AI & INTEL SOLID-STATE MEDIA TECHNOLOGIES

EBOOK CATEGORY: TECHNOLOGY PERSPECTIVE

Release 1.0
MARCH 2021

AI & Intel Solid State Media - First Edition

Published by Brookend Limited.

Document reference number BRKSW0126-01.

CONTENTS

Introduction 1
  Understanding AI Workflows 1
The Modern AI Data Pipeline 2
  Data Sources 2
  Data Lifecycle 2
  The Three V's 3
  Challenges 4
  Storage in AI Workflows 4
Technology Choices 6
  NAND Flash 6
  Intel Optane 7
Software Solutions & Trends 8
Architecture Options 9
  Golden Repository 9
  Data Warehouse 9
  Data Lakes 10
  Vendor Solutions 10
Further Reading and Content 11
More Information 12
  The Author 12

Introduction

Since the development of the first computers in the 1940s and the move to commercial computing in the mid-1950s, data has been at the heart of computing. Although the term may seem antiquated today, IT infrastructure was initially built to do data processing. Early implementations operated mostly in batch mode, crunching payroll or sales information in overnight analysis ready for the next business day.

Today, data processing solutions are much more advanced, with a wealth of data sources that are both human and machine-generated. As IT moves out of the traditional data centre, we're creating data at the edge and with IoT devices. Businesses have progressed to be global 24-hour operations with little or no opportunity for downtime.

Business advantage derives from using data sources in the most efficient way possible. IT organisations need to create solutions that offer real-time processing and analysis of data from corporate websites, bricks and mortar stores, manufacturing operations and research operations, to name just a few. This evolution in IT requires a revolution in technology. AI (Artificial Intelligence) and Machine Learning (ML) are techniques that allow the exploitation of vast quantities of data to create new business insights. The data processing needs of ML/AI are significant and only produce linear improvements for exponential increases of data input.

Analytics has moved from a batch orientated process to one that runs continuously, processing new forms of data that includes real-time streaming content. Graphics processors have evolved into GPUs with thousands of active cores that consume data at high throughput and low latency.

These requirements demand new persistent storage solutions with high endurance, high capacity and low latency at a price point that businesses can afford.

Understanding AI Workflows

Modern analytics has evolved into a real-time model that continually processes and analyses data. Historically, AI solutions operated as batch processes, due to the volumes of data involved and the capabilities of computing resources at the time. Today, business leaders expect analytics to run in real-time, taking feeds from multiple sources and enabling businesses to produce near-instant insights on their data.

At the heart of operations is the data pipeline. Pipelines are automated processes that take input data to develop analytics models and create insights from new data. Vast volumes of data are used to train models in a circular process that continues to iterate over time. The models built by businesses then process new data to generate intelligence and hopefully business advantage in some form.

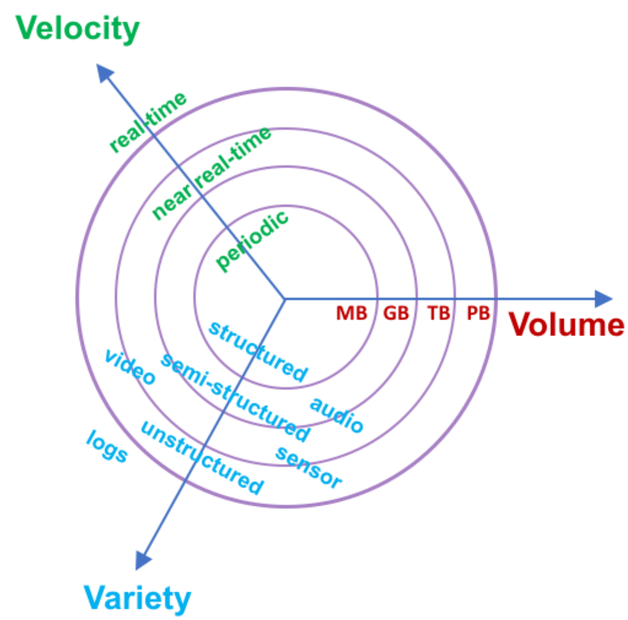
The Modern AI Data Pipeline

In the early days of computing, almost all data came from human input. The data centre was a walled garden, protected from intruders and only accessible through data entry terminals and devices connected directly to the core computing infrastructure. The expansion of the Internet resulted in greater accessibility to new data sources, through publicly exposed websites and then mobile devices.

