Business Strategy of Nucleic Acid Memory for Digital Information Storage

By

Ryan de Ridder

M.S. Mathematics (2008) University of Virginia
B.A. Mathematics & Physics (2007) University of Virginia

Submitted to the System Design & Management Program
In Partial Fulfillment of the Requirements for the Degree of

Master of Science in Engineering & Management
at the
Massachusetts Institute of Technology
February 2017

© 2017 Ryan de Ridder. All Rights Reserved

The author hereby grants to MIT permission to reproduce and to distribute publicly, paper and electronic copies of this thesis document, in whole or in part, in any medium now known or hereafter created.

Written by ________________________________
Signature redacted

Ryan de Ridder
MIT System Design & Management Program
January 20, 2017

Certified by ________________________________
Signature redacted

Aleksandra Kacpetczyk
Thesis Advisor
MIT Sloan School of Management

Accepted by ________________________________
Signature redacted
Warren Seering
Weber-Shaughness Professor of Mechanical Engineering
Business Strategy of Nucleic Acid Memory for Digital Information Storage

by

Ryan de Ridder

Submitted to the System Design & Management Program on January 20, 2017 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Engineering & Management

Abstract

Nucleic acid memory (NAM) is the storage of digital data by encoding the information into the medium of nucleic acids. This is often called DNA storage, as typically, but not necessarily, the information is stored in the nucleobases that comprise DNA. Baum first introduced this idea in 1995, but it wasn't until 2012 that Church proved the idea on a larger scale. NAM has a number of features that make it very promising as a data storage medium. The three typically highlighted are capacity density, data retention (i.e., durability), and energy usage.

NAM should enter the data storage market, as a hardware product, through the ~$4.5B archiving market, by targeting large storage service providers and large data-intensive corporations with on-premise operations. A NAM product has the potential to reduce the capital and operational cost base of these companies, by millions of dollars per year. An architecture strategy should be employed to enter the market, relying on control over underlying ideas and partnerships to barricade the company from competition.

NAM is a decade away from commercialization, making this a very risky early stage venture. The costs need to come down at least 100,000-fold before the technology is cost competitive with current solutions. Additionally, there are a number of scientific and engineering issues that need to be carefully resolved. Due to the risks, the only viable funding source is government grants. If early stage funding were secured, IP should be developed in the core NAM technology of storage and access and an interim revenue source established. This would allow the company a strong chance to thrive in the competitive storage industry, if and when NAM becomes cost competitive.

Thesis Supervisor: Aleksandra Kacperczyk
Title: Fred Kayne (1960) Career Development Professor of Entrepreneurship
# Table of Contents

ABSTRACT ............................................................................................................................................ 3  

TABLE OF CONTENTS .......................................................................................................................... 5  

TABLE OF FIGURES ................................................................................................................................. 7  

DATA STORAGE ........................................................................................................................................ 9  
  
  DATA TYPES ........................................................................................................................................... 9  
  STORAGE ATTRIBUTES ......................................................................................................................... 10  
  STORAGE TERMINOLOGY ..................................................................................................................... 11  
  STORAGE TECHNOLOGY ....................................................................................................................... 13  

STORAGE MARKET .................................................................................................................................... 15  
  
  STORAGE DEMAND ............................................................................................................................... 16  
  MARKET SIZE ......................................................................................................................................... 17  
  ARCHIVING MARKET ............................................................................................................................ 20  
  INDUSTRY DYNAMICS ........................................................................................................................... 22  

NUCLEIC ACID MEMORY ....................................................................................................................... 25  
  
  PROCESS ............................................................................................................................................... 26  
  FEATURES ................................................................................................................................................ 33  
  ISSUES ................................................................................................................................................... 37  
  OUTLOOK .............................................................................................................................................. 38  

NAM STRATEGY ......................................................................................................................................... 39  
  
  FRAMEWORK ....................................................................................................................................... 39  
  MARKET .................................................................................................................................................. 43  
  CUSTOMER ............................................................................................................................................ 44  
  COMPETITION ......................................................................................................................................... 47  
  VALUE CREATION .................................................................................................................................. 50  
  FUNDING ............................................................................................................................................... 54  
  RISKS ...................................................................................................................................................... 56  

CONCLUSION ........................................................................................................................................... 57  

APPENDIX ............................................................................................................................................... 58
# Table of Figures

- **Figure 1**: Key Differences between Archival and Backup ......................................................... 10
- **Figure 2**: Tier Terminology .................................................................................................... 12
- **Figure 3**: Terminology Relationship ....................................................................................... 13
- **Figure 4**: Hard Disk Drive (HDD) ............................................................................................ 14
- **Figure 5**: Digital Video Disc (DVD) .......................................................................................... 14
- **Figure 6**: Solid State Drive (SSD) .............................................................................................. 15
- **Figure 7**: Global Memory Demand Sources ............................................................................. 16
- **Figure 8**: Global Memory Demand ........................................................................................... 17
- **Figure 9**: Worldwide IT Market ............................................................................................... 18
- **Figure 10**: Worldwide Cloud IT Infrastructure Market Forecast by Development Type .......... 19
- **Figure 11**: Worldwide Enterprise Storage Market .................................................................... 19
- **Figure 12**: Worldwide Enterprise Storage Systems Revenue, by Vendor .............................. 20
- **Figure 13**: Worldwide Information Archiving Revenue Projections ...................................... 21
- **Figure 14**: Comparison of Hardware Provider and Cloud Service Provider ......................... 22
- **Figure 15**: Value Curve Comparison ....................................................................................... 23
- **Figure 16**: Data Storage Complaints and Challenges ............................................................... 24
- **Figure 17**: Data Storage Industry Overview ............................................................................ 25
- **Figure 18**: High Level NAM System ....................................................................................... 26
- **Figure 19**: NAM Writing and Reading Process Decomposition ............................................. 27
- **Figure 20**: High Level NAM Process ....................................................................................... 27
- **Figure 21**: Rotating Ternary Scheme ....................................................................................... 28
- **Figure 22**: Encoding Process Decomposition ......................................................................... 28
- **Figure 23**: ATCG Block Representation ................................................................................... 29
- **Figure 24**: Access Process Decomposition ............................................................................ 31
- **Figure 25**: Cost per Megabase of DNA Sequencing ............................................................... 32
- **Figure 26**: Process Diagram .................................................................................................. 33
- **Figure 27**: Comparison of Standard Memory Technologies to Cellular DNA Attributes’ ........ 34
- **Figure 28**: Technology Areal Density Increases & Timescales .............................................. 35
- **Figure 29**: Memory-Retention Times for DNA in Air and Water ............................................ 35
- **Figure 30**: Innovation S-Curve ............................................................................................... 38
- **Figure 31**: Four Entrepreneurial Strategies ............................................................................. 40
- **Figure 32**: Architecture versus Disruption ............................................................................ 41
- **Figure 33**: Data Type Performance Cost Trade-off ................................................................... 43
Data Storage

Computer data storage is the technology of recording digital data in some medium in order to preserve the information. These storage mediums are often discussed and compared across various attributes that describe capability and measure performance. Because of the varying attributes the different storage mediums and technologies are used for unique storage needs. These use cases are described by industry specific terminology, including online, offline, tier, and temperature. The main technologies currently filling the storage requirements are solid state drives, hard disk drives, magnetic tape, and optical discs.

Data Types

Data is prevalent throughout the modern world. It is pervasive in personal lives, research organizations, government agencies, and corporations and is used for countless purposes. For data management, though, there are only three fundamental purposes.

- **Production data** is the primary data an organization uses for daily processes and operations. This essential data is served to clients and users, and must be readily available for frequent and rapid access.
- **Backup data** is a copy of all or part of the production data that is used for recovery purposes. Backups are typically short term as they reflect consistently changing production data.
- **Archive data** is primary data that is no longer needed for production purposes and so has been placed in long-term storage for retrieval when and if necessary.

Production data is the backbone type, and is the source of the other types. Backup data is a copy of some or all of the production data. Archive data is yesterday’s production data that is stored for posterity. Often backup and archive are conflated, so Figure 1\(^1\) displays some of the key differences. In short, archival data is primary information that is rarely accessed and backup data is a copy of primary information used for recovery.

---

\(^1\) "Seven Essential Strategies for Effective Archiving - Seven-Essential-Storage-Strategies-Wp.pdf."
Figure 1: Key differences between archival and backup

<table>
<thead>
<tr>
<th>Archive</th>
<th>Backup</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary information</strong></td>
<td><strong>Copy of information</strong></td>
</tr>
<tr>
<td>Used for <strong>compliance and efficiency</strong>:</td>
<td>Used for <strong>recovery</strong>:</td>
</tr>
<tr>
<td>• data retained in original form</td>
<td>• improves availability, allowing</td>
</tr>
<tr>
<td>• enables response to legal or</td>
<td>applications to be restored to a</td>
</tr>
<tr>
<td>regulatory action</td>
<td>point in time</td>
</tr>
<tr>
<td>• offloads information from</td>
<td></td>
</tr>
<tr>
<td>production systems and storage</td>
<td></td>
</tr>
<tr>
<td><strong>Typically long-term:</strong></td>
<td><strong>Typically short-term:</strong></td>
</tr>
<tr>
<td>• months, years or even decades</td>
<td>• days or weeks</td>
</tr>
</tbody>
</table>

Storage Attributes

Regardless of the data’s purpose, all storage technologies are described through the following attributes\(^2\) (ordered alphabetically). Generally, these attributes either describe a technical capability (accessibility, addressability, capacity, mutability, and volatility) or a performance measure. A technology’s attributes determine what use cases it may fulfill.

- **Accessibility**
  - **Random access** storage allows for any location to be accessed at any moment in roughly the same amount of time.
  - **Sequential access** storage only allows for access in sequence.

- **Addressability**
  - **Location-addressable** storage means each accessible unit of information is selected with its numerical memory address.
  - **File-addressable** storage means the information is divided into files that are selected through human-readable directories.
  - **Content-addressable** storage means each accessible unit of information is selected based on the contents stored therein.

- **Capacity**
  - **Raw capacity** refers to the total number of bits or bytes that a medium can hold.

---

\(^2\) Storage characteristic definitions are taken from "Computer Data Storage."
• **Storage density** is the capacity divided by the unit of length, area or volume.

• **Mutability**
  - Read/write storage or **mutable** storage can be overwritten at any time.
  - Write once read many (WORM) storage or **immutable** storage allows one write.
  - Slow write, fast read storage is mutable, but write is much slower than read.

