Abgeschlossenheit erfüllen*] darstellt und dort noch zusätzliche Prüfsymbole (Redundanz) anhängt.
Auf diese Weise können sowohl mehrere aufeinanderfolgende Bit- und Symbolfehler als auch Verluste ganzer Datenblöcke erkannt und korrigiert werden. Eine Voraussetzung hierfür ist jedoch, dass die Zahl der Fehler die Korrekturkapazität nicht überschreitet. Eingesetzt wird dieses

Verfahren häufig in Speichermedien, CDs/ DVDs und Kommunikationssystemen, um die Datenintegrität auch bei hohen Fehlerraten zu sichern [10].

2.5 Kalte Daten und heiße Daten:

Der Begriff ***kalte Daten*** bezeichnet Daten, die nur selten oder gar nicht aktiv genutzt, aber dennoch langfristig gespeichert werden müssen. Dabei handelt es sich beispielsweise um Daten zu Dokumentations- oder Archivierungszwecken.
Ihr Gegenstück sind die sogenannten ***heißen Daten***, auf die regelmäßig und mit einer geringen Latenzzeit zugegriffen werden muss. Unter Latenzzeit versteht man die Verzögerung zwischen der Eingabe einer Anfrage bzw. eines Befehls und dem Beginn der Antwort oder Ausführung im System. Bei den *heißen Daten* handelt es sich zum Beispiel um einen Datenbankzugriff oder etwaige Cloud-Anwendungen.
Während *heiße Daten* eher auf schnellen, aber energieintensiven Speichermedien wie SSDs (vgl. Abschnitt 3.1) abgelegt werden, eignen sich für *kalte Daten* eher kostengünstige, speicherdichte und langlebige Technologien, da hier die Zugriffsgeschwindigkeit eine eher geringere Rolle spielt und im Vergleich die Haltbarkeit, Speicherdichte und der Energieverbrauch eine höhere Bedeutung erhält.

3 Bestehende Rechenzentren

Rechenzentren sind das Herzstück der digitalen Welt. Sie ermöglichen die Speicherung, Verarbeitung und den Austausch riesiger Datenmengen. Doch mit dem rapiden Wachstum der Datenmenge stehen diese Einrichtungen vor immer größeren Herausforderungen. Besonders mit dem Aufkommen datenintensiver Anwendungen wie dem autonomen Fahren, der Telematik, also die Kombination von Telekommunikation und Informatik zur Erfassung, Übertragung und Verarbeitung von Daten, oder der Cloud-Kollaboration stoßen herkömmliche Speicherlösungen zunehmend an ihre Grenzen.

Dieses Kapitel gibt einen Überblick über die aktuellen Strukturen der Rechenzentren und deren Herausforderung, mit der modernen Welt mitzuhalten.

3.1 Konventionelle Speichertechnologien in Rechenzentren

Dieses Unterkapitel basiert vollständig auf den Ausführungen von Qu, Wu und Li [7].
In den meisten herkömmlichen Rechenzentren greift man auf verschiedene zentrale Speicherkomponenten zurück, die sich hinsichtlich der Kapazität, dem Energieverbrauch und der Zugriffszeit unterscheiden. Die wichtigsten respektive konventionellsten Speicherkomponenten sind die folgenden:

3.1.1) Festplattenlaufwerke (HDDs: Hard Disk Drives)

Mechanische Festplatten zeichnen sich durch große Kapazitäten bei niedrigen Kosten aus.

Ihre Funktionsweise beruht auf rotierenden Platten, die die Daten magnetisch speichern. Trotz höherem Energiebedarf und geringerer Geschwindigkeit gegenüber konventionellen Technologien gehören sie immer noch zu den am häufigsten verwendeten Speichermedien.

3.1.2) Halbleiterlaufwerk (SSDs: Solid State Drives)

Durch den Einsatz von Flash-Speichern sind SSDs wesentlich schneller und langlebiger, zudem verbrauchen sie weniger Energie. SSDs ersetzen zunehmend die HDDs, insbesondere bei Anwendungen, die einen schnellen Datenzugriff erfordern.

3.1.3) Magnetbänder (Tape)

Für die langfristige Archivierung eignen sich nach wie vor Magnetbänder, da diese bei geringen Kosten eine hohe Speicherdichte bieten. Bei dieser Komponente ist die Zugriffszeit relativ lang, dadurch eignen sich die Magnetbänder besonders gut für *kalte Daten*.

iv.) Optische Medien (CD, DVD, Blu-ray Disc)

Optische Medien werden heute überwiegend für die Distribution von Filmen, Musik und Software genutzt, während ihre Rolle als Speichermedium für die private oder geschäftliche Archivierung stark zurückgegangen ist.

Diese traditionellen Technologien sind sehr gut erforscht und zuverlässig, stoßen jedoch durch die ständig wachsende Datenmenge an ihre Grenzen. Zudem sind optische Medien im Vergleich zu der DNA-Speicherung nicht sonderlich gut für die Speicherung *kalter Daten* geeigent, da die Speicherung *kalter Daten* einen hohen Energieverbrauch und Wartungsaufwand mit hohen Kosten zur Folge hat.

3.2 Netzwerktechnologien und Architektur in Rechenzentren

Traditionelle Speicherzentren bestehen aus einer Vielzahl von Speicher- und Recheneinheiten, die über ein spezielles Netzwerk, das sogenannte Data Center Network (DCN), miteinander verbunden sind.
Dieses Netzwerk stellt das zentrale Element eines Rechenzentrums dar. Moderne Rechenzentren benutzen komplexe Architekturen, um den hohen Speicherverkehr und die Speicherdichte zu ermöglichen. Zusammen mit den Netzwerkarchitekturen Leaf-Trunk, Fat-Tree-Strukturen als auch hybride Fat-Tree-Strukturen bilden diese die aktuellen DCNs zur Datenspeicherung.
Mit zunehmender Datenmenge wächst auch die Größe des Rechenzentrums. So betreibt Microsoft etwa weltweit mehr als 100 Rechenzentren mit insgesamt einer Millionen Servern [7]. Diese Größe führt zu hohem Energieverbrauch und bringt enorme Herausforderungen wie Kühlung der Hardware und Optimierung des Netzwerkdesigns mit sich.