In the last 20 years, computing itself has moved outside corporate data centres and into the public cloud, in the form of SaaS, PaaS and IaaS solutions. IoT and edge computing has created new data sources, most of which now generate unstructured and machine-generated information. Direct human input now forms only a small part of new content.

Data Sources

Data is being generated in vast quantities everywhere we look. Mobile devices track our movements, creating new content in the form of video and images. Sensors track everything from seismic activity to traffic patterns. Mapping technologies produce images of the earth, population movements and changes in the climate. Medical systems create information on individuals and data that can be used to study populations as a whole.



Data Lifecycle

Information moves through a lifecycle where the usage of that data changes over time. New content may be pre-filtered at the edge, simply to reduce the volume of data stored. As information is used, the value is high, but over time the importance of data diminishes and eventually content moves to being archived or deleted.

At least, that was the traditional view of the information lifecycle. Today, we can see a different approach, where data continues to offer value, long after creation. The process of developing AI models means data gets retained long after creation and could have some future value yet to be conceived.

The Three V's

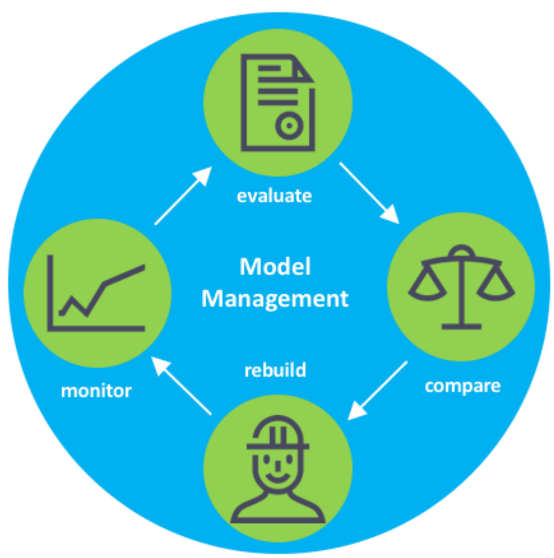
Modern analytics is defined by three vectors, known as the three Vs:

- **Volume** – the amount of data produced by human and machine sources is growing exponentially. Creating value from data requires the analysis of large volumes of information through continued reprocessing. Today we talk in petabytes of content, with exabytes of data on the horizon.
- **Variety** – data derives from many different sources. This can include machine-generated unstructured data, such as logs, semi-structured data including images, videos and data created by medical or seismic platforms or structured data in traditional database applications. Modern AI needs to bring all of this data together.
- **Velocity** – the speed of data processing and analysis. We've already talked about the move towards real-time processing from batch. Other velocity metrics include periodic and near real-time analysis.

The challenge for businesses is how to take these three requirements and translate them into technology and infrastructure. Analytics follows a standard process:

Prepare Data -> Engineer Features -> Train, Test, Build Models -> Deploy Best Performing Model

This is very much the original batch process we discussed earlier in this paper. The challenge in developing any model is that the quality of a model produced through this process risks being immediately outdated by new data. Businesses are moving to a much more dynamic model that results in a "closed loop" process.



The four steps now consist of **monitor**, **evaluate**, **compare** and **rebuild**. This process is continually repeated to iterate over new data and develop more accurate and useful AI models.

Challenges

The challenges of managing a continuously evolving and iterative model are apparent.