• **Performance**
  - **Latency** is the time it takes to access a particular storage location.
  - **Throughput** is the rate at which data can be read from or written to storage.
  - **Granularity** is the size of the largest chunk of data that can be accessed efficiently without introducing additional latency.
  - **Reliability** is the probability of a spontaneous bit value change.
  - **Energy use** is the amount of energy used to store the information.

• **Vatility**
  - **Volatile** storage requires power to maintain the stored information. Typically, RAM (random access memory) is volatile.
  - **Non-volatile** storage retains the stored information even if not powered. Flash, hard disks and tape are all non-volatile.

**Storage Terminology**

The industry uses various sets of related terminology, often loosely, leading to confusion. The following terms are not a comprehensive list of terminology used, but cover the basic components of the vernacular. The terms online, nearline and offline generally refer to how quickly the data is accessible.

• **Online storage** is data storage that is immediately available for input and output.

• **Nearline storage** is not immediately available, but can be made online without human intervention.

• **Offline storage** is not immediately available, and requires some human intervention to become online (typically used for backup and archival).

---

1 "Nearline Storage."
Similarly, primary, secondary, tertiary and offline refer to the hierarchical “proximity” to the CPU and how the data is accessed, as opposed to how quickly.

- **Primary storage** is the only type of storage directly accessible to the CPU. It is often volatile RAM. Access speeds are typically measured in billionths of a second.
- **Secondary storage** is not directly accessible by the CPU. It is non-volatile storage, often stored on hard disks. Access speeds are typically measured in thousands of a second.
- **Tertiary storage** typically involves robotic mechanisms to access the data. It is non-volatile storage, often stored on hard disks or tape. Access speeds are measured in seconds or longer.
- **Offline storage** requires some human intervention for data to become accessible. It is non-volatile storage, typically used for backup and archival.

The data storage tier refers to the use-case-dependent, performance requirements of the data. Unfortunately, there is no definitional standard for tiers. In fact, there is not even an agreement regarding the number of tiers, with some sources defining Tier 1 – Tier 3 and other sources broadening to Tier 0 – Tier 5. The terminology discussed in Figure 2 takes a middle ground, with a total of five tiers, to allow for sufficient segmentation across performance and cost.

**Figure 2: Tier Terminology**

<table>
<thead>
<tr>
<th>Tier</th>
<th>Use Case</th>
<th>Example</th>
<th>Technology</th>
<th>Key Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier 0</td>
<td>Storage of extremely time sensitive, high value, volatile information</td>
<td>Financial transactions</td>
<td>SSD</td>
<td>IOPS / $</td>
</tr>
<tr>
<td>Tier 1</td>
<td>Storage of data used for business operations</td>
<td>Online retail</td>
<td>HDD</td>
<td>IOPS / GBs</td>
</tr>
<tr>
<td>Tier 2</td>
<td>Storage of data where speed is not imperative</td>
<td>Email</td>
<td>HDD / Tape</td>
<td>$ / TB</td>
</tr>
<tr>
<td>Tier 3</td>
<td>Storage of infrequently accessed data</td>
<td>Historical financials</td>
<td>MAID / Tape</td>
<td>$ / TB / Watt</td>
</tr>
<tr>
<td>Tier 4</td>
<td>Storage of data that has been archived</td>
<td>Long-term archival</td>
<td>Tape / Optical</td>
<td>$ / GB / Watt</td>
</tr>
</tbody>
</table>

---

4 "Computer Data Storage."
5 Note, offline is used in two ways but means the same thing in both situations.
6 Solid State Drive (SSD), Hard Disk Drive (HDD), Massive Array of Idle Disks (MAID), Inputs/outputs Operations Per Second (IOPS)
Other common terminology includes “hot” and “cold” data, which refers to how often the data is accessed. Again, there are no standardized definitions, but hot data is accessed the most frequently and cold data the least. In between the extremes are “warm” and “cool” data. These four sets of terms are not identical in meaning, however they are related, as shown in Figure 3.

Figure 3: Terminology Relationship

Storage Technology

The main storage technologies have each been around for at least three decades. Computer data storage, in the form of magnetic tape, was first patented in 1928. Nearly thirty years later, in 1956, hard disk drives were introduced by IBM, who then invented DRAM (dynamic random access memory) in 1966. Also in 1966, the first patent for compact discs was filed. Then in 1984, Toshiba introduced flash memory to the market. Since introduction, these technologies have seen continuous improvement in performance and capacity, and today are still the main storage technologies.

Magnetic memory stores data by writing non-volatile patterns in a magnetized medium. Magnetic tape (an example is cassette tapes), floppy disks, and hard disks are all examples of magnetic memory. The read and write operations are performed by a read-write head that detects and modifies the magnetism of the material. In order to read or write the medium must be spun to provide full access to storage locations. In addition to this sequential accessibility,

---

7 “Hard Disk Drive.”
8 “Dynamic Random-Access Memory.”
9 “Flash Memory.”
10 “Magnetic Storage.”
the characteristics of non-volatility and high density make this medium well suited for secondary and tertiary storage.

Figure 4: Hard Disk Drive (HDD)¹¹

Similar to magnetic memory, optical memory¹² stores data by writing non-volatile patterns on a medium that can be read via light. CDs, DVDs and Blu-rays are all examples of optical memory. The read and write operations are performed by a read-write head that uses laser to etch and detect bumps in the material surface. In order to read or write the medium must be spun to provide full access to storage locations, which allows sequential access. Typically, the medium is used for tertiary and offline storage.

Figure 5: Digital Video Disc (DVD)¹³

Semiconductor memory¹⁴ stores data in electronic integrated circuits on semiconductor materials. All solid-state storage, including flash memory and DRAM, are examples of

¹² "Optical Storage."
¹⁴ "Semiconductor Memory."
semiconductor memory. The read and write operations are performed electronically, meaning there are no moving parts and all semiconductor memory has the feature of random access (i.e., all classify as RAM). Typically, this type of memory is volatile, however, flash memory (NVRAM) is an example of non-volatile memory. This type of memory is ideal for primary storage (in fact RAM is nearly synonymous with primary storage), but non-volatile semiconductor memory is also used as secondary storage.

There are a number of technologies that offer promising storage breakthroughs. In the near term, 3D XPoint (a phase-change memory) is in development by Intel and Micron with claims of much lower latency and much faster throughput all with less energy usage. Another technology that has been in development for decades is holographic memory, which, if realized, promises higher density and throughputs than magnetic or optical. More extreme, research is also being performed to store data at the atomic level, a bit per atom, leading to very dense data storage. These technologies, alongside numerous others, aim to provide performance or density increases over current technologies.

Storage Market

Worldwide demand for data storage has been growing exponentially and is expected to continue into the foreseeable future. This is a healthy indicator for an already massive ~$3.5 trillion worldwide IT market. Within the overall IT market the worldwide enterprise storage

---

16 “Writing Data Onto Single Atoms, Scientists Store The Longest Text Yet.”
market is ~$90 billion, with nearly a third of that dedicated to storage systems. Within the broader storage market, the archival market itself is ~$5 billion and growing. Recently, the data storage market, alongside the broader IT market, has undergone a massive disruption with the introduction of IT as a Service (ITaaS). This new business model offers customers a different way to manage IT requirements and continues to change the landscape of the market.

Storage Demand

Worldwide demand for data storage has been growing exponentially, at a pace that is more than doubling every two years, as shown in Figure 7. Recently, social networks, mobile computing and video consumption have fueled this growth. With these trends enduring, and the looming data explosion from virtual reality and the internet of things, analysts and experts expect this growth to continue, estimating global demand reaching ~40 zettabytes\(^{17}\) (ZBs) by 2020. As an example of the magnitude of a zettabyte; if a gigabyte was a 1-inch cube, then a single zettabyte would fill more than 5 Empire State buildings.

![Figure 7: Global Memory Demand Sources](image)

While 40 ZBs is huge, if the digital universe continues to double roughly every two years then 2040 will see an enormous demand of 27,000 ZBs (27 YBs), as seen in Figure 8. Perhaps unsurprisingly, the worldwide installed storage capacity has not kept pace with this data.

---

17 A bit is a single binary unit. A byte is 8 bits, and is the number of bits needed to represent letters. A kilobyte (kB) is 10^3 bytes. Following a kB and growing by 10^n with each succession is megabyte (MB, 10^6), gigabyte (GB, 10^9), terabyte (TB, 10^12), petabyte (PB, 10^15), exabyte (EB, 10^18), zettabyte (ZB, 10^21), and yottabyte (YB, 10^24).
proliferation, so, much of the generated data cannot be retained. Alarmingly, analysts and experts do not expect this situation will self-correct, and even forecast a widening gap. This gap identifies a potential hole in the market capabilities.

Figure 8: Global Memory Demand

Market Size

The worldwide IT market is massive, estimated to be ~$3.5 trillion in 2016. The market is subdivided into Communication Services, IT Services, Devices, Enterprise Software, and Data Center Systems, as shown in Figure 9. Data storage technologies fit into both the categories of Devices and Data Center Systems. However, data storage technologies are only a very small piece of the Devices category, and the category is not a good fit for nucleic acid memory. For this reason, this paper will only consider the ~$170B data center system market.


19 Segmentation definitions provided by Gartner as follows:
Communications services — Consumer fixed services, consumer mobile services, enterprise fixed services and enterprise mobile services
IT services — Business IT services and IT product support
Enterprise software — Enterprise application software and infrastructure software
Devices — PCs and tablets, mobile phones, and printers
Data center systems — Servers, ECB storage, ENE and UC
An important macro-trend impacting the worldwide IT market is the shift towards cloud services. Cloud services include IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS (Software as a Service). Services include compute and storage, amongst myriad other services. Though many enterprises still prefer to keep their data on-premises (for security, performance, and regulatory reasons), recent years have seen cloud services beginning to supplant the traditional IT model, as seen in Figure 10. From a hardware provider’s perspective, this trend is mostly bad news. Cloud providers attempt to optimize their infrastructure for higher utilization meaning overall the number of units required for the same need decreases, impacting number of units sold. Additionally, the aggregation of hardware buying means that cloud providers can apply more downward price pressure. These two things hit both units and prices, negatively impacting revenues and profits. However, as can be seen in Figure 9, there was healthy growth in the Data Center Systems segment between 2014 and 2015, and growth is expected to continue. This is at least partially due to the continued data (and computing) explosion as seen in Figure 8.