3.3 Herausforderungen bei herkömmlichen Rechenzentren

Trotz jahrelanger Optimierung stehen die konventionellen Rechenzentren heute vor grundlegenden Herausforderungen, die durch das exponentielle Wachstum der Datenmenge und der steigenden Datensicherheits-Anforderungen verschärft werden.

Datenmengen und Skalierbarkeit Mit der wachsenden Verbreitung datenintensiver Technologien wie zum Beispiel dem autonomen Fahren steigt das tägliche Datenvolumen auf bislang ungeahnte Größenordnungen. So erzeugen autonom fahrende Fahrzeuge pro Tag zwischen 20 und 40 Terabytes an Daten [7].
Konventionelle Rechenzentren stoßen dabei an physikalische Grenzen hinsichtlich der Speicherkapazität und der Skalierbarkeit, da die gängigen Speicherlösungen wie HDDs, SDDs, Magnetbänder und optische Medien für die aktuellen Datenmengen nicht ausgelegt sind [4, 7].

Hoher Energieverbrauch und Betriebskosten Ein wesentlicher Nachteil herkömmlicher Speichertechnologien liegt im hohen Energiebedarf für den Betrieb, die Kühlung und die Instandhaltung. Vor allem die Aufrechterhaltung konstanter Temperaturbedingungen sowie der durchgängige Betrieb mechanischer Komponenten wie Festplatten oder Bandlaufwerke verursachen beträchtliche laufende Kosten [7].
Ein hoher Teil der gespeicherten Daten wird nur sehr selten oder fast gar nicht abgerufen und dient somit nur zur Archivierung. Diese *kalten Daten* stellen ein besonders großes Problem dar. Die kontinuierliche Wartung und Instandhaltung der Rechenzentren ist sehr energieaufwändig und damit sehr kostspielig [4, 7].
Die nachhaltige Datenerhaltung wird daher durch *kalten Daten* eingeschränkt.

Zuverlässigkeit und Datensicherheit Digitale Daten sind anfällig für eine Vielzahl von Bedrohungen. Darunter zählen Hardwareausfälle, Systemstörungen, Cyberangriffe oder Datenkorruption. Auch Naturkatastrophen oder elektromagnetische Ereignisse (z.B. Sonnenstürme) können die Integrität der konventionell gespeicherten Daten gefährden [4].
Gleichzeitig steigen die Anforderungen am Datenschutz, insbesondere bei personenbezogenen oder sicherheitsrelevanten Informationen (unter anderem Halterdaten, Standort). Die Einhaltung gesetzlicher Vorgaben sowie technischer Standards stellen die Betreiber der Rechenzentren zunehmend vor operative und regulatorische Hürden [7].

Insgesamt verdeutlichen die beschriebenen Herausforderungen die strukturellen Grenzen herkömmlicher Rechenzentren. Angesichts der steigenden Anforderungen an Datendichte, Energieeffizienz, Langlebigkeit und Sicherheit werden alternative Speicheransätze wie DNA-basierte Speicherlösungen zunehmend attraktiver und zunehmend unverzichtbar.
Eine nähere Betrachtung dieser Ansätze erfolgt direkt im folgenden Kapitel 4.

4 Ansatz und Eigenschaften der DNA-Speicherung

Die DNA-Speicherung ist ein innovativer Ansatz zur langfristigen Speicherung digitaler Daten mit einer hohen Speicherdichte und bietet das Potenzial, konventionelle Speicherlösungen in bestimmten Anwendungsfeldern zu ergänzen oder gar komplett zu ersetzen. Die erfolgreiche Umsetzung der Datenspeicherung in DNA-Molekülen wurde von George Church, Yuan Gao und Sriram Kosuri in der Publikation „Next-Generation Digital Information Storage in DNA“ [2] bewiesen.

Der innovative Ansatz der DNA-Speicherung basiert auf der Nutzung synthetisch hergestellter DNA-Moleküle als Speichermedium und wird zunehmend als Lösung für die Herausforderungen herkömmlicher Rechenzentren diskutiert, die im vorherigen Absatz erläutert wurden. Die schrittweise Skalierung und die praktische Anwendbarkeit der DNA-Speicherung beruhen auf interdisziplinärer Forschung in den Bereichen Biologie, Computer- und Kommunikationssystemen [4, 7, 8].