Data Sprawl

The creation of large volumes of data leads to inevitable challenges to store and manage information. As edge devices create data, decisions must be made on whether or not to centralise the content. Data may exist across many different and diverse platforms, each with unique performance and security challenges.

Data Growth

Repeating what we've already stated, data volumes are growing at almost unmanageable rates. While only linear improvements in AI model accuracy are gained through the processing of exponential volumes of data, we can expect ever-increasing volumes of information to be retained and somehow stored.

This requirement introduces a conundrum. Data ages over time and traditionally moves from high-performance storage platforms to low-cost media and in many cases, eventually to tape. In a traditional view of data utilisation, this model works well. However, modern data analytics solutions need access to ever-larger data sets that may reside across multiple media types. This requirement breaks the rules on data placement because older data may still be needed for processing on low-latency storage. We will return to the problem in a moment, but it's clear that we require an architecture that can both optimise for cost and utilisation.

Storage in AI Workflows

Now we've established some requirements, how does persistent storage fit into the analytics equation? Data storage provides basic requirements of any computing platform, namely:

- **Persistence** – retention of data when the power is removed. In-memory computing provides low latency and high throughput for ML/AI, but the data still needs to reside somewhere.
- **Resilience** – With multiple petabytes of data to store, solutions must protect against component failure and other forms of data corruption or loss. Resiliency is a table-stakes requirement both at the component and system level.
- **Latency** – ML./AI workloads have specific characteristics that differ from OLTP applications and are sensitive to latency times across the entire data set.
- **Scalability** – this applies to both capacity and performance characteristics.

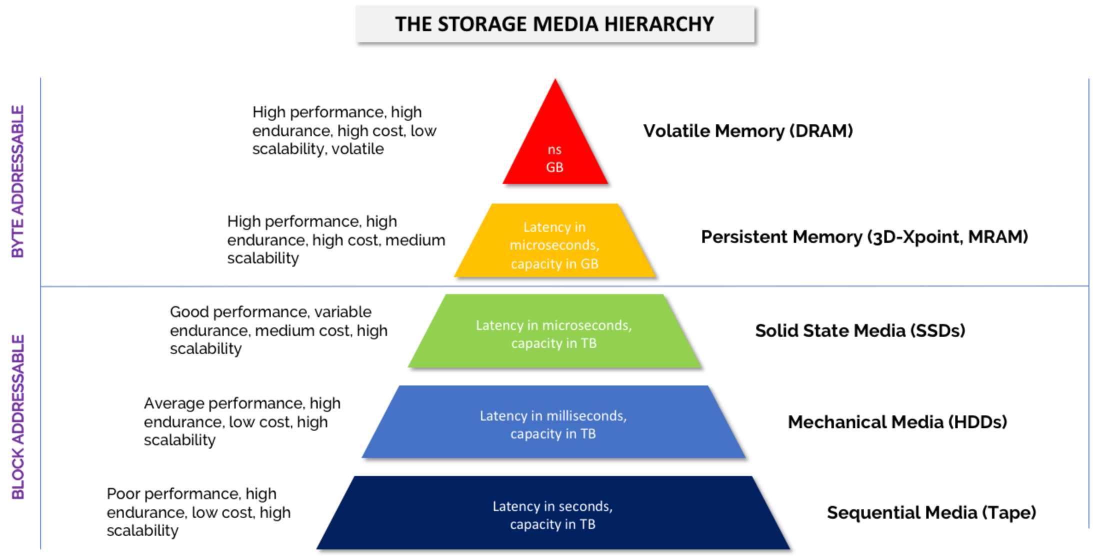
It's worth touching on the final two points in more detail. ML/AI workloads are divided into training and deployment scenarios. In model training, an entire data set will be read and re-read multiple times. It's essential, therefore, that the latency of any individual read request is a low as possible, as this has a direct effect on the time taken to train a model on source data.

From a scalability perspective, the throughput and latency characteristics of the data need to be consistent across the entire data set, as the level of performance achieved will be as quick as the slowest performing part of the data.

