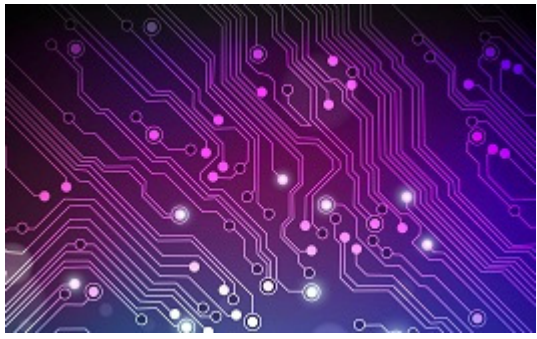
Tractica





WHITE PAPER

Storage Architecture and Implementation Considerations for AI Deployments

Published 1Q 2018

COMMISSIONED BY:

PURESTORAGE

ANAND JOSHI
Principal Analyst

CLINT WHEELOCK
Managing Director

SECTION 1

EXECUTIVE SUMMARY

Artificial intelligence (AI) is shaping up to be the prominent technology of the 21st century. By definition, AI applications are data intensive and storage is a critical aspect of any data center's infrastructure. Much attention is being paid to compute performance in the industry to improve AI application performance; however, storage also deserves a close look and further consideration. Overall system performance is a function of compute, network, and storage capabilities, and a well-planned storage infrastructure can significantly impact overall performance.

This white paper looks at some of the popular use cases in AI enterprise applications and the role that overall storage architecture plays. Tractica has selected five use cases that rely heavily on storage and addresses the challenges pertaining to storage. Factors such as the amount of storage, connectivity, latency, and software infrastructure are considered in the context of the big picture. The goal is to highlight the diversity of applications and industries that are using AI with different storage requirements. This analysis aims to provide a glimpse of how storage is already starting to play a vital role in AI, and the opportunities that this is likely to unleash for both scientific research and businesses.

The covered use cases include:

- Video analytics (retail and surveillance market)
- Medical image analysis (medical)
- Vehicular object detection, identification, and avoidance (automotive)
- Algorithmic trading improvement (finance)
- Content distribution on social media (advertising)

Further analysis covers the technology aspect of storage and how it may impact overall application performance. During the technology analysis, hardware and software have been kept separate from data issues. These are two separate challenges that need individual attention.

Overall system architecture of storage, network, and compute performance, along with their interfaces, plays a key role in any application performance. Storage is more relevant than ever for AI applications due to the large data requirements. Newer storage architecture solutions driven by Flash arrays are a step in right direction to overcoming these challenges.

SECTION 2

INTRODUCTION

AI has been featured prominently in the news headlines for the past few years. AI has enabled major advances in the areas of image, speech, text recognition, and data analysis. This has fueled a wide range of applications in the market that promise to alter the way humans conduct day-to-day dealings.

Although AI has been around in one form or another since the 1960s, many of today's applications are based on deep neural networks (DNNs). The quality of neural network (NN) results is a function of the mathematical model and the data used to build the model. There are two phases to the overall application development: training and inference. During training, the data is used to build the best NN model. During inference, the built model is run against a new set of data and the results are predicted.

The need for compute, storage, and network performance has increased at a rapid pace in the last few years. Figure 2.1 below shows the requirements for science and research and development (R&D) as provided by the Oak Ridge National Laboratory (ORNL). This data was prepared in 2012 when AI was just starting to emerge. If it were prepared today, it would possibly show even higher numbers.

The ORNL projects the performance requirements of 100 petaflops (PFs) to 200 PF compute capacity and 5 petabytes (PBs) to 10 PBs of memory for 2017. In comparison, today's state-of-the-art AI server, NVIDIA's DGX1-V, has compute capacity of 960 Tensor TFLOPS (4X4 matric multiplication of 16-bit arithmetic). The DGX1-V provides 8 terabytes (TBs) of on-board storage, which is typically a small set of overall data size required for an AI application.

Figure 2.1 Roadmap Describing the Need for Computational Performance

Table 1. Computational science platform requirements for the OLCF

	2012	2017	2020	2024
Peak flops	10–20 PF	100–200 PF	500–2000 PF	2000–4000 PF
Memory	0.5–1 PB	5–10 PB	32–64 PB	50–100 PB
Burst storage bandwidth	NA	5 TB/s	32 TB/s	50 TB/s
Burst capacity (cache)	NA	500 TB	3 PB	5 PB
Mid-tier capacity (disk)	20 PB	100 PB	1 EB	5 EB
Bottom-tier capacity (tape)	100 PB	1 EB	10 EB	50 EB
I/O servers	400	500	600	700

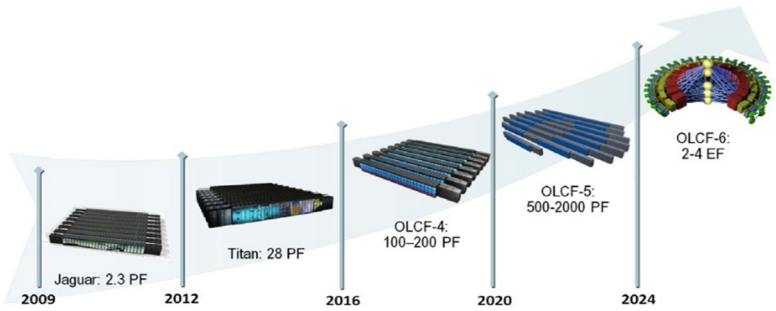


Figure 1. OLCF 2024 roadmap.