4.1 Grundprinzip der DNA-Speicherung

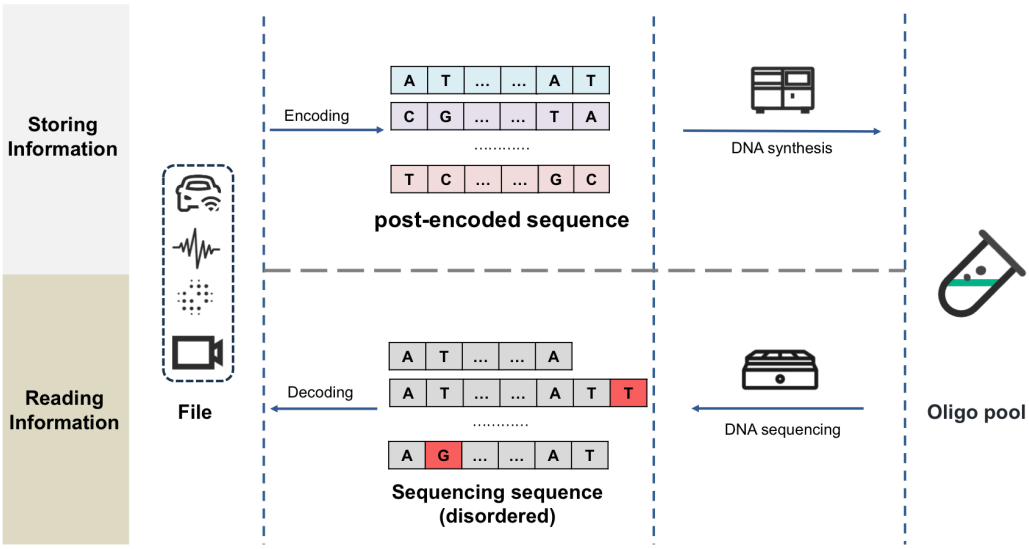


Abbildung 1: Flussdiagramm der DNA-Speicherung, Quelle: Qu et al. (2024) [7]

Abbildung 1 zeigt den gesamten Prozess der DNA-Speicherung. Das Kernprinzip der DNA-Speicherung besteht darin, binäre Daten in die Abfolge der vier Nukleotidbasen Adenin (A), Thymin (T), Guanin (G) und Cytosin (C) zu übersetzen. Anschließend wird diese Sequenz synthetisch als DNA-Molekül hergestellt, physisch gelagert und bei Bedarf wieder ausgelesen. Der gesamte Prozess lässt sich in **fünf Teilschritte** unterteilen [4, 7].

4.1.1 Datenkodierung

Die erste Phase der DNA basierten Speicherung beinhaltet die Umwandlung und Darstellung der digitalen Daten in DNA-Sequenzen. Die digitalen Daten werden zunächst in die binäre Form umgewandelt. Die binären Sequenzen können dann wiederum durch das folgende Mapping umgewandelt werden:

00 <=> Adenin
11 <=> Thymin
10 <=> Guanin
01 <=> Cytosin

Das Ergebnis der Kombination der Mappingregeln als DNA-Basenfolge sind die zuvor umgewandelten Binärdaten.
Fehlerkorrekturverfahren wie das in Kapitel 2 (Theoretische Grundlagen) beschriebene Reed-Solomon-Codes-Verfahren werden in die Datenkodierung integriert, um die Robustheit gegenüber Einsetzungs-, Ersetzungs- und Löschungsfehlern zu gewährleisten. Diese Fehlerbehebung verringert die Wahrscheinlichkeit von Datenfehlern bei der Übertragung der Informationen [4, 7].

4.1.2 DNA-Synthese (Schreiben der Daten)

Die zweite Phase der DNA-Speicherung, dargestellt in Abbildung 1, umfasst den zentralen Schreibprozess mittels DNA-Synthese. Hierbei werden künstliche DNA-Stränge erzeugt, deren Basenabfolge zuvor aus den zu speichernden digitalen Daten berechnet wurde. Üblicherweise erfolgt die Synthese chemisch, beispielsweise durch das phosphoramiditbasierte Verfahren, bei dem die Nukleotide schrittweise zu der gewünschten Sequenz verknüpft werden. Neben diesem etablierten Ansatz existieren auch chipbasierte Verfahren, die höhere Schreibgeschwindigkeiten ermöglichen und potenziell mit geringeren Synthesekosten für die Codierung der Daten verbunden sind.
Da die Synthese derzeit nur in begrenzter Länge (etwa 200-300 Basen) hergestellt werden kann, werden größere Datenmengen in viele kurze Fragmente aufgeteilt, die später beim Auslesen wieder zusammengesetzt werden.

4.1.3 Datenkonservierung

In der dritten Phase des Prozesses geht es um die verschiedenen Arten, wie man die erstellten DNA-Stränge konservieren kann - flüssig, als Trockenpulver oder verkapselt.
Flüssige Proben werden meist in einer speziellen Pufferlösung bei sehr niedrigen Temperaturen gelagert, während Trockenpulver sogar bei Raumtemperatur in luft- und lichtdichten Behältern stabil erhalten bleibt. Besonders verkapselte DNA, etwa in Silica oder speziellen Polymeren, hat sich als eine der zuverlässigsten Methoden für die Langezeitlagerung bewährt.
Indem die DNA-Moleküle von Wasser und Sauerstoff isoliert werden, können diese mehrere

Jahrhunderte erhalten bleiben.
Diese Konservierungsmethoden gewährleisten die Langlebigkeit und Stabilität der in den DNA-Molekülen kodierten digitalen Daten.

4.1.4 DNA-Sequenzierung (Lesen der Daten)

In der vorletzten Phase des Prozesses geht es um die Rückgewinnung der Daten durch die Sequenzierung. Dieses Verfahren ermöglicht die Analyse der Reihenfolge der Nukleotidbasen in den DNA-Strängen [7]. Moderne Sequenzierungstechnologien, wie die sogenannte „Next Generation Sequencing“ (NGS), erlauben die schnelle und präzise Analyse langer DNA-Stränge.
Ein zentrales Konzept der Sequenzierung ist die sogenannte *Coverage*. Damit ist die durchschnittliche Anzahl an Wiederholungen gemeint, mit denen jedes DNA-Fragment gelesen wird. Eine höhere *Coverage* reduziert die Wahrscheinlichkeit von Lesefehlern, da die resultierenden mehrfach gelesenen Fragemente miteinander verglichen und statistisch ausgewertet werden können. Allerdings verlängert sich mit steigender *Coverage* auch die Sequenzierzeit, da mehr Daten verarbeitet werden müssen.