Looking back at the requirements of the “3 Vs”, analytics platforms have a challenge to meet the needs of scalability while continuing to deliver low latency across a range of disparate content formats.

Technology Choices

Solid-state storage has become the de-facto standard for high-performance applications. Over the past decade, hard drive technology has evolved to focus on high-capacity archive devices, while SSDs have replaced the performance HDD market. The trend is clear; SSDs offer greater performance and lower latency than any hard drive, while HDDs continue to win the cost/value argument for large volumes of inactive data.



NAND Flash

The NAND flash market itself is rapidly evolving. Vendors have focused on two vectors for improving the value of SSDs.

Bits Per Cell

We are now in the fourth generation of increasing cell bit-count, moving from SLC (single) through MLC (two), TLC (three) and QLC (four). QLC or quad-level cell offers four bits of data per cell, or storage unit. The technology offers great density, but with compromises on both performance and endurance.

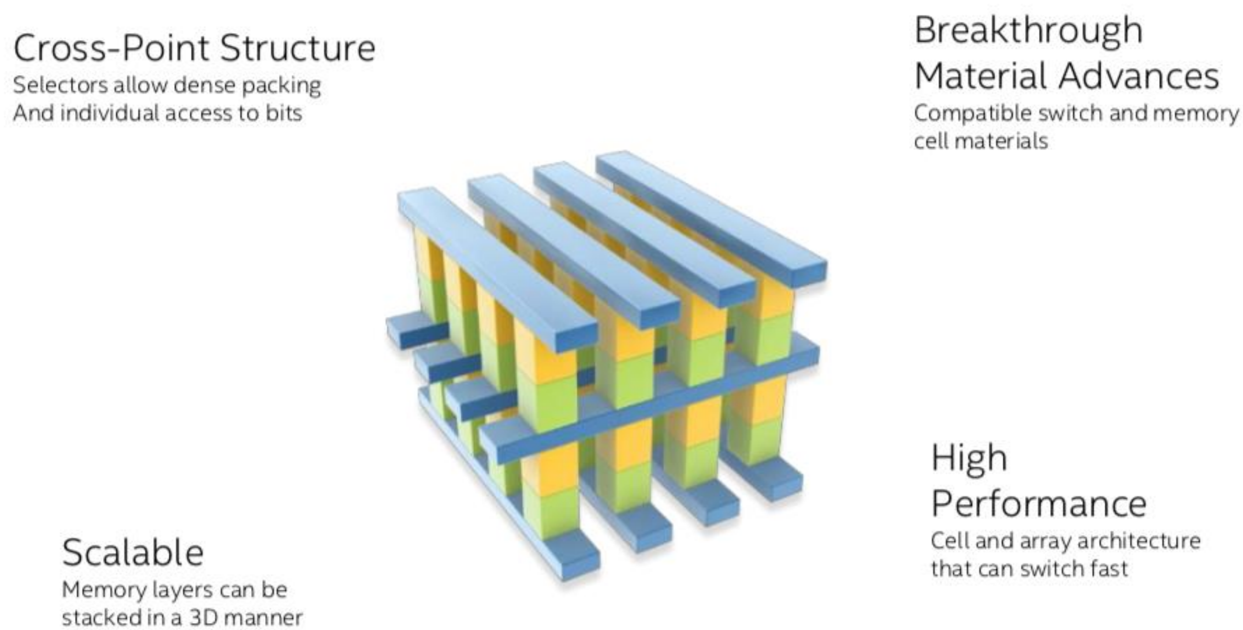
Multiple Layers

The first NAND technology on the market was planar NAND, which stored cells in a two-dimensional grid. Flash manufacturers subsequently developed techniques to increase capacity in the third dimension by layering cells on top of each other in the vertical axis. Intel already offers 64-layer NAND, with 96 and 144-layer products in development.