(Source: Oak Ridge National Laboratory)

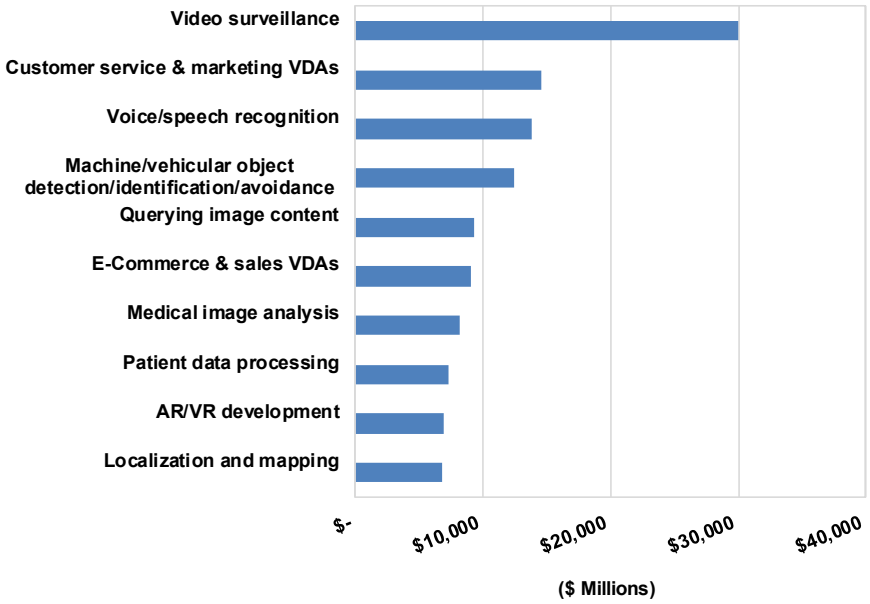
For Big Data applications, the size of the data set for training purposes could run into hundreds of TBs or even PBs. AI algorithms running in conjunction with Big Data applications with such a large data set pose a unique set of challenges for overall system performance. On the software side, challenges with the data management, manipulation, and fetching must be overcome. On the hardware side, challenges with storage architecture, speed, and latency must be overcome. Finally, the compute and network bandwidth must also be tuned so that overall results are generated within the allocated time budget.

This white paper examines the storage aspect of several significant and computationally-intensive AI use cases and highlights their relevant applications. Storage is as important to AI performance as compute capabilities, although most of the headlines usually center around compute. Tractica has conducted extensive research into some of the highest profile use cases and analyzed their storage requirements. Also provided is a description of some of the typical storage-related challenges posed when running Big Data applications.

Tractica has performed comprehensive analysis of more than 200 discrete use cases for AI. Chart 2.1 below shows the highest potential use cases in terms of cumulative revenue. Video surveillance, which consists of using AI techniques for object or facial recognition, is the most popular use case today. Six out of the top 10 use cases focus on imaging and video, one on audio, and the rest on data. Eight out of 10 deal with Big Data and require large amounts of storage space.

While the intersection of AI and storage is an evolving research area, the two are obviously highly correlated. The size of data on the server, the storage network, the connectivity, and the database all play key parts in the overall AI application performance. In addition, the size of the AI workload, the type of data used, and the database used are also considerations in successfully implementing AI applications. By using the proper storage system architecture, the applications and overall use cases stand to benefit immensely.

Chart 2.1 Cumulative AI Revenue, Top 10 Use Cases, World Markets: 2016-2025



(Source: Tractica)

SECTION 3

STORAGE-INTENSIVE AI USE CASES

AI technologies are being used in a wide range of applications. The use cases for AI today can be broadly categorized by considering the type of data with which they deal. The most common use of AI centers on audio, video, text, image, and stored data. In this section, Tractica has selected five use cases that rely heavily on storage and considered factors such as the amount of storage required, connectivity to the system, and overall application performance.

The covered use cases include: video analytics (retail and surveillance market), medical image analysis (medical), vehicular object detection, identification, and avoidance (automotive), algorithmic trading improvement (finance), and content distribution on social media (advertising). The goal is to highlight the diversity of applications and industries that are using AI with different storage requirements. The analysis aims to provide a glimpse of how storage is already starting to play a vital role in AI, and the opportunities that this is likely to unleash for both scientific research and businesses.

3.1 VIDEO ANALYTICS FOR RETAIL AND SECURITY

Video analytics systems extract information from video content that is meaningful as perceived by the human eye. These systems primarily use computer vision technology to extract information about the incoming video. A video analytics system that automatically monitors cameras and generates alerts about events of interest is more effective than relying on a human operator. Human operators can be cost effective at scale, possess limited alertness and attention, and can only monitor a limited number of video feeds. Moreover, the operator's ability to monitor the video and effectively respond to events is significantly compromised over time.

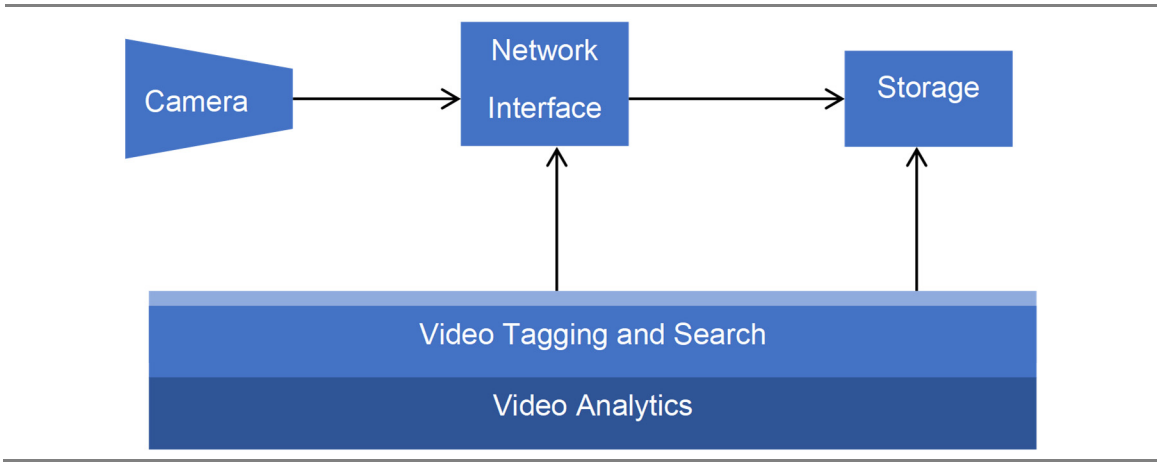
Video analytics, when combined with human monitoring, creates a very powerful value proposition for businesses seeking security and surveillance. Not only do computers exceed humans in terms of attention span, but improved technology means that the results are as accurate (or better) as humans and can be generated in real time.

The extracted data can be stored in a database and analyzed for patterns over a period of time. This provides businesses with unprecedented insight into their businesses not available via other means. This has given rise to a new field of video-driven business intelligence, enabling users to extract statistical data of interest. Retail businesses have capitalized on this technology and it is rapidly evolving into becoming a big market called retail analytics.