Eng mit der *Coverage* verbunden ist die Verwendung von *Redundanz* in den gespeicherten Sequenzen. Diese zusätzliche Information wird bewusst in die DNA-Stränge eingebracht, um Fehler wie Ersetzungen, Einsetzungen oder Löschungen beim Lesen ausgleichen zu können.
Während Redundanz die Zuverlässigkeit des Rekonstruktionsprozesses erhöht, vergrößert sie zugleich den Datenumfang und damit den Rechenaufwand bei der späteren Dekodierung.
Diese beiden Mechanismen – *Coverage* und Redundanz – bilden die Grundlage dafür, dass selbst bei den fehleranfälligen biologischen Prozessen eine robuste Rückgewinnung der Daten möglich ist. Sie stellen allerdings auch zentrale Faktoren dar, die die Gesamtperformance der DNA-Speicherung beeinflussen [8].

4.1.5 Datendekodierung

In der letzten Phase der DNA-Speicherung erfolgt die Entschlüsselung der sequenzierten Daten. Die zuvor sequenzierten Daten werden mithilfe der beschriebenen Mapping-Regel in ihre digitale binäre Darstellung zurückkodiert.

4.2 Eigenschaften der DNA-Speicherung

Die DNA-Speicherung weist eine Reihe von Vorteilen gegenüber den im dritten Kapitel 3 vorgestellten konventionellen Speichertechnologien auf, da sie besonders für die langfristige Archivierung großer, selten genutzter Datenmengen geeignet ist. Die Datenspeicherung in DNA-Strängen weist folgende fünf Merkmale auf, die zeigen, warum man die digitalen Daten in DNA-Strängen speichern sollte:

4.2.1 Hohe Speicherdichte

Wie im zweiten Kapitel 2 bereits erwähnt kann man theoretisch in einem Gramm DNA bis zu 455 Exabytes an digitalen Daten speichern [7].
Selbst wenn man von dieser Zahl die Redundanz und die Fehlerkorrektur miteinbezieht, liegt die Speicherdichte bei DNA-basierter Datenspeicherung mehrere Größenordungen über den konventionellen Rechenzentren - die maximale Speicherkapazität bei einer SSD liegt momentan (Stand: 2025) bei circa 122 Terabytes (122 * 10^{12} Bytes) [11]. Zum Vergleich: 1 Exabyte entspricht 1.000.000 Terabytes.

4.2.2 Langlebigkeit

DNA-Moleküle sind extrem stabil und können unter geeigneten Bedingungen Jahrhunderte bis Jahrtausende überdauern. Die Analyse prähistorischer DNA aus Fossilien belegen die Beständigkeit des Moleküls über Zehntausende von Jahren [4].
Diese Eigenschaft macht die DNA als Speichermedium besonders geeignet zur Archivspeicherung.
Im Gegensatz zu den konventionellen Medien wie Festplatten oder Magnetbändern, welche nach Jahrhunderten schon ausfallen können, bietet die DNA somit eine langfristige Lösung zur Bewahrung digitaler Informationen [7, 8].

4.2.3 Geringer Energieverbrauch

Im Gegensatz zu elektronischen oder magnetischen Speichermedien benötigt die DNA keine permanente Stromversorgung, um ihre Informationsstruktur erhalten zu können.
Bei der Datenspeicherung in DNA-Strängen wird nur Energie bei dem Lese- (Sequenzierung) und Schreibprozess (Synthese) verbraucht, jedoch nicht bei der langjährigen Lagerung.
Dadurch sinken sowohl die Betriebskosten als auch der $CO^2 - Verbrauch$, was ein wichtiger Aspekt im Kontext wachsender Rechenzentren und Datenmengen darstellt [7, 13].

4.2.4 Umweltfreundlichkeit

In der heutigen Zeit ist auch die Umweltfreundlichkeit der DNA-Stränge von großer Bedeutung. Da die DNA ein natürlich vorkommendes Molekül ist, entstehen bei ihrer Nutzung keine elektronischen Abfälle im Gegensatz zur Herstellung herkömmlicher Rechenzentren, bei der solche Abfälle unvermeidbar sind.
Zudem lassen sich die DNA-Stränge mit relativ geringerem Aufwand lagern, da diese keine energieintensive Kühlsysteme erfordern. Damit lässt die DNA-Speicherung ein nachhaltigeres und ressourcenschonenderes Speicherverfahren als herkömmliche Technologien zu [7, 13].

4.2.5 Sicherheit

Die DNA ist unempfindlich gegenüber elektromagnetischen Einflüssen und kann zusätzlich mit Verschlüsselungsmethoden kombiniert werden. Zudem zeigen aktuelle Forschungsarbeiten von Manodee, Pandey, Mishar und Kumar (2024) [4], dass der Einsatz von Fehlerkorrekturverfahren

und Indexierung selbst bei typischen Sequenzierungsfehlern eine zuverlässige Rekronstruktion der zuvor kodierten Daten gewährleistet.
Dadurch eignet sich die Datenspeicherung in DNA-Strängen auch bei Anwendungen, bei denen die Datensicherheit eine übergeordnete Rolle spielt [4, 8].

4.3 Aktuelle Einschränkungen und Herausforderungen

Die Datenspeicherung in DNA-Strängen hat viele Vorteile wie die hohe Datendichte, lange Haltbarkeit und einem niedrigen Energieverbrauch zur Archivierung. Jedoch ist die DNA-Speicherung auch in diversen Bereichen den herkömmlichen Speicherverfahren unterlegen:

4.3.1 Hohe Synthese- und Sequenzierungskosten

Momentan sind die sehr hohen Synthese- und Sequenzierungskosten bei der Datenspeicherung in DNA-Strängen das Kostspieligste an diesem Verfahren [4, 7].