21 In this figure, and in figures throughout unless otherwise noted, an asterisk (*) represents projections.
Regardless of the eventual landscape of cloud versus traditional, data center system hardware is required. Unfortunately, the Data Center Systems market is not solely dedicated to storage; importantly, it also includes servers. Figure 11 shows a slightly different segmentation of the IT market, presenting a picture of the ~$90B Enterprise Storage market, including all Systems, Services and Software. This provides a more accurate estimate of the total addressable market for storage. Within this market, storage hardware resides in the $37B Systems segment.

Figure 11: Worldwide Enterprise Storage Market

---

22 IDC Worldwide Quarterly Cloud IT Infrastructure Tracker Q2 2015
23 "Data Storage - Statista Dossier." Figure 12 and Figure 13 are also reproduced from the same source.
A few vendors dominate this storage systems market, as shown in Figure 12. EMC, HP, Dell, IBM and NetApp make up ~60% of the market, with the newly merged Dell-EMC accounting for nearly 30% of the market and HP holding ~15% market share. Of note, is a trend in the market towards original design manufacturers (ODMs), which is mirrored in the server market. ODMs are also called "white-box storage", and carry the implications of low-cost and generic. One driver of this trend is that large cloud service providers are opting for ODM. Regardless, the market has seen consistent growth, with a 5.4% CAGR over this time-period.

Figure 12: Worldwide Enterprise Storage Systems Revenue, by Vendor

![Figure 12](image)

Archiving Market

The archiving market includes products and solutions for maintaining legacy production data (such as email, documents, website content, and other structured data). Organizations typically archive important documents that need to be maintained but are only infrequently accessed. Cost gains are realized as the tier 4 technologies used for archiving (e.g. tape) trade-off performance characteristics for cheaper capital and operating expenses. Figure 13 shows five year projections for the global market, growing from $4.6B to $6.7B at a 10% CAGR. Assuming the splits seen in Figure 11, systems revenues account for ~40% of the market, yielding a ~$2B 2016 market for archival systems. Historically, archival solutions were implemented by highly regulated customers to maintain information for compliance purposes. However, recently this
has changed as organizations have noted the efficiency and cost gains of this type of data management.

Figure 13: Worldwide Information Archiving Revenue Projections

Gartner identifies three use cases for archives: historical archives, compliance archives, and analysis archives. Historical archives are retained as a record of business, housing content that can't be deleted for business reasons but is infrequently accessed. From a hardware perspective, this type of archive requires very high data integrity. Compliance archives house data that is retained in order to comply with regulations or audit requirements, often containing communications and business records. This type of archive demands high data integrity and low latency. Analysis archives contain high-volume non-production data that is frequently mined, analyzed and used for reporting. This type of archive requires low latency and high capacity. In addition to these type-dependent requirements, the longevity of the archives creates management challenges. Archive data is typically kept for more than ten years and often planned to be kept for up to one hundred years. This length means that the hardware will need to be refreshed multiple times through the period. Each migration of data represents costs as well as risks, which consumers are eager to avoid. As is seen in the broader data storage market, recent years have witnessed the industry trend to archival as a service.

24 "Avoiding the Key Traps of Archival Storage."
Industry Dynamics

In recent years, there have been two massive changes in the data storage industry. Firstly, the cloud created a new business model for providers as well as customers. Secondly, performance has improved to a point that customers are beginning to lose interest in ever faster hardware. These changes have led to a commoditization of the data storage industry.

The cloud is gaining ground on traditional IT because it creates more value for the consumer. As shown in Figure 14, while hardware providers do enable storage capability (the primary need), they also levy onto the consumer a number of unintended costs, burdening operations. Cloud service providers, on the other hand, remove those costs from the consumer and replace them with a simple pay per use operating expense. Furthermore, cloud providers are able to increase the benefits to the consumer. Thus, creating more value to the consumer by removing numerous cons while increasing the pros.

![Figure 14: Comparison of Hardware Provider and Cloud Service Provider](image)

<table>
<thead>
<tr>
<th>Storage Provider</th>
<th>Pros for Consumer</th>
<th>Cons for Consumer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware</td>
<td>Storage Capability</td>
<td>Capital Expenses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Hardware Purchases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Hardware Refreshes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Operating Expenses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Hardware Mgmt</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Energy expenses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Administration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Technology Mgmt</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Capacity Planning</td>
</tr>
<tr>
<td>Service</td>
<td>On-Demand Capacity</td>
<td>Usage Expenses</td>
</tr>
<tr>
<td></td>
<td>Software Functionality</td>
<td>Contract Management</td>
</tr>
</tbody>
</table>

Not only does the cloud service business model increase the value to the consumer but it also accrues value to the vendors. Firstly, due to scale and specialization, cloud providers are able to increase their purchasing power and decrease costs. Secondly, there are more axes on which to differentiate vendors. Vendors are able to offer different services based on pricing structures, performance agreements, and functionalities provided, and can more specifically target customer groups. For example, IBM has a strategy of focusing on building private clouds for...
their enterprise customers, while Google has announced that it will focus on catering to developers. This differentiation allows the vendors to compete on dimensions other than cost, which protects industry profits. Lastly, it helps address an issue with the industry as a whole, in that it increases switching costs and vendor lock-in: once a consumer is set up in one cloud it is costly to migrate to a new cloud. Traditionally, hardware refresh cycles create a natural break point where consumers can fairly easily switch vendors. Lowering costs and increasing differentiation allow service providers to capture more value while higher switching costs allow the vendors to maintain that captured value.

This change in business model is shown in the value curves comparison of Figure 15. Traditional IT reflects the model where hardware is sold to an IT shop which then services its consumers. This model allows maximization of performance and full control over security but sacrifices the ability to easily scale, forces management over all aspects, and has recurring capital expenditures as well as operating expenses (e.g., power and FTE requirements). The cloud service curve reflects the benefits of manageability and scalability ease, with very low capital expenditures. This new value curve is very suggestive of changing customer needs.
Figure 16 shows the top industry challenges in 2015 and 2016. Customers are migrating to the cloud because it provides solutions for: manageability (1) by removing the complications of managing and updating hardware; scale (3) by allowing for immediate capacity increases and decreases; capital expenses (4) by removing the necessity of purchasing and replacing hardware; and operating expenses (5) by eliminating the need for personnel to manage the hardware. Similarly, the data shows customers beginning to move past earlier criticisms of the cloud, namely performance (2) degradation (due to data transmission) and security (7) concerns. Whether this ranking of challenges is the cause or the effect of cloud migration is not clear, but either way it shows that traditional IT no longer best fulfills the customer need.

The other major industry trend is also highlighted in Figure 16, customers are less concerned with performance. This is not an indication of the need for performance decreasing but rather that hardware capabilities are beginning to surpass the need: current performance is good enough. This is not to say that there are no use cases for faster performance, as automated stock trading will always be able to benefit from slightly faster performance. However, the typical enterprise consumer is seeing less benefit from incremental improvements in performance, so that the value is often not worth the cost.

25 “Global Challenges with Data Storage 2015-2016 | Statistic.”
From an industry standpoint, this is good for cloud service providers (whose business model will always innately suffer some performance degradation due to data transmission), and bad for hardware providers who have historically differentiated based on performance, as well as other attributes. Further compounding this differentiation issue is that cloud providers are mature IT departments, with very sophisticated technical expertise and purchasing units. This means that cloud providers are exceedingly aware of their technical requirements and determined to lower costs, which has led to the ODM direct trend identified in Figure 12.

Consumers that are opting for cloud services which abstract the hardware and separate the hardware provider from the consumer; cloud providers that are sophisticated hardware buyers; and performance capabilities that are moving beyond consumer needs have begun to commoditize the storage hardware industry. This has led the storage hardware industry to compete more heavily on price, leading to further commoditization. So, despite the high barriers to entry that the storage hardware market exhibits, the hardware industry is difficult and likely to continue ceding ground to storage service providers.

**Nucleic Acid Memory**

Nucleic acid memory is the storage of data by encoding the information into the medium of nucleic acids. This is often called DNA storage, as typically, but not necessarily, the information is stored in the nucleobases that comprise DNA (deoxyribonucleic acid): adenine (A), thymine (T), cytosine (c), and guanine (G). For simplicity, the rest of this paper will assume DNA as the medium of choice for NAM, and will refer to DNA based NAM as DNAM. Baum first introduced
this idea in 1995,26 with early backing by Clelland in 199927 and Bancroft in 2001.28 The following decade saw experimental and theoretical advancements,29 but it wasn’t until 2012 that Church30 proved the idea on a larger scale, when the book Regenesis was stored and recovered. Today, NAM is at a technology readiness level (TRL) two, the stage where practical applications of the underlying science can be developed, however remains economically unrealistic for mass storage.

Process

At the highest level the system for DNA information storage is simple, as shown in Figure 18.

Figure 18: High Level NAM System31

Digital information is written to DNA strands, which are stored, and then read back as digital information, when required. However, while accurate and immediately similar to other types of data storage processes, this simplistic representation does not provide enough detail into the writing and reading processes. The writing process should be decomposed into encoding and synthesizing, while the reading process should detail sequencing and decoding, as shown in Figure 19.

26 “Building an Associative Memory Vastly Larger Than the Brain on JSTOR.”
27 Clelland, Risca, and Bancroft, “Hiding Messages in DNA Microdots.”
28 Bancroft et al., “Long-Term Storage of Information in DNA.”
29 Yim et al., “The Essential Component in DNA-Based Information Storage System.”
30 Church, Gao, and Kosuri, “Next-Generation Digital Information Storage in DNA.”
31 This figure, and subsequent similar figures, uses Object Process Methodology (OPM) to depict the system. In the OPM ovals represent processes, and rectangles represents objects. Shadows represent physical instances, while a lack of shadow indicates informational instances.
Figure 19: NAM Writing and Reading Process Decomposition

Figure 20 shows this process as a linear progression of events. Firstly, the 0 and 1s of any series of binary digital data is translated to a series of A, T, C and Gs. Once this ATCG series has been designed, it is written to a synthetic strand of DNA. This strand of DNA, containing the information, is then stored until the information needs to be accessed. The strand of DNA is then sequenced to determine the ATCG series. Finally, the series is decoded to translate the ATCGs into binary digital form. The past two decades have seen significant improvements in these areas, due to both concentrated research as well as market advancements.