The NAND industry needs to increase the capacity and reduce the cost of NAND to keep the products competitive in the market. As a result, most vendors are moving the majority of their production towards TLC and QLC designs.

Intel Optane

Optane is the brand-name for 3D-XPoint, a persistent memory technology initially developed jointly by Intel and Micron. Where NAND uses an electrical charge to retain state, 3D-Xpoint uses a change in electrical resistance. The characteristics of 3D-XPoint and NAND are quite different. 3DXPoint delivers much higher endurance, lower latency and greater throughput compared to NAND devices. At this point, QLC NAND has the edge on scalability.



Byte Addressability

3D-XPoint has one other characteristic that differentiates the technology from NAND. In NAND devices, data is written in pages and blocks. By contrast, 3D-XPoint devices provide byte-addressability, the capability to update at the byte level. This feature puts Optane technology on a par with system memory. There are two physical deployment models for Optane that exploit this feature. Optane DIMMs fit right into a traditional memory DIMM slot (subject to processor and motherboard support). Alternatively, Optane deploys as a standard SSD with an NVMe interface.

The DIMM option offers the capability to scale system memory to greater capacities than available with DDR4 DIMMs and at a lower cost. The trade-off is a reduced level of I/O performance. Software developers are already exploiting the capability of Optane DIMMs to increase the capacity of in-memory computing, which has direct benefits for ML/AI workloads.

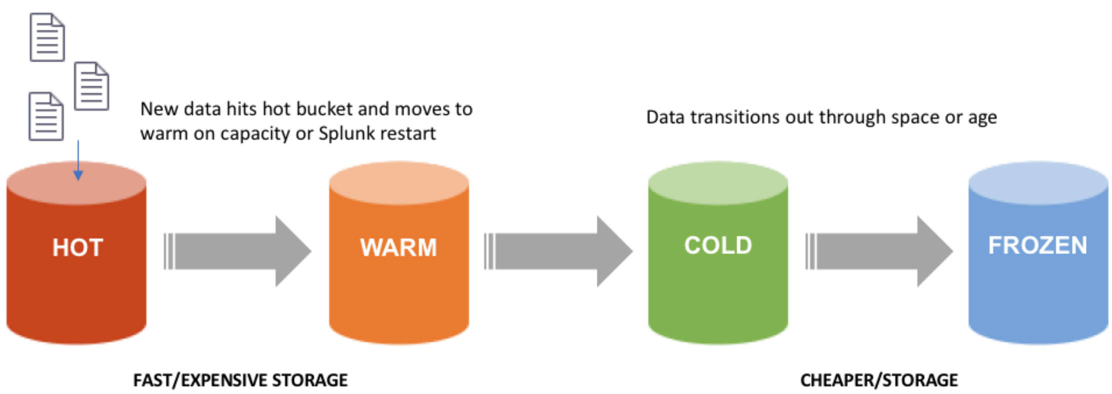
Software Solutions & Trends

Analytics software has moved from batch to real-time processing. Data is now being created and ingested into systems 24x7, with continual analysis. Rather than "collect, load, analyse", new software solutions are designed for continual data ingest and analysis. Features include:

- Parallel processing – data is analysed and processed across multiple concurrent pipelines.
- Real-time – data input and analysis processes effectively operate independently to collate data and provide analysis in real-time.
- Scalability – systems are highly scalable and allow concurrent access to large volumes of data.
- Unstructured content – most of the data analysed day is in unstructured format and stored in file systems or object stores. In the scenarios where structured data is also processed, this can generally be exported as XML or CSV formats.

One aspect of the data processing pipelines to highlight here is the dual nature of storage operations. Ingest processes create large volumes of write I/O as data is stored and indexed. Analysis tasks generate huge amounts of read I/O as data and metadata are searched and analysed.

Solutions need to cater for both types of I/O profile occurring at the same time.