4.3.2 Begrenzte Schreib-/ und Lesegeschwindigkeit

Sowohl die Synthese als auch die Sequenzierung sind sehr zeitaufwändig, da man mit der Begrenzung der Bandbreite immer nur kleine Teile der Daten lesen und schreiben kann.
Dies macht die Technologie vor allem für selten genutzte Archivdaten, den sogenannten *kalten Daten*, interessant [7].

4.3.3 Fehleranfälligkeit

Die DNA-Synthese und -Sequenzierung sind sehr anfällig für Ersetzungs-, Einsetzungs- und Löschungsfehler. Dies erfordert aufwändige Fehlerkorrekturmechanismen, die im zweiten Kapitel 2 erläutert wurden [4, 8].

4.3.4 Fehlende Standardisierung

Für die Kodierung, Speicherung und Lesung der Daten in dem DNA-Speicherungsprozess gibt es noch keine einheitichen Standards, was die Interoperabilität mit anderen Rechenzentren und damit verbunden die breite Anwendung des Systems erschwert [4].

Guanjin Qu und Huaming Wu, Autoren der Publikation „DNA Storage Promotes Long-term Storage of Internet of Vehicles Data“ [7], gehen davon aus, dass die fortschreitende Wissenschaft und Forschung in diesem Bereich die oben genannten Probleme minimieren können, sodass die DNA-Speicherung auch dort mit herkömmlichen Rechenzentren konkurrieren kann.
Diese haben auch ein mit der DNA-Speichertechnologie kombiniertes Rechenzentrum vorgestellt, welches im nächsten Kapitel beschrieben wird.

5 Mögliche Rechenzentrenstruktur

Die Integration von DNA-Speichertechnologien in Rechenzentren eröffnet neue Möglichkeiten, um den steigenden Anforderungen an Speicherkapazität, Energieeffizienz und Datensicherheit gerecht zu werden. Während klassische Rechenzentren weiterhin auf elektronische und magnetische Speicherlösungen angewiesen sind, kann das Verfahren der DNA-Speicherung als zusätzliches Modul vor allem für die langfristige Archivierung genutzt werden [7].

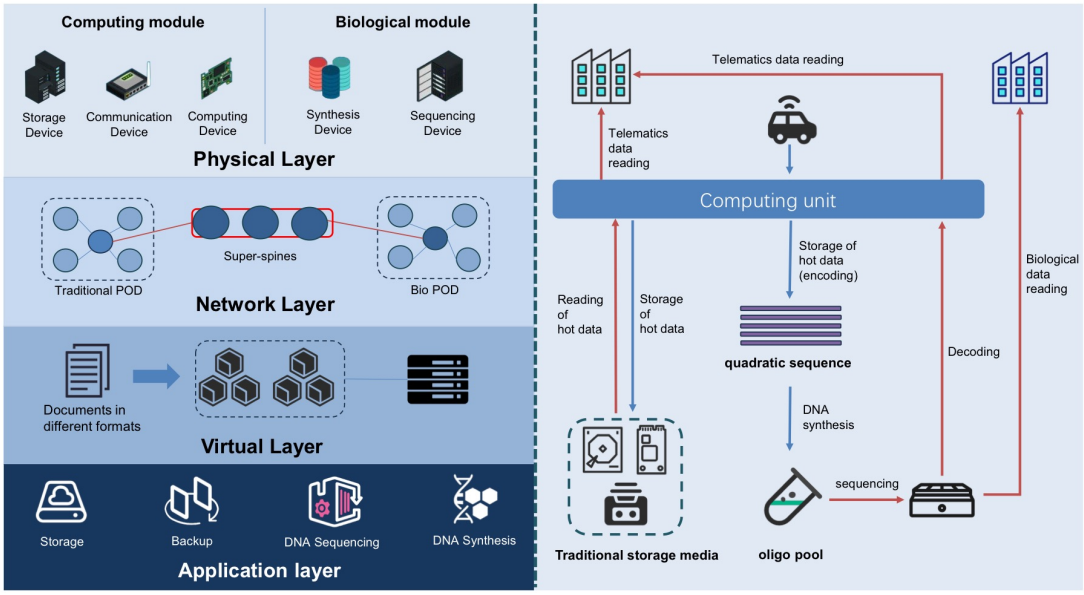


Abbildung 2: Framework Diagramm des Rechenzentrums kombiniert mit DNA-Speicherung, Quelle: Qu et al. (2024) [7]

Das in Abbildung 2 dargestellte Framework-Diagramm [Guanjin Qu und Huaming Wu Publikation, 2024 [7]] zeigt, wie die DNA-Speichertechnologie in bestehende Rechenzentren eingebettet werden kann. Die Abbildung dient zur visuellen Ergänzung, um die nachfolgende Beschreibung der Struktur und die Wechselwirkung der verschiedenen Technologien nachvollziehbar zu machen.

5.1 Strukturebenen des kombinierten Rechenzentrums

Die Struktur lässt sich in zwei zentrale Ebenen unterteilen:

1) Datenerfassung und -eingang

Auf der obersten Ebene gelangen Daten aus verschiedenen Quellen in das Rechenzentrum. Dazu gehören zum Beispiel bei autonomen Fahrzeugen die Fahrzeugsensorikdaten (Video, Punktwolke, Radarwellen, etc.), Fahrzeugzustandsinformationen (Geschwindigkeit,

Beschleunigung, GPS-Position, etc.), aber auch Daten unteschiedlicher Typen, Mengen und auch Nutzungshäufigkeiten.