Figure 20: High Level NAM Process

**Encoding**

Simply, the encoding process takes the digital binary data (a series of 0s and 1s) and translates that binary series into an ATCG series. For example, the series ‘11100001’ could be translated into ‘GCAT’, via the quaternary map: 00->A, 01->T, 10->C, 11->G. This type of function results in an immediate increase in density as the mapping utilizes the full base-4 potential of DNA. However, due to the current state of synthesis and sequencing technologies, this type of

---

32 In OPM the solid triangle indicates a decomposition, in this case, of a process into sub-processes.
33 The Decoding process is the inverse of the Encoding process, so is not separately discussed.
mapping could lead to issues if a series of the same base letter were written, if a segment contained a homopolymer. For this reason, a rotating ternary scheme, similar to that shown in Figure 21, can be used to reduce the error potential.

Figure 21: Rotating Ternary Scheme

<table>
<thead>
<tr>
<th>Previous Nucleotide</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ternary Digit Translate</td>
<td>0</td>
<td>T</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>2</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
</tr>
</tbody>
</table>

In addition to this translation, the encoding process also includes data preparation, occurring prior to translation, as seen in Figure 22. This sub-process can include various types of preparations. For example, a compression process might be applied to the data in order to reduce the number of required synthetic bases. Or preparation might include an encryption step to increase security. Another process that is likely to be included is data tagging, to allow for easier data manipulation. These preparation steps are optional and may in fact be performed by the data owner before providing the data for storage.

Figure 22: Encoding Process Decomposition

34 Augmented from Bornholt et al., "A DNA-Based Archival Storage System."
The final step in the encoding process is designing ATCG blocks. Due to the current state of synthesis, there is a limit on the number of bases that can be synthesized into a strand of DNA, typically representing only kilobytes of information. Therefore, if a piece of information, larger than this limit, needs to be stored the ATCG series is broken down into blocks. Alongside the information to be stored, each block must contain an address indicating its location in the ATCG series. Finally, these blocks can be tagged with other pieces of information to aid data retrieval. In total, each block must be short enough that it can be synthesized into a strand of DNA, and the set of blocks must fully cover the information to be stored, as seen in Figure 23. Another function of this designing process is to increase storage reliability. For example, some schemes use a block cover with overlapping information segments, to provide data redundancy. Another scheme, proposed by Bornholt, incorporates redundancy by taking the exclusive-or of two information payloads to form a third, so that any two of the three blocks can produce the third. The implicit trade-off in increasing reliability, in this way, is a reduction in data density, and generally, much of the designing process suffers the inevitable downside of decreasing data density. Future increases in both synthesis and sequencing technology will mitigate decreases in data density from designing.

Figure 23: ATCG Block Representation

ATCG Blocks

- Information
- Address
- Tag(s)

A few hundred to a few thousand base pairs

Synthesizing

Synthesis is the transcribing of digital data into physical DNA strands. The market for synthetic oligonucleotides is maturing but most use cases are research oriented and so the industry has not yet experienced major growth or the resulting major breakthroughs. Currently, synthesis involves chemically building the strand, nucleotide by nucleotide, akin to printing. One such methodology creates numerous such strands in parallel, organized on a grid. To simplify, at
each step a particular base is chosen, for example T. All strands for which T is not the next base required in the series are masked, leaving all T requiring strands unmasked. A solution of Ts is washed over the grid, allowing the T bases to bind where necessary. This is repeated for the next chosen base, in time growing the strands, base by base. Methods like these can grow millions of different oligonucleotides in parallel, however it is a costly and slow process.

Storing

Once the DNA strands have been synthesized they must be stored. Storage can be envisioned as a vast library of DNA pools, where each pool is contained within a vial and arranged systematically for automated access. While this idea presents a framework for one potential solution, it is far from the only way in which storage could be accomplished. For example, on one extreme, all the data could be stored in the same DNA pool. Or, on the other extreme, each separate block of data could be attached to a specific physical location on a chip. Much work still needs to be done to ensure cost effectiveness, access efficiency, minimal cross-talk, and memory retention. However, it is clear that storing the data in a cool and dry environment enables lasting memory retention, as seen in Figure 29.

Accessing

Accessing refers to the identification and retrieval of specific strands of DNA for reading, upon a data request event, as shown in Figure 24. This step was not represented in high level process maps because it is technically optional. It is possible to not access any one piece of information, but rather to sequence the entirety of the DNA pool in order to retrieve any information. This is, in fact, how many researchers currently perform retrieval. However, in order to commercialize this technology, for efficiency and cost reasons, accessing specific pieces of information is vital.

---

38 Cross-talk refers to two separate strands of DNA, with complimentary segments that bind with one another, leading to entangled strands.
One way that has been proposed to access specific strands of DNA is through polymerase chain reaction (PCR). During the design phase of this method, the DNA strand is tagged with a primer. When the data needs to be accessed an enzyme is released that targets the specific primer, and once found replicates the strand. This process can be repeated until a desired number of copies are created. Then the DNA pool is randomly sampled until the information is fully recovered. While this methodology does mean that the entire pool need not be sequenced, it has a number of downsides. Firstly, with each successive data access the pool becomes further diluted with replicated data. Secondly, each access requires the consumption of a significant amount of raw materials. Lastly, as this relies on random sampling, the method is not precisely targeted and so not very efficient. Much work still needs to be done to identify an access method providing random access, low latency, information integrity, and cost effectiveness.

Sequencing

Sequencing is the process of determining the series of bases in a given DNA strand. Over the past 15 years, and since the Human Genome Project, the sequencing industry has witnessed massive growth and seen technological breakthroughs. This has resulted in a cost reduction of 100,000 – 1,000,000-fold in the past 15 years, as can be seen in Figure 25. With these

---

39 Bornholt et al., “A DNA-Based Archival Storage System.”
breakthroughs, the market size has grown to $4 billion and is expected to reach $10 billion over the next five years.  

Figure 25: Cost per Megabase of DNA Sequencing

There are numerous sequencing methods, but one common methodology uses a polymerase enzyme. Utilizing fluorescent nucleotides, the enzyme creates a complement of the DNA strand. This compliment can then be optically read as each nucleotide emits a different color. This methodology, alongside the others, can be error prone. In particular, due to the specificity of this technology, a series of the same base letter may not be correctly identified, e.g., two bases might be read as a single base. Typically, errors such as these are “designed” away and mitigated by reading at depth, meaning sequencing the same strand multiple times, or sequencing multiple copies of the same strand. These reliability and redundancy measures lead to high information validity.

Process Diagram

While this discussion abstracts away much of the technical detail inherent in the process, it provides a foundation of understanding for the NAM data storage process, as seen in Figure 26.

---

40 “Next Generation Sequencing (NGS) Market by Platforms & Application - 2021 | MarketsandMarkets.”
41 Bornholt et al., “A DNA-Based Archival Storage System.”
Features

NAM has a number of features that make it very promising as a data storage medium. The three that are typically highlighted are capacity density, data retention (i.e., durability), and energy usage. Figure 27 compares standard storage technologies to Cellular DNA, across these traits. While this represents a comparison to cellular activities, it shows the limits for DNAM are orders of magnitude better than standard technologies. In addition to these three traits, NAM has other enticing characteristics, such as non-volatility, manufacturing simplicity, and organic relevance (particularly DNAM).
Figure 27: Comparison of Standard Memory Technologies to Cellular DNA Attributes

<table>
<thead>
<tr>
<th>Metric</th>
<th>SDD (Flash)</th>
<th>HDD</th>
<th>Tape (LTO Gen 7)</th>
<th>Cellular DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volumetric density</td>
<td>~ 10^{16} bit cm^{-3}</td>
<td>~ 10^{13} bit cm^{-3}</td>
<td>&gt; 10^{11} bit cm^{-3}</td>
<td>~ 10^{13} bit cm^{-3}</td>
</tr>
<tr>
<td>Retention</td>
<td>~ 10 years</td>
<td>&gt; 10 years</td>
<td>15 - 30 years</td>
<td>&gt; 100 years</td>
</tr>
<tr>
<td>ON power</td>
<td>~ 0.01-0.04 W per GB</td>
<td>~ 0.04 W per GB</td>
<td>~ 0.005 W per GB</td>
<td>&lt; 10^{-10} W per GB</td>
</tr>
</tbody>
</table>

Density

NAM is tremendously dense, in terms of capacity, for three reasons. Firstly, the medium itself is extremely small: the diameter of a double-stranded DNA molecule is about 2 nanometers. Secondly, the medium is inherently volumetric, unlike other storage mediums, allowing for better utilization of space. Lastly, NAM is not restricted to binary: DNA has 4 bases so could be base-4, while theoretically other nucleic acids structures could contain 6 bases or more. These factors provide DNAM with a raw density of 10^{9} GB/mm^{3}, and NAM with an even greater potential density. If DNAM’s density potential were realized, the global storage needs could fit into a truck-bed with room to spare. Undoubtedly, the technically achievable density will be less than the theoretical limits, but even achievable limits will be a vast improvement over standard technologies, as can be seen in Figure 27.

The increase in data storage density, seen in recent decades, has been one of the driving forces of the technological revolution. Following Moore’s law (~40% / year), density has doubled roughly every two years, with a 1,000-fold increase occurring in roughly 20 years. However, recently that trend has not held because areal density limits are being reached, as shown in Figure 28. If these yearly improvement numbers were to continue (without further slowing) it would take between 23 and 42 years to see a 1,000-fold increase. To combat this, corporations are hoping to begin harnessing volumetric density to increase growth rates.

---

42 Augmented from Zhirnov et al., "Nucleic Acid Memory."
43 LTO Gen 7 numbers from "Areal Density (Computer Storage).", "LTO-Drives-DS.pdf.", "Oracle StorageTek LTO Data Sheet - 033631.pdf."
44 Bornholt et al., "A DNA-Based Archival Storage System."
45 Zhirnov et al., "Nucleic Acid Memory."
Figure 28: Technology Areal Density Increases & Timescales

<table>
<thead>
<tr>
<th>Areal Density Increase (2008 – 2013)</th>
<th>Projected Years for 1000-fold Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAND</td>
<td>35% / year</td>
</tr>
<tr>
<td>LTO Tape</td>
<td>28% / year</td>
</tr>
<tr>
<td>Optical BD</td>
<td>18% / year</td>
</tr>
<tr>
<td>HDD</td>
<td>18% / year</td>
</tr>
</tbody>
</table>

Durability

NAM is also very durable in that it can retain information for long periods of time, as is evidenced by humans collecting and sequencing DNA from ancient fossils. In nature DNA is impacted by mutation, pH, radiation, and mechanical forces amongst other considerations, however many of these issues will be controlled in NAM storage, creating a much more durable medium. This retention capability is especially potent when the NAM is kept in cool dry places, as seen in Figure 29.

