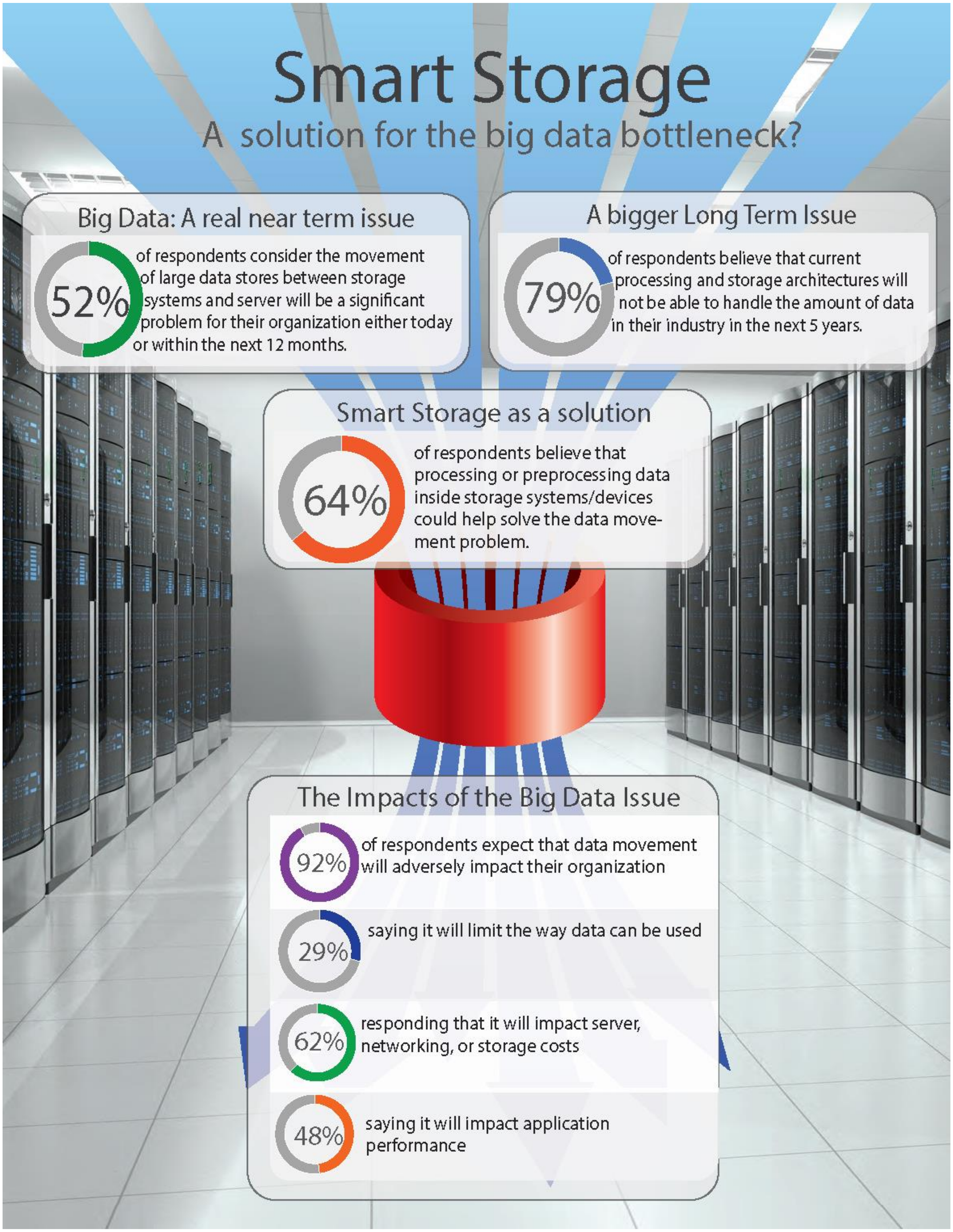
# Large Data Set Analytics, and How Smart Storage Might Reduce the Need for Data Movement

The growth in the use of very large data sets (250 GB or greater) in real-time analytics -- whether it is called Big Data, Artificial Intelligence/Machine Learning (AI/ML), Internet of Things (IoT), or Edge Computing -- is growing rapidly. IDC recently forecasted that the market for Big Data technology and services is expected to grow from $130 billion to $203 billion between 2016 and 2020[1]. Similarly, Statistica projects that the size of the worldwide AI market will grow from $378 million in 2016 to nearly $60 billion in 2025[2]. Cisco's 2015 Global Cloud Index forecasted a compounded annual growth rate (CAGR) of 22% in enterprise database/analytics and IoT between 2015 and 2020, with a growth in the overall data center storage for big data growing from 915 exabytes ($10^{18}$) to 1.8 zettabytes ($10^{21}$) in the same time period. Moreover, Cisco also reported that the data created on devices in 2020 will be 600 zettabytes[3].