2) Rechen- und Speichermodule

Wie man in der Abbildung sehen kann, befinden sich in der Physikalischen Schicht zwei verschiedene Module. Ein Modul besteht aus herkömmlichen Rechenzentren, das zweite Modul beruht auf biologischer Datenspeicherung.
Das Rechenmodul besteht aus den Untereinheiten der Speicherung, der Kommunikation miteinander durch Hochgeschwindigkeits-Switches und Routern und der eigentlichen Recheneinheit, bestehend aus GPUs, CPUs und den Rechenservern.

Das biologisch basierte Modul ist in die Phasen DNA-Synthese, DNA-Konservierung und DNA-Sequenzierung unterteilt.
Je nach Art der zu verarbeitenden Daten kann die kombinierte Rechenarchitektur flexibel das jeweils am besten geeignete Modul einsetzen. Bei *heißen Daten*, welche regelmäßig benötigt werden, würde man den klassischen Ansatz der Speicherung verwenden. Dieses System ist durch seine hohe Geschwindigkeit absolut für die Echtzeit-Verarbeitung der Daten geeignet [4, 8].

Müssen hingegen Daten verarbeitet werden, die selten genutzt und auch langfristig aufbewahrt werden sollen, dann würde sich das System der kombinierten Rechnerarchitektur für das biologisch basierte Rechenmodul entscheiden. Hier würden die Daten dann durch das bereits erklärte DNA-Speicherungsmodell über Synthese in DNA-Strängen gespeichert und auch gelagert werden.

Das Diagramm in Abbildung 2 zeigt deutlich, dass die beiden Module in der Netzwerkschicht miteinander verbunden sind. Damit wird eine hybride Fat-Tree-Struktur, optische Switches der PoD-Knoten (Point of Delivery) verwendet, um sich jeweils für das gewünschte Rechenmodul zu entscheiden.

Gelesen werden die Daten mit dem Verfahren, in welchem die Daten gespeichert wurden. Bei dem biologischen Modul über die Sequenzierung und bei den nicht biologischen Rechenmodulen werden die Daten wie bei herkömmlichen Rechenzentren gelesen.

5.2 Chancen und Grenzen

Die vorgestellte Architektur verbindet die Stärken beider Speicherungstechnologien:

- kurzfristig verfügbare, schnelle Zugriffe durch klassische Speicherlösungen,
- langfristige, nachhaltige Archivierung durch DNA-Speicherung.

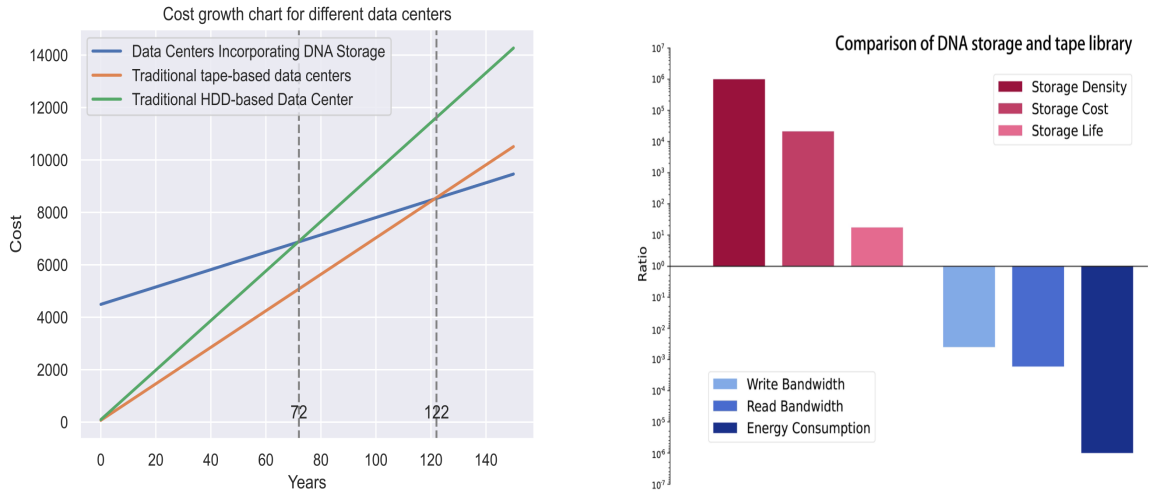
Damit entsteht ein System, welches sowohl hohe Leistung als auch große Langlebigkeit ermöglicht. Gleichzeitig bestehen wie bei jeder Technologie Grenzen: Die Schreib- und Leseprozesse

der DNA sind derzeit noch zeitintensiv und kostenaufwändig, wodurch sich ihr aktueller Einsatz fast nur für die Speicherung von Archivdaten eignet.

Insgesamt zeigt dieses Modell, dass die DNA-Datenspeicherung als **Ergänzung** klassischer Speicherlösungen eine entscheidende Rolle für die Rechenzentren der Zukunft einnehmen kann, insbesondere im Hinblick auf die Energieeffizienz, die Nachhaltigkeit und den Umgang mit stetig wachsenden Datenmengen.

6 Bewertung der Leistung

Die Leistungsbewertung ist ein zentraler Schritt, um die Vorteile und Einschränkungen der DNA-Speicherung im Kontext von herkömmlichen Rechenzentren nachvollziehbar zu machen. In Abbildung 3 wird der Leistungsvergleich in Form der Energiekosten und der Zugriffslatenz der beiden Datenspeicherungsvarianten grafisch dargestellt.
Die Publikation von Guanjin Qu und Huaming Wu „DNA Storage Promotes Long-term Storage of Internet of Vehicles Data“ [7] stellt einen Vergleich zwischen DNA-Speicherung und einem Intel(R) XEON(R) Gold 6348H CPU @ 2.30 GHz mit 1 TB RAM-Speicher dar.