Energy

While the Cellular DNA energy consumption is miniscule, as seen in Figure 27, in practice NAM is not likely to achieve these promises. During storage, the NAM energy consumption

---

47 "Fontana_Volumetric Density Trends for Storage Components -- LOC 09222014.pdf."

48 Altered from Zhirnov et al., "Nucleic Acid Memory."
requirements will be exceedingly small, but implementations of NAM will very likely involve mechanical and biochemical sub-processes during writing, accessing, and reading, which will greatly increase energy consumption. These NAM implementations will likely be similar to tape or MAID libraries, where nearly all energy consumption occurs during write, access and read processes. So, for purposes of this paper, it will be assumed that the energy requirements per GB are similar to competitive technologies. Despite this assumption, there are still energy savings. A NAM implementation will consume less energy than other technologies due to the density of the medium, which will allow for massive physical space savings, thereby reducing the facility’s power consumption to near zero. Since this is a derivative of the density, it will be assumed as a benefit of the density.

Other

NAM possesses other innate characteristics that make it enticing for memory storage. Firstly, NAM is non-volatile, it will maintain the stored information even without power. This is one of the factors that reduces NAM energy consumption, once the information has been written the energy supply can be drastically reduced. Additionally, it provides for information security through potential disaster scenarios. Secondly, while the read and write processes are intensive, the manufacturing of raw memory capacity is simple and inexpensive, in that the bases required for memory can be efficiently produced. This is not the case for standard technologies, where the costs are mainly related to precise capacity manufacturing. Thus, capacity production can be quickly matched to meet capacity demand. Lastly, NAM, and especially DNAM, will remain relevant from a scientific, technological and societal perspective for centuries given the biological relevance of the material. Thus, it is unlikely that society will lose the capability to read the information written to NAM, regardless of technological changes.

Furthermore, a NAM system can be designed to harness other characteristics. NAM can be made to be a random-access memory. Researchers have used PCR to selectively amplify certain strands, information blocks, within a pool proving a RAM architecture. Undoubtedly, this RAM technique will be further improved and likely even eclipsed by other techniques. Additionally, NAM’s reliability can be easily tweaked to the desired level of precision, which can be designed
into the schema in a number of ways (e.g., XOR encoding). A very simple example of this is to use the inherent capability of DNAM to self-rePLICATE in order to create numerous copies of the information. Finally, NAM synthesis, access and sequencing processes can be made massively parallel, increasing throughput and access speeds.

Issues

NAM has been proven on a small scale in research laboratories and is beginning to attract market interest, most noticeably from Microsoft. Currently, the main impediment to commercialization is the cost associated with each additional write and read. Synthesis, in particular, is exceedingly expensive, costing ~$1,000 - ~$10,000 per MB. When comparing to other technologies, like tape on the low-end (~$10 per TB) and SSD on the high-end (~$100 per TB), NAM synthesis alone (~$1,000,000,000 - ~$10,000,000,000 per TB) is prohibitively expensive. Fortunately, the sequencing is not nearly as expensive at ~$1 - ~$1 per MB. This drastic price difference is due to the benefits of the Human Genome Project and a healthy market. However, when scaled these costs are also prohibitive (~$100,000 - ~$1,000,000 per TB). While both of these technologies have been progressing at rates exceeding Moore’s Law, there is still a lot of room for improvement. On the surface, this implies that costs will likely need to drop by more than 10,000,000-fold in order for NAM to compete in the market.

In addition to economic considerations, is the issue of write and read throughput, the time required to write and read. Once again, synthesis is the largest impediment. For example, assuming a 24-hour run can synthesize 1,000,000 strands, with 100 base pairs per strand, at 4 bits per nucleotide, means that it would take 20 days to turn around a GB, and 20,000 days which is roughly 55 years to synthesize a TB. Likewise, gains need to be made on sequencing throughput. For instance, the Complete Genomics NGS platform can sequence 3,000 giga-bases over 11 days. Assuming, again, the 4 bits per base, this is 12,000 gigabits or 1,500 GBs, which

---

49 Bornholt et al., “A DNA-Based Archival Storage System.”
50 Zhirnov et al., “Nucleic Acid Memory.”
51 Goldman et al., “Towards Practical, High-Capacity, Low-Maintenance Information Storage in Synthesized DNA.”
52 Erlich and Zielinski, “Capacity-Approaching DNA Storage.”
53 Zhirnov et al., “Nucleic Acid Memory.”
54 Massive Parallel Sequencing.”
translates to ~5 GB per hour. While neither of these examples account for potential efficiency gains from things like parallelization, they do show that improvements are needed here as well.

Outlook

NAM is a theoretically viable product, however at a TRL two is not close to commercialization. Currently, NAM is still in the gestation period on the innovation S-curve (Figure 30). There is a lot of work that needs to be done to make NAM a feasible product. Most obviously, costs need to drop drastically. There is a lot of hope for this, in that costs have been dropping quickly in recent years. More promising is that there is still a lot of maturity to be realized in the science of synthesis and sequencing. Additionally, there is a lot of automation that can and must be achieved in the technologies. And as the tangential markets grow, economies of scale and learning will be realized. Beyond costs, many of the details of the technology still need to be worked out. For example: What encoding will optimize density? How can strand length be extended to allow for more data per strand? How should the strands be stored to optimize for durability and efficiency? How can specific strands be efficiently accessed and retrieved from a large pool of strands? How can sequencing errors be overcome? The progress that must be made is overwhelming, but once achieved the technology holds great value.

Figure 30: Innovation S-Curve

---

NAM Strategy

In developing a strategy to bring nucleic acid memory technology to the market, it is important to discuss two distinct time periods. These time periods are prior to and after the point where NAM becomes cost competitive with other storage technologies in the market. Prior to this point NAM is not a viable market offering as a mainstream data storage product. That said, there may be other use cases for the underlying technology, and may even be a few extremely fringe customers for whom the features of NAM provide inherent value. After this point, NAM can enter the market as a data storage technology. This section mainly discusses the market entry strategy for after this point. Further, this section does not consider any strategies where a company could enter as tangential to NAM, for example as a software company that will optimize NAM storage.

NAM should enter the data storage market through archiving, by targeting large storage service providers and large data-intensive corporations with on-premise operations. These companies will gain the most value from a NAM product as it reduces their capital and operational cost base by millions of dollars per year. Since these companies depend on the valuable data to be housed by NAM it is unlikely that they will trust any new technology that is not finalized and proven. Therefore, a methodical and controlled entry is preferable to a quick, iterative execution strategy. As NAM is an entirely new paradigm in data storage it will not be possible to enter the market by collaborating with established players. Instead it will be necessary to take a competitive stance and face the incumbents directly. Hence, an architecture strategy should be employed to enter the market, relying on control over underlying ideas and partnerships to barricade the company from competition.

Framework

MIT Professor of Management of Technology Scott Stern has developed an entrepreneurial framework\textsuperscript{54} to help startups determine how to create and capture value. His hypothesis is that there are four choices a founder must make. The founder must choose the technology, the

\textsuperscript{54} Gans, Stern, and Wu, “Foundations of Entrepreneurial Strategy.”
customer, the competition and the company’s identity. To assist in these choices, he has further identified four types of entrepreneurial strategy: intellectual property, architecture, value chain and disruption. Choosing which of these strategies to employ helps to narrow the remaining choices, allowing the company to determine how to create and capture value.

The four types of entrepreneurial strategy can be defined by the company’s choice of whom to compete against and the choice of how to compete. The choice of whom to compete against defines the company’s orientation towards competition versus collaboration and the choice of how to compete defines the company’s investments in execution versus control, as shown in Figure 31. An IP strategy involves investing in control over technology while collaborating with established players to enhance value for the players’ customers. An architecture strategy involves controlling a core idea while developing a brand-new value chain to compete against existing value chains. A value chain strategy involves investing in quick execution and collaborating with existing market players to enhance an established value chain. A disruption strategy involves investing in quick execution in order to compete and ultimately replace existing technologies and value chains.

Figure 31: Four Entrepreneurial Strategies

In considering the options for NAM technology, neither intellectual property nor the value chain strategy fit the circumstances. The IP strategy involves integration into the existing

55 Ibid.
products or services of existing players. However, in the case of NAM, the technology involved is completely different, in terms of process and expertise, from currently used technologies, so integration of IP would not be possible without a completely new product. This strategy would be more applicable for a company that has developed a methodology to increase the capacity of hard disks. The value chain strategy involves integrating a new business into an existing value chain in order to provide more value to the end user. But, in the case of NAM, the end user should see no difference in the storage capability provided. Removing the two collaboration strategies, leaves the competition strategies and the choice between execution or control.

The remaining strategies, architecture and disruption, have similarities but are very different, as Figure 32 shows. Both aim to harness a nascent technology, as it climbs the S-curve, to create value. Both must develop a new value chain in order to capture the value. Both are competitive strategies, where the company is challenging established industry players. The differences between the strategies derive from how they compete. The architectural strategy competes through control while the disruption strategy competes through execution. This difference results in a deliberate and thorough architecture versus a flexible and quick disruption. The architecture strategy prudently picks the ideal customers and shields itself from competition. The disruption strategy targets an under-served customer in order to isolate itself from competitive response until it has an established advantage. The architecture strategy carefully constructs a knowledgeable team to create a differentiated enduring value chain. The disruption strategy rapidly assembles a nimble team to iteratively develop a value chain that will eventually supplant incumbents.

