How to store (petabytes of) machine-generated data

Author: Rainer W. Kaese, Senior Manager Business Development, Storage Products Division, Toshiba Electronics Europe GmbH

The amount of data worldwide grows by several billion terabytes every year because more and more machines and devices are generating data. But where will we put it all? Even in this age of IoT, hard drives remain indispensable.

Data volumes have multiplied in recent decades, but the real data explosion is yet to come. Whereas, in the past, data was mainly created by people, such as photos, videos and documents, with the advent of the IoT age, machines, devices and sensors are now becoming the biggest data producers. There are already far more of them than people and they generate data much faster than us. A single autonomous car, for example, creates several terabytes per day. Then there is the particle accelerator at CERN that generates a petabyte per second, although "only" around 10 petabytes per month are retained for later analysis.

In addition to autonomous driving and research, video surveillance and industry are the key contributors to this data flood. The market research company IDC assumes that the global data volume will grow from 45 zettabytes last year to 175 zettabytes in 2025[1]. This means that, within six years, three times as much data will be generated as existed in total in 2019, namely 130 zettabytes – that is 130 billion terabytes.

Much of this data will be evaluated at the point of creating, for example, in the sensors feeding an autonomous vehicle or production facility (known as edge computing). Here, fast results and reactions in real-time are essential, so the time required for data transmission and central analysis is unacceptable. However, on-site

storage space and computing power are limited, so sooner or later, most data ends up in a data centre. It can then be post-processed and merged with data from other sources, analysed further and archived.

This poses enormous challenges for the storage infrastructures of companies and research institutions. They must be able to absorb a constant influx of large amounts of data and store it reliably. This is only possible with scale-out architectures that provide storage capacities of several dozen petabytes and can be continuously expanded. And they need reliable suppliers of storage hardware who can satisfy this continuous and growing storage demand. After all, we cannot afford for the data to end up flowing into a void. The public cloud is often touted as a suitable solution. Still, the reality is that the bandwidth for the data volumes being discussed is insufficient and the costs are not economically viable.

For organisations that store IoT data, storage becomes, in a sense, a commodity. It is not consumed in the true sense of the word but, like other consumer goods, it is purchased regularly and requires continuing investment. A blueprint of how storage infrastructures and storage procurement models can look in the IoT age is provided by research institutions such as CERN that already process and store vast amounts of data. The European research centre for particle physics is continuously adding new storage expansion units to its data centre, each of which contains several hundred hard drives of the most recent generation. In total, their 100,000 hard disks have attained a total storage capacity of 350 petabytes[2].

The price decides the storage medium

The CERN example demonstrates that there is no way around hard disks when it comes to storing such enormous amounts of data. HDDs remain the cheapest medium that meets the dual requirements of storage space and easy access. By comparison, tape is very inexpensive but is not suitable as an offline medium and is only appropriate for archiving data. Flash memory, on the other hand, is currently still

eight to ten times more expensive per unit capacity than hard disks. Although the prices for SSDs are falling, they are doing so at a similar rate to HDDs. Moreover, HDDs are very well suited to meet the performance requirements of high-capacity storage environments. A single HDD may be inferior to a single SSD, but the combination of several fast-spinning HDDs achieve very high IOPS values that can reliably supply analytics applications with the data they require.

In the end, price alone is the decisive criterion – especially since the data volumes to be stored in the IoT world can only be compressed minimally to save valuable storage space. If at all possible, compression typically takes place within the endpoint or at the edge to reduce the amount of data to be transmitted. Thus, it arrives in compressed form at the data centre and must be stored without further compression. Furthermore, deduplication offers little potential savings because, unlike on typical corporate file shares or backups, there is hardly any identical data.

Because of the flood of data in IoT and the resultant large quantity of drives required, the reliability of the hard disks used is of great importance. This is less to do with possible data losses, as these can be handled using appropriate backup mechanisms, and more to do with maintenance of the hardware. With an Annualised Failure Rate (AFR) of 0.7 per cent, instead of the 0.35 per cent achieved by CERN with Toshiba hard disks, a storage solution using 100,000 hard disks would require that 350 drives are replaced annually – on average almost one drive replacement more per day.

Hard drives will remain irreplaceable for years to come

In the coming years, little will change with the main burden of IoT data storage borne by hard disks. Flash production capacities will simply remain too low for SSDs to outstrip HDDs. To cover the current storage demand with SSDs alone, flash production would have to increase significantly. Bearing in mind that the construction costs for a single flash fabrication facility run to several billion Euros, this is an undertaking that is challenging to finance. Moreover, it would only result in higher

flash output after around two years that would only cover the demand of 2020 and not that of 2022.

The production of hard disks, on the other hand, can be increased much more easily because less cleanroom production is needed than in semiconductor production. Additionally, the development of hard disks is progressing continuously, and new technologies such as HAMR (Heat-Assisted Magnetic Recording) and MAMR (Microwave-Assisted Magnetic Recording) are continuing to deliver capacity increases. Experts assume that HDDs' storage capacity will continue to increase at a rate of around 2 terabytes per year for a few more years at constant cost. Thus, IDC predicts that by the end of 2025, more than 80 per cent of the capacity required in the enterprise sector for core and edge data centres will continue to be obtained in the form of HDDs and less than 20 per cent on SSDs and other flash media[1].

###

Image 1: “Rainer W Kaese Toshiba.jpg”

Rainer W. Kaese, Senior Manager Business Development Storage Products, Toshiba Electronics Europe. (Source: Toshiba Electronics Europe)

Ref. TSH036A_EN

[1] IDC “Data Age 2025” Whitepaper, Update from May 2020
[2] Case Study von Toshiba https://www.toshiba-storage.com/trends-technology/case-study-how-toshiba-hdds-have-helped-cern-keep-track-of-their-generated-data/