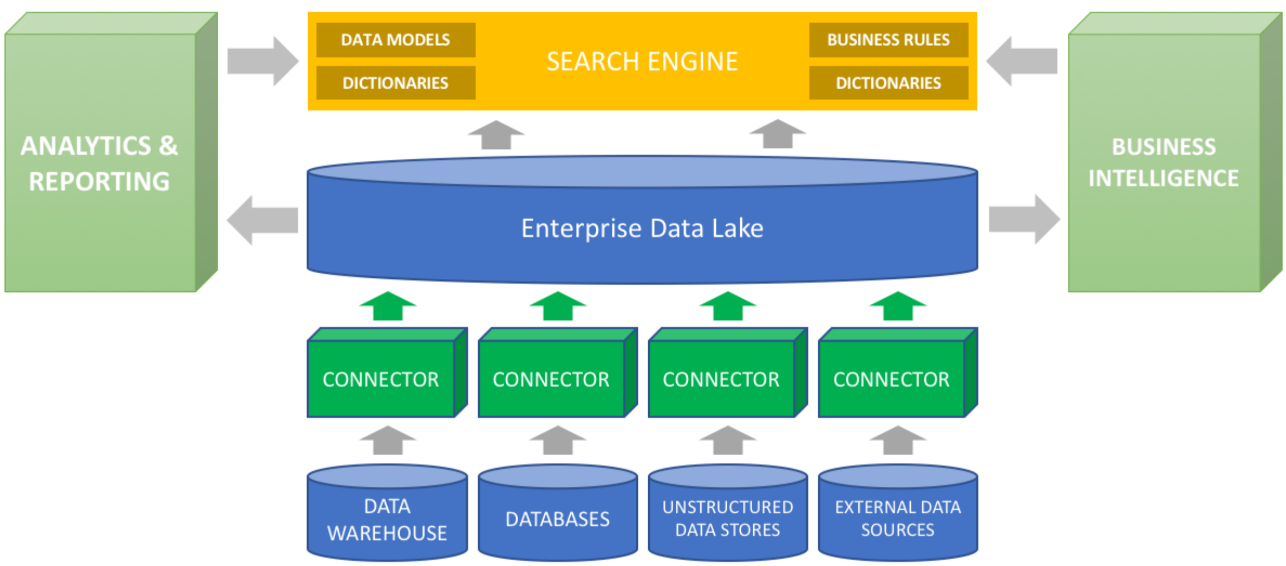
In one model, Splunk uses the concept of buckets to write new and active data. As storage space consumed, active or "hot" buckets move through warm, cold and eventually frozen processes. Data is only ever written to hot buckets, which is aligned to media with high write endurance. As data ages over time, buckets migrate to cheaper media with more read bias, then eventually to archive technology.

Architecture Options

As we've discussed, the requirements of analytics workloads are very different from traditional storage solutions. As enterprises continue to develop new data sources, there's a risk of creating data silos, or isolated islands of multiple data types.

Golden Repository

A single or "golden" repository doesn't have to represent a single platform storing data. Instead, it reflects a single point of truth that identifies where the most current and accurate copy of data resides. The architecture differs from that of a data lake in the sense that the data isn't extracted from other platforms and stored in a single physical repository. While a single logical repository acts as a reference for the location of data, the physical characteristics and requirements of data are still more aligned to the production or primary use of the content.



Data Warehouse

A data warehouse is a structured way in which to store large volumes of data. They are typically built on relational databases with processes for extracting data from primary systems and adding to the warehouse platform. Where primary systems are tuned for OLTP, data warehouses are tuned for search and retrieval functions of long-term archive and non-volatile data. With the increasing volume of unstructured data in the enterprise, data lakes arguably offer more flexibility for bringing together multiple data types.

Data Lakes

Data lakes are proposed as a solution for storing large volumes of data in a single repository, typically in the original native or raw format of the content. A data lake can hold unstructured data in the form of objects and files (for example images and video), semi-structured data such as log files (including standard formats like XML, CSV and JSON) or structured data from traditional databases. Structured content is generally exported into a semi-structured format for processing.