Figure 32: Architecture versus Disruption

<table>
<thead>
<tr>
<th>Technology</th>
<th>Architecture</th>
<th>Disruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competition</td>
<td>Harness nascent technology</td>
<td>Harness nascent technology</td>
</tr>
<tr>
<td>Customer</td>
<td>Orient for competition, invest in control</td>
<td>Orient for competition, invest in execution</td>
</tr>
<tr>
<td>Identity</td>
<td>Select ideal customer</td>
<td>Target an under-served tail of market curve</td>
</tr>
<tr>
<td>Value Creation</td>
<td>Deliberate, thorough, knowledgeable</td>
<td>Rapid, nimble, talented</td>
</tr>
<tr>
<td>Value Capture</td>
<td>New value curve to compete against incumbents</td>
<td>New value curve to eventually supplant incumbents</td>
</tr>
</tbody>
</table>

56 Ibid.
A hallmark of the disruption strategy is speed to market. This speed allows the company to gain a foothold with a customer niche before competitors react. Once the foothold is established, the company leverages it to expand into tangential customer segments. To achieve the speed, the company sacrifices finished products for iterative development. For this strategy to work, the customer must be willing to accept these unrefined products, in exchange for products tailored to that underserved customer’s needs. In the case of NAM, a company would be selling an unproven technology to store valuable data. It seems unlikely that a customer would be willing to trust anything but a proven final product, as a failure could jeopardize business. Additionally, as NAM is not currently economical, speed to market does not seem appropriate.

Alternatively, the architecture strategy is control. This control allows the company to protect the idea as well as thoughtfully develop the product. While typically this strategy takes longer to fruition, the end result is a market ready product that is both protectable and tailored to a specific customer. Though the architecture strategy is more difficult and risky to implement, if a product comes to market the company is usually more stable than one produced by disruption, due to the protection. So, for NAM, this type of product would engender more customer trust as the technology and company have been de-risked through the development process.

In the case of NAM, the architecture strategy is more appropriate than disruption as it will produce a product that customers are more likely to adopt. This might seem surprising given that Clayton Christensen’s acclaimed “The Innovator’s Dilemma” uses the data storage industry as the prime example of a disruptive strategy, but the circumstances are very different. Christensen focuses on disruption within the disk drive industry, which at the time of his example was an already established technology for data storage. The technologies harnessed were innovations of disk drives, altering the performance and density capabilities of the drives but not drastically changing the underlying technology. Fringe and underserved customers were willing risk an early chance on new technologies because the heart of the technology was trusted. NAM, though, is an entirely new data storage technology, requiring a leap of faith.

With the architecture and technology chosen, and the internal identity implied, there are two choices remaining in determining an entrepreneurial strategy for NAM: the customer and the competition. Before these choices can be made the market needs to be determined.
Market

Similar to the tier system shown in Figure 2, different data types require differing levels of performance. As the performance requirements decline, organizations opt to utilize less expensive technologies. SSD and HDD typically dominate the high-performance need, while tape and optical are used where cost is more important than performance. As shown in Figure 33, performance requirements are heavily correlated with the data type. The performance characteristics typically involved are latency, throughput, and density where reliability is assumed. The costs consist of capital costs and operating expenses (e.g., energy consumption). Production data is extensively used and repeatedly served to demanding users and customers, so performance is critical. Backup data is used to restore production capabilities, so although not always needed, it must maintain decent performance characteristics to restore functionality. Archive data is both rarely accessed and typically not rapidly needed, so performance requirements are minimal.

![Figure 33: Data Type Performance Cost Trade-off](image)

Currently, NAM’s performance characteristics (i.e., latency and throughput) are below archiving standards, and NAM’s costs are well above production standards. As the NAM technology matures, presumably both costs will come down and performance will improve. Since performance capabilities are the decisive factor in determining for which use case a technology can be utilized, it follows that as the performance of NAM improves it will first be viable for the least performance intensive use, archival. At that point, if costs are low enough, there will be a potential to enter the archiving market. If the costs are still too high then the market will have to wait until backup performance requirements are met, or perhaps again until production requirements. It is possible that performance does not improve very much. In this case, if costs also do not improve much then there will be no possibility of NAM as a data storage product.
On the other hand, if costs were to improve precipitously one might be able to find an archiving use case for a niche clientele who is willing to accept slow performance for durability, density or some other inherent value trait. Since in all cases archives represents the first possible point for market entry, this paper will assume it as the beachhead market for NAM.

Another reason to choose archives as the beachhead market is that an archive is the terminal location of infrequently accessed data. Technologies that house production or backup data see a constant turnover of data. Data is written, accessed, moved and removed. Any technology utilized in these use cases will undergo continuous write and read operations. For technologies currently deployed for these activities, write and read operations are quite inexpensive. However, for NAM, write (synthesis) and read (sequencing) operations are the most expensive processes, quickly escalating costs. Archives, on the other hand, often see write once, read occasionally, keep forever operations, so synthesis and sequencing will be infrequently faced.

Customer

Having identified the beachhead market for NAM, the customer can be chosen. One of the benefits of the architecture strategy is that it allows the customer to be strategically chosen. To create the most value for the customer, a significant need and fit must be identified. As a beachhead customer, the ideal choice would be an influencer within the industry. Further advantage can be gained if this customer is also underserved by or dissatisfied with current vendors.

In order for customers to elect NAM technology, the benefits of NAM need to create value. As has been previously stated, the main benefits of NAM are durability and density. Within the archive market, NAM’s durability creates more value the longer the data is stored. Specifically, NAM’s value increases with every hardware refresh and data migration that would be required by a competing technology. Every avoided hardware refresh and data migration translates directly into cost savings from hardware and manpower. Similarly, NAM’s density creates more value the larger the volume of data stored. NAM’s value increases with every cubic foot of space that would be required by a competing technology. Every avoided cubic foot translates directly into capital expense and operational expense savings. The capital savings occur because
a large facility is not built (or leased) to house the hardware. The operational savings occur because that facility no longer needs to be powered and because less manpower is required to tend the hardware itself. These NAM features clearly indicate that most value is created for large long-term archives, so this will be the best potential point of entry. Within this customer category, there is still the possibility of capturing value from large customers or many small customers.

Consider selling to many small customers and assume that the economics are such that NAM is only cost competitive for large long-term archives, since that will be the first point of entry for NAM. Because the economics do not work for small archives, no small customer would be willing to purchase NAM hardware. Meaning the company would need to provide storage as a service to many customers requiring long-term archival, whereby amassing a large volume. However, at the outset the company will be without customers or data, and so the economics of NAM, again, do not make sense for multiple years until time and volume accrue (meaning the company itself would probably opt for another technology until that point). To get to that point, the company will need to acquire many customers. To do this, the company will have to differentiate itself from competitors. In the storage service market, customers do not care which technology provides the service, as long as the service level agreements are met, so NAM’s features are no longer a selling point. The main selling points for storage service providers are service level (e.g., archival length, data integrity, data security, retrieval rates), price and functionality (ease of use). There will be numerous companies able to provide similar service levels, leaving the company to compete on price and functionality. Since the NAM costs only make sense with time and scale, this means a NAM-based company would at best be price comparable, so the company will need to compete on functionality. However, to differentiate on functionality the company will need to develop core competencies in software development as well as those required for the hardware development. Thus, initially, it would be very difficult to sell hardware to small customers because the economics do not make sense, and very difficult to enter the market as a service provider because the economics do not make sense, the customers do not care about NAM’s features, and a service provider needs to develop a core competency of software development.
Similarly, consider selling to large customers. Selling storage as a service to large customers is not as difficult as selling to many small customers, but there are still problems. From an economic standpoint, the company would be better off because it would only need a few large customers for the NAM economics to work. While a large business would care more about the technical qualifications of NAM than a small customer (or individual), especially as it relates to cost, the large customer would still care more about the SLA and functionality provided, forcing the company to develop a core competency in software development. On the other hand, NAM as a hardware provider has the benefit of being able to target relatively few large data centers with a clear selling point that utilizes the NAM features, and allows the company to develop NAM related core competencies. So, the NAM technology fits best for customers with large volumes of long-term archive data and a storage hardware need.57

There are a number of potential customer groups that meet these criteria. The most natural choice is large archiving service providers (e.g., HP, Veritas) who amass huge amounts of long-term archive data. Even though many of these service providers are also hardware providers who would be direct competitors and exceedingly unlikely customers (e.g., Dell EMC, HP, IBM), there are numerous service providers that are not hardware providers. Large data-intensive corporations who run on-premises data centers for operational efficiency and cost reasons (e.g., Amazon, Facebook), are another natural choice. Additionally, many of these organizations are also storage service providers (e.g., Amazon, Microsoft, Google). A third potential customer choice is large highly regulated organizations with locally stored data (e.g., government agencies, health care institutions, nuclear power plants). Lastly, data-intensive research organizations (e.g., World Data Center for Climate, CERN) often generate and house massive data that is infrequently accessed.

Of these potential customer groups the most visible and influential customers are archiving service providers and data-intensive corporations. At the intersection of these groups, are companies that have both data-intensive operations and provide archiving services. These companies have massive data storage needs and spend billions of dollars on that storage. These

---

57 Note, positioning the company as a hardware provider does not rule out the possibility of a business model with a centralized service component. For example, it might make economic sense to synthesize centrally and then distribute the stored data to hardware maintained on the customer's site for storage and data retrieval.
companies are leaders that are willing to take a risk on a new technology if long-term financial
gains can be made. These companies are trendsetters that push the boundary of information
technology. These companies are the ideal customers for NAM, as they would be influencers in
the community and because the technology can create a lot of value for them. A final reason to
choose this group as the target customer is they have shown signs of being poorly served by the
current vendors, as witnessed by the trend towards ODM direct providers (Figure 12).

Competition

Two things must be considered when determining the competition: who to compete against
and what core idea to protect. As the architecture strategy involves developing a new value
chain, the company needs to decide in which parts of the chain of activities it plans to compete.
This decision can help the company focus resources in the most critical areas while developing
collaborative partners in other less critical areas. Similarly, the company needs to devote
resources towards developing protection around the core ideas of the venture.

There are two types of potential competitor groups: those against whom the company will
directly compete and those that could begin developing NAM technology. By positioning the
company as a hardware provider, it will be directly competing against other storage hardware
providers (e.g., Dell EMC, HP). Additionally, there are other potential competitors in the NAM
space, as seen in Figure 34. The target customers, service providers, are also potential
competitors. There are potential competitors in the biotech space, in the form of synthesis
companies and sequencing companies. There are research organizations working to develop
the technology and of course there are NAM startups breaking onto the scene. Notably, some
startups have already formed stating their development intent, and Helixworks has already
released a product for sale through Amazon.\(^58\) The “DNADrive from Helixworks Technologies is
the world’s first commercially available DNA storage medium...[offering] 512 Kilobytes of data
storage in specially encoded DNA, encapsulated specially in a custom gold pill.” While this is too
small to be useful for personal storage let alone commercial applications it is significant in being
the first NAM product on the market.