Over the past several years, video analytics has evolved into an ideal solution that meets the needs of surveillance system operators, security officers, and corporate managers. Maturing computer vision algorithms, coupled with increasing computational capacity and increased camera resolution, helped increase the accuracy of results.

Figure 3.1 below shows a typical video analytics system. Today, most analytics system deployments typically feature hundreds of cameras. In some instances, such as a smart city, a video analytics system may consist of thousands of cameras. Such large-scale deployments generate significant amounts of data that must be archived, managed, and made available to the systems that scour through the data to generate analytics.

Figure 3.1 Typical Video Analytics System



(Source: Tractica)

A video analytics system presents many challenges pertaining to ideal storage system design. The video could be streaming 24/7, so enough storage must be made available to store such long streams of content reliably at any given point in time. Camera resolution can vary in such systems and the hardware/software system must be able to handle the bandwidth required for the video stream. A key challenge of storage systems involves optimizing the write times because data is being written the vast majority of the time.

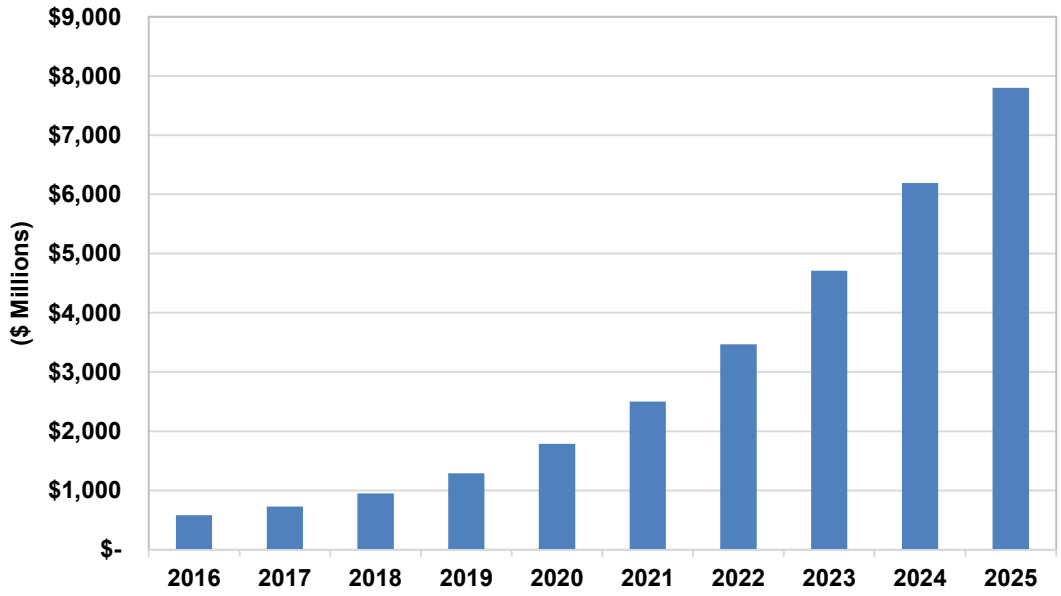
Sometimes, the user may demand real-time results. In such situations, storage system latency becomes important. The system should not only be capable of writing the data within a given latency, but it should also make it available to the compute engine. If the storage system cannot keep up, then frames get dropped, making video less searchable and results inaccurate.

Depending on the retention policy, video streams (which increasingly are reaching higher resolutions), can consume storage rapidly. This means that system administrators must be able to scale (add new storage) their storage infrastructure quickly and non-disruptively.

Video, in essence, represents unstructured data and must be managed so that it can be searched like text for future references. Ensuring that video content can be found, retrieved, and viewed from multiple locations is quickly becoming a key requirement of video analytics solutions. In addition, the level of security (who can watch/access the video streams and when) is an important requirement to ensure the integrity and privacy of the footage.

Video surveillance is the top use case for AI as governments start to deploy AI-based video surveillance systems that can perform real-time facial recognition, pose estimation, and object recognition. In China, almost all municipal police departments, of which there are close to 300, have invested in AI-based video surveillance technology. According to some estimates, as many as 20 million AI-enabled closed-circuit television (CCTV) cameras have been installed, so far, across China.

Chart 3.1 AI-Based Video Analytics Revenue, World Markets: 2016-2025



(Source: Tractica)

3.2 MEDICAL IMAGE ANALYSIS

The medical industry has used a wide range of imaging technologies to investigate the extent of tissue damage or disease. Medical imaging techniques, such as magnetic resonance imaging (MRI), X-ray, and computed tomography (CT) scans, have become commonplace in hospitals today. Analyzing medical images has primarily been a manual process to date. The images are typically taken by a technician in a laboratory and passed on to a doctor for analysis. The doctor manually looks at the image via a photograph or a computer and provides a diagnosis to the patient. This methodology is not only prone to human error, but is also time-consuming and costly.