The benefit of a data lake is that it retains all data in a single repository, reducing the risk of data duplication and inconsistencies. All applications now have a single location from which to derive and access content. With data stored in a single place, it's now possible to continue to create and refine metadata used to manage the content. Naturally, with so much unstructured content, search becomes a key tool in making data lakes work efficiently.

Vendor Solutions

Two vendor solutions worthy of discussion are the new platform architectures from VAST Data and StorONE, both of which use Intel Optane and QLC technology.

VAST Data has developed a petabyte-scale unstructured data solution built for performance, scale and cost optimisation. The architecture is divided into two layers. The bulk of data is stored on QLC NAND media, delivering high read I/O performance across many devices. New data is initially committed to Optane media, which is also used to store the majority of metadata. As new data is ingested, it gets coalesced in Optane before being committed to QLC using highly efficient sequential write processes. Data in the VAST platform is accessed through stateless and lightweight I/O controllers that can be scaled to meet application demand.

StorONE has developed a storage platform for block, file and object content. The StorONE S1:AFAn uses Intel Optane and QLC media as two distinct tiers of cache. New and active data is stored on Optane, while inactive data is eventually migrated down to QLC media over time, based on occupancy thresholds on the Optane tier. The high endurance of Optane enables the platform to deliver I/O for the working set of data from the fastest media, while offering fast read access to data that moves from inactive to active status. The result is a system that is capable of delivering over 1 million IOPS from as little as 30TB of capacity.

Both of these solutions have been designed with the specific characteristics of Intel Optane and QLC in mind while addressing the needs of modern data lakes and ML/AI.

Further Reading and Content

The following blog posts are available on the Architecting IT Website:

- What is Intel Optane? (blog, published 8 September 2020)
- Storage Predictions for 2021 and Beyond (Part I – Media) (blog, published 5 January 2021)
- Storage Predictions for 2021 and Beyond (Part II – Systems) (blog, published 11 January 2021)
- When Will Optane SSDs Replace NAND Flash? (blog, published 17 December 2020)
- Revisiting Scale-up vs Scale-out Architectures (blog, published 8 December 2020)
- StorONE S1: AFAn (blog, published 28 August 2020)
- Persistent Memory in the Data Centre (blog, published 16 July 2020)
- VAST Data launches with new scale-out storage platform (blog, published 14 March 2019)
- Moving to Unstructured Data Stores (blog, published 12 May 2020)
- The Expanding Storage Hierarchy (blog, published 28 August 2019)

The following podcasts are available on the Storage Unpacked website:

- #188 – Is Intel Optane Ready for Primetime? (podcast, published 16 January 2021)
- #184 – MCAS – Memory Centric Active Storage (podcast, published 27 November 2020)
- #171 – Exploiting Persistent Memory with MemVerge (podcast, published 28 August 2020)
- #164 – Introduction to StorONE S1: All-Flash Array.next (podcast, published 26 June 2020)
- #105 – Introduction to VAST Data (Part I) (podcast, published 21 June 2019)
- #106 – Introduction to VAST Data (Part II) (podcast, published 26 June 2019)

More Information

For additional technical background or other advice on replication technologies, contact enquiries@brookend.com for more information. Architecting IT is a brand name of Brookend Ltd and independent consultancy, working for the business value to the end customer.   This report was sponsored by Intel and is free to download from the Architecting IT website. Intel have provided technical background and content, but not editorial review.

Email: architectingit@brookend.com
Twitter: @architectingit

The Author

Chris M Evans has worked in the technology industry since 1987, starting as a systems programmer on the IBM mainframe platform. After working abroad, he co-founded an Internet-based music distribution company during the .com era, returning to consultancy in the new millennium. Chris writes a popular blog at http://blog.architecting.it, attends many conferences and invitation-only events and can be found providing regular industry contributions through Twitter (@chrismevans) and other social media outlets.