\(^{58}\) https://www.amazon.com/Helixworks-dsDNA300-750hw-DNADrive/dp/B01IVSH4OM
In order to succeed in bringing NAM to the market, any player will need four things: storage expertise, scientific expertise (nucleic acid expertise), funding for data storage R&D, and an impetus to enter the market that overrides impediments. Besides NAM startups with the clear intent of entering the market, of the other four potential competitors, storage hardware and service providers are the most likely to enter. Synthesis and sequencing players, while possessing the scientific expertise and potentially being lured by a new, lucrative and tangential market, are unlikely to enter because their own markets are still immature and contain numerous other opportunities for expansion. Research institutes possess the requisite expertise but lack the strong impetus to enter the market. Storage providers, lack the scientific expertise, but have a very large impetus as NAM directly impacts their core businesses. Turning as many of these potential competitors into partners will enable the company to insulate itself from future competition.
A number of pieces need to come together to deliver a NAM hardware product to a customer, as seen in Figure 36. While all these pieces are important to a successful final product, the sequencing, storage, access and synthesis are vital. Of these four pieces, only storage and access are core to NAM. The growing and competitive synthesis market contains numerous technologies and companies. In developing a NAM product, it will be possible to make partnerships with these companies. These partnerships will allow the NAM product to harness the best technologies available. Similarly, the synthesis market is maturing, and should be competitive enough that partnerships can be made in this market without giving too much power to suppliers. This enables the company to focus resources on the core areas of storage and access. Storage is core because it will enable the product to harness the key features of NAM, durability and density. Access is core because it will enable the customer to retrieve, mine and analyze the stored data. These processes are the key to developing a product with a durable competitive advantage, and the core of the company’s technology.
scientific and engineering development efforts and protective measures taken by the company towards storage and access.

Value Creation

NAM’s durability and density create cost savings for data centers. Durability is important for those customers that store their data for long periods of time, meaning more than 10 years. Current technologies have a relatively short lifespan before the hardware needs to be refreshed. For various reasons these technologies lose the capability to store information. For example, with magnetic tape the magnetic charge can be lost with time or the tape layers can degrade with use. Typically, storage technologies have a lifetime of roughly 10 years before they need to be refreshed, which involves purchasing new technology and transferring the data. Magnetic tape, the typical archival medium, has a stated lifetime of 30 years, though in practice they are often refreshed within 10 years. NAM, can reduce this expense as it does not require hardware refreshes or data migrations. Density is important for those customers that store vast amounts of data. For example, Facebook has numerous 60,000 square foot data centers housing 1 exabyte of cold storage each.\footnote{Servers, “Facts and Stats of World’s Largest Data Centers.”} These centers represent a capital expenditure for construction and an operational expenditure for energy (e.g., cooling and power beyond IT equipment energy) that could be nearly entirely avoided with NAM.

This means that when comparing NAM to traditional storage technologies, NAM has a number of cost benefits. Firstly, there is not a need for costly data migrations and hardware refreshes, which occur at least once every ten years. Secondly, there is not a need for physical space. While this is negligible for smaller databases, when considering massive scale this capital expense (or rent) can be quite significant. For example, Facebook pays rent of \(~\$11/\text{month/sqft}, \text{ which is } \sim\$600,000 \text{ per month for an exabyte-sized storage center.}\)\footnote{“Facebook.” <http://www.datacenterknowledge.com/archives/2010/09/16/facebook-50-million-a-year-on-data-centers/>} Thirdly, there is not a need to power the physical space. This is a significant savings, as at small scales the energy costs for space are often as much as the energy costs for the IT equipment (PUE\footnote{Power usage effectiveness (PUE) is a measure of how efficiently a data center consumes power: Total Facility Energy / IT Equipment Energy.} of 2) and at scale despite being more efficient (PUE \sim1.1) the actual expense is quite large.

---

\footnote{Sources, “Facts and Stats of World’s Largest Data Centers.”}
Depending on the PUE and scale, electricity costs for the non-IT equipment in a data center can run into millions of dollars per year.\(^{62}\) Lastly, there will be a lesser need for personnel as there is less hardware to manage; reducing a single person saves \(\sim \$120k\) per year. This equates to multiple millions of savings per customer per year. NAM creates a lot of value for customers, if the technology is cost-competitive. So, the question becomes: how long until NAM becomes cost competitive with other technologies?

In Figure 37, Goldman et al. provide the results of an analytical comparison between on-premises NAM and tape archival, taking into account the durability and associated hardware refresh and data migration savings. The x-axis represents the duration over which data is archived. The y-axis represents the ratio of DNAM’s synthesis costs to tape’s refresh costs through time, so that a value of 1 would indicate they were equivalent. The diagonal lines represent situation for which the two are equivalently costly assuming the stated number of years between hardware refreshes. The area to the right of a given diagonal line represents the situations under which NAM is less expensive than tape archival, and similarly the area to the left of a given line represents tape being less expensive. For example, the green area to the right of the orange 10-year diagonal shows all situations under which NAM is a less expensive archival technology than tape. This analysis indicates that, assuming current synthesis costs, NAM is cost effective for an archival that will be maintained for \(\sim 1,000\) years. Further, this shows that if NAM synthesis were to drop 10-fold then the technology is cost effective for \(\sim 100\) year archival. And, if NAM synthesis became 100-fold less expensive, the technology makes sense for archiving 10 - 100 years.

\(^{62}\) “Facebook Builds Exabyte Data Centers for Cold Storage.”
While these numbers are promising, they are misleading. The analysis makes the assumption that for "archives of a few megabytes, we estimate that the cost in personnel, labour and management of a corresponding tape technology transition might be of the order of $25–100, leading to a current estimate [for the relative cost of DNA-storage writing versus tape transfer fixed cost] in the range 125–500." Unfortunately, this dollar figure is too large. If this were true, then a few PBs of data would require a transfer cost between $25B - $100B (many companies have exabytes of data). If this were true, it would require $25,000 - $100,000 to manage the migration of a few GBs of data, which is something done whenever one replaces a phone or computer. Estimates place the true cost of managing migration at ~$5,000 per TB\textsuperscript{64} (Even this is probably an overestimate for managing PBs, as costs likely do not scale linearly. This would mean a EB would cost $5B, which is absurdly large, so a realistic analysis should not scale linearly.). To simplify, assume transfer costs are $2,500 - $10,000 per TB, $2.5 - $10 per GB, or $0.0025 - $0.010 per MB. This indicates that the number used in the analysis is ~10\textsuperscript{4} too large. Based on the Figure 37 analysis, it follows that DNA synthesis will likely need to drop by 10\textsuperscript{*10\textsuperscript{4}} – 100\textsuperscript{*10\textsuperscript{4}} (100,000 – 1,000,000) before it can economically compete with tape.

\textsuperscript{63} Goldman et al., "Towards Practical, High-Capacity, Low-Maintenance Information Storage in Synthesized DNA."

\textsuperscript{64} "The True 'Cost' of Enterprise Storage - Storage Management."
To verify that a synthesis improvement in the range of 100,000 – 1,000,000 would be needed, a net present value analysis was developed. Since the most likely time a customer would be willing to deploy NAM technology is when contemplating building a warehouse to house cold storage data, the analysis compared a new cold storage facility to a NAM archive. This analysis assumed the facility would ingest 1 PB of data per month with all data stored for 100 years. However, even when accounting for the savings from electricity, warehousing, and manpower generated by NAM, the analysis indicated that synthesis would require at least a 100,000-fold improvement before becoming economically viable. This is not to indicate that the savings from these additional levers are insignificant, but rather that future hardware refresh and migration costs are less significant when discounted and there are other NAM costs to be considered, most significantly sequencing costs.

Unfortunately, this needed improvement is enormous. As previously stated, there is a lot of room for improvement in the underlying NAM technologies, but there is no guarantee that a 100,000-fold synthesis increase will ever be realized. If it is realized, it will take time. Figure 38, shows the timescales required to reach a desired level of improvement assuming various annual rates of improvement. This shows that realizing a 100,000-fold improvement will take more than ten years in even the most optimistic scenario, an improvement rate seen within the sequencing industry, which was spurred on, by the Human Genome project and vast amounts of investments. Another item to consider is that even as NAM improves, so do other substitute technologies, lengthening the time required for NAM to become competitive. The most likely scenario is that NAM will require 15 or more years to become cost competitive, with ten years being very optimistic.
Figure 38: Annual Improvement Rates and Projected Timescales

<table>
<thead>
<tr>
<th>Rate Definition</th>
<th>Goldman et al. Assumption</th>
<th>Moore's Law</th>
<th>Midway Between Moore's Law and Sequencing Trend</th>
<th>Sequencing Trend Since Human Genome Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half the Rate of Moore's Law</td>
<td>2x improvement every 4 years</td>
<td>2x improvement every 2.5 years</td>
<td>2x improvement every 2 years</td>
<td>~6x improvement every 3 years</td>
</tr>
<tr>
<td>Goldman et al. Assumption</td>
<td>32.0%</td>
<td>41.4%</td>
<td>82.0%</td>
<td>10x improvement every ~3 years</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Desired Increase</th>
<th>Years of Improvement Required to Meet Desired Increase, at a Projected Rate of Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x</td>
<td>13  8  7  4  3</td>
</tr>
<tr>
<td>100x</td>
<td>27  17 13 8 6</td>
</tr>
<tr>
<td>1,000x</td>
<td>40  25 20 12 9</td>
</tr>
<tr>
<td>10,000x</td>
<td>53  33 27 15 12</td>
</tr>
<tr>
<td>100,000x</td>
<td>66  42 33 19 14</td>
</tr>
<tr>
<td>1,000,000x</td>
<td>80  50 40 23 17</td>
</tr>
<tr>
<td>10,000,000x</td>
<td>93  58 47 27 20</td>
</tr>
</tbody>
</table>

**Funding**

The above strategy was developed with the assumption that NAM is cost competitive with other archive technologies. Unfortunately, the previous analysis shows the technology to be a decade away from commercialization. Any venture wishing to enter the NAM space could wait to form at a later date when costs have fallen, or the company could form immediately. The benefit of forming immediately is that it provides the company with ample time to develop IP in the core technology and piece together the expertise that will be required for an architecture strategy. However, the large impediment to forming immediately is funding.