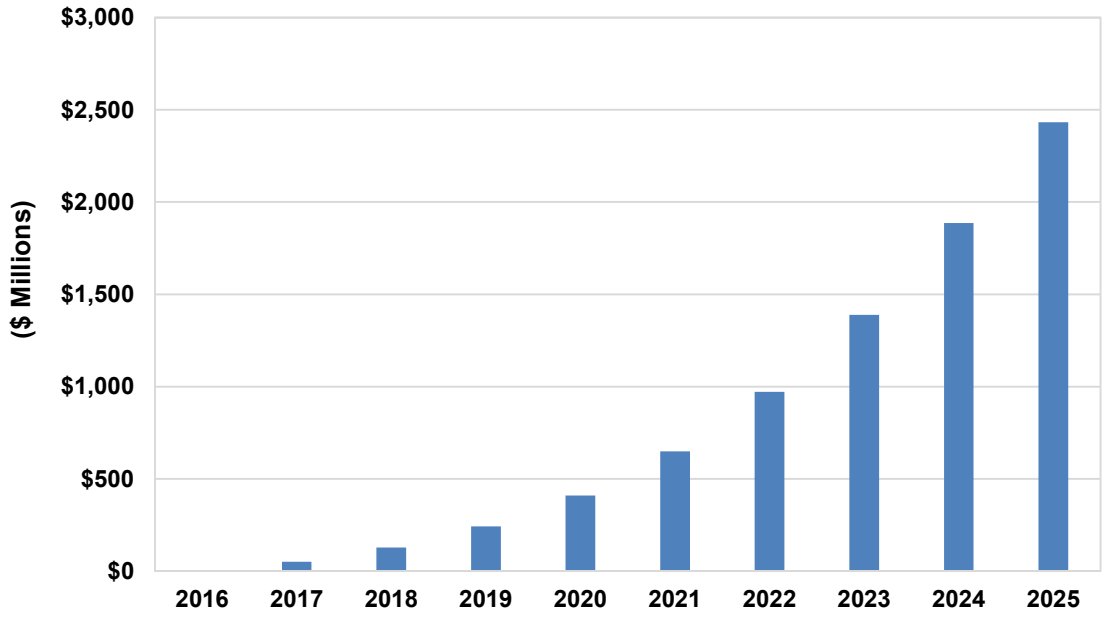
AI has been adapted to analyze such images and is promising to change the way things are done in medical diagnostics. Deep learning (DL) technologies are now being applied to automate the analysis and increase accuracy, precision, and understanding of images down to the pixel. These techniques analyze the incoming image via a trained AI application, make predictions regarding diseases, and provide probability of occurrence. The images can also be enhanced using AI techniques, making it easier for doctors.

AI-based medical image analysis has been used in diagnosing eye diseases, tumor detection, cancer detection, and other applications. The AI techniques have been particularly useful in diagnosing critical conditions, including cancer, neurodegeneration, and heart disease. Faster and smarter results of AI systems are appreciated by both doctors and patients. Tractica sees the market for medical imaging systems growing rapidly and projects that it will increase from $0.07 million worldwide in 2016 to $2.4 billion in 2025.

Researchers in medical imaging face a multitude of challenges regarding storage. The issues revolve around storing, translating, and analyzing the data. Often, images need to be stored in a raw format that increases the file size. This could slow the system down when multitudes of such files are necessary for training. In addition, regulatory requirements may

require anonymizing or encrypting the image. This puts an additional encryption workload on the system. AI-based medical imaging training systems need high compute capacity, integration with databases of different systems, and high memory bandwidth. Genome databases, for instance, could run into hundreds of TBs or higher.

Chart 3.2 AI-Based Medical Image Analysis Revenue, World Markets: 2016-2025



(Source: Tractica)

3.3 AUTOMOTIVE TRAINING AND INFERENCE

Self-driving or autonomous cars have been the media darlings of late. Advanced driver assistance systems (ADAS) became commonplace as of 2017, and almost every major automotive manufacturer has started offering some sort of camera-based system. Although fully autonomous driving remains far out in the future, Level 3 or Level 4 automation is certainly on the horizon.

Object detection and classification, a classic computer vision technique that deals with recognizing images, is the most valuable technology to enable such systems. In today's cars, AI systems read sensor data from cameras and other sensors, and then use pre-trained models to decide a course of action. Such systems use complex NNs to read the incoming video and sensor feed, and generate inference data.

The training part of such AI systems presents a challenge for compute, as well as storage. Training a system that is accurate 99.99999% of the time requires a massive amount of video data. Multiple video feeds may also be required from different angles and locations to compensate for a moving environment. Weather factors (rain, bright sunlight, low lighting, glare, dirt, snow, or any other number of obstructions) can alter the appearance of an object and the training algorithm must account for such issues. In addition, there could be supplemental data in the form of radar, infrared images, or light detection and ranging (LIDAR) to improve the accuracy of the model. The compute engine needs to be able to handle massive amounts of data from different sensors.

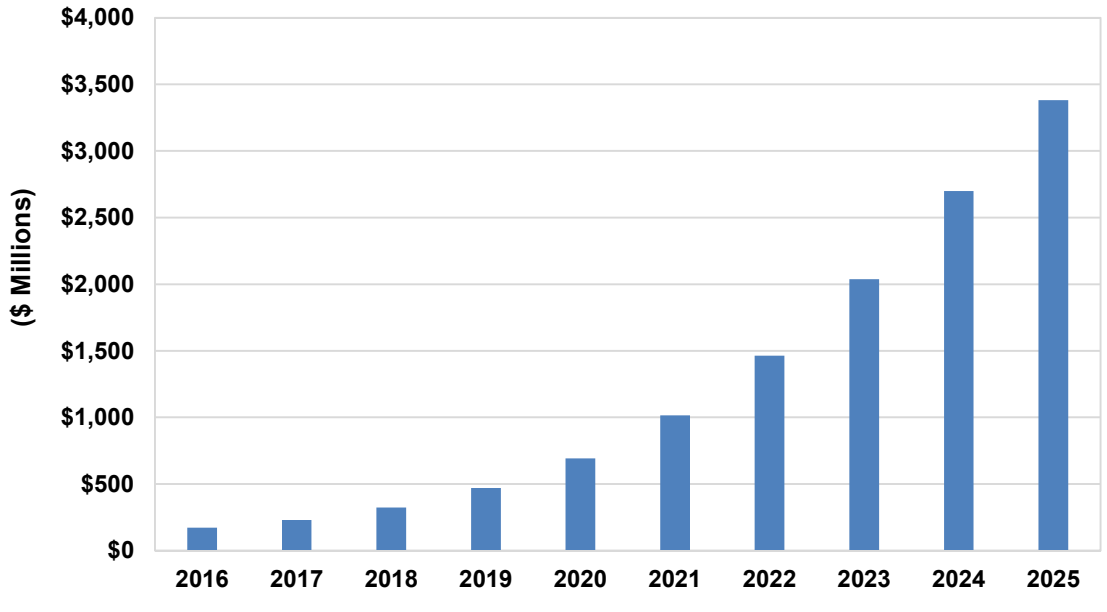
Today, ADAS training is done on a server in a datacenter. The data is collected from an actual car driving and the NN model is trained for days or even months. This model is downloaded to the automobile periodically. Storage is a critical part of the infrastructure for both training and inference. On the training side, the sheer volume of data in an automobile could overwhelm the storage systems. Automobiles have recently started using 360° cameras, which could mean that up to 17 video streams would need to be processed, stored, and then fed into the training system. The volume further increases when considering the data from additional sensors, such as radar.