One of the requirements that this growth drives is the need to move these very large data stores from storage to processors for analysis. This movement is either from local storage to the CPU complex which is limited by PCI Express bandwidth to 31.5 gigabytes/second (GB/s) for PCIe 4.0, or from storage systems to servers which is limited by network bandwidth to 100 Gigabits per second (Gb/s; roughly 12 GB/s). For data stores that are one petabyte (PB) in size, moving an entire data store takes over 32 seconds between local storage and the CPU complex, or over 85 seconds from a storage system to a server. While these transfer times may not seem long, they severely limit the way that data is used for a variety of real-time analytics applications.

To gauge how much of an issue this is, G2M Research launched a survey in October 2017 across application, storage, and network architects in the Big Data, AI/ML, and IoT markets. The survey, which had a total of 112 respondents, indicated a significant concern with the movement of large data stores between storage systems/devices and servers/CPUs, as shown in the infographic below. The extent of this issue is reflected in the fact that 52% of respondents "consider the movement of large data stores between storage systems and servers a significant problem for their organization," either today or within the next twelve months. Of the respondents who believe that data movement would adversely impact their organization, 48% said it will impact application performance, 29% said it will limit the way that the data can be used, and 62% responded that it will impact server, networking, or storage costs. A staggering 79% of the respondents believe that current processing and storage architectures will not be able to handle the amount of data in their industry in the next five years.

In short, today's CPU- and memory-centric architecture will not be able to carry the metaphorical load of the data avalanche coming our way. The most obvious way to minimize the impact of moving large data sets is to increase the bandwidth of both datacenter networks and internal server busses. However, it is likely that these changes will at best just keep up with the increasing data load. Similarly, spreading the large data sets across clusters of servers can reduce the load times, but only by significantly increasing hardware costs. This has prompted a number of companies and organizations to explore the concept of Smart Storage, where some level of processing/pre-processing occurs in the storage system and/or storage devices themselves.



# Smart Storage

## A solution for the big data bottleneck?

### Big Data: A real near term issue

52% of respondents consider the movement of large data stores between storage systems and server will be a significant problem for their organization either today or within the next 12 months.

### A bigger Long Term Issue

79% of respondents believe that current processing and storage architectures will not be able to handle the amount of data in their industry in the next 5 years.

### Smart Storage as a solution

64% of respondents believe that processing or preprocessing data inside storage systems/devices could help solve the data movement problem.

### The Impacts of the Big Data Issue

92% of respondents expect that data movement will adversely impact their organization

29% saying it will limit the way data can be used

62% responding that it will impact server, networking, or storage costs

48% saying it will impact application performance

Smart Storage has taken two primary forms: Content Addressable Storage (CAS) systems, and in-situ processing within solid state drives (SSDs). CAS, such as the Dell/EMC Centera[4], focus on data archiving for environments where governance/compliance and content authenticity are most important, and where data does not change frequently, and hence are not good candidates for large data set analytics. This is one of the reasons why 64% of respondents believed that "processing or pre-processing data inside storage systems/devices could help solve the data movement problem."

These are exactly the capabilities that in-situ processing in SSDs offer by putting multiple ARM cores in the storage controller. A variety of companies such as NGD Systems[5] (the sponsor of this survey), ScaleFlux[6] and others are actively involved in developing in-situ processing solutions. While none of these systems are generally available today, their built-in computation ability, combined with the ability to put 32 to 64 U.2 SSDs in a single server, promise the ability to store and process a petabyte in as little as 2U of rack space.

Since the advent of digital computers, the IT industry has regularly oscillated between convergence and disaggregation, as well as how specific functionality has been packaged and delivered to those who use it. The movement of processing capabilities into storage media, as represented by in-situ processing in SSDs, represents a new evolutionary path in IT that has been made possible by the solid-state nature of SSDs. While the eventual success of in-situ processing will be determined by a variety of factors, the potential it has to address the Big Data movement problem is significant.

References:

1 – Gil Press, "6 Predictions for the $203 Billion Big Data Analytics Market", Forbes.com, January 20, 2017.

2 – https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues, Statistica.com, 2017.

3 – "Cisco Global Cloud Index: Forecast and Methodology, 2015-2020", 2016.

4 - https://www.emc.com/data-protection/centera.htm

5 - http://www.ngdsystems.com/

6 – http://www.scaleflux.com/