For a startup, potential funding sources usually include family and friends, crowdfunding, accelerators, incubators, angels, venture capitalists (VCs), corporate VCs (CVCs). Each of these streams of funding is willing to accept a different level of risk. The risk level of a startup is heavily correlated with how “early” or how far away the venture is from commercialization. Typically, family and friends are the most risk tolerant and VCs the least. Assuming, a lack of funds from family and friends, then crowdfunding, accelerators and incubators are the next most risk tolerant. Unfortunately, each of these sources typically expects work to proceed against a beta product. For NAM, a beta product is likely five or more years from development, so these sources seem very unlikely. Since angels and VCs are even more risk averse, these too are not likely funding sources for a NAM product that is a decade from commercialization. CVCs...
are a slightly different breed and come in two flavors. There are CVCs that are financially motivated and others that are strategically motivated. The financially motivated CVCs, act very similarly to VCs, with the main difference being from where the fund is filled, so are an unlikely funding source. Strategic CVCs aim to invest in startups that are strategically aligned with the mission of the corporation, so that synergies can later be captured. Because of this motivation, strategic CVCs are more risk tolerant if the corporation sees great value to be gained. This means that storage service providers, storage hardware providers, and synthesis or sequencing players might be willing to invest in a NAM startup. However, even these companies would be unlikely to invest in a company that is more than five years from commercialization. If faced with the choice, these corporations would be more likely to invest that money internally towards a related R&D project, because that would at least provide IP, motivation for internal employees, and recruitment. Fortunately, government grants are another more risk tolerant source of funding.

Each year, government grants are awarded to individuals, non-profits, and business. These grants are awarded through many federal agencies for many purposes, including stimulating innovation and small businesses. Two such grants are the Small Business Innovation Research (SBIR) grant and the Small Business Technology Transfer (STTR) grant. The SBIR program is intended to help small businesses conduct research and development. According to the program founder, Roland Tibbetts, it is meant “to provide funding for some of the best early-stage innovation ideas -- ideas that, however promising, are still too high risk for private investors, including venture capital firms.” More than $2.5B is awarded each year to small businesses conducting early stage R&D. Similarly, the STTR is another program that provides funding opportunities for innovative R&D. The STTR is distinct from the SBIR in that it aims to foster R&D between small businesses and research institutions. Numerous agencies participate in these two programs, notably NSF, DoE, DoC NIST, DoD, and DHS. Beyond these two programs many agencies fund grants for specific research requirements.

As grants are meant to be an early-stage funding source, this is a potential route for the NAM company. This would enable to company to develop IP in the core idea while the market

---

65 "Small Business Innovation Research."
matures and costs relax. This would also provide the company with a few years in which to develop a steady source of revenue. While it is unclear what this revenue stream would be there are a number of possibilities. For example, the company could develop IP in and enter the growing synthesis market. Or the company could identify customers for which the features of NAM are valuable in their own right, as opposed to being of value for economic reasons. For example, it might be possible to use NAM as a way to pass highly sensitive messages via human organic transference, utilizing both density and organic compatibility. Whatever its source, this revenue stream should be reinvested in the company to fund further NAM R&D.

To bring NAM to the market will likely require investments in excess of $50 million. So, unless there is a significant revenue source identified, there will need to be outside funding downstream. The most likely and source for this later-stage funding would be a CVC. The company should look first to CVCs that might be potential customers, then to CVCs that might be potential competitors and acquirers.

Risks

At the moment, NAM is a very risky venture. The biggest risk is that the technology is still very early stage. Much progress needs to be made before NAM is close to commercialization, and this will require a long period of time. In that time, a technology could come to market that eclipses the need for NAM. Or, maybe that progress will never be made, as there is no guarantee that costs can be lowered to make the technology cost competitive.

If the technology gains and cost reductions are made, the largest risk is the market. Firstly, the storage industry is very competitive and somewhat commoditized market. Maintaining a profit within this industry would be very challenging for a new entrant. Initially, NAM would be a very differentiated product, compared to standard technology, but that differentiation would fade as other NAM players establish themselves and as the technology matures. Secondly, big industry players (hardware or service providers) may take interest in the technology and dominate the market. For now, this is not much of an issue as most companies (hardware providers at least) are more concerned with technologies that will be viable within a few years, not a decade or more. However, if it were to become a potential issue, the best safeguard
against this is the IP protection the company should have been investing towards. If the IP is strong it will safeguard the company from being chased from the market and likely make the company very attractive for acquisition.

Conclusion

Nucleic acid memory is an intriguing technology with a strong value incentive for large corporations in the $4B data archiving market. The durability and density features of NAM are well suited for long-term archiving and the technology allows for many capital and operational expenses to be avoided. Once the technology is cost competitive with other storage technologies it should be marketed as a hardware solution for massive archives, competing directly with tape and hard disks. However, getting to this point is a very risky venture.

The technology is a decade away from commercialization with progress to be made both economically as well as technically. The costs need to come down at least 100,000-fold before the technology is cost competitive with current solutions. Additionally, there are a number of scientific and engineering issues that need to be carefully resolved. If a company were to undertake this challenge, it would be a difficult path that might never find the market. The first difficulty would be obtaining funding, as most sources would deem the venture too risky at the current stage. The one source that does seem viable is government grants. This would allow the company to grow as a R&D company for multiple years, until a time where it can develop a source of revenue. This revenue should be further devoted to R&D in an effort to develop and safeguard the core technology against future competitors. Specifically, the company should develop IP in the storage and access processes, as these are fundamental to the long-term viability of the company. Additional resources should be given towards synthesis, as this is the main impediment to commercialization. This would allow the company a strong chance to thrive in the competitive storage industry, if and when NAM becomes cost competitive.

Additionally, by fortifying the competitive advantage through a controlled architecture strategy the company will set itself up for a future acquisition by a storage hardware or service provider.
Appendix

Storage Hardware Providers

Dell Technologies: Dell & EMC

In 2016 Dell acquired EMC for $67 billion and restructured the business. Both companies were placed under a parent entity Dell Technologies. EMC was renamed Dell EMC. The original Dell corporation was subdivided, with the PCs portion remaining as a newly formed Dell Inc subsidiary of Dell Technologies, and the server and storage portions being integrated into the newly named Dell EMC. The company is the dominant player in the data storage systems market, selling products and services that enable businesses to store, manage, protect, and analyze data. Products categories include information storage, archiving, backup, recovery, content management, and virtualization.

Hewlett Packard Enterprise (HPE)

In 2015 HP split into two separate companies HP Inc. and HPE. HP Inc. maintained the personal computer and printer businesses while HPE focuses on enterprise products and services. HPE provides hardware, software and cloud solutions for large global enterprises through small business, and government agencies. HPE’s products include servers, storage, software, and networking. HPE’s storage products include HPE 3PAR, HP StorageWorks, and HP XP.

International Business Machines (IBM)

IBM provides products and services, and is a major research organization. Their offerings include cloud computing, IT infrastructure, data and analytics, and security, amongst others. IBM storage products include Storwize All-Flash, FlashSystem, Spectrum storage, DS8000 systems, and magnetic tape.

NetApp

NetApp, founded in 1992, offers systems, services and software for information storage and data management. The company offers various forms of all flash and hybrid storage systems, including their Fabric-Attached Storage.
Storage Service Providers

Amazon

Amazon Web services is the clear leader in the cloud services space with numerous products including EC2 for compute, S3 for storage, and Glacier for archival. In addition, they offer other service categories like management tools, analytics, mobile services, business productivity, and game development. Amazon held 31% market share in IaaS in Q2 2016.66

Microsoft

Microsoft's Windows Azure Storage is the second most widely-used cloud storage service, with the company attaining 11% market share in IaaS in Q2 2016.

IBM

IBM is also the third largest cloud service provider (7% IaaS Q2 2016 market share), through its SmartCloud Enterprise offering. When considering just the private cloud, IBM is the leading service provider in the space as it has focused heavily on this area.

Google

Google entered the market through their Google App offerings, run on their cloud storage platform. Since that time, Google has also released Nearline for archiving, backup and disaster recovery. Google holds 5% market share in IaaS in Q2 2016.

Archiving Market Players

Hewlett Packard Enterprise (HPE)67

The HPE archiving portfolio includes HPE Consolidated Archive (HPE CA) for on-premises archiving, and Digital Safe, for hosted archiving in a private cloud.

---

66 Team, “Amazon Continues To Gain Share In Cloud Infrastructure Services Market.”
67 Archiving companies are a selection of Top Players from the Radicati 2016 Market Quadrant. All text is reproduced from “Information Archiving - Market Quadrant 2016.pdf.”
Veritas Technologies LLC

Veritas Technologies offers information management solutions aimed at the needs of large and complex environments. Veritas’ Information Governance product portfolio includes solutions for: archiving, eDiscover, file analysis and more.

EMC

EMC is a leading provider of enterprise information and virtual infrastructure technologies and solutions. EMC’s product portfolio includes hardware and software for archiving, storage, backup and recovery, disaster recovery, cloud computing, and more.

Smarsh

Smarsh is a provider of cloud-based archiving technology and services aimed at highly regulated industries with strict compliance and eDiscovery requirements, such as financial services (e.g. broker-dealers, investment advisers, banks and lenders), the public sector and healthcare.

Interviews

Aleksandra Kacperczyk: MIT Professor of Entrepreneurship & Strategic Management, Strategy

Anthony Philippakis: Google Ventures, CDO Broad Institute, CVC, Biochemistry

Charles Lambert: IBM Hardware Engineer, Storage Industry

Christopher Costello: Oracle Database Administrator, Storage Technology

Christopher Dwan: Broad Institute Director of IT Architecture & Strategy, Sequencing, Strategy

Dave Sabey: Chairman & President of Sabey Corporation, Cloud Services

Javier Justo: IBM LTO Tape Drive Technical Leader, Archive Technology

Jefferson Clayton: CEO Genesis DNA, Synthesis

Jim Harding: Serial Entrepreneur, Amazon, Microsoft, Technology Strategy

Mark Bathe: MIT Professor of Biological Engineering, Nucleic Acid Memory

Matthew Greenfield: Oracle Engineer, Storage Industry
Mike Rolfes: Dell EMC Global Systems Engineering, Storage Industry

Nathan Thaler: MIT Manager of Cloud Platforms, Storage Technology

Sang-Woo Jun: MIT Full-Stack Database Engineer, Storage Technology

Trish Cotter: MIT Martin Trust Center, Strategy

Tyson Shepherd: MIT Post-doc in Biochemistry, Nucleic Acid Memory

Warren Katz: Entrepreneur, Angel Investor, Government Contracting, Strategy
Bibliography