On the actual automobiles, storage requirements are constrained by the real-time decision-making requirements. Automatic emergency braking (AEB), for instance, must calculate the results within a permissible time, depending on the speed of vehicle, to avoid a collision. The storage system onboard an automobile must cope with multitudes of cameras and sensor data processing in real time.

Automobile companies are looking to make training real time and update the automobile regarding any road issues on the fly. For instance, a pothole could be identified quickly by a training system and passed on to automobiles following behind to avoid accidents. Such systems would require tremendous improvements in the hardware infrastructure for both training and inference beyond where things stand today.

Tractica forecasts that the annual revenue for machine/vehicle object detection/identification/avoidance in automobiles will increase from $172.4 million worldwide in 2016 to $3.4 billion in 2025.

Chart 3.3 AI-Based Automotive Training and Inference, World Markets: 2016-2025



(Source: Tractica)

3.4 Algorithmic Trading Improvement

The financial services industry was one of the early adopters of AI technology. AI is being used in fraud detection, risk analysis, high-frequency trading, derivatives pricing, and providing front-office real-time trading analytics.

Algorithmic trading, sometimes called "algo-trading," has been part of automating investment for years. In this type of trading, a rough schedule for trading is created by an algorithm that provides quantity, price, and a buy or sell signal for a share. When changes in the market occur, the algorithm checks if the situation is applicable and triggers an execution as applicable. The most common application of algo-trading is to enhance trading strategies, including arbitrage, intermarket spreading, market making, and speculation.

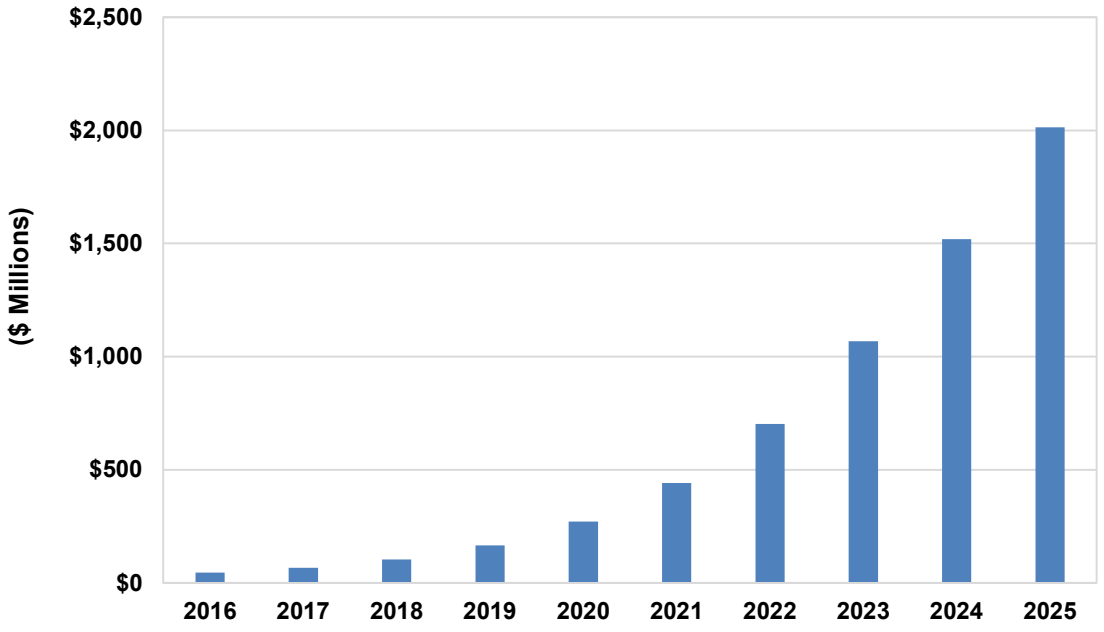
Algorithmic trading applications use a DL network to uncover complex patterns, trends, and relationships. In a high-speed trading environment, the algorithm's speed easily surpasses that of a human's. When a pattern or relationship is successfully identified, it triggers appropriate action deals. Goldman Sachs, Bridgewater Associates, Cerebellum Capital, Euclidean, Man (AHL) Group, and a number of other established investment hedge fund firms are actively using and investigating how and where they can apply AI.

Data sets in financial services can reach multiple PBs, often including real-time statistics, such as economic indicators and news feeds, with execution times generally less than a millisecond. Data input/output (I/O) is a key bottleneck, as the loading time for data sets as large as 50 TBs could take a few days. Successful AI applications in finance require a storage system that can provide high-bandwidth, shared interconnects, and fast workload options.

Given the high stakes, experts point to a number of remaining challenges for DL and AI-enabled algo-trading. These revolve around the limitations of models to fully regard (or disregard) noise, random vectors, and high uncertainty prevalent in financial markets. Furthermore, the very commoditization of such algorithms would erode their competitive predictability, until the algorithms themselves advance in evolutionary computation. Algorithmic trading is in the early stages and the hardware infrastructure will need significant improvements.

Tractica forecasts that the annual revenue for algorithmic trading strategy performance improvement in investment markets will increase from $45.7 million worldwide in 2016 to $2.0 billion in 2025.

Chart 3.4 AI-Based Algorithmic Trading Strategy Performance Improvement, World Markets: 2016-2025



(Source: Tractica)

3.5 Content Distribution on Social Media

Social media has become a key marketing strategy for enterprises and businesses. A vast array of tools has emerged in order to help companies effectively identify, monitor, engage, and learn from user-generated content related to their products. AI is being used widely to "read" the user's mind and generate content that best matches their interest. DL and machine learning techniques are being widely used as tools for mining big unstructured data sets (social media posts, comments, reddit threads, online communities, etc.). This data is often correlated with data from other sources (such as weather, housing price) using clustering algorithms to generate meaningful results.