(a) Energieeffizienz

(b) Zugriffslatenz

Abbildung 3: Vergleichende Leistungsbewertung bei einer Speicherung von 1 ZetaByte: (a) Energieeffizienz und (b) Zugriffslatenz bei DNA-Speicherung gegenüber konventionellen Speichermedien, Quelle: Qu et al. (2024) [7]

6.1 Analyse der Betriebskosten

Wie in **Abbildung 3a** dargestellt, liegen die Betriebskosten für die Speicherung von 1 Zettabyte ($10^{21} Byte$) bei klassischen Speichermedien wie HDDs oder Bandspeichern zunächst deutlich

unter den hohen Einstiegskosten der DNA-basierten Speicherung. Bei den konventionellen Technologien zeigt sich jedoch im Laufe der Zeit ein steiler Anstieg der Betriebskosten.
Ursache hierfür sind die kontinuierlich anfallenden Aufwendungen für Energie, Kühlung, dauerhafte Stromversorgung und die regelmäßige Wartung des Systems.
DNA-Speicher hingegen weisen zwar erheblich höhere Anfangskosten auf, die vor allem durch die Phasen der Synthese und der Sequenzierung verursacht werden, entwickeln sich in ihrer Kostenkurve jedoch deutlich flacher. Da DNA im Gegensatz zu herkömmlichen Speichermedien keine permanente Energiezufuhr benötigt, fallen während der Lagerung kaum zusätzliche Kosten an.
Aus der Abbildung wird zudem ersichtlich, dass sich die Kostengraphen langfristig schneiden: Bei HDDs nach etwa 72 Jahren und bei Bandspeichern nach rund 122 Jahren. Dies verdeutlicht, dass die DNA-Speicherung ab diesem Zeitpunkt eine kostengünstigere Alternative zu herkömmlichen Speichertechnologien darstellt, da sich ihr ökonomischer Vorteil erst über sehr lange Zeiträume hinweg entfaltet.

6.2 Analyse des Leistungsvergleichs

In **Abbildung 3b** ist der Leistungsvergleich zwischen der DNA-Speicherung und seinem wichtigsten alternativen Speichermedium, der Bandbibliothek zu erkennen. Dabei wird deutlich, dass die klassische Technologie insbesondere bei der Lese- und Schreibgeschwindigkeit der DNA-Speicherung aktuell weit überlegen ist, wodurch sich diese besonders für Anwendungen mit häufigem Datenabruf und Echtzeitverarbeitung eignen.
Gleichzeitig zeigt die Abbildung aber auch die klare Überlegenheit der DNA-Speicherung im Hinblick auf die Speicherdichte und der Langlebigkeit.
Auch beim Vergleich der Performance [Abbildung 3b], zeigt sich, dass die Stärken der DNA-Speicherung vor allem in der Speicherung der *kalten Daten* liegt. Wenn Echtzeit-Anwendungen gefordert werden, ist die DNA-basierte Speicherung immer noch den herkömmlichen Technologien unterlegen.
Die längeren Zugriffszeiten lassen sich insbesondere durch die Struktur der Gesamtlaufzeit erklären. Diese setzt sich aus drei Hauptkomponenten zusammen:

$$T_{\text{gesamt}} = T_{\text{Synthese}} + T_{\text{Sequenzierung}} + T_{\text{Dekodierung}}. \tag{1}$$

$\underline{T_{\textbf{Sequenzierung}}}$

$T_{\text{Sequenzierung}}$ ist die Laufzeit der DNA-Synthese. Je mehr Daten gespeichert werden sollen, desto stärker steigen die Synthesezeiten

$\underline{T_{\textbf{Sequenzierung}}\ \textbf{(Lesen)}}$

Beim Auslesen werden die DNA-Fragmente sequenziert. Ein wesentlicher Faktor ist dabei die *Coverage*, welche die Lesezeit proportional erhöht, da pro zusätzlichem Durchlauf mehr Datenmengen verarbeitet werden müssen:

$$T_{\text{Sequenzierung}} = C \cdot n, \tag{2}$$

wobei C die *Coverage* und n die Länge der Sequenz beschreibt.

$T_{\text{Dekodierung}}$ (Rekonstruktion)

Die aus der Sequenzierung gewonnenen Datenströme müssen wieder in die ursprünglichen digitalen Informationen zurückübersetzt werden. Dabei entstehen Verzögerungen durch die notwendige Analyse und den Abgleich der vielen erzeugten und gelesenen Fragmente. Auch wenn dieser Schritt weniger zeitintensiv ist als die Synthese selbst, nimmt er mit wachsendem Datenvolumen zu [8].

7 Zusammenfassung und Fazit

Die vorliegende Arbeit befasst sich mit den Potenzialen und Herausforderungen der DNA-Speicherung im Kontext herkömmlicher Rechenzentren. Ausgangspunkt war die Frage, inwieweit DNA-basierte Datenspeicherung in bestehende Systeme integriert werden kann und ob sie langfristig sogar klassische Speichertechnologien ersetzen könnte.

Es konnte nachgewiesen werden, dass konventionelle DNA-Speicherung insbesondere durch ihre außergewöhnlich hohe Speicherdichte, ihre Langlebigkeit sowie den geringen Energiebedarf für die Konservierung eine vielversprechende Lösung darstellt – vor allem für die Archivierung *kalter Daten*. Für Anwendungen mit Echtzeitverarbeitung bleibt sie jedoch aufgrund hoher Latenzzeiten und geringer Schreib- und Lesegeschwindigkeiten klar im Nachteil. Daher ist die DNA-Speicherung derzeit nicht als Ersatz, sondern eher als Ergänzung zu herkömmlichen Speichertechnologien zu verstehen. Um sie in Rechenzentren praktikabel einsetzen zu können, sind vor allem Effizienzsteigerungen in Synthese, Sequenzierung und Fehlerkorrektur erforderlich.