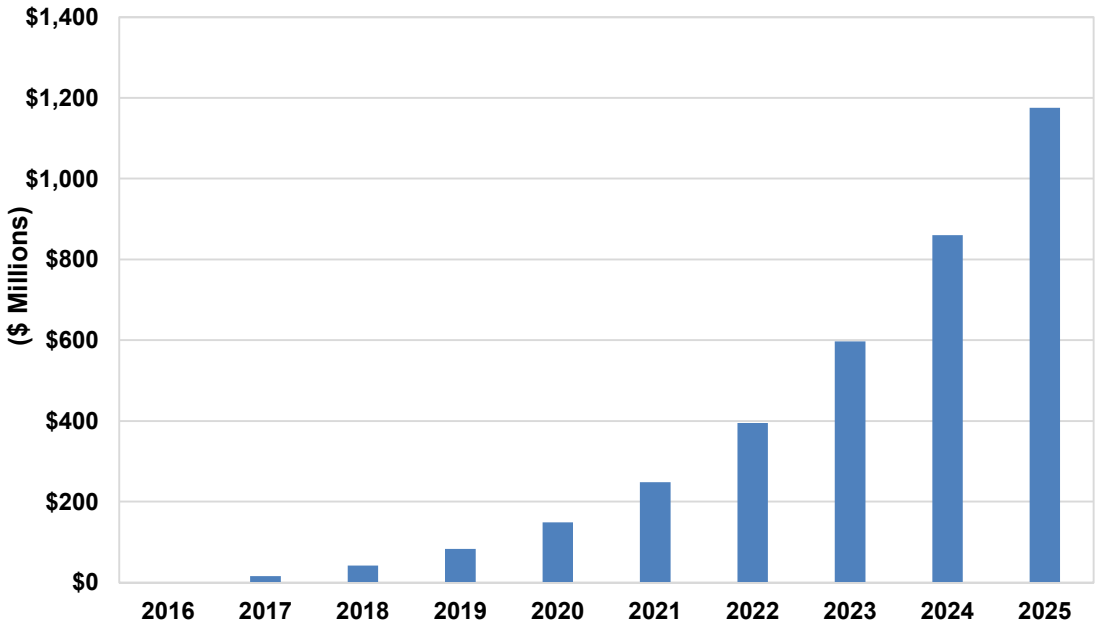
Enterprises and businesses are using AI to support social media publishing and management in several ways, as described below:

- Detect user sentiment by analyzing unstructured data, and his/her social network
- Determine what colors, images, text, hashtags, and other elements resonate most with specific audiences
- Recommend optimal spending for each post or advertisement
- Detect disgruntled customers through sentiment analysis
- Offer alerts, information, product updates, campaign reminders, loyalty incentives, etc. to customers depending on their purchase power
- Recommend specific products and deals
- Analyze competitive content performance to come up with a strategy

The variety of data used to create such solutions could be tremendous. The data sources could be content (colors, keywords, emojis, subject, hashtags), performance related (likes, comments, shares, click-throughs), deployment related (frequency, day, time), or other data types. The data can come from different sources and would need to be converted into a format that best suits the AI system requirements. Databases and systems that can provide such data to a server as and when needed are important in the context of storage in these applications.

Tractica forecasts that the annual revenue for social media content distribution will increase from $0.61 million worldwide in 2016 to nearly $1.2 billion by 2025.

Chart 3.5 AI-Based Content Distribution in Social Media, World Markets: 2016-2025



(Source: Tractica)

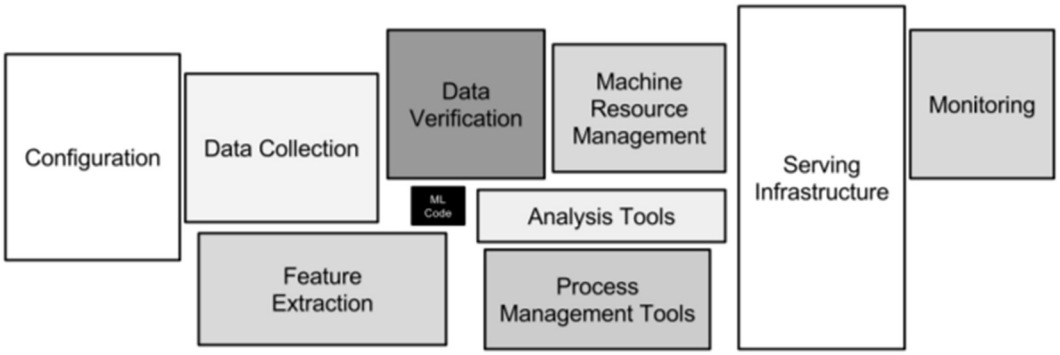
SECTION 4

STORAGE ARCHITECTURE AND IMPLEMENTATION CONSIDERATIONS

AI application is often described in terms of required compute capacity, rather than overall system requirements. When considering a Big Data AI application, overall system performance is important, rather than only the compute infrastructure. A typical data center can be described in terms of compute, network, and storage elements. In addition, the communication protocol that moves data within the elements needs to be considered.

A Big Data system using AI can be described as a data processing pipeline that may consist of data ingestion, conversion, processing, and result generation phases. The data can come from multiple sources, be stored in multiple databases, have different characteristics, etc. Figure 4.1 below, adapted from a Google paper presented at the 2017 Conference on Neural Information Processing Systems (NIPS), states that an AI algorithm is a small part of the overall application. The paper lists machine learning-specific risk factors that must be considered in system design and data-related actions (data collection, verification, feature extraction, analysis, and monitoring). These data considerations ultimately boil down to the design of storage infrastructure, configuration, and management.

Figure 4.1 Hidden Technical Debt in a Machine Learning System



(Source: https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf)

In this section, Tractica analyzes storage from two different viewpoints. The first one deals with the hardware and software aspect, whereas the second one deals with the overall data processing pipeline. Any storage option must consider both viewpoints to ensure that the AI application performance requirements are met.

4.1.1 Hardware and Software Issues

A storage system can be characterized according to hardware and software aspects. The hardware consists of factors such as:

- Capacity
- Latency
- Read-write times
- I/O capacity
- Concurrency
- Connectivity

The software aspects may consist of factors such as:

- File size
- File access pattern (random versus sequential)
- Data ingestion pattern (streaming, real time, low frequency)
- Databases
- Data format (structured versus unstructured)

The two aspects are all interrelated and play different roles at different stages of the AI application processing pipeline. Training and inference systems also require different considerations.

During the training phase, in order to achieve the best performance, it is desired that every step of the pipeline be active and loaded 100%. Any stall or wait time lessens the overall performance. Thus, for training purposes, storage speed considerations for fetching data from the processing element and ensuring that it is available when needed becomes important. The data, in this case, may travel to a system random access memory (RAM) bus via a graphics double data rate (GDDR) or high-bandwidth memory (HBM) interface. It may have to be fetched from a storage element that is connected via Ethernet, Fibre Channel, or another interface. The latency, bus connectivity, and read times are some of the aspects that require consideration.

On the software side, considerations may revolve around the database and how it is stored in the system. For instance, during the training phase for object detection, a database with a large number of image files may be required. A research paper published by the ORNL suggests that when training a model on the ImageNet data set, the image file option was more than 17X slower than the key-value storage option. (Reference: An analysis of image storage systems for scalable training of deep NNs http://www.bafst.com/events/asplos16/bpoe7/wp-content/uploads/analysis-image-storage.pdf.) Thus, the criterion on the software side may involve availability of a database, the type of data, the file access pattern, etc.

A detailed consideration of different parameters for a given application is outside the scope of this white paper. The intersection of storage and AI application performance is an active area of research within academia, with new publications being added daily. Given that the size of datasets will only increase in the future, it is imperative that one consider the hardware and software aspects of storage systems carefully when designing AI infrastructure.

4.1.2 Data-Related Issues

Big Data is often characterized by the four Vs: volume, variety, veracity, and velocity. These considerations apply to AI applications as well and must be considered in designing storage systems.

AI data sets often consist of a large number of examples (inputs), large varieties of class types of outputs), and very high dimensionality (attributes). In addition, a large number of such data sets are necessary to train systems to achieve the desired quality of results. Tractica believes that efficient storage and retrieval of data for AI applications is and will become a growing problem. The problem has two parts. The first is enabling efficient storage and retrieval for a wide range of data formats, such as text, image, video, and audio, as the data is collected and made available to DL algorithms, whether it is a training or an inference system. The second part is enabling the passage of these systems to/from a large number of applications in various domains, such as social networks, shopping, and marketing systems.

In Big Data applications, data can come in a large variety of formats and from a variety of sources. For instance, multimedia data may consist of different file formats from the web, mobile devices, cameras, and other connected appliances. Each may require different bandwidths, resolutions, formats, etc. This requires careful design of both the hardware and the database, and the connectors that convert the data into a form that is acceptable to DL algorithms.

Data ingestion is the process of obtaining and importing data for immediate use or storage in a database. An AI application may use streaming data or ingest it in batches. When numerous Big Data sources exist in diverse formats (as is the case with may AI applications), it can be challenging to ingest data at a reasonable speed and to process it efficiently to achieve the desired performance. The storage hardware performance and the software layer both play key parts in dealing with data variety.

Data cleansing is another important aspect of the overall data pipeline for AI applications. Data cleansing may consist of detecting and correcting (or removing) corrupt or inaccurate records from a data set. Data cleansing may be performed interactively via tools, or as batch processing through scripting. While cleansing the data for inferencing applications, the latency of such systems may add to the overall processing time, thus leading to lower performance.

Data velocity is an important consideration when designing an application that conforms to certain latency. The data can be ingested at varying speeds for different data sources, which could be audio, video, time series data, or simply stored data from a database.

As the volume of data increases, system scaling becomes important. A single server with a central processing unit (CPU) or graphics processing unit (GPU) and storage will not work. Instead, a framework consisting of parallelized machines offering distributed compute, storage, and network will be necessary. Traditional Big Data infrastructure has relied on systems like Hadoop to tackle this problem; however, it is unclear how this would scale up in reference to AI applications at the moment. The distributed computing is of primary importance during the training phase when current systems may take months to generate results. The issues associated with computation, storage, and communication management would need to be addressed to enable scaling up to very large data sets.

Tractica's research suggests that data-related problems dominate the list of application developers' "problem lists" today and could consume a large amount of development time in

comparison with the actual AI algorithm development. Companies have deployed multiple data warehousing systems for their business applications over the years and AI systems often require feeds from such systems in a format that is suitable for AI algorithms. In addition, AI applications may need new sets of data to be combined with the existing data from data warehouses, which leads to new set of challenges.

4.2 CONCLUSIONS AND RECOMMENDATIONS

AI techniques like DNNs are starting to hit performance bottlenecks as data workloads grow exponentially. The intersection of AI and storage is the next step in the evolution of AI hardware infrastructure. While compute performance generally dominates the headlines for pushing the performance, storage is equally as important a consideration to ensure the highest system performance, and it needs to be given the same amount of consideration as compute performance when designing hardware infrastructure.

Flash memory-based storage systems have evolved over the years and are rapidly replacing spinning hard disk-based storage systems. Newer systems driven by Flash arrays in which Flash memory is used to mimic disk arrays are perhaps best suited to run AI applications in terms of performance and flexibility. Industry is lagging behind in terms of providing overall system solutions to optimize storage infrastructure and much remains to be desired. A solid software framework that uses a Flash-based memory storage system to solve problems related to the database, access pattern, buffering, etc. would be a step in the right direction.

SECTION 5

ACRONYM AND ABBREVIATION LIST

Advanced Driver Assistance Systems	ADAS
Automated Emergency Braking	AEB
Central Processing Unit	CPU
Closed-Circuit Television	CCTV
Compound Annual Growth Rate	CAGR
Computed Tomography	CT
Deep Learning	DL
Deep Neural Network	DNN
Graphics Double Data Rate	GDDR
Graphics Processing Unit	GPU
High-Bandwidth Memory	HBM
Input/Output	I/O
Light Detection and Ranging	LIDAR
Magnetic Resonance Imaging	MRI
Neural Information Processing Systems (Conference on)	NIPS
Neural Network	NN
Oak Ridge National Laboratory	ORNL
Petabyte	PB
Petaflop	PF
Random Access Memory	RAM
Research and Development	R&D
Terabyte	TB

SECTION 6

TABLE OF CONTENTS

SECTION 7

TABLE OF CHARTS AND FIGURES

SECTION 8

SCOPE OF STUDY

This white paper examines the storage aspect of several significant and computationally-intensive AI use cases and highlights their relevant applications. Storage is as important to AI performance as compute, although most of the headlines usually center around compute. Tractica has conducted extensive research into some of the highest profile use cases and analyzed their storage requirements. Also provided is a description of some of the typical storage-related challenges posed when running Big Data applications.