Besonders überzeugend ist das in der Arbeit vorgestellte hybride Modell von Guanjin Qu und Huaming Wu (2024) [7]. Dieses kombiniert die Stärken beider Ansätze: klassische Speichertechnologien wie HDDs oder SSDs für *heiße Daten* mit kurzen Zugriffszeiten sowie die DNA-Speicherung für *kalte Daten*, bei denen langfristige Haltbarkeit und niedrige Betriebskosten entscheidend sind. Auf diese Weise entsteht eine Architektur, die sowohl Leistungsfähigkeit als auch Nachhaltigkeit vereint.

Trotzdem ist festzuhalten, dass die ökonomischen Vorteile der DNA-Speicherung erst über sehr lange Zeiträume hinweg sichtbar werden – so schneiden sich die Betriebskostenkurven im Vergleich zu Bandspeichern erst nach rund 122 Jahren, was bedeutet, dass die DNA-Speicherung erst ab diesem Zeitpunkt wirtschaftliche Vorteile bietet. Hier ist weitere Forschung notwendig, um die Technologie zu optimieren und ihre Kosten schneller zu senken.

Insgesamt lässt sich schlussfolgern: Die DNA-Speicherung stellt noch keinen Ersatz klassischer Speichertechnologien dar, birgt jedoch ein erhebliches Zukunftspotenzial als nachhaltige Ergänzung in hybriden Rechenzentren.

Literatur

[1] George Church, Yuan Gao & Sriram Kosuri (2012): *Next-Generation Digital Information Storage in DNA*. doi:10.1126/science.1226355. Available at https://www.researchgate.net/publication/230698422_Next-Generation_Digital_Information_Storage_in_DNA.

[2] George M. Church, Yuan Gao & Sriram Kosuri (2012): *Next-generation digital information storage in DNA*. *Science* 337(6102), pp. 1628–1628, doi:10.1126/science.1226355. Available at https://www.science.org/doi/10.1126/science.1226355.

[3] MVZ Dessau (2025): *DNA-Sequenzierung: Methoden wie Sanger, Pyrosequenzierung und Next-Generation-Sequencing*. https://mvzdessau.de/fachbereiche/kliniken-institute-auenweg/pathologie/leistungsspektrum/molekularpathologie/dna-sequenzierung. Abgerufen am 17. August 2025.

[4] Akarsh Manodee, Neeraj Kumar Pandey, Amit Kumar Mishar & Amit Kumar (2024): *DNA Data Storage: A Solution to the Growing Digital Data Dilemma* 26, pp. 2586–2590. doi:10.1109/ICDT61202.2024.10489686. Available at https://ieeexplore.ieee.org/document/10489686.

[5] OpenAI (2025): *ChatGPT – KI-gestütztes Sprachmodell als Schreibhilfe*. https://chat.openai.com. Eingesetzt als Unterstützung zur grammatikalischen Überprüfung und sprachlichen Umformulierung, nicht zur Erstellung fachlicher Inhalte.

[6] Zhi Ping, Dongzhao Ma, Xiaoluo Huang, Shihong Chen, Longying Liu, Fei Guo, Sha Joe Zhu & Yue Shen (2019): *Carbon-based archiving: current progress and future prospects of DNA-based data storage*. *GigaScience* 8(6), doi:https://doi.org/10.1093/gigascience/giz075. Available at https://academic.oup.com/gigascience/article/8/6/giz075/5521158?searchresult=1.

[7] Guanjin Qu, Huaming Wu & Ruidong Li (2024): *DNA Storage Promotes Long-Term Storage of Internet of Vehicles Data* 39, pp. 220–227. doi:10.1109/MNET.2024.3481262. Available at https://ieeexplore.ieee.org/document/10718314.

[8] Omer Sabary, Han Mao Kiah, Paul Z. Siegel & Eitan Yaakobi (2024): *Survey for a Decade of Coding for DNA Storage*. *IEEE Transactions on Molecular Biological and Multi-Scale Communications* 10(2), pp. 253–271, doi:10.1109/tmbmc.2024.3403488. Available at https://ieeexplore.ieee.org/document/10535435/.

[9] Ilan Shomorony & Reinhard Heckel (2019): *Capacity Results for the Noisy Shuffling Channel*. *IEEE International Symposium on Information Theory (ISIT)*, doi:10.1109/ISIT.2019.8849789. Available at https://ieeexplore.ieee.org/document/8849789.

[10] Priyanka Shrivastava & Uday Pratap Singh (2013): *Error Detection and Correction Using Reed Solomon Codes*. *International Journal of Advanced Research in Computer Science and Software Engineering* 3(11), pp. 965–969. Available at https://www.researchgate.net/publication/305641094_Error_Detection_and_Correction_Using_Reed_Solomon_Codes.

[11] TechRadar (2025): *Largest SSDs and biggest hard drives of 2025*. Available at https://www.techradar.com/best/large-hard-drives-and-ssds.

[12] CS168 Textbook (2025): *Datacenter Topology – Fat-Tree (Clos Topology)*. https://textbook.cs168.io/datacenter/topology.html. Abschnitt „Fat-Tree Clos Topology“ – abgerufen am 17. August 2025.

[13] Zihui Yan, Cong Liang & Huaming Wu (2022): *Upper and Lower Bounds on the Capacity of the DNA-Based Storage Channel* 26, pp. 2586–2590. doi:10.1109/LCOMM.2022.3202961. Available at https://ieeexplore.ieee.org/document/9869772.