SOURCES AND METHODOLOGY

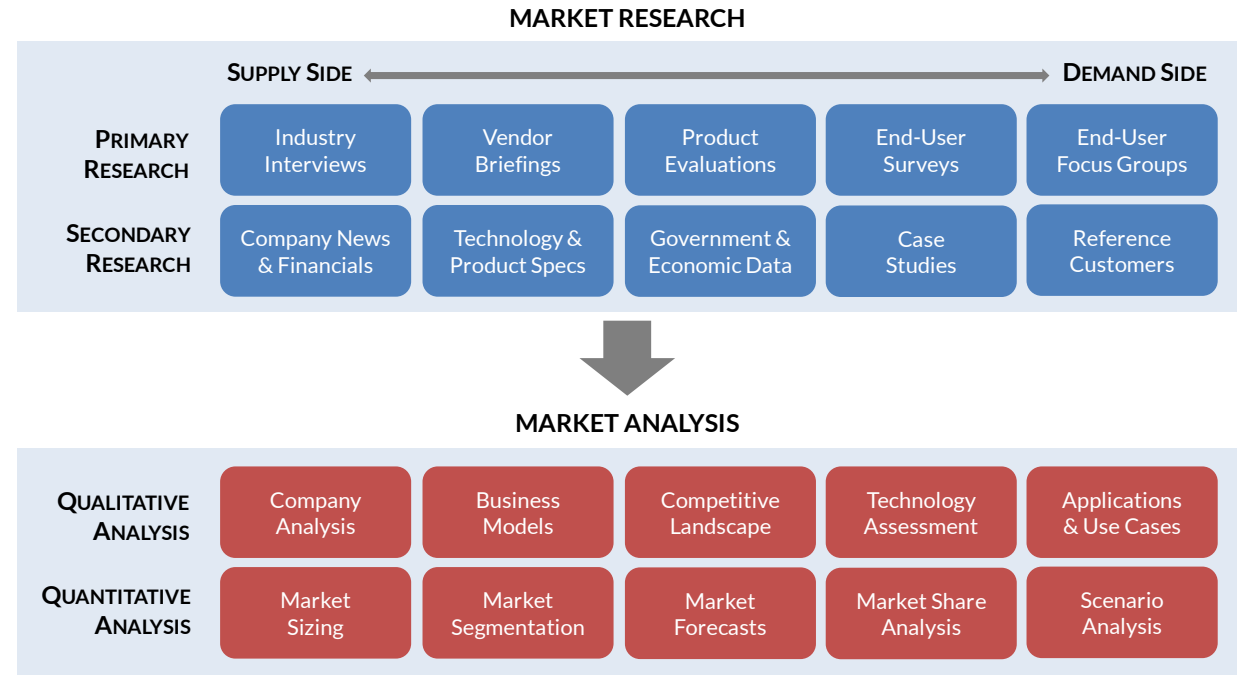
Tractica is an independent market research firm that provides industry participants and stakeholders with an objective, unbiased view of market dynamics and business opportunities within its coverage areas. The firm's industry analysts are dedicated to presenting clear and actionable analysis to support business planning initiatives and go-to-market strategies, utilizing rigorous market research methodologies and without regard for technology hype or special interests including Tractica's own client relationships. Within its market analysis, Tractica strives to offer conclusions and recommendations that reflect the most likely path of industry development, even when those views may be contrarian.

The basis of Tractica's analysis is primary research collected from a variety of sources including industry interviews, vendor briefings, product demonstrations, and quantitative and qualitative market research focused on consumer and business end-users. Industry analysts conduct interviews with representative groups of executives, technology practitioners, sales and marketing professionals, industry association personnel, government representatives, investors, consultants, and other industry stakeholders. Analysts are diligent in pursuing interviews with representatives from every part of the value chain in an effort to gain a comprehensive view of current market activity and future plans. Within the firm's surveys and focus groups, respondent samples are carefully selected to ensure that they provide the most accurate possible view of demand dynamics within consumer and business markets, utilizing balanced and representative samples where appropriate and careful screening and qualification criteria in cases where the research topic requires a more targeted group of respondents.

Tractica's primary research is supplemented by the review and analysis of all secondary information available on the topic being studied, including company news and financial information, technology specifications, product attributes, government and economic data, industry reports and databases from third-party sources, case studies, and reference customers. As applicable, all secondary research sources are appropriately cited within the firm's publications.

All of Tractica's research reports and other publications are carefully reviewed and scrutinized by the firm's senior management team in an effort to ensure that research methodology is sound, all information provided is accurate, analyst assumptions are carefully documented, and conclusions are well-supported by facts. Tractica is highly responsive to feedback from industry participants and, in the event errors in the firm's research are identified and verified, such errors are corrected promptly.

Chart 8.1 Tractica Research Methodology



(Source: Tractica)

NOTES

CAGR refers to compound annual growth rate, using the formula:

CAGR = (End Year Value ÷ Start Year Value)$^{(1/steps)}$ – 1.

CAGRs presented in the tables are for the entire timeframe in the title. Where data for fewer years are given, the CAGR is for the range presented. Where relevant, CAGRs for shorter timeframes may be given as well.

Figures are based on the best estimates available at the time of calculation. Annual revenues, shipments, and sales are based on end-of-year figures unless otherwise noted. All values are expressed in year 2018 U.S. dollars unless otherwise noted. Percentages may not add up to 100 due to rounding.

SECTION 9

ABOUT PURE STORAGE

Pure Storage (NYSE:PSTG) helps customers build a better world with data. The Pure Storage Data Platform, powered by all-Flash storage, offers a simpler, more effective, and more flexible solution for cloud infrastructure and data-rich applications like AI, machine learning and Big Data analytics. With Satmetrix-certified NPS performance in the top 1% of B2B companies, Pure has an ever-expanding range of customers who are some of the happiest in the world.

Published 1Q 2018

© 2018 Tractica LLC
1650 38th Street, Suite 101E
Boulder, CO 80301 USA
Tel: +1.303.248.3000
Email: info@tractica.com
www.tractica.com